9 - INFERÊNCIA ESTATÍSTICA

9.1 - Introdução

Estatística é a ciência que se ocupa de organizar, descrever, analisar e interpretar dados para que seja possível a tomada de decisões e/ou a validação científica de uma conclusão. Vamos rever algumas definições dos Capítulos 6 e 7.

Os dados são coletados para estudar uma ou mais características de uma **POPULAÇÃO**: conjunto de elementos que tem pelo menos uma característica em comum => conjunto das medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m).

Uma população pode ser representada através de um modelo probabilístico: este apresenta condições para uso, uma certa forma para a distribuição de probabilidades, e *parâmetros*.

Os dados necessários para a obtenção do modelo podem ser obtidos através de um CENSO (pesquisa de toda a população), ou através de uma AMOSTRA (subconjunto finito) da população.

Por que usar uma AMOSTRA?
- economia;
- rapidez;
- para evitar a exaustão/extinção da população.

A AMOSTRA deve ser: ***representativa*** da população, ***suficiente*** (para que o resultado tenha confiabilidade), e ***aleatória*** (retirada por sorteio não viciado).

"A Inferência Estatística consiste em fazer *afirmações probabilísticas* sobre as características do modelo probabilístico, que se supõe representar uma população, a partir dos dados de uma amostra *aleatória* (probabilística) desta mesma população".

Fazer uma afirmação probabilística sobre uma característica qualquer é associar à declaração feita uma probabilidade de que tal declaração esteja correta (e portanto a probabilidade complementar de que esteja errada). Quando se usa uma amostra da população SEMPRE haverá uma probabilidade de estar cometendo um erro (justamente por ser usada uma amostra): a diferença entre os métodos estatísticos e os outros reside no fato de que os métodos estatísticos permitem ***calcular*** essa probabilidade de erro. E para que isso seja possível a amostra da população precisa ser aleatória.

As afirmações probabilísticas sobre o modelo da população podem ser basicamente:
=> estimar quais são os possíveis valores dos parâmetros (**Estimação de Parâmetros**):
- qual é o valor da média de uma variável que segue uma distribuição normal?
- qual é o valor da proporção de um dos 2 resultados possíveis de uma variável que segue uma distribuição binomial.

=> testar hipóteses sobre as características do modelo: parâmetros, forma da distribuição de probabilidades, etc. (**Testes de Hipóteses**).
- o valor da média de uma variável que segue uma distribuição é maior do que um certo valor?
- o modelo probabilístico da população é uma distribuição normal?
- o valor da média de uma variável que segue uma distribuição normal em uma população é diferente da mesma média em outra população?

9.1.1- Amostragem Aleatória ou Probabilística

Na amostragem aleatória todos os elementos da população têm chance diferente de zero de pertencer à amostra, uma vez que a seleção dos elementos é feita por sorteio não viciado (ver detalhes no Capítulo 7).

Amostragem Aleatória Simples
=> todos os elementos da população têm a *mesma* chance de pertencer à amostra (escolha por sorteio não viciado);
=> cada elemento sorteado é reposto na população antes do próximo sorteio.

Obviamente nem sempre é viável a amostragem com reposição, se a amostragem for feita SEM REPOSIÇÃO os resultados serão praticamente iguais se o tamanho da amostra não exceder a 5% do tamanho da população[1]. Se a população não for homogênea em relação à variável sob estudo, para garantir a representatividade da amostra somos obrigados a selecionar elementos de cada uma de suas subdivisões.

9.1.2 - Análise Exploratória de Dados[2]

Uma vez tendo coletado os dados, seja através de censo ou por amostragem, é preciso *resumi-los e organizá-los* de maneira a permitir uma primeira *análise*, e posterior uso das informações. As técnicas estatísticas que se ocupam desses aspectos constituem a Análise Exploratória de Dados.

O conjunto de dados pode ser resumido (e apresentado) através das distribuições de freqüências, que relacionam os valores que a variável pode assumir com a freqüência (contagem) com que foram encontrados naquele conjunto. Esta distribuição pode ser apresentada na forma de uma tabela, ou através de um gráfico (estes dois métodos podem ser usados tanto para variáveis qualitativas quanto para variáveis qualitativas).

Há uma terceira forma de resumir o conjunto de dados: as medidas de síntese ou **estatísticas**. As principais estatísticas são a média, o desvio padrão, a variância e a proporção. Serão apresentadas suas fórmulas básicas, e o que cada uma significa.

Média: trata-se de uma estatística que caracteriza o "centro de massa" do conjunto de dados (Valor Esperado); quando é a média populacional recebe o símbolo μ; quando é a média amostral recebe o símbolo $\overline{\mathbf{X}}$; trata-se da média aritmética simples: $\overline{x} = \frac{\sum x_i}{n}$ onde **xi** é o i-ésimo valor do conjunto e **n** é o número de elementos da amostra.

Variância: trata-se de uma estatística que mede a dispersão em torno da média do conjunto (em torno do valor esperado), possuindo uma unidade que é o quadrado da unidade da média (e dos valores do conjunto); quando é a variância populacional recebe o símbolo σ^2; quando é a variância amostral recebe o símbolo $\mathbf{s^2}$; $s^2 = \frac{\sum(x_i^2) - [(\sum x_i)^2 / n]}{n-1}$ (usa-se **n - 1** no denominador quando se

[1] A reposição garante que as probabilidades de selecionar um determinado elemento permanecem constantes, uma vez que o "espaço amostral" permanece o mesmo. Quando o tamanho da amostra é menor ou igual a 5% do tamanho da população, mesmo que não haja reposição supõe-se que as probabilidades não se modificam substancialmente.
[2] Este item foi visto com detalhe no Capítulo 2, na disciplina INE7001 – Estatística para Administradores I.

trata de uma amostra)[3]. O **desvio padrão** é a raiz quadrada positiva da variância, tendo portanto uma unidade que é igual à unidade da média, sendo muitas vezes preferida para efeito de mensuração da dispersão.
Proporção: consiste em calcular a razão entre o número de ocorrências do valor de interesse de uma variável qualitativa e o número total de ocorrências registradas no conjunto (de todos os valores que a variável pode assumir); quando é uma proporção populacional recebe o símbolo π; quando é uma proporção amostral recebe o símbolo **p**.

Os valores das medidas de síntese, além de resumirem o conjunto de dados, constituem uma indicação dos prováveis valores dos parâmetros. Assim, em estudos baseados em amostras, é comum utilizar tais medidas de síntese como *estatísticas* que serão utilizadas para estimar os parâmetros do modelo probabilístico que descreve a população.

9.2 - Distribuição Amostral

Os valores dos parâmetros do modelo populacional calculados em uma amostra são chamados de ***estatísticas***:

Parâmetros (População)	Estatísticas (Amostra)
Média (μ)	$\bar{X}$
Variância (σ^2)	s^2
Desvio Padrão (σ)	**s**
Proporção (π)	**p**
Número de elementos (**N**)	**n**

Seja uma população qualquer com um parâmetro θ de interesse, correspondendo a uma estatística **T** em uma amostra. Amostras aleatórias são retiradas da população e para cada amostra calcula-se o valor **t** da estatística **T**.

Os valores de **t** formam uma nova "população" que segue uma distribuição de probabilidades que é chamada de **<u>distribuição amostral</u>** de **T**.

Seja a população abaixo, constituída pelos pesos em kg de oito pessoas adultas:

População

64	**67**
70	**57**
65	**72**
70	**60**

μ = 65,62 kg

Sortear 3

Amostras

{64, 67, 60} => Média = 63,67 kg

{64, 70, 72} => Média = 68,67 kg

{65, 70, 60} => Média = 65 kg

Observe que há uma **<u>variação</u>** na estatística média, e esta variação precisa ser considerada quando são realizadas as inferências sobre os parâmetros.

Figura 1 - Distribuição Amostral - Exemplo

Assim sendo, o conhecimento das distribuições amostrais das principais estatísticas é necessário para fazer inferências sobre os parâmetros do modelo probabilístico da população.

[3] Há uma razão matemática para isso: garantir que o valor amostral seja um estimador não viciado do valor populacional (maiores detalhes no item Estimação por Ponto).

Por hora, basta conhecer as distribuições amostrais das estatísticas média de uma variável quantitativa qualquer, e proporção de um dos dois únicos resultados de uma variável qualitativa.

Exemplo 9.1[4]- Suponha uma variável quantitativa cujos valores constituem uma população com os seguintes valores: (**2, 3, 4, 5**)
Para esta população, que tem uma distribuição uniforme, podemos observar que os parâmetros são:

μ **= 3,5** σ^2 **= 1,25** (usou-se **n** no denominador por ser uma população)

Se retirarmos todas as amostras aleatórias de 2 elementos (com reposição) possíveis desta população (**n = 2**), teremos os seguintes resultados[5]:

(2, 2)	**(2, 3)**	**(2, 4)**	**(2, 5)**
(3, 2)	**(3, 3)**	**(3, 4)**	**(3, 5)**
(4,2)	**(4, 3)**	**(4, 4)**	**(4, 5)**
(5, 2)	**(5, 3)**	**(5, 4)**	**(5, 5)**

O cálculo das médias de todas as amostras acima resultará na matriz abaixo:

$$\bar{X}\begin{Bmatrix} (2,0) & (2,5) & (3,0) & (3,5) \\ (2,5) & (3,0) & (3,5) & (4,0) \\ (3,0) & (3,5) & (4,0) & (4,5) \\ (3,5) & (4,0) & (4,5) & (5,0) \end{Bmatrix}$$

Se estas médias forem plotadas em um gráfico de freqüências (um histograma):

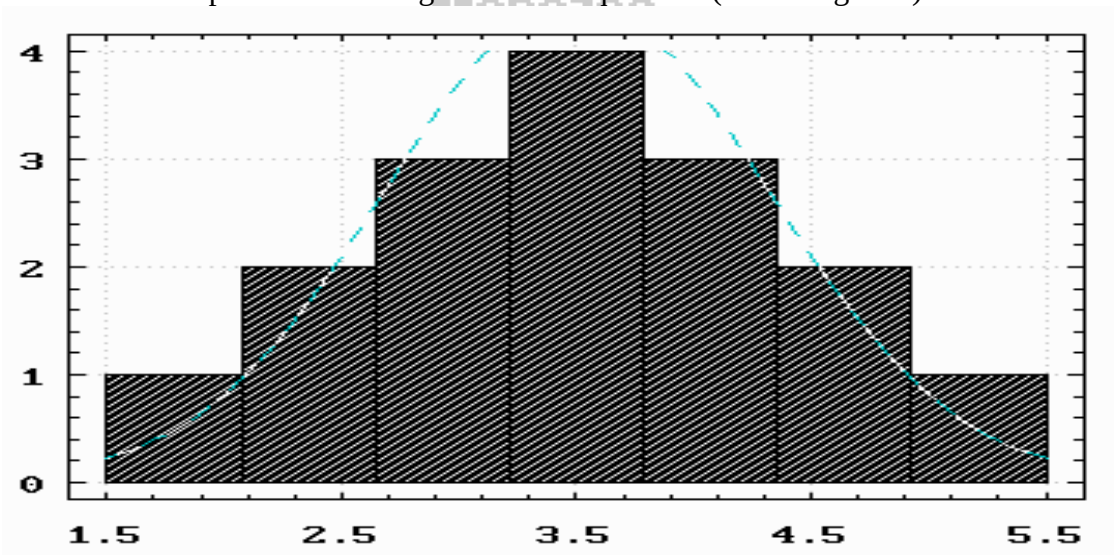


Figura 2 - Histograma de médias amostrais

Se forem calculados a média e a variância das médias de todas as amostras o resultado será:

