# MATH 392 Problem Set 1

*January 26, 2018*

## Problem 1.1

(a) Population would be (probably American) high school students, and the sample is 2000 simple random samples from the population.

The fact that 47% of the high school students watched Glee is a **statistic**.

(b) Population is whole U.S. population.

The fact that 13.9% of the population is between the ages of 15 and 24 is a **parameter**, assuming that the U.S. Census is really a census.

(c) Population is all NBA players.

Average height of 78.93 in is a **parameter**.

(d) Population is all U.S. adults (18 years or older). The sample is 1025 adults selected by Gallup, possibly simple random samples from the population (that Gallup has information).

The fact that 47% would advise their member of Congress to vote for health care legisation is a **statistic**.

## Problem 1.3

(a) As this study is a survey, it is an **observational study**.

(b) No. Because it is not an experiment (i.e. control group and treatment group are not systematically separated) we should not draw conclusion from this study.

(c) No. The study requires to use simple random sampling in order to generalize the result to whole population, but it is only sampled from 3 high schools.

## Problem 1.5 (in the book)

We know that number of $n$ sized subset of $N$ is

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

Among all the elements, choose one and name it $a$. Total number of subsets that includes the element $a$ is

$$\binom{N-1}{n-1} = \frac{(N-1)!}{(n-1)!(N-n)!}.$$

Finally, probability of choosing a subset that includes an element $a$ from an $N$ sized set is

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

which is what we desired to show.

## Problem 1.5 (in the handout)

(a) Since $n = 1000$ and $N = 100000000$, from the above problem the probability I will be selected as a sample is $1/100000 = 0.00001$.

(b) $(1 - 0.00001)^{2000} = 0.9801986$.

(c) Solve following equation:

$$(1 - 0.00001)^x = 0.5$$

which reduces to $x = \log(0.5)/\log(0.99999) = 69314.37$. So we should collect 69315 samples to have a 0.5 probability.

## Problem 2.4

(a)

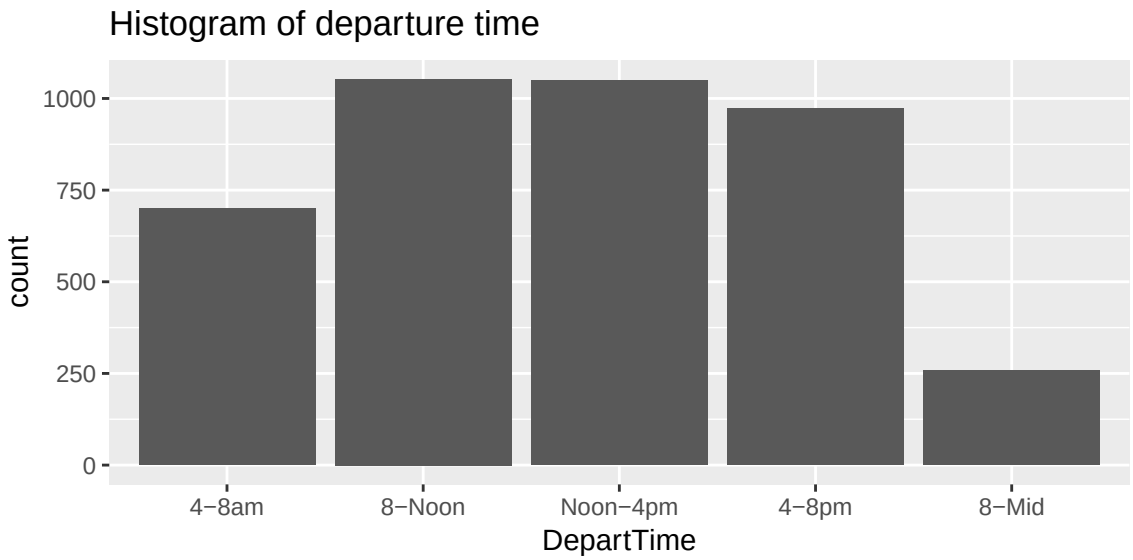
```
departtime_table <- FlightDelays %>%
  group_by(DepartTime) %>%
  summarise(counts= n())

departtime_table
```

```
## # A tibble: 5 x 2
##   DepartTime counts
##   <fct>       <int>
## 1 4-8am         699
## 2 8-Noon       1053
## 3 Noon-4pm     1048
## 4 4-8pm         972
## 5 8-Mid         257
```

Using **group_by**, the dataset was rearragned and **n()** was used to count the number of flights.

