Enap Escola Nacional de Administração Pública

Programa Avaliação Socioeconômica de Projetos

Estatística

Brasília - 2015

Conteúdo para impressão

Fundação Escola Nacional de Administração Pública

Presidente
Gleisson Rubin

Diretor de Desenvolvimento Gerencial
Paulo Marques

Coordenadora-Geral de Educação a Distância
Natália Teles da Mota Teixeira

Conteudista

Alexandre da Costa Pereira (2013)

Diagramação realizada no âmbito do acordo de Cooperação Técnica FUB/CDT/Laboratório Latitude e Enap.

SUMÁRIO

Unidade 1 - ESTATÍSTICA DESCRITIVA

Objetivos da Unidade

Esta unidade tem por objetivo fazer com que você tenha condições de descrever e apresentar os resultados de um conjunto de observações de forma clara, objetiva e passando o máximo de informações possíveis.

Para tal objetivo, serão abordadas questões relacionadas a dados estatísticos, distribuições de frequências, representações gráficas, medidas de posição e dispersão.

Após concluir esta unidade, espera-se que você seja capaz de:

- Reconhecer a importância dos métodos estatísticos para o estudo de variáveis.
- Compreender os conceitos fundamentais da estatística descritiva.
- Analisar distribuições de freqüências para dados estatísticos e suas formas de representação.
- Distinguir e saber aplicar as diversas medidas de posição. Calcular a média, mediana e moda para uma amostra.
- Distinguir e saber aplicar as diversas medidas de dispersão.
- Calcular a variância, o desvio-padrão e coeficiente de variação para uma amostra.

1.1. DADOS ESTATÍSTICOS

A Estatística pode ser definida como sendo “parte da matemática aplicada que fornece métodos para a coleta, organização, descrição, análise e interpretação de dados e para a utilização dos mesmos na tomada de decisões”.

A estatística, portanto, é uma ciência que se dedica à coleta, análise e interpretação de dados, preocupando-se com os métodos de coleta, organização, síntese, apresentação e interpretação dos dados, assim como em tirar conclusões sobre as características das fontes donde estes foram retirados para melhor compreender as situações analisadas.

Como podemos apreciar no portal do Instituto Brasileiro de Geografia e Estatística – IBGE (<http://www.ibge.gov.br/brasil_em_sintese/>), são diversos os dados considerados como de interesse para utilização nos processos de planejamento e gestão de projetos e políticas públicas, apresentando-se de forma destacada no site os dados referentes às temáticas relacionadas à população, educação, trabalho, habitação, agropecuária, indústria, comércio, serviços e contas nacionais, bem como as séries estatísticas à disposição no sítio <http://seriesestatisticas.ibge.gov.br/apresentacao.aspx>.

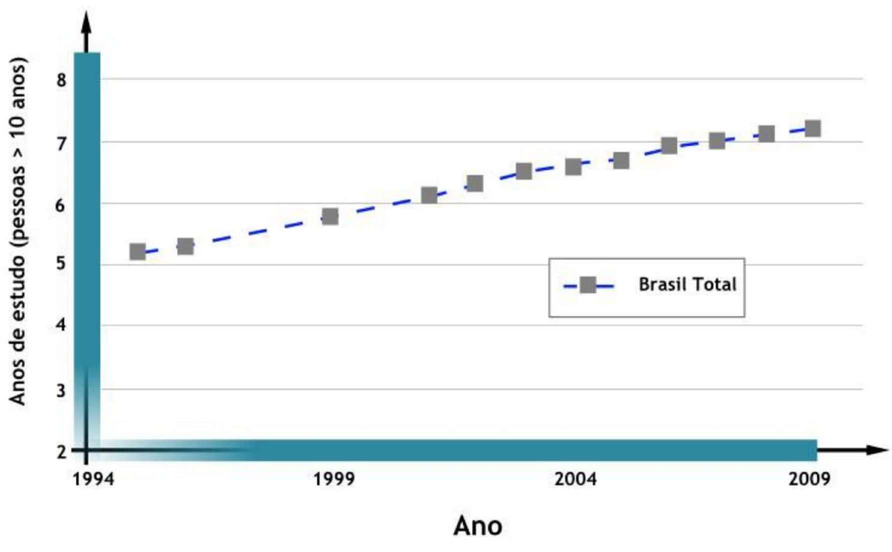
O banco de dados “Séries Estatísticas & Séries Históricas”, à disposição no sítio supracitado, tem por objetivo disseminar, para um público diversificado (instituições governamentais, setor privado, área acadêmica, estudantes, ONGs), informações provenientes de dados oficiais oriundos de pesquisas do IBGE, em sua maior parte, e de outras fontes governamentais.

Ordenadas segundo um intervalo de tempo, essas informações constituem Séries Estatísticas Históricas, sendo, em muitos casos, séries longas (períodos superiores a 20 anos).

Os temas contemplados pelas Séries Estatísticas & Séries Históricas reúnem um acervo de informações sobre a realidade brasileira em suas dimensões social (educação, habitação, trabalho, saúde, organização familiar), demográfica (características da população, dinâmica demográfica e indicadores demográficos) e econômica (sistema de contas nacionais; índices de preços, de produção, de comércio e agropecuária) que, sem excluir os dados de anos recentes, cobrem períodos tão longos quanto possível, importantes enquanto dados estatísticos de interesse em atividades de avaliação socioeconômica de projetos.

Na Figura 01 apresentada abaixo, podemos ver um dos resultados dos levantamentos de séries de dados estatísticos relacionados com a demografia brasileira, referente à série para a evolução dos anos de escolaridade para a parcela da população com idade superior a 10 anos.

Figura 01- Evolução da escolaridade da parcela da população brasileira com idade > 10 anos.



Fonte: <http://seriesestatisticas.ibge.gov.br/series.aspx?no=4&op=0&vcodigo=ECE370&t=media-anos- estudo-pessoas-10-anos>.

Outro exemplo interessante de fonte de dados para séries históricas consiste na compilação de dados realizada por órgãos da Administração Direta, como é o caso do registro feito pelo Ministério da Saúde para o número de óbitos e de internações de vítimas de acidentes do trânsito, cujos dados mais recentes indicam 42.800 óbitos em 2010, 174.000 feridos internados em 2011.

A Figura 02 abaixo mostra a evolução do número de óbitos de 2001 a 2010, com crescimento de 40% no período.

Figura 02- Série de dados de óbitos em acidentes de trânsito no período 2001-2010.

Fonte: <http://www.vias-seguras.com/layout/set/print/os_acidentes/estatisticas/estatisticas_nacionais>.

Para que gráficos como os apresentados nas Figuras 03 e 04 possam ser produzidos, faz-se necessária a mobilização de um esforço importante, com mobilização de recursos humanos, materiais e financeiros de monta para a coleta, tratamento e análise de dados, conforme metodologia adequada com o estudo das distribuições de frequências e representações gráficas, medidas de posição e dispersão para que possam ser adequadamente utilizados para atender, por exemplo, às necessidades do Estado na formulação de políticas públicas, fornecendo dados estatísticos demográficos e econômicos.

Passaremos, então, à apresentação dos fundamentos da estatística descritiva, essenciais para a compreensão do processo estatístico utilizado para a análise de dados e o adequado tratamento dos mesmos.

Como primeiro tópico a ser trabalhado, teremos o estudo das distribuições de frequências para os dados analisados para a adequada compreensão de determinada variável à luz da Estatística.

1.2. DISTRIBUIÇÕES DE FREQUÊNCIAS

Para o estudo do tema relacionado à distribuição de frequências para uma determinada variável, serão apresentados os conceitos referentes à forma de organização dos dados com a adequada determinação do número de classes, sua amplitude, obtenção de suas frequências absoluta e relativa, bem como as formas de representação gráfica e de análise para as variáveis de estudo.

Tipos de Variáveis e a Determinação do Número de Classes (K)

É importante que a distribuição conte com um número adequado de classes. Se o número de classes for excessivamente pequeno, acarretará perda de detalhe e pouca informação se poderá extrair da tabela. Por outro lado, se for utilizado um número excessivo de classes, haverá alguma classe com frequência nula ou muito pequena, não atingindo o objetivo de classificação que é tornar o conjunto de dados supervisionáveis.

Qualquer conjunto de dados, tais como a precipitação pluviométrica em um determinado mês, o PIB de municípios em um Estado, a ocorrência de acidentes em uma rodovia etc., contém informação sobre algum grupo de indivíduos. As possíveis diferenças entre indivíduos determinam a variação que está sempre presente na análise de dados.

Uma característica que pode assumir diferentes valores de indivíduo para indivíduo é denominada variável, pois de outra forma, seria denominada constante. As variáveis podem ser classificadas em qualitativas e quantitativas. Os dados qualitativos são divididos em nominais e ordinais, enquanto os dados quantitativos são divididos em discretas e contínuas.

Portanto, ao conjunto de distintos valores numéricos que adota um caráter quantitativo, denomina-se “variável estatística”, que pode apresentar-se conforme dois tipos:

Variáveis qualitativas ou categóricas: que não se podem medir numericamente (ex: nacionalidade, sexo, escolaridade etc).
Variáveis quantitativas: Possuem valor numérico (ex: idade, preço de um produto, renda etc).

As variáveis também podem se classificar em:

- Variáveis unidimensionais: somente transmitem informação sobre uma característica (ex: idade de alunos de uma turma).
- Variáveis bidimensionais: guardam informação sobre duas características da população (ex: idade e renda de trabalhadores de uma cidade).
- Variáveis pluridimensionais: detém informação sobre três ou mais características (ex: idade, altura e peso de alunos de uma turma).

Por sua parte, as variáveis quantitativas podem classificar-se em discretas e contínuas:

- Discretas: só podem tomar valores inteiros (ex: 1, 2, 8, -4, etc.), como é o caso do número de espécies de orquídeas em um determinado bioma.
- Contínuas: Podem tomar qualquer valor real dentro de um intervalo. Por exemplo, a velocidade de um veículo em uma rodovia, que pode ser 70,4 km/h, 94,57 km/h...

Quando se estuda o comportamento de uma variável, é necessário que se proceda à distinção dos seguintes conceitos:

- Indivíduo: qualquer elemento que porte informação sobre o fenômeno que se estuda. Assim, se estudamos à altura dos alunos em uma turma, cada aluno é um indivíduo; se o objeto de estudo é o preço de uma habitação, cada unidade de habitação é um indivíduo.
- População: conjunto de todos os indivíduos (pessoas, objetos, animais etc) que portem informação sobre o fenômeno que se estuda. Por exemplo, se estudamos o preço de habitações em uma cidade, a população será o total de habitações em dita cidade.
- Amostra: subconjunto que, selecionando uma população, por exemplo no caso do estudo do preço de habitações em uma cidade, o normal seria não obter informações sobre todas as moradias da cidade, pois seria um trabalho muito complexo, mas deveria selecionar um subgrupo (amostra) que se entenda suficientemente representativo.

Convém destacar que quando coletamos dados para uma pesquisa, estas observações são chamadas de dados brutos. Um exemplo de dados brutos corresponde à precipitação pluviométrica mensal, medida em mm, consistindo em dados obtidos conforme metodologia utilizada na hidrologia, na forma que foram coletados em uma determinada estação climatológica, sendo, por este motivo, denominados de dados brutos[1].

Geralmente, este tipo de dado não possui as informações que são necessárias pelo observador para a utilização direta em um projeto, ou conclusões que o auxiliem na tomada de decisão, que somente poderá ser viabilizada após a adequada organização dos dados, visando potencializar a sua capacidade de fornecer informações úteis ao analista e minimizar o erro na tomada de decisão na avaliação de um dado problema a ser enfrentado.

A simples organização dos dados em um rol[2] aumenta muito a capacidade de informação destes. Será possibilitada a verificação da amplitude total[3] de variação para os dados observados e também será possível a organização de rol crescente dos dados, com a identificação dos dados mais frequentes na amostra.

Para a organização de um conjunto de dados, recomenda-se a elaboração de tabela de distribuição de frequências, onde são apresentadas as frequências de cada uma das classes[4], contando o número de observações contidas em cada uma delas, obtendo-se a frequência de classe. Denomina-se distribuição de frequências a disposição tabular dos dados agrupados em classes, associados às suas frequências correspondentes.

