Calculando probabilidades para distribuições normais e não normais

Introdução

Aqui exploramos diferentes formas de se calcular probabilidade, uma vez coletada uma amostra. O primeiro exemplo trata um caso com distribuição normal e o segundo exemplo com distribuição não normal.

Resolvemos os dois exemplos usando três ferramentas diferentes:

- Excel
- Minitab: http://www.minitab.com
- Universal Probability Calculator: https://dunamath.com/calculadora.aspx

Exemplo 1 – Caso com distribuição Normal

<u>Enunciado</u>:

Suponha que você tenha medido algumas vezes o tempo para ir de sua casa ao trabalho (tabela a seguir). Você descobre que em média leva 53 minutos e deseja saber qual a chance de chegar ao trabalho em menos de 1 hora.

Measured values (minutes)		
52.7	43.5	43.3
59.2	47.8	65.2
38.7	51.7	53.6
54.3	67.9	49.7
51.6	63.8	53.6

<u>Solução usando “Universal Provability Calculator (UPC)”:</u>

No passo 1 fica como na figura abaixo:

I wish to know the odds of getting a number < ▼ than the value 60

No passo 2, copiar e colar diretamente do arquivo para o campo do site:

52.7 43.5 43.3 59.2 47.8 65.2 38.7 51.7 53.6 54.3 67.9 49.7 51.6 63.8 53.6

Após clicar em “Calculate” vê-se que a chance de chegar ao trabalho em até 1 hora (60 minutos) é de 81.9%.

<u>Solução usando “Excel - Windows”:</u>

Em Data -> Data Analysis -> Descriptive Statistics obtêm-se os valores abaixo:

Mean	53.10165
Standard Error	2.137853
Median	52.6883
Mode	#N/A
Standard Deviation	8.279871
Sample Variance	68.55626
Kurtosis	-0.36244
Skewness	0.208165
Range	29.1862

Como Kurtosis e Skewness são razoavelmente próximos de zero, podemos assumir que a distribuição é normal, ou pelo menos próxima disto. A amostra tem tamanho 15, relativamente pequena. Deste modo um teste apropriado é o *t-test*.

Temos $t = \frac{(x-\bar{x})}{s} = \frac{(60-53.102)}{8.280} = 0.833$ com grau de liberdade igual a 14.

Minimum	38.7058
Maximum	67.892
Sum	796.5247
Count	15

Usando o comando *T.DIST(0.833,14,1)*, temos que a probabilidade de obter um valor de até 60 é de 79.06%.

Solução usando "Minitab":

Inicialmente fazemos o teste de normalidade. No Minitab: Stat-> Basic Statistics -> Normality Test. Para Anderson-Darling e Kolmogorov-Smirnov temos os resultados ao lado, ambos não rejeitando a hipótese nula de normalidade. Sendo assim, é razoável assumir que a distribuição é normal.

Mean	53.10
StDev	8.280
N	15
AD	0.303
P-Value	0.531

Mean	53.10
StDev	8.280
N	15
KS	0.175
P-Value	>0.150

Como a amostra é pequena, usaremos o *t-test*. No Minitab: Calc -> Probability Distributions-> t. Seleciona-se "cumulative probability", e no campo "input constant" colocamos 0.833 (mesmo t calculado anteriormente para Excel), e obtemos a resposta ao lado, ou seja, a probabilidade de obter um valor de até 60 é de 79.06%.

Cumulative Distribution Function

```
Student's t distribution with 14 DF

       x  P( X ≤ x )
0.833147    0.790626
```

Discussão dos resultados:

Inicialmente, sobre a fonte dos dados, gerou-se 20 mil valores utilizando-se o software Matlab, função: $(randn(20000,1) * 5) + 50$. Assumiu-se que esta é a população. A partir disto coletou-se aleatoriamente 15 valores desta população, listados no enunciado da questão.

