

DEFINIÇÕES BÁSICAS

- Estatística:** estudo dos métodos para coletar, organizar, apresentar e analisar dados.
- Estatística descritiva:** procedimentos usados para organizar e apresentar dados de forma conveniente e comunicativa.
- Inferência estatística:** procedimentos empregados para chegar a grandes conclusões ou inferências sobre populações com base em dados amostrais.
- População:** conjunto completo de observações reais ou potenciais sobre as quais se fazem inferências.
- Amostra:** subconjunto da população selecionado de acordo com um método de amostragem.
- Métodos de amostragem**
 - **Amostra por grupos/cluster:** uma população é dividida em grupos chamados clusters; alguns deles são selecionados aleatoriamente, e cada membro participante é observado.
 - **Amostragem por estrato/camada:** a população é dividida em estratos e um número fixo de elementos de cada estrato é selecionado para a amostra.
 - **Amostra aleatória simples:** uma amostra selecionada de tal forma que cada membro possível da amostra tenha igual oportunidade de ser selecionado; usada em muitas inferências elementares.
- Variável:** atributo de elementos de uma população ou amostra que pode ser medido. Ex.: altura, peso, QI, cor de cabelos e pulsação cardíaca são algumas das muitas variáveis que podem ser medidas em pessoas.

Dados: valores de variáveis observadas.**Tipos de dados**

- Dados qualitativos (ou “categóricos”) são descriptivos porém não numéricos. Ex.: sexo, lugar de nascimento, cor do veículo possuído.
- Dados quantitativos assumem valores numéricos.
- Dados discretos possuem números contáveis (0, 1, 2, ...) como valores. Ex.: o número de pulgas num cachorro, o número de vezes que um professor se atrasa durante um semestre.
- Dados contínuos podem se encontrar em uma faixa de valores, não apenas em valores contáveis. Ex.: altura de crianças, peso de um pacote de feijão, quantidade de vezes que um professor se atrasa.

Níveis de medição

- **Dados qualitativos** podem ser medidos em:
 - **nível nominal:** os valores são apenas nomes, sem nenhuma ordem. Ex.: cor de carro, carreira em faculdade.
 - **nível ordinal:** os valores apresentam uma ordem natural. Ex.: ano da graduação (1º, 2º, 3º, etc.), posto na hierarquia militar.
- **Dados quantitativos** podem ser medidos em:
 - **nível intervalar:** dados numéricos sem ponto zero natural; os intervalos (diferenças) são significativos, mas as razões não são. Ex.: temperatura em graus Fahrenheit; 80°F é 20°F mais quente do que 60°F, mas não é 150% tão quente.
 - **nível racional:** dados numéricos para os quais há um zero verdadeiro; tanto os intervalos como as razões são significativos. Ex.: peso, comprimento, duração, em sua maioria, propriedades físicas.

Estatística: medida numérica resultante de dados amostrais, usada para descrever a amostra e estimar o parâmetro populacional correspondente.

Parâmetro: medida numérica que descreve uma população; em geral os parâmetros não são computados, mas sim inferidos com base em estatística por amostragem.

DISTRIBUIÇÃO DE FREQUÊNCIA

Mostra a frequência (número de vezes observadas) de cada valor de uma variável.

- Tabela nº 1:** alunos de um grupo de uma autoescola, classificados de acordo com os números de acidentes ocorridos:

(nº de acidentes)	(frequência)	(frequência relativa)
<i>x</i>	<i>f</i>	<i>FR</i>
5	3	0,0526
4	2	0,0351
3	9	0,1579
2	15	0,2632
1	16	0,2807
0	12	0,2105

Distribuição de frequência agrupada: os valores da variável são agrupados em classes.

- Tabela nº 2:** as notas de um exame semestral são agrupadas em classes:

Classe	<i>f</i>	frequência acumulada
90-99	4	80
80-89	18	76
70-79	31	58
60-69	19	27
50-59	7	8
40-49	1	1

DISTRIBUIÇÕES DE FREQUÊNCIA ACUMULADA E DE PORCENTAGEM ACUMULADA

Distribuição de frequência relativa: cada frequência é dividida pelo número total de observações para produzir a proporção ou porcentagem do conjunto de dados que apresenta aquele valor. Ex.: terceira coluna da tabela 1.

Distribuição de frequência acumulada: as frequências contam todas as observações em um valor ou classe específico e todos os valores menores. Ex.: terceira coluna da tabela 2.

MEDIDAS DE TENDÊNCIA CENTRAL

Média: medida de tendência central mais usada; sensível a valores extremos.

MÉDIA DA POPULAÇÃO MÉDIA AMOSTRAL

$$\mu = \frac{1}{N} \sum_{i=1}^n x_i \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Média aparada/podada: calculada descartando-se alguns números entre os valores mais altos e mais baixos; é menos sensível do que a média comum.

Média ponderada: calculada multiplicando-se um peso por cada valor, fazendo com que alguns valores influenciem mais fortemente a média do que outros.

Mediana: valor que divide o conjunto de dados ordenados de tal modo que o mesmo número de observações ocorra em cada um dos lados; é uma medida menos sensível aos valores extremos. Para um número ímpar de dados, a mediana é o valor central; para um número par de dados a mediana é a média entre os dois valores centrais. Ex.: na tabela 1 a mediana é indicada pelo 29º elemento que é o valor 2.

Moda: observação que ocorre com a maior frequência. Ex.: na tabela 1, a moda é 1.

MEDIDAS DE DISPERSÃO

Somas de quadrados (SQ): soma dos desvios ao quadrado com base na média:

$$\text{Da população} = \sum (x_i - \mu_x)^2 \text{ ou } \sum x_i^2 - \frac{(\sum x_i)^2}{N}$$

$$\text{Da amostragem} = \sum (x_i - \bar{x})^2 \text{ ou } \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Variância: média das diferenças ao quadrado entre observações e suas médias.

$$\text{Variância da população: } \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

$$\text{Variância da amostragem: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Variâncias para dados agrupados:

$$\text{População: } \sigma^2 = \frac{1}{N} \sum_{i=1}^G f_i (m_i - \mu)^2$$

$$\text{Amostra: } s^2 = \frac{1}{n-1} \sum_{i=1}^G f_i (m_i - \bar{x})^2$$

Desvio padrão: raiz quadrada da variância; ao contrário da variância, apresenta as mesmas unidades dos dados originais e é geralmente mais usada.

$$\text{Ex.: desvio padrão da população } \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

Pontuação padrão: também conhecida como pontuações-Z; a pontuação padrão de um valor é o número direto/direcionado de desvios padrão com base na média na qual o valor foi encontrado; ou seja, $z = \frac{x - \mu}{\sigma}$.

Uma pontuação-z positiva indica um valor maior do que a média; uma pontuação-z negativa indica um valor menor do que a média; uma pontuação-z igual a zero indica o valor da média.

A conversão de todos os valores de um conjunto de dados ou distribuição em pontuação-z é denominada padronização; quando um conjunto de dados ou distribuição é padronizado, apresenta uma nova média $\mu = 0$, e um novo desvio padrão $\sigma = 1$.

