Last Week:

- Sampling Lecture
  - Key terms
  - Sources of error
  - **Probability Samples**
    - Simple Random, Systematic, Stratified, Multistage cluster
    - Importance of sample size

- **Non-Probability samples**
  - Convenience sample
  - Snowball sample
  - Theoretical sample
  - Quota sample

Today:

- Finish sampling lecture
- Move onto Quantitative Data Analysis
  - 3 Types of variables
  - 3 Fundamental conditions for establishing causality
  - 3 types of causes
  - Univariate distributions and statistics
  - Bivariate and Multivariate relationships

- Assignment 2 (SPSS).. Due next week..
- Extension: Friday, 5:00 p.m. (my office:LH208)
- I am primarily interested in your SPSS output!!

Non-PS: Convenience Sample

- Cases are included because they are readily available
  - Ex. The sample of students outside the library..

Non-PS: Snowball Sample

- A form of convenience sampling:
  - Researcher makes contact with one or a few people, who then introduce the researcher to more people, who then introduce the researcher to more people
  - An excellent strategy for ‘hidden’ pops that have strong networks..
    - Ex. Ethnic or Religious Communities

Non-PS: Quota Sample

- Also a convenience sample of sorts
  - –

Qualitative Sampling

- Heavy use of the Non-P sampling discussed already
- Also, ***theoretical sampling***
  –

- Next chapter: Quantitative Analysis!!

- Start with “types of variables”
- And then move on to discuss, the concept of “Causality in greater detail”..

Types of Variables

- ***Nominal*:** There are

Types of Variables

- ***Ordinal*:**

Types of Variables

- ***Interval/ratio:***

Establishing Causality

- Cause and effect

Establishing Causality

1) Temporal order

- the cause of an event must occur before the event being explained

Establishing Causality

1) Temporal Order

Establishing Causality

2) association between variables

–

Establishing Causality

Education (X)	Percent in favour of Capital punishment (Y)	
less than high school	40	
high school grad	28	
some college/college grad	26	
some university	22	
university grad	11	

- As education (**X**) increases, % in favour of capital punishment (**Y**) decreases – they move together
- Appears to be an association between these two variables

Establishing Causality

Education (X)	Percent in favour of Capital punishment (Y)	
less than high school	40	
high school grad	41	
some college/college grad	39	
some university	39	
university grad	40	

- As education increases, % in favour of capital punishment remains roughly the same across categories
- No association here

Establishing Causality

2) association between variables

–

Establishing Causality

2) association between variables

Establishing Causality

3) Eliminate (rule out) Spurious Relationships

Establishing Causality

3) Eliminate (rule out) Spurious Relationships

-

Establishing Causality

3) Eliminate (rule out) Spurious Relationships

Types of Causality

- Not all causes were created equal

- 3 basic types of causes:
  -

Types of Causality

1. ***Necessary cause***:
   - –

Types of Causality

2. ***Sufficient cause:***

   –

Types of Causality

3. ***Contributory cause:***

   –

Inferential statistics

1. ***Univariate***

2. Bivariate

3. Multivariate (many variables)

1. Univariate Distributions

- ***Frequency Distributions:***
  - Simple count across categories of a single variable
    - Mutually exclusive and exhaustive
  - Possible with variables at all levels of measurement (nominal, ordinal, interval/ratio)

Grades	Frequency		
A+	312		
A	419		
B	2789		
C	3505		
D	782		
F	114		
Total	7921		

1. Univariate Distributions

- ***% frequency***
  - Easier to understand the size differences between categories

Grades	Frequency	% Frequency	
A+	312	3.94	
A	419	5.29	
B	2789	35.21	
C	3505	44.25	
D	782	9.87	
F	114	1.44	
Total	7921	100	

1. Univariate Distributions

- % Frequency becomes more useful when making comparisons across more than one group or sample

Grades	Frequency	% Frequency	Cumulative %
A+	312	3.94	3.94
A	419	5.29	9.23
B	2789	35.21	44.44
C	3505	44.25	88.69
D	782	9.87	98.56
F	114	1.44	100.00
Total	7921	100	100

Grades	Frequency	% Frequency	Cumulative %
A+	568	3.33	3.33
A	954	5.59	8.91
B	5432	31.81	40.73
C	7887	46.19	86.92
D	2000	11.71	98.63
F	234	1.37	100.00
Total	17075	100	100

1. Univariate Distributions

- Also, ***cumulative frequency distributions***
  - % of cases in each category added to the % in proceeding categories
  - More useful for rank ordered variables

Grades	Frequency	% Frequency	Cumulative %
A+	312	3.94	3.94
A	419	5.29	9.23
B	2789	35.21	44.44
C	3505	44.25	88.69
D	782	9.87	98.56
F	114	1.44	100.00
Total	7921	100	100

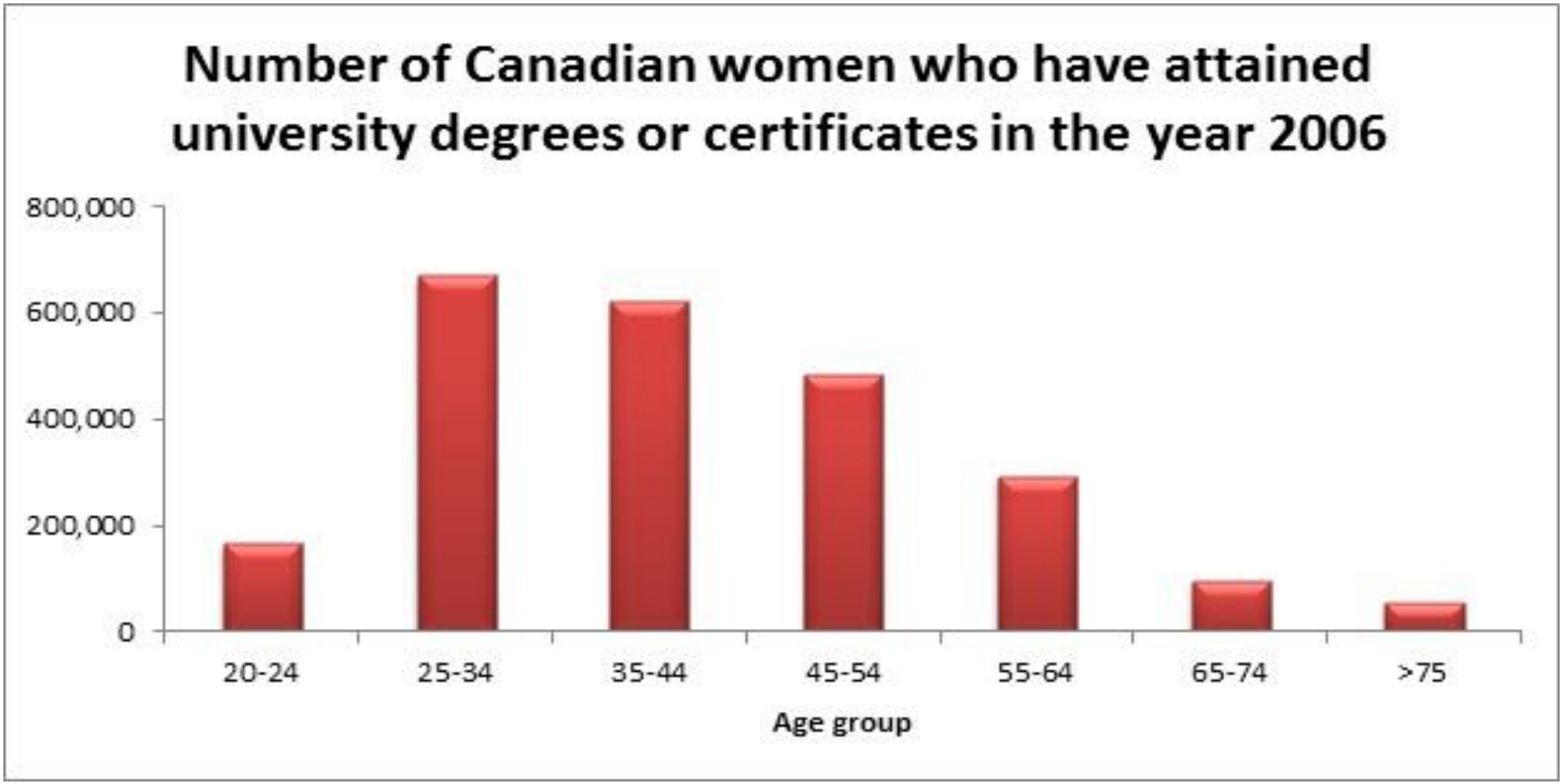
1. Univariate Distributions

- Can express frequencies in graphs as well
  - Many, many types available
  - Type used depends on
    - The level of measurement of the variable you want to display
    - Personal preference

