

CHAPTER 3: WHO PLAYS VIDEO GAMES

math 189 : data analysis and inference : Winter 2018

Jelena Bradic

<http://www.jelenabradic.net>

Assistant Professor, Department of Mathematics, University of California, San Diego

jbradic@ucsd.edu

Introduction

The data

Background

Investigations

Theory

INTRODUCTION

- * Every year, 3,000 – 4,000 students enroll in statistics courses at UC Berkeley.
- * Half of these students take introductory statistics courses to satisfy quantitative reasoning requirement.
- * To aid the instruction of these students, a committee of faculty and students have designed a series of computer labs.

- * The labs are meant to extend the traditional syllabus for a course by providing an interactive learning environment that offers students an alternative method for learning the concepts of statistics and probability.
- * Some have linked labs to video games.
- * To help committee design the labs a survey of undergraduate students who were enrolled in a lower-division statistics course was conducted.
- * The survey's aim was to determine the extent to which the students play video games and which aspects of video games they find most and least fun.

- * Students who were enrolled in advanced statistics course conducted the study.
- * They developed the questionnaire, selected the students to be sampled and collected the data.
- * In this study you will have the opportunity to analyze the results from the sample survey to offer advice to the design committee.

SURVEY METHODOLOGY

What makes a survey a survey?

- Scientific methodology
- Data collection from an individual
- Usually samples from a large population
- Conducted for the purpose of...
 - > Description
 - > Exploration
 - > Explanation

main difference

⇒ population is finite

⇒ population needs to be
corrected

⇒ N ⇒ only goes to finite N
instead of infinity.

SURVEY METHODOLOGY (CONT.)

Characteristics of Good Survey Research

- Quantitative
- Self-monitoring
- Contemporary
- Replicable
- Systematic
- Impartial
- Representative *of population*
- Theory-based \Rightarrow *survey-methodology*
Survey-sampling

SURVEY METHODOLOGY (CONT.)

General Sampling Issues

- Basic rule—all individuals must have equal chance of being selected
- May be more accurate data than a census
- If all members of a population were identical, sampling would not be necessary
- Aim for a sample that is generalizable to total population of interest

confidence interval

randomness is important

Introduction

The data

Background

Investigations

Theory

DESCRIPTION

- * Out of 314 students in Statistics 2, Section 1, during Fall 1994, 95 were selected at random to participate in the survey.
- * Complete surveys were obtained from 91 out of 95 students.
- * The data available here are the students responses to the questionnaire.

The Survey asks students to identify how often they play video games and what they like and dislike about the games. Design a stats lab for 1994

↑ how to connect

- * The answers to these questions were coded numerically as described here.

length

categorical

- ** Time: # of hours played in the week prior to survey
- ** Like to play: 1=never played, 2=very much, 3=somewhat, 4=not really, 5=not at all.
- ** Where play: 1=arcade, 2=home system, 3=home computer, 4=arcade and either home computer or system, 5= home computer and system, 6=all three.
- ** How often: 1=daily, 2=weekly, 3=monthly, 4=semesterly. how are these variables
- ** Play if busy: 1=yes, 0=no. even having other things \Rightarrow yes \Rightarrow desirable \Rightarrow how much they like
- ** Playing educational: 1=yes, 0=no. belief
- ** Sex: 1=male, 0=female.
- ** Age: Students age in years.
- ** Computer at home: 1=yes, 0=no. whether they choose to go to arcade. \Rightarrow goes to somewhere to do certain tasks
- ** Hate math: 1=yes, 0=no.
- ** Work: # of hours worked the week prior to the survey. how long \rightarrow connected with time
- ** Own PC: 1=yes, 0=no.
- ** PS has CD-Rom: 1=yes, 0=no.
- ** Have email: 1=yes, 0=no. few yeses
- ** Grade expected: 4=A,3=B,2=C,1=D,0=F.

① Delete errors

② parallel \Rightarrow are you learning from the lab

hours work for other classes

proportionality

L \rightarrow best T \rightarrow goes up
good: bad \Rightarrow not fairly

SAMPLE OBSERVATIONS

Discrete Data

only 1 and 0 \Rightarrow bernoulli \Rightarrow corr/cov. \Rightarrow large sample cov $>$ true cov. (1)

time	like	where	freq	busy	educ	sex	age	home	math	work	own	cdrom	email	grade	Larger than
2	3	3	2	0	1	0	19	1	0	10	1	0	1	4	
0	3	3	3	0	0	0	18	1	1	0	1	1	1	2	
0	3	1	3	0	0	1	19	1	0	0	1	0	1	3	
0.5	3	3	3	0	1	0	19	1	0	0	1	0	1	3	
0	3	3	4	0	1	0	19	1	1	0	0	0	0	3	
0	3	2	4	0	0	1	19	0	0	12	0	0	0	3	
0	4	3	4	0	0	1	20	1	1	10	1	0	1	3	
0	3	3	4	0	0	0	19	1	0	13	0	0	1	3	
2	3	2	1	1	1	1	19	0	0	0	0	0	0	4	
0	3	3	4	0	1	1	19	1	1	0	1	0	1	4	
0	3	1	4	0	0	0	20	1	0	0	1	0	0	3	
0	3	2	4	0	0	0	19	1	0	0	1	0	1	4	
0	2	4	1	0	1	0	19	1	1	0	0	0	0	4	
3	3	3	2	1	0	0	18	0	0	0	0	0	0	3	
1	3	5	2	0	1	0	18	1	1	14	1	0	1	3	
.	

