Describing data

Although many think of statistics as applying quantitative data to test hypotheses, a large component is *describing* data. In some cases these descriptions can lead to generation of hypotheses.

Accuracy versus Precision

- Accuracy is closeness to the "true" value
  - Bias is the deviation from true value.
- Precision is repeatability

Accuracy and Precision: Discrete versus continuous

- Perfect accuracy is possible when
  - measuring discrete observations **and**
  - there is no measurement error
- Perfect accuracy is impossible
  - When measuring continuous variables **or**
  - there is any measurement error

Significant Digits

The significant digits you report should reflect the process you are measuring and the instrument with which you are measuring.

Derived variables

Most derived variables are in the form of ratios. These include percentages (number observed out of 100), indices, proportions, and rates (number of times an event occurs per time unit).
Even data that appears to be strictly counts can be thought of as ratio data. Table 1 gives counts of numbers of fruits per individual plant.

Counts

	block	plant	fruits
1	4	1	2027
2	5	1	1879
3	6	1	2723
4	7	17	984
5	8	1	442
6	9	1	519
7	10	1	347
8	11	1	666
9	12	1	1882
10	13	1	1111

Table 1: Counts of fruits per individual in *A. pumilis*

Derived II

It is impossible to avoid derived variables in biology.

- heart rate
- density of stomata on a leaf
- frequency of a plant species in a quadrat
- index of diversity
- rate of nucleotide substitution

Derived III

Ratios suffer from statistical problems:

Relative inaccuracy. As you include more separately measured variables in a ratio, their errors can increase multiplicatively (implied error)

Implied error

```
> a <- 1.9
> a

[1] 1.9

> print("implied error of a: 1.85-1.95")
```

```
[1] "implied error of a: 1.85-1.95"
```

```
> b <- 0.9
> b
```

```
[1] 0.9
```

```
> print("implied error of b: 0.85-0.95")
```

```
[1] "implied error of b: 0.85-0.95"
```

Implied error of ratios

```
> a/b
```

```
[1] 2.111111
```

```
> print(paste("implied error of a/b:", (1.85/0.95),
+     "-", (1.95/0.85)))
```

```
[1] "implied error of a/b: 1.94736842105263 - 2.29411764705882"
```

```
> print(paste("midpoint between extremes of a/b",
+     mean(c((1.85/0.95), (1.95/0.85)))))
```

```
[1] "midpoint between extremes of a/b 2.12074303405573"
```

Distribution of ratios

Ratios are often ill-distributed in ways that may violate assumptions of various hypothesis tests.
Because these data are so frequently encountered, transformations have been developed that reduce the effect of non-normal distribution

Methods to look at a single sample

By single sample, we mean data collected on a single variable with no classifying information.

Frequency Dist

The most used initial tool in examining a single sample is the frequency distribution.
A frequency distribution is exactly what it says, a representation of the data in such a way that the number of each type of event is emphasized.

Sexuality	number
Dioecious	147
Non-dioecious	3131

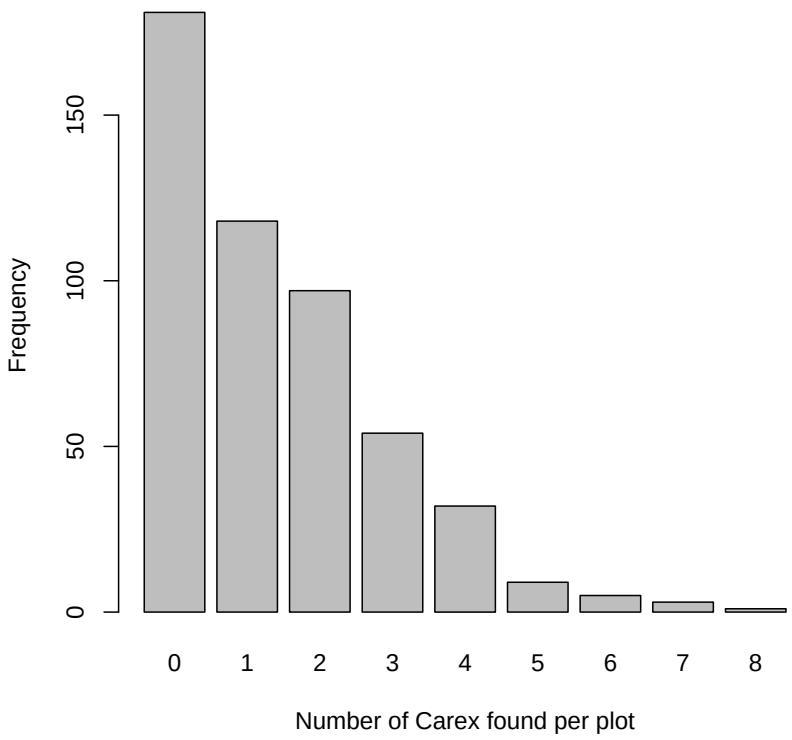
Table 2: Distribution of sexuality in the Carolina Flora

