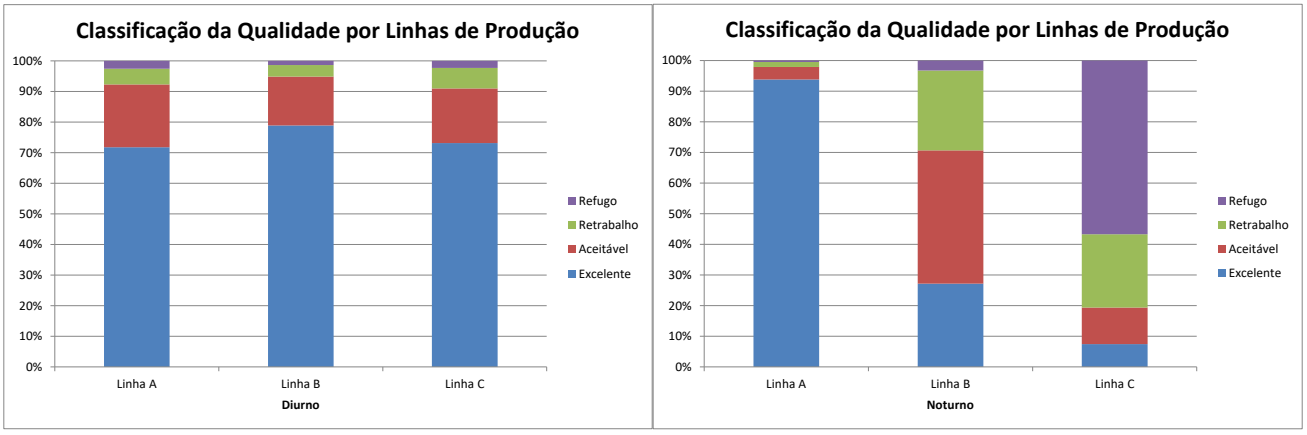
INE6006 – Métodos Estatísticos – Exercícios Prova 1- Semestre 2019.1

1) Certo fabricante de ferramentas dispõe de três linhas de produção (A, B e C), que são operadas em dois turnos de 8 horas (diurno e noturno). Houve reclamações sobre a qualidade dos produtos, então a direção resolveu intensificar a vigilância avaliando a qualidade (classificada como excelente, aceitável, retrabalho ou refugo) das peças produzidas nas três linhas nos dois turnos. Os resultados estão nas tabelas e gráficos a seguir.

Diurno		Qualidade				Total
Linha		Excelente	Aceitável	Retrabalho	Refugo	
A	Frequência	420	120	30	15	585
	% por linha	71,79%	20,51%	5,13%	2,56%	100,00%
	% por coluna	31,44%	37,50%	33,71%	41,67%	32,85%
B	Frequência	568	115	27	10	720
	% por linha	78,89%	15,97%	3,75%	1,39%	100,00%
	% por coluna	42,51%	35,94%	30,34%	27,78%	40,43%
C	Frequência	348	85	32	11	476
	% por linha	73,11%	17,86%	6,72%	2,31%	100,00%
	% por coluna	26,05%	26,56%	35,96%	30,56%	26,73%
Total	Frequência	1336	320	89	36	1781
	% por linha	75,01%	17,97%	5,00%	2,02%	100,00%
	% por coluna	100%	100%	100%	100%	100%
Noturno		Qualidade				Total
Linha		Excelente	Aceitável	Retrabalho	Refugo	
A	Frequência	575	25	10	3	613
	% por linha	93,80%	4,08%	1,63%	0,49%	100,00%
	% por coluna	79,31%	9,43%	4,76%	1,44%	43,54%
B	Frequência	125	200	120	15	460
	% por linha	27,17%	43,48%	26,09%	3,26%	100,00%
	% por coluna	17,24%	75,47%	57,14%	7,21%	32,67%
C	Frequência	25	40	80	190	335
	% por linha	7,46%	11,94%	23,88%	56,72%	100,00%
	% por coluna	3,45%	15,09%	38,10%	91,35%	23,79%
Total	Frequência	725	265	210	208	1408
	% por linha	51,49%	18,82%	14,91%	14,77%	100,00%
	% por coluna	100%	100%	100%	100%	100%



a) Qual é a qualidade predominante no turno diurno? E no noturno? JUSTIFIQUE.

Observar a linha Total nas duas tabelas. No diurno a qualidade predominante é a Excelente com 75,01% (percentual por linha). No noturno é a Excelente com 51,49% do total.

b) O ideal era a produção total distribuir-se igualmente entre as três linhas de produção. Isso ocorre no diurno? E no noturno? JUSTIFIQUE.

Observar a coluna Total nas duas tabelas. No diurno a distribuição é desigual: 32,85% (percentual por coluna) na linha A, 40,43% na B e 26,73% na C. No noturno a distribuição também é desigual: 43,54% na A, 32,67% na B e 23,79% na C. Uma distribuição igual exigiria em torno de 33% de cada linha de produção nos dois turnos.

