Técnicas de Amostragem - Aula 03

Amostragem Aleatória Simples Com Reposição (AASc)

Kaique Matias de Andrade Roberto

Ciências Atuariais

HECSA - Escola de Negócios

FIAM-FAAM-FMU

Conteúdo

Conceitos que aprendemos em Aulas anteriores

Conceitos que aprendemos em Aulas anteriores

Nas Aulas anteriores nós lidamos com as principais definições, nomenclaturas e terminologias da Teoria de Amostragem. A saber:

- população;
- amostra;
- planejamento amostral;
- estatísticas;
- distribuições amostrais;
- estimadores e suas propriedades.

Cálculos com a Distribuição Normal

Cálculos com a Distribuição Normal

A distribuição normal, também conhecida como distribuição Gaussiana, é uma das distribuições de probabilidade mais utilizadas, por pelo menos 2 motivos:

Cálculos com a Distribuição Normal

- permite modelar fenômenos naturais, estudos do comportamento humano, processos industriais, entre outros;

Cálculos com a Distribuição Normal

- possibilita o uso de aproximações para o cálculo de probabilidades de muitas variáveis aleatórias.

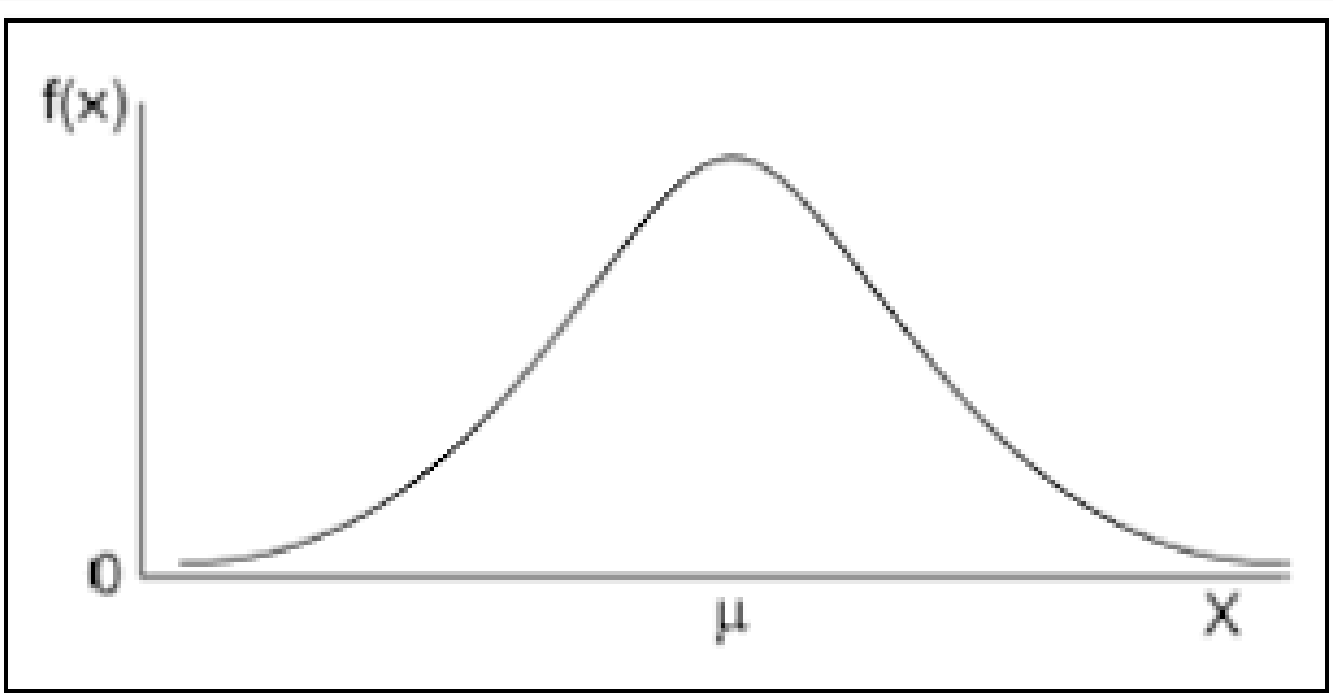
Cálculos com a Distribuição Normal

Definição 2.1

Uma variável aleatória X com média $\mu \in \mathbb{R}$ e desvio padrão $\sigma > 0$ tem **distribuição Normal** denotada $X \sim N(\mu, \sigma^2)$, se a sua função de distribuição de probabilidades for dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}, \; x \in \mathbb{R}.$$

Cálculos com a Distribuição Normal



Cálculos com a Distribuição Normal

Do gráfico da densidade da Normal podemos concluir algumas propriedades básicas:

- $f(x)$ é simétrica em relação a μ;
- $f(x)$ decresce a medida que $|x|$ cresce;
- o valor máximo de $f(x)$ se dá para $x = \mu$.