TÉCNICAS COM GRÁFICOS

Gráfico de barras: representação gráfica que usa barras para indicar a frequência da ocorrência de observações.

Histograma: representação gráfica de barras usado com variáveis quantitativas, contínuas.

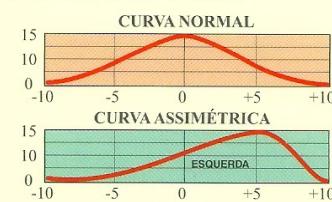
Curva de frequência: representação gráfica que mostra uma distribuição de frequência na forma de uma linha contínua que segue um histograma.

Curva de frequência acumulada: linha contínua que representa um histograma em que as barras de todas as classes inferiores estão empilhadas na classe superior adjacente. Não pode apresentar uma inclinação negativa.

Curva simétrica: a curva da frequência não se altera se sofrer rotação ao redor de seu centro; mediana = média.

Curva normal: curva em forma de sino, simétrica.

Curva assimétrica: parte de uma simetria e desvia-se em uma das extremidades, à esquerda ($\text{média} < \text{mediana}$) ou à direita ($\text{média} > \text{mediana}$).



PROBABILIDADE

Uma medida da possibilidade de um evento aleatório; a frequência relativa de longo prazo na qual um evento ou acontecimento ocorre:

Probabilidade de ocorrência de um evento A

$$p(A) = \frac{\text{Número de ocorrências que favorecem o evento A}}{\text{Número total de ocorrências}}$$

- Espaço amostral** – todas as possíveis ocorrências simples de um experimento.

Relações entre os eventos

- Exaustivo**: dois ou mais eventos recebem o nome de exaustivos se representarem todas as ocorrências possíveis.

Simbolicamente, $P(A \cup B \cup \dots) = 1$.

- Não exaustivo**: dois ou mais eventos recebem o nome de não exaustivos se não esgotarem todas as ocorrências possíveis.

- Mutuamente exclusivos**: eventos que não podem ocorrer simultaneamente: $P(A \cap B) = 0$; e $P(A \cup B) = P(A) + P(B)$. Ex.: homens, mulheres.

- Não mutuamente exclusivos**: eventos que podem ocorrer simultaneamente: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Ex.: homens, olhos castanhos.

- Independente**: eventos cujas probabilidades não são afetadas pela ocorrência ou não-ocorrência um do outro: $P(A|B) = P(A)$; $P(B|A) = P(B)$; e $P(A \cap B) = P(A)P(B)$. Ex.: sexo (masculino/feminino) e cor de olhos.

- Dependente**: eventos cujas probabilidades mudam dependendo da ocorrência ou não-ocorrência um do outro: $P(A|B)$ difere de $P(A)$; $P(B|A)$ difere de $P(B)$; e $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$. Ex.: raça e cor de olhos.

- Probabilidades conjuntas**: probabilidade de que dois ou mais eventos ocorram simultaneamente.

- Probabilidades disjuntas ou incondicionais**: soma de probabilidades.

- Probabilidades condicionais**: probabilidade de A dada a existência de S , ou seja: $P(A|S)$.

- Exemplo**: Dados os números de 1 a 9 como observações em um espaço amostral:

- Eventos mutuamente exclusivos e complementares – Ex.: $P(\text{todos os números ímpares})$; $P(\text{todos os números pares})$.
- Eventos mutuamente exclusivos mas não complementares – Ex.: $P(\text{um número par})$; $P(\text{os números 7 e 5})$.
- Eventos não mutuamente exclusivos e não exaustivos – Ex.: $P(\text{um número par ou um 2})$.

TABELA DE FREQUÊNCIA

	Evento C	Evento D	TOTAL
Evento E	52	36	87
Evento F	62	71	133
Total	114	106	220

Ex.: probabilidade conjunta entre C e E

$$p(C \cap E) = 52/220 = 0,24$$

TABELA DE PROBABILIDADE CONJUNTA, DISJUNTA E CONDICIONAL

	Evento C	Evento D	Probabilidade disjunta	Probabilidade condicional
Evento E	0,24	0,16	0,40	(C/E)=0,60 (D/E)=0,40
Evento F	0,28	0,32	0,60	(C/F)=0,47 (D/F)=0,53
Probabilidade disjunta	0,52	0,48	1,00	
Probabilidade condicional	(E/C)=0,46 (F/C)=0,54	(E/D)=0,33 (F/D)=0,67		

Distribuição amostral: distribuição teórica de probabilidades de uma estatística que será resultado da extração de todas as amostras possíveis de um tamanho de uma população.

VARIÁVEIS ALEATÓRIAS

- Uma **variável aleatória** usa valores numéricos aleatoriamente com probabilidades específicas por **uma função (ou densidade) de distribuição de probabilidade**.

- Variáveis aleatórias discretas**: apenas com valores distintos (assim como dados quantitativos).

- Distribuição binomial**: um modelo para o número (x) de sucessos em uma série de n tentativas independentes em que cada tentativa resulta em sucesso com probabilidade p , ou falha com probabilidade $1 - p$. Ex.: o número (x) de caras ("sucesso") obtidos em 12 (n) disputas imparciais de "cara ou coroa" (probabilidade de cara = $p = 0,5$).

$P(x) = {}_nC_x p^x (1-p)^{n-x}$ em que $P(x)$ é a probabilidade de exatos x sucessos em n tentativas com uma probabilidade constante p de sucesso em cada tentativa;

$${}_nC_x = \frac{n!}{(n-x)!x!}$$

– **Média binomial**: $\mu = np$

– **Variância binomial**: $\sigma^2 = np(1-p)$

– À medida que n aumenta, a distribuição binomial aproxima-se da distribuição normal.

- Distribuição hipergeométrica**:

– Representa o número de sucessos em uma série de n tentativas em que cada tentativa resulta em sucesso ou falha.

– Como a binomial, com a diferença de que cada tentativa é retirada de uma pequena população com N elementos divididos entre N_1 sucessos e N_2 falhas.

– Portanto, a probabilidade de dividir as n tentativas entre x_1 sucessos e x_2 falhas é:

$$P(x_1 \text{ e } x_2) = \frac{N_1!}{x_1!(N_1 - x_1)!} \frac{N_2!}{x_2!(N_2 - x_2)!} \frac{N!}{n!(N - n)!}$$

– **Média hipergeométrica**: $\mu_1 = E(x_1) = \frac{nN_1}{N}$

– **Variância hipergeométrica**: $\sigma^2 = \frac{N - n}{N - 1} \left[\frac{nN_1}{N} \right] \left[\frac{N - n - N_1}{N - 1} \right]$

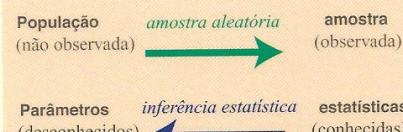
- Distribuição de Poisson** – um modelo para o número de ocorrências de um evento $x = 0, 1, 2, \dots$, contados em um intervalo fixo de espaço e tempo em vez de um número fixo de tentativas; o parâmetro é o número médio de ocorrências, λ , para $x = 0, 1, 2, 3, \dots$ e > 0 ; caso contrário: $P(x) = 0$.

