Chapter 3
Uncover Pattern
Blurred by Sampling Instability

Section 3.1
From a Sample to the Population

David W. Gerbing

The School of Business
Portland State University

3.1a
Samples vs Populations

Problem: All the Data of Interest are Not Available

Analysis is only of some of the observations of interest

- For a study of the number of employees who call in sick on a Friday before a holiday Monday, the data may not exist for many years in the past, and regardless, cannot yet exist for future such Fridays
- A study of the blood chemistry of those who have Type II Diabetes cannot examine all people who have had, who do have, and who will have Type II Diabetes
- The length of time to complete a specific procedure is of interest, but the times of past procedures may not have been recorded, and cannot have been recorded for future instances of the procedure

Populations and Samples

Want the entire population, but typically only get one sample

- **Population**: Set of all existing or potential observations
- The entire population is generally not available
- **Sample**: A subset of the population, the data for a specific analysis
- **Key Concept**: To collect data, randomly gather a subset of the population, and generalize results to the *entire* population

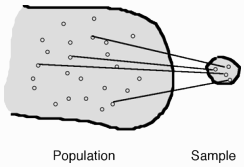


Figure: A population defined by a process that generates the data values, and a sample from the process, where each dot represents a single observation

The Population is of Primary Interest

Generalize beyond the usual one sample of data

- The population contains the desired information
  - May consist of a usually large number of fixed elements in a given location at a given time, such as all registered voters
  - More generally, consists of outcomes of a process ongoing over time, in which case the population and its associated values such as its mean are hypothetical
- **Population Value**: True value based on the *entire* population, such as the population mean
- A population value is not known directly because all the values of the population are not known
- **Key Concept**: A population value is an abstraction, considered real, but not observed, with an unknown value
- **Key Concept**: *Estimate* a population value from sample data

Descriptive Statistics

Analysis of the data from samples

- **Descriptive statistics**: Summarize and display aspects of the sample data drawn from the larger population
- Descriptive statistics are also referred to as summary statistics
- All of the statistics discussed until this point are descriptive statistics, calculated directly from data
  - Mean and standard deviation
  - Median, range, IQR, quartiles and quantiles
- **Key Concept**: *Calculate* the value of a descriptive statistic from sample data
  - Ex: Calculate the median of the class midterm
  - Ex: Calculate the mean ship time from a supplier
- The estimation of unknown population values follows from the calculation of the relevant descriptive statistics

Random Samples

Obtaining the sample

- How are the elements in the sample selected from the larger population?
- **Random Sample**: A sample in which every value of the population has an equal probability of selection
- To select a random sample requires access to randomly generated numbers, today usually accomplished with a computer application such as R (next slide) or Excel
- A random sample is difficult to implement completely, but the essence of randomness is essential to properly generalizing results to the population

Generate a Random Sample

The computer provides the random selection

- Use the R function `sample` to select a sample from a larger set of values
  - Specify the source of these values, numerical or not, and the number of values in the sample, the sample `size`
  - By default, `replace=FALSE`, so sampling is done without replacement, each value in the population can appear only once in the resulting sample
- From the first 100 integers, randomly sample 8 integers

```
> sample(1:100, size=8)
```

- Read Employee data into data frame `d`, draw a sample of 3 people from `d` based on the row names of `d`

```
> d <- Read("Employee")
> sample(row.names(d), size=3)
```

The Sampling Frame

Sampling frame vs. Population of interest

- Distinguish between what is wanted vs what is obtained
- **Sampling Frame**: The *actual* population from which the sample is drawn, distinguished from the *desired* population
- **Key Concept**: The results of an analysis can only be properly generalized to the sampling frame
- The sampling frame determines the scope of generalization of results, not the desired population of interest per se
- The sampling frame should be the population of interest, but sometimes the population actually sampled is not the population that was desired

Generalizing Results from a Sample

An example

- Suppose a researcher conducts a market research survey
- Define the population of interest as all city residents, then
  - Draw a random sample of people listed in the phone book
  - Collect data by calling the people from 9am to 5pm
- The results of this analysis can only be properly generalized to people . . .
  - with a phone
  - with a listed phone number in the phone book
  - who answer all phone calls
  - are available during the daytime hours
- The sampling frame is *not* the population of interest, all city residents, so these survey results cannot be properly generalized to all city residents

