S&DS 220: Homework 9

Due Friday April 12

Instructions

1. Complete the questions below. Upload your knitted PDF solutions to Gradescope by the due date.
2. Your solutions should be a combination of writing and R code. When writing, use complete sentences.
3. Previous homework assignments already had code chunks created for you. Now it is up to you to insert R code chunks within each problem as needed.
4. You should aim for clear and concise communication (in both words and R code).

Problem set questions

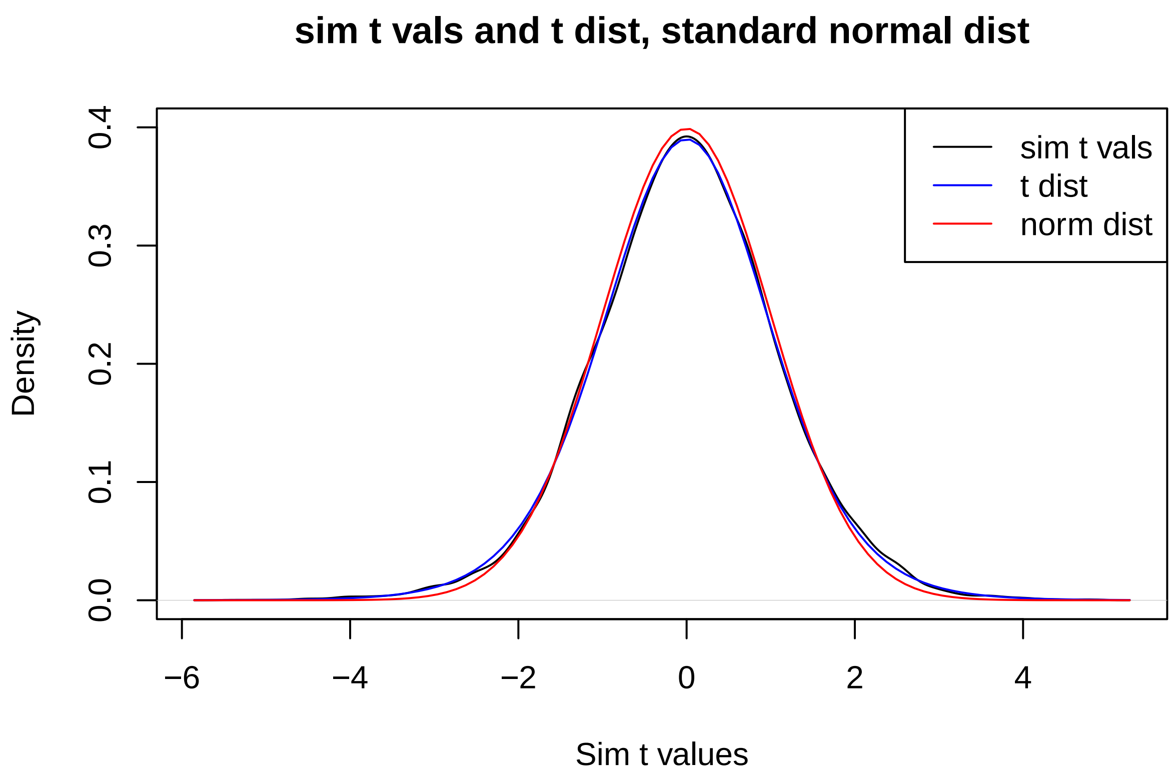
Question 1: (8.1) Density plot of T distribution

Let $X_1, \ldots, X_12$ be independent normal random variables with mean 1 and standard deviation 3. Simulate 10000 values of

