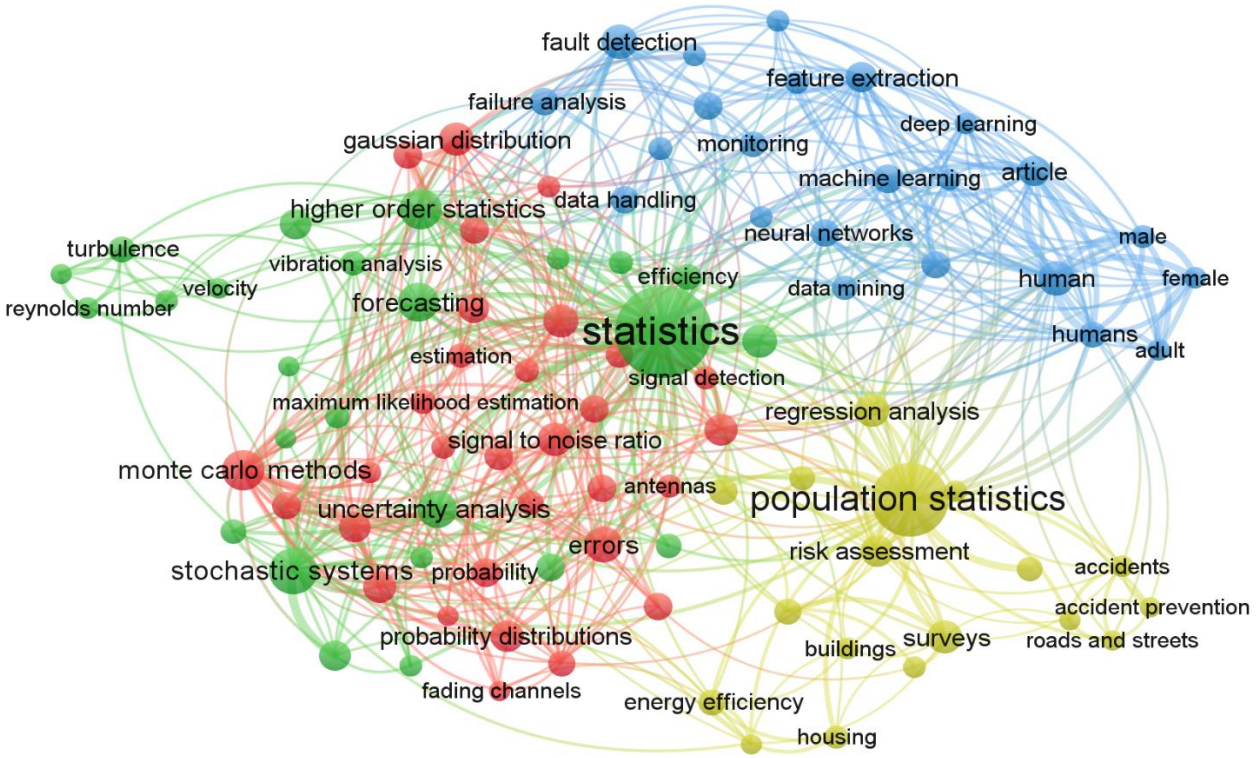

ESTATÍSTICA APLICADA PARA ESTUDANTES DE ENGENHARIA

UM GUIA PRÁTICO

ROGÉRIO CABRAL DE AZEVEDO

CEFET-MG

AZEVEDO ROGÉRIO CABRAL

ESTATÍSTICA APLICADA PARA ESTUDANTES DE ENGENHARIAS

UM GUIA PRÁTICO

CENTRO FEDERAL DE EDUCAÇÃO TÉCNOLÓGICA DE MINAS GERAIS – CEFET-MG

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL – PPGEC

Rogério Cabral de Azevedo, Prof. Dr.
CEFET-MG, Departamento de Engenharia Civil, Programa de Pós-Graduação em Engenharia Civil
http://lattes.cnpq.br/0613519736951060

Janeiro/2021

CAPA: Rogério Cabral de Azevedo (imagem gerada pelo software VOSviewer versão 1.6.15)

ISBN: 978-65-00-15425-2

https://cblservicosprd.blob.core.windows.net/barcode/978-65-00-15425-2.jpeg

CEFET-MG

1ª EDIÇÃO

SUMÁRIO

1 INTRODUÇÃO

Mas afinal, o que é estatística e como podemos nos utilizar de seus conceitos e ferramentas para aprimorar nossas pesquisas de graduação ou pós-graduação? Uma busca em algumas fontes nos fornece diversas definições para o termo "estatística", das quais podemos destacar[1]:

> Um conjunto de técnicas e métodos de pesquisa que, dentre outros tópicos, envolve planejar o experimento a ser realizado, a coleta qualificada dos dados resultantes do experimento, a organização, processamento e análise destes dados, a inferência, ou seja, a capacidade de concluir a partir da análise dos dados processados, a confiabilidade (ou erro) associado a estas conclusões e, por fim, a disseminação das informações.

Esta definição é bem completa, pois abrange todo o cenário no qual um futuro pesquisador está inserido. Envolve como planejar um experimento, como produzir e coletar os resultados do experimento, como organizar, processar e analisar os dados obtidos, como definir ou identificar a confiabilidade ou o erro associado à inferência, e como escolher os métodos mais claros e didáticos para divulgação das informações finais.

O uso da estatística em trabalhos acadêmicos tem aumentado nos últimos anos. Ela tem sido usada principalmente como uma forma de agregar relevância à análise dos resultados obtidos nos experimentos realizados, ao oferecer posições estatisticamente conclusivas sobre esses resultados. A estatística não oferece novos resultados aos experimentos, mas permite que o delineamento dos passos que conduziram aos resultados (o método) seja realizado de forma a separar os fatores de interesse, que foram escolhidos para serem analisados, dos fatores chamados de aleatórios, que embora possuam influência sobre os resultados, devem ser distribuídos de forma a não interferir na análise.

A estatística é fundamental para a análise dos resultados finais de um trabalho acadêmico. Somente apresentar as médias obtidas em um determinado experimento e citar que os resultados obtidos são superiores aos de referência é insuficiente para a realização de uma análise válida, uma vez que diversas medidas influenciam a comparação, como a variância, por exemplo. É mais correto e acadêmico comprovar estatisticamente a existência da diferença e citar, por exemplo que "com 95% de confiabilidade, as médias obtidas no experimento são superiores às médias de referência".

Mas antes de pensarmos em como planejar um experimento ou em como efetuar uma análise estatística dos resultados, é necessário conhecer os conceitos nos quais as ações descritas acima são baseadas. Nesse sentido, alguns conceitos, como os de população, amostra, lote, risco e confiabilidade são fundamentais para o entendimento da estatística.

1.1 População

População representa o conjunto dos todos os elementos, objetos do estudo, que possuem uma ou mais características em comum. Para exemplificar, em uma eleição para presidente, a população seria representada por todos os eleitores habilitados do pais, já para governador, por todos os eleitores habilitados do estado. Dentro das engenharias, o conceito é semelhante: representa todas as peças de mesmo modelo produzidas

[1] www.portalaction.com.br/estatistica-basica

por uma determinada fábrica ou linha de montagem; todo o lote[2] de concreto produzido por uma determinada empresa concreteiras.

1.2 Amostra

Amostra é um subconjunto da população que por ser, na maior parte das vezes, numerosa ou infinita, não pode ser avaliada quantitativamente. O tamanho (quantidade de elementos) da amostra deve ser representativo para o estudo das características de interesse da população. O tamanho e o método de seleção dos elementos da amostra irão depender dos recursos disponíveis e do conhecimento que se tem da população.

Uma dúvida recorrente entre os estudantes de pós-graduação se relaciona aos conceitos de população e amostra. Quando tratamos de coisas concretas, como pessoas, árvores ou carros, o conceito de população e amostra é claro. População representa todas as pessoas, e, nesse grupo, podemos diferenciá-lo ao citar características como faixa etária, sexo, estado civil, estado, cidade ou bairro de residência, dentre outros. Para árvores ou vegetais, temos classificações como ordem, família e gênero. Por fim, para veículos podemos citar características como fabricante, modelo e ano de fabricação.

No entanto, quando se trata dos resultados de experimentos desenvolvidos em pesquisas, a definição fica um pouco mais confusa. Por exemplo, em um experimento abordando a adição de resíduos de construção civil ao concreto, duas variáveis são analisadas: (i) o tipo de resíduo – 3 tipos diferentes (A, B e C) e (ii) o percentual de adição – 4 percentuais (0%, 25%, 50% e 100%). O que define população e amostra neste caso?

Analisando o experimento, concluímos que o mesmo possui 12 composições diferentes, onde o cruzamento entre o tipo de resíduo (três tipos) e o percentual de adição (quatro tipos) resulta nas composições diferentes ($3 \times 4 = 12$) a serem analisadas.

Todo o concreto produzido segundo o método definido para o experimento, com o uso de cada uma das diferentes composições, representa uma população pois possui características diferentes, decorrentes das diferentes composições usadas. Assim, podemos considerar que o experimento produziu 12 populações diferentes.

Claro que dependendo dos objetivos do experimento, os produtos das 12 composições poderiam ser considerados estratos (subgrupos) de uma única população, mas esta consideração não afeta a premissa que queremos expor, de que a população não necessita existir fisicamente, para ser considerada como tal. Basta que possuam características em comum que tornem aquele conjunto único. No caso exposto, todo concreto produzido adotando-se qualquer uma das composições pode ser considerado como população, pois possuem características únicas derivadas de suas composições. Os corpos de prova que foram geradas especificamente para o experimento formam uma amostra dessa população.

Como dito, os corpos de prova produzidos para cada composição representam a amostra. Supondo que, no experimento sejam produzidos quatro corpos de prova por composição, temos 12 amostras compostas por quatro elementos.

[2] Em logística, um lote representa todos os itens produzidos sob as mesmas condições em um determinado período de tempo e com características (físicas, químicas, dimensionais) idênticas. Este conceito é importante para o planejamento de um experimento porque é necessário que todos os materiais, componentes e insumos utilizados para a produção dos corpos de prova a serem testados possuam as mesmas características.

Por meio deste exemplo é também possível compreender como a influência dos recursos disponíveis determina o tamanho da amostra. O senso comum nos leva a crer que quanto maior a quantidade de elementos que compõe a amostra, melhor serão os resultados, o que é uma premissa verdadeira. No entanto, na prática, a premissa é difícil de ser mantida, pois qual a nossa real capacidade em produzir e testar uma grande quantidade de corpos de prova? Teremos material suficiente? Teremos equipamento e tempo suficiente para testar todos os corpos de prova?

O conhecimento da população também é importante. Na hipótese de existir pouco conhecimento dos resultados da adição de determinado tipo de resíduos da construção civil nas propriedades físicas de concretos, amostras com maior número de elementos conduzirão a resultados mais definitivos sobre a população. Contudo, se a literatura já apresenta informações sobre esta influência e queremos apenas comprovar algum direcionamento específico, amostras com menor número de elementos podem ser usadas para esta finalidade.

1.3 Lote

O conceito de lote[2] também é importante para o planejamento de experimentos. No exemplo anterior, se o cimento usado para a produção dos corpos de prova, apesar de serem do mesmo tipo (CP-V, por exemplo) forem oriundos de fabricantes diferentes, suas características físico-químicas podem ter pequenas variações que, por sua vez, podem ter influência nos resultados do experimento.

Os conceitos supracitados, e as formas como eles se correlacionam, demonstram a importância que a amostra (ou a técnica utilizada para sua escolha) possui para a caracterização correta de uma população. Se escolhermos uma amostra de forma errada ou tendenciosa, a inferência (capacidade de transferir para a população como um todo, a análise dos resultados realizada a partir dos dados obtidos com a amostra) é prejudicada, e o trabalho dispendido inutilizado.

1.4 Variáveis

Em estatística, uma variável representa uma característica relativa aos elementos que estão sendo investigados e que nos interessa avaliar em um experimento. De acordo com os valores que essa característica pode assumir (numéricos ou não numéricos), ela pode ser classificada em quantitativa ou qualitativa (abordado no item 2.4Tipos de Variáveis).

Já em relação à um determinado experimento, as variáveis podem ser classificadas em:

Variáveis independentes: são as variáveis que podem ser definidas, controladas, manipuladas e medidas pelo pesquisador em busca de alterações nos valores da variável resposta que está sendo analisada pelo experimento. Também são chamadas de varáveis preditoras ou explicativas.

Variáveis dependentes: são variáveis que podem ser medidas pelo pesquisador e cujos valores dependem do comportamento das variáveis independentes. Normalmente são associadas aos resultados do experimento e, por isso, também denominadas como variáveis resposta.

Variáveis estranhas: são variáveis não controladas nem manipuladas pelo pesquisador e que podem influenciar no comportamento ou na medição das variáveis dependentes. Também conhecidas como ruído, fatores não controláveis, variáveis extrínsecas ou de confusão, seus efeitos devem ser eliminados ou atenuados. Suas principais causas são o viés de seleção, quando as unidades de teste possuem características diferentes entre si (matéria prima de diferentes lotes), variações em fatores não controlados do experimento

(temperatura, umidade, por exemplo) e uso de diferentes equipamentos e/ou instrumentos de medição que podem introduzir alterações na variável dependente.

Uma das principais aplicações da estatística é analisar e predizer o comportamento das variáveis dependentes em função de alterações nos valores das variáveis independentes.

Tratamento: o conceito de tratamento é baseado no cruzamento das variáveis independentes de um experimento. Tratamento representa o conjunto de combinações dos diferentes valores das variáveis independentes que são aplicadas e analisadas em um experimento. No exemplo apresentado no item 1.2 Amostra, duas variáveis independentes, tipo de resíduo e percentual de adição do resíduo, são tratadas e foram identificadas 12 combinações diferentes, resultante do cruzamento dos valores definidos para essas variáveis. Cada combinação diferente representa um tratamento.

1.5 Risco

Antes de abordarmos a questão das técnicas de amostragem, devemos entender o que é o conceito de Risco e como ele se relaciona à composição de uma amostra. O risco relativo à amostragem consiste na margem de erro assumida pelo pesquisador em seu experimento, motivada pelo fato de que a investigação da população é parcial; afinal, a população é investigada a partir de uma amostra, com número de elementos muito inferior ao da população e isto pode gerar conclusões indevidas (risco).

Assim, o risco representa a probabilidade de que as conclusões obtidas a partir da análise da amostra sejam diferentes caso toda a população fosse sujeita ao mesmo procedimento de análise, ou seja, indica a margem de erro assumida na análise. Uma margem de erro de 0,05 indica que há 5% de probabilidade de que a relação entre as variáveis, encontrada na amostra, seja apenas um "acaso feliz" e não seja replicada na população. Assim, se o experimento for repetido várias vezes, pode-se esperar que uma em cada vinte vezes, a relação entre as variáveis em questão seria diferente das observadas nas outras. Uma margem de erro de 5% é considerada como o “limite aceitável” de erro.

1.6 Confiabilidade

O conceito de confiabilidade (margem de acerto) é decorrente do conceito de risco (margem de erro). Se uma determinada análise possui um risco ou uma margem de erro de 5%, isto implica em que a confiabilidade da análise (ou nível de confiança) é de 95%.

Existe também outro tipo de risco a ser considerado, este mais difícil de ser determinado estatisticamente, e que, apesar de não estar associado à amostragem, pode, da mesma forma, conduzir a análises incorretas. Este risco refere-se à adoção de procedimentos inadequados, interpretação errônea de evidências, até mesmo manipulação de resultados. Para evita-los, os conceitos estatísticos relativos ao planejamento do experimento devem ser aplicados e os procedimentos metodológicos adotados devem estar claramente explicitados, permitindo a outros pesquisadores avaliar o método utilizado na pesquisa.

O próximo capítulo apresenta as técnicas de amostragem mais comuns utilizadas em experimentos.

2 TÉCNICAS DE AMOSTRAGEM

Em estudos estatísticos, as técnicas de amostragem referem-se ao modo como selecionamos os elementos de uma população que irão participar de um experimento. Se os elementos de uma amostra não forem selecionados de maneira aleatória, a amostra poderá ser tendenciosa em relação a algum fator e, provavelmente, não representarão a população corretamente.

As técnicas de amostragem podem ser divididas em relação à extração dos elementos e em relação a escolha dos elementos que comporão a amostra.

2.1 Métodos de Extração dos Elementos

A primeira técnica para composição de uma amostra refere-se à **extração dos elementos** que a comporão. A extração dos elementos pode ser realizada com ou sem reposição.

Extração sem reposição: quando um elemento sorteado ou escolhido para compor a amostra não puder ser reposto à população e assim, correr o risco de ser "escolhido" novamente. A extração sem reposição é comum quando se realizam ensaios destrutivos (onde o elemento tem suas características alteradas pelo próprio ensaio).

Extração com reposição: quando um elemento sorteado ou escolhido para compor a amostra pode ser reintegrado à população e, assim, ser sorteado novamente. Neste método, como o elemento é reposto, o método não afeta a probabilidade de retirar qualquer elemento da população, ou seja, as chances serão iguais para sempre.

2.2 Métodos para a Escolha dos Elementos

Quanto à **escolha dos elementos** da amostra, esta pode ser probabilística ou não probabilística. No método Probabilístico cada elemento da população possui determinada probabilidade de ser selecionado para compor a amostra (em geral, a mesma probabilidade). No método não probabilístico há uma escolha deliberada ou direcionada dos elementos que irão compor a amostra.

Os principais **Métodos Não Probabilísticos** são:

Amostragem Acidental: A amostra é composta por elementos que vão "aparecendo" ou pelos elementos que são possíveis de se obter até que se complete o número de elementos da amostra. Esse método é comum, por exemplo, em pesquisa de opinião, nas quais os entrevistados são acidentalmente escolhidos; ou em linhas de produção, onde os elementos são retirados da linha para testes na medida que o teste anterior é finalizado, e enquanto o número de testes previstos não for atingido.

Amostragem Intencional: A amostra é composta por elementos escolhidos por meio de critérios predeterminados, ou seja, escolhe-se intencionalmente um grupo de elementos que irão compor a amostra.

Amostragem por Cotas: Neste caso, a população é classificada em estratos (subgrupos), sendo a definição dos estratos estabelecida em função de propriedades relevantes para a característica a ser estudada. O processo de seleção dos elementos que integram os estratos deve ser previamente estabelecido.

Os principais **Métodos Probabilísticos** são:

Amostragem Aleatória Simples: Nesta técnica de amostragem, cada elemento da população possui uma chance igual e maior que zero de ser selecionado para compor a amostra. Ela é chamada de aleatória porque

a seleção dos elementos é feita sob a forma de sorteio não sendo utilizado nenhum critério ou filtro no processo de seleção. O único problema em relação a este método é que, por ser aleatório, qualquer combinação dos elementos presentes na população pode ser gerada e, com isto, determinada característica desta população pode ser priorizada.

Amostragem Aleatória Estratificada: Para minimizar o problema relatado na amostragem aleatória simples, a população pode ser dividida de acordo com propriedades de interesse para a característica estudada (estratos) e, dentro destes dos estratos, é realizada a amostragem aleatória simples. Há dois tipos de amostragem aleatória estratificada. No primeiro, as amostras parciais retiradas aleatoriamente de cada estrato possuem o mesmo tamanho (amostras com a mesma quantidade de elementos, independentemente do tamanho do estrato). Por sua vez, no segundo método, as amostras parciais possuem tamanho proporcional ao tamanho do estrato. É bem semelhante a amostragem por cotas (não probabilística), mas neste caso, a seleção dos elementos é aleatória.

Amostragem Sistemática: É um tipo de amostragem aleatória simples, com a diferença que os elementos da população são agrupados e ordenados segundo algum critério que não possui influência na característica de interesse. Desta forma, a existência da ordenação facilita o processo de seleção dos elementos. Por exemplo, se temos 50 grupos de 50 elementos e desejamos compor uma amostra de 100 elementos, é possível definir dois números de ordem aleatórios (entre 1 e 50), como por exemplo, o 13º e 27º e assim selecionar estes de cada um dos grupos.

Amostra por Conglomerados: É uma técnica de amostragem realizada em duas ou mais etapas. Na primeira etapa, os grupos ou conglomerados são definidos de acordo com suas características e são sorteados elementos destes conglomerados para representar o próprio conglomerado. Esta etapa pode ser recursiva (grupos dentro de grupos). Na última etapa, são sorteados os elementos que serão testados. É muito utilizada em pesquisas eleitorais com a definição de diversos grupos (cidades, de acordo com seu tamanho ou importância; bairros, de acordo com renda ou situação; e por fim, eleitores dentro de cada grupo escolhido).

2.3 Definição da Amostra

A definição da amostra pode parecer trivial dentro de um experimento, mas seus conceitos são fundamentais para que as conclusões obtidas pela análise dos resultados possam ser transferidas para a população (Inferência).

Podemos entender a importância do processo de amostragem a partir de uma situação bem simples. Em um determinado experimento, foram adquiridos diversos insumos a serem utilizados em um processo construtivo. Estes insumos precisam ser caracterizados. Um deles, agregado fino (areia) foi recebido (doação) em caçambas. Não se tem informações sobre a origem nem sobre a forma de carregamento do insumo na caçamba, mas é necessário caracterizar o insumo para o experimento. Qual o método de amostragem mais adequado para a caracterização?

Antes de propormos uma solução para o problema, é necessário entender corretamente o problema. Diversas questões devem ser esclarecidas para que a solução adotada seja adequada:

- O experimento possui restrições (ou especificações) quanto ao agregado fino, tipo granulometria ou outra?;
- Qual o volume (ou peso) total necessário para o experimento?;
- Quantas caçambas foram entregues?;
- O conteúdo de uma caçamba é suficiente para o experimento?

Bem, vamos supor que existam especificações quanto às características físicas do agregado fino e que foram entregues duas caçambas, sendo que o conteúdo de apenas uma é suficiente para o experimento. Neste caso, como não é possível determinar a origem do conteúdo de cada caçamba e nem como foram preenchidas (pode ser que tenham sido preenchidas com agregados finos com diferentes características físicas), o problema reside em escolher qual das caçambas possui o agregado mais adequado (características físicas) ao experimento, o que nos obriga a testar as duas.

Então, vamos ao método de amostragem. Qual o método mais adequado para esta situação? E, dadas as incertezas presentes, vai haver um 100% correto? Agora entra a questão dos recursos disponíveis (tempo, equipamento e pessoal para testes, recursos financeiros, etc.). Vamos supor que os recursos disponíveis permitam a realização de seis caracterizações, ou seja, podem ser testadas seis amostras. Como escolher seis amostras que representem o conteúdo das duas caçambas?

A primeira parte é mais simples: três amostras para cada caçamba. A segunda parte, de onde retirar as três amostras em cada caçamba, pode ser mais complexa. Pode ser aplicada uma amostragem aleatória estratificada. Dividir o volume da caçamba em 3 estratos verticais, de acordo com a altura da caçamba, e cada um dos estratos, em 4 áreas horizontais, como demonstrado na Figura 1. Um sorteio aleatório de uma das quatro áreas em cada estrato vertical poderia gerar o resultado indicado (áreas 1, 6 e 11). O mesmo processo é repetido na segunda caçamba gerando assim as seis amostras para caracterização.

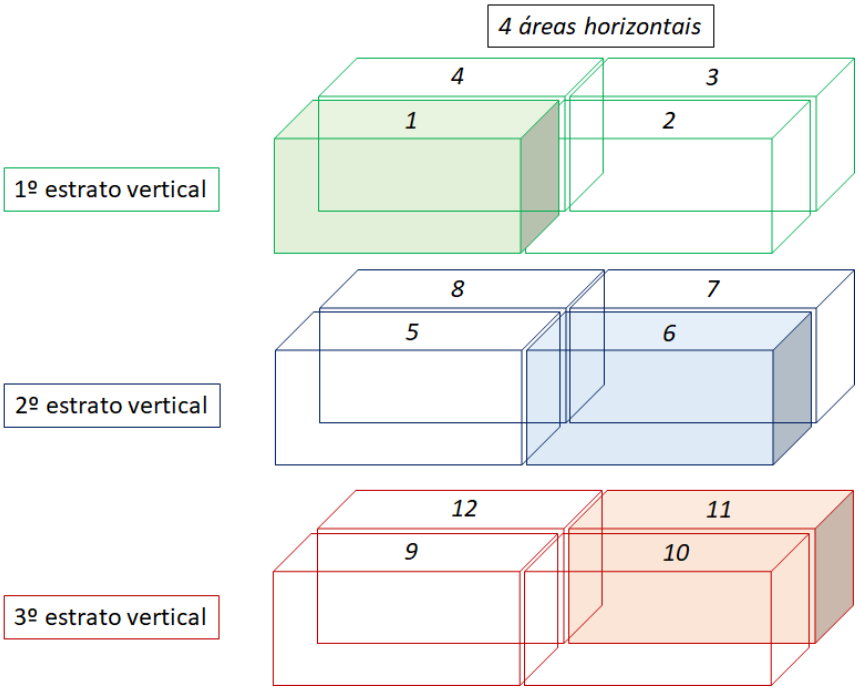

Figura 1 - Exemplo de amostragem estratificada aleatória

Outro processo aleatório válido para a seleção das três amostras seria o sorteio de três das doze áreas, independente do estrato. Isto poderia ocasionar a seleção de mais de uma área por estrato, mas não invalida o método, uma vez que não temos informações sobre a origem do conteúdo de cada caçamba e nem como as mesmas foram preenchidas.

Em princípio, a análise das características físicas de cada amostra indicaria a caçamba mais adequada. Mas tudo vai depender dos resultados obtidos nas caracterizações. Um dos resultados possíveis é que exista uma caçamba cujo conteúdo seja mais adequado ao experimento, devido às características físicas de seu conteúdo. Mas e se todas as amostras das caçambas indicarem diferentes características físicas e todas estiverem dentro dos limites estabelecidos para o experimento? Estas pequenas diferenças irão influenciar o experimento? É

provável que sim, restando ao pesquisador escolher uma das caçambas e adotar procedimentos para homogeneizar seu conteúdo.

2.4 Tipos de Variáveis

Mas quais os tipos de dados que podem ser obtidos a partir de uma amostra? Para compreendermos os tipos de informações que podemos coletar a partir de uma amostra, primeiramente temos que caracterizar o que é um dado e o que é uma variável.

Dado é uma informação coletada e registrada de um elemento da população ou amostra referente a uma variável. Desta forma, por exemplo, o diâmetro da peça selecionada como amostra é um dado, bem como todas as medidas referentes à variável estudada que sejam coletadas na população ou amostra. Podemos entender que o dado representa uma única mensuração ou valor de uma característica de interesse.

Variável é uma característica que pode ser observada (ou medida) em cada elemento de uma população ou em uma amostra desta população. As variáveis assumem valores diferentes em unidades diferentes, associadas à característica que está sendo medida, como por exemplo: diâmetro em mm, peso em quilogramas, resistência a compressão em MPa, etc. Assim, podemos entender que a variável é a característica de interesse que está sendo mensurada na amostra ou população e é representada pelo conjunto de valores mensurados.

As variáveis podem assumir dois tipos básicos: qualitativas e quantitativas, como mostrado na Figura 2. O tipo da variável define a escolha básica da técnica estatística e das interpretações dos resultados.

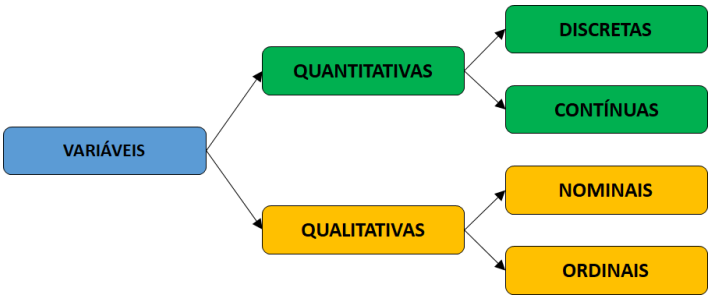

Figura 2 - Tipos de variáveis

Variáveis Qualitativas: correspondem a características que podem ser observadas ou identificadas na população em estudo. Normalmente possuem valores discretos. As variáveis qualitativas são divididas em Nominais e Ordinais. As variáveis qualitativas nominais não possuem ordenação própria, como por exemplo estado civil, cores, cidade ou estado de nascimento. Já as variáveis qualitativas ordinais possuem uma ordem natural pela qual podem ser ordenadas, como classificações de julgamento (péssimo, ruim, regular, bom, muito bom e ótimo). As escalas de likert fazem parte deste tipo de variável e merecem um destaque a parte devido ao seu uso frequente.

Escala de likert: é um tipo de escala de resposta psicométrica usada habitualmente em questionários e é a escala mais usada em pesquisas de opinião. A escala representa a concordância do entrevistado com a afirmação contida na questão. Um tipo comum de escala de likert é: 1 – Discordo totalmente; 2 – Discordo parcialmente; 3 – Indiferente; 4 – Concordo parcialmente; 5 – Concordo totalmente. O problema em relação a estas escalas é o uso indevido do numeral associado à opinião do entrevistado (1, 2, 3, ...). Algumas pesquisas utilizam este numeral (que na realidade representa um valor qualitativo – a opinião do entrevistado) para operações matemáticas como médias, o que é incorreto. Escalas ordinais podem ser utilizada apenas para

operações matemáticas de frequência, contagem, mediana e moda. O Quadro 1 exibe exemplos de escalas de likert.

CONCORDÂNCIA	FREQUÊNCIA	IMPORTÂNCIA	PROBABILIDADE
Concordo totalmente	Sempre	Muito importante	Quase sempre verdade
Concordo	É frequente	Importante	Geralmente verdade
Nem concordo, nem discordo	É ocasional	Moderado	As vezes é verdade
Discordo	É raro	Pouco importante	Geralmente falso
Discordo totalmente	Nunca	Não é importante	Quase sempre falso

Quadro 1 - Escalas de Likert

Variáveis Quantitativas: correspondem a características que podem ser mensuradas na população em estudo. Podem possuir valores discretos ou contínuos. As variáveis quantitativas são ditas discretas quando podem assumir apenas determinados valores do conjunto sendo normalmente associadas a contagens (quantidade); e são ditas contínuas quando podem assumir qualquer valor dentro do conjunto, sendo normalmente associadas a medições (peso, resistência).

A principal diferença entre variáveis qualitativas e quantitativas pode ser vista pelos resultados de sua mensuração. Variáveis qualitativas, por refletirem opiniões, podem obter diferentes mensurações de diferentes observadores sobre o mesmo elemento. Por exemplo, se questionado sobre a importância de um fato, um respondente pode optar pela resposta “muito importante” enquanto outro por “moderado”. São opiniões diferentes sobre o mesmo fato.

Já as variáveis quantitativas, por refletirem medições e não opiniões, devem sempre apresentar o mesmo resultado sempre, excetuando-se diferenças por precisão dos equipamentos de medição. Por exemplo, se dois observadores forem convidados a contar a quantidade de alunos em uma sala de aula em um dado instante, a resposta (quantidade de alunos) deve ser a mesma. Em um outro exemplo, se dois pesquisadores diferentes efetuarem a medição do peso de um determinado corpo de prova em uma mesma balança, o resultado deve ser o mesmo (considerando erros de leitura, precisão da balança e manutenção da integridade do corpo de prova).

Uma vez compreendidos os conceitos de população, amostra, variável e dados, bem como a importância do processo de amostragem, o próximo passo é conhecer os números que resumem e descrevem o conjunto de dados (amostra).

A Estatística Descritiva é usada para descrever os dados que representam a amostra. Inicialmente, os principais conceitos serão apresentados considerando apenas amostras onde todas as observações são conhecidas, ou seja, a variável de interesse foi determinada (observada) para cada elemento da amostra. Amostras cujos valores estão agrupados em classes (representados graficamente por histogramas) não serão tratadas por ora.

3 ESTATISTICA DESCRITIVA

A estatística descritiva é um ramo da estatística que aplica várias técnicas para descrever e sumarizar um conjunto de dados, seja referente a uma amostra ou a uma população. Diferencia-se da inferência estatística, ou estatística indutiva, pelo fato de organizar e sumarizar os dados ao invés de usar os dados em um processo de aprendizado sobre a população.

A estatística descritiva é composta por uma série de medidas básicas que apresentam uma análise descritiva de como os dados estão organizados. São as medidas de posição, medidas de dispersão, quartis, coeficiente de assimetria e coeficiente de curtose.

3.1 Medidas de Posição

As medidas de posição são valores que representam a tendência de concentração (ou distribuição) dos dados observados em relação à característica de interesse. A forma mais usual de representação da tendência de concentração é o gráfico de distribuição de frequência, que apresenta, no eixo horizontal, os valores ou classes agrupadas da característica de interesse e, no eixo vertical, a frequência associada ao valor ou classe.

As medidas de posição mais importantes são a média aritmética, a mediana e a moda.

A **Média Aritmética** ou simplesmente média, pode se referir a população (média populacional - μ) ou a amostra (média amostral - $\bar{x}$) e é calculada pela divisão da soma dos valores observados (x) pela sua quantidade (n). A média retrata a posição central dos valores das observações, mas não apresenta informações sobre sua dispersão.

Amostra A	1	2	3	4	5	6	7	Média $\bar{x}$
Valor	97	98	99	99	99	100	101	99

Amostra B	1	2	3	4	5	6	7	8	Média $\bar{x}$
Valor	90	95	97	97	99	103	105	106	99

Tabela 1 - Valores de amostras e suas respectivas médias

A Tabela 1 apresenta duas amostras ordenadas, uma com sete elementos (A) e outra com oito (B) representando uma característica física destes elementos (comprimento, por exemplo). Ambas possuem média igual a 99, mas os dados da amostra B possuem uma faixa de variação muito maior. A faixa de variação ou amplitude da amostra A é 4 (amplitude é definida como a diferença entre o maior e o menor valor do conjunto ou amostra) enquanto a amplitude da amostra B é 16.

A **Mediana** é uma medida de posição que indica o ponto central dos valores ordenados, ou seja, é o valor que divide um conjunto de dados (ordenados) em duas partes com a mesma quantidade de dados. Se a amostra possui número de observações ímpar, a mediana será a observação central. Se o número de observações for par, a mediana será a média aritmética das duas observações centrais. Para a amostra A (Tabela 1), a mediana é o valor da quarta observação (ponto central, mediana = 99). Já para a amostra B, a mediana é dada pela média aritmética entre os valores da quarta e quinta observações (mediana = (97 + 99) / 2 = 98).

A **Moda** de uma amostra é o valor com maior frequência (número de ocorrências) na amostra. Na amostra A, o valor mais frequente é (99). Assim, a moda desta amostra é igual a 99. Para a amostra B, o valor da moda é 97, pois este é o valor mais frequente. Caso não exista um valor mais frequente (todos os valores das

observações são diferentes, o conjunto é dito “amodal”. Da mesma forma, podem existir amostras com mais de uma moda, quando dois ou mais valores possuem o mesmo número de observações (superior a um).

Além destas três medidas de posição descritas acima, temos as medidas de separação (ou separatrizes) que são valores que ocupam determinadas posições em uma distribuição de frequência: são os quartis, decis e percentis. Os **quartis** dividem uma distribuição de frequência (relação ordenada de observações) em quatro partes iguais, como pode ser visualizado na Figura 3. Podemos notar que o segundo quartil (Q2) corresponde a mediana da distribuição.

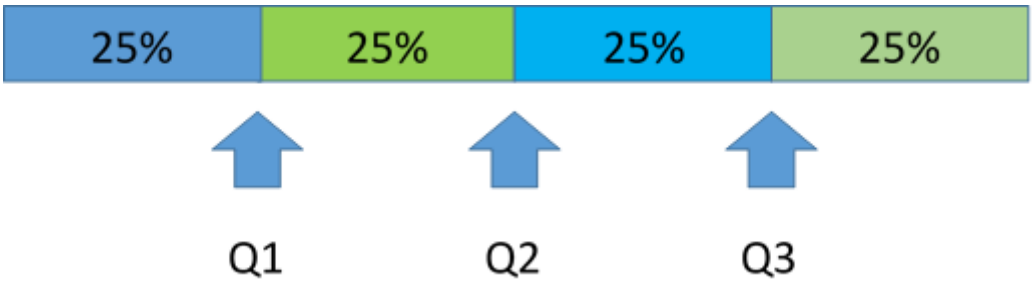

Figura 3 - Quartis

Da mesma forma, os **decis** dividem a distribuição de frequência em 10 partes iguais e os **percentis** em 100 partes iguais.

Como pode ser entendido, as três medidas, média, mediana e moda são medidas de tendência central, pois apontam para três pontos de centralização das observações obtidas. No entanto, elas não demonstram a distribuição dos valores das observações (muito concentrados ou pouco concentrados). Para analisarmos a distribuição dos valores das observações temos as medidas de variabilidade ou de dispersão.

3.2 Medidas de Dispersão ou Variabilidade

As **Medidas de Variabilidade** são medidas estatísticas utilizadas para avaliar o grau de variabilidade ou dispersão dos valores das observações em torno de sua média. Elas são utilizadas para medir a representatividade da média. São elas:

Amplitude: A amplitude (*R*) é o resultado da diferença entre o maior e o menor valor do conjunto de dados. Considerando o conjunto ordenado de dados

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}$$

Temos que a amplitude é dada por $R = X_{(n)} - X_{(1)}$

Variância: A variância (amostral – S^2 ou populacional – σ^2) é a medida de dispersão definida como a média do quadrado dos desvios dos elementos em relação à média. O cálculo da variância considera mais os valores extremos que os valores intermediários, expressando o quanto estes valores estão distantes (dispersos) de sua média. A fórmula da **variância populacional** é:

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} \qquad \textit{Eq. 1}$$

Onde N representa o tamanho da população e μ a média populacional.

Quando tratamos de amostras (parte da população), a média populacional (μ) é substituída pela média amostral ($\bar{x}$) e o tamanho da população (N) pelo tamanho da amostra menos um (*n – 1*). Isto porque ao utilizarmos a média amostral como estimador da média populacional para calcularmos a variância amostral, perdemos 1 grau de liberdade[3] em relação à variância populacional. A fórmula da **variância amostral** é:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad \textit{Eq. 2}$$

Desvio Padrão: Sendo a variância uma medida calculada com valores ao quadrado, seu uso causa uma certa camuflagem dos valores (pois aumenta a medida de dispersão), dificultando um pouco o entendimento. Uma alternativa para solucionar este problema de entendimento é o desvio padrão. O desvio padrão é dado pela raiz quadrada da variância. Assim, o **desvio padrão populacional** é dado por:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}} \qquad \textit{Eq. 3}$$

E o **desvio padrão amostral** é dado por:

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \qquad \textit{Eq. 4}$$

Retomando o exemplo da Tabela 1 (amostras A e B, ambas com média igual a 99) e calculando a amplitude, temos que a amplitude R da amostra A é 4 e da amostra B é 16. Isto demonstra uma maior variação dos valores extremos na amostra B, mas não diz nada sobre o restante das observações da amostra.

Já para a variância, cujo cálculo inclui todas as observações da amostra, temos que a variância amostral de A é igual a 1,67 e a de B igual a 29,43. A variância amostral de B é cerca de 18 vezes maior que a de A (lembre-se que são valores elevados ao quadrado). Isto demonstra uma dispersão dos valores das observações na amostra B muito maior que na amostra A.

Agora, se compararmos o desvio padrão amostral (recordando, igual à raiz quadrada da variância), o da amostra A é 1,29 e o de B é igual a 5,42. Este valor nos apresenta uma medida de dispersão mais próxima dos valores encontrados nas observações, principalmente quando os comparamos com a amplitude R.

A amplitude da amostra A (diferença entre o maior e menor valor da amostra) é 4 e o desvio padrão amostral (raiz quadrada da média do quadrado dos desvios dos elementos em relação à média) é 1,29. Para a amostra B, a amplitude é 16 e o desvio padrão amostral é 5,42. Se fossemos comparar com a variância amostral de B (29,43) teríamos um valor superior a diferença entre o maior e o menor valor das observações na amostra B.

[3] Graus de liberdade de um conjunto de valores representa a quantidade de elementos que podem ter seus valores alterados após terem sido impostas certas restrições a todos os valores. Por exemplo, se a soma de cinco valores é igual a 100, podemos definir os valores de quatro deles, mas o quinto deve obedecer a restrição da soma ser igual a 100. Então temos quatro graus de liberdade para a definição dos cinco valores.

Mas, de qualquer forma, ambas são medidas de dispersão que indicam o quanto os valores observados se distanciam de sua média. No exemplo dado, esta comparação ficou mais fácil, pois as médias das duas amostras são 99. Com as médias iguais, o maior desvio padrão amostral indica a maior dispersão de valores.

E em casos nos quais as médias são diferentes? O desvio padrão e a variância são bastante afetados pela magnitude dos dados e, portanto, pode não oferecer uma medida consistente quando desejamos comparar amostras com médias diferentes, como no exemplo da Tabela 2. Nela são apresentadas quatro amostras com médias bem distintas entre si. Como avaliar qual das amostras possui observações mais coesas?

Amostra	$\bar{X}$	S^2	S
C	10,59	2,53	1,59
D	42,85	26,42	5,14
E	108,21	141,61	11,90
F	321,88	256,32	16,01

Tabela 2 – Média, variância e desvio padrão de amostras

Neste caso, a utilização do **Coeficiente de variação** (CV) apresenta-se como a solução ideal, pois ele oferece uma medida de comparação para a variabilidade de diferentes conjuntos de dados e é definido como a razão entre o desvio padrão e a média (tanto amostrais quanto populacionais):

$$CV = \frac{s}{\bar{x}}100\% \ ou \ CV = \frac{\sigma}{\mu}\ 100\% \qquad \textit{Eq. 5}$$

Assim, para verificarmos qual das amostras possui maior uniformidade entre os valores de suas observações (menor dispersão dos valores em torno da média), basta acrescentar o coeficiente de variação à Tabela 2. Assim, na Tabela 3, podemos ver que a amostra F possui o menor coeficiente de variação (5%) indicando maior concentração das observações em torno da média. Por sua vez, a variância amostral é de 256,32, mostrando que tanto a variância quanto o desvio padrão são afetados pela magnitude dos dados. A amostra mais dispersa, para a qual estes valores mais se afastam de sua média, é a amostra C, cujo coeficiente de variação é 15%.

Amostra	$\bar{X}$	S^2	S	CV
C	10,59	2,53	1,59	15,0%
D	42,85	26,42	5,14	12,0%
E	108,21	141,61	11,90	11,0%
F	321,88	256,32	16,01	5,0%

Tabela 3 - Coeficiente de variação de amostras

O **Coeficiente de Assimetria** é outra medida de dispersão. Ele é usado para distinguir as distribuições assimétricas. Um resultado negativo indica que a cauda do lado esquerdo da distribuição de frequência é maior que a do lado direito. Um resultado positivo para o coeficiente de assimetria indica que a cauda do lado direito é maior que a do lado esquerdo. Um valor nulo indica que os valores são simétricos, ou seja, distribuídos de maneira relativamente iguais em ambos os lados da média (o que não implica necessariamente em uma distribuição simétrica). A Figura 4 ilustra o coeficiente de assimetria.

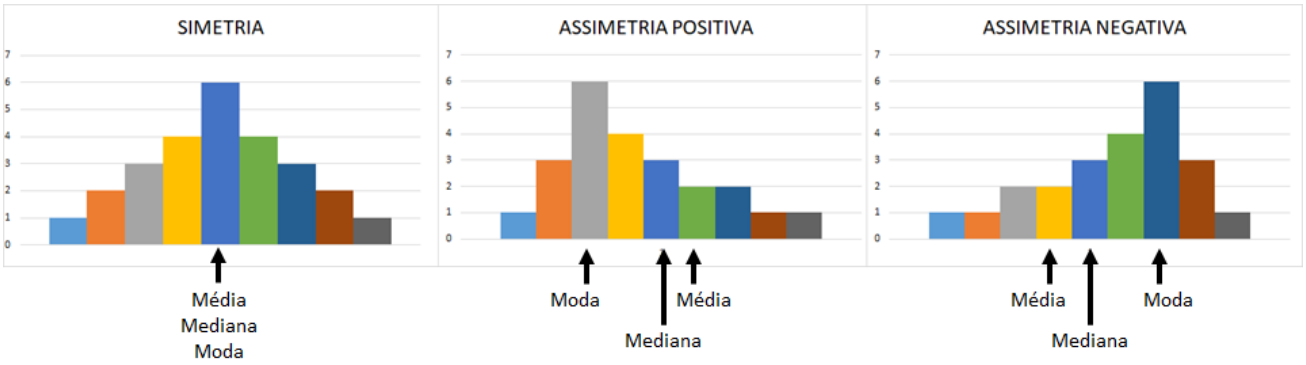

Figura 4 - Coeficiente de assimetria

O coeficiente de assimetria (b_1) é calculado pela fórmula:

$$b_1 = \frac{1}{n}\sum\left[\frac{x_i - \bar{x}}{s}\right]^3 \qquad \textit{Eq. 6}$$

A **Curtose** (b_2) é uma medida de dispersão que caracteriza o achatamento da curva de distribuição de frequência e é dada pela fórmula:

$$b_2 = \frac{1}{n}\sum\left[\frac{x_i - \bar{x}}{s}\right]^4 - 3 \qquad \textit{Eq. 7}$$

Se b_2 = 0, então a função de distribuição tem o mesmo achatamento da distribuição normal[4] e a função é chamada de mesócurtica.

Se b_2 > 0, a função de distribuição possui a curva da função de distribuição mais afunilada com um pico mais alto do que a distribuição normal e é chamada de leptocúrtica.

Se b_2 < 0, a função de distribuição é mais achatada do que a distribuição normal e é chamada de platicúrtica. As curvas que ilustram a curtose são mostradas na Figura 5.

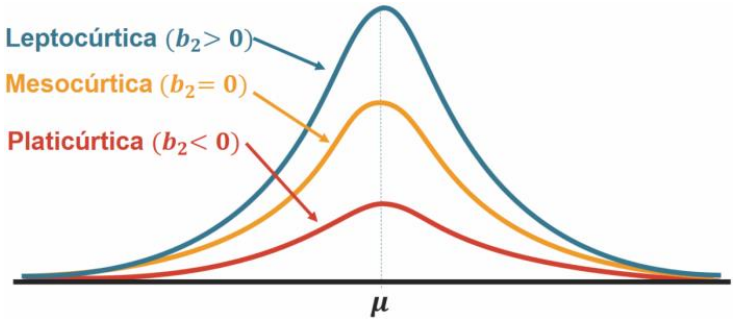

Figura 5 – Curtose – Fonte: PORTALACTION (2020)

[4] A distribuição normal é uma das distribuições de probabilidade mais utilizadas para modelar fenômenos naturais.

3.3 Gráficos

Tão importante quanto conhecer as medidas que representam uma amostra ou população é a forma de apresentação destes valores, ou seja, como apresentar estas informações ao leitor. Os gráficos estatísticos são formas de apresentação dos dados estatísticos cujo objetivo principal é transmitir ao público, de forma simples, clara e objetiva, as informações relativas ao fenômeno em estudo. Diversos tipos de gráficos podem ser utilizados e, dentre estes, destacam-se os histogramas, diagramas de Pareto, boxplots e gráficos de linha.

Histograma: um histograma é um gráfico de barras verticais ou horizontais que representam uma distribuição de frequência de dados agrupados. O histograma pode representar a frequência absoluta (número de observações por classe), frequência relativa (percentual de observações da classe em relação ao total de observações) ou densidade (frequência relativa dividida pela amplitude do intervalo de classes).

A construção de um histograma é relativamente simples. Vamos ver como construí-lo com base no exemplo a seguir.

Exemplo 1 - Os testes de resistência a compressão de 100 corpos de prova de concreto de ultra alta resistência são apresentados na Tabela 4. Monte o histograma relativo ao teste.

Resistência a compressão - Concreto de ultra alta resistência (MPa)									
93	101	99	98	105	101	104	95	94	103
101	102	106	100	95	100	98	104	98	104
97	105	102	99	101	97	103	102	94	101
105	96	101	99	101	101	92	98	102	99
98	101	99	97	101	99	100	98	100	103
100	99	102	101	95	101	100	98	102	100
99	96	101	101	100	98	97	104	100	101
102	97	99	97	98	100	101	99	103	100
96	101	101	100	107	95	99	99	105	94
99	104	98	95	102	103	96	104	102	97

Tabela 4 - Dados de resistência a compressão

O primeiro passo é identificar a amplitude da amostra. Uma rápida leitura dos valores das observações indica o valor de 92 MPa como sendo a menor observação e 107 MPa como a maior observação. Assim:

Amplitude = (maior valor) – (menor valor) = (107) – (92) = 15

Como os valores das observações são discretos (e não contínuos), podemos optar por montar um histograma diretamente com os valores observados ou por meio da criação de classes. Inicialmente vamos trabalhar diretamente com os valores observados. Para isto basta contar a quantidade de observações relativas a cada um dos valores de resistência a compressão, conforme exibido na Tabela 5.

MPa	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107
Qtd	1	1	3	5	4	7	10	13	12	18	9	5	6	4	1	1

Tabela 5 - Quantidade de observações

Com base nos valores observados e na frequência de cada valor (quantidade de vezes que ele aparece), podemos facilmente montar o histograma (Figura 6).

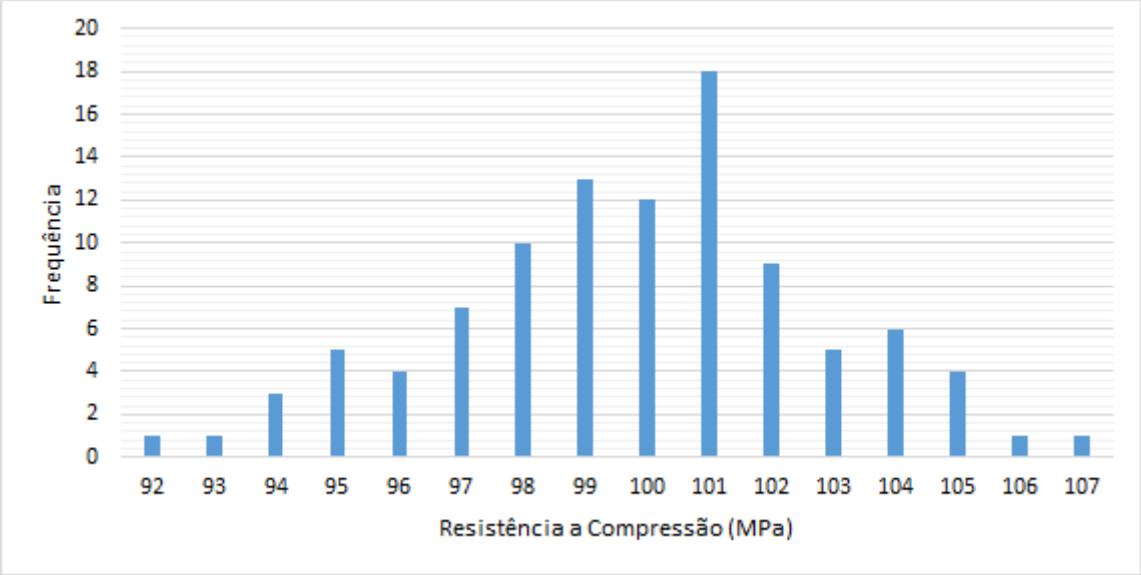

Figura 6 - Distribuição de Frequência

Uma outra forma, mais adequada quando tratamos de valores contínuos, é a criação de classes ao invés de nos utilizarmos dos próprios valores. Neste caso, o primeiro passo é a determinação do número de classes a ser usado. Um dos métodos mais utilizados para a determinação do número de classes é a Regra de Sturges[5], baseada no número de observações e dada pela equação:

$$K = 1 + 3{,}3\ log_{10}\ (n) \qquad \textit{Eq. 8}$$

É importante ressaltar que o número de classes não é um parâmetro rígido. Ele pode ser adequado para melhor representar os valores em função de:

- Na medida do possível, as classes deverão ter amplitudes iguais;
- Escolher os limites dos intervalos entre duas possíveis observações;
- O número de classes não deve ultrapassar 20;
- Escolher limites de classes que facilitem o agrupamento.

Para o nosso exemplo, o número de classes (K) resultado da aplicação da fórmula é 7,6. Assim, podemos escolher o número mais próximo que facilite a organização dos dados. Como a amplitude é 15, o número de classes ideal seria 8, o que resultaria na representação mostrada na Tabela 6:

MPa	[92 ; 93]	[94 ; 95]	[96 ; 97]	[98 ; 99]	[100 ; 101]	[102 ; 103]	[104 ; 105]	[106 ; 107]
Qtd	2	8	11	23	30	14	10	2

Tabela 6 - Distribuição de frequência - classes

A representação utilizada "[92 ; 93]" pode utilizar colchetes e/ou parêntesis. Colchetes indicam a inclusão dos limites ($92 \leq x \leq 93$). Parêntesis indicam a exclusão dos limites. Assim, a expressão "(92 ; 93)" indica ($92 < x < 93$). Já a representação [92 ; 93) indicaria a inclusão do limite inferior "92" e a exclusão do limite superior "93". Assim, a classe conteria todas as observações maiores ou iguais a 92 e menores que 93 (mais adequado a valores contínuos). O histograma é mostrado na Figura 7.

[5] Regra enunciada em 1926 pelo matemático alemão Herbert Sturges.

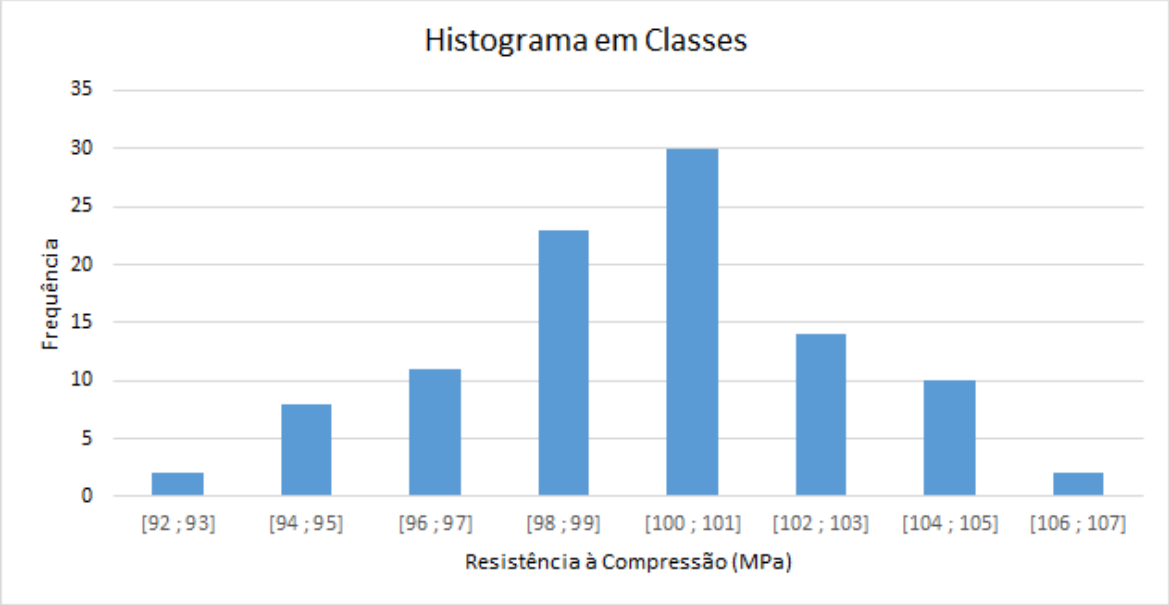

Figura 7 - Distribuição de frequência por classes

Diagrama de Pareto: O diagrama de Pareto é um gráfico de barras que ordena as frequências das ocorrências, da maior para a menor, permitindo a priorização dos problemas. Contém ainda a frequência acumulada. Este diagrama é baseado no Princípio ou Lei de Pareto, também conhecido como princípio 80-20, que afirma que para muitos fenômenos, 80% das consequências advém de 20% das causas.

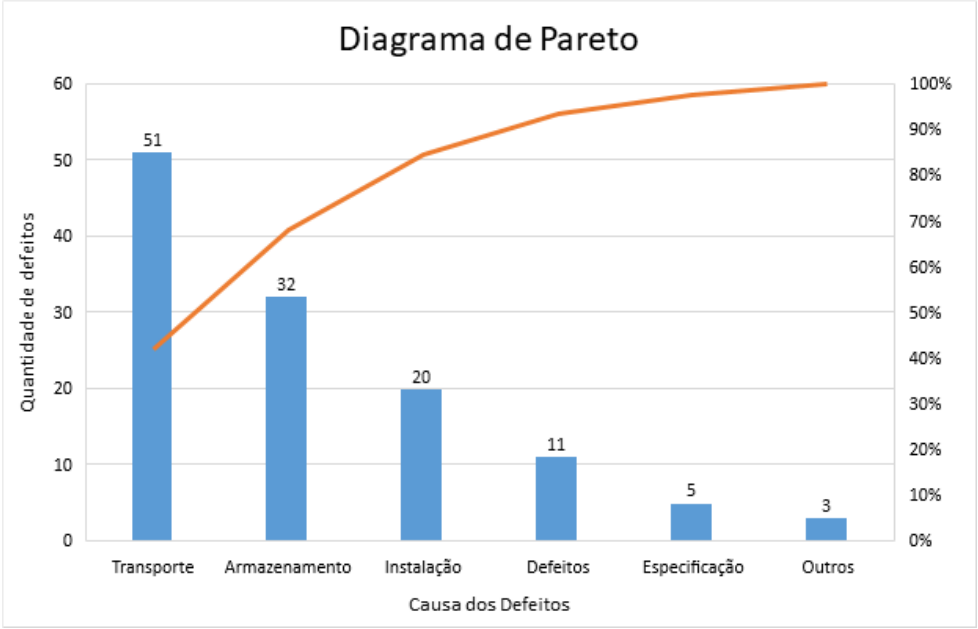

Figura 8 - Diagrama de Pareto

Como exemplo, podemos construir um Diagrama de Pareto para exibir a quantidade e causa de defeitos por lote de peças recebidas, conforme exibido na Figura 8. Pelo Diagrama de Pareto podemos visualizar que as três primeiras causas de defeitos são responsáveis por 84% do total de defeitos encontrados nas peças e devem ser o principal alvo de ações de correção.

Boxplot: Um dos gráficos preferidos e mais usados na comparação entre amostras diferentes e na exibição a distribuição empírica das observações de uma amostra. Seu formato é exibido na Figura 9. O boxplot é montado pela junção de cinco medidas da amostra (ou população). O primeiro quartil (Q1), o segundo quartil ou mediana (Q2), o terceiro quartil (Q3) e dois limites (superior e inferior), dados pelas equações:

$$Limite\ inferior = Q1 - 1{,}5\ x\ (Q3 - Q1) \qquad Eq.\ 9$$

$$Limite\ superior = Q3 + 1{,}5\ x\ (Q3 - Q1) \qquad Eq.\ 10$$

As observações (valores) que estiverem fora destes limites são considerados *outliers* (valores discrepantes) e são representados por asteriscos (*).

