MS-A0503 First course in probability and statistics

2B Standard deviation and correlation

Jukka Kohonen

Deparment of mathematics and systems analysis
Aalto SCI

Academic year 2019–2020
Period III

Contents

Expectation tells only about location of distribution

For a random number X, the expected value (mean) $\mu = \mathbb{E}(X)$:

- is the probability-weighted average of X's possible values, $\sum_x x\,f(x)$ or $\int x\,f(x)dx$
- is roughly a central location of the distribution
- is the approximate average of many independent random numbers that are distributed like X
- tells nothing about the width of the distribution

Example

Some discrete distributions with the same expectation 1:

k	1
$\mathbb{P}(X = k)$	1

k	0	1	2
$\mathbb{P}(Z = k)$	$\frac{1}{2}$	0	$\frac{1}{2}$

k	0	1	2
$\mathbb{P}(Y = k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

k	0	1000000
$\mathbb{P}(W = k)$	0.999999	0.000001

How to measure the difference of X from its expectation?

The absolute difference of X from its mean $\mu = \mathbb{E}(X)$ is a random variable $|X - \mu|$.

If in dice-rolling ($\mu = 3.5$) we obtain $X = 2$, then $X - \mu = -1.5$.

The *mean absolute difference* $\mathbb{E}|X - \mu|$:

- is an approximation to the average $\frac{1}{n}\sum_{i=1}^{n}|X_i - \mu|$, from a large number of independent random numbers distributed like X
- is mathematically slightly inconvenient, because (among other things) the function $x \mapsto |x|$ is not differentiable at zero.

What if we instead use the squared difference $(X - \mu)^2$

Variance

If X has mean $\mu = \mathbb{E}(X)$, then the *squared difference* of X from the mean is a random number $(X - \mu)^2$.

If in dice rolling ($\mu = 3.5$) we obtain $X = 2$, then squared difference is $(2 - 3.5)^2 = (-1.5)^2 = 2.25$.

The expectation *of the squared difference* is called the variance of the random number X: $\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2]$:

- approximates average $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ in many repetitions
- is mathematically convenient, (among other things) because the squaring function $x \mapsto x^2$ has derivatives of all orders
- has the units of *squared* something

	X	$\mathrm{Var}(X)$
Height	m	m^2
Time	s	s^2
Sales	EUR	EUR^2

We go back to the original units by taking the square root.

Standard deviation

Standard deviation, $\mathrm{SD}(X) = \sqrt{\mathbb{E}[(X-\mu)^2]}$ is the *expectation* of the square-difference, returned back to original scale by square root.

Other notations also exist, like $\mathbb{D}(X)$.

It measures:

- (roughly, in cumbersome square-squareroot-way) how much realizations of X are expected to differ from their mean
- width of the distribution of X

For discrete distributions:

$$\mu = \sum_x x\, f(x)$$

$$\mathrm{SD}(X) = \sqrt{\sum_x (x-\mu)^2 f(x)}$$

For continuous distributions:

$$\mu = \int x\, f(x) dx$$

$$\mathrm{SD}(X) = \sqrt{\int (x-\mu)^2\, f(x) dx}$$

Example. Some distributions with mean 1

What are the standard deviations of X, Y, Z?

k	1
$\mathbb{P}(X = k)$	1

k	0	1	2
$\mathbb{P}(Y = k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

k	0	2
$\mathbb{P}(Z = k)$	$\frac{1}{2}$	$\frac{1}{2}$

$$\mathrm{SD}(X) = \sqrt{\sum_k (k-\mu)^2 f_X(k)} = \sqrt{(1-1)^2 \times 1} = 0.$$

$$\mathrm{SD}(Y) = \sqrt{(0-1)^2 \times \frac{1}{3} + (1-1)^2 \times \frac{1}{3} + (2-1)^2 \times \frac{1}{3}} = \sqrt{\frac{2}{3}} \approx 0.82.$$

$$\mathrm{SD}(Z) = \sqrt{(0-1)^2 \times \frac{1}{2} + (1-1)^2 \times 0 + (2-1)^2 \times \frac{1}{2}} = 1.$$

Standard deviation: Alternative (equivalent) formula

Fact

If X has mean $\mu = \mathbb{E}(X)$, then it is also true that

$$\mathsf{SD}(X) \;=\; \sqrt{\mathsf{Var}(X)} \;=\; \sqrt{\mathbb{E}(X^2) - \mu^2}.$$

(This is convenient for calculation, if $\mathbb{E}(X^2)$ is easy to calculate.)

