Problem Section 1

Robustness of the t-test

Exercises

1. Every t test makes the same explicit assumption - namely, that the set of n data points - $X_1, X_2, \ldots, X_n$ - are normally distributed. If the normality assumption is not satisfied, then the ratio

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

will not have a t-distribution. However, whether or not the validity of the t-test is compromised depends on how different the actual distribution of the statistic T is from the t distribution.

In this exercise, you will investigate the sensitivity of the T ratio to violations of the normality assumption by simulating samples of size n from selected distributions and comparing the resulting histogram to the t distribution with $n-1$ degrees of freedom.

a. Simulate $B = 10,000$ samples of size $n = 6, 15$ each from a Unif(0,1) distribution. For each sample, calculate

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$$

where $\mu_0 = \frac{1}{2}$ is the mean of the uniform distribution. Create a histogram of the t ratios and superimpose the t distribution with $n-1$ degrees of freedom. What do you notice?

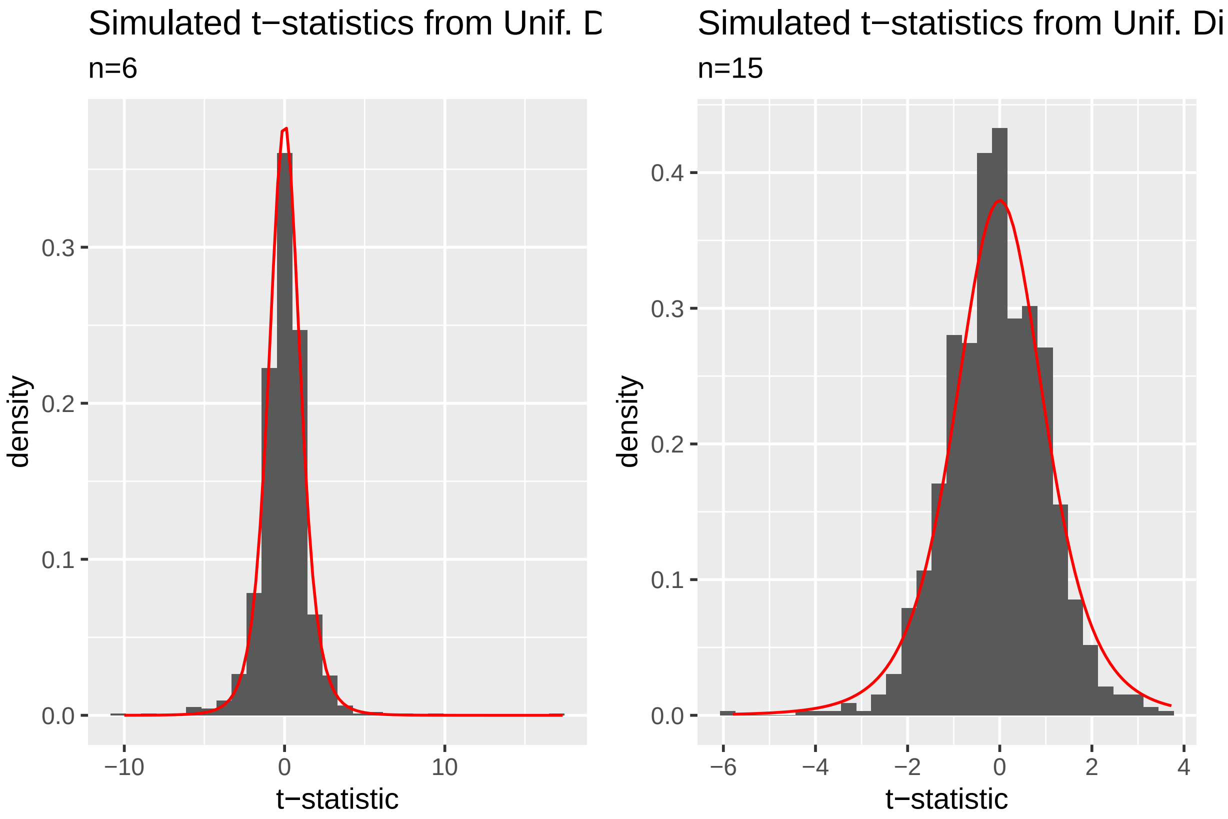
```r
