Coleção de Exercícios de Estatística e Probabilidade para Ciências Sociais
Departamento de Estatística - UFMG

Amanda Xavier[1], Gustavo Narimatsu[2] e Adrian Luna (orientador)[3]

[1]Departmento de Estatística - UFMG

Julho, 2020

Conteúdo

0.1 Introdução

Nas Ciências Sociais a necessidade de exemplos e exercícios tanto estatísticos quanto de probabilidade e que sejam apropriados a área social é uma necessidade permanente na Graduação. Exemplos que permitam melhorar a compreensão na hora de esclarecer as duvidas dos alunos que acudem à monitoria da graduação. Bem seja pela dificuldade do uso de modelos probabilísticos e estatísticos ou pela poca familiaridade com as contas matemáticas ou também por ter que lidar com dados revela-se complicado para quem esta imerso no contexto das Ciências Humanas. Fizemos estes exemplos como suporte, na monitoria da graduação, a tarefa de ensino da Estatística e as Probabilidades nas Ciências Sociais. Aproveitamos para agradecer o apoio, para a conclusão desta coleção, da Chefa do departamento de Estatística, Professora Glaura Franco, e também ao financiamento das bolsas dos monitores pelo Programa de Monitoria da Graduação (PMG) da Pró-Reitoria de Graduação (Prograd).

No capítulo 1 (Exemplos de Pesquisa Empírica) apresentamos exemplos que tem o objetivo de entender o processo de formulação de pesquisa, visando entender o problema a ser abordado a as informações a serem consideradas. No capítulo 2 (Amostragem) apresentamos exemplos que exploram diferentes técnicas de amostragem, como qual tipo de amostragem usar, e a utilização da tabela de números aleatórios para fazer amostragem. No capítulo 3, abordamos exemplos de estatística descritiva, que tem o intuito de mostrar como calcular as medidas de tendência central e os gráfico adequados para diferentes tipos de dados. No capítulo 4 (Probabilidade) apresentamos exercícios que abordam conceitos de suma relevância dentro deste contexto, tais como a noção de espaço amostral, evento de interesse, distribuições de probabilidade contínua e discreta (binomial, normal e qui-quadrado), esperança e variância aplicadas em probabilidade, probabilidade condicional e independência, assim como suas respectivas soluções. No capítulo 5 (Inferência) apresentamos exercícios de teste para a proporção, média e diferença entre médias para grupos pareados e não pareados, intervalo de confiança, independência (qui-quadrado) e teste de hipóteses, incluindo a interpretação dos métodos de resolução destes problemas, como o p-valor, assim como suas respectivas soluções. No capítulo 6 (Regressão Linear) apresentamos exercícios de coeficiente de correlação de Pearson e R^2, visualização dos dados em gráficos de dispersão, cálculo e interpretação dos coeficientes da reta de regressão e suas respectivas soluções.

Capítulo 1

Exemplos de Pesquisa Empírica

Exemplo 1.0.1. *A seguinte notícia apareceu nos jornais: "... o déficit habitacional [no Brasil] de 7,2 milhões de unidades habitacionais. Assistimos o resultado disso nas grandes cidades brasileiras: o adensamento nas favelas e periferias e a sobre oferta de unidades habitacionais para a demanda de renda média que permanecem "encalhadas", com grandes possibilidades de engrossar o número de casas e apartamentos vazios, que hoje já é quase igual ao déficit habitacional do país – 6,7 milhões de unidades. Paradoxo?" Observe que estão sendo relacionados o déficit habitacional e moradia nas favelas e periferias , e a sobre oferta habitacional para a demanda com renda média.*

Considere a problemática mencionada acima,

1. *Como vôçe definiría o problema numa pesquisa empírica* **comparativa** *(população alvo, formulação de objetivos, etc). (máximo 5 linhas)*
2. *Descreva e classifique (qualitativas ou quantitativas) as variáveis que vôçe incluiría no estudo. (máximo 4 variáveis)*
3. *Elabore um questionário para a pesquisa. (máximo 4 questões)*

Solução

1. Como você definiría o problema numa pesquisa empírica **comparativa** (população alvo, formulação de objetivos, etc).

   Problema: Caracterizar o significado do 'deficit habitacional' nas favelas e nos bairros de classe media.

   População Alvo O bairro de classe media *Caiçara* e a favela *Pedreira Prado Lopes* .

   Objetivo Principal Comparar o sentido que tem o 'deficit habitacional' nos setores da população que moram na favela e nos bairros de classe media.

   Objetivos Secundarios: • Relacionar a condição socioeconômica/classe social, o gênero e a raça com a noção de deficit habitacional.

- Relacionar a condição socioeconômica/classe social, o gênero e a raça com os moveis 'desocupados'.
- Relacionar a condição socioeconômica/classe social com aceso ao trabalho

2. Descreva e classifique (qualitativas ou quantitativas) as variáveis que você incluiría no estudo. (máximo 6 variáveis):
   (a) Variável: *Salario*. Nome: SAL , Tipo: Qualitativa ordinal, Valores: A (mais de 10 SM),B (5 à 10 SM),C (2 a 5 SM),D (menos de 2 SM)
   (b) Variável: *Gênero*. Nome: GEN , Tipo: Qualitativa nominal, Valores: H (homem), M (mulher).
   (c) Variável: *Raça*. Nome: RAZ , Tipo: Qualitativa nominal, Valores: B (branca), N (negra), P (parda).
   (d) Variável: *Lugar de Moradia*. Nome: LOCAL , Tipo: Qualitativa nominal, Valores: F (favela), B (bairro).
   (e) Variável: *Deficit habitacional*. Nome: DEF , Tipo: Qualitativa nominal, Valores: S (Sim), N (não).
   (f) Variável: *Distância ao trabalho*. Nome: TRAB , Tipo: Qualitativa nominal, Valores: P (perto), M (media distância), L (longe).
3. Elabore um questionário para a pesquisa. (máximo 6 questões)
   (a) A qual faixa de salario (em salários mínimos) sua família pertence?
      i. De zero até 2 SM.
      ii. De de 2 SM até 5 SM.
      iii. De 5 SM até 10 SM.
      iv. Mais do que 10 SM.
   (b) Qual a seu gênero?
   (c) Qual a sua raça?
   (d) Onde você mora é favela ou bairro?
   (e) Tem falta de moradia onde você mora?
   (f) Seu local de trabalho fica perto, a media distância ou longe?

Exemplo 1.0.2. *A seguinte notícia apareceu no jornal português* **O Observador***: "Estágios atrás de estágios, falsos voluntariados, falsos recibos verdes e salários baixos. São estes alguns dos inimigos comuns dos jovens com formação superior que tentam ingressar no mercado de trabalho. São também temas protagonistas do livro* **Trabalho Igual, Salário Diferente***, de Francisco Fernandes Ferreira.*

Para o autor ... a luta é para que "se cumpra a Constituição e o Código de Trabalho" – isto numa altura em que a "contratação sem termo parece uma utopia para as novas gerações".." Observe que estão sendo relacionados 'características do estagio', e 'trabalho integral'.
Considere a problemática mencionada acima,

1. *Como vôçe definiría o problema numa pesquisa empírica* **comparativa** *(população alvo, formulação de objetivos, etc). (máximo 5 linhas)*
2. *Descreva e classifique (qualitativas ou quantitativas) as variáveis que vôçe incluiría no estudo. (máximo 6 variáveis)*
3. *Elabore um questionário para a pesquisa. (máximo 6 questões)*

Solução

1. Como você definiría o problema numa pesquisa empírica **comparativa** (população alvo, formulação de objetivos, etc).

   Problema: Caracterizar o significado do 'trabalho integral' nos jovens que enfrentam o primeiro trabalho ou estagio.

   População Alvo Os estudantes da UFMG dos cursos de Ciências Sociais, Engenharia e Biologia.

   Objetivo Principal Comparar o sentido que tem o 'trabalho integral' na população jovem que enfrentam o primeiro emprego/estagio.

   Objetivos Secundarios:
   - Relacionar a condição socioeconômica/classe social, o gênero e a raça com a noção de trabalho integral.
   - Relacionar a condição socioeconômica/classe social, o gênero e a raça com os estágios.
   - Relacionar a condição socioeconômica/classe social com os estudos.

2. Descreva e classifique (qualitativas ou quantitativas) as variáveis que você incluiría no estudo. (máximo 6 variáveis):

   (a) Variável: *Salario*. Nome: SAL , Tipo: Qualitativa ordinal, Valores: A (mais de 10 SM),B (5 à 10 SM),C (2 a 5 SM),D (menos de 2 SM)

   (b) Variável: *Gênero*. Nome: GEN , Tipo: Qualitativa nominal, Valores: H (homem), M (mulher).

   (c) Variável: *Raça*. Nome: RAZ , Tipo: Qualitativa nominal, Valores: B (branca), N (negra), P (parda).

   (d) Variável: *Curso de estudo*. Nome: CURSO , Tipo: Qualitativa nominal, Valores: C (Ciências Sociais), E (Engenharia), B (Biologia).

   (e) Variável: *Estagio*. Nome: DEF , Tipo: Qualitativa nominal, Valores: S (Sim), N (não).

(f) Variável: *Opinião sobre o trabalho integral*. Nome: TRAB , Tipo: Qualitativa nominal, Valores: R (ruim), B (bom), E (excelente).

3. Elabore um questionário para a pesquisa. (máximo 6 questões)

   (a) A qual faixa de salario (em salários mínimos) sua família pertence?
      i. De zero até 2 SM.
      ii. De de 2 SM até 5 SM.
      iii. De 5 SM até 10 SM.
      iv. Mais do que 10 SM.

   (b) Qual a seu gênero?

   (c) Qual a sua raça?

   (d) Que área de conhecimento você estuda? Ciências Sociais, Engenharia ou Biologia.

   (e) Tem oportunidade de fazer estagio onde você estuda?

   (f) Acha que ter um trabalho integral no inicio da carreira é ruim, bom ou excelente?

Capítulo 2

Amostragem

Exemplo 2.0.1. *Os chefes de família de uma comunidade de dois bairros são:*

1. **Bairro 1** *Anabela, Manuel, Arcêncio, Elísio, Francisco, Paulo, Manuel, Carlos, António, João, Ana Paula, José Manuel, José António, Jorge, José, Anabela.*

2. **Bairro 2** *Daniela, Douglas, Fabio A., Fabio E., Fabio P., Gabriel, Isabela, Isabella e Jonas.*

Usando a tabela de números aleátorios construa uma amostra estratificada proporcional de tamanho 10 a partir desta população

Solução

- No Bairro 1 há 16 chefes de família e no Bairro 2 há 9 famílias, formando um total de 25. Cada bairro será considerado um estrato, portanto, devemos calcular a proporção de cada bairro no total, para utilizar tal proporção na amostra.

  $$\text{Bairro 1: } \frac{16}{25} * 100 = 64\%$$

  $$\text{Bairro 2: } \frac{9}{25} * 100 = 36\%$$

  Iremos arredondar 64% para 60% e 36% para 40%. Utilizaremos tais proporções em uma amostra de tamanho 10. Ou seja, Do bairro 1 teremos 60% da amostra, que equivale a 6 pessoas, e do bairro 2 selecionaremos 40%, que equivale a 4 pessoas.

  Através da tabela de números aleatórios, escolheremos um critério para selecionar a amostra. iniciaremos obtendo os 6 entrevistados do Bairro 1.

  Um critério de escolha pode ser contar de 7 em 7 e ir pegando sétimo valor:

  – Cada pessoa deverá ser enumerada iniciando do número 1. Por conveniência a numeração será na ordem que os nomes aparecem. No bairro 1 teremos numeração de 1 a 16 e no 2 de 1 a 9. Obtendo a seguinte numeração:

Bairro 1: Anabela(1), Manuel(2), Arcêncio(3), Elísio(4), Francisco(5), Paulo(6), Manuel(7), Carlos(8), António(9), João(10), Ana Paula(11), José Manuel(12), José António(13), Jorge(14), José(15), Anabela(16).
Bairro 2: Daniela(1), Douglas(2), Fabio A.(3), Fabio E.(4), Fabio P.(5), Gabriel(6), Isabela(7), Isabella(8) e Jonas(9).

- Os 6 primeiros dígitos selecionados de 7 em 7 pertencerão ao bairro 1. Como nele há 16 membros, se um dígito selecionado for o número 1, ele deverá ser considerado junto com o seu dígito vizinho - por exemplo: em **1**2572 selecionamos o dígito 1 e, para que haja chances de selecionar algum número maior ou igual a 10 e menor que 16, devemos considerar seu dígito vizinho à nossa direita. Portanto, nesse exemplo, o número selecionado será o número 12: **12**572. Se o dígito vizinho ao número 1, caso selecionemos este valor, for maior que 6, formando assim um número maior que 16, utilizamos o número 1 - por exemplo: em 4 **1**726 não consideramos o dígito vizinho ao um, pois ao considerá-lo obtemos um número (17) que supera o tamanho da primeira amostra, que é igual a 16. Neste caso, como já explicado, consideramos apenas o dígito 1, ou seja, o primeiro membro da primeira amostra. Se o valor selecionado for maior que 1 (2,3,4,5,6,7,8 ou 9), deve-se coletar a pessoa equivalente a este número na amostra - por exemplo: em 56**4**91 o quarto membro da primeira amostra deve ser considerado ;
- Quando todos os 6 membros da primeira amostra forem sorteados, os quatro últimos dígitos selecionados, ainda respeitando o passo de seleção (7 em 7), pertencerão ao bairro 2. Como nele há somente 9 pessoas, deveremos utilizar o número equivalente (1,2,3,4,5,6,7,8 ou 9);
- Se um número sorteado já tiver sido selecionado anteriormente, desconsidere-o e siga o processo.

Seguindo a lógica descrita acima, os membros selecionados foram:

39634 6**2**349 740**8**8 65564 **16**379 197**13** 39153 6**9**459 179**8**6
24537 **14**595 350**5**0 40469 **2**7478 44**5**26 6733**1** 93365 5**4**526
30734 71571 83722 79712 25775 65178 07763 82928 31131

Bairro 1: Manuel, Carlos, Anabela, José Antônio, Antônio e Jorge.
Bairro 2: Fabio P, Douglas, Daniela, Fabio E.

Exemplo 2.0.2. *Considere as seguintes populações alvo, numa pesquisa de trabalho juvenil, escolha um tipo de amostragem (amostragem aleatória simples, sistemática, estratificada ou por conglomerados) para cada uma delas:*

1. *Jovens (20-24 anos) na região sul de Belo Horizonte, os bairros Savassi (4054 jovens), Magabeiras (608 jovens) e Belvedere (410 jovens).*

2. *A Região Pampulha: que se divide em oito Bairros (12362 jovens).*
3. *O conjunto popular Confisco na regional Pampulha: (668 jovens).*
4. *Belo Horizonte: 218000 jovens.*

Solução

1. **Amostragem aleatória simples:** A escolha da amostragem aleatória simples é justificada pela semelhança entre os jovens destes três bairros, ou seja, proporções diferentes de pessoas em cada bairro não afetarão o resultado final possibilitando o modo de amsotragem mais simples que consistem apenas em sortear os jovens dos três bairros.
2. **Amostragem por conglomerados:** É possível considerar que cada bairro da região da Pampulha seja representativo da região inteira, portanto, seria adequado considerar coda bairro um conglomerado, e selecionar apenas alguns bairros. Dentro de cada conglomerado(Bairro) poderiam se pensar em realizar uma amostragem aleatória simples.
3. **Amostragem sistemática:** A facilidade em mapear as residências dessa região, possibilitam uma esquematização da amostragem, basta sortear o primeiro lugar a ser amostrado, e a partir deste, seguir algum padrão pré especificado. Tal técnica nos da propriedades similares a amostragem aleat'oria simples, porém de forma mais rápida e simples.
4. **Amostragem estratificada:** A escolha da amostragem estratificada se dá pela variação das características dos jovens nas diferentes regiões da cidade, é possível definir estratos de acordo com os bairros ou regionais, de tal forma, cada jovem pertenceria apenas a um estrato. Tendo os estratos definidos, basta selecionar a quantidade de jovens nescessária dentro de cada estrato para formar a amostra final.

Exemplo 2.0.3. *Considere as seguintes populações alvo, escolha um tipo de amostragem (amostragem aleatória simples, sistemática, estratificada ou por conglomerados) para cada uma:*

(a) Na região Centro-Sul de Belo Horizonte, os bairros Savassi (46.522 habitantes), Mangabeiras (6.974 habitantes) e Belvedere (4.733 habitantes).
(b) A Região Pampulha: que se divide em oito bairros (141.853 habitantes).
(c) O conjunto popular Confisco na reginal Pampulha: (7.669 habitantes).
(d) Belo Horizonte: 2.412.937 habitantes.

***Resolução**:*

(a) ***Amostragem estratificada*** *- A escolha da amostragem estratificada é justificada pela representatividade ponderada das regiões que deve ser considerada, uma vez que a quantidade de*

habitantes varia bastante de região para região e esses tamanhos são discriminados ao lado do nome de cada uma delas. Em outras palavras, a amostra deve conter quantidades proporcionais de cada área, não sendo cabível, por exemplo, coletar mais dados no Belvedere do que na Savassi.

