e-Tec Brasil
Escola Técnica Aberta do Brasil

Curso Técnico em Meio Ambiente

Estatística Ambiental

Cristiano Poleto




INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
RIO GRANDE DO SUL

UNIVERSIDADE FEDERAL DE SANTA CATARINA

Ministério da Educação

CRISTIANO POLETO

ESCOLA TÉCNICA ABERTA DO BRASIL – E-TEC BRASIL

CURSO TÉCNICO EM MEIO AMBIENTE

Disciplina: Estatística Ambiental

ESCOLA TÉCNICA DA UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Porto Alegre – RS
2008

Presidência da República Federativa do Brasil
Ministério da Educação
Secretaria de Educação a Distância

Este Caderno foi elaborado em parceria entre a Escola Técnica da Universidade Federal do Rio Grande do Sul e a Universidade Federal de Santa Catarina para o Sistema Escola Técnica Aberta do Brasil – e-Tec Brasil.

Equipe de Elaboração
Escola Técnica da Universidade Federal do Rio Grande do Sul

Coordenação Institucional
Eduardo Luiz Fonseca Benites/Escola Técnica da UFRGS
Professor-autor
Cristiano Poleto/Escola Técnica da UFRGS

Comissão de Acompanhamento e Validação
Universidade Federal de Santa Catarina – UFSC

Coordenação Institucional
Araci Hack Catapan/UFSC

Coordenação de Projeto
Silvia Modesto Nassar/UFSC

Coordenação de Design Instrucional
Beatriz Helena Dal Molin/UNIOESTE

Design Instrucional
Dóris Roncarelli/UFSC
Mércia Freire Rocha Cordeiro Machado/ ETUFPR

Web Design
Beatriz Wilges/UFSC

Projeto Gráfico
Beatriz Helena Dal Molin/UNIOESTE
Araci Hack Catapan/UFSC
Elena Maria Mallmann/UFSC
Jorge Luiz Silva Hermenegildo/CEFET-SC
Mércia Freire Rocha Cordeiro Machado/ETUFPR
Silvia Modesto Nassar/UFSC

Supervisão de Projeto Gráfico
Ana Carine García Montero/UFSC

Diagramação
Rafaela Wiele Anton/UFSC
Luís Henrique Lindner/UFSC
Bruno César B. S. de Ávila/UFSC
Juliana Passos Alves/UFSC

Revisão
Lúcia Locatelli Flôres/UFSC

Catalogação na fonte elaborada na DECTI da Biblioteca da UFSC

P765e Poleto, Cristiano

Estatística ambiental / Cristiano Poleto. - Porto Alegre : Escola Técnica da Universidade Federal do Rio Grande do Sul, 2008. 49p. : il

Inclui bibliografia

Curso Técnico em Meio Ambiente, desenvolvido pelo Programa Escola Técnica Aberta do Brasil.

1. Meio ambiente – Estatística. 2. Probabilidades. 3. Amostragem (Estatística). 4. Ensino à distância. I. Título. II.Título: Curso Técnico em Meio Ambiente.

CDU: 577.4:519.2

PROGRAMA E-TEC BRASIL

Amigo(a) estudante!

O Ministério da Educação vem desenvolvendo Políticas e Programas para expansão da Educação Básica e do Ensino Superior no País. Um dos caminhos encontrados para que essa expansão se efetive com maior rapidez e eficiência é a modalidade a distância. No mundo inteiro são milhões os estudantes que frequentam cursos a distância. Aqui no Brasil, são mais de 300 mil os matriculados em cursos regulares de Ensino Médio e Superior a distância, oferecidos por instituições públicas e privadas de ensino.

Em 2005, o MEC implantou o Sistema Universidade Aberta do Brasil (UAB), hoje, consolidado como o maior programa nacional de formação de professores, em nível superior.

Para expansão e melhoria da educação profissional e fortalecimento do Ensino Médio, o MEC está implementando o Programa Escola Técnica Aberta do Brasil (e-Tec Brasil). Espera, assim, oferecer aos jovens das periferias dos grandes centros urbanos e dos municípios do interior do País oportunidades para maior escolaridade, melhores condições de inserção no mundo do trabalho e, dessa forma, com elevado potencial para o desenvolvimento produtivo regional.

O e-Tec é resultado de uma parceria entre a Secretaria de Educação Profissional e Tecnológica (SETEC), a Secretaria de Educação a Distância (SEED) do Ministério da Educação, as universidades e escolas técnicas estaduais e federais.

O Programa apóia a oferta de cursos técnicos de nível médio por parte das escolas públicas de educação profissional federais, estaduais, municipais e, por outro lado, a adequação da infra-estrutura de escolas públicas estaduais e municipais.

Do primeiro Edital do e-Tec Brasil participaram 430 proponentes de adequação de escolas e 74 instituições de ensino técnico, as quais propuseram 147 cursos técnicos de nível médio, abrangendo 14 áreas profissionais. O resultado desse Edital contemplou 193 escolas em 20 unidades federativas. A perspectiva do Programa é que sejam ofertadas 10.000 vagas, em 250 polos, até 2010.

Assim, a modalidade de Educação a Distância oferece nova interface para a mais expressiva expansão da rede federal de educação tecnológica dos últimos anos: a construção dos novos centros federais (CEFETs), a organização dos Institutos Federais de Educação Tecnológica (IFETs) e de seus *campi*.

O Programa e-Tec Brasil vai sendo desenhado na construção coletiva e participação ativa nas ações de democratização e expansão da educação profissional no País, valendo-se dos pilares da educação a distância, sustentados pela formação continuada de professores e pela utilização dos recursos tecnológicos disponíveis.

A equipe que coordena o Programa e-Tec Brasil lhe deseja sucesso na sua formação profissional e na sua caminhada no curso a distância em que está matriculado(a).

Brasília, Ministério da Educação – setembro de 2008.

SUMÁRIO

APRESENTAÇÃO

A Estatística é a área do conhecimento humano que utiliza teorias probabilísticas para explicar eventos, estudos e experimentos; portanto é uma ciência que se desenvolve através do uso de dados empíricos. Tem como objetivos a obtenção, a organização e análise de dados, e a determinação de correlações que sejam capazes de descrever e explicar o que ocorreu e possibilitar uma previsão de futuras ocorrências.

Aplicabilidade da Estatística

O papel da Estatística na investigação científica vai além de indicar a seqüência de cálculos a serem realizados com os dados obtidos. No planejamento, ela auxilia na escolha das situações experimentais e na determinação da quantidade de indivíduos a serem examinados. Na análise dos dados, indica técnicas para resumir e apresentar as informações, bem como para comparar as situações experimentais. Na elaboração das conclusões, os vários métodos estatísticos permitem generalizar a partir dos resultados obtidos. De um modo geral, não existe certeza sobre a correção das conclusões científicas. No entanto, os métodos estatísticos permitem determinar a margem de erro associada às conclusões, com base no conhecimento da variabilidade observada nos resultados (CALLEGARI-JACQUES, 2004).

O que é um experimento? A essência de um experimento está em nos habilitar a comparar os efeitos que dois ou mais "tratamentos" têm sobre alguns atributos das plantas, dos animais ou de qualquer outro material experimental. Para que as comparações sejam válidas, é preciso que o material a ser submetido a tratamentos diferentes seja escolhido sem qualquer predisposição. O experimento, para ser válido, deve não só fornecer informações sobre a natureza e a magnitude dos efeitos aparentes, mas também permitir uma estimativa de variabilidade (HEATH, 1981).

Para compreender melhor a aplicabilidade da estatística, acesse o *link* abaixo para assistir a um vídeo de como realizar uma pesquisa utilizando a estatística como uma ferramenta:

http://br.youtube.com/watch?v=GtL17QiqfOs&feature=related

Elabore uma lista de outros tipos de pesquisas, nas quais a utilização da estatística seja possível.

PALAVRAS DO PROFESSOR-AUTOR

Parabéns e sejam bem-vindos à disciplina de Estatística Ambiental!

Essa disciplina aborda os principais tópicos da área com a finalidade de auxiliar estudos de consistência e interpretação de dados obtidos através de estudos ambientais e de saúde.

A Estatística Ambiental pode ser utilizada como um instrumento para validação de materiais, estudos de clima, estudos populacionais ou de produtos. Além disso, a Estatística é a base para a comprovação de vários estudos e teorias nas áreas de biologia, química e engenharias.

Os conceitos adquiridos nessa disciplina poderão ser utilizados em diversos estudos ambientais, em práticas de laboratório e nos resultados de trabalhos de campo que serão desenvolvidas nas disciplinas de Bioindicadores, Geoprocessamento, Geografia Aplicada, Gestão de Recursos Hídricos e Análise de Impacto Ambiental.

As mídias e os exercícios propostos ao longo do material servirão de apoio e devem ser utilizados para que os seus conhecimentos sejam mais consistentes e aprofundados.

Sucesso e bons estudos!

