Introductory Statistics Lectures

Assessing Normality

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:50:01 2009)

Contents

1 Assessing Normality **1**
1.1 Introduction 1
1.2 Method of assessment . 2
Normal quantile plots . 2
Q-Q plot examples . . . 4
Assessing normality of class data 6
1.3 Summary 8
1.4 Additional Examples . . 8

1 Assessing Normality

1.1 Introduction

Question 1. How could we check to see if the mother heights in our class data set have a normal distribution?

1.2 Method of assessment

Importance of assessing normality
Many statistical tests require that the sample data come from a population with a normal distribution. If we don't satisfy the requirements then the results of the test will not be accurate. It is important to check to ensure we have met the assumptions of normality.

Below are recommendations for an initial assessment of normality.[1]

How to asses normality

1. Make a histogram. Reject normality if dramatically departs from bell shape or more than one outlier exists.
2. Make a normal quantile plot. Reject if plot does not closely follow a line.

Quantiles & Percentiles
If a student who scored 1100 on the SAT was in the 75th percentile then:

$$\overbrace{P_{75}}^{\text{percentile=0.75}} = \overbrace{1100}^{\text{quantile=1100}}$$

The quantile is the data value that has an associated percentile.

NORMAL QUANTILE PLOTS

DEFINITION 1.1 NORMAL QUANTILE PLOT (Q-Q PLOT).
A graph used to assess normality. It plots the sample quantile (vertical axis) against the theoretical quantile (horizontal axis).
sample quantile the original data point x_i value (or z-score).
theoretical quantile the expected z-score for the data point x_i when we assume it comes from a normal distribution.
If the sample quantiles match their theoretical quantiles the graph will be a **straight line indicating the data has a normal distribution.** Normally distributed sample data will have minor deviations from a straight line due to sampling error.

Finding theoretical quantiles

1. Sort the data so that the x_i's are increasing.
2. Find the k percentile (0-1 range) for each x_i

$$k_i = \frac{i - 0.5}{n}$$

[1] Further quantitative methods exist such as the Kolmogorov-Smirnov test and the Shapiro-Wilk normality test. However, these tests must be used with caution because they have very low power when used with small sample sizes and can have a high risk of type II error.

3. Find z_i', the theoretical quantile z-score corresponding to the percentile k_i[2], assuming the data is from a normal distribution:

$$z_i' = \texttt{qnorm(k)}$$

Once you have found each z_i' for each data point x_i you plot the points (z_i', x_i).

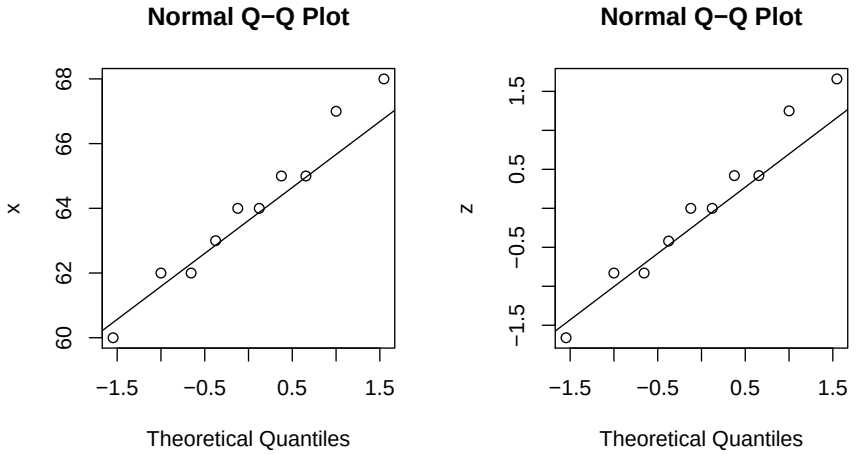
Example

The following sorted data points x_i represent 10 student's mother heights in our class. Also listed are the corresponding z scores:

$$x = \{60, 62, 62, 63, 64, 64, 65, 65, 67, 68\}$$
$$z = \{-1.66, -0.83, -0.83, -0.42, 0, 0, 0.42, 0.42, 1.25, 1.66\}$$

Question 2. Find the theoretical quantile corresponding to the third data point.

