Hypothesis testing and statistical tests

MOL518, Lecture #11

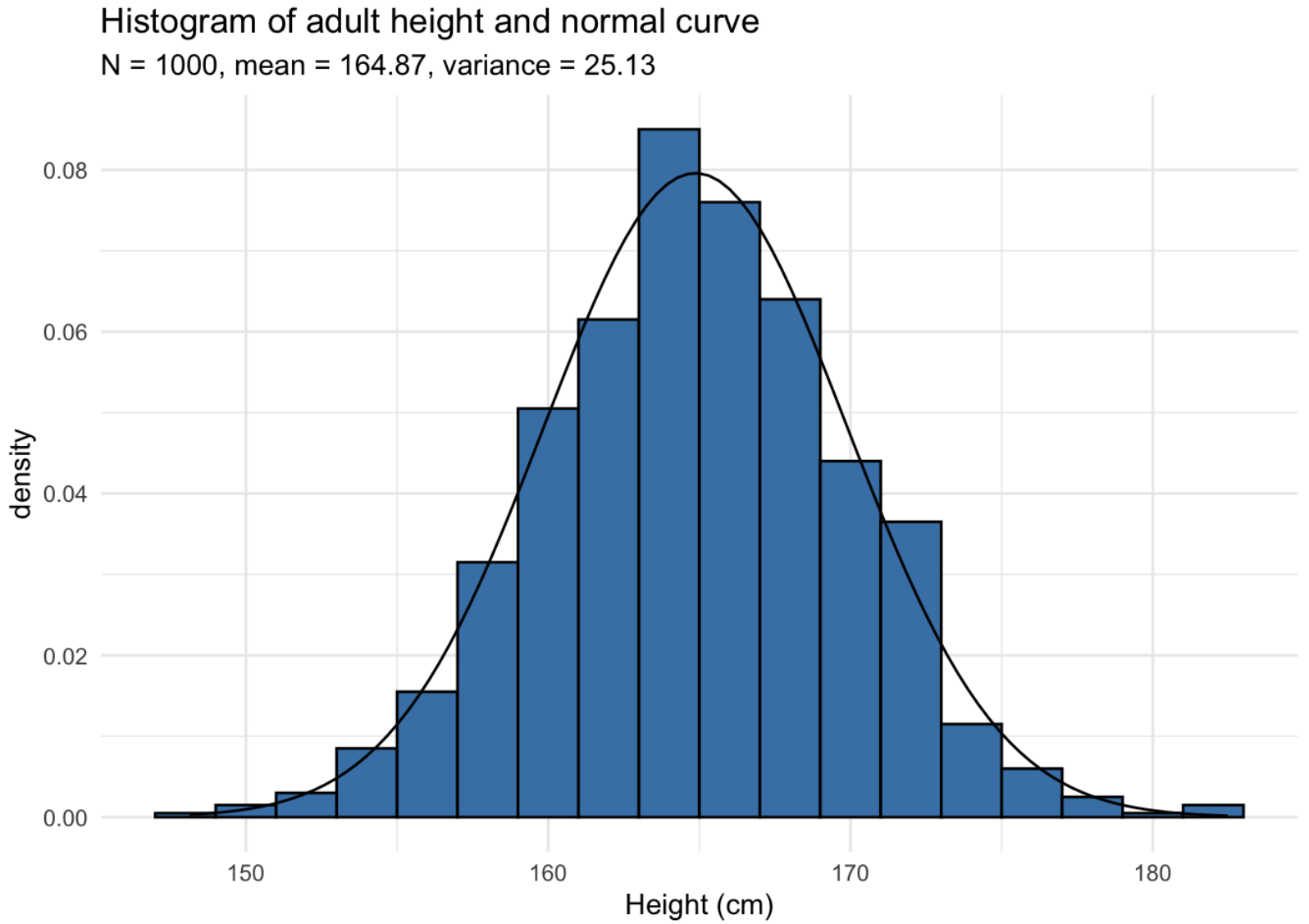
Outline

- What is a statistical test?
- *p*-values
- Types of statistical tests
- Multiple hypothesis testing

Two motivating problems for today

- Assume we have a gene with two alleles, Y and y
- Is the gene under selection, or segregating randomly?
- Assume we have a wild-type (x_1) and a mutant (x_2) bacterium and want to compare the expression level of a GFP reporter
- How do we compare GFP expression between these groups x_1 and x_2?

