STAT 350 Notes

Josh Park

Summer 2024

Chapter 1

An Introduction to Statistics and Statistical Inference

1.1 Intro

1. Data Collection: The process of gathering information.
2. Descriptive Statistics: Summarizing and organizing data.
3. Inferential Statistics: Drawing conclusions from data.

1.2 What Are Statistics and Probability?

Definition 1.1: Claim

a statement that we assume to be true

Definition 1.2: Status quo

existing state/condition

Definition 1.3: Population

set of all things to be studied

Definition 1.4: Sample

subset of population

Definition 1.5: Probability

know everything about population; want to know about sample

Definition 1.6: Inferential Statistics

know everything about sample; want to know about population

Chapter 2

Summarizing Data Using Graphs

2.1 Variables

2.1.1 number of observations

single - univariate
double - bivariate
3+ - multivariate

2.1.2 type

numerical; categorical (nominal[unordered] or ordinal[ordered]); discrete; continuous;

2.2 Basics of graphing

look for overall pattern; deviations; shape; center; variability

2.2.1 Frequency Distribution

Definition 2.1: Bins

intervals that categorical variables are sorted into

Definition 2.2: Frequency distribution

the frequency of number of observations in each class

Definition 2.3: relative frequency

the measured frequency divided by the total data points

$$relativefreq = \frac{freq}{totalcount} \tag{2.1}$$

2.3 Displaying quantitative variables

2.3.1 Histogram - appropriate number of classes

$$\#ofbins \approx \sqrt{\#ofobservations}$$

2.3.2 Identify the shape

peaks: unimodal, bimodal, or multimodal
positively skewed unimodal (peak goes left)
negatively skewed unimodal (peak goes right)

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 4.1: Random Experiment

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 4.1:

By this definition, all experiments are random.

Definition 4.2: outcome

the result of an experiment.

Definition 4.3: Trial

Each time the experiment is done.

4.1.2 Tree Diagrams

Note 4.2:

skip

4.1.3 Sample Spaces

Definition 4.4: sample space

the set of all possible outcomes of an experiment, denoted by S or Ω.

4.1.4 Events

Definition 4.5: An *event* is any collection of outcomes from an experiment.

Example 4.1: The sample space is one possible event.

Definition 4.6: A *simple event* only has one outcome.

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

4.1.5 Set Theory

Definition 4.7: complement

every outcome of an event A in the sample space that is not in A, denoted by A'.

Note 4.3: 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 specifies 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 probability of the empty set is 0. That is, $P(\emptyset) = 0$.

4.2.3 Rules

Definition 4.8: complement rule

for any event A, $P(A') = 1 - P(A)$

Definition 4.9: general additional rule

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Note 4.4:

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), we can use the general additional rule. To calculate an intersection (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 4.5:

skip

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, $A_1, A_2, \dots, A_k$, 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 4.10: Independence

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

$$\begin{align} P(A|B) &= P(A) \tag{4.11}\\ P(B|A) &= P(B) \tag{4.12}\\ P(A \cap B) = P(A)P(B|A) &\implies P(A \cap B) = P(A) \times P(B) \tag{4.13} \end{align}$$

Disjoint vs Independence

Definition 4.11: Disjoint

Two events are *disjoint* if they can not possibly occur at the same time.

Definition 4.12: Independence

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 5.1: random variable

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 5.2: probability mass function (pmf)

The probability distributions of a random variable. 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 5.3: expected value

the weighted average of each value of a discrete random variable X.

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

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

Note 5.1:

$g(x)$ does not have to be linear

5.2.3 Expectation, Variance, and Standard Deviation for a Discrete Variable

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 5.2:

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

1. if X is a random variable and $a, b \in \mathbb{R}$ are fixed then
$$Var(a+bX) = b^2 Var(X) \tag{5.7}$$

Note 5.3: the variance does not rely on where hte center of the distribution is so a is not on RHS

2. if X and Y are independent random variables,
$$Var(X+Y) = Var(X) + Var(Y) \quad and \quad Var(X-Y) = Var(X) + Var(Y) \tag{5.8}$$

   addition rule for variances
   can be added for both addition and subtraction

5.3 Cumulative Distribution Function

skip

5.4 Binomial Random Variable

5.4.1 Binomial Experiment (BInS)

Binary - are there only 2 options?
Independence - is each trial independent?
Number - is the number of trials a contant?
Success - is the probability of success a constant?

