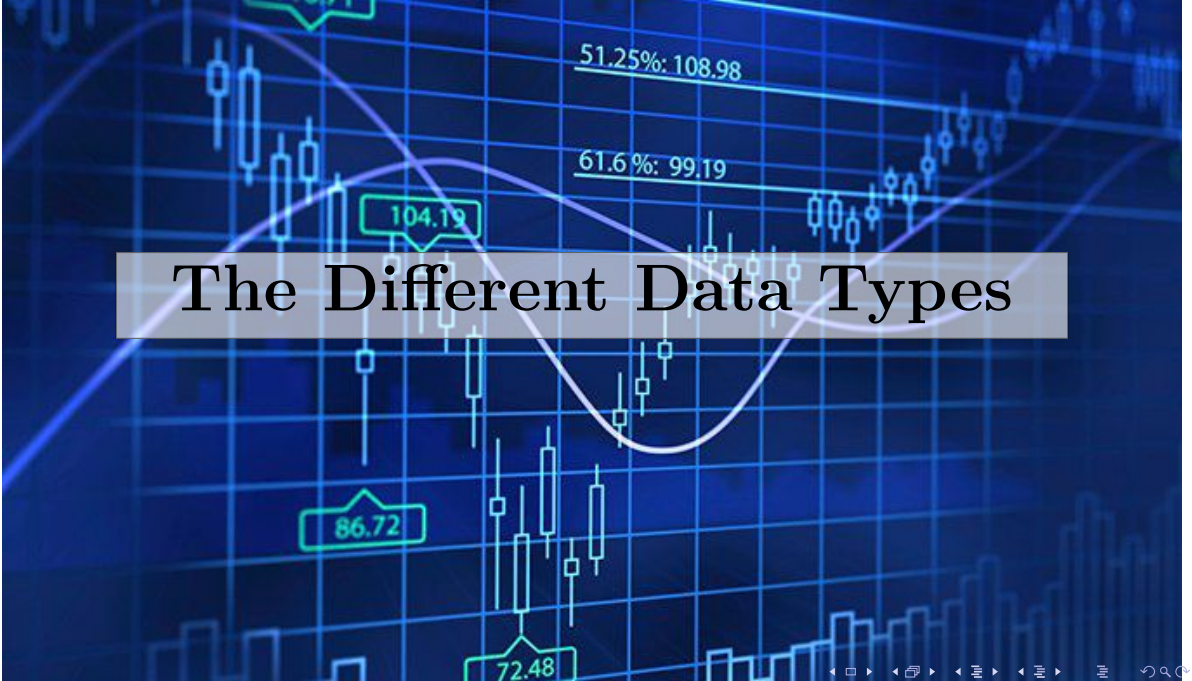


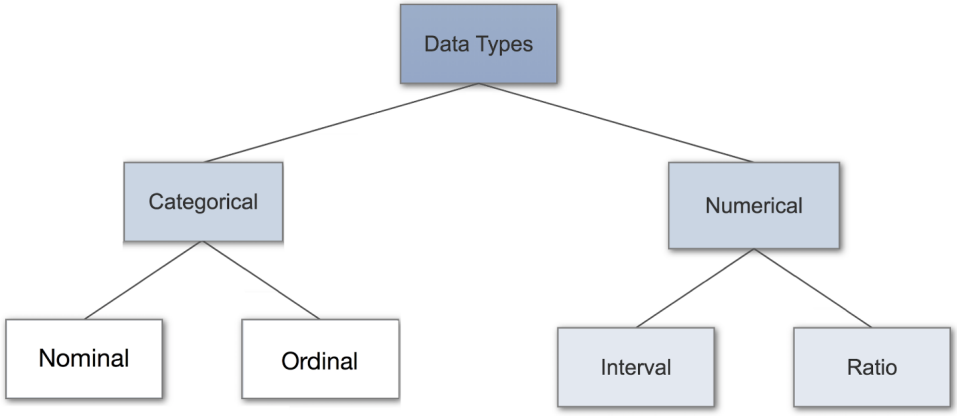
The Different Data Types

Learning Outcomes

The CRISP-DM process.

- Define and give examples of the different types of data;
- Define descriptive statistics;
- Determine which descriptive statistics are appropriate for each type of data;
- Explain why knowing the data type and using the right descriptive statistic is not enough.