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{Média e variância de Poisson: } \lambda.$$

INFERÊNCIA ESTATÍSTICA

- Para se fazer inferências sobre uma população não observada, retira-se uma amostra aleatória.

- A amostra é usada para calcular as estatísticas, que são usadas para tirar conclusões sobre a probabilidade de parâmetros da população.



ESTIMAÇÃO VIESADA E NÃO VIESADA

- Estimação não viésada de um parâmetro**: um estimador (estatística amostral) com um valor médio igual ao valor do parâmetro. Ex.: a média amostral é um estimador não viésado da média populacional; se todos os outros fatores forem iguais, prefere-se um estimador não viésado em relação a um estimador viésado.

- Estimação viésada de um parâmetro**: um estimador (estatística amostral) que não é igual à média do valor do parâmetro. Ex.: a mediana é um estimador viésado, pois a média das medianas amostrais não é sempre igual à mediana populacional; a variância amostral calculada, dividida por n , é um estimador viésado da variância populacional; entretanto, quando calculada com $n-1$, o estimador torna-se não viésado.

- Observe**: os estimadores propriamente ditos apresentam apenas uma fonte de viés: mesmo quando um estimador não viésado é usado o viés ainda estará presente na amostra (elementos não totalmente iguais poderão ser escolhidos).

- Distribuição amostral**: a distribuição de probabilidade de uma estatística amostral que será resultado da retirada de todas as amostras possíveis de um dado tamanho de uma população; como as amostras são retiradas aleatoriamente, **cada estatística amostral é uma variável aleatória**, e tem uma distribuição de probabilidade que pode ser demonstrada por meio da média e do desvio padrão.

- Erro padrão**: o desvio padrão do estimador; **não confunda com o desvio padrão da amostra propriamente dita**; mede a variabilidade das estimativas ao redor de seu valor esperado, ao passo que o desvio padrão da amostra reflete a variabilidade dentro da amostra ao redor da média amostral.

DESVIO PADRÃO DA MÉDIA

- O desvio padrão de todas as médias amostrais possíveis de um dado tamanho amostral, retirado de uma mesma população, é denominado **erro padrão da média amostral**.

- Se o desvio padrão populacional σ for conhecido, o erro padrão é: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

- Em geral, o desvio padrão populacional σ é desconhecido, e é estimado por s ; nesses casos, o erro padrão **estimado** é: $\sigma_{\bar{x}} \approx s_{\bar{x}} = \frac{s}{\sqrt{n}}$.

- Observe**: em qualquer um dos casos, o erro padrão da média amostral diminui à medida que o tamanho da amostra aumenta – uma amostragem maior fornece informação mais confiável sobre a população.

VARIÁVEIS ALEATÓRIAS CONTÍNUAS

- Variáveis aleatórias contínuas**: variáveis que podem assumir qualquer valor de um intervalo ininterrupto em uma linha numérica.

- As probabilidades são medidas apenas em intervalos, nunca em valores isolados; a probabilidade de que uma variável aleatória contínua esteja entre dois valores é exatamente igual à área sob a curva de densidade entre aqueles dois valores.

- Distribuição normal**: curva com formato de sino; uma distribuição cujos valores se reúnem simetricamente em torno da média (assim como a mediana e a moda); é comum e importante ao se fazer inferências.

– A curva de densidade é o gráfico de:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x - \mu)^2 / 2\sigma^2}$$

em que $f(x)$ = frequência em um dado valor; σ = desvio padrão da distribuição normal; μ = a média da distribuição normal; e x = valor da variável da distribuição normal.

- Distribuição normal padronizada**: uma distribuição normal que apresenta média = 0 e desvio padrão = 1; os valores que seguem uma distribuição normal podem ser transformados em distribuição normal padronizada pelo uso de **pontuações-z** (veja *Medidas de Dispersão*, na pág. 1).

TESTANDO HIPÓTESES ESTATÍSTICAS

• Num teste de hipóteses, usa-se um dado amostral para aceitar ou rejeitar a **hipótese nula (H_0)** em favor de uma **hipótese alternativa (H_1)**; o nível de significância ao qual a hipótese nula pode ser rejeitada indica quanta evidência a amostra fornece contra a hipótese nula.

• **Hipótese nula (H_0): sempre especifica um valor (valor da hipótese nula)** para um parâmetro populacional; assume-se que a hipótese nula é verdadeira – essa pressuposição define os cálculos para o teste da hipótese.

Ex.: H_0 = uma moeda é imparcial, ou seja, a proporção de sair "cara" é 0,5; H_1 : $p \neq 0,5$.

• **Hipótese alternativa (H_1): nunca especifica um valor para um parâmetro:** a hipótese alternativa afirma que um parâmetro populacional apresenta um valor diferente do especificado na hipótese nula.

Ex.: H_1 = uma moeda é parcial, ou seja, a proporção de sair cara não é 0,5; H_1 : $p \neq 0,5$.

1. **Hipótese não direcional (ou bilateral):** uma hipótese alternativa (H_1) que declara apenas que o parâmetro da população é simplesmente *diferente* daquele especificado em H_0 , emprega-se um valor de probabilidade bilateral. Ex.: ao empregar um dado amostral para testar se a pulsação cardíaca da média populacional é diferente de 65, usariam-se o teste da hipótese não direcional H_0 : $\mu = 65$ versus H_1 : $\mu \neq 65$.

2. **Hipótese direcional (ou lateral):** uma hipótese alternativa H_1 que declara que o parâmetro populacional é maior (à direita) ou menor (à esquerda) do que o valor especificado em H_0 , emprega-se a probabilidade direcional. Ex.: ao empregar dado amostral para testar se a pulsação cardíaca da média populacional é maior do que 65, usariam-se o teste da hipótese direcional H_0 : $\mu = 65$ versus H_1 : $\mu > 65$.
• A hipótese alternativa H_1 também é denominada "hipótese de pesquisa", pois apenas as considerações expressas como hipóteses alternativas podem ser sustentadas positivamente.

• **Nível de significância:** a probabilidade de se observarem resultados amostrais tão ou mais extremos do que aqueles realmente observados, considerando-se que a hipótese nula seja verdadeira; se a probabilidade for pequena, concluímos que há provas suficientes para se rejeitar a hipótese nula; existem duas abordagens básicas:

1. **Nível de significância fixo (método tradicional):** predetermina-se um nível de significância α ; os níveis de significância comumente usados são 0,01, 0,05 e 0,10.

• **Quanto menor o nível de significância α maior o padrão para se rejeitar H_0 :** o valor crítico para a estatística teste é determinado de tal forma que a probabilidade da estatística teste esteja mais distante de zero do que o valor crítico (em um ou em ambos lados, dependendo de H_1) é α , se a estatística teste recai além do valor crítico – na região de rejeição – H_0 pode ser rejeitada naquele nível de significância α .

2. **Nível de significância observado (método do valor-p):** a estatística teste é calculada usando-se dado amostral, então a distribuição de probabilidade apropriada é usada para encontrar a probabilidade de se observar uma estatística amostral que seja **pelo menos um pouco** diferente do valor da hipótese nula para o parâmetro populacional (o valor da probabilidade ou valor-p); **quanto menor o valor-p, melhor a prova contra a H_0** .

