STA258H5
Statistics with Applied Probability

Al Nosedal and Omid Jazi

Winter 2023

INTRODUCING R AND ASSESSING NORMALITY

Homework

This course uses R. R is an open-source computing package which has seen a huge growth in popularity in the last few years. R can be downloaded from https://cran.r-project.org

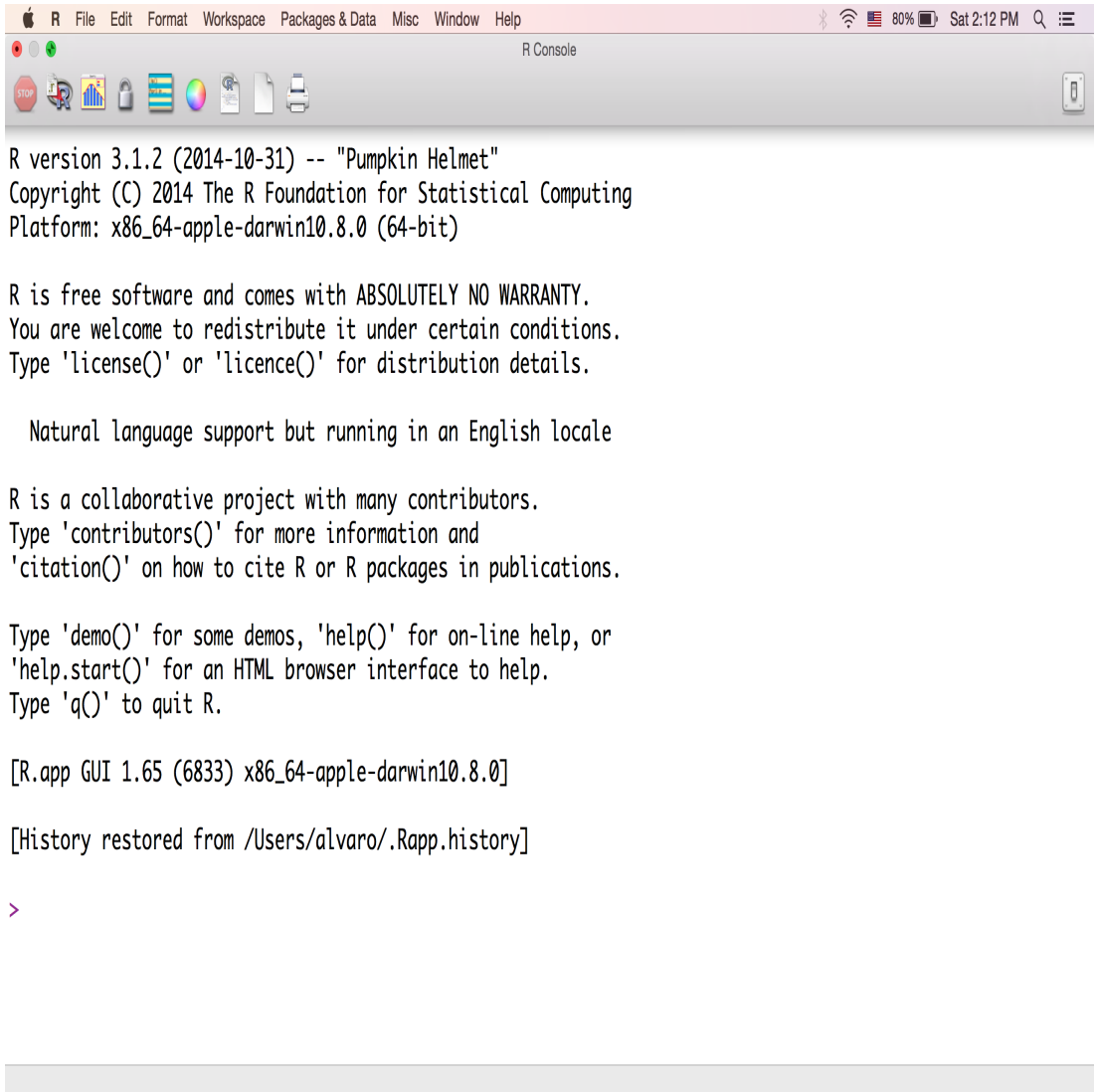
Please, download R and bring your laptop next time.

What exactly is R?

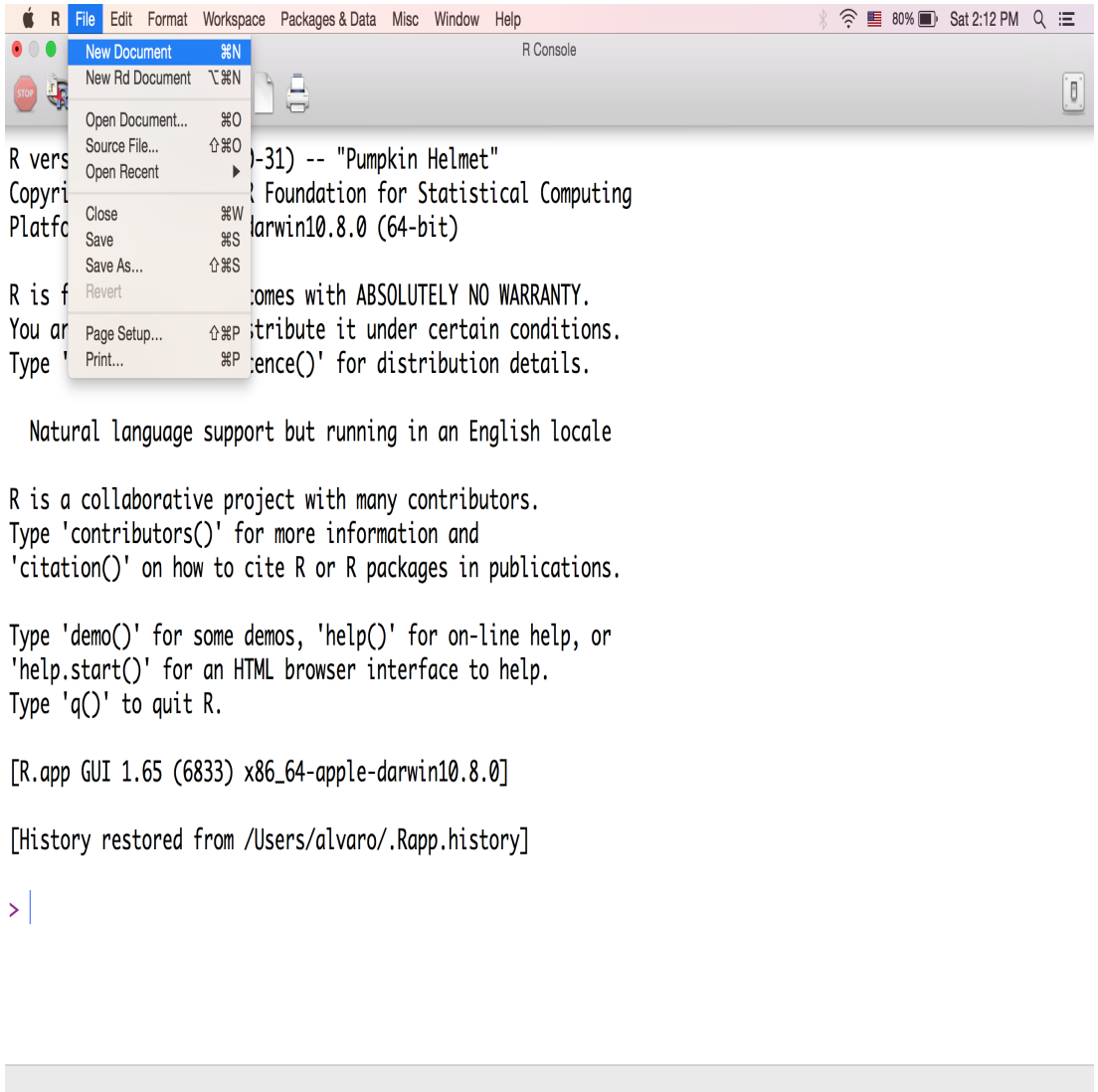
R is used for data manipulation, statistics, and graphics.
It is made of:

- operators ($+ - < - ?$) for calculations on vectors, arrays and matrices
- a huge collection of functions
- facilities for making unlimited types quality graphs
- user contributed packages (sets of related functions)
- the ability to interface with procedures written in C, C+, or FORTRAN and to write additional primitives.

The R GUI


R File Edit Format Workspace Packages & Data Misc Window Help
80% Sat 2:12 PM
R Console
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[R.app GUI 1.65 (6833) x86_64-apple-darwin10.8.0]
[History restored from /Users/alvaro/.Rapp.history]
>

Openning a script in R



What is RStudio?

RStudio is a relatively new editor specially targeted at R. RStudio is cross-platform, free and open-source software.

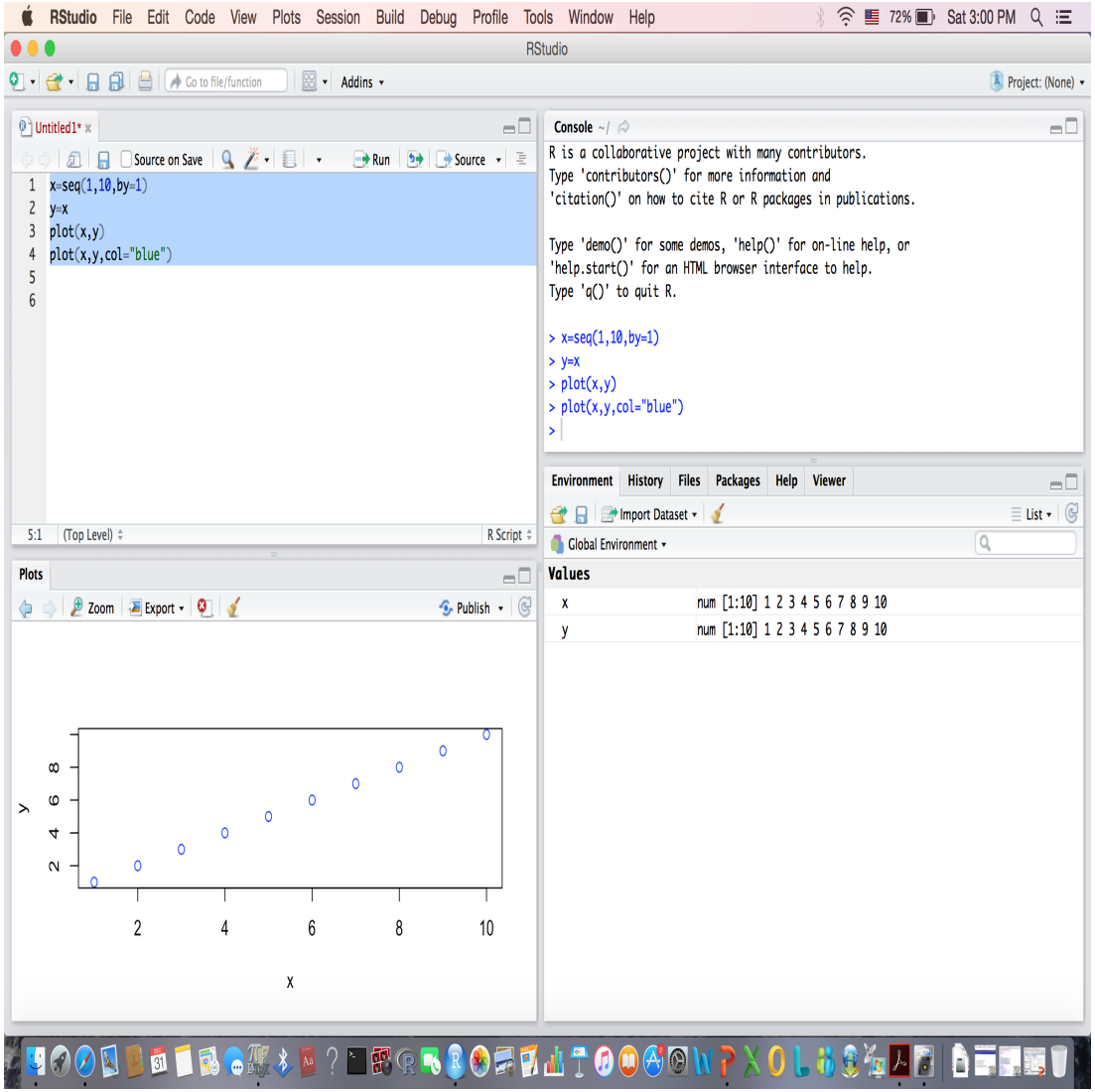
More homework: Obtaining RStudio

Just go to:

http://www.rstudio.com

download the corresponding file, execute it locally and follow the instructions given by the installer.

R Studio


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Sat 3:00 PM
RStudio
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 x=seq(1,10,by=1)
2 y=x
3 plot(x,y)
4 plot(x,y,col="blue")
5:1 (Top Level)
R Script
Console ~/
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
> x=seq(1,10,by=1)
> y=x
> plot(x,y)
> plot(x,y,col="blue")
Environment History Files Packages Help Viewer
Import Dataset
List
Global Environment
Values
x num [1:10] 1 2 3 4 5 6 7 8 9 10
y num [1:10] 1 2 3 4 5 6 7 8 9 10
Plots
Zoom
Export
Publish
y
2 4 6 8
2 4 6 8 10
x

Getting your data into R Studio

1. Save data file as a *something.csv* file.
2. Click on Import Dataset in top right window

Working in R Studio

Enter your commands in the console window
Arithmetic Operations

```
2+3;
```

```
## [1] 5
```

```
3-2;
```

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

```
3*2;
```

```
## [1] 6
```

Working in R Studio

Enter your commands in the console window
Arithmetic Operations

```
3/2;
```

```
## [1] 1.5
```

```
3^2;
```

```
## [1] 9
```

Working in R Studio

Enter your commands in the console window

Assignment

- To assign a value to a variable use $<-$

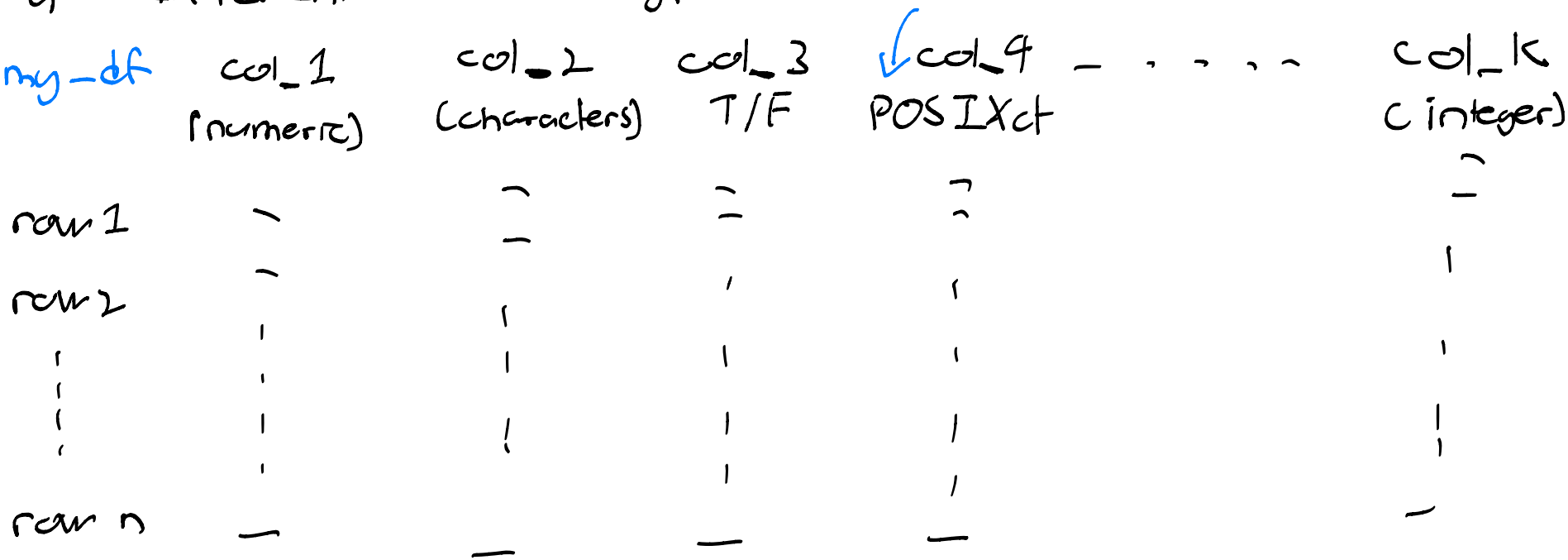
Example:

```
a<-19;
```

Data Frames

A tabular structure in R where each column can be a different data type

date / time



Isolate columns

rows , cols

`my_df$col_1` `my_df[ , 1]` `my_df[ , "col_1"]`

Summarizing Quantitative Data

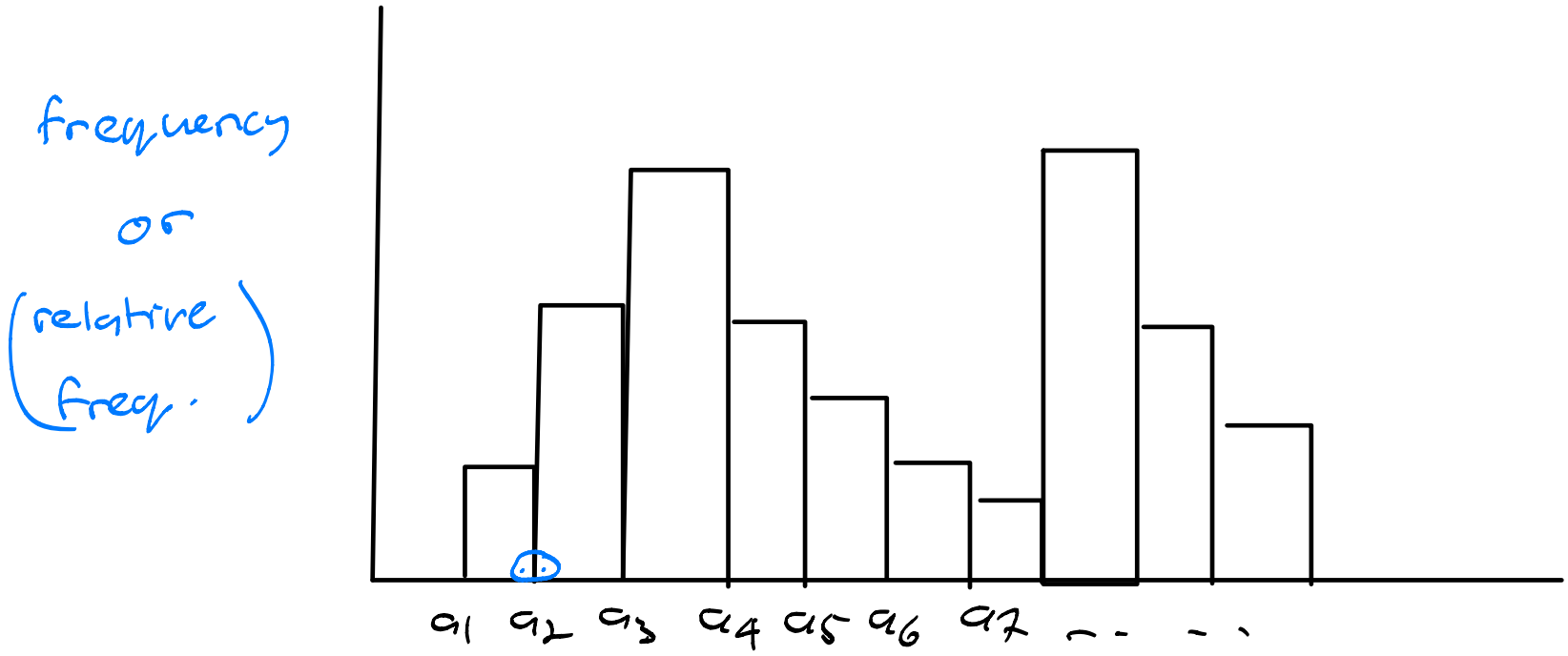
Histograms

A common graphical representation of quantitative data is a histogram. This graphical summary can be prepared for data previously summarized in either a frequency, relative frequency, or percent frequency distribution. A histogram is constructed by placing the variables of interest on the horizontal axis and the frequency, relative frequency, or percent frequency on the vertical axis.

Histograms

In R, bin intervals for histograms are <u>right</u> cts

Bin interval	Frequency	Relative frequency	(%)
$(a_1, a_2]$	f_1	f_1/N	(prop. obs in $[a_1, a_2]$
$(a_2, a_3]$	f_2	f_2/N	(" in $[a_2, a_3]$)
⋮	⋮	⋮	⋮
$(a_{k-1}, a_k]$	f_k	f_k/N	(
	N	1	



Bar plots are for qualitative data

- eg favourite colour, favourite flavour of ice-cream
- The bar widths are not meaningful

Histograms are for quantitative data

- The bin widths are meaningful

Advantages

- Relatively easy and simple way to visualize data (I get an idea of the "shape" of the distribution)
- Flexibility to modify bin widths.

Disadvantages

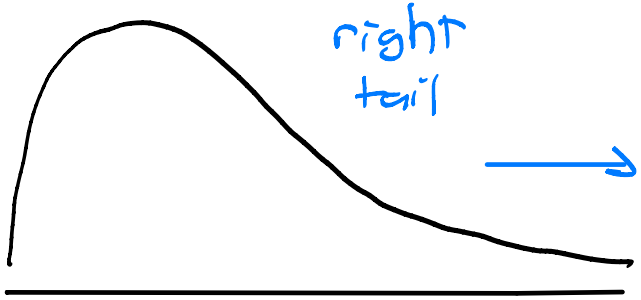
- Not suitable for small data sets
- Values close to break points are likely similar but they may be classified into different bins.

e.g: Student 1 gets 79 → B
Student 2 gets 80 → A

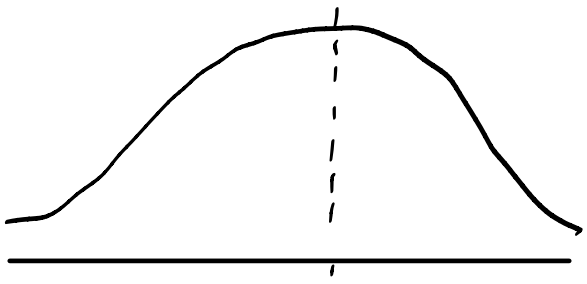
very similar but in 2 different bins

Skewness

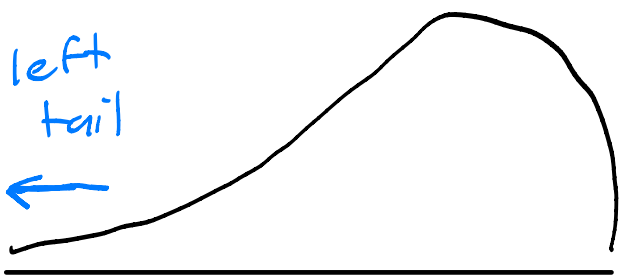
Describes the symmetry or concentration of values in a distribution



right skewed
(+ve ")



Symmetric



left skewed
(-ve skewed)

How to use help in R?

If you know which function you want help with simply use **help**. Example:

```
help(hist);
```

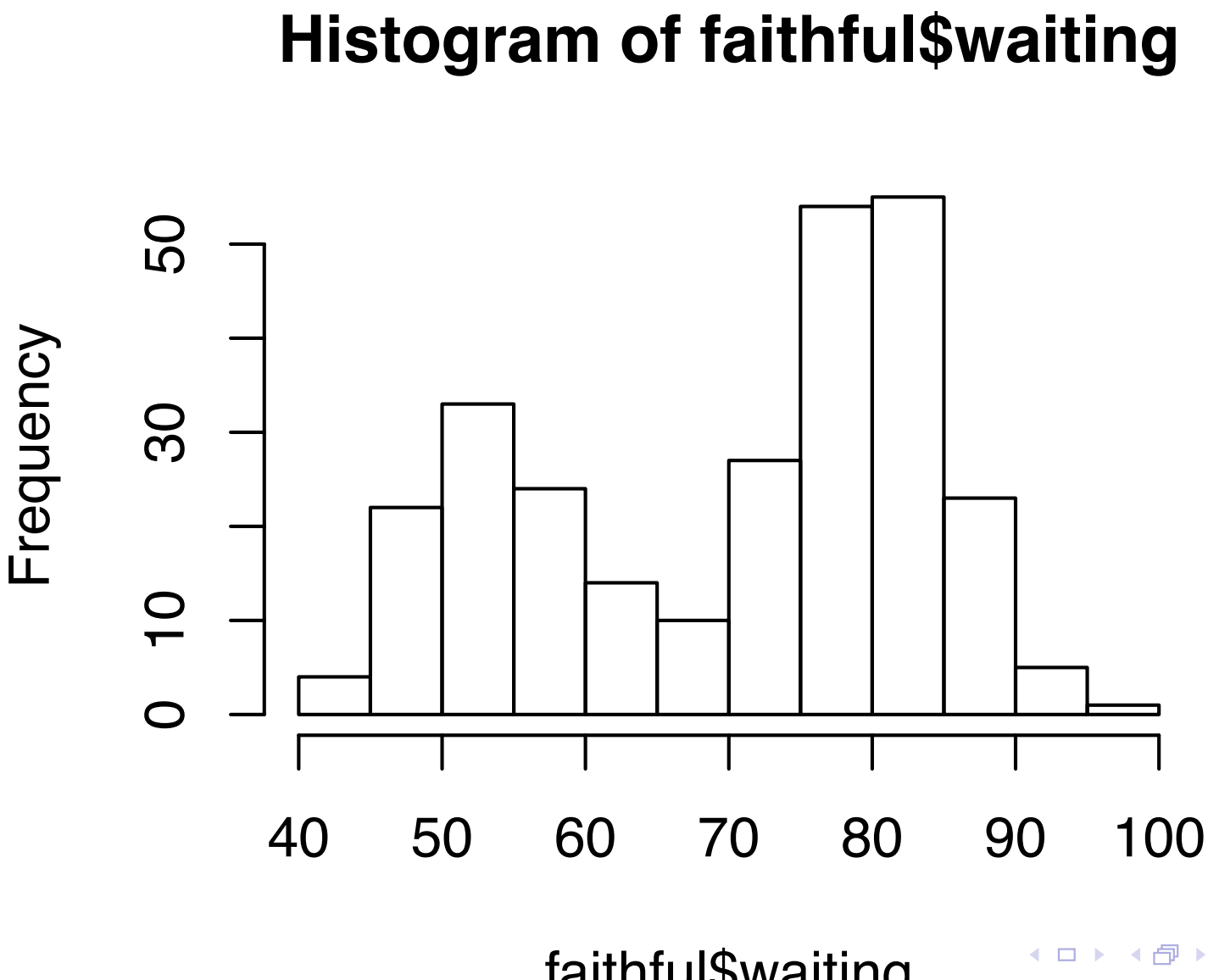
Old Faithful Geyser Data

Description
Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.
(A data frame with 272 observations on 2 variables.)

