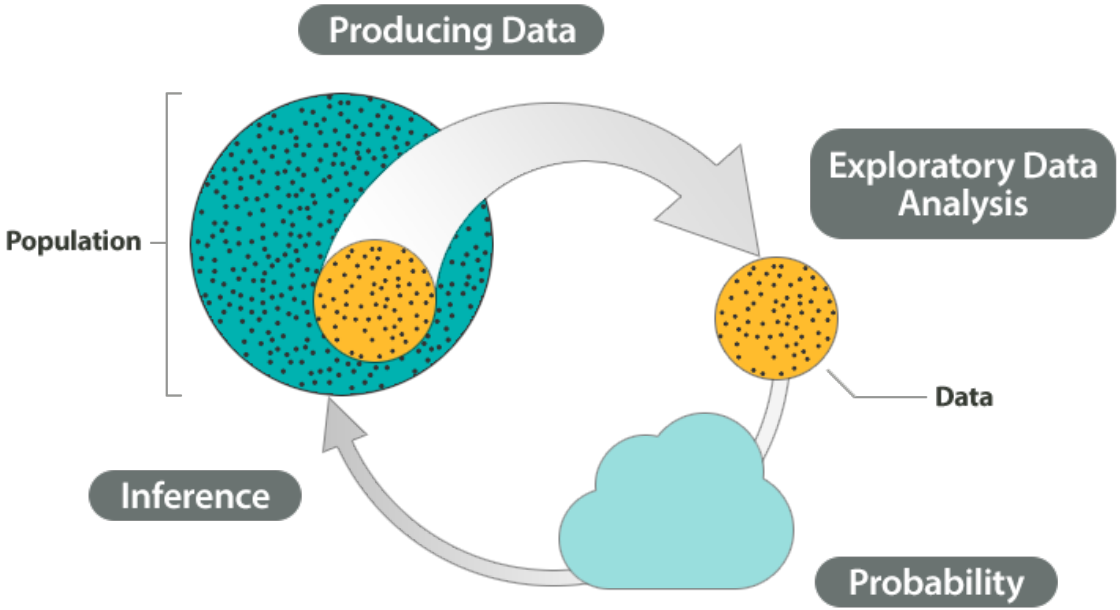
UCLA College | Physical Sciences Statistics

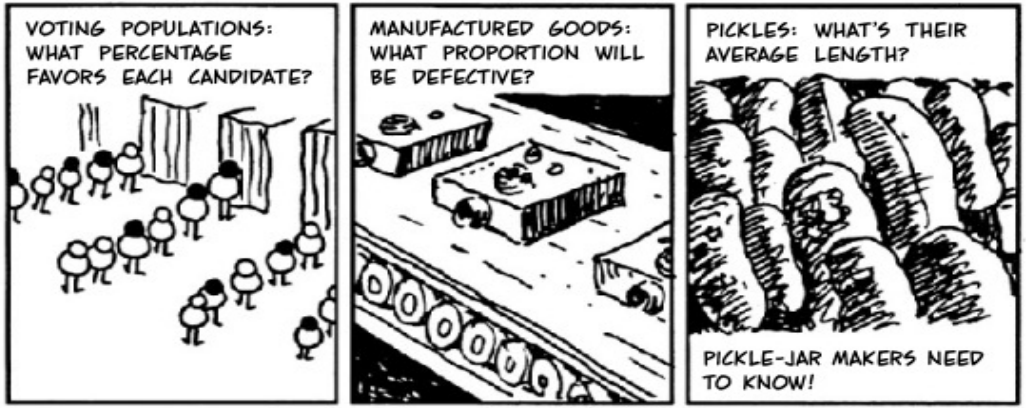
STATS 10: Introduction to Statistical Reasoning
Chapter 7

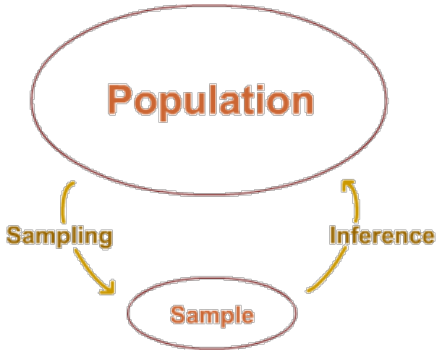
Bingling Wang
Department of Statistics

The Big Picture of Statistics



Statistical Inference





Statistical inference is the art and science of drawing conclusions about a population based on observing characteristics of samples.

