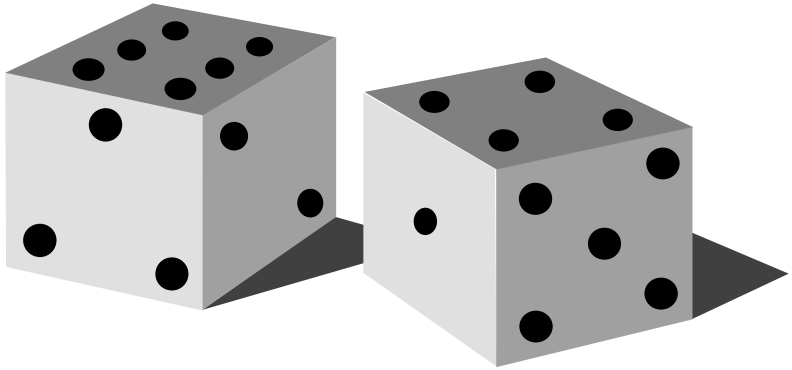
Intro to statistics



Probability:

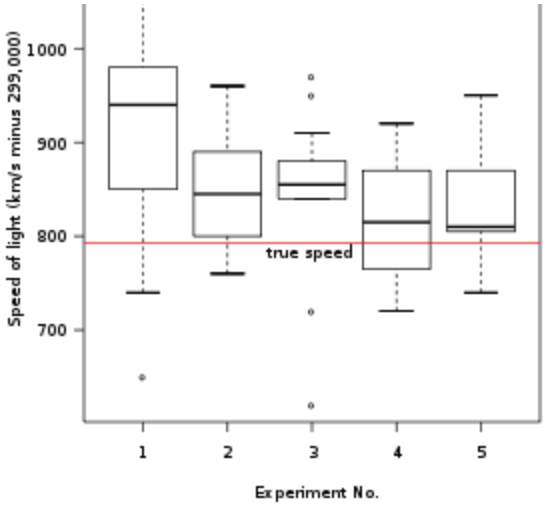
the scientific approach to uncertainty

R-Stats

What do we know for sure?

- Some theories of the past are not believed now…
- Some things considered true by some scientists are not by others…
- Scientists can agree that some scientific knowledge is uncertain...
- **because the world is noisy, chaotic, complex…**

Michelson – Morley experiment, 1887

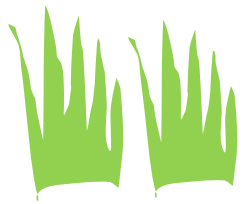



```
> data(morley)
> boxplot(Speed
  ~ Expt, data=
  morley)
```

www.britishhomeopathic.org/research/the_evidence_for_homeopathy.html

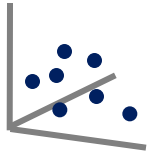
44% of randomised controlled trials in homeopathy have reported positive effects; only 8% have been negative.

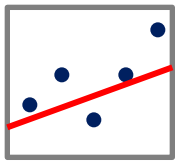
Welcome to biology!

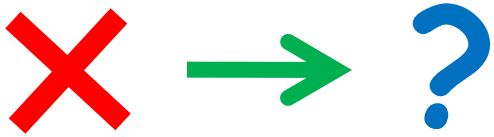


Purposes of statistics

- Explore patterns in data
- Predict or estimate a parameter
  - ML: What value would make data most likely?
- Take subjectivity out of decision-making
  - H_0… when $P < 0.05$, it's time to act.
- Test a null hypothesis
  - H_0… when $P < 0.05$, there's something there...
- Test a hypothesis
  - When $P < 0.05$, we look for a new hypothesis.







Take smoking…

- How do you *know* smoking kills?
  - Theoretical ideas suggest it does more harm than good.
  - But there are smokers who live to 100 and non-smokers who get lung cancer.
- Smokers may be *more likely* to get lung cancer…
  - We could take a sample of people and measure a statistical link (correlation) between smoking and cancer
  - The statistical link will never be exactly 0. But if there's no causal link, the correlations from many samples should cluster around 0. So, we need to find the distribution it would have if there's no real link.
  - This allows us to say exactly how likely a link like the one we observed would be, if there's no real causal link.

Or coins…

We assume coins are symmetrical (theory)…
How would you recognise a biased one?

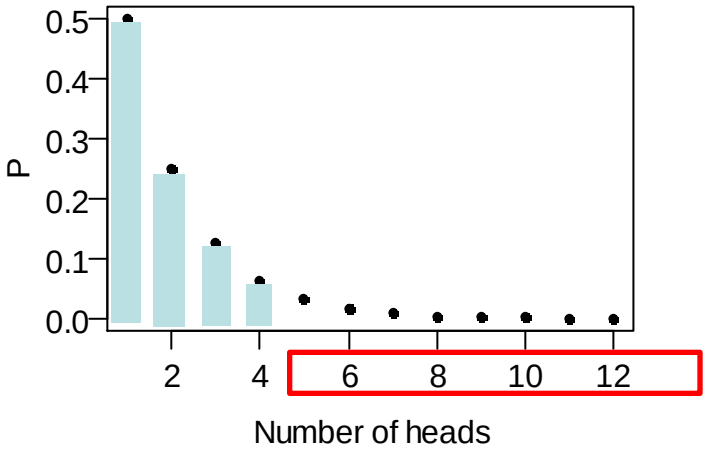
- 2 heads in a row? P=0.25
- 3 in a row? P=0.125
- 4 in a row? P=0.0625
- 5 in a row? P=0.03125
- 10 in a row? P=0.0009766
- 100 in a row?

  P=0.000000000000000000000000000008

All's ok

All's not ok

Interpreting the improbable



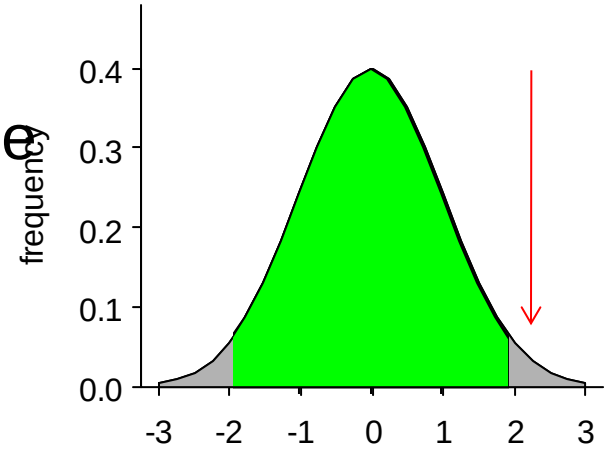
If $P < 0.05$, we say the observed data are so *unlikely* to occur by chance that we're justified in thinking there's a cause (biased coin).

