

# **PSC4375: Descriptive Statistics**

## **Week 2: Lecture 4**

Prof. Weldzius

Villanova University

Slides Updated: 2025-01-29

# Lots of data

- Data from study of the effect of minimum wage

```
library(tidyverse)
data(minwage, package = "qss")
head(minwage)
```

# Lots of data

- Data from study of the effect of minimum wage

##	chain	location	wageBefore	wageAfter	fullBefore
## 1	wendys	PA	5.00	5.25	20
## 2	wendys	PA	5.50	4.75	6
## 3	burgerking	PA	5.00	4.75	50
## 4	burgerking	PA	5.00	5.00	10
## 5	kfc	PA	5.25	5.00	2
## 6	kfc	PA	5.00	5.00	2

  

##	fullAfter	partBefore	partAfter
## 1	0	20	36
## 2	28	26	3
## 3	15	35	18
## 4	26	17	9
## 5	3	8	12
## 6	2	10	9

# Lots and lots of data

```
head(minwage$wageAfter, n = 200)
```

```
##      [1] 5.25 4.75 4.75 5.00 5.00 5.00 4.75 5.00 4.50 4.75 4.5
##     [12] 5.00 4.75 4.75 4.75 4.25 5.00 4.90 5.00 4.75 5.00 4.2
##     [23] 4.75 4.25 4.25 4.25 4.25 4.25 4.25 4.38 4.75 4.25 4.5
##     [34] 4.50 4.25 4.25 4.25 4.25 5.05 4.25 4.25 4.25 4.25 4.3
##     [45] 4.50 4.50 5.00 4.75 5.00 4.35 4.25 4.90 4.50 4.50 4.7
##     [56] 6.25 4.35 4.50 4.50 5.00 4.75 4.50 4.75 4.25 4.91 4.4
##     [67] 4.25 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.0
##     [78] 5.05 5.05 5.05 5.50 5.05 5.05 5.05 5.05 5.05 5.05 5.2
##     [89] 5.25 5.05 5.05 5.50 5.05 5.05 5.05 5.05 5.05 5.05 5.0
##    [100] 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.2
##    [111] 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.67 5.0
##    [122] 5.05 5.05 5.05 5.25 5.25 5.05 5.50 5.05 5.05 5.05 5.5
##    [133] 5.50 5.05 5.05 5.25 5.05 5.05 5.15 5.05 5.05 5.05 5.0
##    [144] 5.00 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.0
```

# How to summarize data

- How should we summarize the wages data? Many possibilities!

# How to summarize data

- How should we summarize the wages data? Many possibilities!
  - Up to now: focus on **averages** or means of variables

# How to summarize data

- How should we summarize the wages data? Many possibilities!
  - Up to now: focus on **averages** or means of variables
- Two salient features of a variable that we want to know:

# How to summarize data

- How should we summarize the wages data? Many possibilities!
  - Up to now: focus on **averages** or means of variables
- Two salient features of a variable that we want to know:
  - **Central tendency**: where is the middle/typical/average value



# How to summarize data

- How should we summarize the wages data? Many possibilities!
  - Up to now: focus on **averages** or means of variables
- Two salient features of a variable that we want to know:
  - **Central tendency**: where is the middle/typical/average value
  - **Spread** around the center: are all values to the center or spread out?

# Center of the data

- “Center” of the data: typical/average value

# Center of the data

- “Center” of the data: typical/average value
- **Mean:** sum of the values divided by the number of observations

# Center of the data

- “Center” of the data: typical/average value
- **Mean:** sum of the values divided by the number of observations

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

# Center of the data

- “Center” of the data: typical/average value
- **Mean:** sum of the values divided by the number of observations

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

- **Median:**

# Center of the data

- “Center” of the data: typical/average value
- **Mean:** sum of the values divided by the number of observations

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

- **Median:**

$$\text{median} = \begin{cases} \text{middle value} & \text{if number of entries is odd} \\ \frac{\text{sum of two middle values}}{2} & \text{if number of entries is even} \end{cases}$$

# Center of the data

- “Center” of the data: typical/average value
- **Mean:** sum of the values divided by the number of observations

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

- **Median:**

$$\text{median} = \begin{cases} \text{middle value} & \text{if number of entries is odd} \\ \frac{\text{sum of two middle values}}{2} & \text{if number of entries is even} \end{cases}$$

- In **R**: `mean()` and `median()`

# Mean vs median

- Median more robust to **outliers**:



# Mean vs median

- Median more robust to **outliers**:
  - Example 1: data = 0, 1, 2, 3, 5. Mean? Median?

