Statistical Analysis

Some of the slides in this presentation are from Bowerman, B. L., O'Connell, R. T., & Murphree, E.S., (2010). *Business Statistics in Practice* (6th Ed.), Copyright © McGraw-Hill Education.

Sampling and Sampling Distributions

1. Random Sampling
2. The Sampling Distribution of the Sample Mean
3. The Sampling Distribution of the Sample Proportion
4. Stratified Random, Cluster, and Systematic Sampling
5. More about Surveys and Errors in Survey Sampling

Random Sampling

- If we select n elements from a population in such a way that every set of n elements in the population has the same chance of being selected, then the elements we select are said to be a **random sample**
- In order to select a random sample of n elements from a population, we make n random selections from the population
  - On each random selection, we give every element remaining in the population for that selection the same chance of being chosen

With or Without Replacement

- We can sample with or without replacement
- With replacement, we place the element chosen on any particular selection back into the population

  We give this element a chance to be chosen on any succeeding selection

- Without replacement, we do not place the element chosen on a particular selection back into the population

  We do not give this element a chance to be chosen on any succeeding selection

- It is best to sample without replacement

Notes on Sampling

- Probability sampling is sampling where we know the chance that each element in the population will be included in the sample

  Allows making statistical inferences

- Convenience sampling is where we select elements because they are easy or convenient to sample
- Voluntary response sampling is where participants self-select
- Judgment sample is where a knowledgeable person selects population elements

Sampling Distribution of the Sample Mean

The **sampling distribution of the sample mean** $\bar{x}$ is the probability distribution of the population of the sample means obtainable from all possible samples of size n from a population of size N

Some Notes

- In many situations, the distribution of the population of all possible sample means looks roughly like a normal curve
- If the population is normally distributed, then for any sample size n the population of all possible sample means is also normally distributed
- The mean, $\mu_{\bar{x}}$, of the population of all possible sample means is equal to μ
- The standard deviation, $\sigma_{\bar{x}}$, of the population of all possible sample means is less than σ

General Conclusions

- If the population of individual items is normal, then the population of all sample means is also normal
- Even if the population of individual items is not normal, there are circumstances when the population of all sample means is normal (**Central Limit Theorem**)

General Conclusions *continued*

- The mean of all possible sample means equals the population mean
  - That is, $\mu = \mu_{\bar{x}}$
- The standard deviation s_x of all sample means is less than the standard deviation of the population
  - That is, $\sigma_{\bar{x}} = \sigma$
  - Each sample mean averages out the high and the low measurements, and so is closer to μ than many of the individual population measurements

The Empirical Rule

The empirical rule holds for the sampling distribution of the sample mean

- 68.4% of all possible sample means are within (plus or minus) one standard deviation $\sigma_{\bar{x}}$ of μ
- 95.5 % of all possible observed values of *x* are within (plus or minus) two $\sigma_{\bar{x}}$ of μ
- 99.7% of all possible observed values of *x* are within (plus or minus) three $\sigma_{\bar{x}}$ of μ

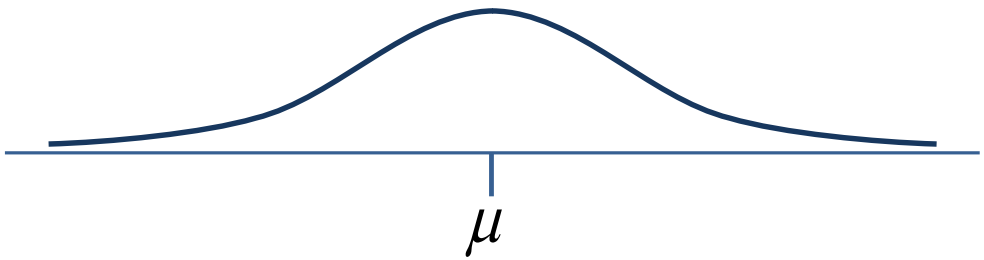
Properties of the Sampling Distribution of the Sample Mean

