Everything you always wanted to know about statistics*

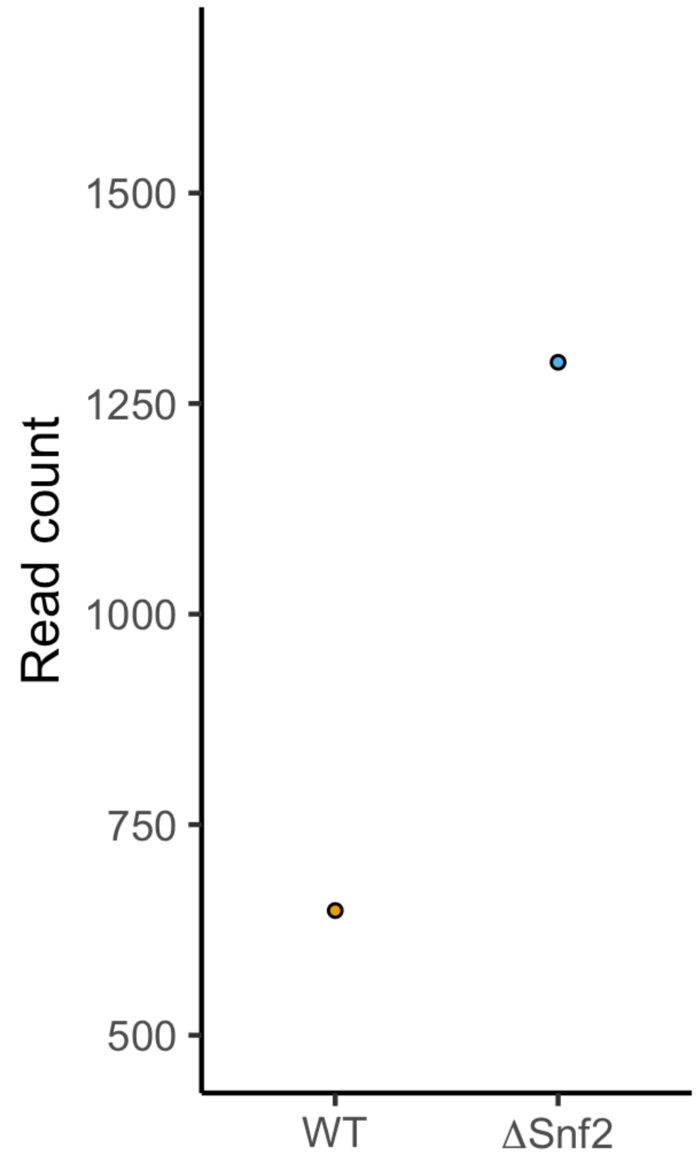
Marek Gierliński
Division of Computational Biology

Hand-outs available at http://is.gd/statlec

*but were afraid to ask

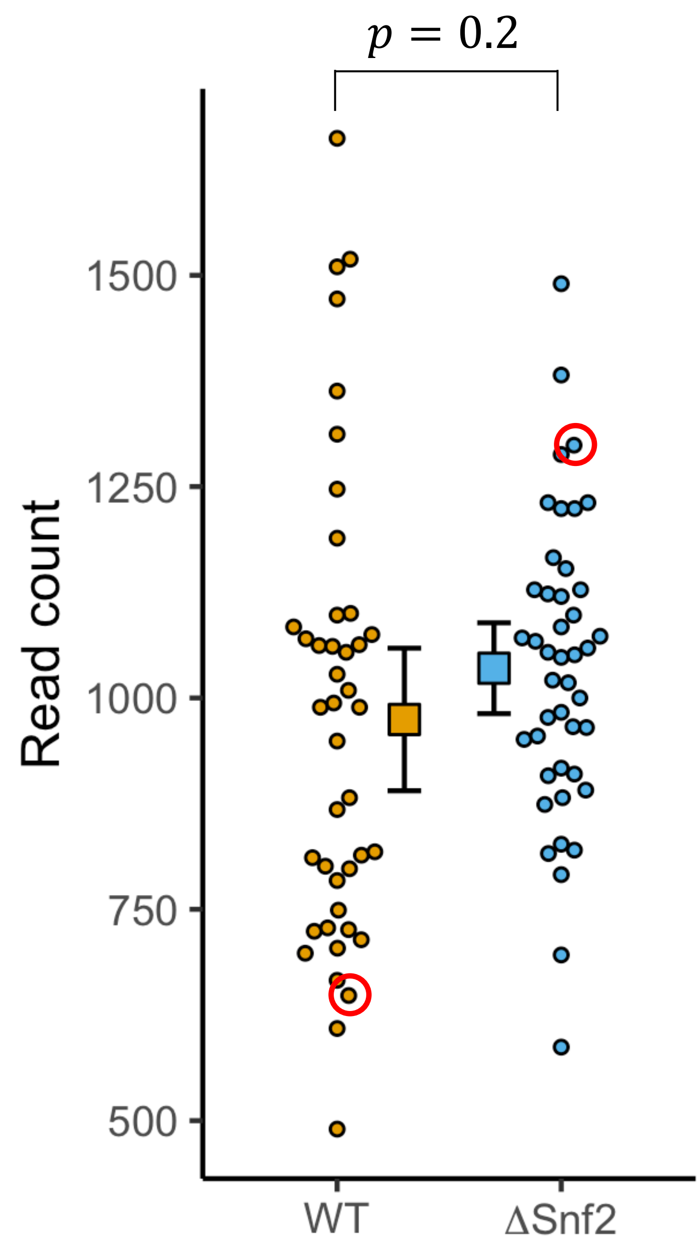
Why do we need statistics?

- Consider an RNA-seq experiment
- Comparing wild type and knock-out
- Expression level of gene IGD1
  - WT = 648
  - ΔSnf2 = 1299
- There is a 2-fold change in intensity
- Great! Gene is upregulated!



Why do we need statistics?

- Consider an RNA-seq experiment
- Comparing wild type and knock-out
- Expression level of gene IGD1
  - WT = 648
  - ΔSnf2 = 1299
- There is a 2-fold change in intensity
- Great! Gene is upregulated!
- Repeat the experiment in 42/44 replicates
  - WT = 975 ± 84
  - ΔSnf2 = 1035 ± 54
- Reveal **variability** of expression
- No difference between WT and knock-out