• Esse método é mais usado em aplicações computadorizadas.

• O valor-p também representa o menor nível de significância ao qual a H_0 pode ser rejeitada; por-

tanto, os resultados dos valores-p podem ser usados com um nível de significância fixo **rejeitando a H_0 se o valor-p $\leq \alpha$** .

• Geralmente, quanto maior o valor da estatística teste (mais distanciado de zero, positivo ou negativo), menor o valor-p, o que fornece uma prova melhor contra a hipótese nula e a favor da alternativa.

- **Noção de prova indireta:** por meio de testes de hipóteses tradicionais, a **hipótese nula nunca poderá ser provada**. Ex.: se jogarmos uma moeda 200 vezes e aparecer coroa exatamente 100 vezes, não temos prova de que a moeda seja parcial, mas tampouco podemos provar que a moeda seja imparcial devido à natureza aleatória da amostragem – é possível tirar cara ou coroa 200 vezes e obter 100 caras, assim como também é possível retirar uma amostra de uma população com média de 104,5 e encontrar uma média amostral de 101; não conseguir rejeitar a hipótese nula não a torna verdadeira e rejeitando-a não provamos que ela seja falsa.

• Dois tipos de erro

- **Erro tipo 1** (erro tipo α): a rejeição da H_0 quando ela é realmente verdadeira. A probabilidade de um erro tipo 1 é dada pelo **nível de significância α** ; erro do tipo 1 é geralmente mais proeminente, pois pode ser controlado.
- **Erro tipo 2** (erro tipo β): a aceitação de H_0 quando ela é realmente falsa. A probabilidade de um erro tipo 2 é dada por β ; erro do tipo 2 é geralmente desconsiderado (o que consiste em um engano): é difícil medir ou controlar, pois β depende do valor real desconhecido do parâmetro em questão, também desconhecido.

HIPÓTESE ESTATÍSTICA		Situação verdadeira de H_0	
		H_0 verdadeira	H_0 falsa
Decisão:	H_0 aceita	Correta ($1-\alpha$)	Erro tipo 2 (β)
	H_0 rejeitada	Erro tipo 1 (α)	Correta ($1-\beta$)

TEOREMA DO LIMITE CENTRAL

(para média amostral \bar{x})

- Se $x_1, x_2, x_3, \dots, x_n$ for uma amostra aleatória simples de n elementos de uma população grande (infinita), com média $\mu(m)$ e desvio padrão σ ; a distribuição de \bar{x} segue a distribuição em forma de sino de uma variável aleatória normal à medida que n aumenta, e a distribuição de razão:

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

se aproxima da distribuição normal padrão enquanto n tende para o infinito. Na prática, uma aproximação normal seria aceitável para amostras que consistam de 30 ou mais observações.

INFERÊNCIA SOBRE MÉDIA POPULACIONAL COM USO DA ESTATÍSTICA-Z (σ CONHECIDO)

Exige que a amostra seja retirada de uma distribuição normal ou tenha um tamanho (n) de pelo menos 30.

- **Usada quando o desvio padrão σ é conhecido:** se σ for conhecido (tratado como uma constante, não aleatório) e as condições acima existirem, então a distribuição da média amostral segue uma distribuição normal, e a estatística teste z segue uma

distribuição normal padronizada: observe que na prática esse caso é raro, a distribuição t é mais comumente usada.

• A estatística teste é $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ em que μ = média populacional (conhecida ou sob a forma hipotética da H_0) e $\sigma/\sqrt{n} = \sigma/\sqrt{n}$.

• **Região crítica:** a porção da área sob a curva que inclui os valores de uma estatística que fornecem provas suficientes para a rejeição da hipótese nula.

– Os níveis de significância mais frequentemente usados são 0,01, 0,05 e 0,1. Para um teste unilateral usando-se a estatística Z corresponde a valores-z de 2,33, 1,65 e 1,21 respectivamente – valores positivos para um teste com cauda à direita, negativo para um teste com cauda à esquerda.

• Para um teste bilateral, a região crítica para $\alpha = 0,01$ é dividida em duas áreas externas iguais marcadas por valores-z de $|2,58|$; para $\alpha = 0,05$, os valores críticos de z são $|1,96|$, e para $\alpha = 0,10$, os valores críticos de z são $|1,65|$.

– **Exemplo 1:** dada uma população com $\sigma = 50$, uma amostra aleatória simples de $n = 100$ tomam-se valores com uma média amostral $\bar{x} = 255$; faça o teste usando o método do valor p H_0 : $\mu = 250$ versus H_1 : $\mu > 250$. Há provas suficientes para a rejeição da hipótese nula?

• Nesse caso, a estatística teste $z = (255 - 250)/(50/\sqrt{100}) = 1,00$.

• Observando a tabela A, a área dada para $z = 1,00$ é 0,3413. A área a sua direita é (pois H_1 é " $>$ ") este teste apresenta cauda à direita) 0,5 - 0,3413 = **0,1587** ou 15,87%.

• Este é o valor p: a probabilidade, se a H_0 for verdadeira (ou seja, $\mu = 250$), de obter uma média amostral igual a 255 ou maior; também representa o menor nível de significância α ao qual a H_0 pode ser rejeitada.

• E ainda, mesmo que H_0 seja verdadeira, a probabilidade de obter uma média amostral ≥ 255 a partir dessa população com um tamanho amostral de $n = 100$ é cerca de 16%, é bastante plausível que H_0 seja verdadeira – não há provas fortes para sustentar essa hipótese alternativa de que a média populacional seja maior do que 250 – portanto, **fracassamos em rejeitar a H_0** .

• Nem mesmo pode ser rejeitada ao nível mais baixo de significância de $\alpha = 0,10$, pois $0,1587 > 0,10$; lembre-se: isso não prova que a média populacional seja igual a 250; apenas não conseguimos acumular provas suficientes contra a pressuposição.

– **Exemplo 2:** retira-se uma amostra aleatória simples de tamanho $n = 25$ de uma população que segue uma distribuição normal com $\sigma = 15$; a média amostral \bar{x} é 95; use o método do valor-p para testar H_0 : $\mu = 100$ versus H_1 : $\mu \neq 100$. Há provas suficientes para a rejeição da pressuposição de que a média populacional seja 100 ao nível de significância α de 0,10? E ao $\alpha = 0,05$?

• Nesse caso, a estatística teste $z = (95 - 100)/(15/\sqrt{25}) = -5/3 = -1,67$.

• Como a curva normal é simétrica, podemos procurar uma pontuação z de 1,67 – o valor na tabela A é 0,4525, ou seja,

$$P(0 < z < 1,67) = P(-1,67 < z < 0) = 0,4525.$$

• Portanto,

$$P(z < -1,67) = P(z > 1,67) = 0,5 - 0,4525 = 0,0475.$$

• Como esse é um teste bilateral (H_1 : $\mu \neq 100$), o valor-p é o dobro desta área, ou 0,095.

• Como o valor-p = 0,095 < 0,10 = α , há provas suficientes para a rejeição da hipótese nula ao nível de significância α de 0,10, mas, no segundo caso, o valor-p = 0,095 > 0,05 = α , portanto o dado amostral não é forte o suficiente para rejeitar a hipótese nula ao nível mais alto de significância (0,05).

