Research Methods

Robert E Simpson

Singapore University of Technology & Design

robert_simpson@sutd.edu.sg

October 9, 2018

Week 5 OBjectives

You should be able to

- Understand when to use the word signifficant
- Form and write a statistical hypothesis
- Perform a hypothesis test
- Calculate the statistical confidence, α-risk, β-risk, and power of a statistical test
- Test whether a sample sample mean is representative of the population mean
- State a confidence limit in your assertion that the sample mean is representative of the population mean
- Know when to use normal, and t-distributions
- Apply: paired difference, pool variances, two independent sample, and single sample hypthesis tests

Review of CLT

The sample means (or sample sums) of independent variables X_1, $X_2, \ldots X_n$, of large enough sample size, n, from any arbitrary distribution is *approximately* normally distributed with an expectation value $\langle \bar{X} \rangle = \mu$.

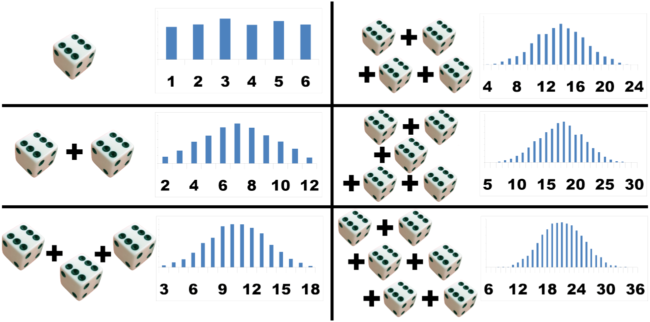
The standard error (standard deviation of the sampled distribution) is:

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

Where σ is the standard deviation of a random sample of n independent variables, and $\sigma_{\bar{x}}$ is the standard error in predicting the mean of the parent population using the random sample of n variable.

Review of CLT

The parent distribution can be any shape and we can still get a *Gaussian* sample distribution.



[When we discuss confidence intervals we are usually referring to the probability of picking a particular sample by chance, and therefore the confidence is referenced against the sample distribution.]

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Use of the word ‘significant’

Science/AAAS: Science Ma...

www.sciencemag.org/site/feature/contribinfo/prep/res/style.xhtml

Search

Most Visited | TierraBooks | untitled folder | MIT | SUTD | finn | Research | teaching | SG | Banks | Googl

- E-Mail Alerts & RSS Feeds
- Information for Authors
  - General Information for Authors
  - Contributors' FAQ
  - Preparing Your Manuscript and Figures
  - Submit an Original Manuscript
- Information for Reviewers
- Reprints & Permissions
- Magazine Help/FAQ
- Advertise in *Science*
- Terms and Conditions
- Privacy Policy

Contact the Journal

- Contact Information
- Feedback

mentioned in figure legends (or in the text where the figure is discussed) are missing from the corresponding figures.

- Avoid jargon; explain obscure terms and define acronyms (keep in mind that many potential readers of your work will not be specialists in your field).
- After introducing an acronym, use only the acronym.
- Use active voice when suitable, particularly when necessary for correct syntax (e.g., "To address this possibility, we constructed a λZap library . . .," not "To address this possibility, a λZap library was constructed . . .").
- Write concisely (e.g., "even though," not "in spite of the fact that").
- When two or more similar terms are used throughout text, either make the usage consistent or clarify the distinction(s), as appropriate.
- Avoid using "-fold" because expressions such as "20-fold smaller" are imprecise; use percentages, proportions, orders of magnitude, or "factor of" instead. (Exception: Usage such as "1000-fold excess" is appropriate.)
- Avoid using "times more," "times less" (see above).
- Use "significant" only when discussing statistical significance.
- Avoid using terms such as "novel," "first," or "Our laboratory has pioneered..." to describe the present work. Do not mention your own work in progress within the text (although a numbered reference to specific work submitted for publication is acceptable; see "in preparation" in reference examples below).
- Omit academic or other titles from affiliations and acknowledgments.

What is a Hypothesis?

Hypotheses

- Statistical tests are used to determine the likelihood that a an observation is related the phenomona
- A proposed explanation for a phenomenon
- For statistical tests **a hypothesis is just a claim**

E.g. An hypothesis that women live longer than men.

