Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Probabilidade e Estatística 2
1^o semestre de 2020
Gustavo L. Gilardoni

Lista de Exercícios 4

1. Um dado foi lançado $n = 200$ vezes obtendo-se 11 vezes a face "1", 34 a "2", 40 a "3", 23 a "4", 38 a "5" e 54 vezes a "6".

Teste, ao nível de significância 5%, a hipótese nula que o dado é não-viesado e calcule o p-valor do teste.

2. No Exemplo 2 das notas de aula foi discutido um conjunto de dados contendo o número de falhas de 50 motores de certas cintas transportadoras, mostrado na seguinte tabela.

# de Falhas	**0**	**1**	**2**	**3**	**Total**
# de Motores	24	10	12	4	50

Uma possibilidade para ajustar esses dados é usar a Distribuição Geométrica. Resumidamente, dizemos que uma variável aleatória X segue a distribuição Geométrica com parâmetro p se $P(X = x) = (1 - p)\, p^x$ para $x = 0, 1, \ldots$. Se $X_1, \ldots, X_n$ é uma amostra aleatória da distribuição Geométrica, o estimador de máxima verossimilhança de p é $\hat{p} = 1/(1 + \bar{X}) = n/(n + \sum_{i=1}^{n} X_i)$.

Teste, ao nível de significância 5%, a hipótese nula que o número de falhas segue uma distribuição Geométrica e calcule o p-valor do teste.

3. (Devore) O artigo "Racial Stereotypes in Children's Television Commercials" ("Estereotipos raciais nos anuncios televisivos de crianças", J. of Adver. Res., 2008: 80–93) reporta as seguintes frequências com as quais caracteres étnicos apareceram em anuncios gravados que foram ao ar em canais de tevê de Filadelfia:

Grupo	**afro-americanos**	**asiáticos**	**caucasianos**	**Latinos**
Frequência	57	11	330	6

Segundo o censo americano do ano 2000, as proporções na população para esses 4 grupos eram 0.177, 0.032, 0.734 e 0.057 respectivamente. Sugerem esses dados que as proporções nos anuncios são diferentes das do censo? Use nível de significância 1%.

4. O seguinte código foi usado na linguagem **R** para gerar $n = 1000$ observações da distribuição Normal(0,1) e construir um histograma.

```
> x<-rnorm(1000,0,1)
> mean(x)
[1] 0.03161167
> sd(x)
[1] 1.003704
> hist(x,breaks=c(min(x),-2,-1,0,1,2,max(x)))
> hist(x,breaks=c(min(x),-2,-1,0,1,2,max(x)))$counts
[1]  32 115 347 340 150  16
```

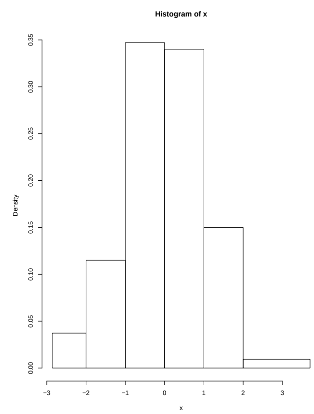

Figura 1: Histograma de 1000 valores gerados usando a função `rnorm` do **R**.

Os valores gerados foram agrupados em intervalos de acordo à seguinte tabela.

Intervalo	$(-\infty, -2]$	$(-2, -1]$	$(-1, 0]$	$(0, 1]$	$(1, 2]$	$(2, \infty]$
Frequências	26	129	358	318	144	25

(i) Teste, ao nível de significância de 1%, se o gerador do **R** gerou efetivamente números aleatórios da distribuição Normal(0,1). Calcule o p-valor. **(ii)** Suponha que não soubéssemos a média μ e variância σ^2 usadas para gerar os dados. Qual seria o procedimento para testar se os dados foram gerados de uma distribuição Normal qualquer? Faça esse teste usando $\alpha = 0.01$ e calcule o p-valor.

5. Se $\mathbf{p} = (p_1, \dots, p_K)$ e $\mathbf{q} = (q_1, \dots, q_K)$ são distribuições de probabilidades tais que $p_i \geq 0$, $q_i > 0$ e $\sum_{i=1}^K p_i = \sum_{i=1}^K q_i = 1$, a *divergência qui-quadrado* de $\mathbf{p}$ com respeito a $\mathbf{q}$ é definida por

$$\chi^2(\mathbf{p}; \mathbf{q}) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i} .$$

(i) Mostre que $\chi^2(\mathbf{p}; \mathbf{q}) \geq 0$ com igualdade se, e somente se, $p_i = q_i$ para todo i. **(ii)** Mostre que $\chi^2(\mathbf{p}; \mathbf{q}) = \sum_{i=1}^K \frac{p_i^2}{q_i} - 1$. **(iii)** Mostre que, em geral, a divergência χ^2 não é simétrica, no sentido que, em geral, não é verdade que $\chi^2(\mathbf{p}; \mathbf{q}) = \chi^2(\mathbf{q}; \mathbf{p})$.