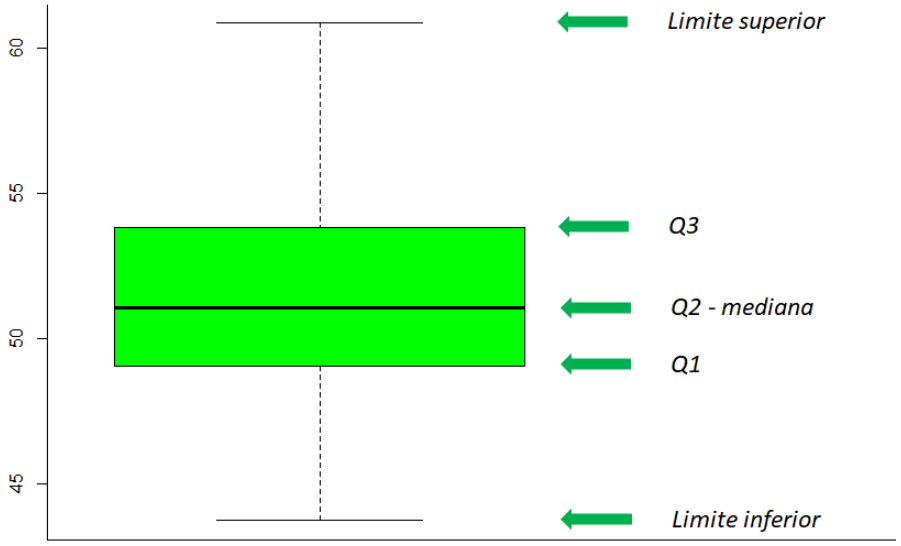

Figura 9 – Gráfico Boxplot

Os boxplot são muito usados como gráfico de comparação entre amostras, como mostra o próximo exemplo.

Exemplo 2: Foram testadas cinco composições diferentes para obtenção de concreto com adição de resíduos de construção e demolição (RCD). Para cada composição foram montados 12 corpos de prova. A Tabela 7 apresenta os resultados da resistência a compressão dos corpos de prova.

		Composições				
		a	b	c	d	e
Corpos de prova	1	54,39	31,89	41,70	42,35	50,45
	2	50,67	30,16	45,99	43,23	46,97
	3	43,40	35,28	38,20	41,38	41,70
	4	53,61	39,52	47,33	47,63	40,95
	5	52,88	33,77	41,71	33,46	44,52
	6	51,71	36,37	34,04	44,74	46,00
	7	55,04	33,69	40,42	25,97	45,15
	8	52,52	35,98	36,57	47,57	52,87
	9	44,53	39,79	41,49	22,60	43,50
	10	47,36	35,21	36,80	38,25	47,82
	11	44,76	33,78	38,26	26,40	42,11
	12	50,03	25,00	45,62	43,55	30,00

Tabela 7 - Resistência a compressão dos corpos de prova

Para montar um boxplot e comparar as amostras basta calcular as cinco medidas (limites superior e inferior, primeiro quartil, segundo quartil ou mediana e terceiro quartil) para cada uma das amostras. O gráfico resultante é o mostrado na Figura 10.

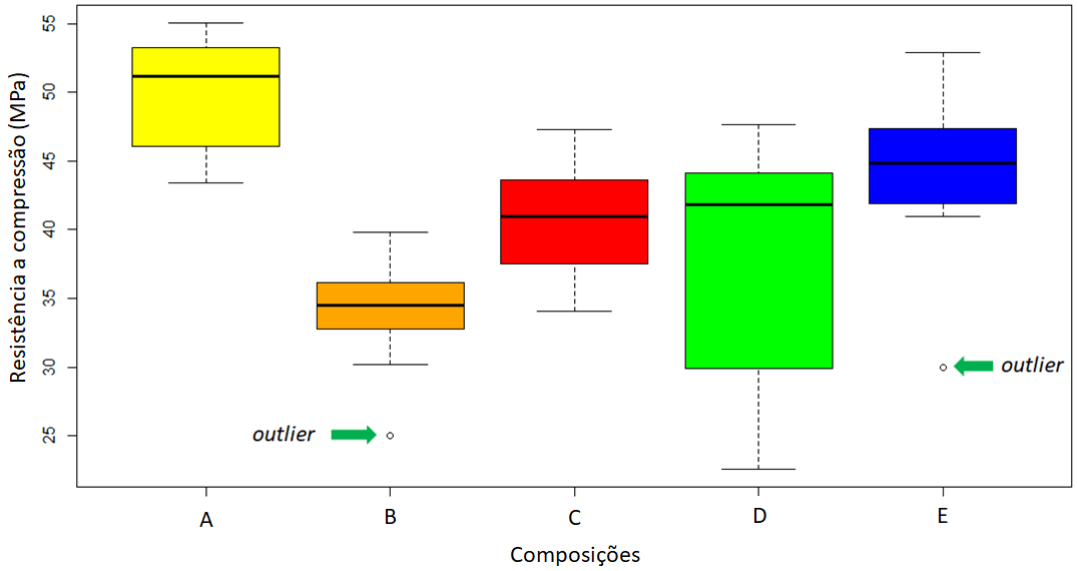

Figura 10 - Gráfico com os boxplots de cada amostra

Analisando o gráfico, pode-se perceber que a amostra A possui os maiores valores de resistência à compressão e a amostra B os menores. A amostra D possui os valores mais dispersos. Pode-se, também, identificar a presença de *outliers* (valores discrepantes) nas amostras B e E.

Gráficos de Linha: Os gráficos de linha são montados a partir de um par de ordenadas x e y. Utilizando os dados do Exemplo 2, podemos montar um gráfico de linha contendo o valor (y) para cada observação (x) de cada amostra das composições (A – E). O gráfico é mostrado na Figura 11.

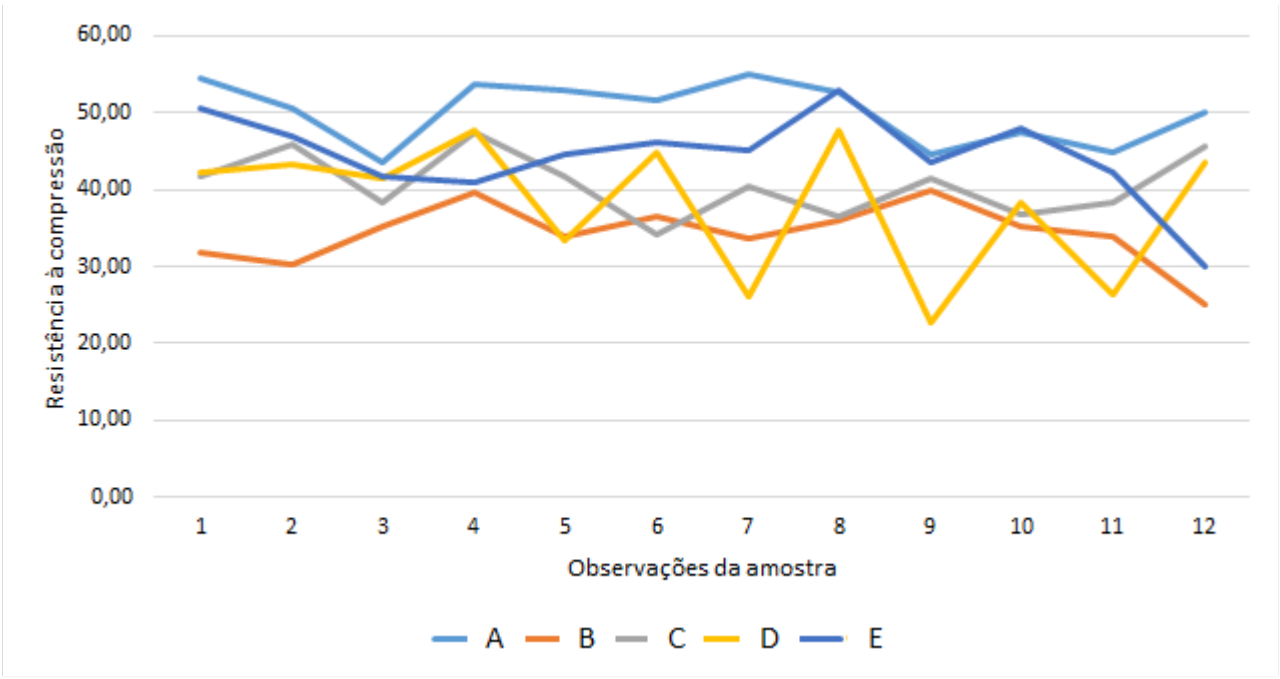

Figura 11 - Gráfico de linhas

A escolha do tipo de gráfico está vinculada ao tipo de informação que queremos transmitir ao leitor. Um boxplot é muito mais eficiente para comparação de valores de amostras (informação a ser transmitida) do que o gráfico de linhas. Agora, se a intenção for exibir tendências ou comportamento de uma variável x em função de outra variável y, o gráfico de linha pode ser muito mais adequado.

Exemplo 3: Neste exemplo, desejamos conhecer o comportamento da resistência a compressão em função da variação do percentual de adição de resíduo de construção e demolição. Os resultados de resistência à compressão em função do percentual de adição são mostrados na Tabela 8.

% adição	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%
MPa	52,87	50,45	47,82	46,97	46,00	45,15	44,52	43,5	42,11	41,7	40,95	30,00

Tabela 8 - Resistência a compressão dos elementos da amostra

O gráfico de linha da Figura 12 exibe o comportamento da resistência a compressão em função do percentual de adição:

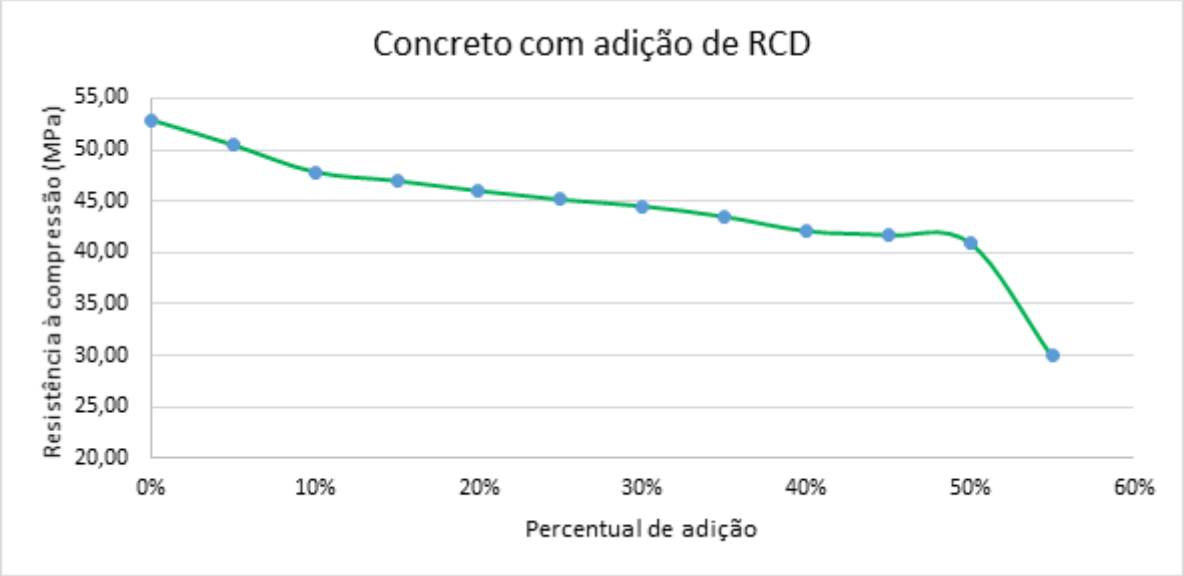

Figura 12 - Gráfico de linha: comportamento da resistência a compressão

O gráfico da Figura 12 ilustra perfeitamente a variação da resistência à compressão em função do aumento da adição de RCD. A escolha do tipo de gráfico mais adequado a informação é fundamental para que a transmissão desta informação seja realizada completamente.

Com isto encerramos esta breve introdução a estatística descritiva, na qual apenas os principais conceitos e medidas foram apresentados. É importante frisar que tudo o conteúdo exposto até o momento é utilizado para caracterizar os valores que foram mensurados (a amostra da população). Estes valores e medidas não podem ser utilizados para caracterizar a população, a não ser quando a amostra seja toda a população (um caso extremamente raro de ser obtido).

Para transferirmos as conclusões de um estudo de amostras para a população que originou a amostra, usamos a Inferência, um ramo da Estatística cujo objetivo é fazer afirmações a partir de um conjunto de valores representativo da população. A inferência estatística faz proposições sobre a população, usando dados da amostra (obtida por um dos métodos de amostragem descritos). Dada uma hipótese sobre a população, para a qual nós queremos fazer inferências, a inferência estatística consiste em escolher um modelo

estatístico adequado ao processo que gerou os dados e, a partir deste modelo, deduzir as proposições (conclusões).

Se por um lado a estatística descritiva detalha precisamente os dados analisados, uma vez que as medidas são obtidas a partir destes mesmos dados e somente deles, por outro a inferência estatística está sempre associada à uma margem de erro (risco), entendida como a probabilidade de que as conclusões, obtidas a partir da análise da amostra, sejam diferentes caso toda a população fosse sujeita ao mesmo procedimento de análise.

4 O SOFTWARE RSTUDIO E A ESTATÍSTICA DESCRITIVA

O uso da estatística em trabalhos acadêmicos foi, em grande medida, facilitado pelo desenvolvimento dos softwares estatísticos. Os cálculos de medidas da estatística descritiva podem ser efetuados até mesmo em planilhas Excel, inclusive alguns cálculos mais avançados de Inferência. Mas, apesar do MS Excel ser de conhecimento e domínio da maior parte dos estudantes, existem softwares específicos para estatística e o que iremos abordar é o software R[6] e o RStudio[7].

O software R é, basicamente, uma interface padrão texto para a linguagem R, uma linguagem de programação multi-paradigma, dinâmica, fracamente tipada e voltada à manipulação, análise e visualização de dados. Já o RStudio é um software de interface para o R com menus e atalhos (padrão Windows) que tornam o uso do R mais simples e amigável. Ambos são softwares de plataforma aberta, em contínuo desenvolvimento e atualização, gratuitos e possuem versões compiladas para Windows, Mac e Linux, motivo pelo qual são adotados como ferramenta estatística por um grande número de pesquisadores.

Existem diversos tutoriais e manuais sobre o uso destes softwares disponíveis na internet que podem ser usados como fonte de informações e para treinamento. A abordagem sobre estes softwares adotada neste texto é restrita a explicações básicas sobre os comandos e funções necessárias para a compreensão e execução dos exemplos e exercícios apresentados.

O primeiro exercício é, claro, a instalação do software R e do RStudio (use a internet, procure-o e instale-o. É fácil). Após a instalação, execute o RStudio. A Figura 13 mostra o layout da tela inicial do RStudio versão 3.6.2.

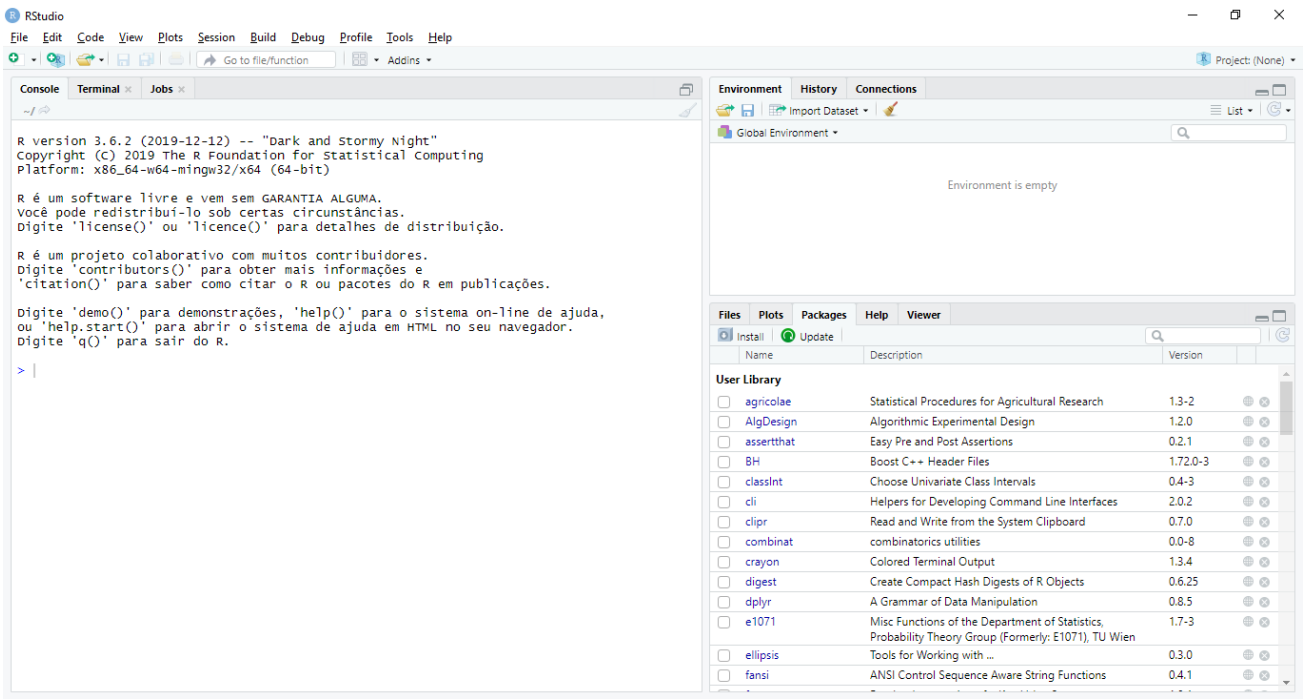

Figura 13 - Software RStudio

[6] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[7] RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc.,Boston, MA URL http://www.rstudio.com/.

RStudio e seu uso

A primeira informação necessária a respeito do RStudio é a de que ele usa notação inglesa para numeração, ou seja, o separador de decimais é o ponto (.) e não a vírgula como nós usamos. A segunda é que, apesar da interface, a maior parte dos comandos e funções são digitadas em janela específica, ainda que o RStudio possua alguns atalhos.

Como pode ser visto na Figura 13, o RStudio possui três janelas. A da esquerda é a janela de comandos com três abas (*console, terminal* e *jobs*). A janela superior direita possui três abas (*environment, history* e *connections*) e a inferior direita possui cinco abas (*files, plot, packages, help* e *viewer*). O uso destas janelas e abas será abordado quando necessário. Por enquanto, vamos usar a janela esquerda (console) para entrada dos comandos.

Para iniciarmos, veremos como citar o RStudio em trabalhos acadêmicos. Digite "*citation()*" na linha de comando. O resultado será:

```
> citation()

To cite R in publications use:

R Core Team (2019). R: A language and environment for statistical
  computing. R Foundation for Statistical Computing, Vienna, Austria. URL
  https://www.R-project.org/.
```

No capítulo anterior, aprendemos sobre as medidas de posição e dispersão. Vamos iniciar o uso do RStudio executando cálculos com estas medidas. Para iniciarmos, em primeiro lugar precisamos conhecer como entrar com dados (valores) no software. Há diversas maneiras[8]: digitação direta, leitura de arquivos contendo dados (em diversos formatos), importação. As mais usuais são a digitação e a leitura de arquivos.

1. Entrada de dados com o comando c(): o comando c() corresponde a "*concatenete*". Seu uso é bem simples. Especifique um nome para o vetor que conterá os dados e relacione os dados a serem inseridos no vetor. Lembre-se, o separador de decimais é o ponto e a virgula separa os valores. Para visualizar o conteúdo do vetor basta digitar o nome do vetor e "enter".

```
> amostra_a = c(97,98,99,99,99,100,101)
> amostra_a
[1]  97  98  99  99  99 100 101
```

   O comando *summary()* exibe um sumário com as estatísticas de posição relativas ao conteúdo do vetor (ou de qualquer outro arranjo) como visto a seguir:

```
> summary(amostra_a)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   97.0    98.5    99.0    99.0    99.5   101.0
```

   O comando exibe o valor mínimo, primeiro quartil, mediana (segundo quartil), média aritmética, terceiro quartil e valor máximo.

2. Outra forma de entrada de dados é via teclado com o comando *scan()*. Este comando abre a digitação de valores que é encerrada digitando-se "enter" duas vezes consecutivas.

[8] Para mais informações consulte os tutoriais disponíveis na internet (sugestão: http://www.leg.ufpr.br/~paulojus/embrapa/Rembrapa/Rembrapase7.html)

```
> amostra_b = scan()
1: 90
2: 95
3: 97
4: 97
5: 99
6: 103
7: 105
8: 106
9:
Read 8 items

> summary(amostra_b)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   90.0    96.5    98.0    99.0   103.5   106.0
```

3. O terceiro modo é o uso do comando *read.xxxx().* Com este comando é possível ler dados como dbf (arquivos DBASE), csv (separado por virgulas) e diversos outros formatos. O formato a ser lido é indicado como complemento do comando (*read.csv2()*). O formado csv pode ser gravado diretamente a partir de planilhas MS Excel. Apenas um cuidado: existem três formatos .csv no MS Excel separado por vírgulas, Macintosh e MS-DOS. Usaremos o primeiro (separado por virgulas) juntamente com o comando *read.csv2()*, pois desta forma o RStudio realiza a conversão de virgula para ponto decimal.

   Para este exemplo, usaremos os dados que foram usados como base para a construção da Tabela 2[9] (amostras c, d, e, f) e, na sintaxe do comando indicaremos a abertura da janela para seleção do arquivo (*file.choose()*) e a *existência* de cabeçalho para os dados (*header = TRUE* ou header = T).

```
> cdef = read.csv2(file.choose(), header=TRUE)
> cdef
       c     d      e      f
1  12.29 43.15 105.63 343.86
2  10.72 45.40 119.70 299.81
3   9.73 43.79  99.93 303.24
4   9.16 51.47  99.45 325.55
5  10.93 38.61 115.77 305.18
6  11.79 44.90  96.95 298.13
7  10.52 45.29  86.31 300.30
8   9.58 44.04 108.36 298.90
9   9.80 42.70 101.76 323.34
10 11.14 46.87 111.93 325.01
11  8.61 49.59  85.68 357.11
12  9.38 41.99  98.91 326.44
13 12.11 35.43 103.43 299.76
14  9.97 48.13 122.39 322.04
15  9.97 48.43 108.32 328.75
16 10.15 47.19 113.43 312.30
17 10.22 41.17  88.40 344.83
18 11.20 46.30 105.74 308.46
19 12.44 32.62  98.92 320.63
20 12.35 55.56 103.15 339.69
> summary(cdef)
       c                d               e                f
 Min.   : 8.610   Min.   :32.62   Min.   : 85.68   Min.   :298.1
 1st Qu.: 9.783   1st Qu.:42.52   1st Qu.: 98.92   1st Qu.:302.5
 Median :10.370   Median :45.09   Median :103.29   Median :321.3
 Mean   :10.603   Mean   :44.63   Mean   :103.71   Mean   :319.2
 3rd Qu.:11.348   3rd Qu.:47.42   3rd Qu.:109.25   3rd Qu.:327.0
 Max.   :12.440   Max.   :55.56   Max.   :122.39   Max.   :357.1
```

[9] Como os números foram gerados aleatoriamente em uma distribuição normal, haverá diferenças nos resultados.

O comando *summary(cdef)* exibe o sumário dos quatro vetores carregados, incluindo a média.

Exercício 1 – Procure os comandos para cálculo do desvio padrão e do coeficiente de variação e execute-os para os vetores c, d, e, f acima.

Como o boxplot é um dos gráficos mais importantes na estatística descritiva, abordaremos sua construção no RStudio. Sua construção é feita a partir do comando *boxplot()* e os argumentos serão os vetores. Como os vetores foram inseridos em uma única variável (*cdef*), usaremos o "$" para identifica-los (o parâmetro $ especifica uma única variável em um vetor) :

Aqui, devido a diferença de grandeza entre os vetores, mostramos apenas os vetores c e d (Figura 14):

```
> boxplot(cdef$c,cdef$d)
```

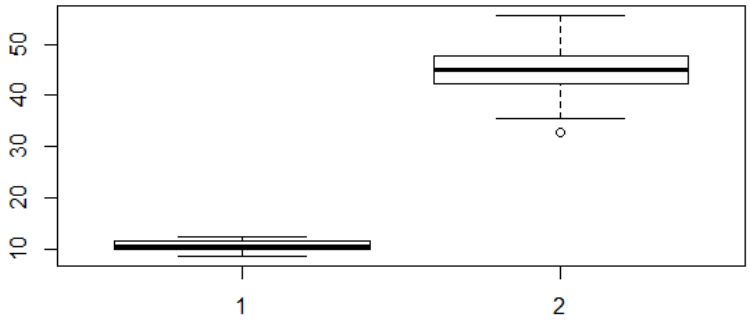

Figura 14 - Gráfico Boxplot das amostras c e d

Se quisermos acrescentar cores, basta complementar o comando boxplot (Figura 15):

```
> boxplot(cdef$c,cdef$d, col=c("yellow","orange"))
```

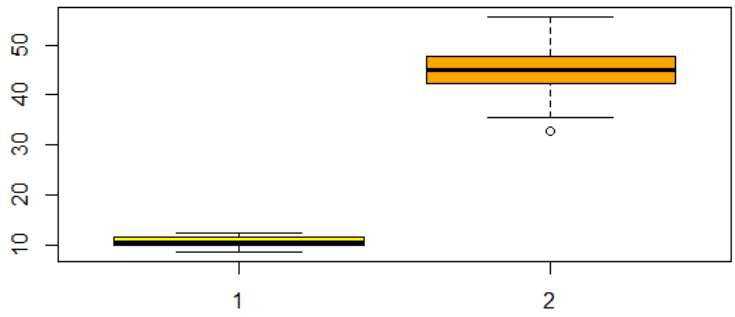

Figura 15 - Gráfico Boxplot colorido das amostras c e d

Exercício 2: Pesquise o comando *boxplot* e relacione as opções de configuração e edição do comando.

Este capítulo foi uma introdução preliminar ao uso do software R e RStudio na resolução de problemas estatísticos. É apenas uma pequena amostra de como ele funciona, de como interagimos com ele e de sua capacidade. Durante os capítulos seguintes, o RStudio será utilizado para a demonstração das funções estatísticas. No entanto, não é objetivo deste texto o aprendizado dos comandos e funções do R. Este é um tema extremamente vasto e para o qual existem diversos bons tutoriais. Alguns destes tutoriais serão indicados em notas de rodapé.

5 MODELOS PROBABILISTICOS E DETERMINISTICOS

Todas as vezes que se estudam fenômenos de observação, o primeiro passo é entender o próprio fenômeno e distinguir o modelo matemático que melhor o explique. Os fenômenos de observação, assim como os modelos matemáticos utilizados para explicar esses fenômenos podem ser divididos em determinísticos e probabilísticos ou aleatórios.

5.1 Modelos Deterministicos

Os fenômenos determinísticos conduzem sempre a um mesmo resultado quando as condições iniciais são as mesmas, ou seja, seguem leis determinísticas e seu resultado é obtido por meio destas leis. Como exemplo podemos tomar o tempo de queda livre de um corpo. Mantidas as mesmas condições, as variações obtidas para o valor do tempo de queda livre de um corpo são extremamente pequenas, e se ocorrem, normalmente tem origem na imprecisão dos mecanismos de medição.

Os fenômenos aleatórios são aqueles cujo resultado, mesmo em condições normais de experimentação, variam de uma observação para outra. Estes fenômenos não possuem uma "lei" ou regra que determine seus resultados ou, se possuem, esta lei ou regra não é conhecida, impossibilitando a previsão de um resultado. Assim, mesmo que haja um grande número de repetições do fenômeno, os resultados não são previsíveis. Por exemplo, podemos considerar os seguintes experimentos conduzidos como fenômenos aleatórios:

- Lançamento de uma moeda;
- Lançamento de um dado;
- Lançamento de duas moedas;
- Retirada de uma carta de um baralho completo, de 52 cartas.

Considerando que os resultados dos experimentos aleatórios estão sujeitos ao acaso, ou seja, são experimentos ou situações em que deve ocorrer um, dentre os vários resultados possíveis, a análise dos resultados dos experimentos relacionados acima revela que:

- Cada experimento pode ser repetido indefinidamente sob as mesmas condições;
- Não se conhece em particular o valor do resultado do experimento "a priori", porém pode-se descrever todos os possíveis resultados;
- Quando o experimento for repetido um grande número de vezes, surgirá uma regularidade.

Os modelos que estudam os fenômenos aleatórios são chamados de probabilísticos, pois, apesar de não podermos prever o resultado, podemos determinar, a priori, a probabilidade de ocorrência de um determinado resultado.

5.2 Modelos Probabilísticos

Podemos então, conceituar Modelo Probabilístico como sendo modelos construídos a partir de certas hipóteses sobre o problema que está sendo estudado. Os modelos probabilísticos são constituídos por duas etapas:

- Da identificação de todos os resultados possíveis de serem obtidos;
- De uma certa lei ou regra que nos informa o quão provável é cada resultado ou grupo de resultados.

Da primeira etapa surge o conceito de **Espaço Amostral** que é o conjunto de todos os resultados possíveis do experimento aleatório e pode ser classificado em:

- Espaço amostral discreto: contém um número finito de possibilidades ou uma sequência infinita com tantos elementos quanto são os números inteiros.
- Espaço amostral contínuo: contém um número infinito de possibilidades igual ao número de pontos em um segmento de reta.

Outro conceito importante é o de **Evento**. Evento é um conjunto de resultados do espaço amostral. Por definição, o evento é sempre um subconjunto do espaço amostral. Por exemplo, para o lançamento de um dado podemos considerar o evento PAR como a ocorrência de um número par (2, 4, 6) e o evento IMPAR como a ocorrência de um número ímpar (1, 3, 5).

Como estamos tratando de modelos probabilísticos, a determinação da lei ou regra que nos informa o quão provável é cada resultado ou grupo de resultados (evento), citada na segunda etapa do modelo probabilístico, é nosso próximo objetivo.

5.3 Probabilidade

A lei que rege o modelo probabilístico é baseada no conceito de probabilidade. Probabilidade é um valor entre 0 (zero) e 1 (um) associada à ocorrência de um determinado evento. A soma das probabilidades de todos os resultados possíveis do experimento deve ser sempre igual a 1.

Para entendermos melhor o conceito de probabilidade, vamos analisar os seguintes exemplos:

- Ocorrência de um número par no lançamento de um dado: Evento A = $\{2, 4, 6\}$ no Espaço amostral = $\{1, 2, 3, 4, 5, 6\}$. O Evento A possui 3 ocorrências num total de 6 ocorrências. A = 3 / 6 = 0,5.
- Ocorrência de um número menor que 3 no lançamento de um dado: Evento B = $\{1, 2\}$ no Espaço amostral = $\{1, 2, 3, 4, 5, 6\}$. O Evento A possui 2 ocorrências num total de 6 ocorrências. B = 2 / 6 = 0,33.
- Ocorrência do número 6: Evento C = $\{6\}$ no Espaço amostral = $\{1, 2, 3, 4, 5, 6\}$. O Evento C possui 1 ocorrência num total de 6 ocorrências. C = 1 / 6 = 0,17.
- Ocorrência de um número maior que 6: Evento D = $\{\emptyset\}$ no Espaço amostral = $\{1, 2, 3, 4, 5, 6\}$. O Evento D possui zero ocorrências num total de 6 ocorrências. D = 0 / 6 = 0.

Dos exemplos acima, podemos entender o **Princípio da Equiprobabilidade**, usado no cálculo da probabilidade de um evento. Ele determina que quando todos os resultados possíveis são igualmente prováveis, isto é, quando as características do experimento sugerem N possíveis resultados, todos com igual probabilidade de ocorrência, a probabilidade de um evento A, contendo N_A resultados, pode ser definida por:

$$P(A) = \frac{N_A}{N} \qquad \textit{Eq. 11}$$

Outro princípio usado para cálculo da probabilidade é o **Princípio da Independência**: Dois eventos são independentes quando a ocorrência de um deles não altera a probabilidade da ocorrência do outro.

Da mesma forma, vamos analisar o princípio da independência a partir dos seguintes exemplos:

- Qual a probabilidade de lançar um dado, duas vezes, e em ambas obtermos números pares? Considerando o Espaço Amostral EA = $\{1, 2, 3, 4, 5, 6\}$ e os eventos desejados E1 = E2 = $\{2, 4, 6\}$a probabilidade P (E1 x E2) = P(E1) x P(E2) = 0,5 x 0,5 = 0,25.

- Numa linha de montagem são produzidas bolas de bilhar em lotes de 10 bolas, sendo que 2 são vermelhas, 2 são verdes, 2 são azuis, 2 são amarelas e 2 brancas. Qual a probabilidade de, em um experimento aleatório, sem reposição, retirarmos 2 bolas brancas?

  Para a primeira retirada temos:
  Espaço Amostral N = (10)
  Evento Bola Branca N_A = (2)
  P1 = Probabilidade de retirada da primeira bola branca = N_A / N = 2 / 10 = 1 / 5 = 0,2.

  Com a retirada da primeira bola branca, o Espaço amostral N foi reduzido de 1 e o número de bolas brancas também. Portanto:
  Espaço Amostral N = (9)
  Evento Bola Branca N_A = (1)
  P2 = Probabilidade de retirada da segunda bola branca = N_A / N = 1 / 9 = 0,11.
  Então, P (P1 P2) = 1/5 x 1/9 = 1/45 = 0,022

Teoria da Contagem: Dados dois eventos, o primeiro dos quais pode ocorrer de *m* maneiras distintas e o segundo pode ocorrer de *n* maneiras distintas, então os dois eventos conjuntamente podem ocorrer de *m.n* maneiras distintas. O cálculo da probabilidade de um evento reduz-se a um problema de contagem.

A **Análise Combinatória** tem fundamental importância para se contar o nº de casos favoráveis e o total de casos por meio dos conceitos e fórmulas de combinações e arranjos. A diferença entre combinação e arranjo é a ordem dos elementos. No arranjo, a ordem de seleção dos elementos é importante e diferencia os resultados, na combinação não. Suponhamos que temos cinco elementos (A, B, C, D e E) e os queremos combinar dois a dois. Para o arranjo, os resultados (Portal Action, 2020) e (Portal Action, 2020) são diferentes. Já para a combinação, como a ordem não importa, (Portal Action, 2020) e (Portal Action, 2020) representam o mesmo resultado.

A fórmula para o cálculo de combinação de *r* elementos *p* a *p* é:

$$C_{r,p} = \frac{r!}{p!\,(r-p)!} \qquad \textit{Eq. 12}$$

A fórmula para o cálculo de arranjo de *r* elementos *p* a *p* é:

$$A_{r,p} = \frac{r!}{(r-p)!} \qquad \textit{Eq. 13}$$

A primeira informação necessária para saber o número total de casos será dada por combinação ou arranjo é, então, saber se a ordem de seleção é importante ou não. Analisemos os dois exemplos a seguir:

- Na confecção de amostras de concreto para testes de resistência, dentre 10 tipos de aditivos diferentes serão escolhidos três para compor cada amostra. Quantos conjuntos diferentes de amostras podem ser formados? Considere que os aditivos serão adicionados sempre no percentual indicado pelo fabricante.

  Bom, temos três aditivos (dentre 10 aditivos) que serão adicionados, juntos, durante o processo de produção do concreto. Neste caso, a ordem não importa, então trata-se de combinação de 10 elementos três a três.

$$C_{10,3} = \frac{10!}{3!\,x\,(10-3)!} = \frac{10!}{3!\,x\,7!} = \frac{10\,x\,9\,x\,8\,x\,7!}{3\,x\,2\,x\,1\,x\,7!} = \frac{720}{6} = 120$$

- Considerando um grupo de dez pessoas, quantas chapas diferentes podemos ter para uma eleição de presidente, tesoureiro e secretário?

  Neste caso, a ordem importa, pois representam resultados diferentes a seleção de uma determinada pessoa para presidente, tesoureiro ou secretário. Trata-se de um arranjo de 10 elementos três a três.

$$A_{10,3} = \frac{10!}{(10-3)!} = \frac{10\ x\ 9\ x\ 8\ x\ 7!}{7!} = \frac{10\ x\ 9\ x\ 8}{1} = 720$$

  Outra forma de analisar o arranjo é: para o primeiro cargo, existem 10 opções, para o segundo cargo, nove opções e para o terceiro, oito opções, pois são 10 pessoas e uma mesma pessoa não pode exercer duas ou mais funções, então temos 10 x 9 x 8 = 720.

6 DISTRIBUIÇÃO DE PROBABILIDADES

Como visto no item anterior, probabilidade é um valor entre 0 (zero) e 1 (um) associada à ocorrência de um determinado evento pertencente ao espaço amostral e que a soma das probabilidades de todos os eventos possíveis (todos os elementos do espaço amostral) é sempre igual a um (1).

A distribuição de probabilidades é uma função que associa uma probabilidade a cada resultado numérico de um experimento, ou seja, fornece a probabilidade associada a cada elemento do espaço amostral. Para que possamos compreender como construir uma distribuição de probabilidades, são necessários alguns conceitos:

Variável aleatória: Muitos experimentos aleatórios produzem resultados não numéricos. Desta forma, é conveniente transformar seus resultados em números, o que é feito por meio de uma **variável aleatória**. Assim, podemos entender uma variável aleatória como uma função que associa um valor numérico a cada ponto do espaço amostral não numérico.

Assim, a variável aleatória é uma variável que tem um valor único para cada resultado aleatório de um experimento. A palavra aleatória indica que em geral só conhecemos aquele valor depois do experimento ser realizado.

Uma vez definida a variável aleatória que irá associar cada elemento do espaço amostral (não numérico), nosso próximo objetivo é o cálculo das probabilidades correspondentes. O conjunto das variáveis e das probabilidades correspondentes é denominado **distribuição de probabilidades**, isto é:

$$P(x) = \{x_i, p(x_i), \qquad i = 1, 2, 3, \dots, n\} \qquad \textit{Eq. 14}$$

A distribuição de probabilidades pode ser mais facilmente visualizada por meio de um exemplo: Considere o lançamento de três moedas. Qual a probabilidade de obtermos zero, uma, duas e três caras?

A busca pela resposta inicia-se com a construção do espaço amostral relativo ao experimento. Cada lançamento pode resultar em cara e coroa. São três lançamentos. Assim, assumindo que CA representa cara e CO coroa, os resultados possíveis e equiprováveis, temos o espaço amostral exibido na Tabela 9:

ESPAÇO AMOSTRAL			
1	CA, CA, CA	5	CO, CA, CA
2	CA, CA, CO	6	CO, CA, CO
3	CA, CO, CA	7	CO, CO, CA
4	CA, CO, CO	8	CO, CO, CO

Tabela 9 - Espaço amostral do experimento

O espaço amostral do experimento possui oito alternativas. Como nosso interesse é a contagem do número de caras, vamos enumerar os eventos (de zero a três caras) no espaço amostral e associar a cada evento sua frequência (Tabela 10):

Evento	Variável aleatória	Elemento do espaço amostral	Frequência	Probabilidade
Zero caras	0	8	1	1/8
Uma cara	1	4, 6 e 7	3	3/8
Duas caras	2	2, 3 e 5	3	3/8
Três caras	3	1	1	1/8

Tabela 10 - Espaço amostral - contagem dos eventos

Na Tabela 10 cada evento foi associado a uma variável aleatória (número real X(e)), sendo a quantidade de ocorrências (frequência) do evento determinada assim como sua probabilidade (N_A/N). A Distribuição de probabilidade associada é mostrada na Figura 16:

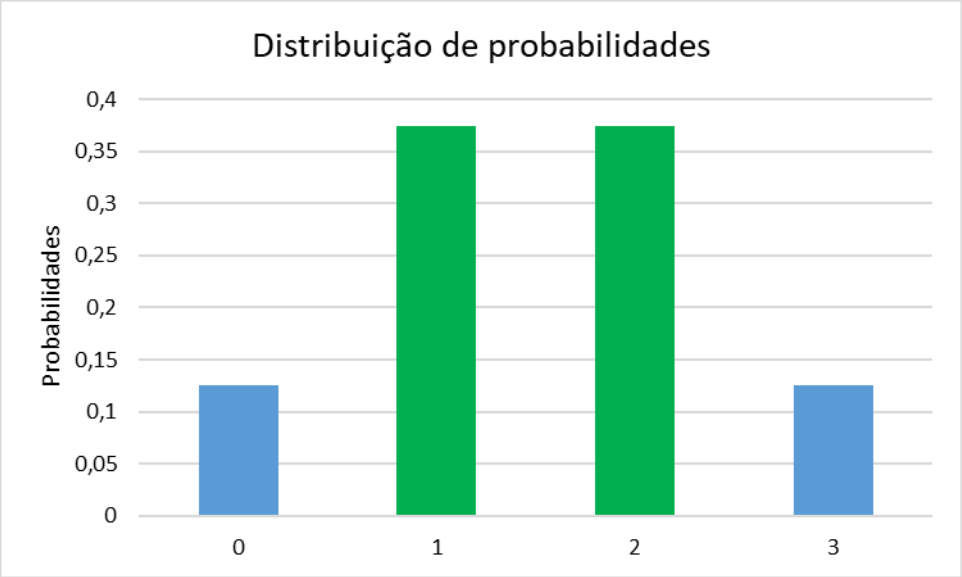

Figura 16 - Distribuição de probabilidades do evento

As distribuições de probabilidades são probabilidades associadas a uma variável aleatória (representando um evento do espaço amostral) e temos duas regras de verificação que se aplicam a qualquer distribuição de probabilidades:

- A soma de todos os valores (probabilidades) de uma distribuição de probabilidades deve ser igual a um (1 = 100%). Assim, $\sum P(x) = 1$, onde x assume todos os valores do espaço amostral ou eventos possíveis.
- A probabilidade de um determinado evento não pode ser negativa. $0 \leq P(x) \leq 1$, para todo x.

As variáveis aleatórias podem ser discretas ou contínuas. No exemplo anterior, temos uma variável aleatória discreta, pois somente pode assumir os valores zero, um, dois ou três. Uma variável aleatória contínua é aquela que pode assumir inúmeros valores num intervalo de números reais e é medida em uma escala contínua. Vamos analisar isto no próximo exemplo:

Exemplo 4: Considere uma roleta, dividida em quatro quadrantes. Seja X a variável aleatória que indica o ponto exato em que o ponteiro para de girar (como existem infinitos pontos em cada quadrante, esta variável aleatória é contínua). Qual a probabilidade de o ponteiro parar no primeiro quadrante (0 a 90º)?

Espaço amostral: para uma roleta dividida em quatro quadrantes, temos um espaço amostral = $\{Q1, Q2, Q3\ e\ Q4\}$

O evento de interesse é o ponteiro parar no primeiro quadrante, então E = $\{Q1\}$. Disto decorre que P(E) = 1/4. Se colocarmos em um gráfico, representando os quadrantes em graus, teremos (Figura 17):

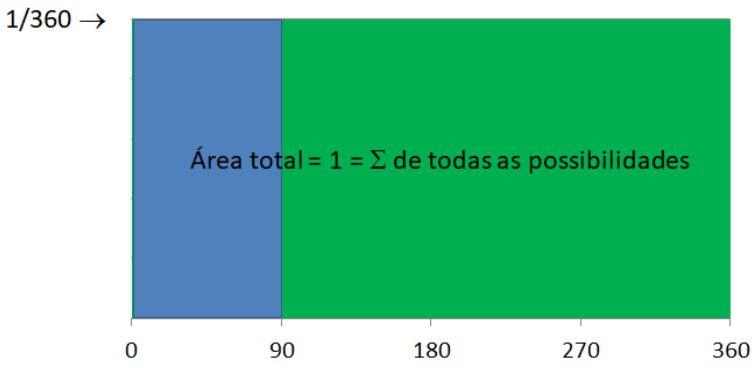

Figura 17 - Distribuição de probabilidades

Vamos aproveitar o gráfico exibido na Figura 17 e analisarmos um pouco as diversas situações que o gráfico pode representar:

- Se tomamos como base os quadrantes, teremos quatro quadrantes e uma probabilidade igual de 1/4 para cada um destes quadrantes. O valor da probabilidade no eixo y seria 1/4 e o eixo x seria numerado de 1 a 4. A área total sob o gráfico corresponderia a 1/4 x 4 = 1. Todas as probabilidades são positivas e estão entre 0 e 1. Nesse caso, atende as duas regras de verificação.
- Se tomarmos como base os graus (de 0^o a 360^o), teremos que o valor da probabilidade no eixo y seria de 1/360 e o eixo x seria numerado de 0 a 360. A área total sob o gráfico corresponderia a 1/360 x 360 = 1. Da mesma forma, todas as probabilidades são positivas e estão entre 0 e 1. Ok, atende as duas regras de verificação.
- Se assumirmos que o ponteiro da roleta pode indicar um valor contínuo no segmento de reta [0 – 360], teríamos uma variável aleatória contínua. O valor da probabilidade no eixo y não seria possível de ser determinado, uma vez que o eixo x possui infinitos valores, mas para ser uma distribuição de probabilidades, a soma de todas as probabilidades continua sendo igual a 1 e todas as probabilidades estariam entre 0 e 1.

A última situação retratada acima mostra que, apesar de termos uma variável aleatória contínua (infinitos valores), as regras das distribuições de probabilidades continuam sendo válidas e podemos nos utilizar delas para o cálculo de probabilidades. Vejamos o exemplo[10] a seguir:

Exemplo 5: A ocorrência de panes em qualquer ponto de uma rede telefônica de 7 km foi modelada por uma distribuição Uniforme no intervalo [0 – 7]. Qual é a probabilidade de que uma pane venha a ocorrer nos primeiros 800 metros? E qual a probabilidade de que ocorra nos 3 km centrais da rede?

A distribuição de probabilidade seria (Figura 18):

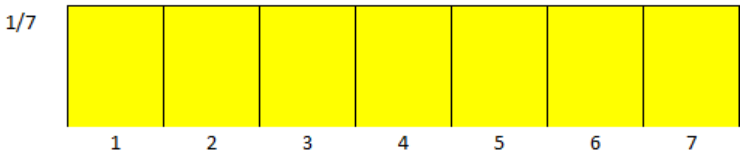

Figura 18 - Distribuição de probabilidades

[10] Reproduzido de http://www.portalaction.com.br/probabilidades/61-distribuicao-uniforme

A função correspondente a curva, chamada de função densidade de probabilidade, é dada por $f(x) = \frac{1}{7}\ se\ 0 \leq x \leq 7$, e zero, caso contrário. Assim, a probabilidade de uma pane ocorrer nos primeiros 800 metros é:

$$P(x < 0{,}8) = \int_0^{0{,}8} f(x)dx = \frac{0{,}8 - 0}{7} = 0{,}1142$$

Já a probabilidade da pane ocorrer nos 3 km centrais seria igual a probabilidade de ocorrência nos 5 km iniciais menos a probabilidade de ocorrência nos 2 km iniciais, ou seja:

$$P(2 \leq x \leq 5) = \int_2^5 f(x)dx = P(x \leq 5) - P(x \leq 2) = \frac{5}{7} - \frac{2}{7} = \frac{3}{7} = 0{,}4285$$

Assim, não interessa qual seja o formato da curva associada a distribuição de probabilidades (dada pela função densidade de probabilidades *F(X)*). Desde que o espaço amostral seja representado no eixo *x*, a probabilidade de um evento pode ser determinada pela relação entre a área total delimitada pela curva e o eixo *x* e a área delimitada correspondente ao evento. A Figura 19 ilustra o exemplo, onde a probabilidade do evento (a, b) é dada pela razão entre a área delimitada pelo evento E(a, b) e a área total.

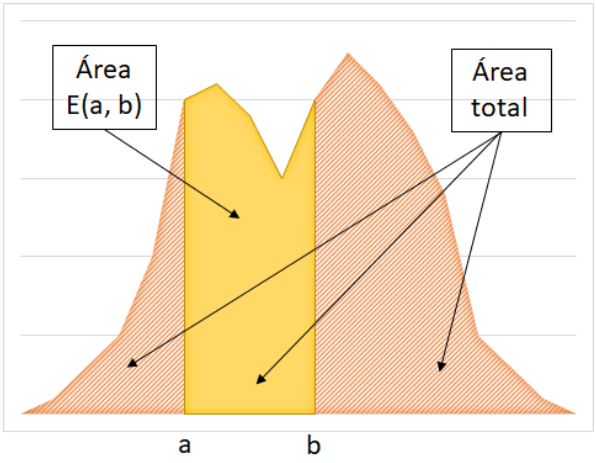

Figura 19 - Probabilidade do evento (a, b)

Desta forma, conseguimos transformar um problema estatístico em um **cálculo matemático**, ou seja, para descobrirmos a probabilidade de um certo evento $P(E_{a,b})$, basta montarmos a distribuição de frequência da variável em estudo, deduzirmos a equação de sua curva, dada pela função densidade de probabilidade $F(X)$, calcularmos a área total sob a curva (de 0 a N) e a área correspondente ao evento $E_{a,b}$. A equação correspondente é:

$$P(E_{a,b}) = \frac{\int_a^b F(X)}{\int_0^N F(X)} \qquad \textit{Eq. 15}$$

Resolvido? Bom, não! Primeiro porque teríamos que deduzir a equação correspondente a função densidade de probabilidade e isto pode não ser tão simples assim, mesmo que tenhamos um espaço amostral que possua um tamanho suficiente (quantidade de elementos) que o permita.

Dependendo da quantidade de valores (resultados de amostras, elementos testados), isto pode ser impossível. Poderíamos sim, estimar com certo grau de precisão o tipo de função de densidade de probabilidade a curva real seguiria e, a partir desta estimativa, fazer aproximações.

Por isto, uma das primeiras atividades para a Inferência é a identificação do **Modelo Probabilístico** a ser usado. O modelo probabilístico é associado ao tipo de distribuição de probabilidades que o espaço amostral em estudo segue. Existem diversos tipos de modelos probabilísticos, sendo que a distribuição uniforme discreta (constante) é a que foi utilizada nos dois exemplos anteriores. Vamos ver a seguir os tipos de Modelos Probabilísticos mais comuns.

Os modelos probabilísticos podem ser divididos em dois tipos básicos: os discretos, baseados em variáveis aleatórias discretas e os modelos contínuos, baseados em variáveis aleatórias contínuas. A diferença entre eles é o valor que suas variáveis aleatórias podem assumir (discretos ou contínuos).

Os modelos apresentados a seguir são modelos discretos.

6.1 Distribuição Uniforme Discreta

O modelo de distribuição uniforme discreta ocorre quando todos os elementos de um espaço amostral definido são igualmente prováveis. Este é o modelo que foi utilizado nos exemplos anteriores. Sua função de distribuição de probabilidades pode ser vista como uma reta (Figura 20).

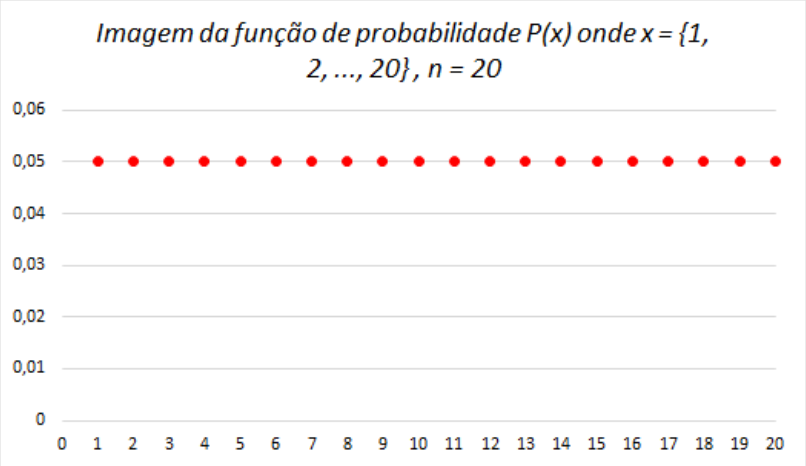

Figura 20 - Distribuição de Probabilidades Uniforme Discreta

6.2 Distribuição de Bernoulli

O modelo de distribuição de Bernoulli é a distribuição mais simples de probabilidades. Corresponde a um único experimento com resultados iguais a sucesso ou fracasso (ou outras variantes como sim ou não; cara ou coroa). Seu espaço amostral corresponde a (Portal Action, 2020) onde o valor um corresponde ao sucesso com probabilidade p e o valor zero ao fracasso com probabilidade $q = 1 - p$ (Figura 21). O experimento é dito justo quando $p = q = 0,5$ (ambos os eventos possuem a mesma probabilidade).

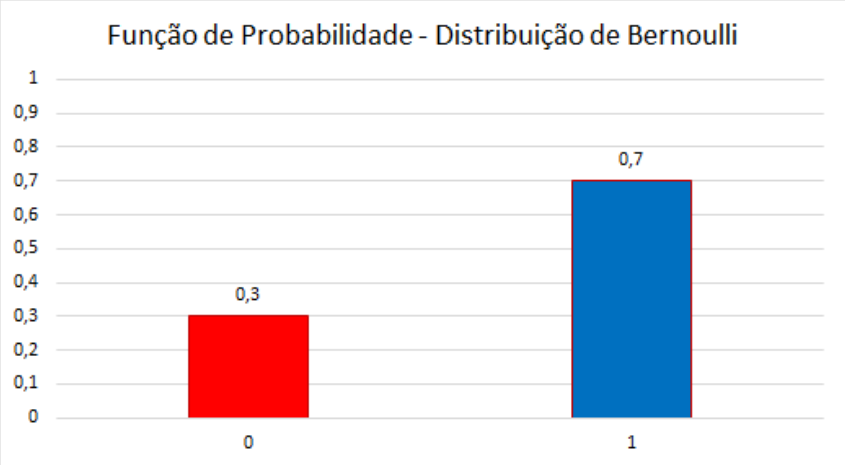

Figura 21 - Distribuição de Bernoulli

6.3 Distribuição binomial

Suponhamos que sejam realizados *N* experimentos cuja distribuição individual é a de Bernoulli, ou seja, uma série de experimentos cujo resultado admite apenas duas classificações (sucesso ou fracasso, masculino ou feminino, cara ou coroa, etc.). Seja *X* uma variável aleatória associada ao número de sucessos (1) obtidos nas *N* realizações do experimento. Se a probabilidade de sucesso de cada um dos experimentos individuais é *p* e a de fracasso é *q* (*q* = 1 – *p*) então dizemos que *X* possui uma distribuição binomial *X* ~ Bin (*N,p*).

Para entendermos melhor, examinemos o seguinte exemplo:

Exemplo 6: Uma linha de montagem ininterrupta produz bolas pretas e brancas, sendo a probabilidade de produção de bolas pretas quatro vezes maior que a de bolas brancas. Em 10 eventos independentes de retirada uma bola, para compor uma amostra de 10 bolas, qual a probabilidade de obtermos três bolas pretas?

Se a probabilidade de retirarmos uma bola preta é quatro vezes maior, vamos assumir que existem quatro vezes mais bolas pretas que brancas, ou seja, a cada cinco bolas, quatro são pretas e uma é branca. Assim, podemos considerar que cada retirada de uma bola (preta ou branca) para compor a amostra de 10 bolas como um experimento de Bernoulli onde *p* (bola preta = sucesso) = 0,8 e *q* = 1 – *p* = 0,2 (bola branca = fracasso).

Para a segunda retirada (experimentos independentes) a probabilidade é a mesma. Assim, a probabilidade de sucesso em duas retiradas (k = 2) é igual a p^2 (0,8 x 0,8 ou p^k). A probabilidade de dois fracassos é igual a (1 – *p*) x (1 – *p*) ou $(1-p)^2$ e generalizando $(1-p)^k$.

A probabilidade de um evento amostral com k sucessos e n – k fracassos é dada pela equação:

$$p^k(1-p)^{n-k} \qquad \textit{Eq. 16}$$

A equação representa a probabilidade de qualquer evento do espaço amostral com k sucessos e n – k fracassos. Assim, temos que determinar quantas combinações diferentes podemos obter de uma amostra de 10 bolas combinando-as três a três. Para relembrar, a equação correspondente (*Eq. 12*) é:

$$C_{(N,k)} = \frac{N!}{k!\,(N-k)!}$$

Assim, para k = 0, 1, 2, ... , N:

$$P(X = k) = C_{N,k}\, p^k (1-p)^{n-k}$$

Então, para três (*k*) sucessos (bola preta) em uma amostra de 10 bolas (*N*) com *p* = 0,8:

$$P(3) = \frac{10!}{3!\,(10-3)!}\,(0{,}8)^3 x (1-0{,}8)^{(10-3)} = 120\; x\; 0{,}512\; x\; 0{,}0000128 = 0{,}0007864$$

Então, supondo agora que a probabilidade de retirada de bolas brancas é igual a 30% (*p* = 0,3) e sendo *k* o número de bolas brancas presentes na amostra, construa o gráfico de distribuição de probabilidades P(*k*).

A equação é a mesma, então: $p(k = 0, 1, \ldots, 10) = C_{(10,k)}\; p^k\; (1-p)^{(10-k)}$. Montando uma tabela da probabilidade (*p(k)*) e probabilidade acumulada (*F(k)*)em função do número de sucessos (k), temos os resultados apresentados na Tabela 11 e o gráfico da distribuição de probabilidades correspondente é apresentado na Figura 22.

k	$C_{(10,k)}$	*p(k)*	*F(k)*
0	1	0,028248	0,028248
1	10	0,121061	0,149308
2	45	0,233474	0,382783
3	120	0,266828	0,649611
4	210	0,200121	0,849732
5	252	0,102919	0,952651
6	210	0,036757	0,989408
7	120	0,009002	0,998410
8	45	0,001447	0,999856
9	10	0,000138	0,999994
10	1	0,000006	1

Tabela 11 - Probabilidades e Probabilidades acumulada

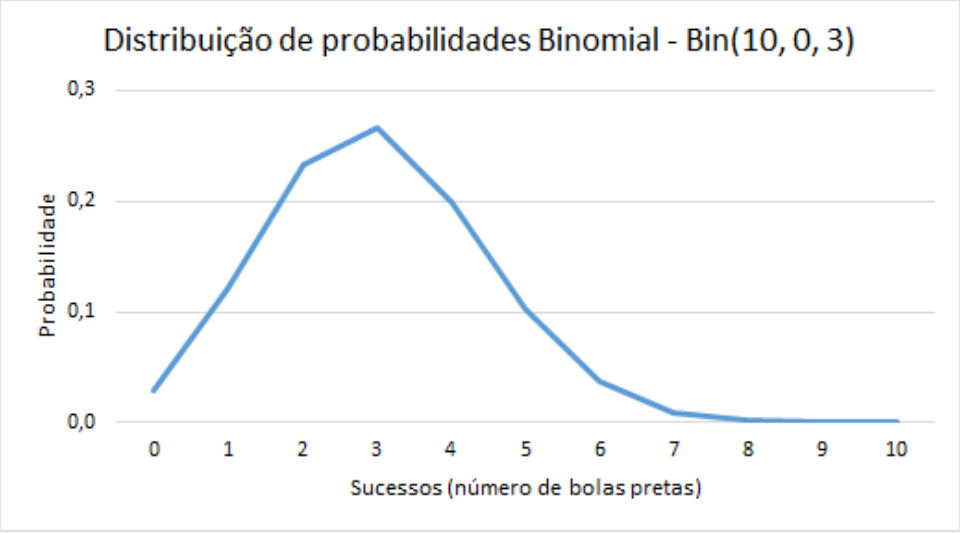

Figura 22 - Distribuição Binomial

A distribuição de Bernoulli é um caso especial da distribuição binomial com *n* = 1.

6.4 Distribuição de Poisson

A distribuição de Poisson é uma distribuição de probabilidade que expressa a probabilidade de uma série de eventos ocorrer num certo intervalo de unidade de medida (unidade de tempo, volume, área, etc.), com a restrição de que estes eventos ocorrem independentemente de quando ocorreu o último evento. A distribuição de Poisson é uma forma limite da distribuição binomial quando N → ∞ e $p \rightarrow 0$ e é usada em casos que envolvem contagem e cuja probabilidade de ocorrência é proporcional ao intervalo de amostragem, como por exemplo em número de defeitos em peças, erros tipográficos por página impressa, mortes por acidente por ano em uma cidade, etc. Neste caso, a variável aleatória é discreta (número de ocorrências) e o espaço amostral é contínuo (tempo, área). A distribuição de Poisson é caracterizada pelo parâmetro λ (derivado de *p* e *q* da distribuição binomial) que é traduzido como a taxa média de ocorrência por unidade de medida. Sua equação é:

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad \textit{Eq. 17}$$

Vejamos como fica o gráfico de distribuição de probabilidades de Poisson a partir do próximo exemplo.

