Introduction to Statistics

Paul M. Magwene

What is statistics?

American Statistical Association

"Statistics is the science of learning from data and of measuring, controlling, and communicating uncertainty."

Whitlock & Schluter (W&S)

"The study of methods to describe and measure aspects of nature from samples."

Statistics involves...

- Exploration – discovering or highlighting trends and patterns using data summarization and visualization
- Description – describing properties of data, such as location, spread, association, using numerical functions
- Estimation – "the process of inferring an unknown quantity of a *population* using *sample data*" (W&S)
- Quantifying uncertainty – determining the magnitude by which estimates of unknown quantities may differ from true values
- Hypothesis testing – the evaluation of "a specific claim regarding a population parameter" based on estimation of parameters and uncertainty from samples

Terminology

Population

The set of "things" we want to study or learn about. Can be concrete (e.g. males over 20 in the United States; brushtail possums in the state of Victoria, Australia), or abstract (e.g. corn plants grown from Monsanto "round up ready" seed; yeast cells synchronized with alpha-pheromone)

Variables

Measureable properties of the "things" (entities) we want to study. Weight, age, expression, sex, abundance, etc. Even things that we think of as "constants" (e.g. speed of light) can vary, as a function of our measurement instruments.

Observation

A discrete entity or thing we have made measurements on. Individuals, genotypes, species, strains, geographic regions, niches, etc.

Sample

A collection of observations for which we have measured one or more variables.

Parameters and statistics

Parameter

A *parameter* is numerical quantity of interest that describes one or more variables in a population.

Statistic (estimate)

A *statistic* is a related numerical quantity calculated by applying a function (algorithm) to a sample

Types of variables

- Categorical or Nominal – labels matter but no mathematical notion of order or distance
  - Sex: Male / Female
  - Species
- Ordinal data – order matters but no distance metric
  - Juvenile, Adult
  - Small, Medium, Large
  - Muddy, Sandy, Gravelly
- Discrete, Integer, Counting
  - Number of vertebrae in a snake
  - Number of pine trees in a specified area
  - Number of heart beats in a minute
  - Number of head bobs during courtship display
- Continuous
  - Body mass
  - Length of right femur
  - Duration of aggressive display

Frequency Distributions

Frequency

Definitions from W&S:
The *frequency* of a measurement is the number of observations in a sample having a that particular value of the measurement

- Relative frequency – proportion of observations having a given measurement (frequency of a measurement/total number of observations)

Frequency distribution

The *frequency distribution* of a variable is the number of times each value of a variable occurs in a sample
-" A *relative frequency distribution* describes the fraction of occurences of each value of a variable" (W&S)

Categorical, ordinal, and discrete numerical variables

- Counting observations with particular values is conceptually straightforward

Continuous variables

- Counts not of particular values, but ranges of values

Tabular representations of frequency distributions: Categorical and ordinal variables

Frequency distribution of sex of babies in NC Births data set:

```
births %>% count(sexBaby)
# A tibble: 2 x 2
  sexBaby     n
  <chr>   <int>
1 female     68
2 male       82
```

Frequency distribution of passenger classes on the Titanic:

```
titanic %>% count(pclass)
# A tibble: 3 x 2
  pclass     n
   <int> <int>
1      1   323
2      2   277
3      3   709
```

Graphical representations of frequency distributions: Bar charts for categorical and ordinal variables

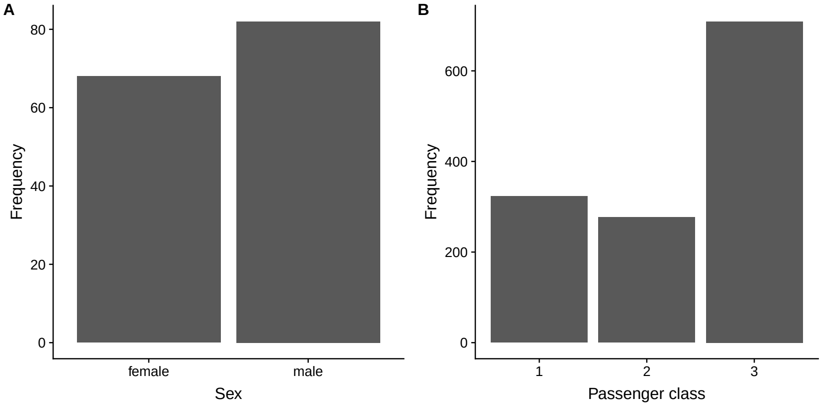


Figure 1: A) Frequency distribution of baby sex, NC births tudy; B) Frequency distribution of passenger class on the Titanic

