STAT231 Final Exam Review

LaTeXer: W. Kong

1 PPDAC

PPDAC = Problem / **P**lan / **D**ata / **A**nalysis / **C**onclusion (See the final page for a summary)

Definition 1.1. The **target population** is the set of animals, people or things about which you wish to draw conclusions. A **unit** is a singleton of the target population.

Definition 1.2. The **sample population** is a specified subset of the target population. A **sample** is a singleton of the sample population and a unit of the study population.

Definition 1.3. A **variate** is a characteristic of a single unit in a target population and is usually one of the following:

1. **Response variates** - interest in the study
2. **Explanatory variate** - why responses vary from unit to unit
   (a) Known - variates that are know to cause the responses
      i. Focal - known variates that divide the target population into subsets
   (b) Unknown - variates that cannot be explained in the that cause responses

Definition 1.4. An **attribute/parameter(T.P.)/statistic(Sample)** is a characteristic of a population which is usually denoted by a function of the response variate. It can have two other names, depending on the population studied.

Definition 1.5. The **aspect** is the goal of the study and is generally one of the following: descriptive, comparative, causative, and predictive.

Note 1. $T.P. \supset S.P. \supset Sample$

Definition 1.6. Let $a(x)$ be defined as an attribute as a function of some population or sample x. We define the **study error** as

$$a(T.P.) - a(S.P.).$$

Definition 1.7. Similar to above, we define the **sample error** as

$$a(S.P.) - a(sample).$$

2 Measurement Analysis

The goal of measures is to explain how far our data is spread out and the relationship of data points.

2.1 Measurements of Spread

Definition 2.1. Coefficient of Variation (CV)

This measure provides a unit-less measurement of spread:$CV = \frac{s}{\bar{x}} \times 100\%$

2.2 Measurements of Association

1. **Covariance**: In theory (a population), the covariance is defined as $\text{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y))$ but in practice (in samples) it is defined as $s_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$.Note that $\text{Cov}(X,Y), s_{XY} \in \mathbb{R}$ and both give us an idea of the direction of the relationship but not the magnitude.

2. **Correlation***:* In theory (a population), the correlation is defined as $\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$ but in practice (in samples) it is defined as $r_{XY} = \frac{s_{XY}}{s_X s_Y}$. Note that $-1 \leq \rho_{XY}, r_{XY} \leq 1$ and both give us an idea of the direction of the relationship AND the magnitude.

   (a) An interpretation of the values is as follows: $|r_{XY}| \approx 1 \implies$ strong relationship, $|r_{XY}| = 1 \implies$ perfectly linear relationship, $|r_{XY}| > 1 \implies$ positive relationship, $|r_{XY}| < 1 \implies$ negative relationship, $|r_{XY}| \approx 0 \implies$ weak relationship

3. **Relative-risk**: From $STAT230$, this the probability of something happening under a condition relative to this same thing happening if the condition is note met. Formally, for two events A and B, it is defined as $RR = \frac{P(A|B)}{P(A|\bar{B})}$. An interesting property is that if $RR = 1$ then $A \perp B$ and vice versa.

4. **Slope**: This will be covered later on.

3 Statistical Models

Recall that the goal of statistics is to guess the value of a population parameter on the basis of a (or more) sample statistic.

3.1 Types of Models

Goal of **statistical models**: explain the relationship between a parameter and a response variate.

The following are the different types of statistical models that we will be examining :

1. **Discrete** (Binary) **Model -** either the population data is within parameters or it is not.

2. **Response Model -** these model the response and *at most* use the explanatory variate implicitly as a focal explanatory variate.

3. **Regression Model -** these create a function that relates the response and the explanatory variate (attribute or parameter); note here that we assume $Y_i = Y_i|X$.

4 Estimates and Estimators

Here, we only review the main ideas of estimates and estimators.