Hypothesis test procedure for single sample mean

i Formulate the hypothesis: the null, H_0, (no difference), and the alternate, H_A, (there is a difference).

ii Identify a test parameter (z-value or t-value) that will be used to assess the *null hypothesis*.

iii Determine the probability (p-value) that the *null hypothesis* can be true

iv Compare this probability against a pre-selected threshold probability significance level α (maybe 5%)

v Rule out the null hypothesis if the p-value $\leq \alpha$ and accept the *alternative hypothesis*.

"If P is low, H_0 must go"

Innocent until proven guilty

The statistical test is set-up such that the null hypothesis is preferentially favoured. There must be overwhelming evidence to support the alternative hypothesis.

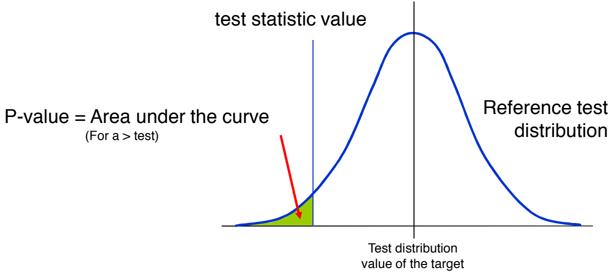
Similar to a jury trial, where the null hypothesis is innocent and it is the burden of the alternative hypothesis during testing to prove that the null hypothesis is guilty.

P-Value

Rigorous Definition The likelihood that the test statistic value observed occurs by chance.

Practical Definition The chance of being wrong when you reject the null hypothesis.

If p is less than α, then reject the null hypothesis!



Test vs Reality

	H_0 True	**H_0 False**
Accept H_0	Correct Acceptance Confidence (1-α)	Type II Error Beta Risk β
Reject H_0	Type I Error Alpha Risk α	Correct Rejection Statistical Power $1-\beta$

Usually the $\beta >> \alpha$. There must be overwhelming statistical evidence for acceptance of H_A

Test vs Reality

Confidence The probability of correctly accepting H_0

漏报 α risk "False Negatives". The risk rejecting the null hypothesis, H_0, when the null hypothesis is true.

β risk "False Positives". The risk of accepting the null hypothesis, H_0, when the alternate hypothesis is true.

Statistical Power The probability of correctly rejecting the null hypothesis.

The only way to decrease both errors is to use a larger sample size such that the standard error becomes smaller and $\bar{x} \to \mu$.

Concept Question 5.1

Car Brakes

There are two types of car brakes, *Type A* and *Type B*. Type A are the standard brakes put on all cars. The manufacturer is considering to switch to Type B, because they heard a rumour that they have a higher performance.

(a) Write an expression for H_0.

(b) Write an espression for H_a.

(c) Describe the β-risk

(d) Describe the α-risk

Concept Question 5.1 Answer

Clearly, in this situation the car manufacturer doesn't want to put worse brakes on the car because it could cause more accidents. Therefore the test needs to be biased toward accepting the null hypothesis, and there needs to be overwhelming evidence to change to Type B brakes.

(a) H_0 :: Type A $\geq$ Type B

(b) H_A :: Type A $<$ Type B

(c) β-risk, the risk that the test shows Type A $\geq$ Type B, when actually Type B $>$ Type A (AKA a false positive)

(d) α-risk, the risk that the test shows Type A $<$ Type B, when actually Type A $\geq$ Type B (AKA a false negative)

The α-risk is set a level decided by the researcher conducting the test. Typically, $\alpha < 0.05$, so there is less than 5% chance that the hypothesis test produces the wrong verdict.

β-risk and sample size

The probability of a β error depends on the sample size.

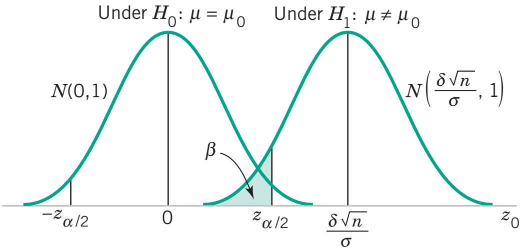
Consider the 2-tail test: $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$.

