Error analysis in biology

Marek Gierliński
Division of Computational Biology

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

Errors, like straws, upon the surface flow;
He who would search for pearls must dive below

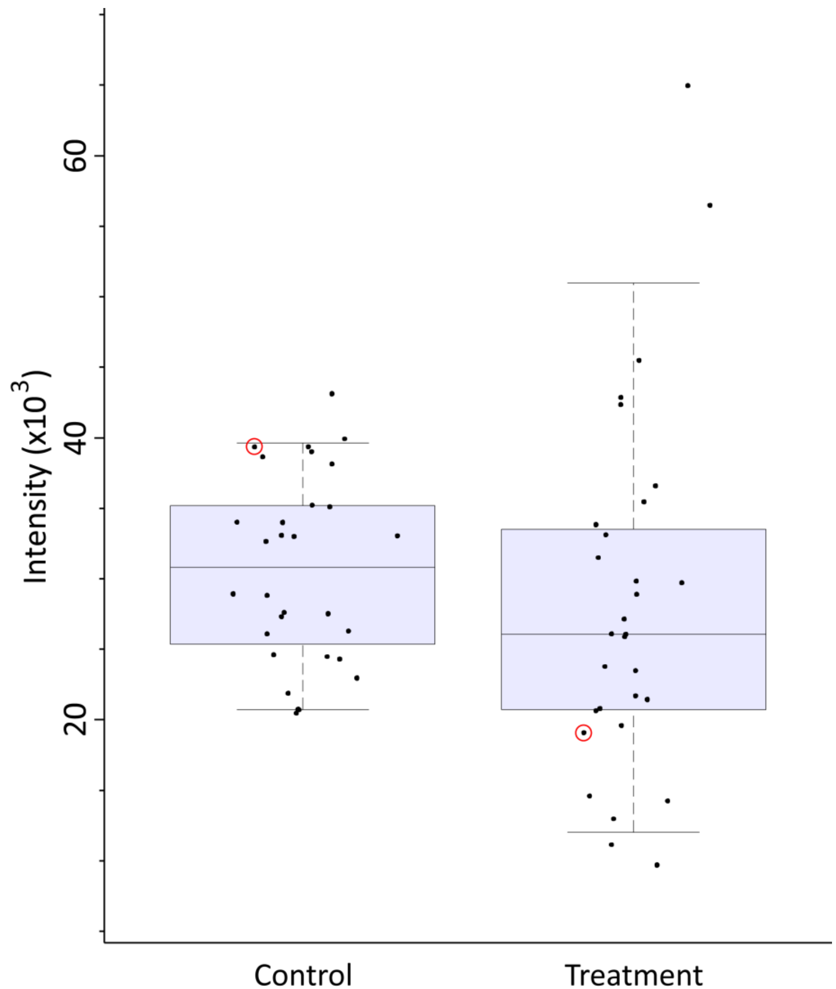
John Dryden (1631-1700)

Why do we need errors (a silly question)?

- 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 errors (a crucial question)!

- 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!
- Now repeat this measurement 30 times
  - control = $(31.5\pm1.6)\times10^3$
  - treatment = $(27.7\pm2.4)\times10^3$
- Reveal **variability** of expression
- Distributions are very similar
- t-test gives $p = 0.2$
- No difference between control and treatment



“A measurement without error is meaningless”

My physics teachers

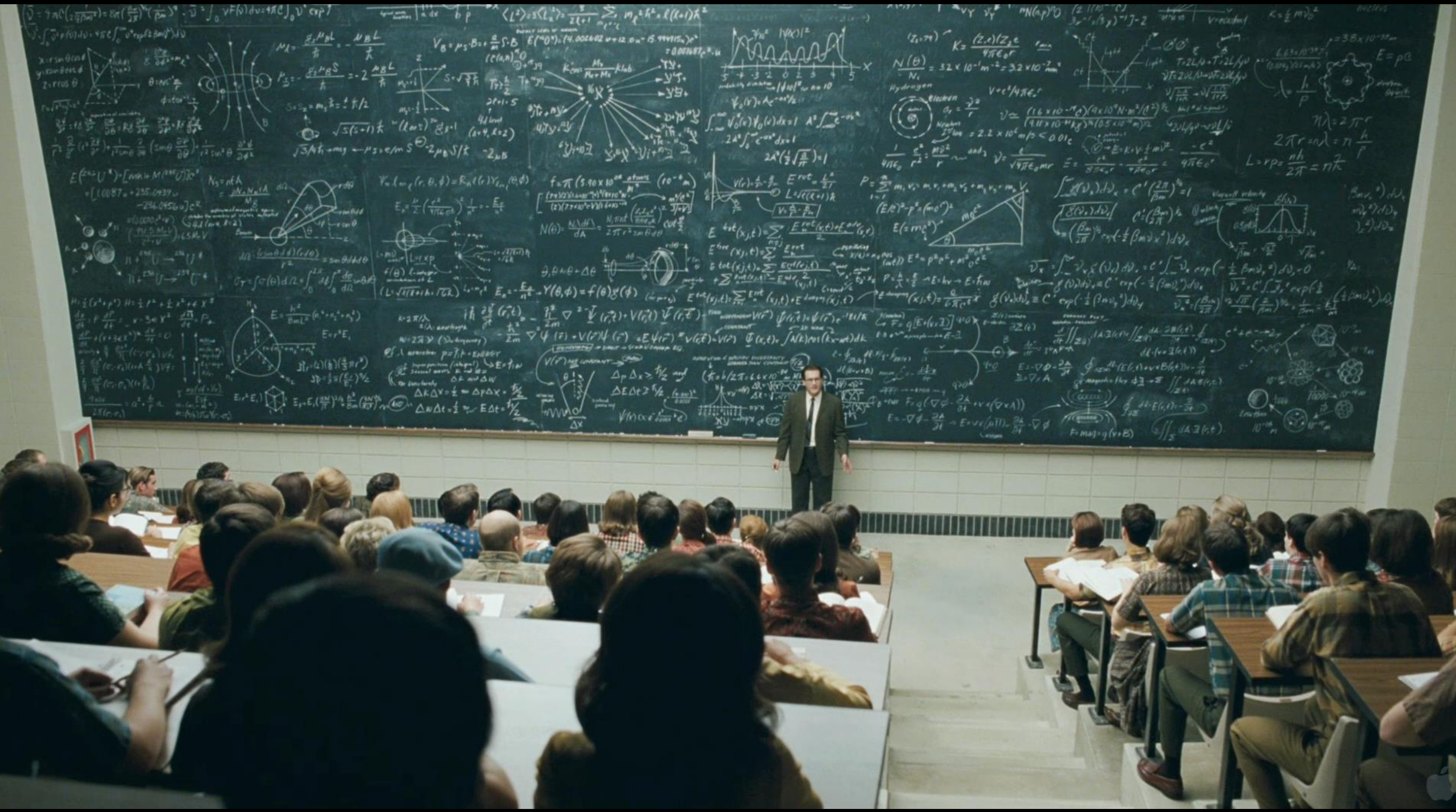


Table of contents

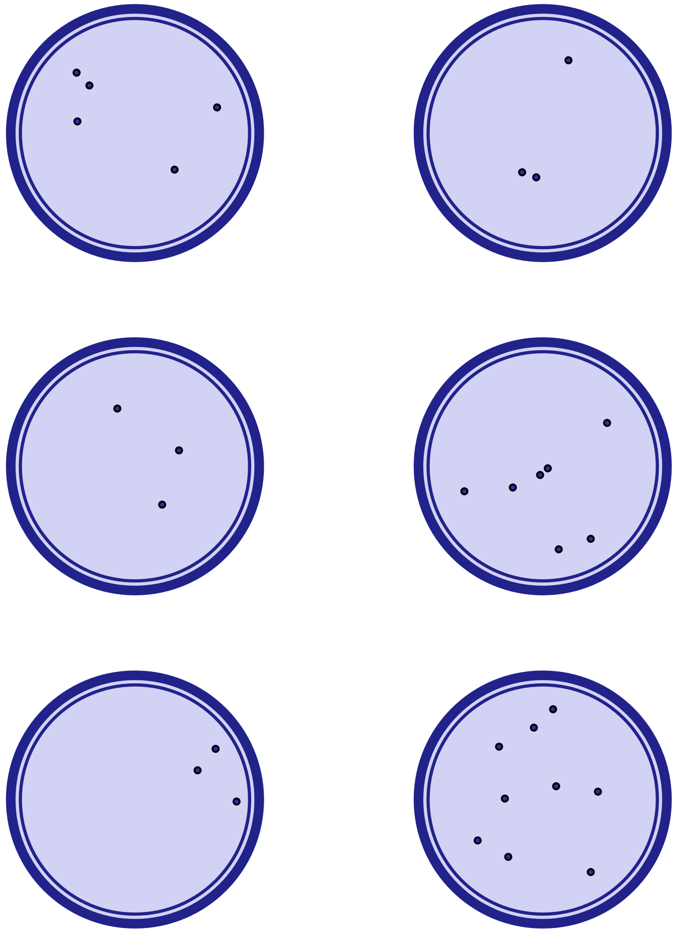
Example

- Experiment: count bacteria in a sample using dilution plating
- 6 replicates
- Find the following numbers of colonies

  5 3 3 7 3 9
- What can we say about these results?