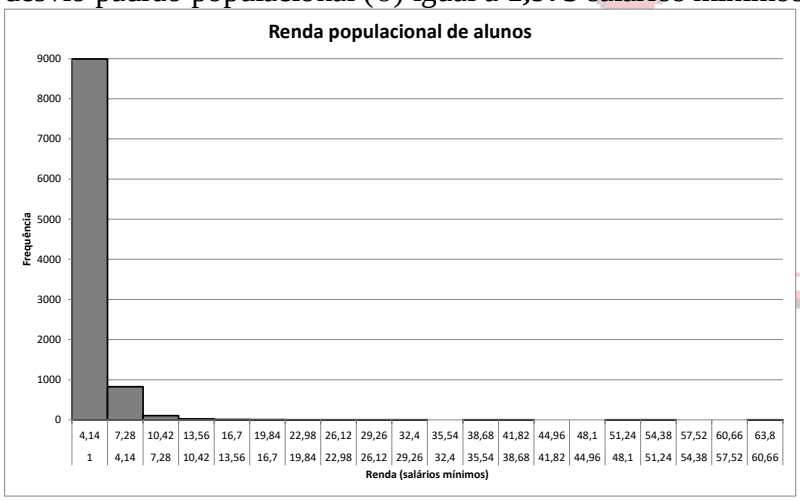
$$\bar{X} = 56/16 = 3{,}5 = \mu \qquad V(\bar{x}) = 0{,}625 = \frac{1{,}25}{2} = \frac{\sigma^2}{n}$$

[4] Elaborado pela professora Carmen Dolores de Freitas de Lacerda.
[5] Há 16 amostras possíveis.

Observe como a distribuição das médias amostrais da variável pode ser aproximada por uma distribuição normal (não obstante a distribuição da variável na população ***não ser normal***), e que o valor esperado das médias amostrais (média das médias) é IGUAL ao valor da média populacional da variável e a variância das médias amostrais é IGUAL ao valor da variância populacional da variável dividida pelo tamanho da amostra[6]. E como consequência desta última afirmação, o desvio padrão das médias amostrais é igual ao desvio padrão populacional dividido pela raiz quadrada do tamanho da amostra[7] ($\sigma(\bar{x}) = \sigma/\sqrt{n}) => \ \sigma(\bar{x}) = \sqrt{1{,}25}/\sqrt{2} = 0{,}7906$. Quanto maior o tamanho da amostra (quanto maior **n**) mais o histograma acima aproximar-se-á de uma distribuição normal, independentemente do formato da distribuição da variável na população.

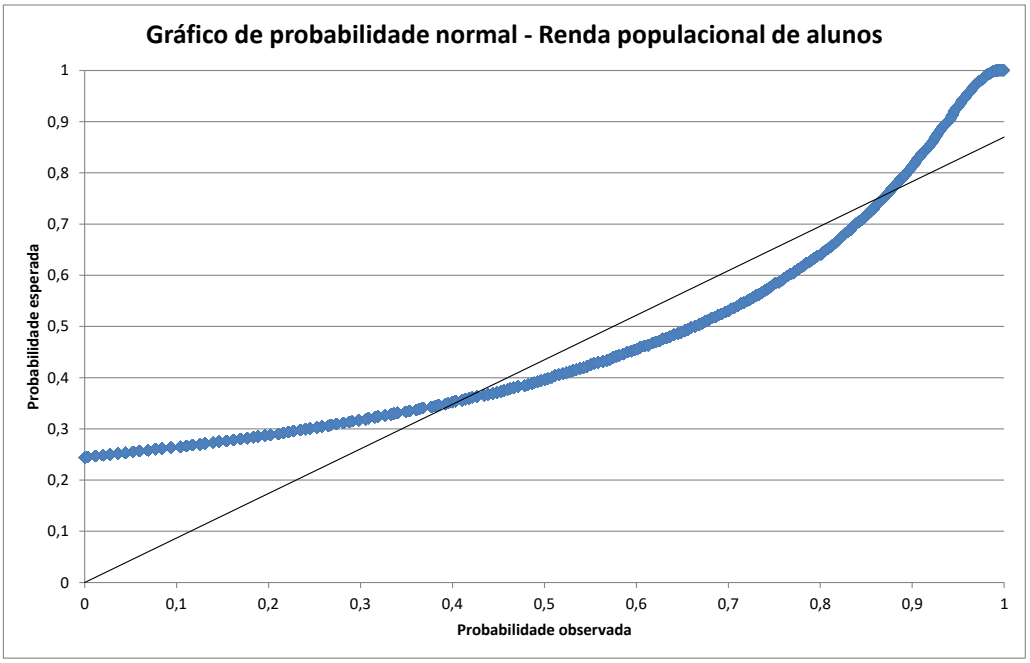
Vamos ver outro exemplo.

Exemplo 9.2 - Na Figura 3 abaixo temos a distribuição populacional (9975 elementos) da variável Renda familiar (em salários mínimos) de alunos de um dos conjuntos de dados usados em semestres anteriores de INE7001. Ela apresenta média populacional (μ) igual a **2,370** salários mínimos, e desvio padrão populacional (σ) igual a **1,973** salários mínimos.



Observe que a distribuição é ASSIMÉTRICA, ou seja, não é normal! Isso pode ser confirmado pela análise do gráfico de probabilidade normal mostrado na Figura 4: percebe-se que os pontos estão distantes da reta teórica, indicando que os dados NÃO seguem uma distribuição normal com média 2,370328 e desvio padrão de 1,972762 salários mínimos.

Figura 3 – Histograma agrupado em classes da Renda familiar populacional de alunos em salários mínimos



[6] Voltaremos a analisar o significado destes resultados quando estudarmos a Estimação por Ponto.
[7] Porque o desvio padrão é a raiz quadrada positiva da amostra.

Figura 4 – Gráfico de probabilidade normal da Renda familiar populacional de alunos em salários mínimos

Imagine que seja possível retirar 1000 amostras aleatórias de n elementos da população mencionada anteriormente. Para cada uma das 1000 amostras pode-se calcular a média das rendas dos alunos: depois calculam-se a média e o desvio padrão das médias amostrais, para comparar com os valores populacionais. Em seguida, constrói-se um histograma e um gráfico de probabilidade normal para as 1000 médias, o que permitirá avaliar a aderência da distribuição das médias amostrais à normal.

Através do suplemento "Análise de Dados" da planilha eletrônica Microsoft Excel ® retiraram-se 1000 amostras aleatórias de 2, 8, 40 e 300 elementos. O procedimento para a retirada das amostras pode ser visto em https://www.youtube.com/watch?v=zkoMzbuYudI&t=340s.

Na Figura 5 está o histograma agrupado em classes das médias das amostras de 2 elementos, e na Figura 6 o gráfico de probabilidade normal das médias amostrais.

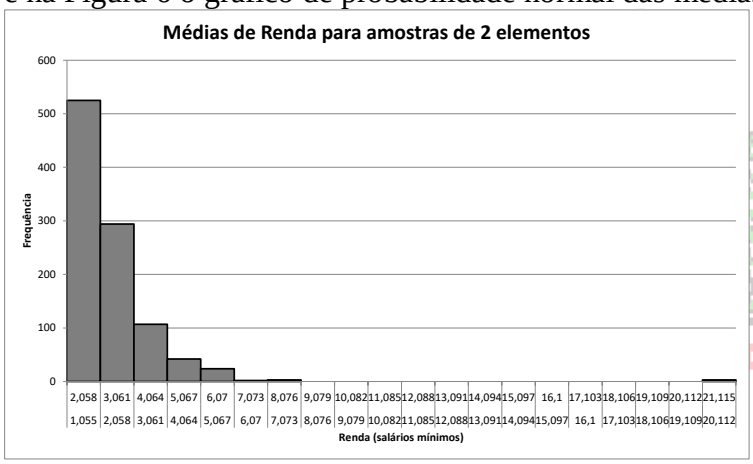
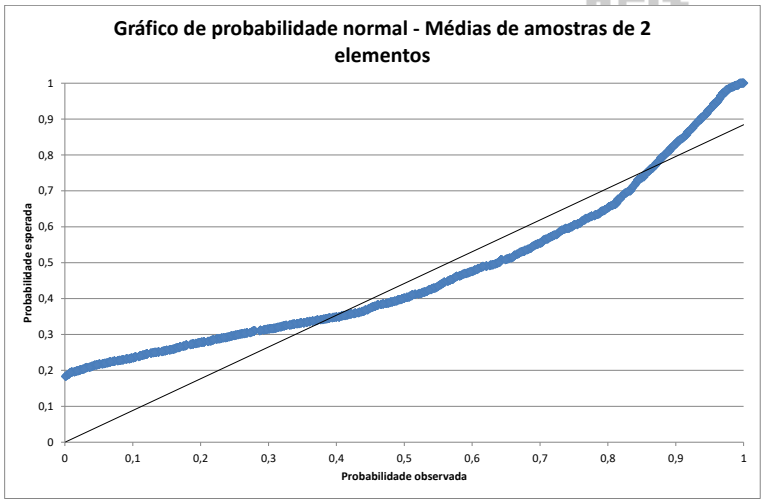


Figura 5 - Distribuição amostral da média (n = 2)

A média das médias amostrais vale 2,370115, e o desvio padrão das médias amostrais vale 1,453793. Calculando $\frac{\sigma}{\sqrt{n}} = 1{,}972762/\sqrt{2} = 1{,}394953$

Observando o histograma vemos que a distribuição das médias, para amostras de 2 elementos, continua assimétrica, mas o valor da média das médias amostrais (2,370115) está muito próximo da média populacional (2,370328), mas o desvio padrão das médias amostrais (1,453793) está um pouco distante de $\sigma/\sqrt{n}$ (1,394953).



Como as amostras têm apenas 2 elementos, a distribuição das médias ainda está muito distante da reta teórica, significando que a distribuição das médias NÃO segue o modelo normal.

Figura 6 – Gráfico de probabilidade normal da média de amostras de 2 elementos da Renda familiar de alunos em salários mínimos

Obviamente o tamanho da amostra utilizada (2 elementos) ainda não foi grande o bastante para levar aos resultados obtidos no Exemplo 9.1 (provavelmente porque a distribuição da população é assimétrica). Por que a média das médias amostrais não coincidiu exatamente com a média populacional, e por que o desvio padrão das médias amostrais não coincidiu exatamente com $\sigma/\sqrt{2}$? A população de onde foram retiradas as amostras têm 9975 elementos, fazendo amostragem com reposição há 9975^2 = 99500625 amostras de 2 elementos possíveis, e foram retiradas apenas 1000...