$$T = \frac{\overline{X} - 1}{S/\sqrt{12}}$$

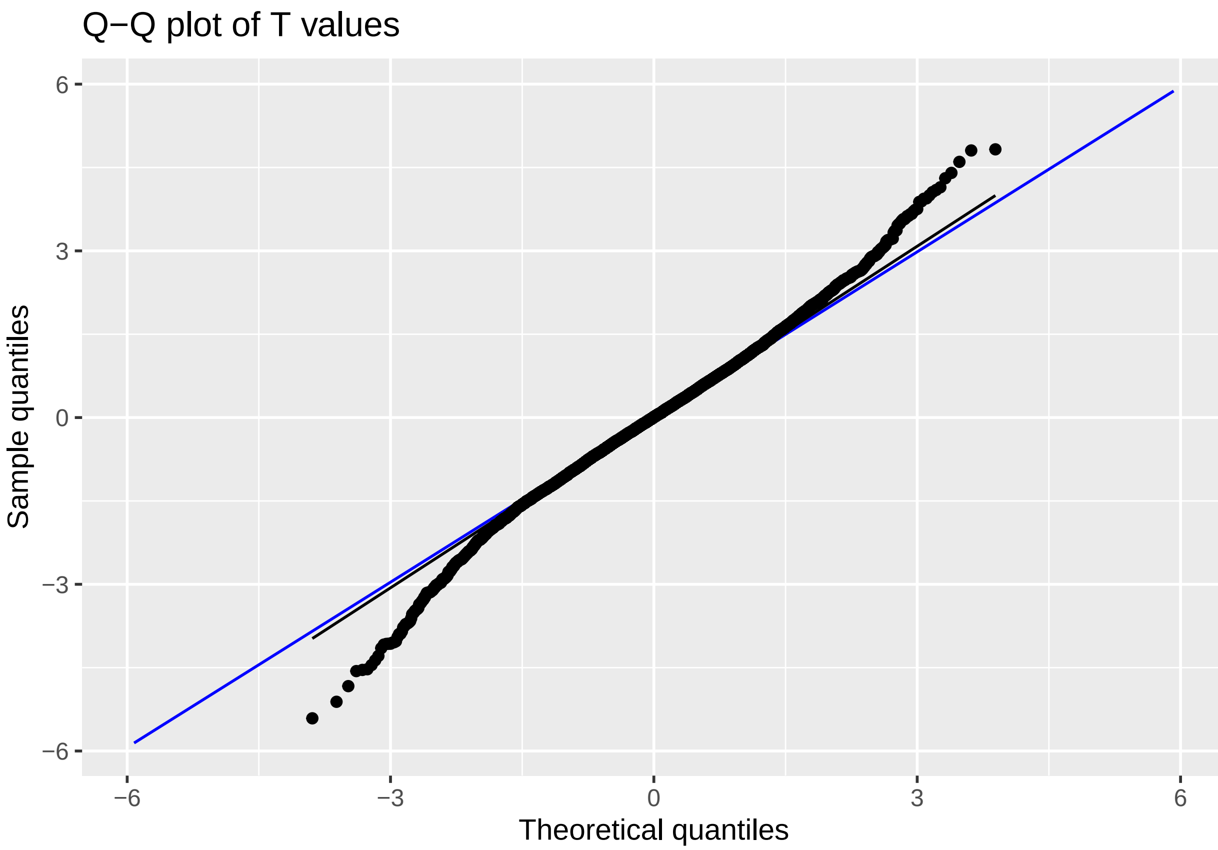
and plot the density function of T. On your plot, add a curve in blue for t with 11 degrees of freedom. Also add a curve in red for the standard normal distribution. Confirm that the distribution of T is t with 11 df.

```
density_func_t <- replicate(1e4, {
  v <- rnorm(12, 1, 3)
  s <- sd(v)
  x <- mean(v)
  t <- (x - 1) / (s / sqrt(12))
})

plot(density(density_func_t),
     ylim = c(0, 0.4),
     xlab = "Sim t values",
     main = "sim t vals and t dist, standard normal dist")
curve(dt(x, 11), col = "blue", add = TRUE)
curve(dnorm(x), col = "red", add = TRUE)
legend("topright",
       legend = c("sim t vals", "t dist", "norm dist"),
       col = c("black", "blue", "red"),
       lty = 1)
```