Para o caso, por exemplo, da precipitação pluviométrica mensal em uma determinada região, poder-se-ia incluir, em uma única classe, os meses em que a precipitação pluviométrica, para um certo período estivesse compreendida no intervalo de 100 e 200mm.

Considerando os dados de chuva referentes ao registro de observações em estação climatológica, tomado como exemplo, será possível identificar conceitos presentes em uma distribuição de frequências. Para a elaboração da distribuição de frequências, é necessário que, primeiramente, se determine o número de classes (k) em que os dados serão agrupados.

Por questões de ordem prática e estética, sugere-se utilizar de 5 a 20 classes, e uma recomendação de ordem prática que poderá ser seguida é a definição do número de classes (k) em função do número de observações (n), segundo a expressão $k = \sqrt{n}$.

Quadro 01: Precipitação pluviométrica média mensal para um período de 4 anos.

Mês/Ano	P (mm)	Mês/Ano	P (mm)	Mês/Ano	P (mm)	Mês/Ano	P (mm)
jan/08	70,9	jan/09	103,3	jan/10	25,5	jan/11	139,8
fev/08	0,9	fev/09	154,2	fev/10	21,7	fev/11	75,9
mar/08	154,4	mar/09	131,8	mar/10	15,3	mar/11	89,5
abr/08	219,6	abr/09	145,6	abr/10	94,3	abr/11	169,1
mai/08	111,3	mai/09	160,6	mai/10	69,7	mai/11	168,9
jun/08	294,5	jun/09	190,6	jun/10	62,7	jun/11	194,1
jul/08	146,1	jul/09	139,5	jul/10	22,8	jul/11	122,8
ago/08	204,8	ago/09	146,7	ago/10	9,9	ago/11	33,8
set/08	25,0	set/09	26,5	set/10	0,0	set/11	2,3
out/08	23,6	set/09	0,2	set/10	0,3	set/11	5,3

1. Dados na forma em que foram coletados, sem nenhum tratamento
2. É a mais simples organização numérica. É a ordenação dos dados em ordem crescente ou decrescente
3. Corresponde à diferença entre o maior e o menor valor observado em um conjunto de dados. Notaremos por A
4. Intervalos nos quais os valores da variável analisada são agrupados.

nov/08	3,5	nov/09	2,0	nov/10	34,7	nov/11	2,7
dez/08	1,8	dez/09	2,6	dez/10	2,6	dez/11	2,6
Obs. P correspondendo a precipitações totais mensais, em mm.							

Para n = 48 observações, teremos, então, o número de classes definido por $k = \sqrt{48} = 6{,}9$, o que implicaria na definição de 7 classes.

Convém destacar que a definição do número de classes, poderá, inclusive, ser realizada conforme critérios fixados pelo próprio analista, respeitando-se o princípio de evitar classes excessivamente ou fracamente populosas.

Em seguida à determinação do número de classes (k) para o agrupamento dos dados, se procederá à determinação da amplitude[5] do intervalo de classe (c), sendo necessária[6], antes, a determinação da amplitude total dos dados (A), que corresponde à diferença entre o maior valor observado e o menor valor observado. Para o cálculo da amplitude total dos dados (A) para o exemplo considerado, que corresponde à diferença entre o maior valor observado e o menor valor observado, teríamos:

$$\boxed{A = 294{,}5 - 0 = 294{,}5 \text{ mm}}$$

Para a amplitude do intervalo de classe (c), considerando o valor da amplitude total (A) calculado, teremos:

$$\boxed{C = 294{,}5 / (7 - 1) = 49{,}1 \text{ mm}}$$

O próximo passo, após conhecida a amplitude de classes, será a determinação dos intervalos de classe, mediante a definição dos limites inferior (LI) e superior (LS) das classes, escolhidos de modo que o menor valor observado esteja localizado no ponto médio (PM) da primeira classe, ou seja:

$$\boxed{PM = (LI + LS) / 2}$$

Para o valor do limite inferior para um intervalo, teríamos:

$$\boxed{\text{Limite Inferior} = \text{Menor Valor} - c / 2}$$

Como exemplo de cálculo, o limite Inferior da 1ª Classe será dado por:

$$\boxed{LI = 0 - 49{,}1/2 = -24{,}5 \text{ mm}}$$

5. Intervalos abertos – os limites da classe (inferior e superior) não pertencem a ela. Intervalos fechados – os limites de classe (superior e inferior) pertencem à classe em questão. Intervalos mistos – um dos limites pertence à classe, e o outro, não.
6. Existem outros procedimentos para determinação da amplitude do intervalo de classe que podem ser encontrados na literatura.

Convém destacar que o ponto médio do intervalo para a primeira classe corresponderá ao valor 0mm e, tendo em vista que não existe valor negativo para precipitação pluviométrica (desconsiderando-se a evapotranspiração), teremos o Quadro 02 abaixo com as classes da nossa distribuição, basta que somemos a amplitude do intervalo de classe a cada limite inferior.

Quadro 02: Definição dos limites inferior e superior de cada uma das classes determinadas para o universo de dados analisado.

Classe	LI	LS
1a	-24,5	24,6
2a	24,6	73,7
3a	73,7	122,8
4a	122,8	171,9
5a	171,9	221
6a	221	270,1
7a	270,1	319,2

Poderemos, então, elaborar um quadro de frequências[7] absolutas e relativas, conforme indicado no quadro apresentado a seguir:

Quadro 03: Frequências absoluta (f_a) e relativa (f_r) para cada uma das classes.

Classe	LI	LS	f_a	f_r
1a	-24,5	24,6	18	0,375
2a	24,6	73,7	8	0,167
3a	73,7	122,8	6	0,125
4a	122,8	171,9	11	0,229
5a	171,9	221	4	0,083
6a	221	270,1	0	0,000
7a	270,1	319,2	1	0,021
Total			48	1,000

Em relação à interpretação das informações contidas no Quadro 03, pode-se observar que os valores para precipitação ocorrida nos 48 meses avaliados estão concentrados na primeira, segunda e quarta classes, decrescendo em direção às classes do fim da tabela.

A apresentação dos dados na forma de distribuição de freqUências facilita bastante o cálculo manual de várias medidas estatísticas de interesse, bem como a sua apresentação gráfica, consistindo em ferramenta à disposição do analista.

Caso o interesse do analista, além da determinação das frequências absolutas e relativas, se dirija à determinação da quantidade de observações que existe acima ou abaixo de um

7. A frequência absoluta (f) corresponde ao número de observações que temos em uma determinada classe ou em um determinado atributo de uma variável qualitativa, e a frequência relativa (fr) corresponde à proporção do número de observações em uma determinada classe em relação ao total de observações que temos. Esta freqência pode ser expressa em termos porcentuais. Para isto, basta multiplicar a frequência relativa obtida por 100.

determinado ponto na distribuição, teríamos o recurso da utilização da frequência acumulada[8]. A frequência acumulada, apresentada no Quadro 04, pode ser obtida da seguinte forma: abaixo do limite superior da primeira classe temos 18 registros que estão presentes nesta classe, como pode ser visto na distribuição de frequências absoluta.

Quando consideramos a segunda classe, a frequência acumulada corresponde ao número de pessoas que temos abaixo do limite superior desta classe, ou seja, os 8 registros da segunda classe mais os 18 da primeira classe totalizam 26 registros mensais de pluviometria abaixo de 122,8mm, correspondendo a 54,2% do total dos registros contidos na amostra. Para as outras classes, o raciocínio é semelhante.

Quadro 04: Frequências acumuladas para cada uma das classes.

Classe	LI	LS	f acum.	f acum.r
1a	-24,5	24,6	0	0,000
2a	24,6	73,7	18	0,375
3a	73,7	122,8	26	0,542
4a	122,8	171,9	32	0,667
5a	171,9	221	43	0,896
6a	221	270,1	47	0,979
7a	270,1	319,2	47	0,979
Total	Total	Total	48	1,000

Um exemplo típico de aplicação das distribuições de frequências acumuladas corresponde à identificação de uma determinada frequência abaixo ou acima de um determinado valor que não corresponde ao limite superior ou inferior de uma classe qualquer. Podemos, então, querer verificar qual a porcentagem de registros mensais de chuva com intensidade superior a 122,8mm. Para isto, basta consultar diretamente a frequência acumulada acima deste valor (100% – 54,2% = 45,8%), pois o valor 122,8mm/mês corresponde a um dos limites de classe apresentados nesta tabela. Para a determinação de frequências acumuladas correspondentes a valores não coincidentes a limites superiores ou inferiores de classes, o procedimento é o mesmo, bastando realizar a interpolação para a classe correspondente para encontrar a frequência acumulada desejada.

Por exemplo, para a determinação da frequência acumulada para registros mensais de pluviometria com intensidade superior a 100mm nos 48 meses que constituem o intervalo de tempo, referente à amostra estudada, teríamos:

73,7 mm --- 62,5%
100,0 mm --- X
122,8 mm --- 45,8%

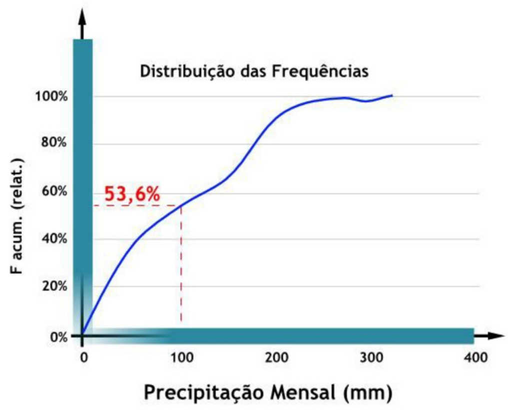
Logo:

$$\boxed{X} = (122{,}8 - 100).(62{,}5 - 45{,}8)/(122{,}8 - 73{,}7) + 45{,}8 = 16{,}7.22{,}8/49{,}1 + 45{,}8 = 53{,}6\%$$

8. Freqüência acumulada é o total acumulado (soma) de todas as classes anteriores até a classe atual.

Ou seja, teríamos a frequência de 46,4% de registros mensais de precipitação pluviométrica superior a 100mm no período de 48 meses em questão. Na Figura 05 apresentada abaixo, teremos a representação gráfica da distribuição de frequências na forma relativa (percentuais) para o caso utilizado no exemplo, sendo destacado o valor correspondente à frequência de registros de precipitação pluviométrica com intensidade superior a 100mm/mês.

Figura 03- Representação gráfica da distribuição de frequências.



Interessante ressaltar que quando se trabalha com variáveis qualitativas, os atributos são as variações nominativas da variável, bem como a construção do quadro, que consiste em contar as ocorrências de cada atributo, e o resultado da contagem definirá a frequência absoluta do atributo em questão. Para exemplificar o caso do trabalho com variáveis qualitativas, teríamos o caso de pesquisa, na qual se procurou avaliar o número de habitações assistidas por esgotamento sanitário em uma determinada população, apresentando-se o Quadro 05 a seguir apresentado:

Quadro 05: Exemplo para frequências absoluta e relativa para variáveis qualitativas.

Esgotamento Sanitário	f_a	f_r
Atendimento	20	0,25
Não Atendimento	60	0,75
Total	80	1,0

Em relação às formas de representação gráfica de tabelas de frequência, é importante destacar que, dependendo do tipo de variável, há um gráfico mais adequado a ser utilizado para a representação. Os diferentes tipos de gráfico (histogramas, polígonos de frequência, ogivas, gráficos de setores, pictogramas e outros) permitem uma melhor visualização de resultados.

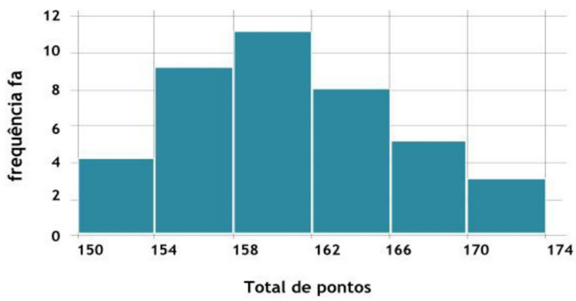
SAIBA MAIS

Estes gráficos podem ser obtidos utilizando planilhas eletrônicas como, por exemplo, o Excel®.

