Sampling Types

Simple Random Sampling – without replacement, every unit has chance of selection
Systematic Sampling – Random starting point, uniform interval. Simple, but lack representation
Stratified Random Sampling – Pop. divided into strata, SRS in each strata. Good representation, but need data to divide
Cluster Sampling – Pop. Divided into clusters, and clusters are selected randomly. Convenient, but costly, maybe high variability
Non-probability - convenience/volunteer sampling, selection bias and non-response bias

Study Design

Experimental Studies – Independent variable manipulated to establish cause-and-effect. **Randomised assignment**(experiment) into treatment/control used to control confounders. **Blinding** and **Placebo** (can be known effect) to reduce bias.

Observational Studies – Observe variables of interest, no manipulation. Cannot establish cause-and-effect. Treatment & control self-assigned.

Basic Rates

Basic rate – rate(x) = $\frac{|x|}{n}$

Joint Rate – rate(x and y) = $\frac{|x \cap y|}{n}$

Conditional Rate – rate(x|y) = $\frac{|x \cap y|}{|y|}$

Symmetry Rule

$rate(A|B) > rate(A|\bar{B}) \leftrightarrow rate(B|A) > rate(B|\bar{A})$ <=> A positively associated to B, B is positively associated to A

$rate(A|B) < rate(A|\bar{B}) \leftrightarrow rate(B|A) < rate(B|\bar{A})$ <=> A negatively associated to B, B is negatively associated to A

$rate(A|B) = rate(A|\bar{B}) \leftrightarrow rate(B|A) = rate(B|\bar{A})$ <=> A not associated to B, B not associated to A

Statistic	Formula	+c	*c	Other notes
Mean $\bar{x}$	$\frac{x_1 + .. x_n}{n}$	$\bar{x}$ + c	$\bar{x}$ * c	
Variance	$\frac{(x_1-\bar{x}) + .. + (x_n - \bar{x})}{n+1}$	same	c^2*variance	Absolute spread in $unit^2$
Std. Dev s_x	$\sqrt{Variance}$	same	$c \times s_x$	Absolute spread in Unit
Coeff. Of Variation	$\frac{s_x}{\bar{x}}$	decreases	same	Degree of spread relative to mean
Median	Middle Value of datapoint (/2 in-between)	median + c	median * c	
IQR	$Q_3 - Q_1$	same	|c| * IQR	Never negative
Mode	Most common data point, peak	Duh	Duh	
Co-variance	-	same	Covar*c (1 axis) Covar*c^2 (2 axis)	Used for correl
correl	$\frac{covariance}{s_x \times s_y}$	No change	No change	Used for linear regression

Basic Rule on Rates

Rate(A) will **always** lie between rate(A|B) and rate (A|$\bar{B}$)

The closer rate(B) is to 100%, the closer rate (A) is to rate(A|B)

If rate(B) = 50%, then rate(A) = $\frac{1}{2}[rate(A|B) + rate(A|\bar{B})]$

If rate(A|B) = rate(A|$\bar{B}$), then rate(A) = rate(A|B) = rate(A|not B)

(Number of observations in B outweigh not B, so naturally rate leans towards number of observations in B). (2/3 rates is all u need)

Simpson's Paradox

When a majority trend reverses upon combining subgroups

rate(X|M) > rate(Y|M), rate(X|F) > rate(Y|FM), but rate(Y)>rate(X)

<u>**Solving Simpson's Paradox Questions:**</u>

1. **Identify** relevant **subgroup** – the group that will be **combined**
2. **Identify** relevant **rate** – the researched rate for the qn
3. Calculate rate in subgroup, but kind of ignore the subgroup
4. **Look for association in subgroups** individually. Does a trend appear?
5. Combine the subgroups and calculate rate. Does association disappear/reverse?

Confounders

Are variables associated to both the independent and dependent variable

<u>Checking for confounders:</u>

1. Is the variable associated to an independent variable?
2. Is the variable associated to the dependent variable
3. If yes, then confounder

Randomised assignement mitigates this.

Five Number Summary & Boxplots

1. Minimum
2. Q1
3. Median
4. Q3
5. Maximum

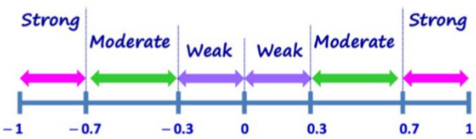
Boxplots are made by drawing a box from **Q1 to Q3**, drawing a line where the median is, and then extending a line from the box to the smallest and largest values that are not outliers. Outliers are marked with dots.