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

- Based on our sample, we can make predictions on future experiments
- We can discuss uncertainty, or error, of the count



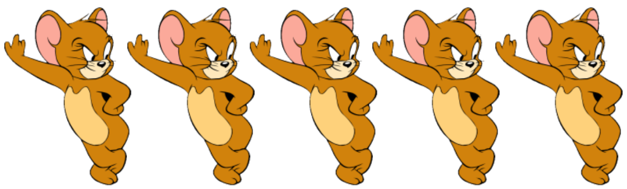
1. Probability distribution

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

Stephen Jay Gould

Random variable

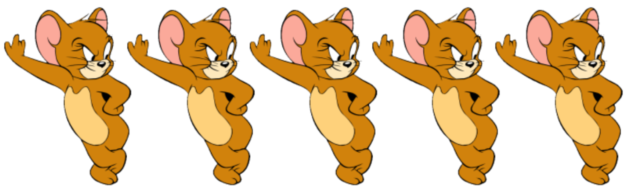
- Random variable can assume random values
- Numerical outcome of an experiment
- Example: result of throwing 2 dice (any number between 2 and 12)
- Non-random variable: number of mice in front of you (5)
- But even the number of mice can be a random variable!
- All values in biological experiments are random variables



Random variable

- Random variable can assume random values
- Numerical outcome of an experiment
- Example: result of throwing 2 dice (any number between 2 and 12)

- Non-random variable: number of mice in front of you (5)

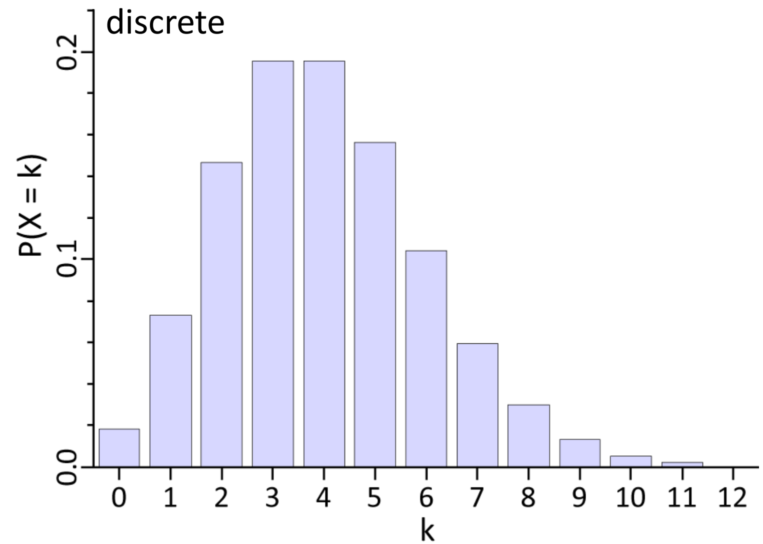


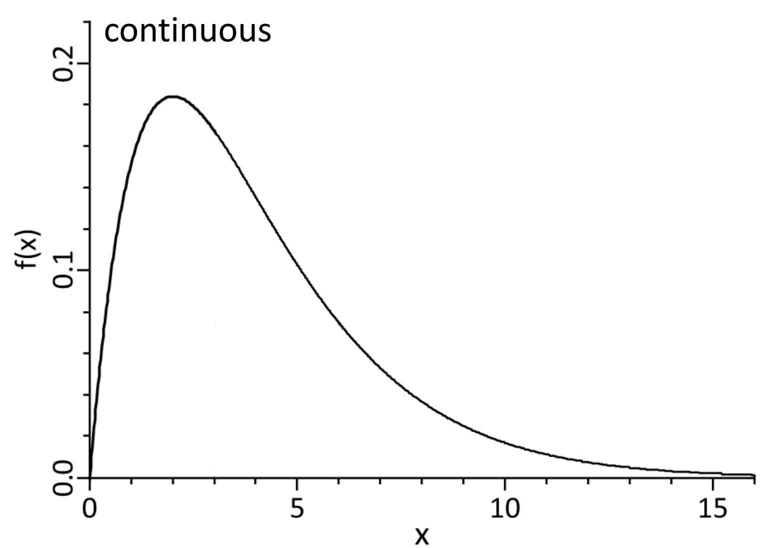
- But even the number of mice can be a random variable!
- All values in biological experiments are random variables

- Two types of random variables
- *discrete* - can assume only certain values
  - number of mice
- *continuous* – can assume any value
  - weight of a mouse

Probability distribution

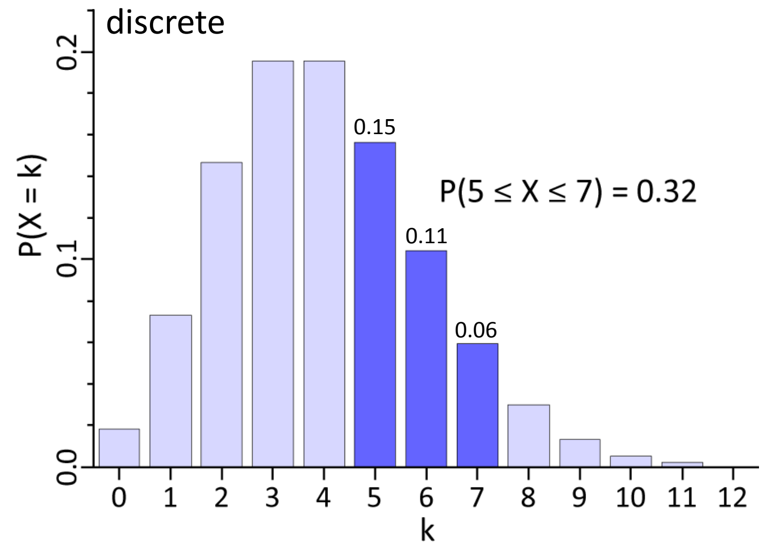
- Probability distribution of a random variable X
- It defines the probability of finding X in a *certain range of values*

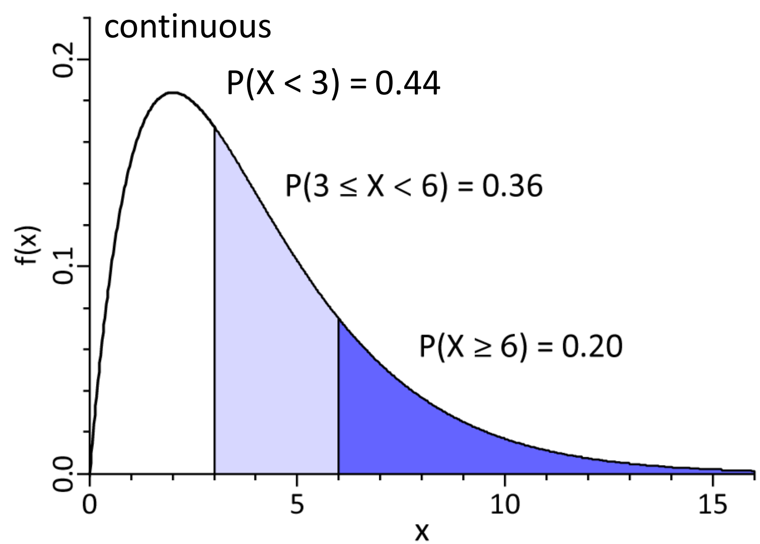




Probability distribution

- Probability distribution of a random variable X
- It defines the probability of finding X in a *certain range of values*
- discrete variable ($k = 0, 1, 2, \ldots$)
  - $P(X = k)$ is a probability of finding $X = k$
  - $P(k_1 \leq X \leq k_2)$ is the sum of individual probabilities
- continuous variable (any value of x)
  - $f(x)$ is a probability density function
  - $P(x_1 \leq X < x_2)$ is the area under the $f(x)$ curve between x_1 and x_2
  - $P(X = 5) = 0$