- We can't "prove" the coin is biased (or that smoking kills)
- We just say it's unlikely that the coin is fair – the more unlikely the observed data, the more certain you become.
- Theory says that a sequence of 10H will appear every 18 mins on average (at one flip per sec); 100H in a row only once every 10^{22} years!

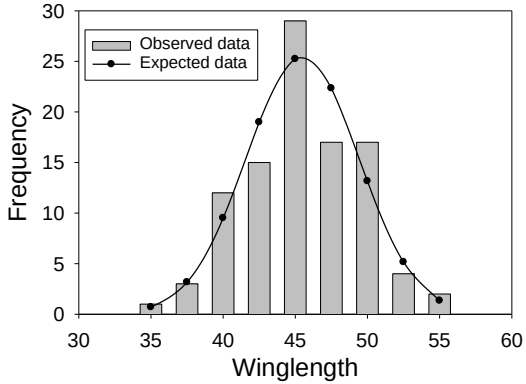
What do we know for sure?

- Find something we can measure (a **variable**) and choose a **statistic** to summarise data into a single number (e.g. mean).
- Make a **precise hypothesis** about the statistic (e.g. mean = 0, or something more interesting…)
- Measure the variable lots of times, look at its **distribution** and calculate the statistic. (It won't be exactly what your hypothesis said!)

- Now suppose your hypothesised value is true for the whole population. Use what you *know* about the distribution to find the **probability** that your observed statistic could come from it by chance.

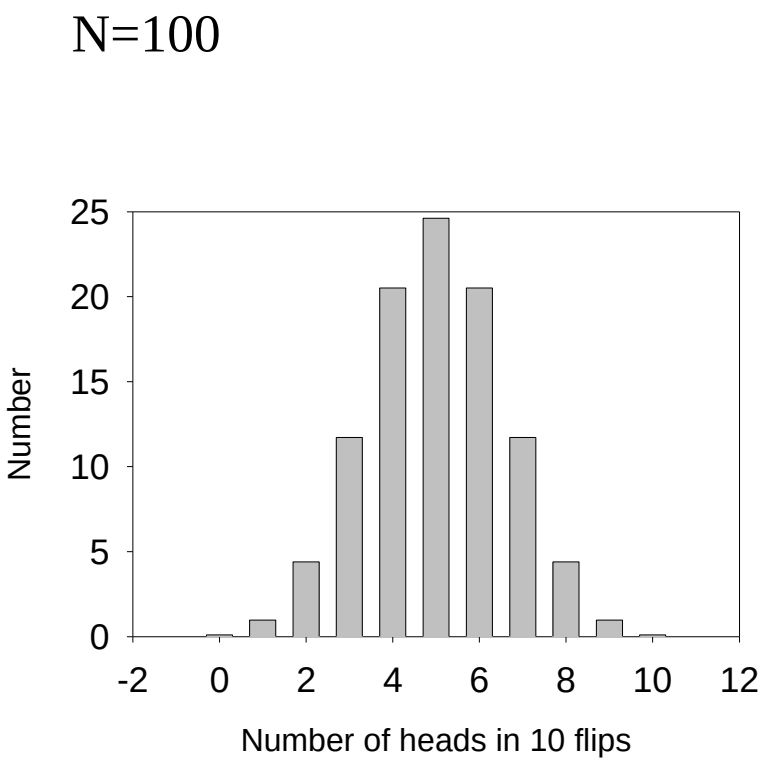


Frequency distributions and models



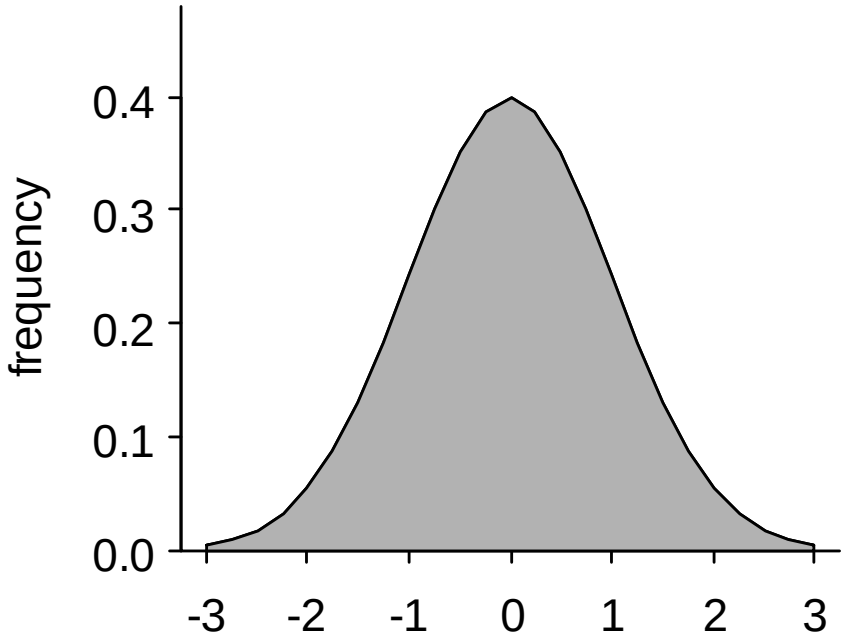
What is Normal?

- If you flip a coin 10 times you would expect 5H and 5T
  - But you would get varying combinations of heads and tails
- flipping 100 sets of 10 would give a *distribution* something like this:



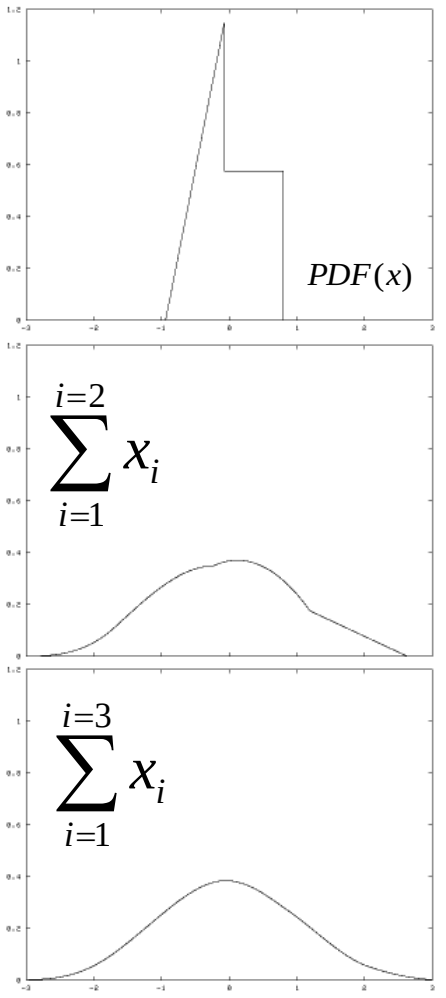
What is Normal?

Measurements of many things have a bell-shaped distribution - where measures around the mean are most common and become rarer away from the mean.

