GEA1000N Cheatsheet

use at your own risk 🎀 @kangzhe

1 Getting Data

1.1 Sampling

Population. A group in which we have interest in drawing conclusions on in a study. Informally, it includes everybody.
Population Parameter. A numerical fact about a population.
Sample. A subset of the population, since it usually is not feasible to study a whole population.
Sampling Frame. List from which the sample was obtained.
Biases.

- **(i) Selection Bias.** Associated with the researcher's biased selection of units into the sample (eg. imperfect sampling frame or non-probability sampling).
- **(ii) Non-response Bias.** Associated with participant's non-disclosure or non-participation in the study.

Probability Sampling. A sampling scheme s.t. the selection process is done via a known randomized mechanism. Every unit in the sampling frame has a known non-zero probability of being selected.

⚠ WARNING probability sampling does not have to be equal chance; that is for SRS (see below).

Types of Probability Sampling.

- **(i) Simple Random Sampling (SRS).** Select randomly (with equal probability) without replacement.
  - ✓ Good representation of population.
  - × Time-consuming.
- **(ii) Systematic Sampling.** Select units from a list of size n by applying a selection interval k and a random starting point a from the first interval. The selected units will be $a, a+n/k, a+2n/k, \ldots$.
  - ✓ Simpler than SRS.
  - × If list is not random, may not be good representation.
- **(iii) Stratified Sampling.** **(1)** Divide the sampling frame into strata. Size of each stratum don't need to be the same, but strata must share similar characteristics. **(2)** Apply SRS to each stratum to generate the overall sample.
  - ✓ Good representation by strata.
  - × Complicated and time-consuming.
  - × Need info about sampling frame and strata.
- **(iv) Cluster Sampling.** **(1)** Divide the sampling frame into clusters. **(2)** Select a fixed number of clusters with SRS. **(3)** All units from selected clusters are included in the overall sample.
  - ✓ Less time-consuming.
  - × Need large sample size to reduce margin of error.

Non-Probability Sampling. Units not chosen by randomization. Examples are convenience sampling and volunteer sampling (self-selected).
Generalizability Conditions.

- ★ good sampling frame ⩾ population,
- ★ probability-based sampling to minimize selection bias,
- ★ large sample to reduce variability or random errors,
- ★ minimize the non-response rate.

1.2 Variables and Summary Statistics

Independent Variable (IV). Variables subject to manipulation either deliberately or spontaneously in a study.
Dependent Variable (DV). Variables hypothesized to change depending on how the IV is manipulated in the study.
Types of Variables.

- **(i) Categorical Variables.** Takes label values.
  - (a) **Ordinal.** There is some natural ordering and numbers can be used to represent the ordering.
  - (b) **Nominal.** There is no intrinsic ordering.
- **(ii) Numerical Variables.** Can do math to them.
  - (a) **Discrete.** There are gaps in the possible numbers taken on the variable.
    ⚠ WARNING Decimals can also be discrete (eg. shoe size).
  - (b) **Continuous.** Can take all possible numerical values in a given range.

Mean, Median, Mode. a.k.a. measures of central tendency.

$$\text{mean} = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

$$\text{median} = Q_2 = 50\text{th percentile}.$$

$$\text{mode} = \text{the data point that appears the most.}$$

Variance and Standard Deviation. Assume sample.

$$\underline{\text{sample}}\ \text{variance} = s_x^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}.$$

$$\underline{\text{sample}}\ \text{standard deviation} = s_x = \sqrt{\text{sample variance}}.$$

Coefficient of Variation. Used to quantify the degree of spread relative to the mean, given by $s_x/\bar{x}$.
Quartiles and IQR. Same as Canvas stats: Minimum, Lower Quartile Q_1 (25th percentile), Median Q_2 (50th percentile), Upper Quartile Q_3 (75th percentile), Maximum. Interquartile range (IQR) is $Q_3 - Q_1$.
Addition and Multiplication to Data.