Os histogramas[9] são gráficos utilizados para representar tabelas intervalares. Já o polígono de frequência[10], você pode obter pela simples união dos pontos médios dos topos dos retângulos de um histograma, conforme mostra a Figura 04, que apresenta o histograma e os polígonos de frequência absoluta e acumulada para a distribuição dos dados referentes às classes correspondentes contidas no Quadro 06, apresentado a seguir:

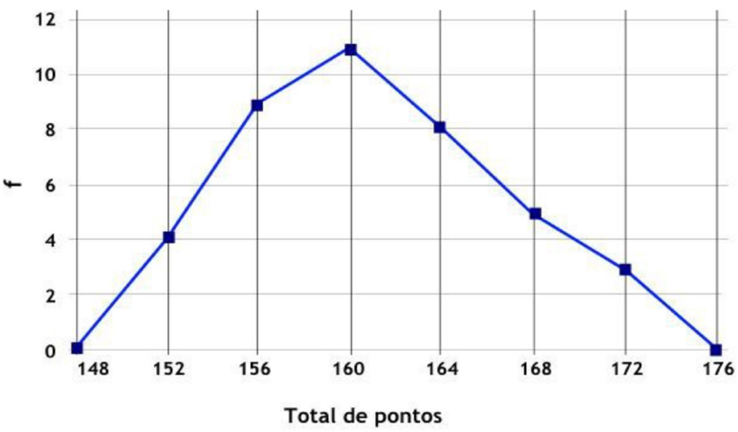
Quadro 06: Dados e frequências absoluta (fa) e acumulada (facum.) para 6 classes.

i	Classes	xi	fai	facumi
1	150 |- 154	152	4	4
2	154 |- 158	156	9	13
3	158 |- 162	160	11	24
4	162 |- 166	164	8	32
5	166 |- 170	168	5	37
6	170 |- 174	172	3	40
Total			40	



a) Histograma

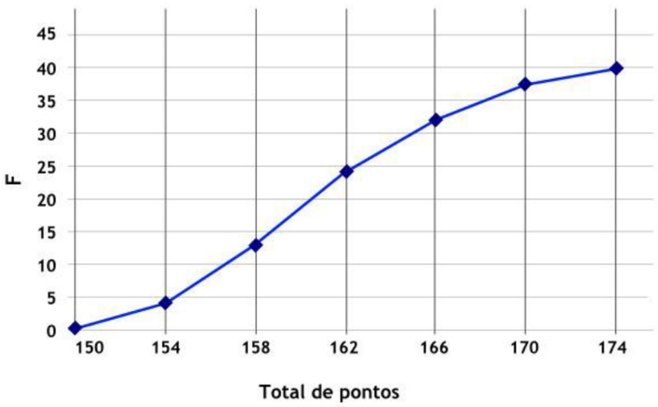
Figura 4a - Exemplos de Histograma



9. Histogramas: são constituídos por um conjunto de retângulos, com as bases assentadas sobre um eixo horizontal, tendo o centro da mesma no ponto médio da classe que representa, e cuja altura é proporcional à frequência da classe.
10. Polígono de frequências é um gráfico de análise no qual as frequências das classes são localizadas sobre perpendiculares levantadas nos pontos médios das classes. Completa-se o polígono unindo as extremidades da linha que ligam os pontos representativos das frequências de classe aos pontos médios das classes imediatamente anterior e posterior às classes extremas, que têm frequência nula.

b) Polígono de frequência absoluta

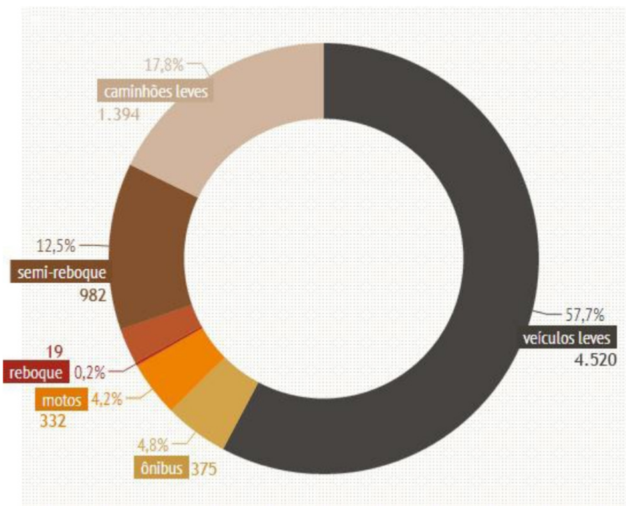
Figura 4b - Exemplos de Polígono de frequência absoluta



Gráficos para a representação de polígono de frequências acumuladas são chamados de ogivas e correspondem a um gráfico onde estas frequências são localizadas sobre perpendiculares, levantadas nos limites inferiores ou superiores das classes.

Para uma tabela de variável qualitativa, há um tipo de gráfico adequado para apresentar os resultados correspondes ao gráfico de setores, também popularmente conhecido como gráfico tipo pizza (Figura 05), com construção simples feita mediante a proporção entre o ângulo central do setor (“fatia da pizza”) e o valor para a variável representada.

Figura 05- Composição média do tráfego para os postos de contagem na BR101-NE[11].



Após o estudo das formas de determinação das distribuições de frequências e gráficos que as representam, você deverá ser capaz de organizar um conjunto de dados por meio de uma distribuição de frequências (absoluta, relativa, e acumuladas), representá-las graficamente e proceder à análise das informações contidas nos mesmos.

11. Pesquisa realizada pelo Exército Brasileiro (2005), mediante parceria com o Departamento Nacional de Infraestrutura de Transportes - DNIT, em 8 (oito) postos de contagem considerados no levantamento estatístico para o projeto de duplicação da BR101-NE.

1.3. MEDIDAS DE POSIÇÃO

As medidas de posição ou de tendência central constituem uma forma mais sintética de apresentar os resultados contidos nos dados observados, pois representam um valor central em torno do qual os dados se concentram. As medidas de tendência central mais empregadas são: média (aritmética, ponderada e geométrica), mediana e moda. Quando se estuda variabilidade, as medidas mais importantes são: amplitude, desvio padrão e variância.

No Quadro 07 mostrado abaixo, são apresentadas as expressões aritméticas para o cálculo das principais medidas de posição:

Quadro 07: Expressões para as principais medidas de tendência central.

Medidas	Expressão
Média aritmética	$\frac{x_1 + \ldots x_n}{n}$
Média aritmética para dados agrupados	$\frac{f_i . X_i \ldots f_1 . x_1}{f_i \ldots f_1}$
Média aritmética ponderada	$\frac{P_i . (X_i) + P_2 . (x_2)}{P_{total}}$
Mediana	1) Se n é impar, o valor é central da série. 2) se n é par, o valor é a média dos dois valores centrais.
Moda	Valor que ocorre com mais freqüência.
Média geométrica	$G = \sqrt[n]{x_i . x}$

A Utilização da Média

Sendo a Distribuição Normal uma das distribuições mais importantes e que surge com mais frequência nas aplicações (o que justifica a grande utilização da média), a média consistirá na melhor medida de localização do centro para uma série de dados. Entretanto, sendo a média uma medida bastante sensível à variabilidade dos dados, é preciso ter cuidado com a sua utilização, tendo em vista que pode propiciar uma imagem distorcida da amostra.

A média possui uma particularidade bastante interessante, que consiste no seguinte: se calcularmos os desvios de todas as observações relativamente à média e somarmos esses desvios, o resultado obtido é igual a zero.

Outra característica da média que torna a sua utilização vantajosa em certas aplicações é quando o que se pretende representar é a quantidade total expressa pelos dados, e então se utiliza a média. Na realidade, ao multiplicar a média pelo número total de elementos, obtemos a quantidade pretendida.

Moda

Define-se moda como sendo o valor que surge com mais frequência se os dados são discretos ou, ainda, o intervalo de classe com maior frequência se os dados são contínuos.

Assim, da representação gráfica dos dados, obtém-se imediatamente o valor que representa a moda ou a classe modal. Esta medida é especialmente útil para reduzir a informação de um conjunto de dados qualitativos, apresentados sob a forma de nomes ou categorias para os quais não se pode calcular a média e por vezes a mediana.

Para dados agrupados com classes, teríamos o seguinte processo para a determinação do valor modal para uma determinada série de dados:

1º. Identificação da classe de maior frequência:

Para o exemplo apresentado no Quadro 06, teríamos a 3ª classe (158 |- 162).

2º passo: Cálculo da Moda:

$$\boxed{Mo} = \boxed{(l_i + l_s) / 2 = (158 + 162) / 2 = 160}$$

Sendo:

l_i: limite inferior da classe modal = 158

l_s: limite superior da classe modal = 162

Mediana

A mediana é uma medida de localização do centro da distribuição dos dados, definida do seguinte modo: Ordenados os elementos da amostra, a mediana é o valor (pertencente ou não à amostra) que a divide ao meio, isto é, 50% dos elementos da amostra são menores ou iguais à mediana e os outros 50% são maiores ou iguais à mediana.

Para sua determinação, utiliza-se a seguinte regra, depois de ordenada a amostra de **n** elementos: Se **n** é ímpar, a mediana é o elemento médio e, se **n** é par, a mediana é a semi-soma dos dois elementos médios.

Teríamos, então, para dados não agrupados, o seguinte processo:

a) Quando o número de valores observados é ímpar:

Exemplo: Considere o conjunto de dados: $X = (5, 2, 7, 10, 3, 4, 1)$

1º) Coloque os valores em ordem crescente ou decrescente:

$$X = (1, 2, 3, 4, 5, 7, 10)$$

2º) Determine a ordem ou posição (P) da Mediana por P = (n+1)/2; portanto:

$$P = (7+1) / 2 = 4^{a} \text{ posição}$$

O número que se encontra na 4ª posição consiste na mediana, portanto, Md = 4

b) Quando o número de valores observados é par:

Exemplo: Considere o conjunto de dados:

$$X = (4, 3, 9, 8, 7, 2, 10, 6)$$

1º) Coloque os valores em ordem crescente ou decrescente:

$$X = (2, 3, 4, 6, 7, 8, 9, 10)$$

2º) Determine a ordem ou posição (P) para n/2 e para n/2 +1

$$P = 8/2 = 4^{a} \text{ posição} \quad \text{e} \quad P = 8/2 + 1 = 5^{a} \text{ posição}$$

Os números são 6 (4ª posição) e 7 (5ª posição): 2, 3, 4, 6, 7, 8, 9, 10

Tira-se a média aritmética entre os dois números:

$$Md = (6+7) / 2 = 6,5$$

Considerações a respeito de Média e Mediana

Como medida de localização, a mediana é mais robusta do que a média, pois não é tão sensível ao valor dos dados que compõem a amostra. Dentre as observações a respeito da comparação entre média e mediana, poderíamos destacar as seguintes:

Quando a distribuição é simétrica, a média e a mediana coincidem.

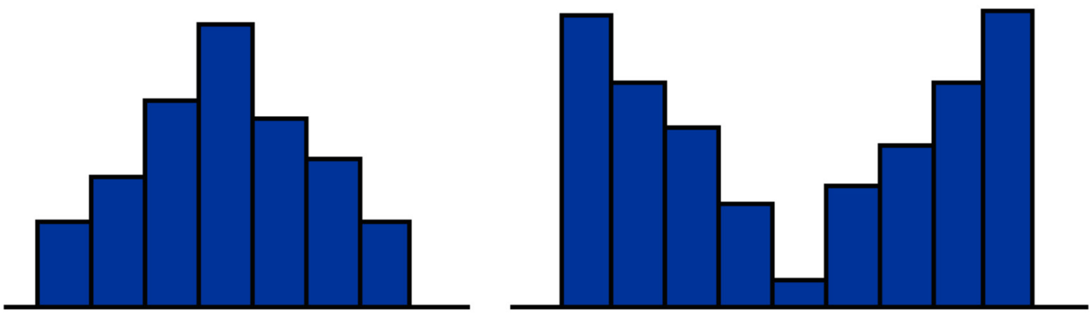
A mediana não é tão sensível, como a média, às observações que são muito maiores ou muito menores do que as restantes (*outliers*). Por outro lado, a média reflete o valor de todas as observações.

Assim, a média, ao contrário da mediana, é uma medida muito influenciada por valores "muito grandes" ou "muito pequenos", mesmo que estes valores surjam em pequeno número na amostra. Estes valores são os responsáveis pela má utilização da média em muitas situações em que teria mais significado utilizar a mediana.

Portanto, teríamos as seguintes considerações sobre a influência da forma da distribuição dos dados:

Figura 05- Exemplos de distribuições simétricas.

