Math 200

OLI Mod 7 study guide

Variables and Classification

Variable classification:

- Categorical (descriptive – e.g. color, race, gender, phone number, . . .)
  - Binary and Non-Binary (categorical variables can be either binary (two options – e.g. yes/no; male/female) or non-binary (more than two – e.g. colors; races).
- Quantitative (numerical – e.g. measures like height and weight, gpa, age, time, . . .)

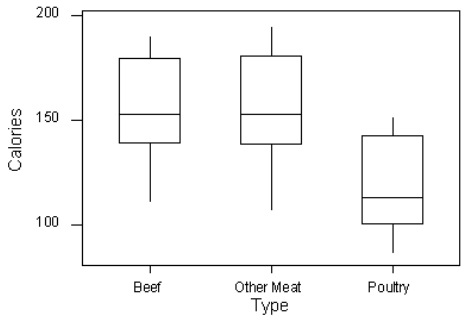
Cause and Effect:

- Explanatory (also called *independent* or *treatment* - the variable that claims to explain, predict, or affect the response).
- Response (also the response variable – is the outcome or result of a study).

C → Q

Compares two or more categorical explanatory elements through quantitative responses. In essence we compare the distributions of the quantitative response for each category of the explanatory variable using side-by-side boxplots supplemented by descriptive statistics.

1. Use medians to compare typical values. Use language like, "on average group A is <difference in medians> bigger than group B."
2. Use IQRs to compare spread of each. Use language that recognizes variability. "Group A has less variability than group B with the middle 50% of group A's attribute <like height or weight> spanning <attribute units such as inches or pounds>, while the middle 50% of group B's attribute ranges over <attribute units>."
3. If the two distributions are more different than they are alike, use this opportunity to emphasize the disparity by comparing the quartile references. e.g. 75% of group A is greater than <some reference value> while 75% of group B is less than <the same reference value>.

Calorie example from OLI: 1. While beef and "other meat" types of hotdogs have about the same average calories – 155; the typical calories for poultry is about 30 points lower. 2. All three categories exhibit about the same variability with the middle 50% of their data within about 40 calories of one another. 3. While beef and "other meat" have about the same caloric content overall, about 75% of beef and "other" have over 140 cal while about 75% of poultry dogs are below 140 cal.	

Other examples: See Rivers handout or LBD activity comparing school programs.

C → C

Comparisons of two categorical variables involves the use of 2-Way tables (contingency).

1. Identify the explanatory variable – label the corresponding table variable.
2. Calculate percentages in the direction of the explanatory variable (if the EV is a column variable, then calculate column percentages using column totals; if the EV is the row variable, then calculate row percentages using row totals).
3. Identify which of the response characteristics you are comparing and compare through division - either by dividing the larger by the smaller (this many times bigger) or by dividing the difference by the larger (relative difference).

Recall the seatbelt example where we compare the incidence of fatal accidents for those wearing seatbelts with those who did not.
Wearing the seatbelt (or not) is the explanatory variable while fatality (or not) is the response variable.

Seat Belt	Injury: Nonfatal Injury	Injury: Fatal Injury	Row Total
Seat Belt	412,368	510	412,878
No Seat Belt	162,527	1,601	164,128
Column Total	574, 895	2,111	577,006

We consider the probabilities P(fatal|seatbelt) and P(fatal|no seatbelt). Note the response variable (first slot – numerator) category remains the same (we're interested in fatality) while the explanatory variable (second slot - denominator) or what we are comparing – covers the two values we are comparing.

P(fatal|seatbelt) = 510/412878 ≈ 0.00124; P(fatal|no seatbelt) = 1601/164128 ≈ 0.00975

Comparing, we have 0.00975/0.00124 = 7.9 so people wearing seat belts are about 8 times more likely to survive an accident than those who do not wear a seat belt.

Alternatively, we can find the *relative* difference: $\frac{0.00975-0.00124}{0.00975} \approx 0.873$ meaning people who wear their seatbelts are 87.3% more likely to survive an accident than those who don't wear a seatbelt.

Examples: see coffee study and dementia (discussed in class) or smallpox quiz.

Q → Q

Scatter plots and Linear Regression. Both variables (in particular the explanatory variable) are quantitative and paired.

1. Identify the explanatory variable and use this as the horizontal axis (*x*-axis)
2. Graph the data as a scatterplot and identify:
    a. Direction (+/–)
    b. Form (Linear/Curvilinear)
    c. Strength (strong/moderate/weak)
    d. Outliers
3. For graphs with linear form find regression formula
    a. Identify and interpret growth (or decay) rate (slope) and initial value (response intercept).
    b. Recognize *Correlation Coefficient* (*r*) as a measure of how well the line approximates the data.
    c. Recognize *Coefficient of Determination* (r^2) as a measure of the percentage of variability in the response variable that can be explained by a linear relationship with the explanatory variable.
    d. Use the linear model as a tool for predicting data values within the bounds of the data (interpolation) and beware using the model as a predictive tool beyond the bounds of the data (extrapolation). In particular, be wary of using the response intercept as a prediction as it is often outside the data set.

Note:

Has r ~ 0 but has a strong association.

Examples include posted examples and the car project.

Causation and Lurking Variables

When we explore the relationship between two variables, there is often a temptation to conclude from the observed relationship that changes in the explanatory variable *cause* changes in the response variable.
Association *does not* imply causation!

Confounding and *lurking* variables both affect the response variable in such a way as to suggest a causative relationship with the explanatory variable. Examples like the firefighters and fire damage or fat consumption and lifespan (from class) are good examples. Typically things like wealth, age, or degree of extremity act as confounding variables.

Simpson's Paradox

Associations between variables that reverse their implication depending on whether data are aggregated or treated separately fall into the category of Simpson's Paradox.
Recall the example from class involving UC Berkeley graduate schools and the disparate admissions rates for men and women.

Examples: See Causation LBD (both) on OLI and class examples.