PROJETO INSTRUCIONAL

UNIDADE	OBJETIVOS	MATERIAL IMPRESSO	RECURSOS DIGITAIS	CARGA HORÁRIA	ESTRATÉGIAS	ATIVIDADES DE AVALIAÇÃO
1	Apresentar de forma simplificada o que é e para que serva a Estatística em estudos ambientais.	Texto contendo a Introdução e as palavras do professor.	Vídeo sobre a utilização da Estatística em pesquisas científicas.	01 hora	Aula expositiva disponibilizada em *PowerPoint*.	Elaboração, pelos estudantes de uma lista de possibilidades de utilização da estatística.
2	Apresentar as principais maneiras de se tratar os dados obtidos por meio de pesquisas quantitativas.	Descrição dos tópicos básicos sobre Estatística Descritiva.	Hipertexto e vídeos abordando os principais itens da unidade.	05 horas	Aula expositiva disponibilizada em *PowerPoint*. Utilização de textos, vídeos e exercícios resolvidos.	Resolução de exercícios aplicados a dados ambientais.
3	Analisar dados, realizando estudos estatísticos capazes de generalizar e obter conclusões sobre uma determinada população.	Apresentação dos principais tópicos relacionados à Probabilidade, Amostragem e Distribuição.	Hipertexto e vídeos abordando os principais itens da unidade.	05 horas	Aula expositiva disponibilizada em *PowerPoint*. Utilização de textos, vídeos e exercícios resolvidos.	Resolução de exercícios aplicados a dados ambientais.
4	Conduzir a aplicação dos conhecimentos de probabilidade e distribuição amostral para realizar testes de hipóteses.	Elaboração de um Teste de Hipóteses e Significância Estatística.	Hipertexto com exercícios resolvidos.	05 horas	Aula expositiva disponibilizada em *PowerPoint*. Utilização de textos, vídeos e exercícios resolvidos.	Resolução de exercícios aplicados a dados gerados por outras disciplinas.
5	Verificar se existe um relacionamento entre duas variáveis em estudo.	Apresentação de análises sobre relações de Correlações Bivariadas.	Hipertexto abordando correlação através do uso de um *software*.	03 horas	Aula expositiva disponibilizada em *PowerPoint*. Utilização de exercícios resolvidos.	Resolução de exercícios aplicados a dados gerados por outras disciplinas.
6	Apresentar como é obtida uma relação de causa-efeito entre duas variáveis quantitativas.	Mostra como se obtém uma Regressão Linear Simples.	Hipertexto sobre Regressão Linear Simples.	03 horas	Aula expositiva disponibilizada em *PowerPoint*. Utilização de exercícios resolvidos.	Resolução de exercícios aplicados a dados gerados por outras disciplinas.
7	Apresentar uma extensão da análise multivariada.	Apresentação da análise de componentes principais e análise de fatores.	Hipertexto abordando os principais itens da unidade.	03 horas	Aula expositiva disponibilizada em *PowerPoint*. Utilização de textos e exercícios resolvidos.	Resolução de exercícios aplicados a dados gerados por outras disciplinas.

ÍCONES E LEGENDAS

Caro estudante! Oferecemos para seu conhecimento os ícones e sua legenda que fazem parte da coluna de indexação. A intimidade com estes e com o sentido de sua presença no caderno ajudará você a compreender melhor as atividades e exercícios propostos (DAL MOLIN, *et al.*,2008).

Saiba mais

Ex: http://www.etecbrasil.mec.gov.br

Este ícone apontará para atividades complementares ou para informações importantes sobre o assunto. Tais informações ou textos complementares podem ser encontrados na fonte referenciada junto ao ícone.

Para refletir...

Ex: Analise o caso... dentro deste tema e compare com..., Assista ao filme...

Toda vez que este ícone aparecer na coluna de indexação indicará um questionamento a ser respondido, uma atividade de aproximação ao contexto no qual você vive ou participa, resultando na apresentação de exemplos cotidianos ou links com seu campo de atuação.

Mídias integradas

Ex.: Assista ao filme... e comente-o.

Quando este ícone for indicado em uma dada unidade significa que você está sendo convidado a fazer atividades que empreguem diferentes mídias, ou seja, participar do ambiente AVEA, assistir e comentar um filme, um videoclipe, ler um jornal, comentar uma reportagem, participar de um chat, de um fórum, enfim, trabalhar com diferentes meios de comunicação.

Avaliação

Este ícone indica uma atividade que será avaliada dentro de critérios específicos da unidade.

Lembre-se

Ex.: O canal de satélite deve ser reservado com antecedência junto à Embratel.

A presença deste ícone ao lado de um trecho do texto indicará que aquele conteúdo significa algo fundamental para a aprendizagem.

Destaque

Retângulo com fundo colorido.

A presença do retângulo de fundo indicará trechos importantes do texto, destacados para maior fixação do conteúdo.

UNIDADE 1 – INTRODUÇÃO

1.1 Objetivo de aprendizagem

Descrever estatisticamente os dados obtidos por meio de pesquisas quantitavas.

1.2 Amostras e população

Uma população pode ser um grupo distinto de pessoas ou seres vivos (homens, mulheres, pessoas destras, etc.) ou de objetos inanimados (carros, computadores, etc.).

Uma amostra é simplesmente uma seleção de alguns elementos de uma determinada população.

1.3 Variáveis

As variáveis são características a respeito dos mais diferentes fenômenos que podem ser mensurados (resultados possíveis de um fenômeno).

É importante identificar que tipo de variável está sendo estudada, uma vez que procedimentos estatísticos diferentes são recomendados em cada situação. A principal divisão ocorre entre variáveis qualitativas e quantitativas.

1.3.1 Variáveis qualitativas

São aquelas que fornecem dados de natureza não numérica (denotam categorias de respostas). Exemplos: Cor de uma flor, raça de um animal, sexo de um paciente, etc.

Acesse no *link* abaixo, a um material complementar sobre Dados, Tabelas e Gráficos. Aproveite para visualizar os exemplos disponibilizados nesse material.

http://alea-estp.ine.pt/html/nocoes/html/cap3_1_i.html

1.3.2 Variáveis quantitativas

São aquelas em que os dados são valores numéricos que expressam quantidades, como a estatura das pessoas, níveis de metais pesados em animais ou o número de sementes integradas em uma vagem. Elas podem ainda ser classificadas em:

a) **Variáveis quantitativas discretas:** decorrem da contagem do número de itens de uma população, assumindo em geral, números reais inteiros como possíveis resultados. Por exemplo: número de espécies encontradas num ecossistema, número de dias secos (sem chuva) em determinada região, número de testes que indicaram níveis de poluição acima do padrão pré-estabelecido, etc.;

b) **Variáveis quantitativas contínuas:** são aquelas cujos dados podem apresentar qualquer valor dentro de um intervalo de variação possível. Por exemplo: medições pluviométricas, concentrações químicas, temperaturas, etc.

De um modo geral, as medições dão origem a variáveis contínuas e as contagens ou enumerações, a variáveis discretas.

1.4 Apresentação de dados (gráficos e tabelas)

Os componentes mais importantes de uma tabela são:

Uma boa organização dos dados, para posterior análise, é fundamental para facilitar a sua interpretação.

- título: explica o que a tabela contém ou expõe;
- corpo: parte da tabela composta por linhas e colunas;
- cabeçalho: especifica o conteúdo das colunas;
- rodapé: é o espaço onde são colocadas as notas de natureza informativa (fonte, notas e chamadas).

Título

Cabeçalho
Corpo
Rodapé

Tabela com distribuição de freqüência por ponto: A cada valor da variável associam-se as freqüências.

Exemplo: Considere os valores seguintes representando a concentração de um metal no sangue (µg/ml) de 15 indivíduos, de uma cidade X num determinado ano.

X = 20 20 21 21 21 22 23 23 24 24 22 20 22 20 21

c) Identifique a população:
Resposta: Indivíduos de uma cidade X.

d) Identifique a amostra:
Resposta: 15 Indivíduos de uma cidade X.

e) Identifique a variável:
Resposta: concentração de um metal no sangue.

f) Construa uma tabela para estes dados:
Resposta:

Concentração de um metal no sangue (µg/ml) de 15 indivíduos de uma cidade X

Concentração de metal	Número de indivíduos	%
19	4	26,67
20	4	26,67
21	3	20,00
22	2	13,33
23	2	13,33
Total	15	100

e) Construa dois gráficos para estes dados (usar gráfico de colunas e de barras):

Obs.: Gráficos de colunas e de barras podem ser utilizados para representar qualquer tipo de variável.

Gráfico de colunas

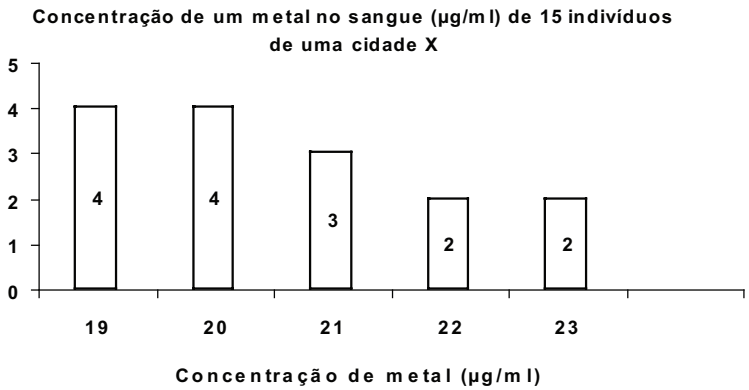


Gráfico de barras

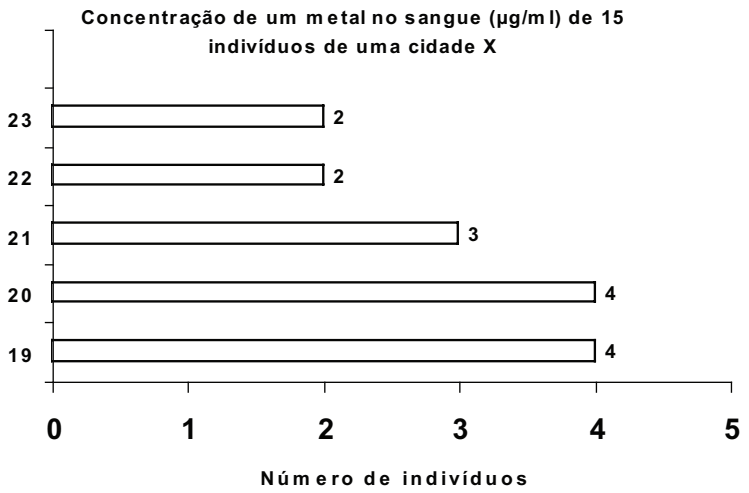


Tabela com distribuição de freqüência por classes (ou intervalo): Quando a variável estudada assumir uma quantidade de valores distintos muito grande, recomenda-se o agrupamento por classes.

