Week 8 Lecture Notes

A. Overview and Learning Objectives

- understand confidence intervals and how it is used to
  - derive an interval estimate
  - understand unknown population parameters (population proportion and mean)
- understand the 4 key steps in hypothesis testing
- apply hypothesis testing in carrying out the chi-squared test and the one-sample t-test

---

B. Statistical Inference

Important to have well-defined generalisability criteria when conducting sampling ⟹ is the result of the study representative of the population of the sample?

- sample statistics are subject to inaccuracies (*bias* of researchers / respondents + *random error*)
  - but want to **minimize these inaccuracies** to be as close as possible to *population parameter* ⟹ which is what we wish to ideally obtain

```
Sample Statistic = population parameter + bias + random error
```

- population parameter is a broad term, which may mean the "target value" we are trying to find out. This could be population mean μ in some cases, and population proportion p in other cases (context-dependent).

An **unbiased sample statistic** does not have selection, non-response and measurement errors/biases

1. Need to know the **survey methodology** used to generate the sample
2. Need to also know the **statistical methods** used to infer finding(s) from the target population in question.
   1. can statistics from the sample level be generalised / lead to similar conclusions at the population level?

Methods to reduce bias (recap)

1. Good Sampling Frame → zero selection bias
2. Use of probability-based sampling methods → zero selection bias
3. 100% response rate → zero non-response bias

> *def:* **Statistical inference** refers to the use of samples to draw **inferences or conclusions** about the population in question.

After EDA is completed, for a given sample, we need to cycle between:

1. Generating Questions
2. Visualization and Analysis of the variables in question
3. Answer Questions and if needed, refine them (fed back into point 1)

Advantages of Sample versus Census

1. **Cost:** census requires measurement of every unit in the population
   1. costly, have a chance of missing out certain groups
   2. very resource intensive
2. **Feasibility:** Instead of taking a small portion for "experiment", require to take everything (i.e. go overboard)
   1. example: Doctor needing to take all of a patient's blood for blood test instead of a small sample

Rule of Inference

> *def:* The *Fundamental Rule of Inference* states that available data can be used to make inferences about a much larger group if the data can be considered to be representative with regards to the question of interest.

- by adopting good sampling methods and good practices (i.e. having a good sampling frame), we can **greatly reduce selection bias** to be insignificant (i.e. selection bias $\implies$ 0).
- random error refers to the small differences arising as a result of *sample variability* when using any probability-based sampling method.

C. Confidence Interval

> *def:* A **confidence interval** is the range of values that is likely to contain a population parameter based on a certain degree of confidence.

- range of values in which the *true mean* may fall within
- allows sampling variability to be taken into consideration
- degree of confidence is
  - represented as a percentage (%)
  - termed as the confidence level (which is typically 95% or 99%)
  - refers to the long-run reliability of the method used to construct the interval (via repeated sampling)

For confidence intervals to be valid, they *have to utilize Simple Random Sampling* (SRS).

Focus is on the construction of *confidence intervals* for the **population proportion and mean.**

- we consider `flat_type` variable in the HDB resale dataset $\implies$ indicates the type of HDB resale flat (i.e. 1-room, 2-room ... 5-room, executive, multi-generational)

Formulas:

$$Raw\ Population\ Proportion_i = \frac{Frequency\ (i)}{Total\ Frequency}$$

$$Actual\ Population\ Proportion_i = Raw\ Population\ Proportion_i\ \pm\ random\ error$$

- *Note:* random error could be negative.

Confidence Interval Formula

$$CI = p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$$

p^* = sample proportion
z^* = z value from standard normal distribution (provides the lower and upper limits)
n = sample size
The $z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$ is known as the margin of *error* which impacts the **width** of the confidence interval

$\sqrt{\frac{p^*(1-p^*)}{n}}$ is the standard deviation of standard error.

z^* increases as the percentage confidence (confidence level) increases

- the wider the margin, the **more meaningless** the interval is

To have a more accurate confidence interval, we can increase the sample size n (to reduce margin of error).