Cálculos com a Distribuição Normal

Note que, para calcular $P(a \leq X \leq b)$ devemos realizar a conta

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}\, dx.$$

Cálculos com a Distribuição Normal

Também pode-se demostrar que se $X \sim N(\mu, \sigma^2)$ então

$$E(X) = \mu \text{ e } \text{Var}(X) = \sigma^2.$$

Cálculos com a Distribuição Normal

Claro, vocês em geral não fizeram cursos de Cálculo Diferencial e Integral. Mas mesmo se esse fosse o caso, essa integral **não é calculável (não existe fórmula fechada)**.

Cálculos com a Distribuição Normal

Apesar disso, existem métodos para aproximação do cálculo integrais com excelente precisão, e para o caso da distribuição Normal, esses valores já estão tabelados.

Cálculos com a Distribuição Normal

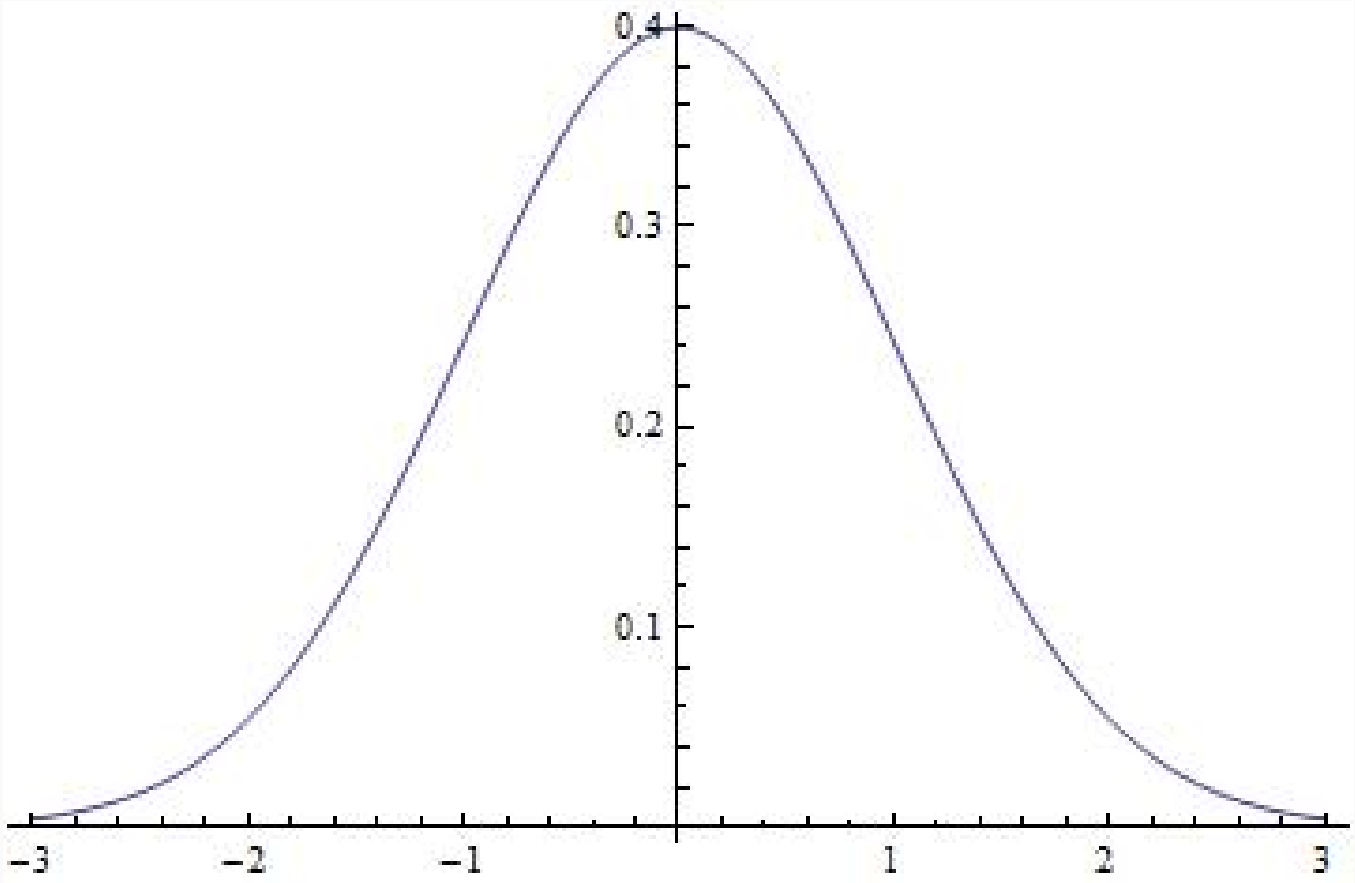
Dito isso, vamos aprender/relembrar como calcular as principais probabilidades envolvendo a distribuição Normal.

