To Do:

Read Chapter 2, Sections 2.5-2.6.
Do end-of-chapter problems 1-19.

Start reading Chapter 3.

Today's Lecture

Invariance Property of Maximum Likelihood Estimates (Section 2.5)

Checking the Fit of a Model (Section 2.6)

Invariance Property of Maximum Likelihood Estimates (Sec. 2.5)

One of the reasons the Method of Maximum Likelihood is so popular.

Theorem 13 (Course Notes, page 61)

If $\hat{\theta}$ is the maximum likelihood estimate of θ then $g(\hat{\theta})$ is the maximum likelihood estimate of $g(\theta)$.

Example

For Binomial data with *y* successes observed in *n* Bernoulli trials the maximum likelihood estimate of P(Success) = *θ* is

$$\hat{\theta} = \frac{y}{n}$$

The maximum likelihood estimate of Var(*Y*) = *nθ*(1 - *θ*) is

$$n\hat{\theta}(1-\hat{\theta}) = n\left(\frac{y}{n}\right)\left(1-\frac{y}{n}\right)$$

Today's Lecture: Checking the Fit of the Model – Section 2.6

In order to study a data set we typically assume a model in which

$Y = (Y_1, Y_2, \ldots, Y_n)$ is a potential random sample from a distribution which is a member of the family of models

$$f(y; \theta) \quad \text{for } \theta \in \Omega.$$

It is important to check that the model adequately represents the variability in Y.

Methods for checking the fit of a model:

(1) Compare a relative frequency histogram of the observed data with a superimposed graph of the probability density function of the assumed (continuous distributions).

(2) Compare a graph of the empirical cumulative distribution function with a superimposed graph of the cumulative distribution function of the assumed (continuous distributions).

You have already seen examples of (1) and (2) on Assignment 1.

Methods for Checking the Fit of a Model Continued

(3) Compare observed frequencies with expected frequencies calculated using the assumed model (discrete and continuous distributions).

(4) Examine a Normal or Gaussian qqplot.

Checking the Fit of the Model (3)

One way to check the fit of the model is to compare the observed relative frequencies based on the data with the expected frequencies calculated using probabilities from the assumed model.

If the model is suitable then the observed and expected frequencies should be "close".

Discrete Data Example: Alpha-Particle Emissions

The Poisson distribution is used to model random events in time.

In a 1910, Ernest Rutherford and Hans Geiger recorded the number of alpha-particles emitted form a polonium source during a fixed period of time (one-eighth of a minute). They made 2608 recordings.

Rutherford and Geiger Experiment 1910

The alpha-particles were detected by a piece of equipment which would eventually be called a Geiger counter.

Geiger Counter
(circa 1908)

Observed Data for 1910 Study of Alpha-Particles

Number of Alpha-Particles Detected	Frequency f_j	Expected Frequency e_j
0	**57**	?
1	**203**	?
2	**383**	?
3	**525**	?
4	**532**	?
5	**408**	?
6	**273**	?
7	**139**	?
8	**45**	?
9	**27**	?
10	**10**	?
11	**6**	?
Total	**2608**	2608

Poisson Model for Alpha-Particle Data

Let Y = number of alpha-particles observed in the time interval of one-eighth minute.

How reasonable is it to assume the model $Y \sim$ Poisson(θ)? What does the parameter θ represent?

To check the model we can compare the observed relative frequencies based on the data with the expected frequencies calculated using probabilities based on the Poisson(θ) model.

Estimating the Parameters

To calculate the expected frequencies we first need to estimate the value of the unknown parameter θ based on the data $y_1, y_2, \ldots, y_n$.

We estimate θ using the maximum likelihood estimate of θ for Poisson data which is the sample mean:

$$\bar{y} = \frac{1}{2608}[57(0) + 203(1) + \cdots + 6(11)] \approx 3.87$$

Note how the sample mean is calculated based on the table of observed frequencies.

Calculating Probabilities

If *Y* ~ Poisson(3.87) then the probability of observing exactly y alpha-particles in the fixed interval is

$$P(Y = y) = \frac{(3.87)^y e^{-3.87}}{y!}, \quad y = 0,1,\ldots$$

For example

$$P(Y = 0) = \frac{(3.87)^0 e^{-3.87}}{0!} = e^{-3.87} = 0.02087$$

Calculating Expected Frequencies

Since P(Y = 0) = 0.02087, then for 2608 time intervals we would expect to see

(2608)(0.02087) = 54.42 intervals

with no alpha-particles observed during the interval.

More generally, we would expect to observe

$$e_j = (2608)P(Y = j) = (2608)\frac{(3.87)^j e^{-3.87}}{j!}, \qquad j = 0,1,\ldots$$

intervals containing *j* alpha-particles.

Observed and Expected Frequencies under assumed Poisson Model

Number of Alpha-Particles Detected	Frequency f_j	Expected Frequency e_j
0	57	54.42
1	203	210.58
2	383	407.43
3	525	525.54
4	532	508.41
5	408	393.47
6	273	253.77
7	139	140.28
8	45	67.86
9	27	29.18
10	10	11.29
11+	6	5.77
Total	2608	2607.99

How well does the model fit the data?

What would you conclude based on this table?

In Chapter 7 we look at a more rigorous statistical method for making such decisions.

How to choose the categories?

For these data the categories

$$y = 0, 1, 2, \ldots, 10, 11+$$

where y = number of alpha particles detected

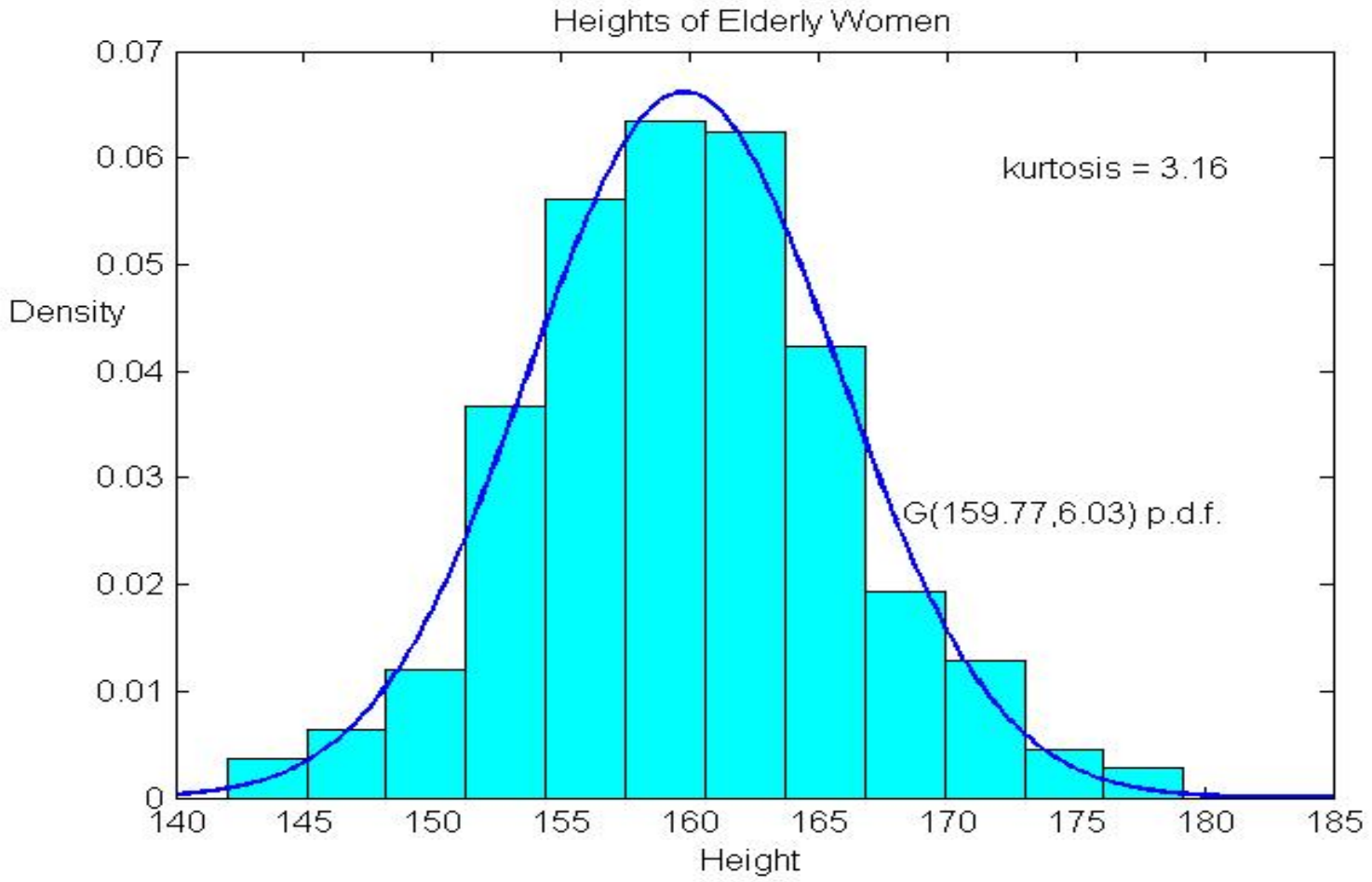
are obvious choices.

In general for discrete data the categories are easier to choose than for continuous data.

Continuous Data Example: Heights in centimeters of a sample of 351 elderly women randomly selected from a community in a study of osteoporosis.