Introducing histograms!



Source: https://statsandr.com/blog/do-my-data-follow-a-normal-distribution-a-note-on-the-most-widely-used-distribution-and-how-to-test-for-normality-in-r/

Outline

- What is a statistical test?
- *p*-values
- Types of statistical tests
- Multiple hypothesis testing

What is a statistical test?

- Statistical tests allow us to decide whether the data support a particular hypothesis (typically we construct null and alternative hypotheses):

- Null hypothesis examples:
  - The coin is fair
  - The gene is not under selection
  - The mutation **does not affect** GFP expression

- Alternative hypothesis examples:
  - The coin is biased
  - The gene is under selection
  - The mutation **does affect** GFP expression

A very simple example

- Suppose we measure 10 pea plants, and all of them are green (yy)
- Null hypothesis is that yy should occur with probability 0.25: (1/2)*(1/2)
- $P(10\ \text{green}) = 0.25^{10} \approx 10^{-6}$
- This is very unlikely, so we can safely reject the null hypothesis
- It gets hard to directly compute probabilities for larger sample sizes and more complex distributions

How do we perform a statistical test?

1. Define a null hypothesis (and typically an alternative hypothesis)
   - Goal is to reject the null hypothesis (we cannot prove the alternative hypothesis!)

2. Select a relevant statistical test with a test statistic T

3. Select a significance level α (maximum acceptable false-positive rate), typically 5% or 1%

4. Using the data, compute the test statistic and the *p*-value

5. Reject the null hypothesis if $p \geq \alpha$

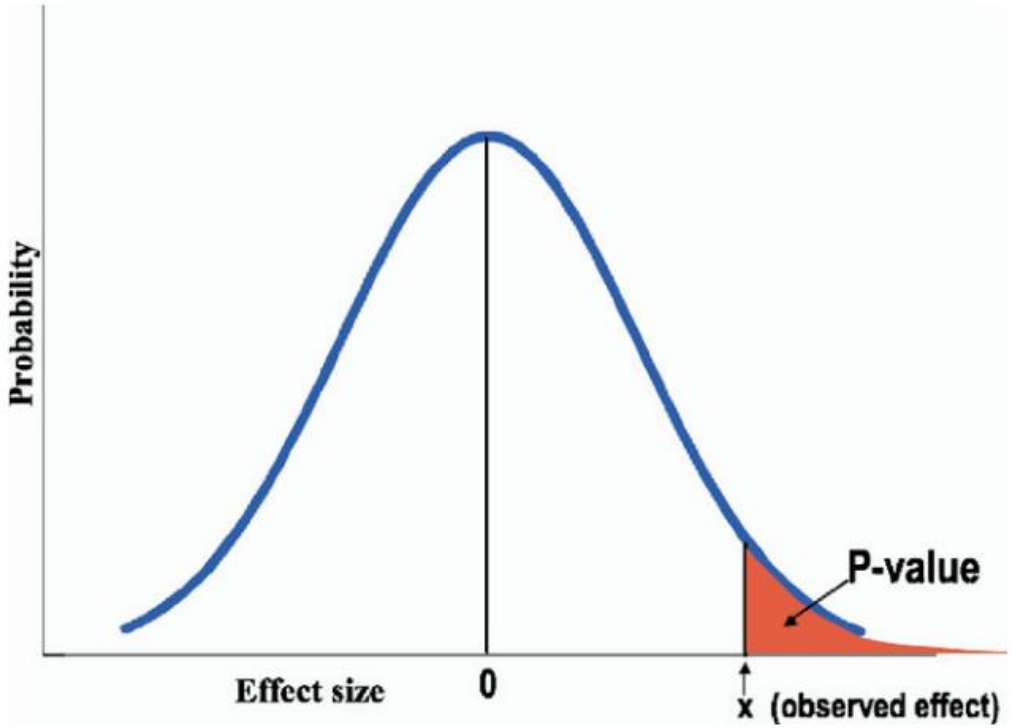
Source: https://en.wikipedia.org/wiki/Statistical_hypothesis_test

Outline

- What is a statistical test?
- *p*-values
- Types of statistical tests
- Multiple hypothesis testing

What is a *p*-value?

