Variables and Descriptive Statistics

Ling250: Data Science for Linguistics

C.M. Downey

Spring 2025

Variable types and scales

Variables

Variables

- A **variable** is some **measurable attribute** of data
  - Sometimes comes from **instrumentation** or **observation** (e.g. vowel formants are measured with a microphone and audio software)
  - Sometimes comes from **human responses** (e.g. a survey asking "do you agree or disagree that this is a useful course?")

Variables

- A **variable** is some **measurable attribute** of data
  - Sometimes comes from **instrumentation** or **observation** (e.g. vowel formants are measured with a microphone and audio software)
  - Sometimes comes from **human responses** (e.g. a survey asking "do you agree or disagree that this is a useful course?")
- Variables are usually stored in **columns** of data frames / CSVs

Variables

- A **variable** is some **measurable attribute** of data
  - Sometimes comes from **instrumentation** or **observation** (e.g. vowel formants are measured with a microphone and audio software)
  - Sometimes comes from **human responses** (e.g. a survey asking "do you agree or disagree that this is a useful course?")
- Variables are usually stored in **columns** of data frames / CSVs
- Variables are **not** all on the **same scale** of measurement
  - Can't directly compare birth year, eye color, and vowel formants

Scales of measurement

Scales of measurement

- Our textbook (Learning Statistics with R) defines **four scales of measurement**:

Scales of measurement

- Our textbook (Learning Statistics with R) defines **four scales of measurement**:
  - **Nominal**: no particular ordering or relationship between values

Scales of measurement

- Our textbook (Learning Statistics with R) defines **four scales of measurement**:
  - **Nominal**: no particular ordering or relationship between values
  - **Ordinal**: values can be ordered, but differences between them is not meaningful

Scales of measurement

- Our textbook (Learning Statistics with R) defines **four scales of measurement**:
  - **Nominal**: no particular ordering or relationship between values
  - **Ordinal**: values can be ordered, but differences between them is not meaningful
  - **Interval**: values are ordered and differences are meaningful, but not ratios

Scales of measurement

- Our textbook (Learning Statistics with R) defines **four scales of measurement**:
  - **Nominal**: no particular ordering or relationship between values
  - **Ordinal**: values can be ordered, but differences between them is not meaningful
  - **Interval**: values are ordered and differences are meaningful, but not ratios
  - **Ratio**: values can be added, subtracted, multiplied, and divided

Scales of measurement

- Our textbook (Learning Statistics with R) defines **four scales of measurement**:
  - **Nominal**: no particular ordering or relationship between values
  - **Ordinal**: values can be ordered, but differences between them is not meaningful
  - **Interval**: values are ordered and differences are meaningful, but not ratios
  - **Ratio**: values can be added, subtracted, multiplied, and divided
- The scale of variables tells us **how we should analyze them**

Scales of measurement

- Our textbook (Learning Statistics with R) defines **four scales of measurement**:
  - **Nominal**: no particular ordering or relationship between values
  - **Ordinal**: values can be ordered, but differences between them is not meaningful
  - **Interval**: values are ordered and differences are meaningful, but not ratios
  - **Ratio**: values can be added, subtracted, multiplied, and divided
- The scale of variables tells us **how we should analyze them**
- We'll go into each of these in more detail

Nominal scale

Nominal scale

- Nominal variables have **no meaningful ordering or relationship** between different values
  - Examples: eye color, sex, citizenship, language, transportation method
  - Neither French nor German is "greater than" the other, and they **can't be averaged**

Nominal scale

- Nominal variables have **no meaningful ordering or relationship** between different values
  - Examples: eye color, sex, citizenship, language, transportation method
  - Neither French nor German is "greater than" the other, and they **can't be averaged**
- "The **only thing** you can say about the different possibilities [values] is that they are different" (LSR)

Nominal scale

- Nominal variables have **no meaningful ordering or relationship** between different values
  - Examples: eye color, sex, citizenship, language, transportation method
  - Neither French nor German is "greater than" the other, and they **can't be averaged**
- "The **only thing** you can say about the different possibilities [values] is that they are different" (LSR)
- I will usually refer to these as **categorical variables**