Exemplo 7: Para um projeto de estrada de rodagem, uma empresa adquiriu um maquinário capaz de executar 1 km de estrada por dia. A especificação do equipamento admite a ocorrência de 0,0001 defeitos por metro quadrado de estrada. Sabendo-se que o edital prescreve cada trecho de estrada com comprimento de 10 km e 12 metros de largura, monte a distribuição de probabilidade correspondente e indique a probabilidade de ocorrência de três defeitos por km linear de estrada.

Em primeiro lugar, devemos calcular o λ tendo como base o km linear de estrada. Temos a ocorrência de 0,0001 defeitos por m^2. Então, para o km linear temos 12 (largura) x 1000 (comprimento) x 0,0001 = 1,2 ocorrências por km linear de estrada (equivalente a 12.000 m^2 de estrada). O resultado do cálculo com a aplicação da equação 17 é mostrado na Tabela 12.

Qtd defeitos	P(x)
0	0,30119
1	0,36143
2	0,21686
3	**0,08674**
4	0,02602
5	0,00625
6	0,00125
7	0,00021
8	0,00003
9	0,00000
10	0,00000

Tabela 12 - Cálculo das probabilidades de defeitos por km de estrada

A probabilidade de ocorrência de 3 defeitos por km linear de estrada seria aproximadamente 8,674%. O gráfico da distribuição de probabilidades é mostrado na Figura 23.

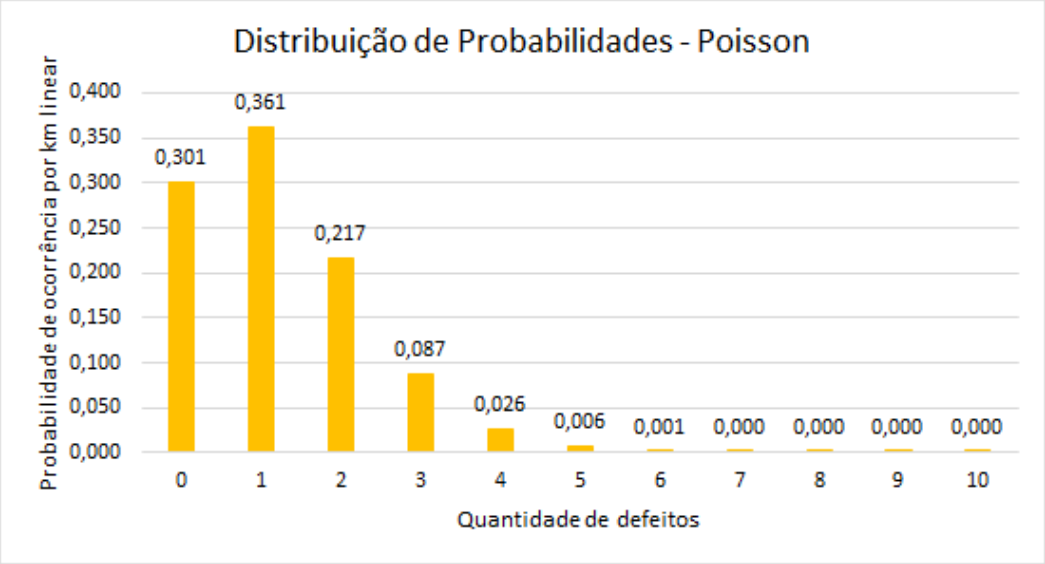

Figura 23 - Distribuição de Poisson

6.5 Distribuição Geométrica

Consideremos uma série de experimentos baseados na distribuição de Bernoulli (fracasso ou sucesso) repetidos até que se obtenha o primeiro sucesso. A probabilidade de sucesso é p e a de fracasso é $q = 1 - p$. Quantos experimentos serão necessários até que se obtenha sucesso? O espaço amostral típico é dado pelo conjunto $\{S, FS, FFS, FFFS, FFFFS. \ldots.\}$. Sendo x o número fracassos antes do primeiro sucesso, a função para a distribuição é:

$$P(x) = q^x p \qquad \textit{Eq. 18}$$

Vamos construir o gráfico de distribuição de probabilidades por meio de um exemplo.

Exemplo 8: Um experimento possui probabilidade de apresentar reação química positiva de 0,3 (30%). Qual a probabilidade de executarmos 5 experimentos antes de obtermos sucesso (cinco fracassos e reação química positiva na sexta tentativa)?

Aplicando a fórmula podemos construir a tabela de probabilidades em função do número de experimentos. O resultado é mostrado na Tabela 13.

x	*P(x)*
0	0,300
1	0,210
2	0,147
3	0,103
4	0,072
5	**0,050**
6	0,035
7	0,025
8	0,017
9	0,012
10	0,008

Tabela 13 - Probabilidades de sucesso após fracassos

Como a Tabela 13 mostra, a probabilidade de sucesso na sexta tentativa (cinco fracassos e um sucesso) é de 0,05 ou 5%. Agora, se alterarmos um pouco o enunciado do problema, teremos outra visão: qual a probabilidade de obtermos sucesso até a sexta tentativa?

Neste caso, a obtenção de sucesso na primeira tentativa (zero fracassos) conta, assim como para a segunda, terceira, quarta, quinta e sexta. A probabilidade total seria a somatória das linhas de zero a cinco. Isto daria 0,882 ou 88,2%.

Observe que o correto entendimento do problema é fundamental para que a análise estatística seja aplicada corretamente. A distribuição de probabilidades associada ao problema é exibida no gráfico da Figura 24.

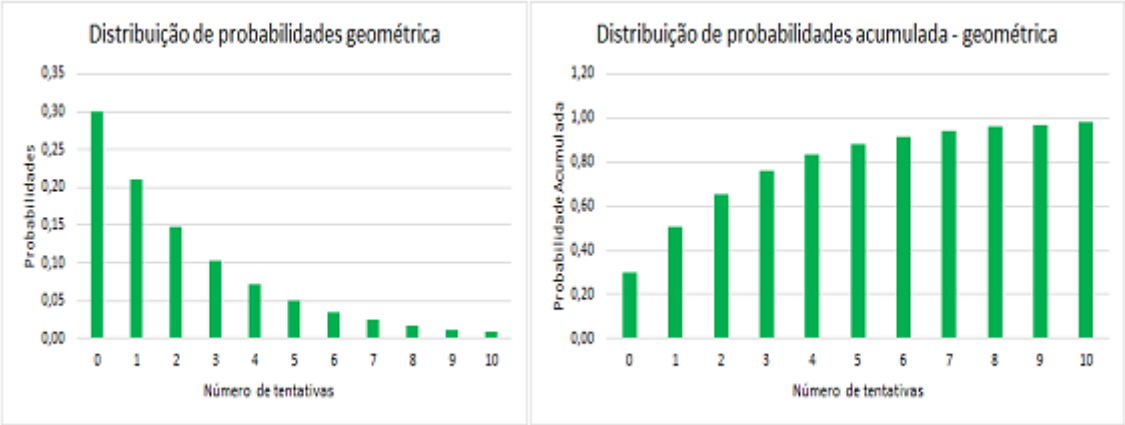

Figura 24 - Distribuições Geométricas

6.6 Distribuição Hipergeométrica

Em alguns casos, a técnica de amostragem a ser aplicada, necessita ser realizada sem a reposição do item, porque o teste de aceitação (sucesso ou fracasso) é realizado à custa do item testado ou porque o processo de seleção da amostra não permite a reposição do item antes da próxima seleção. Para estes casos, a distribuição hipergeométrica é aplicada. O exemplo a seguir ilustra a situação.

Um baralho comum possui 52 cartas, sendo 26 vermelhas e 26 pretas. Se cinco cartas são retiradas aleatoriamente, qual a probabilidade de serem 3 cartas vermelhas e 2 cartas pretas? A retirada é simultânea, então não há possibilidade de reposição das cartas. A quantidade de combinações de cartas vermelhas, três a três, é $C_{(26,3)}$ e a quantidade de combinações possíveis de cartas pretas, duas a duas, é $C_{(26,2)}$, considerando a retirada de cinco cartas. O número total de combinações para a retirada de cinco cartas das 52 cartas do baralho é $C_{(52,5)}$. Assim, a probabilidade de selecionarmos cinco cartas, sem reposição, sendo três vermelhas e duas pretas é:

$$P = \frac{C_{(26,3)}C_{(26,2)}}{C_{(52,5)}} = \frac{[26!/(3!\,23!)][26!/2!\,24!]}{52!/(5!\,47!)} = 0{,}3251$$

Em geral, a distribuição hipergeométrica é aplicada para analisar experimentos para os quais a taxa de sucesso ou fracasso já está determinada, ou seja, para uma população de *N* itens, *k* itens são considerados sucessos e *N* – *k*, fracassos e estamos interessados em determinar a probabilidade de *x* sucessos em uma amostra sem reposição de *n* elementos. A função de densidade de probabilidade $P(x)$ é:

$$h(x;N,n,k) = \frac{C_{(k,x)}C_{(N-k,n-x)}}{C_{(N,n)}} \; onde \; max\{0, n-(N-k)\} \leq x \leq min\{n,k\} \qquad Eq.\ 19$$

Novamente, vamos construir a distribuição de probabilidades por meio de um exemplo.

Exemplo 9: Foram recebidos 30 sacos de cimento CP-V de um determinado fabricante. Considere que cinco sacos possuam cimento com propriedades químicas diferentes. O laboratório possui equipamento e reagentes para testar sete amostras de cimento. Se retirarmos uma amostra de sete elementos colhidas de sacos de cimento diferentes, escolhidos aleatoriamente, qual a probabilidade de termos todas com as mesmas propriedades químicas?

Analisando o exemplo temos:

- N = 30 (população)
- k = 25 (sucesso – propriedades iguais)
- N – k = 5 (fracasso – os cinco sacos de cimento com propriedades diferentes)
- x = n = 7 (tamanho da amostra e quantidade de sucessos esperados na amostra)

$$P(x) = \frac{C_{(25,7)}C_{(30-25,7-7)}}{C_{(30,7)}} = 0{,}2361$$

A probabilidade de termos de duas a sete amostras com as mesmas propriedades, calculadas com a equação 19, é exibida na Tabela 14.

n	*P(x)*
1	0
2	0,00015
3	0,00565
4	0,06214
5	0,26098
6	0,43496
7	0,23612
8	0
9	0
10	0

Tabela 14 - Probabilidades para distribuição hipergeométrica

Com o uso da Tabela 14, a probabilidade de termos pelo menos cinco das sete amostras com propriedades iguais (então, válido para duas, três, quatro e cinco amostras iguais dentre as sete amostras) é igual a probabilidade P(2) + P(3) + P(4) + P(5) = 0,32892 (32,89%). O gráfico de distribuição de probabilidades é mostrado na Figura 25.

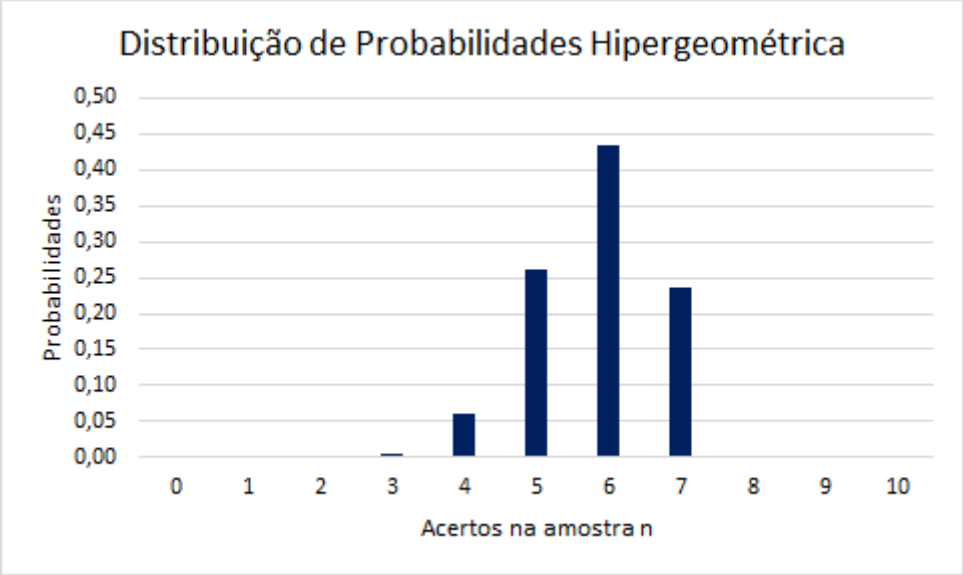

Figura 25 - Distribuição Hipergeométrica

Estes são os principais modelos de distribuição discreta. Existem outros modelos de distribuição discreta que podem ser pesquisados e estudados posteriormente.

A seguir, os principais modelos de distribuição de probabilidades contínuos são apresentados. Recordando, os modelos contínuos são baseados em variáveis aleatórias contínuas, ou seja, em variáveis que podem assumir qualquer valor em um faixa ou segmento pertencente aos números reais.

6.7 Distribuição Normal

A distribuição normal é a mais importante das distribuições de probabilidades. Conhecida como a "curva em forma de sino", a distribuição normal tem sua origem associada aos erros de mensuração. É sabido que quando se efetuam repetidas mensurações de determinada grandeza com um aparelho calibrado, não se chega ao mesmo resultado todas as vezes; obtém-se, ao contrário, um conjunto de valores que oscilam, de modo aproximadamente simétrico, em torno do verdadeiro valor. **Gauss**[11] deduziu matematicamente a distribuição normal como distribuição de probabilidade dos erros de observação, denominando-a então "lei normal dos erros".

A distribuição normal é caracterizada por uma função, cujo gráfico descreve uma curva em forma de sino. Esta distribuição depende de dois parâmetros, a **média** (ou valor esperado) e o **desvio padrão**, conforme mostrado na Figura 26.

[11] Johann Carl Friedrich Gauss nasceu em Braunschweig, Alemanha, no dia 30 de abril de 1777 e faleceu em Göttingen, em 23 de fevereiro de 1855. Foi um matemático, astrônomo e físico alemão que contribuiu muito em diversas áreas da ciência, dentre elas a teoria dos números, estatística, análise matemática, geometria diferencial, geodésia, geofísica, eletroestática, astronomia e ótica

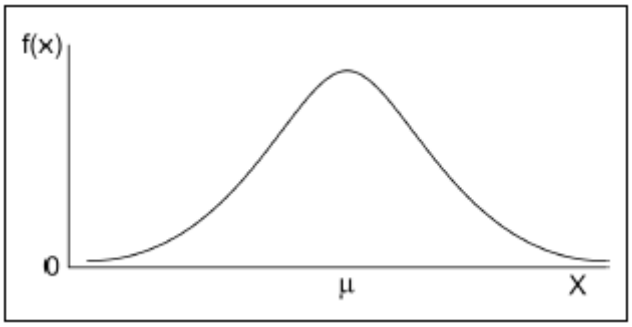

Figura 26 - Gráfico de uma distribuição normal[12]

Uma variável aleatória contínua *X* de média μ e desvio $\sigma^2 > 0$ possui uma distribuição normal se sua função de densidade *f(x)* for:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}, -\infty < x < \infty \qquad \textit{Eq. 20}$$

Propriedades da Distribuição Normal

1. Para uma mesma média **μ** e diferentes desvios padrão **σ**, a distribuição que tem maior desvio padrão se apresenta mais achatada, acusando maior dispersão em torno da média. A que tem menor desvio padrão apresenta "pico" mais acentuado e maior concentração em torno da média, como pode ser visto na Figura 27.

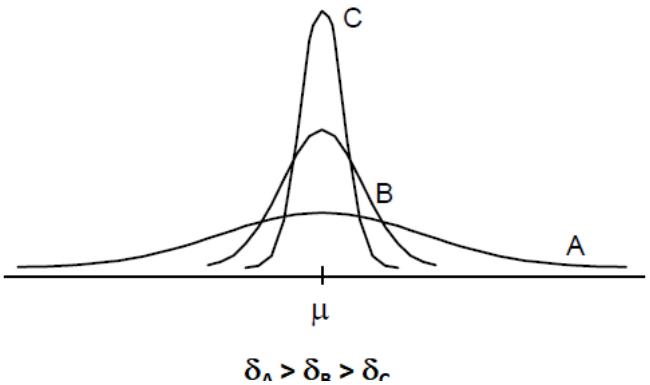

Figura 27 - Curvas normais para mesma média e diferentes desvios padrões

2. Distribuições normais com o mesmo desvio padrão e médias diferentes possuem a mesma dispersão, mas diferem quanto à localização. Quanto maior a média, mais à direita está a curva, como pode ser visto na Figura 28.

[12] Fonte: Portal Action www.portalaction.com.br

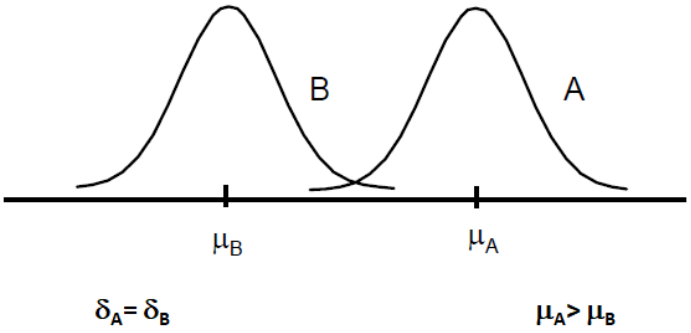

Figura 28 - Curvas normais para diferentes médias com mesmo desvio padrão

3. A probabilidade de uma variável assumir valores entre *a* e *b* é igual à área sob a curva entre esses dois pontos. A determinação dessas probabilidades é realizada matematicamente através da integração da função de densidade de probabilidade entre os pontos *a* e *b* de interesse. No caso da distribuição normal, a Figura 29 apresenta os pontos que pontos que delimitam estas probabilidades.

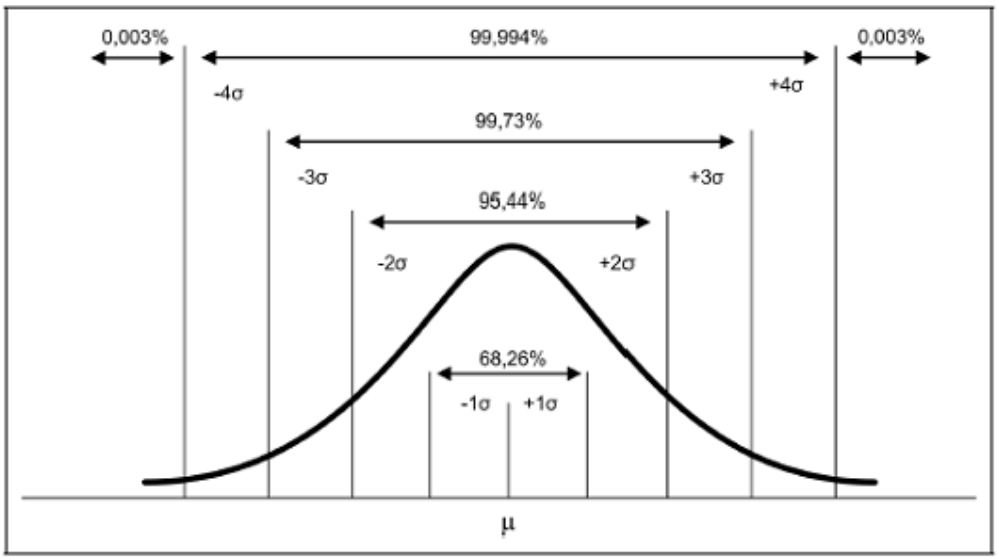

Figura 29 - Propriedades da distribuição normal

O estudo da distribuição normal é importante porque a maioria das variáveis aleatórias de ocorrência natural ou resultante de processos práticos obedece esta distribuição. Desta forma, os resultados de experimentos resultantes de medições (os resultados que normalmente obtemos em nossas pesquisas) seguem uma distribuição normal.

Na maior parte das vezes nas quais o espaço amostral de um experimento envolvendo medições de propriedades não segue uma distribuição normal, as seguintes falhas de planejamento do experimento podem ser encontradas:

- Uso de materiais ou componentes de diferentes fontes, com propriedades físico-químicas diferentes, ocasionando amostras com diferentes características. É como se fosse introduzido um novo fator e este fator não está sendo considerado na análise dos resultados;
- Uso de diferentes equipamentos ou equipes para produzir ou mensurar as amostras. Equipamentos diferentes podem possuir calibragens e precisões diferentes e equipes diferentes podem introduzir

pequenas variações no método, ocasionando diferentes processos de produção ou diferentes resultados em mensuração;
- Uso de métodos não aleatórios para ordenação dos elementos a serem mensurados. Todo equipamento sofre alterações em sua precisão durante o uso. Por exemplo, ao início de um processo de mensuração da resistência à compressão de corpos de prova a prensa pode apresentar uma precisão de 2%. Durante o uso, com o equipamento em funcionamento normal, esta precisão pode variar (1%) e ao final, com os fluidos hidráulicos aquecidos, a precisão pode retornar a 2% (sendo que todos estes valores estão dentro da faixa de trabalho do equipamento).
- Variações no método de produção dos elementos a serem testados por descuido ou desleixo do pesquisador.

Como citado anteriormente, a distribuição normal é a mais importante das distribuições de probabilidades. Nos próximos capítulos voltaremos a abordar, com mais detalhes a inferência e as funções estatísticas aplicadas às distribuições normais.

Além da distribuição normal, é importante conhecer outras formas de distribuição contínuas, seu uso e ocorrências. Vamos apresentar resumidamente as mais importantes.

6.8 Distribuição Qui-Quadrado

Esta distribuição pode ser vista de duas formas diferentes: como a soma de duas distribuições normais ao quadrado ou como um caso particular da distribuição gama (apresentada mais adiante). A distribuição qui-quadrado é utilizada quando estamos analisando a variância de uma amostra que é proveniente de uma população normalmente distribuída.

Definição: Uma variável aleatória contínua *X* segue uma distribuição qui-quadrada com *n* graus de liberdade, denotada por X_n^2, se sua função de densidade for:

$$f(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)}\, x^{n/2-1} e^{-x/2} \;, x > 0, n > 0 \qquad \textit{Eq. 21}$$

Sendo

$$\Gamma(\mathrm{w}) = \int_0^\infty x^{w-1} e^{-x} dx, \;\; w > 0 \qquad \textit{Eq. 22}$$

Podemos notar, pelo gráfico da distribuição qui-quadrado (Figura 30), que esta distribuição é positivamente assimétrica. À medida que os graus de liberdade aumentam, a curva da distribuição aproxima-se da curva normal.

GL – (n) **Graus de liberdade:** este conceito, abordado inicialmente no item 3.2 – Medidas de Dispersão e Variabilidade, é um conceito que deve ser melhor explorado. Graus de liberdade de um conjunto de valores (amostra) representa a quantidade de elementos que podem ter seus valores alterados após terem sido impostas certas restrições a todos os valores. Como exemplo, suponhamos um experimento de resistência à compressão aplicada a uma amostra de oito (8) elementos, cuja média é 40 MPa. Assim, a soma de todas as resistências à compressão é de 240 MPa (restrição → 8 x 40 = 240). Assim, temos um grau de liberdade igual a sete (7 = 8 - 1) pois, sete dos valores podem ser escolhidos aleatoriamente, mas o oitavo deve satisfazer a soma das resistências igual a 240 MPa.

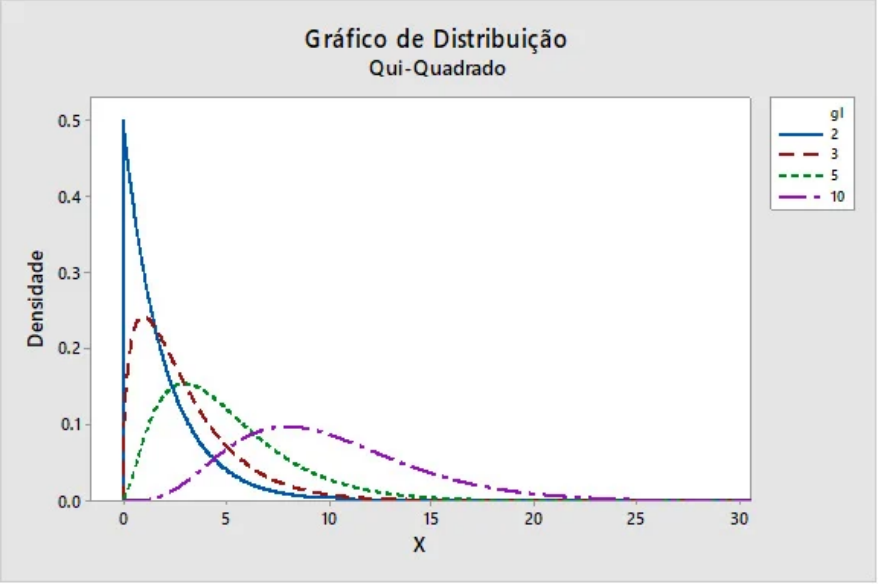

Figura 30 - Distribuição Qui-quadrado

6.9 Distribuição t de Student

Também muito utilizada em estatística, principalmente para modelagem e teste de hipóteses, a distribuição t de Student é uma variação da distribuição normal, com sua característica forma de sino, mas refletindo uma maior variabilidade (com caudas mais alargadas), mais adequada para amostras pequenas (produz valores mais extremos que a distribuição normal).

O único parâmetro que a define e caracteriza é o número de graus de liberdade. Quanto maior for o número de graus de liberdade, mais a curva da distribuição *t* se aproxima da distribuição normal. Sua função de densidade é dada por:

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{(n+1)}{2}} \quad x \in (-\infty, \infty) \qquad \text{Notação } X \sim t_n \qquad \textit{Eq. 23}$$

A Figura 31 apresenta a variação da curva em função da variação dos graus de liberdade.

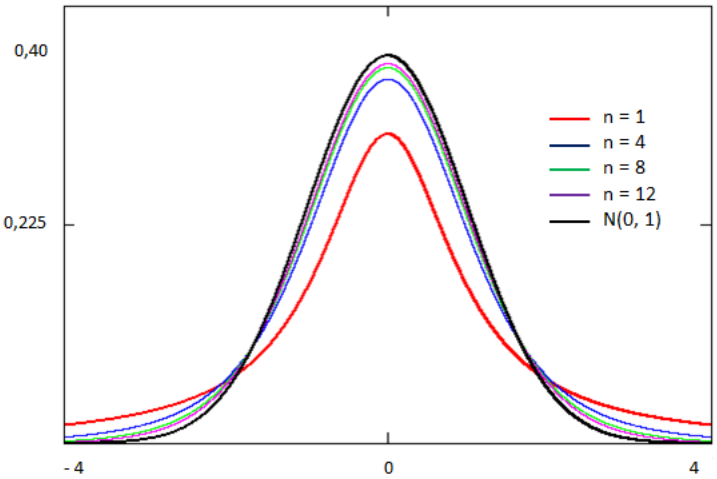

Figura 31 - Distribuição t-Student

6.10 Distribuição Gama

É uma das distribuições mais gerais, pois diversas distribuições são casos particulares dela, como a distribuição exponencial e a distribuição qui-quadrado. Essa distribuição tem como suas principais aplicações à análise de tempo de vida de produtos em engenharia e à distribuição de precipitação de chuva em meteorologia.

A distribuição gama é caracterizada por dois parâmetros: $\alpha > 0$ (denominado parâmetro de forma) e $\beta > 0$, (denominado parâmetro de taxa), denotando-se X ~ Gama (α, β). Sua função de densidade é dada por:

$$f(x) = \frac{\beta^{\alpha} x^{(\alpha-1)} e^{-\beta x}}{\Gamma(\alpha)} \quad se\ x \geq 0 \ \ e\ 0\ caso\ contrário \qquad \textit{Eq. 24}$$

O gráfico da distribuição Gama é apresentado na Figura 32

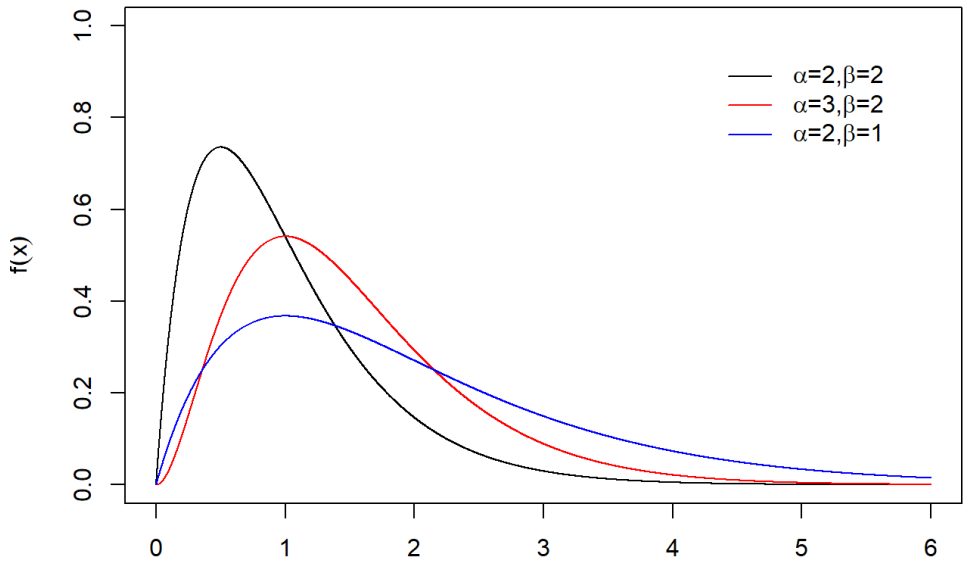

Figura 32 - Distribuição Gama

6.11 Distribuição Exponencial

A distribuição exponencial é caracterizada por ter uma função de taxa de falha constante e é usada como um modelo para o tempo de vida de certos produtos e materiais. Ela descreve adequadamente o tempo de vida de óleos isolantes e dielétricos, entre outros (descreve as probabilidades envolvidas no tempo que decorre para que um determinado evento aconteça, em função de sua vida útil).

Na distribuição exponencial a variável aleatória contínua *x* é definida como o tempo de falha e λ como o tempo médio de vida. Ambos devem ser expressos na mesma unidade, isto é, se o tempo médio de vida é expresso em horas, o tempo de falha também deve ser medido em horas. Sua função de densidade é dada pela equação a seguir e seu gráfico e apresentado na Figura 33.

$$f(x) = 1 - e^{-\lambda x} \quad para\ x \geq 0\ ou\ 0\ para\ x < 0 \qquad \textit{Eq. 25}$$

O exemplo a seguir ilustra o uso da distribuição exponencial: A vida útil de um misturador é estimada em 5 anos (λ = 1/5). Qual a probabilidade de falha nos primeiros dois anos (*x* = 2)?

$$f(x) = 1 - e^{-\frac{1}{5}\cdot 2} = 1 - e^{\frac{-2}{5}} = 0{,}3297 \text{ ou } 32{,}97\%$$

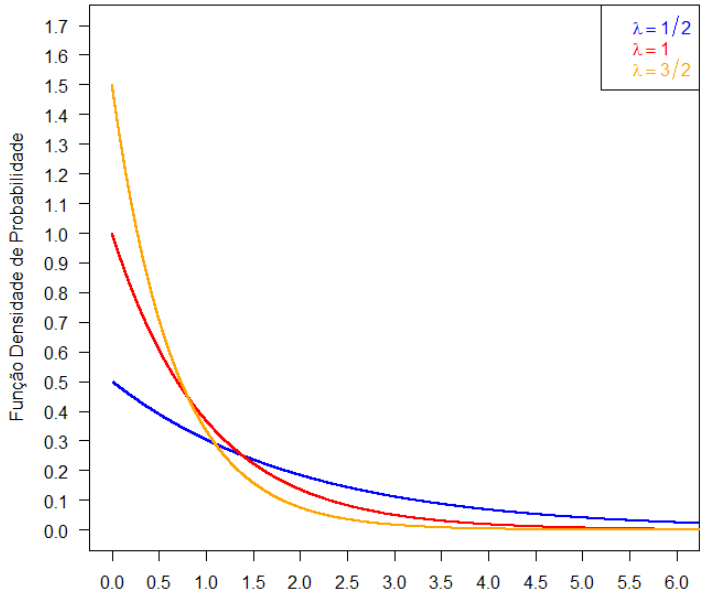

Figura 33 - Distribuição Exponencial – Fonte: www.portalaction.com.br

6.12 Distribuição de Weibull

A distribuição de Weibull é usada em estudos relacionados com o tempo de falha devido a fadiga de metais. Também é frequentemente usada para descrever o tempo de vida de produtos industriais. Seu uso em aplicações práticas é favorecido pelo fato desta distribuição apresentar uma grande variedade de formas, todas com uma propriedade básica: a sua função de taxa de falha é monótona (ou seja, ela é estritamente crescente, estritamente decrescente ou constante). Possui dois parâmetros α, relacionado a escala ou característica da vida e β que é o parâmetro de forma, limite ou inclinação. Sua função de densidade é dada por:

$$f(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}} \quad para\ x > 0 \ \ e\ 0\ caso\ contrário \qquad \textit{Eq. 26}$$

Seu gráfico, para α = 2 e β = 0,5; 1,5 e 3 é mostrado na Figura 34.

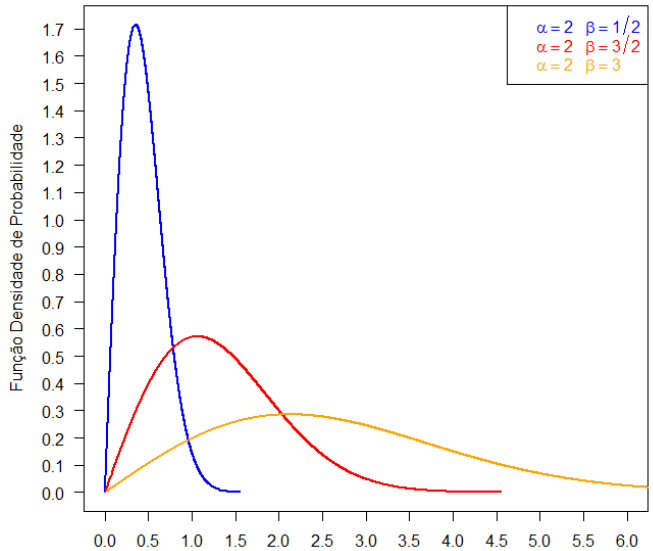

Figura 34 - Distribuição de Weibull – Fonte: www.portalaction.com.br

Existem vários outros tipos de distribuição interessantes, tais como a Log-normal, Logística, Beta e outras. Apresentamos aqui apenas algumas, para que tenhamos uma ideia da forma que suas distribuições de probabilidades se apresentam e, desta forma, identificar a distribuição de probabilidades mais adequada para a análise de seus dados.

Neste processo, o de identificação do tipo de distribuição que está associado aos dados sob análise, o primeiro passo é analisar o próprio experimento e suas respostas (os dados que o experimento gerou). Dados oriundos de medições de propriedades de uma amostra extraída de uma população geralmente possuem distribuição normal. Geralmente, mas nem sempre. Existem testes estatísticos que identificam se os dados, relativos a amostra, possuem distribuição normal. Estes testes são simples de serem aplicados (abordaremos estes testes posteriormente).

Caso estes testes indiquem que a distribuição de probabilidades da variável aleatória, associada a amostra, não é normal, construa a distribuição de probabilidades e analise o tipo de curva que a mesma segue. O tipo de experimento também é uma boa fonte de informações que pode auxiliar. Como visto anteriormente, experimentos associados ao tempo (tempo de vida, ocorrência de falhas) são melhor explicados por outros tipos de distribuição, diferentes da normal. Pesquise.

7 INFERÊNCIA ESTATÍSTICA

Até este ponto do texto, apresentamos os conceitos das medidas de posição (médias, medianas e outras), das medidas de dispersão (variância, desvio padrão, coeficiente de variação), das distribuições de probabilidades dentre outras coisas. Estes são conceitos que caracterizam a amostra e apenas ela.

Agora, como aplicar estes conceitos para a inferência estatística? Como, a partir de uma amostra para qual determinamos a média, o desvio padrão e sua distribuição de probabilidades, transpor estas informações para a população como um todo? Como podemos determinar as probabilidades de um determinado evento?

O primeiro item que o pesquisador deve identificar é a distribuição de probabilidades que os resultados obtidos do experimento seguem. A distribuição de probabilidades é a chave para a determinação correta das funções estatísticas a serem aplicadas e os testes que são usados para identifica-las serão mostrados mais adiante.

As funções estatísticas que apresentaremos agora, cujo objetivo é justamente este, transpor para a população as análises e conclusões retiradas a partir do exame dos dados de uma amostra, em seu conceito, são aplicáveis a qualquer amostra, independente da distribuição de probabilidades que a amostra siga.

No entanto, assim como as funções de probabilidade possuem funções de densidade (equações) diferentes, estas funções também possuem formulações diferentes, adequadas especificamente à cada uma das distribuições de probabilidades. Assim, temos que o conceito da função é sempre o mesmo, mas sua formulação (maneira de ser calculada ou explicitada no software) possui variações para cada uma das distribuições.

As funções serão apresentadas com base na distribuição normal, visto que a maior parte dos resultados (medições) realizadas em experimentos irá seguir este tipo de distribuição e é sobre esta distribuição que encontramos maior quantidade de informações na literatura, facilitando o aprofundamento de sua pesquisa.

Como já foi mostrado na Figura 29, a distribuição de probabilidades da distribuição normal possui uma curva em forma de sino, com as seguintes propriedades, considerando uma característica de interesse *X* medida em uma população com média μ e desvio padrão σ:

- 68,26% dos elementos da população possuem o valor de *x* situado entre μ ± σ;
- 95,46% dos elementos da população possuem o valor de *x* situado entre μ ± 2σ;
- 99,73% dos elementos da população possuem o valor de *x* situado entre μ ± 3σ;
- 99,994% dos elementos da população possuem o valor de *x* situado entre μ ± 4σ.

A partir destas propriedades e do conhecimento das informações de uma **população**, podemos fazer algumas inferências, como no exemplo seguinte.

Exemplo 10: Todos os alunos de Pós-Graduação do CEFET-MG foram mensurados e classificados de acordo com as seguintes variáveis: peso e altura[13], cujas médias e desvios padrões são, respectivamente, 72 kg / 7,2 kg e 175 cm / 17,5 cm. Sabendo-se que estas variáveis seguem uma distribuição normal e são tratadas como independentes, determine as probabilidades de:

1. Alunos com altura inferior a 140 cm;
2. Alunos com peso superior a 93,6 kg;

[13] Dados fictícios

3. Alunos com peso inferior a 79,2 kg e altura superior a 210 cm.

Este é um problema bem fácil de ser resolvido desde que o leiamos com atenção. Não precisamos nem precisamos de funções estatísticas pois todo o conhecimento necessário para sua solução está nos dois parágrafos anteriores.

Primeiro, repare que, no item 1, a diferença de altura desejada (175 – 140 = 35) corresponde a dois desvios padrões, e no item 2 a três desvios padrões. Assim, as próprias propriedades da distribuição normal respondem a estes itens. Para facilitar a visualização, a Figura 35 reapresenta os dados da Figura 29.

1. De acordo com a Figura 35, se 95,46% dos elementos de uma população possuem a altura entre a média e dois desvios, significa que 100% - 95,46% = 4,54% estão fora destes limites, para mais e para menos. Como queremos apenas os com altura inferior a 140 cm, temos que considerar apenas o "para menos", o que leva a divisão do percentual por dois (4,54% / 2 = 2,27%). Assim temos que 2,27% dos alunos de Pós-Graduação do CEFET-MG possuem altura inferior a 140 cm.
2. Neste caso, a diferença de peso corresponde a três desvios padrões (7,2 x 3 = 21,6 kg). 99,73% dos elementos da população estão situados dentro destes limites. Assim, 100 – 99,73 = 0,27% estão fora dele, e, da mesma forma, para cima e para baixo. Como nos interessa apenas os alunos com peso superior, temos que o percentual de alunos com peso superior a 93,6 Kg é de 0,27 / 2 = 0,135 %.
3. Neste item temos uma combinação de probabilidades. A primeira, dos alunos com peso inferior a 79,2 kg o que corresponde à média mais um desvio padrão. Esta probabilidade é melhor visualizada com a ajuda do gráfico da distribuição de probabilidades:

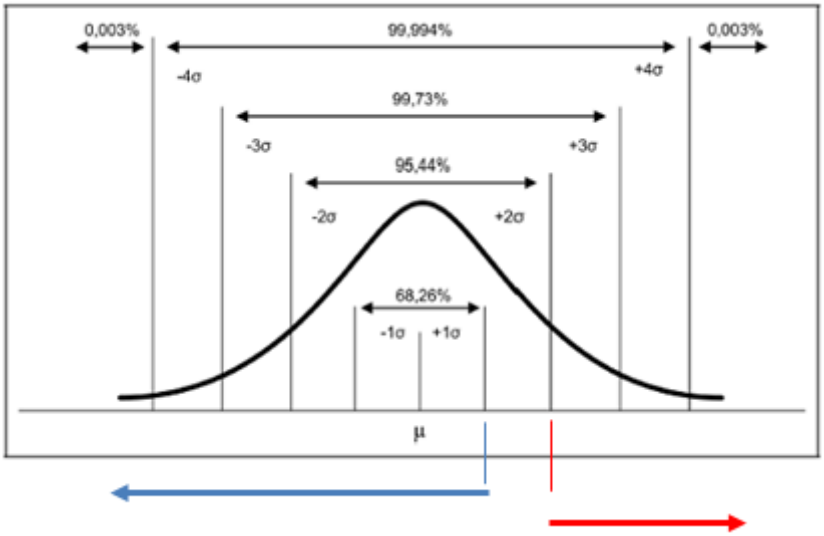

Figura 35 - Distribuição normal - probabilidades

No gráfico podemos visualizar que a linha azul corresponde ao peso inferior a 79,2 kg (a média 72kg somada a um desvio padrão 7,2kg). Como o percentual dos elementos entre a média e um desvio padrão corresponde a 68,26% (novamente, tanto acima quanto abaixo), a metade seria 34,13% (o que corresponde ao percentual entre 72 e 79,2 Kg). Mas como queremos saber o percentual de alunos abaixo de 79,2 kg, temos que incluir os que estão abaixo de 72 kg também (exatamente 50%) o que nos dá: 50 + 34,13 = 84,13%.

No mesmo gráfico, a linha vermelha corresponde aos alunos com altura superior a 210 cm. Para este caso, o raciocínio é o mesmo dos itens 1 e 2. A diferença entre as alturas (35 cm) corresponde a dois desvios padrões, então, como no item 1, apenas 2,27% dos alunos teriam a probabilidade de ter mais do que esta altura.

Agora, tratando-as como variáveis independentes, temos que a probabilidade do evento conjunto seria P1 x P2 = 0,8413 x 0,0227 = 0,0191 ou 1,91%. Assim, teríamos a probabilidade de 1,91 % de encontrarmos alunos da Pós-Graduação com peso inferior a 79,2 kg e altura superior a 210 cm!

O resultado, apesar de estranho (muita altura para pouco peso) seria correto, se não fosse um pequeno problema. E qual é esse problema que invalida a análise realizada?

Bom, para fins didáticos e dentro dos pressupostos apresentados, a análise está correta. Numa situação real, um problema, pequeno, mas extremamente complexo para a análise estatística a torna inválida: no exemplo, as variáveis peso e altura foram consideradas independentes e elas não são. Para um ser humano, o peso está associado à altura. Para a mesma constituição física, quanto maior a altura maior o peso, o que torna estas variáveis dependentes.

Assim, as análises dos itens 1 e 2 estão corretas, mas a do item 3 apresenta o erro grave de considerar independentes duas variáveis dependentes.

Este exemplo "didático" foi apresentado com dois objetivos. O primeiro, de introduzir a questão de probabilidades e o segundo, de mostrar o quão importante é a análise objetiva de todos os fatores envolvidos. Em estatística, uma das principais causas de erro é a não compreensão do problema e, como consequência, a aplicação da técnica ou função incorreta.

7.1 Distribuição Normal Padrão

O exemplo anterior foi bem fácil, com valores determinados para que a solução fosse baseada apenas nas propriedades informadas da distribuição normal. E nos casos reais, onde os valores não são tão "ajustados" assim. Como resolver?

A primeira solução já foi dada anteriormente: "basta montarmos a distribuição de frequência da variável em estudo, deduzirmos a equação de sua curva (função densidade de probabilidade) e calcularmos as áreas totais sob a curva e a área correspondente ao evento [14]".

A segunda é nos aproveitarmos da experiência e conhecimento que nos foram legados por pesquisadores que viveram muito tempo antes de nós (neste caso específico, Johann Carl Friedrich Gauss[11], já citado anteriormente, que em 1809 definiu a lei de Gauss da distribuição normal de erros e sua curva em formato de sino). É o que trataremos a seguir e define os conceitos básicos de inferência estatística.

Uma das (muitas) contribuições de Gauss foi o conceito da distribuição normal padrão. A curva de distribuição normal possui como parâmetros a média e desvio padrão, tornando-a específica para uma população com estas características. Gauss a distribuição padrão, não baseada na média e desvio-padrão e sim na proporção em que os valores se afastam da média, em termos de desvio padrão. Para isto ele propôs uma distribuição normal padrão baseada na seguinte equação:

$$z = \frac{x - \mu}{\sigma} \qquad \textit{Eq. 27}$$

Com esta equação podemos representar a distribuição normal como uma distribuição normal padrão como mostrado Figura 36. A distribuição passa a apresentar as probabilidades em função do desvio dos valores (*X*) em relação à média em função de valores do desvio padrão.

[14] Ver item 6 - DISTRIBUIÇÃO DE PROBABILIDADES

Esta curva especifica de distribuição de frequência (ou padronizada, uma vez que independe dos valores dos nossos dados) possui média 0 (zero) e desvio padrão 1 (um) e é chamada de **distribuição normal padrão**.

Assim, com o objetivo de facilitar a obtenção de determinadas áreas sob a curva normal, podemos transformar qualquer distribuição de probabilidades normal *F(X)* em uma distribuição normal padrão, com média 0 (zero) e desvio padrão 1 (um).

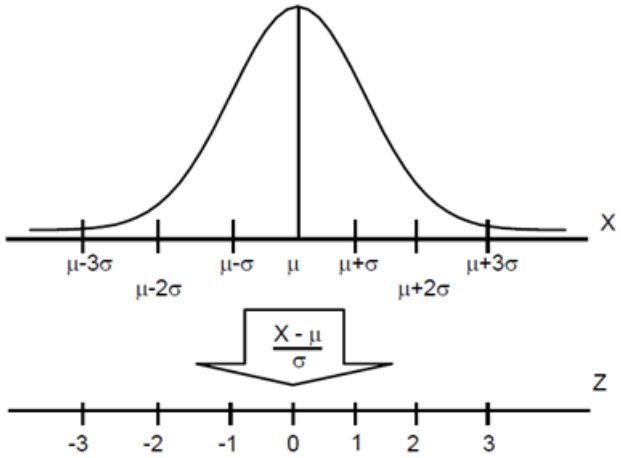

Figura 36 - Distribuição Normal Padrão

E qual a contribuição disto? Simples, Gauss determinou as probabilidades em função da variável *Z* e as apresentou em tabela, facilitando (e muito) o cálculo das probabilidades. O valor *Z* é conhecido como valor padronizado e é uma medida relativa. Mede o quanto *X* se afasta da média, em unidades de desvio padrão. Os valores de *Z* podem ser obtidos a partir de tabelas, como a tabela mostrada na Figura 37.

Como a curva normal (e, logicamente a curva normal padrão também) é simétrica, a tabela também é simétrica. Por exemplo, o valor da probabilidade para z = 1,5 é p = 0,9332. Então o valor de z = -1,5 tem que ser igual a (1 – 0,9332 = 0,0668), o que pode ser conferido facilmente na própria tabela. Desta forma, em algumas fontes, encontramos esta tabela com apenas os valores positivos de *z*.

A cada parte da tabela da Figura 36 (para z positivo e negativo) é dividida em 10 colunas. A primeira coluna apresenta o valor de z com uma casa decimal. As nove colunas seguintes (ver cabeçalho das colunas) acrescentam a segunda casa decimal. O valor de z = -2,75 será encontrado na linha com z = -2,7 na coluna com cabeçalho 0,05 (p = 0,0030). Para mais casas decimais é necessário fazer interpolação entre os valores. É aproximado, mas resolve.

Caso não queiramos interpolar, há diversas outras maneiras de descobrirmos a probabilidade em função do valor z. Com o uso de computador e o software apropriado, há diversas opções. No MS Excel, por exemplo, a função *DISTNORMP(z)* dá a probabilidade associada ao valor z (a tabela anterior foi calculada usando este método). No software estatístico R, a função *pnorm(x, mean, sd, ...)* fornece a probabilidade em função do valor x, da média e do desvio padrão.

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997

Z	0,00	-0,01	-0,02	-0,03	-0,04	-0,05	-0,06	-0,07	-0,08	-0,09
-3,9	0,00005	0,00005	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00003	0,00003
-3,8	0,00007	0,00007	0,00007	0,00006	0,00006	0,00006	0,00006	0,00005	0,00005	0,00005
-3,7	0,00011	0,00010	0,00010	0,00010	0,00009	0,00009	0,00008	0,00008	0,00008	0,00008
-3,6	0,00016	0,00015	0,00015	0,00014	0,00014	0,00013	0,00013	0,00012	0,00012	0,00011
-3,5	0,00023	0,00022	0,00022	0,00021	0,00020	0,00019	0,00019	0,00018	0,00017	0,00017
-3,4	0,00034	0,00032	0,00031	0,00030	0,00029	0,00028	0,00027	0,00026	0,00025	0,00024
-3,3	0,00048	0,00047	0,00045	0,00043	0,00042	0,00040	0,00039	0,00038	0,00036	0,00035
-3,2	0,00069	0,00066	0,00064	0,00062	0,00060	0,00058	0,00056	0,00054	0,00052	0,00050
-3,1	0,00097	0,00094	0,00090	0,00087	0,00084	0,00082	0,00079	0,00076	0,00074	0,00071
-3,0	0,00135	0,00131	0,00126	0,00122	0,00118	0,00114	0,00111	0,00107	0,00104	0,00100
-2,9	0,00187	0,00181	0,00175	0,00169	0,00164	0,00159	0,00154	0,00149	0,00144	0,00139
-2,8	0,00256	0,00248	0,00240	0,00233	0,00226	0,00219	0,00212	0,00205	0,00199	0,00193
-2,7	0,00347	0,00336	0,00326	0,00317	0,00307	0,00298	0,00289	0,00280	0,00272	0,00264
-2,6	0,00466	0,00453	0,00440	0,00427	0,00415	0,00402	0,00391	0,00379	0,00368	0,00357
-2,5	0,00621	0,00604	0,00587	0,00570	0,00554	0,00539	0,00523	0,00508	0,00494	0,00480
-2,4	0,00820	0,00798	0,00776	0,00755	0,00734	0,00714	0,00695	0,00676	0,00657	0,00639
-2,3	0,01072	0,01044	0,01017	0,00990	0,00964	0,00939	0,00914	0,00889	0,00866	0,00842
-2,2	0,01390	0,01355	0,01321	0,01287	0,01255	0,01222	0,01191	0,01160	0,01130	0,01101
-2,1	0,01786	0,01743	0,01700	0,01659	0,01618	0,01578	0,01539	0,01500	0,01463	0,01426
-2,0	0,02275	0,02222	0,02169	0,02118	0,02068	0,02018	0,01970	0,01923	0,01876	0,01831
-1,9	0,02872	0,02807	0,02743	0,02680	0,02619	0,02559	0,02500	0,02442	0,02385	0,02330
-1,8	0,03593	0,03515	0,03438	0,03362	0,03288	0,03216	0,03144	0,03074	0,03005	0,02938
-1,7	0,04457	0,04363	0,04272	0,04182	0,04093	0,04006	0,03920	0,03836	0,03754	0,03673
-1,6	0,05480	0,05370	0,05262	0,05155	0,05050	0,04947	0,04846	0,04746	0,04648	0,04551
-1,5	0,06681	0,06552	0,06426	0,06301	0,06178	0,06057	0,05938	0,05821	0,05705	0,05592
-1,4	0,08076	0,07927	0,07780	0,07636	0,07493	0,07353	0,07215	0,07078	0,06944	0,06811
-1,3	0,09680	0,09510	0,09342	0,09176	0,09012	0,08851	0,08691	0,08534	0,08379	0,08226
-1,2	0,11507	0,11314	0,11123	0,10935	0,10749	0,10565	0,10383	0,10204	0,10027	0,09853
-1,1	0,13567	0,13350	0,13136	0,12924	0,12714	0,12507	0,12302	0,12100	0,11900	0,11702
-1,0	0,15866	0,15625	0,15386	0,15151	0,14917	0,14686	0,14457	0,14231	0,14007	0,13786
-0,9	0,18406	0,18141	0,17879	0,17619	0,17361	0,17106	0,16853	0,16602	0,16354	0,16109
-0,8	0,21186	0,20897	0,20611	0,20327	0,20045	0,19766	0,19489	0,19215	0,18943	0,18673
-0,7	0,24196	0,23885	0,23576	0,23270	0,22965	0,22663	0,22363	0,22065	0,21770	0,21476
-0,6	0,27425	0,27093	0,26763	0,26435	0,26109	0,25785	0,25463	0,25143	0,24825	0,24510
-0,5	0,30854	0,30503	0,30153	0,29806	0,29460	0,29116	0,28774	0,28434	0,28096	0,27760
-0,4	0,34458	0,34090	0,33724	0,33360	0,32997	0,32636	0,32276	0,31918	0,31561	0,31207
-0,3	0,38209	0,37828	0,37448	0,37070	0,36693	0,36317	0,35942	0,35569	0,35197	0,34827
-0,2	0,42074	0,41683	0,41294	0,40905	0,40517	0,40129	0,39743	0,39358	0,38974	0,38591
-0,1	0,46017	0,45620	0,45224	0,44828	0,44433	0,44038	0,43644	0,43251	0,42858	0,42465
0,0	0,50000	0,49601	0,49202	0,48803	0,48405	0,48006	0,47608	0,47210	0,46812	0,46414

Figura 37 - Tabela Normal de Probabilidades

Seguem alguns exemplos de cálculo de probabilidades (para a população, lembre-se) usando estes métodos.

1. Uma concreteira produz um determinado tipo de concreto com **μ = 210** Mpa e **δ = 5** Mpa. Qual a probabilidade de que um corpo de prova cilíndrico tenha resistência inferior a 200 Mpa?

   Para os dois primeiros métodos, o primeiro passo é o cálculo do valor de z. Assim:

   $$z = \frac{x - \mu}{\sigma} = \frac{200 - 210}{5} = -2{,}0$$

   Na tabela normal padrão, o valor da probabilidade para z = $-2{,}0$ é p = 0,0228 indicando uma probabilidade de 2,28% de que um corpo de prova tenha resistência inferior a 200 MPa.

   $$P(X < 200) = p(z) < \frac{x - \mu}{\sigma} \therefore p(z < -2) = -0{,}0228$$

   No MS Excel, basta digitarmos em uma célula de uma planilha a fórmula "=DISTNORMP(-2)", agora usando todas as casas decimais. O resultado é 0,02275 ou 2,275% de probabilidade (pouca diferença do valor obtido usando a tabela, não?).

   No software R, não precisamos de calcular o valor de z. A função pode ser digitada diretamente no console como *pnorm (200, mean = 210, sd = 5)* e o resultado é 0.02275013.

2. Por contrato, o concreto fornecido pela empreiteira X deve ter resistência a compressão superior a 38 Mpa. Sabe-se que a resistência à compressão média é de 40 MPa com desvio padrão de 2,65 MPa. Qual a probabilidade de o concreto fornecido possuir resistência a compressão inferior a 38 Mpa?

   Novamente, iniciamos com o cálculo do valor de z. Assim:

   $$z = \frac{x - \mu}{\sigma} = \frac{38 - 40}{2{,}65} = -0{,}7547 \approx -0{,}75$$

   Na tabela normal padrão, o valor para z = $-0{,}75$ é 0,022663 indicando uma probabilidade de 2,27% de que um corpo de prova tenha resistência inferior a 38 MPa.

   $$P(X < 38) = p(z) < \frac{x - \mu}{\sigma} \therefore p(z < -0{,}75) = -0{,}022663$$

   No MS Excel, basta digitarmos em uma célula de uma planilha a fórmula "=DISTNORMP(-0,75)", agora usando todas as casas decimais. O resultado é 0,0226627 ou 2,227% de probabilidade (novamente, pequena diferença do valor obtido usando a tabela).

   No software R, digitando a função diretamente no console como *pnorm (38, mean = 40, sd = 2.65)* temos o resultado de 0.02252094 o que equivale a 2,252%.

3. O concreto produzido por uma empreiteira possui as seguintes características: μ = 110 MPa e σ = 10 MPa. Qual a probabilidade de obtermos concreto com resistência a compressão de 100 MPa?

   Este exemplo foi colocado aqui estimular um pouco o pensamento. Não desejamos saber a probabilidade de obtenção de concreto com resistência menor que 100 MPa e sim com resistência igual a 100. Como fazer? Bom, podemos tentar com um artifício: considerar que todo concreto com resistência entre 99 e 101 MPa representa o concreto com resistência de 100 MPa. Em teoria, p(99<x<101) = p(x < 101) – p(x < 99). Vamos tentar resolver isto no software R com este intervalo (1 MPa):

```
> pnorm(101,110,10)
[1] 0.1840601
> pnorm(99,110,10)
[1] 0.1356661
> pnorm(101,110,10)-pnorm(99,110,10)
[1] 0.04839406
```

   Bom, a probabilidade seria de 4,84%. Mas vamos tentar reduzir mais o intervalo, para 0,5 MPa e conferir o resultado.

```
> pnorm(100.5,110,10)
[1] 0.1710561
> pnorm(99.5,110,10)
[1] 0.1468591
> pnorm(100.5,110,10)-pnorm(99.5,110,10)
[1] 0.02419707
```

   A probabilidade foi reduzida para a metade (2,42%). Por sorte o software R possui outras funções de probabilidade, tal como a *dnorm()* – densidade de probabilidade – que nos informa a probabilidade em um determinado ponto. Vamos conferir o seu resultado.

```
> dnorm(100,110,10)
[1] 0.02419707
```

Como visto, a probabilidade determinada pela função *dnorm* para os dados do exemplo foi igual à probabilidade obtida quando usamos o intervalo de 99,5 a 100,5 MPa. Assim, podemos concluir que a função *dnorm* não retorna o valor da probabilidade para o valor exato de 100 MPa e sim para o intervalo de 99,5 a 100,5 MPa, pois a probabilidade para o valor exato de 100 MPa é zero.

Este exemplo é interessante pois nos permite discutir o que é um valor para o software *RStudio*. Quando especificamos que a resistência deveria ser igual a 100 MPa o que a função *dnorm* considerou? De 99 a 101 ou de 99,5 a 100,5 (esta, de acordo com os resultados anteriores), ou ainda, de 99,99999 a 100,00001? A probabilidade está associada a faixa definida. Vamos supor que a última faixa seja a faixa solicitada. Qual seria a probabilidade?

```
> pnorm(100.00001,110,10)-pnorm(99.99999,110,10)
[1] 4.839414e-07
```

A probabilidade seria de 0,000000484%. Teoricamente, para o valor exato de 100 MPa, a probabilidade seria de 0%, pois tratamos de valores contínuos e probabilidade de termos um resultado igual ao especificado, com infinitas casas decimais é zero %.