Resumo dos resultados:

UPC-Dunamath	Excel	Minitab	Resposta Correta
81.9%	79.06%	79.06%	97.72%

Observa-se que Excel e Minitab retornaram o mesmo resultado, o que é esperado, já que para ambos usou-se a distribuição de Student, com o mesmo parâmetro t. Em ambos assumiu-se que a distribuição é normal, o que é correto, pois os dados da população foram gerados a partir de uma distribuição normal. No entanto os parâmetros de média e desvio usados para calcular t estão significativamente errados, ou seja, a média da amostra é 53.1 e desvio padrão de 8.28, enquanto que a média da população é 50 com desvio de 5. Isto explica o erro na resposta.

Outro ponto é que, mesmo usando Excel ou Minitab corretamente, provavelmente o tomador de decisão acreditaria nos 79.06% obtido, pois estas ferramentas não fornecem informações sobre o tamanho da incerteza dos cálculos.

O Universal Probability Calculator (UPC) retornou uma probabilidade de 81.9%, um pouco melhor que Excel e Minitab, mas ele também informa que o nível de confiança é baixo (64%), alertando o usuário sore isto.

O UPC, além de calcular a probabilidade de uma forma simples, dá uma estimativa do tamanho da incerteza envolvida. De modo que se o decisor quiser um nível de certeza maior, precisará aumentar o tamanho da amostra. Isto parece ser mais justo com o tomador de decisão.

Exemplo 2 – Caso com distribuição Não Normal

Enunciado:

Como engenheiro de produto, você está estudando o tempo de vida de um disco rígido de computador. Num experimento, você obteve o tempo de vida em horas de 10 discos, como a seguir.

1988.77	2026.69
2074.94	2018.67
1973.65	1921.29
1941.77	1937.22
1895.03	1942.83

A) Qual a probabilidade do tempo de vida ser maior que 1900 horas?

Solução usando "Universal Provability Calculator (UPC)":

No passo 1 fica:

I wish to know the odds of getting a number [> ▾] than the value [1900] .

Note que no campo acima poderia ter-se selecionado ≥. Como os valores tratam-se de variáveis contínuas, isto não é relevante.

No passo 2, copiar e colar diretamente do arquivo para o campo do site:

2027.96 2195.97 2075.78 2038.31 1955.24 1940.5 2070.76 2140.07 1817.75 2006.99

Após clicar em "Calculate" vê-se que a probabilidade do tempo de vida ser maior que 1900 horas é de 90.56%.

Solução usando "Excel - Windows":

Em Data -> Data Analysis -> Descriptive Statistics obtêm-se os valores abaixo:

Mean	1972.086
Standard Error	17.48647
Median	1958.24
Mode	#N/A
Standard Deviation	55.29706
Sample Variance	3057.765
Kurtosis	-0.35796
Skewness	0.552605
Range	179.91
Minimum	1895.03
Maximum	2074.94
Sum	19720.86
Count	10

Como Kurtosis e Skewness são razoavelmente próximos de zero, podemos assumir que a distribuição é normal, ou pelo menos próxima disto. A amostra tem tamanho 10, pequena, e considerando-se que a variância da população é desconhecida, considera-se apropriado o *t-test*.

Temos $t = \frac{(x-\bar{x})}{s} = \frac{(1900-2026.933)}{106.759} = -1.189$ com grau de liberdade igual a 9.

Usando o comando *T.DIST(-1.189,9,1)*, temos que a probabilidade da vida útil ser maior que 1900 é de 86.8%.

Solução usando "Minitab":

Inicialmente fazemos o teste de normalidade. No Minitab: Stat-> Basic Statistics -> Normality Test. Para Anderson-Darling e Kolmogorov-Smirnov temos os resultados ao lado, ambos não rejeitando a hipótese nula de normalidade. Sendo assim, é razoável assumir que a distribuição é normal.

Mean	1972
StDev	55.30
N	10
AD	0.253
P-Value	0.653

Mean	1972
StDev	55.30
N	10
KS	0.202
P-Value	>0.150

Como a amostra é pequena e a variância da população não é conhecida, usaremos o *t-test*. No Minitab: Calc -> Probability Distributions-> t. Seleciona-se "cumulative probability", e no campo "input constant" colocamos -1.304 (mesmo t calculado anteriormente para Excel), e obtemos a resposta ao lado, ou seja, a probabilidade de obter um valor menor que 1900 é de 13.2%. (resultado al lado), ou seja, a probabilidade de ser maior que 1900 é $(100 - 11.2) = 88.8\%$.