1. Univariate Distributions

Bar Charts/histograms

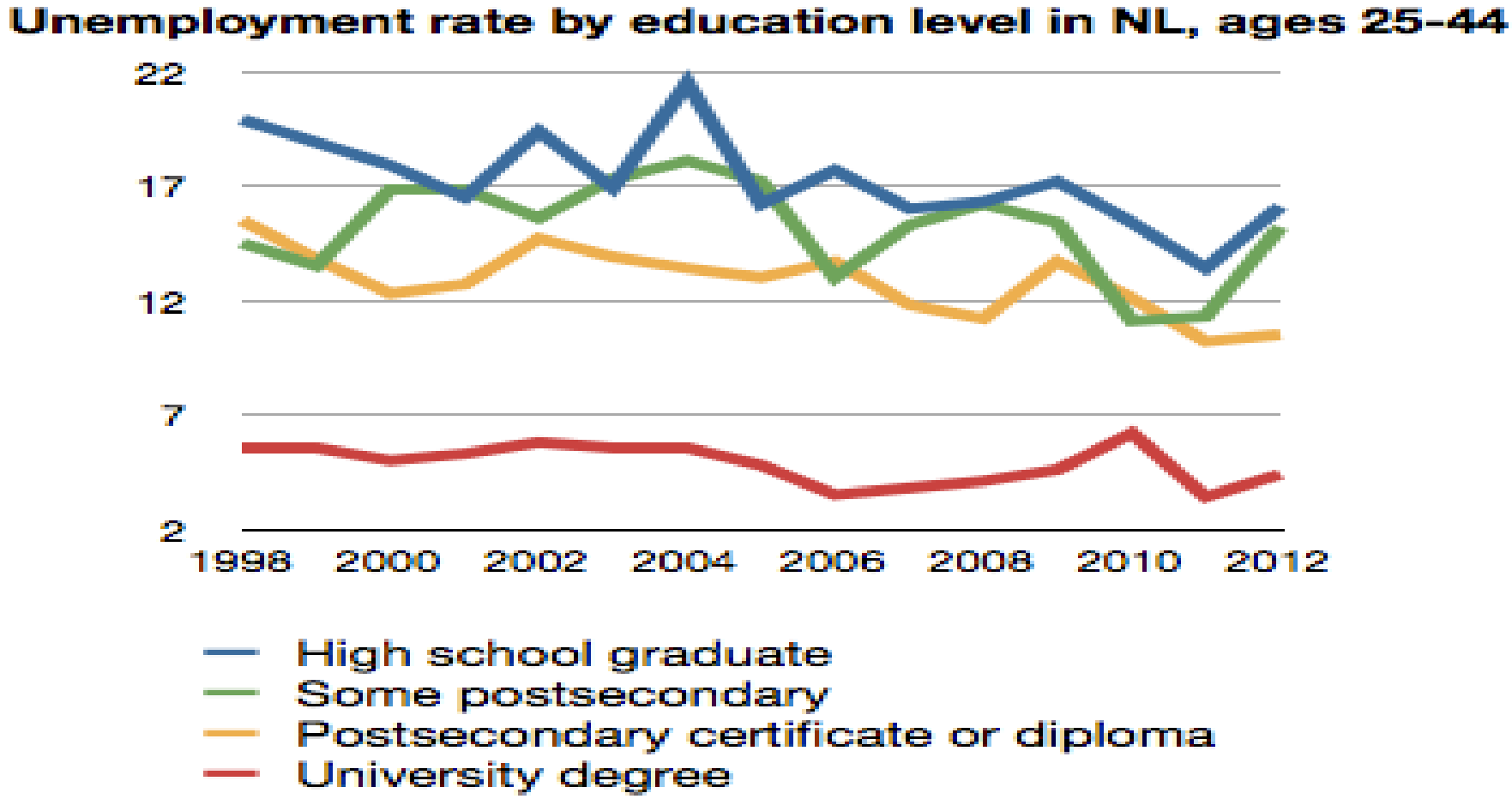
- graph of a frequency distribution
- suitable for nominal and ordinal variables, and interval/ratio variables with a small number of categories



