

Data Summary

STAT 211 - 509

2018/09/11

Descriptive statistics

- We have a sample of data, drawn from some distribution
- How to compute numerical summaries of the data?
- How to visualize the data?

Variables

- **Variable:** any characteristic or quantity to be measured on units in a study
- **Categorical variable:** places a unit into one of several categories
- **Quantitative variable:** takes on numerical values
- **Univariate:** data with one variable
- **Bivariate:** data with two variables
- **Multivariate:** data with three or more variables

Example: US cereal

```
dat <- MASS::UScereal
```

```
str(MASS::UScereal)
```

```
## 'data.frame':   65 obs. of  11 variables:
## $ mfr       : Factor w/ 6 levels "G","K","N","P",...: 3 2 2 1 2 1 6 4 5 1 ...
## $ calories  : num  212 212 100 147 110 ...
## $ protein   : num  12.12 12.12 8 2.67 2 ...
## $ fat       : num  3.03 3.03 0 2.67 0 ...
## $ sodium    : num  394 788 280 240 125 ...
## $ fibre     : num  30.3 27.3 28 2 1 ...
## $ carbo     : num  15.2 21.2 16 14 11 ...
## $ sugars    : num  18.2 15.2 0 13.3 14 ...
## $ shelf     : int   3 3 3 1 2 3 1 3 2 1 ...
## $ potassium : num  848.5 969.7 660 93.3 30 ...
## $ vitamins  : Factor w/ 3 levels "100%","enriched",...: 2 2 2 2 2 2 2 2 2 2 ...
```

Summarizing categorical variable

- **Frequency:** number of times a value occurs in data
- **Relative frequency:** proportion of data that has a value

```
freqs <- table(dat$mfr)
freqs

##
##  G  K  N  P  Q  R
## 22 21  3  9  5  5

props <- freqs/nrow(dat)
props

##
##          G          K          N          P
## 0.33846154 0.32307692 0.04615385 0.13846154
##          Q          R
## 0.07692308 0.07692308

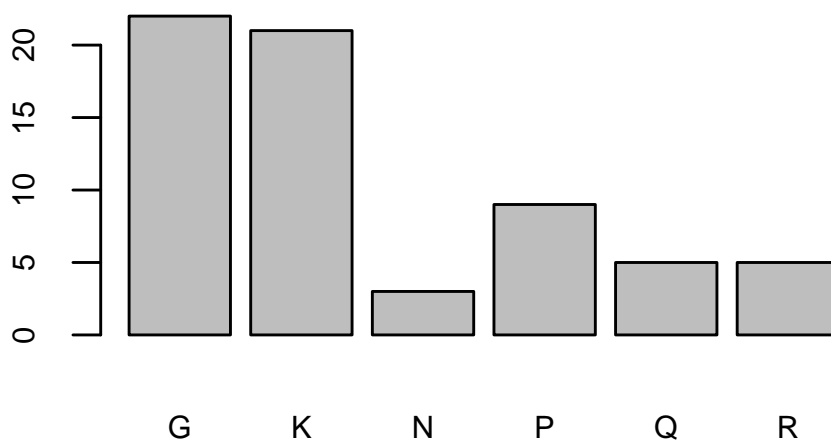
sum(props)

## [1] 1
```

Bar chart

- Compares frequencies
- Unordered

```
barplot(freqs, cex.axis = 0.7, cex.lab = 0.7,
        cex = 0.7)
```



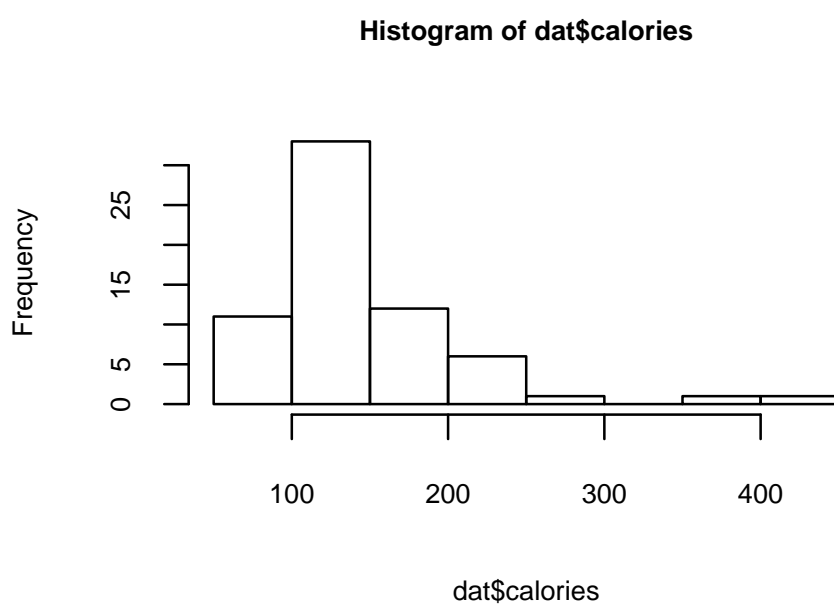
Summarizing quantitative variable

- What is the typical value of the variable?
- What is the spread of the variable?