Qualitative frequency distribution

Data from Sokal and Rohlf (1995)

```
> number.per.quadrat <- 0:8
> obs.freq <- c(181, 118, 97, 54, 32, 9, 5, 3, 1)
> carex <- data.frame(number.per.quadrat, obs.freq)
> carex

  number.per.quadrat obs.freq
1                  0      181
2                  1      118
3                  2       97
4                  3       54
5                  4       32
6                  5        9
7                  6        5
8                  7        3
9                  8        1
```

***Carex* distribution**



Quantitative Freq Dist II

Possum data Maindonald and Braun (2003) have an R package that supports their book. This is called DAAG, after the title.
The possum dataset is provided in the DAAG package

Possum data

```
> library(DAAG)
> data(possum)
> names(possum)

 [1] "case"     "site"     "Pop"      "sex"      "age"
 [6] "hdlngth"  "skullw"   "totlngth" "taill"    "footlgth"
[11] "earconch" "eye"      "chest"    "belly"

> length(possum$totlngth)
```

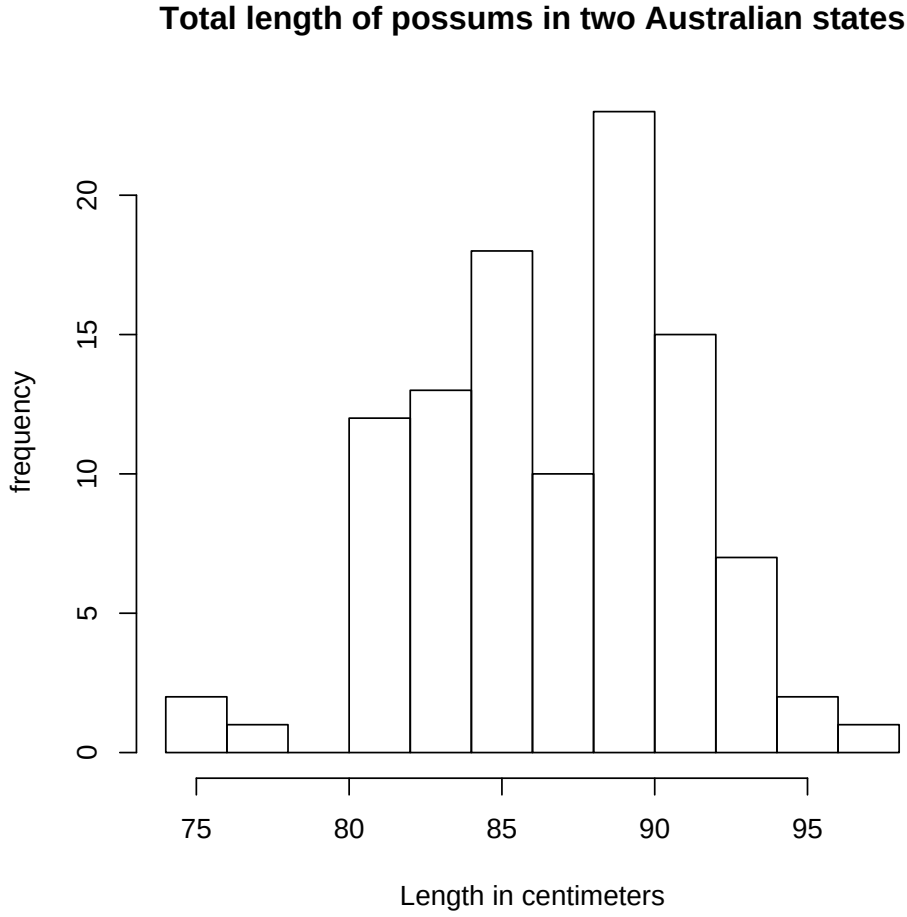
```
[1] 104
```

```
> sort(possum$totlngth)[1:20]

 [1] 75.0 76.0 77.0 80.5 80.5 80.5 81.0 81.0 81.0 81.0 81.5
[12] 81.5 82.0 82.0 82.0 82.5 82.5 82.5 83.0 83.0
```

The sorted data give some idea of the distribution of possum data, but often a graphic method is more useful when looking at continuous data.

