INSTITUTO TECNOLÓGICO DE AERONÁUTICA
DIVISÃO DE ENGENHARIA MECÂNICA

MOQ-14: Projeto e Análise de Experimentos
1^o Semestre/2014
Professor(a): *Denise Beatriz Ferrari*
http://www.mec.ita.br/∼denise
Email: *denise@ita.br*

Covariância e Correlação

Começaremos pelo estudo das relações de apenas um par de variáveis, ou seja, uma variável de resposta (Y) e uma explicativa (X).

1 Tipos de relação

Precisamos, inicialmente, fazer distinção entre dois tipos de relação:

Relação Determinística (ou funcional): É expressa por uma fórmula matemática, que representa uma relação exata entre as variáveis, ou seja, a variável explicativa é *completamente informativa* a respeito da variável de resposta.

$$Y = f(X)$$

Exemplo:

$$T_F = \frac{9}{5}T_C + 32,$$

em que T_F representa a temperatura na escala Fahrenheit e T_C, na escala Celcius. A relação entre as duas temperaturas está representada na figura 1. Note-se que, em uma relação determinística, todos os pontos encontram-se sob a curva que representa a relação. No caso da figura 1, a relação é linear e os pontos encontram-se sob a mesma reta.

Relação Estatística: Diferentemente de uma relação funcional, a relação estatística não é perfeita. Toda relação estatística é caracterizada por gráficos de dispersão. O gráfico de dispersão sugere que parte da variação em Y não pode ser explicada a partir de X, diferentemente do que acontece em uma relação determinística, em que toda alteração no valor de Y corresponde a uma variação no valor de X.

Na figura 2, cada gráfico de dispersão representa a simulação de 500 realizações do par (X, Y). Cada ponto (x_i, y_i) no gráfico, para $i = 1, \ldots, n$, é chamado de *caso* ou *observação*.

A partir da análise da figura 2, é possível notar que relações estatísticas entre as variáveis podem variar em direção e intensidade. O gráfico de dispersão da direita apresenta uma relação fortemente linear e negativa, enquanto o gráfico central apresenta relação positiva não tão forte.

É possível medir *direção* e *intensidade* da relação estatística entre um par de variáveis através de duas grandezas correlacionadas, a saber *covariância* e *coeficiente de correlação*, desenvolvidas a seguir.

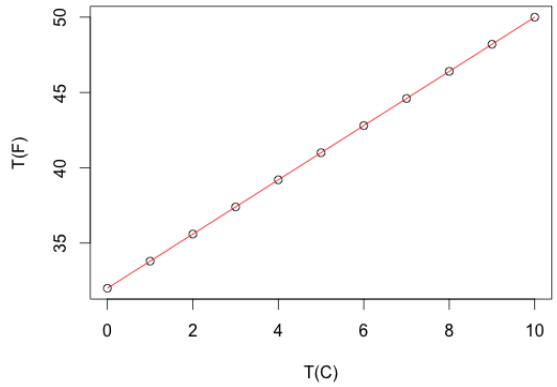

Figura 1: Relação determinística.

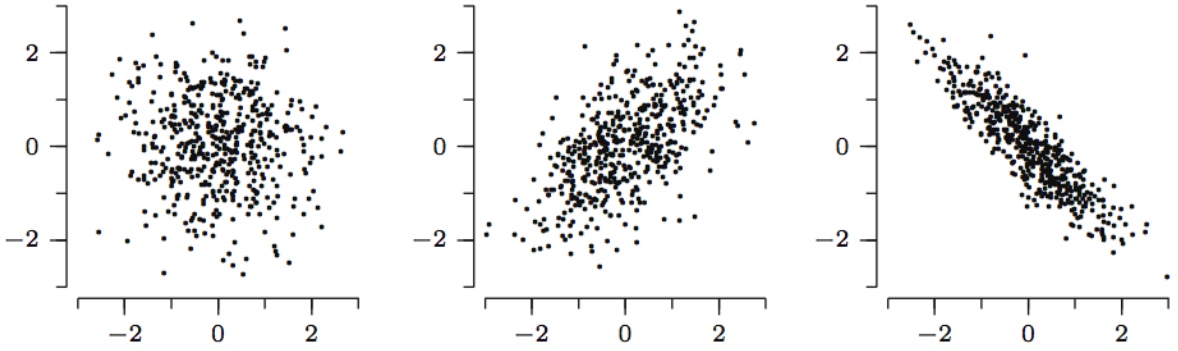

Figura 2: Relação estatística (gráficos de dispersão).