- If the population being sampled is normal, then so is the sampling distribution of the sample mean, $\overline{x}$
- The mean $\sigma_{\overline{x}}$ of the sampling distribution of $\overline{x}$ is $\mu_{\overline{x}} = \mu$
  - That is, the mean of all possible sample means is the same as the population mean

Example of a Sample Distribution

For a population of four students with weekend jobs

Student 1 receives	$100
Student 2 receives	$200
Student 3 receives	$300
Student 4 receives	$400
Population Mean	$250

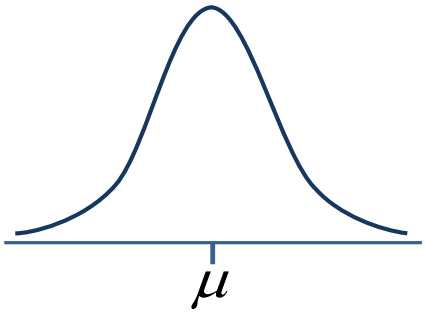


Example of a Sample Distribution *continued*

Let's sample two of the four $_4C_2 = 6$ combinations

Sampling distribution is:

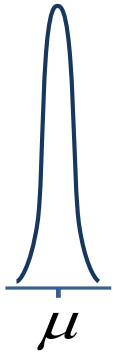
Distribution	$\bar{x}$
100, 200 =	150
100, 300 =	200
100, 400 =	250
200, 300 =	250
200, 400 =	300
300, 400 =	350
$\bar{\bar{x}} = \mu = 250$	



Let's take a larger sample size of three of the four population observations: $_4C_3 = 4$

Sampling distribution is:

Distribution	$\bar{x}$
100, 200, 300 =	200
100, 200, 400 =	233
100, 300, 400 =	266
200, 300, 400 =	300
$\mu = \bar{\bar{x}} = 250$	



Properties of the Sampling Distribution of the Sample Mean

- The variance $\sigma^2_{\bar{x}}$ of the sampling distribution of $\bar{x}$ is

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

- That is, the variance of the sampling distribution of $\bar{x}$ is
  - Directly proportional to the variance of the population
  - Inversely proportional to the sample size

Properties of the Sampling Distribution of the Sample Mean

- The standard deviation $\sigma_{\bar{x}}$ of the sampling distribution of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- That is, the standard deviation of the sampling distribution of $\bar{x}$ is
  - Directly proportional to the standard deviation of the population
  - Inversely proportional to the square root of the sample size

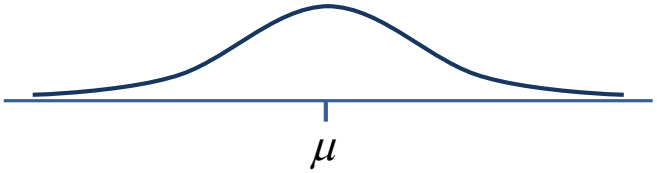
Notes

- The formulas for $\sigma^2_{\bar{x}}$ and $\sigma_{\bar{x}}$ hold if the sampled population is infinite
- The formulas hold approximately if the sampled population is finite but if N is much larger (at least 20 times larger) than the n $(N/n \geq 20)$
  a) $\bar{x}$ is the point estimate of μ, and the larger the sample size n, the more accurate the estimate
  b) Because as n increases, $\sigma_{\bar{x}}$ decreases as $1/\sqrt{n}$
     - Additionally, as n increases, the more representative is the sample of the population

So, to reduce $\sigma_{\bar{x}}$, take bigger samples!