Snapshot of the data

cor \Rightarrow linear dependence on each other

a lot of 0's \Rightarrow not a signal that

they are related

\Rightarrow by randomness

\Rightarrow get a larger sample

positive bias

\hookrightarrow having binary random variables

(2)

Distribution \Rightarrow model for data
prob model \Rightarrow needs to be discrete

\rightarrow not gonna be i.i.d

MISSING DATA

If a question was not answered or improperly answered, then it was coded as 99.

Those respondents who had never played a video game or who did not at all like playing video games were asked to skip many of the questions.

The was a second part of the survey that covers whether the student likes or dislikes playing games and why.

These questions are different from the others in the more than one response may be given.

FOLLOW UP SURVEY

It
Extra Credit Related

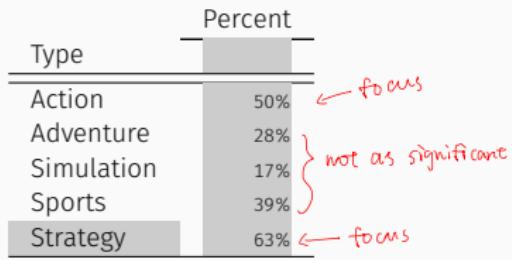


Table: What types of games do you play?
(at most three answers)

Table 1 summarizes the types of games played.

How can you mimic an action / strategy game?

- The student is asked to check all types that he or she plays. For example, 50% of the students responding to this question said that they play action games.
- Not all students responded to this question, in part because those who said they have never played a video game or do not at all like to play video games were instructed to skip this question.

FOLLOW UP SURVEY (CONT.)

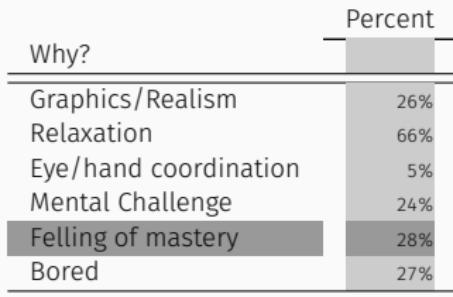


Table 2 summarizes reasons for playing the game.

Table: Why do you play the games you checked above? (at most three answers)

- Students who did answer this question were also asked to provide reasons why they play the games they do. They were asked to select up to three such reasons. Their responses are presented in Table 2

FOLLOW UP SURVEY (CONT.)

Dislikes	Percent
Too much time	48%
Frustrating	26%
Lonely	6%
Too many rules	19%
Costs too much	40%
Boring	17%
Friend's don't play	17%
It is pointless	33%

Table 3 summarizes what students didn't like about the games.

Table: What don't you like about video game playing? (at most three answers)

- Finally, table 3, contains summary of what the students do not like about video games. All students were asked to answer this question, and again they were asked to select up to three reasons for not liking video games.
- Third part of the survey collect general information about the student: age, sex, etc.

Introduction

The data

Background

The survey methodology

Video Games

Investigations

Theory

THE SURVEY METHODOLOGY

- broader idea
- proxy
- journal, journal article

generalize \Rightarrow doesn't include pure math
 \Rightarrow partly include P&S major

All of the population studied were undergraduates enrolled in Introductory Probability and Statistics, Section 1, during Fall 1994.

students who have interests in Business

- The class is a lower-division prerequisite for students intending to major in business
- During the Fall the class met on MWF from 1-2pm in a large lecture hall that seats four hundred.
- In addition to three hours of lecture, students attended a small, one-hour discussion section that met on T/Th.
- There were ten discussion sections for the class, each with approximately 30 students.
△ survey was given a week after the exam

THE SURVEY METHODOLOGY(CONT.)

The list of all students who had taken the second exam of the semester was used to select the students to be surveyed.

- The exam was given a week prior to the survey.
- To choose 95 students for the study, each student was assigned a number from 1 to 314. *only this section*
- (*not uniform*) - A pseudo random number generator selected 95 numbers between 1 to 314.
- To encourage honest responses, the students anonymity was preserved.

91 out of 95 completed the survey

THE SURVEY METHODOLOGY(CONT.)

To limit the number of nonrespondents, a three stage system of data collection was employed.

- Data collectors visited both the Tu and Th meetings of the discussion sections i noted week the survey was conducted.
- The students had taken an exam the week before the survey, and the graded exam papers were returned to them during the discussion section in the week of the survey.
- On Friday, those students who had not been reach during the discussion section were located during the lecture.
- A total of 91 students completed the survey. *out of 95 that are selected*
- To encourage accuracy in reporting, the data collectors were asked to briefly inform the student of the purpose of the survey and of the guarantee of anonymity.

VIDEO GAMES

Video Games can be classified according to the device on which they are played and according to the kind of skills needed to play the game.

coordination

	Eye/hand	Puzzle	Plot	Strategy	Rules
Action	x				
Adventure		x	x		
Simulation				x	x
Strategy				x	x
Role-play		x	x		x

Table 4 summarizes the attributes typically found in each category.

