Correlação e Teorema do Limite Central

Docente: Manuel G. Scotto

Departamento de Matemática
IST, ULisboa

Covariância e Correlação

Importante

A covariância e o coeficiente de correlação medem o **grau de dependência linear** entre as variáveis X e Y.

A covariância entre X e Y é

$$
\begin{aligned}
Cov(X, Y) &= E[(X - E(X))(Y - E(Y))] \\
&= E(XY) - E(X)E(Y),
\end{aligned}
$$

se este valor esperado existir.

Exercício

Calcular a covariância entre as variáveis aleatórias discretas X e Y, sendo a função de probabilidade conjunta

$$P(X=x, Y=y) = \begin{cases} 0.1, & x=1,\, y=3 \\ 0.3, & x=1,\, y=5 \\ 0.4, & x=2,\, y=3 \\ 0.2, & x=2,\, y=5 \\ 0, & \text{outros casos} \end{cases}.$$

Exercício

1. Distribuição marginal de X e valor esperado

$$P(X=x)=\begin{cases}0.4, & x=1\\ 0.6, & x=2\\ 0, & \text{outros casos}\end{cases}.$$

$$E(X)=1\times 0.4+2\times 0.6=1.6$$

2. Distribuição marginal de Y e valor esperado

$$P(Y=y)=\begin{cases}0.5, & y=3\\ 0.5, & y=5\\ 0, & \text{outros casos}\end{cases}.$$

$$E(Y)=3\times 0.5+5\times 0.5=4$$

Exercício

1. Valor esperado $E(XY)$

$$
\begin{aligned}
E(XY) &= 1 \times 3 \times 0.1 + 1 \times 5 \times 0.3 + 2 \times 3 \times 0.4 \\
&+ 2 \times 5 \times 0.2 \\
&= 6.2
\end{aligned}
$$

2. Covariância

$$
\begin{aligned}
Cov(X, Y) &= 6.2 - 1.6 \times 4 \\
&= -0.35
\end{aligned}
$$

Importante

Se X e Y forem variáveis aleatórias **independentes** então a covariância entre X e Y é **nula**. No entanto, a implicação no sentido inverso não é **necessariamente verdadeira**.

Coeficiente de correlação

O coeficiente de correlação entre X e Y é

$$corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}.$$

Exercício

Calcular o coeficiente de correlação entre as variáveis aleatórias contínuas X e Y, sendo a função de densidade conjunta

$$f_{X,Y}(x,y) = \begin{cases} 2 & x \in (0,1),\, 0 < y < x \\ 0 & \text{outros casos} \end{cases} .$$

The NEW ENGLAND JOURNAL of MEDICINE

OCCASIONAL NOTES

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

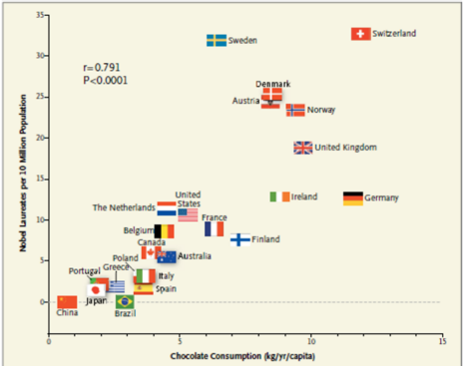
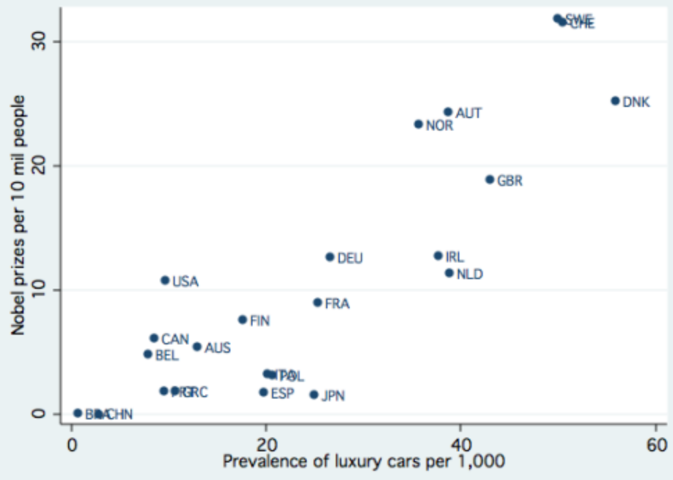
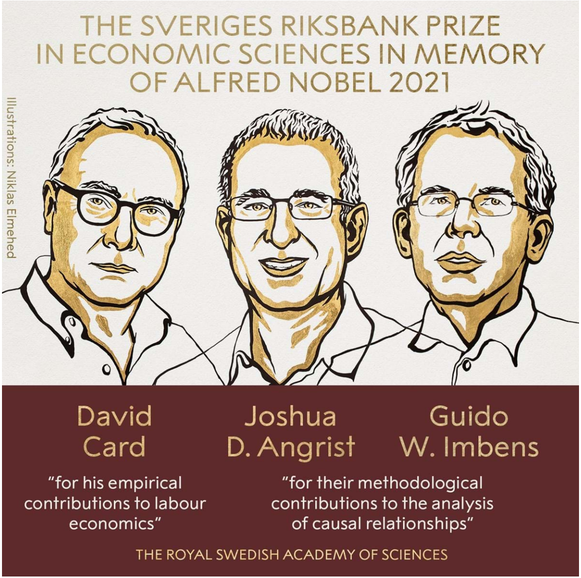


