

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and Brian Caffo. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

Outline

1. Histograms
2. Stem-and-leaf plots
3. Dot charts and dot plots
4. Boxplots
5. Kernel density estimates
6. QQ-plots

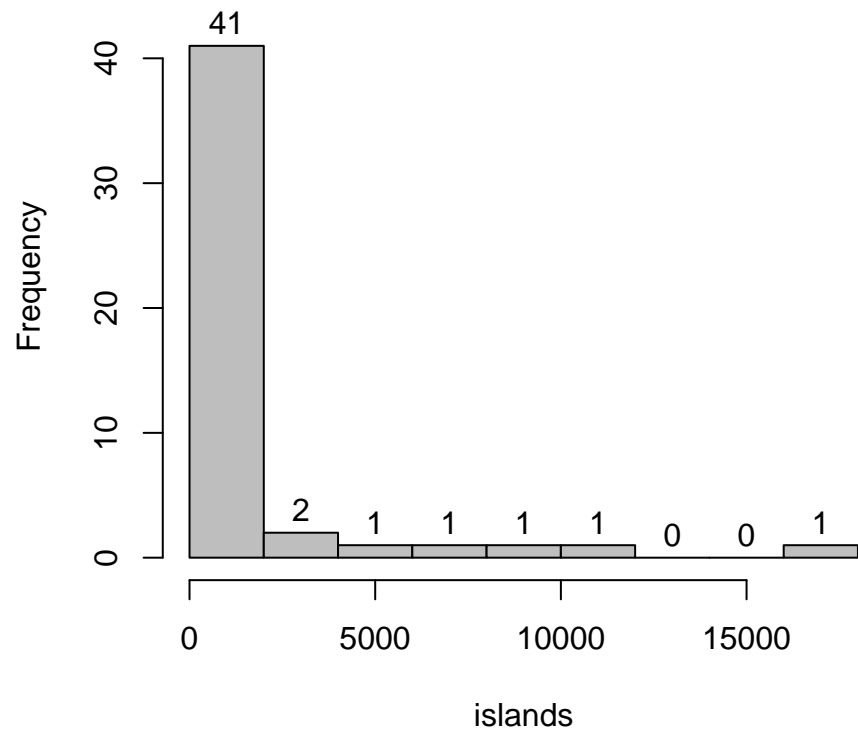
Histograms

- Histograms display a sample estimate of the density or mass function by plotting a bar graph of the frequency or proportion of times that a variable takes specific values, or a range of values for continuous data, within a sample

Example

- The data set `islands` in the R package `datasets` contains the areas of all land masses in thousands of square miles
- Load the data set with the command `data(islands)`
- View the data by typing `islands`
- Create a histogram with the command `hist(islands)`
- Do `?hist` for options

