



# Módulo Básico – Tópicos de Probabilidade e Estatística

## PARTE 2

**Mônica Barros, D.Sc.**

Dezembro de 2006

## Quem sou eu?



### □ Mônica Barros

- Doutora em Séries Temporais – PUC-Rio
- Mestre em Estatística – University of Texas at Austin, EUA
- Bacharel em Matemática – University of Washington, Seattle, EUA
- Professora da PUC-Rio (Depto. De Eng. Elétrica)
- E-mails: [monica@ele.puc-rio.br](mailto:monica@ele.puc-rio.br), [monica@mbarros.com](mailto:monica@mbarros.com)
- Home page: <http://www.mbarros.com>



[monica@mbarros.com](mailto:monica@mbarros.com)

2

## Programa do Curso



- Estatística Descritiva (média, variância, desvio-padrão, covariância, correlação, gráficos)
- Cálculo de Probabilidade (axiomas, probabilidade condicional e teorema de Bayes)
- Variáveis Aleatórias (discretas e contínuas)
- Modelos Probabilísticos (densidades discretas e contínuas e função de distribuição)
- Introdução à Teoria de Estimação e Decisão

[monica@mbarros.com](mailto:monica@mbarros.com)

3

## Variáveis Aleatórias Contínuas



## Conteúdo



- Revisão - As principais distribuições contínuas
- Distribuição Uniforme
- Distribuição Exponencial
- Distribuição Normal
- Combinações Lineares de Variáveis Normais
- Distribuição Lognormal

## Revisão



- É importante lembrar o que caracteriza uma densidade de probabilidade. Seja  $X$  uma variável aleatória contínua com densidade de probabilidade  $f(x)$  e função de distribuição  $F(x)$ . Então:

- 1)  $f(x) \geq 0$  para todo  $x$

- 2)  $\int_{-\infty}^{\infty} f(x)dx = 1$

## Revisão



- A probabilidade de  $X$  estar num intervalo qualquer é:

$$\Pr(c < X < d) = \int_c^d f(x)dx$$

- A função de distribuição e a densidade estão relacionadas através de:

$$f(x) = \frac{dF(x)}{dx}$$

## Distribuição Uniforme



- Se  $X \sim \text{Unif}(a,b)$  então sua densidade é:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{se } x \in (a,b) \\ 0 & \text{se } x \notin (a,b) \end{cases}$$

- A função de distribuição é dada por:

$$F(x) = \Pr(X \leq x) = \begin{cases} 0 & \text{se } x \leq a \\ \frac{x-a}{b-a} & \text{se } x \in (a,b) \\ 1 & \text{se } x \geq b \end{cases}$$

- Note que a função de distribuição é **linear** no intervalo  $(a,b)$ .

## Distribuição Uniforme



- Média e Variância da distribuição Uniforme
- Se  $X \sim \text{Unif}(a,b)$  então:

$$E(X) = \frac{a+b}{2}, \text{VAR}(X) = \frac{(b-a)^2}{12}$$

## Distribuição Exponencial



- A distribuição exponencial é útil na descrição de situações como o tempo entre ocorrências consecutivas de um evento, tempo necessário para realização de uma tarefa etc.
  - Tempo entre chegadas consecutivas de carros em um posto de pedágio
  - Tempo de atendimento em um caixa de banco
- Também, a distribuição Exponencial é frequentemente usada para modelar a vida útil ou duração de componentes eletrônicos.

## Distribuição Exponencial



- Se  $X$  tem densidade Exponencial( $\lambda$ ) então:

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{se } x \geq 0 \text{ e } \lambda > 0 \\ 0 & \text{se } x < 0 \end{cases}$$

- A função de distribuição é:

$$F(x) = \Pr(X \leq x) = \begin{cases} 0 & \text{se } x < 0 \\ 1 - e^{-\lambda x} & \text{se } x \geq 0 \end{cases}$$

## Distribuição Exponencial



- Média e Variância

- Se  $X$  é Exponencial com parâmetro  $\lambda$ , então:

$$E(X) = 1/\lambda$$

$$\text{VAR}(X) = 1/\lambda^2$$

## Distribuição Normal



- A distribuição Normal é talvez a mais importante das distribuições de probabilidade.
- Muitos fenômenos físicos ou econômicos são frequentemente modelados pela distribuição Normal.
- É utilizada para descrever inúmeras aplicações práticas:
  - Altura e peso de pessoas e objetos
  - Nível de chuvas
  - Altura de árvores em uma floresta

## Distribuição Normal



- A distribuição Normal tem a forma de um sino, e possui **dois parâmetros,  $\mu$  e  $\sigma^2$** .
- A distribuição Normal é também chamada de **Gaussiana** em homenagem ao matemático Carl Friederich Gauss (1777 - 1855).
- A distribuição Normal também funciona como uma **boa aproximação para outras densidades**. Por exemplo, sob algumas condições pode-se provar que a densidade Binomial pode ser aproximada pela Normal.

## Distribuição Normal



- Densidade Normal com média  $\mu$  e variância  $\sigma^2$

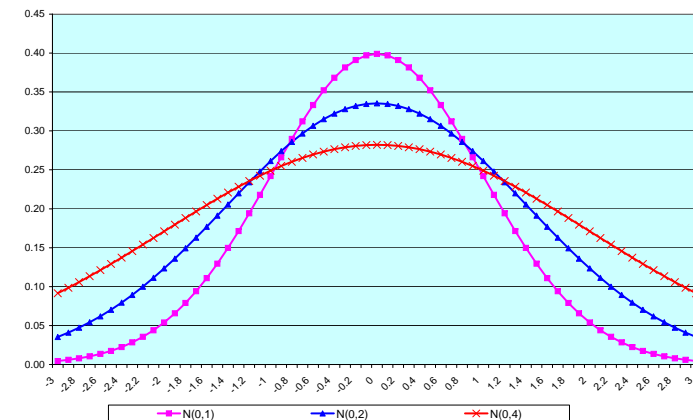
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ onde } \sigma^2 > 0 \text{ e } \mu \in \mathbb{R}$$

- Notação:  $X \sim N(\mu, \sigma^2)$
- A densidade é simétrica em torno de  $\mu$ , e quanto maior o valor da variância  $\sigma^2$ , mais "espalhada" é a distribuição.

## Distribuição Normal



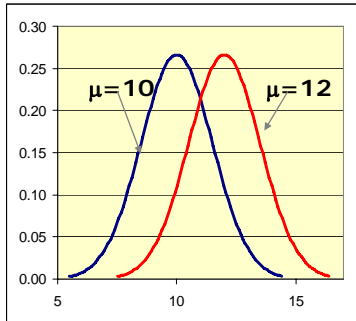
Densidades Normais com média zero e variâncias 1, 2 e 4



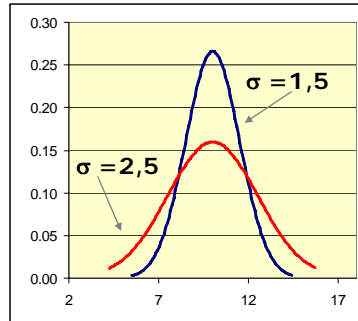
## Distribuição Normal



- A distribuição normal é completamente caracterizada por sua média  $\mu$  e seu desvio-padrão  $\sigma$
- A média define o deslocamento horizontal da curva, enquanto o desvio-padrão define o seu achatamento



monica@mbarros.com



17

## Distribuição Normal



### □ Propriedades

- 1)  $f(x)$  como definida integra a 1.
- 2)  $f(x) > 0$  sempre.
- 3) Os limites de  $f(x)$  quando  $x$  tende a  $+\infty$  e  $-\infty$  são iguais a zero.
- 4) A densidade  $N(\mu, \sigma^2)$  é *simétrica em torno de  $\mu$* , ou seja:  
$$f(\mu + x) = f(\mu - x)$$
- 5) O valor máximo de  $f(x)$  ocorre em  $x = \mu$
- 6) Os pontos de inflexão de  $f(x)$  são  $x = \mu + \sigma$  e  $x = \mu - \sigma$ .

monica@mbarros.com

18

## Distribuição Normal



### □ Média, Variância e função de distribuição

- Se  $X \sim N(\mu, \sigma^2)$  então:

$$E(X) = \mu,$$

$$VAR(X) = \sigma^2$$

- A sua função de distribuição é:

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(u-\mu)^2}{2\sigma^2}\right\} du$$

**Não é possível resolver analiticamente esta integral – precisamos de uma tabela!**

monica@mbarros.com

19

## Distribuição Normal



### □ Tabela: será feita para a distribuição $N(0,1)$

- É possível transformar uma variável  $N(\mu, \sigma^2)$  numa  $N(0,1)$  sem grandes dificuldades, e então podemos **tabelar os valores da função de distribuição de uma  $N(0,1)$** , e esta tabela pode ser usada para encontrar probabilidades envolvendo qualquer variável aleatória Normal.

monica@mbarros.com

20

## Distribuição Normal



### ❑ Problema:

Não é possível criar uma tabela para cada uma das (infinitas) densidades Normais existentes.

### ❑ Solução:

Trabalha-se com a densidade Normal com média 0 e variância 1, e converte-se todas as outras Normais para esta, chamada de **Normal padrão** ou **Normal standard**.

A maioria dos livros de estatística fornece tabelas de probabilidade para a distribuição normal padronizada.

## Distribuição Normal



### ❑ Transformação numa $N(0,1)$

❑ Se  $X \sim N(\mu, \sigma^2)$  então  $Z = (X - \mu)/\sigma$  é uma variável Normal com média 0 e variância 1.

❑ Logo, para transformar uma variável aleatória Normal com quaisquer parâmetros numa Normal (0,1) você deve:

- 1- Subtrair a média
- 2- Dividir o resultado por  $\sigma$ , o desvio padrão

A variável aleatória resultante deste procedimento é uma  $N(0,1)$ .

## Distribuição Normal



❑ Se  $X$  pertence a uma distribuição normal com média  $\mu$  e desvio-padrão  $\sigma$ , seu valor normalizado é dado por:

$$Z = \frac{X - \mu}{\sigma}$$

❑ Existem dois tipos de tabela, que fornecem basicamente a mesma coisa:

- ❑  $\Pr(0 \leq Z \leq z_0)$ , ou seja, a probabilidade do lado direito da curva normal a partir da média até o valor  $z_0$
- ❑  $\Phi(z_0) = \Pr(Z \leq z_0) = 0.5 + \Pr(0 \leq Z \leq z_0)$  (por que?)

❑ Iremos trabalhar com a tabela da função de distribuição, isto é:  $\Phi(z_0)$

## Distribuição Normal



❑ Toda variável Normal pode ser transformada numa Normal com média 0 e variância 1.

❑ Logo, só existe a necessidade de criar uma única tabela para a função de distribuição acumulada.

❑ Se  $X$  é  $N(\mu, \sigma^2)$ . Então a variável  $Z = (X - \mu) / \sigma$  tem distribuição Normal com média zero e variância um, isto é,  $Z$  é  $N(0,1)$ .

## Distribuição Normal



### □ Cálculo de probabilidades

Se  $X$  é uma variável Normal com média  $\mu$  e desvio padrão  $\sigma$  então:

$$\Pr(a \leq X \leq b) = \Pr\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = \Pr\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

□ onde  $\Phi$  é a função de distribuição da  $N(0,1)$ , que é tabelada. Alguns valores importantes são:

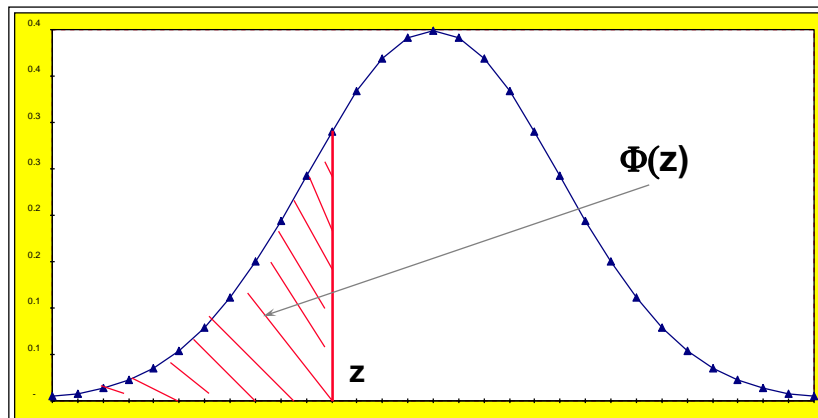
$$\Phi(1.645) = 0.95, \Phi(1.96) = 0.975 \text{ e } \Phi(2.326) = 0.99$$

## Distribuição Normal



- O **Excel** fornece diretamente o valor de  $\Phi(z_0)$  através da função **DIST.NORMP**.
- O **único argumento** para esta função é o valor  $z_0$  para o qual você quer calcular a probabilidade de estar abaixo, pois a função pressupõe que a distribuição usada é a Normal padrão (média 0 e variância 1).

## Tabela da $N(0,1)$ usando $\Phi(z_0)$



## Tabela da $N(0,1)$



- **Simetrias**
- $\Phi(-z) = 1 - \Phi(z)$  se  $z > 0$
- **ISSO É IMPORTANTE POIS A TABELA SÓ CONTÉM VALORES DE z POSITIVOS!**
- Probabilidade de um intervalo simétrico em torno de zero
- $\Pr(-t < Z < t) = 1 - 2\{\Phi(-t)\} = 1 - 2\{1 - \Phi(t)\} = 2 \cdot \Phi(t) - 1$  onde  $Z \sim N(0,1)$

## Tabela da $N(0,1)$ ( $\Phi(z_0) = \Pr(Z \leq z_0)$ )



z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.00	0.50000	0.62	0.7324	1.24	0.8925	1.86	0.9686
0.02	0.50800	0.64	0.7389	1.26	0.8962	1.88	0.9699
0.04	0.51600	0.66	0.7454	1.28	0.8997	1.90	0.9713
0.06	0.52399	0.68	0.7517	1.30	0.9032	1.92	0.9726
0.08	0.53199	0.70	0.7580	1.32	0.9066	1.94	0.9738
0.10	0.53998	0.72	0.7642	1.34	0.9099	1.96	0.9750
0.12	0.54798	0.74	0.7704	1.36	0.9131	1.98	0.9761
0.14	0.55597	0.76	0.7764	1.38	0.9162	2.00	0.9772
0.16	0.56396	0.78	0.7823	1.40	0.9192	2.02	0.9783
0.18	0.57194	0.80	0.7881	1.42	0.9222	2.04	0.9793
0.20	0.57993	0.82	0.7939	1.44	0.9251	2.06	0.9803
0.22	0.58791	0.84	0.7995	1.46	0.9279	2.08	0.9812
0.24	0.59488	0.86	0.8051	1.48	0.9306	2.10	0.9821
0.26	0.60286	0.88	0.8106	1.50	0.9332	2.12	0.9830
0.28	0.61083	0.90	0.8159	1.52	0.9357	2.14	0.9838
0.30	0.61779	0.92	0.8212	1.54	0.9382	2.16	0.9846
0.32	0.62575	0.94	0.8264	1.56	0.9406	2.18	0.9854
0.34	0.63371	0.96	0.8315	1.58	0.9429	2.20	0.9861
0.36	0.64166	0.98	0.8365	1.60	0.9452	2.22	0.9868
0.38	0.64960	1.00	0.8413	1.62	0.9474	2.24	0.9875
0.40	0.65754	1.02	0.8461	1.64	0.9495	2.26	0.9881
0.42	0.66548	1.04	0.8508	1.66	0.9515	2.28	0.9887
0.44	0.67341	1.06	0.8554	1.68	0.9535	2.30	0.9893
0.46	0.67722	1.08	0.8599	1.70	0.9554	2.32	0.9898
0.48	0.68444	1.10	0.8643	1.72	0.9573	2.34	0.9904
0.50	0.69165	1.12	0.8686	1.74	0.9591	2.36	0.9909
0.52	0.69885	1.14	0.8729	1.76	0.9608	2.38	0.9913
0.54	0.70544	1.16	0.8770	1.78	0.9625	2.40	0.9918
0.56	0.71223	1.18	0.8810	1.80	0.9641	2.42	0.9922
0.58	0.71900	1.20	0.8849	1.82	0.9656	2.44	0.9927
0.60	0.72577	1.22	0.8888	1.84	0.9671	2.46	0.9931

## Tabela da $N(0,1)$



- ❑ Dicas
- ❑ Você precisa explorar as simetrias da  $N(0,1)$  pois a tabela só é dada para valores positivos de  $z_0$ . Por causa da simetria em torno de zero,  $\Phi(0) = 0.5$  e  $\Phi(z_0)$  é menor que 0.5 se  $z_0$  for um número negativo.
- ❑ Se você tiver dúvidas, faça um desenho!
- ❑ Lembre-se sempre que  $\Phi(z_0)$  é uma função de distribuição, ou seja, mede a probabilidade de estarmos **ABAIXO** do ponto  $z_0$ .