If the H_0 is false, then the true mean is $\bar{X} = \mu_0 + \delta$, where $\delta > 0$, then the expected text statistic Z_0 is:

$$E(Z_0) = E\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}\right) = \frac{\mu_0 + \delta - \mu_0}{\sigma/\sqrt{N}} = \frac{\delta\sqrt{N}}{\sigma}$$

β-risk and sample size

Suppose $Var(Z_0) = 1$, then $Z_0 \sim N\left(\frac{\delta\sqrt{N}}{\sigma}, 1\right)$



The β-error is the probability that $-z_{\alpha/2} \geq Z_0 \leq z_{\alpha/2}$ given that H_A is true.

$$\beta = P(z_{\alpha/2} - \frac{\delta\sqrt{N}}{\sigma}) - P(-z_{\alpha/2} - \frac{\delta\sqrt{N}}{\sigma})$$

Case problem 5.1

Light Bulbs

The traditional process for light bulb manufacture results in bulbs with a mean lifetime of $\mu = 1200$ h and a standard deviation of $\sigma = 300$ h. A new manufacturing process is developed. Assuming a significance level for the test to be $\alpha = 0.05$. For a sample size of 100 bulbs, what would be a statistically significant longer lifetime for bulbs manufactured by the new process?

If for a similar sample of 100 bulbs, $\bar{x} = 1260$ h. Do we accept null or alternative hypothesis?

Case problem 5.1

Light Bulbs

1. The *null hypothesis* is that the new process is no better than the previous one. i.e. H_0: $\mu = 1200$ h. The alternative hypothesis is $H_a > 1200$ h.
2. Set the significance level to 0.05 with a one tailed test (the new process should have a longer life) which corresponds to z=1.64
3. The null hypothesis is true if:

$$z_c \geq \frac{\bar{x}_c - \mu_0}{\sigma/\sqrt{n}}$$

$$\bar{x}_c \leq \mu_0 + \frac{S}{\sqrt{n}} z_c$$

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Case problem 5.1

Light Bulbs

4. We estimate the standard deviation of the sample is the same as the standard deviation of the population

$$\sigma \sim S$$

and therefore

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

5. From the standard normal table: $z_c = z_{0.05} = 1.64$. Therefore the null hypothesis is true if:

$$x_c \leq 1200 + \frac{300}{\sqrt{100}} \times 1.64 = 1249.2$$

6. Therefore the alternative hypothesis would be accepted if the sample has a mean lifetime greater than 1250 hours.

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Case problem 5.1

Light Bulbs

If we have incorrectly accepted the alternative hypothesis, we would have been extremely unlucky. We took a sample of 100 bulbs that on average performed **significantly** better than the population mean.

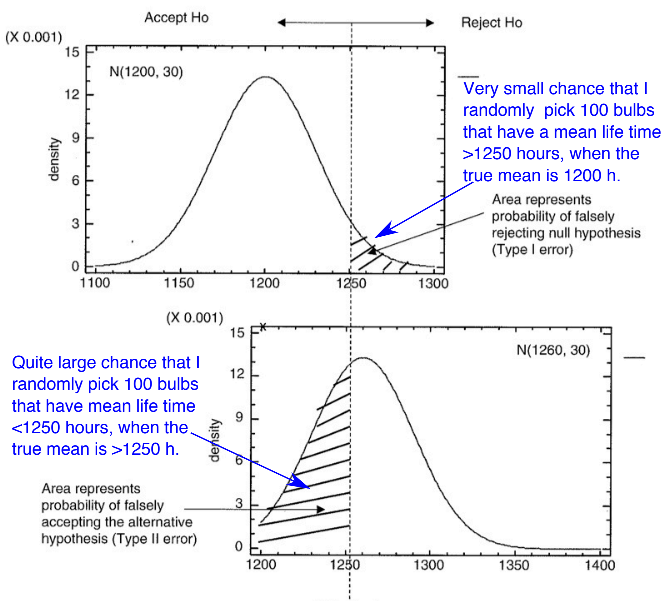
Case problem 5.1

Light Bulbs

Let's assume that we measured 100 bulbs and found $\bar{x} = 1260$, since this is greater than x_c, we would accept the alternative hypothesis.