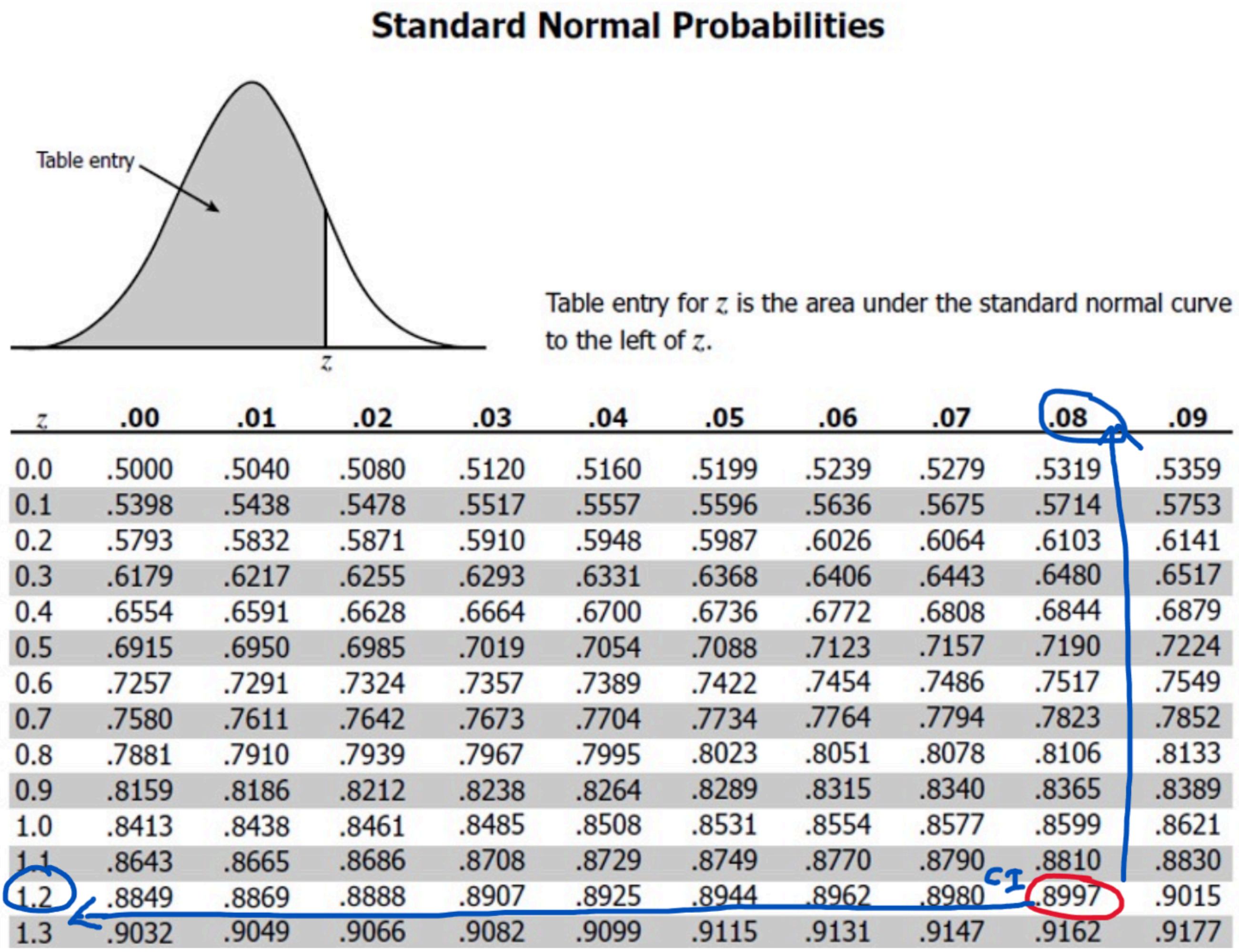
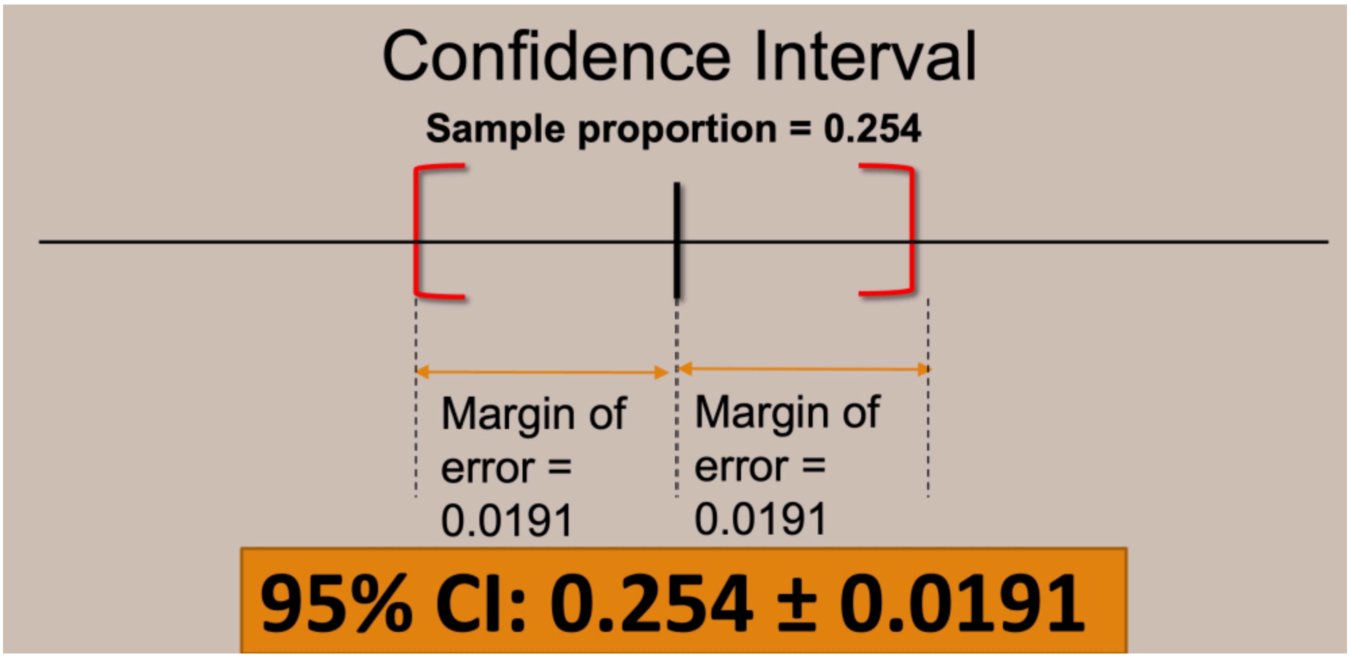