Data Types: the 'Levels of Measurement'



Data Types: the **categorical** 'Levels of Measurement'

- **Nominal**;
  - Qualitative classification of different objects by names – measures membership;
  - No quantitative value or order information;
  - Examples: Gender, nationality, zip code, eye color, error code;

- **Ordinal**;
  - Categories with a natural ordering, but no well-defined scale – measures rank;
  - No quantitative value;
  - Examples: Party membership, polling agreement (Likert) scales, ed level, class;

Data Types: the **numerical** 'Levels of Measurement'

- **Discrete**;
  - Difference btwn units on scale is constant but can only take certain values;
  - Can be counted;
  - Examples: Age in years, Turn 1, Turn 2, etc.

- **Interval**;
  - Difference btwn units on scale is constant, but no zero point – measures exact difference;
  - Examples: Time of day, date, temperature (F or C), test scores, IQ;

- **Ratio**;
  - Difference btwn units on scale is constant/has a zero point – measures exact difference +;
  - Examples: Height and weight, earnings, spending, tax rate, temperature (K).

What are descriptive statistics?!

- In general, two kinds of statistics:
  - **Descriptive Statistics** – what we'll talk about today;
  - **Inferential Statistics** – what we'll spend much of the rest of the semester on;

- Typically, descriptive statistics are always reported even if main focus is on something more sophisticated.

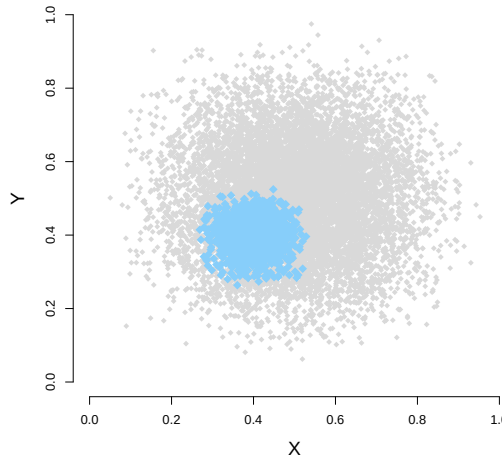
What are descriptive statistics: key terms

- **Population**: a 'complete' group of N objects, items, entities, or events of interest – e.g. all adults living in the US;
- **Sample**: a selected subset of n individuals from a population – e.g. 5,000 US adults appearing in a poll;
- **Summary Statistic**: a summary of the information in a set of observations – e.g. mean, median, mode, etc.;
- **Parametric**: derived from a probability distribution – e.g. a z-score is related to a normal distribution;
- **Nonparametric**: NOT derived from a probability distribution – e.g. descriptive statistics, histograms, etc.;
- **Univariate**: dealing with a single variable;
- **Multivariate**: dealing with relationships between several variables;

What job do they do?

- The main job of descriptive statistics is to **summarize** the information in a **sample**;
  - ...describe the data in the sample;
  - ...assess data quality (e.g. variation, correlation btwn variables, etc.);
  - ...support later inferential analysis;

- The main job of inferential statistics is to **learn** about the **population** that the sample comes from.



MSOE UNIVERSITY


Population
Sampling
Sample
Inferential statistics
Descriptive statistics

Measures of Central Tendency (for dataset x: $0, 1, 4, 4, 5, 5, 7, 7, 7, 9, 9$)

- **Mode** – the **sample** mode is written as $\overline{x}_{mode}$ and is the element that occurs most often in the sample. In our example $\overline{x}_{mode} = 7$.

Measures of Central Tendency (for dataset x: $0, 1, 4, 4, 5, 5, 7, 7, 7, 9, 9$)

- **Median** – the **sample** median is written as $\overline{x}_{med}$:

$$\overline{x}_{med} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ x_{(n/2)} + x_{(n/2)+1} & \text{if } n \text{ is even} \end{cases} \implies \overline{x}_{med} = x_{(11+1)/2} = x_6 = 5.$$

Measures of Central Tendency (for dataset x: $0, 1, 4, 4, 5, 5, 7, 7, 7, 9, 9$)

- **Arithmetic mean** – the **sample** mean is written as $\overline{x}$:

$$\overline{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \Longrightarrow \frac{0+1+4+4+5+5+7+7+7+9+9}{11} \approx 5.273.$$

