Modelos Probabilísticos

Andreas Kneip

9 de maio de 2018

Definições básicas

Os *modelos probabilísticos* são construídos a partir de certas premissas ou hipóteses e consistem de: (1) os possíveis resultados e (2) uma lei que diz quão provável é um determinado resultado (Barbetta 2012).

População e amostra

Uma *população* é um conjunto de elementos sobre os quais queremos tirar conclusões. Para isso fazemos inferências a partir de uma *amostra* dessa população.

Exemplos de população:

- a população carcerária de um país
- a medida da altura de um time de basquete
- nível de instrução dos chefes de família de um determinado bairro da capital

Exemplos de amostra:

- uma parte da população carcerária estratificada por renda familiar, grau de instrução, sexo e local de residência
- jogadores de um determinado time de basquete que terminaram em quadra o último jogo
- moradores da quadra XXX

Inferência estatística

Ao estudar as características de uma determinada amostra de uma população estamos interessados em estimar o valor de um determinado *parâmetro*. Um parâmetro é uma medida que descreve uma determinada característica presente na população.

A *inferência estatística* é o processo de generalizar os resultados obtidos a partir da amostra para toda a população. Pode envolver também os *testes de hipóteses* o *testes de significância*.

Distribuições de probabilidade

- R possui diversos modelos de distribuição de probabilidade.
- Funções para o cálculo de probabilidade, função densidade de probabilidade e sorteios.
- Exemplos:
- beta, binomial, Cauchy, chi-squared, exponential, F, gamma, geometric, hypergeometric, log-normal, logistic, negative binomial, normal, Poisson, Student's t, uniform, Weibull, Wilcoxon e muito mais!!!!

Funções de distribuição de probabilidade do R

Nas várias funções de probabilidade do R, a regra de nomenclatura das funções tem

- d para a função densidade
- p para a função distribuição acumulada
- q para a função quantil
- r gera números aleatórios (random)

Distribuição de probabilidade

Em teoria da probabilidade e em estatística, uma distribuição de probabilidade descreve o comportamento aleatório de um fenômeno dependente do acaso. O estudo dos fenômenos aleatórios começou com o estudo dos jogos de azar – jogos de dados, sorteios de bolas de urna e cara ou coroa eram motivações para compreender e prever os experimentos aleatórios. Essas abordagens iniciais são fenômenos discretos, o que significa que o número de resultados possíveis é finito ou contável. Entretanto, certas questões revelam distribuições de probabilidade com suporte infinito não contável. Por exemplo, quando o lançamento de uma moeda tende ao infinito, o número de coroas aproxima-se de uma distribuição normal.

Flutuações e variabilidade estão presentes em quase todo valor que pode ser medido durante a observação de um fenômeno, independente de sua natureza, além disso quase todas as medidas possuem uma parte de erro intrínseco. A distribuição de probabilidade pode modelar incertezas e descrever fenômenos físicos, biológicos, econômicos, entre outros. O domínio da estatística permite o encontro das distribuições de probabilidade adaptadas aos fenômenos aleatórios.

Há muitas distribuições de probabilidade diferentes. Entre as distribuições de probabilidade, a distribuição normal tem uma importância particular. De acordo com o teorema central do limite, a distribuição normal representa o comportamento assintótico de várias distribuições de probabilidade.

O conceito de distribuição de probabilidade é formalizado matematicamente pela teoria da medida – uma distribuição de probabilidade é uma medida muitas vezes vista como uma distribuição que descreve o comportamento de uma variável aleatória discreta ou contínua. Uma medida é uma distribuição de probabilidade se sua massa total for 1. O estudo de uma variável aleatória de acordo com uma distribuição de probabilidade discreta revela o cálculo de somas e de séries, enquanto que o estudo de uma variável aleatória de acordo com uma distribuição de probabilidade absolutamente contínua revela o cálculo de integrais. As funções particulares permitem caracterizar as distribuições de probabilidade como a função de distribuição e a função característica (Wikipédia 2017).

