Recitation Section 6: Q–Q Plots

1 Review

Suppose that we are given data sample $x_{(1)}, \ldots, x_{(n)}$ that have been sorted in increasing order.

- We assign to $x_{(i)}$, i.e. the i^{th} smallest observation, the number

  $$q_i = \frac{i - 0.5}{n}.$$

  Then, $x_{(i)}$ is called the q_i sample quantile.
  (*Why* $\frac{i-0.5}{n}$*? Imagine dividing the interval* $[0, 1]$ *into* n *equal sized subdivisions. Then, the* i*-th subdivision would start at* $\frac{i-1}{n}$ *and end at* $\frac{i}{n}$*. The center of the subdivision is exactly* $\frac{i-0.5}{n}$*. In the context of the data, the interval* $[0, 1]$ *represents the idea of 0% to 100% of your datapoints.* q_i *is a convention for conveying what percentage of all the datapoints that* $x_{(i)}$ *is bigger than. In other words, we can use it to say that* $x_{(i)}$ *is larger than about* $(q_i \times 100)$*% of all datapoints.*)

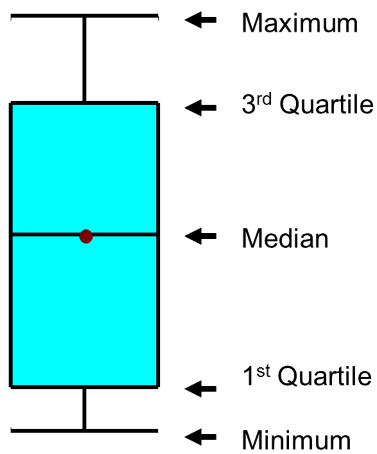
- Given a random variable with CDF $F(x)$ that we think may fit the data, a *Q–Q plot* is a scatterplot containing the points

  $$\left(F^{-1}\Big(\frac{i - 0.5}{n}\Big), x_{(i)}\right) \qquad i = 1, \ldots, n$$

  where $F^{-1}(x)$ is the *inverse CDF*: the function satisfying $F(F^{-1}(x)) = x$. The points should lie near the line $y = x$ if the fit is good.
  (*Remember that the CDF,* $F(x)$*, gives the answer to "what is the probability that a random sample is less than* x*?" The inverse CDF,* $F^{-1}(q)$*, then gives the answer to "what is the* x *value that the random sample would be less than exactly* $(q \times 100)$*% of the time?" If you read that to yourself a few times, you will realize that the two questions are exactly the reverse (inverse!) of each other. You will also notice that the argument of (i.e. input of the function)* F^{-1} *corresponds exactly to the idea of* q_i *from above.*)

2 Exercises

1. In earlier classes, you were introduced to the *box and whiskers plot.* That looks like this:



Five components of the above plots are the following:

- *Maximum:* The largest datapoint.
- *Minimum:* The smallest datapoint.
- *Median:* The data point that is in the middle of the sorted values. If you have an even number of data points, it is the average of the middle two values.
- *First Quartile:* The data point that is in the middle of the values between the median and the smallest value. If there are two values in the middle, take their average.
- *Third Quartile:* The data point that is in the middle of the values between the median and the largest value. If there are two values in the middle, take their average.

Roughly, what *quantiles* do these five components correspond to?

Minimum is 0 quantile, first quartile is 0.25 quantile, median is 0.5 quantile, third quartile is 0.75 quantile, maximum is 1 quantile.

2. Consider the following 10 data points:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
-141	246	62	131	94	-81	-14	-78	177	-52

(a) Sort these data points in ascending order and call the sorted data points $x_{(1)}, ..., x_{(10)}$. Fill out the table below:
(This is to get you familiar with the notation)

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
-141	-81	-78	-52	-14	62	94	131	177	249

(b) Suppose that associated with each x_i is a rank r_i that tells you the position of x_i when the data is sorted in ascending order. For example, $r_1 = 1$ because x_1 is the smallest data point, and $r_2 = 10$ because x_2 is the largest data point. Fill out the table below:

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
1	10								

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
1	10	6	8	7	2	5	3	9	4

(c) Write a simple relationship between r_i, x_i, and $x_{(i)}$.
$x_i = x_{(r_i)}$

(d) Use the notation q_i to indicate that x_i is the q_i-quantile. Write q_i in terms of r_i:
$q_i = \frac{r_i - 0.5}{10}$. So we can interpret q_i as the "relative/normalized rank".