Below are Normal Q-Q plots for the above 10 mother heights. Not that plotting (z_i', x_i) is equivalent to (z_i', z_i).

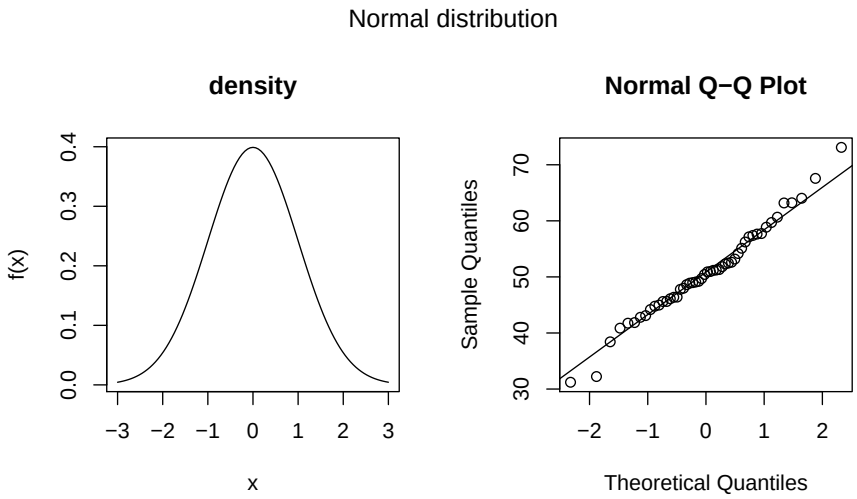


Q-Q PLOTS:
`qqnorm(x); qqline(x)`
Where `x` is a vector of data.

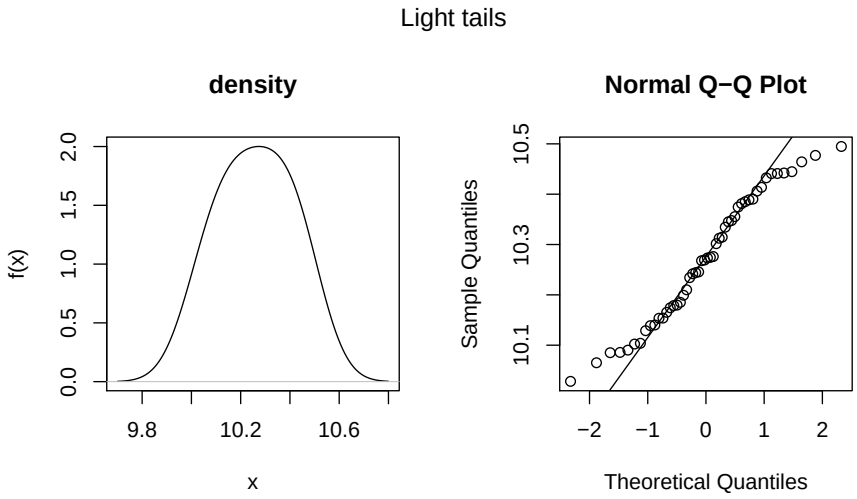
R COMMAND

[2]Percentiles are a type of probability.