Histogram

```
# names of variables in faithful dataset;
names(faithful)


## Basic plot;
hist(faithful$waiting)
```

```
## [1] "eruptions" "waiting"
```



Histogram

```
# breakpoints between histogram cells;
hist(faithful$waiting,plot=FALSE)$breaks

##  [1]  40  45  50  55  60  65  70  75  80  85  90  95
## [13] 100

hist(faithful$waiting,plot=FALSE)$counts

##  [1]  4 22 33 24 14 10 27 54 55 23  5  1
```

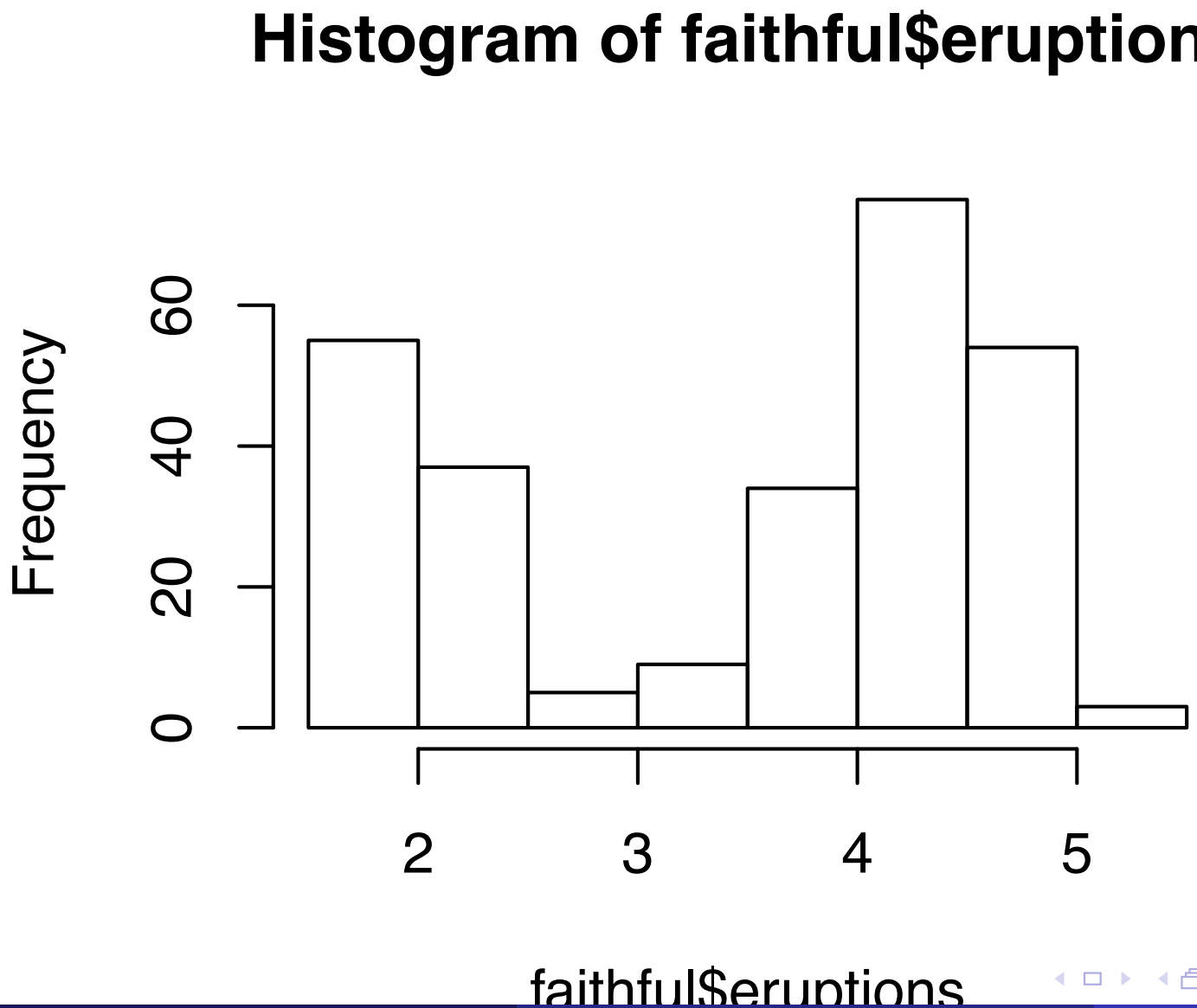
Histogram

```
## First six observations of data set;

head(faithful);

## Basic plot;

hist(faithful$eruptions);
```

First six observations

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

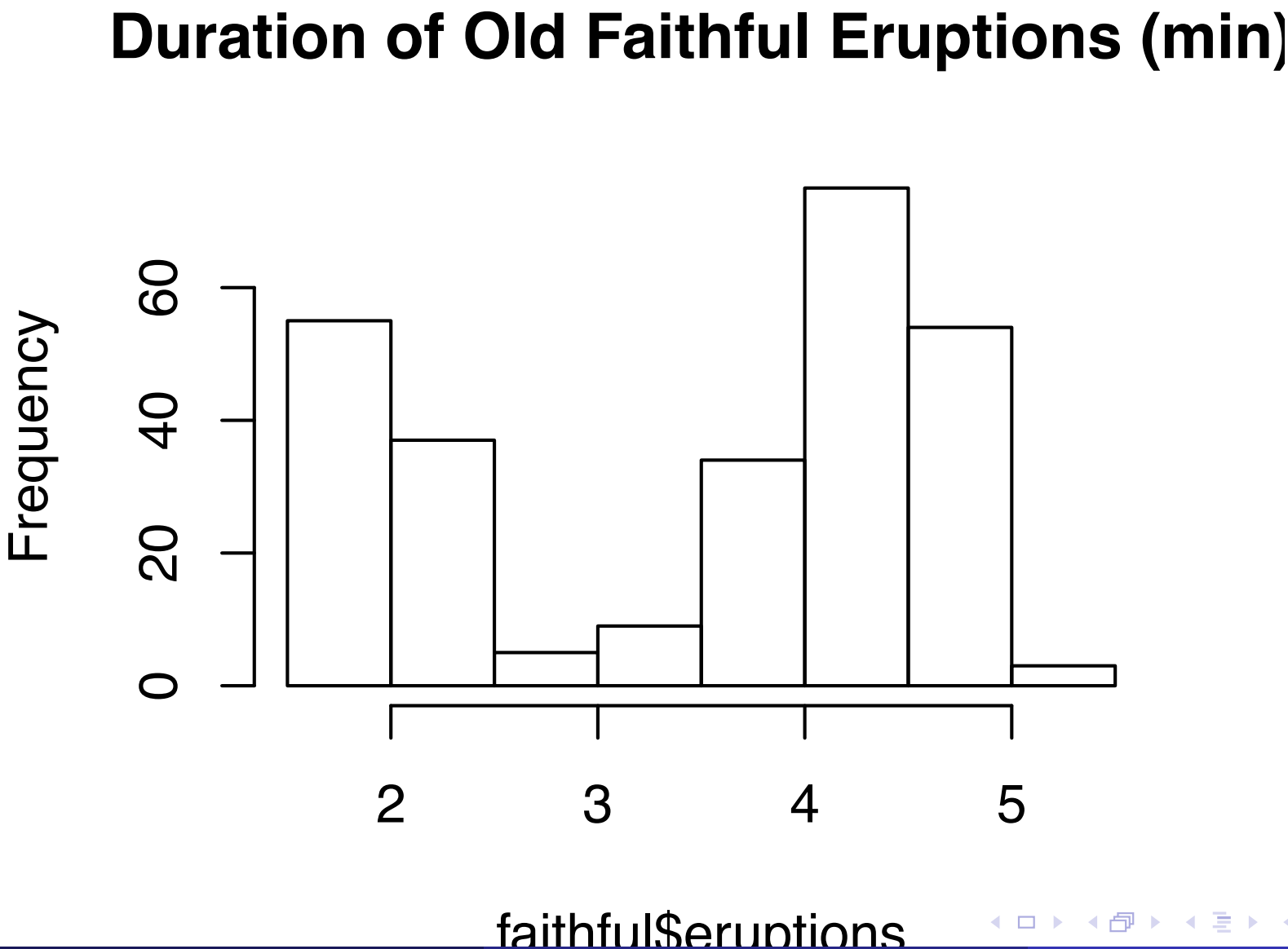
Histogram


Histogram of faithful$eruptions
Frequency
0
20
40
60
2
3
4
5
faithful$eruptions

Histogram (with title)

```
## Nicer plot.

hist(faithful$eruptions,
main="Duration of Old Faithful Eruptions (min)");
```

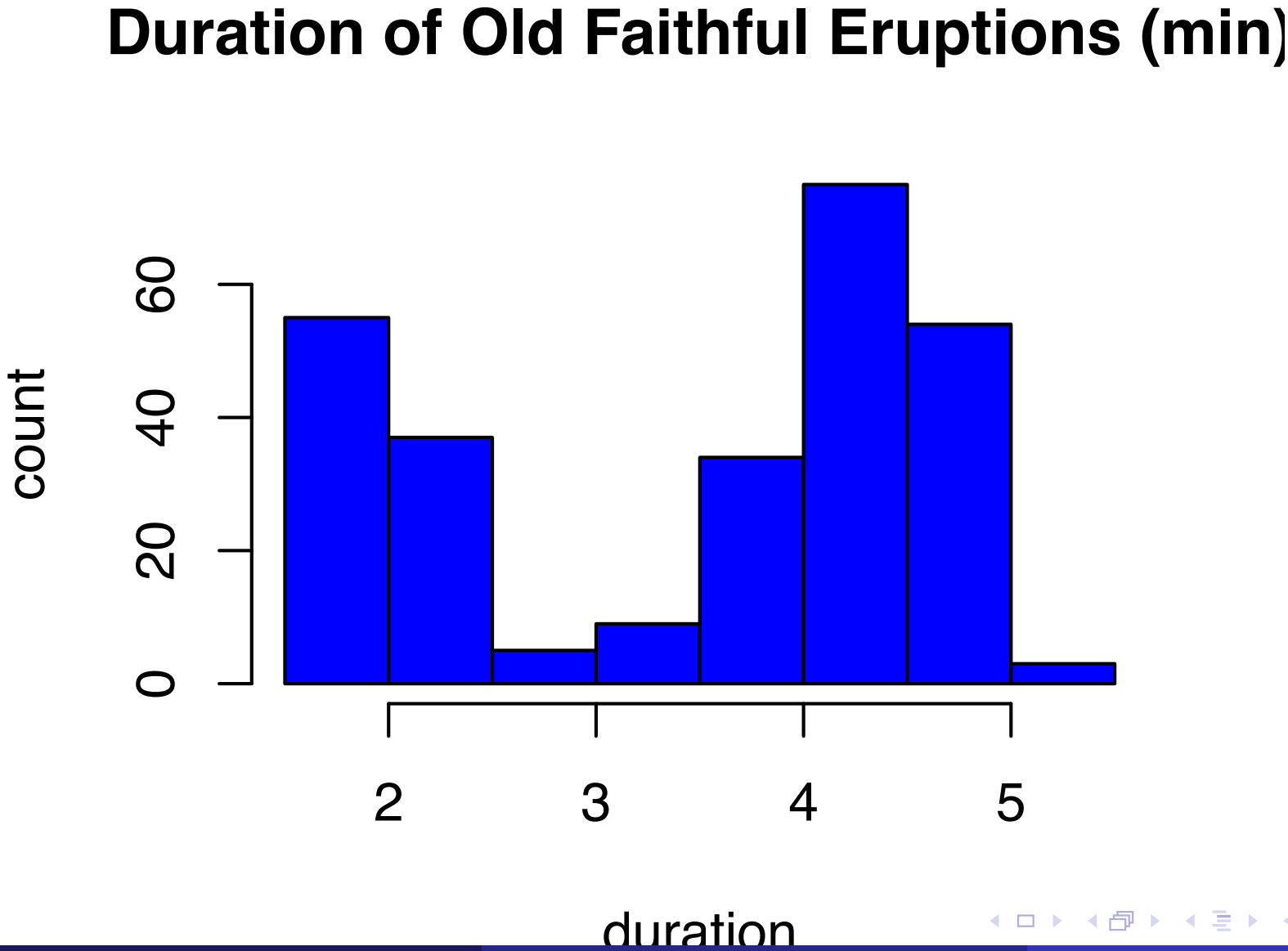
Histogram (with title)



Histogram

```
## Add axes labels and color.

hist(faithful$eruptions,
main="Duration of Old Faithful Eruptions (min)",
xlab="duration",ylab="count", col="blue");
```

Histogram


Duration of Old Faithful Eruptions (min)
count
0
20
40
60
2
3
4
5
duration

ggplot2

The **ggplot2** package is designed to have a syntax that is consistent across all graphic types; that is to say, the command language is very similar from one type of graph to another.

R Code (installing library)

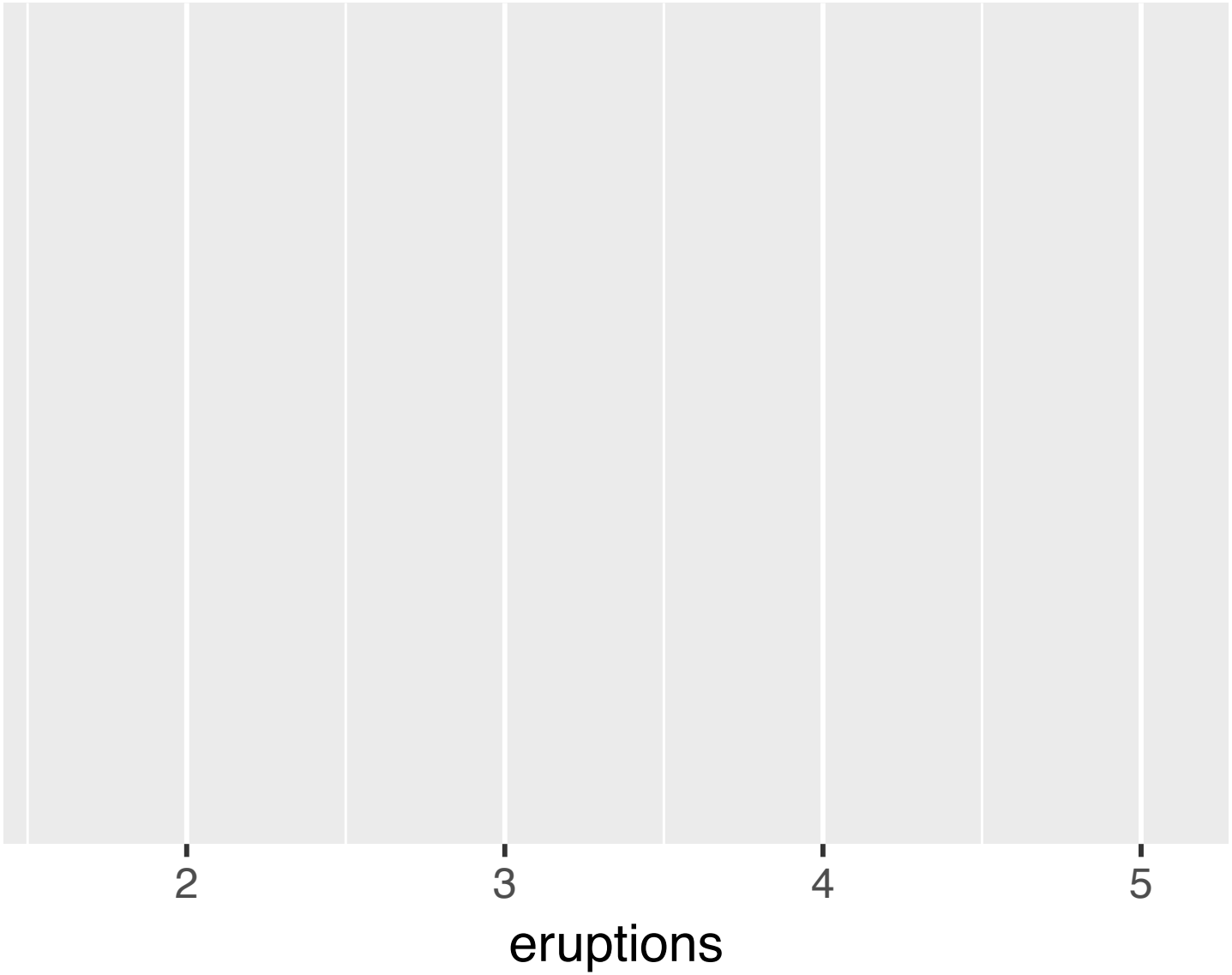
```
install.packages("ggplot2");
```

R Code (loading library)

```
library(ggplot2);
```

Histogram (again)

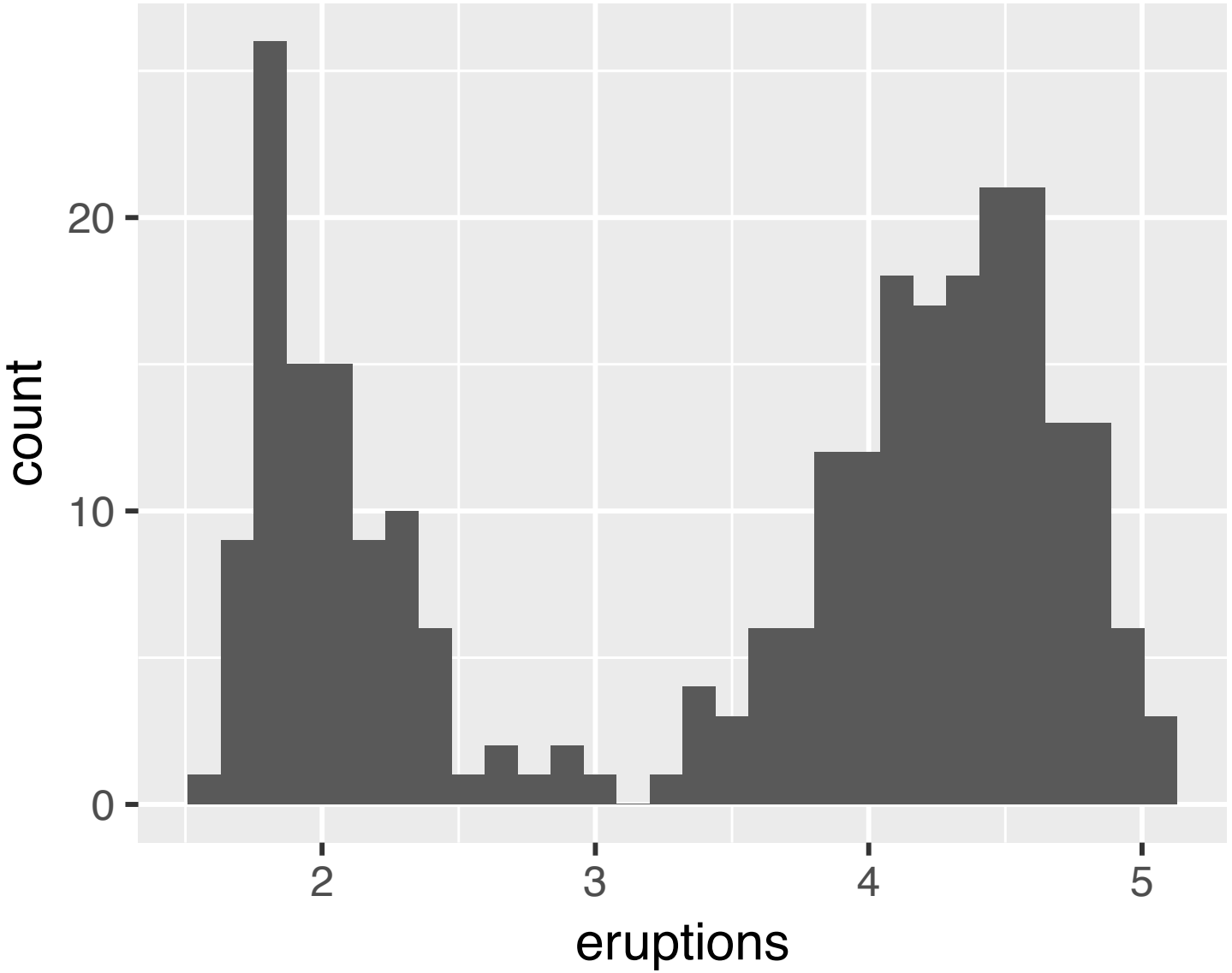
```
# loading library;
library(ggplot2);

# preparing canvas;
ggplot(data=faithful,mapping=aes(x=eruptions));
```


2
3
4
5
eruptions

Histogram (again)

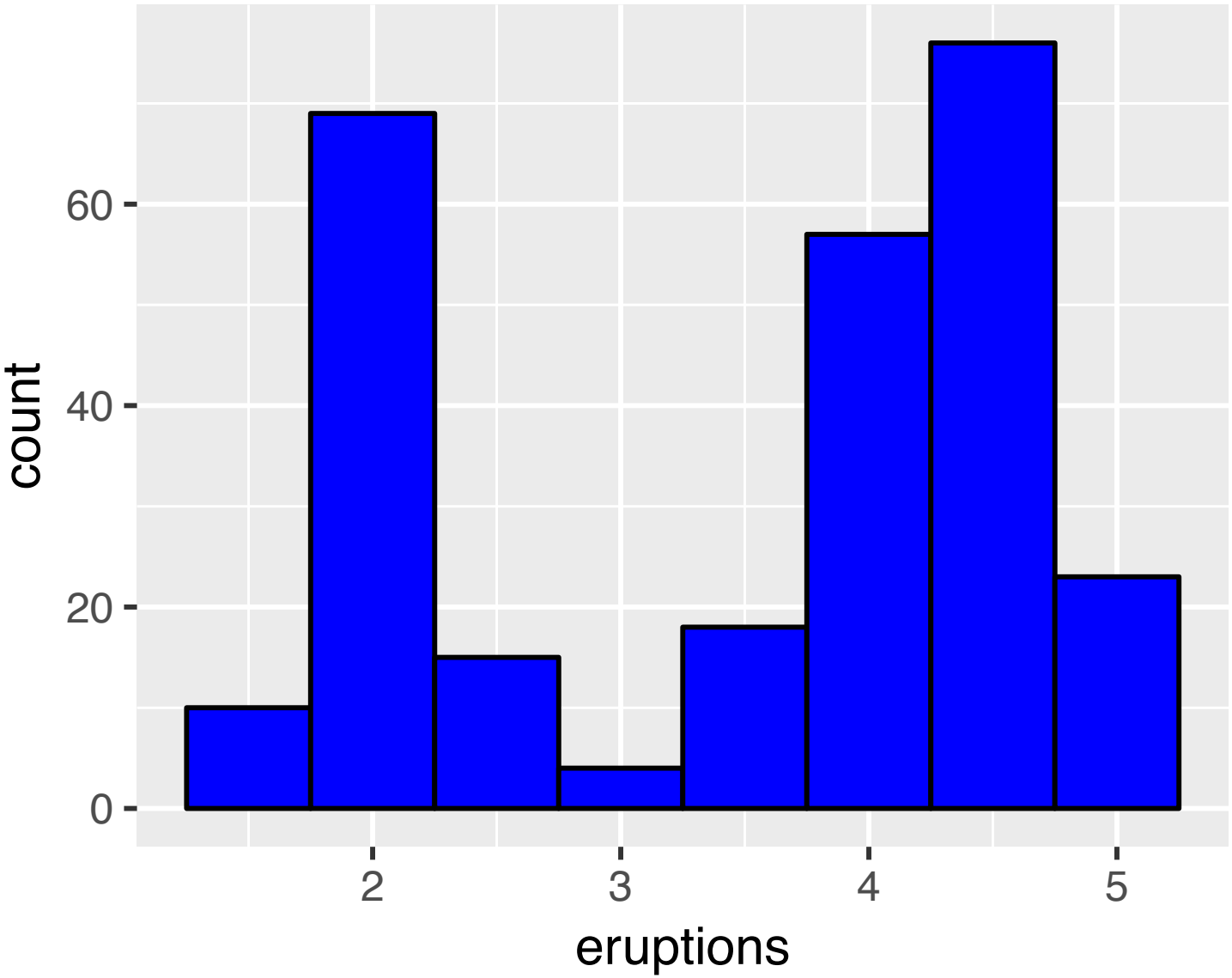
```
# adding layer with histogram;

ggplot(data=faithful,mapping=aes(x=eruptions))+
geom_histogram()
```

```
## `stat_bin()` using `bins = 30`.  Pick better value
## with `binwidth`.
```



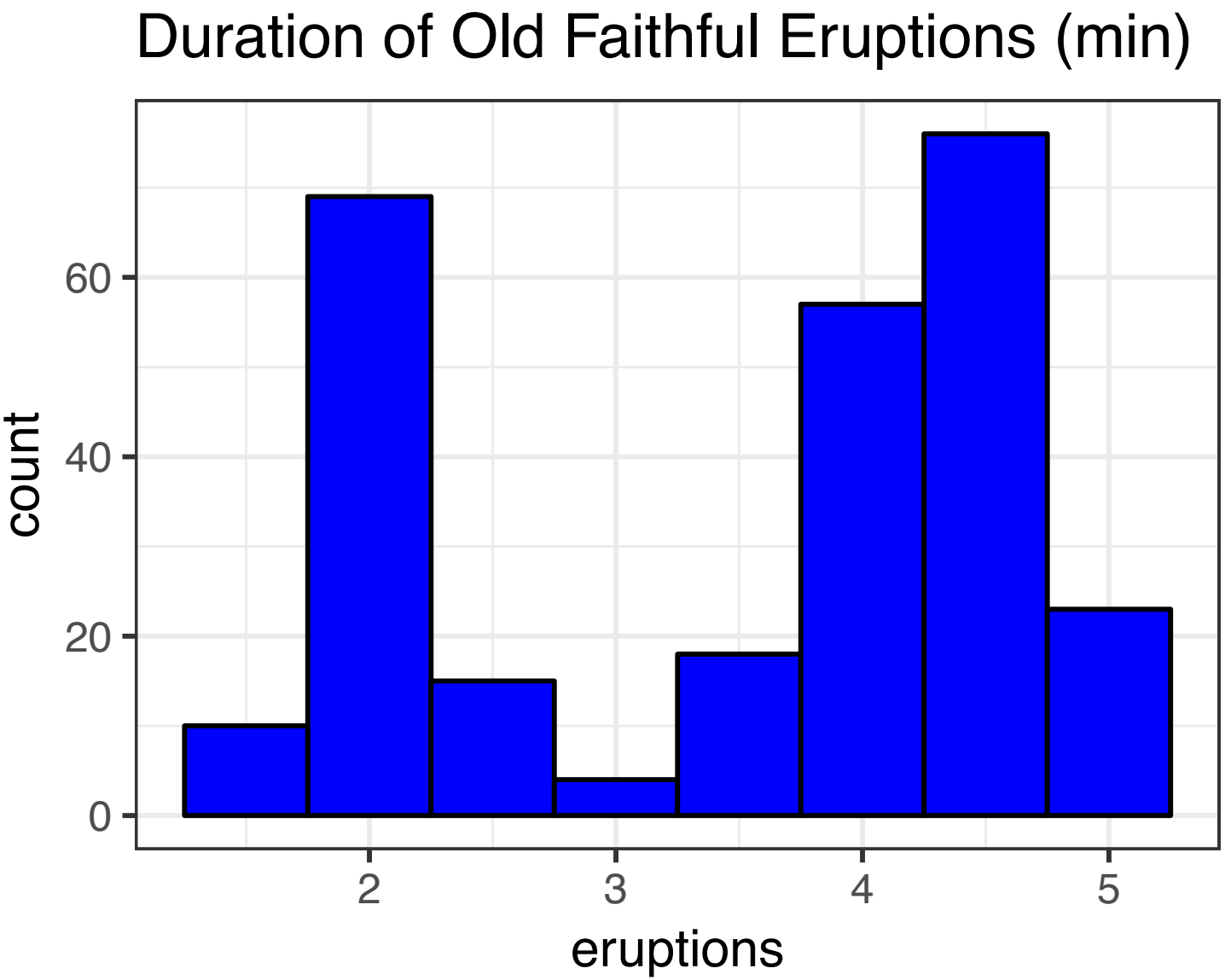
Histogram (again)

```
# changing binwidth and color;
ggplot(data=faithful,mapping=aes(x=eruptions))+
geom_histogram(color="black",fill="blue",binwidth=0.5)
```


count
60
40
20
0
2
3
4
5
eruptions

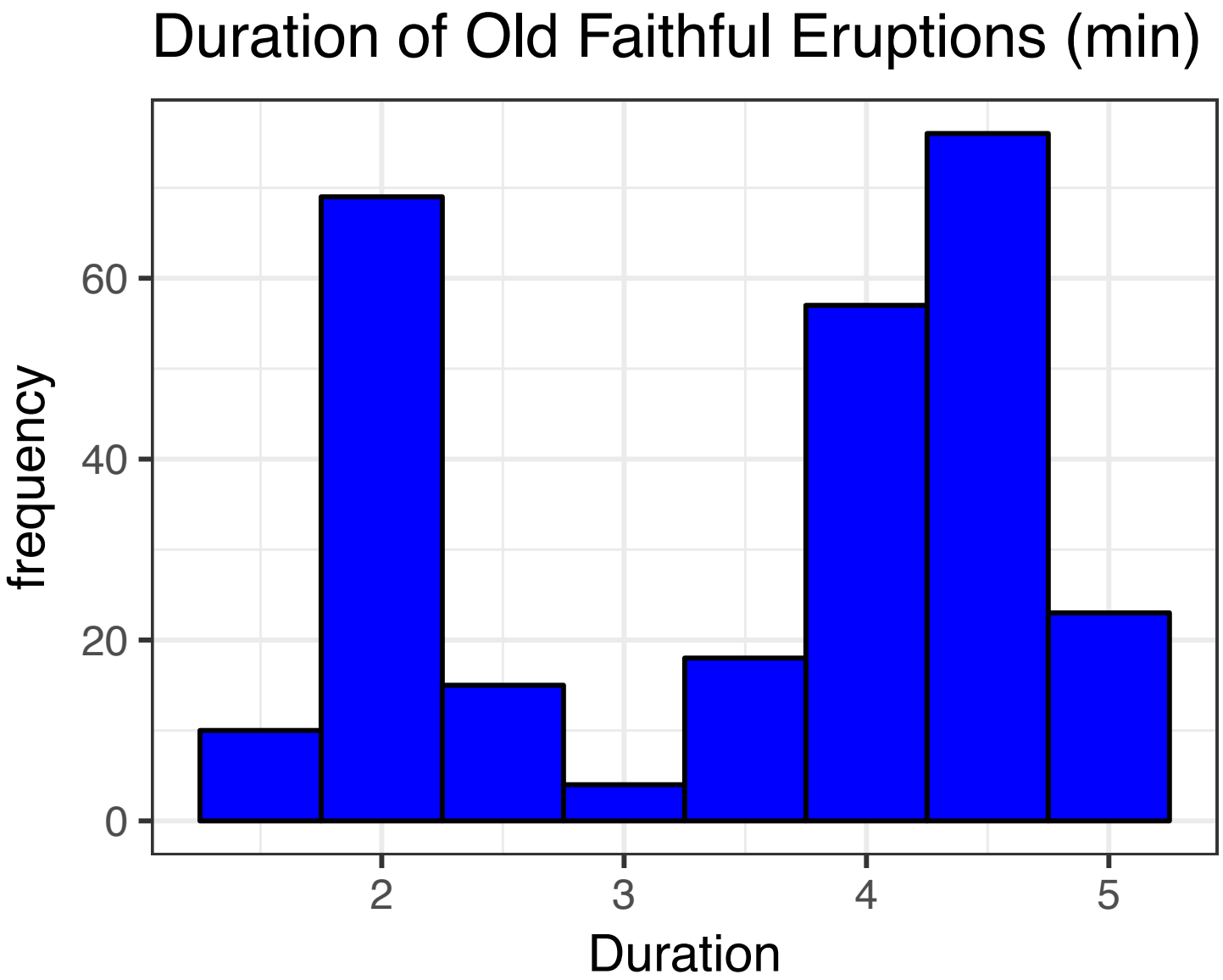
Histogram (again)

```
# adding main title and using theme_bw;
ggplot(data=faithful,mapping=aes(x=eruptions))+
geom_histogram(color="black",fill="blue",binwidth=0.5)+
labs(title="Duration of Old Faithful Eruptions (min)")+
theme_bw()
```


Duration of Old Faithful Eruptions (min)
count
60
40
20
0
2
3
4
5
eruptions

Histogram (again)

```
# changing labels for x-axis and y-axis;

ggplot(data=faithful,mapping=aes(x=eruptions))+
geom_histogram(color="black",fill="blue",binwidth=0.5)+
labs(title="Duration of Old Faithful Eruptions (min)",
x="Duration",y="frequency")+
theme_bw()
```


Duration of Old Faithful Eruptions (min)
frequency
60
40
20
0
2
3
4
5
Duration

Let $X_1, X_2, \dots, X_n$ be a sample of data points which are indep draws from the same distrib

Sample mean ($\bar{x}$)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

centre of mass

Sample Variance (s^2)

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

squared dispersion

$(\text{units})^2$

The average squared distance relative to the mean

Standard Deviation (s)

$$s = \sqrt{s^2}$$

units

Measure of dispersion relative to mean in same units.

Example	Section A	Section B
	50, 60, 70, 80, 90	68, 69, 70, 71, 72
Sample mean ($\bar{x}$)	70	70
Sample Var (s^2)	← large →	↔ small
Sample SD (s)		

Median (M)

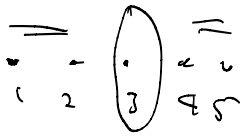
Midpoint (middle value). Half of all values above M
" " " " below M

Order data from smallest to largest (ascending)

$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$

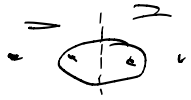
Case 1: n is odd

Median is value in $\left(\frac{n+1}{2}\right)$ position



Case 2: n is even

Median is average of $\left(\frac{n}{2}\right)$ and $\left(\frac{n+1}{2}\right)$ values



Percentiles

The p^{th} percentile is a value such that p% of observations are below it

Quartiles

Special cases of percentiles

Q_1: The 25^{th} percentile (25% of values are below it)

median Q_2: 50^{th} " (50% " " " " ")

Q_2: 75^{th} " (75% " " " " ")

Interquartile Range (IQR)

$$IQR = Q_3 - Q_1$$

Outliers

unusually large or small data points relative to other

Example (Slide 67)

4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Median (M)

$n = 15$ is odd

M is value in $\frac{n+1}{2} = \frac{15+1}{2} = 8^{th}$ pos

$M = 35$

use to find Q_3

Q_1

4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.