Below is the bar chart of the departure times.

```
FlightDelays %>%
  ggplot(aes(x=DepartTime)) +
  geom_bar() +
  ggtitle("Histogram of departure time")
```




(b)

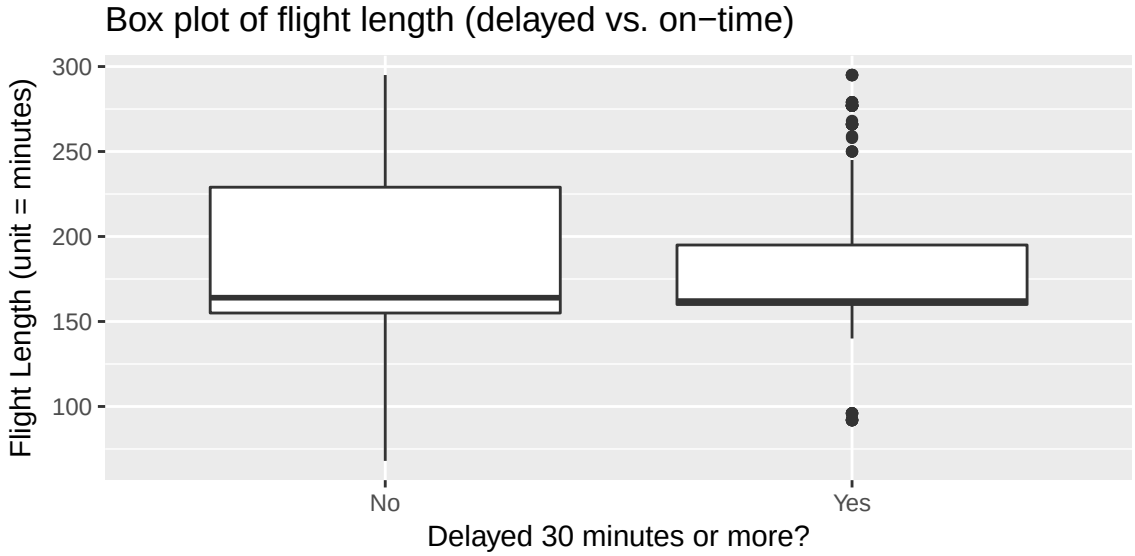
```
contingency <- table(FlightDelays$Day, FlightDelays$Delayed30)

contingency_new <- cbind(contingency, contingency[,1] +
contingency[,2], contingency[,2]/(contingency[,1] + contingency[,2]))

contingency_new
```

```
##      No Yes
## Sun 507  44 551 0.07985
## Mon 569  61 630 0.09683
## Tue 535  93 628 0.14809
## Wed 488  76 564 0.13475
## Thu 434 132 566 0.23322
## Fri 493 144 637 0.22606
## Sat 406  47 453 0.10375
```

Above is the contingency table, and the last column represents the proportion of flights delayed at least 30 minutes. So Sunday - 8%, Monday - 10%, Tuesday - 15%, Wednesday - 13%, Thursday - 23%, Friday - 23%, Saturday - 10% are the proportions of flight delays for each day in a week.

(c)

```
FlightDelays %>% ggplot(aes(x = Delayed30, y = FlightLength)) +
  geom_boxplot() +
  ylab("Flight Length (unit = minutes)") + xlab("Delayed 30 minutes or more?") +
  ggtitle("Box plot of flight length (delayed vs. on-time)")
```




(d)

While I cannot be too sure, there seems to exist a relationship that for the flights that were delay, less flights tend to fly more than 200 minutes (fly slower), unlike when the flight was not delay. It is reasonable as the airlines would prefer their planes to arrive on-time.