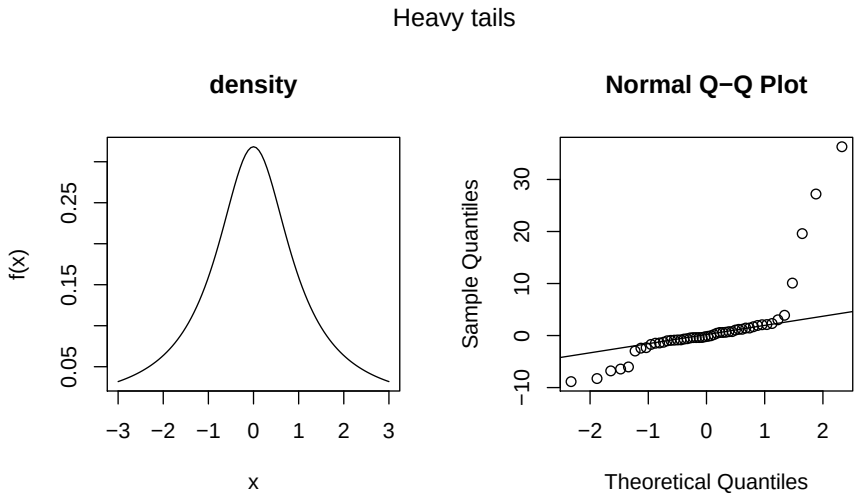
Q-Q PLOT EXAMPLES



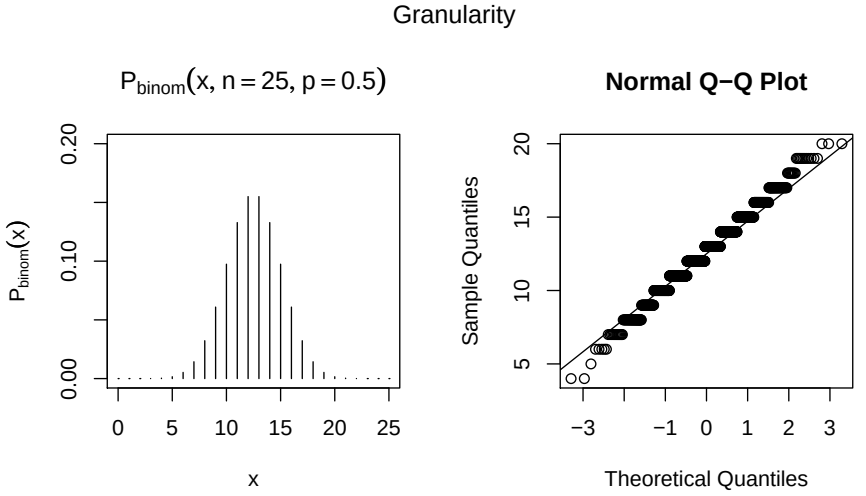
Q-Q plot of 50 data points randomly selected from a normal density.



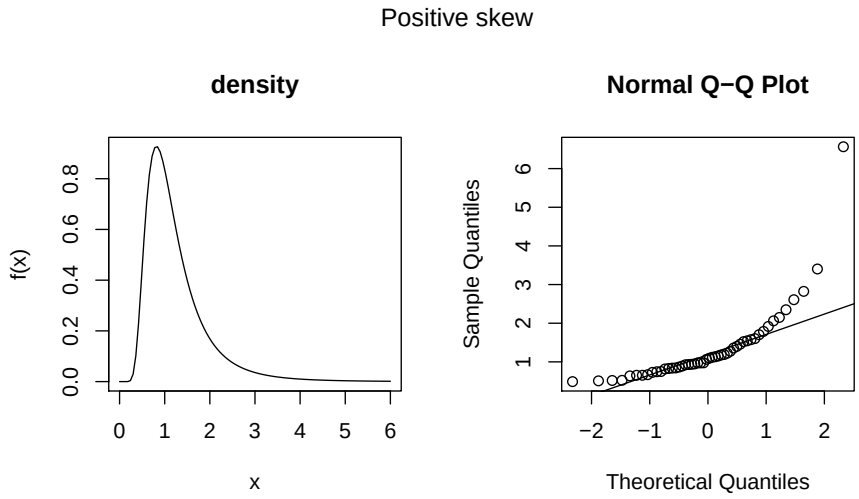
A light tailed density has less area in the tails making them appear shorter.



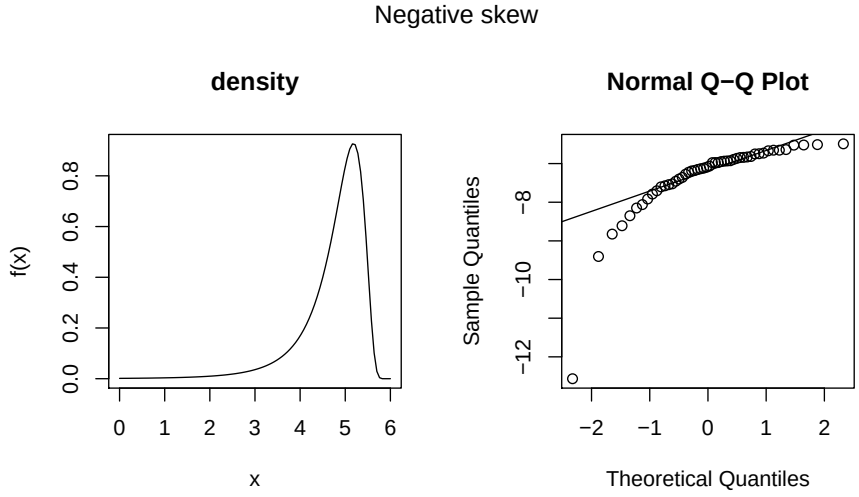
A heavy tailed density has more area in the tails making them appear longer.