Histogram of islands

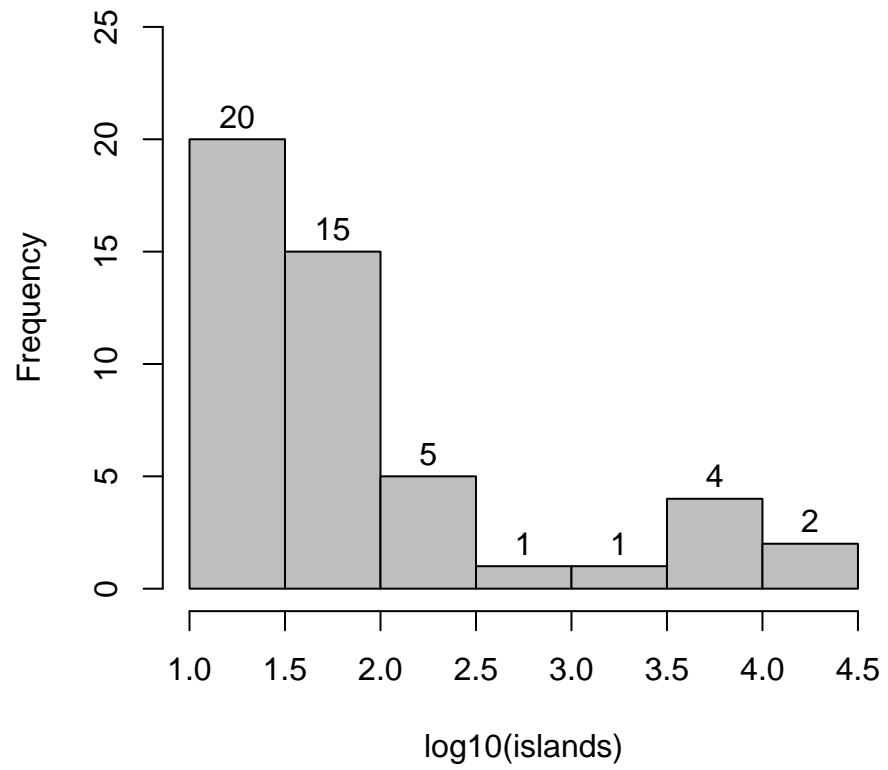


Pros and cons

- Histograms are useful and easy, apply to continuous, discrete and even unordered data
- They use a lot of ink and space to display very little information
- It's difficult to display several at the same time for comparisons

Also, for this data it's probably preferable to consider log base 10 (orders of magnitude), since the raw histogram simply says that most islands are small

Histogram of log10(islands)



Stem-and-leaf plots

- Stem-and-leaf plots are extremely useful for getting distribution information on the fly
- Read the text about creating them
- They display the complete data set and so waste very little ink
- Two data sets' stem and leaf plots can be shown back-to-back for comparisons
- Created by John Tukey, a leading figure in the development of the statistical sciences and signal processing

Example

```
> stem(log10(islands))
```

The decimal point is at the |

```
1 | 1111112222233444
```

```
1 | 5555556666667899999
```

```
2 | 3344
```

```
2 | 59
```

```
3 |
```