1. Univariate Distributions

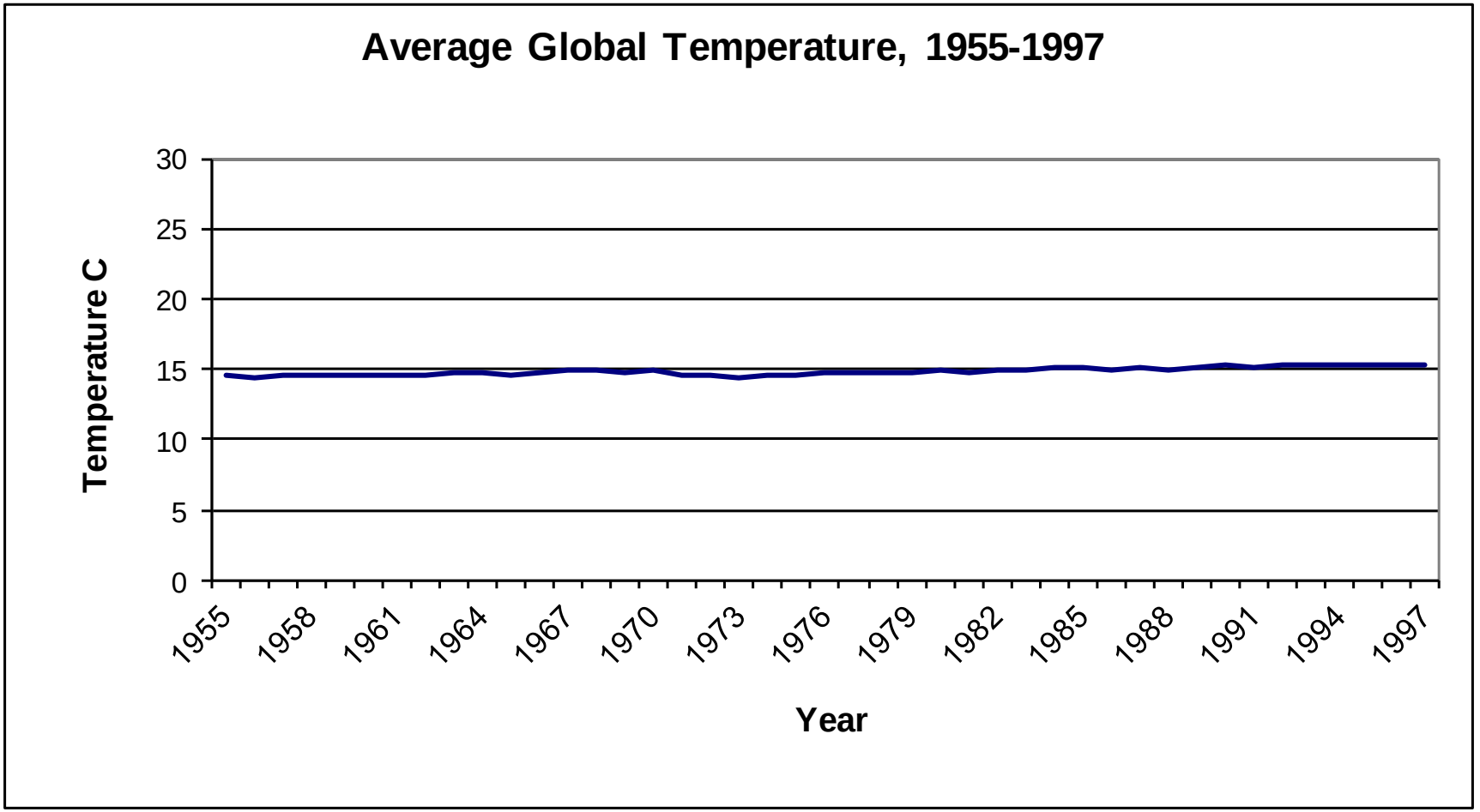
Line Graph/Frequency Polygon

- Suitable for ordinal variables with many categories, and interval/ratio data



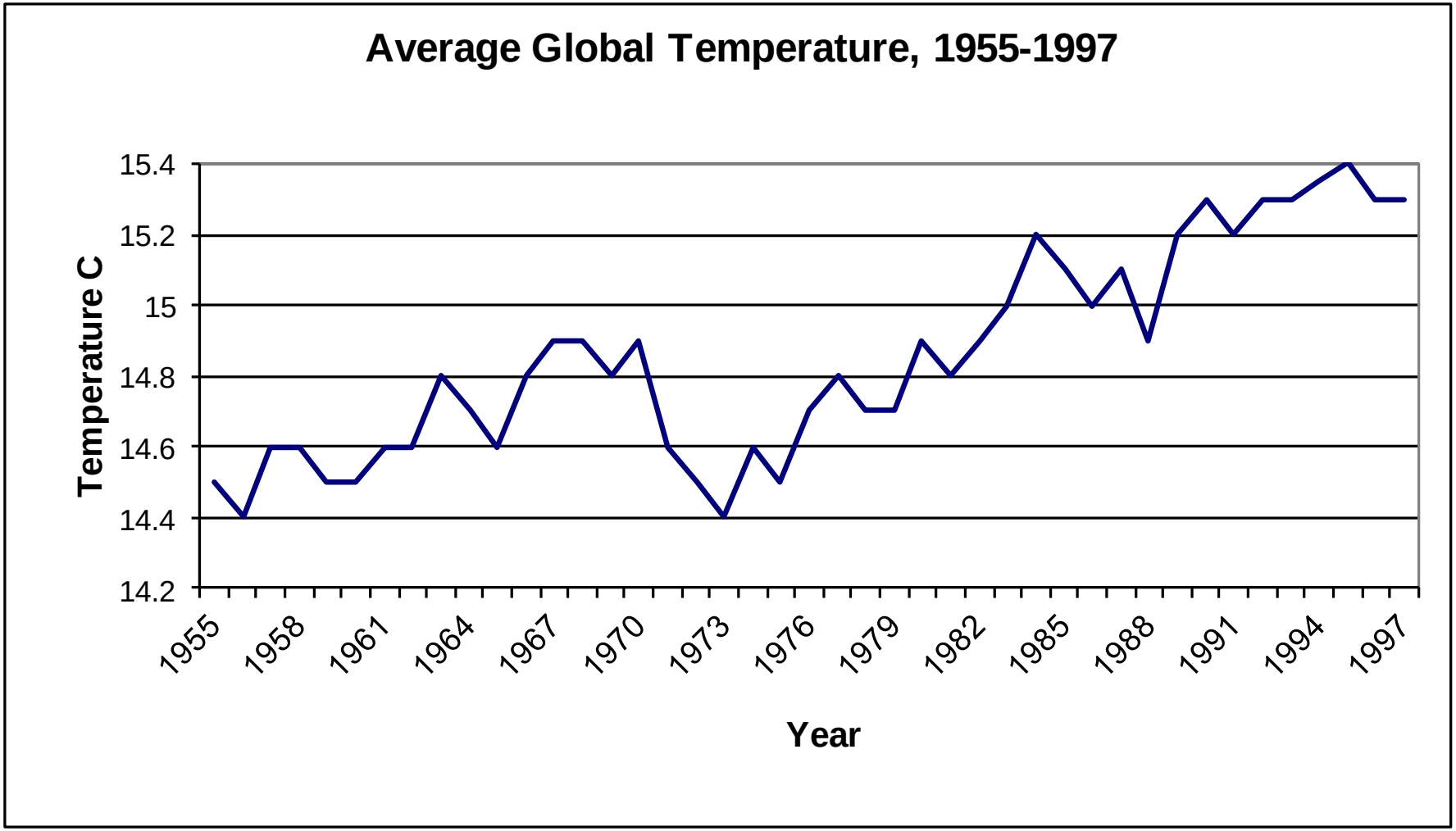
1. Univariate Distributions

- N.B. Lies, damn lies, and statistics! Always be cautious when interpreting graphs !



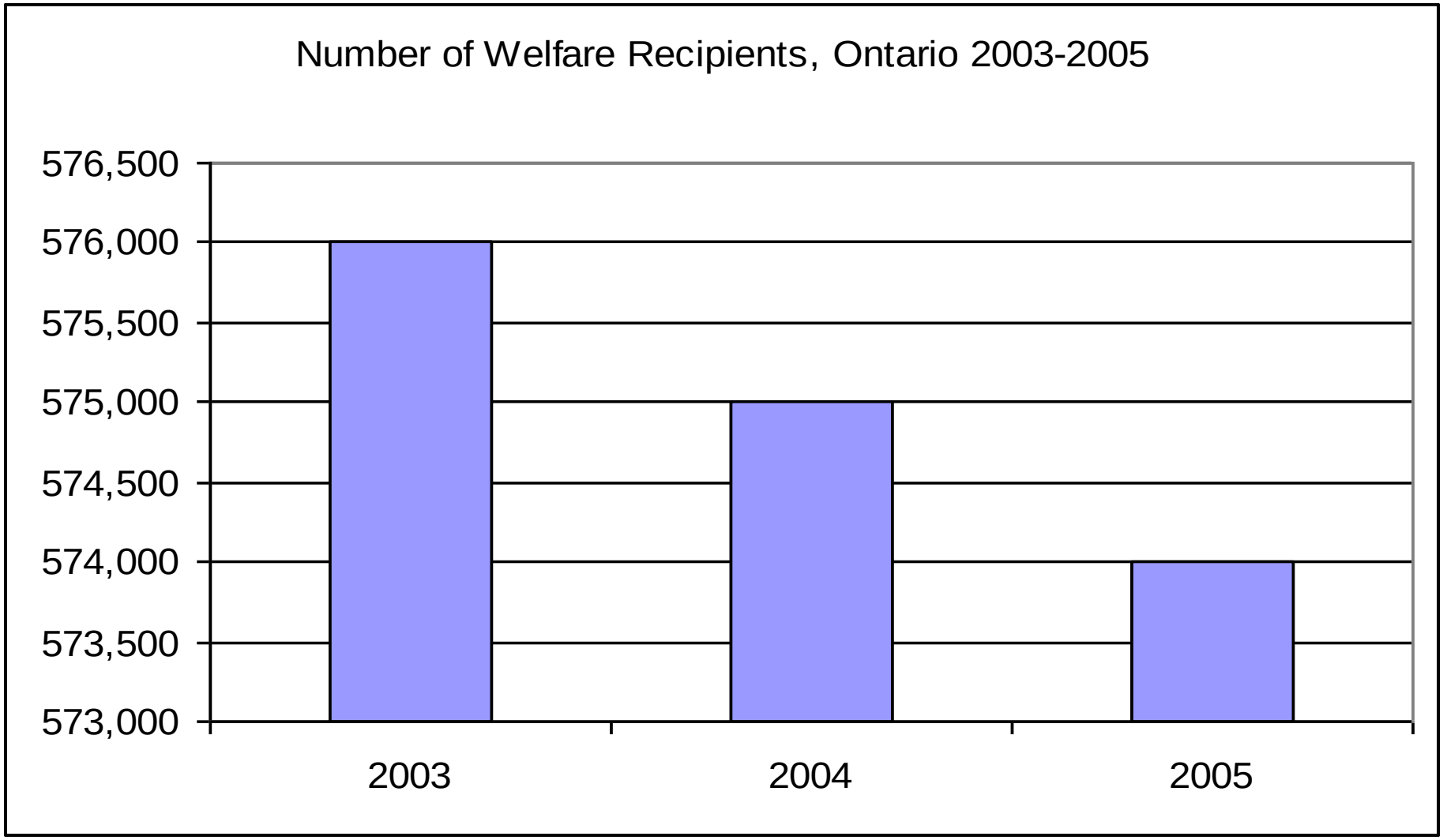
1. Univariate Distributions

- Same data, different scale!
- Try to be honest about this in your own work, as well as watching out for it in others' work



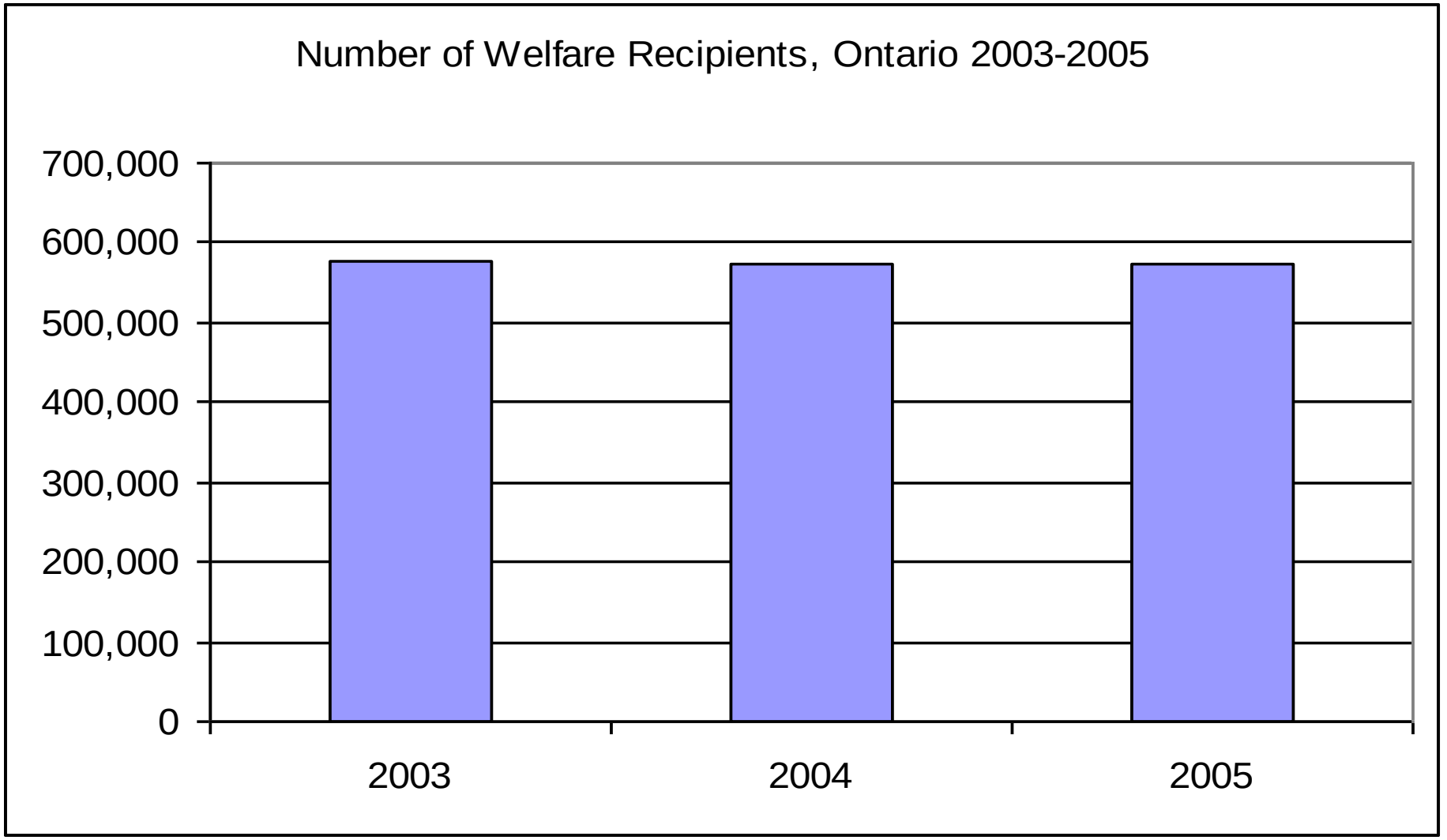
1. Univariate Distributions

- Another example …



1. Univariate Distributions

- Compared to ..