5.4.2 Binomial Random Variables

Definition 5.4: Binomial Random Variable

The *binomial random variable* X must come from a binomial experiment and maps every outcome to some $k \in \mathbb{R}$. X has parameters n (num trials) and p (prob success).

5.4.3 PMF of Binomial Random

If X is a binomial random bariable with n trials and probility of a success p, then

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0, 1, \dots, n \tag{5.9}$$

5.4.4 Shapes of Binomial Distributions

skewedness of a binom random variable depends on p.
$p < 0.5$; distribution is skewed right
$p = 0.5$; distribution is symmetric
$0.5 < p$; distribution is skewed left

5.4.5 Binomial Distribution: Mean and Standard Deviation

If $X \sim B(n, p)$, then

$$\mathbb{E}(X) = \mu_X - = np \tag{5.10}$$

$$Var(X) = np(1-p) \tag{5.11}$$

$$\sigma_X = \sqrt{np(1-p)} \tag{5.12}$$

5.5 Poisson Random Variables

Definition 5.5: the *poisson random variable* is used to count the number of events that happen during some interval or in some area.

Example 5.1: # of people who enter the PMU per day; # of radioactive particles emitted from some material in 1 minute

5.5.1 Poisson Experiment

The poisson distribution also has an experiment, but it is more complex than binomial

1. the probability that an event occurs in some given interval is equal for any unit of equal size; the *rate* is proportional to the size
2. the number of events that occur within an interval i s independent of the number that occur in any other non-overlapping interval
3. the probaiblity that more than one event occurs within a unit of measure is negligible for small units

5.5.2 Poisson Distribution

only one param: λ
Read $X \sim Poisson(\lambda)$ as "X is distributed as a Poisson random variable with parameter λ"

$$p(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x \in \mathbb{N}_0 \tag{5.13}$$

5.5.3 Mean and Variance

The mean and variance of the Poisson random variable are the same as λ. So,

$$\mu_X = \sigma^2 \tag{5.14}$$

$$\sigma_X = \sqrt{\lambda} \tag{5.15}$$

λ used instead of μ to maintain consistency with exponential distribution.
If units for λ were different than the units given in the problem, the problem can still be oslved due to the independence property (2) ofthe Poisson experiment.

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

$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x \tag{6.1}$$

Note 6.1:

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 6.2:

Notice that $f(x) \leq 1$ need not be true; consider the function $g(x) = 4$ on the interval [0, 0.25].

Note 6.3:

In the case of $g(x)$, the bounds on the integral for property 2 must be adjusted

6.1.3 Mean and Variance

Note 6.4: Rules

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.

$$\text{Discrete:} \qquad E(X) = \mu_X = \sum xp(x) \qquad E(g(X)) = \sum g(x)p(x) \tag{6.2}$$

$$\text{Continuous:} \qquad E(X) = \mu_X = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x \qquad E(g(X)) = \int_{-\infty}^{\infty} g(x)p(x)\mathrm{d}x \tag{6.3}$$

Recall the formula for variance of discrete random variables

$$\begin{aligned} Var(X) &= E[(X-\mu_X)^2] = \sum (x_i - \mu_X)^2 \cdot p_i \\ &= E[(X-\mu_X)^2] = \int_{-\infty}^{\infty} (x_i - \mu_X)^2 \cdot p_i \end{aligned} \tag{6.4}$$

$$Var(X) = E(X^2) - (E(X))^2 \tag{6.5}$$

$$\sigma_X = \sqrt{Var(X)} \tag{6.6}$$

Equation 6.5 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

$$F(x) = P(X \leq x) = \int_{-\infty}^{s} f(x)ds \tag{6.7}$$

Note 6.5:

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

$$p(x) = \text{probability mass function} \tag{6.8}$$

$$f(x) = \text{probability density function} \tag{6.9}$$

$$F(x) = \text{cumulative distribution function} \tag{6.10}$$

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

$$p = \int_{-\infty}^{x} f(x)ds \tag{6.11}$$

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 6.6:

The median occurs when $p = 0.5$. Hence,

$$p = 0.5 = \int_{-\infty}^{\mu'} f(x)\mathrm{d}x = F(\mu') \text{ where } \mu' = \tilde{\mu} \tag{6.12}$$

6.2 Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{where} - \infty < \mu < \infty \tag{6.13}$$

don't need to know closed form
defined as long as we know the variance σ^2 and the mean μ
thus $X \sim N(\mu, \sigma^2)$