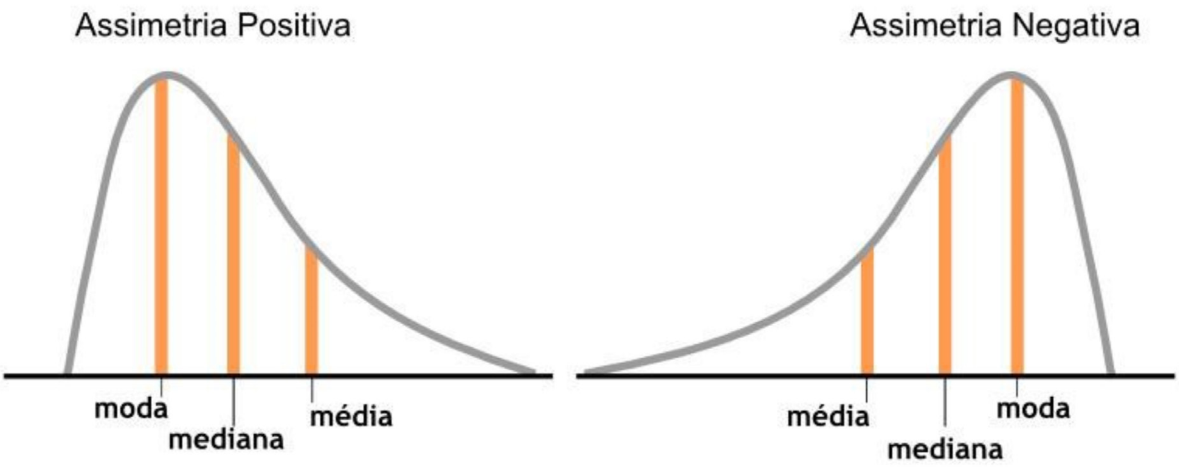
- Quando for aproximadamente simétrica, a média aproxima-se da mediana.



Quando se apresentar de forma enviesada para a direita (alguns valores grandes como *outliers*), a média tende a ser maior que a mediana.

Figura 06- Exemplos de distribuições assimétricas.

Caso a distribuição seja enviesada para a esquerda (alguns valores pequenos como *outliers*), a média tende a ser inferior à mediana.



1.4. MEDIDAS DE DISPERSÃO

No item anterior foram apresentados conceitos de algumas medidas de localização do centro de uma distribuição de dados. Veremos agora como medir a variabilidade presente num conjunto de dados. As medidas de dispersão são utilizadas para medir o grau de variabilidade (dispersão) dos valores observados em torno da média aritmética. Servem para medir a representatividade da média e proporcionam conhecer o nível de homogeneidade ou heterogeneidade dentro de cada grupo analisado.

Assim, um aspecto importante no estudo descritivo de um conjunto de dados é o da determinação da variabilidade ou dispersão desses dados, relativamente à medida de localização do centro da amostra. Supondo ser a média a medida de localização mais importante, será relativamente a ela que se define a principal medida de dispersão – a variância, apresentada a seguir.

Variância

Define-se a variância como sendo a medida que se obtém somando os quadrados dos desvios das observações da amostra, relativamente à sua média, e dividindo pelo número de observações da amostra menos um.

$$\boxed{\sum \frac{(x1 - \text{média})^2}{n - 1}}$$

Desvio-padrão

É a raiz quadrada da variância. Na fórmula original para o cálculo da variância, observa-se que é uma soma de quadrados, a unidade em que se exprime não é a mesma que a dos dados Por exemplo, se a unidade original for metro (m), o resultado será metro ao quadrado (m2).

Para retornar à unidade de medida original, extrai-se a raiz quadrada da variância, passando a chamar-se de desvio-padrão. Assim, para obter uma medida da variabilidade ou dispersão com as mesmas unidades que os dados, tomamos a raiz quadrada da variância e obtemos o desvio padrão.

$$\sqrt{\Sigma\frac{(x1 - \text{média})^2}{n - 1}}$$

O desvio padrão, portanto, é uma medida que só pode assumir valores não negativos e, quanto maior for, maior será a dispersão dos dados. O desvio padrão será maior, quanto mais variabilidade houver entre os dados.

Coeficiente de Variação

O coeficiente de variação (CV) consiste em uma medida relativa de dispersão, útil para a comparação em termos relativos ao grau de concentração em torno da média de séries distintas. Para uma amostra, teríamos a seguinte expressão:

$$CV = \text{Desvio Padrão / Média x 100 (\%)}$$

IMPORTANTE

O coeficiente de variação é expresso em porcentagem, avaliado para amostras segundo a seguinte referência:

Baixa dispersão: CV $\leq$ 15%

Média dispersão: 15% < CV < 30%

Grande dispersão: CV $\geq$ 30%

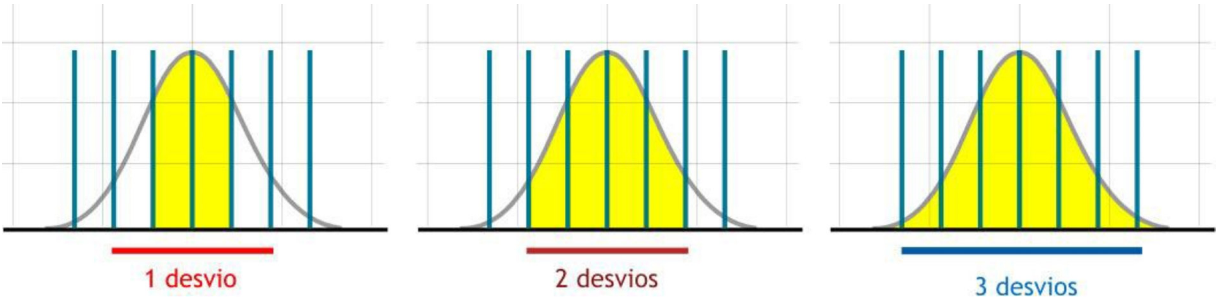
Distribuição Normal

A distribuição normal é a mais importante distribuição estatística, considerando a questão prática e teórica, apresentando-se em formato de sino, unimodal, simétrica em relação a sua média.

Considerando a probabilidade de ocorrência, a área sob sua curva soma 100%. Isso quer dizer que a probabilidade de uma observação assumir um valor entre dois pontos quaisquer é igual à área compreendida entre esses dois pontos.

Na figura apresentada a seguir, com as barras situadas logo abaixo do eixo das abscissas, representando os desvios-padrão, quanto mais afastado do centro da curva normal, mais área compreendida abaixo da curva haverá, ou seja, a um desvio-padrão, temos 68,26% das observações contidas, a dois desvios-padrões, possuímos 95,44% dos dados compreendidos e, finalmente, a três desvios, temos 99,73% de probabilidade de ocorrência.

Figura 07- Relação entre o desvio-padrão e a probabilidade de ocorrência de um evento.



p(x) para n desvios-padrão 68,26% => 1 desvio

95,44% => 2 desvios

99,73% => 3 desvios

O desvio-padrão, quando analisado isoladamente, não dá margem a muitas conclusões. Por exemplo, para uma distribuição cuja média é 79,7, como é o caso do exemplo com a série de dados de precipitação pluviométrica mensal, visto no início do curso, que apresentou desvio padrão de 76,9, considerado como bastante elevado, um desvio-padrão de 5mm/mês seria pequeno, mas para uma distribuição cuja média fosse 10, este desvio-padrão já não seria tão pequeno.

IMPORTANTE

Condições para se usar o desvio-padrão ou variância para comparar a variabilidade entre grupos:

- mesmo número de observações;
- mesma unidade;
- mesma média.

Além disso, se quisermos comparar duas ou mais amostras de valores expressos em unidades diferentes, não será ser possível fazer a comparação por meio do desvio-padrão, pois ele é expresso na mesma unidade dos dados.

UNIDADE 2 – PROBABILIDADES E TÉCNICAS DE AMOSTRAGEM

OBJETIVOS DA UNIDADE

Após concluir esta unidade, espera-se que você seja capaz de:

- Reconhecer a importância do estudo das probabilidades e das técnicas de amostragem na avaliação de projetos.
- Conhecer e saber aplicar os conceitos fundamentais de probabilidade e amostragem.
- Calcular a probabilidade de ocorrência de um evento.
- Associar o estudo da probabilidade à atividade de análise de risco em projetos.
- Distinguir técnicas de amostragem e identificar vantagens de determinada técnica em relação às demais.

2.1. INTRODUÇÃO AO ESTUDO DAS PROBABILIDADES

Estudaremos agora a probabilidade, que é uma ferramenta usada e necessária para se fazer correlações entre a amostra e a população, de modo que a partir de informações da amostra se possam fazer afirmações sobre características da população, entendendo-se a probabilidade como ferramenta básica da inferência estatística.

Interessante observar que utilizamos, no nosso cotidiano, o conceito de probabilidade em situações como:

- É pouco provável que amanhã chova.   O risco do projeto é elevado.
- Provavelmente fulano não conseguirá se eleger.
- Diminuiu a chance do paciente se recuperar.

São fenômenos ou eventos com resultados não completamente conhecidos a priori. Mesmo que a chance da ocorrência seja alta, os resultados não são conhecidos antes de ocorrer mas, de certa forma, mantém certa regularidade, o que permite determinar a chance de ocorrência: a Probabilidade.

A probabilidade é um número entre zero e um, inclusive, o que significa que, no mínimo, não há nenhuma hipótese do evento acontecer e, no máximo, o evento sempre ocorrerá:

$$\boxed{0 \leq P(E) \leq 1}$$

Normalmente representamos probabilidades através de frações, mas também podemos representá-las por números decimais ou por porcentagens.

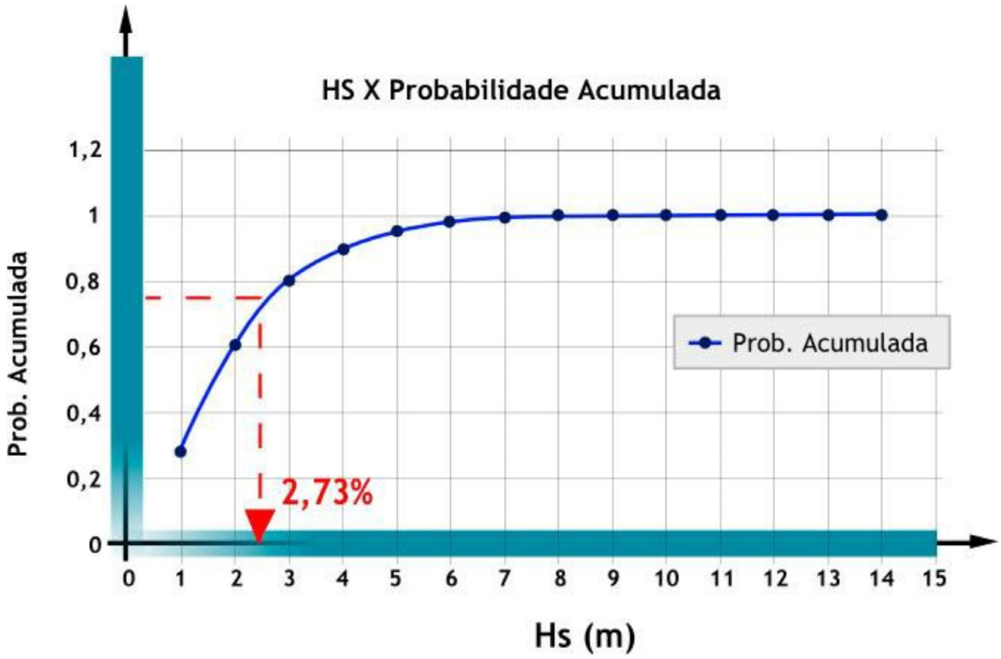
Por exemplo, em um projeto de embarcações, a estimativa da altura da onda de projeto é feita, geralmente, considerando-se 0,75 a probabilidade do navio encontrar ondas maiores que sua onda de projeto. A probabilidade acumulada é a de ocorrência de todas as alturas menores que a determinada.

