Distribuição amostral de estatísticas importantes

Prof. Wagner Hugo Bonat

Departamento de Estatística
Universidade Federal do Paraná

Distribuição amostral da média

- Sejam $Y_1, \ldots, Y_n$ v.a.'s independentes e identicamente distribuídas (iid) com distribuição desconhecida, porém com média $E(Y_i) = \mu$ e variância $V(Y_i) = \sigma^2 < \infty$. Para amostras grandes o TLC nos diz que

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

- E para outras estatísticas de interesse?

Distribuição amostral da média

- Sejam $Y_1, \ldots, Y_n$ v.a.'s independentes e identicamente distribuídas (iid) com distribuição desconhecida, porém com média $E(Y_i) = \mu$ e variância $V(Y_i) = \sigma^2 < \infty$. Para amostras grandes o TLC nos diz que

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

- E para outras estatísticas de interesse?
- De forma geral é dificil obter a distribuição amostral de outras estatísticas.
- Porém para v.a.'s Normais temos alguns resultados importantes.

Amostragem de v.a.'s Normais e estatísticas relacionadas

Sejam $Y_1, \ldots, Y_n$ v.a.'s iid com distribuição $N(\mu, \sigma^2)$.

Algumas estatísticas relacionadas são:

- Média amostral $\rightarrow \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- Variância amostral $\rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ ou $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$.
- Estatística t-Student $\rightarrow t = \frac{\overline{Y}-\mu}{S/\sqrt{n}}$.
- Sendo duas v.a.'s Normais com variância S^2_{Y1} e S^2_{Y2}, respectivamente. A razão $\frac{S^2_{Y1}}{S^2_{Y2}}$ é chamada de estatística F.

Distribuição χ^2

- Sendo $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, então

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \text{onde} \quad n-1 \quad \text{são os graus de liberdade.}$$

- Função densidade probabilidade $Y_s \sim \chi^2_k$

$$f(y_s; k) = \frac{y_s^{\frac{k}{2}-1} e^{-\frac{y_s}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \quad \text{para} \quad k \in N \quad \text{e} \quad y_s > 0.$$

- Cálculo de probabilidades $\rightarrow$ tabelas (só para as caudas) ou softwares estatísticos.

Ilustração computacional

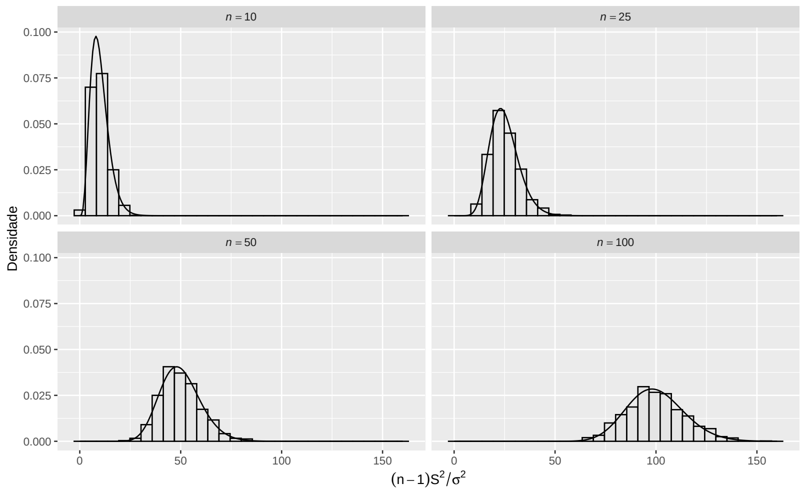


Figura 1. Distribuição amostral da estatística qui-quadrado.

Aplicações e propriedades da distribuição χ^2

- Muito comum em testes de hipóteses:
  - Independência em tabelas de contingência.
  - Bondade de ajuste.
  - Razão de verossimilhanças.
  - Log-rank.
  - Cochran-Mantel-Haenszel.
- Soma de quadrados de $n-1$ Normais padrão independentes.
- Caso particular da distribuição Gama.