(b) ***Amostragem por conglomerados*** *- A escolha da amostragem por conglomerados é justificada pela homogeneidade da região da Pampulha. Para estudarmos tal região, basta coletar um grupo (cluster) que a represente, ou seja, coletamos todos os dados de apenas um dos oito bairros para representar a Pampulha.*

(c) ***Amostragem sistemática*** *- A escolha da amostragem sistemática é justificada pela facilidade em sistematizar a escolha dos elementos da amostra. Para realizar tal amostragem, conseguimos estabelecer quais elementos podemos colocar em nossa amostra de forma ordenada, devido à numeração dos apartamentos. Conseguimos, com isso, escolher o primeiro apartamento a ser entrevistado e definir os próximos com base em intervalos arbitrários. Por exemplo, podemos escolher o primeiro sendo o próximo o nono apartamento à frente, depois dele o décimo oitavo, após este o vigésimo sétimo à frente e assim por diante. Neste exemplo foi definido um intervalo de tamanho igual a nove, mas, como dito, podemos escolher essa amplitude.*

(d) ***Amostragem estratificada*** *- A escolha da amostragem estratificada é justificada pela heterogeneidade da população de Belo Horizonte. Ao fazer uma amostragem estratificada estamos considerando a representatividade de regiões com diferentes estruturas socioeconômicas.*

Capítulo 3

Estatística Descritiva

Exemplo 3.0.1. *A idade dos 10 ingressantes num certo ano no curso de pós-graduação em jornalismo de uma universidade foi o seguinte:*

$$22,\ 23,\ 23,\ 25,\ 21,\ 24,\ 20,\ 23,\ 20,\ 21$$

1. *Ache as frequências absolutas, relativas, e relativas acumuladas.*
2. *Grafique o histograma.*
3. *Ache a media, moda, mediana, variância e esvio padrão da idade.*
4. *Construa o boxplot.*

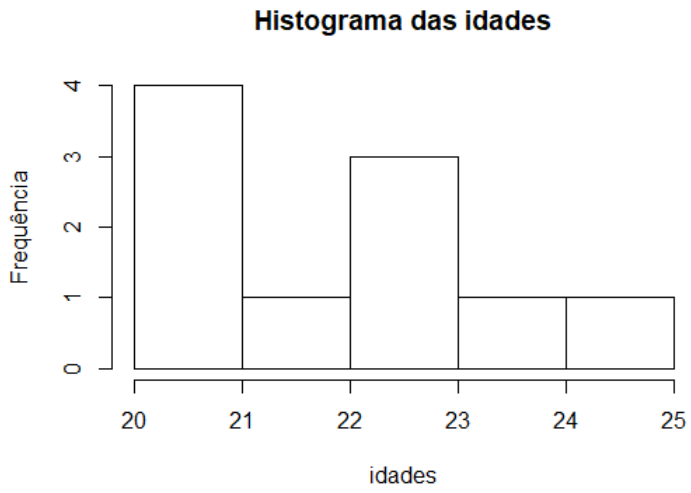
Solução

1. Ache as frequências absolutas, relativas, e relativas acumuladas.

   - A **frequência absoluta** é obtida através da **contagem** de vezes que cada elemento aparece na amostra, por exemplo, a idade 23 anos, apareceu 3 vezes, portanto, sua frequência é 3.
   - A **frequência relativa** é a **proporção** de vezes em que o elemento aparece, ou seja, a frequência absoluta divida pelo total. A idade 23, aparece 3 vezes, dividindo pelo total obtemos $\frac{3}{10} = 0.3$.
   - **frequência relativa acumulada** é **a soma cumulativa das frequência relativas**. Na da idade 23 é a soma das frequência das idades menores ou iguais a 23 (20,21,22,23); 0.2+0.2+0.1+0.3= 0.8.

Idades	Freq Absoluta	Freq Relativa	Freq Cumulativa
20	2	0.2	0.2
21	2	0.2	0.4
22	1	0.1	0.5
23	3	0.3	0.8
24	1	0.1	0.9
25	1	0.1	1.0

2. Grafique o histograma. O gráfico "histograma"deve conter as classes no eixo x e a frequência em que elas aparecem no eixo y. As barras devem ter a altura da frequência a que elas correspondem. Por exemplo, há quatro pessoas entre 20 e 21 anos, portanto a caixa correspondente a este intervalo corresponde à altura 4, especificada no eixo Y.



3. Ache a média, moda, mediana, variância e desvio padrão da idade.

   - A média éncontrada através da fórmula: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, que descreve a soma de todos os valores, divididos pela quantidade de valores. Temos então:
     Soma de todos os valores: $22 + 23 + 23 + 25 + 21 + 24 + 20 + 23 + 20 + 21 = 222$
     Soma dividida pela quantidade de valores: $\frac{222}{10} = 22.2$.
   - A moda é o elemento **mais frequente**, podemos observar a partir da tabela de frequêcias que o valor mais frequente é a idade 23, portanto a moda é igual a **23**.

- Para achar a mediana devemos **ordenar o conjunto de dados em ordem crescente**. A mediana é o valor que separa a metade maior e a metade menor da amostra, em termos mais simples,é o valor do meio de um conjunto. Para encontrar a mediana, devemos seguir a seguinte regra:

  Se o número de elementos na amostra for **par**, então a mediana é a **média dos dois valores centrais**. Se o numero de elementos na amostra for **ímpar**, a mediana é o **valor central**. Ordenando os valores temos:

  $$20,20,21,21,\mathbf{22,23},23,23,24,25.$$

  Os valores centrais são 22 e 23, a média entre eles é igual a $\frac{22+23}{2} = \mathbf{22.5}$.
- A variância é dada através da fórmula: $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  Primeiro calculamos todos os valores menos a média obtendo os seguintes valores: $(x_i - \bar{x})$

  $$\text{-0.2, 0.8, 0.8, 2.8, -1.2, 1.8, -2.2, 0.8, -2.2, -1.2;}$$

  Em seguida, elevamos todos os valores ao quadrado: $(x_i - \bar{x})^2$

  $$0.04,\ 0.64,\ 0.64,\ 7.84,\ 1.44,\ 3.24,\ 4.84,\ 0.64,\ 4.84,\ 1.44.$$

  Depois somamos todos os valores $\sum_{i=1}^{n}(x_i - \bar{x})^2$

  $$0.04+0.64+0.64+7.84+1.44+3.24+4.84+0.64+4.84+1.44 = 25.6$$

  Por fim dividimos pela quantidade de elementos(10),$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

  $$\frac{25.6}{10} = \mathbf{2.56}$$
- O desvio padrão é a **raiz quadrada da variância**. A raiz quadrada de 2.56 é igual é: igual a **1.6**.

4. Construa o boxplot.

   Para construir um boxplot, precisamos do **primeiro quartil**, da **mediana** e do **terceiro quartil**. A mediana já foi encontrada anteriormente, o primeiro e o terceiro quartil podem ser obtidos através da tabela de frequências. O primeiro quartil, é o valor que deixa 25% dos valores abaixo dele, ou seja, assim que ultrapassamos uma frequência acumulada de 0.25. O terceiro quartil é o valor que deixa 75% dos valores abaixo dele, ou seja, assim que ultrapassamos a frequência acumulada de 0.75. Portanto o primeiro quartil é igual a 21 e o terceiro é igual a 23.

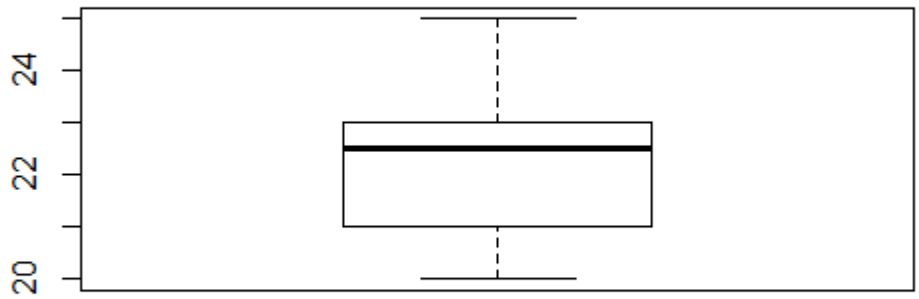
   A caixa do boxplot vai do primeiro quartil até o terceiro, e a mediana deve ser traçada dentro dela. **As linhas devem ser traçadas até os limites estabelecidos**:

   - O limite inferior é dado pelo maior valor entre o mínimo dos dados e $Q_1 - 1,5(Q_3 - Q_1)$.
   - O limite superior é dado pelo menor valor entre o máximo dados e $Q_3 + 1,5(Q_3 - Q_1)$.

Onde Q1 indica o primeiro quartil e Q3 indica o terceiro

Temos que o menor valor dos nossos dados é 20, e $21 - 1,5(23 - 21) = 18$, portanto nossa linha inferior é traçada até o número 20.

Temos que o maior valor dos nossos dados é 25, e $23 + 1,5(23 - 21) = 26$, portanto nossa linha superior será traçada até o número 25.



Exemplo 3.0.2. *O tempo, em horas, de trabalho diário dos 10 estagiários na área de comunicação de uma grande empresa foi o seguinte:*

$$4, 3, 5, 8, 6, 4, 4, 3, 6, 4$$

1. *Ache as frequências absolutas, relativas e relativas acumuladas.*
2. *Desenhe o histograma.*
3. *Ache a média, moda, mediana, variância, desvio padrão, coeficiente de variação do tempo.*
4. *Construa o boxplot.*

Solução

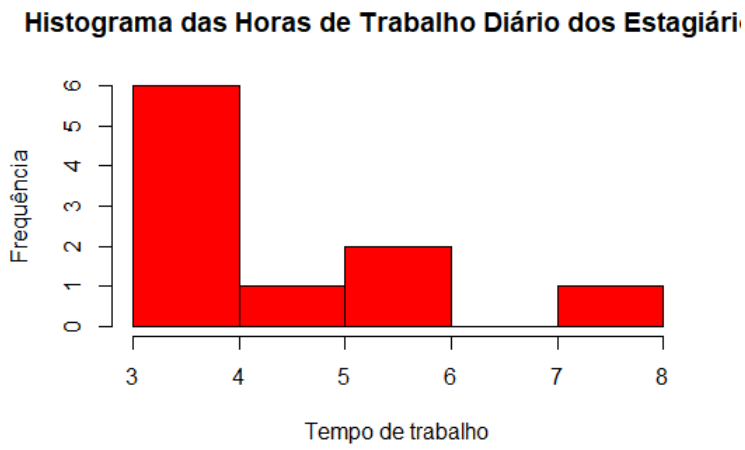
1. Ache as frequências absolutas, relativas e relativas acumuladas.

Horas	Frequência absoluta	Frequência relativa	Frequência relativa acumulada
3	2	0.2	0.2
4	4	0.4	0.6
5	1	0.1	0.7
6	2	0.2	0.9
8	1	0.1	1

   A **frequência absoluta** é o número de vezes que algo aparece em um conjunto. Se o número 3 apareceu duas (2) vezes em nosso conjunto, então sua frequência absoluta é igual a 2. A **frequência relativa** é a representatividade que a frequência absoluta tem no conjunto o qual ela pertence. Então, se o número 3 apareceu duas (2) vezes em um conjunto de dez (10) observações, sua frequência relativa é igual a $\frac{2}{10} = 0.2$. Faça esse procedimento para todos os outros valores e você irá conseguir achar suas frequências absolutas e relativas. Já a **frequência relativa acumulada** é a soma da frequência relativa atual com as frequências relativas anteriores. O número 3, como é o primeiro, possui frequência relativa acumulada igual a 0.2. O número quatro (4) possui frequência relativa acumulada igual a sua frequência relativa somada com a frequência relativa do número anterior (o número 3). Então a frequência relativa acumulada do número 4 é igual a $0.4 + 0.2 = 0.6$. Faça o mesmo procedimento para o restante dos números.

2. Desenhe o histograma.

   O gráfico "histograma"deve conter as classes no eixo x e a frequência em que elas aparecem no eixo y. A altura das barras devem corresponder à frequência em que as classes que elas representam aparecem. Por exemplo, há seis (6) estagiários que trabalham entre 3 e 4 horas por dia, portanto, a barra correspondente a este intervalo atinge à altura 6, especificada no eixo Y.



3. Ache a média, moda, mediana, variância, desvio padrão, coeficiente de variação do tempo.

   - Média:
     A média é uma medida de resumo que se calcula da seguinte forma: some todos os valores do conjunto e divida pelo número de itens que há nele. Portanto, a média do tempo de trabalho dos estagiários é: $\frac{4+3+5+8+6+4+4+3+6+4}{10} = 4,7$ horas
   - Moda:
     Moda é o valor que mais se repete no banco de dados. Portanto, a moda do tempo de trabalho dos estagiários é: 4 horas
   - Mediana:
     A mediana é o valor que divide o conjunto ordenado ao meio, ou seja, é o valor em que 50% dos números do conjunto ordenado são maiores que ele e 50% são menores. Para calculá-lo, **primeiramente ordenamos** o conjunto, depois **determinamos sua posição**. Para isto, se a quantidade total de elementos do conjunto for par, dividimos essa quantidade por 2 e consideramos o resultado dessa divisão por dois mais o próximo número. Por exemplo, o nosso conjunto possui 10 elementos, 10 é um número par, então $\frac{10}{2} = 5$ e, assim, a mediana será a média entre o quinto e o sexto elemento. Conjunto ordenado: 3,3,4,4,4,4,5,6,6,8. Quinto elemento: 4; Sexto elemento: 4; Mediana do tempo de trabalho dos estagiários: $\frac{4+4}{2} = \frac{8}{2} = 4$ horas.
     Se a quantidade total de elementos fosse ímpar, dividiríamos o número por 2 e arredondamos para cima. Por exemplo, se a quantidade total de elementos fosse igual a 9, a mediana seria o quinto elemento do conjunto ordenado, pois $\frac{9}{2} = 4,5$ e, arredondando para cima, encontramos o número 5.
   - Variância:
     A variância é uma medida que nos informa o quanto os dados variam em torno da média. Para encontrar a variância devemos calcular a média dos desvios de cada número com

relação à média do conjunto. Porém, cada desvio (a distância de cada número do conjunto até a média do mesmo) deve ser elevado ao quadrado para que, quando forem somados, não resulte em zero, já que a soma dos desvios negativos sempre cancela os desvios positivos. A fórmula da variância é: $Var = \sum_{x=1}^{n} \frac{1}{n}(x_i - \bar{x})^2$. Portanto, a variância do tempo de trabalho diário dos estagiários é: $\frac{1}{10}((4-4,7)^2+(3-4,7)^2+(5-4,7)^2+(8-4,7)^2+(6-4,7)^2+(4-4,7)^2+(4-4,7)^2+(3-4,7)^2+(6-4,7)^2+(4-4,7)^2) = \frac{22,1}{10} = 2,21h^2$

- Desvio Padrão:

  O desvio padrão é a raiz quadrada da variância. O sentido de calcular essa medida se justifica pelo fato de que a unidade de medida da variância é a unidade de medida original elevada ao quadrado, já que elevamos todos os desvios ao quadrado. Portanto, o desvio padrão do tempo de trabalho diário dos estagiários é: $\sqrt{var} = \sqrt{2,21} = 1,48$ horas

- Coeficiente de variação:

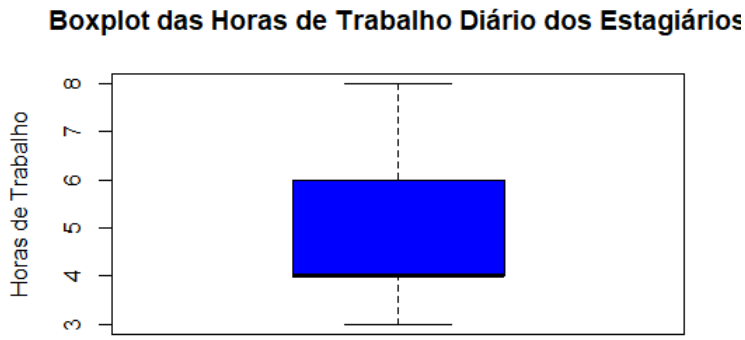
  O coeficiente de variação é uma medida que nos informa a representatividade do desvio padrão com relação à média. Para calcular essa medida basta dividir o desvio padrão da média e multiplicar esse resultado por 100. Ou seja, $cv = \frac{dp}{media}100$. Portanto, o coeficiente de variação do tempo diário dos estagiários é: $cv = \frac{1,48}{4,7}100 = 31,49\%$. Ou seja, o desvio padrão é igual a 31,49% da média.

  Observação: uma das vantagens de calcular o coeficiente de variação é poder comparar a variabilidade de conjuntos numéricos que possuem médias diferentes. Por exemplo, qual conjunto varia mais em torno da média: $c1 = 2, 4, 6$ ou $c2 = 20, 40, 60$? Se calcularmos o desvio padrão de cada conjunto chegaremos à conclusão de que c2 possui um desvio padrão maior. Mas tome cuidado, pois em compensação,sua média também é maior. Ao calcularmos o coeficiente de variação de cada um descobrimos que eles são iguais e, portanto, cada um desses conjuntos (c1 e c2) possuem a mesma variabilidade com relação à média.

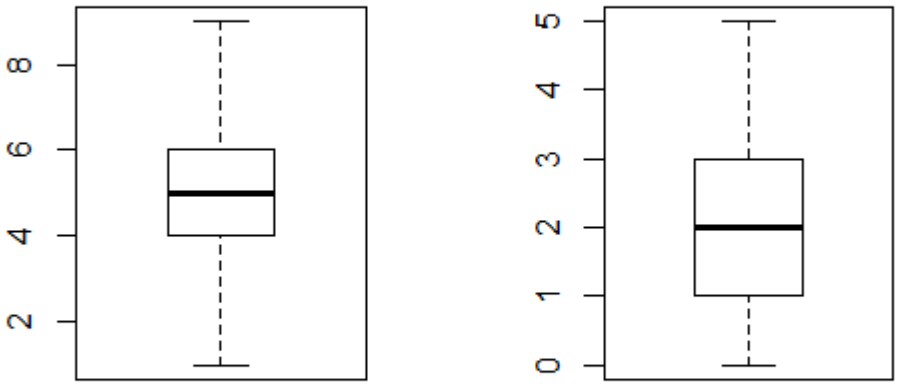
4. Construa o boxplot.

   O Boxplot é um gráfico que nos informa medidas de posição: o valor mínimo, os quartis, valor máximo. A primeira linha horizontal de todas representa o menor valor do conjunto (valor mínimo), que no nosso caso é o número 3: **3**,3,4,4,4,4,5,6,6,8. O primeiro quartil é a primeira linha horizontal da caixa e é o valor em que 25% dos números do conjunto ordenado estão abaixo dele. Para determinar sua posição basta calcular quanto é 25% de 10 (pois o conjunto numérico possui 10 números) e arredondar para cima, caso a posição não seja exata. Daí, o primeiro quartil é o número cuja posição no conjunto ordenado é $0,25x10 = 2,5$ mas, como arredondamos para cima, como dito, sua posição é o número 3, ou seja, o terceiro número no conjunto ordenado: 3,3,**4**,4,4,4,5,6,6,8. O segundo quartil é a mediana, valor em que 50% dos números do conjunto ordenado estão abaixo dele, e é representado pela linha mais escura (em negrito) dentro da caixa. Em nosso caso esse valor coincide com o primeiro quartil, ou seja, a mediana é igual a 4: 3,3,4,4,**4**,**4**,5,6,6,8, daí $\frac{4+4}{2} = \mathbf{4}$. O terceiro quartil é a última linha horizontal da caixa e é o valor em que 75% dos números do conjunto ordenado estão abaixo dele. Para determinar sua posição basta calcular quanto é 75% de 10 (pois o conjunto numérico

possui 10 números) e arredondar para cima, caso a posição não seja exata. Daí, o último quartil é o número cuja posição no conjunto ordenado é $0,75x10 = 7,5$ mas, como arredondamos para cima, como dito, sua posição é o número 8, ou seja, o oitavo número no conjunto ordenado: 3,3,4,4,4,4,5,**6**,6,8.Já a última linha horizontal de todas representa o maior valor do conjunto (valor máximo), que no nosso caso é o número 8: 3,3,4,4,4,4,5,6,6,**8**. Portanto, o boxplot das horas de trabalho dos estagiários é representado da seguinte maneira:



Exemplo 3.0.3. *Numa pesquisa sobre o tempo de trabalho dos estagiários de duas empresas A e B obteve-se os seguintes boxplots:*



1. *Qual a mediana em ambos gráficos ? Comente o significado da diferencia entre as medianas.*
2. *Compare a distancia interquartílica(DI) entre os dois gráficos, descreva o significado desta (DI) em ambos gráficos.*
3. *Comente a assimetria (se tiver) entre os gráficos.*

Solução

1. Qual a mediana em ambos gráficos ? Comente o significado da diferencia entre as medianas.

   A mediana é sempre a linha que está dentro da caixa do boxplot. A caixa, no caso, é o retângulo. Portanto, a mediana do primeiro gráfico (empresa A) é igual a 5. Já a mediana do segundo gráfico (empresa B) é igual a 2.