Frequency Dist II

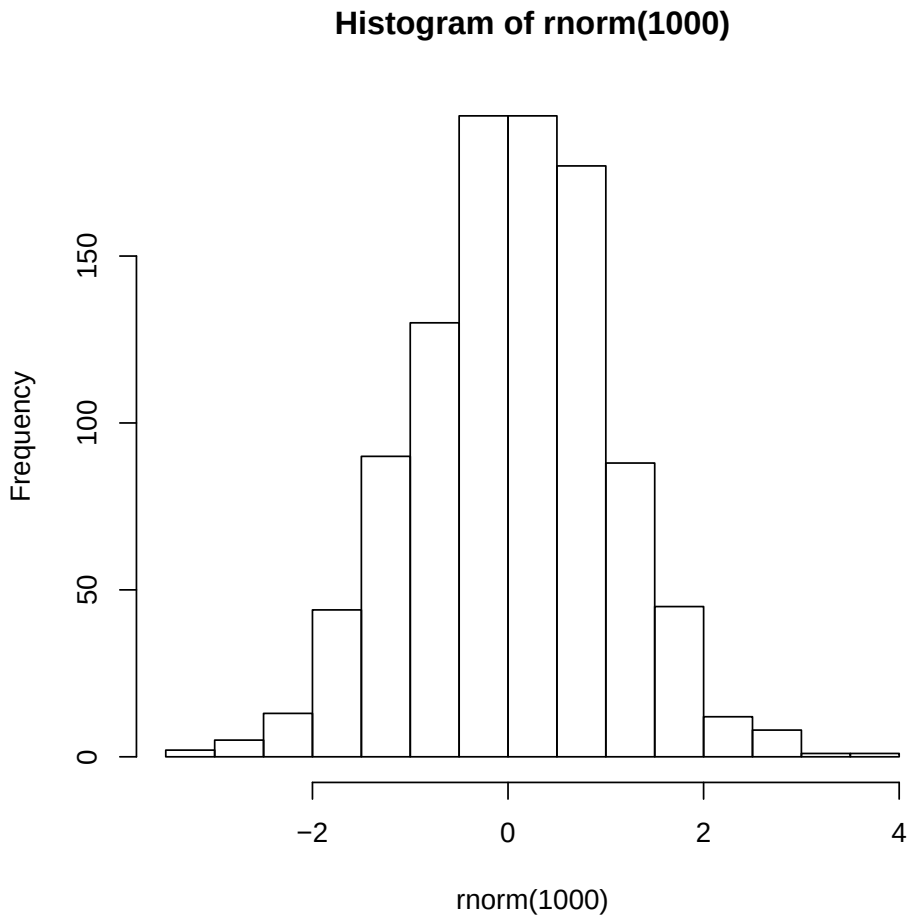
```
> hist(possum$totlngth, main = "Total length of possums in two Australian states",
+     xlab = "Length in centimeters", ylab = "frequency")
```



Normal Dist

Just so you know what a histogram of a more familiar distribution looks like:

```
> hist(rnorm(1000))
```



Hand Generated Freq Distributions

It is possible to take a vector of observations for a variable and quickly get a good feeling of the distribution of the data by a couple of hand methods.
One problem with hand methods is that most data are recorded on computers now and once the data is computer readable, more sophisticated methods are available. Because they have a long history, it is worth understanding how they work and what their results look like.

Discrete distributions

Generally the number of observations in each category are tallied and reported as number of observations per discrete category. In R, `table()` performs this function

```
> q <- c(1, 2, 1, 1, 2, 3, 3, 4, 5, 4, 5, 6, 7,
+     7, 6, 5, 4, 3, 4, 5, 7, 5)
> table(q)

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

> data(iris)
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
```

Continuous distributions

To count members sampled from continuous distributions, the data need to be broken up into some sort of categories. These categories can be defined by the invesitgator and tallies performed. Tukey's stem and leaf plots perform the task of categorizing and tallying in a convenient fashion. Stem and leaf plots categories are determined by placement of decimal points.

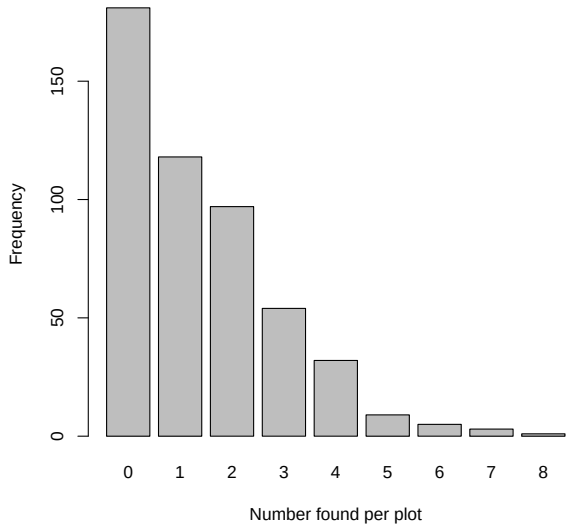
Stem and leaf plot

```
> data(iris)
> stem(iris$Petal.Width[iris$Species == "setosa"])

  The decimal point is 1 digit(s) to the left of the |

  1 | 00000
  2 | 00000000000000000000000000000
  3 | 0000000
  4 | 0000000
  5 | 0
  6 | 0
```

