# Error analysis in biology

Marek Gierliński
Division of Computational Biology

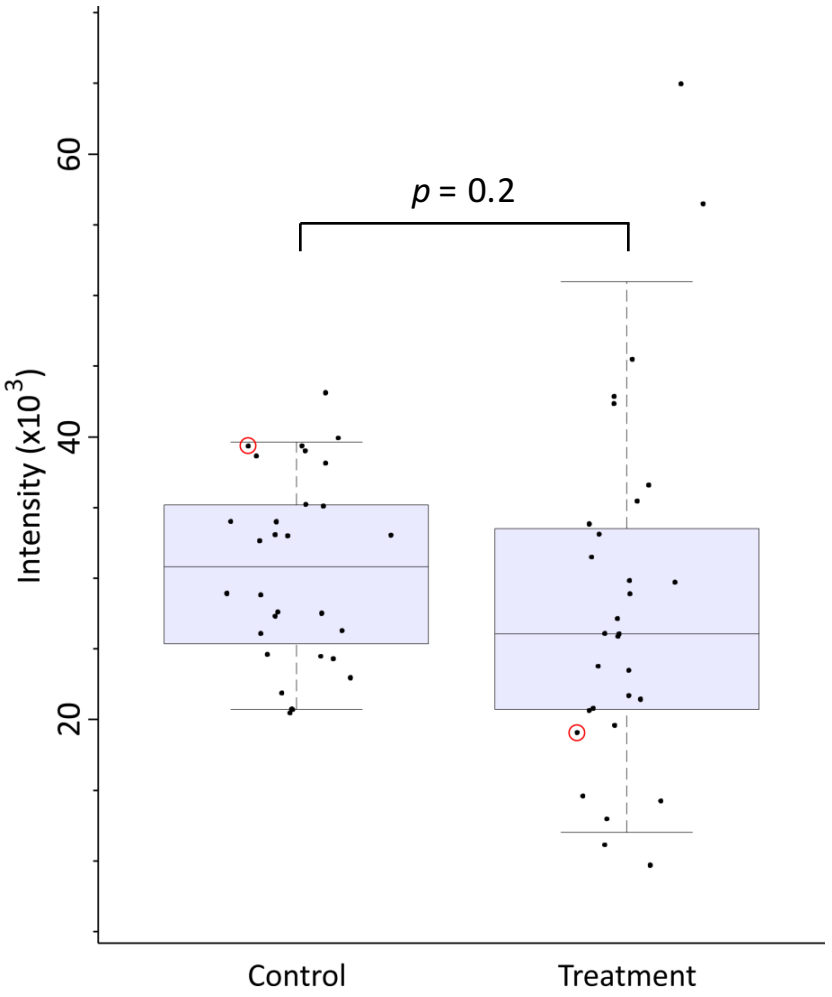
Slides available at http://is.gd/statlec

# Why do we need error analysis?

- Consider a microarray experiment
- Comparing control and treatment
- Expression level of FLG
  - control = 41,723
  - treatment = 19,786
- There is a 2-fold change in intensity
- Great! Gene is repressed in our treatment!

# Why do we need error analysis?

- Consider a microarray experiment
- Comparing control and treatment
- Expression level of FLG
  - control = 41,723
  - treatment = 19,786
- There is a 2-fold change in intensity
- Great! Gene is repressed in our treatment!
- Repeat the experiment in 30 replicates
  - control = $(31.5 \pm 1.6) \times 10^3$
  - treatment = $(27.7 \pm 2.4) \times 10^3$
- Reveal **variability** of expression
- No difference between control and treatment



“A measurement without error is meaningless”

*My physics teachers*

# Data Analysis Group

Chris Cole

Pietà Schofield

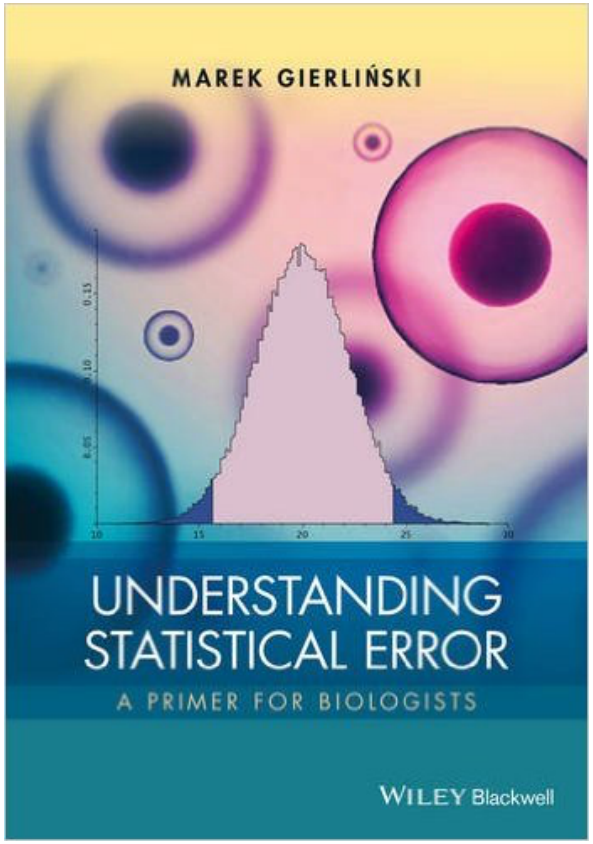
Marek Gierliński

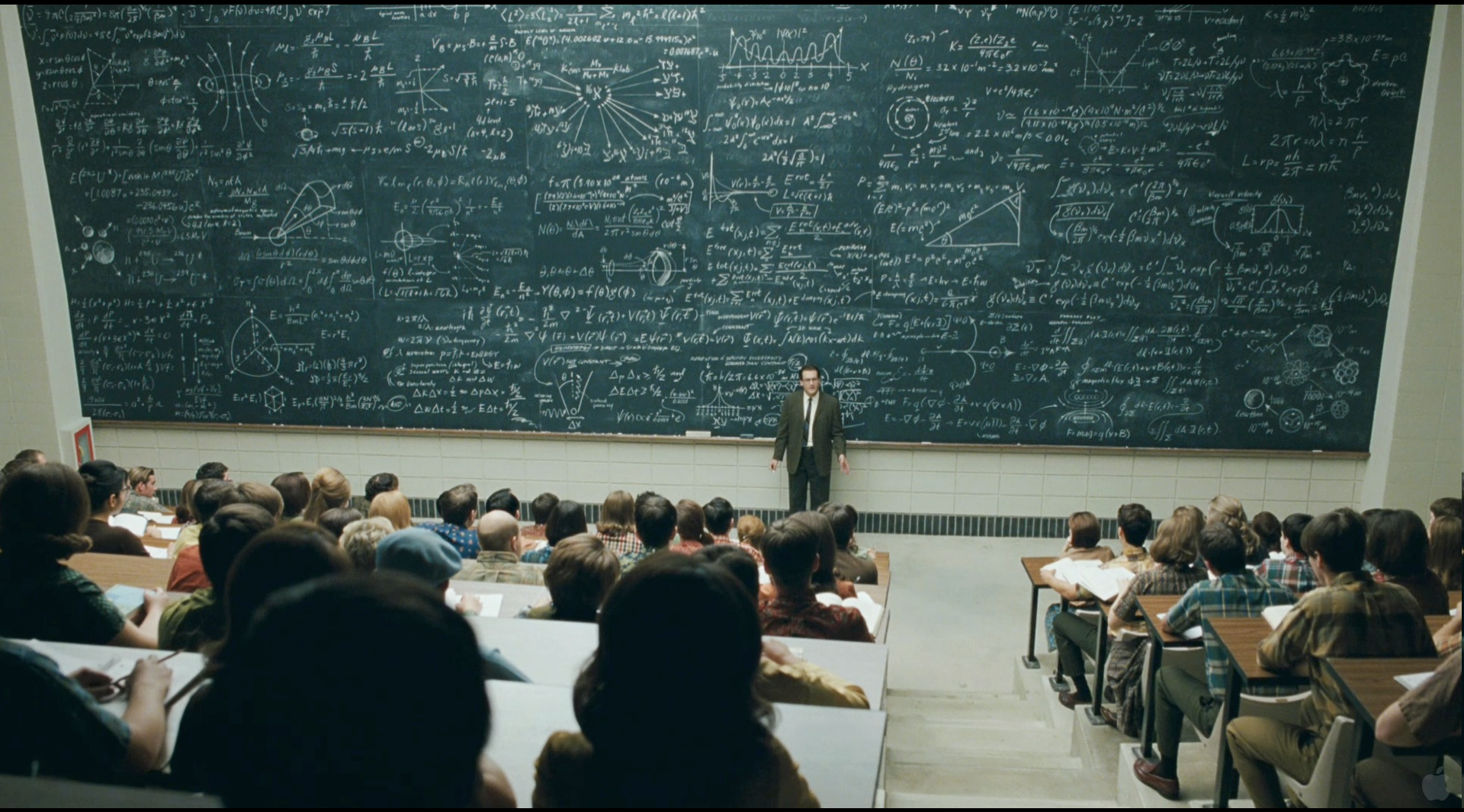
Computational Biology
Barton Group
Level 2 CTIR