2. Compare a distancia interquartílica(DI) entre os dois gráficos, descreva o significado desta (DI) em ambos gráficos.

   A distância interquartílica é a distância entre o primeiro quartil e o terceiro quartil. O termo quartil diz respeito à valores que dividem o conjunto numérico ordenado em quatro partes. Existem, portanto, 3 quartis. O primeiro quartil é o valor que está acima de 25% dos dados ordenados, o segundo quartil é o valor que está acima de 50% dos dados ordenados e o terceiro e último quartil é o valor que está acima de 75% dos dados ordenados. Para que a explicação fique mais clara, imagine um conjunto numérico ordenado de 0 a 100, ou seja:

0,1,2,3,4,...,96,97,98,99,100. O primeiro quartil é 25, o segundo quartil é 50 e o terceiro quartil é 75. Nem todo conjunto, como este, acharemos quartis inteiros. Se o conjunto ordenado for, por exemplo: 2,3,6,9,12,34, que é um conjunto de tamanho igual a 6, o primeiro quartil será a o número da posição $0,25x6 = 1,5$; daí arrendondamos 1,5 para 2 e, dessa forma, atente-se, o primeiro quartil é o segundo dado ordenado, que é o número 3. Sempre arredondamos para cima. O segundo quartil será o número da posição $0,5x6 = 3$, que é o terceiro dado ordenado (6) e o terceiro quartil será o número da posição $0,75x6 = 4,5$, arredondamos para 5, e será o quinto dado ordenado (12). Visto o conceito de quartil, agora vamos identificá-los no boxplot. A primeira linha horizontal da caixa (a linha que fecha a parte de baixo do retângulo) é o primeiro quartil, a linha horizontal que se localiza dentro da caixa é o segundo quartil, ou seja, a mediana. A última linha horizontal (a linha que fecha a parte de cima do retângulo) é o terceiro quartil. Como, repito, a distância interquartílica é a distância entre o primeiro quartil e o terceiro quartil, a distância interquartílica da empresa A é $6 - 4 = 2$, enquanto que da empresa B é $3 - 1 = 2$. Ou seja, as distâncias interquartílicas entre as empresas são iguais.

3. Comente a assimetria (se tiver) entre os gráficos.

   Podemos ver que a Empresa A apresenta dados simétricos. Chegamos a essa conclusão ao observar a equidistância entre o valor mínimo e o máximo à mediana. No caso da Empresa B o mesmo não acontece, há uma assimetria, pois podemos ver que a mediana está mais próxima do valor mínimo.

Capítulo 4

Probabilidade

Exemplo 4.0.1. *Numa pesquisa feita com $n = 10$ amantes do cinema de uma cidade do interior encontramos que 5 deles preferem um filme de drama, 1 de ação e 4 de terror.*

1. *Ache a probabilidade p de preferir um filme de drama.*
2. *Numa exibição especial, nessa cidade, do filme 'Que horas ela volta?' se venderam 20 ingressos, qual a probabilidade de no máximo 3 dos que iram assistir o filme tenham preferencia pelo drama.*

Solução

1. **Ache a probabilidade p de preferir um filme de drama.**

   Inicialmente, precisamos definir o espaço amostral e o nosso Evento de interesse. O espaço amostral consiste em todos os possíveis resultados.Se estamos selecionando pessoas no cinema, e estamos observando qual o gênero preferido dela, nosso espaço amostral consiste no possíveis gêneros que a pessoa pode ter preferência.

   Espaço Amostral = {Drama, Ação e Terror}.

   Um evento nada mais é que eu subconjunto do espaço amostral, pode ser apenas um elemento ou vários. Por exemplo, se estivermos interessados em ação e terror, nosso evento seria as pessoas preferirem terror ou preferir ação. No nosso problema, estamos interessados no gênero drama.

   Evento de interesse = Drama.

   A nossa amostra é o resultado das observações. Ou seja, o gênero de preferência de cada pessoa que foi perguntada é um elemento da amostra.

   Amostra:{...}.

   A probabilidade será calculada através da divisão entre o número de vezes que o evento de interesse acontece pelo número total de elementos na amostra. De tal forma, faremos a seguinte divisão:

$$\frac{\text{Número de vezes que Drama aparece na amostra}}{\text{Numero total de elementos na amostra}}$$

Temos então que: $p =$ Probabilidade (drama) $= \frac{5}{10} = 0.5$

2. **Numa exibição especial, nessa cidade, do filme 'Que horas ela volta?' se venderam 20 ingressos, qual a probabilidade de no máximo 3 dos que iram assistir o filme tenham preferencia pelo drama.**

   Novamente precisamos definir nosso evento de interesse e o nosso espaço amostral. Nesta caso estamos interessados não mais na preferência de uma pessoa, e sim no número de pessoas que tem preferência pelo gênero drama.

   Evento de interesse = Nº de pessoas na amostra que preferem drama;

   Agora, o nosso espaço amostral, deixa de ser os gêneros e passa a ser a quantidade de pessoas que preferem o gênero Drama, pode ser que nenhuma pessoa tenha preferência por drama, ou que todos os 20 tenham preferência por drama. de tal forma, todos os valores entre 0 e 20 são possíveis de acontecer.

   Espaço amostral = { 0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20}

   Nosso interesse no momento é a variável aleatória **X = nº de pessoas que preferem drama**. Vamos destacar alguns pontos importantes:

   - È razoável considerar que todos as pessoas presentes tem a mesma probabilidade de ter preferencia pelo gênero Drama.(observações identicamente distribuídas)
   - É razoável considerar também que o gênero preferido de uma pessoas não influencia no gênero preferido de outra pessoa(independência nas observações.)

   Considerando os dois pontos descritos acima, é razoável pensar na distribuição binomial. Um modelo binomial consiste numa sequência de observações que tem apenas duas possibilidades(preferir drama ou não preferir) e que todas observações tenham a mesma probabilidade. Usaremos então um modelo binomial com parâmetros p = 0.5 (probabilidade de uma pessoa preferir drama) e n= 20 (número de pessoas).

   $$X \sim bin(n = 20, p = 0.5)$$

   A função de de probabilidades desta variável com distribuição binomial será dada por:

   $$P(X = k) = \frac{20!}{k!(20-k)!} 0.5^k (1 - 0.5)^{20-k}, k = 0 \ldots, 20$$

A função acima indica que a probabilidade da variável aleatória X ser igual a um valor k é o resultado da operação.

O símbolo "!"após um valor indica a operação fatorial, que consiste em multiplicar o valor por todos os valore anteriores, por exemplo: $5! = 5 \times 4 \times 3 \times 2 \times 1$.

0.5^k indica a probabilidade de observarmos k vezes o evento de interesse na amostra.

$(1 - 0.5)^{20-k}$ indica a probabilidade do evento que não é de interesse ocorrer 20 - k vezes, completando a amostra de tamanho 20.

$\frac{20!}{k!(20-k)!}$ serve para contar de quantas maneiras distintas as observações podem ocorrer.

Como desejamos calcular a probabilidade de no máximo 3 pessoas preferirem drama, deveremos considerar todas as possibilidades. Ou seja, quando nenhuma pessoa preferir drama e quando 1 , 2 e 3 pessoas preferirem.A probabilidade desejada será então a soma da probabilidade de todas as possibilidades:

$$P(X \leq 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$

- $P(x = 0) = \frac{20!}{0!(20-0)!} 0.5^0 (1 - 0.5)^{20-0} = (1)(1)(0.0000009)0.0000009$
- $P(x = 1) = \frac{20!}{1!(20-1)!} 0.5^1 (1 - 0.5)^{20-1} = (20)(0.5)(0.00000195) = 0.000019$
- $P(x = 2) = \frac{20!}{2!(20-2)!} 0.5^2 (1 - 0.5)^{20-2} = (190)(0.25)(0.000038) = 0.00018$
- $P(x = 3) = \frac{20!}{3!(20-3)!} 0.5^3 (1 - 0.5)^{20-3} = (1140)(0.125)(0.0000076) = 0.00108$

Somando tais probabilidades, obtemos a probabilidade de no máximo 3 preferirem drama.

$$P(X \leq 3) = 0.0000009 + 0.000019 + 0.00018 + 0.00108 = 0.0012$$

Temos então que $P(X \leq 3 = 0.0012)$, ou seja, numa amostra de 20 pessoas que foram assistir a edição especial, a probabilidade de que no máximo 3 tenham preferência pelo gênero drama é 0.0012.

Exemplo 4.0.2. *O tempo diário de ocupação por pessoa da biblioteca de uma faculdade tem distribuição Gaussiana com media 2,5 horas e desvio padrão 0.8. Calcule a probabilidade de que as pessoas que usam a biblioteca fiquem nela entre 1,5 e 2,5 horas.*

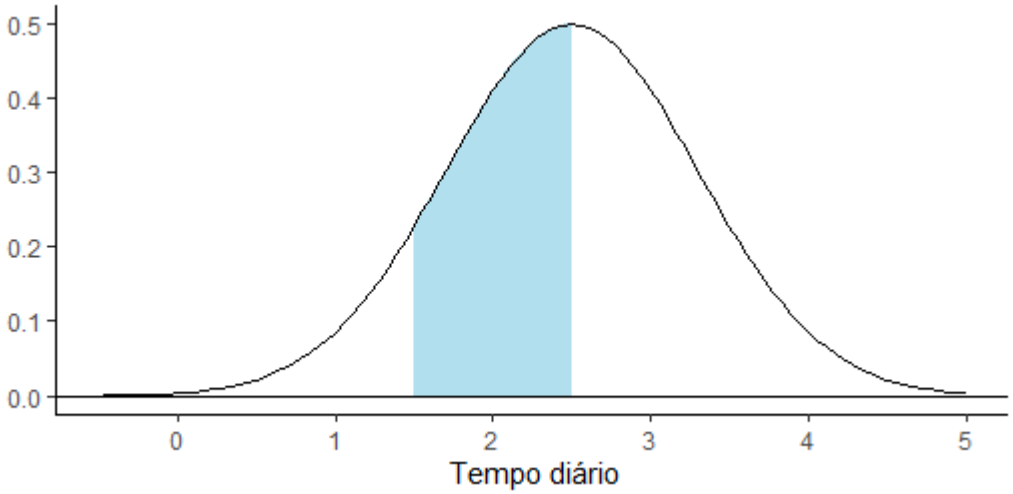
Solução

O primeiro passo é definir a variável aleatória do problema:

X = O tempo diário de ocupação por pessoa da biblioteca de uma faculdade

$$X \sim Normal(\mu = 2.5, \sigma = 0.8)$$

Quando estamos interessados em calcular uma probabilidade, estamos na verdade calculando uma área. Como desejamos encontrar $P(1.5 \leq X \leq 2.5)$, estamos interessados em encontrar a área entre os pontos 1.5 e 2.5, como destacadona figura abaixo:



Com as especificações da distribuição normal, tal área seria calculada através de uma integral muito complicada. Para facilitar este processo utiliza-se de uma padronização, que tem como objetivo, transformar a distribuição atual em uma distribuição normal, com média 0 e variância 1, para a qual as probabilidade estão tabeladas. Tal distribuição é comumente denotada por Z.

A padronização é feita através da seguinte forma:

$$Z = \frac{X - \mu}{\sigma}$$

Todo valor que tivermos terá seu valor equivalente na distribuição normal padronizada, e através deste valor equivalente, utilizamos a tabela para descobrir a probabilidade.

Como estamos interessados em encontrar a probabilidade de um intervalo de valores acontecer, podemos utilizar a subtração de probabilidades. Se pegarmos tudo que vem antes de um valor A (o maior valor), e subtraímos tudo que vem antes de um valor B(o menor valor), obtemos um intervalo entre o valor A e o valor B. Ou seja se pegarmos toda a área antes do valor 2.5 e subtrairmos a área antes do valor 1.5, ficamos com a área do intervalo entre 1.5 e 2.5. $P(1.5 \leq X \leq 2.5) = P(X \leq 2.5) - P(X \leq 1.5)$. Para realizar o cálculo das probabilidades

deveremos padronizar os valores, como mostrado anteriormente e procurar pelas probabilidade ta tabela da distribuição normal padrão.

Teremos então :

$$
\begin{aligned}
&P(1.5 \leq X \leq 2.5) \\
&= P(\frac{1.5-2.5}{0.8} \leq \frac{X-\mu}{\sigma} \leq \frac{1.5-2.5}{0.8}) \\
&= P(-1 \leq Z \leq 0) \\
&= P(Z \leq 0) - P(Z \leq -1)
\end{aligned}
$$

Através da tabela da distribuição normal padrão(Tabela Z) é possível obter as probabilidades. È importante frisar que a tabela normal pode ser apresentada em diversos formatos, então é importante observar qual área ela está especificando. Se estivermos utilizando a tabela que nos dá a área anterior á um ponto, basta procurar o valor x unindo extremidades e a probabilidade estará no interior da tabela:

A área abaixo do valor 0 é encontrada na tabela a seguir:

Tabela da Distribuição Normal Padrão
P(Z<z)

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177

Vale ressaltar que a valor Z deverá ser encontrado, unindo a primeira coluna com a primeira linha, A primeira coluna exibe o valor e sua primeira casa decimal, e a primeira linha especifica a segunda casa decimal(Basta somar a linha com a coluna). Por exemplo, a probabilidade de z ser menor que 1.27, é encontrada cruzando a linha equivalente ao valor 1.2, com a coluna equivalente a 0.07 (obtendo probabilidade igual a 0.8980).

A área abaixo do valor -1 é encontrada na tabela a seguir:

$P(Z<z)$

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367

Temos então que a probabilidade de Z ser menor que 0 é igual0.5,e a probabilidade de Z ser menor que -1 é 0.1587. Realizando subtração especificada teremos: 0.5-0.1587 = 0.3414.

Cada valor de Z foi obtido através da nossa distribuição com média 2.5 e desvio padrão 0.8. Como o valor 2.5 gerou o valor 0 na distribuição normal padrão, $P(X \leq 2.5 = 0.5)$, e como o valor 1.5 gerou o valor -1, $P(X \leq 1.5 = 0.1587)$, então, $P(1.5 \leq X \leq 2.5) = P(X \leq 2.5) - P(X \leq 1.5) = 0.5 - 0.1587 = 0.3414$

Exemplo 4.0.3. *Em um curso de graduação em uma faculdade sabemos que o número de favoráveis ao trabalho de campo é igual a 70% e 30% são contra.*

1. *Apresente a distribuição de probabilidade do número de alunos favoráveis quando selecionamos uma amostra aleatória (simples) de 6 alunos.*

2. *Chamando de X o número de alunos favoráveis nessa amostra de tamanho igual a 6, ache a esperança de X.*

Fórmulas

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, k = 0, \ldots, n$$

Solução

1. **Apresente a distribuição de probabilidade do número de alunos favoráveis quando selecionamos uma amostra aleatória (simples) de 6 alunos**

   Uma distribuição de probabilidade descreve o quão provável é encontrar os possíveis valores de uma variável aleatória. Em nosso exemplo, a **variável aleatória** é o **número de alunos favoráveis** que podemos encontrar em uma amostra de 6 alunos. Podemos encontrar dentro dela: 0 favoráveis, 1 favorável, 2 favoráveis, 3 favoráveis, 4 favoráveis, 5 favoráveis ou 6 favoráveis, não sendo possível, portanto, encontrar outros valores (ex: -1,7,8,...). Nesse sentido, basta encontrarmos a função que nos informa as probabilidades de ocorrência dos possíveis números de favoráveis (0 a 6) e calcularmos essas probabilidades para cada um desses possíveis valores (P(X=0), P(X=1),...,P(X=6)). Como encontrar essa função? Bom, sabemos que a distribuição de probabilidade que nos informa as chances de encontrar uma soma (em nosso caso, a soma de favoráveis dentro da nossa amostra, que pode resultar em 0,1,2...,6) é a **distribuição binomial**. Sua fórmula é:

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, k = 0, \ldots, 6.$$

   Antes de usá-la, vamos entendê-la um pouco melhor. Primeiramente, vamos identificar e descrever cada elemento nessa função.