Distribuição normal

Em probabilidade e estatística, a distribuição normal é uma das distribuições de probabilidade mais utilizadas para modelar fenômenos naturais. Isso se deve ao fato de que um grande número de fenômenos naturais apresenta sua distribuição de probabilidade tão proximamente normal, que a ela pode ser com sucesso referida, e, portanto, com adequado acerto por ela representada como se normal fosse. A distribuição normal é ligada a vários conceitos matemáticos como movimento browniano, ruído branco, entre outros. A distribuição normal também é chamada distribuição gaussiana, distribuição de Gauss ou distribuição de Laplace–Gauss, em referência aos matemáticos, físicos e astrônomos francês Pierre–Simon Laplace (1749 – 1827) e alemão Carl Friedrich Gauss (1777 – 1855).

O papel central da distribuição normal decorre do fato de ser o limite de um grande número de distribuições de probabilidade como mostra o teorema central do limite, o qual permite estudar probabilisticamente a média das variáveis independentes de uma amostra aleatória simples de tamanho grande n. A distribuição normal corresponde ao comportamento do efeito agregado de experiências aleatórias independentes e semelhantes em certas circunstâncias quando o número de experiências é muito alto. Com esta propriedade, a distribuição normal pode aproximar–se da distribuição de efeito agregado de outras distribuições e modelar vários estudos científicos como erros de medição ou testes estatísticos com as tabelas de distribuição normal.

É dada pela função de densidade de probabilidade

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

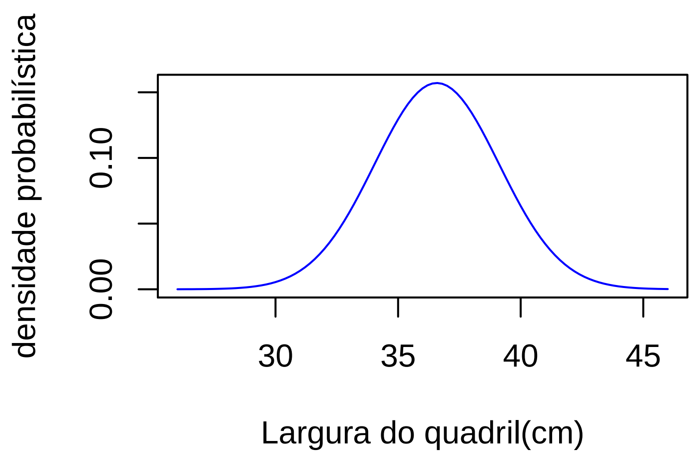
Exemplo

Nos EUA, a média (μ) de quadril dos passageiros masculinos de uma companhia aérea é 36,58 cm, com desvio padrão (σ) de 2,54 cm. Podemos simular uma amostra de 1000 homens, com esta média e com este desvio padrão usando a distribuição normal:

```
rn1 <- rnorm(1000, mean = 36.58, sd = 2.54); head(rn1); tail(rn1)
```

```
## [1] 41.48092 36.80387 35.28899 29.50232 37.14174 38.51800
```

```
## [1] 37.81016 37.68587 32.03897 40.04208 32.58774 36.69597
```



Se a companhia aérea fixar a largura do assento em 38 cm, qual o percentual de passageiros que cabem e que não cabem? Que não cabem

```
pnorm(q=38,mean=36.58,sd=2.54,lower.tail=FALSE)
```

