

Sampling Distributions

Jeffrey Woo

School of Data Science, University of Virginia

Introduction

We will explore a few concepts in statistical theory that will allow us to assess how we can use data from our sample to make inferences about the larger population of interest.

1 Histograms and Probability Density Functions

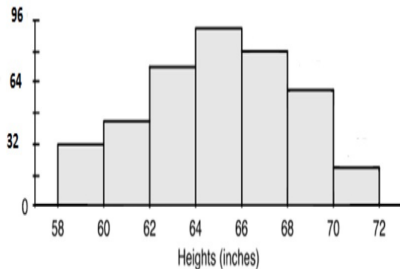
2 Population and Samples

3 Sampling Distribution of Sample Mean

Histograms

Graphical tools are often used to summarize data to give us an idea about our data. For example, with quantitative data, we often use a histogram

Heights	Frequency	Relative Frequency
$58 \leq x < 60$	32	0.08
$60 \leq x < 62$	44	0.11
$62 \leq x < 64$	72	0.18
$64 \leq x < 66$	92	0.23
$66 \leq x < 68$	80	0.20
$68 \leq x < 70$	60	0.15
$70 \leq x < 72$	20	0.05



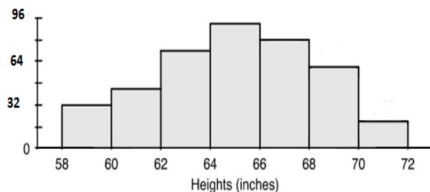
Histograms

With a histogram, we can have an idea about

- the center of the data
- the variability of the data
- the distribution of the data

The **distribution** of the histogram informs us the possible values of the variable of interest, as well as how often various values occur in our data.

Histograms

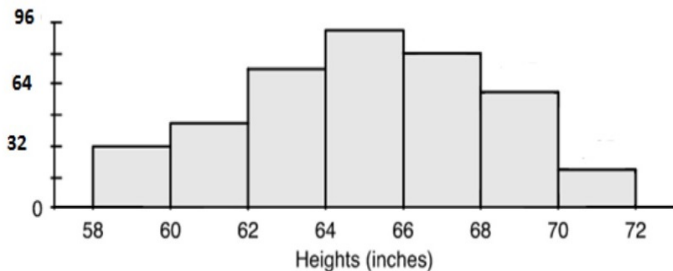


The distribution of this data

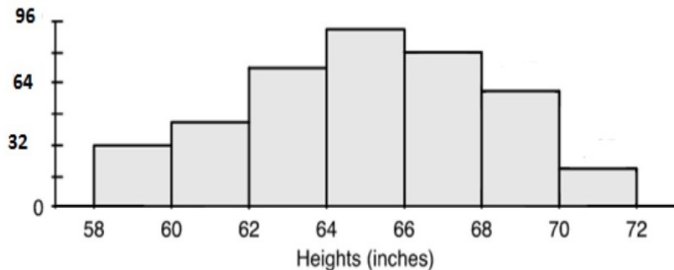
- The most common heights are around 64 to 66 inches.
- Heights are between 58 and 72 inches.
- As heights are further away 65 inches, they become less likely to occur.

Histograms

Question: Based only on this histogram, what are some ways you can estimate the proportion of girls who are at least 68 inches tall?



Histograms



We could also use a mathematical function to approximate the histogram, and use areas under the mathematical function to estimate proportions.

Probability Density Functions

These functions are called **probability density functions (pdf)**.

These functions must

- non-negative, and
- integrate to 1.

The density function is a mathematical representation of the distribution of the data.

Common Probability Density Functions

- Normal distribution.
- t distribution.
- χ^2 distribution.
- F distribution.

Normal distribution

A normal distribution is a symmetric, bell-shaped distribution. A normal distribution with mean μ and standard deviation σ is denoted by $N(\mu, \sigma)$. Its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi} \exp(-\frac{1}{2} (\frac{x-\mu}{\sigma})^2)} \quad (1)$$

If (1) is a good approximation for the distribution of the data, we can estimate probabilities by integrating (1) over the relevant range(s).

Normal distribution

- A normal distribution with mean 0 and standard deviation 1 is called a **standard normal distribution**.
- It turns out that any normal distribution X with mean μ and standard deviation σ can be standardized by

$$Z = \frac{X - \mu}{\sigma}.$$

- Then Z follows a standard normal distribution.

