Problem Set 6

Troy Jennings

1. Location of the mean (Crash Fatality Data)

The data set "USSeatBelts", data for the years 1983–1997 from 50 US States, plus the District of Columbia, for assessing traffic fatalities and seat belt usage, is in the "AER" package. Further details are available in the help for "USSeatBelts". These questions use the "state", "year", "fatalities", and "drinkage" variables. As detailed in the documentation, "fatalities" is the number of fatalities per million traffic miles and "drinkage" is a binary variable that is "yes" if the state had a minimum drinking age of 21 years and "no" otherwise.

As can be seen from the tabulation below, by 1988, all the jurisdictions adopted a minimum drinking age of 21 years. The data can be reformatted as shown to have columns for each year's values of "fatalities" and "drinkage".

```
data("USSeatBelts")
#table(USSeatBelts$year, USSeatBelts$drinkage)
dat <- USSeatBelts
dat <- pivot_wider(dat,id_cols=state,names_from = year,values_from = c(fatalities,drinkage))
```

1a.

Using the data frame "dat", perform a visual check of whether the value of "fatalities" in 1983 minus the value of "fatalities" in 1988 among the 29 jurisdictions that had a value of "no" for "drinkage" in 1983 could be considered Normally distributed. The function "ggqqplot" in the "ggpubr" package may help.

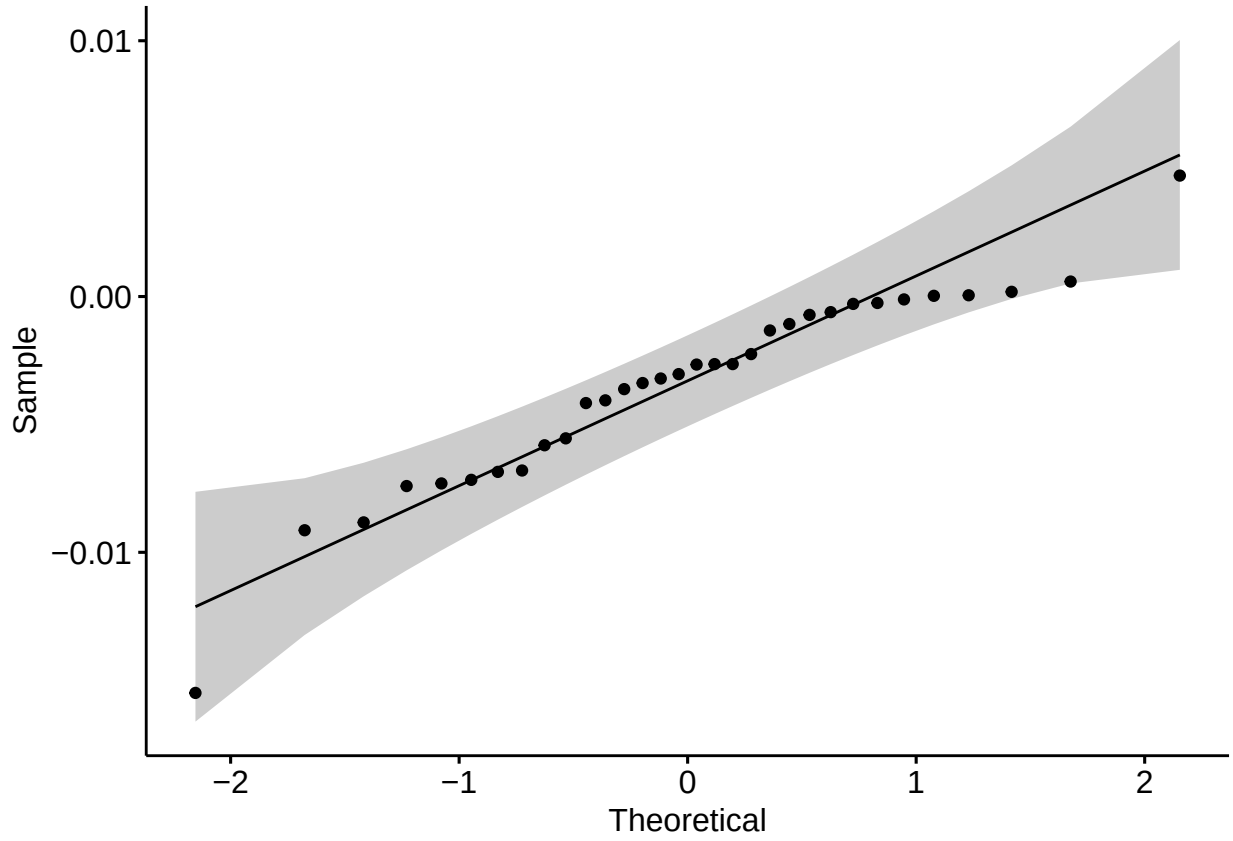
Question 1A Solution

```
## generate our dataframe with appropriate conditions
cols <- c("fatalities_1983", "fatalities_1988", "drinkage_1983", "drinkage_1988")
dat.2 <- select(dat, cols)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(cols)‘ instead of ‘cols‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
dat.2 <- filter(dat.2, drinkage_1983 == 'no')
measurement <- dat.2$fatalities_1988 - dat.2$fatalities_1983

## plot qq plot
g <- ggqqplot(data= measurement)
g
```