Tabela A
Áreas sob a curva normal

área da média até z 

Z	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	,0000	,0040	,0080	,0120	,0160	,0199	,0239	,0279	,0319	,0359
0,1	,0398	,0438	,0478	,0517	,0557	,0596	,0636	,0675	,0714	,0753
0,2	,0793	,0832	,0871	,0910	,0948	,0987	,1026	,1064	,1103	,1141
0,3	,1179	,1217	,1255	,1293	,1331	,1368	,1406	,1443	,1480	,1517
0,4	,1554	,1591	,1628	,1664	,1700	,1736	,1772	,1808	,1844	,1879
0,5	,1915	,1950	,1985	,2019	,2054	,2088	,2123	,2157	,2190	,2224
0,6	,2257	,2291	,2324	,2357	,2389	,2422	,2454	,2486	,2517	,2549
0,7	,2580	,2611	,2642	,2673	,2704	,2734	,2764	,2794	,2823	,2852
0,8	,2881	,2910	,2939	,2967	,2995	,3023	,3051	,3078	,3106	,3133
0,9	,3159	,3186	,3212	,3238	,3264	,3289	,3315	,3340	,3365	,3389
1,0	,3413	,3438	,3461	,3485	,3508	,3531	,3554	,3577	,3599	,3621
1,1	,3643	,3665	,3686	,3708	,3729	,3749	,3770	,3790	,3810	,3830
1,2	,3849	,3869	,3888	,3907	,3925	,3944	,3962	,3980	,3997	,4015
1,3	,4032	,4049	,4066	,4082	,4099	,4115	,4131	,4147	,4162	,4177
1,4	,4192	,4207	,4222	,4236	,4251	,4265	,4279	,4292	,4306	,4319
1,5	,4332	,4345	,4357	,4370	,4382	,4394	,4406	,4418	,4429	,4441
1,6	,4452	,4463	,4474	,4484	,4495	,4505	,4515	,4525	,4535	,4545
1,7	,4554	,4564	,4573	,4582	,4591	,4599	,4608	,4616	,4625	,4633
1,8	,4641	,4649	,4656	,4664	,4671	,4678	,4686	,4693	,4699	,4706
1,9	,4713	,4719	,4726	,4732	,4738	,4744	,4750	,4756	,4761	,4767
2,0	,4772	,4778	,4783	,4788	,4793	,4798	,4803	,4808	,4812	,4817
2,1	,4821	,4826	,4830	,4834	,4838	,4842	,4846	,4850	,4854	,4857
2,2	,4861	,4864	,4868	,4871	,4875	,4878	,4881	,4884	,4887	,4890
2,3	,4893	,4896	,4898	,4901	,4904	,4906	,4909	,4911	,4913	,4916
2,4	,4918	,4920	,4922	,4925	,4927	,4929	,4931	,4932	,4934	,4936
2,5	,4938	,4940	,4941	,4943	,4945	,4946	,4948	,4949	,4951	,4952
2,6	,4953	,4955	,4956	,4957	,4959	,4960	,4961	,4962	,4963	,4964
2,7	,4965	,4966	,4967	,4968	,4969	,4970	,4971	,4972	,4973	,4974
2,8	,4974	,4975	,4976	,4977	,4977	,4978	,4979	,4979	,4980	,4981
2,9	,4981	,4982	,4982	,4983	,4984	,4984	,4985	,4985	,4986	,4986
3,0	,4987	,4987	,4987	,4988	,4988	,4989	,4989	,4989	,4990	,4990

INFERÊNCIA SOBRE MÉDIA POPULACIONAL USANDO A ESTATÍSTICA T (σ DESCONHECIDO)

Exige que a amostra seja retirada de uma distribuição normal ou tenha um tamanho (n) de pelo menos 30.

- Quando o desvio padrão σ é desconhecido – e em geral esse é o caso mais frequente – ele é estimado a partir de s , o desvio padrão amostral.
- Devido à variabilidade de ambas as estimativas – a média amostral e o desvio padrão amostral – a estatística teste segue uma distribuição t, e não uma distribuição z.
- Comparação entre as distribuições t e z
 - Embora ambas as distribuições sejam simétricas em torno de uma média zero, a distribuição t é mais espalhada do que a distribuição normal, o que produz um valor crítico maior de t como limite para a região de rejeição.
 - A distribuição t é caracterizada por seus **graus de liberdade** (gl), que se referem ao número de valores que são livres para variar após estabelecerem-se algumas restrições aos dados.
 - Por exemplo, se sabermos que uma amostra de tamanho 4 produz uma média de 87, sabemos que a soma dos números é $4 * 87 = 348$; isso não nos diz nada sobre os valores individuais da amostra – há números infinitos de formas para se obter 4 números que somem 348; mas quando escolhemos três deles, o **quarto é determinado**.
 - Por exemplo, o primeiro número pode ser 84, o segundo 98 e o terceiro 81; mas se os primeiros três números forem 84, 98 e 81, então o quarto tem de ser 85, o único número que produzirá a média amostral conhecida – ou seja, existe $n-1$ ou 3 graus de liberdade nesse exemplo.
- Para um teste sobre uma média populacional, a estatística t segue uma distribuição t com $n-1$ gl.
- À medida que o gl diminui, a distribuição t aproxima-se da distribuição z normal padronizada.
- A estatística teste t usada para testar hipóteses sobre uma média populacional é:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \text{ em que } \mu = \text{média populacional na } H_0 \text{ e } s_x = \frac{s}{\sqrt{n}}.$$

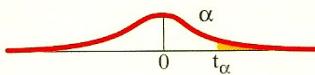
Observe: não é tão diferente da estatística teste z usada quando o σ é conhecido!

- Ex.: uma amostra aleatória simples de tamanho 25 é retirada de uma população que segue uma distribuição normal, com uma média amostral de 42, e desvio padrão amostral de 7,5; faça o teste ao nível de significância fixo $\alpha = 0,05$: $H_0: \mu = 45$ versus $H_1: \mu > 45$.

- Esse teste apresenta cauda à esquerda ($H_1: \mu > 45$), portanto o valor crítico e a região de rejeição serão negativos.
- Consultando a tabela B para encontrar o valor crítico apropriado, com $gl = n - 1 = 24$, produz-se um valor crítico de -1,711; a hipótese nula pode ser rejeitada ao $\alpha = 0,05$ se o valor da estatística teste $t < -1,711$.
- A estatística teste $t = (42 - 45) / (7,5/\sqrt{25}) = -3/1,5 = -2$; como isso é menor do que o valor crítico de -1,711, a H_0 é rejeitada ao $\alpha = 0,05$.

Tabela B
Valores críticos de t

Os valores indicam a área à direita de t_α



A*	0,1	0,05	0,025	0,01	0,005
B*	0,2	0,1	0,05	0,02	0,01
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
inf	1,282	1,645	1,960	2,326	2,576

A* = Nível de significância para teste unilateral.

B* = Nível de significância para teste bilateral.