1 2 3 4 5 6 7 1 2 3 4 5 6 7

M

$n = 7$ is odd

Q_1 is value in $\frac{7+1}{2} = 4^{th}$ pos $\quad Q_1 = 30$

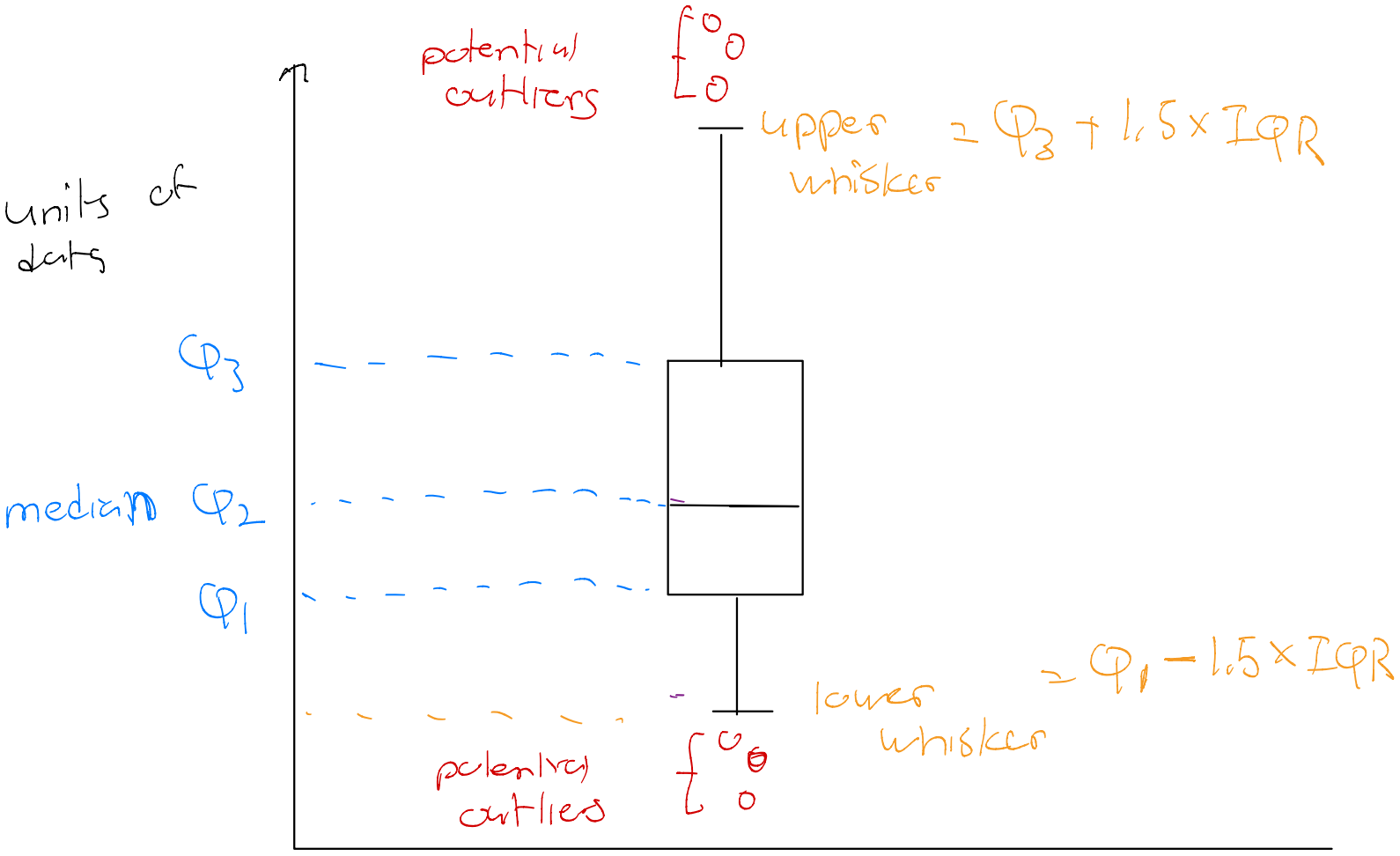
Note: R may give slightly different values for quartiles (R may perform interpolation)

Five Number Summary

min, Q_1, Q_2, Q_3, max

Boxplot

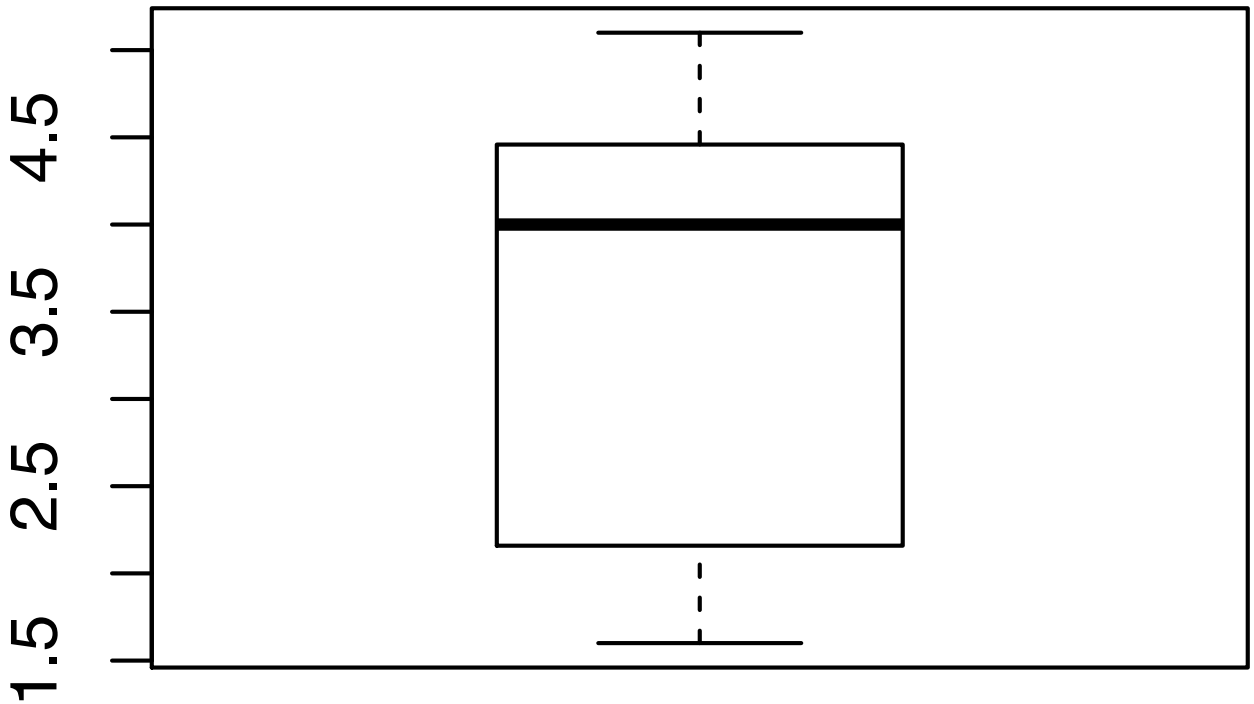
A relatively simple visualization of data used to determine skewness and potential outliers



Boxplot

```
## Basic plot.

boxplot(faithful$eruptions);
```

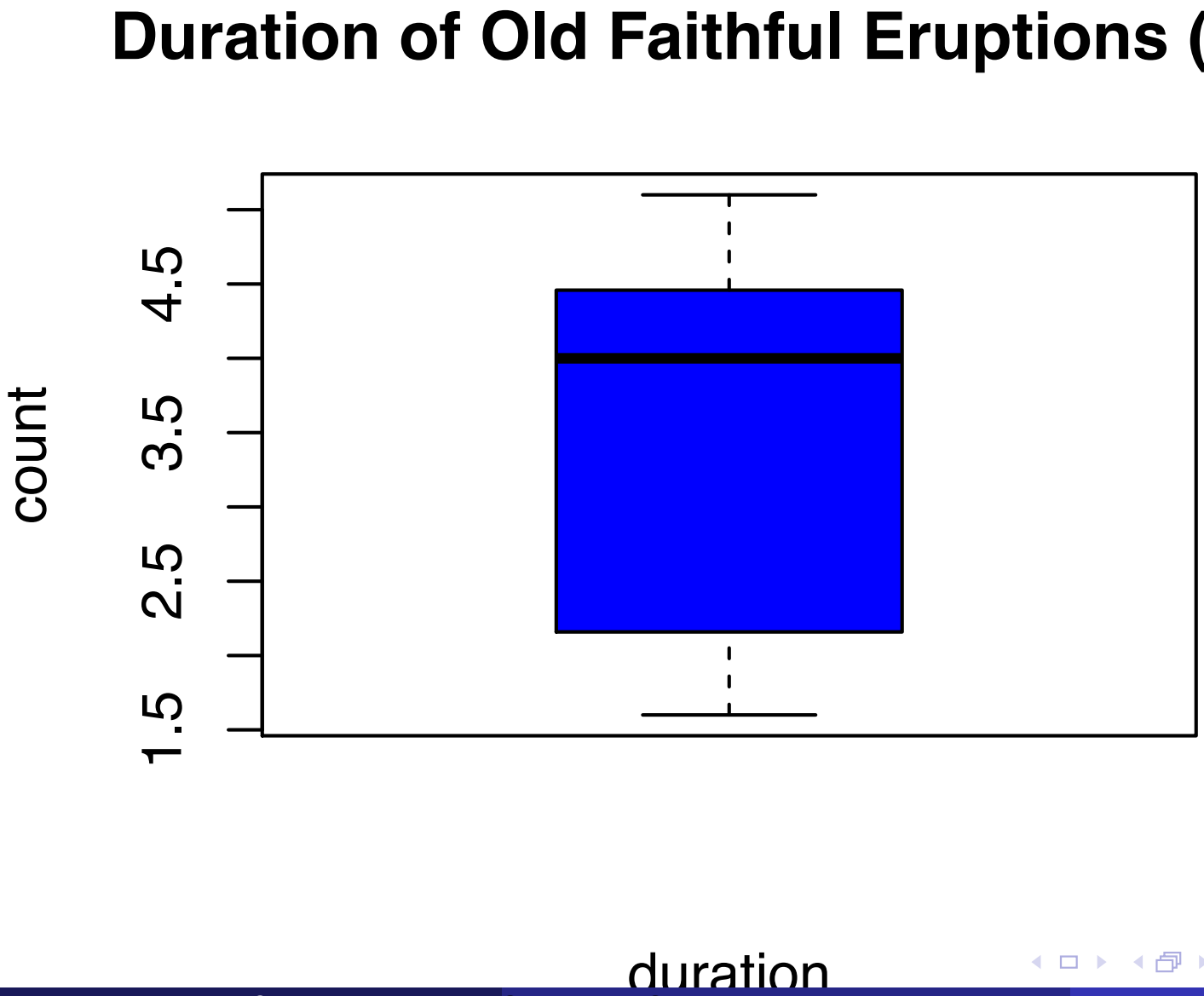
Boxplot



Boxplot

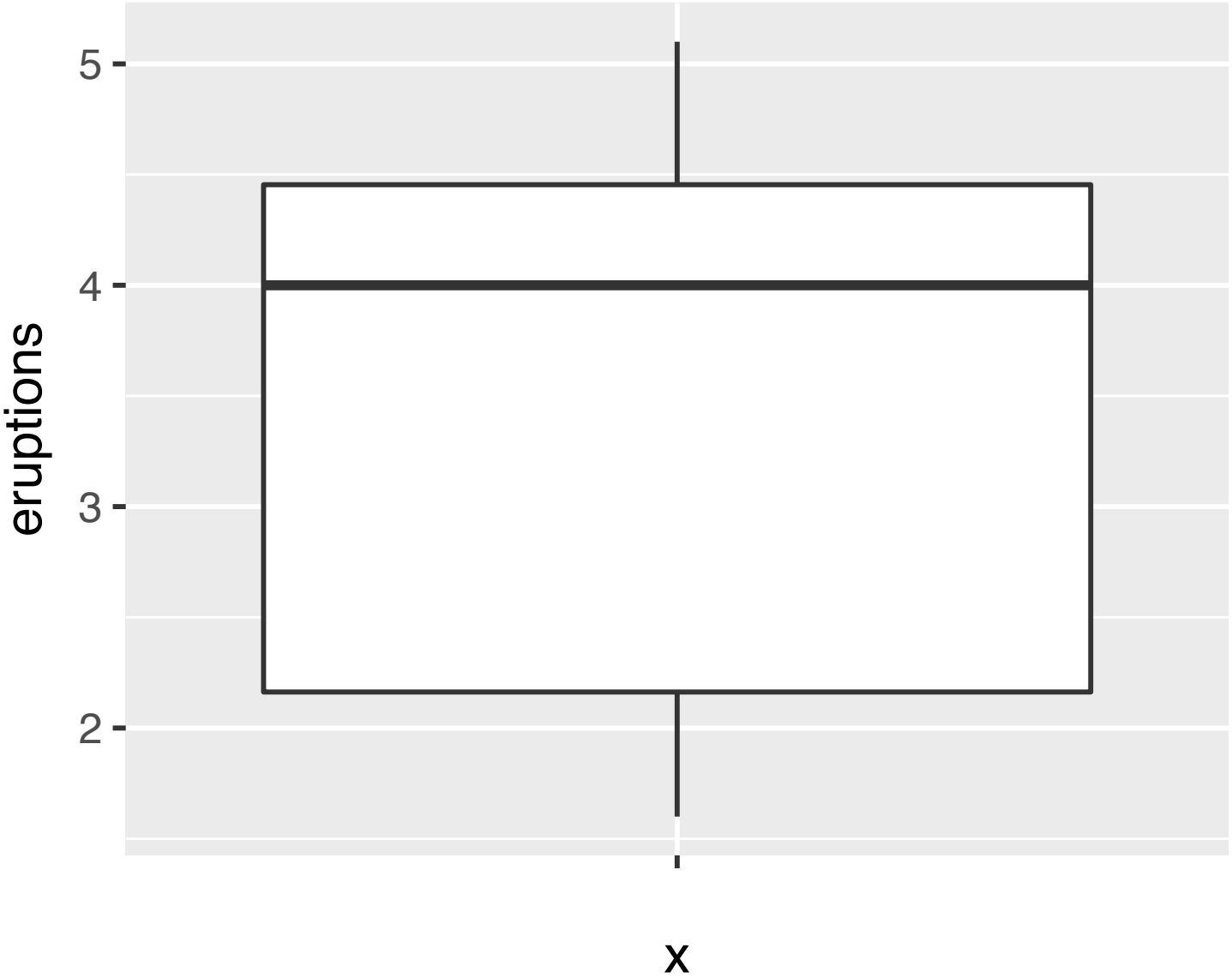
```
## Add axes labels and color.

boxplot(faithful$eruptions,
main="Duration of Old Faithful Eruptions (min)",
xlab="duration",ylab="count", col="blue");
```

Boxplot


Duration of Old Faithful Eruptions (min)
count
1.5
2.5
3.5
4.5
duration

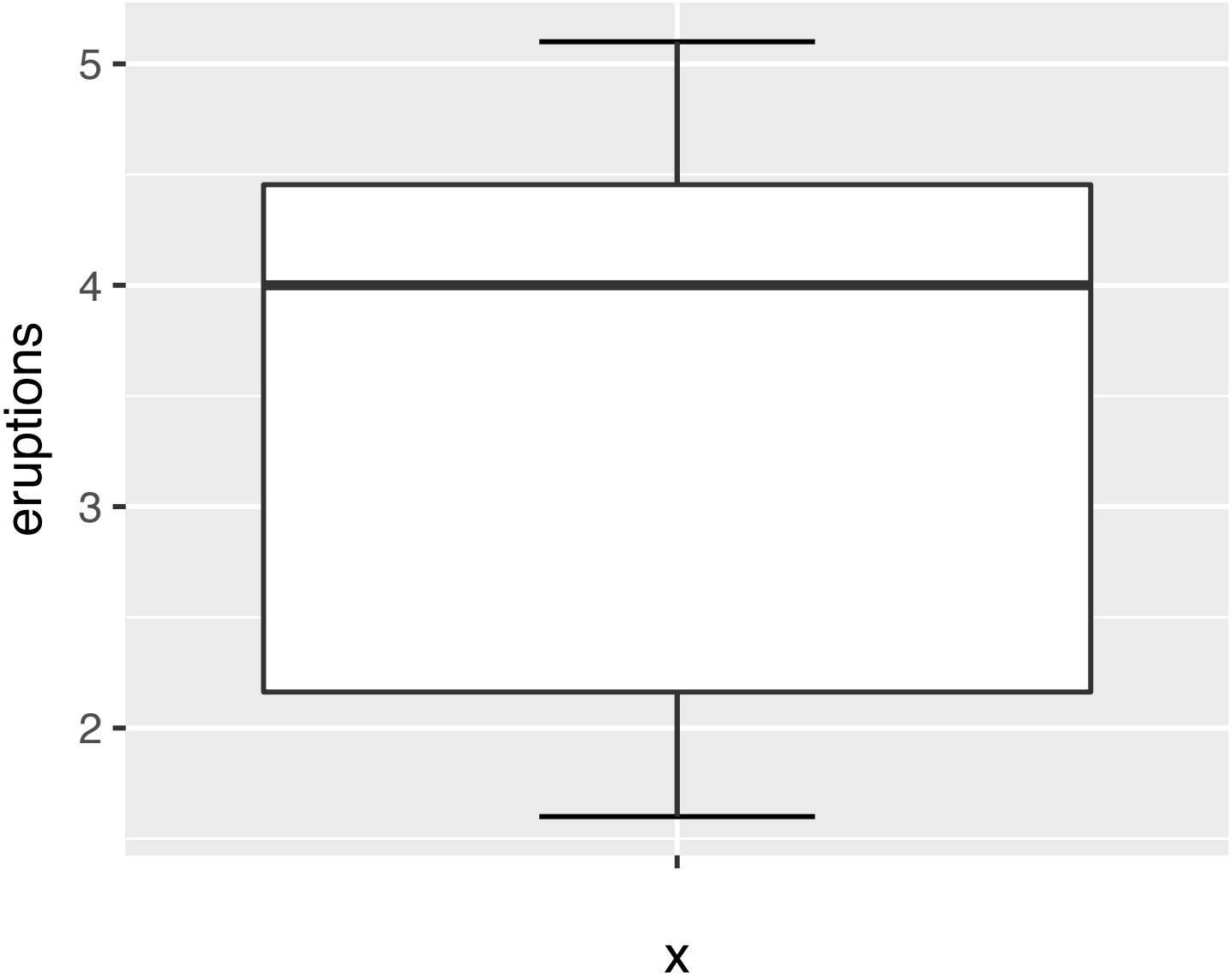
Boxplot (again)

```
# basic boxplot;
ggplot(data=faithful,mapping=aes(x=" ",y=eruptions))+
geom_boxplot()
```


5
4
3
2
eruptions
x

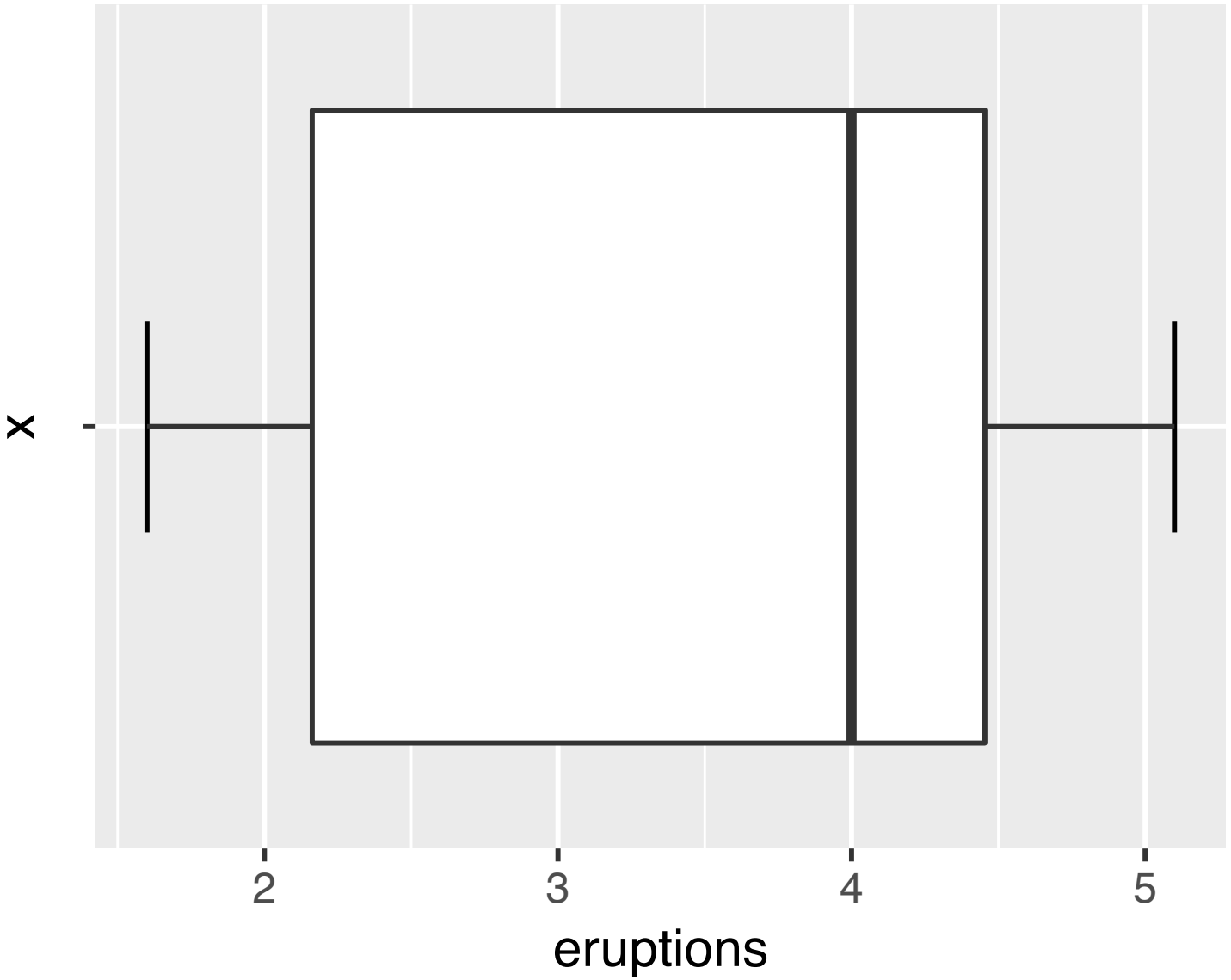
Boxplot (again)

```
# with whiskers;
ggplot(data=faithful,mapping=aes(x=" ",y=eruptions))+
stat_boxplot(geom="errorbar",width=0.25)+
geom_boxplot()
```


5
4
3
2
eruptions
x

Boxplot (again)

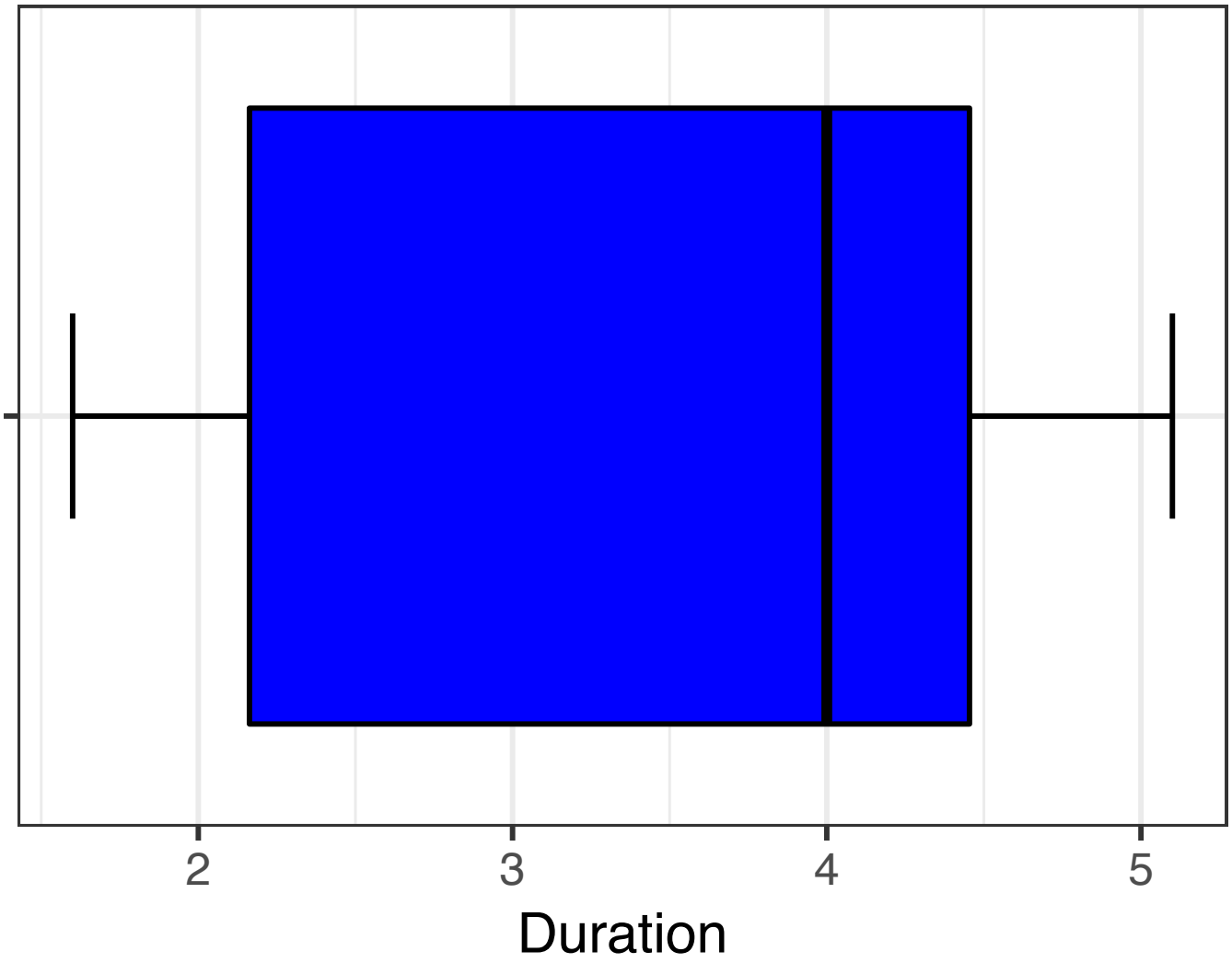
```
# horizontal;
ggplot(data=faithful,mapping=aes(x=" ",y=eruptions))+
stat_boxplot(geom="errorbar",width=0.25)+
geom_boxplot()+
coord_flip()
```


x
2
3
4
5
eruptions

Boxplot (again)

```
# with titles and color;
ggplot(data=faithful,mapping=aes(x=" ",y=eruptions))+
stat_boxplot(geom="errorbar",width=0.25,color="black")+
geom_boxplot(color="black",fill="blue")+
coord_flip()+
labs(title="Duration of Old Faithful Eruptions (min)",
x=" ",y="Duration")+
theme_bw()
```

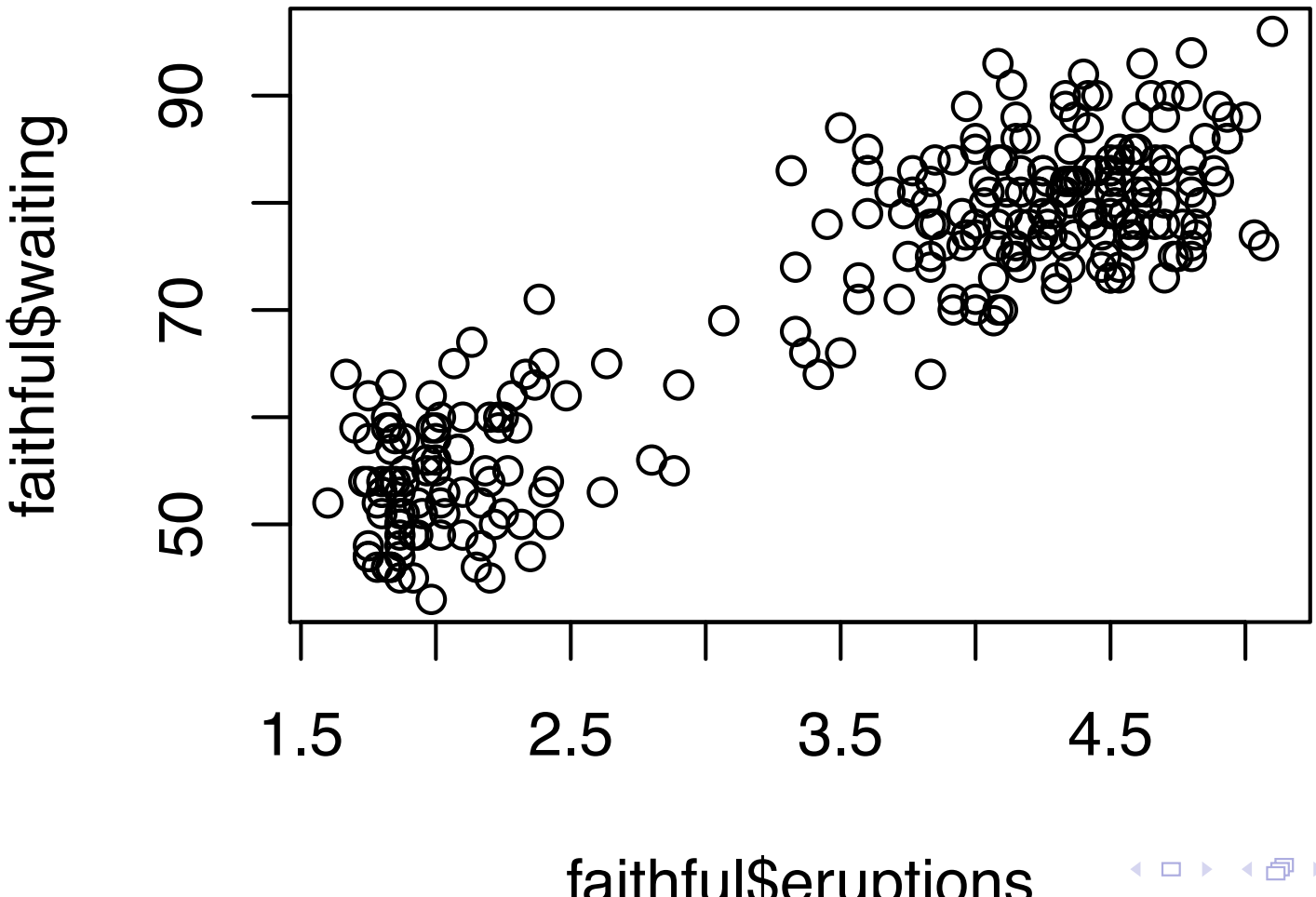
Duration of Old Faithful Eruptions (min)



