

Relationship Between Variables

More EDA

Women in Data Science Conference (WIDS)

UCR will host a regional event on March 2nd (more info to follow)

Consider participating in the Datathon <https://www.widsconference.org/datathon.html>

Reminders

Did you fill-out the Google sheet with your team and github account info??

Did you submit your project proposal??

- Have you identified the website , datasets, APIs you plan to use??

Midterm part 1 on Friday 2/7 (a short mini quiz really)

Midterm part 2 during lab on Tuesday and Thursday (short notebook assignment to be turned in by end of lab).

Correlation Test


- When performing EDA, we would like to examine the relationship between variables.
- We can usually infer relationships between variables by looking at various visualizations (like scatter plots).
- We can also examine the type and strength of a relationship using basic statistics. Lets take a look at some approaches today

Part 1 - Correlation between quantitative variables

- **Data Description:** A Pennsylvania research firm conducted a study in which 30 drivers (of ages 18 to 82 years old) were sampled, and for each one, the maximum distance (in feet) at which he/she could read a newly designed sign was determined. The goal of this study was to explore the relationship between a driver's age and the maximum distance at which signs were legible, and then use the study's findings to improve safety for older drivers. (Reference: Utts and Heckard, Mind on Statistics (2002). Originally source: Data collected by Last Resource, Inc, Bellfonte, PA.)

Here is what the raw data looks like:

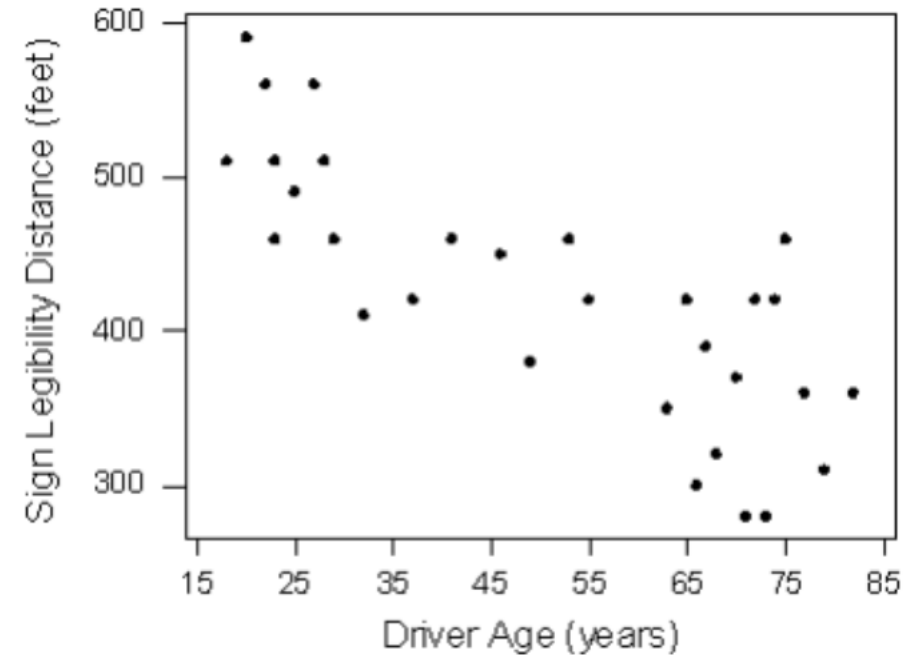
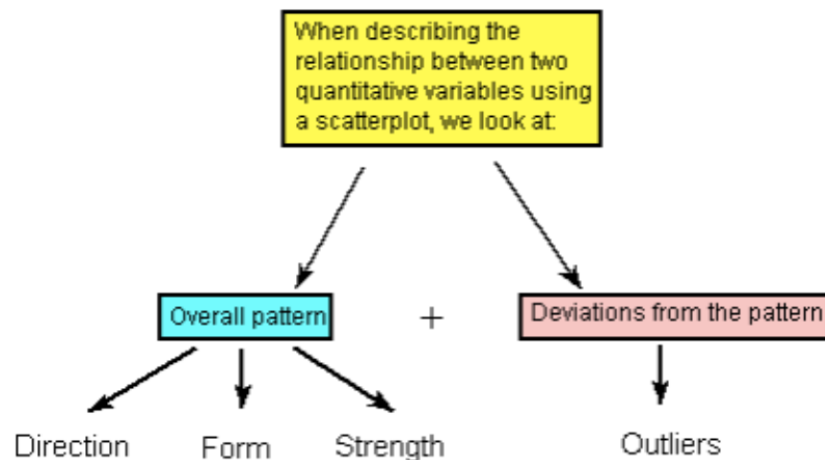
- the explanatory (X) variable is Age, and
- the response (Y) variable is Distance