Exemplo: Os dados abaixo referem-se às precipitações (mm), de uma amostra de 15 cidades, que ocorreram no mês de agosto de 2005:

35	26	39	25	39
21	40	16	32	39
23	15	27	44	50

Como construir a tabela com distribuição de freqüência por classes:

- Determinar o valor máximo (Vmáx.), valor mínimo (Vmín.) e a amplitude total (Vmáx- Vmín);

- Determinar o número de classes (K) que é dado por Sturges 1+3,3.log(n), onde n é o tamanho da amostra;

- Dividir a amplitude total pelo número de classes desejado (tamanho do intervalo "pulo").

a) Quem é a população alvo desde estudo?

Resposta: Precipitações que ocorreram no mês de agosto de 2005.

b) Construa uma tabela com 5 classes.

Precipitação (mm) em 15 cidades no mês de agosto

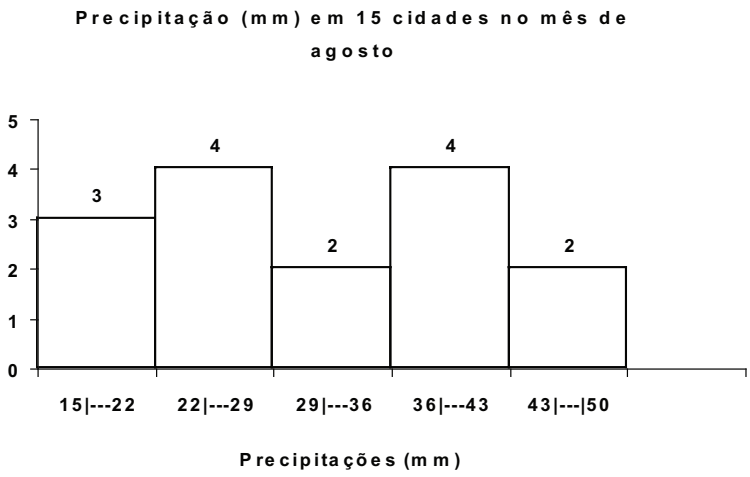
Precipitação(mm)	Número de cidades	%
15|---22	3	20
22|---29	4	26,67
29|---36	2	13,33
36|---43	4	26,67
43|---50	2	13,33
Total	15	100

Sobre o exercício ao lado:

Qual é a amostra estudada do exercício ao lado?
Resposta: 15 cidades.

Qual é a variável de estudo?
Resposta: Precipitações (mm).

c) Histograma



Obs.: Histogramas só devem ser utilizados para representar variáveis quantitativas agrupadas por classes.

1.5 Medidas de tendência central ou de posição

As medidas de tendência central são as formas mais comumente encontradas da estatística descritiva. Uma medida de tendência central de um conjunto de dados ou amostra pode fornecer uma boa indicação sobre as informações de uma população.

Média Aritmética simples (média): É a soma de todos os valores de uma variável dividida pelo número total de observações (não disposto em distribuição de freqüência).

Notação: μ → média populacional

$\bar{x}$ → média amostral

Fórmula: $\bar{x} = \frac{\sum x_i}{n}$, onde:

$\sum$ = somatório; x_i = variável em estudo;

n = tamanho da amostra.

Exemplo: Considere o número de amostras de água coletadas no período de 1 ano em 5 corpos d'água.

42 43 36 32 40

Determine:

a) Qual é a média?

Resposta: $\bar{x} = \frac{42 + 43 + 36 + 32 + 40}{5} = 38{,}6$

Média Ponderada: É usada para cálculos em que os valores dados têm pesos diferentes. É o método apropriado para distribuição de freqüência por ponto ou por intervalo.

Fórmula: $\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$, onde:

x_i é a variável em estudo;

f_i é a freqüência absoluta (repetições associadas a cada valor de x_i) e para a distribuição por intervalo usa-se:

$x_i = (I_i + L_i)/2$, onde:

x_i é a variável em estudo;

É importante que seja dada uma maior ênfase nos seguintes tópicos:
- Média;
- Mediana;
- Moda;
- Média da população;
- Erro amostral.

Para saber mais, acesse o site abaixo para ver uma animação sobre Média e Mediana: http://matefixe.blogspot.com/search/label/Estatística

Sobre o exercício ao lado:

Qual é a amostra?
Resposta: 5 corpos d'água.

Qual é a variável?
Resposta: Número de amostras de água coletadas.

I_i = limite inferior;
L_i = limite superior.

Exemplo: A tabela abaixo apresenta a distribuição do número de análises diárias realizadas por 79 funcionários de um determinado laboratório.

Número de análises	Número de funcionários	%
5	3	3,8
10	23	29,1
15	43	54,4
20	10	12,7
Total	79	100

Resposta:

A Média Ponderada pode ser apresentada para dados agrupados por pontos, como no exemplo ao lado.

Número de análises (x_i)	Número de funcionários (f_i)	%	$x_i.f_i$
5	3	3,8	**15**
10	23	29,1	**230**
15	43	54,4	**645**
20	10	12,7	**200**
Total	79	100	**1090**

Sobre o exercício ao lado:

Qual é a amostra?
Resposta: 79 funcionários.

Qual é a variável?
Resposta: Número de análises.

a) Calcule a média.

Resposta: $\bar{x} = \frac{15 + 230 + 645 + 200}{79} = 13{,}8$

Outra possibilidade é a utilização da **Média Ponderada para dados agrupados em classes.**

Exemplo: A tabela abaixo apresenta o tempo de duração (dias) para se realizar análises granulométricas de solos em 28 laboratórios credenciados em todo o Brasil.

Tempo (dias)	N° de análises	%
4 |-- 6	20	**71,4**
6 |-- 8	3	**10,7**
8 |-- 10	5	**17,9**
Total	28	**100**

Resposta:

Tempo (dias) (x_i)	N° de análises(f_i)	%	$x_i f_i$
4 |-- 6	20	71,4	**100**
6 |-- 8	3	10,7	**21**
8 |-- 10	5	17,9	**45**
Total	28	100	**166**

Sobre o exercício ao lado:

Qual é a amostra?
Resposta: 28 análises granulométricas.

Qual é a variável?
Resposta: Tempo em dias.

a) Calcule a média.

Resposta: $\bar{x} = \frac{100 + 21 + 45}{28} = 5{,}9$

Mediana: É uma medida de posição; encontra-se exatamente no centro de um conjunto de dados.

Notação: "md"

Acesse o *site* abaixo e saiba mais sobre Mediana:
http://pt.wikipedia.org/wiki/Mediana_(estatística)

Para calcular, pode-se seguir a ordem abaixo:

1º) Ordenar o conjunto de dados em ordem crescente;

2º) Encontrar a posição da mediana:

se "n" for par → P.md = n/2

se "n" for impar → P.md = (n + 1) / 2

Exemplo: Considere o Carbono Orgânico Total (COT) em g/kg de 6 amostras de solos: 89 89 90 92 100 120. Calcule a mediana (md).

Resposta: n=6, logo P.md = n/2 = 6/2 = 3

Pega-se a 3ª e 4ª posições: md= (90+92)/2 = 91 g/kg

Moda: é o valor que ocorre com maior freqüência entre os dados.

Veja mais detalhes sobre Moda no *link* abaixo:
http://pt.wikipedia.org/wiki/Moda_(estatística)

Moda e mediana para dados agrupados por classes (ou intervalo)

Para o cálculo da Moda pode-se utilizar o Método de Czuber que é considerado o método mais preciso.

$M_o = l_i + c\frac{f_{mo} - f_{ant}}{2.f_{mo} - (f_{ant} + f_{post})}$, onde:

l_i = Limite inferior da classe modal; c = Tamanho do intervalo de classe; f_{mo} = Freqüência absoluta da classe modal; f_{ant} = Freqüência absoluta anterior à classe modal; f_{post} = Freqüência absoluta posterior à classe modal.

Para o cálculo da mediana pode-se seguir os seguintes passos:

1º) Determina-se as freqüências acumuladas;

2º) Calcula-se o $P_{md} = \frac{n}{2}$;

3º) Marca-se a classe correspondente à freqüência acumulada imediatamente superior a $\frac{n}{2}$ classe mediana e, em seguida, emprega-se a fórmula:

$M_d = l_i + c\frac{\frac{n}{2} - F_{ant}}{f_{Md}}$, onde:

l_i = Limite inferior da classe mediana;

c = Tamanho do intervalo de classe;

$\frac{n}{2}$ = posição da mediana, onde n é o tamanho da amostra;

f_{ant} = Freqüência acumulada anterior à classe mediana;

f_{Md} = Freqüência absoluta da classe mediana.

Exemplo: As análises de pH de 100 amostras de efluentes industriais de uma determinada empresa obteve os seguintes resultados.

pH	número de amostras
2|---4	25
4|---6	35
6|---8	20
8|---10	15
10|---12	5
Total	100

a) Calcule a moda e a mediana.

