Midterm 2
Takehome Portion

Zoe Farmer

March 21, 2014

1 Premise

You are inspecting computer printers for damage, and the manufacturer has told you that 10% of their printers are damaged.

Sidenote, all `R` code is included inline, using KnitR and LaTeX. If you have any question about the contents of this report, or the methods used to create it, please reference `www.will-farmer.com` for my email address and other information.

2 Questions and Answers

1. *You are planning to inspect n printers from the manufacturer's warehouse. If the actual percentage of damaged printers is 10%, what does n need to be so that the width of your 95% confidence interval for the true proportion of damaged printers is less than* 0.11*?*

   (a) Looking at this as a Binomial Random variable, we can denote p to be the percentage of "successes", in this case the number of damaged printers that we find. We can thereby express this as

   $$\sigma_X = \sqrt{np(1-p)}$$

   We have an estimate for p, which we will denote as $\hat{p}$, which we know has approximately normal distribution. This let's us establish a similar definition as above.

   $$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

   With all these facts we can establish our confidence interval to be

   $$P\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_{\alpha/2}\right) = 0.95$$
   $$\Rightarrow \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

   Establishing our width as w and solving for n we obtain

   $$\begin{aligned} \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= w \\ 2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= w \\ -\frac{4\left(\hat{p}^2 z_{\alpha/2}^2 - \hat{p} z_{\alpha/2}^2\right)}{w^2} &= n \end{aligned}$$

```
w     <- 0.11
p_hat <- 0.10
alpha <- 0.05
z     <- qnorm(alpha / 2)
n     <- round(- (4 * (p_hat^2 * z^2 - p_hat * z^2))/(w^2))
```

The above computation yields an n of $\boxed{114}$, we will use this n later for better analysis.

2. *You go to the warehouse and sample n printers. If the actual percentage of damaged printers is 10%, on average, how many printers do you expect to count with damage?*
   (a) Using our $n = 114$ from before, we would expect to see 10% of them damaged, or about 11.
3. *You go to the warehouse 100 times, and each time you sample n printers and count the number of them that are damaged. At each visit, you calculate a 95% confidence interval for the percent of damaged printers you counted. How many times do you expect the percentage of printers you calculate to fall in these intervals? How many times do you expect 10% to fall in these intervals?*

   (a) First let us establish our confidence interval.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

We know each of these elements, and we can calculate our interval.

```
error <- abs(z * sqrt((p_hat * (1 - p_hat))/n))
lower <- p_hat - error
upper <- p_hat + error
```

Thusly our interval is $\boxed{(0.0449298, 0.1550702)}$. If we calculate the 95% confidence interval for the percentage 100 times based on the sample percentage, said percentage will fall within the interval every time, since the confidence interval is based off of this value. On the flipside, 10%, our true percentage will only fall within these intervals 95 out of 100 times, since it is the 95% confidence interval.

4. *Simulate data to match part (3), assuming that at each visit, the number of damaged printers you count are distributed* $Bin(n, 0.10)$. *Make a histogram of the percentage of damaged printers you counted at each visit.*

   (a) Simulating appropriate data, we can create a histogram of our damaged printer percentage using our previously calculated n and with our established definition of a confidence interval. Please reference Figure 1, and note that for this plot and all subsequent plots the dashed line indicates the sample mean.

```
db100 <- rbinom(100, n, 0.10)
interval_data100 <- data.frame(printers=db100, percent=db100/n)
ggplot(interval_data100, aes(percent)) +
          geom_histogram(aes(y=..density..), binwidth=0.02,
                         color='black', fill='white') +
          geom_vline(aes(xintercept=mean(percent, na.rm=T)),
                     color="black", linetype="dashed", size=0.5) +
          geom_density(alpha=0.2)
```