Bar charts are visually "heavy" when there are many categories or discrete values

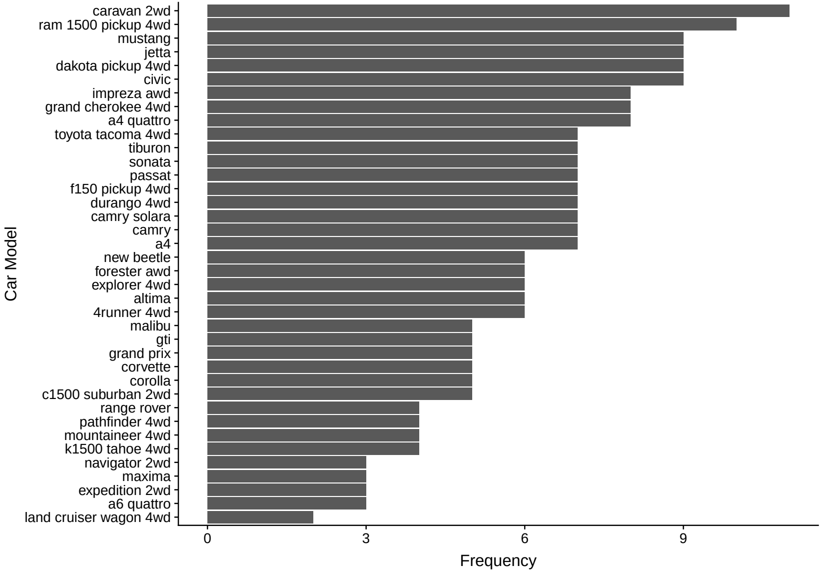


Figure 2: Frequency distribution of car models in the mpg data set

“Lollipop” charts are an aternative to bar charts for discrete frequency distributions

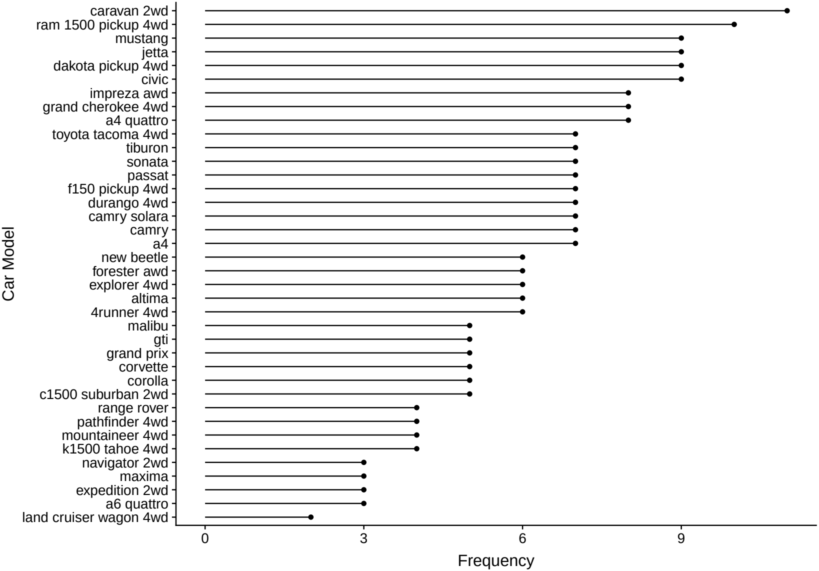


Figure 3: Frequency distribution of car models in the mpg data set

“Lollipop” charts are also appropriate for discrete numerical data with a modest number of values

