Lab 7: Hypothesis Testing

August 9, 2016

1 Goals

Today in lab, we will review hypothesis testing concepts

- Examine the relationship between level and power
- Examine data from the NFL "Deflategate" saga

2 Level vs Power

In the first section of the lab, we will look at the relationship between a hypothesis test's level and power. In particular, we will see that there is no free lunch and that trying to minimize the Type I error often comes at the cost of the probability of a Type II error.

The level of a test is a (hopefully) pre-determined cut-off. When the P-value of the data is less than the level, we reject the null hypothesis. If the p-value is larger than the level, we fail to reject the null hypothesis. This cut-off is typically set at .05.

$$\text{p-value} = P(\text{Test statistc as or more extreme than what we actualy observed } | \text{Null Hypothesis})$$

If we reject the null hypothesis anytime a p-value is less than .05, then we will actually incorrectly reject the null hypothesis .05 of the time when the null hypothesis is actually true. This is because the rejection region is unlikely under the null hypothesis, but not impossible, and even unlikely events happen sometimes. This is a Type I error.

So to minimize the probability of committing a Type I error when the null hypothesis is true, why don't we simply lower the cut-off so that I only reject the null hypothesis .025 or .01 or .001 of the time. This would lower my probability of a Type I error. Let's take a look at why things are not that simple.

2.1 Setting up the Null Distribution

I am working for a pharmeceutical comapny interested in producing a new drug for combatting cancer. It is impossible to prevent side-effects in all patients, but suppose I am willing to release the drug if the proportion of individuals who do not experience side effects is higher than .9. Thus, if p represents the proportion of individuals who do not experience side effects, I am interested in testing

$$H_0 : p = .9$$
$$H_A : p > .9$$

Suppose I get a sample of 1200 individuals representative of the population and test what proportion of the individuals experience side effects.

2.2 Questions

- What is a Type I error in this context? What is the practical cost of committing that error in this context?
- What is a Type II error in this context? What is the practical cost of committing that error in this context?
- What is the distribution of $\hat{p}$ under the null hypothesis?
- Given a cut-off of .05, for what observed values of $\hat{p}$ would I reject the null hypothesis?

2.3 Simulating Data

The null distribution, given that $n = 1200$ and $p = .9$ is approximately

$$\hat{p} \dot{\sim} \mathcal{N}\left(.9, \frac{.9(1-.9)}{1200}\right)$$

Under a standard normal distribution, the cut-off with .05 to the right is 1.645. Note that here we need a cut-off to the right, because that is the alternative hypothesis we are interested in.

Thus, we can translate the z-score of 1.645 to a value of $\hat{p}$

$$\frac{\hat{p}_{cutoff} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = 1.645$$

$$\Rightarrow \hat{p}_{cutoff} = 1.645 \times \sqrt{p_0(1-p_0)/n} + p_0 = 0.914$$

So we would reject the null hypothesis whenever $\hat{p} > \hat{p}_{cutoff} = .914$. Let's simulate 5000 random samples assuming that the null distribution is true, and confirm that criteria leads to rejecting a Type I error rate of .05.

```
# number of samples
sim.size <- 5000
# placeholder for results
p.hat <- rep(0, sim.size)
for(i in 1:sim.size){
  # simulate a sample of individuals who do not experience side effects
  p.hat[i] <- rbinom(n = 1, size = 1200, prob = .9) / 1200
}

### Note that you could do this without a for loop by simly using
# p.hat <- rbinom(n = sim.size, size = 700, prob = .9) / 700


# check to see if we would reject the null hypothesis
reject <- p.hat > .914

# See what proportion of the time we reject the null hypothesis
mean(reject)
```

```
## [1] 0.0526
```