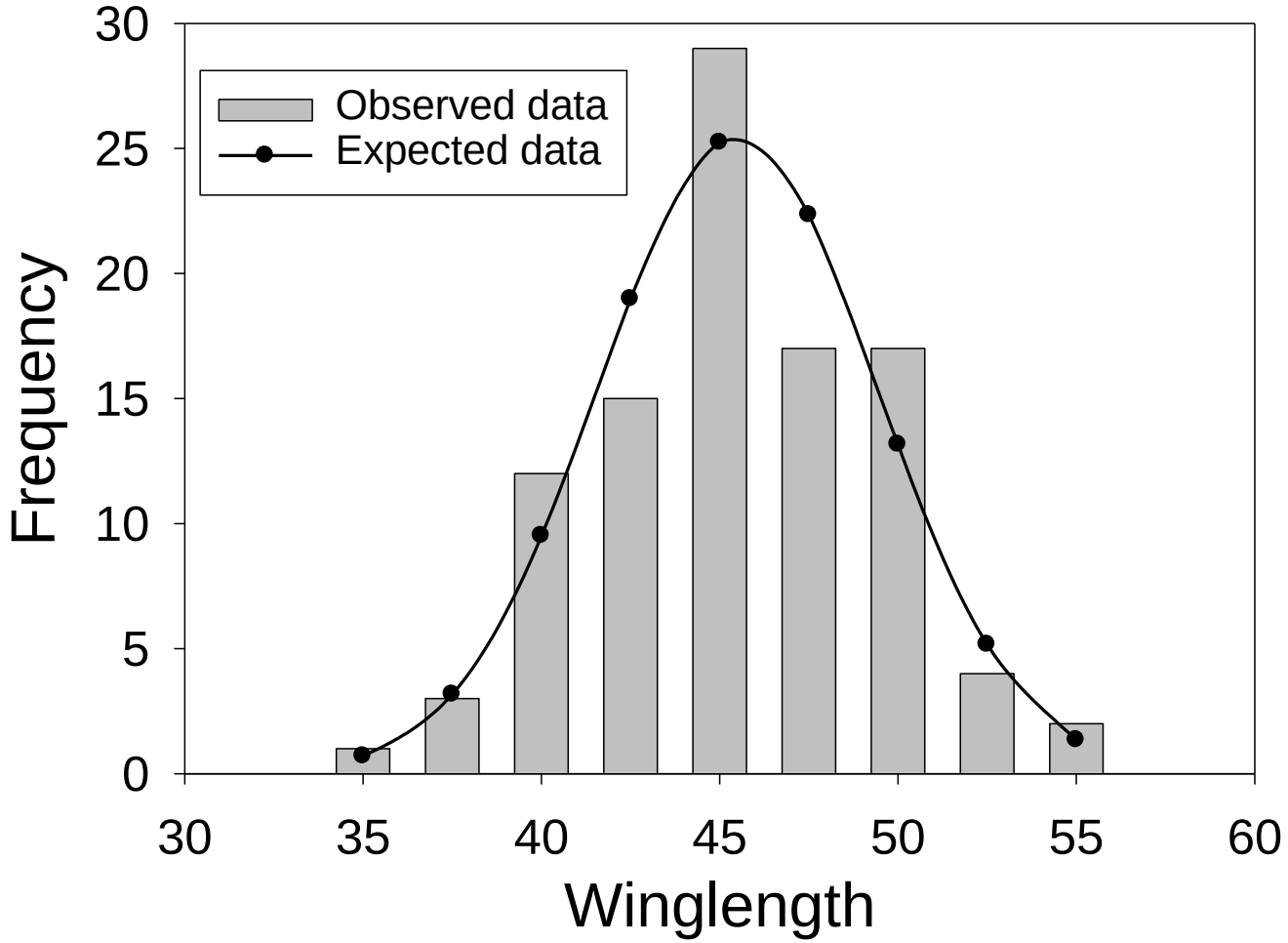


A model of normality…

- For many data there is an appropriate theoretical model
- The model is an *ideal distribution* – the one we would get with a very large sample
- For a continuous variable with no limits, the model is often the *Normal* or *Gaussian* distribution



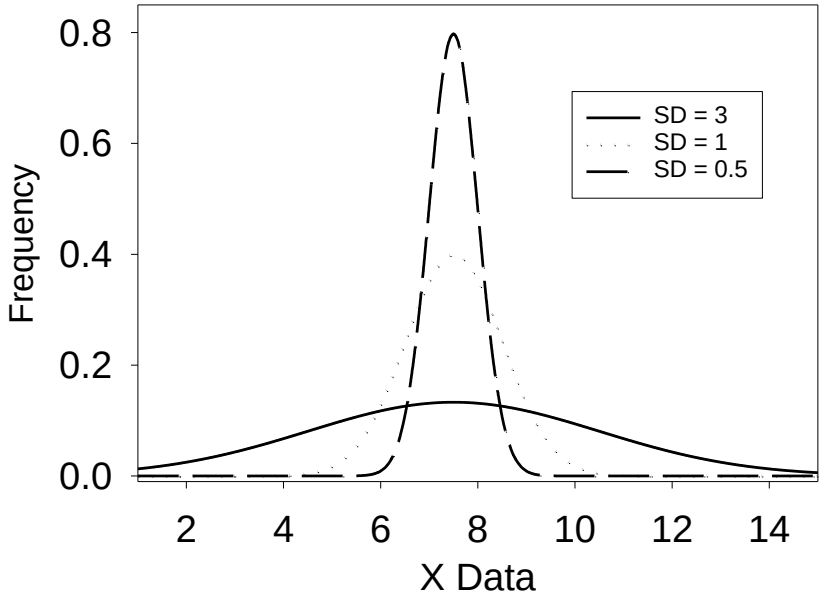
Central limit theorems are a set of weak-convergence results in probability theory. They basically prove that any sum of many independent identically-distributed random variables is approximately normally distributed. These results explain the ubiquity of the normal distribution.


Observed data
Expected data
Frequency
Winglength
0
5
10
15
20
25
30
30
35
40
45
50
55
60

A Normal distribution is defined by:

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

- two constants: π & e
- two *parameters* – the expectation (μ) and standard deviation (σ)
  - μ is the average ("location")
  - σ describes the spread ("dispersion")



also called "Gaussian"

If it looks Normal, estimate μ and σ

- To estimate μ, just calculate the mean of the data
- To estimate σ we want an "average" distance between each point and the mean

  Find the distance between the mean and each point

  "deviation" or "residual"

  Square it (removes the sign)

  Add them up

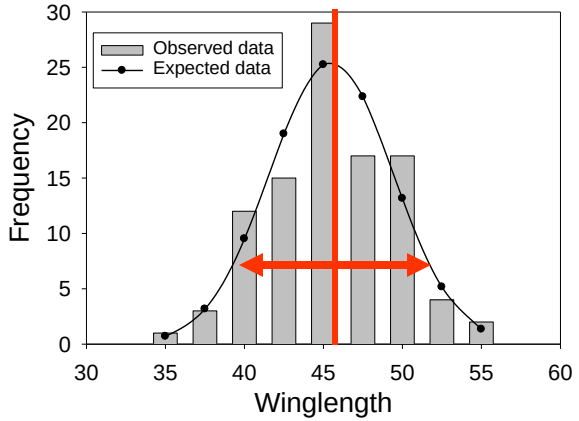
  this is the sum of squares (SS)

  (sum of the squared **deviations)**

  Average it

  this is the mean square (MS)

  also known as **variance**



Degrees of freedom

Wait a moment!