Now, we need to show that the T follows a t-dist with 11 df. How can we show this? Using a QQ-plot. The t distribution fits the Q-Q plot for our sample very well, as seen below.

```
T_vals = density_func_t
q_plot <- tibble(T_vals = T_vals) %>%
  ggplot(aes(sample = T_vals)) +
  geom_qq_line(distribution = qt, dparams = list(df = 11), col = "blue") +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Q-Q plot of T values", x = "Theoretical quantiles", y = "Sample quantiles")

q_plot
```



Q−Q plot of T values
Sample quantiles
Theoretical quantiles
6
3
0
−3
−6
−6
−3
0
3
6

Question 2: (8.7) Confidence interval for 19th century speed of light experiments

The data set `morley` is built into R. The Speed variable contains 100 measurements of the speed of light, conducted by A. A. Michelson in 1879. Measurements have 299000 km/s subtracted from them.

Compute the 95% confidence interval for the speed of light. Does your confidence interval contain the modern accepted speed of light of 299792 km/s?

```r
N <- nrow(morley)
xb <- mean(morley$Speed)
s <- sd(morley$Speed) / sqrt(N)
tval <- qt(0.025, df = N-1, lower.tail = FALSE)
c <- xb + c(-1, 1) * tval * s
c
```

```
## [1] 836.7226 868.0774
```

Question 3: (8.13) Performance of confidence intervals

Suppose a population has an exponential distribution with $\lambda = 0.5$. We can simulate drawing a sample of size 10 from this population with `rexp(10, 0.5)` and compute a 95% confidence interval with `t.test(rexp(10, 0.5), mu = 2)$conf.int`.

(a) What is the population mean μ?

the population mean is $1/(0.5) = 2$

(b) Write code to simulate 10000 confidence intervals and determine what percent of the time μ is in the 95% confidence interval.

```
m <- 2

m_c <- replicate(1e4, {
  x <- t.test(rexp(10, 0.5), m = 2)$conf.int
  (x[1] < m) & (m < x[2])

})

mean(m_c)
```

```
## [1] 0.8968
```

(c) Why is your answer different from 95%?

the solution is different from 95% b/c the normality assumption is violated. The exponential distribution is not normal, so the confidence interval is not accurate. a sample size of 10 is simply not sufficient to be normal.

Question 4: (8.17) Analyzing health data

This problem uses the `bp.obese` data set from the `ISwR` package. Consider the `obese` variable. What is the natural null hypothesis? Is there evidence to suggest that the obesity level of the population differs from the null hypothesis?

```
library(ISwR)
```

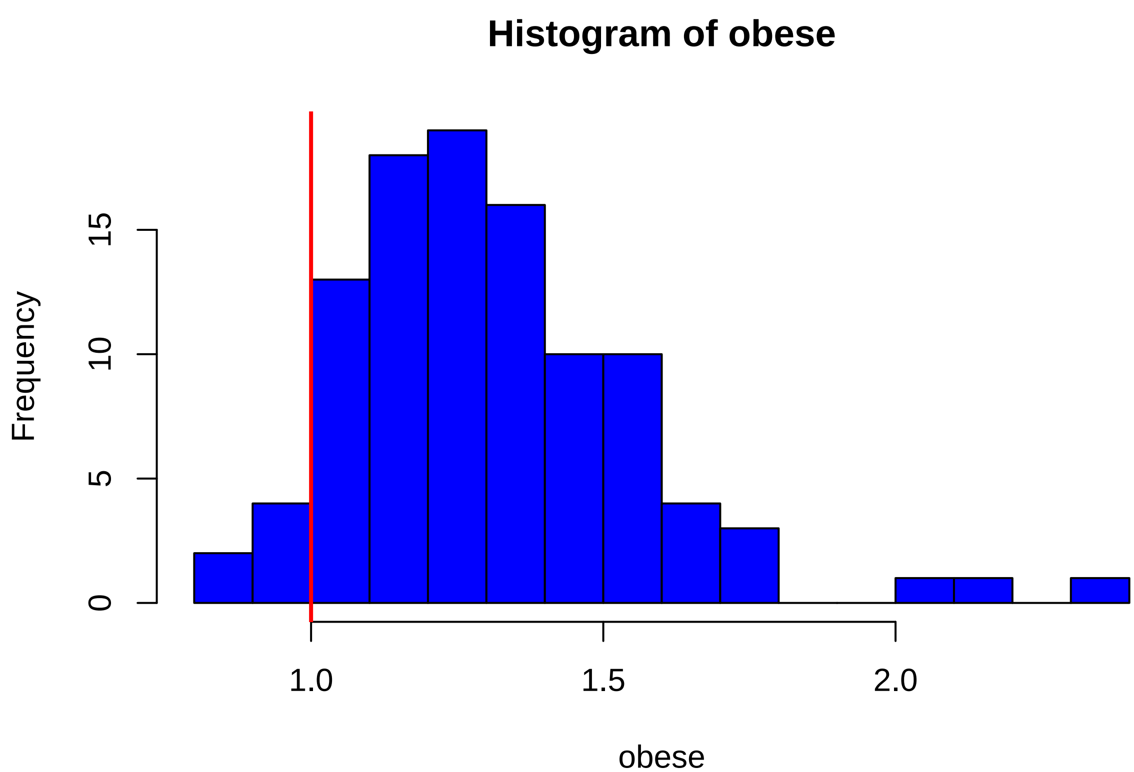
```
##
## Attaching package: 'ISwR'

## The following object is masked from 'package:fosdata':
##
##     malaria
```

```
glimpse(bp.obese)
```

```
## Rows: 102
## Columns: 3
## $ sex   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ obese <dbl> 1.31, 1.31, 1.19, 1.11, 1.34, 1.17, 1.56, 1.18, 1.04, 1.03, 0.88~
## $ bp    <int> 130, 148, 146, 122, 140, 146, 132, 110, 124, 150, 120, 114, 136,~
```