Lembrete: Nós trabalhamos neste capítulo com populações. Os símbolos μ e σ significam média populacional e desvio padrão populacional. Ou seja, temos informações sobre a população como um todo. Apesar de estarmos calculando probabilidades, estamos fazendo isto com dados populacionais. A distribuição padrão normal e o valor *z* apresentam informações sobre a população. No próximo capítulo, abordaremos amostras e a inferência, a transposição das conclusões obtidas a partir da análise dos dados da amostra para a população.

Sugestão de Pesquisa para ampliar conhecimento: O software R possui funções associadas as distribuições de probabilidades. Duas foram vistas nos exemplos anteriores: *pnorm* e *dnorm*. Existem outras e podem ser aplicadas a outras distribuições de probabilidades. As funções são indicadas pela primeira letra (*p*, *d*, *q* e *r*) *seguidas* pelo tipo de distribuição a ser aplicada (no exemplo, *norm*). As funções são:

- Função densidade (ou probabilidade): calcula o valor da densidade, para funções contínuas, ou da probabilidade, para funções discretas, para cada elemento x. Indicada pela letra d.
- Função distribuição: calcula a distribuição acumulada ($p \leq x$). Indicada pela letra p.
- Função probabilidade: calcula o valor de x correspondente a probabilidade acumulada (inverso da função distribuição). Indicada pela letra q.
- Função gerador aleatório: gera números aleatórios para a distribuição escolhida. Indicada pela letra r.

Os tipos de distribuição que podem ser associadas a estas funções (normalmente já pré-carregadas no R) são apresentadas a seguir. Cada uma destas funções possui parâmetros distintos. Pesquise os parâmetros de cada uma delas e teste com exemplos:

- (norm) – distribuição normal;
- (binom) – distribuição binomial;
- (pois) – distribuição de Poisson;
- (geom) – distribuição geométrica;
- (hyper) – distribuição hipergeométrica;
- (unif) – distribuição uniforme;
- (exp) – distribuição exponencial;
- (gamma) – distribuição gama;
- (chisq) – distribuição qui-quadrado;
- (t) – distribuição t-Student.

7.2 Distribuição t-Student

A distribuição normal aplica-se quando temos informações sobre a população ou quando nossa amostra contém quantidade de elementos suficiente para que possamos considera-la como "representativa da população". Em estudos que envolvem populações que podem ser definidas (classificadas e contadas), existem fórmulas específicas, baseadas no grau de confiabilidade, no erro máximo de estimativa admitido, na média e desvio padrão populacional que nos permitem calcular a quantidade mínima de elementos necessárias para a amostra.

Quando não temos estas informações podemos adotar outras estratégias para a inferência sobre a população. A distribuição t-Student é uma delas. Quando usamos a distribuição normal para amostras pequenas (**n ≤ 30**) são obtidos valores de probabilidades menos precisos. Assim, adota-se distribuição t-Student.

A distribuição t possui a mesma forma da distribuição normal (em forma de sino) e é simétrica sobre a média. A diferença é que a distribuição t tem caudas mais largas (mais áreas nas caudas), fazendo com que seus valores críticos sejam maiores que os da distribuição normal. É como pagar um preço maior por trabalhar com pequenas amostras.

Outro fator importante sobre a distribuição t é que ela é construída em função dos graus de liberdade (já visto anteriormente) e estes estão diretamente relacionados com o tamanho n da amostra. Para cada grau de liberdade há uma curva diferente. Quanto menor os graus de liberdade, mais larga será a cauda. Quanto maior, mais a curva se aproxima da curva normal (recomenda-se, para n > 30, usar a curva normal).

Assim, a tabela t-Student é construída com muito menos dados e é necessário muito mais interpolações. Imagine, se ela fosse construída igual a tabela da distribuição normal, abrangendo amostras de 2 a 31 elementos (de 1 a 30 graus de liberdade) teríamos 30 tabelas de distribuição, similares a tabela de distribuição normal.

A tabela t-Student é exibida na Figura 38 e a a seguir explicaremos suas propriedades e seu uso.

Como pode ser visto (e deve ser entendido, claro), é como se cada linha desta tabela (cada grau de liberdade) representasse todas as informações que foram apresentadas na tabela normal padrão. São 600 dados resumidos em 12. Então, é óbvio que teremos que fazer interpolações para encontrar valores diferentes dos que constam nos cabeçalhos de linhas e colunas.

E quais são as diferenças e similaridades entres estas tabelas. Existem algumas, mas são fáceis de serem assimiladas. Iniciando do cabeçalho, temos:

Na tabela normal padrão, a combinação do cabeçalho de linha com o cabeçalho de coluna representa um valor de probabilidade. Na tabela t-Student, o valor de probabilidade é representado no cabeçalho das colunas apenas. E temos duas linhas de cabeçalho, a primeira unicaudal e a segunda bicaudal, que podemos interpretar como mostrado a seguir.

O valor apresentado no conteúdo da tabela representa o valor de $|t(n-1; 1-\alpha/2)|$, ou seja, o módulo do valor encontrado na linha correspondente a n -1 graus de liberdade (sendo n igual à quantidade de elementos na amostra) e $\alpha/2$ a probabilidade especificada.

α - Unicaudal	0,25	0,20	0,15	0,1	0,07	0,06	0,05	0,04	0,03	0,025	0,02	0,015	0,01	0,005
α - Bicaudal	0,50	0,40	0,30	0,20	0,14	0,12	0,10	0,08	0,06	0,05	0,04	0,03	0,02	0,01
GL (n-1)														
1	1,000	1,376	1,963	3,078	4,474	5,242	6,314	7,916	10,579	12,706	15,895	21,205	31,821	63,657
2	0,816	1,061	1,386	1,886	2,383	2,620	2,920	3,320	3,896	4,303	4,849	5,643	6,965	9,925
3	0,765	0,978	1,250	1,638	1,995	2,156	2,353	2,605	2,951	3,182	3,482	3,896	4,541	5,841
4	0,741	0,941	1,190	1,533	1,838	1,971	2,132	2,333	2,601	2,776	2,999	3,298	3,747	4,604
5	0,727	0,920	1,156	1,476	1,753	1,873	2,015	2,191	2,422	2,571	2,757	3,003	3,365	4,032
6	0,718	0,906	1,134	1,440	1,700	1,812	1,943	2,104	2,313	2,447	2,612	2,829	3,143	3,707
7	0,711	0,896	1,119	1,415	1,664	1,770	1,895	2,046	2,241	2,365	2,517	2,715	2,998	3,499
8	0,706	0,889	1,108	1,397	1,638	1,740	1,860	2,004	2,189	2,306	2,449	2,634	2,896	3,355
9	0,703	0,883	1,100	1,383	1,619	1,718	1,833	1,973	2,150	2,262	2,398	2,574	2,821	3,250
10	0,700	0,879	1,093	1,372	1,603	1,700	1,812	1,948	2,120	2,228	2,359	2,527	2,764	3,169
11	0,697	0,876	1,088	1,363	1,591	1,686	1,796	1,928	2,096	2,201	2,328	2,491	2,718	3,106
12	0,695	0,873	1,083	1,356	1,580	1,674	1,782	1,912	2,076	2,179	2,303	2,461	2,681	3,055
13	0,694	0,870	1,079	1,350	1,572	1,664	1,771	1,899	2,060	2,160	2,282	2,436	2,650	3,012
14	0,692	0,868	1,076	1,345	1,565	1,656	1,761	1,887	2,046	2,145	2,264	2,415	2,624	2,977
15	0,691	0,866	1,074	1,341	1,558	1,649	1,753	1,878	2,034	2,131	2,249	2,397	2,602	2,947
16	0,690	0,865	1,071	1,337	1,553	1,642	1,746	1,869	2,024	2,120	2,235	2,382	2,583	2,921
17	0,689	0,863	1,069	1,333	1,548	1,637	1,740	1,862	2,015	2,110	2,224	2,368	2,567	2,898
18	0,688	0,862	1,067	1,330	1,544	1,632	1,734	1,855	2,007	2,101	2,214	2,356	2,552	2,878
19	0,688	0,861	1,066	1,328	1,540	1,628	1,729	1,850	2,000	2,093	2,205	2,346	2,539	2,861
20	0,687	0,860	1,064	1,325	1,537	1,624	1,725	1,844	1,994	2,086	2,197	2,336	2,528	2,845
21	0,686	0,859	1,063	1,323	1,534	1,621	1,721	1,840	1,988	2,080	2,189	2,328	2,518	2,831
22	0,686	0,858	1,061	1,321	1,531	1,618	1,717	1,835	1,983	2,074	2,183	2,320	2,508	2,819
23	0,685	0,858	1,060	1,319	1,529	1,615	1,714	1,832	1,978	2,069	2,177	2,313	2,500	2,807
24	0,685	0,857	1,059	1,318	1,526	1,612	1,711	1,828	1,974	2,064	2,172	2,307	2,492	2,797
25	0,684	0,856	1,058	1,316	1,524	1,610	1,708	1,825	1,970	2,060	2,167	2,301	2,485	2,787
26	0,684	0,856	1,058	1,315	1,522	1,608	1,706	1,822	1,967	2,056	2,162	2,296	2,479	2,779
27	0,684	0,855	1,057	1,314	1,521	1,606	1,703	1,819	1,963	2,052	2,158	2,291	2,473	2,771
28	0,683	0,855	1,056	1,313	1,519	1,604	1,701	1,817	1,960	2,048	2,154	2,286	2,467	2,763
29	0,683	0,854	1,055	1,311	1,517	1,602	1,699	1,814	1,957	2,045	2,150	2,282	2,462	2,756
30	0,683	0,854	1,055	1,310	1,516	1,600	1,697	1,812	1,955	2,042	2,147	2,278	2,457	2,750
31	0,682	0,853	1,054	1,309	1,515	1,599	1,696	1,810	1,952	2,040	2,144	2,275	2,453	2,744
32	0,682	0,853	1,054	1,309	1,513	1,597	1,694	1,808	1,950	2,037	2,141	2,271	2,449	2,738
33	0,682	0,853	1,053	1,308	1,512	1,596	1,692	1,806	1,948	2,035	2,138	2,268	2,445	2,733
34	0,682	0,852	1,052	1,307	1,511	1,595	1,691	1,805	1,946	2,032	2,136	2,265	2,441	2,728
35	0,682	0,852	1,052	1,306	1,510	1,594	1,690	1,803	1,944	2,030	2,133	2,262	2,438	2,724
36	0,681	0,852	1,052	1,306	1,509	1,593	1,688	1,802	1,942	2,028	2,131	2,260	2,434	2,719
37	0,681	0,851	1,051	1,305	1,508	1,592	1,687	1,800	1,940	2,026	2,129	2,257	2,431	2,715
38	0,681	0,851	1,051	1,304	1,507	1,591	1,686	1,799	1,939	2,024	2,127	2,255	2,429	2,712
39	0,681	0,851	1,050	1,304	1,506	1,590	1,685	1,798	1,937	2,023	2,125	2,252	2,426	2,708
40	0,681	0,851	1,050	1,303	1,506	1,589	1,684	1,796	1,936	2,021	2,123	2,250	2,423	2,704
50	0,679	0,849	1,047	1,299	1,500	1,582	1,676	1,787	1,924	2,009	2,109	2,234	2,403	2,678
60	0,679	0,848	1,045	1,296	1,496	1,577	1,671	1,781	1,917	2,000	2,099	2,223	2,390	2,660
70	0,678	0,847	1,044	1,294	1,493	1,574	1,667	1,776	1,912	1,994	2,093	2,215	2,381	2,648
80	0,678	0,846	1,043	1,292	1,491	1,572	1,664	1,773	1,908	1,990	2,088	2,209	2,374	2,639
90	0,677	0,846	1,042	1,291	1,489	1,570	1,662	1,771	1,905	1,987	2,084	2,205	2,368	2,632
100	0,677	0,845	1,042	1,290	1,488	1,568	1,660	1,769	1,902	1,984	2,081	2,201	2,364	2,626

Figura 38 - Tabela t-Student

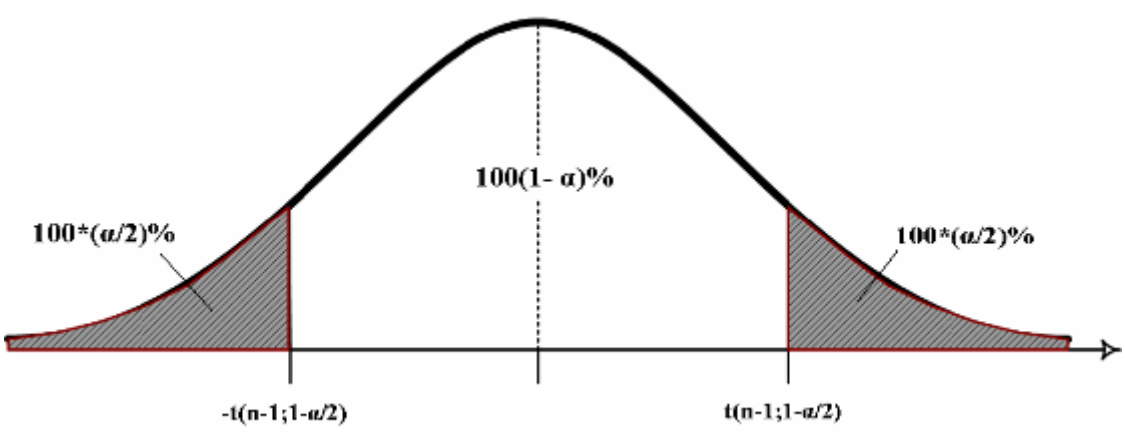

Figura 39 - Áreas de Probabilidades na Distribuição t-Student

O valor de t é calculado de forma similar ao valor z (da tabela de distribuição normal padrão). Sua equação é dada por:

$$T = \frac{x - \bar{X}}{s/\sqrt{n}} \qquad \textit{Eq. 28}$$

Cada linha da tabela, dada por um valor de grau de liberdade, representa uma distribuição de probabilidades, específica para aquele grau de liberdade. Por exemplo:

1. Uma concreteira produz um determinado tipo de concreto com $\bar{X}$ = 210 Mpa e **s = 5** Mpa. Qual a probabilidade de que um corpo de prova cilíndrico tenha resistência inferior a 200 Mpa, sabendo-se que a amostra possui 4 elementos?

   Com estes dados, o valor de T é: $T = \frac{200-210}{5/\sqrt{4}}$ = 4,0. Consultando o valor t(3;$-\alpha/2$) = 4 correspondente a linha de três graus de liberdade, temos que está entre 3,896 (correspondendo a 0,015) e 4,541 (correspondendo a 0,01). Fazendo a interpolação conseguimos um valor próximo de 0,014 (1,4%).

   Este valor (1,4%) é inferior ao encontrado quando usamos a distribuição normal (2,275%). É o preço a se pagar por trabalhar com amostras pequenas.

   Bom, então vamos aumentar o número de elementos da amostra. Suponhamos uma amostra de 8 elementos. Assim: $T = \frac{200-210}{5/\sqrt{8}} = 5{,}657$. Agora trabalhamos com 7 graus de liberdade (n – 1). Consultando a linha correspondente (7 graus de liberdade) temos que o maior valor de t é 3,499 correspondendo a 0,005 (0,5%). Isto significa que a probabilidade de encontrarmos um corpo de prova com resistência inferior a 200 MPa é inferior a 0,5%, considerando que a média e o desvio foram obtidos a partir de uma amostra de 8 elementos.

   O MS Excel possui função para cálculo da probabilidade associada à distribuição t-Student. É a função DISTT(t; graus de liberdade; número de caudas). Se a usarmos para o valor acima (t = 5,657 → DISTT(5,657;7;1)) o resultado será 0,000383, indicando a probabilidade de 0,0383%.

   No software R, a função é *pt(t, graus de liberdade, lower.tail = TRUE)*. O resultado é 0,0003773162 (0,0377%).

Refaça os outros exemplos usando a distribuição de t-Student.

7.3 Identificação da Distribuição de Probabilidades

Como citado anteriormente, o primeiro passo para o emprego de funções estatísticas na análise dos resultados de um experimento é a identificação da distribuição de probabilidades que os resultados seguem. Como, a distribuição normal é a mais comum entre os resultados de experimentos, iniciamos por ela.

Os testes utilizados para identificar se a distribuição de probabilidades associada a um conjunto de dados pode ser aproximada pela distribuição normal são chamados de testes de normalidade. As principais técnicas são: o papel da probabilidade, o teste de Kolmogorov – Smirnov, o teste de Anderson – Darling e o teste de Shapiro – Wilk. Existem vários outros, cada um com características próprias de uso e diferentes capacidades de associação com uma curva normal padrão.

OUTLIERS

Antes de verificarmos se uma amostra de dados pode ser considerada como uma distribuição normal é conveniente verificarmos se, dentre os dados da amostra, não há nenhum valor que se distancie do restante (valores anormais, espúrios, contaminantes, extremos, aberrantes). Estes valores são denominados *outliers* e podem mascarar a verdadeira distribuição dos dados.

A preocupação com a identificação e eliminação de valores *outliers* é antiga e data das primeiras tentativas de analisar um conjunto de dados. A primeira análise a ser feita, antes mesmo da identificação de um valor *outlier* é analisar o experimento, com o objetivo de prever a origem de um possível valor *outlier*, pois sua provável origem pode determinar a forma como eles devem ser tratados.

As principais causas da existência de valores *outliers* em uma amostra são erros de medição, erros de execução e a própria variabilidade inerente dos elementos da população.

O principal método gráfico para identificação de valores *outliers* em uma amostra é o boxplot (apresentado em capítulo anterior e representado na Figura 9 e Figura 10). Com o uso de boxplot, temos as seguintes regras para identificação de outliers:

1. Consideram-se valores suspeitos de serem *outliers* os valores *X's* situados na faixa dada pela equação dada a seguir. Estes valores podem ser aceitos na população após análise de sua origem.

$$x < Q1 - 1{,}5\,(Q3 - Q1)\ ou\ x > Q3 + 1{,}5\,(Q3 - Q1) \qquad \textit{Eq. 29}$$

2. Já são considerados valores extremos (*outliers*) os valores X que ultrapassam a faixa definida pela equação abaixo e que devem ser investigados e identificada a origem da dispersão. A Figura 40 ilustra o processo.

$$x < Q1 - 3\,(Q3 - Q1)\ ou\ x > Q3 + 3\,(Q3 - Q1) \qquad \textit{Eq. 30}$$

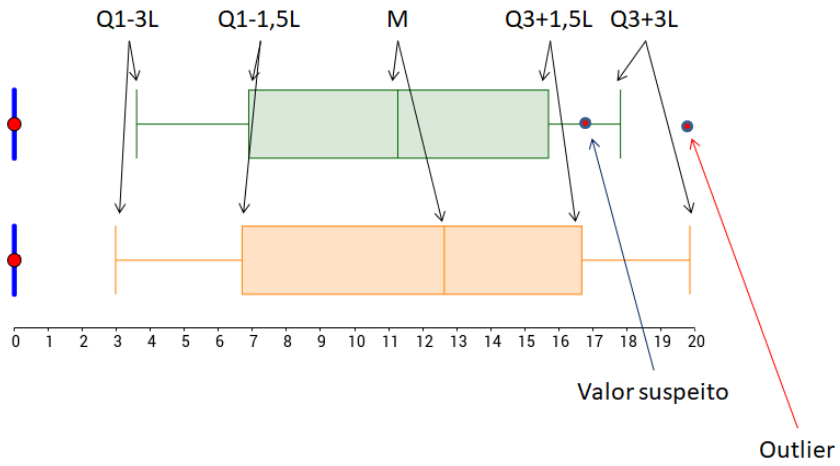

Figura 40 - Identificação de valores outliers por Boxplot

Vamos utilizar o boxplot do RStudio para verificação de *outliers* na sequência de dados apresentada Tabela 15.

N	X	Y
1	111,0	68,0
2	92,0	46,0
3	90,0	50,0
4	107,0	59,0
5	98,0	50,0
6	150,0	66,0
7	118,0	54,0
8	110,0	51,0
9	117,0	59,0
10	97,0	97,0
11	112,0	65,0

Tabela 15 - Valores X e Y para identificação de outliers

Carregando o vetor X no RStudio e criando um boxplot a partir do conjunto de dados, temos o gráfico exibido na Figura 41:

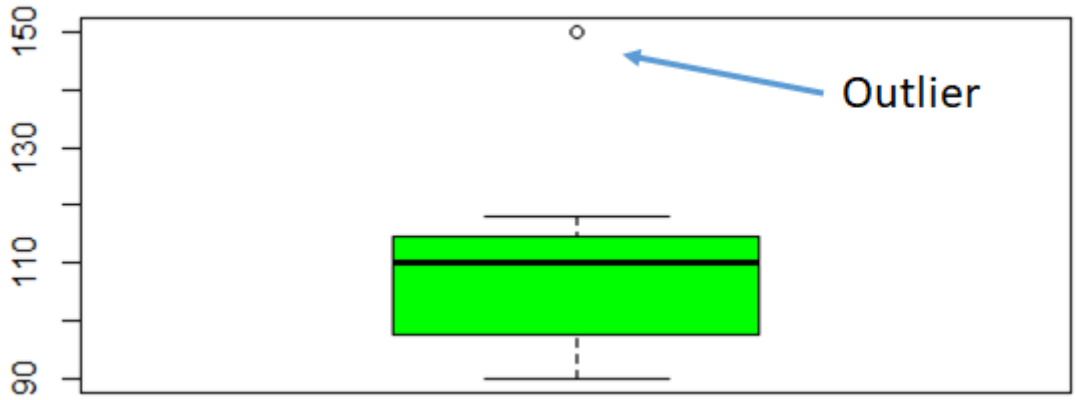

Figura 41 - Identificação de outliers pelo Boxplot

Para sabermos se o valor identificado no gráfico da Figura 41 é realmente um valor outlier, temos que, além de investigar sua origem e aplicar as regras descritas anteriormente, pois o gráfico fornecido pelo RStudio não identifica os limites de suspeição e certeza. As barras horizontais limites apresentadas no gráfico mostram os valores máximo (118) e mínimo (90), já excluindo o que ele considerou como outlier (150).

Efetuando o cálculo para os limites de suspeição, temos $x < Q1 - 1{,}5\ (Q3 - Q1)\ ou\ x > Q3 + 1{,}5\ (Q3 - Q1)$, correspondendo a x < 67 ou x > 147.

Para os limites de certeza temos $x < Q1 - 3\ (Q3 - Q1)\ ou\ x > Q3 + 3\ (Q3 - Q1)$, correspondendo a x < 37 ou x > 177.

Desta forma, o valor 150 está fora da faixa de suspeição e pode ser considerado um outlier. Valores dentro da faixa de suspeição também podem (ou devem) ser excluídos da amostra. Tudo depende da precisão desejada e da quantidade de elementos que a amostra contém.

Eliminado o valor 150 da amostra e recriando o gráfico do boxplot, podemos verificar que não foram identificados novos valores outliers, conforme mostrado na Figura 42.

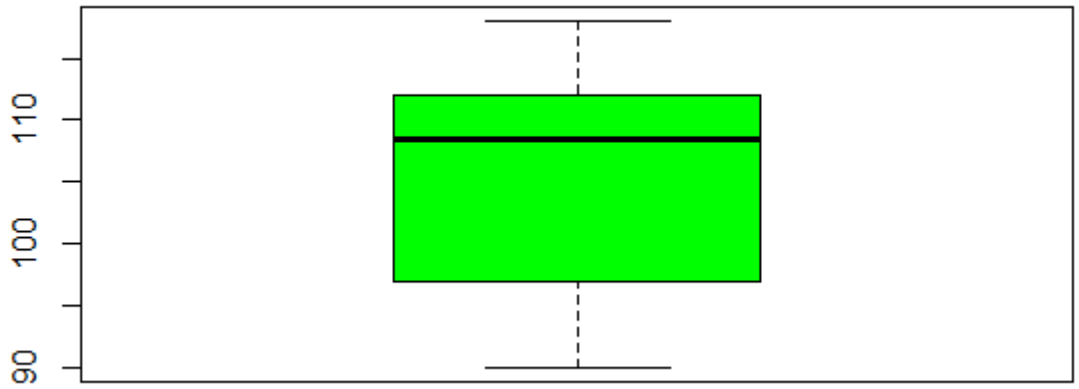

Figura 42 - Boxplot com a remoção de valores outliers

TESTE DE GRUBBS

Uma outra forma de identificar valores outliers em uma amostra é o Teste de Grubbs. É bem simples e fácil de ser executado. O Teste de Grubs é baseado na seguinte fórmula: $G = |x_i - \bar{x}|/s$, ou seja, ele é baseado na razão entre o desvio de um determinado valor da amostra da média da amostra e o desvio padrão da amostra. O valor G encontrado é comparado o valor crítico relacionado com o número de elementos da amostra fornecido pela tabela apresentada na Figura 43, onde α indica o erro aceitável ou o nível de confiança (1 – α).

Para o mesmo exemplo anterior, podemos usar o teste de Grubbs para conferir os valores outliers presentes. Para 11 elementos na amostra e com 95% de confiabilidade (1 – α), o valor crítico para G apresentado na tabela da Figura 43 é 2,23.

	Alpha				
N	**0.1**	**0.075**	**0.05**	**0.025**	**0.01**
3	1.15	1.15	1.15	1.15	1.15
4	1.42	1.44	1.46	1.48	1.49
5	1.6	1.64	1.67	1.71	1.75
6	1.73	1.77	1.82	1.89	1.94
7	1.83	1.88	1.94	2.02	2.1
8	1.91	1.96	2.03	2.13	2.22
9	1.98	2.04	2.11	2.21	2.32
10	2.03	2.1	2.18	2.29	2.41
11	2.09	2.14	2.23	2.36	2.48
12	2.13	2.2	2.29	2.41	2.55
13	2.17	2.24	2.33	2.46	2.61
14	2.21	2.28	2.37	2.51	2.66
15	2.25	2.32	2.41	2.55	2.71
16	2.28	2.35	2.44	2.59	2.75
17	2.31	2.38	2.47	2.62	2.79

Figura 43 - Valores Críticos para o Teste de Grubbs

Verificando os valores de G calculados e apresentados na Tabela 16, podemos verificar que o único valor acima do valor crítico de 2,23 é o valor de G = 2,45, correspondente ao elemento com valor 150, o que corrobora a identificação de valores outliers realizada por meio do gráfico de boxplot.

	x	G
1	90,00	1,16
2	92,00	1,04
3	97,00	0,74
4	98,00	0,68
5	107,00	0,14
6	110,00	0,04
7	111,00	0,10
8	112,00	0,16
9	117,00	0,47
10	118,00	0,53
11	150,00	2,45
Média	**109,27**	
Desvio	**16,61**	

Tabela 16 - Identificação de valores outliers pelo Teste de Grubbs

Retirando o valor 150 e recalculando os valores G para a amostra (agora com 10 elementos), descobriremos que o maior valor G encontrado para os valores da Tabela 16 é 1,49 (correspondendo ao elemento com valor 90). O valor de Grubbs crítico para amostras com 10 elementos e 95% de confiabilidade é 2,18. Assim, podemos considerar que o valor 150 é o único valor outlier presente na amostra.

Z-SCORES

O Z-score é uma variação do Teste de Grubbs. Para este teste, utilizamos os valores z-standardizados dos dados, conforme a fórmula abaixo. Da mesma forma que o teste de Grubbs, mensuramos o desvio da média em unidades do desvio padrão.

$$Z = |x - \mu|/\sigma \ ou \ z = |x - \bar{x}|/s \qquad \textit{Eq. 31}$$

1. Para amostras cujo conjunto dos dados é pequeno (inferior a 50), valores que tenham *z-scores* inferiores a -2.5 ou superiores a 2.5 devem ser considerados *outliers*.
2. Se o conjunto dos dados é grande (entre 50 e 1000), valores que tenham *z-scores* inferiores a -3.3 ou superiores a 3.3 são tipicamente considerados *outliers*.
3. Para grandes amostras (> 1000), valores com *z-scores* extremos (+/- 3,3) podem ser considerados normais.

7.4 Testes de Normalidade

Os testes de normalidade são utilizados para verificar se a distribuição de probabilidade associada a um conjunto de dados pode ser aproximada pela distribuição normal. As principais técnicas a serem discutidas são:

- Papel da probabilidade
- Teste de Kolmogorov – Smirnov
- Teste de Anderson – Darling
- Teste de Shapiro – Wilk
- Teste de Ryan-Joiner

Papel da Probabilidade

O papel da probabilidade é uma técnica gráfica utilizada para verificar a adequação de um determinado modelo estatístico aos dados. Os passos para sua construção são:

1. Considere uma amostra F(x) = $X_1, X_2, ..., X_n$;
2. Ordene, em ordem crescente, os *N* elementos da amostra;
3. Simule uma distribuição Normal de *N* elementos (*d(i)*), onde *D* = 1, 2, ..., *N*, tal que

$$d(i) = (D - 0{,}3)/(N + 0{,}4) \qquad \textit{Eq. 32}$$

A correção no numerador de -0,3 e +0,4 no denominador é necessário para que não tenhamos di = 1. Estas constantes não são padrão, dependendo do autor ou software;

4. Simule uma distribuição normal de N elementos;
5. Calcule a função *Z*, tal que

$$(Z = \left(d(i) - \overline{d(i)}\right)/s_{d(i)}) \quad \text{para } d(i) \qquad \textit{Eq. 33}$$

6. Monte o gráfico de dispersão F(x) e Z.

Exemplificando para os valores F(x) exibidos na Tabela 17, temos:

F(x)	D	d(i)	Z
1,42738	1	0,06731	-1,4863
1,52229	2	0,16346	-1,1560
1,69742	3	0,25962	-0,8257
1,90642	4	0,35577	-0,4954
1,98492	5	0,45192	-0,1651
1,99568	6	0,54808	0,1651
2,10288	7	0,64423	0,4954
2,22488	8	0,74038	0,8257
2,61826	9	0,83654	1,1560
3,15435	10	0,93269	1,4863
	Média	**0,5**	
	Des.padrão	**0,29112**	

Tabela 17 - Dados para determinação do Gráfico Papel da probabilidade

No gráfico gerado (Figura 44), podemos avaliar o quanto a distribuição de probabilidades normal ideal Z(d(i)), representada pela linha vermelha, se distancia dos valores plotados (linha azul). É uma análise visual e subjetiva, sujeita a interpretação do pesquisador e, por isto mesmo, pouco utilizada em trabalhos acadêmicos.

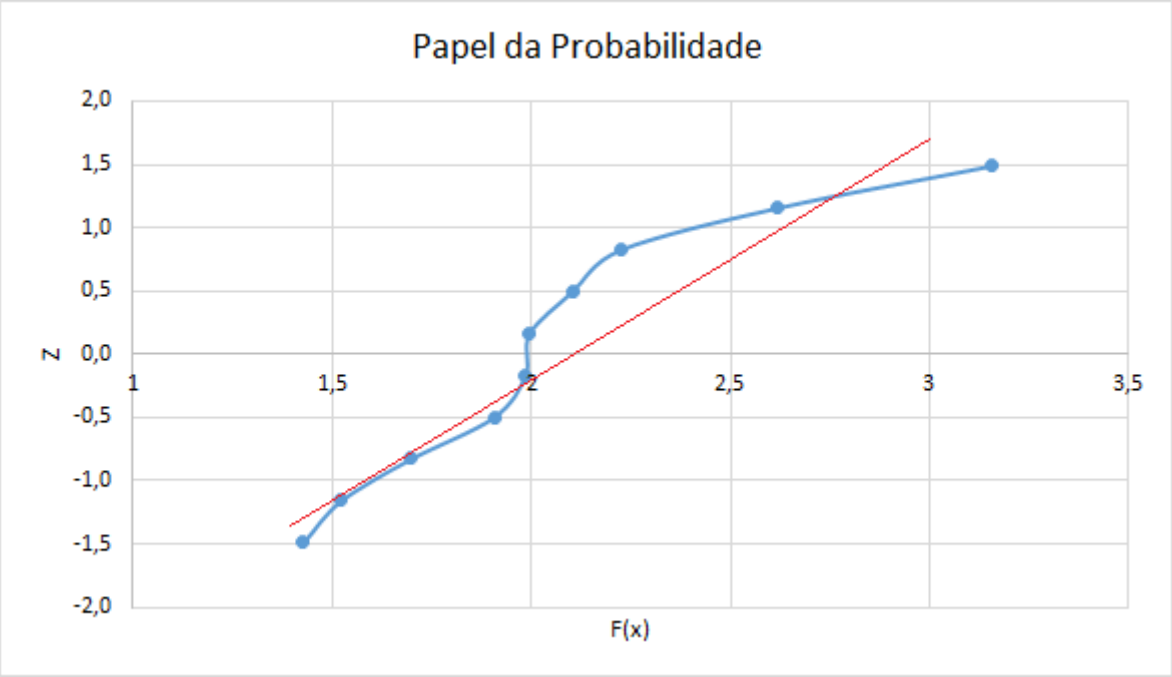

Figura 44 - Gráfico Papel da Probabilidade

Teste De Kolmogorov - Smirnov

Grande parte dos problemas que encontramos em estatística são tratados com a hipótese que os dados são retirados de uma população com uma distribuição de probabilidade específica. Por exemplo, suponha que um pequeno número de observações foi retirada de uma população com distribuição desconhecida e que estamos interessados em testar hipóteses sobre a média desta população.

O Teste de Kolmogorov – Smirnov é um teste de hipóteses e é usado para verificar se a hipótese de os dados de uma determinada amostra seguirem uma distribuição normal pode ser rejeitada ou não. Este teste observa a máxima diferença absoluta entre a função normal de distribuição acumulada para os dados e a função de distribuição empírica dos dados. Como critério, comparamos esta diferença com um valor crítico, para um dado nível de significância. A Figura 45 ilustra o funcionamento do processo.

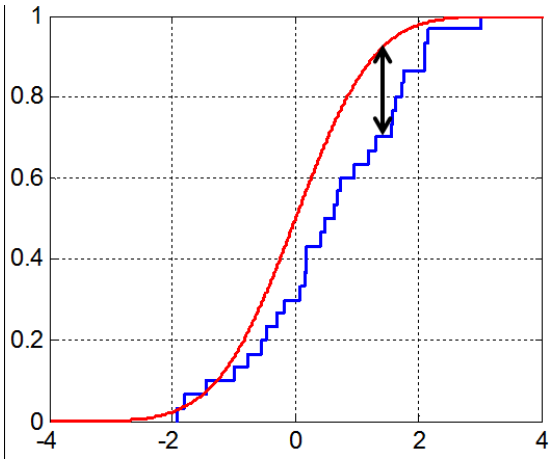

Figura 45 - Máxima distância entre a função normal e a função empírica dos dados

Para exemplificarmos o teste, considere uma amostra aleatória simples $X_1, X_2, X_3, \ldots, X_N$ de uma população com função de distribuição acumulada contínua desconhecida. A estatística utilizada para o teste é:

$$D_n = Sup_x|F(x) - Fn(x)| \qquad \textit{Eq. 34}$$

Onde: *F(X)* representa a função de distribuição acumulada assumida para os dados e *Fn(X)* representa a função de distribuição acumulada empírica dos dados.

Esta função corresponde a distância máxima vertical entre os gráficos de *F(X)* e *Fn(X)* sobre a amplitude dos possíveis valores de *x* . Como a função de distribuição empírica é descontínua e a função de distribuição hipotética é contínua, vamos considerar duas outras estatísticas:

$$D^+ = Sup_{x(i)}\left|F(x_i) - Fn(x_{(i)})\right| \qquad \textit{Eq. 35}$$

$$D^- = Sup_{x(i)}\left|F(x_i) - Fn(x_{(i-1)})\right| \qquad \textit{Eq. 36}$$

Essas estatísticas medem as distâncias (vertical) entre os gráficos das duas funções, teórica e empírica, nos pontos *x(i-1)* e *x(i)* . Com isso, podemos utilizar como estatística de teste:

$$D_n = \max(D^+, D^-)$$

Se D_n é maior que o valor crítico para a estatística do teste (Figura 46), rejeitamos a hipótese de normalidade dos dados com (1-α) 100% de confiança. Caso contrário, não rejeitamos a hipótese de normalidade.

	α				
n	0.20	0.10	0.05	0.02	0.01
1	0.900	0.95	0.975	0.990	0.995
2	0.684	0.776	0.842	0.900	0.929
3	0.565	0.636	0.708	0.785	0.829
4	0.493	0.565	0.624	0.689	0.734
5	0.447	0.509	0.563	0.627	0.669
6	0.410	0.468	0.519	0.577	0.617
7	0.381	0.436	0.483	0.538	0.576
8	0.358	0.410	0.454	0.407	0.542
9	0.339	0.387	0.430	0.480	0.513
10	0.323	0.369	0.409	0.457	0.489
11	0.308	0.352	0.391	0.437	0.468
12	0.296	0.338	0.375	0.419	0.449
13	0.285	0.325	0.361	0.404	0.432
14	0.275	0.314	0.349	0.390	0.418
15	0.266	0.304	0.338	0.377	0.404
16	0.258	0.295	0.327	0.366	0.392
17	0.250	0.286	0.318	0.355	0.381
18	0.244	0.279	0.309	0.346	0.371
19	0.237	0.271	0.301	0.337	0.361
20	0.232	0.265	0.294	0.329	0.352

	α				
n	0.20	0.10	0.05	0.02	0.01
21	0.226	0.259	0.287	0.321	0.344
22	0.221	0.253	0.281	0.314	0.337
23	0.216	0.247	0.275	0.307	0.330
24	0.212	0.242	0.269	0.301	0.323
25	0.208	0.238	0.264	0.295	0.317
26	0.204	0.233	0.259	0.290	0.311
27	0.200	0.229	0.254	0.284	0.305
28	0.197	0.225	0.250	0.279	0.300
29	0.193	0.221	0.246	0.275	0.295
30	0.190	0.218	0.242	0.270	0.290
31	0.187	0.214	0.238	0.266	0.285
32	0.184	0.211	0.234	0.262	0.181
33	0.182	0.208	0.231	0.258	0.277
34	0.179	0.205	0.227	0.254	0.273
35	0.177	0.202	0.224	0.251	0.269
36	0.174	0.199	0.221	0.247	0.265
37	0.172	0.196	0.218	0.244	0.262
38	0.170	0.194	0.215	0.241	0.258
39	0.168	0.191	0.213	0.238	0.255
40	0.165	0.189	0.210	0.235	0.252

Figura 46 - Kolmogorov-Smirnov - Valores Críticos para a estatística do teste

Exemplo 11: Para a amostra abaixo (Tabela 18), retirada de testes de resistência a compressão, verifique se a distribuição dos dados corresponde à distribuição normal com 95% de confiabilidade (a amostra não possui *outliers*):

N	1	2	3	4	5	6	7	8	9	10
Xi	38,5	37,5	37,6	37,8	39	40,1	40,8	41,5	42,3	42,5

Tabela 18 - Valores de resistência à compressão de uma amostra

A Tabela 19 apresenta os passos necessários para o teste de Kolmogorov-Smirnov, já com os valores máximos D^+ e D^- identificados.

N	**Xi**	**Z(i)**	**P(Z*i*)**	**Fn(Xi) = 1/n**	**D+ = |P(Zi) - Fn(xi)|**	**D- = |P(Zi) – Fn(X i-1)|**
1	37,5	-1,163	0,1223	0,1000	0,0223	**0,1223**
2	37,6	-1,112	0,1331	0,2000	0,0669	0,0331
3	37,8	-1,009	0,1565	0,3000	0,1435	0,0435
4	38,5	-0,649	0,2583	0,4000	0,1417	0,0417
5	39	-0,391	0,3478	0,5000	**0,1522**	0,0522
6	40,1	0,175	0,5695	0,6000	0,0305	0,0695
7	40,8	0,535	0,7038	0,7000	0,0038	0,1038
8	41,5	0,896	0,8148	0,8000	0,0148	0,1148
9	42,3	1,308	0,9045	0,9000	0,0045	0,1045
10	42,5	1,410	0,9208	1,0000	0,0792	0,0208
Média	**39,76**					
D.Padrão	**1,94**					

Tabela 19 - Exemplo do teste de Kolmogorov

A tabela da Figura 46 nos dá o valor crítico para D considerando a amostra com 10 elementos (n = 10) e 95% de confiabilidade (α = 0,05). O valor crítico para a estatística do teste é 0,409. Como a estatística do teste D_n = max (D+, D –) = 0,1522 é menor que o valor crítico, não podemos rejeitar a hipótese de normalidade dos dados com (1-α) 100% (95%) de confiança.

Exercício: Dada as amostras abaixo (Tabela 20), verificar se os dados seguem distribuição normal. Os dados não foram verificados quanto a presença de outliers.

N	1	2	3	4	5	6	7	8	9	10	11
X	111	92	90	107	98	150	118	110	117	97	112
Y	68	46	50	59	50	66	54	51	59	97	65

Tabela 20 - Dados para teste de normalidade

Os testes de normalidade são testes de hipóteses, onde a hipótese base (chamada de hipótese zero ou H0) é de que os dados da amostra seguem a distribuição normal. A hipótese contrária (H1 – os dados **não** seguem a distribuição normal) é aceita quando conseguimos rejeitar H0 e H1 é considerada como a hipótese forte (a H0 é a hipótese fraca). Quando o teste consegue a rejeição de H0, temos certeza que os dados não seguem a distribuição normal. Quando não se consegue rejeitar H0 e, consequentemente, H0 é aceita, não podemos afirmar com certeza que a distribuição é normal, simplesmente não conseguimos provar o contrário.

Costuma-se chamar o teste de hipóteses de teste de presunção de inocência. Todo réu é inocente (H0) até que se prove o contrário (H1). Se não conseguimos provar a culpa, temos que aceitar que o réu é inocente (aceitar H0). No caso contrário, quando conseguimos provar que o réu é culpado (H1) dizemos que conseguimos rejeitar H0. Assim, podemos entender que os testes de hipóteses ou conseguem rejeitar H0 (provar a culpa com certeza, por isso H1 é chamada de hipótese forte) ou são obrigados a aceitar H0 (aceitar a inocência, uma vez que não conseguiram provar a culpa e por isso, chamada de hipótese fraca).

Espero que a lógica por trás dos testes de normalidade tenha sido entendida, pois todos os demais testes seguem o mesmo princípio: a comparação com a distribuição normal. O tipo de comparação varia de um teste para outro, alterando a precisão e a confiabilidade com a qual a hipótese da normalidade dos dados (H0) é rejeitada ou aceita.

Os testes de normalidade papel da probabilidade e Kolmogorov possuem cálculo mais simplificado e foram apresentados acima para que a lógica envolvida em sua análise possa ser entendida. Os demais testes de normalidade, mais complexos e precisos, serão apresentados a partir do RStudio.

7.5 Testes De Normalidade No Rstudio

Neste tópico vamos nos centrar na execução dos testes de normalidade no RStudio e não na matemática ou estatística que compõe estes testes. Vamos compará-los quanto aos resultados e verificar quais são os mais rigorosos e os menos rigorosos.

Faremos isto a partir de exemplos, para facilitar e permitir que os testes sejam replicados como exercícios práticos. Consideremos uma amostra de 20 elementos representando a resistência a compressão de corpos de prova (Tabela 21), cujos outliers não foram identificados. A amostra possui distribuição normal?

N	1	2	3	4	5	6	7	8	9	10
Xi	105,8	110,8	72	101,3	102,4	125,8	99,7	103,5	104,6	139
N	11	12	13	14	15	16	17	18	19	20
Xi	105,2	109,3	107,3	105,9	103,2	101,2	102,1	99,8	103,2	105,7

Tabela 21 - Resistência a compressão de 20 corpos de prova

Identificação dos outliers usando boxplot

Inicialmente, vamos digitar os dados em uma planilha com formato csv (separado por vírgulas) tendo como cabeçalho as letras "res" (resistência). Todos os dados devem ser digitados na coluna A. A seguir, usando o comando "read.csv2 (file.choose(), header = TRUE)" vamos carregar a planilha no RStudio (o primeiro parâmetro indica a abertura de janela para a seleção do arquivo e o segundo, a existência de *header* – cabeçalho). Para verificar se os dados foram corretamente carregados, podemos executar o comando "summary", como mostrado a seguir. Os dados foram carregados na variável (vetor) "dados".

```
> dados = read.csv2(file.choose(), header = TRUE)
> summary(dados)
       x
 Min.   : 72.0
 1st Qu.:101.9
 Median :104.0
 Mean   :105.4
 3rd Qu.:106.2
 Max.   :139.0
```

O próximo passo é a verificação da existência de outliers. Podemos fazer isto executando o teste de Grubbs ou montando um boxplot com o vetor. Uma vez que o objetivo é usar o RStudio, vamos optar pelo boxplot (Figura 47).

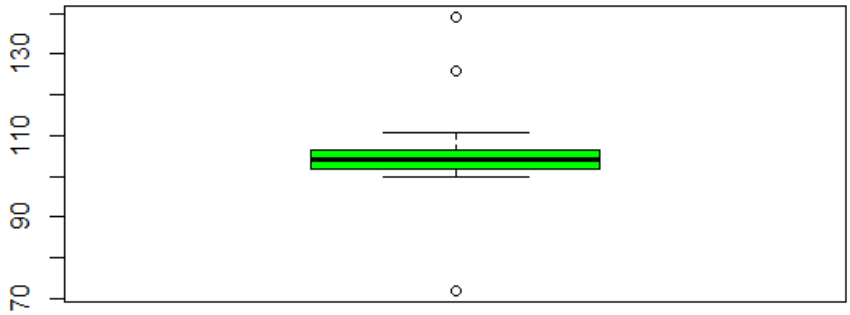

Figura 47 - Identificação de valores outliers

Na Figura 47 podemos visualizar a identificação de três valores considerados como outliers. O menor valor e dos dois maiores valores. Vamos retirá-los da amostra, recarregar a planilha csv e reexecutar o boxplot (Figura 48).

```
> dados = read.csv2(file.choose(), header = TRUE)
> summary(dados)
       x
 Min.   : 99.7
 1st Qu.:102.1
 Median :103.5
 Mean   :104.2
 3rd Qu.:105.8
 Max.   :110.8
```

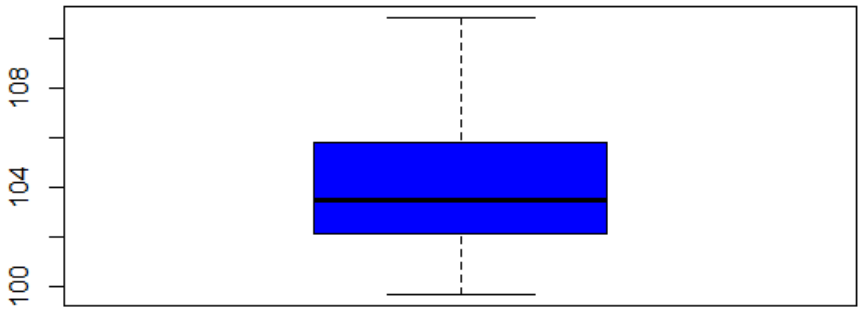

Figura 48 - Dados do vetor com a remoção dos valores outliers

Testes de Normalidade no RStudio

Uma vez removido os outliers presentes na amostra, podemos executar os testes de normalidade e comparar seus resultados. O primeiro será o teste gráfico papel da probabilidade (Figura 49):

```
> qqnorm(dados)
> qqline(dados, lty = 2, col = "red")
```

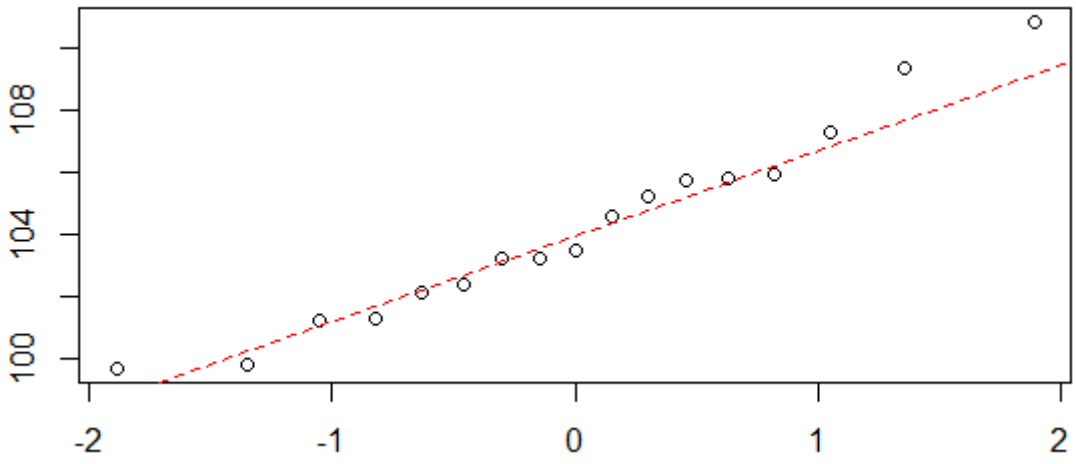

Figura 49 - Gráfico qqnorm (papel da probabilidade)

No gráfico da Figura 49, o eixo y representa os valores de resistência a compressão e o eixo x os quantis[15] teóricos da distribuição normal N(0, 1). A interpretação é subjetiva, mas como os pontos estão bem próximos da reta tracejada vermelha, podemos ter um bom indicativo da normalidade dos dados.

Os testes seguintes são testes de hipóteses. Vamos aplicar os testes de Kolmogorov-Smirnov, Lilliefors (variação do teste de Kolmogorov-Smirnov), Cramer-von Mises, Shapiro-Wilk, Shapiro-Francia e Anderson-Darling. Todos estes testes estão disponíveis no pacote "Nortest"[16]. Vamos também usar um pouco de código para agrupar os resultados.

[15] Em estatística, é comum o uso do termo quantil para referir-se a percentis. A diferença é que o quantil é expresso sob a forma decimal (quantil 0,5 = percentil 50%).
[16] Se você ainda não tem conhecimento do significado de "pacote" para o RStudio, está na hora de começar a estudar e pesquisar um pouco mais.

```
> x <- mean(dados)
> s <- sd(dados)
> cat("\n Média amostral =", x, "\n Desvio padrão amostral =", s)
```

```
Média amostral = 104.1765
Desvio padrão amostral = 3.114388
```

O primeiro teste, de Kolmogorov-Smirnov, precisa, como parâmetros, da média e desvio amostrais.

```
> t1 <- ks.test(dados, "pnorm", 104.1765, 3.114388)
> t2 <- lillie.test(dados)
> t3 <- cvm.test(dados)
> t4 <- shapiro.test(dados)
> t5 <- sf.test(dados)
> t6 <- ad.test(dados)
```

Os resultados dos testes de normalidade foram armazenados nas variáveis t_n. Podemos exibi-los digitando o nome da variável (t5, por exemplo), mas vamos continuar com o agrupamento dos resultados criando uma tabela para exibi-los.

```
> testes <- c(t1$method, t2$method, t3$method, t4$method, t5$method,
+             t6$method) # descrição do método
> estt <- as.numeric(c(t1$statistic, t2$statistic, t3$statistic,
+                      t4$statistic, t5$statistic, t6$statistic)) # estat.
> valorp <- c(t1$p.value, t2$p.value, t3$p.value, t4$p.value, t5$p.value,
+             t6$p.value) # valor p
> result_testes <- cbind(estt, valorp) # inserindo na tabela resultados
> rownames(result_testes) <- testes # nome das linhas
> colnames(result_testes) <- c("Estatística", "p") # nome das colunas
```

Agora basta digitar o nome da tabela com os resultados (result_testes) e analisar cada linha da mesma.

```
> print(result_testes, digits = 5)
                                          Estatística      p
One-sample Kolmogorov-Smirnov test            0.115392 0.97737
Lilliefors (Kolmogorov-Smirnov) normality test    0.115389 0.79205
Cramer-von Mises normality test               0.032028 0.80417
Shapiro-Wilk normality test                   0.961168 0.65354
Shapiro-Francia normality test                0.966677 0.66747
Anderson-Darling normality test               0.233047 0.76041
```

Interpretação dos resultados:

Os testes de hipóteses nos softwares estatísticos (incluindo o RStudio) são dados em termos de **p-valor**. Então, antes de analisarmos os resultados, vamos entender o significado do p-valor.

O p-valor representa a probabilidade de obter um efeito pelo menos tão extremo quanto aquele em seus dados amostrais, assumindo que a hipótese nula é verdadeira. Os p-valores abordam apenas uma questão: quão provável são seus dados, assumindo-se que a hipótese nula é verdadeira.

O p-valor, também denominado nível descritivo do teste, representa a probabilidade de que a estatística do teste (como variável aleatória) tenha valor igual ou mais extremo que aquela observada em uma amostra sob a hipótese nula, ou seja, quando a hipótese H_0 é verdadeira.

Tradicionalmente, o valor de corte para rejeitar a hipótese nula é de 0,05 (nível de significância $\alpha = 0,05$, mas pode ser alterado em qualquer dos testes, de acordo com a necessidade do pesquisador), o que significa que,

quando não há nenhuma diferença, um valor tão extremo para a estatística de teste é esperado em menos de 5% das vezes.

Um p-valor inferior ao valor pré-determinado para o nível de significância (vamos considerar 0,05, indicando um nível de confiabilidade de 95%), por exemplo, um p-valor de 0,03, conduz a rejeição da hipótese nula (H0) e consequente aceitação da hipótese alternativa (H1). Com o p-valor = 0,03, temos que há apenas uma probabilidade de 3% de se observar a condição imposta sob a hipótese nula. Como essa probabilidade inferior a probabilidade arbitrada para o teste (0,05 ou 5%), rejeitamos a hipótese nula.

Lembre-se que é uma análise estatística. Sob as condições descritas no parágrafo anterior, o p-valor de 0,03 pode ser interpretado como a possibilidade de, em cada 100 amostras iguais extraídas da população, 3 amostras confirmarão a hipótese nula e 97 não a confirmarão. Como 3 em cada 100 representa um percentual inferior ao estabelecido como nível de significância para o teste (5 em cada 100), a conclusão estatística é pela rejeição da hipótese nula.

Ainda temos que considerar que estamos trabalhando com amostras teoricamente retiradas aleatoriamente de uma população. Assim, um p-valor inferior ao nível de significância estabelecido para o teste indica o quanto os dados são improváveis assumindo-se que a hipótese nula é verdadeira. Isto conduz a duas prováveis constatações concorrentes: (1) a hipótese nula é verdadeira, mas a amostra é incomum e não representa a população; ou (2) a hipótese nula é falsa e a amostra é representativa da população.

Voltando aos resultados dos testes de normalidade, o maior p-valor encontrado foi para o teste de Kolmogorov-Smirnov (p-valor = 0,9774) e o menor p-valor foi para o teste de Shapiro-Wilk (p-valor = 0,6535). Todos os p-valor encontrados são superiores a 0,05, indicando que a hipótese nula (H0 – normalidade dos dados da amostra) não pode ser rejeitada. Podemos também entender (e a literatura corrobora) que os testes mais rigorosos são os de Shapiro-Wilk e Shapiro-Francia e o menos rigoroso o de Kolmogorov-Smirnov.

Para um detalhamento maior, vamos executar alguns desses testes de forma isolada. O teste de Kolmogorov-Smirnov e o teste de Shapiro-Wilk.

```
> ks.test(dados, "pnorm", 104.2, 3.1144)

   One-sample Kolmogorov-Smirnov test

data:  dados
D = 0.11833, p-value = 0.9712
```

```
> shapiro.test(dados$res)

   Shapiro-Wilk normality test

data:  dados$res
W = 0.96117, p-value = 0.6535
```

Executando o teste de Kolmogorov-Smirnov para o exemplo da Tabela 18, temos:

```
> dados = read.csv2(file.choose(), header = T)
> summary(dados)
      res
 Min.   :37.50
 1st Qu.:37.98
 Median :39.55
 Mean   :39.76
 3rd Qu.:41.33
```

```
 Max.   :42.50
> x = mean(dados$res)
> sx = sd(dados$res)
> ks.test(dados$res, "pnorm", x, sx)

   One-sample Kolmogorov-Smirnov test

data:  dados$res
D = 0.15218, p-value = 0.9485
alternative hypothesis: two-sided
```

O valor da estatística D (0,15218) resultante do teste executado no RStudio é o mesmo encontrado quando executamos os cálculos no MS Excel (0,1522). O RStudio não nos mostra o valor crítico (Figura 46), mas o p-valor nos dá a confiabilidade (ou amplitude) com a qual podemos aceitar a hipótese nula (normalidade dos dados).

7.6 Intervalo De Confiança

Um intervalo de confiança (IC) é um intervalo estimado de um parâmetro de interesse de uma população (a média, por exemplo). Em vez de estimar o parâmetro por um único valor, é dado um intervalo de estimativas prováveis, centralizado no valor do parâmetro de interesse ($\bar{X} \mp \Delta x$), por exemplo.

Intervalos de confiança são usados para indicar a confiabilidade de uma estimativa em relação ao valor de um parâmetro de interesse. Por exemplo, em dois experimentos, ao compararmos os intervalos de confiança, calculados com o mesmo nível de significância α, para a média da resistência a compressão obtidos, sendo o primeiro 100 ± 15 MPa e o segundo 100 ± 7 MPa, podemos concluir que o segundo experimento ofereceu resultados mais confiáveis, com menor variação. Isto significa que, sendo todas as estimativas iguais, pesquisas que resultem num IC menor é mais confiável do que uma que resulte num IC maior.

Um dos principais parâmetros associados ao intervalo de confiança é o coeficiente de confiança ou nível de confiança ou simplesmente confiança $(1 - \alpha)$. É o valor complementar do erro esperado: se temos 5% de chances de errar uma estimativa ($\alpha = 0{,}05$), temos, consequentemente, 95% de confiança em acertar a mesma estimativa (1 - α).

Outra forma de entendermos o coeficiente de confiança é a repetição do experimento. Com o nível de confiança $(1 - \alpha)$ podemos afirmar que, se repetirmos muitas vezes o experimento, aproximadamente em $100\ (1 - \alpha)$ das vezes a média populacional estará no intervalo encontrado.

Uma das principais interpretações do intervalo de confiança consiste em avaliar a incerteza que temos a respeito de estimarmos um determinado parâmetro populacional a partir de uma amostra aleatória de tamanho *n*.

Intervalo De Confiança Para A Média

Quando queremos estimar (inferir) a média de uma população por meio da análise dos valores de uma amostra, ou seja, queremos inferir valores para a população a partir dos dados da amostra, temos dois casos distintos a considerar: quando a variância da população é conhecida e quando ela é desconhecida.

Podemos considerar que no primeiro caso, variância da população conhecida, temos algumas informações sobre a população e podemos adotar métodos que considerem que a amostra é próxima da população. No segundo caso, variância desconhecida, não sabemos nada sobre a população que originou a amostra. Neste caso, métodos cujo resultado seja mais "abrangente" são os indicados.

Variância Conhecida: Consideremos uma amostra aleatória simples *X1 .... Xn*, obtida de uma população com distribuição normal, com média μ e variância σ conhecidas. A variável *Z*, nestas situações, é dada pela equação:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \qquad \textit{Eq. 37}$$

Consideremos que a probabilidade da variável Z (Figura 50) tomar valores entre $-Z_{\alpha/2}$ e $Z_{\alpha/2}$ é de (1 – α). Então, de acordo com a curva da distribuição normal padrão, temos que $P\left[-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\right] = (1 - \alpha)$.

