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.

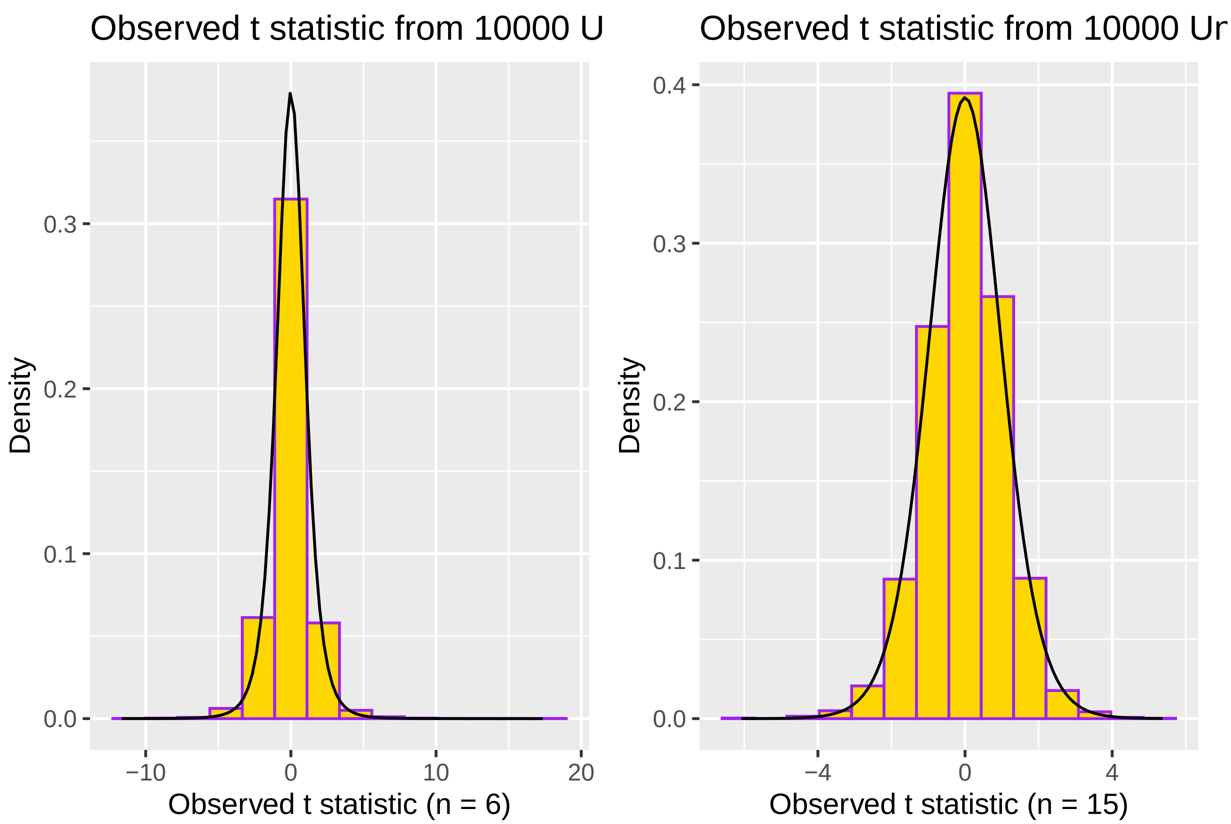
a. Simulate $B = 10,000$ samples of size $n = 6, 15$ each from a Unif(0,1) distribution. For each sample, calculate

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

   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?

```
set.seed(2737)
B = 10000
mu0 = 1/2

get_tobs <- function(x, n) {
  xbar = mean(x)
  sd = sd(x)
  tobs = sqrt(n) * (xbar - mu0) / sd
}

unif_tobs_df <- tibble(tobs_6 = replicate(n = B, expr = get_tobs(runif(6, 0, 1), 6)),
                       tobs_15 = replicate(n = B, expr = get_tobs(runif(15, 0, 1), 15))
                       )

p1 <- ggplot(data = unif_tobs_df, mapping = aes(x = tobs_6)) +
      geom_histogram(mapping = aes(y = after_stat(density)), bins = 14, color = "purple", fill = "gold")
      geom_function(fun = dt, args = list(df = 6 - 1)) +
      labs(x = "Observed t statistic (n = 6)",
           y = "Density",
           title = "Observed t statistic from 10000 Uniform(0,1), t overlaid")

p2 <- ggplot(data = unif_tobs_df, mapping = aes(x = tobs_15)) +
      geom_histogram(mapping = aes(y = after_stat(density)), bins = 14, color = "purple", fill = "gold")
      geom_function(fun = dt, args = list(df = 15 - 1)) +
      labs(x = "Observed t statistic (n = 15)",
```

```
         y = "Density",
         title = "Observed t statistic from 10000 Uniform(0,1), t overlaid")

p1 + p2
```

