Sample Proportions

$$\hat{p} = \frac{\text{\# of yes responses}}{n}$$

Variables & Definitions

p – population proportion (parameter)

$\hat{p}$ – sample proportion (statistic)

n – sample size

Sampling Distribution of $\hat{p}$

* The sampling distribution of $\hat{p}$ is approximately normal.

Assumption: $np \geq 10$ and $n(1-p) \geq 10$ must be verified before using a normal approximation.

* The mean of the sampling distribution is exactly p.

$$\mu_{\hat{p}} = p \quad (\hat{p} \text{ is an unbiased estimator of } p)$$

* The standard deviation of the sampling distribution:

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

Assumption: *population size* $\geq 10n$ must be verified before using standard deviation formula.

Normal approximations allow the use of:

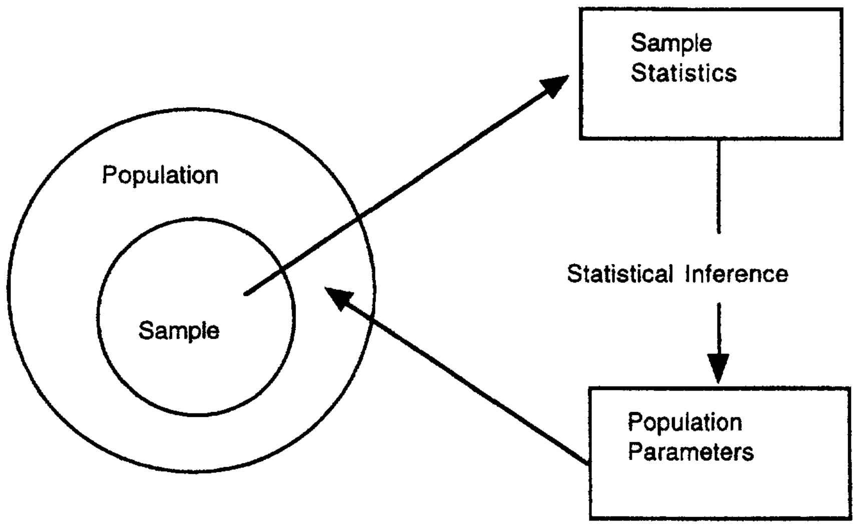
z – tables to find probabilities under a normal distribution curve or the use of calculator function Normalcdf (min, max, μ, σ)

Parameter

A numerical value that describes a charateristic in a population.

Statistic

A number calculated from a sample of a population.



In statistical practice:

The value of a parameter is not known.
Statistics are often used to estimate an unknown parameter.

Sampling Distribution of a Statistic

The distribution of values taken by the statistic in all possible samples of the same size from a population of interest.

Unbiased Estimator

A statistic in which the mean of its sampling distribution is equal to the true value of the paramater being estimated.

Variability of a Statistic

Described by the spread of the statistic's sampling distribution.
Determined by the sampling design and size of the sample.

* Larger samples have distributions with smaller spread (variability).
* Samples of the *same size* from large populations of *different size* will have sampling distributions with approximately the same variability.

Confidence Intervals

An interval, calculated from a sample, which is believed to contain the true mean of a population (μ).

Standard form : estimate ± margine of error

Variables & Definitions

CI – confidence interval

C – confidence level (in %)

z^* – upper critical value $(1-C)/2$

Confidence Interval for μ (σ known)

$$CI = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Note : z^* is used to determine confidence intervals when σ is known for a population (unrealistic) or when working with a sample proportion $\hat{p}$.

Smaller CIs give more accurate approximations of μ.

Example :

Given : a 95% CI for μ is calculated to be 6.2 ± 1.1 or (5.1, 7.3)
We are 95% confident that the population mean μ lies within the interval (5.1, 7.3).
This interval was calculated by a method that gives correct results in 95% of all samples.
Does not imply :
The probability is 95% that the population mean lies in (5.1, 7.3).

Note : $z^* \frac{\sigma}{\sqrt{n}}$ is called the margin of error

If the confidence level C increases, z^*, the margin of error, and the CI decrease.

If the sample size n increases, the margin of error and the CI decrease.

Sample Mean

Variables & Definitions

μ – population mean (parameter)

σ – population standard deviation (parameter)

$\bar{x}$ – sample mean (statistic)

s – sample standard deviation (statistic)

n – sample size

Sampling Distribution of $\bar{x}$

* The sampling distribution of $\bar{x}$ has a normal distribution when the sample is from a population that has a normal distribution.

  or

  The Central Limit Theorem states that the sampling distribution of $\bar{x}$ is *close* to a normal distribution when the sample is taken from *any population* whatsoever.