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Covariância e Correlação




THE SVERIGES RIKSBANK PRIZE IN ECONOMIC SCIENCES IN MEMORY OF ALFRED NOBEL 2021
Illustrations: Niklas Elmehed
David Card
"for his empirical contributions to labour economics"
Joshua D. Angrist
Guido W. Imbens
"for their methodological contributions to the analysis of causal relationships"
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Definição

Sejam $X_1, \ldots, X_n$ variáveis aleatórias e $c_1, \ldots, c_n$ constantes reais. Então a variável aleatória

$$Y = \sum_{i=1}^{n} c_i X_i,$$

diz-se uma combinação linear das v.a's $X_1, \ldots, X_n$.

Valor esperado e variância

- Valor Esperado: $E(Y) = E\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i E(X_i)$;
- Variância:

$$V(Y) = \sum_{i=1}^{n} c_i^2 V(X_i) + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} c_i c_j Cov(X_i, X_j);$$

- $V(X_1 + X_2) = V(X_1) + V(X_2) + 2Cov(X_1, X_2)$;
- **Importante:** $V(X_1 - X_2) = V(X_1) + V(X_2) - 2Cov(X_1, X_2)$.

Combinações Lineares de Variáveis Aleatórias

V.a.	Combinação linear
$X_i \sim_{indep}$ binomial(n_i, p), $i = 1, \ldots, k$	$\sum_{i=1}^{k} X_i \sim$ binomial $\left(\sum_{i=1}^{k} n_i, p\right)$
$X_i \sim_{indep}$ Poisson(λ_i), $i = 1, \ldots, n$	$\sum_{i=1}^{n} X_i \sim$ Poisson $(\sum_{i=1}^{n} \lambda_i)$
$X_i \sim_{indep.}$ normal(μ_i, σ_i^2), $i = 1, \ldots, n$	$\sum_{i=1}^{n} X_i \sim$ normal $(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$
	$\sum_{i=1}^{n} c_i X_i \sim$ normal $(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2)$
$X \sim$ exponencial(λ)	$cX \sim$ exponencial $\left(\frac{\lambda}{c}\right)$

Exercício

Na cidade de Lisboa o número de novos casos de uma certa doença que ocorrem diariamente, X, tem uma distribuição de Poisson de parâmetro $\lambda = 2$. Admite-se que as ocorrências da doença são independentes de dia para dia. Seja $Y =$"número de novos casos que se verificam num ano". Calcule $P(700 < Y < 800)$.

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

Exercício

Na cidade de Lisboa o número de novos casos de uma certa doença que ocorrem diariamente, X, tem uma distribuição de Poisson de parâmetro $\lambda = 2$. Admite-se que as ocorrências da doença são independentes de dia para dia. Seja $Y =$"número de novos casos que se verificam num ano". Calcule $P(700 < Y < 800)$.

