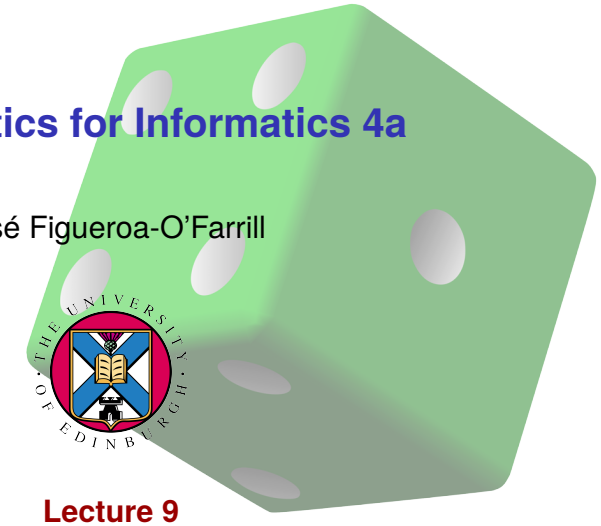
Mathematics for Informatics 4a

José Figueroa-O'Farrill



Lecture 9
15 February 2012

The story of the film so far...

- Discrete random variables $X_1, \dots, X_n$ on the same probability space have a **joint probability mass function**:

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \mathbb{P}(\{X_1 = x_1\} \cap \dots \cap \{X_n = x_n\})$$

- $f_{X_1,\dots,X_n} : \mathbb{R}^n \to [0,1]$ and $\sum_{x_1,\dots,x_n} f_{X_1,\dots,X_n}(x_1,\dots,x_n) = 1$
- $X_1, \dots, X_n$ are **independent** if for all $2 \leqslant k \leqslant n$ and $x_{i_1}, \dots, x_{i_k}$,

$$f_{X_{i_1},\dots,X_{i_k}}(x_{i_1},\dots,x_{i_k}) = f_{X_{i_1}}(x_{i_1}) \dots f_{X_{i_k}}(x_{i_k})$$

- $h(X_1,\dots,X_n)$ is a discrete random variable and

$$\mathbb{E}(h(X_1,\dots,X_n)) = \sum_{x_1,\dots,x_n} h(x_1,\dots,x_n) f_{X_1,\dots,X_n}(x_1,\dots,x_n)$$

- Expectation is linear: $\mathbb{E}(\sum_i \alpha_i X_i) = \sum_i \alpha_i \mathbb{E}(X_i)$

Expectation of a product

Lemma

If X and Y are independent, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Proof.

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_{x,y} xy f_{X,Y}(x,y) \\
&= \sum_{x,y} xy f_X(x) f_Y(y) && \text{(independence)} \\
&= \sum_x x f_X(x) \sum_y y f_Y(y) \\
&= \mathbb{E}(X)\mathbb{E}(Y)
\end{aligned}$$

□

$\mathbb{E}(XY)$ is an inner product

The expectation value defines a real inner product.
If X, Y are two discrete random variables, let us define $\langle X, Y\rangle$ by

$$\langle X, Y\rangle = \mathbb{E}(XY)$$

We need to show that $\langle X, Y\rangle$ satisfies the axioms of an inner product:

1. it is symmetric: $\langle X, Y\rangle = \mathbb{E}(XY) = \mathbb{E}(YX) = \langle Y, X\rangle$
2. it is bilinear:
   - $\langle aX, Y\rangle = \mathbb{E}(aXY) = a\mathbb{E}(XY) = a\langle X, Y\rangle$
   - $\langle X_1 + X_2, Y\rangle = \mathbb{E}((X_1 + X_2)Y) = \mathbb{E}(X_1Y) + \mathbb{E}(X_2Y) = \langle X_1, Y\rangle + \langle X_2, Y\rangle$
3. it is positive-definite: if $\langle X, X\rangle = 0$, then $\mathbb{E}(X^2) = 0$, whence $\sum_x x^2 f(x) = 0$, whence $xf(x) = 0$ for all x. If $x \neq 0$, then $f(x) = 0$ and thus $f(0) = 1$. In other words, $\mathbb{P}(X = 0) = 1$ and hence $X = 0$ almost surely.

Additivity of variance for independent variables

How about the variance $\mathrm{Var}(X+Y)$?