Effect of Sample Size

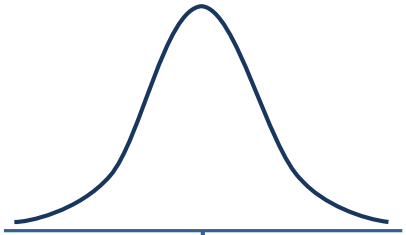
(a) The population of individual mileages



The normal distribution describing the population of all individual car mileages, which has mean μ pool and standard deviation $\sigma = .8$

Scale of mileages

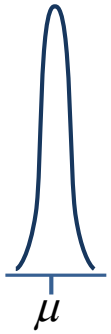
(b) The sampling distribution of the sample mean $\bar{x}$ when $n = 5$



The normal distribution describing the population of all possible sample means when the sample size is 5, where $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.8}{\sqrt{5}} = .358$

Scale of sample means, $\bar{x}$

(c) The sampling distribution of the sample mean $\bar{x}$ when $n = 50$



The normal distribution describing the population of all possible sample means when the sample size is 50, where $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.8}{\sqrt{50}} = .113$

Scale of sample means, $\bar{x}$

Central Limit Theorem

- Now consider a non-normal population
- Still have: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma\sqrt{n}$
  - Exactly correct if infinite population
  - Approximately correct if population size N finite but much larger than sample size n
- But if population is non-normal, what is the shape of the sampling distribution of the sample mean?
  - The sampling distribution is approximately normal if the sample is large enough, even if the population is non-normal (Central Limit Theorem)

The Central Limit Theorem *continued*

- No matter the probability distribution that describes the population, if the sample size n is large enough, the population of all possible sample means is approximately normal with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma\sqrt{n}$
- Further, the larger the sample size n, the closer the sampling distribution of the sample mean is to being normal

In other words, the larger n, the better the approximation

How Large?

- How large is “large enough?”
- If the sample size is at least 30, then for most sampled populations, the sampling distribution of sample means is approximately normal
  - Here, if *n* is at least 30, it will be assumed that the sampling distribution of $\bar{x}$ is approximately normal

    If the population is normal, then the sampling distribution of $\bar{x}$ is normal regardless of the sample size

Unbiased Estimates

- A sample statistic is an unbiased point estimate of a population parameter if the mean of all possible values of the sample statistic equals the population parameter

- $\bar{x}$ is an unbiased estimate of μ because $\mu_{\bar{x}} = \mu$
  - In general, the sample mean is always an unbiased estimate of μ
  - The sample median is often an unbiased estimate of μ but not always

Unbiased Estimates *continued*

- The sample variance s^2 is an unbiased estimate of σ^2

  That is why s^2 has a divisor of $(n-1)$ and not n

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

- However, s is not an unbiased estimate of σ

  Even so, the usual practice is to use s as an estimate of σ

Minimum Variance Estimates

Want the sample statistic to have a small standard deviation

- All values of the sample statistic should be clustered around the population parameter

Then, the statistic from any sample should be close to the population parameter

Minimum Variance Estimates *continued*

Given a choice between unbiased estimates, choose one with smallest standard deviation

- The sample mean and the sample median are both unbiased estimates of μ
- The sampling distribution of sample means generally has a smaller standard deviation than that of sample medians

Finite Populations

If a finite population of size N is sampled randomly and without replacement, must use the “finite population correction” to calculate the correct standard deviation of the sampling distribution of the sample mean

- If N is less than 20 times the sample size, that is,

$$if \;\; N < 20 \times n$$

- Otherwise

$$\sigma_{\bar{x}} \neq \frac{\sigma}{\sqrt{n}} \;\; \text{but instead} \;\; \sigma_{\bar{x}} < \frac{\sigma}{\sqrt{n}}$$

Finite Populations *continued*

The finite population correction is

$$\sqrt{\frac{N-n}{N-1}}$$

The standard error is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Sampling Distribution of the Sample Proportion

The probability distribution of all possible sample proportions is the sampling distribution of the sample proportion

If a random sample of size n is taken from a population then the sampling distribution of the sample proportion is

- Approximately normal, if n is large
- Has a mean that equals p
- Has standard deviation