Resposta:

pH	número de amostras	F_i
2|---4	25	**25**
4|---6	35	**60**
6|---8	20	**80**
8|---10	15	**95**
10|---12	5	**100**
Total	100	**360**

Determinação da classe modal (classe com maior freqüência absoluta): **4 |----6**

Na seqüência, calcula-se:

$$M_o = l_i + c\frac{f_{mo} - f_{ant}}{2.f_{mo} - (f_{ant} + f_{post})} = 4 + 2 \times \frac{35 - 25}{2 \times 35 - (25 + 20)} = 4{,}8$$

Para a determinação da mediana:

1º) Determina-se as freqüências acumuladas;

2º) Calcula-se $P_{md} = \frac{n}{2}$ = 50ª posição;

3º) Marca-se a classe correspondente à freqüência acumulada imediatamente superior a $\frac{n}{2}$ classe mediana (4 |----6) e, em seguida, emprega-se a fórmula:

$$M_d = l_i + c\frac{\frac{n}{2} - F_{ant}}{f_{Md}} = 4 + 2 \times \frac{50 - 25}{35} = 5{,}4$$

1.6 Gráficos circulares ou de Setores (Pizza)

É a representação gráfica da freqüência relativa (percentagem) de cada categoria da variável. Este gráfico é utilizado, preferencialmente, para representar variáveis **qualitativas**. Não se recomenda a sua utilização quando a variável assumir muitas categorias de respostas distintas.

A construção do gráfico de setores segue uma regra de três simples, onde as freqüências de cada classe correspondem ao ângulo que se deseja representar em relação à freqüência total que representa o total de 360º.

Obs.: Uma das melhores formas de se analisar e explorar os dados de uma pesquisa é através de técnicas gráficas.
Assim, sugere-se a utilização de outras ferramentas gráficas, tais como:
- Histograma de freqüências;
- Diagrama de *box plot*.

Exemplo: A tabela a seguir apresenta o número de peixes contaminados e não-contaminados com PCBs em um determinado lago.

Peixes	Amostras	%
Contaminados	32	38
Não-contaminados	52	62
Total	84	100

As suas principais características são:
- a área do gráfico equivale à totalidade de casos (360° = 100%);
- cada "fatia" representa a percentagem de cada categoria.

Resposta:

Para saber mais sobre os tipos de gráficos utilizados em estatística, veja o vídeo postado no *site* abaixo:

http://br.youtube.com/watch?v=ZRnwhwfsK_w&feature=related

1.7 Variação ou dispersão de distribuições

Apenas as medidas de tendência central são insuficientes para representar de forma adequada todos os conjuntos de dados, pois não são capazes de revelar a sua variabilidade.

Embora a amplitude total seja de fácil solução, esta medida é muito limitada, pois só utiliza os valores extremos, independentemente da forma que se distribuem os demais valores.

Amplitude total (A_t): A amplitude total é a diferença entre o maior e o menor valor de um conjunto de dados.

$$A_t = V_{máx} - V_{mín}$$

Exemplo: Considere a concentração de poluentes em 10 amostras de um determinado rio.

15 20 17 30 40 35 23 37 17 20

Calcule a amplitude total para as 10 amostras.

Resposta: $V_{máx} = 37$ e $V_{mín} = 15$

$$A_t = V_{máx} - V_{mín} = 37 - 15 = 22$$

Amplitude total para dados agrupados por classes (ou intervalo)

Neste caso, a amplitude total é a diferença entre o limite superior da última classe e o limite inferior da primeira classe, isto é:

$A_t = L_i - l_i$

Geralmente faz-se uso da amplitude total quando se quer determinar a amplitude da temperatura em um dia ou de um determinado ano, no controle de qualidade de um processo ou como uma medida de cálculo rápido.

Exemplo: Foram analisadas 30 amostras do rio Paranapanema, em diferentes pontos de coleta, para avaliar a concentração de um tipo de poluente. Os resultados foram agrupados em uma tabela de distribuição de freqüências por classes, como a apresentada a seguir.

Concentração	Número de amostras
10|---20	15
20|---30	10
30|---40	2
40|---50	2
50|---60	1

a) Determine a amplitude.

Resposta: $A_t = 60 - 10 = 50$

Variância Absoluta

Notação: S^2

Fórmula: $S^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$, onde:

x_i = variável;

$\bar{x}$ = média;

n = tamanho da amostra.

Exemplo: Os dados abaixo representam as concentrações de alumínio (mg/kg) de amostras de 4 solos.

400 350 450 300

Sobre o exercício ao lado:

Qual é a amostra?
Resposta: 4 solos.

Qual é a variável?
Resposta: Concentração de alumínio.

a) Calcule a variância.

Resposta:

$$S^2 = \frac{(400-375)^2 + (350-375)^2 + (450-375)^2 + (300-375)^2}{4-1} = 4166$$

1.8 Desvio-padrão

Para saber mais, acesse o *link* abaixo e veja um vídeo sobre Média e Desvio-padrão: http://br.youtube.com/watch?v=8X9apoqlbgs

Notação: S

Fórmula: $S = \sqrt{S^2}$ (raiz quadrada da variância)

Exemplo: Utilizando os dados do exemplo anterior, calcule o desvio-padrão da amostra.

Resposta: $S = \sqrt{S^2} = 64{,}55$

Portanto, a concentração média de alumínio é de 375 mg/kg com uma variação de 64,55.

Quanto ao desvio-padrão para dados agrupados, utiliza-se a mesma fórmula apresentada anteriormente ($S = \sqrt{S^2}$).

1.9 Variância e desvio-padrão para dados agrupados

Para dados agrupados, calcula-se a variância com a seguinte equação:

$$S^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{n - 1}$$

Exemplo: A tabela a seguir apresenta o número de microorganismos encontrados em amostras de comida industrializada.

Número de microorganismos	Número de indústrias	%
5	4	33,3
6	5	41,7
7	3	25
Total	12	100

a) Calcule a média e o desvio padrão.

Resposta:

Número de microorganismos (x_i)	Número de indústrias (f_i)	%	$x_i f_i$	$(x_i - \bar{x})2.f_i$
5	4	33,3	20	3,2
6	5	41,7	30	0,05
7	3	25	21	3,6
Total	12	100	71	6,85

$$\bar{x} = \frac{71}{12} = 5{,}9; \qquad S^2 = \frac{6{,}85}{11} = 0{,}62; \qquad S = \sqrt{0{,}62} = 0{,}8$$

1.10 Coeficiente de variação (CV)

É a relação percentual que o desvio-padrão tem sobre a média

Fórmula: $CV = \frac{S}{\bar{x}} \times 100$, onde:

S = Desvio-padrão; $\bar{x}$ = média.

Exemplo: Utilizando os dados do exemplo anterior, calcule o Coeficiente de Variação (CV).

Acesse o *site* abaixo e veja mais sobre Coeficiente de Variação: http://pt.wikipedia.org/wiki/Coeficiente_de_variação

Resposta: $CV = \frac{0{,}8}{5{,}9} \times 100 = 13{,}56\%$

1.11 Variável reduzida ou padronizada

Mede a magnitude do desvio-padrão em relação à média (dados atípicos Z>3).

Fórmula: $Z = \frac{x_i - \bar{x}}{S}$, onde:

Z = número de desvios (S) a contar da média;
S = Desvio-padrão;
x_i = variável em estudo;
$\bar{x}$ = média.

Exemplo: Com os dados do exemplo anterior, calcule a variável reduzida (Z).

Para praticar e verificar se você compreendeu todos os conceitos, refaça todos os exercícios resolvidos desta Unidade.

Resposta: $Z_1 = \frac{5-5{,}9}{0{,}8} = -1{,}125$ desvios

$Z_2 = \frac{6-5{,}9}{0{,}8} = 0{,}125$ desvios e $Z_3 = \frac{7-5{,}9}{0{,}8} = 1{,}3755$ desvios

UNIDADE 2 – PROBABILIDADE, AMOSTRAGEM E DISTRIBUIÇÃO

2.1 Objetivo de aprendizagem

Compreender o processo de análise de dados por meio de estudos estatísticos capazes de generalizar e obter conclusões sobre uma determinada população.

Para se entender os estudos estatísticos, é necessário que haja uma boa compreensão do conceito de probabilidade. Assim, deve-se dar uma maior ênfase nos tópicos abaixo:
- probabilidades condicionadas;
- aplicabilidade da probabilidade.

2.2 Probabilidade

Existe no nosso cotidiano uma série de situações de incerteza, das quais, embora não saibamos efetivamente o que vai ocorrer, pode-se listar os resultados possíveis e suas respectivas probabilidades. Costumamos chamar estas situações de incerteza de fenômenos aleatórios.

Experimento aleatório: É qualquer fenômeno aleatório que possa ser executado pelo homem.

Espaço amostral (S): É o conjunto de todos os possíveis resultados de um experimento.

2.3 Estudos de Probabilidade

São as seguintes as propriedades que devem ser respeitadas:
0 % e,
$P(\overline{A})$ = 100% - P(A), onde:
$P(\overline{A})$ = Probabilidade de não acontecer.

$$P(A) = \frac{\text{Número de eventos favoráveis a "A"}}{\text{Total de possíveis resultados}}$$

Exemplo: Lançamento de um dado com S = {1, 2, 3, 4, 5, 6}

- P(1) = $\frac{1}{6}$ = 0,17 = 17% (Probabilidade de sair o número 1)
- P(6) = $\frac{1}{6}$ = 0,17 = 17% (Probabilidade de sair o número 6)
- P(face par) = = 0,5 = 50% (Probabilidade de sair uma face par)
- P(múltiplo de 3) = = 0,33 = 33% (Probabilidade de sair um múltiplo de 3)

Acesse o *site* abaixo e assista ao vídeo sobre probabilidade que utiliza moedas como exemplo.
http://br.youtube.com/watch?v=vPDI8Gz4saY&feature=related

Operações com probabilidade: são os estudos mais comumente encontrados e podem ser descritos como dois eventos independentes (A e B):

$$P(A \text{ e } B) = P(A) \times P(B)$$

$$P(A \text{ ou } B) = P(A) + P(B)$$

Exemplo: A probabilidade de um homem ter artrite daqui a 30 anos é 30% e de sua mulher 40%. Qual é a probabilidade de que daqui a 30 anos:

a) Ambos tenham artrite?

