MS-A0503 First course in probability and statistics

3A Distributions of sums and averages

Jukka Kohonen

Deparment of mathematics and systems analysis
Aalto SCI

Academic year 2019–2020
Period III

Contents

Sum of two random variables: Mean and SD

If X, Y are random variables, and $S = X + Y$, we already know how to calculate its

- mean: $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- variance: $\text{Var}(X+Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X,Y) + \text{Var}(Y)$
- standard deviation: $\sqrt{\text{Var}}$

That is, *location* and *width* of the distribution of $X + Y$.

But we still do not know the shape of the distribution. This may be very different from the distributions of X and Y. (Examples will follow.)

Knowing the shape would be useful for calculating good estimates of e.g. tail probabilities. (Chebyshev gives only loose bounds, recall last lecture. Knowing the shape is better.)

Sum of several random variables: Mean and SD

Before going to shapes, let's note that for a sum of *three* random variables, we can just apply the summation formulas recursively.

$$
\begin{aligned}
\mathbb{E}(X+Y+Z) &= \mathbb{E}((X+Y)+Z) = \mathbb{E}(X+Y)+\mathbb{E}(Z) \\
&= \mathbb{E}(X)+\mathbb{E}(Y)+\mathbb{E}(Z).
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{Var}(X+Y+Z) &= \mathrm{Var}(X)+\mathrm{Var}(Y)+\mathrm{Var}(Z) \\
&\quad + 2\,\mathrm{Cov}(X,Y)+2\,\mathrm{Cov}(X,Z)+2\,\mathrm{Cov}(Y,Z).
\end{aligned}
$$

In particular, if all variables are independent, then all covariances are zero, so

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

Generalization to more than 3 variables goes as you can expect.

Sum of two random variables: Shape

If X, Y are random variables, their sum $S = X + Y$ is also a random variable. Its distribution can be determined from the joint distribution $f_{X,Y}(x, y)$. How?

Like the distribution of any transformation $g(X, Y)$:

1. Study the joint distribution of (X, Y).
2. Find the possible values of $g(X, Y)$.
3. For each possible value s, find out, *which* values of the pair (X, Y) lead to $g(X, Y) = s$.
4. Add up their probabilities, to find $\mathbb{P}(g(X, Y) = s)$.

In step 3, one might really go through the possibilities (one by one), or try to find a general rule.

Example

The sum of two 100-sided dice S takes integer values $2 \ldots 200$. Let us find *all* of their probabilities (on blackboard).

Sum of two random variables: Shape

The distribution of $X + Y$ can be determined by summing over the "diagonals" of the joint distribution $f_{X,Y}(x, y)$.

$$f_{X+Y}(s) \;=\; \sum_x f_{X,Y}(x,\; s-x)$$

$$f_{X+Y}(s) \;=\; \int_{-\infty}^{\infty} f_{X,Y}(x,\; s-x)\, dx.$$

If X and Y are independent:

$$f_{X+Y}(s) \;=\; \sum_x f_X(x)\; f_Y(s-x)$$

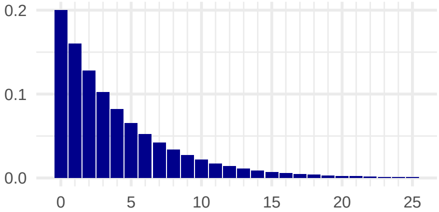
$$f_{X+Y}(s) \;=\; \int_{-\infty}^{\infty} f_X(x)\; f_Y(s-x)\, dx.$$

(This is called the *convolution* of the two distributions.)

Example: Sum of two geometric

Let X_1 and X_2 be independent, each following the geometric distribution with parameter $a = 4/5$, and density

$$f(x) \ = \ (1 - a)a^x.$$



Application: Roll a five-sided die until you get a five. The number of "failed" rolls has this geometric distribution.

Determine the distribution of $X_1 + X_2$.

