CH. 1: Data Analysis in an Evidence-Based Practice Environment

Independent variable:
Influence on, an outcome. eg. smoking.

Dependent variable:
The outcome of interest, hypothesized to depend on, or be caused by, the independent variable. eg. Lung cancer, Gender (IV) impact self-esteem (DV).

Level of Measurement:
A system of classification with four types of measurement rules that affect the kind of statistical analysis that is appropriate:

- **Nominal**: Eg. hair colour, gender, religion
- **Ordinal**: Eg. degree of pain, scale of self-esteem
- **Interval**: Eg. temperture
- **Ratio**: Eg. medication dose (number of milligrams, number of pills) , pulse

Descriptive statistics:
Describe and summarize data about the sample
Eg. Percent female in the sample, average weight of participants

Inferential Statistics: ***(what we learn from this course)***

- Parameter 參數: Eg: Average daily caloric intake of all 10-year-old children in New York
- Statistic: Eg: Average daily caloric intake of 300 10-year-old children from three particular NY schools

Inferential statistics, based on laws of probability, help researchers draw objective conclusions about a population, using data from a sample.
Often used to test hypotheses (predictions) about relationships between variables

CH. 2: Frequency Distributions: Tabulating and Displaying Data

Frequency distribution: Help impose order on the data.

Data Value	Frequency (f)	Percentage (%)	Cumulative Percentage
1	10	10.0	10.0
2	20	20.0	30.0
3	40	40.0	70.0
4	15	15.0	85.0
5	15	15.0	100.0
TOTAL	100	100.0	

10+20
10+20+40
10+20+40+15
10+20+40+15+15
*

Shapes of Distributions:

- **Modality:** most frequely number
- **Symmetry:**


Positive skew:
Longer tail trails off to right (fewer people with high values, like for income)


Negative skew:
Longer tail trails off to left (fewer people with low values, like age at death)

Eg. Skewness Index Examples *
Skewness index = 0.80
Standard error = 0.33 → significant skew
0.33 x 2 = 0.66

Skewness index = -0.72
Standard error = 0.34 → 0.34x2 = 0.68 significant skew

CH. 3: Central Tendency, Variability, and Relative Standing

Central Tendency: (specific)

- Three alternative indexes:
  - The mode:
  - The median:
  - The mean (N):

Scale	Central Tendency Index	Variability Index
Nominal	Mode	--
Ordinal	Median	Range, *IQR*
Interval and ratio	Mean	Standard deviation, Variance

The Mean
Most frequently used measure of central tendency—usually preferred for interval- and ratio-level data.*

$M = \Sigma X \div N$ (Σ: sum; N: number of people)

Age = 26, 27, 27, 28, 29, 30, 31
$M = 198 \div 7$
$= 28.28$
$\doteq 28.3$

$M = \bar{X}$ (simple mean, x bar)
μ (population mean, mu)

The Range
The difference between the highest and lowest value in the distribution*
110 120 130 140 150 150 160 170 180 190
range = 190 - 110 = 80

The Interquartile Range
Based on *quartiles*

- Lower quartile (Q_1): Point below which 25% of scores lie
- Upper quartile (Q_3): Point below which 75% of scores lie
- Important in evaluating outliers

11 | 22 | 33 | 44
Q_1 Q_2 (median) Q_3

$IQR = Q_3 - Q_1$

The Standard deviation (SD - Group data):
An index that conveys how much, on average, scores in a distribution vary.
"How far it can go from mean".*

$$SD = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}}$$

ex.

