Three S's

Alan T. Arnholt

Tuesday, August 29, 2017 - 11:18:44.

Using R Markdown

Using R Markdown allows us to write both text and code in the same document. Use R code chunks to insert code:

````
```{r}
# Some code
x <- 1:10
```
````

Use inline `R` to write answers inline using the following format: `` `r R_CODE` ``. For example, to compute the mean of the values 1, 3, 5, and 7, one might can use `` `r mean(c(1, 3, 5, 7))` ``. The mean of 1, 3, 5, and 7 is 4.

Using packages

To use functions in packages such as **`psych`**, one must either specify the package by prepending the function with the package name and two colons or load the package using the command `library(PackageName)`.

Consider using the function `describe` on the `mtcars` data set.

```r
describe(mtcars) # psych has not been loaded!
```

```
Error in describe(mtcars): could not find function "describe"
```

```r
psych::describe(mtcars)
```

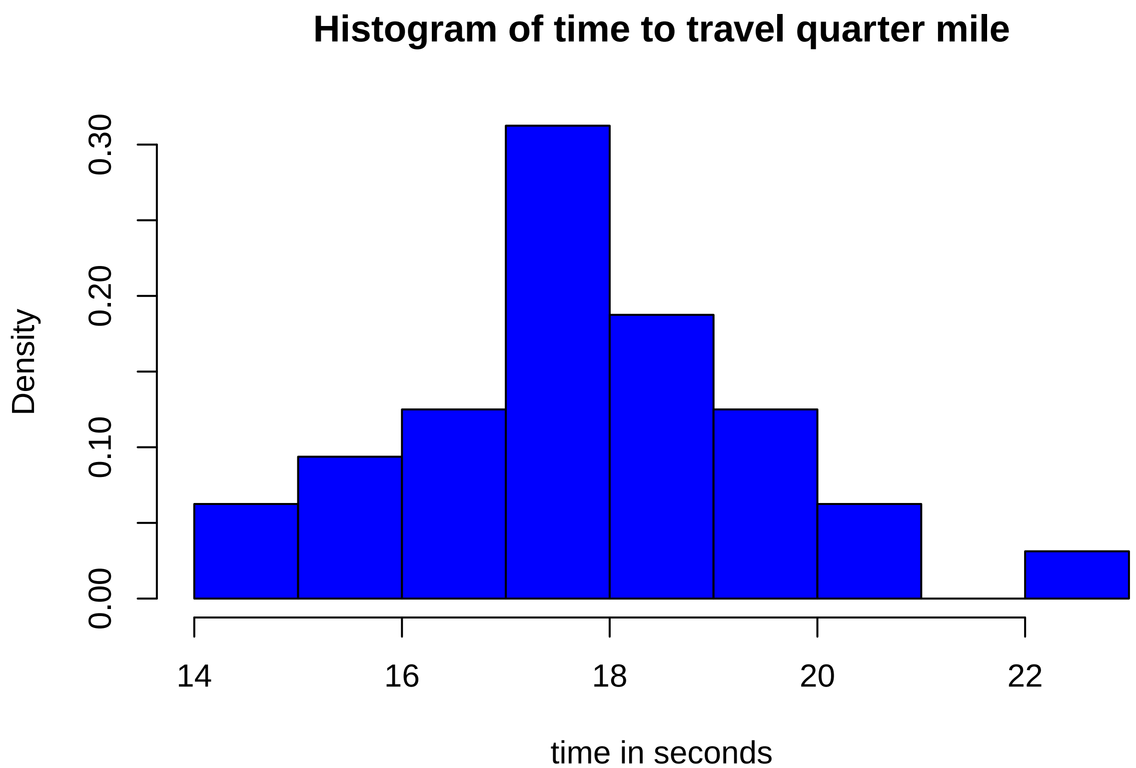
```
     vars  n   mean     sd median trimmed    mad   min    max  range  skew
mpg     1 32  20.09   6.03  19.20   19.70   5.41 10.40  33.90  23.50  0.61
cyl     2 32   6.19   1.79   6.00    6.23   2.97  4.00   8.00   4.00 -0.17
disp    3 32 230.72 123.94 196.30  222.52 140.48 71.10 472.00 400.90  0.38
hp      4 32 146.69  68.56 123.00  141.19  77.10 52.00 335.00 283.00  0.73
drat    5 32   3.60   0.53   3.70    3.58   0.70  2.76   4.93   2.17  0.27
wt      6 32   3.22   0.98   3.33    3.15   0.77  1.51   5.42   3.91  0.42
qsec    7 32  17.85   1.79  17.71   17.83   1.42 14.50  22.90   8.40  0.37
vs      8 32   0.44   0.50   0.00    0.42   0.00  0.00   1.00   1.00  0.24
am      9 32   0.41   0.50   0.00    0.38   0.00  0.00   1.00   1.00  0.36
gear   10 32   3.69   0.74   4.00    3.62   1.48  3.00   5.00   2.00  0.53
carb   11 32   2.81   1.62   2.00    2.65   1.48  1.00   8.00   7.00  1.05
     kurtosis    se
mpg     -0.37  1.07
cyl     -1.76  0.32
disp    -1.21 21.91
hp      -0.14 12.12
drat    -0.71  0.09
wt      -0.02  0.17
qsec     0.34  0.32
vs      -2.00  0.09
am      -1.92  0.09
```

```
gear      -1.07   0.13
carb       1.26   0.29
```

