Transformations and outliers

Patrick Breheny

April 23

Problems with t-tests

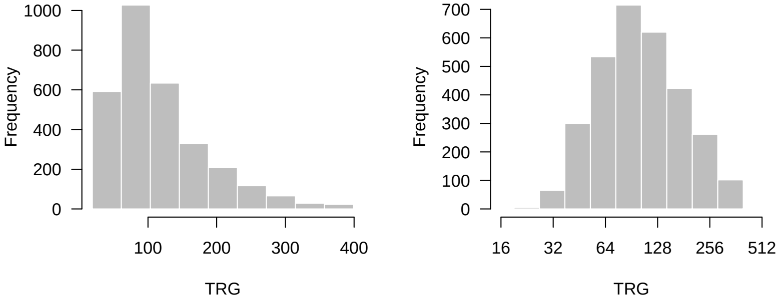
- In the last lecture, we covered the standard way of analyzing whether or not a continuous outcome is different between two groups: the t-test
- However, the focus of the t-test is entirely upon the mean
- As you may recall from our lecture on descriptive statistics towards the beginning of the course, the mean is very sensitive to outliers, and strongly affected by skewed data
- In cases where the mean is an unreliable measure of central tendency, the t-test will be an unreliable test of differences in central tendencies

Transforming the data

- When it comes to skewed distributions, the most common response is to transform the data
- Generally, the most common type of skewness is right-skewness
- Consequently, the most common type of transformation is the log transform
- We have already seen one example of a log transform, when we found a confidence interval for the log odds ratio instead of the odds ratio

Example: Triglyceride levels

As an example of the log transform, consider the levels of triglycerides in the blood of individuals, as measured in the NHANES study:



Low-carb diet study

- Putting this observation into practice, let's consider a 2003 study published in the *New England Journal of Medicine* of whether low-carbohydrate diets are effective
- The investigators studied obese individuals for six months, randomly assigning one group to a low-fat diet and another group to a low-carb diet
- One of the outcomes of interest was the reduction in triglyceride levels over the course of the study

Analysis of untransformed data

- The group on the low-fat diet reduced their triglyceride levels by an average of 7 mg/dl, compared with 38 for the low-carb group
- The pooled standard deviation was 70 mg/dl, and the sample sizes were 36 and 43, respectively
- Thus, $\mathrm{SE} = 70\sqrt{1/43 + 1/36} = 15.8$
- The difference between the means is therefore $31/15.8 = 1.96$ standard errors away from the expected value under the null
- This produces the moderately significant p-value $(p = .06)$

Analysis of transformed data

- On the other hand, let's analyze the log-transformed data
- Looking at log-triglyceride levels, the group on the low-carb diet saw an average reduction of 0.30, compared with 0.03 for the low-fat group
- The pooled standard deviation of the difference in log-triglyceride levels was 0.39
- Thus, $\mathrm{SE} = 0.39\sqrt{1/43 + 1/36} = 0.088$
- The difference between the means is therefore $0.30/0.088 = 3.41$ standard errors away from the expected value under the null
- This produces a much more powerful analysis: $p = .001$

Confidence intervals

- It's also worth discussing the implications of transformations on confidence intervals
- In the low-carbohydrate diet group, the mean reduction on the log scale was 0.30 with a standard deviation of 0.36
- This results in $\mathrm{SE} = 0.36/\sqrt{43} = 0.055$ (note that we're working with one sample here) and a 95% confidence interval of $(0.19, 0.41)$, but what does this mean in terms of the original units: triglyceride levels?

