HW5_110078509

110078509 & this guy help me: 110078506, 110078501

20220319

Question 1

i. Would this scenario create systematic or random error (or both or neither)?

ii. Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?

iii. Will it increase or decrease our power to reject the null hypothesis?

iv. Which kind of error (Type I or Type II) becomes more likely because of this scenario?

a. *only collected data from a pool of young consumers, and missed many older customers who you suspect might use the product much less every day.*

i.*systematic error*

ii.*sd, diff are affected.*

Explains:

Because we missed many older customers. Therefore, the alternative distribution we obtained is too narrow compared to the real one. Hence, the sd is affected.

The total nnumber of sample size (n) is not affected. However, we sample incorrectly from a group which is not i.i.d to the popuation we want to estimate. This matters.

The diff is (x mean -μo), in this scenario, the x means is not accurate anymore, because we missed plenty of samples lower than the average. Hence the diff is affected also. And it supposed to increase the diff.

iii.*Increase the power*

Explains:

This misleading sample error results in decreasing of std, which increase the power. Due to the neglecting of many older customers, we overestimate the average usage

time of the device, which made the alt distribution shifted rightward and increase the power.

Explains:

It also called false Positive. As we missed many elder customers who you suspect might use the product much less every day, which caused H0 much easier to be rejected. However, if we considered this group in our sampling, it supposed to be harder to reject the H0. As the definition of false positive, the so called " false positive" is represented the rejection of the H0 as the the H0 is real. Hence, it type I error.

---

b.Find that 20 of the respondents are reporting data from the wrong wearable device, so they should be removed from the data.

---

i. random error

---

ii. n would be lower.

Explains:

Because these noisy data shall not be considered in.

---

iii. It decrease the power to reject the null hypothesis

Reason:

Manipulate the sliding window to figure out.

---

iv. Type II Error

Explains: Because the power$(1-\beta)$ is decrease c, therefore, the type II error (β) is rising.

---

c. A very annoying professor visiting your company has criticized your colleague's "95% confidence" criteria, and has suggested relaxing it to just 90%.

Explains:

It just the change of the confidence level based on the professor's suggestion.

---

ii. *alpha*

Explains:

From 0.05 to 0.1

---

iii. *Increase the power to reject the null hypothesis.*

Explains:

Because the critical point of the right tail shifted leftward.

---

iv. *Type I error*

Explains: Type 1 errors has a probability of "α" correlated to the level of confidence we set. Originally, 95% confidence level means that there is a 5% chance of getting a type I error. According to the suggestion, we set the confidence level to 90% , the chance of getting a type 1 error increase (10%).

---

d. Your colleague has measured usage times on five weekdays and taken a daily average. But you feel this will underreport usage for younger people who are very active on weekends, whereas it over-reports usage of older users.

---

i. *systematic error*

---

ii. *diff , sd will be affected.*

Explains:

Because the population means we want to inference includes the user behavior from Monday to Sunday. If we only choose the workday data as our sample, we will overemphasize the behavior of elder users and neglect the younger people who are very active on weekends. Therefore, the mean sample mean is underestimated. According to the formula of diff(sample mean - mu 0), as the mean is affected and the mu0 is still, the diff would be affected for sure.

And for the part of standard deviation, we obtained the narrower std compared to the unbaised sampling result. Therefore, sd would be affected.

---

iii. *Depends.*

Explains:

H0 :the mean usage time of the new smartwatch is the same or less than for the previous smartwatch.

H alt :The mean usage time is greater than that of our previous smartwatch.

Because we skipped the frequent user in the weekend, which made the sample mean lower than the hypothesis mu.

The increasing of diff increase the power.

However, the increasing of sd, decrease the power.

Therefore, it depends.

---

iv. *Type II*

Explains: Because we should count weekend users in, but we don't. And false negative means that we should consider specific targets is negative, but it doesn't.

