

Modelos Matemáticos e Aplicações (18/19)

Módulo I – Revisão dos fundamentos de Probabilidade e Estatística
com apoio do R – Aulas 3 e 4

Manuela Neves

26 & 28 Fev 2019

- 1 Probabilidade e Inferência
- 2 Em jeito de resumo...
- 3 Introdução à Teoria da Amostragem
- 4 Teoria da Estimação—introdução
- 5 Estimação pontual: estimador e estimativa
- 6 Propriedades dos estimadores
- 7 Métodos de Estimação
 - O método dos momentos
 - O Método da Máxima Verosimilhança
 - Propriedades dos estimadores de Máxima Verosimilhança
 - A máxima verosimilhança no 
- 8 Estimadores mais usuais e distribuições de amostragem –resumo

Probabilidade e Inferência

Pode dizer-se que a **Probabilidade e a Inferência** têm objectivos diferentes: enquanto na Probabilidade se parte de um dado esquema ou modelo para calcular a probabilidade de certos resultados ou acontecimentos se realizarem; na Inferência parte-se de dados ou observações e procura saber-se ou inferir-se algo sobre o modelo.

A Inferência é a “passagem do particular ao geral.”

A Inferência Estatística tem como objectivo **definir procedimentos** que, aplicados a uma amostra extraída da população, **nos permitam estimar parâmetros desconhecidos dessa população ou algo sobre o modelo da população.**

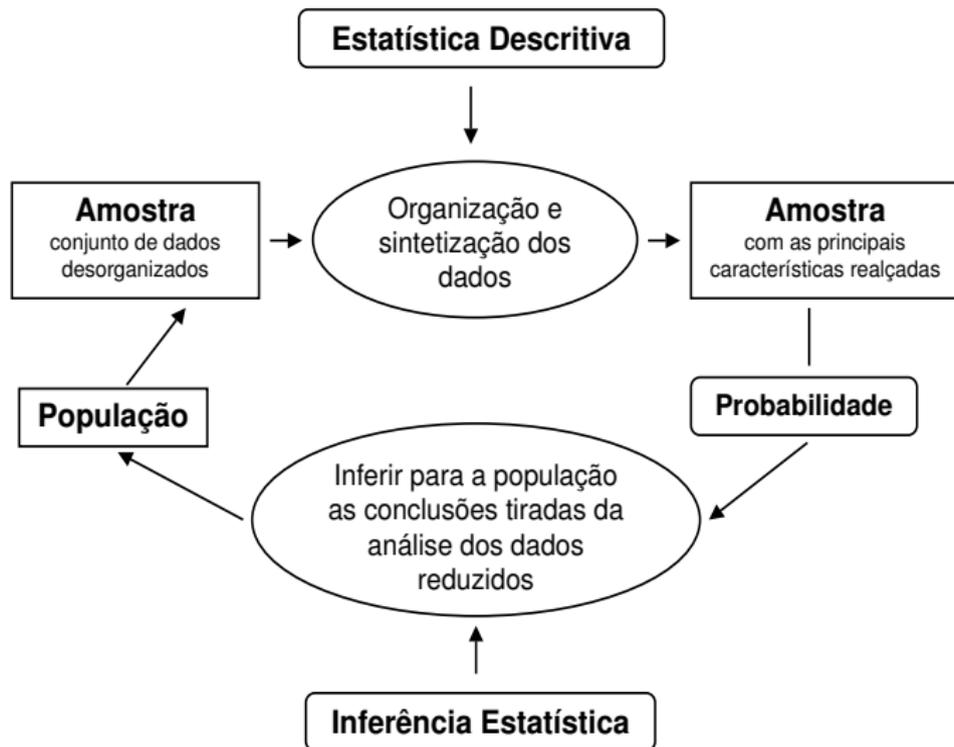
De facto uma amostra particular é apenas uma das muitas amostras (em n° infinito se a população for infinita) que se podem obter por um **processo de amostragem.**

A extensão do particular ao geral chama-se **inferência indutiva**. É o caminho para a aquisição de novos conhecimentos.