Aumentando o tamanho de amostra para 8 elementos obtiveram-se os resultados a seguir.

Na Figura 7 está o histograma agrupado em classes das médias das amostras de 8 elementos, e na Figura 8 o gráfico de probabilidade normal das médias amostrais.

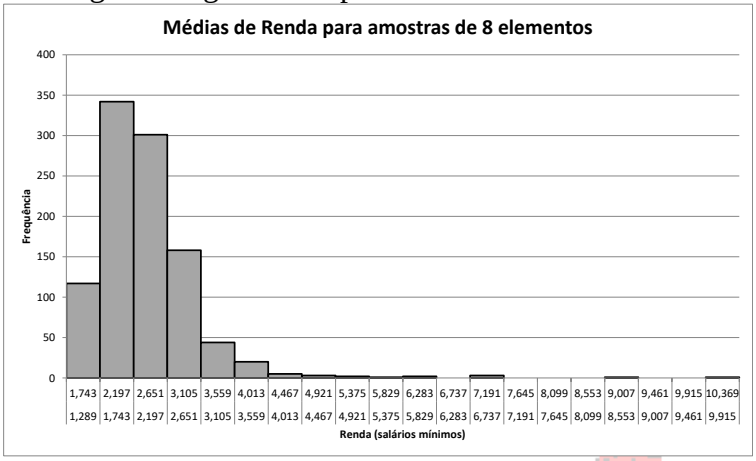
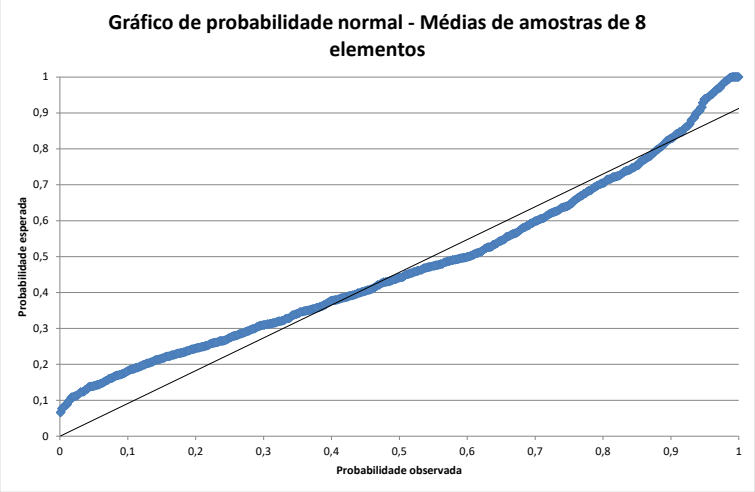


Figura 7 - Distribuição amostral da média (n = 8)

A média das médias amostrais vale 2,359600, e o desvio padrão das médias amostrais vale 0,709258. Calculando $\frac{\sigma}{\sqrt{n}} = 1{,}972762/\sqrt{8} = 0{,}697477$

Observando o histograma vemos que a distribuição das médias, para amostras de 8 elementos, continua assimétrica, o valor da média das médias amostrais (2,359600) está próximo da média populacional (2,370328), e o desvio padrão das médias amostrais (0,709258) está mais próximo de $\sigma/\sqrt{n}$ (0,697477) do que no caso anterior.



Houve uma melhoria em relação ao caso anterior, mas a distribuição das médias ainda está distante da reta teórica, significando que a distribuição das médias NÃO segue o modelo normal.

Figura 8 – Gráfico de probabilidade normal da média de amostras de 8 elementos da Renda familiar de alunos em salários mínimos

Novamente o tamanho da amostra utilizada (8 elementos) ainda não foi grande o bastante para levar aos resultados obtidos no Exemplo 9.1 Vamos agora ver os resultados obtidos para 1000 amostras aleatórias de 40 elementos cada. Na Figura 9 está o histograma agrupado em classes das médias das amostras de 40 elementos, e na Figura 10 o gráfico de probabilidade normal das médias amostrais.

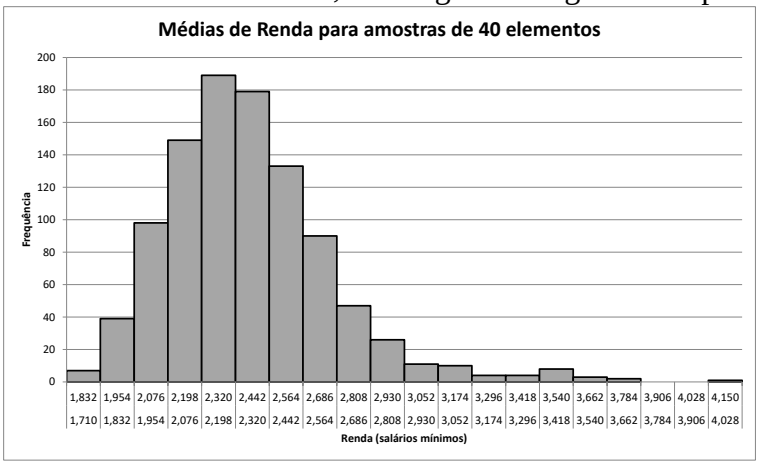
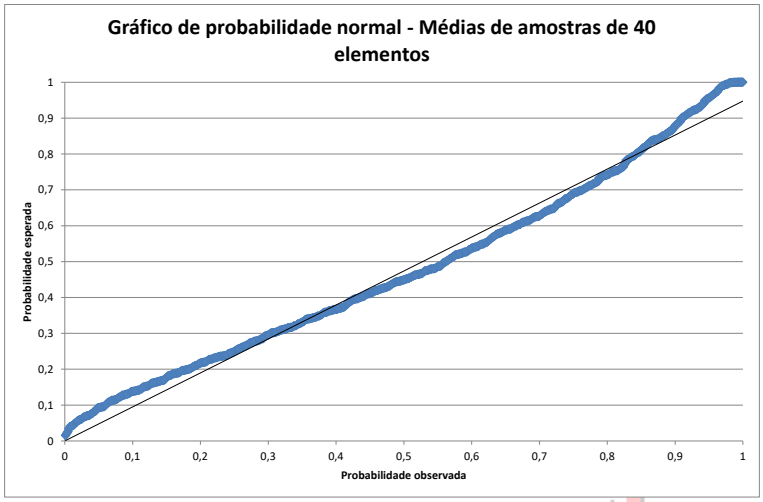


Figura 9 - Distribuição amostral da média (n = 40)

A média das médias amostrais vale 2,368340, e o desvio padrão das médias amostrais vale 0,305290. Calculando $\frac{\sigma}{\sqrt{n}} = 1{,}972762/\sqrt{40} = 0{,}311921$

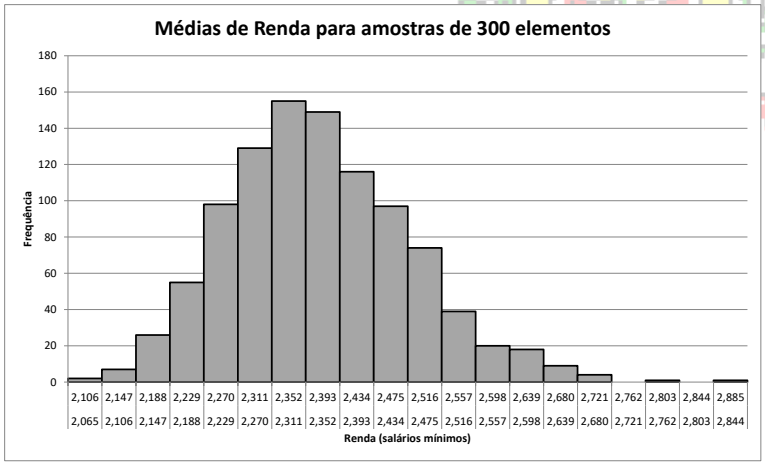
Observando o histograma vemos que a distribuição das médias, para amostras de 40 elementos, continua assimétrica, o valor da média das médias amostrais (2,368340) está próximo da média populacional (2,370328), e o desvio padrão das médias amostrais (0,305290) está próximo de $\sigma/\sqrt{n}$ (0,311921).



Houve uma grande melhoria em relação ao caso anterior, mas os pontos ainda apresentam certa distância da reta teórica, indicando que a distribuição das médias de amostras com 40 elementos ainda não pode ser considerada normal.

Figura 10 – Gráfico de probabilidade normal da média de amostras de 40 elementos da Renda familiar de alunos em salários mínimos

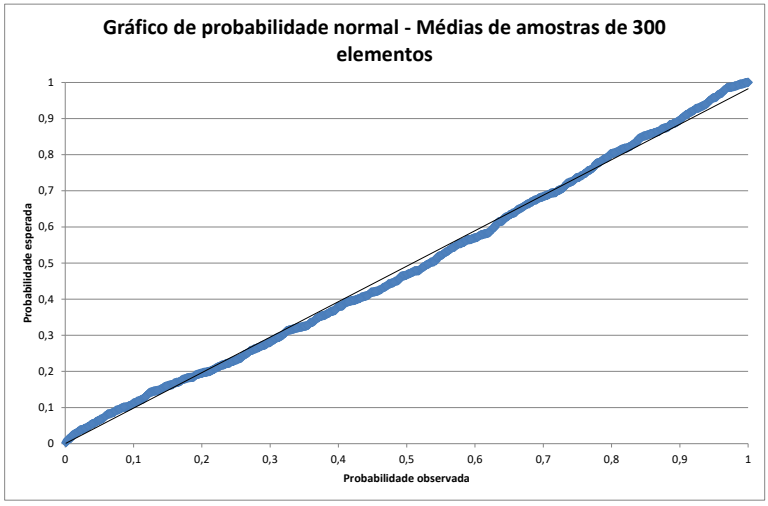
Aumentando a amostra para 300 elementos os resultados obtidos para 1000 amostras são mostrados nas Figuras 11 e 12.



A média das médias amostrais vale 2,368823, e o desvio padrão das médias amostrais vale 0,110282. Calculando $\frac{\sigma}{\sqrt{n}} = 1{,}972762/\sqrt{300}$ = 0,113897

Observando o histograma vemos que a distribuição das médias, para amostras de 300 elementos, está quase simétrica, o valor da média das médias amostrais (2,368823) está próximo da média populacional (2,370328), e o desvio padrão das médias amostrais (0,110282) está próximo de $\sigma/\sqrt{n}$ (0,113897).

Figura 11 - Distribuição amostral da média (n = 300)



Houve uma grande melhoria em relação ao caso anterior, os pontos estão praticamente sobre a reta teórica, indicando que a distribuição das médias de amostras com 300 elementos pode ser considerada normal.