Measures of Central Tendency (for dataset x: $0, 1, 4, 4, 5, 5, 7, 7, 7, 9, 9$)

- **Midrange** – the **sample** midrange is written as $\overline{x}_{mid}$:

$$\overline{x}_{mid} = \frac{\max\{x\} + \min\{x\}}{2} \Longrightarrow \overline{x}_{mid} = \frac{9+0}{2} = 4.5.$$

Measures of Variability (for dataset x: $0, 1, 4, 4, 5, 5, 7, 7, 7, 9, 9$)

- **Range** – written as σ_{max}, the distance between the min and max:

$$\sigma_{max} = \max\{x\} - \min\{x\} \implies 9 - 0 = 9.$$

Measures of Variability (for dataset x: $0, 1, 4, 4, 5, 5, 7, 7, 7, 9, 9$)

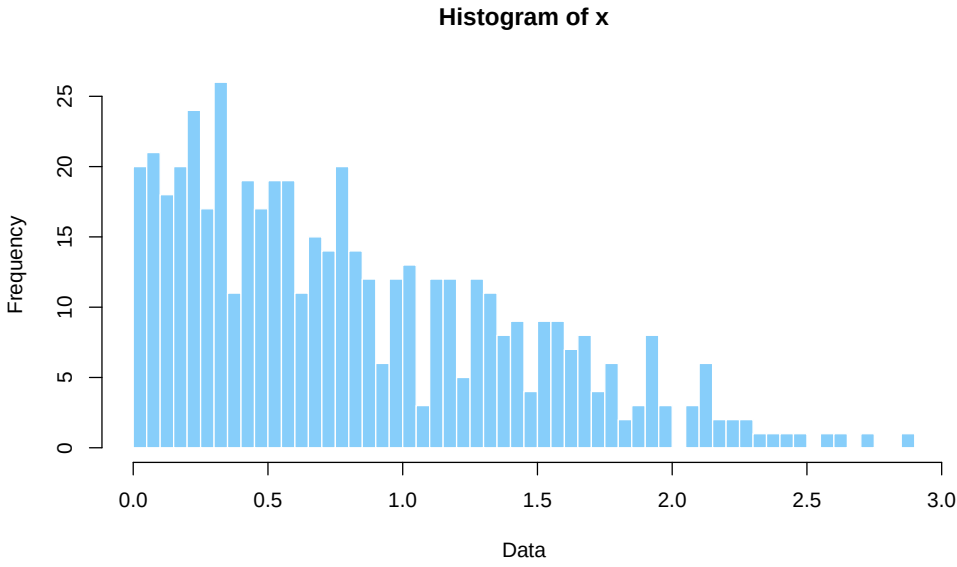
- **Variation ratio** – written as σ_{vr}, the proportion of cases NOT in the modal category:

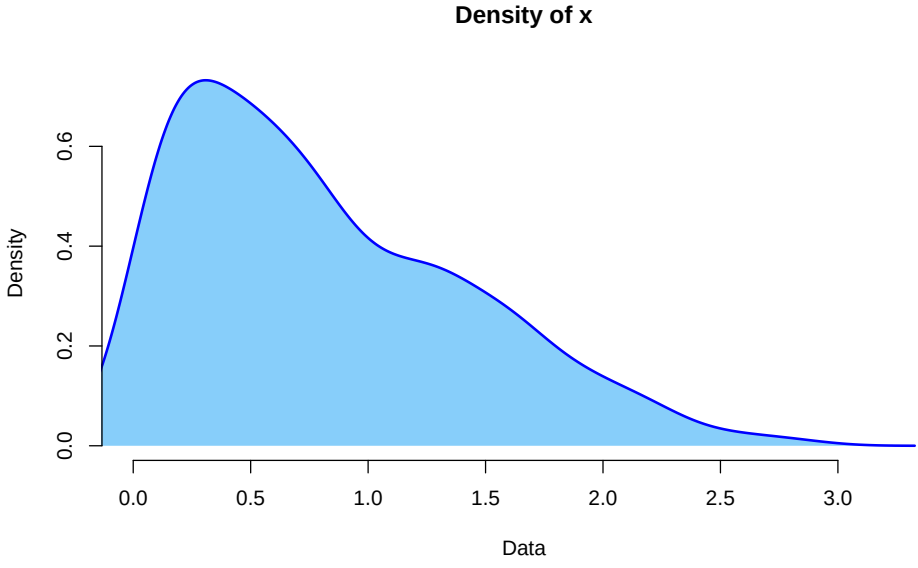
$$\sigma_{vr} = 1 - \frac{f_m}{n} \Longrightarrow 1 - \frac{3}{11} \approx 0.727, \text{where } f_m = \text{ \# of cases IN the modal category.}$$