Notation:
- $X, Y, W, \ldots$ - random variables (symbols)
- $x, y, k, \ldots$ - actual numbers

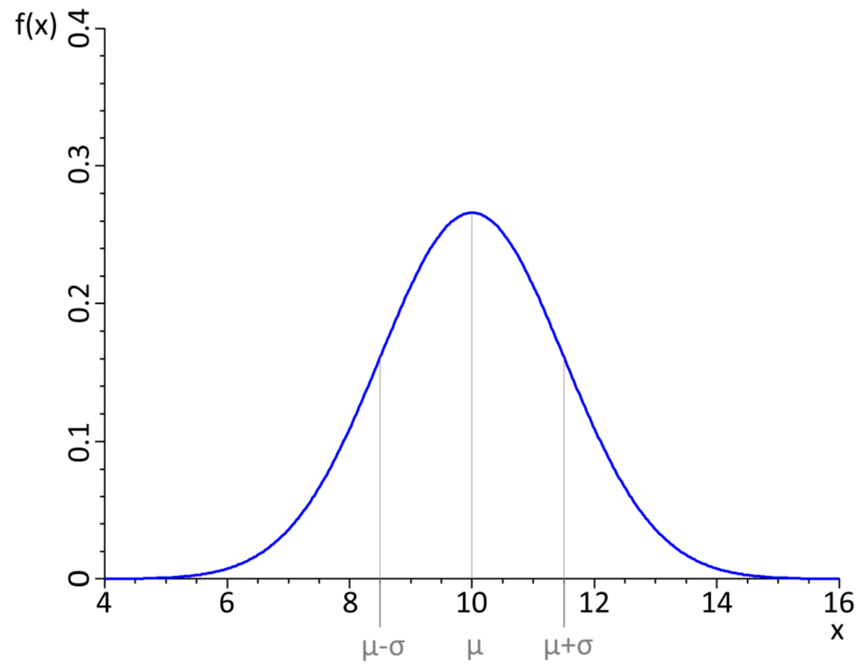
Gaussian distribution

- Gaussian (or normal) 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
- Many observables are normally distributed (central limit theorem)

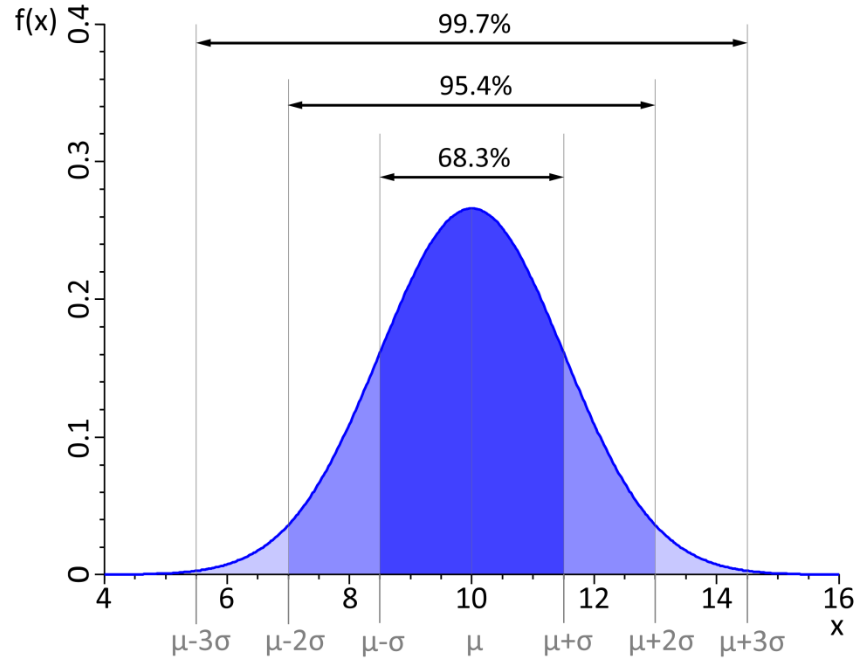


$\mathcal{N}(10, 1.5)$ - normal distribution with $\mu = 10$ and $\sigma = 1.5$

Gaussian distribution: a few numbers

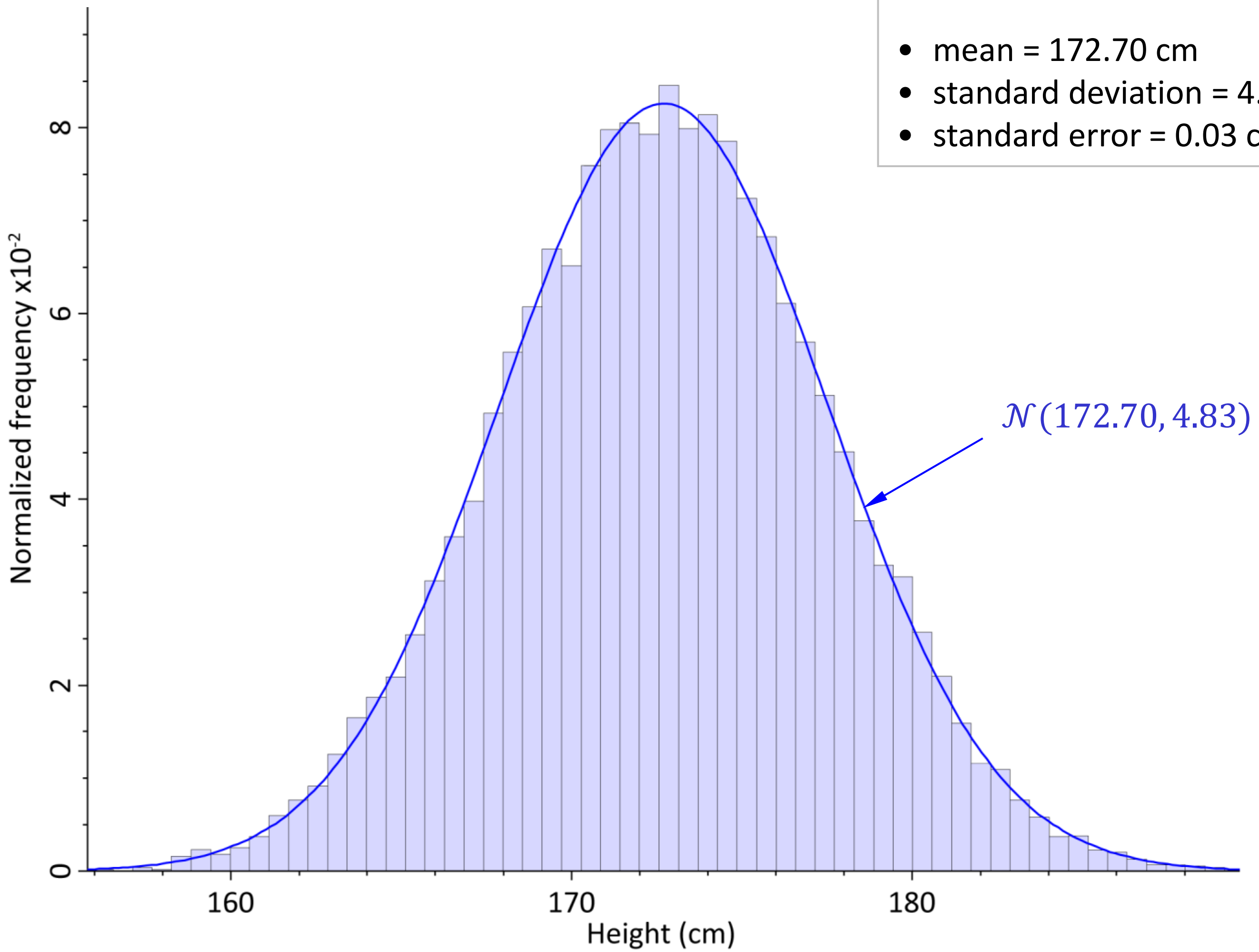
- Area under the curve = probability
- Probability of being within one sigma of the mean is about ⅔ (68.3%)
- Terminology "one sigma", "three sigma": probability of being outside a range (tail)
- 95% confidence intervals are traditionally used: correspond to about 1.96σ

	In	Out	Odds
$\pm 1\sigma$	68.3%	31.7%	1:3
$\pm 1.96\sigma$	95.0%	5.0%	1:20
$\pm 2\sigma$	95.4%	4.6%	1:20
$\pm 3\sigma$	99.7%	0.3%	1:400
$\pm 4\sigma$	99.994%	0.006%	1:16,000
$\pm 5\sigma$	99.99993%	0.00007%	1:1,700,000



