Problem Set 1 – MATH 392

Theodore Dounias

1/25/2018

1.1

a) Population is all high school students, sample is 2000 students, 47% is a statistic.

b) From what I recall the 2000 census was still just basically asking everyone, so the US population is the Population, and 13,9% is a parameter.

c) This only draws a reslt for the 2006-2007 season, so the population is all players in that season, and 78,93 is a parameter. If we were using this season as a sample for all seasons, then it would be a statistic.

d) The Population is all US adults, 1.025 is the sample size, 47% is a statistic.

1.3

a) An experiment, as the researchers involved had input into the environment of the study.

b c) No to b, because no to c. Generally a study needs to be valid and ratifiable. This means that the methods and processes used should be scientificaly rigorous, and that it can be reproduced in a different environment. Here we have one study, on a sample that is significantly smaller than the population, and lacking a relevant control group (people on a normal–for a diabetic at least–diet). Therefore, the study is neither valid or ratifiable at this point, so no conclusions can be safely drawn.

1.5

a) 0,00001

b)

```
(1 - 0.00001)^2000
```

```
## [1] 0.9801986
```

c) About 69.314 samples must be selected

```
(1 - 0.00001)^69314
```

```
## [1] 0.5000019
```

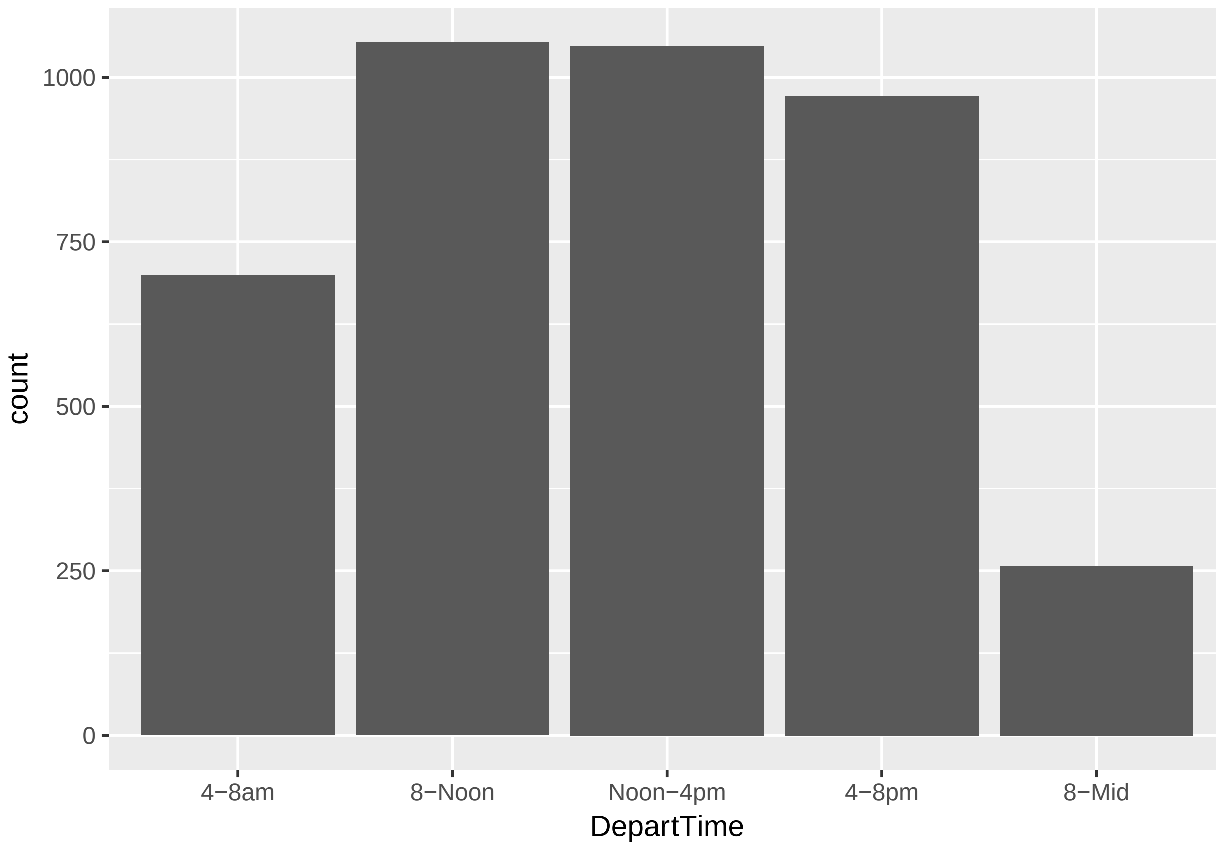
2.4

a)

```
depart <- FlightDelays %>%
  group_by(DepartTime) %>%
  mutate(count = 1) %>%
  summarise(n = sum(count))

depart
```

