GEOG3023: Statistics and Geographic Data

Study Guide for Midterm

Kylen Solvik

The following is a study guide for the midterm. It only includes the topics that will be potentially covered in the exam and what is important that I expect you to understand. It follows the order of the lectures. It doesn't intend to be in details and complete, please refer to the lecture slides or textbook if there is anything unclear to you. Hopefully it can make your preparation of the exam a little easier.

1 Introduction to Statistics

- Population vs. sample and why do we need to do sampling?
  - Population: the entire group you want to study
  - Sample: a subset of the population taken because the entire population is usually too large or impossible to analyze. The characteristics of a sample are taken to be representative of the population.
- Sample Error
  - Sample error occurs because samples are not *perfect* representations of a population.
  - Large, random samples will have low error, BUT sample error is unavoidable
- What are the difference between descriptive statistics and inferential statistics?
  - Descriptive statistics is to describe the sample data
  - Inferential statistics is to infer the population parameters of interest through sample data
- What are the special characteristics of spatial data?
  - Waldo Tobler's first law of geography
  - Spatial heterogeneity, spatial version of Simpson's paradox
  - Scale effects

2 Descriptive statistics

- Descriptive statistics is to describe and analyze *sample* data (review how to do them all in R)
- Histogram
  - What can affect the appearance of a histogram?
- Measures of data: mean, median, mode, standard deviation, quartiles (quantiles)
- Measure central tendency:
  - Mean: $\bar{x}$ (stable value and useful in analysis though sensitive to outliers)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

    - ∗ *Note:* Sample mean is an unbiased estimator of population mean μ
  - Mode: Data value that occurs most often, corresponding to the highest peak in the histogram
  - Median: 50th percentile (tolerant to outliers though not as useful in statistical inference)

- Measure dispersion: variances and standard deviation and coefficient of variance
  - Range: Max - Min (crude and inaccurate)
  - Interquartile Range: 75th - 25th quartile
  - Sample standard deviation, s, an estimate of population standard deviation σ:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n-1}}$$

    - *Note:* The standard deviation is calculated on n-1 degrees of freedom rather than n because dividing by n would yield a biased estimator of σ
  - Variance: sd^2. SD is used more often
  - Coefficient of variation: ratio of standard deviation and mean

$$\frac{s}{\bar{x}}$$

    - A unitless measure of dispersion, meaning the value does not dependent on unit of the measurements (unlike standard deviation σ or variances σ^2)
- Measure skewness:
  - Data are skewed right if the histogram points right and there are lots of outliers on the high end
  - Data are skewed left if the histogram points left and there are lots of outliers on the low end
  - skewness is a statistic that measures these
    - skewness close to 0: symmetric/unskewed
    - skewness $<$ 0: skewed left
    - skewness $>$ 0: skewed right
- Z-Score:
  - A measurement describing how far some observation is from the mean
    - $z = \frac{value - mean}{standard\ deviation}$
- Applications of descriptive statistics to spatial data:
  - Mean center of spatial point pattern → mean
  - Median center of spatial point pattern → median
  - Standard distances of spatial point pattern → standard deviation
  - Quadrat method → histogram

3 Graphics and plots in statistics

- “a picture is worth of thousands of words”
- Box-plot
  - Useful to compare multiple sets of sample data
  - Easy to compare the degree of dispersion and skewness and to identify outliers
- Scatter plot
  - Show the correlation between two variables
- Line plot
  - Good for time-series observations
- Bar chart and pie chart
  - Why pie chart is the worst chart in the world?
- **The Principle of Proportional Ink**

  - "The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented"
  - Examine plots and identify possible misconception it could lead to
  - There are lots of ways to mislead with plots without actually "lying" or falsifying data. Think about the motivations of who made the plot!

4 Statistical relationship

- Statistical relationship between two variables
- Scatter plot is often used to display the relationship
  - up trend → positive
  - down trend → negative
  - tend to flat → no strong relationship
- Sample correlation coefficient can quantify the strength of relationship
- Correlation coefficient (formal name is Pearson's correlation coefficient)
  - It is a standardized version of sample covariance:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

    * $-1 \leq r \leq 1$
    * $r > 0$ indicates positive correlation
    * $r < 0$ indicates negative correlation
    * $r = 0$ indicates no correlation
    * Should review how to do in R: cor()
- Interpretation of correlation coefficient
  - Correlation coefficient reflects the strength and direction of a linear relationship, but not the slope of that relationship, nor many aspects of nonlinear relationships
  - It cannot tell the spurious relationship
- Simple linear regression
  - $y = a + bx$
  - y: response variable or dependent variable
  - x: explanatory variable or independent variable
  - Correlation coefficient explains how well the data close to the regression line
  - Should review how to do in R: lm()
- Applications in spatial data:
  - Modifiable areal unit problem: the geographic aggregation of spatial data can affect the results of correlation coefficients dramatically

5 Probability

- Probability defined: $P(A) = \frac{n(A)}{n}$
- Intersection: Multiplication rule (assuming independence):
  - $P(A\ and\ B) = P(A) * P(B)$
  - Also called $P(A \cap B)$
- Union: Addition rule (assuming independence):

  - $P(A\ or\ B) = P(A) + P(B) - P(AandB)$
  - Also called $P(A \cup B)$

- Conditional Probability:
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - If B happens, what is the prob that A also happens?