156	163	169	161	154	156	163	164	156	166	177	158
150	164	159	157	166	163	153	161	170	159	170	157
156	156	153	178	161	164	158	158	162	160	150	162
155	161	158	163	158	162	163	152	173	159	154	155
164	163	164	157	152	154	173	154	162	163	163	165
160	162	155	160	151	163	160	165	166	178	153	160
156	151	165	169	157	152	164	166	160	165	163	158
153	162	163	162	164	155	155	161	162	156	169	159
159	159	158	160	165	152	157	149	169	154	146	156
157	163	166	165	155	151	157	156	160	170	158	165
167	162	153	156	163	157	147	163	161	161	153	155
166	159	157	152	159	166	160	157	153	159	156	152
151	171	162	158	152	157	162	168	155	155	155	161
157	158	153	155	161	160	160	170	163	153	159	169
155	161	156	153	156	158	164	160	157	158	157	156
160	161	167	162	158	163	147	153	155	159	156	161
158	164	163	155	155	158	165	176	158	155	150	154
164	145	153	169	160	159	159	163	148	171	158	158
157	158	168	161	165	167	158	158	161	160	163	163
169	163	164	150	154	165	158	161	156			
154	158	162	164	158	165	158	156	162			
157	167	142	166	163	163	151	163	153			
169	154	155	167	164	170	174	155	157			
155	168	152	165	158	162	173	154	167			
158	167	164	170	164	166	170	160	148			
150	165	165	147	162	165	158	145	150			
163	166	162	163	160	162	153	168	163			
158	155	168	160	153	163	161	145	161			
161	155	158	161	163	157	156	152	156			
160	152	153									

Heights of Elderly Women Data



$$\bar{y} = 159.77, \quad s^2 = 36.36, \quad s = 6.03$$

Checking Gaussian Model for Heights of Elderly Women

Let Y = height of a randomly selected elderly woman in a large population. How reasonable is it to assume the model $Y \sim G(\mu,\sigma)$?

To check the model we can compare the observed relative frequencies based on the data with the expected frequencies calculated using probabilities based on the $G(\mu,\sigma)$ model.

Observed Frequencies for Heights of Elderly Women

How do we choose these intervals?

Interval $I_j = [a_j, a_{j-1})$	Observed Frequency f_j	Expected Frequency e_j
(-∞,148.2)	11	?
[148.2,151.3)	13	?
[151.3,154.4)	40	?
[154.4,157.5)	61	?
[157.5,160.6)	69	?
[160.6,163.7)	68	?
[163.7,166.8)	46	?
[166.8,169.9)	21	?
[169.9,173.0)	14	?
[173,+ ∞)	8	?
Total	351	351

How to choose the intervals

The intervals are typically selected so that there are 10 to 15 intervals and each interval contains at least one y-value from the sample.

If a software package like R is used to produce the frequency histogram then the intervals are usually chosen automatically. An option for user specified intervals is also usually provided.

Choosing the intervals usually involves a number of attempts until a good set of intervals are found.

Estimating the Parameters

To calculate the expected frequencies we need to estimate the values of the unknown parameters μ and σ based on the data $y_1, y_2, \ldots, y_n$.

We estimate μ and σ using the sample mean (the maximum likelihood estimate) and the sample standard deviation (not the maximum likelihood estimate):

$$\bar{y} = 159.77, \quad s^2 = 36.36, \quad s = 6.03$$

Calculating Probabilities

For example if $Y \sim G(159.77, 6.03)$ then the probability a randomly chosen elderly woman has a height between 155 and 156 cm is

$$
\begin{aligned}
& P(155 < Y \leq 156) \\
&= P\left(\frac{155-159.77}{6.03} < Z \leq \frac{156-159.77}{6.03}\right) \\
&= P(-0.79 < Z \leq -0.63) \quad \text{where } Z \sim G(0,1) \\
&= 0.0514
\end{aligned}
$$

Calculating Expected Frequencies

Since P(155 < Y ≤ 156) = 0.0514, then in a sample of 351 elderly women we would expect to see (351)(0.0514) = 18.04 women with heights between 155 and 156 cm.

More generally, for the interval $I_j = [a_j, a_{j-1})$ we would we expect to observe $e_j = (351)(p_j)$ women with heights between a_{j-1} and a_j cm where

$$p_j = P\big(Y \in [a_{j-1}, a_j)\big) = P\big(a_{j-1} < Y \le a_j\big)$$

$$= P\left(\frac{a_{j-1} - 159.77}{6.03} < Z \le \frac{a_j - 159.77}{6.03}\right)$$

$$\text{where } Z \sim G(0,1)$$

Observed Frequencies and Expected Frequencies under Assumed Gaussian Model

Interval $I_j = [a_j, a_{j-1})$	Observed Frequency f_j	Expected Frequency e_j
$(-\infty, 148.2)$	11	9.64
[148.2,151.3)	13	18.42
[151.3,154.4)	40	37.35
[154.4,157.5)	61	58.49
[157.5,160.6)	69	70.00
[160.6,163.7)	68	66.00
[163.7,166.8)	46	47.58
[166.8,169.9)	21	26.48
[169.9,173.0)	14	11.38
$[173, +\infty)$	8	4.76
Total	351	350.8

How well does the model fit the data?

What would you conclude based on this table?

How well does the model fit the data?

What would you conclude based on this table?

For continuous random variables we have another (graphical) way to decide whether or not the assumed model fits the data.

Checking the Fit of the Model (4)

Another way to check the fit of the model is to compare the cumulative distribution function based on the model and the empirical cumulative distribution based on the data.

See the video:

"*The empirical cdf and the qqplot* at watstat.ca"

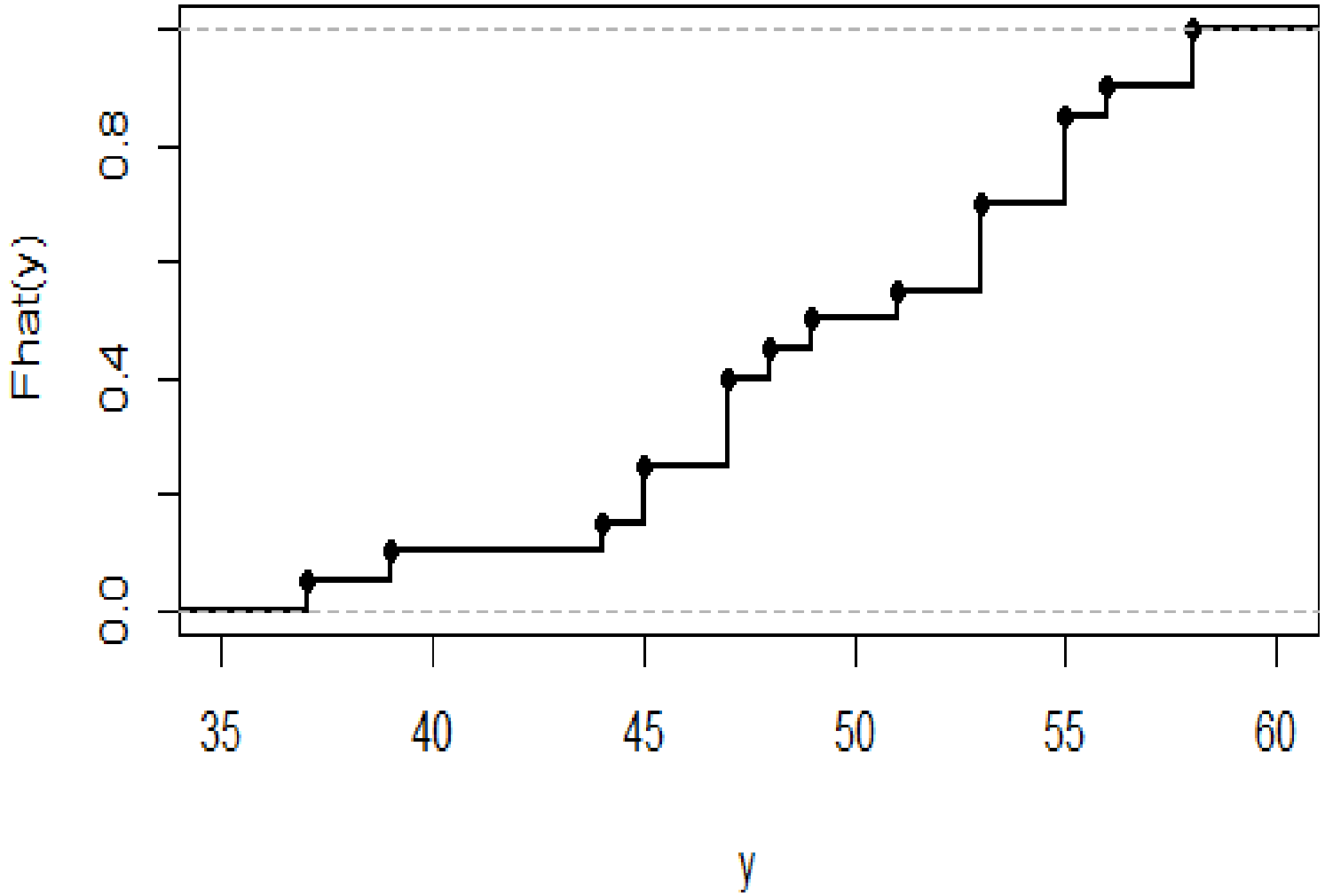
Recall Definition of Empirical Cumulative Distribution Function

$$\hat{F}(y) = \frac{\text{number of values in } \{y_1, y_2, \ldots, y_n\} \leq y}{n}$$