Histogram

- **Histogram:** bar graph of binned data where height of bar above each bin denotes frequency or relative frequency of values in the bin

```
hist(dat$calories, cex.axis = 0.7, cex.lab = 0.7,
      cex.main = 0.7)
```



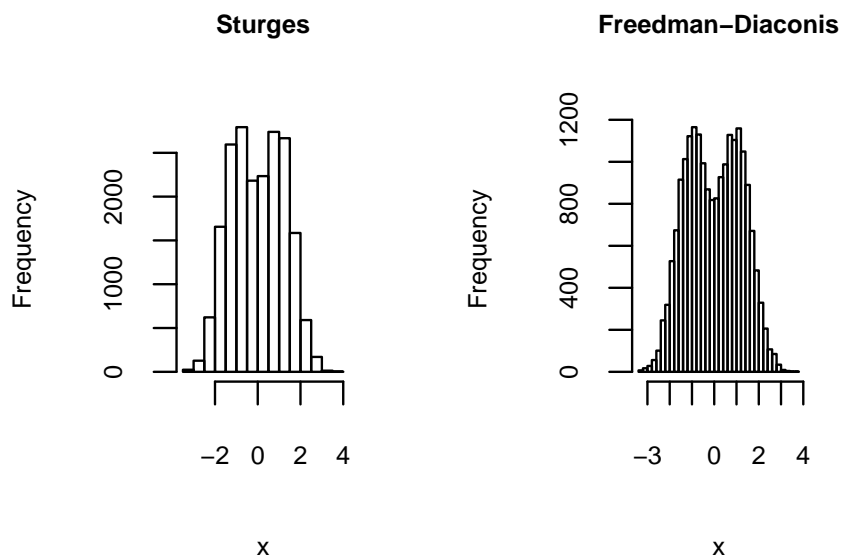
- Need to choose number of bins among which we divide the n data points
- General rule: number of bins $\approx \sqrt{n}$
- `breaks` argument in `hist()`. Can be a string that specifies a built in algorithm for binning. Good default is "FD", for Freedman-Diaconis rule

Breaks example

- Data drawn from a distribution with two modes

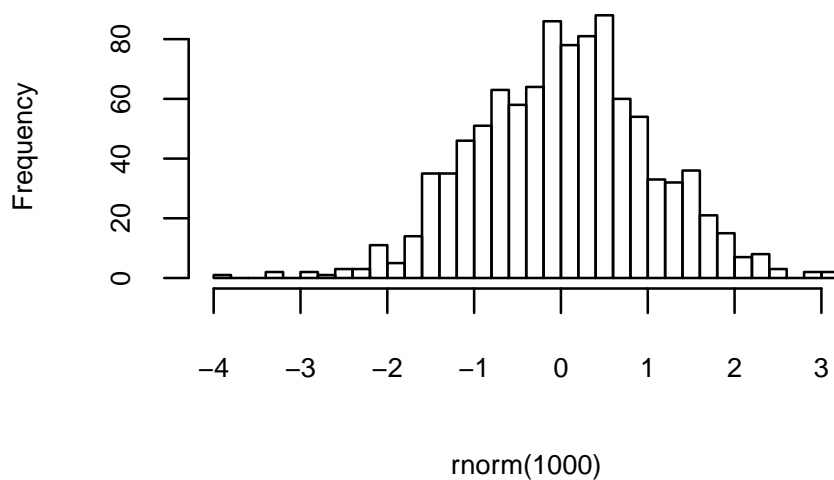
```
set.seed(1)
z <- rbinom(20000, 1, 0.5) + 1
means <- c(-1, 1)
```

```
x <- rnorm(20000, mean = means[z], 0.7)
par(mfrow = c(1, 2))
hist(x, main = "Sturges", cex.main = 0.7, cex.axis = 0.7,
     cex.lab = 0.7)
hist(x, breaks = "FD", main = "Freedman-Diaconis",
     cex.main = 0.7, cex.axis = 0.7, cex.lab = 0.7)
```