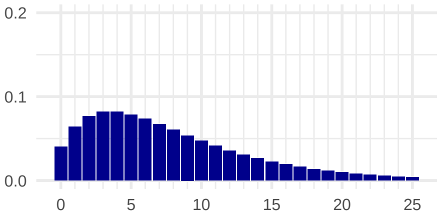
Example: Sum of two geometric

The possible values of $X_1 + X_2$ are $\{0, 1, 2, \dots\}$, and the density is obtained from

$$f_{X_1+X_2}(s) \;=\; \sum_x f(x)f(s-x) \;=\; \sum_{x=0}^{s}(1-a)a^x(1-a)a^{s-x}$$

The density of the sum is

$$f_{X_1+X_2}(s) \;=\; (1-a)^2(s+1)a^s$$



Application: Roll a five-sided die, until you have got *two* fives in total. The number of “failed” (non-five) rolls has this distribution, called *negative binomial distribution*.

Sum of several random variables: Shape

Let X, Y, Z be independent random variables.
What is the distribution of their sum $X + Y + Z$?

Apply the previous formula recursively.

- Let $U = X + Y$, and find f_U by the convolution formula.
- Let $S = U + Z$, and find f_S by the convolution formula.

This gives the exact distribution of a sum, but the repeated summations/integrals may be cumbersome.

In many cases the exact distribution is well known (so you may find it in the literature).

For example, the sum of *several* geometric random variables is again a known thing (negative binomial).

Contents

What does the Law of Large Numbers say?

If X_i are many independent numbers from the same distribution, with mean μ, then their average is with high probability

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx \mu.$$

The Law of Large Numbers *does not tell*

- How good is this approximation? (What is the probability?)
- Does the standard deviation of X play some role?

Some idea of the approximation is gained by the standard deviation

$$\mathrm{SD}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{SD}\left(\sum_{i=1}^{n} X_i\right).$$

So we need a formula for SD of a large sum.

Standard deviation of $X + Y$

Calculate $\sigma_{X+Y} = \mathrm{SD}(X+Y)$, when we know means $\mu_X = 1$ ja $\mu_Y = 1$ and standard deviations $\sigma_X = 2$ ja $\sigma_Y = 3$.

Solution

From the linearity of covariance,

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{Cov}(X+Y,\ X+Y) \\
&= \mathrm{Cov}(X,X) + \mathrm{Cov}(Y,X) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,Y) \\
&= \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y),
\end{aligned}$$

thus

$$\mathrm{SD}(X+Y) = \sqrt{\sigma_X^2 + 2\,\mathrm{Cor}(X,Y)\,\sigma_X\sigma_Y + \sigma_Y^2}.$$

We *cannot* calculate the SD of the sum without knowing the correlation.

- Because $-1 \le \mathrm{Cor}(X,Y) \le 1$, we do get bounds $|\sigma_X - \sigma_Y| \ \le\ \mathrm{SD}(X+Y) \ \le\ \sigma_X + \sigma_Y$, eli $1 \le \sigma_{X+Y} \le 5$.
- If X and Y are independent, then $\mathrm{Cor}(X,Y) = 0$ and $\sigma_{X+Y} \ = \ \sqrt{\sigma_X^2 + \sigma_Y^2} \ = \ \sqrt{13} \ \approx \ 3.6$.

Standard deviation of a long sum

Fact

If $X_1, \ldots, X_n$ are random variables, the standard deviation of their sum is

$$\mathrm{SD}\left(\sum_i X_i\right) = \sqrt{\sum_i \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j \rho_{i,j}},$$

where $\sigma_i = \mathrm{SD}(X_i)$ and $\rho_{i,j} = \mathrm{Cor}(X_i, X_j)$.