6.2.1 Normal Distribution

shape: symmetric; bell shaped; unimodal
variance σ^2 is spread; mean is also median since symmetric
curve is concave down when x is $\mu \pm \sigma$ otherwise concave down

6.2.2 Standardization

empirical rule - 68% of data within 1 stdev; 95 within 2; 99.7 within 3
standard normal distribution: $\mu = 0; \sigma = 1$

Definition 6.1: z

$z = f(z)$

probability of certain outcome $x =$ cdf of z; values are in z table

6.2.3 Using the z-table

$$P(-2.77 < Z < 1.54) = P(Z < 1.54) - P(Z < -2.77) = 0.9382 - 0.0028 = 0.9354 \tag{6.14}$$

6.2.4 Probabilities

normalize any given distribution to be standard normal with

$$z = \frac{x - \mu}{\sigma} \tag{6.15}$$

then z is z-score; tells how many standard deviations x is from the mean μ

6.2.5 Percentiles

Example 6.1:

89th percentile $= P(Z < b) = 0.89$

must find corresponding value in z-table (don't interpolate; use printed value)
now un-normalize z score by

$$z = \frac{x-\mu}{\sigma} \implies x = \mu + \sigma z \tag{6.16}$$

Symmetry

How do we find $P(\mu - b \leq X \leq \mu + b)$ (symmetry around the mean)?
Let $P(\mu - b \leq X <\leq \mu + b) = C$. Then each little "wedge" on each end is $\frac{1-c}{2}$
Then upper bound of C is $P(X < \mu + b) = 1 - \frac{1-C}{2}$; and lower bound is $\frac{1-C}{2}$

6.3 Determining Normality

6.3.1 Normal probability plots

Ways to check normality

1. graph and compare
2. backward empirical rule to see if proportions match
3. ratio of IQR:s shoudl be about 1.4
4. normal probability plot (QQ) (best method)
5. inference

6.3.2 Conceptual Procedure for Normal Probability Plots

idea: if data fits distribution, the percentiles should correspond with the percentiles of that distribution

1. arrange from smallest to largest
2. record corresponding percentiles (complicated)
3. find z value
4. plot original data point vs new z point

If data is normal, we can use x and z to obtain the mean and standard

6.4 Uniform Distribution

continuous distribution; all points are distributed evenly between a, b.
density curve:

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & else \end{cases} \tag{6.17}$$

Mean is

$$\mathbb{E}(X) = \frac{a+b}{2} \tag{6.18}$$

standard deviation

$$\sigma_X = \frac{b-a}{\sqrt{12}} \tag{6.19}$$

Example 6.2:

200 boxes an hour; distribution is uniform from 18.2 to 20.4 (nearest tenth)

1. what is the probability that the package weighs less than 20 lbs?

$$P(X < 20) = \int_{18.2}^{20} \frac{1}{20.4 - 18.2} dx = \int_{18.2}^{20} \frac{1}{2.2} dx = 0.818 \tag{6.20}$$

2. what are the mean and standard deviation of the weights?

$$\mathbb{E}(X) = \frac{18.2 + 20.4}{2} = 19.3 \tag{6.21}$$

$$\sigma_X = \frac{20.4 - 18.2}{\sqrt{12}} = 0.635 \tag{6.22}$$

6.5 Exponential Distribution

exponential distribution pdf:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & else \end{cases} \tag{6.23}$$

cdf:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{\lambda x} & x \geq 0 \end{cases} \tag{6.24}$$

$$E(X) = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2} \qquad \sigma_X = \frac{1}{\lambda} \tag{6.25}$$

Chapter 7

Sampling Distributions

7.1 Parameters and Statistics

Definition 7.1: Parameter

a *parameter* is a number that represents a characteristics of a population

Example 7.1: Parameter

μ, σ

Definition 7.2: Statistic

a *statistic* is any number calculated from a sample

Example 7.2: Statistic

$\bar{x}, s$

statistics are random variables because the results of each random experiment is unknown

7.2 Sampling Distribution of a Sample Mean and CLT

Definition 7.3: Sampling distribution

the probability distribution of a statistic is called a *sampling distribution*

Note 7.1:

now have 2 distributions: sampling (for a statistic) and population (for the population)

7.2.1 What is a sampling distribution?

sampling distr is theoretical; can never take all samples. also called probability distribution of the statistic

