7 aula

COVARIÂNCIA E CORRELAÇÃO

META

Apresentar os conteúdos de covariância e correlação

OBJETIVOS

Ao final desta aula, o aluno deverá:
identificar variáveis correlacionadas;
representar graficamente as variações das variáveis;
calcular covariância e correlação;
visualizar através de exemplos e atividades as aplicações destes parâmetros.

PRÉ-REQUISITOS

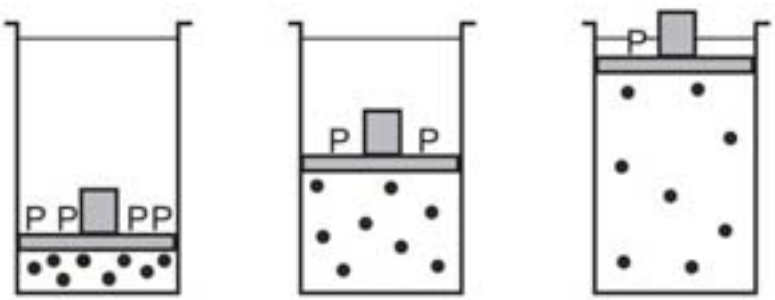
O aluno deverá compreender Teorema do limite central e Intervalo de confiança.

(Fonte: http://www.lumesoft.com).

(Fonte: http://cathywebster.com).

INTRODUÇÃO

Até agora, vínhamos falando de variáveis cujos valores se distribuem aleatoriamente dentro de uma população. Neste caso, vale lembrar que dizemos que os valores são independentes entre si, pois, por exemplo, se tirarmos uma azeitona de dentro de um pote cheio que acabamos de comprar e a pesarmos, o peso obtido não nos permite prever qual será o peso de uma segunda azeitona retirada aleatoriamente do mesmo pote. Por outro lado, se conseguirmos medir o volume das azeitonas (um pouco mais difícil de se medir), vamos notar que, de certa forma, quando uma variável aumenta, a outra também segue nesta mesma direção. Há variáveis que são correlacionadas de forma inversa, ou seja, quando uma aumenta, a outra tende a diminuir, como por exemplo, quanto maior a pressão de um gás (a temperatura constante e para a mesma quantidade de matéria), menor o volume ocupado. Portanto, a importância disto é que podemos, adiante, propor equações relacionando as variáveis e realizar previsões muito seguras.



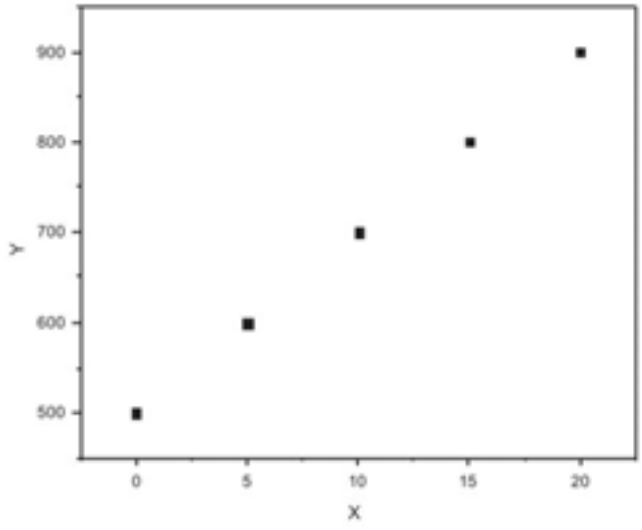
Lei de Boyle-Mariotte

Listamos abaixo outros exemplos de variáveis correlacionadas: a) freqüência de fumar e capacidade pulmonar; b) concentração de uma substância colorida em solução e intensidade de absorção de radiação eletromagnética na região do visível (luz); c) endimento de uma reação e temperatura.