If $X_1, \ldots, X_n$ are independent (so $\rho_{i,j} = 0$) and identically distributed (so $\mu_i = \mu$ and $\sigma_i = \sigma$), we can simplify

$$\mathrm{SD}\left(\sum_{i=1}^n X_i\right) = \sqrt{\sum_{i=1}^n \sigma_i^2} = \sqrt{n\sigma^2} = \sigma\sqrt{n}.$$

Standard deviation of a long sum: Proof

From the linearity of covariance,

$$\begin{aligned}
\operatorname{Var}\left(\sum_i X_i\right) &= \operatorname{Cov}\left(\sum_i X_i, \sum_j X_j\right) \\
&= \sum_i \sum_j \operatorname{Cov}(X_i, X_j) \\
&= \sum_i \left(\operatorname{Cov}(X_i, X_i) + \sum_{j \neq i} \operatorname{Cov}(X_i, X_j) \right) \\
&= \sum_i \operatorname{Var}(X_i) + \sum_i \sum_{j \neq i} \operatorname{Cov}(X_i, X_j) \\
&= \sum_i \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j \rho_{i,j},
\end{aligned}$$

thus

$$\operatorname{SD}\left(\sum_i X_i\right) = \sqrt{\operatorname{Var}\left(\sum_i X_i\right)} = \sqrt{\sum_i \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j \rho_{i,j}}.$$

Standard deviation of a long sum, with independent terms

Fact

If $X_1, \ldots, X_n$ are independent and have the same standard deviation $\sigma = \sigma_i$ for all $i = 1, \ldots, n$, then

$$\mathrm{SD}\left(\sum_{i=1}^{n} X_i\right) = \sqrt{\sum_{i=1}^{n} \sigma_i^2} = \sigma\sqrt{n}.$$

Proof.

Follows from the previous slide, because (by independence) $\rho_{i,j} = \mathrm{Cor}(X_i, Y_j) = 0$ for all $i \neq j$. $\square$

Mean and SD of a sum: Summary

If $X_1, \ldots, X_n$ have means, standard deviations and correlations $\mu_i = \mathbb{E}(X_i)$, $\sigma_i = \mathrm{SD}(X_i)$ ja $\rho_{i,j} = \mathrm{Cor}(X_i, X_j)$, then:

If terms are	$\mathbb{E}(\sum_i X_i)$	$\mathrm{SD}(\sum_i X_i)$
Anything	$\sum_i \mu_i$	$\sqrt{\sum_i \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j \rho_{i,j}}$
Independent	$\sum_i \mu_i$	$\sqrt{\sum_i \sigma_i^2}$
Independent, same distribution	μn	$\sigma \sqrt{n}$

Example. Sum of many dice

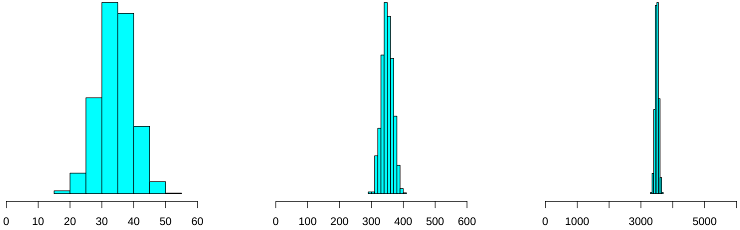
Play n rounds, gaining X_i (die result) on each round. Let us look at mean, std.dev. and distribution of total gains

$S_n = X_1 + \cdots + X_n$ for $n = 10, 100, 1000$.

Gain from one round has $\mu = 3.5$ and std.dev.

$$\sigma = \sqrt{\mathbb{E}(X_i^2) - \mu^2} = \sqrt{\tfrac{1}{6}(1^2 + \cdots + 6^2) - (3.5)^2} \approx 1.7.$$

Independent rounds $\implies \mathbb{E}(S_n) = \mu n$ ja $\mathrm{SD}(S_n) = \sigma\sqrt{n}$.



$\mathbb{E}(S_{10}) = 35$ $\quad$ $\mathbb{E}(S_{100}) = 350$ $\quad$ $\mathbb{E}(S_{1000}) = 3500$

$\mathrm{SD}(S_{10}) \approx 5.4$ $\quad$ $\mathrm{SD}(S_{100}) \approx 17$ $\quad$ $\mathrm{SD}(S_{1000}) \approx 54$

Another example. Sum of many indicator variables

300 tickets are sold for a flight that has 290 seats. We estimate that 5% of the passengers won't show up (independently). Probability that we can seat all passengers who show up?

