Two-sample inference: Continuous data

Patrick Breheny

October 20

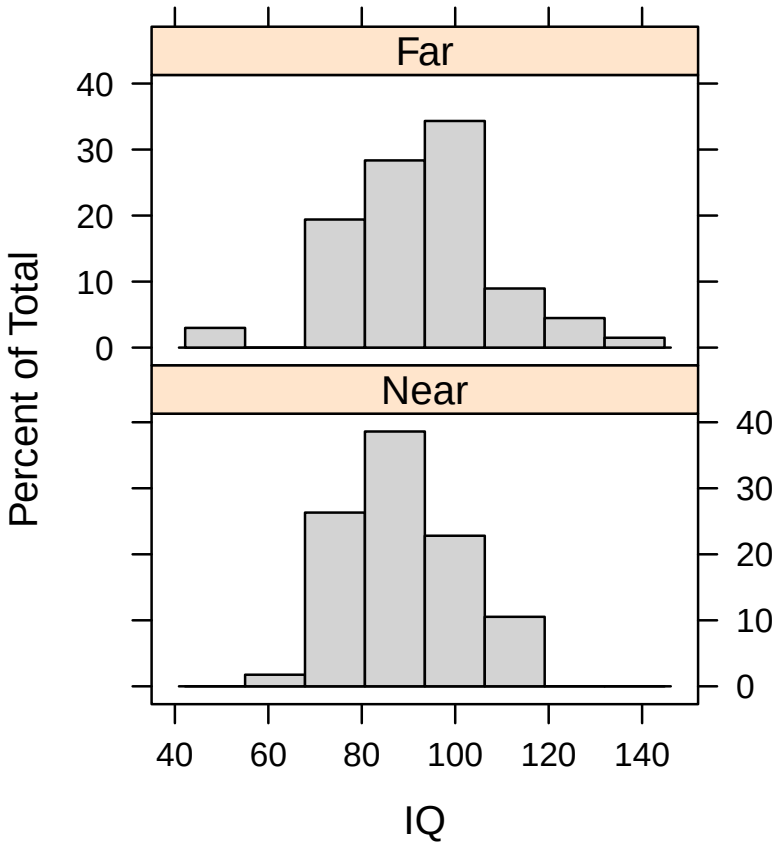
Introduction

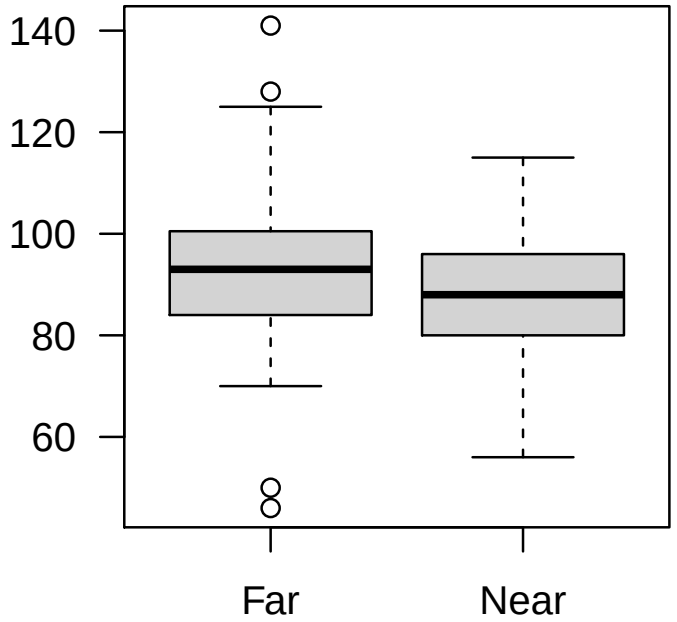
- We now turn our attention to two-sample studies, starting with continuous data
- As you might expect, some procedures work very well when the continuous variable follows a normal distribution, but work poorly when it doesn't
- This week we will discuss procedures that work well for data that (at least approximately) follows a normal distribution