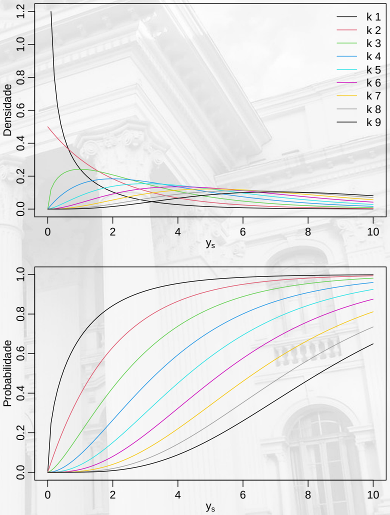


Figura 2. Distribuição qui-quadrado.

Exemplo: Bateria para celular

Uma empresa desenvolveu uma nova bateria para celular. Em média a bateria dura 60 horas com desvio-padrão de 4 horas. Suponha que o fabricante efetua um controle da qualidade das baterias onde são selecionadas aleatoriamente 7 baterias. Supondo que a duração das baterias pode ser adequadamente modelada pela distribuição Normal. Calcule

- Probabilidade da variância amostral ser maior que 16 horas.
- Probabilidade da variância amostral estar entre 4 e 36 horas.
- Probabilidade da variância amostral ser menor do que 4 horas.

Figura 3. Foto de Tyler Lastovich no Pexels.

Exemplo: Bateria para celular (cont.)

- Probabilidade da variância amostral ser maior que 16 horas.

$$\mathrm{P}(S^2 > 16) = \mathrm{P}\left((n-1)\frac{s^2}{\sigma^2} > (7-1)\frac{16}{16}\right) = P(\chi^2_{7-1} > 6) \approx 0{,}423.$$

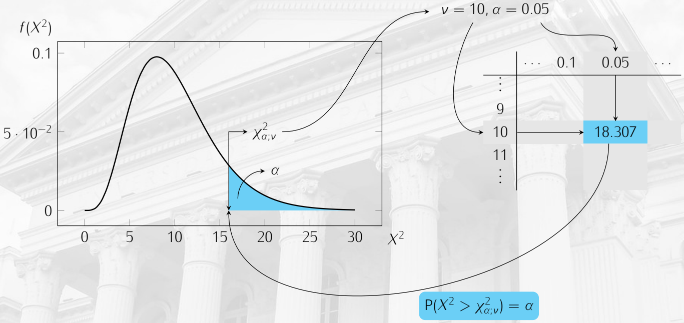
- Probabilidade da variância amostral estar entre 4 e 36 horas.

$$\mathrm{P}(4 < S^2 < 36) = \mathrm{P}\left((7-1)\frac{4}{16} < \chi^2_{7-1} < (7-1)\frac{36}{16}\right) = P(1{,}5 < \chi^2_{7-1} < 13{,}5) \approx 0{,}923.$$

- Probabilidade da variância amostral ser menor do que 4 horas.

$$\mathrm{P}(S^2 < 4) = \mathrm{P}\left(\chi^2_{7-1} < (7-1)\frac{4}{16}\right) = P(\chi^2_{7-1} < 1{,}5) \approx 0{,}040.$$

Consulta da tabela χ^2



Pontos percentuais da distribuição χ^2 com áreas na calda direita.

ν/α	$\alpha = 0.995$	0.99	0.975	0.95	0.9	0.5	0.1	0.05	0.025	0.01	0.005
$\nu = 1$	0.000	0.000	0.001	0.004	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	3.357	7.779	9.488	11.143	13.277	14.860

Figura 4. Consulta da tabela qui-quadrado.

Distribuição t-Student

- Sendo $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ e $\overline{Y}$ e S^2 a média e a variância amostral a v.a.

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

  t_{n-1} denota a distribuição t-Student com $n-1$ graus de liberdade.

- Função densidade probabilidade

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

  onde $\nu \in \mathbb{N}$ é o número de graus de liberdade e $\Gamma(\cdot)$ é a função gama.

- Cálculo de probabilidades $\rightarrow$ tabelas similares às da distribuição Normal ou softwares estatísticos.