Resposta: P(H e M) = 0,3 x 0,4 = 0,12 = 12%

b) Somente a mulher tenha artrite?

Resposta: P($\bar{H}$ e M) = 0,7 x 0,4 = 0,28 = 28%

c) Ambos não tenham artrite?

Resposta: P($\bar{H}$ e $\bar{M}$) = 0,7 x 0,6 = 0,42 = 42 %

d) Pelo menos 1 deles tenha artrite?

Resposta: P(H e M) ou P(H e $\bar{M}$) ou P($\bar{H}$ e M)

0,12 + 0,18 + 0,28 = 0,58 = 58%

Probabilidade condicional: A probabilidade é dita condicional quando a probabilidade de um evento depende da condição em que ele está sendo considerado.

Exemplo: Faça os estudos de probabilidade para os dados referentes ao sexo *versus* aquisição de planos de saúde.

Sexo	Plano de Saúde (p.s.)		Total
	Sim	Não	
Masculino	240	414	654
Feminino	323	100	423
Total	563	514	1077

a) Qual é a probabilidade de se ter plano de saúde (p.s.)?

Resposta: P(p.s.) = $\frac{563}{1077}$ = 0,52 = 52%

b) Qual é a probabilidade de um homem ter plano de saúde?

Resposta: P(p.s./H) = $\frac{240}{654}$ = 0,37 = 37%

c) Qual é a probabilidade de uma mulher ter plano de saúde?

Resposta: P(p.s./M) = $\frac{323}{423}$ = 0,76 = 76 %

2.4 Distribuição binomial

Utiliza-se o termo binomial para designar situações em que os resultados de uma variável aleatória possam ser agrupados em:

"Sucesso" ou "Fracasso" que ocorrem independentemente.

Modelo binomial

$$P(X = x) = C_n^x . p^x . (1 - p)^{n-x} \quad , \textit{ onde:}$$

$$C_n^x = \frac{n!}{x!(n - x)!}$$

A distribuição binomial é útil para determinar a probabilidade de um certo número de sucessos num conjunto de observações e segue estas determinações:
- o experimento consiste em n tentativas em iguais condições;
- cada tentativa tem dois possíveis resultados Sucesso ou Fracasso;
- as probabilidades de Sucesso (p) e Fracasso (1-p) permanecem constantes.

Exemplo: A probabilidade de erros em leituras de coliformes fecais é de 10% em 5 placas selecionadas para contagem; qual a probabilidade de:

a) 1 leitura estar errada?

Resposta: P(x = 1) = $C_5^1 . 0,1^1 . 0,9^4$ = 0,33 = 33%

b) 3 leituras estarem erradas?

Resposta: P(x = 3) = $C_5^1 . 0,1^1 . 0,9^4$ = 0,008 = 0,8%

c) Todas as leituras estarem erradas?

Resposta: P(x = 5) = $C_5^1 . 0,1^1 . 0,9^4$ = 0,00001 = 0,001%

d) Todas as leituras estarem corretas?

Resposta: P(x = 0) = $C_5^1 . 0,1^1 . 0,9^4$ = 0,59 = 59 %

Veja como utilizar uma planilha de Excel para cálculos de uma distribuição binomial no *link* abaixo:
http://br.youtube.com/watch?v=kbbP_sxclNE

2.5 Distribuição de Poisson

A distribuição de Poisson descreve as probabilidades do número de ocorrências num campo ou intervalo contínuo (em geral espaço ou tempo).

Deve-se admitir que exista uma taxa média constante dada em ocorrência por unidades chamada de λ (lambda).

Modelo Poisson: $P(X = x) = \frac{e^{-\lambda} . \lambda^x}{x!}$, *onde:*

$\lambda = n.p$

São exemplos de variáveis que seguem uma distribuição de Poisson: número de acidentes por dia, número de telefonemas por hora, número de defeitos por metro, etc.

Exemplo: Uma máquina produz 9 pipetas defeituosas a cada 1000 peças produzidas. Calcule a probabilidade de que em um lote com 200 pipetas ocorram:

a) 2 defeituosas.

Resposta: $P(x=2) = \frac{e^{-1,8}.1,8^2}{2!} = 0,26 = 26\ \%$

b) Nenhuma defeituosa.

Resposta: $P(x=0) = \frac{e^{-1,8}.1,8^2}{2!} = 0,16 = 16\ \%$

c) 1 defeituosa.

Resposta: $P(x=1) = \frac{e^{-1,8}.1,8^2}{2!} = 0,288 = 28,8\ \%$

2.6 Distribuição normal

É muito importante saber a forma com que os dados se distribuem, pois muitos testes estatísticos só são válidos se os dados se distribuem de uma determinada maneira.

A distribuição Normal é uma das distribuições mais importantes. Para que uma população possa ser assim classificada, esta deverá apresentar características de simetria em torno da média, as caudas devem encontrarem o eixo x no infinito, e a média, a mediana e a moda devem ser coincidentes.

Considere uma variável aleatória "x" com média μ e desvio padrão σ que apresenta as seguintes características:

- é simétrica em torno da média;
- prolonga-se de -∞ a +∞;
- sua área total é 100%;
- sua distribuição de probabilidade produz uma curva em forma de sino.

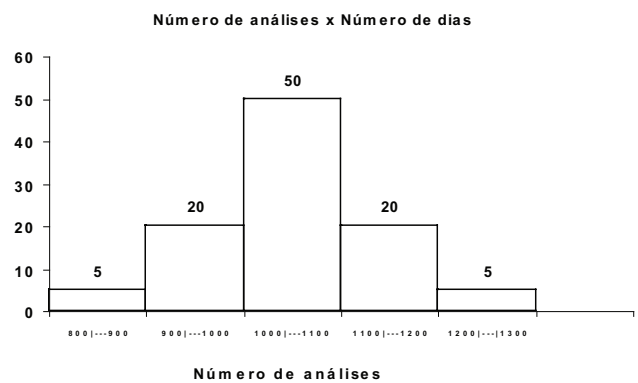


Figura 2.1 – Gráfico de colunas do número de dias versus número de análises que segue uma Distribuição Normal.

A área sob a curva entre a média e um ponto arbitrário (Figura 2.2) é função do número de desvios-padrão entre a média e aquele ponto.

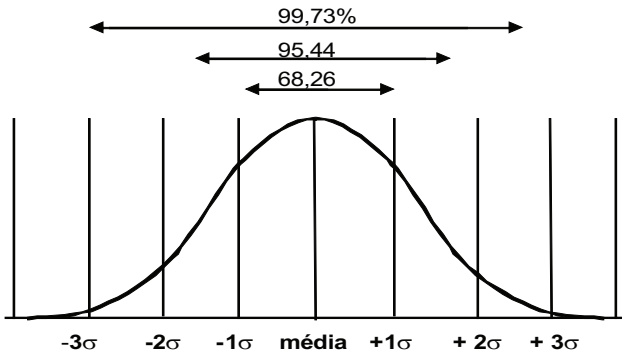


Figura 2.2 – Exemplo de uma curva de Distribuição Normal

Acesse ao link abaixo para ver um vídeo sobre como se forma uma curva Normal: http://br.youtube.com/watch?v=BMJGZuB1HCE

Para o cálculo da probabilidade na curva normal, usa-se:

$$z = \frac{x - \mu}{\sigma}$$

Onde:

z = número de desvios (σ) a contar da média;

x = valor arbitrado, diferente da média;

μ = média da distribuição normal;

σ = desvio-padrão.

Calculado o valor de z, acha-se na tabela (em anexo) a probabilidade de que ocorra certo resultado entre dois pontos (limites), como se pode observar na tabela abaixo.

Z	Área entre a média e Z	Z	Área
1,00	0,3413	1,65	0,4505
1,96	0,4750	2,00	0,4772

Exemplo: O consumo de água de uma comunidade tem distribuição Normal com média de 1000 litros e desvio-padrão de 200 litros. Calcule a probabilidade de um indivíduo desta comunidade ter um consumo:

a) Inferior a 800 litros.

Resposta: Z = -1, logo P(x<800) = 0,158655 = 15,86 %

b) Inferior a 1300 litros.

Resposta: Z=1,5, logo P(x<1300) = 0,933193 = 93,31%

c) Superior a 1100 litros.

Resposta: Z= 0,5, logo P(x < 1100) = 0,3085 = 30,85%

d) Entre 900 e 1200 litros.

Resposta: Z1 = -0,5 e Z2 = 1, logo P(900<x<1200) = 0,53 = 53 %

Distribuição normal padrão: Os tópicos anteriores abordaram a importância de se identificar a forma de distribuição e se esta é uma distribuição normal. A distribuição normal padrão é uma forma de distribuição normal com média igual a "0" e desvio-padrão igual a "1". Essa distribuição é muito importante, pois permite que se estabeleçam comparações entre valores de amostras diferentes e valores de uma mesma amostra.

Distribuição amostral: A distribuição amostral é calculada a partir da retirada do maior número possível de amostras de uma dada população.

Intervalos de confiança e erro-padrão: Os intervalos de confiança são estimativas intervalares como, por exemplo, em relação à média de uma amostra, na qual estes são capazes de fornecer um raio de valores (ou intervalo) em torno da média amostral. Este intervalo viabiliza a constatação com uma determinada confiança, se a média populacional está contida, ou não, em sua abrangência.