Cálculos com a Distribuição Normal

Definição 2.2

Cálculos com a Distribuição Normal é denominada **Normal Padrão**. Denotamos $Z \sim N(0, 1)$.

Cálculos com a Distribuição Normal



Cálculos com a Distribuição Normal

Exemplo 2.3

Seja $Z \sim N(0,1)$ a Normal Padrão. Calcule:

a - $P(Z < 1)$;

b - $P(Z < 1.5)$;

c - $P(Z > 2)$;

d - $P(Z > 3)$;

e - $P(-1 < Z < 1)$;

f - $P(0.5 < Z < 1.7)$;

g - $P(Z < -0.7 \text{ ou } Z > 0.7)$;

h - $P(Z < -0.4 \text{ ou } Z > 1)$.

Cálculos com a Distribuição Normal

Exercício 2.1

Seja $Z \sim N(0,1)$ a Normal Padrão. Calcule:

a - $P(-1 < Z < 1)$;

b - $P(-2 < Z < 2)$;

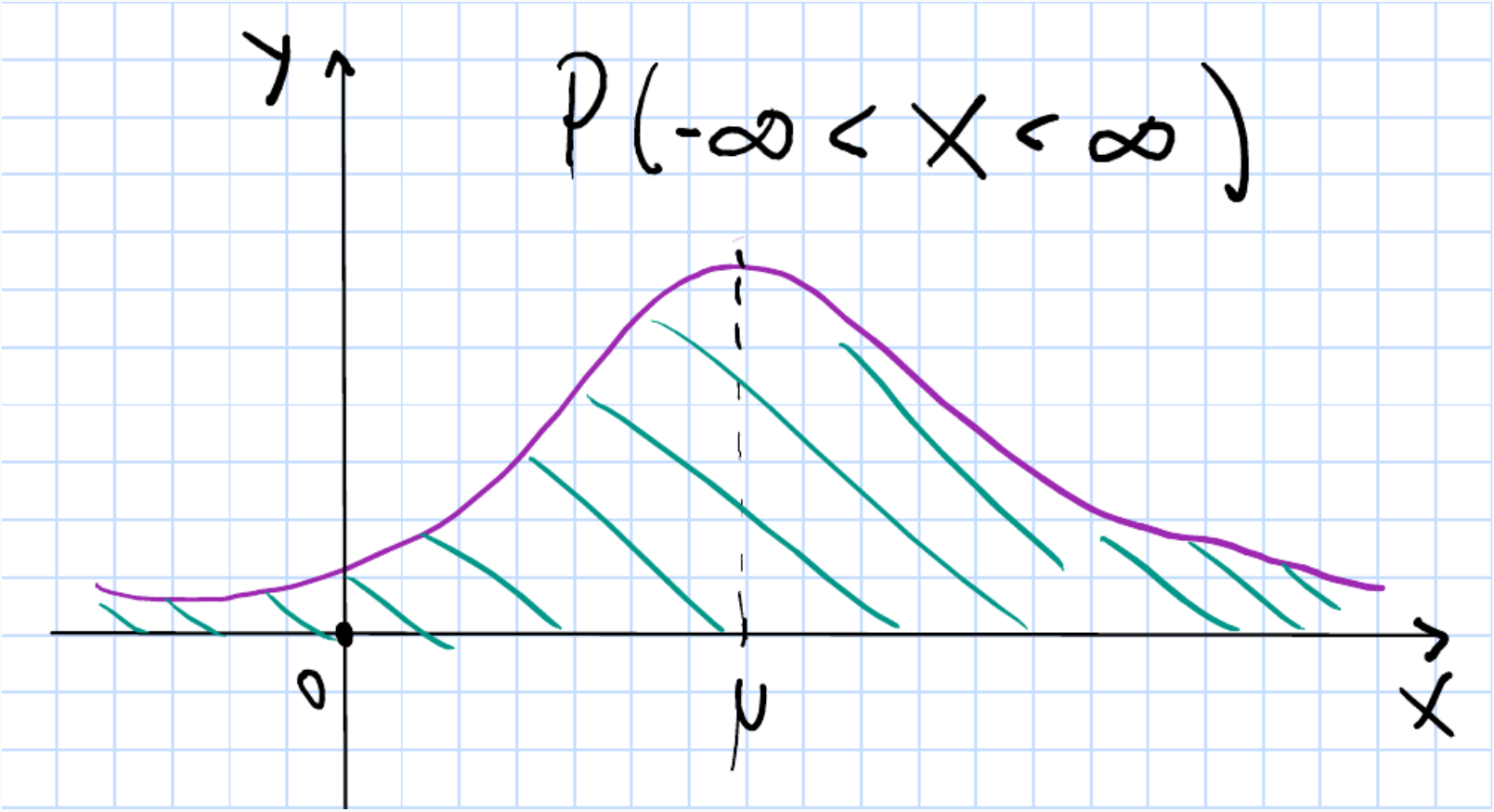
c - $P(-3 < Z < 3)$;

