GER1000 (Analysis of Statistics)

Jia Cheng

January 2021

1 Definitions

QR Framework

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

2 Associations

2.1 Rates

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

Definitions and Propositions

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

2.2 Groups

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

2.3 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.

2.4 Simpson's Paradox

See Brilliant.com

3 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)

3.1 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

3.2 Ecological Correlation

Definition 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

3.3 Attenuation Effect/Data Removal

Data removal can cause the attenuation effect. Removal of data causes attenuation effect when it result in **under**stating strength of correlation.

4 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.

$$\begin{aligned}
\text{Risk, } R(D|A) &= P(D|A) \\
\text{Odds, } O(D|A) &= \frac{P(D|A)}{P(D^c|A)} = \frac{P(D|A)}{1-P(D|A)} \\
\text{Risk ratio, } RR &= \frac{R(D|A)}{R(D|B)} = \frac{P(D|A)}{P(D|B)} \\
\text{Odds ratio, } OR &= \frac{O(D|A)}{O(D|B)} = \frac{\frac{P(D|A)}{P(D^c|A)}}{\frac{P(D|B)}{P(D^c|B)}} = RR \cdot \frac{1-P(D|B)}{1-P(D|A)}
\end{aligned}$$

Note that the above assumes that B is used as a baseline, such that $P(D|B)$ is the denominator in RR and OR.

$$\begin{aligned}
RR > 1 &\iff OR > 1 \\
RR < 1 &\iff OR < 1 \\
RR = 1 &\iff OR = 1
\end{aligned}$$

This is a standard contingency table

	D	D^c
A		
B		

where A, B are exposure events (e.g. A represents exposed, B represents not exposed), and D is disease event (and D^C is the event representing no disease)

Cohort study

	D	D^c	Sample size
A			100
B			100

Proposition Sample RR can be used to estimate population RR.

Intuition Denote P_s as the sample probability/rate function, P_p as the population probability/rate function. Denote n_s as the cardinality of an event with regards to the sample. Denote n_p as the

cardinality of an event with regards to the population.
Then,

$$\begin{aligned}
\text{Sample RR} &= \frac{P_s(D|A)}{P_s(D|B)} \\
&= \frac{\frac{P_s(DA)}{P_s(A)}}{\frac{P_s(DB)}{P_s(B)}} \\
&= \frac{\frac{n_s(DA)}{n_s(A)}}{\frac{n_s(DB)}{n_s(B)}} \\
&\approx \frac{\frac{n_p(DA)}{n_p(A)}}{\frac{n_p(DB)}{n_p(B)}} \\
&= \frac{\frac{P_p(DA)}{P_p(A)}}{\frac{P_p(DB)}{P_p(B)}} \\
&= \frac{P_p(D|A)}{P_p(D|B)} = \text{Population RR}
\end{aligned}$$

The above derivation relies on the idea that

$$\frac{n_s(DA)}{n_s(A)} \approx \frac{n_p(DA)}{n_p(A)} \text{ and}$$
$$\frac{n_s(DB)}{n_s(B)} \approx \frac{n_p(DB)}{n_p(B)}$$

Intuitively, the first $\approx$ is true since our first (exposure) sample of 100 was drawn directly from the exposure group A, hence it is natural to believe that the proportion of $n_p(DA)$ as a fraction of $n_p(A)$ would be approximately equal in our sample.
The same argument for the non-exposure group B gives the second $\approx$.

Proposition Sample OR can be used to estimate population OR.

Case control study

	D	D^c
A		
B		
Sample size	100	100

Sample RR **cannot** usually be used to estimate population RR. This is especially true if the disease D is rare.

Intuition To make things easier, let's first assume that the disease is extremely rare, such that we expect $P_p(D|A), P_p(D|B)$ to be very close to 0, where P_p is defined in the previous section on cohort study.

Our samples are: 100 ppl from those having disease D, 100 from those without disease D.
Note that in the derivation for cohort study, we assume that $\frac{n_s(DA)}{n_s(A)} \approx \frac{n_p(DA)}{n_p(A)}$ (likewise for B). In other words, we assumed that $P_s(D|A) \approx P_p(D|A)$.
But this is clearly untrue in the case of a case control study of a very rare disease. Half of our total

sample comes from those with the disease, so either of $P_s(D|A), P_s(D|B)$ must be quite close to 0.5, much larger than $P_p(D|A), P_p(D|B)$!

Proposition Like in cohort study, sample OR can be used to estimate population OR.

5 Testing

5.1 Null hypothesis, alternative hypothesis, p-value

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

5.2 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.

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

In other words,

$$\begin{aligned}
\text{P(true positive)} &= P(AD) = P(A|D)P(D) = \text{Sensitivity} \cdot P(D) \\
\text{P(true negative)} &= P(A^cD^c) = P(A^c|D^c)P(D^c) = \text{Specificity} \cdot P(D) \\
\text{P(false positive)} &= P(AD^c) = (1 - \text{Specificity}) \cdot P(D^c) \\
\text{P(false negative)} &= P(A^cD) = (1 - \text{Sensitivity}) \cdot P(D)
\end{aligned}$$