GER1000 Wk1-2

- QR Framework (Quantitative Reasoning)
  - offers precision

1) Frame
2) Specify
3) Collect
4) Analyse
5) communicate

- Causal Relationships

(Chpt 1) controlled experiments
(Chpt 2) Observational studies

assign subj into grps

- compare rates
  (eg.) Treatment vs Control
- [groups possibly having different risks]
  - → social economic status
  - → severity of disease

Q: Are groups different aside from treatment

- Randomised assignment from group w parental control ⟹ Randomised Controlled Double Blind Experiment
  - Placebo
  - Binding Subj
  - Binding Doctors

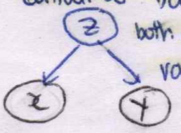
precautions agst bias! *
(confounding bias)

⟹ can reduce confounders (using randomized)
∴ can suggest causation

Diff from experimental study in that groups have been preassigned subjects
(eg. can't ask someone smoke 30 yrs)

⟹ ∴ cannot reduce confounders by randomization (but can slice)
∴ cannot suggest/prove causation
(only correlation)

Confounder: var that influences both depend & indep variable



- A, B popn characteristics w $0 < r(A), r(B) < 1$ "rate"

	A and B	Condition		
eg. Smoking & Lung prob	positively associated	r(A|B) > r(A | not B)	or	r(B| A) > r(B| not A)
eg. Poverty & High IQ	negatively associated	r(A|B) < r(A | not B)	or	r(B|A) < r(B| not A)
	not associated	r(A|B) = r(A | not B)	or	r(B|A) = r(B| not A)

B is more common among people w A than among people without A

*always compare R(A|B) & R(A|~B)
NOT R(A|B) & R(~A|B)

~~Confounding~~

Confounder

- Smoking & Heart Disease → Age & Sex confounder (exaggerates effect of smoking)
  2 common confounders
- confounder (3rd var) associated w both exposure & disease

* impt to control for confounder in Observational studies
  - ↳ "slicing" → splitting grp into heart disease amongst males & h.d. amst females
  [Separating Data sets]

Techniques for controlling confounder
- Slicing (basic but cumbersome)
- Statistical Technique (eg. regression)

thinking of confounders.
Are __ & __ diff in some ways other than __

SUMMARY

- Association is not causation
- Observational Studies prone to confounding
  → knowledge & thinking useful for spotting potential confounders.
- Slicing is effective for controlling confounders.

Quiz learning

- Yule-Simpson Paradox: when there is a disparity btw
  - direction of association in most subgroups
  - & ~~dir of asso~~ overall association (when subgroups are combined)

(Chpt 1 slide 83)

→ R/nships in subgroups are reversed when rates are combined
⟹ Due to confounding (eg. size of kidney stones)
∴ treat paradox by slicing into grps / confounders

GER wk 3 & 4

- **Association → Statistical Relationship**
  1. Bivariate Data & Scatter Diagram → closeness to meanline
  2. Explore Relationship
  3. correlation coefficient
  4. some limitations
  5. Ecological correlation (aggregate measures)
  6. Cautionary Notes
  7. Simple Linear Regression

- correlation relation r
  (direction & strength) of linear relation

-1 —— 0 —— 1

∴ if var indep → r = 0
r = 0 ↛ var indep (r only measure linear dependencies)

- r close to ±1 → strong association (above 0.7)
- " ±0.5 → moderate " (0.3–0.7)
- " 0 → weak " (0–0.3)

} positive/ Negative

- computing correlation coefficient (r)
  1. Convert each var to Standard Unit (SU) $SU = \frac{x_i - \bar{x}}{sd_x}$
  2. Take pdt of SU for father-son pair
  3. r ⇒ Avg of all products

$$\boxed{r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{sd_x}\right)\left(\frac{y_i - \bar{Y}}{sd_y}\right)}$$

⇒ =CORREL(rangeX, rangeY)

- no units
- not affected by interchange
- not affected by change of scale (addition/ multiplication) or units of measurement

Limitations w correlation Causation

1. Causality
2. Outliers in data set → very sensitive, may deflate/inflate correlation
3. non-linear association

Ecological correlation

- Data of Aggregate data (not individual data) like groups

→ generally if measured btw 2 variables, strength would be overstated (cos same line)
*(similar if var are in same direction)