Ordinal scale

Ordinal scale

- Ordinal variables have a **natural, meaningful ordering** between values, but **no other structure**
  - Examples: position in a race, ranked choice voting, other rankings
  - Finishing 1st- and 2nd-place tells you who was faster, but **not by how much**

Ordinal scale

- Ordinal variables have a **natural, meaningful ordering** between values, but **no other structure**
  - Examples: position in a race, ranked choice voting, other rankings
  - Finishing 1st- and 2nd-place tells you who was faster, but **not by how much**
- Addition and subtraction are **not well-defined** (1st place isn't 3rd place minus 2nd place)

Ordinal scale

- Ordinal variables have a **natural, meaningful ordering** between values, but **no other structure**
  - Examples: position in a race, ranked choice voting, other rankings
  - Finishing 1st- and 2nd-place tells you who was faster, but **not by how much**
- Addition and subtraction are **not well-defined** (1st place isn't 3rd place minus 2nd place)
- Ordinal values also **can't be meaningfully averaged**

Interval scale

Interval scale

- Interval variables are **ordered** and **differences between them are meaningful** (you can do addition and subtraction)
  - Examples: Fahrenheit and Celsius, calendar year
  - The **difference** between 60° and 65° is the **same** as between 20° and 25°

Interval scale

- Interval variables are **ordered** and **differences between them are meaningful** (you can do addition and subtraction)
  - Examples: Fahrenheit and Celsius, calendar year
  - The **difference** between 60° and 65° is the **same** as between 20° and 25°
- Interval values **don't have a natural zero-value**
  - 0° is **not** actually a lack of heat; 40° is **not** twice as hot as 20°
  - 2008 is **not** "1.0024 times later" than 2003

Interval scale

- Interval variables are **ordered** and **differences between them are meaningful** (you can do addition and subtraction)
  - Examples: Fahrenheit and Celsius, calendar year
  - The **difference** between 60° and 65° is the **same** as between 20° and 25°
- Interval values **don't have a natural zero-value**
  - 0° is **not** actually a lack of heat; 40° is **not** twice as hot as 20°
  - 2008 is **not** "1.0024 times later" than 2003
- Lack of zero means intervals are **not meaningful to multiply & divide**

Ratio scale

Ratio scale

- Ratio variables have **ordering, meaningful differences,** and **a true zero**
  - Examples: race finish time, weight, age, sports scores (usually)
  - Meaningful to say a 30 y/o is **5 years older** than a 25 y/o (30 - 25)
  - or **1.2x older** (30 / 25)

Ratio scale

- Ratio variables have **ordering, meaningful differences,** and **a true zero**
  - Examples: race finish time, weight, age, sports scores (usually)
  - Meaningful to say a 30 y/o is **5 years older** than a 25 y/o (30 - 25)
  - or **1.2x older** (30 / 25)
- Ratio variables can be **meaningfully averaged**

Ratio scale

- Ratio variables have **ordering, meaningful differences,** and **a true zero**
  - Examples: race finish time, weight, age, sports scores (usually)
  - Meaningful to say a 30 y/o is **5 years older** than a 25 y/o (30 - 25)
  - or **1.2x older** (30 / 25)
- Ratio variables can be **meaningfully averaged**
- These are what you might think of as "regular" numbers

Continuous and discrete variables

Continuous and discrete variables

- **Continuous**: for any two values, it's **possible to have a value in-between**
  - E.g. between 2.0 and 3.0 is 2.5; between 2.0 and 2.1 is 2.05, etc.

Continuous and discrete variables

- **Continuous**: for any two values, it's **possible to have a value in-between**
  - E.g. between 2.0 and 3.0 is 2.5; between 2.0 and 2.1 is 2.05, etc.
- **Discrete**: any variable that's **not continuous**
  - Example: the number that comes up after **rolling a die**

Continuous and discrete variables

- **Continuous**: for any two values, it's **possible to have a value in-between**
  - E.g. between 2.0 and 3.0 is 2.5; between 2.0 and 2.1 is 2.05, etc.
- **Discrete**: any variable that's **not continuous**
  - Example: the number that comes up after **rolling a die**
- **Nominal and ordinal** variables are **discrete by definition**

	continuous	discrete
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

LSR (p.17)