3. Match each of the following terms with their corresponding properties:
"left skewed", "right skewed", "heavy tailed", "light tailed"

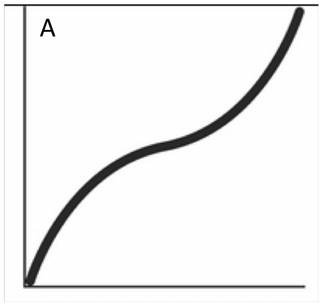
(a) Right tail of data heavier than theoretical distribution;
Left tail of data lighter than theoretical distribution.
right skewed

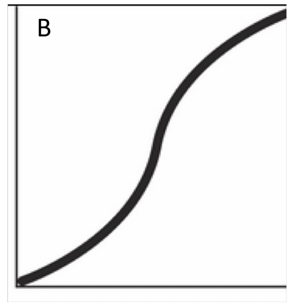
(b) Right tail of data lighter than theoretical distribution;
Left tail of data heavier than theoretical distribution.
left skewed

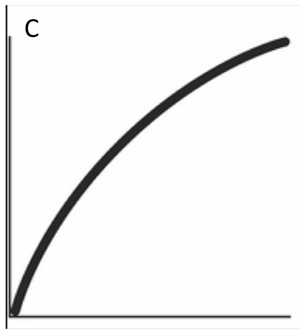
(c) Right tail of data lighter than theoretical distribution;
Left tail of data lighter than theoretical distribution.
light tailed

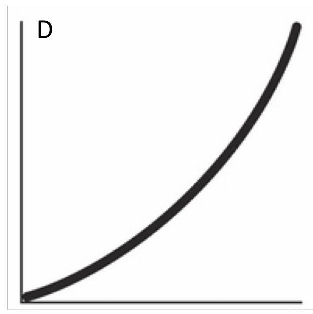
(d) Right tail of data heavier than theoretical distribution;
Left tail of data heavier than theoretical distribution.
heavy tailed

4. Match each of the following plots with their corresponding properties:
"left skewed", "right skewed", "heavy tailed", "light tailed"









A: Heavy tailed, B: light tailed, C: left-skewed, and D: right-skewed

5. Consider the data sample 0.08, 0.54, 1.13, 1.57, 1.74. We hypothesize that these observations may originate from the Uniform(0, 2) distribution. Let $F(x)$ denote the CDF of this random variable.

   (a) Find $F^{-1}(x)$ for $0 \leq x \leq 2$.

   Since, F is the CDF of a Uniform(0, 2) distribution, we have

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{2} & 0 \leq x \leq 2 \\ 1 & x > 2 \end{cases}.$$

   Hence, $F^{-1}(x) = 2x$ for $0 \leq x \leq 2$.

   (b) Compute the theoretical and sample quantiles. Try to manually sketch a Q–Q plot of the data.

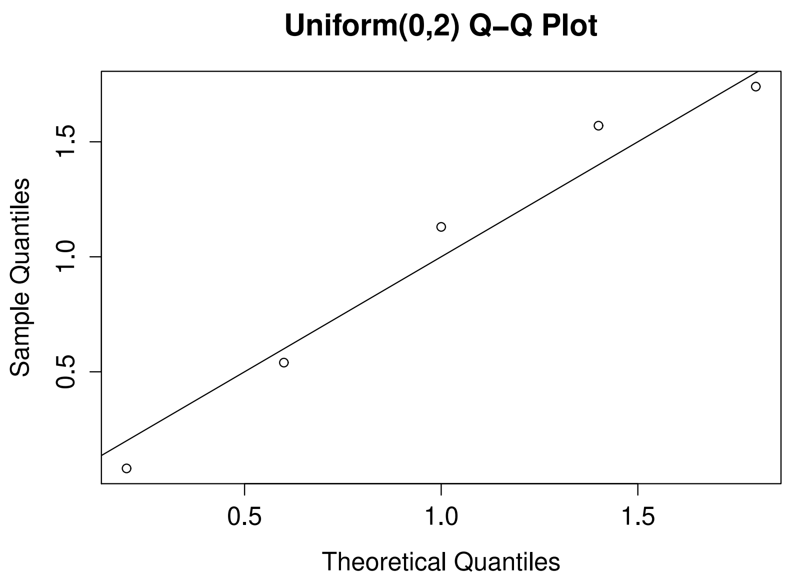
   The theoretical and sample quantiles are provided in the following table.

