GER1000 Quantitative Reasoning
Help Sheet for Final Examination

Disclaimer: I hereby affirm that any errors found in this sheet is solely due to my own human error and are not committed on purpose in order to "snake".

Joshua of House Ursaia

Controlled Experiments

Experiment that compares between the response of a treatment group and a controlled group.

Randomized Control	Double-Blinding
Experiment that assigns subjects into the **control** and **treatment group** randomly. • Large number of subjects → likely that the two groups are similar in all aspect.	Experiment where the subjects do not know whether they are in the treatment or control groups; neither do those who evaluate the responses. • Guards against bias either in the responses or evaluations. • Minimizes confounding

Observational Studies

Experiment where the investigators **do not assign** the subjects into the treatment or control group.

Consequence	Effects of treatment may be confounded with the effects of factors that got the subjects into the treatment and control groups in the first place.

Rate

$rate(A \mid B)$	Rate of A among people with B
$rate(A \mid not\ B)$	Rate of A among people without B

Basic Rule on Rates

$rate(B)$ is always between $rate(B \mid A)$ and $rate(B \mid not\ A)$.

$rate(A) \to 100\%$	$rate(B) \to rate(B \mid A)$
$rate(B \mid A) = rate(B \mid not\ A)$	$rate(B) = rate(B \mid A)$
$rate(A) = 50\%$	$rate(B)$ is exactly halfway between $rate(B \mid A)$ and $rate(B \mid not\ A)$.

Yule-Simpson Paradox

Suppose that the population consists of several subgroups, and in each subgroup, $rate(B \mid A) > rate(B \mid not\ A)$.

When the **subgroups are combined**, it may happen that $rate(B \mid A) \leq rate(B \mid not\ A)$.

Association

A and B are positively associated	$rate(A \mid B) > rate(A \mid not\ B)$ $rate(B \mid A) > rate(B \mid not\ A)$
A and B are negatively associated	$rate(A \mid B) < rate(A \mid not\ B)$ $rate(B \mid A) < rate(B \mid not\ A)$
No Association	$rate(A \mid B) = rate(A \mid not\ B)$ $rate(B \mid A) = rate(B \mid not\ A)$

Confounders

A **confounder** is a third variable associated with both exposure and disease.

Methods to control for confounding factors

Slicing Method	Compare smaller groups which are relatively homogeneous with respect to the confounding factor.
Regression	Explained in detail later.

Types of Relationships

Deterministic	The value of the dependent variable **can be determined** from the value of the independent variable.
Statistical	The **average pattern** of one variable can be described with the value of the other variable.

Linear Regression

Linear Regression is used to investigate association between two continuous variables.

Linear Regression Line	The line of best fit to data, where the sum of squares of the distance of each data point to the line is minimum.

Linear Regression Coefficient, r $(-1 \leq r \leq 1)$

Sign	$r > 0$ → Positive association $r < 0$ → Negative association $r = 0$ → **No linear association** (may still have non-linear association)
Magnitude	The closer the value to 1 *or* − 1, the stronger the **linear association**
Not affected by	1. Interchange of the two variables 2. Adding a number to all values of a variable 3. Multiplying a positive number to all values of a variable

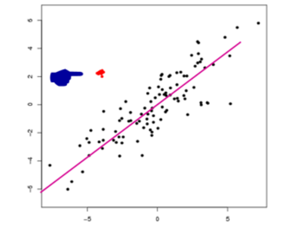
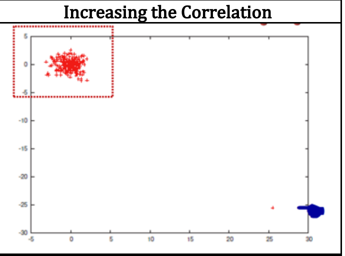
Points to take note:

- In general, the slope of the linear regression line ≠ the correlation coefficient.
- A high linear correlation coefficient does not necessarily mean that the relationship is linear.

- **Extrapolation:** Predicting the value of the dependent variable beyond the observed range of the independent variable is dangerous.

Impact of Outliers on Correlation

It is dangerous to exclude outliers from the analysis without understanding the causes of their occurrence.