Table entry
z
CI

Standard Normal Probabilities

Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177



Confidence Interval

Sample proportion = 0.254

Margin of error = 0.0191

Margin of error = 0.0191

95% CI: 0.254 ± 0.0191

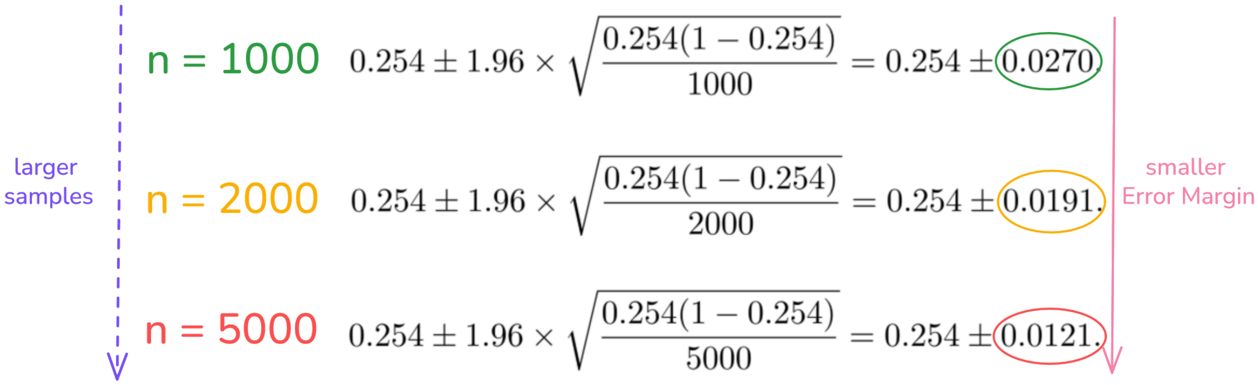
Common Mistake

Claim that there is a 95% chance that the population proportion of a 5-room resale HDB flat lies between 0.235 and 0.273.