- We use limited data to make inference about the population. It involves uncertainty because the entire population is not being measured.
- An important component of statistical inference is measuring that uncertainty.

Learning about the World through Surveys

Terminology

A **survey** is any activity that collects or acquires statistical data.
Usually some type of questionnaire (e.g. in-person, phone or internet survey).

- **Population**: a group of objects we wish to study.
  - All students enrolled in stats 10

- **Parameter**: a numerical value that characterizes some aspect of the population.
  - The mean GPA of stats 10 students

- **Sample**: a collection of people or objects taken from the population.
  - The students who came to class today

- **Statistic/Estimator**: a numerical characteristic of a sample of data and used to estimate the value of a characteristic of the population.
  - The mean GPA of the students came to class today

Notation for Parameters and Statistics

- Statistics -- quantities based on data from a sample.
- Parameters -- quantities based on the population.

In general, Latin letters are used to represent statistics, Greek characters are used to represent population parameters.

Statistics (based on data)		Parameters (typically unknown)	
Sample mean	$\bar{x}$ (x-bar)	Population mean	μ (mu)
Sample standard deviation	s	Population standard deviation	σ (sigma)
Sample variance	s^2	Population variance	σ^2
Sample proportion	$\hat{p}$ (p-hat)	Population proportion	p

Example

In Oct. 2020, the Pew Research Center surveyed 5,858 U.S. employed adults (full time or part time) to understand how the coronavirus outbreak has changed the way Americans work.

The survey found that about 38% of these employed adults say that, for the most part, the responsibilities of their job can be done from home.

Research question:

What proportion of U.S. employed adults can do their job mostly from home?

Identify **population**, **sample**, **parameter** of interest and **statistic/estimator**.

Bias

A survey method is **biased** if it has a tendency to produce an untrue value.

- **Selection bias:** taking a sample that is NOT **representative** of the population.
- **Measurement bias:** bias that results from problems in the measurement process
- **Estimator bias:** using statistics that tend to over or underestimate the parameter.

Ex. A senator conducted a poll in her state by calling 100 people whose names were randomly sampled from the phone book (mobile phones and unlisted numbers aren't in phone books). The senator's office called those numbers until they got a response from all 100 people chosen.

Simple Random Sampling

One way that should (but is not guaranteed) give a representative sample is

Simple random sampling (SRS)

A procedure for sampling from a population in which

(a) Every single object of the population is equally likely to be chosen

(b) Every possible sample has an equal chance of being selected.

Steps

1. Define the **target population**
2. Start with a **sampling frame**: a list of every objects in the population
3. Decide on the **sample size**
4. With the sampling frame, **randomly** draw a sample of objects with specified size **without replacement (SRSWOR)**
   -- no object can be repeated in a sample

Example

Suppose you want to determine the musical preferences of all students at your university, based on a sample of students.

Here are some examples of possible ways to pursue this problem.

1. Post a music-lovers' survey on the university's pop music forum, asking students to vote for their favorite type of music.

2. Stand outside the Fine Arts Building and ask students around to respond to your question about musical preference.

3. Obtain a student directory with email addresses of all the university's students and send the music poll to every 10th name on the list.

4. Obtain a student directory with email addresses of all the university's students, and randomly sample some addresses and send your music poll to these sampled students.