O grau de incerteza que acompanha as inferências indutivas pode ser medido rigorosamente em termos de probabilidade, se a experiência foi conduzida segundo determinados princípios (probabilísticos ou aleatórios).

Os procedimentos que levam à obtenção de amostras são do domínio da **Teoria da Amostragem**.

Enquanto a **Amostragem** se refere, regra geral, a procedimentos de recolha de dados representativos de populações finitas, o **Planeamento de Experiências** refere-se à “produção” de dados, quando se controla a variação de algumas variáveis, ficando uma ou mais livres



Seja X a **variável de interesse** numa população em estudo.

Na **teoria da amostragem** usam-se procedimentos de recolha de elementos da população para obter uma **amostra**.

O objectivo dessa amostra é usá-la para inferir sobre propriedades da variável em estudo. É preocupação da teoria da amostragem:

- a recolha de amostras representativas da população;
- a determinação da dimensão da amostra (para a qual é crucial **estimar a variabilidade da característica**) que permita obter estimativas de parâmetros de interesse com uma dada precisão.

Antes de referirmos alguns **conceitos preliminares** em amostragem, vejamos um exemplo:

Exemplo 1

Vamos supor que temos um campo rectangular de $5 \text{ km} \times 4 \text{ km}$ e pretendemos escolher ao acaso áreas de $100 \text{ m} \times 100 \text{ m}$. Podemos definir um par de coordenadas (i, j) , com $i = 1, 2, \dots, 50$ e $j = 1, 2, \dots, 40$. Então a escolha aleatória de, por exemplo, 10 parcelas **podia ser assim feita**:

```
> ndim<-10; nlinhas<-seq(1:50);ncol<-seq(1:40)
> i<-sample(nlinhas,ndim);j<-sample(ncol,ndim)
> parcelas<-cbind(i,j);parcelas
```

```
      i  j
[1,] 27 34
[2,] 14  3
[3,] 36  4
...
```

Definição 1

As variáveis aleatórias X_1, X_2, \dots, X_n constituem uma amostra aleatória de dimensão n , que se costuma representar por $\underline{X} = (X_1, X_2, \dots, X_n)$, retirada de uma população X , se são mutuamente independentes e têm a mesma distribuição que X

Seja (x_1, x_2, \dots, x_n) uma amostra de n observações da característica, obtidas após um processo de amostragem.

Cada um daqueles valores é uma realização de n variáveis (X_1, X_2, \dots, X_n) que são “réplicas” da variável X

A **Inferência Estatística** pretende responder a dois grandes problemas:

- calcular valores aproximados (**estimativas**) e obter **intervalos de confiança** para parâmetros desconhecidos da população.
- formular hipóteses e verificar se há concordância entre o que se supõe e os factos – **testes de hipóteses**

O primeiro problema é do domínio da **Teoria da Estimação**. Se;

- o valor desconhecido do parâmetro θ , usando o conhecimento, mesmo aproximado, da distribuição da população (**estimação paramétrica**);
- a função distribuição desconhecida da variável em estudo, F ou parâmetros sem o pressuposto do conhecimento de um modelo para a população (**estimação não paramétrica**).

Estimação pontual: estimador e estimativa

Dada uma amostra aleatória (X_1, X_2, \dots, X_n) chama-se **estatística** a uma função da amostra aleatória que não envolve parâmetros desconhecidos.

Definição

Chama-se **estimador** de θ a uma estatística que a cada amostra observada faz corresponder um valor que estima θ , a que chamamos uma **estimativa** de θ .