Ilustração computacional

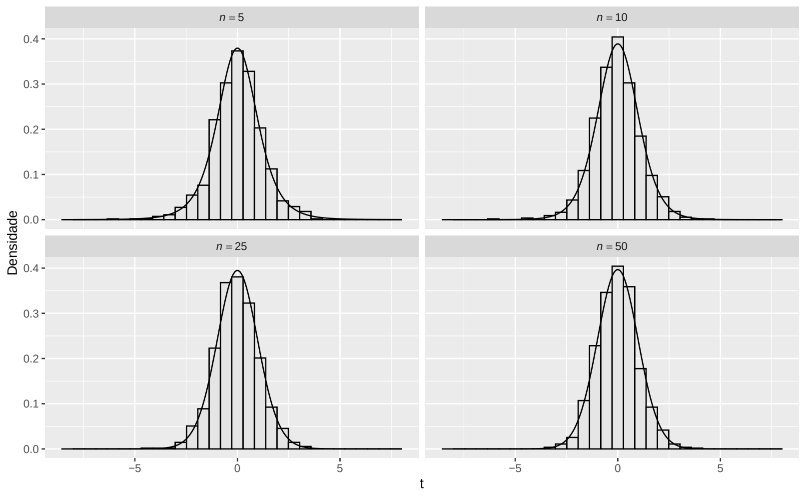


Figura 5. Distribuição amostral da estatística t-Student.

Aplicações e propriedades da distribuição t-Student

- Teste t e suas variações.
- Intervalo de confiança para a média.
- Simétrica em forma de sino (igual à Normal).
- Caudas mais pesadas que a Normal.
- Descreve o comportamento da razão de algumas v.a.'s.
- Desenvolvida por William Sealy Gosset (sob o pseudônimo de Student).

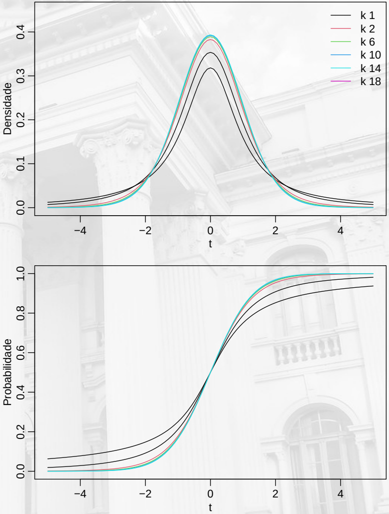


Figura 6. Distribuição t-Student.

Exemplo: Acupuntura

Suponha que um experimento foi realizado para avaliar a efetividade do uso da acupuntura para aliviar a dor. A taxa sensorial de 15 pacientes foi medida resultando em uma média de 8,22 e um desvio-padrão de 1,67. Supondo que a distribuição Normal é adequada para a variável de interesse obtenha um intervalo t_1 e t_2, tal que a probabilidade deste intervalo conter a média populacional seja de aproximadamente 0,95.

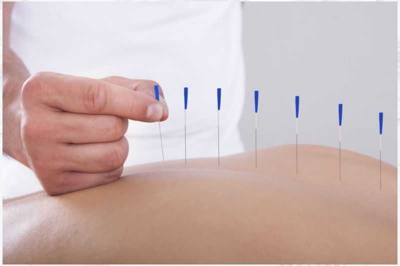


Figura 7. Extraído de www.health.harvard.edu.

Exemplo 2 (cont.)

- Precisamos obter valores $\overline{y}_{li}$ e $\overline{y}_{ls}$, tal que

$$\mathrm{P}(\overline{y}_{li} < \mu < \overline{y}_{ls}) = 0{,}95.$$

- Vamos padronizar

$$\mathrm{P}\left(\frac{\overline{y}_{li} - \mu}{s/\sqrt{n}} < \frac{\overline{y} - \mu}{s/\sqrt{n}} < \frac{\overline{y}_{ls} - \mu}{s/\sqrt{n}}\right) = \mathrm{P}\left(t_1 < \frac{\overline{y} - \mu}{s/\sqrt{n}} < t_2\right) = 0{,}95$$

- Note que a distribuição t—Student é simétrica, então vamos focar apenas em intervalos simétricos, o que implica que $t = t_2 = -t_1$. Assim os limites são dados por

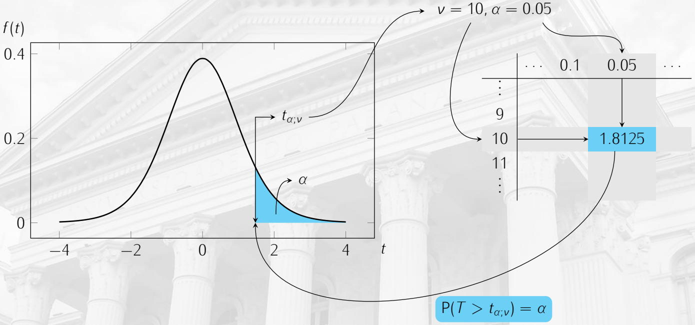
$$\overline{y} \pm t_{0{,}05/2}\frac{s}{\sqrt{n}},$$

onde $t_{0{,}05/2}$ é o valor da distribuição t-Student com $n - 1$ graus de liberdade.