Distribuição de médias amostrais: é uma distribuição de probabilidade que indica o quanto prováveis são diversas médias amostrais; é uma distribuição em função da média, do desvio-padrão da população e do tamanho da amostra, ou seja, para cada combinação desses três elementos haverá uma única distribuição amostral de médias amostrais.

Segundo o Teorema do Limite Central, a média e, por conseqüência, a soma de n valores amostrais tende a seguir o modelo normal, independentemente da distribuição de origem que tais valores assumem. Portanto, a média das médias amostrais é igual à média dos valores individuais, e o desvio padrão das médias é menor do que o desvio padrão dos valores individuais na razão de $\frac{1}{\sqrt{n}}$

Média: $\bar{x} = \mu$ Desvio-padrão: S

As estimativas podem ser:
- Estimativa por ponto: dada por um único valor (parâmetro) populacional;
- Estimativa por intervalo: fornece um intervalo de possíveis valores, dentre os quais se admite que esteja o parâmetro populacional.

Estimação: é o processo que consiste em utilizar dados amostrais para estimar parâmetros populacionais desconhecidos.

Erro de estimação: para o caso da média, significa a diferença entre a média amostral e a verdadeira média populacional.

$$\bar{x} \pm \text{erro} \longrightarrow \bar{x} \pm z\frac{s}{\sqrt{n}}, \quad \text{onde:}$$

$\bar{x}$ = Média da amostra e, também, média estimada para a população;

z = Variável padronizada, confiança desejada em % (valores da tabela curva normal); s = Desvio-padrão; n = Tamanho da amostra.

2.7 Intervalo de Confiança (IC)

(IC) = ($\bar{x}$ – erro; $\bar{x}$ + erro)

Intervalos de confiança para estimação da média, quando o σ populacional é conhecido.

Exemplo: Cálculo de intervalo para: n = 36; S = 3; $\bar{x}$ = 24,2

Confiança desejada	Z(valores da Tabela)	Fórmula	Cálculo
90%	1,65	$\bar{x} \pm 1{,}65 \frac{s}{\sqrt{n}}$	$\bar{x} \pm 1{,}65 \frac{3}{\sqrt{36}}$
95%	1,96	$\bar{x} \pm 1{,}96 \frac{s}{\sqrt{n}}$	$\bar{x} \pm 1{,}96 \frac{3}{\sqrt{36}}$
99%	2,58	$\bar{x} \pm 2{,}58 \frac{s}{\sqrt{n}}$	$\bar{x} \pm 2{,}58 \frac{3}{\sqrt{36}}$

Continuação:

erro	intervalo
24,2 ± 0,825	23,375 a 25,025
24,2 ± 0,98	23,22 a 25,18
24,2 ± 1,29	23,11 a 25,69

2.8 Distribuição "t de Student" (n < 30 e/ou σ desconhecido)

Usa-se essa distribuição quando o desvio padrão da população S é desconhecido e a amostra é igual a 30 ou menor de 30 (pequenas amostras), desde que a população submetida à amostragem seja normal. Diferentemente da distribuição normal, que é padrão para qualquer amostra, há uma distribuição t para cada tamanho de amostra.

Obs.: Utiliza-se essa distribuição para pequenas amostras.

$\bar{x} \pm$ erro $\qquad\qquad \bar{x} \pm t\frac{s}{\sqrt{n}}$

Para encontrar t, usa-se α e g.l., onde:

α = nível de significância (1 – confiança)

g.l. = graus de liberdade = n - 1

Assim, quando a confiança desejada for 95%:

α = 1-0,95 e o g.l. = graus de liberdade = n - 1

Exemplo: Sabendo-se que uma amostra tem 25 unidades, que a sua média é 20 e que possui um desvio-padrão igual a 1,5; represente um intervalo de confiança ao nível de 90%, 95% e 99%.

Resposta: n = 25, portanto graus de liberdade = n − 1 = 24; S = 1,5; $\bar{x} = 20$

Procure uma pesquisa científica em que o autor faça referência a intervalos de confiança e refaça os cálculos com dados disponíveis.

Confiança desejada	t (ver tabela)	Fórmula
90%	1,711	$\bar{x} \pm t\frac{s}{\sqrt{n}}$
95%	2,064	$\bar{x} \pm t\frac{s}{\sqrt{n}}$
99%	2,797	$\bar{x} \pm t\frac{s}{\sqrt{n}}$

Cálculo	Intervalo
$\bar{x} \pm 1{,}711\frac{s}{\sqrt{n}}$	$20 \pm 0{,}5739$
$\bar{x} \pm 2{,}064\frac{s}{\sqrt{n}}$	$20 \pm 0{,}6922$
$\bar{x} \pm 2{,}797\frac{s}{\sqrt{n}}$	$20 \pm 0{,}9381$

Erro-padrão

Em alguns casos, é conveniente trabalhar com a média das médias amostrais. Nesse caso, o desvio-padrão de uma distribuição de médias ou de diferenças entre médias é também chamado de erro-padrão.

UNIDADE 3 – TESTE DE HIPÓTESES E SIGNIFICÂNCIA ESTATÍSTICA

3.1 Objetivo de aprendizagem

Conhecer como aplicar a probabilidade e distribuição amostral para realizar testes e hipóteses.

3.2 Teste de hipóteses

Ao se realizar um determinado estudo ou idealizar um experimento científico, deve-se ter objetivos ou metas bem estabelecidos por meio de afirmações sobre o que se deseja verificar. Essas afirmações provisórias são denominadas de hipóteses.

Com a obtenção de dados e a aplicação dessa análise estatística, é possível verificar se os resultados confirmam ou não essas hipóteses.

Designa-se por Ho, chamada hipótese nula, a hipótese estatística a ser testada, e por H1, a hipótese alternativa. A hipótese nula expressa uma igualdade, enquanto que a hipótese alternativa é dada por uma desigualdade (≠, <, >). Os testes utilizados podem ser observados nas Figuras 3.1, 3.2 e 3.3:

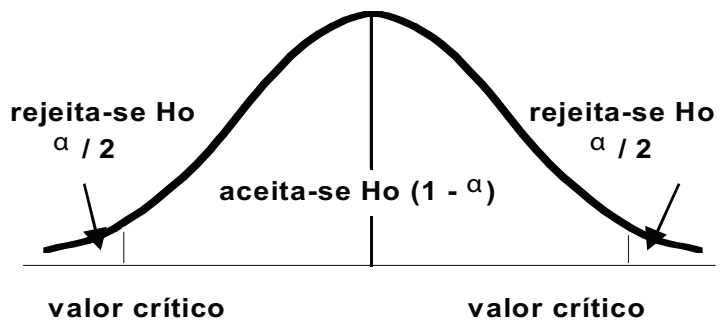


Figura 3.1 – Teste Bicaudal ou Bilateral

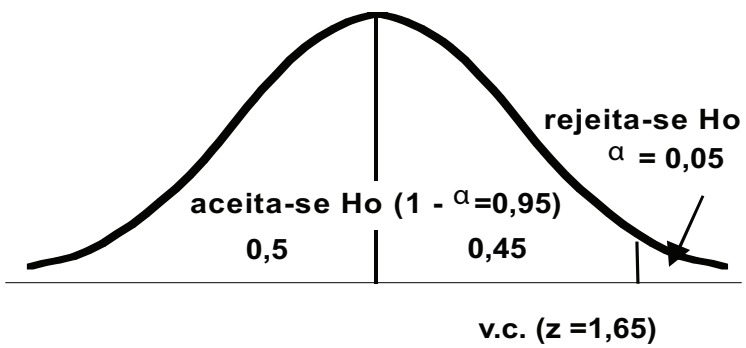


Figura 3.2 – Teste Unilateral Direito

Por que formular uma hipótese?

É interessante destacar que a maior parte dos estudos e experimentos que se iniciam sem um objetivo determinado tornam-se estéreis. Sem uma hipótese clara a nos impulsionar, guiar e a sugerir o que observar e anotar, talvez todo o esforço despendido durante o trabalho, ou pesquisa, resulte em nada.

Acesse o *link*, a seguir, para ver um exercício sobre testes de hipóteses:

http://br.youtube.com/watch?v=UsffYlsUOU8&feature=related_

É comum encontrar trabalhos científicos nos quais os pesquisadores relatam seus resultados como significativos ou não-significativos.

Complemente seus estudos pesquisando na Internet outros exercícios sobre Testes de Hipóteses.

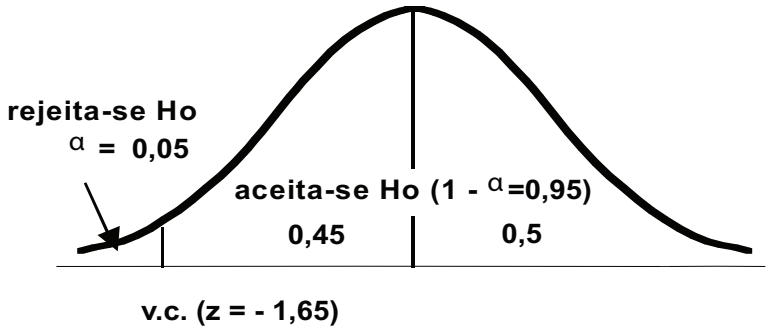


Figura 3.3 – Teste Unilateral Esquerdo

Assim, o risco de rejeitar incorretamente a hipótese nula é chamado de erro do Tipo I e tem uma probabilidade de ocorrer igual a α. No entanto, também é possível que se cometa o erro de aceitar a hipótese nula quando não se deveria e, portanto, cometer um erro do Tipo II com a probabilidade de ocorrer igual a β.