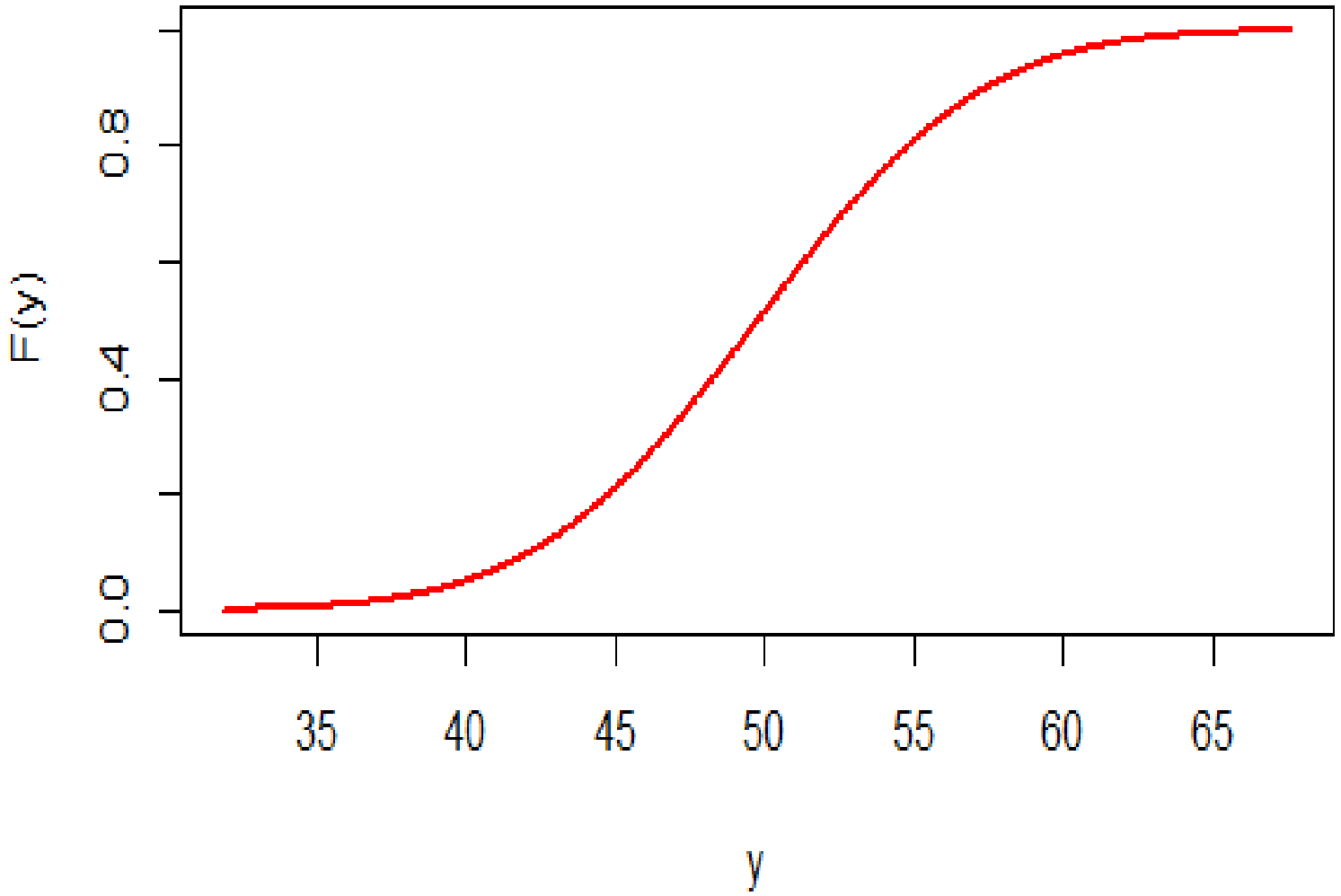
defined for all $y \in \Re$

is an estimate, based on the observed data, of the true c.d.f $F(y) = P(Y \leq y)$ from which the observed data arise.

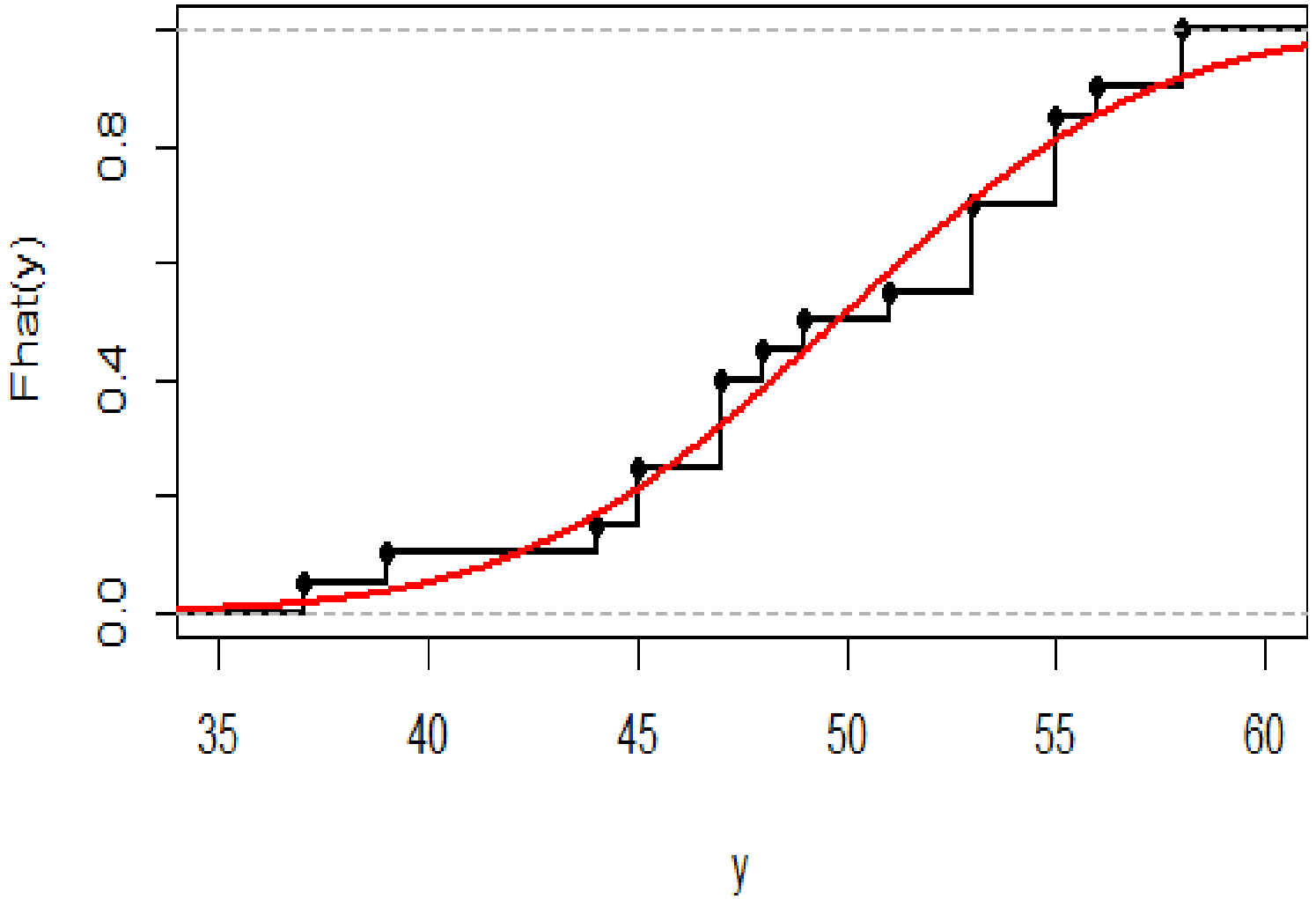
Empirical c.d.f.



Gaussian c.d.f.



Empirical c.d.f. and Gaussian c.d.f.



Comparing c.d.f.'s

Is it reasonable to assume a Gaussian model for these data?

How to compare the graphs?

It is easier to make comparisons when dealing with straight lines.

How can we turn the previous graph of two curves into a straight line comparison?

In other words, we want to create a plot, based on the data, which will be approximately a straight line if a Gaussian model fits the data well.

Normal or Gaussian Qqplot

See discussion page 69 of the Course Notes.

Suppose $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ are the observed data ordered from smallest to largest.

A plot of the points

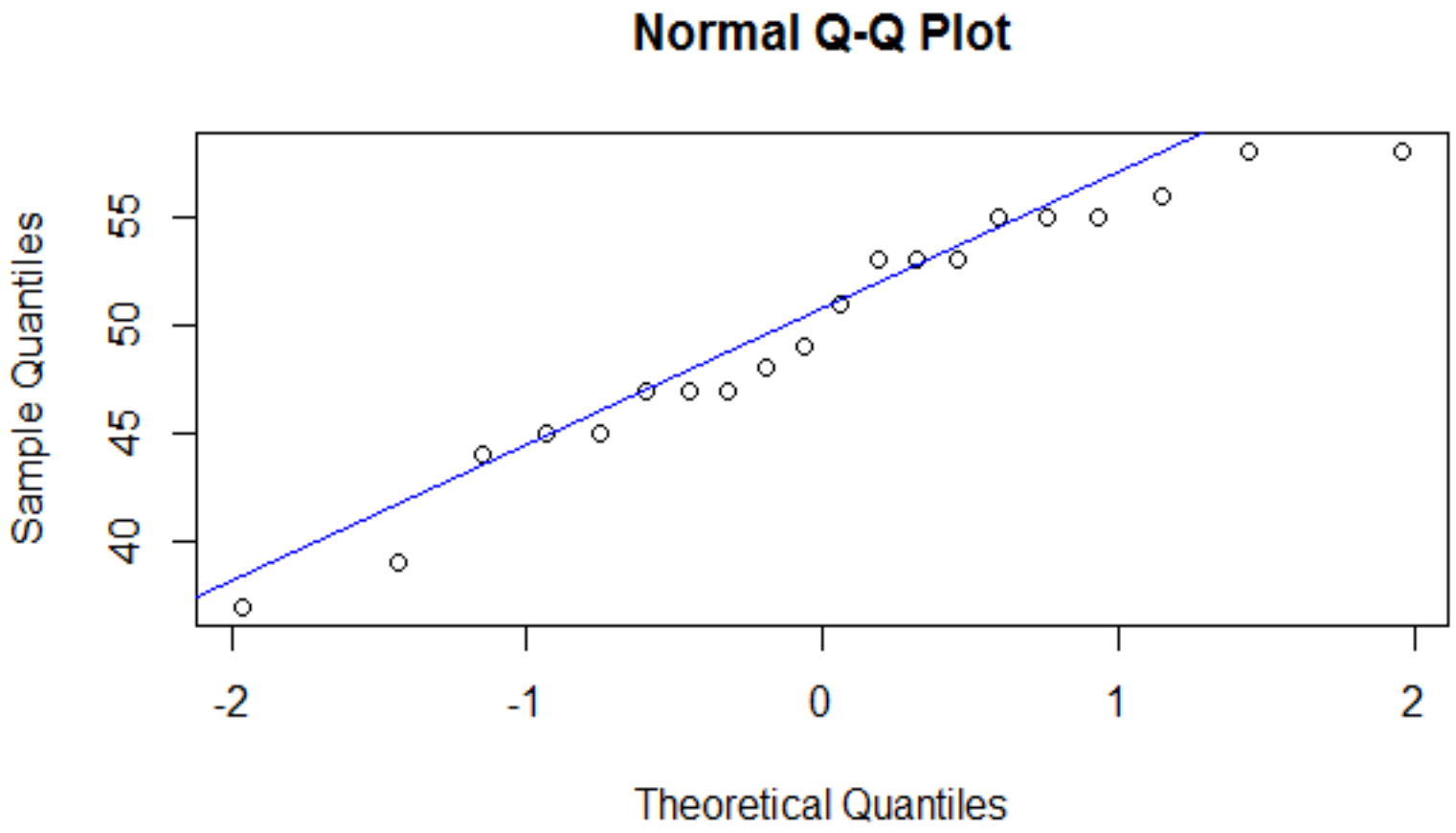
$$\left(\Phi^{-1}\left(\frac{i}{n+1} \right), y_{(i)} \right) \quad \text{for } i = 1, 2, \ldots, n$$