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

Where p is the population proportion and $\hat{p}$ is the sampled proportion

Stratified Random, Cluster, and Systematic Sampling

- Divide the population into non-overlapping groups (strata) of similar units
- Select a random sample from each stratum
- Combine the random samples to make full sample
- Appropriate when the population consists of two or more different groups so that:
  a) The groups differ from each other with respect to the variable of interest
  b) Units within a group are similar to each other

  Divide population into strata by age, gender, income

Cluster Sampling

- “Cluster” or group a population into subpopulations
- Each cluster is a representative small-scale version of the population (i.e. heterogeneous group)
- A simple random sample is chosen from each cluster
- Combine the random samples from each cluster to make the full sample
- Appropriate for populations spread over a large geographic area so that…
  - There are different sections or regions in the area with respect to the variable of interest
  - A random sample of the cluster

Combination

It is sometimes a good idea to combine stratification with multistage cluster sampling

For example, we wish to estimate the proportion of all registered voters who favor a presidential candidate

- Divide United States into regions
- Use these regions as strata
- Take a multistage cluster sample from each stratum

Systematic Sampling

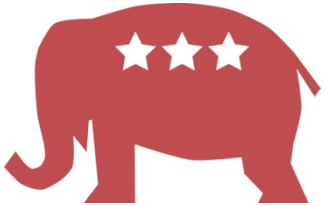
- To systematically select n units without replacement from a frame of N units, divide N by n and round down to a whole number
- Randomly select one unit within the first N/n interval
- Select every N/n^{th} unit after that

More about Surveys and Errors in Survey Sampling

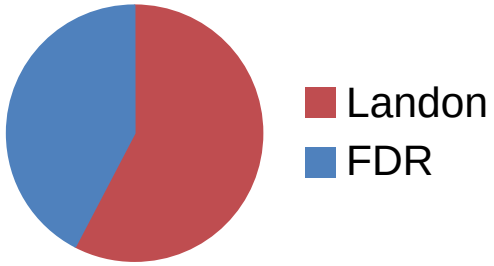
- Dichotomous questions ask for a yes/no response
- Multiple choice questions give the respondent a list of choices to select from
- Open-ended questions allow the respondent to answer in their own words

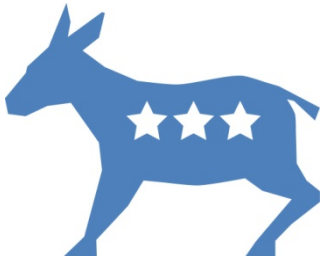
Errors in Sampling *example*

The Literary Digest poll of 1936 predicted that Landon would beat FDR in the presidential election. The poll's sample was based on prospective voters from magazine subscription lists, automobile registration lists, phone and club membership lists. During the depths of the Great Depression only the wealthy could afford these items. Typically wealthy voters were Republican, making the poll biased towards Landon and grossly inaccurate.

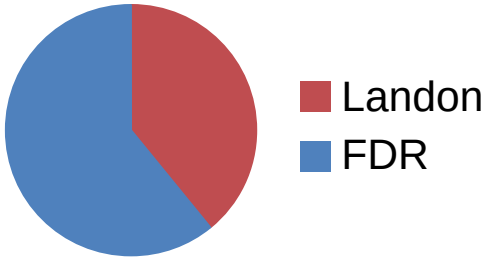


1936 Literary Digest Poll





1936 Actual Results



Errors Occurring in Surveys

- Must be well designed
  - *Target population:* entire population of interest
  - *Sample frame:* a list of sampling elements
  - *Sampling error:* difference between population and sample
- Random sampling should eliminate bias
- Random sample may not be representative:
  - *Under-coverage*: too few sampled units or some of the population was excluded
  - *Non-response*: when a sampled unit cannot be contacted or refuses to participate
  - *Response bias*: responses of selected units are not truthful