STAT 350 Notes

Josh Park

Summer 2024

Chapter 1

An Introduction to Statistics and Statistical Inference

1.1 Book goals

1.2 What Are Statistics and Probability?

1.3 How to Solve Word Problems

1.4 Brief Introduction to R

Chapter 2

Summarizing Data Using Graphs

2.1 Variables

2.1.1 Classification of variables

2.1.2 Determining if the chosen variables are appropriate for the question of interest

2.2 Basics of graphing

2.2.1 What to look for in graphs

2.2.2 Definition of terms

2.3 Visualization of numeric variables

2.3.1 Histogram - appropriate number of classes

2.3.2 Identify the shape

2.3.3 Determination of outliers

Chapter 3

Numerical Summary Measures

3.1 Center of a distribution

3.1.1 Notation

x = random variable
x_i = specific observation
n = sample size

3.1.2 Sample mean

$$\bar{x} = \frac{sum\ of\ observations}{n} = \frac{1}{n}\sum x_i \tag{3.1}$$

R command: mean(variable)

3.1.3 Sample median

$$\tilde{x} = centermost\ value\ in\ ordered\ dataset \tag{3.2}$$

R command: median(variable)

3.2 Spread or variability of the data

three common ways to measure spread:

1. sample range
2. sample variance (or stdev)
3. interquartile range (IQR)

3.2.1 Range

range = $\max(x) - \min(x)$
completely depends on extreme values, so not very reliable
no R command for this

3.2.2 Sample Variance (sample standard deviation)

Variance

$variance = s_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$
R command: var(variable)

Standard Deviation

$standard\ deviation = s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$
R command: sd(variable)

if var = sd = 0, there is no spread (all data is the same)

3.2.3 Interquartile range (IQR)

Quartile

quartile = 1/4 of the data
R command = quantile(variable)
R command for % = quantile(variable, prob=c (p1, p2))

IQR

IQR = $Q_3 - Q_1$

3.3 Boxplots

fast way to vizualize five-number summary
five number summary: minimum, first quartile, median, third quartile, maximum

3.3.1 Outliers

IF = inner fence
OF = outer fence
subscript L = lower bound
subscript H = higher bound

$$IF_L = Q_1 - 1.5(IQR) \qquad IF_H = Q_3 + 1.5(IQR) \qquad \text{mild} \tag{3.3}$$
$$OF_L = Q_1 - 3(IQR) \qquad OF_H = Q_3 + 3(IQR) \qquad \text{extreme} \tag{3.4}$$

3.4 Choosing Measures of Center and Spread

if data is skewed, use median and IQR.
if symmetric, use mean and standard deviation.

3.5 z-score

3.5.1 z-score

the z-score of a data point x_i quantifies distance from the mean value in terms of standard deviations.

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.5}$$

Chapter 4

Probability

4.1 Experiments, Sample Spaces, Events

4.1.1 Experiments

Definition. A random *experiment* is any activity in which there are at least two possible outcomes and the result of the activity can not be predicted with absolute certainty.

Note. By this definition, all experiments are random.

Definition. An *outcome* is the result of an experiment.

Definition. Each time the experiment is done is called a *trial*.

4.1.2 Tree Diagrams

Note. This section is trivial

4.1.3 Sample Spaces

Definition. The *sample space* of an experiment is the set of all possible outcomes, denoted by S or Ω.

4.1.4 Events

Definition. An *event* is any collection of outcomes from an experiment. The sample space is one possible event.

Definition. A *simple event* only has one outcome.

We say that an event has occurred if the resulting outcome is containied in the event.

4.1.5 Set Theory

Definition. The *complement* of an event A contains every outcome in the sample space that is not in A, denoted by A'.

Note. Remainder of section is trivial.

