Two-sample inference: Continuous data

Patrick Breheny

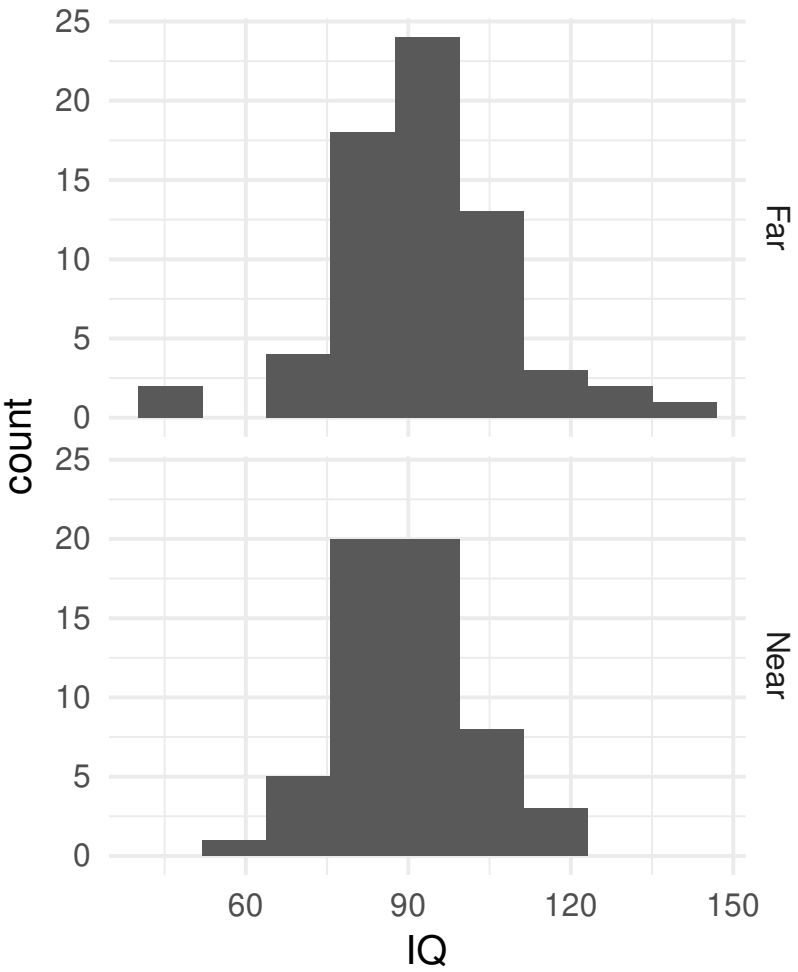
Introduction

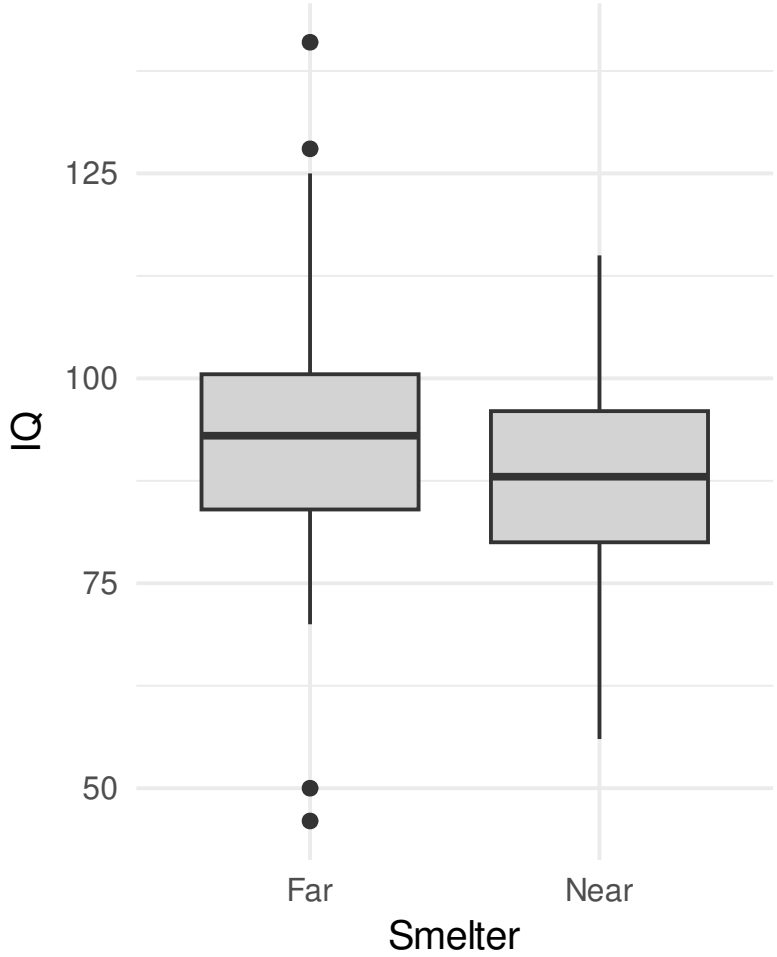
- Our next several lectures will deal with two-sample inference for continuous data
- As we will see, some procedures work very well when the continuous variable follows a roughly normal distribution, but work poorly when it doesn't
- We will begin with the procedures that work well for data that (at least approximately) follows a normal distribution

