Transformations and rank-based tests

Patrick Breheny

November 8

Problems with t-tests

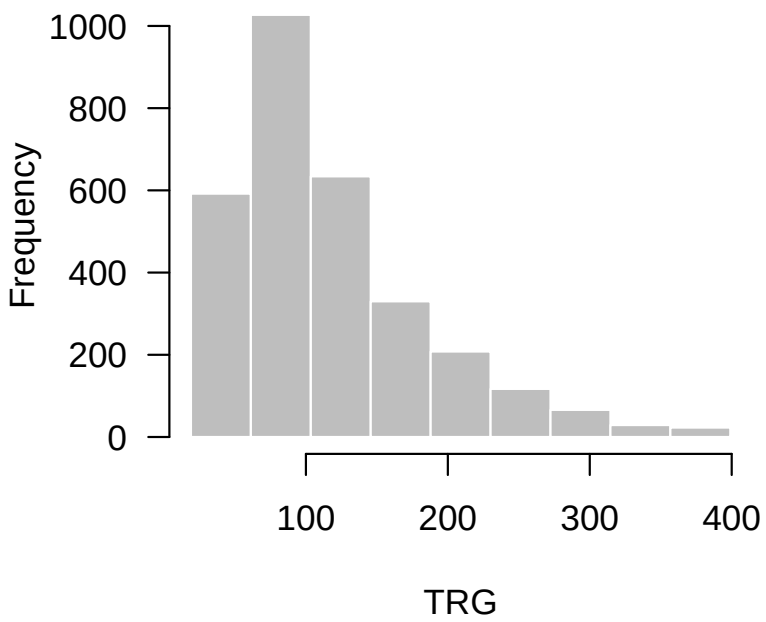
- Our previous discussion of comparing continuous outcomes in two-group studies focused on t-tests
- The derivation of t-tests assumes normality, although as we saw in lab, the approach is fairly robust to departures from normality
- A more fundamental limitation of the t-test is that it focuses entirely on the mean
- When the data is skewed or contains outliers, the mean itself is an unreliable measure of central tendency and the t-test will be an unreliable test of differences between two groups

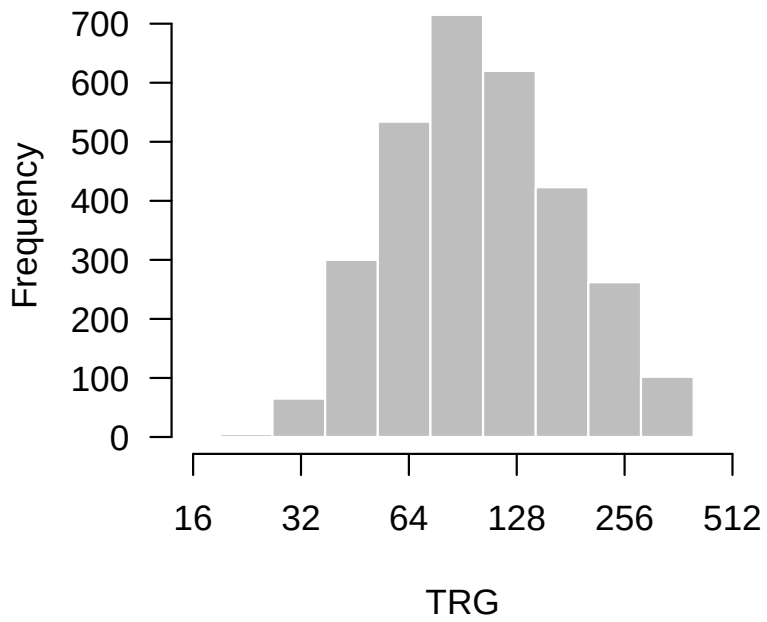
Transforming the data

- 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 at reducing serum triglyceride levels
- The investigators studied overweight 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 66 mg/dl, and the sample sizes were 43 and 36, respectively
- Thus, $\mathrm{SE} = 66\sqrt{1/43 + 1/36} = 15$
- The difference between the means is therefore $31/15 = 2.08$ standard errors away from the expected value under the null
- This produces the moderately significant p-value ($p = .04$)