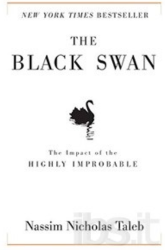
Proof.

$$\begin{aligned} \mathsf{Var}(X) \;=\; \mathbb{E}[(X-\mu)^2] \;&=\; \mathbb{E}[X^2 - 2\mu X + \mu^2] \\ &=\; \mathbb{E}[X^2] - \mathbb{E}[2\mu X] + \mathbb{E}[\mu^2] \\ &=\; \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 \\ &=\; \mathbb{E}[X^2] - \mu^2 \end{aligned}$$

$$\Longrightarrow \qquad \mathsf{SD}(X) \;=\; \sqrt{\mathsf{Var}(X)} \;=\; \sqrt{\mathbb{E}[X^2] - \mu^2}$$

$\square$

Example: Black swan — Two-valued distribution



k	0	10^6
$\mathbb{P}(X = k)$	$1 - 10^{-6}$	10^{-6}

$$\mu = \mathbb{E}(X) = 1$$

Calculate the standard deviation.

Method 1 (straight from the definition):

$$\begin{aligned} \mathrm{SD}(X) &= \sqrt{\sum_x (x-\mu)^2 f(x)} \\ &= \sqrt{(0-1)^2 \times (1-10^{-6}) + (10^6-1)^2 \times 10^{-6}} \approx 1000. \end{aligned}$$

Method 2 (alternative formula):

$$\mathbb{E}(X^2) = \sum_x x^2 f(x) = 0^2 \times (1-10^{-6}) + (10^6)^2 \times 10^{-6} = 10^6.$$

$$\implies \quad \mathrm{SD}(X) = \sqrt{\mathbb{E}(X^2) - \mu^2} = \sqrt{10^6 - 1^2} \approx 1000.$$

Example: Metro — Continuous uniform distribution

Waiting time X is uniformly distributed in interval $[0, 10]$. Then it has mean $\mu = 5$ (minutes). What is the standard distribution?

Method 1 (from definition):

$$\mathrm{SD}(X) = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx} = \sqrt{\int_{0}^{10} (x-5)^2 \frac{1}{10} dx} = \cdots$$

Method 2 (by alternative formula):

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_{0}^{10} x^2 \frac{1}{10} dx = \frac{1}{10}\left[\frac{1}{3}x^3\right]_0^{10} \approx 33.33.$$

$$\Longrightarrow \quad \mathrm{SD}(X) = \sqrt{\mathbb{E}(X^2) - \mu^2} = \sqrt{33.33 - 5^2} \approx 2.89 \text{ minutes.}$$

Std. dev. of shifted and scaled random numbers

Fact (Previous lecture)

(i) $\mathbb{E}(a) = a$.

(ii) $\mathbb{E}(bX) = b\mathbb{E}(X)$.

(iii) $\mathbb{E}(a + bX) = a + b\mathbb{E}(X)$.

Fact

(i) $\mathrm{SD}(a) = 0$.

(ii) $\mathrm{SD}(bX) = |b|\,\mathrm{SD}(X)$.

(iii) $\mathrm{SD}(a + bX) = |b|\,\mathrm{SD}(X)$.

Proof.

$$\begin{aligned}
\mathrm{Var}(a + bX) &= \mathbb{E}[\big(a + bX - \mathbb{E}[a + bX]\big)^2] \\
&= \mathbb{E}[\big(a + bX - a - b\mu\big)^2] \\
&= \mathbb{E}[\big(bX - b\mu\big)^2] \\
&= \mathbb{E}[b^2(X - \mu)^2] \;=\; b^2\mathbb{E}[(X - \mu)^2] \;=\; b^2\,\mathrm{Var}(X),
\end{aligned}$$

$$\mathrm{SD}(a + bX) \;=\; \sqrt{\mathrm{Var}(a + bX)} \;=\; \sqrt{b^2\,\mathrm{Var}(X)} \;=\; |b|\,\mathrm{SD}(X).$$

This proves (iii). Items (i) and (ii) follow as special cases. $\square$

Contents

Chebyshev's inequality: probability of large differences

Fact (Chebyshev's inequality)

For any random variable that has mean μ *and standard deviation* σ*, it is true that the event* $\{X = \mu \pm 2\sigma\} = \{X \in [\mu - 2\sigma,\ \mu + 2\sigma]\}$ *has probability at least*

$$\mathbb{P}(X = \mu \pm 2\sigma) \geq \frac{3}{4}.$$

Pafnuty Chebyshev
1821–1894

More generally $\mathbb{P}(X = \mu \pm r\sigma) \geq 1 - \frac{1}{r^2}$ for any $r \geq 1$.