Example: lead exposure and IQ

- Our motivating example for today deals with a study that we've looked at a few times concerning lead exposure and neurological development for a group of children in El Paso, Texas
- The study compared the IQ levels of 57 children who lived within 1 mile of a lead smelter and a control group of 67 children who lived at least 1 mile away from the smelter
- In the study, the average IQ of the children who lived near the smelter was 89.2, while the average IQ of the control group was 92.7

Looking at the data





Could the results have been due to chance?

- Looking at the raw data, it appears that living close to the smelter may hamper neurological development
- However, as always, there is sampling variability present — just because, in this sample, the children who lived closer to the smelter had lower IQs does not necessarily imply that the population of children who live near smelters have lower IQs
- We need to ask whether or not our finding could have been due to chance, and what other explanations are consistent with the data

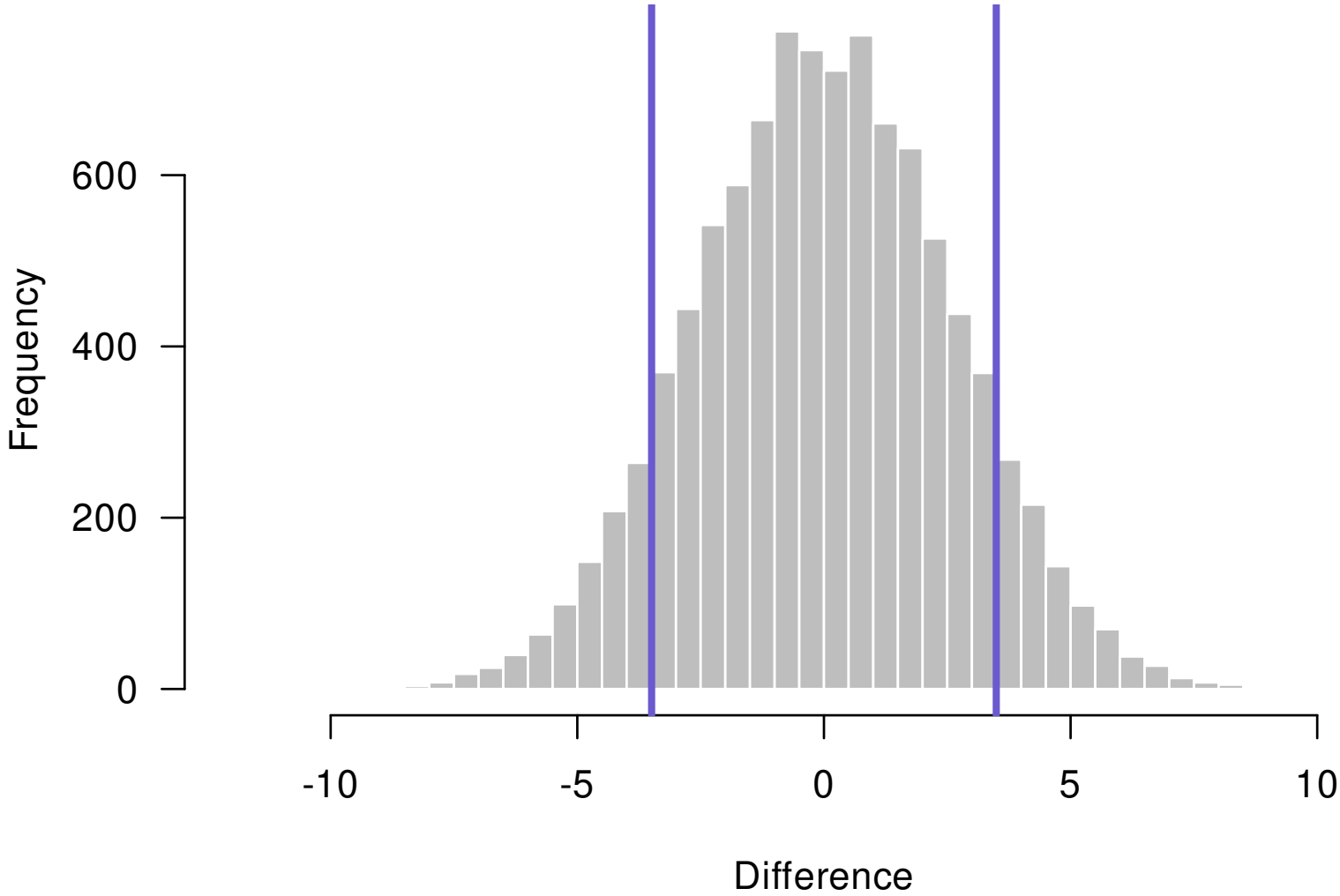
The difference between two means

- We can calculate separate confidence intervals for the two groups using one-sample t-distribution methods
- However, as we have said several times, the better way to analyze the data is to use all of the data to answer the single question: is there a difference between the two groups?
- Denoting the two groups with the subscripts 1 and 2, we will attempt to test the hypothesis that $\mu_1 = \mu_2$ by looking at the random variable $\hat{\mu}_1 - \hat{\mu}_2$ (i.e., the difference in sample means) and determining whether it is far away from 0 with respect to variability (standard error)
- First, we will go over an exact, albeit computer-intensive, approach (the "permutation test"), then an approximate approach that can be done by hand (the "two-sample t-test")

Viewing our study as balls in an urn

- The same concept that we encountered in the two-sample Fisher's exact test can be used for continuous data also
- If the smelter had no effect on children's neurological development, then it wouldn't matter if the child lived near it
- Under this null hypothesis, then, it would be like writing down each child's IQ on a ball, putting all the balls into an urn, then randomly picking out 57 balls and calling them the "near" group, and calling the other 67 balls the "far" group
- If we were to do this, how often would we get a difference as large or larger than 3.5, the actual difference observed?

