

Homework #5

Carson Crenshaw (cgc8gdt)

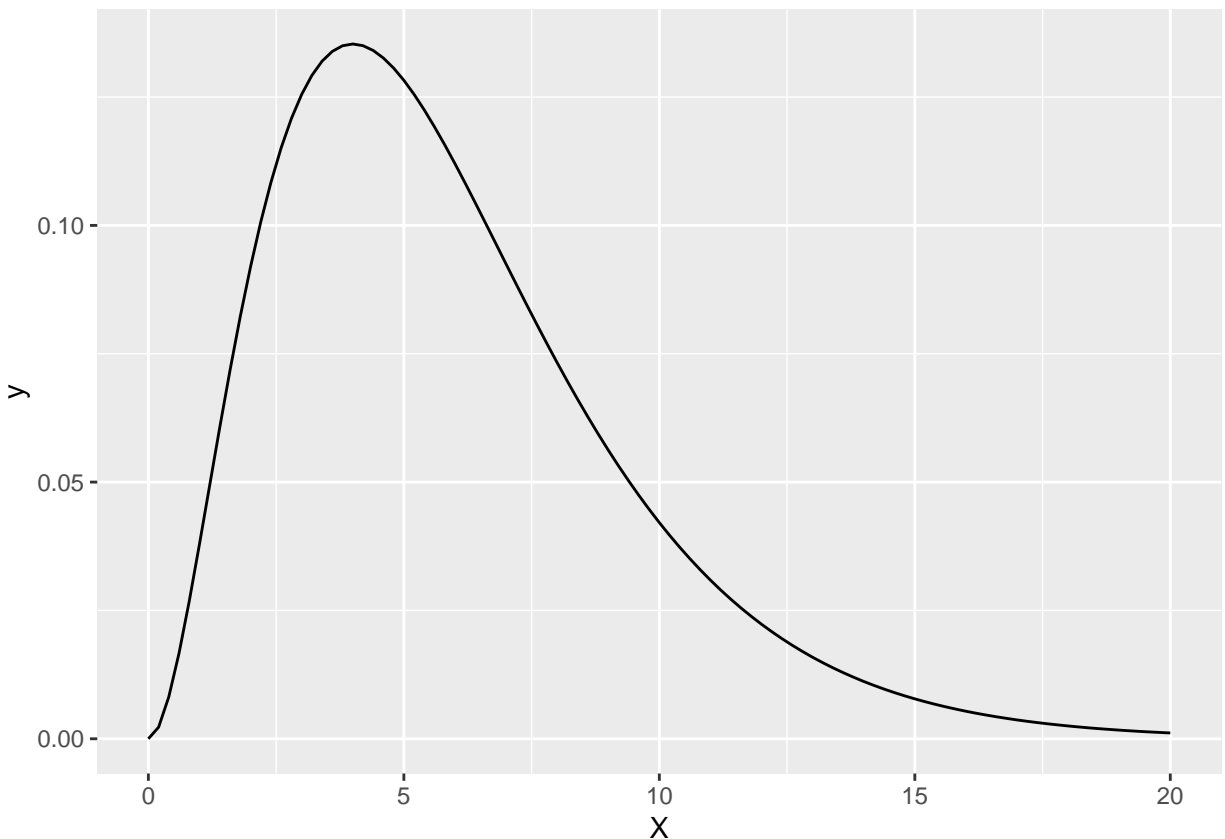
Insert packages/Set seed

```
# Set Seed
set.seed(10062002)
#Library
library(ggplot2)
```

Problem 1

The plot of the population chi-squared ($df = 6$) distribution density curve is shown below. The shape of the distribution is right (positively) skewed. Regardless of how many degrees of freedom there are, the chi-squared distribution is always skewed right. However, as the degree of freedom (k) increases, the shape of the distribution becomes closer to a normal distribution with a symmetrical bell shape. In other words, this graph is more bell-shaped than a chi-square distribution would be with only three degrees of freedom.

```
# Plotting the chi-squared population distribution
ChiSquareGraph <- data.frame(X=c(0,20))
ggplot(ChiSquareGraph, aes(x=X)) + stat_function(fun=dchisq, args=list(df=6))
```



Problem 2

Part a

For all sample sizes, the two-sided chi-squared distribution z-test results in an empirical Type I error close to what we might expect (0.05). There is a little observable variation between the different sample sizes because the sample is not coming from a normal distribution (values get closer to 0.05 as the sample size increases).

```
# Definition of Variables
popmean <- 6
popstd <- 3.53
siglevel <- 0.05

# Two-sided test
ztestfunc <- function(x){
  samp <- rchisq(x, 6)
  zscore <- ((mean(samp) - popmean) / (popstd / sqrt(x)))
  pvalue <- 2*pnorm(-abs(zscore))
  pvalue <= siglevel
}

# Sample Size of 8
emptytype1errortz1 <- replicate(10000,ztestfunc(8))
sum(emptytype1errortz1)/10000
```

```
## [1] 0.0394
```

```
# Sample Size of 26
emptytype1errortz2 <- replicate(10000,ztestfunc(26))
sum(emptytype1errortz2)/10000
```

```
## [1] 0.0447
```

```
# Sample Size of 53
emptytype1errortz3 <- replicate(10000,ztestfunc(53))
sum(emptytype1errortz3)/10000
```

```
## [1] 0.046
```

Part b

In contrast to the two-sided test, the left-sided chi-squared distribution z-test results in an empirical Type I error that clearly increases in proximity to the expected value of 0.05 as the sample size increases. All values are below 0.05 but slowly increase and approach the expected value as the sample size grows.