1. Univariate Statistics

- Univariate summary indicators are also available
  - Summarize the variable with a single number
  - Two types:
    - ***measures of central tendency***
      - The 'average' or midrange case
      - Mode, Median, Mean
    - ***measures of variability/dispersion***
      - how widely dispersed the cases are
      - IQV, range, standard deviation

1. Univariate Statistics

Measures of Central Tendency

- ***Mode:*** most common response
  - Here it is the 'Romantic era'
  - For nominal level variables or ordinal level with a very small # of categories

What is your favorite era of classical music?			
	Frequency	% Frequency	
Gregorian	28	7.89	
Renaissanc	47	13.24	
Baroque	49	13.80	
Classical	67	18.87	
Romantic	89	25.07	
Modern	75	21.13	
Total	355	100.00	

1. Univariate Statistics

Measures of Central Tendency

- ***Median***: the middle score
  - half of all respondents fall above the score
  - half fall below

  – In this case, the median is 65

  – Good for any ranked variable
    - Best if # of categories is large

Results from a Test
91
86
86
83
82
79
78
76
72
69
68
67
67
65
61
60
58
58
57
56
56
55
52
50
50
48
42

1. Univariate Statistics

Measures of Central Tendency

- ***Mean:*** arithmetic average
  - sum of the case scores divided by the total number of cases

Rank	Results from a Test
1	91
2	86
3	86
4	83
5	82
6	79
7	78
8	76
9	72
10	69
11	68
12	67
13	67
14	65
15	61
16	60
17	58
18	58
19	57
20	56
21	56
22	55
23	52
24	50
25	50
26	48
27	42
Mean	65.63

1. Univariate Statistics

Measures of central tendency

- Mean and median often quite close
  - 65 and 65.63
  - Except when there are ***case outliers***

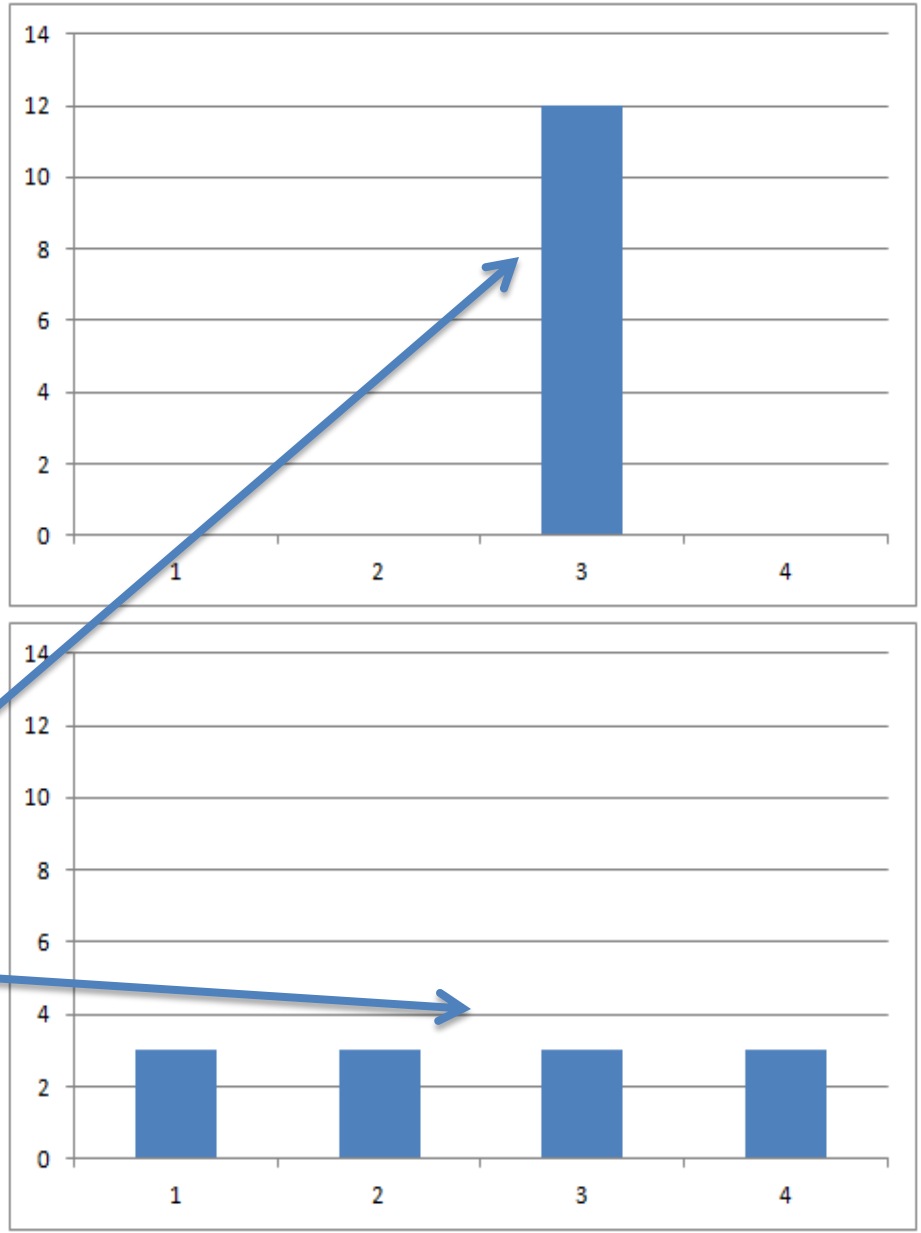
- Ex. say a couple of students had a really hard time and the bottom scores were 5 and 15 instead of 42 and 48
  - Median is still 65
  - But now mean is now 63.04

Rank	Results from a Test
1	91
2	86
3	86
4	83
5	82
6	79
7	78
8	76
9	72
10	69
11	68
12	67
13	67
14	65
15	61
16	60
17	58
18	58
19	57
20	56
21	56
22	55
23	52
24	50
25	50
26	15
27	5
Mean	63.04

1. Univariate Statistics

Measures of variability/dispersion

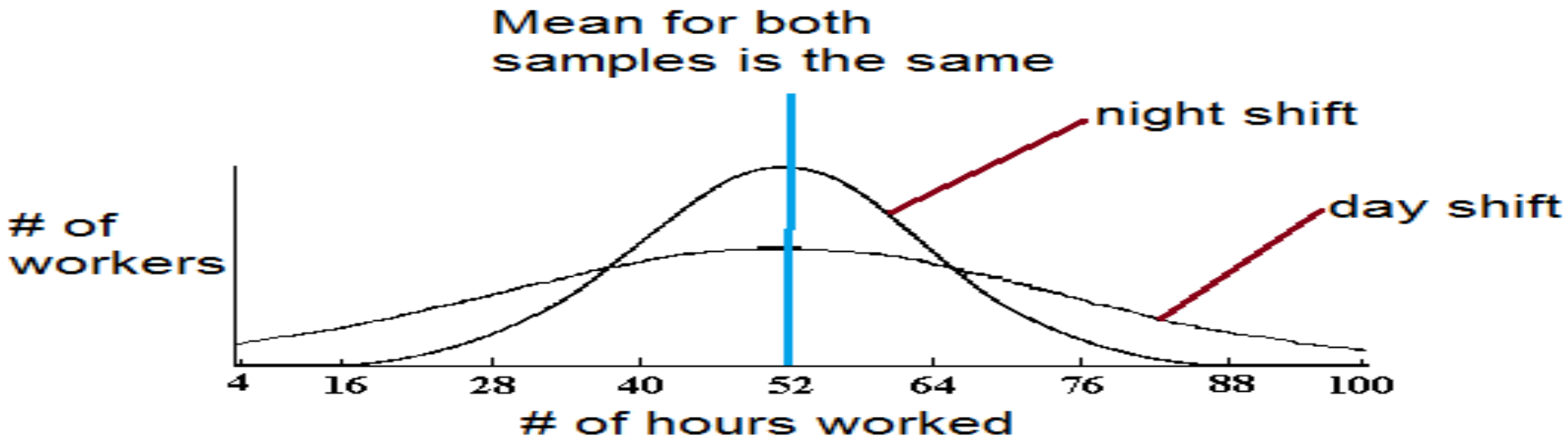
- ***Index of Qualitative Variation (IQV)***
  - For nominal variables
  - Ranges between 0 and 1
  - 0 if all of the cases fall into a single category
  - 1 if all of the cases are equally distributed across categories