Neste exemplo, mediante interpolação linear, pode-se obter a altura de onda correspondente à probabilidade de 75%. Portanto, para a probabilidade acumulada da ocorrência correspondente à altura de onda de projeto, ter-se-ia o valor da altura de onda a ser utilizado no dimensionamento da estrutura naval em questão, como mostra a tabela a seguir:

Tabela 1 – Dimensionamento da estrutura naval

Altura de Onda			
Hs(m)	Total	Probabilidade Relativa	Probabilidade Acumulada
1,0	26287	0,263	0,263
2,0	34001	0,340	0,603
3,0	20092	0,201	0,804
4,0	10482	0,105	0,909
5,0	5073	0,051	0,959
6,0	2323	0,023	0,983
7,0	1018	0,010	0,993
8,0	432	0,004	0,997
9,0	178	0,002	0,999
10,0	70	0,001	1,000
11,0	28	0,000	1,000
12,0	11	0,000	1,000
13,0	4	0,000	1,000
14,0	1	0,000	1,000
Total	100000		
Probabilidade de Ocorrência			
% Ocorrência		Altura de Onda(m)	
75,0%	**0,750**	**2,73**	(valor interpolado)
60,3%	0,603	2,00	
80,4%	0,804	3,00	

Figura 08- Procedimento para a determinação da altura de onda de projeto14.



Fonte: Exercício contido no site <www.oceanica.ufrj.br > [consulta em 10/04/2012].

2.2. CONCEITOS FUNDAMENTAIS: EXPERIMENTOS, ESPAÇO AMOSTRAL E TIPOS DE EVENTOS[12]

O estudo da probabilidade resulta na necessidade de, em certas situações, prevermos a possibilidade de ocorrência de determinados fatos. Ao começarmos o estudo da probabilidade, normalmente a primeira ideia que nos vem à mente é a da sua utilização em jogos, mas podemos utilizá-lo em muitas outras áreas. Um bom exemplo é na área da avaliação do risco de projetos, bem como na utilização de métodos de análise de sensibilidade em processos de tomada de decisão na seleção de projetos. Para iniciarmos o estudo da probabilidade, vamos a seguir definir alguns conceitos importantes sobre a matéria, fazendo uso de exemplos clássicos para a compreensão dos fundamentos do cálculo de probabilidades.

Experimento Aleatório

Se lançarmos uma moeda ao chão para observarmos a face que ficou para cima, o resultado é imprevisível, pois tanto pode dar cara quanto coroa. Se, ao invés de uma moeda, o objeto a ser lançado for um dado, o resultado será mais imprevisível ainda, pois aumentamos o número de possibilidades de resultado. Experimentos como estes, ocorrendo nas mesmas condições ou em condições semelhantes, que podem apresentar resultados diferentes a cada ocorrência, damos o nome de experimentos aleatórios.

Espaço Amostral

Ao lançarmos uma moeda, não sabemos qual será a face que ficará para cima, no entanto, podemos afirmar com toda certeza que, ou será cara, ou será coroa, pois uma moeda só possui estas duas faces.

12. Conteúdo sobre noções de probabilidade adaptado do material instrucional contido no sítio <http://www.matematicadidatica.com.br/ProbabilidadeConceitos.aspx> [consulta em 13/04/2012].

Neste exemplo, ao conjunto {cara, coroa} damos o nome de espaço amostral, pois ele é o conjunto de todos os resultados possíveis de ocorrer neste experimento. Representamos um espaço amostral, ou espaço amostral universal como também é chamado, pela letra S.

Figura 09- Experimento de lançar moeda, com espaço amostral S = {cara, coroa}[13].

Se, ao invés de uma moeda, o objeto a ser lançado for um dado, o espaço amostral será:

$$S = \{1, 2, 3, 4, 5, 6\}$$

Evento

Quando lançamos um dado ou uma moeda, chamamos a ocorrência deste fato de evento. Qualquer subconjunto de um espaço amostral é um evento. Em relação ao espaço amostral do lançamento de um dado, veja o conjunto a seguir:

$$A = \{2, 3, 5\}$$

Note que **A** está contido em **S**, sendo um subconjunto de **S**. O conjunto **A** é a representação do evento do lançamento de um dado, quando temos a face para cima igual a um número primo.

Classificação de Eventos

Podemos classificar os eventos por vários tipos. Vejamos alguns deles:

Evento Simples: Classificamos assim os eventos que são formados por um único elemento do espaço amostral. Desta forma, **A = {5}** é a representação de um evento simples do lançamento de um dado cuja face para cima é divisível por 5. Nenhuma das outras possibilidades são divisíveis por **5**.

13. Fonte: <http://revistaescola.abril.com.br> [consulta em 13/04/2012]

Evento Certo: Ao lançarmos um dado, é certo que a face que ficará para cima, terá um número divisor de 720. Este é um evento certo, pois **720 = 6! = 6 . 5 . 4 . 3 . 2 . 1**, obviamente qualquer um dos números da face de um dado é um divisor de **720**, pois **720** é o produto de todos eles. O conjunto **A = {2, 3, 5, 6, 4, 1}** representa um evento certo, pois ele possui todos os elementos do espaço amostral **S = {1, 2, 3, 4, 5, 6}**.

Evento Impossível: No lançamento do conjunto de dois dados, qual é a possibilidade de a soma dos números contidos nas duas faces para cima, ser igual a **15**? Este é um evento impossível, pois o valor máximo que podemos obter é igual a doze. Podemos representá-lo por **A = ???**, ou ainda por **A = {?? }.**

Evento União: Seja **A = {1, 3}**, o evento de ocorrência da face superior no lançamento de um dado, ímpar e menor ou igual a **3 e B = { 3, 5 }**, será o evento de ocorrência da face superior, ímpar e maior ou igual a 3, então teremos **C = { 1, 3, 5 }**, que representa o evento de ocorrência da face superior ímpar, que é a união dos conjuntos A e B, ou seja, $C = A \cup B$. Note que o evento **C** contém todos os elementos de **A** e **B**.

Evento Intersecção: Seja **A = {2, 4**}, o evento de ocorrência da face superior no lançamento de um dado, par e menor ou igual a **4** e **B = {4, 6}**, será o evento de ocorrência da face superior, par e maior ou igual a **4**, então teremos **C = {4}**, que representa o evento de ocorrência da face superior par, que é a intersecção dos conjuntos A e B, ou seja,

Veja que o evento **C** contém apenas os elementos comuns a **A** e **B**.

Eventos Mutuamente Exclusivos: Seja **A = {1, 2, 3, 6}** o evento de ocorrência da face superior no lançamento de um dado, um número divisor de **6 e B = {5}**, o evento de ocorrência da face superior, um divisor de 5, os eventos **A** e **B** são mutuamente exclusivos, pois $A \cap B = \emptyset$, isto é, os eventos não possuem elementos em comum.

Evento Complementar: Seja **A = {1, 3, 5}** o evento de ocorrência da face superior no lançamento de um dado, um número ímpar, o seu evento complementar é **A = {2, 4, 6}** o evento de ocorrência da face superior no lançamento de um dado, um número par. Os elementos de **A** são todos os elementos do espaço amostral S que não estão contidos em A, então temos que **A = S - A** e ainda que **S = A + A**.

2.3. PROBABILIDADE DE OCORRÊNCIA DE UM EVENTO

Passaremos a apresentar considerações sobre os fundamentos para o estudo da probabilidade de ocorrência de um determinado evento.

IMPORTANTE

O ponto central em todas as situações onde usamos probabilidade é a possibilidade de quantificar quão provável é determinado EVENTO. As probabilidades são utilizadas para exprimir a chance de ocorrência de determinado evento .

Experimentos aleatórios, espaço, amostral e evento

Encontramos na natureza dois tipos de fenômenos: determinísticos e aleatórios. Os fenômenos determinísticos são aqueles em que os resultados são sempre os mesmos, qualquer que seja o número de ocorrência dos mesmos.

Se considerarmos o fenômeno, por exemplo, da flutuação de objeto imerso na água, sabemos que caso o objeto possui peso específico menor que a água, ele flutuará. Esse exemplo caracteriza um fenômeno determinístico. Nos fenômenos aleatórios, os resultados não serão previsíveis, mesmo que haja várias repetições do mesmo fenômeno.

Podemos considerar, como experimentos aleatórios, os seguintes exemplos de fenômenos produzidos pelo homem:

- lançamento de uma moeda;
- lançamento de um dado;
- determinação da vida útil de um componente eletrônico;
- previsão do tempo.

A cada experimento aleatório está associado o resultado do mesmo, que não é previsível, chamado de evento aleatório.

Um conjunto “S” que consiste de todos os resultados possíveis de um experimento aleatório é denominado espaço amostral.

Definição de Probabilidade

A probabilidade de um evento A, denotada por P(A), é um número de 0 a 1 que indica a chance de ocorrência do evento A. Quanto mais próxima de 1 é P(A), maior é a chance de ocorrência do evento A, e quanto mais próxima de zero, menor é a chance de ocorrência do evento A. A um evento impossível, atribui-se probabilidade zero, enquanto que um evento certo tem probabilidade 1,0.

As probabilidades podem ser expressas de diversas maneiras, inclusive decimais, frações e percentagens. Por exemplo, a chance de ocorrência de um determinado evento pode ser expressa como 20%; 2 em 10; 0,20 ou 1/5.

Assim, sendo E é um evento, n(E) é o seu número de elementos, S é o espaço amostral não vazio e n(S) é a quantidade de elementos do mesmo, e assim temos que a probabilidade de E ocorrer é igual a:

$$\boxed{P(E) = n(E) \,/\, n(S), \text{ para } n(S) \neq 0}$$

Recorreremos ao exemplo clássico do lançamento de um dado para a utilização da equação para o cálculo de probabilidades: Qual é a probabilidade de obtermos um número divisor de 6 no lançamento de um dado?

Como vimos acima, o espaço amostral do lançamento de um dado é:

$$\boxed{S = \{1, 2, 3, 4, 5, 6\}}$$

Como estamos interessados apenas nos resultados divisores de 6, o evento E é representado por:

$$E = \{ 1, 2, 3, 6 \}$$

Então n(E) = 4 e n(S) = 6, portanto:

$$P(E) = \frac{n(E)}{n(S)} \Rightarrow P(E) = \frac{4}{6} \Rightarrow P(E) = \frac{2}{3}$$

Podemos também apresentar o resultado na forma de uma porcentagem, portanto, a probabilidade de se obter um número divisor de 6 é 2/3 ou 66,67%.

Considerando a ótica da avaliação de projetos, admitindo que três projetos, Alfa, Beta e Gama, lançaram um produto no mercado, e supondo que as condições de retorno econômico para os três projetos são as mesmas, qual é a probabilidade da empresa Alfa de obter primeiro a meta do retorno socioeconômico?

Como 3 é o número total de projetos, então o projeto Alfa tem 1 chance em 3 de obter primeiro o indicador econômico considerado, o que resultaria em uma probabilidade de 1/3 em obter primeiro o retorno.

A probabilidade de um evento ocorrer (de o projeto Alfa obter retorno econômico primeiro), considerando-se um espaço amostral (Alfa, Beta e Gama), é igual à razão do número de elementos do evento (1, apenas Alfa) para o número de elementos do espaço amostral (3, o número de projetos), desde que o espaço amostral seja um conjunto equiprovável, ou seja, todos os seus elementos tenham a mesma possibilidade de ocorrer (as condições de êxito econômico são as mesmas para as três empresas). Vamos, agora, exercitar a lógica da probabilidade conforme a classificação dos eventos aplicáveis à situação.

Vimos que a probabilidade da união de dois eventos pode ser calculada através da fórmula:

$$P = (A \cup B) = P(A) + P(B) - P(A \cap B)$$

No caso da intersecção dos eventos, resulta vazia, ou seja, não há elementos em comum aos dois eventos, pois poderíamos simplesmente utilizar:

$$P(A \cup B) = P(A) + P(B)$$

Vamos a um exemplo:

Na construção de uma usina hidrelétrica, trabalham 2.000 profissionais. Destes trabalhadores, 500 possuem mais de dez anos de experiência, 300 são pedreiros e 200 são pedreiros com mais de dez anos de experiência profissional. Selecionando-se um trabalhador com mais de dez anos de experiência, qual a probabilidade de este ser um pedreiro?

Chamemos de A o evento que representa o ofício do pedreiro e B o evento que representa mais de dez anos de experiência para o profissional.