Bar Graphs

The graphical representation of tallies from discrete data usually depends on bar graphs.

```
> barplot(carex[, 2], names.arg = carex[, 1], xlab = "Number found per plot",
+     ylab = "Frequency")
```



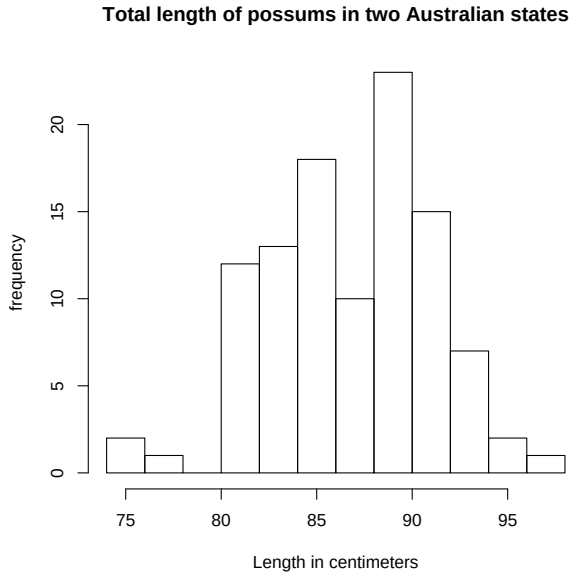
Histogram

A histogram is not a bar graph. In a histogram, the x-axis has quantitative meaning. The height of bars are related to the number of observations found within the x-interval

Frequencies

Most histograms are plotted with the actual counts of individuals on the y-axis.

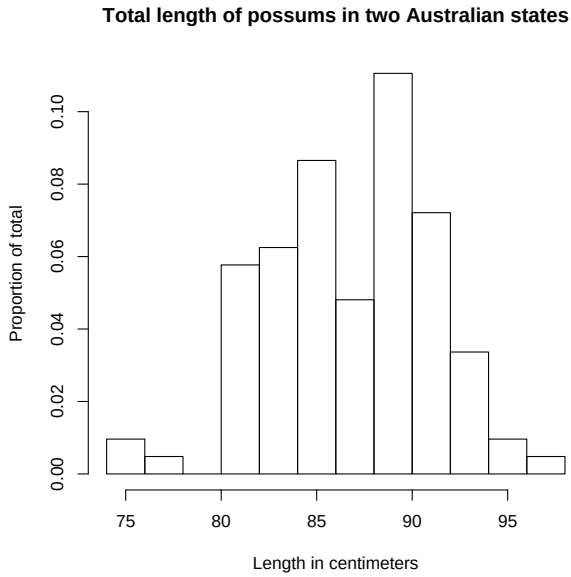
```
> hist(possum$totlngth, main = "Total length of possums in two Australian states",
+     xlab = "Length in centimeters", ylab = "frequency")
```



Histograms

Some histograms provide the proportion or density of the total observations within an x-axis interval.

```
> hist(possum$totlngth, main = "Total length of possums in two Australian states",
+     xlab = "Length in centimeters", ylab = "Proportion of total",
+     freq = F)
```

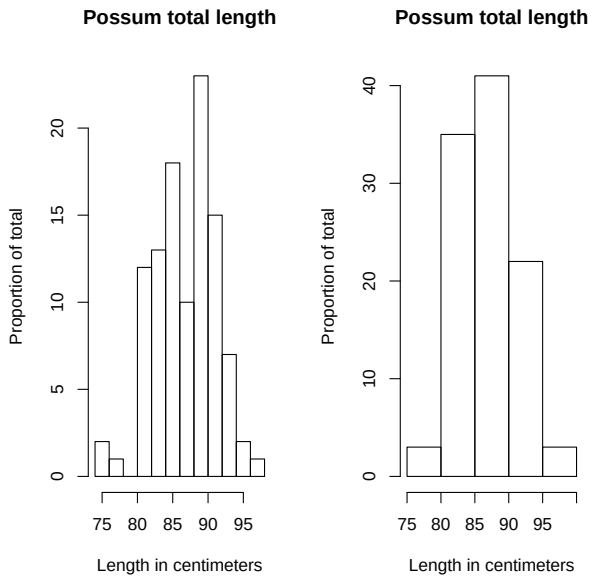


Determining breakpoints

This is a difficult subject. Changing the breakpoints can provide very different views of the data. Often, the range of the data is broken into x different equally spaced *bins* or categories.