- How did we find the mean?
- If we calculated it from the data, we've used the data to help minimise our mean square!
- so for estimating one parameter from the data, we "pay" one data point – one *degree of freedom* – so we actually divide by N – 1

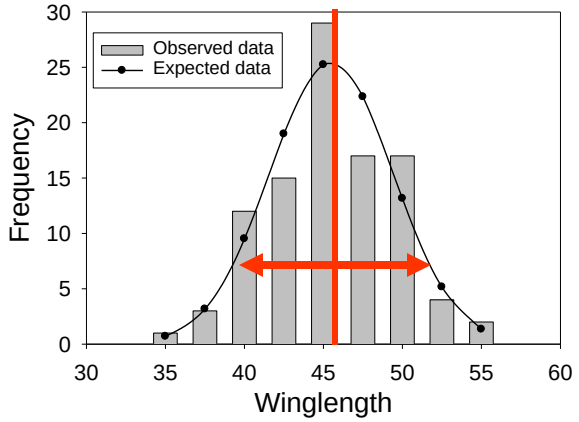
Each data point is a piece of information – so total "bits" of information in data = N

Calculating a mean, etc, makes use of this information: if you know the mean and N-1 data you can calculate the final point.

3 data: 3, 5 & 13 → sum = 21; mean = 7.

So if you have the mean and N-1 data points you automatically know the last datum (there is no "freedom" in its value).

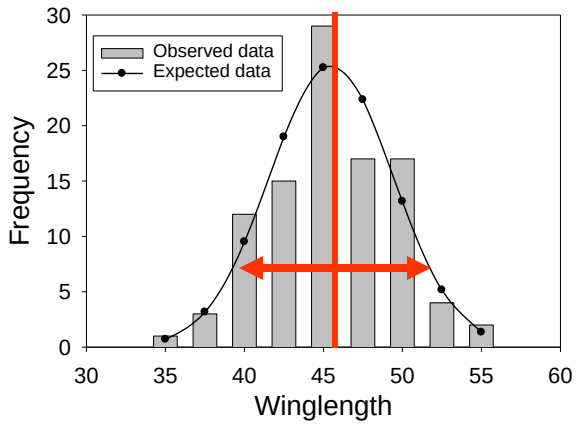
Generally, d.f. = N – number of parameters calculated



Estimating σ

- SS/(N-1) gives the "*mean square*" (MS) or *variance.*
- √Var gives the "standard deviation" (SD), in units the same as the mean
- and this is our best estimate of σ

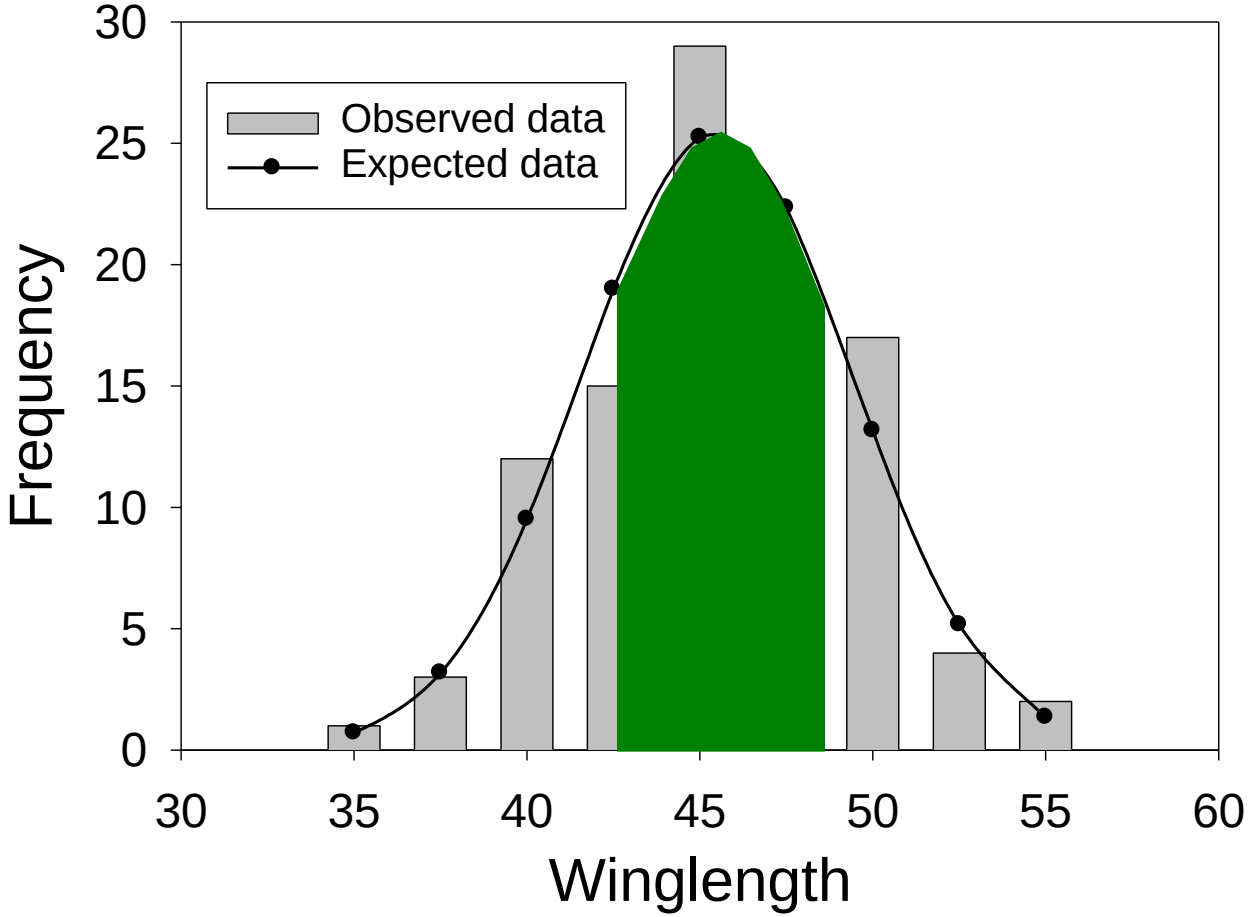
$$SD = \sqrt{\frac{\sum_{1}^{n}\left(x-\bar{x}\right)^2}{n-1}}$$



From a frequency distribution come probabilities….