## Problem 2.8

(a)

The CDF of the exponential distribution is $F(x) = 1 - e^{-\lambda x}$. $q_1, q_2(= m), q_3$ are first quartile, median, and third quartile if and only if $F(q_n) = 0.25 \times n$. Therefore

$$1 - e^{-\lambda q_1} = 0.25,$$

so

$q1 = -log(0.75)/\lambda$. Similarly, $m = -log(0.5)/\lambda$ and $q_3 = -log(0.25)/\lambda$.

(b)

We first need to find its CDF.

$$F(x) = \int_1^x \frac{\alpha}{t^{\alpha+1}} dt = -\frac{1}{x^\alpha} + 1.$$

Then next step is same as (a).

$$-\frac{1}{{q_1}^\alpha} + 1 = 0.25,$$

which makes $q_1 = (\frac{4}{3})^{1/\alpha}$. Similarly, $m = 2^{1/\alpha}$ and $q_3 = 4^{1/\alpha}$.

## Problem 2.9

By slightly modifying the definition of a quantile, we obtain

$$F(q_p) = 1 - \frac{9}{q_p^2} = p,$$

and therefore,

$$q_p = \frac{3}{\sqrt{1-p}} \text{ (since } x \text{ is positive).}$$

We obtained an expression for $p$th quantile.

## Problem 2.10

PMF of $X$ is

$$f(x) = \binom{20}{x}(0.3)^x(0.7)^{20-x}.$$

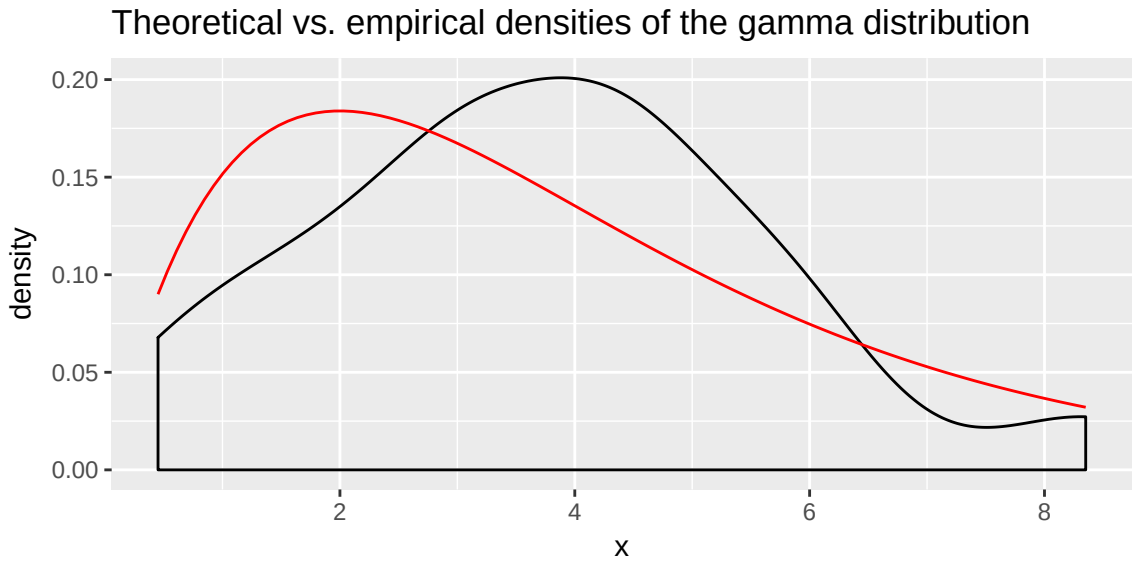
We can list first few probabilities of $X$: $P(X=0) = 0.0008, P(X=1) = 0.007, P(X=2) = 0.029, P(X=3) = 0.071, P(X=4) = 0.13$.

As we add from $P(X=0)$ (to obtain CDF $F(x)$), wee see that $F(2) = 0.035$ and $F(3) = 0.11$. Because $X$ is a discrete random variable, $F(x)$ is a step function, which means that there exists no $q$ such that $F(q) = 0.05$ since CDF is always increasing.

