Relationships

Stephen B. Holt, Ph.D.

Rockefeller College
of Public Affairs & Policy
University at Albany State University of New York

October 30, 2022

Summarizing Data

Most analyses begin with summarizing key variables in the sample used in the study. A good description of data will achieve a few purposes:

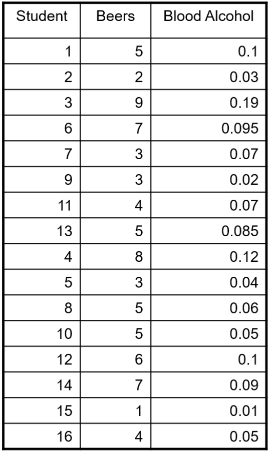
1. Assess generalizability of the sample
2. Assess potential differences between treatment and control status
3. Provide a sense of the central tendency and variation in the primary outcome
4. Examine trends over time to establish the importance of the outcome

Summary Statistics

	All	Low	High
Waiting for services (T in mins.)	1.88	2.24***	1.24
	(15.47)	(16.87)	(10.02)
Waiting for services ($T \mid T > 0$)	39.96	44.69***	28.89
	(59.62)	(61.62)	(39.36)
Travel time for services	27.15	25.69***	30.21
	(44.88)	(44.68)	(47.48)
HH income $20K or less	0.20	1.00	0.00
HH income $150K or more	0.08	0.00	1.00
White	0.82	0.75***	0.85
Black	0.12	0.20***	0.05
$<$ HS diploma	0.17	0.29***	0.10
College degree +	0.29	0.13***	0.63
Num. of HH children	0.78	0.71***	0.88
Observations	210,586	49,688	14,852

Two Variable Example

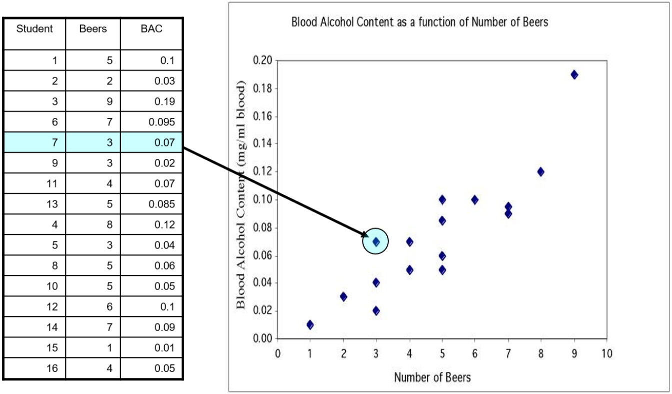
- Here, we have two quantitative variables for each of 16 students.
  1. How many beers they drank, and
  2. Their blood alcohol level (BAC)
- We are interested in the relationship between the two variables: How is one affected by changes in the other one?



Student	Beers	Blood Alcohol
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05

Scatterplots

In a **scatterplot**, one axis is used to represent each of the variables, and the data are plotted as points on the graph.