Results of the experiment

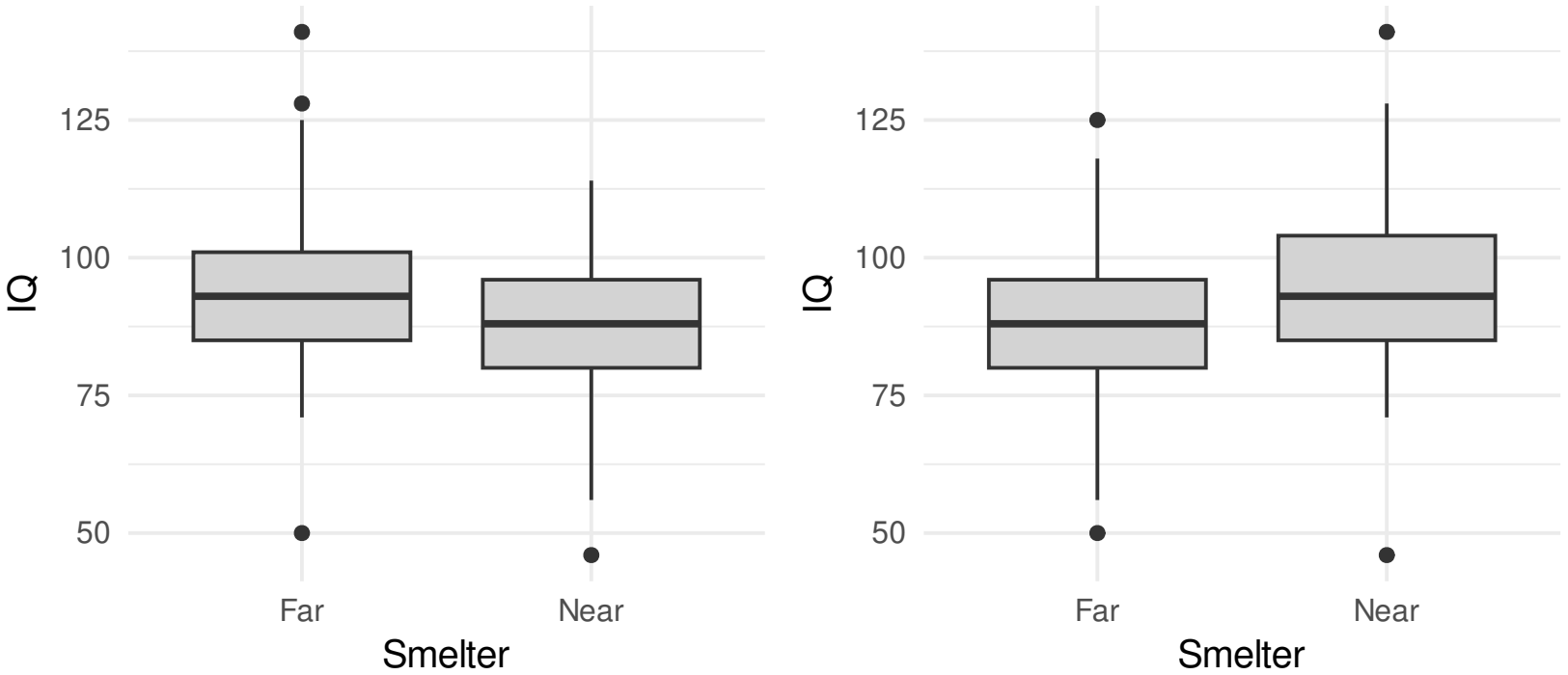


Results of the experiment (cont'd)

- In the experiment, I obtained a random difference in means larger than the observed one 17676 times out of 100,000
- My "experimental" p-value is $17676/10^5 = 0.177$
- Thus, our suggestive-looking box plots could easily have come about entirely due to chance

Experiment examples

Examples of recreations of the experiment under the null hypothesis in which a difference as large or larger was obtained:



Permutation tests

- This approach to carrying out a hypothesis test is called a *permutation test*
- The different orders that a sequence of numbers can be arranged in are called permutations; a permutation test is essentially calculating the percent of random permutations under the null hypothesis that produce a result as extreme or more extreme than the observed value
- Unlike Fisher's exact test, exact solutions are very time-consuming to calculate (this problem is much harder than Fisher's exact test)
- Unless the number of observations is small enough that we can count the permutations by hand, we need a computer to perform a permutation test

Approximate results based on the normal distribution

- As you might guess from the look of the histogram, a much easier way to obtain an answer is to use the normal distribution as an approximation
- Letting $\hat{\delta} = \bar{x}_1 - \bar{x}_2$, consider the test statistic:

$$\frac{\hat{\delta} - \delta_0}{\mathrm{SE}_d} = \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{SE}_d}$$

- But what's the standard error of the difference between two means, SE_d?