1. Not correct because the **population proportion p is fixed**, does not vary $\implies$ no probabilistic element in what the proportion is going to be
    1. p is either (a) INSIDE the interval or (b) NOT INSIDE the interval
2. For a particular sample, the confidence interval constructed only depends on the sample proportion and the corresponding z^* value and therefore the CI is also "**fixed**" and there is no probabilistic element to it

Properties of Confidence Intervals

1. When a sample is taken with *the same sampling frame, sample sampling method (SRS)* but **smaller sample size**
    1. The resultant CI will be **larger** than the one with the larger sample size
    2. Larger Sample size = Smaller Random Error (Margin)



larger samples

n = 1000 $0.254 \pm 1.96 \times \sqrt{\frac{0.254(1-0.254)}{1000}} = 0.254 \pm 0.0270$

n = 2000 $0.254 \pm 1.96 \times \sqrt{\frac{0.254(1-0.254)}{2000}} = 0.254 \pm 0.0191$

n = 5000 $0.254 \pm 1.96 \times \sqrt{\frac{0.254(1-0.254)}{5000}} = 0.254 \pm 0.0121$

smaller Error Margin

2. **Confidence Level** impacts the confidence intervals
    1. i.e. Confidence level of 90% vs 95% affects the z^* value and hence the overall computation of the confidence interval.
        - z^* for 95% is `1.96`
        - z^* for 90% is `1.645`
    2. Lower z^* value results in a *narrower interval*

- 95% confidence interval: when we repeat the experiment again and again, **about 95 out of 100** of the intervals contain the population parameter

Population Mean μ Formula

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

- $\bar{x}$: sample mean
- t^* : "t-value" from t-distribution table
- s : sample S.D.
- n : sample size
- the margin of error (i.e. the stuff behind $\pm$) is a way to **quantify the random error**

Summary

- confidence intervals are used to quantify random error present in **every sample**
    - including SRS experiments where a level or bias can be reduced or considered as negligible or insignificant
- confidence intervals and the confidence level used to compute the interval can be understood using repeated sampling
    - avoid using terms like "chance" or "probability" when considering if population parameter lies within the **confidence interval** constructed from a **single sample**
- properties of confidence intervals (see above)
    - intervals are based on the sample (could vary from sample to sample)
    - population parameter of interest is an **unknown constant value** (it doesn't change and there is no probabilistic element in it).
- how are confidence intervals constructed using **two** population parameters (i.e. population proportion and population mean μ) $\implies$ using software (refer to the labs)

D. Hypothesis Testing

- can try to use a sample statistic to infer a population parameter using the formula:

  ```
  Sample Statistic = population parameter + bias + random error
  ```

- when $bias \to 0$, $Sample\ Statistic\ = population\ parameter\ + random\ error$
- assumption that sample is taking from population using SRS technique, from a perfect sampling frame and no non-response bias
- we will only need to do a hypothesis test for a *sample*, **NOT** for an entire population!

> *def:* A **hypothesis test** is a statistical inference method used to decide if the data from a random sample is **sufficient** to support *a particular hypothesis* about a population.

- enables us to ask if the observed sample population deviates from the hypothesized population $\implies$ explainable via *chance variation*?

> def: A **typical hypothesis** about a population could be anything that we want to know about the population.

- can only prove what the null hypothesis is not

Types of Hypotheses

1. Is a population parameter x?
2. In the population, are the categorical variables A and B **associated** with each other?

Questions:

Do we need to reject out null hypothesis and does the sample proportion warrant it (is it sufficient to reject H_0)

Steps in Hypothesis Testing

1. Identifying the **question and the context**, stating the null (H_0) and alternative (H_1) hypotheses.
   1. H_0 the statement being tested, which makes a claim about current or historical population mean (μ)
   2. H_1 the statement to be adopted if the evidence disproves H_0.
2. Set the **significance level** of the test, which measures the threshold / tolerance of deviation from what is hypothesized
   1. can the deviation from the hypothesis to the actual sample reading be explained by chance variation?
   2. usually set as $5\%, 1$ or 10%.
   3. unless stated otherwise, take sig level to be 5% or 0.05
   4. the probability of observing value or more extreme in the direction of alternative hypothesis, given that the null hypothesis is true
      1. can be computed as $P(Reject\ H_0 \mid H_0\ is\ true)$
3. Use the sample to find the **relevant sample statistic**
4. With the sample statistic and the hypothesis, **calculate** the **p-value**
5. Make the **conclusion** of the hypothesis test. Reject or accept H_0?
   1. conclusion depends on **calculated p-value versus significance level** for the test

> *def:* The $p\text{-}value$ is the **probability** of obtaining a result as extreme or more extreme than our observation in the direction of the alternative hypothesis H_1, assuming H_0 is true.