Outliers

1. Data point **greater** than Q3 + 1.5 * IQR
2. Data point **lesser** than Q1 - 1.5 * IQR

Don't be myopic and focus on upper-bound outliers. Lower-bound outliers exist too.

Correlation Coefficient



Correlation coeffients tells us how **linearly** associated 2 variables are. Coefficients of 1/-1 are perfectly correlated variables. **Association is not causation**. Correlation Does not tell us anything about non-linear associations. Outliers may or may not affect the value

Standard Units

Used in calculation of correl.

$$\frac{x - \bar{x}}{s_x}$$

It standardizes the units such that it is independent of units/scales of the variables themselves. This ensures their association is captured accurately.

Modelling non-linear associations

We can manipulate non-linear associations into linear ones by manipulating the axes. If a relationship is exponential, we can plot ln y against ln x. So, we get a regression line like y = lnc + x ln b

Linear Regression

Fitting an independent variable (x-axis) and dependent variable (y-axis) onto a straight line. Allows for **prediction** of y given x, however should **only be used** within the span of the x-axis.

Slope of regression line is calculated using:

$$\frac{Covariance(X,Y)}{Variance_x}$$

If one axis*c, gradient of slope changes by c * slope. No change for addition. If two axis * c, unchanged.

Fallacies

Ecological Fallacy – Using a correlation noted <u>among subgroups at the aggregate</u> level (*neighbourhoods that have higher income vote blue)*, to make inferences about the <u>association at the individual</u> level (*wealthy **individuals** vote blue)*

Atomistic Fallacy – Using correlation noted <u>among subgroups at the individual level</u> (*individually, a researcher sees that wealthier individuals vote blue*), to make inferences about the association at the aggregate level (*researcher concludes wealthy **regions** vote blue*)

Neither – If no subgroup association was looked at, there's a good change this is not ecological/atomistic fallacy – overgeneralization? Additionally, if it's not between **TWO** variables, it is not this fallacy.

Conditional Independence

Two events are conditionally independent iff:

$$P(A \cap B \mid C) = P(A|C) * P(B|C)$$

Probability

Conditional Probability – P(E|F) = $\frac{P(E \cap F)}{P(F)}$

Joint Probability – $P(E \cap F) = P(E|F) * P(F)$

Simultaneous Probability - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Mutually Exclusive Union - $P(A \cup B) = P(A) + P(B)$

Independence & Joint Probability – $P(A) = P(A|B) \rightarrow P(A \cap B) = P(A) * P(B)$

Law of Total Probability

If for some sample space S, E and F are mutually exclusive and $E \cup F = S$, law states:

$$P(G) = P(G|F) * P(F) + P(G|E) * P(E)$$

Looks intimidating but all it's saying if that E and F are the **only two outcomes**, and E and F can **never happen at the same time**, that is the overall probability of G.

*E.g: P(defective | A) = 0.05 and P(defective | B) = 0.1. P(A) = 0.7 and P(B) = 0.3. P(defective)=0.05*0.7 + 0.3 * 0.1 =0.065*

More Fallacies

Conjunction Fallacy – Believing P(A and B) > P(A)

Base Rate Fallacy – Ignoring rate of occurrence of some trait when making a decision

Prosecutor's Fallacy – Taking P(A|B) = P(B|A). This is only true when P(A)=P(B), or P(A and B) = 0

Sensitivity and Specificity

Sensitivity – True positive rate, P(Positive | Infected)

Specificity – True negative rate, P(Negative | Not infected)

To get one from another you need **base rates**.

Random Variables

Discrete random variables – have gaps, like dice roll

Continuous random variables – Area under graph, integration over interval

Fundamental Rule for using Data for Inference

Data from samples can be used to make inferences about a larger group if **bias** is minimized (good sampling frame, SRS). If there is little bias, then:

Sample statistic = pop. param + random error

Confidence Intervals

CI is given: $p^* \pm z^* \sqrt{\frac{p^* * (1-p^*)}{n}}$

Z-value is based on **confidence level**, 95%=1.960

Confidence interval tells us "**We are 95% confident that population param lies within this confidence interval**"

If we take another sample using the same way and construct another confidence interval, 95/100 samples will contain the <u>**true**</u> population parameter.

Properties: **Larger** sample size, **smaller** random error, **narrower** CI. **Higher** confidence, **wider** CI.