- Usando a tabela da distribuição t-Student com 14 graus de liberdade, temos

$$8{,}22 \pm 2{,}14\frac{1{,}67}{\sqrt{15}} \approx [7{,}30; 9{,}14].$$

Consulta da tabela t-Student



Pontos percentuais da distribuição t de Student com áreas na calda direita.

ν/α	$\alpha = 0.4$	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
$\nu = 1$	0.3249	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567	127.3213	318.3088	636.6192
2	0.2887	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248	14.0890	22.3271	31.5991
3	0.2767	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409	7.4533	10.2145	12.9240
4	0.2707	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041	5.5976	7.1732	8.6103
5	0.2672	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321	4.7733	5.8934	6.8688

Figura 8. Consulta da tabela t-Student.

Distribuição F de Snedecor

- Sejam $Y_{1i} \sim \mathrm{N}(\mu_1, \sigma_1^2)$ e $Y_{2i} \sim \mathrm{N}(\mu_2, \sigma_2^2)$. Com média e variância amostrais $\overline{Y}_1$, $\overline{Y}_2$, S_1^2 e S_2^2. Suponha ainda que amostras de tamanho n_1 e n_2 estão disponíveis de Y_1 e Y_2. Se $\sigma_1^2 = \sigma_2^2$, então temos que a v.a.

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1},$$

em que F_{n_1-1,n_2-1} denota a distribuição F com $n_1 - 1$ e $n_2 - 1$ graus de liberdade.

- Função densidade probabilidade

$$f(y) = \sqrt{\frac{(d_1 y)^{d_1} d_2^{d_2}}{(d_1 y + d_2)^{d_1+d_2}}} \Big/ y\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right) \quad \text{para} \quad y > 0,$$

onde d_1 e d_2 são os graus de liberdade do numerador e denominador e $\mathrm{B}(\cdot)$ é a função beta.

- Cálculo de probabilidades $\rightarrow$ tabelas ou softwares estatísticos.

Ilustração computacional

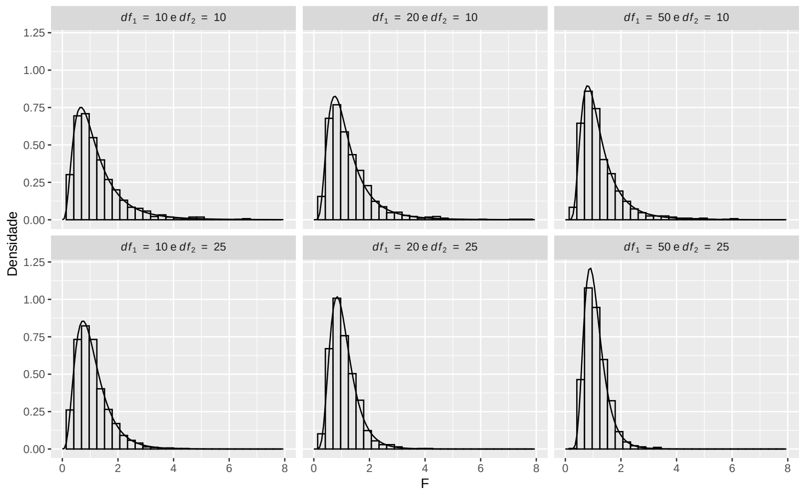


Figura 9. Distribuição amostral da estatística F.

Aplicações e propriedades da distribuição F

- Teste F para igualdade de variâncias.
- ANOVA - Analise de variância.
- Modelos de regressão.
- Razão entre v.a.'s qui-quadrado.
- Também conhecida como distribuição de Fisher-Snedecor's.
- Se $\sigma^2_{Y_1} \neq \sigma^2_{Y_2}$ a estatística F ainda tem distribuição F, porém não central.