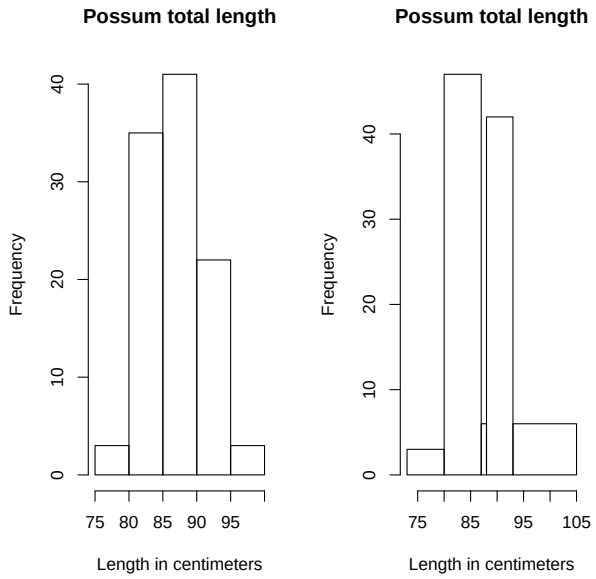
Numbers of bins

Choosing different numbers of categories can have a large effect



Irregularly spaced bins

So can choosing irregularly shaped intervals:



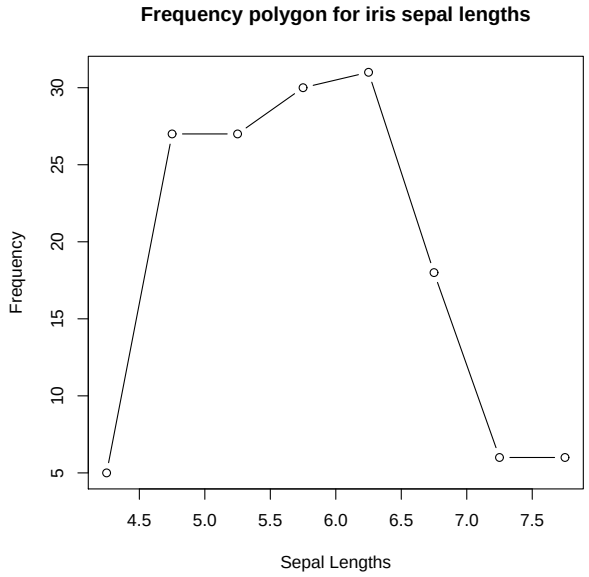
Interval spacing Disadvantage

Irregularly spaced intervals can seriously bias the way that data are visually interpreted. Make sure you have a good justification before using irregular-sized bins

Frequency Polygons

Frequency polygons try to give a continuous appearance to histograms. R does not provide a single command to illustrate these. It takes 2 commands.

```
> h <- hist(iris$Sepal.Length, plot = F)
> plot(h$mids, h$counts, type = "b", main = "Frequency polygon for iris sepal lengths"
+     ylab = "Frequency", xlab = "Sepal Lengths")
```



Summary statistics

Summary statistics summarize the information contained in a sample. This has the effect of throwing out information, yet at the same time it may make it easier to make *general statements* about the sample.

Statistics of location

These are also known as measures of *central tendency*. Such statistics are intended to place a sample along a particular quantitative dimension or axis.

Arithmetic mean

$$\bar{Y} = \frac{\Sigma_{i=1}^{n} Y_i}{n} \tag{1}$$

Where the Y_i's are actual observations and n is the number of observations

Alternative notations

This notation is useful for discrete data:

$$\bar{Y} = \frac{1}{n} \Sigma_{i=1}^{n} f_i * q_i \tag{2}$$

Where the f_i's are counts and the q_i are values for the different categories
A weighted mean can also be defined:

$$\bar{Y} = \frac{\Sigma_{i=1}^{n} w_i Y_i}{\Sigma_{i=1}^{n} w_i}$$

Where the w_i's are weights. When all w_i are equal to 1, this reduces to equation (1).