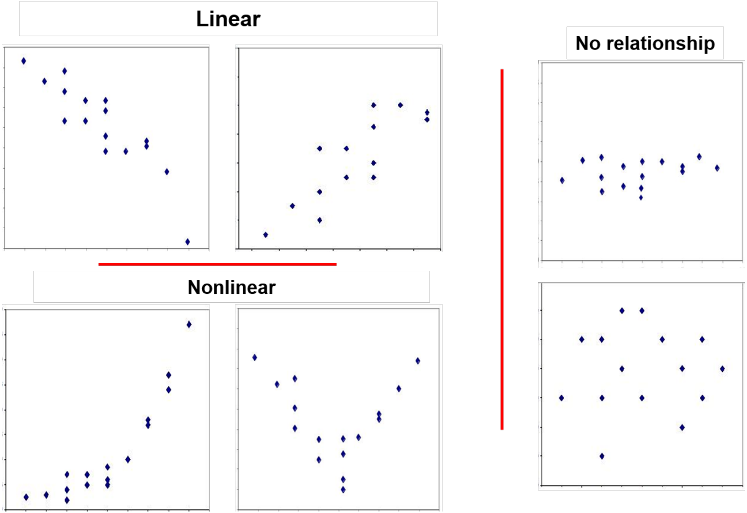


Student	Beers	BAC
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05

Interpreting scatterplots

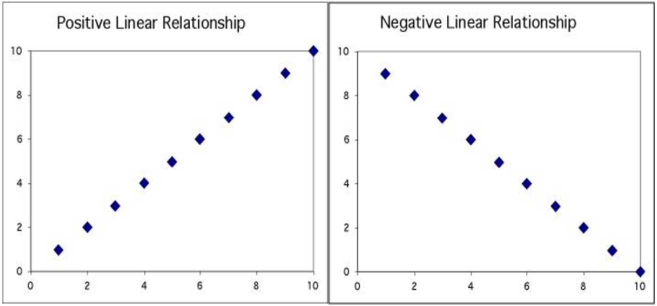
- After plotting two variables on a scatterplot, we describe the relationship by examining the **form**, **direction**, and **strength** of the association. We look for an overall pattern . . .
    - Form: linear, curved, clusters, no pattern
    - Direction: positive, negative, no direction
    - Strength: how closely the points fit the “form”
- . . . and deviations from that pattern.
    - Outliers

Form and Direction of an Association



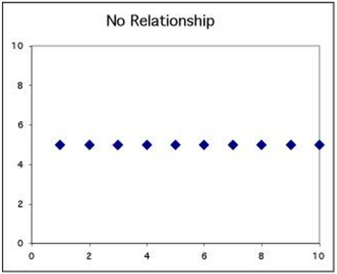
Direction of a Relationship

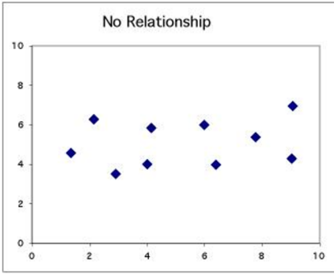
Positive association: High values of one variable tend to occur together with high values of the other variable.
Negative association: High values of one variable tend to occur together with low values of the other variable.



Direction of a Relationship

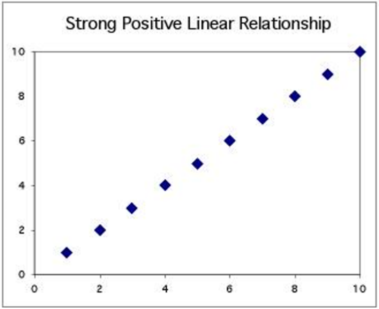
No relationship: X and Y vary independently. Knowing X tells you nothing about Y.

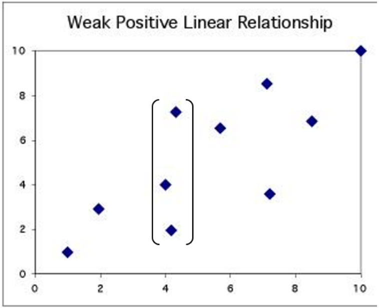




Strength of a Relationship

The **strength** of the relationship between the two variables can be seen by how much variation, or **scatter**, there is around the main form.

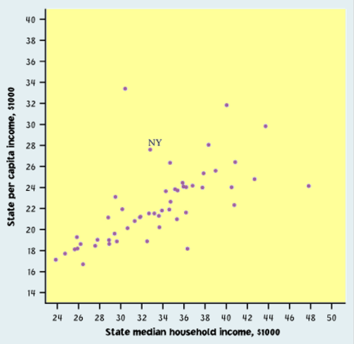


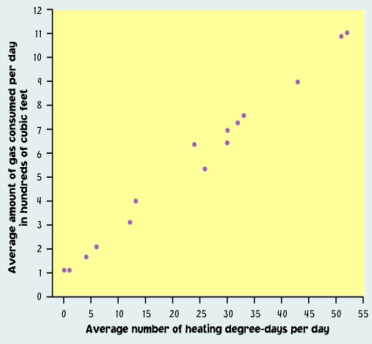


With a strong relationship, you can get a pretty good estimate of y if you know x.