Evaluating Survey Method

Suppose we have a survey that asks a random sample of 100 UCLA students whether they are concerned with getting a job after graduation. The pollster (the one asking the question) then produces an estimate of the proportion of UCLA students who are concerned with getting a job after graduation.

Suppose we send out a group of pollsters and they all use the same method to conduct the survey. Each pollster produces an estimate of the proportion in the same way.

- How did the group perform as a whole?
- How different will the estimates be?

In evaluating the estimation method, we are interested in the

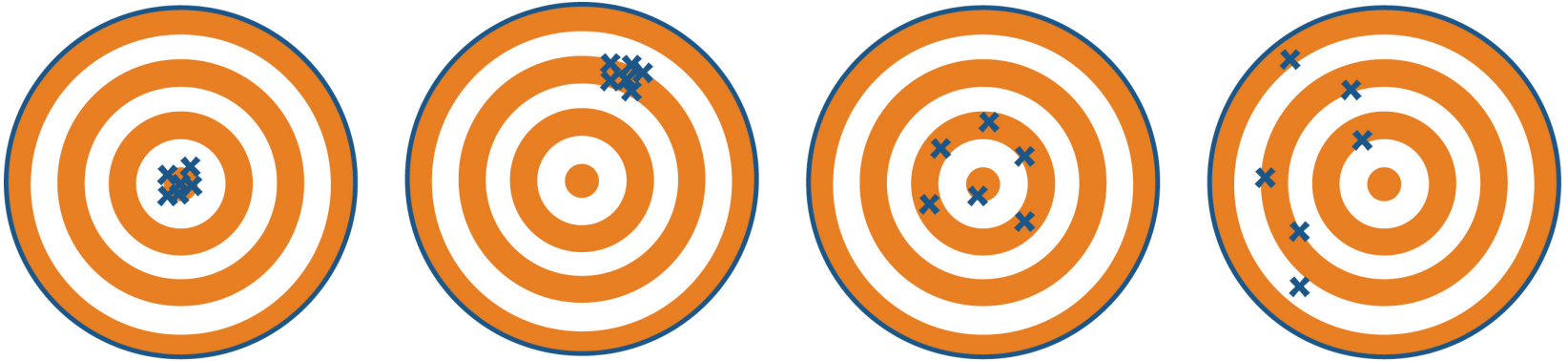
accuracy and **precision**

of our estimates.

Quality of a Survey

Accuracy Closeness of estimates to the true parameter

Precision Closeness of estimates to each other



High Accuracy
High Precision

Low Accuracy
High Precision

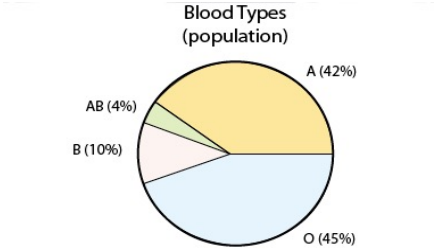
High Accuracy
Low Precision

Low Accuracy
Low Precision

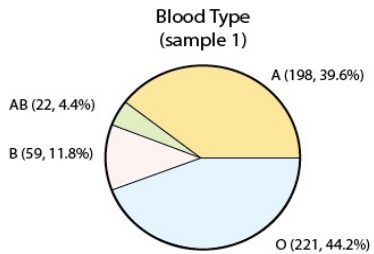
Sampling Distributions

Example

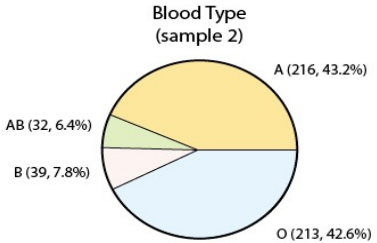
Distribution of blood types in the entire U.S. population and in samples of size 500.