While one could see that the values of the two-sided test also increase slightly in value as the sample size increases, this occurrence is more clearly demonstrated in the left-sided test wherein the difference in empirical Type I error values are much greater (the values start much lower, and the difference between the value for the smallest and largest sample is much larger).

This test also seems to demonstrate the principle of the Central Limit Theorem (as greater sample sizes result in values closer to 0.05) because the sample is not taken from a normal distribution.

```
# Left-sided
ztestfuncleft <- function(x){
  samp <- rchisq(x, 6)
  zscore <- ((mean(samp) - popmean) / (popstd / sqrt(x)))
  pvalue <- pnorm(zscore, lower.tail=TRUE)
  pvalue <= siglevel
}
```

```
# Sample Size of 8
emptytype1errorlz1 <- replicate(10000,ztestfuncleft(8))
sum(emptytype1errorlz1)/10000
```

```
## [1] 0.0347
```

```
# Sample Size of 26
emptytype1errorlz2 <- replicate(10000,ztestfuncleft(26))
sum(emptytype1errorlz2)/10000
```

```
## [1] 0.0395
```

```
# Sample Size of 53
emptytype1errorlz3 <- replicate(10000,ztestfuncleft(53))
sum(emptytype1errorlz3)/10000
```

```
## [1] 0.0437
```

Part c

To compare this right-sided z-test to the previous tests, the right-sided test outputs are closer to the expected empirical Type I error value (0.05) than the left-sided test outputs. The values start above 0.05 and slowly decrease towards the expected value as the sample size increases.

The right-sided test, however, is another demonstration of the Central Limit Theorem as greater sample sizes result in values closer to 0.05. It must be noted, however, that even the smallest value is still exceptionally close to the expected value of 0.05.

```
# Right-sided
ztestfuncright <- function(x){
  samp <- rchisq(x, 6)
  zscore <- ((mean(samp) - popmean) / (popstd / sqrt(x)))
  pvalue <- pnorm(zscore, lower.tail=FALSE)
  pvalue <= siglevel
}
```

```
# Sample Size of 8
emptytype1errorrrz1 <- replicate(10000,ztestfuncright(8))
sum(emptytype1errorrrz1)/10000
```

```
## [1] 0.0565
```

```
# Sample Size of 26
emptytype1errorrrz2 <- replicate(10000,ztestfuncright(26))
sum(emptytype1errorrrz2)/10000
```

```
## [1] 0.0541
```

```
# Sample Size of 53
emptytype1errorrrz3 <- replicate(10000,ztestfuncright(53))
sum(emptytype1errorrrz3)/10000
```

```
## [1] 0.0505
```

Problem 3

Illustrated below, the pattern for the t-test and the z-test are relatively similar. For example, the two-sided tests are pretty close to the expected Empirical Type I error value of 0.05. The behavior of the right- and left-tailed tests, however, are reversed here. For the t-test, the right-tailed test gives values lower 0.05, and the left-tailed test gives values above 0.05.

Part a The two-sided t-test is very similar to that of the two-sided z-test as the values remain close to 0.05. There is an observable amount of variation between the resulting values, however, because the sample is not a normal distribution. The error values found during the t-test are also slightly more extreme than those of the z-test.

```
# Two-sided test
ttestfunc <- function(x){
  samp <- rchisq(x, 6)
  tstat <- ((mean(samp) - popmean) / (sd(samp) / sqrt(x)))
  pvalue <- 2*pt(-abs(tstat), x-1)
  pvalue <= siglevel
}

# Sample Size of 8
emptytype1errortt1 <- replicate(10000,ttestfunc(8))
sum(emptytype1errortt1)/10000
```

```
## [1] 0.0718
```

```
# Sample Size of 26
emptytype1errortt2 <- replicate(10000,ttestfunc(26))
sum(emptytype1errortt2)/10000
```

```
## [1] 0.06
```

```
# Sample Size of 53
emptytype1errortt3 <- replicate(10000,ttestfunc(53))
sum(emptytype1errortt3)/10000
```