- X is rather probably ($\geq 75\%$) *within* two std. deviations from its mean
- X is very probably ($\geq 99\%$) within ten std. deviations from its mean

Chebyshev's inequality gives a *lower bound* for the "near mean" probability. For particular distributions, the real probability may be larger. (For "tail probability" we have an *upper bound*.)

Example: Document lengths

In a certain journal, word counts of articles have mean 1000 and standard deviation 200. We don't know the exact distribution. Is it probable that a randomly chosen article's word count is

(a) within $[600, 1400]$? (two std.dev. from mean)

(b) within $[800, 1200]$? (one std.dev. from mean)

Solution

(a) From Chebyshev's inequality

$$\mathbb{P}(X \in [600, 1400]) \;=\; \mathbb{P}(X = \mu \pm 2\sigma) \;\geq\; 1 - \frac{1}{2^2} \;=\; 75\%,$$

so at least 75% of articles are like this.

(b) Here Chebyshev says nothing very useful. All it says is

$$\mathbb{P}(X \in [800, 1200]) \;=\; \mathbb{P}(X = \mu \pm \sigma) \;\geq\; 1 - \frac{1}{1^2} \;=\; 0.$$

We would need better information about the actual distribution.

Example: Document lengths (take two)

In a certain journal, word counts of articles have mean 1000 and standard deviation 200. We also happen to know they are normally distributed. Is it probable that a randomly chosen article's word count is

(a) within $[600, 1400]$ (two std.dev. from mean)

(b) within $[800, 1200]$ (one std.dev. from mean)

Solution

(a) From the CDF of normal distribution (e.g. in R: `1-2*pnorm(-2)`)

$$\mathbb{P}(X \in [600, 1400]) = \mathbb{P}(X = \mu \pm 2\sigma) = \mathbb{P}(\frac{X - \mu}{\sigma} = 0 \pm 2) \approx 95\%.$$

(b) From the CDF of normal distribution (e.g. in R: `1-2*pnorm(-1)`)

$$\mathbb{P}(X \in [800, 1200]) = \mathbb{P}(X = \mu \pm \sigma) = \mathbb{P}\left(\frac{X - \mu}{\sigma} = 0 \pm 1\right) \approx 68\%.$$

We got much higher probabilities because we knew the distribution.

Example: Document lengths (take three)

In a certain journal, word counts of articles have mean 1000 and standard deviation 200; in fact, they have distribution

k	750	1000	1250
$\mathbb{P}(X = k)$	32%	36%	32%

Is it probable that a randomly chosen article's word count is

(a) within $[600, 1400]$ (two std.dev. from mean)

(b) within $[800, 1200]$ (one std.dev. from mean)

Solution

Directly from the distribution table, we see that the word count is

(a) *certainly* (100%) within $[600, 1400]$

(b) but not very probably (only 36%) within $[800, 1200]$

Food for thought: How was this example generated? We wanted a distribution that has SD=200, and two possible values symmetric around the mean. But how to choose their probabilities so that we get the SD we wanted?

Proving Chebyshev (continuous; dicrete similar)

Let $r > 0$. Suppose X has density $f(x)$, mean μ and standard deviation σ. Let MID be the interval $[\mu - r\sigma, \mu + r\sigma]$ and TAIL its complement. Now

$$\begin{aligned}
\mathrm{Var}(X) = \sigma^2 &= \int_{\mathbb{R}} (x-\mu)^2 f(x) dx = \int_{\mathrm{MID}} (\dots) + \int_{\mathrm{TAIL}} (\dots) \\
&\geq \int_{\mathrm{TAIL}} (x-\mu)^2 f(x) dx \quad \geq \quad \int_{\mathrm{TAIL}} (r\sigma)^2 f(x) dx \\
&= r^2\sigma^2 \int_{\mathrm{TAIL}} f(x) dx = r^2\sigma^2\, \mathbb{P}(x \in \mathrm{TAIL}).
\end{aligned}$$