Population

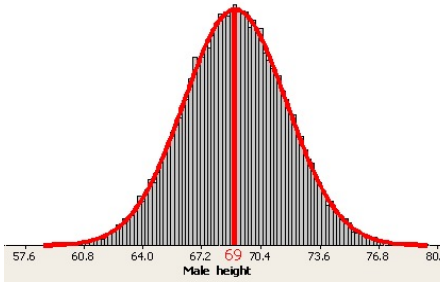


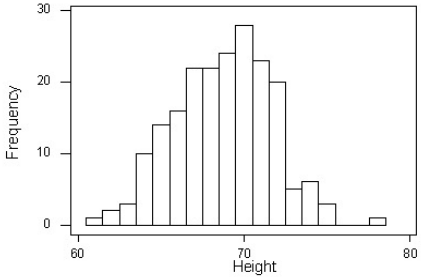
Sample 1

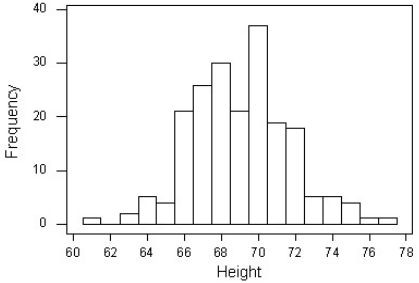


Sample 2

Distribution of heights of adult males in the entire population and in samples of size 200.







Sampling Distribution

A different random sample will produce different results and thus different estimates.

The sample statistic is used to estimate the population parameter.

- The value of the parameter is always the same.
- The value of the statistic vary from sample to sample.
- Statistic value is random as it depends on a random sample.
- The statistic has a probability distribution.

A **sampling distribution** is the **distribution** of the values of the **statistic** in repeated **samples**.

- Gives probabilities for values of the statistic.
- Gives important characteristics of the statistic, e.g., bias and precision.
- Used for making inferences about a population.

Bias and Standard Error

Measuring accuracy and precision of a statistic/estimator

- The **Bias** is a measure of the **accuracy** of a statistic.
  - The bias of an estimator is the distance between the mean value of the estimator and the population parameter.
  - An estimator is **unbiased** if its mean value is equal to the population parameter (bias is 0)

- The **Standard Error (SE)** is a measure of the **precision** of a statistic.
  - Standard deviation of the sampling distribution

Simulation -- Behavior of Sample Proportion

Suppose that in a class of 100 college students, 70% of the students sleep after 11:00 pm.

What would you expect to see in terms of the behavior of sample proportion ($\hat{p}$) if many random samples of size n were taken from the population?

Simulation steps:

1. Using SRS, we take a random sample of n students.
2. Ask the students in the sample if they go to sleep after 11:00 pm.
3. Compute the sample proportion $\hat{p} = \frac{\text{\# of people slept after 11:00 pm}}{\text{\# of people in the group}}$.
4. Repeat steps 1-3 a total of 1000 times. Each time calculate $\hat{p}$ and record its value.

For example, selecting a random sample of $n = 4$ students, the result might be:

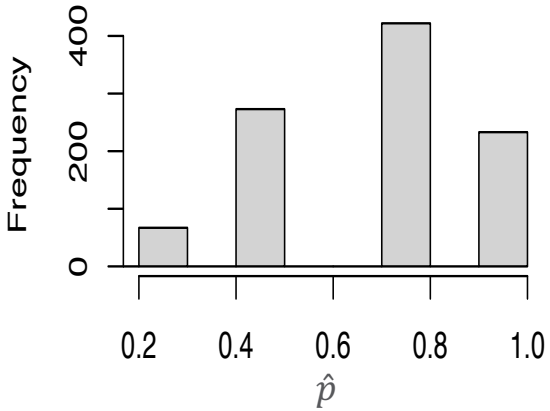
1 1 0 1

Then the sample proportion is $\hat{p} = \frac{3}{4} = 0.75$.