$\Theta^*(X_1, X_2, \dots, X_n)$ é um estimador
 $\theta^*(x_1, x_2, \dots, x_n)$ é uma estimativa

Não confundir

- estimador– variável aleatória
- estimativa– valor aproximado do parâmetro, obtido pelo estimador usando uma amostra particular

Estimação pontual: estimador e estimativa

Exemplo: São estimadores

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

Observada, por exemplo, a amostra **(1, 2, 0, 3, 1, 5)**

As estimativas associadas àqueles estimadores são:

$$\text{i) } \bar{x} = \frac{1}{6} \sum_{i=1}^6 x_i = 2 \quad \text{e} \quad \text{ii) } s^2 = \frac{1}{5} \sum_{i=1}^6 (x_i - \bar{x})^2 = 3.2$$

Consideremos o caso de termos uma dada população com distribuição $F(x|\theta)$, com um **parâmetro desconhecido θ** .

Como se podem definir vários estimadores de um parâmetro, põe-se o problema de escolher, se possível “o melhor”. Há então que considerar certas **propriedades que um estimador deve verificar**.

Vejamos algumas:

1. Consistência

Um estimador Θ^* diz-se convergente ou consistente para o parâmetro θ se $\Theta^* \xrightarrow{P} \theta$.

Prova-se que é condição suficiente de convergência de um estimador Θ^* para θ que

$$\text{Var}(\Theta^*) \rightarrow 0 \quad \text{e} \quad E(\Theta^*) \rightarrow \theta$$

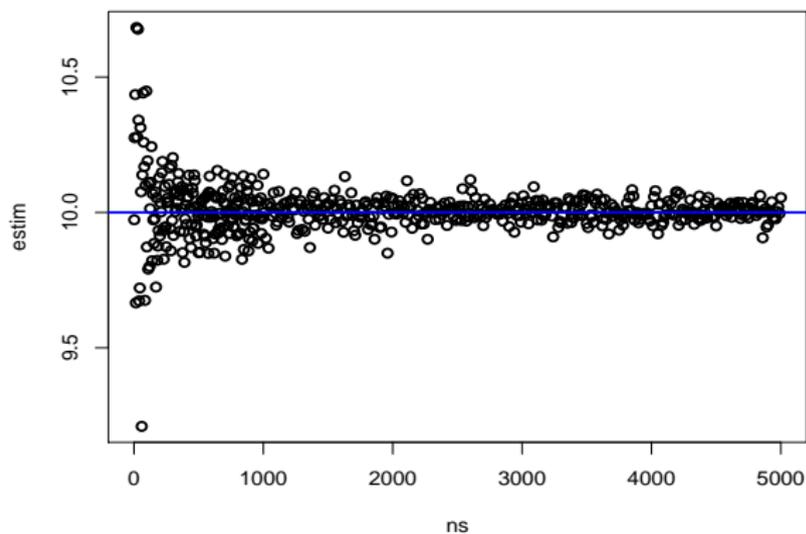
Vejamos o seguinte exemplo ilustrando a convergência de um estimador.

Exemplo

Consideremos uma população $\mathcal{N}(10, 2)$ e vejamos, utilizando o , o comportamento da **média amostral** para várias dimensões de amostras, $n = 2, 5, 10, 15, 20, \dots, 1000, 1010, 1020, \dots, 5000$

```
> ns <- c(2, seq(5, 1000, by = 5), seq(1010, 5000, by = 10))
> estim <- numeric(length(ns)) # tb podia fazer-se estim<-c()
> for (i in 1:length(ns)) {
+ amostra <- rnorm(ns[i], 10, 2)
+ estim[i] <- mean(amostra)
+ }
> plot(ns, estim, lwd=2)
> abline(h = 10, lwd=2, col=4)
```

Ilustrando a convergência



2. Não enviesamento

O estimador Θ^* diz-se centrado ou não enviesado se $E(\Theta^*) = \theta$.

Na verdade é uma propriedade que muitos estimadores não possuem e daí ter mais interesse definir **uma medida da diferença entre o estimador e o parâmetro** que ele pretende estimar.