Nível de significância: Como os dados obtidos provêm de amostras, a decisão sobre aceitar ou não uma determinada hipótese está associada a uma probabilidade de erro e este pode ser controlado ou mensurado através do nível de significância (ou ponto de corte) utilizado durante os estudos.

Erro do Tipo I e do Tipo II: Durante os testes de hipóteses, existe a possibilidade de que se cometa um erro ao se considerar uma hipótese verdadeira quando, na verdade, ela não é. Nesse caso, pode-se concluir que tal efeito ocorrido deve-se apenas ao acaso.

UNIDADE 4 – CORRELAÇÕES BIVARIADAS

4.1 Objetivo de aprendizagem

Compreender como se analisa o relacionamento entre duas variáveis para verificar se existem correlações bivariadas.

Leia mais sobre correlação bivariada na apresentação de um *software* de estatística:

http://www2.dce.ua.pt/leies/pacgi/SPSS_21_03_07/sessao_2.pdf

4.2 Caracterizando correlações bivariadas

Considera-se uma correlação bivariada quando da ocorrência de relacionamento entre duas variáveis e, se as mesmas estiverem associadas, é usual dizer que estas são correlacionadas. A Figura 4.1, a seguir, apresenta o diagrama de dispersão para duas variáveis (X e Y).

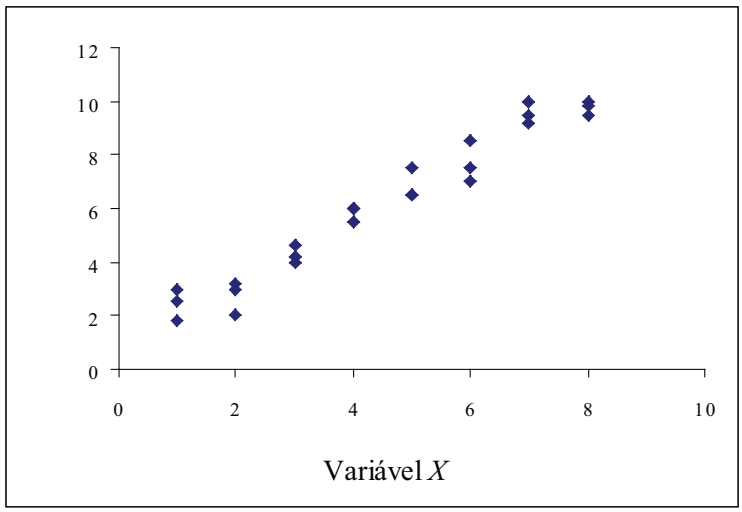


Figura 4.1 – Diagrama de dispersão para as variáveis X e Y

O coeficiente de correlação tem duas propriedades que caracterizam a natureza de uma relação entre duas variáveis. Uma é o seu sinal (+ ou -) e a outra é sua magnitude. O sinal é o mesmo do coeficiente angular de uma reta imaginária que se "ajustaria" aos dados se essa reta fosse traçada num diagrama de dispersão, e indica se esta reta é crescente (+), relacionamento positivo, ou decrescente (-), relacionamento negativo. A magnitude de R indica quão próximos da "reta" estão os pontos individuais. Por exemplo, valores de R próximos de –1 (correlação negativa perfeita) ou +1 (correlação positiva perfeita) indicam que os valores estão muito próximos da reta ou mesmo sobre a reta, enquanto que os valores mais próximos do 0 (zero) sugerem maior dispersão. Quando as variáveis caminham ora no mesmo sentido, ora em sentidos opostos, diz-se que não há correlação. A forma mais simples e intuitiva de verificar a existência de correlação entre duas variáveis é através do diagrama de dispersão.

Aplicabilidade das correlações bivariadas: A aplicação de análises de correlação permite descobrir se existe um relacionamento entre as variáveis em estudo. Entretanto, essa não é a única informação que essas análises oferecem, já que elas permitem determinar a direção do relacionamento e sua força ou magnitude.

Entre as direções de relacionamento pode-se citar os positivos perfeitos, os positivos imperfeitos, os negativos perfeitos (Figura 4.2), os negativos imperfeitos e os não-lineares.

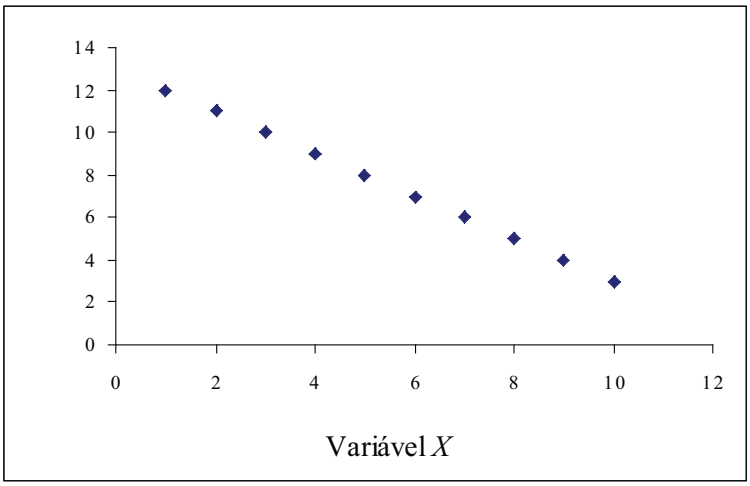


Figura 4.2 – Relacionamento negativo perfeito

4.3 Coeficiente de correlação de Pearson (r)

O coeficiente de correlação produto-momento (r) ou coeficiente de correlação de Pearson, assim designado porque sua fórmula foi proposta por Karl Pearson em 1896, é uma proporção entre a covariância das duas variáveis e uma medida das variáveis separadamente. Apresenta uma grande vantagem por ser um número puro, ou seja, independe da unidade de medida das variáveis (pode-se ter duas unidades de medida diferentes).

A Figura 4.3 apresenta um diagrama de dispersão para duas variáveis (X e Y) com problemas de homocedasticidade, isto é, presença de heterocedastidade.

Este tipo de análise é amplamente utilizado. Mas a utilização dessa análise de forma indiscriminada pode resultar em erros de interpretação e conduzir a conclusões equivocadas, como é o caso da violação da pressuposição de homocedasticidade (Figura 4.3).

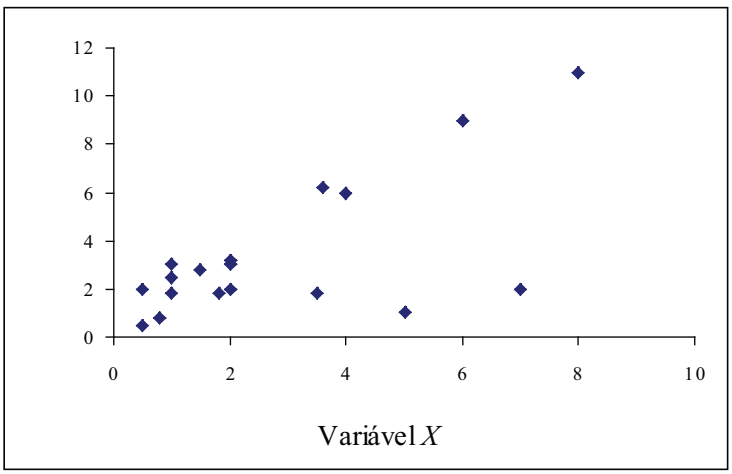


Figura 4.3 – Homocedasticidade

Cálculo do coeficiente de correlação (r):

$$r_{xy} = \frac{covariância\,(x,y)}{desvio\ \ padrão\ (x) \times desvio\ \ padrão\ (y)}$$

$$r_{xy} = \frac{\sum x.y - \frac{(\sum x).(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Variação do valor (r):

r_{xy} varia entre -1 e +1

Tente aplicar os estudos de correlação aos dados utilizados nas unidades anteriores.

UNIDADE 5 – REGRESSÃO LINEAR

5.1 Objetivo de aprendizagem

Entender como se obtém um modelo de relação entre duas variáveis quantitativas.

Utilidades da regressão linear:
- Estudar a existência de dependência de y em relação a x;
- Expressar matematicamente esta relação através de uma equação.

A sua principal utilidade é representar a dependência de uma variável quantitativa em relação a outra através de uma equação simples.

Obs.: Para que estes conceitos básicos e necessários para o prosseguimento do aprendizado sejam elucidados, sugere-se uma ênfase nos seguintes tópicos:
- Variável dependente (y) e variável independente (x);
- Obtenção da reta.

5.2 Regressão linear simples

A análise de regressão linear simples é uma extensão da análise de correlação e aplica-se para se obter uma relação de causa-efeito entre duas variáveis quantitativas que seja expressa matematicamente.

A regressão linear simples é um procedimento que fornece equações de linhas reta, por isso o termo linear, que descrevem fenômenos nos quais há apenas uma variável independente, por isso simples.

Podem-se prever valores para a variável dependente (y) em relação a valores não observados da variável independente (x). Isto é permitido dentro da faixa de valores estudados para x ou mesmo fora, desde que a extrapolação não seja exagerada, isto é, não haja um afastamento muito grande entre o valor de x desejado e o último (ou primeiro) valor de x estudado.

Linha ou reta de regressão linear: A análise de regressão linear permite a confecção de uma linha real (conforme Figura 5.1) que possibilita prever um valor de y a partir de x.

É muito comum, na prática, termos duas variáveis x e y, cujos valores se admitem relacionados.