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

```
ggplot(FlightDelays, aes(DepartTime)) +
  geom_bar()
```



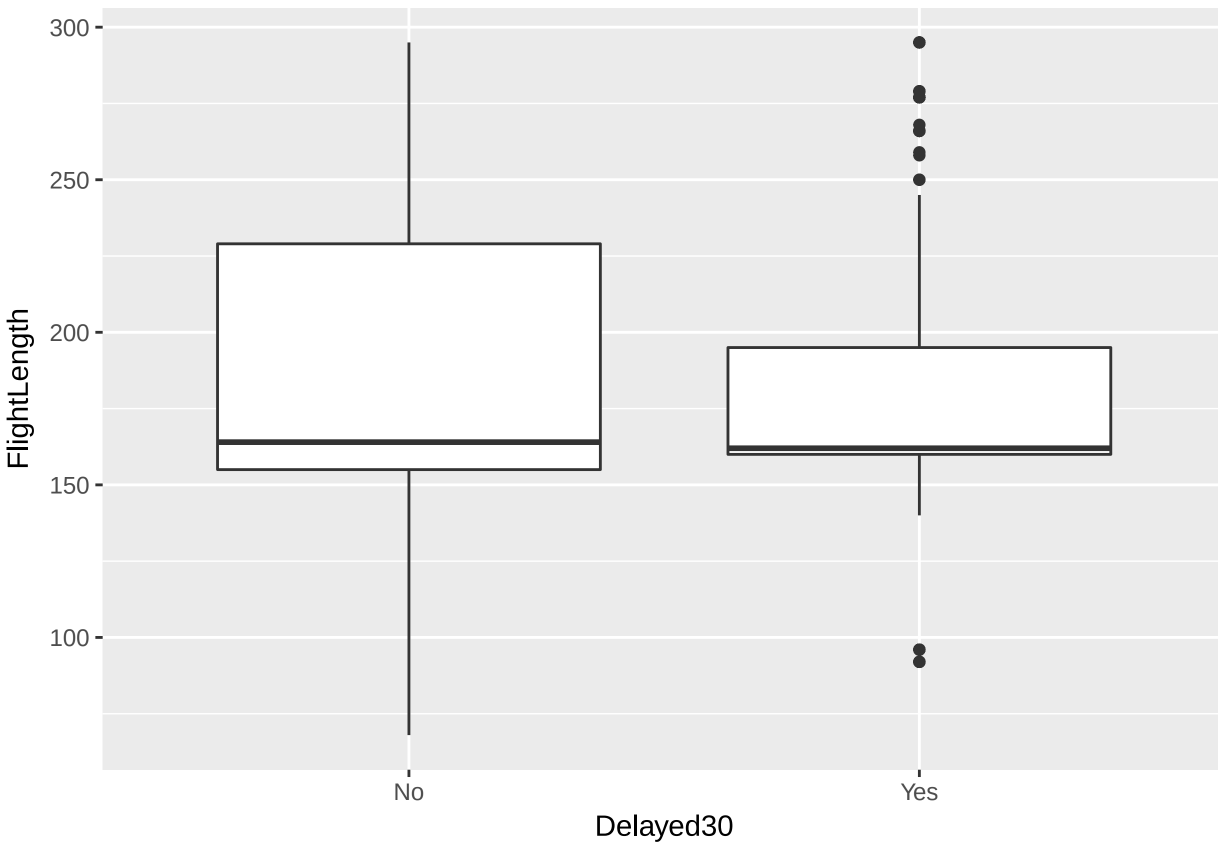
b)

```
day_con <- FlightDelays %>%
  select(Day, Delayed30)

day_con$Delayed30 <- as.integer(day_con$Delayed30)

day_con$Delayed30 <- day_con$Delayed30 - 1

day_con <- day_con %>%
  group_by(Day) %>%
  mutate(count = 1) %>%
  summarise(Delayed = sum(Delayed30), Not_Delayed = sum(count) - sum(Delayed30),
            Proportion = sum(Delayed30)/sum(count))

day_con
```

```
## # A tibble: 7 x 4
##   Day   Delayed Not_Delayed Proportion
##   <fct>   <dbl>       <dbl>      <dbl>
## 1 Sun      44.0         507     0.0799
## 2 Mon      61.0         569     0.0968
## 3 Tue      93.0         535     0.148
## 4 Wed      76.0         488     0.135
## 5 Thu     132           434     0.233
## 6 Fri     144           493     0.226
## 7 Sat      47.0         406     0.104
```

