

Problem Set 5

Mashhood Syed

There are 8 questions. Please answer all of them in this Rmd file. Submit both the Rmd file and the corresponding PDF. Change the name above from “You” to your name.

1. After conducting an experiment, Samantha finds her p-value is 3.4%. What does that mean?

This means the probability of obtaining a sample mean, given that the value in the null hypothesis is true, is 3.4%. When the p-value is $< .05$ (5%), we reject the null hypothesis. Therefore, Samantha will reject the null hypothesis of her experiment. According to Fisher, the value is statistically significant and would warrant further investigation.

2. After comparing the means of two models, Gregory finds the 95% confidence interval is (34.51, 98.57). What does that mean?

This means that Gregory can be 95% certain that the true population mean is between 34.51 and 98.57.

We discussed 4 general types of problems. In the remaining exercises, you can reference the problems by letter.

A. Test of Population Mean B. Test of Two Sample Means C. Test of Population Proportion D. Test of Two Sample Proportions

3. Adam heard that projects using WhizBang take about 2.3 weeks on average to complete. He randomly sampled 27 projects using WhizBang with the number of weeks listed below.

1. What kind of problem is this? A
2. Calculate Fisher’s p-value and the 95% confidence interval for this data. Is the result statistically significant?

p-value is .8150435 <- The result is statistically significant (less than 5%) 95% conf. int is (2.24, 2.34)
<- No the result is not statistically significant. The status quo lies in the conf. interval.

4. What else do we need to know in order to determine if the result has practical significance?

Maybe some domain knowledge about WhizBang (like what is it maybe :)).

```
whizbang <- list()
whizbang$data <- c( 2.73, 2.39, 2.29, 2.21, 2.17, 2.52, 2.42, 2.43, 2.21, 2.29, 2.14, 2.30, 2.25, 2.19,
whizbang$mu = 2.3
whizbang$n = length(whizbang$data)
whizbang$xbar = mean(whizbang$data)
whizbang$sd = sd(whizbang$data)
whizbang
```

```
## $data
## [1] 2.73 2.39 2.29 2.21 2.17 2.52 2.42 2.43 2.21 2.29 2.14 2.30 2.25 2.19
## [15] 2.47 2.20 2.16 2.23 2.23 2.28 2.22 2.30 2.19 2.27 2.34 2.29 2.22
##
## $mu
## [1] 2.3
```

```
##
## $n
## [1] 27
##
## $xbar
## [1] 2.294
##
## $sd
## [1] 0.1303
```

```
whizbang$t = (whizbang$xbar - whizbang$mu)/(whizbang$sd/sqrt(whizbang$n))
whizbang$t
```

```
## [1] -0.2363
```

```
whizbang$p.value = 2 * pt(-abs(whizbang$t), df = whizbang$n - 1)
whizbang$p.value
```

```
## [1] 0.815
```

```
whizbang$se = whizbang$sd / sqrt(whizbang$n)
whizbang$ci = c(whizbang$xbar - 1.96 * whizbang$se, whizbang$xbar + 1.96 * whizbang$se)
whizbang$ci
```

```
## [1] 2.245 2.343
```

4. Sarah thinks she's found the cure for the 48 hour flu—or at least a way to shorten it. In order to test her treatment, she has two random samples of individuals who have just been diagnosed with the 48 hour flu. One group she gives chicken soup (the traditional treatment) and the other group she gives her cure, Jameson's Irish Whiskey. She then observes the number of hours for each group until symptoms abate.

1. What kind of problem is this?

B

2. Calculate all values for this problem using The Modern Synthesis. Start with Neyman-Pearson then calculate the p-value and confidence interval. Is the result statistically significant?

Neyman-Pearson: Welch Two Sample t-test data: *fludata.1* and *fludata.2* $t = 6.307$, $df = 32.633$, $p\text{-value} = 4.113e-07$ alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: 3.029552 5.916640 sample estimates: mean of x mean of y: 46.94941 42.47632

The result is statistically significant. based on a p-value of .0000004113

p-value: 3.44888e-07 95% conf. interval: (3.033372, 5.912820)

Since the interval does not include zero we can say that this result is statistically significant.

3. What else do we need to know in order to determine if the result has practical significance?

Domain knowledge


```
## $data
## [1] 1 1 1 1 0 0 1 1 1 0 1 1 1 0 1 1 1 0 0 0 1 0 0 1 0 0 1 0 1 1 0 1 1 1
## [36] 1 0 1 1 1 1 1 1 1 1
##
## $theta
## [1] 0.69
##
## $n
## [1] 45
##
## $p
## [1] 0.6889
##
## $se
## [1] 0.06901
```

```
## [1] -0.0161
```

```
## [1] 0.9872
```

1. What kind of problem is this?

2. Frame the inference problem using Fisher's Point and Interval approaches to statistical significance. Was the result statistically significant?