where Φ^{-1} is the inverse c.d.f. of a G(0,1) random variable

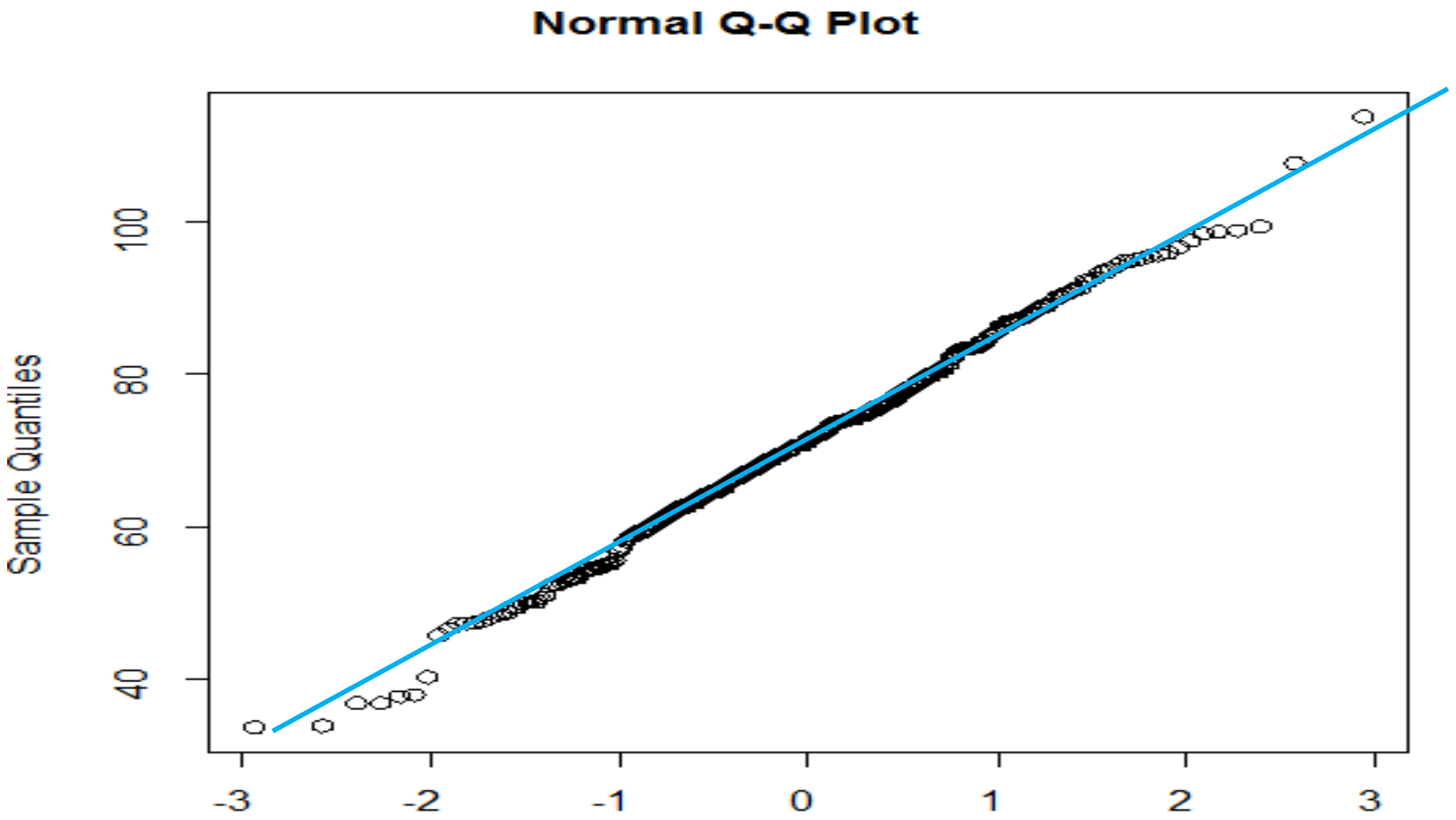
should be approximately a straight line if the data are well modelled by a Normal or Gaussian distribution.

Qqplot of data

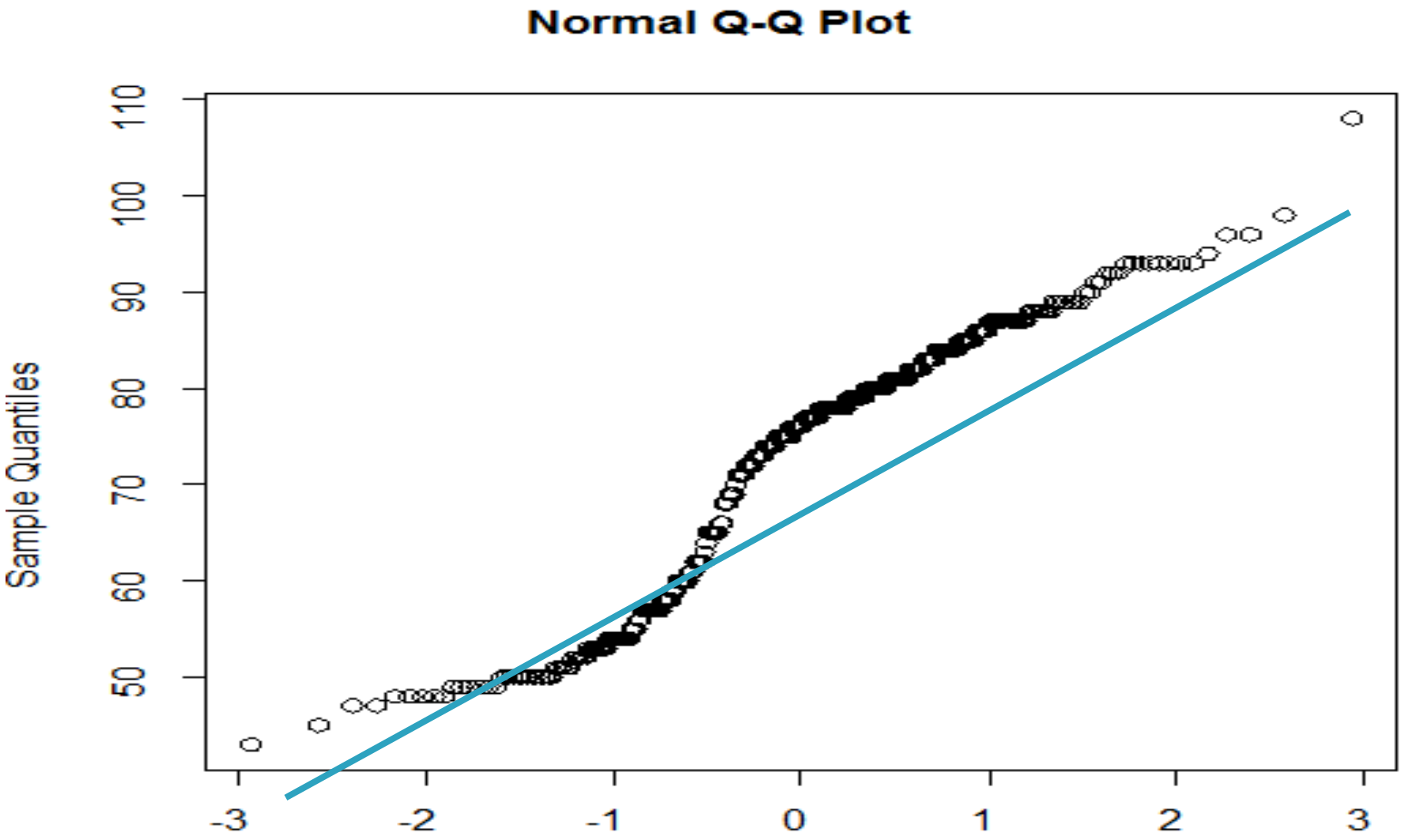
Data: 37 39 44 45 45 47 47 47 48 49
51 53 53 53 55 55 55 56 58 58