If you have a mathematical model to describe the distribution, you can estimate future probabilities based on past frequencies

- *if you assume the data come from a Normal distribution, you can calculate the mean and SD and predict the probability of getting a value in any "slice" of the distribution (probability of any exact value is 0!)*
- *The further away from the mean, the more unlikely it is to sample that point at random*



- 50% of the area under the curve is shaded...
- There is a 50% chance that any point selected at random will be this close to the mean (42.9 - 48.1)

Mean=45.5 SD=3.92

A fly picked at random from this “population” described by the normal distribution will be…

Range	**Probability**
Greater or Less than the mean	50%
Mean ± 0.67 standard deviations	**50%**
Mean ± 1 standard deviation	68.3%
Mean ± 1.96 standard deviations	**95%**
Mean ± 2 standard deviations	95.5%
Mean ± 2.58 standard deviations	**99%**
Mean ± 3 standard deviations	99.7%
Mean ± 3.31 standard deviations	**99.9%**

Statistics can compare populations…

when you can't measure every individual in the population

- so you take a "*sample*" from the population
- and try and work out what the population characteristics are; the bigger the sample the better are your estimates.

[NB: Population parameters have Greek symbols, sample parameters have Roman ones. SD is σ when it refers to a population, *s* when it refers to a sample.]

Parametric models

- when the data in a sample are used to estimate the *parameters* of an assumed distribution (e.g. mean and SD for Normal), which in turn are used to estimate probabilities

- Non-parametric models do not estimate distributional parameters (though may make distributional assumptions)

Sampling error

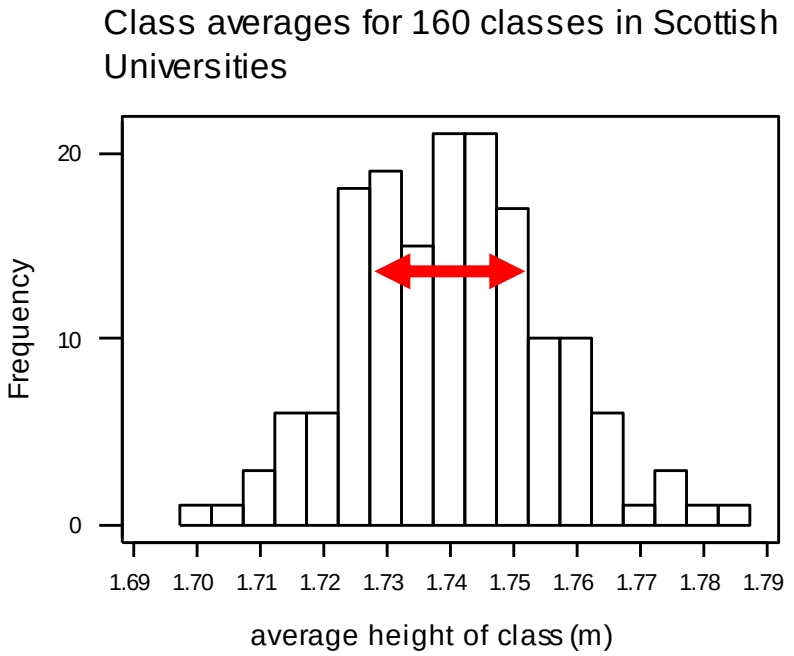
- When you sample, the sample will not be completely representative (no individual is truly average!)
- A *sample mean* will not be exactly the true average of the population
  - but the larger the sample, the more the differences average out
- Sampling “error” means sampling “noise”

For example…

If you sampled many classes and measured the heights of males, you'd get a **mean for each sample** and lots of sample means would give a distribution

- The SD of a **distribution of means** is called the Standard Error of the mean.
- It can also be estimated from a single sample as SD/√N

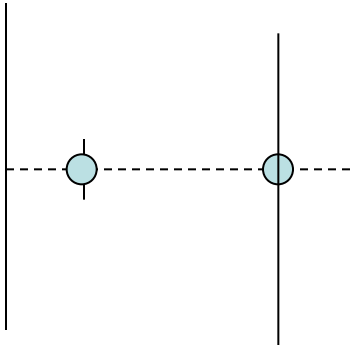
or √(Var/N), because of the definition of the variance



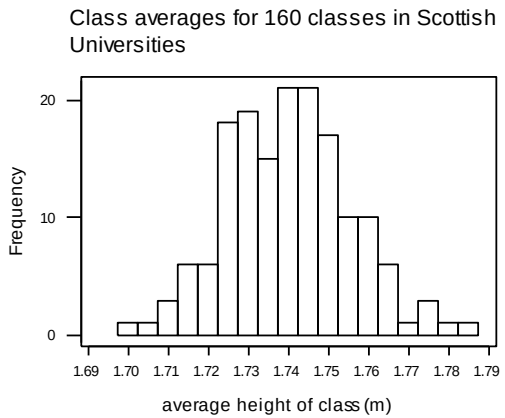
The SE

- The *standard error of the mean* measures how "error-prone" your estimate of the mean is
  - i.e. it estimates how repeatable your estimate of the mean is

Data	*120, 189, 195 vs 167, 168, 169*
Means	*168 vs 168*
SDs	*41.7 vs 1*
SEs	*24 vs 0.6*



As the SE is simply a SD of means…



- it tells us about the probability of getting a mean at any given distance from the true mean
- most samples are going to have means which are close to the true mean

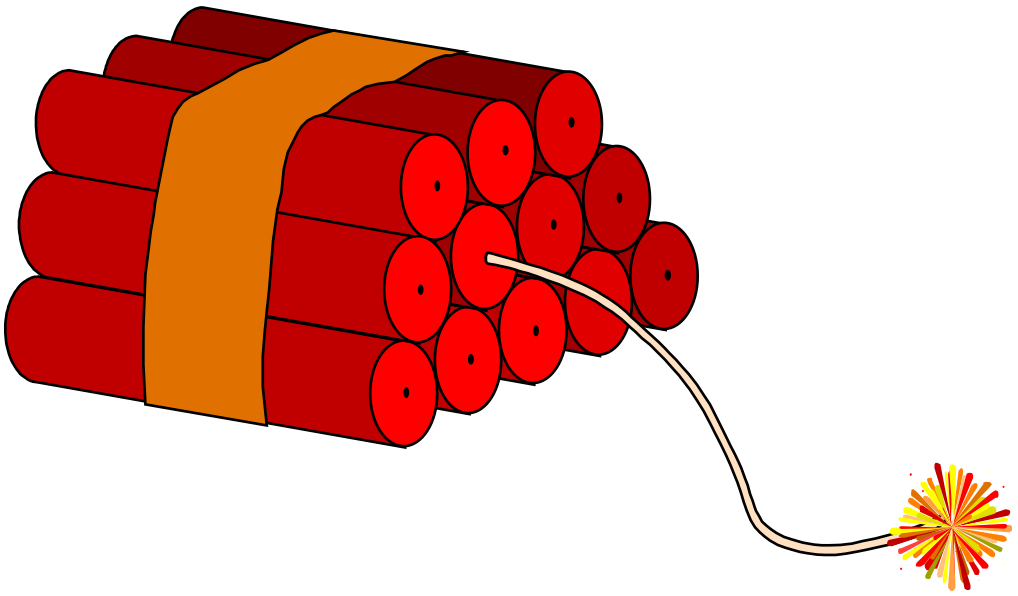
  (within 1-2 SEs)
- samples with poor means are going to be rare