c)

```r
ggplot(FlightDelays) +
  geom_boxplot(aes(y = FlightLength, x = Delayed30))
```



d) Large–I assume intercontinental and international–flights seem to be delayed significantly less, and all the delays centered on 3-4 hour flights. This might be because of international conections, or because of airports prioritizing longer flights, and not caring that much about small local ones, when they assign runways to planes.

2.8

a) Integrating by parts we get:

$$E(X) = \int \lambda x e^{-\lambda x} dx = -xe^{-\lambda x}\Big|_0^\infty - \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}$$

Here we can use the integral definition of quantile to get:

$$\int_0^a \lambda e^{-\lambda x} dx = p \Rightarrow 1 + e^{-\lambda a} = p \Rightarrow a = \frac{-ln(1-p)}{\lambda}$$

So the quantiles are $\frac{ln(\frac{4}{3})}{\lambda}$ and $\frac{ln(4)}{\lambda}$

b) We have:

$$E(X) = \int_1^\infty \frac{a}{x^a} dx = \frac{ax^{1-a}}{1-a}\Big|_1^\infty = \begin{cases} \infty & a \leq 1 \\ \frac{a}{a-1} & a < 1 \end{cases}$$

Again for the quantile:

$$\int_1^b \frac{a}{x^{a+1}} dx = p \Rightarrow 1 - k^{-a} = p \Rightarrow k = (1-p)^{-1/a}$$

So we have the quantiles $3/4^{-1/a}$ and $1/4^{-1/a}$.

2.9

Using exactly the same method used above, which I will omit due to LaTeX repetitive typing, I end up with the expression $p_{th} quantile = \frac{3}{\sqrt{1-p}}$

2.10

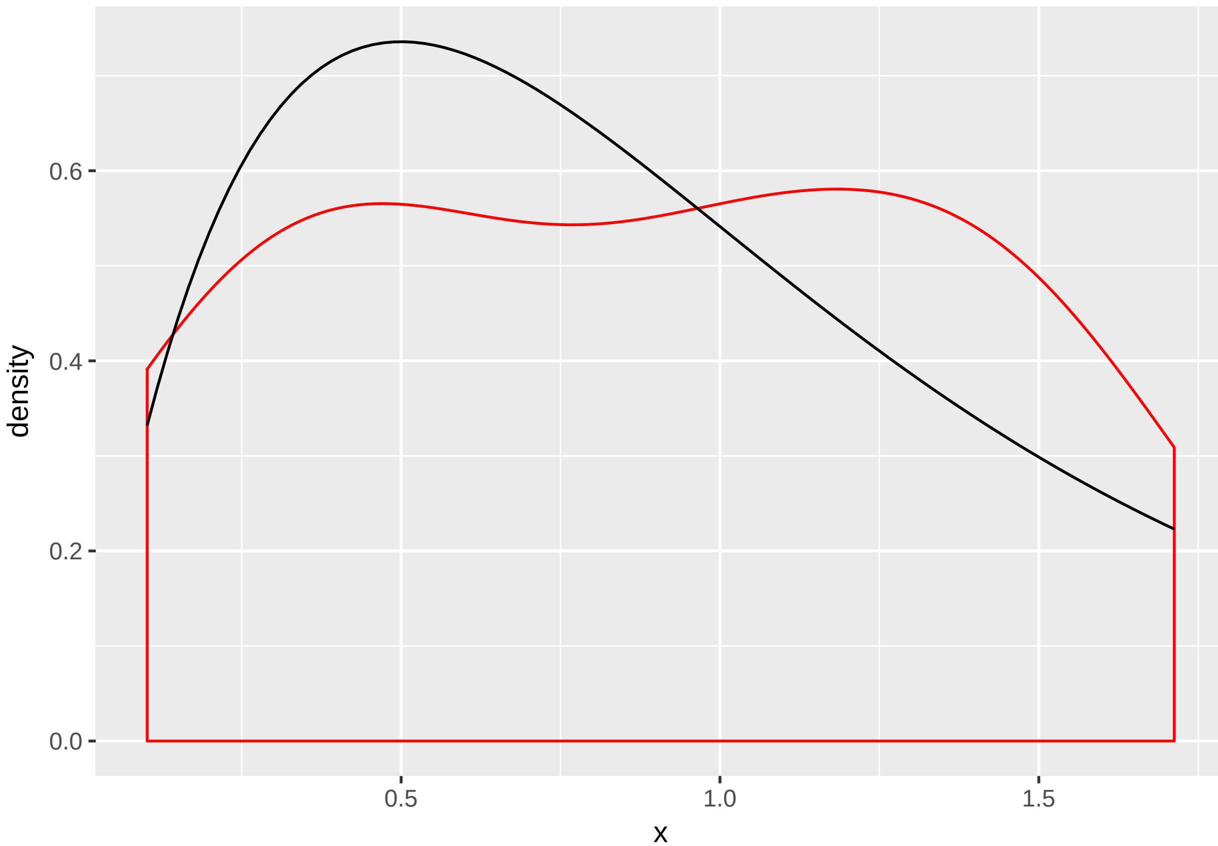
```r
qbinom(0.05, 20, 0.3)
```

```
## [1] 3
```