d - $P(-4 < Z < 4)$.

Cálculos com a Distribuição Normal

Vamos usar a Normal Padrão $Z \sim N(0,1)$ para calcular probabilidades para qualquer normal $X \sim N(\mu, \sigma^2)$.

Cálculos com a Distribuição Normal



Cálculos com a Distribuição Normal

Para obtermos, a partir da distribuição normal, a distribuição normal padrão ou distribuição normal reduzida, a variável original X é transformada em uma nova variável aleatória Z, com média zero ($\mu = 0$) e variância 1 ($\sigma^2 = 1$):

$$\frac{X - \mu}{\sigma} \sim N(0, 1).$$

Cálculos com a Distribuição Normal

Este tipo de transformação, conhecida por **zscore**, é muito utilizada para a padronização de variáveis, pois não altera a forma da distribuição da variável original e gera uma nova variável com média zero e variância 1.

Cálculos com a Distribuição Normal

Logo, para calcular $P(X < 0)$ para uma normal $N(\mu, \sigma^2)$ basta aplicar o zscore e calcular

$$P\left(\frac{X-\mu}{\sigma} \leq \frac{0-\mu}{\sigma}\right)$$

e proceder da mesma maneira que fizemos para o caso $Z \sim N(0,1)$.

Cálculos com a Distribuição Normal

Vamos ilustrar as contas com a Normal $X \sim N(8, 36)$.

Cálculos com a Distribuição Normal

Exemplo 2.4

Seja $Z \sim N(8, 36)$ a Normal Padrão. Calcule:

a - $P(X < 0)$;
b - $P(X \leq 12)$;
c - $P(X \leq 20)$;
d - $P(X \geq 2)$;
e - $P(X \geq 5)$;
f - $P(6 \leq X \geq 11)$;
g - $P(10 \leq X \geq 25)$;
h - $P(X \leq 3 \text{ ou } X \geq 9)$;
i - $P(X \leq 9 \text{ ou } X \geq 20)$.

Cálculos com a Distribuição Normal

Exercício 2.2

Seja $X \sim N(\mu, \sigma^2)$. Calcule:

a - $P(-\sigma + \mu < Z < \sigma + \mu)$;

b - $P(-2\sigma + \mu < Z < 2\sigma + \mu)$;

c - $P(-3\sigma + \mu < Z < 3\sigma + \mu)$;

d - $P(-4\sigma + \mu < Z < 4\sigma + \mu)$.