4.2 Introduction to Probability

4.2.1 What is Probability?

Frequentist POV

In the frequentist interpretation of probability, we say the probability of any outcome of any random experiment is the long term proportion of times that the outcome occurs over the total number of trials.

$$P(A) = \lim_{N \to \infty} \frac{n}{N} \tag{4.1}$$

Bayesian POV

In Bayesian probability, the probabilist specifices some *prior probability*, which is then updated upon collection of *relevant data*.

4.2.2 Properties

1. Given any event A, it must be that $0 \leq P(A) \leq 1$.
2. Assuming ω is an outcome of A, then $P(A) = \sum P(\omega)$. That is, the sum of probabilities of all outcomes in an event is equal to the probability of the event.
3. The probability of the sample space is 1. That is, $P(\Omega) = 1$.
4. The proability of the empty set is 0. That is, $P(\emptyset) = 0$.

4.2.3 Rules

Complement rule. For any event A, $P(A') = 1 - P(A)$
General additional rule. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Note. When adding disjoint probabilities we need not subtract the last term, as the intersection will be empty.

4.3 Conditional Probability and Independence

4.3.1 What is conditional probability?

A *conditional probability* is written $P(A|B)$ and is read 'the probability of A, given that B occurs'.

4.3.2 General Multiplication Rule

To calculate a union (or 'or'), we can use the general additional rule. To calculate an intersection (or 'and'), we can use the general multiplication rule.

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \implies P(A \cap B) = P(A)P(B|A) \tag{4.2}$$

Additionally, this rule can be applied to an arbitrary number of unions.

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) \tag{4.3}$$

4.3.3 Tree Diagrams revisited

Note. Trivial

4.3.4 Bayes' Rule using Tree Diagrams

Bayes' rule

We use Bayes' rule when calculating a conditional probability in one direction, but you only know the conditional probability in the other direction. This method is not needed when the probability of the intersection is known.
To find the probability of A given B,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{4.4}$$

If we don't know $P(A \cap B)$, we use the general multiplication rule to write

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4.5}$$

If we know what $P(B)$ is, we are done. Otherwise, we use the fact that

$$P(B) = P(B \cap A) + P(B \cap A') = P(B|A)P(A) + P(B|A')P(A'). \tag{4.6}$$

This is called the *Law of Total Probabilities for Two Variables.*
Subbing eqn 4.6 into eqn 4.5, we get *Bayes' Rule for two variables*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')} \tag{4.7}$$

For more than two variables, suppose the sample space is partitioned into k disjoint events, $A_1, A_2, \ldots, A_k$, none of which have a probability of 0, such that

$$\sum_{i=1}^{k} P(A_i) = 1 \tag{4.8}$$

Then, the *Law of Total Probability* is

$$\sum_{i=1}^{k} P(B|A_i)P(A_i) \tag{4.9}$$

and Bayes' rule is

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)} \tag{4.10}$$

Bayesian Statistics

To summarize Bayesian Statistics, we first begin with a prior probability A. Then, given additional context, B, we can improve our prediction of the probability of A by calculating $P(A|B)$. This is called the *posterior* probability. In the example from the book, after someone tested positive for the disease, the probability that they have the disease increased from 0.01 to 0.165.

4.3.5 Independence

Definition. Two events are *independent* if knowing the outcome of one does not affect the outcome of the other. Mathematically, we write

$$P(A|B) = P(A) \tag{4.11}$$

$$P(B|A) = P(B) \tag{4.12}$$

or

$$P(A \cap B) = P(A)P(B|A) \implies P(A \cap B) = P(A) \times P(B) \tag{4.13}$$

Disjoint vs Independence

Definition. Two events are *disjoint* if they can not possibly occur at the same time.
Definition. Two events are *independent* if the outcome of one does not impact the other.

