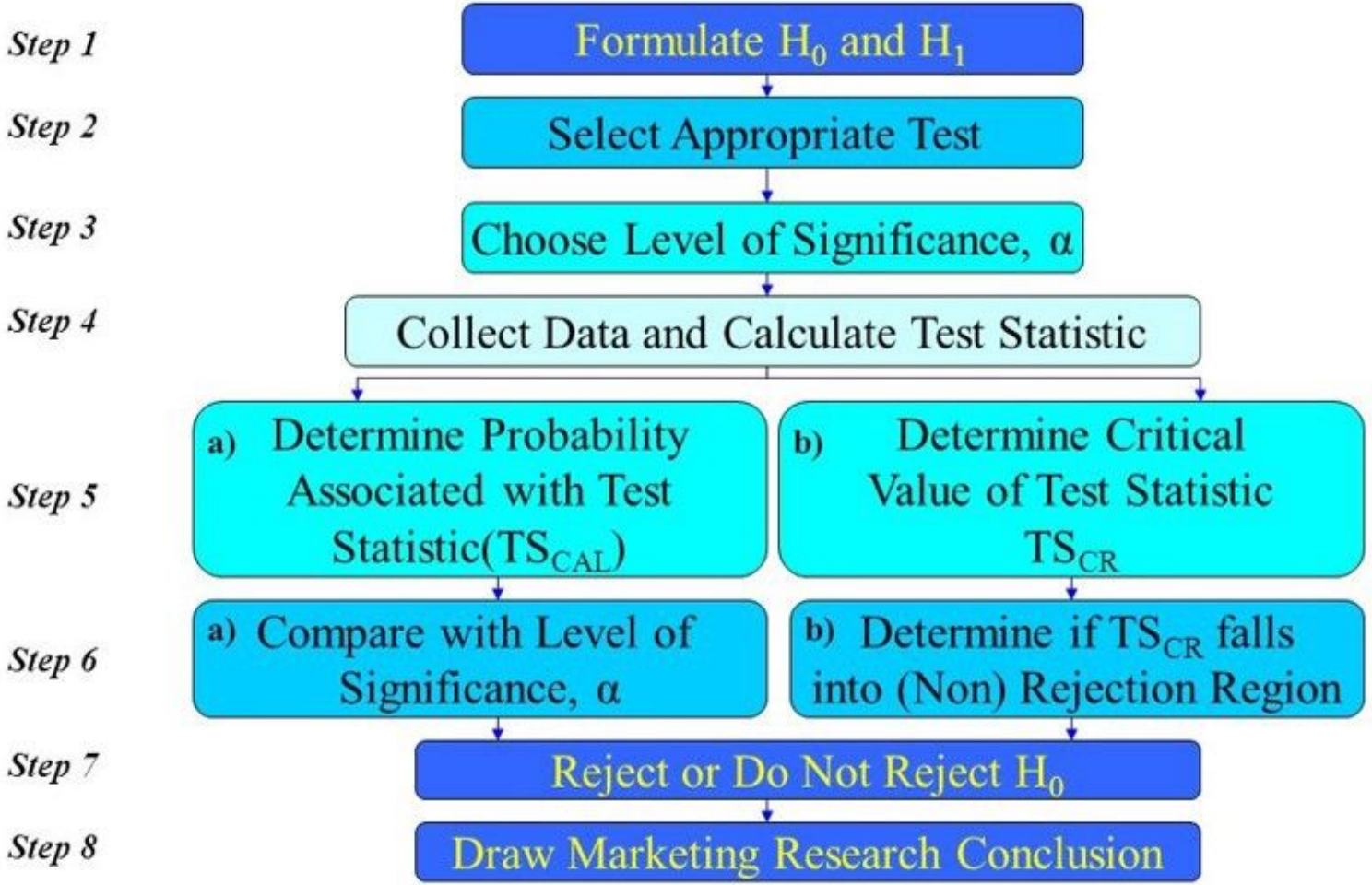
Statistics Module Review

Innis Cohort

Agenda

1. **Probability Distribution Functions** - 20 min
2. **When to Use Which Probability Distribution** - 10 min
3. **Probability Distribution Word Problems** - 20 min
4. **Null & Alternative Hypotheses** - 10 min
5. **Type I & Type II Errors** - 10 min
6. **When to Use Which Hypothesis Test** - 20 min
7. **How to Perform Hypothesis Tests** - 20 min

- There will be 5-10 minutes between each topic to ask any questions or get clarification.