- Mathematical definition: $p = P(\text{Data}|\text{Null})$
- Not just probability of the observed data, but also for more extreme data!



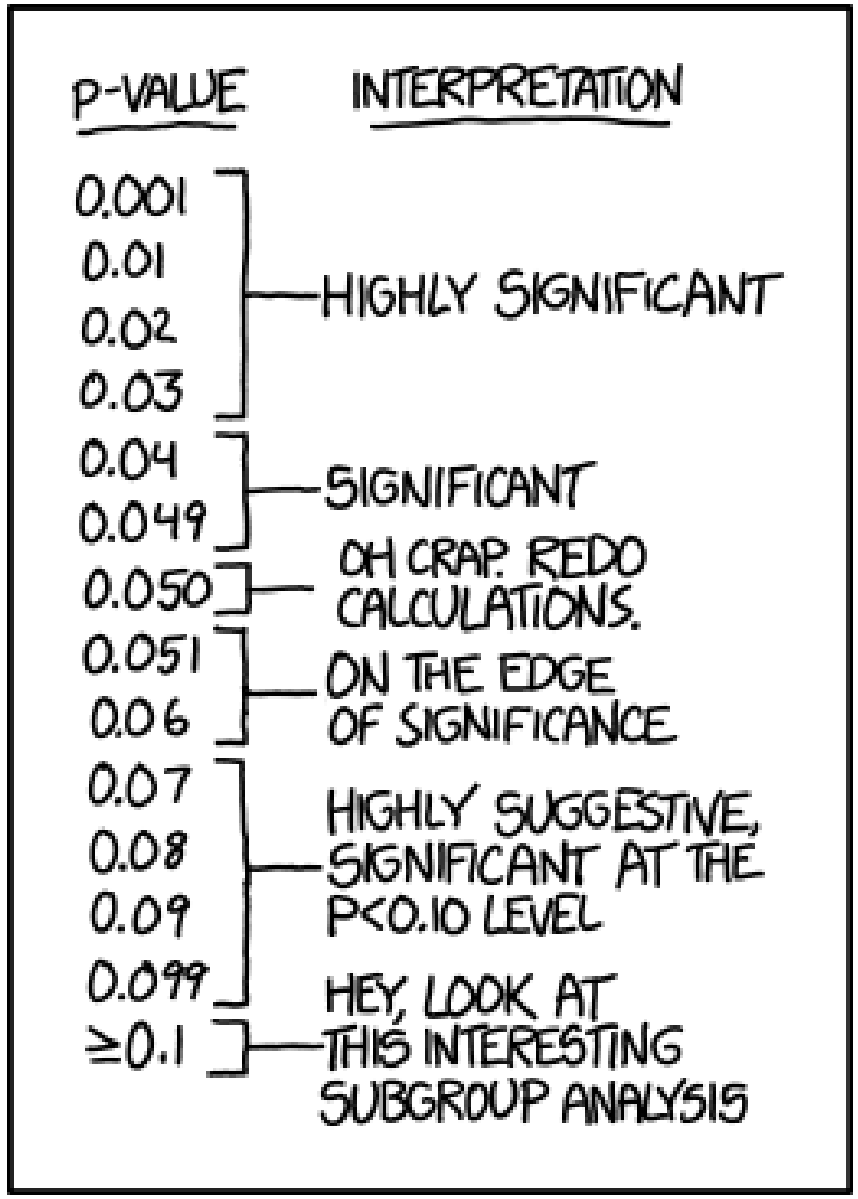
Source: https://www.ohri.ca/newsroom/seminars/SeminarUploads/1829%5CSuggested%20Reading%20-%20Nov%203,%202014.pdf

How do we interpret a *p*-value?