A v.a Y tem distribuição Poisson de parâmetro $\lambda_Y = 2 \times 365$.

$$\begin{aligned} P(700 < Y < 800) &= \sum_{m=701}^{799} e^{-\lambda_Y} \cdot \frac{\lambda_Y^m}{m!} = F_Y(799) - F_Y(700) \\ &= \texttt{ppois(799, 730)} - \texttt{ppois(700, 730)} \\ &= 0.857. \end{aligned}$$

Exercício

Numa determinada linha de fabrico uma máquina enche sacos de adubo. O peso de cada saco obedece a uma lei Normal com média 10 Kg e desvio padrão 0.5 Kg.

À medida que vão sendo cheios, os sacos são empilhados, em grupos de 10, num tabuleiro mecânico que os transporta para o armazém de expedição. Este tabuleiro não suporta um peso superior a 103 Kg; quando tal sucede o tabuleiro não arranca e os sacos desperdiçam-se. Por outro lado, se o peso dos sacos for inferior a 96 Kg o tabuleiro ganha velocidade excessiva, deixando cair um saco pelo caminho. Qual a probabilidade de o tabuleiro chegar ao armazém tal como foi carregado?

Resolução

A v.a X tem distribuição Normal de média $\mu = 10Kg$ e variância $\sigma^2 = 0.5^2 Kg^2$. Seja Y a variável aleatória

$$Y = X_1 + \cdots + X_{10} = \sum_{m=1}^{10} X_m.$$

Esta nova variável Y tem distribuição Normal de média

$$\mu_Y = 10 \times 10 = 100Kg$$

e variância $\sigma_Y^2 = 10 \times 0.5^2 Kg^2$.

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

Resolução

$$\begin{aligned}
P(96 \leq Y \leq 103) &= P\left(\frac{96-100}{\sqrt{10 \times 0.5^2}} \leq Z \leq \frac{103-100}{\sqrt{10 \times 0.5^2}}\right) \\
&= P(-2.52 \leq Z \leq 1.89) \\
&= F_Z(1.89) - F_Z(-2.52) \\
&= F_Z(1.89) - [1 - F_Z(2.52)] \\
&= F_Z(1.89) - 1 + F_Z(2.52) \\
&= 0.97 - 1 + 0.99 \\
&= 0.96.
\end{aligned}$$

Resolução

$$
\begin{aligned}
P(-2.52 \leq Z \leq 1.89) &= \texttt{pnorm(1.89, 0,1)} - \\
&- \texttt{pnorm(-2.51, 0,1)} \\
&= 0.96
\end{aligned}
$$

$$
\begin{aligned}
P(96 \leq Y \leq 103) &= \texttt{pnorm(103, 100, sqrt(2.5))} - \\
&- \texttt{pnorm(96, 100, sqrt(2.5))} \\
&= 0.96.
\end{aligned}
$$

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

Teorema do Limite Central

A presente secção destina-se a apresentar um dos resultados **mais importantes** da teoria da probabilidade, o *teorema de Lindberg-Levy* mais conhecido como **Teorema do Limite Central**.

Este resultado garante que a soma de n variáveis aleatórias independentes - todas com a mesma média e a mesma variância - tem, depois de estandardizada e para valores de n suficientemente grandes, distribuição aproximada $N(0, 1)$.

Definição

Dada uma sucessão de variáveis aleatórias i.i.d. com valor esperado μ e variância σ^2. Considere-se ainda $S_n = \sum_{i=1}^{n} X_i$ e $\bar{X}_n = S_n/n$. Então

$$\begin{aligned}\lim_{n\to\infty} P\left(\frac{S_n - E(S_n)}{\sqrt{V(S_n)}} \leq z\right) &= \lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq z\right)\\ &= \Phi(z).\end{aligned}$$

Definição (cont)

De forma equivalente,

$$\begin{aligned}\lim_{n\to\infty} P\left(\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{V(\bar{X}_n)}} \le z\right) &= \lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \le z\right) \\ &= \Phi(z).\end{aligned}$$

Exercício

Um computador, ao adicionar números, arredonda cada parcela para o inteiro mais próximo. Admita-se que todos os erros de arredondamento sejam independentes e uniformemente distribuídos em $(-0.5, 0.5)$. Se 1500 números forem adicionados, qual é a probabilidade do erro total ultrapassar 15?

