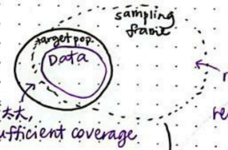
Chapter 1

- population = grp. of individual /subject we want know smtg abt
  - ↓ population parameter 想研究的指标
  - ↓ sample ⟷ census
    - sample: pop. 里 proportion 参加 study, sample size ↑, 准确, random error ↓, ② 用大 sample size
    - census: pop. 里全部人都参与调查; ① cost ↑ ② time ↑ ③ speed ↓; 有人 refuse 也算 (non-response bias)
  - ↓ estimate



- 什太大, insufficient coverage
- redundant ← remove from analysis

→ sampling frame ⊇ target pop. 才能 generalise
- ↳ include all sampling units?
- ↳ 有 irrelevant unit from other pop. 吗?
- ↳ 有 duplicated sampling units 吗?
- ↳ 有 sampling in clusters 吗?

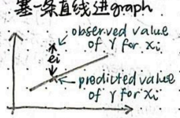
在做 sampling frame 时, 可能有 bias
- non-response bias (non-disclosure of info, will happen naturally 无论 method) ④ minimise non-response; * non-truthful ≠ random error
- selection bias — sampling 注意 ① 选好 sampling f., imperfect selection frame (+p. / insuff! / frame)
- non-probability sampling
  - convenience s. → proximity & availability → selection & non-response bias
  - volunteer s. → researchers seek volunteers 但有强烈意见的人更可能参加
- ② 选 probability sampling: probability sampling — know & randomised selection (可计算出被选中机会)
  - simple random s. → all units equal chance → select randomly w/o replacement
  - stratified random s. → 将 similar characteristic 分组, 再在每个组 simple random (好比更决定每个组选多少人) — strata 之间的 characteristic 不同
  - cluster sampling → 把 unit 分成 similar cluster, 再 P随机 抽出几个 cluster 做代表 → 组内所有人参与 → if clusters are dissimilar, high variability ⇒ need larger sample size so moE ↓
  - systematic s. → 在 n 个 unit 里, 选间距 k, 再从 1 到 k 选个起始点, 开选! (起点, k=3) ↳ but if 机器每 3 个坏, 那不准确

Fallacies
- ecological fallacy: 用 aggregate (group) 的 correlation 判断 individual 的
  - ecological correlation: consider grp characteristic → compute grp average 之后把他们连来算 r
  - aggregate points ⇒ 看好是否只是一个大组 /w slicing (≠ ecological), 还是真的有几个 aggregate point
  - → 可能和 individual 反方向, 也可能 overstate assoc. of individuals
- atomistic fallacy: 用 individual 的 correlation 判断 group 的

Studying non-linear 2 variables
- * r 未必准, 看 scatter plot 的形状才是关键!
- ↳ log → $y = cb^t$ ⇒ 两边加 ln: $\ln y = \ln(cb^t)$, $\ln y = \ln c + t \ln b$ (Y = ln y, x = t, m = ln b)
  1. compute $(t, \ln y)$
  2. plot $\ln y$ against t
  3. find linear regression line
  4. 有 line 就有 m & intercept, 可以反推出 ln c & ln b 里 c & b 是什么 也

☆ r ≠ gradient of regression line, $m = \frac{s_y}{s_x} r$

Linear Regression
- $Y = mx + b$ → 可以从 ↓ predict data; 塞一条直线进 graph



- e_i = residual of ith observation (error of i-th ")
- obtaining regression line: method of least squares, 找让 Σe_i^2 最小的线; 一定 pass through ave $(\bar{x}, \bar{y})$
- 只能用 X 判断 Y (因为线是 minimise x 的 e_i, 要用 Y 判断 x 就要找 e_i: Y min 的线)
- 超出已有的 x 的范围去做 prediction 会不准

Chapter 2 : Categorical Variable

- Analyse 1 variable → table & plots (dodge; stacked: no., 100%) → rate
- Analyse 2 variables → 2×2 table

outcome / treatment	Success	Fail	Row T.
X			
Y			
Column T.			