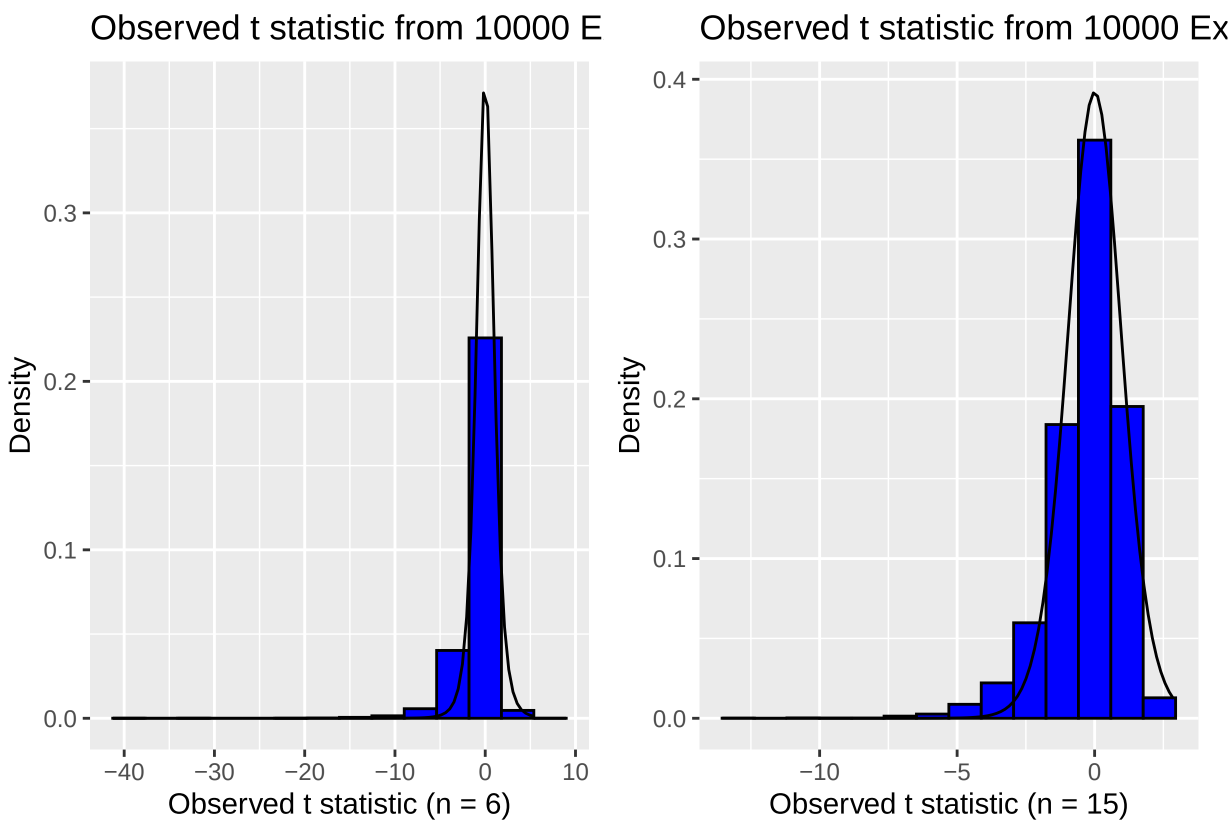


For both $n = 6$ and $n = 15$ we see that the t distribution fits the distribution very well.

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)
exp_tobs_df <- tibble(tobs_6 = replicate(n = B, expr = get_tobs(rexp(6, 2), 6)),
                       tobs_15 = replicate(n = B, expr = get_tobs(rexp(15, 2), 15))
                       )

g1 <- ggplot(data = exp_tobs_df, mapping = aes(x = tobs_6)) +
      geom_histogram(mapping = aes(y = after_stat(density)), bins = 14, color = "black", fill = "blue")
      geom_function(fun = dt, args = list(df = 6 - 1)) +
      labs(x = "Observed t statistic (n = 6)",
           y = "Density",
           title = "Observed t statistic from 10000 Exp(2), t overlaid")

g2 <- ggplot(data = exp_tobs_df, mapping = aes(x = tobs_15)) +
      geom_histogram(mapping = aes(y = after_stat(density)), bins = 14, color = "black", fill = "blue")
      geom_function(fun = dt, args = list(df = 15 - 1)) +
      labs(x = "Observed t statistic (n = 15)",
           y = "Density",
```

```r
         title = "Observed t statistic from 10000 Exp(2), t overlaid")

g1 + g2
```



As n increases, the t distribution fits the distribution of the data increasingly well.

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?

As n increases, the distribution of $\bar{X} \approx$ Norm by the CLT. Thus, the distribution of T will be approximately t_{n-1}.

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

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.

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

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

		With replacement				Without replacement			
		X			f_Y	X			f_Y
		1	2	3		1	2	3	
	1	1/9	1/9	1/9	1/3	0	1/6	1/6	1/3
Y	2	1/9	1/9	1/9	1/3	1/6	0	1/6	1/3
	3	1/9	1/9	1/9	1/3	1/6	1/6	0	1/3
	f_X	1/3	1/3	1/3	1.00	1/3	1/3	1/3	1.00