i	$F^{-1}\left(\frac{i-0.5}{n}\right)$	x_i
1	0.2	0.08
2	0.6	0.54
3	1	1.13
4	1.4	1.57
5	1.8	1.74

(c) Build this Q–Q plot in R using the `plot(x, y)` command. Remember that you can construct vectors in R using, e.g., `y = c(0.08, 0.54, 1.13, 1.57, 1.74)`. Overlay the line $y = x$ onto your plot using `abline(0, 1)`. (The first argument specifies the intercept, and the second specifies the slope.)

The relevant R code is:

```
y=c(0.08,0.54,1.13,1.57,1.74)
x=2*(1:5-0.5)/5
plot(x,y,xlab = "Theoretical quantiles",ylab = "Sample quantiles",
  main="Uniform(0,2) Q-Q Plot")
abline(0,1)
```



(d) Does the Uniform(0, 2) distribution appear to fit the data well?

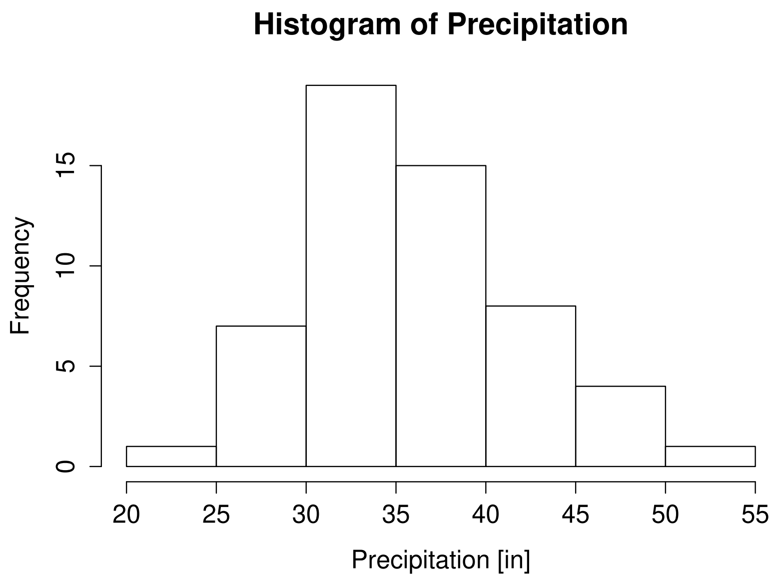
The Uniform(0, 2) seems to provide an okay fit.

6. The file `precip.csv` contains 55 samples of yearly precipitation totals (in inches) in Ithaca from 1960 to 2014.

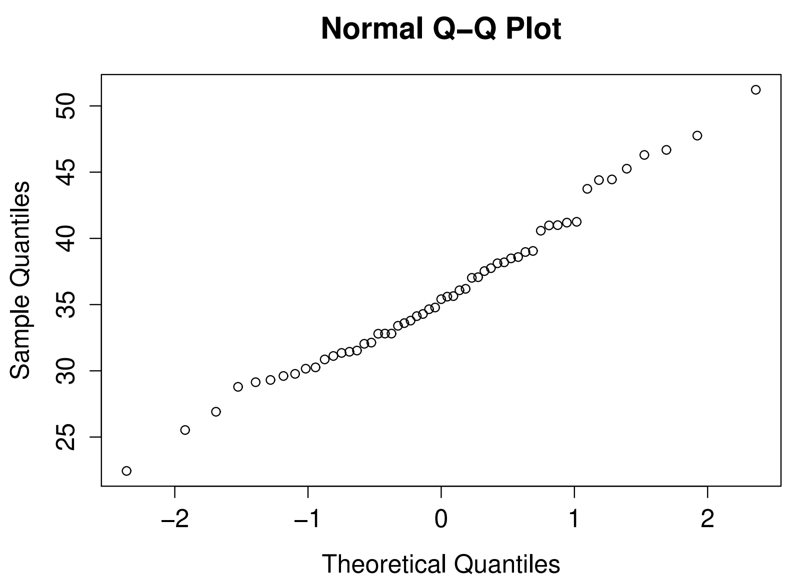
(a) Import the dataset into RStudio and create a histogram of `precip`.