Step 1
Formulate H_0 and H_1
Step 2
Select Appropriate Test
Step 3
Choose Level of Significance, α
Step 4
Collect Data and Calculate Test Statistic
Step 5
a) Determine Probability Associated with Test Statistic(TS_{CAL})
b) Determine Critical Value of Test Statistic TS_{CR}
Step 6
a) Compare with Level of Significance, α
b) Determine if TS_{CR} falls into (Non) Rejection Region
Step 7
Reject or Do Not Reject H_0
Step 8
Draw Marketing Research Conclusion

Probability Distribution Functions

Probability Distribution Functions

Where can I find information on scipy.stats distribution functions?

https://docs.scipy.org/doc/scipy/reference/stats.html?highlight=scipy%20stats

Scipy documentation is not clear about how to use these functions.
Seems like numpy is used under the hood.
Look at equivalent numpy functions for usage details.

Probability Distribution Functions

Why not use the numpy functions instead?

- You can, but consider your requirements.

Scipy distribution functions come with their own built in functions.

- .pdf, .pmf, .cdf, .ppf, .sf, .isf, .rvs, and more

Scipy distribution functions are actually objects.

- This isn’t important, but worth noting.

Probability Distribution Functions

For now we only need to know about these:

- Uniform distribution
- Normal distribution
- Binomial distribution
- Poisson distribution

Let’s see some code!

https://github.com/alegarcia-dev/statistics-exercises/blob/main/stats_review.ipynb

Probability Distribution Functions

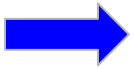
I **have** a given value

I **want** the probability

My data is **continuous** (decimal)

.pdf

My data is **discrete** (whole number)



.pmf

EXAMPLES:

.pdf (probability density functions) - probability of it taking a computer 2.3 sec to process data

.pmf (probability mass functions) - probability of rolling 3 with a dice.

Probability Distribution Functions

Less than or equal to:

I have a **value** and need a **probability** → **.cdf**

I have a **probability** and need a **value** → **.ppf**

EXAMPLES:

.cdf (cumulative density functions) - probability of rolling a **3 or lower** **.cdf(3)**
- probability of rolling l**ower than 3** **.cdf(2)**

.ppf (percent point functions) - there is a ⅓ chance a roll will produce **lower than what number**.

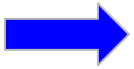
Probability Distribution Functions

Greater than or equal to:

I have a **value** and need a **probability**

.sf

I have a **probability** and need a **value**



.isf

EXAMPLES:

.sf (survival functions) - probability of rolling **greater than 5. .sf(5)**
- probability of rolling a **greater than or equal to 5 .sf(4)**

.isf (inverse survival functions) - there is a ⅓ chance a roll will produce **higher than** what number.

When to Use Which Probability Distribution Function

When to Use Uniform Distribution

- The probability of getting any value is equal.

- Examples
  - Rolling a die
  - Flipping a coin
  - Drawing a card from a shuffled deck

When to Use Normal Distribution

- When we are given a mean and standard deviation.

- When we are given values that can represent mean and standard deviation.

- Examples
  - Store sells on average 5000 products plus or minus 1000.
  - Steve usually eats 5 doughnuts with a standard deviation of 1 doughnut.

When to Use Binomial Distribution

- We have a probability of success and a number of trials.
- Examples
  - 100 true or false questions and we are guessing randomly.
  - ... there's a 3% chance that any one student cleans the break area ... 3 active cohorts of 22 students visit the break area.
  - There are approximately 3 web development cohorts for every 1 data science cohort at Codeup. Assuming that Codeup randomly selects an alumni to put on a billboard, what are the odds that the two billboards I drive past both have data science students on them?
  - There's a 70% chance on any given day that there will be at least one food truck at Travis Park. However, you haven't seen a food truck there in 3 days. How unlikely is this?

When to Use Poisson Distribution

- We have some average number of events happening over a given time interval.