	Age	Distance
Driver 1	18	510
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360

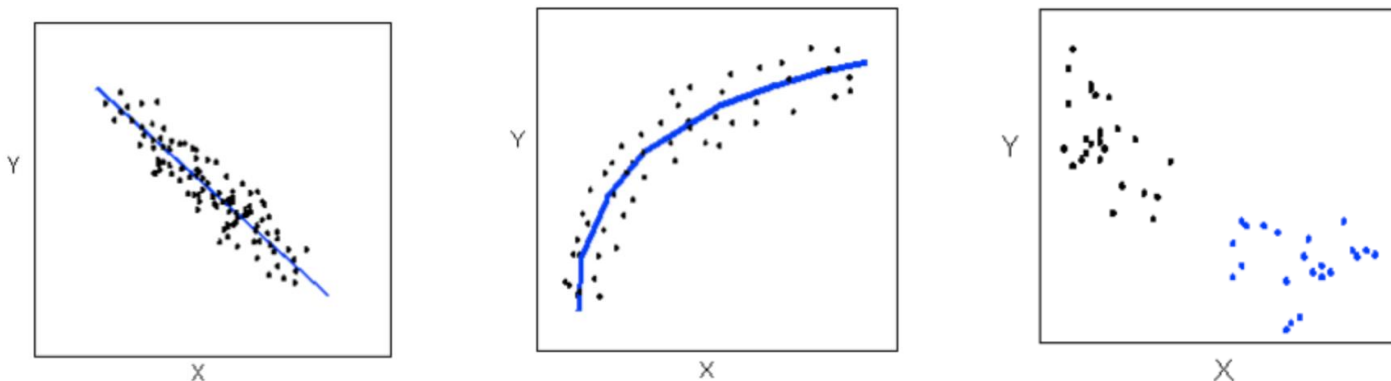
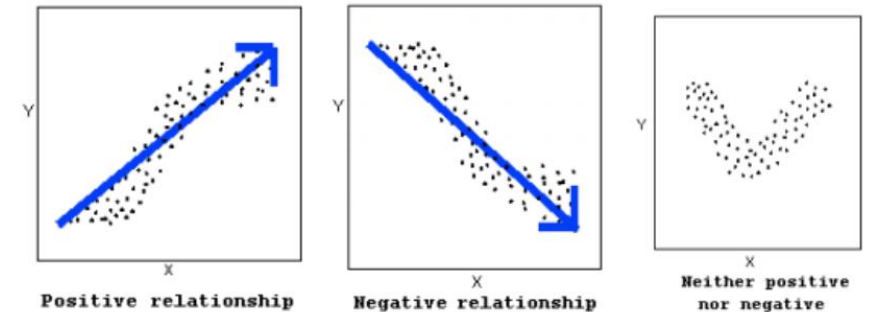
Part 1 - Correlation between quantitative variables (Cont.)

- Scatter plot of the data
- How should we interpret the scatter plot?
- When describing the overall pattern of the relationship we look at its direction, form and strength.