3. What else would we need to know in order to determine if the result is practically significant?

```
insomnia <- list()
insomnia$data.1 <- c(1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1)
insomnia$data.2 <- c(0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1)

insomnia$n.1 = length(insomnia$data.1)
insomnia$n.2 = length(insomnia$data.2)

insomnia
```

```
## $data.1
## [1] 1 0 0 0 0 0 1 1 0 1 0 0 1 1 1 1 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 1 1 0
## [36] 0 0 1 1 0 0 1 1 1 1 1 1 0 1 1 1 0 0 1 0 1
##
## $data.2
## [1] 0 1 1 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0
## [36] 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0
##
## $n.1
## [1] 56
##
## $n.2
## [1] 53
```

```
insomnia$p.1 = mean(insomnia$data.1)
insomnia$p.2 = mean(insomnia$data.2)

insomnia$p = (insomnia$p.1 * insomnia$n.1 + insomnia$p.2 * insomnia$n.2)/(insomnia$n.1 + insomnia$n.2)
insomnia$se <- insomnia$p * (1 - insomnia$p) * ((1/insomnia$n.1) + (1/insomnia$n.2))
insomnia$d <- insomnia$p.1 - insomnia$p.2
insomnia$z.score <- insomnia$d/insomnia$se
insomnia
```

```
## $data.1
## [1] 1 0 0 0 0 0 1 1 0 1 0 0 1 1 1 1 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 1 1 0
## [36] 0 0 1 1 0 0 1 1 1 1 1 1 0 1 1 1 0 0 1 0 1
##
## $data.2
## [1] 0 1 1 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0
## [36] 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0
##
## $n.1
## [1] 56
##
## $n.2
## [1] 53
##
## $p.1
## [1] 0.5893
##
## $p.2
## [1] 0.2075
##
## $p
## [1] 0.4037
##
## $se
## [1] 0.00884
##
## $d
## [1] 0.3817
##
## $z.score
## [1] 43.18
```

```
insomnia$p.value <- 2 * pnorm(-abs(insomnia$z.score))
insomnia$p.value
```

```
## [1] 0
```

```
insomnia$ci <- c(insomnia$d - 1.96 * insomnia$se, insomnia$d + 1.96 * insomnia$se)
insomnia$ci
```

```
## [1] 0.3644 0.3991
```

7. Redo question #3 using the Bootstrap. Provide both Frequentist and Bayesian interpretations. Use the same data structure for your results but add the “bootstrap” qualifier as we did in class.

```
sample(whizbang$data, whizbang$n, replace = TRUE)
```

```
## [1] 2.52 2.42 2.43 2.23 2.16 2.21 2.29 2.14 2.19 2.21 2.28 2.47 2.28 2.34
## [15] 2.14 2.73 2.29 2.19 2.29 2.34 2.16 2.21 2.29 2.21 2.23 2.29 2.14
```

```
t = rep(1, 1000)
```

```
whizbang$bootstrap.samples <- lapply(t, function(x) sample(whizbang$data,
whizbang$n, replace = TRUE))
```

```
whizbang$bootstrap.means <- unlist(lapply(whizbang$bootstrap.samples,
mean))
```

```
min(whizbang$bootstrap.means)
```

```
## [1] 2.235
```

```
max(whizbang$bootstrap.means)
```

```
## [1] 2.381
```

```
mean(whizbang$bootstrap.means)
```