- **Addition.** If you add some number c,
  - **(i)** mean $\bar{x} \to \bar{x} + c$,
  - **(ii)** variance, standard deviation, and IQR don't change.
- **Multiplication.** If you multiply by some constant c,
  - **(i)** mean $\bar{x} \to c\bar{x}$,
  - **(ii)** variance $s_x^2 \to c^2 s_x^2$,
  - **(iii)** standard deviation $s_x \to |c| s_x$,
  - **(iv)** IQR $(Q_3 - Q_1) \to |c|(Q_3 - Q_1)$.

1.3 Study Designs

Experimental Design. Manipulate one variable to find evidence for cause-and-effect relationships.
Control and Treatment Group. Control group must exist for us to test the effect of the IV.
Random Assignment. An impartial procedure that uses chance to allocate subjects into treatment and control groups.
Placebo. An inactive substance or intervention that looks the same and is given the same way as an active drug or treatment being tested.
Blinding.

- **(i)** Subjects don't know if they are in the treatment or control group.
- **(ii)** Assessors don't know if a data point is from the treatment or control group.

Do one of the above, we have single-blinding. Do both and we have double-blinding.
Observational Study. Observes individuals and measures the variables of interest, usually without any direct manipulation of the variables by the researchers.

2 Categorical Data Analysis

2.1 Rates

Rate. For some event E,

$$\text{rate}(E) = \frac{\text{number of times } E \text{ happened}}{\text{total number of data points}}.$$

Conditional Rate. For an event E given that F also happened, we denote that rate with $\text{rate}(E \mid F)$. A similar notation is used for probabilities $\text{P}(E \mid F)$.

2.2 Association

positive association	negative association
$\text{r}(A \mid B) > \text{r}(A \mid \neg B)$	$\text{r}(A \mid B) < \text{r}(A \mid \neg B)$
$\text{r}(B \mid A) > \text{r}(B \mid \neg A)$	$\text{r}(B \mid A) < \text{r}(B \mid \neg A)$
$\text{r}(\neg A \mid \neg B) > \text{r}(\neg A \mid B)$	$\text{r}(\neg A \mid \neg B) < \text{r}(\neg A \mid B)$
$\text{r}(\neg B \mid \neg A) > \text{r}(\neg B \mid A)$	$\text{r}(\neg B \mid \neg A) < \text{r}(\neg B \mid A)$

Symmetry Rule of Rates. The sign must be the same.

$$\text{r}(A \mid B) \gtreqless \text{r}(A \mid \neg B) \iff \text{r}(B \mid A) \gtreqless \text{r}(B \mid \neg A).$$

Basic Rule on Rates. The overall $\text{rate}(A)$ will always lie between $\text{rate}(A \mid B)$ and $\text{rate}(A \mid \neg B)$.

- **(i)** The closer $\text{rate}(B)$ is to 100%, the closer $\text{rate}(A)$ is to $\text{rate}(A \mid B)$.
- **(ii)** If $\text{rate}(B) = 0.5$, then
$$\text{rate}(A) = 0.5[\text{rate}(A \mid B) + \text{rate}(A \mid \neg B)].$$
- **(iii)** If $\text{rate}(A \mid B) = \text{rate}(A \mid \neg B)$, then
$$\text{rate}(A) = \text{rate}(A \mid B) = \text{rate}(A \mid \neg B).$$

2.3 Simpson's Paradox

Simpson's Paradox. A phenomenon in which a trend appears in more than half of the groups of data but disappears or reverses when the groups are combined. Here, "disappears" means the two variables in question (say A and B) are no longer associated, that is $\text{rate}(A \mid B) = \text{rate}(A \mid \neg B)$.
Confounder. A third variable that is associated with both the IV and DV. We do not care if it is positively or negatively associated, as long as the variable is associated in some way.
Dealing with Confounders (Slicing). Slicing is where the data is segregated by the confounding variable. Then consider each group separately.