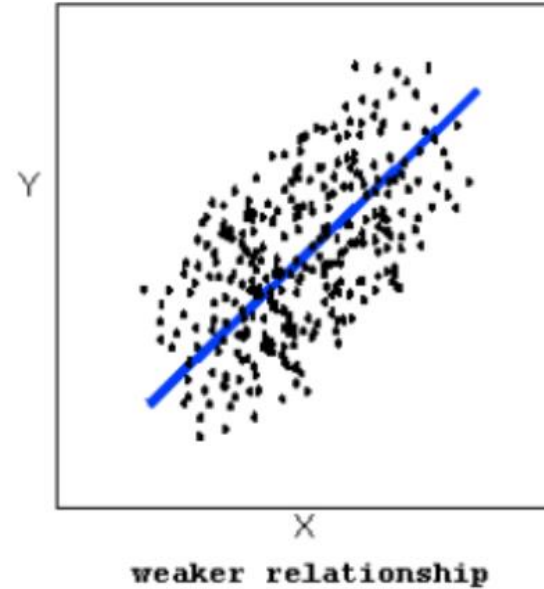
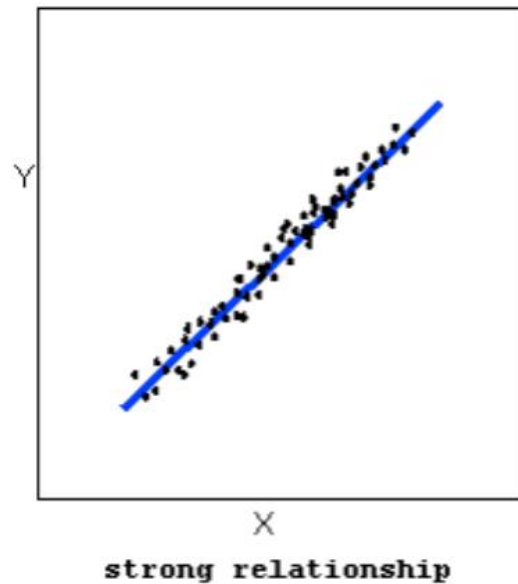
Part 1 - Correlation between quantitative variables (Cont.)

- The **direction** of the relationship can be positive, negative, or neither:
 - A *positive (or increasing) relationship* means that an increase in one of the variables is associated with an increase in the other.
 - A *negative (or decreasing) relationship* means that an increase in one of the variables is associated with a decrease in the other.
 - Not all relationships can be classified as either positive or negative.
- The **form** of the relationship is its general shape. When identifying the form, we try to find the simplest way to describe the shape of the scatterplot.



Part 1 - Correlation between quantitative variables (Cont.)

- The **strength** of the relationship is determined by how closely the data follow the form of the relationship. Let's look, for example, at the following two scatterplots displaying positive, linear relationships:



Part 1 - Correlation between quantitative variables (Cont.)

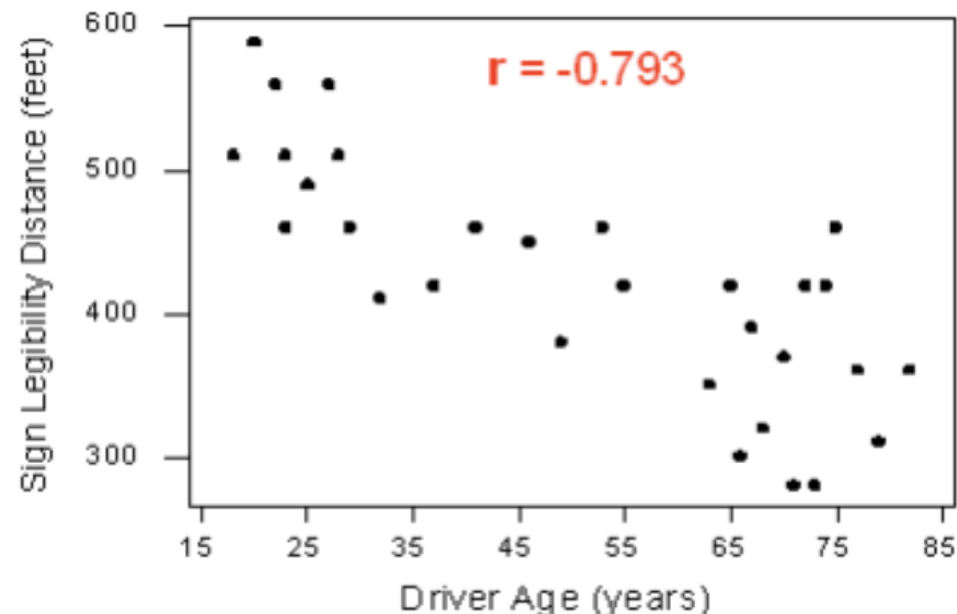
- **The Correlation Coefficient— r** (Pearson correlation)
 - The numerical measure that assesses the strength of a linear relationship is called the correlation coefficient, and is denoted by r .
 - *Calculation:* r is calculated using the following formula:
 - $$r = \frac{\text{Covariance}(X, Y)}{\sigma_x \sigma_y}$$
- **The covariance** of two variables x and y in a data set measures how the two are linearly related.
 - $$s_{xy} = \text{Covariance}(X, Y) = \frac{1}{n-1} \sum_1^n (x_i - \tilde{x}) * (y_i - \tilde{y})$$