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}((X+Y)^2) - \mathbb{E}(X+Y)^2 \\ &= \mathbb{E}(X^2+2XY+Y^2) - (\mathbb{E}(X)+\mathbb{E}(Y))^2 \\ &= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - \mathbb{E}(X)^2 - 2\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)^2 \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\left(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\right)\end{aligned}$$

Theorem

If X and Y are independent discrete random variables

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Covariance

Definition

The **covariance** of two discrete random variables is

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Letting μ_X and μ_Y denote the means of X and Y, respectively,

$$\mathrm{Cov}(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y))$$

Indeed,

$$\begin{aligned}\mathbb{E}((X-\mu_X)(Y-\mu_Y)) &= \mathbb{E}(XY) - \mathbb{E}(\mu_X Y) - \mathbb{E}(\mu_Y X) + \mathbb{E}(\mu_X\mu_Y) \\ &= \mathbb{E}(XY) - \mu_X\mu_Y\end{aligned}$$

Example (Max and min for two fair dice)

We roll two fair dice. Let X and Y denote their scores. Independence says that $\mathrm{Cov}(X,Y)=0$. Consider however the new variables $U=\min(X,Y)$ and $V=\max(X,Y)$:

U	1	2	3	4	5	6
1	1	1	1	1	1	1
2	1	2	2	2	2	2
3	1	2	3	3	3	3
4	1	2	3	4	4	4
5	1	2	3	4	5	5
6	1	2	3	4	5	6

V	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

$$\mathbb{E}(U)=\tfrac{91}{36},\ \mathbb{E}(U^2)=\tfrac{301}{36},\ \mathbb{E}(V)=\tfrac{161}{36},\ \mathbb{E}(V^2)=\tfrac{791}{36},\ \mathbb{E}(UV)=\tfrac{49}{4}$$

$$\Longrightarrow \mathrm{Var}(U)=\mathrm{Var}(V)=\tfrac{2555}{1296} \quad \text{and} \quad \mathrm{Cov}(U,V)=\left(\tfrac{35}{36}\right)^2$$

Definition

Two discrete random variables X and Y are said to be **uncorrelated** if $\mathrm{Cov}(X,Y)=0$.

Warning

Uncorrelated random variables need **not** be independent!

Counterexample

Suppose that X is a discrete random variable with probability mass function symmetric about 0; that is, $f_X(-x)=f_X(x)$. Let $Y=X^2$. Clearly X,Y are not independent: $f(x,y)=0$ unless $y=x^2$ even if $f_X(x)f_Y(y)\neq 0$. However they are uncorrelated:

$$\mathbb{E}(XY)=\mathbb{E}(X^3)=\sum_x x^3 f_X(x)=0$$

and similarly $\mathbb{E}(X)=0$, whence $\mathbb{E}(X)\mathbb{E}(Y)=0$.

An alternative criterion for independence

The above counterexample says that the following implication cannot be reversed:

$$X, Y \quad \text{independent} \implies \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$

However, one has the following

Theorem

Two discrete random variables X and Y are independent if and only if

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

for all functions g, h.

The proof is not hard, but we will skip it.

The Cauchy–Schwarz inequality

Recall that $\langle X, Y\rangle = \mathbb{E}(XY)$ is an inner product.
Every inner product obeys the ▸ Cauchy–Schwarz inequality:

$$\langle X, Y\rangle^2 \leqslant \langle X, X\rangle \langle Y, Y\rangle$$

which in terms of expectations is

$$\mathbb{E}(XY)^2 \leqslant \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

Now,

$$\mathrm{Cov}(X, Y)^2 = \mathbb{E}((X - \mu_X)(Y - \mu_Y))^2 \leqslant \mathbb{E}((X - \mu_X)^2)\mathbb{E}((Y - \mu_Y)^2)$$

whence

$$\mathrm{Cov}(X, Y)^2 \leqslant \mathrm{Var}(X)\,\mathrm{Var}(Y)$$

Correlation

Let X and Y be two discrete random variables with means μ_X and μ_Y and standard deviations σ_X, σ_Y. The **correlation** $\rho(X, Y)$ of X and Y is defined by

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

From the Cauchy–Schwarz inequality, we see that

$$\rho(X, Y)^2 \leqslant 1 \implies -1 \leqslant \rho(X, Y) \leqslant 1$$

Hence the correlation is a number between -1 and 1:

- a correlation of 1 suggests a linear relation with positive slope between X and Y,
- whereas a correlation of -1 suggests a linear relation with negative slope.

Example (Max and min for two fair dice – continued)

Continuing with the ▸ previous example, we now simply compute

$$\rho(U, V) = \frac{\mathrm{Cov}(U, V)}{\sqrt{\mathrm{Var}(U)\,\mathrm{Var}(V)}} = \frac{35^2}{36^2} \Big/ \frac{2555}{36^2} = \frac{35}{73}.$$

Remark

The funny normalisation of $\rho(X, Y)$ is justified by the following:

$$\rho(\alpha X + \beta, \gamma Y + \delta) = \mathrm{sign}(\alpha\gamma)\rho(X, Y)$$

which follows from

$$\mathrm{Cov}(\alpha X + \beta, \gamma Y + \delta) = \alpha\gamma\,\mathrm{Cov}(X, Y)$$

and $\sigma_{\alpha X+\beta} = |\alpha|\sigma_X$ and $\sigma_{\gamma Y+\delta} = |\gamma|\sigma_Y$.