Decreasing the Correlation	Increasing the Correlation
	

Attenuation Effect

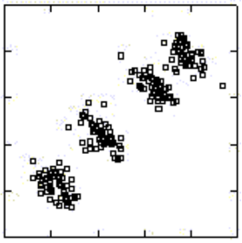
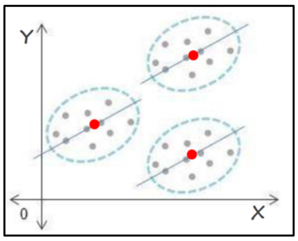
Range restriction in one variable could cause the correlation coefficient obtained to "**understate**" the strength of association between two variables.

Check for an "oval shape" in your scatter diagram. If you see it, you are less likely to have the attenuation effect.

Ecological Correlation

Correlation based on **aggregated data**, such as group average or rates. When the associations for both individuals and aggregates are in the same direction, the ecological correlation will typically "**overstate**" the strength of association in individuals, because the variability among individuals have been eliminated.

Fallacies

Ecological Fallacy	Atomistic Fallacy
	
Deducing the correlation of individuals from aggregate data.	Deducing the correlation of aggregate data from the correlation based on individuals.

Estimations

$Estimate = Parameter + Bias + Random\ Error$

Population Parameter	A numerical fact about a population.
Bias	Depends on the method of sampling.
Random Error	Depends on the size of sampling.

Sampling Frame

A list of sampling units intended to identify all units of a population.

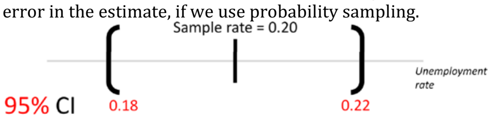
- Has to cover exactly or bigger than target population and has to be up-to-date.

Biases and their Causes

Selection Bias	Systematic tendency to **exclude** one kind of person or another from the sample, caused by: • Imperfect sampling frame (which excludes certain desired units) • Non-probability sampling methods
Non-Response Bias	Differences between non-respondents and respondents
Response Bias	Phrasing of the questions, tone, or attitude of the interviewers.
Other Types	Subjects may have a tendency to understate responses about undesirable social habits (ashamed).

Confidence Intervals

Confidence Interval is the range of values in which we are reasonably certain our unknown parameter lies in. It is helpful in providing information about the random error in the estimate, if we use probability sampling.



95% CI: 0.20 ± 0.02

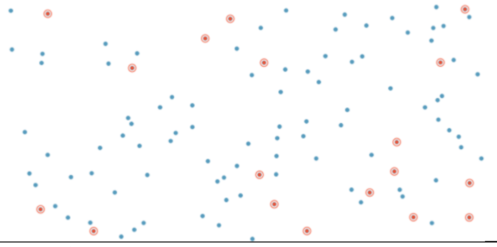
Interpretations	The experimenters are 95% confident that the range 0.18 and 0.22 contains the population parameter.
	95% of the researchers who repeat the experiment will have intervals that contain the population parameter.

Link to Random Error	For an experiment with a **larger sample size** → likely to have **smaller random error** → at the same 95% confidence interval, the experiment will have a **smaller range**.

Types of Probability Sampling

Simple Random Sampling

Draw units at random from a sampling frame without replacement, with every unit having an equal chance of being selected.

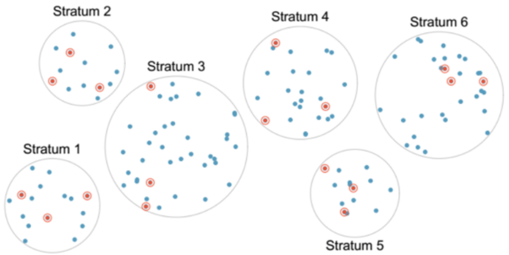


Systematic Sampling

Select a random starting point, r. Include every k^{th} unit after r into the sample.

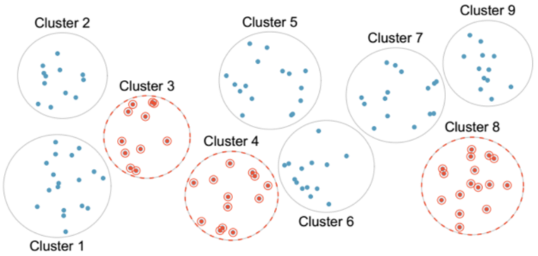
Stratified Sampling

Divide the population into homogeneous subgroups (strata). Take a random sample from each subgroup.