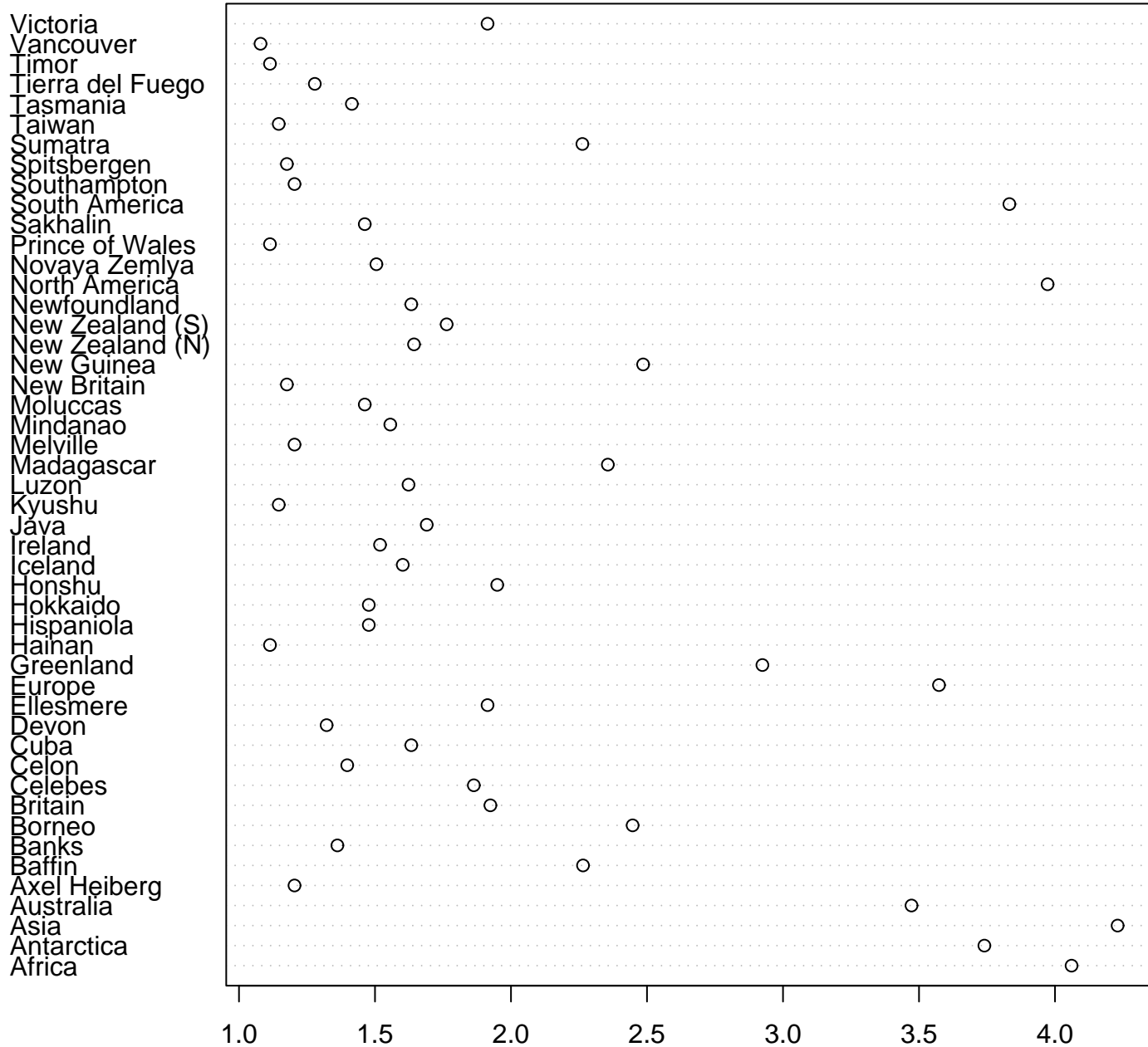
```
3 | 5678
```

```
4 | 012
```

Dotcharts

- Dotcharts simply display a data set, one point per dot
- Ordering of the of the dots and labeling of the axes can the display additional information
- Dotcharts show a complete data set and so have high data density
- May be impossible to construct/difficult to interpret for data sets with lots of points

islands data: log10(area) (log10(sq. miles))



Discussion

- Maybe ordering alphabetically isn't the best thing for this data set
- Perhaps grouped by continent, then nations by geography (grouping Pacific islands together)?

Dotplots comparing grouped data

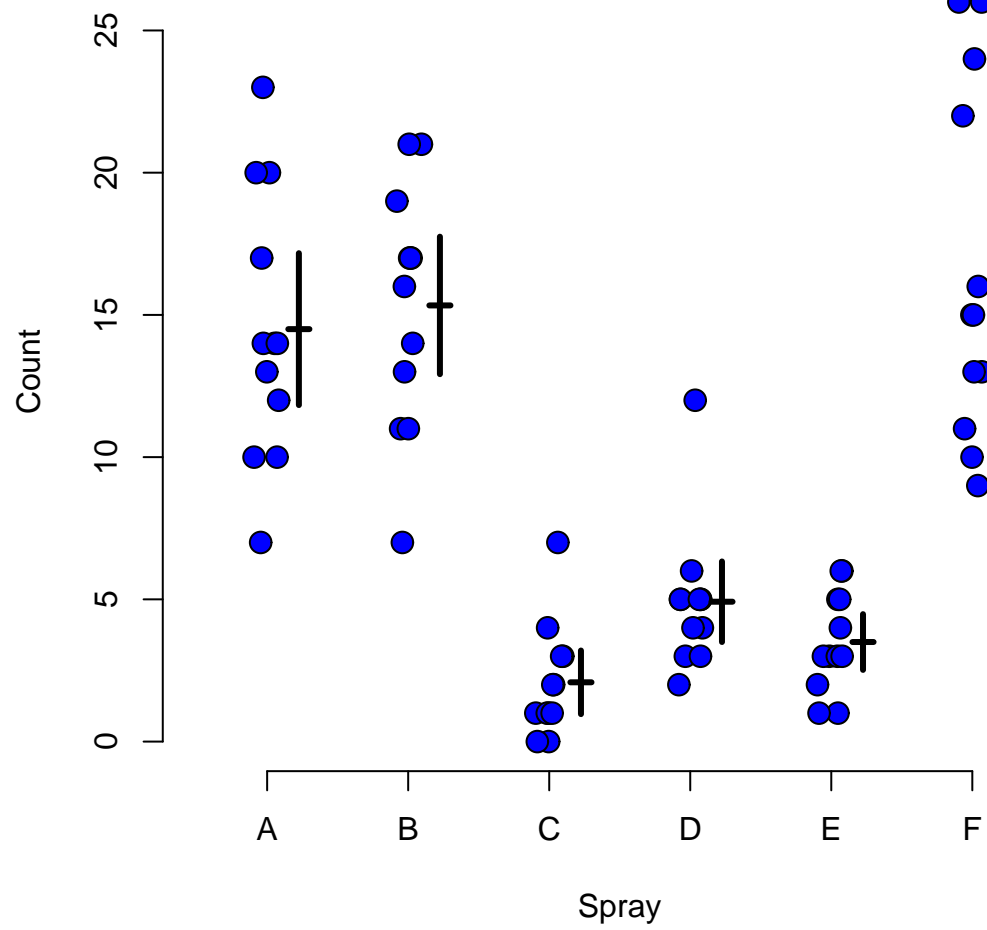
- For data sets in groups, you often want to display density information by group
- If the size of the data permits, displaying the whole data is preferable
- Add horizontal lines to depict means, medians
- Add vertical lines to depict variation, show confidence intervals interquartile ranges
- Jitter the points to avoid overplotting (`jitter`)