Number of passengers showing up is $N = X_1 + \cdots + X_{300}$, where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th passenger shows up,} \\ 0, & \text{otherwise.} \end{cases}$$

Because $\mu_X = \mathbb{E}(X_i) = 0.95$ and $\sigma_X = \mathrm{SD}(X_i) = \sqrt{\mu_X(1-\mu_X)} \approx 0.22$, we get $\mu_N = \mu_X \times 300 = 285$ and $\sigma_N = \sigma_X \times \sqrt{300} \approx 3.8$.

From Chebyshev, we could have the bound

$$\mathbb{P}(N \in [280, 290]) \approx \mathbb{P}(N = \mu_N \pm 1.32\sigma_N) \geq 1 - \frac{1}{1.32^2} \approx 42.6\%.$$

So we have at least probability 42.6% of seating everybody.
However, we can do much better by looking at the distribution shape.

Sum of indicators: Exact distribution

What is the exact distribution of N, the number of passengers showing up?

$$N = X_1 + \cdots + X_{300}$$

The possible values of N are $\{0, 1, 2, \ldots, 300\}$.

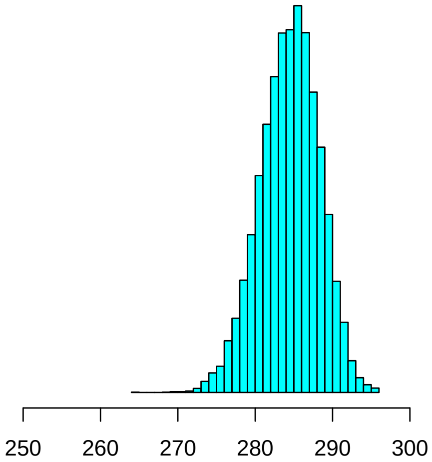
$$\mathbb{P}(N = 0) \;=\; (1 - 0.95)^{300} \;\leq\; 0.1^{300} \;=\; 10^{-300}$$

$$\mathbb{P}(N = k) \;=\; \binom{300}{k}(1 - 0.95)^{300-k}0.95^{k}$$

- N has the binomial distribution with parameters $n = 300$ and $p = 0.95$.
- We can simply calculate the individual densities and add them up.
- R does it for us:
  $\mathbb{P}(N \leq 290) \;=\;$ `pbinom(290,300,0.95)` $\approx\;$ 93.5% and
  $\mathbb{P}(N \in [280, 290]) \;=\;$
  `pbinom(290,300,0.95) - pbinom(279,300,0.95)` $\approx\;$ 85.7%.

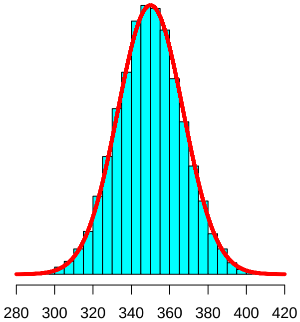
Sum of indicators: Simulated distribution

Let us simulate the numbers of show-up passengers (N) on 10 000 flights, that is, numbers from $\mathrm{Bin}(300, 0.95)$, and draw a histogram.

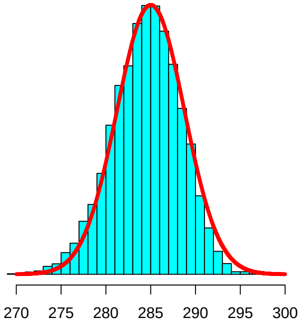


100 dice vs. 300 flight tickets

We observe: The distributions of these two random variables (sum of dice; and number of show-up passengers) seem to have the same shape, although different location and scale.



Sum of 100 independent dice



Sum of 300 independent indicators

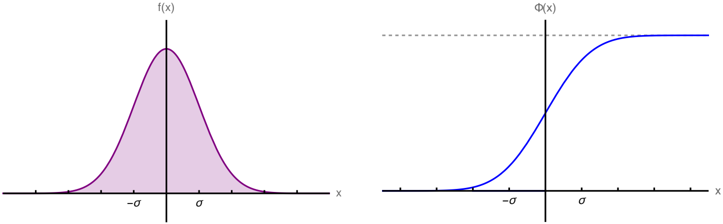
This is not a coincidence! Moreover, this holds more generally.