Podemos calcular a probabilidade de ocorrer A, tendo ocorrido B através da fórmula:

$$P\left(\frac{A}{B}\right) = \frac{n(A \cap B)}{n(B)}$$

Segundo o enunciado

$$\boxed{n(A \cap B) = 200} \text{ e } \boxed{n(B) = 500}$$

Portanto:

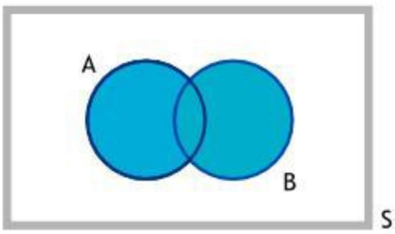
$$\boxed{P(A/B) = 200 \ / \ 500 = 2/5}$$

IMPORTANTE

Note que, no caso da probabilidade condicional, ao invés de calcularmos a probabilidade em função do número de elementos do espaço amostral, a calculamos em função do número de elementos do evento que já ocorreu. Portanto, a probabilidade de o pedreiro também possuir mais de dez anos de experiência profissional é de 40%.

Probabilidade no Caso da União de Eventos

Considere A e B como dois eventos de um espaço amostral S, finito e não vazio. Temos:



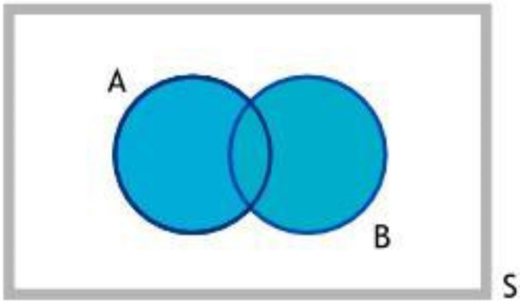
$$n(A \cup B) = n(A) + n(B) - (A \cap B) \Leftrightarrow$$

$$\Rightarrow \frac{n(A \cup B)}{n(S)} = \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)}$$

Logo:

$$\boxed{P(A \cup B) = P(A) + P(B)}$$

Probabilidade no Caso de Eventos Mutuamente Exclusivos



Considerando que $A \cap B = \emptyset$, nesse caso A e B serão denominados mutuamente exclusivos. Observe que $P(A \cap B) = 0$, pois:

Caso da Probabilidade Condicionada

Considere dois eventos A e B de um espaço amostral S, finito e não vazio. A probabilidade de B condicionada a A é dada pela probabilidade de ocorrência de B, sabendo que já ocorreu A. É representada por P(B/A).

Situação da Intersecção de Eventos

Considerando A e B como dois eventos de um espaço amostral S, finito e não vazio, temos:

$$P(B/A) = \frac{n(A \cap B)}{n(A)} = \frac{n(A \cap B) \div n(S)}{n(A) \div n(S)} = \frac{P(A \cap B)}{P(A)}$$

$$P(A/B) = \frac{n(A \cap B)}{n(B)} = \frac{n(A \cap B) \div n(S)}{n(B) \div n(S)} = \frac{P(A \cap B)}{P(B)}$$

Portanto:

$$P(A \cap B) = P(A) \,.\, P(B/A)$$

$$P(A \cap B) = P(B) \,.\, P(A/B)$$

Considerando A e B como eventos independentes, logo P(B/A) = P(B), P(A/B) = P(A), sendo assim:

$$P(A \cap B) = P(A) \,.\, P(B)$$

Para saber se os eventos A e B são independentes, podemos utilizar a definição ou calcular a probabilidade de $A \cap B$.

Representação:

$$A \text{ e } B \text{ independentes} \Leftrightarrow P(A/B) = P(A)$$

ou

$$A \text{ e } B \text{ independentes} \Leftrightarrow P(A \cap B) = P(A) \,.\, P(B)$$

2.4. CONSIDERAÇÕES SOBRE ANÁLISE DE RISCO NA AVALIAÇÃO DE PROJETOS

Evento de Risco

De acordo com o PMBOK[14] (2008), um evento de risco, é uma ocorrência discreta ou distinta que pode afetar o projeto para melhor ou pior. Quanto ao aspecto da probabilidade de ocorrência, a mesma pode ter valores atribuídos de forma subjetiva e de forma objetiva.

Subjetivamente, a probabilidade é uma porcentagem indicando o grau de confiança ou a estimativa pessoal quanto à possibilidade de ocorrência de um evento (probabilidade subjetiva). Como exemplo, temos afirmações do tipo "eu acho que há 50% de chance de perda", ou "eu acredito que há somente uma chance em mil de uma inundação atingir nossa fábrica".

Podemos, por outro lado, entender a probabilidade objetiva como sendo um número real, associado a um evento (E), destinado a medir sua possibilidade de ocorrência.

IMPORTANTE

Todos os projetos têm algum risco, ou não seriam projetos. Os projetos são iniciados quando há algum elemento de risco e a gerência busca, então, o enfoque de um plano de projeto e uma equipe para realizar o trabalho. Às vezes assume-se um grande risco quando não se tem a compreensão necessária dos elementos que podem levar ao fracasso do projeto, podendo até ser interpretado como um pequeno risco, que não pressiona as barreiras de custo, tempo e desempenho técnico.

O risco deve ser analisado por três componentes: pelo evento, a probabilidade de ocorrência e o impacto do evento. Normalmente é considerado como um evento de consequências negativas, quanto ao custo, ao tempo e à qualidade. Por este motivo, os gerentes concentram suas forças em evitá-los e em como lidar com eles. Eles esquecem que, eventualmente, os riscos podem ter consequências positivas.

Riscos Internos

O risco interno é inerente ao projeto, controlado pelo líder que pode reduzi-lo mediante ações diretas, como o desenvolvimento de planos de contingência. Riscos internos constituem uma parte das limitações estabelecidas para o projeto por meio da criação de metas. A data de

14. *Project Management Body of Knowledge*

entrega do produto pode ser otimista, e o plano precisa refletir esta data. O planejamento terá por base a data de entrega e o cronograma de trabalho visará assegurar a entrega. O custo constitui, em geral, uma área com limitações.

O planejamento irá direcionar o tipo e a quantidade de recursos a serem utilizados no projeto, mesmo quando o orçamento for menor do que o desejado. Soluções técnicas estarão em perigo quando o tempo de entrega for otimista, sendo, entretanto, os recursos financeiros, limitados.

Dentre os riscos internos, podemos citar: custo, prazo, gerenciais, perda de potencial e fluxo de caixa.

Como exemplo dos riscos internos, devemos considerar que fluxos de caixa podem sofrer alterações, e que estas devem ser avaliadas constantemente durante o andamento do projeto para que possamos classificá-las como relevantes e se irão adicionar valor à empresa.

Riscos Externos

O risco externo encontra-se fora do controle dos líderes de projeto, como, por exemplo, as interfaces do projeto, desconhecidas e cujas definições são feitas por terceiros. Os riscos externos do projeto podem ser influenciados pelo líder de projeto e, portanto previstos.

No entanto, não há controle direto sobre as ocorrências de riscos, que podem receber a influência de acordos e contratos com terceiros.

O grau de influência exercido pelo líder do projeto é determinado pela identificação de riscos externos e pela colaboração ampliada por terceiros, os quais poderiam ser outros líderes de projetos, gerentes funcionais, distribuidores e entidades contratantes. Dentro os riscos externos previsíveis, podemos citar:

- **Taxa de Câmbio:** é o risco associado a operações internacionais em um mundo no qual o valor relativo das moedas varia. Para um grande projeto, a gestão do risco de câmbio é complicada pelo fato de que pode haver uma grande variedade de diferentes moedas envolvidas em suas inúmeras atividades. É muito provável que uma variação de alguma taxa beneficie algumas áreas e prejudique outras.
- **Inflação:** é o risco macroeconômico da perda de renda disponível de pessoas físicas e jurídicas, acarretando prejuízos ao fluxo de caixa e recusa do projeto.
- **Impactos Sociais:** é o risco relacionado aos possíveis efeitos sociais, econômicos, culturais, sobre pessoas, grupos de pessoas ou comunidades.
- **Impactos Ambientais:** os riscos ambientais impactam diretamente os aumentos dos custos do projeto devido, principalmente, a: possíveis impactos ambientais, com a aplicação de multas, licenças e regulamentações, o uso de novos materiais desenvolvidos para não agredir o meio ambiente, risco de denegrir a imagem da empresa junto ao mercado devido a um possível dano ambiental, limpeza de locais poluídos, transporte e manuseio de materiais e o aumento do preço de recursos não renováveis ou ainda aumento de impostos sobre produtos poluentes.
- **Riscos Operacionais:** referem-se às perdas potenciais resultantes de sistemas inadequados, má administração, controles defeituosos ou falha humana. Também incluem as fraudes e os riscos de modelo, onde há o perigo de imperfeição nos modelos de controle.
- Riscos do Mercado: são aqueles que podem gerar resultados adversos em função de instabilidade em taxas de juros, taxas de câmbio, preços de ações, etc. Para

analisarmos, devemos considerar: a existência de um mercado local ou internacional, as projeções de preço para os produtos, tarifas e barreiras alfandegárias, o acompanhamento de projetos e de produtos concorrentes, o acesso aos mercados em termos logísticos, comunicativos e comerciais, a obsolescência, caso haja uso de tecnologia de grande inovação.

Já para os riscos externos imprevisíveis, podemos citar: medidas reguladoras, efeitos colaterais e desastres da natureza.

Gerenciando os riscos do projeto com a matriz de probabilidade e impacto

> "Risco é um evento ou uma condição incerta, que se ocorrer, tem um efeito em pelo menos um objetivo do projeto" (PMBOK, 2008).

Entende-se por objetivos do projeto as áreas de escopo, cronograma, custo e qualidade, que podem ser afetadas, como já comentado, negativamente ou positivamente, a depender da natureza do risco.

Pode parecer estranho, mas são diversas as formas de um risco afetar um objetivo do projeto de forma positiva. Por exemplo, poderíamos citar o caso de obra em que o tempo de duração foi definido em 24 meses, com custo orçado em R$20 milhões. Devido à eficiência do processo executivo utilizado, a obra foi entregue em 18 meses, consumindo 85% dos recursos planejados.

Portanto, o impacto foi direto nos objetivos de custo e cronograma do projeto, e o cliente ficou imensamente satisfeito pelos resultados alcançados. O resultado do risco em utilizar processos tecnológicos mais avançados foi positivo, pois menos recursos foram consumidos e a obra conduzirá a benefícios mais rapidamente.

Por outro lado, os riscos negativos estão diretamente associados a impactos não desejáveis em algum objetivo do projeto. Por exemplo, o atraso na concessão de uma autorização ambiental para inícios das obras, o que acarreta umimpacto negativo direto no cronograma do projeto.

Os riscos podem ter maior ou menor grau de impacto e probabilidade de ocorrência. Diante disso, torna-se necessário priorizá-los a fim de aumentar o desempenho do projeto.

É necessário se concentrar nos riscos de alta prioridade, porém sem ignorar os riscos de menor grau de prioridade. Estes, por sua vez, devem compor uma lista de observação.

Mas essa análise de impacto e probabilidade dos riscos não é bastante subjetiva? Ora, um risco X pode ter um impacto ALTO no projeto, enquanto que, para outra pessoa, esse mesmo risco é considerado apenas MODERADO.

A fim de mitigar esse risco da subjetividade da análise, o PMBOK afirma que o estabelecimento de definições dos níveis de probabilidade e impacto pode reduzir a influência de parcialidade. Uma sugestão é a definição de uma escala de impactos para os quatro objetivos do projeto. Observemos a figura apresentada a seguir:

Condições definidas para escalas de impacto de um risco em objetivos importantes do projeto (os exemplos são mostrados somente para impactos negativos)					
Objetivo do projeto	São mostradas escalas relativas ou numéricas				
	Muito baixo / 0,05	Baixo / 0,10	Moderado / 0,20	Alto / 0,40	Muito alto / 0,80
Custo	Aumento de custo não significativo	Aumento de custo < 10%	Aumento de custo de 10% a 20%	Aumento de custo de 20% a 40%	Aumento de custo > 40%
Tempo	Aumento de tempo não significativo	Aumento de tempo < 5%	Aumento de tempo de 5% a 10%	Aumento de tempo de 10% a 20%	Aumento de tempo > 20%
Escopo	Diminuição do escopo quase imperceptível	Áreas menos importantes do escopo afetadas	Áreas importantes do escopo afetadas	Redução do escopo inaceitável para o patrocinador	Item final do projeto sem nenhuma utilidade
Qualidade	Degradação da qualidade quase imperceptível	Somente as aplicações mais críticas são afetadas	Redução da qualidade exige a aprovação do patrocinador	Redução da qualidade inaceitável para o patrocinador	Item final do projeto sem nenhuma utilidade

Figura 03- Exemplo de condições para definição de escalas de risco em projeto18.

