Distribuições e Teorema do Limite Central

Distribuições:
Normal,
Binomial,
Log-normal,
Uniforme

Roteiro

- Distribuição empírica de freqüência
- Distribuições
- Teorema do Limite Central

Teorema do Limite Central

- "Se tomarmos amostras grandes de uma população, as médias amostrais terão distribuição Normal mesmo que os dados originais não tenham distribuição Normal."

Simulações:
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Distribuição Empírica de Freqüência

- Dados observados / histograma

Distribuições de Probabilidades

- Princípio teórico:

" Existe uma função que governa a probabilidade de obtermos determinados valores na observação de uma grandeza " (Helene e Vanin)

⇩

Função (Densidade) de Probabilidade (fdp)

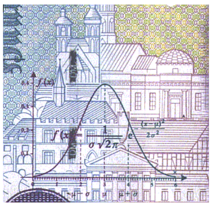
- " Podemos entender uma distribuição de probabilidades como um equivalente teórico de uma distribuição empírica de freqüências " (Petrie e Watson)

Variáveis aleatórias

- variável aleatória: pode assumir diferentes valores, cada qual com uma dada probabilidade
- quantitativas (discreta/contínua)
- variável aleatória binária (0 ou 1, positivo ou negativo)

Distribuição Normal (ou Gaussiana)

Em homenagem a C. F. Gauss, matemático alemão do séc. XVIII



Nota de 10 marcos alemães

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

fdp Normal ou Gaussiana

- " função de variável contínua que parece ajustar muitas das funções densidade de probabilidade observadas (...) também em situações e objetos do dia-a-dia, tais como o tamanho de pregos fabricados por uma máquina ou a massa de pães que, presumivelmente, seriam sempre iguais" (Helene e Vanin)

Características da Normal

- descrita por 2 parâmetros:
  - média μ, desvio padrão σ
- unimodal
- simétrica em torno da média ("forma de sino")
- média = mediana = moda

Distribuição Normal Padrão (ou Reduzida)

- média=0 e desvio padrão=1

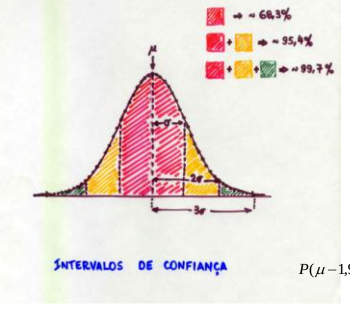
$$z=\frac{x-\mu}{\sigma}$$

$$x=\mu+\sigma \Rightarrow z=1$$

$$x=\mu+2\sigma \Rightarrow z=2$$

$$x=\mu+3\sigma \Rightarrow z=3$$

Intervalos de confiança



$$P(\mu-1{,}96\sigma<x<\mu+1{,}96\sigma)=95\%$$

$$z=1{,}96$$

Distribuição Normal

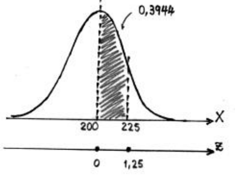
Qual a probabilidade de um indivíduo apresentar nível de colesterol com valor entre 200 e 225mg/100ml?

x: nível de colesterol no plasma

Valores hipotéticos:

$\mu=200\,mg/100ml$

$\sigma=20\,mg/100ml$



$x=225$

$$z=\frac{x-\mu}{\sigma}=\frac{225-200}{20}=1{,}25$$

$$P(200<x<225)=P(0<z<1{,}25)=0{,}3944 \text{ ou } 39{,}44\%$$

Distribuição Normal

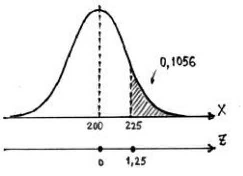
Qual a probabilidade de um indivíduo apresentar nível de colesterol superior a 225mg/100ml?

x: nível de colesterol no plasma

Valores hipotéticos:

$\mu = 200\, mg/100\, ml$

$\sigma = 20\, mg/100\, ml$



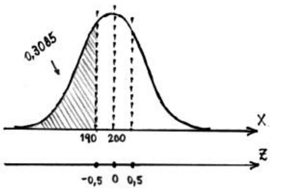
$x = 225$

$$z = \frac{x-\mu}{\sigma} = \frac{225-200}{20} = 1{,}25$$

$$P(x>225) = P(z>1{,}25) = 0{,}1056 \text{ ou } 10{,}56\%$$

Distribuição Normal

Qual a probabilidade de um indivíduo apresentar nível de colesterol inferior a 190 mg/100ml?