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

Example: Gaussian distribution

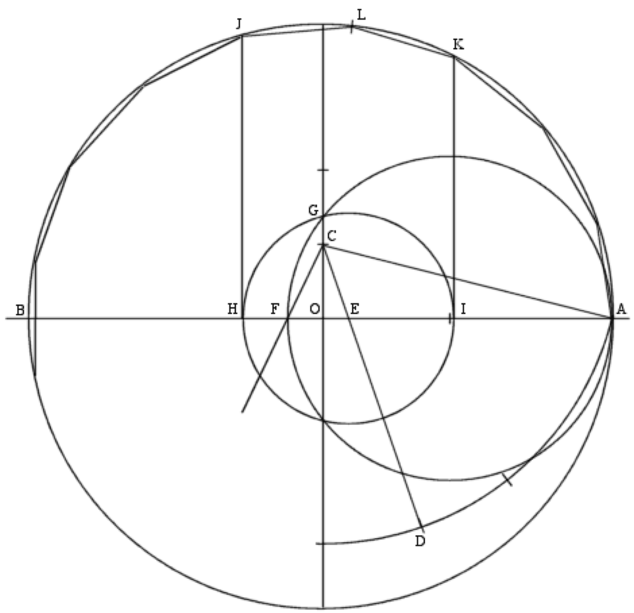


Height of 25,000 individuals from Hong Kong

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

Carl Friedrich Gauss (1777-1855)

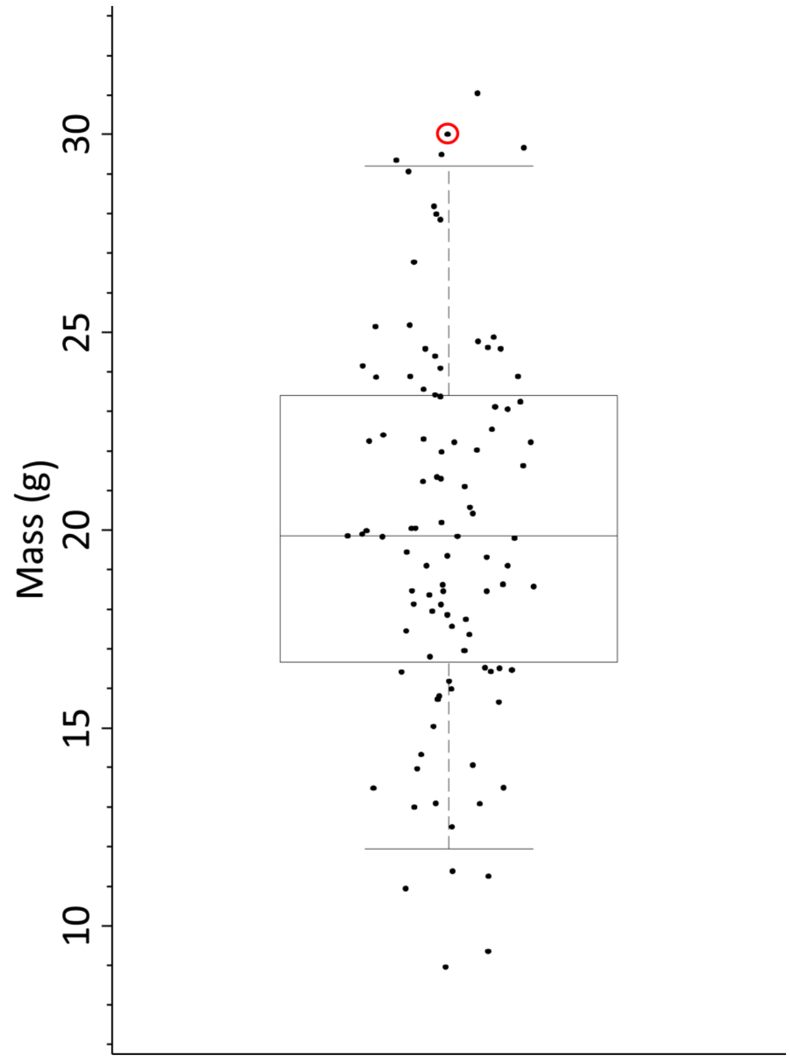
- Brilliant German mathematician
- Constructed a regular heptadecagon with ruler and compass
- He requested that a regular heptadecagon be inscribed on his tombstone
- However, it was Abraham de Moivre (1667-1754) who first formulated “Gaussian” distribution



Exercise: estimate an outlier

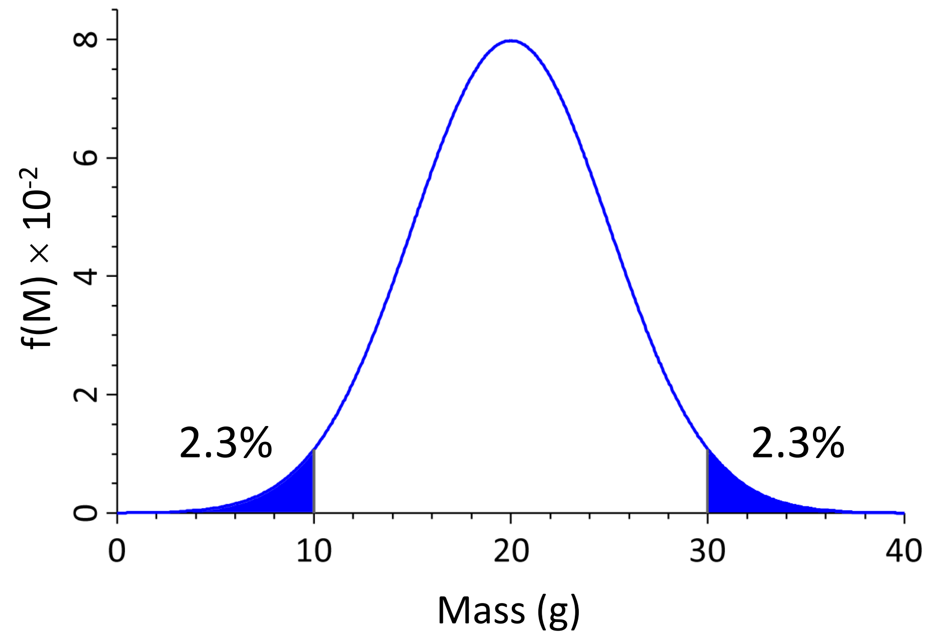
- Obesity study in mice
- Sample of 100 mice, find body weight
    - mean = 20 g
    - standard deviation = 5 g
- Jerry's weight is 30 g
- What is the probability of Jerry being that fat?

	In	Out	Odds
$\pm 1\sigma$	68.3%	31.7%	1:3
$\pm 1.96\sigma$	95.0%	5.0%	1:20
$\pm 2\sigma$	95.4%	4.6%	1:20
$\pm 3\sigma$	99.7%	0.3%	1:400
$\pm 4\sigma$	99.994%	0.006%	1:16,000
$\pm 5\sigma$	99.99993%	0.00007%	1:1,700,000



Exercise: estimate an "outlier"

- What is the probability of Jerry being that fat?
- 30 g is 2σ from the mean:
  - $P(X = 30\text{ g}) = 0$
  - $P(X \geq 30\text{ g}) = 2.3\%$
  - $P(X \geq 30\text{ g} \cup X < 10\text{ g}) = 4.6\%$
- One-tail or two-tail probability?