Discrete data (such as binomial) will be clumped or granular in a Q-Q plot.



For positive skew, Q-Q plot has increasing slope from left to right.

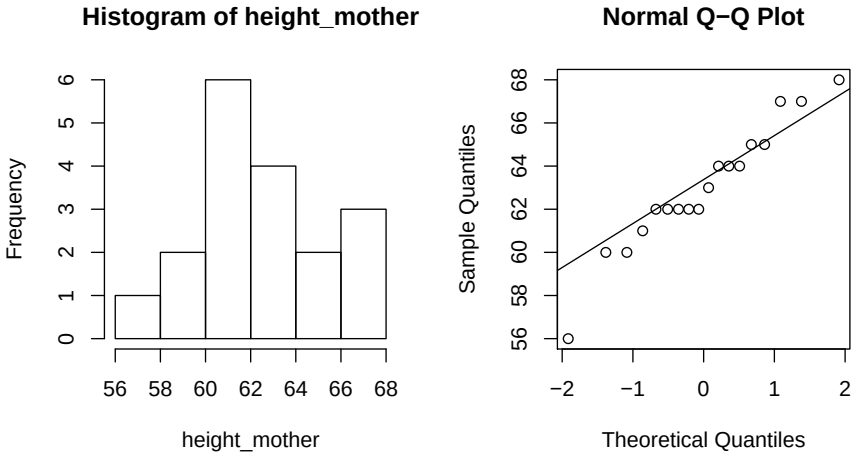


For negative skew, Q-Q plot has decreasing slope from left to right.

ASSESSING NORMALITY OF CLASS DATA

Mother heights

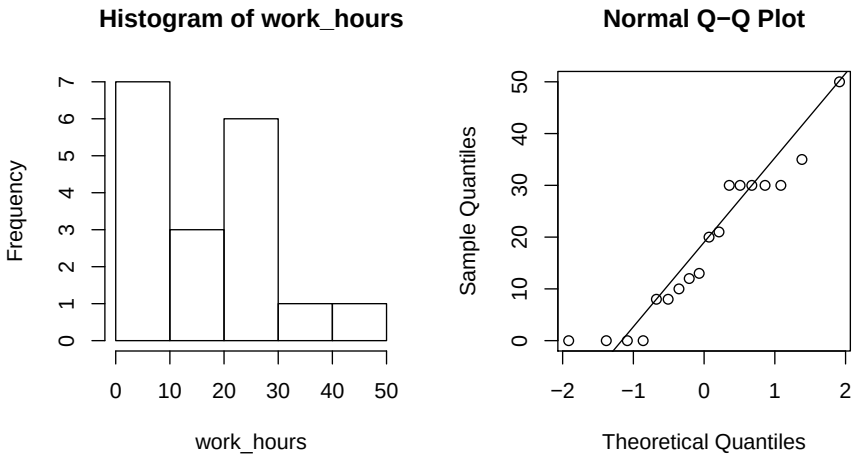
```
R: par(mfrow = c(1, 2))
R: hist(height_mother)
R: qqnorm(height_mother)
R: qqline(height_mother)
```



Question 3. Do the mother heights appear to be normally distributed?

Work hours

```
R: par(mfrow = c(1, 2))
R: hist(work_hours)
R: qqnorm(work_hours)
R: qqline(work_hours)
```



Question 4. Do the student work hours appear to be normally distributed?

1.3 Summary

How to asses normality

1. Make a histogram. Reject normality if dramatically departs from bell shape or more than one outlier exists.
   `R: hist(x)`
2. Make a normal quantile plot. Reject if plot does not closely follow a line.
   `R: qqnorm(x); qqline(x)`

1.4 Additional Examples

Question 5. Assess the normality of the men's weight and cholesterol data in the `Mhealth` table from the book.