Cancel σ^2 and move r^2 to other side:

$$\mathbb{P}(X \in \mathrm{TAIL}) \leq \frac{1}{r^2}.$$

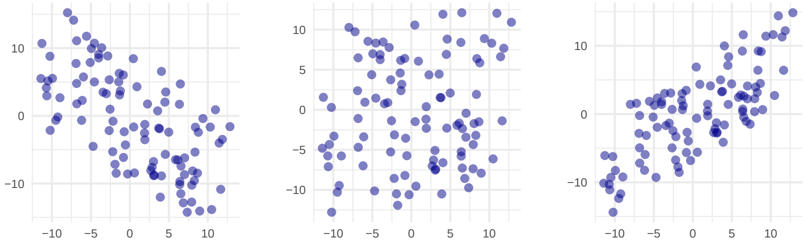
Note: From Chebyshev, one can actually prove the (Weak) Law of Large Numbers. One extra ingredient is needed, namely the variance of a sum; see next lecture and https://en.wikipedia.org/wiki/Law_of_large_numbers

Contents

Shape of the joint distribution

Standard deviation measures the dispersion of *one* r.v. around its mean.

For two random variables, we would like to know X and Y typically differ (from their means) *to the same direction* and how strong this effect is.



Covariance

$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$, measures how strongly X and Y vary in the same direction.

Discrete

$$\sum_x \sum_y (x - \mu_X)(y - \mu_Y)\, f(x, y)$$

Continuous

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)\, f(x, y) dxdy.$$

The covariance

- is > 0, if $X - \mu_X$ and $Y - \mu_Y$ have often the same sign
- is < 0, if $X - \mu_X$ and $Y - \mu_Y$ have often opposite signs
- like variance, is in square units (m^2, s^2, EUR^2, ...)

But now we do not want to take the square root (why)? (Can be neg.)

Covariance: Alternative formula

Often more convenient in calculations than the definition.

Fact

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

Proof.

$$
\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= \mathbb{E}[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\
&= \mathbb{E}[XY] - \mu_X \mathbb{E}[Y] - \mu_Y \mathbb{E}[X] + \mathbb{E}[\mu_X \mu_Y] \\
&= \mathbb{E}[XY] - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\
&= \mathbb{E}[XY] - \mu_X \mu_Y .
\end{aligned}
$$

$\square$

Symmetry and linearity of covariance

Fact

The covariance $\mathrm{Cov}(X, Y)$ *is symmetric and linear in each of its arguments:*

$\mathrm{Cov}(Y, X) = \mathrm{Cov}(X, Y)$
$\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y).$
$\mathrm{Cov}(X, Y_1 + Y_2) = \mathrm{Cov}(X, Y_1) + \mathrm{Cov}(X, Y_2).$
$\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y)$

More generally:

$$\mathrm{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j\,\mathrm{Cov}(X_i, Y_j)$$

Proving linearity of covariance

Let's denote $Y = \sum_{j=1}^{n} b_j Y_j$. Using the "alternative formula" of covariance, and linearity of expectation,

$$\begin{aligned}
\mathrm{Cov}\left(\sum_i a_i X_i, Y\right) &= \mathbb{E}[(\sum_i a_i X_i) Y] - \mathbb{E}[(\sum_i a_i X_i)]\mathbb{E}[Y] \\
&= \sum_i a_i \mathbb{E}[X_i Y] - \left(\sum_i a_i \mathbb{E}[X_i]\right) \mathbb{E}[Y] \\
&= \sum_i a_i \mathbb{E}[X_i Y] - \sum_i a_i \mathbb{E}[X_i]\mathbb{E}[Y] \\
&= \sum_i a_i \left(\mathbb{E}[X_i Y] - \mathbb{E}[X_i]\mathbb{E}[Y]\right) = \sum_i a_i \,\mathrm{Cov}(X_i, Y).
\end{aligned}$$

By symmetry and the above, we obtain

$$\begin{aligned}
\sum_i a_i \,\mathrm{Cov}(X_i, Y) &= \sum_i a_i \,\mathrm{Cov}(Y, X_i) \\
&= \sum_i a_i \,\mathrm{Cov}(\sum_j b_j Y_j, X_i) \\
&= \sum_i a_i \sum_j b_j \,\mathrm{Cov}(Y_j, X_i) \\
&= \sum_i \sum_j a_i b_j \,\mathrm{Cov}(X_i, Y_j).
\end{aligned}$$

Covariance: Summary

The covariance of random variables X and Y is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

where $\mu_X = \mathbb{E}(X)$ ja $\mu_Y = \mathbb{E}(Y)$.

Discrete

$$\sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y)$$

Continuous

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy.$$

Covariance is symmetric and linear:

$$\text{Cov}(Y, X) = \text{Cov}(X, Y)$$

$$\text{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \, \text{Cov}(X_i, Y_j)$$

Correlation (coefficient)

It would be awkward to "normalize" covariance by square root (because covariance can be negative).