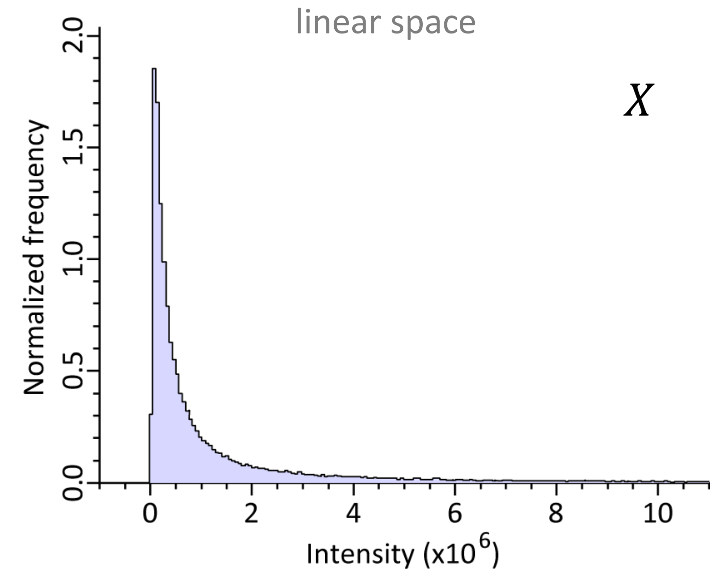
- But even with probability of 2.3% you will expect on average about 2 fat mice in a sample of a 100
- Rare events are expected in large samples
- Jerry is fat, but he is not a statistical outlier

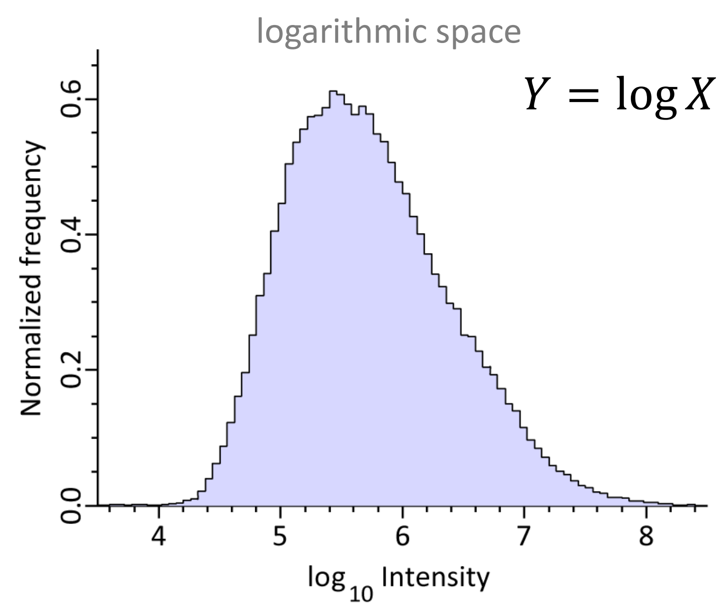
Log-normal distribution

- Log-normal distribution is a probability distribution of a random variable whose logarithm is normally distributed

Log-normal	X	$X = e^Y$
Normal	$Y = \ln X$	Y

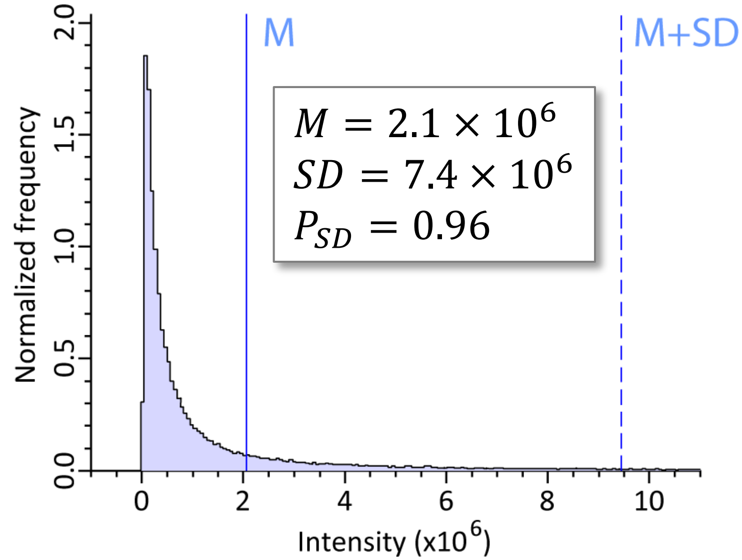
- Log-normal distribution can be very asymmetric!

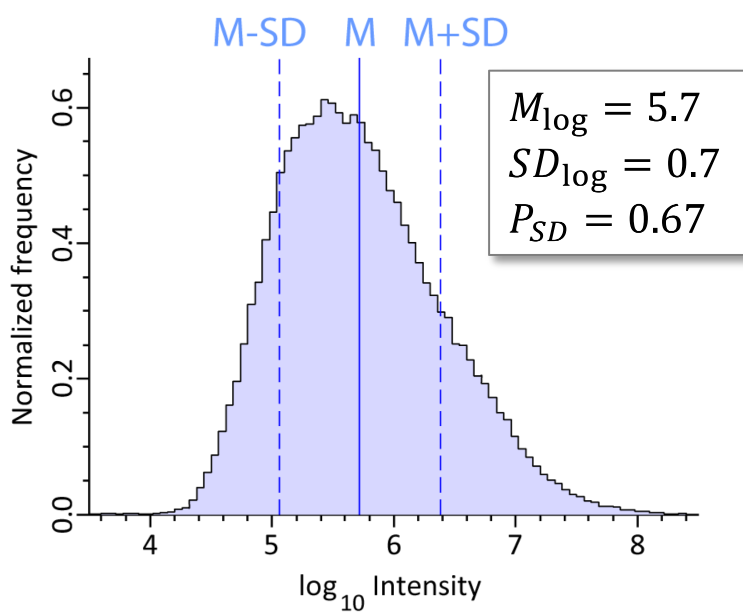




Example: log-normal distribution

- Peptide intensities from 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

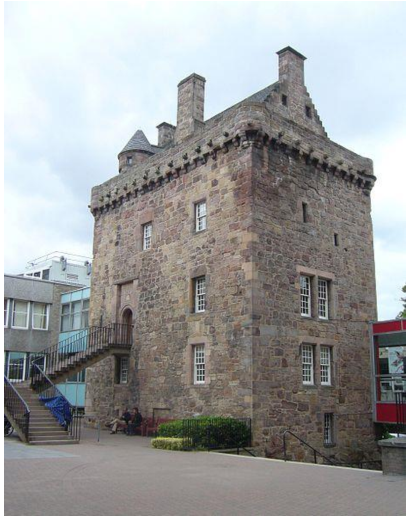
- Examples of log-normal distributions
  - gene expression (RNA-seq, microarrays)
  - mass spectrometry data
  - drug potency IC_{50}
- Difference in log space is a ratio in linear space

$$\log x_1 - \log x_2 = \log \frac{x_1}{x_2}$$

- This is why you should use ratios, not differences, to compare results in these experiments
- 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
  - $10^6 = 1{,}000{,}000$
  - $2^{12} = 4096$

John Napier (1550-1617)

- Scottish mathematician and astronomer
- *Mirifici Logarithmorum Canonis Descriptio* (1614)
- 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



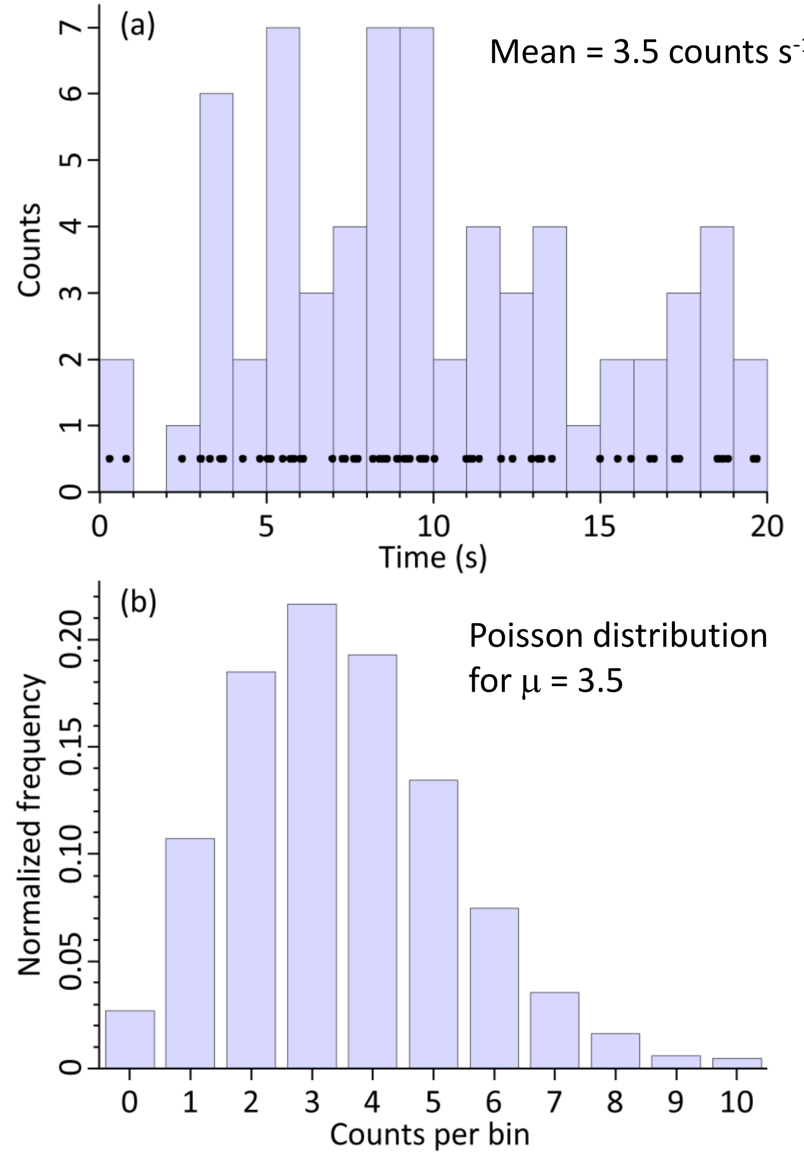
Merchiston Castle, Edinburgh

Poisson distribution

- Consider radioactive decay
- Atomic nucleus can decay spontaneously
- We don't know when it is going to happen