# Mean vs median

- Median more robust to **outliers**:
  - Example 1: data = 0, 1, 2, 3, 5. Mean? Median?
  - Example 2: data = 0, 1, 2, 3, 100. Mean? Median?

# Mean vs median

- Median more robust to **outliers**:
  - Example 1: data = 0, 1, 2, 3, 5. Mean? Median?
  - Example 2: data = 0, 1, 2, 3, 100. Mean? Median?
- What does Mark Zuckerberg do to the mean vs. median income?

# Spread of the data

- Are the values of the variable close to the center?

# Spread of the data

- Are the values of the variable close to the center?
- **Range:**  $[\min(X), \max(X)]$

# Spread of the data

- Are the values of the variable close to the center?
- **Range:**  $[\min(X), \max(X)]$
- **Quantile** (quartile, percentile, etc.): divide data into equal sized groups.

# Spread of the data

- Are the values of the variable close to the center?
- **Range:**  $[\min(X), \max(X)]$
- **Quantile** (quartile, percentile, etc.): divide data into equal sized groups.
  - 25th percentile: lower quartile (25% of the data below this value)

# Spread of the data

- Are the values of the variable close to the center?
- **Range:**  $[\min(X), \max(X)]$
- **Quantile** (quartile, percentile, etc.): divide data into equal sized groups.
  - 25th percentile: lower quartile (25% of the data below this value)
  - 50th percentile: median (50% of the data below this value)



# Spread of the data

- Are the values of the variable close to the center?
- **Range:**  $[\min(X), \max(X)]$
- **Quantile** (quartile, percentile, etc.): divide data into equal sized groups.
  - 25th percentile: lower quartile (25% of the data below this value)
  - 50th percentile: median (50% of the data below this value)
  - 75th percentile: upper quartile (75% of the data below this value)

# Spread of the data

- Are the values of the variable close to the center?
- **Range:**  $[\min(X), \max(X)]$
- **Quantile** (quartile, percentile, etc.): divide data into equal sized groups.
  - 25th percentile: lower quartile (25% of the data below this value)
  - 50th percentile: median (50% of the data below this value)
  - 75th percentile: upper quartile (75% of the data below this value)
- **Interquartile range (IQR):** a measure of variability

# Spread of the data

- Are the values of the variable close to the center?
- **Range:**  $[\min(X), \max(X)]$
- **Quantile** (quartile, percentile, etc.): divide data into equal sized groups.
  - 25th percentile: lower quartile (25% of the data below this value)
  - 50th percentile: median (50% of the data below this value)
  - 75th percentile: upper quartile (75% of the data below this value)
- **Interquartile range (IQR):** a measure of variability
  - How spread out is the middle half of the data?

# Spread of the data

- Are the values of the variable close to the center?
- **Range:**  $[\min(X), \max(X)]$
- **Quantile** (quartile, percentile, etc.): divide data into equal sized groups.
  - 25th percentile: lower quartile (25% of the data below this value)
  - 50th percentile: median (50% of the data below this value)
  - 75th percentile: upper quartile (75% of the data below this value)
- **Interquartile range (IQR):** a measure of variability
  - How spread out is the middle half of the data?
  - Is most of the data really close to the median or are the values spread out?