With a weak relationship, for any x you might get a wide range of y values.

Strength of a Relationship



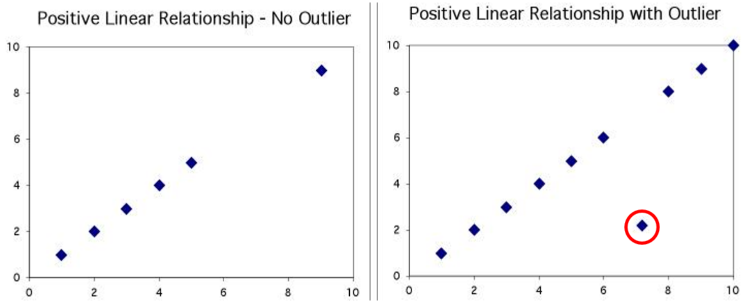


This is a **weak** relationship. For a particular state median household income, you can't predict the state per capita income very well.

This is a **very strong** relationship. The daily amount of gas consumed can be predicted quite accurately for a given temperature value.

Outliers

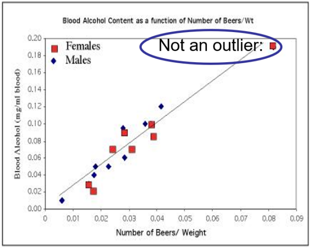
An outlier is a data value that has a very low probability of occurrence (i.e., it is unusual or unexpected).



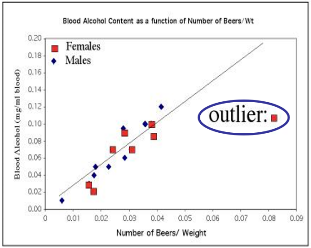
In a scatterplot, outliers are points that fall outside of the overall pattern of the relationship.

Outliers

- The upper right-hand point here is *not* an outlier of the relationship—It is what you would expect for this many beers given the linear relationship between beers/weight and blood alcohol.



- This point is not in line with the others, so it *is* an outlier of the relationship.



Measure of spread: the standard deviation

The standard deviation "s" is used to describe the variation around the mean. Like the *mean*, it is not resistant to skew or outliers.
Recall that there are two steps in the calculation of s: calculate the variance (s^2) and take the square root.

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

The purpose of the standard deviation is to create a measure of spread that is *standardized*. For instance, the range of blood alcohol content values observed in a sample cannot be compared to the range of drinks consumed because they are measured with different units (parts of alcohol per 1000 parts of blood versus a count of beverages). Standard deviations use observations' distance from the average to create a measure of spread comparable across variables (i.e., a standard deviation increase has the same interpretation for both BAC and beers).

Correlation Coefficient: Pearson's "r"

The Pearson's "r" provides a way to more precisely measure the relationship between two variables with a measure that can be compared across relationships.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i - \overline{x}}{s_x})(\frac{y_i - \overline{y}}{s_y}) \tag{2}$$

1. The steps to calculating the r coefficient begins with computing the mean and standard deviation of both variables you believe are related.
2. Then calculate the percent of a standard deviation each observation falls on both variables.
3. Multiplying these together for each variable provides a measure of the relationship between x and y for each observation in the sample.
4. Adding these factors and dividing them by the degrees of freedom $(n-1)$ provides the average strength of the relationship between x and y in the sample, or the r coefficient.