Covariância É possível incluir no gráfico de dispersão as retas vertical e horizontal que passam, respectivamente, pelas médias amostrais $\bar{x}$ e $\bar{y}$, dadas por

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{e} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Tais retas definem quatro quadrantes no gráfico de dispersão, conforme ilustra a figura 3. Pode-se calcular, para cada observação $i = 1, \ldots, n$, as quantidades dadas por:

- $(x_i - \bar{x})$ (desvio da média da variável explicativa)
- $(y_i - \bar{y})$ (desvio da média da variável de resposta)
- $(x_i - \bar{x})(y_i - \bar{y})$ (produto dos desvios)

Se a relação entre X e Y é aproximadamente linear e positiva, é fácil perceber que as observações no primeiro quadrante terão $(x_i - \bar{x}) > 0$, $(y_i - \bar{y}) > 0$ e, consequentemente,

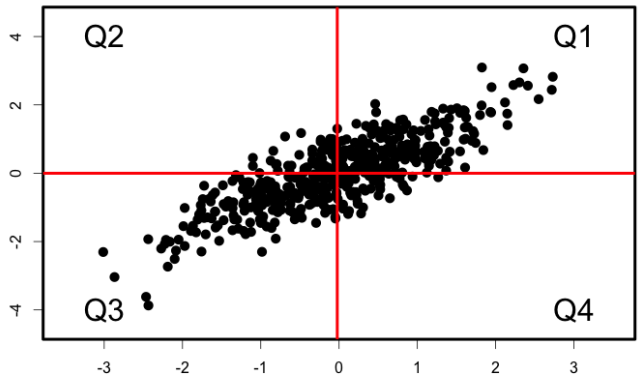

Figura 3: Quadrantes num gráfico de dispersao.

$(x_i - \bar{x})(y_i - \bar{y}) > 0$. No segundo quadrante, $(x_i - \bar{x}) < 0$, $(y_i - \bar{y}) > 0$ e $(x_i - \bar{x})(y_i - \bar{y}) < 0$. No terceiro quadrante, $(x_i - \bar{x}) < 0$, $(y_i - \bar{y}) < 0$ e $(x_i - \bar{x})(y_i - \bar{y})>0$. E, finalmente, no quarto quadrante, $(x_i - \bar{x}) > 0$, $(y_i - \bar{y}) < 0$ e $(x_i - \bar{x})(y_i - \bar{y}) < 0$. Ainda, é de se esperar que haja mais observações no primeiro e terceiro quadrantes do que no segundo e quarto. Sendo assim, espera-se que a soma dos produtos dos desvios seja positiva.

Por outro lado, para uma relação aproximadamente linear e negativa, tende-se a observar maior concentração de pontos no segundo e quarto quadrantes. Sendo assim e seguindo o raciocínio anteriormente desenvolvido, conclui-se que a soma dos desvios tende a ser negativa.

A covariância entre X e Y, definida por

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

indica a direção da relação entre as duas variáveis. Desta forma, quando $Cov(X,Y) > 0$, conclui-se que a relação entre X e Y é positiva. Se $Cov(X,Y) < 0$, a relação é negativa.

Coeficiente de correlação A covariância, no entanto, não fornece uma medida da intensidade da relação, já que depende das unidades em que as variáveis são expressas. Uma maneira de contornar este problema é através da padronização dos dados, subtraindo de cada x_i e y_i, para $i = 1, \ldots, n$ sua respectiva média e dividindo pelo desvio-padrão amostral correspondente. Temos as seguintes variáveis padronizadas:

$$u_i = \frac{x_i - \bar{x}}{s_x} \quad \text{e} \quad v_i = \frac{y_i - \bar{y}}{s_y}, \tag{2}$$

em que

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \text{e} \quad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

correspondem aos desvios-padrão amostrais. As variáveis aleatórias padronizadas, U e V, têm média zero e desvio-padrão unirário. A covariância entre U e V é chamada *coeficiente de*

correlação, dado por:

$$Cor(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right). \tag{3}$$

Outras formas para o coeficiente de correlação são:

$$r_{XY} = Cor(X,Y) = \frac{Cov(X,Y)}{s_x s_y} \tag{4}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{5}$$

Propriedades:

- Medida da associação *linear* entre duas variáveis (intensidade)
- Sinal indica se a relação entre duas variáveis é positiva ou negativa (direção)
- É adimensional
- X, Y independentes $\Longrightarrow r_{XY} = 0$ (o contrário não necessariamente vale)
- $-1 \leq r_{XY} \leq 1$
- Pode ser fortemente influenciado por "outliers".

Embora o coeficiente de correlação seja bastante útil em medir a força e a direção da relação linear entre duas variáveis, não pode ser utilizado para previsão (qual o valor de Y, para um determinado valor de X?). Ainda, o coeficiente de correlação avalia apenas a relação entre pares de variáveis.

Análise de regressão, que veremos a seguir, pode ser encarada como uma extensão interessante de análise de correlação pois, não limita a avaliação da relação a pares de variáveis e permite a construção de um modelo que, além de medir intensidade e direção da relação entre resposta e variáveis explicativas, também pode ser usada para previsão, pois descreve quantitativamente esta relação.