# Spread of the data

- Are the values of the variable close to the center?
- **Range:**  $[\min(X), \max(X)]$
- **Quantile** (quartile, percentile, etc.): divide data into equal sized groups.
  - 25th percentile: lower quartile (25% of the data below this value)
  - 50th percentile: median (50% of the data below this value)
  - 75th percentile: upper quartile (75% of the data below this value)
- **Interquartile range (IQR):** a measure of variability
  - How spread out is the middle half of the data?
  - Is most of the data really close to the median or are the values spread out?
- **R function:** `range()`, `summary()`, `IQR()`

# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Steps:



# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Steps:
  - 1 Subtract each data point by the mean

# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Steps:
  - 1 Subtract each data point by the mean
  - 2 Square each resulting difference

# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Steps:
  - 1 Subtract each data point by the mean
  - 2 Square each resulting difference
  - 3 Take the sum of these values

# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Steps:
  - 1 Subtract each data point by the mean
  - 2 Square each resulting difference
  - 3 Take the sum of these values
  - 4 Divide by  $n - 1$  (or  $n$ , doesn't matter much)

# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Steps:
  - 1 Subtract each data point by the mean
  - 2 Square each resulting difference
  - 3 Take the sum of these values
  - 4 Divide by  $n - 1$  (or  $n$ , doesn't matter much)
  - 5 Take the square root

# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Steps:
  - 1 Subtract each data point by the mean
  - 2 Square each resulting difference
  - 3 Take the sum of these values
  - 4 Divide by  $n - 1$  (or  $n$ , doesn't matter much)
  - 5 Take the square root
- **Variance:** standard deviation<sup>2</sup>

# Standard deviation

- **Standard deviation:** On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Steps:
  - 1 Subtract each data point by the mean
  - 2 Square each resulting difference
  - 3 Take the sum of these values
  - 4 Divide by  $n - 1$  (or  $n$ , doesn't matter much)
  - 5 Take the square root
- **Variance:** standard deviation<sup>2</sup>
- Why not just take the average deviations from mean without squaring?

# How large is large?

- Is a wage of 5.30 an hour large?



# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?

# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!

# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!
  - Different units, difference spreads of the data, etc.

# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!
  - Different units, difference spreads of the data, etc.
- Need a way to put any variable on **common units**

# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!
  - Different units, difference spreads of the data, etc.
- Need a way to put any variable on **common units**
- **z-score:**

# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!
  - Different units, difference spreads of the data, etc.
- Need a way to put any variable on **common units**
- **z-score:**

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!
  - Different units, difference spreads of the data, etc.
- Need a way to put any variable on **common units**
- **z-score:**

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

- Interpretation:

# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!
  - Different units, difference spreads of the data, etc.
- Need a way to put any variable on **common units**
- **z-score**:

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

- Interpretation:
  - Positive values above the mean, negative values below the mean



# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!
  - Different units, difference spreads of the data, etc.
- Need a way to put any variable on **common units**
- **z-score:**

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

- Interpretation:
  - Positive values above the mean, negative values below the mean
  - Units now on the scale of **standard deviations away from the mean**

# How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!
  - Different units, difference spreads of the data, etc.
- Need a way to put any variable on **common units**
- **z-score:**

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

- Interpretation:
  - Positive values above the mean, negative values below the mean
  - Units now on the scale of **standard deviations away from the mean**
  - Intuition: data more than 3 SDs away from mean are rare

## z-score example

- Jane works at The Grog where there's a tip jar.

## z-score example

- Jane works at The Grog where there's a tip jar.
- She's been keeping track of her daily tips:

## z-score example

- Jane works at The Grog where there's a tip jar.
- She's been keeping track of her daily tips:
  - Average tip of \$1.56 with a standard deviation of 20 cents.

## z-score example

- Jane works at The Grog where there's a tip jar.
- She's been keeping track of her daily tips:
  - Average tip of \$1.56 with a standard deviation of 20 cents.
- Yesterday, Jane got a \$1.86 tip. How big is this?

## z-score example

- Jane works at The Grog where there's a tip jar.
- She's been keeping track of her daily tips:
  - Average tip of \$1.56 with a standard deviation of 20 cents.
- Yesterday, Jane got a \$1.86 tip. How big is this?
- Today she got \$0.56, what about that?