Example Case Study 1 -- H-Test for Population Proportion

H_0

- case where observation explainable by chance variation
- population prop = 0.5
- can write as $H_0 : p = 0.5$

H_1

- population proportion < 0.5
- can write as $H_1 : p < 0.5$

Important Note that $H_0 \cap H_1 = \emptyset \implies$ one or the other true only!

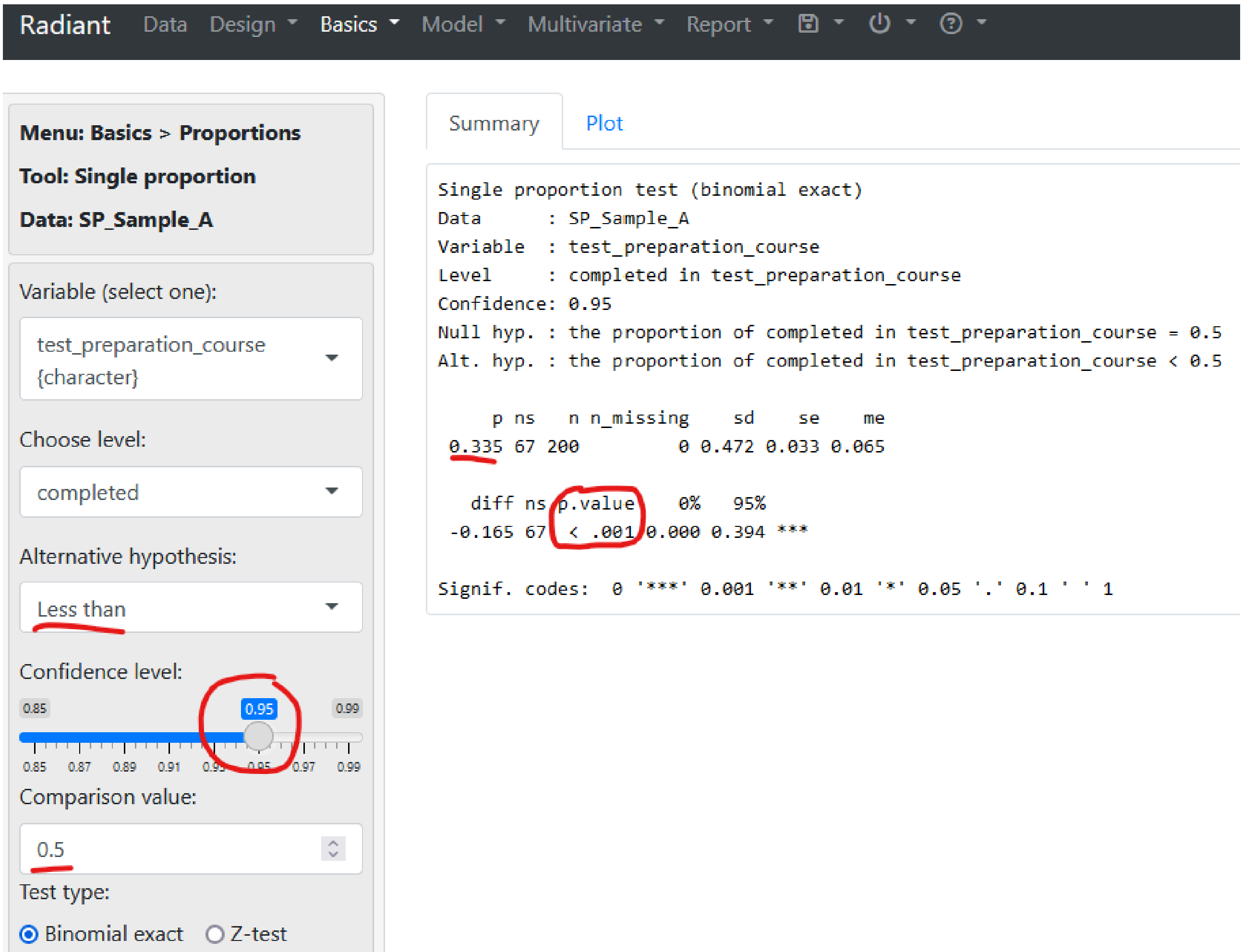
"null value" is the value you want to disprove" with hypo testing