3.1b
Sampling Fluctuations

Sampling Fluctuations

The sample is only the starting point of statistical analysis

- Statistics such as the sample mean, m, provide a summary of this distribution of values
- The specific values in a sample differ from sample to sample
- Accordingly, the value of a sample statistic of interest, such as m, *arbitrarily* varies from sample to sample
- Each sample outcome, such as m, is an arbitrary result, which only hints at the true, underlying population value
- **Key Concept**: Typically, only one sample is taken and so only one m is observed, but the following reality is the basic motivating concern addressed by statistical inference

> IF many samples were taken, then a different value of m WOULD BE observed for each sample

Sampling Fluctuations with a Sample Size of $n = 10$

To illustrate, flip a fair coin 10 times

- Encode the outcome of each flip as a value of the variable Y
- Score each Head as Y=1 and each Tail as Y=0
- Calculate the sample mean

$$m = \frac{\sum Y_i}{n}$$

Now get 1 sample of 10 coin flips!

Figure: Get 5 heads.

Result

$$m = \frac{0+0+1+1+0+1+1+0+0+1}{10} = .5$$

Gather Another Sample of Size $n = 10$

Flip *same* fair coin another 10 times

- Again, score Y=1 for a Head and Y=0 for a Tail
- Again, calculate m for $n = 10$

Now get 1 sample of 10 coin flips!

Figure: Get 6 heads.

Result

$$m = \frac{1+1+1+0+0+1+1+0+1+0}{10} = .6$$

- One sample of ten flips yields $m = .5$ and another sample yields $m = .6$

Outcomes of Repeated Samples from Same Population

Variation of sampling results applies to all sample data

- Mean length of time to complete a procedure
  - Sample 1: $m = 14.53$ minutes
  - Sample 2: $m = 13.68$ minutes
- **Key Concept**: The value of a statistic, such as the sample mean, m, randomly varies from sample to sample

---

Why? Sampling variation of a statistic results from random variation of the data values from random sample to random sample

---

- A descriptive statistic only describes a sample, but is limited in its generality to describe the corresponding population
- The focus of statistical analysis is on the stable population values

3.1c
Better Estimates from More Information
The Law of Large Numbers

Population Mean μ is an Abstraction

Calculate sample data values, estimate population values

- To understand reality, management wants to know the value of the population mean, μ
  - The difficulty is that μ is an abstraction, a hypothetical, never directly calculated
  - To estimate the unknown requires information from which to provide the basis for the estimate
- Only one kind of information is considered in traditional statistical analysis
- **Classical or frequentist model** of statistics[1]: Obtain knowledge of a population value, such as μ, *only* from data
- **Key Concept**: More data, all randomly sampled from the *same* process – generally provides better estimation

[1] The primary competing model is the Bayesian model, in which other types of information are also considered in the estimation of a population value

Illustrate the Benefit of Larger Samples

Plot a mean to compare over many sample sizes

- Consider a coin flip in which the truth is that the coin is fair
  - Collect the data: Flip a coin a specified number of times and record if a Head or a Tail after each flip
  - Create a variable Y to indicate the number of Heads obtained
  - Score each Head as Y=1 and each Tail as Y=0
  - Over the *entire* population, one-half of all flips result in a 1 and one-half of all flips result in a 0
  - Compute the true value: $\mu = \frac{1}{2}(1) + \frac{1}{2}(0) = 0.5$
  - **Key Concept**: Can data analysis reveal this reality of a fair coin?
- **Running Mean**: Re-calculate the sample mean, m, after each new data value is collected
- Resulting plot of the running mean shows the value of the sample mean, m, as the sample size increases

Simulate Repeated Coin Flips

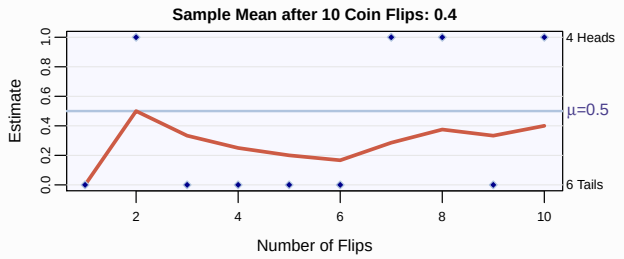
The computer provides the data *as if* we have many coin flips