## Tabela da $N(0,1)$



- ❑ **Dica - uso da calculadoras da série HP48**
- ❑ A função **UTPN** fornece  $1 - \Phi(z_0)$  (na verdade, esta função é até mais geral, pois pode ser usada com uma média e variância qualquer; consulte o manual da sua calculadora).
- ❑ O algoritmo que irei mostrar é válido apenas para a  $N(0,1)$ .
  - ❑ Menu MTH
  - ❑ Submenu PROB
  - ❑ Função UTPN
  - ❑ Argumentos 0, 1,  $z_0$
  - ❑ Retorna a probabilidade de uma variável  $N(0,1)$  estar **ACIMA** do ponto  $z$ , ou seja,  $1 - \Phi(z_0)$

## Distribuição Normal



- ❑ Seja  $X \sim N(\mu, \sigma^2)$  e  $k > 0$ . Mostre que  $\Pr\{\mu - k\sigma < X < \mu + k\sigma\}$  só depende de  $k$  (não depende de  $\mu$  e  $\sigma$ ).
- ❑ Solução
- ❑ Note que a probabilidade desejada é a probabilidade de  $X$  estar a uma distância menor ou igual a  $k$  desvios padrões da sua média.

## Distribuição Normal



$$\begin{aligned}\Pr(\mu - k\sigma < X < \mu + k\sigma) &= \Pr(-k\sigma < X - \mu < +k\sigma) = \\ &= \Pr\left(-\frac{k\sigma}{\sigma} < \frac{X - \mu}{\sigma} < +\frac{k\sigma}{\sigma}\right) = \Pr\left(-k < \frac{X - \mu}{\sigma} < k\right) = \Pr(-k < Z < +k) = \\ &= 2 \cdot \Phi(k) - 1\end{aligned}$$

- **As probabilidades para alguns valores k estão abaixo:**

$$\Pr(\mu - \sigma < X < \mu + \sigma) = 2 \cdot \Phi(1) - 1 = 0.6826$$

$$\Pr(\mu - 1.645\sigma < X < \mu + 1.645\sigma) = 2 \cdot \Phi(1.645) - 1 = 0.90$$

$$\Pr(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 2 \cdot \Phi(1.96) - 1 = 0.95$$

$$\Pr(\mu - 2.57\sigma < X < \mu + 2.57\sigma) = 2 \cdot \Phi(2.57) - 1 = 0.99$$

## Distribuição Normal



- Na verdade, aquela **“regra de bolso”** que diz que **68%** dos valores estão a uma distância de 1 d.p. da média e **95%** dos valores estão a dois desvios da média acabou de ser mostrada no slide anterior.
- **Mas note que isso só é realmente verdade para a distribuição Normal!**

## Distribuição Normal



- **Exemplo**

Numa agência bancária localizada numa grande cidade brasileira, verificou-se que os clientes pessoa física mantêm, em média, um volume de R\$ 4800,00 aplicados no banco.

A dispersão entre os volumes de recursos, medida pelo desvio padrão, é R\$ 1600,00. Além disso, pode-se encarar os saldos dos correntistas como independentes entre si e Normalmente distribuídos.

## Distribuição Normal



- O banco pretende abrir uma nova agência e seus executivos imaginam que o poder aquisitivo nesta nova área é semelhante ao dos clientes desta agência.
- a) Um cliente é VIP se está entre os 5% com maior volume de recursos. Quanto uma pessoa deveria manter no banco para ser considerada cliente VIP?

## Distribuição Normal



- b) O banco pretende cobrar tarifas mais altas dos clientes que têm um baixo volume de recursos aplicados na instituição.

Os clientes cujos volumes de recursos estão entre os 10% mais baixos terão de pagar esta tarifa mais alta. Abaixo de qual volume um cliente será alvo desta tarifa diferenciada?

## Distribuição Normal



### □ Solução

Seja  $X$  a variável que mede o volume de recursos de um cliente típico da agência. Então  $X$  é Normal  $(4800, (1600)^2)$ . Daí:  $Z = \frac{X - 4800}{1600}$

tem densidade Normal padrão.

Para estar entre os 5% mais “ricos”, precisamos encontrar  $z_0$  tal que  $\Phi(z_0) = 95\%$ . Usando a função INV.NORMP do Excel, encontramos  $z_0 = 1.645$ .

Logo,  $\frac{X - 4800}{1600} = 1.645 \Rightarrow X = 4800 + 1.645(1600) = 7432$

## Distribuição Normal



### □ Solução (continuação)

b) Para estar entre os 10% mais “pobres” precisamos encontrar  $z_0$  tal que  $\Phi(z_0) = 10\%$ . A função INV.NORMP do Excel fornece  $z_0 = -1.281$ . Logo,

$$\frac{X - 4800}{1600} = -1.281 \Rightarrow X = 4800 - 1.281(1600) = 2750.40$$

- Ou seja, clientes com volume de recursos abaixo de R\$ 2750 estarão sujeitos a uma tarifa mais alta, e aqueles com volume de aplicações acima de R\$ 7432 terão tratamento VIP.

## Distribuição Normal



### □ Exemplo

- O saldo devedor dos usuários de um certo cartão de crédito é uma variável aleatória Normal com média R\$ 200 e desvio padrão R\$ 75.

- a) Qual a probabilidade do saldo devedor de um usuário estar entre R\$ 100 e R\$ 300?
- b) Qual deve ser o seu saldo devedor para que você esteja entre os 5% mais endividados?

### □ Solução

$X$  é Normal com média 200 e desvio padrão 75 e assim  $Z = (X - 200)/75$  é  $N(0,1)$ .

## Distribuição Normal



### □ Solução (continuação)

$$\Pr(100 < X < 300) =$$

$$\Pr\left(\frac{100-200}{75} < Z < \frac{300-200}{75}\right) = \Pr(-1.333 < Z < +1.333) = \\ = \Phi(1.333) - \Phi(-1.333) = 2\Phi(1.333) - 1 = 0.8176$$

b) Para que você esteja entre os 5% mais endividados, o saldo devedor padronizado deve ser igual a 1.645 (veja tabela da Normal). Daí:

$$Z = \frac{X-200}{75} = 1.645 \Rightarrow X = 200 + 1.645(75) = 323.38$$

é o saldo para estar entre os 5% com maior saldo devedor.

## Distribuição Normal



### □ Exemplo (para casa)

□ O consumo médio residencial de energia elétrica nos meses de verão numa certa cidade é uma variável Normal com média 210 kWh e desvio padrão 18 kWh.

a) Qual a probabilidade de que o consumo no verão exceda 225 kWh?

b) Calcule a probabilidade de que o consumo no verão seja inferior a 190 kWh.

c) Quanto você deve consumir para estar entre os 2.5% que mais gastam energia?

## Distribuição Normal



### □ Exemplo (para casa)

□ Numa certa empresa de informática, o salário *anual* médio dos funcionários com menos de 5 anos de experiência é R\$ 24000, com desvio padrão de R\$ 3000. Suponha que os salários têm distribuição Normal e calcule os valores pedidos a seguir.

## Distribuição Normal



□ a) Qual a probabilidade do salário anual de um funcionário qualquer com menos de 5 anos de experiência ser menor que R\$ 20000?

□ b) Qual deve ser o valor do salário anual de um funcionário com menos de 5 anos de experiência se 95% dos funcionários (com menos de 5 anos de experiência) tem salário abaixo dele?

## Distribuição Normal



- c) Toma-se uma amostra de 36 funcionários com menos de 5 anos de experiência. Qual a probabilidade do salário médio na amostra exceder R\$ 24500?
- d) Toma-se uma amostra de 12 funcionários com menos de 5 anos de experiência. Qual a probabilidade do maior salário na amostra exceder R\$ 28000?

## Combinações Lineares de Variáveis Normais



- Sejam  $X_1, X_2, \dots, X_n$  variáveis aleatórias independentes, onde  $X_i \sim N(\mu_i, \sigma_i^2)$  e seja  $Y = X_1 + X_2 + \dots + X_n$ .
- Então  $Y$  tem distribuição Normal com média  $\mu_y$  e variância  $\sigma_y^2$  dadas por:

$$\mu_y = \sum_{i=1}^n \mu_i$$
$$\sigma_y^2 = \sum_{i=1}^n \sigma_i^2$$

## Combinações Lineares de Variáveis Normais



- Um **caso particular** importante é: se os  $X_i$ 's forem iid  $N(\mu, \sigma^2)$ , então sua **soma** é Normal com média  $n \cdot \mu$  e variância  $n \cdot \sigma^2$  e a **média amostral** é Normal com média  $\mu$  e variância  $\sigma^2/n$ .

## Distribuição Normal



- Exemplo (continuação)
- Considere o exemplo dos saldos em aplicações bancárias. Suponha que tomamos uma amostra de 16 clientes da agência.
- Qual a probabilidade de que o saldo médio das aplicações dos clientes na amostra exceda R\$ 4900?

Seja  $\bar{X}$  a média dos saldos dos clientes na amostra.

$$\bar{X} \text{ tem distribuição } N\left(4800, \frac{(1600)^2}{16}\right)$$

## Distribuição Normal



□ Então:

$$\begin{aligned}\Pr(\bar{X} > 4900) &= \Pr\left(\frac{\bar{X} - 4800}{\frac{1600}{4}} > \frac{4900 - 4800}{\frac{1600}{4}}\right) = \Pr\left(Z > \frac{100}{400}\right) = \\ &= \Pr(Z > 0.25) = 1 - \Phi(0.25) = 1 - 0.599 = 0.401\end{aligned}$$

## Distribuição Normal (para casa)



- Um estudante universitário gasta em média R\$ 600,00 em livros por ano. A dispersão entre os valores gastos, medida pelo desvio padrão, é R\$ 240,00.
- Além disso, pode-se encarar os valores gastos pelos universitários como independentes entre si e Normalmente distribuídos. Também, a maioria dos estudantes adquire livros pela Internet.

## Distribuição Normal (para casa)



- a) Uma grande livraria na Internet pretende oferecer um cartão VIP aos clientes que mais compram livros. Apenas os 1% que mais consomem livros num período de um ano receberão o cartão. Acima de qual volume anual de compras um consumidor se candidata ao cartão VIP?
- b) Considere 16 estudantes universitários. Qual a probabilidade do gasto médio anual em livros destas 16 pessoas ultrapassar R\$ 660,00?
- c) Dentre as 16 pessoas nesta mesma amostra, qual a probabilidade do estudante que menos consumiu livros ter gasto mais de R\$ 650 no ano?

## Distribuição Normal (para casa)



- Um apartamento de 2 quartos numa certa região da cidade custa, em média R\$ 260 mil. A dispersão entre os valores, medida pelo desvio padrão, é R\$ 100 mil.
- Além disso, pode-se encarar os preços dos apartamentos como independentes entre si e Normalmente distribuídos.

## Distribuição Normal (para casa)



- ❑ a) Uma imobiliária pretende oferecer uma viagem de presente aos compradores de apartamentos de 2 quartos neste bairro que comprem os apartamentos situados na faixa dos 10% mais caros. A partir de quanto deve custar o seu apartamento para que você ganhe a viagem de “presente”?
- ❑ b) Considere 16 compradores de apartamentos de 2 quartos neste bairro. Qual a probabilidade do preço médio pago por eles ser inferior a R\$ 300 mil?
- ❑ c) Dentre as 16 pessoas nesta mesma amostra, qual a probabilidade do comprador que pagou mais caro por um apartamento ter pago menos de R\$ 285 mil?

monica@mbarros.com

53

## A distribuição Lognormal



- ❑ A distribuição Lognormal é uma distribuição de probabilidade contínua usada para dados positivos.
- ❑ Esta distribuição é freqüentemente usada na modelagem do preço de ações e outros ativos financeiros, e também pode modelar o tempo até a ocorrência de um defeito de uma máquina.

monica@mbarros.com

54

## A distribuição Lognormal



- ❑ **Veja o link:**  
<http://www.inf.ethz.ch/personal/gut/lognormal/> para um simulador interessante de variáveis lognormais e normais.
- ❑ Se você se interessar, o artigo do link:
- ❑ <http://stat.ethz.ch/~stahel/lognormal/bioscience.pdf>
- ❑ discute o uso da lognormal nas ciências.

monica@mbarros.com

55

## A Distribuição Lognormal



- ❑ Como criar uma variável lognormal?
- ❑ Seja  $X \sim N(\mu, \sigma^2)$ . Seja  $Y = \exp(X)$ . Então Y tem densidade Lognormal com parâmetros  $\mu$  e  $\sigma^2$ .
- ❑ A densidade de Y pode ser facilmente encontrada pelos métodos usuais (por exemplo, o método do Jacobiano), e é dada por:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \left(\frac{1}{y}\right) \cdot \exp\left(-\frac{(\log(y) - \mu)^2}{2\sigma^2}\right) \quad \text{onde } y > 0$$

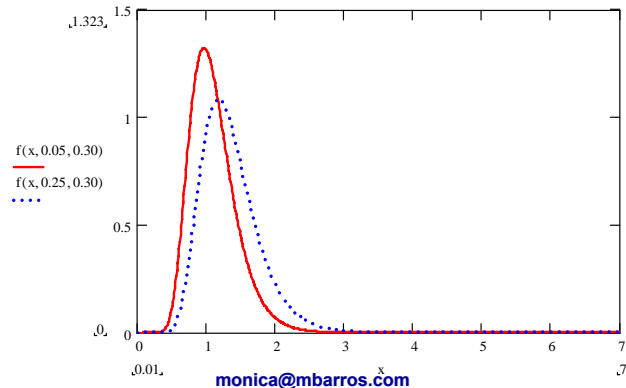
monica@mbarros.com

56

## A Distribuição Lognormal



- Exemplo – Lognormais com  $\mu = 0.05$  e  $0.25$  e  $\sigma = 0.30$



57

## A Distribuição Lognormal



- Atenção:
- A distribuição Lognormal, ao contrário do que o nome indica, não significa a densidade do logaritmo de uma variável Normal, pois uma variável Normal admite valores negativos, onde o logaritmo não está definido. Uma variável aleatória com densidade **Lognormal** é encontrada tomando-se a **exponencial** de uma variável aleatória **Normal**!

monica@mbarros.com

58

## Lognormal como modelo para o preço de uma ação



- Uma forma de descrever a incerteza sobre o preço de uma ação é supor que as variações no preço entre os instantes  $t$  e  $t+\Delta t$  podem ser divididas em 2 componentes, uma aleatória e a outra determinística, como a seguir:

$$S_{t+\Delta t} = S_t \cdot \left\{ \exp\left(\mu \cdot \Delta t + \sigma \cdot Z \cdot \sqrt{\Delta t}\right) \right\}$$

- onde  $Z$  é uma variável  $N(0,1)$  e  $\mu$  e  $\sigma > 0$  são parâmetros conhecidos. O parâmetro  $\mu$  representa a taxa média de crescimento do preço ao longo do tempo.

monica@mbarros.com

59

## Lognormal como modelo para o preço de uma ação



- Note que, se  $\sigma = 0$ , a evolução dos preços é **puramente determinística**, e então:
$$S_{t+\Delta t} = S_t \cdot \left\{ \exp(\mu \cdot \Delta t) \right\}$$
- Nesta expressão percebemos que a tendência determinística dos preços é crescente desde que  $\mu > 0$ .
- Se  $\sigma > 0$  então existe uma **componente aleatória** no comportamento dos preços. Esta componente aleatória é dada por uma variável aleatória  $N(0,1)$ , e assim o efeito desta variável pode ser o de atenuar o crescimento determinístico no preço, pois  $Z$  pode ser negativo. **Note que a variável  $\exp(Z)$  é Lognormal.**

monica@mbarros.com

60

## Propriedades da distribuição Lognormal



### □ Teorema (média e variância da Lognormal)

□ Se  $Y \sim \text{Lognormal}(\mu, \sigma^2)$  então:

□  $E(Y) = \exp(\mu + \sigma^2/2)$

$$\text{VAR}(Y) = \exp(2\mu + \sigma^2) \cdot (e^{\sigma^2} - 1)$$

□ Demonstração – faça em casa – use a fgm de uma Normal.