Possible outcomes/train of thoughts
$T1$ - H_0 is valid despite the low sample proportion $p^* = 0.335$ (using `SP_Sample_A.csv`), as there is a chance of variation due to fewer students who completed the `test_prepration_course` being selected

$T2$ - H_1 is valid (H_0 invalid) because $p < 0.5$ and thus $p^* < 0.5$ as well.

As shown in the results below, we eventually reject H_0 and accept H_1 because

- The $p\text{-}value$ is smaller than 0.001



Menu: Basics > Proportions

Tool: Single proportion

Data: SP_Sample_A

Variable (select one): test_preparation_course {character}

Choose level: completed

Alternative hypothesis: Less than

Confidence level: 0.95

Comparison value: 0.5

Test type: ● Binomial exact ○ Z-test

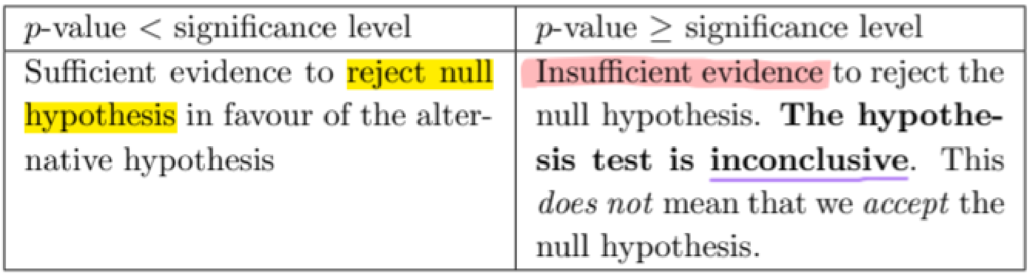
```
Single proportion test (binomial exact)
Data      : SP_Sample_A
Variable  : test_preparation_course
Level     : completed in test_preparation_course
Confidence: 0.95
Null hyp. : the proportion of completed in test_preparation_course = 0.5
Alt. hyp. : the proportion of completed in test_preparation_course < 0.5

     p ns   n n_missing    sd    se    me
 0.335 67 200         0 0.472 0.033 0.065

   diff ns p.value    0%   95%
 -0.165 67  < .001 0.000 0.394 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What to do based on `p-value` versus significance level.



p-value $<$ significance level	p-value $\geq$ significance level