Part 1 - Correlation between quantitative variables (Cont.)

- **The Correlation Coefficient— r** (Pearson correlation)
 - The numerical measure that assesses the strength of a linear relationship is called the correlation coefficient, and is denoted by r .
 - *Calculation:* r is calculated using the following formula:
 - $r = \text{covariance}(X, Y) / (\text{stdv}(X) * \text{stdv}(Y))$
 - *Interpretation and Properties:* Once we obtain the value of r , its interpretation with respect to the strength of linear relationships is quite simple, as this walkthrough will illustrate:
 - The value of r ranges from -1 to 1
 - Negative values of r indicate a negative direction for a linear relationship
 - Positive value of r indicate a positive direction for a linear relationship
 - Values of r close to zero indicate a weak linear relationship.
 - Values of r close to 1/-1 indicate a strong linear relationship

Part 1 - Correlation between quantitative variables (Cont.)

- Example : Earlier, we used the scatterplot below to find a negative linear relationship between the age of a driver and the maximum distance at which a highway sign was legible.
 - What about the strength of the relationship?
 - It turns out that the correlation between the two variables is $r = -0.793$.



Part 2 - Correlation between categorical variables

- We can observe the relationship between two categorical variables using graphs.
- We will perform statistical inference for two categorical variables, using the sample data to draw conclusions about whether or not we have evidence that the variables are related in the larger population from which the sample was drawn.
- In other words, we would like to assess whether the relationship between X and Y that we observed in the data is due to a real relationship between X and Y in the population, or if it is something that could have happened just by chance due to sampling variability.
 - We always assume that the data we have is a sample as we are unable to collect all instances / observations

Part 2 - Correlation between categorical variables

- The statistical test that will answer this question is called the **chi-square test of independence**.
- Chi is a Greek letter that looks like this: χ , so the test is sometimes referred to as: The χ^2 test of independence.

Part 2 - Correlation between categorical variables

- Example:** In the early 1970s, a young man challenged an Oklahoma state law that prohibited the sale of 3.2% beer to males under age 21 but allowed its sale to females in the same age group. The case (Craig v. Boren, 429 U.S. 190 [1976]) was ultimately heard by the U.S. Supreme Court. The main justification provided by Oklahoma for the law was traffic safety. One of the 3 main pieces of data presented to the Court was the result of a “random roadside survey” that recorded information on gender and whether or not the driver had been drinking alcohol in the previous two hours. There were a total of 619 drivers under 20 years of age included in the survey. Here is what the collected data looked like:

Driver	Gender	Drove Drunk?
Driver 1	M	N
Driver 2	M	N
Driver 3	M	Y
Driver 4	F	N
.	.	.
Driver 619	F	N

Drank Alcohol in Last 2 Hours?			
Gender ↓	Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Note that we are looking to see whether drunk driving is related to gender, so the explanatory variable (X) is gender, and the response variable (Y) is drunk driving.