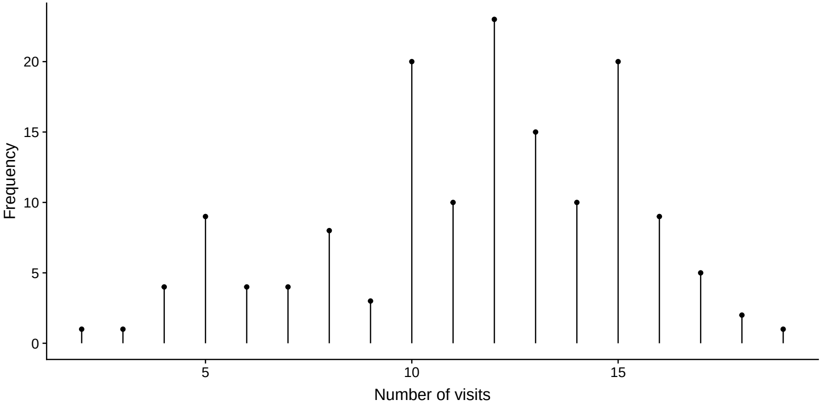
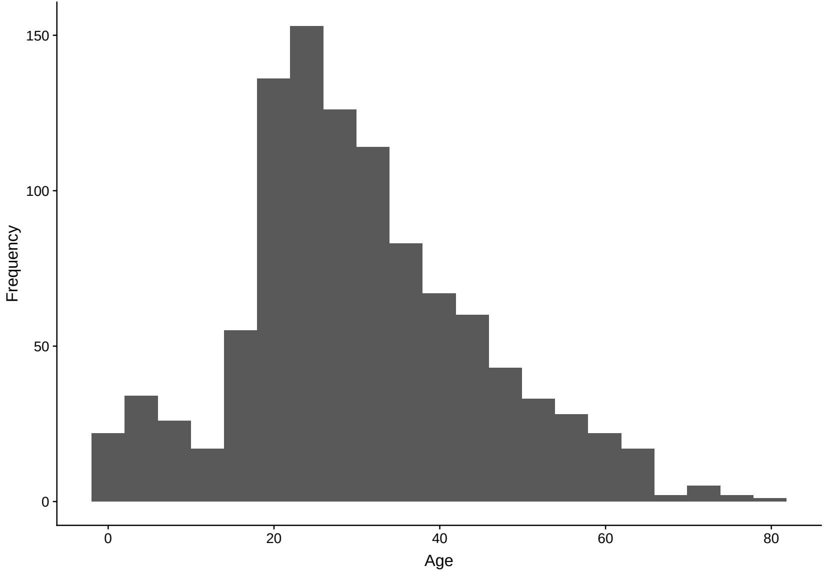


Figure 4: Frequency distribution for the number of prenatal visits in the NC Births data

Graphical representations of frequency distributions: Histograms for continuous variables or discrete numerical variables with many values



Descriptive statistics for frequency distributions

Measures of location or central tendency

Three most common statistics to describe the "central tendency" of a variable are:

1. Mean
2. Median
3. Mode

Mean

The arithmetic mean of a variable.

R function

- `mean()`

Algorithm

- Sum up all the observations for the variable of interest, and divide the sum by the total number of observations

Mathematical notation

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

Mathematical notation: Sums

- The mathematical notation for summing is represented by the Greek symbol sigma ($\sum$)
- The sum symbol is short-hand for a "for-loop" where each time through the loop you evaluate the statement to the right of the $\sum$, and accumulate the values
- Example:

$$\sum_{i=1}^{5} i$$

- Notice that there are two numbers – one above and one below the $\sum$. These are the upper and lower indices of the for loop.
- In this example we're simply adding up the numbers from 1 to 5.

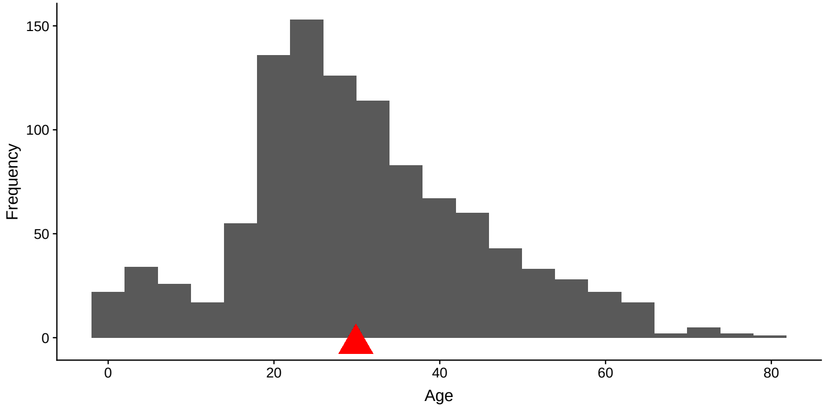
Mathematical notation: Sums, cont

- The indices of the sum notation can be used to index a variable.