Case problem 5.1

Light Bulbs

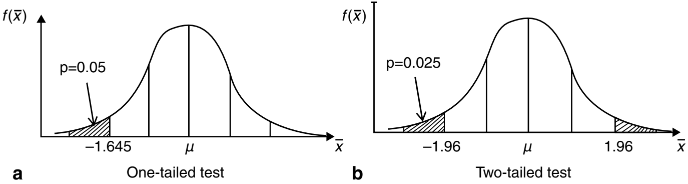


One and two tail tests

Two tail test Used to judge whether a sample is different from the specified population value

One tail test Used to judge whether a sample is greater than or less than the specified population value

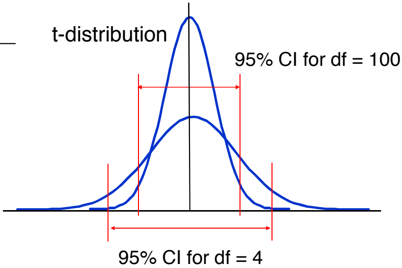
Both of the shaded areas in (a) and (b), below, represent a significance level of 0.05. Notice the different $\bar{X}_c$ that is used for different types of test.



Student t-distribution

- The t-distribution is a family of normal-shaped distributions that are dependent on sample size.
- The shape depends on the sample size n. The smaller n, the "flatter" the t-distribution, compared to z (standard normal).
- Degrees of freedom (usually, $df = n - 1$). The variables from a sample that are free to vary after the sample statistic has been calculated.
- 95% Confidence Intervals for different sample sizes:

Sample	t-value
5	2.78
10	2.26
20	2.09
30	2.05
100	1.98
1000	1.96



t-statistic

The t-statistic is the statistic associated with a t-distribution.

- Similar to a z-score for the normal distribution
- The t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{1}$$

All test statistics are very similar:

- They always have an interesting gap in the numerator (related to the deviation)
- They always have a measure of noise in the denominator
- A statistic is always a “signal to noise” measure

Selection of test statistic

z-statistic: If the population variance is known, and for sample sizes, $n > 30$, use the z-statistic with the standard normal tables

t-statistic: if the population variance is unknown or for sample sizes, $n < 30$, use the t-statistic and the Student t-tables with the appropriate degree of freedom. The sample statndard deviation, S, is used instead of the population variance, σ

Two independent sample tests on the means

Use to compare the means of two samples that are taken from two different populations.
It is assumed:

- the standard deviations of the two populations are reasonably close
- the populations are approximately normally distributed

The z-score is given by:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\bar{\mu}_1 - \bar{\mu}_2)}{\sqrt{\left(\frac{s_1^2}{n1} + \frac{s_2^2}{n2}\right)}} \tag{2}$$

Notice that in the denominator we have added errors according to the PoE. 分母

Two independent sample tests on the means

The confidence intervals at a particular critical level z_c for the difference in population means can be determined by:

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm z_c \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)} \tag{3}$$

where the standard error, $\sigma(\bar{x}_1, \bar{x}_2) = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm z_c \sigma(\bar{x}_1, \bar{x}_2) \tag{4}$$

Thus, analysing the difference between the **two** populations we have combined the statistical differences into **one** random variable, z, which can be analysed in the normal way.

Pooled variances

For small sample sizes where the variances of both populations are similar, we have the option to use a pooled variance, s_p, which is defined as:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{5}$$

with degrees of freedom (d.f$= n_1 + n_2 - 2$)

- The pooled variance is simply the weighted average of the two variances
- The use of pooled variances results in tight confidence intervals (hence its appeal)

now

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t_c \sigma(\bar{x}_1, \bar{x}_2) \tag{6}$$

where $\sigma(\bar{x}_1, \bar{x}_2) = \left[s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{1/2}$

Paired difference test

Used when measurements of samples from two different populations are correlated (called *paired samples*), i.e. when a one-to-one (dependent) correspondence exists between the factors that influence both populations.

For example:

- Measurements taken at the same time that are affected by the same time-varying variable
  - Differences at specific time have a smaller deviation that differences in collected at different times
- Measurements before and after an operation
- Measurements with and without treatments on the same units
- Measurements by operator one or operator two

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Paired difference tests method

- Rather than take the difference of the means of the two samples, **differences of paired observations are used**
- One can **perform a t-test on these differences** to determine if the mean is significantly different than a specified resolution (or z-test if the standard deviation of the population difference is known)
- The confidence interval that is calculated from paired data is narrower (because the time/operator/treatment- varying factor is reduced).