3 Dealing with Numerical Data

3.1 Univariate EDA

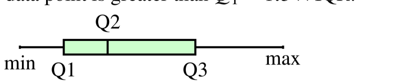
Distribution. An orientation of data points broken down by their observed number of frequency or occurrence.
Histogram. A graphical representation that organizes data points into ranges or bins.
Histogram vs Bar Graph. Histogram is only for ranged data, bar graph can do anything you want.
Describing Univariate.

- **(i) Shape.** Talk about the peaks and skewness (left-skewed means "tail" on the left).
- **(ii) Center.** Talk about modes, mode median mean.
  - (a) For left-skewed distribution, we usually (not always) have mean < median < mode.
  - (b) For right-skewed distribution, we usually (not always) have mean > median > mode.
- **(iii) Spread.** Talk about IQR and standard deviation.

Outliers. An observation that falls well above/below the overall bulk of data.

- Value of data point is greater than $Q_3 + 1.5 \times \text{IQR}$, or
- Value of data point is greater than $Q_1 - 1.5 \times \text{IQR}$.

Boxplots.



3.2 Bivariate EDA

Bivariate Relationships.

- **(i) Deterministic.** The value of one variable can be determined exactly if we know the value of the other variable. Example: relationship between Fahrenheit and degrees Celsius.
- **(ii) Statistical.** The value of one variable can tell us the average value of another variable.

Scatter Plot. IV is plotted on x-axis and DV on y-axis.
Describing Bivariate. **(i) Strength:** weak/strong/none, **(ii) Direction:** positive/negative/neither, **(iii) Form:** linear/non-linear.

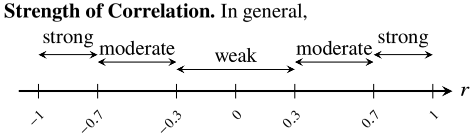
$$\underbrace{\text{weak}}_{\text{strength}}\ \underbrace{\text{positive}}_{\text{direction}}\ \underbrace{\text{linear}}_{\text{form}}\ \text{correlation.}$$

3.3 Correlation Coefficient r

Correlation Coefficient. A measure of the linear association between two numerical variables, denoted by r.

$$\text{association is} \begin{cases} \text{positive} & \text{if } r > 0, \\ \text{negative} & \text{if } r < 0, \\ \text{no linear association} & \text{if } r = 0. \end{cases}$$

Perfect Linear Association. If $r = 1$, we have perfect positive association; if $r = -1$, perfect negative association.
Strength of Correlation. In general,



Algorithm for Calculating r. Not required to know, but might be useful for intuition.

Step 1. Find mean and standard deviation of x and y.

Step 2. Convert each value of x and y into standard units:

$$x_{\text{std}} = \frac{x - \bar{x}}{s_x} \quad \text{and} \quad y_{\text{std}} = \frac{y - \bar{y}}{s_y}.$$

Step 3. Compute the product $x_{\text{std}} y_{\text{std}}$ for each data point.

Step 4. Use the formula

$$r = \frac{1}{n-1}\left(\sum_{\text{all } x_{\text{std}}, y_{\text{std}}} x_{\text{std}} y_{\text{std}}\right).$$

Transforming Data. The correlation coefficient is not affected when:

- swapping the x and y variables,
- adding a number to all values of a variable,
- multiplying a positive number to all values of a variable.

WARNING For multiplication, if you multiply all values of a variable by a negative number, the sign of r will flip (positive to negative and vice versa), but the value does not change.

Limitations of r.

× Association is not causation.

× r does not tell us anything about non-linear associations. To deal with this, the data needs to be linearized accordingly to a linear relationship.

× Outliers can affect r significantly.

× So can confounders.

Ecological Correlation. Represents relationships observed at the aggregate level, considering the characteristics of groups rather than individuals.

Fallacy	Using	To conclude
Ecological	Ecological correlation (aggregate level)	Individual level correlation
Atomistic	Individual level correlation	Ecological correlation (aggregate level)

3.4 Linear Regression

Linear Regression. If we believe two variables X and Y are linearly associated, we model the relation by the equation of a straight line $Y = mX + b$.