*Ecological fallacy: deduce inferences on corr. abt individuals based on aggregate data

* Atomistic Fallacy: Generalize corr. based on individuals toward aggregate

} Only causal* correlated relations

∴ proceed carefully w scatter diagram

Cautionary Notes on Correlation

Attenuation $\xrightarrow{\text{means}}$ Reduction in Value

- Attenuation Effect: (Phenomenon due to Range Restriction in 1 var)
  ↳ ~~restriction of range~~ Range Restriction
  - bivariate data set formed from restriction of one var
  - data for other variable available for a limited range

  *tend to have diminishing effect of correlation (understate)

- Removal of Data (eg.)
  (eg. space shuttle removal of non-damaged data obj)

Linear Regression

- using regression line (Y = a + bX)
- specify independent (X) and dependent (Y) var.

a → y-intercept
b → gradient

| excel name ~~excel name~~ |
| " intercept " |
| " slope " |

} - f(x) get
- "statistical"
...

✓ Determined by Least Square Method

GER wk5

Sampling

- How Data is generated
- Generate own Data set / use public ones?
- Terminology : "unit", "population", "sample"

census — measurement from evy unit in pop^n

sample — measurement from some selected unit in pop^n

(eg. Drug Testing (Disease)

unit — individual w disease

pop^n — collection of all individuals defined above

sample — collection of individuals placed under experiment

Adv of sampling over taking census : when census not possible (eg. Disease / Blood sample) ; Speed, cost, accuracy (∵ btr task force to take measurement)

Sampling frame → list of sampling units used to identify all units in pop^n (eg house address → people in HDB houses)

- Good sample is extendable
- Evy unit in pop^n has possibility to be sampled
- No selection bias
- simpleste sampling frame : list of units in pop^n

Characteristics of good frame

- Good covrage
- Up-to-date & complete

→ inclusion of undesired units will not increase cost

→ exclusion of desired units will not have major impacts on outcome of study

Probability Sampling Plan

(eg.) Simple Random Sampling (evy unit same prob of selection)

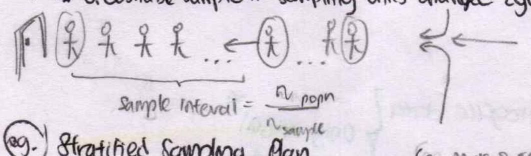
(eg.) Systematic Sampling (application of selection interval K)

- can apply to situation when exact size of popn not known (rough estimate only)

→ treated like Simple Random Sampling IF sampling units arranged randomly

(eg. 110 units in popn,
Sampling unit desired 10.
110/10 = 11 . rand(1,11) = 6
Sampling unit = [6, 17, 28 ...])

× ~~~~ Undesirable sample if sampling units arranged cyclically (147147...)



sample interval = $\frac{n_{popn}}{n_{sample}}$

(eg.) Stratified Sampling Plan

→ divide popn into grps ("strata")

→ take prob sample from each group

— can include setting quotas

(eg. Male & Female)

(eg. Country as strata so local organisations help take results → find effectiveness of drug in each country (more acurate))

(eg) Multistage Sampling Plan

- In many studies, several stages of sampling req. for ~~~~ reaching popn units.

(eg.) list of houses [Simple Random] →

→ everyone interviewed [Cluster Sampling]

→ one member randomly interviewed [2-stage Sampling]

(eg) Drug testing of particular disease

- Sampling frame = list of physicians
- From each physician, only some of patients included in study.