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

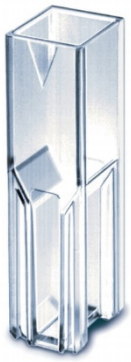




# Table of contents

# 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**

# 1. Random variables and probability distributions

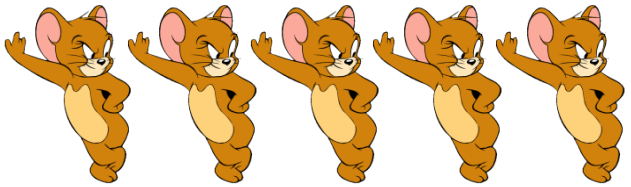
"Misunderstanding of probability may be the greatest of all general impediments to scientific literacy"

*Stephen Jay Gould*

# 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)

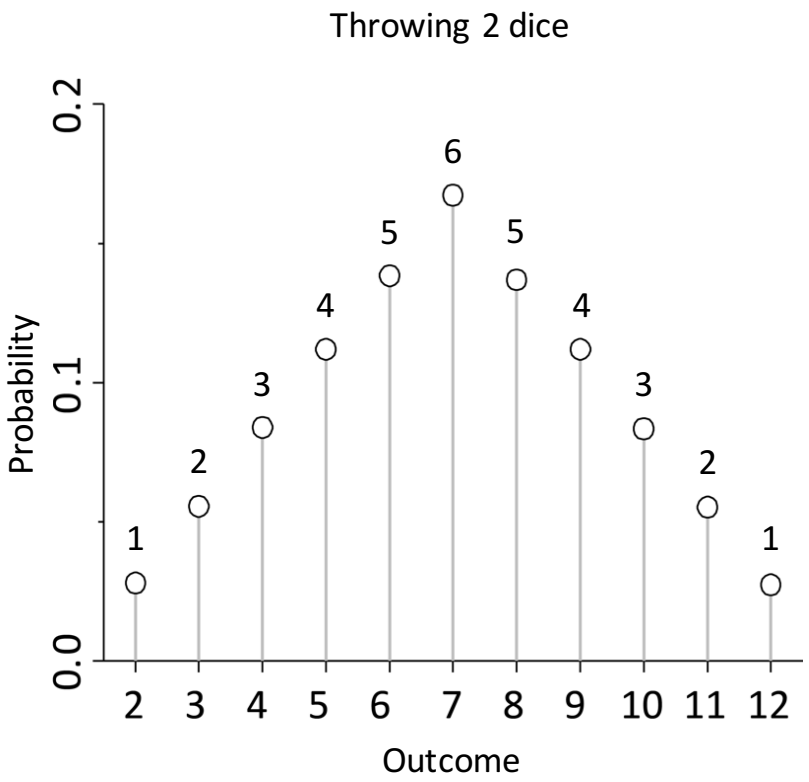


- Continuous variables:
  - weight of a mouse
  - height of a person
  - fluorescent marker luminosity
  - protein abundance

# Probability distribution

- 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 |

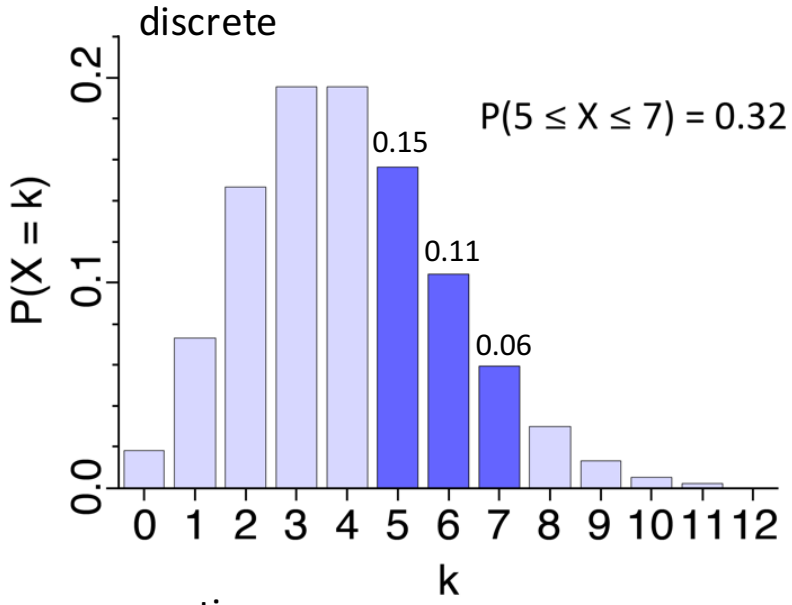


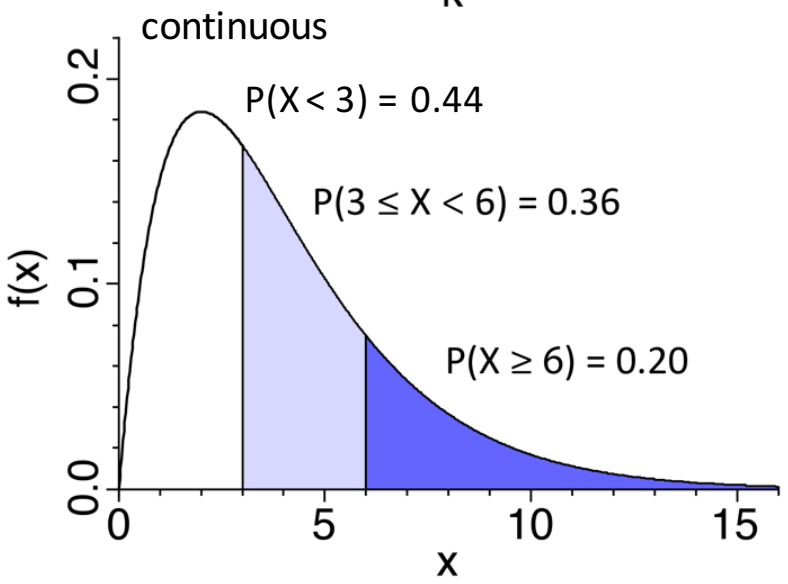
# Probability distribution

- Assigns a probability to each of the possible outcomes
- $X$ – random variable
- $k, x$ – possible values of $X$

- $P(X = 5)$ – probability of $X$ being 5
- $P(5 \leq X \leq 7)$ – probability of $X$ between 5 and 7 (sum of probabilities)

- $f(x)$ – probability density function
- $P(X < 3)$ – area under the curve $f(x)$
- $P(X = 5) = 0$