Sufficient evidence to reject null hypothesis in favour of the alternative hypothesis	Insufficient evidence to reject the null hypothesis. **The hypothesis test is inconclusive.** This *does not* mean that we *accept* the null hypothesis.

Example Case Study 2 -- H-Test for Population Mean

H_0

- Population mean (μ) reading score $= 69$

H_1

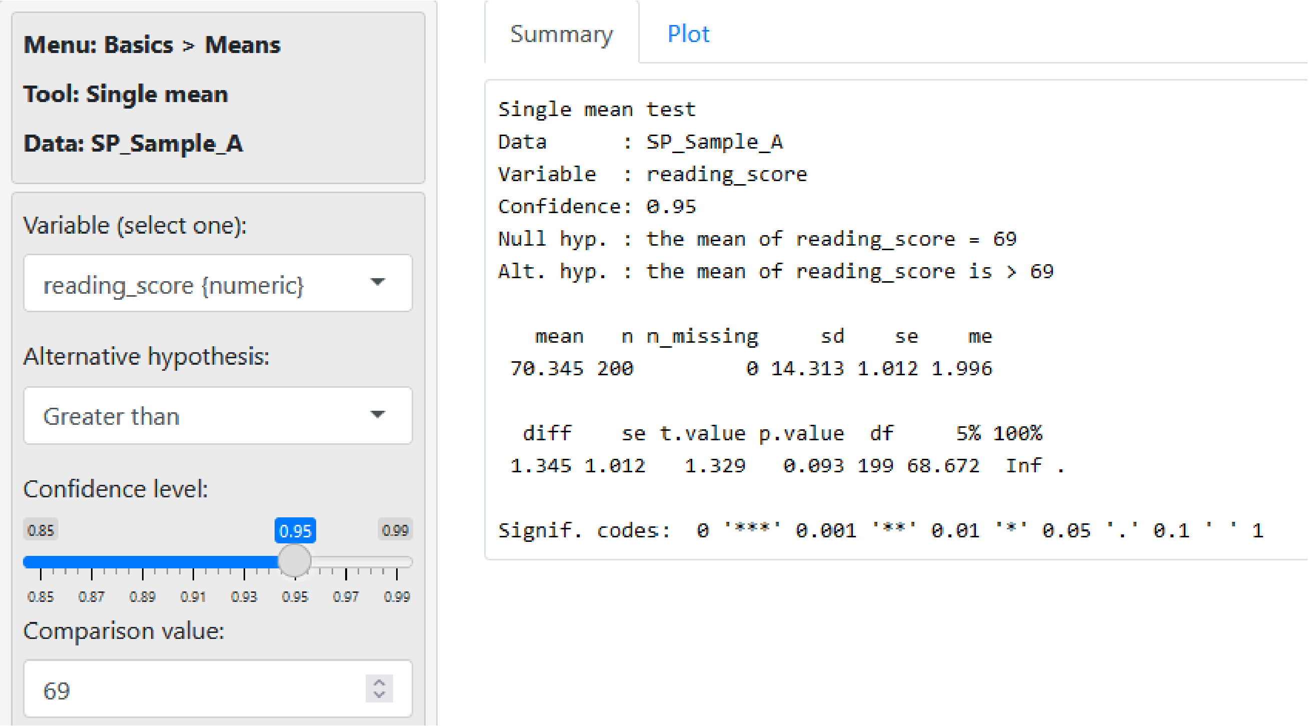
- Population mean (μ) reading score > 69

Possible outcomes/train of thoughts

$T1$ - H_0 is valid despite high sample mean of $\mu = 70.345$ observed due to chance variation and simply because there were more students who scored better in the reading test in the sample

$T2$ - H_1 is valid (H_0 invalid) because $\mu > 69 \implies \therefore$ sample mean $\bar{x} = 70.345 > 69$

$\therefore$, cannot reject H_0 since $p = 0.093 > 0.05$.



Menu: Basics > Means

Tool: Single mean

Data: SP_Sample_A

Variable (select one): reading_score {numeric}

Alternative hypothesis: Greater than

Confidence level: 0.95

Comparison value: 69

```
Single mean test
Data       : SP_Sample_A
Variable   : reading_score
Confidence: 0.95
Null hyp.  : the mean of reading_score = 69
Alt. hyp.  : the mean of reading_score is > 69

   mean   n n_missing     sd    se    me
 70.345 200         0 14.313 1.012 1.996

  diff    se t.value p.value  df     5% 100%
 1.345 1.012   1.329   0.093 199 68.672  Inf .

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example Case Study 3 -- H-Test for Association