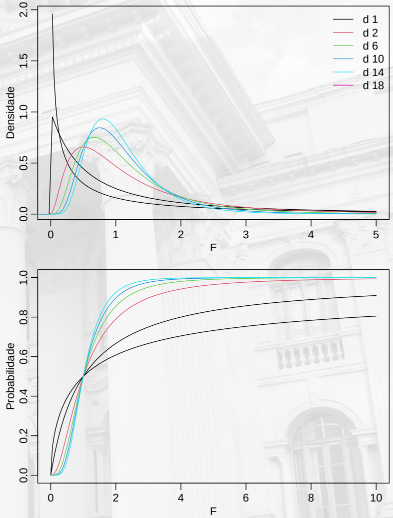
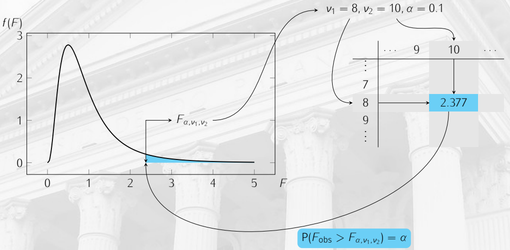


Figura 10. Distribuição F de Snedecor.

Consulta da tabela F de Snedecor



Pontos percentuais da distribuição F com áreas na calda direita para $\alpha = 0.1$.

	$\nu_2 = 1$	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
$\nu_1 = 1$	39.863	8.526	5.538	4.545	4.060	3.776	3.589	3.458	3.360	3.285	3.177	3.073	2.975	2.927	2.881	2.835	2.791	2.748
2	49.500	9.000	5.462	4.325	3.780	3.463	3.257	3.113	3.006	2.924	2.807	2.695	2.589	2.538	2.489	2.440	2.393	2.347
3	53.593	9.162	5.391	4.191	3.619	3.289	3.074	2.924	2.813	2.728	2.606	2.490	2.380	2.327	2.276	2.226	2.177	2.130
4	55.833	9.243	5.343	4.107	3.520	3.181	2.961	2.806	2.693	2.605	2.480	2.361	2.249	2.195	2.142	2.091	2.041	1.992
5	57.240	9.293	5.309	4.051	3.453	3.108	2.883	2.726	2.611	2.522	2.394	2.273	2.158	2.103	2.049	1.997	1.946	1.896
6	58.204	9.326	5.285	4.010	3.405	3.055	2.827	2.668	2.551	2.461	2.331	2.208	2.091	2.035	1.980	1.927	1.875	1.824
7	58.906	9.349	5.266	3.979	3.368	3.014	2.785	2.624	2.505	2.414	2.283	2.158	2.040	1.983	1.927	1.873	1.819	1.767

Figura 11. Consulta da tabela F de Snedecor

Exemplo: Acunputura

Suponha que um experimento foi realizado com dois grupos para avaliar a efetividade do uso da acupuntura para aliviar a dor. A taxa sensorial foi medida para o grupo 1 em 5 pacientes e para o 2 em 8 pacientes. Suponha que as variâncias amostrais foram $s_1^2 = 4{,}44$ e o $s_2^2 = 1{,}5$. Assumindo que as variâncias populacionais são iguais, qual a probabilidade de ocorrer a razão $\frac{s_1^2}{s_2^2}$ ou uma mais extrema?

$$P\left(\frac{s_1^2}{s_2^2} > \frac{4{,}44}{1{,}5}\right) = \mathrm{P}(F_{5-1,8-1} > 2{,}96) \approx 0{,}1.$$

Você considera a suposição de igualdade de variâncias plausível dado o resultado do experimento?

Relações entre as distribuições

- Uma v.a. Normal padrão ao quadrado tem distribuição χ^2 com $gl = 1$.
- Uma v.a. t-Student ao quadrado tem distribuição F com $gl = 1$.
- Razão de duas v.a. χ^2 divida pelos seus gl's tem distribuição F_{n_1,n_2}.
- Distribuição F converge para a χ^2 com $n \to \infty$.
- Existem extensões não-centrais (modelo de locação e escala).
- Distribuição t-Student é uma alternativa robusta a Normal.
- Todas são relacionadas a Normal e quando gl cresce vão convergir para a Normal.

Resumo

- Distribuição amostral de estatísticas importantes.
  - Distribuição χ^2;
  - Distribuição $t-$Student;
  - Distribuição F.

Figura 12. Retirada do Google imagens.