Exercise

Exercise

- What are the **measurement scales** for each variable in our vowels dataset? (speaker, word, vowel, F1, F2, sex, height)

Exercise

- What are the **measurement scales** for each variable in our vowels dataset? (speaker, word, vowel, F1, F2, sex, height)
  - Speaker: **nominal/categorical**

Exercise

- What are the **measurement scales** for each variable in our vowels dataset? (speaker, word, vowel, F1, F2, sex, height)
  - Speaker: **nominal/categorical**
  - Word: **nominal/categorical**

Exercise

- What are the **measurement scales** for each variable in our vowels dataset? (speaker, word, vowel, F1, F2, sex, height)
  - Speaker: **nominal/categorical**
  - Word: **nominal/categorical**
  - Vowel: **nominal/categorical** (for our purposes)

Exercise

- What are the **measurement scales** for each variable in our vowels dataset? (speaker, word, vowel, F1, F2, sex, height)
  - Speaker: **nominal/categorical**
  - Word: **nominal/categorical**
  - Vowel: **nominal/categorical** (for our purposes)
  - F1: **ratio**

Exercise

- What are the **measurement scales** for each variable in our vowels dataset? (speaker, word, vowel, F1, F2, sex, height)
  - Speaker: **nominal/categorical**
  - Word: **nominal/categorical**
  - Vowel: **nominal/categorical** (for our purposes)
  - F1: **ratio**
  - F2: **ratio**

Exercise

- What are the **measurement scales** for each variable in our vowels dataset? (speaker, word, vowel, F1, F2, sex, height)
  - Speaker: **nominal/categorical**
  - Word: **nominal/categorical**
  - Vowel: **nominal/categorical** (for our purposes)
  - F1: **ratio**
  - F2: **ratio**
  - Sex: **nominal/categorical**

Exercise

- What are the **measurement scales** for each variable in our vowels dataset? (speaker, word, vowel, F1, F2, sex, height)
  - Speaker: **nominal/categorical**
  - Word: **nominal/categorical**
  - Vowel: **nominal/categorical** (for our purposes)
  - F1: **ratio**
  - F2: **ratio**
  - Sex: **nominal/categorical**
  - Height: **ratio**

Briefly: factors in R

R factors

R factors

- Sometimes a **categorical** variable is **coded as a number**, e.g. in a CSV
  - A column for "sex" might use 0 for male, 1 for female
  - **Trial groups** in an experiment might be numbered 1, 2, 3...

R factors

- Sometimes a **categorical** variable is **coded as a number**, e.g. in a CSV
  - A column for "sex" might use 0 for male, 1 for female
  - **Trial groups** in an experiment might be numbered 1, 2, 3...
- R will naively **assume these are numeric**. This can cause unintended problems

R factors

- Sometimes a **categorical** variable is **coded as a number**, e.g. in a CSV
  - A column for "sex" might use 0 for male, 1 for female
  - **Trial groups** in an experiment might be numbered 1, 2, 3...
- R will naively **assume these are numeric**. This can cause unintended problems

```
> sex = c(0, 1, 0, 0, 1, 0, 0, 0)
> sex
[1] 0 1 0 0 1 0 0 0
> sex < 2
[1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> sex / 8
[1] 0.000 0.125 0.000 0.000 0.125 0.000 0.000 0.000
>
```

R factors

R factors

- The R data type **factor** is used to designate **categorical variables**

R factors

- The R data type **factor** is used to designate **categorical variables**
- Data can be **converted to factors** with the `as.factor()` function
  - This prevents them being **manipulated in inappropriate ways**

```
> sex = c(0, 1, 0, 0, 1, 0, 0, 0)
> sex + 1
[1] 1 2 1 1 2 1 1 1
> sex = as.factor(sex)
> sex
[1] 0 1 0 0 1 0 0 0
Levels: 0 1
> sex + 1
[1] NA NA NA NA NA NA NA NA
Warning message:
In Ops.factor(sex, 1) : '+' not meaningful for factors
```