- We know how likely it is to happen in a given period of time
- Collect counts in 1-s bins
- Create distribution of counts per bin

- This applies to any counts in time or space
  - 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:

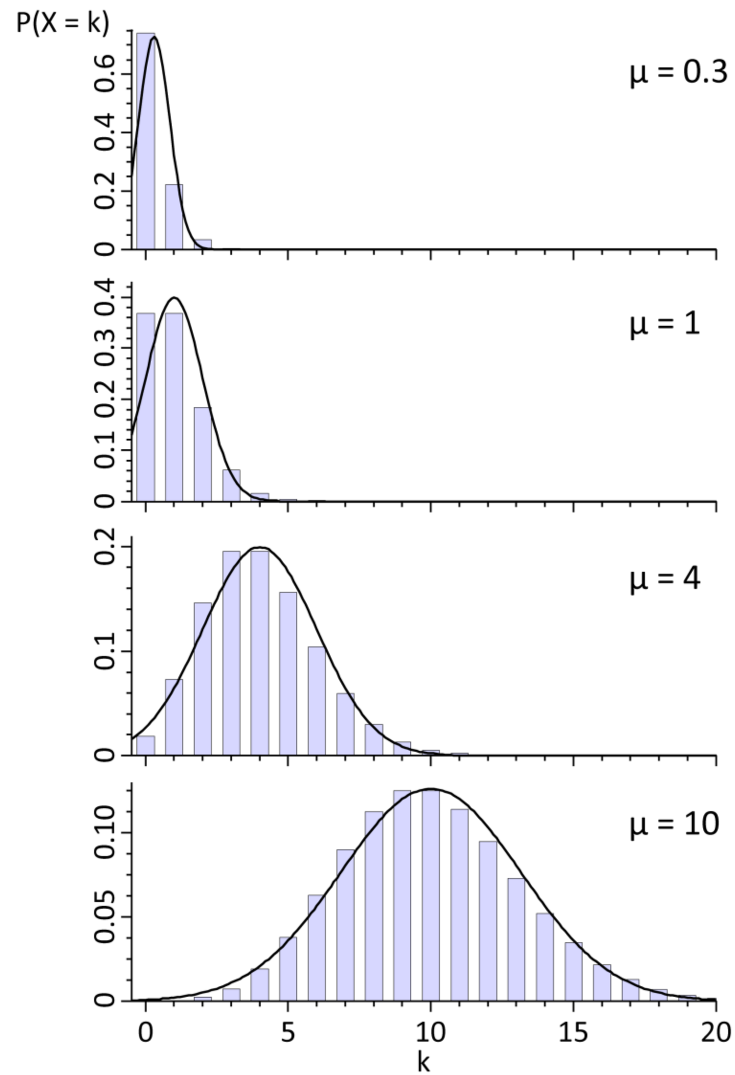
$$P(X = k) = \frac{\mu^k e^{-\mu}}{k!}$$

- Poisson distribution is characterized by the mean count rate, μ (not integer!)
- Standard deviation is not a free parameter:

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

- For large μ Poisson distribution approximates Gaussian

Poisson and corresponding Gaussian distributions

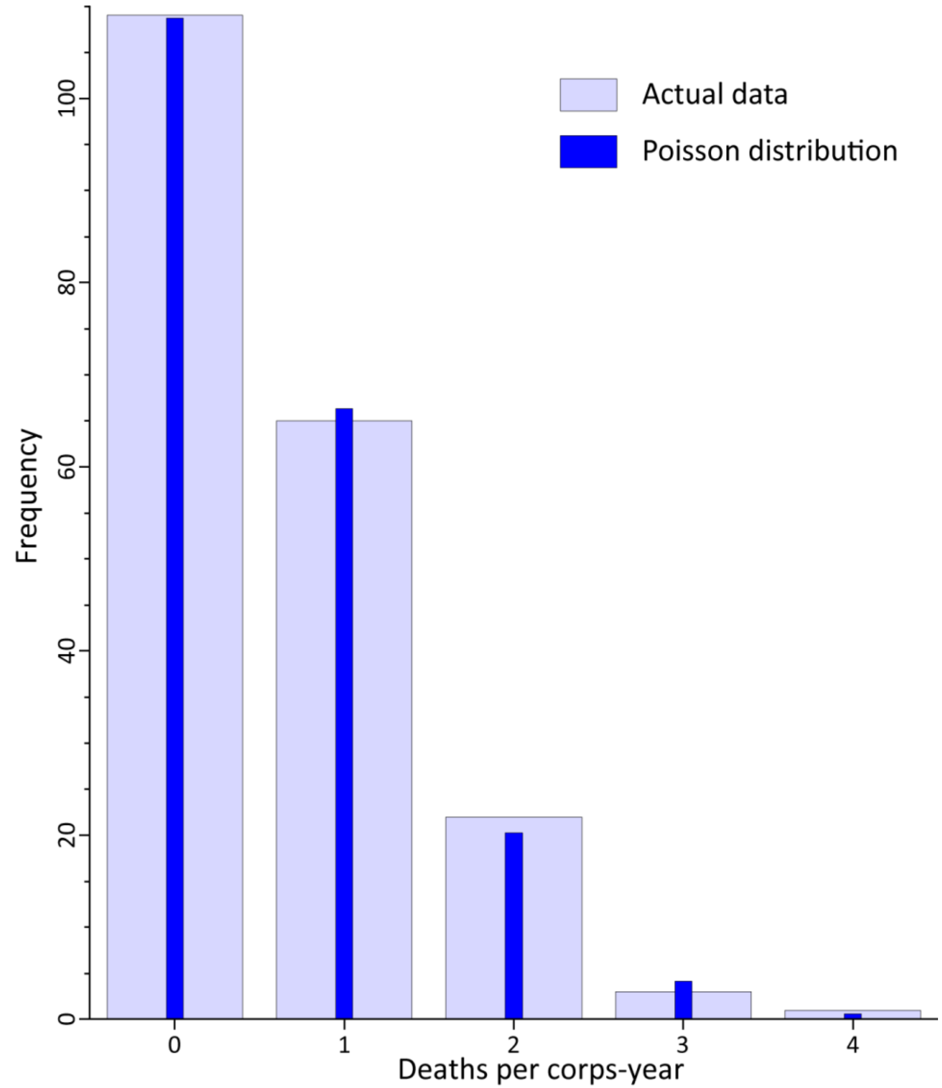


Example: Poisson distribution

- von Bortkiewicz (1898) "*Das Gesetz der kleinen Zahlen*"
- Number of soldiers in the Prussian army killed by horse kicks
  - 10 army corps, 20 years of data
  - Deaths per year per army corps
- One year in one corps there were four deaths – investigation started

- Death distribution follows Poisson law
- mean = 0.61 deaths / corps / year

- 4 deaths in a corps-year are expected to happen from time to time!
- $P(X = 4) \; = \; 0.035$ in 10 corps
- On average it should happen once in 29 years

