Formulário

Variáveis Aleatórias (v.a.):

v.a.'s inteiras	v.a.'s contínuas
Função massa de probabilidade (f.m.p.)	**Função densidade de probabilidade (f.d.p.)**
$f : D \longrightarrow [0,1]$	$f : \mathbb{R} \longrightarrow \mathbb{R}_0^+$
$x \longmapsto f(x) = P(X = x)$	$x \longmapsto f(x)$
$\sum_x f(x) = 1$	$\int_{-\infty}^{+\infty} f(x)dx = 1$
Função distribuição (f.d.)	**Função distribuição (f.d.)**
$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$	$F(x) = P(X \leq x) = P(X < x) = \int_{-\infty}^{x} f(t)dt$
$P(a < X \leq b) = \sum_{x=a+1}^{b} f(x) =$	$P(a < X \leq b) = P(a < X < b) = P(a \leq X < b) =$
$= F(b) - F(a)$	$= P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$
Valor esperado - $E[X] = \sum_x xf(x)$	**Valor esperado -** $E[X] = \int_{-\infty}^{+\infty} xf(x)dx$
Variância - $Var[X] = E[(X - E[X])^2] =$	**Variância -** $Var[X] = E[(X - E[X])^2] =$
$= \sum_x (x - E[X])^2 f(x)$	$= \int_{-\infty}^{+\infty} (x - E[X])^2 f(x)dx$

Propriedades Gerais: **1-**Se $F(x_p) = p$, então x_p é o quantil de ordem p $(0 < p < 1)$. **2-**$E[aX + b] = aE[X] + b$ $\forall a, b \in \mathbb{R}$. **3-**Sejam X e Y v.a. $E[X+Y] = E[X]+E[Y]$. **4-**A variância da v.a. X é dada por: $Var[X] = E[X^2]-E^2[X]$. **5-**$Var[aX + b] = a^2 Var[X]$ $\forall a, b \in \mathbb{R}$. **6-**Sejam X e Y duas v.a. independentes $Var[X + Y] = Var[X] + Var[Y]$ e $E[XY] = E[X]E[Y]$.

Distribuições:

Bernoulli, $B(p)$: $f(x) = p^x(1-p)^{1-x},\ x = 0, 1.\ E[X] = p,\ Var[X] = p(1-p).$

Binomial, $B(n, p)$: $f(x) = \binom{n}{x} p^x(1-p)^{n-x},\ x = 0, 1, 2, \ldots, n.\ E[X] = np,\ Var[X] = np(1-p).$

Geométrica, $G(p)$:

X — número de provas até ao primeiro sucesso (p - probabilidade de sucesso)

$$f(x) = P(X = x) = (1-p)^{x-1}p,\ x = 1, 2, \ldots.\ E[X] = 1/p,\ Var[X] = (1-p)/p^2.$$

Y — número de insucessos até ao primeiro sucesso $(Y = X - 1)$

$$f(y) = P(Y = y) = (1-p)^y p,\ y = 0, 1, 2, \ldots.\ E[Y] = (1-p)/p,\ Var[Y] = (1-p)/p^2.$$

Poisson, $P(\lambda)$: $f(x) = e^{-\lambda}\lambda^x/x!,\ x = 0, 1, 2, \ldots.\ E[X] = \lambda,\ Var[X] = \lambda.$

Uniforme, $U(a, b)$: $f(x) = 1/(b-a),\ x \in [a, b].\ E[X] = (a+b)/2,\ Var[X] = (b-a)^2/12.$

Exponencial, $E(\lambda)$: $f(x) = \lambda e^{-\lambda x},\ \ x \geq 0.\ E[X] = 1/\lambda,\ Var[X] = 1/\lambda^2.$

Normal, $N(\mu, \sigma^2)$: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \ x \in \mathbb{R},\ \ \mu \in \mathbb{R},\ \ \sigma > 0.\ E[X] = \mu,\ Var[X] = \sigma^2.$

Método dos Momentos para dois parâmetros desconhecidos: $\left\{ \begin{array}{l} E[X] = \bar{X} \\ Var[X] = S^2 \end{array} \right.$

Teorema do Limite Central: Sejam $X_1, X_2, \ldots, X_n$ v.a.'s i.i.d. com média μ e variância σ^2.

$$\frac{(\sum_{i=1}^n X_i) - n\mu}{\sigma\sqrt{n}} \underset{n\to\infty}{\longrightarrow} Z \frown N(0,1) \quad \Leftrightarrow \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \underset{n\to\infty}{\longrightarrow} Z \frown N(0,1).$$

Em particular: Se $X \frown B(n,p) \Rightarrow \frac{X-np}{\sqrt{np(1-p)}} \overset{\circ}{\frown} N(0,1),\ n \to \infty$; Se $X \frown P(\lambda) \Rightarrow \frac{X-\lambda}{\sqrt{\lambda}} \overset{\circ}{\frown} N(0,1),\ \lambda \to \infty.$

Características amostrais: Numa amostra aleatória $(X_1, \ldots, X_n)$

$$\bar{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i \qquad S^2 = \tfrac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \qquad S_c^2 = \tfrac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$$

Distribuições de amostragem: Sejam $(X_1, X_2, \ldots, X_n)$ e $(Y_1, Y_2, \ldots, Y_m)$ duas amostras aleatórias de populações Normais com médias μ_X, μ_Y e variâncias σ_X^2, σ_Y^2 respectivamente. Seja S_{c_D} o desvio padrão da amostra de diferenças $X_i - Y_i$ quando as amostras estão emparelhadas, e defina-se $S_p = \sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_{c_X}^2 + (m-1)S_{c_Y}^2}{n+m-2}}$ quando as amostras são independentes.