Amostragem Aleatória Simples

Amostragem Aleatória Simples

A amostragem aleatória simples (AAS) é o método mais simples e mais importante para a seleção de uma amostra.

Amostragem Aleatória Simples

Além de servir como plano próprio, o seu procedimento é usado de modo repetido em procedimento de múltiplos estágios.

Amostragem Aleatória Simples

A principal caracterização para o uso do plano AAS é a existência de um sistema de referências completo, descrevendo cada uma das unidades elementares.

Amostragem Aleatória Simples

Deste modo tem-se bem listado o universo

$$\mathcal{U} = \{1, 2, ..., N\}.$$

O plano é descrito do seguinte modo:

Amostragem Aleatória Simples

- Utilizando algum procedimento aleatório (tabela de números aleatórios, urna, etc) sorteia-se com igual probabilidade um elemento da população $\mathcal{U}$;

Amostragem Aleatória Simples

- repete-se o processo anterior até que sejam sorteados n unidades, tendo sido este número pré-fixado anteriormente;

Amostragem Aleatória Simples

- caso seja permitido o sorteio de uma unidade mais de uma vez, tem-se o processo AAS com reposição (AASc). Quando o elemento sorteado é removido de $\mathcal{U}$ antes do sorteio próximo, tem-se o plano AAS sem reposição (AASs).

Amostragem Aleatória Simples

Do ponto de vista prático, o plano AASs é muito mais interessante, pois vai de encontro ao princípio intuitivo que "não se ganha mais informação se uma mesma unidade aparece mais de uma vez na amostra".

Amostragem Aleatória Simples

Por outro lado, o plano AASc, introduz vantagens matemáticas e estatísticas, como a independência entre as unidades sorteadas, que facilita em muito a determinação das propriedades estatísticas dos estimadores das quantidades populacionais de interesse.

Amostragem Aleatória Simples

Basta observar na maioria dos assuntos tratados em livros de inferência há imposição de que as unidades que fazem parte da amostra sejam independentes.

Amostragem Aleatória Simples

Também vamos considerar inferências para os seguintes parâmetros de interesse: considere para cada unidade i, uma característica populacional unidimensional de interesse, Y_i, $i \in \mathcal{U}$. Vamos considerar:

Amostragem Aleatória Simples

- total populacional,

$$\tau = \sum_{i=1}^{N} Y_i;$$

Amostragem Aleatória Simples

- média populacional,

$$\mu = \overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i;$$

- variância populacional, representada por

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2 \text{ e } S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu)^2.$$

Amostragem Aleatória Simples Com Reposição (AASc)

Amostragem Aleatória Simples Com Reposição (AASc)

A AASc opera da seguinte forma:

Amostragem Aleatória Simples Com Reposição (AASc)

- A população está numerada de 1 a N, de acordo com o sistema de referências, ou seja,

$$\mathcal{U} = \{1, 2, ..., N\}.$$

Amostragem Aleatória Simples Com Reposição (AASc)

- Utilizando um software ou tabela de números aleatórios, sorteia-se, com igual probabilidade, uma das N unidades da população;

Amostragem Aleatória Simples Com Reposição (AASc)

- repõe-se essa unidade na população e sorteia-se um elemento seguinte;

Amostragem Aleatória Simples Com Reposição (AASc)

- repete-se esse procedimento até que n unidades tenham sido sorteadas.

Amostragem Aleatória Simples Com Reposição (AASc)

Agora vamos listar os resultados teóricos envolvendo AASc:

Amostragem Aleatória Simples Com Reposição (AASc)

Teorema 4.1

Para o plano amostral AASc, a variável f_i, número de vezes que a unidade i aparece na amostra segue uma distribuição binomial

$$f_i \sim b\left(n, \frac{1}{N}\right).$$

Além disso,

$$E[f_i] = \frac{n}{N},\ \text{Var}[f_i] = \frac{n}{N}\left(1 - \frac{1}{N}\right)$$

$$\text{Cov}[f_i, f_j] = -\frac{n}{N} \text{ para } i \neq j.$$

Amostragem Aleatória Simples Com Reposição (AASc)

Teorema 4.2

A estatística $t(\boldsymbol{s})$, total da amostra, definida por

$$t(\boldsymbol{s}) = \sum_{i \in \boldsymbol{s}} Y_i$$

tem, para o plano AASc, as seguintes propriedades:

$$E[t] = n\mu \text{ e } \text{Var}[t] = n\sigma^2.$$

Amostragem Aleatória Simples Com Reposição (AASc)

Dos resultados acima, derivam-se estimadores não-viesados para μ e σ^2, resumidos no seguinte teorema.