The standard error of the difference between two means

- Suppose we have $\bar{x}_1$ with standard error SE_1 and $\bar{x}_2$ with standard error SE_2 (and that $\bar{x}_1$ and $\bar{x}_2$ are independent)
- Then the standard error of $\bar{x}_1 - \bar{x}_2$ is

$$\mathrm{SE}_d = \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$$

- Note the connections with both the root-mean-square idea and the square root law from earlier in the course
- Note also that $\sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2} < \mathrm{SE}_1 + \mathrm{SE}_2$ — a two-sample test is a more powerful way to look at differences than two separate analyses

The split

- This equation would be perfect if we knew SE_1 and SE_2
- But of course we don't — all we have are estimates
- There are two ways of settling this question, and they have led to two different forms of the two-sample t-test

Approach #1: Student's t-test

- The first approach was invented by W.S. Gosset (Student)
- His approach was to assume that the standard deviations of the two groups were the same
- If you do this, then you only have one extra source of uncertainty to worry about: the uncertainty in your estimate of the common standard deviation

Approach #2: Welch's t-test

- The second approach was invented by B.L. Welch
- He generalized the two-sample t-test to situations in which the standard deviations were different between the two groups
- If you don't make Student's assumption, then you have two extra sources of uncertainty to worry about: uncertainty about SD_1 and uncertainty about SD_2

Student's test vs. Welch's test

- We'll talk more about the difference between the two tests at the end of class
- In most situations, the two tests provide similar answers
- However, Student's test is much easier to do by hand

The pooled standard deviation

- In order to estimate a standard error, we will need an estimate of the common standard deviation
- This is obtained by *pooling* the deviations
- To calculate a standard deviation, we took the root-mean-square of the deviations (only with $n - 1$ in the denominator)
- To calculate a pooled standard deviation, we take the root-mean-square of all the deviations from both groups (only with $n_1 + n_2 - 2$ in the denominator)
- Essentially, the pooled standard deviation is a weighted average of the standard deviations in the two groups, where the weights are the number of observations in each group

The pooled standard deviation and the standard error

- Letting SD_p denote our pooled standard deviation, our estimated standard error is

$$\begin{aligned}\mathrm{SE}_d &= \sqrt{\frac{\mathrm{SD}_p^2}{n_1} + \frac{\mathrm{SD}_p^2}{n_2}} \\ &= \mathrm{SD}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\end{aligned}$$

- This equation is similar to our earlier one, only now the amount by which the SE is reduced in comparison to the SD depends on the sample size in each group

Sample size and standard error

- So, let's say that we have 50 subjects in one group and 10 subjects in the other group, and we have enough money to enroll 20 more people in the study
- To reduce the SE as much as possible, should we assign them to the group that already has 50, or the group that only has 10?
- Let's check:

$$\sqrt{\frac{1}{70} + \frac{1}{10}} = 0.34 \qquad \sqrt{\frac{1}{50} + \frac{1}{30}} = 0.23$$

The advantages of balanced sample sizes

- This example illustrates an important general point: the greatest improvement in accuracy/reduction in standard error comes when the sample sizes of the two groups are balanced
- Occasionally, it is much easier (or cheaper) to obtain (or assign) subjects in one group than in the other
- In these cases, one often sees unbalanced sample sizes
- However, it is rare to see a ratio that exceeds 3:1, as the study runs into diminishing returns — no matter how much you reduce the standard error of $\bar{x}_1$, the standard error of $\bar{x}_2$ will still be there

The degrees of freedom of the two-sample t-test

- We have said that if we make the assumption of equal standard deviations, then we only need to worry about the variability of $\bar{x}_1 - \bar{x}_2$ and the variability in your estimate of the common standard deviation
- How variable is our estimate of the common standard deviation?
- Well, it's now based on $n_1 + n_2 - 2$ degrees of freedom (we lose one degree of freedom for each mean that we calculate)
- Thus, to perform inference, we will look up results on Student's curve with $n_1 + n_2 - 2$ degrees of freedom

Student's t-test: procedure

The procedure of Student's two-sample t-test should look quite familiar:

(1) Estimate the standard error: $\mathrm{SE}_d = \mathrm{SD}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

(2) Calculate the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{SE}_d}$$

(3) Calculate the area under the Student's curve with $n_1 + n_2 - 2$ degrees of freedom curve outside $\pm t$