- Examples
  - The number of phone calls received by a call center **per** hour
  - The number of decay events **per** second from a radioactive source
  - The number of emails sent by a mail server in a (**per)** day

Probability Distribution Word Problems

Probability Distribution Word Problems

Look for the relevant information.

- Are we given an average?
- What about a standard deviation?
- Do we have a number of trials?
- What about a probability of success?
- Are the odds fair?

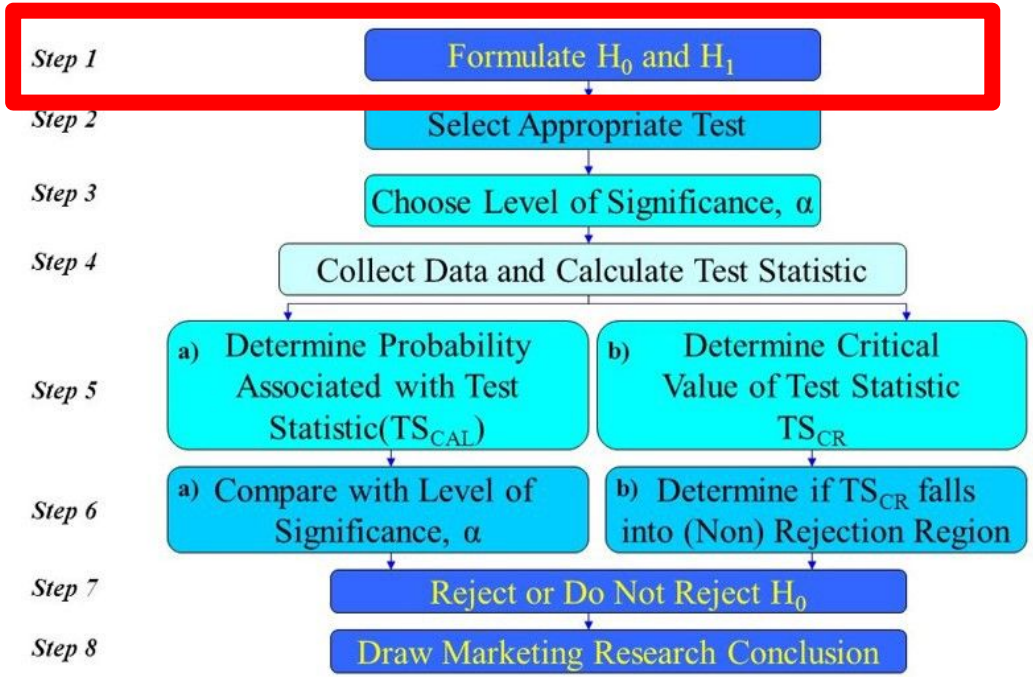
Probability Distribution Word Problems

Let’s see some code!

https://github.com/alegarcia-dev/statistics-exercises/blob/main/stats_review.ipynb

Null & Alternative Hypotheses

Null & Alternative Hypotheses



Null & Alternative Hypotheses

A hypothesis test evaluates two mutually exclusive statements about a population and informs us which statement is best supported by our sample data.

1. **H_0:** There is **no difference** between smokers' tips and the overall population's tip average. ⇨ **Null**

2. **H_a or H_1:** There **is a difference** between smokers' tips and the overall population's tip average. ⇨ **Alternative**

Null & Alternative Hypotheses

NULL (H_0)

There is ***NO*** difference
"status-quo"

- H_0: There **is no difference** in right-handed people and left-handed individual's heights.
- H_0: The amount of sleep a student gets the night before an exam makes **no difference** on the student's exam score.

ALTERNATIVE (Ha)

There ***IS*** a difference

- H_a: There **is a difference** in right-handed people and left-handed individual's heights.
- H_a: Less sleep the night before an exam leads to **a lower** exam score.

What does “no difference” mean

H_0	H_a
equal (=)	not equal (≠) **or** greater than (>) **or** less than (<)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)

Examples

Resource Link - Rachel's GitHub

1. Has the network latency gone up since we switched internet service providers?

1. - Null Hypothesis

There is no difference in network latency since changing internet providers.

1. - Alternative Hypothesis

There is a difference in network latency since changing internet providers.