- Estatísticas envolvendo a média amostral

$\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}} \frown N(0,1)$; IC para μ_X : $\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$;

$\frac{\bar{X}-\mu_X}{S_{c_X}/\sqrt{n}} \frown t_{n-1}$; IC para μ_X : $\bar{X} \pm t_{1-\frac{\alpha}{2},n-1}\ \frac{S_c}{\sqrt{n}}$;

Se as amostras forem independentes

$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \frown N(0,1)$; IC para $\mu_X - \mu_Y$: $\bar{X}-\bar{Y} \pm z_{1-\frac{\alpha}{2}}\ \sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$;

$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_p} \frown t_{n+m-2}$; IC para $\mu_X - \mu_Y$: $\bar{X}-\bar{Y} \pm t_{1-\frac{\alpha}{2},n+m-2}\ S_p$

Se as amostras forem emparelhadas

$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_{c_D}/\sqrt{n}} \frown t_{n-1}$ IC para $\mu_X - \mu_Y$: $\bar{X}-\bar{Y} \pm t_{1-\frac{\alpha}{2},n-1} S_{c_D}/\sqrt{n}$.

Se a distribuição não for Normal, pelo TLC,

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{\circ}{\frown} N(0,1)$; IC para μ_X : $\bar{X} \pm z_{1-\frac{\alpha}{2}}\ \frac{\sigma}{\sqrt{n}}$;

$\frac{\bar{X}-\mu}{S_c/\sqrt{n}} \overset{\circ}{\frown} N(0,1)$; IC para μ_X : $\bar{X} \pm z_{1-\frac{\alpha}{2}}\ \frac{S_c}{\sqrt{n}}$;

- Estatísticas envolvendo a variância amostral corrigida

$\frac{(n-1)S_{c_X}^2}{\sigma_X^2} \frown \chi^2_{(n-1)}$; IC para σ_X^2 : $\left(\frac{(n-1)S_c^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{(n-1)S_c^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right)$

$\frac{S_{c_X}^2/\sigma_X^2}{S_{c_Y}^2/\sigma_Y^2} \frown F_{n-1,m-1}$; IC para $\frac{\sigma_X^2}{\sigma_Y^2}$: $\left(\frac{S_{c_X}^2}{S_{c_Y}^2} f_{\frac{\alpha}{2},m-1,n-1}, \frac{S_{c_X}^2}{S_{c_Y}^2} f_{1-\frac{\alpha}{2},m-1,n-1}\right)$

com $f_{1-\alpha,\nu,\omega} = \frac{1}{f_{\alpha,\omega,\nu}}$

Proporções: $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \overset{\circ}{\frown} N(0,1)$ $\hat{p} = \frac{X}{n}$; IC para p : $= \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

ANOVA: Comparação de médias (ou medianas) de g grupos.

ANOVA paramétrica: Todos os grupos devem ter distribuição Normal com a mesma variância. Os grupos devem ter dimensão n.

$SS_T = \sum_{i=1}^{g}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{..})^2$ $SS_G = n\sum_{i=1}^{g}(\bar{Y}_{i.}-\bar{Y}_{..})^2$ $SS_E = \sum_{i=1}^{g}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i.})^2$ $SS_T = SS_G + SS_E$

Fonte de Variação	Soma de quadrados	graus de liberdade	Média dos quadrados	F_0	$p-value$
Entre Grupos	SS_G	$g-1$	$MS_G = \frac{SS_G}{g-1}$	$\frac{MS_G}{MS_E}$	$(\cdot)$
Dentro dos grupos	SS_E	$g(n-1)$	$MS_E = \frac{SS_E}{g(n-1)}$		
Total	SS_T	$gn-1$			

Efeitos aleatórios: $\hat{\sigma}_\tau^2 = (MS_G - MS_E)/n$

ANOVA não-paramétrica: teste de Kruskal-Wallis.

Comparações múltiplas (pares de médias) Consideram-se todas as comparações de pares de médias envolvidas na ANOVA. Método de **Bonferroni** e de **Tukey** solucionam a problemática associada ao nível de significância do conjunto de comparações (m). **Método de Bonferroni** - reduz o tamanho individual para que o tamanho total seja o desejado: $\alpha = m\alpha_m$ onde α_m é o tamanho de cada comparação individual. **LSD** (Least Significant Difference): comparações múltiplas sem qualquer correcção.