- Common misconceptions:
  - Probability the null hypothesis is true (Fisher himself got this wrong!)
  - Non-significance (e.g. $p > 0.05$) means there is no difference between groups
  - Statistical significance equals importance

- Actually, context matters:
  - What is the effect size?
  - What is the sample size?

Scientists are obsessed with p < 0.05



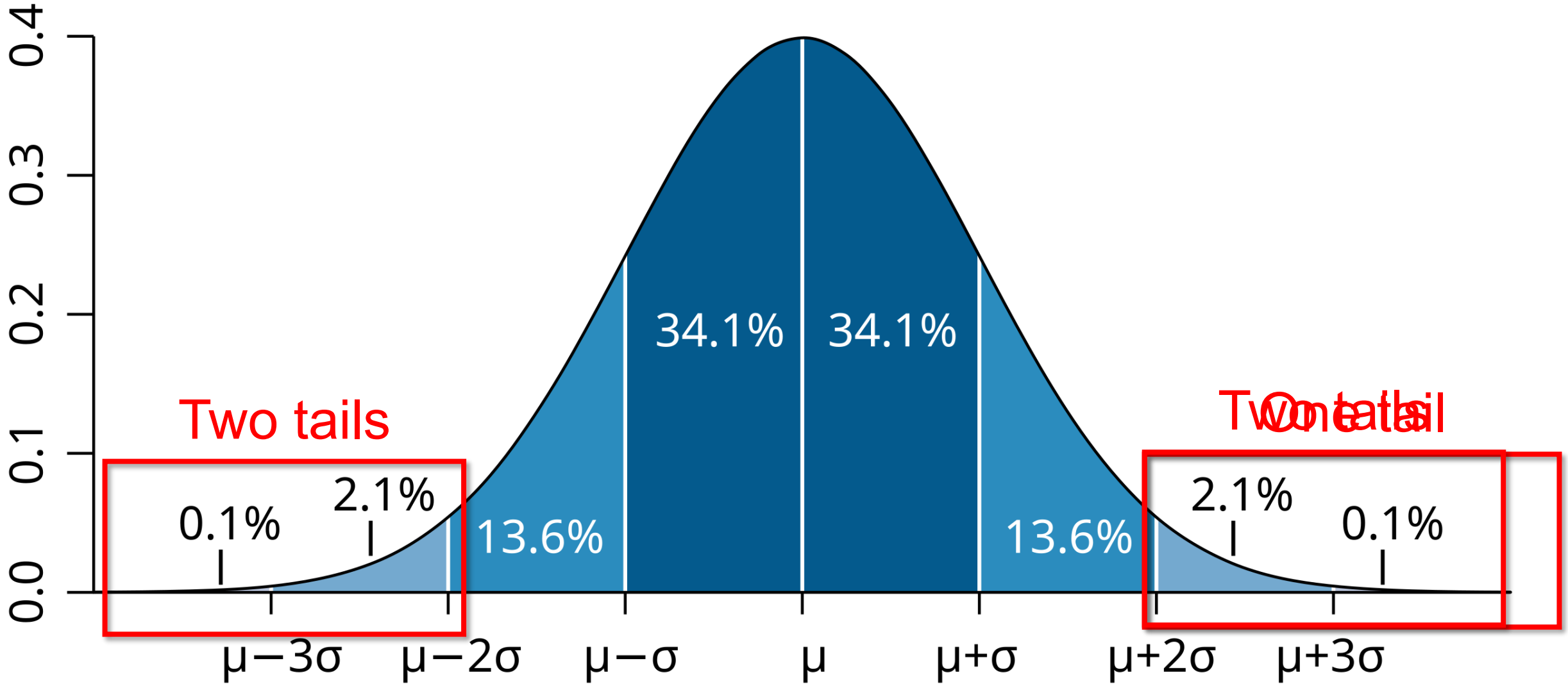
Source: https://xkcd.com/1478/

p-values can be one-tailed and two-tailed

- One-tailed: consider the possibility that a coin is biased towards **heads**
- Two-tailed: consider the possibility that a coin is biased towards **heads or tails**