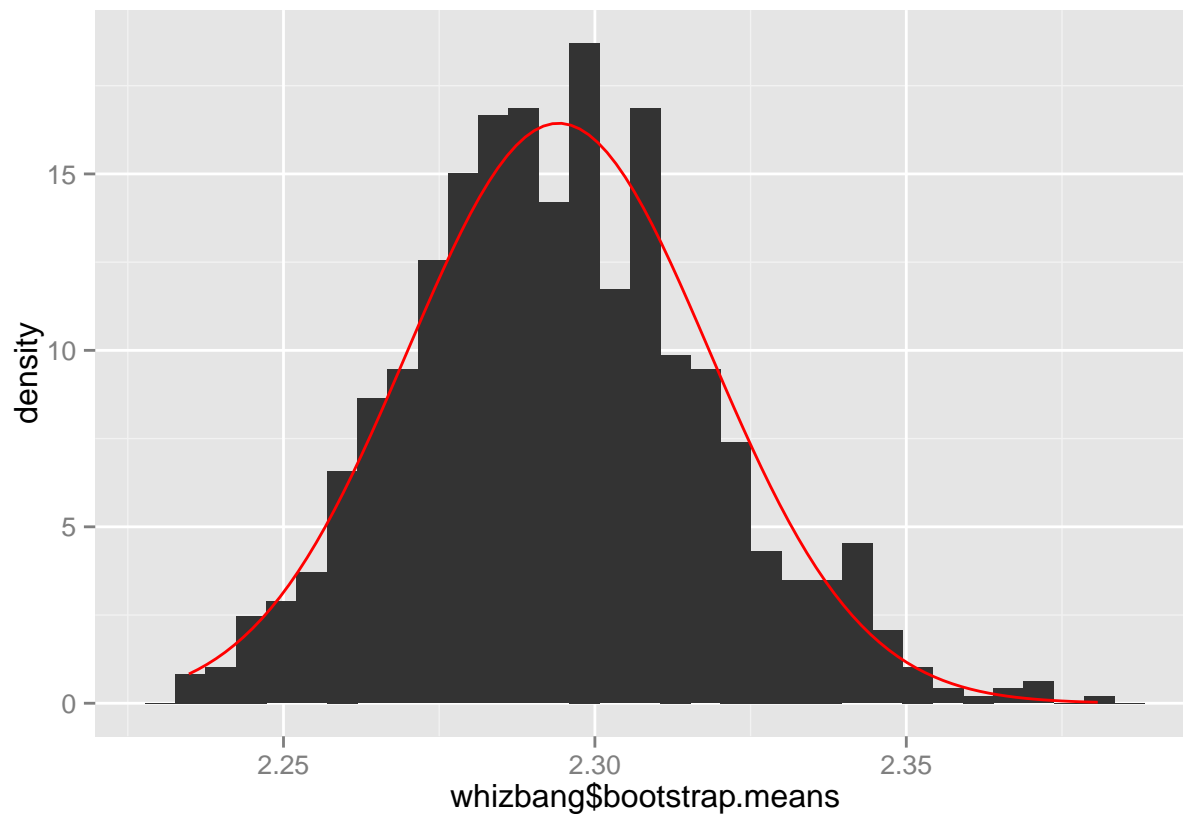
```
## [1] 2.294
```

```
sd(whizbang$bootstrap.means)
```

```
## [1] 0.02427
```

```
library(ggplot2)
ggplot(as.data.frame(whizbang$bootstrap.means), aes(x = whizbang$bootstrap.means)) +
geom_histogram(aes(y = ..density..)) + stat_function(fun = dnorm, args = list(mean = mean(whizbang$boot
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
whizbang$bootstrap.mean <- mean(whizbang$bootstrap.means)
whizbang$bootstrap.sd <- sd(whizbang$bootstrap.means)

whizbang$bootstrap.ci <- c(whizbang$bootstrap.mean - 2.31 *
whizbang$bootstrap.sd, whizbang$bootstrap.mean + 2.31 *
whizbang$bootstrap.sd)

whizbang$ci
```

```
## [1] 2.245 2.343
```

```
whizbang$bootstrap.ci
```

```
## [1] 2.238 2.350
```

```
quantile(whizbang$bootstrap.means, c(0.025, 0.975))
```

```
## 2.5% 97.5%
## 2.249 2.344
```

8. Redo question #6 using the Bootstrap. Provide both Frequentist and Bayesian interpretations. For the Bayesian interpretation discuss whether or not an uninformative prior is justified. Use the same data structure for your results but add the “bootstrap” qualifier as we did in class.

The use of prior probability is based on having a set of hypotheses where the prior probability of each is the same (p.15). In this problem, there is no evidence of having a set of hypothesis from previous experiments. So in this case I don't think it's justified.


```

insomnia$bootstrap.samples.1 <- lapply(t, function(x) sample(insomnia$data.1, insomnia$n.1, replace = T))
insomnia$bootstrap.samples.2 <- lapply(t, function(x) sample(insomnia$data.2, insomnia$n.2, replace = T))
insomnia$bootstrap.means.1 <- unlist(lapply(insomnia$bootstrap.samples.1, mean))
insomnia$bootstrap.means.2 <- unlist(lapply(insomnia$bootstrap.samples.2, mean))
insomnia$bootstrap.d <- insomnia$bootstrap.means.1 - insomnia$bootstrap.means.2

```

```
min(insomnia$bootstrap.d)
```

```
## [1] 0.1068
```

```
max(insomnia$bootstrap.d)
```

```
## [1] 0.6378
```

```
mean(insomnia$bootstrap.d)
```

```
## [1] 0.381
```

```
sd(insomnia$bootstrap.d)
```