We use the following code to extract and create the histogram:

```
precipTable = read.csv("precip.csv")
precip = precipTable$precip
hist(precip,xlab = "Precipitation [in]",main="Histogram of precipitation")
```



(b) Build a normal Q–Q plot of preciptation in R using `qqnorm(precip)`.
This yields the following normal Q-Q plot

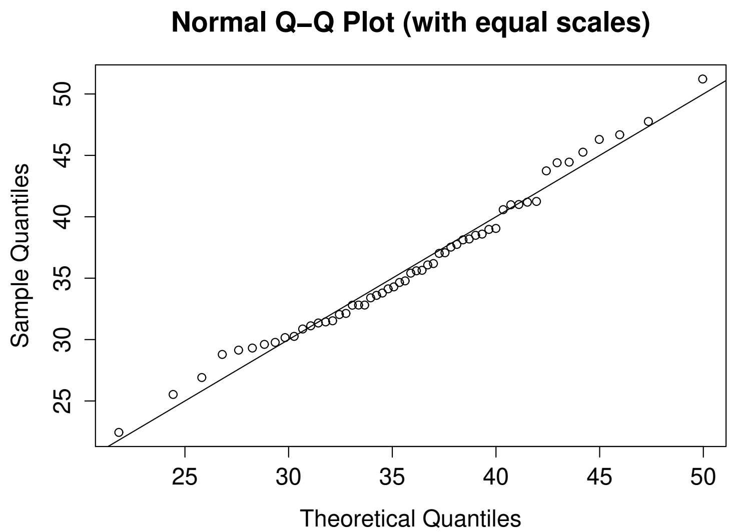


(c) Note that in the plot from part (b), the x-axis and y-axis are on different scales. This is because R's `qqnorm` function generates *standard* normal Q–Q plots. It may be better to generate a Q-Q plot with respect to a normal distribution whose mean and variance are equal to the sample mean and variance. This can be done manually as follows:

```
n = length(precip)
qi = (1:n - 0.5)/n
mu = mean(precip)
sigma = sd(precip)
x = qnorm(qi, mu, sigma)
plot(x, sort(precip))
abline(0, 1)
```

Note that `qnorm` computes the inverse CDF of the normal distribution.

Executing the above code gives us the following plot



7. You may have noticed that a normal random variable may not provided the best fit, as the data appear to have a slight positive skew. Consider fitting the data to the *gamma distribution*, which has PDF

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x \geq 0$$

where α is a *shape parameter*, β is a *rate parameter*, and

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, e^{-x}\, dx.$$

We'll try $\alpha = 38.5$ and $\beta = 1.08$. (Don't worry about how these numbers were obtained. You'll learn in a couple of weeks how parameters can be *estimated* from the data.)

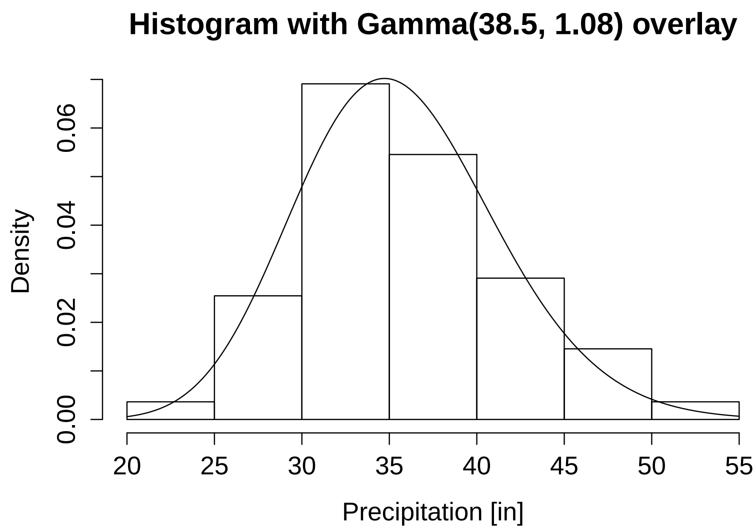
(a) Begin with a histogram of the `precip`, and overlay it with a plot of the Gamma(38.5, 1.08) PDF. R's `lines` command comes in handy here. Apply the argument `freq = FALSE` so that your histogram displays frequencies on the y-axis, rather than counts.