4.1 Maximum Likelihood Estimation (MLE) Algorithm

1. Define $L = f(y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f(y_i)$ where we call L a **likelihood function**. Simplify if possible. Note that $f(y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f(y_i)$ because we are assuming random sampling, implying that $y_i \perp y_j$, $\forall i \neq j$.
2. Define $l = \ln(L)$. Simplify l using logarithmic laws.
3. Find $\frac{\partial l}{\partial \theta_1}, \frac{\partial l}{\partial \theta_2}, ..., \frac{\partial l}{\partial \theta_n}$, set each of the partials to zero, and solve for each θ_i, $i = 1, ..., n$. The solved $\theta_i's$ are called the **estimates** of f and we add a hat, $\hat{\theta}_i$, to indicate this.

4.2 Estimators

$\hat{\theta}$ is the realization (from a sample) of a distribution of estimates. The distribution is called an **estimator** and is denoted by $\tilde{\theta}$.

4.3 Biases in Statistics

Definition 4.1. We say that for a given estimator, $\tilde{\theta}$, of an estimate for a model is **unbiased** if $E(\tilde{\theta}) = \theta$ holds.Otherwise, we say that our estimator is **biased**.

5 Distribution Theory

We introduce the following new distributions.

- If $X \sim N(0, 1)$ then $X^2 \sim \chi_1^2$ which we call a **Chi-squared** (pronounced "Kai-Squared") **distribution** on one degree of freedom
- Let $X \sim \chi_m^2$ and $Y \sim \chi_n^2$. Then $X + Y \sim \chi_{n+m}^2$ which is a Chi-squared on $n + m$ degrees of freedom
- Let $N \sim N(0, 1)$, $X \sim \chi_v^2$, $X \perp N$. Then $\frac{N}{\sqrt{\frac{X}{v}}} \sim t_v$ which we call a **student's t-distribution** on v degrees of freedom

Properties of the Student's t-Distribution

- This distribution is symmetric
- For distribution $T \sim t_v$, when $v > 30$, the student's t is almost identical to the normal distribution with mean 0 and variance 1
- For $v \ll 30$, T is very close to a uniform distribution with thick tails and very even, unpronounced center

5.1 Least Squares Method

There are two ways to use this method. First, for a given model Y and parameter θ, suppose that we get a best fit $\hat{y}$ and define $\hat{\epsilon}_i = |\hat{y} - y_i|$. The least squares approach is through any of the two

1. (Algebraic) Define $W = \sum_{i=1}^{n} \hat{\epsilon}_i^2$. Calculate and minimize $\frac{\partial W}{\partial \theta}$ to determine θ.
2. (Geometric) Define $W = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \hat{\epsilon}^t \hat{\epsilon}$. Note that $W \perp \text{span}\{\overrightarrow{1}, \overrightarrow{x}\}$ and so $\hat{\epsilon}^t \overrightarrow{1} = 0$ and $\hat{\epsilon}^t \overrightarrow{x} = 0$. Use these equations to determine θ.

6 Intervals

Here, we deviate from the order of lectures and focus on the various types of constructed intervals. However, in this section, I will only provide the formulas and not the motivation.

Name	Formula	Properties
Confidence Intervals	$EST \pm cSE = \hat{\theta} \pm c\sqrt{Var(\tilde{\theta})}$	If $Var(\tilde{\theta})$ is known, then $C \sim N(0,1)$ and if it is unknown, we replace $Var(\tilde{\theta})$ with $\hat{Var}(\tilde{\theta})$ and $C \sim t_{n-q}$. When $\alpha = 5\%$, which affects our value of c, we are constructing a 95% confidence interval.
Predicting Intervals	$EST \pm cSE = f(\hat{\theta}) \pm \sqrt{Var(Y_p)}$	Same as above except note that Y_p is different from a standard model $Y_i = f(\theta) + \epsilon_i$ in that the first component is random (i.e. $Y_p = f(\tilde{\theta}) + \epsilon_p$). When $\alpha = 5\%$, which affects our value of c, we are constructing a 95% confidence interval.
Likelyhood Intervals	Solution of $R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$ where $L(\theta, y_i) = \prod_{i=1}^{n} f(y_i, \theta)$	When computing the solution of $R(\theta) \approx 0.1$, this will give the 95% likelyhood interval for θ. This interval is particularly useful for models that are not necessarily normal