As you can see from the simulated data, we reject the null hypothesis rougly .05 of time, even though the null hypothesis is actually correct. Now, let's keep the same procedure, where we reject the null hypothesis when $\hat{p} > .914$. However, now assume that the null hypothesis is not actually correct and that the true proportion is actually .91. Let's check and see how often I am able to correctly determine that the null hypothesis is incorrect.

```
# number of samples
sim.size <- 5000
# placeholder for results
p.hat <- rep(0, sim.size)
for(i in 1:sim.size){
  # simulate a sample of individuals who do not experience side effects
  # notice that prob is now .93, not .9 as under the null hypothesis
  p.hat[i] <- rbinom(n = 1, size = 1200, prob = .91) / 1200
}

### Note that you could do this without a for loop by simly using
# p.hat <- rbinom(n = sim.size, size = 700, prob = .9) / 700


# check to see if we would reject the null hypothesis
reject <- p.hat > .914

# See what proportion of the time we reject the null hypothesis
mean(reject)
```

```
## [1] 0.3176
```

We can see that now, we are able to correctly identify that the true proportion is not .9 in about 33% of the samples we take. This is the **power** of the test when $p = .91$.

2.4 Questions

- Try the same simulation, but set the true $p = .92$ or .95. How does the power change as the true state of the world is further away from the null hypothesis?
- Using the qnorm function, calculate the cut-off value ($\hat{p}_{cutoff}$) when the level of the test is .025 and .01 respectively.

2.5 Decreasing the Type I error rate

Now suppose, if we mistakenly release a drug which does not have a non-side effect proportion that is larger than .9, there will be a large fine from the Food and Drug Administration (FDA). Thus, I want to be really careful and decrease the Type I error rate.

2.6 Questions

- Using the values you calculated above, re-run the simulations and see what the power of the new cut-offs is when $p = .91$.
- How does this change when the true $p = .92$ or .95.

- In a few sentences, summarize the take away message. Specifically, what happens to power as the true parameter get further away from the null hypothesis. Also, what happens to power as you decrease the level of the test?

3 Deflategate

For the past decade, the New England Patriots have been a dominant team in the NFL. However, they have often been accused of taking unethical advantages and sometimes down right cheating. In particular, in January 2015, the Indianopolis Colts accused the New England Patriots of using footballs which were not inflated to the proper NFL standard. The NFL requires that footballs be inflated to somewhere between 12.5 and 13.5 pounds per square inch (PSI). However, Tom Brady, the Patriots quarterback and former Uggs model (seriously, look it up), prefers slightly deflated footballs because it makes the balls easier to grip, throw and catch. Although the referees typically check the balls before each game to make sure they fall within the NFL standards, the Patriots were accused (and convicted) of secretly tampering with the balls after the official check and letting additional air out of the ball.

For a comprehensive description of the Deflategate saga, read the Vox writeup. The NFL commissioned a detailed investigation into the matter which was lead by attorney Ted Wells. The resulting "Wells Report" is available here. In particular, we will be using data gathered from page 168.

Before the January 18, 2015 game, the referees measured the balls for both the Patriots and Colts and noted that the Colts balls all seemed to be roughly 13 psi, while the Patriots balls were right around 12.5 (the lower limit). Two Patriot balls were under the legal limit, and were inflated slightly to bring them up to the limit. However, during the game, the Colts complained, so the referees re-measured the PSI of 11 Patriot balls and 4 of the Colt's balls during half-time. We will be analyzing that data today. For a variety of reasons (air pressure decreases when it is cold outside, footballs deflate slightly during normal game conditions, etc) the NFL expects a normal amount of deflation during the course of a game even without any foul play. So, the question of interest is not necessarily were the Patriot's balls less than the 12.5 psi lower bound (if they started at 12.5, we would expect them to be legally less than 12.5 by halftime), but if the Patriot's balls deflated more than the Colt's balls by a statistically significant amount. Thus, let $\mu_{patriots}$ and μ_{colts} be the true amount of deflation (pressure drop) for each team. We are interested in

$$H_0 : \mu_{patriots} - \mu_{colts} = 0$$
$$H_A : \mu_{patriots} - \mu_{colts} > 0$$

3.1 Questions

- What type of test will we be using? (ie what line of table from the lecture notes should we use?)
- What is the test statistic?
- What is the Null Distribution of the test statistic?

3.2 Data Analysis

First let's read in the data. In particular, there were 2 pressure gauges used by the referees at halftime, we will be using the readings from gauge 2.