Example of Pearson's R: Airfare

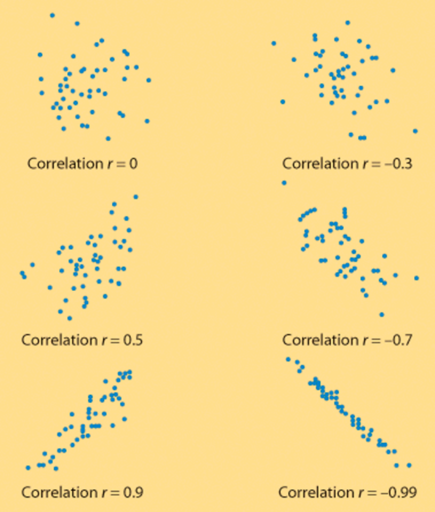
What's the relationship between the price of a plane ticket and the distance of the flight?

Sample: 15 flights
Avg. distance ($\overline{x}$): 1145 miles
s_x: 706.8 miles
Avg. price ($\overline{y}$): \$218.7
s_y: \$61.93
r ($\frac{\sum times}{n-1}$): 0.70
...a strong relationship!

idn	dist (x)	fare (y)	(x-xbar)/s	(y-ybar)/s	times
1	2310	361	1.65	2.30	3.79
2	656	132	-0.69	-1.40	0.97
3	904	274	-0.34	0.89	-0.30
4	444	182	-0.99	-0.59	0.59
5	2458	271	1.86	0.84	1.57
6	1050	153	-0.13	-1.06	0.14
7	1710	210	0.80	-0.14	-0.11
8	624	183	-0.74	-0.58	0.43
9	957	213	-0.27	-0.09	0.02
10	1334	167	0.27	-0.84	-0.22
11	444	186	-0.99	-0.53	0.52
12	769	203	-0.53	-0.25	0.14
13	810	253	-0.47	0.55	-0.26
14	453	190	-0.98	-0.46	0.45
15	2254	303	1.57	1.36	2.13

Visualizing r coefficients

- "r" quantifies the **strength** and **direction** of a linear relationship between 2 quantitative variables.
- **Strength**: how closely the points follow a straight line.
- **Direction**: is positive when individuals with higher X values tend to have higher values of Y.



Lurking Variables

A **lurking variable** is a variable not included in the study design that does have an effect on the variables studied.

Lurking variables can *falsely suggest* a relationship.

What is the lurking variable in these examples?

How could you answer if you didn't know anything about the topic?

Examples:

- Strong positive association between number of firefighters at a fire site and the amount of damage a fire does.
- Negative association between moderate amounts of wine drinking and death rates from heart disease in developed nations.

Categorical Variables

Categorical variables don't necessarily make sense in scatter plots. Observations stack into a limited number of values, and often those values stand-in for a different meaning than the number represented in the dataset (e.g., race or generation or education level).

Often, researchers are interested in the relationship between two categorical variables. For instance, have education levels changed across generations?

To answer this question, a researcher would use a two-way, or block, study design. A two-way design uses two categorical factors with several levels for both factors to answer the question. Here, generations are often defined using categories of ages (a proxy for birth cohorts) and education can be categorized by the highest degree a person has completed.

Two-way Tables

The researcher would descriptively answer the research question using a **two-way table**.
First factor, age grouping, defines the columns.
Second factor, education level, defines the rows.

			Age Group		
Education level	18 to 24	25 to 34	35 to 54	55 or older	Total
Less than HS	26994	26698	69389	116669	239750
HS diploma	123462	116768	258297	428349	926876
Some college	94738	94191	181058	223464	593451
4-year college+	30534	146423	284465	309227	770649
Total	275728	384080	793209	1077709	2530726

Reading Two-Way Tables

- We call education the **row variable** and age group the **column variable**.
- Each combination of values for these two variables is called a cell.
- For each cell, we can compute a proportion by dividing the cell entry by the total sample size. The collection of these proportions would be the **joint distribution** of the two variables.