Qqplot for Simulated Normal Data



Qqplot for Old Faithful Data

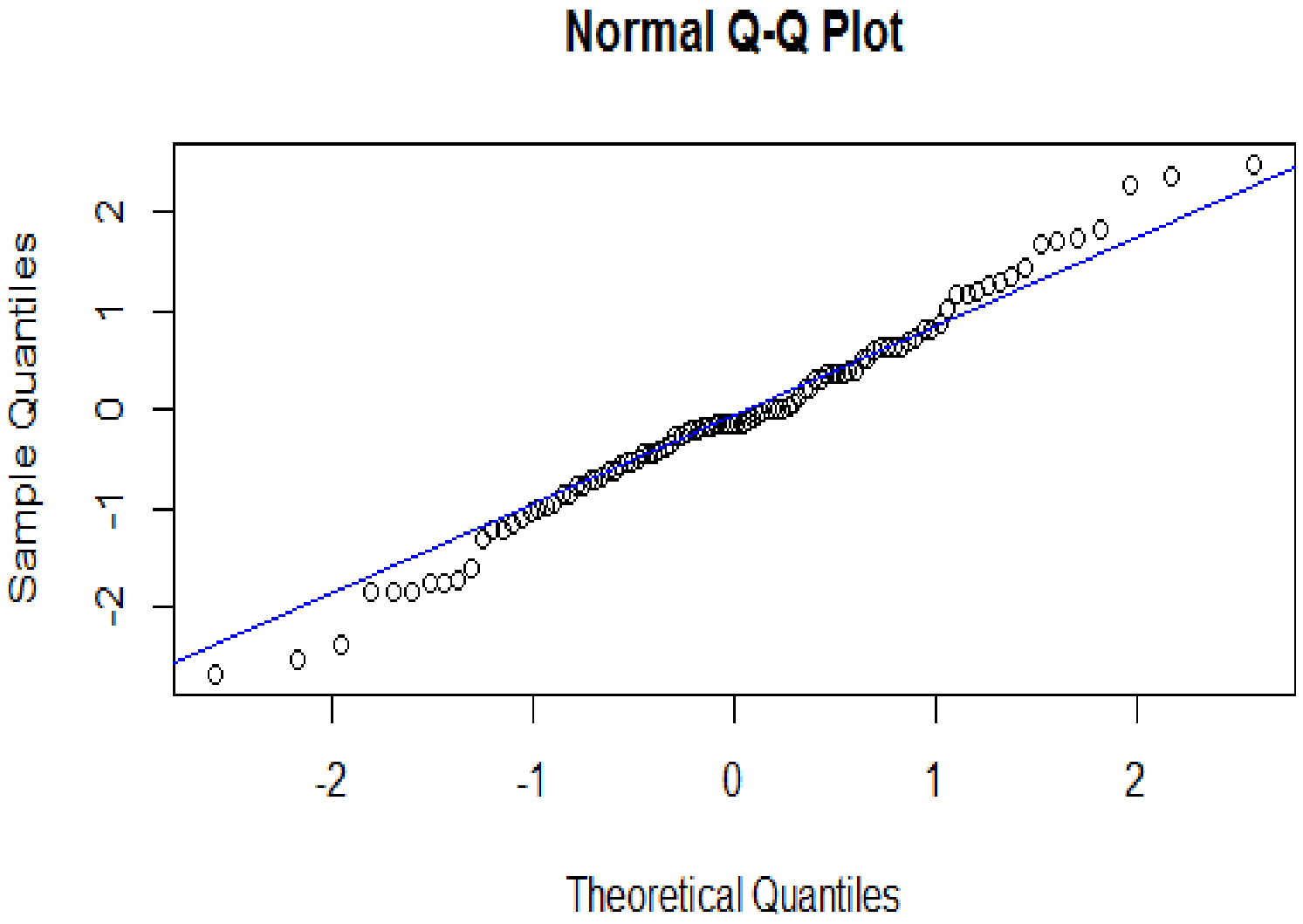


Sample qqplots in R

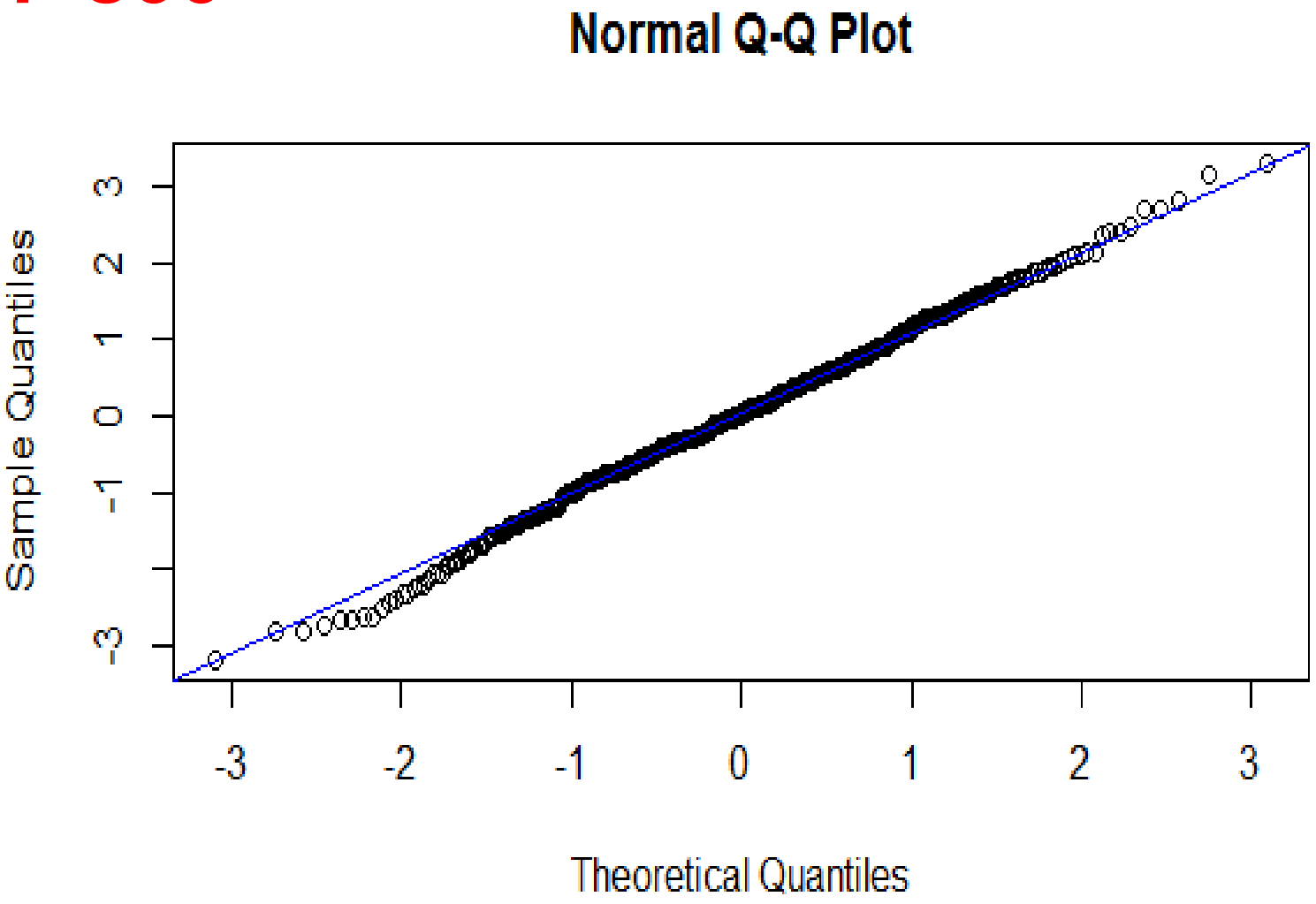
Are they symmetric? Heavy tailed? Light tailed? Skewness? Kurtosis > 3?

qqnorm(rnorm(100))	qqnorm(rnorm(500))
qqnorm(runif(100))	qqnorm(runif(500))
qqnorm(rexp(100))	qqnorm(rexp(500))
qqnorm(rgamma(100,4,1))	qqnorm(rgamma(500,4,1))
qqnorm(rt(100,3))	qqnorm(rt(500,3))
qqnorm(rcauchy(100))	qqnorm(rcauchy(500))

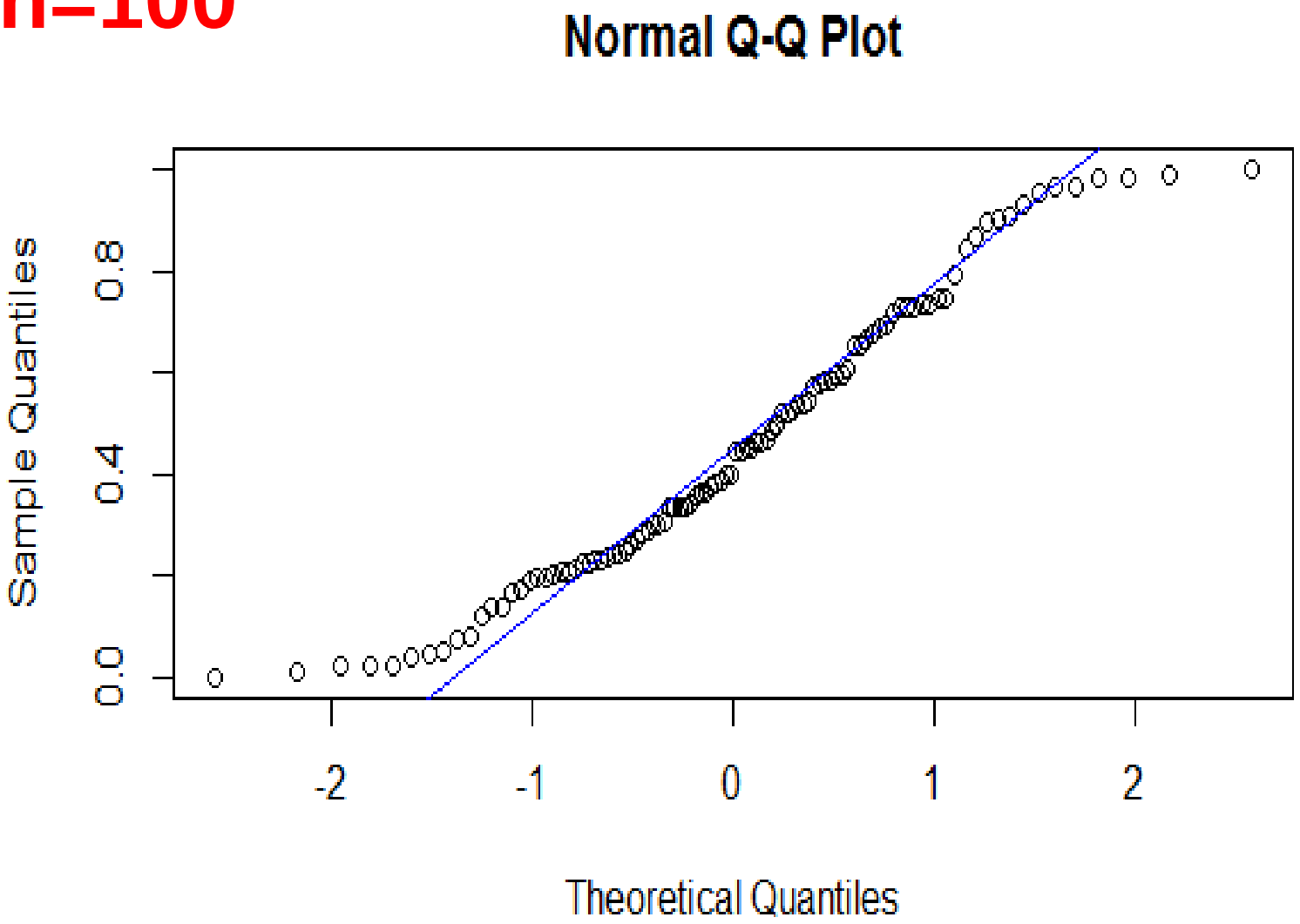
Qqplot for Simulated Gaussian Data, n=100