Fonte: Sítio <}http://danielettinger.files.wordpress.com/2011/06/> [consulta em 01/05/2012]

As definições dos níveis de probabilidade e impacto podem ser adaptadas a cada projeto de acordo com o ambiente organizacional. Tabelas semelhantes podem ser definidas para os riscos positivos do projeto.

Diante dos riscos a que qualquer projeto encontra-se exposto, torna-se necessário encontrar uma ferramenta que possibilite priorizá-los de forma, no mínimo, aceitável.

É nesse cenário que surge a matriz de probabilidade e impacto. Elas especificam as combinações de probabilidade e impacto que resultam em uma classificação dos riscos como de prioridade BAIXA, MODERADA ou ALTA, conforme poderemos observar na tabela apresentada a seguir:

Tabela 2 - Exemplo de definição de matriz de probabilidade e impacto em projetos.

Prob.	Ameaças					Oportunidades				
0.90	0.05	0.09	0.18	0.36	0.72	0.72	0.36	0.18	0.09	0.05
0.70	0.04	0.07	0.14	0.28	0.56	0.56	0.28	0.14	0.07	0.04
0.50	0.03	0.05	0.10	0.20	0.40	0.40	0.20	0.10	0.05	0.03
0.30	0.02	0.03	0.06	0.12	0.24	0.24	0.12	0.06	0.03	0.02
0.10	0.01	0.01	0.02	0.04	0.08	0.08	0.04	0.02	0.01	0.01
	0.05	0.10	0.20	0.40	0.80	0.80	0.40	0.20	0.10	0.05

Assim, as células em cinza escuro, com maiores valores, representam alto risco, enquanto que as preenchidas com cinza médio são riscos baixos. As células em tonalidade cinza claro representam riscos moderados. Por exemplo, um risco negativo com probabilidade de ocorrência estimada em 0,70 e impacto de 0,10 representa uma ameaça de valor 0,07, o que é considerada como MODERADA.

Neste exemplo, ter-se-ia a definição de valores aceitáveis para as probabilidades, para ameaças e oportunidades, graduadas no eixo das abscissas conforme o seu impacto no projeto (0,05 a 0,8), conduzindo a código de cores para as classes de riscos em função da prioridade (baixa, moderada ou alta) .

Através dessa ferramenta, é possível ainda identificar áreas do projeto que sofrem mais ameaças ou oportunidades. Para isso, torna-se necessária a categorização dos riscos identificados no projeto, gerando assim uma estrutura hierárquica de riscos.

A avaliação de probabilidade de riscos investiga a probabilidade de cada risco específico ocorrer. A avaliação de impacto de riscos investiga o efeito potencial sobre um objetivo do projeto, como tempo, custo, escopo ou qualidade, inclusive os efeitos negativos das ameaças e os efeitos positivos das oportunidades.

A probabilidade e o impacto seriam avaliados para cada risco identificado. Os riscos podem ser avaliados em entrevistas ou reuniões com participantes selecionados por sua familiaridade com as categorias de risco da pauta. São incluídos os membros da equipe do projeto e, talvez, especialistas de fora do projeto. A opinião especializada é necessária, pois podem existir poucas informações sobre riscos no banco de dados de projetos passados da organização. Um facilitador experiente pode liderar a discussão, pois os participantes podem ter pouca experiência em avaliação de riscos.

A probabilidade de cada risco e seu impacto em cada objetivo são avaliados durante a entrevista ou reunião. Os detalhes da explanação, inclusive as premissas que justificam os níveis atribuídos, também são registrados.

As probabilidades e impactos de riscos são classificados de acordo com as definições fornecidas no plano de gerenciamento de riscos.

Às vezes, os riscos com probabilidade e impacto visivelmente baixos não serão classificados, mas serão incluídos em uma lista de observação para monitoramento futuro.

SAIBA MAIS

Para ampliar seus conhecimentos sobre gerenciamento de riscos em projetos, sugerimos o estudo de métodos estatísticos para análise, como o método de Monte Carlo, uma das ferramentas tratadas no curso “Avaliação de Riscos” do curso presencial do Programa Avaliação Socioeconômica de Projetos.

2.5. CONCEITOS FUNDAMENTAIS SOBRE AMOSTRAGEM

O processo de retirada de amostras de uma população, denominado de amostragem, é uma das etapas fundamentais na tomada de decisões nos diversos níveis gerenciais, tendo em vista que uma amostragem mal executada geralmente conduz a estatísticas pouco confiáveis, bem como a tomada de decisão associada a uma margem de erro importante.

Este tópico tem por objetivo apresentar a você alguns conceitos e definições essenciais para conduzir convenientemente uma operação de amostragem, visando principalmente à coleta de dados socioeconômicos. Amostragem é o processo que procura extrair da população elementos que, através de cálculos probabilísticos, ou não, consigam prover dados inferenciais da população-alvo.

Tipos de Amostragem	Não Probabilística
	Acidental ou conveniência
	Intencional
	Quotas ou proporcional
	Desproporcional
	Probabilística
	Aleatória Simples
	Aleatória Estratificada
	Conglomerado

Não Probabilística

A escolha de um método não probabilístico, via de regra, sempre encontrará desvantagem frente ao método probabilístico. No entanto, em alguns casos, se faz necessária a opção por este método. Convém ressaltar que não há formas de se generalizar os resultados obtidos na amostra para o todo da população quando se opta por este método de amostragem.

Amostragem Não Probabilística

Quando trabalhamos com a amostragem não probabilística, não conhecemos a priori a probabilidade que um elemento da população tem de pertencer à amostra. Neste caso, não é possível calcular o erro decorrente da generalização dos resultados das análises estatísticas da amostra para a população de onde a amostra foi retirada. Utiliza-se geralmente a amostragem não probabilística pela simplicidade intrínseca ao processo, bem como pela impossibilidade de se obter uma amostra probabilística, como seria desejável. Os principais tipos de amostragem não probabilística que temos são amostragem sem norma ou a esmo, intencional e por cotas.

Amostragem a Esmo

Imagine um rebanho com 1.000 animais, integrantes de um lote de gado bovino a ser destinado para corte. A enumeração destes animais ficaria muito difícil, tornando-se a amostragem aleatória simples inviável. Assim, em situações deste tipo, supondo que a população de animais seja homogênea, escolhemos a esmo a quantidade relativa ao tamanho da amostra, e quanto mais homogênea for a população, mais podemos supor a equivalência com uma amostragem aleatória simples. Os animais, desta forma, serão escolhidos para compor a amostra de um determinado tamanho sem nenhuma norma ou a esmo, o que origina o nome deste tipo de amostragem.

Acidental ou conveniência

Indicada para estudos exploratórios. Frequentemente utilizados em supermercados para testar produtos.

Intencional

A amostragem intencional corresponde àquela em que o responsável pela amostragem escolhe, deliberadamente, determinados elementos para pertencer à amostra, em função de seu juízo de valor sobre a representatividade da população. Um exemplo deste tipo de amostragem corresponde à situação em que se deseja saber a aceitação em relação a uma nova marca de automóvel de passeio a ser inserida no mercado de uma cidade. Somente entrarão para compor a amostra pessoas que façam uso do automóvel e que tenham condições financeiras de comprar esta nova marca (público-alvo).

O entrevistador dirige-se a um grupo em específico para saber sua opinião. Por exemplo, quando de um estudo sobre automóveis, o pesquisador procura apenas oficinas.

Quotas ou proporcional

Neste tipo de amostragem é procedida a divisão da população em grupos, selecionando-se uma cota proporcional ao tamanho de cada grupo. Entretanto, dentro de cada grupo não é feito sorteio, sendo procurados os elementos até que a cota de cada grupo seja alcançada.

Em pesquisas eleitorais, a divisão de uma população em grupos (considerando, por exemplo, o sexo, o nível de escolaridade, a faixa etária e a renda) pode servir de base para a definição dos grupos, partindo da suposição de que estas variáveis definem grupos com comportamentos diferenciados no processo eleitoral. Para se ter uma ideia do tamanho destes grupos, pode-se recorrer a pesquisas feitas anteriormente pelo IBGE (Instituto Brasileiro de Geografia e Estatística).

Na realidade, trata-se de uma variação da amostragem intencional. Necessita-se ter um prévio conhecimento da população e sua proporcionalidade. Por exemplo, deseja- se entrevistar apenas indivíduos da classe A, que representam 12% da população. Esta será a quota para o trabalho. Comumente também substratifica-se uma quota obedecendo a uma segunda proporcionalidade.

Desproporcional

Muito utilizada quando a escolha da amostra for desproporcional à população. Atribui- se pesos para os dados, e assim obtém-se resultados ponderados representativos para o estudo.

Probabilística

Para que se possam realizar inferências sobre a população, é necessário que se trabalhe com amostragem probabilística. É o método que garante segurança quando se investiga alguma hipótese. Normalmente os indivíduos investigados possuem a mesma probabilidade de ser selecionado na amostra.

Aleatória Simples

Você deve utilizar a amostragem aleatória simples somente quando a população for homogênea em relação à variável que se deseja estudar. Geralmente, atribuímos uma numeração a cada indivíduo da população, e através de um sorteio aleatório, os elementos que vão compor a amostra são selecionados.

Todos os elementos da população têm a mesma probabilidade de pertencer à amostra. Imagine que você queira amostrar um número de pessoas que estão fazendo um determinado concurso com n inscritos. Como a população é finita, devemos enumerar cada um dos n candidatos e sortear n deles. É o mais utilizado processo de amostragem. Prático e eficaz, confere precisão ao processo de amostragem. Normalmente utiliza-se uma tabela de números aleatórios e nomeiam-se os indivíduos, sorteando-se um por um até completar a amostra calculada. Uma variação deste tipo de amostragem é a sistemática. Em um grande número de exemplos, o pesquisador depara-se com a população ordenada. Neste sentido, tem-se os indivíduos dispostos em sequência, o que dificulta a aplicação exata desta técnica.

Amostragem Sistemática

Porém, é de fundamental importância que a variável de interesse não apresente ciclos de variação coincidente com os ciclos de retirada, pois este fato tornará a amostragem não aleatória.

A técnica de amostragem caracterizada pela coleta na população de elementos que vão compor a amostra, de forma cíclica (em períodos), se chama amostragem sistemática, interessante de ser utilizada quando convém obter a amostra, por exemplo, quando os elementos da população apresentam-se ordenados.

Uma situação típica é quando se necessita retirar uma amostra de população onde a variável tempo condiciona características presentes nos elementos estudados, como é o caso da análise da evolução dos índices de preços para a consideração do efeito da inflação, durante certo período, sobre uma série de pagamentos.

Quanto ao procedimento, o primeiro passo será o cálculo da constante **(K)**, que atuará como fator de ciclo para a definição do momento e período a considerar para a realização da amostragem o que possibilitará, em um segundo momento, a definição do intervalo de tempo para a coleta da amostra.

Após a definição do valor de **K**, é definido o ponto inicial da amostragem, ou seja, um dos elementos do primeiro intervalo é constituído pelos elementos populacionais numerados de 1 até n. Escolhe-se o seguinte, que será o elemento de ordem **(i + K)**; , sempre somando-se K à ordem do elemento anterior, até completar a escolha dos n elementos que vão compor a amostra.

Amostragem Aleatória Estratificada

Quando se deseja guardar uma proporcionalidade na população heterogênea e a variável de interesse apresenta uma heterogeneidade na população, permitindo a identificação de grupos homogêneos, poderá a população ser dividida em grupos (estratos) e proceder-se à amostragem dentro de cada estrato, garantindo, assim, a representatividade de cada estrato na amostra. A estratificação em cada subpopulação poderá ser realizada mediante critérios como classe social, renda, idade, sexo, entre outros.

Conforme comenta Tavares (2007), podemos verificar que pesquisas eleitorais apresentam uma grande heterogeneidade em relação à intenção de votos, quando consideramos, por exemplo, a faixa salarial ou o nível de escolaridade.