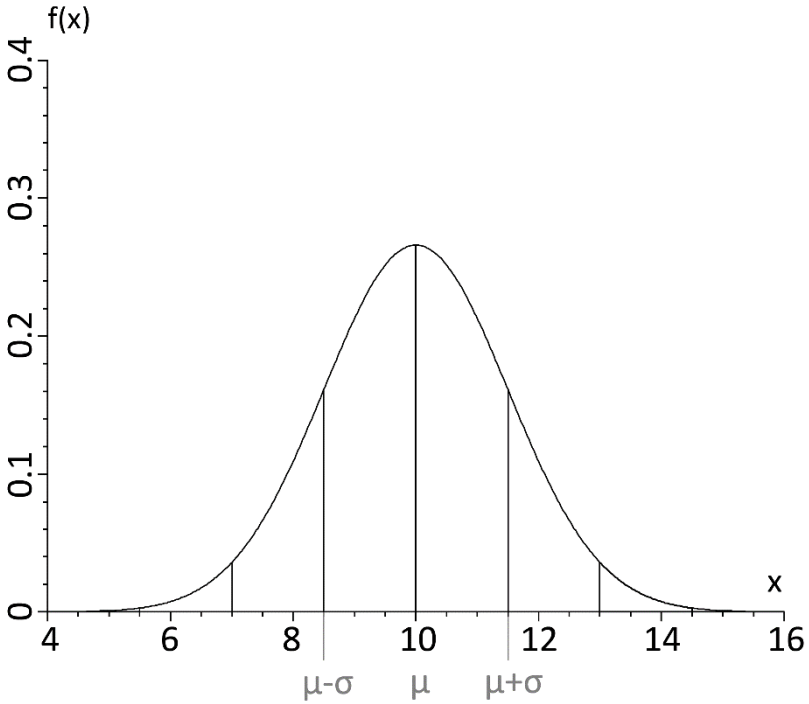
# Gaussian distribution

- Gaussian (or normal) probability distribution

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

  - $\mu$ - mean
  - $\sigma$ - standard deviation
  - $\sigma^2$ - variance

- It is called “normal” as it often appears in nature

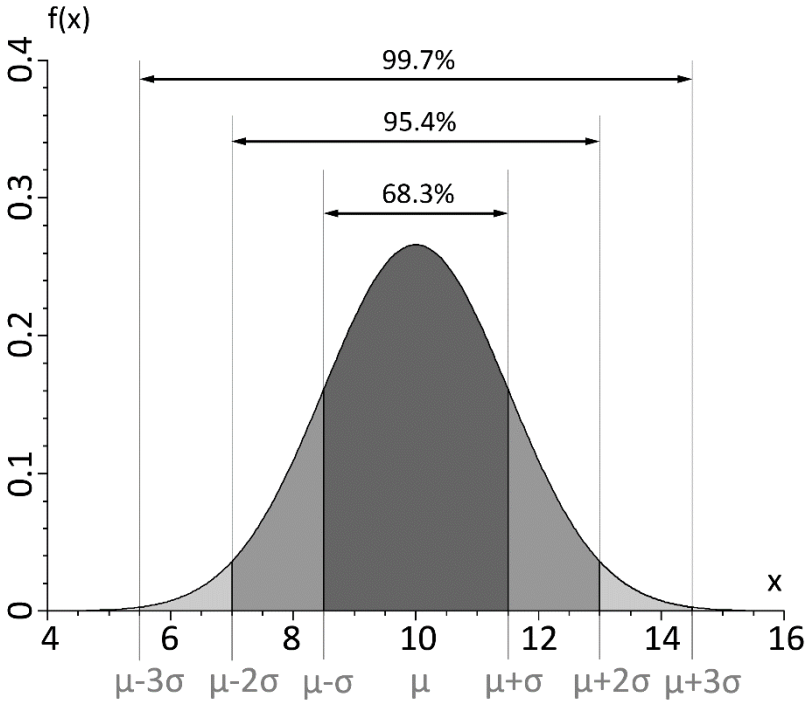



$\mathcal{N}(10, 1.5)$ - normal distribution with $\mu = 10$ and $\sigma = 1.5$

# Gaussian 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 $1.96\sigma$

| | 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 |

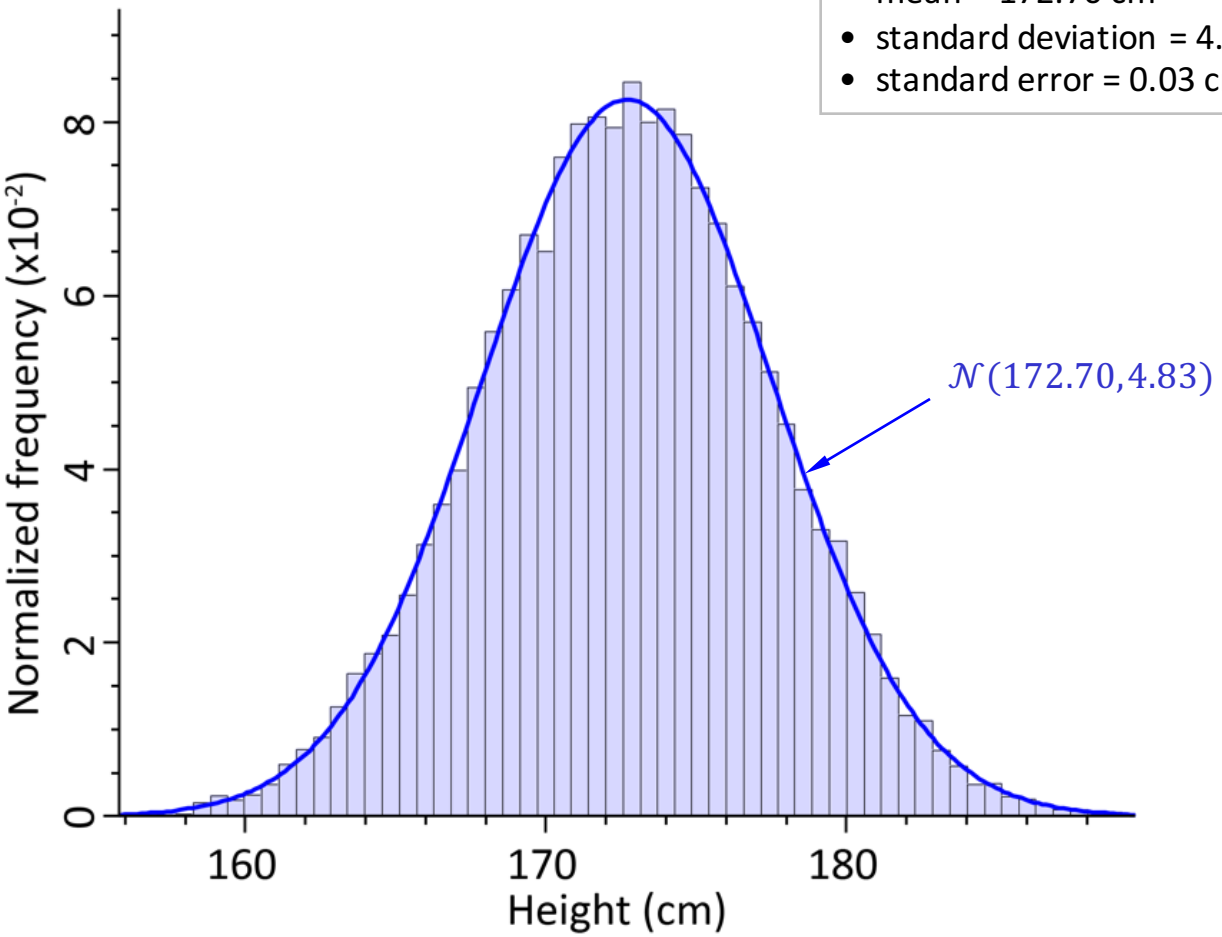



$\mathcal{N}(10, 1.5)$ - normal distribution with $\mu = 10$ and $\sigma = 1.5$

# Example: Gaussian distribution

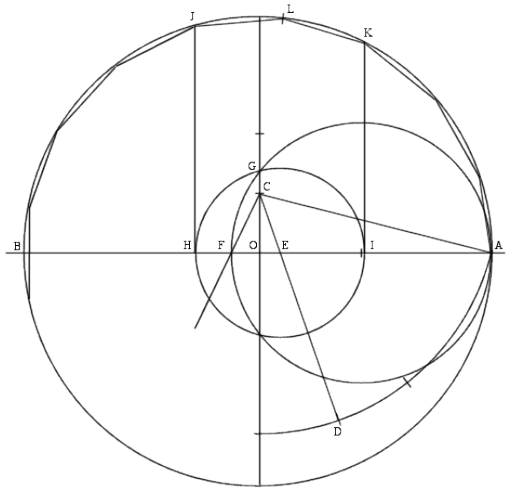
Height of 25,000 individuals from Hong Kong