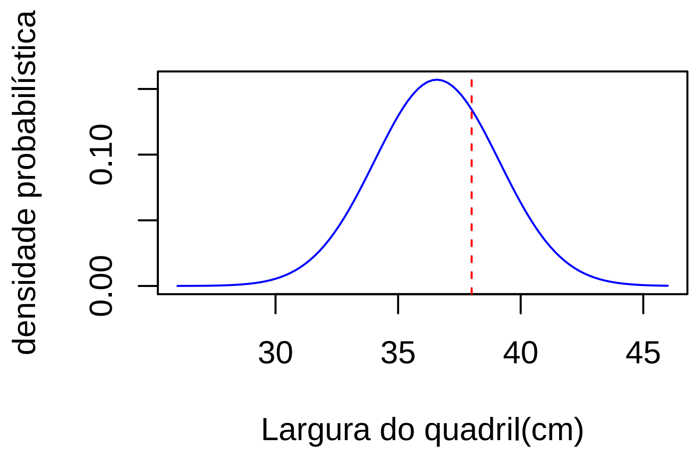
```
## [1] 0.2880621
```

E os que cabem

```
pnorm(q=38,mean=36.58,sd=2.54)
```

```
## [1] 0.7119379
```

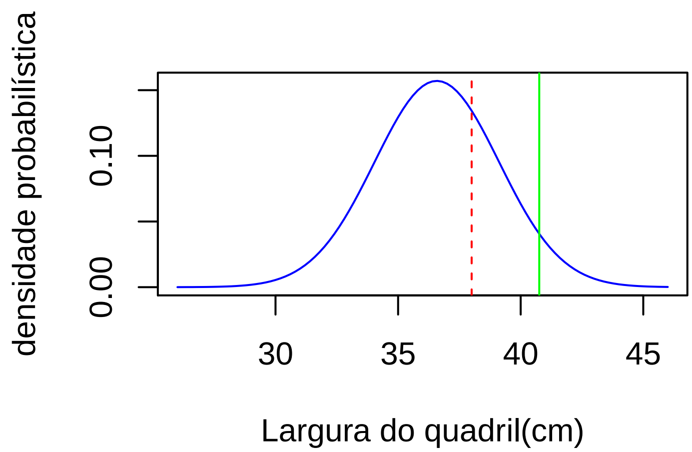
```
# ou 1 - pnorm(q=38,mean=36.58,sd=2.54, lower.tail=FALSE)
```



Qual o tamanho do assento de modo que 95% dos passageiros masculinos caibam?

```r
qnorm(p=0.95,mean=36.58,sd=2.54)
```

```
## [1] 40.75793
```



Distribuição binomial

Em teoria das probabilidades e estatística, a distribuição binomial é a distribuição de probabilidade discreta do número de sucessos numa sequência de n tentativas tais que:

- Cada tentativa tem exclusivamente como resultado duas possibilidades, sucesso ou fracasso (binomial, a que se chama de tentativa de Bernoulli);
- Cada tentativa é independente das demais;
- A probabilidade de sucesso a cada tentativa p permanece constante independente das demais;
- A variável de interesse, ou pretendida, é o número de sucessos k nas n tentativas.

A distribuição binomial de obter k sucessos em n tentativas de probabilidade p é

$$b(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \text{ , para } n = 1, 2, \ldots$$

onde

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Exemplo 1

Em 10 lançamentos de um dado, qual a probabilidade de sair o número 1 duas vezes?

```
dbinom(2,10,1/6)
```

```
## [1] 0.29071
```

Exemplo 2

Numa pessoa normal 60% dos leucócitos são neutrófilos. Se observarmos uma amostra de 10 leucócitos de uma pessoa supostamente normal, qual a probabilidade de encontrarmos pelo menos 8 neutrófilos?

```
sum(dbinom(8:10,10,0.6))
```

```
## [1] 0.1672898
```

Exemplo 3

Qual distribuição de probabilidade de um fazendeiro ter 12 filhas?

Probabilidade de 0 a 12 "sucessos" em 12 tentativas.

Probabilidade de cada sucesso: 0.5

```
probs <- dbinom(0:12, size = 12, prob = 0.5); probs
```

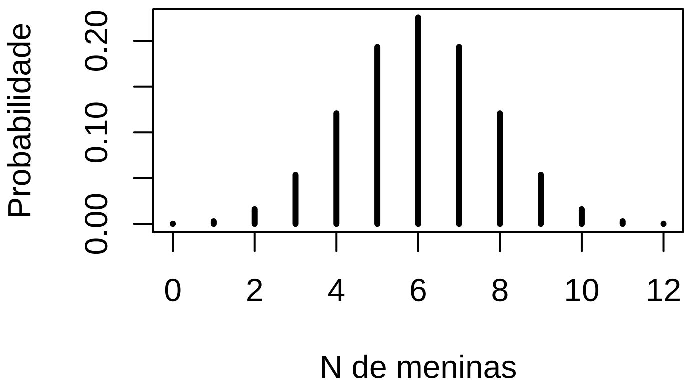
```
##  [1] 0.0002441406 0.0029296875 0.0161132813 0.0537109375 0.1208496094
##  [6] 0.1933593750 0.2255859375 0.1933593750 0.1208496094 0.0537109375
## [11] 0.0161132813 0.0029296875 0.0002441406
```