Student's t-test: example

- For the lead study,
  - The mean IQ for the children who lived near the smelter was 89.2
  - The mean IQ for the children who did not live near the smelter was 92.7
  - The pooled standard deviation was 14.4
- The individual standard deviations were 12.2 and 16.0 — note that the pooled standard deviation is close to the average of the individual standard deviations, but slightly closer to the SD for the "far" group since that group had the larger sample size

Student's t-test: example

(1) Estimate the standard error: $\text{SE}_d = 14.4\sqrt{\frac{1}{57} + \frac{1}{67}} = 2.59$

(2) Calculate the test statistic:

$$t = \frac{92.7 - 89.2}{2.59} = 1.35$$

(3) For Student's curve with $57 + 67 - 2 = 122$ degrees of freedom, 17.9% of the area lies outside 1.35

Thus, if there was no difference in IQ between the two groups, we would have observed a difference as large or larger than the one we saw about 18% of the time; the study provides very little evidence of a population difference

Carrying out t-tests in R

As always, it is useful to know how to carry out these tests using statistical software; here's how to carry out t-tests in R:

- Student's t-test:

```
t.test(IQ ~ Smelter, lead_iq, var.equal = TRUE)
#     Two Sample t-test
# t = 1.3505, df = 122, p-value = 0.1793
```

- Welch's t-test:

```
t.test(IQ ~ Smelter, lead_iq)
#     Welch Two Sample t-test
# t = 1.38, df = 120.62, p-value = 0.1702
```

Student's t-test, Welch's t-test, and the permutation test

- Note that Student's t-test agrees quite well with our permutation test from earlier, and with Welch's test
  - Permutation: $p = 0.17676$
  - Student's: $p = 0.179$
  - Welch's: $p = 0.170$
- This is usually the case when the sample sizes are reasonably large: the approximations work well, agreeing both with each other and with exact approaches

Confidence intervals

- Our hypothesis test suggests that there may be no difference in IQ between the two groups
- Is it possible that there is a difference, even a large difference?
- This is the kind of question answered by a confidence interval

Confidence intervals: Procedure

The procedure for calculating confidence intervals is straightforward:

(1) Estimate the standard error: $\mathrm{SE}_d = \mathrm{SD}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

(2) Determine the values $\pm t_{x\%}$ that contain the middle $x\%$ of the Student's curve with $n_1 + n_2 - 2$ degrees of freedom

(3) Calculate the confidence interval:

$$(\bar{x}_1 - \bar{x}_2 - t_{x\%} \cdot \mathrm{SE}_d, \bar{x}_1 - \bar{x}_2 + t_{x\%} \cdot \mathrm{SE}_d)$$

Confidence intervals: Example

For the lead study:

(1) The standard error is $\text{SE}_d = 14.4\sqrt{\frac{1}{57} + \frac{1}{67}} = 2.59$, and the difference between the two means was $92.7 - 89.2 = 3.5$

(2) The values ± 1.98 contain the middle 95% of the Student's curve with $57 + 67 - 2 = 122$ degrees of freedom

(3) Thus, the 95% confidence interval is:

$$(3.5 - 1.98 \times 2.59, 3.5 + 1.98 \times 2.59) = (-1.63, 8.61)$$

So, although we cannot rule out the idea that lead has no effect on IQ, it's possible that lead reduces IQ by as much as 8 and a half points (or increases it by a point and a half)

Should I use a permutation test or a t-test?