$$\sum_{i=1}^{5} x_i$$

- x_i is how mathematicians typically represent indexing of a variable, whereas in R we'd write `x[i]` instead.
- This sum says "add up the first 5 values of the variable x".

Mean illustrated



Median

The middle value of a set of observations.

R function:

- `median()`

Algorithm:

- Sort all the observations from smallest to largest. If an odd number of observations, take the middle value. If an even number of observations, take the mean of the two middle values.

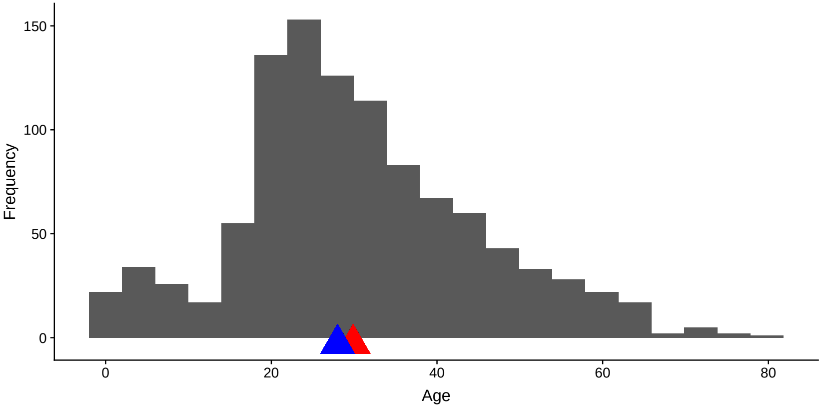
Mathematical notation:

- No simple notation

Notes:

- The median is a *robust* estimator of location. Robust statistics are those that are not strongly affected by outliers our violations of model assumptions.
- Mean and median are usually very similar for symmetrical distributions without outliers.

Median illustrated



Robustness of median: Example

There are five people in a bar. Their 20126 incomes are: \$42,000, \$60,000, \$75,000, \$80,000, and \$84,000.

```
income <- c(42000, 60000, 75000, 80000, 84000)
median(income)
[1] 75000
mean(income)
[1] 68200
```

Now Warren Buffet walks into the bar. His income in 2016 was ~\$12.7 billion dollars.

```
income <- c(42000, 60000, 75000, 80000, 84000, 1270000000)
median(income)
[1] 77500
mean(income)
[1] 211723500
```

Mode

The most common value of a variable.

R function:

- use `tally()` or `count()`for categorical or ordinal variables
- no built-in in algorithm for continuous variables
- **NOT** `mode()` which gives the storage mode of an object

Algorithm:

- For each discrete value, count the number of observations with that value. The value with the most observations is the mode.

Mathematical notation:

- No simple notation

Spread / Variation

The most common statistics to describe the spread or variation of a variable of interest are:

1. Range
2. Inter-quartile range
3. Variance and standard deviation

Range

The difference between the largest and smallest value of a set of observations.

R function:

- `range()` returns min and max. So to get the range as we define it here, you can do `diff(range(x))` [Lookup what `diff` does!]

Algorithm:

- Find the largest and smallest observations. Substract the smallest value from the largest value.