Regressão linear $Y_i = b_0 + b_1 x_{1,i} + \ldots + b_k x_{k,i} + \epsilon_i, \quad \epsilon_i \frown N(0, \sigma^2), \quad iid.$ Seja $p = k + 1$.

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$S_{YY} = SS_R + SS_E \qquad R^2 = \frac{SS_R}{S_{YY}} = 1 - \frac{SS_E}{S_{YY}} \qquad R_a^2 = 1 - \frac{SS_E/(n-p)}{S_{YY}/(n-1)}$$

Estimadores dos parâmetros e suas propriedades

$$\hat{\sigma}^2 = \frac{SS_E}{n-p} \qquad \frac{(n-p)\hat{\sigma}^2}{\sigma^2} = \frac{SS_E}{\sigma^2} \frown \chi^2_{(n-p)} \qquad T_i = \frac{\hat{b}_i - b_i}{\hat{\sigma}_{b_i}} \frown t_{n-p}$$

Tabela de regressão (contém os testes individuais $H_0: \ b_i = 0 \ \ vs \ \ H_1: \ b_i \neq 0$)

		Coeficientes não-estandardizados	Coeficientes estandardizados		
Coeficiente	b	Erro padrão	β	t	$p - value$
Constante	$\hat{b}_0$	$\hat{\sigma}_{b_0}$		t_{0obs}	$(\cdot)$
x_1	$\hat{b}_1$	$\hat{\sigma}_{b_1}$	$\hat{\beta}_1$	t_{1obs}	$(\cdot)$
x_2	$\hat{b}_2$	$\hat{\sigma}_{b_2}$	$\hat{\beta}_2$	t_{2obs}	$(\cdot)$
$\vdots$					
x_k	$\hat{b}_k$	$\hat{\sigma}_{b_k}$	$\hat{\beta}_k$	$t_{k\,obs}$	$(\cdot)$

ANOVA da regressão (contém o teste às hipóteses $H_0: \ b_1 = b_2 = \ldots = b_k = 0 \ \ vs \ \ H_1: \ \exists b_i \neq 0$)

Fonte de Variação	SS	$g.l.$	MS	F_0	$p - value$
Regressão	SS_R	k	$MS_R = \frac{SS_R}{k}$	$\frac{MS_R}{MS_E}$	$(\cdot)$
Erros	SS_E	$n-p$	$MS_E = \frac{SS_E}{n-p}$		
Total	S_{YY}	$n-1$			

Outras quantidades de interesse:

Valor predito $E[Y|\boldsymbol{x}] = \mu(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{b}$, $\hat{\mu}(\boldsymbol{x}) = \boldsymbol{x}'\hat{\boldsymbol{b}}$: $\dfrac{\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})}{\hat{\sigma}\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}} \frown t_{n-p}$

Observação futura $Y|\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{b} + \boldsymbol{\epsilon}$. $\hat{Y}|\boldsymbol{x} = \boldsymbol{x}'\hat{\boldsymbol{b}}$: $\dfrac{\hat{Y}|\boldsymbol{x} - Y|\boldsymbol{x}}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}} \frown t_{n-p}.$

Avaliação da qualidade e validação dos pressupostos da regressão:

Avaliação: Diagramas de dispersão; Gráfico de dispersão entre os valores observados Y_i e os preditos $(\hat{Y}_i)$; Coeficiente de determinação R^2; Comparação entre a variância dos erros $\hat{\sigma}^2$ e a variância de Y_i; Teste ao significado da regressão (ANOVA).

Validação: os resíduos (e_i) devem ser Normais com variância constante e independentes: QQ-plot ou PP-plot; Gráfico de resíduos versus valores preditos $\hat{Y}_i$.

. . . . . .

Boxplots: barreiras de outliers: $Q_{1/4} - 1.5df$ e $Q_{3/4} + 1.5df$; **barreiras de extremos (outliers severos):** $Q_{1/4} - 3df$ e $Q_{3/4} + 3df$, com $df = Q_{3/4} - Q_{1/4}$.

Testes de hipóteses: Num teste de hipóteses a hipótese nula deve ser sempre simples.

p-value do teste: é a probabilidade de observar um valor da estatística de teste tanto ou mais afastado que o valor observado na amostra, assumindo que H_0 é verdadeira.

Num teste de tamanho α rejeita-se a hipótese nula (de igualdade) quando p-value $< \alpha$.

Para transformar um p-value bilateral em unilateral divide-se por dois desde que a(s) amostra(s) aponte(m) no sentido da hipótese alternativa. Caso contrário calcula-se ($1-p$-value/2).

Há 3 procedimentos para realizar um teste de hipóteses:

1. Cálculo da região crítica
2. Através do p-value.
3. Através de intervalos de confiança (válido apenas para testes bilaterais). Neste caso rejeita-se H_0 se o valor em teste não pertencer ao IC construído para o parâmetro.