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-fat diet saw an average reduction of 1.8, compared with 3.5 for the low-carb group
- The pooled standard deviation of the log-triglyceride levels was 2.2
- Thus, $\text{SE} = 2.2\sqrt{1/43 + 1/36} = 0.5$
- The difference between the means is therefore $1.7/0.5 = 3.4$ 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
- The (Student's) confidence interval for the difference in log-triglyceride levels is $3.5 - 1.8 \pm 1.99(0.5) = (0.71, 2.69)$; this is fairly straightforward
- But what does this mean in terms of the original units: triglyceride levels?
- Recall that differences on the log scale are ratios on the original scale; thus, when we invert the transformation (by exponentiating, also known as taking the "antilog"), we will obtain a confidence interval for the ratio between the two means

Confidence intervals (cont'd)

- Thus, in the low-carb diet study, we see a difference of 1.7 on the log scale; this corresponds to a ratio of $e^{1.7} = 5.5$ on the original scale – in other words, subjects on the low-carb diet reduced their triglycerides 5.5 times more than subjects on the low-fat diet
- Similarly, to calculate a confidence interval, we exponentiate the two endpoints (note the similarity to constructing CIs for the odds ratio):

$$(e^{0.71}, e^{2.69}) = (2, 15)$$

Geometric vs. arithmetic means

- Note that in this approach, we are examining differences between the means of the log-transformed values, not the logs of the means
- Thus, we have not actually estimated and constructed an interval for the ratio of means . . . what have we constructed an interval for?
- The exponentiated mean of the log-transformed values is known as the *geometric mean*; thus, what we have actually constructed a confidence interval for is the ratio of the geometric means, as opposed to the usual *arithmetic mean*
- This is desirable, for as we mentioned at the outset, means are not particularly good measures of central tendency when data is skewed; geometric means are much more stable for skewed data

Tailgating study

- Thus, if there exists a transformation that makes our data look normally distributed (a *normalizing transformation*), analysis is straightforward: we just transform the data and we can then use the t-test approaches we've already developed
- But what if no normalizing transformation exists?
- As a concrete example, consider 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