Substituindo *Z* na equação de probabilidade acima, temos:

$$P\left[-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\alpha/2}\right] = (1 - \alpha) \qquad \textit{Eq. 38}$$

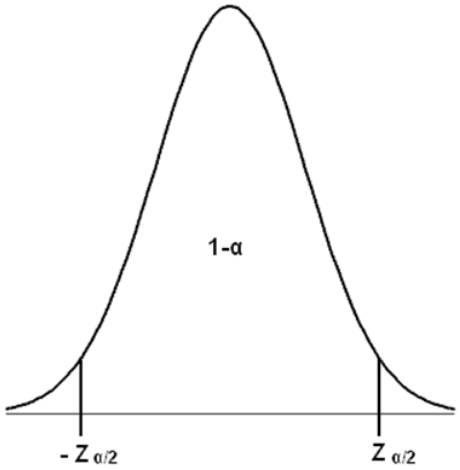

Figura 50 - Intervalo de confiança para a média - Variância Conhecida

Isolando a média populacional μ, a equação passa a ser:

$$P\left[\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = (1 - \alpha) \qquad \textit{Eq. 39}$$

A equação acima corresponde ao Intervalo de Confiança para a média com um nível de confiabilidade (1 – α) e pode ser reescrita como:

$$IC(\mu, 1 - \alpha) = \left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \; ; \; \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \qquad \textit{Eq. 40}$$

Como citado anteriormente, o intervalo de confiança significa que, repetindo o experimento muitas vezes, em aproximadamente 100 (1 – α)% das vezes a média populacional estará no intervalo encontrado.

Exemplo 12: Em um experimento para testes de diferentes compostos para produção de concreto de alta resistência com variância conhecida (para o exemplo, considerar o desvio padrão populacional σ igual ao desvio padrão amostral *s*), foram testadas quatro composições A, B, C e D diferentes, com 4, 6, 8 e 10 elementos por amostra, respectivamente. A partir dos resultados de resistência a flexão (Tabela 22) observados para os elementos da amostra, determine o intervalo de confiança com 95% de confiabilidade.

Para a solução do problema apresentado, vamos inicialmente carregar os dados no RStudio a partir de uma planilha .csv. Com a planilha carregada podemos verificar a existência de valores outliers (se existirem, devem ser excluídos da amostra) e executar o teste de normalidade Shapiro-Wilk que nos parece ser o mais crítico dos testes estudado.

N	A	B	C	D
1	63,73	71,01	96,45	95,24
2	72,15	65,38	82,52	95,13
3	58,22	81,93	92,62	85,44
4	58,03	72,97	90,82	86,13
5		58,68	94,36	79,5
6		52,53	81,68	86,55
7			81,49	84,44
8			93,67	108,37
9				94,39
10				94,19
Média	63,03	67,08	89,20	90,94
Desv.P	6,63	10,54	6,26	8,22
CV	0,11	0,16	0,07	0,09

Tabela 22 - Dados de resistência a flexão das amostras

Carga dos dados no RStudio e execução do comando para geração do gráfico de boxplot (com cores diferentes para cada tratamento) mostrado na Figura 51.

```
> library(nortest)
> dados = read.csv2(file.choose(), header=T)
> dados
       a     b     c      d
1  63.73 71.01 96.45  95.24
2  72.15 65.38 82.52  95.13
3  58.22 81.93 92.62  85.44
4  58.03 72.97 90.82  86.13
5     NA 58.68 94.36  79.50
6     NA 52.53 81.68  86.55
7     NA    NA 81.49  84.44
8     NA    NA 93.67 108.37
9     NA    NA    NA  94.39
10    NA    NA    NA  94.19

> boxplot(dados, col=c("red","blue","yellow","gray"))
```

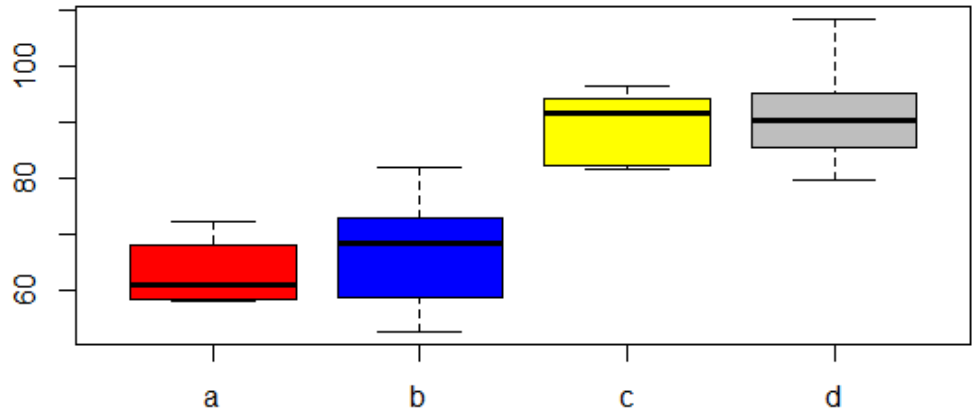

Figura 51 - Boxplot dos dados das amostras A, B, C e D

O gráfico de boxplot gerado pelo RStudio não indica a presença de valores outliers. Assim, prosseguimos com o teste de normalidade para os elementos das amostras.

```
> shapiro.test(dados$a)
    Shapiro-Wilk normality test
data:  dados$a
W = 0.85583, p-value = 0.2456

> shapiro.test(dados$b)
    Shapiro-Wilk normality test
data:  dados$b
W = 0.98613, p-value = 0.9777

> shapiro.test(dados$c)
    Shapiro-Wilk normality test
data:  dados$c
W = 0.83366, p-value = 0.06479

> shapiro.test(dados$d)
    Shapiro-Wilk normality test
data:  dados$d
W = 0.90921, p-value = 0.2756
```

Todos os p-valores são superiores a 0,05, de onde não podemos rejeitar a hipótese de que os dados das amostras seguem a distribuição normal. Assim, passamos ao cálculo dos intervalos de confiança para a média para cada uma das amostras. Vamos calcular pela fórmula dada anteriormente e pelo RStudio.

Como o valor do nível de confiança foi definido como 95% (1 – α), isto implica que α é igual a 0,05 e α/2 = 0,025. Com o uso da tabela da distribuição normal padronizada, obtemos que $Z_{0,025}$ = 1,96 e com a aplicação da fórmula a seguir podemos calcular o intervalo de confiança (*Eq. 40*) para todas as amostras, conforme mostrado na Tabela 23.

$$IC(\mu, 1-\alpha) = \left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \; ; \; \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

	A	B	C	D
n	4	6	8	10
Média	63,03	67,08	89,20	90,94
Desv.P	6,63	10,54	6,26	8,22
CV	0,11	0,16	0,07	0,09

$Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	6,50	8,44	4,34	5,09

IC -	56,54	58,65	84,86	85,84
IC +	69,53	75,52	93,54	96,03

Tabela 23 - Intervalo de confiança para as amostras

Colocando no formato padrão de intervalo de confiança temos:

Amostra A $IC(\mu, 0{,}95) = (56{,}54; 69{,}53)$

Amostra B $IC(\mu, 0{,}95) = (58{,}65; 75{,}52)$

Amostra C $IC(\mu, 0{,}95) = (84{,}86; 93{,}54)$

Amostra D $IC(\mu, 0{,}95) = (85{,}84; 96{,}03)$

Neste exemplo trabalhamos com diferentes coeficientes de variação (razão entre o desvio padrão e a média amostral). Se todas as amostras tivessem o mesmo coeficiente de variação, poderíamos notar, mais explicitamente que, à medida que o número de elementos na amostra aumenta, a relação entre a amplitude do intervalo de confiança e a média diminui, pois, com o aumento do tamanho da amostra, conseguimos representar melhor a população e assim, obter estimativas mais precisas.

Da mesma forma, se considerarmos amostras com a mesma quantidade de elementos, quanto maior for o desvio padrão, maior será a relação entre a amplitude do intervalo de confiança e a média, pois, maior variabilidade nos elementos da amostra implica em menor precisão nas estimativas para a população.

Variância Desconhecida: Quando não temos informações sobre a população, somente os dados da amostra para a análise, a diferença é que usamos a distribuição t-Student ao invés da distribuição normal padrão.

Consideremos, por exemplo, uma amostra aleatória simples *X1 .... Xn*, obtida de uma população com distribuição normal, com média e variância desconhecidas. Como neste caso a variância é desconhecida, utilizaremos a variância amostral S^2 no lugar de δ^2 . Assim, temos que a fórmula apresentada anteriormente para *Z* passa a ser:

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t(n-1) \qquad \textit{Eq. 41}$$

Ou seja, a variável obedece a **distribuição t de Student** com (n-1) graus de liberdade. Então, ao fixarmos o nível de significância α, obtemos da Tabela da distribuição t de Student com (n-1) graus de liberdade, o valor $t_{((n-1),\,\alpha/2)}$, que satisfaz a probabilidade P, tal que:

$$P\left[-T_{(n-1,\frac{\alpha}{2})} \leq T \leq T_{(n-1,\ \frac{\alpha}{2})}\right] = (1 - \alpha) \qquad \textit{Eq. 42}$$

Repetindo o mesmo raciocínio empregado anteriormente, na dedução do intervalo de confiança para a média, temos:

$$IC(\mu, 1 - \alpha) = \ (\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}\ ;\ \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}) \qquad \textit{Eq. 43}$$

Reforçando novamente, o intervalo de confiança significa que, repetindo o experimento muitas vezes, em aproximadamente 100 (1 – α)% das vezes a média populacional estará no intervalo encontrado. Só que, desta

vez, como usamos a distribuição t-Student como base para a inferência da média populacional, o intervalo de confiança terá maior amplitude.

Para visualizarmos isto, vamos repetir o exemplo anterior, com o pressuposto que não possuímos informações sobre a população (variância desconhecida). Os dados são apresentados na Tabela 24.

Se compararmos com os dados apresentados na Tabela 23 (cálculo do IC usando a distribuição normal, considerando a variância populacional conhecida), podemos facilmente identificar que a amplitude do intervalo da IC para a média aumentou. Uma vez que não temos informações sobre a população e vamos inferir usando apenas os dados das amostras, as inferências são mais conservadoras.

As observações feitas anteriormente permanecem. Quanto maior o número de elementos da amostra, menor a amplitude do IC e, quanto maior o desvio padrão, maior a amplitude do IC, considerando-se o outro fator constante.

	A	B	C	D
n	4	6	8	10
Média	63,03	67,08	89,20	90,94
Desv.P	6,63	10,54	6,26	8,22
CV	0,11	0,16	0,07	0,09
t(n-1, α/2)	3,182	2,571	2,365	2,262
$t_{\alpha/2}\frac{s}{\sqrt{n}}$	10,55	11,07	5,23	5,88
IC -	52,49	56,02	83,97	85,07
IC +	73,59	78,16	94,44	96,82

Tabela 24 - Cálculo do IC usando a distribuição t-Student

7.7 Testes de Hipóteses – Comparação de Médias

Neste item vamos apresentar as ideias fundamentais sobre testes de hipóteses. Podemos considerar que um dos principais objetivos de um experimento é confirmar uma determinada afirmação sobre uma população, ou, mais especificamente, sobre um parâmetro dessa população. Assim, torna-se também objetivo do experimento comprovar se os resultados experimentais provenientes de uma amostra contrariam, ou não, tal afirmação.

Esta é a função do teste de hipóteses. Vamos supor que um pesquisador deseja saber se a inclusão de um determinado elemento na produção do concreto permite melhorar suas propriedades, como resistência mecânica, porosidade, dentre outras. Podemos entender que esta pesquisa levanta hipóteses sobre as propriedades (por exemplo, a média μ da resistência à compressão, tração por compressão diametral, absorção por imersão e permeabilidade) do material a ser produzido (população).

O pesquisador poderia fazer suposições ou afirmativas sobre a variável aleatória que representa as propriedades de interesse do material produzido, qual o percentual de incremento ou decremento em cada uma das propriedades, por exemplo. Estas afirmações ou suposições são chamadas hipóteses estatísticas. Assim, podemos dizer que hipótese estatística é uma conjectura sobre um parâmetro ou propriedade a ser comprovada (ou rejeitada) por meio da análise dos resultados de experimentos.

A hipótese base para o teste de hipóteses é chamada de **hipótese nula** (*H0*) e ela é usualmente caracterizada pela igualdade[17]. No exemplo anterior a hipótese nula seria a de que as médias das propriedades de interesse são as mesmas com ou sem a adição do novo elemento, ou seja, a adição do novo elemento não produz melhorias significativas nas propriedades do concreto produzido.

A hipótese contrária, que usamos como alternativa à hipótese nula, isto é, a hipótese que será aceita quando a hipótese nula é rejeitada é denominada **hipótese alternativa** (*H1*), também chamada de hipótese do pesquisador. Para o exemplo, como o pesquisador está interessado em comprovar melhorias nas propriedades, a hipótese alternativa seria que as médias das propriedades de interesse são maiores para o concreto com a adição do elemento do que sem a adição. Assim poderíamos ter:

$$H0: \mu_0 P_X = \mu P_X$$
$$H1: \mu_0 P_X > \mu P_X$$

Onde μ_0 é a média da propriedade para o concreto de referência (produzido sem a adição do elemento) e μ a média com a adição do elemento para a propriedade (ou parâmetro) P_x.

Outros tipos de formulação de hipóteses também são comuns, tais como:

$H0: \mu_0 = \mu$ $H1: \mu_0 \neq \mu$	$H0: \mu_0 = \mu$ $H1: \mu_0 > \mu$	$H0: \mu_0 = \mu$ $H1: \mu_0 < \mu$	$H0: \mu_0 = \mu$ $H1: \mu_0 > \mu + x$

Os testes de hipóteses podem ser bilaterais, quando desejamos saber se a média é diferente (neste caso, se a média for maior ou menor, não importa, pois ela é diferente) ou unilaterais, quando a hipótese H1 é construída com a suposição de aumento (maior) ou diminuição (menor) da média.

Para os testes bilaterais, o nível de confiança estipulado para o teste (normalmente α = 0,05) deve ser dividido entre as caudas, pois queremos ter 95% de certeza de que a média é diferente, como mostrado na Figura 52. As hipóteses para o teste são:

$$H0: \mu_0 = \mu$$
$$H1: \mu_0 \neq \mu$$

Independente de usarmos a distribuição normal padrão ou a distribuição t-Student, o valor de α estipulado para o teste será dividido entre as duas caudas (α/2). A região central da curva indica a área de aceitação de *H0* e as caudas a área de rejeição de *H0* e consequente aceitação de *H1* (Figura 52).

[17] Ver comentário sobre testes de hipóteses na Página 65

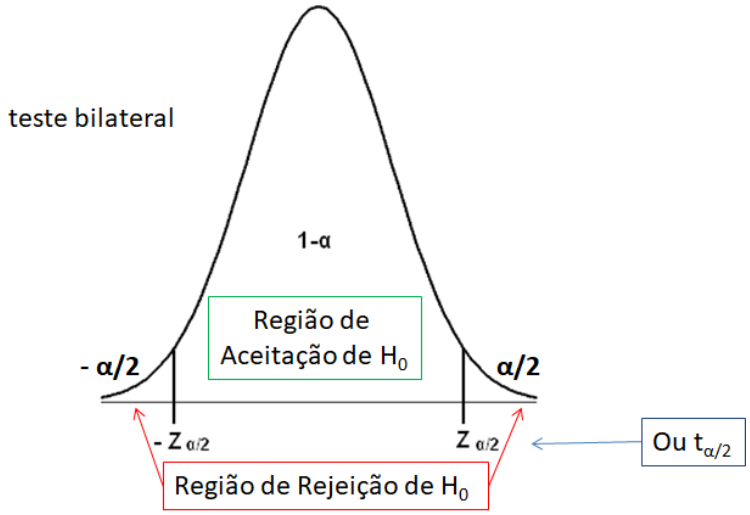

Figura 52 - Teste bilateral - Regiões de rejeição

Nos testes unilaterais (Figura 53), onde as hipóteses são formuladas com a suposição de que a média é maior ou menor, as hipóteses formuladas podem ser:

1	$H0: \mu_0 = \mu$ $H1: \mu_0 < \mu$	2	$H0: \mu_0 = \mu$ $H1: \mu_0 > \mu$

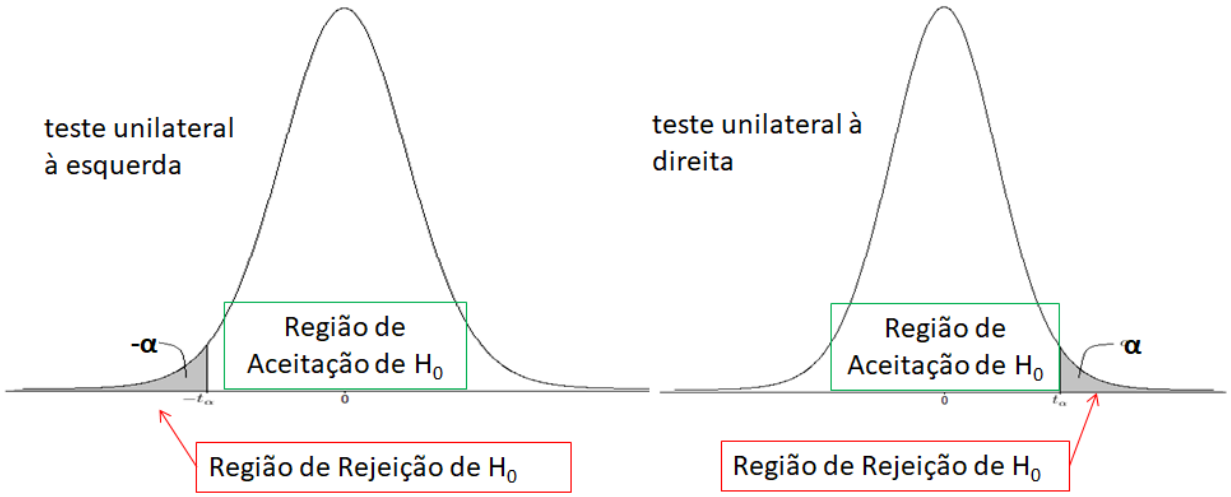

Figura 53 - Testes unilaterais - regiões de rejeição

No teste unilateral com as hipóteses estipuladas acima (menor ou maior que), também independente da distribuição que usarmos, o valor α estipulado para o teste será alocado na região correspondente à região de rejeição de *H0*. Afinal, agora nos interessa saber, com 100% (1 – α) de confiabilidade, se rejeitamos ou não a hipótese nula.

Análise dos dados a serem comparados

Antes de iniciarmos os testes de hipóteses, temos que analisar os dados e verificar o conhecimento que possuímos sobre os mesmos. Quantas amostras desejo comparar? As amostras são independentes?

Possuímos alguma informação sobre a população? Quantos elementos possui cada amostra? Todas estas são questões que irão direcionar o cálculo da estatística que será utilizada para a comparação com a probabilidade extraída da distribuição normal padronizada ou da distribuição de t-Student. Podemos dizer que o procedimento estatístico a ser usado na análise dos dados é dependente das questões formuladas acima.

Amostras Independentes: Quando os elementos da amostra são distintos e independentes ou quando não há informações suficientes para determinar similaridades entre os elementos.

Amostra com Dados Pareados: Quando os elementos da amostra são analisados em situações diferentes (antes e depois), ou seja, cada elemento está associado a um par de medidas: uma antes de um determinado tratamento e outra depois deste tratamento. Outra situação ocorre quando podemos formar pares de elementos tão similares quanto possível e garantindo que os elementos do par sejam direcionados a amostras diferentes. Assim poderemos aplicar tratamentos diferentes em cada elemento do par.

Os procedimentos estatísticos para dados pareados somente devem ser utilizados quando se tem segurança de que, no período entre as mensurações, o único valor que afeta os dados é o fator em estudo (tratamento). Caso contrário, é mais recomendado um delineamento como amostras independentes.

Estatística a ser usada

Da mesma forma que foi utilizado na determinação do Intervalo de Confiança, temos que identificar se temos ou não informações sobre a população para a qual queremos inferir. Novamente temos duas situações: variância conhecida (temos informações sobre a população) e variância desconhecida (não temos informações sobre a população).

Se conhecemos a média e a variância populacional, usamos a distribuição normal padrão (Z). Se os dados populacionais são desconhecidos (situação que irá abranger a maioria dos experimentos inovadores de engenharia), usamos a distribuição t-Student.

Outro fator que influi na escolha de qual distribuição utilizar para o teste de hipóteses é a quantidade de elementos que a amostra contém. Se a amostra contiver mais de 30 elementos (30 mensurações, descontando-se os valores identificados como outliers), podemos usar a distribuição normal padronizada. Caso contrário, para amostras com 30 ou menos elementos, devemos usar a distribuição t-Student.

A Tabela 25 apresenta o resumo dos conceitos que definem a estatística do teste a ser usada.

Condição	Estatística
Variância conhecida **OU** n > 30	$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
Variância desconhecida **E** n ≤ 30	$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$

Tabela 25 - Escolha da estatística a ser usada

Testes de Comparação de Médias

A seguir apresentamos os testes de comparação de médias, construídos com base nos conceitos citados anteriormente.

Uma amostra: quando citamos a comparação dos dados de uma amostra, na realidade nos referimos a uma única variável aleatória associada a esta amostra. A comparação é feita para a variável aleatória que representa a propriedade ou parâmetro de interesse.

O critério de avaliação é:

- Teste bilateral: se Tobs > Tα/2 ou se Tobs < -T-α/2 , rejeitamos ***H0*** . Caso contrário, não rejeitamos ***H0***.
- Teste unilateral à direita: se Tobs > Tα, rejeitamos ***H0***. Caso contrário, não rejeitamos ***H0***.
- Teste unilateral à esquerda: se Tobs < -T-α, rejeitamos ***H0***. Caso contrário, não rejeitamos ***H0***.

OBS. Se a variância for conhecida ou a amostra possuir mais de 30 elementos, usamos a distribuição normal padronizada e a estatística para o teste é dada por ***Z***

Exemplo 13: De um lote de 1000 dormentes de concreto foram selecionados aleatoriamente 35 dormentes para testes de resistência a flexão. É exigido que a resistência a flexão seja igual 54 MPa. A média e o desvio padrão amostrais foram de 56,81 e 7,4 MPa, respectivamente. O lote atende as especificações, com 95% de confiabilidade?

Como o objetivo é determinar se a média populacional é igual a 54 MPa, com confiabilidade de 95% (α = 0,05), o teste de hipótese é:

$$\begin{cases} H0: \mu_0 = \mu \\ H1: \mu_0 \neq \mu \end{cases}$$

O valor base para a estatística é α/2 = 0,025 e como a amostra é formada por 35 elementos, vamos usar a distribuição normal padrão e a estatística Z.

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \ \therefore Z = \frac{56{,}81 - 54}{\frac{7{,}4}{\sqrt{35}}} = 2{,}25$$

Assim temos que $Z_{obs} = 2{,}25$. O valor de $Z_{\alpha/2}$ na tabela da distribuição normal padrão é 1,96. Então temos que $|Z_{obs}|$ = 2,25 e $|Z_{\alpha/2}|$= 1,96 o que nos mostra que Z_{obs} está na Zona de rejeição de ***H0***, conforme pode ser observado na Figura 54.

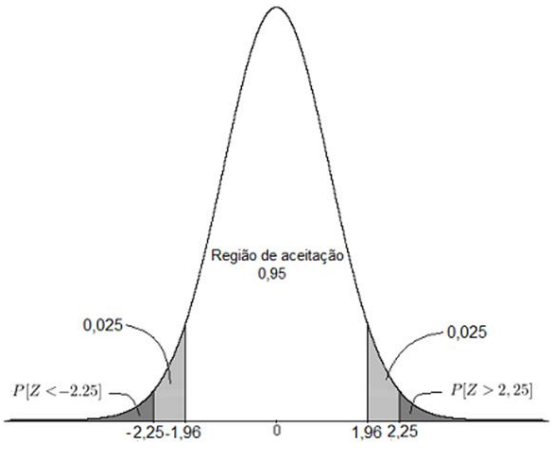

Figura 54 - Teste bilateral - Z observado

Também podemos calcular a probabilidade de a média amostral ser igual ao valor proposto para o teste (54 MPa). Como temos um teste bilateral e o valor de α foi dividido entre as duas caudas, temos que o p-valor é dado por:

$$P[Z > |Z_{obs}|] + P[Z < -|Z_{obs}|] = P[Z > 2{,}25] + P[Z < -2{,}25] = 0{,}0122 + 0{,}0122 = 0{,}0244$$

Assim, temos a probabilidade de 2,44% para a hipótese ***H0***. Como a confiabilidade foi estabelecida em 95%, implicando em α = 0,05, a probabilidade para ***H0*** é inferior a estabelecida, levando a rejeição de ***H0***.

A mesma comparação pode ser realizada em termos do p-valor. O p-valor resultante do teste é 0,0244 e o p-valor estabelecido para o teste é 0,05. Como o p-valor resultante é inferior ao estabelecido, rejeita-se ***H0***.

Exemplo 14: Os dados abaixo representam a resistência a ruptura por tração de 10 amostras de um cabo de aço. Com base nos resultados, deseja-se saber se esse cabo obedece a especificação: carga média de ruptura superior a 1500 kgf, com 95% de confiabilidade (não foram identificados valores outliers na amostra).

Valores ensaios: 1508 / 1518 / 1492 / 1505 / 1515 / 1507 / 1510 / 1505 / 1496 / 1498

Desta vez o objetivo é determinar se a média populacional é superior a 1500 Kgf, com confiabilidade de 95% (α = 0,05), o teste de hipótese é:

$$\begin{aligned} &H0: \mu_0 = \mu \\ &H1: \mu_0 > \mu \end{aligned}$$

E usaremos o teste unilateral à direita, onde se *Tobs > Tα,* rejeitamos ***H0***. Caso contrário, não rejeitamos ***H0.***

O valor base para a estatística é α = 0,05 e como a amostra é formada por 10 elementos, vamos usar a distribuição t-Student.

A partir dos dados da amostra, os seguintes valores foram calculados:

- X = 1.505,4 kgf
- S = 8,1948 kgf
- N = 10
- α = 0,05

O valor da estatística do teste é:

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t(n-1) \ \therefore T = \frac{1505{,}4 - 1500}{\frac{8{,}2}{\sqrt{10}}} = 2{,}0825 \ \therefore T_{obs} = 2{,}0838$$

A tabela da distribuição t-Studend (Figura 38) nos fornece o valor da estatística para α = 0,05 e GL = n – 1 = 10 – 1 = 9. Assim, temos que $T_{(\alpha,9)} = 1{,}833$.

A Figura 55 nos permite visualizar as estatísticas do teste. Como $T_{obs} > T_{(\alpha,9)}$ (ou seja, *2,0838 > 1,833*) a hipótese ***H0*** pode ser rejeitada. Isto indica que a resistência média de ruptura é superior a 1500 kgf, com 95% de confiabilidade.

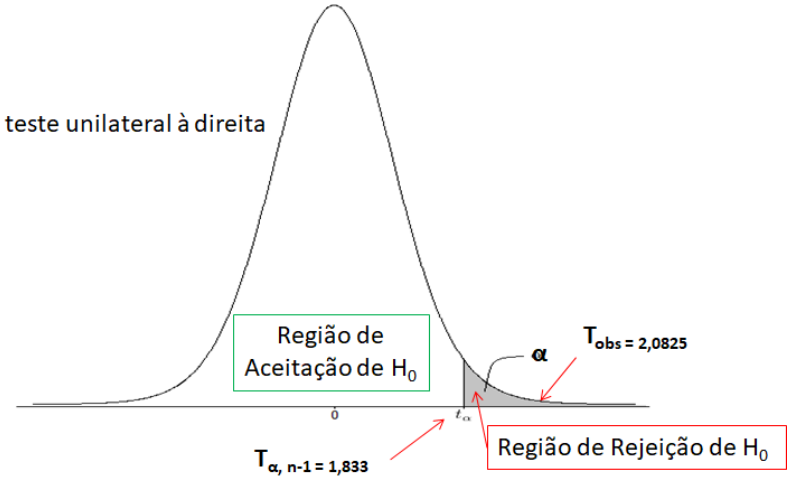

Figura 55 - Teste unilateral a direita

Testes de Comparação de Médias com Duas Amostras

Quando comparamos duas amostras (Figura 56), passamos a ter quatro tipos de situações que devem ser consideradas: (i) a variância populacional de ambas as amostras é conhecida; (ii) a variância populacional das amostras é igual mas desconhecida; (iii) as variâncias populacionais são desconhecidas; e (iv) os dados são pareados. Para cada tipo há uma fórmula diferente para a estatística.

Figura 56 - Teste de comparação de médias com duas amostras

Variâncias conhecidas

Suponha que queremos comparar a diferença nas médias μ_1 e μ_2 ($\mu_1 - \mu_2 = \Delta_0$) de duas populações normais e independentes, sendo suas variâncias conhecidas. A Estatística do teste é dada por:

$$Z_{obs} = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad ou \quad \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \textit{Eq. 44}$$

A hipótese nula é dada por: **$H_0 : \mu_1 - \mu_2 = \Delta_0$.** O teste pode ser resumido pela Tabela 26.

$H_0: \mu_1 - \mu_2 = \Delta_0$		
Hipóteses Alternativas	**Valor P**	**Critérios de rejeição de H_0**
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probabilidade acima de $\|Z_0\|$ e abaixo de $-\|Z_0\|$ $P = 2[1 - \phi(\|Z_0\|)]$	$Z_0 > Z_{\alpha/2}$ ou $Z_0 < -Z_{\alpha/2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probabilidade acima de Z_0 $P = 1 - \phi(Z_0)$	$Z_0 > Z_{\alpha}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probabilidade abaixo de Z_0 $P = \phi(Z_0)$	$Z_0 < -Z_{\alpha}$

Tabela 26 - Comparação de duas médias com variância conhecida

Exemplo 15: Uma empresa está interessada em desenvolver produtos para aceleração da cura do concreto. Uma nova formulação é proposta e um experimento de comparação com a formulação antiga é preparado com duas amostras: a primeira usa a composição padrão e a segunda tem novo ingrediente para aceleração da cura. Espera-se que a adição do novo ingrediente não altere a variância da resistência a compressão (2,7MPa). Dez amostras com a formulação 1 foram testadas com 168 horas e tiveram uma resistência média a compressão de 15,5 MPa. Outras 15 amostras com a formulação 2 foram testadas também com 168 horas e tiveram uma resistência a compressão de 17,2 MPa. Sabendo-se que as condições de preparação e teste foram homogêneas, podemos afirmar com 95% de confiabilidade que a adição do novo ingrediente foi benéfica para a cura do concreto (aumento da resistência a compressão)?

Neste experimento o objetivo é determinar se a nova formulação (adição do novo ingrediente) melhora o tempo de cura do concreto. Temos duas amostras (variância populacional conhecida e igual a 2,7 e suposta distribuição normal) e a confiabilidade exigida é de 95% (α = 0,05). Então temos:

- $\sigma_1 = \sigma_2 = \sigma = 2{,}7$ Mpa
- n1 = 10
- n2 = 15
- $\overline{X1} = 15{,}5$
- $\overline{X2} = 17{,}2$
- α = 0,05

O teste de hipótese proposto para o problema é o **teste unilateral à direita**, com as seguintes hipóteses:

$H0: \mu_2 - \mu_1 = \Delta_0 = 0$ ∴ A resistência a compressão permanece igual

$H1: \mu_2 - \mu_1 > \Delta_0 > 0$ ∴ Rejeitar **H_0** se o novo ingrediente aumentar a resistência

O teste a ser aplicado é o teste unilateral a direita e o critério de rejeição da hipótese nula, estipulado na Tabela 26, é $Z_{0bs} > Z_{\alpha}$. Usando a estatística do teste, temos:

$$Z_{obs} = \frac{\bar{X}_2 - \bar{X}_1 - \Delta_0}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}} = \frac{17{,}2 - 15{,}5 - 0}{\sqrt{\frac{2{,}7^2}{15} + \frac{2{,}7^2}{10}}} = 1{,}5423 \;\therefore\; Z_{obs} = 1{,}5423$$

O valor de Z na tabela normal padronizada (Figura 37) para α = 0,05 (1 – α = 0,95) é $Z_{\alpha} = 1{,}6449$. Assim temos que $Z_{obs} < Z_{\alpha}$, o que nos coloca na região de aceitação de *H0*, como pode ser visualizado na Figura 57.

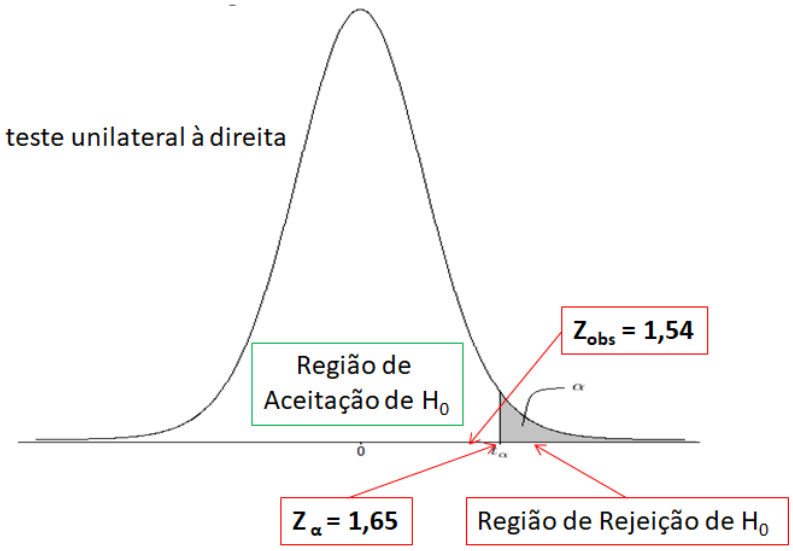

Figura 57 - Comparação de duas médias com variância conhecida

Outra forma de análise é o p-valor (probabilidade da estatística). $P(Z_{obs}) = 1 - P(1{,}5423) = 1 - 0{,}9385 = 0{,}0615$. O p-valor é igual a 0,0615 e é maior que a estatista proposta para o teste (0,05) conduzindo a aceitação de H0 (a subtração é realizada porque nos interessa a área de rejeição de *H0* e o valor 0,9385 corresponde à área da curva até o $Z_\alpha = 1{,}65$).

Em termos de probabilidade, há uma probabilidade de 6,15% de encontrarmos médias de resistências a compressão iguais para a população, o que é superior ao limite de 5% estabelecido.

Com base no exposto, aceita-se **H_0: $\mu_1 - \mu_2 = \Delta_0 = 0$** e fica estabelecido que não há diferenças estatísticas significativas entre as médias das formulações propostas no experimento.

Variâncias iguais e desconhecidas

Suponha que queremos comparar a diferença nas médias μ_1 e μ_2 ($\mu_1 - \mu_2 = \Delta_0$) de duas populações normais e independentes, sendo suas variâncias iguais mas desconhecidas; ($\sigma_1^2 = \sigma_2^2 = \sigma^2$).

Como sabemos que as variâncias são iguais, mas desconhecidas, precisamos combinar as duas variâncias das amostras, a partir dos desvios padrões calculados S_1 e S_2 para formar um estimador da variância σ. Este estimador é denominado S_P^2 e é definido por:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \qquad \textit{Eq. 45}$$

O número de graus de liberdade para a comparação da média, neste caso, será dado por $n_1 + n_2 - 2$. A estatística do teste é:

$$T = \frac{(\overline{X_1} - \overline{X_2}) - \Delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \textit{Eq. 46}$$

A hipótese nula é dada por: **H_0 : $\mu_1 - \mu_2 = \Delta_0$.** O teste pode ser resumido pela Tabela 27.

$H_0: \mu_1 - \mu_2 = \Delta_0$		
Hipóteses Alternativas	**Valor P**	**Critérios de rejeição de H_0**
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probabilidade acima de $\lvert T_0 \rvert$ e abaixo de $-\lvert T_0 \rvert$ $P = 2[1 - \phi(\lvert T_0 \rvert)]$	$T_0 > T_{\alpha/2, N1+N2-2}$ ou $T_0 < -T_{\alpha/2, N1+N2-2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probabilidade acima de T_0 $P = 1 - \phi(T_0)$	$T_0 > T_{\alpha, N1+N2-2}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probabilidade abaixo de T_0 $P = \phi(T_0)$	$T_0 < -T_{\alpha, N1+N2-2}$

Tabela 27 - Comparação de duas médias com variância iguais e desconhecidas

Exemplo 16: A adição de agregados de resíduos de concreto deve ser testada. Para tanto, foram testadas duas amostras: a primeira, com agregados naturais (AN) e a segunda com substituição de 25% dos agregados naturais por agregados de resíduos de concreto (ARC). Não houve alteração dos demais fatores. Os resultados dos testes de compressão são apresentados a seguir: A1 (38,76; 40,18 e 41,89) e A2 (41,66; 41,16 e 42,70). Supondo-se que a variância populacional para os tipos de concreto é igual (mas desconhecida), analise as amostras no nível de significância de 0,05.

Este experimento pede para que os resultados das amostras sejam analisados. São duas formulações diferentes, a segunda com adição de resíduos. Vamos então analisar se as médias de resistência a compressão são iguais. Inicialmente vamos exibir os valores das amostras com o uso do boxplot apresentado na Figura 58. A visualização dos boxplots permite supor que a amostra B possui valores superiores, então vamos construir a hipótese para o teste baseado nesta suposição.

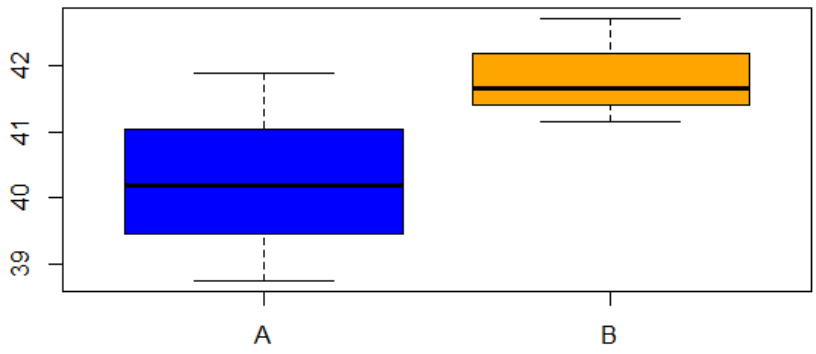

Figura 58 - Boxplot com os dados do exemplo 16

Cada amostra possui 3 elementos, portanto temos GL = 3 + 3 – 2 = 4 e devemos usar a distribuição de t-Student. As hipóteses para o teste (unilateral à direita) são:

$H0: \mu_B - \mu_A = \Delta_0 = 0$ ∴ A resistência a compressão é igual para as duas amostras

$H1: \mu_B - \mu_A > \Delta_0 > 0$ ∴ Rejeitar H_0 se o novo ingrediente aumentar a resistência

Em primeiro lugar, vamos calcular a média e o desvio padrão amostral (Tabela 28):

				Média	Desv.Pad.
CR (A)	38,76	40,18	41,89	40,28	1,57
25% ARC (B)	41,66	41,16	42,70	41,84	0,79

Tabela 28 - Cálculo da média e desvio padrão amostral

Antes de calcularmos a estatística do teste, temos que calcular o estimador da variância s_p:

$$s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} = \sqrt{\frac{(3-1)1{,}57^2 + (3-1)0{,}79^2}{3+3-2}} = 1{,}2428$$

Com a estatística do teste, temos:

$$T_{obs} = \frac{(\overline{X_B} - \overline{X_A}) - \Delta_0}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{(41{,}84 - 40{,}28) - 0}{1{,}2428\sqrt{\frac{1}{3} + \frac{1}{3}}} = 1{,}5373 \quad e \quad T_{obs} = 1{,}5373$$

O valor de $T_{(\alpha,4)}$ na tabela t-Student (Figura 38) para α = 0,05 (1 – α = 0,95) é $T_{(\alpha,4)}$ = 2,132. Assim temos que $T_{obs} < T_{(\alpha,4)}$ o que nos coloca na região de aceitação de *H0* ("A resistência a compressão é igual para as duas amostras"), como pode ser visualizado na Figura 59.

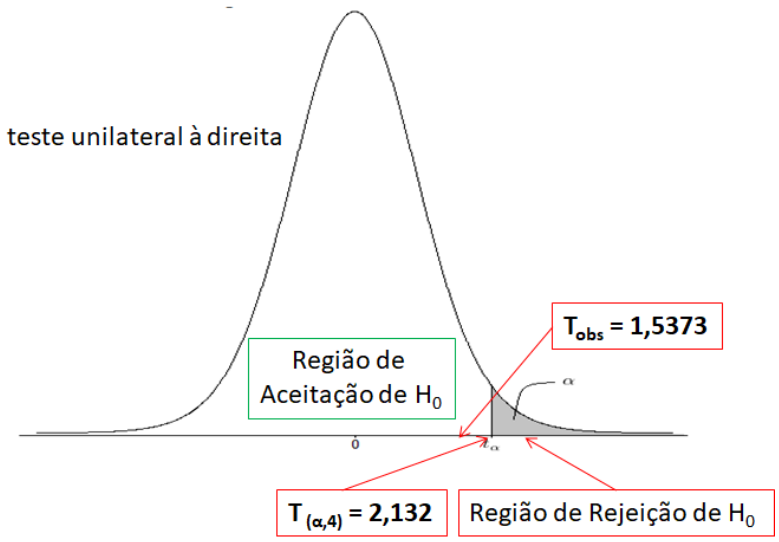

Figura 59 - Comparação de duas médias com variâncias iguais e desconhecidas

Usando o p-valor (probabilidade da estatística) temos $P(T_{obs}) = P(1{,}5373) = 0{,}0995$. O p-valor é igual a 0,0995 e é maior que a estatista proposta para o teste (0,05) conduzindo a aceitação de *H0.*

Em termos de probabilidade, há uma probabilidade de 9,95% de encontrarmos médias de resistências a compressão iguais para a população, o que é superior ao limite de 5% estabelecido.

Com base no exposto, aceita-se **H_0: $\mu_1 - \mu_2 = \Delta_0 = 0$** e fica estabelecido que não há diferenças estatísticas significativas entre as médias das amostras, apesar da suposição inicial feita pela interpretação dos boxplots exibidos na Figura 58.

Variâncias desconhecidas

Suponha que queremos comparar a diferença nas médias μ_1 e μ_2 ($\mu_1 - \mu_2 = \Delta_0$) de duas populações normais e independentes, sendo suas variâncias desconhecidas; ($\sigma_1^2 \neq \sigma_2^2$). Neste caso, como as variâncias são desconhecidas e, possivelmente desiguais, precisamos estimar os graus de liberdade (υ) com o uso da equação:

$$\upsilon = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} \qquad \textit{Eq. 47}$$

Com o número de graus de liberdade para a comparação da média dado pela expressão acima, a estatística do teste é:

$$T_{obs} = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \qquad \textit{Eq. 48}$$

A hipótese nula também é dada por: **$H_0 : \mu_1 - \mu_2 = \Delta_0$.** O teste pode ser resumido pela tabela apresentada na Tabela 29.

$H_0: \mu_1 - \mu_2 = \Delta_0$		
Hipóteses Alternativas	**Valor P**	**Critérios de rejeição de H_0**
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probabilidade acima de $\lvert T_0 \rvert$ e abaixo de - $\lvert T_0 \rvert$ $P = 2[1 - \phi(\lvert T_0 \rvert)]$	$T_0 > T_{\alpha/2, v}$ ou $T_0 < -T_{\alpha/2, v}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probabilidade acima de T_0 $P = 1 - \phi(T_0)$	$T_0 > T_{\alpha, v}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probabilidade abaixo de T_0 $P = \phi(T_0)$	$T_0 < -T_{\alpha, v}$

Tabela 29 - Hipóteses para variâncias desconhecidas

Exemplo 17: Segundo o fabricante, a adição de determinado aditivo aumenta a resistência a compressão do concreto em, no mínimo, 10%. Para testar este aditivo, uma empresa produziu, usando o mesmo método, duas amostras com 10 elementos, mostrados na Tabela 30. A primeira amostra (A), com a formulação padrão usada pela empresa e a segunda amostra (B), com a inclusão do aditivo nas proporções indicadas pelo fabricante. Os testes da primeira amostra resultaram em uma média amostral $\overline{X_A}$ de 44,80 MPa e desvio padrão de 3,93 MPa. A segunda amostra obteve média amostral $\overline{X_B}$ de 50,36 MPa e desvio padrão amostral de 4,96 MPa. Verifique se o aditivo atinge os objetivos propostos com nível de significância de 0,05.

Amostra 1 (A)	46,46	45,79	39,14	39,38	49,53	42,20	47,31	44,94	50,62	42,64
Amostra 2 (B)	53,90	53,04	55,94	47,97	52,00	51,01	42,53	57,04	45,38	44,83

Tabela 30 - Valores das amostras A e B

Esta é uma situação diferente. Temos que verificar se o concreto produzido com o aditivo apresentará um aumento na resistência a compressão de, no mínimo, 10%. São duas formulações diferentes, a segunda com o aditivo. Novamente, iremos usar o gráfico de boxplot para auxiliar a definição das hipóteses. A Figura 60 exibe as informações sobre as amostras.

Como as informações do gráfico de boxplot não permitem suposições, vamos nos ater ao enunciado do exemplo.

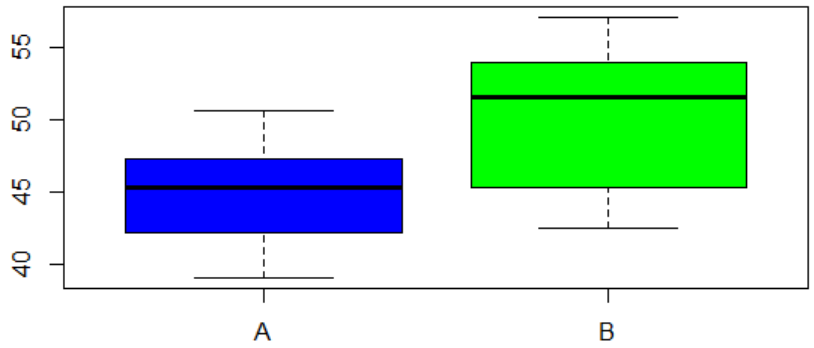

Figura 60 - Boxplot com os dados do exemplo 17

Como a quantidade de elementos das amostras é inferior a 30 e não temos informações sobre a variância populacional, devemos usar a distribuição de t-Student com a suposição de variâncias desconhecidas. Como o fabricante afirma que o aumento na resistência é superior a 10%, as hipóteses para o teste podem ser definidas como (unilateral à direita):

$H0: \mu_B - \mu_A = \Delta_0 = 10\% \mu_A$ ∴ A resistência a compressão atingiu o aumento
$H1: \mu_B - \mu_A > \Delta_0 > 10\% \mu_A$ ∴ Rejeitar **H_0** se o aumento for superior a 10% (fabricante)

Usando os valores fornecidos para a média e desvio padrão amostrais, o teste de hipóteses pode ser transcrito para α = 0,05 como:

$H0: \mu_B - \mu_A = 4{,}48$ ∴ A resistência a compressão atingiu o aumento
$H1: \mu_B - \mu_A > 4{,}48$ ∴ Rejeitar **H_0** se o aumento for superior a 10%

Inicialmente vamos calcular a estimativa dos graus de liberdade υ:

$$\upsilon = \frac{(s_A^2/n_1 + s_B^2/n_B)^2}{(s_A^2/n_A)^2/(n_A-1) + (s_B^2/n_B)^2/(n_B-1)} = \frac{(3{,}93^2/10 + 4{,}96^2/10)^2}{\frac{(3{,}93^2/10)^2}{9} + \frac{(4{,}96^2/10)^2}{9}} = 17{,}1159 \cong 17$$

A estatística do teste é:

$$T_{obs} = \frac{\bar{X}_B - \bar{X}_A - \Delta_0}{\sqrt{s_A^2/n_A + s_B^2/n_B}} = \frac{50{,}36 - 44{,}8 - 4{,}48}{\sqrt{3{,}93^2/10 + 4{,}96^2/10}} = 0{,}5399$$

O critério para rejeição da hipótese nula (Tabela 29) é $T_{obs} > T_{(\alpha,17)}$. O valor de $T_{(\alpha,17)}$ na tabela t-Student (Figura 38) para α = 0,05 (1 – α = 0,95) é $T_{(\alpha,17)}$ = 1,74. Assim temos que $|T_{obs}| < T_{(\alpha,17)}$ o que nos coloca na região de aceitação de *H0*.

Calculando a probabilidade com base no valor da estatística do teste (T_{obs}), temos $P(|T_{obs}|) = P(0{,}5397) = 0{,}298134$ (valor calculado no MS Excel pela função *DISTT(0,5399; 17; 1)*, respectivamente, valor T observado, graus de liberdade e unicaudal). O p-valor é igual a 0,298134 corresponde a uma probabilidade de 29,81% e é maior que a estatista proposta para o teste (0,05 = 5%) conduzindo a aceitação de *H0.*

Em termos de probabilidade, há uma probabilidade de 29,81% de encontrarmos médias de resistências a compressão inferiores ao aumento de 10% prometido para o concreto com o aditivo (considerando-se a população), o que é bem superior ao limite de 5% estabelecido.

Com base no exposto, aceita-se **H_0: $\mu_1 - \mu_2 = \Delta_0 = 4{,}48$** e, consequentemente, rejeita-se **H1** (aumento superior a 10%).

Comentários: Comparando as médias da amostra sem o aditivo (X_B = 44,8 MPa) e após o uso do aditivo (X_A = 50,36 MPa) temos a impressão que o objetivo do experimento foi atingido pois as médias demonstram o aumento de 10% (44,8 + 4,48 = 49,28), pois o valor de X_2 é maior que 49,28 MPa. No entanto, não podemos nos esquecer que a representa apenas o valor central da distribuição.

Vamos considerar apenas a amostra com o aditivo (X_1 = 50,36 MPa) e levantar a probabilidade de encontramos valores superiores a 49,28 MPa usando o RStudio. Para o teste o vetor *dados$a1* (criado com base nos dados da Tabela 30) foi carregado com os dez valores de resistência à compressão da amostra:

```
> t.test(dados$a1,mu=49.28, alternative = "greater")

        One Sample t-test

data:  dados$a1
t = 0.69178, df = 9, p-value = 0.2533
alternative hypothesis: true mean is greater than 49.28
95 percent confidence interval:
 47.49156      Inf
sample estimates:
mean of x
   50.364
```

Conforme o teste comprova, a hipótese $H0: \mu_1 = 49{,}28$ não pode ser rejeitada e, consequentemente, a hipótese que nos interessa $H1: \mu_1 > 49{,}28$ não pode ser comprovada. O p-valor nos indica o percentual de ocorrências de médias superiores a 49,28 MPa (25,33%) e isto considerando apenas a amostra com adição.

Outra maneira de entendermos o teste é plotarmos a distribuição de probabilidades populacional das duas amostras (Figura 61). No gráfico estão destacadas as médias das duas amostras e o valor (49,28) que 95% das resistências a compressão dos elementos deveriam superar, uma vez que o enunciado pede um nível de significância de 0,05.

Podemos visualizar que um pouco menos que 50% da área sob o gráfico da distribuição de frequência está à esquerda da linha verde que delimita o valor de 49,28 MPa, indicando que um percentual correspondente à esta área possui resistência a compressão inferior a este limite.

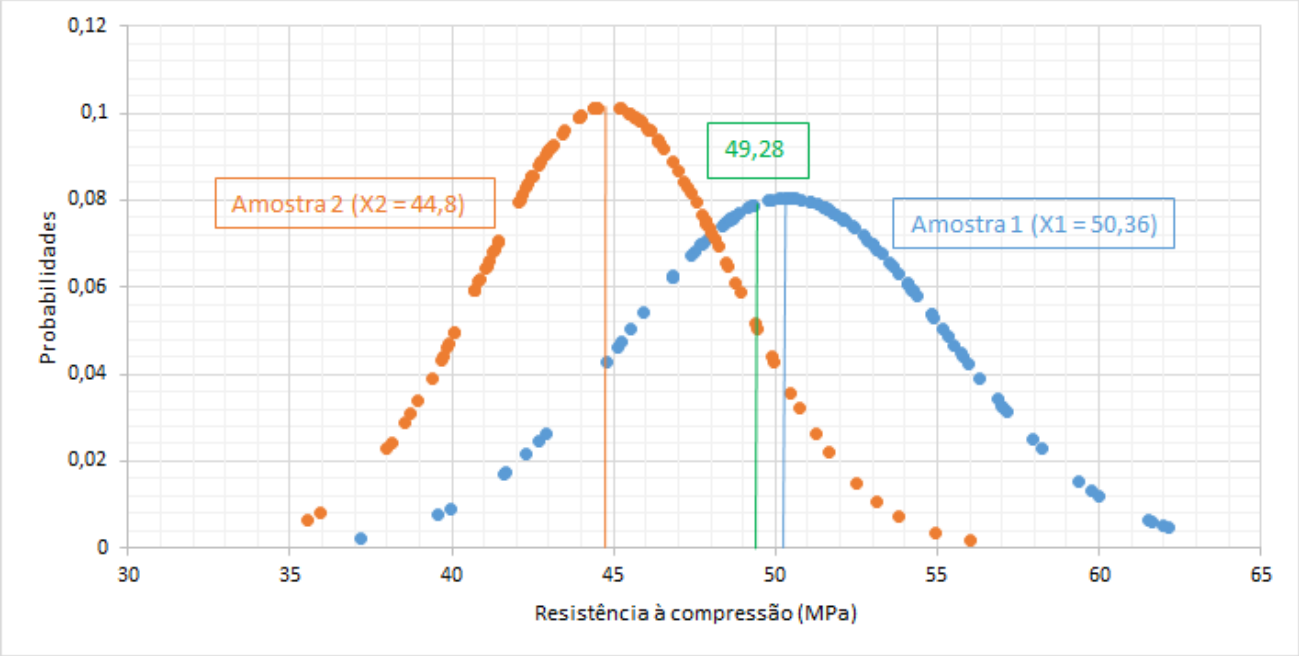

Figura 61- Distribuição de probabilidades populacional das duas amostras

Também podemos executar o teste *t* do RStudio para compararmos a média juntamente com a diferença esperada (os vetores a1 e a2 contém os valores de resistência a compressão das amostras):

```
> t.test(dados$a1,dados$a2,mu = 4.48, var.equal=F,alternative="greater")

        Welch Two Sample t-test

data:  dados$a1 and dados$a2
t = 0.54141, df = 17.116, p-value = 0.2976
alternative hypothesis: true difference in means is greater than 4.48
95 percent confidence interval:
 2.084536      Inf
sample estimates:
mean of x mean of y
   50.364    44.801
```

As hipóteses para o teste as mesmas usadas na solução do exemplo:

$$H0: \mu_1 - \mu_2 = 4{,}48$$
$$H1: \mu_1 - \mu_2 > 4{,}48$$

Como o p-valor do teste é 0,2979 não podemos rejeitar a hipótese H0 (ela é aceita) e consequentemente não conseguimos comprovar que o aumento na resistência a compressão é superior a 10%. O p-valor encontrado é bem próximo ao que foi calculado no MS Excel (0,2981).

Dados Pareados

Um caso especial de teste *t* para duas amostras ocorre quando as observações nas duas populações são coletadas em pares. Cada par de observações é tomado em condições homogêneas, mas que podem mudar de uma observação para outra. É o caso de termos o mesmo corpo de prova submetido a duas observações, sendo que o **único** fator que as diferencia é o tratamento ao qual o corpo foi submetido. Assim, podemos considerar que temos uma amostra de pares (X_1,Y_1; X_2,Y_2; ...; X_N,Y_N). Neste caso, o valor de interesse não são as mensurações das amostras, mas a **diferença** entre elas.

Para entendermos melhor o significado de dados pareados, imagine um grupo de pessoas submetidas ao mesmo regime e controladas durante o regime. Teríamos o peso de cada elemento antes e depois de um certo período de tempo. Considerando-se que o único fator que pode influenciar a alteração do peso é o regime, teríamos dados pareados (peso antes e peso depois).

Denominando de D a diferença entre as mensurações da característica de interesse, o teste usa a distribuição t-Student com (*n – 1*) graus de liberdade. A estatística do teste é dada por:

$$T = \frac{\overline{D} - \Delta_0}{s_D/\sqrt{n}} \qquad \textit{Eq. 49}$$

Onde:

- D : média da diferença das mensurações dos pares
- Δ_0 : valor esperado na comparação
- S_D : desvio padrão da diferença das mensurações dos pares
- n : número de elementos nas amostras

O teste de hipóteses para amostras pareadas é baseado na análise da diferença entre as mensurações dos pares $D_i = X_i - Y_i, para\ i = 1, 2, \dots, n$, sendo μ_D a média destas diferenças. As hipóteses para o teste são:

$H0: \mu_D = \Delta_0$	$H0: \mu_D = \Delta_0$	$H0: \mu_D = \Delta_0$
$H1: \mu_D \neq \Delta_0$	$H1: \mu_D < \Delta_0$	$H1: \mu_D < \Delta_0$

A hipótese nula é dada por: **$H_0 : \mu_D = \Delta_0$.** O teste resumido é apresentado na Tabela 31.

H0: μD = Δ0		
Hipóteses Alternativas	**Valor P**	**Critérios de rejeição de H0**
H1: μD ≠ Δ0	Probabilidade acima de |T0| e abaixo de - |T0| P = 2[1 – ϕ(|T0|)]	Tobs > T α/2, n-1 ou Tobs < - T α/2, n-1
H1: μD > Δ0	Probabilidade acima de T0 P = 1 – ϕ(T0)	Tobs >Tα, n-1
H1: μD < Δ0	Probabilidade abaixo de T0 P = ϕ(T0)	Tobs < - Tα, n-1

Tabela 31 - Hipóteses para dados pareados

Exemplo 18: Dois métodos (A e B) diferentes de previsão da resistência à compressão de corpos de prova de concreto estão sendo avaliados em uma pesquisa. Os dois métodos foram aplicados em 9 corpos de prova e a resistência à compressão prevista foi calculada. Em seguida, os corpos de prova foram rompidos e sua resistência a compressão foi mensurada. A resistência a compressão mensurada (RCM) e a prevista pelos métodos (RPA e RPB) é apresentada na Tabela 32.

Com um nível de significância α de 0,05 e partindo do pressuposto que RC_M (resistência mensurada) representa o valor real da resistência a compressão, determine: (a) os métodos A e B podem ser considerados estatisticamente diferentes? (b) comprove qual o método mais adequado.

As médias e desvios padrão amostrais são exibidos na Tabela 33.

	RC_M	RP_A	RP_B
1	45,30	45,00	53,10
2	45,24	46,08	66,45
3	46,88	49,77	58,91
4	46,58	47,94	62,7
5	54,07	44,26	59,43
6	49,38	43,88	59,00
7	46,54	50,41	48,46
8	49,05	44,39	51,69
9	41,64	46,01	56,44

Tabela 32 - Resistência a compressão real e prevista pelos métodos A e B

	RC_M	RP_A	RP_B
Média	47,19	46,42	57,35
Desv.P.	3,43	2,42	5,56

Tabela 33 - Média e desvio padrão amostrais

a) O método de comparação de dados pareados compara dados de duas amostras pareadas, e, neste caso, temos três amostras. Assim, os dados devem ser tratados em função do objetivo, a saber, determinar se os métodos oferecem respostas diferentes em prever RC_M. A solução é comparar os métodos A e B em função da razão entre a previsão e o valor mensurado.