H_0

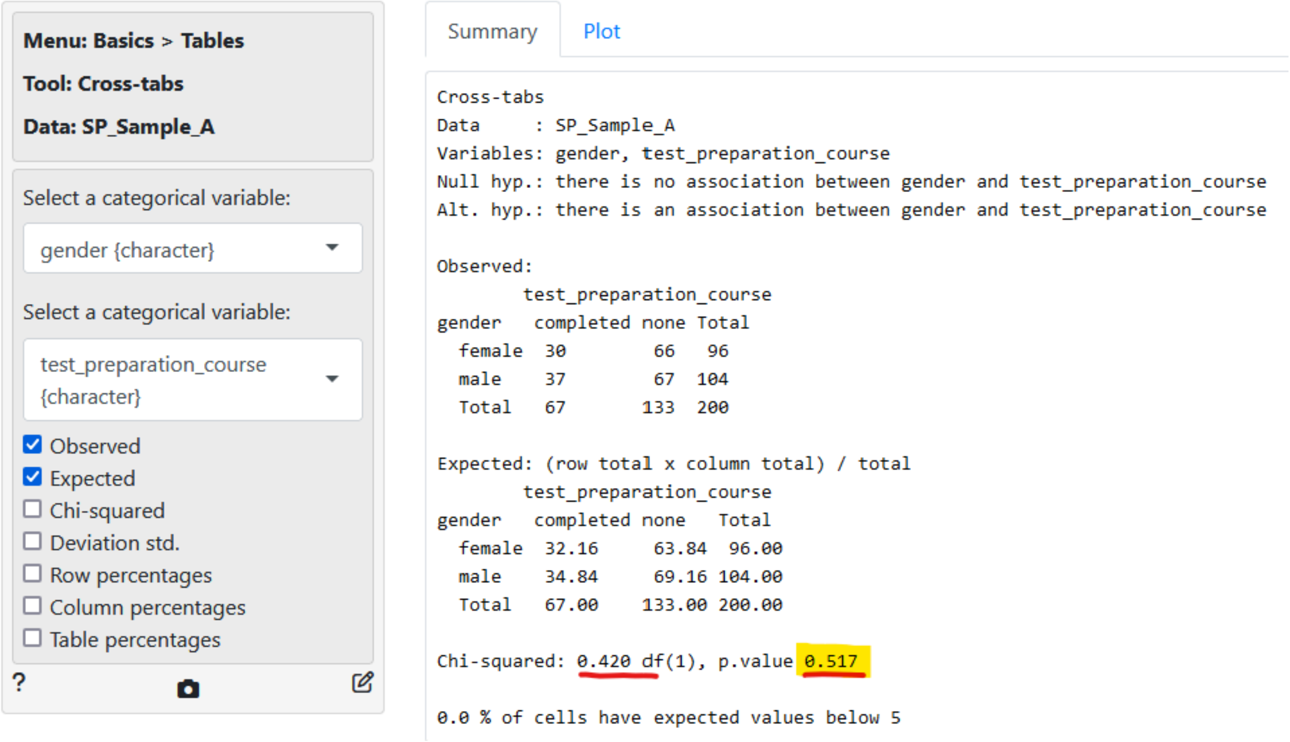
- no association between two variables at population level (i.e. `gender` and `test_preparation_course`)

H_1

- there is an association between two variables at population level (i.e. `gender` and `test_preparation_course`)

Do this using the chi-squared test for association (`Basics > Cross-tabs` in Radiant)'

- conclude that we cannot reject H_0 that there is no association (since insufficient evidence as per table above), since $p = 0.517 > 0.05/5\%$ (significance value)



Menu: Basics > Tables

Tool: Cross-tabs

Data: SP_Sample_A

Select a categorical variable: gender {character}

Select a categorical variable: test_preparation_course {character}

☑ Observed
☑ Expected
☐ Chi-squared
☐ Deviation std.
☐ Row percentages
☐ Column percentages
☐ Table percentages

```
Cross-tabs
Data     : SP_Sample_A
Variables: gender, test_preparation_course
Null hyp.: there is no association between gender and test_preparation_course
Alt. hyp.: there is an association between gender and test_preparation_course

Observed:
        test_preparation_course
gender   completed none Total
  female  30         66   96
  male    37         67  104
  Total   67        133  200

Expected: (row total x column total) / total
        test_preparation_course
gender   completed none   Total
  female  32.16      63.84  96.00
  male    34.84      69.16 104.00
  Total   67.00     133.00 200.00

Chi-squared: 0.420 df(1), p.value 0.517

0.0 % of cells have expected values below 5
```

Chi-squared tests

- how to look at the expected value versus what I have in practise
  - the further the difference, the more evidence to say that there is a relation of one categorical variable over another
- chi square variable $\implies$ sum of all 4 values in a `2 x 2` table.
- the lower the p-value, the increase of the likelihood where we reject H_0

Error Margin Calculation Formula (Excel)

```
=AVERAGE($F:$F)
=STDEV.S($F:$F)
=CONFIDENCE.T(0.05, J2, 200) # args are (1 - interval, cell, sample_size)
```