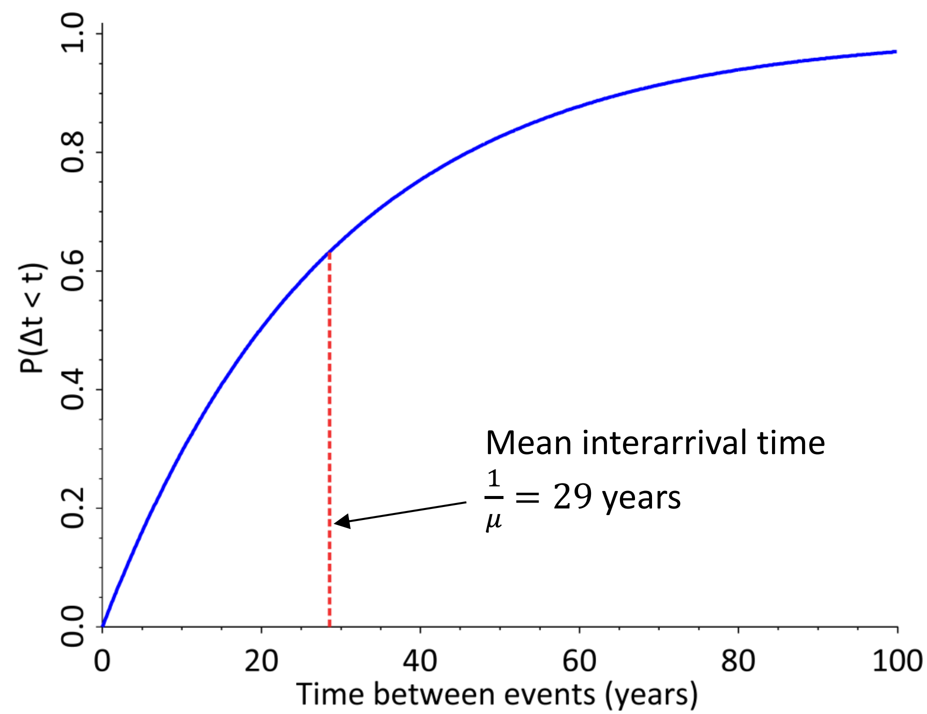


Interarrival times

- How long do we need to wait for the next event to happen?
- Time between two events, ΔT, is called interarrival time
- It is a random variable with cumulative distribution

$$P(\Delta T < t) = 1 - e^{-\mu t}$$

- Probability of observing at least one event in time t
- Mean interarrival time is $\frac{1}{\mu}$
- However, random events occur randomly, so there is no periodicity!
- "On average once in 29 years" does not mean "every 29 years"



Cumulative distribution of interarrival times between 4 deaths in one corps-year ($\mu = 0.035$ per year)

If you play National Lottery once a week, the mean interarrival time between the jackpots is $\frac{1}{\mu} \approx 269{,}000$ years.

Exercise: Poisson distribution

- Poisson law:

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

- You transfect a marker into a population of $n = 3 \times 10^5$ cells
- It functionally integrates with the genome at a rate of $r = 10^{-5}$
- What is the probability of having at least one cell with the marker?

- First calculate the mean (expected) number of marked cells:

$$\mu = nr = 3$$

- Now we can use the Poisson law to find $P(X = 0)$

$$P(X = 0) = \frac{3^0 e^{-3}}{0!} = \frac{1 \times 0.05}{1} = 0.05$$

- Hence, the solution

$$P(X > 0) = 1 - P(X = 0) = 0.95$$

Binomial distribution

- A series of n "trials"
- Probability of "success" in one trial is p
- Probability of "failure" in one trial is $1 - p$
- What is the probability of having exactly k successes in n trials?

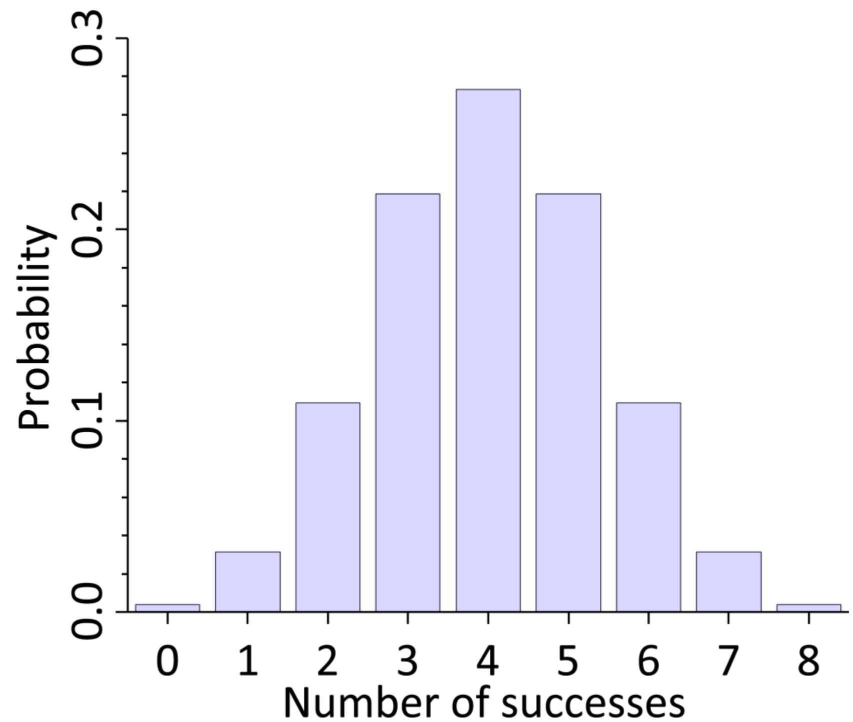
- Binomial distribution

$$\mu = np$$

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

- For large n binomial distribution approximates a Gaussian

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

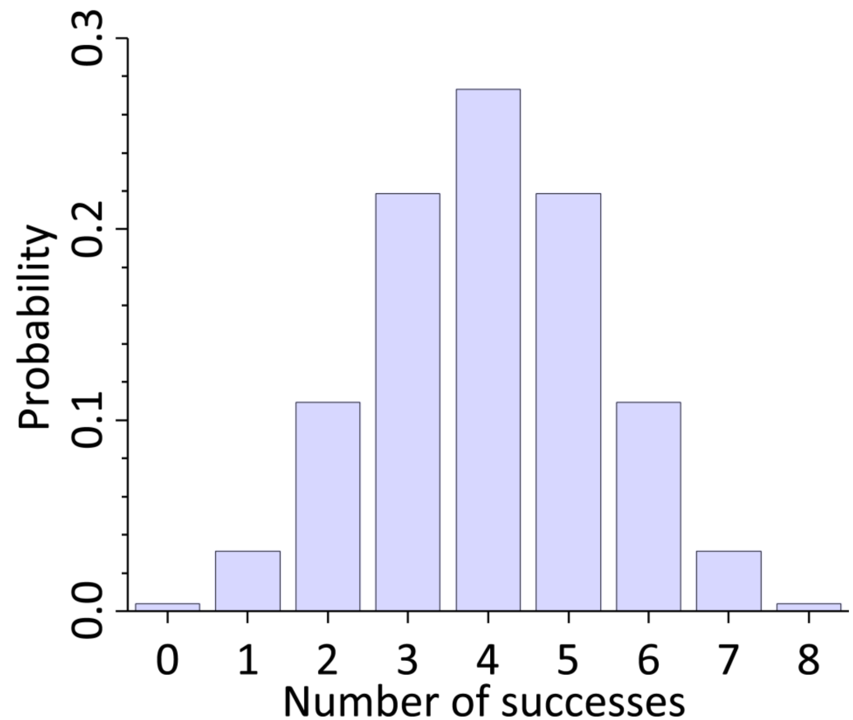


Example: toss a coin
heads = success ($p = 0.5$)
tails = failure ($1 - p = 0.5$)

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

Exercise: tossing a coin

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



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

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

Exercise: tossing a coin

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

- There is only one way of having 8 heads

  H H H H H H H H

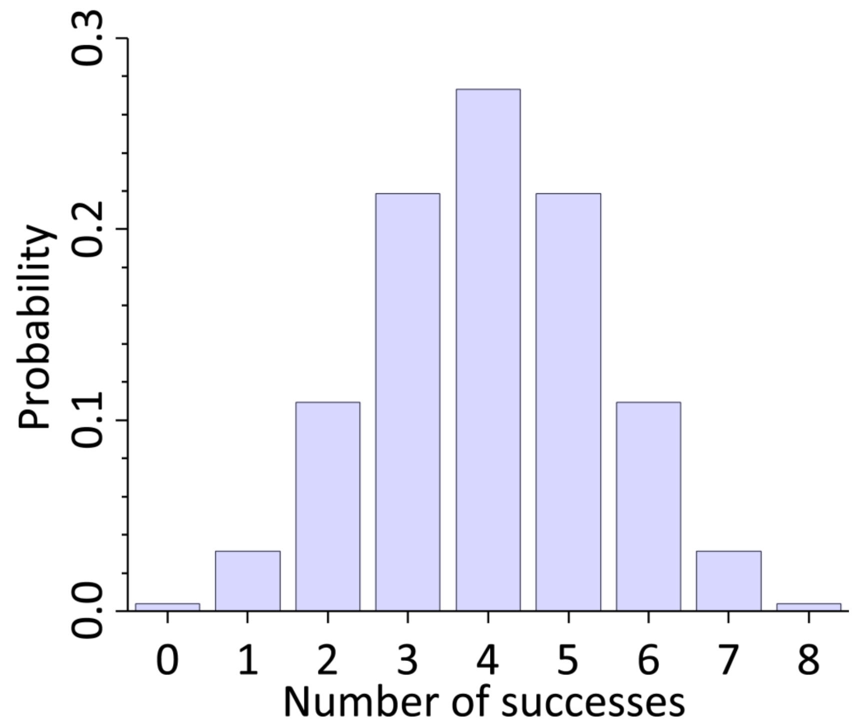
- There are $\binom{8}{4} = 70$ ways of getting 4 heads and 4 tails