Paired difference tests

Let d_i be the difference between two individual readings of two small paired samples ($n < 30$). $\bar{d}$ is then the mean of the differences, tus the t-statistic is:

$$t = \frac{\bar{d}}{\sigma} \tag{7}$$

$$\text{where the standard error, } \sigma = \frac{s_d}{\sqrt{n}} \tag{8}$$

$$\tag{9}$$

and the confidence interval around $\bar{d}$ is:

$$\mu_d = \bar{d} \pm t_c(s_d/\sqrt{n}) \tag{10}$$

Case Problem 5.2

Energy efficiency of two similar buildings

In many countries buildings must meet energy efficiency regulations. Building B2 was awarded the energy efficiency certification but building B1 was deemed too energy inefficient to receive the certification. The owner of building B1 requested an energy consultant to compare the energy bills for the two buildings. See the table on the right. According to the data should the energy certificate be awarded to B1?

Month	Building B1 Utility cost ($)	Building B2 Utility cost ($)	Difference in Costs (B1 − B2)	Outdoor temperature (°C)
1	693	639	54	3.5
2	759	678	81	4.7
3	1005	918	87	9.2
4	1074	999	75	10.4
5	1449	1302	147	17.3
6	1932	1827	105	26
7	2106	2049	57	29.2
8	2073	1971	102	28.6
9	1905	1782	123	25.5
10	1338	1281	57	15.2
11	981	933	48	8.7
12	873	825	48	6.8
Mean	*1,349*	*1,267*	*82*	
Std. Deviation	*530.07*	*516.03*	*32.00*	

Case Problem 5.2

Energy efficiency of two similar buildings

Incorrect Analysis

- Choose to compare the monthly mean utility bills over the course of a year
- The dataset is small ($n < 30$) therefore use t-test
- The null hypothesis, H_0, is that $\mu_1 = \mu_2$
- Since the sample sizes are small we can pool the variances to get a better defined variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{11}$$

$$= \frac{(12 - 1)(530.07)^2 + (12 - 1)(516.03)^2}{12 + 12 - 2} \tag{12}$$

$$= 273,630.6 \tag{13}$$

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Case Problem 5.2

Energy efficiency of two similar buildings

Incorrect Analysis

- Compute the t-statistic using:
  $\mu_1 - \mu_2 = (x_1 - x_2) \pm t_c \sigma(\bar{x_1}, \bar{x_2})$

$$t = \frac{\bar{d}}{\sigma} = \frac{\bar{d}}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \text{ (notice that } \sigma \text{ comes from PoE)} \qquad (14)$$

$$t = \frac{(1349 - 1267) - 0}{\left[273,630.6\left(\frac{1}{12} + \frac{1}{12}\right)\right]^{1/2}} = 0.38 \qquad (15)$$

- The t-value is very small.
- the number of degrees of freedom is 22 (d.f=12+12-2)
- For d.f=22, the 90% CI is t=1.321
- The consultant would **mistakenly** report that insufficient statistical evidence exists to state that the two buildings are different

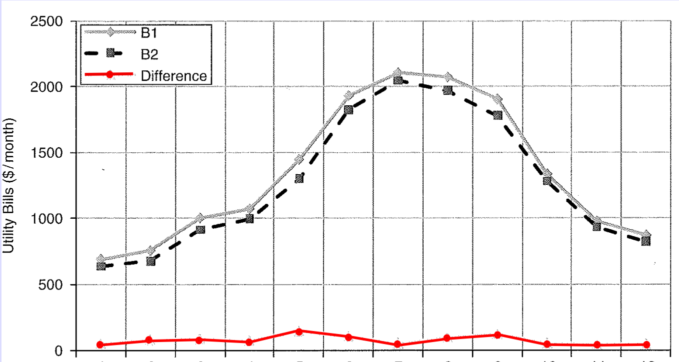
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Case Problem 5.2