- One-tailed: consider that the mean of the mutant is **greater than** the wild-type
- Two-tailed: consider that the mean of the mutant could be **greater or less than** the wild-type

Visualizing one- vs. two-tailed p-values



Source: https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Standard_deviation_diagram_micro.svg

When should you use one vs. two tails?

- It depends on the outcomes you are interested in
- Often in science we care about both significant increases *and* decreases
- Examples:
  - A mutation could cause a GFP reporter to go up or down
  - Selection for a gene could be positive or negative
  - A treatment could make a disease better or worse(!)
- One-tailed tests should be used only if the distribution has a single tail:
  - Exponential
  - Chi-squared test

Some key caveats to remember

- No statistical test is perfect, we expect false-positive results some fraction of the time
- A 5% significance threshold is totally arbitrary
- Significance testing does not consider the effect size!
- As a result, significance testing (and *p* values) are somewhat controversial

Defining one more term: SEM

- Standard error of the mean, or SEM
- $s = \frac{\sigma}{\sqrt{N}}$, where σ is the standard deviation and N is the sample size
- **SEM is a weighted standard deviation based on sample size!**
- SEMs are useful for estimating the mean of a distribution, which is critical for performing statistical tests

Outline

- What is a statistical test?
- *p*-values
- Types of statistical tests
- Multiple hypothesis testing

Some common univariate statistical tests in biology

- *t*-tests are useful for normal distributions
  - Unpaired (Student's) *t*-test
  - Welch's *t*-test
  - Paired *t*-test
- Chi-squared test
  - Useful for proportions or frequencies
- Rank-based tests:
  - Useful when the data are not normal!
  - Mann-Whitney U test (non-parametric analogue of unpaired *t*-test)
  - Wilcoxon signed-rank test (non-parametric analogue of paired *t*-test)

Sources: https://www.biostathandbook.com/testchoice.html, https://en.wikipedia.org/wiki/Ranking_(statistics)

The unpaired (Student's) *t*-test compares the means of two datasets

- Assumptions:
  - Data (x_1, x_2) are normally distributed
  - x_1, x_2 have equal variances
  - However, it's not too sensitive to these assumptions if the distributions are the same, and the sample sizes are similar between groups

- Null hypothesis: **Equal means**
- Alternative hypothesis: **Different means**

- Formula for equal n: $t = \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{SEM}(x_1, x_2)}$
- Denominator is basically a pooled SEM

Source: https://en.wikipedia.org/wiki/Student%27s_t-test

Welch's *t*-test does not require equal variances

- Assumptions:
  - Data (x_1, x_2) are normally distributed
  - ~~x_1, x_2 have equal variances~~

- Null hypothesis: **Equal means**
- Alternative hypothesis: **Different means**

- Formula: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2}}$, where s denotes the SEM of x_1 and x_2
- **Denominator is not based on pooled data!**

Paired *t*-tests are great for before and after measurements

- Assumptions:
  - Data (x_1, x_2) **are paired** (therefore have same sample size)
  - Differences $d = x_1 - x_2$ are normally distributed; check with a histogram

- Null hypothesis: **Mean difference is zero**
- Alternative hypothesis: **Mean difference is not zero**

- Formula: $t = \frac{\bar{d}}{s_d}$
- Increased power relative to a standard *t*-test because each sample now has 2 measurements; cost is having to do 2x the measurements

What if my data are not normal!?!?!?!?