- **Key Concept**: How close is m, based on data, to the truth of $\mu = 0.5$ as the sample size increases?
- Although in practice: Only one sample with one sample size ...
- Computer simulation allows us to explore what happens for different sample sizes, here different numbers of coin flips
- Obtain a different result each time the simulation is run

- `lessR` function `simFlips` plots the running mean as sample size increases as if repeated coin flips with the same coin
  - *Required*: `n`, maximum sample size, the number of flips
  - *Default*: `prob=.5`, probability of a Head, coin is fair
  - *Optional*: `?simFlips`, e.g., `pause=TRUE` for each flip

```
> simFlips(n=10)
```

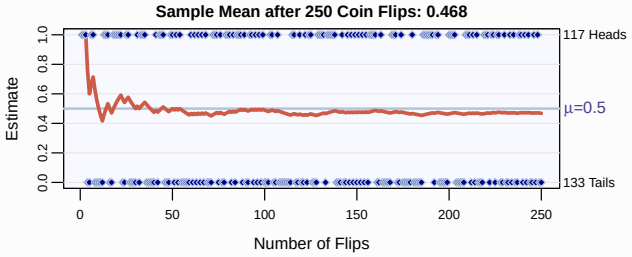

m estimates μ from 1 to 10 Coin Flips



Estimate for $\mu = 0.5$: After 10 coin flips,

$$m = \frac{\sum Y_i}{n} = \frac{\text{Number of Heads}}{\text{Number of Flips}} = \frac{4}{10} = 0.4$$

m estimates μ from 1 to 250 Coin Flips

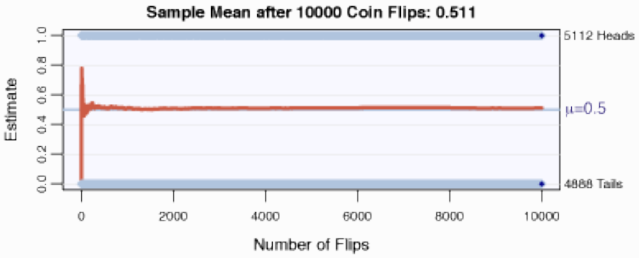


m is relatively unstable for the first 50 or so coin flips ...
and then the estimate nicely settles down after 200 or so flips

Estimate for $\mu = 0.5$: After 250 coin flips,

$$m = \frac{\sum Y_i}{n} = \frac{\text{Number of Heads}}{\text{Number of Flips}} = \frac{117}{250} = 0.468$$

m estimates μ from 1 to 10000 Coin Flips



Even after 10000 coin flips ...
the true value of μ is not attained, but a closer estimate resulted

Estimate for $\mu = 0.5$: After 10000 coin flips,

$$m = \frac{\sum Y_i}{n} = \frac{\text{Number of Heads}}{\text{Number of Flips}} = \frac{5112}{10000} = 0.511$$

More Information, Better Estimation

How close is the estimate to the actual population value?

- **Key Concept**: Generalize information obtained from a sample, such as the sample mean, to the population as a whole
- Sample all data values from the same population so that a common μ and σ underlie each data value Y_i
- **Law of Large Numbers**: The larger the random sample of data values all from the same population, the closer the estimate *tends* to be to the underlying population value
  - This result of large samples is only a tendency
  - There is no guarantee that a statistic from any one larger sample yields a closer estimate to the population value

Appendix:
Probability Distributions

Purpose of Statistical Inference

Need a method to estimate an unknown value

- **Central Purpose of Statistical Analysis**: Analyze data to better understand reality in terms of the underlying population values as a basis for making business decisions
- For the typical data analysis scenario, the population values such as μ are not known, but instead *estimated* from the data
- **Key Concept**: The development of how to estimate population values from the data begins with understanding the type of samples found from *known* population values
- The statistical principles illustrated with coin flipping are
  - the same core principles involved with the application of statistical inference to business applications
  - easier to visualize in the more sterile and some ways simpler context of coin flipping