We collaborate on various types of projects

Anything involving data analysis

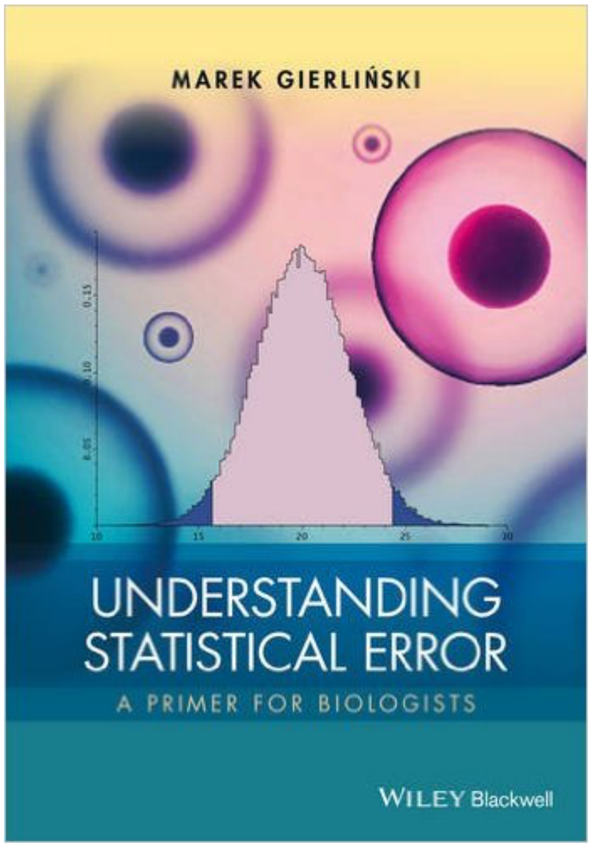
Marek Gierliński

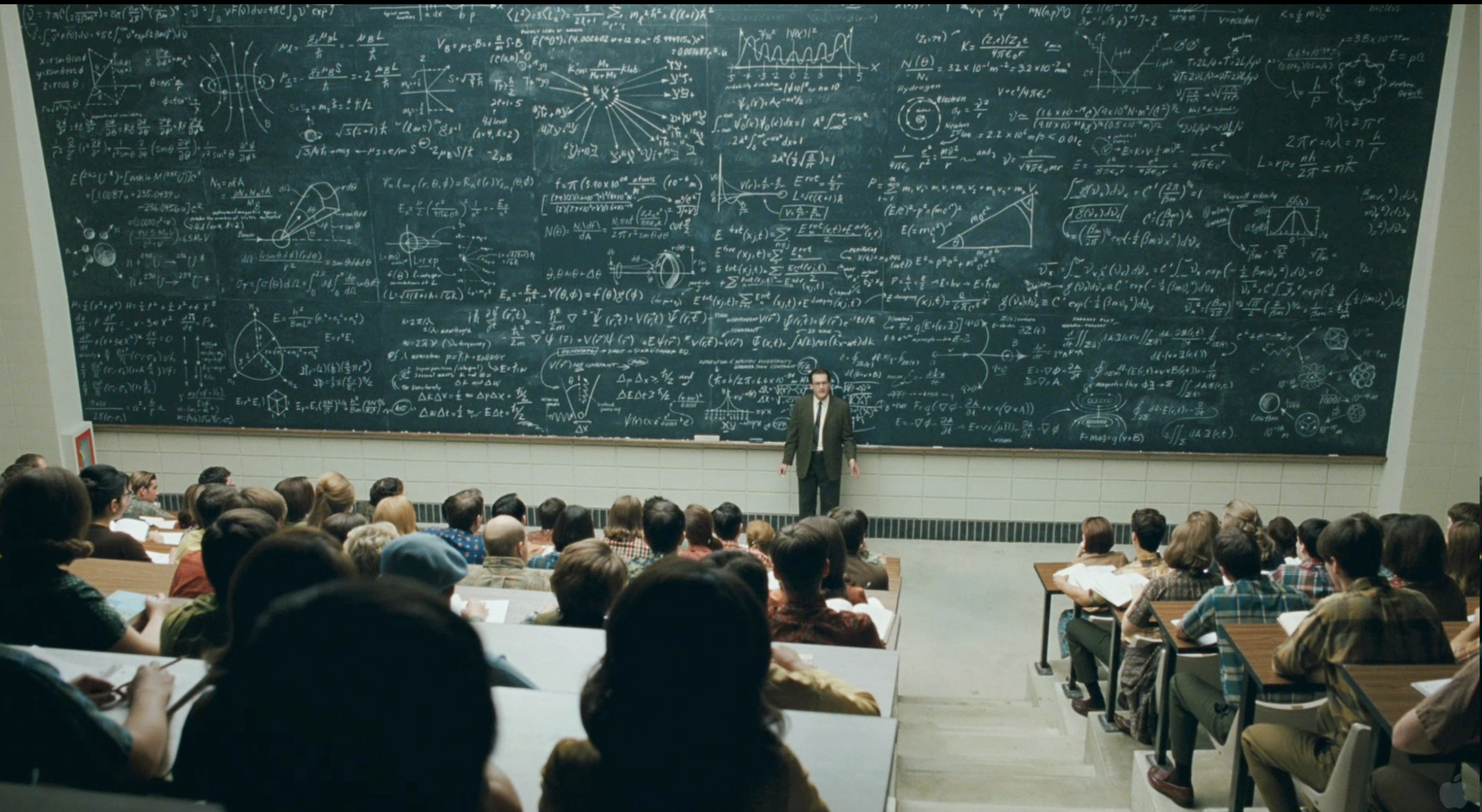
James Abbott

http://www.compbio.dundee.ac.uk/dag.html

Course materials

- Lecture slides available (one day before each lecture) at http://is.gd/statlec
- "Understanding statistical error: a primer for biologists", Wiley





1. Probability distributions

Random variables
Normal, log-normal, Poisson, Binomial

2. Errors and statistical estimators

Measurement and random errors
Population and sample
Standard deviation, standard error

3. Confidence intervals 1

Sampling distribution
Confidence interval of the mean, median

4. Confidence intervals 2

Confidence interval of count data, correlation, proportion

5. Data presentation

How to make a good plot

6. Introduction to p-values

Null hypothesis, statistical test, p-value
Fisher's test

7. Contingency tables

Chi-square test
G-test

8. T-test

One- and two-sample, paired
One-sample variance test

9. ANOVA

One-way
Two-way

10. Non-parametric methods

Mann-Whitney
Wilcoxon signed-rank
Kruskal-Wallis
Kolmogorov-Smirnov

11. Statistical power

Effect size
Power in t-test
Power in ANOVA

12. Multiple test corrections

Family-wise error rate
False discovery rate
Holm-Bonferroni limit
Benjamini-Hochberg limit

13. What's wrong with p-values?

A lot

1. Probability distributions

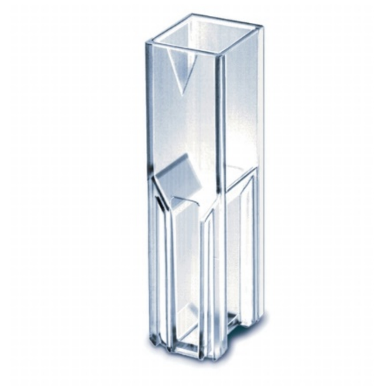
"Misunderstanding of probability may be the greatest of all general impediments to scientific literacy"

Stephen Jay Gould

Hand-outs available at http://is.gd/statlec

Example

- Experiment: estimate bacterial concentration using a spectrophotometer



- 6 replicates
- Find the following OD600

  0.37 0.34 0.41 0.40 0.30 0.33

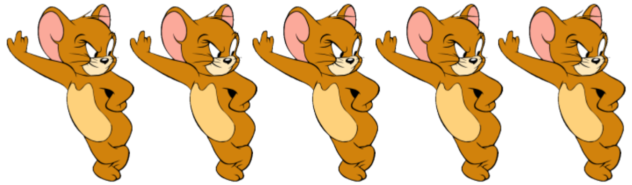
- Experimental result is a **random variable**
- It follows a certain **probability distribution**

Random variable: random numbers

Discrete and continuous random variables

- Discrete variables:
  - sum of 2 dice (2, 3, 4, …, 12)
  - categorical outcome
  - number of mice (5, non random?)
  - number of mice in survival experiment (random)

- Continuous variables:
  - weight of a mouse
  - height of a person
  - fluorescent marker luminosity
  - protein abundance

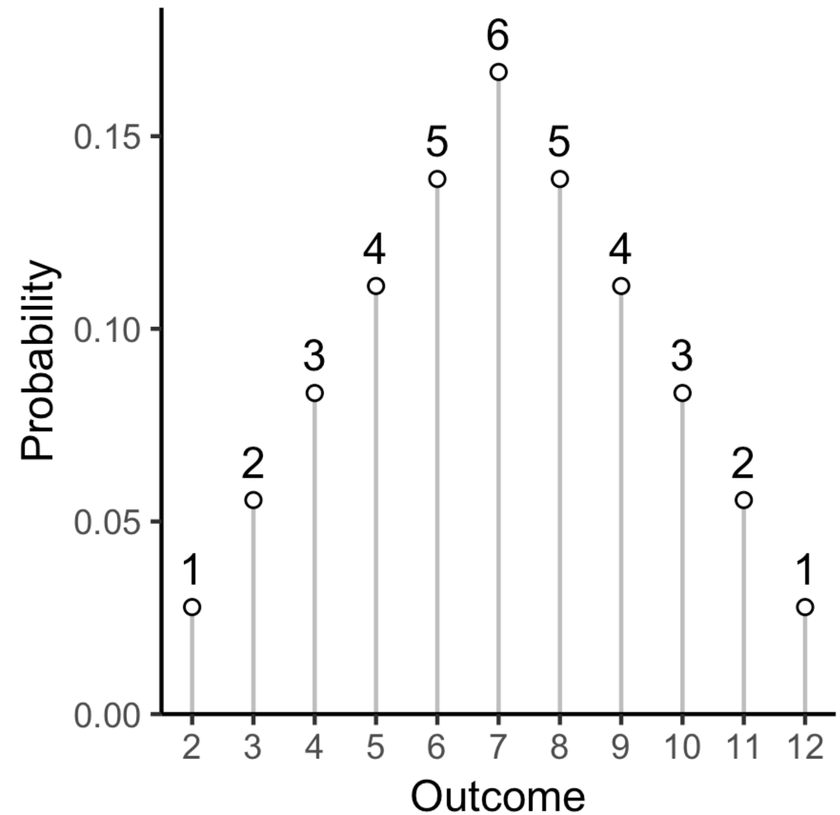


Probability distribution (2 dice)

- Assigns a probability to each of the possible outcomes
- Throwing 2 dice

Outcome	Combinations
2	1+1
3	1+2, 2+1
4	1+3, 2+2, 3+1
5	1+4, 2+3, 3+2, 4+1
6	1+5, 2+4, 3+3, 4+2, 5+1
7	1+6, 2+5, 3+4, 4+3, 5+2, 6+1
8	2+6, 3+5, 4+4, 5+3, 6+2
9	3+6, 4+5, 5+4, 6+3
10	4+6, 5+5, 6+4
11	5+6, 6+5
12	6+6