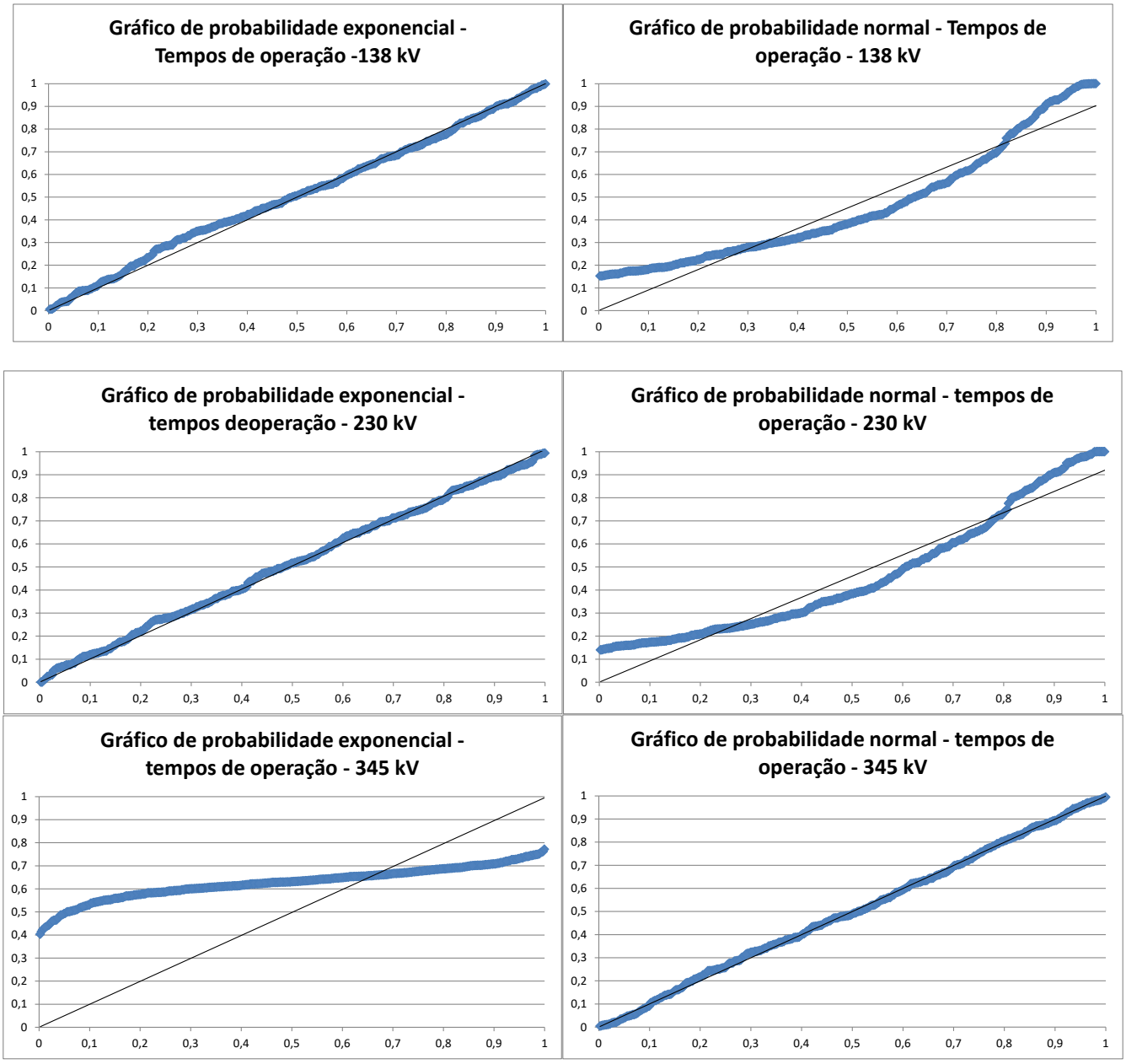
c) Existe associação entre a qualidade das peças e a linha de produção no diurno? E no noturno? JUSTIFIQUE.
Observar a linha total nas duas tabelas. Se não houvesse relação os percentuais por linha nas linhas de produção deveriam ser semelhantes aos do total.
No turno diurno deveriam ser 75,01% de excelente, 17,97% de aceitável, 5% de retrabalho e 2,02% de refugo: isso realmente ocorre, os % por linha são semelhantes aos citados (por exemplo, 71,79% de excelente na linha A, 78,89% na linha B e 73,11% na C, não se afastando 5% de 75,01%[1]). Então NÃO HÁ relação entre as variáveis linha de produção e qualidade no turno diurno.
No turno noturno deveriam ser 51,49% de excelente, 18,82% de aceitável, 14,91% de retrabalho e 14,77% de refugo: há grandes diferenças de uma linha para outra, na A há 93,80% de excelente, contra 17,24% na B e apenas 7,46% de excelente na C. Então HÁ relação entre as variáveis linha de produção e qualidade no turno noturno.
É possível visualizar as semelhanças das linhas de produção do diurno pelo gráfico de colunas 100% empilhadas à esquerda: observem a grande semelhança entre as barras, com mais de 70% da produção considerada excelente. Já no gráfico do noturno as barras são muito diferentes entre si: quase a totalidade da produção da linha A é considerada excelente, e mais de 50% da produção da linha C é considerada refugo.

2) Uma concessionária de transmissão de energia elétrica observou os tempos de operação (em horas) de três tipos de suas linhas de transmissão (138, 230 e 345 kV). Os resultados referentes a 400 registros de cada tipo são mostrados nas tabelas e gráficos a seguir.