Sampling Distribution Simulation

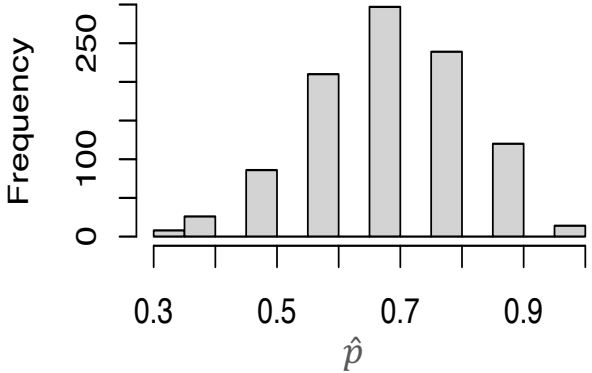
Distribution of sample proportions ($\hat{p}$) obtained from 1000 random samples of size n

sample size $n = 4$



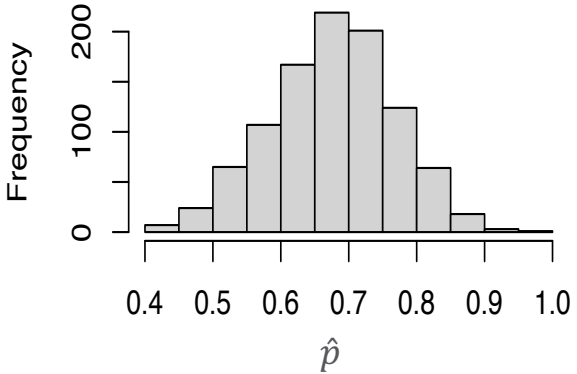
Mean: 0.7027; SD: 0.2258

sample size $n = 10$



Mean: 0.7029; SD: 0.1327

sample size $n = 20$



Mean: 0.7013; SD: 0.0925

Simulation Summary

Simulation	Sample Size	Mean	SE
1	4	0.7027	0.2258
2	10	0.7029	0.1327
3	20	0.7013	0.0925
Population proportion $p = 0.7$			

- The estimator $\hat{p}$ is unbiased regardless of the sample sizes.
- The standard error decreases (better precision) as the sample size increases.
- The shape of the sampling distribution is more symmetric for larger sample sizes.

Finding Bias and Standard Error

If the following two conditions are met:

1. The sample is **randomly** selected from the population of interest.
2. The population is much **larger** than the **sample size** (at least 10 times larger than the sample, as a rule of thumb).

Then,

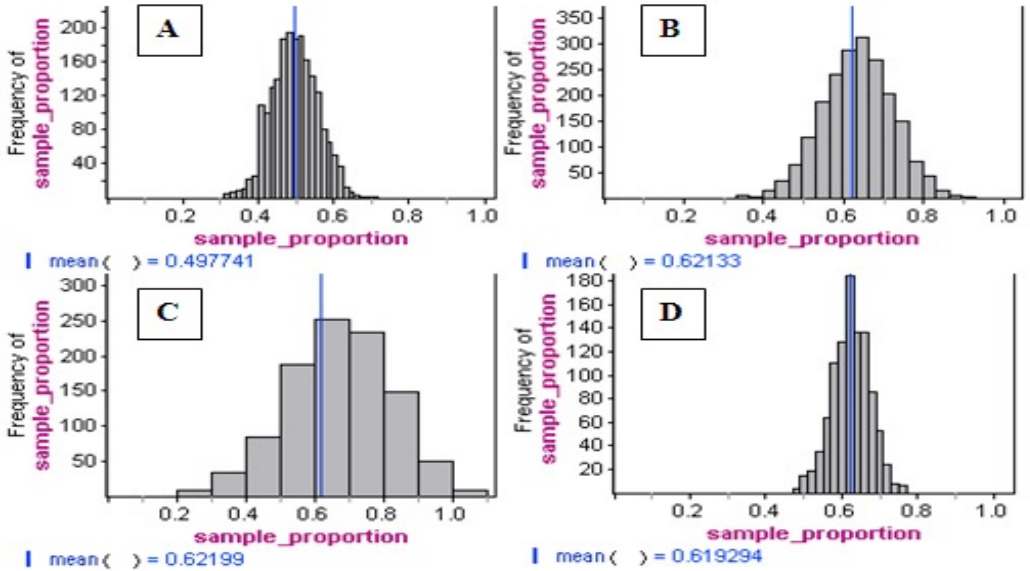
- The bias of $\hat{p}$ is 0
- The standard error is:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Example

According to the National Postsecondary Student Aid Study conducted by the U.S. Department of Education in 2008, 62% of graduates from public universities had student loans.