the natural null hypothesis is $H0 = 1$

```
with(bp.obese, hist(obese, breaks = 20, col = "blue"))
abline(v = 1, col = "red", lwd = 2)
```



```r
t.test(bp.obese$obese, mu = 1)
```

```
##
##  One Sample t-test
##
## data:  bp.obese$obese
## t = 12.262, df = 101, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  1.262395 1.363684
## sample estimates:
## mean of x
##  1.313039
```

Question 5: (8.18) Significance level and rejection region

Suppose you collect a random sample of size 20 from a normal population with unknown mean and standard deviation. You wish to test $H_0 : \mu = 2$ versus $H_a : \mu \neq 2$.

(a) The region $|T| > 1.6$ is a rejection region for this hypothesis test. What is the α level of the rejection region?

```
val <- pt(1.6, df = 19, lower.tail = FALSE)
2 * val
```

```
## [1] 0.1260951
```

(b) Find a rejection region that corresponds to $\alpha = 0.005$.

```
val_2 <- qt(0.0025, df = 19, lower.tail = FALSE)
val_2
```

```
## [1] 3.173725
```

the level of rejection is when $|T| > 3.17$

Question 6: (8.20) Bad statistical practice, determing H_a after collecting data

Suppose that a dishonest statistician is doing a t-test of $H_0 : \mu = 0$ at the $\alpha = 0.05$ level. The statistician waits until they get the data to specify the alternative hypothesis. If $\overline{X} > 0$, then they choose $H_a : \mu > 0$ and if $\overline{X} < 0$, then they choose $H_a : \mu < 0$.

Suppose the statistician collects 20 independent samples and the underlying population is standard normal. Use simulation to confirm that the null hypothesis is rejected 10% of the time.

We can simulate the statistician's behavior by generating 10000 samples of size 20 from a standard normal distribution and computing the p-value for each sample. We can then determine the proportion of samples for which the null hypothesis is rejected. As seen below, we will reject the null hypothesis.

```
v <- replicate(1e4, {
  x <- rnorm(20)
  xb <- mean(x)
  t <- t.test(x)$p.value
  if (xb >= 0)
    {
    p <- t.test(x, alternative = "greater")$p.value
  }
  else
    {
    p <- t.test(x, alternative = "less")$p.value
    }
  p < 0.05
})
mean(v)
```

```
## [1] 0.1027
```