7 Hypothesis Testing

While our confidence interval does not tell us in a yes or no way whether or not a statistical estimate is true, a hypothesis test does. Here are the steps:

1. State the hypothesis, $H_0 : \theta = \theta_0$(this is only an example), called the null hypothesis (H_1 is called the alternative hypothesis and states a statement contrary to the null hypothesis).
2. Calculate the discrepancy (also called the test statistic), denoted by $d = \frac{\hat{\theta} - \theta_0}{\sqrt{Var(\tilde{\theta})}} = \frac{estimate - H_0\, value}{SE}$ assuming that $\tilde{\theta}$ is unbiased and the realization of d, denoted by D, is $N(0,1)$ if $Var(\tilde{\theta})$ is known and t_{n-q} otherwise. Note that d is the number of standard deviations θ_0 is from $\hat{\theta}$.
3. Calculate a $p-$value given by $p = 2P(D > |d|)$. It is also the probability that one sees a value worse than $\hat{\theta}$, given that the null hypothesis is true. The greater the $p-$value, the more evidence against the model in order to reject.

4. Reject or not reject (note that we do not "accept" the model)

The following the table that subjectively describes interpretations for $p-$values:

P value	Interpretation
p-value<1%	A ton of evidence against H_0
1%≤p-value<5%	A lot of evidence against H_0
5%≤p-value≤10%	Some evidence against H_0
p-value>10%	There is virtually no evidence against H_0

Note that one model for which $D \sim N(0,1)$ is $Y_i = \epsilon_i$ where $\epsilon_i \sim \text{Bin}(1, \Pi)$ since $\sqrt{Var(\tilde{\theta})} = \sqrt{\frac{\hat{\Pi}_0(1-\hat{\Pi}_0)}{n}}$ by our null hypothesis and central limit theorem.

8 Comparative Models

The goal of a comparative model is to compare the mean of two groups and determine if there is a causation relationship between one and the other.

Definition 8.1. If x causes y and there is some variate z that is common between the two, then we say z is a **confounding variable** because it gives the illusion that z causes y. It is also sometimes called a **lurking variable**.

There are two main models that help determine if one variate causes another and they are the following.

Experimental Study

1. For every unit in the T.P. set the F.E.V. (focal explanatory variate) to level 1
2. We measure the attribute of interest
3. Repeat 1 and 2 but with set the F.E.V. to level 2
4. Only the F.E.V. changes and every other explanatory variate is fixed
5. If the attribute changes between steps 1 and 4, then causation occurs

Problems?

- We cannot sample the whole T.P.
- It is not possible to keep all explanatory variates fixed
- The attributes change (on average)

Observational Study

1. First, observe an association between x and y in many places, settings, types of studies, etc.
2. There must be a reason for why x causes y (either scientifically or logically)
3. There must be a consistent dose relationship
4. The association has to hold when other possible variates are held fixed

9 Experimental Design

There are three main tools that are used by statisticians to improve experimental design.

1. Replication
   (a) Simply put, we increase the sample size
      i. This is to decrease the variance of confidence intervals, which improves accuracy
2. Randomization
   (a) We select units in a random matter (i.e. if there are 2+ groups. we try to randomly assign units into the groups)
      i. This is to reduce bias and create a more representative sample
      ii. It allows us to assume independence between $Y_i's$
      iii. It reduces the chance of confounding variates by unknown explanatory variates
3. Pairing
   (a) In an experimental study, we call it blocking and in an observational study, we call it matching
   (b) The actual process is just matching units by their explanatory variates and matched units are called twins
      i. For example in a group of 500 twins, grouped by gender and ages, used to test a vaccine, one of the twins in each group will take the vaccine and another will take a placebo
      ii. We do this in order to reduce the chance of confounding due to known explanatory variates
   (c) Pairing also allows us to perform subtraction between twins to compare certain attributes of the population
      i. Note that taking differences does not change the variability of the difference distribution

10 Model Assessment

We usually want the following four assumptions to be true, when constructing a model $Y_i = f(\theta) + \epsilon_i$ to fit a sample. We also use certain statistical tools to measure how well our model fits these conditions. Note that these tools/tests require subjective observation.