Rates
- marginal r. = category no. / grand total ↳ Rate (success)
- joint r. = 交集 no. / grand total ↳ Rate (Y & failure)
- conditional r. (among / given) = 交集 no. / subgroup after 1
  - * rate(eat | watch) + rate(× eat | watch) = 1, 所以不能用来看 association (eat | ×eat, watch)
  - ↳ find the rate of success for each treatment: Rate(success | X) vs Rate(success | Y)

Rules
- Symmetry Rule (A true B, 那 B 也成立)
  - R(A|B) > R(A|NB) ⟺ R(B|A) > R(B|NA)
  - R(A|B) < R(A|NB) ⟺ R(B|A) < R(B|NA)
  - = "   = "
- Basic Rule of R (subgroup rate & overall rate 的关系)
  - → overall R(A) lies btw R(A|B) & R(A|NB)
  - R(B) 越大, R(A) 越靠近 R(A|B)
  - R(B) = 50%, R(A) = $\frac{R(A|B) + R(A|NB)}{2}$
  - → R(A|B) = R(A|NB), 那 R(A) 也一样

Associations ≠ causation
- association present
  - R(A|B) > R(A|NB): +ve btw A & B — presence of A is stronger when B is present than when B is absent
    - ↳ R(A|B) > R(A|NB)
    - ↳ R(B|A) > R(B|NA)
    - ↳ R(NA|NB) > R(NA|B)
    - ↳ R(NB|NA) > R(NB|A)
  - R(A|B) < R(A|NB): -ve btw A & B, 也可以讲 +ve btw A & NB)
- association absent → R(A|B) = R(A|NB) 因果前面的都 OK

- Simpson's Paradox → trend appears in majority of several grps of data but disappears / reverses when groups combined (no association)
  - ↳ R(success | X) < R(success | Y) -ve (后果只看这个)
  - R(Large | X) > R(Large | Y) ∴ +ve
  - R(success | large) < R(success | small) ∴ -ve
  - treatment X — success; Large stone: confounding variable

- confounders → third variable assoc. to both dependent & independent variable
  - ∴ check confounder ⇒ 看和另外两个 variable 有没有 assoc.
  1. draw table of 2 variables we 调查
  2. calc. conditional rates btw variables
  3. compare assoc. btw variables
  4. 重复 切换 & 反都写出 assoc. → 第 3 个 variable 和 2 个都有联系! = confounder ⇓ slicing 讨论
  - ∴ 在 study 时要考虑 background info

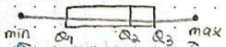
- * simpson ⇒ confounder; 没 simpson ≠ 没 confounder

Q: x is positively associated with y
A: Rate (y | x) vs. Rate (y | nx)

Chapter 3: Numerical Data

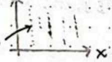
Univariate Data → orientation of data points (frequency distribution)

- summary statistics = descriptive statistics
- visualisation
  - ✓ illustrate great dif. accross frequencies of data points
  - ✓ 看shape of distribution
  - ✓ show no. of data points.
  - HISTOGRAM → center → mean, median, mode
    - bins
      - • avoid too big, data info lost
      - • avoid too small, too few data points in a bin
    - spread → outlier → deviation
      - std dev: S大, variability大, data far from center
      - range = largest − smallest
    - shapes
      - peaks → unimodal → bimodal → multimodal
      - skewed vs symmetrical (包括normal distribution) mean, med, mode close & near peak
      - Left vs Right
      - 拉拽mean, 一般mean < med < mode
      - robust statistics (not good measure of central tendancy, 易被outlier影响)
  - BOXPLOTS
    - ✓ compare distribution of dif. data sets
    - ✓ identify outliers.
    - outlier, min, Q_1, Q_2, Q_3, max, whiskers
    - outlier if (a) value > $Q_3 + 1.5IQR$ (b) value < $Q_1 - 1.5IQR$
    - shape → left skewed → right skewed; variability $\frac{(max-med)}{(med-min)}$ → 哪里tail longer (outlier更多)
    - center → compare relative position of med. & mean
    - spread → 用IQR 在不同数据间比较