1. Draw a card. A: card is a heart, B: card is not a heart

   disjoint; not independent

2. Toss 2 coins. A: first coin is head, B: second coin is head

   not disjoint; independent

3. Roll 2 4-sided die. A: first die is 2, B: sum of die is 3

   not disjoint; not independent

Chapter 5

Random Variables and Discrete Probability Distributions

5.1 Random Variables

5.1.1 Random Variables

Definition. A *random variable* is a numerical characteristic obtained from a random experiment. So, random variables are functions and follow all properties of mathematical functions.

5.1.2 Probability Distributions - pmf

Definition. The probability distributions of a random variable is called the *probability mass function (pmf)*. In symbols, $p(x) = P(X = x)$

5.1.3 Properties

Pmfs are valid probability distributions, so they follow the axioms of probability.

1. $0 \leq p_i \leq 1$. Each probability lies between 0 and 1.
2. $\sum_i p_i(x) = 1$. The sum of all probabilities is 1.

5.2 Expected Value and Variance

5.2.1 Expected Value

Definition. The *expected value* of a discrete random variable X is the weighted average of each value.

$$E(X) = \mu_X = \sum_{i=1}^{m} x_i p_i \tag{5.1}$$

5.2.2 Rules of Expected Values

1. If X is a random variable and a and b are fixed, then

   $$E(a + bX) = a + bE(X)$$

2. If X and Y are random variables, then

   $$E(X + Y) = E(X) + E(Y)$$

3. If X is a random variable and g is a function of X, then

$$E(g(X)) = \sum_{i=1}^{m} g(x_i)p_i$$

5.2.3 Variance and Standard Deviation

Recall sample variance measures spread by taking the average of the squared differences between observations and their center

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{5.2}$$

Note. Define the population variance of X by $Var(X)$, σ^2, or σ_X^2.

$$Var(X) = \sigma^2 = \sigma_X^2 \tag{5.3}$$

The population variance is the expected squared difference between X and μ_X.

$$Var(X) = E[(X - \mu_X)^2] = \sum (x_i - \mu_X)^2 \cdot p_i \tag{5.4}$$

Simplify.

$$Var(X) = E[(X - \mu_X)^2] = E(X^2) - (E(X))^2 \tag{5.5}$$

Then, the standard deviation is the sqaure root of the variance

$$\sigma_X = \sqrt{Var(X)} \tag{5.6}$$

5.2.4 Rules of Variance

5.3 Cumulative Distribution Function

5.4 Binomial Random Variable

5.4.1 Binomial Experiment

5.4.2 Binomial Probabilities

5.4.3 Mean and Variance

5.5 Poisson Random Variables

5.5.1 Poisson Experiment

5.5.2 Poisson Probabilities

5.5.3 Mean and Variance

Chapter 6

Continuous Probability Distributions

6.1 Probability Distribution for Continuous Random Variables - General

Objectives

1. Describe the basis of the probability density function (pdf).
2. Use the probability density function (pdf) and cumulative distribution function (cdf) of a continuous random variable to calculate probabilities and percentiles (median) of events.
3. Be able to use a pdf to find the mean of a continuous random variable.
4. Be able to use a pdf to find the variance of a continuous random variable.

6.1.1 Density curves and probabilities (pdf)

Define the pdf $f(x)$ such that $\int_{\infty}^{\infty} f(x)\mathrm{d}x = 1$.
Then, the probability that $a < X < b$ is

$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x$$

Note. When $X = a$, $\int_a^a f(x)\mathrm{d}x = 0$, so $P(X \leq a) = P(X < a)$

6.1.2 Properties

A valid density curve must have the two following properties:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$

Note. Notice that $f(x) \leq 1$ need not be true; consider the function $g(x) = 4$ on the interval [0, 0.25].
Note. In the case of $g(x)$, the bounds on the integral for property 2 must be adjusted

6.1.3 Mean and Variance

Note. The rules for the means and variances are the same for both discrete and continouous random variables; the only difference is how the values are computed.