Observe: a distribuição t é uma alternativa **consistente** para a distribuição z quando a média populacional estiver sendo testada: as inferências provavelmente são válidas mesmo nos casos em que a distribuição populacional estiver longe de ser normal; entretanto, quanto maior o ponto de partida da normalidade na população, maior o tamanho da amostra necessária para um teste de hipótese válido usando qualquer um dos tipos de distribuição.

INTERVALOS DE CONFIANÇA

Intervalo de confiança: um intervalo dentro do qual um parâmetro populacional é provável de ser encontrado; determinado por dado amostral e um **nível de confiança** escolhido ($1 - \alpha$ – [α se refere ao nível de significância]).

- Os níveis de confiança comuns são 90%, 95% e 99%, assim como os níveis de significância são 0,10, 0,05 e 0,01.
- ($1 - \alpha$) intervalo de confiança para μ :

$$\bar{x}_i - z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) \leq \mu \leq \bar{x} + z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) \text{ em que } z_{\alpha/2} \text{ é o valor da variável } z \text{ normal padronizada que produz uma área } \alpha/2 \text{ em cada cauda da distribuição.}$$

- Uma estatística t deve ser usada no lugar da estatística z quando σ for desconhecido e s deve ser usado como uma estimativa

- Exemplo:** dado $\bar{x} = 108$, $s = 15$ e $n = 26$, estime o intervalo de confiança de 95% para a média populacional.
 - Como não se conhece a variância populacional, usa-se a distribuição t.
 - O intervalo resultante, usando um valor t de 2,060 da tabela B linha 25 da coluna do meio, é aproximadamente de 102 a 114.
 - Consequentemente, qualquer hipótese nula de que μ esteja entre 102 e 114 é sustentável baseado nessa amostra.
 - Qualquer μ hipotético abaixo de 102 ou acima de 114 será rejeitado ao nível de significância 0,05.

COMPARAÇÃO ENTRE MÉDIAS POPULACIONAIS

Distribuição amostral da diferença entre médias:

se um número de pares de amostras for tirado da mesma população ou de duas populações diferentes, então:

- a distribuição das diferenças entre pares de médias amostrais tende a ser normal (distribuição z);
- a média dessas diferenças entre médias $\mu_{\bar{x}_1 - \bar{x}_2}$ é igual à diferença entre as médias populacionais, ou seja, $\mu_1 - \mu_2$.

Amostras independentes

– Estamos testando se duas amostras foram retiradas de populações com a mesma média, ou seja, $H_0: \mu_1 = \mu_2$, versus uma alternativa lateral ou bilateral.

– Quando σ_1 e σ_2 são conhecidos, a estatística teste z segue uma distribuição normal padronizada sob a hipótese nula.

– O erro padrão da diferença entre as médias ($\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{(\sigma_1^2)/n_1 + (\sigma_2^2)/n_2}$).

– Na qual ($\mu_1 - \mu_2$) representa a diferença hipotética em médias, a seguinte estatística pode ser usada para testes de hipóteses:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)(\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}.$$

– Quando σ_1 e σ_2 são desconhecidos, o que representa o caso mais frequente, substitua s_1 e s_2 por σ_1 e σ_2 respectivamente nas fórmulas acima e use a distribuição t com $gl = n_1 + n_2 - 2$.

Teste t para um n pequeno (pequenas amostras)

– Ambas as populações apresentam distribuição normal.

– $n < 30$.

– Exige homogeneidade na variância: σ_1 e σ_2 são desconhecidos mas considerados iguais – *consideração arriscada*.

– Muitos estatísticos não recomendam a distribuição t com erros padrão em amostras pequenas, a abordagem acima é mais conservadora.

♦ O teste de hipótese pode apresentar duas caudas (= versus ≠) ou uma cauda: $\mu_1 \leq \mu_2$ e a alternativa é $\mu_1 > \mu_2$ (ou $\mu_1 \geq \mu_2$ e a alternativa é $\mu_1 < \mu_2$).

♦ Graus de liberdade (gl):

$$(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2.$$

♦ Use a fórmula dada abaixo para estimar o $\sigma_{\bar{x}_1 - \bar{x}_2}$ e determinar $s_{\bar{x}_1 - \bar{x}_2}$.

♦ Determine a região crítica para a rejeição dando um nível de significância aceitável e considerando a tabela t com grau de liberdade = $n_1 + n_2 - 2$.

♦ Use a seguinte fórmula para o erro padrão estimado:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \left[\frac{n_1 + n_2}{n_1 n_2} \right]$$

• Pares combinados: ao se fazerem medições repetidas dos mesmos elementos, podemos testar a diferença da média.

– Por exemplo, clientes de um programa de perda de peso antes e depois do programa, e uma diferença de média significativa atribuída à sua eficácia.

Erro padrão da diferença de média

$$\text{Fórmula geral: } s_d = s_{\bar{x}} = \frac{s_d}{\sqrt{n}}$$

em que s_d é o desvio padrão das diferenças:

$$s_d = \sqrt{\frac{\sum(d^2)}{n} - \frac{(\sum d)^2}{n^2}}$$

– Podemos testar $H_0: \mu_d = 0$ versus uma alternativa lateral ou bilateral usando uma estatística teste t.

COMPARAÇÃO DE VARIÂNCIAS

- A heterogeneidade das variâncias (um critério para teste t de duas amostras): a condição de que as variâncias de duas populações são iguais; para estabelecer heterogeneidade de variâncias, teste $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_1: \sigma_1^2 \neq \sigma_2^2$ (observe que isto é equivalente a testar $H_0: \sigma_1^2 / \sigma_2^2 = 1$, versus $H_1: \sigma_1^2 / \sigma_2^2 \neq 1$).
- Sob a hipótese nula, a estatística teste s_1^2/s_2^2 segue uma distribuição F com graus de liberdade: ($n_1 - 1$, $n_1 - 1$); se a estatística teste exceder o valor crítico na tabela C, então a hipótese nula pode ser rejeitada ao nível de significância indicado.

Tabela C
Valores críticos de F

Linha superior = .05, linha inferior = .01, pontos para distribuição de F

Graus de liberdade para o numerador

	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242
	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,38	19,39	19,40
	98,49	99,01	99,17	99,25	99,30	99,33	99,34	99,36	99,38	99,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78
	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23
4	7,71	6,94	6,59	6,39	6,26	6,16	6,08	6,04	6,00	5,96
	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74
	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63
	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34
	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13
	10,56	8,02	6,99	6,42	6,06	5,8	5,62	5,47	5,35	5,26
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97
	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,08	4,95	4,85

ANÁLISE DA VARIÂNCIA (ANOVA)

Finalidade: determinar se existe uma diferença significativa entre mais de duas médias grupais.

– Indica a possibilidade do efeito da média geral dos tratamentos experimentais: **não** especifica qual das médias é diferente.

ANOVA: consiste em obter estimativas independentes de subgrupos populacionais.

– A soma total dos quadrados é dividida em componentes de variação conhecidos.

Tipos de variância

– **Variância entre grupos:** reflete a magnitude da diferença entre as médias dos grupos.

– **Variância dentro de grupos:** reflete a dispersão dentro de cada grupo de tratamento; também recebe o nome de **estimativa dentro**.