Markov's inequality

Theorem (Markov's inequality)

Let X be a discrete random variable taking non-negative values. Then

$$\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}(X)}{a}$$

Proof.

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x \geqslant 0} x\mathbb{P}(X = x) = \sum_{0 \leqslant x < a} x\mathbb{P}(X = x) + \sum_{x \geqslant a} x\mathbb{P}(X = x) \\ &\geqslant \sum_{x \geqslant a} x\mathbb{P}(X = x) \geqslant \sum_{x \geqslant a} a\mathbb{P}(X = x) = a\mathbb{P}(X \geqslant a)\end{aligned}$$

□

Example

A factory produces an average of n items every week. *What can be said about the probability that this week's production shall be at least $2n$ items?*
Let X be the discrete random variable counting the number of items produced. Then by Markov's inequality

$$\mathbb{P}(X \geqslant 2n) \leqslant \frac{n}{2n} = \frac{1}{2}\,.$$

So I wouldn't bet on it!

Markov's inequality is not terribly sharp; e.g.,

$$\mathbb{P}(X \geqslant \mathbb{E}(X)) \leqslant 1\,.$$

It has one interesting corollary, though.

Chebyshev's inequality

Theorem

Let X be a discrete random variable with mean μ and variance σ^2. Then for any $\varepsilon > 0$,

$$\mathbb{P}(|X - \mu| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{\varepsilon^2}$$

Proof.

Notice that for $\varepsilon > 0$, $|X - \mu| \geqslant \varepsilon$ if and only if $(X - \mu)^2 \geqslant \varepsilon^2$, so

$$\begin{aligned}\mathbb{P}(|X - \mu| \geqslant \varepsilon) &= \mathbb{P}((X - \mu)^2 \geqslant \varepsilon^2) \\ &\leqslant \frac{\mathbb{E}((X - \mu)^2)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \qquad \text{(by Markov's)}\end{aligned}$$

□

Example

Back to the factory in the previous example, let the average be $n = 500$ and the variance in a week's production is 100, then what can be said about the probability that this week's production falls between 400 and 600?
By Chebyshev's,

$$\mathbb{P}(|X - 500| \geqslant 100) \leqslant \frac{\sigma^2}{100^2} = \frac{1}{100}$$

whence

$$\begin{aligned}\mathbb{P}(|X - 500| < 100) &= 1 - \mathbb{P}(|X - 500| \geqslant 100) \\ &\geqslant 1 - \tfrac{1}{100} = \tfrac{99}{100}\,.\end{aligned}$$

So pretty likely!

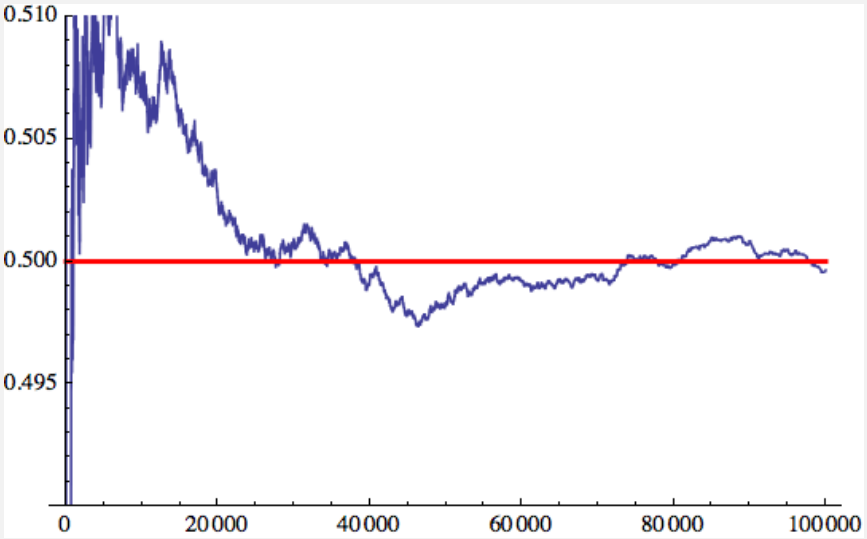
The law of large numbers I

Consider a number n of independent discrete random variables $X_1, \ldots, X_n$ with the same probability mass function. One says that they are "**independent and identically distributed**", abbreviated "**i.i.d.**". In particular, they have the same mean and variance.
The law of large numbers says that in the limit $n \to \infty$,

$$\frac{1}{n}(X_1 + \cdots + X_n) \to \mu$$

in probability.
The law of large numbers justifies the "relative frequency" interpretation of probability. For example, it says that tossing a fair coin a large number n of times, the proportion of heads will approach $\frac{1}{2}$ in the limit $n \to \infty$, in the sense that deviations from $\frac{1}{2}$ (e.g., a long run of heads or of tails) will become increasingly rare.