1b.

Using Student's t, test the hypothesis that the differences in "fatalities" between 1983 and 1988 for jurisdictions that went from "no" to "yes" in "drinkage" during this period are consistent with samples drawn from a Normal distribution with mean equal to 0. Please state your conclusions from the Student's t test including whether the test is a valid test of the location of the mean at 0.

In 1983, a lower drinking age than 21 was used by the states not having a minimum drinking age of 21. This analysis could be one step in examining the association between raised drinking age and traffic fatalities per million miles.

Question 1B Solution

```
## obtain the mean of the data
m <- mean(measurement)
mu <- 0
## obtain the standard deviation of the data
n <- length(measurement)
s <- (sd(measurement)/sqrt(n))

## automatically calculate the t-test
t.test(measurement - mu)
```

```
##
##  One Sample t-test
##
## data:  measurement - mu
## t = -5.0541, df = 31, p-value = 1.839e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.004861200 -0.002065882
## sample estimates:
##    mean of x
## -0.003463541
```

From the results of the t-test, we can see that the mean of the samples the differences in fatalities between 1983 and 1988 for jurisdictions that went from "no" to "yes" in drinking during the period are consistent with a Normal distribution, given the true mean for the data is very close to 0 (our hypothesized value).

1c.

What is the 99% confidence interval for the mean of these differences? Is this confidence interval consistent with a drop in the fatality rate between 1983 and 1988?

```
# ## manually calculate the confidence interval
a <- qt(.99, n-1)
lower <- m-a*s-mu
upper <- m+a*s-mu
cat("(", lower, ", ", upper, ")\n", sep= "")
```

```
## (-0.005144437, -0.001782644)
```

TODO: Answer the confidence interval question

1d.

Can you conclude that the increased drinking age caused a reduction in the fatality rate? The calculation below may help you think about this question.

```
fatal.diff.yes<-dat$fatalities_1983[dat$drinkage_1983=="yes"]-
          dat$fatalities_1988[dat$drinkage_1983=="yes"]
mean(fatal.diff.yes)
```

```
## [1] 0.003302584
```

2. Robustness of the z-test

A type 1 error in a hypothesis test is the rejection of the null hypothesis when it is true. For the z-test and the t-test, suppose the sampled population has the null distribution and you have a threshold p-value p below which you will reject the null hypothesis. For both tests, the probability of a type one error is exactly p.

The *power* of a hypothesis test is the probability of rejecting the null hypothesis when it is false. The power of the test depends on the way and extent to which the null hypothesis is false.

In the work below, you will compute the power of the z-test on data for which the population is $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ and the null hypothesis is that the sample is from a $Normal(\mu = 2\sqrt{2} - 1, \sigma^2 = 4)$ population. Then you will estimate the power of the z-test if the data are from a *Gamma* distribution with mean $2\sqrt{2}$ and variance 4 given the null hypothesis that the sample is from a $Normal(\mu = 2\sqrt{2} - 1, \sigma^2 = 4)$ population. You will estimate the power of the z-test if the sample is from a $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ distribution but the values are rounded using the null hypothesis is that the sample is from a $Normal(\mu = 2\sqrt{2} - 1, \sigma^2 = 4)$ population.

You will also estimate the probability of a type 1 error using the z-test on data from a *Gamma* distribution with mean $2\sqrt{2}$ and variance 4 given the null hypothesis that the sample is from a $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ population. You will estimate the probability of a type 1 error using the z-test on data from a $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ distribution given the null hypothesis that the sample is from a $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ population.

The goal is to gain an understanding of the extent to which the z-test remains a valid test of the location of the mean under these violations of the assumptions of the z-test as a test of the location of the mean.

The shape, scale, mean, and variance variables defined below are arranged so that changing the shape value will allow you to explore other *Normal* and *Gamma* distributions while retaining the property that they have the same mean and both have variance equal to 4. Thus a difference of 1 in the true and the hypothesized mean remains a difference of 0.5 standard deviations. A difference of this size is considered medium-sized by some rules of thumb.