## Distribuições derivadas da Normal



## Distribuições Derivadas da Normal



### □ Densidade Qui-quadrado com k graus de liberdade

Seja X uma variável aleatória contínua e positiva com densidade dada por:

$$f(x) = \frac{1}{2^{k/2} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot x^{\frac{k}{2}-1} \cdot e^{-x/2} \quad \text{onde } x > 0$$

□ Então X tem densidade Qui-quadrado com k graus de liberdade, e escrevemos:  $X \sim \chi^2_k$

## Distribuições Derivadas da Normal



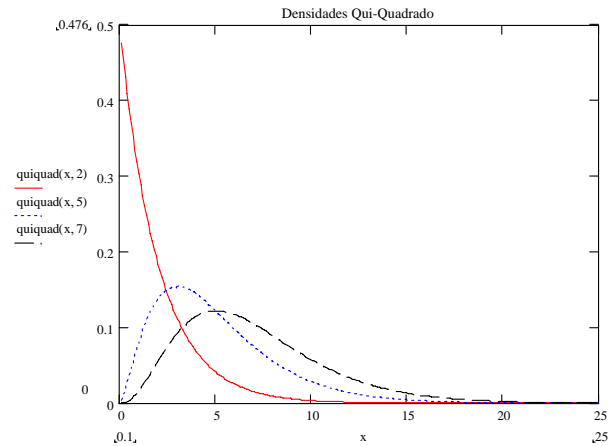
□ A densidade Qui-quadrado é apenas um caso particular de uma outra densidade chamada densidade Gama, que também inclui a Exponencial como caso particular.

□ Se X é Qui-quadrado com n graus de liberdade então:

$$E(X) = \frac{n/2}{1/2} = n$$

$$\text{VAR}(X) = \frac{n/2}{(1/2)^2} = 2n$$

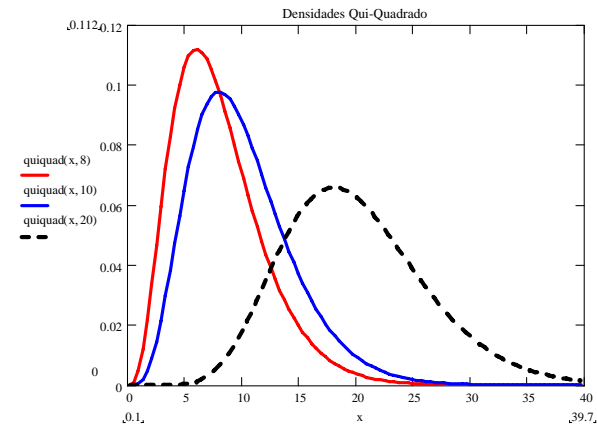
## Distribuição Qui-Quadrado



monica@mbarros.com

65

## Distribuição Qui-Quadrado



monica@mbarros.com

66

## Distribuição Qui-quadrado



### □ Tabelas da função de distribuição Qui-quadrado

- A densidade Qui-quadrado é tabelada para diversos graus de liberdade.
- As tabelas geralmente fornecem o valor  $x_{1-\alpha}$  tal que  $\Pr(X < x_{1-\alpha}) = 1 - \alpha$  para  $\alpha = 1\%$ ,  $5\%$ ,  $10\%$ . Também existem tabelas que apresentam o valor  $x_\alpha$  tais que  $\Pr(X < x_\alpha) = \alpha$ , isto é,  $\Pr(X > x_\alpha) = 1 - \alpha$ .

monica@mbarros.com

67

## Distribuição Qui-quadrado



probabilidade →	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
graus de liberdade ↓									
1	0.000	0.004	0.016	0.102	0.455	1.323	2.706	3.841	6.635
2	0.020	0.103	0.211	0.575	1.386	2.773	4.605	5.991	9.210
3	0.115	0.352	0.584	1.213	2.366	4.108	6.251	7.815	11.345
4	0.297	0.711	1.064	1.923	3.357	5.385	7.779	9.488	13.277
5	0.554	1.145	1.610	2.675	4.351	6.626	9.236	11.070	15.086
6	0.872	1.635	2.204	3.455	5.348	7.841	10.645	12.592	16.812
7	1.239	2.167	2.833	4.255	6.346	9.037	12.017	14.067	18.475
8	1.647	2.733	3.490	5.071	7.344	10.219	13.362	15.507	20.090
9	2.088	3.325	4.168	5.899	8.343	11.389	14.684	16.919	21.666
10	2.558	3.940	4.865	6.737	9.342	12.549	15.987	18.307	23.209
11	3.053	4.575	5.578	7.584	10.341	13.701	17.275	19.675	24.725
12	3.571	5.226	6.304	8.438	11.340	14.845	18.549	21.026	26.217
13	4.107	5.892	7.041	9.299	12.340	15.984	19.812	22.362	27.688
14	4.660	6.571	7.790	10.165	13.339	17.117	21.064	23.685	29.141
15	5.229	7.261	8.547	11.037	14.339	18.245	22.307	24.996	30.578
16	5.812	7.962	9.312	11.912	15.338	19.369	23.542	26.296	32.000
17	6.408	8.672	10.085	12.792	16.338	20.489	24.769	27.587	33.409
18	7.015	9.390	10.865	13.675	17.338	21.605	25.989	28.869	34.805
19	7.633	10.117	11.651	14.562	18.338	22.718	27.204	30.144	36.191
20	8.260	10.851	12.443	15.452	19.337	23.828	28.412	31.410	37.566

monica@mbarros.com

68

## Distribuição Qui-quadrado



- ❑ Função de Distribuição Qui-quadrado no Excel
- ❑ Use as funções **DIST.QUI** e **INV.QUI**
- ❑ **A tabela anterior foi produzida usando INV.QUI** – dada uma probabilidade e o grau de liberdade, a função INV.QUI retorna o ponto correspondente da densidade tal que a probabilidade de estar **ACIMA** do ponto é a especificada como argumento da função.

monica@mbarros.com

69

## Distribuição Qui-quadrado



- ❑ Função de Distribuição Qui-quadrado no Excel
- ❑ Por exemplo, para uma Qui-quadrado com 10 graus de liberdade:
- ❑  $INV.QUI(0.99, 10) = 2.558$
- ❑  $INV.QUI(0.01, 10) = 23.209$
- ❑ Ou seja, a probabilidade de uma v.a. Qui-quadrado com 10 graus de liberdade exceder 2.558 é 0.99, e a probabilidade da mesma variável exceder 23.209 é 0.01.

monica@mbarros.com

70

## Distribuição Qui-Quadrado



Valor de x para o qual desejamos

$Pr(X > x)$

Graus de liberdade da Qui-quadrado

Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

CHIDIST   =CHIDIST()

1 A B C D E F G H

2 CHIDIST

3 x = number

4 Deg\_freedom = number

5 =

6 Returns the one-tailed probability of the chi-squared distribution.

7 X is the value at which you want to evaluate the distribution, a nonnegative number.

8 Formula result =

9 OK Cancel

monica@mbarros.com

71

## Distribuição Qui-Quadrado



Da figura segue que, a  $Pr(X > 15)$  quando X é uma Qui-quadrado com 12 graus de liberdade é 0.2414

Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

CHIDIST   =CHIDIST(15;12)

1 A B C D E F

2 CHIDIST

3 x = 15

4 Deg\_freedom = 12

5 =

6 Returns the one-tailed probability of the chi-squared distribution.

7 X is the value at which you want to evaluate the distribution, a nonnegative number.

8 Formula result = 0.241436451

9 OK Cancel

monica@mbarros.com

72

## Distribuição Qui-Quadrado



- Por exemplo:
- Supondo que  $X$  seja uma variável aleatória com densidade qui-quadrado com 6 graus de liberdade, a probabilidade de  $X$  exceder 0.87 é 99%.
- Analogamente, a probabilidade de  $X$  exceder 12.59 é 5% e a probabilidade de  $X$  estar acima de 16.81 é apenas 1%.

## Distribuição Qui-Quadrado



- Podemos estar interessados na pergunta “ao contrário”. Dada uma Qui-Quadrado com  $k$  graus de liberdade e uma probabilidade  $\alpha$ , qual é o ponto tal que a probabilidade de estar ACIMA dele é  $\alpha$ ?
- O Excel também nos dá esta resposta, através da função INV.CHI.

## Distribuição Qui-Quadrado



Da figura segue que, a  $\Pr(X > 31.4104)$  quando  $X$  é uma Qui-quadrado com 20 graus de liberdade é 0.05

## Distribuição Qui-quadrado



- A densidade Qui-quadrado é importante no contexto de amostras aleatórias Normais, na estimação da variância.
- Também pode-se provar que o quadrado de uma variável Normal padrão (que estudaremos a seguir) tem densidade Qui-quadrado com um grau de liberdade.

## Distribuições Derivadas da Normal



- Uma propriedade muito importante da densidade Qui-quadrado é a **preservação da mesma família** de densidades **quando somamos** variáveis independentes.
- Ou seja, se  $X_1, X_2, \dots, X_n$  são variáveis independentes, cada uma com distribuição Qui-quadrado, a soma de  $X_1, X_2, \dots, X_n$  também é uma variável aleatória qui-quadrado.

## Distribuições Derivadas da Normal



- **Teorema (aditividade da densidade Qui-quadrado)**
- Sejam  $X_1, X_2, \dots, X_n$  v.a. aleatórias independentes, e suponha que  $X_i$  tem densidade qui-quadrado com  $k_i$  graus de liberdade. Seja  $Y = X_1 + X_2 + \dots + X_n$ .
- Então  $Y$  tem também uma densidade Qui-quadrado, mas com  $k = k_1 + k_2 + \dots + k_n$  graus de liberdade.
- O próximo teorema exhibe a relação existente entre as densidades  $N(0,1)$  e Qui-quadrado.

## Distribuições Derivadas da Normal



- **Teorema**
- Seja  $Z \sim N(0,1)$ . Então  $V = Z^2$  tem densidade Qui-quadrado com 1 grau de liberdade.
- **A combinação dos dois últimos teoremas leva a um resultado importante.**
- Sejam  $Z_1, Z_2, \dots, Z_n$  v.a. independentes e identicamente distribuídas com densidade  $N(0,1)$ . Então:

## Distribuições Derivadas da Normal



$$V = \sum_{i=1}^n Z_i^2 = Z_1^2 + Z_2^2 + \dots + Z_n^2$$

- tem densidade Qui-quadrado com  $n$  graus de liberdade.
- Este resultado segue trivialmente dos dois últimos, se lembrarmos que cada  $Z_i^2$  tem densidade qui-quadrado com 1 grau de liberdade (e são todos independentes).

## Distribuições Derivadas da Normal



- ❑ **Por que a densidade Qui-quadrado é importante?**
- ❑ Porque está relacionada com a distribuição da variância amostral de uma amostra aleatória Normal, como indicado no próximo teorema.
- ❑ Por exemplo, se desejarmos encontrar um intervalo baseado na variância amostral que contenha, com alta probabilidade, a variância (desconhecida) da distribuição Normal, este intervalo será construído a partir da distribuição Qui-quadrado.

## Distribuições Derivadas da Normal



- ❑ Teorema
- ❑ Sejam  $X_1, X_2, \dots, X_n$  uma amostra aleatória da distribuição  $N(\mu, \sigma^2)$ . Seja  $S^2$  a variância amostral:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- ❑ Então:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$$

- ❑ **tem distribuição Qui-quadrado com (n-1) graus de liberdade.**

## Distribuições Derivadas da Normal



- ❑ A partir deste teorema podemos deduzir facilmente a média e variância de  $S^2$ .
- ❑ Teorema
- ❑ Sejam  $X_1, X_2, \dots, X_n$  uma amostra aleatória da distribuição  $N(\mu, \sigma^2)$ . Seja  $S^2$  a variância amostral. Então:

$$E(S^2) = \sigma^2$$

$$VAR(S^2) = \frac{2\sigma^4}{n-1}$$

## Distribuições Derivadas da Normal



- ❑ A distribuição t de Student
- ❑ Tem apenas um parâmetro k, o número de graus de liberdade, e é definida como:

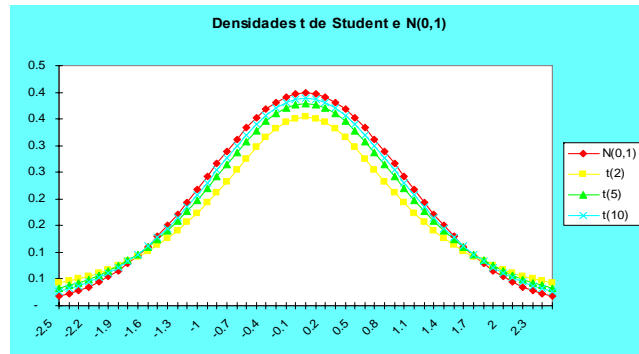
$$T = \frac{Z}{\sqrt{V/k}}$$

- ❑ Onde Z é  $N(0,1)$  e V é Qui-Quadrado com k graus de liberdade, e ambos são independentes.
- ❑ Esta distribuição é simétrica em torno de zero, também tem forma de sino e, à medida que o número de graus de liberdade cresce, se aproxima da  $N(0,1)$ .