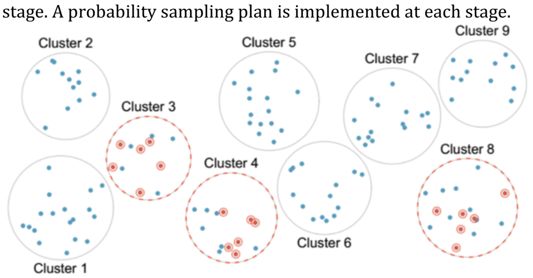
Cluster Sampling

Divide population into naturally occurring subgroups (clusters). Take random sample of clusters.



Multi-Stage Sampling

Sampling is carried out in stages with smaller and smaller sampling units at each stage. A probability sampling plan is implemented at each stage.



Types of Non-Probability Sampling

Volunteer / Self-Selected Sampling

Only those people interested in the study would respond.
→ Hard to talk about the degree of non-response bias.

Convenience / Haphazard Sampling

Using the most convenient group available or deciding on the sample on the spot.

Judgement Sampling

Using human judgement to choose "representative" units.

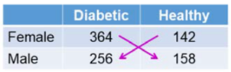
Quota Sampling

Interviewers are free to select anyone they like until they meet a fixed quota of units to interview for a fixed number for certain categories (eg. sex, age)

Risk and Risk Ratio

Risk (Disease | Male)	$\frac{Number\ of\ Males\ with\ Disease}{Number\ of\ Males}$
Risk Ratio of Disease between Males to Females	$\frac{Risk\ (Disease\ \vert\ Male)}{Risk\ (Disease\ \vert\ Female)}$

Odds and Odds Ratio

Odds (Disease) among Males	$\frac{Risk}{1-Risk}$
	$\frac{Number\ of\ Males\ with\ Disease}{Number\ of\ Males\ without\ Disease}$
Odds Ratio (Disease) between Females and Males (Females is the "first" group, Males is the "baseline" group)	$\frac{Odds\ (Disease)\ among\ Females}{Odds\ (Disease)\ among\ Males}$
	You can also use the **cross-product ratio**, by setting up a 2x2 contingency table: • **Event of interest** on the first column • The **first group (Females)** are on the first row.  $\frac{364\times158}{142\times256}\approx1.58$

	Diabetic	Healthy
Female	364	142
Male	256	158

Interpretation of Odds Ratio

OR = 1	No difference in disease risk in the two groups, and RR = 1
OR < 1	Higher risk in the first group, and RR > 1
OR > 1	Lower risk in the first group, and RR < 1

Types of Observational Studies

In observational studies, there is Association between the exposure and disease if $Risk\ Ratio\ \neq 1$ or $Odds\ Ratio\ \neq 1$.

	Can estimate Population Odds Ratio?	Can estimate Population Risk Ratio?
Cohort Study	Yes	Yes
• Starts from the exposure by investigating 100 smokers and 100 non-smokers. • Looks into the future and observes how many people from each group eventually have cancer.		
Case-Control Study	Yes	No!
• Starts from the outcome, or the disease statistics by investigating 100 people with cancer and 100 people without cancer. Looks into the background and observes how many people from each group were smokers.		
Cross Sectional Study	Yes	Yes
• Starts from a population sample by investigating 1000 random people. • Looks at their current disease and exposure status.		

Probability Rules

Complement Rule	$P(Complement) = 1 - P(A)$
"At Least One"	$P(At\ least\ one) = 1 - P(None)$
Addition Rule	$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$
Mutually Exclusive Events	$P(A\ or\ B) = P(A) + P(B)$
Multiplication Rule	$P(A\ and\ B) = P(A) \times P(B \mid A)$
Independent Events	$P(A\ and\ B) = P(A) \times P(B)$

Average Value of an Action

Assumption: Outcome A and B are mutually exclusive.

$$Value\ of\ Action = V(A) \times P(A) + V(B) \times P(B)$$

Hypothesis Testing

p-value	The sum of probabilities of events similar to and more extreme than our observation.
Null Hypothesis	Corresponds to the idea that an observation is due to chance.
Alternative Hypothesis	Corresponds to the idea that an observation is due to chance.

If the p-value is very small → our observed result of the experiment is likely not due to chance → reject the Null Hypothesis.

Testing for Rare Events

Base Rate	$P(Disease)$
Sensitivity	$P(Positive \mid Disease)$
Specificity	$P(Negative \mid No\ Disease)$