Figura 12 – Gráfico de probabilidade normal da média de amostras de 300 elementos da Renda familiar de alunos em salários mínimos

Podemos então enunciar os teoremas:

TEOREMA DAS COMBINAÇÕES LINEARES

Se a variável de interesse segue uma distribuição normal na população a distribuição amostral das médias de amostras aleatórias retiradas desta população também será normal, independentemente do tamanho destas amostras.

TEOREMA DO LIMITE CENTRAL

Se a variável de interesse não segue uma distribuição normal na população (ou não se sabe qual é a sua distribuição) a distribuição amostral das médias de amostras aleatórias retiradas desta população será normal se o tamanho destas amostras for suficientemente grande, com uma média igual à média populacional e uma variância igual à variância populacional dividida pelo tamanho da amostra.

Em muitos livros afirma-se que com um tamanho de amostra de 30 elementos a distribuição amostral da média já pode ser considerada normal. Como vimos nas Figuras 9 e 10, mesmo para uma amostra com 40 elementos a distribuição das médias não pode ser considerada normal. Já no caso do Exemplo 9.1 para uma amostra com apenas 2 elementos é suficiente para considerar as médias com aderência ao modelo normal. Tudo dependerá da distribuição da população: se a distribuição for simétrica, ou ligeiramente assimétrica, quase que seguramente uma amostra de 30 elementos permitiria constatar o teorema central do limite, com a distribuição das médias amostrais sendo normal. Porém, se a população tiver uma distribuição muito assimétrica (como a mostrada na Figura 3), ou tiver mais de um "pico", 30 elementos podem não ser suficientes. O problema é que o comportamento da população usualmente é *desconhecido*, então precisamos ter alguma outra informação sobre a variável de interesse: por exemplo, variáveis como renda, gastos geralmente têm distribuições assimétricas (recomenda-se então amostras maiores), enquanto outras como medidas corpóreas, dimensões de peças geralmente têm distribuições simétricas (pode-se trabalhar com amostras de 30 elementos, provavelmente a distribuição das médias seguirá uma normal).

Exemplo 9.3[8] - Pense agora em uma variável qualitativa que pode assumir apenas dois valores, e que constitui a seguinte população: (□, □, □, □, ■)

Vamos supor que há interesse no valor ■ (este valor seria o nosso "sucesso"). A proporção deste valor na população (o valor do *parâmetro*) será $\pi = 1/5$.

Se retirarmos todas as amostras aleatórias de 2 elementos (com reposição) possíveis desta população (**n = 2**), teremos os seguintes resultados[9]:

(□,□)	(□,□)	(□,□)	(□,□)	(□, ■)
(□,□)	(□,□)	(□,□)	(□,□)	(□, ■)
(□,□)	(□,□)	(□,□)	(□,□)	(□, ■)
(□,□)	(□,□)	(□,□)	(□,□)	(□, ■)
(□, ■)	(□, ■)	(□, ■)	(□, ■)	(■, ■)

Figura 13 - Amostras de tamanho 2 para proporção

Observe que se definirmos a variável como o número de "sucessos" (número de ■) esta terá uma distribuição **binomial**: há apenas dois resultados possíveis para cada realização, há um número

[8] Elaborado pela professora Carmen Dolores de Freitas de Lacerda.
[9] Há 25 amostras possíveis.

limitado de realizações (**n** = 2 no caso), e cada realização independe da outra (porque a amostra é aleatória com reposição).

Calculando a proporção de ■ em cada uma das amostras, e chamando esta proporção amostral de **p**, teremos os seguintes resultados:

$$p = \begin{matrix} (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (1/2) & (1/2) & (1/2) & (1/2) & (1) \end{matrix}$$

Calculando a média (valor esperado) e a variância das proporções acima teremos:

$$\bar{X} = E(p) = \frac{1}{5} = \pi \qquad s^2 = 0{,}08 = \frac{\left(\frac{1}{5}\right) \times \left(1 - \frac{1}{5}\right)}{2} = \frac{\pi \times (1-\pi)}{n}$$

Observe que o valor esperado (média) das proporções amostrais é IGUAL ao valor da proporção populacional de ■, e que a variância das proporções amostrais é IGUAL ao produto da proporção populacional de ■ por seu complementar, dividido pelo tamanho da amostra[10].

Lembrem-se de que uma distribuição binomial pode ser *aproximada* por uma distribuição normal se algumas condições forem satisfeitas: se o produto do número de realizações pela probabilidade de "sucesso" (**n** x **p**) E o produto do número de realizações pela probabilidade de "fracasso" (**n** x [**1 - p**]) forem ambos maiores ou iguais a 5[11]. E esta distribuição normal teria média igual a **n** x **p** e variância igual a **n** x **p** x (**1 - p**). Se estamos interessados apenas na proporção (probabilidade de "sucesso") e não no número de "sucessos" as expressões anteriores podem ser divididas por **n** (o tamanho da amostra): média igual a **p** e variância igual a [**p** x (**1- p**) / **n**].

Por causa do Teorema do Limite Central é que a distribuição normal é tão importante. Ela representa muito bem uma grande variedade de fenômenos, mas é devido à sua utilização generalizada em Inferência Estatística que o seu estudo é imprescindível. Ressalte-se porém que a sua aplicação costuma resumir-se ao que se chama de Inferência Paramétrica, inferências sobre os parâmetros dos modelos probabilísticos que descrevem as variáveis na população. Para fazer inferências sobre outros aspectos que não os parâmetros, ou quando as amostras utilizadas não forem suficientemente grandes para se assumir a validade do Teorema do Limite Central, é preciso usar técnicas de *Inferência Não Paramétrica* (que nós não veremos nesta disciplina).

9.3 - Estimação por Ponto de Parâmetros

Uma vez tendo decidido que modelo probabilístico é mais adequado para representar a variável de interesse na População resta obter os seus parâmetros. Nos estudos feitos com base em amostras é preciso escolher qual das estatísticas da amostra será o melhor *estimador* para cada parâmetro do modelo. A Estimação por Ponto consiste em determinar qual será o melhor estimador para o parâmetro de interesse.

---

[10] Voltaremos a analisar o significado deste resultado quando estudarmos Estimação por Ponto.
[11] Isto também é decorrência do Teorema Central do Limite.

Como os parâmetros serão estimados através das estatísticas (estimadores) de uma amostra aleatória, e como para cada amostra aleatória as estatísticas apresentarão diferentes valores, os estimadores também terão valores aleatórios. Em outras palavras um Estimador é uma *variável aleatória* que segue uma distribuição de probabilidades.

Naturalmente haverá várias estatísticas **T** que poderão ser usadas como estimadores de um parâmetro θ. Como escolher qual das estatísticas será o melhor estimador para o parâmetro?

Há basicamente três critérios para a escolha de um estimador: o estimador precisa ser justo, consistente e eficiente[12].

*Um Estimador **T** é um estimador justo (não tendencioso) de um parâmetro θ quando o valor esperado de **T** é igual ao valor do parâmetro θ a ser estimado: **E(T) = θ**.*

Um Estimador T é um estimador consistente de um parâmetro θ quando além ser um estimador justo a sua variância tende a zero à medida que o tamanho da amostra aleatória aumenta: $\lim_{n\to\infty} V(T) = 0$.

Se há dois Estimadores justos de um parâmetro o mais eficiente é aquele que apresentar a menor variância.

9.3.1 - Estimação por Ponto dos principais parâmetros

Os principais parâmetros que vamos avaliar aqui são: média de uma variável que segue uma distribuição normal (ou qualquer distribuição se a amostra for suficientemente grande) em uma população (média populacional - μ) e proporção de ocorrência de um dos valores de uma variável que segue uma distribuição de Bernoulli/Binomial[13] em uma população (proporção populacional - π). Em suma escolher quais estatísticas amostrais são mais adequadas para estimar estes parâmetros, usando os critérios definidos acima.

Lembrando dos Exemplos 9.1, 9.2 e 9.3, algumas constatações que lá foram feitas passarão a fazer sentido agora.

Vamos supor que houvesse a intenção de estimar a média populacional da variável do Exemplo 9.1. Qual das estatísticas disponíveis seria o melhor estimador?

Lembrem-se que após retirar todas as amostras aleatórias possíveis daquela população, calcularmos a média de cada amostra, e posteriormente calcularmos a média dessas médias constatou-se que o valor esperado das médias amostrais (média das médias) é IGUAL ao valor da média populacional da variável e a variância das médias amostrais é IGUAL ao valor da variância populacional da variável dividida pelo tamanho da amostra:

$$E(\bar{x}) = \mu \qquad V(\bar{x}) = \frac{\sigma^2}{n}$$

O melhor estimador da média populacional μ é a média amostral $\bar{X}$, pois trata-se de um estimador justo e consistente:

---

[12] Na realidade há mais critérios, mas estes são os mais importantes, maiores detalhes em COSTA NETO, P.O. Estatística, Ed. Edgard Blücher, 1978.

[13] Ambas exigem que experimento seja um experimento de Bernoulli: que tenha (ou possa ser reduzido) a apenas 2 resultados possíveis complementares.

- Justo porque o valor esperado da média amostral será a média populacional;
- Consistente porque se o tamanho da amostra **n** tender ao infinito a variância da média amostral (do Estimador) tenderá a zero.

Agora vamos supor que houvesse a intenção de estimar a proporção populacional do valor ■ da variável do Exemplo 9.3. Qual das estatísticas disponíveis seria o melhor estimador?

Lembrem-se que após retirar todas as amostras aleatórias possíveis daquela população, calcularmos a proporção de ■ em cada amostra, e posteriormente calcularmos a média dessas proporções constatou-se que o valor esperado das proporções amostrais (média das proporções) é IGUAL ao valor da proporção populacional do valor ■ da variável e a variância das proporções amostrais é IGUAL ao valor do produto da proporção populacional do valor ■ da variável pela sua complementar dividida pelo tamanho da amostra:

$$E(p) = \pi \qquad V(p) = \frac{\pi \times (1-\pi)}{n}$$

O melhor estimador da média populacional π é a proporção amostral **p**, pois se trata de um estimador justo e consistente:

- Justo porque o valor esperado da proporção amostral será a proporção populacional;
- Consistente porque se o tamanho da amostra **n** tender ao infinito a variância da proporção amostral (do Estimador) tenderá a zero.

Poderíamos fazer um procedimento semelhante para estimar outros parâmetros, como, por exemplo, a variância populacional de uma variável. Este procedimento não será demonstrado, mas o melhor estimador da variância populacional será a variância amostral SE FOR USADO **n - 1** NO DENOMINADOR DA EXPRESSÃO DE CÁLCULO[14]. Somente assim a variância amostral será um estimador justo (não viciado) da variância populacional.

Como o desvio padrão é a raiz quadrada da variância é comum estimar o desvio padrão populacional extraindo a raiz quadrada da variância amostral.

9.4 - Estimação por Intervalo de Parâmetros

Geralmente uma inferência estatística é feita com base em uma única amostra: na maior parte dos casos é totalmente inviável retirar todas as amostras possíveis de uma determinada população.