   - **P(X=k)** significa a probabilidade de encontrarmos k observações dentro de um grupo de elementos cujo tamanho é o maior valor de k possível ,em nosso caso o maior valor possível para k é 6. Sendo assim, k pode ser igual a 0,1,...,6. Então, P(X=0), por exemplo, é a probabilidade de encontrarmos 0 favoráveis em um grupo de 6 alunos. P(X=1) é a probabilidade de encontrarmos 1 favorável em um grupo de 6 alunos e assim por diante até P(X=6), que é a probabilidade de encontrarmos 6 favoráveis em um grupo de 6 alunos (todos favoráveis).
   - $p^k(1-p)^{n-k}$ significa acontecer um evento k vezes (p^k) e não acontecer esse mesmo evento n-k vezes ($(1-p)^{n-k}$). Esse evento tem probabilidade igual a p de acontecer, portanto, tem probabilidade 1-p de não acontecer. No caso, n é o número total de itens na amostra, o que evidencia também o maior k possível, que é igual a n. Em nosso problema, n=6. Para ficar mais clara a explicação, vamos dar valor ao p e ao k. Para que observemos 2 favoráveis (k=2) em um grupo de 6 alunos (n=6), sendo que a probabilidade de encontrar um favorável na universidade é de 0,7 (p=0,7), é necessário acontecer o seguinte: observamos um favorável **e** um favorável **e** um desfavorável **e** um desfavorável **e** um desfavorável **e** um desfavorável =

$$0,7 \times 0,7 \times 0,3 \times 0,3 \times 0,3 \times 0,3 = 0,7^2 0,3^4 = p^k(1-p)^{n-k} \tag{4.0.1}$$

     Repare que, como a chance de observar um favorável na universidade é de 0,7 (70%), a chance de não observar é $1 - 0,7 = 0,3$ (30%). Como queremos determinar a chance de,

em um grupo com n=6 pessoas, observar k=2 favoráveis (p=0,7) e n-k = 6-2 = 4 não favoráveis (p=0,3), devemos calcular a multiplicação apresentada acima que é generalizada pela expressão $p^k(1-p)^{n-k}$

- Por fim, veja que a multiplicação apresentada no item anterior pode ser escrita de outras formas, por exemplo:

$$0,7 \times 0,3 \times 0,3 \times 0,3 \times 0,7 \times 0,3 = p^k(1-p)^{n-k} \tag{4.0.2}$$

Mas quantas de formas possíveis podemos escrever essa multiplicação ? Essa quantidade é:

$$\frac{\underline{\textbf{n!}}}{\underline{\textbf{k!(n - k)!}}} \tag{4.0.3}$$

Esse é o primeiro termo da fórmula, que multiplica $p^k(1-p)^{n-k}$, como podem ver. Essa expressão nos informa o número de combinações possíveis de $p^k(1-p)^{n-k}$. É importante notar que esse número de combinações depende de n e k. No item anterior exemplifiquei com k=2 alunos favoráveis e, como nossa amostra tem tamanho n=6 alunos, o número de combinações possíveis para $p^k(1-p)^{n-k} = 0,7^2 0,3^4$ neste caso é igual a $\frac{6!}{2!(6-2)!} = \frac{6!}{(2!)(4!)} = \frac{6\times5\times4\times3\times2\times1}{(2\times1)\times(4\times3\times2\times1)} = 15$. Lembre-se que a sintaxe "!"significa fatorial. 0! = 1, por definição. $1! = 1$, $2! = 2 \times 1$, $3! = 3 \times 2 \times 1$, $4! = 4 \times 3 \times 2 \times 1$, $5 = 5 \times 4 \times 3 \times 2 \times 1$ e assim por diante. Agora, sabendo usar a fórmula, basta calcular todas as probabilidades de números de alunos favoráveis (0 a 6) e assim encontramos a **distribuição de probabilidade de X, que é:**

- $P(X=0) = \frac{6!}{0!(6-0)!}0,7^0(1-0,3)^{6-0} = 0,0007$
- $P(X=1) = \frac{6!}{1!(6-1)!}0,7^1(1-0,3)^{6-1} = 0,01$
- $P(X=2) = \frac{6!}{2!(6-2)!}0,7^2(1-0,3)^{6-2} = 0,06$
- $P(X=3) = \frac{6!}{3!(6-3)!}0,7^3(1-0,3)^{6-3} = 0,18$
- $P(X=4) = \frac{6!}{4!(6-4)!}0,7^4(1-0,3)^{6-4} = 0,32$
- $P(X=5) = \frac{6!}{5!(6-5)!}0,7^5(1-0,3)^{6-5} = 0,30$
- $P(X=6) = \frac{6!}{6!(6-6)!}0,7^6(1-0,3)^{6-6} = 0,11$
- 0, caso contrário (caso k seja diferente de 0,1,2,3,4,5 ou 6)

Atenção:

- Qualquer número elevado a zero é igual a um. Ex: $1^0 = 1$, $2^0 = 1$, $50^0 = 1$, ...
- $0! = 1$, $1! = 1$, $2! = 2 \times 1$, $3! = 3 \times 2 \times 1$, ...

2. **Chamando de X o número de alunos favoráveis nessa amostra de tamanho igual a 6, ache a esperança de X.**

   A esperança de uma variável aleatória é o **resultado esperado** dessa variável considerando sua probabilidade de ocorrência e o tamanho da amostra observada. Por exemplo, se a probabilidade de observar um aluno favorável é igual a 70%, em uma turma com 100 alunos **esperamos** que haja 70 favoráveis. Já que em nosso problema n = 6, a esperança de X é igual a $6 \times 0,7 = \mathbf{4,2}$. Outra forma de calcular a esperança é: $\sum_{k=1}^{n} k_i.P(X = k_i)$, ou seja, a soma das observações multiplicadas pela probabilidade da ocorrência de cada uma delas. Em nosso exemplo, seria: $0.P(X = 0) + 1.P(X = 1) + 2.P(X = 2) + 3.P(X = 3) + 4.P(X = 4) + 5.P(X = 5) + 6.P(X = 6) = 4,2$

Exemplo 4.0.4. *Numa pesquisa com um conjunto de $n = 100$ candidatos a um estagio obteve-se a seguinte informação:*

Sexo	Número de estágios			
	Nenhum	Um ou dois	Três ou mais	Total
Masculino	5	10	35	50
Feminino	35	10	5	50
Total	40	20	40	100

1. *Ache as seguintes probabilidades:*
   - *P(Ter feito pelo menos um estágio | Ser mulher)*
   - *P(Ter feito no máximo dois estágios e ser homem)*
   - *P(Ser homem | Nenhum estágio anterior)*
2. *Os eventos 'Ser mulher' e 'Nenhum estágio anterior' são independentes?*

Solução

1. **Ache as seguintes probabilidades:**
   - *P(Ter feito pelo menos 1 estágio | Ser mulher)*
     Queremos calcular qual é a probabilidade de uma pessoa ter feito mais de um estágio dado que essa pessoa é uma mulher. Em outras palavras, imagine que selecionamos ao acaso uma pessoa dentre os 100 candidatos e descobrimos que ela é do sexo feminino. Todavia, não sabemos de seu histórico como estagiária, temos conhecimento apenas de seu sexo. A pergunta é: qual é a probabilidade dessa mulher ter feito no mínimo um estágio (um estágio ou mais) ? Para calcularmos essa probabilidade devemos considerar apenas o total de mulheres, que é igual a 50 (veja o total da linha das 'mulheres'), como podemos ver na

tabela, já que sabemos que a pessoa selecionada é do sexo feminino. Repare que o grupo das mulheres que fizeram um ou dois estágios possui 10 membros, e o grupo das mulheres que fizeram três ou mais estágios possui 5 membros. Portanto, o total de mulheres que fizeram no mínimo um estágio é igual a 10 + 5 = 15. Ora, a probabilidade de que mulher selecionada tenha feito no mínimo um estágio é a probabilidade de que ela pertença a esse grupo de 15 mulheres, sendo que, como dito, 10 dessas 15 fizeram um ou dois estágios e 5 dessas 15 fizeram três ou mais estágios. Como o total de mulheres é igual a 50, a probabilidade de uma pessoa ter feito no mínimo um estágio sabendo que essa pessoa é uma mulher é igual a $\frac{15}{50} =$ ***0.3*** *ou* ***30%***

- *P(Ter feito no máximo dois estágios e ser homem)*

  Repare que o grupo de pessoas que satisfazem essas duas condições, ter feito no máximo dois estágios e ser homem, tem tamanho igual 15. Observe que 5 desses 15 são homens que nunca estagiaram (veja o primeiro quadradinho da linha dos 'homens') e 10 desses 15 são homens que tiveram um ou dois estágios (veja o segundo quadradinho da linha dos 'homens'). Portanto, como 15 pessoas satisfazem essas duas condições, a probabilidade de selecionar ao acaso uma pessoa pertencente a esse grupo é de $\frac{15}{100} =$ ***0.15*** *ou* ***15%***. *Talvez você esteja se perguntando por quê não é* $\frac{15}{50}$*, assim como na questão anterior. Entretanto, é de suma importância identificar a diferença entre as duas questões. Na questão anterior já sabemos que a pessoa selecionada é uma mulher, o que nos obriga a restringir o total a 50, pois há 50 mulheres entre os 100 candidatos. Já nesta questão queremos saber qual é a chance de nesse grupo total de 100 candidatos encontrar um indivíduo que satisfaça a condição de ter feito no máximo dois estágios e ser homem.*

- *P(Ser homem | Nenhum estágio anterior)*

  Queremos calcular qual é a probabilidade de uma pessoa ser homem dado que essa pessoa nunca fez estágio. Em outras palavras, imagine que selecionamos ao acaso uma pessoa dentre os 100 candidatos e descobrimos que ela nunca fez estágio. Todavia, não sabemos seu sexo, temos conhecimento apenas de sua experiência com estágio (esta suposição parece estranha, mas não se preocupe, o objetivo dela é apenas deixar a explicação mais clara). A pergunta é: qual é a probabilidade de que essa pessoa que nunca estagiou ser um homem ? Para calcularmos essa probabilidade devemos considerar apenas o total de pessoas que nunca estagiaram, que é igual a 40 (veja o primeiro quadradinho da última linha), como podemos ver na tabela, já que sabemos que a pessoa selecionada nunca fez estágio. Repare que do grupo dos que nunca tiveram estágio, 5 são homens . Portanto, a chance de que essa pessoa seja homem sabendo que ela nunca estagiou é $\frac{5}{40} =$ ***0.125*** *ou* ***12,5%***

2. **Os eventos 'Ser mulher' e 'Nenhum estágio anterior' são independentes?**

   Sabemos que a probabilidade de dois eventos independentes acontecerem ao mesmo tempo é igual ao produto (resultado de uma multiplicação) entre as probabilidades individuais. Em outras palavras, se 'ser mulher' e 'nenhum estágio anterior' forem eventos independentes, a probabilidade de que esses dois eventos ocorram ao mesmo tempo é igual a P('ser mulher')

x P('nenhum estágio anterior'). Sabemos que, como há 50 mulheres entre 100 candidatos, a probabilidade de 'ser mulher' é igual a 0,5 (50%). Além disso, como há 40 pessoas que nunca estagiaram entre 100 candidatos, a probabilidade do evento 'nenhum estágio anterior' é igual a 0,4 (40%). Portanto, a probabilidade de que esses dois eventos ocorram ao mesmo tempo ***se forem independentes*** *é igual a* ***0,5x0,4 = 0,2*** *ou* ***20%****. Como podemos ver, o tamanho grupo de pessoas que são mulheres e nunca estagiaram é igual a 35 (veja o primeiro quadradinho da linha das 'mulheres'). Então,* ***a probabilidade real*** *de observarmos, nesse grupo de 100 pessoas, esses dois eventos acontecerem ao mesmo tempo é de* $\frac{35}{100} =$ ***0.35*** *ou* ***35%****. Por fim, como* ***0,2*** *é diferente de* ***0,35****, os eventos 'ser mulher' e 'nenhum estágio anterior'* ***não são independentes****.*

Exemplo 4.0.5. *Suponha 20% das vezes você contesta uma mensagem postada num grupo de Whatsapp. Suponha que foram postados $n = 10$ mensagens no grupo e chame de $X =$'número de contestações feitas'.*

1. *Qual a probabilidade de que não tenha nenhuma contestação.*
2. *Qual a probabilidade de que tenha contestado todas as mensagens.*
3. *Apresente a distribuição de probabilidades da variável aleatória X.*
4. *Ache o valor esperado e a variância da variável aleatória X?*

Solução

1. **Qual a probabilidade de que não tenha nenhuma contestação.**

   Se em 20%das vezes a mensagem pode é contestada a **probabilidade de contestar uma mensagem é igual 0.2**, por consequência, a **probabilidade da mensagem não ser contestada é 0.8**, pois estas são as únicas duas possibilidades. È razoável supor que a contestação de uma mensagem não depende de outra.

   Para calcular a probabilidade de não haver nenhuma contestação, é preciso considerar a probabilidade cada uma das 10 mensagens não ser contestada. Como a probabilidade de uma mensagem não ser contestada é 0.8 basta multiplicas este valor 10 vezes:

   $$0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.8 = 0.8^{10} = 0.107$$

   Portanto a probabilidade de não haver contestação é 0.107

2. **Qual a probabilidade de que tenha contestado todas as mensagens.**

   Para calcular a probabilidade de contestar todas as mensagens é preciso considerar a probabilidade a probabilidade de cada uma das 10 mensagem ser contestada. Como esta probabilidade é igual a 0.2, basta multiplicar 0.2 dez vezes.

$$0.2 \times 0.2 \times 0.2 \times 0.2 \times 0.2 \times 0.2 \times 0.2 \times 0.2 \times 0.2 \times 0.2 = 0.2^{10} = 0.0000001024$$

   A probabilidade de não haver contestação é 0.0000001024.

3. **Apresente a distribuição de probabilidades da variável aleatória X.**

   A variável aleatória X é definida como X= "Número de constatações feitas". Trata-se de 10 mensagens independentes em que só é possível observar duas respostas em cada observação. Como X é a contagem de constatações feitas, podemos considerar que X segue **distribuição binomial com parâmetros p= 0.2 e n=10.** A função de probabilidade da distribuição binomial é dada por:

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, k = 0, .., n$$

   - Lembrando que o símbolo fatorial indica que o valor deve ser multiplicado por todos os seus antecessores($3! = 3 \times 2 \times 1 = 6$).
   - p^k indica a probabilidade de observarmos o evento de interesse k vezes
   - $(1-p)^k$ indica a probabilidade de não observarmos o evento de sucesso no restante das vezes.
   - O termo $\frac{n!}{k!(n-k)!}$, serve para contar de quantas maneiras podemos obter k sucessos.

   Substituindo os valores temos que a nossa função de probabilidade é dada por:

$$P(X = k) = \frac{10!}{k!(10-k)!} 0.2^k (1-0.2)^{10-k}, k = 0, .., 10$$

   Para encontrar a probabilidade de encontrar cada quantidade de contestações, basta substituir a quantidade por k (refaça os itens 1 e 2 usando a formula para conferir)

4. **Ache o valor esperado e a variância da variável aleatória X?**

   Para encontrar o valor esperado e a variância podemos apenas aplicar as fórmulas, no entanto conhecer a distribuição de probabilidade da variável nos dá informação sobre ela, e isso facilita o cálculo de tais medidas. A **esperança** de uma variável aleatória binomial é dada por **n x p** e a **variância** é dada por **n x p x (1-p)**, portanto aqui conseguimos calcular de forma mais rápida tais valores:

$$E(X) = n \times p = 10 \times 0.2 = 2$$
$$Var(x) = n \times p \times (1-p) = 10 \times 0.2 \times (1-0.2) = 1.6$$

Exemplo 4.0.6. *Numa pesquisa com um conjunto de $n = 10$ estudantes usuários do Whatsapp eles foram classificados em três categorias: Gosta muito, Gosta e Acha ruim; obteve-se a seguinte informação:*

Gosto pelo Whatsapp

Sexo	*Tipo*			
	Gosta muito	*Gosta*	*Acha ruim*	*Total*
Homem	*1*	*1*	*3*	*5*
Mulher	*3*	*1*	*1*	*5*
Total	*4*	*2*	*4*	*10*

1. *Ache as seguintes probabilidades:*
   - *(a)* $P(\ Acha\ ruim \mid Ser\ mulher\ )$
   - *(b)* $P(\ Gosta\ muito\ ou\ Gosta\ e\ ser\ homem\ )$
   - *(c)* $P(\ Ser\ Homem \mid Gosta\ muito)$
2. *Os eventos 'Ser mulher' e 'Gosta muito' são independentes? explique a resposta*

soluções

1. *Antes de começar a fazer as contas vamos relembrar um pouco sobre probabilidade condicional. Quando temos $P(A|B)$ (Prob de A dado B), estamos interessados em calcular a probabilidade do evento A acontecer Sabendo que o evento B já aconteceu. Tal probabilidade pode ser calculada através da redução de espaço amostral, que consiste em calcular a probabilidade de evento A acontecer dentro das observações do evento B.*

   Outra forma de calcular tal probabilidade é através da fórmula de probabilidade condicional, que é dada por:

   $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

   Tal formula indica que a probabilidade do evento A acontecer sabendo que o evento B aconteceu, pode ser calculada através da divisão entre a probabilidade dos dois ocorrerem simultaneamente pela probabilidade do evento B acontecer (A mesma ideia da redução do espaço amostral).

   Sabendo isso, vamos calcular as probabilidades pedidas.

(a) $P($ **Acha ruim | Ser mulher** $)$

Temos que:

$$P(\ Acha\ ruim \mid Ser\ mulher\ ) = \frac{P(\ Acha\ ruim \cap Ser\ mulher\ )}{P(\ Ser\ mulher\ )}$$

A probabilidade será encontrada através da divisão pela quantidade de vezes que o evento acontece pelo tamanho da amostra. A intercessão é onde os eventos ocorrem ao mesmo tempo, olhando na tabela podemos ver que os dois evento de interesse só acontecem juntos uma vez, portanto $P(\ Acha\ ruim\ \cap\ Ser\ mulher\ ) = 1/10$ *. Há 5 mulheres na amostra, então* $P(\ Ser\ mulher\ ) = 5/10$*. Temos então:*

$$P(\ Acha\ ruim \mid Ser\ mulher\ ) = \frac{\frac{1}{10}}{\frac{5}{10}} = \frac{1}{10} \times \frac{10}{5} = \frac{1}{5} = 0.2$$

encontramos então a probabilidade de interesse: $P(\ \boldsymbol{Acha\ ruim} \mid \boldsymbol{Ser\ mulher}\ ) = 0.2$

(b) $P($ **Gosta muito ou Gosta e ser homem** $)$

Estamos interessados em calcular a probabilidade de gostar muito do whatsapp ou ser homem, neste caso não é necessário que isso aconteça simultaneamente. Portanto, devemos considerar todas as pessoas que gostam muito, e todos os homens. Temos então 4 pessoas que gostam muito e 5 homens. A ideia inicial seria somar estes valores, e dividir pelo total, no entanto, quando fazemos isso estamos contando os homens que gostam muito duas vezes. Torna-se necessário subtrair deste total o a quantidade em que os dois acontecem ao mesmo tempo (5 + 4 -1= 8).

Para ficar mais claro, vamos definir exatamente quais valores devem ser considerados: Temos, 1 homem que gosta muito, 1 homem que gosta, 3 homens que acham ruim e 3 mulheres que gostam muito(1+1+3+3 = 8).

Note que nos dois casos é importante tomar cuidado para não contar a intercessão duas vezes. Agora que já temos a quantidade de vezes em que o evento de interesse acontece, basta dividir este valor pelo tamanho total da amostra: $8/10 = 0.8$.

A probabilidade de alguém gostar muito ou ser homem é igual á 0.8

(c) $P($ **Ser Homem |Gosta muito**$)$

Temos que:

$$P(\ Ser\ Homem \mid Gosta\ muito\ ) = \frac{P(\ Ser\ Homem \cap\ Gosta\ muito\ )}{P(\ Gosta\ muito)}$$

A probabilidade será encontrada através da divisão pela quantidade de vezes que o evento acontece pelo tamanho da amostra. A intercessão é onde os eventos ocorrem ao mesmo tempo, olhando na tabela podemos ver que os dois evento de interesse só acontecem juntos uma vez, portanto $P(\ Ser\ Homem \cap\ Gostar\ muito\ ) = 1/10$ *. Há 5 homens na amostra, então* $P(\ Ser\ Homem\ ) = 5/10$*. Temos então:*

$$P(\ Ser\ Homem \mid\ Gosta\ muito) = \frac{\frac{1}{10}}{\frac{5}{10}} = \frac{1}{10} \times \frac{10}{5} = \frac{1}{5} = 0.2$$

encontramos então a probabilidade de interesse: $P(\ \boldsymbol{Ser\ Homem} \mid\ \boldsymbol{Gosta\ muito}\ ) = 0.2$

2. **Os eventos 'Ser mulher' e 'Gosta muito' são independentes? explique a resposta**

   Se os dois eventos são independentes, o ***fato de um deles ocorrer, não deverá interferir na probabilidade do outro ocorrer****. Portanto para que os eventos "Ser mulher"e "Gostar muito"seja independentes, é preciso que:*

   $P(\ Ser\ mulher\ |Gosta\ muito)$ *seja igual á* $P(ser\ mulher)$ *e*
   $P(\ Gostar\ muito\ |ser\ mulher)$ *seja igual á* $P(Gostar\ muito)$

   Vamos então calcular tais probabilidades e conferir se as igualdades são satisfeitas.

