Section 4: Assumptions, Transformations, Rank-Sum Tests

TF: Reagan Rose (rrose@g.harvard.edu)

September 29, 2015

Assumptions

To carry out hypothesis tests, we have been focusing on the **t-test**. But to do the t-test, we had to derive the **t-distribution**, which was based on the *chi-squared distribution*, which was based on the *normal distribution*.

In other words, **the entire t-test is dependent on the assumption that**

$$X_1, \ldots, X_n \sim^{iid} \mathcal{N}(\mu, \sigma^2).$$

So to run a t-test, we must first assume that

1. The observations X_i are independent
2. The observations X_i come from a normal distribution

Can you think of some situations where these two assumptions may be violated? It happens often! But sometimes, even though our assumptions are violated, we don't have the tools to do any other type of test. This is the concern of **robustness**, which tells us how well a test performs when its assumptions are violated. We have two main measures of robustness:

1. **Type I Error:** $P(\text{Reject } H_0 | H_0 \text{ true}) = \alpha$
2. **Power:** $P(\text{Reject } H_0 | H_0 \text{false}) = 1 - P(\text{Don't reject } H_0 | H_0 \text{false}) = 1$ - Type II Error

Violations of Independence

What are some common circumstances in which independence of observations may be violated?

- **Natural clustering****: Suppose we're interested in how students do on standardized tests, so we sample students from several schools in Massachusetts. It may be unreasonable to assume that students *across* schools are as independent as students *within* schools. Sampling two students from the same family, or a set of twins may also violate independence assumptions.
- **Time dependence**: Stock market prices today are certainly *not* independent from the prices of yesterday. This is why special methods for time series exist!
- **Spatial dependence**: The amount of rainfall in Cambridge, MA in 2014 and the amount of rainfall in Somerville, MA in 2014 are probably not independent.

The main problem with independence violations is that they lead to **incorrect variance estimates**. Independence violations are best when accounted for *a priori* - meaning we think about how our data may be dependent and control for that when sampling. *A posteriori* or *ad hoc* independence violations can be a serious problem, and are generally dealt with by using more complex statistical methods.

Violations of Normality

In general, no distribution is 100 % normal. Whether the distribution underlying your data is "normal enough" for the t-test to work is something we can figure out by looking at:

1. **Histograms**: This is the go-to method. Check out the histogram, does it look *normal-ish*?
2. **QQ-plot**: A more refined graphical method, essentially plots the empirical quantiles of the data against what we'd expect to see if our data was truly normal.

While t-test are **generally pretty robust** to most normality violations, severely non-normal or skewed data can lead to invalid results. To deal with these issues, we try to make our data look "more normal" using **transformations**.

Transformations

The main data transformations that people use are (in order of popularity):

1. **Log**: $X_i = ln(Y_i)$
2. **Square root**: $X_i = \sqrt{Y_i}$
3. **Reciprocal**: $X_i = 1/Y_i$
4. **Logit**: $X_i = logit(Y_i) = ln\left(\frac{Y_i}{1-Y_i}\right)$

There are many "rules of thumb" out there that can help you determine which transformation to use given what your data looks like, but in practice many people just try out some (or all) of the above and find the best fit!

Wilcoxon Rank-Sum Test

In practice, people ***really*** like to use t-tests and statistical methods that rely on **normal distribution theory**, because these methods are more reliable and give us the best way to *generalize* our results and to use our data to make *predictions* about the future.

But what if our normality assumptions are severely violated, and we can't transform or change our data to make it better fit assumptions of the normal theory procedures? Then we have to rely on **nonparametric tests**, which are tests that do not make distributional assumptions about the underlying data (remember the permutation test). Cue the **Wilcoxon rank-sum test**, a nonparametric test that can bypass the normality assumptions and give us results similar to those of a *two-sample t-test*.

The idea: We have two samples, say A and B, with $n_A < n_B$ observations in each samples. We want to test whether these two samples are drawn from the same distribution. Then we definethe rank of an observation as the place that it occupies after we order all the observations (both groups), denoted rank(YA,i) or rank(YB,i). The the rank-sum test statistic is:

$$T = \sum_{i=1}^{n_A} rank(Y_{A,i}).$$

To get a p-value, we permute our data a bunch of times, and see how likely it is that we would have seen our observed test statistic T_{obs}, if the samples, A and B, were truly drawn from the same distribution. If it is unlikely, we reject the null and conclude that groups A and B are different!

Exercises

Question 1. Suppose we're interested in whether Starbucks or Dunkin Donuts has better tasting coffee. We go to 7 different Starbucks and 10 different Dunkin Donuts locations around Harvard and score the coffee based on a number of factors. The data is below:

Starbucks	8.50	9.48	8.65	8.16	8.83	7.76	8.63		
Dunkin Donuts	8.27	8.20	8.25	8.14	9.00	8.10	7.20	8.32	7.70

Conduct a hypothesis test using a rank-sum test with 1000 simulations, and state your conclusions. Which coffee should I drink?

Question 2. Reddit data Download the reddit data from Canvas or my Github.

(a) Conduct a t-test of whether the number of upvotes significantly differs between the 'pics' and 'funny' groups.

(b) Use graphical approaches to analyze the normality assumption between the groups. If normality is violated, use different transfors to try generating an approximate normal distribution.

(c) Use the best transform from (b) to conduct a new t-test and compare your conclusions to what you found in (a)

Question 3. Download the wine data and load it into R.

(a) What are your hypotheses and test statistic for the rank-sum test? Which group should be used for the test statistic?

(b) Evaluate T_{obs} for this data.

(c) Do $n = 1000$ simulations of the rank orders, and compute the simulated p-value for your T_{obs}, and determine your conclusion.

(d) Try doing the t-test for the same hypotheses, and compare your conclusions from part (c).