Example

- The InsectSprays dataset contains counts of insect deaths by insecticide type (A, B, C, D, E, F)
- You can obtain the data set with the command
`data(InsectSprays)`

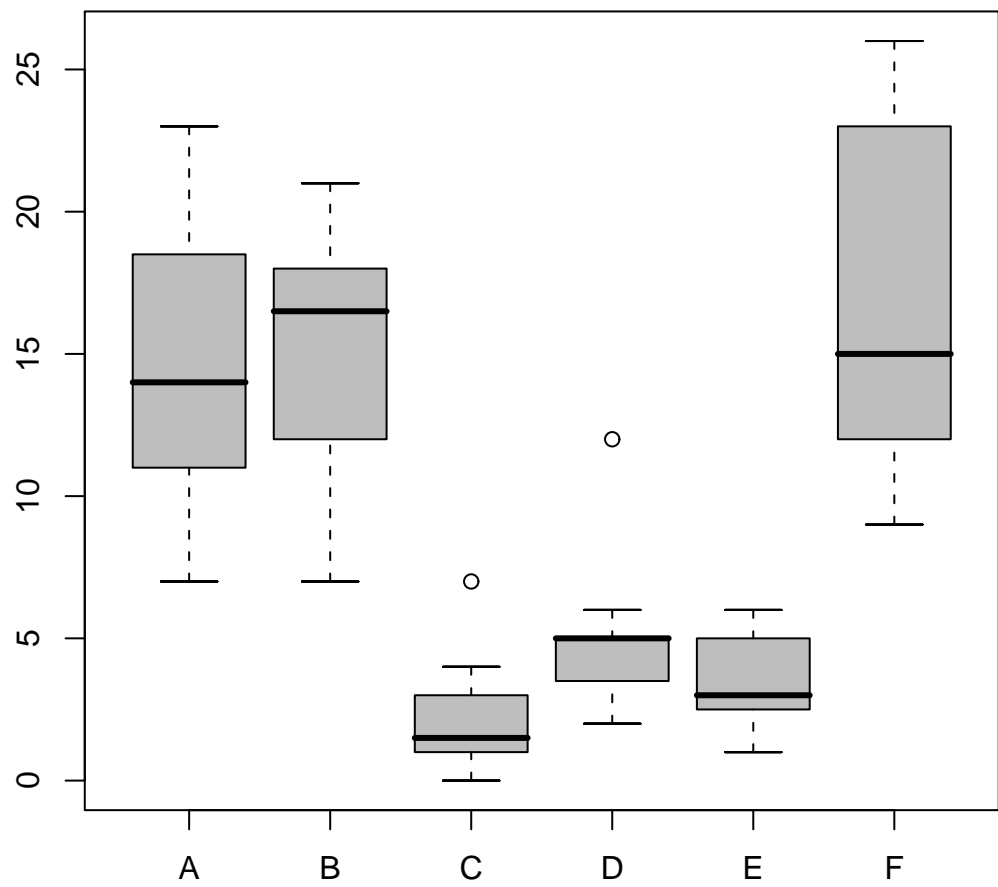
The gist of the code is below

```
attach(InsectSprays)
plot(c(.5, 6.5), range(count))
sprayTypes <- unique(spray)
for (i in 1 : length(sprayTypes)){
  y <- count[spray == sprayTypes[i]]
  n <- sum(spray == sprayTypes[i])
  points(jitter(rep(i, n), amount = .1), y)
  lines(i + c(.12, .28), rep(mean(y), 2), lwd = 3)
  lines(rep(i + .2, 2),
        mean(y) + c(-1.96, 1.96) * sd(y) / sqrt(n)
        )
}
```