- The difference between a permutation test or a t-test is whether one trusts the normal approximation to the sampling distribution
- Thus, if n is large, it doesn't matter; the sampling distribution will always be approximately normal (because of the Central Limit Theorem)
- If n is small, then this is a tough choice — the difference between the two tests can be sizable, and there is little data with which to determine normality
- In practice, however, permutation tests are not very common
- The reason is that, if one is concerned about normality, it is usually better to choose one of the methods that we will talk about in the next lecture

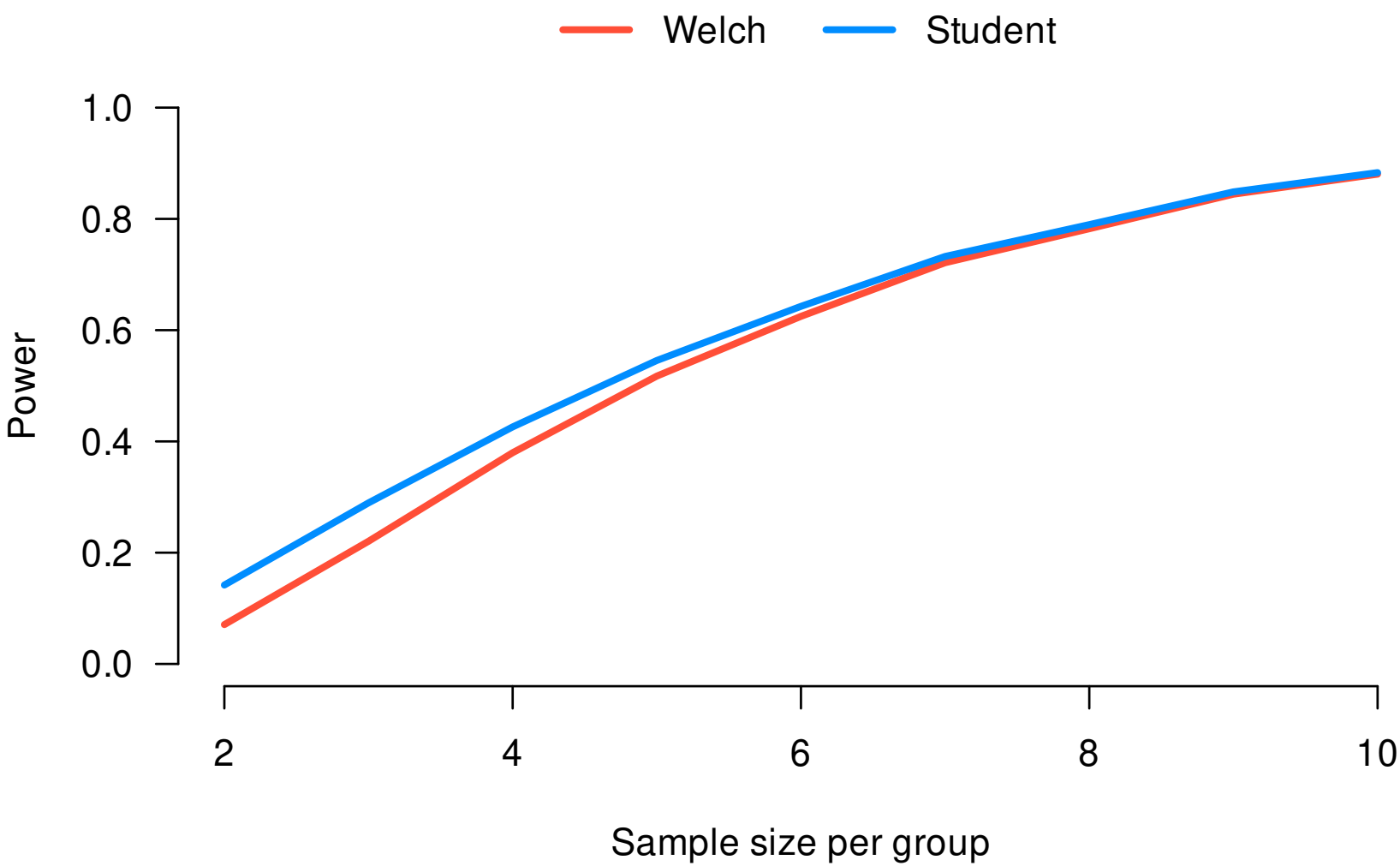
Should I use Student's test or Welch's test?