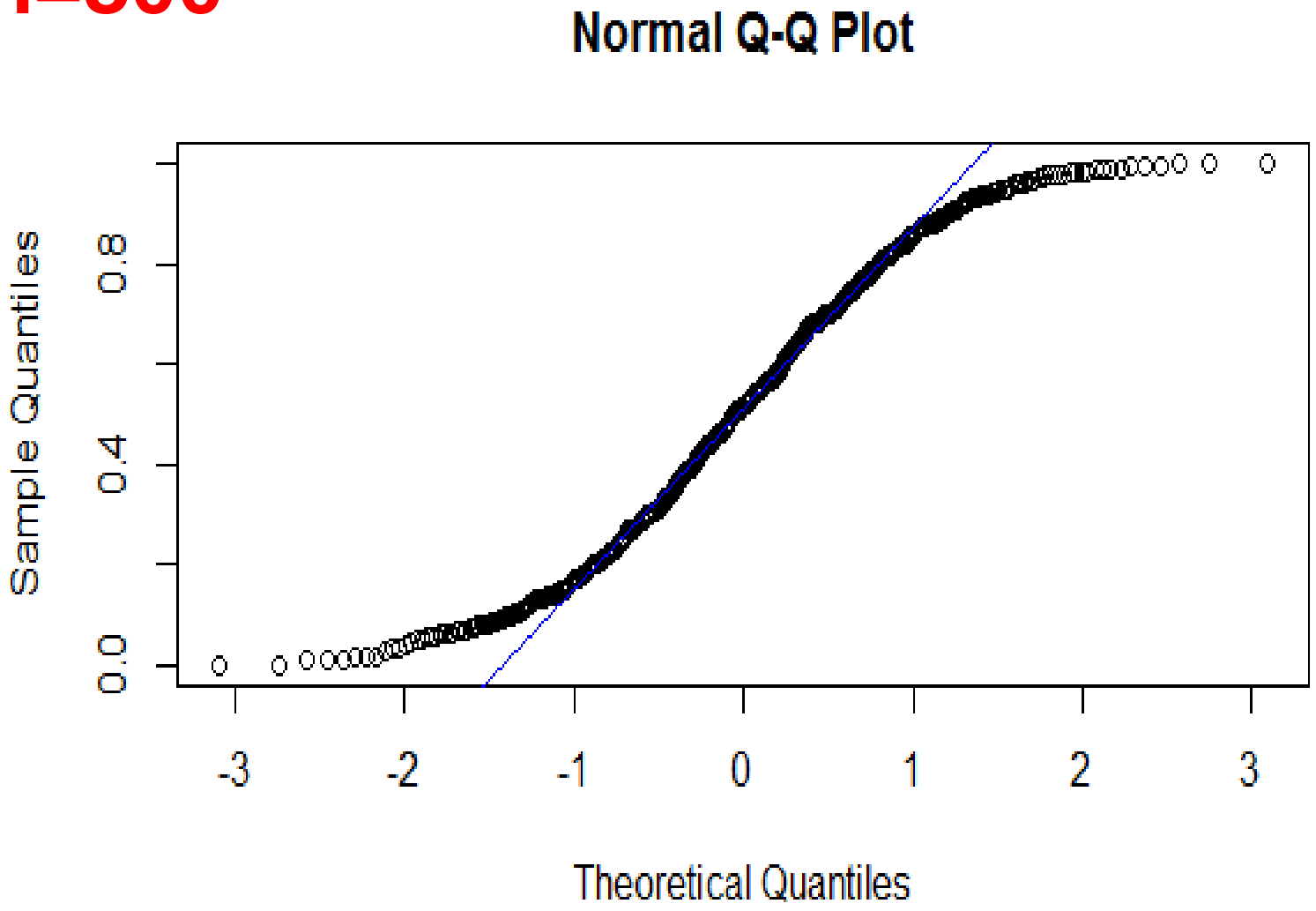
Qqplot for Simulated Gaussian Data, n=500



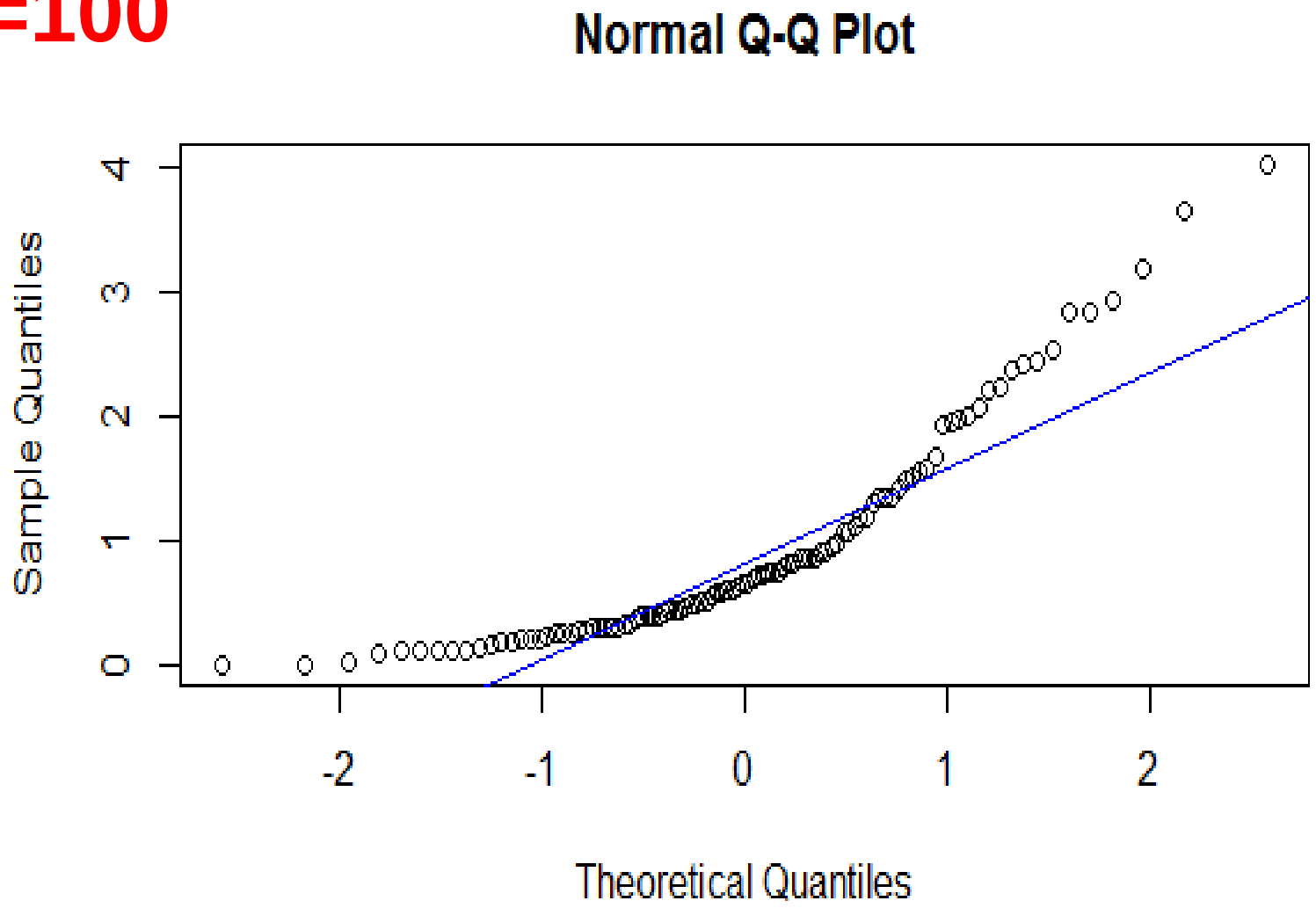
Qqplot for Simulated Uniform Data, n=100



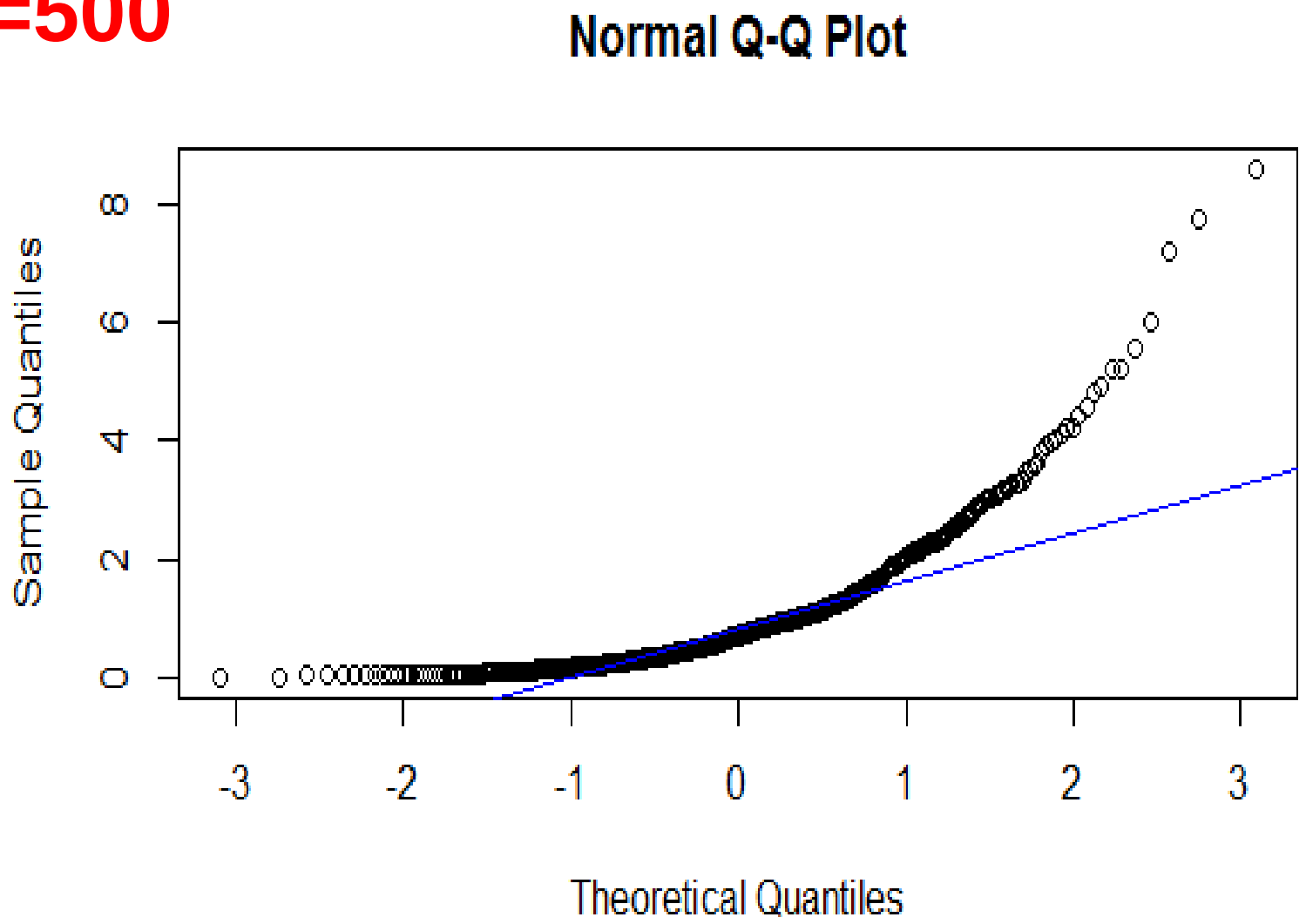
Qqplot for Simulated Uniform Data, n=500