As the SE is simply a SD of means…

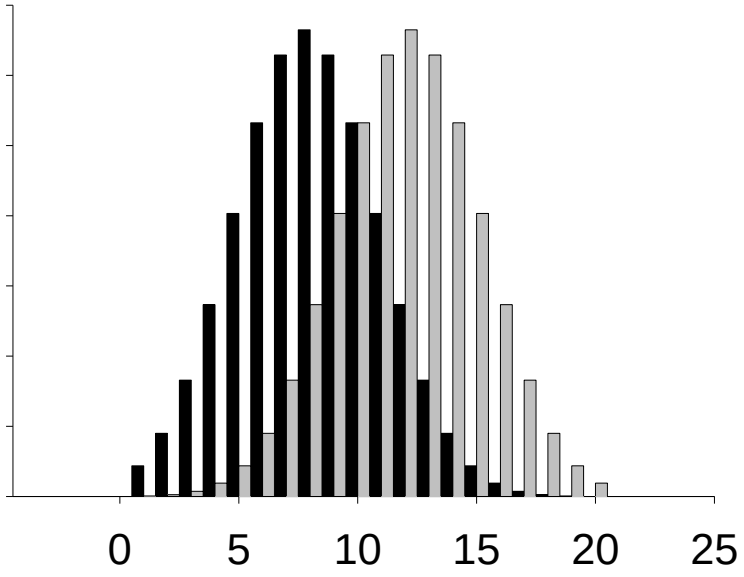
- the "distance" from the "real" mean to the sample mean will occur with these probabilities:

Range	**Probability**
Greater or Less than the mean	50%
Mean ± 0.67 SEs	50%
Mean ± 1 SE	68.3%
Mean ± 1.96 SEs (95% CI)	95%
Mean ± 2 SEs	95.5%
Mean ± 2.58 SEs	99%
Mean ± 3 SEs	99.7%
Mean ± 3.31 SEs	99.9%

This is dynamite



Is this sample different from that?



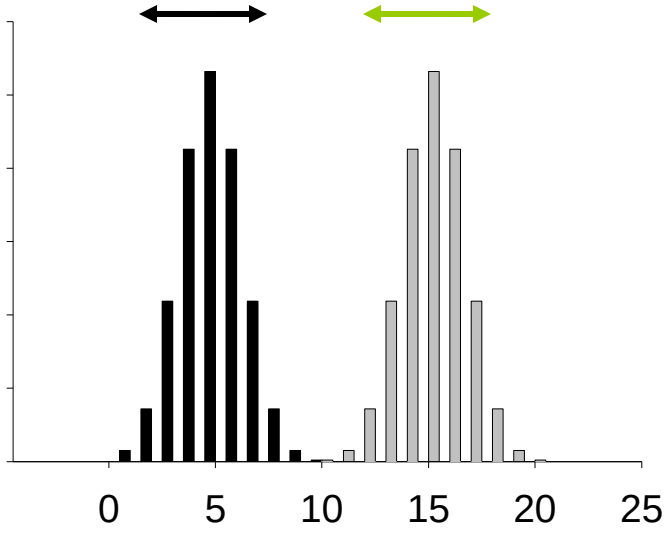
Comparing two (or more) samples…

Any two samples will differ, so how can we decide if they're from the same population?

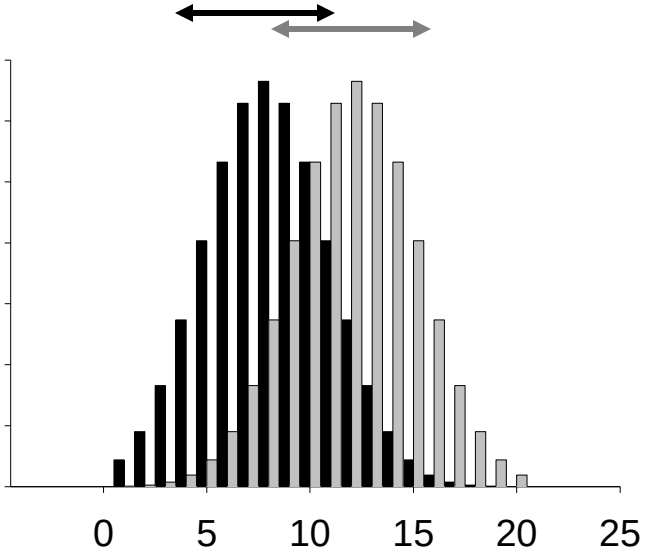
- We look at the variation *within* samples (due to sampling error) and compare it to the *variation between (or among)* samples
- comparing the *variation between* sample means with sampling error*, we ask:
- how *likely* would it be for a single population to yield two samples that differ by the observed amount?

*Compare the variance of the means with the mean variance!

- If the variation within the samples is small compared to the variation between samples then the difference is *unlikely* to occur from a single population.



- Conversely, if the variation between samples is small compared to the variation within the samples then the difference *could be* due to chance.