			Age Group		
Education level	18 to 24	25 to 34	35 to 54	55 or older	Total
Less than HS	26994	26698	69389	116669	239750
HS diploma	123462	116768	258297	428349	926876
Some college	94738	94191	181058	223464	593451
4-year college+	30534	146423	284465	309227	770649
Total	275728	384080	793209	1077709	2530726

Marginal Distributions

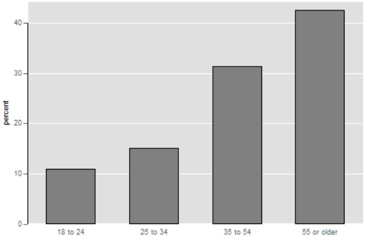
We can look at each categorical variable separately in a two-way table by studying the row totals and the column totals. They represent the **marginal distributions**, expressed in counts or percentages. (They are written as if in a margin.)

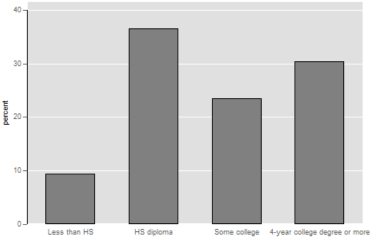
			Age Group		
Education level	18 to 24	25 to 34	35 to 54	55 or older	**Total**
Less than HS	26994	26698	69389	116669	**239750**
HS diploma	123462	116768	258297	428349	**926876**
Some college	94738	94191	181058	223464	**593451**
4-year college+	30534	146423	284465	309227	**770649**
Total	*275728*	*384080*	*793209*	*1077709*	2530726

Marginal Distributions

When we use bar graphs (or pie graphs) to show the distribution of a categorical variable, it captures the equivalent of the marginal distribution of that variable, and the marginal distribution is typically expressed in terms of percent of the total rather than a strict count of observations.

	Age Group				
Education level	18 to 24	25 to 34	35 to 54	55 or older	**Total**
Less than HS	26994	26698	69389	116669	**239750**
HS diploma	123462	116768	258297	428349	**926876**
Some college	94738	94191	181058	223464	**593451**
4-year college+	30534	146423	284465	309227	**770649**
Total	*275728*	*384080*	*793209*	*1077709*	2530726





Conditional Distribution

- In the table below, the 25 to 34 age group occupies the second column. To find the complete distribution of education in this age group, look only at that column. Compute each count as a percent of the column total.
- These percents should add up to 100% because all persons in this age group fall into one of the education categories. These four percents together are the conditional distribution of education, given the 25 to 34 age group.

		Age Group			
Education level	18 to 24	25 to 34	35 to 54	55 or older	Total
Less than HS	26994	26698	69389	116669	239750
HS diploma	123462	116768	258297	428349	926876
Some college	94738	94191	181058	223464	593451
4-year college+	30534	146423	284465	309227	770649
Total	275728	384080	793209	1077709	2530726

Conditional Distributions

The percents within the table represent the conditional distributions. Comparing the conditional distributions allows you to describe the "relationship" between both categorical variables. $C.D. = \frac{cell}{columntotal}$

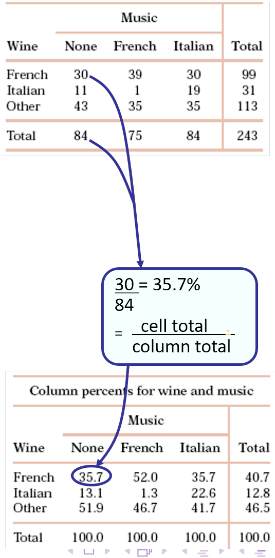
	Age Group				
Education	18 to 24	25 to 34	35 to 54	55 or older	Total
Less than HS	26994	26698	69389	116669	239750
	(9.79)	(6.95)	(8.75)	(10.83)	(9.47)
HS diploma	123462	116768	258297	428349	926876
	(44.78)	(30.40)	(32.56)	(39.75)	(36.62)
Some college	94738	94191	181058	223464	593451
	(34.36)	(24.52)	(22.83)	(20.74)	(23.45)
4-year college+	30534	146423	284465	309227	770649
	(11.07)	(38.12)	(35.86)	(28.69)	(30.45)
Total	275728	384080	793209	1077709	2530726
	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)