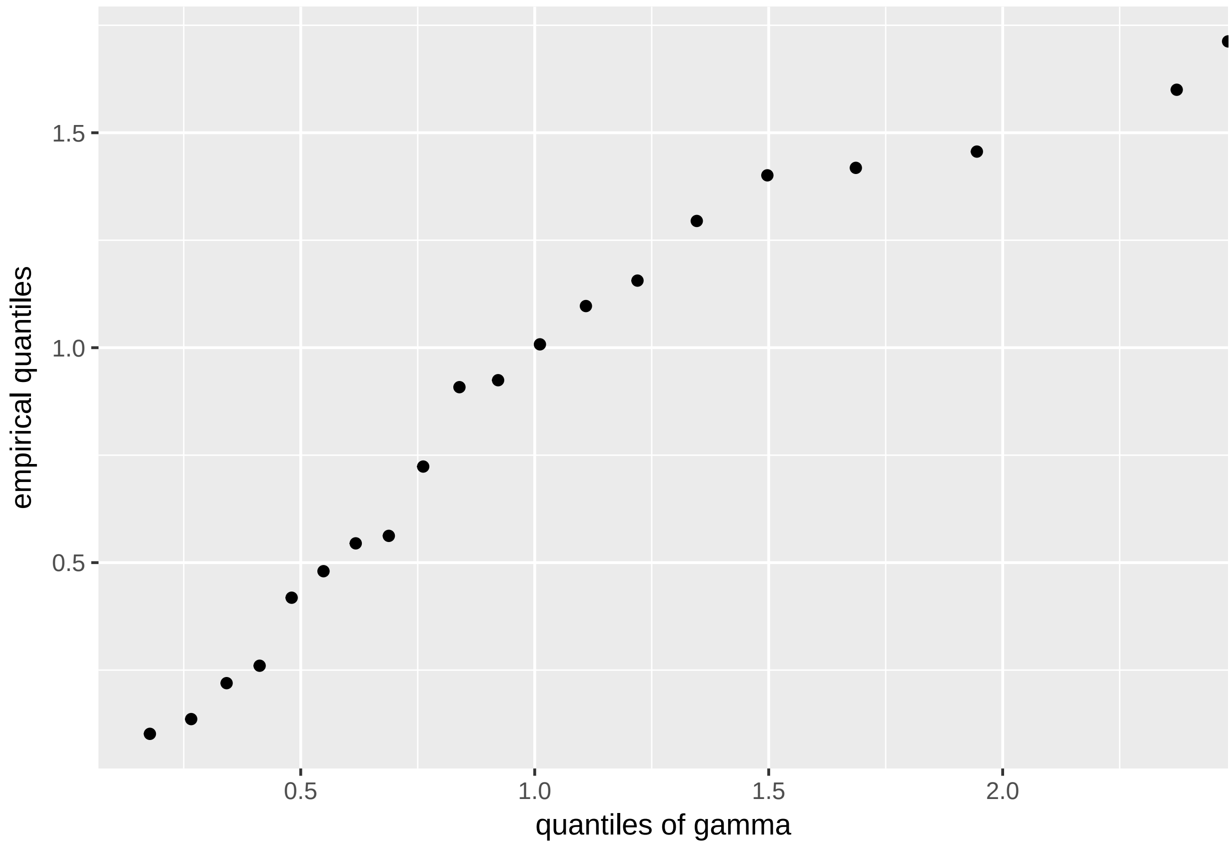
2.11

a)

```r
s <- rgamma(20, 2, 2)

df <- data.frame(x = s)

ggplot(df, aes(x = x)) +
  geom_density(col = "red") +
  stat_function(fun = dgamma, args=list(shape=2, rate=2))
```