1. Univariate Statistics

Measures of variability/dispersion

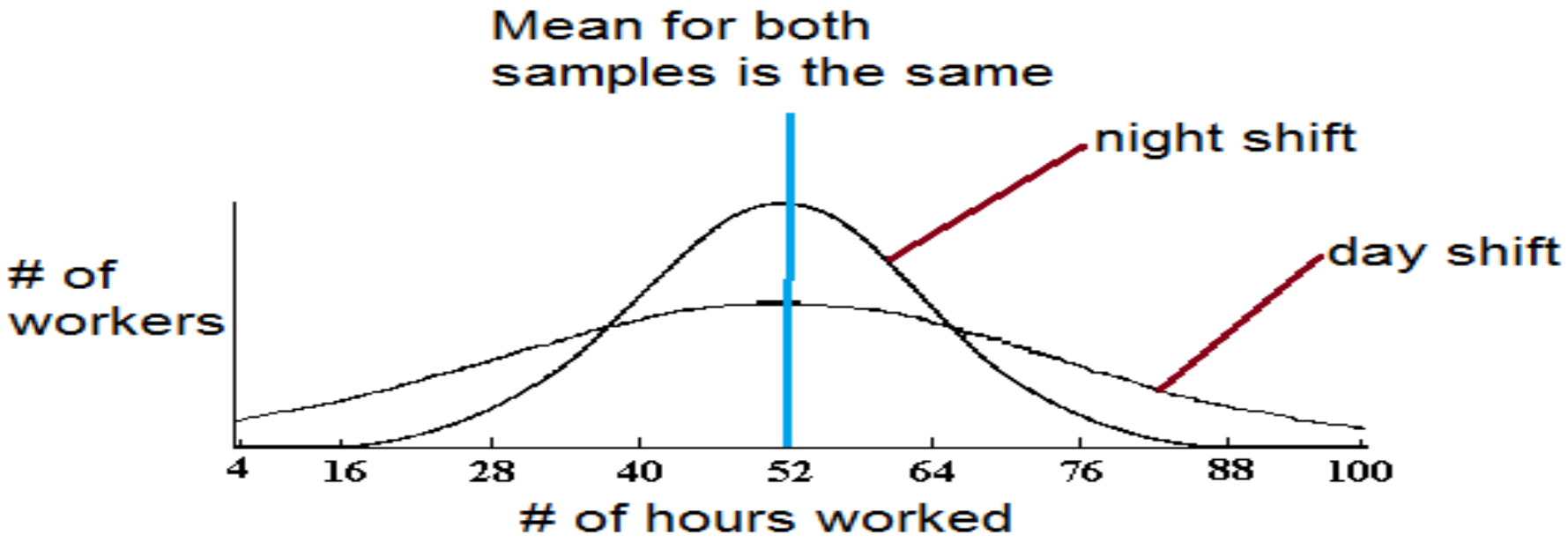
- For interval/ratio variables, and ordinal variables with a large # of categories:
  - we can describe the ***range*** of the variables and *how the scores are distributed about the mean* (the ***standard deviation***)
    - Ex. Night shift and day shift at an auto parts factory



1. Univariate Statistics

Ex. Night shift and day shift at an auto parts factory

- Both shifts work an average of 52 hours per week
- Night shift have smaller ***range*** of hours worked compared to the day shift
  - 88-16 = 72 hours compared to 100-4 = 96 hours
- ***Standard deviation*** (SD) – the higher the standard deviation, the higher the variability about the mean
  - SD is about 15 hours for the day shift, and about 9 hours for the night shift
  - There are more cases further from the mean for the day shift



2. Bivariate Distributions/Statistics

- Distributions
  - contingency tables/cross tabs
- Statistics
  - tests of association

2. Bivariate Distributions/Statistics

Contingency Tables (aka Bivariate Tables, Cross tabulation)

- Useful to produce ***conditional distributions***
  - The distribution of one variable (**Y**), for each category of another (**X**)
  - Here, the distribution of smokers (**Y**) for each category of gender (**X**)

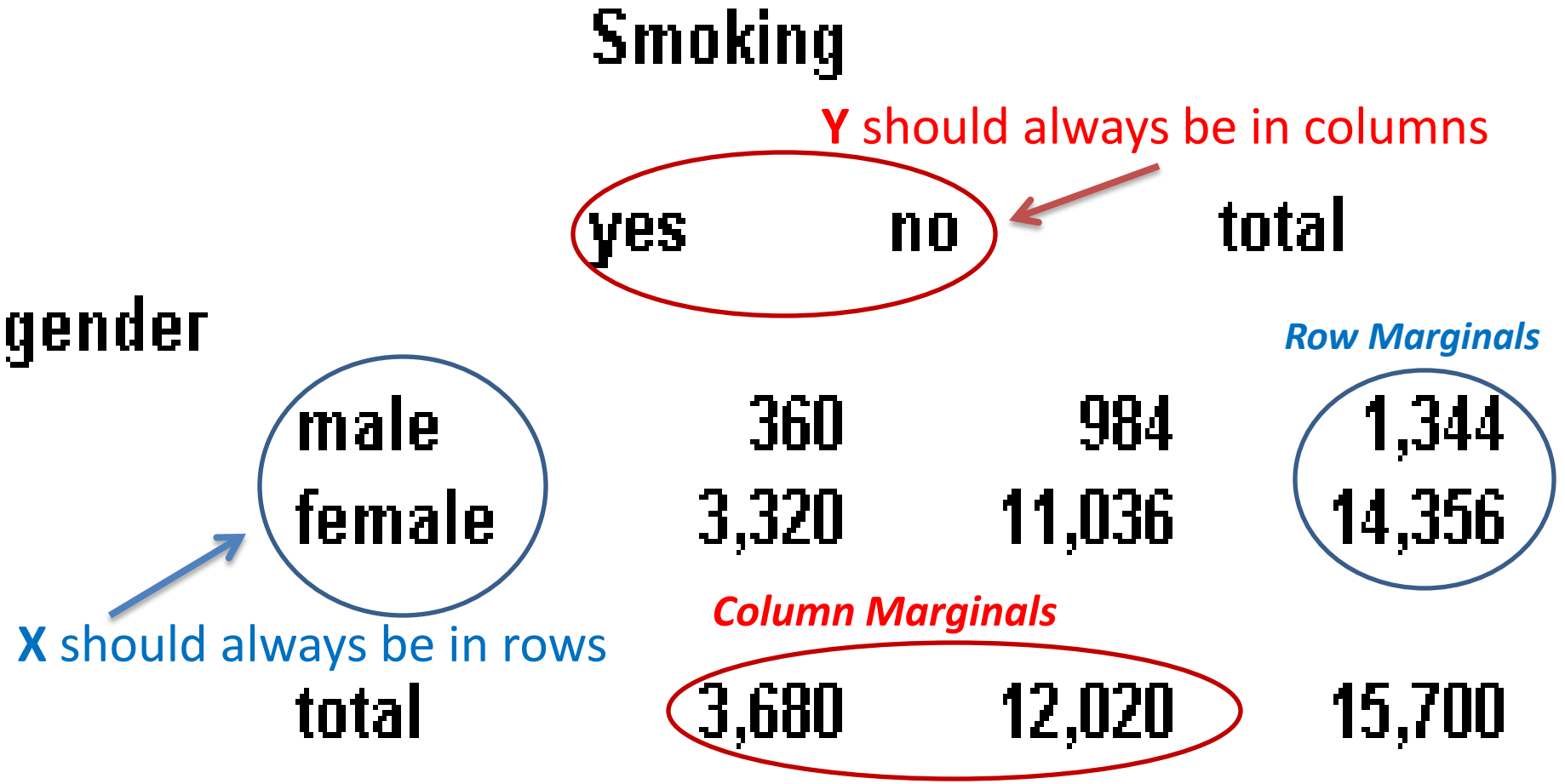
The Relationship Between Smoking and Gender

	Smoking		
gender	yes	no	total
male	360	984	1,344
female	3,320	11,036	14,356
total	3,680	12,020	15,700

2. Bivariate Distributions/Statistics

Contingency Tables

The Relationship Between Smoking and Gender



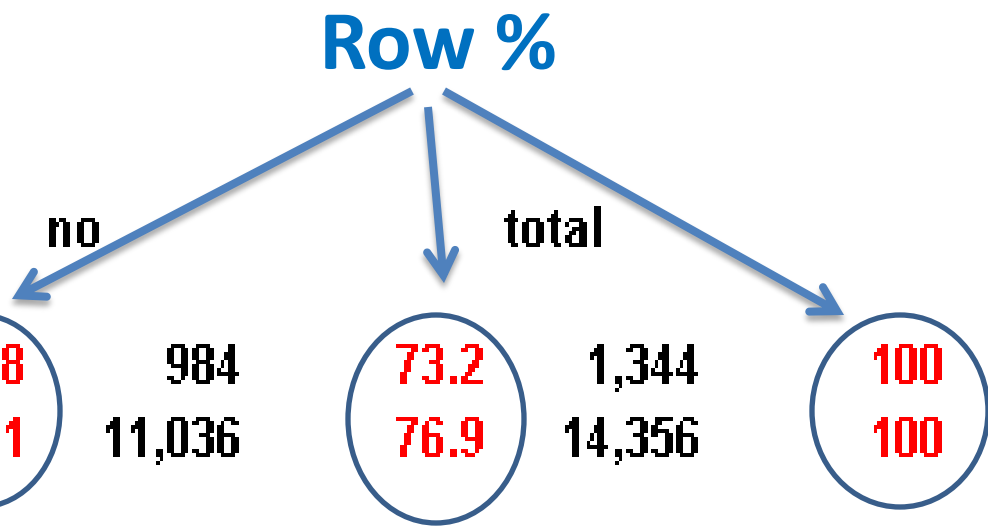
2. Bivariate Distributions/Statistics

Contingency Tables

- Like frequency tables
  - Easier to understand when working with % than raw frequencies
- If table set up properly (**X** in rows), work with ***row %***
  - Here more men smoke than women
    - 26.8% compared to 23.1%

The Relationship Between Smoking and Gender

		Smoking				
		yes	no		total	
gender						
	male	360	26.8	984	73.2	1,344	100
	female	3,320	23.1	11,036	76.9	14,356	100
	total	3,680		12,020		15,700	



2. Bivariate Distributions/Statistics

Contingency Tables

- What about ***column %*** ?
  - distribution of **X** for each category of **Y**?
    - Offers less useful, or misleading information
      - 90.2 % of smokers are female, 9.8% are men
    - Dependent on the distribution of gender in the sample rather than smokers
      - Interpretation is different as well (as though smoking causes gender)