Seja $X_i \sim \mathrm{U}(-0.5, 0.5)$, para $i = 1, \ldots, 1500$. Seja a variável aleatória Y "erro total"

$$Y = \sum_{i=1}^{1500} X_i.$$

Para esta variável $E(Y) = E(\sum_{i=1}^{1500} X_i) = \sum_{i=1}^{1500} E(X_i) = 0$ e

Resolução (cont)

a variância é igual a

$$V(Y) = V\left(\sum_{i=1}^{1500} X_i\right) = \sum_{i=1}^{1500} V(X_i) = \frac{1500}{12}.$$

Não esquecer que $V(X_i) = (0.5 - (-0.5))^2/12 = 1/12$. Pede-se para calcular

$$\begin{aligned} P(Y > 15) &= P\left(Z > \frac{15-0}{\sqrt{1500/12}}\right) \\ &= 1 - P\left(Z \leq \frac{15-0}{\sqrt{1500/12}}\right) = 1 - P(Z \leq 1.34) \\ &\overset{tlc}{\approx} 1 - \Phi(1.34) = 1 - 0.9099. \end{aligned}$$

Exercício

O peso médio dos indivíduos duma certa espécie de bivalves é 31g e o respetivo desvio padrão é 2.4g. Recolhe-se uma amostra aleatória de 36 indivíduos desta espécie.

(a) Qual a probabilidade, aproximada, da média da amostra ser inferior a 30g?

(b) Qual a probabilidade, aproximada, da média da amostra estar compreendida entre 30g e 32g?

(c) Qual a probabilidade, aproximada, de o peso total da amostra ser superior a 1100g?

Resolução

O peso médio dos indivíduos duma certa espécie de bivalves é 31g e o respetivo desvio padrão é 2.4g. Recolhe-se uma amostra aleatória de 36 indivíduos desta espécie.

(a) Qual a probabilidade, aproximada, da média da amostra ser inferior a 30g?

Seja $\bar{X} = (X_1 + \cdots + X_{36})/36$. Neste caso $E(\bar{X}) = 31$ e $V(\bar{X}) = 2.4^2/36$. Então

$$\begin{aligned} P(\bar{X} < 30) &= P\left(Z < \frac{30-31}{\sqrt{2.4^2/36}}\right) = P(Z < -2.50) \\ &\overset{tlc}{\approx} \Phi(-2.50) = 1 - \Phi(2.50) = 1 - 0.9938. \end{aligned}$$

Importante

- A utilização do **teorema do limite central** para distribuições **discretas** apresenta um problema, uma vez que se vai *aproximar um fenómeno discreto por uma distribuição contínua*.
- Embora não exista uma solução ótima para todas as situações, é prática comum adotar a chamada **correção de continuidade** que, em geral, permite obter aproximações de boa qualidade.

$$P(a \leq X \leq b) = P(a - \epsilon \leq X \leq b + \epsilon), \; 0 \leq \epsilon < 1.$$

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

Importante (cont)

- A correção de continuidade consiste em fazer $\epsilon = 0.5$.

  Seja $X_1, \ldots, X_n$ uma sucessão de variáveis aleatórias com distribuição de Bernoulli de média p e variância $p(1-p)$. Seja $S_n = \sum_{i=1}^{n} X_i$. Então

$$P(a \leq S_n \leq b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right).$$

Exercício

O número de doentes que se apresentam na urgência do hospital de Santa Maria, por hora, pode ser modelado por uma distribuição de Poisson com taxa 25. Qual a probabilidade aproximada para $P(599 < X < 680)$, onde X denota o número de doentes por dia?

Exercício

Variável aleatória $X = \sum_{i=1}^{24} Y_i$. Então, $X \sim \text{Po}(600)$ e

$$
\begin{aligned}
P(599 < X < 680) &= P(600 \leq X \leq 679) \\
&= P(600 - 0.5 \leq X \leq 679 + 0.5) \\
&= P(X \leq 679.5) - P(X \leq 599.5) \\
&= P\left(\frac{X - 600}{\sqrt{600}} \leq \frac{679.5 - 600}{\sqrt{600}}\right) \\
&- P\left(\frac{X - 600}{\sqrt{600}} \leq \frac{599.5 - 600}{\sqrt{600}}\right) \\
&= F_Z(3.24) - F_Z(-0.02) \\
&\approx \Phi(3.24) - \Phi(-0.02)
\end{aligned}
$$

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA