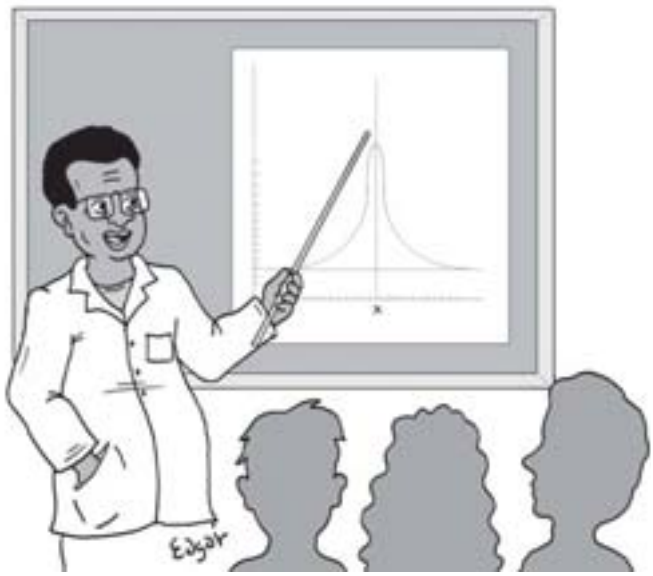
DISTRIBUIÇÃO NORMAL - PARTE I

4 aula

META
Apresentar o conteúdo de distribuição normal

OBJETIVOS
Ao final desta aula, o aluno deverá:
determinar a média e a variância para uma função contínua;
padronizar uma variável;
determinar intervalos de confiança.

PRÉ-REQUISITOS
O aluno deverá saber interpretar os histogramas como representação da distribuição de valores de uma amostra.



INTRODUÇÃO

Na aula anterior, vimos que quando precisamos obter informações sobre uma população, referentes a um determinado dado que varia dentro dessa população, podemos manipular dados de uma amostra representativa. Vimos também que um número relativamente grande de dados pode ser analisado com muita clareza através de uma representação gráfica conhecida como histograma. Nesta aula, veremos que as características básicas da distribuição representadas pelo histograma, ou seja, sua localização no eixo horizontal (faixa de valores possíveis) e sua dispersão dentro desta faixa, podem ser abreviadas por várias grandezas ou parâmetros estatísticos. Dentre essas grandezas, as mais estudadas nas ciências físicas são a média aritmética e o desvio padrão.

MÉDIA ARITMÉTICA

A média aritmética de um conjunto de dados, que indica sua localização ou tendência central no histograma, é simplesmente a soma de todos os valores divididos pelo número total de elementos do conjunto. Este é o conceito de média que a maioria das pessoas conhece, mesmo intuitivamente. Utilizamos aqui este mesmo conceito e, muitas vezes, simplesmente nos referimos a ele como média, dispensando a qualificação de média aritmética, pois caso seja necessário definir outro tipo de média, as definições serão destacadas apropriadamente.

Resumindo a forma matemática destes parâmetros temos:

média (normalmente traz uma barra sobre a letra x):

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

x_i = i-ésima observação

N = Número total de observações na amostra

Lembre-se de que na nossa amostra de 140 alturas, a média foi de 1,70 m. Se, no caso, obtivemos essa média a partir de uma amostra de uma população, trata-se da média amostral e não da média populacional.

Para expressar a dispersão dos valores medidos em torno da média, calculamos o parâmetro chamado de desvio padrão. Para calcular o desvio padrão, primeiro calculamos a diferença, ou simplesmente desvio, de cada valor individual em relação à média de todos os valores da amostra.

Desvio – é uma medida do afastamento de cada ponto em relação à média.

$$d_i = x_i - \bar{x}$$

Em seguida, somamos os quadrados de todos os desvios e dividimos o total por N-1. O resultado é a variância, que é uma medida do espalhamento das observações em torno da média.

Variância s^2:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

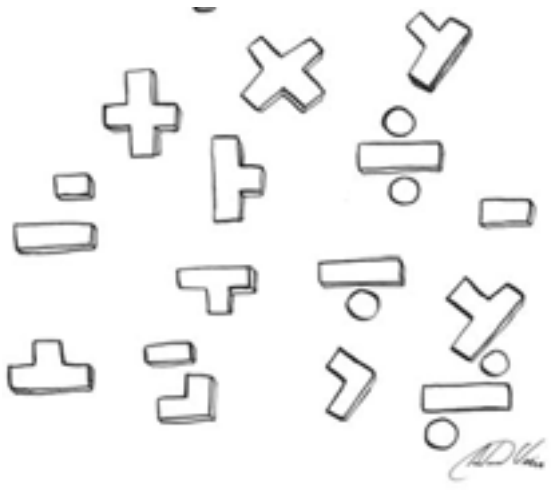
Note que a variância é uma espécie de média dos quadrados dos desvios.

Enquanto a média tem as mesmas unidades das medidas originais, as unidades da variância são, pela própria definição, o quadrado das unidades originais. Para que as medidas de dispersão e de posição tenham as mesmas unidades, costuma-se substituir a variância pela sua raiz quadrada, que é chamada de desvio padrão.