set.seed(2737)
B <- 1000
sim_func_unif <- function(x){
  samp1 = runif(n=6,0,1)
  samp2 = runif(n=15,0,1)
  mu_0 <- 1/2
  t1 = (mean(samp1) - mu_0)/(sd(samp1)/sqrt(6))
  t2 = (mean(samp2) - mu_0)/(sd(samp2)/sqrt(15))
  return(c("t1" = t1,"t2" = t2))
}
sims <- lapply(1:B, sim_func_unif)
sims_df <- data.frame(do.call(rbind,sims))

p1 <- ggplot(data=sims_df) +
  geom_histogram(aes(x = t1 , y=..density..), bins=30)+
  stat_function(fun = dt, args = list(df = 5), col='red')+
  labs(x = "t-statistic",
       title = "Simulated t-statistics from Unif. Dist.",
       subtitle = "n=6")

p2 <- ggplot(data=sims_df) +
  geom_histogram(aes(x = t2 , y=..density..), bins=30)+
  stat_function(fun = dt, args = list(df = 5), col='red')+
  labs(x = "t-statistic",
       title = "Simulated t-statistics from Unif. Dist.",
       subtitle = "n=15")

p1+p2
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```



We notice that the simulated values fit quite well with the theoretical t-distribution.

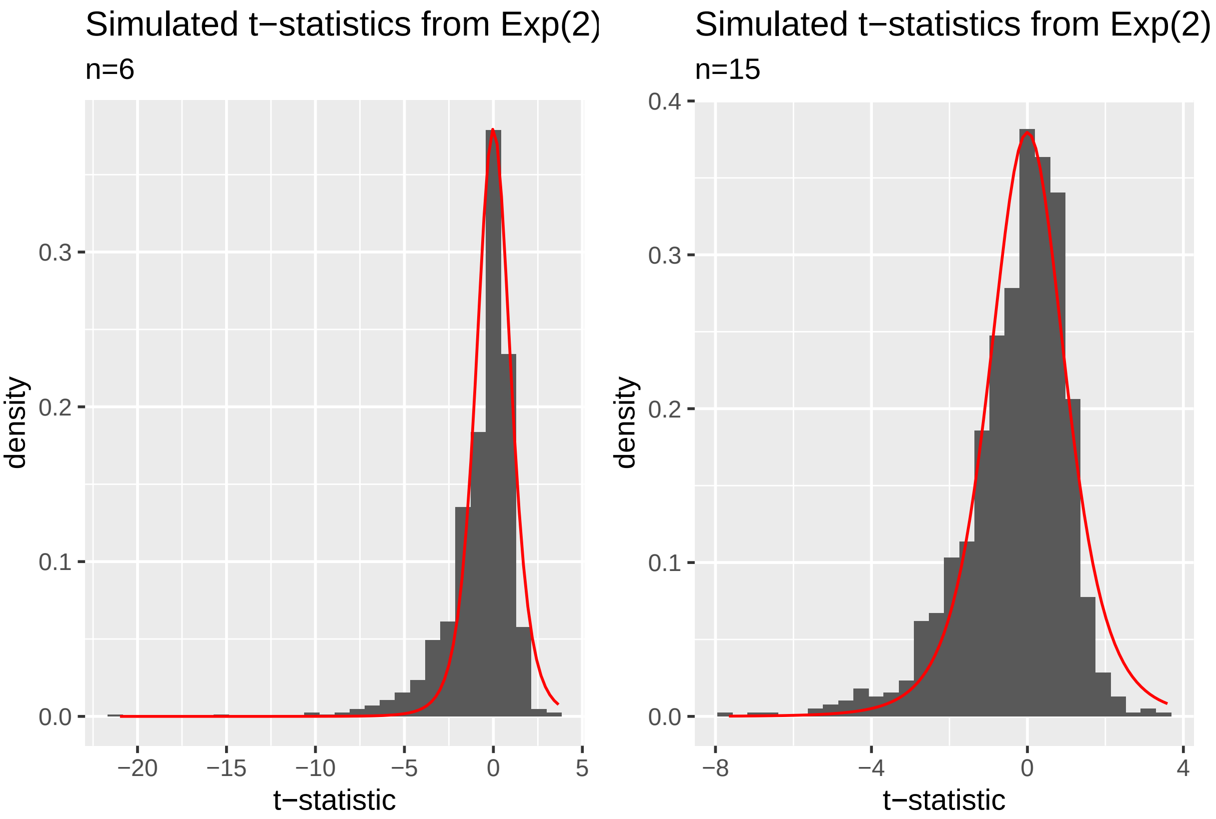
b. Repeat part a. for samples drawn from an exponential distribution with rate $\lambda_0 = 2$. (Note: $\mu_0 = 1/2$ for this distribution also) What do you notice?