X	$\bar{X}$	$X-\bar{X}$	$(X-\bar{X})^2$
6	7	-1	1
4	7	-3	9
7	7	0	0
8	7	1	1
10	7	3	9
35		0	Σ=20

5 people

$SD = \sqrt{\frac{20}{5-1}} = \sqrt{\frac{20}{4}}$
$= \sqrt{5}$
$= 2.236$

Variance 變動
An important variability concept in inferential statistics, but not used descriptively.*

$$\frac{\Sigma(X-\bar{X})^2}{N-1} = SD^2$$

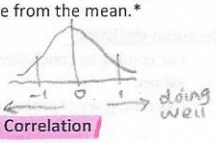
Percentiles and Outliers
Outliers are often defined in relation to percentiles

- A mild outlier is a score that is between 1.5 and 3.0 times the value of the *IQR*, below Q_1 or above Q_3
- An extreme outlier is a score that is greater than 3.0 times the value of the *IQR*, below Q_1 or above Q_3

Standard Score (Z score – Individual data)(measure standard deviation)
Another index of relative standing helpful in interpreting raw scores.
A score expressed in standard deviation units, in relative distance from the mean.*

individual score
$Z = \frac{X - \bar{X}}{SD}$ (X: individual score; $\bar{X}$: mean)

ex.
$\bar{X} = 500$
$SD = 100$
$X = 600$
$Z = \frac{600-500}{100} = 1$



CH. 4: Bivariate Description: Crosstabulation, Risk Indexes, and Correlation

Bivariate descriptive statistics:
Used to describe relationships between two variables.
Eg. Height and weight, Smoking status and lung cancer incidence

Crosstabulation 交叉製表
Example of a Crosstab table*:

10
50.0% → Row
20.0% ↓ column
10.0% whole group

Risk Index Scenarios
Eg*.

smoker \ cancer (column)	Yes	No
Yes (control)	a	b
No (experimental)	c	d

Row: Independant v.
cancer: dependent v.

swelling	Yes	No	
ex control 16mm	a 20	b 80	100
NE exp 25mm	c 10	d 90	100
	30	170	

Absolute Risk: Absolute risk is the proportion of people with a negative outcome.*

$AR_E = \frac{a}{a+b}$ $AR_E = \frac{20}{20+80} = 0.2 = 20\%$ ← more risk

$AR_{NE} = \frac{c}{c+d}$ $AR_{NE} = \frac{10}{10+90} = 0.1 = 10\%$

Absolute risk reduction: The absolute difference between the two risk groups.*

$ARR = AR_E - AR_{NE}$ $ARR = 0.2 - 0.1 = 0.1 = 10\%$ reduction in risk of swelling with the longer needle.

Relative Risk: The ratio of absolute risks (adverse outcomes) in the two groups.*

$RR = AR_E \div AR_{NE}$ $\frac{0.2}{0.1} = 2.00$ children immunized with the shorter needle were twice as likely to have swelling relative with the longer needle.

Relative Risk Reduction: The proportion of baseline risk that is reduced through non exposure (or receipt of an intervention).*

$RRR = ARR \div AR_E$ $\frac{0.1}{0.2} = 0.5$ being immunized with the longer needle reduced the relative risk of swelling by 50%.

Odds: The proportion of people in each risk group who have the adverse outcome, relative to the proportion who do not.*

$Odds_E = \frac{a}{b}$ $\frac{20}{80} = 0.25$ outcome

$Odds_{NE} = \frac{c}{d}$ $\frac{10}{90} = 0.111$ outcome

16mm needle odds of swelling 1/4
25mm needle, the odds of swelling 1/9

Odds Ratio: The ratio of the two odds.*

$OR = \frac{Odds_E}{Odds_{NE}}$ $\frac{0.25}{0.111} = 2.25$ odds of swelling are 2.25 times higher with the shorter one as with longer one.

Number Needed to Treat: Estimate of how many people would need to avoid the exposure (or get a treatment) to prevent one negative outcome.*

$NNT = \frac{1}{ARR}$ $\frac{1}{0.1} = 10$ children would need to be immunized with 25mm needle rather than 16mm one to prevent one case of swelling

Correlation: Correlations between two quantitative variables can be graphed in a scatterplot.*
maximum number: +1 positive correlation
-1 negative correlation
if number is 1.67 wrong!!

Correlation Coefficients

- Relationships between two variables.
- Most widely used correlation coefficient: Pearson's product moment correlation coefficient. → tell you +1 or -1
- Often called **Pearson's r** * r = .75 "stronger than" r = .50
- Pearson's r is computed with variables that are interval- or ratio-level measures.
  - 1.00 = Perfect positive relationship
  - .35 = Weak/moderate positive relationship
  - .00 = No relationship
  - -.20 = Weak negative relationship
  - -.70 = Strong negative relationship

r^2
Indicates the proportion of variability in one variable accounted for or explained by the second variable.
Eg. *r* between height and weight = .60*
$r^2 = .36$ → 36% the variation in weight is accounted for by height.