Geometric mean

The geometric mean is occasionally used in statistics:

$$GM_Y = (\prod_{i=1}^{n} Y_i)^{1/n} \tag{3}$$

Harmonic mean A harmonic mean is used in population biology and physics. It is the reciprocal of the arithmetic mean of the reciprocals

$$\frac{1}{H_Y} = \frac{1}{n}\Sigma\frac{1}{Y} \tag{4}$$

Comparison of means

```
> sum(possum$totlngth)/length(possum$totlngth)

[1] 87.08846

> mean(possum$totlngth)

[1] 87.08846

> prod(possum$totlngth)^(1/length(possum$totlngth))

[1] 86.98152

> exp(sum(log(possum$totlngth))/length(possum$totlngth))

[1] 86.98152

> 1/(mean(1/possum$totlngth))

[1] 86.87323
```

median

The median is the value halfway along the list of sorted observations. It is also known as the 50th percentile and corresponds to the value which is greater than 50% of observations. A stem and leaf plot makes it relatively easy to find the median. In general, order the observations and the observation found in the middle for odd n and the average of the two middle observations for even n is the median.

```
> mean(possum$totlngth[possum$sex == "f"])

[1] 87.90698

> median(possum$totlngth[possum$sex == "f"])

[1] 88.5

> stem(possum$totlngth[possum$sex == "f"])

  The decimal point is at the |

  74 | 0
  76 |
  78 |
  80 | 05
  82 | 0500
  84 | 05005
  86 | 05505
  88 | 0005500005555
  90 | 5550055
  92 | 000
  94 | 05
  96 | 5
```

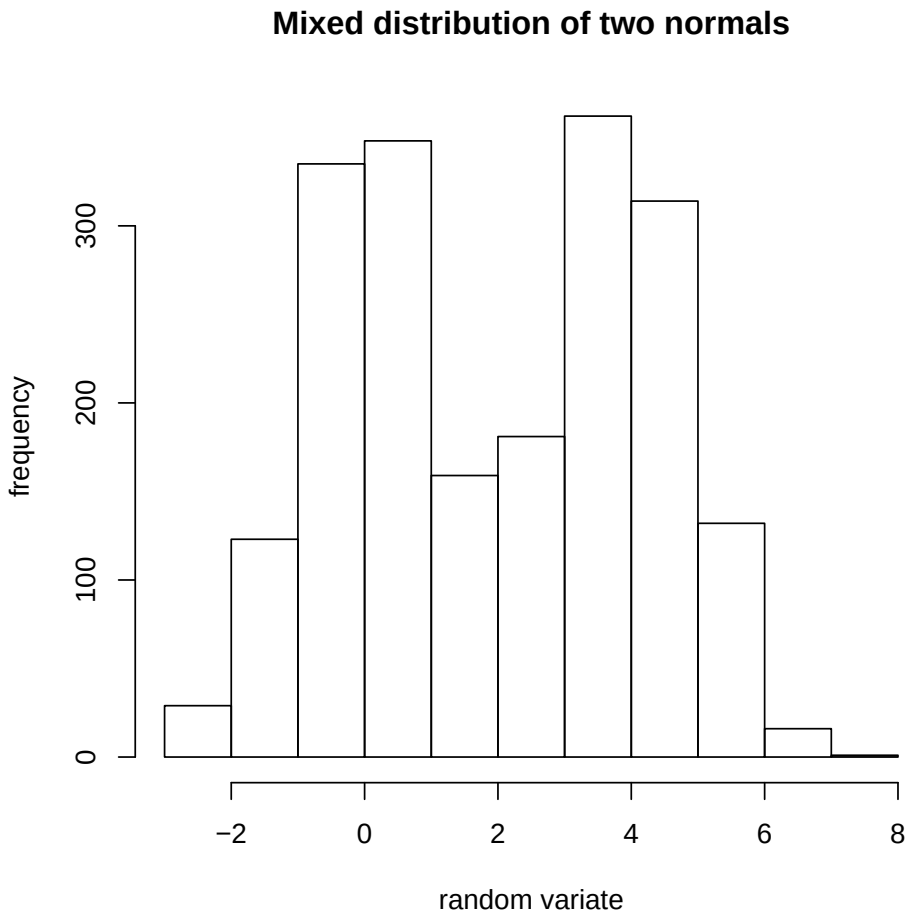
mode

The mode is a concept that emerges naturally from examination of frequency distributions. It is the most common value of a discrete distribution.
Modes aren't well defined on a continuous scale because no two observations take the same value, but the most 'full' bin in a histogram is often termed the mode.

Multimodality

When there are more than one mode, the distribution is considered bi- or multi-modal.

```
> hist(c(rnorm(1000, mean = 0), rnorm(1000, mean = 4)),
+     xlab = "random variate", ylab = "frequency",
+     main = "Mixed distribution of two normals")
```



Statistics of dispersion

Knowing where the sample sits on an axis says nothing about the distributional shape of the sample. Measurements of dispersion provide this insight.

