MAT121: Final Exam Review

Cheng Peng

1 Descriptive Statistics

1.1 Basic concepts

- Population and sample
- random sample
- parameter and statistic
- categorical and numerical data

1.2 Summarizing Data Using Tables and Chart

- **Categorical Data**
  1. Frequency tables
  2. Bar chart and pie chart
- **Numerical data**
  1. Frequency tables - need to group data to create class labels (small data windows should have equal width).
  2. Histogram (shapes - symmetric or skewed?)

1.3 Numerical Summary of a Numerical Data

You should be able to calculate the following measures if you are given a tiny data set.

- **Central tendency**
  1. mean
  2. median
  3. model
- **Variation**
  1. variance (sample and population)
  2. standard deviation (sample and population)
- **Location**
  1. z-score transformation
  2. percentiles
  3. Five number summary
  4. Box-plot

2 Probability Distributions

- Normal distribution and density curves
- z-score transformation to convert general normal distribution to the standard normal distribution.
- finding the area of the region under the density curve define by two given values (including the infinity)
- finding the value for a given area and a given value.
- Use of normal, t, and chi-square table.

3 Sampling Distributions

3.1 Sampling distribution of sample means $\bar{x}$

Let $\{x_1, x_2, \ldots, x_n\}$ be a random sample taken from a population with mean μ and standard deviation σ.

Depending on the amount of information available in the data and the assumptions of the population, there are three cases listed in the following

- **Case 1**: If $n > 30$, then $\bar{x} \to N(\mu, \sigma/\sqrt{n})$. If σ is unknown, replace it with the sample standard deviation.
- **Case 2**: If $n \leq 30$, the population is normal, and σ is known, then $\bar{x} \to N(\mu, \sigma/\sqrt{n})$
- **Case 3**: If $n \leq 30$, the population is normal, and σ is unknown, then $(\bar{x} - \mu)/(s/\sqrt{n}) \to t_{n-1}$

3.2 Sampling distribution of sample proportions $\hat{y}$

- if $n\hat{p} > 5$ and $n(1 - \hat{p}) > 5$, then $\hat{p} \to N(p, \sqrt{p(1-p)/n})$.

4 Confidence Intervals

4.1 Confidence Intervals of Single Mean

- The general formulation of interval at 95% confidence level.

$$\bar{x} \pm CV \times \frac{s}{\sqrt{n}}$$

where $E = s/\sqrt{n}$ is the margin of error. E is equal to half of the confidence interval.

1. **Case 1**: using the standard normal table to find $CV = Z_{\alpha/2}$.
2. **Case 2**: using the standard normal table to find $CV = Z_{\alpha/2}$. s should be replaced by the known σ.
3. **Case 3**: using the t table to find $CV = t_{n-1,\alpha/2}$.

- Things need to know about the confidence interval of a `population mean` μ
    1. Need to know how to choose an appropriate case based on the given information to construct the confidence interval.
    2. Need to know how to use normal and t tables to find the critical value based on the given confidence level.
    3. The margin of error is half of the width of the confidence interval. For example, if a confidence interval is $(100, 150)$, the margin of error $E = (150 - 100)/2 = 25$.
    4. The sample mean $\bar{x}$ is always inside the confidence interval (actually at the center of the confidence interval).

5. Fix the fixed standard error and confidence level, as sample size increases, the margin of error decreases.
6. For a given sample (i.e., fixed sample size and standard deviation), as the confidence level increases, the margin of error also increases.
7. Finally need to know how to interpret the confidence interval.

4.2 Confidence Intervals of A Single population Proportion

- The general formulation of the confidence interval of a population proportion.

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

if $n\hat{p} > 5$ and $n(1-\hat{p}) > 5$.

- Things need to know about the confidence interval of a population proportion p.
  1. Check the conditions for constructing the confidence interval.
  2. Should always turn percentages to decimals in all calculations.
  3. The sample proportion is always inside the confidence interval.
  4. Both confidence limits are non-negative.