Cumulative Distribution Function

```
Student's t distribution with 9 DF

     x  P( X ≤ x )
-1.304    0.112298
```

B) Quão certo você está disto? (seu grau de certeza)
Pelo UPC-Dunamath, temos:
"We are 68% confidence that the true value is between 85.56% and 95.56%".

Ou seja, estamos 68% confiantes que o verdadeiro valor seja entre 85.56% e 95.56%. Isto também significa dizer que, se você coletar outras 15 amostras, e fizer isto um número muito grande de vezes, pelo menos 68% das amostras, a probabilidade calculada estará entre 85.56% e 95.56%.

Note que Excel e Minitab não disponibilizam esta informação.

C) Para aumentar seu grau de certeza na análise, você esperou a conclusão do teste de vida de mais 30 discos (teste complementar) e refez o cálculo de probabilidade com as 2 amostras juntas. Qual a probabilidade do tempo de vida ser maior que 1900 horas e qual a confiabilidade do novo resultado?

1988.77	2026.69	2053.48	2140.11	2132.87	2062.56	1970.53	2164.22
2074.94	2018.67	1982.92	1924.92	2154.11	1788.89	2046.63	2019.41
1973.65	1921.29	1968.29	1753.65	1972.47	2028.2	2000.97	1960.72
1941.77	1937.22	1943.67	1957.47	1909.35	2018.27	2102.17	1695.47
1895.03	1942.83	2063.94	1678.59	1948.96	2050.25	1899.61	2058.53

Solução usando "Universal Provability Calculator (UPC)":

No passo 1 fica:

I wish to know the odds of getting a number [> ▼] than the value [1900] .

No passo 2, copiar e colar diretamente do arquivo para o campo do site:

1988.77 2026.69 2053.48 2140.11 2132.87 2062.56 1970.53 2164.22 2074.94 2018.67 1982.92 1924.92

Note que no campo acima há mais valores no lado direito, basta percorrer com o cursor do mouse.

Após clicar em "Calculate" vê-se que a probabilidade do tempo de vida ser maior que 1900 horas é de 85.09%, com 79% de certeza que o valor correto está entre 80.09% e 90.09%.

Solução usando "Excel - Windows":

Em Data -> Data Analysis -> Descriptive Statistics obtêm-se os valores abaixo:

Mean	1979.302
Standard Error	17.43848
Median	1978.285
Mode	#N/A
Standard Deviation	110.2906
Sample Variance	12164.02
Kurtosis	1.321178
Skewness	-0.90893
Range	485.63
Minimum	1678.59
Maximum	2164.22
Sum	79172.09
Count	40

Kurtosis e Skewness não são próximos de zero, mas nem tão longe. É mais seguro assumir que a distribuição não é normal.

No Excel não há uma forma imediata de tratar distribuições não normais. Uma alternativa em nosso caso, em que Kurtosis e Skewness não estão longe de zero, seria optar pela distribuição de Student com $t = \frac{(x-\bar{x})}{s} = \frac{(1900-1979.30)}{110.29} = -0.719$, Excel comando *T.DIST(-0.719,39,1)*, resultado em 76.2%.

Uma outra alternativa ainda simples no Excel seria usar a distribuição empírica, como mostrado abaixo.

