Section 2: Probability and Hypothesis Testing

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

September 16, 2015

Hypothesis Testing

Suppose you'd like to answer a question with respect to some *scientific hypothesis*, and you conduct an experiment to do so. How do we know whether to accept or reject the hypothesis?

1. Formulate the null H_0 and alternative H_a hypotheses. The null generally corresponds to a "status quo" or "no effect", and the alternative is that there is indeed an effect.
2. Calculate a *test statistic*.
3. Compute the *p-value* of the observed value of the test statistic, based on some sampling distribution.
4. Compare the p-value to the desired significance level α. If $p < \alpha$, we reject H_0 in favor of H_a. Otherwise, we **cannot reject** H_0.

To make these steps clear, we clarify some terms.

Test statistic: As noted above, a statistic is a function of the data, and a test statistic is a specific type of statistic that is formulated to answer the scientific question of interest (note that much of this is flexible and under the discretion of the analyst).

p-value: Arguably the most important (and controversial?) aspect of hypothesis testing, the p-value denotes the probability that we will see a value of the test statistic as high as or higher than than the observed.

Significance level: After computing the p-value, we need some reference probability to compare it to. We denote this as or the significance level; if $p < \alpha$, then we reject H_0.

Randomization + Permutation Tests

Fisher's Randomization Test

The key point of Fisher's Randomization Test is that probability (or uncertainty) only seeps into our analysis when we use randomization for the assignment. The reasoning goes as follows:

1. Everyone has a "true effect" under control and treatment.
2. If we knew exactly which individual was in which group (control/treatment), then we would know exactly the size of the measured effect (i.e. we could measure again and again and obtain the same results, because the same individuals are in the same groups).
3. When we randomize assignment of individuals to groups, we lose the distinction of which individual is in which group.

Thus, all we need to model probabilistically is how the randomization affects our average effect size. We have a number of assumptions:

- Random assignment to groups
- Under H_0, units are exchangable
- Treatment effect is additive $Y_{treatment,i} = Y_{control,i} + \delta$

Our null hypothesis is that $H_0 : \delta = 0$ and $H_a : \delta \neq 0$. The test statistics is then

$$\hat{\delta} = \bar{Y}_{treat} - \bar{Y}_{control}.$$

Now under H_0, it shouldn't matter whether a subject is assigned to control or treatment; the estimated $\hat{\delta}$ should stay the same. Thus, the idea of Fisher's Randomization Test is the following:

1. Permute the units put into the control/treatment groups, while keeping the observed outcomes the same.
2. Record the new simulated $\hat{\delta}$ under each permutation.
3. Using the empirical (observed) distribution of $\hat{\delta}$, compute the proportion of simulated estimates that are at least as high as the observed $\hat{\delta}$ in the actual data.
4. This is the estimate of the p-value, so compare to α.

Permutation Test

The permutation test is essentially a generalization of the randomization test to observational studies. In this case, we do not randomize subjects into groups (since this is only possible in an experiment, by definition).

However, subjects may be naturally grouped (i.e. smoker/non-smoker). To examine whether this grouping has an effect on some outcome (i.e. lung cancer), we can compute the difference of some variable between groups 0,1:

$$\hat{\delta} = \bar{Y}_1 - \bar{Y}_0.$$

Now, under the corresponding $H_0 : \delta = 0$, we should expect that relabeling units between groups should have no effect. Thus, we can again permute group labels as we did in the randomization test, and compute $\hat{\delta}$ under each permutation (note that we permute *group labels*, and keep the *outcomes* constant).

Exercises

Question 1. Bayes Rule
The test for a rare disease is known to be 98% accurate at predicting the disease for those who are carriers, and 97% accurate at classifying non-carriers (i.e. for those who don't have the disease, it is correct 97% of the time). This disease is rare and occurs randomly in 5% of the population. You test positive for the disease. What is the probability that you are a carrier?

Question 2. Horseshoe Crabs
Download the horseshoe crab data, available at *https://github.com/reaganrose/Stat139*

(a) Compute the following summary statistics for the weight of the horseshoe crabs in the dataset: mean, median, 1st quartile, 3rd quartile, minimum, and maximum.

(b) Create boxplots for weights depending on whether the crab has color denoted 1 or 0.

(c) Compute the correlation between color and weight. Why might this not be so meaningful? What does the relationship suggest?

(d) Conduct a permutation test using color as a grouping, and report your results for whether there is a significant difference in weights between the color 1 and 0 groups.

Question 3. 'Funny' or 'Pics'
Now we have an actual dataset consisting of Reddit posts in various subreddits, and their number of upvotes. This data is available at *https://github.com/reaganrose/Stat139*. We are interested in whether the posts in the 'funny' subreddit (labeled 1) garner more upvotes than those in the 'pics' subreddit (labeled 0).

(a) State clearly the null and alternative hypotheses to answer the question provided.

(b) Use descriptive statistics and graphical output ('exploratory data analysis') to make an argument for or against the given hypothesis.

(c) Use an appropriate hypothesis test and compute the p-value to determine whether posts in the 'funny' subreddit do indeed garner more upvotes, for a significance level of $\alpha = 0.05$.