100,000 tosses of a fair (?) coin



The law of large numbers II

Theorem (The (weak) law of large numbers)

Let $X_1, X_2, \ldots$ be i.i.d. discrete random variables with mean μ and variance σ^2 and let $Z_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Then

$$\forall \varepsilon > 0 \qquad \mathbb{P}(|Z_n - \mu| < \varepsilon) \to 1 \quad \text{as } n \to \infty$$

Proof.

By linearity of expectation, $\mathbb{E}(Z_n) = \mu$, and since the X_i are independent $\mathrm{Var}(Z_n) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) = \frac{\sigma^2}{n}$. By Chebyshev,

$$\mathbb{P}(|Z_n - \mu| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{n\varepsilon^2} \implies \mathbb{P}(|Z_n - \mu| < \varepsilon) \geqslant 1 - \frac{\sigma^2}{n\varepsilon^2}$$

□

The law of large numbers III

We will now justify probability as relative frequency.
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A \in \mathcal{F}$ be an event.
Let I_A denote the **indicator variable** of A, a discrete random variable defined by

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

The probability mass function f of an indicator variable is very simple: $f(1) = \mathbb{P}(A)$ and hence $f(0) = 1 - \mathbb{P}(A)$. Its mean is given by

$$\mu = \mathbb{E}(I_A) = 0 \times f(0) + 1 \times f(1) = f(1) = \mathbb{P}(A)$$

and its variance by

$$\begin{aligned} \sigma^2 = \mathrm{Var}(I_A) &= (0-\mu)^2 f(0) + (1-\mu)^2 f(1) \\ &= \mathbb{P}(A)^2(1 - \mathbb{P}(A)) + (1 - \mathbb{P}(A))^2\mathbb{P}(A) \\ &= \mathbb{P}(A)(1 - \mathbb{P}(A)) \end{aligned}$$

The law of large numbers IV

- Now imagine repeating the experiment and counting how many outcomes belong to A.
- Let X_i denote the random variable which agrees with the indicator variable of A at the ith trial.
- Then the $X_1, X_2, \ldots$ are i.i.d. discrete random variables, with mean $\mathbb{P}(A)$ and variance $\mathbb{P}(A)(1-\mathbb{P}(A))$.
- Let $Z_n = \frac{1}{n}(X_1 + \cdots + X_n)$. What does Z_n measure?
- Z_n measures the proportion of trials with outcomes in A after n trials. This is what we had originally called $N(A)/n$.
- The law of large numbers says that in the limit as $n \to \infty$, $Z_n \to \mathbb{P}(A)$ in probability.
- This makes precise our initial hand-waving argument of $N(A)/n$ "converging in some way" to $\mathbb{P}(A)$.

Summary

- X, Y independent: $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$
- $\mathbb{E}(XY)$ defines an inner product
- X, Y independent: $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
- In general: $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$
- **covariance**: $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. If $\mathrm{Cov}(X,Y) = 0$ we say X, Y are **uncorrelated**
- **correlation**: $\rho(X,Y) = \mathrm{Cov}(X,Y)/(\sigma(X)\sigma(Y))$ measures "linear dependence" between X, Y
- We proved two inequalities:
  - **Markov**: $\mathbb{P}(|X| \geqslant a) \leqslant \mathbb{E}(|X|)/a$
  - **Chebyshev**: $\mathbb{P}(|X-\mu| \geqslant \varepsilon) \leqslant \sigma^2/\varepsilon^2$
- The **law of large numbers** "explains" the relative frequency definition of probability: it says that if X_i are i.i.d. discrete random variables, then as $n \to \infty$, $\frac{1}{n}(X_1 + \cdots + X_n) \to \mu$ *in probability*; i.e., deviations from μ are still possible, but they are increasingly improbable

Proof of the Cauchy–Schwarz inequality

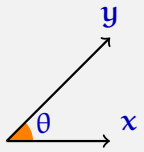
The Cauchy–Schwarz inequality says that if x, y are any two vectors in a positive-definite inner product space, then

$$|\langle x, y\rangle| \leqslant |x||y|\,, \qquad \text{where } |x| = \sqrt{\langle x, x\rangle} \text{ is the length.}$$

Any two vectors lie on a plane, so let's pretend we are in $\mathbb{R}^2$, and diagonalising $\langle -, -\rangle$, we take it to be the dot product. In that case,

$$x \cdot y = |x||y| \cos\theta\,,$$

where θ is the angle between x and y. Since $|\cos\theta| \leqslant 1$, the inequality follows.



Back to the main story.