Graphs and Sampling Distributions

So far, we have studied probability, the science of formulating models for abstract experiments and computing probabilities. Statistics is the science of (collecting), presenting and interpreting data. To do that, we need the models constructed using probability and apply them to "real" (observed) data. In the next few lectures we will look at some general properties of data and how data is usually recorded and described.

Population: a well defined group of objects (or individuals) that the researchers are interested in learning about.

Sample: a subset of the population that will be examined to gather information (data).

Census: the attempt to gather information from every individual in the population.

Variable: the characteristic whose values are observed (measured) in the study.

Example: Characteristics such as average word length, average sentence length and word length frequency distribution can be used to detect plagiarism and determine the authorship of documents. Pseudepigrapha (greek: "False ascriptions") are falsely attributed works of writing. Are some of the works attributed to Shakespeare really written by him or maybe by Francis Bacon? To study this claim, a linguist randomly selects ten pages from a work whose authorship is in question. He then records the length of every word on those ten pages. What is the population in this study? What is the sample? What is the variable?

Population：All the words in the boos

Sample: words on 10 selected pages

Variable: word length

How can this method be helpful to determine authorship?

Univariate Data are observations made on a single variable. They can be Quantitative or Categorical in nature. Quantitative data assumes numerical values (e.g., 8 letters or 2.3 inch) while categorical data takes on label values (e.g., blond, brunette etc.) If two (or more) observations are made on each individual in the population we have Bivariate (or Multivariate) data.

Example: In the above context of Pseudography and Plagiarism, name at least one categorical and one quantitative variable that may be of interest to the investigators.

Poll Question 10.1

Descriptive Statistics

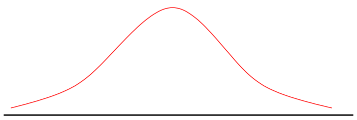
Recall that Statistics is the science of collecting, presenting and interpreting data. Descriptive Statistics focuses on the "Presentation" point. How can a large quantity of data be visually and numerically represented to allow a viewer to draw conclusions? Some of the most commonly used display methods for univariate quantitative data include histograms and box-plots.

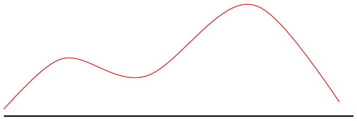
Histograms

In a histogram, the possible values of the variable are marked on the x-axis and the y-axis shows the observed frequencies of the values. For continuous data, the possible values are grouped together in small intervals, for which relative frequencies are computed.

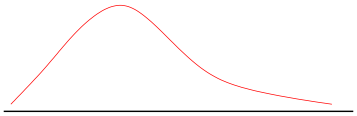
Histograms are intended to provide the observer with a quick visual overview of the data. Characteristics to look out for are:

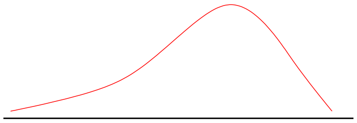
- Number of (large) peaks. Data with one peak is called UNIMODAL, data with two peaks is called BIMODAL and three or more peaks are said to be MULTIMODAL.
- Location. What is a representative value in the "center" of the histogram? "Center" is not necessarily to be equated with "middle" in this context.
- Spread. How many measurement units are the observations spread out around this center?
- Symmetry. A unimodal histogram is positively skewed if the right "tail" is longer and thicker compared to the left tail. The histogram is negatively skewed if the left tail is heavier than the right.


Unimodal


Bimodal

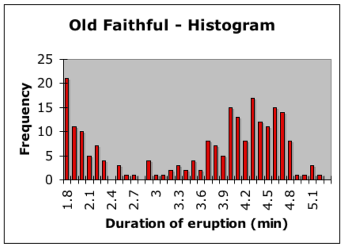
multi-modal

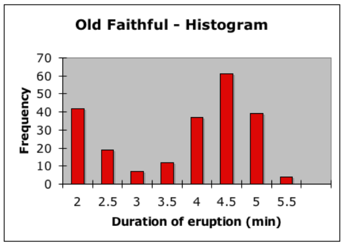

Positively Skewed

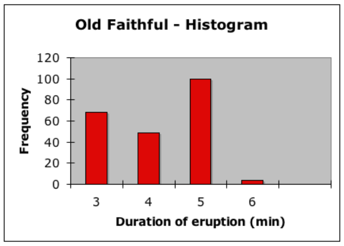

Negatively Skewed

Note: It is assumed that you still remember (from middle-school) how to create and interpret a histogram. If you forgot, please remind yourself using the review materials available on Canvas under modules.

Example: Old Faithful is a geyser in Yellowstone National Park in Wyoming. Check out the Old Faithful live-stream to see real-time pictures of the geyser. The geyser erupts approximately every 60 - 80 minutes, shooting hot water 20-75 feet high into the air. Park rangers have collected data on the duration of eruptions and time between eruptions to predict future eruptions for park visitors. The following three histgrams display the same data on 222 eruption durations collected over 16 consecutive days.







(a) Is "eruption duration" a quantitative or categorical variable? Is it discrete or continuous?

Poll Question 10.2

(b) Comment on the differences between the three histograms. Where do they come from? different bin width and bin numbers

(c) Which of the three histograms is (in your opinion) best suited to display the data?

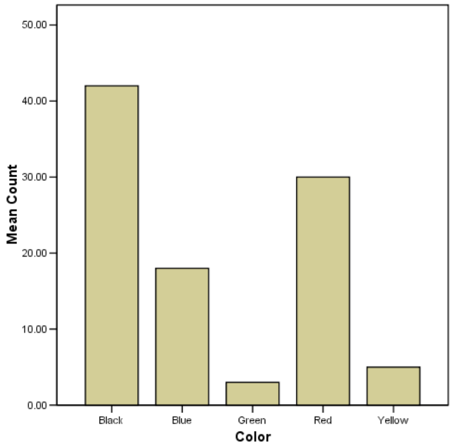
(d) Describe the features of the data set in words, referring to the histogram you chose in (b).

Bar Graph

Exclusively for categorical data, bar graphs are used to graphically display frequencies of occurrences of possible values.

Example: The number of cars (in thousands) sold in a European country is recorded by color and displayed in a bar plot

Color	Counts	percentage
Black	42	42.86
Red	30	30.61
Blue	18	18.37
Green	3	3.06
Yellow	5	5.10



What are the differences between a bar graph and a histogram?

- A bar graph is used to display non-numerical categorical data and a histogram is used to display quantitative data.

- In histograms, the values on the x-axis have to be in order. Bar graphs can have the categories listed on the x-axis in any order.

- The intervals of a histogram should cover the whole range with a sensibly chosen number of classes. Bar graphs need to have as many bars as there are categories.

Boxplots

Histograms convey a general sense about quantitative data but without numerical precision. A pictorial summary called a Boxplot combines several numerical data summaries in a graphical display. These features include

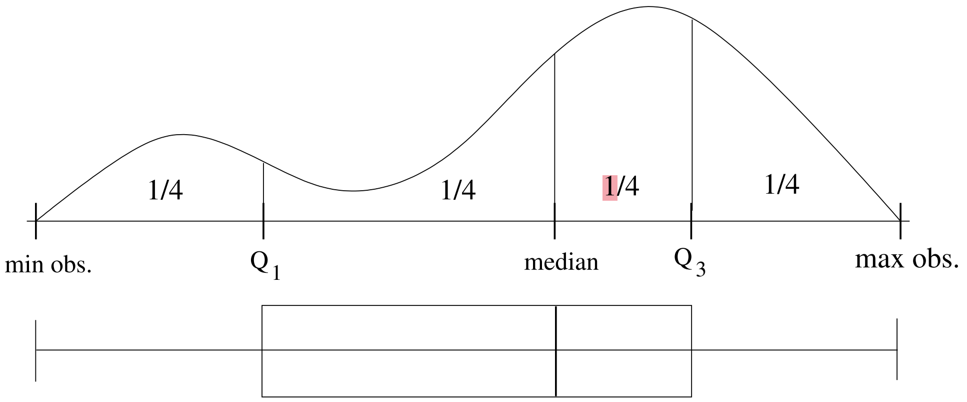
- The center;

- The spread;

- The extent and nature of departure from symmetry;

- The identification of outliers.

Boxplots are based on the following five summary statistics: The minimum observation, the first quartile (25th percentile), the median, the third quartile (75th percentile, and the maximum observation. The difference between the third and first quartile $Q_3 - Q_1$ is sometimes referred to as the fourth spread f_s or the interquartile range.



Outliers in boxplots are sometimes indicated as stars or dots. In this case the lines are extended to the largest and smallest observation that is *not* an outlier. When should an observation be considered an outlier?

An observation can be considered an outlier if it is either larger than $Q_3 + 1.5 \cdot f_s$ or smaller than $Q_1 - 1.5 \cdot f_s$. An outlier is extreme if it is either larger than $Q_3 + 3 \cdot f_s$ or smaller than $Q_1 - 3 \cdot f_s$.

Boxplots can be especially useful to observe similarities and differences of more than one distribution: Vital signs of patients on medication and placebo, fuel efficiency of different types of automobiles, yields of different breeds of corn etc.

Other Graphs

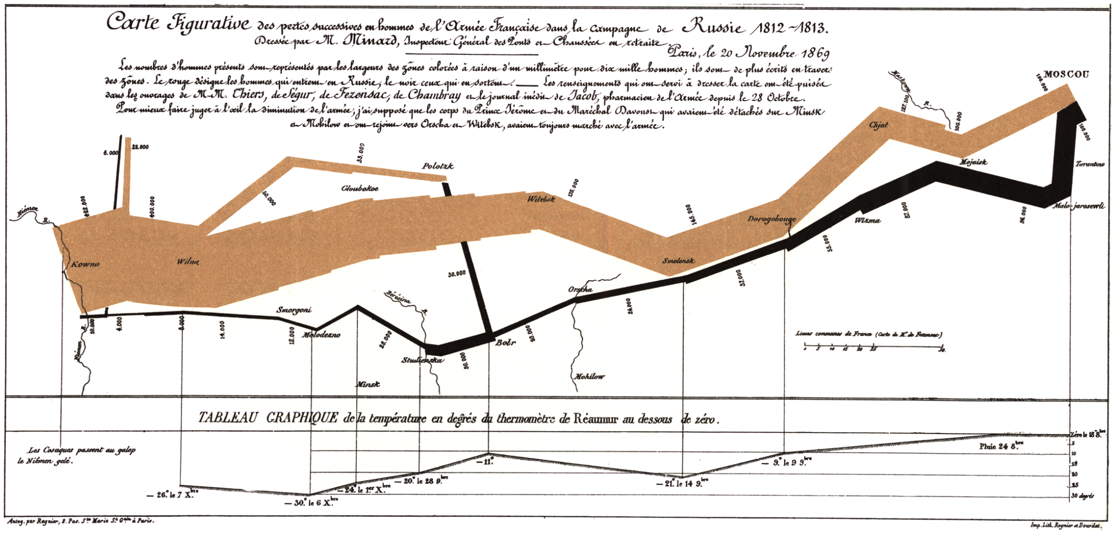
There are many other graphs that can be used to describe data. In fact, data visualization is its own field at the intersection of computer science and data science. If you take subsequent statistics classes (for instance Math 161B), you will learn more about graphical displays such as line-plots and scatter-plots and how to create them. SJSU also has a graduate statistics class on the topic (Math 250 - Data Visualization). The goal of data visualization is to clearly and effectively communicate information derived from your data. Good visualizations help viewers to make comparisons or draw conclusions about the data. To make a good visualization, one has to decide which aspects of the data to focus on, whether to show absolute counts or percentages (probabilities). Data can be shown as overall percentages (probabilities) or separated by groups and shown as percentages within each group (conditional probabilities).

Edward Tufte is an American Statistics professor who is famous for his work on data visualization. In his famous book “The Visual Display of Quantitative Information” he starts out by presenting guidelines for what a good graphical display should do.

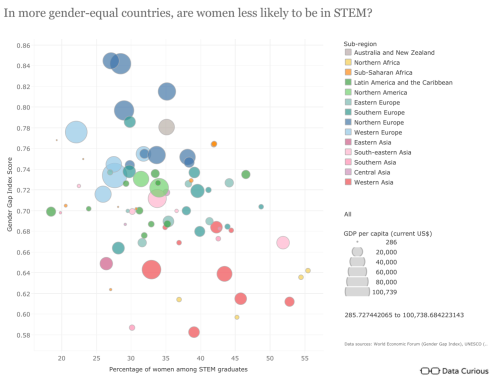
- Above all else show the data
- Design cannot rescue failed content

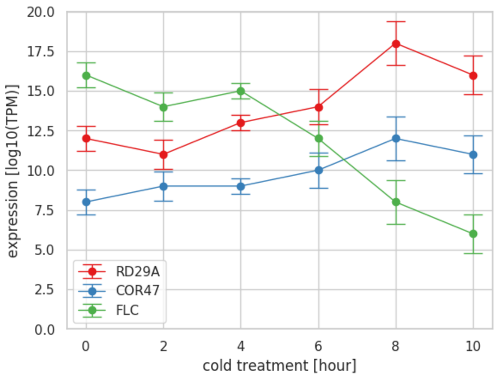
- Avoid distorting what the data have to say
- Present many numbers in a small space
- Encourage the eye to compare different pieces of data
- Reveal the data at several levels of detail, from a broad overview to a fine structure
- Serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- Be closely integrated with the statistical and verbal descriptions of a data set

Famous example of a good graphic: (which you can see on top of our math-stats department website). Produced in 1869 by Charles Minard. Shows Napoleon's March to Russia in the campaign of 1812. The thick band shows the size of the army. Also shown are geographical information, date, and temperature.

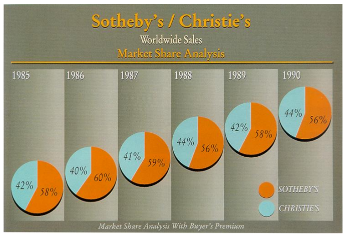


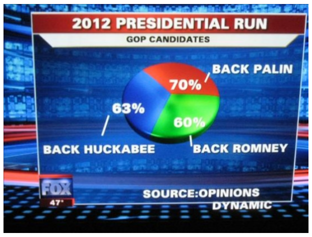
Below are a few more examples of good graphs. On the left you see a scatterplot, which shows the relationship between two quantitative variables (gender gap, women in STEM). In this plot, color and dot-size are used to incorporate additional information about the data. On the right, you see a line-graph in which the change of one variable (gene expression) is shown as a function of treatment time. In this graph, uncertainty is shown through error bars, which represent variability in the data.

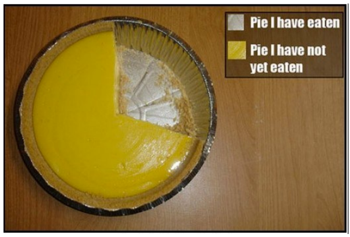




While it is possible condense a lot of data into visually appealing and informative graphs, it is also possible to use graphs to distort information or to mislead readers. Shown below are three exaples of pie-charts, of which only one is not misleading.







Poll Question 10.3

Numerical Ways to describe Data

Graphical displays of data are an excellent way to obtain a quick visual overview. Numerical measures can provide additional accuracy. Note, that all numerical measures on a sample depend on the selection of individuals that the sample data was derived from. If another sample were selected, those measurements would likely be different and thus all values would change. In fact, all quantities computed based on sample data are RANDOM VARIABLES.

Location

To describe the center or location of a data set the most common numerical measures are the mean and median.

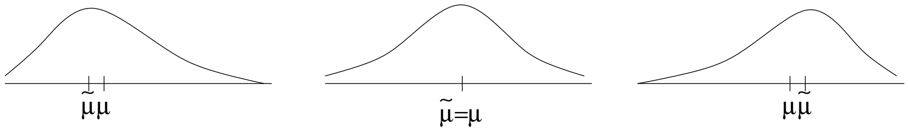
SAMPLE MEAN: For a given set of observations $x_1, \ldots, x_n$ the sample mean is the arithmetic average of the observations:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

It is usually reported with one decimal digit more accuracy than the observations themselves.

SAMPLE MEDIAN: The median is the "middle" observation. Its computation depends on the number n of observations. Order the observations by size and then read off the median as

$$\tilde{x} = \begin{cases} \text{the middle observation} & \text{if } n \text{ is odd} \\ \text{the average of the two middle observations} & \text{if } n \text{ is even} \end{cases}$$



Poll Question 10.4

Other measures of location include Quartiles, Percentiles and Trimmed Means. Recall that the median divides the data set into two parts of equal size. The same idea is used in the computation of quartiles and percentiles.

The sample mean is more susceptible to outliers in a data set than the sample median. A compromise between these two popular location measures is the trimmed mean. In it, the most extreme observations (on both ends of the spectrum) are discarded before the mean of the remaining observations is computed. A trimmed mean with a moderate trimming percentage (between 5% and 25%) will yield a measure of location that is not as sensitive to outliers as the mean and not as insensitive to distribution shape as the median.

Variability

We know that the variance or standard deviation are a measure for the variability of a random variable. Recall, that the original formula for the variance of a discrete random variable was

$$V(X) = \sum_{\text{all } x} (x-\mu)^2 p(x)$$

Even though we do not usually use this formula to compute variances, it can be used in the case of real data to compute the sample variance.

X bar

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

Note: The factor of $\frac{1}{n-1}$ makes the sample variance an unbiased estimate of the population variance. That means that $E(s^2) = \sigma^2$.

The Sample Standard Deviation is the square root of the variance

$$s = \sqrt{s^2}.$$

Poll Question 10.5

Statistics and Their Distributions

We have spent a lot of time discussing random variables. In experiments, we collect observations on those random variables (which we call DATA). The values of the observations are subject to chance (exact time of observation, selection of study participants etc.) and we can use probability theory to describe their distributions. Often, the data are used to compute a summary which is used to represent a particular aspect of the observations.

DEFINITION: A STATISTIC is any quantity whose value can be computed from the sample data. Prior to obtaining the data, there is uncertainty about the particular value the statistic will take on. Therefore, a statistic is a random variable (denoted by an uppercase letter). After the data has been collected, the value of the statistic can be computed (denoted by a lower case letter).

Example: Let X be the number of siblings of a randomly chosen student in this class. We are interested in the average $E(X) = \bar{X} = \mu$ of X.

(a) Randomly select three students and collect data from them.

(b) Compute the sample average $\bar{x}$.

The Distribution of the Sample Mean

The above example shows that the value of the sample mean can depend on the selection of subjects in a study. Thus the sample mean is a random variable and we can investigate its distribution.

PROPOSITION: Let $X_1, \ldots, X_n$ be a random sample from some distribution. This means that the observations $X_1, \ldots, X_n$ are independent and have the same distribution (in particular, they have the same mean μ and the same variance σ^2). Then

- $E(\bar{X}) = \mu$ $= \frac{\sum_{1}^{n} X_i}{n} = \frac{1}{n}E(X_1+\cdots+X_n) = \frac{E(X_1)+\cdots+E(X_n)}{n} = \mu$
- $V(\bar{X}) = \frac{\sigma^2}{n}$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. $= V\left(\frac{X_1+\cdots+X_n}{n}\right) = \frac{1}{n^2} \cdot \left(\frac{V(X_1)}{\sigma^2} + \cdots \frac{V(X_n)}{\sigma^2}\right)$

Poll Question 10.6

The Central Limit Theorem

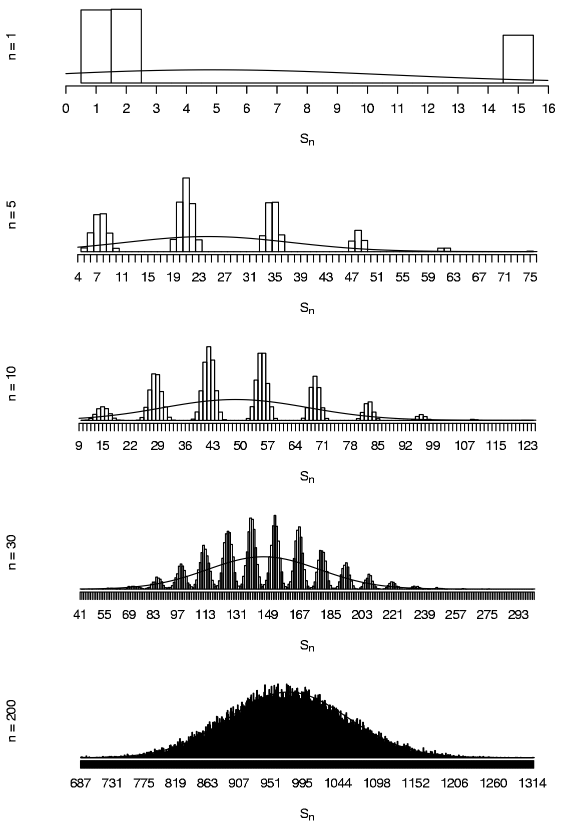
Remark: The Central Limit Theorem (CLT) is one of the most important results in probability theory. It was first stated by Abraham DeMoivre in 1733 and generalized by Pierre Laplace in 1812. Because of that, it is also sometimes called the De Moivre - Laplace Theorem. It took statisticians over 150 years (until 1901) to prove the Central Limit Theorem in its current most general form.

Situation: Consider an experiment where a certain variable is measured repeatedly many times. Let X_i be the result of the i^{th} measurement. The X_i's may have some discrete or continuous (not necessarily known) distribution. What can we say about the sum or the average of the X_i's?

Fact: If the X_i's have normal distributions, then the sum or average of the X_i's is exactly normally distributed.

Amazing Fact: If a measurement is repeated often, then the sum or the average of any kind of random variable (discrete or continuous, not necessarily normal) is approximately normally distributed.

Example: ~~Shown is the PMF of the sum of n discrete X_i for various n-values.~~



Poll Question 10.7

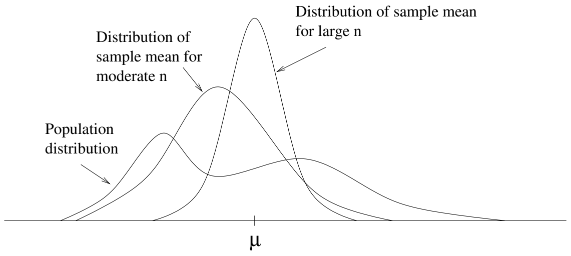
Theorem: CENTRAL LIMIT THEOREM

Let $X_1, \dots, X_n$ be independent random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$ for $i = 1, \dots, n$. Then for large n ($n > 30$), the sum and the average of the X_i's has approximately a Normal distribution

n: the sample size

$$\boxed{\begin{array}{llcl} \text{Average:} & \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i & \sim & \text{Normal}(\mu, \frac{\sigma^2}{n}) \\ \text{Sum:} & \sum_{i=1}^{n} X_i & \sim & \text{Normal}(n\mu, n\sigma^2) \end{array}}$$

The larger n is, the more closely will the PDF of the sum or average resemble a Normal PDF.



Fact: The Normal approximation to the Binomial is an application of the Central Limit Theorem. Recall, that a Binomial(n, p) random variable can be expressed as a sum of n independent Bernoulli(p) random variables.

Example: I have a 20-sided fair die. Suppose I roll the die 50 times and compute the average outcome. What is the probability that the average falls between 10 and 11?

X_i= outcome of the ith die roll

X_i~DiscreteUniform(1,...,20)

E(X_i)=(1+20)/2=10.5

V(X_i)=E(X^2)-E(X)^2

$\bar{X}$ ~Normal(10.5,33.25/50)

How often should I roll the 20-sided die to be at least 90% certain that the average of the outcomes will be between 10 and 11?

n=# of rolls

$\bar{X}$ = ~Normal(10.5,33.25/50)

0.9=P(10< $\bar{X}$ <11)

Point Estimation

Recall: The goal of statistics is to use data to draw conclusions about populations. In particular, we are interested in one or more specific characteristics of the population. These (unknown) population parameters are usually denoted by Greek letters. Since it is often not feasible to investigate every member of the population, a random sample is taken instead.

Population with unknown parameter μ

Take a sample $X_1, \ldots, X_n$

Compute a statistic $\hat{\mu} = \bar{x}$

Definition: A point estimate of a parameter θ is a number that can be regarded as a sensible value for θ. The point estimate is obtained by choosing a suitable statistic and computing the value of the statistic for a random sample taken from the population.

Examples: For each of the following scenarios suggest a sensible statistic that would lead to a point estimate for the population parameter.

(a) Let p be the proportion of teenagers in New York state who have been infected with the HPV virus. Suppose you have blood samples from n randomly chosen teenagers in the state.

Let X_i=1(individual who has HPV) or 0(else)

then, $\hat{p} = \bar{x} = \frac{1}{n}\sum x_i$

$\sum x_i \sim Bino(n, p)$

there are multiple ways of point estimate for one problem

(b) Let μ be the average yield (in tons) produced from one acre of cotton plants. You randomly select n cotton farmers who report the yields on one acre parcels of land.

X_i= yield on 1 acre for farmer i

$\hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i$

(c) Let σ be the standard deviation of annual rate of return of a particular investment. You have data on the annual rate of return for the past n years.

Popular point estimates: Some often used point estimates include

Sample proportion: $\hat{p} = \dfrac{\#\text{ of individuals in sample which exhibit a trait}}{n}$

Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Note: A point estimate $\hat{\theta}$ for θ is a random variable. It depends on the choice of the statistic and on the selection of the sample. Hence, we cannot always expect that our "best guess" $\hat{\theta}$ will be exactly equal to the true population parameter.

$$\hat{\theta} = \theta + \text{ error of estimation}$$

Unbiased Estimators

DEFINITION: A point estimate $\hat{\theta}$ is said to be UNBIASED, if

$$E(\hat{\theta}) = \theta$$

This means that if we keep obtaining "best guesses" $\hat{\theta}$ for θ, their average will converge to the true answer θ. If a point estimate $\hat{\theta}$ is not unbiased, then the difference $E(\hat{\theta}) - \theta$ is called the BIAS of $\hat{\theta}$.

Example: Suppose X is Binomial random variable. We are interested in estimating the population parameter p. In order to do this, we perform n independent Bernoulli trials and count the number X of successes. Then $\hat{p} = \frac{X}{n}$ is an unbiased estimate for p.

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = p$$

PROPOSITION: Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then the estimator

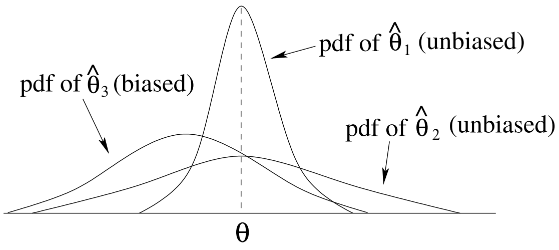
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is an unbiased estimator of the variance σ^2.

PROPOSITION: If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ, then $\bar{X}$ is an unbiased estimator of μ. If the distribution is symmetric and continuous, then the sample median $\tilde{X}$ or a trimmed mean are also unbiased estimators of μ.

Estimators with Minimum Variance

Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ. Since the estimates are both random variables, they have variances. Among all possible estimates for a population parameter, the unbiased estimates are usually preferred over biased estimates $\hat{\theta}_3$. Among the unbiased estimates, the best estimate is the one which has the lowest variance.


pdf of $\hat{\theta}_1$ (unbiased)
pdf of $\hat{\theta}_3$ (biased)
pdf of $\hat{\theta}_2$ (unbiased)
θ

Poll Question 10.8

Definition: The unbiased estimator with the lowest variance is called the minimum variance unbiased estimator (MVUE) of θ.

Standard Error

every statistic has its own standard error

Definition: The standard error of an estimator $\hat{\theta}$ is its standard deviation $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. If the standard error itself involves unknown parameters that can be estimated the substitution of those estimates into $\sigma_{\hat{\theta}}$ yields the estimated standard error $s_{\hat{\theta}}$.

Example: The Central Limit Theorem states that the distribution of the sample mean $\bar{x}$ of independent and identically distributed random variables $X_1, \ldots, X_n$ is

$$\bar{X} \sim \text{ Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

where $\mu = E(X_i)$ and $\sigma^2 = V(X_i)$. What can we say about the point estimate $\bar{x}$ for μ?

Note: Every point estimate $\hat{\theta}$ has its own standard error. The standard error of the sample mean $(\sigma/\sqrt{n})$ is the most well-known example. Other point estimates have *different* standard errors! The standard error of a sample proportion $\hat{p}$, for instance is $\sqrt{p(1-p)/n}$.

Example: If an investigator needs to obtain accurate information about a question which an interviewed person may be reluctant to answer honestly ("Have you ever smoked pot?", "Do you cheat on your taxes?" etc.) a technique called RANDOMIZED RESPONSE can be utilized. The investigator gives the subject two questions to answer - the delicate question and a completely harmless one.

- Question 1: Have you ever smoked pot?
- Question 2: Is the last digit of your social security number even?

The subject now rolls a die and will answer question 1 (Yes/No) if the die roll is even and question 2 (Yes/No) if the die roll is odd. Since a "Yes" answer does not stigmatize anyone, we can assume the responses to be truthful. Let p denote the percentage of Americans who have smoked pot.

(a) How would you use the Yes/No responses from n participants to estimate p? Let's call the number of "Yes" responses from our n participants X...

X_i=1,0

P("Yes")=p*0.5+0.5*0.5

$\sum X_i \sim$ Binomial $(n, P(yes))$

E(Binomial)=nP("yes")

$\sum x_i$

(b) Show that your estimator $\hat{p}$ is unbiased.

(c) Find the variance of your estimate $\hat{p}$.

(d) Would this method also work, if we would exchange the second harmless question for "Is your favorite color purple?"

Poll Question 10.9