64% of variation in weight is accounted for by other factors.

significance : reject (null hypothesis)
not significance : retain

Correlation Matrix

An efficient way to display several correlation coefficients.

scale: interval + ratio
sig = significance

significance correlation	0.05
	0.04
	0.03
	0.02
	0.01
even more significance ←	0.000001

positive
thick!

Pearson correlation = .179**
sig = .000 (significance)
N = 919 (number of people in study.)

CH. 5: Statistical Inference

Descriptive statistics: For describing samples.

Inferential statistics:

- For coming to conclusions about what is probably true in a population, based on sample values.
- Uses the laws of probability to provide guidance on what is *probably true*.

Probability of an event (*p*) is expressed as a proportion

The probability of drawing a red card from a normal shuffled deck:
Ways "red" event can occur = 26 = 0.50
Total number of possibilities = 52
p = .50: There is a 50-50 chance that the card will be a red suit

if throw away 1 red card: → 25/51

Multiplicative law provides this formula:

p (A then B) = *p* (A) × *p* (B)

The probability of drawing two red cards consecutively: or (adding) or then (multiply)
p (red, then red) = *p* (red) × *p* (red)
p (red, then red) = .50 × .50 = .25
p = .25: There is a one in four chance of drawing two red cards in a row

What is the probability of drawing 10 red cards in a row if the null hypothesis of a fair deck is true?
p = .50 × .50 × .50 × .50 × .50 × .50 × .50 × .50 × .50 × .50 = .001
Only 1 in 1,000 draws of 10 cards from a fair deck would yield all red cards

Probability distributions

- Similar to frequency polygons (or histograms).
- They graph the probabilities of all events that could occur
  So, the total area of a probability distribution = 1.0

Probability density function = Probability distribution for continuous variables

Eg.* (2) tail (3) mean

Z = 1 or greater → A = 0.159 (Probability = 0.159)

Z = 1 or less → A = 0.341 + 0.5 = 0.841

percentail: 84.1%

Z = -0.3 → A = 0.382 → percentail: 38.2%
(who do worse than me.)

Mean as statistic in population called:

Stable

Sampling Error

- Sample means from a population tend to fluctuate from one sample to another because of sampling error.
- Never be exactly 100.0
- Sample mean might be 98.2 or 101.6 or 99.7

Sampling distribution is a type of probability distribution

Sampling distribution of the mean

- A type of probability distribution.
- It's the distribution of an infinite number of sample means from the population.

age of 2nd year of nursing student across Canada

Eg.* - sample = 50
- mean = 20yrs

test group	2nd	3rd
	26	23
22		

plugin → 24

infinite of data (mean)

Central Limit Theorem

Mathematic formulation, shows that the mean of a sampling distribution of the mean always equals the population mean.

Standard error of the mean (*SEM*)

- Standard deviation of a theoretical sampling distribution.
- Larger the *SEM*, the less likely it is that a sample mean is a good estimate of the population mean. (Less: estimate will be good)
- *SEM*s are never known (due to theoretical), but can be estimated.

Formula for *SEM* Estimate*

$SEM = \frac{SD}{\sqrt{N}}$ → sample's standard deviation; → sample size

SEM estimate: if sample was larger SEM would be smaller!

ex. N = 25, $\bar{x}$ = 101, SD = 10.5

$SEM = \frac{10.5}{\sqrt{25}} = 2.1$

95% ish (probability 95% ish (.95) fall between ±1, -1 deviation)

96.8 | 101 | 105.2

(101 - 2.1 - 2.1) (101 + 2.1 + 2.1)

∴ 95% of infinite number is between 96.8 and 105.2

Statistical Inference 推理 Approaches

- Hypothesis testing:
  Used to for nursing research, now has change.
- Parameter 參數 estimation:
  Used to estimate a population value—e.g., a mean, percentage, or odds ratio.
  - **Point estimate:**
    - Calculation of a "single value" as the estimate of the parameter.
    - Simply the value of the descriptive statistic, like a mean.
  - **Interval estimate:**
    - More accurate! 95% of population fall between these values!
    - Provides a range of values within which the population value has a specified probability of lying.
    - involves constructing "confidence intervals" around the point estimate.

confidence interval: 95% CI = $\bar{x} \pm 1.96 \times \sigma_{\bar{x}}$ (this number never change; will given this number; symbol actual SEM in pop.)