Intuitivamente percebemos que as estatísticas calculadas nessa única amostra, mesmo sendo os melhores estimadores para os parâmetros de interesse, terão uma probabilidade *infinitesimal* de coincidir exatamente com os valores reais dos parâmetros. Então a Estimação por Ponto dos parâmetros é insuficiente, e as estimativas assim obtidas servirão apenas como referência para a **Estimação por Intervalo**.

"A Estimação por Intervalo consiste em colocar um ***Intervalo de Confiança*** (I.C.) em torno da estimativa obtida através da Estimação por Ponto".

O Intervalo de Confiança terá uma certa probabilidade chamada de ***Nível de confiança*** (que costuma ser simbolizado como **1 - α**) de conter o valor real do parâmetro: fazer uma Estimação por

[14] Esta é a "razão matemática" a que se referia a nota que tratava de variância amostral no Capítulo 2 de INE7001.

Intervalo de um parâmetro é efetuar uma afirmação probabilística sobre este parâmetro, indicando uma faixa de possíveis valores, e a probabilidade de que esta faixa realmente contenha o valor real do parâmetro. A probabilidade de que o Intervalo de Confiança **não** contenha o valor real do parâmetro é chamada de ***Nível de Significância*** (α), e o valor desta probabilidade será o complementar do Nível de Confiança. É comum definir o Nível de Significância como uma probabilidade máxima de erro, um risco máximo admissível.

A determinação do Intervalo de Confiança para um determinado parâmetro resume-se basicamente a definir o Limite Inferior e o Limite Superior do intervalo, supondo um determinado Nível de Confiança (ou Significância). A definição dos limites dependerá também da *distribuição amostral* da estatística usada como referência para o intervalo e do tamanho da amostra utilizada.

Para os dois parâmetros em que temos maior interesse (média populacional μ e proporção populacional π) a distribuição amostral dos estimadores (média amostral $\bar{x}$ e proporção amostral **p**, respectivamente) pode ser aproximada por uma distribuição normal:[15] o Intervalo de Confiança será então simétrico em relação ao valor calculado da estimativa (média ou proporção amostral), com base na amostra aleatória coletada.

Figura 14 - Intervalo de Confiança para uma distribuição normal

Onde: **Li** é o limite inferior e **Ls** é o limite superior do Intervalo de Confiança; **1 - α** é o Nível de Confiança estabelecido, observando que o valor do Nível de Significância α é dividido igualmente entre os valores abaixo de **Li** e acima de **Ls**.

Para obter os limites em função do Nível de Confiança devemos utilizar a *distribuição normal padrão* (variável **Z** com média zero e variância um): fixar um certo valor de probabilidade, obter o valor de Z correspondente, e substituir o valor em **Z = (x - "média")/ "desvio padrão"**[16], para obter o valor **x** (valor correspondente ao valor de **Z** para a probabilidade fixada). Observe a figura abaixo:

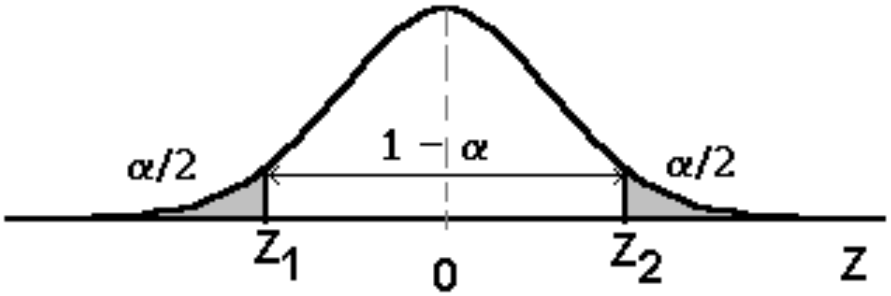


Figura 15 - Intervalo de Confiança para a distribuição normal padrão

[15] Isso não ocorre para todos os parâmetros, para a variância, por exemplo, se a variável tiver distribuição normal na população a distribuição amostral será a Quiquadrado, que é assimétrica.
[16] Foram colocados entre aspas porque os valores dependerão dos parâmetros sob análise e de outros fatores.

O limite **Li** (inferior) corresponde a Z_1 e o limite **Ls** (superior) corresponde a Z_2. O ponto central 0 (zero) corresponde ao valor calculado da Estimativa. Como a variável **Z** tem distribuição normal com média igual a zero (lembrando que a distribuição normal é simétrica em relação à média) os valores de Z_1 e Z_2 serão iguais em módulo (Z_1 será negativo e Z_2 positivo):

Z_1 será um valor de **Z** tal que $P(Z \le Z_1) = \frac{\alpha}{2}$, e Z_2 será um valor tal que $P(Z \le Z_2) = 1 - \frac{\alpha}{2}$

Então obteremos os valores dos limites através das expressões:

Z_1 = (Li - "média")/ "desvio padrão" => Li = "média" + Z_1 × "desvio padrão"

Z_2 = (Ls - "média")/ "desvio padrão" => Ls = "média" + Z_2 × "desvio padrão"

Como $Z_1 = -Z_2$, podemos substituir:

Li = "média" - Z_2 × "desvio padrão" Ls = "média" + Z_2 × "desvio padrão"

E este valor Z_2 é chamado de $Z_{crítico}$, porque corresponde aos limites do intervalo[17]:

Li = "média" - $Z_{crítico}$ × "desvio padrão"

Ls = "média" + $Z_{crítico}$ × "desvio padrão"

Reparem que o mesmo valor é somado e subtraído da "média". Este valor é chamado de semi-intervalo ou precisão do intervalo, e recebe símbolo e_0:

$$e_0 = Z_{crítico} \times \text{"desvio padrão"}$$

Resta agora definir corretamente o valor da "média" e do "desvio padrão" para cada um dos parâmetros em que estamos interessados (média e proporção populacional). Com base nas conclusões obtidas na Estimação por Ponto isso será simples. Contudo, há alguns outros aspectos que precisarão ser esmiuçados.

9.4.1 - Estimação por Intervalo da Média Populacional μ

Lembrando das expressões anteriores:

Li = "média" - $Z_{crítico}$ × "desvio padrão" = "média" - e_0

Ls = "média" + $Z_{crítico}$ × "desvio padrão" = "média" + e_0

Neste caso a "**média**" será a média amostral $\bar{x}$ (ou mais precisamente o seu valor):

$$P(\bar{x} - e_0 \le \mu \le \bar{x} + e_0) = 1 - \alpha$$

O valor de e_0 dependerá de outros aspectos.

a) Se a variância populacional σ^2 da variável (cuja média populacional queremos estimar) for conhecida.

Neste caso a variância amostral da média poderá ser calculada através da expressão:

$V(\bar{x}) = \frac{\sigma^2}{n}$, e, por conseguinte, o "desvio padrão" será $\text{desvio padrao} = \frac{\sigma}{\sqrt{n}}$

[17] Esta notação é a utilizada na apostila de Roteiros e Tabelas.

E $\mathbf{e_0}$ será: $$e_0 = Z_{critico} \times \frac{\sigma}{\sqrt{n}}$$

Bastará então fixar o Nível de Confiança (ou de Significância) para obter $\mathbf{Z_{crítico}}$ e calcular $\mathbf{e_0}$.

b) Se a variância populacional σ^2 da variável for *desconhecida*.

Nestes casos procede-se como no item anterior, apenas fazendo com que σ = **s**, ou seja, considerando que o desvio padrão da variável na população é uma boa estimativa do desvio padrão da variável na amostra (suposição razoável para grandes amostras). Contudo, não será possível usar a distribuição normal padrão (**Z**) através da distribuição **t de Student**. Trata-se de uma distribuição de probabilidades que apresenta média igual a zero (como a normal padrão), é simétrica em relação à média, mas apresenta uma variância igual a **n / (n - 2)**, ou seja seus valores dependem do tamanho da amostra, apresentando maior variância para menores valores de amostra[18]. Quanto maior o tamanho da amostra mais a variância de **t** aproxima-se de 1,00 (variância da normal padrão)[19]. A distribuição t de Student está na figura abaixo:

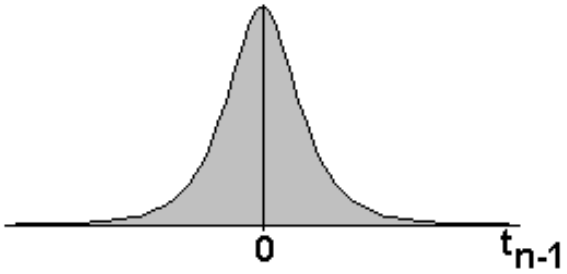


Figura 16 - Distribuição t de Student

Observe que tal como a distribuição normal padrão a distribuição **t** de Student é *simétrica* em relação à média (que é igual a zero).

O valor **n -1** (tamanho da amostra menos 1) é chamado de número de *graus de liberdade* da estatística. Quando a variância amostral é calculada supõe-se que a média já seja conhecida, assim apenas um determinado número de elementos da amostra poderá ter seus valores variando livremente, este número será igual a **n - 1**, porque um dos valores não poderá variar livremente, pois terá que ter um valor tal que a média permaneça a mesma calculada anteriormente. Assim, a estatística terá **n - 1** graus de liberdade.

A distribuição **t** de Student encontra-se tabelada no apêndice para vários graus de liberdade e valores de probabilidade.

Quando a variância populacional da variável é desconhecida e a amostra tem até 30 elementos substitui-se σ por **s** e **Z** por $\mathbf{t_{n-1}}$ em todas as expressões para determinação dos limites do intervalo de confiança, obtendo:

$$\mathbf{Li = \text{“média”} - t_{n-1,crítico} \times \text{“desvio padrão”} = \text{“média”} - e_0}$$
$$\mathbf{Ls = \text{“média”} + t_{n-1,crítico} \times \text{“desvio padrão”} = \text{“média”} + e_0}$$

[18] Esta é a correção propriamente dita, pois ao usar pequenas amostras o risco de que a variância amostral da variável seja diferente da variância populacional é maior, podendo levar a intervalos de confiança que não correspondem à realidade. A não utilização desta correção foi a fonte de muitos erros no passado, e, infelizmente, de ainda alguns erros no presente.

[19] Para tamanhos de amostra maiores do que 50 supõe-se que a variância de **t** é igual a 1: o que permite a utilização da distribuição normal padrão.