É o chamado **viés** (ou em inglês '**bias**') que representaremos por $b_\theta(\Theta^*)$ e se define como

$$b_\theta(\Theta^*) = E(\Theta^*) - \theta$$

3. Erro quadrático médio

Chama-se **erro quadrático médio** do estimador Θ^* a **$EQM(\Theta^*) = E[(\Theta^* - \theta)^2]$** .

Esta propriedade é um dos critérios mais usados para comparar estimadores.

É muito fácil mostrar que

$$EQM(\Theta^*) = Var[\Theta^*] + [b_{\theta}(\Theta^*)]^2$$

Se Θ^* é um **estimador centrado** de um parâmetro então o **erro quadrático médio** \equiv **variância**, i.e.

$$E[(\Theta^* - \theta)^2] = Var(\Theta^*)$$

Ilustração da precisão e da exactidão

O termo **exactidão (accuracy)** refere-se à proximidade de uma medição ou estimativa ao verdadeiro valor.

O termo **precisão ou variância (precision)** refere-se ao “grau de concordância numa sucessão de medições”.



Accuracy & Precision



Accurate
but , not precise



Precise
but , not accurate



Accurate
and **Precise**

Situação desejável: estimador centrado com variância mínima.

Diz-se que temos o **estimador mais eficiente**.

Um estimador de um parâmetro θ é o mais eficiente se tiver a menor probabilidade de se afastar de θ , i.e, com dispersão mínima.

Se não conseguirmos estimadores centrados interessa-nos procurar **estimadores que tenham um EQM mínimo**, i.e.,

$$\forall \tilde{\Theta} \quad E[(\Theta^* - \theta)^2] \leq E[(\tilde{\Theta} - \theta)^2] \quad \forall \theta$$

Existe um critério estabelecido na chamada **desigualdade de Fréchet–Cramér–Rao** que, sob certas condições, permite obter um limite inferior para o EQM de um estimador (para mais informações ver Murteira e Antunes (2012) e Casella e Berger (2002))

Até aqui falámos em **estimadores** e nas propriedades que devem possuir. Interessa ter procedimentos que construam estimadores com boas propriedades.

Vamos então falar dos **principais métodos de estimação paramétrica**.

Dos **métodos de estimação paramétrica** vamos referir:

- o **Método dos momentos** e
- o **Método da Máxima verosimilhança**

O método dos momentos

Introduzido por Karl Pearson no início do século XX, foi o primeiro método de estimação a ser apresentado e que tem uma filosofia muito simples.

O método consiste em:

– considerar como estimadores dos parâmetros desconhecidos as soluções das equações que se obtêm igualando os momentos teóricos aos momentos empíricos.

É um método de aplicação geral, tendo como única condição que a distribuição tenha um número suficiente de momentos (teóricos).

O método dos momentos

Sejam $\theta_1, \dots, \theta_k$ parâmetros desconhecidos de uma v.a. X .
O método dos momentos consiste em igualar momentos teóricos e momentos empíricos, i.e.,

$$\begin{aligned} E[X] &= m'_1 & \text{c/} & & m'_1 &= \frac{1}{n} \sum_{i=1}^n x_i \\ E[X^2] &= m'_2 & \text{c/} & & m'_2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 \\ &\vdots & & & & \\ E[X^k] &= m'_k & \text{c/} & & m'_k &= \frac{1}{n} \sum_{i=1}^n x_i^k \end{aligned}$$

$m'_k = \frac{1}{n} \sum_{i=1}^n x_i^k$ são os chamados **momentos empíricos**, calculados à custa da amostra (x_1, \dots, x_n) .

Aquelas igualdades dão-nos **estimativas** que são concretização dos **estimadores** com as expressões correspondentes.

O método dos momentos— exemplos

Exemplo

Consideremos $X \sim \mathcal{N}(\mu, \sigma)$. Quais os **estimadores** de momentos de μ e σ^2 ?