There are 36 combinations possible



Discrete random variable

probability

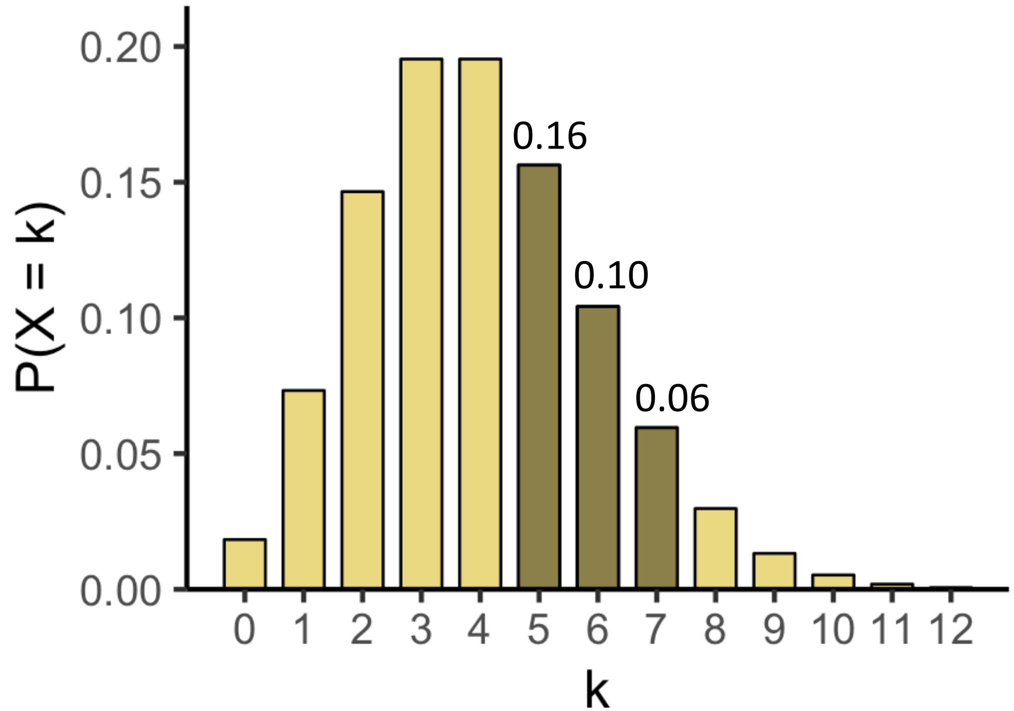
$$P(X = k)$$

random variable

outcome

$P(X = 6) = 0.10$

$P(5 \leq X \leq 7) = 0.32$



Continuous random variable

probability

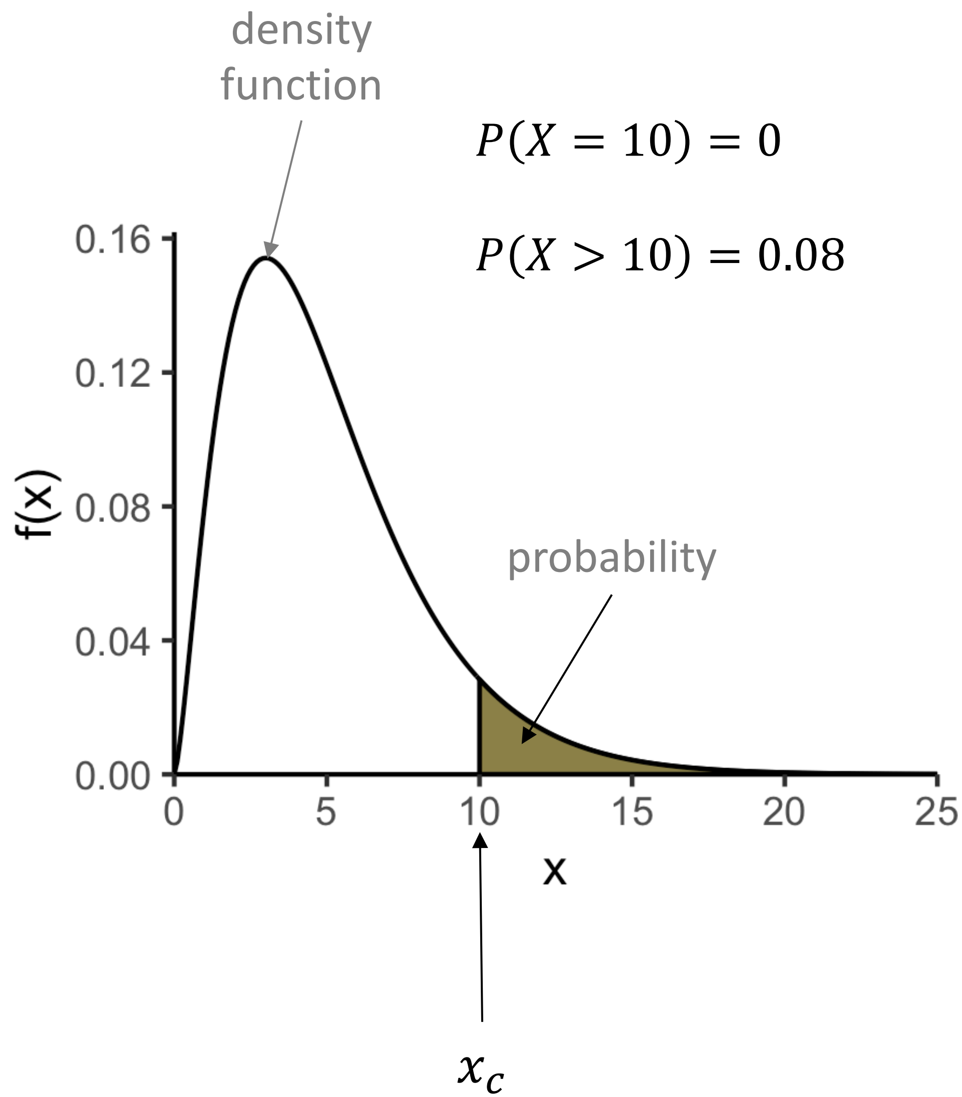
density function

$$P(X > x_c) = \int_{x_c}^{\infty} f(x)dx$$

random variable

limit

area under the curve



Normal distribution

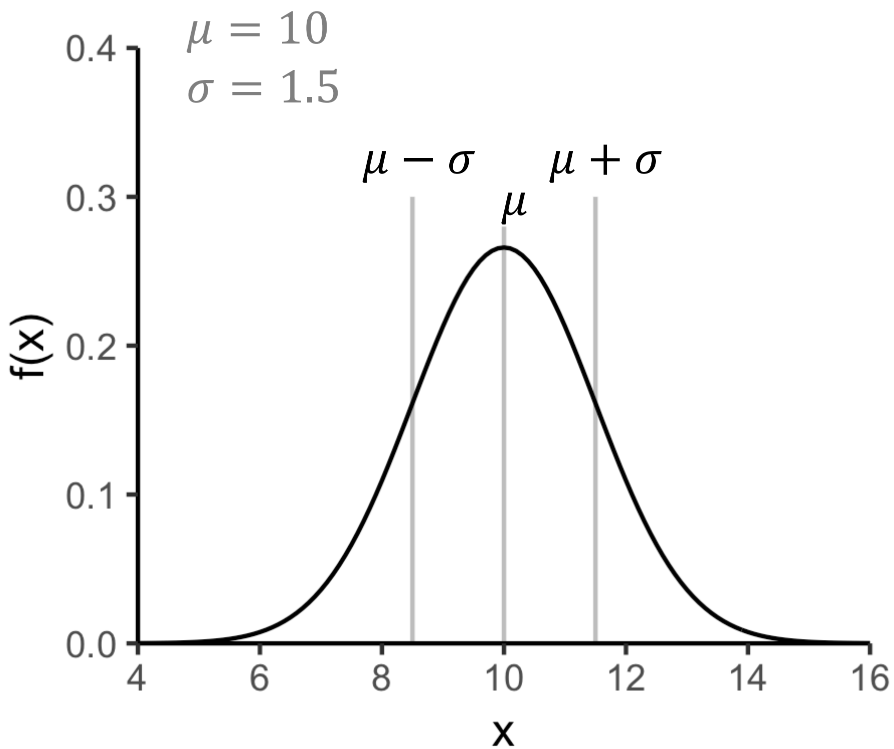
Normal distribution

- Normal (or Gaussian) probability distribution

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

  - μ - mean
  - σ - standard deviation
  - σ^2 - variance

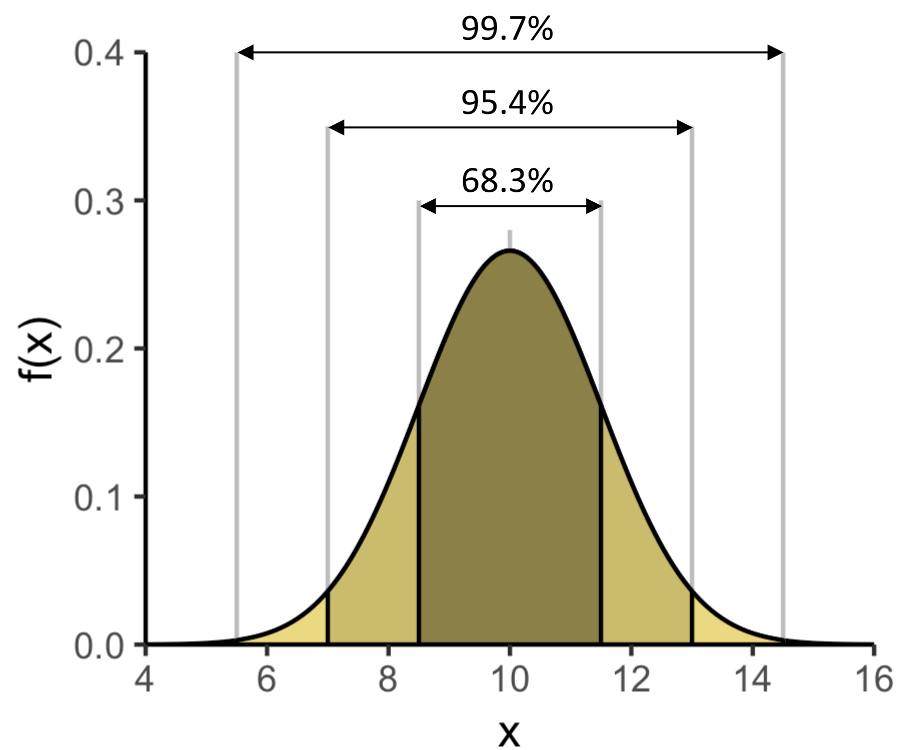
- It is called "normal" as it often appears in nature