Which distribution is a plausible representation of the sampling distribution for random samples of 30 students?



The Central Limit Theorem for Sample Proportions

Central Limit Theorem

By approximating the sampling distribution of $\hat{p}$ through simulation, we saw that the shape of the sampling distribution changed as sample size increased. The shape is more symmetric for larger sample sizes.

Central limit theorem (CLT) is a mathematical theorem that gives a good approximation of the sampling distribution.

Let p denote the population proportion. If some basic conditions are met:

1. Random and Independent; 2. Large Sample; 3. Big Population.

Then the sampling distribution of $\hat{p}$ approximately follows a normal distribution:

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

If value of p is unknown, we substitute the value of sample proportion $\hat{p}$ in calculation.

$$N\left(\hat{p}, \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Checking Conditions for the CLT

Condition 1: Random and Independent. The sample is collected randomly from the population, and observations are independent of each other

Check: There is no way to check this by looking at the data. We need to trust (or assume) that the survey is well designed, and that sampling is done randomly.

Condition 2: Large Sample. The sample size n is large enough that the sample expects at least 10 successes and 10 failures (textbook uses 10 as a rule of thumb).

Check: If the sample size is n, and success probability is p, we expect np successes and $n(1-p)$ failures. If p is known, we check that $\mathbf{np \geq 10}$ and $\mathbf{n(1-p) \geq 10}$. If p is unknown, we check that $\mathbf{n\hat{p} \geq 10}$ and $\mathbf{n(1-\hat{p}) \geq 10}$.

Condition 3: Big Population. If the sample is collected without replacement, then the population size must be much (at least ten times) bigger than the sample size.

Check: If N is the population size, we check that $\boldsymbol{N \geq 10n}$

Example

It is known that 24% of all American drivers admit to texting while driving.

A researcher randomly select 200 American drivers and asked if they text while driving. The researcher found that 80 of 200 drivers who text while driving.

1. What do you expect the distribution for the sample proportion of drivers who text while driving among 200 sample drivers to be?

2. Is 0.40 an unusual value as $\hat{p}$ according to the distribution?

3. What is the probability that 55 or more among the 200 drivers text while driving?

Example

1. What do you expect the distribution for the sample proportion of drivers who text while driving among 200 sample drivers to be?

$$p = 0.24, \qquad SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.24 \times 0.76}{200}} = 0.03$$

Check CLT conditions:

1. The drivers were randomly selected.
2. $\mathrm{np} = 48$ and $\mathrm{n}(1-\mathrm{p}) = 152$.
3. Population size is much more than 10 times the sample size.

The sampling distribution of the proportion is approximately normal based on the CLT.

$$\hat{p} \sim N(0.24, 0.03)$$

Example

2. Is 0.40 an unusual value as $\hat{p}$ according to the distribution?

$$\hat{p} \sim N(0.24, 0.03)$$
$$z = \frac{0.4 - 0.24}{0.03} = 5.33$$

The z score is much greater than 2, therefore 0.40 is an unusually large value according to the distribution.

3. What is the probability that 55 or more among the 200 drivers text while driving?

$$P\left(\hat{p} \geq \frac{55}{200}\right) = P(\hat{p} \geq 0.275) = P\left(z \geq \frac{0.275 - 0.24}{0.03}\right) = P(z \geq 1.167) = 0.1216$$

If we were to take a sample of 200, there is about a 12.16% chance that 55 or more among the sampled 200 American drivers text while driving.