Teste

– Quando a variância entre grupos é grande em relação à variância dentro de grupo, a razão F também será grande.

$$\text{Variância entre grupos} = \frac{n \sum (\bar{x}_i - \bar{x}_{tot})^2}{k - 1}$$

em que x_i = média do i-ésimo grupo de tratamento e \bar{x}_{tot} = média de todos os n valores entre todos os k grupos de tratamento.

$$\text{Variância dentro de grupos} = \frac{SQ_1 + SQ_2 + \dots + SQ_k}{n - k}$$

em que os SQ são a soma dos quadrados (veja *Medidas de tendência central na pág. I*) de cada valor dos subgrupos ao redor da média do subgrupo.

Usando a razão f: f = Variância entre grupos/variância dentro de grupos

– Os graus de liberdade são k-1 para o numerador e n-k para o denominador.

– Se a variância entre grupos > variância dentro de grupos, os tratamentos experimentais são responsáveis pelas grandes diferenças entre as médias dos grupos.

♦ **Hipótese nula:** as médias amostrais dos grupos são todas estimativas de uma média populacional comum; ou seja, $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$, para todos os grupos de tratamento k, versus $H_1: \text{ao menos um par de médias é diferente}$ (a determinação de qual par (ou pares) é diferente, requer um teste de *follow-up*).

PROPORÇÕES

Em **amostras aleatórias** de tamanho n , a proporção amostral p flutua em torno da média da proporção p com uma variância de $\frac{p(1-p)}{n}$,

erro padrão da proporção de $\sqrt{p(1-p)/n}$.

A medida que o tamanho amostral aumenta, ele se concentra mais próximo da média-alvo. E também se aproxima da distribuição normal, no caso:

$$z = \frac{p - \pi}{\sqrt{p(1-p)/n}}.$$

CORRELAÇÃO

Correlação

– Correlação refere-se à relação entre duas variáveis.

– O coeficiente de correlação r (também conhecido como “Coeficiente de Correlação Produto-Momento de Pearson”) é uma medida da relação linear entre duas variáveis quantitativas.

– Ex.: dadas as observações de duas variáveis X e Y, podemos calcular a soma de seus quadrados:

$$SQ_x = \sum (x - \bar{x})^2 \text{ e } SQ_y = \sum (y - \bar{y})^2.$$

Fórmulas para a correlação (r) de Pearson:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{SQ_x \cdot SQ_y}} =$$

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{N}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{N}} \sqrt{\sum y^2 - \frac{(\sum y)^2}{N}}}$$

Observe que $-1 \leq r \leq 1$ para qualquer conjunto de dados; quando $r = 1$, os dados são considerados de correlação positiva perfeita – se forem representados graficamente, formariam uma linha reta com curva positiva (para cima); se $r = -1$, os dados são considerados de correlação negativa perfeita – se representados graficamente, formariam uma curva negativa (para baixo); se $r = 0$, os dados são considerados **sem correlação linear** (é possível, logicamente, que eles estejam relacionados de outra forma). *Observe:* é possível que uma amostra aleatória de uma população com correlação zero produza por acaso uma amostra com $r \neq 0$.

TESTES QUI-QUADRADO χ^2

• É o teste não paramétrico mais usado.

• A média $\chi^2 =$ seus graus de liberdade.

• A variância $\chi^2 =$ o dobro de seus graus de liberdade.

• Pode ser usado para testar independência, homogeneidade e aderência.

• O quadrado de uma variável normal padronizada é uma variável qui-quadrado com $gl = 1$.

• Assim como a distribuição t, a forma da distribuição depende do valor do gl.

Cálculo dos graus de liberdade (gl)

- Se o método qui-quadrado testa a aderência de uma distribuição hipotética (usa a distribuição de frequência), $gl = g - 1$, em que g = número de grupos, ou classes, na distribuição de frequência.
- Se o método qui-quadrado testa a homogeneidade ou a independência (tabela de contingência bilateral):
 $gl = (\text{número de linhas}-1) (\text{número de colunas}-1)$

Teste de aderência: para aplicar a distribuição qui-quadrado desta forma, o valor de qui-quadrado crítico é representado por: $\sum \frac{(f_o - f_e)^2}{f_e}$

em que f_0 = frequência observada da variável, f_e = frequência esperada (com base na distribuição populacional hipotética).

Testes de contingência: aplicação de testes qui-quadrado em duas populações separadas para testar independência estatística de atributos.

Testes de homogeneidade: aplicação de testes qui-quadrado em duas amostras para testar se vieram de populações com distribuições semelhantes.

Teste de sequência: testa se uma sequência (a ser incluída numa amostra) é aleatória. Aplicam-se as seguintes equações:

$$(\bar{R}) = \frac{2n_1 n_2}{n_1 + n_2} + 1 \cdot S_R \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} \text{ em que}$$

\bar{R} = número médio de sequências;

n_1 = número de ocorrências de um tipo;

n_2 = número de ocorrências de outro tipo;

S_R = desvio padrão da distribuição do número de sequências.

TESTE DE HIPÓTESE PARA CORRELAÇÃO LINEAR

Com uma amostra aleatória simples de tamanho n que produz um coeficiente de correlação amostral r , é possível testar a correlação linear na população, ρ . Ou seja, empregamos o teste de hipótese $H_0: \rho = \rho_0$, versus uma alternativa lateral com cauda à direita, à esquerda ou bilateral; em geral, estamos interessados em determinar se existe alguma correlação linear; portanto, $\rho_0 = 0$.

$$(r - \rho_0)$$

A estatística teste é: $t = \frac{r - \rho_0}{\sqrt{(1 - r^2)/(n - 2)}}$

a qual segue uma distribuição t com $n - 2$ graus de liberdade sob H_0 ; este teste de hipótese assume que a amostra foi tirada de uma população com uma distribuição normal bivariada.

Exemplo: Uma amostra aleatória simples de tamanho 27 produz um coeficiente de correlação $r = -0,41$; com um $\alpha = 0,05$, há evidência suficiente para a existência de uma relação linear negativa?

Como estamos testando uma relação linear negativa, necessitamos um teste bilateral com cauda à esquerda: $H_0: \rho = 0$, versus $H_1: \rho < 0$; o valor crítico pode ser encontrado com base na distribuição t com gl de $n - 2 = 25$, e $\alpha = 0,05$ lateral; como este é um teste com cauda à esquerda, tomamos o valor negativo: $-1,708$; ou seja, se a estatística for menor que $-1,708$, concluimos que há suficiente prova para uma relação linear negativa.

A estatística teste é: $t = \frac{-0,41}{\sqrt{(1 - (-0,41)^2)/(27 - 2)}} = -2,248$,

permitindo a rejeição da hipótese nula de que não existe correlação linear e aceitando a hipótese alternativa de uma correlação linear negativa ao $\alpha = 0,05$.

REGRESSÃO

A regressão é um método para a previsão de valores de uma variável (variável de resultado/consequência ou dependente) com base nos valores de uma ou mais variáveis independentes ou prognosticadoras; montar um modelo de regressão é o processo no qual usamos dados amostrais para determinar uma equação e assim representar a relação.