## Problem 2.11

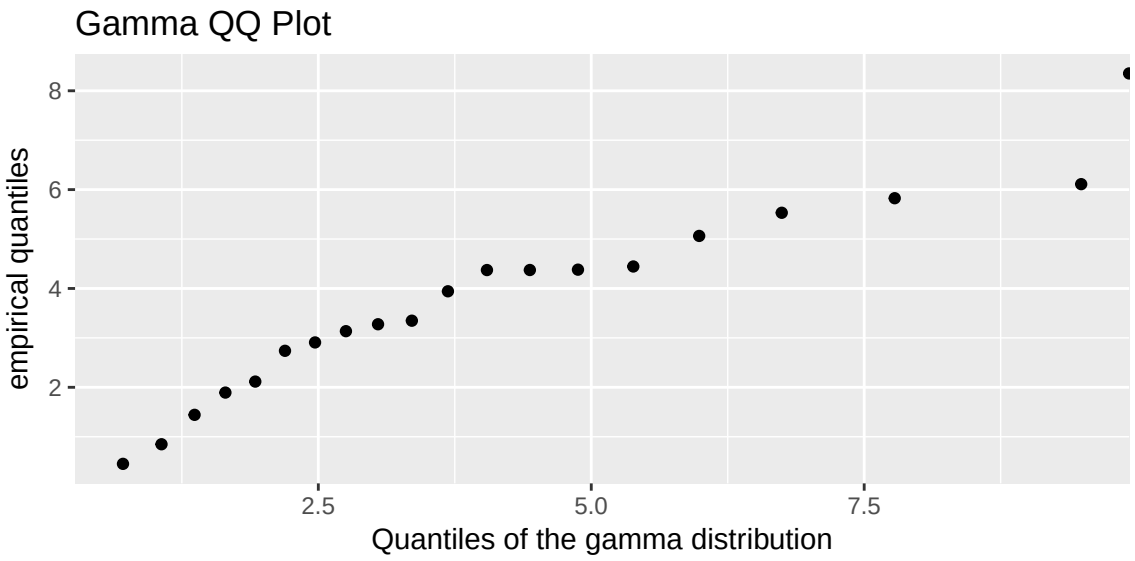
1.

```
n1 <- 20
x1 <- rgamma(n1, shape = 2, scale = 2)
df1 <- data.frame(x = x1)
ggplot(df1, aes(x=x)) + geom_density() +
  stat_function(fun=dgamma, args = list(shape = 2, scale =2), col = "red") +
  ggtitle("Theoretical vs. empirical densities of the gamma distribution")
```



**rgamma** was used to generate 20 random samples from gamma distribution. **ggplot** was used to draw the density plot. The black line represents the simulated density plot and the red line represents the PDF of the gamma distribution. Two plots do not coincide very well, but I'd say they look quite OK since only 20 samples were taken.

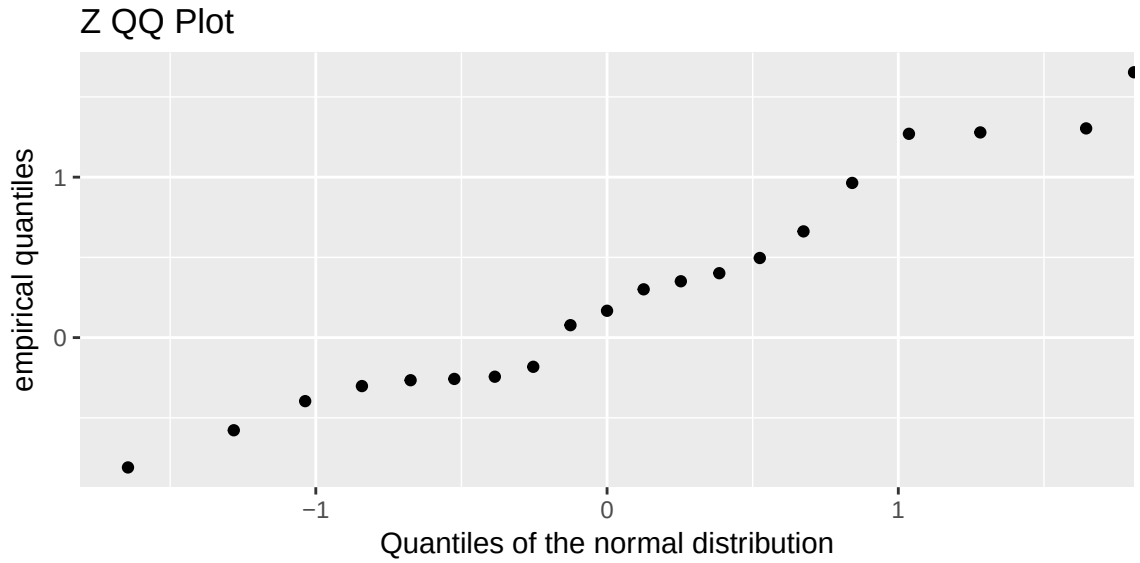
```
df2 <- data.frame(x= qgamma(1:n1/n1, shape = 2, scale = 2), y = sort(x1))

ggplot(df2, aes(x=x, y=y)) + geom_point() +
  xlab("Quantiles of the gamma distribution") + ylab("empirical quantiles") +
  ggtitle("Gamma QQ Plot")
```