Question 2

Imagine this time that Verizon claims that they take no more than 7.6 minutes on average (single-tail test)

Import data

```
verizon <- read.csv("verizon.csv")
time <- verizon$Time
```

a. Recreate the traditional hypothesis test of last week using high-level built-in functions of R:

i. Use the t.test() function to conduct a one-sample, one-tailed t-test:

report 99% confidence interval of the mean, t-value, and p-value Use the power.t.test() function to tell us the power of the test

Right- tail testing

```
t.test(time, mu=7.6, alternative="greater", conf.level=0.99)
```

```
##
##   One Sample t-test
##
## data:   time
## t = 2.5608, df = 1686, p-value = 0.005265
## alternative hypothesis: true mean is greater than 7.6
## 99 percent confidence interval:
##   7.683604         Inf
## sample estimates:
## mean of x
##   8.522009
```

---

ii. Use the power.t.test() function to tell us the power of the test

```
#it's 99% one-sided
hyp_mean <- 7.6
power.t.test(n = length(time),
             delta = mean(time) - hyp_mean,
             sig.level = 0.01,
             sd = sd(time),
             alternative = "one.sided")
```

```
##
##        Two-sample t test power calculation
##
##                n = 1687
##            delta = 0.9220095
##               sd = 14.78848
##        sig.level = 0.01
##            power = 0.3028078
##      alternative = one.sided
##
## NOTE: n is number in *each* group
```

Ans:: power = 0. 3028078

---

b. Let's use bootstrapped hypothesis testing to re-examine this problem:

i. Retrieve the original t-value from traditional methods (above)
```
sample_se <- sd(time) / sqrt(length(time))
t_value <- (mean(time) - hyp_mean) / sample_se
t_value
```

```
## [1] 2.560762
```

Ans:

---

ii. Bootstrap the null and alternative t-distributions

```r
bootstrap_null_alt <- function(sample0, hyp_mean) {
resample <- sample(sample0, length(sample0), replace=TRUE)
resample_se <- sd(resample) / sqrt(length(resample))

t_stat_alt <- (mean(resample) - hyp_mean) / resample_se
t_stat_null <- (mean(resample) - mean(sample0)) / resample_se
c(t_stat_alt, t_stat_null)
}

set.seed(42)
boot_stats <- replicate(20000, bootstrap_null_alt(time, hyp_mean))

boot_t_stats <- replicate(10000,bootstrap_null_alt(time, hyp_mean))

t_alt <- boot_t_stats[1,]
t_null <- boot_t_stats[2,]


#Making the plot here
plot(density(t_alt),  xlim=c(-4,5), col = 'red', lwd = 1, main = "The n
ull and alternative t-distributions")
lines(density(t_null), lty = "solid", col = "black")
null_text <- 't_null'
text(x = 0, y =0.365 , null_text, col = "black", cex = 1)
alt_text <- 't_alt'
text(x = 2.6, y =0.39 , alt_text, col = "red", cex = 1)
```

The null and alternative t-distributions

N = 10000   Bandwidth = 0.1248

---
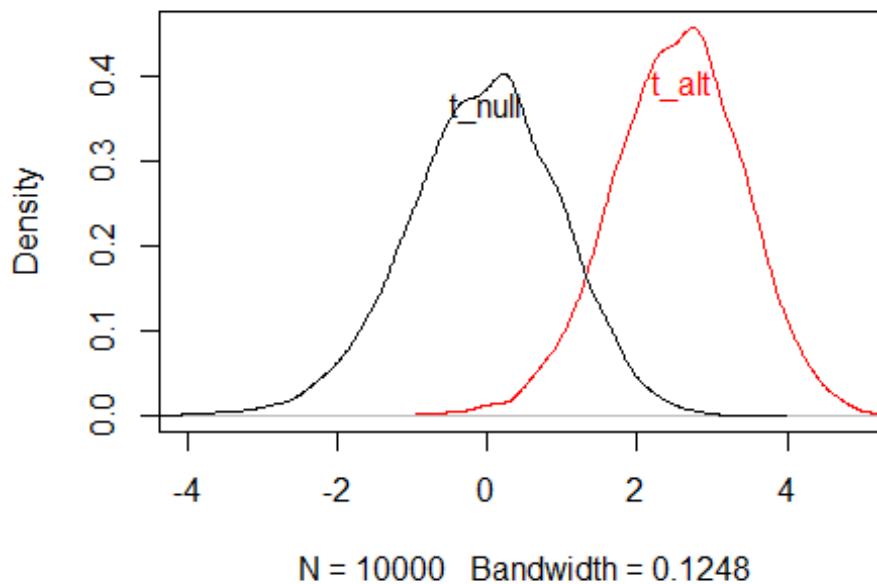
iii. Find the 99% cutoff value for critical null values of t (from the bootstrapped null); What should our test conclude when comparing the original t-value to the 99% cutoff value?

```
ci_99 <- quantile(t_null, probs=0.99) ; ci_99
```

```
##       99%
## 2.148639
```