E **e_0** será: $$e_0 = t_{n-1,critico} \times \frac{s}{\sqrt{n}}$$

Os valores de **$t_{n-1,crítico}$** podem ser obtidos de forma semelhante aos de **$Z_{crítico}$**, definindo o Nível de Confiança (ou de Significância), mas precisam também da definição do número de graus de liberdade (**n - 1**): tendo estes valores basta procurar o valor em uma tabela ou em um programa computacional.

c) Se a amostragem for probabilística estratificada proporcional

Vimos na seção 7.2.3.1 do Capítulo 7 que quando a amostragem é estratificada proporcional a obtenção da estimativa da média total da amostra precisa ponderar as médias da variável em cada estrato pelas frações deles na população. E, é preciso fazer uma estimativa da variância (para depois obter o desvio padrão) total da amostra, também ponderando as variâncias em cada estrato pelas frações deles na população. Naquela seção obteve-se:

$$\bar{x}_{Estratificada} = \sum_{i=1}^{E} W_i \times \bar{x}_i$$

Onde:
- E é o número de estratos identificados na população;
- $\bar{x}_i$ é a média de uma variável quantitativa na parte da amostra retirada de um estrato i qualquer;
- W_i é a proporção que o estrato i têm na população (N_i/N, onde N_i é o tamanho do estrato i e N é o tamanho total da população).

De acordo com autores[20] é possível obter uma estimativa para a variância de $\bar{x}_{estratificada}$ (e portando do seu desvio padrão):

$$s^2{}_{\bar{x}\,estratificada} = \sum_{i=1}^{E} \left[\frac{W_i{}^2 \times s^2{}_i}{n_i} - \frac{W_i \times s^2{}_i}{N} \right]$$

Onde:
- s^2_i é a variância amostral da variável no estrato i;
- n_i é o tamanho da amostra no estrato i.

Então **e_0** será:

$$e_0 = Z_{crítico} \times \sqrt{s^2{}_{\bar{x}\,estratificada}}$$

Como geralmente em amostragem estratificada proporcional as amostras em cada estrato são grandes é possível considerar a aproximação pela normal, e usar Z da distribuição normal. Lembre-se que o **e_0** será somado e subtraído à $\bar{x}_{estratificada}$ para obter os limites do intervalo de confiança.

Se o tamanho da amostra (n) for superior a 5% do tamanho da população (N) os valores de e_0 precisam ser **corrigidos**. Caso contrário os limites dos intervalos não serão acurados. A correção é mostrada na equação a seguir:

[20] COCHRAN, W.G. Técnicas de Amostragem, Rio de Janeiro: Fundo de Cultura, 1965 COCHRAN, W.G. Técnicas de Amostragem, Rio de Janeiro: Fundo de Cultura, 1965 e BOLFARINE, H., BUSSAB, W.O. Elementos de Amostragem, São Paulo: Edgar Blücher, 2005.

$$e_{0_{corrigido}} = e_0 \times \sqrt{\frac{N-n}{N-1}}$$

9.4.2 - Estimação por Intervalo da Proporção Populacional π

No item **9.3** declaramos que o melhor estimador para a proporção populacional $\boldsymbol{\pi}$ é a proporção amostral **p**. E que esta proporção amostral teria média igual a $\boldsymbol{\pi}$ e variância igual a $[\boldsymbol{\pi} \times (\mathbf{1} - \boldsymbol{\pi})]/\mathbf{n}$ onde **n** é o tamanho da amostra aleatória. A distribuição da proporção amostral **p** é binomial, e sabe-se que a distribuição binomial pode ser aproximada por uma normal se algumas condições forem satisfeitas, no caso se:

$$\mathbf{n} \times \boldsymbol{\pi} \geq 5 \text{ E } \mathbf{n} \times (\mathbf{1} - \boldsymbol{\pi}) \geq 5.$$

Ora, se $\boldsymbol{\pi}$ fosse conhecido não estaríamos aqui nos preocupando com a sua Estimação por Intervalo, assim vamos verificar se é possível aproximar a distribuição binomial de **p** por uma normal se:

$\mathbf{n} \times \mathbf{p} \geq 5$ E $\mathbf{n} \times (\mathbf{1} - \mathbf{p}) \geq 5$, ou seja usando o próprio valor da proporção amostral observada (trata-se de uma aproximação razoável).

SE E SOMENTE SE estas duas condições forem satisfeitas poderemos usar as expressões abaixo (lembrando das expressões anteriores):

Li = "média" - $Z_{crítico}$ × "desvio padrão" = "média" - e_0

Ls = "média" + $Z_{crítico}$ × "desvio padrão" = "média" + e_0

Neste caso a "**média**" será a proporção amostral (ou mais precisamente o seu valor):

$$P(p - e_0 \leq \mu \leq p + e_0) = 1 - \alpha$$

E o valor do "desvio padrão" será igual a $\sqrt{\frac{\pi \times (1-\pi)}{n}}$. Novamente, como $\boldsymbol{\pi}$ é desconhecido, usaremos a proporção amostral **p** como aproximação.

Então $\mathbf{e_0}$ será: $$e_0 = Z_{critico} \times \sqrt{\frac{p \times (1-p)}{n}}$$

Bastará então fixar o Nível de Confiança (ou de Significância), $\mathbf{Z_{crítico}}$ e calcular $\mathbf{e_0}$.

Novamente, para o caso de amostragem estratificada proporcional, recorda-se a seção 7.2.3.1, que mostrava uma estimativa para a proporção amostral total:

$$p_{Estratificada} = \sum_{i=1}^{E} W_i \times p_i$$

Onde:
- E é o número de estratos identificados na população;
- p_i é a proporção do valor de interesse de uma variável qualitativa na parte da amostra retirada de um estrato i qualquer;
- W_i é a proporção que o estrato i têm na população.

E, tal como no caso da média, é possível obter uma estimativa para a variância (e posteriormente o desvio padrão) de $p_{estratificada}$:

$$s^2{}_{p\ estratificada} = \sum_{i=1}^{E}\left[W_i{}^2 \times \frac{(N_i - n_i)}{(N_i - 1)} \times \frac{p_i \times (1 - p_i)}{n_i - 1}\right]$$

Onde:
- n_i é o tamanho da amostra no estrato i.

Então $\mathbf{e_0}$ será:

$$e_0 = Z_{crítico} \times \sqrt{s^2{}_{p\ estratificada}}$$

Lembre-se que o $\mathbf{e_0}$ será somado e subtraído à $p_{estratificada}$ para obter os limites do intervalo de confiança.

Novamente, precisamos corrigir o valor de e_0 para o caso de população finita:

$$e_{0_{corrigido}} = e_0 \times \sqrt{\frac{N-n}{N-1}}$$

Em suma a Estimação por Intervalo da média e da proporção populacional consiste basicamente em calcular a amplitude do semi-intervalo (o $\mathbf{e_0}$), de acordo com as condições do problema sob análise.
- Para a média, observar se é viável considerar que a distribuição da variável na população é normal, ou que a amostra seja suficientemente grande para que a distribuição das médias amostrais possa ser considerada normal. Se isso for verificado, identificar se a variância populacional da variável é conhecida: caso seja deverá ser usada a variável **Z** da distribuição normal padrão, para qualquer tamanho de amostra. Se variância populacional da variável é desconhecida formalmente deve ser usada a variável $\mathbf{t_{n-1}}$ da distribuição de Student.
- Para a proporção, observar se é possível fazer a aproximação pela distribuição normal.
- Se a amostragem for estratificada proporcional é preciso ponderar os valores pelas proporções de cada estrato na população.

9.5 - Tamanho Mínimo de Amostra para Estimação por Intervalo de parâmetros

Como foi observado a determinação dos limites de um Intervalo de Confiança (determinação do $\mathbf{e_0}$) dependem do tamanho da amostra aleatória coletada, além do Nível de Confiança e da distribuição amostral do estimador utilizado. Nada podemos fazer quanto à distribuição amostral do estimador, o Nível de Confiança nós podemos controlar, seria interessante definir então uma precisão (um valor para $\mathbf{e_0}$) para o Intervalo de Confiança: é muito comum querermos estabelecer previamente qual será a faixa de variação de um determinado parâmetro, com uma certa confiabilidade.

Contudo, para um *mesmo* tamanho de amostra:
- se aumentarmos o Nível de Confiança (reduzirmos o Nível de Significância) teremos um valor crítico maior, o que aumentará o valor de $\mathbf{e_0}$, resultando em um Intervalo de Confiança mais "largo", com menor precisão.
- se resolvermos aumentar a precisão (menor valor de $\mathbf{e_0}$), obter um Intervalo de Confiança mais "estreito", teremos uma queda no Nível de Confiança.

A solução é obter um **tamanho mínimo de amostra** capaz de atender simultaneamente ao Nível de Confiança (ou de Significância) e à precisão (**e_0**) especificados. Como as expressões de **e_0** são em função do tamanho de amostra (**n**), é razoável pensar em reordená-las de forma a fazer com que o tamanho de amostra seja função do Nível de Confiança e da precisão (**e_0**).

9.5.1 - Tamanho Mínimo de Amostra para Estimação por Intervalo da média populacional

a) Variância populacional conhecida

$$e_0 = Z_{critico} \times \frac{\sigma}{\sqrt{n}} \qquad \text{isolando } \mathbf{n}: \qquad n = \left(\frac{Z_{critico} \times \sigma}{e_0} \right)^2$$

Neste caso basta especificar o valor de **e_0** (na mesma unidade do desvio padrão populacional **σ**), e o Nível de Confiança (que será usado para encontrar o **$Z_{crítico}$**) e calcular o tamanho mínimo de amostra.

b) Variância populacional desconhecida

$$e_0 = t_{n-1,critico} \times \frac{s}{\sqrt{n}} \qquad \text{isolando } \mathbf{n}: \qquad n = \left(\frac{t_{n-1,critico} \times s}{e_0} \right)^2$$

O procedimento neste caso seria semelhante exceto por um pequeno problema: "se estamos calculando o tamanho da amostra como podemos conhecer **n - 1** e o desvio padrão amostral **s**?"

Quando a variância populacional da variável é desconhecida o usual é retirar uma amostra piloto com um tamanho **n*** arbitrário. A partir dos resultados desta amostra são calculadas as estatísticas (entre elas o desvio padrão amostral **s**) que são substituídas na expressão acima.

Se **n** ≤ **n*** então a amostra piloto é **suficiente** para o Nível de Confiança e a precisão exigidos.

Se **n** > **n*** então a amostra piloto é **insuficiente** para o Nível de Confiança e a precisão exigidas, sendo então necessário retornar à população e retirar os elementos necessários para completar o tamanho mínimo de amostra. O processo continua até que a amostra seja considerada suficiente.

9.5.2 - Tamanho Mínimo de Amostra para Estimação por Intervalo da proporção populacional

Para a proporção populacional teremos:

$$e_0 = Z_{critico} \times \sqrt{\frac{p \times (1-p)}{n}} \qquad \text{isolando n:} \qquad n = \left(\frac{Z_{critico}}{e_0} \right)^2 \times p \times (1-p)$$

É necessário especificar o Nível de Confiança (ou de Significância) que será usado para encontrar o **$Z_{crítico}$**, e o valor de **e_0** (tomando o cuidado de que tanto **e_0** quanto **p** e **1- p** estejam **todos** como proporções adimensionais ou como percentuais) para que seja possível calcular o valor do tamanho mínimo de amostra.