VARIÁVEIS

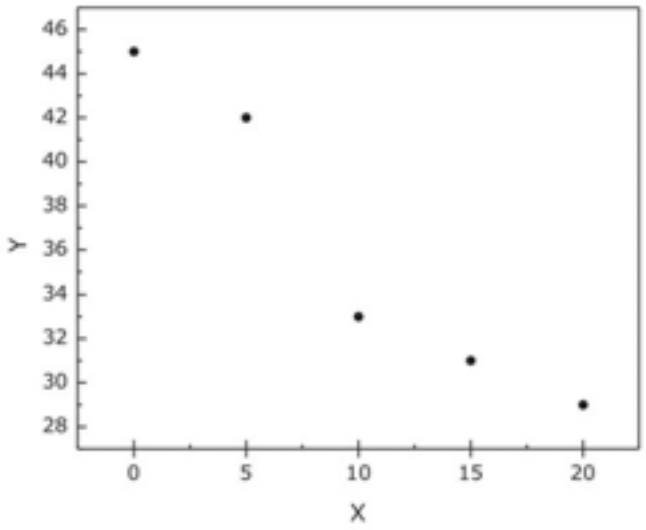
Outra situação em que dispomos de dados é a retração linear de peças cerâmicas, que ocorre durante o processo de aquecimento. Inicialmente, a argila é misturada a uma certa quantidade de água para a moldagem, seguida de aquecimento a temperaturas relativamente altas para formação das peças. Se medirmos o comprimento da peça antes da queima em uma certa direção, antes e depois do tratamento térmico, veremos que o comprimento diminui e é isso que chamamos de retração linear.

Ao lado, representamos como X a porcentagem de relação linear, inicialmente com Y sendo a temperatura. Vemos que quanto mais alta a temperatura, maior a porcentagem de retração linear. Durante o aquecimento, os grãos de argila sofrem um processo chamado de sinterização, quando os pontos em que suas superfícies se tocam acabam por se "soldarem". Os espaços vazios entre as partículas permanecem na forma de poros. Quanto mais alta a temperatura, maior é o grau de sinterização, o que reduz o volume dos poros, causando a retração da peça.



Por outro lado, representamos abaixo a porcentagem de retração linear como X, e como Y para o índice de absorção de água, que se relaciona com a porosidade do material. Esta medida é feita depois da peça aquecida a várias temperaturas, pesando-se a peça seca e

após ser imersa em água. Quanto mais poros a peça tiver, mais água irá absorver. Isto ocorre com as peças menos retraídas, pois estas têm mais poros.



Observações:

a) Se calcularmos o valor médio de cada uma delas, em cada caso os demais valores se distanciarão de suas médias de acordo com as respectivas variâncias (como já vimos). Como há uma correlação entre as variações dos valores das duas variáveis, em ambos os casos dizemos que as duas variáveis aleatórias apresentam uma certa covariância, ou seja, uma tendência a se desviarem de forma mais ou menos conjunta em relação às respectivas médias.

b) Em *a,* tendem a se desviar de forma oposta da média, enquanto em b, de forma parecida.

COVARIÂNCIA E CORRELAÇÃO

Expressamos a covariância, ou seja, a tendência das duas variáveis se desviarem de suas médias de forma conjunta, inicial-

mente representando o desvio de cada valor x_i ou y_i (correlacionados entre si) em relação às suas médias:

$$(x_i - \bar{x})$$

$$(y_i - \bar{y})$$

A medida numérica da tendência a variar conjuntamente é obtida como a média dos produtos dos dois desvios. Assim, a covariância amostral das variáveis x e y é:

$$Cov(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

em que:

(x_i, y_i): observações individuais.

$\bar{x}, \bar{y}$: médias amostrais

N: número de elementos na amostra.

Os valores que resultam da fórmula dependem dos valores medidos para as variáveis, ou seja, se calcularmos a covariância para um conjunto de valores de retração linear com temperatura, todos obtidos a temperaturas abaixo de 700 °C, vamos obter um certo valor. Se depois disso, para a mesma amostra, calcularmos a covariância entre retração linear e temperatura para um conjunto de dados obtidos acima de 700 °C, vamos obter outro valor, o que não tem muito sentido, pois estamos tratando de dados com a mesma natureza. Isto é simplesmente um problema de escala, cuja solução é obtida ao se fazer uma espécie de normalização para a covariância, recebendo agora o nome de coeficiente de correlação. Esta normalização envolve dividir cada desvio individual pelo desvio padrão da variável correspondente.