Characterizing `qsec`

- Shape

```r
hist(mtcars$qsec, col = "blue", freq = FALSE,
      main = "Histogram of time to travel quarter mile",
      xlab = "time in seconds")
```



- Center

```r
Mean <- mean(mtcars$qsec)
Mean
```

```
[1] 17.84875
```

- Spread

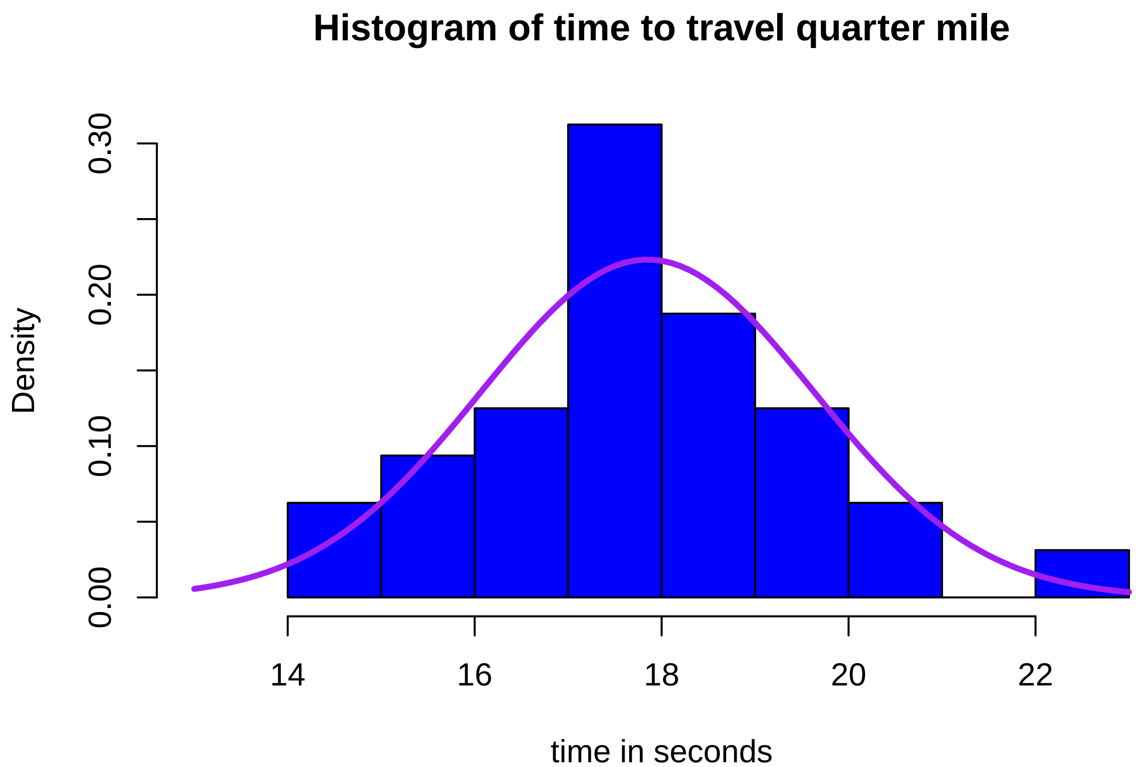
```r
SD <- sd(mtcars$qsec)
SD
```

```
[1] 1.786943
```

The distribution of `qsec` is unimodal and symmetric with a mean of 17.85 seconds and a standard deviation of 1.79 seconds.