Da mesma forma que no caso da Estimação da média quando a variância populacional é desconhecida teremos que recorrer à uma amostra piloto, procedendo de forma semelhante à letra b) do item **9.5.1**. No cálculo do tamanho mínimo de amostra para a Estimação por Intervalo da

proporção populacional há porém uma solução alternativa: utiliza-se uma estimativa exagerada[21] da amostra, supondo o máximo valor possível para o produto **p** × (**1 - p**), que ocorrerá quando ambas as proporções forem iguais a 0,5 (50%).

Conforme visto no Capítulo 7, se o tamanho da população for conhecido é recomendável corrigir o tamanho da amostra obtida, seja para o intervalo de confiança de média ou proporção, através da seguinte fórmula:

$$n_{corrigido} = \frac{N \times n}{N + n} \qquad \text{onde } N \text{ é o tamanho da população}$$

Assim procedendo, evitamos o inconveniente de obter um tamanho de amostra superior ao tamanho da população, o que pode ocorrer se N não for muito grande.

Exemplo 9.4 - Retirou-se uma amostra aleatória de 4 elementos de uma produção de cortes bovinos no intuito de estimar a média do peso do corte. Obteve-se média de 8,2 kg e desvio padrão de 0,4 kg. Supondo população normal.

a) Determinar um intervalo de confiança para a média populacional com 1% de significância.

b) Para estimar a média, com 1% de significância e precisão de 0,2 kg, esta amostra é suficiente?

a) Seguindo o roteiro de Estimação de Parâmetros do apêndice:

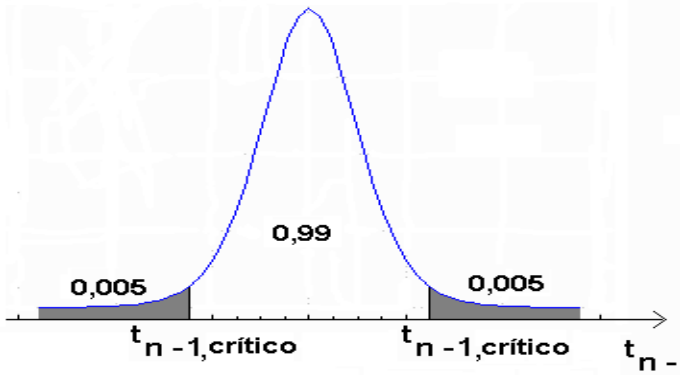
1) O parâmetro de interesse é a média populacional μ *do peso do corte.*

2) Adotou-se um nível de significância de 1%, então $\alpha = 0{,}01$ e $\mathbf{1 - \alpha} = 0{,}99$[22].

3) As estatísticas disponíveis são: ***média amostral*** *= 8,2 kg* **s** = 0,4 kg **n** = 4 *elementos.*

4) Definição da variável de teste: como a variância populacional é DESCONHECIDA, não obstante a população ter distribuição normal, a distribuição amostral da média será **t** *de Student, e a variável de teste será* $\mathbf{t_{n-1}}$.

5) Encontrar o valor de $\mathbf{t_{n-1,crítico}}$ *: como o Intervalo de Confiança para a média é bilateral, teremos uma situação semelhante à da figura abaixo:*



Para encontrar o valor crítico devemos procurar na tabela da distribuição de Student, na linha correspondente a **n-1** *graus de liberdade, ou seja, em 4 - 1 = 3 graus de liberdade. O valor da probabilidade pode ser visto na figura ao lado: os valores críticos serão* $\mathbf{t_{3;0,005}}$ e $\mathbf{t_{3;0,995}}$ *os quais serão iguais em módulo. E o valor de* $\mathbf{t_{n-1,crítico}}$ *será igual a* **5,84** *(em módulo)*

6) Determinam-se os limites do intervalo, através da expressão abaixo (cujo resultado será somado e subtraído da média amostral) para determinar os limites do intervalo:

$$e_0 = \frac{t_{n-1,crítico} \times s}{\sqrt{n}} = \frac{5{,}84 \times 0{,}4}{\sqrt{4}} = 1{,}168\ kg$$

$$L_I = \bar{x} - e_0 = 8{,}2 - 1{,}168 = 7{,}032\,kg \qquad L_S = \bar{x} + e_0 = 8{,}2 + 1{,}168 = 9{,}368\,kg$$

7) Então o intervalo de 99% de confiança para a média populacional da dimensão é [7,032;9,368] kg.

[21] Esta solução somente é usada quando a natureza da pesquisa é tal que não é possível retirar uma amostra piloto: a retirada de uma amostra piloto e a eventual retirada de novos elementos da população poderia prejudicar muito o resultado da pesquisa. Paga-se, então, o preço de ter uma amostra substancialmente maior do que talvez fosse necessário.

[22] Este valor pode ser arbitrado pelo usuário ou pode ser uma exigência do problema sob análise, ou até mesmo uma exigência legal. Os níveis de significância mais comuns são de 1%, 5% ou mesmo 10%.

Interpretação: há 99% de probabilidade de que a verdadeira média populacional do peso de corte esteja entre 7,032 e 9,368 kg.

b) Como a variância populacional é DESCONHECIDA, não obstante a população ter distribuição normal, a distribuição amostral da média será **t** *de Student, e a variável de teste será* $\mathbf{t_{n-1}}$. *Assim será usada a seguinte expressão para calcular o tamanho mínimo de amostra para a estimação por intervalo da média populacional.*

$$n = \left(\frac{t_{n-1,critico} \times s}{e_0} \right)^2$$

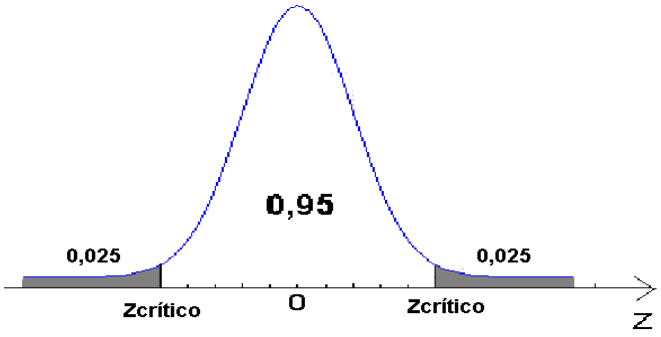
O nível de significância é o mesmo do item a. Sendo assim, o valor crítico continuará sendo o mesmo: $\mathbf{t_{n-1,crítico}}$ = 5,84. *O desvio padrão amostral vale 0,4 kg, e o valor de* $\mathbf{e_0}$, *a precisão, foi fixado em 0,2 kg. Basta então substituir os valores na expressão:*

$$n = \left(\frac{t_{n-1,critico} \times s}{e_0} \right)^2 = \left(\frac{5,84 \times 0,4}{0,2} \right)^2 = 136,42 \cong 137 \ \ elementos$$

Observe que o tamanho mínimo de amostra necessário para atender a 1% de significância e precisão de 0,2 kg deveria ser de 137 elementos. Como a amostra coletada possui apenas 4 elementos ela é INSUFICIENTE para a significância e precisão exigidas. Recomenda-se o retorno à população para a retirada aleatória de mais 133 cortes.

Exemplo 9.5 - Retirou-se uma amostra aleatória de 1000 peças de um lote. Verificou-se que 35 eram defeituosas.
a) Determinar um intervalo de confiança de 95% para a proporção peças defeituosas no lote.
b) Supondo 99% de confiança e precisão de 1%, esta amostra é suficiente para estimar a proporção populacional?

1) O parâmetro de interesse é a proporção populacional $\boldsymbol{\pi}$ *de peças defeituosas.*
2) Adotou-se um nível de significância de 5%, então $\boldsymbol{\alpha}$ = 0,05 e **1 -** $\boldsymbol{\alpha}$ = 0,95
3) As estatísticas são: ***proporção amostral de peças defeituosas*** **p** = 35/1000 **n** = 1000 *elementos.*
4) Definição da variável de teste: precisamos verificar se é possível fazer a aproximação pela normal, então **n** x **p** = 1000 x 0,035 = 35 > 5 e **n** x (**1- p**) = 1000 x 0,965 = 965 > 5. *Como ambos os produtos satisfazem as condições para a aproximação podemos usar a variável* **Z** *da distribuição normal padrão*
5) Encontrar o valor de $\mathbf{Z_{crítico}}$: *como o Intervalo de Confiança para a média é bilateral, teremos uma situação semelhante à da figura abaixo:*



Para encontrar o valor crítico devemos procurar na tabela da distribuição normal padrão pela probabilidade 0,975 (0,95+0,025) O valor da probabilidade pode ser visto na figura ao lado: os valores críticos serão $\mathbf{Z_{0,025}}$ e $\mathbf{Z_{0,975}}$ *os quais serão iguais em módulo. E o valor de* $\mathbf{Z_{crítico}}$ *será igual a 1,96 (em módulo).*

6) Passa-se agora a determinação dos limites do intervalo, através da expressão abaixo (cujo resultado será somado e subtraído da proporção amostral de peças defeituosas) para determinar os limites do intervalo:

$$e_0 = Z_{critico} \times \sqrt{\frac{p \times (1-p)}{n}} = 1,96 \times \sqrt{\frac{0,035 \times 0,965}{1000}} = 0,0114$$

$$L_I = p - e_0 = 0,035 - 0,0114 = 0,0236 \qquad L_S = p + e_0 = 0,035 + 0,0114 = 0,0464$$

7) Então, o intervalo de 95% de confiança para a proporção populacional de peças defeituosas é [2,36%;4,64%].
Interpretação: há 95% de probabilidade de que a verdadeira proporção populacional de plantas atacadas pelo fungo esteja entre 2,36% e 4,64%.

b) *De acordo com o item anterior é possível utilizar a aproximação pela distribuição normal. A expressão para o cálculo do tamanho mínimo de amostra para a proporção populacional será:*

$$n = \left(\frac{Z_{critico}}{e_0} \right)^2 \times p \times (1-p)$$

Os valores de **p** *e* **1 - p** *já são conhecidos*: **p** = 0,035 **1 - p** = 0,965
O nível de confiança exigido é de 99%: para encontrar o valor crítico devemos procurar na tabela da distribuição normal padrão pela probabilidade 0,995 (0,99+0,005); os valores críticos serão $\mathbf{Z_{0,005}}$ e $\mathbf{Z_{0,995}}$ *os quais serão iguais em módulo. E o valor de* $\mathbf{Z_{crítico}}$ *será igual a 2,58 (em módulo). A precisão foi fixada em 1% (0,01). Substituindo os valores na expressão acima:*

$$n = \left(\frac{Z_{critico}}{e_0} \right)^2 \times p \times (1-p) = \left(\frac{2,58}{0,01} \right)^2 \times 0,035 \times 0,965 = 2248,14 \cong 2249$$

Observe que o tamanho mínimo de amostra necessário para atender a 99% de confiança e precisão de 1% deveria ser de 2249 elementos. Como a amostra coletada possui apenas 1000 elementos ela é INSUFICIENTE para a confiança e precisão exigidas. Recomenda-se o retorno à população para a retirada aleatória de mais 1249 peças.