The Relationship Between Smoking and Gender

Column %

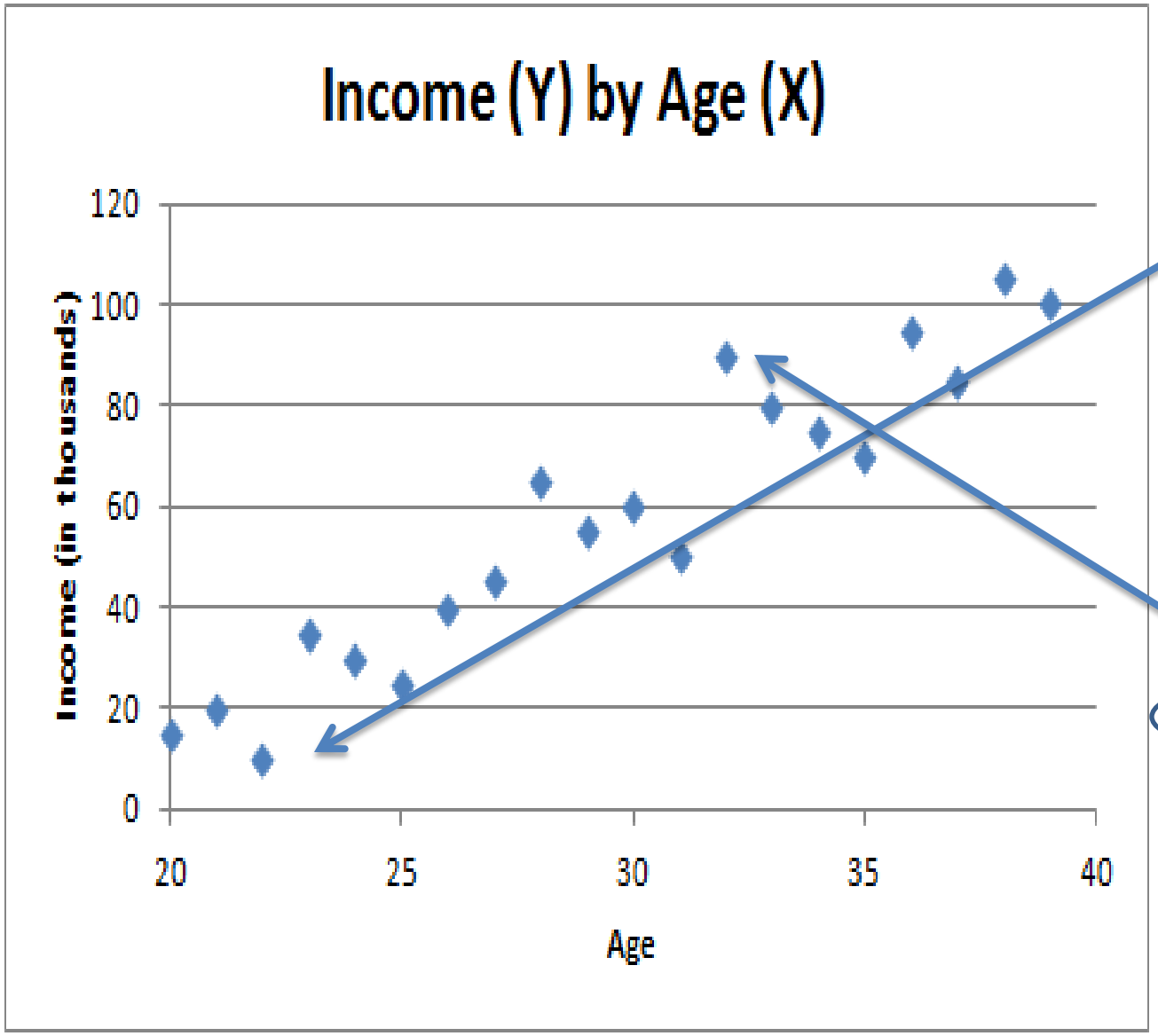
	Smoking				
	yes		no		total
gender					
male	360	9.8	984	8.2	1,344
female	3,320	90.2	11,036	91.8	14,356
total	3,680	100	12,020	100	15,700

2. Bivariate Distributions/Statistics

- Contingency tables are unweildy for interval/ratio (or even large ordinal) variables

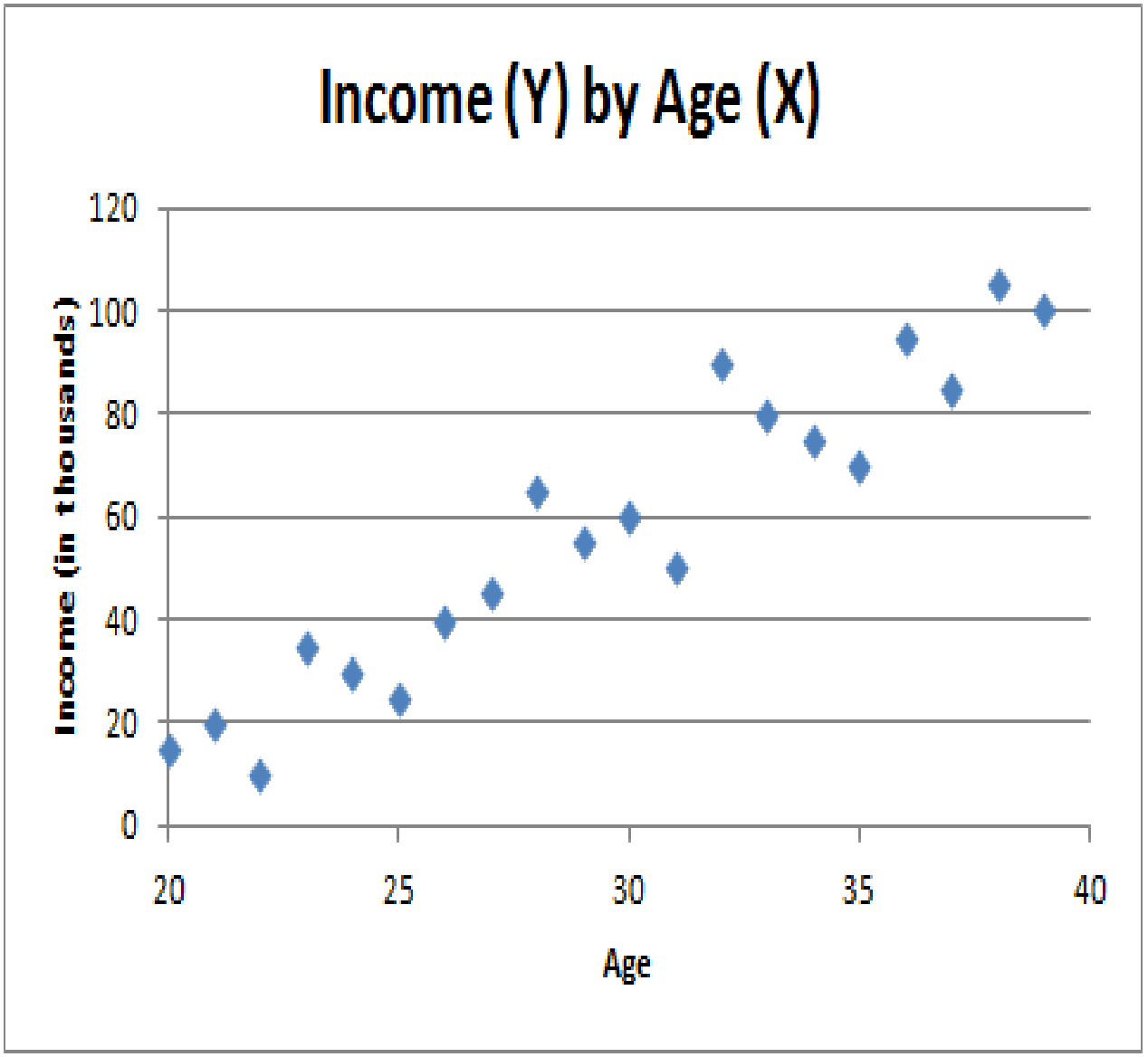
- Scatter plots
  - A simple graphic way to display distributions of **Y** across categories of **X** when there are many categories of each
  - Can not be used with nominal variables