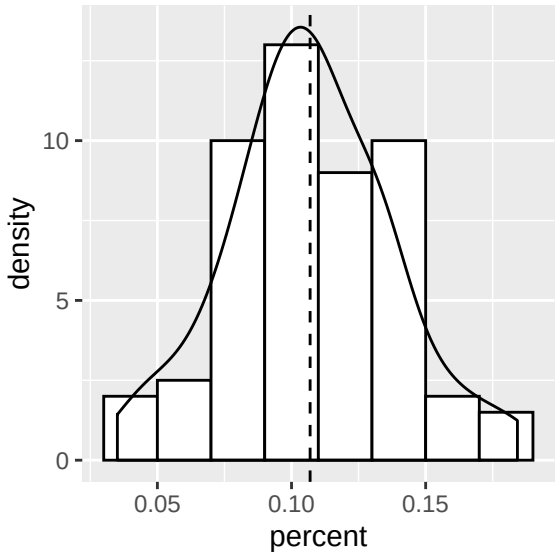


Figure 1: Percentage of Damaged Printers per Trial (100 Visits)

5. *For each of the 100 visits, make a 95% confidence interval for the percentage of damaged printers you count. How many of these confidence intervals contain the true percentage of* 0.10*? Is this close to what you expected?*

   (a) We can create a function that logically indexes our data frame in order to determine which generated confidence intervals contain our true percentage, and then determine the length.

```
in_interval <- function(m, n) {
    clower <- m - abs(z * sqrt((m * (1 - m)) / n))
    cupper <- m + abs(z * sqrt((m * (1 - m)) / n))
    (clower < 0.10) & (0.10 < cupper)
}
count100 <- length(interval_data100$percent[
                in_interval(interval_data100$percent, n)])
```

   This results in 91 intervals that contain the true percentage. We expected 95 of them, but this is close enough.[1] Note, this goes back to the definition of a 95% confidence interval as we can see that 91/100 contained the true percentage. This is as it should be. Since the data is randomly generated however, we won't always get exactly 95 for this number.

[1] Note, this count is regenerated every time this paper is recompiled. If my above statement makes no sense, please forgive the grammatically insensitive random data generator.

6. *Repeat parts (4) and (5) with 500 visits and 1000 visits. What do you notice about your histogram? What theory are you observing? What do you notice about the number of confidence intervals that contain the true percentage?*

   (a) First, we establish our data using the `rbinom` command, and then we create a data frame with the given data. This data frame's first column is the number of damaged printers counted, and the second column is the percentage. Next we count how many confidence intervals generated using the sample percentage contain the true percentage.

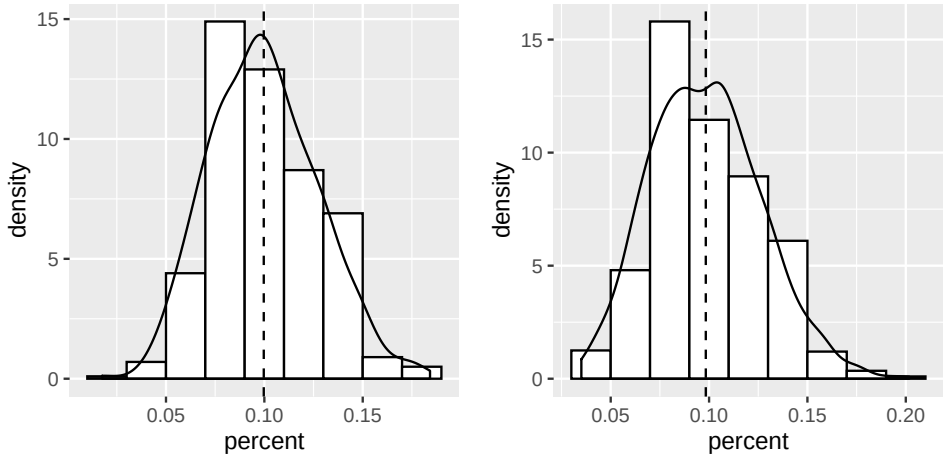
```
# Establish Data
db500 <- rbinom(500, n, 0.10)
db1000 <- rbinom(1000, n, 0.10)

# Create Data Frames
interval_data500 <- data.frame(printers=db500, percent=db500/n)
interval_data1000 <- data.frame(printers=db1000, percent=db1000/n)

# Plot our distributions
ggplot(interval_data500, aes(percent)) +
          geom_histogram(aes(y=..density..), binwidth=0.02,
                         color='black', fill='white') +
          geom_vline(aes(xintercept=mean(percent, na.rm=T)),
                     color="black", linetype="dashed", size=0.5) +
          geom_density(alpha=0.2)

ggplot(interval_data1000, aes(percent)) +
          geom_histogram(aes(y=..density..), binwidth=0.02,
                         color='black', fill='white') +
          geom_vline(aes(xintercept=mean(percent, na.rm=T)),
                     color="black", linetype="dashed", size=0.5) +
          geom_density(alpha=0.2)

# Count instances of percentage within interval
count500 <- length(interval_data500$percent[
                   in_interval(interval_data500$percent, n)])
count1000 <- length(interval_data1000$percent[
                in_interval(interval_data1000$percent, n)])
```