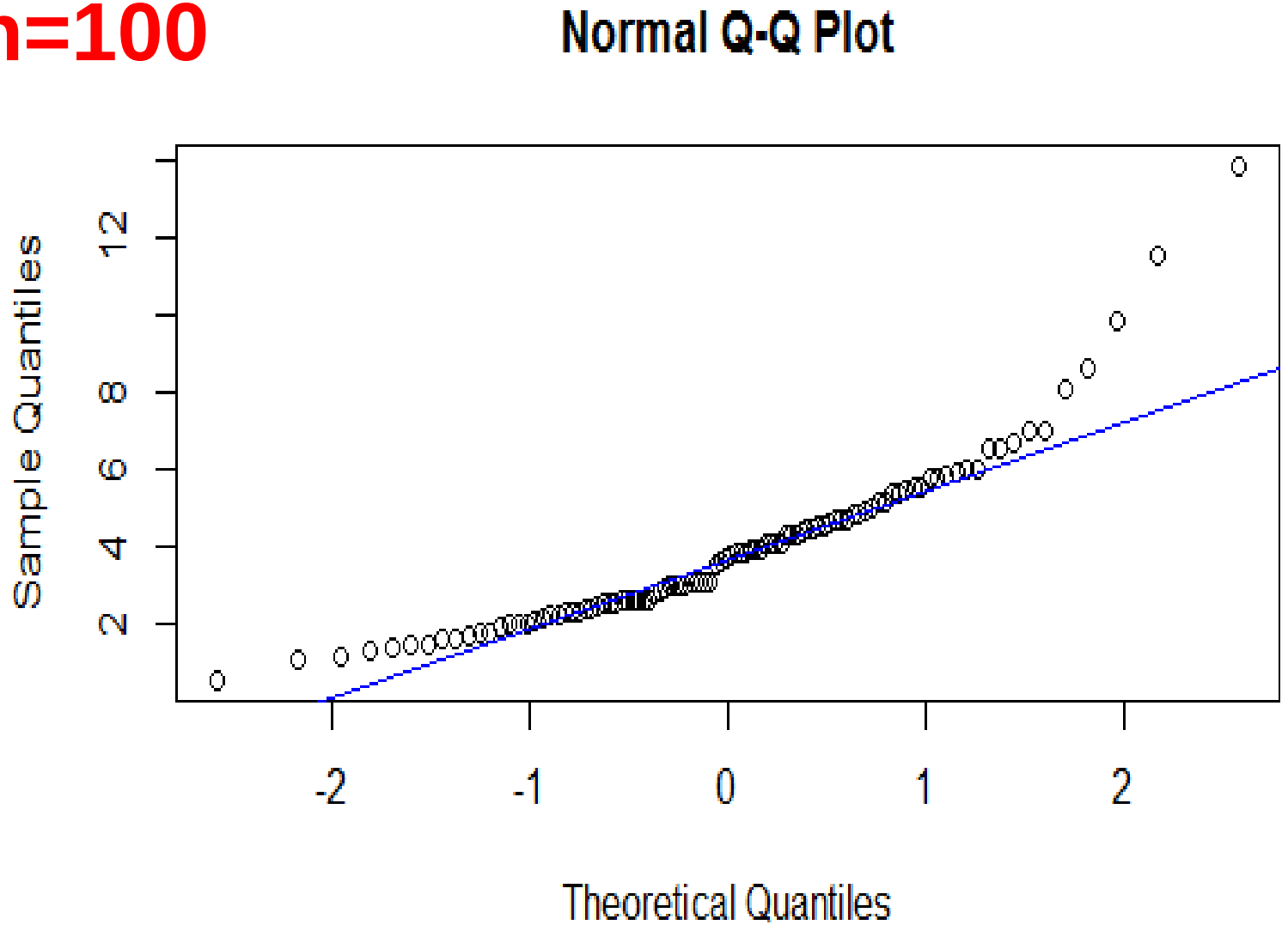
Qqplot for Simulated Exponential Data, n=100



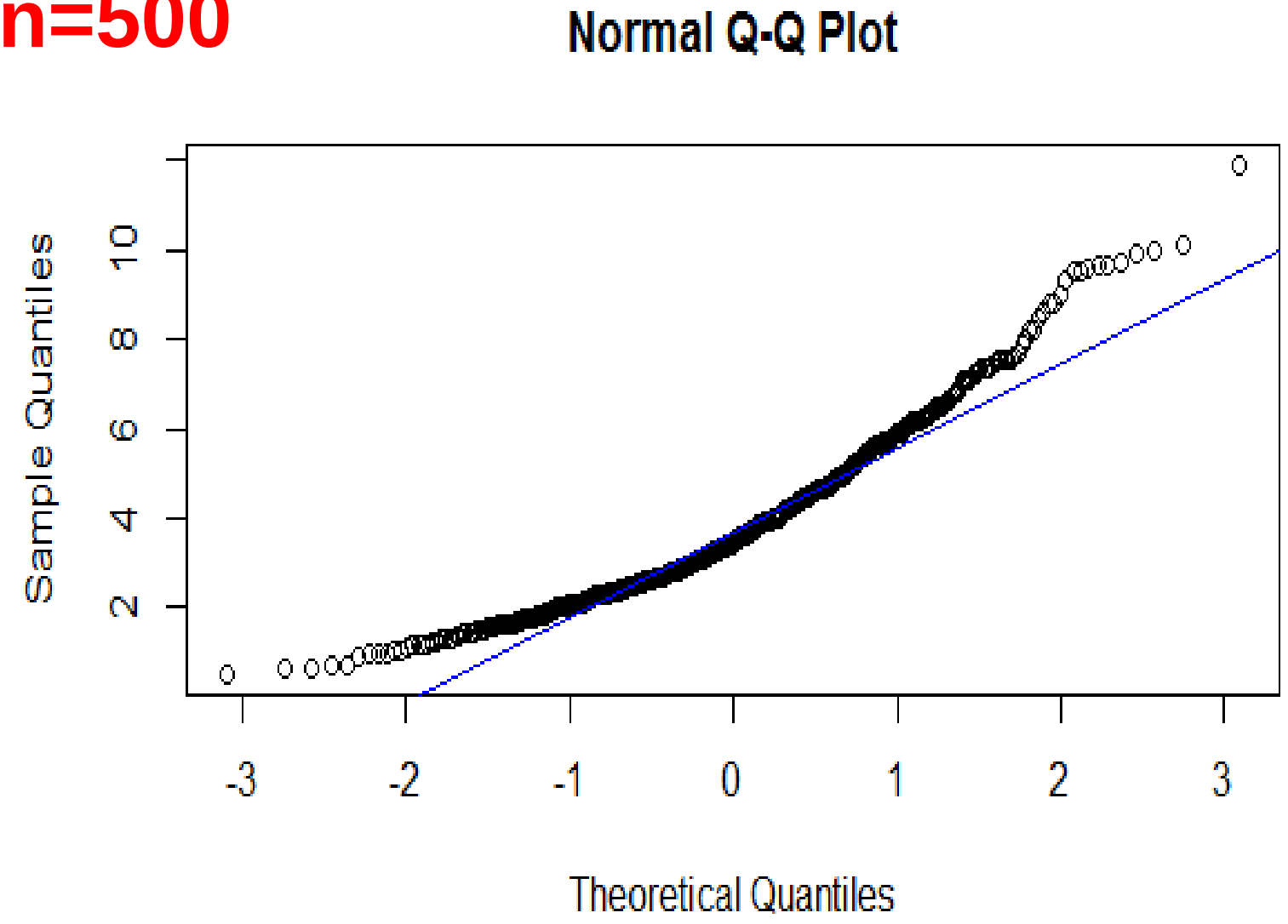
Qqplot for Simulated Exponential Data, n=500



