Homework 3

Your Name

Due: November 12, 2019 at 11:59pm

Homework Policies:

You are encouraged to discuss problem sets with your fellow students (and with the Course Instructor of course), but you must write your own final answers, in your own words. Solutions prepared "in committee" or by copying someone else's paper are not acceptable. This violates the Brown standards of plagiarism, and you will not have the benefit of having thought about and worked the problem when you take the examinations.

All answers must be in complete sentences and all graphs must be properly labeled.

In this homework you will be required to use .Rmd to do it., you can then knit to a word document of PDF to turn it in.

For the PDF Version of this assignment: PDF

For the R Markdown Version of this assignment: RMarkdown

Turning the Homework in:

Please turn the homework in through canvas. You may use a pdf, html or word doc file to turn the assignment in.

PHP 1510 Assignment Link

PHP 2510 Assignment Link

Central Limit Theorem and Confidence Intervals.

1. (3 points) What does the Law of Large Numbers tell us?

The Law of Large numbers tells us that as our sample size increases, our sample mean converges (becomes) the true population mean. This shows us that if we have a large enough sample size, we believe that our estimate is the truth or close to it.

2. (3 points) What does the Central Limit Theorem tell us?

The central limit theorem discusses what happens we we run a study over and over again. Each time we draw a sample, we calculate a mean. These means then follow a normal distribution.

3. (4 points) How do the law of large numbers and the Central Limit Theorem work together?

If we have a large sample size, our estimate is close to the true population mean. What happens when we sample over and over again with a large sample size is that because each of these is a better estimate of the mean, they tend to have less variation. So as your sample size increases your normal distribution of means shrinks with a smaller variance.

4. (6 points) What are 3 reasons we would want to create a confidence interval?

*** We may wish to create a confidence interval because: - We know a point estimate is wrong and wish to use an interval estimate. - We desire to understand the amount of error in our estimate. - We wish to test a hypothesis based on the confidence interval. ***

5. (3 points) How do we know whether or not we can get our critical value in a confidence interval from the Z distribution of a t distribution.

We use the z disribution if we know the population variance. If we do not know the population variance we use the t distribution.

6. Generate 1000 random values from a $N(2, 4)$ distribution.
   a. (4 points) Create and interpret a 90% confidence interval using the Z distribution.

```
set.seed(123)
x <- rnorm(1000, 2, 4)
mn <- mean(x)
z <- qnorm(0.975)
std.dev <- sd(x)
std.error <- z*std.dev/sqrt(1000)

low = mn - std.error
high = mn + std.error
low
```

```
## [1] 1.818652
```

```
high
```

```
## [1] 2.310371
```

   b. (4 points) Create and interpret a 90% confidence interval using the t distribution.

```
set.seed(123)
x <- rnorm(1000, 2, 4)
mn <- mean(x)
t <- qt(0.975, df=999)
std.dev <- sd(x)
std.error <- t*std.dev/sqrt(1000)

low = mn - std.error
high = mn + std.error
low
```

```
## [1] 1.818354
```

```
high
```

```
## [1] 2.310669
```

   c. (2 points) How do these compare?

There is very little difference from these. This is due to the fact that with a sample size of 1000 our degrees of freedom are 999, this means our critical value in the t is almost identical to the one in the z.

Data Problems

We will work with the same data in which we used for lectures 13 and 15 as well as your midterm.

7. (2 points) What is the mean number of poor mental health days (`menthlth`)?

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
brfss2 %>%
  summarise(mean(menthlth, na.rm=T))
```