Bivariate Data → statistical → 只能给出ave. value → association

- deterministic: 知道x, 可以算出y的
- SCATTER PLOT
  - describe relation: = strength + direction + form of association
  - ordered pair (x, y)
  - direction: +ve, -ve, neither
  - strength: strong, weak (因为没有trendline 看overall关系)
  - form (general shape): linear, non-linear (quadratic, exponential)
- CORRELATION COEFFICIENT (measure of linear assoc. btw 2 variables)
  - properties
    - • $-1 \le r \le 1$
    - • sign = direction ↳ +ve ↳ -ve ↳ = 1/= −1 ⇒ perfect ↳ = 0 ⇒ no linear (可能非线性/random)
    - • magnitude = strength: $0 \sim \pm 0.3$ = weak; $\pm 0.3 \sim \pm 0.7$ = moderate; $\pm 0.7 \sim \pm 1$ = strong
  - computing r
    - ① $\bar{x}, S_x$ → 加C不影响r
    - ② convert x & y 值 all to std units. $\frac{x-\bar{x}}{S_x}$ 乘正数不影响r值
    - ③ compute (xy) for 每组
    - ④ $r = \frac{1}{n-1}\sum xy$ → x, y互换不影响r
  - limitations
    - ① assoc. ≠ causation, 只是statistical relationship
    - ② × non-linear association ↳ cubic, quadratic, logarithmic
    - ③ outlier affect r significantly (可能强/削弱, 可能没影响, 取决于outlier在哪里)



Variables: attribute that can be measured → Identifier v. ↳ names

- Categorical v.
  - ordinal ↳ 1 to 5, how happy are you? — difference is subjective, algebraic calc. ✗ use
  - nominal ↳ cats, dogs, colour
- Numerical v. — ✓ arithmetic operations
  - Discrete ↳ no. of children
  - Continuous ↳ mass, time.

Chapter 4: Statistical Inference

Probability → repeatable & 可写下所有 possible outcome

- $P(A \cup B) = P(A) + P(B)$; $P(A \cap B)$, 交集 = 0
- sample space → all possible outcomes → P(sample space) = 1 (rule ②)
- uniform probability → assign probability so equal probability is assigned to every outcome
  - 之前的结果会影响下个的结果 gambler's fallacy
- events → subcollection of sample space → $P(E) = \frac{\text{no. of event发生}}{\text{no. of repetition of experiment}}$ → $0 \le P(E) \le 1$ (rule ①)
  - * outcome can be considered event but not all events are outcomes
  - rule ③: if E & F mutually exclusive, then $P(E) + P(F) = P(E \cup F)$
- independence (不同于mutually exclusive)
  - 用rate的想法, 相当于A & B are not associated
  - A & B independent, $P(A) = P(A|B)$, $P(A|B) = \frac{P(A \cap B)}{P(B)}$, $P(A) = \frac{P(A \cap B)}{P(B)}$
  - $P(A) \times P(B) = P(A \cap B)$ — 只有A & B independent 才能用的公式
- conditional probability: $P(E|F)$ = probability of E given F (mcm. new sample space)
  - if $E \cap F = 0$, $P(E \cap F) = 0$
  - $= \frac{P(E \cap F)}{P(F)}$
  - if $P(F) = 0$, $P(E|F) = 0$
  - conditional prob = conditional rates: $P(A|B) = \text{Rate}(A|B)$
  - prosecutor's fallacy: 把 $P(E|F) = P(F|E)$, 实际上有条件: $P(A \cap B) = 0$ 或 $P(A) = P(B)$; 才会 $\frac{P(E \cap F)}{P(F)} = \frac{P(F \cap E)}{P(E)}$
- conditionally independent: A & B are conditionally independent given an event C, $P(C) > 0$ if $P(A \cap B|C) = P(A|C) \times P(B|C)$

Summary statistics

- Measures of central tendency
  - mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$; $\frac{A}{A+B}\bar{x}_A + \frac{B}{A+B}\bar{x}_B = \bar{x}$ — is closer to subgrp mean w/ higher weightage
  - 可以把proportion想成mean
  - mode
  - median: 1, 2, 3, 4, 5; 1, 2, 3 | 4, 5, 6; 由 ±k, m' = m ± k; 由 ×c, m' = m × c
- Measures of dispersion
  - Standard deviation: $S_x = \sqrt{\text{variance}}$, $S_x = \sqrt{\frac{\sum(x_i与\bar{x}之差平方)}{n}}$, $S_x \ge 0$
    - 由 always non-negative; 由 ±k, S_x不变; 由 ×c, $S_x' = |c|S_x$
    - coefficient of variation $= \frac{S_x}{\bar{x}}$, $\bar{x} \neq 0$
  - Interquartile range: $IQR = Q_3 - Q_1 \ge 0$ — spread of middle 50%, $IQR \ge 0$
    - 由 always non-negative; 由 ±k, IQR不变; 由 ×c, $IQR' = |c|IQR$
    - * odd no. data, exclude median for Q_1 & Q_3
    - * even no. data, include the two no. used to calc. median