Difficulties in Sampling

- Using imperfect sampling frame
  - exclude desired & include unwanted
  - ① Redefine target pop^n ↳ increase cost of study
  - ② Assess impact of excluding these units in our study
- Non-Response
  - Not anyone willing take ~~~~ part in study
  - ↳ incentive increase response rate (eg. Medicine Study < payment / free checkup)
- Getting Volunteer / Self-Selected Sampling [Aka non-prob sampling]
  - (eg.) News / Media conduct polls on website
  - ↳ bias as only people w strong views willing to give answers
- Using a Convenience or Haphazard Sample
- Taking a Judgement Sample : interviewers choose units by discretion
- Selecting a Quota Sample (typical in Market Research) (distinguish from Stratified) — ~~~~ convenience sample to hit Quota
  - ↳ compare prop of quta sample to category units / census units

Statistician John Tukey "I would trade all your 18,000 case histories for 400 in a prob sample"

GER1000 WK6

Ch3 Estimating Parameter

- Parameter : numerical fact bout popⁿ (usually unknown)
  → ~~[illegible]~~ estimated from sample

{ Estimate = Param + Random Error + Bias }

Random Error ↓ Easy to Quantify

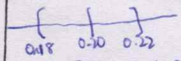
Bias ↓ Hard to Quantify

→ Random Error

Larger sample size ⟶ likely smaller random error

Confidence Interval ⟵ ~~Bate~~ confidence level, range

- range of valus we are reasonably certain unknown param lies in



95% CI : 0.20 ± 0.02

Confidence level x% → x% ~~chance~~ likelihood that ~~[illegible]~~ actual param value in range

(aka x% of researches will have intervals containing popn param) → resp & unique

↳ Larger sample size $\xrightarrow{\text{likely}}$ smaller range of CI (width of CI) also affected by variation in popn

- We are 95% confident that [0.18, 0.22] contains the popc param
  not 'chance'

→ Types of Bias
1) Selection Bias — Systematic tendency ~~[illegible]~~ on the part of sampling procedure to exclude one kind of person from sample
   Caused By : → imperfect sampling frame
   → Non probability sampling methods
2) Non-response Bias
   … other types

→ Systematic tendency from subjects who do not respond to survey / questionaire

Caused By : ⟶ Diff btw non-respondent & respondents

** non-response rate ↑ ~~[illegible]~~ $\xrightarrow{\text{likely}}$ Non-response bias significance ↑

other types of Bias

- Phrasing of qn / tone / attitude of interviewer (including order)
- When subj have tendency to understate undesirable social habits (like smoking)

⇒ Conclusion (to minimize bias in survey)

1. Include every popⁿ unit in frame
2. Use prob sampling frame
3. Get 100% response rate

~~Chapter 7 Uncertainty~~

~~Unit : measuring uncertainty~~

~~Probability — the measure of likelihood~~

~~- Interpretation (Relative Freq vs Personal Prob)~~

~~Relative Freq~~	~~Personal Prob~~
~~can be quantified~~	~~can't be exactly quantified~~
~~Based on repeated observation of outcomes~~	~~based on own personal belief~~

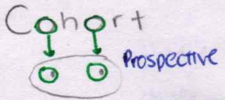
~~→ Proportion of times over the long run~~

Geri000 wk7

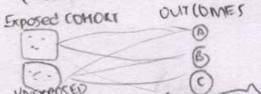
(Chpt4) more on observational studies

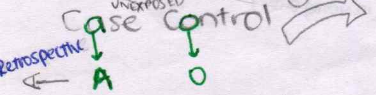
PARTE 1: Risks (Rate in popⁿ)

Two kinds of observational studies
- > Risks : rate in population → Risk Ratio (RR) (aka Relative Risk)
- > odds

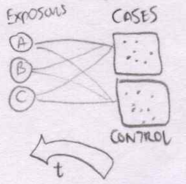


"Relatives are the same"
(Relative Risk)





"at Odds w one another"
(odd ratio)



(eg.) risk(Diabetes | Female) = $\frac{D \cap F}{F} = \frac{72,000}{216,000} \approx 0.33$

$risk(D|M) = 0.25$

$\text{Risk Ratio} = \frac{risk(D|F)}{risk(D|M)} \approx \frac{0.33}{0.25} \approx 1.33$

	Diabetic	Healthy	Row total
Females	72k	144k	216k
Males	52k	156k	208k
Col total	124 k	300k	424k

- Use Probability Samples to estimate
  2 ~~samples~~ strategies : Simple Random Sample (SRS)
  from
  - each exposure group (female & males)
  - each disease group (diabetic & healthy)

Cohort & Case Control studies

Study	Samples from	Advantage
Cohort	Exposure grp	Risks & RR can be estimated from sample table
Case-Control	Disease grp	Good for Rare Diseases

↳ estimation NOT POSSIBLE
~~(risk estimates too high)~~
(depends on sampling % from each group)

Remarks on Sampling
- > Randomised Exp : subj need not resemble popⁿ
  Extrapolation to popⁿ is an issue
- > Observational studies : Both extrapolation & confounders are impt issues
  - cohort studies rely on random samples for accurate estimation of risks

Summary (Risks)
- Risks like rates are affected by confounding
- Risk Ratio / Relative Risk measures association
- Cohort Study has accurate estimation of popⁿ risks & RR w random samples
- Case-Control Study does not.