A tabela de distribuição empírica (EDF) pode ser montada como a seguir:

X(i)	q < X(i)	EDF <x	EDF > x	X(i)	q < X(i)	EDF <x	EDF > x
1678.59	1	0.025	0.975	1982.92	21	0.525	0.475
1695.47	2	0.05	0.95	1988.77	22	0.55	0.45
1753.65	3	0.075	0.925	2000.97	23	0.575	0.425
1788.89	4	0.1	0.9	2018.27	24	0.6	0.4
1895.03	5	0.125	0.875	2018.67	25	0.625	0.375
1899.61	6	0.15	0.85	2019.41	26	0.65	0.35
1909.35	7	0.175	0.825	2026.69	27	0.675	0.325
1921.29	8	0.2	0.8	2028.2	28	0.7	0.3
1924.92	9	0.225	0.775	2046.63	29	0.725	0.275
1937.22	10	0.25	0.75	2050.25	30	0.75	0.25
1941.77	11	0.275	0.725	2053.48	31	0.775	0.225
1942.83	12	0.3	0.7	2058.53	32	0.8	0.2
1943.67	13	0.325	0.675	2062.56	33	0.825	0.175
1948.96	14	0.35	0.65	2063.94	34	0.85	0.15
1957.47	15	0.375	0.625	2074.94	35	0.875	0.125
1960.72	16	0.4	0.6	2102.17	36	0.9	0.1
1968.29	17	0.425	0.575	2132.87	37	0.925	0.075
1970.53	18	0.45	0.55	2140.11	38	0.95	0.05
1972.47	19	0.475	0.525	2154.11	39	0.975	0.025
1973.65	20	0.5	0.5	2164.22	40	1	0

Na tabela de distribuição empírica, primeira coluna tem os valores em ordem crescente, a segunda coluna tem para cada valor a quantidade de valores menores ou iguais a ele (coincide com o número da linha), a terceira coluna tem o valor da segunda coluna dividido pelo tamanho da amostra (ou seja, a frequência cumulativa), e finalmente, a quarta coluna tem o complemento da terceira coluna.

Queremos a probabilidade de obter um valor maios que 1900. O valor 1900 está entre o valor das linhas 6 e 7, ou seja, 1899.61 and 1909.35. Desta forma, pode-se dizer que a probabilidade de ser maior que 1900 é alguma coisa entre 82.5% e 85%. Note-se que não há garantias de que o valor correto esteja neste intervalo. Mas como a distribuição provavelmente não é normal, este método permite ter alguma noção da probabilidade desejada.

Solução usando "Minitab":

Inicialmente fazemos o teste de normalidade. No Minitab: Stat-> Basic Statistics -> Normality Test. Para Anderson-Darling rejeitou-se a hipótese nula de normalidade. Sendo assim, não é razoável assumir que a distribuição é normal.	Mean 1979 StDev 110.3 N 40 AD 0.866 P-Value 0.024	Mean 1979 StDev 110.3 N 40 KS 0.126 P-Value 0.104

Como a distribuição não é normal, precisamos estimar o tipo de distribuição. No Minitab: *Stat > Quality Tools > Individual Distribution Identification.*

Obtem-se a tabela ao lado, com o teste Anderson-Darling aplicado a vários tipos de distribuição. Em geral, todos aqueles com $P < 0.05$ são imediatamente descartados. Dentre os restantes, escolhe a distribuição com maior valor de P.

Em nosso caso seria Johnson Transformation, a seguir Box-Cox Transformation, e a seguir Weibull. Como as duas primeiras são transformações e não distribuição nativas, e não possuem uma forma de uso direta no Minitab, fiquemos aqui com a Weibull.

Goodness of Fit Test

Distribution	AD	P
Normal	0.866	0.024
Box-Cox Transformation	0.431	0.292
Lognormal	1.080	0.007
3-Parameter Lognormal	0.854	*
Exponential	16.522	<0.003
2-Parameter Exponential	7.955	<0.010
Weibull	0.481	0.229
3-Parameter Weibull	0.473	0.149
Smallest Extreme Value	0.521	0.192
Largest Extreme Value	2.498	<0.010
Gamma	0.991	0.014
3-Parameter Gamma	15.045	*
Logistic	0.469	0.199
Loglogistic	0.559	0.101
3-Parameter Loglogistic	0.469	*
Johnson Transformation	0.291	0.591