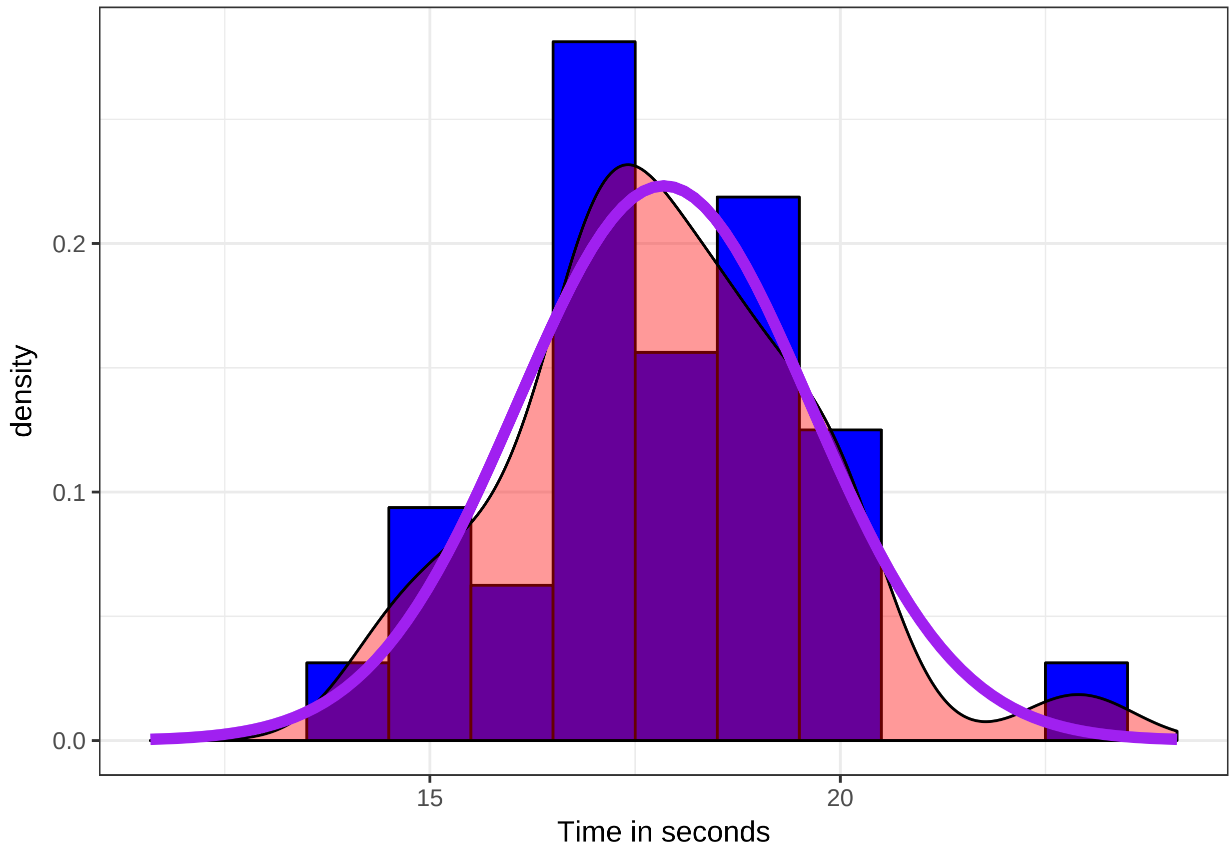
Superimposing a Normal Distribution

```
hist(mtcars$qsec, col = "blue", freq = FALSE,
      main = "Histogram of time to travel quarter mile",
      xlab = "time in seconds", xlim = c(13, 23))
curve(dnorm(x, Mean, SD), 13, 23, col = "purple", add = TRUE, lwd = 3)
```