Example: lead exposure and IQ

- Our motivating example for today deals with a study 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, 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 $\bar{x}_1 - \bar{x}_2$ and determining whether it is far away from 0 with respect to variability (standard error)

Sum and difference of two normal random variables

- To consider the difference of two means, we will need the following result concerning the sum and difference of two normally distributed random variables:
- **Theorem:** Suppose $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ are independent. Then $X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.
- **Corollary:** Suppose $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ are independent. Then $X - Y \sim \mathrm{N}(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Distribution of the difference between two means

- Now, we already know that if sample 1 is drawn from a $\mathrm{N}(\mu_1, \sigma_1^2)$ distribution and sample 2 is drawn from a $\mathrm{N}(\mu_2, \sigma_2^2)$ distribution,

$$\bar{x}_1 \sim \mathrm{N}(\mu_1, \sigma_1^2/n_1)$$
$$\bar{x}_2 \sim \mathrm{N}(\mu_2, \sigma_2^2/n_2)$$

  where n_1 and n_2 denote the sample size for each group
- Thus,

$$\bar{x}_1 - \bar{x}_2 \sim \mathrm{N}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

- By the central limit theorem, this is also the approximate distribution of $\bar{x}_1 - \bar{x}_2$, with accuracy depending on n_1, n_2, and the shape of the underlying distribution

The standard error of the difference between two means

- In other words, 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

- So,

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\mathrm{SE}_d} \sim \mathrm{N}(0, 1)$$

- We would be done at this point if we knew σ_1 and σ_2
- But of course we don't – all we have are estimates – and as we have seen, ignoring the uncertainty in these estimates causes problems
- There are two ways of addressing this issue, 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 σ

Approach #2: Welch's t-test

- The second approach was developed by B.L. Welch (with related independent contributions from Franklin Satterthwaite)
- 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 σ_1 and uncertainty about σ_2

Combining the information

- Let's start with Gosset's approach in which we assume that $\sigma_1 = \sigma_2 = \sigma$ (and again, that the data are coming from normal distributions)
- In this case,

$$(n_1 - 1)S_1^2/\sigma^2 \sim \chi^2_{n_1-1}$$
$$(n_2 - 1)S_2^2/\sigma^2 \sim \chi^2_{n_2-1}$$

- Since both of these quantities contain information about σ, it makes sense to combine, or pool, them:

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

The pooled variance

- Or, to rewrite the previous equation,

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2},$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is the *pooled variance*

- Note that the pooled variance is a weighted average of the two sample variances, with weights $n_1 - 1$ and $n_2 - 1$

The pooled standard deviation and the standard error

- We can therefore rewrite the standard error of $\bar{x}_1 - \bar{x}_2$ in terms of S_p, the pooled standard deviation:

$$\begin{aligned} \text{SE}_d &= \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \\ &= S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned}$$

- This equation is similar to our earlier square root law, 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

Student's t-test

- Returning to the problem of inference concerning $\mu_1 - \mu_2$, we have:

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\text{SE}_d} \sim t_{n_1+n_2-2}$$

- Thus, we have a pivotal quantity and it is straightforward to develop hypothesis tests and confidence intervals based on the above relationship
- This test is usually referred to as "Student's t-test" or the "equal variance t-test"; simply referring to it as the "two-sample t-test" is a little vague, as that label could equally apply to the Welch test

Lead study

- For the lead study,
  - The mean IQ for the 57 children who lived near the smelter was 89.2 (SD=12.2)
  - The mean IQ for the 67 children who did not live near the smelter was 92.7 (SD=16.0)
- Thus, the pooled standard deviation was 14.4 – close to the simple average of the individual standard deviations, but somewhat 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: $SE_d = 14.4\sqrt{\frac{1}{57} + \frac{1}{67}} = 2.59$

#2 Calculate the test statistic:

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

#3 For the t distribution 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 systematic difference

Confidence interval: example

We proceed similarly to obtain confidence intervals:

#1 As before, 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:

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

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)