## Distribuições Derivadas da Normal



- Quando  $n$  (número de graus de liberdade) cresce, a densidade  $t$  de Student se torna cada vez mais parecida com uma  $N(0,1)$



monica@mbarros.com

85

## Distribuições Derivadas da Normal



- Exemplo (uso de uma tabela t)

graus de liberdade	0.9	0.95	0.975	0.99	0.995
1	3.078	6.314	12.706	31.821	63.657
5	1.476	2.015	2.571	3.365	4.032
10	1.372	1.812	2.228	2.764	3.169
15	1.341	1.753	2.131	2.602	2.947
20	1.325	1.725	2.086	2.528	2.845

- Por exemplo, se  $T$  tem 10 graus de liberdade, a probabilidade de  $T$  ser menor que 1.372 é 90%. Se o número de graus de liberdade passa a 15, o valor tal que a probabilidade de  $T$  ser menor que ele é 90% passa a ser 1.341.

monica@mbarros.com

86

## Distribuições Derivadas da Normal



- Função do Excel para a distribuição  $t$

Função	Descrição
<code>invt(p; gl)</code>	Para a distribuição $t$ de Student, calcula o valor $t$ para $p = 2\alpha$ , com $gl$ graus de liberdade

- Por exemplo,  $INVT(0.05, 20) = 2.086$  é o valor da distribuição  $t$  com 20 graus de liberdade tal que  $Pr(T > 2.086) = 0.05/2 = 0.025$ .

- CUIDADO** com a especificação da probabilidade para esta função, a função `INVT` fornece as probabilidades “bi-laterais”.

monica@mbarros.com

87

## Distribuições Derivadas da Normal



- Refazemos a seguir o exemplo anterior com a função `INVT` do Excel. Note a especificação da probabilidade como  $0.20 = 2\alpha$ , enquanto na nossa tabela as colunas referem-se a  $1 - \alpha$ .

monica@mbarros.com

88

## Distribuições Derivadas da Normal



- **Por que a densidade t é importante?**
- Ela é essencial no contexto de intervalos de confiança e testes de hipóteses, como veremos posteriormente. A justificativa vem, em parte, do próximo resultado.
- Teorema
- Sejam  $X_1, X_2, \dots, X_n$  uma amostra aleatória da distribuição  $N(\mu, \sigma^2)$ . Sejam  $\bar{X}$  e  $S^2$  a média e variância amostrais. Então:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

## Distribuições Derivadas da Normal



- **A distribuição F**
- Sejam  $V$  e  $W$  variáveis aleatórias independentes com densidades Qui-quadrado com  $p$  e  $q$  graus de liberdade respectivamente. Construa uma nova variável aleatória  $X$  como:

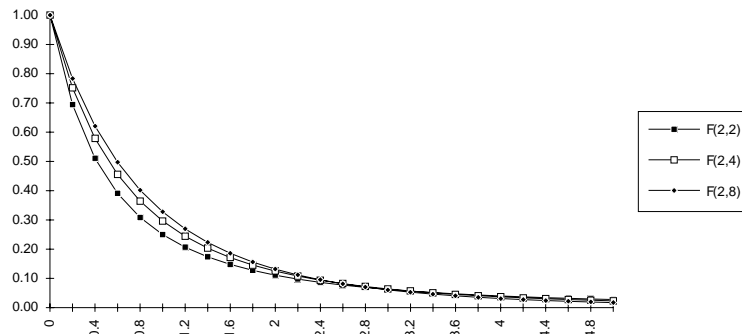
$$X = \frac{V/p}{W/q} = \frac{qV}{pW}$$

- Então  $X$  tem densidade  $F$  com  $p$  graus de liberdade no numerador e  $q$  graus de liberdade no denominador, e escrevemos:  $X \sim F(p, q)$ .

## Distribuições Derivadas da Normal



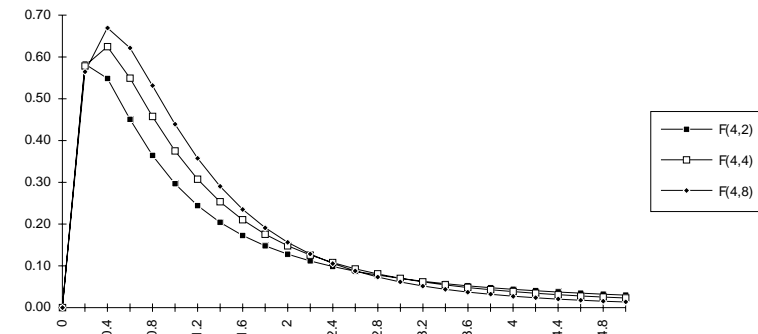
Densidades  $F(2,2)$ ,  $F(2,4)$  e  $F(2,8)$



## Distribuições Derivadas da Normal



Densidades  $F(4,2)$ ,  $F(4,4)$  e  $F(4,8)$



## Distribuições Derivadas da Normal



- O primeiro parâmetro da densidade F indica o número de graus de liberdade do numerador, enquanto o segundo parâmetro refere-se aos graus de liberdade do denominador.
- A densidade F **não é simétrica** em torno de qualquer ponto, e dependendo do número de graus de liberdade no numerador, ela pode ter um comportamento "exponencial" ou então pode ter um máximo global.

## Distribuições Derivadas da Normal



- **Resultado importante**
  - Se  $X \sim F(p,q)$  então  $1/X \sim F(q, p)$ .
  - A demonstração disso é trivial se você sabe como uma distribuição F é criada.
- A importância da distribuição F ficará evidente quando estudarmos intervalos de confiança para a variância da Normal. Por enquanto, iremos apenas enunciar o próximo resultado, que é importante para demonstrações futuras.

## Distribuições Derivadas da Normal



- Teorema
- Considere duas amostras independentes de tamanhos  $n_1$  e  $n_2$  obtidas a partir de duas populações Normais com variâncias  $\sigma_1^2$  e  $\sigma_2^2$ . Sejam  $S_1^2$  e  $S_2^2$  as variâncias amostrais. Então:

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

- tem densidade F com  $n_1 - 1$  graus no numerador e  $n_2 - 1$  graus no denominador.

## Distribuições Derivadas da Normal



- Funções do Excel para a distribuição F

Valor cuja prob. de estar ABAIXO dele queremos encontrar

Graus no numerador

Graus no denominador

## Distribuições Derivadas da Normal



### □ Funções do Excel para a distribuição F

Da figura segue que  $\Pr(X > 9.01) = 0.05$  quando  $X$  é uma variável  $F(5,3)$

monica@mbarros.com

97



## Estatística

## Conteúdo



- Diferença entre Probabilidade e Estatística
- Amostra Aleatória
- Objetivos da Estatística
- Distribuição Amostral
- Estimação Pontual
- Estimação Bayesiana X Clássica
- Método de Máxima Verossimilhança

monica@mbarros.com

99

## Probabilidade e Estatística – Qual a Diferença?



- Até agora tivemos estivessemos interessados em Probabilidade, ou seja, nosso objetivo era:
  - apresentar alguns dos modelos probabilísticos mais usuais e as situações em que eles surgem na prática.

monica@mbarros.com

100

## Diferença entre Probabilidade e Estatística



- ❑ A partir de agora começamos realmente a falar de Estatística.
- ❑ Os capítulos anteriores lidavam com Probabilidade. Qual a diferença?
- ❑ **Em Probabilidade, a densidade (ou função de probabilidade) era inteiramente conhecida.**
- ❑ Em Estatística, teremos uma amostra aleatória de uma distribuição com certos **parâmetros desconhecidos**, e procuraremos descobrir alguma coisa sobre estes parâmetros.

monica@mbarros.com

101

## Amostra Aleatória (a.a.)



- ❑ **É apenas um conjunto de variáveis aleatórias iid (independentes e identicamente distribuídas).**
- ❑ Se  $X_1, X_2, \dots, X_n$  formam uma a.a. então, em particular, todas as variáveis têm a mesma densidade ou função de probabilidade, e portanto suas médias são todas iguais (o mesmo ocorre com suas variâncias).

monica@mbarros.com

102

## Objetivos - Estatística



- ❑ A distribuição da amostra é conhecida exceto por alguns parâmetros que buscamos estimar.
- ❑ **Objetivo: obter maneiras de encontrar estimadores ("chutes") destes parâmetros. Estes estimadores serão pontuais (e começaremos a estudar um importante método de estimação pontual hoje) ou por intervalos (nas próximas aulas).**

monica@mbarros.com

103

## Objetivos - Estatística



- ❑ Também é preciso ter uma idéia clara das **propriedades** desejáveis **destes estimadores**, e saber, segundo algum critério, se o estimador encontrado é bom ou ruim.
- ❑ Finalmente, em Estatística estamos interessados também em **testar hipóteses** sobre parâmetros desconhecidos.

monica@mbarros.com

104

## Distribuição Amostral



- Uma “estatística” é qualquer função das observações numa amostra aleatória.
- Por exemplo, duas das estatísticas mais usadas são  $\bar{X}$  (a média amostral) e  $S^2$  (a variância amostral).
- Já vimos que:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{e} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

## Distribuição Amostral



- Dada uma amostra aleatória  $X_1, X_2, \dots, X_n$  com uma densidade (ou função de probabilidade), podemos tentar encontrar a densidade da média e da variância amostral, e usá-las para inferir sobre a média e variância verdadeiras (e desconhecidas) de  $X_1, X_2, \dots, X_n$ .

## Estimação Pontual



- Problemas de estimação de parâmetros surgem frequentemente em Ciências e Engenharia. Por exemplo, muitas vezes desejamos estimar os seguintes parâmetros:
- a média de uma população,
- a variância ou desvio padrão de uma população,
- a proporção de itens numa população que pertencem a uma classe de interesse,
- a diferença entre as médias de duas populações.

## Estimação Pontual



- Como estimar estas quantidades? Alguns estimadores razoáveis nestas situações são:
- a média amostral,
- a variância ou desvio padrão amostrais,
- a proporção de itens na amostra que pertencem à classe de interesse,
- a diferença entre as médias amostrais de duas amostras independentes, cada uma representando uma das populações.

## Estimação Pontual



- $X_1, X_2, \dots, X_n$  variáveis aleatórias.
- $x_1, x_2, \dots, x_n$  valores observados das variáveis aleatórias.
- Seja  $X$  uma variável aleatória com densidade  $f(x, \theta)$ , onde  $\theta$  é um parâmetro, e  $\theta \in \Omega$ .
- O conjunto  $\Omega$  é chamado de **espaço paramétrico**.
- Objetivo: estimar  $\theta$ .

## Estimação Pontual



- A densidade de  $X$ ,  $f(x, \theta)$ , tem uma forma conhecida, **exceto pelo parâmetro  $\theta$**  que varia no conjunto  $\Omega$ .
- Assim, não temos apenas uma densidade, mas uma família de densidades. A cada valor de  $\theta$  em  $\Omega$ . corresponde um membro da família.
- Aqui adotaremos o **enfoque "clássico"** de estimação, no qual  $\theta$  é um **parâmetro desconhecido**, suposto constante, e não uma variável aleatória.

## Estimação Bayesiana X Clássica



- Na **estimação Bayesiana**,  $\theta$  será encarado como uma **variável aleatória**, e a ele associaremos uma distribuição de probabilidade.
- A distribuição de probabilidade de  $\theta$  antes de observarmos os dados será chamada de distribuição a priori, e muitas vezes representa o nosso conhecimento subjetivo sobre o parâmetro  $\theta$ .
- A distribuição de  $\theta$  após observarmos a amostra é conhecida como distribuição a posteriori de  $\theta$ .

## Estimação Bayesiana versus Clássica



- Em estatística Bayesiana a verossimilhança (que iremos definir em breve) "carrega" a informação sobre  $\theta$  contida na amostra, e resulta na atualização da densidade de  $\theta$ , passando de uma priori para uma posteriori.
- A densidade a posteriori combina a "informação" subjetiva trazida pela priori com a "informação" proveniente da amostra.
- Os dois enfoques, Clássico e Bayesiano, concordam se o tamanho da amostra é grande.

## Definição do Problema de Estimação Pontual



- *O problema geral aqui é ...*
- A partir dos dados observados  $x_1, x_2, \dots, x_n$  precisamos *escolher um membro* de uma família de densidades para representar estes dados.
- Ou seja, precisamos de um *estimador pontual* de  $\theta$  (um "chute educado" para o valor desconhecido de  $\theta$ ).

## Definição do Problema de Estimação Pontual



- Seja  $X_1, X_2, \dots, X_n$  uma amostra aleatória da densidade  $f(x, \theta)$ .
- O **objetivo** agora é **definir uma estatística**  $T = T(X_1, X_2, \dots, X_n)$  de tal modo que, após observarmos  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$   $t = t(x_1, x_2, \dots, x_n)$  seja uma boa estimativa pontual de  $\theta$ .
- Na verdade, a cada amostra obtida, encontraremos um valor para a estatística usada para "chutar"  $\theta$ , pois esta estatística é também uma variável aleatória

## Exemplo



- No próximo exemplo exibimos a média amostral de 5 amostras de tamanho 50 geradas a partir da densidade  $N(0,1)$  no Excel . A média amostral serve para estimar a média da distribuição (zero, neste caso) e portanto deve ser, para todas as amostras, um valor próximo de zero.
- Os resultados para as 5 amostras geradas estão a seguir.

	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5
Média	0,076	0,150	0,180	-0,199	0,055
Desvio Padrão	1,108	1,060	1,020	1,017	0,923
Mediana	0,168	0,179	0,241	-0,206	0,072

## Exemplo

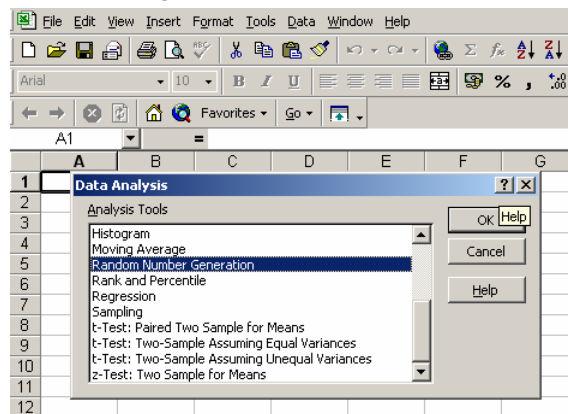


- Note que os valores estimados da média em cada amostra são todos diferentes entre si, e diferentes do valor real da média da população, que é = 0.
- Da mesma maneira, as estimativas do desvio padrão (cujo valor real é 1) são todas diferentes do valor real, e diferentes entre si. Note que , na prática, os valores de  $\mu$  e  $\sigma$  são desconhecidos, o que não acontece neste exemplo, onde geramos amostras de uma distribuição conhecida.

## Exemplo



- ❑ Como fazer a geração destas variáveis Normais no Excel? Lembre-se que o suplemento de análise de dados deve estar previamente instalado.