Amostragem Aleatória Simples Com Reposição (AASc)

Teorema 4.3

A média amostral

$$\overline{y} = \hat{\mu} = \frac{t(\boldsymbol{s})}{n} = \frac{1}{n}\sum_{i \in \boldsymbol{s}} Y_i$$

é um estimador não viesado da média populacional μ dentro do plano AASc, e ainda

$$\text{Var}[\overline{y}] = \frac{\sigma^2}{n}.$$

Amostragem Aleatória Simples Com Reposição (AASc)

Teorema 4.4

Dentro do plano AASc, a estatística

$$s^2 = \frac{1}{n-1}\sum_{i \in \boldsymbol{s}}(Y_i - \overline{y})^2,$$

é um estimador não viesado da variância populacional σ^2.

Amostragem Aleatória Simples Com Reposição (AASc)

Exemplo 4.5

Considere novamente os dados na planilha "aula-02-exemplo", e considere a variável renda familiar, onde o universo é $\mathcal{U}$ e o parâmetro populacional é $\boldsymbol{D} = (12, 30, 18)$. Vamos verificar como se comportam $\overline{y}$ e s^2 com relação as funções paramétricas μ e σ^2 de $\boldsymbol{D}$ para o plano amostral AASs com $n = 2$.

Normalidade e Intervalo de Confiança

Normalidade e Intervalo de Confiança

Conforme o tamanho da amostra aumenta, as distribuições de $\overline{y}$ e T vão se aproximando da distribuição normal, de acordo com o Teorema do Limite Central (TLC), tanto para o caso da AASc como para a AASs.

Teorema 5.1 (Limite Central)

Para amostras aleatórias simples $(X_1, ..., X_n)$, retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média $\overline{X}$ aproxima-se, para n grande, de uma distribuição normal, com média μ e variância σ^2/n.

Normalidade e Intervalo de Confiança

Então para n suficientemente grande, temos, com relação à AASc, que

$$\frac{\overline{y} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \text{ e } \frac{T - \tau}{N\sqrt{\sigma^2/n}} \sim N(0,1),$$

onde $N(0,1)$ denota uma variável aleatória com distribuição normal com média zero e variância 1.

Normalidade e Intervalo de Confiança

Os resultados acima possibilitam a obtenção de intervalos de confiança aproximados para $\overline{y}$ e T.

Normalidade e Intervalo de Confiança

Então com relação à média populacional, temos para n suficientemente grande que

$$P\left(\frac{|\overline{y}-\mu|}{\sqrt{\sigma^2/n}} \leq z_\alpha\right) \simeq 1-\alpha,$$

onde z_α é a ordenada da $N(0,1)$ de tal forma que a área na densidade da $N(0,1)$ no intervalo $(-z_\alpha, z_\alpha)$ é igual a $1-\alpha$.

Quando σ^2 for desconhecido, iremos substitui-lo por seu estimador não viciado s^2, que para n grande é bem próximo de σ^2. Assim, a expressão acima pode ser escrita como

$$P\left(\overline{y} - z_\alpha\sqrt{\frac{s^2}{n}} \leq \mu \leq y + z_\alpha\sqrt{\frac{s^2}{n}}\right) \simeq 1 - \alpha.$$

Normalidade e Intervalo de Confiança

Disso segue que

$$\left(\overline{y} - z_\alpha \sqrt{\frac{s^2}{n}}, y + z_\alpha \sqrt{\frac{s^2}{n}}\right)$$

é um intervalo de confiança para μ com coeficiente de confiança aproximadamente igual a $1 - \alpha$.

Normalidade e Intervalo de Confiança

A interpretação usual do intervalo de confiança está baseada no fato de que se forem observadas 100 amostras AAS, e construídos 100 intervalos de confiança baseados nestas amostras, então, aproximadamente $100(1-\alpha)\%$ dos intervalos devem conter μ.