Normal distribution: a few numbers

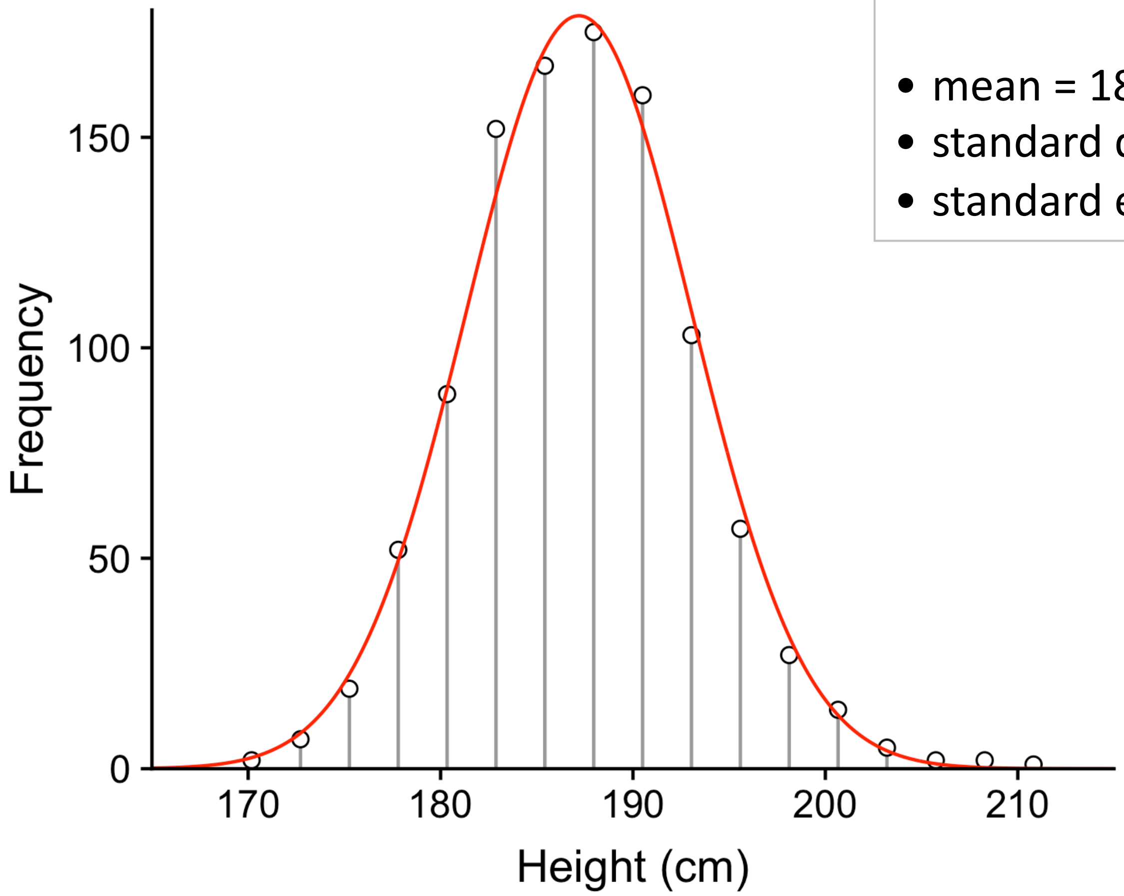
- Area under the curve = probability
- Probability within one sigma of the mean is about ⅔ (68.3%)
- 95% confidence intervals are traditionally used: correspond to about 1.96σ

	In	Out	Odds of out
±1σ	68.3%	31.7%	1:3
±2σ	95.4%	4.6%	1:20
±3σ	99.7%	0.3%	1:400
±4σ	99.994%	0.006%	1:16,000
±5σ	99.99993%	0.00007%	1:1,700,000
±1.96σ	95.0%	5.0%	1:20



$\mu = 10$
$\sigma = 1.5$

Example: normal distribution



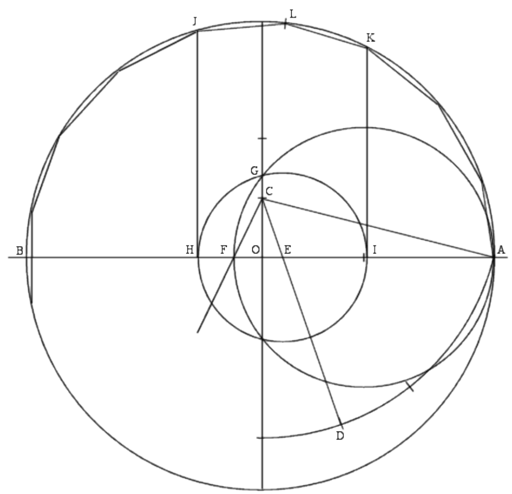
Height of 1034 Major League baseball players

- mean = 187.2 cm
- standard deviation = 5.9 cm
- standard error = 0.2 cm

http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_MLB_HeightsWeights

Carl Friedrich Gauss (1777-1855)

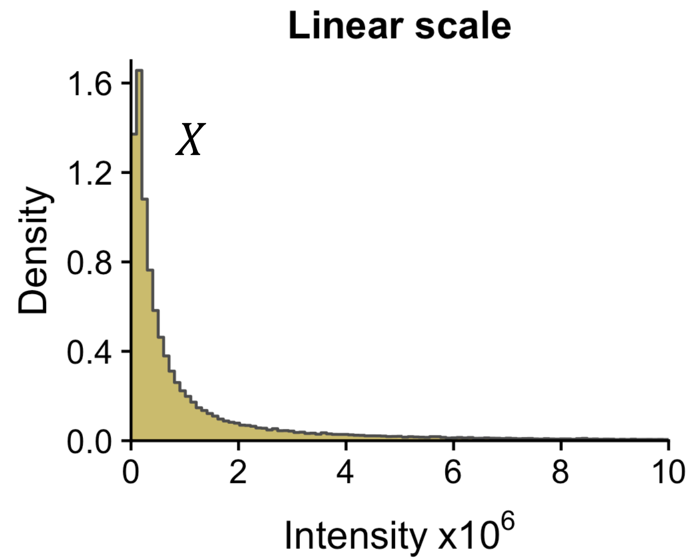
- Brilliant German mathematician
- Constructed a regular heptadecagon with a ruler and a compass
- He requested that a regular heptadecagon should be inscribed on his tombstone
- However, it was Abraham de Moivre (1667-1754) who first formulated “Gaussian” distribution

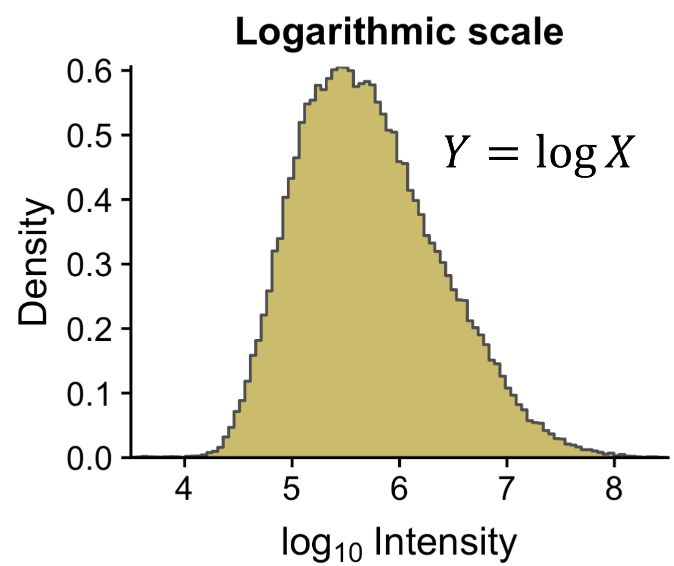


Log-normal distribution

- Probability distribution of a random variable whose logarithm is normally distributed

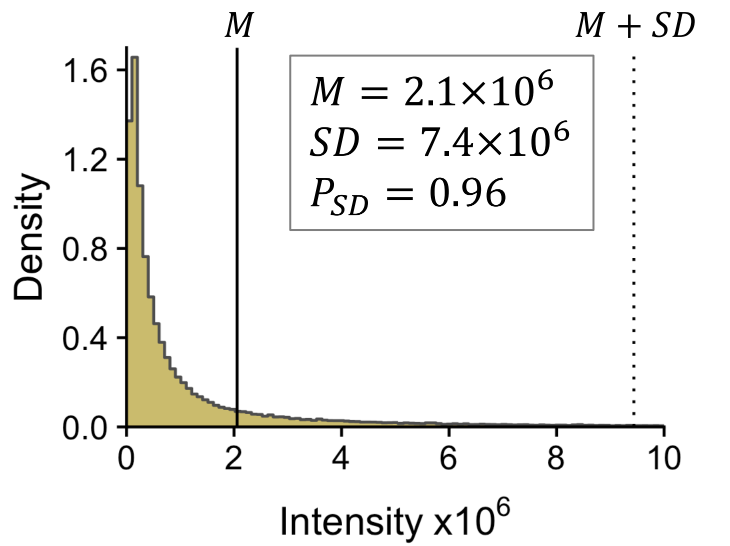
- Log-normal distribution can be very asymmetric!