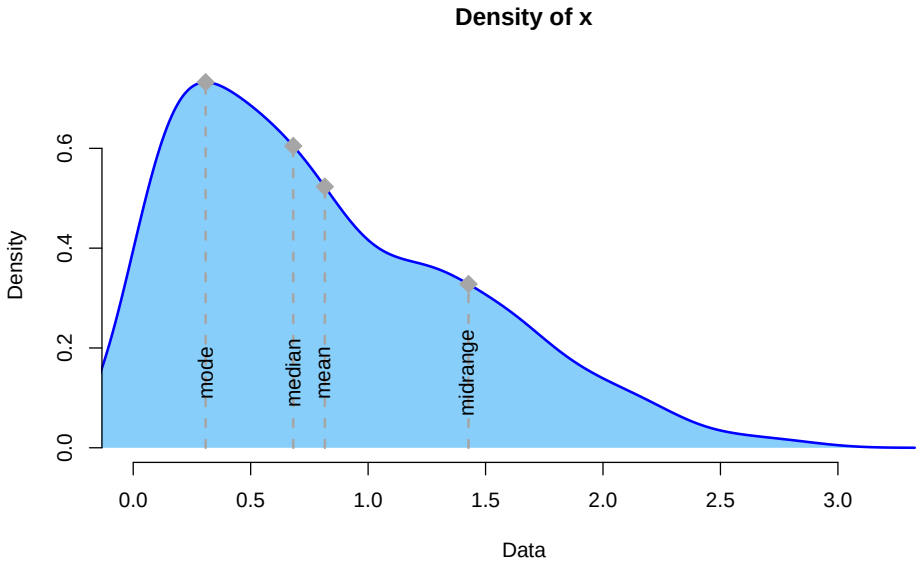
Measures of Variability (for dataset x: $0, 1, 4, 4, 5, 5, 7, 7, 7, 9, 9$)

- **Standard deviation/Variance** – written as σ, the sum of squared distance from mean:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{(0-5.273)^2 + (1-5.273)^2 + \ldots + (9-5.273)^2}{11}} \approx 2.799$$


Histogram of x
Frequency
0
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Data


Density of x
Density
0.0
0.2
0.4
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Data


Density of x
Density
0.0
0.2
0.4
0.6
mode
median
mean
midrange
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Data

Data Types: the **categorical** 'Levels of Measurement'

- **Nominal**;
  - Qualitative classification of different objects by names – measures membership;
  - No quantitative value or order information;
  - Examples: Gender, nationality, zip code, eye color, error code;
  - Appropriate: equality, mode, Variation ratio;

- **Ordinal**;
  - Categories with a natural ordering, but no well-defined scale – measures rank;
  - No quantitative value;
  - Examples: Party membership, polling agreement (Likert) scales, ed level, class;
  - Appropriate: above plus $>$ and $<$, median, range;

Data Types: the **numerical** 'Levels of Measurement'

- **Discrete**;
  - Difference btwn units on scale is constant but can only take certain values;
  - Can be counted;
  - Examples: Age in years, Turn 1, Turn 2, etc.
  - Appropriate: above plus + and −;

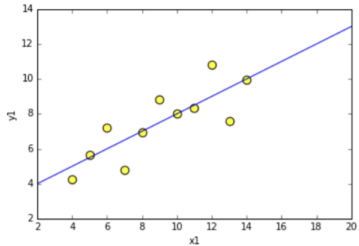
- **Interval**;
  - Difference btwn units on scale is constant, but no zero point – measures exact difference;
  - Examples: Time of day, date, temperature (F or C), test scores, IQ;
  - Appropriate: above plus + and −, mean, standard deviation;

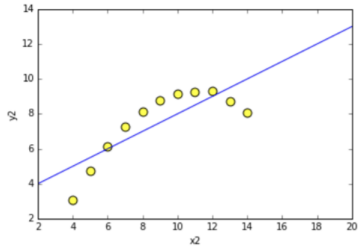
- **Ratio**;
  - Difference btwn units on scale is constant/has a zero point – measures exact difference +;
  - Examples: Height and weight, earnings, spending, tax rate, temperature (K).
  - Appropriate: above plus $*$ and /.

