Stat 139: Statistical Sleuthing through Linear Models

Section 5: Power + Multiple Comparisons

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

October 6, 2015

Type I and Type II Error

Remember the classic table for understanding Type I and Type II error:

	H_0 true	H_0 false
Fail to reject H_0	-	Type II Error
Reject H_0	Type I Error	-

In particular, these correspond to the **probabilities**:

$$P(\text{Type I Error}) = P(\text{Reject } H_0 | H_0 \text{ true}) = \alpha$$

$$P(\text{Type II Error}) = P(\text{Fail to reject } H_0 | H_0 \text{ false}) = \beta.$$

In layman's terms, Type I error can be interpreted as the **probability of picking up a random effect**, and Type II error can be interpreted as the **probability of missing a truly significant effect**. Note that there is *always a tradeoff between Type I and Type II error.* Allowing for more Type I error reduces Type II error, and vice-versa.

Exercise: Can you think of a situation where it is more important to reduce Type I Error? What about Type II Error?

Power and Effect Size

So far in this course, we've focused a lot on Type I error, but haven't really discussed the Type II error. This is discussed in relation to **statistical power**, which tells us our **power to detect a significant effect in the data**.

Power $= P(\text{Reject } H_0 | H_0 \text{ true}) = 1 - P(\text{Don't reject } H_0 | H_0 \text{ true}) = 1 - P(\text{Type II Error}) = 1 - \beta.$

The way we examine power is as follows:

1. Compute the critical value x^* so that the probability under the null of values more extreme than x^* is α (i.e. $P(X > x^*|H_0) = \alpha$).
2. Compute the probability $P(X > x^*|H_a)$ - this is the power

The power of a test depends a lot on **how much we can trust our data**. This is influenced by the following:

- **Sample size**: In general, as sample size increases, we're more sure that our data reflects the true distribution, so statistical power is higher.
- **Effect size:** As the effect size (i.e. the true difference between groups (under H_a)) increases, it becomes easier to detect the effect, so power increases.
- **Standard deviation:** As the standard deviation of observations increases, we become less sure of our data and the distributions of the data under the null and altnerative hypotheses widen, so statistical power decreases.

When we use a test statistic such as the sample mean to test our hypotheses (as in a t-test), since the sample mean is normally distributed, the power computation for the hypotheses $H_0 : \mu = \mu_0$, and $H_a : \mu = \mu_a$ becomes:

$$P(\text{Reject } H_0|H_0 \text{ false}) = P(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > 1.96|H_0 \text{ false}) = P(\frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} > x^*)$$

where $x^* = \frac{\mu_0 - \mu_a + 1.96\sigma/\sqrt{n}}{\sigma/\sqrt{n}}$. So if you want a power of at least $1 - \beta$, then we can solve for this by noting that $x^* \leq z_{1-\beta}$.

Multiple Comparisons

The primary problem of multiple comparisons is that having a significance level of α only bounds the Type I error **for one study**. Suppose that we see the results of 10 independent studies that rejected the null, each with the same significance level of α. What is the probability that at least one of them reported a result due to Type I error (purely due to chance)? For any given study, the probability is α. However, for the 10 studies considered jointly, the probability is

$$P(\text{at least one Type I error}) = 1 - P(\text{no Type I error}) = 1 - (1 - \alpha)^{10}.$$

So if $\alpha = 0.05$ and we are considering 10 studies that reported significant results, the probability is

$$1 - (1 - .05)^{10} = 40.1\%$$

If we're considering 100 studies that reported significant results, this probability is 99.5%! There are many methods out there that can be used to correct for this problem. They do so by **bounding the overall Type I error across multiple comparisons**. One of the most popular (also most naive, and most conservative) is called the **Bonferroni correction**, which says that if we compare pairs of I groups, then we should se

$$\alpha^* = \frac{\alpha}{\binom{I}{2}}$$

Exercises

Question 1. Understanding Power. Suppose we are interested in knowing the mean GPA of all Harvard students, and we want to design a study to find out. We begin by assuming that the mean GPA is 3.35, with a known standard deviation of $\sigma = 0.15$.

(a) State the null and alternative hypotheses for the purpose of calculating power.

(b) If we sample $n = 25$ students, what is the rejection region for our null hypothesis (i.e. at what value of $\bar{X}$ do we reject H_0)?

(c) If the true mean if actually $\mu = 3.3$, what is the power of the test in part (a)?

(d) What would the power be if the true mean was actually $\mu = 3.5$?

(e) Now suppose again that the true mean is $\mu = 3.3$ and we want to have power of at least 80% at the $\alpha = 0.05$ signfiicance level of two-sided test. What is the minimum number of students we need to sample?

(f) Would the necessary sample size increase or decrease if we wanted a power of 90%? Suppose, due to time or resource constraints, we cannot sample any more than $n = 25$ students. What else about the study could we change to try to increase the power?

Question 2. Simulating Multiple Comparisons. Suppose we would like to explore whether any of K genes are linked to cancer at the $\alpha = 0.05$ significance level. Here, we are doing **one sided tests** for each of the genes. Unfortunately, none of the genes are actually related to cancer.

(a) Calculate the probability of rejecting H_0 for at least one of the K tests. What is this when $K = 20$?

(b) Using the Bonferroni correction, calculate the α^* needed such that the overall Type I error rate is α.

(c) Simulate the following in R:

  (i) Assume cancer rates are distributed $X \sim N(0, 1)$ and we are testing the hypothesis $\mu = 0$ for each of $K = 20$ different genes. Find the proportion of times that H_0 is rejected on average for a **single test**. Is it what you expect?

  (ii) Now simulate all $K = 20$ random variables and test the null hypothesis that all means are simultaneously 0. What is the proportion of significant results? (Note: this should be close to your answer in (a)).

  (iii) Use the Bonferroni correction on the above simulation. Now what is your observed Type I error rate?