117

## Exemplo



5 variáveis

50 valores por variável

Pasta onde armazenar resultados (opcional)

Semente do gerador (opcional)

monica@mbarros.com

118

## O que é um bom estimador?



- ❑ Existem potencialmente milhares de estimadores para um certo parâmetro.
- ❑ Por exemplo, para estimar a média de uma população poderíamos usar a média amostral, a mediana amostral, a média entre a menor e a maior observação na amostra e uma infinidade de outros estimadores "razoáveis".

monica@mbarros.com

119

## O que é um bom estimador?



- ❑ Como escolher dentre eles? Quais serão os critérios usados para comparar estimadores e caracterizar os bons estimadores?
- ❑ Por enquanto não responderemos a esta questão, mas começaremos a estudar o (talvez) mais tradicional método de estimação pontual.

monica@mbarros.com

120

## Método da Máx. Verossimilhança



- ❑ A função de verossimilhança (likelihood function)
- ❑ Esta é uma função relativamente simples com um nome indigesto!
- ❑ "Likelihood" em inglês é uma palavra de uso corrente, que indica "plausibilidade". Ao contrário, "verossimilhança" é uma coisa meio obscura.
- ❑ Seja  $X_1, X_2, \dots, X_n$  uma amostra aleatória da densidade  $f(x, \theta)$ .

monica@mbarros.com

121

## Método da Máx. Verossimilhança



- ❑ A **função de verossimilhança** é a densidade conjunta encarada como função do parâmetro  $\theta$ . Isto é:

$$L(\theta) = f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i, \theta)$$

- ❑ A partir da verossimilhança podemos encontrar um estimador, o estimador de máxima verossimilhança (MLE = maximum likelihood estimator).
- ❑ O MLE é obtido a partir da maximização da verossimilhança, geralmente feita através da equação  $dL(\theta)/d\theta = 0$ .

monica@mbarros.com

122

## Método da Máx. Verossimilhança



- ❑ É equivalente maximizar  $L(\theta)$  ou seu logaritmo natural,  $l(\theta) = \log L(\theta)$  onde  $\log(\cdot)$  indica o logaritmo na base e.
- ❑ Esta última função é chamada **log-verossimilhança** e é freqüentemente mais fácil de maximizar do que  $L(\theta)$ , pois as verossimilhanças muitas vezes podem ser escritas como  $\exp\{ \dots \}$ .
- ❑ A equivalência da maximização de  $L(\theta)$  e  $l(\theta)$  decorre do fato de  $L(\theta)$  ser sempre maior que 0 (pois é o produto de densidades) e do logaritmo ser uma função bijetora.

monica@mbarros.com

123

## Método da Máx. Verossimilhança



- ❑ *Por que maximizar a verossimilhança?*
- ❑ Suponha que temos uma amostra aleatória  $X_1, X_2, \dots, X_n$  de uma densidade qualquer, completamente conhecida exceto pelo parâmetro  $\theta$ .
- ❑ Ao observarmos cada  $x_i$ , a densidade conjunta fica completamente especificada exceto pelo valor de  $\theta$ . Então, por que não "chutar" para  $\theta$  o valor que torna esta função um máximo?
- ❑ Este "chute" para  $\theta$  é o valor que **mais concorda com os dados** observados.

monica@mbarros.com

124

## Exemplo 1 - MLE (Poisson)



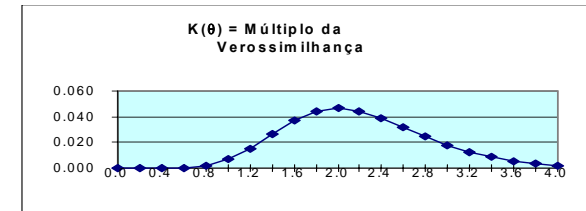
- Suponha que obtemos uma amostra aleatória de tamanho 5 da distribuição Poisson com média  $\theta$ .
- Os valores observados na amostra são: 0, 6, 1, 2 e 1.
- Então a função de probabilidade conjunta é:

$$L(\theta) = \prod_{i=1}^n f(x_i, \theta) = \prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{e^{-5\theta} e^{\sum x_i}}{\prod_{i=1}^5 x_i!}$$
$$L(\theta) = \frac{e^{-5\theta} \theta^{10}}{0!6!1!2!1!} = \frac{\theta^{10} e^{-5\theta}}{1440}$$

## Exemplo 1 - MLE (Poisson)



- Seja  $K(\theta) = 1440.L(\theta) = \theta^{10}e^{-5\theta}$
- Podemos fazer um gráfico de  $K(\theta)$  e ver qual o valor que aparentemente maximiza esta função, ou, alternativamente, fazer um gráfico de  $L(\theta)$  ou  $l(\theta)$ . O gráfico de  $K(\theta)$  é:



- O máximo aparente ocorre em  $\theta = 2$ .

## Método da Máx. Verossimilhança



- Podemos confirmar se este valor realmente corresponde ao máximo através de técnicas simples do Cálculo.
- Lembre-se que uma **condição necessária** (mas não suficiente) para a existência de um **máximo local** é que a **primeira derivada** da função de interesse seja **zero**.
- Isso nos leva à idéia de "equação de máxima verossimilhança", discutida a seguir.

## Método da Máx. Verossimilhança



- Para maximizar  $L(\theta)$ , uma condição necessária é que sua primeira derivada seja igual a zero.
- Assim, a **equação de máxima verossimilhança** é:
$$\frac{dL(\theta)}{d\theta} = 0$$
- e esta equação deve ser resolvida, por métodos analíticos ou numéricos para  $\theta$ .
- Para assegurar que a solução de  $dL/d\theta = 0$  seja realmente um máximo da verossimilhança, precisamos garantir que a segunda derivada seja  $\leq 0$ .

## Método da Máx. Verossimilhança



- ❑ A equação de máxima verossimilhança pode ser reescrita em termos da log-verossimilhança. Assim, é equivalente resolver:

$$\frac{d(\log L(\theta))}{d\theta} = 0 \Leftrightarrow \frac{dl(\theta)}{d\theta} = 0 \text{ para } \theta.$$

- ❑ O estimador obtido pela maximização da função de verossimilhança é chamado **de estimador de máxima verossimilhança (MLE)**.
- ❑ Notação:
- ❑ Geralmente denotaremos o MLE por  $\hat{\theta} = T(X_1, X_2, \dots, X_n)$

## Método da Máx. Verossimilhança



- ❑ **Atenção**
- ❑ Em muitos casos o estimador de máxima verossimilhança é único e pode ser obtido por métodos analíticos.
- ❑ Em outros casos, a equação de máxima verossimilhança  $dL/d\theta = 0$  (ou  $dl/d\theta = 0$ ) não nos dá o resultado correto e precisaremos encontrar o máximo da verossimilhança por outros métodos (por exemplo, graficamente)

## Exemplo 2 - MLE (Poisson)



- ❑ Considere o exemplo 1. A log verossimilhança é:

$$l(\theta) = -5.\theta + 10.\log \theta - \log(1440)$$

- ❑ Derivando esta última expressão com relação a  $\theta$  e igualando a zero leva a:

$$\frac{dl}{d\theta} = 0 \rightarrow -5 + \frac{10}{\theta} = 0 \rightarrow -5.\theta = -10 \rightarrow \hat{\theta} = 2$$

- ❑ é o estimador de máxima verossimilhança para  $\theta$ .
- ❑ Compare este resultado com o exemplo 1. Este resultado **não** é mera coincidência.

## Exemplo 3 (Bernoulli)



- ❑ Sejam  $X_1, X_2, \dots, X_n$  iid Bernoulli( $\theta$ ).
- ❑ A função de probabilidade de cada  $X_i$  é:

$$f(x_i, \theta) = \theta^{x_i} (1 - \theta)^{1 - x_i} \quad \theta \in (0, 1), \quad x_i = 0, 1$$

- ❑ A função de verossimilhança o produto das funções de probabilidade individuais, isto é:

$$\begin{aligned} L(\theta) &= f(x_1, x_2, \dots, x_n, \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1 - x_i} = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} = \\ &= \theta^{n\bar{X}} (1 - \theta)^{n - n\bar{X}} = \exp\{n\bar{X} \cdot \log \theta + (n - n\bar{X}) \log(1 - \theta)\} \end{aligned}$$

### Exemplo 3 (Bernoulli)



□ A log verossimilhança é:

$$l(\theta) = \log(L(\theta)) = \left(\sum x_i\right) \log \theta + (n - \sum x_i) \log(1 - \theta) = n\bar{X} \cdot \log \theta + (n - n\bar{X}) \log(1 - \theta)$$

□ Resolvendo a equação de verossimilhança leva a:

$$\frac{dl}{d\theta} = 0 \Rightarrow \frac{n\bar{X}}{\theta} - \frac{(n - n\bar{X})}{1 - \theta} = 0 \\ \Leftrightarrow (1 - \theta)n\bar{X} = n\theta - n\bar{X}\theta \Leftrightarrow n\theta = n\bar{X}$$

□ E então o MLE para  $\theta$  é a média amostral.

□ Verifique que  $\left. \frac{d^2l}{d\theta^2} \right|_{\theta = \bar{X}} < 0$

□ de tal forma que a média amostral realmente MAXIMIZA a verossimilhança.

### Exemplo 4 (Normal)



□ Sejam  $X_1, X_2, \dots, X_n$  iid Normal( $\mu, 1$ ), ou seja, uma Normal com média desconhecida e variância conhecida (e suposta igual a um sem perda de generalidade).

□ Mostre que o MLE de  $\mu$  é a média amostral.

### Exemplo 4 (Normal)



□ A verossimilhança é:

$$L(\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(X_i - \mu)^2\right] = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2\right\}$$

□ Note que a verossimilhança é máxima quando  $Q(\mu) = \sum (X_i - \mu)^2$  é mínimo.

□ Então é equivalente maximizar  $L(\mu)$  ou minimizar  $Q(\mu)$ .

### Exemplo 4 (Normal)



$$Q(\mu) = \sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i^2 - 2\mu X_i + \mu^2) = \sum_{i=1}^n X_i^2 - 2\mu n\bar{X} + n\mu^2$$

□ Derivando  $Q(\mu)$  em relação a  $\mu$  e igualando a zero nos leva a um ponto crítico:

$$\frac{dQ(\mu)}{d\mu} = 0 \rightarrow -2 \sum_{i=1}^n X_i + 2n\mu = 0 \Leftrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

□ Logo, o MLE de  $\mu$  é a média amostral.

## Estimador não tendencioso



- ❑ A primeira característica desejável num estimador é a não tendenciosidade. Esta é uma característica desejável, mas **não é a única desejável** ou sequer a mais importante.
- ❑ **Definição (Estimador não tendencioso)**
- ❑ Seja  $\tilde{\theta} = T(X_1, X_2, \dots, X_n)$  um estimador para o parâmetro  $\theta$  de uma densidade  $f(x, \theta)$ .  $\tilde{\theta}$  é chamado de **não tendencioso** se  $E(\tilde{\theta}) = \theta$ , do contrário  $\tilde{\theta}$  é dito tendencioso.

## Exemplo 5 - MLE (Poisson)



- ❑ Considere novamente os exemplos da distribuição de Poisson.
- ❑ O estimador de máxima verossimilhança de  $\theta$  é  $\hat{\theta} = \bar{X}$
- ❑ Mas, os  $X_i$  são iid Poisson( $\theta$ ), para  $i = 1, 2, \dots, n$ , e então  $E(X_i) = \theta$ . Mas,

$$\hat{\theta} = \bar{X} \text{ e } E(\hat{\theta}) = E(\bar{X}) = E(X_i) = \theta$$

e assim  $\hat{\theta} = \bar{X}$  é **não tendencioso** para  $\theta$ .

## Exemplo 6 - MLE (Normal)



- ❑ Sejam  $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ .
  - ❑ Sejam  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  e  $S^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$
  - ❑ Mas,
- $$\bar{X} \sim N(\mu, \sigma^2 / n)$$
- ❑ e então  $\bar{X}$  é um estimador **não tendencioso** para  $\mu$ .
  - ❑ Também,

$$E(S^{*2}) = \frac{(n-1)\sigma^2}{n} \neq \sigma^2$$

- ❑ e assim  $S^{*2}$  é um estimador **tendencioso** para  $\sigma^2$ .

## Intervalos de Confiança



## Conteúdo



- Intervalos de Confiança – Motivação
- Intervalos de Confiança para Médias
- Intervalos de Confiança para Diferenças entre Médias (Variâncias supostas iguais)
- Intervalo de Confiança para a variância de uma Normal
- Intervalos de Confiança para a razão de variâncias
- Intervalo de Confiança aproximado para a proporção uma Binomial

## Intervalos de Confiança



- Até agora estivemos interessados em encontrar uma estimativa pontual para um parâmetro desconhecido  $\theta$ .
- Também enumeramos algumas propriedades desejáveis de estimadores pontuais.
- **Agora tentaremos obter não apenas uma estimativa pontual, mas um intervalo** que contenha o parâmetro de interesse com uma **probabilidade especificada**. Este intervalo será chamado de “Intervalo de Confiança”.

## Intervalos de Confiança



O intervalo de confiança  $100(1-\alpha)\%$  para  $\theta$  é dado por:

$$L(\tilde{X}) \leq \theta \leq U(\tilde{X})$$

Onde  $L(\tilde{X})$  (limite inferior) e  $U(\tilde{X})$  (limite superior) são tais que:

$$Prob[L(\tilde{X}) \leq \theta \leq U(\tilde{X})] = 1 - \alpha$$

**Onde  $\alpha$  é um número especificado pelo usuário.**

## Intervalos de Confiança



- Note que o intervalo  $[L(\tilde{X}), U(\tilde{X})]$  é **aleatório**, e a cada amostra obtida iremos encontrar valores diferentes para os limites L e U.
- A notação  $\tilde{X}$  indica todos os elementos da amostra aleatória, isto é:

$$\tilde{X} = (X_1, X_2, \dots, X_n)$$

## Intervalos de Confiança – Média da Normal



- ❑ Consideraremos agora o caso mais comum na prática onde os dados são supostos **NORMAIS** e  $\theta$  é **média** da distribuição.
  
- ❑ Serão abordados dois casos: variância do modelo conhecida e variância do modelo desconhecida.

## Intervalos de Confiança – Média da Normal



- ❑ **Argumento intuitivo....**
- ❑ Suponha que você tem uma amostra aleatória da Normal, em que a média é desconhecida.
  
- ❑ Se você precisasse achar um estimador pontual de  $\theta$  (a média), usaria a média amostral  $\bar{X}$ .

## Intervalos de Confiança – Média da Normal



- ❑ E se agora você precisar encontrar um intervalo que contenha  $\theta$  com uma probabilidade especificada?
  
- ❑ Parece natural que este intervalo tenha a forma:  $(\bar{X} - c, \bar{X} + c)$  onde  $c$  é uma constante a ser especificada. Veremos que os intervalos encontrados para a média da Normal têm exatamente esta forma!

## Intervalo de Confiança – Média da Normal



### Caso I

$X \sim \text{NORMAL}(\theta, \sigma^2)$ ;  $\sigma^2$  conhecido

- ❑ Seja  $\tilde{X} = (X_1, \dots, X_n)$  uma a.a. de tamanho  $n$  da distribuição Normal acima.
- ❑ Já vimos que  $\bar{X} = \sum \frac{X_i}{n}$  é o estimador de máxima verossimilhança de  $\theta$ . Além disto, é fácil provar que:

$$\bar{X} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

## Intervalo de Confiança – Média da Normal



- Logo, podemos padronizar a média amostral, transformando-a numa v.a. com densidade  $N(0,1)$  da seguinte maneira:

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

- Usando uma tabela da Normal podemos encontrar, **por exemplo**, a probabilidade desta nova variável estar **entre -2 e +2**.

## Intervalo de Confiança – Média da Normal



$$\text{Prob}(-2 < Z < 2) = \Phi(2) - \Phi(-2) = 0.954$$

- Substituindo Z na expressão anterior leva a:

$$-2 < \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} < +2 \Leftrightarrow \bar{X} - \frac{2\sigma}{\sqrt{n}} < \theta < \bar{X} + \frac{2\sigma}{\sqrt{n}}$$

- Daí:

$$\text{Prob}\{-2 < Z < +2\} = \text{Prob}\left\{\bar{X} - \frac{2\sigma}{\sqrt{n}} < \theta < \bar{X} + \frac{2\sigma}{\sqrt{n}}\right\} = 0.954$$

- O intervalo que acabamos de encontrar é um **intervalo de confiança 95.4% para  $\theta$** .