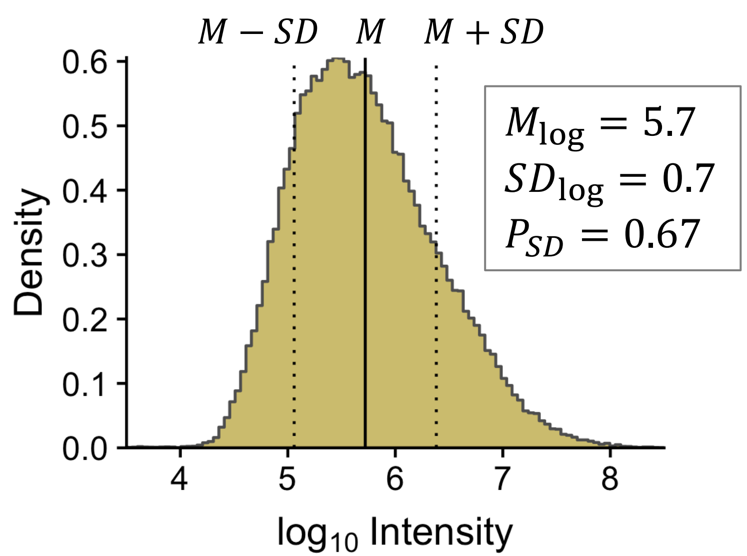




Example: log-normal distribution

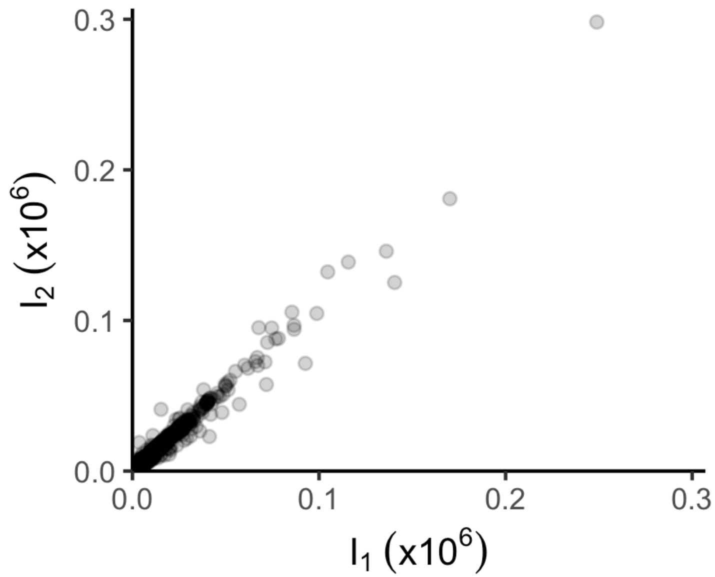
- Peptide intensities from a mass spectrometry experiment
- P_{SD} - fraction of data within $M \pm SD$
- Data look better in logarithmic space
- Always plot the distribution of your data before analysis
- About two-thirds of data points are within one standard deviation from the mean **only** when their distribution is approximately Gaussian





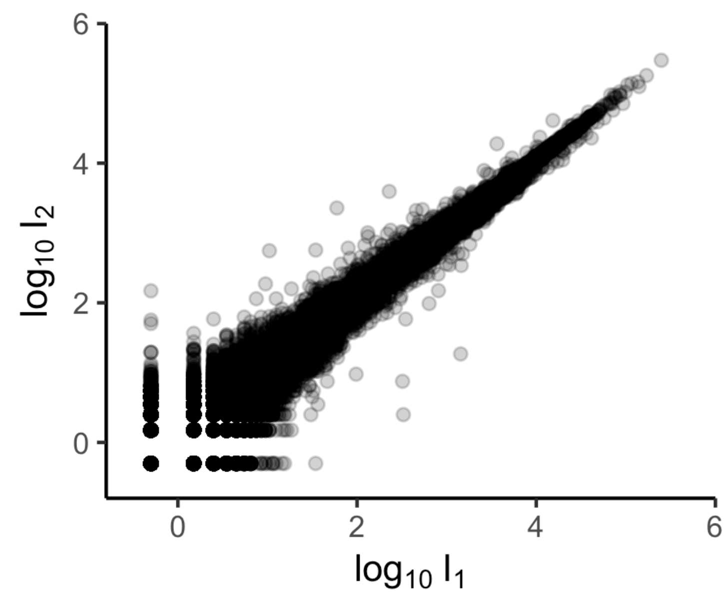
A few notes on log-normal distribution

- Examples of log-normal distributions
  - gene expression (RNA-seq, microarrays)
  - mass spectrometry data
  - drug potency IC_{50}
- Plot these data in logarithmic scale!



- It doesn't matter if you use $\log_2$, $\log_{10}$ or $\ln$, as long as you are consistent

- $\log_{10}$ is easier to understand in plots
  - $10^5 = 100{,}000$
  - $2^{10} = 1024$



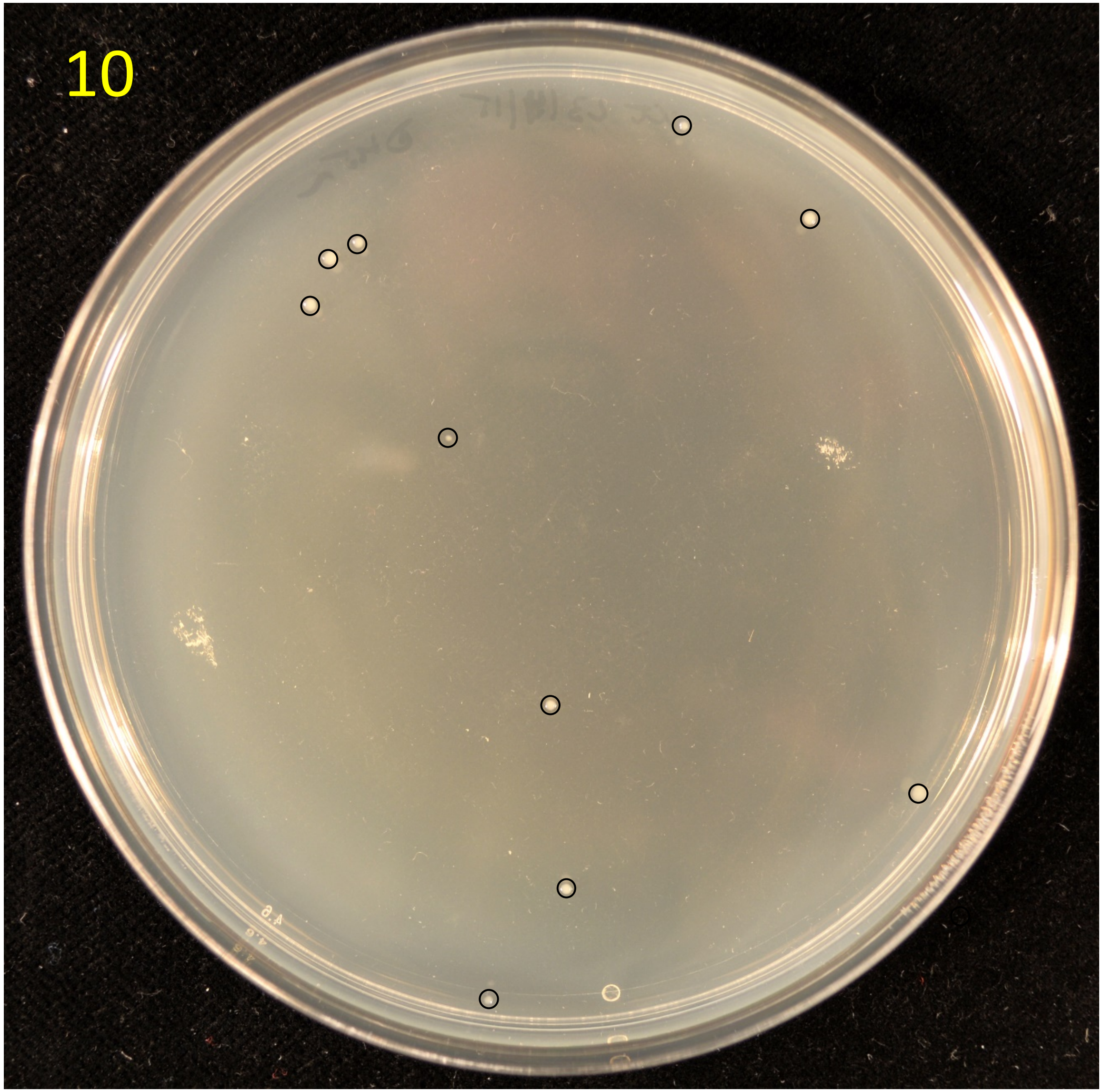
John Napier (1550-1617)

- Scottish mathematician and astronomer
- Invented logarithms and published first tables of natural logarithms
- Created “Napier’s bones”, the first practical calculator
- Had an interest in theology, calculated the date of the end of the world between 1688 and 1700
- Apparently involved in alchemy and necromancy

Merchiston Castle, Edinburgh

Poisson distribution

Counting bacterial colonies



Courtesy of Katharina Trunk

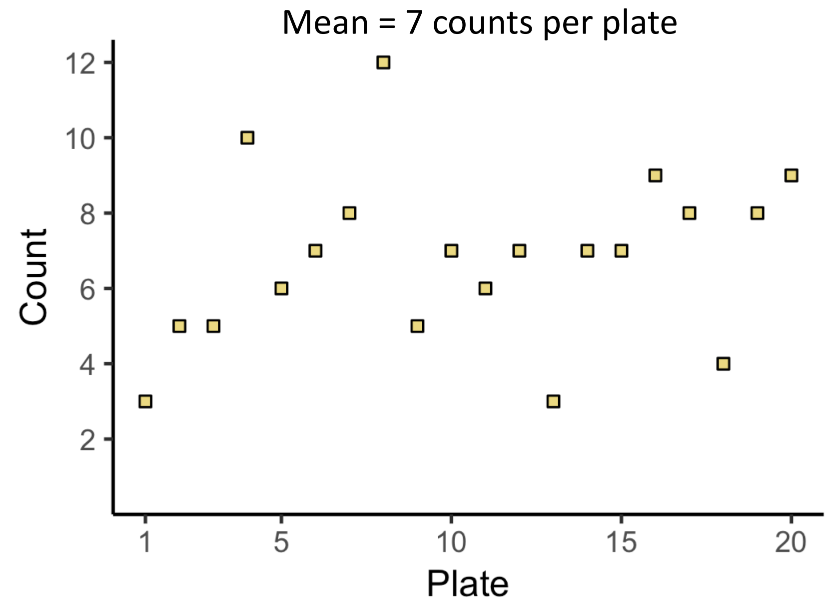
100 µl of 10^{-7} dilution of OD_{600} = 2.0