Classes (h)	138 kV		230 kV		345 kV		Total	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
1,92|-3940,295	114	28,50%	214	53,50%	0	0,00%	328	27,33%
3940,295|-7878,67	115	28,75%	104	26,00%	183	45,75%	402	33,50%
7878,67|-11817,045	57	14,25%	48	12,00%	216	54,00%	321	26,75%
11817,045|-15755,42	41	10,25%	23	5,75%	1	0,25%	65	5,42%
15755,42|-19693,795	24	6,00%	2	0,50%	0	0,00%	26	2,17%
19693,795|-23632,17	19	4,75%	6	1,50%	0	0,00%	25	2,08%
23632,17|-27570,545	10	2,50%	3	0,75%	0	0,00%	13	1,08%
27570,545|-31508,92	6	1,50%	0	0,00%	0	0,00%	6	0,50%
31508,92|-35447,295	2	0,50%	0	0,00%	0	0,00%	2	0,17%
35447,295|-39385,67	5	1,25%	0	0,00%	0	0,00%	5	0,42%
39385,67|-43324,045	3	0,75%	0	0,00%	0	0,00%	3	0,25%
43324,045|-47262,42	0	0,00%	0	0,00%	0	0,00%	0	0,00%
47262,42|-51200,795	2	0,50%	0	0,00%	0	0,00%	2	0,17%
51200,795|-55139,17	1	0,25%	0	0,00%	0	0,00%	1	0,08%
55139,17|-59077,545	0	0,00%	0	0,00%	0	0,00%	0	0,00%
59077,545|-63015,92	1	0,25%	0	0,00%	0	0,00%	1	0,08%
TOTAL	400	100%	400	100%	400	100%	1200	100%

Medidas	138 kV	230 kV	345 kV	Total
Média (horas)	9652,47	5065,16	8017,09	7578,24
Mediana (horas)	6838,95	3679,76	7981,62	7024,02
Qi (horas)	3416,90	1650,28	7083,04	3608,23
Qs (horas)	12644,98	6946,82	9014,95	9280,02
D.padrão (horas)	9355,95	4692,55	1474,39	6386,47
CV%	96,93%	92,64%	18,39%	84,27%
Qs-Md (horas)	5806,03	3267,06	1033,33	2256,00
Md-Qi (horas)	3422,04	2029,48	898,58	3415,79
Qs-Qi (horas)	9228,07	5296,54	1931,90	5671,79
Qs+1,5 × (Qs-Qi) (horas)	26487,09	14891,62	11912,80	17787,71
Qi-1,5 × (Qs-Qi) (horas)	-10425,21	-6294,52	4185,19	-4899,45
Mínimo (horas)	50,82	1,92	4130,52	1,92
Máximo (horas)	63015,52	25684,39	11829,40	63015,52
Assimetria	2,05	1,65	-0,09	2,84
Curtose	5,75	3,33	-0,12	13,99

[1] Raciocínio semelhante pode ser obtido pelos percentuais por coluna.



a) Com base apenas na tabela agrupada em classes há diferenças entre os valores dos tempos de operação de uma tensão para outra? JUSTIFIQUE.

Observem a coluna Total. Se não houver diferença entre os tipos de linha de transmissão os percentuais de cada classe devem ser semelhantes aos da coluna Total. Não é o que ocorre. Por exemplo, na classe 1,92|-3940,295 horas os percentuais deveriam ser em torno de 27,33%, mas são muito diferentes: 28,50% nas linhas de 138 kV está próximo, mas 53,50% nas linhas de 230kV e 0% mas linhas de 345 kV estão muito distantes. A diferença é especialmente gritante nas linhas de 345 kV onde <u>todos</u> os tempos estão concentrados 3940,295 e 11817,045 horas. Então HÁ diferenças entre os valores dos tempos de operação de um nível de tensão para outro.

b) Com base apenas nas medidas de síntese do TOTAL de tempos, caracterize a tendência central, dispersão, assimetria, curtose e existência de discrepantes do tempo de operação.

Tendência central

O valor típico de tempo de operação oscila entre 7024,02 (mediana – 50% dos tempos abaixo e 50% acima deste valor) e 7578,24 horas (média).

Dispersão

A variação total do tempo de operação é de 1,92 (mínimo) a 63015,52 horas (máximo), com um desvio padrão de 6386,47 horas, que representa 84,27% (CV%) da média (o que parece uma grande dispersão).

Assimetria

A assimetria vale 2,84 (se igual a zero significa simetria), média e mediana apresentam diferenças (7578,24 e 7024,02 horas), e a diferença entre quartil superior e mediana (2256,00 horas) é diferente da existente entre mediana e quartil inferior (3415,79 horas, quase 1200 horas de diferença). Tudo isso aponta para uma distribuição assimétrica.

Curtose

A distribuição deve ser leptocúrtica, pois o valor de curtose (13,99) está consideravelmente acima de zero. Se fosse mesocúrtica o valor seria próximo a 0, e se platicúrtica, menor.

Existência de discrepantes

Valores menores do que -4899,45 horas seriam discrepantes inferiores, o que é impossível. Como o valor mínimo é 1,92 horas, não há discrepantes inferiores. Valores maiores do que 17787,71 horas seriam discrepantes superiores. Como o valor máximo é 63015,52 horas, portanto maior do que 17787,71 horas, pode-se afirmar que há no mínimo um valor discrepante superior de tempo de operação (o próprio valor de máximo).

c) Com base apenas nas medidas de síntese há evidência de diferença nos tempos de operação em função da tensão das linhas? JUSTIFIQUE.

