Probability and Inferential Statistics

Parameter	A number you derive from a population
Statistic	A number you derive from a sample
Census	A survey of the whole population

Symbols

	Population Parameter (Greek Letter)	Sample Statistic (English Letter)
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Variance	σ^2	s^2

Probability & Non-Probability Samples

Probability Samples	Every case in the population has the same chance of being selected
Non-Probability Samples	A specific group is being used as your sample. *Surveying students enrolled in a class*

Example

We want to know what % of students work during the semester.
We draw a sample of 500 from a list of all students at the university
N = 20,000 (all students at university)
P = 500/20,000
Use a table of random numbers to selected 500 ID numbers with 6 digits
6 digits will be chosen 500 times until they match up with student numbers
After questioning each of these 500 students, we find that 368 (74%) work during the semester.
Population – 20,000

Example (cont)

Sample – 500
Statistic – 74%
Parameter – Doesn't directly appear (it's implicit)
(% of all students in the population who held a job)

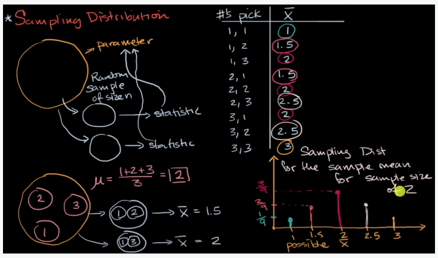
Sampling Variation

Sample Statistics	Variables (e.g., sample mean, sample proportion)
Sampling Error	The sample will differ from the population purely by chance
Positive Sampling Error	Making the statistic exceed the population
Negative Sampling Error	Making the statistic less than the population parameter

Sample statistic = population parameter + sampling error

Sampling Distribution
The theoretical, probabilistic distribution of a statistic for all possible samples of a given size (n).

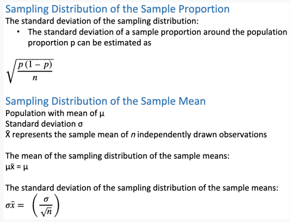
Construction of a Sampling Distribution



Statistic is used to estimate a parameter.
Not all statistics will have the same value.
What is the distribution of the values that we can get for the statistic?

Standard Error = population standard error / square root of the population size

Sampling Distribution


Sampling Distribution of the Sample Proportion
The standard deviation of the sampling distribution:
- The standard deviation of a sample proportion around the population proportion p can be estimated as

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

Sampling Distribution of the Sample Mean
Population with mean of µ
Standard deviation σ
X̄ represents the sample mean of *n* independently drawn observations

The mean of the sampling distribution of the sample means:
$\mu\bar{x} = \mu$

The standard deviation of the sampling distribution of the sample means:
$\sigma\bar{x} = \left(\frac{\sigma}{\sqrt{n}}\right)$

Practice Question

The average age for a population of doctors in a hospital is 51.6 years, What does this mean value represent?
A parameter
What does it mean for a sample to be representative
The sample reproduces the important characteristics of the population
Which set of symbols represents the standard deviation of the sampling distribution?
Which of these terms is synonymous with the standard error of the mean?
The standard deviation of a sampling distribution

Two Estimation Procedures

Point Estimate	A sample statistic used to estimate a population parameter
Confidence Intervals	Consist of a range of values instead of a single point

Example of point estimate:
50% of Canadians drive less because of gas.

Example of confidence:
Between 47% and 53% of Canadian drivers drive less due to high gas prices.

Confidence Intervals
- Point estimate is in the middle
- Lower and upper bound of C.I: 47% and 53%
- Margin of Error: radius or spread of the confidence interval (3%)

Criteria for Choosing Estimators

Bias	An estimator is unbiased if the mean of its sampling distribution is equal to the population value of interest
Efficiency	The extent to which the sampling distribution is clustered around its mean

Bias

% of sample means or proportions	Fall within
68%	± 1 standard deviation
95%	± 2 standard deviations
99%	± 3 standard deviations

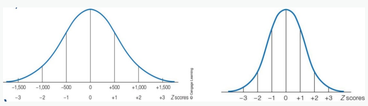
If n is large, we know that the sample mean/proportion is equal to the population parameter and: (image)

Very good (68 out of 100 chances) that our sample outcome is within +/- 1 standard deviation of the true population parameter

Excellent (95 out of 100) that it is within +/- 3 standard deviations

In less than 1% of cases, a sample outcome will lie further away than +/- 3 standard deviations

Efficiency



Getting back to the matter of dispersion: standard error $\sigma\bar{x}$ (standard deviation of the sampling distribution) = $\sigma/(\sqrt{n})$

Standard error is an inverse function of n: as sample size increases, $\sigma\bar{x}$ will decrease

The smaller the standard deviation of a sampling distribution, the greater the clustering and the higher the efficiency.