Então, se fizéssemos uma amostragem aleatória simples, poderíamos incluir na amostra uma maior quantidade de elementos de um grupo, e, proporcionalmente, este grupo será pequeno em relação à população. Desta forma, não teríamos uma amostra representativa da população a ser estudada. Então, podemos dividir a população em grupos (estratos) que são homogêneos para a característica que estamos avaliando, ou seja, neste caso, a intenção de votos.

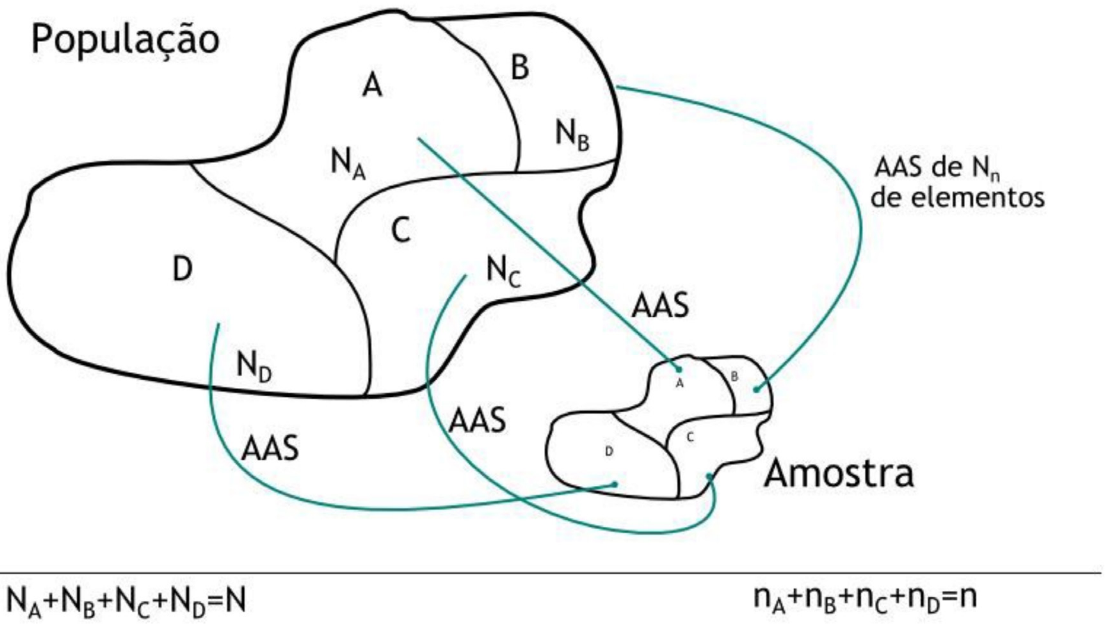
Destaca-se que, como estamos dividindo a população em estratos (grupos) que são homogêneos dentro de si, podemos, então, caracterizar a amostragem estratificada. Para efetuarmos a amostragem estratificada de forma proporcional, precisamos primeiramente definir a proporção do estrato em relação à população.

Proporção do estrato h será igual ao número de elementos presentes neste estrato (Nh) dividido pelo tamanho da população (N) --- (Nh/N).

Após obter esta proporção do estrato em relação à população, deve-se multiplicar o tamanho total da amostra (n) pela proporção de cada estrato na população (Nh/N). Assim, teremos um tamanho de amostra em cada estrato, proporcional ao tamanho do estrato em relação à população.

A figura apresentada a seguir mostra como é feita a escolha dos elementos de cada estrato (A, B, C, D), salientando o que você pode fazer usando amostragem aleatória simples devido ao fato de os estratos serem homogêneos individualmente, considerando a variável de interesse.

Figura 11- Critério para a escolha dos elementos de cada estrato (A, B, C, D).



Fonte: Tavares, 2007.

Conglomerado

Em corriqueiras situações, torna-se difícil coletar características da população. Nesta modalidade de amostragem, sorteia-se um conjunto e procura-se estudar todo o conjunto. É exemplo de amostragem por conglomerado, famílias, organizações e quarteirões.
Destaca Tavares (2007) que apesar de a amostragem estratificada apresentar resultados satisfatórios, a sua implementação é dificultada pela falta de informações sobre a população para fazer a estratificação, o que poderá ser contornado mediante a utilização de esquema de amostragem chamado “amostragem por conglomerados”, definidos em função da experiência do gestor ou pesquisador.

Comenta ainda Tavares (2007) que a utilização da amostragem por conglomerados possibilita uma redução significativa do custo do processo de amostragem. Portanto, um conglomerado é um subgrupo da população, que individualmente reproduz a população, ou seja, individualmente, os elementos que o compõem são muito heterogêneos entre si. Ressalta ainda que este tipo de amostragem é muito útil quando a população é grande, por exemplo, no caso de uma pesquisa em nível nacional. Para efetuarmos a amostragem por conglomerados, primeiramente os definimos e depois dividimos a população nos conglomerados, procedendose ao sorteio destes por meio de um processo aleatório onde são avaliados todos os indivíduos presentes em um estágio.

IMPORTANTE

Conglomerados são definidos geralmente segundo fatores geográficos. . A Pesquisa Nacional por Amostra de Domicílios (PNAD) do IBGE, por exemplo, é feita por conglomerados, em três estágios.

2.6. DIMENSIONAMENTO DA AMOSTRA

Efetivamente, um dos temas mais importantes em uma análise estatística trata da determinação de qual o melhor tamanho de amostras que devemos ter, sabendo-se que amostras grandes são dispendiosas e demandam mais tempo de manipulação e estudo, bem como amostras pequenas são menos precisas e pouco confiáveis. Quando se deseja dimensionar o tamanho da amostra, o procedimento desenvolve-se em três etapas distintas:

- Avaliar a variável mais importante do grupo e a mais significativa.
- Analisar se é ordinal, intervalar ou nominal.
- Verificar se a população é finita ou infinita.

Abaixo podemos observar diversas expressões propostas para a determinação do tamanho da amostra:

Variável intervalar e população infinita $$n = \left(\frac{Z.\sigma}{d}\right)^2$$

Variável intervalar e população finita $$n = \frac{Z^2.\sigma^2.N}{d^2(N-1) + z^2.\sigma^2}$$

Variável nominal ou ordinal e população infinita $$n = \left(\frac{z^2.pq}{d^2}\right)$$

Variável nominal ou ordinal e população finita $$n = \frac{z^2.p.q.N}{d^2(N-1)+z^2.p.q}$$

Observa-se que a variável **d** corresponde ao erro admissível, sendo **Z** função do nível de confiança a ser considerado. Para os casos de populações consideradas como finitas, a variável N corresponde ao tamanho da população. A proporção (p) será a estimativa da verdadeira proporção de um dos níveis escolhidos para a variável adotada. Por exemplo, 55% dos integrantes da população é do sexo feminino, então p será 0,55. A proporção (q) será sempre α . Neste exemplo q, será 0,45. O erro é representado por d. Para casos em que não se tenha como identificar as proporções confere-se 0,5 para **p** e **q**.

Como exemplo, realizaremos o cálculo do tamanho de amostra para o caso de população considerada como infinita, recorrendo-se à equação correspondente a variável intervalar. Observa-se que o tamanho da amostra depende do grau de confiança desejado, da margem de erro pretendida e do desvio padrão (σ).
Vamos ao exemplo:

Queremos estimar a renda média no primeiro ano de um egresso de escola técnica profissional. Para um nível de confiança de 95% [α = 0,05 $\rightarrow$Z = 1,96; conforme relação Z = f(α)], quantas amostras deveremos ter para que a média esteja a menos que R$250,00 da renda média verdadeira da população. Suponha σ conhecido e igual a R$800,00?

$$n = [\ Z.\sigma/d\ ]^2 = [1{,}96\ .\ 800/250]^2 = 39{,}34 \Rightarrow 40 \text{ amostras}$$

Caso fosse admissível uma margem de erro de R$500, teríamos:

$$n = [\ Z.\sigma/d\ ]^2 = [1{,}96\ .\ 800/500]^2 = 9{,}83 \Rightarrow 10 \text{ amostras}$$

Ou seja, caso o erro admissível fosse o dobro, o tamanho da amostra poderia ser reduzido à quarta parte.

IMPORTANTE

A utilização de uma amostra probabilística é melhor para garantir a representatividade da amostra, pois o acaso será o único responsável por eventuais discrepâncias entre população e amostra.

Estas discrepâncias são levadas em consideração nas inferências estatísticas.

REFERÊNCIAS

UNIDADE 1

BARBETTA, P. A. **Estatística Aplicada às Ciências Sociais**. 4 ed. Florianópolis: Ed. da UFSC, 2002.

BRAULE, Ricardo. **Estatística Aplicada com Excel: para cursos de Administração e Economia**. Rio de Janeiro: Campus, 2001.

BUSSAB, W. O.; MORETTIN, P. **Estatística Básica**. São Paulo: Atual, 2002.

COSTA NETO, P. L. de O. **Estatística**. São Paulo: Edgard Blücher, 2002.

DOWNING, D.; CLARK, J. **Estatística Aplicada**. São Paulo: Saraiva, 2000.

FONSECA, Jairo Simon da; MARTINS, Gilberto de Andrade. **Curso de Estatística**. Rio de Janeiro: LTC, 1982.

FREUD, J. E.; SIMON, G. A. **Estatística Aplicada**. Bookman, 2000, 403 p.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. **Estatística: teoria e aplicações (usando o Microsoft Excel em português)**. LTC, 2000, 812 p.

MORETTIN, L. G. **Estatística Básica:** Probabilidade. V. 1. São Paulo: Makron Books, 1999.

SOARES, J. F.; FARIAS, A. A.; CESAR, C. C. **Introdução à Estatística**. Rio de Janeiro: LTC, 1991.

SPIEGEL, M. **Probabilidade e Estatística**. Mc Graw Hill. 1993.

TAVARES, M. **Estatística Aplicada à Administração. Secretaria de Educação à Distância do Ministério da Educação Sistema Universidade Aberta do Brasil** – UAB. Diretoria do Departamento de Políticas em Educação a Distância – DPEAD. Brasília: MEC, 2007.

TRIOLA, M. F. Introdução à Estatística. Rio de Janeiro: LTC, 1999.

WONNACOTT, T. H., WONNACOTT, R. J. **Estatística Aplicada à Economia e à Administração**. Rio de Janeiro: LTC, 1981.

UNIDADE 2

BARBETTA, P. A. **Estatística Aplicada às Ciências Sociais**. 4 ed. Florianópolis: Ed. da UFSC, 2002.

BRAULE, Ricardo. **Estatística Aplicada com Excel: para cursos de Administração e Economia.** Rio de Janeiro: Campus, 2001.

BUSSAB, W. O.; MORETTIN, P. **Estatística Básica**. São Paulo: Atual, 2002.

COSTA NETO, P. L. de O. **Estatística**. São Paulo: Edgard Blücher, 2002.

DOWNING, D.; CLARK, J. **Estatística Aplicada**. São Paulo: Saraiva, 2000.

FONSECA, Jairo Simon da; MARTINS, Gilberto de Andrade. **Curso de Estatística**. Rio de Janeiro: LTC, 1982.

FREUD, J. E.; SIMON, G. A. **Estatística Aplicada.** Bookman, 2000, 403 p.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. **Estatística: teoria e aplicações (usando o Microsoft Excel em português)**. LTC, 2000, 812 p.

MORETTIN, L. G. **Estatística Básica: Probabilidade**. V. 1. São Paulo: Makron Books, 1999.

SOARES, J. F.; FARIAS, A. A.; CESAR, C. C. **Introdução à Estatística.** Rio de Janeiro: LTC, 1991.

SPIEGEL, M. **Probabilidade e Estatística.** Mc Graw Hill. 1993.

STEVENSON, William J. **Estatística Aplicada à Administração.** São Paulo: Harper, 1981.

TAVARES, M. **Estatística Aplicada à Administração. Secretaria de Educação à Distância do Ministério da Educação Sistema Universidade Aberta do Brasil – UAB.** Diretoria do Departamento de Políticas em Educação a Distância – DPEAD. Brasília: MEC, 2007.

TRIOLA, M. F. **Introdução à Estatística**. Rio de Janeiro: LTC, 1999.

WONNACOTT, T. H., WONNACOTT, R. J. **Estatística Aplicada à Economia e à Administração.** Rio de Janeiro: LTC, 1981.