Table: Classification of five main types of video games

Device: arcade, console, PC
little strategy → more action → anything → role-play

Arcade games: fast and emphasize eye/hand coordination

Console games: action, adventure or strategy games

PC games: simulation and role-play exclusively and other types as well

learn, memorize

Introduction

The data

Background

Investigations

Investigations

Theory

The objective of this study is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab?

- ① CLT
- ② Finite sample correction
- ③ Bootstrap

Build Confidence Interval

Exploratory design

Confidence interval

Optional

- * [Scenario 1:] Begin by providing an estimate for the fraction of students who played a video game in the week prior to the survey. Provide an interval estimate as well as a point estimate for this proportion.
- * [Scenario 2:] Check to see how the amount of time spent playing vide games in the week prior to the survey compares to the reported frequency of play (daily, weekly, etc). How might the fact that there was an exam in the week prior to the survey affect your previous estimates and this comparison?
- * [Scenario 3:] Consider making an internal estimate for the average amount of time spent playing video games in the week prior to the survey. Keep in mind the overall shape of the sample distribution. A simulation study may help determine the appropriateness of an interval estimate.
- * [Scenario 4:] Next consider the "attitude" questions. In general, do you think the students enjoy playing video games? If you had to make a short list of the most important reasons why students like/dislike video games, what would you put on the list? Don't forget that those students who say that they have never played video games or do not at all like video games are asked to skip over some of these questions. So, there may be many nonrespondents to the questions as to whether they think video games are educational, where they play video games, etc.

The objective of this study is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab?

proportion

① estimate ③ different confidence interval

c1

three different
interval estimators

① WT

② Finite sample

③ Bootstrap

bootstrap

what does that
tell you?
which interval
is the best?

drawn from the
data.

- * [Scenario 1:] Begin by providing an estimate for the fraction of students who played a video game in the week prior to the survey. Provide an interval estimate as well as a point estimate for this proportion.
- * [Scenario 2:] Check to see how the amount of time spent playing video games in the week prior to the survey compares to the reported frequency of play (daily, weekly, etc). How might the fact that there was an exam in the week prior to the survey affect your previous estimates and this comparison? → negatively skewed?
- * [Scenario 3:] Consider making an interval estimate for the average amount of time spent playing video games in the week prior to the survey. Keep in mind the overall shape of the sample distribution. A simulation study may help determine the appropriateness of an interval estimate.
- * [Scenario 4:] Next consider the "attitude" questions. In general, do you think the students enjoy playing video games? If you had to make a short list of the most important reasons why students like/dislike video games, what would you put on the list? Don't forget that those students who say that they have never played video games or do not at all like video games are asked to skip over some of these questions. So, there may be many nonrespondents to the questions as to whether they think video games are educational, where they play video games, etc. why yes/no? statistical variable selection

The objective of this study is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab?

Compare two distributions → save tools as HW!

- * [Scenario 5:] Look for the differences between those who like to play video games and those who don't. To do this, use the questions in the last part of the survey, and make comparisons between male and female students, those who work for pay and those who don't, those who own a computer and those who don't. Graphical display and cross-tabulations are particularly helpful in making these kinds of comparisons. Also, you may want to collapse the range of responses to a question down to two or three possibilities before making these comparisons.

represent
data in
tables

2x2 max
3x3

Table ⇒ best way to represent



The objective of this study is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab?

Summary
of HWI

- * [Scenario 5:] Look for the differences between those who like to play video games and those who don't. To do this, use the questions in the last part of the survey, and make comparisons between male and female students, those who work for pay and those who don't, those who own a computer and those who don't. Graphical display and cross-tabulations are particularly helpful in making these kinds of comparisons. Also, you may want to collapse the range of responses to a question down to two or three possibilities before making these comparisons.
- * [Scenario 6:] Just for fun, further investigate the grade that students expect in the course. How will it match the target distribution used in grade assignment of 20% A's, 30% B's, 40% C's and 10% D's or lower? If the nonrespondents were failing students who no longer bothered to come to the discussion section, would this change the picture ?

Optional

[EC]

reason why

Introduction

take data

from
this
finite
population

Background
is for randomness

Investigations

probability

Theory

Goals

The Probability Model

Sample Statistics

Estimators for Standard Error

Population Total and Percentages

Normal Approximation and Confidence Intervals

An Example

An Example

The Bootstrap

chances
of
being in
the
sample

Design Stats Lab



being in the
sample
 \Rightarrow dependent

independence.

knowing info for one sample
doesn't change the chance of
other sample being selected in
the data

once you know
one is in the sample,
the chances of knowing
others are in is
reduced

because finite
population
dependence
 \hookrightarrow sample of
finite population

depends on
the probability
model, how
good is the estimate



- ① how to estimate, quality of the estimate
- ② CI, why they look
- ③ how to correct them for finite population
- ④ simulations bootstrap \leftarrow data drawn from data
 \uparrow
finite population



Survey data \Rightarrow independence?

\hookrightarrow simple, random sample?
each column doesn't change other column \leftarrow independent
As a collection of individual
 \hookrightarrow independent from each other?