$$\text{Discrete:} \qquad E(X) = \mu_X = \sum xp(x) \qquad E(g(X)) = \sum g(x)p(x) \tag{6.1}$$

$$\text{Continuous:} \qquad E(X) = \mu_X = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x \qquad E(g(X)) = \int_{-\infty}^{\infty} g(x)p(x)\mathrm{d}x \tag{6.2}$$

Recall the formula for variance of discrete random variables

$$\begin{aligned} Var(X) &= E[(X-\mu_X)^2] = \sum (x_i - \mu_X)^2 \cdot p_i \\ &= E[(X-\mu_X)^2] = \int_{-\infty}^{\infty} (x_i - \mu_X)^2 \cdot p_i \end{aligned} \tag{6.3}$$

$$Var(X) = E(X^2) - (E(X))^2 \tag{6.4}$$

$$\sigma_X = \sqrt{Var(X)} \tag{6.5}$$

Equation 6.4 is recommended as it is computationally much easier to evaluate.

6.1.4 Cumulative Distribution Function (cdf)

The cumulative distribution function (cdf) is the probability that the random variable will be less than or equal to some value. It is written $F(X)$ and the formula is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)ds \tag{6.6}$$

Note. The variable of integration is changed to be some dummy variable s, as the bounds of a definite integral can not be a function of the variable of integration.
To recap, we now have

$$p(x) = \text{probability mass function} \tag{6.7}$$

$$f(x) = \text{probability density function} \tag{6.8}$$

$$F(x) = \text{cumulative distribution function} \tag{6.9}$$

6.1.5 Percentiles

For continuous distributions, percentiles are much simpler to compute. Given that $0 < p < 1$, the $100p$th percentile for a value x can be computed with the integral

$$p = \int_{-\infty}^{x} f(x)ds \tag{6.10}$$

Note that this integral is the same as the cdf. Thus if the cdf is already known, we can simply find when $F(x) = p$. Again, the $100p^{\text{th}}$ percentile is when $100p$ percent of the data is less than p, and the rest is above
Note. The median occurs when $p = 0.5$. Hence,

$$p = 0.5 = \int_{-\infty}^{\mu'} f(x)dx = F(\mu') \text{ where } \mu' = \tilde{\mu} \tag{6.11}$$