Contents

Standard normal distribution

Random variable Z has standard normal distribution (with mean $\mu = 0$ and standard deviation $\sigma = 1$), if it has density

$$f(t) \;=\; \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \;=\; \texttt{dnorm(x)}$$



Then its cumulative distribution function is

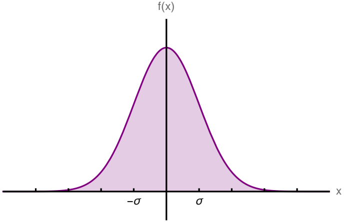
$$\Phi(z) = F_Z(z) \;=\; \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt \;=\; \texttt{pnorm(z)}$$

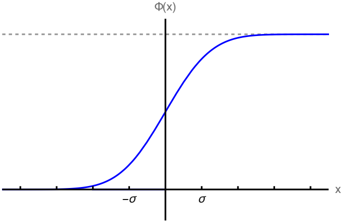
The integral is a bit cumbersome, but if you have z, you can look up $\Phi(z)$ from tables (see e.g. Ross or course page); or you can use a calculator or computer.

Normal distribution (general)

Random variable X has normal distribution with mean μ and standard deviation σ, if it has density

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} = \texttt{dnorm(x, mu, sigma)}$$





The general CDF is also easily calculated in R:
`pnorm(x, mu, sigma)`

But if you need to use tables, you can use scaling and shifting.

Normal distribution: Scaling and shifting

Fact

If Z has a standard normal distribution, and μ and $\sigma > 0$ are constants, then the transformation $X = \mu + \sigma Z$ also has a normal distribution.

Now *which* normal distribution does X have? Let us calculate its parameters:

$$\mathbb{E}(X) = \mathbb{E}(\mu + \sigma Z) = \mu + \sigma \mathbb{E}(Z) = \mu,$$
$$\mathrm{SD}(X) = \mathrm{SD}(\mu + \sigma Z) = \sigma \cdot sd(Z) = \sigma.$$

We can also go the other way:

Fact

If X has a normal distribution with parameters μ and σ, then $Z = (X - \mu)/\sigma$ has standard normal distribution.

This is called *standardization* of X, and useful for calculating the CDF $F_X(x)$.

Using standardization for CDF

If X is normal with parameters μ and σ, and then the transformation

$$Z = \frac{X - \mu}{\sigma}$$

has standard normal distribution.

Then

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) \\ &= \mathbb{P}(\mu + \sigma Z \leq x) \\ &= \mathbb{P}(\sigma Z \leq x - \mu) \\ &= \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) \\ &= F_Z\left(\frac{x - \mu}{\sigma}\right). \end{aligned}$$

The values of $F_z(\ldots)$, also denoted $\Phi(\ldots)$, can be looked up in tables, or calculated e.g. with R.

Finding CDF directly / by standardization

Let X be normally distributed with mean $\mu = 10$ and standard deviation $\sigma = 3$.

What is $F_X(16) = \mathbb{P}(X \leq 16)$, that is, the probability that X is at most two standard deviations (2σ) above its mean?

Method 1. Directly with R.

```
> pnorm(16,10,3)
[1] 0.9772499
```

Method 2. By standardization. Because $Z = (X - 10)/3$ has standard normal distribution, we calculate $F_Z((16 - 10)/3) = F_Z(2)$ by ...

```
> pnorm((16-10)/3)
[1] 0.9772499
```

Normal distribution: More useful facts

Fact
If X, Y are normally distributed random variables, and independent, then $S = X + Y$ is also normally distributed.

Fact
If X is a normally distributed random variable, then any scaling and shifting $Y = a + bX$ also has normal distribution.

In both cases, the *parameters* of the new distribution can be calculated by the already known formulas (linearity of mean and covariance).