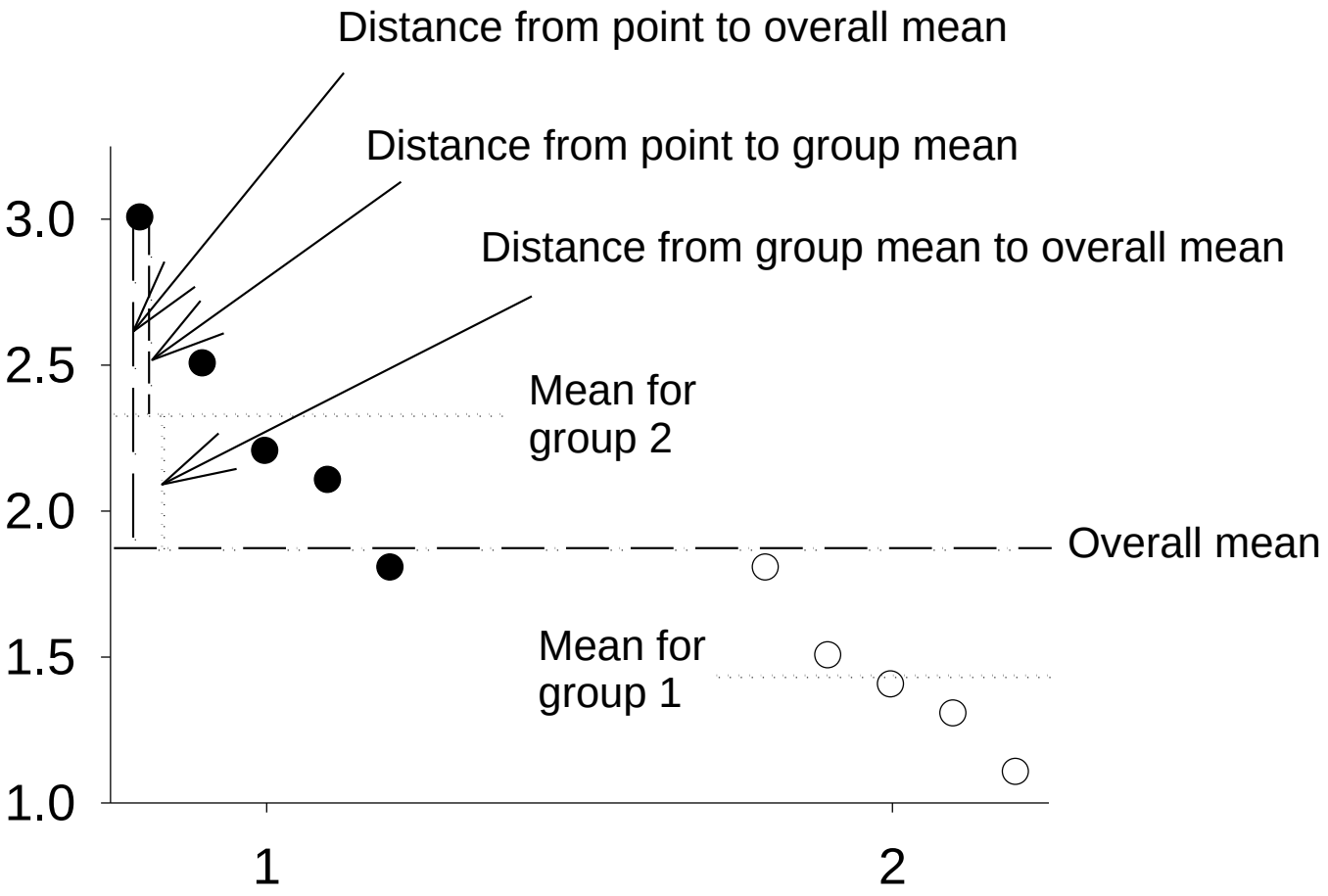
Partitioning the variation...

- To compare the *within*-samples variation to the *among*-samples variation we separate
  - **total variance** (overall MS)
    - using distances from each point to overall mean

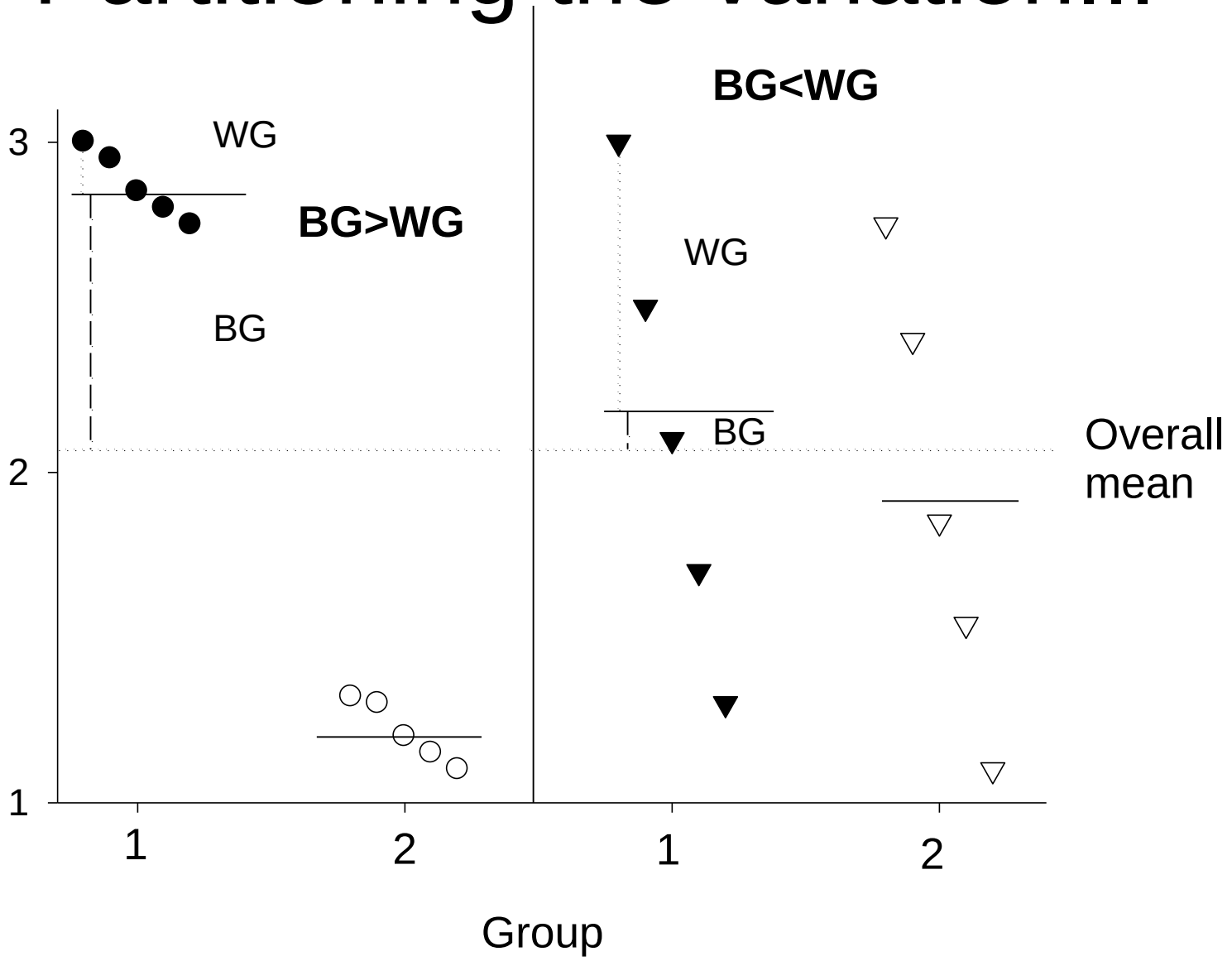
  into two components:

  - **variation among groups** (explained MS)
    - using “among-groups” distances = distance from the overall mean to a group mean
  - **variation within the groups** (error MS)
    - using “within-groups” distances = distance from a group mean to a point (error)

Partitioning the variation...