Mathematical notation:

$$\max(x) - \min(x)$$

Quantiles, quartiles, interquartile range

- **Quantiles** – points that will divide a frequency distribution into equal sized groups
  - quartiles – points dividing a distribution into 4 equal groups
  - Median is the 2nd quartile
  - deciles – points dividing a distribution into 10 equal groups
  - percentiles – points dividing a distribution into 100 equal groups
- **Interquartile range (IQR)** – range of values that captures the central 50% of the distribution
  - Q1 = lower quartile (25th percentile), Q3 = upper quartile (75th percentile)

Boxplots revisited

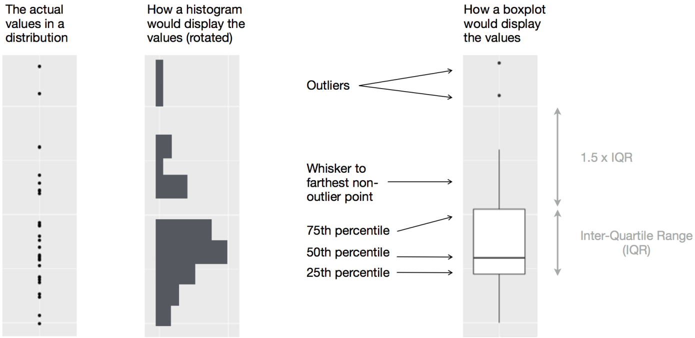
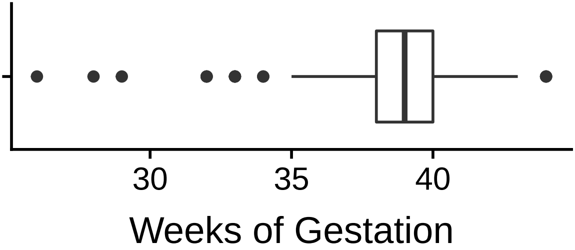


Figure 5: Features of a boxplot and their relationship to the quartiles and IQR

The R summary function

The summary() function provides a quick mechanism for generating some key statistics of location and spread.

```
births %$% summary(weeks)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  26.00   38.00   39.00   38.55   40.00   44.00 
```



Variance and standard deviation

Deviate – the difference between an observation and the mean; can be negative or positive. Units same as the x_i.

$$x_i - \overline{x}$$

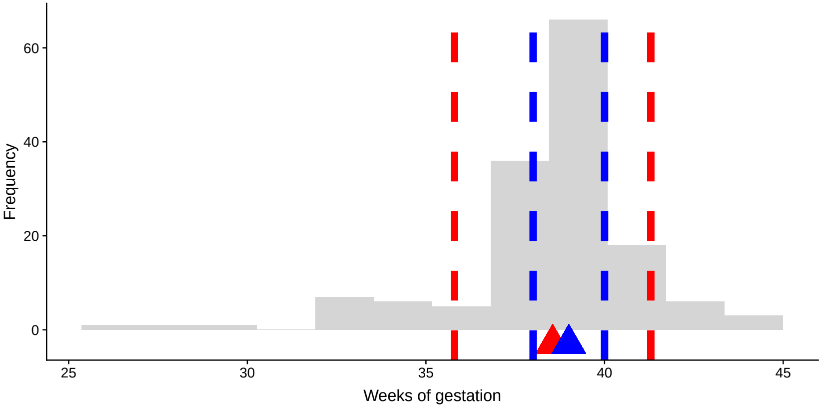
Variance – the mean (approximately) squared deviation (units^2).

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

Standard deviation – the square root of the variance (units same as the x_i).

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Visualization of standard deviation and interquartile range



Coefficient of variation

- Standard deviation expressed as percentage of mean
- Unitless measure that is ueful for comparing spread of data measured in different units or magnitudes
- Only works for variables with positive means

$$CV_x = \frac{s_x \times 100}{\overline{X}}$$

Skewness

- Skewness describes asymmetry of distributions

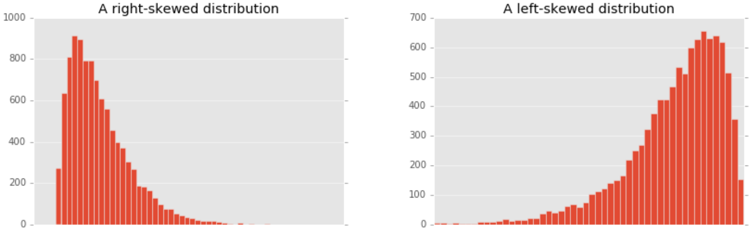


Figure 6: Skewed distributions

Cumulative frequency and cumulative frequency distribution

- "*Cumulative relative frequency* at a given measurement is the fraction of observations less than or equal to that measurement" (W&S)

- A *cumulative frequency distribution* is a graph displaying the cumulative relative frequency over the range of a variable