Scatterplot

```
## Basic plot.

plot(faithful$eruptions,faithful$waiting);
```

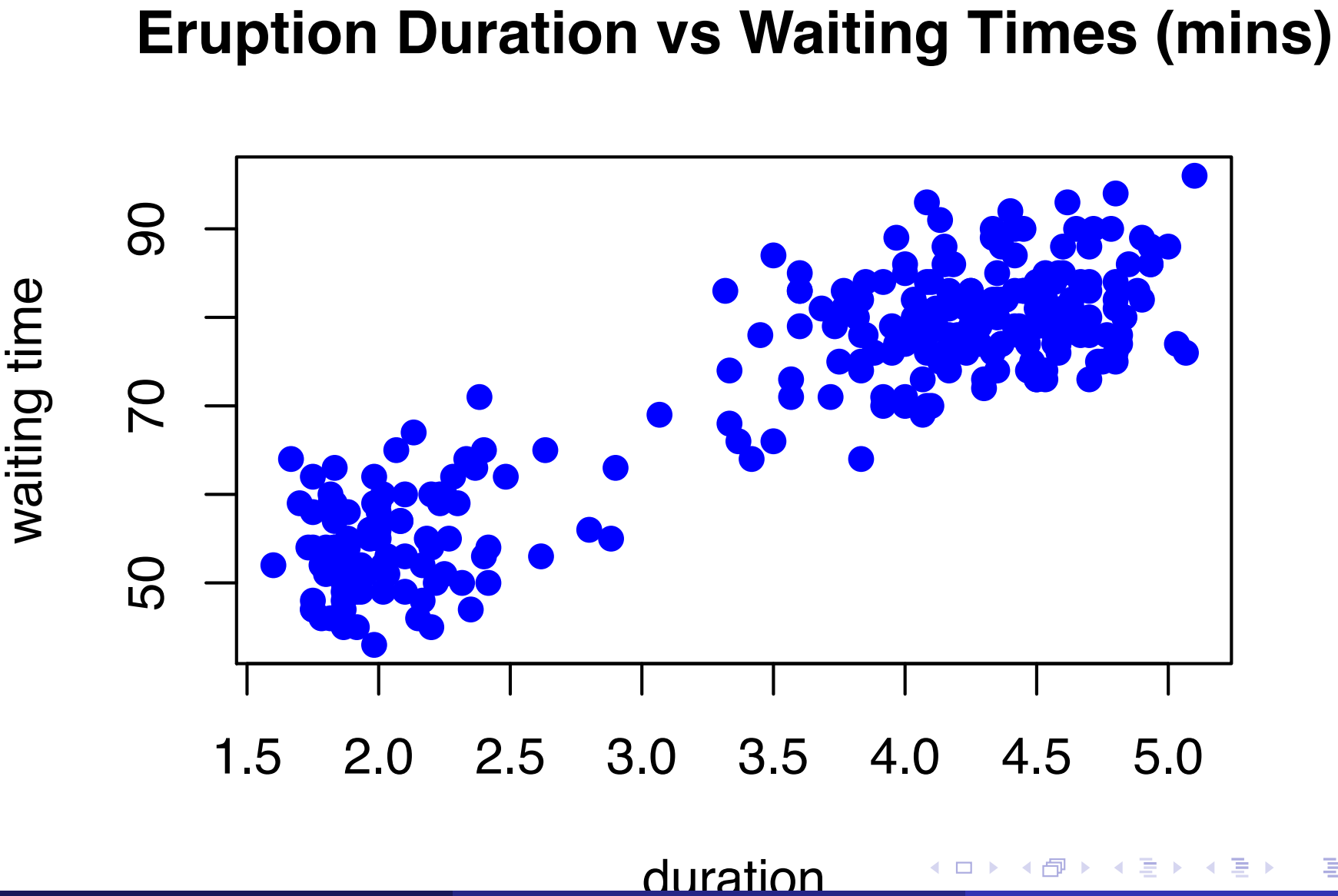
Scatterplot



Scatterplot

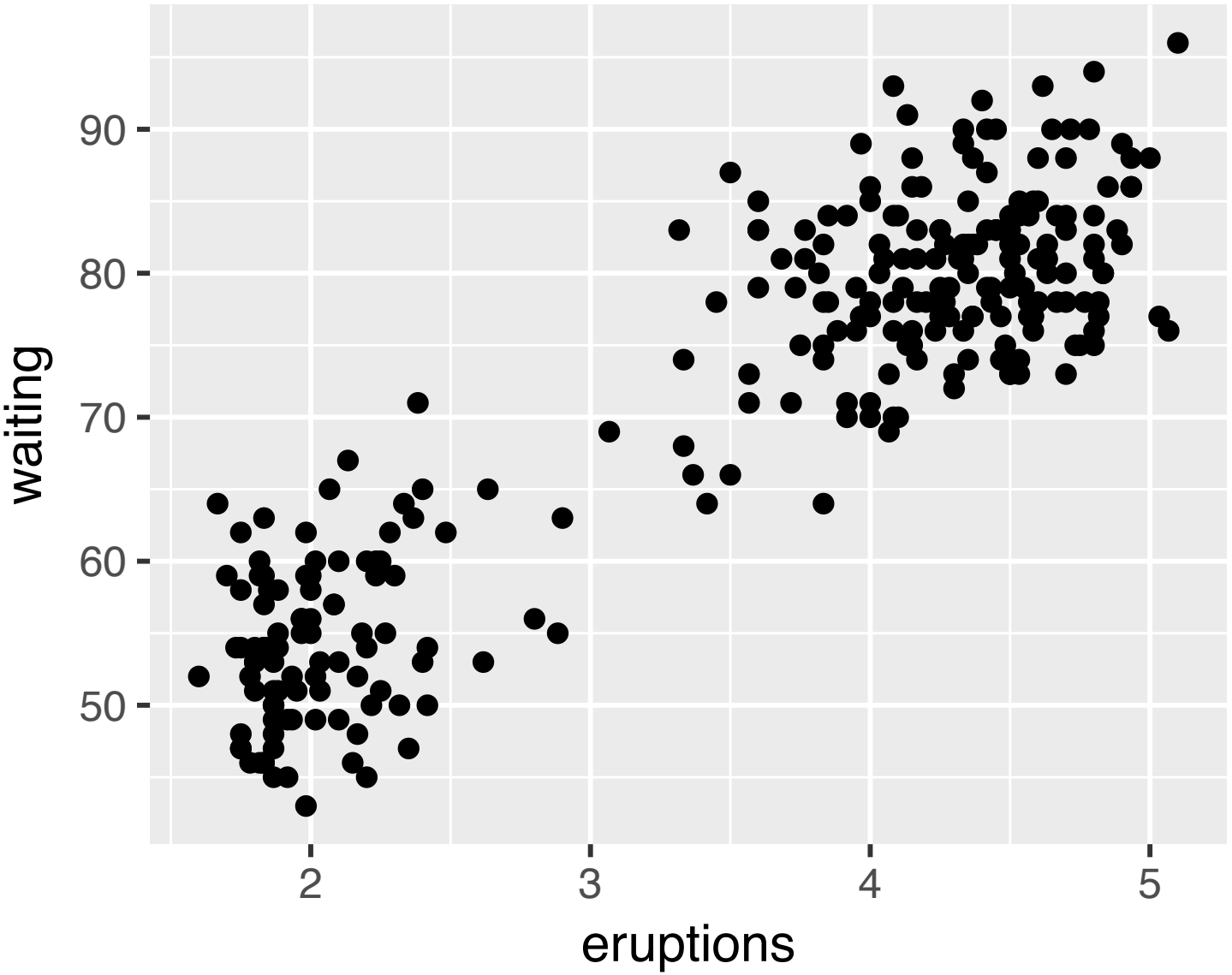
```
## Nicer plot.

plot(faithful$eruptions,faithful$waiting,
main="Eruption Duration vs Waiting Times (mins)",
xlab="duration",ylab="waiting time",
pch=19, col="blue");
```

Scatterplot



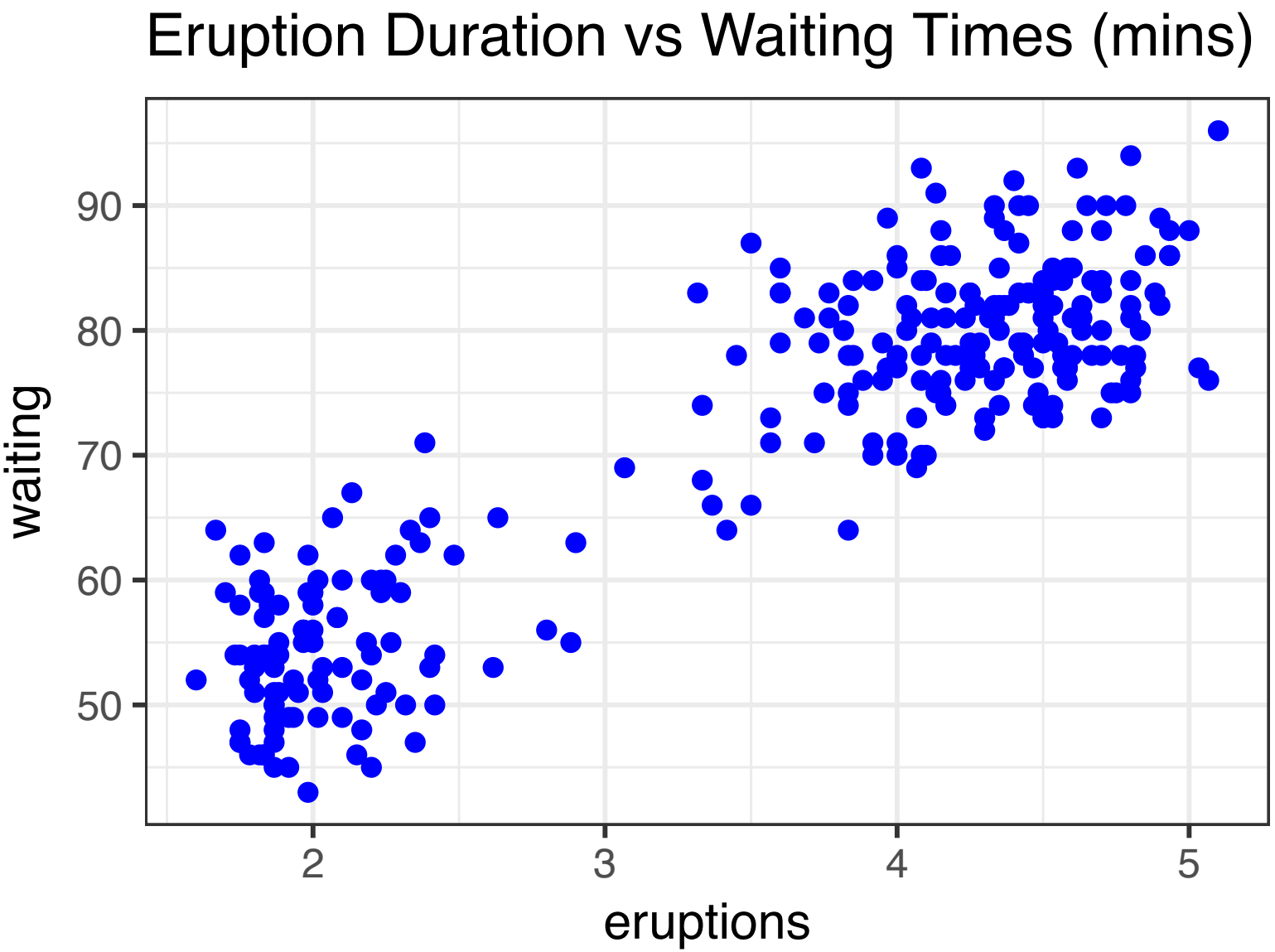
Scatterplot (again)

```
# basic;
ggplot(data=faithful,mapping=aes(x=eruptions,y=waiting))+
geom_point()
```


waiting
90
80
70
60
50
2
3
4
5
eruptions

Scatterplot (again)

```
# with titles;
ggplot(data=faithful,mapping=aes(x=eruptions,y=waiting))+
geom_point(color="blue")+
labs(title="Eruption Duration vs Waiting Times (mins)",
xlab="duration",ylab="waiting time")+
theme_bw()
```


Eruption Duration vs Waiting Times (mins)
waiting
90
80
70
60
50
2
3
4
5
eruptions

Making a panel of graphs (base R)

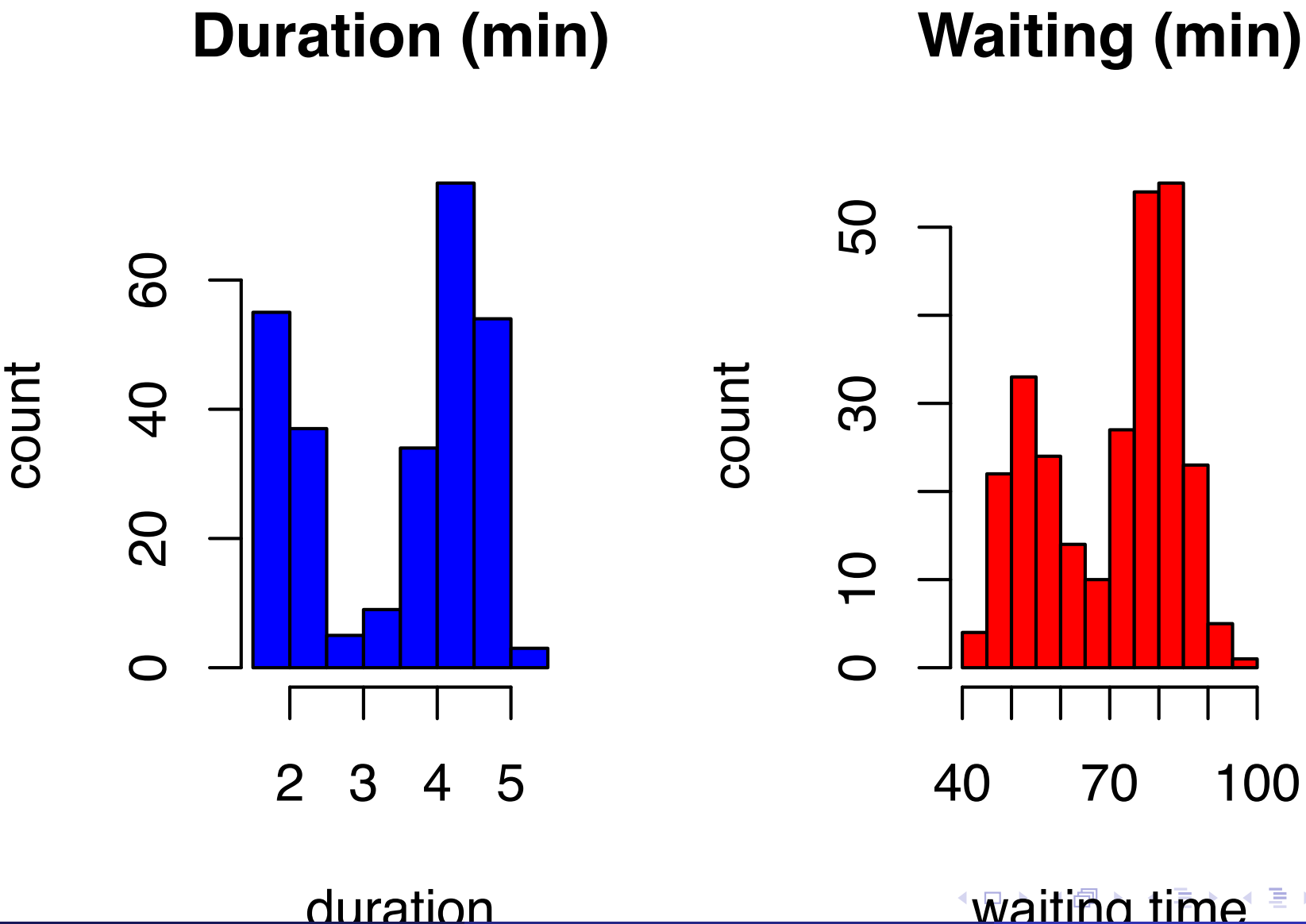
If you want more than one graph in a panel.

```
par(mfrow=c(nrow,ncol ) )

# where nrow= number of rows
# and ncol=number of columns;
```

Panel of graphs

```
par(mfrow=c(1,2) )

hist(faithful$eruptions,
main="Duration (min)",
xlab="duration",ylab="count", col="blue");

hist(faithful$waiting,
main="Waiting (min)",
xlab="waiting time",ylab="count", col="red");
```

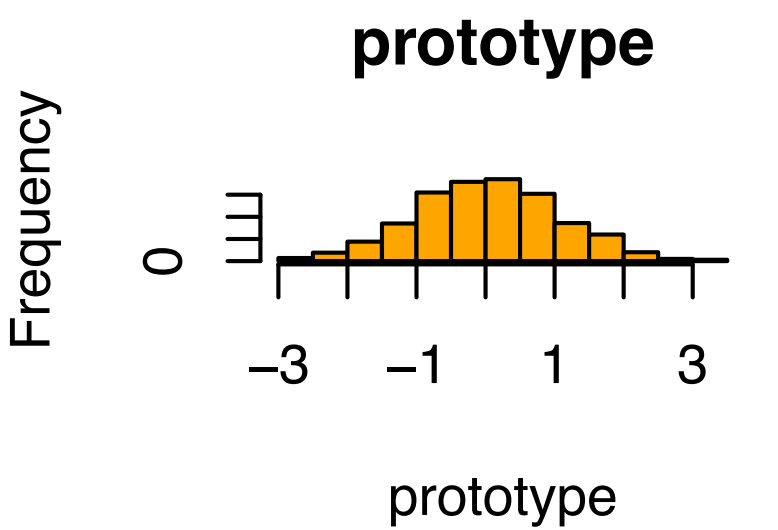
Panel of graphs

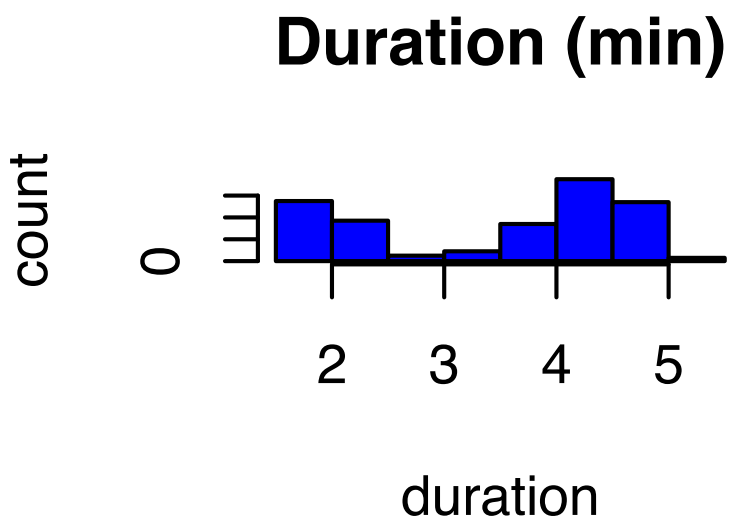


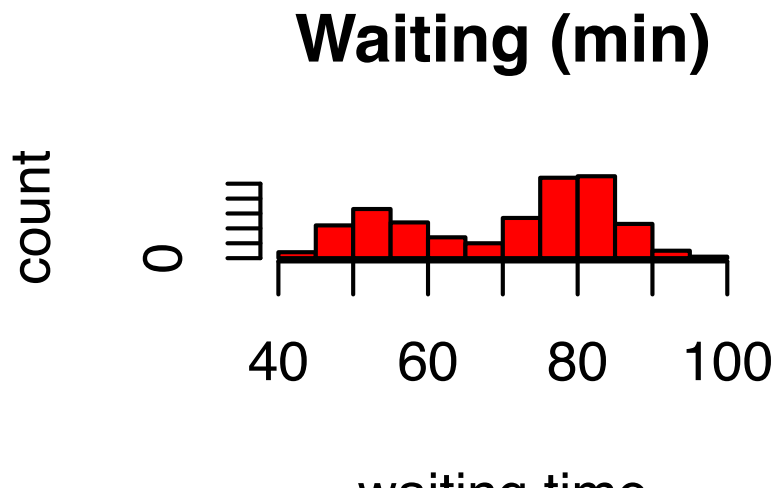
Panel of graphs

```
prototype<-rnorm(1000,mean=0,sd=1);

par(mfrow=c(2,2) )

hist(prototype,
main="prototype",
col="orange");

hist(faithful$eruptions,
main="Duration (min)",
xlab="duration",ylab="count", col="blue");

hist(faithful$waiting,
main="Waiting (min)",
xlab="waiting time",ylab="count", col="red");
```

Panel of graphs







Problem

How much do people with a bachelor's degree (but no higher degree) earn? Here are the incomes of 15 such people, chosen at random by the Census Bureau in March 2002 and asked how much they earned in 2001. Most people reported their incomes to the nearest thousand dollars, so we have rounded their responses to thousands of dollars: 110 25 50 50 55 30 35 30 4 32 50 30 32 74 60.
How could we find the "typical" income for people with a bachelor's degree (but no higher degree)?

Measuring center: the median

The **median** M is the **midpoint** of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of the distribution:
Arrange all observations in order of size, from smallest to largest.
If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $\frac{n+1}{2}$ observations up from the bottom of the list.
If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. Find the location of the median by counting $\frac{n+1}{2}$ observations up from the bottom of the list.

Income Problem (Median)

We know that if we want to find the median, M, we have to order our observations from smallest to largest: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110. Let's find the location of M

location of $M = \frac{n+1}{2} = \frac{15+1}{2} = 8$

Therefore, $M = x_8 = 35$ (x_8= 8th observation on our ordered list).

The quartiles Q_1 and Q_3

To calculate the quartiles:
Arrange the observations in increasing order and locate the median M in the ordered list of observations.
The first quartile Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
The third quartile Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Income Problem (Q_1)

Data:
4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.
From previous work, we know that $M = x_8 = 35$.
This implies that the first half of our data has $n_1 = 7$ observations. Let us find the location of Q_1:
location of $Q_1 = \frac{n_1+1}{2} = \frac{7+1}{2} = 4$.
This means that $Q_1 = x_4 = 30$.

Income Problem (Q_3)

Data:
4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.
From previous work, we know that $M = x_8 = 35$.
This implies that the first half of our data has $n_2 = 7$ observations. Let us find the location of Q_3:
location of $Q_3 = \frac{n_2+1}{2} = \frac{7+1}{2} = 4$.
This means that $Q_3 = 55$.

Five-number summary

The five-number summary of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is
$min\ Q_1\ M\ Q_3\ MAX$.

Income Problem (five-number summary)

Data: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110. The five-number summary for our income problem is given by:
4 30 35 55 110

R Code

```
# Step 1. Entering Data;

income=c(4,25,30,30,30,32,32,35,50,50,50,55,60,74,110);
```

R Code

```
# Step 2. Finding five-number summary;

fivenum(income);
```

R Code

```
## [1]   4.0  30.0  35.0  52.5 110.0
```

Note. Sometimes, R will give you a slightly different five-number summary.

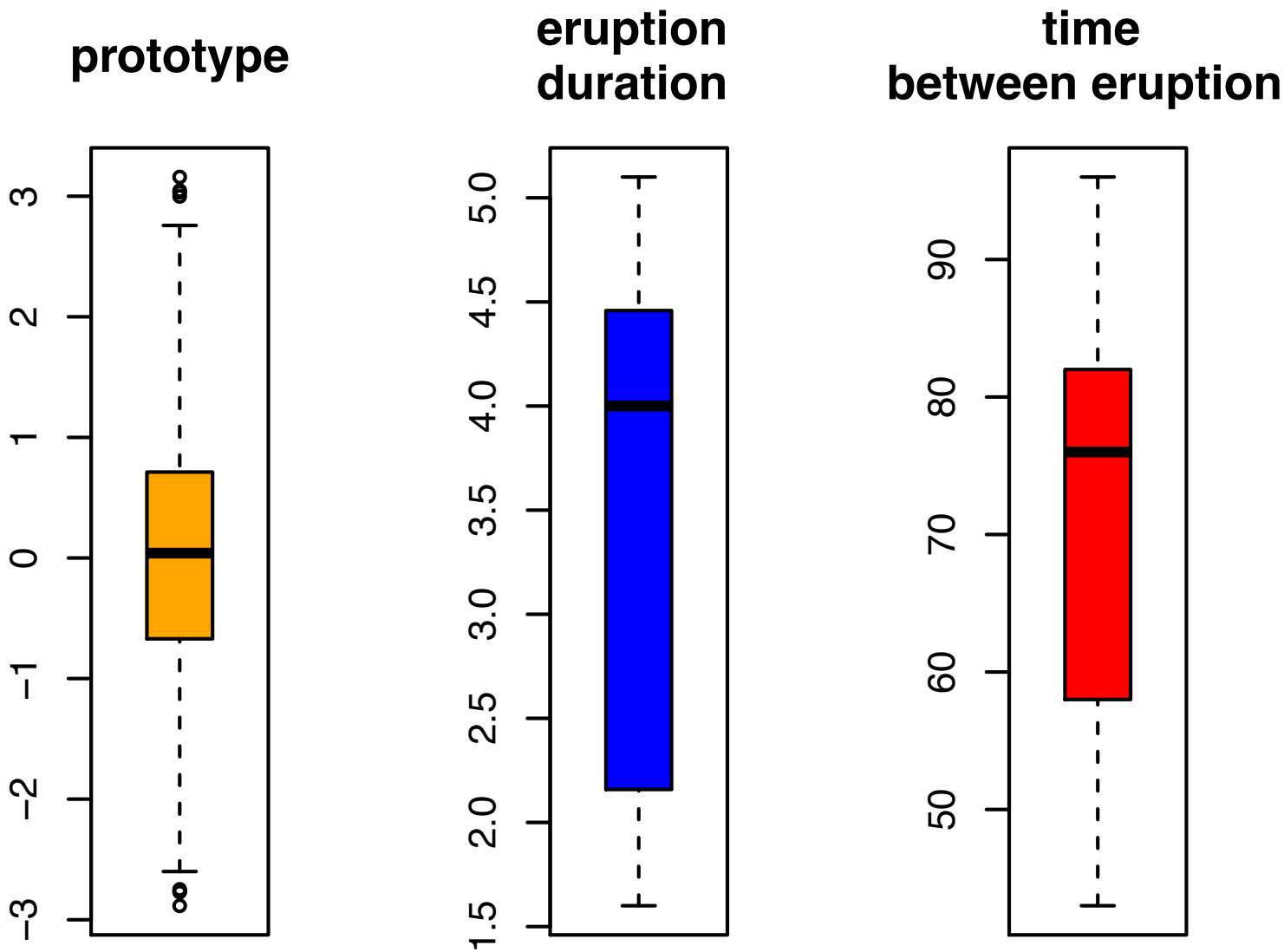
Box plot

A boxplot is a graph of the five-number summary.
A central box spans the quartiles Q_1 and Q_3.
A line in the box marks the median M.
Lines extended from the box out to the smallest and largest observations.

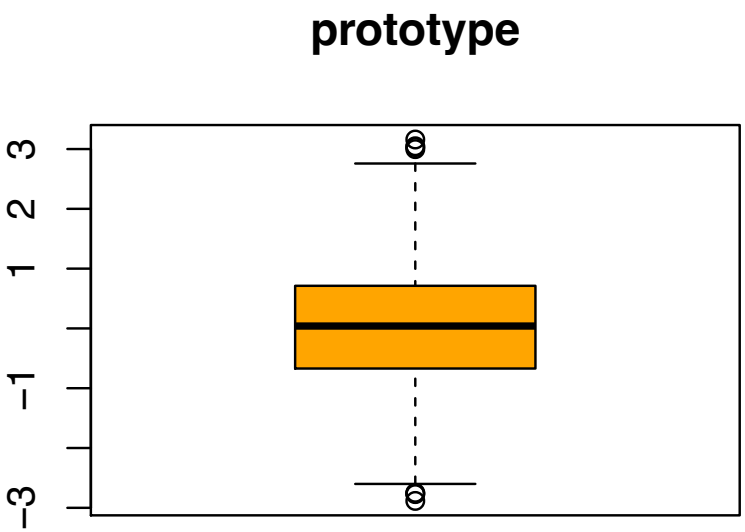
Boxplot

```
par(mfrow=c(1,3) )

boxplot(prototype,
main="prototype",
col="orange");

boxplot(faithful$eruptions,
main="eruption duration ",
col="blue");

boxplot(faithful$waiting,
main="time between eruptions",
 col="red");
```

Boxplot


prototype
eruption duration
time between eruption
−3
−2
−1
0
1
2
3
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
50
60
70
80
90

Boxplot



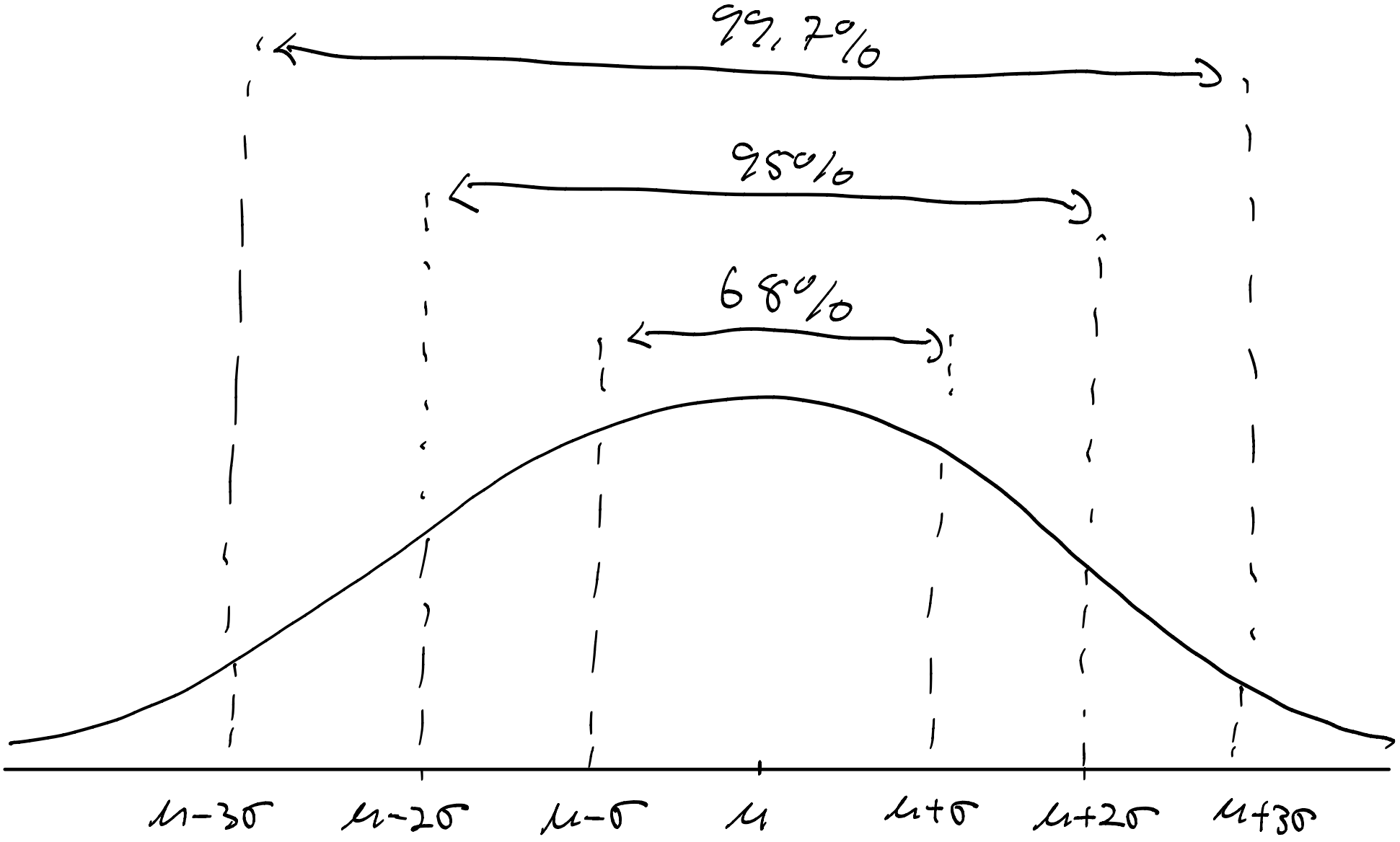
Should be:

- Symmetric
- $\approx$ 7 out of 1000 outliers
- Tails $\approx$ 1.5 IQR

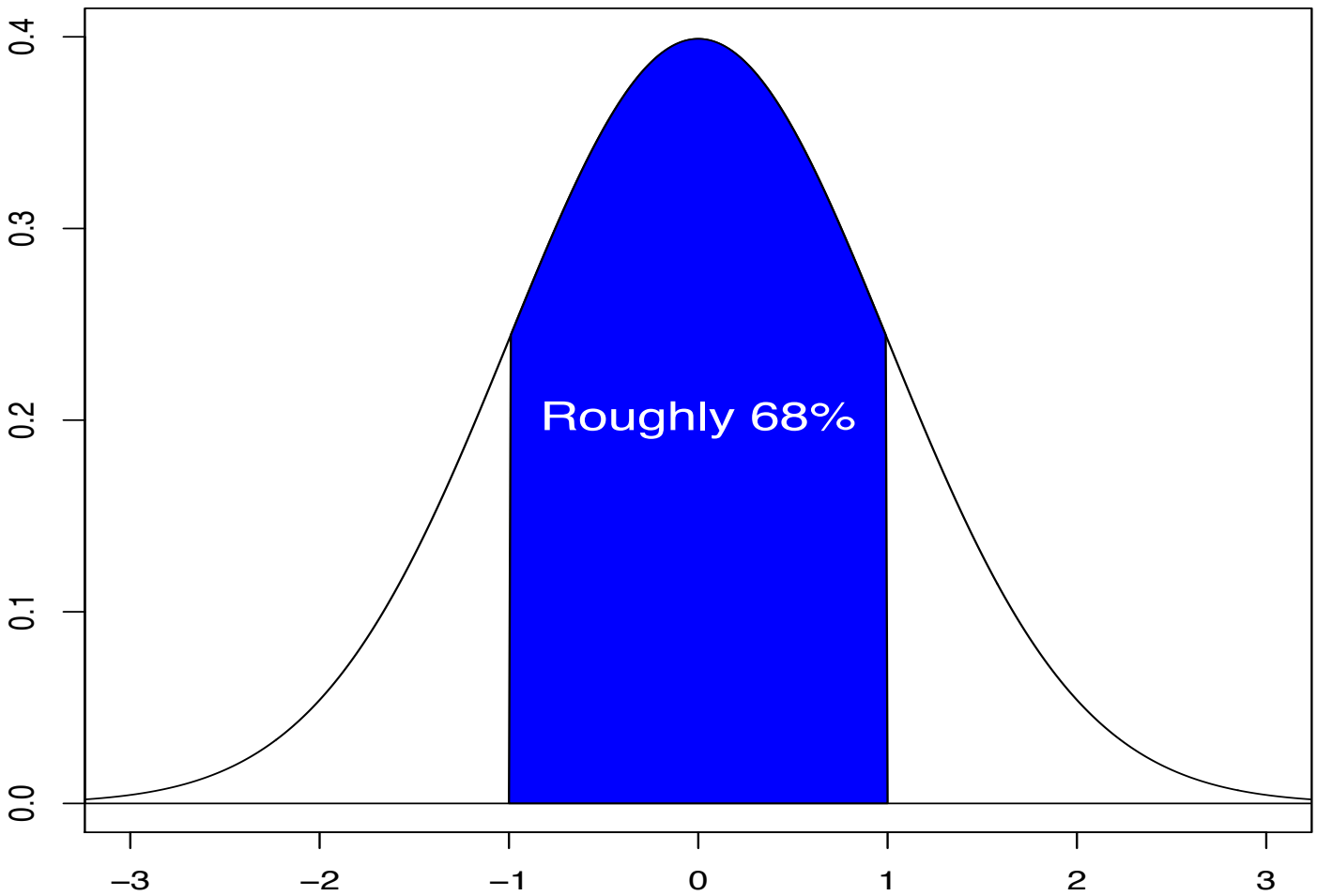
The 68-95-99.7 rule

The Empirical Rule works for approx bell-shaped distributions

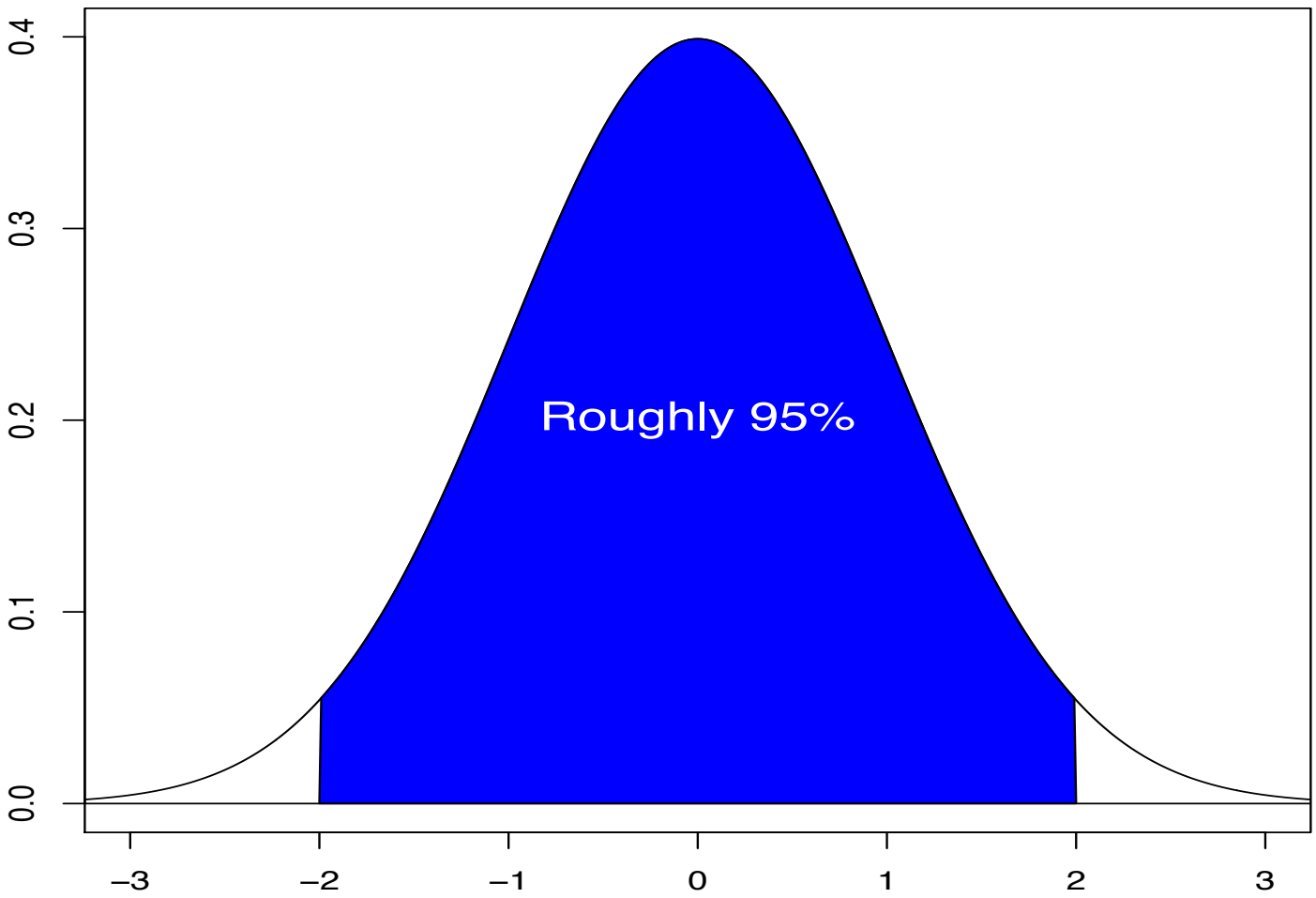
In the Normal distribution with mean μ and standard deviation σ:
Approximately 68% of the observations fall within σ of the mean μ.
Approximately 95% of the observations fall within 2σ of μ.
Approximately 99.7% of the observations fall within 3σ of μ.
Note. The 68-95-99.7 rule is also know as the empirical rule.


99,7%
95%
68%
μ-3σ
μ-2σ
μ-σ
μ
μ+σ
μ+2σ
μ+3σ

Example $N(\mu = 0, \sigma = 1)$



Example $N(\mu = 0, \sigma = 1)$



R Code

```
meanP<-mean(prototype);

sdP<-sqrt(var(prototype));

lower<-meanP-sdP;

upper<-meanP+sdP;

N<-length(prototype);

100*length(prototype[ lower<prototype & prototype<upper])/N;

## [1] 68.8
```

R Code

```
meanP<-mean(prototype);

sdP<-sqrt(var(prototype));

lower<-meanP-2*sdP;

upper<-meanP+2*sdP;

N<-length(prototype);

100*length(prototype[ lower<prototype & prototype<upper])/N;

## [1] 95.6
```

R Code

```
meanP<-mean(prototype);

sdP<-sqrt(var(prototype));

lower<-meanP-3*sdP;

upper<-meanP+3*sdP;

N<-length(prototype);

100*length(prototype[ lower<prototype & prototype<upper])/N;

## [1] 99.9
```

R Code

```
eruptions<-faithful$eruptions;

meanE<-mean(eruptions);

sdE<-sqrt(var(eruptions));

lower<-meanE-sdE;

upper<-meanE+sdE;

N<-length(eruptions);

100*length(eruptions[ lower<eruptions & eruptions<upper])/N;

## [1] 55.14706
```

R Code

```
eruptions<-faithful$eruptions;

meanE<-mean(eruptions);

sdE<-sqrt(var(eruptions));

lower<-meanE-2*sdE;

upper<-meanE+2*sdE;

N<-length(eruptions);

100*length(eruptions[ lower<eruptions & eruptions<upper])/N;

## [1] 100
```

R Code

```
eruptions<-faithful$eruptions;

meanE<-mean(eruptions);

sdE<-sqrt(var(eruptions));

lower<-meanE-3*sdE;

upper<-meanE+3*sdE;

N<-length(eruptions);

100*length(eruptions[ lower<eruptions & eruptions<upper])/N;

## [1] 100
```

Homework?

Modify R Code given above and see what happens with waiting time.

Quiz 1

Released Fri 17 ~ 12: noon

open 24 hr window

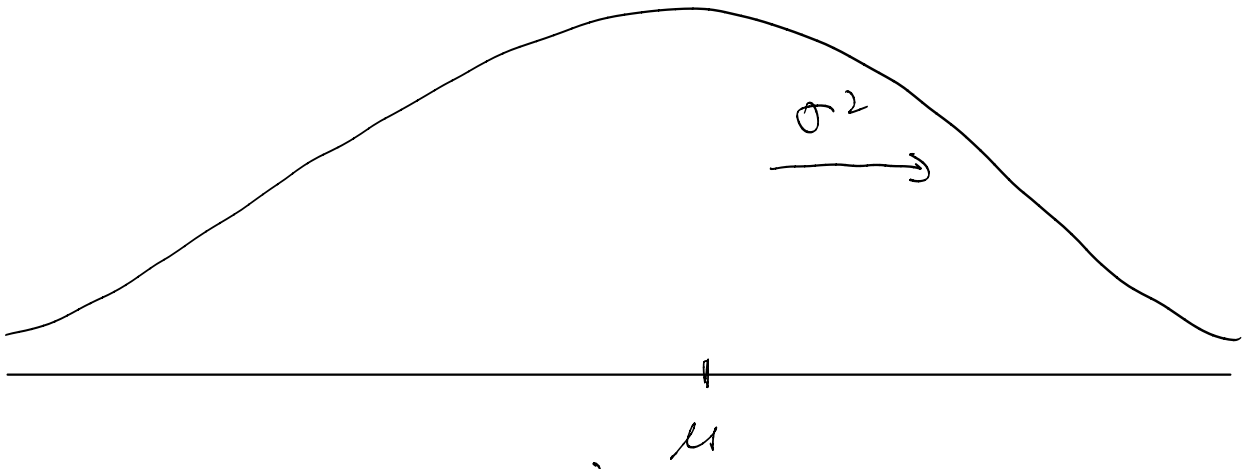
1 hr

3 attempts (highest)

Require R

Normal Distribution (Gaussian Distribution)

$X \sim N(\mu, \sigma^2)$



$$f(x) = \frac{1}{(\sqrt{2\pi})\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad \sigma^2 > 0$$

Good model of processes in our reality

Standard Normal

Special case of Normal. $\mu = 0$, $\sigma^2 = 1$

Notation: $Z \sim N(\mu = 0, \sigma^2 = 1^2)$

$Z \sim N(0, 1^2)$

Normal Quantile Quantile (QQ) Plots

Used to determine whether data may be from a population which is normal

Observed data is plotted on y-axis

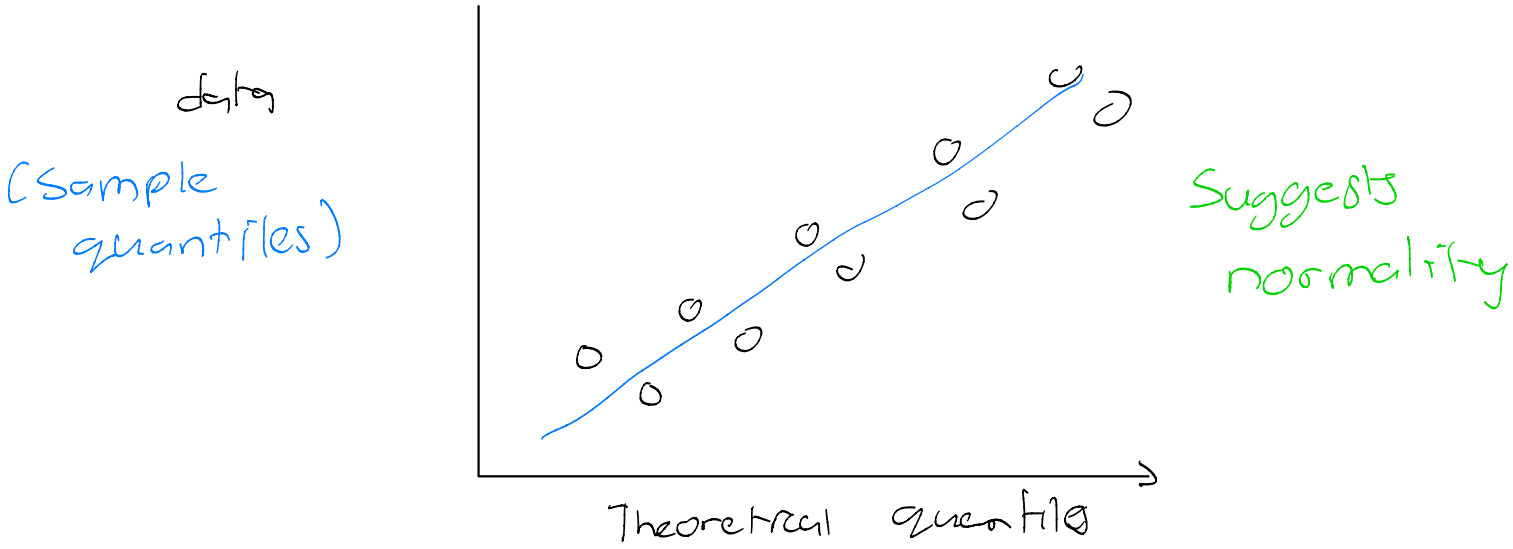
Theoretical quantiles are plotted on x-axis.

$$Z = \frac{X - \mu}{\sigma}$$

estimate with $\bar{x}$

estimate with s

If points appear to follow a straight line, suggests data is from a population which is normal



R: qqnorm(data) ①

qqline(data) ②

Q-Q Plot (Example)

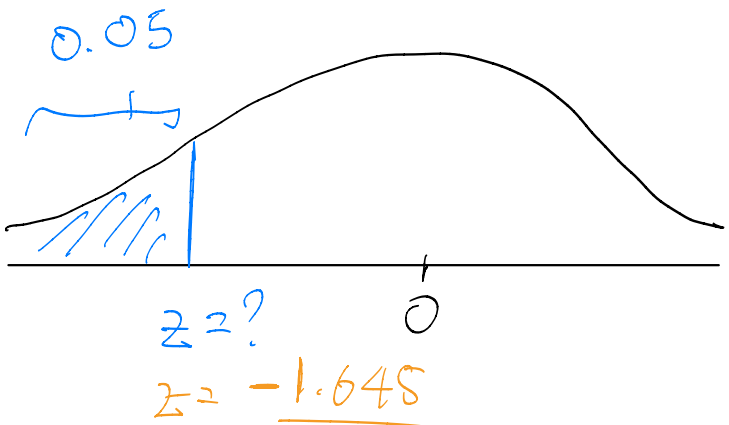
A sample of $n = 10$ observations gives the values in the following table:

	Ordered Observations $x_{(j)}$	*Probability levels* $(j - 1/2)/n$	*Standard Normal Quantiles* $q_{(j)}$
1	-1	0.05	-1.645
2	-0.10	0.15	-1.036
3	0.16	0.25	-0.674
4	0.41	0.35	-0.385
5	0.62	0.45	-0.125
6	0.80	0.55	0.125
7	1.26	0.65	0.385
8	1.54	0.75	0.674
9	1.71	0.85	1.036
10	2.30	0.95	1.645

Here, for example, $P[Z \leq 0.385] = \int_{-\infty}^{0.385} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.65$.

$j=1 \quad x_{(1)} = -1 \quad n=10 \quad (j-1/2)/n = (1-1/2)/10 = 0.05$

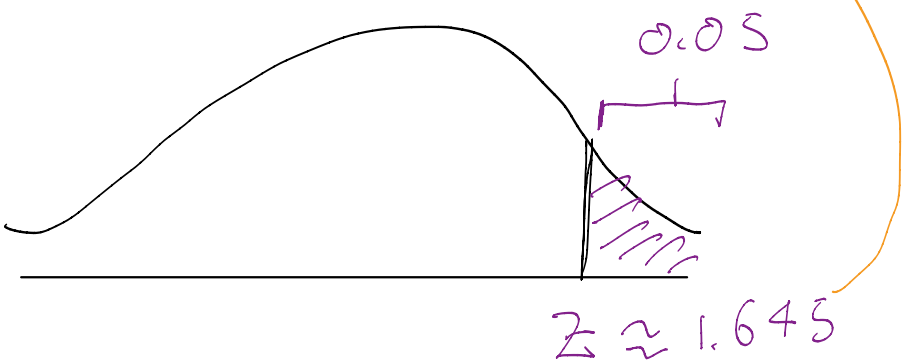
Area to left



$z = -1.645$

Use Symmetry

Table



point on qq plot $(-1.645, -1)$

Q-Q Plot (Example)

Let us now construct the Q-Q plot and comment on its appearance. The Q-Q plot for the foregoing data, which is a plot of the ordered data $x_{(j)}$ against the normal quantiles is shown below. The pairs of points $(q_{(j)}, x_{(j)})$ lie very nearly along a straight line, and we would not reject the notion that these data are Normally distributed-particularly with a sample size as small as $n = 10$.

R Code

using procedure on slide 90

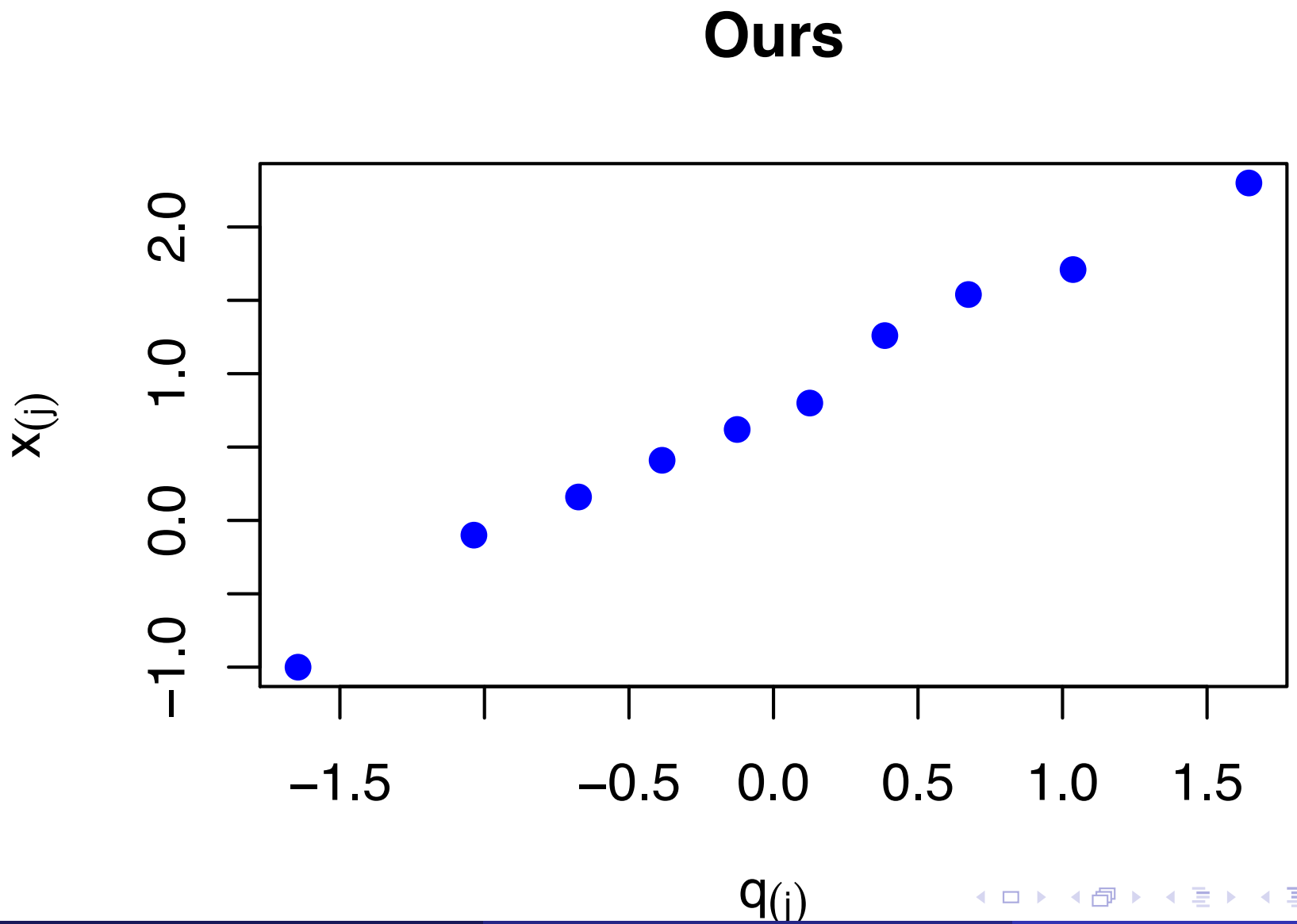
```
## Ordered observations;

obs<-c(-1,-0.1,0.16,0.41,0.62,0.80,1.26,1.54,1.71,2.30);

n<-length(obs);

## Corresponding probability values;

prob.levels<-(seq(1:n)-0.5)/n;

## Standard Normal Quantiles;

norm.quantiles<-qnorm(prob.levels);
```

2 values from standard normal

R Code

```
## Q-Q plot;

plot(norm.quantiles,obs,
xlab=expression(q[(j)]),
ylab=expression(x[(j)]),
main="Ours",col="blue",pch=19);

## Q-Q plot (using R function);

qqnorm(obs,col="blue",pch=19);
```

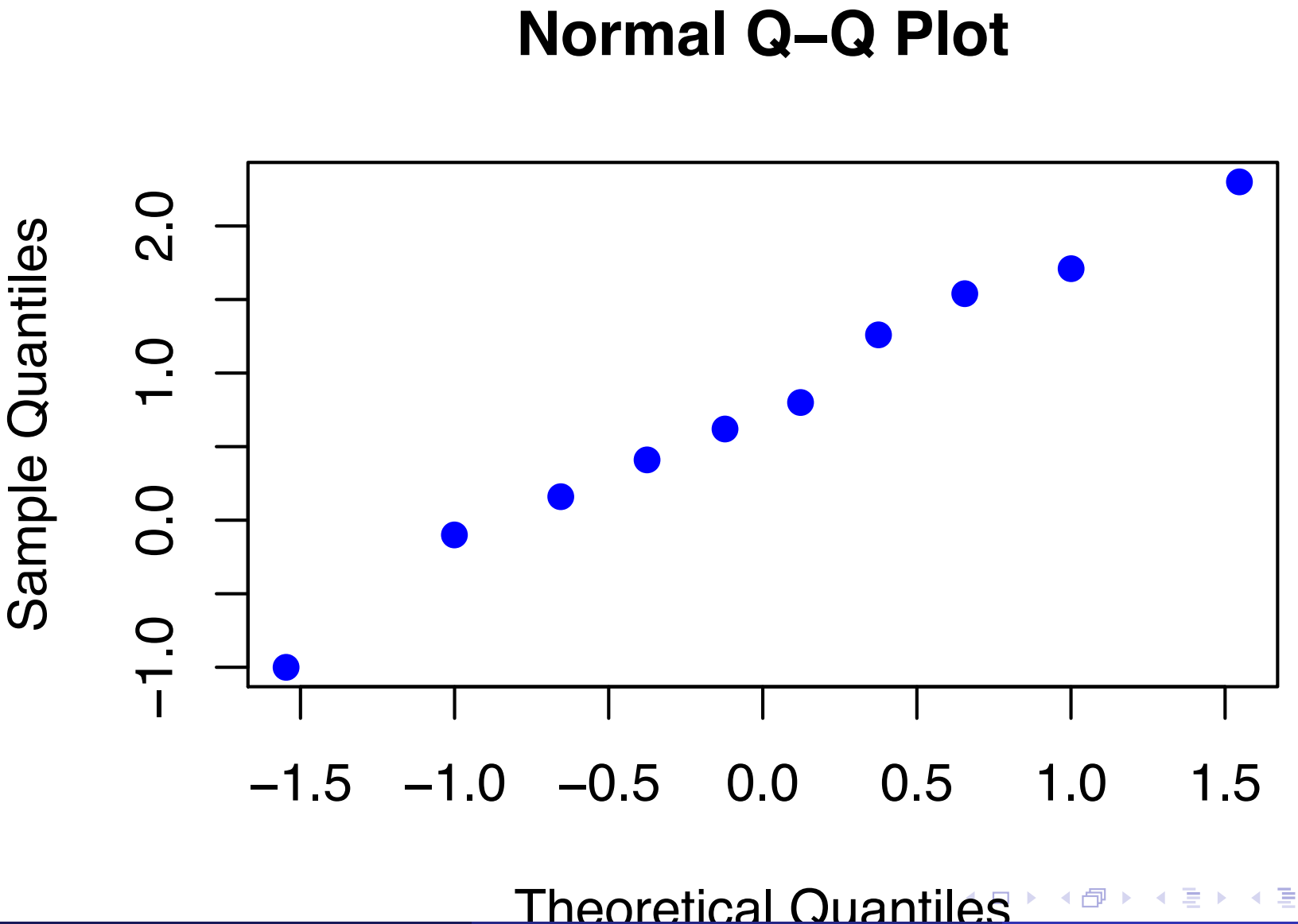
R Code



R Code (base R)

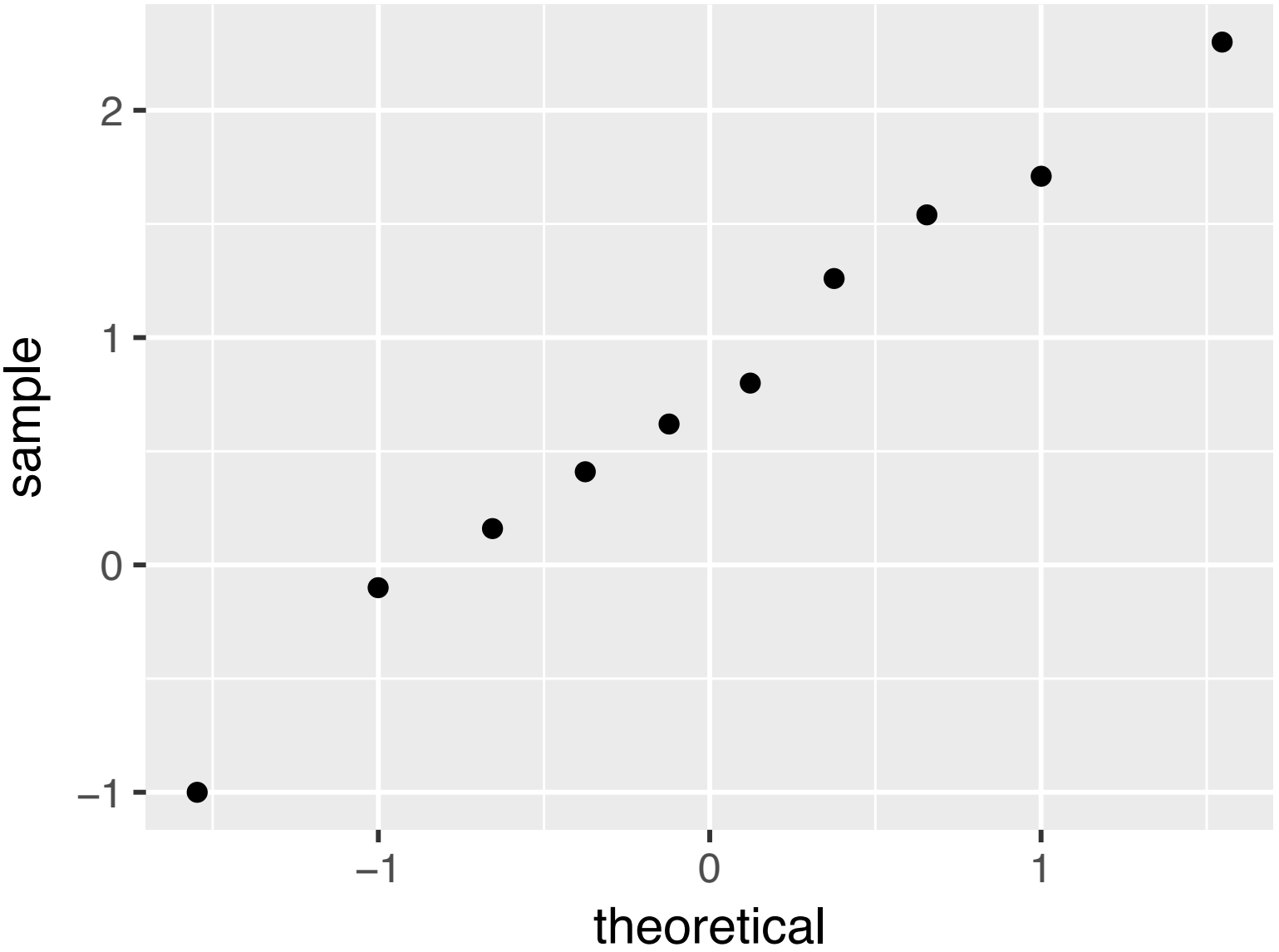
```r
## Q-Q plot (using R function);

qqnorm(obs,col="blue",pch=19);
```

R Code


Normal Q–Q Plot
Sample Quantiles
−1.0
0.0
1.0
2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles

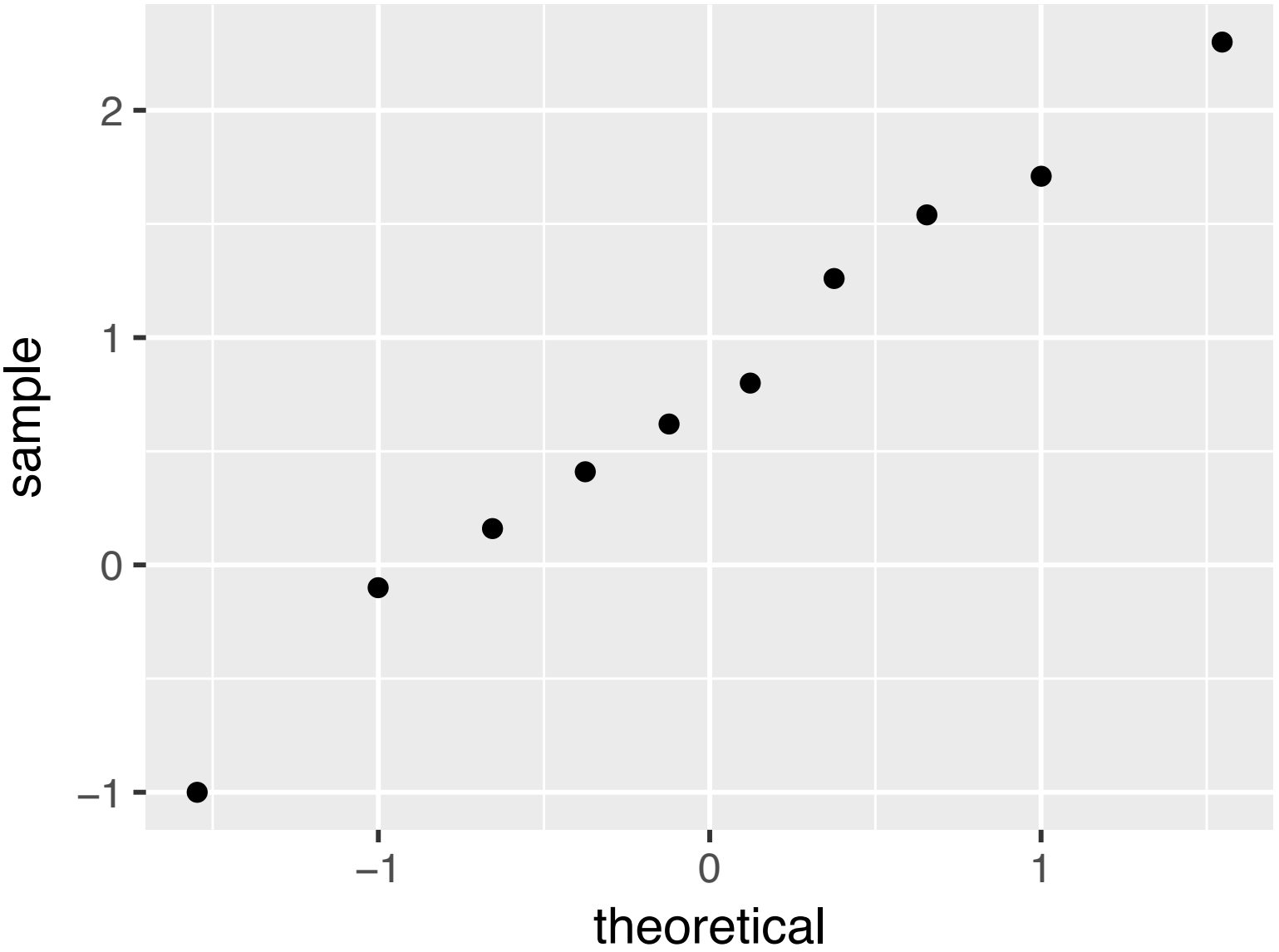
Q-Q plot (again)

```
# basic;
# one way;
obs=c(-1,-0.1,0.16,0.41,0.62,0.80,1.26,1.54,1.71,2.30);
df=data.frame(obs);
ggplot(data=df,mapping=aes(sample=obs))+
geom_qq( )
```


sample
2
1
0
−1
−1
0
1
theoretical

Q-Q plot (again)

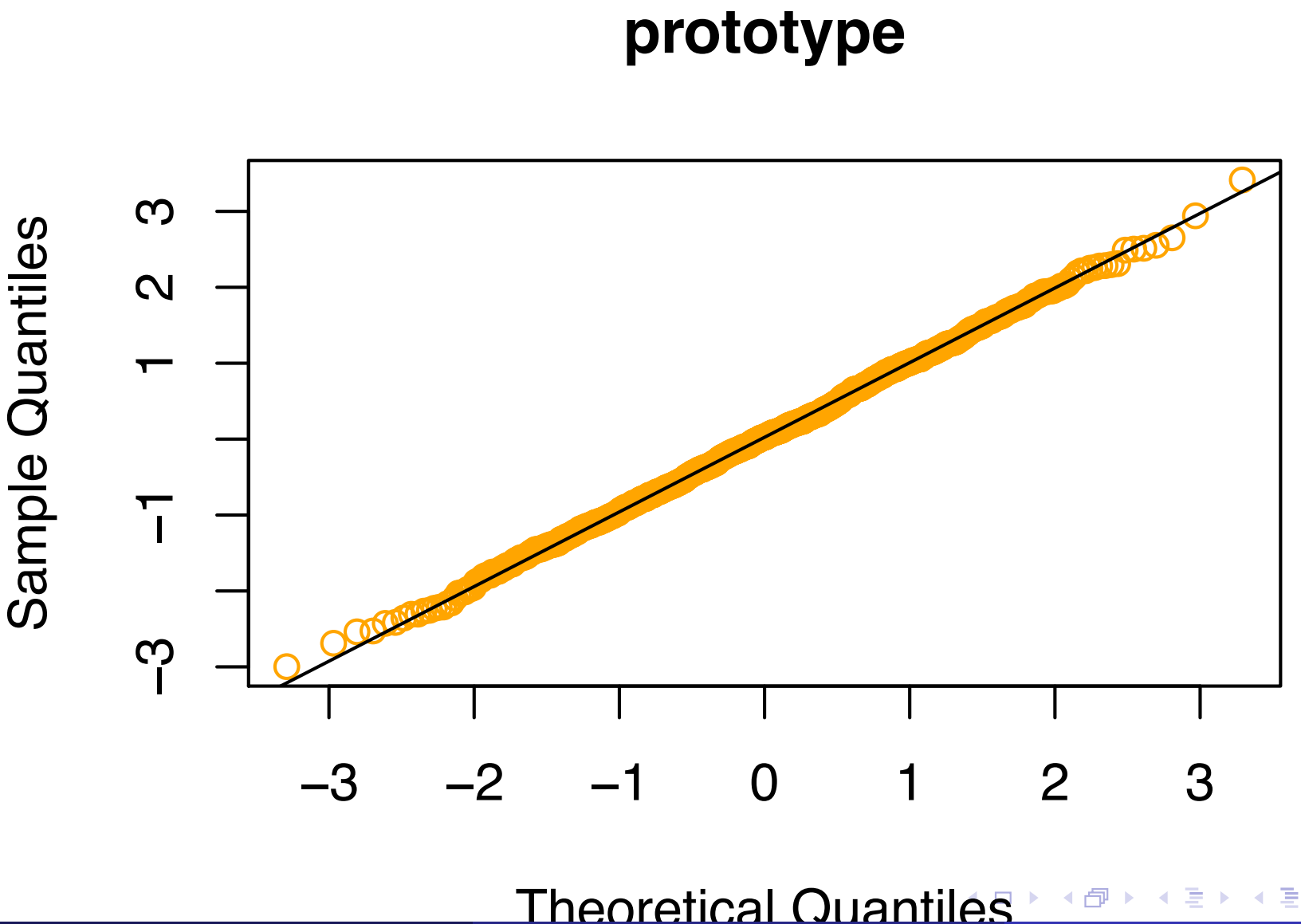
```
# basic;
# another way;
obs=c(-1,-0.1,0.16,0.41,0.62,0.80,1.26,1.54,1.71,2.30);
df=data.frame(obs);
ggplot(data=df)+
geom_qq(mapping=aes(sample=obs))
```


2
1
0
−1
sample
−1
0
1
theoretical

Q-Q plots

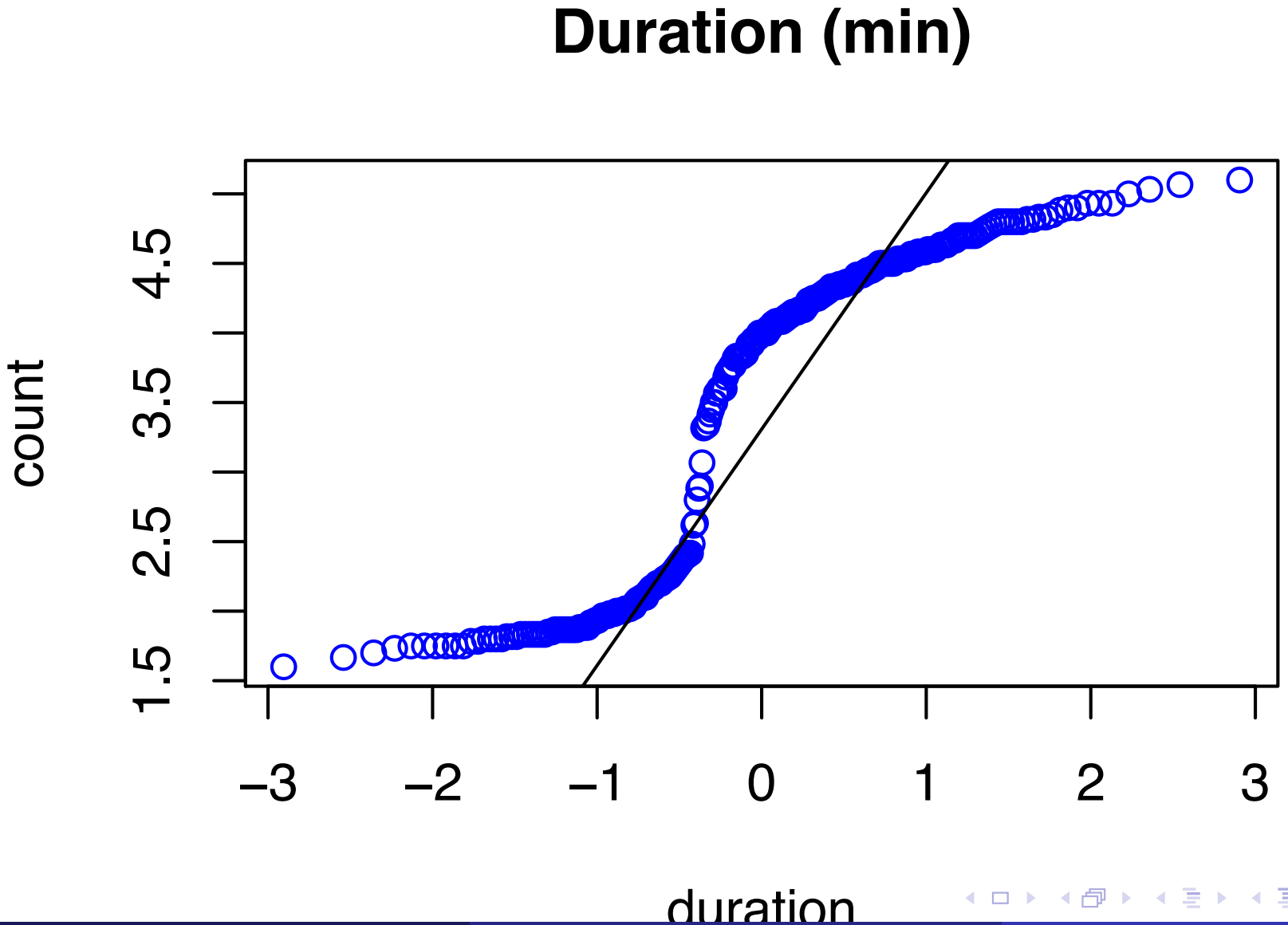
```
prototype<-rnorm(1000,mean=0,sd=1);

par(mfrow=c(1,1) )

qqnorm(prototype,
main="prototype",
col="orange");
qqline(prototype);
```

Q-Q plots

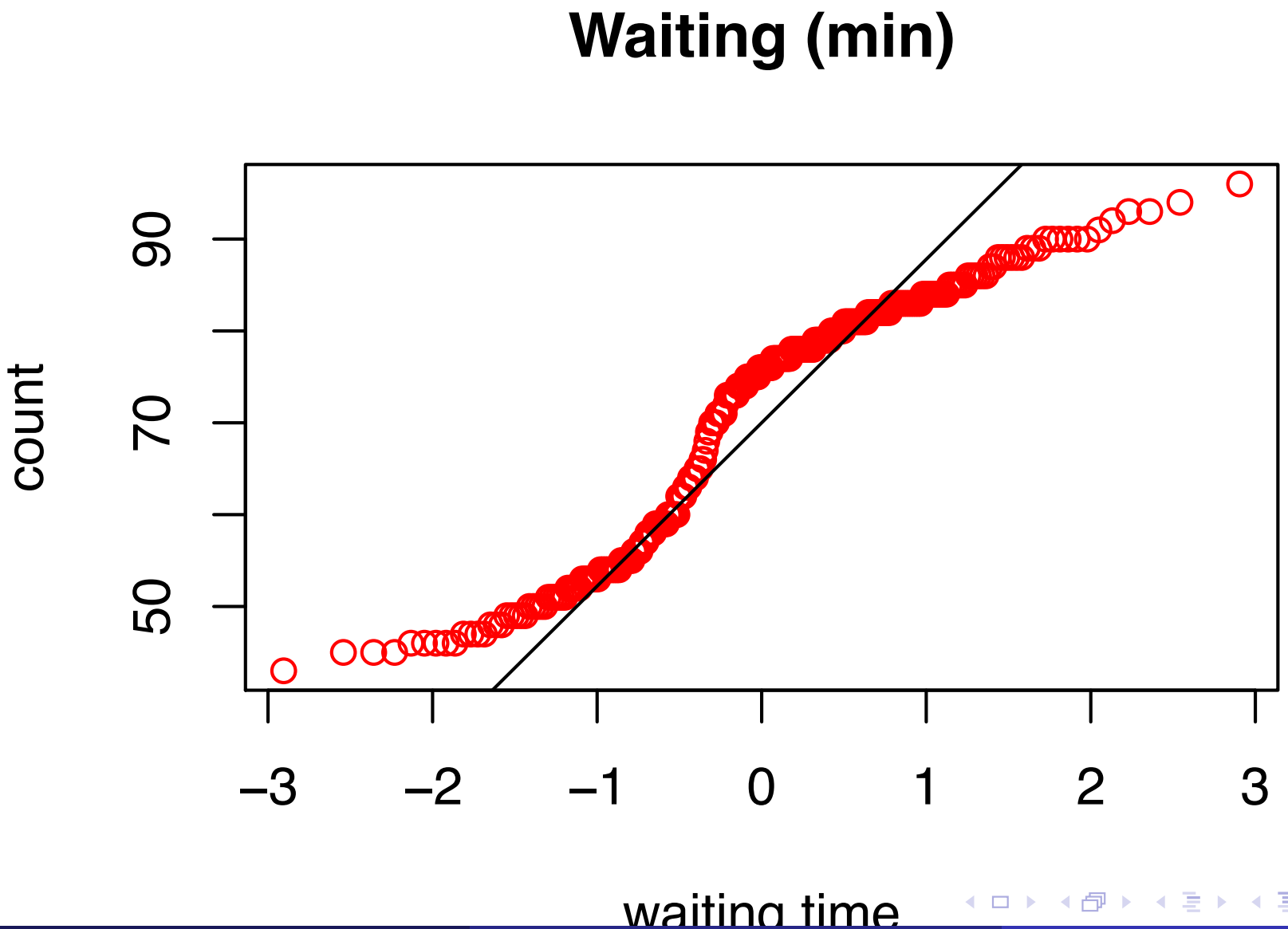

prototype
Sample Quantiles
−3
−1
1
2
3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Q-Q plots

not normal

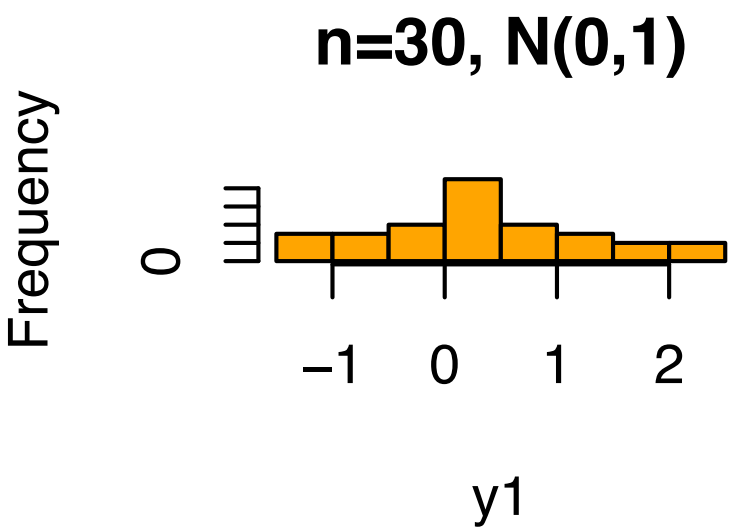


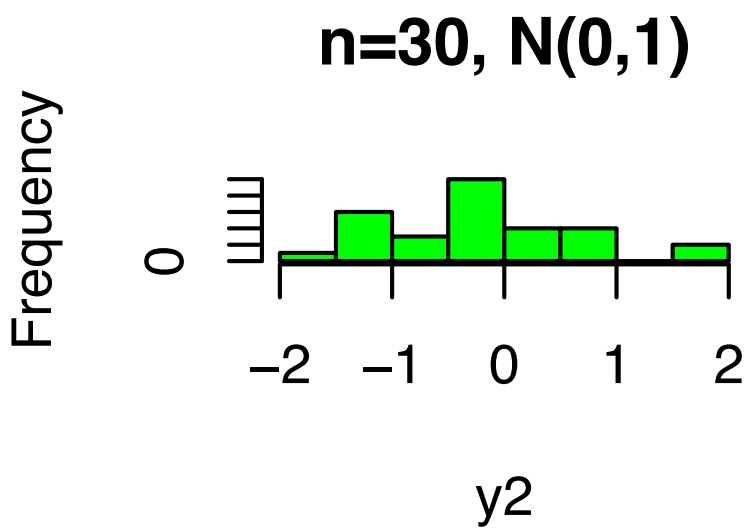
Q-Q plots

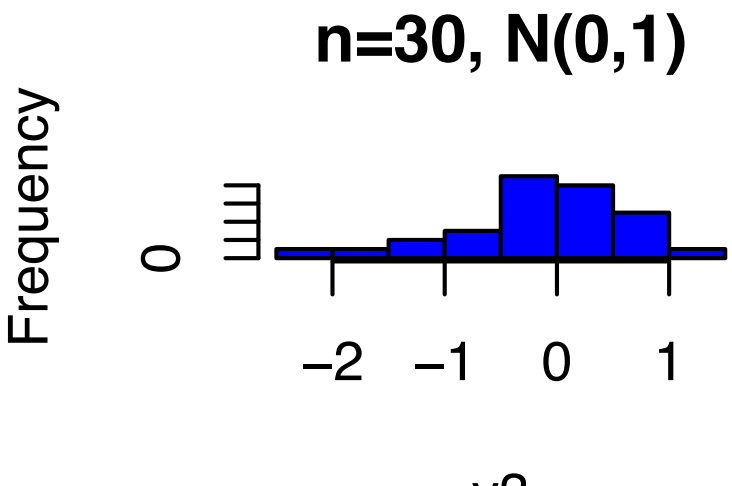


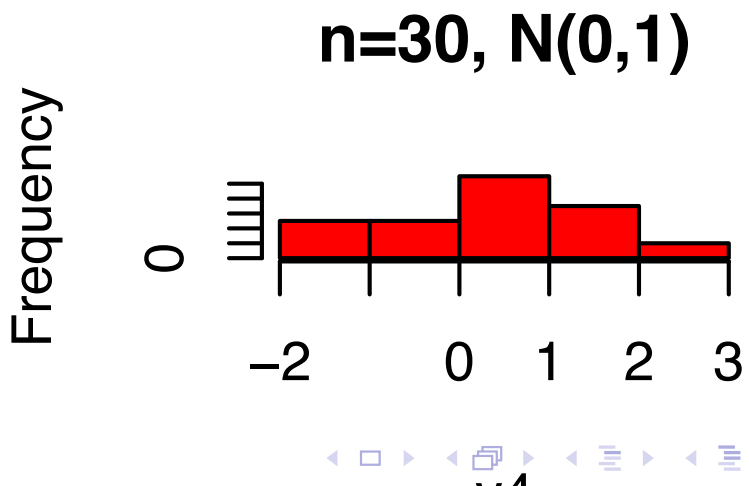
- If plots are OK $\rightarrow$ data could come from a Normal distribution . . . but it could come from some other distribution. So good plots don't prove data came from a Normal distribution
- If plots are not OK $\rightarrow$ data probably does not come from a Normal distribution, we can't assume data is from Normal population