5 Hypothesis Tests of A Single Population Mean and Proportion

5.1 Testing A Single Population Mean μ

- Things you need to know about testing hypothesis of population mean (μ)
  1. Identifying the claim correctly. 6 possible claims. Pay attention to key words such as `at least` ($\geq$), `at most` ($\leq$), `no more than` ==> ($\leq$), `is` ($=$), `differ` ($\neq$), etc.
  2. Setting up Ho and Ha. "$=$" sign must be included in Ho and "$\neq$" must be included in Ha.
  3. Evaluating the test statistics: $TS = (\bar{x} - \mu_0)/(s/\sqrt{n})$, where μ_0 is the claimed population mean.
  4. Types of test: Ha "$\neq$" ==> `two-tailed test`; "$>$" ==> `right-tailed test`; "$<$" ==> `left-tailed test`.
  5. Knowing how to use the critical value method: (1) need to know when using a normal table; (2) need to know when use t- table; and (3) know the location of the rejection region.
  6. Knowing how to calculate the p-value: (1) for a left-tailed (right-tailed) test, the p-value is equal to the left (right) tail area; (2) for a two-tailed test, the p-value is equal to 2 $\times$ the smaller test area.
  7. Decision rules: (1) For the critical value method, if TS is in the RR, reject Ho, otherwise, fail to reject Ho; (2) for the p-value method, if $p < \alpha$, reject Ho, otherwise, fail to reject Ho.
  8. Type I and Type II Errors: (1) If Ho is true, but was rejected ==> Type I error;

(2) If Ho is false, but it was not rejected ==> Type II error.

5.2 Testing A Single Population Proportion

- Things need to know about testing proportion: in addition to the above-mentioned points for testing population mean (μ), the following are also expected to know.
  1. all percentages MUST be converted to decimals in all calculations.
  2. key word `majority` means $p > 50\%$.

6 Two Sample Tests of Population Means (μ_1 and μ_2)

6.1 Paired Sample (before-after samples)

Two measurements were taken from each subject in the study (before and after samples).

- Calculate the difference between the paired observations to obtain a single sample.
- Use the standard one-sample test procedure to test the difference.
- Ho is set up based on μ_d:

6.2 Independent Samples - Large Samples or Normal Population with Known Variances

- Identifying claim about the difference between two populations means $\mu_1 - \mu_2$
- Setting up Ho and Ha
- Evaluating the test statistic

$$TS = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Finding the critical value and p-value using the normal table.

6.3 Independent Samples - Small Samples with Unknown Population Variances

- **Assumptions**: (1) both populations are normal; (2) Both population variances (standard deviations) are unknown; (3) Both population variances are equal
- Calculating the pooled sample variances

$$s_{pool}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- Evaluating test statistic

$$TS = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_{pool}^2(1/n_1 + 1/n_2)}}$$

- Find either the critical value(s) or p-value from the t-table with degrees of freedom $(n_1 + n_2 - 2)$.

7 Correlation and Coefficient

- The the structure of the data set

X	Y
x_1	y_1
x_2	y_2
x_3	y_3
$\vdots$	$\vdots$
x_n	y_n

- Scatter plots for the types of association: need to tell linear and non-linear association between x and y.
- **Correlation coefficient of two numerical variables** - interpretation of the correlation coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- **Least square regression line**: $y = \beta_0 + \beta_1 x$, where β_0 and β_1 are called intercept and slope respectively (both are unknown). The objective is to estimate β_0 and β_1, denoted by b_0 and b_1 from the given data set.
- **Fitted regression line** $\hat{y} = b_0 + b_1 X$
  1. **Example** ice cream sale$(\hat{y}) = -169.75 + 4.275 \times$ temperature(x)
  2. Interpretation of the slope (b_1): change of y as x increased by **one** unit. From the above ice cream sale example, slope $b_1 = 4.275$ can be interpret as `when the temperature increases 1 degree, the ice cream sale will increase 4.275 dollors`.
  3. **Coefficient of determination**. Assume the coefficient of determination $R^2 = 0.694$. The interpretation is that the regression line explains about 69.4% of the variation of ice cream sales.
  4. Testing the significance of the slope: Ho $\beta_1 = 0$ vs $\beta_1 \neq 0$. If you are given the value of the test statistic and the p-value, you should be able to make a statistical decision and draw a conclusion.
  5. **Prediction**: if you are given a fitted regression line, you should be able to predict y using the prediction equation. For example, given ice cream sale$(\hat{y}) = -169.75 + 4.275 \times$ temperature(x). If we want to predict the ice cream sale when the temperature is 80 degrees, we simply plug in the 80 degrees to the above equation to get the predicted ice cream sale.
  6. The interpretation of the confidence interval of the slope: For example, given 95% confidence interval for the slope associated with the temperature is (1.44, 7.11). The interpretation is that we are 95% confident that, as the temperature increases by 1 degree, the change ice cream sale will be between 1.44 and 7.11 dollars.
  7. We can also similarly interpret the predicted mean interval of the ice cream sale at a given temperature.

8 Independence and Goodness-of-fit

- For a given distribution of a categorical variable from an observed table, we should be able to find the expected table. The chi-squared test is used to test whether the data supports the claimed distribution.
  1. Given information: (1) the distribution of the categorical variable; (2) an observed distribution table based on the sample.
  2. Need to find the expected table.