Consider a Model of Flipping a Fair Coin

With a model of reality, can calculate the population mean

- Consider a model of reality: Flipping a fair coin
- Let the variable Y represent the outcomes of the coin flips, with a Head scored as Y=1 and a Tail scored as Y=0
- Now consider the complete, hypothetical *population* of all the data, all possible fair coin flips
- In the population, exactly $\frac{1}{2}$ of the values of Y are Y=1 and exactly $\frac{1}{2}$ are Y=0
- Because the probabilities of the data outcomes are known, here we can calculate the value of μ as a weighted mean

$$\mu = \tfrac{1}{2}(1) + \tfrac{1}{2}(0) = .5$$

- For a fair coin, the value of this abstraction μ is known
- $\mu = .5$ is the underlying reality with the presumption of a fair coin, manifested in the data with values of either 0 or 1

Heads or Tails?

Evaluate the fairness of the coin when μ is not known

- The decision maker realizes that in a particular context, such as when gambling, that the coin may or may not be fair
- To illustrate the need for inference, consider a situation in which the decision maker places a single bet, investment, on a Head or a Tail
- To provide a basis for making the investment, the decision maker wishes to assess as to whether a coin is either fair, or biased toward landing on either a Head or a Tail
- That is, if a Head is scored a 1 and a Tail a 0, the decision maker wishes to know if
  - $\mu = .5$, so no basis for a bet on either Head or Tail
  - $\mu > .5$, so bet on a Head
  - $\mu < .5$, so bet on a Tail

Heads or Tails?

Evaluate the fairness of the coin

- This situation illustrates several aspects of the real world application of statistical inference
  - The outcome is unknown and so an investment is a bet against a future reality
  - The rational decision for a bet on the outcome is based on the value of the μ, the population mean
  - The true value of μ is unknown
  - Estimate, that is, infer, μ from the data
  - Only one decision/investment will occur
  - Even if the coin is biased towards a Head, so that the best decision is to bet on a Head, the outcome could still be a Tail, what might be referred to as "bad luck"

Probability

To assess fairness, gather a sample, here 10 coin flips

- To assess reality, begin with a model of reality, here a fair coin
- IF $\mu = .5$, THEN how many Heads will result?
- Because of sampling fluctuations, for any one sample
  - There is no specific answer, so could obtain 5 Heads or 4 Heads or 6 Heads or anywhere from 0 to 10 Heads
  - Can only know *generally* how many Heads will result
- **Probability**: A number from 0 to 1, inclusive, that indicates how likely it is that a specified event will occur
  - 0 indicates certainty that the event will *not* occur
  - 1 indicates certainty that the event will occur
  - .5 indicates that the likelihood the event will occur is the same as it will not occur
  - Intermediate values are scaled accordingly

Probability Distributions

Assessing probability over all possibilities

- With only two outcomes, Heads or Tails, scored as 0 and 1 respectively, the population mean μ is also the population proportion, specified by the Greek letter π, which rhymes with "pie", and spelled as "pi"
- Coin flipping probabilities are expressed here by x, the number of obtained Heads over the $n = 10$ trials (coin flips), with the probability of each outcome, here $\pi = 0.5$, the same for each trial
- **Binomial Probability**: Probability of achieving a specified outcome x times from one of two possibilities over n independent trials, each such outcome with probability π
  - Each trial results in one of only two outcomes
  - The word "independent" means that the outcome of one trial has no influence on the outcome of any other trial

Logic of Binomial Probabilities

What is the probability of 5 Heads on 10 tosses (trials)

- Probability of one Head or one Tail for this fair coin is $P(H) = P(T) = 0.5$
- Each trial is independent, so for a pattern of two flips, such as HT or TT, the probability of any one pattern is the product of the individual probabilities for each trial, $(.5)(.5) = .5^2$
- Similarly, the probability of any one pattern of H's and T's for 10 flips, such as HTTHHTHTHT, is $.5^{10}$
- The probability of all patterns of 5 H's and 5 T's is the probability for any one pattern, $.5^{10}$, multiplied by the total combination of the number of all possible patterns such as HHHHHTTTTT, HTHTHTHTHT, THHTTTHTHH, etc
- There is a formula for calculating the number of such combinations, but it is tedious, and, fortunately, can use the computer for these computations