## Intervalo de Confiança – Média da Normal



- Ou seja, na notação mostrada antes:

$$L(\tilde{X}) = \bar{X} - \frac{2\sigma}{\sqrt{n}}$$

$$U(\tilde{X}) = \bar{X} + \frac{2\sigma}{\sqrt{n}}$$

$$1 - \alpha = 0.954$$

- A seguir exibimos uma “receita de bolo” para obter o IC da média de uma Normal com variância conhecida.

## Intervalo de Confiança – Média da Normal



- **Receita de Bolo**

- Seja  $\tilde{X} = (X_1, \dots, X_n)$  uma a.a. de tamanho n da distribuição Normal com **média desconhecida  $\theta$**  e **variância conhecida  $\sigma^2$** .

- Um intervalo de confiança  $100(1 - \alpha)\%$  para  $\theta$  é dado por:

$$\left[\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}; \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$$

- Onde  $z_{1-\alpha/2}$  é obtido da função de distribuição  $N(0,1)$  e é tal que  $\text{Pr}(Z < z_{1-\alpha/2}) = 1 - \alpha/2$ .

## Intervalo de Confiança – Média da Normal



- Note que, pela simetria em torno de zero da distribuição  $N(0,1)$ :
- $z_{1-\alpha/2}$  é o ponto tal que, a **probabilidade de estar ACIMA dele é  $\alpha/2$**  usando uma distribuição  $N(0,1)$ .
- Também é fácil perceber que, se  $Z$  é  $N(0,1)$ :

$$\Pr\left\{-z_{1-\frac{\alpha}{2}} < Z < +z_{1-\frac{\alpha}{2}}\right\} = 1 - \alpha$$

- E esta última expressão foi empregada para obter o IC para a média.

## IC para a média da Normal com $\sigma$ conhecido



- **Exemplo**
- Considere a população de alunos da PUC. Para uma amostra de 50 alunos obtivemos uma altura média de 1,68m.
- Sabe-se que o desvio-padrão da altura da população de alunos da PUC é o mesmo que o da população de jovens cariocas com menos de 25 anos: 0,11m.
- Suponha que as alturas dos alunos são Normalmente distribuídas.
- Determine, com um **nível de confiança de 95%**, o intervalo onde a real altura média da população de alunos da PUC deve estar localizada.

## IC para a média da Normal com $\sigma$ conhecido



### □ Solução

- Note que a amostra é Normal com variância conhecida, e assim a distribuição de  $\bar{X}$  também é Normal.
- Da tabela da Normal, ou usando a função **INV.NORMP** do Excel, procuramos um valor  $z_0$  tal que  $\Pr(Z < z_0) = 1 - \alpha/2 = 97.5\%$ , isto é,  $\Phi(z_0) = 97.5\%$ . A função INV.NORMP fornece  $z_0 = 1.96$ .

## IC para a média da Normal com $\sigma$ conhecido



### □ Solução

- O IC 95% (para as alturas em cm) é então:

$$\left(\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = \left(168 - 1.96 \frac{11}{\sqrt{50}}, 168 + 1.96 \frac{11}{\sqrt{50}}\right) \\ = (164.95 \text{ cm}, 171.05 \text{ cm})$$

## IC para a média da Normal com $\sigma$ conhecido



❑ Receita de bolo – qual valor de  $z_{\alpha/2}$  usar?

Coefficiente de Confiança	valor tabelado de z
80.0%	1.282
90.0%	1.645
95.0%	1.960
97.0%	2.170
97.5%	2.241
99.0%	2.576

Estes pontos são encontrados através da função INV.NORMP do Excel – Note que, se o coeficiente de confiança é  $1 - \alpha$ , devemos buscar um ponto na tabela da Normal tal que a probabilidade de estar **ACIMA** dele é  $\alpha/2$ , ou seja, a probabilidade de estar **ABAIXO** dele é  $1 - \alpha/2$  (o argumento da função INV.NORMP é  $1 - \alpha/2$ ).

## IC para a média da Normal com $\sigma$ conhecido



1.96 (a “resposta da função” é tal que a probabilidade de estar abaixo deste valor é 0,975)

## IC para a média da Normal com $\sigma$ conhecido



❑ Exemplo

- ❑ Numa amostra de 36 postos de gasolina no Rio de Janeiro, o preço médio do litro da gasolina aditivada foi de R\$ 1.78. Sabe-se, por experiências anteriores, que o desvio padrão é R\$ 0.20.
- ❑ Encontre intervalos de confiança 90%, 95% e 99% para o preço médio da gasolina aditivada no Rio de Janeiro supondo que a amostra é Normal.

❑ Solução

- ❑ Aqui estamos supondo que o desvio padrão é conhecido, e assim podemos usar um intervalo baseado na densidade Normal.

## IC para a média da Normal com $\sigma$ conhecido



- ❑ Os IC têm a forma geral:  $\left( \bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$
- ❑ O IC 90% é:  $\left( 1.78 - 1.645 \frac{(0.20)}{6}, 1.78 + 1.645 \frac{(0.20)}{6} \right) = (\text{R\$ } 1.725, \text{R\$ } 1.835)$
- ❑ O IC 95% é:  $\left( 1.78 - 1.96 \frac{(0.20)}{6}, 1.78 + 1.96 \frac{(0.20)}{6} \right) = (\text{R\$ } 1.715, \text{R\$ } 1.845)$
- ❑ O IC 99% é:  $\left( 1.78 - 2.576 \frac{(0.20)}{6}, 1.78 + 2.576 \frac{(0.20)}{6} \right) = (\text{R\$ } 1.694, \text{R\$ } 1.866)$

Note que, à medida que o coeficiente de confiança aumenta, a largura do intervalo também aumenta!

## IC para a média da Normal com $\sigma$ conhecido



- ❑ Exemplo (para casa)
- ❑ O preço médio de um automóvel Palio ELX 1.0 4 portas ano 2001 é R\$ 17727 (segundo o Jornal Valor Econômico de 07/07/2003).
- ❑ Suponha que o desvio padrão REAL dos preços seja R\$ 1500 e o tamanho da amostra é  $n = 25$  carros.
- ❑ Encontre intervalos de confiança 95% e 99% para os preços de Palios ELX 1.0 quatro portas ano 2001 supondo que os preços são Normalmente distribuídos.

## IC para a média da Normal com $\sigma$ conhecido



- ❑ Exemplo (para casa)
- ❑ Toma-se uma amostra de 25 usuário de um cartão de crédito e observa-se que o gasto médio mensal é R\$ 600.
- ❑ O desvio padrão é conhecido e igual a R\$ 250.
- ❑ Encontre intervalos de confiança 95 e 99% para o gasto médio com cartão na população de usuários.

## PIVOT



- ❑ Seja  $\tilde{X} = (X_1, \dots, X_n)$  uma a.a. de tamanho  $n$  de uma densidade (ou função de probabilidade)  $f(x, \theta)$ .
- ❑ Seja  $Q = q(X_1, \dots, X_n, \theta)$  uma função dos elementos da amostra e do parâmetro desconhecido  $\theta$ .
- ❑ **Q é chamado de PIVOT se sua distribuição não depende de  $\theta$ .**
- ❑ Um PIVOT é usado para encontrar intervalos de confiança para parâmetros desconhecidos.

## PIVOT



- ❑ No exemplo do IC da média da Normal com variância conhecida, a quantidade:

$$Z = \frac{\sqrt{n}(\bar{X} - \theta)}{\sigma}$$

- ❑ é um PIVOT, pois depende de  $\tilde{X} = (X_1, \dots, X_n)$  e  $\theta$ , sua distribuição não depende de  $\theta$  (pois é  $N(0,1)$ ) e assim pode ser usada na construção de um IC para  $\theta$ .

## IC para a média da Normal com $\sigma$ desconhecido



### Caso II

### $X \sim \text{NORMAL}(\theta, \sigma^2)$ ; $\sigma^2$ DESCONHECIDO

□ Seja  $X = (X_1, \dots, X_n)$  uma a.a. de tamanho  $n$  da distribuição Normal acima.

□ Os estimadores **não tendenciosos** de  $\theta$  e  $\sigma^2$

são:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  e  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

onde  $\bar{X} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$  e  $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$

## IC para a média da Normal com $\sigma$ desconhecido



□ Também,  $\bar{X}$  e  $S^2$  são independentes.

□ Pela definição de uma v.a. t de Student:

$$T = \frac{\sqrt{n}(\bar{X} - \theta)}{\frac{\sigma}{\sqrt{(n-1)S^2}}} = \sqrt{n} \cdot \frac{\bar{X} - \theta}{S} \sim t_{n-1}$$

□ Onde:  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

□ Assim da tabela da distribuição t de Student com  $n-1$  graus de liberdade podemos obter dois números  $\underline{a}$  e  $\underline{b}$  tais que:  $\Pr(a < T < b) = 1 - \alpha$

## IC para a média da Normal com $\sigma$ desconhecido



□ Para encontrar um intervalo simétrico fazemos  $a = -b$  e assim:

$$\text{Prob}[a < T < b] = \text{Prob}\{-b < T < +b\} = \text{Prob}\left\{-b < \sqrt{n} \left( \frac{\bar{X} - \theta}{S} \right) < b\right\} = 1 - \alpha$$

$$\Leftrightarrow \text{Prob}\left(-b \frac{S}{\sqrt{n}} < \bar{X} - \theta < +b \frac{S}{\sqrt{n}}\right) =$$

$$= \text{Prob}\left(-\bar{X} - b \frac{S}{\sqrt{n}} < -\theta < -\bar{X} + b \frac{S}{\sqrt{n}}\right) =$$

$$= \text{Prob}\left(\bar{X} - b \frac{S}{\sqrt{n}} < \theta < \bar{X} + b \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

## IC para a média da Normal com $\sigma$ desconhecido



□ Portanto:

□ O intervalo  $\left(\bar{X} - b \frac{S}{\sqrt{n}}, \bar{X} + b \frac{S}{\sqrt{n}}\right)$

□ é um intervalo aleatório com probabilidade  $1 - \alpha$  de incluir o parâmetro desconhecido  $\theta$ .

□ O ponto  $\underline{b}$  que aparece na definição do IC é obtido da distribuição t com  $n-1$  graus de liberdade, e é tal que  $\Pr(T > b) = \alpha/2$ .

## IC para a média da Normal com $\sigma$ desconhecido



- ❑ **Receita de Bolo**
- ❑ Seja  $X_1, X_2, \dots, X_n$  uma a.a. de tamanho  $n$  da distribuição Normal com **média desconhecida  $\theta$**  e **variância desconhecida  $\sigma^2$** .
- ❑ Um intervalo de confiança  $100(1 - \alpha)\%$  para  $\theta$  é dado por: 
$$\left( \bar{X} - b \frac{S}{\sqrt{n}}, \bar{X} + b \frac{S}{\sqrt{n}} \right)$$
- ❑ Onde  $b$  é obtido da função de distribuição t de Student com  $n-1$  graus de liberdade e é tal que  $\Pr(T > b) = \alpha/2$ .

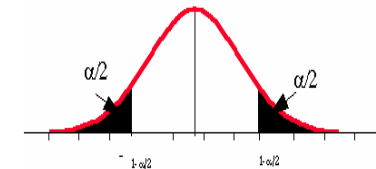
## IC para a média da Normal com $\sigma$ desconhecido



- ❑ O IC  $100(1-\alpha)\%$  para  $\theta$  é:

$$\left( \bar{X} - t_{n-1;1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1;1-\alpha/2} \frac{S}{\sqrt{n}} \right)$$

- ❑ Onde  $S$  é o desvio padrão amostral e  $t_{n-1,1-\alpha/2}$  é um ponto da distribuição t de Student com  $n-1$  graus de liberdade tal que  $\Pr(T > t_{n-1,1-\alpha/2}) = \alpha/2$ , como no gráfico a seguir:



## IC para a média da Normal com $\sigma$ desconhecido



- ❑ O valor  $t_{n-1,1-\alpha/2}$  é obtido de uma tabela da distribuição t com  $n-1$  graus de liberdade. Pode-se, alternativamente, usar a função INVT do Excel.

## IC para a média da Normal com $\sigma$ desconhecido



- ❑ **Exemplo**
- ❑ Numa amostra de 16 postos de gasolina no Rio de Janeiro, o preço médio do litro da gasolina aditivada foi de R\$ 1.78.
- ❑ O desvio padrão dos preços **estimado** na amostra é R\$ 0.20. Encontre intervalos de confiança 90%, 95% e 99% para o preço médio da gasolina aditivada no Rio de Janeiro e compare-os com os encontrados no exemplo da página 18.

## IC para a média da Normal com $\sigma$ desconhecido



- ❑ Solução
- ❑ Aqui deve-se usar a distribuição t para encontrar o IC, pois o desvio padrão é desconhecido. A forma do intervalo é:

$$IC = \bar{X} \pm t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} = \left( \bar{X} - t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \right)$$

- ❑ Pela função INVT do Excel com 15 graus de liberdade obtemos os pontos percentuais para os IC 90, 95 e 99%, que são, respectivamente: 1.753, 2.131 e 2.947.

## IC para a média da Normal com $\sigma$ desconhecido



- ❑ O IC 90% é:  $\left( 1.78 - 1.753 \frac{(0.20)}{\sqrt{16}}, 1.78 + 1.753 \frac{(0.20)}{\sqrt{16}} \right) = (\text{R\$ } 1.692, \text{ R\$ } 1.868)$
- ❑ O IC 95% é:  $\left( 1.78 - 2.131 \frac{(0.20)}{\sqrt{16}}, 1.78 + 2.131 \frac{(0.20)}{\sqrt{16}} \right) = (\text{R\$ } 1.673, \text{ R\$ } 1.887)$
- ❑ O IC 99% é:  $\left( 1.78 - 2.947 \frac{(0.20)}{\sqrt{16}}, 1.78 + 2.947 \frac{(0.20)}{\sqrt{16}} \right) = (\text{R\$ } 1.633, \text{ R\$ } 1.927)$

**Note que os intervalos de confiança são mais largos que os correspondentes para a Normal**

## Nota IMPORTANTE – uso de INVT no Excel



- ❑ Suponha que você quer encontrar um intervalo de confiança  $100*(1 - \alpha)\%$ .
  - ❑ Então para obter o ponto  $t_{1-\alpha/2}$  que entra no cálculo do IC, use a função INVT com os argumentos:
    - ❑  $\alpha$  e
    - ❑  $n - 1$  graus de liberdade
    - ❑ Pois a função INVT do Excel fornece a o ponto tal que a probabilidade de estar ACIMA dele é especificada.
  - ❑ Isso se deve ao fato do primeiro argumento da função no Excel ser, na verdade, o valor para o intervalo bilateral.