- In the earlier example, Student's test and Welch's test were basically the same, even though the standard deviations of the two groups differed by about 30%
- This is often the case
- However, the two tests will provide different answers when:
  - The standard deviations of the two groups are very different, **and**
  - The number of subjects in each group is also very different

Tradeoff

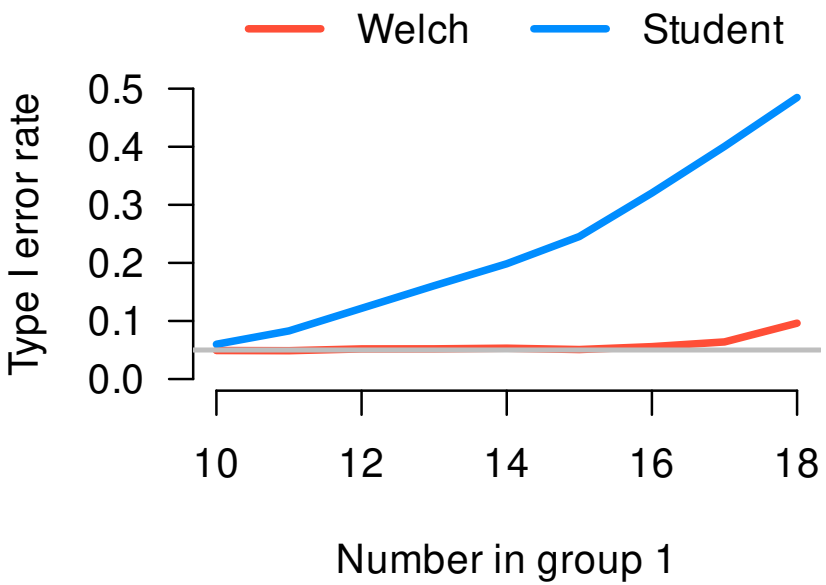
- So which test is better?
- This is a complicated question, but the basic tradeoff is this:
  - When the standard deviations are fairly close, Student's t-test is (slightly) more powerful for small sample sizes
  - When they are not, Student's t-test is unreliable — it is making an assumption of common variability and that assumption is inaccurate
- In practice, the decision depends more on sample size than standard deviation: if you've got enough data, there is no need to make unnecessary assumptions

Power in equal SD setting

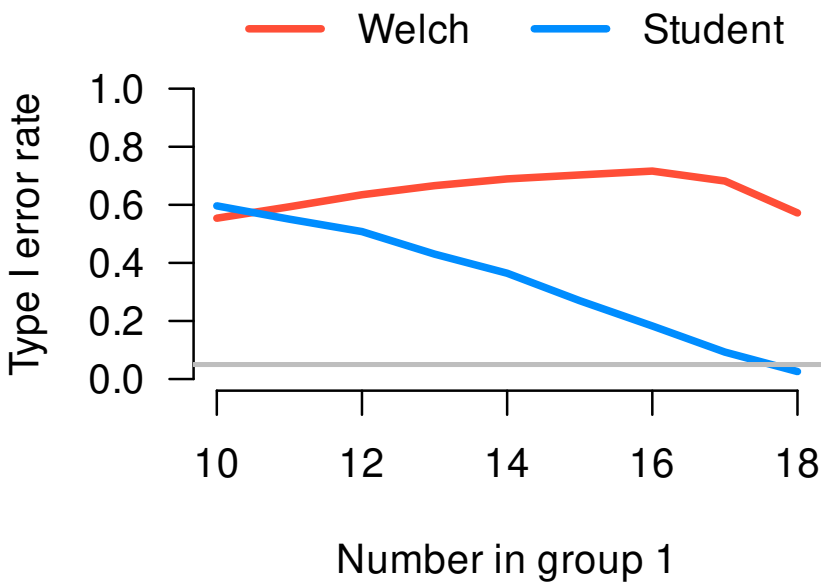


Unequal SD setting (4:1 ratio, total $n = 20$)

More samples from the low-variance group means inflated type I error:



More samples from the high-variance group means low power:



Summary

- We discussed three tests that work well when the data is normally distributed:
  - Permutation tests
  - Student's t-test
  - Welch's t-test
- Student's test assumes equal standard deviations; Welch's test does not
- Know how to carry out hypothesis tests and construct confidence intervals based on Student's approach