- How to check if two variables are independent with each other?
  - Using conditional probability: $P(A|B) = P(A)$
  - Using joint probability: or $P(AandB) = P(A) \times P(B)$

- Bayes' rule: inverting conditional probabilities**
  - $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
  - $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{\underbrace{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}_{\text{law of total probability}}}$
  - How to use it to solve problems? Please see the slides and lab 5 for examples using Bayes' rule

- Example of a probability distribution: Normal distribution:
  - Standard normal distribution with 0 mean $\mu = 0$ and 1 as standard deviation $\sigma = 1$
  - Z-score:
    $$Z = \frac{x - \mu}{\sigma}$$
    * It measures the distance of a number x from the majority
  - three sigma rules

6 Sampling

- We only focus on probability based sampling methods:
  - Simple random sampling
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling

- Convenience sampling is NOT probability based and so should not be used

- Sampling distributions
  - Since sampling process is random, the statistics derived from samples (including sample mean, standard deviation) are also random and follow a sampling distribution
  - The law of large numbers and the central limit theorem make it possible to infer population parameters based on samples
  - We only focus on central limit theorem:
    * If we draw n samples of the population with mean μ and standard deviation σ. If the sampling size n is large enough, the distribution of the sample mean is approximately normally distributed with mean μ and standard deviation of $\frac{\sigma}{\sqrt{n}}$

7 Point and interval estimation

- Point estimation:
  - Sample mean $\bar{x}$ is an unbiased estimator of population mean μ

  - Sample standard deviation s is an unbiased estimator of population standard deviation σ

- Interval estimation of population mean μ when standard deviation σ is known:
  - For 95% confidence level, the confidence interval of population mean μ is the following ($\bar{x}$ is the sample mean and n is the sample size):

$$[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$$

  - Generally for confidence level $1 - \alpha$, the confidence interval for population mean is the following ($\bar{x}$ is the sample mean and n is the sample size):
  - What is $z_{\frac{\alpha}{2}}$?

$$[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

- Interval estimation of population mean μ when standard deviation σ is unknown:
  - When population standard deviation σ is unknown, we can use the unbiased estimator sample standard deviation s to estimate it, and it leads to Student's t-distribution
  - Generally for confidence level $1 - \alpha$, the confidence interval for population mean is the following ($\bar{x}$ is the sample mean and n is the sample size$).
  - What is $t_{\frac{\alpha}{2},n-1}$?

$$[\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}]$$

- Interpretation of confidence interval
  - It means that if we keeping repeat the sampling process, 95% of the samples will yield an interval containing the value of population mean μ. Or we are 95% confident that the population mean lies within the interval

8 Hypothesis testing principles

- Formation of hypothesis:
  - H_0 vs. H_a
  - Null hypothesis is a statement of NON DIFFERENCE
    - E.g. for one-sample test: The mean height of CU Boulder students is 5'8"
    - E.g. for two-sample test: The mean height of CU Boulder students is the same as CSU
  - Alternative hypothesis is a statement of difference
    - E.g. for one-sample test: The mean height of CU Boulder students is NOT 5'8"
    - E.g. for two-sample test: The mean height of CU Boulder students is NOT the same as CSU

- Test statistics
  - To perform a hypothesis testing, first you calculate a statistic that describes your sample (e.g. z-score)

- P-value
  - The probability of getting the test statistic we got assuming H_0 is true
  - Type I error
  - If the probability is really small, then H_0 is probably not true!!
  - Before running our test, we set an 'alpha level' (usually 0.05 or 0.01)
  - If p $<$ alpha, we reject the null hypothesis!

- Hypothesis testing of population mean μ when σ is known and given: z-test
    - The z-score of sample mean is standard normal distribution

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

    - P-value can be obtained based standard normal distribution
- Hypothesis testing of population mean μ when σ is unknown: t-test
    - The normalized version of sample mean is t-distribution with degree of freedom n-1

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim T_{n-1}$$

    - P-value can be obtained based Student's t-distribution
- Interpretation of p-value
    - We can interpret it as the strength of evidence or degree of belief against H0.
        - Smaller p-values → more evidence and confident against H0.
        - But we typically shouldn't *compare* p-values because they're very test-specific
        - Type I error: probability of error by rejecting H_0. Determined by the alpha level

9 t-tests: One-sample and two-sample

- t-tests are used to compare sample means (see above on concepts)
- There are 3 main types:

1. **One-Sample**: Whether a sample came from a population with population mean "x" (**one-sample t-test**)
    - E.g. Our sample average height was 164 cm, can we reject the claim that the population mean is 167?
2. **Two-Sample**: Whether two samples come from populations with the same mean (**two-sample t-test**)
    - E.g. Are CU Boulder or CSU students taller?
3. **Paired**: Whether individual elements experience change before/after a treatment/event/time period
    - E.g. Are the same CU Boulder students happier before/after midterms?

- When we run t.test in R, it will report a p-value:
    - If that p-value is less than our alpha (usually 0.05 or 0.01, depending on how confident we want to be in our result):
        - Reject the null hypothesis that the means are the same
    - If that p-value greater than our alpha:
    - Fail to reject the null hypothesis (*not* the same as proving the null hypothesis is true)