Executing the following code:

```
hist(precip,freq=FALSE,xlab = "Precipitation [in]",
    main="Histogram of Precipitation with Gamma(38.5,1.08) overlayed")
```

```
y = 20:55
alpha = 38.5
beta = 1.08
gammaPDF = dgamma(y, shape=alpha, rate=beta)
lines(y,gammaPDF)
```

yields the following plot

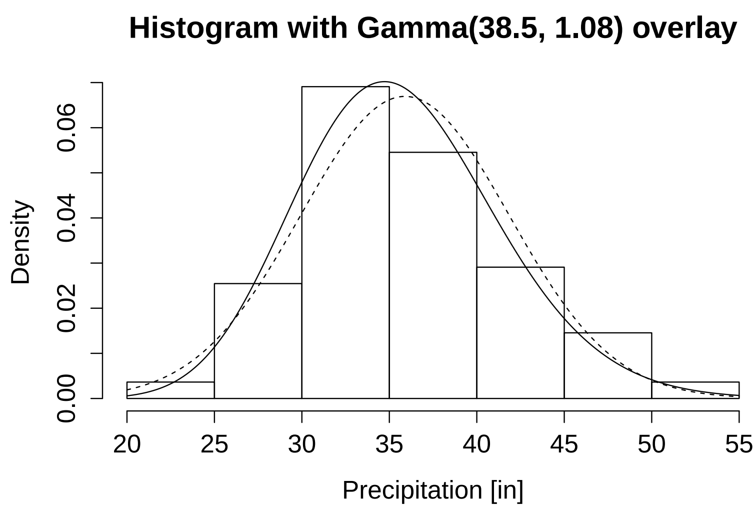


(b) Just for comparison, overlay onto your graph from part (a) the PDF of the normal distribution you used in Problem 2. Apply the argument `lty=2` to create a dashed line. How do the fits compare?

Executing the code (assuming you have `mu` and `sigma` stored from previous parts)

```
normalPDF = dnorm(y, mean=mu, sd=sigma)
lines(y, normalPDF, lty=2)
```

gives us the plot

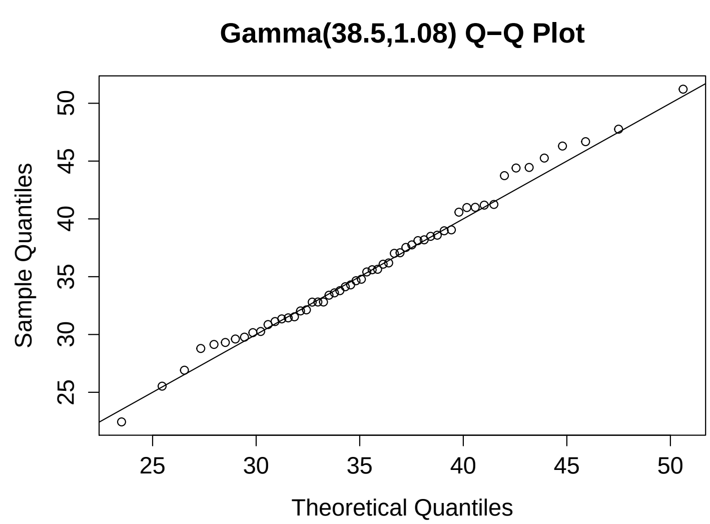


Qualitatively the gamma distribution seems to fit the data better.

(c) Construct a gamma Q–Q plot of the data. R's `qgamma` command computes the inverse CDF of the gamma distribution.
Using the following code

```
plot(qgamma(qi,alpha,beta),sort(precip),
    xlab = "Theoretical Quantiles", ylab="Sample Quantiles",
    main="Gamma(38.5,1.08) Q-Q Plot")
abline(0,1)
```

we obtain the following Q–Q plot:



(d) Does using a gamma distribution instead of a normal distribution significantly improve the fit? Explain.
We get a slightly better fit with the gamma distribution.