In this section we will use as our primary example the problem of estimating the average amount of time students in the class spent playing video games in the week prior to the survey.

- * To determine **the exact amount of time for the entire class** we would need to interview all of the students (over 3000 of them).
- * Alternatively, a subset of them can be interviewed, and the information collected from this subset could provide an approximation to the full group.
 - * In this section we discuss one rule for selecting a subset of student to be surveyed, the **simple random sample**.
 - * **The simple random sample** is a probability method for selecting the students..
 - * Probability methods are useful because through chance we can make useful statements about the relation between the sample and the entire group.
 - * With a probability method, we know the chance of each possible sample.

Terminology

- * **Population units** make up the group that we want to know more about
 - * In this lab, the units are the students enrolled in the 1994 Fall semester class of Introductory Probability and Statistics.
- * **Population size**, usually denoted by N , is the total number of units in the population. For very large population, often the exact size of the population is not known. Here we have 314 students in the class.
- * **Unit characteristic** is a particular piece of information about each member of the population.
 - * The characteristic that interests us in our example is the amount of time the student played video games in the week prior to the survey.
- * **Population parameter** is a summary of the characteristic for all units in the population, such as the average value of the characteristic.
 - * The population parameter of interests to us here is the average amount of time students in the class spent playing video games in the week prior to the survey.

GOALS (CONT.)

Terminology

- * **Sample units** are those members of the population selected for the sample.
- * **Sample size** usually denoted by n , is the number of units chosen for the sample. We will use 91 for our sample size, and ignore the four who did not respond.
- * **Sample statistic** is a numerical summary of the characteristic of the units sampled. The statistic estimated the population parameter. Since the population parameter in our example is the average time spent playing video games by all students in the class in the week prior to the survey, a reasonable sample statistic is the average time spent playing video games by 11 students in the sample.

The simple random sample is a very simple probability model for assigning probabilities to all samples of size n from a population of size N .

on finite population

$$N = 314$$

$$n = 91$$

In general with N population units and a sample of size n , there are N choose n possible samples. The probability rule that defined the simple random sample is that each one of the $\binom{N}{n}$ samples is equally likely to be selected. That is, each unique sample of n units has the same chance, $1/\binom{N}{n}$, of being selected. From this probability, we can make statements about the variations we would expect to see across repeated samples.

of people participated in the survey

N - population size
 n - sample size

P (each of us part of sample
= $\frac{1}{\binom{N}{n}}$)

- Assign each unit a number from 1 to N .
- Write each number on a ticket, put all of the tickets in the box, mix them up,
- Draw n tickets one at a time from the box without replacement.

or draw simultaneously

NUMBER OF DIFFERENT UNITS IN THE SAMPLE

$P(\text{observed data})$

Unit # 1

- * The chance that the unit # 1 is the first to be selected for the sample is $1/N$.
- * Likewise, the unit # 1 has chance $1/N$ of being the second unit chosen for the sample.
- * All together, the unit #1 has the chance of n/N of being in the sample.

order matters

overall
data
↑
population

Unit # 2

- * By symmetry, The chance that the unit # 2 is the first to be selected for the sample is $1/N$.
- * Likewise, the unit # 2 has chance $1/N$ of being the second unit chosen for the sample.
- * All together, the unit #2 has the chance of n/N of being in the sample.

Each unit has the same chance of being in the sample.

DEPENDENCE

$$\max(P(\text{observes})) \quad \text{maximum likelihood}$$

Pr(observing what I want to observe)

There is dependence between selections.

- * The chance that the unit # 1 is chosen first and the unit # 2 is chosen second is $\frac{1}{N(N-1)}$. *(conditional probability)*
- * This chance is the same for any two units in the population, hence the chance that # 1 and # 2 are both in the sample, is $\frac{n(n-1)}{N(N-1)}$.
- * In our example

$$\mathbb{P}(\text{unit } \# 1 \text{ in the sample}) = 91/314,$$

$$\mathbb{P}(\text{unit } \# 1 \text{ and unit } \# 2 \text{ are in the sample}) = \frac{91 \times 90}{314 \times 313}. \quad \text{joint event (independent)}$$

DEPENDENCE

Discrete

Probability distribution of the units chosen in the sample

unit number

- * Let $I(1), I(2), \dots$ represent the first, second, ... number drawn from the list $1, 2, \dots, N$.
- * Then,

$$\mathbb{P}(I(1) = 1) = 1/N$$

$$\mathbb{P}(I(1) = 1 \text{ and } I(2) = N) = \frac{1}{N(N-1)}$$

and in general, for $1 \leq j_1 \neq j_2 \neq \dots \neq j_n \leq N$

$$\mathbb{P}(I(1) = j_1, I(2) = j_2, \dots, I(n) = j_n) = \frac{1}{N(N-1)\dots(N-n+1)}$$



Why is this equation true?

Explain in the report

The simple random sample method puts a probability structure on the sample. Different samples have different characteristic and different sample statistics.

.....

Sample statistic has a probability distribution related to the sampling procedure.

SAMPLE STATISTICS

Computing expected value of the sample statistic.....