   ** Há 5 mulheres na amostra, portanto P(ser mulher) = 5/10 = 0.5*
   *** $P(\ Ser\ mulher\ |Gosta\ muito) = \frac{3/10}{4/10} = 3/4 = 0.75$
   ** Há 4 pessoas que gostam muito, portanto P(gostar muito) = 4/10 = 0.4*
   *** $P(\ Gosta\ muito\ |Ser\ mulher) = \frac{3/10}{5/10} = 3/5 = 0.6$

   Como$P(ser\ mulher)$ *é diferente de* $P(\ Ser\ mulher|Gosta\ muito)$ *e* $P(gostar\ muito)$ *é diferente de* $P(\ Gosta\ muito\ |Ser\ mulher)$ *, podemos concluir que* ***os eventos não são independentes****.*

Exemplo 4.0.7. *Numa pesquisa sobre o primeiro emprego, com um conjunto de* $n = 100$ *jovens (18-22 anos), obteve-se a seguinte informação:*

Setor	Educação			
	Ensino médio completo	Técnico	Graduação incompleta	Total
Comércio	40	10	10	60
Indústria	5	25	10	40
Total	45	35	20	100

1. *Ache as seguintes probabilidades:*
   - *P(Ter Graduação Incompleta ou Técnico|Primeiro emprego no Comércio)*
   - *P(Ter apenas Ensino médio completo e primeiro emprego na Industria)*
   - *P(Primeiro emprego no Comércio|Graduação Incompleta)*
2. *Os eventos 'Primeiro emprego no Comercio' e 'Graduação Incompleta' são independentes?*

Solução

1. *Ache as seguintes probabilidades:*

   - *P(Ter Graduação Incompleta ou Técnico|Primeiro emprego no Comércio)*
     O que estamos calculando nesse exercício é a probabilidade de encontrar alguém que tenha graduação incompleta ou técnico sabendo que o primeiro emprego desse alguém foi no comércio. Sabemos que o total de pessoas que tiveram o primeiro emprego no comércio é igual a 60, como pode ver na tabela. Calculamos, então, a quantidade de pessoas dentro desse universo de 60 pessoas o total dos que possuem graduação incompleta mais os que possuem curso técnico. Somamos esses dois grupos, uma vez que ambos satisfazem a condição (graduação incompleta ou técnico). Assim, como o total de pessoas que possuem graduação incompleta é igual a 10 e o total de pessoas que possuem técnico é igual a 10, existem $10 + 10 = 20$ *pessoas que satisfazem uma ou outra condição dentro dos 60 que tiveram o primeiro emprego no comércio. Então, a probabilidade de encontrar alguém que tenha graduação incompleta ou técnico sabendo que o primeiro emprego desse alguém foi no comércio é igual a* $\frac{20}{60} = 0.333$ *ou 33.3%.*
   - *P(Ter apenas Ensino médio completo e primeiro emprego na Industria)*
     Repare que nesse exercício devemos calcular o total de pessoas que satisfazem ambas as condições (Ter apenas Ensino médio completo e primeiro emprego na Industria) e dividir pelo total de jovens (n = 100). O total de pessoas que satisfazem ambas as condições é igual a 5, como pode ver na tabela. Então, a probabilidade de encontrarmos alguém desse grupo dentro do total de 100 jovens é igual a $\frac{5}{100} = 0.05$ *ou 5%.*
   - *P(Primeiro emprego no Comércio|Graduação Incompleta)*
     O que estamos calculando nesse exercício é a probabilidade de encontrar alguém que teve o primeiro emprego no comércio sabendo que esse alguém possui graduação incompleta. Sabemos que o total de pessoas que tiveram seu primeiro emprego no comércio dentro do grupo dos que possuem graduação incompleta é igual a 10. Sabemos também, olhando na tabela, que o total de pessoas que possuem graduação incompleta é igual a 20. Então, a probabilidade de encontrar alguém que teve o primeiro emprego no comércio sabendo que esse alguém possui graduação incompleta é igual a $\frac{10}{20} = 0.5$ *ou 50%.*

2. *Os eventos 'Primeiro emprego no Comercio' e 'Graduação Incompleta' são independentes?*

   Se são independentes, então a probabilidade de ambos ocorrerem ao mesmo tempo é igual a probabilidade de um vezes a probabilidade do outro. (P(Primeiro emprego no Comercio) x P(Graduação Incompleta) = P(Primeiro emprego no Comercio,Graduação Incompleta)). Basta verificar se isto ocorre. Se ocorre, são independentes, caso contrário, não são.

   - *P(Primeiro emprego no Comercio,Graduação Incompleta) =* $\frac{10}{100} = 0.1$
   - *P(Primeiro emprego no Comercio) =* $\frac{60}{100} = 0.6$
   - *P(Graduação Incompleta) =* $\frac{20}{100} = 0.2$

P(Primeiro emprego no Comercio) x P(Graduação Incompleta) = $0.6x0.2 = 0.12$. *Como P(Primeiro emprego no Comercio,Graduação Incompleta) = 0.1, concluímos que os eventos 'Primeiro emprego no Comercio' e 'Graduação Incompleta' não são independentes, pois* $0.1 \neq 0.12$.

Exemplo 4.0.8. *O tempo, em meses, no primeiro emprego nesta população de jovens tem distribuição Binomial com parâmetros* $n = 6$ *e* $p = 0.8$. *Calcule a probabilidade de que um jovem fique no primeiro emprego 1 ou 2 meses*

Solução

A probabilidade de que um jovem fique no primeiro emprego 1 ou 2 meses é igual a probabilidade de que ele fique 1 mês **somado** *à probabilidade de que ele fique 2 meses.*

Para calcular ambas as probabilidades para depois somá-las vamos aplicar estes valores (1 e 2) na fórmula da binomial. Sua fórmula é:

$$P(X = k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}, k = 0, \ldots, n$$

Como o exercício nos fornece o valor de n e p ($n = 6$ *e* $p = 0.8$*), basta substituir esses valores na fórmula e k será igual a 1, para calcular a probabilidade de que um jovem fique 1 mês, e depois igual a 2 para calcular a probabilidade de que um jovem fique 2 meses.*

- $P(X = 1) = \frac{6!}{1!(6-1)!}0.8^1(1-0.8)^{6-1} = \frac{6!}{1!5!}0.8^1(0.2)^5 = 0.001$
- $P(X = 2) = \frac{6!}{2!(6-2)!}0.8^2(1-0.8)^{6-2} = \frac{6!}{2!4!}0.8^2(0.2)^4 = 0.01$

Observação: Lembre-se que $6! = 6x5x4x3x2x1$, $1! = 1$ *e todo número elevado a 1 é igual a ele mesmo.*

$P(X = 1) + P(X = 2) = 0.011$. *Portanto, a probabilidade de que um jovem fique no primeiro emprego 1 ou 2 meses é igual a 0.011 ou 1.1%.*

Exemplo 4.0.9. *O salario pago na Indústria no primeiro emprego para jovens tem distribuição Normal com media 1.5 salários mínimos (SM) e variância 0.5 (SM). Salario* $\sim$ *N(1.5,0.5).*

1. *Ache a probabilidade de um jovem ter um Salario maior do que 2 SM no primeiro emprego.*
2. *Ache a probabilidade de um jovem ter um Salario menor do que 1 SM no primeiro emprego.*

Solução

1. *Ache a probabilidade de um jovem ter um salário maior do que 2 salários mínimos (SM) no primeiro emprego.*

   Como na tabela da distribuição normal encontramos apenas as probabilidades de uma distribuição normal com média igual a 0 e desvio padrão igual a 1, vamos achar o valor padronizado de 2 SM. Lembre-se que o desvio padrão é a raiz quadrada da variância

   $$Z = \frac{2-1.5}{\sqrt{0.5}} = 0.7071$$

   Nesse sentido, descobrimos que 2 SM se distancia em 0.7071 desvios padrão da média. Isso significa, além disso, que a observação de número 2 em uma distribuição normal com média 1.5 e variância 0.5 equivale a uma observação de número 0.7071 em uma distribuição normal com média 0 e variância 1. Pronto, agora é só olhar na tabela Z qual é a probabilidade de encontrarmos algo maior que 0.7071 e, assim, descobrimos a probabilidade de um jovem ter um salário maior do que 2 salários mínimos (SM) no primeiro emprego. Essa probabilidade é igual a $P(Z > 0.7071) = 0.2397$ *ou 23.97%.*

2. *Ache a probabilidade de um jovem ter um salário menor do que 1 salário mínimo (SM) no primeiro emprego.*

   Como a tabela Z nos fornece apenas as probabilidades acima de 0 (acima da média) vamos ter que lembrar uma propriedade da distribuição normal: a propriedade da simetria. Essa propriedade é muito útil, pois como sabemos que a probabilidade de encontrarmos uma observação abaixo da média é a mesma de encontrar uma observação acima da média, também sabemos que a probabilidade de encontrar uma algo menor que uma observação que se distancia x desvios padrão da média é a mesma de encontrar uma algo maior que uma observação que se distancia x desvios padrão da média. Sabendo disso, vamos padronizar a nossa observação (1 SM) para saber quantos desvios padrão essa observação (1 SM) se distancia da média e, assim, encontrar seu valor equivalente em uma distribuição normal com média 0 e desvio padrão igual a 1:

   $$Z = \frac{1-1.5}{\sqrt{0.5}} = -0.7071$$

   Ora, como explicado anteriormente, $P(Z < -0.7071) = P(Z > 0.7071) = 0.2397$ *ou 23.97%. Assim, a probabilidade de um jovem ter um salário menor do que 1 salário mínimo (SM) no primeiro emprego é igual a 0.2397 (23.97%) e inclusive é igual a probabilidade de um jovem ter um salário maior do que 2 salários mínimos (SM) no primeiro emprego (questão anterior).*

Capítulo 5

Inferência

Exemplo 5.0.1. *Duas questões sobre o aplicativo que mais gostam (Whatsapp ou Facebook) foram aplicadas a dez alunos de uma turma de adolescentes. Os resultados dos questionários (***0** *não gosta,* **1** *gosta) do Whatsapp foram: 1, 1, 1, 0, 1, 1, 0, 0, 1, 1 e os resultados do Facebook: 1, 0, 0, 1, 1, 0, 0, 0, 0, 1. A proporção dos que gostam dos aplicativos das redes sociais na população jovem é de* $p = 0.7(70\%)$.

1. *Acredita-se que a proporção dos que gostam do Whatsapp é* **maior** *que a da população. Faça um teste de hipóteses para verificar esta suposição (formule as hipótese correspondentes, ache a estatística adequada, ache o p-valor e obtenha o resultado do teste)*

2. *No caso do Facebook, acredita-se que a proporção é* **diferente** *da população. Faça um teste de hipóteses para verificar isto.*

Solução

Desejamos fazer alguma conclusão sobre a população com base na amostra, e por termos uma amostra pequena, um teste exato será mais adequado, a estrutura dos dados, nos faz pensar na distribuição binomial, pois trata-se de uma sequência independente de observações dicotômicas. Consideramos ***X = "número de pessoas que gostam da rede social mencionada"****, obtemos observações provenientes de distribuições binomiais.*

1. **Acredita-se que a proporção dos que gostam do Whatsapp é maior que a da população. Faça um teste de hipóteses para verificar esta suposição (formule as hipótese correspondentes, ache a estatística adequada, ache o p-valor e obtenha o resultado do teste)**

   Desejamos testar se a proporção de pessoas que gostam do Whatsapp é ***maior*** *que a da população que é igual á 0.7. Neste caso, estamos testando as seguintes hipóteses:*

$$H_0 : p = 0.7$$
$$H_1 : p > 0.7$$

Para prosseguir com o teste, devemos ***supor que a hipótese nula seja verdadeira****. Assumindo a suposição de que a proporção é igual a 0.7, os dados seriam provenientes de uma distribuição binomial com parâmetros n= 10 e p = 0.7.*

A decisão do teste é dada com base na probabilidade da ocorrência do valor observado ou um valor mais extremo(esta probabilidade é o nosso p-valor). Na amostra referente ao Whatsapp, foram observados, 7 pessoas que marcaram que gostam , portanto, nossa estatística de teste é igual á 7.

O p-valor será dado então por: $P(X \geq 7|p = 0.7)$*. A especificação* $|p = 0.7$*, serve para lembrar que estamos supondo a hipótese nula como verdadeira (é comum aparecer* $|H_0)$*.*

$$P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$$

Vamos lembrar aqui a formula da distribuição binomial já com as especificações da nossa hipótese. Neste ponto supomos que você já tenha familiaridade com a distribuição binomial e o uso de sua fórmula.

$$P(X = k) = \frac{10!}{k!(10-k)!}0.7^k(1-0.7)^{10-k}, k = 0, \ldots, 10$$

Para cada probabilidade desejada, basta substituir o valor k na fórmula.
$P(X = 7) = \frac{10!}{7!(10-7)!}0.7^7(1-0.7)^{10-7} = 0.267$
$P(X = 8) = \frac{10!}{8!(10-8)!}0.7^8(1-0.7)^{10-8} = 0.234$
$P(X = 9) = \frac{10!}{9!(10-9)!}0.7^9(1-0.7)^{10-9} = 0.121$
$P(X = 10) = \frac{10!}{10!(10-10)!}0.7^10(1-0.7)^{10-10} = 0.028$
Teremos então:

$$P(X \geq 7) = 0.267 + 0.234 + 0.121 + 0.028 = 0.65$$

Nosso p-valor é igual á 0.65, ou seja, A probabilidade de observamos 7 sucessos ou mais, se a hipótese nula for verdadeira, é 0.65(Uma probabilidade bem alta), ao compararmos com um nível de significância $\alpha = 0.05$*, temos que o p-valor é maior(Não rejeitamos* H_0*), isso indica que não temos evidências para decidir a favor da hipótese alternativa. Portanto, a proporção* ***não é maior que 0.7****.*

2. **No caso do Facebook, acredita-se que a proporção é diferente da população. Faça um teste de hipóteses para verificar isto.**

   Agora estamos interessados em testar se a proporção de pessoas que gostam do Facebook é ***Diferente*** *da proporção da população total, que é igual á 0.7. Neste caso, estamos testando as seguintes hipóteses:*

$$H_0 : p = 0.7$$
$$H_1 : p \neq 0.7$$

*Novamente precisamos **supor que a hipótese nula seja verdadeira**. A partir de tal suposição, diremos que a amostra referente ao facebook é proveniente de uma distribuição binomial com parâmetros p= 0.7 e n=10.*

A decisão do teste será dada com base na probabilidade de ocorrência do valor observado ou um valor mais extremo(p-valor) Nessa amostra foram observadas 4 pessoas que gostam do Facebook, portanto, nossa estatística de teste é igual a 4.

*O nosso p-valor é dado pela probabilidade de encontrar um valor igual ou mais extremo. Como o nosso teste é **bilateral**, devemos calcular a probabilidade observar o valor 4, e **considerar todas as probabilidades menores ou iguais essa**, pois são valore mais extremos ou seja menos ou igualmente prováveis que 4. A seguir é exibida uma tabela com as probabilidades para cada valor, e a partir dela analisaremos e encontraremos o p-valor:*

P(X= 0) = 0.0000	P(X=6)=0.200
P(X= 1)= 0.0001	P(X=7)=0.266
P(X= 2)=0.0014	P(X=8)=0.233
P(X= 3)=0.009	P(X=9)=0.121
P(X=4)=0.036	**P(X=10)=0.028**
P(X=5)=0.102	

Os cálculos foram feitos com base na formula da distribuição binomial especificado no primeiro item desta questão.

*A probabilidade de observamos 4 valores é 0.036. Portanto todos os valores com probabilidade menor que esta serão considerados extremos, e deverão ser incorporados no nosso p-valor. Na tabela estão destacados em vermelho, os valores de menor probabilidade. Portanto, **o p-valor será a soma destas probabilidades**.*

$$pvalor = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 10)$$
$$pvalor = 0.00000 + 0.0001 + 0.0014 + 0.009 + 0.036 + 0.028 = 0.075$$

O p-valor encontrado foi 0.075. Ao compararmos com um nível de significância de 5% (0.05) nós não temos evidências para rejeitar a hipótese nula. Ou seja, podemos dizer que a proporção de pessoas que gostam do facebook pode ser considerada igual a proporção de pessoas que gostam de redes sociais na população.

Um fato importante de se lembrar é que, caso o valor a ser suposto na hipótese nula fosse 0.5, estaríamos diante de uma distribuição simétrica, portanto bastaria calcular a probabilidade de ser menor que 4, e multiplicar este valor por 2.

Exemplo 5.0.2. *Um sociólogo acredita que o número de redes sociais utilizadas pode interferir na chance de encontrar velhas amizades. Através de uma amostra ele obteve a seguinte tabela para as variáveis $Y \equiv$ número de redes sociais utilizadas e $X \equiv$ número de velhas amizades.*

$X \backslash Y$	1	2	3
0	3	2	4
1	1	1	2
2	0	1	3

O número de redes sociais utilizadas(Y) e o número de velhas amizades(X) são independentes?? (formule as hipótese correspondentes, ache a estatística adequada, ache o p-valor e conclua o teste).