- mean = 172.70 cm
- standard deviation = 4.83 cm
- standard error = 0.03 cm



# Carl Friedrich Gauss (1777-1855)

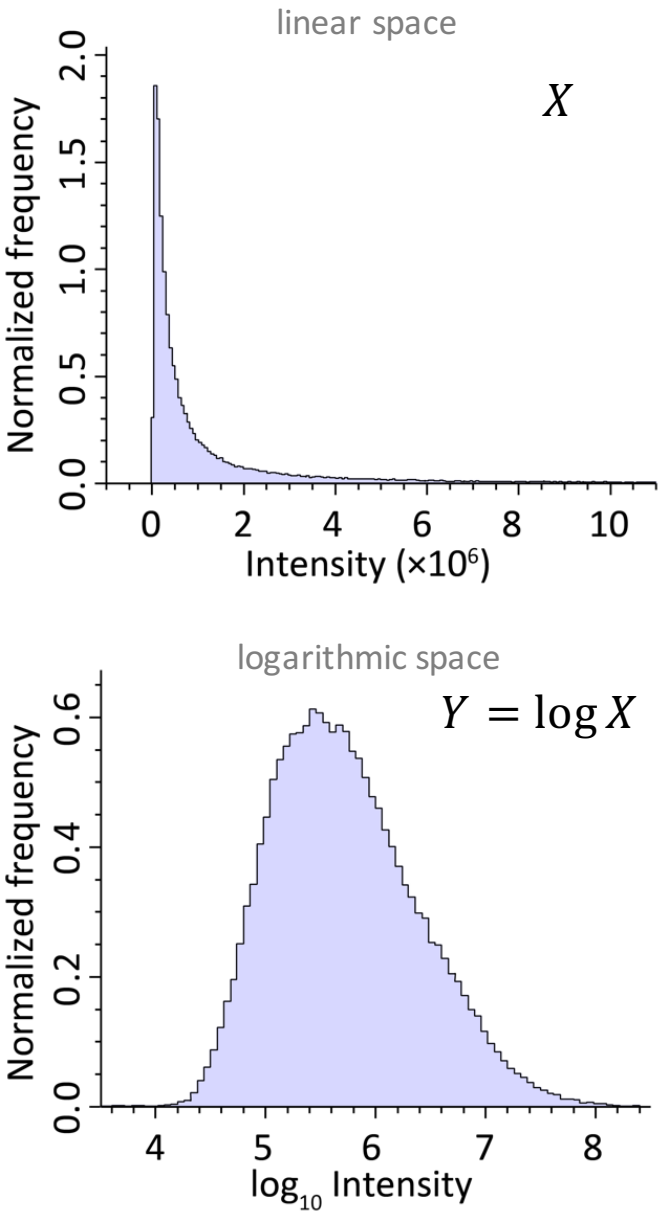
- Brilliant German mathematician
- Constructed a regular heptadecagon with a ruler and a compass
- He requested that a regular heptadecagon should be inscribed on his tombstone
- However, it was Abraham de Moivre (1667-1754) who first formulated "Gaussian" distribution




# Log-normal distribution

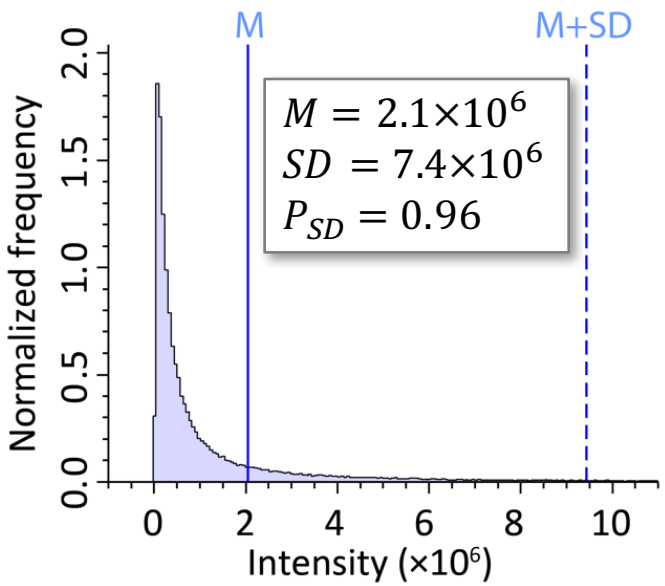
- Probability distribution of a random variable whose logarithm is normally distributed

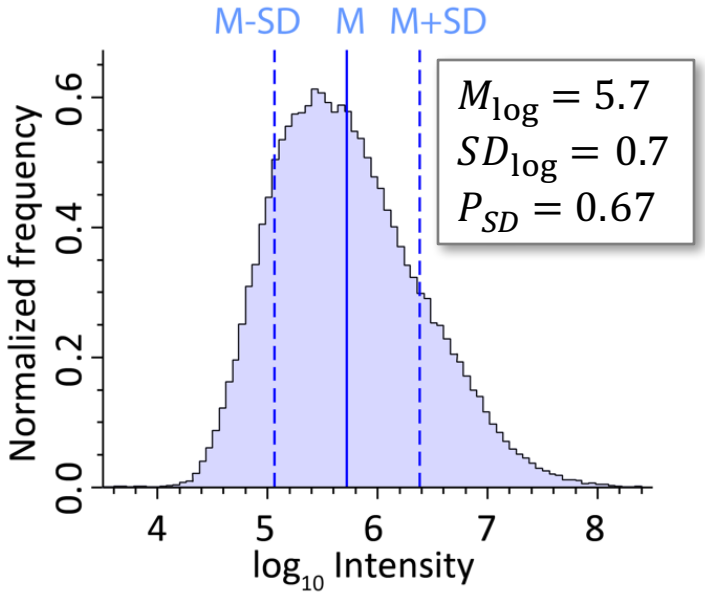
- Log-normal distribution can be very asymmetric!



# Example: log-normal distribution

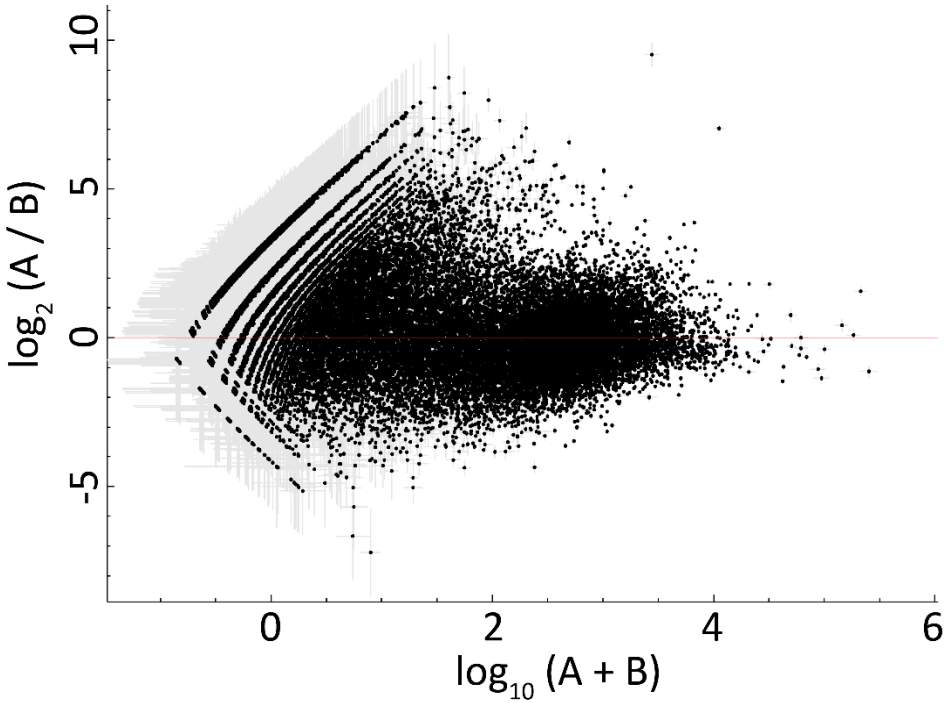
- Peptide intensities from a mass spectrometry experiment
- $P_{SD}$ - fraction of data within $M \pm SD$
- Data look better in logarithmic space
- Always plot the distribution of your data before analysis
- About two-thirds of data points are within one standard deviation from the mean **only** when 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 $IC_{50}$