- * Let x_1 be the value of the characteristic for unit # 1. x_2 for unit #2, ...
- * In our example x_i is the time spent playing video games by the student # i: $i = 1, \dots, 314$.
- * Population average is

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

for each unit

as our population parameter.

- * Let $\underline{x_{I(j)}}$ represent the value of the characteristic for the j-th unit sampled.
unit that is selected

Note: $x_{I(1)}$ is a random variable \rightarrow because j is randomly selected to be $I(1)$

- * In our example $x_{I(j)}$ represent the value of the time spent playing video games by the j-th unit sampled: $j = 1, \dots, 91$.

allowed to

$$\mathbb{E}(x_{I(j)}) = \sum_{i=1}^N x_i \mathbb{P}(I(j) = i) = \sum_{i=1}^N x_i \frac{1}{N} = \mu$$

i is selected to be in the j th position
Anytime in the population

SAMPLE STATISTICS (CONT.)

- * The sample average is the sample statistic that estimates the population parameter μ ,

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_{I(j)}.$$

round.
var.
because
is rand. var.

can just use

$$E[\bar{x}] = \sum E[x_i]$$

Note: \bar{x} is a random variable

- * $E(\bar{x}) = \mu$

check if this is
a good estimator → unbiased

Note: We have shown that the sample average is an **unbiased estimator** of the population parameter

- * Next we find the standard deviation of \bar{x} . To do this, we first find the variance of $x_{I(j)}$

T_{rand. var.} I_(j)
the value can be selected

$$\text{Var}(x_{I(j)}) = \mathbb{E} (x_{I(j)} - \mu)^2 \stackrel{?}{=} \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \sigma^2$$

first unit
in the population is the j-th position in
the sample

where we used σ^2 to denote **population variance**.

SAMPLE STATISTICS (CONT.)

Variance
Definition & Property

Computing variance of the sample statistic....

Use Slides

- * Then we compute the variance of the sample average \bar{x} as follows

$$\begin{aligned} \text{Var}(\bar{x}) &= \frac{1}{n^2} \text{Var} \left(\sum_{j=1}^n x_{I(j)} \right) && \text{Cov}(x_{I(j)}, x_{I(k)}) \\ &= \frac{1}{n^2} \sum_{j=1}^n \text{Var}(x_{I(j)}) + \frac{1}{n^2} \sum_{j=1, j \neq k}^n \text{Cov}(x_{I(j)}, x_{I(k)}) && (2) \\ &= \frac{1}{n} \sigma^2 + \frac{n-1}{n} \text{Cov}(x_{I(1)}, x_{I(2)}) && (3) \end{aligned}$$

is independence from $j \rightarrow k$ (1)

Explain

The last equality follows from noting that all pairs $(x_{I(j)}, x_{I(k)})$ are identically distributed. The covariance between any two sampled units $x_{I(j)}$ and $x_{I(k)}$ is not 0 because the sampling procedure makes them dependent.

$$\text{Cov}(x_{I(1)}, x_{I(2)}) = -\frac{\sigma^2}{N-1}$$

SAMPLE STATISTICS (CONT.)

finite population
correction factor

$$\text{Var}(\bar{x}) = \frac{1}{n} \sigma^2 \frac{N-n}{N-1}, \quad \text{SD}(\bar{x}) = \frac{1}{\sqrt{n}} \sigma \sqrt{\frac{N-n}{N-1}}$$

The factor

$$\frac{N-n}{N-1} = 1 - \frac{n-1}{N-1} \sim 1 - \frac{n}{N}$$

← sampling fraction

in the variance and SD is called the finite population correction factor. The ratio n/N is called the sampling fraction. It is very small when the sample size is small relative to the population size.

If $n=N \Rightarrow \text{Var}(\bar{x})=0 \Rightarrow$ no randomness in the data

This is frequently the case in sampling, and when this happens

$$\text{Var}(\bar{x}) \sim \sigma^2/n$$

and the finite population correction factor is often ignored.

In our example, it cannot be ignored as

$$\frac{\sqrt{314-91}}{\sqrt{314-1}} = 0.84$$

* needs the randomness to analyze the data

VARIANCE....

Notice that without the correction factor, the variance is the same as if we made the draws with replacement (or sampling from an infinite population)

With a simple random sample, the standard deviation for the estimator can be computed in advance, dependent on the population variance σ^2 . If σ^2 is known approximately, then the sample size can be chosen to give an acceptable level of accuracy for the estimator.

Often a pilot study, results from a related study, or a worst-case estimate of σ^2 is used in planning the sample size for the survey.

ESTIMATORS FOR STANDARD ERRORS

Standard deviations for estimators are typically called standard errors (SEs)

- . They indicate the size of the deviation of the estimator from its expectation.
- * When σ^2 is unknown, a common estimator for it is

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

- * To estimate $\text{Var}(\bar{x})$, we can then use

$$\frac{s^2}{n} \frac{N-n}{N-1}$$

- * The reason for using s^2 is that the sample, when chosen by the simple random sample method, should look roughly like a small -scaled version of the population, so we plug in s^2 for σ^2 in the variance of \bar{x} .

ESTIMATORS FOR STANDARD ERRORS (CONT.)