(a) 500 Visits (b) 1000 Visits

Figure 2: Percentage of Damaged Printers per Trial

In Figure 2a we see that not only is the histogram density becoming more and more smooth, but that the Central Limit Theorem is holding. The Central Limit Theorem roughly states that expected value of our estimators will approach the true value as the amount of data grows. If we examine our count, we see that it is 0.942, which is also close to 95%[2]. We can now move to 1000 visits.

In Figure 2b we see that our histogram is a very accurate model of the situation with again, the Central Limit Theorem as to the reason why. If we now examine our count, we see that it is 0.934, which is also close to 95%.[3]

[2]Ibid.
[3]Ibid.

7. *Make a histogram of the p-values for the 1000 visits, testing if the manufacturer's claim is true. What does the distribution of your p-values look like?*

   (a) In this case we assume our null hypothesis is that the percent of damaged printers is 10%. We can examine each generated element to determine whether or not our null hypothesis is valid. In this case our test is two-tailed, as the data could be either too high, or too low, and we need to just test that our sample percentage is not equal to our true percentage. In other words, we need to test

$$H_0 \neq H_a$$

```
        z_test <- (p_hat - interval_data1000$percent)/(sqrt((
                 interval_data1000$percent *
                 (1 - interval_data1000$percent))/n))
        interval_data1000p <- data.frame(printers=interval_data1000$printers,
                                         percent=interval_data1000$percent,
                                         pvalues=2 * (1 - pnorm(abs(z_test))))
        ggplot(interval_data1000p, aes(pvalues)) +
                   scale_x_continuous(limits = c(0, 1)) +
                   geom_histogram(aes(y=..density..), binwidth=0.1,
                                  color='black', fill='white') +
                   geom_vline(aes(xintercept=0.1),
                              color="red", linetype="solid", size=0.5) +
                   geom_vline(aes(xintercept=0.05),
                              color="red", linetype="solid", size=0.5) +
                   geom_vline(aes(xintercept=0.01),
                              color="red", linetype="solid", size=0.5)

## Warning:  Removed 2 rows containing missing values (geom_bar).
```

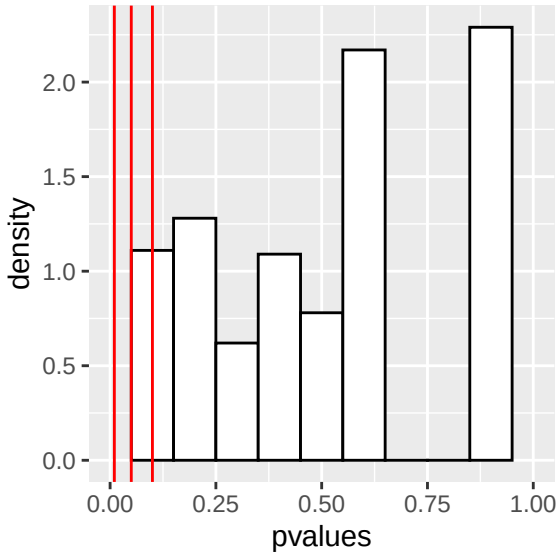


Figure 3: Distribution of our P-Values

In Figure 3 we mark $p = 0.01, p = 0.05$, and $p = 0.1$ to indicate rejection zones. These p-values are distributed fairly uniformly, and as a result this graph does not tell us much about whther or not our null hypothesis is valid; all it tells us is that the null *might* be invalid. Based on our other parts we've previously calculated however, we know that the manufacturer's claim is most likely true.