A Tabela 34 exibe a relação $RP_A/RC_M \;\; e \;\; RP_B/RC_M$ calculada a partir dos dados da Tabela 32.

rpa	0,9934	1,0186	1,0616	1,0292	0,8186	0,8886	1,0832	0,9050	1,1049
rpb	1,1722	1,4688	1,2566	1,3461	1,0991	1,1948	1,0413	1,0538	1,3554
Di	-0,1788	-0,4503	-0,1950	-0,3169	-0,2806	-0,3062	0,0419	-0,1488	-0,2505

Tabela 34 - Relação entre resistência prevista e resistência mensurada para os métodos A e B

Onde *Dj* é a diferença entre a razão das mensurações da característica de interesse (no caso a relação entre as resistências à compressão prevista e mensurada). Então temos:

$\overline{D} = -0{,}2317$

$S_D = 0{,}1366$

$n = 9 \; e \;\; GL = n - 1 = 8$

A hipótese inicial para o teste pode ser verificar se os métodos oferecem respostas diferentes para a previsão de resistência à compressão (teste bilateral). Assim, temos, considerando $\Delta_0 = 0$:

$H0 : \mu_D = 0$
$H1 : \mu_D \neq 0$

Onde μ_D representa a média das diferenças D_i. A Tabela 31 nos dá os critérios de rejeição de H0, a saber, $T_{obs} > T_{(0{,}025,8)} \; ou \; T_{obs} < -T_{(0{,}025,8)}$. O valor de $T_{(0{,}025,8)}$ pode ser obtido na tabela da Figura 38 e é igual a 2,306.

A estatística do teste é dada por:

$$T = \frac{\overline{D} - \Delta_0}{S_d/\sqrt{n}} = \frac{-0{,}2317 - 0}{0{,}1366/\sqrt{9}} = -5{,}0896$$

Como T_{obs}é menor que $-T_{(0{,}025,8)}$, isto é, -5,0896 < -2,306, a hipótese nula é rejeitada e temos que os resultados gerados pelos métodos são estatisticamente diferentes, sendo que a probabilidade associada a estatística T é de 99,9% (muito superior aos 5% permitido pelo teste).

b) Para identificarmos o método mais adequado podemos realizar o teste *t* pareado comparando o valor real com o previsto em cada um dos métodos.

Para o método A temos:

mensurado	45,30	45,24	46,88	46,58	54,07	49,38	46,54	49,05	41,64
método A	45,00	46,08	49,77	47,94	44,26	43,88	50,41	44,39	46,01
Dj	0,30	-0,84	-2,89	-1,36	9,81	5,50	-3,87	4,66	-4,37

Tabela 35 - Resistência mensurada, resistência calculada (A) e diferença entre elas

Assim temos:

$\overline{D} = 0{,}77$

$S_D = 4{,}8467$

A hipótese do teste é $H0: \mu_D = 0$, ou seja, o método A representa o valor real da resistência a compressão. Assim, considerando $\Delta_0 = 0$, as hipóteses são:

$$H0 : \mu_D = 0$$
$$H1 : \mu_D \neq 0$$

Novamente, a Tabela 31 nos dá os critérios de rejeição de H0, a saber, $T_{obs} > T_{(0{,}025,8)}\ ou\ T_{obs} < -T_{(0{,}025,8)}$. O valor de $T_{(0{,}025,8)}$ pode ser obtido na tabela da Figura 38 e é igual a 2,306. A estatística do teste é:

$$T_{obs} = \frac{\overline{D} - \Delta_0}{S_d/\sqrt{n}} = \frac{0{,}77 - 0}{4{,}8467/\sqrt{9}} = 0{,}4766$$

Como $T_{obs}\ (0{,}4766) < T_{(0{,}025,8)}(2{,}306)$, a hipótese nula não pode ser rejeitada e temos que os resultados gerados pelo método A podem ser considerados similares as mensurações efetuadas.

Para o método B (Tabela 36):

mensurado	45,30	45,24	46,88	46,58	54,07	49,38	46,54	49,05	41,64
método B	53,10	66,45	58,91	62,70	59,43	59,00	48,46	51,69	56,44
Dj	-7,80	-21,21	-12,03	-16,12	-5,36	-9,62	-1,92	-2,64	-14,80

Tabela 36 - Resistência mensurada, resistência calculada (B) e diferença entre elas

Assim temos:

$\overline{D} = -10{,}17$

$S_D = 6{,}481$

A hipótese do teste continua a mesma, ou seja, $H0: \mu_D = 0$, ou seja, o método B representa o valor real da resistência a compressão. Assim, considerando $\Delta_0 = 0$, as hipóteses são:

$$H0: \mu_D = 0$$
$$H1: \mu_D \neq 0$$

Considerando a mesma situação da comparação anterior, temos os mesmos critérios de rejeição de H0 e o valor de $T_{(0,025,8)}$ igual a 2,306. A estatística do teste é:

$$T_{obs} = \frac{\overline{D} - \Delta_0}{S_d/\sqrt{n}} = \frac{-10{,}17 - 0}{6{,}481/\sqrt{9}} = -4{,}7076$$

Como $T_{obs}\ (-4{,}7076) < -T_{(0,025,8)}(-2{,}306)$, o critério de rejeição de H0 é satisfeito e a hipótese nula é rejeitada. Assim, podemos concluir que os resultados gerados pelo método B são estatisticamente diferentes das mensurações efetuadas.

Analisando os resultados das comparações, temos que o método A apresenta resultados similares aos dos testes reais e o método B não, de onde podemos concluir que o método A é adequado.

7.8 Erros Cometidos nos Testes de Hipóteses

Como estamos tratando de hipóteses e probabilidades de acerto, nenhum teste é 100% confiável, pois há sempre a probabilidade de chegarmos à conclusão errada. A realização de um teste de hipóteses conduz a dois tipos de erros possíveis: erro tipo I e erro tipo II. Os riscos de ocorrência desses dois tipos de erro são inversamente proporcionais, ou seja, quanto mais nos esforçamos para diminuir um, aumentamos o outro. Os tipos de erro são determinados pelo nível de significância (α) do teste e pelo poder do teste (β).

Erro Tipo I

Quando a hipótese nula (**H0**) é verdadeira e o teste realizado indica sua rejeição, é cometido um erro do tipo I. A probabilidade de cometer um erro do tipo I é dada pelo nível de significância α definido para o teste de hipóteses. Um α de 0,05 indica que é aceito uma chance de 5% de que o teste pode errar ao rejeitar a hipótese nula. Para reduzir este risco, pode ser usado um valor inferior para α. Entretanto isto acarreta que o teste terá uma menor probabilidade de detectar uma diferença verdadeira (rejeição de **H0)**, quando ela realmente existe.

Erro Tipo II

Quando a hipótese nula (**H0**) é falsa e o teste realizado não a rejeita, é cometido um erro de tipo II. A probabilidade de cometer um erro de tipo II é dada por β. A probabilidade de ocorrência do erro tipo II pode ser diminuída com o aumento do poder do teste. Isto pode ser feito, por exemplo, garantindo-se que o tamanho da amostra seja grande o suficiente para detectar uma diferença, quando ela realmente existir. Como a probabilidade de não rejeitar uma hipótese nula falsa é dada por β, o valor 1 – β refere-se à probabilidade de realmente rejeitar a hipótese nula falsa (**H0**). Esse valor (1 – β) é denominado poder ou potência do teste.

Para entendermos a relação entre os erros tipo I e tipo II e para determinar qual dos tipos de erro terá consequências mais danosas em um determinado teste, vamos considerar a seguinte situação:

Um pesquisador deseja comparar a eficácia de dois aditivos na cura do concreto e estabeleceu as seguintes hipóteses:

$HO: \mu_1 = \mu_2$ Os dois aditivos são igualmente eficazes

$H1: \mu_1 \neq \mu_2$ Os aditivos não são igualmente eficazes

Um erro do tipo I ocorre se o teste realizado pelo pesquisador rejeita a hipótese nula (***H0***) e conclui que os dois aditivos possuem eficácia diferente, quando, na realidade, a eficácia é a mesma. Se os aditivos tiverem a mesma eficácia, a pesquisa poderá não considerar este erro muito severo porque a cura do concreto será similar, independentemente de qual aditivo for usado.

Contudo, se ocorrer um erro do tipo II, o teste realizado pelo pesquisador não irá rejeitar a hipótese nula (H0), quando essa hipótese deveria ter sido rejeitada. Assim, a pesquisa irá concluir que os aditivos possuem a mesma eficácia quando, na realidade, não possuem. Este erro possui potencial para invalidar uma pesquisa, pois termina por recomendar um aditivo que não é eficaz para o que se propõe. Agora imagine a mesma situação para um medicamento prestes a ser comercializado para o público.

Poder ou Potência do teste

O poder ou potência do teste tem como objetivo conhecer o quanto o teste de hipóteses controla um erro do tipo II, ou seja, qual a probabilidade de não rejeitar a hipótese nula se esta for falsa.

O poder de um teste de hipóteses é afetado por três fatores: tamanho da amostra, nível de significância e a diferença entre o valor real e o valor suposto para o teste.

Tamanho da amostra: Como já citado anteriormente, quanto maior o tamanho da amostra, maior a confiabilidade da análise, ou seja, com os outros parâmetros constantes, quanto maior o tamanho da amostra, maior o poder do teste.

Nível de Significância: Se o nível de significância (α) é aumentado, a área de rejeição do teste também aumenta. Da mesma forma, a região de aceitação (1 – α) é proporcionalmente reduzida. Como resultado, aumentam as chances de rejeitar a hipótese nula. Isto significa que o teste tem menos chance de aceitar (não rejeitar) a hipótese nula quando ela é falsa, e consequentemente, menor chance de cometer um erro do tipo II. Então, o poder do teste aumenta.

O valor real do parâmetro a ser testado: Quanto maior a diferença entre o valor real do parâmetro e o valor especificado pela hipótese nula, maior o poder do teste, pois é mais fácil para o teste detectar essa diferença.

Para entendermos melhor o poder do teste, consideremos a estatística (Eq. 37):

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

E o teste de hipóteses

$HO: \mu = \mu_0$

$H1: \mu \neq \mu_0$

O erro do tipo II (β) é cometido ao não rejeitar (aceitar) a hipótese nula (H0) quando ela é falsa (H1 é verdadeira). Então, suponha que a média real é $\mu = \mu_0 + \Delta$ o que leva a hipótese nula ser falsa. Considerando isto, a estatística do teste passa a ser:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - (\mu_0 + \Delta)}{\sigma/\sqrt{n}} + \frac{\Delta}{\sigma/\sqrt{n}}$$

A distribuição de Z_0 quando $\mu = \mu_0 + \Delta$ é:

$$Z_0 \sim N\left(\frac{\Delta}{\sigma/\sqrt{n}}, 1\right)$$

Para um teste bilateral, a probabilidade do erro tipo II (não rejeitar H_0) é a probabilidade de que Z_0 esteja entre $-Z_{\alpha/2}$ e $Z_{\alpha/2}$ uma vez que H_1 é verdadeira. Esta probabilidade é dada por:

$$\beta = \Phi\left(Z_{\alpha/2} - \frac{\Delta\sqrt{n}}{\sigma}\right) - \Phi\left(-Z_{\alpha/2} - \frac{\Delta\sqrt{n}}{\sigma}\right) \qquad \textit{Eq. 50}$$

Onde Φ é a função distribuição acumulada da distribuição normal padrão. Para os testes unilaterais à esquerda e à direita, as probabilidades do erro tipo II (β) são, respectivamente:

$$1 - \Phi\left(-Z_{\alpha/2} - \frac{\Delta\sqrt{n}}{\sigma}\right) \quad e \quad \Phi\left(Z_{\alpha/2} - \frac{\Delta\sqrt{n}}{\sigma}\right) \qquad \textit{Eq. 51}$$

E o poder do teste é dado por: Poder = 1 – β.

Exemplo 19: Uma empresa quer testar, com base em uma amostra aleatória de 30 elementos, com um nível de significância de 0,05, se o diâmetro das barras de aço produzidas é de 8,0 mm. A amostra obteve um diâmetro médio de 8,09 mm e se sabe, de experimentos anteriores que o desvio populacional é de 0,22 mm. Verifique a hipótese e determine o poder do teste.

As hipóteses são:

$HO: \mu = 8{,}0$

$H1: \mu \neq 8{,}0$

Como α = 0,05 e usaremos o teste bilateral, temos que $Z_{\alpha/2} = Z_{0{,}025} = 1{,}96$ o critério a ser aplicado é rejeitar H0 se $Z_{obs} < -1{,}96 \; ou \; Z_{obs} > 1{,}96$. Assim temos:

$$Z_{obs} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{8{,}09 - 8{,}0}{0{,}22/\sqrt{30}} = 2{,}24$$

Como $Z_{obs} = 2{,}24 \; > 1{,}96$, a hipótese nula (H0) é rejeitada, ou seja, o teste não indica que a média populacional μ seja igual a 8,0 mm. A diferença entre 8,0 e 8,09 é significativa.

O poder do teste bilateral é dado por:

$$1 - \beta = 1 - \Phi\left(Z_{\alpha/2} - \frac{\Delta\sqrt{n}}{\sigma}\right) - \Phi\left(-Z_{\alpha/2} - \frac{\Delta\sqrt{n}}{\sigma}\right) = 1 - \Phi\left(1{,}96 - \frac{0{,}09\sqrt{30}}{0{,}22}\right) - \Phi\left(1{,}96 - \frac{0{,}09\sqrt{30}}{0{,}22}\right)$$

$$1 - \beta = 1 - \Phi(-0{,}2807) - \Phi(-4{,}2007) = 1 - 0{,}3895 - 0{,}00001 = 0{,}6105$$

Assim, temos que o poder do teste em detectar diferença de 0,09 mm no diâmetro da barra é de 61,05%

7.9 Testes de Hipóteses – RStudio

No capítulo anterior foi visto os testes de hipóteses com suas respectivas fórmulas e exemplos. Neste capítulo vamos dedicar atenção para a execução destes mesmos testes estatísticos no RStudio.

O teste mais usado em comparações de amostras é o *t.test*, baseado na distribuição de t-Student, cuja teoria já foi apresentada. Os principais parâmetros para a execução deste teste são apresentados no Quadro 2. Para maiores informações, acesse a função *HELP* (tecla F1) do RStudio.

SINTAXE: t.test(x, y = NULL, alternative = ("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, …)	
x	um vetor numérico (não vazio) de valores de dados.
y	um vetor numérico (não vazio) de valores de dados. OPCIONAL
alternative	uma sequência de caracteres ("two.sided", "greater" ou "less") associada à hipótese alternativa ($\mu_0 \neq \mu$, $\mu_0 > \mu$ ou $\mu_0 < \mu$). Apenas a letra inicial pode ser usada. OPCIONAL. DEFAULT = "two.sided".
mu	valor real da média (ou diferença de médias se você estiver executando um teste de duas amostras). OPCIONAL. DEFAULT = 0 (zero)
paired	Variável lógica (TRUE / FALSE) indicando se é um teste com dados pareados ou não. OPCIONAL. DEFAULT = FALSE
var.equal	Variável lógica (TRUE / FALSE) indicando se as variâncias são iguais (TRUE) ou não (FALSE) . Se, TRUE então, a variância combinada é usada para estimar a variância, caso contrário, a aproximação de Welch (ou Satterthwaite) é usada. OPCIONAL. DEFAULT = FALSE
conf.level	Nível de confiança do intervalo. OPCIONAL. DEFAULT = 0,95 (α = 0,05)

Quadro 2 - Parâmetros para o t.test no RStudio

Vamos iniciar a execução dos testes de comparação de médias no RStudio com o último exemplo visto, amostras pareadas, usando o mesmo exemplo (Exemplo 18) do capítulo anterior.

Teste *t* com Dados Pareados: Vamos continuar com o exemplo 6, só que agora no RStudio. Para facilitar, os dados apresentados na Tabela 32, Tabela 34 e Tabela 35, foram carregados em planilha MS Excel no formato csv e importados para o RStudio, com os comandos abaixo:

```
> dados = read.csv2(file.choose(),header=T)
> dados
      m     a     b       rpa      rpb           d
1 45.30 45.00 53.10 0.9933775 1.172185  0.14027014
2 45.24 46.08 66.45 1.0185676 1.468833  0.33775500
3 46.88 49.77 58.91 1.0616468 1.256613  0.26585657
4 46.58 47.94 62.70 1.0291971 1.346071  0.28629437
5 54.07 44.26 59.43 0.8185685 1.099131 -0.09124134
6 49.38 43.88 59.00 0.8886189 1.194816  0.05166972
7 46.54 50.41 48.46 1.0831543 1.041255  0.12277458
8 49.05 44.39 51.69 0.9049949 1.053823 -0.04393139
9 41.64 46.01 56.44 1.1049472 1.355427  0.36717254
```

Para o comando *t.test* (teste de t-Student para comparação de médias), não precisamos calcular o valor D (diferença entre as mensurações), pois isto é feito internamente. Para o teste entraremos com os valores diretamente. Entretanto, para compararmos os métodos A e B diretamente, precisamos da razão entre as mensurações, dadas pelas variáveis aleatórias *dados$rpa* e *dados$rpb*.

Assim, para o item (a), a comparação dos métodos A e B, o comando no RStudio é:

```
> t.test(dados$rpa, dados$rpb, paired = TRUE, alternative = 'two.sided')

        Paired t-test

data:  dados$rpa and dados$rpb
t = -5.0896, df = 8, p-value = 0.0009419
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3366442 -0.1267071
sample estimates:
mean of the differences
             -0.2316757
```

O resultado do teste apresenta a estatística T (t = -5,0896) é igual a estatística calculada pela fórmula anterior e o p-valor = 0,0009419 é inferior a 0,05 fazendo com que a hipótese ***H0*** possa ser rejeitada. Os dois métodos apresentam resultados diferentes.

Para o item (b), podemos comparar diretamente os valores das amostras m/a e m/b e analisarmos os resultados do teste. Comparando a amostra m (mensuração da resistência a compressão) com a amostra a (método A), temos:

```
> t.test(dados$m, dados$a, paired = TRUE, alternative = 'two.sided')

        Paired t-test

data:  dados$m and dados$a
t = 0.4773, df = 8, p-value = 0.6459
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.954365  4.496588
sample estimates:
mean of the differences
              0.7711111
```

O resultado do teste apresenta a estatística T (t = 0,4773) que é praticamente a estatística calculada pela fórmula anterior (0,4766) e o p-valor = 0,6459 é superior a 0,05 fazendo com que a hipótese ***H0*** não possa ser rejeitada. Assim, os resultados obtidos pelo método A podem ser considerados iguais aos resultados reais.

Já na comparação da amostra m (mensuração da resistência a compressão) com a amostra b (método B), temos:

```
> t.test(dados$m, dados$b, paired = TRUE, alternative = 'two.sided')

        Paired t-test

data:  dados$m and dados$b
t = -4.7063, df = 8, p-value = 0.001529
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.148201  -5.185133
sample estimates:
mean of the differences
              -10.16667
```

O resultado do teste apresenta a estatística T (t = -4,7063) que é a mesma a estatística calculada pela fórmula anterior (-4,7076) e o p-valor = 0,001529 é inferior a 0,05 fazendo com que a hipótese H0 possa ser rejeitada. Assim, os resultados obtidos pelo método B não são iguais aos resultados reais.

Testes de Comparação de Médias com Duas Amostras no RStudio

Uma das vantagens do uso de softwares estatísticos, tipo o RStudio, é não precisar consultar tabelas para encontrar os valores críticos para os testes. Mas não se esqueça de que existem testes baseados na distribuição t-Student (*t.test*) e testes baseados na distribuição normal padronizada (*z.test*).

Exemplo 20: Pretende-se comparar amostras de duas concreteiras diferentes. Para tanto cada concreteira produziu amostras com 20 elementos que foram testados quanto a resistência a compressão. Os dados obtidos dos testes são mostrados na Tabela 37.

Sabendo-se que a Concreteira A informa que o valor médio da resistência a compressão é 40 MPa e a Concreteira B, 50 MPa, verifique se as informações são corretas, compare as amostras e determine a probabilidade das Concreteiras fornecerem material inferior ou superior ao da concorrente.

Concreteira A		Concreteira B	
36,63	55,17	52,00	49,37
48,23	30,76	52,89	47,76
32,04	38,59	53,01	51,97
44,26	29,99	53,11	51,29
39,22	57,77	58,57	49,42
37,18	35,31	49,09	50,26
27,86	26,73	52,49	54,16
41,50	34,36	44,31	48,39
20,13	49,99	47,55	43,00
37,20	60,01	54,82	47,76

Tabela 37 - Exemplo 20 - Resultados de resistência a compressão

Em primeiro lugar, vamos inserir os dados no RStudio a partir de planilha MS Excel no formato csv:

```
> dados = read.csv2(file.choose(),header=T)
> dados
       a     b
1  36.63 52.00
2  48.23 52.89
3  32.04 53.01
4  44.26 53.11
5  39.22 58.57
6  37.18 49.09
7  27.86 52.49
8  41.50 44.31
9  20.13 47.55
10 37.20 54.82
11 55.17 49.37
12 30.76 47.76
13 38.59 51.97
14 29.99 51.29
15 57.77 49.42
16 35.31 50.26
17 26.73 54.16
18 34.36 48.39
19 49.99 43.00
20 60.01 47.76
```

Com os dados carregados, podemos executar o teste *t* para comparação das duas amostras:

```
> t.test(dados$a, dados$b, alternative = "two.sided", conf.level = 0,05)

   Welch Two Sample t-test

data:  dados$a and dados$b
t = -6.501, df = 23.356, p-value = 1.15e-06
alternative hypothesis: true difference in means is not equal to 5
0 percent confidence interval:
 -11.4145 -11.4145
sample estimates:
mean of x mean of y
  39.1465   50.5610
```

O resultado do teste mostra que a hipótese ***H0*** (as amostras possuem médias iguais) é rejeitada (p-valor = 0,00000115). Além disto, as médias amostrais calculadas são de 39,15 MPa para a Concreteira A e 50,56 MPa para a Concreteira B. Para exemplificar, podemos executar o teste *t* para verificar a igualdade destas médias com os valores declarados no enunciado.

```
> t.test(dados$a, mu = 40, alternative = "two.sided", var.equal=T, conf.level= 0,05)

   Two Sample t-test

data:  dados$a and 5
t = -0.53447, df = 19, p-value = 0.5992
alternative hypothesis: true difference in means is not equal to 40
0 percent confidence interval:
 34.1465 34.1465
sample estimates:
mean of x mean of y
  39.1465    5.0000

> t.test(dados$b, mu= 50, alternative = "two.sided", var.equal=T, conf.level= 0,05)

   Two Sample t-test

data:  dados$b and 5
t = -1.1891, df = 19, p-value = 0.249
alternative hypothesis: true difference in means is not equal to 50
0 percent confidence interval:
 45.561 45.561
sample estimates:
mean of x mean of y
   50.561     5.000
```

Em ambos os testes, o p-valor é superior a 0,05. Com isto podemos aceitar a hipótese H0 estabelecida para os testes. Podemos aceitar que a média da resistência a compressão do material fornecido pela Concreteira A é igual a 40 MPa e que o mesmo ocorre para a Concreteira B (média = 50 MPa).

Agora vamos ver a probabilidade da Concreteira A fornecer um material com resistência a compressão superior ao da Concreteira B (média > 50 MPa):

```
> t.test(dados$a, mu= 50, alternative = "greater")

   One Sample t-test

data:  dados$a
t = -4.5414, df = 19, p-value = 0.9999
alternative hypothesis: true mean is greater than 50
95 percent confidence interval:
 35.01402      Inf
sample estimates:
mean of x
```

```
   39.1465
```

Com o p-valor = 0,9999, a probabilidade da Concreteira A fornecer um material com resistência a compressão superior a 50 MPa é igual 1 – P(x ≤ 50) = 1 – 0,9999 = 0,0001. Verificando no RStudio com base na estatística do teste e dos graus de liberdade, temos:

```
> pt(-4.5414,19)
[1] 0.000111605
```

Quanto a probabilidade da Concreteira B fornecer um material com resistência a compressão inferior a 40 MPa, basta executarmos o mesmo teste, alterando os parâmetros:

```
> t.test(dados$b, mu=40, alternative = "less")

   One Sample t-test

data:  dados$b
t = 12.965, df = 19, p-value = 1
alternative hypothesis: true mean is less than 40
95 percent confidence interval:
     -Inf 51.96954
sample estimates:
mean of x
   50.561
```

Com o p-valor = 1, a probabilidade de fornecimento de concreto com resistência a compressão inferior a 40 MPa é praticamente nula. Mas, para exemplo, vamos verificar a probabilidade associada a estatística do teste:

```
> 1 - pt(12.965,19)
[1] 3.470702e-11
```

O resultado é $3{,}47\ x\ 10^{-9}$%. Acredito que isto pode ser considerado como uma probabilidade praticamente nula[18].

Outros usos para o teste *t*

No Exemplo 12 foram dados os valores de resistência a compressão de quatro amostras com quantidades diferentes de elementos e foi pedido o intervalo de confiança para a média. Vamos calcular este intervalo usando a função *t.test* e *z.test* do RStudio, para fins de comparação. Inicialmente, vamos carregar o vetor que contém os dados das amostras e o pacote "**TeachingDemos**".

```
> dados = read.csv2(file.choose(), header=T)
> dados
          a        b        c         d
1  63.73392 71.01935 96.45293  95.24954
2  72.15981 65.38353 82.52394  95.13334
3  58.22972 81.93491 92.62981  85.44964
4  58.03466 72.97790 90.82530  86.13743
5        NA 58.68078 94.36778  79.50997
6        NA 52.53909 81.68666  86.55196
7        NA       NA 81.49332  84.44111
8        NA       NA 93.67926 108.37138
9        NA       NA       NA  94.39628
10       NA       NA       NA  94.19306
```

[18] Um leitor curioso teria notado que, no caso anterior, era só a probabilidade. Neste foi 1 – probabilidade. Porque?

Como foi informado no enunciado, vamos considerar a variância populacional como conhecida, o que nos leva ao *z.test* (baseado na distribuição normal). Agora, observe que como as amostras possuem tamanho diferente, teremos que delimitar o vetor em seu uso. Além disto, o *z.test* exige que o desvio padrão seja informado. Assim, teremos que calculá-lo no RStudio também.

```
> sda = sd(dados$a[1:4])
> sda
[1] 6.629346
> z.test(dados$a[1:4], sd=sda)
   One Sample z-test
data:  dados$a[1:4]
z = 19.018, n = 4.0000, Std. Dev. = 6.6293, Std. Dev. of the sample mean =
3.3147, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 56.54289 69.53617
sample estimates:
mean of dados$a[1:4]
            63.03953
```

Os dados de interesse foram ressaltados no resultado apresentado pelo RStudio. Temos o valor da estatística Z, o número ***n*** de elementos da amostra, o p-valor e o intervalo de confiança. O valor calculado anteriormente foi $IC_A(\mu, 0{,}95) = (56{,}54; 69{,}53)$, o que confere com o resultado do teste. Fazendo o mesmo para as outras amostras, temos:

```
> sdb = sd(dados$b[1:6])
> sdb
[1] 10.54357
> z.test(dados$b[1:6], sd=sdb)

   One Sample z-test

data:  dados$b[1:6]
z = 15.586, n = 6.0000, Std. Dev. = 10.5436, Std. Dev. of the sample mean
= 4.3044, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 58.65280 75.52572
sample estimates:
mean of dados$b[1:6]
            67.08926
```

O valor calculado para o intervalo de confiança foi $IC_B(\mu, 0{,}95) = (58{,}65; 75{,}52)$

```
> sdc = sd(dados$c[1:8])
> sdc
[1] 6.258621
> z.test(dados$c[1:8], sd=sdc)

   One Sample z-test

data:  dados$c[1:8]
z = 40.315, n = 8.0000, Std. Dev. = 6.2586, Std. Dev. of the sample mean =
2.2128, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 84.87045 93.54430
sample estimates:
mean of dados$c[1:8]
            89.20737
```

O valor calculado para o intervalo de confiança foi $IC_C(\mu, 0{,}95) = (84{,}86; 93{,}54)$

```
> sdd = sd(dados$d)
> sdd
[1] 8.218424
> z.test(dados$d, sd=sdd)

   One Sample z-test

data:  dados$d
z = 34.993, n = 10.0000, Std. Dev. = 8.2184, Std. Dev. of the sample mean
= 2.5989, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 85.84963 96.03711
sample estimates:
mean of dados$d
       90.94337
```

O valor calculado para o intervalo de confiança foi $IC_D(\mu, 0{,}95) = (85{,}84; 96{,}03)$

Da mesma forma que no Exemplo 12, podemos repetir os cálculos supondo que a variância populacional é desconhecida. Assim, sem informações sobre a população, usaremos o t.test.

```
> t.test(dados$a[1:4])

   One Sample t-test

data:  dados$a[1:4]
t = 19.018, df = 3, p-value = 0.0003174
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 52.49076 73.58830
sample estimates:
mean of x
 63.03953
```

O valor calculado para o intervalo de confiança foi $IC_{A(\alpha/2,3)} = (52{,}49; 73{,}59)$

```
> t.test(dados$b[1:6])

   One Sample t-test

data:  dados$b[1:6]
t = 15.586, df = 5, p-value = 1.975e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 56.02446 78.15406
sample estimates:
mean of x
 67.08926
```

O valor calculado para o intervalo de confiança foi $IC_{B(\alpha/2,3)} = (56{,}02; 78{,}16)$

```
> t.test(dados$c[1:8])

   One Sample t-test

data:  dados$c[1:8]
t = 40.315, df = 7, p-value = 1.506e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 83.97504 94.43971
```

```
sample estimates:
mean of x
 89.20737
```

O valor calculado para o intervalo de confiança foi $IC_{C(\alpha/2,3)} = (83{,}97; 94{,}44)$

```
> t.test(dados$d[1:10])

   One Sample t-test

data:  dados$d[1:10]
t = 34.993, df = 9, p-value = 6.281e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 85.06426 96.82248
sample estimates:
mean of x
 90.94337
```

O valor calculado para o intervalo de confiança foi $IC_{D(\alpha/2,3)} = (85{,}07; 96{,}82)$

Os valores dos intervalos de confiança obtidos a partir da distribuição normal e da distribuição de t-Student são exibidos na Tabela 38.

Amostra	Elementos	Dist.Normal	p-valor	Dist. t.Student	p-valor
$IC_A(\mu, 0{,}95)$	4	$(56{,}54; 69{,}53)$	2.2e-16	$(52{,}49; 73{,}59)$	3,174e-04
$IC_B(\mu, 0{,}95)$	6	$(58{,}65; 75{,}52)$	2.2e-16	$(56{,}02; 78{,}16)$	1,975e-05
$IC_C(\mu, 0{,}95)$	8	$(84{,}86; 93{,}54)$	2.2e-16	$(83{,}97; 94{,}44)$	1,506e-09
$IC_D(\mu, 0{,}95)$	10	$(85{,}84; 96{,}03)$	2.2e-16	$(85{,}07; 96{,}82)$	6,281e-11

Tabela 38 - Comparação dos IC´s obtidos com base na distribuição normal e de t-Student

Como pode ser visualizado na Tabela 38, os intervalos de confiança calculados com base na distribuição normal são menores que quando calculados com a distribuição de t-Student (como seria esperado). A medida que o número de elementos na amostra aumenta, o tamanho do intervalo de confiança diminui (para ambas as distribuições), sendo que quanto maior a quantidade de elementos da amostra, mais o intervalo de confiança calculado pela distribuição t-Student se aproxima do calculado pela distribuição normal.

Em relação ao p-valor, temos o mesmo p-valor para os cálculos do intervalo de confiança calculados com a distribuição normal, uma vez a curva da distribuição normal é única (em teoria, igual a curva da distribuição de t-Student com graus de liberdade tendendo ao infinito) e o aumento da quantidade de elementos reflete na diminuição do intervalo em torno da média.

Já para a distribuição de t-Student, o aumento da quantidade de elementos da amostra altera os graus de liberdade (número de elementos da amostra – 1) e, consequentemente, a curva da distribuição. Assim, o aumento da quantidade de elementos da amostra diminui tanto o intervalo de confiança em torno da média quanto o p-valor.

O Poder do teste no RStudio

A função que permite o cálculo do poder do teste pertence ao pacote "**TeachingDemos**". Ela depende da quantidade de elementos da amostra, do nível de significância e a diferença entre o valor real e o valor suposto para o teste. Aplicando os dados do Exemplo 19 onde temos:

- N = 30;
- Nível de significância α de 0,05
- Desvio populacional de 0,22
- Diferença entre o valor real e o valor suposto = 8,09 – 8,0 = 0,09

```
> power.t.test(n = 30, delta = 0.09, sd = 0.22, sig.level = 0.05,power = NULL, type
= "one.sample",alternative = "two.sided",strict = TRUE)

     One-sample t test power calculation

              n = 30
          delta = 0.09
             sd = 0.22
      sig.level = 0.05
          power = 0.5816798
    alternative = two.sided
```

O teste retorna o poder do teste como sendo de 58,17% (erros de arredondamento justificam a diferença entre este valor e o calculado de 61,05%).

8 ANÁLISE DE VARIÂNCIA (ANOVA)

No capítulo anterior, a Inferência estatística, foram analisados casos de estimação e testes de hipóteses. Foi o caso dos testes de comparação de médias baseadas na distribuição normal (teste z) e distribuição de t-Student (teste *t*).

Assim, nos exemplos vistos, analisamos a variação da resistência a compressão (característica de interesse) de amostras criadas com e sem a adição de resíduos de construção e demolição, RCD's (exemplo 16); de amostras criadas com e sem o uso de aditivos (exemplo 17); e oriundas de concreteiras diferentes (exemplo 20). Em cada um destes exemplos, temos um fator (respectivamente, RCD, aditivo e concreteiras) e o fator possui dois níveis (com e sem RCD, com e sem aditivo, concreteiras A e B). Análises envolvendo inferência entre uma ou duas amostras e um fator podem ser chamados de problemas de um único fator com dois níveis ($k = 2$).

Agora, se diferentes situações tivessem que ser analisados no mesmo experimento, como a comparação da resistência a compressão do concreto produzido por mais de duas concreteiras ou experimentos envolvendo a análise de amostras com diversos percentuais de substituição de agregados por RCD, o experimento envolveria um fator (concreteiras ou RCD respectivamente) com mais de dois níveis (quantidade de concreteiras ou os diferentes percentuais de adição de RCD).

Em experimentos de um fator com mais de dois níveis ($k > 2$) é assumido que é necessário K tratamentos (amostras), cada um com populações de N elementos. Por exemplo, se a substituição de agregado grosso por RCD fosse testada em cinco percentuais diferentes (0%, 25%, 50%, 75% e 100%) teríamos cinco níveis ($k = 5$) e seriam necessárias cinco amostras (tratamentos) de N elementos, uma para cada um dos cinco níveis.

Comparar os resultados das cinco amostras pelos métodos já vistos (que permitem comparar duas amostras) seria trabalhoso e pouco prático. É neste ponto que entra a Análise de Variância ou ANOVA. A análise de variância é um modelo estatístico usado para comparar a distribuição de três ou mais grupos de amostras independentes.

Também podemos entende-la como um conjunto de modelos estatísticos nos quais a variância amostral é fracionada em componentes associados aos diferentes fatores (variáveis) de um experimento, sendo que estes fatores que podem estar relacionados à característica de interesse (resultado) do processo, produto ou serviço, objeto de estudo do experimento. Por meio desse fracionamento a análise de variância estuda a influência dos fatores na característica de interesse.

A definição acima nos mostra que a ANOVA não somente se aplica a experimentos de um fator com vários níveis, mas também é capaz de analisar vários fatores, cada um em diferentes níveis. Além disto, a ANOVA é capaz de identificar a influência que um fator exerce em outro fator (interação), mas primeiro, vamos conhecer a análise de variância com um único fator.

8.1 ANOVA – Um Fator

Um procedimento de análise de variância possui como pressupostos as seguintes suposições:

- As observações são independentes, ou seja, cada mensuração da característica de interesse de um elemento da amostra deve ser independente;
- As amostras possuem a mesma variância populacional;
- Os erros (variações entre uma mensuração e a média da amostra) são independentes e provenientes de uma distribuição normal padrão com média igual a zero e variância constante.

Isto porque, é claro, existem variações entre as mensurações e entre as médias das amostras. Estas variações podem ser divididas em dois grupos: (i) variações entre as mensurações de uma amostra; e (ii) variações entre as médias das amostras.

As variações entre as mensurações de uma amostra podem ser produzidas por diversos fatos, tais como, diferenças de temperatura ou umidade no momento do preparo da amostra, preparo da amostra por pesquisadores diferentes, heterogeneidade nas matérias primas empregadas, mensuração da característica de interesse por diferentes equipamentos ou em momentos diferentes, dentre muitas outras.

Em qualquer proporção, a variação observada entre as mensurações deve ser considerada ou como uma variável aleatória ou como fruto do acaso. É parte da função da análise de variância determinar se essa variação observada são as que esperaríamos ter em função do acaso ou se alguma variável foi provavelmente negligenciada.

As variações entre as médias das amostras (ou tratamentos) são o objeto do estudo. A função da análise de variância é esta: verificar se os níveis do fator (ou dos fatores) envolvidos no experimento são os responsáveis pelas variações da média encontradas nas amostras. Isto remete a própria definição da ANOVA: estudar a influência dos fatores na característica de interesse.

Contextualizando a Aplicação da ANOVA

Exemplo 21: Uma empresa fabricante de cimento está testando aditivos para melhoria da resistência mecânica do concreto, com o objetivo de incorporá-lo ao cimento. Decidiu-se testar, com nível de significância de 0,05, cinco aditivos diferentes na proporção recomendada (tratamento) e seis amostras aleatórias de cada tratamento foram selecionadas para preparo e teste, gerando um total de 30 elementos a serem testados. Os dados obtidos estão registrados na Tabela 39.

n / tratamento	1	2	3	4	5
1	42,80	41,25	47,49	49,33	43,93
2	56,68	44,76	44,72	55,58	45,76
3	48,70	45,24	44,02	46,33	43,07
4	41,84	45,09	53,36	48,92	50,36
5	37,62	36,83	48,63	50,40	46,70
6	46,42	36,27	54,28	51,07	41,88
média	45,68	41,57	48,75	50,27	45,28
desvio	6,61	4,16	4,29	3,07	3,04

Tabela 39 - Resistência mecânica dos tratamentos com aditivos

Agora, como comparar os resultados (observações) produzidos por cada um dos aditivos? A preparação dos elementos pode ter influenciado algum resultado (maior temperatura ou pequenas diferenças na dosagem dos insumos)? Para verificar se a resistência mecânica realmente variou em função do tipo de aditivo devemos utilizar um teste estatístico que além de considerar as médias dos tratamentos, também leve em conta a variação da resistência dentro de cada tratamento.

Em primeiro lugar, vamos representar estes dados sob a forma de gráfico de boxplot usando o RStudio (Figura 62). Por meio do gráfico, podemos comparar a distribuição dos valores de cada amostra, mas isto não nos confirma se são iguais ou diferentes.

Para descobrirmos se os aditivos influenciam na resistência mecânica do concreto precisamos de análises estatísticas mais complexas, tais como a análise de variância, que veremos a seguir.

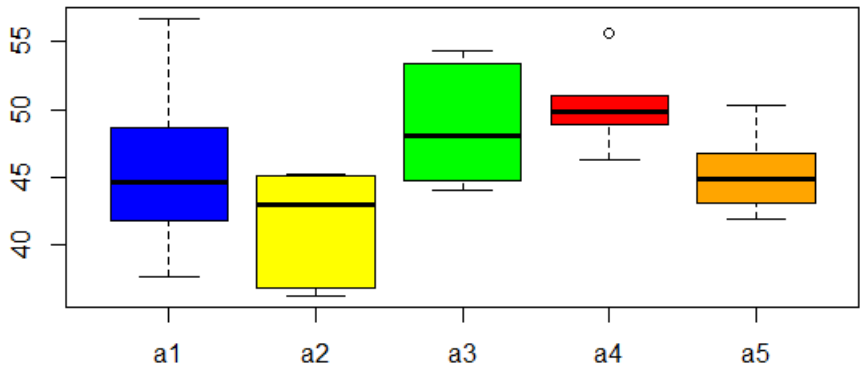

Figura 62 - BoxPlot dos dados

A Análise de Variância Simples

Para a análise de variância assume-se que as *K* populações são independentes e normalmente distribuídas com médias $\mu_1, \mu_2, \ldots, \mu_K$ e variância comum σ^2. Isto pode ser assumido desde que a aleatorização seja critério para o experimento, garantindo uma distribuição uniforme do erro experimental por todo o tratamento.

As hipóteses padrões para o teste são:

$$H0: \mu_1 = \mu_2 = \ldots = \mu_K$$

$$H1: pelo\ menos\ duas\ médias\ não\ são\ iguais$$

Vamos denotar como y_{ij} como a j-ésima observação do i-ésimo tratamento e vamos organizar os dados como mostrado na Tabela 40. Nela, Y_i é a soma das observações na amostra do i-ésimo tratamento, $\overline{y}_i$ é a média das observações do i-ésimo tratamento, $Y_{..}$ é a soma de todas as *nk* observações e $\overline{y}_{..}$ é a média de todas as *nk* observações.

Tratamento	**1**	**2**	**...**	***i***	**...**	***k***	
	y_{11}	y_{21}	...	y_{i1}	...	y_{k1}	
	y_{12}	y_{22}	...	y_{i2}	...	y_{k2}	
	...	...	...	...	...	...	
	y_{1n}	y_{2n}	...	y_{in}	...	y_{kn}	
Total	Y_1	Y_2	...	Y_i	...	Y_k	$Y_{..}$
Média	$\overline{y}_1$	$\overline{y}_2$	...	$\overline{y}_i$	...	$\overline{y}_k$	$\overline{y}_{..}$

Tabela 40 - Amostras aleatórias do experimento

Onde cada observação pode ser escrita da forma:

$$y_{ij} = \overline{y}_i + \epsilon_{ij} \qquad \textit{Eq. 52}$$

Onde ϵ_{ij} mede o desvio da j-ésima observação da i-ésima média amostral do tratamento correspondente. O termo ϵ_{ij} representa o erro aleatório. Da mesma forma, considerando-se que as médias de cada tratamento desviam-se da média geral $\overline{y}_{..}$ devido à influência deste tratamento (*i*) e denotando α_i como o efeito do i-ésimo tratamento, podemos reescrever a fórmula acima como:

$$y_{ij} = \overline{y}_{..} + \alpha_i + \epsilon_{ij} \qquad \textit{Eq. 53}$$

Desta forma, a hipótese nula de que todas as k médias são iguais e a hipótese alternativa de que pelo menos duas das médias são diferentes pode ser escrita como:

$H0: \alpha_1 = \alpha_2 = \ \ldots = \ \alpha_K = 0$

$H1: pelo\ menos\ dos\ \alpha_i\ não\ é\ igual\ a\ zero$

O teste da análise de variância é baseado na comparação de duas estimativas independentes da variância populacional σ^2, dada pela equação:

$$\sigma^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{..}\right)^2 \qquad \textit{Eq. 54}$$

Estas duas estimativas independentes são obtidas dividindo-se a variabilidade total dos dados em dois componentes:

$$\sigma^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{y}_{i} - \overline{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{i}\right)^2 \qquad \textit{Eq. 55}$$

Ou simplesmente: $SQT = SQA + SQE$ de onde passaremos a denotar:

$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{..}\right)^2$	soma dos quadrados total, responsável por medir a variabilidade total dos dados
$SQA = \sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{y}_{i} - \overline{y}_{..})^2$	soma dos quadrados dos desvios dos tratamentos, é o desvio das médias estimadas em cada tratamento em torno da média geral dos dados e representa a variabilidade devido ao tratamento
$SQE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{i}\right)^2$	Soma dos quadrados dos erros, é o desvio das observações em torno da média estimada do seu tratamento e representa a variabilidade de das observações dentro do tratamento

Uma equação alternativa para SQA é mostrada a seguir. A segunda somatória é substituída por uma multiplicação, uma vez que o termo da somatória não varia em função de *n*:

$$SQA = n\sum_{i=1}^{k}(\overline{y}_{i} - \overline{y}_{..})^2 \qquad \textit{Eq. 56}$$

Como citado anteriormente, estamos tratando de estimativas independentes da variância populacional σ^2. Suposto que a variância amostral pode ser obtida dividindo-se SQT pelos seus graus de liberdade (*n – 1*), o mesmo pode ser realizado com seus componentes para se obter as duas estimativas independentes:

$$s_1^2 = \frac{SQA}{k-1} \ \ e \ \ s^2 = \frac{SQE}{k(n-1)} \qquad \textit{Eq. 57}$$

Assim temos que s_1^2 é uma estimativa não viciada de σ^2, pois se ***H0*** for verdadeira, a somatória dos α_i será zero o que faz $s_1^2 = \sigma^2$. Entretanto, se ***H1*** for verdadeira, s_1^2 estima σ^2 e mais um termo adicional, que mensura a variação devido a efeitos sistemáticos. Desta forma, quando ***H0*** é falsa, s_1^2 superestima σ^2 $(s_1^2 > \sigma^2)$.

Já a estimativa s^2 é uma estimativa não viciada, independente da verdade ou da falsidade da hipótese nula. Disto decorre que a razão entre s_1^2 e s^2, denotada razão *f* pode ser usada para avaliar a igualdade das médias.

Desta forma, a razão $f = s_1^2/s^2$ é um valor da variável aleatória F com $k - 1$ e $k(n - 1)$ graus de liberdade. Assim, temos que a hipótese nula é rejeitada no nível de significância α quando:

$$f_c > f_\alpha[k-1, k(n-1)] \qquad \textit{Eq. 58}$$

A Tabela 41 resume a análise de variância ANOVA simples e a Figura 63 apresenta a tabela F com os valores críticos com $k - 1$ e $k(n - 1)$ graus de liberdade.

Fonte da variação	Soma dos Quadrados	Graus de liberdade	Quadrado médio	F calculado
Tratamento	SQA	*k - 1*	$s_1^2 = \frac{SQA}{k-1}$	$f_c = s_1^2/s^2$
Erro	SQE	*k (n – 1)*	$s^2 = \frac{SQE}{k(n-1)}$	
Total	SQT	*kn – 1*		

Tabela 41 - Análise de variância ANOVA simples

	ν_1										
ν_2	1	2	3	4	5	6	7	8	9	10	11
1	161,45	199,50	215,70	224,58	230,16	234,0	236,8	238,9	240,5	241,9	242,98
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40
3	10,13	9,55	9,27	9,11	9,01	8,94	8,88	8,84	8,81	8,78	8,76
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46

Figura 63 - Distribuição F com v1 graus de liberdade do numerador e v2 graus de liberdade do numerador para alpha = 0,05

Retornando ao Exemplo 21, vamos realizar os cálculos do ANOVA em uma planilha MS Excel, para acompanhamento do processo:

	A1	A2	A3	A4	A5	
1	42,80	41,25	47,49	49,33	43,93	
2	56,68	44,76	44,72	55,58	45,76	
3	48,70	45,24	44,02	46,33	43,07	
4	41,84	45,09	53,36	48,92	50,36	
5	37,62	36,83	48,63	50,40	46,70	
6	46,42	36,27	54,28	51,07	41,88	
média	**45,68**	**41,57**	**48,75**	**50,27**	**45,28**	**46,31**

Tabela 42 - Dados para o cálculo do ANOVA

Parte 1: Cálculo do SQT, onde cada célula contém o valor $\left(y_{ij} - \bar{y}_{..}\right)^2$.

$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2$	12,33	25,61	1,39	9,11	5,67
	107,52	2,41	2,53	85,91	0,30
	5,71	1,15	5,25	0,00	10,50
	19,99	1,49	49,69	6,81	16,39
	75,53	89,89	5,38	16,72	0,15
	0,01	100,82	63,50	22,65	19,63
SQT =	**764,06**				

Parte 2: Cálculo do SQA, onde cada célula contém o valor $(\bar{y}_i - \bar{y}_{..})^2$.

$SQA = n\sum_{i=1}^{k}(\bar{y}_i - \bar{y}_{..})^2$	0,40	22,45	5,95	15,69	1,06
SQA =	**273,24**				

Parte 3: Cálculo do SQE, onde cada célula contém o valor $\left(y_{ij} - \bar{y}_i\right)^2$.

$SQE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2$	8,28	0,10	1,59	0,89	1,83
	121,07	10,15	16,24	28,18	0,23
	9,14	13,44	22,37	15,54	4,90
	14,72	12,37	21,25	1,83	25,77
	64,91	22,50	0,01	0,02	2,01
	0,55	28,13	30,58	0,64	11,58
SQE =	**490,82**				

Parte 4: Tabela resumo da ANOVA:

Fonte da variação	Soma dos Quadrados	Graus de liberdade	Quadrado médio	F calculado	F(0,05, 4, 25)
Tratamento	273,24	*4*	68,31	**3,48**	**2,76**
Erro	490,82	*25*	19,63		
Total	764,06	*29*			**p-valor=0,0216**[19]

Conclusão: Como $f_c > f_\alpha[k-1, k(n-1)]$, isto é **3,48 > 2,76** temos que, com nível de significância igual a 0,05, podemos rejeitar a hipótese nula (igualdade das médias). Assim, temos constatação estatística que pelo menos duas das médias são diferentes.

Bom, agora a pergunta: e é só? Pelo menos duas das médias são diferentes e o que isto significa? O principal uso da análise de variância não é apenas a comparação de médias, mas a análise da significância do tratamento nos resultados do experimento. Se pelo menos duas das médias são diferentes significa que os aditivos influenciam, de forma diferente, na resistência mecânica do concreto, ou seja, eles influenciam os resultados.

Ainda não sabemos quais são as médias diferentes nem se a maior delas difere das outras (afinal, procuramos o melhor tratamento). Para isto são necessários outros testes de comparação de médias, como o teste de

[19] Calculado com a função DISTF(F calculado; Graus de liberdade do tratamento; Graus de liberdade do erro) do MS Excel DISTF(3,48; 4; 25)

Tukey ou o teste de Ducan que serão apresentados após o ANOVA. O teste *t* também pode ser usado para complementar a análise. Com o uso do teste *t* cada uma das médias dos tratamentos pode ser comparada com a média geral ou entre elas mesmas.

A ANOVA é a principal forma de avaliar, estatisticamente, a influência de um tratamento nos resultados de um experimento. No texto anterior, foi apresentada a ANOVA de um fator com *k* níveis. O mesmo raciocínio pode ser aplicado para experimentos com diversos fatores, cada um deles com número de níveis diferentes.

8.2 ANOVA – Dois Fatores

Na grande maioria dos experimentos, estamos interessados em avaliar a influência que dois ou mais fatores podem exercer sobre a característica de interesse (resposta). Quando o experimento envolve dois fatores em diferentes níveis, diz-se que temos uma ANOVA de dois fatores ou ANOVA *two way*. Se envolve mais de dois fatores é chamada de ANOVA Fatorial.

Um fato interessante a ser notado é que, quando temos mais de um fator, existe sempre a possibilidade da influência mútua entre os fatores, ou seja, a possibilidade de interação entre os fatores do experimento. Assim, além da influência que cada fator exerce sobre a característica de interesse (fato que pode ser identificado por meio da ANOVA de um fator), a análise de variância deve considerar a possibilidade que um dos fatores atue como catalizador ou bloqueador da influência do outro fator.

Exemplo 22: Consideremos o seguinte experimento. Um pesquisador deseja avaliar o impacto da substituição parcial de dois insumos na produção de cimento. Para tanto, realizou um experimento de dois fatores (A e B) cada um com dois níveis de substituição, representados por "+" e "-" e, para cada tratamento, foram elaborados 4 elementos para teste. O resultado é apresentado na Tabela 43.

		A			
		-		+	
B	-	58,97	57,65	67,12	65,16
		65,10	66,33	64,92	66,80
	+	71,92	66,89	75,34	73,20
		68,27	68,94	75,20	72,15

Tabela 43 - Estudo de interação entre fatores

Interação entre Fatores

A interação entre os fatores corresponde à diferença de comportamento de um fator (fator A, por exemplo) nos diferentes níveis do outro fator (fator B) com respeito a característica de interesse (resposta). Uma das primeiras e mais simples formas de avaliação da interação entre os fatores são o gráfico de interação e o gráfico dos efeitos principais.

Gráfico de interação: O gráfico de interação é montado a partir das médias amostrais dos fatores agrupados em seus níveis. Para entendermos o processo, vamos resumir o quadro apresentado na Tabela 43 substituindo os valores dos elementos das amostras pela média amostral (Tabela 44).

		A	
		-	+
B	-	62,01	66,00
	+	69,00	73,97

Tabela 44 - Efeito dos Fatores – média amostral

O gráfico é montado com os níveis do fator A no eixo x e o fator B como resposta (ou ao contrário). Nele analisamos se a diferença na resposta entre os níveis de um fator não é a mesma em todos os níveis dos outros fatores. Quando isto ocorre há uma interação entre os fatores.

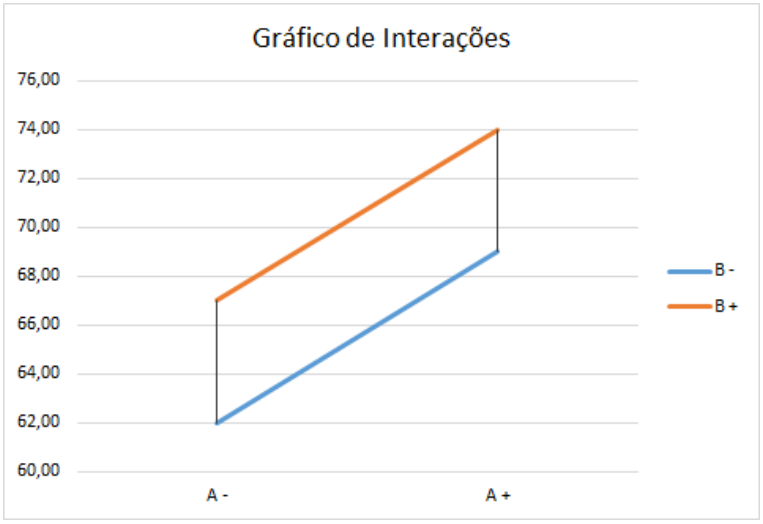

Figura 64 - Análise das interações entre os fatores

Ao analisarmos o gráfico da Figura 64 vemos que a diferença entre os níveis do fator B é a mesma para A+ e A-, indicando não haver interação entre os fatores.

No caso de não haver interações, podemos interpretar o gráfico dos efeitos principais. O gráfico dos efeitos principais é montado com as médias de cada fator em cada nível, a exemplo Tabela 45. O gráfico correspondente é exibido na Figura 65.

	-	+
A	65,51	69,99
B	64,01	71,49

Tabela 45 - Dados para gráfico dos efeitos principais – médias amostrais

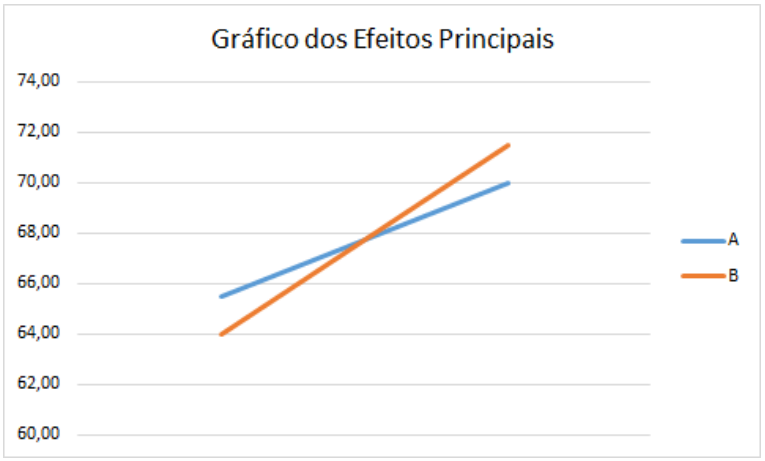

Figura 65 - Gráfico dos efeitos principais

A análise do gráfico (Figura 65) nos mostra que ambos os fatores impactam na característica de interesse, embora o fator B possa possuir maior contribuição para o resultado. Neste exemplo, foi analisado um experimento de dois fatores, cada um com dois níveis, ambos com influência positiva na característica de interesse e sem interação entre os fatores.

Exemplo 23: Agora, vamos analisar uma nova situação. Novamente um experimento com dois fatores de dois níveis, um com influência positiva e outro com influência negativa na característica de interesse e sem interação entre os fatores. A Tabela 46 apresenta os valores das médias amostrais de cada tratamento, para facilitar a montagem dos gráficos

		A	
		-	+
B	-	46,28	54,00
	+	41,79	49,54

Tabela 46 - Efeito dos Fatores (influências positiva e negativa dos fatores)

O gráfico de interação (Figura 66) foi novamente montado com o fator A no eixo x. Nele podemos visualizar que a diferença entre os níveis do fator B é a mesma para A+ e A-, indicando não haver interação entre os fatores. Note que o gráfico é montado a partir das médias amostrais, que podem possuir desvios em relação as médias populacionais. Assim, pequenas diferenças entre os níveis dos fatores são admissíveis.

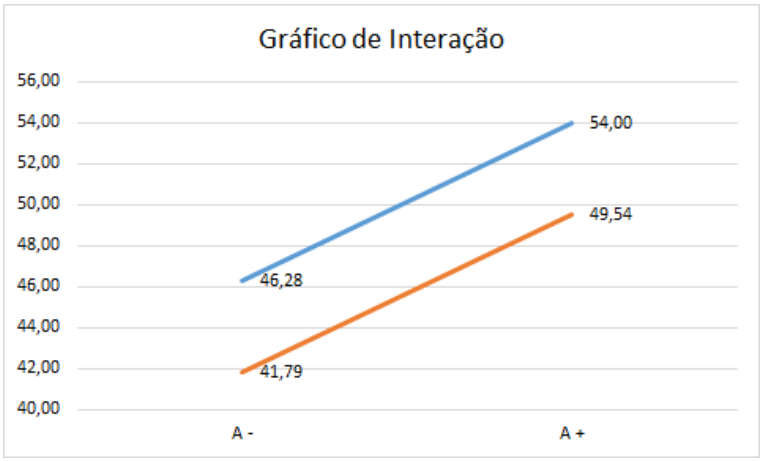

Figura 66 - Gráfico de Interação com fatores com influências diferentes na característica de interesse

Já o gráfico dos efeitos principais (Figura 67), baseado nos dados da Tabela 47, mostra a contribuição de cada fator para a característica de interesse. Nele é possível ver claramente que o fator A possui influência positiva na característica de interesse enquanto o fator B possui influência negativa.

	-	+
A	44,04	50,14
B	51,77	45,67

Tabela 47 - Dados para gráfico dos efeitos principais

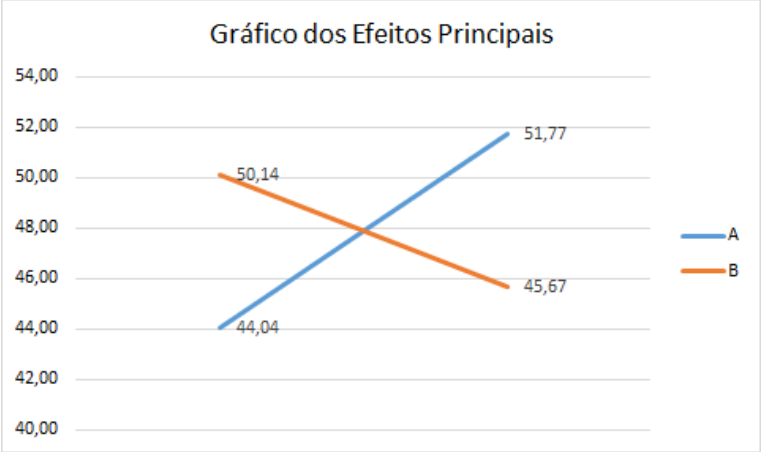

Figura 67 - Gráfico dos Efeitos Principais - Influências positiva e negativa na característica de interesse

Exemplo 24: E, por último, antes de iniciarmos o ANOVA propriamente dito, vamos analisar um terceiro exemplo. Um experimento com dois fatores de dois níveis, ambos com influência na característica de interesse e com interação entre os fatores (que para ambos pode ser positiva ou negativa). A Tabela 48 apresenta os valores das médias amostrais de cada tratamento. Neste caso, haverá interação entre os fatores.