Acrescentando um nome para cada elemento do objeto criado acima

```
names(probs) = 0:12; round(probs, 5)
```

```
##       0       1       2       3       4       5       6       7       8
## 0.00024 0.00293 0.01611 0.05371 0.12085 0.19336 0.22559 0.19336 0.12085
##       9      10      11      12
## 0.05371 0.01611 0.00293 0.00024
```

```
plot(0:12, probs, type= "h", lwd=3, xlab="N de meninas",
     ylab="Probabilidade")
```

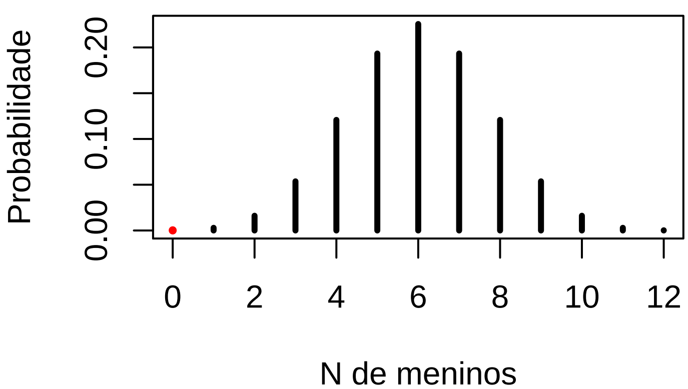


Qual a probabilidade de não ter meninos?

```r
dbinom(0, size=12, prob=0.5)
```

```
## [1] 0.0002441406
```

Adicionando isto ao gráfico



Distribuição uniforme

Em estatística e probabilidade, a distribuição uniforme é a distribuição de probabilidades contínua mais simples de conceituar: a probabilidade de se gerar qualquer ponto em um intervalo contido no espaço amostral é proporcional ao tamanho do intervalo, visto que na distribuição uniforme a $f(x)$ é igual para qualquer valor de x no intervalo considerado. Por exemplo, se considerarmos um intervalo em x de zero à dez positivo

($x \in [0, 10]$), e assumirmos que temos uma distribuição uniforme nesse intervalo, a probabilidade de $f(x)$ no intervalo $ 2= 5/6) p <- p+1 }; p

[1] 1

“‘

Distribuição geométrica

Em teoria das probabilidades e estatística, a distribuição geométrica é constituída por duas funções de probabilidade discretas:

- a distribuição de probabilidade do número X de tentativas de Bernoulli necessárias para alcançar um sucesso, suportadas pelo conjunto { 1, 2, 3, ... }, ou
- a distribuição de probabilidade do número $Y = X - 1$ de insucessos antes do primeiro sucesso, suportadas pelo conjunto { 0, 1, 2, 3, ... }.

Se a probabilidade de sucesso de cada tentativa é p, então a probabilidade de n tentativas serem necessárias para ocorrer um sucesso é

$$P(X = n) = (1 - p)^{n-1}p$$

para n = 1, 2, 3, ...

De forma equivalente, a probabilidade de serem necessários n insucessos antes do primeiro sucesso é

$$P(Y = n) = (1 - p)^n p$$

para n = 0, 1, 2, 3, ...

Em qualquer caso, a sequência de probabilidades é uma progressão geométrica.

Por exemplo, suponha um dado que é atirado repetidamente até à primeira vez que aparece um “1”. A probabilidade de distribuição do número de vezes que o dado é atirado é suportado pelo conjunto infinito { 1, 2, 3, ... } e é uma distribuição geométrica com $p = 1/6$.

O valor esperado de uma variável aleatória geometricamente distribuída X é $1/p$ e a variância é $(1 - p)/p^2$;

$$E(X) = \frac{1}{p}, \quad \mathrm{var}(X) = \frac{1-p}{p^2}.$$

De forma equivalente, o valor esperado de uma variável aleatória geometricamente distribuída Y é $(1 - p)/p$, e a sua variância é $(1 - p)/p^2$.