* The mean of the sampling distribution is exactly μ.

$$\mu_{\bar{x}} = \mu \qquad (\bar{x} \text{ is an unbiased estimator of } \mu)$$

* The standard deviation of the sampling distribution:

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

Assumption: *population size* $\geq 10n$ must be verified before using standard deviation formula.

Notes: a sample of size $4n$ will decrease the standard deviation by 1/2.
larger samples are needed for non-normal populations.

Normal approximations allow the use of:

z – tables to find probabilities under a normal distribution curve or the use of calculator function Normalcdf (min, max, μ, σ)

Determining Sample Size

Given a "desired" margin of error m, find the sample size needed to assure that the CI will have a margin of error less than or equal to m.

1) Set the margin of error less than or equal to m.
2) Solve the inequality for n.
3) The minimum sample size is the integer greater than or equal to n.

When σ is known use: $$z^* \frac{\sigma}{\sqrt{n}} \leq m$$

When σ is unknown use: $$z^* \frac{s}{\sqrt{n}} \leq m$$

With proportions use: $$\sqrt{\frac{p^*(1-p^*)}{n}} \leq m$$

p^* is a guessed value for the sample proportion.
If no guess is appropriate, use $p^* = 0.5$

Significance Testing

Assesses the evidence provided by data against a null hypothesis in favor of an alternative hypothesis H_a.

Outline of a significance test

1) Verify assumptions
2) State hypotheses
3) Calculate the test statistic and p - value
4) Reject or do not reject H_0
5) Conclusion statement

Chosing Test Statistics

Use if σ is known or when using proportions.

z – test statistic

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Use when σ is unknown.

t – test statistic

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \qquad df = n - 1$$

p – values

The area under the density curve outside of the critical value(s).

The p - value represents how likely results at least as extreme as the ones observed would be if the null hypothesis were true.

Reject H_0 (rules)

Reject H_0 if the p - value is less than the significance level.

Reject H_0 if the p - value is extremely small (no α given).

Reject H_0 if the test statistic to the outside of the critical value.

Inference for the Mean of a Population

Using the sample mean $\bar{x}$ from an SRS to calculate confidence intervals and significance tests for the population mean μ.

Assumptions for inference about a mean :

* Data are obtained using a SRS.
  (usually indicated in the problem)
* Observations from the population are normally distributed.
  (check using a stem & leaf plot or histogram)
* Parameters μ and σ are not known.

When to use t procedures :

$n < 15$ - use t procedures if data are close to normal.
$n \geq 15$ - use t procedures unless an outlier or strong skewness.
$n \geq 40$ - use t procedures for any data unless an outlier exists.

Single Sample Confidence Interval

$$CI = \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$df = n - 1$ (degrees of freedom)
t^* - based on right tail area

Single Sample Test Statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

To test $H_0 : \mu = \mu_o$
against $H_a : \mu > \mu_o$
$H_a : \mu < \mu_o$
or $H_a : \mu \neq \mu_o$

Type I error

Falsely rejecting H_0.

If we reject H_0, when in fact H_0 is true.

Type II error

Falsely accepting H_0.

If we do not reject H_0, when in fact H_0 is false.

		Truth about the population.	
		H_0 true	H_a true
Decision based on sample	reject H_0	Type I error α	Power $1-\beta$
	do not reject H_0	$1-\alpha$	Type II error β

The significance level α of any fixed level test is the Type I error.

Power

The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true.

The higher the power (probability), the more sensitive the test is in detecting that an alternative value of the parameter is true.

The graph below has a fixed μ_0 and μ_a (this is one possible setting).

Matched Pairs Procedures

(for Two - Sample Inference)

A matched pairs t - procedure is performed when two sets of data are produces for each individual (before & after).

Procedure :

The two sets of data must be dependent.

A single set of differences (for each individual) is produced.

A single sample t - procedure is performed on the set of differences.

Alternative Two - Sample Test Statistic

Variances (or s) equal (or very close) - pooled

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{where } \sigma = \sqrt{\frac{(n_1 - 1)s_1^2 - (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

To test $H_0 : \mu_1 = \mu_2$

against $H_a : \mu_1 < \mu_2$

$H_a : \mu_1 > \mu_2$

or $H_a : \mu_1 \neq \mu_2$

Pooling is used when the variances of two independent samples are the same or very close to the same. This procedure "pools" (averages) the two variances to estimate the common population variance.

Two - Sample Inference for the Mean

To compare the responses to two treatments or to compare the characteristics of two populations.