- Other approaches:
  - Bootstrapping
  - Rank-based statistical tests

- Bootstrapping resamples your data (with replacement) many times and computes e.g. the mean from the resampled data!
- Then, make a histogram

- Still requires **independence** and a **large sample size**
- Trade-offs:
  - It's computationally intensive, compared to other methods
  - Bootstrapping is largely empirical so some statisticians don't like it

Rank-based tests

- Convert each number in your dataset to a rank (e.g. 1, 2, 3, 4…)
- Perform statistics on the ranks rather than the actual data(!)

- **This is basically a test for the median rather than the mean!!!**

- However, you sacrifice some power by converting the data to ranks

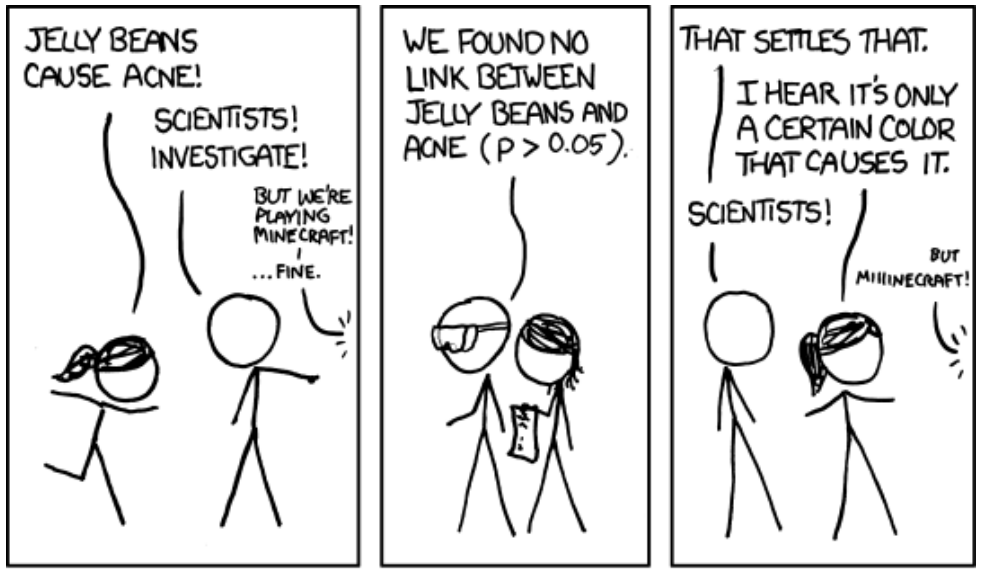
- Null hypothesis: distributions are identical
- Alternative hypothesis: distributions are not identical