Factor levels

Factor levels

- The **same numbers** might be used for **different factor variables**
  - This also causes **unintended equivalence** between these variables

```
> sex = as.factor(c(0, 1, 0, 0, 1, 0, 0, 0))
> placebo_group = as.factor(c(1, 0, 0, 1, 1, 0, 1, 1))
> sex == placebo_group
[1] FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE
> levels(sex) = c("male", "female")
> levels(placebo_group) = c("placebo", "drug")
> sex
[1] male   female male   male   female male   male
[8] male
Levels: male female
> placebo_group
[1] drug    placebo placebo drug    drug    placebo
[7] drug    drug
Levels: placebo drug
> sex == placebo_group
Error in Ops.factor(sex, placebo_group) :
  level sets of factors are different
```

Factor levels

- The **same numbers** might be used for **different factor variables**
  - This also causes **unintended equivalence** between these variables
- The **levels** (values) of a factor can be **given names** with `levels()`
  - This can **disambiguate** mixups

```
> sex = as.factor(c(0, 1, 0, 0, 1, 0, 0, 0))
> placebo_group = as.factor(c(1, 0, 0, 1, 1, 0, 1, 1))
> sex == placebo_group
[1] FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE
> levels(sex) = c("male", "female")
> levels(placebo_group) = c("placebo", "drug")
> sex
[1] male   female male   male   female male   male
[8] male
Levels: male female
> placebo_group
[1] drug    placebo placebo drug    drug    placebo
[7] drug    drug
Levels: placebo drug
> sex == placebo_group
Error in Ops.factor(sex, placebo_group) :
  level sets of factors are different
```

Descriptive statistics

What are descriptive statistics?

What are descriptive statistics?

- Statistics is usually divided into **descriptive** and **inferential** versions
  - Descriptive: goal is to **concisely and meaningfully summarize** some data
  - Inferential: goal is to well... make **inferences** from some data (more in a week or two)

What are descriptive statistics?

- Statistics is usually divided into **descriptive** and **inferential** versions
  - Descriptive: goal is to **concisely and meaningfully summarize** some data
  - Inferential: goal is to well... make **inferences** from some data (more in a week or two)
- Two broad types of descriptive statistics are **central tendencies** (mean, median, mode, etc.) and **variability** (variance, standard deviation, etc.)
  - Central tendency: where the **"middle" of the data** is
  - Variability: **how much** the data **deviates from the central tendency**

Central tendencies

Mean

Mean

- What most people refer to as the **"average"**: the sum of all data points, divided by the total number of data points

Mean

- What most people refer to as the **"average"**: the sum of all data points, divided by the total number of data points

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- ($\overline{X}$ is the mean, X is the vector of all datapoints, X_i is a particular datapoint, N is the total number of datapoints)

Mean

- What most people refer to as the **"average"**: the sum of all data points, divided by the total number of data points

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

- ($\overline{X}$ is the mean, X is the vector of all datapoints, X_i is a particular datapoint, N is the total number of datapoints)

- Can be calculated easily in R with `mean()`

```
> mean(c(1, 2, 3, 4, 5))
[1] 3
```

Median

Median

- The median is the **middle datapoint**, when arranged in order
  - 3 is the median of **both** [1, 2, **3**, 4, 5] and [1, 2, **3**, 8, 10]
  - If there's an **even** number of datapoints, it's the **average of the middle two** (the median of [1, 2, 3, 4] is 2.5

Median

- The median is the **middle datapoint**, when arranged in order
  - 3 is the median of **both** [1, 2, **3**, 4, 5] and [1, 2, **3**, 8, 10]
  - If there's an **even** number of datapoints, it's the **average of the middle two** (the median of [1, 2, 3, 4] is 2.5
- Notice that median is **not as susceptible to outliers**

Median

- The median is the **middle datapoint**, when arranged in order
  - 3 is the median of **both** [1, 2, **3**, 4, 5] and [1, 2, **3**, 8, 10]
  - If there's an **even** number of datapoints, it's the **average of the middle two** (the median of [1, 2, 3, 4] is 2.5
- Notice that median is **not as susceptible to outliers**
- It is equivalent to the **50th percentile** (more on percentiles in a second)
  - **50% of the datapoints** are above and below the median