Exemplo 5.2

Uma máquina enche pacotes de café com uma variância igual a $100g$. Ela estava regulada para encher os pacotes com $500g$, em média. Agora, ela se desregulou, e queremos saber qual a nova média μ. Uma amostra de 25 pacotes apresentou uma média igual a $485g$. Vamos construir um intervalo de confiança com 95% de confiança para μ.

Determinação do Tamanho da Amostra

Determinação do Tamanho da Amostra

Agora vamos determinar o tamanho da amostra n de tal forma que o estimador obtido tenha um erro máximo de estimação igual a ε, com determinado grau de confiança (probabilidade).

Determinação do Tamanho da Amostra

Para a determinação do tamanho da amostra é preciso:

- fixar o erro máximo desejado ε;
- fixar o grau de confiança z_α;
- possuir algum conhecimento da variabilidade da população σ^2.

Determinação do Tamanho da Amostra

Os dois primeiros são fixados pelo pesquisador, e quanto ao terceiro, a resposta exige mais trabalho. O uso de pesquisas anteriores ou amostras piloto são os critérios mais utilizados.

Determinação do Tamanho da Amostra

De maneira mais específica, o problema consiste em determinar n de modo que

$$P(|\overline{y} - \mu| \leq \varepsilon) \simeq 1 - \alpha.$$

Determinação do Tamanho da Amostra

Para n grande, tem-se que

$$P\left(|\overline{y} - \mu| \leq z_\alpha \sqrt{\frac{\sigma^2}{n}}\right) \simeq 1 - \alpha.$$

Determinação do Tamanho da Amostra

Então para ε fixado, a solução do problema acima consiste em determinar n de tal forma que

$$n \geq \frac{\sigma^2}{(\varepsilon/z_\alpha)^2}.$$

Determinação do Tamanho da Amostra

Exemplo 6.1

Considere a população de moradores de um condomínio ($N = 540$). Deseja-se estimar a idade média dos condôminos. Com base em pesquisas passadas, pode-se obter a estimativa para σ^2 de 463.32. Suponha que será retirada da população uma amostra segundo AASc. Admitindo que a diferença entre a média amostral e a verdadeira média populacional seja, no máximo, de 4 anos, com um nível de confiança de 95%, determine o tamanho da amostra a ser coletada.

Estimação da Proporção

Estimação da Proporção

De maneira geral, em muitas situações, existe interesse em estudar a proporção de elementos em certa população que possuem determinada característica, como ser ou não um item defeituoso, ser ou não eleitor de um determinado partido, etc.

Estimação da Proporção

Nesta situação, a cada elemento da população está associada a variável aleatória Y_i da seguinte maneira:

$$Y_i = \begin{cases} 1 \text{ se o indivíduo for portador da característica} \\ 0 \text{ se o indivíduo não for portador da característica} \end{cases}$$

Distribuição Amostral da Proporção

Daí

$$P = \mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

é a proporção de unidades da população que possuem a característica de interesse, e podemos escrever

$$\sigma^2 = P(1 - P).$$

Estimação da Proporção

A seguir segue um resumo dos resultados teóricos obtidos para a estimação da proporção:

Estimação da Proporção

Teorema 7.1

Um estimador não-viciado de P baseado na AASc é dado por:

$$p = \hat{P} = \overline{y} = \frac{m}{n},$$

com

$$\text{Var}(\hat{P}) = \frac{PQ}{n},$$

sendo $Q = 1 - P$. Além disso, um estimador não-viciado de $\text{Var}(P)$ é

$$\text{var}(p) = \frac{\hat{P}\hat{Q}}{n-1}.$$

Estimação da Proporção

Utilizando-se a aproximação normal da binomial, um intervalo de confiança aproximado para P é dado por

$$\left(\hat{P} - z_\alpha \sqrt{\frac{\hat{P}\hat{Q}}{n-1}}; \hat{P} + z_\alpha \sqrt{\frac{\hat{P}\hat{Q}}{n-1}} \right).$$