```
set.seed(88)
B <- 1000
rate_param <- 2
sim_func_exp <- function(x){
  samp1 = rexp(n=6, rate=rate_param)
  samp2 = rexp(n=15,rate=rate_param)
  mu_0 = 1/2
  t1 = (mean(samp1) - mu_0)/(sd(samp1)/sqrt(6))
  t2 = (mean(samp2) - mu_0)/(sd(samp2)/sqrt(15))
  return(c("t1" = t1,"t2" = t2))
}
sims <- lapply(1:B, sim_func_exp)
sims_df <- data.frame(do.call(rbind,sims))

p1 <- ggplot(data=sims_df) +
  geom_histogram(aes(x = t1 , y=..density..), bins=30)+
  stat_function(fun = dt, args = list(df = 5), col='red')+
  labs(x = "t-statistic",
       title = "Simulated t-statistics from Exp(2)",
       subtitle = "n=6")

p2 <- ggplot(data=sims_df) +
  geom_histogram(aes(x = t2 , y=..density..), bins=30)+
  stat_function(fun = dt, args = list(df = 5), col='red')+
  labs(x = "t-statistic",
       title = "Simulated t-statistics from Exp(2)",
       subtitle = "n=15")

p1+p2
```



Here we notice, similar to the uniform case, that the simulated t-values follow the theoretical distribution quite well. This points to the robustness of the t-test even when normality assumptions are violated.

2. Your simulations in problem 1 should show that the distribution of

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$$

will become increasingly similar to a t_{n-1} distribution as n increases, regardless of which distribution you sample from. Can you explain why this happens?

We see this happens because we know by the CLT that sums of random variables converge to a normal distribution. Since we are dealing with sample means, the CLT lets us relax our normal assumptions.

3. What is $Cov(X, X)$?

$Cov(X, X) = Var(X)$ by definition. We see this occurs because:

$$Cov(X, X) = E[(X - E(X))(X - E(X))] = E[(X - E(X))^2] = Var(X)$$

4. Two draws are made at random from the box below $\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}}$.

Let X denote the number on the first randomly selected ticket and Y the second. The joint PMF of $\langle X, Y\rangle$ is shown below.

		With replacement				Without replacement			
		X			f_Y	X			f_Y
		1	2	3		1	2	3	
Y	1	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$
	2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{3}$
	3	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{3}$
	f_X	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1.00	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1.00

a. When the draws are made with replacement, $Cov\,[X, Y] = 0$. Why?

When draws are made with replacement, the number on the first and second ticket will be independent. This means that they will thus have zero covariance. This occurs because if X and Y are independent, then $E(XY) = E(X)E(Y)$. Thus we have for the covariance:

$$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

b. Find $Cov\,[X, Y]$ when the draws are made without replacement. Does the sign make sense?

We know that $Cov(X, Y) = E(XY) - E(X)E(Y)$.

We have that $E(X) = E(Y) = (1/3) \times 1 + (1/3) \times 2 + (1/3) \times 3 = 2$

For $E(XY)$ we must consider their joint distribution $f(x, y)$ as given by the table in the without replacement section. We see that since the joint distribution is symmetric (think about why) we can utilize that in making our calculations easier. Specifically we can use that $P(X = i, Y = j) = P(X = j, Y = i)$

We have that this will be:

$$\begin{aligned} E(XY) &= (1 \times 1) \times 0 + [(1 \times 2) \times \frac{1}{6} \times 2] + [(1 \times 3) \times \frac{1}{6} \times 2] + [(2 \times 3) \times \frac{1}{6} \times 2] \\ &= \frac{2}{6} \times (1 \times 2 + 1 \times 3 + 2 \times 3) \end{aligned}$$

Using R we see that this is equal to:

```r
E_xy <- (2/6)*sum(c(1*2+1*3+2*3))
E_xy
```

```
## [1] 3.666667
```

Combining this with $E(X) = E(Y) = 2$ we have that the covariance is equal to:

```r
#note I name the covariance cov_xy, since cov is an R function and
#we do not want to overwrite it with a variable
cov_xy <- E_xy - 2*2
cov_xy
```

```
## [1] -0.3333333
```

We see that the covariance is negative in this case. This implies that X and Y are negatively correllated, i.e. that on average as X increases, Y decreases and vice versa. This makes sense as since the draws are without replacement, if X is drawn as a 1 (low number), then Y must be larger. Similarly if X is drawn as a 3 (high number) then Y will be smaller. In the case of X=2, it is equally likely that Y could be smaller or larger, but since we observe a clear increasing/decreasing relationship 2/3 of the time we still have a negative covariance.