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

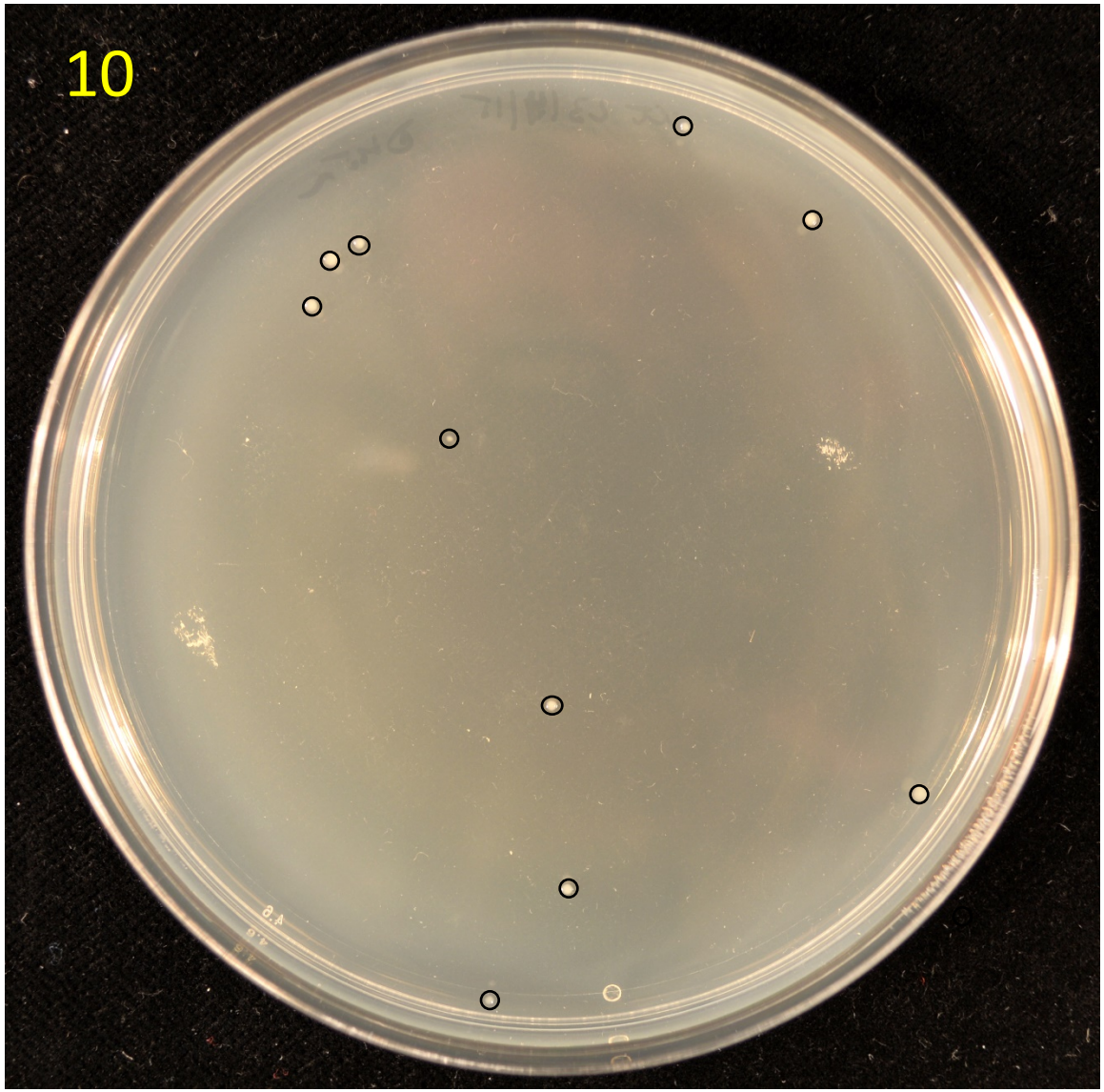
- $\log_{10}$ is easier to understand in plots
  - $10^5 = 100{,}000$
  - $2^{10} = 1024$

# 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

# Counting bacterial colonies

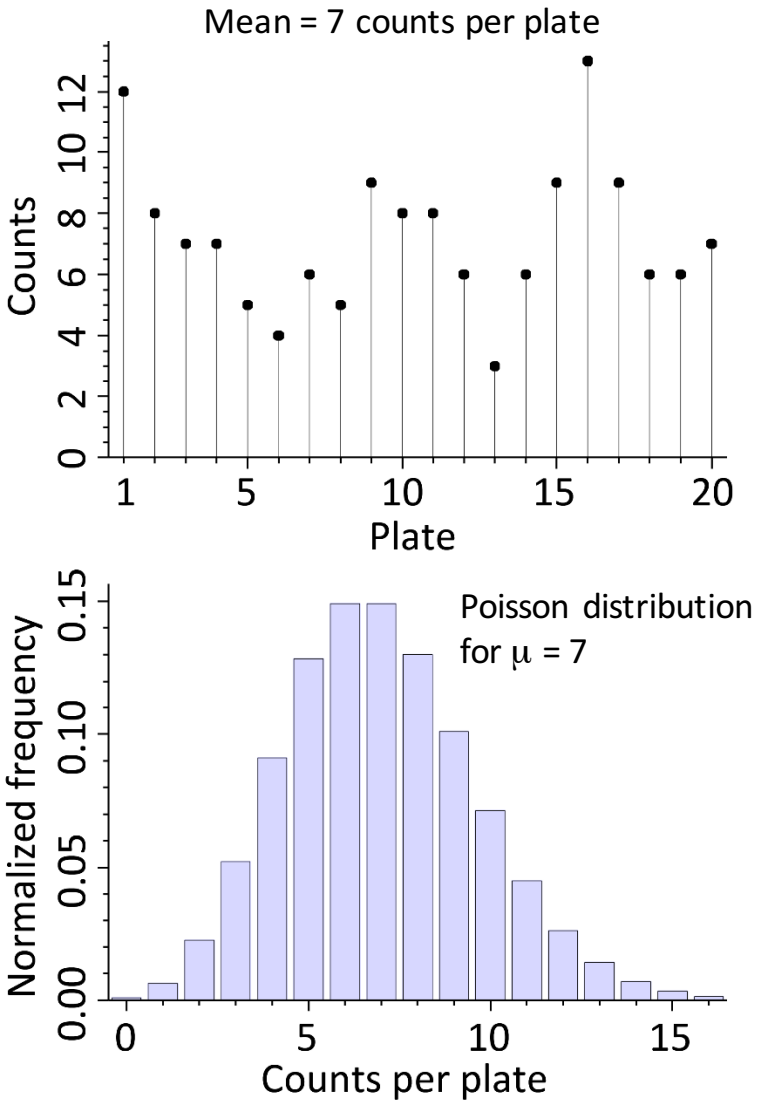



Courtesy of Katharina Trunk, Molecular Biology

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:

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

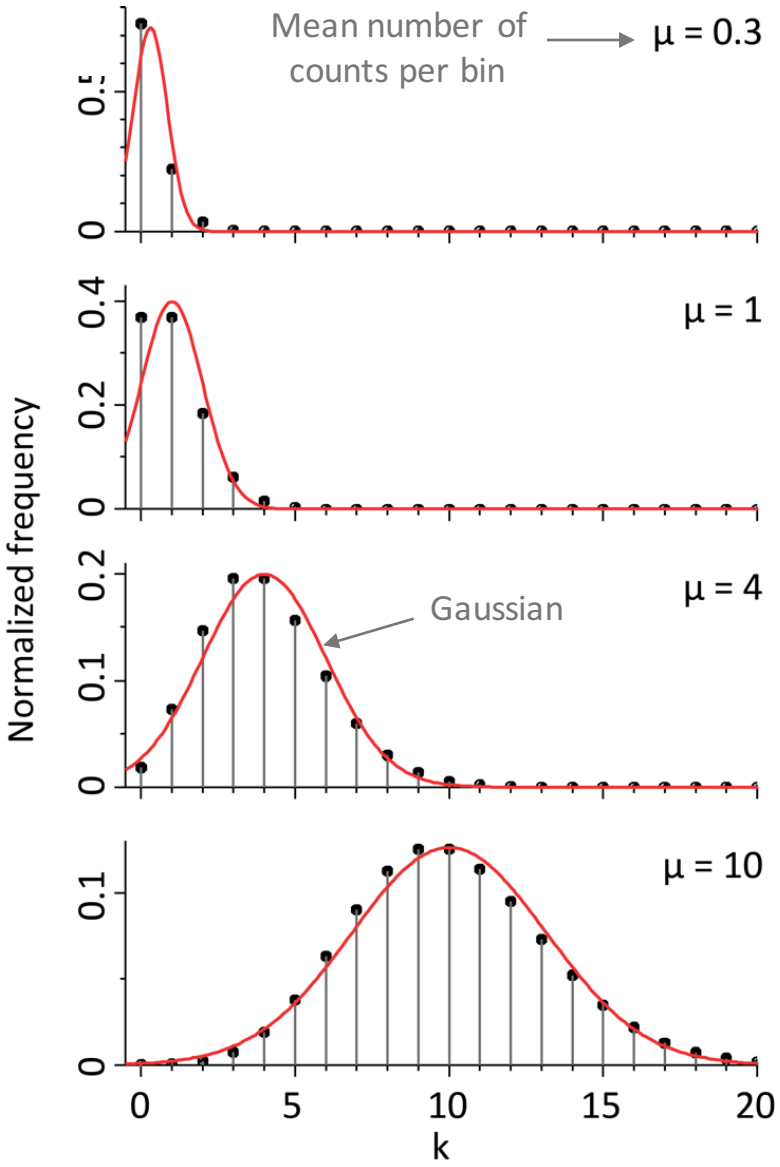
- One parameter: mean count rate, $\mu$
- Standard deviation:

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

$$\sigma^2 = \mu$$

- For large $\mu$ 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

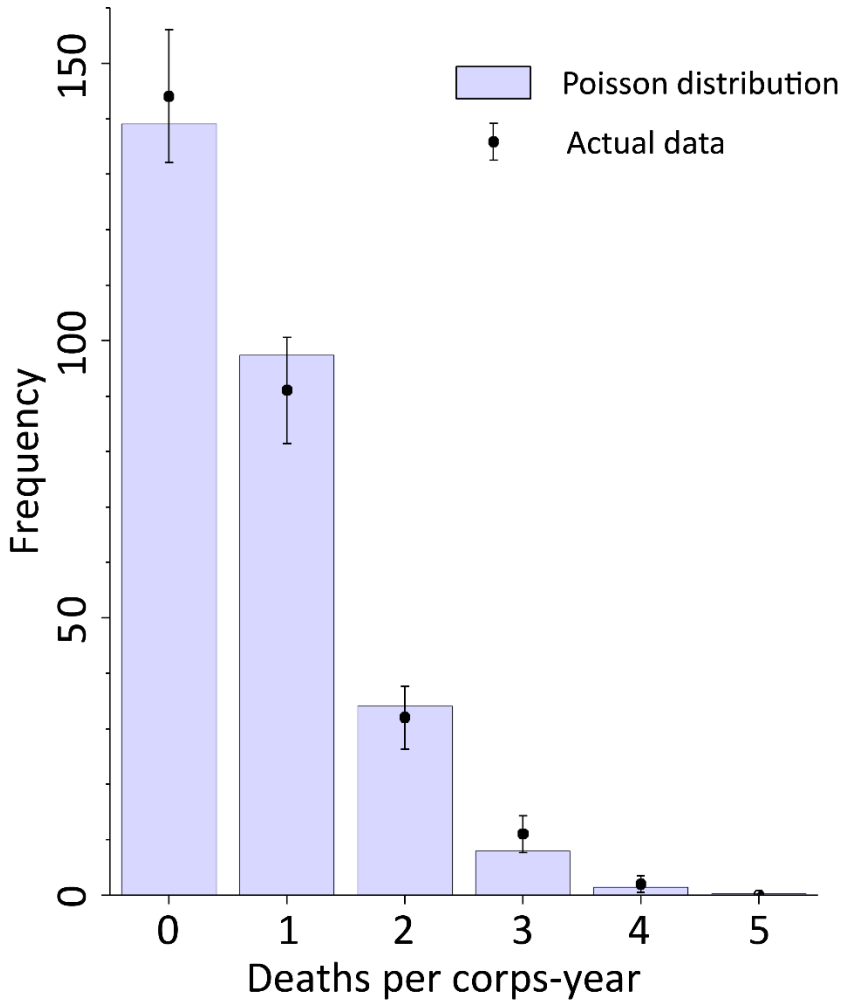
In nachstehender Tabelle sind die Zahlen der durch Schlag eines Pferdes verunglückten Militärpersonen, nach Armeecorps („G." bedeutet Gardecorps) und Kalenderjahren nachgewiesen.[1])

| | 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

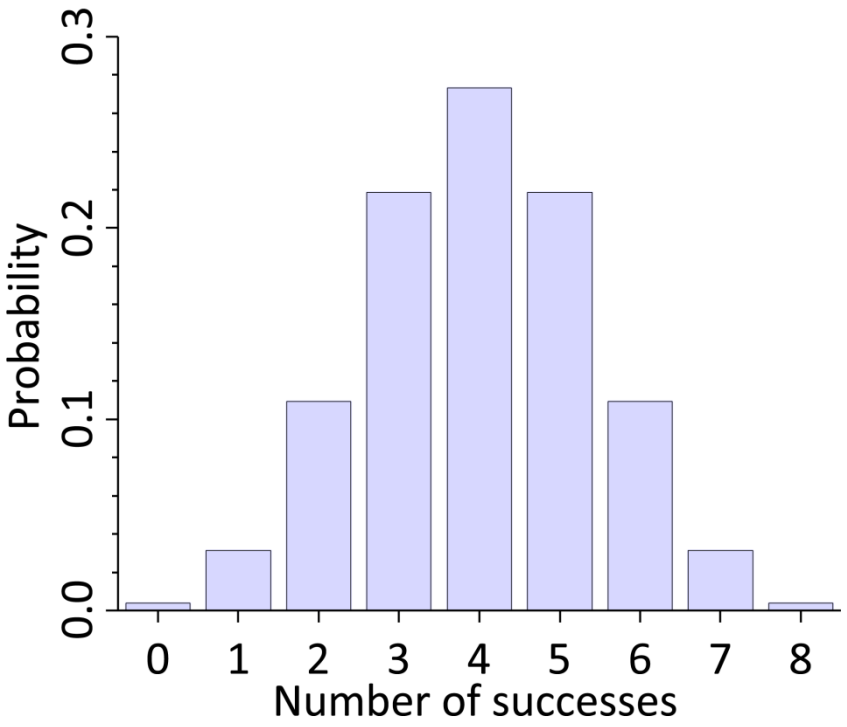
- 4 deaths in a corps-year are expected to happen from time to time!
- $P(X = 4) = 0.078$ in 14 corps
- On average it should happen once in 13 years