Panel of graphs

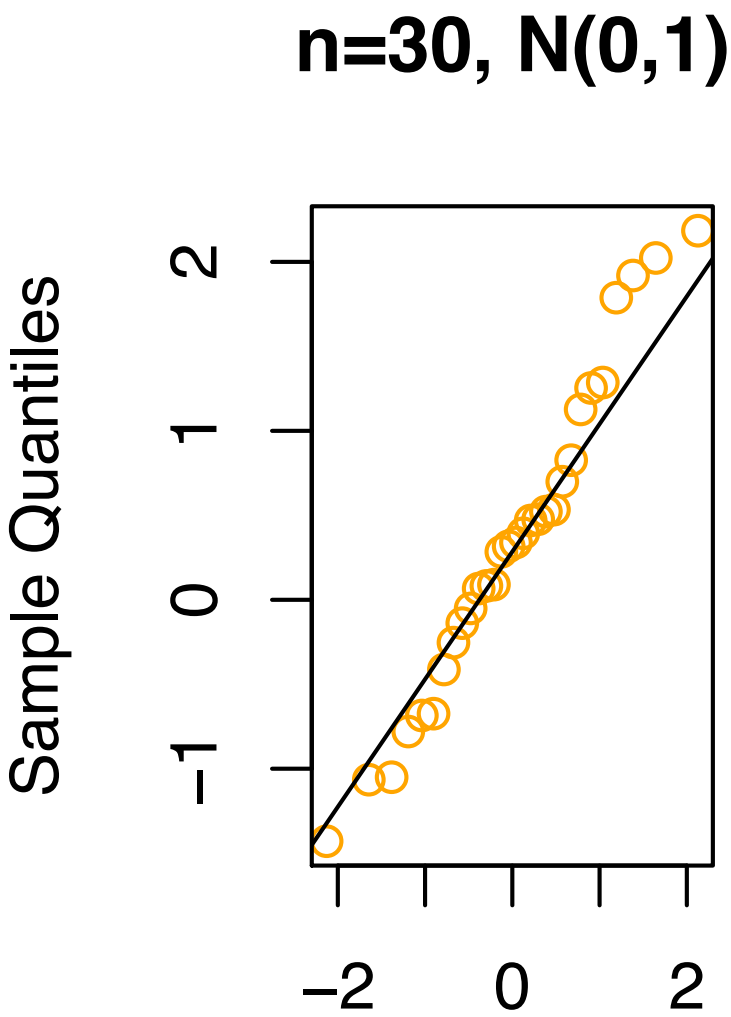


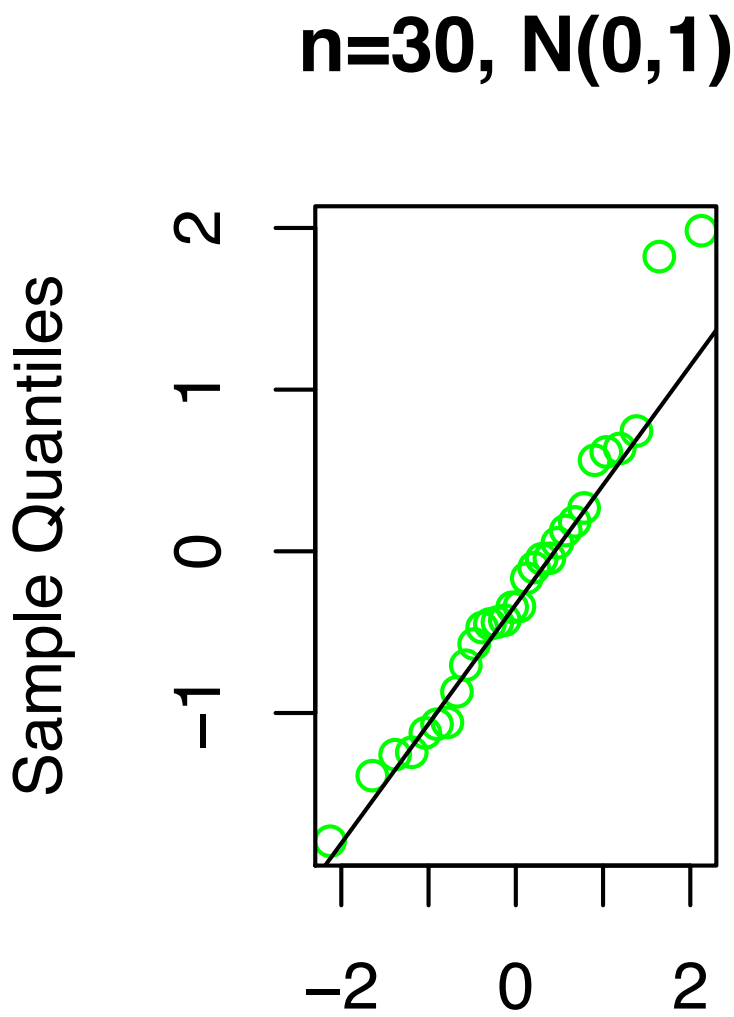




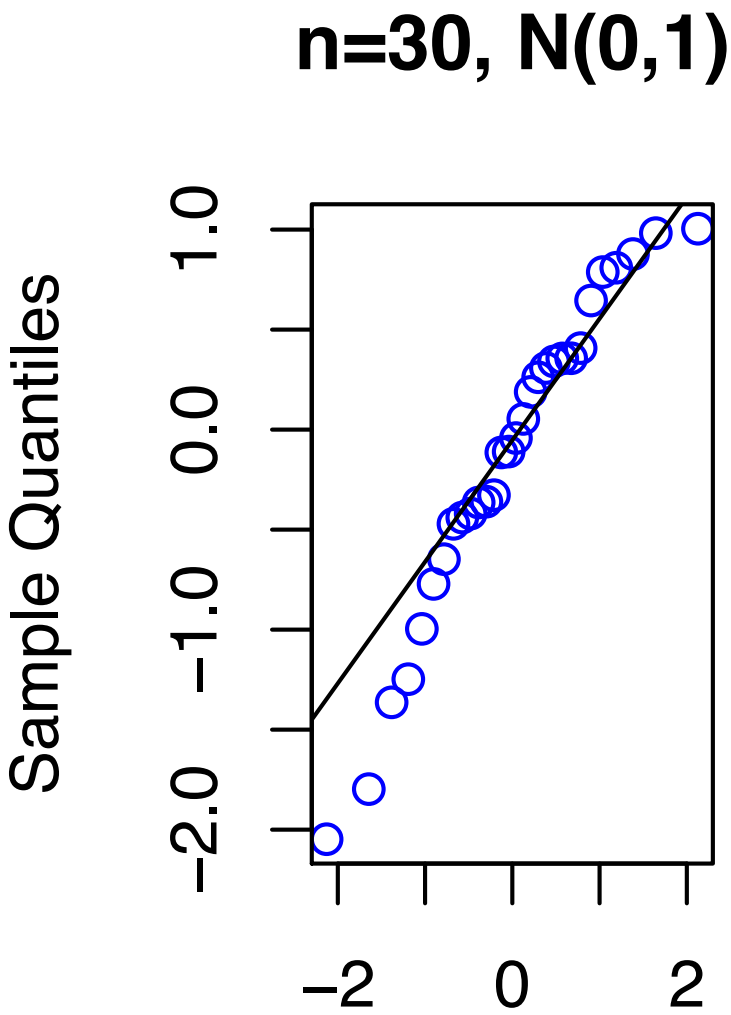


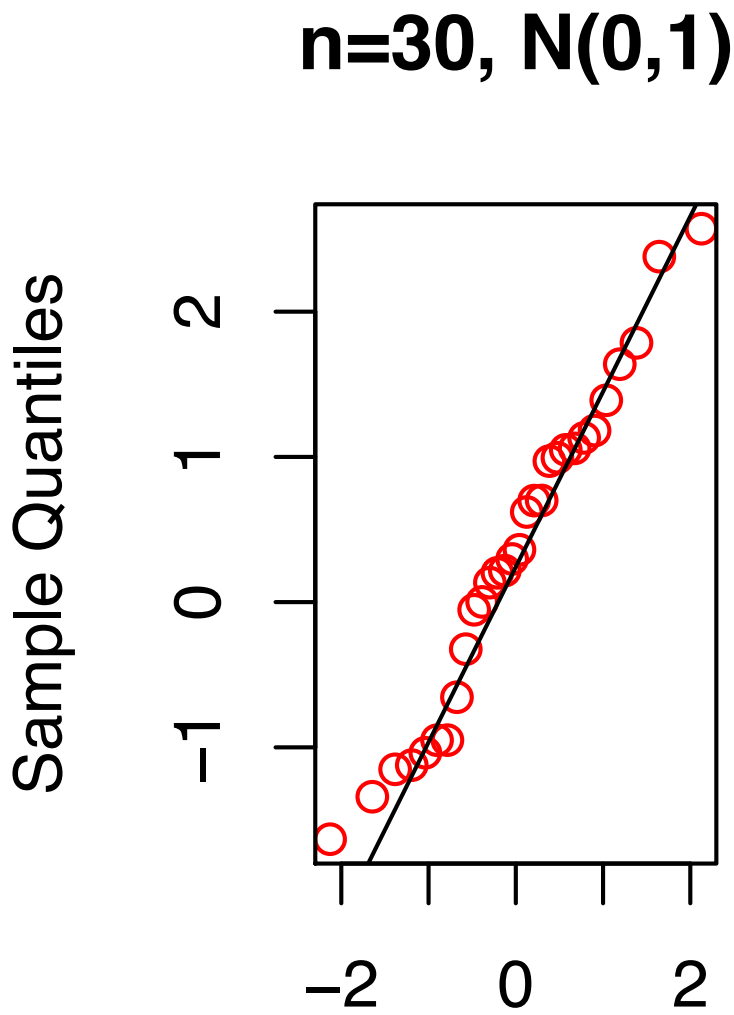
Q-Q plots


n=30, N(0,1)
Sample Quantiles
−1
0
1
2
−2
0
2
Theoretical Quantiles

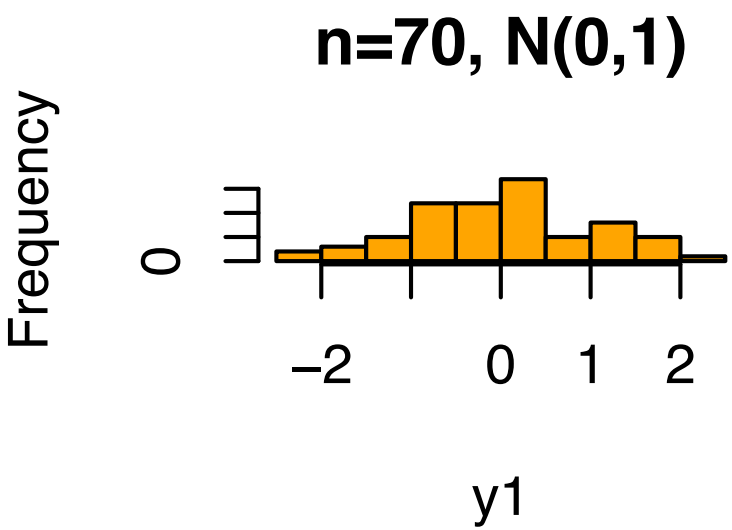

n=30, N(0,1)
Sample Quantiles
−1
0
1
2
−2
0
2
Theoretical Quantiles

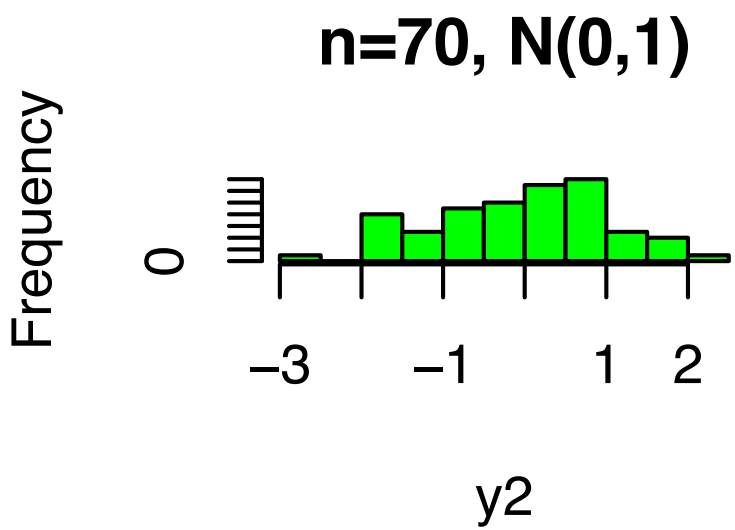
Q-Q plots

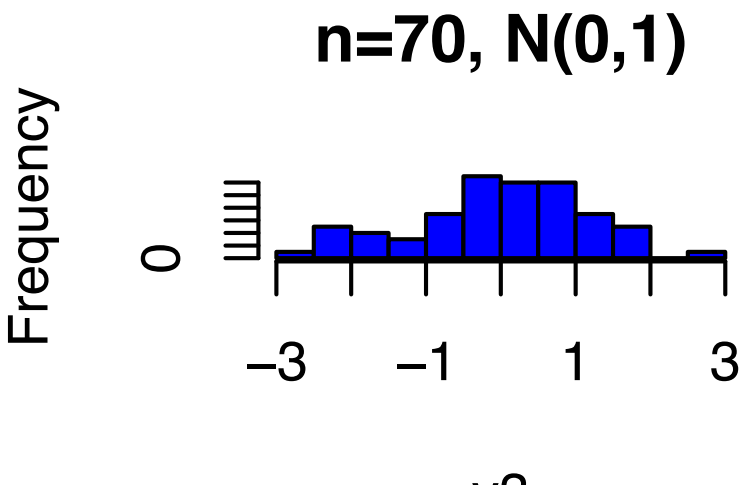

n=30, N(0,1)
Sample Quantiles
−2.0
−1.0
0.0
1.0
−2
0
2
Theoretical Quantiles

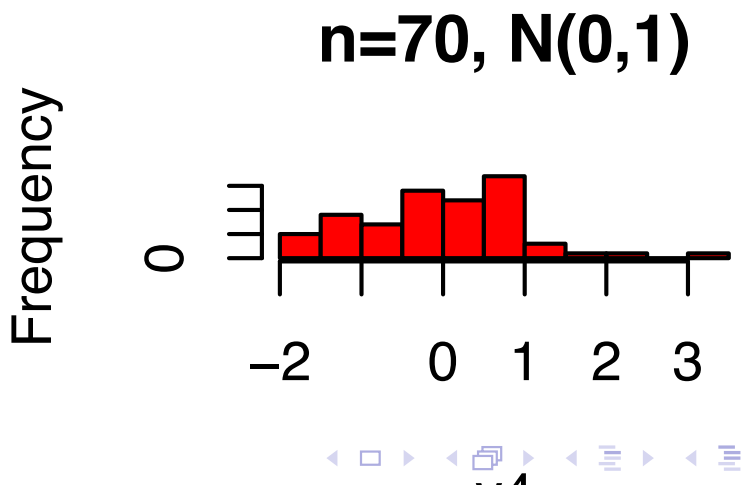

n=30, N(0,1)
Sample Quantiles
−1
0
1
2
−2
0
2
Theoretical Quantiles

Panel of graphs

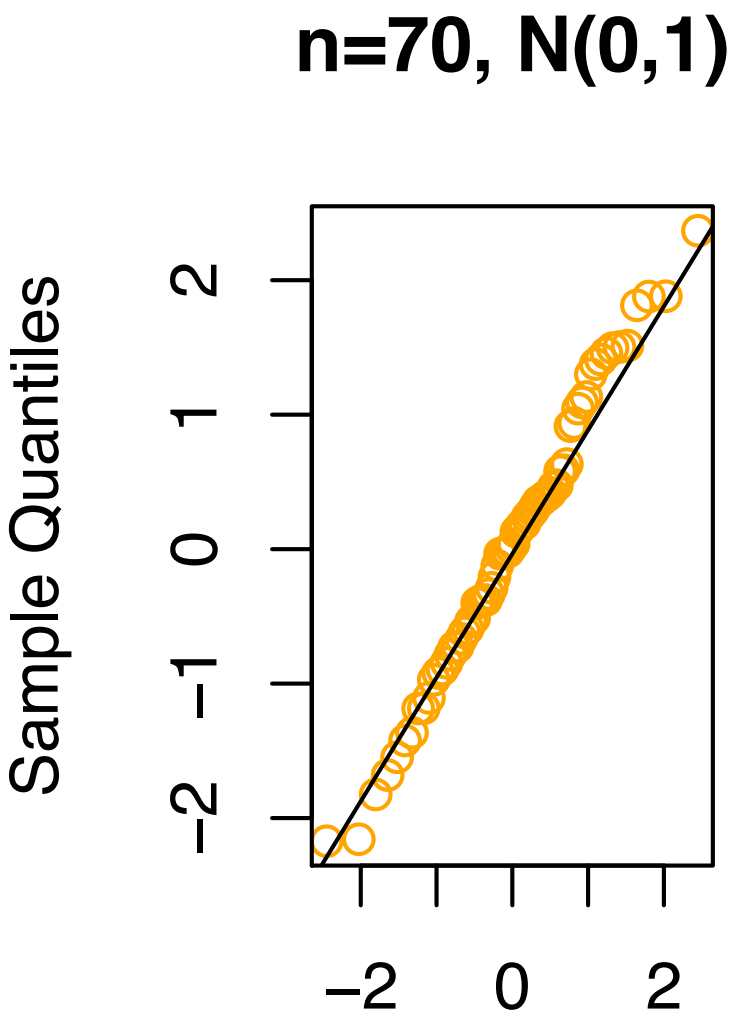


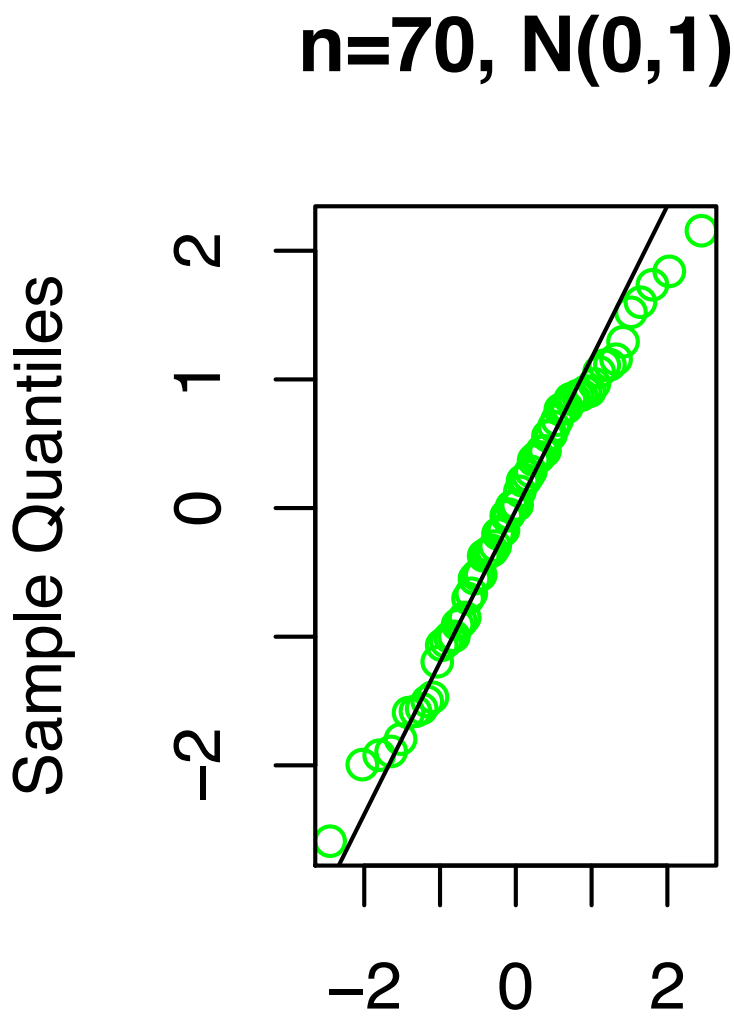




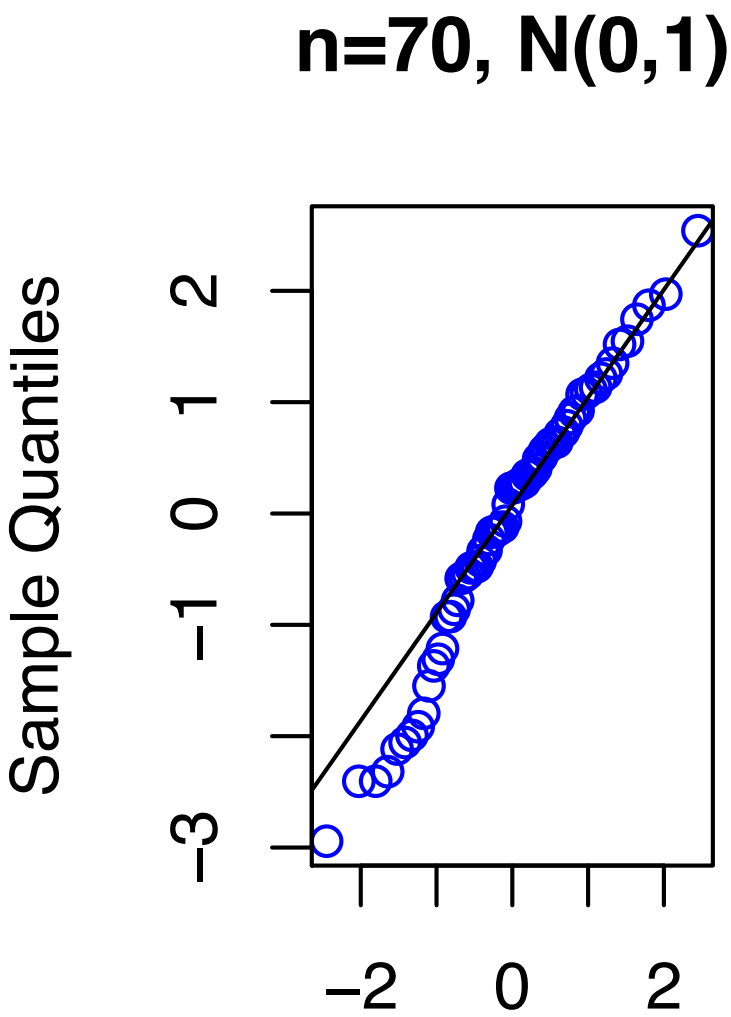


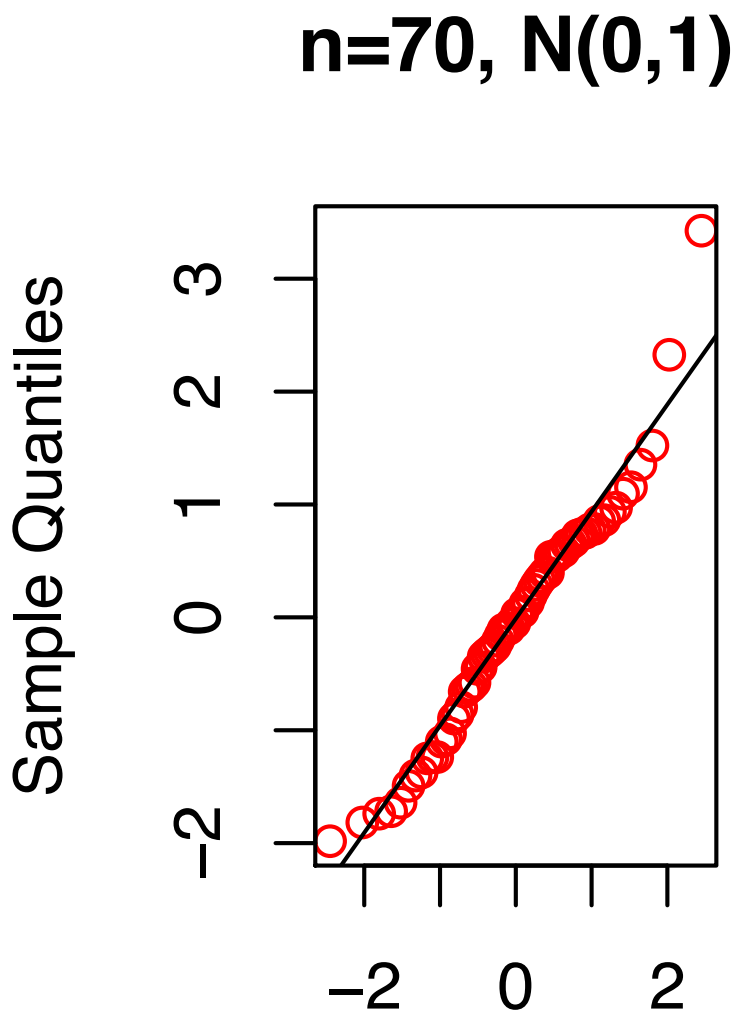
QQ plots


n=70, N(0,1)
Sample Quantiles
−2
−1
0
1
2
−2
0
2
Theoretical Quantiles

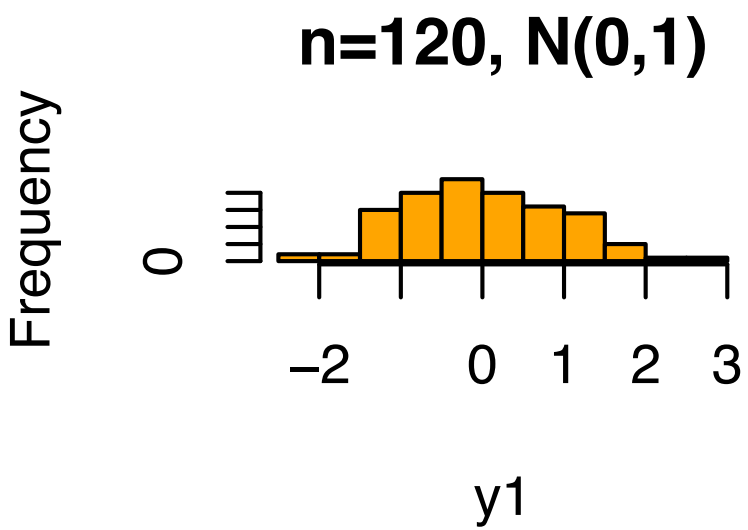

n=70, N(0,1)
Sample Quantiles
−2
0
1
2
−2
0
2
Theoretical Quantiles

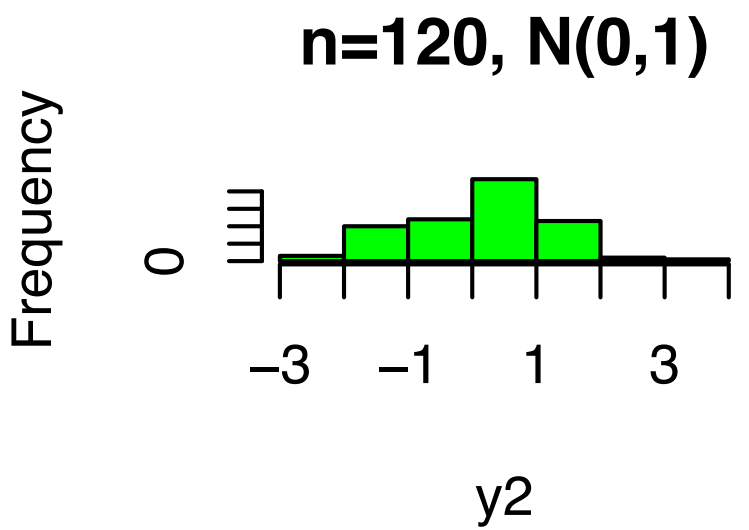
QQ plots

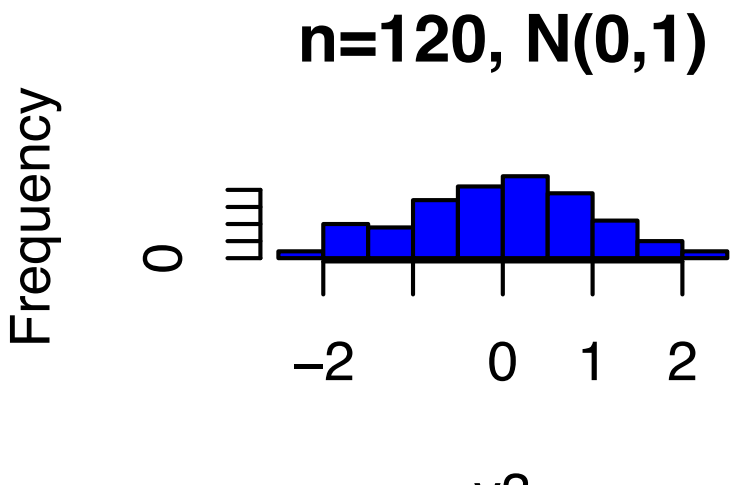

n=70, N(0,1)
Sample Quantiles
−3
−1
0
1
2
−2
0
2
Theoretical Quantiles

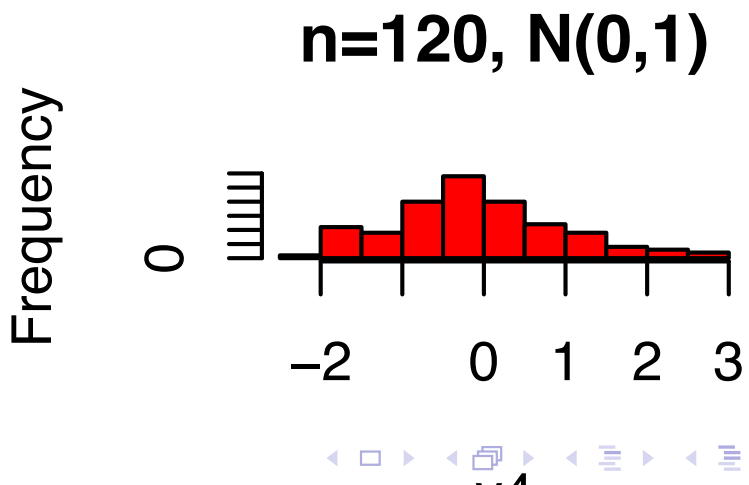

n=70, N(0,1)
Sample Quantiles
−2
0
1
2
3
−2
0
2
Theoretical Quantiles

Panel of graphs

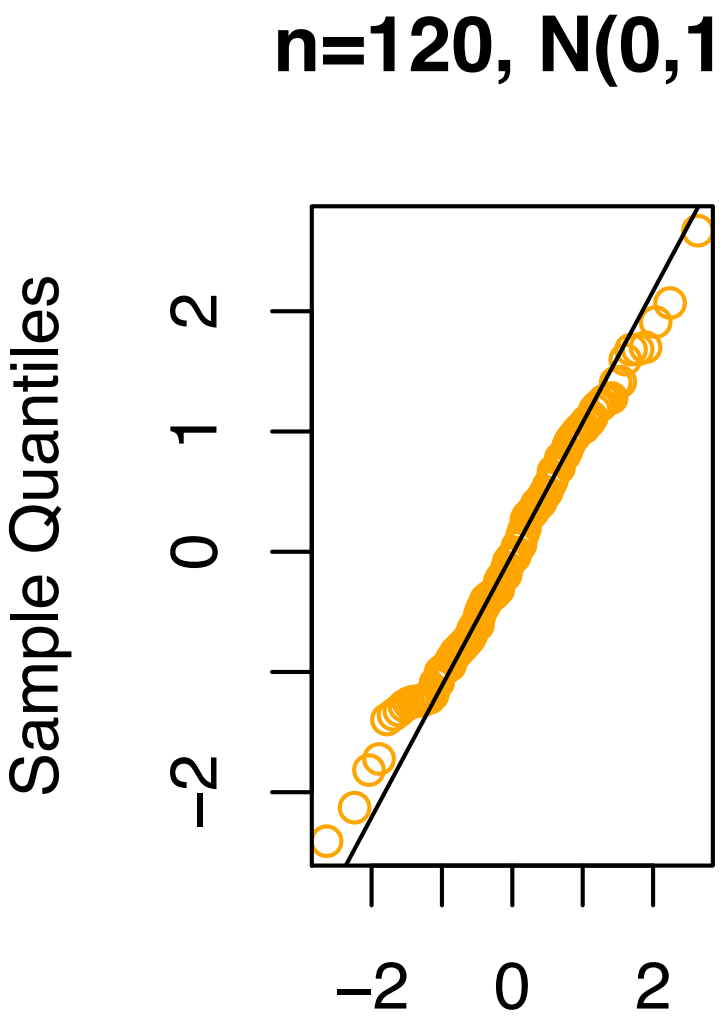


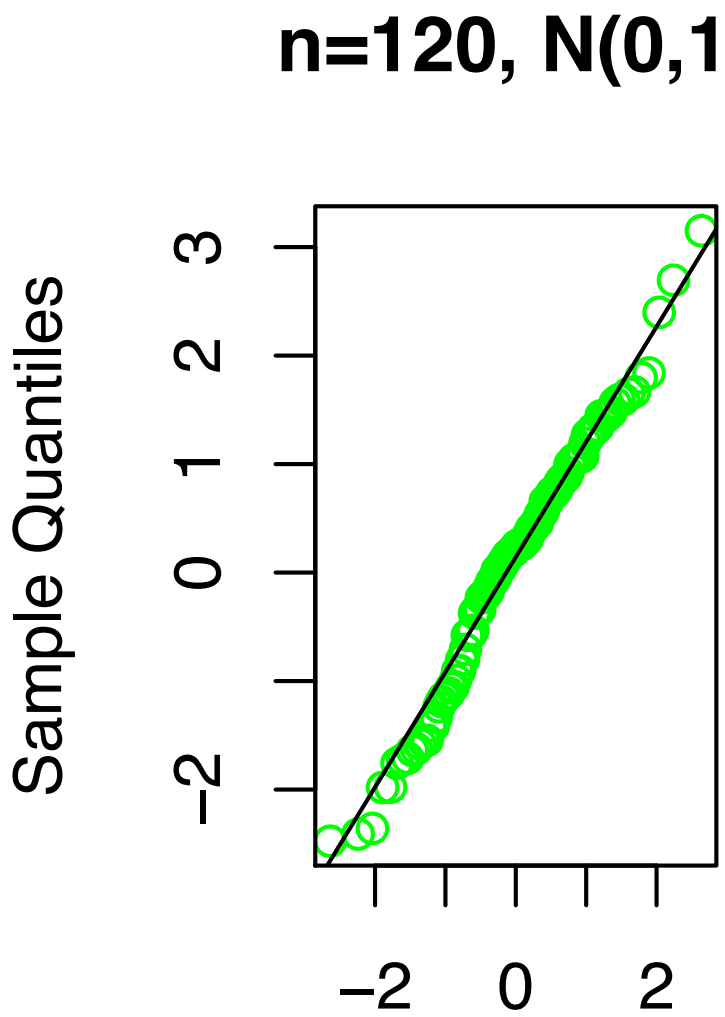




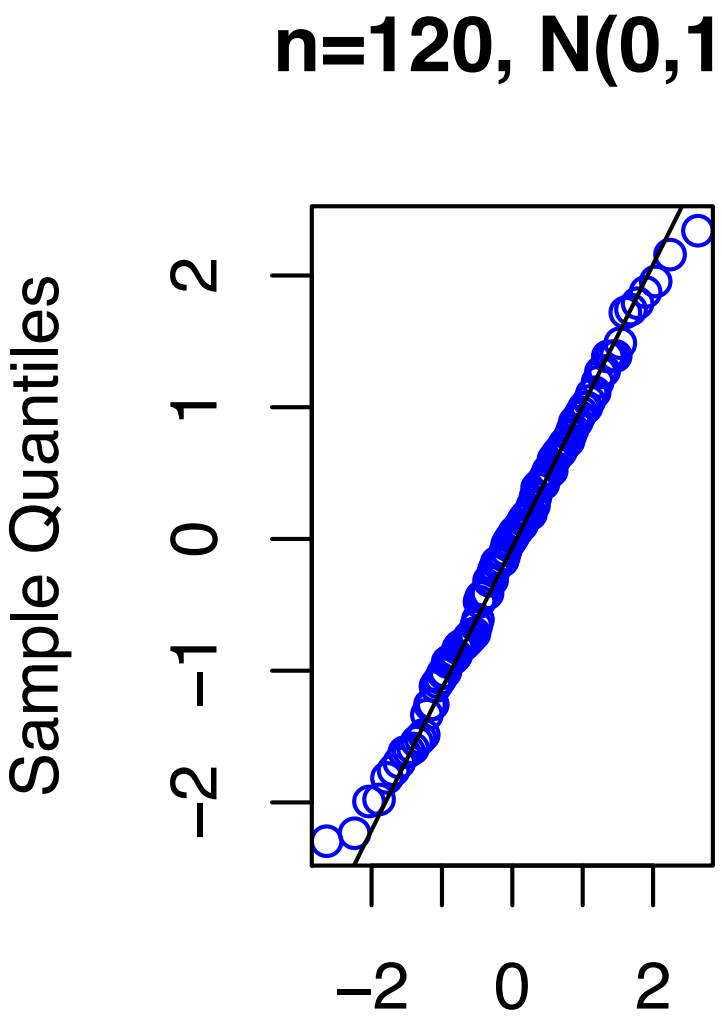


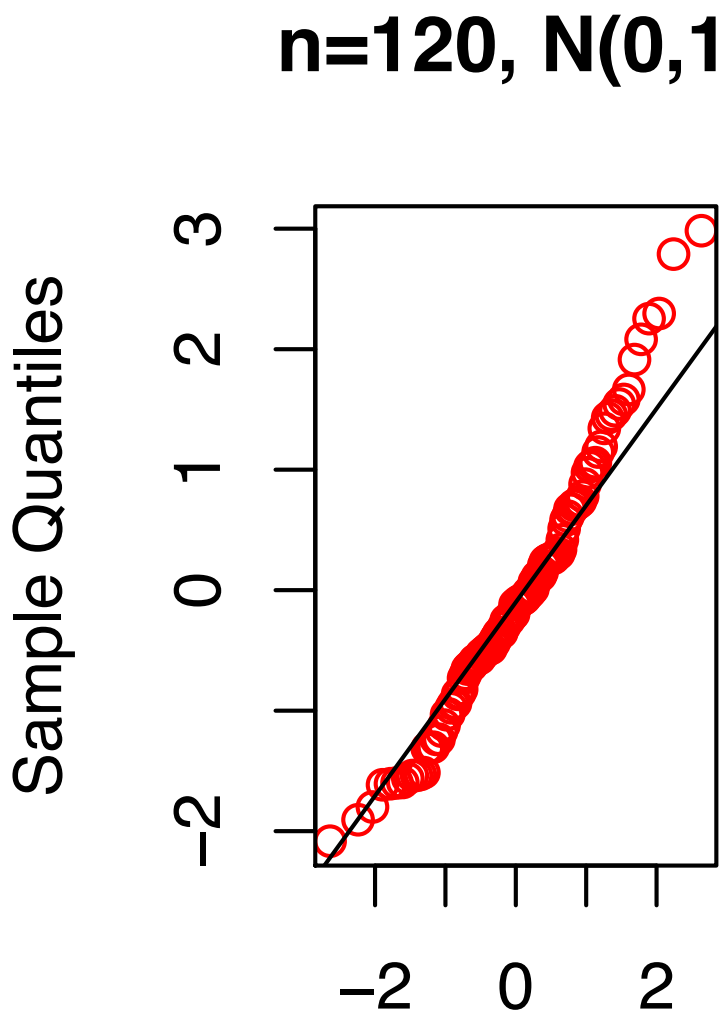
QQ plots


n=120, N(0,1)
Sample Quantiles
−2
0
1
2
−2
0
2
Theoretical Quantiles


n=120, N(0,1)
Sample Quantiles
−2
0
1
2
3
−2
0
2
Theoretical Quantiles

QQ plots


n=120, N(0,1)
Sample Quantiles
−2
−1
0
1
2
−2
0
2
Theoretical Quantiles


n=120, N(0,1)
Sample Quantiles
−2
0
1
2
3
−2
0
2
Theoretical Quantiles

APPENDIX

Analysis of Long-Distance Telephone Bills

As part of a larger study, a long-distance company wanted to acquire information about the monthly bills of new subscribers in the first month after signing with the company. The company's marketing manager conducted a survey of 200 new residential subscribers and recorded the first month's bills. The general manager planned to present his findings to senior executives. What information can be extracted from these data?

Reading data from txt files

```
# Step 1. Entering data;
# url of long-distance data;
phone_url =
"https://mcs.utm.utoronto.ca/~nosedal/data/phone.txt"

# import data in R;
phone_data= read.table(phone_url, header = TRUE);

phone_data[1:5, ];

names(phone_data);
```

Reading data from txt files

```
## [1]  42.19  38.45  29.23  89.35 118.04
## [1] "Bills"
```

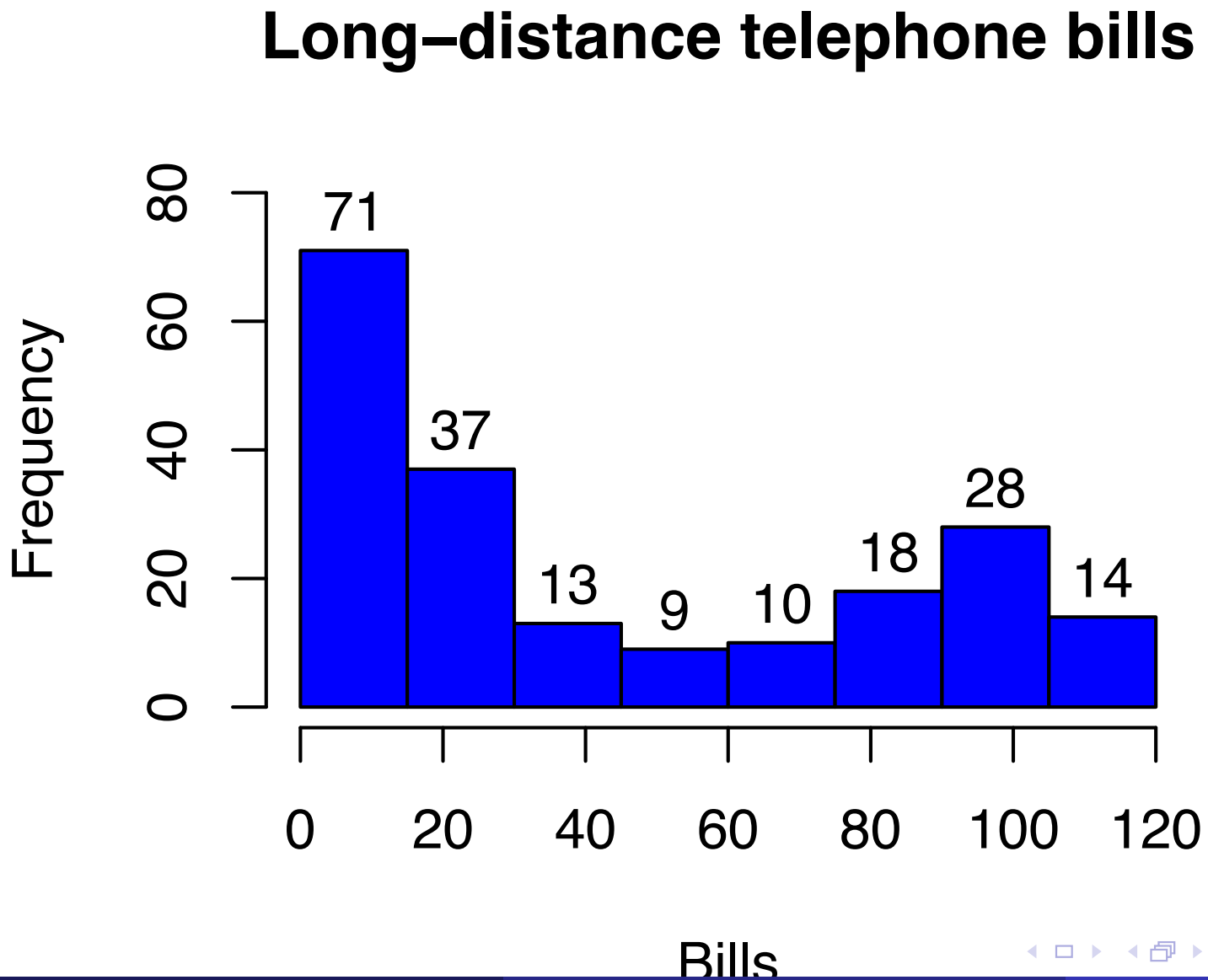
Making a histogram

Let us make a histogram that shows frequency counts. (This could provide useful information). As we already know, we create a frequency distribution for interval data by counting the number of observations that fall into each of a series of intervals, called classes, that cover the complete range of observations. We define our classes as follows:
Amounts that are less than or equal to 15.
Amounts that are more than 15 but less than or equal to 30.
Amounts that are more than 30 but less than or equal to 45.
⋮
Amounts that are more than 105 but less than or equal to 120.