## Utilizando o Excel



- ❑ Funções do Excel para a distribuição t

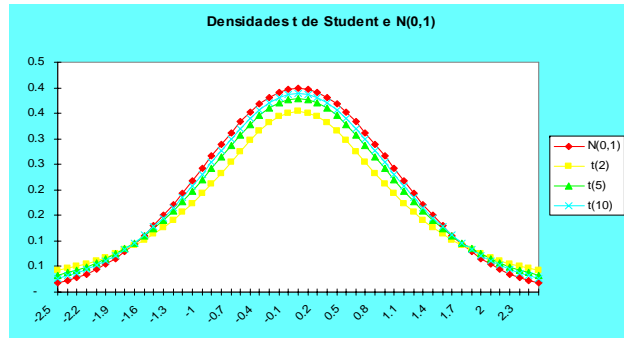
Função	Descrição
invtp; gl)	Para a distribuição t de Student, calcula o valor t para $p = 2.\alpha$ , com gl graus de liberdade

- ❑ Por exemplo,  $INVT(0.05, 20) = 2.086$  calcula o valor na tabela t com 20 graus de liberdade e é tal que  $Pr(T > 2.086) = 0.05/2 = 0.025$

## Distribuição t de Student



- Quando  $n$  (número de graus de liberdade) cresce, a densidade t de Student se torna cada vez mais parecida com uma  $N(0,1)$

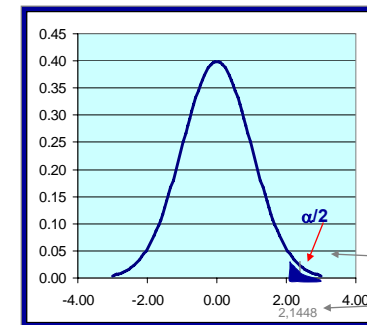


177

## A distribuição t de Student



- Exemplo: para uma amostra com 15 elementos (14 graus de liberdade) e para um nível de confiança de 5% ( $\alpha/2 = 0,025$ ),  $t$  é igual a 2,1448



G.L	0.100	0.075	0.050	0.025	0.020
1	3.0777	4.1653	6.3137	12.7062	15.8945
2	1.8856	2.2819	2.9200	4.3027	4.8487
3	1.6377	1.9243	2.3534	3.1824	3.4819
4	1.5332	1.7782	2.1318	2.7765	2.9985
5	1.4759	1.6994	2.0150	2.5706	2.7565
6	1.4398	1.6502	1.9432	2.4469	2.6122
7	1.4149	1.6166	1.8946	2.3646	2.5168
8	1.3968	1.5922	1.8595	2.3060	2.4490
9	1.3830	1.5737	1.8331	2.2622	2.3984
10	1.3722	1.5592	1.8125	2.2281	2.3593
11	1.3634	1.5476	1.7959	2.2010	2.3281
12	1.3562	1.5380	1.7823	2.1788	2.3027
13	1.3502	1.5299	1.7709	2.1604	2.2816
14	1.3450	1.5231	1.7613	2.1448	2.2638
15	1.3406	1.5172	1.7531	2.1315	2.2485
16	1.3368	1.5121	1.7459	2.1199	2.2354

monica@mbarros.com

178

## Comparação: IC Normais x IC t de Student



- A distribuição t nos fornece intervalos de comprimento maior que os intervalos Normais com a mesma probabilidade.
- À medida que o número de graus de liberdade da densidade t cresce, a densidade se torna mais e mais parecida com uma  $N(0,1)$ , e conseqüentemente, os intervalos se tornam mais próximos dos encontrados através da distribuição  $N(0,1)$ .

179

## Comparação: IC Normais x IC t de Student



- Também, o comprimento dos intervalos diminui à medida que aumentamos o número de observações.
- Isto é intuitivamente razoável, pois à medida que o tamanho da amostra cresce,  $\bar{X}$  “converge” para  $\mu$  e temos cada vez mais “certeza” de que a média amostral está num intervalo de pequeno comprimento em torno de  $\mu$  com alta probabilidade (este resultado é conhecido como “lei dos grandes números”).

monica@mbarros.com

180

## Utilizando o Excel



- ❑ O Excel também pode ser utilizado para o cálculo do intervalo de confiança para  $\sigma$  desconhecido (para qualquer tamanho de amostra)
  - ❑ Selecione no menu **Ferramentas** a opção **Análise de Dados**;
  - ❑ Escolha a opção **Estatística Descritiva**;
  - ❑ Na caixa **Intervalo de Entrada**, selecione os dados da amostra;
  - ❑ Selecione a opção **Intervalo de Confiança para a Média** e coloque o intervalo de confiança desejado;
  - ❑ Na caixa **Intervalo de Saída**, selecione o local da planilha onde os resultados serão colocados;
  - ❑ Clique em Ok.

monica@mbarros.com

181

## Utilizando o Excel



- ❑ A saída **Erro padrão** fornece o valor de  $\sigma/\sqrt{n}$  para  $n$  grande.
- ❑ Para obter o intervalo de confiança baseado na Normal, calcule  $z_{1-\alpha/2}$  utilizando a função apropriada, multiplique pelo Erro padrão, e faça: média amostral + e - o resultado encontrado.
- ❑ A saída **Intervalo de Confiança** já fornece o valor de  $(t_{1-\alpha/2, n-1})\sigma/\sqrt{n}$  (ou seja, já fornece o que deve ser somado e subtraído da média amostral), bastando apenas subtrair e somar à média.

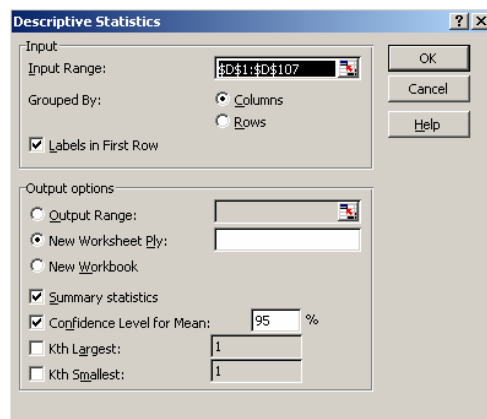
monica@mbarros.com

182

## Utilizando o Excel



- ❑ A seguir aplicamos esta análise para o preço da gasolina em 106 postos do Rio de Janeiro em Agosto de 2002.



monica@mbarros.com

183

## Utilizando o Excel



Gas. Comum	
Média	1.725
Erro Padrão	0.007
Mediana	1.725
Moda	1.749
Desvio Padrão	0.075
Variância Amostral	0.006
Curtose	1.082
Assimetria	0.386
Amplitude (Máx - Mín)	0.410
Mínimo	1.520
Máximo	1.930
Soma	182.847
n	106
IC 95%	0.014

O erro padrão é apenas o desvio padrão dividido por  $\sqrt{n} = \sqrt{106}$

$(t_{0.025})\sigma/\sqrt{n}$  – basta subtrair e somar este valor à média para encontrar o IC 95%

monica@mbarros.com

184

## Utilizando o Excel



- **Nota:**
- Como o tamanho da amostra é grande, poderíamos ter usado um IC baseado na distribuição Normal.
- Na verdade, a diferença praticamente inexistente, pois o número de graus de liberdade da distribuição t neste caso (105) a torna, para todos os efeitos, indistingüível da Normal.

## Forma Alternativa para um IC baseado na distribuição t



- Se definirmos a variância amostral como:

$$S^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

e então  $\frac{(n)S^{*2}}{\sigma^2} \sim \chi_{n-1}^2$

- Daí a variável T torna-se:

$$T = \frac{\sqrt{n}(\bar{X} - \theta)}{\frac{\sigma}{\sqrt{(n)S^{*2}}}} = \sqrt{n-1} \cdot \frac{\bar{X} - \theta}{S^*} \sim t_{n-1}$$

## Forma Alternativa para um IC baseado na distribuição t



- E aí o intervalo de confiança torna-se:

$$IC = \bar{X} \pm t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{S^*}{\sqrt{n-1}} = \left( \bar{X} - t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{S^*}{\sqrt{n-1}}, \bar{X} + t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{S^*}{\sqrt{n-1}} \right)$$

- **Qual intervalo é “melhor”? Nenhum** – são equivalentes, o importante é saber se você está calculando a variância amostral com denominador n ou (n-1), para ser coerente na sua escolha.

## IC para a média de uma distribuição qualquer – GRANDES AMOSTRAS



- Intervalo de confiança aproximado para as médias de distribuição não-normais (**baseado no Teorema Central do Limite**).
- Considere a v.a. X com densidade ou função de probabilidade f(x), não necessariamente Normal.
- Tome uma a.a. de tamanho n desta densidade.

## IC para a média de uma distribuição qualquer – GRANDES AMOSTRAS

- Se  $n$  (o tamanho da amostra) é grande o Teorema Central do Limite estabelece que:

$$S^2 \xrightarrow{P} \sigma^2 \quad \sqrt{n} \frac{(\bar{X} - \theta)}{\sigma} \xrightarrow{d} N(0,1)$$

$$\frac{\sqrt{n}(\bar{X} - \theta) / \sigma}{\sqrt{(n-1)S^2 / (n-1)\sigma^2}} = \sqrt{n} \frac{(\bar{X} - \theta)}{S} \xrightarrow{d} N(0,1)$$

## IC para a média de uma distribuição qualquer – GRANDES AMOSTRAS

- Daí, um intervalo de confiança aproximado para  $\theta$  quando a variância é desconhecida e  $X_i$  é não- Normal é:

$$\left( \bar{X} - z_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}}; \bar{X} + z_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}} \right)$$

onde  $z_{1-\alpha/2}$  é obtido de uma  $N(0,1)$  tal que:

$$\text{Prob} [- z_{1-\alpha/2} < Z < z_{1-\alpha/2} ] = 1 - \alpha \text{ sendo } Z \sim N(0,1)$$

## IC para diferenças entre médias

- Objetivo**
- Comparação das médias de duas amostras aleatórias Normais.
- Exemplos: Agricultura, Medicina, Energia, Veterinária, Marketing, Produção, Finanças, etc...

## IC para diferenças entre médias

- Aplicações - Medicina
- Deseja-se medir o efeito da dieta sobre a pressão sanguínea e a taxa de colesterol de uma pessoa. Toma-se duas amostras “parecidas” de pessoas (mesmas idades, pesos, nível de atividade, etc... ).
- Uma das amostras é submetida a uma dieta com alto teor de gordura e carnes vermelhas.
- O outro grupo ingere uma dieta consistindo principalmente em vegetais, carnes brancas e grãos.

## IC para diferenças entre médias



- ❑ Os pacientes são acompanhados por um período de 3 meses, no qual são feitas medições quinzenais da pressão sanguínea e da taxa de colesterol.
- ❑ Como a dieta afeta estas 2 quantidades? A pressão sanguínea no grupo que ingere mais gordura é significativamente maior que no outro grupo?
- ❑ E a taxa de colesterol?

## IC para diferenças entre médias



- ❑ Aplicações - Veterinária
- ❑ A empresa produtora da ração “Baby Dog” decide lançar no mercado uma nova marca de ração, “”Super Baby Dog”, que supostamente tem maior teor nutritivo.
- ❑ Toma-se uma amostra de 200 cachorrinhos com 2 meses de idade, 100 deles alimentados com “Baby Dog” e 100 alimentados com “Super Baby Dog”.

## IC para diferenças entre médias



- ❑ Ao completarem 6 meses de idade, os cães são novamente examinados e registra-se o aumento de peso no período de 2 a 6 meses de idade.
- ❑ Pergunta-se: a ração “Super Baby Dog” fez os cachorrinhos crescerem mais que a “Baby Dog”? Qual a diferença no aumento de peso médio dos cães submetidos às duas rações?

## IC para diferenças entre médias



- ❑ Aplicações – Marketing
- ❑ A empresa ABC concentra seus anúncios de TV no horário nobre, gastando uma imensa fortuna em publicidade. Como forma de conter as despesas, a companhia decide direcionar seus anúncios para um horário mais tardio, e para programas vistos por um público principalmente das classes A e B. A questão de interesse para a empresa é: esta mudança foi eficaz? Ou seja, será que a empresa economizou dinheiro e ainda manteve o mesmo nível de vendas após a mudança do horário de seus anúncios?

## IC para diferenças entre médias



### □ **Formulação Matemática**

- Considere duas populações Normais com médias ( $\mu_1$  e  $\mu_2$ ) possivelmente distintas e com a **mesma variância (esta hipótese é essencial para resolver o problema!)**. Isto é:

$$X_i \sim N(\mu_1, \sigma^2) \text{ e } Y_j \sim N(\mu_2, \sigma^2)$$

Onde  $i = 1, 2, \dots, m$  e  $j = 1, 2, \dots, n$

## IC para diferenças entre médias



- Considere as duas amostras aleatórias de X e Y com tamanhos m e n respectivamente, isto é:

$$\tilde{X} = (X_1, \dots, X_m); \quad \tilde{Y} = (Y_1, \dots, Y_n)$$

- Suponha que todos os parâmetros ( $\mu_1$ ,  $\mu_2$  e  $\sigma^2$ ) são desconhecidos. Então o nosso objetivo é:

**Achar um intervalo de confiança 100(1- $\alpha$ )% para ( $\mu_1 - \mu_2$ ).**

## IC para diferenças entre médias



- Intuitivamente, este intervalo deverá ser baseado nas respectivas médias amostrais e terá a forma:

$$(\bar{X} - \bar{Y} - c, \bar{X} - \bar{Y} + c)$$

- A questão que devemos responder é: como achar esta constante c?

## IC para diferenças entre médias



**Solução:**

**Sabemos que:**

$$\bar{X} \sim N(\mu_1; \sigma^2 / m); \quad \bar{Y} \sim N(\mu_2; \sigma^2 / n)$$

e estas médias amostrais são independentes. Então qualquer combinação linear de  $\bar{X}$  e  $\bar{Y}$  é Normal e, em particular:

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

## IC para diferenças entre médias



Além disso, temos que:

$$\frac{(m-1)S_1^2}{\sigma^2} \sim \chi_{m-1}^2 \quad \frac{(n-1)S_2^2}{\sigma^2} \sim \chi_{n-1}^2$$

Onde  $S_1^2$  é a variância amostral da 1a. amostra (X's) e  $S_2^2$  a variância amostral dos Y's, ambas independentes.

Daí:

$$\frac{1}{\sigma^2} ((m-1)S_1^2 + (n-1)S_2^2) \sim \chi_{n+m-2}^2$$

## IC para diferenças entre médias



### Revisão:

□ Seja  $Z \sim N(0,1)$  e  $V \sim \chi_p^2$ , ambas independentes.

□ Então:

$$T = Z / \sqrt{V/p} \sim t_p,$$

Tem uma distribuição t de Student com p graus de liberdade

## IC para diferenças entre médias



Combinando os resultados temos:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left( \frac{1}{m} + \frac{1}{n} \right)}} \sim N(0,1)$$

$$V = \frac{1}{\sigma^2} ((m-1)S_1^2 + (n-1)S_2^2) \sim \chi_{n+m-2}^2$$

## IC para diferenças entre médias



Além disso, Z e V são independentes, então a variável T dada por:

$$T = \frac{Z}{\sqrt{\frac{V}{n+m-2}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\left( \frac{1}{n} + \frac{1}{m} \right) \left( \frac{(m-1)S_1^2 + (n-1)S_2^2}{n+m-2} \right)}} \sim t_{n+m-2}$$

Tem distribuição t de Student com (m+n-2) graus de liberdade.

## IC para diferenças entre médias



Dado um nível de significância  $100*(1-\alpha)\%$  podemos achar um número “b” tal que:

$$\text{Prob}\{-b < T < b\} = (1-\alpha)$$

b é obtido a partir da distribuição t com  $n+m-2$  graus de liberdade, onde T é a variável mostrada no “slide” anterior, calculada a partir da diferença entre as médias das duas amostras.

## IC para diferenças entre médias



□ Para simplificar a notação, seja:

$$R = \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \left(\frac{(m-1)S_1^2 + (n-1)S_2^2}{n+m-2}\right)}$$

□ O IC  $100*(1-\alpha)\%$  para a diferença das médias é:

$$\left(\bar{X} - \bar{Y} - bR; \bar{X} - \bar{Y} + bR\right)$$

## IC para diferenças entre médias



- Exemplo
- Estuda-se um certo processo químico com o objetivo de tentar aumentar a produção de um certo composto. Atualmente usa-se na produção um certo tipo de catalisador A, mas um outro tipo de catalisador B é aceitável.
- Faz-se uma experiência com  $n = 8$  tentativas para o catalisador A e o mesmo nº de repetições para o catalisador B.