Using ggplot2

```
library(ggplot2)
ggplot(data = mtcars, aes(x = qsec, ..density..)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  xlim(Mean - 3.5*SD, Mean + 3.5*SD) +
  labs(x = "Time in seconds") +
  geom_density(fill = "red", alpha = 0.4) +
  stat_function(fun = dnorm, args = list(mean = Mean, sd = SD),
                inherit.aes = FALSE, size = 2, color = "purple") +
  theme_bw()
```



Tests of Significance

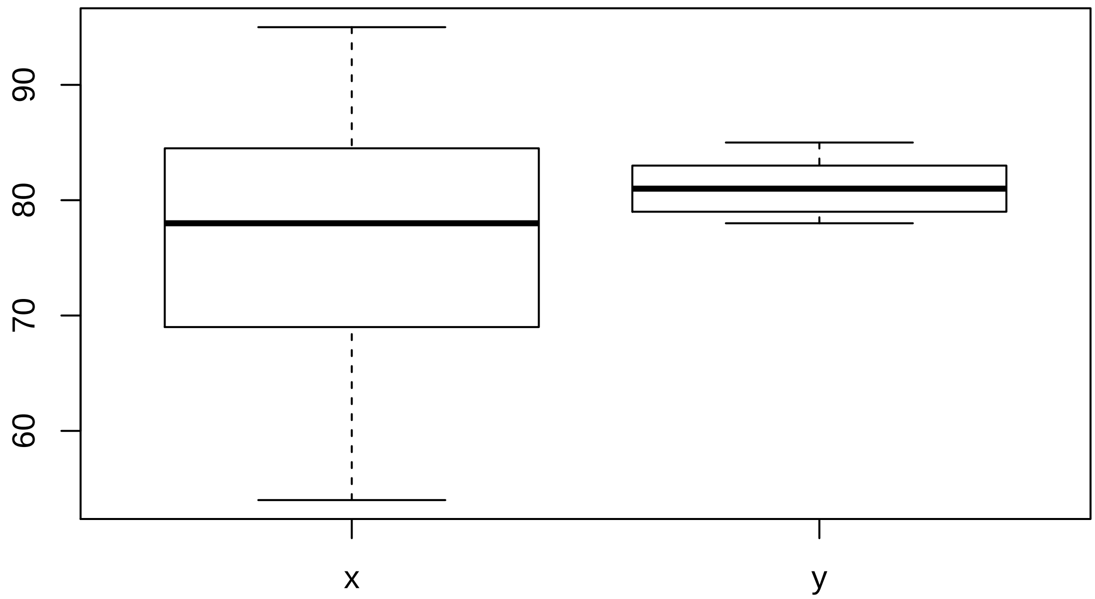
1. Hypotheses — State the null and alternative hypotheses.
2. Test Statistic
3. Rejection Region Calculations
4. Statistical Conclusion
5. English Conclusion