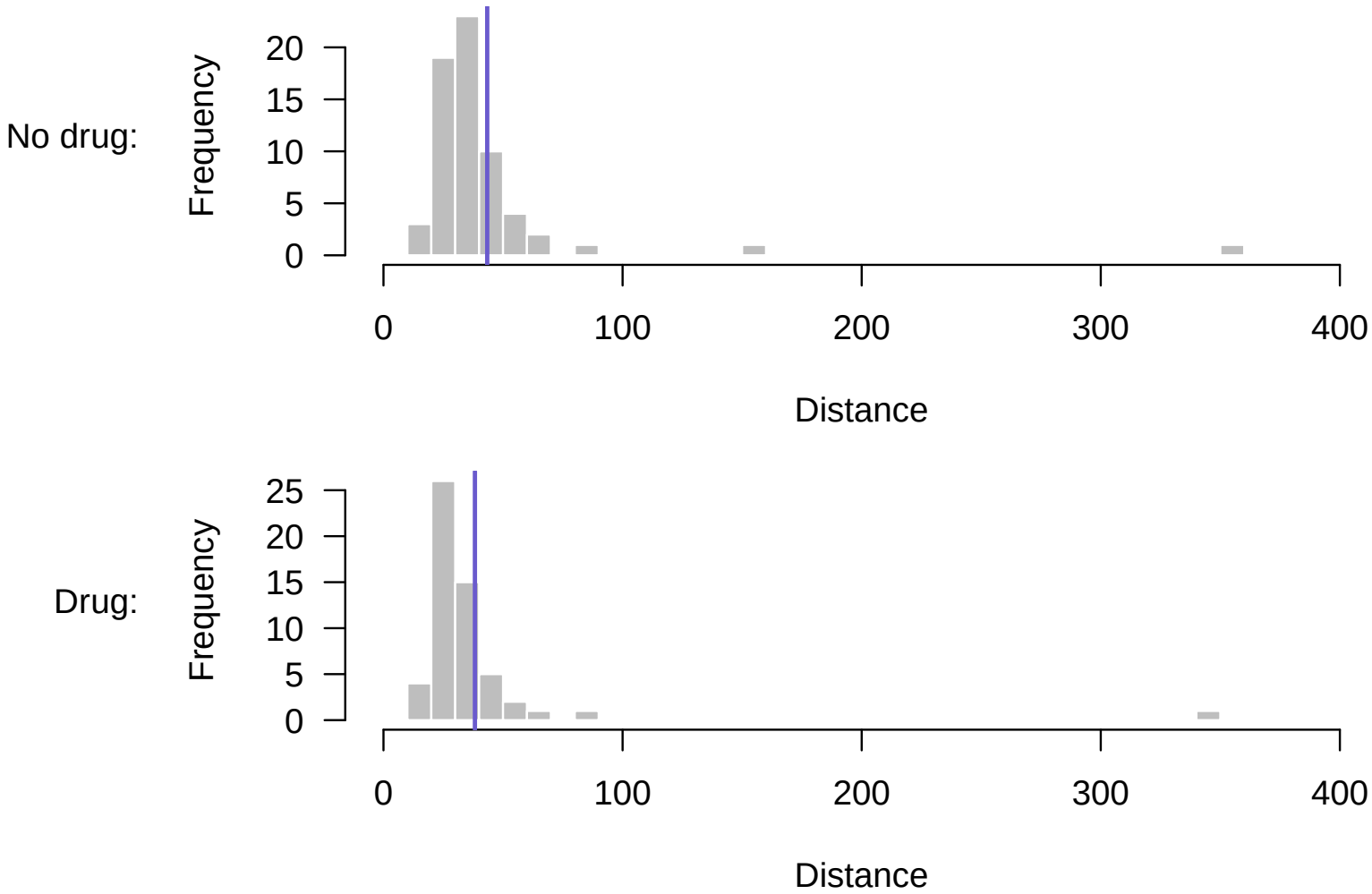
Goal of the study

- 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 ($p = .53$)

What the data look like



Outliers

- As we easily see from the graph, huge outliers are present in our data
- As we know, 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

- Occasionally, there are valid reasons for throwing away outliers
- For example, a measurement resulting from a computer glitch or human error, or if, say, further investigation reveals that experimental protocols were not followed for that subject
- For example, in the tailgating study, the subjects were not told that this was a study of tailgating behavior; if the three individuals with the extreme following distances somehow learned this and didn't drive as they normally would because they knew their tailgating distance was being measured, including them may do more harm than good