O passo anterior também fornece a tabela a seguir, com os parâmetros de cada tipo de distribuição. Em nosso caso, para Weibull, há 2 parâmetros: 22.30053 (shape) and 2027 (scale).

```
ML Estimates of Distribution Parameters

Distribution                  Location       Shape        Scale     Threshold
Normal*                     1979.30225                110.29062
Box-Cox Transformation*    3.12590E+16              7.95334E+15
Lognormal*                     7.58893                  0.05736
3-Parameter Lognormal         13.66510                  0.00013  -8.58377E+05
Exponential                                          1979.30225
2-Parameter Exponential                               308.42270    1670.87943
Weibull                                  22.30053    2027.25391
3-Parameter Weibull                      15.06220    1377.13817     649.03970
Smallest Extreme Value      2029.44972                 90.07752
Largest Extreme Value       1920.56168                128.21685
Gamma                                   318.17708      6.22076
3-Parameter Gamma                      6641.41982      1.34126   -7015.18695
Logistic                    1987.39914                 57.97085
Loglogistic                    7.59408                  0.02962
3-Parameter Loglogistic       13.66441                  0.00007  -8.57775E+05
Johnson Transformation*        0.00497                  0.87821
```

A seguir, no Minitab: *Calc -> Probability Distributions->Weibull*. Selecionar "cumulative probability", digitar os 2 parâmetros da distribuição, e no campo "input constant", entrar com o valor 1900 desejado.

Assim, obteremos a resposta abaixo:

Cumulative Distribution Function

```
Weibull with shape = 22.3005 and scale = 2027.25

   x  P( X ≤ x )
1900    0.209888
```

Como queremos a probabilidade para valores maiores que 1900, temos $1 - 0.2098 = 0.7902 = 79.02\%$. Ufa, finalmente.

Discussão dos resultados:

Inicialmente, sobre a fonte dos dados, gerou-se 20 mil valores utilizando-se o software Matlab, função: $wblrnd(2042.6, 25.8773, 20000, 1)$ gerando-se uma população com distribuição Weibull, média 2000.3 e desvio-padrão 97.192. A partir disto coletou-se aleatoriamente os valores desta população, listados no enunciado da questão.

Para 10 amostras iniciais:

UPC-Dunamath	Excel	Minitab	Resposta Correta
90.56%	88.8%	88.8%	85.82%

Para 40 amostras:

UPC-Dunamath	Excel (using Student Distribution)	Excel (empirical distribution)	Minitab	Resposta Correta
85.09%	76.2%	[82.5% - 85%]	79.02%	85.82%

Para tabela com 10 amostras, observa-se que Excel e Minitab retornaram o mesmo resultado, o que é esperado, já que para ambos usou-se a distribuição de Student, com o mesmo parâmetro t. Embora aprovado no teste de normalidade, a distribuição da população é Weibull.

Os parâmetros de média e desvio usados para calcular t são 1972.09 e 55.30 respectivamente, enquanto que a média da população é 2000.3 com desvio de 97.192. Note que embora tenha assumidose a distribuição errada, o erro da probabilidade foi pequeno, apenas uma coincidência influenciada pela relação entre a média e o desvio-padrão da amostra.

Em relação ao caso com 40 amostras, tanto no Excel como Minitab, o teste de normalidade rejeitou a suposição de normal. No Excel, uma forma de ter uma noção do resultado foi usar distribuição empírica, dando algo em torno do intervalo de 82.5% a 85%, mostrando-se ser uma estimativa até razoável.

Já no Minitab, depois de todo o trabalho identificando-se a melhor distribuição e seus parâmetros, o resultado até piorou em relação ao caso com 10 amostras, o que é possível, por se tratarem de amostras pequenas, talvez as novas amostras fossem menos representativas da população ou apenas uma coincidência numérica.

Finalmente, observa-se quão complicado estas análises podem ser. Calcular o valor de probabilidade já é complicado, e no final você obtém um resultado cuja incerteza é desconhecida. O *Universal Probability Calculator (UPC)*, torna o cálculo imediato, e ainda lhe dá a informação adicional da incerteza, sem precisar se preocupar com todas as premissas e picuinhas estatísticas, pois tudo isto é analisado pelo algoritmo sem envolver o usuário. *Autor: Douglas Moura Miranda*