- Exposure → Disease
- R(Disease | Exposure) = P(getting Disease if you are exposure grp)

PARTE 2: Odds

Risk & Odds

$$\boxed{odds = \frac{risk}{1-risk}}$$

- odds > risk
- if risk is small, odds is small

~ODDS RATIO~

Odds (Diabetes) among females = $\frac{72k}{144k} \approx 0.50 \longleftrightarrow \frac{0.33}{1-0.33}$

Odds (Diabetes) among males = $\frac{52k}{156k} \approx 0.33 \longleftarrow \frac{0.25}{1-0.25}$

OR(Diabetes) btw Female & Male (baseline)

$\approx \frac{0.50}{0.33} \approx 1.50$

*
- OR = 1 → No diff. in disease risk btw 2 groups: RR = 1 (only situation where value can be deduced from OR)
- OR > 1 → Higher Risk in first group : RR > 1
- OR < 1 → Lower Risk in first group : RR < 1

CROSS-PRODUCT RATIO to calculate OR

	Diabetic	Healthy
Female	378	702
Male	53	155

$\frac{378 \times 155}{702 \times 53} \approx 1.57$

	Diabetic	Healthy
Female	364	142
Male	256	158

$\frac{364 \times 158}{142 \times 256} \approx 1.58$

Table must be set proper
- Event of intest (1st Col)
- 1st Group (1st Row)

✓ Risk Ratio
→ Only Cohort Studies

✓ Odds Ratio
→ Both Cohort & Case-Control studies

✓ Cohort Study
① All subjects Disease-Free @ beginning
② Sometimes, OR is used by researches
→ Compare odds of developing disease between 2 exposure groups

Multi-Level contigency Table
→ Form 2x2 table (choose interest & baselines)

* error for observational studies, correlation ↛ causation

Chapter: Uncertainty (measuring it)

Probability - measure of likelihood

Interpretations	Relative freq.	Subjective Prob Personal Prob
	can be quantified exactly	can't be quantifies exactly
	Based on repeated observation of outcomes	Based on own personal beliefs

Relative Freq vs Prob

Tossing coin	Weather forecast
Precise Prob	imprecise Prob
Assumption abt physical reality	circumstances repeat & outcome observed
Prob used to predict relative frequency	Relative Freq used to predict Prob

(Diabetes eg.)

Risk_female = 0.33 = $\frac{72k}{216k}$ = ~~Prob~~ (female has Diabetes)

Risk-male = 0.25 = P(Male has Diabetes)

Simple Probability Rules
- Mutually Exclusive A, B → $P(A \cup B) = P(A) + P(B)$
- Complement Rule

odds revisited

$odds = \frac{\#\ of\ event}{\#\ of\ non\text{-}event} = \frac{P(event)}{P(event\ not\ occuring)}$

- Indepent Event A, B → $P(A \cap B) = P(A)P(B)$
  (ocurence of one does not affect ocurence of the other)

Ger1000 wk8

Unit 3) Expected value

Unit 4) Uncetainty p-values

Overview:

p-values

Null Hypothesis — Hypothesis test — Statistical Significance

eg. cointossing, is coin fair?

Assume Null Hypothesis: the coin is fair → assumes that observation is due to chance

HHHTH — outcome
THHHH
HTHHH
HHTHH
HHHHT — equivalent to outcome
HHHHH — more extreme than outcome

Evidence that coin is biased in favor of head

$P\text{-value} = P(THHHH) + P(HTHHH) + \ldots = 0.18$

$P\text{-value} = 0.18 > 0.05$

→ Do not reject null hypothesis at the 5% significance level

→ cannot conclude that the coin is not fair

p-value :

- The probability of obtaining an outcome equivalent to or more extreme than the observed.