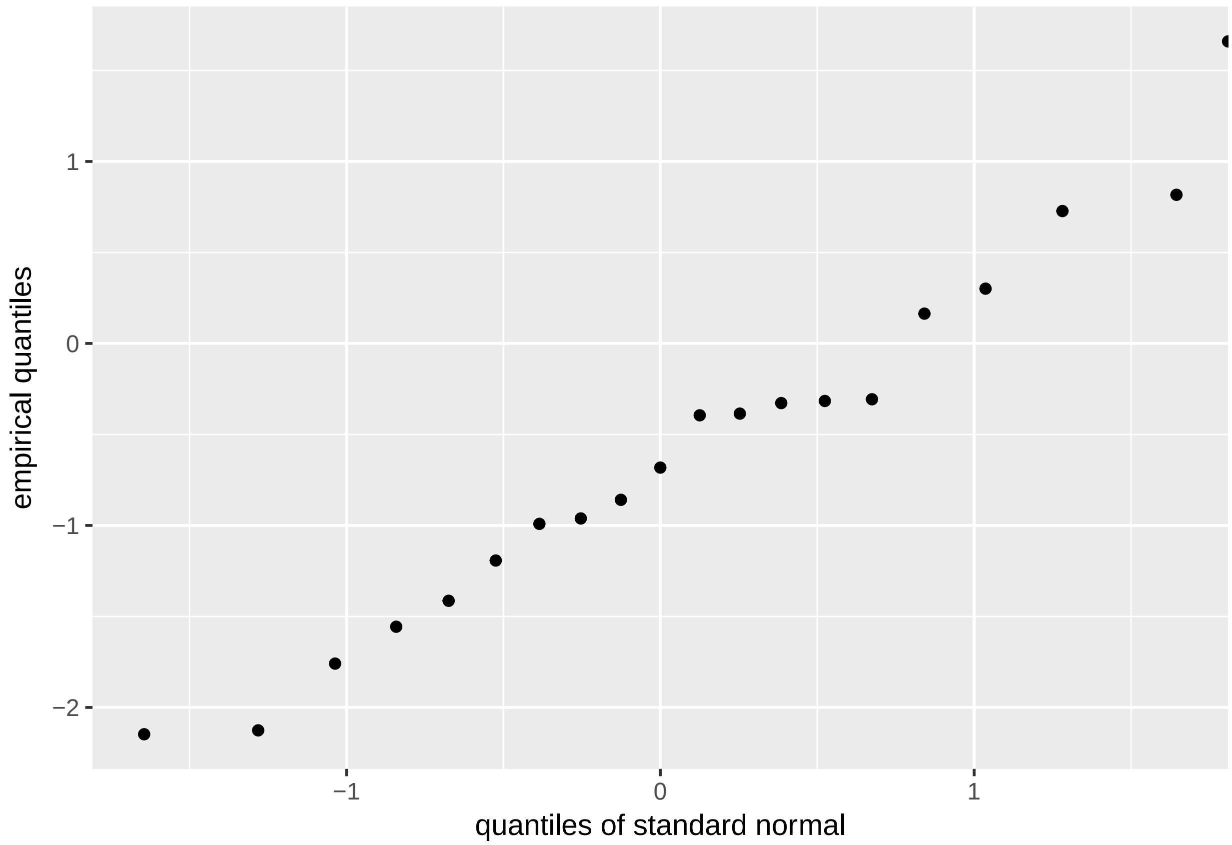



```r
n<-20

df <- data.frame(x = qgamma(1:n/n, 2, 2),
                 y = sort(s))
ggplot(df, aes(x = x, y = y)) +
  geom_point() +
  xlab("quantiles of gamma") +
  ylab("empirical quantiles")
```