8. *Repeat parts (4)-(7), but simulate the number of damaged printers assuming a* $Bin(n, 0.25)$ *distribution. As the inspector, you do not know if the manufacturer is telling the truth, but you are a great statistician who knows how to analyze data properly. After examining your histogram, the confidence intervals, and the p-values, what would you conclude about the veracity of the manufacturer's claim? And, more importantly, Why?*

   (a) First we can simulate the data using the new distribution in the same manner as part (4). Please reference Figure 4. Note the mean-line remains, but is markedly different from before due to the new distribution.

```
db100_new <- rbinom(100, n, 0.25)
interval_data100_new <- data.frame(printers=db100_new, percent=db100_new/n)
ggplot(interval_data100_new, aes(percent)) +
          geom_histogram(aes(y=..density..), binwidth=0.02,
                         color='black', fill='white') +
          geom_vline(aes(xintercept=mean(percent, na.rm=T)),
                     color="black", linetype="dashed", size=0.5) +
          geom_density(alpha=0.2)
```

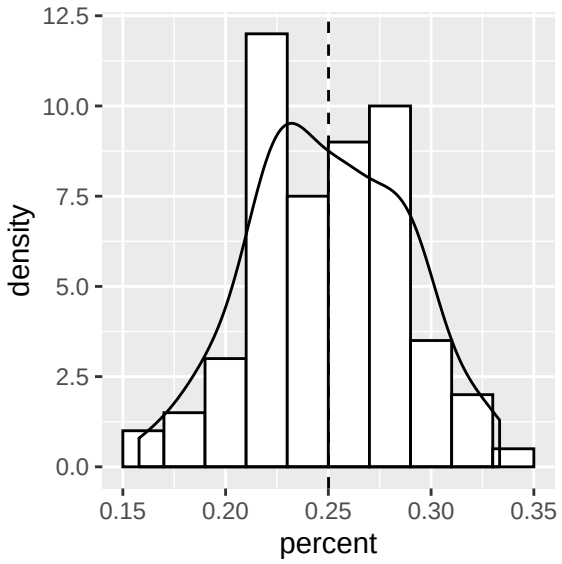


Figure 4: Percentage of Damaged Printers per Trial (100 Visits)

Now we can re-use our function from above to determine how many sample percentages are within the sample's calculated confidence interval.

```
count100_new <- length(interval_data100_new$percent[
                in_interval(interval_data100_new$percent, n)])
```

And we see that the percent of confidence intervals that contain our true percentage under the null hypothesis is 0.02. We can now extend the experiment using 500 and 1000 visits as before.

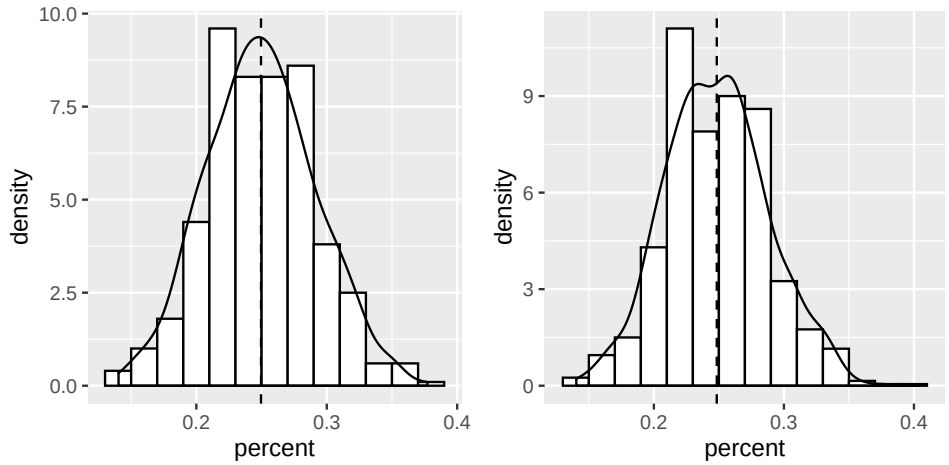
```
# Establish Data
db500_new <- rbinom(500, n, 0.25)
db1000_new <- rbinom(1000, n, 0.25)

# Create Data Frames
interval_data500_new <- data.frame(printers=db500_new,
                                   percent=db500_new/n)
interval_data1000_new <- data.frame(printers=db1000_new,
                                    percent=db1000_new/n)

# Plot our distributions
ggplot(interval_data500_new, aes(percent)) +
           geom_histogram(aes(y=..density..), binwidth=0.02,
                          color='black', fill='white') +
           geom_vline(aes(xintercept=mean(percent, na.rm=T)),
                      color="black", linetype="dashed", size=0.5) +
           geom_density(alpha=0.2)

ggplot(interval_data1000_new, aes(percent)) +
           geom_histogram(aes(y=..density..), binwidth=0.02,
                          color='black', fill='white') +
           geom_vline(aes(xintercept=mean(percent, na.rm=T)),
                      color="black", linetype="dashed", size=0.5) +
           geom_density(alpha=0.2)

# Count instances of percentage within interval
count500_new <- length(interval_data500_new$percent[
                    in_interval(interval_data500_new$percent, n)])
count1000_new <- length(interval_data1000_new$percent[
                  in_interval(interval_data1000_new$percent, n)])
```