Tem-se :

$$\begin{aligned} E[X] &= \mu & \text{e} & \quad M'_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \\ E[X^2] &= \sigma^2 + \mu^2 & \text{e} & \quad M'_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{aligned}$$

donde

$$\begin{aligned} \mu^* &= \bar{X} \\ (\sigma^2)^* &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \mu^{*2} \Rightarrow (\sigma^2)^* = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{aligned}$$

Exercício

Considere $X \sim \mathcal{P}(\lambda)$. Determine o **estimador** de momentos de λ .

O método dos momentos

Os cálculos não são complicados, mas no cálculo dos momentos empíricos aparecem potências de expoente elevado quando há muitos parâmetros, conduzindo a estimativas instáveis.

Por isso, **como regra prática deve evitar-se recorrer ao método dos momentos para mais de quatro parâmetros.**

Nota: Os estimadores obtidos pelo método dos momentos são menos eficientes do que os estimadores de máxima verosimilhança, que passamos já a apresentar.

Há uma função no  – `uniroot()` que pesquisa num dado intervalo o “zero” de uma função, relativamente ao seu 1^o argumento.

A Verosimilhança

Seja X uma v.a. cuja distribuição depende de um **parâmetro** θ , desconhecido, e (X_1, \dots, X_n) uma amostra aleatória. Seja (x_1, \dots, x_n) a amostra observada.

Definição

Chama-se **verosimilhança da amostra** e representa-se por $\mathcal{L}(\theta | x_1, x_2, \dots, x_n)$ a

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta) \quad \text{caso contínuo}$$

$$P(X_1 = x_1, \dots, X_n = x_n | \theta) = \prod_{i=1}^n P(X_i = x_i | \theta) \quad \text{caso discreto}$$

O Método da Máxima Verosimilhança (MMV)

O **método da máxima verosimilhança**, proposto por Fisher em 1922 e desenvolvido em 1925 consiste em escolher como **estimativa de θ** o valor que **maximiza a verosimilhança** $\mathcal{L}(\theta | \mathbf{x}_1, \dots, \mathbf{x}_n)$, face a uma amostra observada (x_1, \dots, x_n) .

Vejamos uma ilustração no .

Exemplo muito simples

Num pequeno lago que temos no jardim há peixes. Eu gostava de saber quantos peixes terá o lago. Como posso “estimar”?

Um procedimento comum é o chamado método de captura-recaptura, de que já falaram certamente quando deram o modelo de probabilidade hipergeométrico. Como se descreve o método?

- Apanhámos $M = 7$ peixes, marcámos cada um e libertámo-lo de seguida.
- Aguarda-se uns dias para voltarem “aos seus hábitos” e vamos pescar uns quantos de novo.
- Seja X a v.a. que conta o número de capturados-marcados, i.e. são os recapturados.
- Suponhamos que da segunda vez apanhámos $n = 4$, dos quais 3 estão marcados, então $x = 3$.

O Método da Máxima Verosimilhança (MMV)

Ora pretendemos saber o número N de peixes que existem no lago

$$N = M + (N - M);$$

X designa o número de peixes recapturados.

$$X \sim \mathcal{H}(N, n = 4, M = 7) \quad \text{então} \quad P[X = x] = \frac{\binom{7}{x} \binom{N-7}{4-x}}{\binom{N}{4}}$$

Se **encontrámos 3 marcados**, qual o número de peixes que com maior probabilidade lá existirá?

Sabemos que $N \geq 7$, mas poderá ser $N = 8, 9, 10, 11, 12, 13, 14, \dots$?