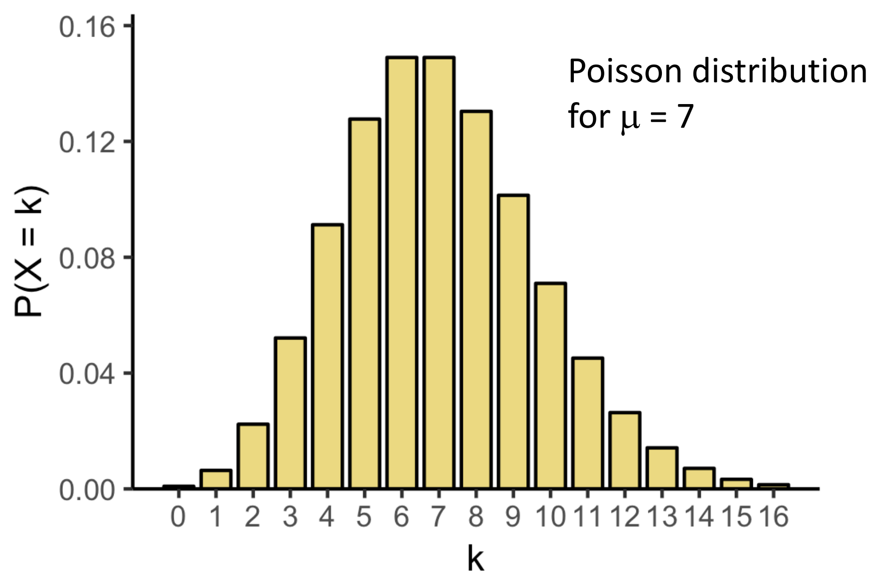
Poisson distribution

- Measure of bacterial count per unit volume

- Poisson count: always per bin

- This applies to any counts in time or space
  - radioactive decays per second
  - number of deaths in a population
  - number of cells in a counting chamber
  - number of mutations in a DNA fragment





Poisson distribution

- *Random* and *independent* events
- Probability of observing exactly k events:

$$P(X = k) = \frac{\mu^k e^{-\mu}}{k!}$$

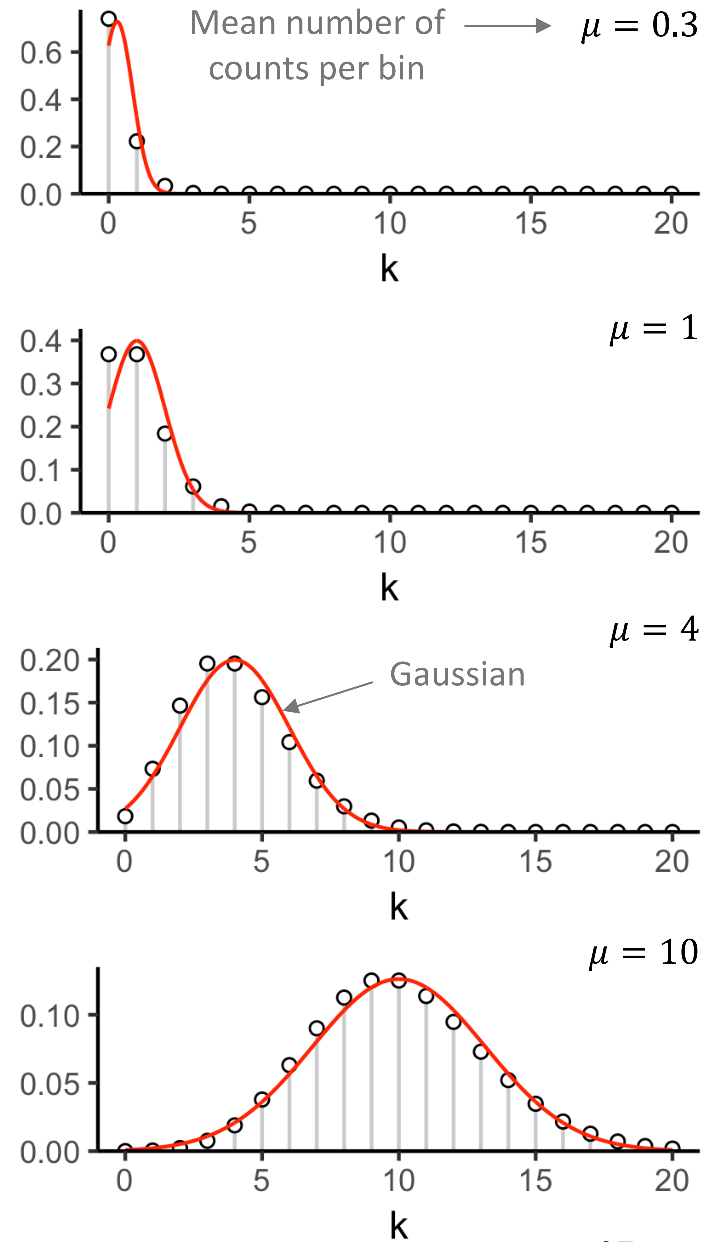
- One parameter: mean count rate, μ
- Standard deviation:

$$\sigma = \sqrt{\mu}$$

$$\sigma^2 = \mu$$

- For large μ Poisson distribution approximates Gaussian
- Example, $\mu = 4$:

$$P(X = 2) = \frac{4^2 e^{-4}}{2!} = \frac{16 \times 0.0183}{2} = 0.147$$



Classic example: horse kicks

- Ladislaus von Bortkiewicz (1898) “*Das Gesetz der kleinen Zahlen*”
- Number of soldiers in the Prussian army killed by horse kicks
  - 14 army corps, 20 years of data
  - Deaths per year per army corps

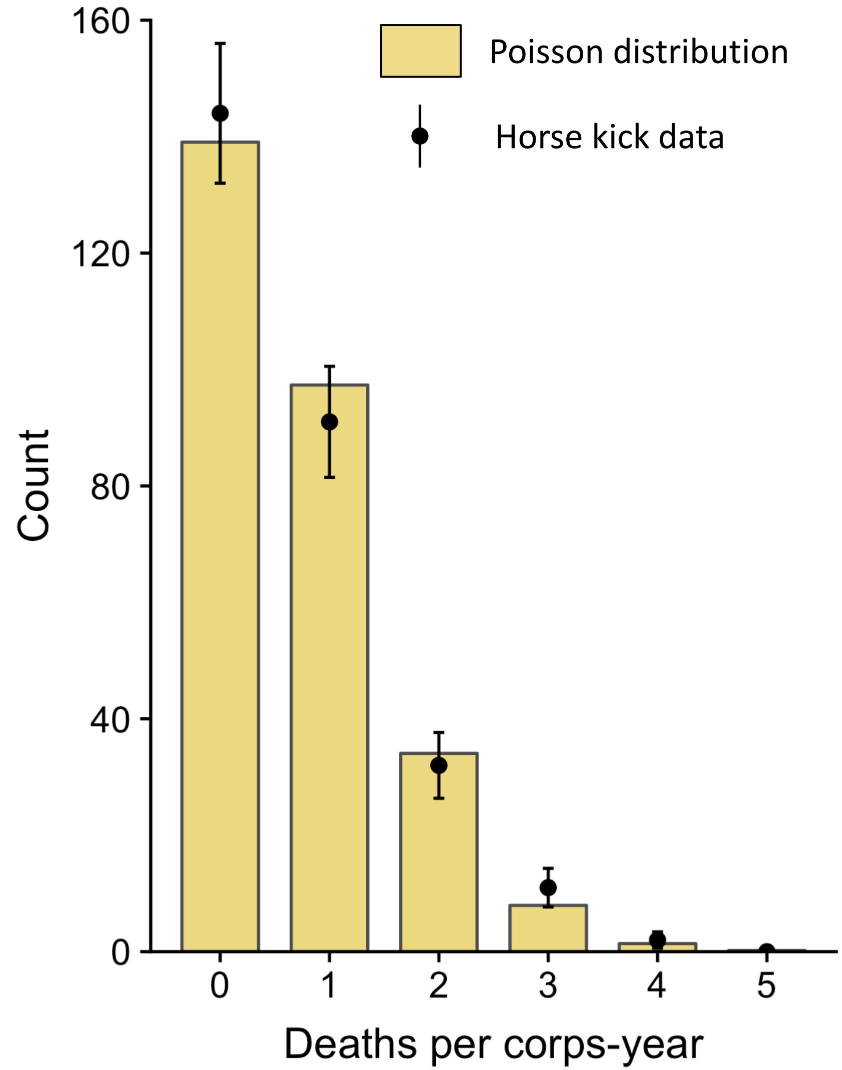
In nachstehender Tabelle sind die Zahlen der durch Schlag eines Pferdes verunglückten Militärpersonen, nach Armeecorps („G.“ bedeutet Gardecorps) und Kalenderjahren nachgewiesen.[1])