Assumptions for two - sample inference :

The two SRSs are from two distinct populations.
(The samples are independent)
Both populations are normaly distributed.
Parameters μ and σ are not known for either population.

Two – Sample Confidence Interval

Variances (or s) unequal - not pooled

$$CI = \left(\bar{x}_1 - \bar{x}_2\right) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$df =$ the smaller of $n_1 - 1$ or $n_2 - 1$
or df by TI - 83 (formula).

Two - Sample Test Statistic

Variances (or s) unequal - not pooled

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

To test $H_0 : \mu_1 = \mu_2$
against $H_a : \mu_1 < \mu_2$
$H_a : \mu_1 > \mu_2$
or $H_a : \mu_1 \neq \mu_2$

Single Sample
Inference for a Population Proportions

Variables & Definitions

p – population proportion (parameter)

$\hat{p}$ – sample proportion (statistic)

$\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ – sample standard deviation (statistic)

Assumptions for Confidence Intervals

Data are a SRS from the population of interest
(or observations are independent)

Population size $> 10n$

$n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

Single Sample Confidence Interval

$$CI = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \hat{p} = \frac{x}{n} = \frac{\text{number of } \textit{successes}}{\text{sample size}}$$

z^* - is the upper $\alpha/2$ critical value.

Assumptions for Test Statistic

Data are a SRS from the population of interest
(or observations are independent)

Population size $> 10n$

$np_0 \geq 10$ and $n(1-p_0) \geq 10$

Single Sample Test Statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

Two - Sample
Inference for a Population Proportions

Assumptions for Confidence Intervals

Data are a SRS from the population of interest
(or observations are independent)
Population size $> 10n$
$n\hat{p}_1 \geq 5,\ n(1-\hat{p}_1) \geq 5,\ n\hat{p}_2 \geq 5,$ and $n(1-\hat{p}_2) \geq 5$

Two - Sample Confidence Interval

Variances (or s) unequal - not pooled

$$CI = (\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Assumptions for Test Statistic

Data are a SRS from the population of interest
(or observations are independent)
Population size $> 10n$
$n\hat{p}_1 \geq 5,\ n(1-\hat{p}_1) \geq 5,\ n\hat{p}_2 \geq 5,$ and $n(1-\hat{p}_2) \geq 5$

Two - Sample Test Statistic

Variances (or s) equal - pooled
(all samples come from the same population)

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\hat{p} = \frac{\text{count of successes in both samples combined}}{\text{count of observations in both samples combined}}$$

For a Test of $H_0: p_1 = p_2$ and $H_a: p_1 \neq p_2$

Inference for Categorical Data

Chi - square Test of Independence

Test of Independence - the two - way table application is typically used to test whether there is a relationship between two categorical variables (also refered to as a test for independence - 2 variables)

Assumptions for a χ^2 – test

Data are an SRS from the population of interest.
All expected values are at least 1.
No more than 20% of the expected counts are less than 5.

Hypotheses used in Chi - square Test

H_0 – there is no relationship between 2 categorical variables. (*Independence*)
H_1 – there is a relationship between two categorical variables. (*Dependence*)

Calculating a Chi - square Test Statistic

* Use a Calculator with Matrix mode to find χ^2 – test statistic

By hand - use a table to find test value

$$\chi^2 = \sum \frac{(O-E)^2}{E} \qquad df = (\#\,rows - 1)(\#\,columns - 1)$$

Expected Value Formula $E = \dfrac{\text{(row sum)(column sum)}}{\text{grand total}}$

	Variable #2 choice 1 choice 2 choice 3 etc.	Total
Variable #1 choice 1 choice 2 choice 3 etc.		
Totals		**Grand Total**

Inference for Categorical Data

Chi - square Testing

Goodness of Fit Test - used to determine if a population (sample) fits a given specified form usually stated in percents.

Assumptions for a χ^2 – test

Data are an SRS from the population of interest.
All expected values are at least 1.
No more than 20% of the expected counts are less than 5.

Hypotheses used in Chi - square Test

H_0 : the actual population percents are *equal* to the hypothesized percentages.
H_a : the actual population percents are *different from* the hypothesized percentages.

Chi - square Test Statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

with degrees of freedom df = number of categories - 1

O - observed counts (frequency)
Usually stated as counts from a sample.

E - expected counts
Usually stated as a percent (change to a count)

Calculating a Chi - square Test Statistic

(Use a table to find test value)

O (observed)	E (expected)	$\frac{(O-E)^2}{E}$
		The sum of this row. $\chi^2 = \sum \frac{(O-E)^2}{E}$