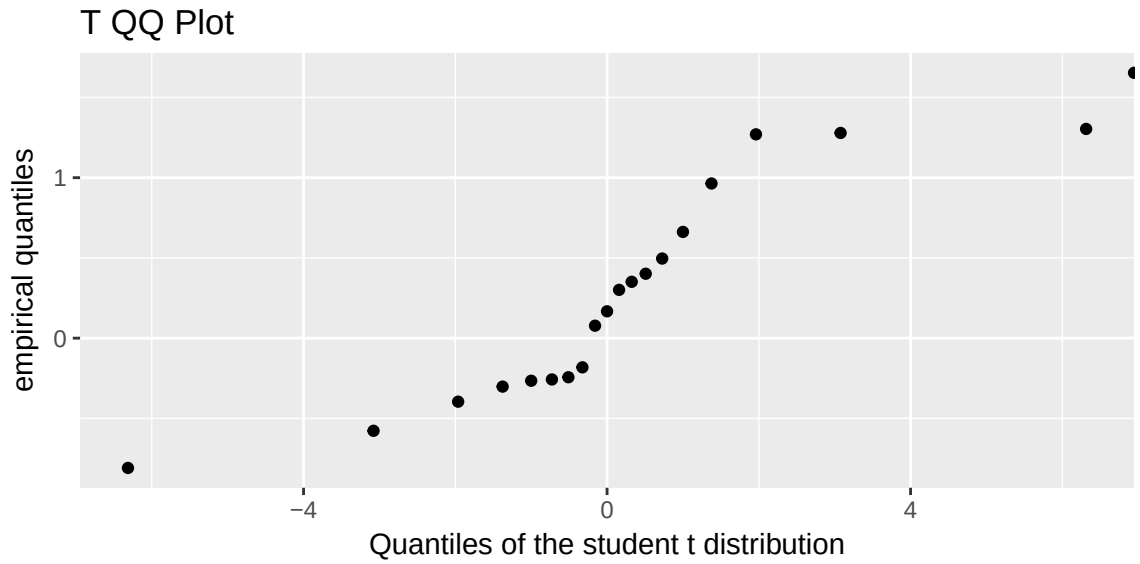

The code provided in the slide was modified; instead of drawing normal QQ plot, gamma QQ plot was drawn by replacing **qnorm** with **qgamma**. We observe that the QQ plot is vaguely linear.

2.

```
n2 <- 20
x2 <- rnorm(n2)
df3 <- data.frame(x = qnorm(1:n2/n2), y = sort(x2))

ggplot(df3, aes(x=x, y=y)) + geom_point() +
  xlab("Quantiles of the normal distribution") + ylab("empirical quantiles") +
  ggtitle("Z QQ Plot")
```




```
df4 <- data.frame(x = qt(1:n2/n2, df= 1), y = sort(x2))

ggplot(df4, aes(x=x, y=y)) + geom_point() +
  xlab("Quantiles of the student t distribution") + ylab("empirical quantiles") +
  ggtitle("T QQ Plot")
```




20 random samples from the normal distribution were drawn. Among the two QQ plots, the normal QQ plot seems more linear than the t QQ plot. t QQ plot might have been more accurate if we had smaller sample size.