Arguments against disregarding outliers

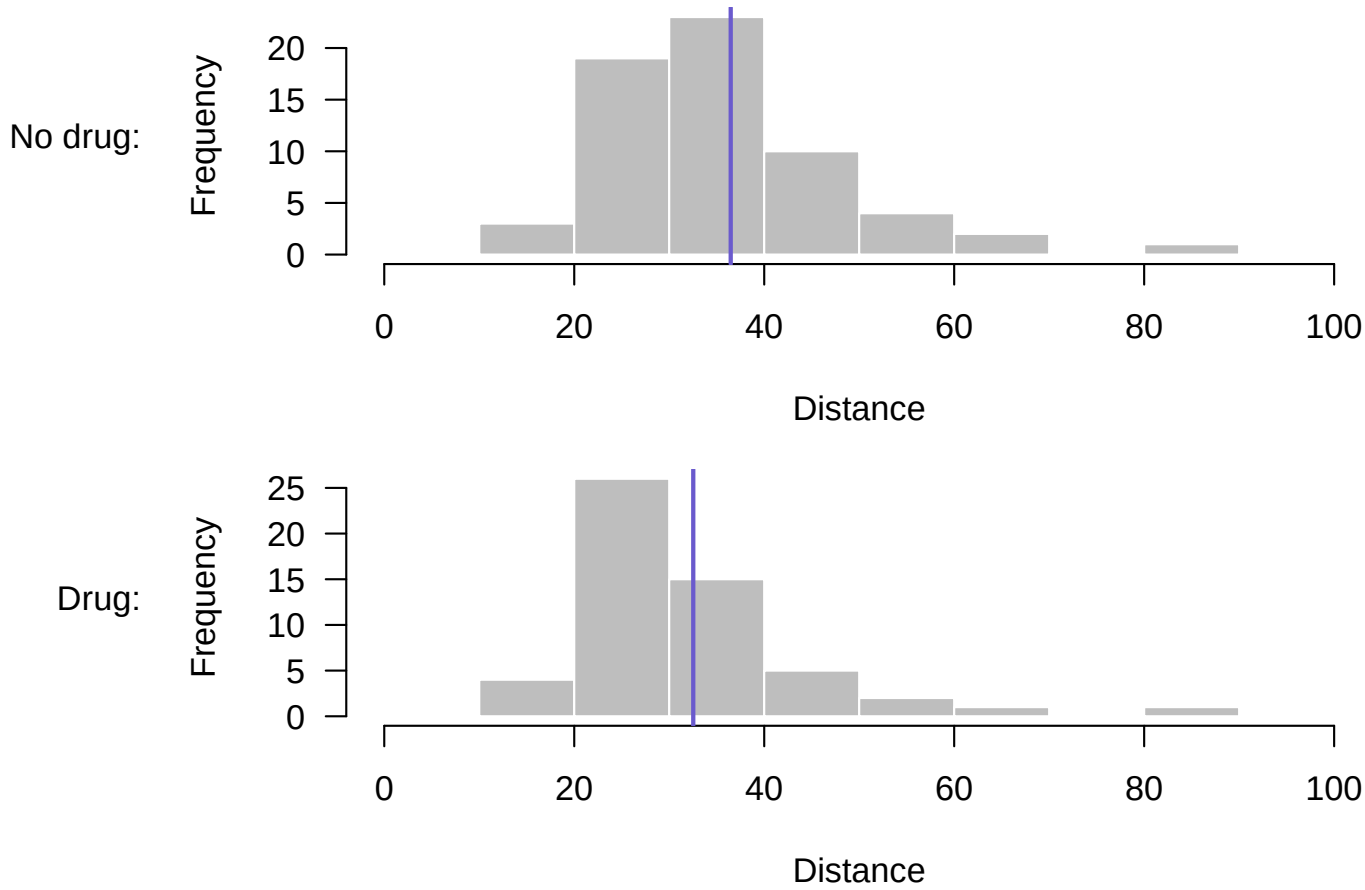
- However, throwing away observations is a questionable practice
- Perhaps computer glitches, human errors, or subjects not taking the study seriously were problems for other observations, too, but they just didn't stand out as much
- Throwing away outliers often produces a distorted view of the world 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, they might have a hard time publishing this study in certain journals because the p-value is above .05
- They could go back, collect more data and refine their study design, but that would be a lot of work
- An easier solution would be 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 p-value of .045
- This manner of picking and choosing which data we are going to allow into our study is at best questionable, and worst scientific dishonesty
- A much better approach is to keep all subjects in the data set, but analyze the data using a method that is robust to the presence of outliers

Outliers: Remarks

Before deriving such an approach, let me just make a few final remarks about outliers

- Outliers have a large impact on many types of analyses, and are without question worthy of attention and investigation
- 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
- Finally, outliers can be the most important and interesting observations of all

How to measure "extreme-ness"

- Recall the definition of a p-value:

$$p(x) = P\{T(X) \geq T(x)|H_0\}$$

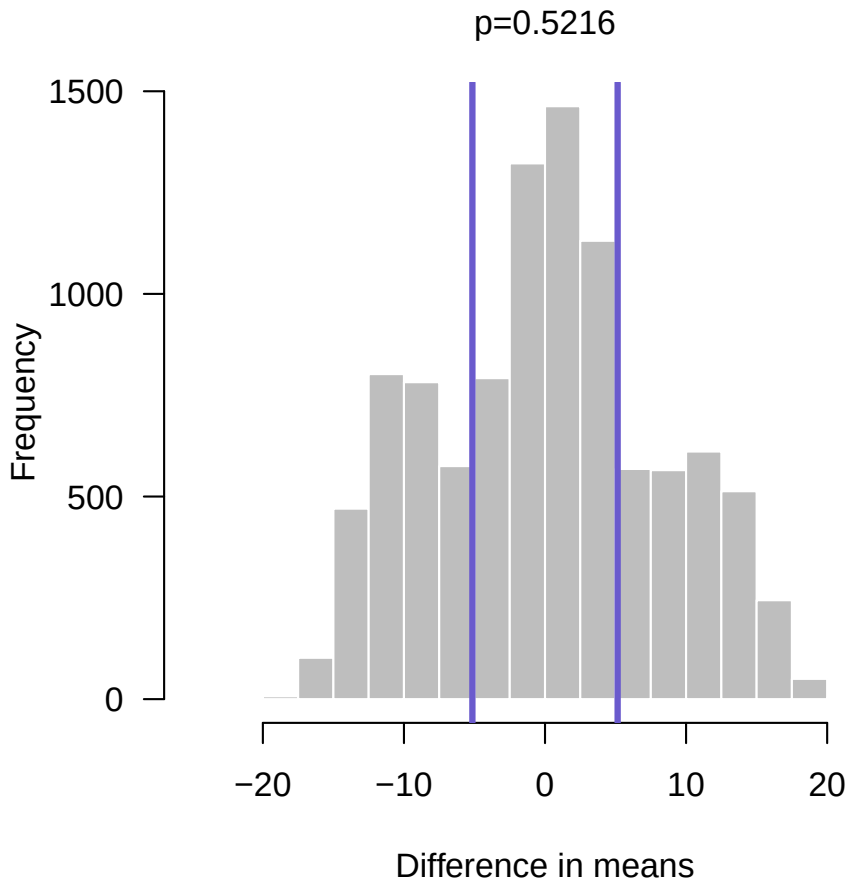
- The t-test uses $T(x) = |\bar{x}_1 - \bar{x}_2|$ as a measure of "extreme-ness", but suppose we wanted to use some other measure, such as the median, that is more robust to outliers
- How could we calculate the probability above?
- One powerful approach is to employ the same concept that we encountered in the Fisher's exact test: condition on the observed values of X and view those outcomes as balls in an urn; then the null hypothesis of no difference between groups becomes the hypothesis that all the balls are drawn from the same urn

Viewing our study as balls in an urn

- Specifically, for the tailgating study, suppose we wrote down each subject's average following distance on a ball
- If drug use is independent of tailgating behavior, then our experiment is equivalent to putting all 119 balls in the same urn, drawing out 55 and calling them the "illegal drug users" group, and letting the 64 balls remaining in the urn represent the "non-illegal drug users" group
- We could do this repeatedly and measure the fraction of time we see the event $T(X) \geq T(x)$