Median

- The median is the **middle datapoint**, when arranged in order
  - 3 is the median of **both** [1, 2, **3**, 4, 5] and [1, 2, **3**, 8, 10]
  - If there's an **even** number of datapoints, it's the **average of the middle two** (the median of [1, 2, 3, 4] is 2.5
- Notice that median is **not as susceptible to outliers**
- It is equivalent to the **50th percentile** (more on percentiles in a second)
  - **50% of the datapoints** are above and below the median
- R also has a `median()` function

Mean vs. median

Mean vs. median

- Advantages of using the **mean:**
  - It takes **every datapoint into account** (median ignores all but the middle)
  - It's **mathematically important** (will come up describing distributions)
  - **Almost everyone understands it** as the "average"

Mean vs. median

- Advantages of using the **mean:**
  - It takes **every datapoint into account** (median ignores all but the middle)
  - It's **mathematically important** (will come up describing distributions)
  - **Almost everyone understands it** as the "average"
- Advantages of using the **median:**
  - It is **resistant to outlier datapoints** (thus why most people refer to "median salary" or "median home price"; ultra-high earners skew the mean)
  - It is appropriate for **ordinal data** whereas mean is not!

Mean vs. median intuition

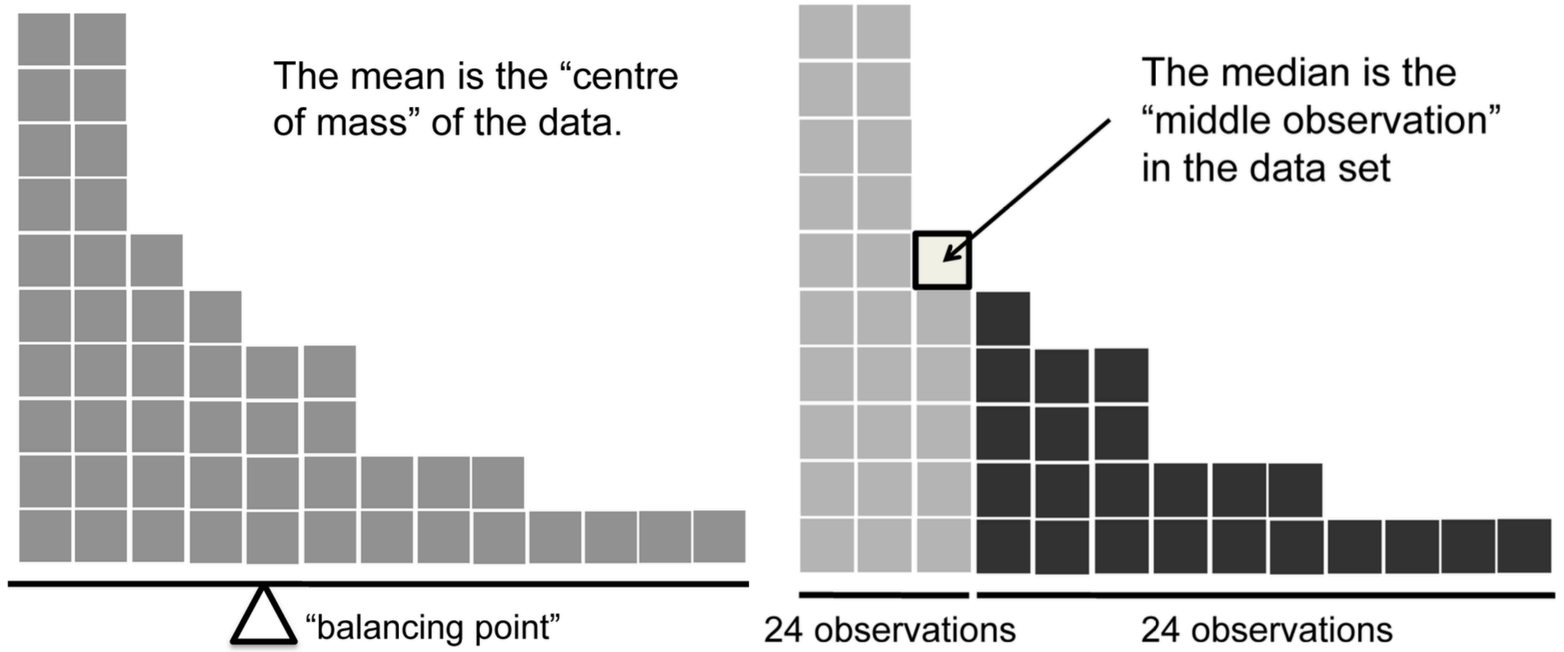

Figure 5.2: An illustration of the difference between how the mean and the median should be interpreted. The mean is basically the "centre of gravity" of the data set: if you imagine that the histogram of the data is a solid object, then the point on which you could balance it (as if on a see-saw) is the mean. In contrast, the median is the middle observation. Half of the observations are smaller, and half of the observations are larger.

LSR (p.118)

Variability

Mean Absolute Deviation

Mean Absolute Deviation

- **Deviation** refers to how far a datapoint is from the mean $(X_i - \overline{X})$

Mean Absolute Deviation

- **Deviation** refers to how far a datapoint is from the mean $(X_i - \overline{X})$
- **Absolute** deviation: we don't care if the difference is **positive** or **negative** (we take the absolute value $|X_i - \overline{X}|$)

Mean Absolute Deviation

- **Deviation** refers to how far a datapoint is from the mean $(X_i - \overline{X})$
- **Absolute** deviation: we don't care if the difference is **positive** or **negative** (we take the absolute value $|X_i - \overline{X}|$)
- Mean Absolute Deviation: the **average deviation** between the mean and all datapoints

$$\frac{1}{N}\sum_{i=1}^{N}|X_i - \overline{X}|$$

Mean Absolute Deviation

- **Deviation** refers to how far a datapoint is from the mean $(X_i - \overline{X})$
- **Absolute** deviation: we don't care if the difference is **positive** or **negative** (we take the absolute value $|X_i - \overline{X}|$)
- Mean Absolute Deviation: the **average deviation** between the mean and all datapoints

$$\frac{1}{N}\sum_{i=1}^{N}|X_i - \overline{X}|$$