- 1 Histograms and Probability Density Functions
- 2 Population and Samples
- 3 Sampling Distribution of Sample Mean

Motivation

In many studies, we want to get answers to questions regarding a population of interest. For example, what is the average annual income of American adults?

- Ideally, we would like to obtain the data from every single American adult.
- However, due to constraints (e.g. time and money), we are unable to obtain the data from every single American adult.
- We then typically collect data from a random sample of American adults.
- We then use the characteristics of the sample to estimate the characteristics of the population.

Population Vs Sample

- **Population:** The group of all items of interest in our study.
- **Sample:** The items from which we actually collect data on.



Population Vs Sample: Example

A manufacturing company produces 5 million parts. To estimate the proportion of parts that are defective, 300 parts are randomly selected and carefully inspected for defects. What is the

- population of interest?
- sample?

Parameters Vs Statistics

- A **parameter** is a number describing a characteristic of the population. Parameters are fixed values, but in practice we do not know their numerical values.
- A **statistic** is a number describing a characteristic of a sample. Statistics vary from sample to sample.

We often use a statistic to estimate an unknown parameter.

Variability in Statistics

Each time we take a random sample from a population, we are likely to get a different set of individuals and calculate a different statistic. There is **variability** in the statistics.

Question: Can we quantify this variability without having to obtain many different random samples?

Variability in Statistics

- If we take lots of random samples of the same size from a given population, the distribution of the sample statistics, **the sampling distribution**, will follow a predictable shape.
- Under some circumstances, the sampling distribution can be well-approximated by a specific distribution and its pdf.
- The variance of statistics generally decrease as the sample size increase.

1 Histograms and Probability Density Functions

2 Population and Samples

3 Sampling Distribution of Sample Mean

Sampling Distribution of Sample Mean

When a continuous variable, X , in a population follows a $N(\mu, \sigma)$ distribution, the sampling distribution of the sample mean, \bar{x} , for all possible samples of size n is $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Central Limit Theorem

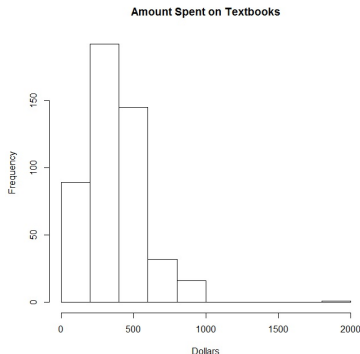
Consider a quantitative variable, X , in a population that has mean μ and standard deviation σ , and is not necessarily normally distributed. If n is **large enough**, the sampling distribution of the sample mean, \bar{x} , for all possible samples of size n is approximately $N(\mu, \frac{\sigma}{\sqrt{n}})$.

This is known as the **Central Limit Theorem**.

Implication: With a large enough sample size, we can use the normal distribution to find probabilities associated with sample means.

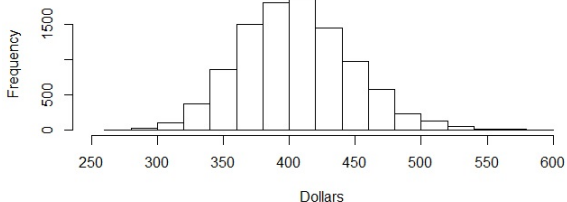
Worked Example: Textbook Spending

Question: Based on data from the Spring 2017 semester, the mean amount spent on textbooks for the semester is \$405.17 with standard deviation \$210.59. The histogram for the variable amount spent on textbooks that semester is displayed below. How would you describe the shape of this histogram?

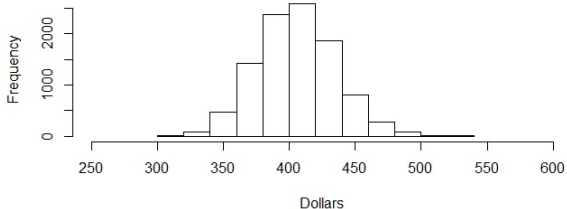


Worked Example: Textbook Spending

Histogram of Sample Means with $n=25$



Histogram of Sample Means with $n=50$



Worked Example: Textbook Spending

Question: Suppose I have a random sample of 25 students. What is the probability that the sample mean is less than \$415? What if I have a random sample of 50 students instead?

Worked Example: Textbook Spending

Question: Suppose I have a random sample of 50 students. What is the probability that the sample mean is more than \$400?