Desvio padrão – também é uma medida do espalhamento entre os diversos valores obtidos .

$$s = \sqrt{s^2}$$

O desvio padrão é geralmente utilizado para definir intervalos em torno da média. Na nossa amostra, o desvio padrão foi de 0,03641. O intervalo definido por um desvio padrão em torno da média tem como limites, portanto, 1,66596 e 1,73878. Obtemos o limite inferior, 1,66596, subtraindo o desvio padrão 0,03641 da média 1,70237; e o limite superior, somando o desvio padrão ao valor da média. A região compreendida entre estes dois valores corresponde a 66,6% da área total do histograma ou 2/3 de todos os valores. Já a região definida por dois desvios padrão tem como limites 1,62955 e 1,77519 e contém 96,8% da área total. A boa notícia é que você não vai precisar calcular todas estas somatórias, pois qualquer calculadora científica já vem programada para calcular estes parâmetros estatísticos ao



toque de uma tecla. Como tarefa, descubra como fazer isto usando sua calculadora preferida (use, para isto, o manual do fabricante).

Agora que definimos as grandezas estatísticas mais comuns, obtidas para a descrição estatística de amostras, voltemos um pouco nossa atenção para a forma do histograma obtido. Vimos que ele representa graficamente a distribuição de valores ao longo da amostra e que esses valores se distribuíram dentro de uma faixa. Nos limites desta faixa as freqüências foram baixas e aumentaram simetricamente até um certo valor, na região central. Nosso próximo passo é classificar os tipos de distribuição, partindo justamente do tipo observado em nossa amostra que, aliás, é a distribuição mais comum, e por isso recebeu a denominação de distribuição normal.

A DISTRIBUIÇÃO NORMAL

Vamo-nos concentrar na nossa amostra de pessoas que efetivamente tiveram sua altura medida e esquecer um pouco a população total. Nessa amostra, conhecemos a distribuição exata de valores, então vamos tratá-la como uma nova população e buscar um modelo matemático que descreva a distribuição de valores dentro desta população. Imaginemos que temos à nossa disposição um modelo que possa ser adequado para isto. O procedimento a ser adotado inicialmente é testar esse modelo, ou seja, verificar se ele realmente representa nossos dados de forma adequada. Em caso positivo, usamos esse modelo, caso contrário, procuramos um novo modelo.

Um dos modelos estatísticos mais importantes é a distribuição normal ou Gaussiana, que é uma distribuição de probabilidades de ocorrência de erros em medições, proposta no início do século XIX por Carl F. Gauss. Tantos foram – e são – os dados adequadamente descritos por ele que se chegou a pensar que os conjuntos de dados que não o seguissem estavam errados com relação ao modo como

foram medidos. É daí que vem o nome de distribuição normal. Hoje já se conhecem exceções à obediência da distribuição normal.

A distribuição normal é uma distribuição contínua, ou seja, uma distribuição em que a variável pode assumir qualquer valor dentro de uma faixa como, por exemplo, pesos de um legume qualquer. Por exemplo, podemos ter, dentro de um pacote de batatas, tubérculos pesando desde 100 a 650 g, com quaisquer outros valores possíveis dentro deste intervalo.

O agora, lembre-se agora novamente do nosso histograma. Nele há intervalos de valores que ocorreram mais do que outros (aqueles valores cujos blocos são mais "altos"). Por outro lado, os valores próximos às extremidades estão associados a blocos mais baixos, ou seja, ocorreram menos. Podemos associar esta freqüência relativa à probabilidade de ocorrência dos referidos valores, pois os valores que ocorrem mais são conseqüentemente mais prováveis.

Uma distribuição da variável contínua x é definida pela sua densidade de probabilidade f(x), que é uma expressão matemática que relaciona, no caso da distribuição normal, a probabilidade de ocorrência com parâmetros como a média e a variância populacionais. Não vamos deduzir a expressão específica a seguir, mas sim somente apresentá-la. Contudo, a dedução completa pode ser encontrada em livros de estatística.

Distribuição Normal:

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

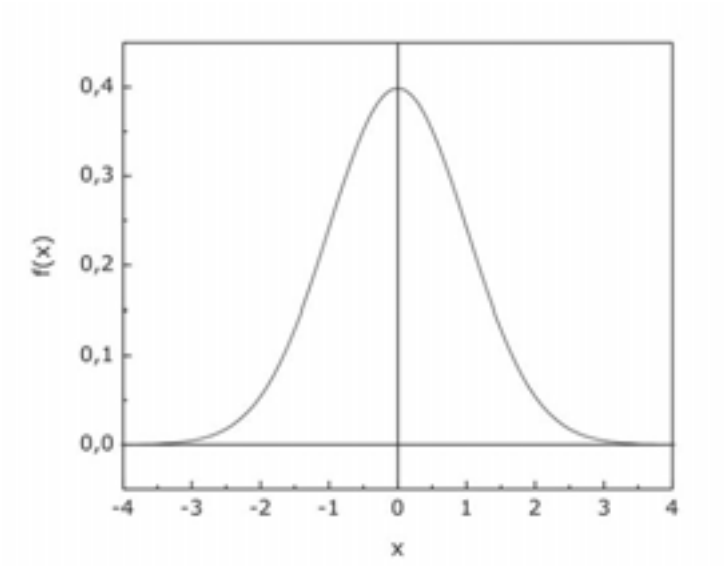
em que f(x) é a densidade de probabilidade associada a ocorrência de um certo valor;

f(x)dx é a probabilidade de ocorrência de um valor da variável no intervalo que vai de x a x+dx;

m é a média populacional;

s^2 é a variância populacional.