6. Num teste de aderência, sejam $n_1^{obs}, \dots, n_K^{obs}$ as frequências observadas, $n = \sum_{i=1}^K n_i^{obs}$ o tamanho amostral e $n_1^{esp}, \dots, n_K^{esp}$ as frequências esperadas ($n_i^{esp} = np_{0i}$ ou $n_i^{esp} = n\hat{p}_{0i}$, dependendo do caso). Mostre que o estatístico χ^2_{obs} pode ser escrito como

$$\chi^2_{obs} = \sum_{i=1}^{K} \frac{(n_i^{obs})^2}{n_i^{esp}} - n .$$

Nos exemplos, qual das duas formas de χ^2_{obs} é mais fácil de calcular?

7. (Bussab e Morettin). Num laboratório foi realizada uma pesquisa de mercado em que se estudou a preferência com relação a dois adoçantes artificiais, A e B, obtendo-se os resultados seguintes.

Sexo	**Preferem A**	**Preferem B**	**Indiferentes**
Feminino	50	110	40
Masculino	150	42	8

Teste, ao nível de significância de 1%, se a distribuição de preferências pelos dois sexos é a mesma. Calcule o p-valor.

8. Uma companhia mineradora opera uma frota de caminhões *off-road* equipados com motores de 3 fabricantes diferentes. Um engenheiro de produção classificou as falhas desses motores em três tipos diferentes e coletou os dados a seguir.

	Tipo de falha			
Fabricante	**1**	**2**	**3**	**4**
A	14	18	8	10
B	20	48	12	20
C	10	16	14	10

Ao nível de significância de 5%, existe evidência que a distribuição dos tipos de falha é diferente segundo o fabricante do motor? Calcule o p-valor do teste.

9. (Devore). No artigo "Alcohol use in students seeking primary care treatment at university health services" ("Consumo de bebidas alcóolicas por estudantes que procuram as unidades de saúde universitárias", J. of Amer. College Health, 2012: 217–225) cada aluno do sexo masculino numa amostra foi classificado de acordo a grupo etário e ao número de episodios de consumo excessivo de álcool durante os últimos 30 dias.

	Grupo etário		
# episodios	**18-20**	**21-23**	**$\geq$ 24**
Nenhum	357	293	592
1-2	218	285	354
3-4	184	218	185
$\geq$ 5	328	331	147

De acordo a esses dados, parece existir associação entre uso de álcool e grupo etário na população da qual a amostra foi selecionada? Use $\alpha = 0.01$ e calcule o p-valor.

10. Uma amostra de 1000 alunos de uma certa universidade foi classificada de acordo ao seu uso de tabaco (fumante, não fumante) e ao uso de tabaco pelos seus pais (nenhum, somente um e os dois fumantes). Os dados são apresentados na seguinte tabela.

	Fumante		
# pais fumantes	**Sim**	**Não**	**Total**
Nenhum	75	257	332
Só 1	77	339	416
Os dois	35	217	252
Total	187	813	1000

Teste, ao nível de significância 5%, se existe associação entre uso de tabaco por alunos dessa universidade e por seus pais. Calcule o p-valor.

Algumas soluções

1. $\chi^2_{obs} \doteq 32.98$, p-valor $\doteq 3.8 \times 10^{-6}$, ao nível de 5% rejeitamos a H_0 do dado ser não-viesado.

8. $\chi^2_{obs} \doteq 6.477$, p-valor $\doteq 0.043$, ao nível de 5% rejeitamos a H_0 da distribuição ser geométrica.

3. $\chi^2_{obs} \doteq 19.599$, p-valor $\doteq 0.0002$, ao nível de 5% rejeitamos a H_0 da distribuição seguir as proporções do censo.

4. **(i)** $\chi^2_{obs} \doteq 3.929$ com 5 gl, p-valor $\doteq 0.560$, não rejeitamos H_0 que as observações foram geradas da distribuição Normal Padrão; **(ii)** $\chi^2_{obs} \doteq 4.659$ com 3 gl, p-valor $\doteq 0.199$, não rejeitamos H_0 que as observações foram geradas de uma distribuição Normal.

7. $\chi^2_{obs} \doteq 101.75$ com 2gl, p-valor $\doteq 2.2 \times 10^{-16}$, ao nível de 1% rejeitamos a H_0 de homogeneidade.

8. $\chi^2_{obs} \doteq 8.638$ com 6gl, p-valor $\doteq 0.195$, ao nível de 5% não existe evidência que a distribuição sea diferente para cada fabricante.

9. $\chi^2_{obs} \doteq 212.91$ com 6gl, p-valor $\doteq 2.2 \times 10^{-16}$, ao nível de 5% rejeitamos a H_0 de não associação (i.é. de independência entre grupo etário e número de episódios).

10. $\chi^2_{obs} \doteq 7.1517$ com 2gl, p-valor $\doteq 0.028$, ao nível de 5% rejeitamos a H_0 de não associação (i.é. de independência entre fumante/não fumante e número de pais fumantes).