	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
G	—	2	2	1	—	—	1	1	—	3	—	2	1	—	—	1	—	1	—	1
I	—	—	—	2	—	3	—	2	—	—	—	1	1	1	—	2	—	3	1	—
II	—	—	—	2	—	2	—	—	1	1	—	—	2	1	1	—	—	2	—	—
III	—	—	—	1	1	1	2	—	2	—	—	—	1	—	1	2	1	—	—	—
IV	—	1	—	1	1	1	1	—	—	—	—	1	—	—	—	—	1	1	—	—
V	—	—	—	—	2	1	—	—	1	—	—	1	—	1	1	1	1	1	1	—
VI	—	—	1	—	2	—	—	1	2	—	1	1	3	1	1	1	—	3	—	—
VII	1	—	1	—	—	—	1	—	1	1	—	—	2	—	—	2	1	—	2	—
VIII	1	—	—	—	1	—	—	1	—	—	—	—	1	—	—	—	1	1	—	1
IX	—	—	—	—	—	2	1	1	1	—	2	1	1	—	1	2	—	1	—	—
X	—	—	1	1	—	1	—	2	—	2	—	—	—	—	2	1	3	—	1	1
XI	—	—	—	—	2	4	—	1	3	—	1	1	1	1	2	1	3	1	3	1
XIV	1	1	2	1	1	3	—	4	—	1	—	3	2	1	—	2	1	1	—	—
XV	—	1	—	—	—	—	—	1	—	1	1	—	—	—	2	2	—	—	—	—

Example: Poisson distribution

- Death distribution follows Poisson law
- mean = 0.70 deaths / corps / year

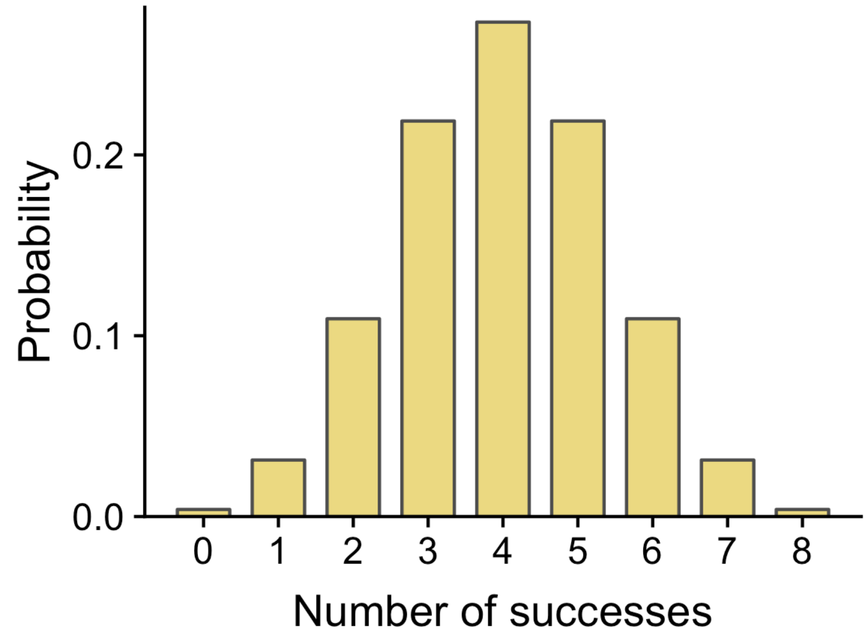
- 4 deaths in a corps-year are expected to happen from time to time!
- $P(X = 4) = 0.078$ in 14 corps
- On average it should happen once in 13 years



Binomial distribution

Binomial distribution

- A series of n "trials"
- In each trial, the probability of:
  - "success" $= p$
  - "failure" $= 1 - p$
- What is the probability of having exactly k successes in n trials?



- Applications:
  - random errors
  - error of the proportion
  - error of the median

Example: toss a coin
heads = success ($p = 0.5$)
tails = failure ($1 - p = 0.5$)

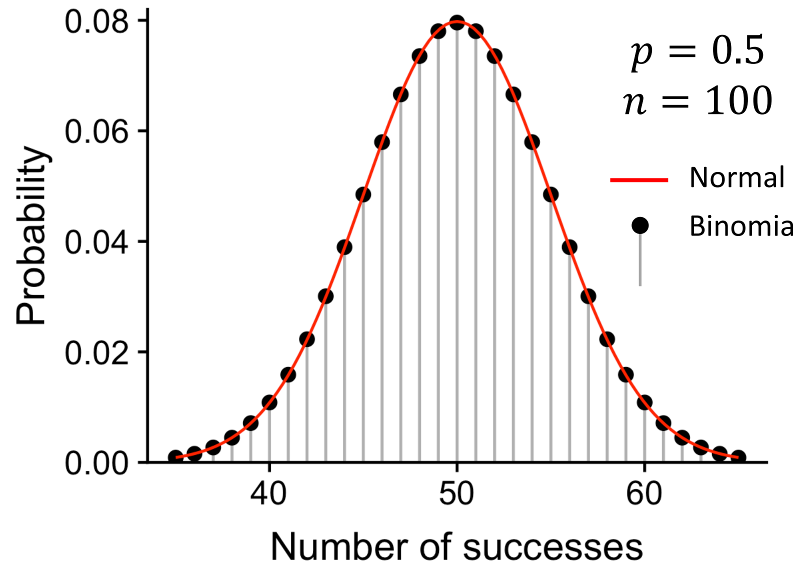
Probability of getting k heads from 8 coins

Binomial distribution

- Mean and standard deviation

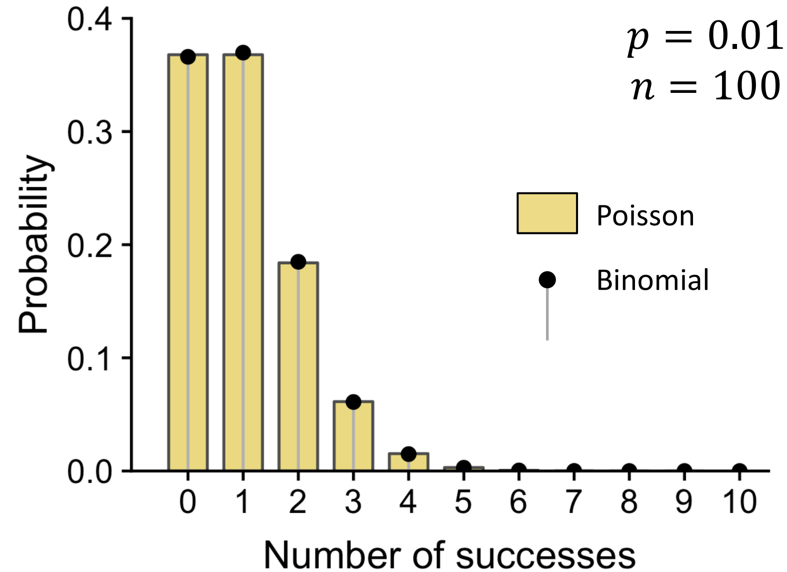
$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$



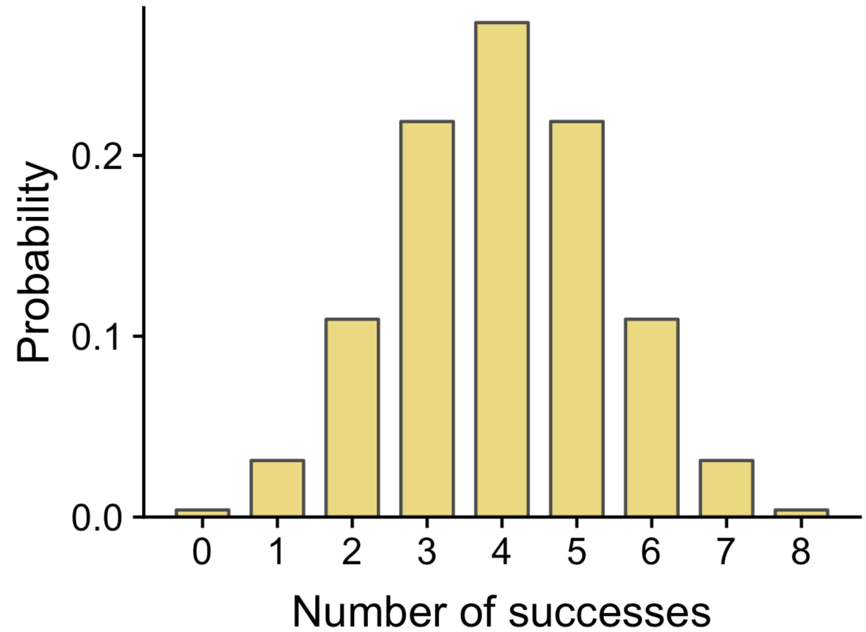
- For large n can be approximated by normal distribution

- For large n and small p it becomes Poisson



Example: tossing a coin

- Toss 8 coins
- Question: why is the probability having heads 4 times much larger than the probability of heads 8 times?



Example: toss a coin
heads = success ($p = 0.5$)
tails = failure ($1 - p = 0.5$)

What is the probability of obtaining heads k times from 8 coins?