As médias e medianas são diferentes nos três níveis de tensão, em 138 kV a média está em 9652,47 horas (a maior de todas), enquanto em 230 kV está em 5065,16 h (a menor). As medianas também são diferentes, mas chama a atenção que nas linhas de 345 kV média (8017,09 h) e mediana (7981,62h) estão próximas, mas também diferentes das medidas dos outros níveis de tensão. Comportamento semelhante ocorre nos quartis: inferiores (3416,90 h em 138 kV, 1650,28 h em 230 kV e 7083,04 h em 345 kV), e superiores (12644,98 h em 138 kV, 6946,82 h em 230 kV e 9014,95 h em 345 kV). Conclui-se então que há evidência de diferenças entre os tempos, podendo chegar a quase 3000 h entre a mediana das linhas de 138 kV (6838,95 h) e 230 kV (3679,76 h). Novamente, em uma situação real precisaríamos de gráficos, e possivelmente realizar testes de hipóteses (para comparar as médias dos tempos dos níveis de tensão 2 a 2, ou uma análise de variância, para avaliar ser há diferenças significativas entre as médias dos três níveis de tensão).

d) Com base apenas nas medidas de síntese há evidência de que os tempos de operação dos três níveis de tensão sigam a distribuição normal? JUSTIFIQUE.

Para seguir exatamente uma distribuição normal as medidas de assimetria e de curtose precisam ambas ser iguais a zero, ou bastante próximas de zero. Isso ocorre apenas nas linhas de 345 kV, assimetria = -0,09 e curtose = -0,12. Nas demais os valores são substancialmente diferentes de zero, indicando distribuições diferentes da normal. É possível ver isso nos gráficos de probabilidade em seguida.

e) Em estudos de confiabilidade geralmente supõe-se que os tempos de operação sigam a distribuição exponencial. Observando os gráficos de probabilidade dos tempos de operação dos diferentes níveis de tensão pode-se concluir que a suposição é satisfeita? JUSTIFIQUE.

No gráfico de probabilidade comparam-se os valores obtidos de probabilidade (através da frequência relativa) dos dados com os valores esperados se os mesmos dados seguissem uma distribuição teórica (exponencial, normal, etc.). Se os pontos se distribuírem sobre uma reta teórica com coeficiente linear igual a zero admite-se a aderência dos dados à distribuição teórica em análise. Nas linhas de 138 kV e 230 kV os pontos apresentam-se sobre as retas nos gráficos da distribuição exponencial, mas não nos da distribuição normal (coincide com a conclusão do item d, onde estes dois tipos de linha tinham medidas de assimetria e curtose diferentes de zero, implicando que suas distribuições não eram normais). Já nas linhas de 345 kV os pontos estão muito distantes da reta no gráfico da exponencial, mas com muito boa aderência no da distribuição normal (novamente coincidindo com o resultado do item d em que as medidas de assimetria e curtose eram próximas de zero).

3) O laboratório de análises clínicas DIAGNOSIS tem 5 filiais (I, II, III, IV, V) que enviam material para avaliação na sede. Em uma semana típica 35% do material avaliado vêm da filial I, 25% da II, 10% da III, 9% da IV e 21% da V. Estatísticas anteriores mostram que 0,8% do material da filial I apresenta alguma contaminação por bactérias, 0,3% do material da filial II apresenta contaminação, 0,2% do material da filial III apresenta contaminação, 0,4% do material da filial IV apresenta contaminação e 0,1% do material da filial V apresenta contaminação por bactérias.

a) Para uma semana típica, construa a expressão para cálculo da probabilidade de que haja contaminação por bactérias no material avaliado na sede do laboratório, desenvolva-a até a forma mais simplificada possível.

Haverá contaminação se o material oriundo da filial I estiver contaminado OU o material oriundo da filial II estiver contaminado OU ... Transformando em operação com eventos chamando contaminado de C:

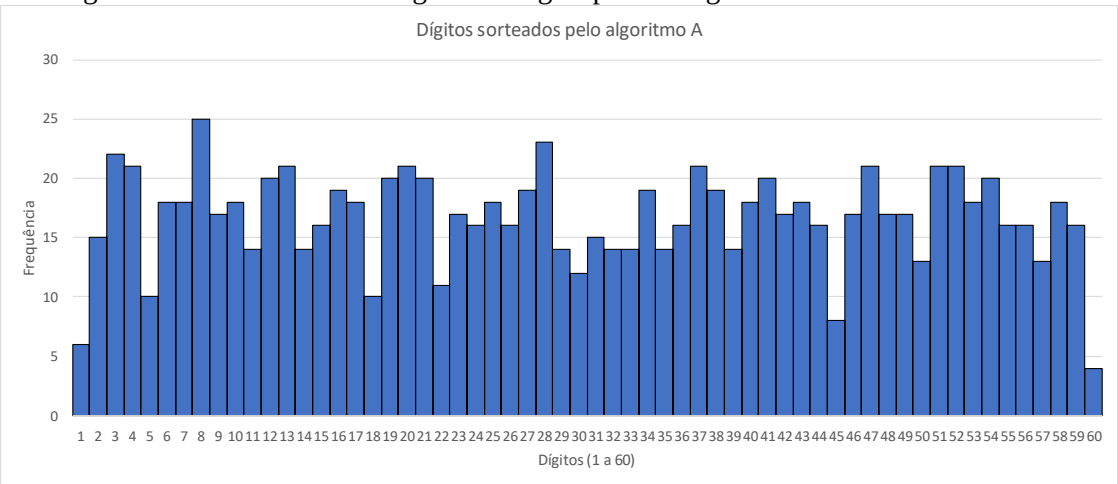
$P(C) = P[(I \cap C) \cup (II \cap C) \cup (III \cap C) \cup (IV \cap C) \cup (V \cap C)]$

Os cinco eventos são mutuamente exclusivos, não há informações indicando mistura de material:

$P(C) = P(I \cap C) + P(II \cap C) + P(III \cap C) + P(IV \cap C) + P(V \cap C)$.