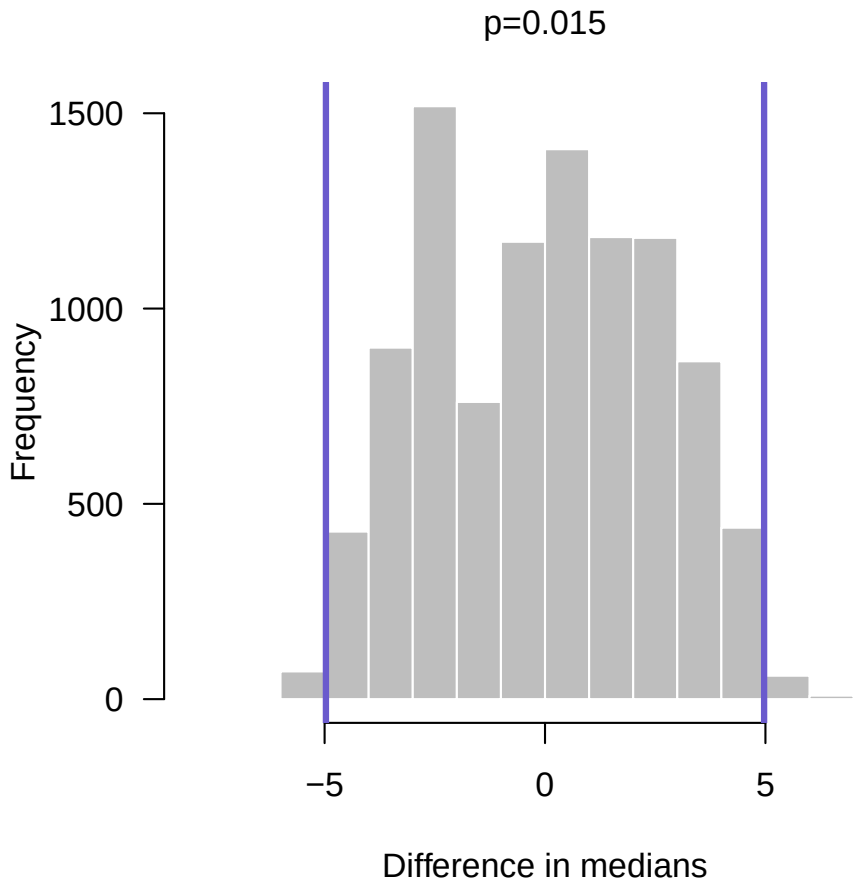
Results of the experiment: Means

Using $T(x) = |\bar{x}_1 - \bar{x}_2|$:



Results of the experiment: Medians

Using $T(x) = |\tilde{x}_1 - \tilde{x}_2|$, where $\tilde{x}$ denotes the median of x:



Remarks

- We have obtained two very different results here; keep in mind that this has *nothing* to do with any distributional assumptions and *everything* to do with our choice of $T(x)$
- This approach to carrying out a hypothesis test is called a *permutation test*; another way of thinking about the test is that we are 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 to the permutation test are rather time-consuming to calculate, and unless the number of observations is small, we must typically approximate the exact answer by using a large number of random permutations (I used 10,000 in this illustration)

Rank-based methods

- Rather than investigate the difference in medians, a related, but slightly different, approach is to consider the ranks of the observations
- Essentially, ranking the data is another kind of transformation, one that works quite well with almost any distribution
- By ranking the data, the impact of outliers is mitigated: regardless of how extreme an outlier is, it receives the same rank as if it were just slightly larger than the second-largest observation
- Also, any problem of skewness is eliminated, because all ranks are equally far apart from each other

Tailgating ranks

For example, instead of looking at the actual following distances, we could look at the ranks of the following distances:

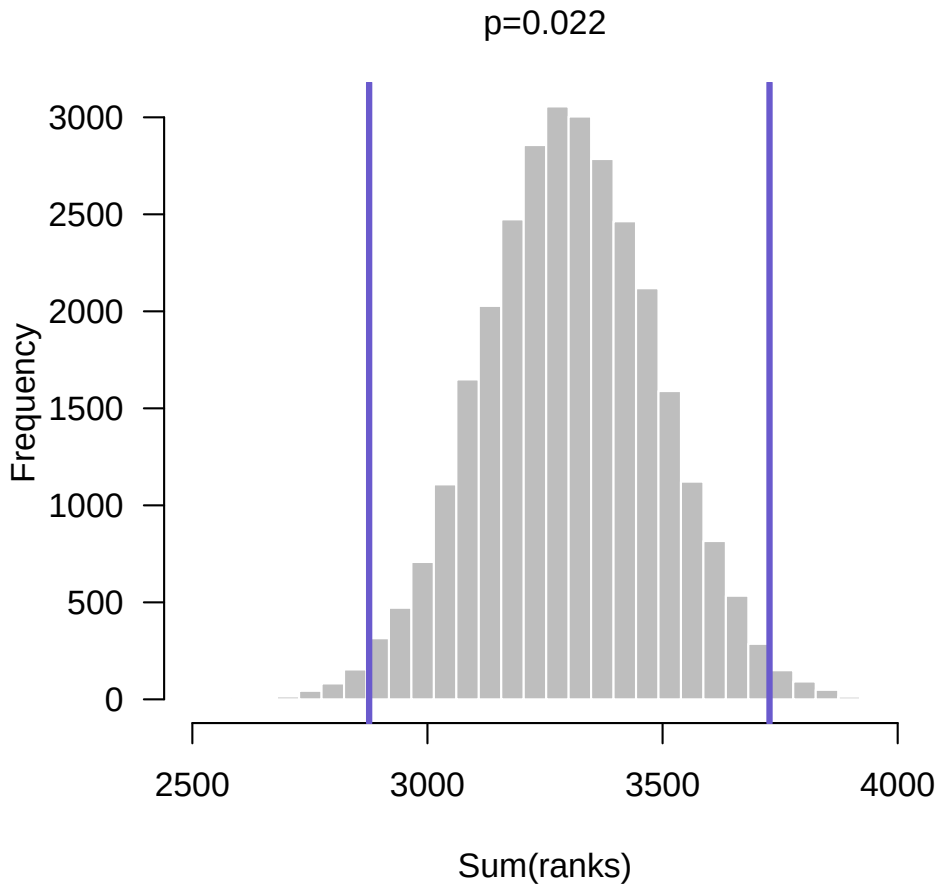
Following distance	Rank
17.89	2
38.96	88
38.31	85
28.58	40
27.70	33
49.76	104
28.91	44
20.38	9
34.03	68
68.34	114
...	...