- $\epsilon_i \sim N(0, \sigma^2)$
  - Why? Because Y_i is not normal if ϵ_i is not normal. However, $\tilde{\theta}$ is still likely to be normal by CLT.
  - Tests:
    * Histogram of residuals $\hat{\epsilon_i} = |\hat{y} - y_i|$ (should be bell-shaped)
    * QQ Plot, which is the plot of theoretical quartiles versus sample quartiles (should be linear with intercept ˜0)
      · Usually a little variability at the tails of the line is okay
- $E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2, \epsilon_i's$ are independent
  - Why? All of our models, tests and estimates depend on this.

- Tests:
  - Scatter plot of residuals (y-axis) versus fitted values (x-axis)
    - We hope that it is centered on 0, has no visible pattern and that the data is bounded by two parallel lines (constant variance)
    - If there is not a constant variance, such as a funnel (funnel effect), we usually transform the fitted values (e.g. $y \to \ln y$)
    - If the plot seems periodic, we will need a new model (STAT 371/372)
  - Scatter plot of fitted values (y-axis) versus explanatory variates (x-axis)
    - This is used mainly in regression models
    - We hope to see the same conditions in the previous scatter plot

11 Chi-Squared Test

The purpose of a Chi-squared test is to determine if there is an association between two random variables X,Y, given that they both contain only counting data. The following are the steps

1. State the null hypothesis as H_0 : X and Y are not associated.
2. If there are m possible observations for X and n possible observations for Y, then define
$$d = \sum_{j=1}^{n}\sum_{i=1}^{m} \frac{(expected - observed)^2}{expected} = \sum_{j=1}^{n}\sum_{i=1}^{m} \frac{(e_{ij} - o_{ij})^2}{e_{ij}}$$
where $e_{ij} = P(X = x_i) \cdot P(Y = y_j)$, $o_{ij} = P(X = x_i, Y = y_j)$, for $i = 1, ..., m$ and $j = 1, ..., n$.
3. Assume that $D \sim \chi^2_{(m-1)(n-1)}$.
4. Calculate the p-value which in this case is $Pr(D > d)$ since $d \geq 0$, which means we are conducting a one-tailed hypothesis.
5. Interpret it as always (see the table in Section 8)

Data Quality (Analysis)

There are 3 factors that we look at:

1. Outliers
2. Missing Data Points
3. Measurement Issues

Characteristic of a Data Set (Analysis)

Outliers could be found in any data set but these 3 always are:

1. **Shape**
   (a) Skewness and Kurtosis, Bell-shaped, Skewed left (negative), Skewed right (positive), Uniform
2. **Center** (location)
   (a) The "middle" of our data
      i. Mode: statistic that asks which value occurs the most frequently
      ii. Median ($Q2$): the middle data value
      iii. Mean: the sample mean is $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$
   (b) **Robustness**: The median is less affected by outliers and is thus *robust.*
3. **Spread** (variability)
   (a) Range**:** By definition, this is $x_{(n)} - x_{(1)}$
   (b) IQR (Interquartile range): The middle half of your data

Problem

The problem step's job is to clearly define the

1. Goal or Aspect of the study
2. Target Population and Units
3. Unit's Variates
4. Attributes and Parameters

Plan

1. Define the **Study Protocol**
2. Define the **Sampling Protocol**
3. Define the **Sample**
4. Define the **measurement system**

Data Types:

- **Discrete Data***:* Simply put, there are "holes" between the numbers
- **Continuous** (CTS) **Data***:* We **assume** that there are no "holes"
- **Nominal Data***:* No order in the data
- **Ordinal Data***:* There is some order in the data
- **Binary Data***:* e.g. Success/failure, true/false, yes/no
- **Counting Data***:* Used for counting the number of events

Conclusion

In the conclusion, there are only two aspects of the study that you need to be concerned about:
Did you answer your problem
Talk about **limitations** (i.e. study errors, samples errors)