COEFICIENTE DE CORRELAÇÃO AMOSTRAL DAS VARIÁVEIS X E Y:

$$r(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

em que:

r(x,y): coeficiente de correlação entre as variáveis x e y.

s_x e s_y: desvios padrão das variáveis x e y, respectivamente.

PROPRIEDADES DA COV(X,Y)

Cov(x,x) = s_x^2: se usarmos os dados de somente uma das variáveis, iremos obter a própria variância desta variável. Isto nos dá a dimensão de que a covariância realmente representa a variância conjunta de variáveis correlacionadas.

PROPRIEDADES DO R(X,Y):

r(x,y) encontra-se sempre no intervalo [-1;+1], como resultado da normalização.

O coeficiente de correlação de variáveis independentes é zero. A recíproca não é verdadeira, ou seja, quando obtemos um coeficiente de correlação igual a zero, podemos descartar somente correlações lineares. Correlações com outras ordens (equação com variáveis quadradas etc.) não podem ser descartadas.

ATIVIDADES

Os valores a seguir correspondem aos pesos e volumes de azeitonas retiradas aleatoriamente de um pote. Calcule a covariância e o coeficiente de correlação entre os pesos e volumes.

M (g)	V (mL)
0,9504	1,08
2,1384	2,14
1,436	1,43
1,8952	1,95
1,4608	1,48
1,488	1,44
1,636	1,74

CONCLUSÃO

Podemos concluir a partir do que estudamos nesta aula, que muitas vezes estamos frente a variáveis que estão correlacionadas (a variação de uma delas acompanha inversamente ou diretamente a variação da outra). Estas correlações não são meras artificialidades, podendo nos dizer muito sobre a teoria dos processos que ocorrem mediante a variação destas variáveis. Por fim, conhecemos os parâmetros que podem ser utilizados para expressar a correlação entre variáveis: a covariância e o coeficiente de correlação.

RESUMO

Após verificarmos como é possível representar distribuições de valores que são independentes entre si, no sentido de que ao sabermos o valor de uma observação, nada podemos dizer a respeito do valor da observação seguinte, aprendemos, nesta aula, as vantagens que podem aparecer quando temos em mãos duas distribuições de valores que não são propriamente independentes entre si. Um exemplo disso é o conjunto dos pesos (uma população) e dos volumes (outra população) de um determinado vegetal (no caso, exemplificamos com azeitonas). Os valores dos pesos são independentes entre si, ocorrendo o mesmo com os valores dos volumes. Todavia, se examinamos se há uma correspondência geral entre valores de pesos e de volumes, veremos que há uma tendência de que pesos altos ocorram em conjunto com volumes altos, observando-se também que quando examinamos pesos baixos, encontramos volumes também baixos. Expressamos a covariância, ou seja, a tendência de as duas variáveis se desviarem de suas médias de forma conjunta, inicialmente representando o desvio de cada valor x_i ou y_i (correlacionados entre si) em relação às suas médias e, em seguida, a medida numérica da tendência a variar conjuntamente é obtida como a média dos produtos dos dois desvios. Na prática, a importância de se determinar variáveis correlacionadas é que estas observações fornecem informações importantes a respeito dos mecanismos que governam os processos que estamos estudando. É possível, com a construção de modelos matemáticos representando a dependência mútua entre as variáveis, propor equações de aplicabilidade geral e aprofundar teorias e modelos mecanísticos. Muito do que se sabe em ciência foi determinado a partir da observação das relações de proporcionalidade entre variáveis.

PRÓXIMA AULA

Na próxima aula, vamos comparar os parâmetros amostrais obtidos quando realizamos apenas uma amostragem numerosa e quando realizamos muitas amostras pequenas e tiramos a média de cada uma, fazendo a distribuição das médias.

REFERÊNCIAS

BARROS NETO, B.; SCARMINIO, I. E.; BRUNS. R. E. **Planejamento e otimização de experimentos**. Campinas: 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: Atual, 1985.