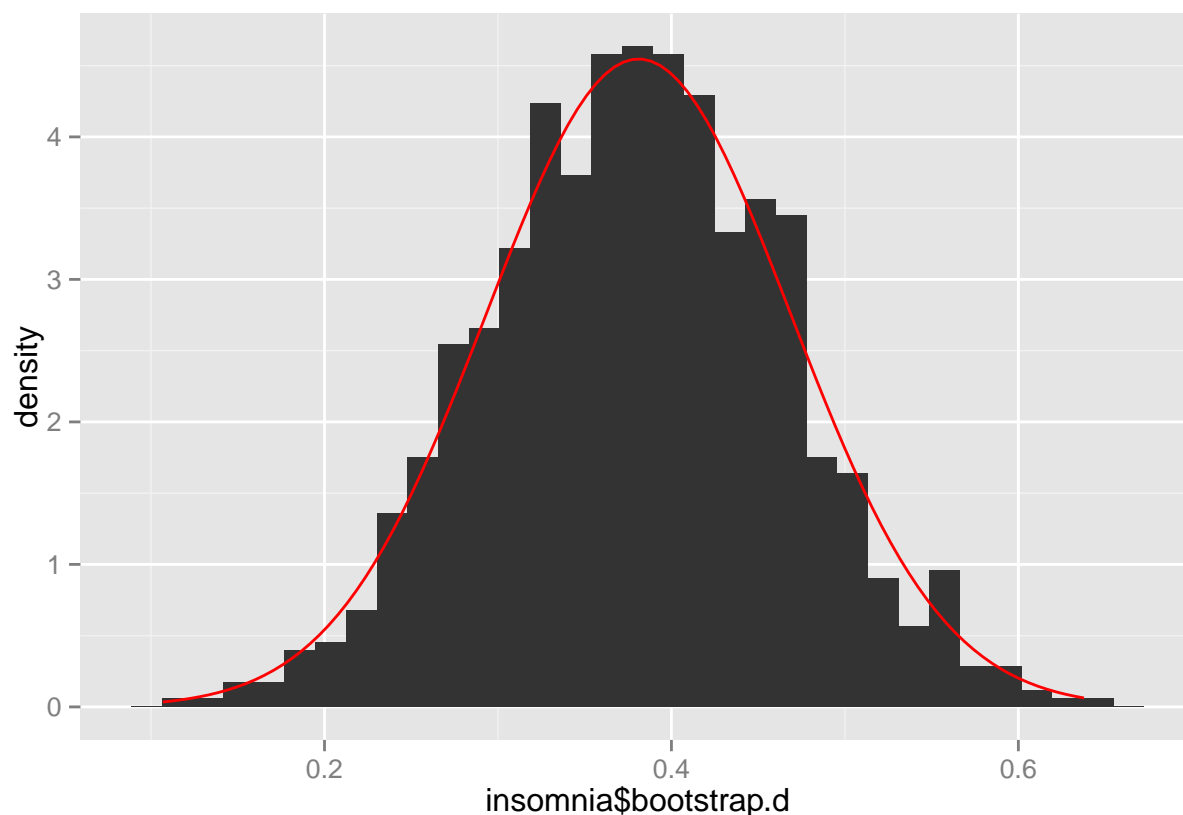
```
## [1] 0.08774
```

```

ggplot(as.data.frame(insomnia$bootstrap.d), aes(x = insomnia$bootstrap.d)) +
  geom_histogram(aes(y = ..density..)) + stat_function(fun = dnorm, args = list(mean = mean(insomnia$bootstrap.d),
  sd = sd(insomnia$bootstrap.d)), color = "red")

```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
insomnia$bootstrap.mean <- mean(insomnia$bootstrap.d)
insomnia$bootstrap.mean
```

```
## [1] 0.381
```

```
insomnia$bootstrap.sd <- sd(insomnia$bootstrap.d)
insomnia$bootstrap.sd
```

```
## [1] 0.08774
```

```
insomnia$bootstrap.ci <- c(insomnia$bootstrap.mean - 1.96 * insomnia$bootstrap.sd, insomnia$bootstrap.m
1.96 * insomnia$bootstrap.sd)
```

```
insomnia$bootstrap.ci
```

```
## [1] 0.209 0.553
```

```
insomnia$ci
```

```
## [1] 0.3644 0.3991
```

```
quantile(insomnia$bootstrap.d, c(0.025, 0.975))
```

```
## 2.5% 97.5%
```

```
## 0.2170 0.5624
```