Confidence intervals (cont'd)

- As we saw with odds ratios, differences on the log scale are ratios on the original scale; thus, when we invert the transformation (by exponentiating), we will obtain a confidence interval for the ratio between the two means
- Thus, in the low-carb diet study, we saw a difference of 0.30 on the log scale; this corresponds to $e^{0.30} = 1.35$ on the original scale – in other words, subjects on the low-carb diet reduced their triglycerides by 35%
- Similarly, to calculate a confidence interval for that percent reduction, we exponentiate the two endpoints:

$$(e^{0.19}, e^{0.41}) = (1.21, 1.51),$$

  in other words, the true reduction at the population level would likely be somewhere between 21% and 51%

Technical notes

These bullet points are just for your information and for completeness; for the purposes of this course you don't need to know them:

- The mean of the log-transformed values is not the same as the log of the mean. The (exponentiated) mean of the log-transformed values is known as the *geometric mean*. Technically, what we have constructed is a confidence interval for the ratio of geometric means.
- You can calculate a confidence interval for difference between the two diets as well and exponentiate; in this case that would represent a confidence interval for the ratio of ratios (of geometric means), but this is getting a little esoteric

The big picture

- If the data looks relatively normal after the transformation, we can simply perform a t-test on the transformed observations
- The t-test assumes a normal distribution, so this transformation will generally result in a more powerful, less error-prone test
- Transformations are a sound statistical practice – we're not manipulating data, just measuring it in a different way

Tailgating study

- Let us now turn our attention to a study done at the University of Iowa investigating the tailgating behavior of young adults
- In a driving simulator, subjects were instructed to follow a lead vehicle, which was programmed to vary its speed in an unpredictable fashion
- As the lead vehicle does so, more cautious drivers respond by following at a further distance; riskier drivers respond by tailgating

Goal of the study

- The outcome of interest is the average distance between the driver's car and the lead vehicle over the course of the drive, which we will call the "following distance"
- The study's sample contained 55 drivers who were users of illegal drugs, and 64 drivers who were not
- The average following distance in the drug user group was 38.2 meters, and 43.4 in the non-drug user group, a difference of 5.2 meters
- Is this difference statistically significant?

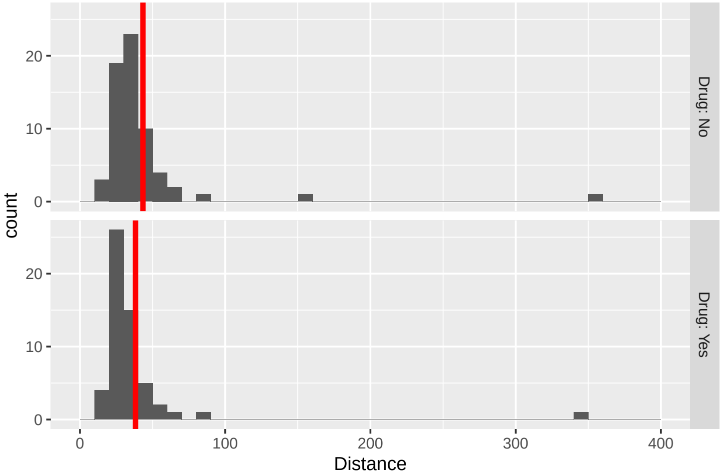
Analysis using a t-test

- No, says the t-test
- The pooled standard deviation is 44, producing a standard error of 8.1
- The difference in means is therefore less than one standard error away from what we would expect under the null
- There is virtually no evidence against the null ($p = .53$)

Always look at your data

- Nothing interesting here; let's move on, right?
- Not so fast!
- Remember, we should always look at our data (this is especially true with continuous data)
- In practice, we should look at it first – before we do any sort of testing – but today, I'm trying to make a point

What the data look like



Outliers

- As we easily see from the graph, huge outliers are present in our data
- And as mentioned earlier, the mean is sensitive to these outliers, and as a result, our t-test is unreliable
- The simplest solution (and unfortunately, probably the most common) is to throw away these observations
- So, let's delete the three individuals with extremely large following distances from our data set and re-perform our t-test (NOTE: I am not in any way recommending this as a way to analyze data; we are doing this simply for the sake of exploration and illustration)

Removing outliers in the tailgating study

- By removing the outliers, the pooled standard deviation drops from 44 to 12
- As a result, our observed difference is now 1.7 standard errors away from its null hypothesis expected value
- The p-value goes from 0.53 to 0.09