Boxplots

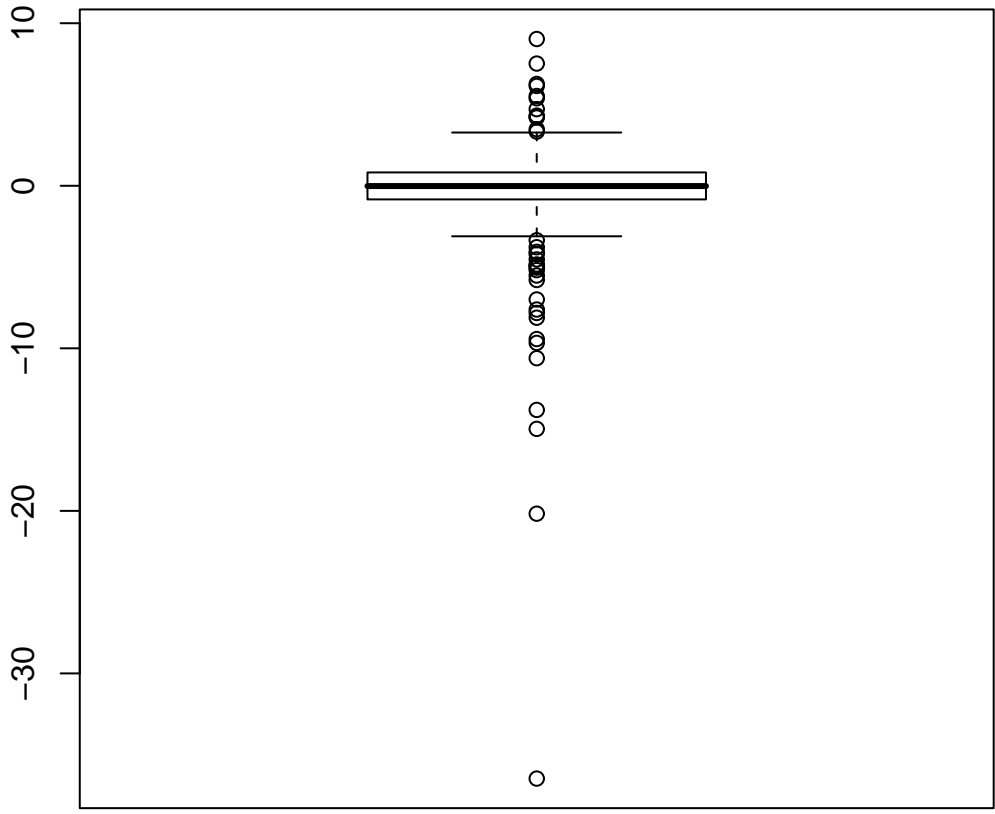
- Boxplots are useful for the same sort of display as the dot chart, but in instances where displaying the whole data set is not possible
- Centerline of the boxes represents the median while the box edges correspond to the quartiles
- Whiskers extend out to a constant times the IQR or the max value
- Sometimes potential outliers are denoted by points beyond the whiskers
- Also invented by Tukey
- Skewness indicated by centerline being near one of the box edges