- This measure is **easy to understand**, but **not used much** (to my knowledge)

Variance

Variance

- **Variance** is can also be defined as the **mean squared deviation**

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

Variance

- **Variance** is can also be defined as the **mean squared deviation**

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

- The squaring is similar to absolute value in **making all the numbers positive**

Variance

- **Variance** is can also be defined as the **mean squared deviation**

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

- The squaring is similar to absolute value in **making all the numbers positive**
- Very **important to math**, but not as intuitively interpretable to people
  - Standard deviation is more intuitive, and **derived from variance** (next slide)

Variance

- **Variance** is can also be defined as the **mean squared deviation**

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

- The squaring is similar to absolute value in **making all the numbers positive**
- Very **important to math**, but not as intuitively interpretable to people
  - Standard deviation is more intuitive, and **derived from variance** (next slide)
- R command: `var()`

Standard Deviation

Standard Deviation

- **Standard Deviation** is defined as the **square root of the variance**:

$$StDev(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

Standard Deviation

- **Standard Deviation** is defined as the **square root of the variance**:

$$StDev(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

- It is often abbreviated as σ (sigma), and variance as σ^2 (sigma squared)

Standard Deviation

- **Standard Deviation** is defined as the **square root of the variance**:

$$StDev(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

- It is often abbreviated as σ (sigma), and variance as σ^2 (sigma squared)
- Like variance, it has **nice mathematical properties**, but it is also **easier to interpret**, because it's in the same units as the original variable

Standard Deviation

- **Standard Deviation** is defined as the **square root of the variance**:

$$StDev(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

- It is often abbreviated as σ (sigma), and variance as σ^2 (sigma squared)
- Like variance, it has **nice mathematical properties**, but it is also **easier to interpret**, because it's in the same units as the original variable
- Probably the **most common** variability measure. You'll often see a distribution described with $\overline{X} \pm StDev(X)$

Standard Deviation

- **Standard Deviation** is defined as the **square root of the variance**:

$$StDev(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

- It is often abbreviated as σ (sigma), and variance as σ^2 (sigma squared)
- Like variance, it has **nice mathematical properties**, but it is also **easier to interpret**, because it's in the same units as the original variable
- Probably the **most common** variability measure. You'll often see a distribution described with $\overline{X} \pm StDev(X)$
- R command: `sd()`