③ choose a good design — Study Designs

- Experimental = controlled (ethical?) → treatment vs control grp (混杂)
  - Random assignment (overall status is similar btw each grp, minimise confounders)
  - Blinding → avoid placebo effect (subjects) → avoid bias (assessors)
- Observational
  - → no direct manipulation
  - → exposure vs non-exposure grp
  - → subjects assign themselves
  - → provides evidence on association not causation

- • Law of total probability → if E & F are mutually exclusive, $E \cup F = S$, then $P(G) = P(G|E) \times P(E) + P(G|F) \times P(F)$
- • conjunction fallacy → if one believes that $P(A \cap B) > P(A)$ or $P(A \cap B) > P(B)$ — 实际上, chance of one event happening > chance of two events happening
- • base rate fallacy → 没有考虑base rate就做计算 — rate of occurence of some trait in a population

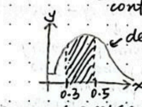
Contingency table来avoid fallacy & 算conditional p.

	+ve test	-ve test	Total
Drunk	④	⑧	③ 用base rate得出
Sober	⑤	⑦	②
Total	⑦	⑨	① assume

(check是否)

- sensitivity of test = true positive rate ↓ P(tested +ve | infected)
- specificity of test = true negative rate ↓ P(tested -ve | not infected)

- discrete random ↳ dice 点数, P(Y=1), P(Y=2)...
- continuous random → density curve → represented by interval of values of area under density curve; $P(0.3 \le Y \le 0.5)$; $P[a, b]$ = Area $[a, b]$



statistical inference: 用sample to draw inferences / conclusions abt pop. — 如果available data is representative 才可以做inference

- sample statistics = pop. parameter + bias + random error → 尽量少bias后 = pop. parameter + random error (∵ sampling variability 所有的small difference)
- 尽量减少
  - selection → good sampling methods (simple random) → good sampling practices (good frame)
  - → non-response bias → high response rate
- tools for statistical inference (conducted on probability-based samples)
  - confidence interval (CI) → range of values that likely contain a pop. parameter based on certain confidence level (CL) % deg. of confidence
    - proportion 只有取census才懂 pop. proportion unknown; 用sample之后breakdown
      - ① simple random. x个单位, 得到其proportion为y
      - ② y = pop. proportion ± random error
      - ③ to construct CI, margin of error: $p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$ (sample pop., z value, sample size)
      - z^*的值取决于CL: ↳ 95% = 1.96, ↳ 90% = 1.645
    - mean: $\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$ (sample std dev, ↓ t-distribution, sample mean)
    - properties: 由 sample size↑, random error↓, CI↓; 由 CL↑, CI↑
    - 如何解读CI: $p^* \pm$ moE; [|] → length of CI = 2MOE respective 的; 若CL = 95%, 就有95% confident p lies within CI — 如果sample 100次, 95次p会在CI里
  - hypothesis testing → decide if there is sufficient support for a particular hypothesis (*reject claim时, null = claim is true; collect evidence to reject null for alt)
    - ① identify question, state null & alternative hypothesis — 假设刚好chance 达到才有实验里的结果, 真实pop.不是; H_1: pop. parameter 不等于 null; H_0: pop parameter = null value
    - ② set significance lvl of test, 一般5% (α) — 如果p-value < α, reject null in favour of alt; p-value > α, not enough evidence to reject null (不代表accept null)
    - ③ find sample stats ↳ sample proportion, sample mean
    - ④ calculate p-value → 用software算 (P(obtain result as / more extreme than observation in direc. of alt assuming null is true))
    - ⑤ conclude
    - → 也可以test association: Null: no association; Alt: ✓; 之后算p value for chi² test, 若 p-value ≮ α, insufficient evidence to reject null