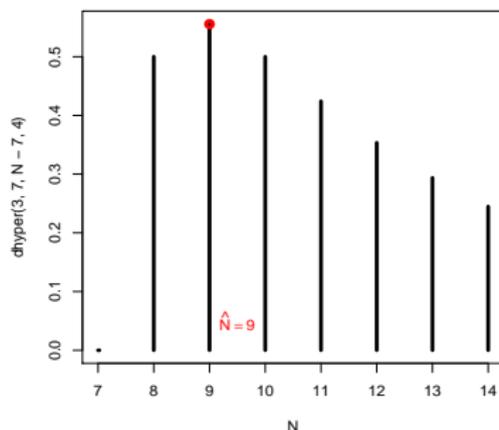
Foi Sir Ronald Fisher fez a pergunta: *“What is the value of N which has the highest likelihood?”*

Ou seja, para os valores possíveis para N qual o que torna máxima a probabilidade acima?

Resolver no , calculando as probabilidades e fazendo um gráfico

Exemplo muito simples

```
N<-seq(7,14);N  
plot(N,dhyper(3,7,N-7,4),type="h",lwd=4)  
points(9,dhyper(3,7,2,4),type="p",col=2,lwd=4)  
text(9.5,0.05,expression(hat(N)==9),col=2)
```



O Método da Máxima Verosimilhança (MMV)

Em muitas situações as funções de verosimilhança satisfazem condições que permitem que o valor para o qual a verosimilhança é máxima seja obtido por derivação.

Porém, e como a função logarítmica é monótona, regra geral é mais cómodo trabalhar com a função **log-verosimilhança**,

$$\log \mathcal{L}(\theta|\underline{x})$$

O Método da Máxima Verosimilhança (MMV)

Então, no caso de existirem derivadas (e para um único parâmetro θ), o valor do maximizante é obtido de modo que:

$$\frac{d \log \mathcal{L}}{d\theta} = 0 \quad \text{e} \quad \frac{d^2 \log \mathcal{L}}{d\theta^2} < 0$$

Note-se que
$$\frac{d \log \mathcal{L}}{d\theta} = \sum_{i=1}^n \frac{d \log f(x_i | \theta)}{d\theta}$$

A solução, $\hat{\theta}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ é **a estimativa de máxima verosimilhança**, que é uma realização da v.a. $\hat{\Theta} = \hat{\Theta}(X_1, \dots, X_n)$.

Vamos fazer um **exercício** muito simples – determinar o estimador de máxima verosimilhança do parâmetro β do modelo exponencial.

Propriedades dos estimadores de Máxima Verosimilhança

Embora possa não ser centrado, o estimador de máxima verosimilhança de um parâmetro genérico θ é o **único estimador eficiente** (caso exista estimador eficiente) e

— é **assintoticamente normal** com valor médio θ e variância assintótica dada por

$$\frac{1}{E \left[\left(\frac{\partial \log \mathcal{L}}{\partial \theta} \right)^2 \right]} = \frac{1}{nE \left[\left(\frac{\partial \log f}{\partial \theta} \right)^2 \right]} = \frac{1}{-E \left[\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2} \right]}$$

A máxima verosimilhança no R

No  existem funções para obter as **estimativas de máxima verosimilhança**, usando procedimentos iterativos.

Temos, por exemplo, as funções

`mle()` do *package* **stats4**

`fitdistr()` do *package* **MASS**

`optimize ()` – adequado quando há só um parâmetro

`optim ()` – adequado quando há dois ou mais parâmetros

Exemplo

Consideremos o ficheiro de dados do  `PlantGrowth`.
Vamos usar a amostra de valores observados da variável “weight”.
Vamos supor que as observações são a concretização de uma a.a.
 (X_1, X_2, \dots, X_n) , i.i.d. $\mathcal{N}(\mu, \sigma)$

Primeiro necessitamos de escrever a função de log-verosimilhança

Observação importante a função de optimização, por omissão calcula mínimos, pelo que devemos escrever a **log-verosimilhança negativa**

```
>minuslogL <- function(mu, sigma){  
  -sum(dnorm(x, mean = mu, sd = sigma, log = TRUE)) }  
}
```

Para obter a estimativa de máxima verosimilhança, o algoritmo de optimização requer valores iniciais para os parâmetros a estimar.