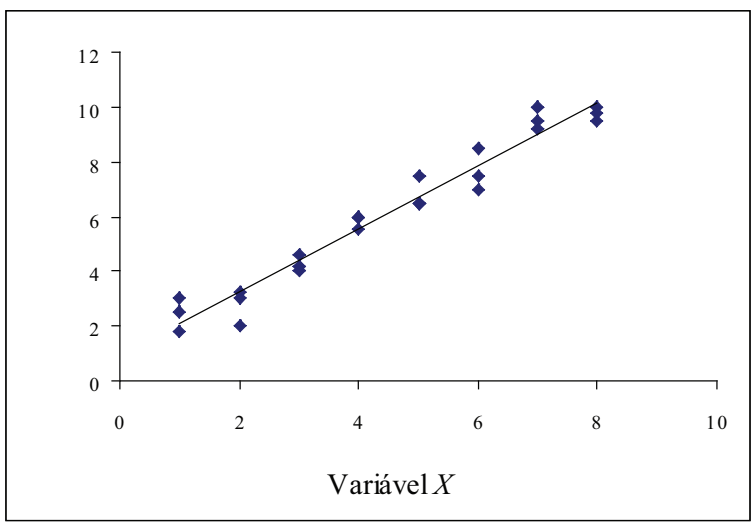


Figura 5.1 – Linha de regressão linear

Pode-se citar, como um exemplo, o espaçamento entre as árvores de um bosque. A relação entre a produção de lenha (y) e a área, a área ocupada por cada árvore (x) será uma função de x:

$$Y = f(x)$$

Se a relação entre as variáveis puder ser expressa por uma equação de primeiro grau, isto é: $y = a + bx$, onde:

x = variável independente;

y = variável dependente;

a = parâmetro ou coeficiente linear;

b = parâmetro ou coeficiente angular,

tem-se uma regressão linear simples.

5.2.1 Equação de regressão

Cálculo da equação de regressão linear e valores de a e b: A representação matemática de uma linha reta é dada por: y = a + bx.

Para facilitar o uso das análises de regressão linear, pode-se obter uma equação de regressão, conforme Figura 5.1, que possibilitará a realização de previsões conforme previamente apresentado pela reta de regressão linear.

Para se encontrar a equação y = a + bx, que descreve o relacionamento entre as variáveis x e y, temos que estimar os valores de a e b. Para isso, vamos aplicar o chamado método dos mínimos ou dos mínimos quadrados, que tem como objetivo tornar mínima a soma dos quadrados desvios.

Leia mais sobre Regressão Linear Simples no arquivo do *site*: http://www.inf.ufsc.br/~ogliari/arquivos/Analise_de_Regressao_linear_simples.ppt

Método dos mínimos quadrados

$$y = a + bx$$

$$\sum y = an + b\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2$$

Resolvendo o sistema determina-se a e b, conforme as equações a seguir:

$$b = \frac{n.\sum(x.y) - (\sum x).(\sum y)}{n.\sum(x^2) - (\sum x)^2}$$

$$a = \frac{\sum y - b.\sum x}{n}$$

onde:

n = número de observações realizadas;

= somatório dos produtos dos pares de valores observados;

soma dos valores observados de x;

soma dos valores observados de y;

somatório dos quadrados dos valores observados de x;

quadrado da soma dos valores observados de x.

Regressão Linear Múltipla

http://www.ufv.br/saeg/saeg43.htm

5.3 Regressão linear múltipla

Muitas vezes, uma variável y depende de um conjunto de outras variáveis (x_1, x_2, ..., x_k) independentes. Então, a relação entre as variáveis y e x_1, x_2 e x_k pode ser expressa por uma equação polinomial mostrada a seguir:

$$y = a_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k + \varepsilon$$

onde a_0 e b_1,b_2 e b_k são parâmetros a serem estimados a partir dos dados. Além dos valores de x_1, x_2 e x_k, consideramos que também depende de outros fatores, representados no modelo por ε, chamado de efeito aleatório. O modelo desta relação é denominado de Modelo de Regressão Linear Múltipla.

Os conceitos vistos anteriormente devem ser adequandamente generalizados, acrescidos da suposição de que as variáveis x_1, x_2 e x_k são independentes, isto é, a correlação entre elas deve ser baixa.

UNIDADE 6 – ANÁLISE FATORIAL

As técnicas estatísticas utilizadas para analisar um conjunto de variáveis $x_1, x_2, x_3, \ldots, x_k$ simultaneamente, compõem a chamada Estatística Multivariada.

As técnicas exploratórias multivariadas são orientadas por dois grandes princípios: a redução do número de variáveis e a descoberta de uma estrutura de associação entre as variáveis originais. O propósito da análise é substituir um conjunto de variáveis correlacionadas, por um conjunto de novas variáveis não correlacionadas. Essas novas variáveis representam combinações lineares das iniciais e estão ordenadas de maneira que suas variâncias decresçam da primeira à última. É com esta estratégia que se obtém uma redução do número de variáveis sem perda considerável de informação. Nesta unidade introduziremos os conceitos de duas técnicas: Componentes Principais e Análise Fatorial.

Componentes Principais é utilizada para analisar dados organizados em tabelas de freqüência, com o objetivo de identificar alguma correspondência entre linhas e colunas. É útil para analisar um conjunto de variáveis categorizadas.

A Análise Fatorial é utilizada para analisar as correlações entre as variáveis originais e encontrar padrões de associação, chamados de fatores, que identificam variáveis não observáveis na realidade pesquisada. Essa técnica é útil para analisar um conjunto de variáveis numéricas.

Para fazer uma análise utilizando essas técnicas, é necessário dispor de programas computacionais específicos.

Acesse o *link* abaixo para saber mais sobre Análise Fatorial e veja também os outros tópicos disponíveis nesta página:

http://www.ufv.br/saeg/saeg43.htm

REFERÊNCIAS

CALLEGARI-JACQUES, S. M. **Bioestatística:** princípios e aplicações. Porto Alegre: Artmed, 2004. 255p.

CASTRO, Lauro Sodré Viveiros. **Exercícios de Estatística.** Rio de Janeiro: Científica,1994.

CRESPO, A. **Estatística Fácil.** 6. ed. São Paulo: Saraiava, 1989.

DAL MOLIN, Beatriz Helena et al. **Mapa Referencial para Construção de Material Didático - Programa e-Tec Brasil.** 2ª ed. revisada. Florianópolis: Universidade Federal de Santa Catarina - UFSC, 2008.

DORIA, U. **Introdução à Estatística.** São Paulo: Negócio Ed., 1999.

DOWNING, D.; CLARK, J. **Estatística Aplicada.** São Paulo: Saraiva, 1999.

GRANER, E. A. **Estatística:** bases para o seu emprego na experimentação agronômica e em outros problemas biológicos. 2.ed. Sao Paulo: Melhoramentos, 1966. 184p.

HEATH, O. V. S. A **Estatística na Pesquisa Científica.** São Paulo: EDUSP, 1981. 95p.

HOEL, P. G. **Estatística Elementar.** 4.ed. Rio de Janeiro: Fundo da Cultura, 1972. 311p.

LOPES, A. **Probabilidade Estatística.** Rio de Janeiro: Reichman, 1999.

MEYER, P. L. **Probabilidade:** Aplicações à Estatística**.** Rio de Janeiro: Ao Livro Técnico S.A./EDUSP, 1969. 391p.

PARADINE, C. G.; RIVETT, B. H. P. **Métodos Estatísticos para Tecnologistas.** São Paulo: EDUSP, 1974. 350p.

REIDY, J.; DANECY, C. **Estatística sem Matemática usando SPSS para Windows.** Porto Alegre: Artmed, 2006.

VICENT, W. **Statistic in Kinesiology.** Champaing: Human Kineties, 1999.

YouTube. Disponível em: <http://br.youtube.com>. Acesso em: 09 set. 2008.

Wikipédia. Disponível em: <http://pt.wikipedia.org/wiki/P%C3%A1gina_principal>. Acesso em: 09 set. 2008.

GLOSSÁRIO

Amplie o Glossário e, conseqüentemente, o seu vocabulário técnico, com pesquisas na Internet e baixando o arquivo do "Vocabulário Básico de Recursos Naturais e Meio Ambiente" no *site* do IBGE:

http://www.ibge.gov.br/home/geociencias/recursosnaturais/vocabulario.shtm?c=13

DBO: Demanda Bioquímica de Oxigênio

PCBs: Bifenilas Policloradas

COT: Carbono orgânico total

pH: potencial hidrogeniônico é o índice que indica a acidez, a neutralidade ou a alcalinidade de um meio.

CURRÍCULO SINTÉTICO DO PROFESSOR-AUTOR

Cristiano Poleto possui graduação em Engenharia Civil (1996) e especialização em Engenharia de Segurança do Trabalho pela Universidade Estadual de Maringá (2001), mestrado em Engenharia Civil com ênfase em Recursos Hídricos e Tecnologias Ambientais pela Universidade Estadual Paulista Júlio de Mesquita Filho (2003) e doutorado pela Universidade Federal do Rio Grande do Sul em Recursos Hídricos e Saneamento Ambiental (2007). Tem experiência nas áreas de Meio Ambiente, Engenharia Sanitária e Recursos Hídricos, atuando principalmente nos seguintes temas: qualidade da água, bacias hidrográficas urbanas, sedimentos urbanos e qualidade dos sedimentos. É docente na Universidade Federal do Rio Grande do Sul em cursos técnicos e na pós-graduação em Recursos Hídricos e Saneamento Ambiental. Tem experiência na organização de cursos de extensão e eventos científicos. É autor de trabalhos científicos publicados em jornais e revistas nacionais e internacionais, e de três livros na área de sedimentos e meio ambiente.

ISBN 85-7519-189-6