7.2.2 The mean and standard deviation of a sampling distribution

we use $\bar{x}$ (sample mean) to make inferences about the population mean
$\bar{x}$ varies from different samples so we must consider the distribution of $\bar{X}$

Note 7.2:

$\bar{X}$ is the sample mean, so we are interested in $\mathbb{E}(\bar{X})$

7.2.3 The shape of a sampling distribution

If a population distribution is normal, the sampling distribution is also normal

$$\text{If } X \sim N(\mu, \sigma^2)\text{, then the sample distribution of } \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

If a population is not normal, the sampling distribution is approximately normal (given large sample size)

Let $\bar{X}$ be the mean of observations in a random sample space of size n drawn from a population with mean μ and finite variance σ^2. If the sample size n is large enough, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Chapter 8

Experimental Design

8.1 Sources of Data

8.1.1 Anecdotes

Definition 8.1: Anecdote

A story about a single data point

8.1.2 Available data

Definition 8.2: Available data

Data that has already been taken

8.1.3 Experiments versus Observational Studies

Definition 8.3: Experiment

A situation where a researcher chooses the conditions and manipulates variables

Definition 8.4: Observational study

A situation where a researcher only observes and does not influence the variables

8.2 Designing Studies

8.2.1 Identify parts of the study

Definition 8.5: Experimental unit/sample unit

the objects of interest in a study (e.g. humans/subjects)

Definition 8.6: factor

the specific experimental conditions applied to the units

Definition 8.7: level

the factor under consideration

8.2.2 Principles of study design

control

bias

randomization

other designs

replication

8.2.3 Problems with studies

8.2.4 Matched pairs and block designs

8.2.5 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

interested in population mean, so we will use the statistic $\bar{X}$

9.2 Point Estimation

9.2.1 Definitions

Definition 9.1: point estimate

a single mumber computer from a sample; best guess for parameter (denoted by θ)

Definition 9.2: estimator

statistic of interest; random variable

Definition 9.3: estimate

value produced by estimator

Note 9.1:

Do not need to estimate statistic since it has already been calculated, only need to estimate population parameter

9.2.2 Which estimator to use?

look for certain properties in distribution
ideal distribution has

1. symmetry
2. low variance

Note 9.2: warning

low variance doesn't always mean estimator is accurate

9.2.3 Biased/Unbiased Estimators

Definition 9.4: unbiased estimator

a statistic $\hat{\theta}$ of a population parameter θ if $\mathbb{E}(\hat{\theta}) = \theta$

Definition 9.5: biased estimator

a statistic $\hat{\theta}$ of a population parameter θ if $\mathbb{E}(\hat{\theta}) \neq \theta$

unbiased is usually better; can be rigorously shown that $\bar{X}$ is an unbiased estimator for μ of a normal distribution.

Example 9.1: other unbiased estimators

Sample proportion $\hat{P}$ for estimating p because $\mathbb{E}(P) = p$
Sample variance S^2 for estimating population variance σ^2 because $\mathbb{E}(S^2) = \sigma^2$

$$\mathbb{E}(S) = \mathbb{E}(\sqrt{S^2}) = \sqrt{\mathbb{E}(S^2)} = \sqrt{\sigma^2} = \sigma \tag{9.1}$$

9.2.4 Estimators with minimum variance

Definition 9.6: Minimum variance unbiased estimator (MVUE)

the unbiased estimator of θ with least variance

Note 9.3:

the MVUE does not necessarily have to be symmetric

9.3 Confidence Interval (CI) for μ (σ known)

9.3.1 Assumptions

Note 9.4: Necessary assumptions

1. data is sampled appropriately and generated by a random sample from a properly randomized experiment
2. sampling distribution is normal (follows from CLT when n is large enough)

9.3.2 Motivation

smaller confidence interval gives more accurate idea of the true value of the parameter

9.3.3 Derivation

$$\mu_{\bar{X}} = \mu_X \tag{9.2}$$

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \tag{9.3}$$

Note 9.5: Naive solution

We could use $\bar{x} \pm \sigma_{\bar{X}}$ as interval, since 68% of data is within 1 standard deviation. Alternatively, use $\bar{x} \pm 2\sigma_{\bar{X}}$ as interval, since 95% of data is within 2 standard deviations.

9.3.4 Definitions

Note 9.6: Invoking CLT

remember that $n > 30$ is only a rule of thumb and exact number will depend on each distribution; extremely skewed distributions may require higher value of n.