Valid reasons for disregarding outliers

- There are certainly valid reasons for throwing away outliers
- For example, a measurement resulting from a computer glitch or human error
- Or, in the tailgating study, if we had reason to believe that the three individuals with the extreme following distances weren't taking the study seriously, including them may be doing more harm than good

Arguments against disregarding outliers

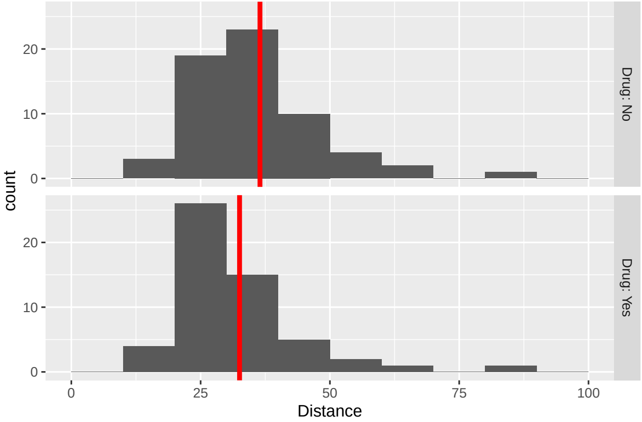
- However, throwing away observations is a questionable practice
- Perhaps computer glitches, human errors, or subjects not taking the study seriously were problems for other observations, too, but they just didn't stand out as much
- Throwing away outliers often produces a distorted view of the world in which nothing unusual ever happens, and overstates the accuracy of a study's findings

Throwing away outliers: a slippery slope

- Furthermore, throwing away outliers threatens scientific integrity and objectivity
- For example, the investigators put a lot of work into that driving study, and they got (after throwing out three outliers) a t-test p-value of 0.09
- Unfortunately, this is not a very convincing p-value
- They could go back, collect more data and refine their study design, but that would be a lot of work
- One can imagine that the temptation would be high to keep throwing away outliers

Throwing away outliers: a slippery slope (cont'd)

Now that we've thrown away the three largest outliers, the next two largest measurements kind of look like outliers:



Throwing away outliers: a slippery slope (cont'd)

- What if we throw these measurements away too?
- Our pooled standard deviation drops now to 10.7
- As a result, our observed difference is now 2.03 standard errors away from 0, resulting in a "significant" p-value of .045

Data snooping

- This manner of picking and choosing which data we are going to allow into our study, and which data we are going to conveniently discard, is highly dubious, and any p-value that is calculated in this manner is questionable (even, perhaps, meaningless)
- This activity is sometimes referred to as "data snooping" or "data dredging"
- Unfortunately, this goes on all the time, and the person reading the finished article has very little idea of what has happened behind the scenes resulting in that "significant" p-value

The ozone layer

- Furthermore, outliers are often the most interesting observations – instead of being thrown away, they deserve the opposite: further investigation
- As a dramatic example, consider the case of the hole in the ozone layer created by the use of chlorofluorocarbons (CFCs) and first noticed in the middle 1980s
- As the story garnered worldwide attention, investigators from around the world started looking into NASA's satellite data on ozone concentration
- These investigators discovered that there was appreciable evidence of an ozone hole by the late 1970s
- However, NASA had been discounting these sudden, large decreases in Antarctic ozone layers as outliers – at what turns out to have been considerable environmental cost

The big picture

- Sometimes, there are good reasons for throwing away misleading, outlying observations
- However, waiting until the final stages of analysis and then throwing away observations to make your results look better is both dishonest and grossly distorts one's research
- It is usually better to keep all subjects in the data set, but analyze the data using a method that is robust to the presence of outliers
- Also, don't forget that outliers can be the most important and interesting observations of all

Summary

- A common way of analyzing data that is not normally distributed is to transform it so that it is
- In particular, it is common to analyze right-skewed data using the log transformation
- Differences on the log scale correspond to ratios on the original scale
- Outliers have a dramatic effect on t-tests – but that doesn't necessarily mean you should remove them