Constructing Confidence Intervals

1. Set the alpha, *a*
2. Find the Z score (or critical value) associated with alpha
3. Construct the confidence interval (we will substitute values into the appropriate formulas for confidence interval)

Constructing Confidence Intervals - Set the Alpha

1. Alpha = the probability that the interval will be wrong, I.e., it doesn't include the population parameter.

The commonly used alpha level 0.05 corresponds to a 95% confidence level.

If an infinite number of intervals were constructed at the 0.50 alpha level (all other things being equal). 95% of them would contain the population value; 5% would not.

Constructing Confidence Intervals - Find Z Score

Confidence Level (%)	Alpha	$\alpha/2$	Z Score
90	0.10	0.0500	±1.65
95	0.05	0.0250	±1.96
99	0.01	0.0050	±2.58
99.9	0.001	0.0005	±3.29

For an interval estimate based on +/-1.96 Z's:

The probabilities are that 95% of all such interval will include or overlap the population value

We can be 85% confident that the interval around our one sample outcome contains the population value

Confidence Interval

Point Estimate +/- Margin of Error

Point Estimate +/- (Critical Value * Standard Error)

The margin of error depends on:

(1) the standard error for statistic AND

(2) a "critical value/Z score" based on the confidence level

Constructing Confidence Intervals for Proportions

$$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{n}}$$

Point Estimate +/- (Critical Value/Score) x Standard Error)

for large samples (interval estimation for proportions based on small samples) (n<100) not covered)

Example

$$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{n}}$$

$$\textbf{C.I.} = P_s \pm 1.96\left(\sqrt{\frac{P_u(1-Pu)}{n}}\right) = .30 \pm 1.96\left(\sqrt{\frac{(0.5)(0.5)}{200}}\right)$$

$$= .30 \pm 1.96\left(\sqrt{\frac{0.25}{200}}\right) = .30 \pm 1.96(.035) = \mathbf{.30 \pm .07}$$

What proportion of students at your university missed at least one day of classes because of illness last semester?

Out of a random sample of 200, 60 reported having missed classes: Ps = 60/200 = .30

Confidence Intervals for Means

$$\text{c.i.} = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

where c.i. = confidence interval

$\bar{X}$ = the sample mean

Z = the Z score as determined by the alpha level

$\frac{\sigma}{\sqrt{n}}$ = the standard deviation of the sampling distribution or the standard error of the mean

formula for large samples (n≥100)

Example

$$\text{C.I.} = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\text{C.I.} = 105 \pm 1.96\left(\frac{15}{\sqrt{200}}\right)$$

$$\text{C.I.} = 105 \pm 1.96\left(\frac{15}{14.14}\right)$$

$$\text{C.I.} = 105 \pm (1.96)(1.06)$$

$$\textbf{C.I.} = \mathbf{105 \pm 2.08}$$

You want to estimate the average IQ of a community using a random sample of 200 residents
- with a sample mean IQ of 105
- assuming a population standard deviation for IQ scores of 15

Alpha set at .05 (i.e. we are willing to run a 5% chance of being wrong).

What is the corresponding Z score ?
What is the formula?

Conf

$$\text{c.i.} = \bar{X} \pm t\left(\frac{s}{\sqrt{n-1}}\right)$$

where c.i. = confidence interval
$\bar{X}$ = the sample mean
t = the t score as determined by the alpha level and $n - 1$ degrees of freedom
$\frac{s}{\sqrt{n-1}}$ = the estimated standard error of the mean, when σ is unknown

Three differences to Formula 6.1:
- σ is replaced by s
- n is replaced by **n–1** to correct for the fact that s is a biased estimator of σ

Three differences to Formula 6.1:
- σ is replaced by s
- n is replaced by n–1 to correct for the fact that s is a biased estimator of σ

To construct confidence intervals from sample means when s is unknown, we must use a different theoretical distribution, called the **Student's t distribution.**

T Distribution

The shape of the t distribution varies as a function of sample size.
- Distribution is a family of curves, each curve is defined by its degrees of freedom – a value indicating the number of scores in a sample that are "free to vary" when calculating statistics.
- **Degrees of freedom (df = n–1).**

T Distribution (cont)

- As n increases, s becomes a more and more reliable estimator of the population standard deviation (σ)

t distribution becomes more and more like the Z distribution.

Smaller samples: t distribution is flatter and has heavier tails than Z distribution.

The Z and t distribution are essentially identical when the sample size is greater than 100.

T-Table Practice

Find t score for alpha = 0.05 for n=30
Answers:
Degrees of freedom (df = n-1): 30 – 1 = 29
t score: ±2.045