Because it's single tail testing, 99% critical of the t distribution is 2.148639.

```
plot(density(t_alt),  xlim=c(-4,5), col = 'red', lwd = 1, main = "The n
ull and alternative t-distributions With 99% cutoff value  ")
lines(density(t_null), lty = "solid", col = "black")

# Labeled it
null_text <- 't_null'
text(x = 0, y =0.365 , null_text, col = "black", cex = 1)
alt_text <- 't_alt'
text(x = 3, y =0.39 , alt_text, col = "red", cex = 1)

abline(v=t_value, col = "blue") # 2.560762
text(x = 3.25, y =0.05 , 't_value', col = "blue", cex = 1)
abline(v=ci_99, lty ="dotted")
text(x = 1.65, y =.0025 , '99% cutoff value', col = "black", cex = 1)
```
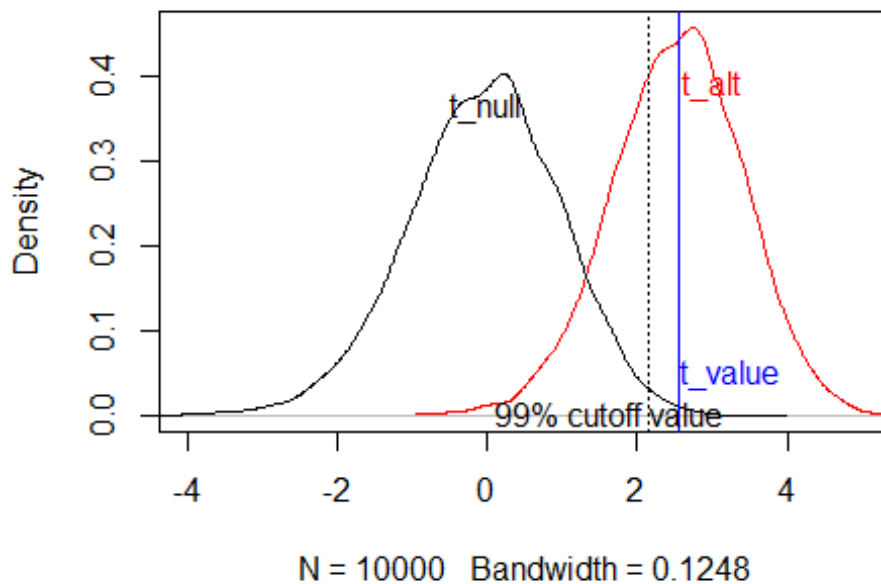
e null and alternative t-distributions With 99% cutoff

N = 10000   Bandwidth = 0.1248

It seems that our t value(2.560762) is higher than the 99% cutoff value (2.148639). Therefore, we should reject H0. (I made a plot for more clear explanation)

---

iv.   Compute the p-value and power of our bootstrapped test

```
null_probs <- ecdf(t_null)
one_tailed_p_value <- 1-null_probs(t_value);
one_tailed_p_value
```

```
## [1] 0.0024
```

```
# [1] 0.0024 -> right-tailed
```

For the part of p value, right-tailed: 0.0024

```
alt_probs <- ecdf(t_alt)
 (1 - alt_probs(ci_99))
```

```
## [1] 0.6776
```

(right)single-tailed power: 0.6776