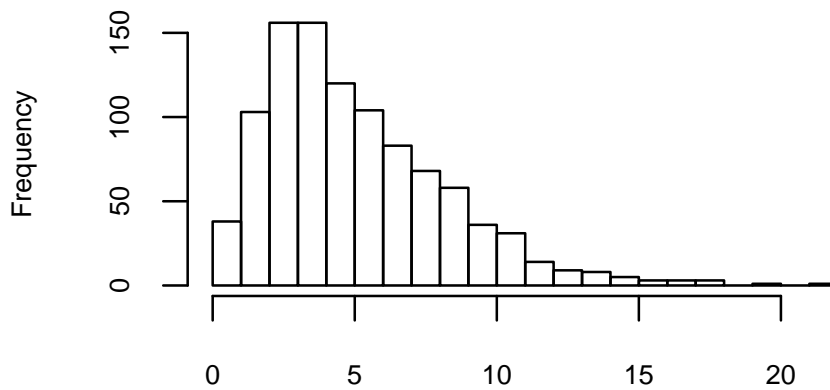
The shape of quantitative data

- **Symmetric** data is mirrored about each side of a center value



- **Skewed** data has one side much longer than the other

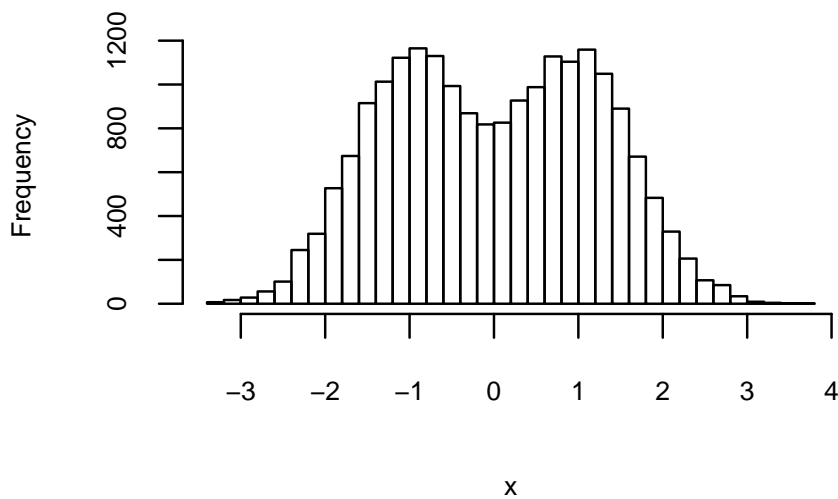
Right skewed data



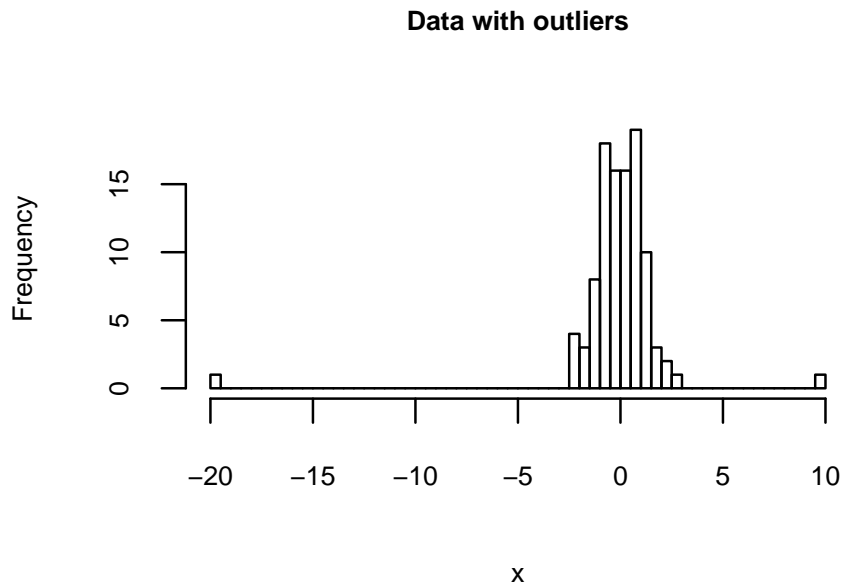
`rchisq(1000, df = 5)`

- The **mode** is the peak value of the distribution
- **Multimodal** data has multiple modes

Bimodal data



- **Outliers** are data points “far” from most other data
- Determination of outliers is subjective
- *Do not* remove outliers if you don’t know for sure that the data is erroneous



Summary statistics for quantitative data

Measures of central tendency

- **Sample median:** value separating lower 50% of data from upper 50% of sample
 - For finite set of numbers, the middle value
 - If even number of values, then mean of middle two numbers
- **Sample mean:** Given sample values x_1, \dots, x_n ,

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

```
mean(dat$calories)
```

```
## [1] 149.4083
```