# Binomial distribution

- 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

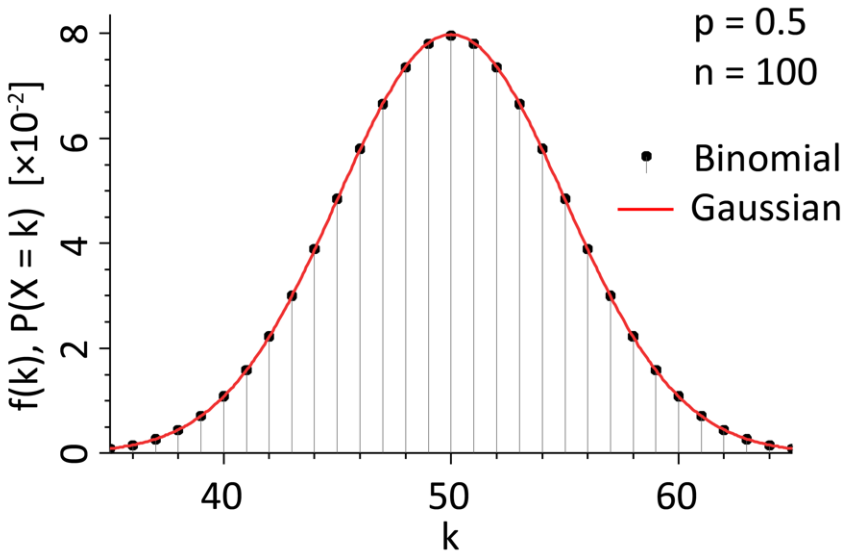
# Binomial distribution

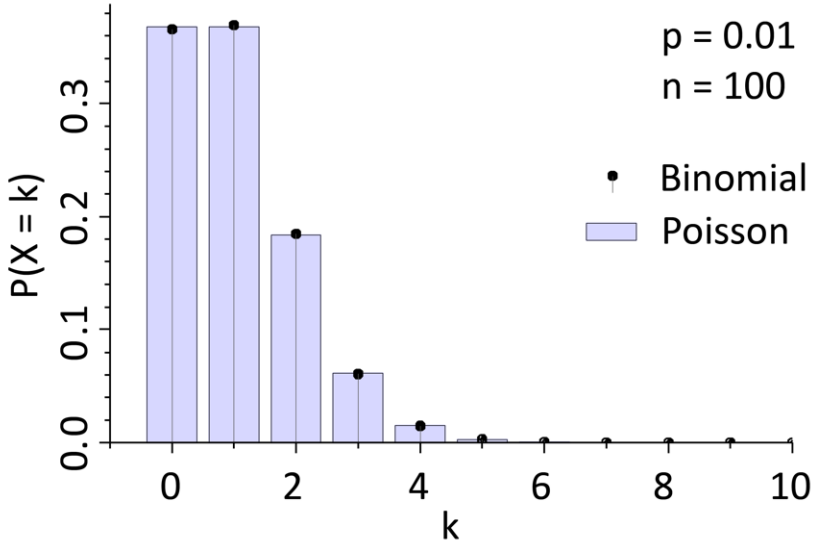
- Mean and standard deviation

$$\mu = np$$

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

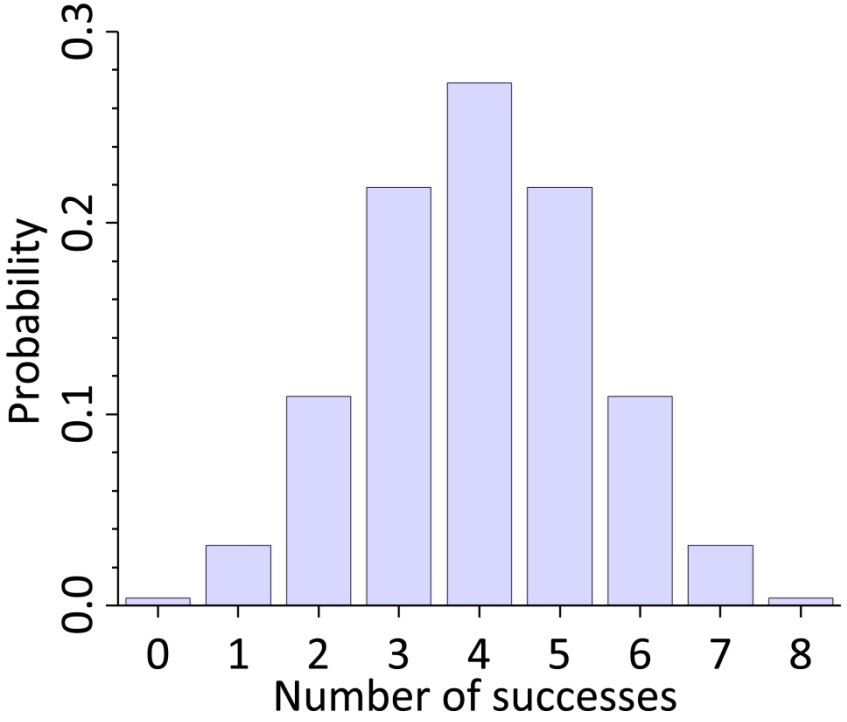
- For large $n$ it approximates Gaussian
- For large $n$ and small $p$ it becomes Poisson



# 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?

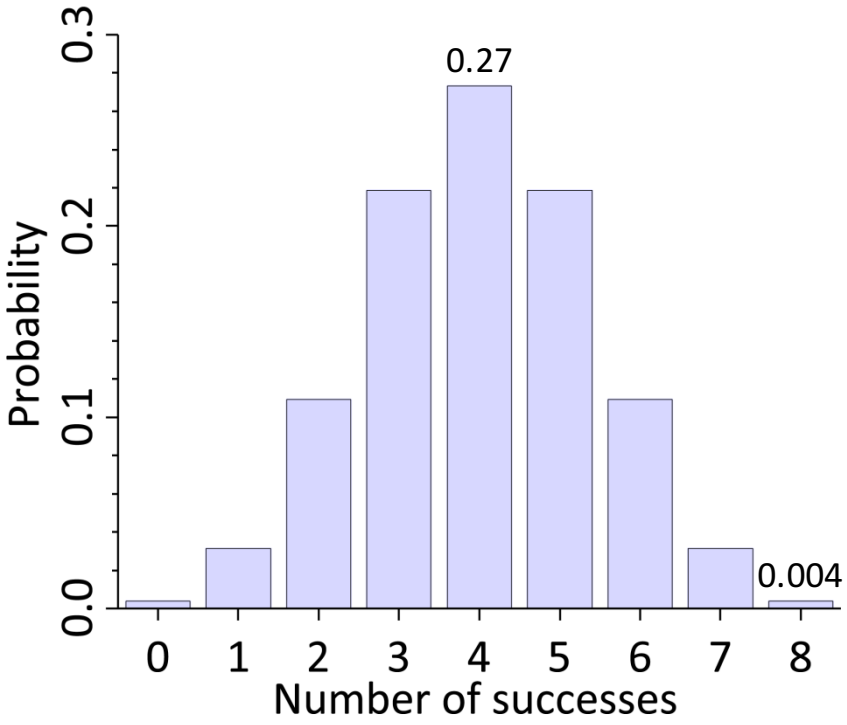
# Example: tossing a coin

- There is only one way of having heads 8 times

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

...