$x = 190$

$$z = \frac{x-\mu}{\sigma} = \frac{190-200}{20} = -0{,}5$$

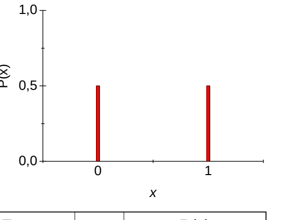
$$P(x<190) = P(z<-0{,}5) = 0{,}3085 \text{ ou } 30{,}85\%$$

Nos exemplos da distribuição Normal, vimos que, a partir da função densidade de probabilidade, podemos calcular a probabilidade de obter um valor em um determinado intervalo.

Distribuição binomial

- Distribuição de valores discretos mais conhecida
- Surge quando observamos um conjunto de *n* variáveis aleatórias <u>binárias</u> independentes

Distribuição binomial (*n*=1, *p*=1/2)



Evento	*x*	P(*x*)
F	0	*q*=1/2
M	1	*p*=1/2

x: variável aleatória que representa o número de bezerros nascidos do sexo masculino (M) em *n* nascidos

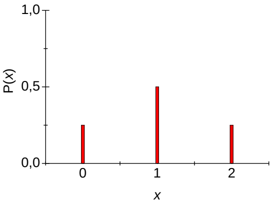
p: probabilidade de nascer um bezerro macho

q=1-p: probabilidade de nascer fêmea

n=1

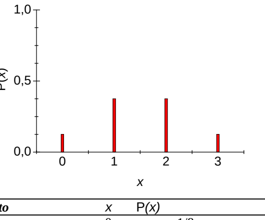
p=1/2

Distribuição binomial (*n*=2, *p*=1/2)



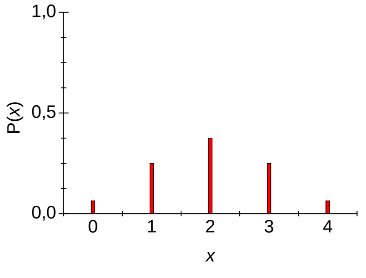
Evento	*x*	P(*x*)	
FF	0	q.q=1/4	
FM	1	q.p	2q.p=1/2
MF	1	p.q	
MM	2	p.p=1/4	

Distribuição binomial (*n*=3, *p*=1/2)



Evento	*x*	P(*x*)	
FFF	0	q.q.q = 1/8	
FFM	1	q.q.p	$3q^2p = 3/8$
FMF	1	q.p.q	
MFF	1	p.q.q	
FMM	2	q.p.p	$3qp^2 = 3/8$
MFM	2	p.q.p	
MMF	2	p.p.q	
MMM	3	p.p.p=1/8	

Distribuição binomial (n=4, p=1/2)



Distribuição binomial

- *n*: número de observações
- *x*: número de eventos de um certo tipo ("sucesso")
- *p*: probabilidade de ocorrência do evento que nos interessa

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

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

- média: $\mu = n\,p$
- variância: $\sigma^2 = n\,p\,q$

Aproximação da binomial pela Normal

- A distribuição binomial se aproxima de uma distribuição Normal quando...

