GER1000 (Analysis of Statistics)

Jia Cheng

January 2021

1 Definitions

QR Framework

1. Frame
2. Specify
3. Collect
4. Analyze
5. Communicate

2 Misc

TP-SR5 Seat number 12

3 Associations

3.1 Rates

Essentially conditional probability. From here on, define $rate := P$ (probability function)

Definitions and Propositions

- Positive Association of A with B: If $P(A|B) > P(A|B^c) \vee P(B|A) > P(B|A^c)$
- $P(A|B) = P(A|B^c) \implies P(B|A) = P(B) = P(B|A^c)$

3.2 Groups

- Observational group (non-assigned)
- Control Group
- Treatment Group

3.3 Confounders

Confounding variables are associated with both independent and dependent variables.

To reduce confounding, use slicing, so as to compare smaller groups with are relatively homogeneous w.r.t. the factors.

3.4 Simpson's Paradox

See Brilliant.com

3.5 Tutorial 1

Imagine that you are an intern at a large tuition centre catered to students of age 11 and 12 years. Your employer wants to know if it is worthwhile to invest in iPads to improve students' proficiency in English. He gives you authority and resources, and asks you to design an experiment on the thousands of customers.

(a) How would you enrol subjects and assign them into two groups?

Ask for parental consent, then randomly assign consenting students to control and "treatment" groups.

(b) How feasible is it to use a placebo, or to implement double-blinding?

Placebo is not possible, since you can't just fake an ipad.

Blinding students is difficult, since knowledge of having/not having ipad in lesson is easily known by students and by parents.
Single blinding is possible however. For e.g. when testing the students at the end of the trial, do not inform the accessors about which group they belong to.

4 Association

QR Framework: **Analyze, Communicate**

Def: Relationship between 2 numerical variables
Can be deterministic (described by mathematical equation/equality) or statistical (described by trends/patterns)

4.1 Linear Regression

Regression Line

- Also called line of least squares (minimises squared errors of data points in y-direction)
- Predicts average/approximate values
- Cannot (should not) be extrapolated to predict values of dependent variable outside the line

Coefficient of correlation

$$\frac{1}{n}\sum_i \frac{X_i - \frac{\sum_j X_j}{n}}{SD_X} \cdot \frac{Y_i - \frac{\sum_j Y_j}{n}}{SD_Y}$$

- Measures linear association between variables
- Unitless
- No linear association only when $r = 0$
- Descriptors:
  - Strong: $|r| \in [0.7, 1]$

  - Moderate: $|r| \in [0.3, 0.7]$
  - Weak: $|r| \in [0, 0.3]$
- Descriptors:
  - Positive/Negative
- Example: There is a (1)moderate (2)positive linear association between X and Y

Outlier

- Should not always be removed
- When removed, has serious consequences on r-value

4.2 Ecological Correlation

Def: Correlation computed based on aggregate data
Used when individual data is more difficult to obtain

When association for both individuals and aggregates are in the same direction, ecological correlation based on aggregates will typically overstate strength of association in individuals.

Fallacies

- Ecological Fallacy: When we deduce inferences on individual-level correlation based on aggregate data
- Atomistic Fallacy: When we generalize correlation based on individuals toward aggregate-level correlation

4.3 Attenuation Effect/Data Removal

Tends to understate strength of correlation

4.4 Chocolate Consumption

Confounding factors; Said factors are positively associated with both the dependent and independent variables.

He brushed them under the carpet lol. If I were to collect aggregate data, I would get those from similar socioeconomic backgrounds and make use of table-splitting.

The present data are based on country averages, and the specific chocolate intake of individual Nobel laureates of the past and present remains unknown. The cumulative dose of chocolate that is needed to sufficiently increase the odds of being asked to travel to Stockholm is uncertain. This research is evolving, since both the number of Nobel laureates and chocolate consumption are time-dependent variables and change from year to year.

Instead of collecting aggregate data (of which Nobel laureates only make up a exceedingly tiny proportion of), I would collect data directly from the Nobel Laureates. Ecological fallacy. Population data may not apply to individuals. Even if it does apply to a certain proportion of individuals, Nobel laureates are so few and far between. Population may like chocolate. Laureates themselves may not even consume chocolate.

Change in population units with respect to time; Tastes and preferences, economic situation would all affect annual chocolate consumption. Considering how few Nobel Laureates are produced per year, even getting one more Laureate per country would lead to large changes in the population unit.

5 Odds, Risk ratios

Given 2 groups, A and B, we will abuse notation and let A, B refer to the events that a person is in A and B respectively.

$$\text{Odds, } O(A) = \frac{P(D|A)}{P(D^c|A)}$$
$$\text{Risk, } R(A) = P(D|A)$$
$$\text{Odds ratio, } OR(A/B) = \frac{O(A)}{O(B)} = \frac{\frac{P(D|A)}{P(D^c|A)}}{\frac{P(D|B)}{P(D^c|B)}} = \frac{P(D|A)P(D^c|B)}{P(D|B)P(D^c|A)}$$
$$\text{Risk ratio, } RR(A/B) = \frac{R(A)}{R(D)} = \frac{P(D|A)}{P(D|B)}$$

$RR(A/B) > 1 \iff P(D|A) > P(D|B) \iff P(D^c|A) < P(D^c|B) \iff \frac{P(D^c|A)}{P(D^c|B)} < 1 \iff \frac{P(D^c|B)}{P(D^c|A)} > 1$
Hence, $RR(A/B) > 1 \implies OR(A/B) > 1$
Since $P(D|A) > P(D|A) \iff P(D^c|A) < P(D^c|B)$, the converse is clearly true as well.
An analogous derivation shows that $RR(A/B) < 1 \iff OR(A/B) < 1$.
Finally, $RR(A/B) = 1 \iff OR(A/B) = 1$

6 Testing

6.1 Null hypothesis, alternative hypothesis, p-value

p-value is calculated with the assumption that the null hypothesis is true.

6.2 Sensitivity and specificity

Let A be the event that a person is tested positive for disease. Let D be the event that the person has the disease.

$$\text{Base rate} = P(D)$$
$$\text{Sensitivity} = P(A|D)$$
$$\text{Specificity} = P(A^c|D^c)$$