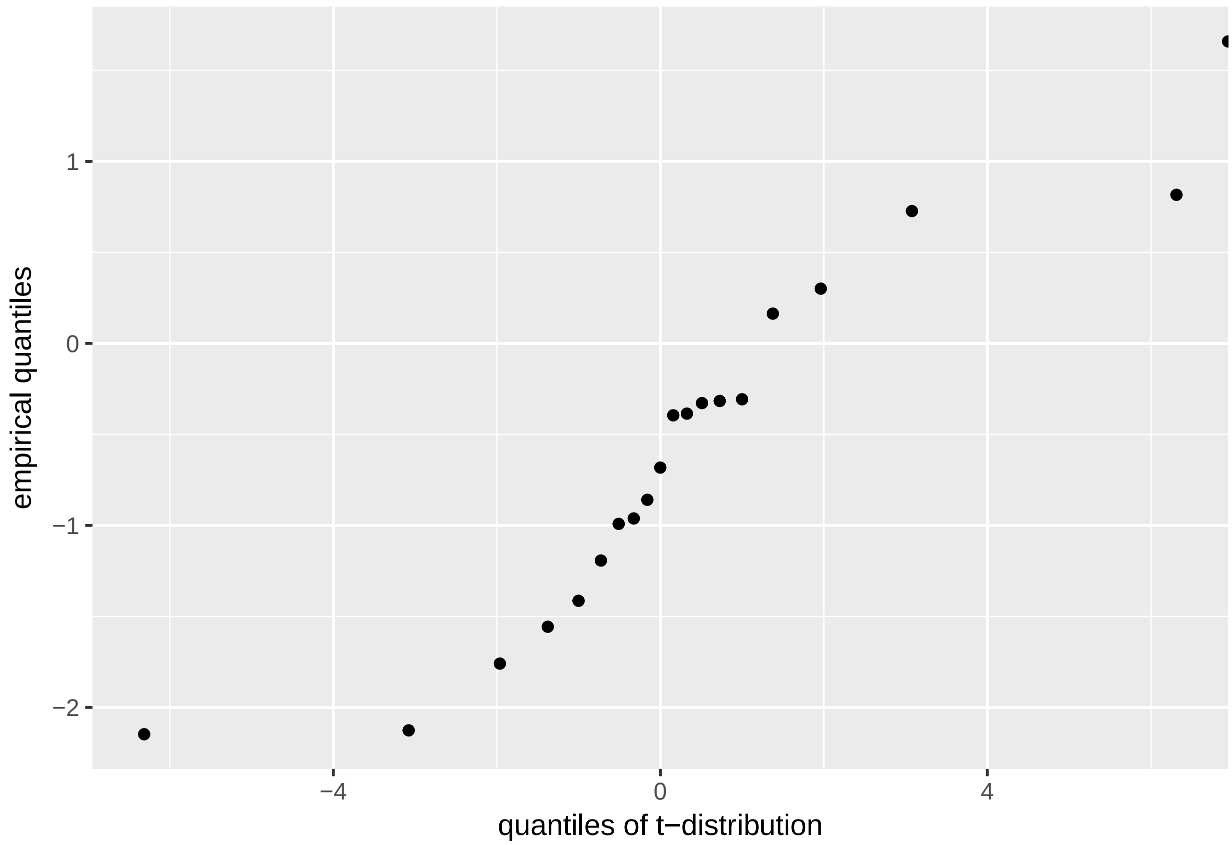


b)

```
#I know this is your code but it is literally just the same thing asked for

n <- 20
x <- rnorm(n)
df <- data.frame(x = qnorm(1:n/n),
                 y = sort(x))
ggplot(df, aes(x = x, y = y)) +
  geom_point() +
  xlab("quantiles of standard normal") +
  ylab("empirical quantiles")
```




```r
df <- data.frame(x = qt(1:n/n, 1),
                 y = sort(x))
ggplot(df, aes(x = x, y = y)) +
  geom_point() +
  xlab("quantiles of t-distribution") +
  ylab("empirical quantiles")
```



empirical quantiles
1
0
−1
−2
−4
0
4
quantiles of t−distribution