2.a

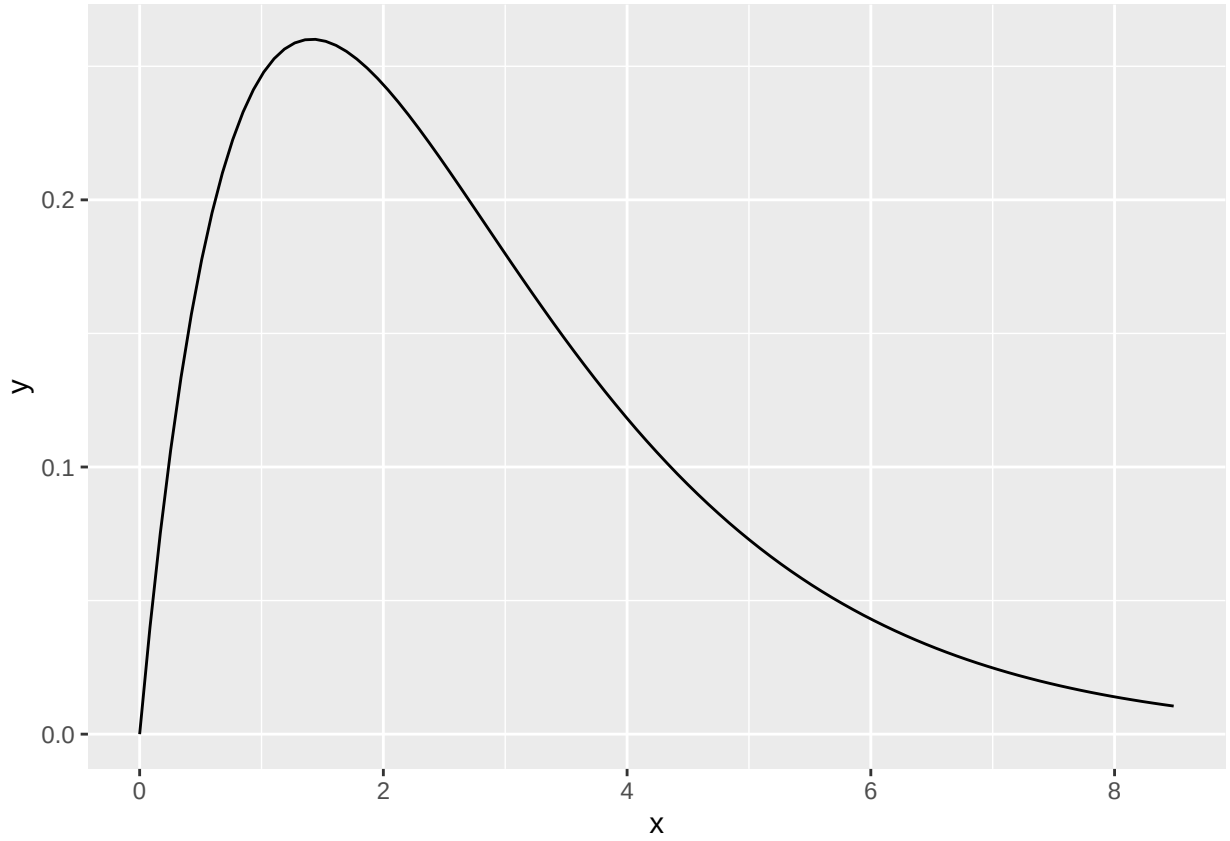
Suppose an iid sample of size 50 is drawn from a population with a $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ distribution. Let the null hypothesis be that the sample is from a $Normal(\mu = 2\sqrt{2} - 1, \sigma^2 = 4)$ population. What is the probability that a two-sided z-test of performed on the sample will have a p-value that is less than or equal to 0.01? Please compute the result exactly using "pnorm", though you may check your answer using simulations. You may want to start by finding the set of values for the z-statistic that result in a p-value of less than or equal to 0.01.

```
shp<-2
scl<-sqrt(4/shp)
sig<-sqrt(shp*scl^2) # sigma in 2a
mu<-shp*scl # 2*sqrt(2), mu in 2a
```