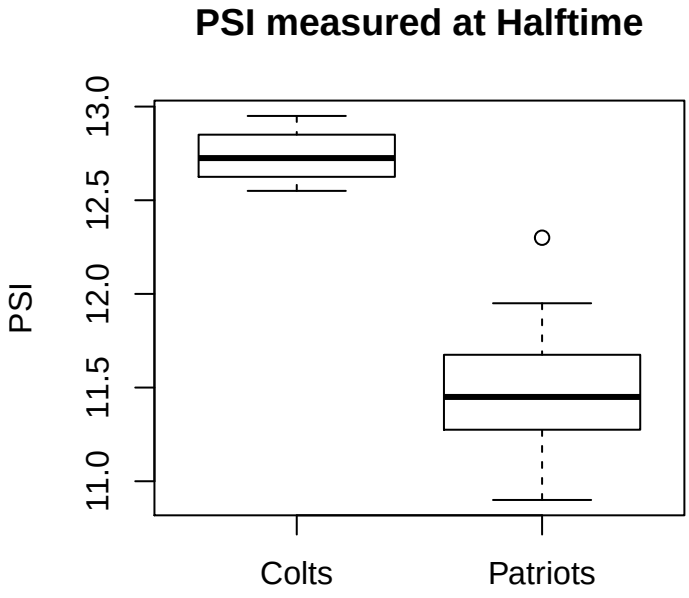
```
# read in data
deflate <- read.csv("http://www.stat.washington.edu/~ysamwang/notes/deflategate.csv")
names(deflate)
```

```
## [1] "Team"         "Ball"         "Gauge1"       "Gauge2"
## [5] "InitialPSI"   "PressureDrop"
```

Team	Ball	Gauge1	Gauge2	InitialPSI	PressureDrop
Patriots	1	11.50	11.80	12.50	0.70
Patriots	2	10.85	11.20	12.50	1.30
Patriots	3	11.15	11.50	12.50	1.00
Patriots	4	10.70	11.00	12.50	1.50
Patriots	5	11.10	11.45	12.50	1.05
Patriots	6	11.60	11.95	12.50	0.55
Patriots	7	11.85	12.30	12.50	0.20
Patriots	8	11.10	11.55	12.50	0.95
Patriots	9	10.95	11.35	12.50	1.15
Patriots	10	10.50	10.90	12.50	1.60
Patriots	11	10.90	11.35	12.50	1.15
Colts	1	12.35	12.70	13.00	0.30
Colts	2	12.30	12.75	13.00	0.25
Colts	3	12.50	12.95	13.00	0.05
Colts	4	12.15	12.55	13.00	0.45

Just by graphing the measured PSI, we can see that the Patriot balls had a much lower PSI on average than the Colt's footballs. However, since we expect a normal amount of deflation due to game play and weather, we want to actually see if the Patriots balls deflated by more than the Colt's balls. To measure this, we assume (as does the NFL's Well's Report) that the Colt's balls started at 13 PSI and the Patriots started at 12.5 PSI.

```
boxplot(Gauge2 ~ Team, data = deflate, main = "PSI measured at Halftime",
        ylab = "PSI")
```



3.3 Questions

- What type of test will we be using? (ie what line of table from the lecture notes should we use?)
- What is the test statistic?
- What is the null distribution of the test statistic?
- What is the average amount of deflation for the Patriots?
- What is the average amount of deflation for the Colts?
- What is $s_{patriots}$ of the amount of deflation?
- What is s_{colts} of the amount of deflation?
- Calculate the test statistic.
- Calculate the p-value.
- What would you conclude based on the p-value? Explain it in words.

You can check your work above with the code below.

```
# patriot deflation
avg.pat <- mean(deflate$PressureDrop[deflate$Team == "Patriots"])
s.pat <- sd(deflate$PressureDrop[deflate$Team == "Patriots"])

# colts deflation
avg.colts <- mean(deflate$PressureDrop[deflate$Team == "Colts"])
s.colts <- sd(deflate$PressureDrop[deflate$Team == "Colts"])

# Test statistic
test.stat <- (avg.pat - avg.colts) / sqrt(s.pat^2 / 11 + s.colts^2 / 4)

# P-value; Note that we take complement because we want the area to
# the right of our test statistic because that is the direction of
# the alternative hypotheis
1 - pt(test.stat, df = 3)
```

```
## [1] 0.007477918
```