Resolução do Exemplo

Considerarei a **média e o desvio padrão observados** e também **valores próximos**, para ilustrar.

```
>data(PlantGrowth)
>x <- PlantGrowth$weight
>media<-mean(x); dp<-sd(x);media;dp
>library(stats4)
library(stats4)
>MLest1 <- mle(minuslogL, start = list(mu = media, sigma = dp))
# outra tentativa
>MLest2 <- mle(minuslogL, start = list(mu = 5.1, sigma = 0.5))
>summary(MLest1);summary(MLest2)
Maximum likelihood estimation
Call:
mle(minuslogl = minuslogL, start = list(mu = media, sigma = dp))
Coefficients:
      Estimate Std. Error
mu      5.0730000 0.12586801
sigma  0.6894075 0.08900152
-2 log L: 62.82084
```

Resumo dos principais estimadores e estimativas que iremos usar

Parâmetros que vamos referir , seus estimadores e estimativas

Parâmetro a estimar

Estimador

Estimativa

$$\mu$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

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

$$\sigma^2$$

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

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$\rho$$

$$\hat{P} = \frac{X^{(a)}}{n}$$

$$\hat{p} = \frac{x^{(b)}}{n}$$

$$\mu_1 - \mu_2$$

$$\bar{X}_1 - \bar{X}_2$$

$$\bar{x}_1 - \bar{x}_2$$

$$\sigma_1^2 / \sigma_2^2$$

$$S_1^2 / S_2^2$$

$$s_1^2 / s_2^2$$

$$p_1 - p_2$$

$$\hat{P}_1 - \hat{P}_2$$

$$\hat{p}_1 - \hat{p}_2$$

(a) X - v.a. que conta ... e $\tilde{}$ (b) x - número observado de sucessos na amostra de dimensão n .

Sampling Distributions—summary

Estimator	Assumption	Variable	Distribution
\bar{X}	$X_i \sim \mathcal{N}(\mu, \sigma)$ σ known	$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	$\mathcal{N}(0, 1)$
\bar{X}	$X_i \sim \mathcal{N}(\mu, \sigma)$ σ unknown	$\frac{\bar{X} - \mu}{S/\sqrt{n}}$	$t_{(n-1)}$
\bar{X}	X_i unknown distribution n “large”	$\frac{\bar{X} - \mu}{s/\sqrt{n}}$	$\sim \mathcal{N}(0, 1)$
S^2	$X_i \sim \mathcal{N}(\mu, \sigma)$	$\frac{(n-1)S^2}{\sigma^2}$	$\chi^2_{(n-1)}$
\hat{P}	$X \sim B(n, p)^{(a)}$ n “large”	$\frac{X - np}{\sqrt{npq}}$	$\sim \mathcal{N}(0, 1)$

^(a) X is the n. of sucesses in n Bernoulli trials (experiments).

Sampling Distributions—summary

Estimator	Assumption	Variable	Distribution
$\bar{X} - \bar{Y}$	$X_i \sim \mathcal{N}(\mu_1, \sigma_1)$ $Y_i \sim \mathcal{N}(\mu_2, \sigma_2)$ σ_1 e σ_2 unknown but that is possible to suppose equal	$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_P \sqrt{1/n_1 + 1/n_2}}$ $S_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$t_{(n_1 + n_2 - 2)}$
S_1^2 / S_2^2	$X_i \sim \mathcal{N}(\mu_1, \sigma_1)$ $Y_i \sim \mathcal{N}(\mu_2, \sigma_2)$	$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$	$F_{(n_1 - 1, n_2 - 1)}$

$X_i, i = 1, \dots, n_1$ and

$Y_i, i = 1, \dots, n_2$ are independent random samples.