Como a sua distribuição contínua análoga (a distribuição exponencial), a distribuição geométrica tem a propriedade de perda de memória. Isto significa que se se tentar repetir uma experiência antes do primeiro sucesso, então, dado que o primeiro sucesso ainda não ocorreu, a função de distribuição condicional do número de tentativas adicionais não depende de quantos insucessos foram observados até então. A distribuição geométrica é, de facto, a única distribuição discreta com esta propriedade

Exemplo 1

Sabe-se que 10% das peças produzidas por certa máquina apresentam defeitos. Num determinado dia, qual a probabilidade de que a terceira peça produzida por esta máquina seja a primeira defeituosa? Lembre-se que a distribuição geométrica é dada pela expressão

$$P(X = n) = (1 - p)^{n-1} * p$$

Neste caso, o "sucesso" é apenas a terceira peça estar defeituosa, então n=3 e p=0,1.

```
dgeom(2,0.1)
```

```
## [1] 0.081
```

Exemplo 2

O custo na fabricação de um produto é R$ 1.000,00. Supostamente terminado, o produto é submetido a um teste. Se aprovado, encerra-se o processo, caso contrário perde-se R$ 400 das peças do produto, que são substituídas e em seguida é feito um novo teste, assim sucessivamente até a obtenção de êxito no teste. Sabendo-se que em 90% das vezes o teste é realizado com sucesso, determine a probabilidade de que o custo de um produto supere R$ 1800,00.

```
(1-sum(dgeom(0:2, 0.9))) * 100
```

```
## [1] 0.1
```

Distribuição de Pascal

(Ou binomial negativa)

A distribuição binomial negativa ou distribuição de Pascal é uma distribuição de probabilidade discreta. Esta distribuição indica o número de tentativas necessárias para obter k sucessos de igual probabilidade θ ao fim de n experimentos de Bernoulli, sendo a última tentativa um sucesso. A sua função de probabilidade é dada por:

$$b(n; k, p) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \text{ , para } n = k, k+1, \ldots$$

Exemplo

Um dado é lançado sucessivamente até se obter a face "1" pela terceira vez. Qual a probabilidade de serem necessários exatemente cinco lançamentos?

```
dnbinom(2,3,1/6)
```

```
## [1] 0.01929012
```

Distribuição hipergeométrica

Em teoria das probabilidades e estatística, a distribuição hipergeométrica é uma distribuição de probabilidade discreta que descreve a probabilidade de k sucessos em n retiradas, sem reposição, de uma população de tamanho N que contém exatamente K sucessos, sendo cada retirada um sucesso ou um fracasso. Em contraste, a distribuição binomial descreve a probabilidade de k sucessos em n retiradas com reposição.

Exemplo 1: Suponhamos que num grupo composto por 12 pessoas existam 7 mulheres. Escolhendo-se simultaneamente 6 pessoas do grupo, qual a probabilidade de serem sorteadas 4 mulheres?

```r
dhyper(4, 7, 12-7, 6)
```

```
## [1] 0.3787879
```

Exemplo

Pequenos motores elétricos são expedidos em lotes de 50 unidades. Antes de uma remessa ser aprovada, inspeciona-se uma amostra com 5 motores. Se nenhum dos 5 for defeituoso, o lote é aprovado. Caso contrário todo os motores são inspecionados. Suponha que existam 3 motores defeituosos no lote. Qual a probabilidade de que a inspeção 100% seja necessária?

```r
dhyper(0, 3, 50-3,5)
```

```
## [1] 0.7239796
```

Distribuição de Poisson

Na teoria da probabilidade e na estatística, a distribuição de Poisson é uma distribuição de probabilidade de variável aleatória discreta que expressa a probabilidade de uma série de eventos ocorrer num certo período de tempo se estes eventos ocorrem independentemente de quando ocorreu o último evento.

A distribuição foi descoberta por Siméon Denis Poisson (1781–1840) e publicada, conjuntamente com a sua teoria da probabilidade, em 1838 no seu trabalho *Recherches sur la probabilité des jugements en matières criminelles et matière civile* ("Inquérito sobre a probabilidade em julgamentos sobre matérias criminais e civis"). O trabalho focava em certas variáveis aleatórias N que contavam, entre outras coisas, o número de ocorrências discretas de um certo fenômeno durante um intervalo de tempo de determinada duração. A probabilidade de que existam exatamente k ocorrências (k sendo um inteiro não negativo, $k = 0, 1, 2, \ldots$) é

$$f(k; \lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}$$

onde

- e é base do logaritmo natural ($\mathrm{e} = 2.71828\ldots$),
- k! é o fatorial de k,
- λ é um número real, igual ao número esperado de ocorrências que ocorrem num dado intervalo de tempo. Por exemplo, se o evento ocorre a uma média de 4 minutos, e estamos interessados no número de eventos que ocorrem num intervalo de 10 minutos, usariámos como modelo a distribuição de Poisson com $\lambda = 10/4 = 2.5$.