Tabela da Distribuição Qui-quadrado:

	$p-$valor								
$g.l.$	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
4	5,39	7,78	9,49	11.14	13,28	14,86	16,42	18,47	20,00

Solução

Quando desejamos identificar se existe independência entre duas variáveis que foram medidas na **mesma unidade experimental**, o teste sugerido é o teste Qui-Quadrado. No nosso caso, desejamos identificar se há independência entre o número de redes sociais e o número de velhas amizades, sendo que as duas variáveis são observadas em cada pessoa, portanto, o teste Qui-quadrado é adequado. Portanto estamos testando as seguintes hipóteses:

$$H_0: \text{ X e Y são independentes}$$
$$H_1: \text{ X e Y não são independentes}$$

Antes de prosseguir, vamos definir algumas notações para facilitar o entendimento:

- **n** é o tamanho total da amostra = 17
- O_{ij} é o valor observado na casela que corresponde a linha i e coluna j, por exemplo: $O_{11} = 3$ e $n_{23=2} = 3$
- $n_{i.}$ corresponde a soma da linha i, portanto:
  $n_{1.} = 3 + 2 + 4 = 9$; $n_{2.} = 1 + 1 + 2 = 4$ e $n_{3.} = 0 + 1 + 3 = 4$
- $n_{.j}$ corresponde a soma da coluna j, portanto:
  $n_{.1} = 3 + 1 + 0 = 4$; $n_{.2} = 2 + 1 + 1 = 4$ e $n_{.3} = 4 + 2 + 3 = 9$

Como em todos os testes de hipótese, o primeiro passo é supor a hipótese nula como verdadeira, vamos pensar então em, quantas observações teríamos em cada casela se X e Y fossem independentes. para isso vamos lembra que , se dois eventos A e B são independentes, então

$P(a \cap b) = P(a) \times P(b)$. Portanto, teremos $P(X_i \cap Y_i) = P(X_i) \times P(X_j) = p_{ij}$, para i e j variando de 1 até 3(3 linhas e 3 colunas).

Tendo definido isso, as probabilidades são encontradas por:

$$p_{ij} = \frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$$

Para encontrar o **número esperado de observações em cada casela(i,j)**, entre o total de observações e com a hipótese de independência devemos multiplicar a probabilidade pelo tamanho da amostra. Teremos nossos valores esperados definidos por:

$$E_{ij} = n \times p_{ij} = n \times \frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$$

Simplificando ficamos com:

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

Com a fórmula acima, conseguimos calcular o valor esperado para cada casela(combinação ij) da nossa tabela. por exemplo:

$E_{12} = \frac{9 \times 4}{17} = 2.11$
$E_{21} = \frac{4 \times 4}{17} = 0.94$

Se seguimos supondo independência, a distância entre os valores observados e esperados, é dada por:

$$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{i=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Sendo que χ^2_{obs} tem distribuição qui-quadrado com (r-1)(s-1) graus de liberdade. Em nosso exercício, temos 3 categoria em X e 3 na variável Y, portanto r = 3, e s = 3, e χ^2_{obs} terá distribuição qui-quadrado com 4 graus de liberdade$((3-1) \times (3-1))$.

Entendo que a formula acima, pode ser confusa para muitas pessoas, no entanto ela nos indica que devemos pegar cada casela, subtrair dela seu valor esperado e elevar este valor ao quadrado, e em seguida, dividir pelo valor esperado. Vamos calcular tais valores:

Para facilitar a visualização, utilizaremos tabelas para isso:

Observado				***Esperado***			
X\Y	**1**	**2**	**3**	**X\Y**	**1**	**2**	**3**
0	3	2	4	**0**	2.11	2.11	4.76
1	1	1	2	**1**	0.94	0.94	2.11
2	0	1	3	**2**	0.94	2.11	2.11

Os valores observados são os valores originais, e os valores observados, foram calculados conforme foi explicado anteriormente.

Agora que temos os valores observados e esperados, devemos subtrair o valor esperado do observado e em seguida elevar o resultado ao quadrado e em seguida dividir pelo valor esperado nas caselas equivalente, por exemplo: Nas linha 1 e coluna 1, teremos: $\frac{(3-2.11)^2}{2.11}$, e procederemos somando os resultados para todas as 9 caselas:
$\frac{(3-2.11)^2}{2.11}+\frac{(2-2.11)^2}{2.11}+\frac{(4-4.76)^2}{4.76}+\frac{(1-0.94)^2}{0.94}+\frac{(1-0.94)^2}{0.94}+\frac{(2-2.11)^2}{2.11}+\frac{(0-0.94)^2}{0.94}+\frac{(1-2.11)^2}{2.11}+\frac{(3-2.11)^2}{2.11}$
Prosseguindo com a conta:

$\chi^2_{obs} = 0.36 + 0.006 + 1.12 + 0.003.0.003 + 0.006 + 0.941 + 0.003 + 0.036 = 1.82$

Obtemos então o nosso valor $\chi^2_{obs} = 1.82$, Agora, precisamos então concluir nosso teste.

O teste é baseado na distância entre os valores esperados e os observados, portanto, se as distâncias forem grandes, as variáveis não são independentes. Quando comparamos com a distribuição qui-quadrado, estamos comparando valores positivos, portanto, um χ^2_{obs} pequeno indica independência. O cálculo do p-valor nos dá a probabilidade de uma valor χ^2_q ser maior que o nosso valor χ^2_{obs}, se essa probabilidade for menor que o nosso nível de significância, nós rejeitamos a hipótese nula, pois estamos com uma observação muito diferente do esperado em caso de independência. Nossa regra de decisão fica da seguinte forma:

Se $P(\chi^2_q \geq \chi^2_{obs}) < \alpha$**, então rejeitamos** H_0

A tabela da distribuição χ^2_4 foi mostrada anteriormente, nela devemos procurar na linha do grau de liberdade o valor mais próximo do nosso valor observado, e através dele, obter o p-valor. Por exemplo, se tivéssemos observado o valor 13,28, nosso p-valor seria 0,01. No nosso caso, nosso valor observado foi 1.82, Na tabela, o menor valor, é 5.39, com p-valor equivalente a 0,25. Neste caso, podemos concluir que nosso p-valor é maior que 0,25. Sendo 0,25 maior que um nível de significância de 0,05. Temos o suficiente para não rejeitar a hipótese nula(p-valor maior que o nível de significância). Portanto, não há evidências para dizer que o número de redes sociais utilizadas e o número de velhas amizades são relacionados. Ou seja, as duas variáveis são independentes.

O p-valor exato pode ser obtido através do uso de um computador. Através do software R, obtivemos um p-valor exato igual á : 0.7682

Exemplo 5.0.3. *Em um estudo sobre um curso ofertado nas agências de telemarketing, para duas amostras de* $n = 10$ *trabalhadores, obtiveram-se os seguintes resultados:*

Sujeito	1	2	3	4	5	6	7	8	9	10
Salário antes do curso	7.00	8.40	8.30	8.60	8.40	6.90	8.30	11.80	9.30	10.70
Salário depois do curso	7.20	84.0	8.20	9.00	8.70	7.00	8.00	12.00	9.50	10.80

- *Ache os intervalos de 95% de confiança para a média dos salários nas duas amostras.*
- *A partir da análise dos valores obtidos no item anterior qual seria a conclusão que teríamos? (explique o porquê da conclusão)*

Solução

1. **Ache os intervalos de 95% de confiança para a média da duração da jornada de trabalho das duas amostras.**

   Para construir um intervalo de confiança devemos, primeiramente, encontrar a estimativa pontual. Esta é o centro do intervalo de confiança. Em nosso caso, a estimativa pontual é a média. A margem de erro é o valor o qual subtraímos da média para encontrar o limite inferior do intervalo e somamos a média para encontrar o limite superior. A margem de erro é dada por: erro padrão x $t_{\frac{\alpha}{2}}$. Sendo que o erro padrão é igual ao desvio padrão da nossa amostra dividido pela raiz do tamanho da mesma.

   Sendo assim, a estimativa pontual do "Salário Antes do Curso"é igual a 8.77. Sua margem de erro é igual a $\frac{1.51}{\sqrt{10}}2.262 = 1.08$; já que o desvio padrão desse grupo é 1.51, o tamanho dessa amostra é igual a 10 e o valor $t_{\frac{\alpha}{2}}$ (o que delimita, na distribuição t de Student, uma área de 95% entre $-t_{\frac{\alpha}{2}}$ e $t_{\frac{\alpha}{2}}$) é igual a 2.262. Por isso, o intervalo de confiança para a média desse grupo é: $8.77 + -1.08$ ou [7.69 ; 9.85]

   Fazemos o mesmo procedimento para o outro grupo, "Salário Depois do Curso". Estimativa pontual: 8.88. Margem de erro: $\frac{1.55}{\sqrt{10}}2.262 = 1.10$. Por isso, o intervalo de confiança para a média desse grupo é: $8.88 + -1.10$ ou [7.78;9.98].

2. **A partir da análise dos valores obtidos no item anterior qual seria a conclusão que teríamos a respeito da eficácia do curso? Ele influencia no salário? (explique o porquê da conclusão)**

   Para concluir se o curso faz alguma diferença ou não no salário dos funcionários não podemos simplesmente verificar a diferença média entre os salários, pois não testamos o curso em todos os trabalhadores do mundo nesse ramo laboral. Afinal, queremos concluir sobre a eficácia do curso, e não seu resultado em um grupo específico de pessoas. Mas como temos disponível apenas esse grupo específico, devemos tirar a conclusão que queremos a partir dele por meio da inferência estatística. Nesse sentido, vamos construir um intervalo de confiança para a média das diferenças de salário. Vamos adotar um nível de significância igual a 5%, ou seja, nosso intervalo de confiança terá 95% de confiança. Se o valor zero (0) estiver contido dentro do nosso intervalo não rejeitamos a hipótese de que a média das diferenças salariais antes e depois do curso é igual a zero, ou seja, neste caso concluímos que o curso não influencia em um aumento ou em uma perda salarial. Por outro lado, se o valor zero (0) não estiver contido dentro do nosso intervalo rejeitamos a hipótese de que a média das diferenças salariais antes e depois do curso é igual a zero, ou seja, neste caso concluímos que o curso influencia de alguma forma no salário dos trabalhadores.

$$H_0 : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

Sendo que μ_d é a média populacional das diferenças, ou seja, a verdadeira média a qual queremos estimar por meio da nossa amostra.

Como nossas amostras são pareadas (dependentes uma da outra), ou seja, para cada membro da primeira amostra existe o seu par na segunda amostra, uma vez que ambas possuem os mesmos indivíduos, porém a primeira evidencia o salário antes e a segunda evidencia o salário depois, iremos construir o intervalo de confiança baseado em dez (10) valores amostrais, que são as 10 diferenças salariais. Se as amostras fossem independentes não faríamos dessa maneira.

O vetor das diferenças é: [-0.2,0,0.1,-0.4,-0.3,-0.1,0.3,-0.2,-0.2,-0.1]

Seguindo o mesmo procedimento do item anterior, temos que a estimativa pontual é igual a -0.11 (média das diferenças apresentadas acima) e a margem de erro é: $\frac{0.2}{\sqrt{10}}2.262 = 0.14$. Assim, o intervalo de 95% de confiança para a diferença média salarial (antes - depois) é: $-0.11 + -0.14$ ou [-0.25;0.03]. Como o número zero (0) está contido no intervalo não rejeitamos H0. Ou seja, concluímos que o curso não influencia na média salarial dos trabalhadores desse ramo.

Exemplo 5.0.4. *Em um estudo sobre a eficácia dos cursos de pré-vestibulares numa turma de $n = 10$ alunos obteve-se os seguintes resultados:*

Sujeito	1	2	3	4	5	6	7	8	9	10
Pontos antes do curso	700	840	830	860	840	690	830	1180	930	1070
Pontos depois do curso	720	840	820	900	870	700	800	1200	950	1080

Teste a afirmativa de que o curso não teve nenhum efeito sobre os conceitos obtidos:

1. *Estabeleça as hipóteses nula e alternativa*
2. *Qual o valor da estatística apropriada para este teste (suponha que o desvio padrão das diferenças é $\sigma = 15$)*
3. *Ache o p-valor*
4. *Qual é o resultado do teste ?*

Solução

1. *Estabeleça as hipóteses nula e alternativa*

H0: $\mu_a - \mu_d = 0$
H1: $\mu_a - \mu_d \neq 0$

No caso, H0 supõe que a média populacional dos pontos antes do curso é igual à média populacional dos pontos depois do curso, ou seja, que a diferença entre eles é igual a zero. Esta hipótese afirma, portanto, que o curso não possui eficácia. H1, a hipótese alternativa, afirma o contrário.

2. *Qual o valor da estatística apropriada para este teste (suponha que o desvio padrão das diferenças é $\sigma = 15$)*

   A estatística de teste é uma medida que nos informa o quando H0 discorda do que observamos em nossa amostra. Podemos perceber que a média das diferenças amostrais é igual a

   $$\frac{-20+0+10+(-40)+(-10)+30+(-20)+(-20)+(-10)}{10} = -8$$

   O tanto que este valor se distancia do que H0 supõe em número de erros padrão (desvio padrão da distribuição das diferenças amostrais) é igual a $\frac{-8-0}{\frac{15}{\sqrt{10}}} = -1.68$

3. *Ache o p-valor*

   O p-valor é a probabilidade de encontrarmos um valor mais atípico do que encontramos dado que H0 é verdadeira. Então basta encontrar na tabela t de Student a probabilidade de encontrarmos uma observação menor que -1.68 e multiplicar por dois (pois o teste é bilateral). Portanto, o p-valor é $2.P(X < -1.68) = 2x0.063 = 0.126$.

4. *Qual é o resultado do teste a um nível de 5% de significância ?*

   Não rejeitamos H0, pois o p-valor é maior que a significância (0.126 > 0.05). Ou seja, concluímos estatisticamente que o curso não exerce influência nos pontos obtidos pelos alunos de pré-vestibular.

Exemplo 5.0.5. *Numa amostra de 200 jovens, 25 deles fazem estágio.*

1. *Ache um intervalo de 95% de confiança para a proporção $\hat{p}$ dos que fazem estagio.*
2. *Ache um intervalo de 90% de confiança para a proporção $\hat{q}$ dos que não fazem estagio.*
3. *Teste a hipótese de que a probabilidade de fazer estagio é 0.10*

Solução

1. *Ache um intervalo de 95% de confiança para a proporção $\hat{p}$ dos que fazem estagio.*

   Sabemos que a proporção de pessoas que fazem estágio $\hat{p}$ é igual a $\frac{25}{200} = 0.125$ ou 12.5%. No entanto, como estamos estudando uma amostra, e não a população inteira de jovens, não significa que 0.125 é a real proporção de jovens que fazem estágio. Sabemos, entretanto, que a

distribuição de probabilidade que gera estas proporções amostrais é uma distribuição normal com média igual a p (sendo p a proporção populacional, a qual estamos interessados em descobrir) e desvio padrão igual a $\sqrt{\frac{p(1-p)}{n}}$. *Neste caso, n é igual a 200 e p, por não conhecermos, o substituímos por* $\hat{p}$, *que é igual a 0.125. Dessa forma, como conhecemos o desvio padrão da distribuição normal que gera as proporções amostrais de amostras de tamanho igual a 200 (***erro padrão***) e também sabemos que 1.96 é a quantidade de desvios padrão distante da média que delimita uma área de 95% em tal distribuição (basta olhar na tabela da distribuição normal), a margem de erro é igual a 1.96 vezes o erro padrão:* $\sqrt{\frac{p(1-p)}{n}}1.96 = \sqrt{\frac{0.125(0.875)}{200}}1.96 = 0.001$. *Assim, o limite inferior do intervalo é* $0.125 - 0.001 = 0.124$ *e o limite superior do intervalo é* $0.125 + 0.001 = 0.126$.

2. *Ache um intervalo de 90% de confiança para a proporção* $\hat{q}$ *dos que não fazem estagio.*

   Faça exatamente o mesmo procedimento do item anterior, porém neste caso a proporção $\hat{q}$ *é igual a* $\frac{175}{200} = 0.875$. *Assim, a margem de erro é igual a* $\sqrt{\frac{p(1-p)}{n}}1.96 = \sqrt{\frac{0.875(0.125)}{200}}1.96 = 0.001$. *Portanto, o limite inferior do intervalo é* $0.875 - 0.001 = 0.874$ *e o limite superior do intervalo é* $0.875 + 0.001 = 0.876$.

   Repare que a margem de erro foi igual a do exercício anterior, pois ter e não ter estágio são eventos complementares em uma amostra finita. Note que $0.125 + 0.875 = 1 = 100\%$

3. *Teste a hipótese de que a probabilidade de fazer estágio é igual a 0.10. Use uma significância igual a 0.05.*

$$H_0 : p = 0.10$$
$$H_1 : p \neq 0.10$$

   Vamos encontrar a **estatística de teste**. *Para tal, basta calcular a quantidade de erros padrão que a nossa observação se distancia de H0. Assim, a estatística de teste é igual a* $\frac{0.125-0.10}{\sqrt{\frac{0.1(0.9)}{200}}} = 1.17$. *Repare que o erro padrão é calculado supondo que H0 seja verdade, pois estamos verificando o quão atípico é observarmos uma proporção amostral* $\hat{p} = 0.125$ *em uma amostra de 200 jovens em um contexto onde a proporção real (populacional) é igual a 0.10 e, dessa forma, decidimos se rejeitamos ou não H0. Como encontramos 1.17 para a nossa estatística de teste, basta verificarmos qual é a probabilidade de encontrarmos algo mais atípico que isto e, assim, decidir se nossa observação (*$\hat{p} = 0.125$*) é ou não é improvável em um contexto onde H0 é verdade (*$p = 0.10$*). Se olharmos na tabela Z podemos notar que a probabilidade de encontrarmos algo maior que 1.17 é 0.12. Como nosso teste é bilateral (ser diferente pode ser tanto maior quanto menor) o valor-p é igual a* $2P(Z > 1.17) = 2x0.12 = 0.24$. *Por fim, como a probabilidade de encontrarmos algo mais atípico que nossa observação supondo que H0 é verdade (p-valor) é igual a 0.24 e este valor respeita a nossa tolerância estabelecida para ele (significância, que é igual a 0.05), pois* $0.24 > 0.05$, *não rejeitamos H0. Ou seja, admitimos estatisticamente que é plausível que a média populacional seja igual a 0.10.*

Exemplo 5.0.6. *Numa empresa encontramos que as posições ocupadas pelos trabalhadores se distribuem como na tabela abaixo:*

Gênero	Posições Ocupadas			Total
	Altas	Médias	Baixas	
Mulheres	4	21	69	94
Homens	15	30	100	145
Total	19	51	169	239

Acredita-se que as posições ocupadas não dependem do gênero. Para testar esta afirmação faça o que se pede:

1. *Construa as hipóteses para este teste.*
2. *Ache a estatística Q^2 para este teste.*
3. *Encontre o p-valor e, através dele, responda qual é a conclusão do teste de hipóteses. Use uma significância igual a 0.05 (5%)*

Solução

1. *Construa as hipóteses para este teste.*

   H_0: *'Gênero' e 'Posição ocupada' são independentes*
   H_1: *'Gênero' e 'Posição ocupada' não são independentes*

   Como estamos querendo verificar se as posições não dependem do gênero, H0 está supondo, neste caso, que a probabilidade de **ocupar uma determinada posição** *e* **ser de um determinado gênero** *(mulher, por exemplo) são eventos* **independentes**. *É importante saber que existe um teorema em probabilidade que nos informa que a probabilidade de acontecer simultaneamente dois eventos independentes é igual a multiplicação de suas probabilidades individuais. Por exemplo, se A e B são eventos independentes, então a probabilidade de que eles ocorram ao mesmo tempo é:* $P(A, B) = P(A).P(B)$.