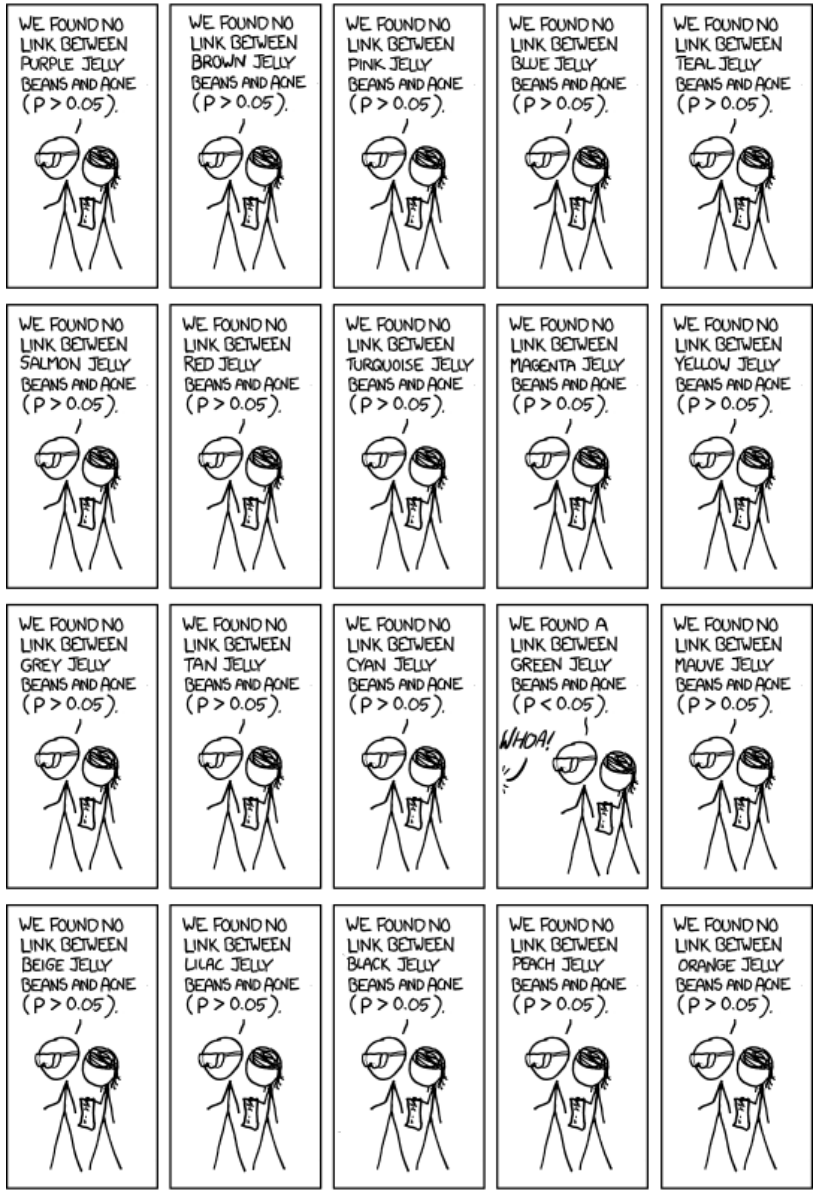
- Still requires independence

Outline

- What is a statistical test?
- *p*-values
- Types of statistical tests
- Multiple hypothesis testing