REGRESSÃO LINEAR SIMPLES

Num modelo de regressão linear simples, usamos apenas uma variável prognosticadora e assumimos que a relação para a variável de resultado seja linear; ou seja, o gráfico para a equação de regressão é de uma linha reta; (em geral nos referimos à "linha de regressão"); para toda a população, o modelo pode ser indicado como:

$$y = \beta_0 + \beta_1 x + e$$

y é chamado de variável dependente (ou variável de resultado), pois pressupõe-se que depende de uma relação linear com x;

x é a variável independente, também chamada de variável prognosticadora;

β_0 é a intersecção da linha de regressão; isto é, o valor previsto para y quando x = 0;

β_1 é a curva da linha de regressão – a mudança marginal/disjunta em y por unidade de mudança em x;

e se refere ao erro padrão; pressupõe-se que a estimativa dentro segue uma distribuição normal com uma média de zero e variação constante – ou seja, não deve haver aumento nem diminuição na dispersão em regiões diferentes ao longo da linha de regressão; além disso, pressupõe-se que as estimativas dentro são independentes para observações diferentes (x, y).

Com base nos dados amostrais, encontramos estimativas b_0 e b_1 da intersecção β_0 e curva β_1 ; o que nos fornece a equação de regressão estimada (ou amostral) $\hat{y} = b_0 + b_1 x$.

As estimativas de parâmetro b_0 e b_1 podem ser derivadas usando-se várias formas; uma das mais comuns é o método dos mínimos quadrados; as estimativas dos mínimos quadrados minimizam as diferenças das somas dos quadrados previstos e os valores reais da variável dependente y.

Em um modelo de regressão linear simples, as estimativas dos mínimos quadrados da intersecção e da curva são:
curva estimada = $b_1 = \frac{SQxy}{SQx}$
intersecção estimada = $b_0 = \bar{y} - b_1 \bar{x}$.

Essas estimativas – e outros cálculos de regressão – envolvem somas dos quadrados:

$$SQxy = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - (\sum x)(\sum y)/n$$

$$SQx = \sum(x - \bar{x})^2 = \sum(x^2) - (\sum x)^2/n$$

$$SQy = \sum(y - \bar{y})^2 = \sum(y^2) - (\sum y)^2/n$$

Ex.: uma amostra aleatória simples de 8 carros fornece o seguinte dado sobre deslocamento do motor (x) e distância percorrida em estradas (y); monta-se um modelo de regressão linear simples.

(deslocamento)	(distância em milhas)			
x	y	x^2	y^2	xy
5,7	18	32,49	324	102,6
2,5	19	6,25	361	47,5
3,8	20	14,44	400	76,0
2,8	19	7,84	361	53,2
4,6	17	21,16	289	78,2
1,6	32	2,56	1024	51,2
1,6	29	2,56	841	46,4
1,4	30	1,96	900	42,0
Soma:	24,0	184	89,26	4500 497,1

Montar um modelo envolve calcular as estimativas dos mínimos quadrados b_0 e b_1 ; observe que há 8 observações – isto é, $n = 8$.

Em primeiro lugar, $SQxy = \sum xy - (\sum x)(\sum y)/n = -54,9$, $SQx = \sum(x^2) - (\sum x)^2/n = 17,26$, e $SQy = \sum(y^2) - (\sum y)^2/n = 268$

Então a curva estimada é $b_1 = \frac{SQxy}{SQx} = -3,18$, e a intersecção estimada é $b_0 = -b_1 \bar{x} = 32,54$.

O modelo de regressão estimado, portanto, é:
milhagem = $32,54 - 3,18 \text{ deslocamento}$.

SIGNIFICÂNCIA DE UM MODELO DE REGRESSÃO

Podemos avaliar a significância do modelo por meio de teste para ver se a amostra fornece prova suficiente de uma relação linear da população; ou seja, aplicamos o teste $H_0: \beta_1 = 0$, versus $H_1: \beta_1 \neq 0$; isso equivale exatamente ao teste da correlação linear na população: $H_0: \rho = 0$, versus $H_1: \rho \neq 0$; o teste da correlação é mais simples:

$$\text{O coeficiente de correlação } r = \frac{SQxy}{\sqrt{SQx \cdot SQy}} = -0,8072.$$

$$\text{A estatística teste } t = \frac{(r - 0)}{\sqrt{(1 - r^2)/(n - 2)}} = -3,350.$$

Consultando a tabela B, com graus de liberdade = $n - 2 = 6$, obtemos um valor crítico de 3,143 ao $\alpha = 0,02$, e um valor crítico de 3,707 ao $\alpha = 0,01$; como temos um teste bilateral, devemos considerar o valor absoluto da estatística teste, que supera 3,143, mas não supera 3,707; dessa forma, rejeitamos a H_0 ao $\alpha = 0,02$ mas não ao $\alpha = 0,01$, portanto o valor p está entre 0,02 e 0,01; (o valor p real, que pode ser encontrado por meio de cálculos computadorizados, é 0,0154); esse é um modelo significante razoável.

DETERMINAÇÃO LINEAR

Os modelos de regressão também são avaliados pelo coeficiente de determinação linear, r^2 ; isso representa a proporção da variação total em y que é explicado pelo modelo de regressão; o coeficiente de determinação linear pode ser calculado por várias formas; a mais fácil é usando: $r^2 = (r)^2$; ou seja, o coeficiente de determinação é o quadrado do coeficiente de correlação.

RESÍDUOS

A diferença entre um valor observado e um valor "acomodado" de y ($y - \hat{y}$) é chamado de resíduo; examinar os resíduos é útil para identificar valores fora da linha (observações distorcidas da linha de regressão, que representam valores não usuais para x e y) e avaliar as pressuposições do modelo.



Barros, Fischer & Associados

ESTATÍSTICA

4ª edição – Junho / 2010

Autor: John Mijares

Tradução: Mônica Koehler Sant'Ana

Consultoria: Nabor Monteiro

Edição: Andréa Barros

Arte: Cláudio Scalzite e Maurício Cioffoli

Revisão: Paulo Roberto Pompéo

Resumo – Estatística (série de Ciências Exatas, nº 13) é uma publicação da Barros, Fischer & Associados Ltda., sob licença editorial de Spring Publishing Group, Inc. Copyright © 2010 Barcharts, Inc. USA. Todos os direitos da edição são reservados Barros, Fischer & Associados Ltda. A série de resumos de Ciências Exatas, devido a seu formato condensado, contém os conceitos básicos das matérias de que trata, sendo excelente ferramenta para estudantes e profissionais da área.

Endereço: Rua Ulpiano, 86

Lapa, São Paulo, CEP 05050-020

Telefone/fax: 0 (xx) 11 3675-0508

Site: www.resumao.com.br

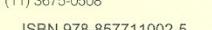
E-mail: contato@resumao.com.br

Impressão: Tarco Indústria Gráfica Ltda.

Distribuição e vendas: Bafisa, tel.: (11) 3675-0508



1147819



ISBN 978-857711002-5

Reprodução proibida

É expressamente proibida a reprodução total ou parcial do conteúdo desta publicação sem a prévia autorização do editor.

9 788577 110025