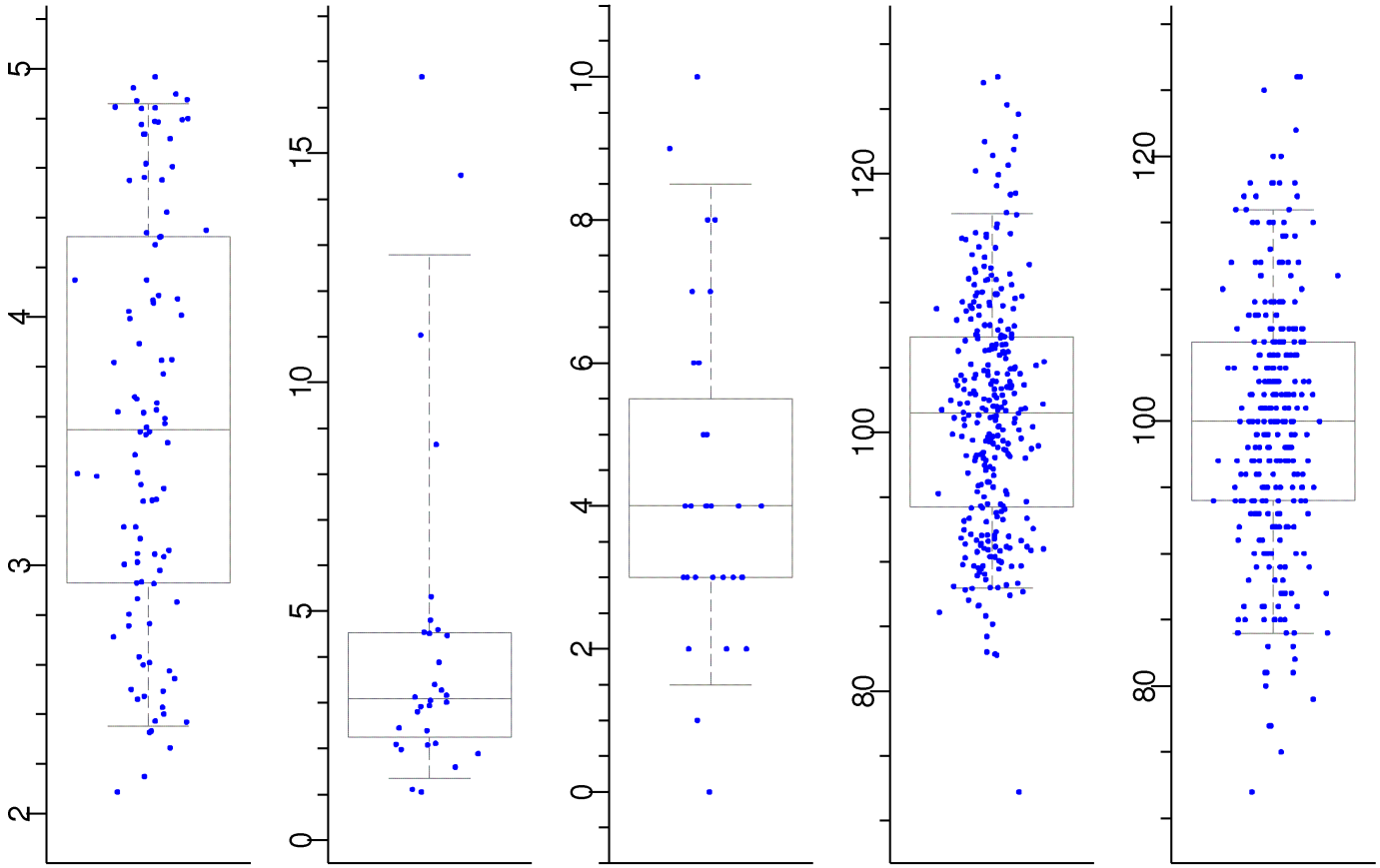
$$\binom{8}{4} = 70$$




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

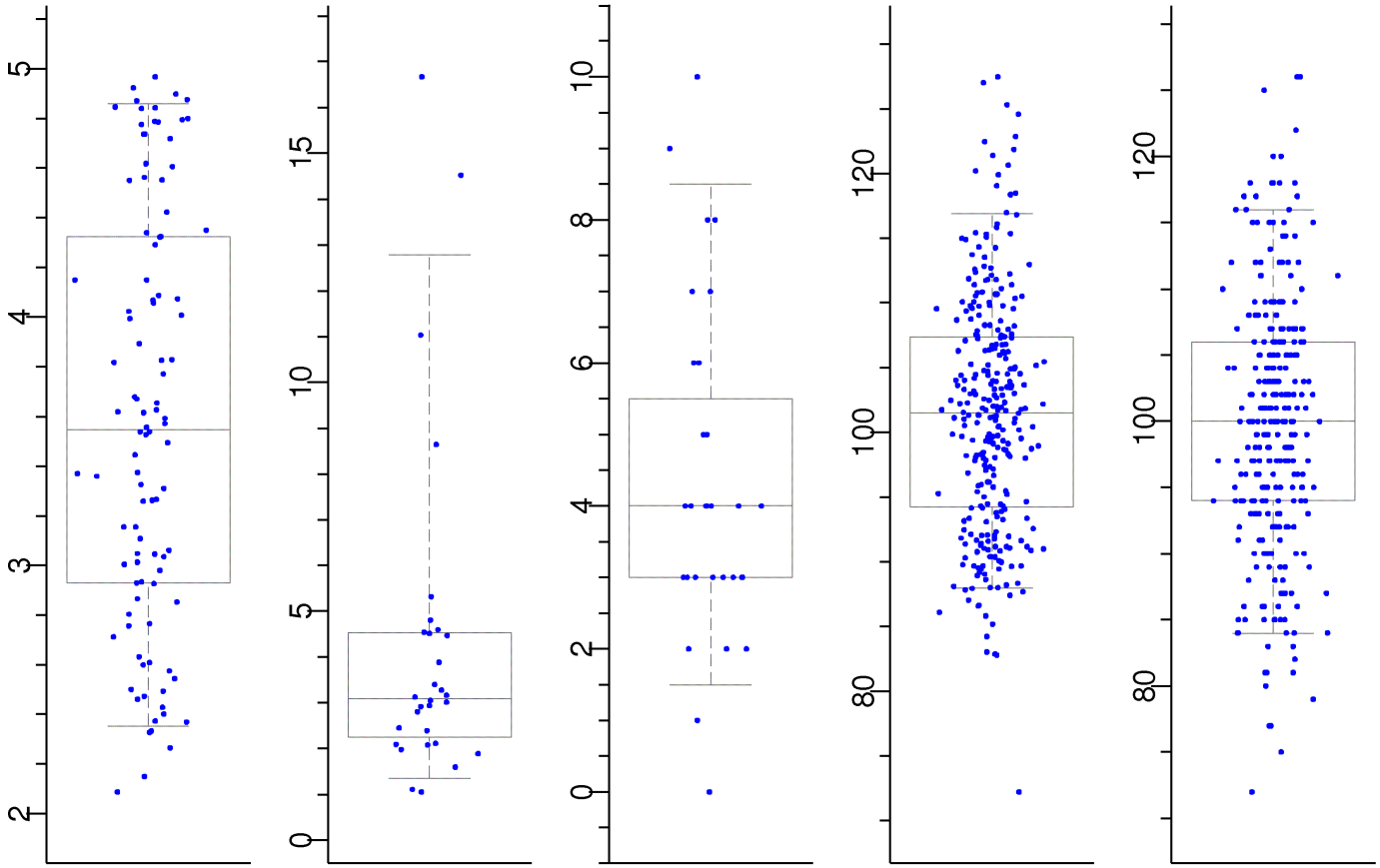
What is the probability of obtaining heads $k$ times from 8 coins?

# Exercise: recognize these distributions




| Distribution | | | | | |
|---|---|---|---|---|---|
| Mean | | | | | |
| *SD* | | | | | |

# Exercise: recognize these distributions




| Distribution | Uniform | Log-normal | Poisson | Gaussian | Poisson |
|---|---|---|---|---|---|
| Mean | 3.5 | 3.5 | 4 | 100 | 100 |
| *SD* | 0.87 | 0.90 | 2 | 10 | 10 |

# Homework

- In an experiment a marker was transfected into a population of $n = 3 \times 10^5$ cells. The marker functionally integrates into the genome at a rate of 1 in $10^5$. What is the probability of obtaining at least one marked cell after this procedure?

- Hint: use binomial or Poisson distribution.




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