Quantiles/Percentiles

Quantiles/Percentiles

- A **quantile** or **percentile** is a value below which a **certain percentage of the data** falls

Quantiles/Percentiles

- A **quantile** or **percentile** is a value below which a **certain percentage of the data** falls
  - If the 50th percentile of height is 169cm, that means **50% of the datapoints** are shorter than 169cm

Quantiles/Percentiles

- A **quantile** or **percentile** is a value below which a **certain percentage of the data** falls
  - If the 50th percentile of height is 169cm, that means **50% of the datapoints** are shorter than 169cm
  - The **median** is the same thing as the **50th percentile**

Quantiles/Percentiles

- A **quantile** or **percentile** is a value below which a **certain percentage of the data** falls
  - If the 50th percentile of height is 169cm, that means **50% of the datapoints** are shorter than 169cm
  - The **median** is the same thing as the **50th percentile**
- Can view quantiles in R with `quantile()`

```
> quantile(vowels$HEIGHT, probs=0.5)
50%
169
> quantile(vowels$HEIGHT, probs=c(0.25, 0.5, 0.75))
   25%    50%    75%
163.00 169.00 175.25
```

Interquartile Range

Interquartile Range

- The Interquartile Range (IQR) is the **difference** between the **75th percentile** and the **25th percentile**

Interquartile Range

- The Interquartile Range (IQR) is the **difference** between the **75th percentile** and the **25th percentile**
- The idea is this shows how wide the **middle 50%** of the data is
  - More useful **in comparison** to the **total range**

Interquartile Range

- The Interquartile Range (IQR) is the **difference** between the **75th percentile** and the **25th percentile**
- The idea is this shows how wide the **middle 50%** of the data is
  - More useful **in comparison** to the **total range**
- Can view in R with `IQR()`

```
> IQR(vowels$HEIGHT)
[1] 12.25
> range = max(vowels$HEIGHT) - min(vowels$HEIGHT)
> range
[1] 36
```

Summary statistics

- R has a handy function to get several summary statistics in a **single command**, and can be called on **either** a vector or a data frame

```
> summary(vowels$HEIGHT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  157.0   163.0   169.0   169.9   175.2   193.0
> summary(vowels)
   SPEAKER              WORD              VOWEL                F1
 Length:484         Length:484         Length:484         Min.   : 177.6
 Class :character   Class :character   Class :character   1st Qu.: 447.0
 Mode  :character   Mode  :character   Mode  :character   Median : 549.2
                                                          Mean   : 581.3
                                                          3rd Qu.: 710.2
                                                          Max.   :1059.1

       F2            SEX                HEIGHT
 Min.   : 379   Length:484         Min.   :157.0
 1st Qu.:1256   Class :character   1st Qu.:163.0
 Median :1541   Mode  :character   Median :169.0
 Mean   :1576                      Mean   :169.9
 3rd Qu.:1853                      3rd Qu.:175.2
 Max.   :2875                      Max.   :193.0
```

Summary statistics

- R has a handy function to get several summary statistics in a **single command**, and can be called on **either** a vector or a data frame

Notice the summary of the character columns isn't very useful

```
> summary(vowels$HEIGHT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  157.0   163.0   169.0   169.9   175.2   193.0
> summary(vowels)
   SPEAKER              WORD              VOWEL                F1
 Length:484         Length:484         Length:484         Min.   : 177.6
 Class :character   Class :character   Class :character   1st Qu.: 447.0
 Mode  :character   Mode  :character   Mode  :character   Median : 549.2
                                                          Mean   : 581.3
                                                          3rd Qu.: 710.2
                                                          Max.   :1059.1

       F2            SEX                HEIGHT
 Min.   : 379   Length:484         Min.   :157.0
 1st Qu.:1256   Class :character   1st Qu.:163.0
 Median :1541   Mode  :character   Median :169.0
 Mean   :1576                      Mean   :169.9
 3rd Qu.:1853                      3rd Qu.:175.2
 Max.   :2875                      Max.   :193.0
```