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

Part b

In contrast to the z-test, the behavior of the left-sided t-test results in values above 0.05. Following the expectations of a distribution that is not normal, the values get closer to 0.05 as the sample grows larger. These values are also more extreme than those seen in the results of the z-test.

```
# Left-sided test
ttestfuncleft <- function(x){
  samp <- rchisq(x, 6)
  tstat <- ((mean(samp) - popmean) / (sd(samp) / sqrt(x)))
  pvalue <- pt(tstat, x-1, lower.tail=TRUE)
  pvalue <= siglevel
}
```

```
# Sample Size of 8
emptyerrorlt1 <- replicate(10000,ttestfuncleft(8))
sum(emptyerrorlt1)/10000
```

```
## [1] 0.0968
```

```
# Sample Size of 26
emptyerrorlt2 <- replicate(10000,ttestfuncleft(26))
sum(emptyerrorlt2)/10000
```

```
## [1] 0.0842
```

```
# Sample Size of 53
emptyerrorlt3 <- replicate(10000,ttestfuncleft(53))
sum(emptyerrorlt3)/10000
```

```
## [1] 0.0682
```

Part c

Finally, the right-sided t-test follows an inverse distributional pattern when compared to the z-test and results in more extreme values. The right-sided test results in less error (all values are below 0.05), yet the values get closer to 0.05 as the sample size increases.

Conceptually this is understandable because when you visualize the chi-square distribution in comparison to the t distribution the t distribution is a little below the skew of the chi-square distribution on the left side. This results in more error on the left side while the right side sees the opposite. The t-tests also see more extreme error values than the z-test.

```
# Right-sided test
ttestfuncright <- function(x){
  samp <- rchisq(x, 6)
  tstat <- ((mean(samp) - popmean) / (sd(samp) / sqrt(x)))
  pvalue <- pt(tstat, x-1, lower.tail=FALSE)
  pvalue <= siglevel
}
```

```
# Sample Size of 8
emptyerrorrt1 <- replicate(10000,ttestfuncright(8))
sum(emptyerrorrt1)/10000
```

```
## [1] 0.0224
```

```
# Sample Size of 26
emptyerrorrt2 <- replicate(10000,ttestfuncright(26))
sum(emptyerrorrt2)/10000
```

```
## [1] 0.032
```

```
# Sample Size of 53
emptyerrorrt3 <- replicate(10000,ttestfuncright(53))
sum(emptyerrorrt3)/10000
```

```
## [1] 0.0333
```

Theoretically, Central Limit Theorem should apply to all tests from questions 2 and 3 because the sample comes from a chi-squared and not a normal distribution. This is understood in the variation between sample sizes, with the larger sample sizes often holding the Empirical Type I error value closest to 0.05.

Problem 4

Using 10,000 repitons, the 95% bootstrap confience interval for the variance of the lengths of Delta Airline's arrival delays is shown below. The variance illustrates the degree of spread in the data set, with larger variances representing data that is more spread out (in relation to the mean). The large confidence interval suggests that the airline data does not provide a precise representation of the population variance. In other words, the confidence interval is not a precise population estimate because of the wide margin of values that can be found within this interval.

```
# Load in the data set
Delta_delays <- read.csv("Delta delays.csv")

# Bootstrap confidence interval
A <- Delta_delays[['Arrival.Delay']]
B<-10000
samp_var <- var(A)
boot_samp <- replicate(B, sample(A, replace=T))
boot_var <- apply(boot_samp,2,var)
boot_err <- boot_var - samp_var
boot_err_sort <- sort(boot_err)
p2.5 <- B*0.025
p97.5 <- B*0.975
boot_ci <- samp_var - boot_err_sort[c(p97.5,p2.5)]
boot_ci
```

```
## [1] 7.578348 999.663818
```

Problem 5

Part a

A right-tailed test will be used to determine if the unknown population 95th percentile of arrival delay lengths is greater than 0 (late arrival). - Null: 95% percentile = 0 - Alt: 95% percentile > 0

The resulting p-value is approximately 0.07. At a significance level of 0.1, the null hypothesis is rejected. Therefore there is sufficient evidence that the 95th percentile of arrival delay lengths on these holiday travel days is a late arrival (a positive value above 0).

```
samp_data <- Delta_delays[['Arrival.Delay']]
samp_95percentile <- quantile(samp_data, 0.95)
B<-10000
boot_95percentile <- replicate(B, quantile(sample(samp_data, replace=T), 0.95))
percent95 <- 0
boot_percentile_null <- boot_95percentile - mean(boot_95percentile) + percent95
sum(boot_percentile_null >= samp_95percentile)/B
```

```
## [1] 0.0793
```

Part b

Bootstrapping is a reliable way to evaluate the estimated sampling distribution of a parameter of interest. Despite this, the conclusion given in part A is not indisputable. With a p-value of 0.07, the conclusion will change depending on the set significance level that is determined. Because I used an alpha value of 0.10, the null hypothesis is rejected and I can conclude that there is significant evidence that the 95th percentile of arrival delay lengths is a late arrival. If I were to have used 0.05 as my significance level, however, my test would fail to reject the null and I would not have significant evidence that the 95th percentile is a late arrival.