Estimação da Proporção

Notando-se que o produto PQ (e portanto $\hat{P}\hat{Q}$) é sempre menor que $1/4$, um intervalo de confiança conservador para P é dado por

$$\left(\hat{P} - z_\alpha \sqrt{\frac{1}{4(n-1)}}; \hat{P} + z_\alpha \sqrt{\frac{1}{4(n-1)}}\right).$$

Estimação da Proporção

Como no caso da Média amostral, pode-se considerar o tamanho da amostra n de tal forma que

$$P(|\hat{P} - P| \leq \varepsilon) \cong 1 - \alpha.$$

Estimação da Proporção

Pode-se mostrar que o valor de n é dado por

$$n = \frac{PQ}{\varepsilon^2 / z_\alpha^2},$$

e para usarmos esta fórmula, é necessário um valor estimado para P.

Estimação da Proporção

Uma forma alternativa, que produz um valor conservador para n consiste em utilizar o fato de que $PQ \leq 1/4$. Neste caso, tem-se

$$n = \frac{z_\alpha^2}{4\varepsilon^2}.$$

Estimação da Proporção

Exemplo 7.2

Suponha que $p = 30\%$ dos estudantes de uma escola sejam mulheres. Colhemos uma AAS de $n = 10$ estudantes e calculamos

$$\hat{p} = \text{ proporção de mulheres na amostra.}$$

Qual a probabilidade de que $\hat{p}$ difira de p em menos de $0,01$?

Comentários Finais

Comentários Finais

Em resumo, na aula de hoje nós lidamos com os aspectos da AASc:

- definição da AASc;
- propriedades das principais estatísticas;
- normalidade e intervalo de confiança;
- tamanho da amostra;
- estimação da proporção.

Comentários Finais

Nas próximas aulas nós vamos:

- fazer exercícios envolvendo os conceitos da aula de hoje;
- começar a lidar com os aspectos da AASs (Amostragem Aleatória Simples Sem Reposição).

Comentários Finais

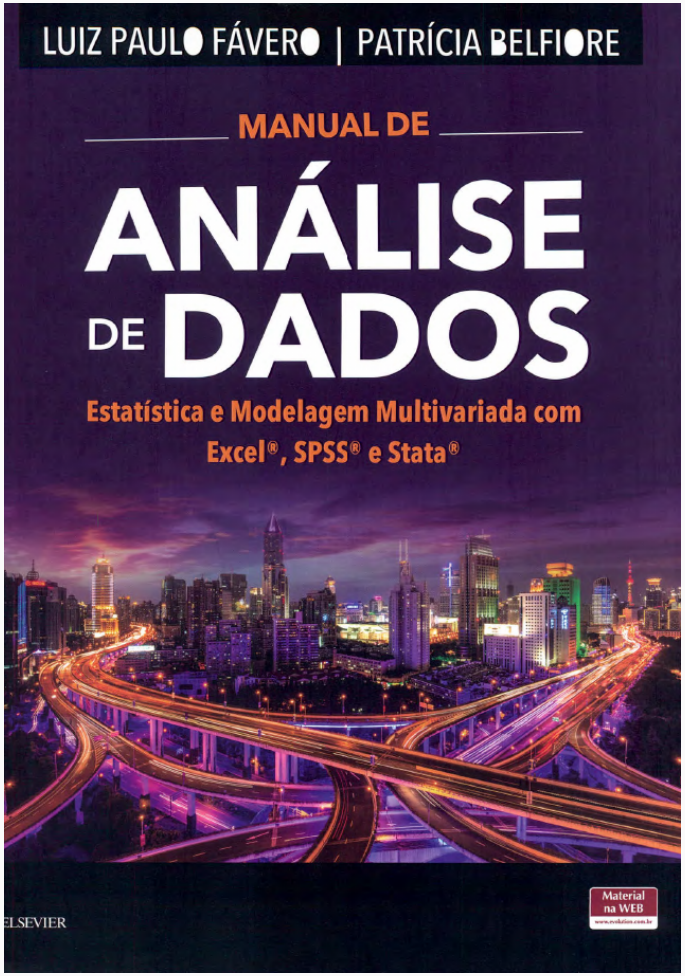
ATIVIDADE PARA ENTREGAR (E COMPOR A NOTA N1)

Resolva em grupos de até 4 integrantes os Exercícios 3.5-3.9.

Referências

Referências

Referências



Bons Estudos!