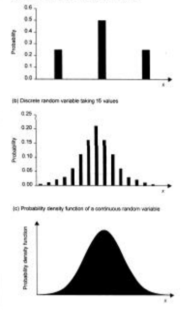
Critérios apresentados por diferentes autores:

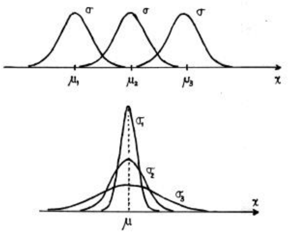
$np > 5$ e $nq > 5$ (Vieira, 1988)

$np > 15$ e $nq > 15$ (Johnson e Bhattacharyya, 1996)

$npq > 3$ (Noether, 1991)

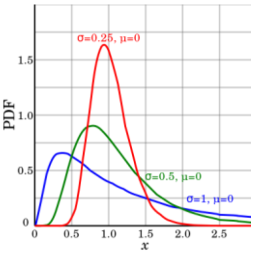
Distribuição Normal





Distribuição Log-Normal

Distribuição cujo logaritmo possui distribuição normal



Distribuição Log-Normal

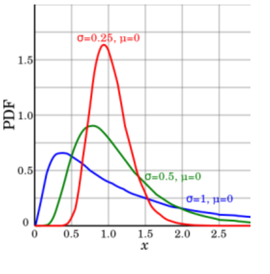
Distribuição cujo logaritmo possui distribuição normal

Se:

$\ln(X) \sim \mathcal{N}(\mu, \sigma^2).$

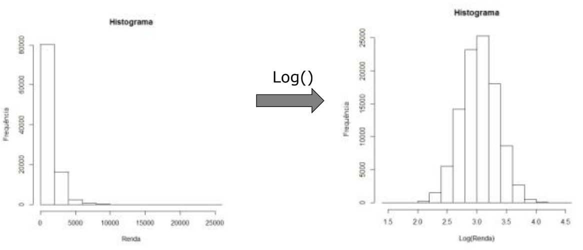
Fácil deduzir que:

$$\begin{aligned} f_X(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\Pr(X \le x) = \frac{\mathrm{d}}{\mathrm{d}x}\Pr(\ln X \le \ln x) \\ &= \frac{\mathrm{d}}{\mathrm{d}x}\Phi\left(\frac{\ln x - \mu}{\sigma}\right) \\ &= \varphi\left(\frac{\ln x - \mu}{\sigma}\right)\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\ln x - \mu}{\sigma}\right) \\ &= \varphi\left(\frac{\ln x - \mu}{\sigma}\right)\frac{1}{\sigma x} \\ &= \frac{1}{x}\cdot\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right). \end{aligned}$$



Distribuição Log-Normal

Distribuição cujo logaritmo possui distribuição normal

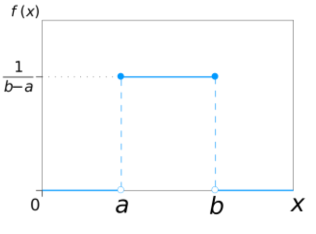


Características da Log-Normal

- descrita por 2 parâmetros:
  - Média da normal correspondente μ,
  - Desvio padrão da normal correspondente σ
- unimodal
- assimétrica
- média > mediana > moda

- Muito comum (não tanto quanto normal)

Distribuição Uniforme



$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

Características da Uniforme

- descrita por 2 parâmetros:
  - Mínimo *a*, máximo *b*

- Todos os valores no intervalo são igualmente prováveis
- simétrica em torno da média
- média = mediana
- pouco comum nas ciências biológicas

Extra: Distribuição Poisson

Distribuição de Poisson

- se *p* é muito pequeno (evento raro) e *n* (número de observações) tende para infinito, a distribuição binomial se aproxima de uma distribuição de Poisson.
  - média e variância: $\lambda = n\,p$

$$P(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

$e = 2{,}7182818\ldots$ é conhecido como algarismo neperiano

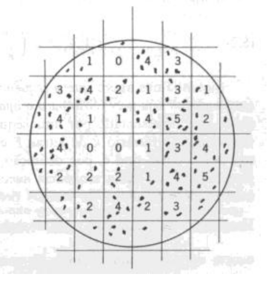
Distribuição de Poisson

- exemplos:
  - desintegração radioativa
  - lançamento de bombas sobre Londres
  - ligações erradas
  - contagem de bactérias em placa de Petri

Fonte: W. Feller, *Introdução à Teoria das Probabilidades e Suas Aplicações*, Edgar Blücher, 1976.