Range

Range is the difference between the highest and lowest values for a set of observations. Two problems with the range:

- influenced strongly by outliers (they determine the range)
- influenced strongly by sample size (always stays same or increases)

Interquartile range

One improvement upon the range is the distance between the 25th and 75th percentiles. This is much less sensitive to outliers and sample sizes and with computers it is easy to calculate
In R the quantiles can be calculated using `quantile()`

```
> quantile(possum$totlngth, 0.25)

25%
 84

> quantile(possum$totlngth, 0.75)

75%
 90
```

Range versus interquartile range

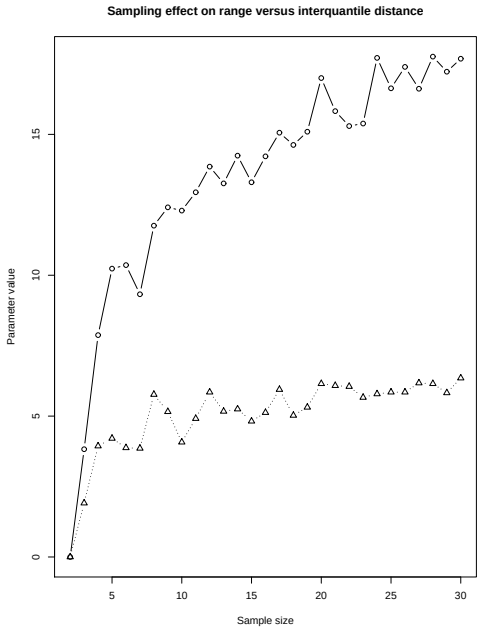


Figure 1: Figure showing the relationship between sample size and measures of dispersion. The data are body lengths from the possum dataframe. Circles represent ranges and triangles represent interquartile distances

Variance

The standard measure for dispersion used in statistics is *variance* (and its square-root, standard deviation). The variance of a population of observations is the average squared deviation of elements from the mean. Based on that verbal description, it is pretty easy to define variance:

$$\sigma^2 = \frac{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}{n} \tag{5}$$

Computational formula

Usually the following formula is given as a computational formula

$$\sigma^2 = \frac{\Sigma_{i=1}^{n}Y_i^2 - n\bar{Y}^2}{n} \tag{6}$$

Difference in expectations

Another measure of variance is the difference in the expectation for the squared values in the population and the expectation for the values squared. Because an expected value is a mean:

```
> sigma <- sum((mean(possum$totlngth) - possum$totlngth)^2)/length(possum$totlngth)
> sigma
[1] 18.40217
> sigma2 <- mean(possum$totlngth^2) - mean(possum$totlngth)^2
> sigma2
[1] 18.40217
```

This should make sense given the computational formula.

Dependence of Std Dev on sample size

Figure 1 illustrates some of the problems with ranges versus interquantile distance with respect to sample size effects. Figure 2 does the same analysis but also includes the square root of variance, standard deviation (σ).

CV

There is a built-in relationship between variance and the scale used to measure a variable. There is also often a relationship between mean and variance in biological samples.
The *coefficient of variation* (CV) scales measures of variance by the mean. Specifically, the standard deviation is divided by the mean and multiplied by 100. Therefore, CV always exists on a scale of 0-100 and may in some instances be used to compare measures of variance for two samples or variables.

$$CV = 100 \times \frac{s}{\bar{Y}} \tag{7}$$

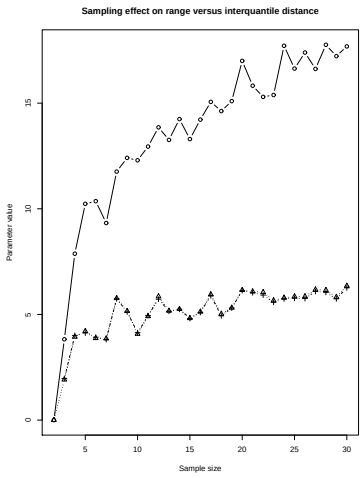
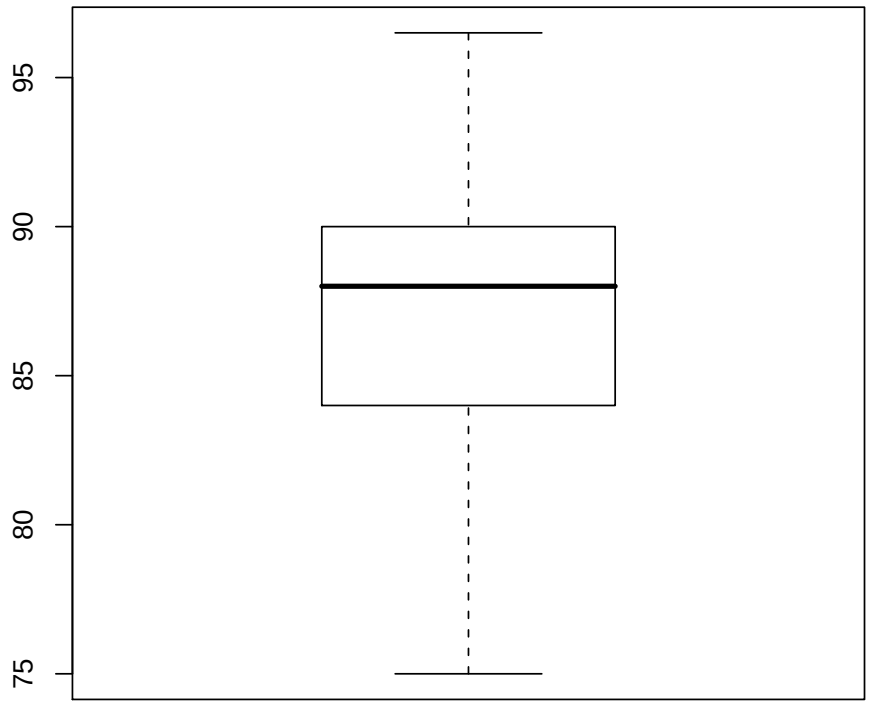