Boxplots discussion

- Don't use boxplots for small numbers of observations, just plot the data!
- Try logging if some of the boxes are too squished relative to other ones; You can convert the axis to unlogged units (though they will not be equally spaced anymore)
- For data with lots and lots of observations omit the outliers plotting if you get so many of them that you cant see the points
- Example of a bad box plot

```
boxplot(rt(500, 2))
```



Kernel density estimates

- Kernel density estimates are essentially more modern versions of histograms providing density estimates for continuous data
- Observations are weighted according to a “kernel”, in most cases a Gaussian density
- “Bandwidth” of the kernel effectively plays the role of the bin size for the histogram
 - a. Too low of a bandwidth yields a too variable (jagged) measure of the density
 - b. Too high of a bandwidth oversmooths
- The R function `density` can be used to create KDEs

Example

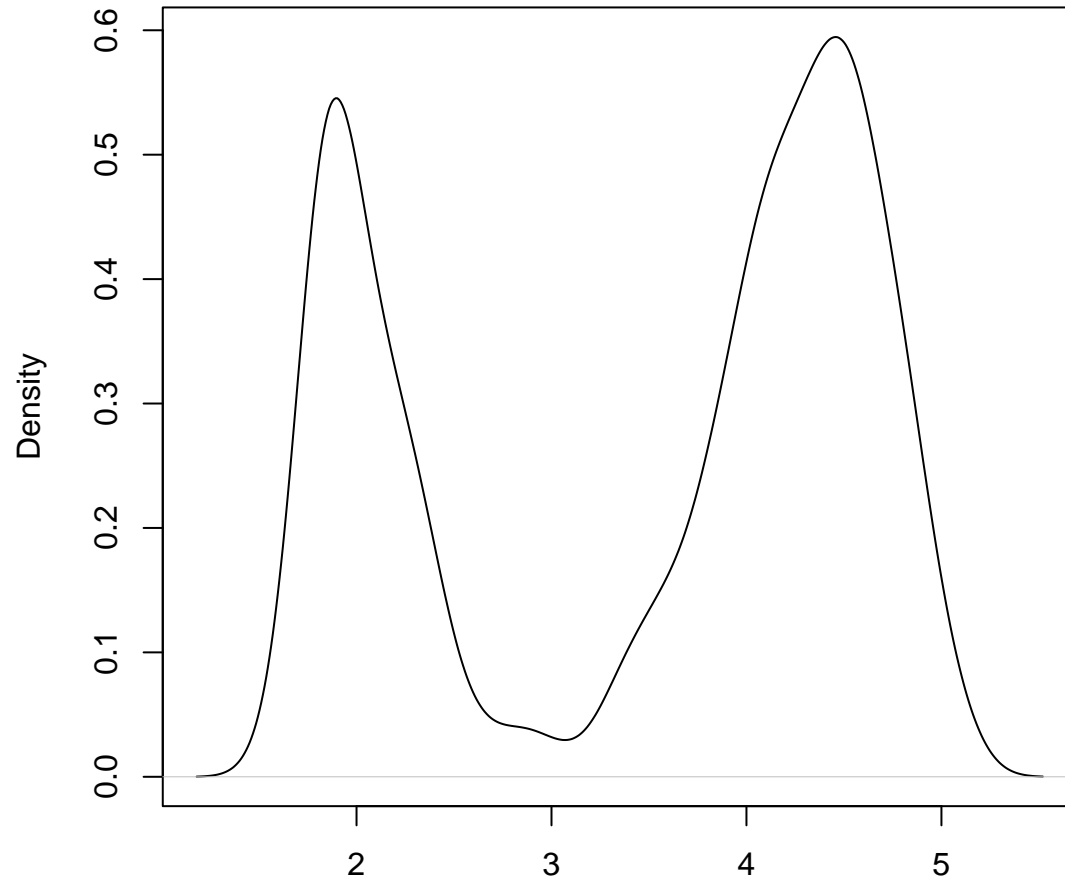
Data is the waiting and eruption times in minutes between eruptions of the Old Faithful Geyser in Yellowstone National park

```
data(faithful)
```

```
d <- density(faithful$eruptions, bw = "sj")
```

```
plot(d)
```