2. *Ache a estatística Q^2 para este teste.*

   Para calcular a estatística Q^2 (Qui-Quadrado) basta encontrar os valores esperados considerando que H0 seja verdade (que os eventos "posição"e "gênero"são independentes) e aplicá-los na fórmula: $\sum \frac{(e_i - o_i)^2}{e_i}$. *Sendo que o_i é cada valor observado da tabela que não faça parte de algum total (total da linha ou total da coluna). Para encontrar a quantidade esperada de mulheres em altas posições, por exemplo, considerando que os eventos "posição"e "gênero"são independentes, primeiro multiplicamos a probabilidade de ser mulher pela probabilidade de alguém ocupar altos cargos e encontrar . Essa probabilidade é igual a* $\frac{94}{239}\frac{19}{239} = 0.031$.

Depois, encontramos a quantidade esperada que desejamos multiplicando esta probabilidade pelo total de pessoas: $0.031x239 = 7.4$, *daí arredondamos para 8. Ou seja,* **se** *os eventos "Posição"e "Gênero"são independentes,* **esperamos observar 8 pessoas** *que são mulheres e ocupam altas posições.*

Basta fazer o mesmo procedimento para o restante dos cargos para as mulheres e o mesmo para os homens.

Valor esperado para*:*

- *Ser mulher e ocupar altas posições: 8 pessoas (já calculado acima)*
- *Ser mulher e ocupar médias posições:* $\frac{94}{239}\frac{51}{239} = 0.083$*; daí* $0.083x239 = 19.83$ *e arredondamos para 20.*
- *Ser mulher e ocupar baixas posições:* $\frac{94}{239}\frac{169}{239} = 0.278$*; daí* $0.278x239 = 66.44$ *e arredondamos para 67.*
- *Ser homem e ocupar altas posições:* $\frac{145}{239}\frac{19}{239} = 0.048$*; daí* $0.048x239 = 11.47$ *e arredondamos para 12.*
- *Ser homem e ocupar médias posições:* $\frac{145}{239}\frac{51}{239} = 0.129$*; daí* $0.129x239 = 30.83$ *e arredondamos para 31.*
- *Ser homem e ocupar baixas posições:* $\frac{145}{239}\frac{169}{239} = 0.429$*; daí* $0.429x239 = 102.53$ *e arredondamos para 103.*

Por fim, aplicamos estes valores na fórmula apresentada no início da resolução desse problema. $Q^2 = \sum \frac{(8-4)^2}{8} + \frac{(20-21)^2}{20} + \frac{(67-69)^2}{67} + \frac{(12-15)^2}{12} + \frac{(31-30)^2}{31} + \frac{(103-100)^2}{103} = 2.97$

3. *Encontre o p-valor e, através dele, responda qual é a conclusão do teste de hipóteses. Use uma significância igual a 0.05 (5%)*

   Para encontrarmos o valor-p basta olhar na tabela Qui-Quadrado com 6 graus de liberdade (pois temos 2 linhas e 3 colunas, sem contar com as linhas e colunas do total. Assim $2x3 = 6$*) qual é a probabilidade de encontrar algo maior que 2.97. Esta probabilidade (p-valor) é igual a 0.8126. Como o p-valor é maior que a significância (*$0.8126 > 0.05$*) não rejeitamos H0. Ou seja, não rejeitamos a hipótese de que as posições ocupadas não dependem do gênero.*

Exemplo 5.0.7. *Duas empresas agrícolas estão interessadas em descobrir se a produtividade media dos funcionários é a mesma.*
Para isso, foi feira uma medição de produtividade em cada trabalhador, numa escala de 20 a 40 pontos. Considere as variáveis, $X \equiv$ *"Produtividade na primeira empresa"e* $Y \equiv$ *"Produtividade na segunda empresa". Os resultados são exibidos a seguir:*

Funcionários Empresas	1	2	3	4	5	6	7	8	9	10
X	22	21	28	30	33	33	26	24	31	22
Y	25	28	25	36	32	39	28	33	37	27

- *Supondo que as produtividade nas duas empresas siga uma distribuição normal, teste a hipótese de que a produtividade media é maior na segunda empresa: estabeleça as hipóteses adequadas, ache o valor da estatística do teste e ache o resultado do teste usando o p-valor*

Solução

O primeiro passo é pensar na distribuição dos dados e em seus parâmetros. O enunciado nos diz que a produtividade nas duas empresas segue uma distribuição normal, e estamos interessados em fazer testes sobre a média. A distribuição normal, no entanto, é caracterizada por dois parâmetros, a média e a variância. Como não sabemos a variância de nenhuma das duas populações, precisaremos estimá-las através da amostra. Neste ponto devemos pensar se a variância das duas populações é igual ou diferente (Há testes adequados para testar isso).

Fazendo os teste adequados, é razoável dizer que a **variância das duas populações é igual**, portanto realizaremos o teste para comparação de médias de duas populações que tem a mesma variância.

Antes de iniciarmos o teste de hipótese, vamos calcular as nossas estimativas de média e variância para x e y. Denotaremos por: $\bar{X}$ = média da variável X e $\bar{Y}$ a média da variável Y.

A média é calculada pela seguinte formula:$\bar{X} = \frac{\sum_{i=1}^{10}(x_i)}{n}$, que indica que devemos somar todos os valor observados e dividir a soma pelo tamanho amostra(no nosso caso, n=10). Portanto teremos:

$$\bar{X} = \frac{\sum_{i=1}^{10}(x_i)}{n} = 22 + 21 + 28 + 30 + 33 + 33 + 26 + 24 + 31 + 22 = \frac{270}{10} = 27$$

$$\bar{Y} = \frac{\sum_{i=1}^{10}(y_i)}{n} = 25 + 28 + 25 + 26 + 32 + 39 + 28 + 33 + 37 + 27 = \frac{310}{10} = 31$$

Denotaremos a variância da amostra da variável X por s_x^2 , e a variância da amostra da variável Y com s_y^2. A fórmula para cálculo da variância amostra é dada por: $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$, portanto deveremos fazer cada valor menos a média amostral e elevar cada resultado ao quadrado, em seguida, somar tudo e depois dividir por n-1(no nosso caso 10-1 = 9)

$s_x^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{1}{9} \times [(22-27)^2 + (21-27)^2 + (28-27)^2 + (30-27)^2 + (33-27)^2 + (33-27)^2 + (26-27)^2 + (24-27)^2 + (31-27)^2 + (22-27)^2]$.

$= \frac{1}{9} \times [25 + 36 + 1 + 9 + 36 + 36 + 1 + 9 + 16 + 25] = \frac{194}{9} = 21.555.$

$s_y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{1}{9} \times [(25-31)^2 + (28-31)^2 + (25-31)^2 + (36-31)^2 + (32-31)^2 + (39-31)^2 + (28-31)^2 + (33-31)^2 + (37-31)^2 + (27-31)^2]$.

$= \frac{1}{9} \times [36 + 9 + 36 + 25 + 1 + 64 + 9 + 4 + 36 + 16] = \frac{236}{9} = 26.222.$

Temos então que: $\bar{X}$ = 27, $\bar{Y}$ =31, s_x^2=21.555 e s_y^2 = 26.222.

Tendo calculado as estimativas, podemos pensar na hipótese, O nosso interesse é saber se a produtividade média na segunda empresa é maior do que primeira, ou seja, se a média de Y(μ_y) é maior que a média de X(μ_x). Se as médias fossem iguais, a diferença entre elas seria 0, portanto, se a média de Y for maior que a de X, e fizermos a média de Y menos a média de X, nosso resultado seria maior que 0. Teremos então nossas hipóteses definidas por:

$$\begin{cases} H_0 : \mu_y - \mu_x = 0 \\ H_1 : \mu_y - \mu_x \neq 0 \end{cases}$$

Temos que a variável aleatória $T = \frac{(\bar{Y} - \bar{x}) - (\mu_y - \mu_x)}{s_p \times \sqrt{1/n_1 + 1/n_2}}$**, segue distribuição t-student com n1+n2-2 graus de liberdade**, onde $n_1 e n_2$ são os tamanho amostrais de y e x, e s_p**, é o desvio padrão agrupado de X e Y**. O desvio padrão agrupado, pode ser calculado através da fórmula:

$$s_p = \sqrt{\frac{(n_1 - 1) \times s_x^2 + n_2 - 1) \times s_y^2}{n_1 + n_2 - 2}}$$

Calculando o desvio agrupado, teremos:

$s_p = \sqrt{\frac{(n_1-1)\times s_1^2 + n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(10-1)\times 21.555 + (10-1)\times 26.222}{10+10-2}} = \sqrt{\frac{430}{18}} = \sqrt{23.88} = 4.88$

A ideia do teste de hipótese é **supor que a hipótese nula é verdadeira**, e com base nisso observar se o resultado obtido pela amostra está dentre os resultados mais prováveis de ser observado, caso a hipótese nula seja verdadeira. Se supomos que H_0 é verdadeira($\mu_y = \mu_x$), teremos:

$$T_{obs} = \frac{\bar{y} - \bar{x}}{s_p \times \sqrt{1/n_1 + 1/n_2}}$$

Portanto:

$T_{obs} = \frac{\bar{y}-\bar{x}}{s_p \times \sqrt{1/n_1+1/n_2}} = \frac{31-27}{4.88 \times \sqrt{1/10+1/10}} = \frac{4}{2.18} = 1.83$

Agora precisamos definir se o valor é provável ou não de acontecer. Através da tabela t, conseguimos saber qual valor tem probabilidade 5% de que aconteça ele ou uma valor maior(Valor crítico), portanto qualquer valor acima deste estará dentro dos 5% valores menos prováveis. Podemos encontrar o valor crítico na tabela t-student, no cruzamento entre o grau de liberdade(no caso 18) e o nível de significância desejado.

	Área na cauda superior								
gl	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,000	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819

Na tabela acima, temos destacado o valor obtido, portanto qualquer valor acima 1.734 está dentro dos valores menos prováveis.

Nosso valor observado foi 1.83, um valor maior que 1.73, portanto a probabilidade de uma observação dessa forma ocorra caso a hipótese nula seja verdadeira é muito baixa. Isso nos leva a concluir que a hipótese nula não é verdadeira. Ou seja, diremos que a produtividade média na segunda empresa(Y) é maior que na primeira(X).

Capítulo 6

Regressão Linear

Exemplo 6.0.1. *Acredita-se que um bom desempenho em matemática resulta em um bom desempenho em português. Para testar isso uma prova de matemática e uma prova de português foram aplicadas á 10 alunos de uma turma. Os resultados obtidos por cada aluno são mostrados a seguir:*

Aluno	1	2	3	4	5	6	7	8	9	10
Português	9	9	8	5	9	8	6	4	8	8
Matemática	7	8	5	6	8	5	8	9	4	3

1. *Suponha que as notas de português e matemática seguem uma distribuição normal e encontre o intervalos de confiança para a media de cada prova:* $IC(\mu, 95\%)$.
2. *Faça um diagrama de dispersão com os pontos observados*
3. *Faça uma regressão linear simples(use as notas de português como variável dependente) e represente a reta de regressão no diagrama de dispersão.*
4. *Calcule o coeficiente* R^2.
5. *A partir da analise dos valores obtidos nos itens anteriores qual seria a conclusão que teríamos?*
6. *Usando os itens anteriores ache o valor do coeficiente de correlação entre as duas notas.*

Solução

1. **Suponha que as notas de português e matemática seguem uma distribuição normal e encontre o intervalos de confiança para a media de cada prova:** $IC(\mu, 95\%)$ **.**

   O primeiro passo passo para se encontrar um intervalo de confiança é pensar na distribuição dos dados e em seus parâmetros. Sabemos que os dados seguem distribuição normal, e que o

parâmetro de interesse é média. A variância da população é desconhecida, portanto precisaremos utilizar a variância da amostra(s^2) para estimar a variância da população.

Quando utilizamos a **variância da amostra para estimar a da população** teremos que a padronização dos dados seguirá a **distribuição t-Student com n-1 graus** de liberdade:

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

Sabendo disto, podemos fixar o nível e significância e calcular a margem de erro(ME) . Lembrando que a margem de erro é o valor máximo esperado da diferença entre os resultados da amostra e a população.

$$ME = t_{\alpha/2}\sqrt{\frac{s^2}{n}}$$

O valor $t_{\alpha/2}$ é encontrado na tabela t-Student. Por se tratar de um intervalo, o nível de significância deverá ser dividido por 2(por isso a simbologia $t_{\alpha/2}$), portanto fixando um nível de significância de 5% , precisaremos encontrar na tabela o encontro entre os graus de liberdade e 0.025 (0.05/2). Nossas duas amostras são de tamanho 10, como são n-1 graus de liberdade, precisaremos procurar na tabela a linha referente a 9 graus de liberdade.

	Área na cauda superior								
gl	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,000	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587

A tabela acima mostra o valor obtido na tabela t, portanto $t_{\alpha/2} = 2.262$. Agora que sabemos como encontrar a margem de erro. O intervalo será dado pela estiva pontual(Média amostral) menos a Margem de erro até a estimativa pontual mais a margem erro.

- **Português**

  A média e a variância da amostra das notas de português são respectivamente $\bar{x} = 7.4$ e $s^2 = 3.155$. Nossa margem de erro será:

  $$ME = t_{\alpha/2}\sqrt{\frac{s}{n}} = 2.262 \times \sqrt{\frac{3.155}{10}} = 1.27$$

  O intervalo de confiança é dado pela média mais ou menos a margem de erro:

$$IC = \bar{x} \pm ME = 7.4 \pm 1.27 = [7.4 - 1.27; 7.4 + 1.27]$$
$$IC = [6.13; 8.67]$$

Interpretação: Com 95% de confiança, o valor real da média das notas de português está contido no intervalo entre 6.13 e 8.67.

- **Matemática** A média e a variância da amostra das notas de matemática são respectivamente $\bar{x} = 6.3$ e $s^2 = 4.011$. Nossa margem de erro será:

$$ME = t_{\alpha/2}\sqrt{\frac{s}{n}} = 2.262 \times \sqrt{\frac{4.011}{10}} = 1.43$$

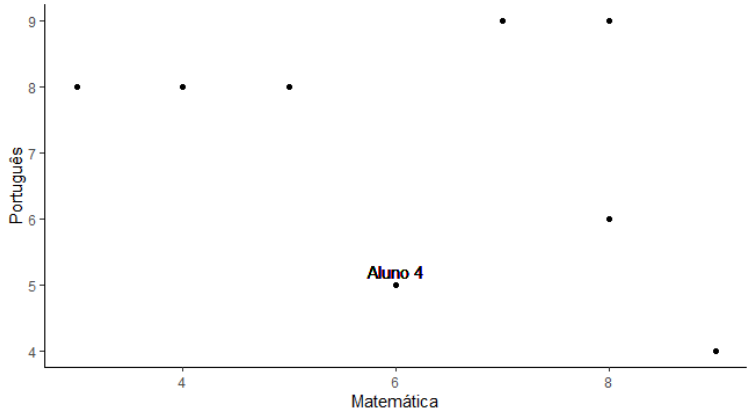
O intervalo de confiança é dado pela média mais ou menos a margem de erro:

$$IC = \bar{x} \pm ME = 6.3 \pm 1.43 = [6.3 - 1.43; 6.3 + 1.43]$$
$$IC = [4.867; 7.732]$$

Com 95% de confiança, o valor real da média das notas de português está contido no intervalo entre 4.867 e 7.732.

2. **Faça um diagrama de dispersão com os pontos observados**

Para criar o diagrama de dispersão, basta traçar sobre o plano cartesiano os pontos nos encontros das notas do mesmo aluno. No eixo y traçaremos as notas de português e no eixo y, as notas de matemática. O aluno 4 por exemplo, será marcado no encontro dos pontos 5 e 6(Destacaremos este ponto no gráfico para exemplificar).



3. **Faça uma regressão linear simples(use as notas de português como variável dependente) e represente a reta de regressão no diagrama de dispersão.**

Numa regressão linear simples, possuímos duas variáveis. A **variável dependente é a que está sendo explicada e a independente é a que queremos utilizar para explicar a variação da variável dependente.**

A ideia é estimar a função que determina a relação entre as duas variáveis, e essa relação é definida através da equação de uma reta.

$$Y = \alpha + \beta \times x + \epsilon$$

Onde Y será a variável dependente e X a variável independente(Portanto português = Y e matemática = X), e os valores α e β são respectivamente o intercepto(ponto onde x =0) e a inclinação da reta, e ϵ indica o efeito aleatório(aquilo que não conseguimos controlar) . A nossa **reta de regressão** é a reta que minimiza o quadrado da diferença entre os valores reais da variável dependente e os valores que estimamos para ela. Então é preciso encontrar os valores de α e β que satisfazem isso.

Os valores que satisfazem isso são encontrados a seguir.

A primeira estimativa a ser encontrada é a de β e ela é dada por:

$$\hat{\beta} = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - (\sum_{i=1}^{n} x_i) \times (\sum_{i=1}^{n} y_i)}{(n \times \sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

Na formula mencionada acima, temos que :

- $\sum_{i=1}^{n}(x_i \times y_i)$ indica que devemos olhar todos os pares x,y (os pares de notas de português e matemática) e multiplicar-los, e em seguida, somar os resultados, temos portanto:
  $(9\times7)+(9\times8)+(8\times5)+(5\times6)+(9\times8)+(8\times5)+(6\times8)+(4\times9)+(8\times4)+(8\times3) = 457$
- $\sum_{i=1}^{n} x_i$ indica que devemos somar todas os valores de X, no caso, todas as notas de matemática:
  $7 + 8 + 5 + 6 + 8 + 5 + 8 + 9 + 4 + 3 = 63$
- $\sum_{i=1}^{n} y_i$ indica que devemos somar todas os valores de y, no caso, todas as notas de português:
  $9 + 9 + 8 + 5 + 9 + 8 + 6 + 4 + 8 + 8 = 74$
- $\sum_{i=1}^{n}(x_i)^2$ indica que devemos elevar todos os valores de x(notas de matemática) ao quadrado depois somar os resultados:
  $7^2 + 8^2 + 5^2 + 6^2 + 8^2 + 5^2 + 8^2 + 9^2 + 4^2 + 3^2 = 433$

Substituindo os valores na formula temos:

$$\hat{\beta} = \frac{10 \times 457 - 63 \times 74}{(10 \times 433) - (63)^2} = \frac{4570 - 4662}{4330 - 3969} = \frac{-92}{361} = -0.2548$$

Portanto, temos que nossa estimativa para beta $\hat{\beta} = -0.2548$. Agora vamos encontrar a estimativa de α, conhecida por $\hat{\alpha}$, que tem sua formula dada por:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \hat{\beta} \times \sum_{i=1}^{n} x_i}{n}$$

Todas as medidas necessárias já foram encontradas anteriormente, portanto, podemos apenas substituir os valores e calcular o valor de $\hat{\alpha}$

$$\hat{\alpha} = \frac{74 - (-0.2548) \times 63}{10} = \frac{74 + 16.0554}{10} = \frac{90.0554}{10} = 9.0055$$

Lembrando que Colocamos esse "chapéu"em cima dos valores por que estamos estimando a reta, então nossa **reta de regressão** estimada será dada por:

$$\hat{Y} = \hat{\alpha} + \hat{\beta} \times X_i$$

Para traçar a reta de regressão precisamos calcular os valores estimados de Y, marcar no gráfico seu encontro com o valor x e traçar a reta que liga estes pontos.