Hypothesis Testing

Null Hypothesis – Corresponds to the baseline assumption/status quo

Alternative Hypothesis – corresponds to what you're trying to prove

Example – *In our sample, treatment X is positively associated to recovery (vs treatment Y). Our* ***null****: P(Recovery|X) = P(Recovery|Y). Our* ***alternate****: P(Recovery|X) > P(recovery|Y)*

Steps for Hypothesis testing:

1.State null & alternate

2.Set sig. level *(10%/5%/1%)*

3.Calc P-value

4. Make conclusion

Conclusions from Hypothesis test:

A)P-value **<** sig. level (*0.1/0.05/0.01)*

Reject null in favour of alternate

B) P-value **>** sig. level

Inconclusive, Insufficient evi. for **alternate**

Chi-Squared and T-test

For Chi-squared and t-test, steps are basically the same.

For chi-squared note that the **null** hypothesis is that there is **no association** between X and Y, and our alternate is that there exists an association.

Chi-squared test only tells us that the association exists, it doesn't say anything about direction or strength

Scatterplots

Shows relations between 2 variables

Independent usually on X axis, dependent on Y.

Direction tells us how the graph looks. Positive slope? Negative? Erratic?

Form is the general shape of the scatterplot. Straight line? Curved?

Strength indicates how closely the data follow the form. Scattered or tight?

P-value

P-value is the the porbability of obtaining a test result **at least as extreme** as the result observed, assuming the **null hypothesis is true**. Note that *how extreme* a result is is based on which direction the tail of the test is pointing to.

If your hypothesis are null: P(E) = 0.5 and alternate: P(E)<0.5, and your observed result is 0.7, anything less than 0.7 is at least as extreme.

If your value follows the arrow, it is more extreme. But always **consider the result**.

T-test formula (pop. Mean)

$$x \pm z^* \times \frac{s_x}{\sqrt{n}}$$

Z-values for CIs

90%	95%	99%
1.645	1.960	2.576

Boxplots

Tells us five number summary. Note that 25% of data should be above and below the boxes, as they represent Q1 and Q3

Shape: Allows us to see variability. More outliers on upper end?

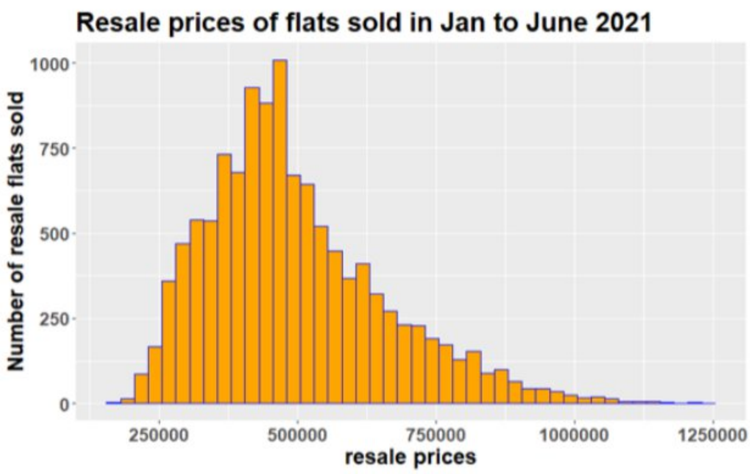
Center: Easily see median and compare with mean

Spread: IQR gives good visualization of spread

Making Associations

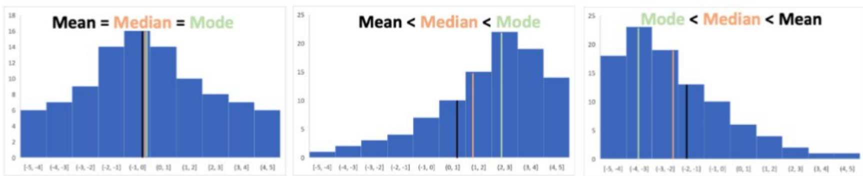
Positive Association	Negative Association
$R(A\|B) > R(A\|\bar{B})$	$R(A\|B) < R(A\|\bar{B})$
$R(B\|A) > R(B\|\bar{A})$	$R(B\|A) < R(B\bar{A})$
$R(\bar{A}\|\bar{B}) > R(\bar{A}\|B)$	$R(\bar{A}\|\bar{B}) < R(\bar{A}\|B)$
$R(\bar{B}\|\bar{A}) > R(\bar{B}\|A)$	$R(\bar{B}\|\bar{A}) < R(\bar{B}\|A)$

Histograms



Shape – Peaks and Skewness: Histogram can have one peak (unimodal) or two peaks (bimodal). Skewness is the direction that the tail (if any) is pointing towards. Above diagram is right skewed.

Central Tendency:



Spread – Standard Deviation and Range: Lower standard deviation = more defined peak. Range gives info on variation.