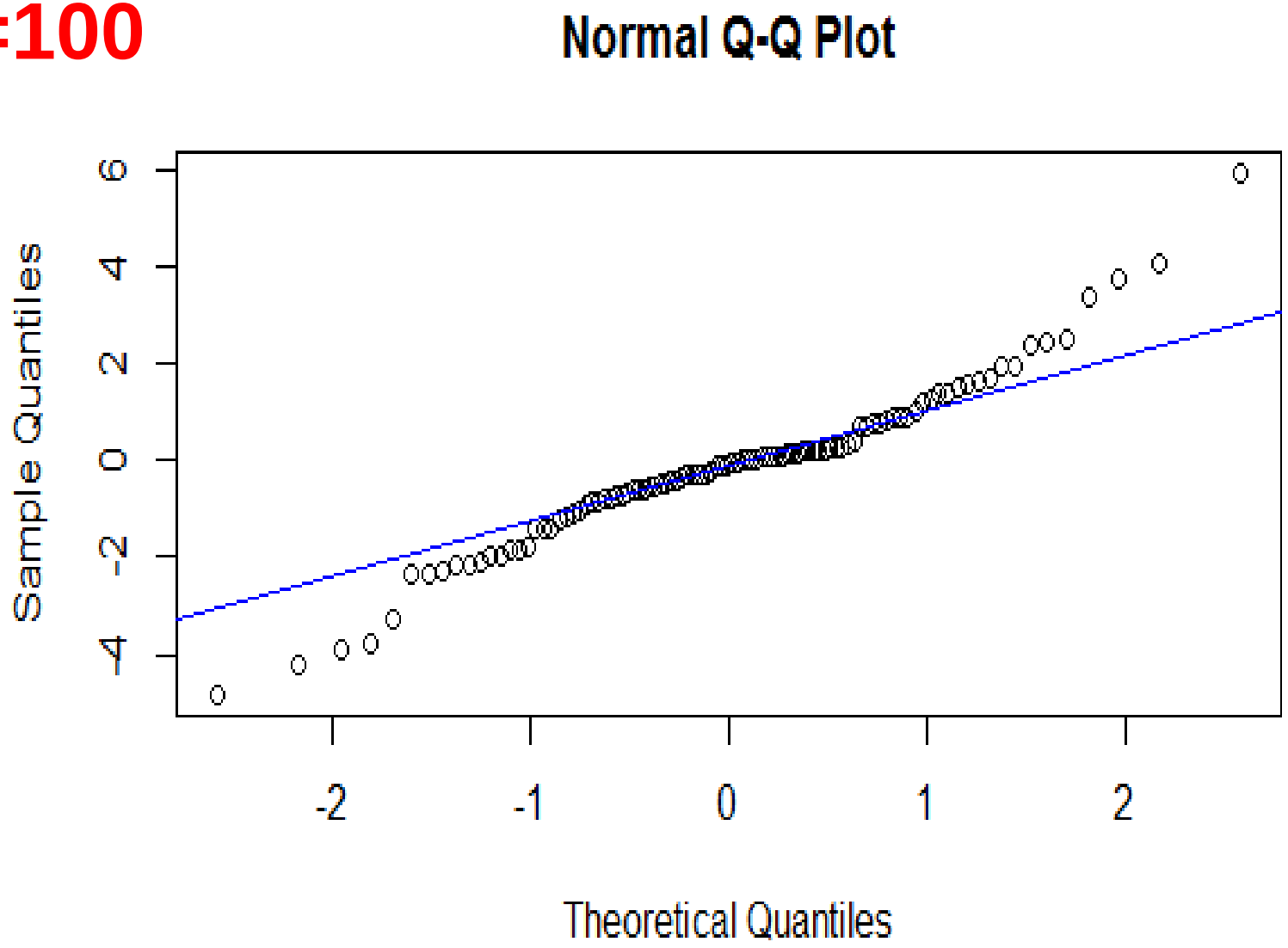
Qqplot for Simulated Gamma Data, n=100



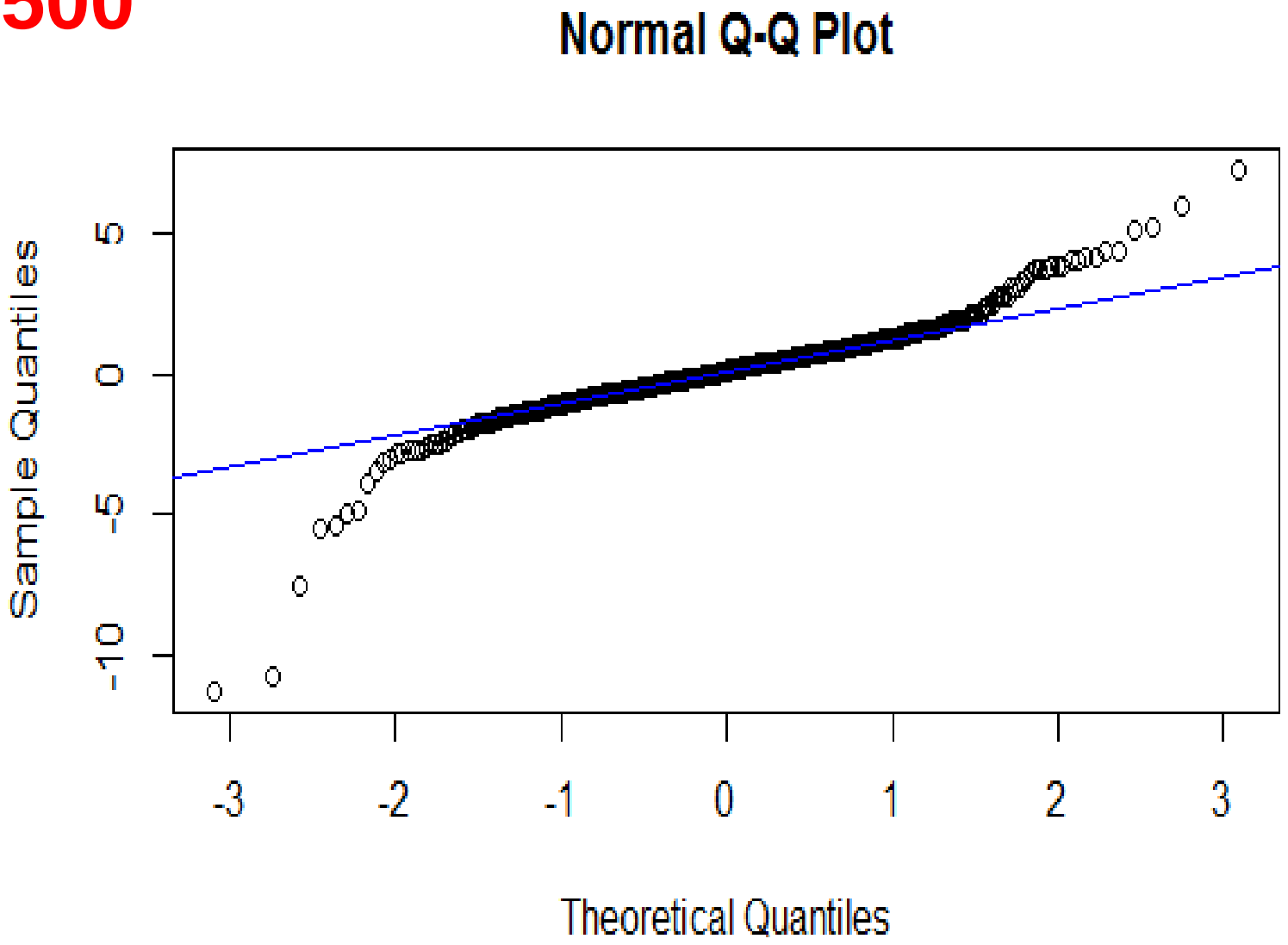
Qqplot for Simulated Gamma Data, n=500



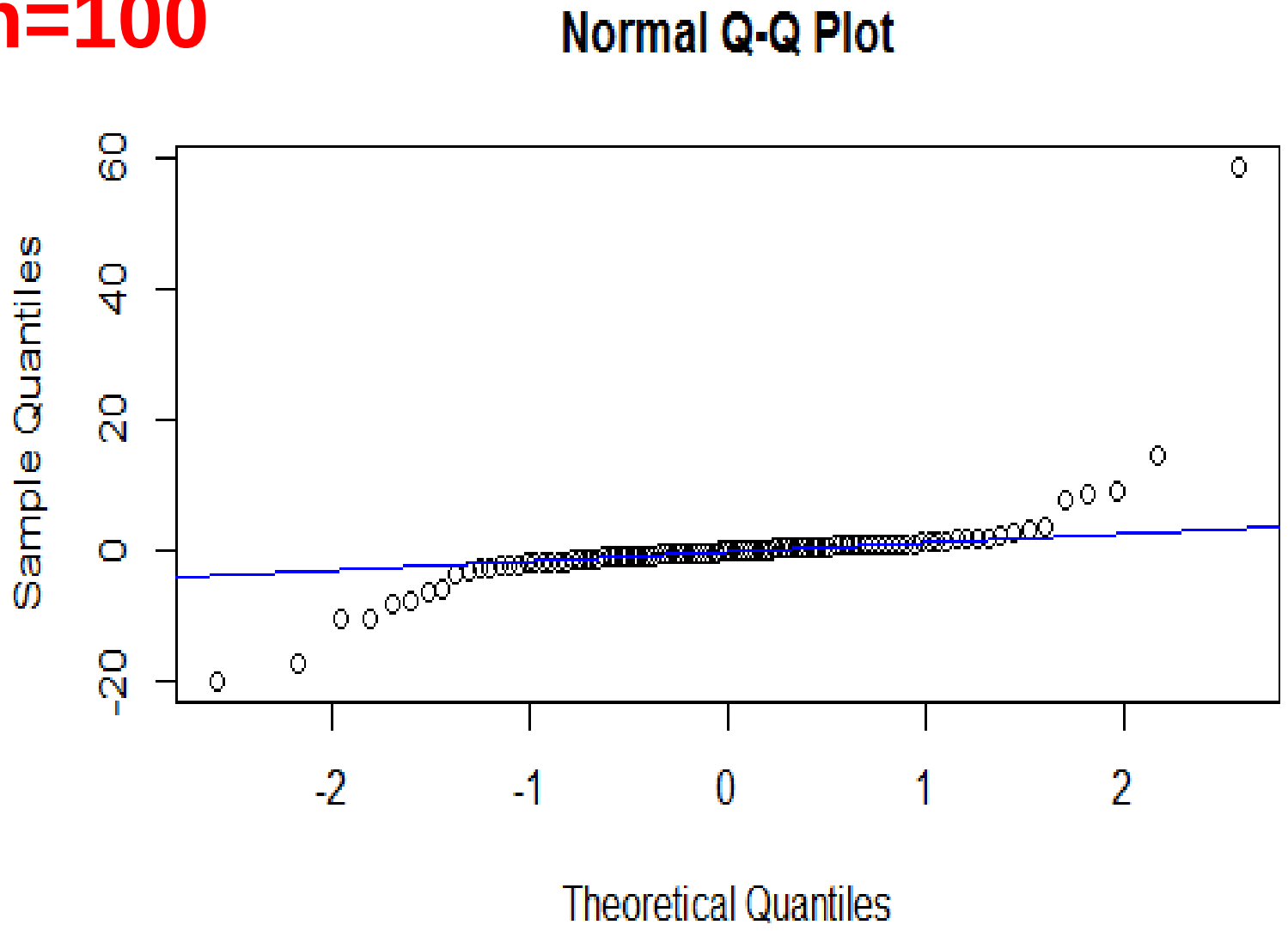
Qqplot for Simulated t Distribution Data, n=100



Qqplot for Simulated t Distribution Data, n=500



Qqplot for Simulated Cauchy Data, n=100



Qqplot for Simulated Cauchy Data, n=500