Energy efficiency of two similar buildings

Correct Analysis

- Close observation of the data reveals that the utility bills rise and fall together because of seasonal variations in weather.
- The condition that the two samples are independent is violated.
- A paired difference test should be used



Case Problem 5.2

Energy efficiency of two similar buildings

Correct Analysis

- The test is meant to determine whether the monthly mean of the *differences* in utility charges if is zero or not.
- Hypothesise that there is no difference, H_0: $\bar{d} = 0$.
- Alternative hypothesis is that it is different from zero, H_A: $\bar{d} \neq 0$.

$$t = \frac{\bar{x}_D - 0}{S_D/\sqrt{n_D}} = \frac{82}{32/\sqrt{12}} = 8.88 \tag{16}$$

With degrees of freedom d.f=12-1=11

- A significance level of 0.05, suggests a critical level of $t_c = 2.201$ (two tailed test).

Case Problem 5.2

Energy efficiency of two similar buildings

Correct Analysis

- 8.88 $>>$ 2.201, therefore we can reject the null hypothesis and accept the alternative hypothesis.
- The buildings show a different energy performance, but how confident are we that Building B2 is **more efficient** than Building B1? –we should use a one tailed test
- Building B1 is less efficient than Building B2, even at significance level of 0.0005 or Confidence Level=99.95%. (at a significance level of 0.0005, t=4.437)

Misleading results can be obtained if inferential tests are misused!

Single Sample test

分数

Typically used to determine the fraction of a population who have a particular preference or own a particular piece of equipment.

For example SUTD campus development (CD) might perform a survey on a random sample of graduate students to ascertain the proportion that require an air conditioned working environment past 8pm.

SUTD CD should then extrapolate the results to understand the requirements of the full graduate population. They should state the 95% confidents limits of the proportion of student population that require A/C.
(This is an example of a binomial experiment, the answer is either ‘yes’ or ‘no’ for requiring A/C)

Single sample Test

We can use a Binomial Distribution. The number of successes, x, in n trials is the Bernouli random variable:

$$B(x; , n, p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{17}$$

with a mean: $\mu = np$ and variance: $\sigma^2 = np(1-p)$

If the population mean μ is known, then the proportion, p, is: $p = \frac{\mu}{n}$, where n is the total population. Similarly, the proportion variance is also divided by the population size:

$$S^2 = \frac{np(1-p)}{n} = p(1-p)$$

However, our survey is of a sample of the population, which gives us an estimator, $\hat{p}$, of the proportion, p: $\hat{p} = \frac{\text{number of successes}}{\text{number of trials}}$. Thus,

$$S^2 = \hat{p}(1-\hat{p})$$

Standard error of the proportion

Apply the CLT to the sample to find the standard error.

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

The standard error for a proportion is therefore:

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

Case Problem 5.3

In a random sample of n=1000 new condominium residences in Singapore, it was found that 630 had bathtubs. Find the 95% confidence interval for the fraction of condominiums that have bathtubs.

Case Problem 5.3 Solution

The proportion with bathtubs in the sample is: $\hat{p} = \frac{630}{1000} = 0.63$
The standard deviation is:

$$\sigma = \sqrt{\hat{p}(1-\hat{p})} = \sqrt{(0.63 \times 0.37)} = 0.48$$

The Standard Error is

$$\sigma_{SE} = \frac{\sigma}{\sqrt{n}} = \frac{0.48}{\sqrt{1000}} = 0.015$$

We can use Normal Distribution since $n > 30$, and the binomial

distribution is approximately normal. From the Normal Distribution tables: $z_{0.025} = 1.96$. Therefore the two tailed CI for p is:

$$0.63 - 1.96\left[\frac{0.63(1-0.63)}{1000}\right]^{1/2} < p < 0.63 + 1.96\left[\frac{0.63(1-0.63)}{1000}\right]^{1/2}$$

$$0.63 - (1.96 \times 0.015) < p < 0.63 + (1.96 \times 0.015)$$

$$0.60 < p < 0.66$$

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Summary

- In scientific writing Use the word ‘significant’ only when discussing a statistical difference. Otherwise use the word ‘substantial’
- Hypothesis testing allows one to establish a confidence in whether a sample is representative of the population
- There can be major differences between difference tests (paired diff vs sample tests on the means).