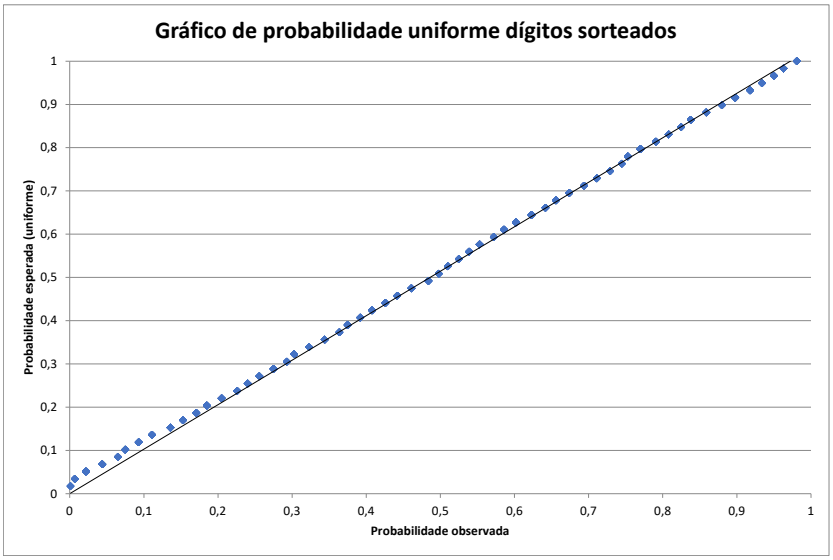
Usando a propriedade do produto:
$P(C) = P(I)\times P(C|I) + P(II)\times P(C|II) + P(III)\times P(C|III) + P(IV)\times P(C|IV) + P(V)\times P(C|V)$
E este seria o desenvolvimento mais simplificado.
Substituindo as probabilidades (não precisam fazer isso, é meramente ilustrativo):
$P(C) = 0{,}35\times 0{,}008 + 0{,}25\times 0{,}003 + 0{,}1\times 0{,}002 + 0{,}09\times 0{,}004 + 0{,}21\times 0{,}001 = 0{,}00432\ (0{,}432\%).$
b) Para uma semana típica, construa a expressão para cálculo da probabilidade que o material avaliado tenha vindo da filial I, supondo que tenha apresentado contaminação por bactérias, desenvolva-a até a forma mais simplificada possível.
Trata-se de um caso de probabilidade condicional: $P(I|C)$[2]
$P(I|C) = P(I\cap C)/P(C) = P(I)\times P(C|I)/P(C)$ *E este seria o desenvolvimento mais simplificado.*
Substituindo as probabilidades (não precisam fazer isso, é meramente ilustrativo):
$P(I|C) = 0{,}35\times 0{,}008/0{,}00432 = 0{,}648148\ (64{,}8148\%)$. *Se o material estiver contaminado há 64,8148% de probabilidade de ter vindo da filial I.*

4) Para os casos a seguir identifique qual é o modelo probabilístico mais apropriado. JUSTIFIQUE sua resposta.
a) Estudos históricos mostram que o número de falhas em uma linha de transmissão de 440 kV possui uma taxa aproximadamente constante de 0,005 falhas por ano. Há interesse em calcular a probabilidade de que ocorra mais de uma falha na linha em um período de 5 anos.
Trata-se de uma variável aleatória discreta, número de falhas em uma linha de transmissão em um período de 5 anos. Mas, não há limite superior para as suas realizações, apenas um período de tempo contínuo para a análise (5 anos), e uma taxa de ocorrência considerada constante. Conclui-se então que o modelo de Poisson é o mais apropriado para este caso: procura-se $P(X > 1)$.
b) A empresa Monte Carlo foi contratada para desenvolver um algoritmo de sorteio para a nova loteria do país sul-americano Pindorama, que permitirá apostar em seis números inteiros entre 1 e 60. Após 1000 simulações de sorteio com o algoritmo A foi obtido o histograma a seguir para os dígitos sorteados.



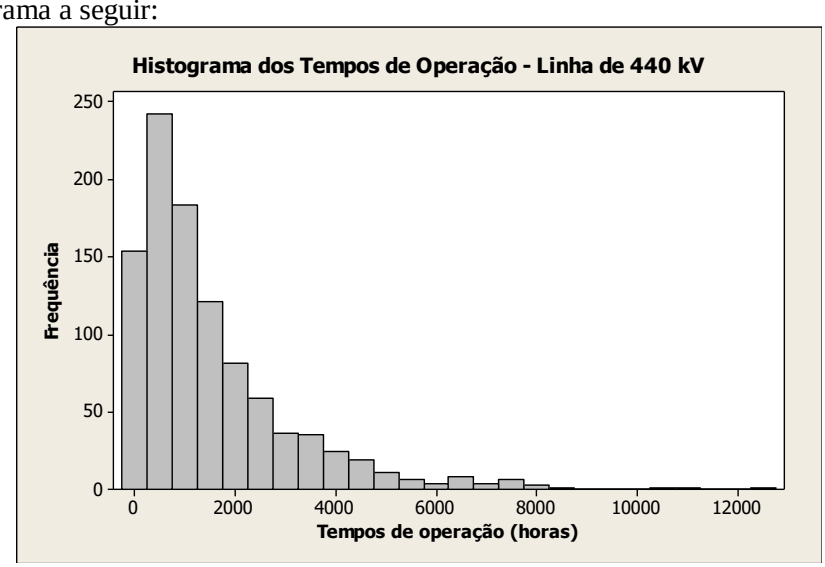
Embora haja algumas flutuações (os dígitos 1, 45 e 60 têm frequências menores) a maioria dos dígitos teve frequência entre 15-20 sorteios, ou seja, frequências <u>semelhantes</u>, e próximas ao valor que deveria ter cada uma delas se os 1000 sorteios fossem divididos igualmente pelos 60 dígitos (1000/60 = 16,66666). Portanto, e observando o formato "retangular" do histograma NÃO agrupado[3] *acima, conclui-se que a distribuição dos dígitos pode ser considerada uniforme.*
Apenas para fins ilustrativos vejam o gráfico de probabilidade uniforme a seguir, feito com os dígitos sorteados:

[2] Trata-se de um caso de aplicação do teorema de Bayes.
[3] Ou seja, não há manipulação com os limites das classes.

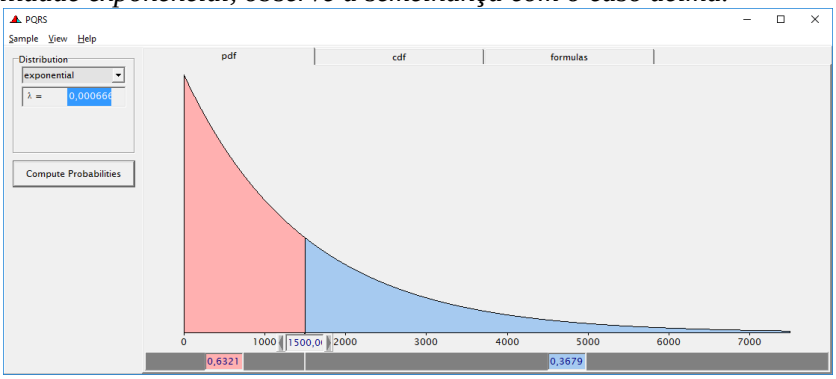


Observe como os pontos praticamente estão sobre a reta esperada, se os dados seguissem a distribuição uniforme entre 1 e 60.

c) Os tempos de operação (tempos para a falha) da linha de transmissão do item a foram monitorados também, resultando no histograma a seguir:

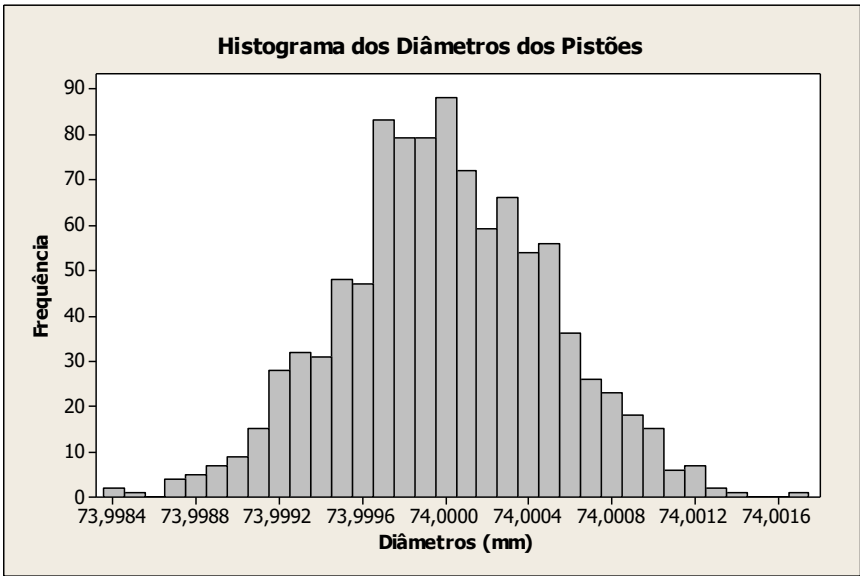


Pelo histograma percebe-se que é uma distribuição assimétrica à direita, com uma concentração maior de frequências nos valores mais baixos dos tempos. O problema é que há vários modelos que têm este mesmo comportamento, mas o modelo exponencial parece ser uma boa opção, veja abaixo um gráfico da função densidade de probabilidade exponencial, observe a semelhança com o caso acima.



Apenas para fins ilustrativos: a decisão acima pode ser considerada um pouco subjetiva, a melhor solução seria construir um gráfico de probabilidade exponencial para uma conclusão mais apurada.

d) Os pistões de motores de um tipo de motocicleta tiveram seus diâmetros medidos, e o resultado pode ser visto no histograma a seguir:



Existe uma maior concentração de medidas em torno de determinados valores centrais (no caso acima, em torno de 74 mm), e à medida que os valores se afastam do centro as frequências vão diminuindo, de forma equilibrada, tanto para valores maiores quanto menores do que 74 mm. Pode-se então concluir que a distribuição é aproximadamente simétrica, e o formato do histograma lembra a "curva de sino" da distribuição normal. Portanto, pode-se concluir que a variável diâmetro dos pistões tem distribuição normal.

Novamente, um gráfico de probabilidade normal permitiria uma decisão mais objetiva.