Example

	Level of Student				
Pet preferences	Freshmen	Sophomore	Junior	Senior	Total
Cat	0	3	3	2	8
	(0.00)	(75.00)	(37.50)	(33.33)	(40.00)
Dog	2	0	3	4	9
	(100.00)	(0.00)	(37.50)	(66.67)	(45.00)
Fish	0	1	0	0	1
	(0.00)	(25.00)	(0.00)	(0.00)	(5.00)
Other	0	0	1	0	1
	(0.00)	(0.00)	(12.50)	(0.00)	(5.00)
Reptile	0	0	1	0	1
	(0.00)	(0.00)	(12.50)	(0.00)	(5.00)
Total	2	4	8	6	20
	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)
N	20				

Music and Wine Purchase Decisions

- What is the relationship between type of music played in supermarkets and type of wine purchased?
- We want to compare the conditional distributions of the response variable (wine purchased) for each value of the explanatory variable (music played). Therefore, we calculate column percents.
- Calculations: When no music was played, there were 84 bottles of wine sold. Of these, 30 were French wine. $30/84 = 0.357 \rightarrow 35.7\%$ of the wine sold was French when no music was played.



	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

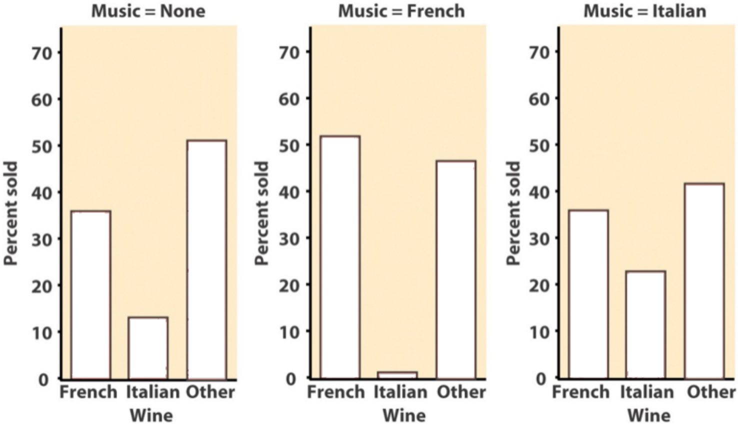
$\frac{30}{84} = 35.7\%$

$= \frac{\text{cell total}}{\text{column total}}$

Column percents for wine and music

	Music			
Wine	None	French	Italian	Total
French	35.7	52.0	35.7	40.7
Italian	13.1	1.3	22.6	12.8
Other	51.9	46.7	41.7	46.5
Total	100.0	100.0	100.0	100.0

Does background music affect wine purchases?



Caution with Association

- As we introduced last week, associations can be biased. This is true for categorical variables as well. Simpson's paradox provides one example of how relationships alone can be unintentionally misleading.
- **Simpson's Paradox**: An association or comparison that holds for all of several groups can reverse direction when the data are combined (aggregated) to form a single group.

	Day 1	Day 2	Total
Person A	63/90 (70%)	4/10 (40%)	67/100 (67%)
Person B	8/10 (80%)	45/90 (50%)	53/100 (53%)

Simpon's Paradox Examples

- Some analyses show men accepted to colleges at higher rates then women. However, each college accepts a higher share of women than men.
- A political party can receive more overall votes in a state and still lose the majority of individual districts in the state legislature.
- Generally, these incidents have to do with how much weight (i.e., the relative number of observations) a particular category has in an analysis.