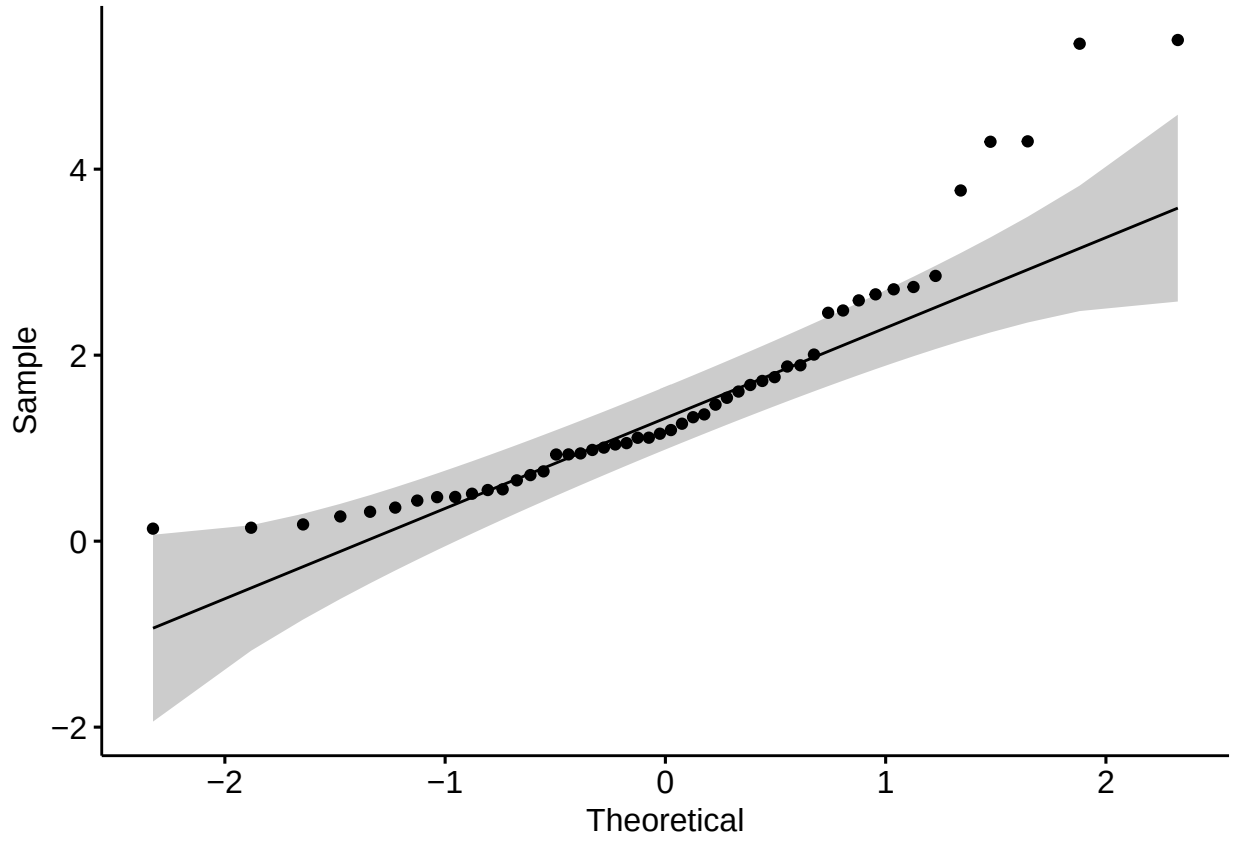
2.b

Suppose an iid sample of size 50 is drawn from population with a $Gamma(shape = 2, scale = \sqrt{2})$ distribution. Note that the mean of this distribution is $2\sqrt{2}$ and the variance is 4, as in the Normal population above. Let the null hypothesis be that the sample is drawn from a $Normal(\mu = 2\sqrt{2} - 1, \sigma^2 = 4)$ population. Please use 100,000 samples to estimate the probability that a two-sided z-test of performed on the sample will have a p-value that is less than or equal to 0.01.

```
dat.plot<-data.frame(x=c(0,3*mu))
ggplot(data=dat.plot,aes(x=x))+
  stat_function(fun=dgamma, args=list(shape=shp,scale=scl))
```