Example 9.12 from PASWR2

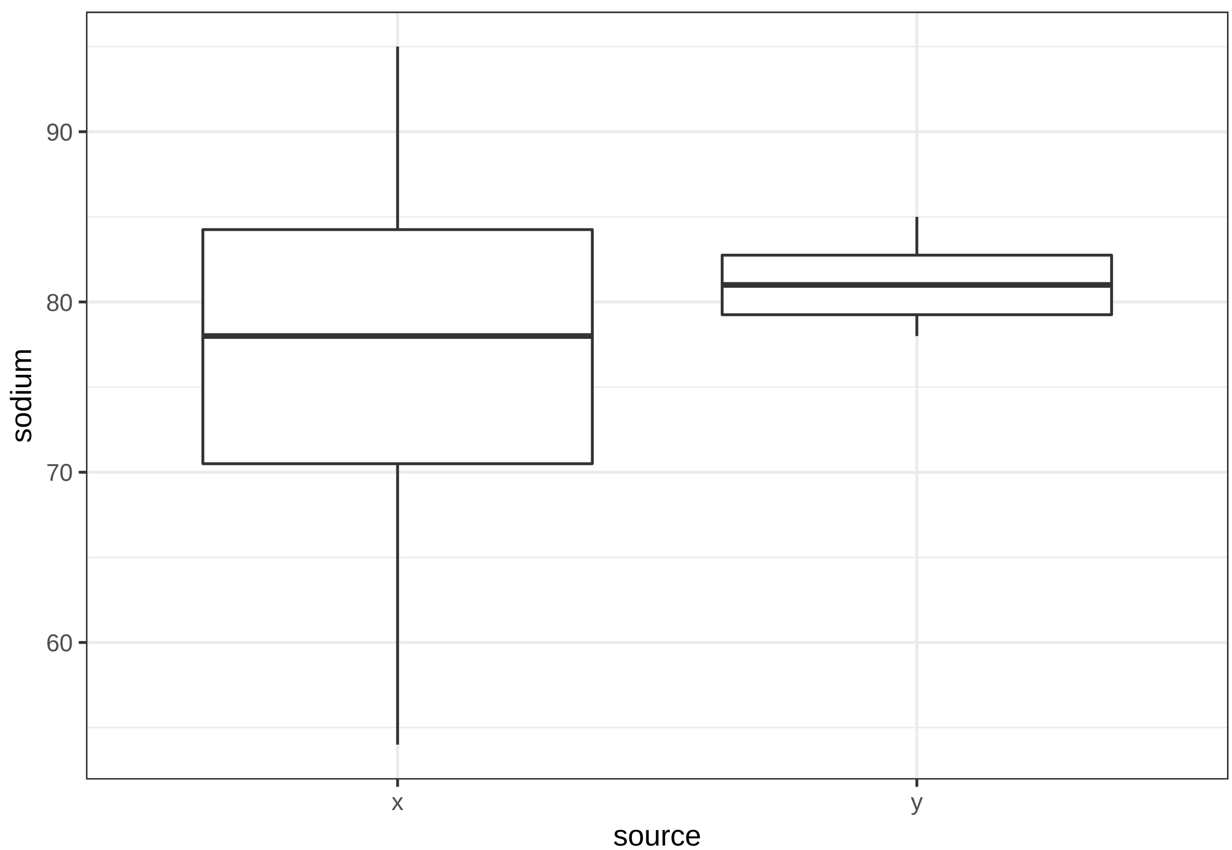
A bottled water company acquires its water from two independent sources, X and Y. The company suspects that the sodium content in the water from source X is less than the sodium content from source Y. An independent agency measures the sodium content in 20 samples from source X and 10 samples from source Y and stores them in data frame `WATER` of the `PASWR2` package. Is there statistical evidence to suggest the average sodium content in the water from source X is less than the average sodium content in Y?

Solution: To solve this problem, start by verifying the reasonableness of the normality assumption.

```
library(PASWR2)      # load the PASWR2 package
library(ggplot2)     # load the ggplot2 package
library(lsr)         # load the lsr package
library(DescTools)   # load the DescTools package
boxplot(sodium ~ source, data = WATER)
```




```r
ggplot(data = WATER, aes(x = source, y = sodium)) +
geom_boxplot() +
theme_bw()
```