  H H H H T T T T

  H H H T H T T T

  H H H T T H T T

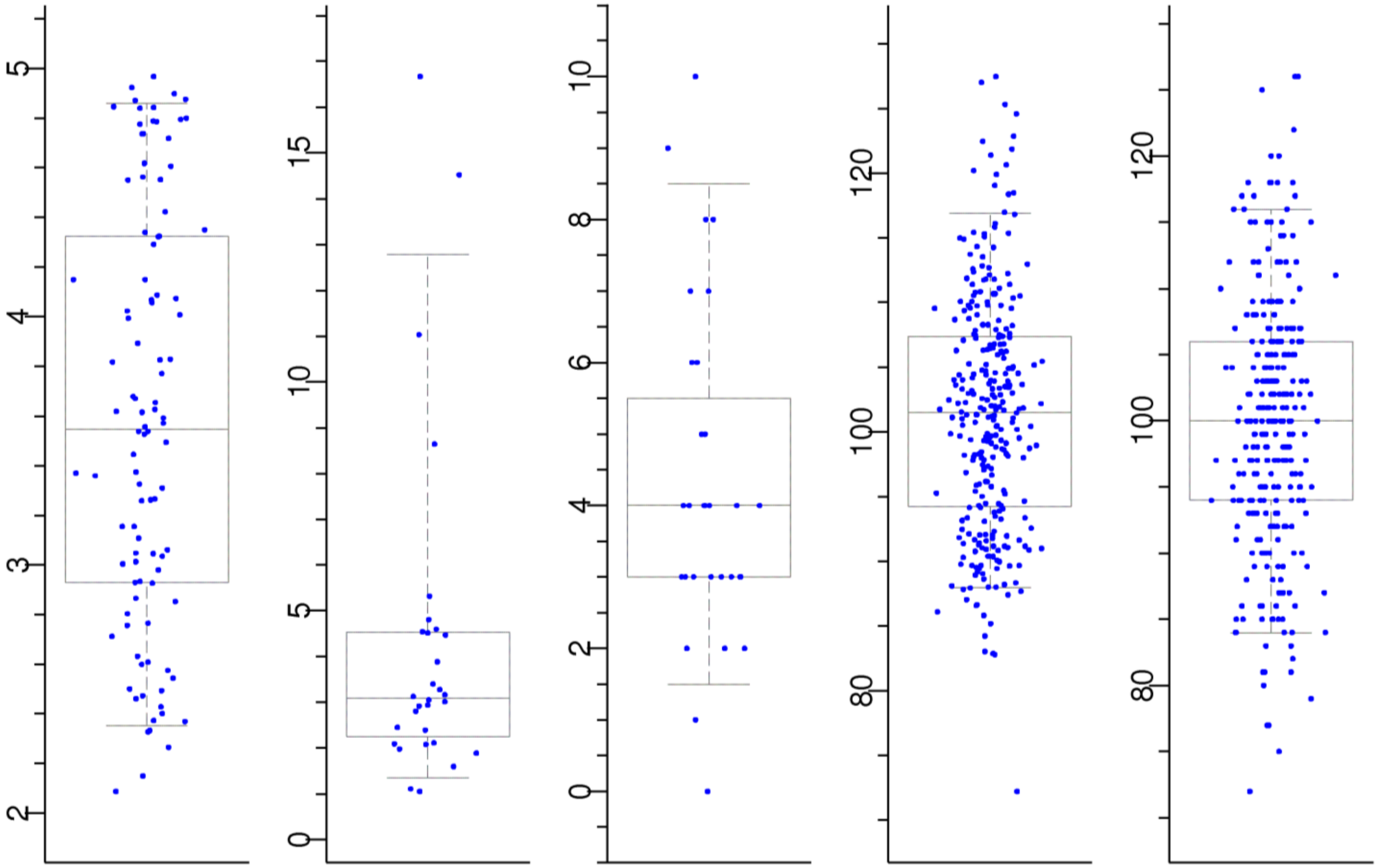
  …



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

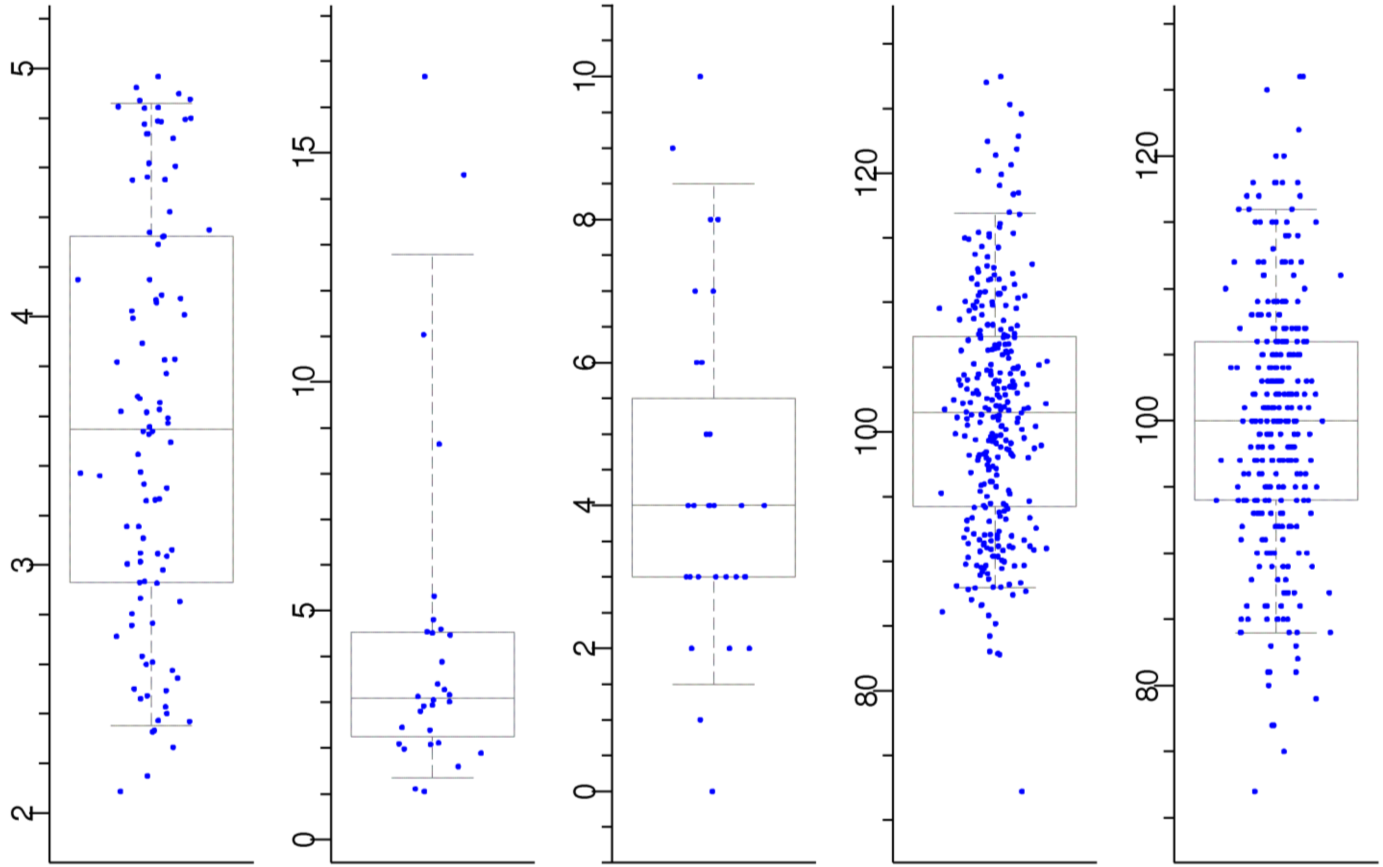
What is the probability of obtaining k heads from 8 coins?

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

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

Please leave your feedback forms on the table by the door