95% CI = $\bar{x} \pm$ (t) × SEM → $SEM = \frac{SD}{\sqrt{x}}$

t table: N-1=df

ex. N = 25 sample, interval estimate?
sample mean $\bar{x}$ = 50
SD = 3

$SEM = \frac{SD}{\sqrt{N}} = \frac{3}{\sqrt{25}} = 0.6$

N-1=24 (df)

a: 95% CI = $\bar{x} \pm$ (t) × SEM
= 50 ± 2.06 × 0.6 = 50 + 1.236 = 51.236 (higher confidence)
= 50 − 1.236 = 48.764 (lower confidence)

∴ 95% confidence population mean fall between 48.764 and 51.236
∴ Point estimate = 50

t Distribution

- Bell shaped and symmetric.
- A sample size increase, the t distribution is very close to a "normal" distribution.
- Small samples, the tails of a *t* distribution are "fatter".

*CI*s around Proportions

- Frequently computed around proportions/percentages and risk indexes like Relative Risk and the Odds Ratio.
- Binomial distribution.
- Almost always done by computer.
- The larger the sample size, the smaller the *CI*.

Hypothesis Testing

- Common use (retain/reject).
- Second broad approach to statistical inference.
- Help researchers make objective decisions about accepting or rejecting a null hypothesis.
  Eg.
  Cigarette smoking is unrelated to lung cancer (always start assumed to be innocent!)
  Cigarette smoking is related to lung cancer

There is always a risk of error (2 types)

- ***Type I error (*the only type of error can make):***
  - The null hypothesis is really "true" in the population, but the researcher "rejects" it.
  - Incorrect reject, which shouldn't!
  - Controlled through the level of significance (alpha, α)* α = 0.05, 5% error (can be α = 0.01, 1% error)
- ***Type II error:***
  - The null hypothesis is really "false" in the population, but the researcher "accepts" it.
  - Incorrect retain, which shouldn't!
  - Probability of committing a Type II error is called beta (β)
  *

β = P(Type II)
power = 1 − β

bigger sample size = more power!

Power

- **3 factors influence power:**
  - Sample size → the easiest one to control.
  - Effect size
  - Alpha level
- Power and the probability of making a Type II error are inversely related.
- ↓ the probability of making a Type II error, ↑the probability of making a Type I error

determine type II error ← Power = 1 − β → type II error

ex. Power = 0.8 (good) Power = 1 − β, 0.8 = 1 − ? → 0.2 (20% type II error)

One-Sample *t* Tests

- A statistical test that tests the null hypothesis that the population mean is a "specific value".
- Tests whether the sample mean differs from a real or hypothesized population mean. (compare "Sample mean" to "Population mean")

t statistic calculation for one sample t test (hypothesis test)

$t = \frac{\bar{x} - M(\mu)}{SEM}$

ex. population mean M = 7, N = 25, sample mean $\bar{x}$ = 8, SD = 1

$SEM = \frac{D}{\sqrt{N}} = \frac{1}{\sqrt{25}} = 0.2$

① $t = \frac{\bar{x} - M}{SEM} = \frac{8 - 7}{0.2} = 5$

calculated T: 5

② df = N − 1 = 25 − 1 = 24 (check t table (look at significance) level 0.05)

③ critical T: 2.06 (t table T)

④ (tail) reject ← Retain → (tail) reject (2.06)
compare = $\frac{\bar{x} - M}{SD} = \frac{8 - 7}{1} = 1$ → significance higher

∴ Reject: significance more