Figure 2: Figure showing the relationship between sample size and measures of dispersion. The data are body lengths from the possum dataframe. Circles represent ranges, triangles represent interquartile distances, and plus symbols represent the standard deviation

Boxplots

One graphical method to characterize the distribution of a set of data is the Box and Whisker plot devised by Tukey. These plots focus on the median, plot hinges (essentially the interquartile distance) and a truncated range: R's base package implements `boxplot()` to produce these graphs.

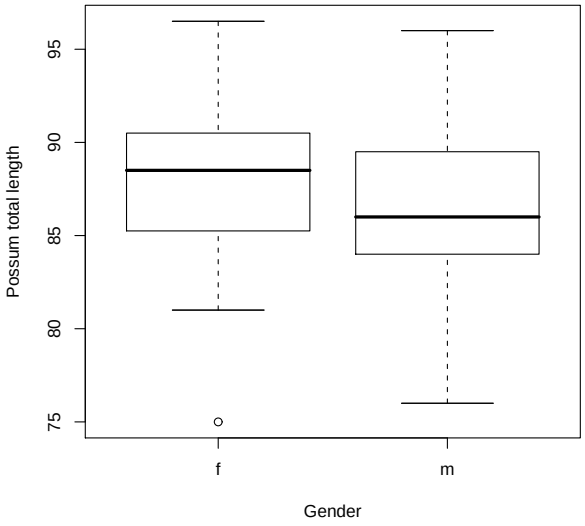
```
> boxplot(possum$totlngth)
```



Box plots on subsets of data

Note that you can look at the effect of another discrete variable (like sex) using the formula interface:

```
> boxplot(totlngth ~ sex, ylab = "Possum total length",
+     xlab = "Gender", data = possum)
```



Parameters versus sample statistics

When introducing the variance, we were not very specific and formal in the language. Populations include all realizations of a particular variable. Samples are a subset of the observations in a population. The sample mean and variance do a good job of describing the sample.
Most people don't want to characterize the sample as much as use it to sample to infer something about the population.

Population parameters

There is a population mean and variance. These are essentially unknowable, and we use sample statistics to estimate these quantities. In general the parametric values are denoted by Greek letters and sample statistics, by Roman letters. For example the population mean is μ and the sample mean is $\bar{Y}$. The population variance is σ^2 whereas the sample variance is s^2. So formally the population variance is:

$$\sigma^2 = \frac{\Sigma_{i=1}^{n}(Y_i - \mu)^2}{n} \tag{8}$$

Where the Y_i's are **all** observations in a population and μ is the parametric mean.

Sample Variance

Equation (9) gives the correct form to estimate the population variance from a sample. This is known as the *sample variance*

$$s^2 = \frac{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} \tag{9}$$

The $n-1$ mentioned above is know as the *degrees of freedom* .

Degrees of freedom

Sokal and Rohlf (1995) say that "statisticians have found that s^2 underestimates σ^2 unless degrees of freedom are used as a denominator"
That's true, but another way of thinking about it is that to calculate the variance, you are actually estimating two quantities, the population mean and the populations variance. The estimate of the variance depends upon the estimate of the mean. In general, degrees of freedom equal the sample size minus the number of additional parameters you are trying to estimate.

References

Maindonald, J. and Braun, J. (2003). *Data Analysis and Graphics Using R*, Cambridge Univ. Press.

Sokal, R. R. and Rohlf, F. J. (1995). *Biometry. The principles and practice of statistics in biological research*, third edn, W. H. Freeman, San Francisco, California.