Contents

Normal approximation

Fact (Central Limit Theorem, CLT)

If $X_1, \ldots, X_n$ are independent and identically distributed random variables, with same mean μ and same standard deviation σ, then their sum S has approximately a normal distribution, if n is large.

The parameters of the distribution we already know:

$$\mathbb{E}(S) = n\mu$$
$$\mathrm{SD}(S) = \sqrt{n}\sigma.$$

It follows that the average S/n also has a normal distribution.

Note

This is a universal law of nature: it holds *whatever distribution* the individual terms have (discrete/continuous, symmetric/skewed etc; recall sums of dice, and sums of indicators.) However, the independence of the terms is rather important (but there are variations of CLT).

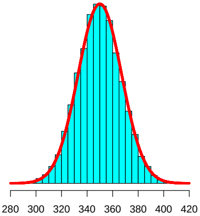
de Moivre 1733, Laplace 1812, Lyapunov 1911, Lindeberg 1922, Turing 1934

Example: Sum of dice, normal approximation

After 100 rounds, probability that gains are

(a) in the interval $[316, 384]$?

(b) over 500 EUR?



One round has $\mu_X = 3.5$ and $\sigma_X \approx 1.7$, so sum has $\mu_S = 350$ and $\sigma_S \approx 17$. Normal approximation

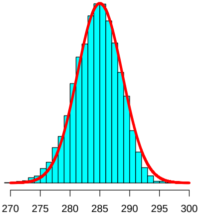
$$\frac{S-350}{17} \overset{\mathrm{d}}{\approx} Z.$$

$$\begin{aligned}\mathbb{P}(316 \le S \le 384) &= \mathbb{P}\left(-2 \le \frac{S-350}{17} \le 2\right) \\ &\approx \mathbb{P}(-2 \le Z \le 2) = 1 - 2\mathbb{P}(Z \le -2) \approx 95.4\%.\end{aligned}$$

$$\begin{aligned}\mathbb{P}(S_{100} > 500) &= \mathbb{P}\left(\frac{S-350}{17} > 8.82\right) \\ &\approx \mathbb{P}(Z > 8.82) = \mathbb{P}(Z \le -8.82) \approx 6 \times 10^{-19}.\end{aligned}$$

Example: Sum of indicators, normal approximation

Probability that we can seat everybody? (Sold 300 tickets, but 290 seats.)



Number of passengers showing up $N = X_1 + \cdots + X_{300}$. Each term X_i has $\mu_X = 0.95$ and $\sigma_X = 0.218$, so sum has $\mu_N = 285$ and $\sigma_S = 3.77$. Normal approximation:

$$\frac{N - 285}{3.77} \overset{\mathrm{d}}{\approx} Z.$$

$$\begin{aligned}
\mathbb{P}(N \le 290) \;=\; \mathbb{P}(N \le 290.5) \;&=\; \mathbb{P}\left(\frac{N-285}{3.77} \le 1.46\right) \\
&\approx\; \mathbb{P}(Z \le 1.46) \\
&=\; 1 - \mathbb{P}(Z \le -1.46) \;\approx\; 92.8\%.
\end{aligned}$$

(Exact prob was: `pbinom(290,300,0.95)` $= 93.5\%$)

Contents

Sum of exponentially distributed r.v.

Driving a car, flies hit windscreen with rate $\lambda = 1/100$ (one fly in 100 seconds), randomly and independently.

Let $X_i \sim \text{Exp}(\lambda)$ be the waiting time for the ith fly (after the previous fly), or for the first fly if $i = 1$. The individual waiting times have the *exponential distribution*.

The waiting time for n flies, $S = X_1 + X_2 + \ldots + X_n$, does *not* have exponential distribution. Try the following R code with, e.g. $n = 2$, $n = 5$ or $n = 50$.

```
rate    <- 1/100
repeats <- 1000000
n       <- 5
X       <- matrix(rexp(repeats*n, rate), repeats, n)
S       <- rowSums(X)
hist(S,100)
```

Next lecture is about empirical distributions in observed data...