- * In fact we can make a slightly better estimate for $\text{Var}(\bar{x})$
- * To estimate $\text{Var}(\bar{x})$, we can then use

$$\mathbb{E}s^2 = \frac{N}{N-1}\sigma^2$$

- * Hence, an unbiased estimator of σ^2 is then

$$s^2 \frac{N-1}{N}$$

plug in $\text{Var}(\bar{x}) = \frac{1}{n}s^2 \frac{N-n}{N-1}$

- * Hence an unbiased estimator of $\text{Var}(\bar{x})$ is

$$CI: \left(\bar{x} \pm z_{\alpha/2} \sqrt{s^2 \frac{N-n}{n}} \right)$$

Note: There is essentially no difference between these two estimators of $\text{Var}(\bar{x})$ for any reasonably sized population.

POPULATION TOTALS AND PERCENTAGES

Sometimes the population parameter is a proportion or percentage, such as the proportion of students who played a video game in the week prior to the survey or the percentage of students who own PCs.

POPULATION TOTALS AND PERCENTAGES

fraction

Sometimes the population parameter is a proportion or percentage, such as the proportion of students who played a video game in the week prior to the survey or the percentage of students who own PCs.

When the parameter is a proportion, it makes sense for the characteristic value x_i to be 0 or 1 to denote the absence or presence of the characteristic, respectively. For example, for $i = 1, \dots, 314$

discrete $x_i = \begin{cases} 1 & \text{if the } i\text{th student in the population owns a PC} \\ 0 & \text{otherwise} \end{cases}$ (4)

↳ value ↑ is different

POPULATION TOTALS AND PERCENTAGES

Then $\tau = \sum x_i$ counts all of the students who own PCs in the population, and

$$\pi = \frac{1}{N} \sum x_i$$

is the proportion of students in the population who owns PCs.

In this case \bar{x} remains an unbiased estimate of π , **the population average**, and $N\bar{x}$ estimates τ .

A simpler form for the population variance and the unbiased estimator of $\text{Var}(\bar{x})$ can be obtained because

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \pi)^2 \stackrel{\text{Bernoulli}}{\longrightarrow} \stackrel{\text{Binomial}}{\longrightarrow} \pi(1 - \pi).$$

Then the **estimator for the standard error** is

$$\hat{SE}(\bar{x}) = \frac{\sqrt{\bar{x}(1 - \bar{x})}}{\sqrt{n - 1}} \frac{\sqrt{N - n}}{\sqrt{N}}$$

*s and Var are
slightly different*

POPULATION TOTALS AND PERCENTAGES

Often the symbols $\hat{\mu}$, $\hat{\pi}$ and $\hat{\tau}$ are used in place of \bar{x} , $N\bar{x}$ to denote sample estimates of the parameters μ , π and τ . The following table contains the expectations and standard errors for estimators of a population average, proportion and total.

Confidence intervals are above μ

	Average	Proportion	Total
Parameter	μ	π	τ
Estimator	\bar{x}	\bar{x}	$N\bar{x}$
Expectation	μ	π	τ
Standard Error	$\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$	$\sqrt{\frac{\pi(1-\pi)}{n}} \sqrt{\frac{N-n}{N-1}}$	$N \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
Estimator of SE	$\frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$	$\frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n-1}} \sqrt{\frac{N-n}{N}}$	$N \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$

Determine if CI is good

If D is normal

- (1) correct
- (2) smallest width

test for normality

① point Estimate of μ : \bar{x} (unbiased)

② Distribution of point estimate of \bar{x} ? \Rightarrow Distribution of \bar{x} ? \Rightarrow CI: around point estimate \bar{x}

Table: Properties of sample statistics

Assumptions

- ① if $\frac{N}{n} \rightarrow \infty$: CLT: $D \sim N(\mu, \frac{\sigma^2(N-n)}{n(N-1)})$
- ② $n < \infty$: Bootstrap \Rightarrow whether \bar{x} resemble normal
- ③ if $N = \infty$, if X_i , iid $\Rightarrow n \rightarrow \infty$

If the sample size is large, then the probability distribution of the sample average is often well approximated by the normal curve.

Central Limit Theorem

If X_1, \dots, X_n are independent, identically distributed with mean μ and variance σ^2 then, for n large, the probability distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is approximately standard normal

NORMAL APPROXIMATION

Assumption : Finite IID

If the sample size is large, then the probability distribution of the sample average is often well approximated by the normal curve.

Central Limit Theorem

* our data is not independent

If X_1, \dots, X_n are independent, identically distributed with mean μ and variance σ^2 then, for n large, the probability distribution of

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

$n \rightarrow \infty$
 $N \rightarrow \infty$
 $\frac{n}{N} \rightarrow 0$

is approximately standard normal

In simple random sampling, the $x_{(j)}$ are identically distributed but nonindependent. However, the normal approximation can still hold if, in addition to the sample size being large, it is not too large relative to the population size.

applies to
our survey data
* simple sampled data

IS IT NORMAL?

- * If the sampling proportion n/N is small, then the $x_{l(j)}$ are nearly independent.
- * There are no hard rules for how large n must be or how small n/N must be before we can use the normal approximation.
- * You can use simulation to check it.
- * The central limit theorem is very powerful result. It implies that for any population distribution, under simple random sampling (for appropriate n and n/N), the sample average has an approximate normal distribution.