```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.
```

```
##   mean(menthlth, na.rm = T)
## 1                  5.508428
```

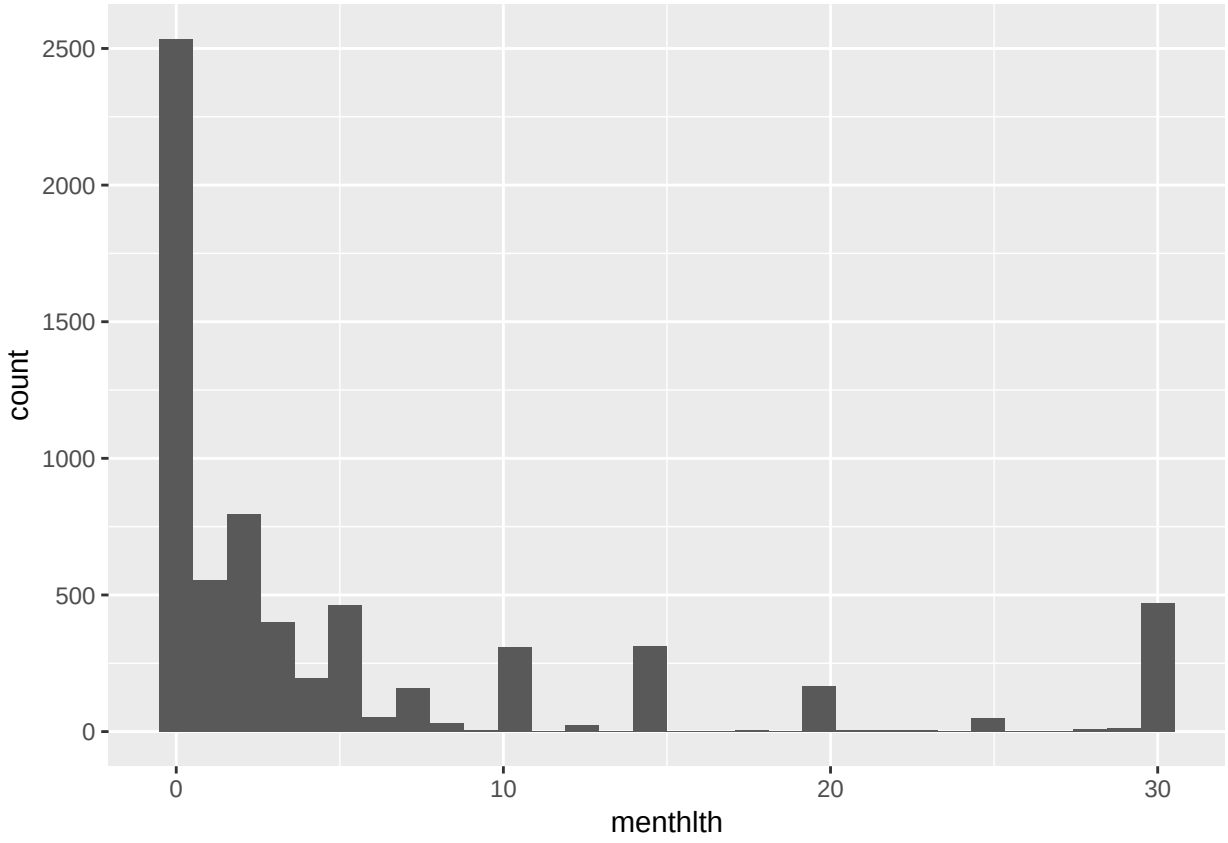
The mean is 5.51 days.

8. (3 points) Graph and describe the distribution of poor mental health days.

```
library(ggplot2)
ggplot(brfss2, aes(menthlth)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 121 rows containing non-finite values (stat_bin).
```



We can see that most of the observations are less than 10 days, over 2500 of these are actually at 0 of the past 30 days where there was poor mental health. There are around 500 that have had poor mental health 30 out of the last 30 days.

9. Create and interpret a confidence interval for poor mental health days
   a. (3 points) Using the t distribution.

```
set.seed(123)
x <- brfss2$menthlth
mn <- mean(x, na.rm=T)
t <- qt(0.975, df=999)
std.dev <- sd(x, na.rm=T)
std.error <- t*std.dev/sqrt(1000)

low = mn - std.error
high = mn + std.error
low
```

```
## [1] 4.971629
```

```
high
```

```
## [1] 6.045228
```

If we were to perform this study over and over 95% of the intervals constructed in this fashion would cover the truth. We feel that the true number of mental health days for hte population falls between 4.97 and 6.05 days out of the past 30.

   b. (6 points) Using the bootstrap approach with 1000 bootstraps.