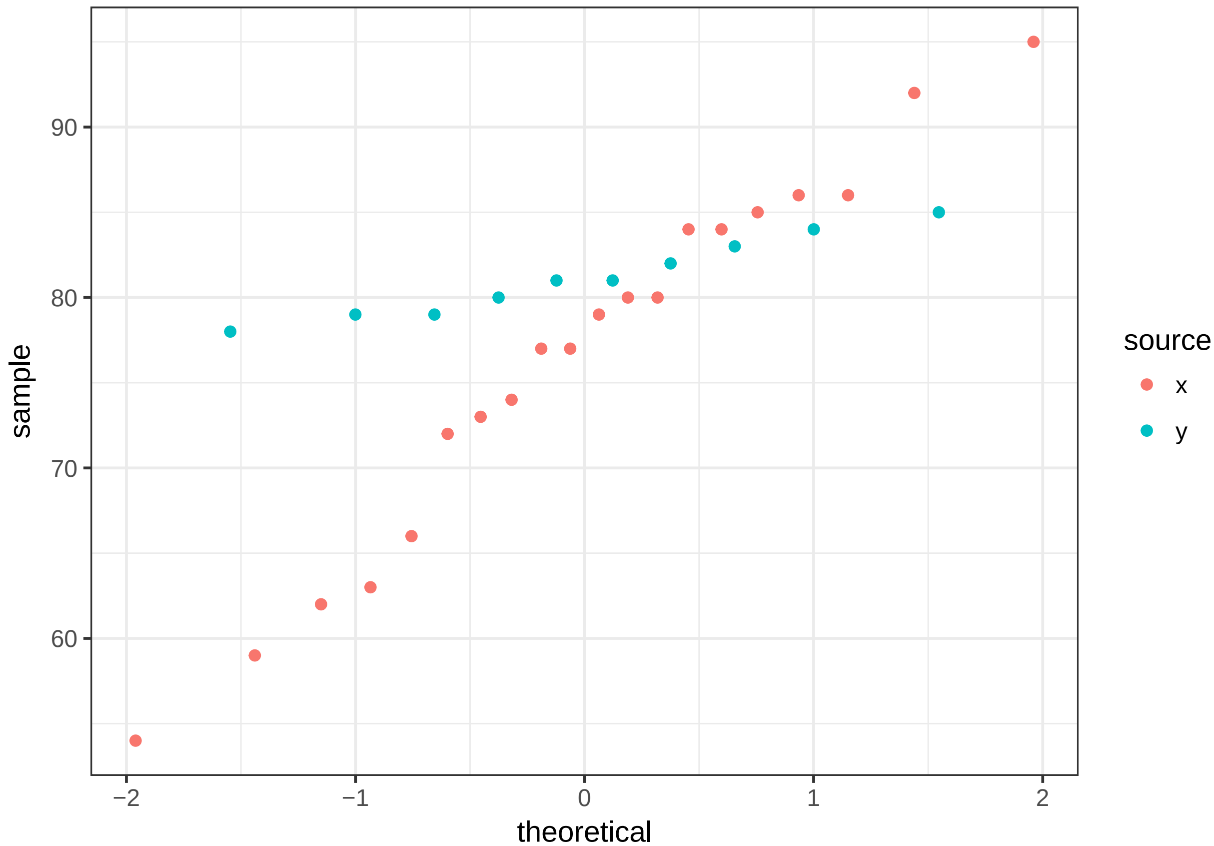



```r
LeveneTest(sodium ~ source, data = WATER)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value   Pr(>F)
```

```
group   1  10.033 0.003697 **
      28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ggplot(data = WATER, aes(sample = sodium, color = source)) +
stat_qq() +
theme_bw()
```




```r
CohenD(WATER$x, WATER$y, na.rm = TRUE)
```

```
[1] -0.5205894
attr(,"magnitude")
[1] "medium"
```

```r
cohensD(formula = sodium ~ source, data = WATER)
```

```
[1] 0.5205894
```

1. **Hypotheses** — $H_0 : \mu_X - \mu_Y = 0$ versus $H_1 : \mu_X - \mu_Y < 0$
2. **Test Statistic** —The test statistic is $\bar{X} - \bar{Y}$. The standardized test statistic under the assumptioon that H_0 is true and its approximate distribution are

$$\frac{\left[(\bar{X} - \bar{Y} - \delta_0)\right]}{\sqrt{\frac{S_X^2}{n_x} + \frac{S_Y^2}{n_Y}}} \overset{\bullet}{\sim} t_\nu$$

3. **Rejection Region Calculations** — $P(t_{obs} < t_{0.05,22.069}) = -1.7169086$.

```r
TR <- t.test(sodium ~ source, data = WATER, alternative = "less")
TR
```

```

	Welch Two Sample t-test

data:  sodium by source
t = -1.8589, df = 22.069, p-value = 0.03822
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.3665724
sample estimates:
mean in group x mean in group y 
           76.4            81.2 
```

4. **Statistical Conclusion** — Since the p-value is 0.0382165, reject the null hypothesis.
5. **English Conclusion** — There is evidence to suggest the average sodium content for source X is less than the average sodium content for source Y.