CONFIDENCE INTERVALS

Normal distribution can be used to provide confidence intervals for the population parameter. One **interval estimate** of μ , called **68% confidence interval** is

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}}, \bar{x} + \frac{\sigma}{\sqrt{n}} \right)$$

A **95% confidence interval** is

$$\left(\bar{x} - 2 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \frac{\sigma}{\sqrt{n}} \right)$$

Interpretation

By Central Limit Theorem the chance that \bar{x} is within one (or two) standard error(s) of μ is approximately 68% (or 95%).

Note: Sample statistic \bar{x} is random, so we can think of confidence intervals as random intervals.

Different samples lead to different confidence intervals. If we take many simple random samples where for each sample we compute CI, then we expect 95% of CI's to contain μ .

CONFIDENCE INTERVALS

That is, ignoring the finite sample population correction factor,

$$\mathbb{P}\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) \quad (5)$$

$$= \mathbb{P}\left(2 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 2\right) \sim 0.95 \quad (6)$$

In practice, we often don't know σ , and we substitute s in place of σ in order to make the confidence interval. With this substitution we often call the interval, [approximate confidence interval](#).

AN EXAMPLE

To summarize the ideas introduced in this section, consider the problem of estimating the proportion of females in the class.

In this case we know that there are 131 females in the class. Therefore, we have all of the population information

$$\pi = 131/314, \quad \sigma^2 = 0.2431, \quad N = 314$$

Because we have simple random sample, the probability distribution for $x_{I(j)}$ matches the population distribution, i.e.

$$\mathbb{P}(x_{I(j)} = 1) = 131/314 = 0.4172, \quad \mathbb{P}(x_{I(j)} = 0) = 183/314 = 0.5818.$$

This means that

$$\mathbb{E}(x_{I(j)}) = 0.4172, \quad \text{Var}(x_{I(j)}) = 0.2431, \quad \mathbb{E}(\bar{x}) = 0.4172$$

$$SE(\bar{x}) = \sqrt{\frac{0.2431}{91} \times \frac{223}{313}} = 0.044$$

AN EXAMPLE (CONT.)

The exact distribution of \bar{x} can be found:

$$\mathbb{P}(\bar{x} = m/91) = \mathbb{P}(\text{the sample has } m \text{ females}) = \frac{\binom{131}{m} \binom{183}{91-m}}{\binom{314}{91}}$$

This is known as [hypergeometric distribution](#).

In a real sampling problem, the exact distribution of \bar{x} is not known.
However, it is known

- the distribution of $x_{(j)}$ matches the propagation distribution
- \bar{x} is the unbiased estimator of the population parameter
- provided n is large and n/N small, the distribution of \bar{x} is roughly normal

This is enough to construct confidence intervals.

AN EXAMPLE (CONT.)

In this example n is large but $n/N = 91/314$ is not small. We will use bootstrap to check if we can use normal approximation.

In our example, 38 out of 91 students were females.

- Our estimate for the population parameter is $\bar{x} = 38/91 = 0.42$
- Our estimate of the standard error is

$$\frac{\sqrt{\frac{38}{91}(1 - \frac{38}{91})}}{\sqrt{91 - 1}} \times \frac{\sqrt{314 - 91}}{\sqrt{314}} = 0.044$$

- Hence, our CI is $(0.33, 0.51)$

AN EXAMPLE (CONT.)

Finally, in this example. the actual proportion of women in the sample was very close to the expected proportion of women, $\mathbb{E}(\bar{x})$.

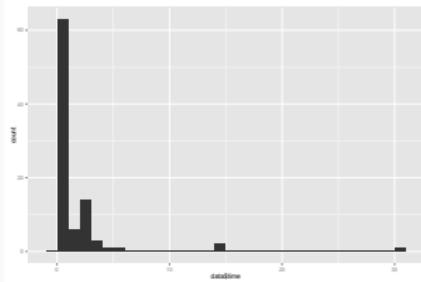
- We may want to calculate the probability that we would get as close or closer to the expected proportion.
- Sometimes, samples seem to be too close to what we expect: if the chance of getting data as close or closer to the expected value is small, say 1/100, then we might suspect the sampling procedure.
 - Expected number of women in the sample is $91 \times \pi = 37.97$
 -

$$\mathbb{P}(\text{exactly 38 of 91 students were women}) = \frac{\binom{131}{38} \binom{183}{91-38}}{\binom{314}{91}} = 0.10$$

THE BOOTSTRAP

From the histogram of the time spent playing video games by the students in the sample, we see that the sample distribution is extremely skewed.

This observation raises a question of whether the probability distribution of the sample average follows normal curve.



- * Without knowledge of the population, we cannot answer this question completely.
- * Bootstrap can help.

THE BOOTSTRAP (CONT.)

Bootstrap algorithms are simple and general tools for

- (a) assessing estimators' accuracy via variance estimation, and
- (b) producing confidence intervals and p-values.

Bootstrap applies to finite samples and provides numerical solutions for non-standard situations so that it is particularly appealing when dealing with finite populations and complex sampling designs

Bootstrap for survey data

THE BOOTSTRAP (CONT.)