(a) 500 Visits (b) 1000 Visits

Figure 5: Percentage of Damaged Printers per Trial

Repeating with 500 and 1000 visits respectively, we obtain 0.028 and 0.024 percent of sample confidence intervals that contain our true percentage of 0.10.

We can now look at the p-values.

```
z_test <- (p_hat - interval_data1000_new$percent)/(sqrt((
          interval_data1000_new$percent *
          (1 - interval_data1000_new$percent))/n))
interval_data1000p_new <- data.frame(
                          printers=interval_data1000_new$printers,
                          percent=interval_data1000_new$percent,
                          pvalues=2 * (1 - pnorm(abs(z_test))))
ggplot(interval_data1000p_new, aes(pvalues)) +
            scale_x_continuous(limits = c(0, 0.11)) +
            geom_histogram(aes(y=..density..), binwidth=0.02,
                           color='black', fill='white') +
            geom_vline(aes(xintercept=0.1),
                       color="red", linetype="solid", size=0.5) +
            geom_vline(aes(xintercept=0.05),
                       color="red", linetype="solid", size=0.5) +
            geom_vline(aes(xintercept=0.01),
                       color="red", linetype="solid", size=0.5)
```

```
## Warning:  Removed 5 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```

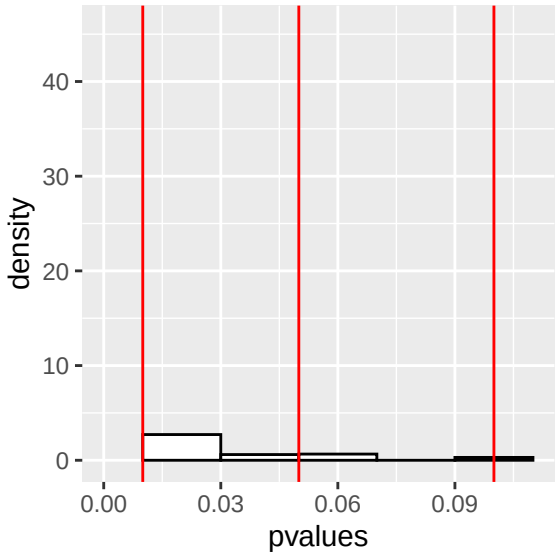


Figure 6: Distribution of our P-Values

After all the analysis is finished we can conclude that the printer manufacturer is incorrect in assuming that only 10% of their printers are damaged. This conclusion is established from several different points. First, we can see just by plotting a histogram of the data that the average percent of damaged printers is about 25%, which is a huge mark against the null hypothesis that our percentage is equal to 0.10.

Second, if we take a look at how many of our calculated confidence intervals contain our true mean, we see that this number is effectively zero, meaning that by definition we no longer have a 95% confidence interval established, and something is incorrect.

Lastly, by examining our p-values we see that almost every single value is less than 0.10, and most are below 0.01, which idicates a very strong presumption against our null hypothesis.

Based on all the evidence we should reject our null hypothesis in favor of something else. Based on the experimental data it would not be incorrect to assume a contending null hypothesis to be 0.25, which is indeed correct.