STAT 350 Notes

Josh Park

Summer 2024

Chapter 1

An Introduction to Statistics and Statistical Inference

Chapter 2

Summarizing Data Using Graphs

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

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

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.

$$z_i = \frac{x_i - \bar{x}}{s} \qquad (3.5)$$

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 $0 \le P(A) \le 1$.

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

Chapter 6

Continuous Probability Distributions

Chapter 7

Sampling Distributions

Chapter 8

Experimental Design

Chapter 9

Confidence Intervals based on a Single Sample

Chapter 10

Hypothesis Tests Based on a Single Sample

Chapter 11

CI and HT Based on Two Samples or Treatments

Chapter 12

The Analysis Of Variance (ANOVA)

Chapter 13

Correlation and Linear Regression: Simple Linear Regression

Chapter 14

Correlation and Linear Regression: Correlation, Diagnostics, Inference