		A	
		-	+
B	-	31,48	49,15
	+	31,87	42,96

Tabela 48 - Efeito dos Fatores (influências positiva e interação entre os fatores)

O gráfico de interação (Figura 68) foi montado com o fator A no eixo x. Nele podemos visualizar que a diferença entre os níveis do fator B **não** é a mesma para A+ e A-, indicando haver interação entre os fatores.

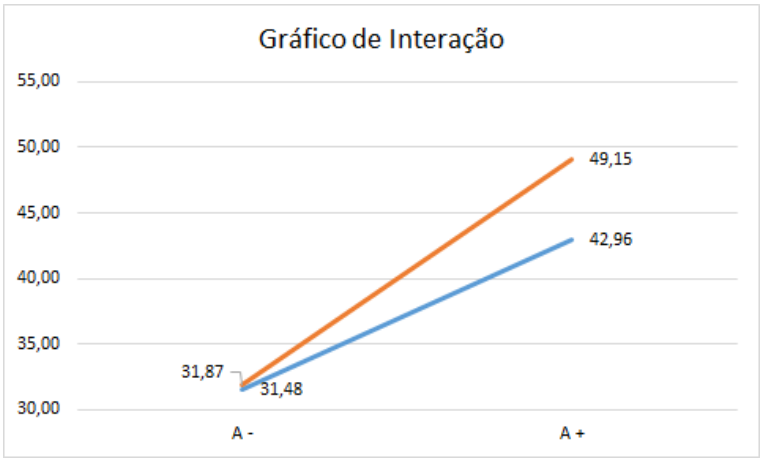

Figura 68 - Gráfico de interação - influência da interação entre os fatores

Os dados para a montagem do gráfico dos efeitos principais é mostrado na Tabela 49 e o gráfico é mostrado na Figura 69.

	-	+
A	31,68	37,22
B	46,06	40,51

Tabela 49 - Dados para gráfico dos efeitos principais

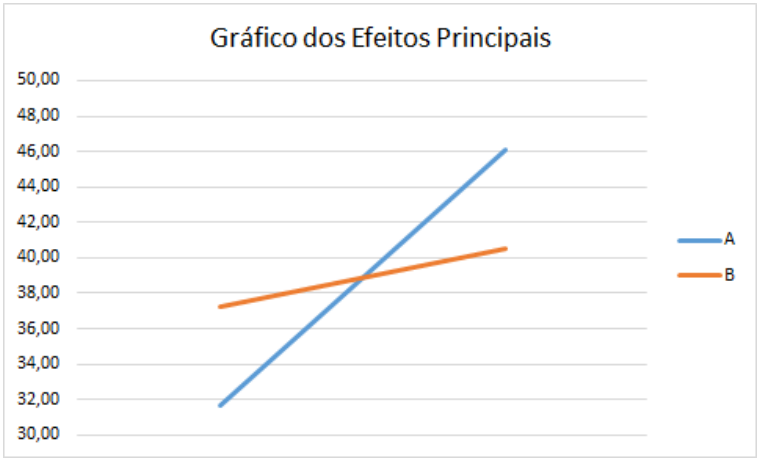

Figura 69 - Gráfico dos efeitos principais com interação entre os fatores

Pela análise do gráfico (Figura 69), poderíamos supor que ambos os fatores possuem efeito positivo na característica de interesse, mas como foi dito anteriormente, existe interação entre os fatores, ou seja, uma nova influência foi estabelecida e o gráfico indica apenas os efeitos principais. Assim, pode ser que a interação entre eles esteja ocultando a real influência de um dos fatores. Para analisarmos a influência de cada um dos fatores e da interação entre eles, precisamos da ANOVA.

Modelo da ANOVA – Dois Fatores

Consideremos um experimento com dois fatores A e B, no qual o fator A tem *a* níveis e o fator B tem *b* níveis. Para cada combinação de níveis, temos *n* elementos. Na Tabela 50, apresentamos os dados do experimento:

Fator A	Fator B				Média
	1	2	...	b	
1	$y_{111}, \cdots, y_{11n}$	$y_{121}, \cdots, y_{12n}$	...	$y_{1b1}, \cdots, y_{1bn}$	$\overline{y_{1..}}$
2	$y_{211}, \cdots, y_{21n}$	$y_{221}, \cdots, y_{22n}$	...	$y_{2b1}, \cdots, y_{1bn}$	$\overline{y_{2..}}$
...	...	...	...	...	
a	$y_{a11}, \cdots, y_{a1n}$	$y_{a21}, \cdots, y_{a2n}$	...	$y_{ab1}, \cdots, y_{abn}$	$\overline{y_{a.}}$
Média	$\overline{y_{.1.}}$	$\overline{y_{.2.}}$	...	$\overline{y_{.b.}}$	$\overline{y_{...}}$

Tabela 50 - Dados para ANOVA de dois fatores

Da mesma forma que na ANOVA de um fator, cada observação pode ser descrita da forma:

$$y_{ijk} = \bar{y}_{...} + \epsilon_{ijk} \qquad \textit{Eq. 59}$$

Onde ϵ_{ijk} mede os desvios dos valores dos elementos y_{ijk} da média da população $\overline{y_{...}}$. Ainda, repetindo o raciocínio empregado no ANOVA de um fator, podemos considerar que o valor de cada elemento desvia-se da média geral $y_{...}$ devido à: (i) influência do efeito do nível *i* do fator A, denotando como α_i; (ii) influência do efeito do nível *j* fator B, denotado β_j; e (ii) influência da possível interação *ij* dos fatores A e B, denotada $\alpha\beta_{ij}$, e assim reescrever a fórmula acima como:

$$y_{ijk} = \bar{y}_{...} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \qquad \textit{Eq. 60}$$

Na qual temos que impor as seguintes restrições:

$$\sum_{i=1}^{a} \alpha_i = 0, \sum_{j=1}^{b} \beta_j = 0, \sum_{i=1}^{a} (\alpha\beta)_i = 0, \sum_{j=1}^{b} (\alpha\beta)_j = 0 \qquad \textit{Eq. 61}$$

Como citado anteriormente, em um experimento com dois fatores, precisamos avaliar se existe interação entre os fatores. O gráfico de interação nos mostra evidências da existência de interação. O ANOVA avalia o efeito da interação por meio de um teste de hipóteses. Caso o efeito da interação não seja significativo, O ANOVA avalia os efeitos principais (individuais), também por meio de testes de hipóteses apropriados. Os testes de hipóteses são apresentados a seguir.

Objetivo	Hipótese
Efeito do Fator A	H_0': $\alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_a = 0$ H_1': $Pelo\ menos\ um\ dos\ \alpha_i$ é $diferente\ de\ zero$
Efeito do Fator B	H_0'': $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_b = 0$ H_1'': $Pelo\ menos\ um\ dos\ \beta_j$ é $diferente\ de\ zero$
Efeito da Interação A B	H'''_0: $\alpha\beta_{11} = \alpha\beta_{12} = \alpha\beta_{13} = \cdots = \alpha\beta_{ab} = 0$ H'''_1: $Pelo\ menos\ um\ dos\ \alpha\beta_{ij}$ é $diferente\ de\ zero$

Alertamos para o fato de que, caso a interação tenha grande influência sobre a característica de interesse, ela pode mascarar os efeitos dos fatores principais. Por isto é recomendável que a análise da interação seja realizada primeiro. Caso seja constatado que a interação entre os fatores é desprezível, as hipóteses 1 e 2 podem ser testadas e a interpretação é simples. Caso a interação seja significativa, a análise pode se tornar mais complexa.

Da mesma forma que no ANOVA de um fator, vamos decompor a variabilidade total dos dados σ^2, denotada "soma dos quadrados" em quatro componentes, tais que:

$$\sigma^2 = SQT = SQA + SQB + SQAB + SQE \qquad \textit{Eq. 62}$$

De onde passaremos a denotar:

$SQT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \overline{y_{...}}\right)^2$ soma dos quadrados total, responsável por medir a variabilidade total dos dados

$SQA = bn\sum_{i=1}^{a}(\overline{y_{i..}} - \overline{y_{...}})^2$ soma dos quadrados do tratamento A, é o desvio das médias estimadas em no tratamento A em torno de sua média geral e representa a variabilidade devido ao tratamento A

$SQB = an\sum_{j=1}^{b}\left(\overline{y_{.J.}} - \overline{y_{...}}\right)^2$ soma dos quadrados do tratamento B, é o desvio das médias estimadas em no tratamento B em torno de sua média geral e representa a variabilidade devido ao tratamento B

$SQAB = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(y_{ij.} - \overline{y_{i..}} - \overline{y_{.J.}} + \overline{y_{...}}\right)^2$ Soma dos quadrados da interação AB

$SQE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \overline{y_{...}}\right)^2$ Soma dos quadrados dos erros, é o desvio das observações em torno da média estimada e representa a variabilidade de das observações

Como estamos tratando de estimativas independentes da variância populacional σ^2, podemos supor que a variância amostral pode ser obtida dividindo-se SQT pelos seus graus de liberdade (*n – 1*), o mesmo pode ser realizado com seus componentes para se obter as duas estimativas independentes:

$$s_1^2 = \frac{SQA}{a-1}, \qquad s_2^2 = \frac{SQB}{b-1}, \qquad s_{12}^2 = \frac{SQAB}{(a-1)(b-1)} \quad e \quad s^2 = \frac{SQE}{ab(n-1)} = \sigma^2 \qquad \textit{Eq. 63}$$

Assim temos que estas estimativas de variância são estimativas independentes e não viciadas de σ^2 com a condição de que o somatório dos efeitos $\alpha_i,\ \beta_j e\ \alpha\beta_{ij}$ são nulos. Assim, para testar as hipóteses podemos comparar cada desvio com o desvio σ^2, como detalhado a seguir na Tabela 51:

Objetivo		Estimador	Critério
Efeito do Fator A	H_0':	$f_1 = \frac{s_1^2}{s^2}$	$f_1 > f_\alpha[a-1, ab(n-1)]$
Efeito do Fator B	H_0'':	$f_2 = \frac{s_2^2}{s^2}$	$f_2 > f_\alpha[b-1, ab(n-1)]$
Efeito da Interação A B	H'''_0:	$f_3 = \frac{s_{12}^2}{s^2}$	$f_3 > f_\alpha[(a-1)(b-1), ab(n-1)]$

Tabela 51 - Teste e critérios para ANOVA dois fatores

A Tabela 52 apresentada na a seguir resume a análise de variância ANOVA dois fatores

Fonte da variação	Soma dos Quadrados	Graus de liberdade	Quadrado médio	F calculado
Tratamento A	SQA	*a - 1*	$s_1^2 = \frac{SQA}{a-1}$	$f_c = s_1^2/s^2$
Tratamento B	SQB	*b - 1*	$s_2^2 = \frac{SQB}{b-1}$	$f_c = s_2^2/s^2$
Interação AB	SQAB	*(a – 1)(b – 1)*	$s_{12}^2 = \frac{SQAB}{(a-1)(b-1)}$	$f_c = s_{12}^2/s^2$

Erro	SQE	*ab (n – 1)*	$s^2 = \frac{SQE}{ab(n-1)}$
Total	SQT	*abn – 1*	

Tabela 52 - Análise de variância ANOVA dois fatores

Para exemplificar o processo, vamos retomar o **Exemplo 24**, cujos dados foram apresentados na Tabela 48. A análise gráfica nos mostrou que havia interação entre os fatores. Os dados originais são apresentados na Tabela 53:

		B			
		-		+	
A	-	31,11	34,83	31,79	30,92
		29,72	30,29	33,44	31,35
	+	42,61	42,68	51,17	48,21
		41,32	45,24	49,52	47,72

Tabela 53 - Dados para cálculo do ANOVA

Com base nos dados da tabela apresentada acima, foram calculados os valores de SQA, SQB, SQAB, SQE e SQT, apresentados na Tabela 54.

Fonte da variação	Soma dos Quadrados	Graus de liberdade	Quadrado médio	F calculado	Fα	P-valor
Tratamento A	826,95	*1*	826,95	285,73	4,75	$9{,}85 \times 10^{-10}$
Tratamento B	43,32	*1*	43,32	14,97	4,75	0,00223
Interação AB	33,68	*1*	33,68	11,64	2,51	0,00516
Erro	34,73	*12*	2,89			
Total	938,68	*15*				

Tabela 54 - Resultado Anova dois fatores

Análise dos resultados: Em primeiro lugar, podemos verificar que todos os F calculados são superiores ao Fα (para encontrar o Fα, foi utilizado 0,05 como nível de confiança e os graus de liberdade dados pela coluna "critérios" da Tabela 51). Com isto, para todas as hipóteses, a ***H0*** (igualdade) pode ser rejeitada e temos que ambos os fatores A e B e também sua interação são significativos para a característica de interesse.

Outra forma de vermos isto é fornecida pelo p-valor. Em todas as três hipóteses, o p-valor é inferior a 0,05, levando a rejeição da hipótese nula nas três situações.

Bom, sabemos que ambos os fatores e sua interação são significativos, mas como eles influenciam a característica de interesse? O ANOVA não nos dá essa informação. Apenas podemos concluir o que foi expresso acima. Se não houvesse interação, o próprio gráfico dos efeitos principais nos daria a resposta. Mas a interação existe e a análise fica mais complexa.

Para que possam ter uma ideia do significado desta complexidade, vamos apresentar a base que foi utilizada para geração dos dados usados no exemplo:

- A característica de interesse (*X*) foi determinada a partir da equação $x = 25 + 60A - 5B + 55AB$, onde os níveis de A foram (0,1 / 0,2) e os níveis de B foram (1,0 / 2,0). Com isto conseguimos os valores esperados para a média de cada tratamento;
- A partir do valor esperado para a média, foi gerada uma distribuição aleatória de frequência com 4 elementos para compor a amostra, mantendo-se o desvio padrão inferior a 10% do valor esperado para a média. Estes foram as medidas utilizadas para os elementos da amostra.

Este conjunto de passos gerou os dados usados no exemplo. A análise da equação empregada mostra que o efeito do fator B é negativo, ou seja, ele influencia negativamente a característica de interesse. No entanto, o valor da interação AB é positivo e é superior à contribuição do fator B.

O gráfico de interações (Figura 68) mostra que há interação entre os fatores, o que foi confirmado pela ANOVA. O gráfico dos efeitos principais (Figura 69), que analisa apenas os efeitos destes fatores na característica de interesse, foi mascarado pela interação entre os fatores (que é positiva e maior que a influência negativa do fator B). A ANOVA, apesar de extremamente útil, nos informa sobre a significância dos fatores sobre a característica de interesse e não sobre como os fatores atuam sobre esta característica.

Assim, se não temos informação prévia sobre o tipo de contribuição do fator sobre o resultado, informação esta que poderia ser obtida de estudos anteriores (análise da literatura), temos que tomar outras providências que nos auxiliarão a definir o tipo de contribuição, como:

- Pesquisar mais, afinal é difícil encontrar algo que é tão inédito e inovador a ponto de nunca ter sido tentado anteriormente;
- Aumentar o número de níveis nos fatores, incluindo o nível zero (sem a inclusão do fator), para podermos analisar separadamente a influência do fator sobre a característica de interesse;
- Fracionar o experimento, realizando experimentos prévios menores, com o objetivo de descobrir como cada fator contribui para os resultados, quanto temos mais de dois fatores

E também, sempre podemos avançar no estudo da estatística, pois existem outras funções estatísticas que podem nos auxiliar a identificar como cada fator contribui para a característica de interesse, como a Análise de Regressão.

Por enquanto, vamos continuar com a ANOVA e apresentar mais exemplos de sua utilização e importância, desta vez com o auxílio do software RStudio.

8.3 ANOVA e o RStudio

A teoria base da Análise de Variância (ANOVA) já foi apresentada. A partir desse ponto acreditamos ser mais simples e fácil compreender e avaliar a importância da ANOVA a partir da análise de seu uso em experimentos e da forma como ela contribui para o entendimento dos resultados.

Antes é necessário um esclarecimento sobre a função ANOVA no RStudio. A base para os cálculos da análise de variância é uma só, mas as fórmulas variam, como pode ser visto para o ANOVA de um fator e de dois fatores. Quanto maior o número de fatores, mais complexas se tornam as fórmulas. A função ANOVA (função *aov* ou *lm* no RStudio) é uma só e atende a todas as variações. Apenas seus parâmetros irão variar, se usada para um fator, dois fatores ou mais de dois fatores.

Antes de apresentarmos problemas mais complexos, vamos repetir os exemplos anteriores da ANOVA:

Exemplo 21: neste experimento tivemos o teste de cinco aditivos, com amostras de seis elementos. Vamos carregar os dados no RStudio e, depois, com os dados carregados, podemos executar a análise de variância. Vamos designar uma variável para armazenar seus resultados (*dados_an*) e logo após a execução, exibir o resultados (pode ser pelo comando *anova* ou *summary*).

Carga dos dados:

```
> dados = read.csv2(file.choose(), header = T)
> summary(dados)
  a          res
 a1:6   Min.   :36.27
 a2:6   1st Qu.:43.28
 a3:6   Median :46.05
 a4:6   Mean   :46.31
 a5:6   3rd Qu.:49.23
        Max.   :56.68
```

Execução da ANOVA:

```
> dados_an = aov(res~a, data = dados)
> anova(dados_an)

Analysis of Variance Table

Response: res
          Df Sum Sq Mean Sq F value  Pr(>F)
a          4 273.24  68.309  3.4794 0.02165 *
Residuals 25 490.82  19.633
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resgatando aqui o resumo da ANOVA apresentado anteriormente, para fins de comparação, podemos ver que os resultados são os mesmos:

Tabela resumo da ANOVA:

Fonte da variação	Soma dos Quadrados	Graus de liberdade	Quadrado médio	F calculado	F(0,05, 4, 25)
Tratamento	273,24	*4*	68,31	**3,48**	**2,76**
Erro	490,82	*25*	19,63		
Total	764,06	*29*			**p-valor=0,0216**

A função *aov* que executa a ANOVA nos traz como resultado o p-valor (a ser comparado com o nível de significância estabelecido para o teste, sendo que o valor padrão é 0,05). Também apresenta um resumo similar ao quadro estudado anteriormente, com a soma dos quadrados (*Sum sq*), graus de liberdade (*Df*), quadrado médio (*Mean sq*) e o valor da estatística F calculada (*F value*).

A função *lm* também executa a ANOVA da mesma forma. A diferença entre elas é que com o uso da função *lm*, podemos extrair informações mais detalhadas com o uso da função *summary*. Já a função *aov* permite o uso do teste de tukey (comparação múltipla de médias). Abaixo, a execução da ANOVA é repetida com a função *lm*, para conferência dos resultados.

```
> dados_lm = lm(res~a, data=dados)
> anova(dados_lm)
Analysis of Variance Table
```

```
Response: res
          Df Sum Sq Mean Sq F value  Pr(>F)
a          4 273.24  68.309  3.4794 0.02165 *
Residuals 25 490.82  19.633
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Exemplo 22: Neste experimento (ANOVA de dois fatores) foi avaliado o impacto da substituição parcial de dois insumos na produção de cimento. Vamos carregar e exibir os dados no RStudio, antes de executar a análise de variância.

```
> dados = read.csv2(file.choose(), header = T)
> dados
    a  b   res
1  a- b- 58.97
2  a- b- 65.10
3  a- b- 57.65
4  a- b- 66.33
5  a- b+ 71.92
6  a- b+ 68.27
7  a- b+ 66.89
8  a- b+ 68.94
9  a+ b- 67.12
10 a+ b- 64.92
11 a+ b- 65.16
12 a+ b- 66.80
13 a+ b+ 75.34
14 a+ b+ 75.20
15 a+ b+ 73.20
16 a+ b+ 72.15
```

Repare que os dados a serem carregados foram organizados de forma diferente, com os fatores e a característica de interesse organizadas em colunas. Esse é o padrão para a entrada de dados no RStudio (os dados do exemplo anterior também foram carregados neste formato).

Uma outra informação: a função *aov* permite executar a análise de variância com ou sem a análise das interações entre os fatores. A diferença é a forma de entrada dos parâmetros relativos aos fatores e a característica de interesse:

- res ~ a + b indica execução da análise de variância sem a análise da interação e;
- res ~ a * b indica execução da análise de variância com a análise da interação.

Vamos executar primeiramente com a análise da interação:

```
> dados_an = aov(res ~ a * b, data = dados)

> anova(dados_an)

Analysis of Variance Table

Response: res
          Df  Sum Sq Mean Sq F value    Pr(>F)
a          1  80.192  80.192 11.8747  0.004841 **
b          1 223.951 223.951 33.1624 9.039e-05 ***
a:b        1   0.960   0.960  0.1422  0.712674
Residuals 12  81.038   6.753
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A análise do resultado da análise de variância nos mostra que tanto o fator A quanto o Fator B são significativos para a determinação do valor da característica de interesse, isto é, ambos os fatores influenciam na característica de interesse (p-valor << 0,05).

Já a interação entre os fatores (a:b) não possui influência na característica de interesse (p-valor = 0,71 >> 0,05), como já havia mostrado o gráfico de interações exibido na Figura 64 - Análise das interações entre os fatores.

A execução do ANOVA sem interações mostra:

```
> dados_an = aov(res ~ a + b, data=dados)

> anova(dados_an)

Analysis of Variance Table

Response: res
          Df  Sum Sq Mean Sq F value    Pr(>F)
a          1  80.192  80.192  12.714  0.003452 **
b          1 223.951 223.951  35.505 4.757e-05 ***
Residuals 13  81.998   6.308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Obviamente há uma pequena diferença entre os resultados, pois como as medidas dos elementos da amostra foram geradas por meio da distribuição de frequências normalizada (valores aleatórios) existe ruído, oriundo do desvio da média amostral em relação ao valor esperado (dados da geração dos valores). Assim, saber de antemão se há interação entre os fatores ou não, direciona a execução correta da função *aov*, tornando a análise mais precisa.

Exemplo 23: Neste exemplo, foi analisado um experimento de dois fatores de dois níveis, um com influência positiva e outro com influência negativa na característica de interesse e sem interação entre os fatores. O processo de carregamento dos dados e execução do ANOVA é o mesmo.

```
> dados = read.csv2(file.choose(), header = T)

> dados_an = aov(res~a+b,data=dados)

> anova(dados_an)

Analysis of Variance Table

Response: res
          Df  Sum Sq Mean Sq F value    Pr(>F)
a          1 238.780 238.780  41.707 2.139e-05 ***
b          1  79.968  79.968  13.968  0.002487 **
Residuals 13  74.428   5.725
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A análise do resultado acima nos mostra que ambos os fatores são significativos para a determinação do valor da característica de interesse, isto é, ambos os fatores influenciam na característica de interesse (p-valor muito menor que 0,05). No entanto, a análise não mostra o tipo de contribuição (positiva ou negativa) que foi exibida na Figura 67 - Gráfico dos Efeitos Principais - Influências positiva e negativa na característica de interesse.

Exemplo 24: Este exemplo abordou um experimento com dois fatores de dois níveis, ambos com influência na característica de interesse e com interação entre os fatores.

Como já sabemos de antemão que há interação entre os fatores, opta-se por usar a formulação da função *aov* que considera a interação (res ~ a * b). O resultado confirma o resumo apresentado na Tabela 54. Tanto os fatores quanto a interação são significativos para a característica de interesse (p-valor < 0,05) e, novamente é ressaltado que a análise não nos mostra o tipo de contribuição de cada fator ou da interação (se é positiva ou negativa), conforme foi discutido anteriormente.

```
> dados = read.csv2(file.choose(), header = T)
> dados_an = aov(res~a*b, data=dados)
> anova(dados_an)
Analysis of Variance Table
Response: res
          Df Sum Sq Mean Sq F value    Pr(>F)
a          1 826.85  826.85 285.518 9.885e-10 ***
b          1  43.30   43.30  14.951  0.002242 **
a:b        1  33.70   33.70  11.636  0.005162 **
Residuals 12  34.75    2.90
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

8.4 ANOVA – Análises de Validação

No Capítulo 8.1 foram apresentados os requisitos relativos aos dados para a execução da análise de variância, a saber: observações independentes; variâncias iguais; e distribuições normais. A independência das observações é um pressuposto que o planejamento do experimento deve garantir e as normalidade das distribuições dos tratamentos deve ser testada como mostrado anteriormente (testes de Shapiro-Wilk e Shapiro-Francia). Quanto a igualdade da variância, ela pode ser verificada como mostrado mais adiante.

Estes requisitos existem para garantir que os resultados da ANOVA expressem de forma correta a realidade da correlação e influência dos fatores em relação à característica de interesse. Uma das formas de verificarmos isto é o coeficiente de determinação (R^2).

Coeficiente de Determinação

O coeficiente de determinação (R^2) mede o quanto a característica de interesse é explicada pelo modelo. Quanto maior o valor de R^2 melhor o modelo explica a variação da característica de interesse. Um valor acima de 0,70 indica que o modelo proposto está explicando bem a relação entre os fatores e a característica de interesse. A equação usada para calcular o R^2 é dada por:

$$R^2 = 1 - \frac{SQE}{SQT} \qquad \textit{Eq. 64}$$

Para verificarmos o quanto cada modelo estatístico apresentado nos exemplos de 21 a 24 explica a relação entre os fatores e a característica de interesse, vamos calcular o valor de R^2 para cada um deles. O resultado é apresentado na Tabela 55.

Exemplo	SQE	SQT	$R^2 = 1 - SQE/SQT$
21	490,82	764,06	0,3576
22	81,038	386,141	0,7901
23	74,428	393,176	0,8107
24	34,730	938,680	0,9630

Tabela 55 - Cálculo de R2 para os exemplos anteriores

Como pode ser visto acima, o modelo do exemplo 21 é o único que, considerando o coeficiente de determinação R^2, não explica adequadamente a relação entre os fatores e a característica de interesse. Mas isto tem um motivo e esse motivo será entendido quando complementarmos nossa análise com o uso de outras ferramentas estatísticas.

Gráficos de Diagnóstico do Modelo Estatístico

A ANOVA possui quatro gráficos para diagnóstico do modelo estatístico. Esses gráficos, são fornecidos pela função *plot* associada ao nome da variável que armazena os resultados da ANOVA (ex.: *plot(dados_an)* onde *dados_an* é a variável indicada para armazenar o resultado do ANOVA). A seguir, os quatro gráficos gerados para o modelo estatístico do Exemplo 21, são exibidos e explicados:

O gráfico 1 (resíduos vs. Valores ajustados ou *Residual vs. Fitted*[20]) mostra indícios sobre o comportamento da variância dos resíduos com relação aos valores ajustados (preditos pelo modelo), sendo ideal para analisar a presença de não-linearidades no modelo. A linha vermelha no gráfico (Figura 70) denota a média dos resíduos e deve se ajustar ao valor zero. Para o Exemplo 21, os valores dos resíduos estão uniformemente distribuídos em torno do valor zero. Assim, o modelo é considerado como linear e válido.

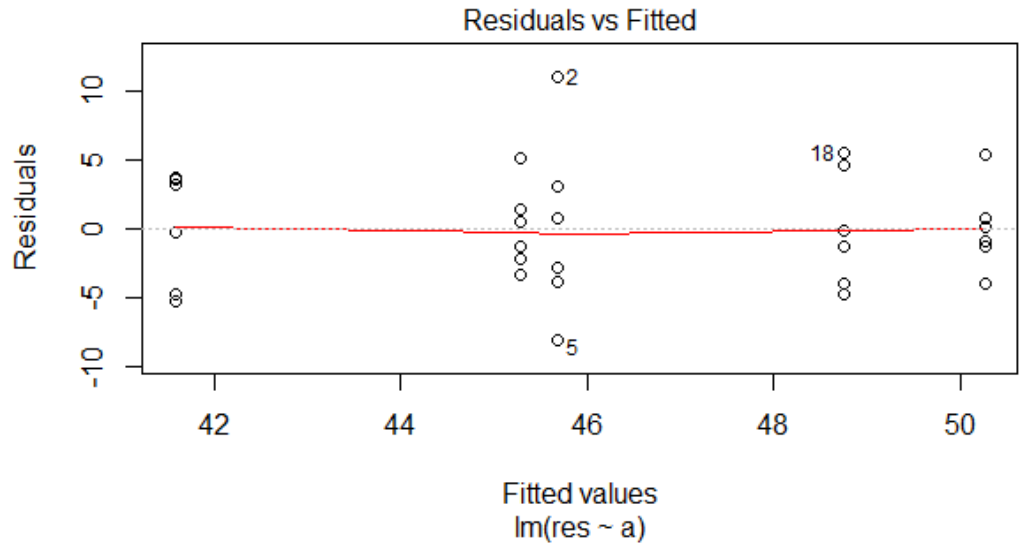

Figura 70 - Gráfico Residuos x Valores ajustados

O gráfico 2 (Q-Q[21]) dos resíduos padronizados, é usado para verificação da normalidade dos resíduos, verificando-se o afastamento da curva ideal. Um certo afastamento, principalmente no início e final (caudas da distribuição normal) é esperado. Para o Exemplo 21, tendo como hipótese nula a normalidade dos resíduos, o gráfico (Figura 71) indica a aceitação da hipótese, uma vez que não há afastamentos extremos da curva.

[20] O gráfico resíduos vs. valores ajustados deve exibir uma nuvem de pontos aleatórios e homogêneos distribuídos em torno do eixo horizontal (y = 0)

[21] O gráfico Q-Q (quantil-quantil ou qq-plot) é um recurso gráfico exploratório usado para verificar a validade de um pressuposto de distribuição para um dado conjunto de dados

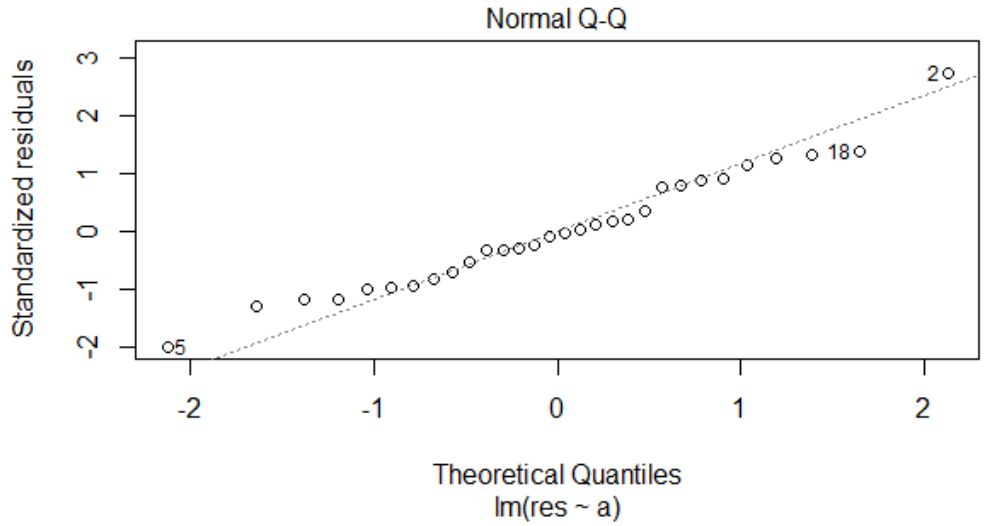

Figura 71 - Grafico Normal Q-Q

O gráfico 3 (Scale-Location) é semelhante ao gráfico 1 (Residual x Fitted), mas simplifica a análise da variação constante dos resíduos. Usa a raiz quadrada do valor absoluto dos resíduos padronizados ao invés do valor do próprio resíduo. A linha vermelha, quando horizontal, indica que a magnitude média dos resíduos padronizados não muda muito em função dos valores ajustados. No caso do Exemplo 21 (Figura 72), existe uma variação mínima nos intervalos, entre 0,6 e 1,0.

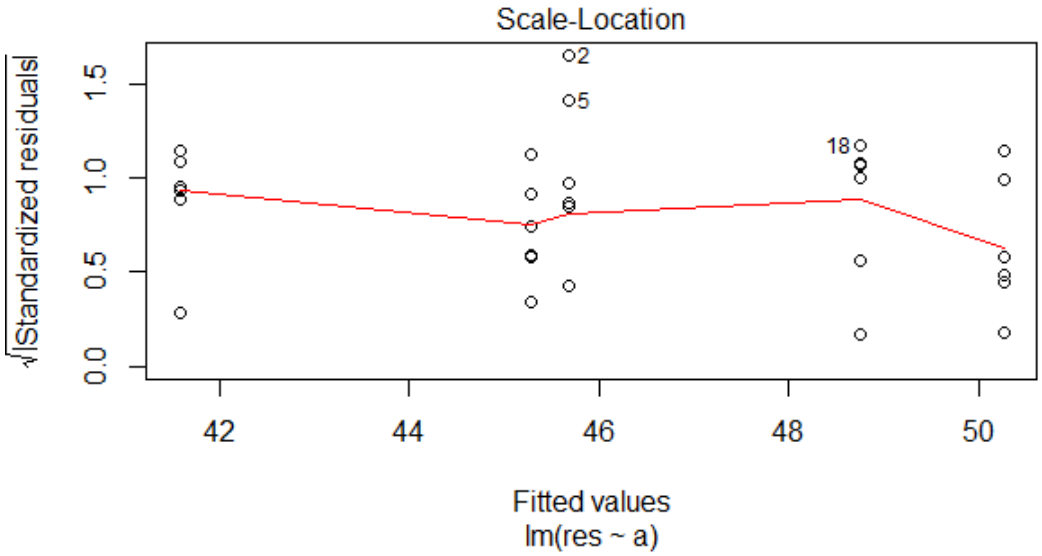

Figura 72 - Gráfico Scale - Location

O gráfico 4 (constante de Leverage) pode ser útil para detectar a presença de pontos com alta influência no modelo estatístico. No gráfico, quando uma linha tracejada vermelha delimita a distância de Cook (indicada pelo nome de "*Cook's distance*") e os pontos situados além desta linha são pontos com maior influência no modelo estatístico e sua exclusão pode melhorar o coeficiente de determinação. No caso do Exemplo 21 (Figura 73), não há a representação da linha tracejada vermelha (*Cook's distance*), indicando que os resíduos padronizados estão distantes da linha de Cook e que não existem pontos com maior influência sobre o modelo estatístico.

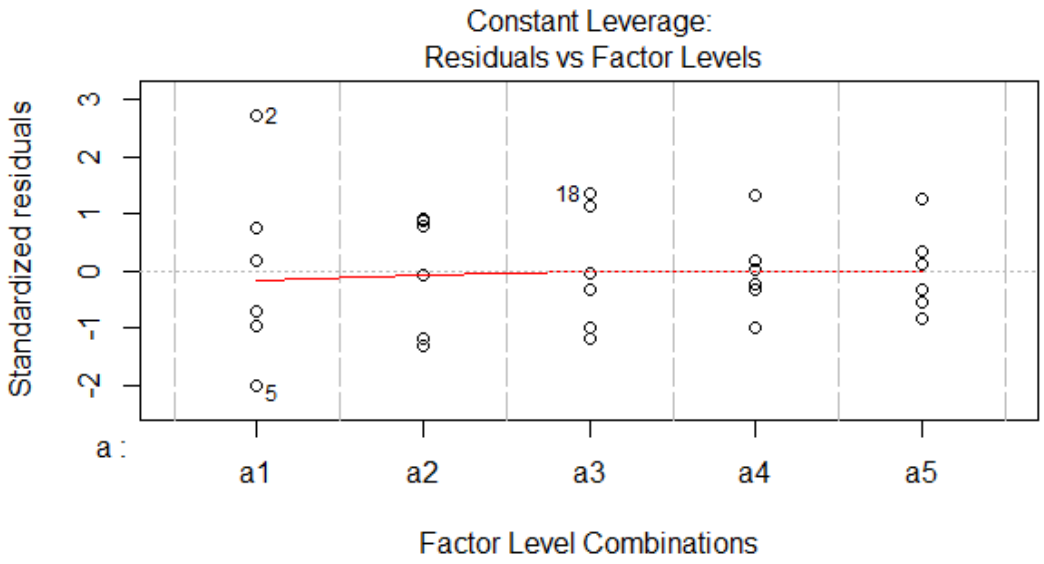

Figura 73 - Gráfico Constante de Leverage

Ainda como exemplo, os quatro gráficos para a análise do modelo apresentado no Exemplo 24 são exibidos na Figura 74. Como pode ser notado, os gráficos não indicam discrepâncias que possam invalidar o modelo ANOVA desenvolvido no exemplo.

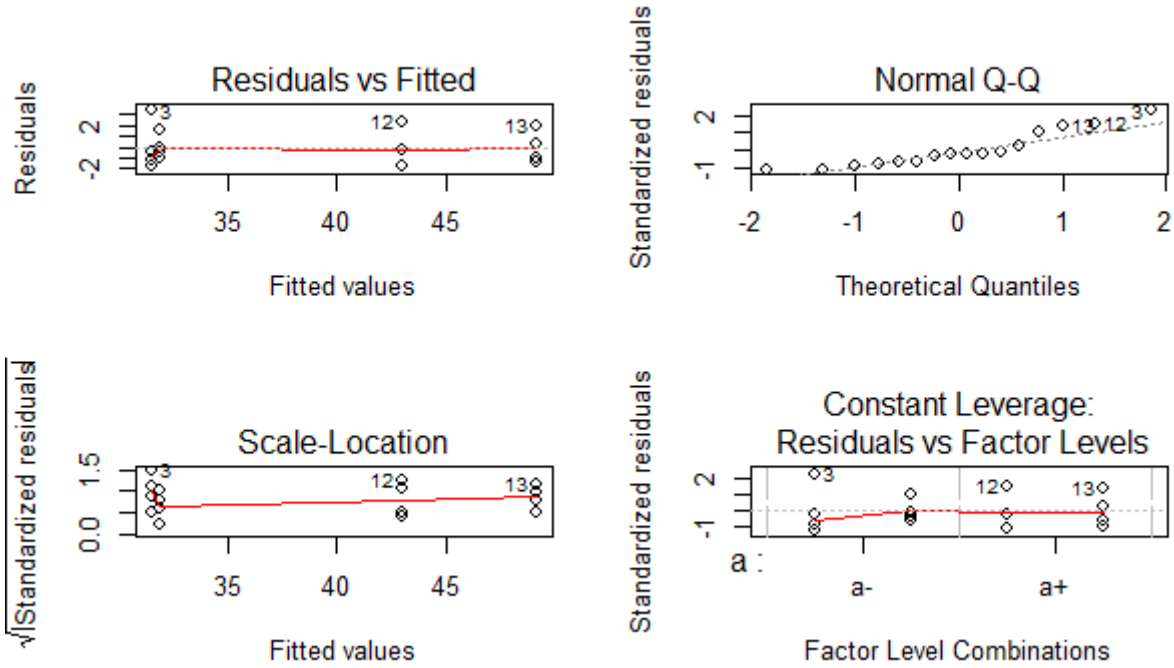

Figura 74 - Gráficos para diagnóstico do modelo ANOVA do exemplo 11

8.5 ANOVA – Complementando a análise com o Teste de Tukey

Com o uso da análise de variância ANOVA, muitas informações são esclarecidas, como quais fatores são significantes para a característica de interesse, mas ainda existem dúvidas a serem solucionadas. O Exemplo 21, onde foram analisados cinco diferentes tipos de aditivos, a ANOVA nos mostrou que podíamos rejeitar a hipótese nula, e, com isto, rejeitamos a igualdade das médias dos cinco tratamentos. No exemplo temos pelo menos duas médias que não são iguais.

Também temos que o coeficiente de determinação (R^2 = 0,3576) nos mostrou que o modelo ANOVA não explica bem a relação entre o fator (aditivo) e a característica de interesse (resistência mecânica). Além disto, a pergunta principal (qual aditivo produziu os melhores resultados) ainda não foi respondida.

Se olharmos novamente o gráfico boxplot da Figura 62, que descreve as variações dos tratamentos, poderíamos escolher entre o aditivo A3 ou o A4, que apresentam os melhores resultados. Mas eles serão estatisticamente diferentes e diferentes dos outros resultados? Para responder isto poderíamos fazer uma série de comparações de médias usando o teste *t* (t-Student) ou uma única comparação múltipla de médias.

Os testes de Tukey e Duncan fazem exatamente isto e possuem o mesmo suporte teórico do teste *t*. Portanto, vamos abordá-los diretamente no RStudio.

Teste de Tukey

Dentre os testes de comparações múltiplas existentes, o Teste de Tukey se destaca por fazer comparações entre todos os pares possíveis (médias dos tratamentos) e também por apresentar resultados rigorosos. Também conhecido como Teste de Tukey HSD (Teste de Tukey da Diferença Honestamente Significativa), é calculado pela seguinte equação:

$$DMS = q_{\alpha}(g, N-g)\sqrt{QME/n} \qquad \textit{Eq. 65}$$

Onde:

- DMS = diferença mínima significativa
- q_{α} = valor tabelado (Tabela Teste de Tukey)
- g = número de grupos a serem comparados
- N = número total de elementos dos tratamentos
- n = número de elementos no tratamento
- QME = quadrado médio do erro (SQE/Graus de liberdade ou s^2)

Com o teste, rejeita-se a igualdade de dois grupamentos de médias (*i* e *j*), se:

$$|y_i - y_j| > DMS$$

Resgatando os dados do Exemplo 21, temos:

tratamento	1	2	3	4	5
médias	45,68	41,57	48,75	50,27	45,28

- α = 0,05
- g = 5 grupamentos (correspondente as amostras dos cinco níveis – aditivos)
- n = 6 (seis elementos por amostra – grupamento)
- N = 30 elementos (seis elementos por cinco amostras)
- QME = 19,63

- $q_{0,05}(5, 25) \cong 4{,}16$

Assim, temos: $DMS = 4{,}16\sqrt{19{,}63/6} = 7{,}52$

Calculando a diferença entre as médias $|y_i - y_j|$, temos os valores exibidos na Tabela 56, onde podemos verificar que apenas a diferença entre as médias dos grupos 2 e 4 (aditivos 2 e 4) são superiores ao DMS. Então temos apenas duas médias estatisticamente diferentes.

Grupo	Diferença	**Grupo**	Diferença	**Grupo**	Diferença	**Grupo**	Diferença
Y_{12}	4,11	**Y_{23}**	7,18	**Y_{34}**	1,52	**Y_{45}**	4,99
Y_{13}	3,07	**Y_{24}**	8,7	**Y_{35}**	3,47		
Y_{14}	4,59	**Y_{25}**	3,71				
Y_{15}	0,4						

Tabela 56 - Diferença entre as médias dos tratamentos

A interpretação dos resultados é simples: o aditivo 4 somente apresenta resultados significativos (melhoria) quando comparado ao aditivo 2. Nas demais comparações, não há diferenças estatísticas significativas. Esta conclusão justifica o baixo valor do coeficiente de determinação (R^2 = 0,3576) obtido para o exemplo, afinal, apenas uma comparação de grupos apresentou diferença estatística.

No modelo ANOVA, a significância obtida (p-valor = 0,0216) advém unicamente desta diferença. Os outros aditivos (1, 3 e 5) não resultam em melhoria significativa e mesmo os aditivos 2 e 4 não apresentam diferença quando comparados com os aditivos 1, 3 e 5.

Teste de Tukey no RStudio

O teste de Tukey é executado no RStudio por meio da função *TukeyHSD(var)*, onde *var* é o nome da variável que armazena o resultado da ANOVA, ou seja, a execução do teste de tukey exige execução prévia da análise de variância.

Existem outras funções que também executam o teste de Tukey, mas são fornecidas por outros pacotes que devem ser previamente instalados, como a função *HSD.test ()*, fornecida pelo pacote "agricolae" e a função TukeyC, fornecida pelo pacote "TukeyC". As funções de pacotes específicos costumam oferecer respostas mais completas. Para o procedimento vamos carregar os dados do Exemplo 21, executar a ANOVA e, em seguida, o teste de Tukey padrão:

```
> dados = read.csv2(file.choose(),header=T)
> dados_an = aov(res ~ a, data=dados)
> TukeyHSD(dados_an)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = res ~ a, data = dados)

$a
            diff         lwr       upr     p adj
a2-a1 -4.1033333 -11.6163534  3.409687 0.5086528
a3-a1  3.0733333  -4.4396867 10.586353 0.7505567
a4-a1  4.5950000  -2.9180200 12.108020 0.3978124
a5-a1 -0.3933333  -7.9063534  7.119687 0.9998650
a3-a2  7.1766667  -0.3363534 14.689687 0.0662092
a4-a2  8.6983333   1.1853133 16.211353 0.0175529
a5-a2  3.7100000  -3.8030200 11.223020 0.6025015
```

```
a4-a3  1.5216667  -5.9913534  9.034687 0.9745999
a5-a3 -3.4666667 -10.9796867  4.046353 0.6605841
a5-a4 -4.9883333 -12.5013534  2.524687 0.3185660
```

Como pode ser visualizado, a primeira coluna contém a diferença entre as médias dos grupos, a segunda e a terceira os valores inferior e superior do intervalo de confiança da diferença entre as médias e a quarta o p-valor. A única comparação onde o p-valor é inferior a 0,05 (nível de confiança) é a4 – a2, confirmando o resultado anterior.

Em seguida, vamos executar o teste de Tukey fornecido pelo pacote "TukeyC":

```
> TukeyC(dados_an)
Results
   Means G1 G2
a4 50.27  a
a3 48.75  a  b
a1 45.68  a  b
a5 45.28  a  b
a2 41.57     b

Sig.level
 0.05

Diff_Prob
      a4    a3    a1    a5    a2
a4 0.000 1.522 4.595 4.988 8.698
a3 0.975 0.000 3.073 3.467 7.177
a1 0.398 0.751 0.000 0.393 4.103
a5 0.319 0.661 1.000 0.000 3.710
a2 0.018 0.066 0.509 0.603 0.000

MSD
      a4    a3    a1    a5    a2
a4 0.000 7.513 7.513 7.513 7.513
a3 7.513 0.000 7.513 7.513 7.513
a1 7.513 7.513 0.000 7.513 7.513
a5 7.513 7.513 7.513 0.000 7.513
a2 7.513 7.513 7.513 7.513 0.000
```

Este pacote acrescenta o agrupamento das médias. Os tratamentos a1, a3 e a5 foram colocados nos dois grupos (A e B). O tratamento a2 somente no grupo A e o tratamento a4 somente no grupo B. A interpretação é que os aditivos 1, 2, 3 e 4 possuem médias iguais e os aditivos 2, 3, 4 e 5 também possuem médias iguais. Somente as médias dos aditivos 2 e 4 são diferentes.

O RStudio também permite plotar a análise gráfica do resultado, como mostrado na Figura 75.

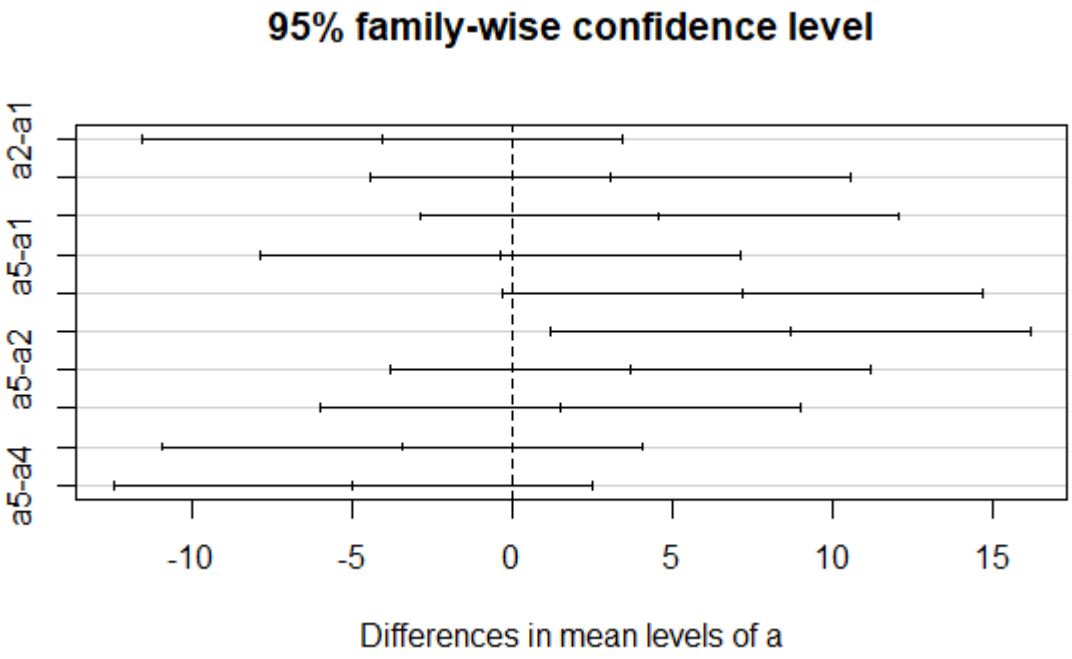

Figura 75 - Análise gráfica do teste de Tukey

E, por último, o teste de Tukey do pacote "agricolae":

```
> library(agricolae)
> dados = read.csv2(file.choose(), header=T)
> dados_an = aov(res ~a, data=dados)
> tukey_an <- HSD.test(dados_an, c("a"), main="res ~ a", console=TRUE)

Study: res ~ a

HSD Test for res

Mean Square Error:  19.63272

a,  means

        res      std r   Min   Max
a1 45.67667 6.613192 6 37.62 56.68
a2 41.57333 4.164020 6 36.27 45.24
a3 48.75000 4.290660 6 44.02 54.28
a4 50.27167 3.068638 6 46.33 55.58
a5 45.28333 3.043673 6 41.88 50.36

Alpha: 0.05 ; DF Error: 25
Critical Value of Studentized Range: 4.153363

Minimun Significant Difference: 7.51302

Treatments with the same letter are not significantly different.

        res groups
a4 50.27167      a
a3 48.75000     ab
a1 45.67667     ab
a5 45.28333     ab
a2 41.57333      b
```

Este pacote também fornece o agrupamento das médias e apresenta o valor crítico da tabela do teste de Tukey (4,153363).

Para complementarmos as análises dos outros exemplos, precisamos avançar um pouco mais, pois estes exemplos abordavam situações mais complexas que comparação de médias. Para estes, precisamos ver as Análises de Regressão (linear e múltipla). Para encerrar este item, vamos apresentar um estudo de caso que ilustra muito bem o uso da ANOVA para identificar os fatores significantes em um experimento.

8.6 ANOVA – Estudo de Caso

Análise da influência de fatores físicos (localização) no consumo energético mensal médio das unidades habitacionais[22]

O objetivo do estudo é analisar a influência dos fatores físicos relativos à localização e posicionamento das unidades habitacionais no consumo energético mensal médio destas unidades, para posterior desenvolvimento de um modelo termoenergético de uma edificação no software EnergyPlus. Para tanto foi selecionado um conjunto habitacional localizada na cidade do Rio de Janeiro (identificada como uma das três capitais estaduais com os piores cenários climáticos frente ao conforto dos usuários). O conjunto habitacional selecionado é composto por quatro condomínios e possui 900 apartamentos. Para objeto de estudo foi selecionado o Condomínio 2, com 200 apartamentos de 2 quartos (Figura 76).

Figura 76 - Localização dos Condomínios (Fonte: PRJ (2013))

O Condomínio 2 (indicado pela seta vermelha) é composto por 10 blocos de 5 andares com 4 apartamentos por andar (Figura 77). Dos 200 apartamentos do condomínio, foram levantados os consumos mensais de 67 apartamentos, representando 33,5% do total de unidades, quantidade considerada significativa para um estudo preliminar, cujo objetivo é identificar os fatores que podem ter influência no consumo de energia das unidades habitacionais.

[22] Dados fornecidos pela Doutoranda Fernanda Dutra Mourão de Oliveira (PPGEC/CEFET-MG), obtidos de sua dissertação de Mestrado (os dados foram modificados)

Os dados de consumo mensal foram levantados juntamente com a quantidade de moradores de cada unidade. Assim, temos como informações preliminares: (i) bloco; (ii) unidade; (iii) quantidade de habitantes; (iv) mês; e (v) consumo.

O levantamento dos dados foi realizado presencialmente, mediante de solicitação e permissão para registrar os dados de consumo. Também foram solicitadas informações sobre equipamentos instalados, quantidade de moradores, rotina de uso. As unidades que não forneceram estes dados foram excluídas do levantamento.

A característica de interesse, o consumo de energia mensal por unidade, segundo pesquisa bibliográfica realizada, está relacionado principalmente à quantidade de habitantes por unidade, ao perfil de uso dos equipamentos eletroeletrônicos, destacando-se chuveiros, ar-condicionado/aquecedores, fornos elétricos, dentre outros e ao gradiente de temperatura interno/externo.

Destes fatores, o perfil de uso não pode ser adequadamente classificado por falta/incorreção das informações obtidas. No entanto, acredita-se que o mesmo pode ser representado pela quantidade de habitantes. Já o gradiente de temperatura interno/externo está relacionado às médias mensais de temperatura da cidade, e, por consequência, ao mês do consumo registrado. Todas as unidades possuem preparação para instalação de ar-condicionado nos quartos e salas, mas não foi possível levantar as situações de uso deste equipamento.

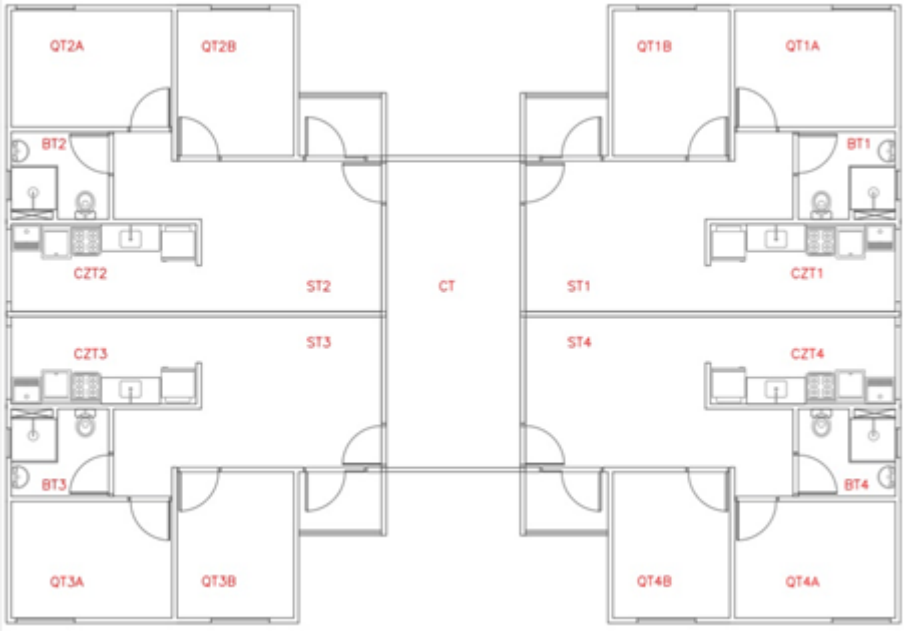

Figura 77 - Planta baixa (andar) dos blocos

Assim, a análise inicial da significância dos dados foi realizada com os fatores bloco (blc - número do bloco), mês (mês) e quantidade de habitantes (qha), sendo a quantidade de habitantes representada por: “a” (1 habitante), “b” (2 habitantes), “c” (3 habitantes), “d” (4 habitantes) e “e” (mais de 4 habitantes). Os dados foram carregados no RStudio por meio de planilha MS Excel no formato .csv e inicialmente analisados sem considerar interações:

```
>dados = read.csv2(file.choose(), header=T)
> dados_an = aov(formula=res ~ blc  + mes + cls, data =dados)
> summary(dados_an)
             Df  Sum Sq Mean Sq F value Pr(>F)
blc           1   14754   14754   3.673 0.0557 .
mes          11  531815   48347  12.035 <2e-16 ***
qha           4 1626052  406513 101.198 <2e-16 ***
Residuals   671 2695424    4017
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A análise inicial confirma a significância já prevista para o mês e para a quantidade de habitantes (p-valor próximo de zero). O bloco (localização do prédio no terreno) possui pouca significância (p-valor = 0,0557). O coeficiente de determinação R^2 foi calculado e foi igual a 0,4463 (inferior a 0,70), demonstrando que o modelo pode não representar corretamente o problema.

Para analisar a interação entre os fatores, a ANOVA foi executada novamente, desta vez com interação.

A execução da ANOVA com interação mostrou que a interação entre os fatores mês e quantidade de habitantes possui significância (p-valor = 0,000495) e a interação entre os outros fatores não.

```
> dados_an = aov(formula=res ~ blc  + mes + qha, data =dados)
> summary(dados_an)
             Df  Sum Sq Mean Sq F value Pr(>F)
blc           1   14754   14754   3.673 0.0557 .
mes          11  531815   48347  12.035 <2e-16 ***
qha           4 1626052  406513 101.198 <2e-16 ***
Residuals   671 2695424    4017
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
> dados_an = aov(formula=res ~ blc  * mes * qha, data =dados)
> summary(dados_an)
             Df  Sum Sq Mean Sq F value   Pr(>F)
blc           1   14754   14754   3.772 0.052585 .
mes          11  531815   48347  12.362  < 2e-16 ***
qha           4 1626052  406513 103.940  < 2e-16 ***
blc:mes      11    8263     751   0.192 0.998006
blc:qha       4   18090    4522   1.156 0.329127
mes:qha      38  296234    7796   1.993 0.000495 ***
blc:mes:qha  32   80974    2530   0.647 0.934612
Residuals   586 2291864    3911
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como o objetivo principal do estudo é analisar a influência dos fatores físicos relativos à localização e posicionamento de apartamentos no consumo energético e existem estudos comprovando esta influência, a primeira suspeita é de que o fator bloco não esteja representando corretamente a localização e posicionamento da unidade. Novas pesquisas bibliográficas conduzidas a fim de identificar quais os fatores relativos ao posicionamento podem ser influenciadores do consumo indicaram os seguintes aspectos a serem considerados:

- Andar (and): unidades localizadas no andar térreo (1º andar) possuem, normalmente, temperaturas internas mais baixas, devido ao contato com o solo. As unidades localizadas no último andar (5º andar) possuem temperaturas internas mais elevadas, devido ao aquecimento direto da cobertura do prédio pela irradiação solar.
- Orientação (dir): para a zona bioclimática do Rio de Janeiro, unidades habitacionais com fachada voltada para oeste possuem temperatura interna mais elevada, uma vez que recebem maior irradiação solar na fachada.

Estes novos fatores foram acrescentados aos fatores já usados, sendo que o fator bloco foi mantido para verificar, se, com a adição dos novos fatores, este fator como representante da posição do prédio no terreno

seria significante. O fator orientação (dir) foi associado à orientação da fachada principal da unidade (Leste, Norte, Sul, Oeste). O fator andar foi associado ao andar de localização da unidade (de 1 a 5).

Os dados alterados foram novamente carregados no RStudio e a ANOVA executada.

Antes da realização da análise do resultado da ANOVA, o coeficiente de determinação R^2 foi calculado e o seu valor (R^2 = 0,8413) mostrou que o modelo pode ser considerado como um modelo que representa bem o problema (R^2 > 0,70), ou seja, a introdução dos novos fatores aprimorou o modelo.

A análise do resultado da ANOVA confirma que os fatores anteriormente identificados como significativos, a saber, mês (mes), quantidade de habitantes (qha) e a interação entre estes fatores (mês:qha), são significativos neste novo modelo também. O fator bloco (blc) teve seu p-valor ligeiramente aumentado (de 0,052585 para 0,05968) o que basicamente não altera sua significância.