1. - True Positive

There IS a difference in network latency and we conclude there is a difference in network latency.

1. - True Negative

There is NO difference in network latency and we conclude there is no difference in network latency

1. - Type I Error

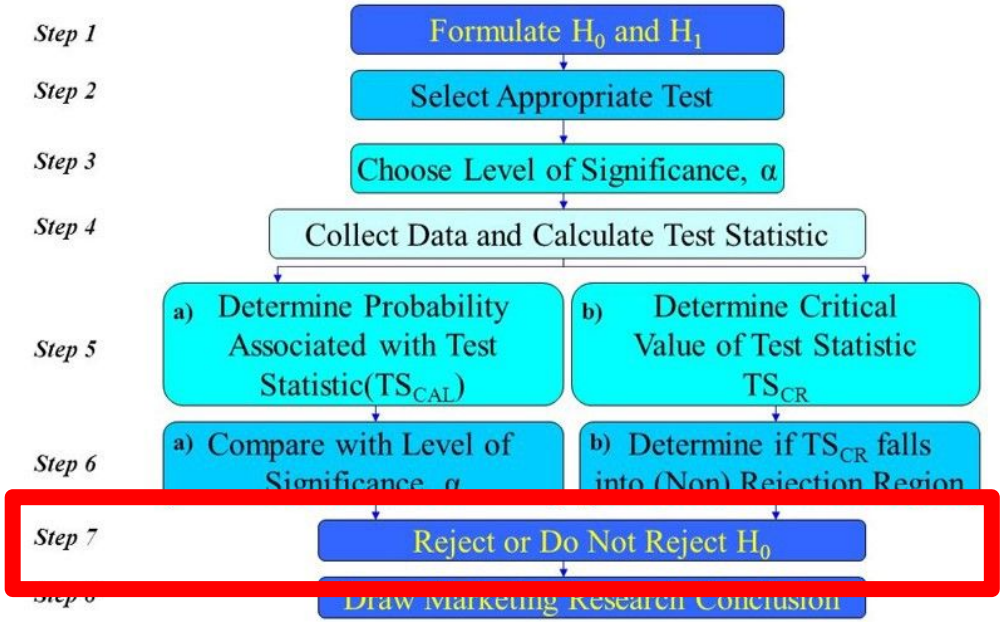
We reject (accepting as false) the hypothesis that there was no difference in latency, when in fact, there was no difference in latency, and the hypothesis was true.

1. - Type II Error

We failed to reject (accepting as true) the hypothesis that there was no difference in latency, when in fact, there was a difference in latency, and the hypothesis was false.

Type I & Type II Errors

Type I & Type II Errors



- After running and interpreting the values returned by the appropriate statistical test for my data, I will either **fail to reject** (accept) or **reject** the **Null Hypothesis.**

- In doing this I could lake an error.

Part One: Resource

Part Two: Resource

Type I & Type II Errors

Type I Error:

We REJECT the Null Hypothesis when it is TRUE.