Method of Least Squares. To find the line of best fit, we minimize the sum of errors by minimizing the distance between all data points to the line.

$$\text{task: minimize } e^2 = e_1^2 + e_2^2 + \cdots + e_n^2.$$

Every error is squared so we remove negative signs.

Swapping x and y. You will get different regression lines.

Correlation Coefficient and Regression. For a model with the equation $Y = mX + b$,

$$m = \frac{s_Y}{s_X} r.$$

Validity Range of Model. We can use the model equation to only predict values within the given range. That is to say, extrapolations are not valid.

4 Statistical Inference

4.1 Probability

Probability Experiment. A procedure that is repeatable and allows for the exact listing of all possible outcomes.

Sample Space. The collection of all possible outcomes of a probability experiment.

Probability (garbage definition). For a probability experiment with an associated sample space, the probability of an event of the sample space is the total probability that the outcome of the experiment is an element of the event.

Rules of Probabilities.

(i) The probability of each event E, denoted by $\mathrm{P}(E)$ is a number between 0 and 1 (inclusive).

(ii) If we denote the entire sample space as S then the probability of S is $\mathrm{P}(S) = 1$.

(iii) If E and F are mutually exclusive events, then $\mathrm{P}(E \cup F) = \mathrm{P}(E) + \mathrm{P}(F)$.

Uniform Probability. The way of assigning probabilities to outcomes such that equal probability is assigned to every outcome in the finite sample space. This, if the sample space contains N different outcomes, then the probability assigned to each outcome is $1/N$.

4.2 Conditional Probability and Independence

Conditional Probability. The probability where E occurs given that F also has occurred is written as $\mathrm{P}(E \mid F)$.

Intersection. $E \cap F$ means E and F has happened.

Equation for Conditional Probability.

$$\mathrm{P}(E \mid F) = \frac{\mathrm{P}(E \cap F)}{\mathrm{P}(F)}.$$

Comparing Rates and Probabilities.

Random Sampling	Probability Experiment
Sampling Frame	Sample Space
A subgroup E of the sampling frame	An event E of the sample space
rate(A)	$\mathrm{P}(A)$

Independence. Two events E and F are independent if $\mathrm{P}(E) = \mathrm{P}(E \mid F)$.

Conditionally Independent. Two events E and F are conditionally independent given an event G with $\mathrm{P}(G) > 0$ if $\mathrm{P}(E \cap F \mid G) = \mathrm{P}(E \mid G) \times \mathrm{P}(F \mid G)$.

Law of Total Probability. If E, F, and G are events from the same sample space S s.t. **(i)** E and F are mutually exclusive, and **(ii)** $E \cup F = S$, then

$$\mathrm{P}(G) = \mathrm{P}(G \mid E) \times \mathrm{P}(E) + \mathrm{P}(G \mid F) \times \mathrm{P}(F).$$

4.3 Fallacies

A Useful Inequality from 4.3.2. (Conjunction Fallacy). For any events A and B, we must have

$$\mathrm{P}(A \cap B) \leqslant \mathrm{P}(A) \quad \text{and} \quad \mathrm{P}(A \cap B) \leqslant \mathrm{P}(B).$$

Prosecutor's Fallacy. In general, $\mathrm{P}(A \mid B) \neq \mathrm{P}(B \mid A)$.

Base Rate Fallacy. A decision making error in which information about the rate of occurrence of some trait in a population, called the base rate information, is ignored or not given appropriate weight.

Sensitivity and Specificity. Sensitivity is the true positive rate, that is $\mathrm{P}(+ \mid \text{infected})$. **Specificity** is the true negative rate, that is $\mathrm{P}(- \mid \text{not infected})$. We cannot deduce $\mathrm{P}(\text{infected} \mid +)$ from the above.

4.4 Statistical Inference

Random Variable. A numerical variable with probabilities assigned to each of the possible numerical values taken by the numerical variable. Random variables are either discrete or continuous.

Density Curve. Given the density curve of a continuous random variable, we can find the probability that the random variable takes a certain range of values with the area under the curve within that range.