The Behrens-Fisher problem

- What about the unequal variance case? What's the distribution of

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}?$$

- This problem has been considered by many brilliant statisticians, but unfortunately, we do not have an exact solution to this problem
- The difficulty arises in the denominator; we have seen that $\chi^2_{\nu_1} + \chi^2_{\nu_2} \sim \chi^2_{\nu_1+\nu_2}$, but in the above expression, we have a linear combination $a\chi^2_{\nu_1} + b\chi^2_{\nu_2}$, which does not follow a χ^2 distribution (or any other known, tractable distribution)

The approximation

- Several statisticians have proposed solutions to this problem (all of which are approximate), but the most widely used approach is to approximate the distribution of $a\chi^2_{\nu_1} + b\chi^2_{\nu_2}$ by a χ^2_ν distribution, and find the value of ν that best approximates the actual distribution of the linear combination
- This is a bit messy, but one can estimate ν by "moment matching" (i.e., solving for the value of ν that gives the correct mean and variance) to obtain:

$$\hat{\nu} = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{(S_1^2/n_1)^2/(n_1 - 1) + (S_2^2/n_2)^2/(n_2 - 1)}$$

Welch pivot

- The appeal of approximating the distribution of the denominator using a χ^2 distribution is, of course, that we again wind up with a $\mathrm{N}(0,1)$ random variable divided by the square root of a χ^2 random variable divided by its degrees of freedom, so

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\mathrm{SE}_d} \sim t_{\hat{\nu}},$$

  where $\mathrm{SE}_d^2 = S_1^2/n_1 + S_2^2/n_2$
- Tests and intervals then proceed just as in the Student approach, with two differences:
  - Degrees of freedom are different
  - The sample variances have not been pooled in estimating SE_d

Welch's t-test: example

Returning to the lead example (recall that $\bar{x}_1 = 92.7, \text{SD}_1 = 16.0, n_1 = 67$ and $\bar{x}_2 = 89.2, \text{SD}_2 = 12.2, n_2 = 57$):

#1 Estimate the standard error: $SE_d = \sqrt{\frac{16.0^2}{67} + \frac{12.2^2}{57}} = 2.53$

#2 Calculate the test statistic:

$$t = \frac{92.7 - 89.2}{2.53} = 1.38$$

#3 For the t distribution with 120.6 degrees of freedom, 17.0% of the area lies outside ± 1.38

This p-value (0.17) is very similar to the one we obtained using the Student approach (0.18)

Confidence interval: example

Similarly for confidence intervals:

#1 Again, $\text{SE}_d = 2.53$ and $\bar{x}_1 - \bar{x}_2 = 3.5$

#2 The values ± 1.98 contain the middle 95% of the t distribution with 120.6 degrees of freedom

#3 Thus, the 95% confidence interval is:

$$(3.5 - 1.98(2.53), 3.5 + 1.98(2.53)) = (-1.52, 8.51)$$

Which is again very similar to the Student interval

Student's test or Welch's test?

- In our 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, especially when both groups have reasonably large sample sizes
- However, in this week's lab will explore how robust Student's test is to the equal variance assumption and in which situations it fails to produce proper tests and intervals

Student's test or Welch's test? (cont'd)

- So which test should one use?
- Different textbooks will recommend different approaches, but my general recommendation would be:
  - If n_1 and n_2 are reasonably large (say, > 10), use Welch's approach
  - Otherwise, use Student's approach (unless for some reason you have reason to believe that the variances are wildly different between the two groups)
- The reason behind this recommendation is that the only cost associated with the Welch approach is essentially to spend a degree of freedom to estimate an extra standard deviation – unless your sample size is very small, you can easily afford this

Summary

- Today we considered the quantity

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\mathrm{SE}_d};$$

  know how SE_d is estimated, what is being assumed/approximated, and what distribution the above quantity follows under the Student and Welch approaches (although for the sake of exams, I would provide the Welch/Satterthwaite formula for $\hat{\nu}$)
- Given sample means, standard deviations, and sample sizes, know how to carry out hypothesis tests and construct confidence intervals based on the Student and Welch approaches