0.05%

small ← → Large

- unlikely for the obsered to occur by chance
- unlikely that null hyp is true

- More likely that null hyp is true
- more likely that obsered occured by chance

conclusion: case 1) reject H_0 (accept [redacted] H_1 at x% level of significance)
case 2) fail to reject H_0 (≠ accept H_0) cannot

* H_0 is never accepted

Null hypothesis rejected at 5% level of statistical significance

Hypothesis Testing

1. Identify the question (Frame)
2. State the null hypothesis (Specify)
3. Conduct the experiment (Collect)
4. Compute P-value (Analyse)
5. Make conclusion abt the null hypothesis (Communicate)

(summary)

small p-value
- prob too small for obsered data to be purely by chance
- Statistically significant to reject null hypothesis
- Obserd effect in sample is likely to reflect effect in population

Large p-val
- obsered data may be due to chance
- Not statistically significant to rej H_0
- obsered effect may or may not reflect effect in popn

Drug Testing eg.

1) Is drug effective in some popn ?
2) ~~H~~ H_0 : drug has no effect in popn
3) Give drug to three patients and obseve the number of patients who survived
4) Compute P-value
5) Conclude (rej H_0?)

Disease D, 40% fatal

- 40% fatal → 60% survive → P(survive) = 0.6
- P(all 3 patients survive) = $0.6^3 = 0.216$
- $P\text{-val} = 0.216 > 0.05$ ⟹ Do not reject null hypothesis at 5% significance level

If ~~4~~ 4 out of 6 survived

$$P\text{-val} = P(4\text{ survive}) + P(5\text{ survive}) + P(6\text{ survive})$$
$$= \binom{6}{4}(0.6)^4(0.4)^2 + \binom{6}{5}(0.6)^5(0.4)^2 + \binom{6}{6}(0.6)^6$$

Unit 5) Conditional Prob

- independent A, B: $P(A|B) = P(A)$ / $P(A\cap B) = P(A)P(B)$
- $P(A\cap B) = P(A|B)P(B)$

interesting eg.

D: double cot death

E: Sally is innocent

$P(D\|E)$	vs $P(E\|D)$
○ Prob quoted by expet	○ Prob required by prosecutor

Unit 6) Uncertainty Testing Rare Events

eg. Framing qn : [tested positive ⟹ Treatment?] P(disease | positive)

Specifying what to measure :

Base Rate	P(disease)	= A/N
Sensitivity of Test	P(positive | disease)	= C/A
Specificity of Test	P(negative | no disease)	= D/B

Collecting Data:

1. Carry out study on random sample (N) from popn
2. Record no of pep w disease (A) & wo disease (B)
3. Among those w disease, record no. of pep tested positive (C)
4. Among those wo disease, record no of people tested negative (D).

Analyze ~~Tables~~ Data:

Base rate : 0.001
Sensitivity : 0.95
Specificity : 0.9

fill in table

Contigency Table: assume popn of 100,000

	Test positive	Test negative	Row sum
Have Disease	95	5	100
Do not have Disease	9990	89910	99900
Col sum	10085	89915	100 000

True positive

False positive

False negative

True negative

$P(\text{disease} | \text{positive}) = \dfrac{\text{No. of true positive}}{\text{No. of positives}} = 0.0094$ & $P(\text{no disease} | \text{positive}) = \dfrac{\text{No. of false positive}}{\text{No. of pep tested pos}} = 0.9906$

Communicate the Findings

- Test has high sensitivity & specificity: P(~~dis | pos~~ pos | dis) = 0.95, P(neg | no dis) = 0.9
- Less than 1% tested positive have the disease : P(disease | positive) = 0.0094
- More than 99% tested positive have no disease : P(no disease | positive) = 0.9906

→ if tested ~~confident~~ positive, not confident that test is correct

→ ALWAYS happens, when disease is rare.

SUMMARY

- P (event happening | event suspected)
- Base rate, sensitivity, specificity REQUIRED!
- Contigency Table set up
- True positive & False positive