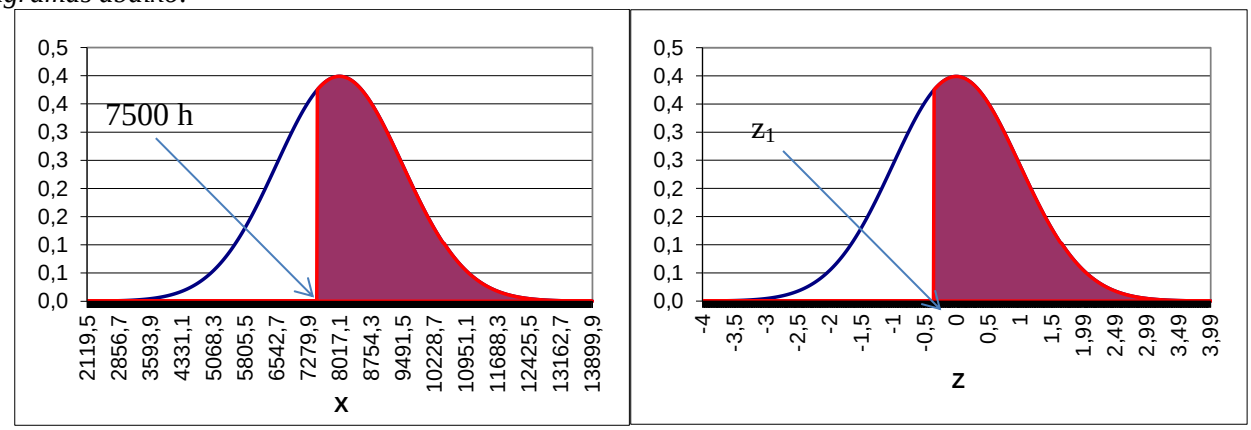
e) Em um sistema de transmissão de dados existe uma probabilidade igual a 0,02 de um dado ser transmitido erroneamente. Ao se realizar um teste para analisar a confiabilidade do sistema foram transmitidos 10 dados. Deseja-se calcular a probabilidade de haver erro na transmissão.

Trata-se de uma variável aleatória discreta, número de dados com erro em 10 transmitidos. Há limite superior para as suas realizações, 10, e menciona-se que há uma probabilidade de erro na transmissão de um dado. Cada transmissão pode ter apenas dois resultados possíveis: erro e não erro. Como não há nenhuma informação a respeito, admite-se que as transmissões são independentes, com a probabilidade de erro p (igual a 0,02) sendo considerada constante. Conclui-se então que o modelo binomial é o mais apropriado para este caso: procura-se P(X > 0).

5) Para as linhas de transmissão de 345 kV da questão 1, imagine que os seus tempos de operação sigam a distribuição normal com a média e desvio padrão mostrados na tabela lá apresentada. Considerando uma linha qualquer escolhida aleatoriamente.

a) Faça o diagrama da distribuição normal mostrando a probabilidade de que o tempo de operação seja maior do que 7500 horas.

Basicamente é preciso construir um diagrama (pode ser aproximado, à mão) mostrando a área procurada (maior do que 7500 horas). Na questão 1, os tempos de operação das linhas de 345 kV foram considerados aproximadamente normais, com média igual a 8017,09 horas e desvio padrão de 1474,39 horas. Veja os diagramas abaixo:

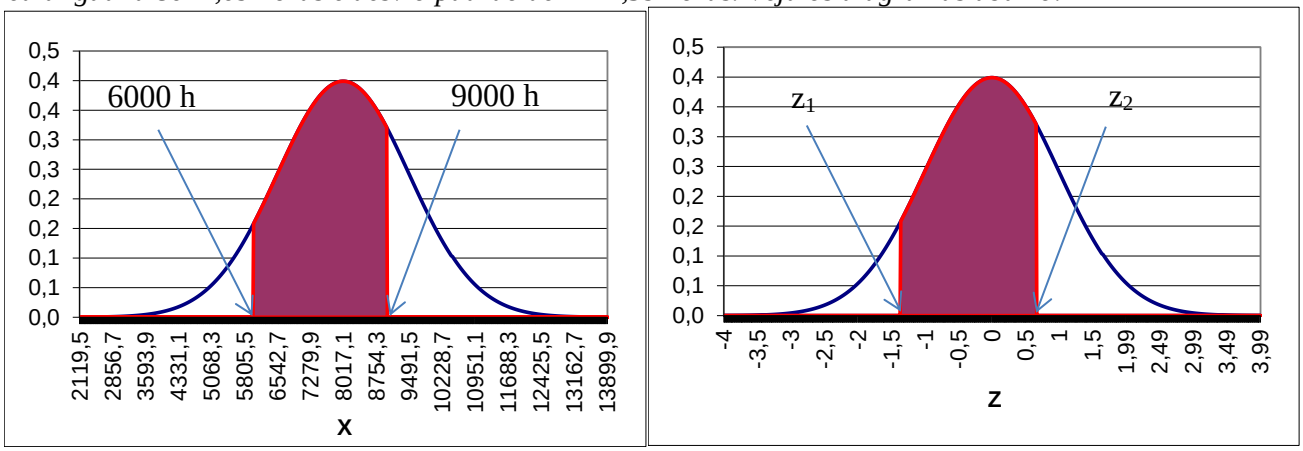


O valor 7500 h é menor do que a média, por isso fica à esquerda de 8017,09, e o z_1, correspondente a 7500 será NEGATIVO, pois 7500 é menor do que a média. Então, procura-se P(X > 7500h) = P(Z > z_1).