Como função de k, esta é a função de probabilidade. A distribuição de Poisson pode ser derivada como um caso limite da distribuição binomial.

Exemplo 1

Suponhamos que os acidentes de trabalho numa fábrica ocorram segundo as hipóteses de Poisson com uma média de 3 acidentes a cada 5 meses. Qual a probabilidade de ocorrer: exatamente um acidente num mês? E dois acidentes em 2 meses? E algum acidente no próximo mês, sabendo-se que há um mês não ocorre acidente algum.

```r
dpois(1, 3/5); dpois(2, 2*3/5); 1 - dpois(0, 3/5)
```

```
## [1] 0.329287
```

```
## [1] 0.2168598
```

```
## [1] 0.4511884
```

Exemplo 2

A chegada de navios petroleiros, numa refinaria, obedece uma distribuição de Poisson. O número médio de navios que chegam por dia é 2. As instalações da refinaria permitem aportar até 3 navios por dia, sendo que os excedentes a 3 deverão seguir para outro porto. Suponha-se ainda que cada navio permaneça por um dia na refinaria. Pede-se:

1. Em um dia, qual a probabilidade de ser necessário mandar algum petroleiro para outro porto?
2. Qual a probabilidade de que a partir de um dado instante passe 8 horas sem chegar navio algum?
3. Em quanto as atuais instalações devem ser aumentadas para permitir manobrar todos os petroleiros em 94% dos dias?

Solução do exemplo 2

1

```
1 - sum( dpois( 0:3, 2 ) )
```

```
## [1] 0.1428765
```

2

```
dpois( 0, 2 * 1/3 )
```

```
## [1] 0.5134171
```

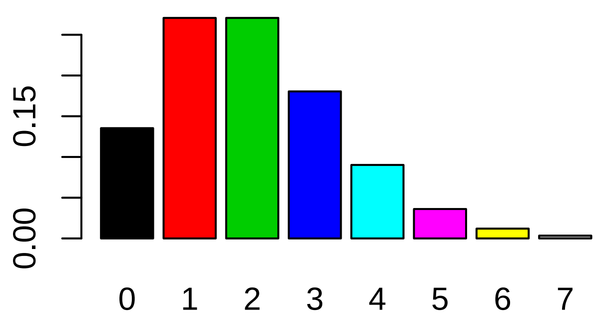
3

```
qpois(.94, 2)
```

```
## [1] 4
```

Qual o número mais provável de petroleiros que chegarão num certo dia?

```
barplot( dpois(0:7, 2), names=0:7, col=1:8)
```



Qual o número médio de petroleiros atentidos por dia?

```
0 * dpois(0,2) + 1 * dpois(1,2) + 2 * dpois(2,2) +
 3 * (1 - sum(dpois(0:2, 2)))
```

```
## [1] 1.781982
```