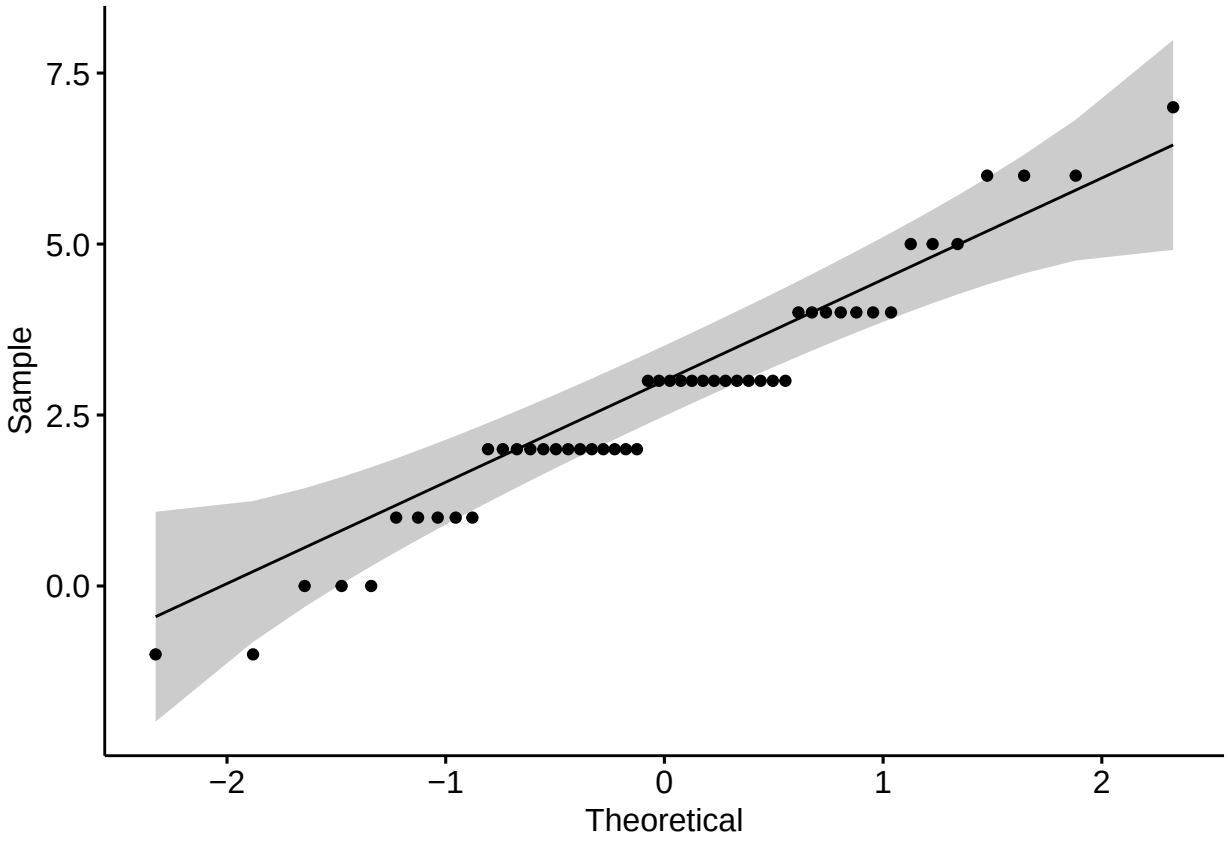

```
ggqqplot(rgamma(50,shp,scl))
```



2.c

Suppose an iid sample of size 50 is drawn from population with a $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ distribution but that the values are rounded to the nearest 0.1 (see the "round" function). Let the null hypothesis be that the sample is drawn from a $Normal(\mu = 2\sqrt{2} - 1, \sigma^2 = 4)$ population. Please use 100,000 samples to estimate the probability that a two-sided z-test of performed on the sample will have a p-value that is less than or equal to 0.01.

```
n<-50
set.seed(1234567)
ggqqplot(round(rnorm(n,mu,sig)))
```



2.d

Suppose an iid sample of size 50 is drawn from population with a $Gamma(shape = 2, scale = \sqrt{2})$ distribution. Note that the mean of this distribution is $2\sqrt{2}$ and the variance is 4, as in the Normal population above. Let the null hypothesis be that the sample is drawn from a $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ population. Please use 100,000 samples to estimate the probability that a two-sided z-test performed on the sample will have a p-value that is less than or equal to 0.01?

2.e

Suppose an iid sample of size 50 is drawn from population with a $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ distribution but that the values are rounded to the nearest 0.1 (see the "round" function). Let the null hypothesis be that the sample is drawn from a $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ population. Please use 100,000 samples to estimate the probability that a two-sided z-test performed on the sample will have a p-value that is less than or equal to 0.01? (Note that this probability is, by definition, 0.01 if the values weren't rounded.)

2.f

Nonnormality

Does the correctness of the p-value and the power of the test seem to be strongly affected by the change from the $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ population to the $Gamma(shape = 2, scale = \sqrt{2})$ population?

Rounding

Does the correctness of the p-value and the power of the test seem to be strongly affected by the change from the $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ population to the rounded $Normal(\mu = 2\sqrt{2}, \sigma^2 = 4)$ values?