Obtain Binomial Probability Distribution

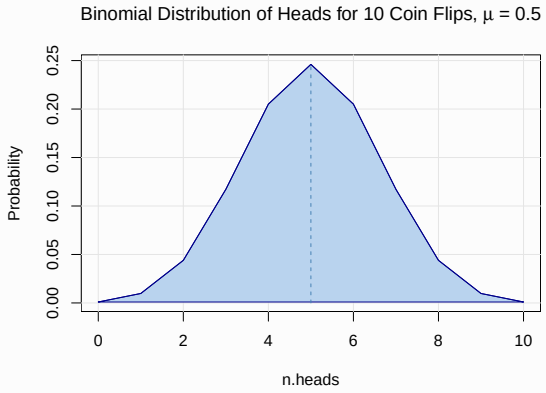
The distribution of probabilities for all possibilities

- To obtain the specific binomial probability for five Heads out of ten flips with a fair coin, set x=5, n=10 and pi=.5 in
  - R: `> dbinom(x,n,pi)`
  - Excel: `=BINOMDIST(x,n,pi,FALSE)`
- Now consider the probability for each possible number of Heads, such as over 10 coin flips with a fair coin
- **Probability Distribution**: Probability of each possible outcome from a specified procedure
- To calculate all 11 binomial probabilities at once, specify a list of integers from 0 to 10 with 0:10, then plot

```
> x <- 0:10
> probs <- dbinom(x,10,0.5)
> Plot(x, probs, col.area="lightsteelblue2")
```

Binomial Probability Distribution for a Fair Coin

- Probability of 5 Heads on 10 flips of a fair coin is just 0.246
- Probability of 4 Heads is not much lower, 0.205



Other Possibility is that the Coin is Biased

Consider a bias towards Tails

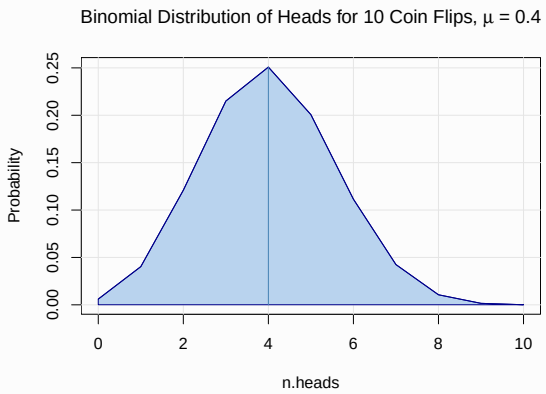
- Now consider a model of reality in which the coin is biased, with a probability of a Head on each trial of only 0.4, $\pi = 0.4$
- However, the person placing the bet does not know the true state of reality, so he or she must do what all of us must do when we wish to understand some aspect of reality without direct knowledge of that reality

  Estimate the value of the population parameter from sample data

- To illustrate one of the issues with this estimation, consider the corresponding probability distribution for the biased coin with $\pi = 0.4$

Binomial Probability Distribution for Biased Coin

- Probability of 4 Heads for $\pi = .4$ is 0.251
- Probability of 5 Heads on 10 flips of this coin is 0.201



Dilemma Posed by Sampling Error

Sampling error blurs our view of the underlying reality

- **Key Concept**: The existence of sampling error, the random fluctuations of sample results from sample to sample, means that what we observe does not unequivocally reveal the truth
- The most often result for a fair coin is that 0.5 of the flips result in a Head, and so also 0.5 for a Tail, however ...
  - If the coin is fair, $\pi = 0.5$, then there is about a 1 in 5 chance of getting only 4 Heads
  - If the coin is biased towards Tails, $\pi = 0.4$, then there is about a 1 in 5 chance of getting 5 Heads
- The difficulty is if our data is 5 Heads or 4 Heads, in *either* case it is reasonable that $\pi = 0.5$ or $\pi = 0.4$ or ...
- Estimation of a population value can be difficult because a sample result is consistent with multiple possibilities

Index Subtract 2 from each listed value to get the Slide

The End