6.2 Normal Distribution

6.2.1 Distribution

6.2.2 Standardization

6.2.3 Using the z-table

6.2.4 Probabilities

6.2.5 Percentiles

6.3 Determining if a distribution is normal

6.3.1 Normal probability plots

6.4 Uniform Distribution

6.4.1 Distribution

6.5 Exponential Distribution

6.5.1 Distribution

6.6 Other continuous distributions (Optional)

6.6.1 Gamma Distribution

6.6.2 Beta Distribution

6.6.3 Weibull Distribution

6.6.4 Lognormal Distribution

Chapter 7

Sampling Distributions

7.1 Parameters and Statistics

7.2 Sampling Distribution of a Sample Mean

7.2.1 What is a sampling distribution?

7.2.2 The mean and standard deviation of a sampling distribution

7.2.3 The shape of a sampling distribution

Chapter 8

Experimental Design

8.1 Sources of Data

8.1.1 Anecdotes

8.1.2 Available data

8.1.3 Experiments versus Observational Studies

8.2 Designing Studies

8.2.1 Identify parts of the study

8.2.2 Comparative studies

8.2.3 Principles of study design

8.2.4 Problems with studies

8.2.5 Matched pairs and block designs

8.2.6 Good designs

8.3 Sampling

8.3.1 SRS

8.3.2 Stratified random sample

8.3.3 Bad sampling techniques

8.3.4 Good techniques

8.4 Causality

8.4.1 Type of lurking variable

Common response

Confounding

8.4.2 The best way to determine causality

8.4.3 Problems of lurking variables

8.5 Ethics

Chapter 9

Confidence Intervals based on a Single Sample

9.1 Introduction to Statistical Inference

9.2 Point Estimation

9.2.1 Definitions

9.2.2 Which estimator to use?

9.2.3 Biased/Unbiased Estimators

9.2.4 MVUE

9.3 Confidence Interval (CI) for the population mean when the standard deviation is known

9.3.1 Assumptions

9.3.2 Why an interval is needed

9.3.3 Definitions

9.3.4 Derivation

9.3.5 Interpretation

9.3.6 Precision of confidence interval

9.3.7 Sample size determination

9.3.8 Methodology

9.3.9 Confidence bounds

9.3.10 Summary

9.3.11 Cautions

9.4 Confidence Interval (CI) for the population mean when the standard deviation is not known

9.4.1 Assumptions

9.4.2 Changes from when the standard deviation is known

9.4.3 t-distribution

Chapter 10

Hypothesis Tests Based on a Single Sample

10.1 Assumptions for Inference

10.2 The parts of the hypothesis test

10.2.1 Parts of the hypothesis test

10.2.2 The Hypotheses

10.3 Hypothesis Test Errors and Power

10.3.1 Type I and Type II errors

10.3.2 Power: meaning and calculations

10.4 Hypothesis tests for a population mean when σ is known

10.4.1 The test statistic

10.4.2 p-value

10.4.3 Conclusion

10.4.4 General procedure for Hypothesis Testing

10.4.5 Relationship between confidence intervals and hypothesis tests

10.5 Hypothesis tests for a population mean when σ is unknown

10.5.1 Summary

10.6 Inference

10.6.1 Inference for non-normal distributions

10.6.2 Final Comments

p-value and significance levels

Practical significance

Other

Chapter 11

CI and HT Based on Two Samples or Treatments

11.1 Introduction

11.1.1 Notation

11.1.2 Definitions

11.2 Two Independent Samples

11.2.1 Assumptions

11.2.2 Mean and standard deviation - known σ

11.2.3 Hypothesis test - known σ

11.2.4 Confidence Interval - known σ

11.2.5 Standard deviation - unknown σ

Pooling

Formula

Satterthwaite Approximation

11.2.6 Summary

11.2.7 Robustness

11.3 Two Sample Matched Pair

11.3.1 Assumptions

11.3.2 Mean and standard deviation

11.3.3 Summary

11.4 Which method to use?

Chapter 12

The Analysis Of Variance (ANOVA)

12.1 One-Way ANOVA

12.1.1 Definitions

12.1.2 Background - Why use variance?

12.1.3 Notation

12.1.4 Hypotheses

12.1.5 Assumptions

12.1.6 Model

12.1.7 Statistics

12.1.8 Test Statistic - conceptual

12.1.9 ANOVA Table

Formulas

Test Statistic - calculation

p-value

ANOVA Table - summary

12.1.10 Hypothesis Test - summary

12.1.11 t-test versus F-test

12.2 Determining Differences

12.2.1 How to determine which means are different?

12.2.2 Why not use two-sample independent?

12.2.3 General method

12.2.4 LSD (Fishers)

12.2.5 Bonferroni

12.2.6 Tukey

12.2.7 Dunnett

12.2.8 Graphical Display

12.2.9 Procedure

Chapter 13

Correlation and Linear Regression: Simple Linear Regression

13.1 Simple Linear Regression

13.1.1 Purpose

13.1.2 Definitions

13.1.3 Scatterplots

13.1.4 Theory

13.1.5 Results

13.1.6 ANOVA table

13.1.7 Coefficient of Determination

Chapter 14

Correlation and Linear Regression: Correlation, Diagnostics, Inference

14.1 Correlation

14.1.1 Calculations

14.1.2 Properties

14.2 Diagnostics

14.2.1 Scatterplot

14.2.2 Residual plot

14.2.3 Normality tests

14.3 Inference 1

14.3.1 Background

14.3.2 F-test

14.3.3 Inference for the slope

14.4 Inference 2

14.4.1 Confidence Interval for the mean at a point

14.4.2 Prediction Interval