2. Bivariate Distributions/Statistics



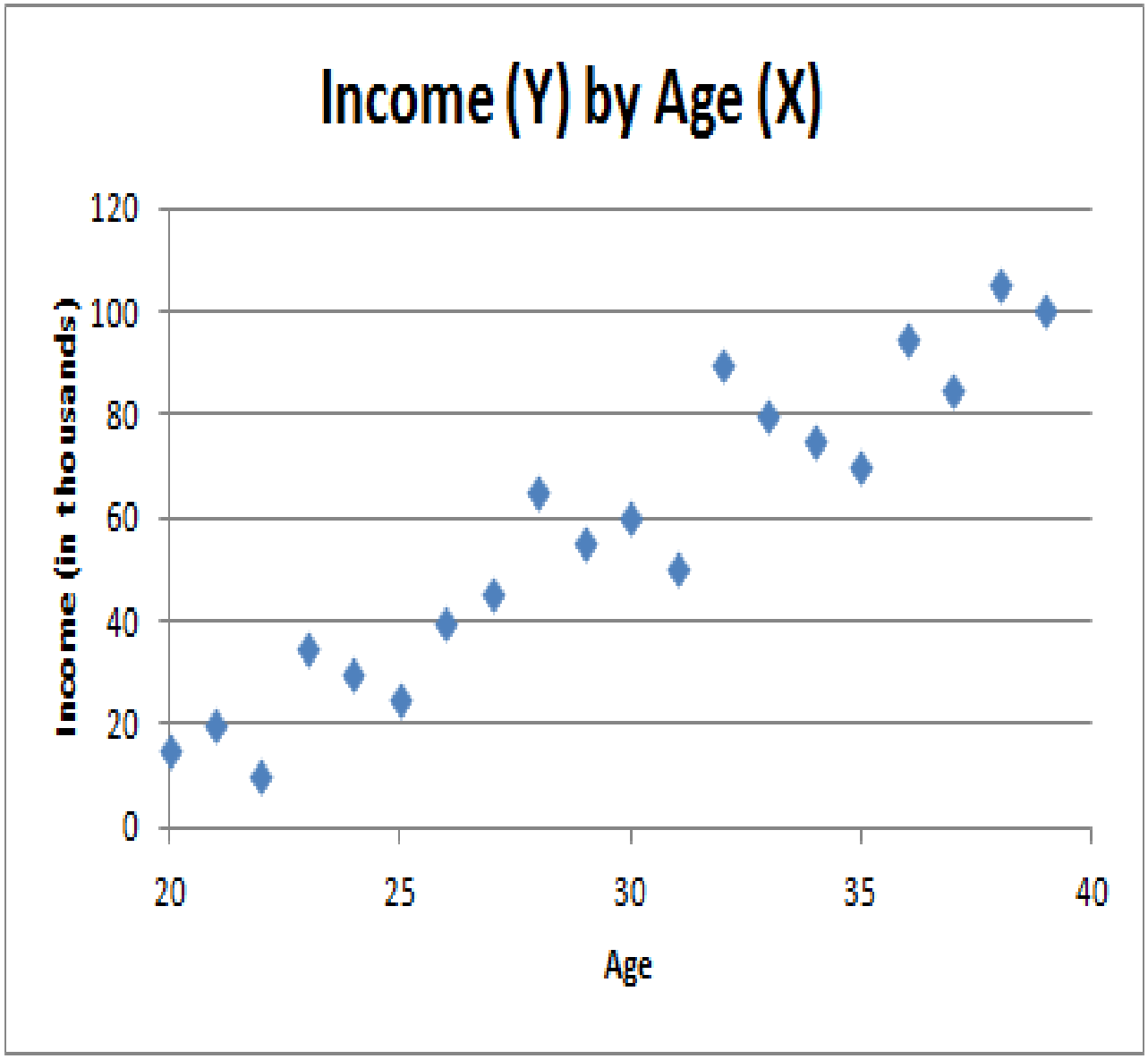
Age	Income (in thousands)
20	15
21	20
22	10
23	35
24	30
25	25
26	40
27	45
28	65
29	55
30	60
31	50
32	90
33	80
34	75
35	70
36	95
37	85
38	105
39	100

2. Bivariate Distributions/Statistics



- Can see change in the distribution of **Y** for each change in **X**
- In this case, as age (**X**) increases, income (**Y**) also increases

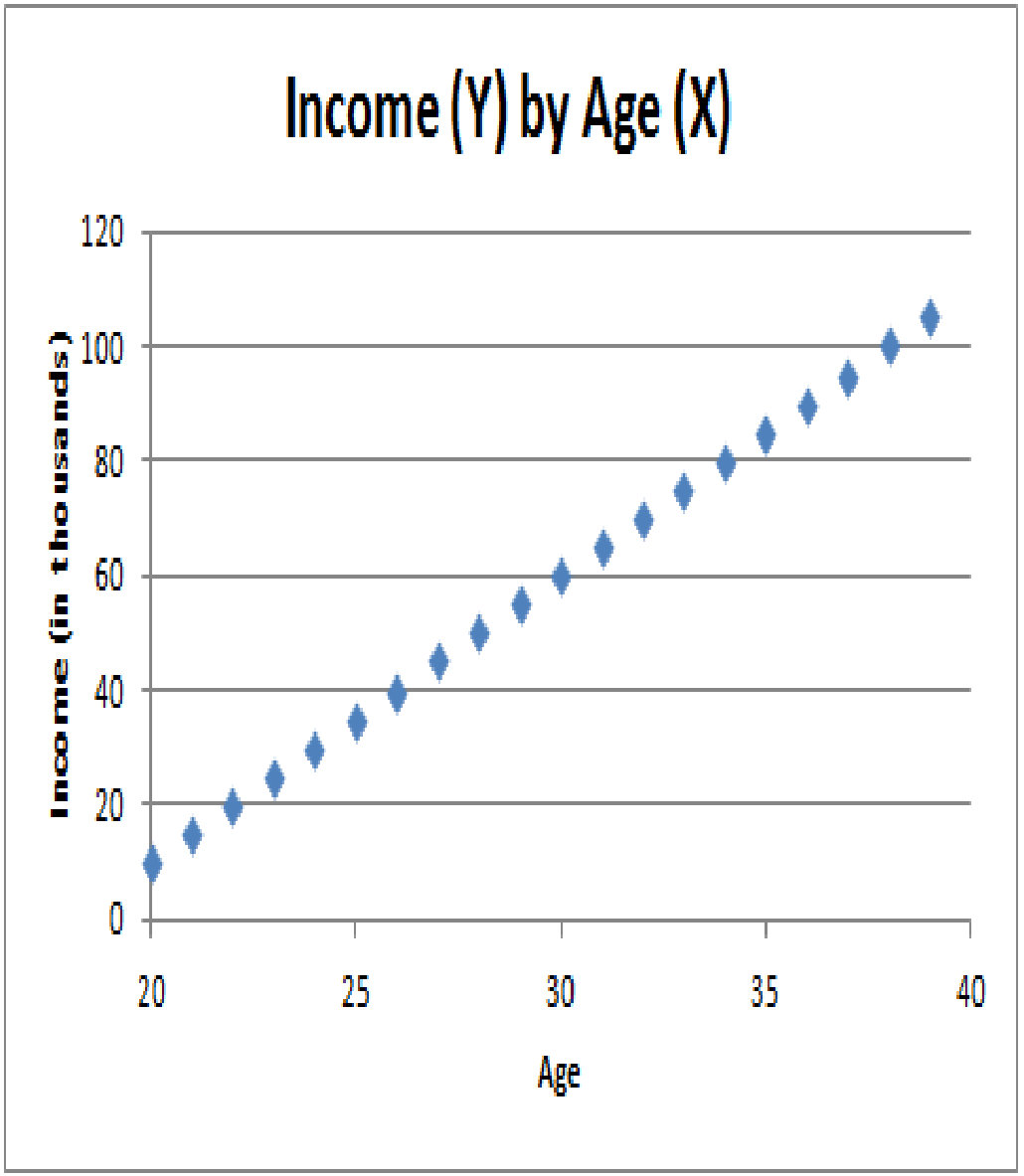
2. Bivariate Distributions/Statistics



- This looks like a pretty ***strong association***
- we can clearly see that as age increases, income increases as well

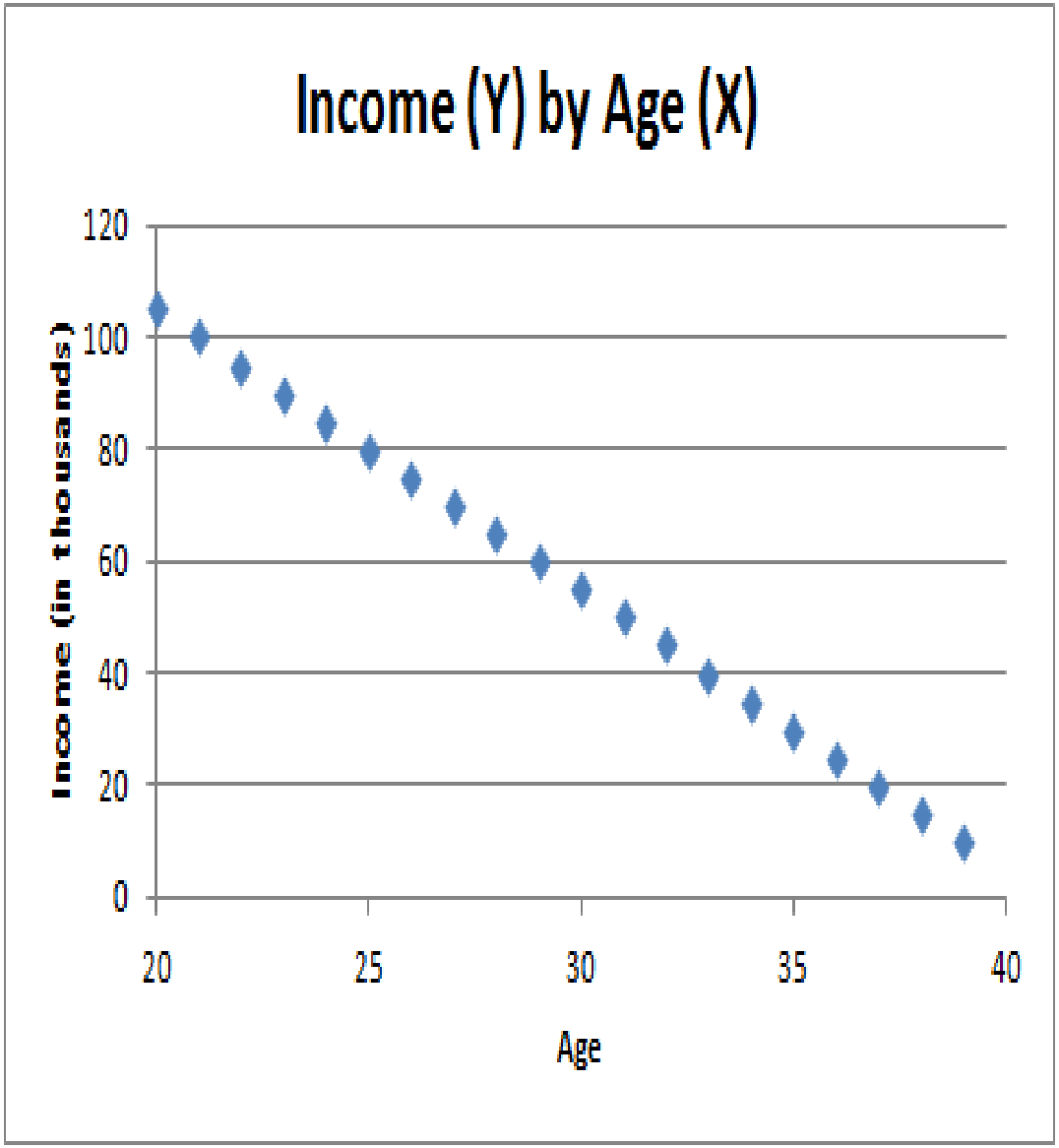
2. Bivariate Distributions/Statistics

- This is a **perfect association**
  - The strongest you can have
  - For each increase in **X** we get an exact increase in **Y**
  - The direction of this association is ***positive***