According to the simple random sample probability model, the distribution of the sample should look roughly similar to that of the population.

We could create a new population of 314 based on the sample and use this population, which we call the [the bootstrap population](#), to find the probability distribution of the sample average.

The following table helps:

Time	Count	Bootstrap Population
0	57	197
0.1	1	3
0.5	5	17
1	5	17
1.5	1	4
2	14	48
3	3	11
4	1	3
5	1	4
14	2	7
30	1	3

THE BOOTSTRAP (CONT.)

Efron



Step 1: generate population
Step 2: draw sample

For every unit in the sample, we make $314/91 = 3.45$ units in the bootstrap population with the same time value and round off to the nearest integer.

Next, to determine the probability distribution of the sample average when the sample is taken from the bootstrap population, we use a computer.

- * Select a simple random sample of 91 from the bootstrap population, called a bootstrap sample, and take its average.
- * Then we take another sample of 91 and take its average.
- * A histogram of bootstrap sample averages, each constructed from a simple random sample of 91 from the bootstrap population, appears in the Figure below. $(\frac{N}{n})$

non-integer



3.4

use different ways to process

Step 1: replicate each data point 3.4 times

→ use different ways to process

Step 2: structure of sample 1 \rightarrow structure of original survey sample

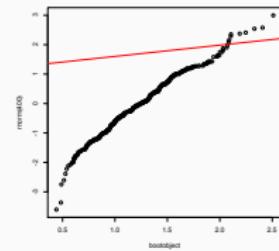
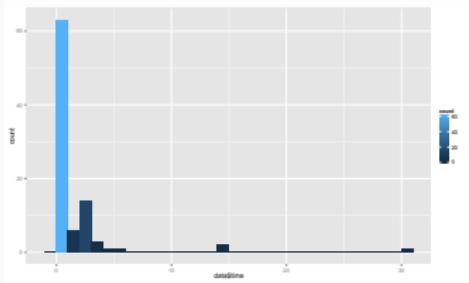
For example: survey sample \leftarrow simple Random Sample

survey \leftarrow without replacement

THE BOOTSTRAP (CONT.)

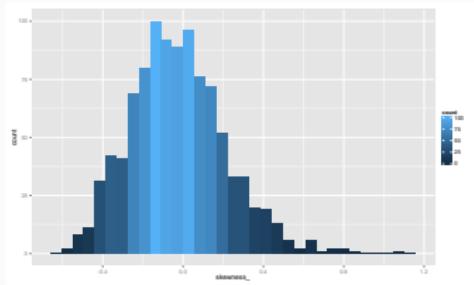
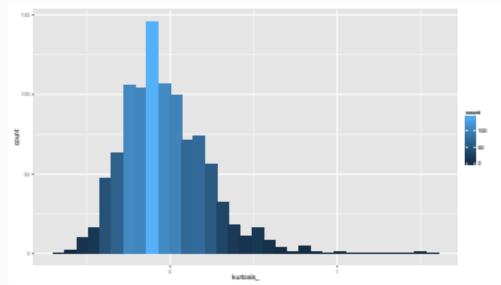
We took 400 bootstrap samples from the bootstrap population, in order to make a reasonable simulation of a probability distribution of the bootstrap average.

From the qq plot we see it to be close to normal distribution



THE BOOTSTRAP (CONT.)

- * To further validate this claim we can compare the kurtosis - 2.67 and skewness - 0.19 for the 400 bootstrap sample averages to a simulated distribution of skewness and kurtosis for samples of size 400 from a normal distribution
- * See figure bellow.



THE BOOTSTRAP (CONT.)

The method described here is one version of the bootstrap.

The bootstrap technique derives its name from the expression "to pull yourself up by your own bootstraps",

In the sampling context, we study the relation between bootstrap samples and the bootstrap population, where both the samples and the population are known, in order to learn about the relationship between our actual sample and the population, where the latter is unknown.

BOOTSTRAP (CONT.)-R CODE FOR THE PLOTS ABOVE

```
bootobject= NULL
for ( i in 1:400)
{
bootobject[i]=mean(sample(as.vector(data$time),size=91,replace=TRUE))
}
m=qplot(bootobject, geom="histogram")
m + geom_histogram(aes(fill = ..count..))
require(e1071)
kurtosis_=NULL
for (i in 1:1000)
{
kurtosis_[i]=kurtosis(rnorm(400))
}
m=qplot(kurtosis_, geom="histogram")
m + geom_histogram(aes(fill = ..count..))
skewness_=NULL
for (i in 1:1000)
{
skewness_[i]=kurtosis(rnorm(400))
}
m=qplot(skewness_, geom="histogram")
m + geom_histogram(aes(fill = ..count..))
```

The single-elimination jackknife re-sampling technique

- * Somewhat similar to the bootstrap re-sampling technique
- * Do experiment once (here: 1000 measurements)
- * Make 1000 samples of 999 measurements each ? with the i-th measurement eliminated in the i-th sample
- * Compute the mean (or the wanted quantity) of each sample

Contrasts:

- * The bootstrap method handles skewed distributions better
- * The jackknife method is suitable for smaller original data samples