`density.default(x = faithful$eruptions, bw = "sj")`



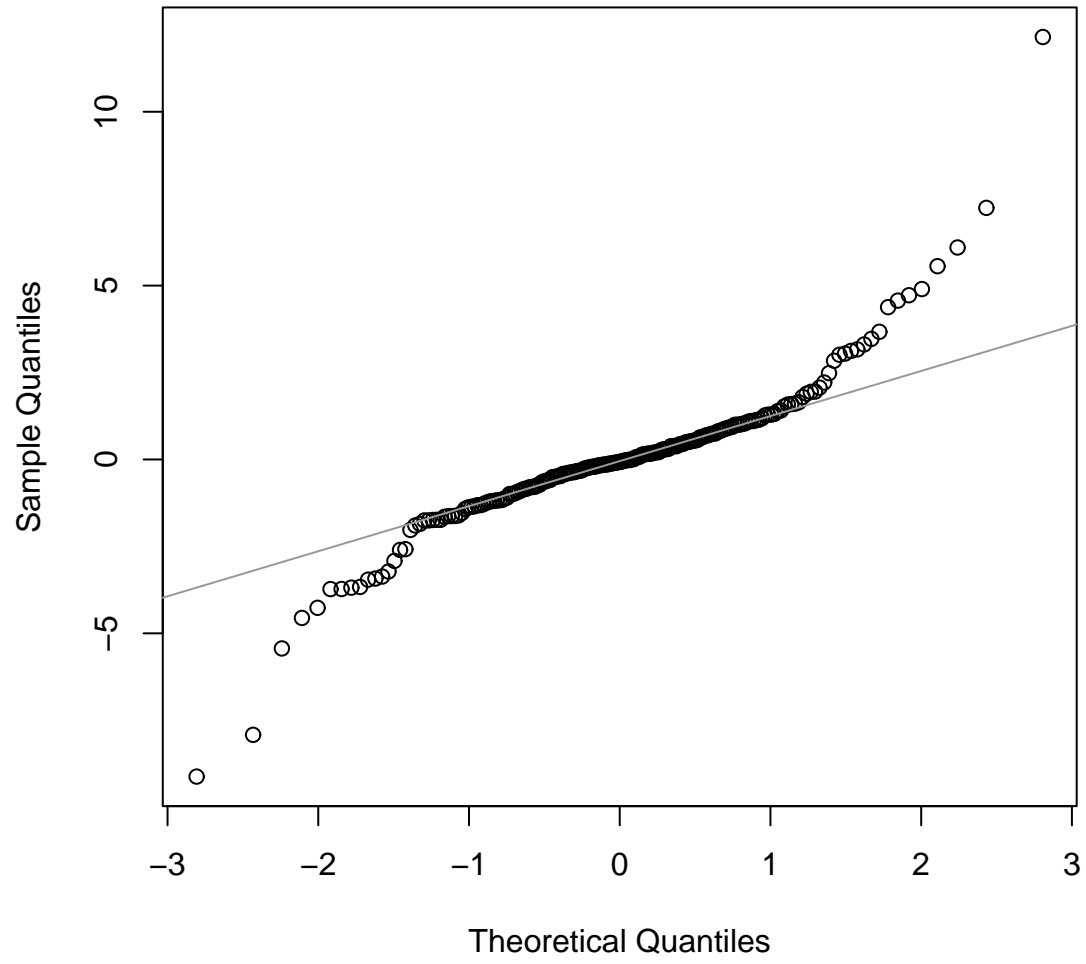
N = 272 Bandwidth = 0.14

QQ-plots

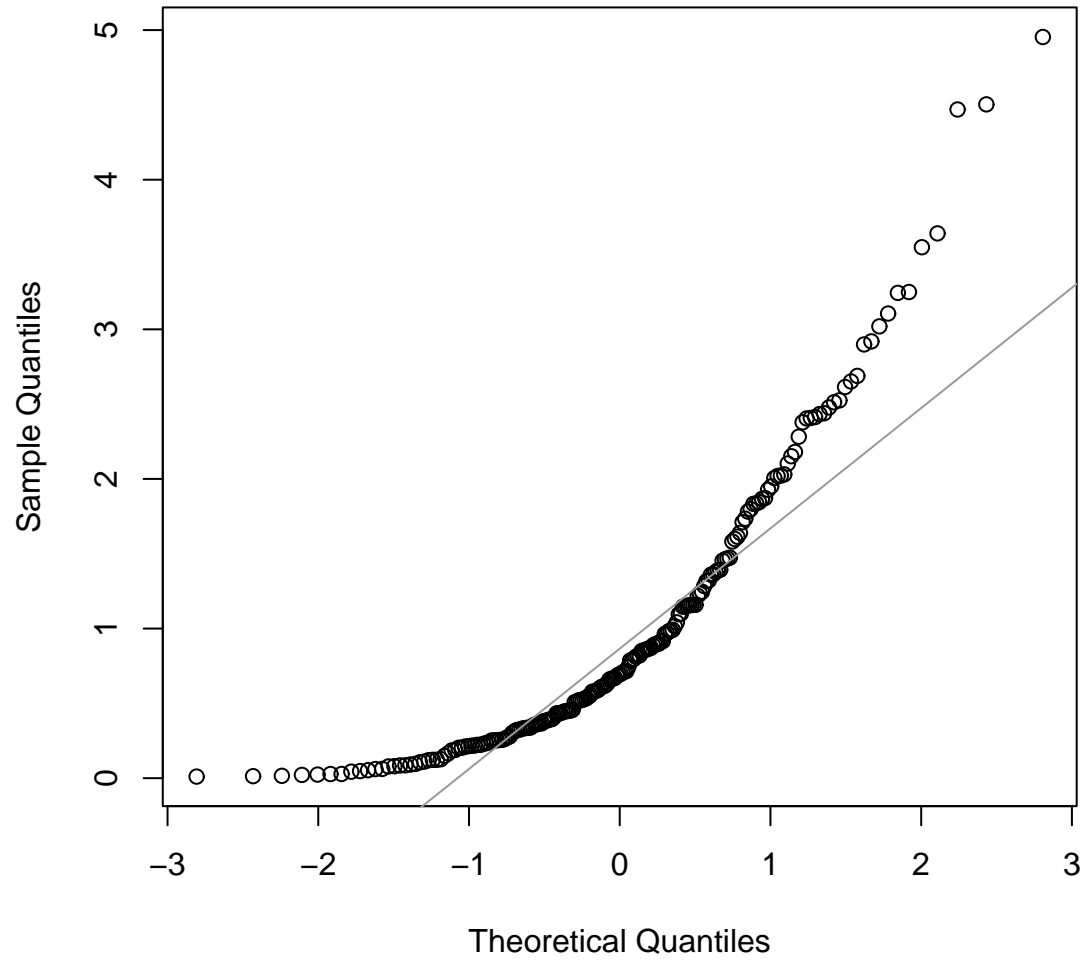
- QQ-plots (for quantile-quantile) are extremely useful for comparing data to a theoretical distribution
- Plot the empirical quantiles against theoretical quantiles
- Most useful for diagnosing normality

- Let x_p be the p^{th} quantile from a $N(\mu, \sigma^2)$
- Then $P(X \leq x_p) = p$
- Clearly $P(Z \leq \frac{x_p - \mu}{\sigma}) = p$
- Therefore $x_p = \mu + z_p\sigma$ (this should not be news)
- Result, quantiles from a $N(\mu, \sigma^2)$ population should be linearly related to standard normal quantiles
- A normal qq-plot plot the empirical quantiles against the theoretical standard normal quantiles
- In R `qqnorm` for a normal QQ-plot and `qqplot` for a qq-plot against an arbitrary distribution

Normal Q-Q Plot



Normal Q-Q Plot



Normal Q-Q Plot