```
set.seed(123)
library(rsample)
```

```
## Warning: package 'rsample' was built under R version 3.5.3
## Loading required package: tidyr
```

```
library(purrr)
```

```
## Warning: package 'purrr' was built under R version 3.5.3
```

```
brfss3 <- brfss2 %>% select(menthlth)
bt_data <- bootstraps(brfss3, times = 1000)
get_mean <- function(split) {
  # access to the sample data
  split_data <- analysis(split)
  # calculate the sample mean value
  split_mean <- mean(split_data$menthlth, na.rm=T)
  return(split_mean)
}
bt_data$bt_means <- map_dbl(bt_data$splits, get_mean)
bt_ci <- round(quantile(bt_data$bt_means, c(0.025, 0.975)), 3)
bt_ci
```

```
##  2.5% 97.5%
## 5.300 5.711
```

We are 95% confidence that the number of poor mental health days in the past 30 days would be between 5.3 and 5.7 days in the population.

10. Consider the variable of Binary General Health, `genhlth_bin`, and the relationship it has with poor mental health days, `menthlth`.
    a. (3 points) Bootstrap a confidence interval for the mean number of poor mental health days for those who are generally healthy.

```
set.seed(123)
library(rsample)
library(purrr)
brfss4 <- brfss2 %>%
  filter(genhlth_bin == "Excell/VG/G") %>%
  select(menthlth)

bt_data <- bootstraps(brfss4, times = 1000)
get_mean <- function(split) {
  # access to the sample data
  split_data <- analysis(split)
  # calculate the sample mean value
  split_mean <- mean(split_data$menthlth, na.rm=T)
  return(split_mean)
}
bt_data$bt_means <- map_dbl(bt_data$splits, get_mean)
bt_ci <- round(quantile(bt_data$bt_means, c(0.025, 0.975)), 3)
bt_ci
```

```
##  2.5% 97.5%
## 4.464 4.898
```

    b. (3 points) Bootstrap a confidence interval for the mean number of poor mental health days for those who are not generally healthy.

```
set.seed(123)
library(rsample)
library(purrr)
brfss5 <- brfss2 %>%
  filter(genhlth_bin == "Fair/Poor") %>%
  select(menthlth)

bt_data <- bootstraps(brfss5, times = 1000)
get_mean <- function(split) {
  # access to the sample data
  split_data <- analysis(split)
  # calculate the sample mean value
  split_mean <- mean(split_data$menthlth, na.rm=T)
  return(split_mean)
}
bt_data$bt_means <- map_dbl(bt_data$splits, get_mean)
bt_ci <- round(quantile(bt_data$bt_means, c(0.025, 0.975)), 3)
bt_ci
```

```
##  2.5% 97.5%
## 8.132 9.353
```

c. (3 points) Interpret and discuss what connections you see between these confidence intervals. ***We can see that for those who self classify as generally healthy, that we are 95% confident that they have between 4.5 and 4.9 days out of the last 30 where they had poor mental health. For those who self classify as poorer health, they have between 8.1 and 9.4 days out of the last 30. This would suggest that these 2 groups have a different number of average poor mental health days. Furthermore, we could say that those who are generally healthy have less poor mental health days.***

d. (4 points) Bootstrap and interpret a confidence interval for the difference in mean poor mental health days between those who are generally healthy and those who are not.

```
set.seed(123)
library(rsample)
library(purrr)
brfss6 <- brfss2 %>% select(menthlth, genhlth_bin)
bt_data <- bootstraps(brfss6, times = 1000)
get_diff <- function(splits) {
  d <- analysis(splits)
  mean_yes <- mean(d$menthlth[d$genhlth_bin == "Excell/VG/G"], na.rm=T)
  mean_no <- mean(d$menthlth[d$genhlth_bin == "Fair/Poor"], na.rm=T)
  mean_yes - mean_no
}
bt_data$bt_diffs <- map_dbl(bt_data$splits, get_diff)
bt_ci <- round(quantile(bt_data$bt_diffs, c(0.025, 0.975)), 3)
bt_ci
```

```
##   2.5%  97.5%
## -4.683 -3.432
```

We are 95% confident that those who are generally healthy have between 3.4 and 4.7 less poor mental health days on average than those who do not have good overall health.