Distribuição de Poisson

- ex. Bactérias em uma placa de Petri



Bactérias em uma placa de Petri

x: número de bactérias (colônias) em cada quadrante

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**
número observado	5	19	26	26	21	13	8
valor teórico	6,3	18,4	27,0	26,4	19,4	11,4	5,5

número de quadrantes observados:

5+19+26+26+21+13+8 = 118

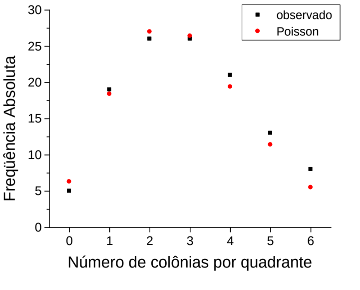
número de colônias observadas:

0x5 + 1x19 + 2x26 + 3x26 + 4x21 + 5x13 + 6x8 = 346

número médio de colônias por quadrante:

346/118 = 2,9322

Bactérias em uma placa de Petri



Teorema do Limite Central

- "Se tomarmos amostras grandes de uma população, as médias amostrais terão distribuição Normal mesmo que os dados originais não tenham distribuição Normal."

Simulações:

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Teorema do Limite Central

- Média das médias:

$$\bar{\bar{x}} = \bar{x}_m = \mu$$

μ = média da população
σ = desvio padrão da população
n = tamanho da amostra

- Desvio padrão das médias (Erro padrão):

$$s_m = \frac{\sigma}{\sqrt{n}}$$

Simulações:

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Estimativas de uma média populacional

Da mesma maneira que podemos calcular as médias amostrais possíveis de uma população,

Podemos calcular as populações possíveis que geraram uma média amostral

Da normal: $P(\mu - 1{,}96\,\sigma < x < \mu + 1{,}96\,\sigma) = 95\%$

No TLC: $P(\mu - 1{,}96 * \sigma_m < \bar{x} < \mu + 1{,}96 * \sigma_m) = 95\%$

$$P(\bar{x} - 1{,}96 * \sigma_m < \mu < \bar{x} + 1{,}96 * \sigma_m) = 95\%$$

Estimativas de uma média populacional

$$P(\bar{x} - 1{,}96 * \sigma_m < \mu < \bar{x} + 1{,}96 * \sigma_m) = 95\%$$

Precisão será dada pelo erro: $\varepsilon = 1{,}96 * s_m$

O tamanho do intervalo de confiança (IC), ou do erro, depende de:

$$s_m = \frac{\sigma}{\sqrt{n}}$$

Portanto:
Quanto maior o n (tamanho da amostra), menor o IC
Quanto maior o desvio-padrão da população, maior o IC

Estimativas de uma média populacional

Cálculo do tamanho de amostra

$$\varepsilon = 1{,}96 * s_m$$

$$\varepsilon = 1{,}96 * \frac{\sigma}{\sqrt{n}} \Longrightarrow n = 1{,}96^2 * \frac{\sigma^2}{\varepsilon^2}$$

Estimativas de uma média populacional

Próximo slide
só para quem tem estômago forte

Estimativas de uma média populacional

Da normal: $P(\mu - 1{,}96\,\sigma < x < \mu + 1{,}96\,\sigma) = 95\%$

No TLC: $P(\mu - 1{,}96 * \sigma_m < \bar{x} < \mu + 1{,}96 * \sigma_m) = 95\%$

Adicionando (-μ-média) em cada parte:

$$P(\mu - \mu - \bar{x} - 1{,}96 * \sigma_m < \bar{x} - \mu - \bar{x} < \mu - \mu - \bar{x} + 1{,}96 * \sigma_m) = 95\%$$

$$P(-\bar{x} - 1{,}96 * \sigma_m < -\mu < -\bar{x} + 1{,}96 * \sigma_m) = 95\%$$

* (-1):

$$P(\bar{x} + 1{,}96 * \sigma_m > \mu > \bar{x} - 1{,}96 * \sigma_m) = 95\%$$

$$P(\bar{x} - 1{,}96 * \sigma_m < \mu < \bar{x} + 1{,}96 * \sigma_m) = 95\%$$

Recompensa