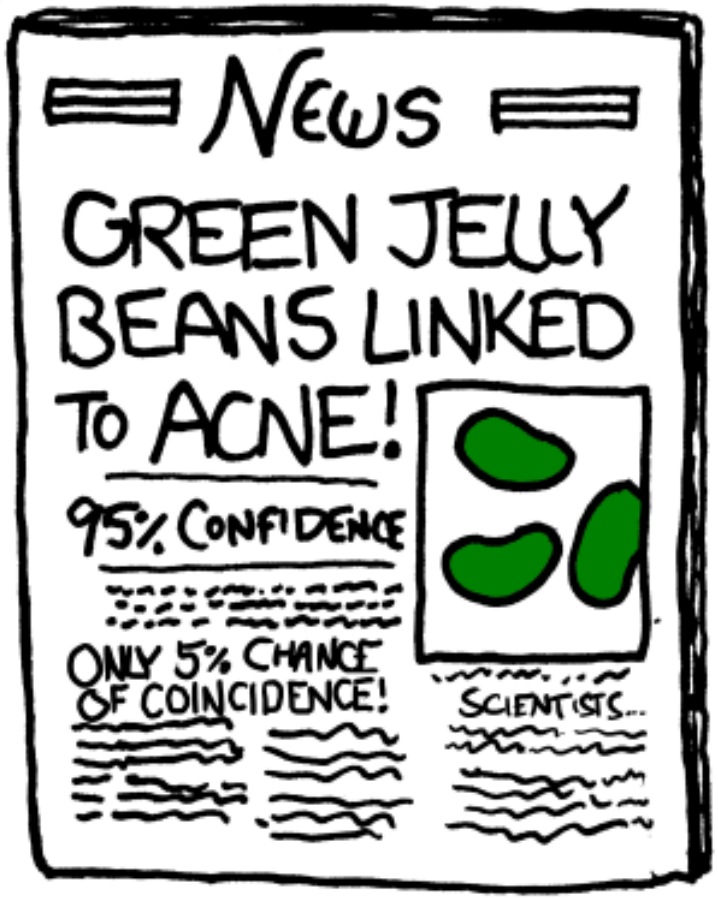
A poignant example





Source: https://xkcd.com/882/.html

A poignant example



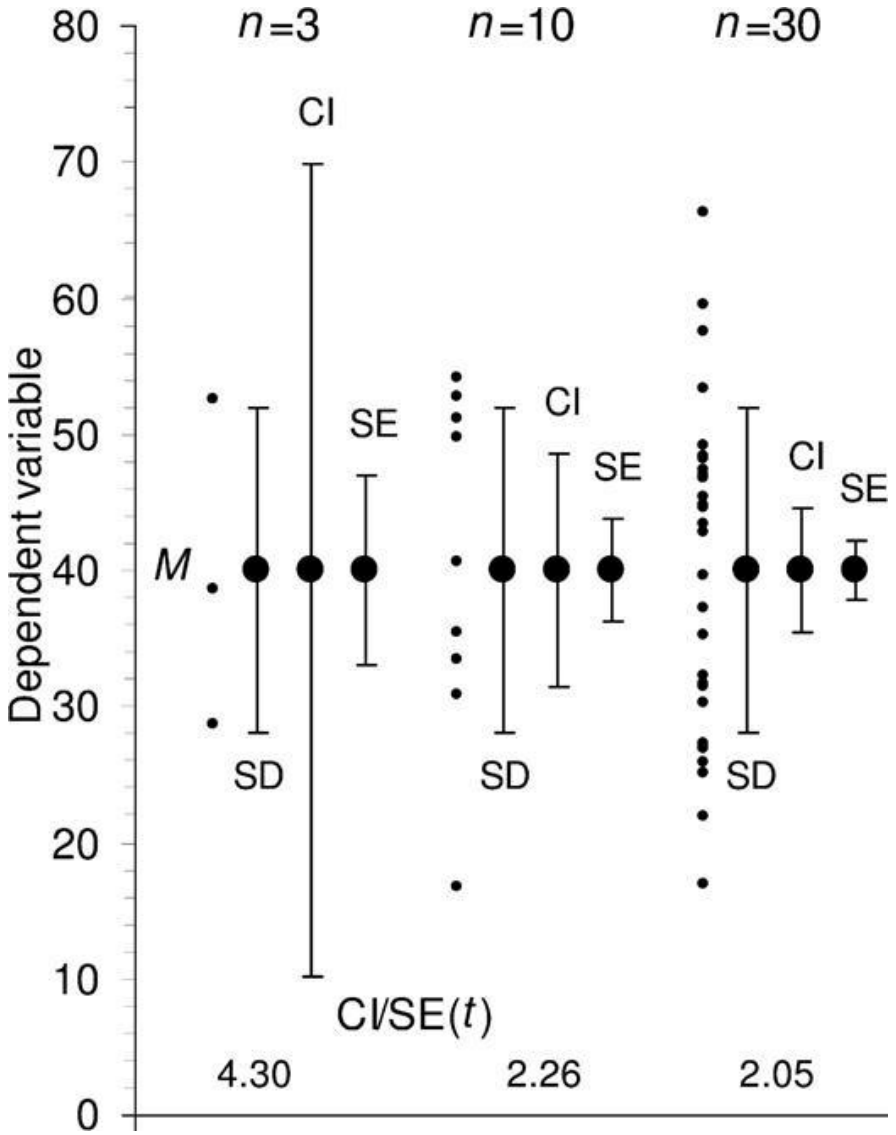
Source: https://xkcd.com/882/.html

There are several ways to correct for this

- The Bonferroni correction:
  - When performing m tests, divide α by m
  - e.g. for ten tests, use $p < 0.005$
- Instead of adjusting α, you can also adjust p (multiply it by m)
- Other corrections let you specify a false discovery rate independent of α:
  - Benjamini-Hochberg procedure
  - Many other methods
- This will be discussed further in the genomics unit

Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_distribution_pdf_-_public_domain.svg

Reading for next time: Error Bars



https://rupress.org/jcb/article/177/1/7/34602/Error-bars-in-experimental-biology

Extra Slides