A equação anterior é especificamente a forma da função matemática que reproduz a curva de distribuição estatística da variável. O que chamamos de curva de distribuição estatística da variável ou distribuição de freqüência é uma curva como esta abaixo, que nos mostra a variação das probabilidades dos valores possíveis para a referida variável. Não se preocupe qual seria esta variável do exemplo, somente veja que a probabilidade é máxima na média, que é igual a zero neste exemplo.



Quando uma variável aleatória (com média dada por m e variância ó) se distribui de acordo com a distribuição normal, usamos a seguinte notação: x H" N(m,$ó^2$). O sinal H" aqui significa: "distribui-se de acordo com". No exemplo acima, x H" N(0,1) e veremos na próxima aula que na distribuição normal, praticamente toda a área da curva está contida dentro do intervalo entre ± 3 ó (desvios-padrão) em torno da média (3 acima e 3 abaixo). Por fim, o fato de termos os valores de m = 0 e ó = 1 permite que este exemplo possa ser classificado como uma situação especial de distribuição normal: a distribuição normal padronizada, que é o assunto da nossa próxima aula.

CONCLUSÃO

Como conclusão desta aula, podemos destacar que existe um tipo de distribuição estatística capaz de descrever um número tão grande de conjuntos de dados e populações diferentes que ficou conhecida como distribuição normal. A própria distribuição se relaciona com a probabilidade de ocorrência de valores e ambas podem ser representadas por expressões matemáticas.

(Fonte: http://amadeo.blog.com).

RESUMO

Quando precisamos obter informações sobre uma população, referentes a um determinado dado que varia dentro dela, podemos manipular dados de uma amostra representativa desta população e que um número relativamente grande de dados pode ser analisado com muita clareza através de uma representação gráfica conhecida como histograma. Nele, a ocorrência de valores é maior quanto mais próximos de um determinado valor, e vemos que os valores que apareceram vão de um certo valor mais baixo até um mais alto. Assim, começamos a ver que, olhando para o modo como os valores se distribuem, podemos começar a encontrar maneiras de descrever a nossa amostra: há um valor que, quanto mais próximo a ele, maior a ocorrência; todos os valores que foram observados estão dentro de um limite que vai do valor menor que foi observado até o maior; próximo destes limites há cada vez menos ocorrência. Estas características são próprias daquela população e podem ser usadas para qualificá-la. Aqui podemos introduzir os conceitos de média e desvios. O conceito de média que a maioria das pessoas conhece, mesmo intuitivamente, é simplesmente a soma de todos os valores divididos pelo número total de elementos do conjunto. Este conceito está correto. Os desvios dão uma medida do afastamento de cada ponto em relação à média. A forma do histograma que vimos também é muito importante. Nele há intervalos de valores que ocorreram mais do que outros (aqueles valores cujos blocos são mais "altos"). Por outro lado, os valores próximos às extremidades estão associados a blocos mais baixos, ou seja, ocorreram menos. Este comportamento é muito comum para tipos muito diferentes de dados, como por exemplo, dados científicos, demográficos etc., tão comum que foi batizado de distribuição normal.

A DISTRIBUIÇÃO NORMAL REALMENTE É MUITO UTILIZADA?

Para dar uma idéia do quanto a distribuição normal é considerada válida, ela é utilizada para se comparar as ocorrências médias de uma determinada doença, em uma dada população, em períodos anteriores àquele para o qual seja necessário identificar a possível ocorrência de uma epidemia. Estes são os chamados diagramas de controle e são utilizados para se estabelecer um intervalo de variação considerado normal.

Para saber mais, acesse a página: http://www.saude.sc.gov.br/gestores/sala_de_leitura/saude_e_cidadania/ed_07/pdf/09_04.pdf (Acessada em 07/02/2008).

PRÓXIMA AULA

Na próxima aula iremos aprender a usar a forma matemática da distribuição normal para calcular a probabilidade de que um certo valor ou intervalo de valores ocorram a partir da densidade de probabilidade.

REFERÊNCIAS

BARROS NETO, B.; SCARMINIO, I. E.; BRUNS. R. E.; **Planejamento e otimização de experimentos**. Editora da Unicamp, 1995.

BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. **Statistics for experimenters. An introduction to design, data analysis and model building**. New York, Wiley: 1978.

BUSSAB, W. O.; MORETIN, P. A. **Estatística básica**, São Paulo: Ed. Atual, 1985.