Exercícios

1. Suponha que o tempo necessário para atendimento de clientes em uma central de atendimento telefônico siga uma distribuição normal de média de 8 minutos e desvio padrão de 2 minutos.
   (a) Qual é a probabilidade de que um atendimento dure menos de 5 minutos?
   (b) E mais do que 9,5 minutos?
   (c) E entre 7 e 10 minutos?
   (d) 75% das chamadas telefônicas requerem pelo menos quanto tempo de atendimento?
2. A distribuição dos pesos de coelhos criados numa granja pode muito bem ser representada por uma distribuição Normal, com média 5 kg e desvio padrão 0,9 kg. Um abatedouro comprará 5000 coelhos e pretende classificá-los de acordo com o peso do seguinte modo: 15% dos mais leves como pequenos, os 50% seguintes como médios, os 20% seguintes como grandes e os 15% mais pesados como extras. Quais os limites de peso para cada classificação?
3. Uma enchedora automática de refrigerantes está regulada para que o volume médio de líquido em cada garrafa seja de 1000 cm3 e desvio padrão de 10 m3. Admita que o volume siga uma distribuição normal.
   (a) Qual é a porcentagem de garrafas em que o volume de líquido é menor que 990 cm3?
   (b) Qual é a porcentagem de garrafas em que o volume de líquido não se desvia da média em mais do que dois desvios padrões?
   (c) Se 10 garrafas são selecionadas ao acaso, qual é a probabilidade de que, no máximo, 4 tenham volume de líquido superior a 1002 cm3?
   (d) Se garrafas vão sendo selecionadas até aparecer uma com volume de líquido superior a 1005 cm3, qual é a probabilidade de que seja necessário selecionar pelo menos 5 garrafas?
4. Uma empresa produz televisores de 2 tipos, tipo A (comum) e tipo B (luxo), e garante a restituição da quantia paga se qualquer televisor apresentar defeito grave no prazo de seis meses. O tempo para ocorrência de algum defeito grave nos televisores tem distribuição normal sendo que, no tipo A, com média de 10 meses e desvio padrão de 2 meses e no tipo B, com média de 11 meses e desvio padrão de 3 meses. Os televisores de tipo A e B são produzidos com lucro de 1200 u.m. e 2100 u.m. respectivamente e, caso haja restituição, com prejuízo de 2500 u.m. e 7000 u.m. Respectivamente.
   (a) Calcule as probabilidades de haver restituição nos televisores do tipo A e do tipo B.
   (b) Calcule o lucro médio para os televisores do tipo A e para os televisores do tipo B.
   (c) Baseando-se nos lucros médios, a empresa deveria incentivar as vendas dos aparelhos do tipo A ou do tipo B?
5. Suponha que a variável aleatória X tenha uma distribuicão geométrica com $p = 0.5$. Determine as seguintes probabilidades:
   (a) $P(X = 1)$
   (b) $P(X = 4)$
   (c) $P(X = 8)$
   (d) $P(X \leq 2)$
   (e) $P(X > 2)$
6. Suponha que cada uma de suas chamadas telefônicas para um serviço de reclamações tenha uma probabilidade de 0.02 de se completar, ou seja, não obter o sinal de ocupado. Considere que suas chamadas são independentes.
   (a) Qual é a probabilidade de sua primeira chamadaa se completar seja a décima chamada?
   (b) Qual é a probabilidade de se necessitar de mais de cinco chamadas para que a ligação se complete?
   (c) Qual o número médio necessário de chamadas para que a ligação se complete?

7. Uma companhia emprega 800 homens com 55 anos de idade. Suponha 30% carregue um marcador no cromossomo masculino, que indique um risco crescente de alta pressão sanguínea.
    (a) Se 10 homens na companhia forem testados em relação ao marcador, qual será a probabilidade de exatamente um homem ter esse marcador?
    (b) Se 10 homens na companhia forem testados com relação ao marcador, qual será a probabilidade de mais de um homem ter esse marcador?
8. Os astrônomos tratam o número de estrelas em um dado volume do espaço como sendo uma variável aleatória de Poisson. A densidade na Via Láctea, na vizinhança do nosso sistemasolar, é uma estrela por 16 (anos-luz)^3.
    (a) Qual a probabilidade de duas ou mais estrelas em 16(anos-luz)^3?
    (b) Quantos (anos-luz)^3 de espaço devem ser estudados de modo que a probabilidade de uma ou mais estrelas exceda 95%?
9. A espessura de um flange em um componente de espaçonave é uniformemente distribuídaentre 0.95 e 1.05 mm.
    (a) Determine a proporção de flanges que excedem1.02 mm.
    (b) Qual o valor da espessura que é excedida por 90% dos flanges?
    (c) Determine a média e a variância da espessura do flange.
10. A resistência à compressão de amostras de cimento podemser modeladas por uma distribuição normal, com média de 6000 kg/cm^2 e um desvio padrão de 100 kg/cm^2.
    (a) Qual a probabilidade da resistência da amostra ser menor do que 625 kg/cm^2?
    (b) Qual a probabilidade da amostra estar entre 5800 e 5900kg/cm^2?
    (c) Que a resistência é excedida por 95% das amostras?

Referências

Barbetta, Pedro Alberto. 2012. *Estatística Aplicada às Ciências Humanas*. Florianópolis/SC: Editora da UFSC.

Wikipédia. 2017. "Distribuição de Probabilidade." https://pt.wikipedia.org/wiki/Distribuiç~ao_de_probabilidade.