Dos novos fatores introduzidos no modelo (direção e andar), a direção de orientação da fachada (dir) mostrou-se significante (p-valor = 0,0000382) bem como sua interação com o bloco (posicionamento do prédio no terreno – blc:and). Isto confirma a pesquisa bibliográfica realizada e nos permite supor que uma melhor representação do posicionamento do bloco (representado por seu número), possa trazer melhores resultados.

```
> dados_an = aov(formula=res ~ blc  * and * dir * mes * qha, data =dados)
> summary(dados_an)
                    Df  Sum Sq Mean Sq F value   Pr(>F)
blc                  1   14754   14754   3.590  0.05968 .
and                  1   10122   10122   2.463  0.11826
dir                  3  100816   33605   8.176 3.82e-05 ***
mes                 11  533120   48465  11.791  < 2e-16 ***
qha                  4 1576094  394023  95.863  < 2e-16 ***
blc:and              1   23771   23771   5.783  0.01715 *
blc:dir              3   40649   13550   3.297  0.02165 *
and:dir              3   20487    6829   1.661  0.17681
blc:mes             11    7650     695   0.169  0.99883
and:mes             11   25413    2310   0.562  0.85775
dir:mes             33   72278    2190   0.533  0.98300
blc:qha              4   11959    2990   0.727  0.57428
and:qha              4   27207    6802   1.655  0.16229
dir:qha             11   64872    5897   1.435  0.16017
mes:qha             38  276920    7287   1.773  0.00679 **
blc:and:dir          3   29772    9924   2.414  0.06798 .
blc:and:mes         11   15703    1428   0.347  0.97351
blc:dir:mes         33  118812    3600   0.876  0.66439
and:dir:mes         33   40646    1232   0.300  0.99994
blc:and:qha          4   11629    2907   0.707  0.58785
blc:dir:qha         10   29483    2948   0.717  0.70758
and:dir:qha          9   43778    4864   1.183  0.30774
blc:mes:qha         31   94335    3043   0.740  0.83856
and:mes:qha         29  160938    5550   1.350  0.12125
dir:mes:qha         70  219145    3131   0.762  0.90529
blc:and:dir:mes     33  154231    4674   1.137  0.29135
blc:and:dir:qha      7   26990    3856   0.938  0.47833
blc:and:mes:qha     20   83957    4198   1.021  0.43915
blc:dir:mes:qha     41  155124    3784   0.921  0.61164
and:dir:mes:qha     22   93140    4234   1.030  0.42975
blc:and:dir:mes:qha  4   11519    2880   0.701  0.59242
Residuals          188  772730    4110
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Já o fator andar (and) não se mostrou significante para o consumo mensal de energia por unidade (p-valor = 0,11826). No entanto, a interação entre bloco e andar (blc:and) e entre bloco, andar e direção (blc:and:dir) mostram certa significância, reforçando a suposição anterior (melhor representação do posicionamento do bloco, talvez por orientação, similar à direção).

Como resultado da análise os seguintes fatores e interações foram identificados como significativos para o consumo mensal de energia de cada unidade habitacional e devem ser considerados no modelo termoenergético a ser desenvolvido no software EnergyPlus[23]:

- Mês, uma vez que as médias mensais de temperatura da cidade influenciam no gradiente de temperatura interno/externo e os usuários se valem de meios de aquecimento/resfriamento para compensar o gradiente de temperatura.
- Quantidade de habitantes: a quantidade de pessoas e seus padrões de uso vão influenciar diretamente o consumo mensal de energia, portanto, é fundamental que o modelo do EnergyPlus os represente corretamente.
- Direção: a orientação de fachada de cada unidade habitacional influi na quantidade de irradiação solar recebida e, consequentemente, na temperatura interna da unidade.
- A interação entre os fatores bloco e andar (p-valor = 0,01715) é significativa, mas necessita ser melhor explicitada no modelo, uma vez que a representação numérica do bloco pode não ser a melhor a ser adotada no modelo a ser desenvolvido.
- A interação entre os fatores bloco e direção (p-valor = 0,02165) é significativa e as considerações anteriores são válidas para ela também.
- A interação entre os fatores bloco, andar e direção (p-valor = 0,06798) possui significância superior a 0,05 e poderia ser desprezada. No entanto, esta interação merece ser investigada novamente, após alteração da representação do fator bloco.

Os gráficos de validação do modelo ANOVA são exibidos na Figura 78. A interpretação dos gráficos pode ser conferida no item 8.4 ANOVA – Análises de Validação, onde os gráficos são explicados e não indicam problemas no modelo estatístico.

[23] EnergyPlus ™ é um programa de simulação de energia de edifícios para modelagem doo consumo de energia (aquecimento, resfriamento, ventilação, iluminação)

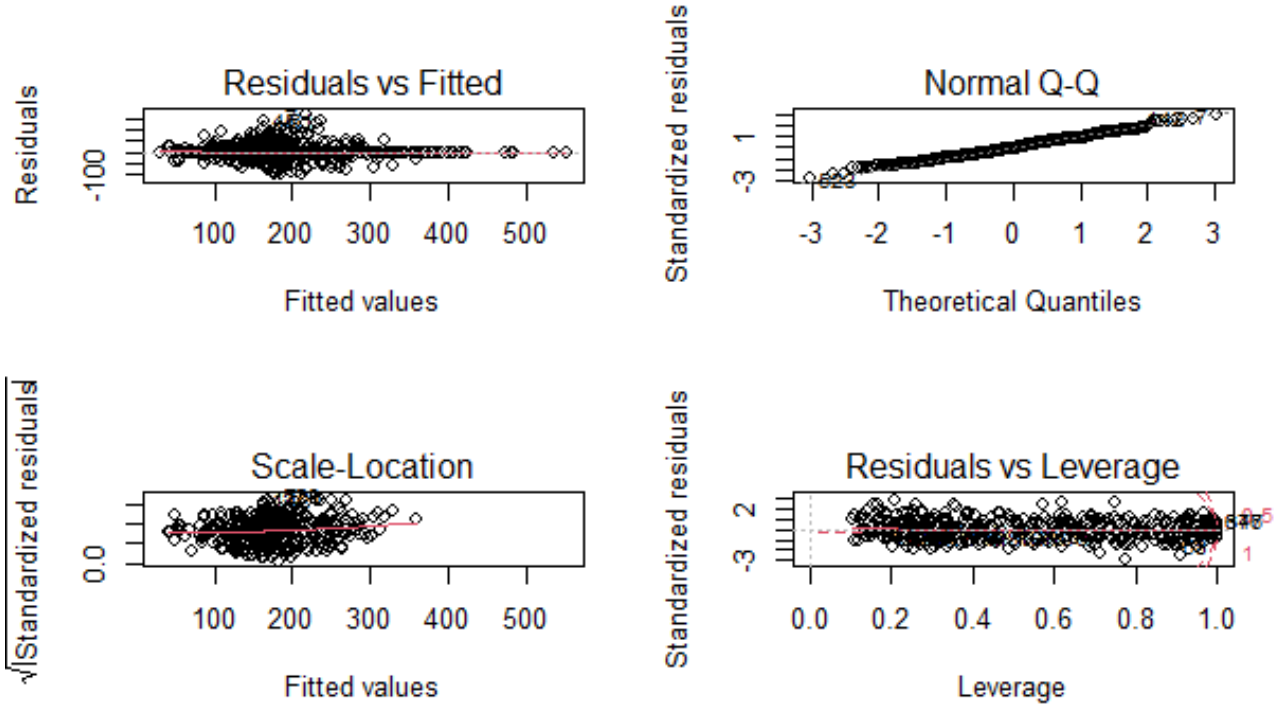

Figura 78 - Gráficos de Validação do modelo ANOVA

9 ANÁLISE DE REGRESSÃO

No capítulo anterior, a Análise de Variância, estudamos ferramentas estatísticas que nos permitiram identificar quais os fatores (ou variáveis de entrada) que influenciavam a característica de interesse (ou variável de saída). No entanto, não nos foi possível identificar como os fatores influenciam a característica de interesse (se positiva ou negativamente).

Em diversas situações é necessário identificar como as entradas de um processo estão influenciando os resultados obtidos. Nestes casos é necessário estabelecer um modelo matemático que explique a relação entre as variáveis de entrada e a de saída. Este tipo de modelagem é denominado **REGRESSÃO** e ajuda a entender como o comportamento das variáveis de entrada pode mudar o comportamento da variável de saída.

Como exemplo, vamos supor que o valor de um imóvel possa ser determinado unicamente pela relação (R) entre a área construída (a_c) e a área do terreno (a_t). Assim, um terreno totalmente construído teria uma relação de um (1) e um com nada construído teria uma relação de zero (0). Uma forma razoável de expressar a relação entre a entrada e a saída seria:

$$Valor = \alpha + \beta R, \ \ onde\ R = a_c/a_t \qquad \textit{Eq. 66}$$

Ou, chamando a variável de saída de Y e a variável de entrada de X, temos: $Y = \alpha + \beta x$ e sua representação gráfica seria dada pela Figura 79.

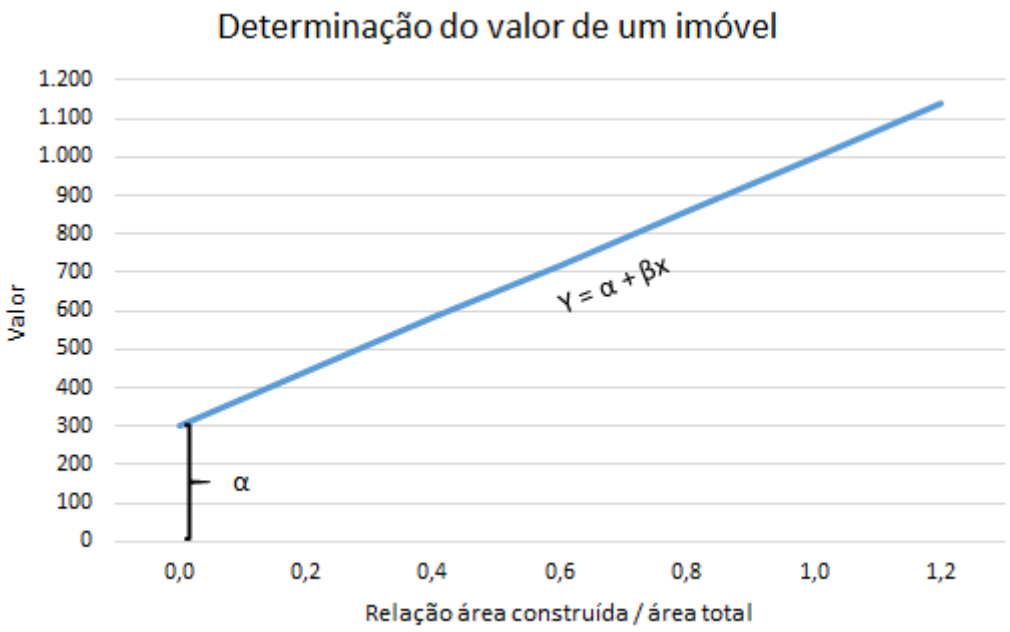

Figura 79 - Gráfico de uma relação linear

Onde α é o intercepto, representando o valor do terreno (sem construções) e β a inclinação da reta.

Na regressão, as variáveis de saída (característica de interesse, resposta ou saída do processo) são denominadas varáveis dependentes, porque seus valores são determinados pelas variáveis de entrada (fatores) que, por sua vez, são denominadas variáveis independentes ou regressores naturais.

Se a relação entre a variável dependente e seu regressor for exata, trata-se de uma relação determinística e não há componente aleatório ou probabilístico nela. No entanto, nos exemplos estudados e em praticamente

todos os experimentos de engenharia e outras ciências, esta relação não é determinística. Ela é probabilística e, desta forma, para um dado valor de x, nem sempre obtemos o mesmo valor de Y. O conceito de Análise de Regressão tenta encontrar o melhor modelo matemático que explique a relação entre x e Y, quantificando a força desta relação e permitindo a previsão dos valores de Y em função dos valores possíveis do regressor x.

A previsão dos valores de Y em função de x é um dos atributos mais importantes da regressão, uma vez que podemos utilizar o modelo para obtermos os valores de Y correspondentes aos valores de x que não estavam entre os dados usados para gerar o modelo. Este procedimento é chamado **predição** e, em geral, é válida para os valores de x que estão dentro do intervalo de x estudado. A utilização de valores fora do intervalo estudado recebe o nome de **extrapolação** e deve ser usada com cuidado, pois, o modelo é válido no intervalo estudado. Fora deste intervalo, não podemos ter certeza de sua acuracidade. A predição é a aplicação mais comum para os modelos de regressão.

Além da predição, a regressão nos permite identificar os regressores mais significativos para a variável dependente. O modelo matemático resultante nos permite visualizar os regressores que mais contribuem e eliminar aqueles cuja contribuição não seja importante, em processo similar ao que a ANOVA realiza.

A análise de regressão depende da coleta de dados e da quantidade de níveis de cada tratamento. Se tivermos apenas dois níveis, independentemente da quantidade de elementos na amostra de cada nível, a resposta obtida será sempre uma linha reta unido os pontos médios (média amostral) de cada nível. Com mais de dois níveis, podemos avaliar se a resposta é realmente linear ou não e existem artifícios que podem ser empregados caso a resposta obtida não seja linear.

O estudo dos modelos de regressão podem ser divididos em: **Regressão Linear Simples**, onde apenas uma variável de entrada (regressor) possui influência sobre a variável dependente (resposta); **Regressão Linear Múltipla**, onde a variável dependente está relacionada com mais de um regressor (vários fatores influenciam a resposta); e **Regressão Logística**, onde a variável dependente é uma variável qualitativa e apresenta valores como possíveis realizações uma qualidade (ou atributo) e não mais como resultado de uma mensuração.

9.1 Regressão Linear Simples

O modelo da regressão linear simples pressupõe que apenas um regressor afete a variável dependente, assim, a resposta Y está relacionada com o regressor x (variável independente) por meio da equação

$$Y = \alpha + \beta x + \epsilon \qquad \textit{Eq. 67}$$

Onde α e β são os parâmetros desconhecidos do intercepto e da inclinação respectivamente e Є é uma variável aleatória assumida como sendo distribuída com $\sum \epsilon = 0$. Da equação que representa o modelo podemos intuir que:

- A variável dependente Y também é aleatória, já que Є é aleatório.
- O valor da variável regressora x não é aleatório e pode ser mensurado com erro desprezível.
- O valor de Є, chamado de erro aleatório ou distúrbio aleatório (ruído), evita que o modelo se torne um modelo determinístico.

Como Є está distribuído de forma que $\sum \epsilon = 0$, temos que para um valor de x específico, os valores de Y estão distribuídos ao redor da reta de regressão real $Y = \alpha + \beta x$. Se o modelo matemático for bem determinado, ou seja, se não houver regressores adicionais não considerados e a suposição de linearidade for adequada

dentro do intervalo de valores estudados, a somatória dos erros positivos e negativos ao redor da regressão real será próxima de zero.

Na prática, não conhecemos a reta da regressão real, mas podemos supor que ela exista e podemos desenhar uma reta estimada que satisfaça da melhor forma possível a suposição $\sum \epsilon = 0$. A Figura 80 apresenta a reta de regressão real de um caso hipotético com os erros de cada observação enfatizados.

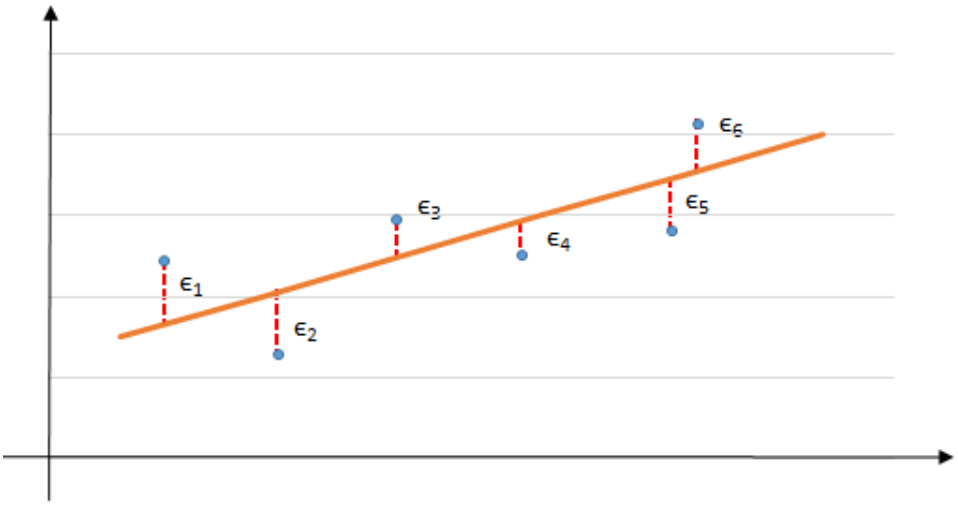

Figura 80 - Diagrama de dispersão dos dados hipotéticos (x,y) ao redor da reta de regressão real

Voltamos a reafirmar que a reta representada na Figura 80 é uma idealização. Em uma situação real, desconhecemos a regressão real e precisamos determina-la com as observações disponíveis, o que pode resultar em uma ótima representação ou não. Isto depende principalmente da qualidade dos dados disponíveis.

Para melhor entendermos isto, vamos plotar o gráfico de dispersão de uma outra situação, envolvendo um experimento de um fator com quatro níveis, com amostra de três elementos para cada nível. Esta situação é ilustrada na Figura 81.

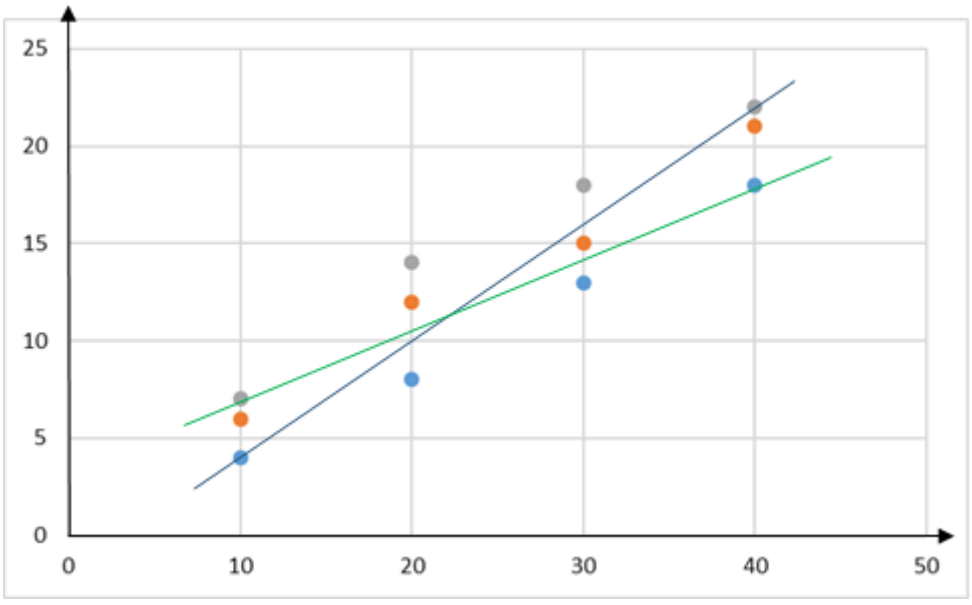

Figura 81 - Gráfico de dispersão de um experimento de 4 níveis com amostras de 3 elementos

Os níveis do experimento estão representados no eixo x e a resposta no eixo Y. Como temos três respostas para cada nível, qualquer suposição de reta entre os pontos que representam as amostras poderia ser a regressão real, como as duas retas exibidas no gráfico.

A questão passa a ser, então, como determinar a melhor aproximação linear que represente a regressão real. Assim como usado na ANOVA, o método dos mínimos quadrados é o modelo matemático utilizado para determinar os valores de α e β.

Da mesma forma que para a ANOVA, temos que estabelecer os pressupostos que orientam o modelo de regressão:

- A relação matemática entre x e Y é linear no intervalo de estudo.
- A variável independente x não é uma variável aleatória, ou seja, seus valores são fixos (controlados).
- A média do erro é nula, ou seja $\sum \epsilon = 0$.
- Para um dado valor de x, a variância do erro (ϵ) é sempre σ^2, ou seja, a variância dos erros é sempre igual.
- Os erros (ϵ) são aleatórios e seguem a distribuição normal e o erro de uma observação não está correlacionado com o erro de outra observação.

Método dos Mínimos Quadrados

Supondo que a relação entre x e Y é linear no intervalo estudado, podemos estimar os parâmetros α e β para obter a melhor reta que represente a relação entre as variáveis. O Método dos Mínimos Quadrados é uma estratégia de estimação dos parâmetros da regressão e sua aplicação não se limita apenas às relações lineares.

Para a análise de regressão, o primeiro passo é obter as estimativas dos parâmetros α e β. Os valores das estimativas são obtidos a partir dos desvios de cada elemento (x_i, Y_i, i = 1, ..., n) da amostra (ϵ_i) em relação a uma reta arbitrária $\alpha + \beta x$ passando por estes pontos, como mostrado no gráfico da Figura 82.

Para o valor *xi* do regressor, o valor predito por esta reta é $\alpha + \beta x_i$, enquanto o valor observado é *Yi*. Os desvios entre estes dois valores é $\epsilon_i = Y_i - (\alpha + \beta x_i)$, que corresponde a distância vertical do ponto à reta arbitrária.

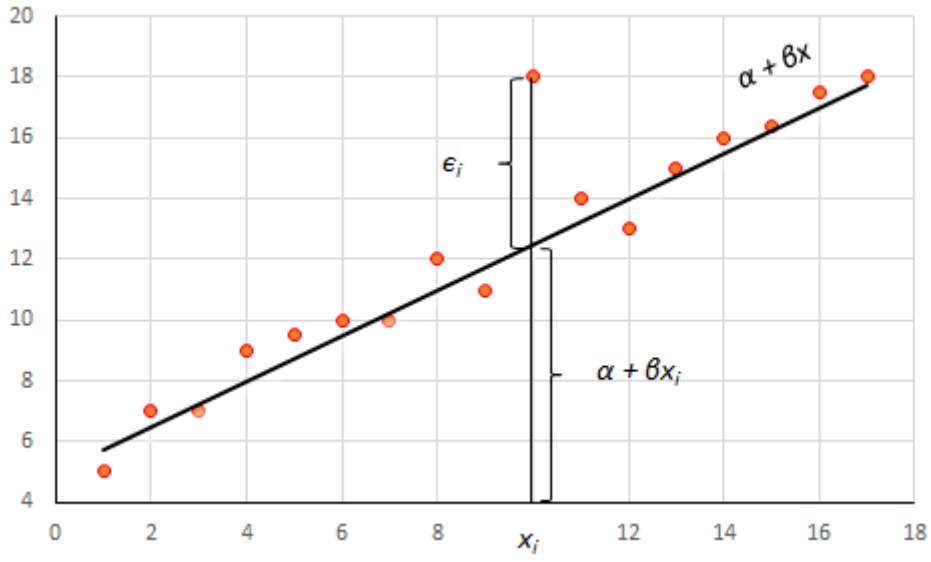

Figura 82 - Reta de regressão

O objetivo do modelo de regressão é estimar os parâmetros α e β de modo que o quadrado dos desvios (ϵ_i) entre os valores observados e estimados sejam os menores possíveis. O método de mínimos quadrados, usado no modelo de regressão, é baseado na minimização da soma dos quadrados dos erros em torno da reta de regressão, denominada SQE. Assim, devemos determinar α e β de forma que o valor de SQE seja o menor possível:

$$SQE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2 = \sum_{i=i}^{n} (Y_i - \alpha - \beta x_i)^2 \qquad \textit{Eq. 68}$$

Deixando as deduções matemáticas para aqueles que queiram se aprofundar no estudo da Estatística, a equação acima pode ser decomposta em três fatores principais, a soma dos quadrados dos desvios das médias de x e de Y e a soma dos produtos cruzados de x e Y, conforme expresso a seguir:

$Sxx = \sum_{i=1}^{n} (x_i - \bar{x})^2$ — somatório dos quadrados dos desvios de x_i em relação à média de $\bar{x}$

$Syy = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ — Somatório dos quadrados dos desvios de Y_i em relação à média de $\bar{Y}$

$Sxy = \sum_{i=1}^{n} [(x_i - \bar{x})(Y_i - \bar{Y})]$ — Somatório dos quadrados do produto cruzado de x_i e Y_i em relação ao produto da média de $\bar{x}$ e $\bar{Y}$

Ou ainda, prosseguindo com a dedução matemática:

$$Sxx = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \qquad Syy = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \qquad Sxy = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \qquad \textit{Eq. 69}$$

Desta forma, as estimativas de mínimos quadrados de α e β, em termos desta notação são:

$$\beta = Sxy/Sxx \qquad \textit{Eq. 70}$$

$$\alpha = \bar{Y} - \alpha\bar{x} \qquad \textit{Eq. 71}$$

Coeficiente de Determinação

Da mesma forma que para a ANOVA, o coeficiente (R^2) mede o quanto a característica de interesse é explicada pela curva de regressão linear. Quanto maior o valor de R^2 melhor a equação da curva traduz a variação da característica de interesse. Um valor acima de 0,70 indica que o modelo proposto está explicando bem a relação entre os fatores e a característica de interesse. A expressão usada para calcular o R^2 é dada por:

$$R^2 = \frac{S_{xY}^2}{S_{xx}\, S_{YY}} \qquad \textit{Eq. 72}$$

Exemplo 25: A influência da adição de cinza de bagaço de cana de açúcar na resistência de compressão diametral de peças queimadas de cerâmica vermelha foi testada por meio de um experimento de um fator com cinco níveis (respectivamente, 0%, 5%, 10%, 15% e 20% de adição de cinzas). As outras matérias primas foram mantidas constantes. Para cada tratamento, foram feitas amostras de cinco elementos, cujos resultados de resistência são mostrados na Tabela 57. Monte o gráfico de dispersão e determine a curva de regressão linear correspondente.

	Tratamentos				
	0%	5%	10%	15%	20%
1	2,91	2,55	1,39	1,04	0,85
2	2,89	2,40	1,48	1,10	0,97
3	2,76	2,59	1,51	1,17	0,92
4	2,90	2,34	1,50	1,06	0,89
5	3,11	2,41	1,56	1,13	0,93

Tabela 57 - Resultados dos ensaios de resistência (MPa)

O gráfico de dispersão é bem simples de ser montado. Basta plotar os tratamentos no eixo x e os valores da resistência de cada tratamento no eixo y, resultando no gráfico mostrado na Figura 83.

Observando o gráfico, podemos verificar que a suposição de linearidade da curva é válida, principalmente no intervalo de 0% a 15% de adição. A dispersão do tratamento com 20% de adição de cinza de bagaço de cana foge um pouco da linearidade para este tratamento, mas não impede que a análise seja realizada.

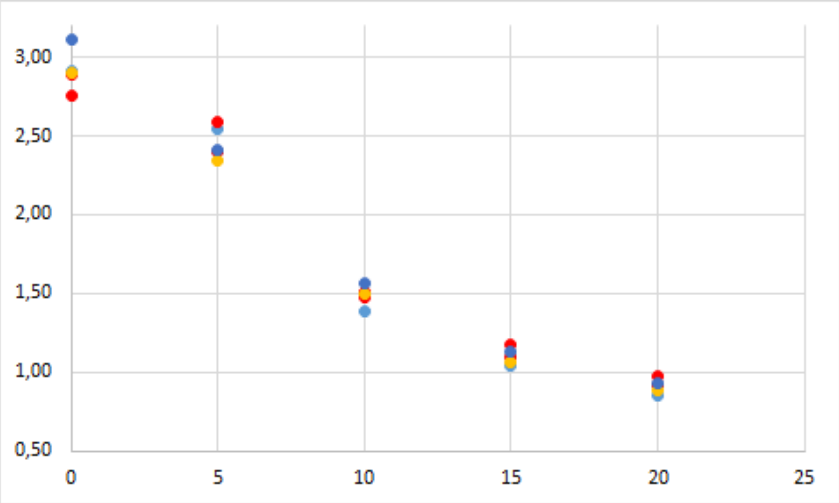

Figura 83 - Gráfico de dispersão do Exemplo 12

Para facilitar os cálculos de Sxx, SYY e SxY, podemos organizar os resultados dos ensaios em duas colunas, a primeira (x) com os valores dos percentuais de adição (0, 5, 10, 15 e 20) e a segunda com o valor da resistência à compressão diametral. Neste formato, os valores do percentual de adição irão se repetir para cada elemento da amostra. Com este formato, fica mais fácil de calcularmos os valores base para a equação, conforme mostrado na Tabela 58.

Calculando os parâmetros α e β:

$$\beta = S_{xY}/S_{xx} = -134{,}05/1250 = -0{,}10724$$

$$\alpha = \bar{Y} - \beta\bar{x} = 1{,}7744 - (-0{,}10724 \times 10) = 2{,}8468$$

O que se traduz na equação da curva da regressão linear:

$$Y = 2{,}8468 - 0{,}10724\,x$$

x_i	Y_i	$(x_i - \overline{x})^2$	$(Y_i - \overline{Y})^2$	$(x_i - \overline{x})^2 (Y_i - \overline{Y})^2$
0	2,91	100	1,289587	-11,356
0	2,89	100	1,244563	-16,0041
...	...	...	...	...
0	3,11	100	1,783827	-18,4302
5	2,55	25	0,601555	-3,83086
5	2,40	25	0,391375	-8,99222
...	...	...	...	...
5	2,41	25	0,403987	-10,4505
...	...	...	...	...
20	0,85	100	0,854515	4,347486
....	...	...	...	...
20	0,93	100	0,713011	1,478366
$\overline{x}$	$\overline{Y}$	S_{xx}	S_{YY}	S_{xY}
10	1,7744	1250	15,3758	- 134,05

Tabela 58 - Valores para cálculo de regressão

Para sabermos se a equação acima representa bem o comportamento da característica de interesse (resposta *Y*) em função da variável independente (*x*), vamos determinar o coeficiente de determinação (R^2):

$$R^2 = \frac{S_{xY}^2}{S_{xx}\, S_{YY}} = \frac{(-134{,}05)^2}{(1250 \times 15{,}3758)} = 0{,}934944$$

O valor de R2 é superior a 0,70, significando uma boa representatividade para a curva de regressão linear apresentada.

9.2 Regressão Linear Múltipla

Regressão múltipla é uma coleção de técnicas estatísticas usadas para construir modelos que descrevem as relações entre as várias variáveis independentes de entrada e a saída de um determinado processo. A diferença entre a regressão linear simples e a múltipla é que a regressão múltipla possui duas ou mais variáveis independentes relacionadas à uma única resposta.

Na maioria dos problemas em que a análise de regressão é aplicada é necessário de mais de uma variável independente no modelo de regressão, ou seja, a resposta *Y* é influenciada por mais de um fator. Um modelo de regressão linear múltipla com k variáveis independentes $x_1, x_2, \ldots, x_k$, associadas a uma resposta *Y* é dado pela equação:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad \textit{Eq. 73}$$

Onde cada coeficiente β é estimado com base nos dados da amostra, usando o método dos mínimos quadrados. Para um modelo de regressão linear múltipla com duas variáveis independentes x_1 *e* x_2 e sem interação entre si, a equação pode ser transcrita como:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 \qquad \textit{Eq. 74}$$

Se formos considerar a possibilidade de interação entre as variáveis independentes, ou seja, o efeito de x_1 na resposta média depende do nível de x_2 e, analogamente, o efeito de x_2 na resposta média depende de x_1, o modelo de regressão passa a ser:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \qquad \textit{Eq. 75}$$

Os pressupostos necessários para o desenvolvimento do Modelo de Regressão Linear Múltipla são:

- O erro tem média zero e variância σ^2, desconhecida.
- Os erros são não correlacionados;
- Os erros têm distribuição normal;
- Os valores da variáveis independentes $x_1, x_2, \ldots, x_k$ não são aleatórios e podem ser mensurados com erro desprezível.

Para o desenvolvimento do modelo, suponha um experimento com *n* observações da variável resposta e das *p* variáveis independentes (*n* > *p*). Sendo Y_i o valor da variável resposta na i-ésima observação e x_{ij} o valor da variável independente x_j também na i-ésima observação, para *j* = 1, 2, ..., *p*. O modelo pode ser representado como mostrado na Tabela 59.

Y	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$		$\boldsymbol{x_p}$
$\boldsymbol{Y_1}$	$\boldsymbol{x_{11}}$	$\boldsymbol{x_{22}}$		$\boldsymbol{x_{1p}}$
$\boldsymbol{Y_2}$	$\boldsymbol{x_{21}}$	$\boldsymbol{x_{23}}$		$\boldsymbol{x_{2p}}$
$\boldsymbol{Y_n}$	$\boldsymbol{x_{n1}}$	$\boldsymbol{x_{n2}}$		$\boldsymbol{x_{np}}$

Tabela 59 - Representação dos dados para modelo de regressão linear múltipla

Cada observação Y_i deve satisfazer a equação:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \qquad \textit{Eq. 76}$$

O objetivo do método dos mínimos quadrados é fazer com que a somatória de ϵ_i tenda a zero, ou seja, minimizar a equação:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \alpha - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}\right)^2 \qquad \textit{Eq. 77}$$

O que podemos obter derivando a equação em função de todos os β's, o que vai conduzir a uma representação matricial, cuja equação simplificada é:

$$Y = x\beta + \epsilon \qquad \textit{Eq. 78}$$

Onde:

$$Y = \begin{bmatrix} Y_1 \\ \ldots \\ Y_n \end{bmatrix}, \quad x = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \ldots \\ \beta_p \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \ldots \\ \epsilon_p \end{bmatrix} \qquad \textit{Eq. 79}$$

O cálculo e determinação dos valores dos coeficientes α e β's envolve cálculos matriciais um pouco mais complexos que os desenvolvimentos anteriores apresentados para a ANOVA e para a Regressão Linear Simples. Como nosso objetivo é o uso prático dos recursos estatístico, sem detrimento da teoria que orienta o raciocínio do pesquisador, acreditamos ser mais produtivo apresentar o uso da Regressão Linear Múltipla por meio do RStudio, nosso próximo item.

9.3 Regressão Linear No RStudio

A execução da regressão linear no RStudio é realizada pela função *lm()*. Sim, a mesma função que também executa a ANOVA (afinal, ambas as análises estatísticas são baseadas no método dos mínimos quadrados). A função *lm()* é utilizada tanto para Regressão Linear Simples quanto Múltipla (assim como a função *aov* da ANOVA).

Regressão Linear Simples

Em primeiro lugar, vamos ver como funciona a Regressão Linear Simples, com os dados do Exemplo 25- A influência da adição de cinza de bagaço de cana de açúcar na resistência de compressão diametral de peças queimadas de cerâmica vermelha.

Os dados devem ser fornecidos ao software em colunas, uma para a variável independente (*x*) e outra para os resultados (*Y*). O quadro abaixo exibe a entrada dos dados e a execução da regressão linear.

```
> dados = read.csv2(file.choose(), header = T)
> summary(dados)
       x               y         
 Min.   : 0   Min.   :0.850  
 1st Qu.: 5   1st Qu.:1.060  
 Median :10   Median :1.500  
 Mean   :10   Mean   :1.774  
 3rd Qu.:15   3rd Qu.:2.550  
 Max.   :20   Max.   :3.110  
> dados_lm = lm(y ~ x, data = dados)
> summary(dados_lm)

Call:
lm(formula = y ~ x, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3844 -0.1782  0.0432  0.1880  0.2794 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.846800   0.072242   39.41  < 2e-16 ***
x           -0.107240   0.005899  -18.18 3.82e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2085 on 23 degrees of freedom
Multiple R-squared:  0.9349,     Adjusted R-squared:  0.9321 
F-statistic: 330.5 on 1 and 23 DF,  p-value: 3.825e-15
```

Se resgatarmos a equação calculada anteriormente ($Y = 2{,}8468 - 0{,}10724\,x$), veremos que os coeficientes α e β apresentam os mesmos valores, assim como o coeficiente de determinação ($R^2 = 0{,}934944$).

Além do coeficiente de determinação, os mesmos gráficos de diagnóstico do modelo, apresentados para a ANOVA, podem ser utilizados para verificar a acuracidade do modelo gerado. Recordando seus conceitos, temos:

O Gráfico Residual vs. Fitted **(**Figura 84**)** apresenta o comportamento da variância dos resíduos com relação aos valores ajustados (preditos pelo modelo), sendo ideal para analisar a presença de não-linearidades no modelo. A linha vermelha no gráfico denota a média dos resíduos, e deve se aproximar de uma linha reta (considerar a escala utilizada).

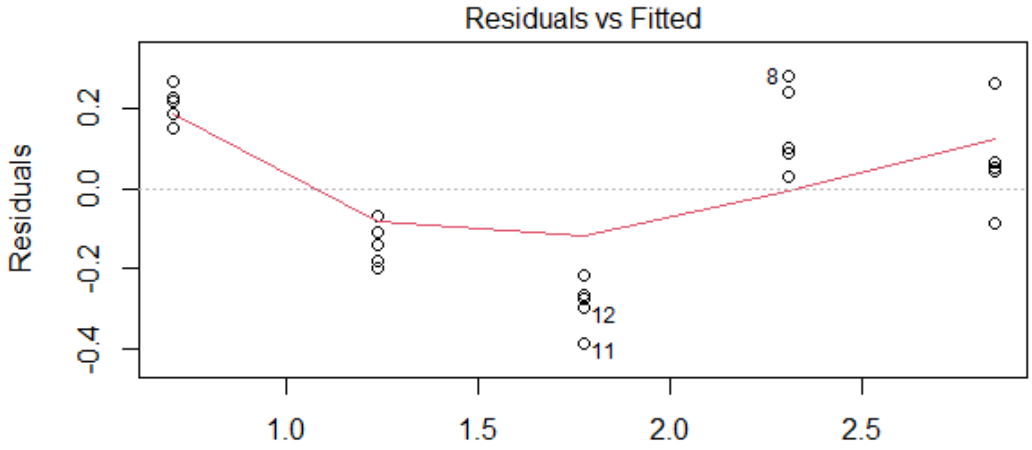

Figura 84 - Gráfico Residuos x Valores ajustados

O gráfico Normal Q-Q **(**Figura 85**)** dos resíduos padronizados analisa a normalidade dos resíduos, verificando o afastamento da curva ideal.

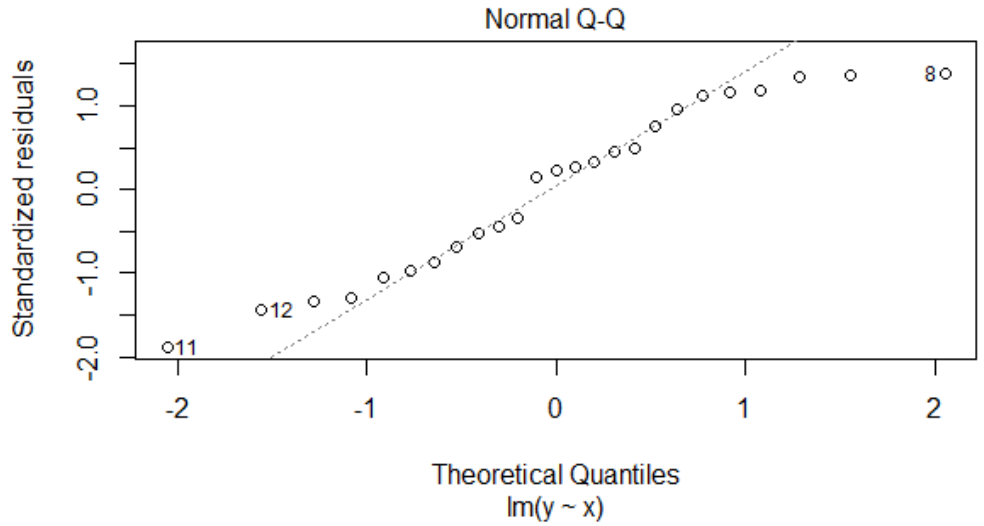

Figura 85 - Grafico Normal Q-Q

O gráfico Scale-Location (Figura 86) é semelhante ao gráfico Residual x Fitted, mas usa a raiz quadrada do valor absoluto dos resíduos padronizados ao invés do valor do próprio resíduo. A linha vermelha, quando horizontal, indica a perfeita ausência de variação.

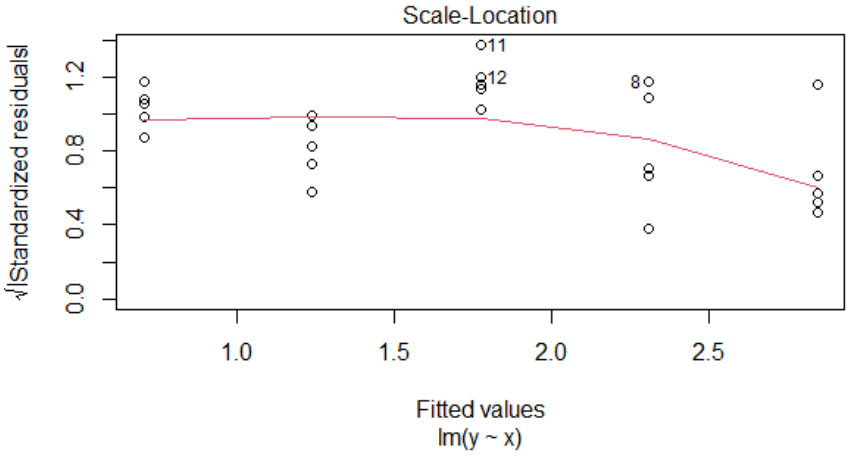

Figura 86 - Gráfico Scale - Location

O gráfico da Constante de Leverage (Figura 87) é útil para detectar a presença de pontos influenciadores.

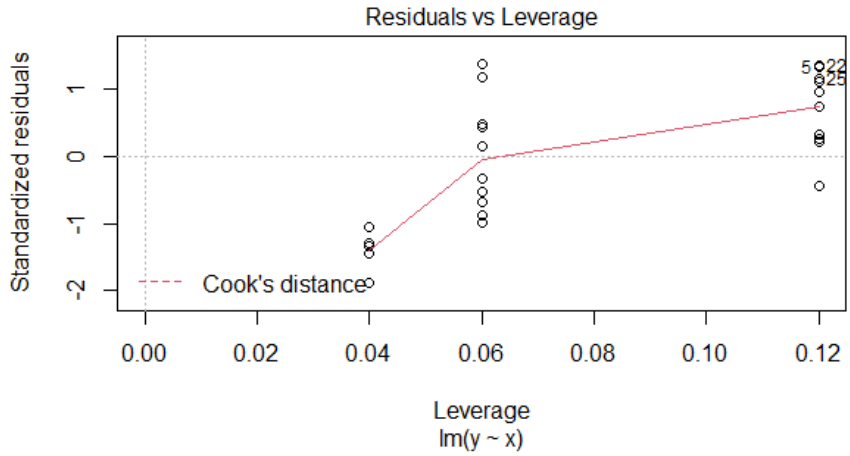

Figura 87 - Gráfico Constante de Leverage

Como dissemos anteriormente, na análise do gráfico de dispersão, o tratamento com 20% de adição de cinza de bagaço de cana foge um pouco da linearidade ideal da curva. Assim, podemos determinar a curva de regressão, somente com os tratamentos de 0% a 15%, e assim identificar as diferenças nos parâmetros α e β e verificar se o coeficiente de determinação (R^2) apresenta melhoria.

Vamos montar uma nova entrada de dados, excluindo os dados relativos ao tratamento com 20% de adição dos dados de entrada e reexecutar a análise estatística.

O coeficiente de determinação aumentou (de 0,9349 para 0,9589), indicando uma melhora na representatividade da curva da regressão linear. Também podemos comparar os valores médios das amostras de cada tratamento com os valores preditos pelas equações das curvas e analisar os resíduos (Tabela 60).

```
> dados = read.csv2(file.choose(), header = T)
> dados_lm = lm(y ~ x, data = dados)
> summary(dados_lm)

Call:
lm(formula = y ~ x, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2794 -0.1219  0.0206  0.1000  0.2794

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.951800   0.058564   50.40  < 2e-16 ***
x           -0.128240   0.006261  -20.48 6.38e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1565 on 18 degrees of freedom
Multiple R-squared:  0.9589,     Adjusted R-squared:  0.9566
F-statistic: 419.6 on 1 and 18 DF,  p-value: 6.376e-14
```

x_i	$\bar{Y}$	Y(eq.1)	Y(eq.2)	$\bar{Y} - Y(eq.1)$	$\bar{Y} - Y(eq.2)$
0	2,914	2,8468	2,9518	0,067	0,038
5	2,458	2,3106	2,3106	0,147	0,147
10	1,488	1,7744	1,6694	0,286	0,181
15	1,1	1,2382	1,0282	0,138	0,072
20	0,912	0,702	0,387	0,210	0,525

Tabela 60 - Valores preditos para Y

Como pode ser visto, a segunda equação apresenta valores um pouco mais próximos da média amostral do que a primeira, para os valores preditos de x_i de 0 a 15, mas a diferença é pequena. Também podemos observar que, para a segunda equação, o valor de Y_i para 20% de adição está bem distante da média amostral (extrapolação – cálculo para valor de x_i fora do intervalo de estudo). Isto deve-se ao fato da curva de regressão ter sido construída para o intervalo de 0 a 15, que representa a parte mais linear das médias amostrais.

O gráfico com os valores das médias amostrais e os valores preditos (Figura 88) também permite uma visualização da proximidade das curvas das regressões lineares.

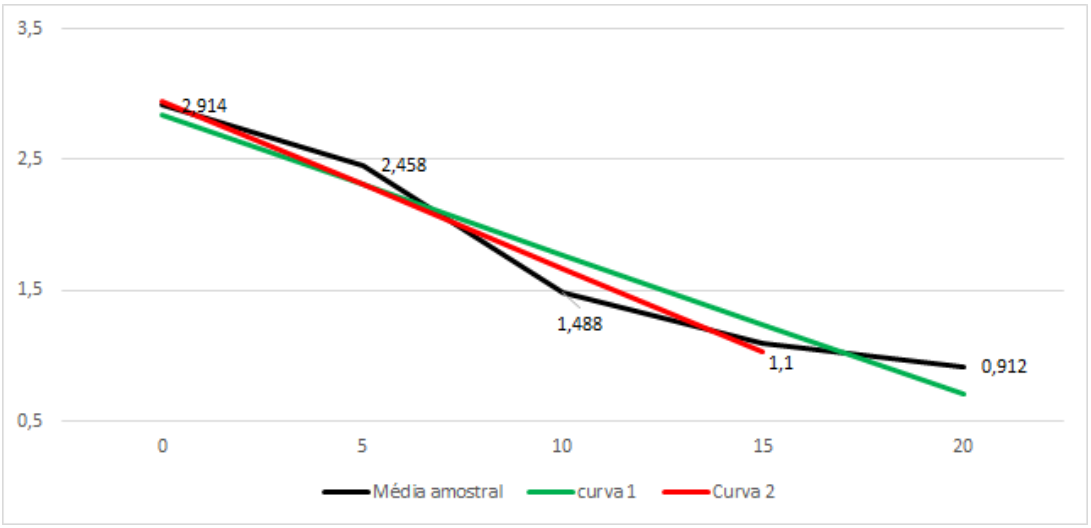

Figura 88 - Exemplo 25 - Gráfico com as médias amostrais e os valores preditos

Regressão Linear Múltipla

Agora finalmente podemos retomar o Exemplo 24, onde foi apresentado um experimento com dois fatores de dois níveis, ambos com influência na característica de interesse e com interação entre os fatores (que para ambos pode ser positiva ou negativa). A ANOVA nos confirmou que ambos os fatores e sua interação são significantes para a variável resposta e cuja equação que originou os dados (curva de regressão original $x = 25 + 60A - 5B + 55AB$) foi apresentada logo após a Tabela 54.

Vamos carregar os dados no RStudio e verificar o quão próximo à curva de regressão proposta é da curva original. Lembre-se que agora, devemos representar os valores dos fatores com seus valores reais e não como a+, a-, b+ e b-. Os valores dos níveis de A foram (0,1 / 0,2) e os níveis de B foram (1,0 / 2,0).

A fórmula da Regressão Linear Múltipla deve ser adequada para refletir a interação entre os fatores (variáveis independentes). Assim, usaremos a notação "res ~ a * b" ao invés de "res ~ a + b", que é usada quando temos certeza de que não há interações entre os fatores.

```
> dados = read.csv2(file.choose(), header = T)
> dados_lm = lm(res ~ a * b, data=dados)
> summary(dados_lm)

Call:
lm(formula = res ~ a * b, data = dados)

Residuals:
   Min     1Q Median     3Q    Max
-1.768 -1.016 -0.365  0.665  3.342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.430      4.254   5.977 6.44e-05 ***
a             56.700     26.907   2.107  0.05681 .
b             -5.417      2.691  -2.013  0.06706 .
a:b           58.050     17.018   3.411  0.00516 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.702 on 12 degrees of freedom
Multiple R-squared:  0.963,      Adjusted R-squared:  0.9537
F-statistic:   104 on 3 and 12 DF,  p-value: 7.434e-09
```

A aplicação da Regressão Linear Múltipla resultou na equação abaixo e para efeito de comparação, vamos repetir a equação original que foi usada para o cálculo da média predita para cada tratamento (considere que a média amostral não refletiu perfeitamente a média predita, devido a aleatorização dos valores dos elementos):

$$res = 25{,}43 + 56{,}7\ a - 5{,}417\ b + 58{,}05\ ab$$ - Equação da curva de regressão

$$x = 25{,}00 + 60{,}00\ a - 5{,}00\ b + 55{,}00\ ab$$ - Equação original

A equação da curva de regressão apresentada confirma a premissa apresentada anteriormente, da influência positiva do fator a, negativa para o fator b e positiva para a interação entre os fatores a e b (positiva e superior a influência negativa do fator b).

Tendo-se em conta que os valores de cada tratamento foram gerados aleatoriamente (quatro elementos por tratamento) e a média amostral não reflete exatamente o valor predito determinado pela equação original, podemos dizer que a Regressão Linear determinou com a maior exatidão possível a equação da curva de regressão. Os coeficientes obtidos estão extremamente próximos dos usados na equação original e a Regressão Linear foi capaz de determinar com precisão o tipo de contribuição de cada fator (e da interação) para a resposta.

Se apresentarmos em uma tabela (Tabela 61), os valores de a, b, da média amostral (MA), dos valores calculados pela equação original X (ori) e os valores preditos pela curva de regressão RES(rlm), veremos o quão próximo eles são:

a	b	MA	X(ori)	RES(rlm)
0,1	1	31,49	31,50	31,49
0,1	2	31,88	32,00	31,88
0,2	1	42,96	43,00	42,96
0,2	2	49,15	49,00	49,16

Tabela 61 - Valores calculados e preditos pela Regressão Linear Múltipla

O gráfico da Figura 89 praticamente sobrepõe as curvas que representam os valores originais tomados como base para as quatros amostra de 4 elementos representados por sua média amostral, os valores calculados pela equação original e os valores preditos pela curva de regressão linear.

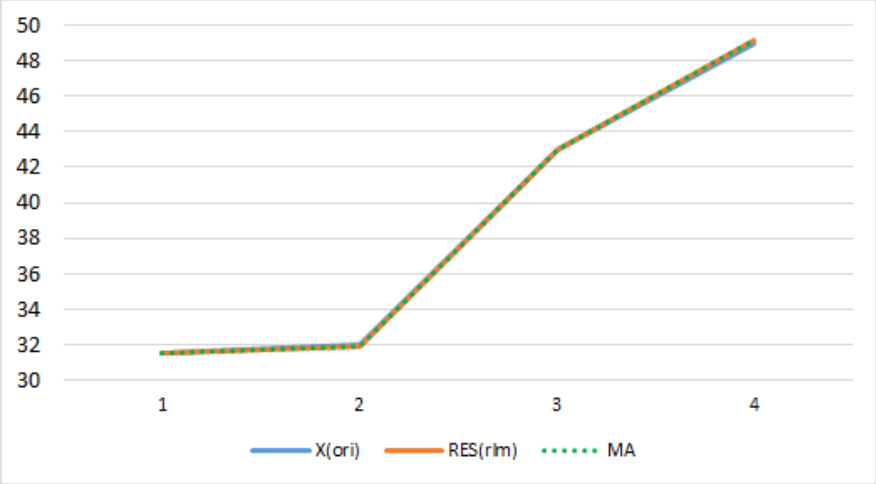

Figura 89 - Gráfico com as curvas original, de regressão e média amostral

Podemos ver que as curvas não são bem lineares e sim formadas por segmentos de reta unindo os pontos referenciados dos quatro tratamentos. E, já que temos a equação original e a equação dada pela Regressão, podemos inserir mais pontos, para ver o formato real da curva, mostrado na Figura 90.

Nas curvas da Figura 90 inserimos no eixo x a indicação de extrapolação (E) quando os valores calculados pela curva de regressão estão fora do intervalo de estudo e de predição (P) quando os valores calculados estão dentro do intervalo de estudo.

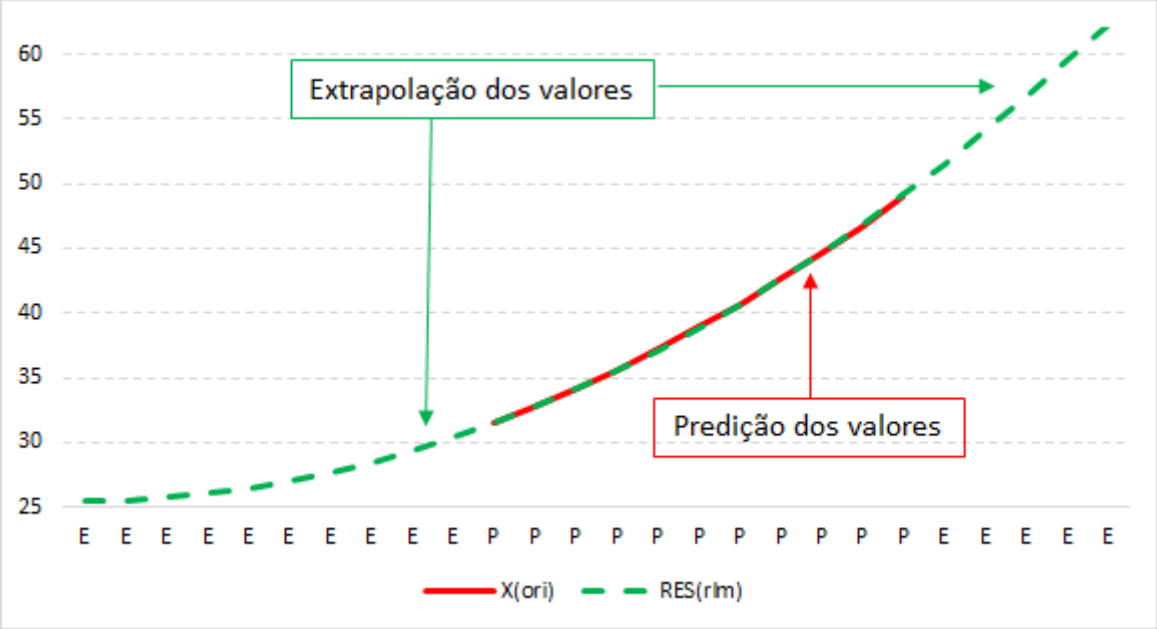

Figura 90 - Curvas original e da regressão linear múltipla, com predição e extrapolação de valores

Agora, devemos ressaltar que este é um **exemplo teórico** e, portanto, não foi influenciado por outros fatores. Os valores dos elementos das amostras foram aleatorizados mas ajustados para refletir, da melhor forma possível, a média amostral desejada.

Em um experimento real, dificilmente conseguiríamos uma situação assim. Diversos fatores não previstos (os ditos fatores aleatórios) iriam influenciar o experimento, tais como:

- Diferenças de dosagem das matérias primas.
- Fadiga de equipamento (tanto na mensuração quanto no preparo).
- Fatores não controláveis, como temperatura, pressão, umidade e outros.
- Falta de planejamento do experimento, cansaço ou desatenção do pesquisador
- E muitos outros.

Além disto, o exemplo teórico não considera fatos que normalmente aconteceriam em um experimento, como a **saturação**, que ocorre quando o aumento na adição de um componente não influencia mais o resultado ou assume comportamento contrário ao anterior (passa a influenciar negativamente ao invés de positivamente).

Assim, este exemplo deve ser visto apenas como explicativo para o poder da Regressão Linear Múltipla em representar e facilitar a análise da influência dos fatores e de sua interação na característica de interesse, ou seja, a resposta do experimento.

REFERÊNCIAS

ALVES, M. C. **Teste t de Student**. Seção Técnica de Informática. Piracicaba. 2017

BARBETTA, P. A. **Estatística Aplicada às Ciências Sociais**. 8. Florianópolis, SC: 2012. ISBN 978-85-328-0604-8.

FARIAS, A. M. L. D.; DEMARQUI, F. N. **Análise de Variância de um Fator**. 2017

FUKUCHI, R. K. **Análise de Variância (ANOVA)**. RStudio Pubs 2019.

GREENWOOD, M.; BANNER, K. **ANOVA model diagnostics including QQ-plots**. Statistics with R: Creative Commons 2015a.

_____. **Histograms, boxplots, and density curves**. Statistics with R: Creative Commons 2015b.

_____. **Multiple (pair-wise) comparisons using Tukey's HSD and the compact letter display - Statistics with R**. Statistics with R: Creative Commons 2015c.

_____. **Summary of importance R-code**. Statistics with R: Creative Commons 2015d.

GUIMARÃES, A. M. **Análise de Variância (ANOVA) one-way e Tukey usando R**. Medium. California, US: A Medium Corporation 2019.

MINITAB, L. **Entendendo Análise de Variância (ANOVA) e o teste F**. Editor Minitab: Minitab 2019a.

_____. **Interpretar os principais resultados para ANOVA para 1 fator**. Editor Minitab: Minitab 2019b.

_____. **Quais são os erros do tipo I e II?** Editor Minitab: Minitab 2019c.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. 6. Rio de Janeiro, RJ: John Wiley & Sons, Inc., 2016. 629 ISBN 130978-11-185-3971-2.

PANOSSO, A. R.; MALHEIROS, E. B. **Estatística Experimental Aplicada - Software R**. Jaboticabal, SP: FCAV / UNESP - Campus de Jaboticabal.

Portal Action. 2020. Disponível em: < http://www.portalaction.com.br/ >.

PORTALACTION. Portal Action. São Carlos - SP, 2020. Disponível em: < http://www.portalaction.com.br/ >.

PROVETE, D. B. **Intervalo de confiança z.test e t.test**. RPubs: RPubs 2017.

REIS, M. M. **Simulação e Cálculo do Poder do Teste e de Tamanho de Amostra para Testes no aplicativo R**. INE6006 - Procedimentos. Florianópolis:

RODRIGUES, É. C. **Modelos de Regressão Linear Simples**. 2016a

_____. **Modelos de Regressão Linear Simples - Análise de Resíduos**. 2016b

_____. **Modelos de Regressão Múltipla**. 2016c

SIMON, L.; YOUNG, D.; PARDOE, I. **STAT 462**. Applied Regression Analysis: The Pennsylvania State University 2019.

TRIOLA, M. F. **Introdução à estatística**. LTC Rio de Janeiro, 2005.

WALPOLE, R. E.; MYERS, R. H.; MYERS, S. L.; YE, K. **Probabilidade & Estatística para engenharias e ciências**. 8. São Paulo, SP: Pearson Prentice Hall, 2009. 491 ISBN 978-85-7605-199-2.