Part 2 - Correlation between categorical variables

- Let us do basic EDA and look at percentages of drunk driving between genders:

Gender (X)	Drank Alcohol in Last 2 Hours (Y)?		Total
	Yes	No	
Male	77/481=16.0%	404/481=84.0%	100%
Female	16/138=11.6%	122/138=88.4%	100%

- For the 619 sampled drivers, a larger percentage of males were found to be drunk than females, hence, the data provides some evidence that drunk driving is related to gender.
- However, this in itself is not enough to conclude that such a relationship exists in the larger population of drivers under 20.
- We need to further investigate the data and decide between the following two points of view:
 - The evidence provided by the roadside survey (16% vs. 11.6%) is strong enough to conclude (beyond a reasonable doubt) that it must be due to a relationship between drunk driving and gender in the population of drivers under 20.
 - The evidence provided by the roadside survey (16% vs. 11.6%) is not strong enough to make that conclusion and could just have happened by chance due to sampling variability, and not necessarily because a relationship exists in the population.

Part 2 - Correlation between categorical variables

- **The Chi-Square Test of Independence**

- The chi-square test of independence examines our observed data and tells us whether we have enough evidence to conclude beyond a reasonable doubt that two categorical variables are related.
- *Step 1: Stating the Hypothesis*
 - H_0 : (They are independent.) There is no relationship between gender and drunk driving.
 - H_a : (They are not independent.) There is a relationship between gender and drunk driving.
- Step 2: Measure how far the data are from what is claimed in the null hypothesis.
 - For the null hypothesis, we need to calculate the counts that we would expect to see if drunk driving and gender were really independent (i.e., if H_0 were true).
 - For example, we actually observed 77 males who drove drunk; if drunk driving and gender were indeed independent (if H_0 were true), how many male drunk drivers would we expect to see instead of 77?
 - So, we need to calculate:
 - The observed counts (the data)
 - The expected counts (if H_0 were true)

Part 2 - Correlation between categorical variables

- **The Chi-Square Test of Independence (Example continued)**
- So, we need to estimate the size of the discrepancy between what we observed and what we would expect to observe if H_0 were true.
- $P(\text{Drunk}) = 93 / 619$ and
- $P(\text{Male}) = 481 / 619$, and so,
- $P(\text{Drunk and Male}) = P(\text{Drunk}) * P(\text{Male}) = (93 / 619) (481 / 619)$
- Therefore, since there are total of 619 drivers, if drunk driving and gender were independent, the count of drunk male drivers that I would expect to see is:

$$\text{Expected Count} = 619 * P(\text{Drunk and Male}) = 619 * (93/619) * (481/619) = (93 * 481) / 619$$

		Drank Alcohol in Last 2 Hours?		
Gender ↓	Yes	No	Total	
Male	77	404	481	
Female	16	122	138	
Total	93	526	619	

Observed counts

- Again, the expected count equals the product of the corresponding column and row totals, divided by the overall table total. This will always be the case and will help streamline our calculations:
- Expected Count = (Column Total x Row Total)/(Table Total)

Expected Counts

Gender	Drank Alcohol in Last 2 Hours?		Total
	Yes	No	
Male	$(93 \times 481) / 619 = 72.3$	$(526 \times 481) / 619 = 408.7$	481
Female	$(93 \times 138) / 619 = 20.7$	$(526 \times 138) / 619 = 117.3$	138
Total	93	526	619

Observed Counts

Gender	Drank Alcohol in Last 2 Hours?		Total
	Yes	No	
Male	77	404	481
Female	16	122	138
Total	93	526	619

Part 2 - Correlation between categorical variables

- Step 3: Finding the P-Value
 - The p-value for the chi-square test for independence is the probability of getting counts like those observed, assuming that the two variables are not related (which is claimed by the null hypothesis).
 - The smaller the p-value, the more surprising it would be to get counts like we did if the null hypothesis were true.
- Step 4: Stating the Conclusion in Context
 - A small p-value indicates that the evidence provided by the data is strong enough to reject H_0 and conclude (beyond a reasonable doubt) that the two variables are related.
 - In particular, if a significance level of .05 is used, we will reject H_0 if the p-value is less than .05