Partitioning the variation...



From distances to variances…

- All these distances are squared, then added up with other corresponding distances… to give 3 sums of squares (SS)
- To get the 3 mean squares (MS) we need to divide by the appropriate degrees of freedom, which are:
- *N*-1 for the total MS
- *k*-1 for the among-groups MS (*k* groups)
- (*N*-1) - (*k*-1) for the within-groups MS

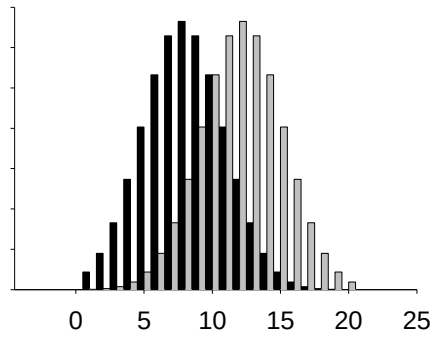
What are these MS?

- A MS is a variance, a measure of spread
- The total MS measures the overall variation within the data.
- If your treatment has an effect, it will increase the overall variation by moving the treatment means apart. The *among-groups* MS measures the spread of treatment means.
- The variation around the treatment means (due to sampling error) is measured by the within-groups MS
  - Error or residual variance
  - A "residual" = distance from a datum to its group mean

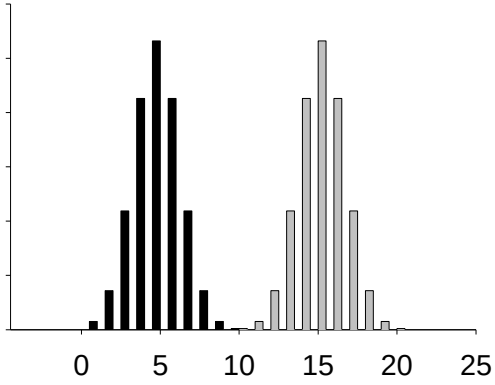
Comparing these variances (MS)

Whether or not a difference between group means is consistent with a single population can be assessed by the *ratio* of between-groups variance to within-groups variance

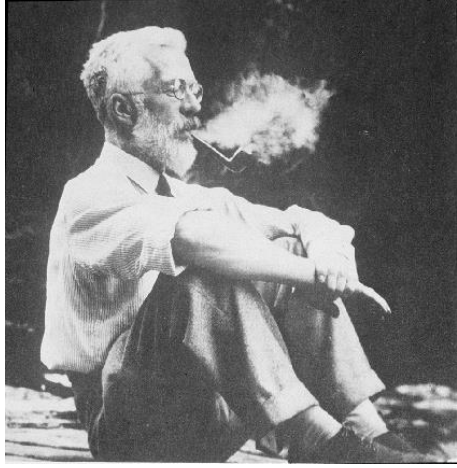
B-G small, W-G big



B-G big, W-G small



The F-statistic



- *F* = among-gps MS ÷ within-gps MS
- If *F* is large, it is *unlikely* the observed variation among means is due to chance
- If *F* is small, the observed variation *could be* due to chance
- *F* is the *variance ratio statistic* (called *F* after Fisher)
- This method of comparing variances is called *Analysis of Variance* (ANOVA)

The ANOVA table

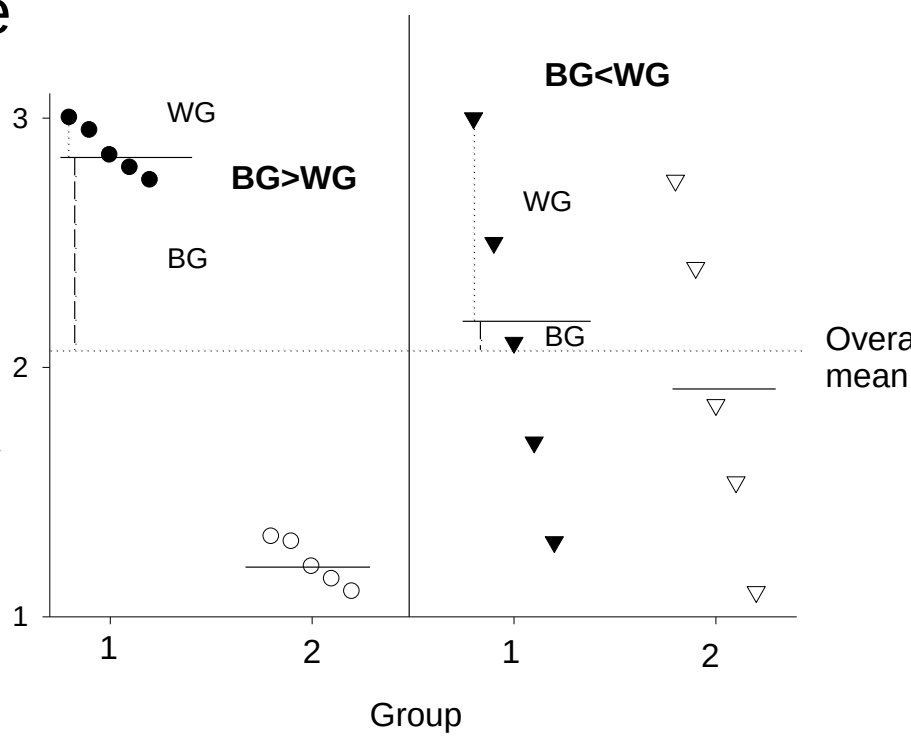
Source	df	SS	MS	F
Treatment	(*N*-1) - (*k*-1)	$\sum_{1}^{K} n_k \left(\overline{x_k} - \overline{\overline{x}}\right)^2$	$=\frac{SS}{df}$	$=\frac{MS_T}{MS_E}$
Error	k – 1	$\sum_{1}^{N} \left(x - \overline{x}\right)^2$		
Total	N – 1	$\sum_{1}^{N} \left(x - \overline{\overline{x}}\right)^2$		

And we can do the same analysis for all kinds of linear models.

Summary

- If a sample's data vary a lot, then a second sample of the same data would be likely to have quite a different mean
- Conversely, if all data were the same, a second sample's mean would be identical
- Therefore, the variability within a sample informs us about the reliability (or variability) of the estimate of the underlying population mean

- So, if there is little variability within the groups, but a large difference between the groups, then it is likely that the means of any future samples will also be different…
  - the likelihood of the observed difference coming from a single population is small ($P < 0.05$)
- Conversely, if groups are very variable and the difference between group means is small, then the means of any future samples may not be different from each other
  - the likelihood of the observed difference coming from a single population is large ($P > 0.05$)



Practical

- Start “Linear Models Part 1”