Note 9.7: Confidence Interval Formula

All CIs are given by the formula $estimate \pm margin\ of\ error$

Definition 9.7: Confidence interval (CI)

An interval of values constructed so that with a specified degree of confidence, the value of the population parameter lies in the interval

Definition 9.8: Confidence coefficient (C)

The probability that the confidence encloses the popoulation parameter in repeated samplings. note that $0 \leq C \leq 1$.

Definition 9.9: Confidence level

C expressed as a percentage. confidence level = 100C%

9.3.5 Interpretation

Definition 9.10: Critical value ($z_{\alpha/2}$)

a value on the measurement axis in a standard normal distribution such that

$$P(Z \geq z_{\alpha/2}) = \alpha/2 \tag{9.4}$$

Thus $z_{\alpha/2}$ is the z-value such that the area is $\alpha/2$ to the right of the curve

Note 9.8: useful relationshiops

From the symmetry of the normal distribution, it follows that

$$P(Z \leq -z_{\alpha/2}) = \alpha/2 \tag{9.5}$$

$$P(Z \leq z_{\alpha/2}) = 1 - \alpha/2 \tag{9.6}$$

Note 9.9: R code for $z_{\alpha/2}$

z $<-$ qnorm(alpha/2, lower.tail=FALSE)

9.3.6 Interpretation of CI

Note 9.10: Randomness of μ

μ is NOT a random value; it represents a fixed parameter that we are solving for.

Example 9.2: Confidence

CORRECT: "We are 95% confident that the interval contains μ" (implies μ is random)
INCORRECT: "We are 95% confident that μ lies in the interval" (implies interval is random)

9.3.7 Precision of CI

Recall the formula for margin of error

$$ME = z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}} \tag{9.7}$$

Some possible methods for minimizing the margin of error are

1. reduce critical value

   recall that higher confidence means wider interval, increasing confidence implies lowering precision. Thus, 100% confidence implies the interval includes all possible outcomes, giving us no additional information.

2. reduce σ

   lower variance means the data is more centralized, so a smaller interval will be able to capture a greater percentage of values.

3. increase n

   recall that n is in the denominator of the equation for standard error, so a greater number of trials will decrease the standard deviation of the sampling distribution

Determining sample size

Standard deviation is already as minimized as possible and confidence level depends on the specific experiment, so the only viable way to reduce the interval is increasing sample size. The sample size has to be an integer, so some algebraic manipulation of equation 9.7 gives

$$n = \left\lceil \left(\frac{z_{\alpha/2} \cdot \sigma_X}{ME} \right)^2 \right\rceil \tag{9.8}$$

9.3.8 Practical procedure

1. plan experiment to minimize standard deviation as much as possible
2. determine lowest acceptable confidence level
3. determine largest acceptable width of confidence interval
4. compute n with equation 9.8
5. perform experiment

9.3.9 Confidence bounds

Definition 9.11: Confidence bound

a one sided confidence interval; useful when only one direction matters

$$\text{upper confidence bound:} \quad \mu < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \tag{9.9}$$

$$\text{lower confidence bound:} \quad \mu > \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \tag{9.10}$$

9.4 Confidence Interval (CI) for μ (σ unknown)

9.4.1 Assumptions

don't know μ or σ, so use sample standard deviation s to estimate the population standard deviation σ

9.4.2 Changes from when the standard deviation is known

normalize critical value

$$z = \frac{\bar{x} - \mu}{\sigma/n} \implies t = \frac{\bar{x} - \mu}{s/n} \tag{9.11}$$

no longer standard normal; called t-distribution

9.4.3 t-distribution

t-distribution has more than one curve because the shape depends on "degrees of freedom", denoted by ν.

$$\nu = n - 1 \tag{9.12}$$

Note that as $n \to \infty$, $s \to \sigma$ and $\mu_X \to \mu_{\bar{X}}$. Thus as $n \to \infty$, the curve becomes the standard normal distribution

9.4.4 Summary confidence interval/bounds for t distributions

Confidence interval

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \tag{9.13}$$

upper confidence bound

$$\mu < \bar{x} \pm t_{\alpha,n-1} \frac{s}{\sqrt{n}} \tag{9.14}$$

lower confidence bound

$$\mu > \bar{x} \pm t_{\alpha,n-1} \frac{s}{\sqrt{n}} \tag{9.15}$$

sample size

$$n = \left(\frac{t'_{\alpha/2,n'-1} s}{ME} \right)^2 \tag{9.16}$$