Exemplo 9.6 – Lembrando do Exemplo 7.3 do Capítulo 7. Uma pesquisa de opinião busca conhecer melhor o comportamento da renda familiar e da opinião de eleitores de uma cidade em relação a um projeto de lei. Para os objetivos da pesquisa é apropriado dividir a população em estratos definidos por região, e as proporções de cada um em relação ao total de habitantes, de acordo com a tabela abaixo:

Região	Número de habitantes (Ni)	Wi (peso do estrato) = Ni/N
Nordeste	35000	35000/350000 = 0,1
Noroeste	70000	70000/350000 = 0,2
Sudoeste	140000	140000/350000 = 0,4
Sudeste	17500	17500/350000 = 0,05
Central	87500	87500/350000 = 0,25
Total (N)	3500000	-

Retirou-se uma amostra probabilística estratificada proporcional de 700 elementos, e foram registradas as rendas familiares e as opiniões dos eleitores sobre o projeto de lei, sendo calculadas as médias e variâncias[23] amostrais de renda (em salários mínimos) e as proporções amostrais de eleitores favoráveis ao projeto de lei, registradas na tabela a seguir:

Região	Número de elementos na amostra (n_i)	Média amostral de renda por estrato ($\bar{x}_i$) em salários mínimos	Variância amostral de renda por estrato (s^2_i) em (salários mínimos)^2	Proporção amostral de favoráveis por estrato (p_i)
Nordeste	70	3,35	1,56	0,20
Noroeste	140	2,76	1,32	0,25
Sudoeste	280	1,89	0,42	0,10
Sudeste	35	4,78	0,67	0,42
Central	175	3,05	1,04	0,52
Total (n)	700	-	-	-

[23] As variâncias não haviam sido apresentadas no Exemplo 7.3.

Obtenha os intervalos de confiança de 95% para a média populacional de renda e para a proporção populacional de favoráveis ao projeto de lei.

No Exemplo 7.3 foram obtidos os valores ponderados das medidas amostrais em cada estrato e depois os totais, conforme visto na tabela a seguir:

Região	W_i	n_i	$\bar{x}_i$	$W_i \times \bar{x}_i$ *(salários mínimos)*	p_i *(favoráveis)*	$W_i \times p_i$
Nordeste	*0,1*	*70*	*3,35*	*0,1×3,35 = 0,335*	*0,20*	*0,1×0,20 = 0,02*
Noroeste	*0,2*	*140*	*2,76*	*0,2×2,76 =0,952*	*0,25*	*0,2×0,25 = 0,05*
Sudoeste	*0,4*	*280*	*1,89*	*0,4×1,89 =0,756*	*0,10*	*0,4×0,10 = 0,04*
Sudeste	*0,05*	*35*	*4,78*	*0,05×4,78 =0,239*	*0,42*	*0,05×0,42 = 0,021*
Central	*0,25*	*175*	*3,05*	*0,25×3,05 =0,7625*	*0,52*	*0,25×0,52 = 0,13*
Total	*1,0*	*700*	-	*3,0445*	*1,0*	*0,261 (26,1%)*

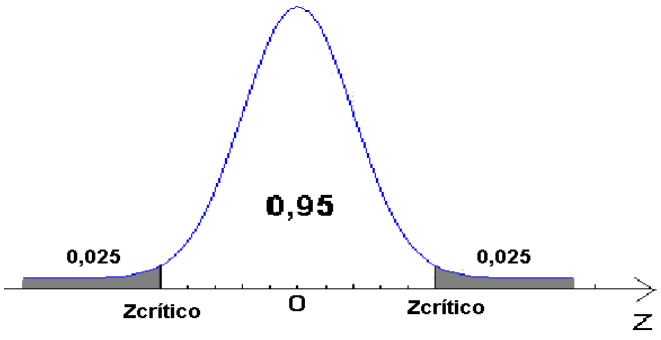
Então a média amostral (estratificada) de renda é de 3,0445 salários mínimos, e a proporção amostral (estratificada) de favoráveis ao projeto de lei é de 26,1%.

Agora é preciso estimar as variâncias da média e da proporção amostral (estratificadas):

Região	W_i	n_i	$s^2{}_i$	$\frac{W_i^2 \times s_i^2}{n_i} - \frac{W_i \times s_i^2}{N}$
Nordeste	*0,1*	*70*	*1,56*	$\frac{0,1^2 \times 1,56}{70} - \frac{0,1 \times 1,56}{350000}$
Noroeste	*0,2*	*140*	*1,32*	$\frac{0,2^2 \times 1,32}{140} - \frac{0,2 \times 1,32}{350000}$
Sudoeste	*0,4*	*280*	*0,42*	$\frac{0,4^2 \times 0,42}{280} - \frac{0,4 \times 0,42}{350000}$
Sudeste	*0,05*	*35*	*0,67*	$\frac{0,05^2 \times 0,67}{35} - \frac{0,05 \times 0,67}{350000}$
Central	*0,25*	*175*	*1,04*	$\frac{0,25^2 \times 1,04}{175} - \frac{0,25 \times 1,04}{350000}$
Total	*1,0*	*700*	-	$s^2{}_{\bar{x}\ estratificada} = \sum_{i=1}^{E}\left[\frac{W_i{}^2 \times s^2{}_i}{n_i} - \frac{W_i \times s^2{}_i}{N}\right]$ *= 0,00715*

Região	W_i	N_i	n_i	p_i	$W_i{}^2 \times \frac{(N_i - n_i)}{(N_i - 1)} \times \frac{p_i \times (1 - p_i)}{n_i - 1}$
Nordeste	*0,1*	*35000*	*70*	*0,20*	$0,1^2 \times \frac{(35000 - 70)}{(35000_i - 1)} \times \frac{0,20 \times (1 - 0,20)}{70 - 1}$
Noroeste	*0,2*	*70000*	*140*	*0,25*	$0,2^2 \times \frac{(70000 - 140)}{(70000 - 1)} \times \frac{0,25 \times (1 - 0,25)}{140 - 1}$
Sudoeste	*0,4*	*140000*	*280*	*0,10*	$0,4^2 \times \frac{(140000 - 280)}{(140000 - 1)} \times \frac{0,10 \times (1 - 0,10)}{280 - 1}$
Sudeste	*0,05*	*17500*	*35*	*0,42*	$0,05^2 \times \frac{(17500 - 35)}{(17500 - 1)} \times \frac{0,42 \times (1 - 0,42)}{35 - 1}$
Central	*0,25*	*87500*	*175*	*0,52*	$0,25^2 \times \frac{(87500 - 175)}{(87500 - 1)} \times \frac{0,52 \times (1 - 0,52)}{175 - 1}$
Total	*1,0*	*350000*	*700*	-	$s^2{}_{p\ estratificada} = \sum_{i=1}^{E}\left[W_i{}^2 \times \frac{(N_i - n_i)}{(N_i - 1)} \times \frac{p_i \times (1 - p_i)}{n_i - 1}\right]$ *=0,00015*

Como é possível considerar que a distribuição de $\bar{x}\ estratificada$ e $p\ estratificada$ podem ser aproximadas por uma normal, é possível encontrar o o valor de $\mathbf{Z_{crítico}}$ *: como o Intervalo de Confiança para a média é bilateral, teremos uma situação semelhante à da figura abaixo:*



Para encontrar o valor crítico devemos procurar na tabela da distribuição normal padrão pela probabilidade 0,975 (0,95+0,025) O valor da probabilidade pode ser visto na figura ao lado: os valores críticos serão $\mathbf{Z_{0,025}}$ e $\mathbf{Z_{0,975}}$ *os quais serão iguais em módulo. E o valor de* $\mathbf{Z_{crítico}}$ *será igual a 1,96 (em módulo).*

Para o caso da média estratificada:

$$e_0 = Z_{crítico} \times \sqrt{s^2{}_{\bar{x}\ estratificada}} = 1{,}96 \times \sqrt{0{,}00715} = 0{,}1657$$

$$L_I = \bar{x}_{estratificada} - \sqrt{s^2{}_{\bar{x}\ estratificada}} = 3{,}0445 - 0{,}1657 = 2{,}8788$$

$$L_S = \bar{x}_{estratificada} + \sqrt{s^2{}_{\bar{x}\ estratificada}} = 3{,}0445 + 0{,}1657 = 3{,}2102$$

Então o intervalo de 95% de confiança para a média populacional da renda é [2,8788;3,2102] salários mínimos[24].

Para o caso da proporção estratificada:

$$e_0 = Z_{crítico} \times \sqrt{s^2{}_{p\ estratificada}} = 1{,}96 \times \sqrt{0{,}00015} = 0{,}024$$

$$L_I = p_{estratificada} - \sqrt{s^2{}_{p\ estratificada}} = 0{,}261 - 0{,}024 = 0{,}237$$

$$L_S = p_{estratificada} + \sqrt{s^2{}_{p\ estratificada}} = 0{,}261 + 0{,}024 = 0{,}285$$

Então o intervalo de 95% de confiança para a proporção populacional de favoráveis é [23,7%;28,5%][25].

"EMPATE TÉCNICO"

Estamos acostumados a ouvir declarações do tipo "os candidatos A e B estão tecnicamente empatados na preferência eleitoral". O que significa isso? Geralmente as pesquisas de opinião eleitoral consistem em obter as proporções de entrevistados que declara votar neste ou naquele candidato, naquele momento. Posteriormente as proporções são generalizadas estatisticamente para a população, através do cálculo de intervalos de confiança para as proporções de cada candidato. Se os intervalos de confiança das proporções de dois ou mais candidatos apresentam grandes superposições declara-se que há um "empate técnico": as diferenças entre eles devem-se provavelmente ao acaso, e para todos os fins estão em condições virtualmente iguais, naquele momento.

[24] 95% de probabilidade de que a verdadeira média populacional de Renda esteja entre 2,8788 e 3,2102 salários mínimos.
[25] 95% de probabilidade de que a verdadeira proporção populacional de renda esteja entre 23,7% e 28,5%.

Exemplo 9.7 - Imagine que uma pesquisa de opinião eleitoral apresentasse os seguintes resultados (intervalos de confiança para a proporção que declara votar no candidato) sobre a prefeitura do município de Tapioca. Quais candidatos estão tecnicamente empatados?

Opinião	Limite inferior %	Limite superior %
Godofredo Astrogildo	31%	37%
Filismino Arquibaldo	14%	20%
Urraca Hermengarda	13%	19%
Salustiano Quintanilha	22%	28%
Indecisos	11%	17%

Filismino e Urraca estão tecnicamente empatados, pois seus intervalos de confiança apresentam grande sobreposição. Godofredo está muito na frente, pois o limite inferior de seu intervalo é maior do que o limite superior de Salustiano, que está em segundo lugar. É importante ressaltar que o número de indecisos é razoável, variando de 11 a 17%, quando eles se decidirem poderão mudar completamente o quadro da eleição, ou garantir a vitória folgada de Godofredo.