R code

```
# Step 2. Making histogram;
classes=seq(0, 120, by =15);
# seq creates a sequence that starts at 0
# and ends at 120
# in jumps of 15;

hist(phone_data$Bills,breaks=classes,
col="blue",right=TRUE, labels=TRUE,
main="Long-distance telephone bills",
xlab="Bills",ylim=c(0,80));
# phone_bills£Bills tells R to use that column;
# main adds title to our histogram;
# xlab adds title to x-axis;
```

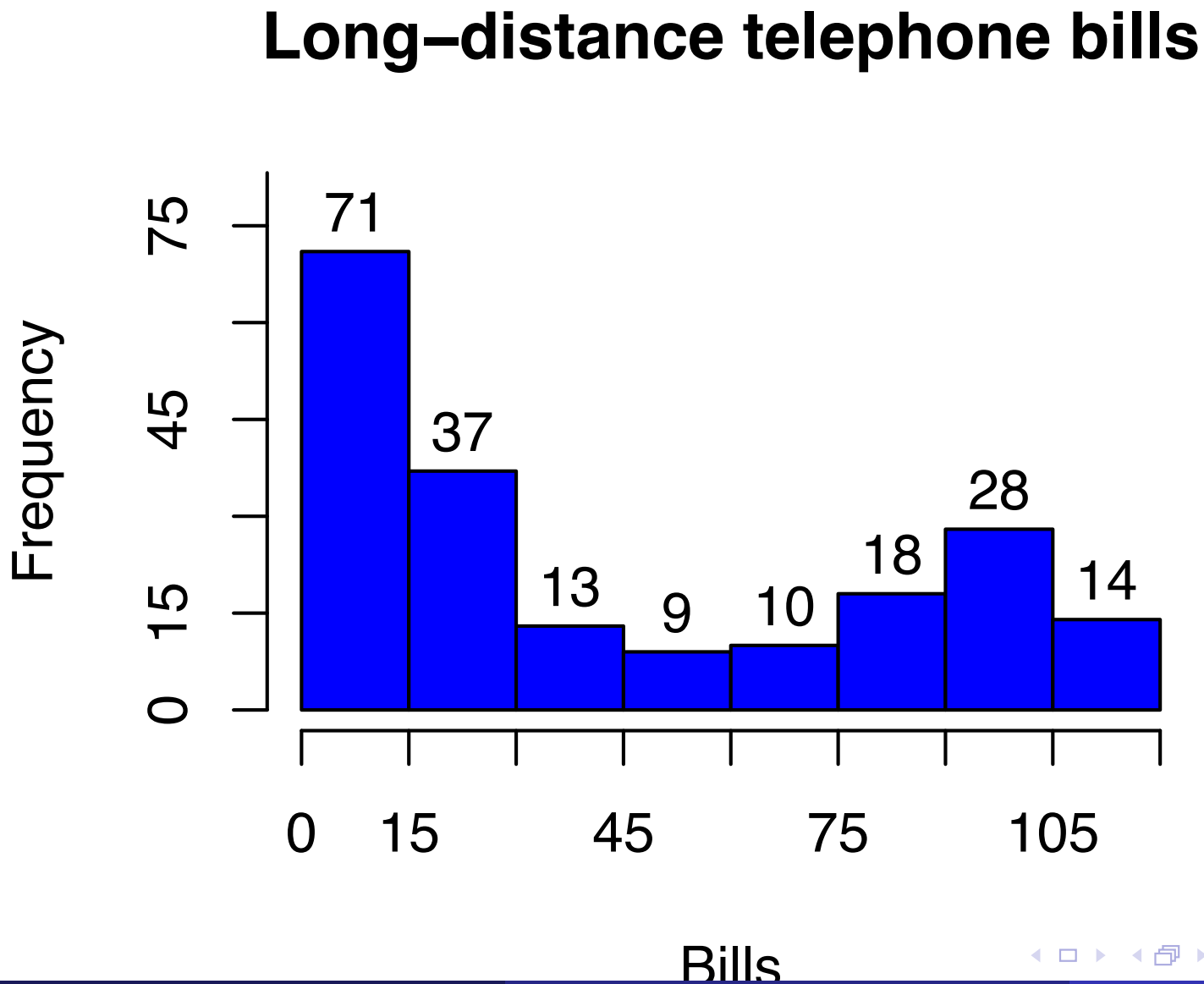
R code


Long–distance telephone bills
Frequency
0
20
40
60
80
71
37
13
9
10
18
28
14
0
20
40
60
80
100
120
Bills

R code (another version)

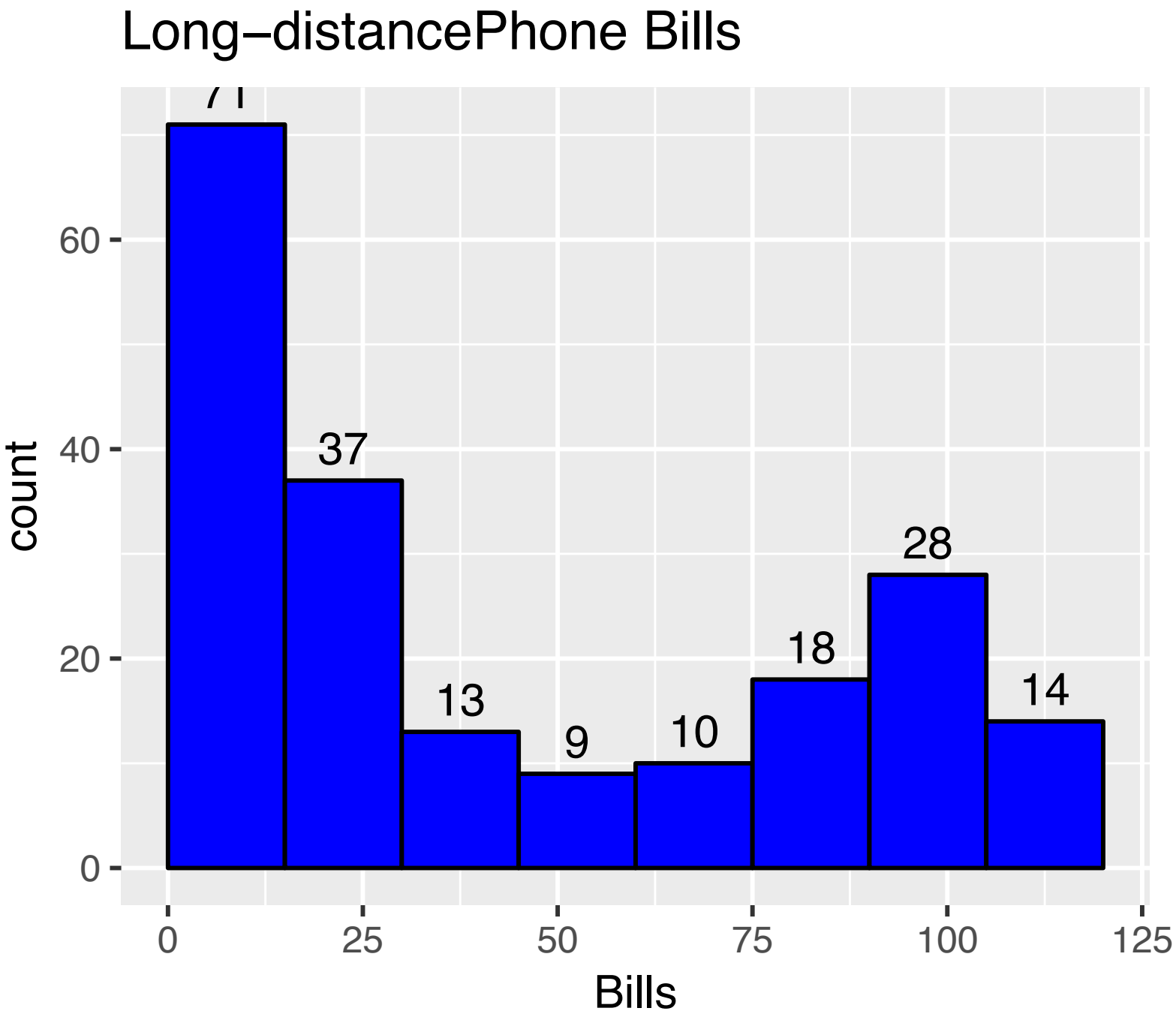
```r
# Step 2. Making histogram;
classes=seq(0, 120, by =15);
# seq creates a sequence that starts at 0
# and ends at 120
# in jumps of 15;

hist(phone_data$Bills,breaks=classes,
col="blue",right=TRUE, labels=TRUE, axes=FALSE,
main="Long-distance telephone bills",
xlab="Bills",ylim=c(0,80));
axis(1,at=seq(0,120,by=15));
# "new" scale for x axis;
axis(2,at=seq(0,90,by=15));
# "new" scale for y axis;
```

R code (another version)



Another version (from file in local folder)

```
phone_data= read.table(file="phone.txt", header = TRUE);
df=data.frame(phone_data);
library(ggplot2);
names(df);
ggplot(data=df,mapping=aes(x=Bills))+
geom_histogram(breaks=seq(0,120,by=15),
col="black",fill="blue")+
stat_bin(aes(y=..count.., label=..count..),
geom="text", vjust=-.5,
breaks=seq(0,120,by=15))+
ggtitle("Long-distancePhone Bills")
```


Long−distancePhone Bills
count
60
40
20
0
71
37
13
9
10
18
28
14
0
25
50
75
100
125
Bills

Comments

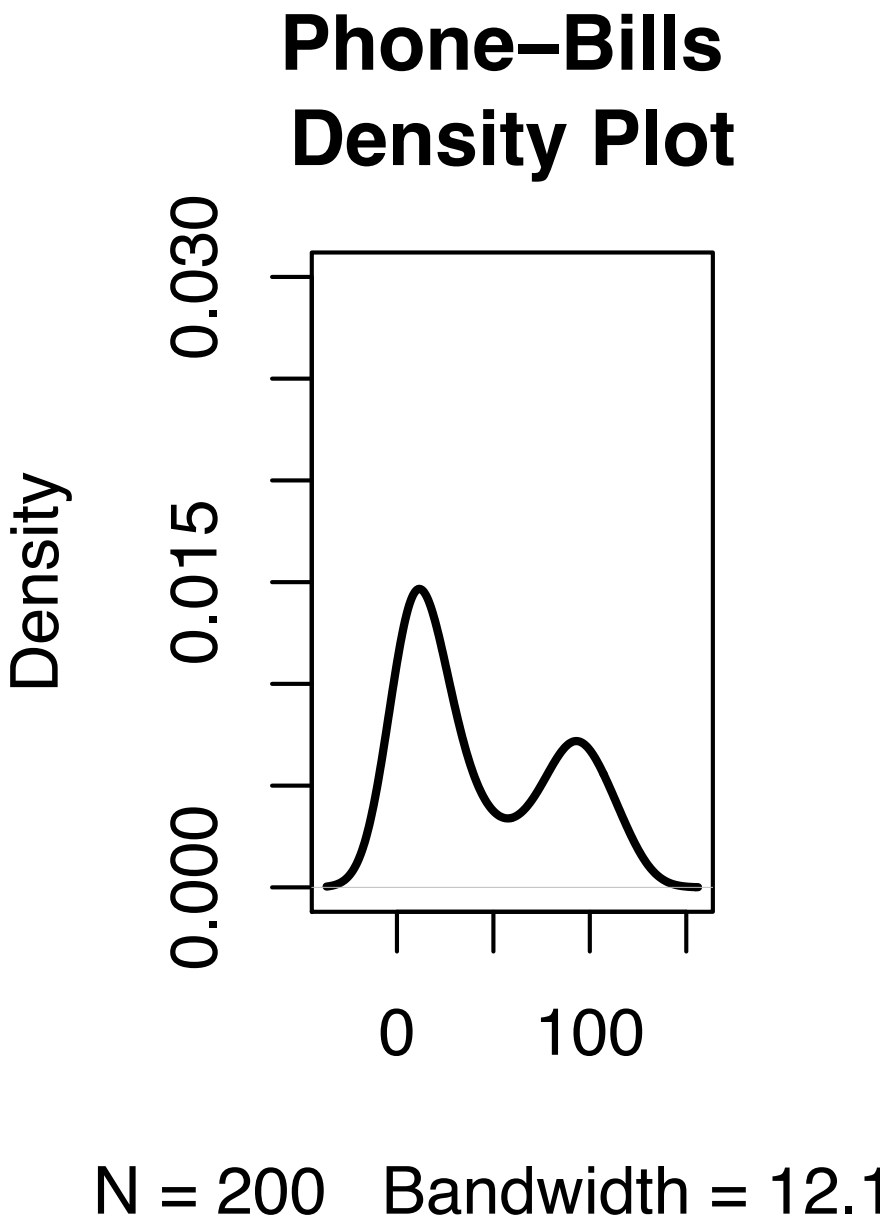
The histogram gives us a clear view of the way the bills are distributed. About half the monthly bills are small ($ 0 to $30), a few bills are in the middle range ($30 to $75), and a relatively large number of long-distance bills are at the high end of the range. It would appear from this sample of first-month long-distance bills that the company's customers are split unevenly between light and heavy users of long-distance telephone service.

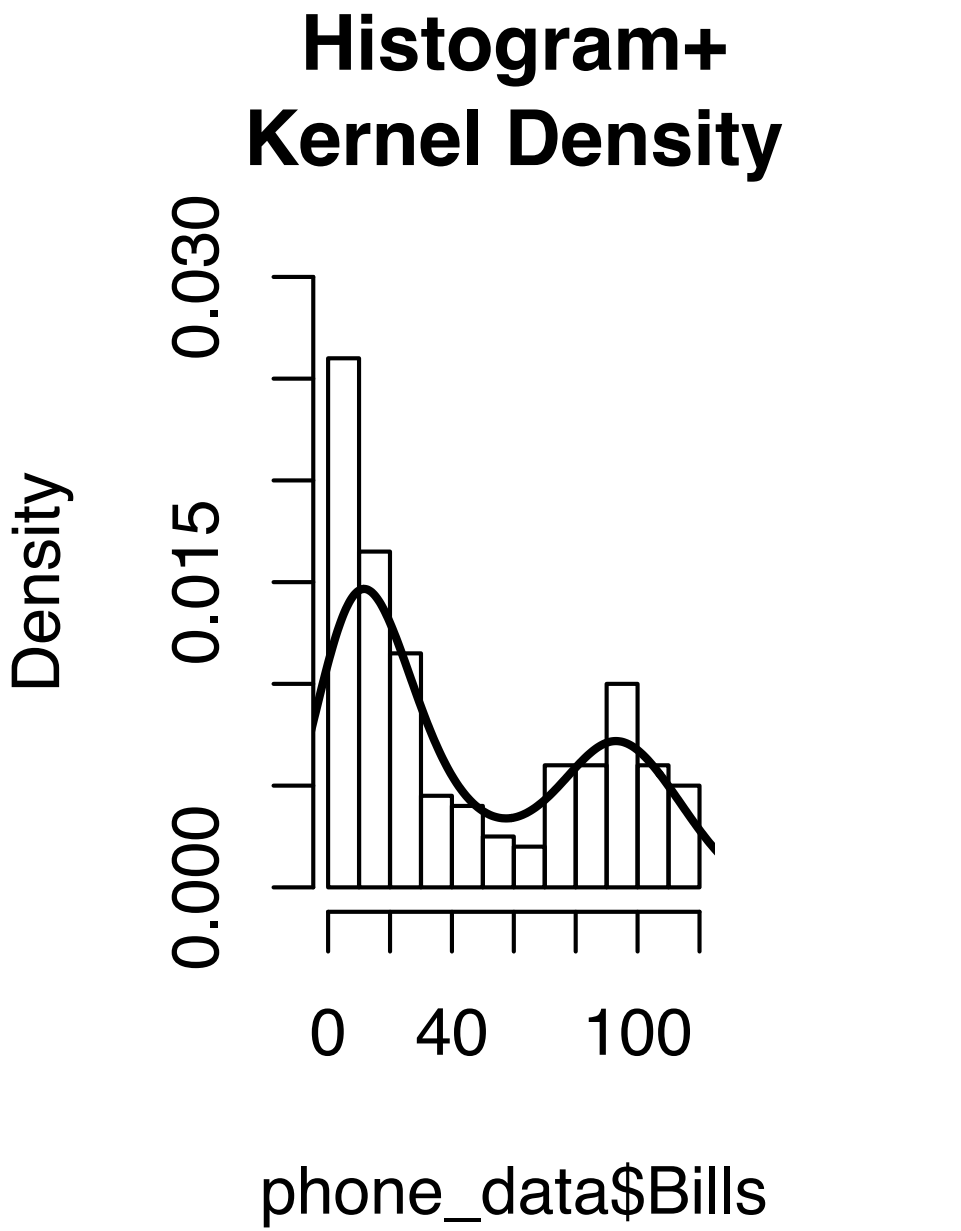
Kernel densities

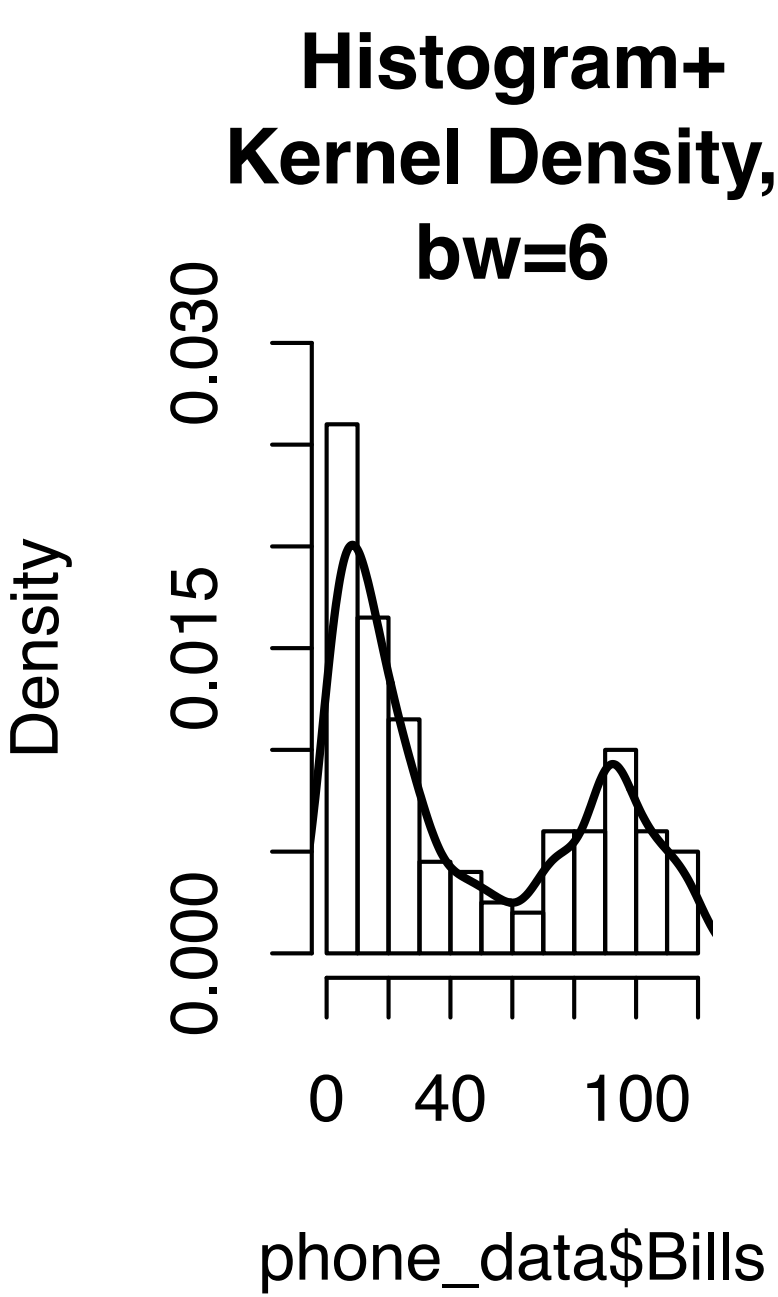
A common problem in science is to estimate, from a data sample, a mathematical function that describes the relative likelihood that a variable (such as long-distance bills) takes a particular value. The rule, or formula, that gives the likelihood of a given value of, for example, long-distance bills is called the density function.
The graphs shown below are kernel density plots, smooth line approximations of the density function.

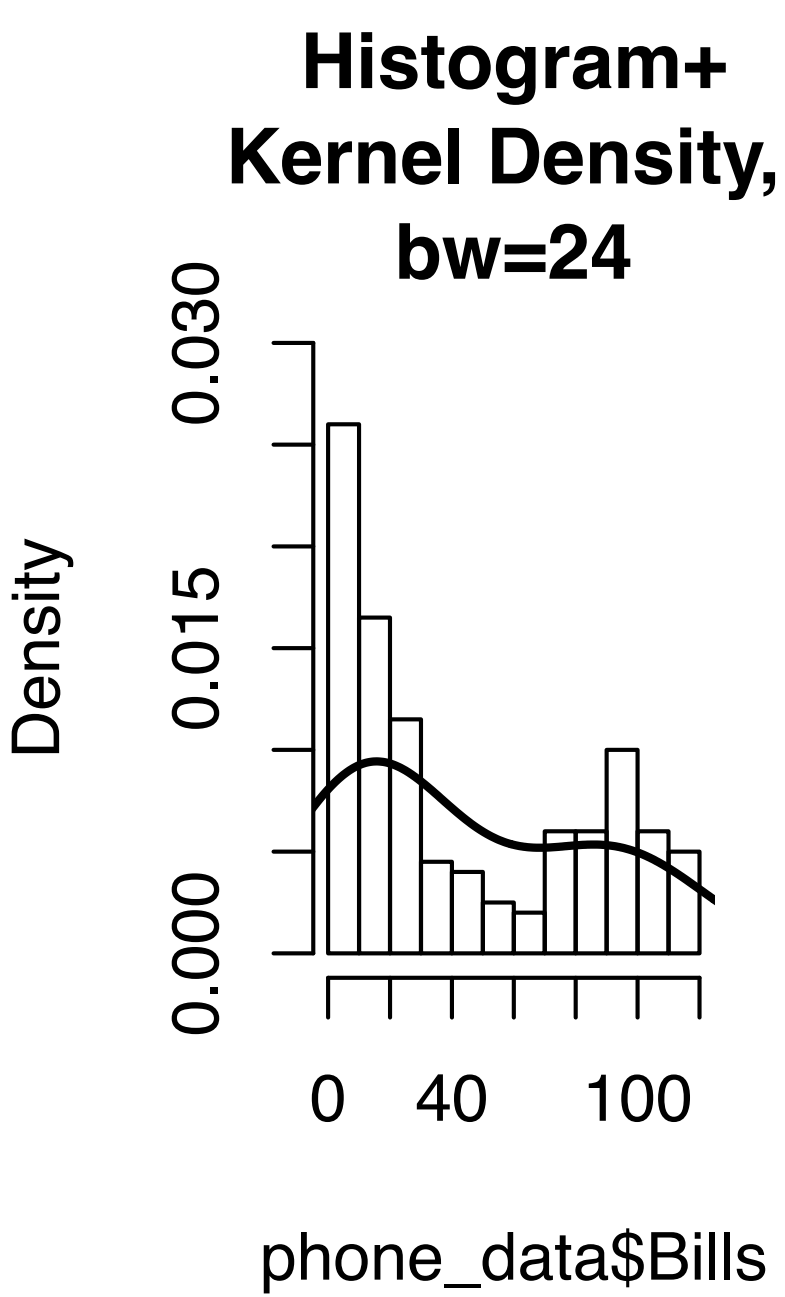
```
den=density(phone_data$Bills);
par(mfrow=c(1,2) )

plot(den,main="Phone-Bills
Density Plot",ylim=c(0,0.03),
lwd=2)

hist(phone_data$Bills,freq=F,ylim=c(0,0.03),
main="Histogram+
Kernel Density");
lines(den, lwd=2)
```

```
den2=density(phone_data$Bills,bw=6);
den3=density(phone_data$Bills,bw=24);

par(mfrow=c(1,2) )

hist(phone_data$Bills,freq=F,ylim=c(0,0.03),
main="Histogram+
Kernel Density,
bw=6");
lines(den2, lwd=2)

hist(phone_data$Bills,freq=F,ylim=c(0,0.03),
main="Histogram+
Kernel Density,
bw=24");
lines(den3, lwd=2)
```


Phone–Bills
Density Plot
Density
0.000
0.015
0.030
0
100
N = 200 Bandwidth = 12.1

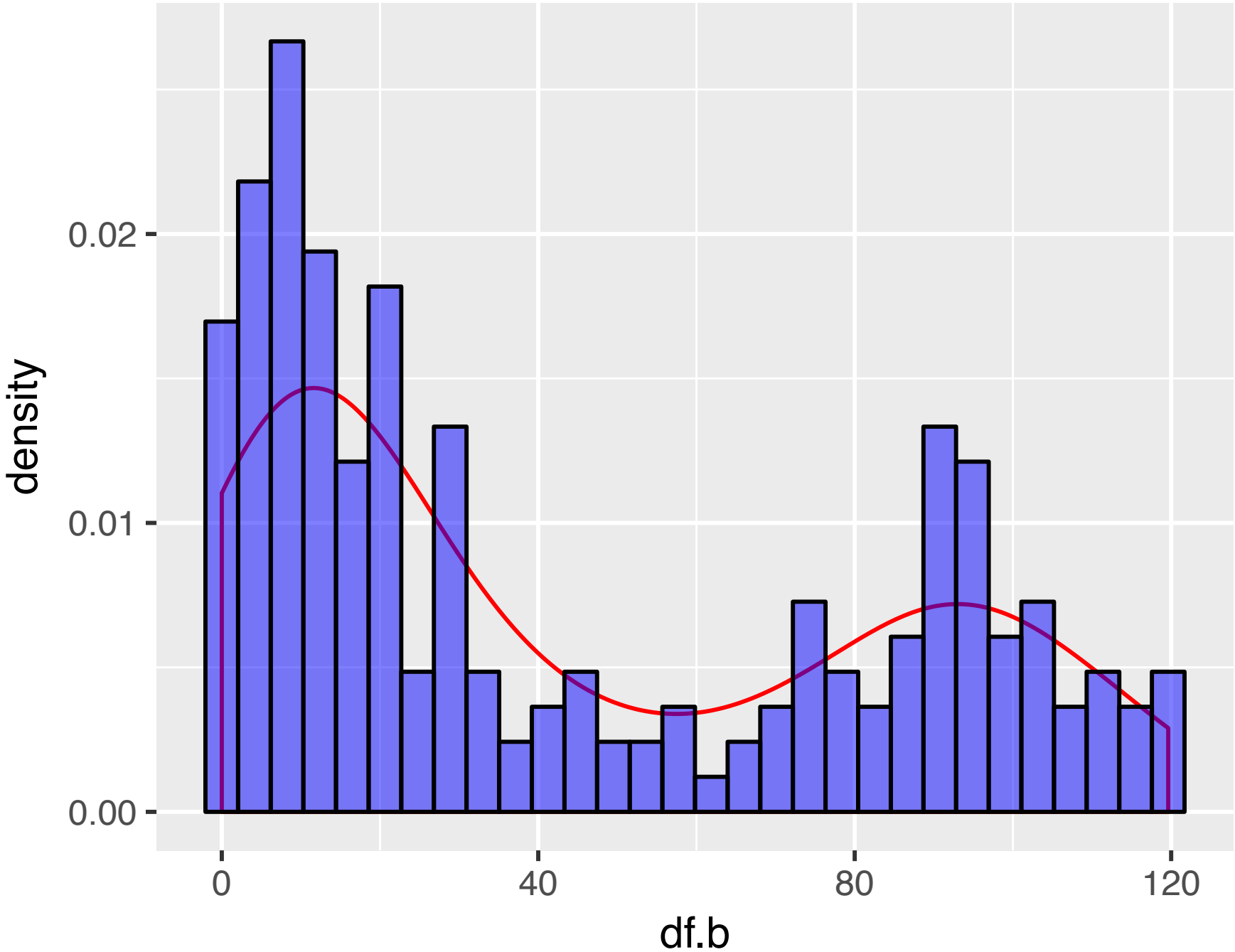

Histogram+
Kernel Density
Density
0.000
0.015
0.030
0
40
100
phone_data$Bills


Histogram+
Kernel Density,
bw=6
Density
0.000
0.015
0.030
0
40
100
phone_data$Bills


Histogram+
Kernel Density,
bw=24
Density
0.000
0.015
0.030
0
40
100
phone_data$Bills

Using ggplot2

```
df.b=data.frame(phone_data$Bills);

ggplot(data=df.b,mapping=aes(x=df.b,y=..density..))+
geom_density(alpha=0.5,col="red")+
geom_histogram(col="black",fill="blue",alpha=0.5)
```


density
0.02
0.01
0.00
0
40
80
120
df.b

1.5 IQR Rule

Identifying suspected outliers. Whether an observation is an outlier is a matter of judgement: does it appear to clearly stand apart from the rest of the distribution? When large volumes of data are scanned automatically, however, we need a rule to pick out suspected outliers. The most common rule is the 1.5 IQR rule. A point is a suspected outlier if it lies more than 1.5 IQR below the first quartile Q_1 or above the third quartile Q_3.

A high income.

In our income problem, we noted the influence of one high income of \$110,000 among the incomes of a sample of 15 college graduates. Does the 1.5 IQR rule identify this income as a suspected outlier?

Solution

Data: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.
Q_1 and Q_3 are given by:
Q_1 =30 and Q_3 =55
$Q_3 + 1.5$ IQR $= 55 + 1.5(25) = 92.5$
Since $110 > 92.5$ we conclude that 110 is an outlier.

R code

```r
# Step 1. Entering data;
income=c(4, 25, 30, 30, 30, 32, 32, 35,
         50, 50, 50, 55, 60, 74, 110);

# Step 2. Making boxplot

boxplot(income,col="blue",
        ylab="Income (thousands of dollars)")
# this version identifies
# suspected outliers.
```

R code