Apenas para fins ilustrativos: $z_1 = (7500 – 8017,09) / 1474,39 = -0,35$.
Usando uma tabela de distribuição normal padrão (como a mostrada em aula):
$P(Z > z_1) = 1 – P(Z > -z_1) = 1 – 0,3632 = 0,6338$ (63,38%, repare que é maior do que 50%, coerente com os diagramas). Podemos trocar o sinal de z_1 devido à simetria da normal padrão, com média zero.
No Excel, usando as funções mais antigas:
= 1-DISTNORM(7500;8017,09;1474,39;VERDADEIRO) = 0,637098739. É preciso fazer este procedimento porque a função DISTNORM (com a opção VERDADEIRO) calcula a probabilidade acumulada ATÉ 7500 h, e há interesse em calcular a probabilidade dos tempos estarem ACIMA de 7500 h.

b) Faça o diagrama da distribuição normal mostrando a probabilidade de que o tempo de operação esteja entre 6000 e 9000 horas.

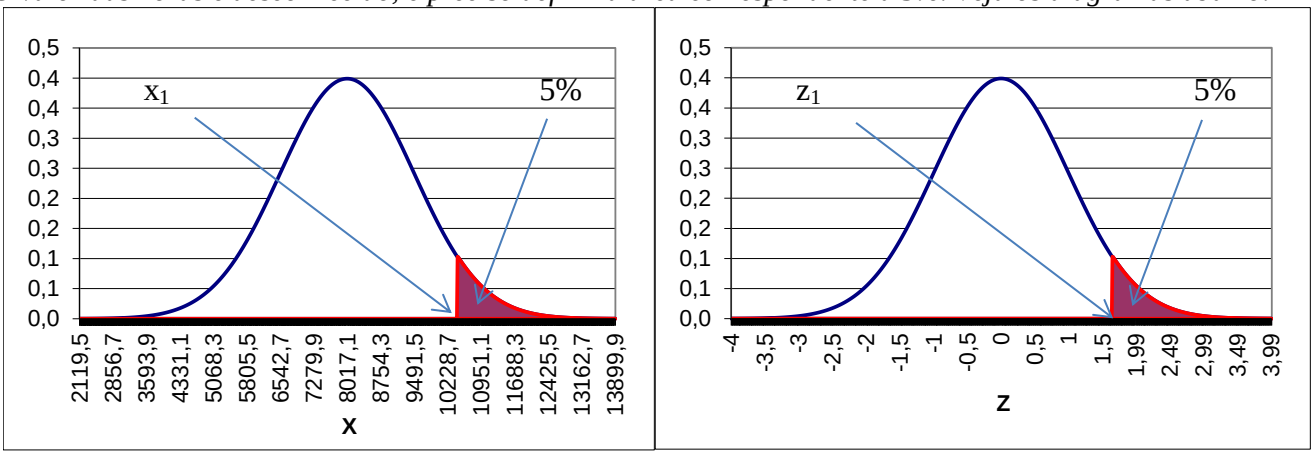
Basicamente é preciso construir um diagrama mostrando a área procurada (entre 6000 e 9000 horas). Na questão 1, os tempos de operação das linhas de 345 kV foram considerados aproximadamente normais, com média igual a 8017,09 horas e desvio padrão de 1474,39 horas. Veja os diagramas abaixo:



O valor 6000 h é menor do que a média, por isso fica à esquerda de 8017,09, e o z_1, correspondente a 6000 será NEGATIVO, pois 6000 é menor do que a média. O valor 9000 h é maior do que a média, por isso fica à direita de 8017,09, e o z_2, correspondente a 9000 será NEGATIVO, pois 9000 é menor do que a média. Então, procura-se $P(6000 h < X < 9000 h) = P(z_1 < Z < z_2)$.
Para fins ilustrativos: $z_1 = (6000 – 8017,09) / 1474,39 = -1,37$ $z_2 = (9000 – 8017,09) / 1474,39 = 0,67$
Usando uma tabela de distribuição normal padrão (como a mostrada em aula):
$P(z_1 < Z < z_2) = 1- P(Z > z_2) – P(Z > -z_1) = 1 – 0,2514 – 0,0853 = 0,6633$ (valor maior do que 50%, e coerente com os diagramas, trocamos o sinal de z1 por que a tabela tem apenas valores de z positivos, e isso pode ser feito por ser a normal padrão simétrica, e com média zero).
No Excel, usando as funções mais antigas:
= DISTNORM(9000;8017,09;1474,39;VERDADEIRO) - DISTNORM(6000;8017,09;1474,39;VERDADEIRO) = 0,661861026. É preciso fazer este procedimento porque se deseja calcular a probabilidade dos tempos assumirem valores entre 6000 e 9000 h, então, calcula-se a probabilidade acumulada até 9000 h e depois se subtrai a probabilidade acumulada até 6000 h.

c) Faça o diagrama da distribuição normal mostrando o tempo de operação que é ultrapassado em apenas 5% dos casos.

O valor das horas é desconhecido, é preciso definir a área correspondente a 5%. Veja os diagramas abaixo:



Então $P(X > x_1) = 0,05 = P(Z > z_1) = 0,05$
Para fins ilustrativos, buscando em uma tabela da distribuição normal:
$P(Z > 1,64) = 0,0505$ e $P(Z > 1,65) = 0,0495$, então z_1 está ENTRE 1,64 e 1,65, pois a probabilidade associada ($P(Z>z_1)$) é igual a 0,05. Por interpolação linear, $z_1 = 1,645$. Usando a equação de z para isolar o x:
$x_1 = 8017,09 + z_1 \times 1474,39 = 8017,09 + 1,645 \times 1474,39 = 10442,46$ horas. Observe que está coerente, um valor consideravelmente acima da média.
No Excel, usando as funções mais antigas:
= INV.NORM(0,95;8017,09;1474,34) = 10442,1635, esta função recupera o valor do tempo associado à probabilidade acumulada 0,95 (sobram 0,05 acima do valor).