Percentiles

- **Percentile:** the p th percentile is the value such that $p \times 100\%$ of sample data is below it and $(1 - p) \times 100\%$ are above it.
- **First quartile (Q1)** is 25th percentile
- **Second quartile (Q2)** is 50th percentile
- **Third quartile (Q3)** is 75th percentile
- **Five-number summary**

```
fivenum(dat$calories)
```

```
## [1]  50.0000 110.0000 134.3284 179.1045
```

```
## [5] 440.0000
```

```
summary(dat$calories)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.
```

```
##      50.0   110.0   134.3   149.4   179.1
```

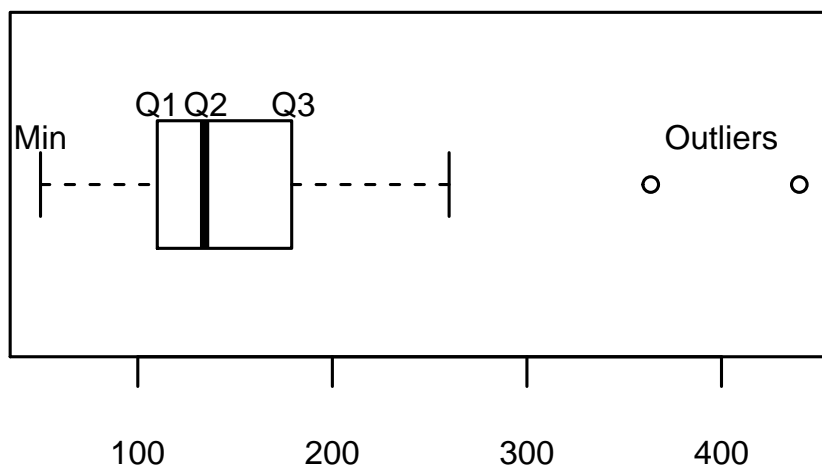
```
##      Max.
```

```
##     440.0
```

Boxplot

- Visualize the 5 number summary
- In R: `boxplot()`
- **Interquartile range:** $IQR = Q3 - Q1$
- **Outliers:** values greater than $Q3 + IQR$ or less than $Q1 - IQR$ are represented with a point

Calories data



Measures of spread

- **IQR:** $Q3 - Q1$, the range of the middle 50% of the data
- **Sample variance, s^2 :** sum of squared deviations from the mean divided by $n - 1$:

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

- **Sample standard deviation, s:** square root of sample variance.

Has same units as data

```
var(dat$calories)
```

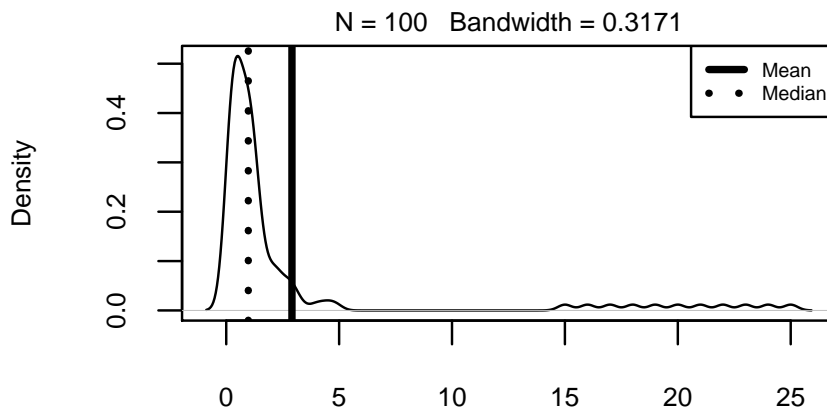
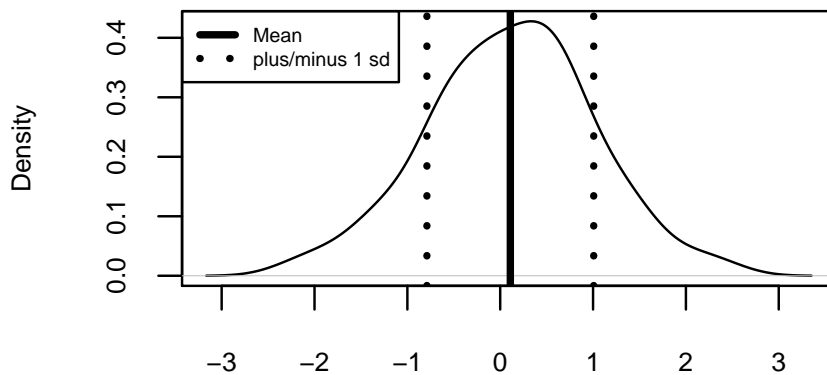
```
## [1] 3895.242
```

```
sd(dat$calories)
```

```
## [1] 62.41187
```

Choosing measure of central tendency and spread

- Sample mean and sample standard deviation good for symmetric data
- For skewed data or data with outliers, sample median and interquartile range may be more appropriate



N = 111 Bandwidth = 0.307