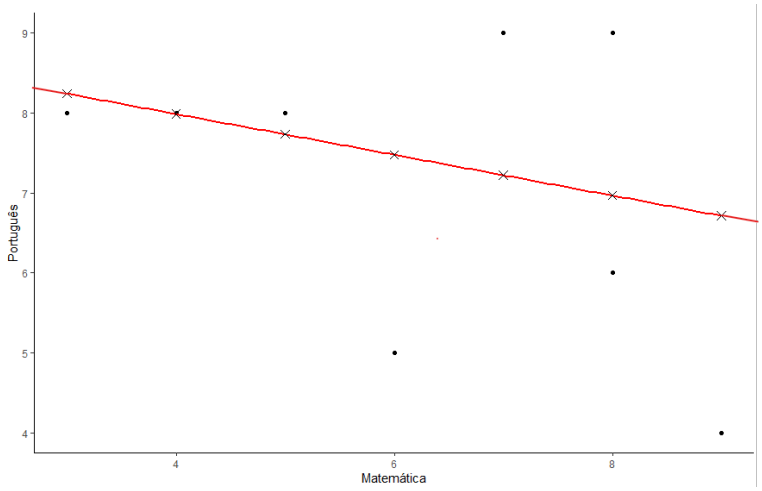
Anteriormente encontramos $\beta_0 = 9.0055$ e $\beta_1 = -0.2548$. Portanto a nossa reta é definida por:

$$\hat{Y} = 9.0055 - 0.2548 \times X_i$$

Agora basta aplicar estes valores na formula da reta estimada e encontrar $\hat{y}$ para cada valor de x.por exemplo, se x = 7 $\hat{Y} = 9.0055 - 0.2548 \times 7 = 7.22$, Fazendo isso, obtemos os seguintes valores:

Aluno	1	2	3	4	5	6	7	8	9	10
Português estimado	7.22	6.96	7.73	7.47	6.96	7.73	6.96	6.71	7.98	8.24
Matemática	7	8	5	6	8	5	8	9	4	3

Agora basta traçar os novos pontos sobre o gráfico e passar a reta sobre eles.



4. **Calcule o coeficiente de determinação R^2**

   O coeficiente de determinação, denotado por R^2 indica o quanto o modelo foi capaz de explicar os dados coletados. Note,que no gráfico que contém a **reta de regressão** que alguns pontos estão longe da reta, isso porque o modelo não explica 100 % da variação das notas de português.

   Antes de calcular o coeficiente , vamos pensar na variação do modelo, e na variação total. Sem nenhum modelo especificado, é razoável pensar na média aritmética como a melhor estimativa e quanto temos um modelo, é razoável pensar que a estimativa feita pelo modelo é mais precisa que a média. Portanto, a **variação do modelo** é calculada pela soma de quadrados da diferença entre o **valor real e o valor predito**. E a **variação total** é calculada pela soma dos quadrados da diferença entre os **valores reais e a média aritmética**, portanto teremos:

$$\text{variação do modelo} = \sum_{i=1} n(y - \hat{y})^2 =$$
$$(9 - 7.22)^2 + (9 - 6.96)^2 + (8 - 7.73)^2 + (5 - 7.47)^2 + (9 - 6.96)^2 + (8 - 7.73)^2 +$$
$$(6 - 6.96)^2 + (4 - 6.71)^2 + (8 - 7.98)^2 + (8 - 8.24)^2 = 26.0554$$

$$\text{variação total} = \sum_{i=1} n(y - \bar{y})^2 =$$
$$(9 - 7.4)^2 + (9 - 7.4)^2 + (8 - 7.4)^2 + (5 - 7.4)^2 + (9 - 7.4)^2 + (8 - 7.4)^2 +$$
$$(6 - 7.4)^2 + (4 - 7.4)^2 + (8 - 7.4)^2 + (8 - 7.4)^2 = 28.4$$

   Para saber o quanto o modelo explica, devemos considerar a diferença entre a variação total e variação explicada, e dividir pela variação total para obter a proporção. teremos então:

$$R^2 = \frac{\text{variação total} - \text{variação do modelo}}{\text{variação total}} = \frac{28.4 - 26.0554}{28.4} = 0.0825$$

   Através da expressão acima, o valor R^2 encontrado foi de 0.0825, que é equivalente a 8.25%, ou seja, no nosso modelo 8.25% da variação das notas de português pode ser explicada com as notas de matemática

5. **A partir da analise dos valores obtidos nos itens anteriores qual seria a conclusão que teríamos?**

   A análise dos resultados de uma regressão deve levar em conta vários fatores. Um deles é o fato do modelo estar bem ajusto, verificamos isso através da análise de resíduos do modelo, mas por hora, vamos supor que o modelo está bem ajustado.

   Usualmente o intercepto não é muito útil, a maior interpretação vem do coeficiente β, que indica o que espera que se aconteça em média com o y ao acrescentarmos 1 unidade em x. Como temos $\beta = -0.2548$, podemos concluir que, o aumento de um ponto nas notas de matemática indicam a diminuição em 0.2548 nas notas de português. E que, por mais que haja essa relação, apenas 8.25% da variação das notas de português está sendo explicado pela nota de matemática, portanto deve-se considerar que há outros fatores que influenciam nas notas de português.

6. **Usando os itens anteriores ache o valor do coeficiente de correlação entre as duas notas.**

   O coeficiente de determinação é o quadrado do coeficiente de correlação, portanto, para encontrar módulo |r| do coeficiente de correlação basta obter a raiz quadrada do coeficiente R^2

   $$|r| = \sqrt{R^2}$$

   Anteriormente, encontramos um coeficiente de determinação igual a 0.0825, portanto, o coeficiente de correlação será : $r = \sqrt{0.0825} = 0.2873$. Basta descobrir se o coeficiente é negativo ou positivo, por quando um valor está elevado ao quadrado, perdemos a informação do sinal. O coeficiente de correlação carregará o mesmo sinal do coeficiente β_1. Temos então que o coeficiente de correlação é -0.2873.

Exemplo 6.0.2. *Após uma reestruturação numa empresa a produtividade média da empresa mudou.A empresa fez uma medição da produtividade de cada trabalhador (antes e depois da reestruturação), numa escala de 20 a 40 pontos. Considere as variáveis, X = "Produtividade Anterior" e Y = "Produtividade Posterior".*

Funcionário	João	Maria	José	Pedro	Rita	Joana	Flávio	Paulo	Catarina	Felipe
Antes	22	21	28	30	33	33	26	24	31	22
Depois	25	28	25	36	32	39	28	33	37	27

1. *Faça um diagrama de dispersão para X e Y.*

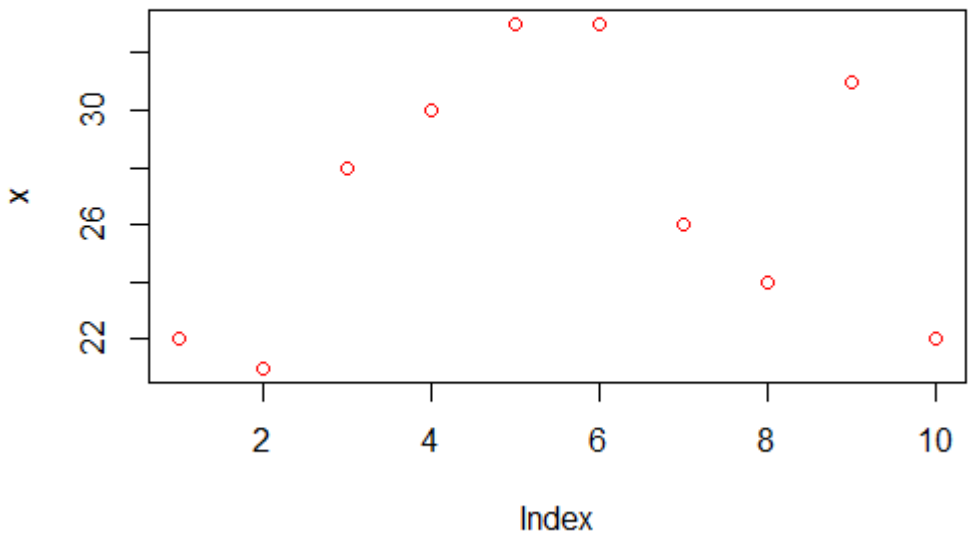
2. *Faça uma regressão linear simples para Y como função de X e represente ela no gráfico da dispersão.*

3. *Ache o coeficiente de correlação r.*

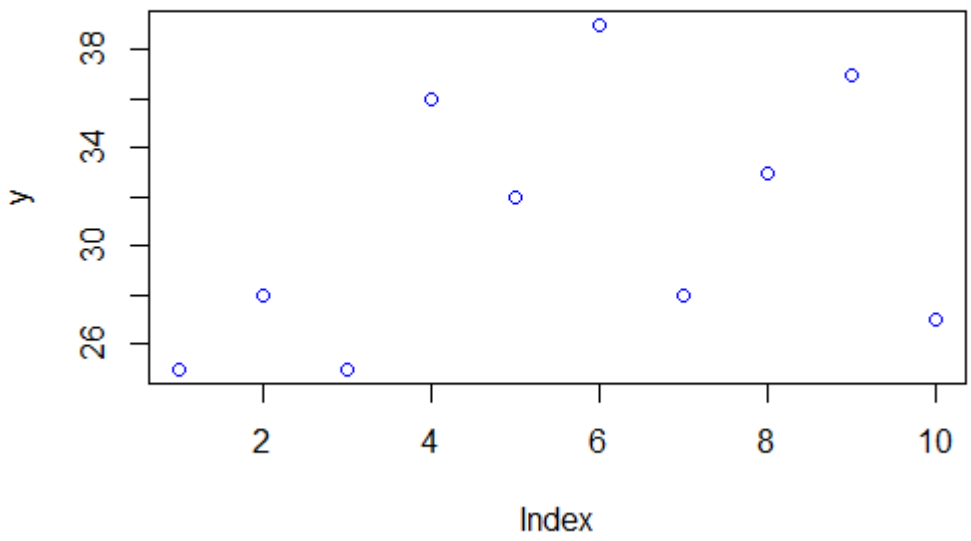
Solução

1. *Faça um diagrama de dispersão para X e Y.*

   Para fazer o diagrama de dispersão basta desenhar, no plano cartesiano, onde se encontra os valores da produtividade de cada funcionário. Para X:



   Para Y:



2. *Faça uma regressão linear simples para Y como função de X e represente ela no gráfico da dispersão.*

   Numa regressão linear simples, possuímos duas variáveis. A ***variável dependente é a que está sendo explicada e a independente é a que queremos utilizar para explicar a variação da variável dependente.***

   A ideia é estimar a função que determina a relação entre as duas variáveis, e essa relação é definida através da equação de uma reta.

$$Y = \alpha + \beta \times x + \epsilon$$

   Onde Y será a variável dependente e X a variável independente(Portanto produtividade posterior = Y e produtividade anterior = X), e os valores α e β são respectivamente o intercepto(ponto onde x =0) e a inclinação da reta, e ϵ indica o efeito aleatório(aquilo que não conseguimos controlar) . A nossa **reta de regressão** *é a reta que minimiza o quadrado da diferença entre os valores reais da variável dependente e os valores que estimamos para ela. Então é preciso encontrar os valores de α e β que satisfazem isso.*

   Os valores que satisfazem isso são encontrados a seguir.

   A primeira estimativa a ser encontrada é a de β e ela é dada por:

$$\hat{\beta} = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - (\sum_{i=1}^{n} x_i) \times (\sum_{i=1}^{n} y_i)}{(n \times \sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

Na formula mencionada acima, temos que :

- $\sum_{i=1}^{n}(x_i \times y_i)$ *indica que devemos olhar todos os pares x,y (os pares de notas de produtividade posterior e produtividade anterior) e multiplicar-los, e em seguida, somar os resultados, temos portanto:*
  $(22 \times 25) + (21 \times 28) + (28 \times 25) + (30 \times 36) + (33 \times 32) + (33 \times 39) + (26 \times 28) + (24 \times 33) + (31 \times 37) + (22 \times 27) = 8522$
- $\sum_{i=1}^{n} x_i$ *indica que devemos somar todos os valores de X, no caso, todos os valores da produtividade anterior:*
  $22 + 21 + 28 + 30 + 33 + 33 + 26 + 24 + 31 + 22 = 270$
- $\sum_{i=1}^{n} y_i$ *indica que devemos somar todos os valores de y, no caso, todos os valores da produtividade posterior:*
  $25 + 28 + 25 + 36 + 32 + 39 + 28 + 33 + 37 + 27 = 310$
- $\sum_{i=1}^{n}(x_i)^2$ *indica que devemos elevar todos os valores de x(produtividade anterior) ao quadrado depois somar os resultados:*
  $22^2 + 21^2 + 28^2 + 30^2 + 33^2 + 33^2 + 26^2 + 24^2 + 31^2 + 22^2 = 7484$

Substituindo os valores na formula temos:

$$\hat{\beta} = \frac{10 \times 8522 - 270 \times 310}{(10 \times 7484) - (270)^2} = \frac{85220 - 83700}{74840 - 72900} = \frac{1520}{1940} = 0.7835$$

Portanto, temos que nossa estimativa para beta $\hat{\beta} = 0.7835$ *. Agora vamos encontrar a estimativa de* α*, conhecida por* $\hat{\alpha}$*, que tem sua formula dada por:*

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \hat{\beta} \times \sum_{i=1}^{n} x_i}{n}$$

Todas as medidas necessárias já foram encontradas anteriormente, portanto, podemos apenas substituir os valores e calcular o valor de $\hat{\alpha}$

$$\hat{\alpha} = \frac{310 - 0.7835 \times 270}{10} = \frac{310 - 211.54}{10} = \frac{98.46}{10} = 9.84$$

Lembrando que Colocamos esse "chapéu"em cima dos valores por que estamos estimando a reta, então nossa **reta de regressão** *estimada será dada por:*

$$\hat{Y} = \hat{\alpha} + \hat{\beta} \times X_i$$

Para traçar a reta de regressão precisamos calcular os valores estimados de Y, marcar no gráfico seu encontro com o valor x e traçar a reta que liga estes pontos.

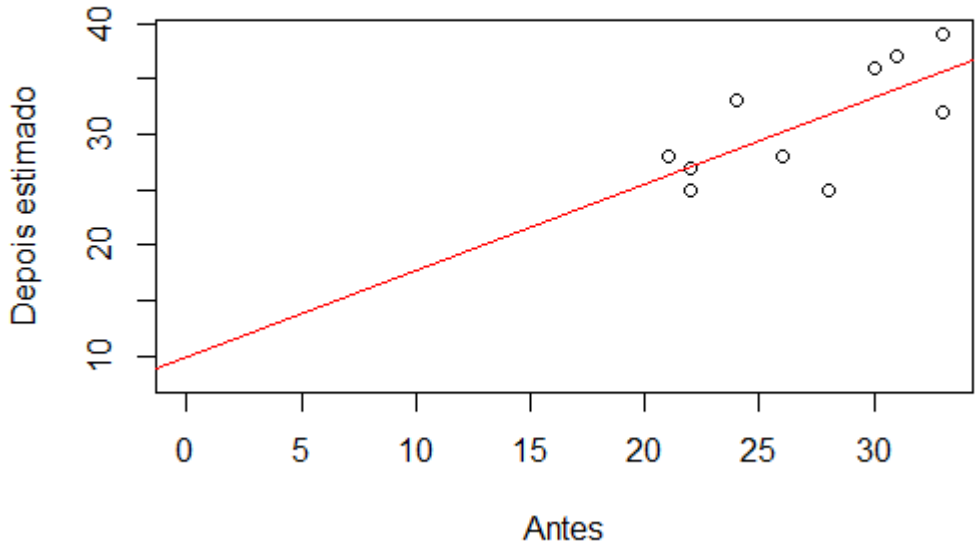
Anteriormente encontramos $\beta_0 = 9.84$ *e* $\beta_1 = 0.7835$ *. Portanto a nossa reta é definida por:*

$$\hat{Y} = 9.84 + 0.7835 \times X_i$$

Agora basta aplicar estes valores na formula da reta estimada e encontrar $\hat{y}$ *para cada valor de x. Por exemplo, se x = 22, então* $\hat{Y} = 9.84 + 0.7835 \times 22 = 27.07$*, Fazendo isso, obtemos os seguintes valores:*

Funcionário	João	Maria	José	Pedro	Rita	Joana	Flávio	Paulo	Catarina	Felipe
Antes	22	21	28	30	33	33	26	24	31	22
Depois estimado	27.07	26.29	31.77	33.34	35.69	35.69	30.21	28.64	34.12	27.07

Agora basta traçar os novos pontos sobre o gráfico e passar a reta sobre eles.



3. *Ache o coeficiente de correlação R.*

   Para encontrar o coeficiente de correlação usamos a seguinte fórmula:

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

   - $\sum(x_i - \bar{x})(y_i - \bar{y})$ *significa somar a multiplicação das diferenças entre cada número do conjunto X e a média de X com as diferenças entre cada número do conjunto Y e a média de Y . Como X significa a produtividade anterior e Y significa a produtividade posterior, então a média de X é* $\bar{x} = 27$ *e a média de Y é* $\bar{y} = 31$. *Logo,* $\sum(x_i - \bar{x})(y_i - \bar{y}) = (22-27).(25-31) + (21-27).(28-31) + (28-27).(25-31) + (30-27).(36-31) + (33-27).(32-31) + (33-27).(39-31) + (26-27).(28-31) + (24-27).(33-31) + (31-27).(37-31) + (22-27).(27-31) = 122$
   - *No denominador (a parte de baixo da fração), n-1 significa o comprimento da coluna menos uma unidade. Como temos 10 colunas de observações então n-1 é igual a* $10 - 1 = 9$.
   - s_x *é o desvio padrão da produtividade anterior. Então* $s_x = 4,64$.
   - s_y *é o desvio padrão da produtividade posterior. Então* $s_y = 5,12$.

   Então, o coeficiente de correlação é igual a

$$r = \frac{122}{9.4,64.5,12} = \frac{122}{213,81} = 0,57$$