Also, we would like to know the covariance *relative* to the scaling of the two variables. (Think what happens to covariance if both variables multiplied by 1000.)

Here we apply a different kind of normalization ...

Correlation (coefficient)

$$\mathrm{Cor}(X, Y) \;=\; \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)}$$

measures how X and Y vary jointly, in *normalized* units.

Independent random numbers are uncorrelated

Fact

If X and Y are (stochastically) independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ *and* $\mathrm{Cor}(X, Y) = 0$.

Proof.

In the discrete case:

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_x \sum_y xy \; f_{X,Y}(x, y) \\
&= \sum_x \sum_y xy \; f_X(x) f_Y(y) \\
&= \left(\sum_x x \; f_X(x) \right) \left(\sum_y y \; f_Y(y) \right) \; = \; \mathbb{E}(X)\mathbb{E}(Y).
\end{aligned}$$

Applying the covariance formula
$\mathrm{Cov}(X, Y) \; = \; \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \; = \; \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) \; = \; 0.$
Thus also $\mathrm{Cor}(X, Y) = 0$. $\square$

Example. Two binary random variables

X and Y are both uniformly distributed among two values $\{-1, +1\}$.

More over $\mathbb{P}(X = +1, Y = +1) \;=\; c$.

Find joint distribution and correlation.

		Y: -1	Y: $+1$	Sum
X	-1	c	$\frac{1}{2} - c$	$\frac{1}{2}$
	$+1$	$\frac{1}{2} - c$	c	$\frac{1}{2}$
Sum		$\frac{1}{2}$	$\frac{1}{2}$	

$\mathbb{E}(X) = 0$
$\mathbb{E}(X^2) = (-1)^2 \times \frac{1}{2} + (+1)^2 \times \frac{1}{2} = 1$
$\mathrm{SD}(X) = \sqrt{\mathbb{E}(X^2) - (\mathbb{E}(X))^2} = \sqrt{1 - 0^2} = 1$

$\mathbb{E}(Y) = \mathbb{E}(X) = 0$, $\mathrm{SD}(Y) = \mathrm{SD}(X) = 1$.

$\mathbb{E}(XY) \;=\; (-1)^2 \times c \;+\; 2 \times (-1)(+1) \times (\frac{1}{2} - c) \;+\; (+1)^2 c \;=\; 4c - 1$
$\mathrm{Cov}(X, Y) \;=\; \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \;=\; 4c - 1$

$$\mathrm{Cor}(X, Y) \;=\; \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)} \;=\; 4c - 1$$

Example. Linear *deterministic* dependence

Suppose we have two random variables X, Y such that always $Y = a + bX$ (exactly!), and X has some distribution with mean $\mathbb{E}(X) = \mu$ and standard deviation $\mathrm{SD}(X) = \sigma$.

Calculate the correlation of X and Y.

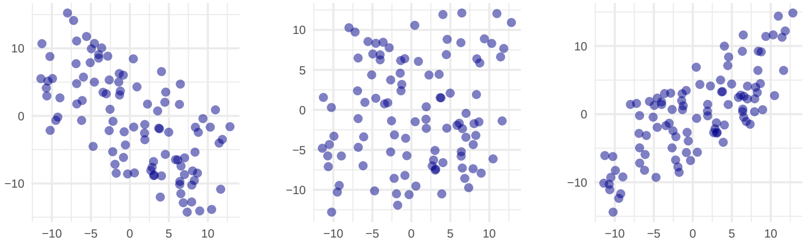
$$\mathrm{Cov}(X, Y) \;=\; \mathrm{Cov}(X, a + bX) \;=\; \mathrm{Cov}(X, a) + \mathrm{Cov}(X, bX) \;=\; b\,\mathrm{Var}(X).$$

$$\mathrm{SD}(Y) \;=\; \mathrm{SD}(a + bX) \;=\; |b|\,\mathrm{SD}(X)$$

$$\mathrm{Cor}(X, Y) \;=\; \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)} \;=\; \frac{b\,\mathrm{Var}(X)}{|b|\,\mathrm{SD}(X)^2} \;=\; \frac{b}{|b|}.$$

$$\mathrm{Cor}(X, Y) \;=\; \begin{cases} +1, & \text{if } b > 0, \\ 0, & \text{if } b = 0, \\ -1, & \text{if } b < 0. \end{cases}$$

(x, y) pairs drawn from some correlated distributions



$\rho = -0.60$ $\rho = 0.28$ $\rho = 0.80$

Next lecture is about sums of (many) random variables, and normal approximation...