Statistical Inference. The use of samples to draw inferences or conclusions about the population in question.

Accounting for Inaccuracies.

$$\text{sample statistic} = \frac{\text{population}}{\text{parameter}} + \text{bias} + \text{random error}.$$

Fundamental Rule for using Data for Inference. Available data can be used to make inferences about a much larger group if the data can be considered to be representative with regards to the question of interest.

Confidence Interval. A confidence interval (CI) is a range of values that is likely to contain a population parameter of a certain degree of confidence. This degree of confidence is called the confidence level and is usually expressed as a percentage.

WARNING It is a common mistake to say that there is a 95% chance that the population proportion lies in the confidence interval. This statement is incorrect because:

- population proportion p is "fixed", although unknown.
- for any particular sample, the confidence interval constructed only depends on the sample proportion and the value of z^* corresponding to a chosen confidence level. Thus, the confidence interval is also "fixed" and there is no probabilistic element in it.

Either the population is in the interval or it is not, but we cannot associate any kind of chance to it.

Proportion. The proportion p (denoted as p^* for samples) is the fraction of that event's frequency in the entire dataset.

z-value (informal). The z-value is the numerical value of the standard normal variable. For example, $\mathrm{P}(Z = 1.96) = 0.95$ the z-value of a 95% confidence interval has an associated z-value of 1.96.

Confidence Interval (Proportion). Given p^* is the sample proportion, z^* is the z-value from the standard normal distribution, and n is the sample size,

$$\mathrm{CI} = p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}.$$

Example for calculating CI (4.4.4). Given that the proportion of 5-room flats is $508/2000 = 0.254$, we find that using the formula, the confidence interval of 95% is

$$0.254 \pm 1.96 \times \sqrt{\frac{(0.254)(1-0.254)}{2000}} = 0.254 \pm \underbrace{0.0191}_{\substack{\text{margin}\\ \text{of error}}}.$$

Confidence Interval (Sample Mean). Given that $\bar{x}$ is the sample mean, t^* is the t-value from the t-distribution, s is the sample standard deviation, and n is the sample size,

$$\mathrm{CI} = \bar{x} \pm t^* \times \frac{s}{\sqrt{n}}.$$

Effects on Confidence Intervals. With a smaller sample size, the confidence interval is wider. With a larger sample size, the confidence interval is narrower.

4.5 Hypothesis Testing

Steps in Hypothesis Testing.

(i) Identify the question, state the null hypothesis H_0 and the alternative hypothesis H_1.

(ii) Set a significance level α. Usually $\alpha = 5\%$.

(iii) Find the relevant sample statistic (usually mean).

(iv) Calculate the p-value.

(v) Make a conclusion of the hypothesis test.

p-value. The p-value is the probability of obtaining a result as extreme or more extreme than our observation in the direction of the alternative hypothesis, assuming the null hypothesis is true.

Conclusions of Hypothesis Testing

- If p-value < significance level, then we say that we have sufficient evidence to reject the null hypothesis in favour of the alternative hypothesis.
- If p-value $\geqslant$ significance level, we say that we have insufficient evidence to reject the null hypothesis. The hypothesis test is inconclusive.

WARNING this does not mean that we accept the null hypothesis!

Hypothesis Tests. We can conduct hypothesis tests on:

- population proportion,
- population mean, using the t test or z test,
- association, using a chi-squared (χ^2) test.

Hypothesis Test Format. We usually have to test

$$H_0: \text{population parameter} = \text{null value}$$

against some inequality $>$, $<$, or $\neq$ on the population parameter. For example:

$$H_1: \text{population parameter} > \text{null value}.$$

Let the significance level $\alpha = 0.05 = 5\%$. After calculating the p-value, we compare this against α. We make the appropriate conclusions according to the notes above.

5 /THANKYOUVICTOR

Wishing you

$\text{rate}(\text{get an A+} \mid \text{you take this paper}) = 100\%$!

Good luck for your finals!