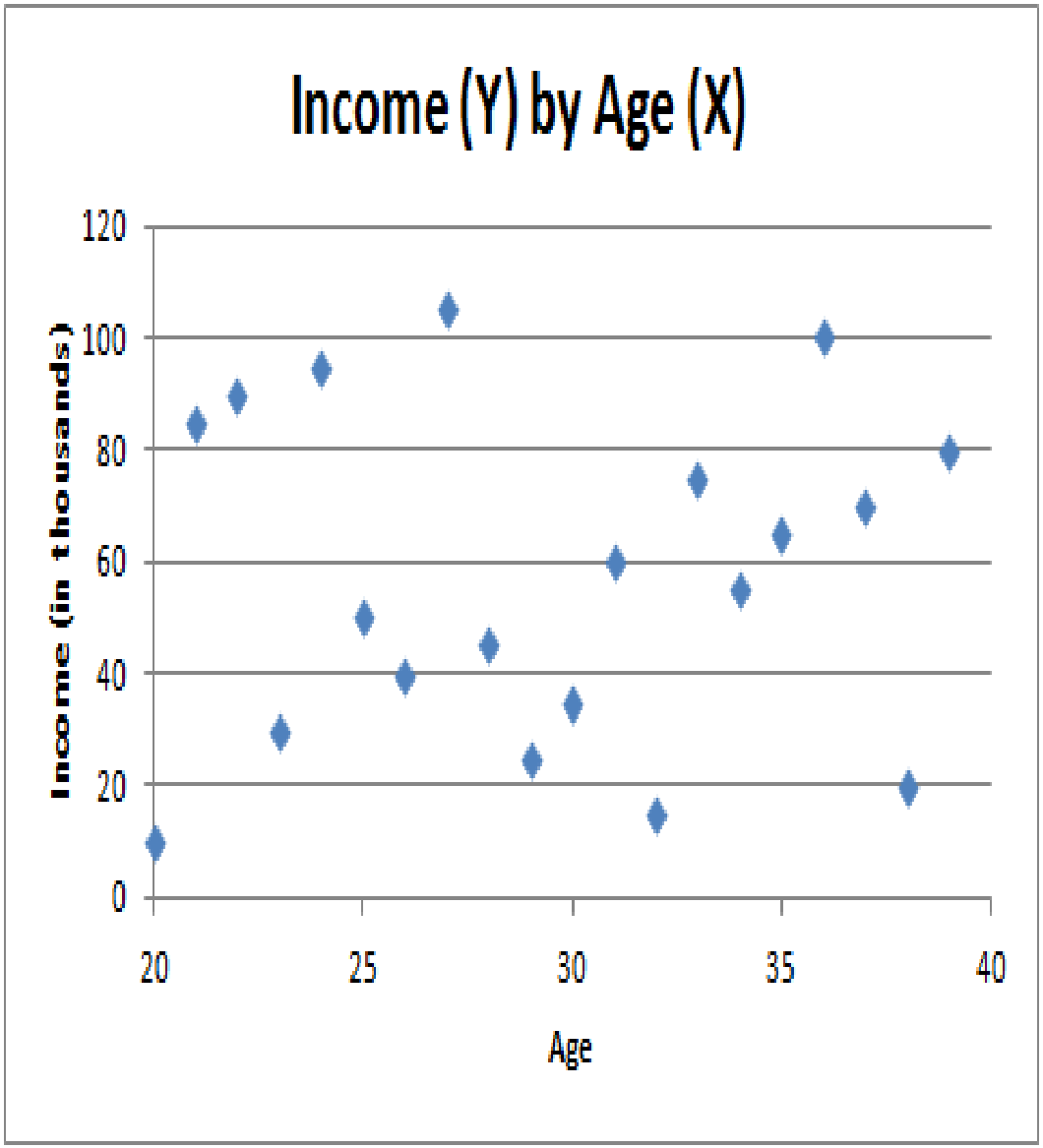
2. Bivariate Distributions/Statistics

- This is another perfect association
  - This time, the direction of the association is ***negative***
    - As **X** increases, **Y** decreases



2. Bivariate Distributions/Statistics

- Here we don't seem to have any association at all



3. Multivariate Relationships and Distributions/Statistics

3. Multivariate (many variables)

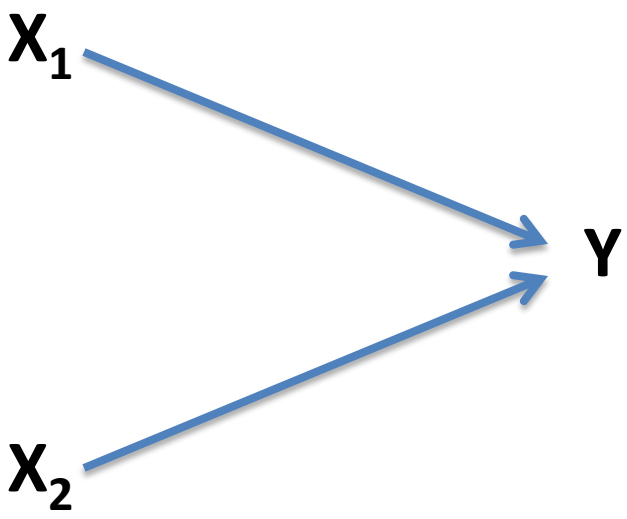
- Types of associations
  - Independent
  - Intervening
  - Interactive
  - Spurious

- Distributions and Statistics
  - 3 way contingency tables
  - Complex models for 3306 and grad school

3. Multivariate Relationships

1) Independent Relationships

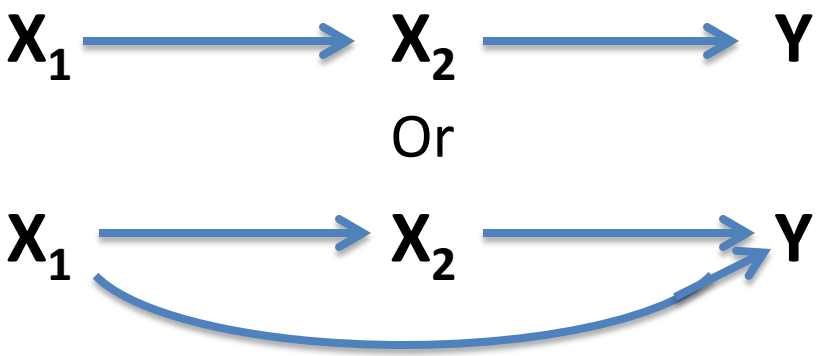
- Each X has an independent effect on Y



3. Multivariate Relationships

2) Intervening Relationships

- All or part of an association between one variable to another is indirect
  - A third variable 'intervenes' between the two

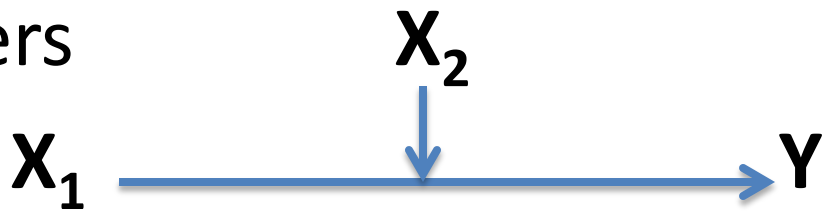


  - Ex.

3. Multivariate Relationships

3) Interactive Relationships

- One variable changes the relationship between two others



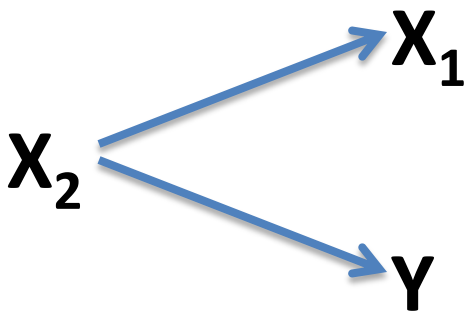
- Effect of X_1 on **Y** varies by category of X_2
  - Ex. The effect of poverty (X_1) on mental health **(Y)** varies by gender (X_2)
    - Maybe poverty effects the mental health of men more than women

3. Multivariate Relationships

4) Spurious Relationships

- a 'false' association between $\mathbf{X_1}$ and **Y**, caused by an antecedent variable ($\mathbf{X_2}$) related to both

this is hiding …



This!