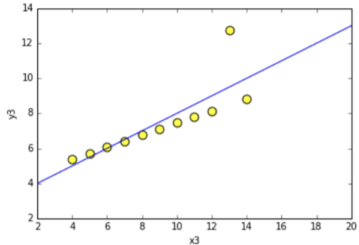
Descriptive stats are not enough: Anscombe's Quartet

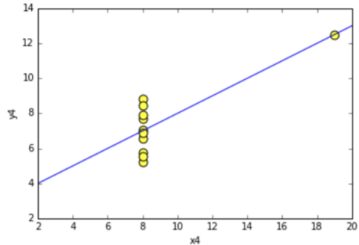
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.31	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Corr.	0.816		0.816		0.816		0.816	

Descriptive stats are not enough: Anscombe's Quartet





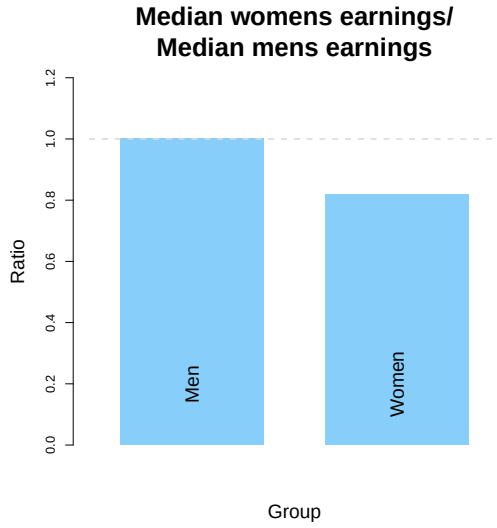




MSOE UNIVERSITY

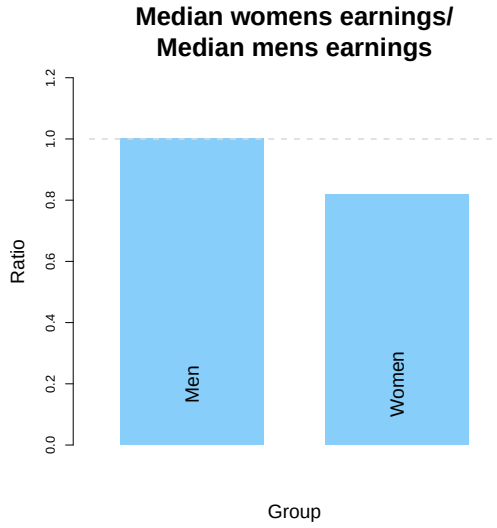
Descriptive stats are not enough: Gender Wage Gap

- Lots of progress – ever more women in labor market w/ higher education;
- Refers to the earnings difference between women and men:
  - Women consistently earn less than men in US;
  - How to measure how much less?
  - ...and what drives the gap???
- Simple descriptive statistics:
  - Compute median annual earnings for full time women and men;
  - Take the ratio.



Descriptive stats are not enough: Gender Wage Gap

- Lots of progress – ever more women in labor market w/ higher education;
- Refers to the earnings difference between women and men:
  - Women consistently earn less than men in US;
  - How to measure how much less?
  - ...and what drives the gap???
- Simple descriptive statistics:
  - Compute median annual earnings for full time women and men;
  - Take the ratio.
  - Is this enough?



Descriptive stats are not enough: Gender Wage Gap

- Lots of progress – ever more women in labor market w/ higher education;
- Refers to the earnings difference between women and men:
  - Women consistently earn less than men in US;
  - How to measure how much less?
  - ...and what drives the gap???
- Simple descriptive statistics:
  - Compute median annual earnings for full time women and men;
  - Take the ratio.
  - Is this enough? NO!!!