The Mann-Whitney/Wilcoxon test

- Now consider a permutation test using as a test statistic $T(X)$ the sum of the ranks in the drug user group
- This approach to hypothesis testing (rank-then-permutation-test) is called either the *Mann-Whitney U test* or the *Wilcoxon rank-sum test*; I'll use the two interchangeably, or use the abbreviation "MWW test"
- It is a very common approach to testing for differences between two groups when one is concerned about normality/skewness/outliers – any of the things that can cause problems with the t-test

Wilcoxon rank sum test

Using $T(x) = \sum_{i:g_i=1} r_i$, where r_i is the rank and g_i is the group membership of the ith observation



Calculating the Mann-Whitney/Wilcoxon test

- Working with ranks has two important practical advantages over the "difference of medians" permutation test:
  - It is reasonably straightforward to carry out an approximate version of the test by hand based on the idea that the sum of the ranks approximately follows a normal distribution
  - Statisticians have developed clever ways of calculating exact p-values for permutation tests in the special case where the data are consecutive positive integers (i.e., ranks) that are much faster than the brute force permutation test approach
- Thus, many software packages will offer an option to calculate exact p-values for the Mann-Whitney/Wilcoxon test; this is usually quite fast, although for large sample sizes it can still be computer-intensive, so software packages also take various shortcuts and approximations; for this reason, MWW p-values may differ slightly depending on the program you are using

Tailgating study: Mann-Whitney test

- Applying the Mann-Whitney test to the tailgating study, we obtain a p-value of .02, very similar to what we obtained with the "difference of medians" permutation test
  - Exact p-value: 0.0236
  - Approximate p-value: 0.0238 or 0.0240, depending on the approximation
- Note that by ranking the data, we have minimized the impact of the outliers, conducted a test that doesn't rely on any assumptions about the distribution of the data, avoided arbitrary decisions about which observations to throw away, and even obtained a more significant p-value
- This is a very sound, safe approach to analyzing this data; indeed, it was the approach chosen by the investigators when they published this study

Nonparametric statistics

- Statistical methods like the t-test may be called "parametric", since unknown parameters (i.e., μ) and their effect on the distribution of data are central to the approach
- In contrast, the Mann-Whitney/Wilcoxon test involves no parameters whatsoever; such methods are referred to as *nonparametric* to highlight this fundamental difference
- The advantage of nonparametric methods is that they make fewer assumptions and don't get derailed when those assumptions go wrong – for example, when outliers are present
- The disadvantage of nonparametric methods is that we are often interested in estimating and obtaining confidence intervals for parameters, and nonparametric methods are not always helpful in this regard

Permutation tests have low power when n is small

- The Mann-Whitney/Wilcoxon test is an essential alternative to the t-test, and requires no assumptions about the population distribution
- However, it is a permutation test, and like any permutation test, it has little to no power for very small sample sizes (as you will see in the next homework assignment)

Power and nonparametric tests

- Don't read too much into this, however
- The difference in power is far less dramatic when the sample size is larger (for large sample sizes, the Mann-Whitney/Wilcoxon test is about 95% as powerful as the t-test, even when the outcome is normally distributed)
- Furthermore, as we saw in the driving study, when outliers/skewness are present, nonparametric methods can be much more powerful than t-tests

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
- Permutation tests are a flexible and useful method for testing differences without making distributional assumptions
- Rank-based methods are a powerful way to analyze data when distributional assumptions are questionable, and particularly effective in the presence of outliers
- Parametric vs. nonparametric:
  - Parametric advantages: More powerful when parametric assumptions hold, straightforward confidence intervals
  - Nonparametric advantages: Minimal assumptions, more powerful when parametric assumptions are wrong