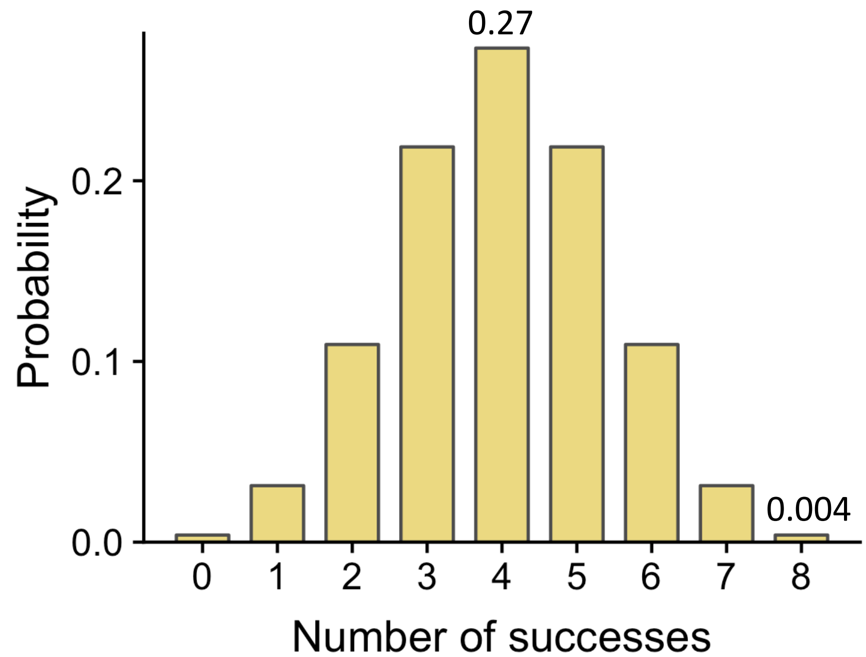
Example: tossing a coin

- There is only one way of having heads 8 times

- There many are ways of getting 4 heads and 4 tails

…

$$\binom{8}{4} = 70$$

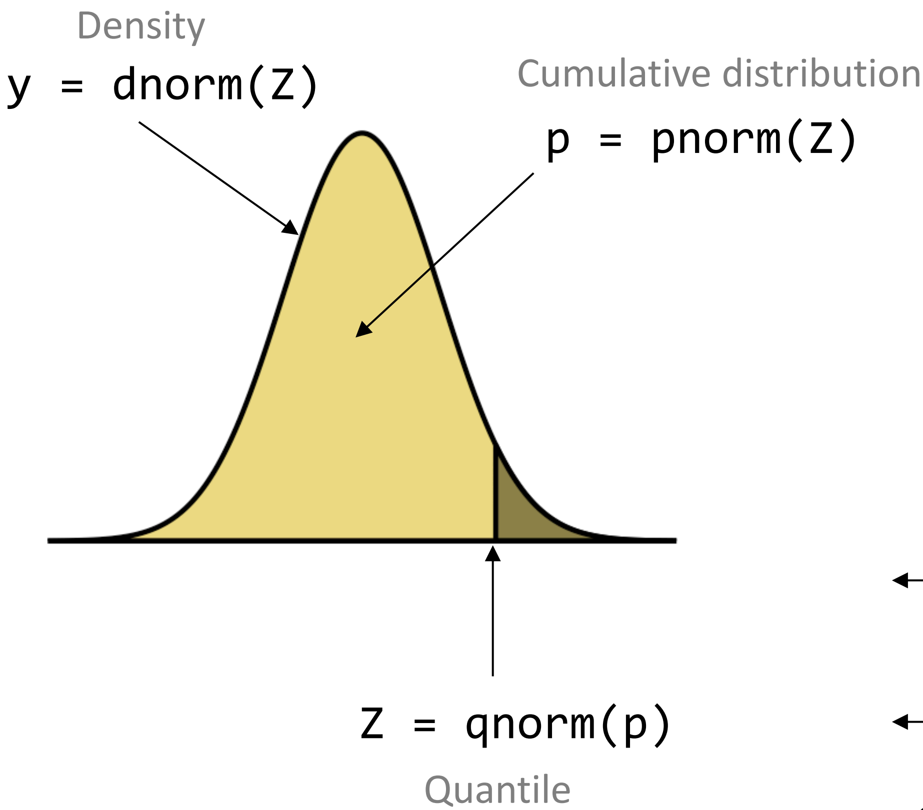


Example: toss a coin
heads = success ($p = 0.5$)
tails = failure ($1 - p = 0.5$)

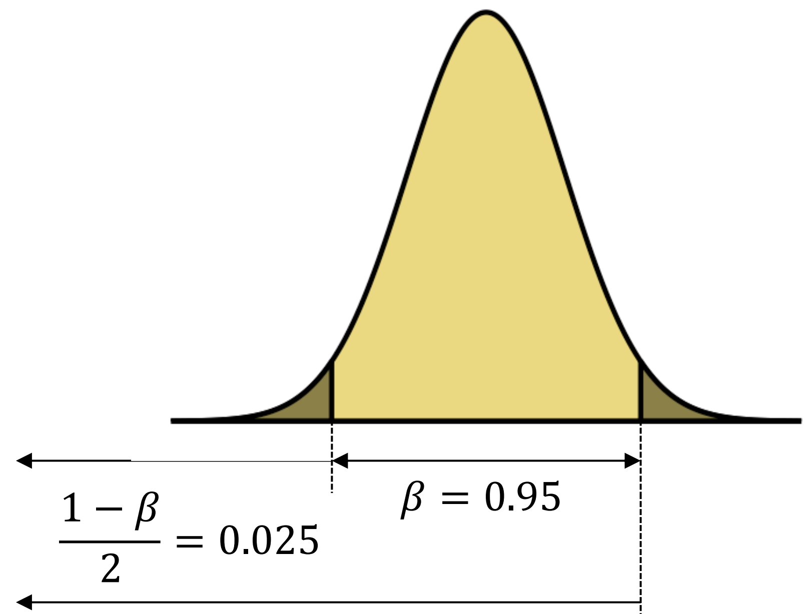
What is the probability of obtaining heads k times from 8 coins?

Probability distributions in R

Probability distributions in R



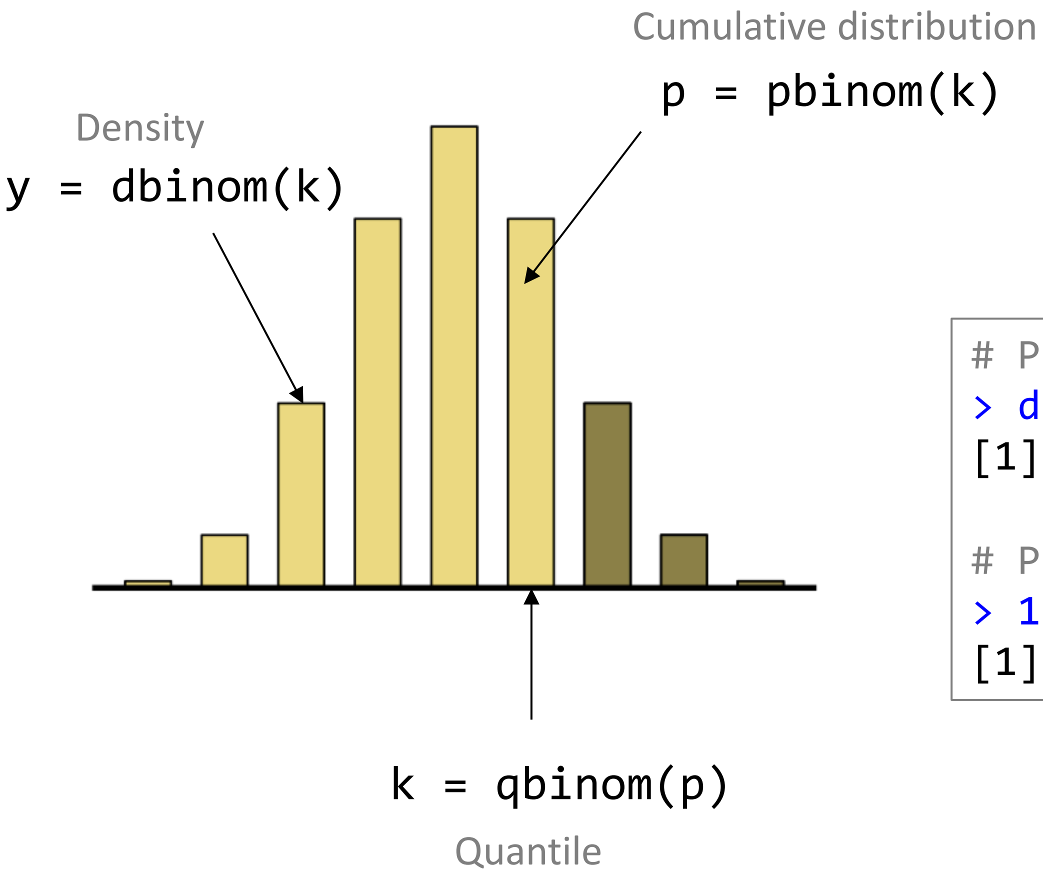
Find Z corresponding to CI = 95%



$$p = \frac{1-\beta}{2} + \beta = \frac{1+\beta}{2} = 0.975$$

```
Z = qnorm(0.975)
[1] 1.959964
```

Probability distributions in R




```
# Probability of exactly 2 heads
> dbinom(2, size=8, prob=0.5)
[1] 0.109375

# Probability of at least 6 heads
> 1 - pbinom(5, size=8, prob=0.5)
[1] 0.14453122
```

Probability distributions in R

Distribution	Density	Cumulative	Quantiles
Normal	`dnorm`	`pnorm`	`qnorm`
Poisson	`dpois`	`ppois`	`qpois`
Binomial	`dbinom`	`pbinom`	`qbinom`
Log-normal	`dlnorm`	`plnorm`	`qlnorm`
Uniform	`dunif`	`punif`	`qunif`
Student t	`dt`	`pt`	`qt`
Chi-square	`dchisq`	`pchisq`	`qchisq`
Hypergeometric	`dhyper`	`phyper`	`qhyper`
F	`df`	`pf`	`qf`

Summary

Distribution	Description	Examples
Normal	Bell-shaped	Often seen in nature, e.g. human height
Log-normal	Logarithm of this is normal	High-throughput experiments
Poisson	Count distribution	Counts of cells per plate
Binomial	Success vs failure	Male/female distribution

Hand-outs available at http://is.gd/statlec