## IC para diferenças entre médias



□ As médias e variâncias amostrais são:

$$\bar{X} = 91.73, \bar{Y} = 93.75 \text{ e } S_1^2 = 3.89, S_2^2 = 4.02.$$

- Construa um intervalo de confiança 95% para  $\mu_1 - \mu_2$ .
- Solução
- $n = m = 8$

$$R = \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \left(\frac{(m-1)S_1^2 + (n-1)S_2^2}{(n+m-2)}\right)} = \sqrt{\left(\frac{1}{4}\right) \left(\frac{7(3.89) + 7(4.02)}{14}\right)} = 0.989$$

## IC para diferenças entre médias



- $b = 2.145$  da tabela  $t_{14}$ . O intervalo de confiança é:

$$(\bar{X} - \bar{Y}) \pm bR = -2.02 \pm 2.121 = (-4.141, 0.101)$$

- Note que este *intervalo inclui zero*. Isso indica que pode não existir diferença real na produção média usando os catalisadores A e B. Assim, baseado apenas neste teste, parece não haver razão para mudar do catalisador A para o B com o objetivo de aumentar a produção.

## IC para a variância da Normal



- Sejam  $X_1, X_2, \dots, X_n$  iid  $N(\mu, \sigma^2)$  onde ambos  $\mu$  e  $\sigma^2$  são desconhecidos. Este é o caso usual na prática, onde desejamos inferir sobre um dos parâmetros quando ambos são desconhecidos.

- A variância amostral é  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

- Também sabemos que  $nS^2/\sigma^2$  tem distribuição Qui-quadrado com  $n-1$  graus de liberdade.

## IC para a variância da Normal



- Dado  $\alpha \in (0,1)$  ache  $a$  e  $b$  da tabela Qui-quadrado com  $(n - 1)$  graus de liberdade tais que:

- $\Pr(a < (n-1)S^2/\sigma^2 < b) = 1 - \alpha$  e

- $\Pr((n-1)S^2/\sigma^2 < a) = \alpha/2 = \Pr((n-1)S^2/\sigma^2 > b)$

- Logo:  $\Pr[(n-1)S^2/b < \sigma^2 < (n-1)S^2/a] = 1 - \alpha$ .

## IC para a variância da Normal



- O intervalo  $((n-1)S^2/b, (n-1)S^2/a)$  é um intervalo aleatório com probabilidade  $1 - \alpha$  de incluir o parâmetro desconhecido  $\sigma^2$ .

- Exemplo

- Sejam  $X_1, X_2, \dots, X_9$  iid Normais com média  $\mu$  e variância  $\sigma^2$ .

- Observa-se  $s^2 = 7.63$ . Encontre um intervalo de confiança 95% para  $\sigma^2$ .

## IC para a variância da Normal



- ❑ Solução
- ❑ Neste caso precisamos encontrar a e b de uma tabela Qui-quadrado com 8 graus de liberdade.
- ❑ O ponto a tal que a probabilidade de estar abaixo dele é 2.5% é: 2.180
- ❑ O ponto b tal que a probabilidade de estar abaixo dele é 97.5% (ou seja, a probabilidade de estar acima dele é 2.5%) é: 17.535.

## IC para a variância da Normal



- ❑ O intervalo de confiança 95% para a variância da distribuição é:

$$\left( \frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a} \right) = \left( \frac{8(7.63)}{17.535}, \frac{8(7.63)}{2.180} \right) = (3.481, 28.004)$$

## IC para a razão das variâncias



- ❑ A princípio pode parecer estranho encontrar um intervalo de confiança para **a razão** entre as variâncias de duas amostras.
- ❑ Mas, existem resultados distribucionais apropriados para lidar com este problema, enquanto não existem distribuições apropriadas para testar, por exemplo, a diferença entre as variâncias das 2 amostras.

## IC para a razão das variâncias



- ❑ No exemplo do IC para a diferença entre médias foi necessário supor que a variância das duas amostras era igual.
- ❑ Como verificar isso? Podemos fazer um intervalo de confiança para a RAZÃO das variâncias.
- ❑ Se este intervalo incluir 1, existe evidência a favor da igualdade das variâncias. Do contrário, se o intervalo não incluir 1, ficaremos (no mínimo) desconfiados sobre a validade do teste t proposto anteriormente.

## IC para a razão das variâncias



### □ Situação

$$X_i \sim N(\mu_1, \sigma^2) \text{ e } Y_j \sim N(\mu_2, \sigma^2)$$

Onde  $i = 1, 2, \dots, m$  e  $j = 1, 2, \dots, n$

- As variâncias amostrais para as duas amostras são os estimadores de  $\sigma_1^2$  e  $\sigma_2^2$ , dadas por:

$$S_1^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2 \quad \text{e} \quad S_2^2 = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y})^2$$

## IC para a razão das variâncias



- Sabemos também que  $S_1^2$  e  $S_2^2$  são independentes, e múltiplos destas variâncias têm distribuição Qui-quadrado, ou seja:

$$\frac{(m-1)S_1^2}{\sigma^2} \sim \chi_{m-1}^2$$

e

$$\frac{(n-1)S_2^2}{\sigma^2} \sim \chi_{n-1}^2$$

## IC para a razão das variâncias



- Também, estas duas variáveis Qui-quadrado são independentes, o que nos permite usar a definição de uma variável aleatória com distribuição F:

$$F = \frac{\chi_p^2/p}{\chi_q^2/q} = \frac{q}{p} \frac{\chi_p^2}{\chi_q^2} \sim F(p, q)$$

## IC para a razão das variâncias



- Assim, a variável aleatória:

$$F = \frac{\frac{(m-1)S_1^2}{\sigma_1^2}/(m-1)}{\frac{(n-1)S_2^2}{\sigma_2^2}/(n-1)} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}$$

- Tem distribuição F com  $m-1$  graus de liberdade no numerador e  $n-1$  graus no denominador.

## IC para a razão das variâncias



- ❑ Como encontrar um intervalo de confiança  $(1-\alpha)\%$  para a razão de variâncias?
- ❑ Dado  $\alpha \in (0,1)$ , ache  $\underline{a}$  e  $\underline{b}$  tais que:  $\Pr(a < F < b) = 1-\alpha$  e  $F \sim F(n-1, m-1)$
- ❑ Por convenção escolhemos  $\underline{a}$  e  $\underline{b}$  tais que:
- ❑  $\Pr(F \leq a) = \alpha/2$ ,  $\Pr(F \geq b) = \alpha/2 \Rightarrow \Pr(F < b) = 1-\alpha/2$ ,
- ❑ e este valor é encontrado a partir de uma tabela da função de distribuição F.

## IC para a razão das variâncias



- ❑ Frequentemente  $\alpha$  é um valor pequeno, e não existe na tabela, e daí temos que usar um truque, que decorre da maneira como uma variável F é criada.
- ❑ Lembre-se que se  $F \sim F(p,q)$ , F é a razão de 2 variáveis aleatórias Qui quadrado independentes, divididas pelos seus graus de liberdade.

## IC para a razão das variâncias



- ❑ Logo, se  $F \sim F(p,q)$  então  $F = (V_1/p)/(V_2/q) = qV_1/pV_2$  onde  $V_1$  e  $V_2$  são independentes. Então  $W = 1/F = (pV_2)/(qV_1) = (V_2/q)/(V_1/p)$  tem densidade  $F(q,p)$ .
- ❑ Logo:

$$\Pr(F \leq a) = \Pr\left(\frac{1}{F} \geq \frac{1}{a}\right) = 1 - \Pr\left[\frac{1}{F} \leq \frac{1}{a}\right] = \frac{\alpha}{2}$$

## IC para a razão das variâncias



- ❑ Também, os seguintes eventos são equivalentes:

$$a < F < b \Leftrightarrow a < \frac{S_1^2 \sigma_2^2}{\sigma_1^2 S_2^2} < b$$
$$\Leftrightarrow a \frac{S_2^2}{S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < b \frac{S_2^2}{S_1^2}$$

- ❑ Logo, o intervalo:  $\left(a \frac{S_2^2}{S_1^2}, b \frac{S_2^2}{S_1^2}\right)$
- ❑ é um intervalo aleatório com probabilidade  $1-\alpha$  de incluir o valor desconhecido  $\sigma_2^2/\sigma_1^2$

## IC para a razão das variâncias



- Exemplo
- Considere duas amostras Normais tais que  $m = 10$  (tamanho da 1a. amostra),  $n = 5$  (tamanho da 2a. amostra),  $S_1^2 = 20$  e  $S_2^2 = 35.6$ .
- Encontre um intervalo de confiança 95% para a razão de variâncias.

$$\frac{S_2^2}{S_1^2} = \frac{(35.6)}{(20)} = 1.78$$

## IC para a razão das variâncias



- Precisamos achar  $a$  e  $b$  tais que:
- Se  $F \sim F(m-1, n-1) = F(9, 4)$  então  $\Pr(F \leq a) = \alpha/2 = 0.025$  e  $\Pr(F \geq b) = \alpha/2 = 0.025$ .
- Logo:  $\Pr(F \leq b) = 0.0975 \Rightarrow b = 8.90$ .
- E:  $\Pr(F \leq a) = 0.025 \Leftrightarrow \Pr(F > a) = 0.975$

$$\Leftrightarrow \Pr\left(\frac{1}{F} < \frac{1}{a}\right) = 0.975 \quad \text{onde} \quad \frac{1}{F} \sim F(4, 9)$$

- Então, olhando para a tabela  $F(4, 9)$  segue que:

## IC para a razão das variâncias



$$\frac{1}{a} = 4.72 \Rightarrow a = \frac{1}{4.72}$$

- O intervalo de confiança 95% para  $\sigma_2^2/\sigma_1^2$  é:

$$(1.78a, 1.78b) = \left(\frac{1.78}{4.72}, 1.78(8.90)\right) = (0.376, 15.842)$$

## Exemplo (para casa)



- Toma-se duas amostras de marcas de pneus para testar a sua durabilidade média (em milhares de km). Os resultados estão a seguir.

	Marca 1	Marca 2
Tamanho da amostra	15	13
Durabilidade média do pneu (em mil km)	50	45
Desvio padrão da durabilidade média (em mil km)	9	13

- Encontre um intervalo de confiança 95% para  $\mu_1 - \mu_2$  onde  $\mu_1$  é a durabilidade média dos pneus da marca 1 e  $\mu_2$  é a mesma coisa para a marca 2.
- Com 95% de probabilidade existe a chance de  $\mu_1$  e  $\mu_2$  serem iguais? Por que?
- Encontre um intervalo de confiança 95% para a razão das variâncias. As variâncias das duas amostras podem ser iguais com este grau de confiança?

## IC aproximado para a proporção de uma Binomial



- Seja  $Y \sim \text{Bin}(n,p)$  onde  $n$  é conhecido e  $0 < p < 1$  é desconhecido.

- Assim,  $E(Y) = np$ ,  $\text{VAR}(Y) = np(1-p)$ , e  $\hat{p} = \frac{Y}{n}$  é o estimador de máxima verossimilhança para  $p$ .

- Pelo Teorema Central do Limite:

$$\frac{Y - np}{\sqrt{np(1-p)}} \underset{\text{aprox}}{\sim} N(0,1) \quad \text{se } n \text{ é grande.}$$

## IC aproximado para a proporção de uma Binomial



- Mas, precisamos de uma estimativa do desvio padrão de  $Y$  para calcular o intervalo de confiança para  $\mu = E(Y) = np$ , e então substituímos  $p$  no denominador pelo seu estimador de máxima verossimilhança.

- Ou seja, um *intervalo de confiança*  $1-\alpha$  aproximado para  $p$  é:

$$\left( \hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

## IC aproximado para a proporção de uma Binomial



Este intervalo foi obtido da seguinte maneira:

$$\frac{Y - np}{\sqrt{np(1-p)}} \underset{\text{aprox}}{\sim} N(0,1)$$

- Dividindo o numerador e o denominador acima por  $n$  leva a:

$$Z = \frac{(Y/n) - p}{\frac{1}{n} \sqrt{np(1-p)}} = \frac{(Y/n) - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

## IC aproximado para a proporção de uma Binomial



- E como  $Z$  definido acima é aproximadamente  $N(0,1)$  então:

$$\Pr[-z_{1-\alpha/2} < Z < z_{1-\alpha/2}] = 1-\alpha$$

e obtemos o intervalo indicado.

## IC aproximado para a proporção de uma Binomial



- Exemplo
- Uma pesquisa do governo afirma que 10% dos homens com idade inferior a 25 anos estão desempregados.
- Encontre a probabilidade de que, ao tomarmos uma amostra de 400 homens com menos de 25 anos, a proporção estimada de desempregados seja superior a 12%.

## IC aproximado para a proporção de uma Binomial



- Solução
- A probabilidade real (segundo o governo) de um homem desta faixa etária estar desempregado é  $p = 10\%$ .
- Toma-se uma amostra de tamanho 400 e estima-se  $p$  a partir desta amostra. Podemos utilizar o Teorema Central do Limite e encontramos:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \sqrt{n} \frac{\hat{p} - p}{\sqrt{p(1-p)}} \approx \sqrt{n} \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}} \text{ é aproximadamente } N(0,1)$$

## IC aproximado para a proporção de uma Binomial



- A probabilidade desejada é:

$$\begin{aligned} \Pr(\hat{p} > 0.12) &= \Pr\left(\sqrt{\frac{400}{(1/10)(9/10)}}(\hat{p} - 0.10) > \sqrt{\frac{400}{(1/10)(9/10)}}(0.12 - 0.10)\right) = \\ &= \Pr\left(\left(\frac{200}{3}\right)(\hat{p} - 0.10) > \left(\frac{200}{3}\right)(0.02)\right) = \Pr\left(Z > \frac{4}{3}\right) = \Pr(Z > 1.33) = 0.0918 \end{aligned}$$

- Logo, existe uma probabilidade de cerca de 9% de que a estimativa amostral ultrapasse 12%, mesmo que o valor real seja 10%.

## IC aproximado para a proporção de uma Binomial



- Exemplo
- Considere novamente a situação do exemplo anterior.
- Suponha que a probabilidade de um homem com menos de 25 estar desempregado é desconhecida, e será estimada a partir de uma amostra de 400 homens.
- Suponha que observamos  $\hat{p} = 0.12$ . Encontre um intervalo de confiança 90% aproximado para  $p$ .

## IC aproximado para a proporção de uma Binomial



- Solução
- Pelo exemplo anterior:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \sqrt{n} \frac{\hat{p} - p}{\sqrt{p(1-p)}} \approx \sqrt{n} \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}} = \frac{\sqrt{400}}{\sqrt{(0.12)(0.88)}} (\hat{p} - p) = 61.546(\hat{p} - p)$$

- É aproximadamente  $N(0,1)$ . Usando a tabela da Normal leva a:

$$\Pr(-1.645 < Z < +1.645) = 0.90 \Rightarrow \Pr(-1.645 < 61.546(\hat{p} - p) < +1.645) = 0.90$$

## IC aproximado para a proporção de uma Binomial



- Logo:

$$\begin{aligned} \Rightarrow \Pr\left(\hat{p} - \frac{1.645}{61.546} < p < \hat{p} + \frac{1.645}{61.546}\right) &= \Pr\left(0.12 - \frac{1.645}{61.546} < p < 0.12 + \frac{1.645}{61.546}\right) = \\ &= \Pr(9.33\% < p < 14.67\%) \end{aligned}$$

- Ou seja, nestas condições há 90% de probabilidade da taxa de desemprego real estar entre 9.33% e 14.67%.