- **We *should have* accepted (failed to reject) the Null Hypothesis.**

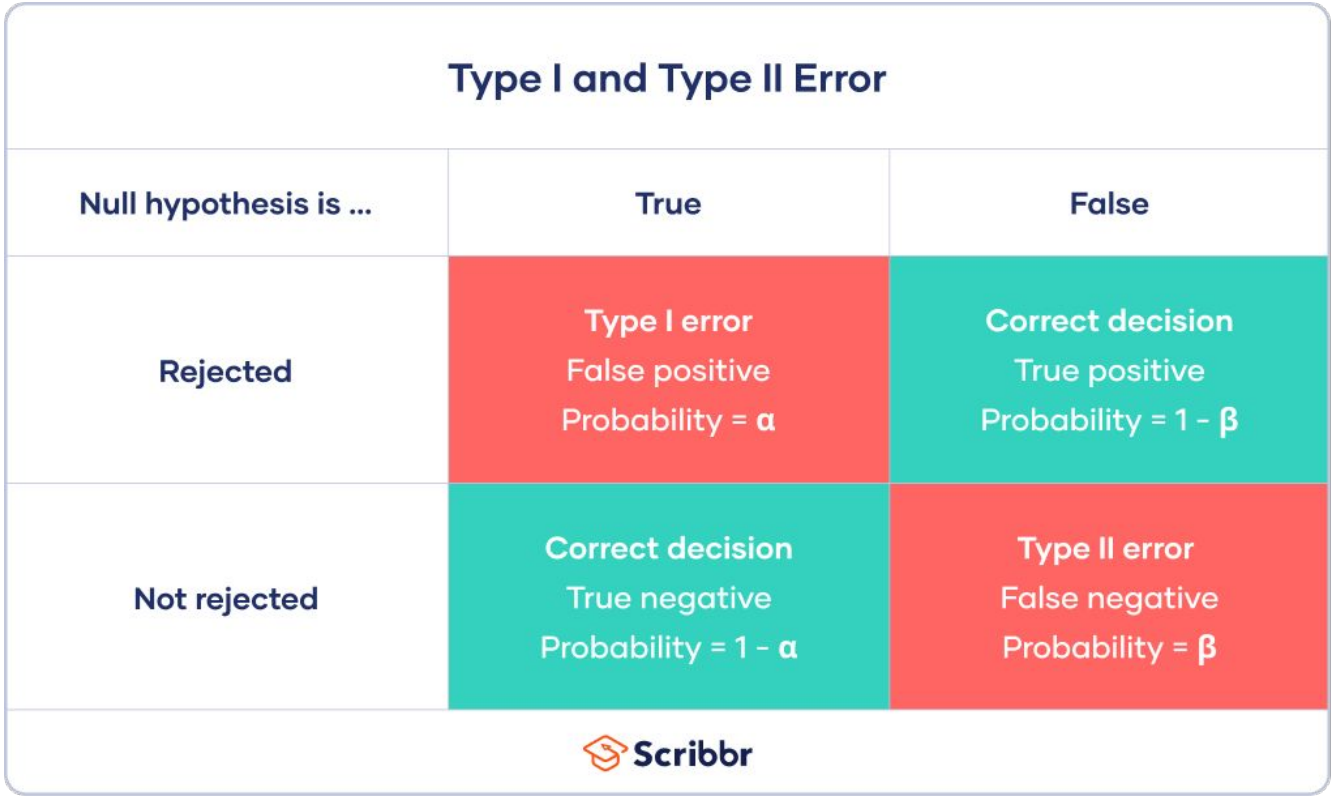
Type II Error:

We ACCEPT (*fail to reject)* the Null Hypothesis when it is FALSE.

- **We *should have* rejected the Null Hypothesis.**

Examples

[Resource]



Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	**Type I error** False positive Probability = α	**Correct decision** True positive Probability = 1 - β
Not rejected	**Correct decision** True negative Probability = 1 - α	**Type II error** False negative Probability = β

Scribbr

Practice

[Resource Link - Jesse's GitHub]

Is the website redesign any good?

H0: The old and new website versions are no better than each other, say for attracting traffic or clicks

Ha: The new website version is better than the old, say at attracting traffic/clicks

True Negative (Accept H0 and H0 is True): Accepted that the two versions of the website are no better than each other and that is true

Accept H0 but H0 is false (False negative): Accepted that the two versons are no better from each other but actually the redesign (new version) is better

Type II

Reject H0 but H0 is True (False positive): We rejected that the two versions are no better from each other (so we thought the redesign is better) but actually they are no better from each other

Reject H0 and H0 is False (True positive): We rejected that the two versions are no better from each other (accepted that the redesign is better) and the new version is in fact better

Type I

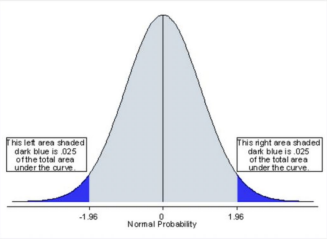
When to Use Which Hypothesis Test

Hypothesis Tests

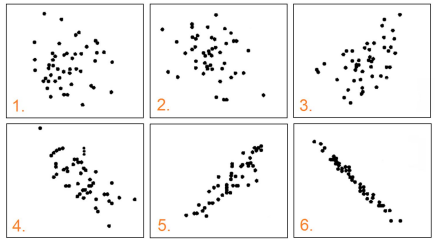
Comparing means

Determining linear relationships

T-Test

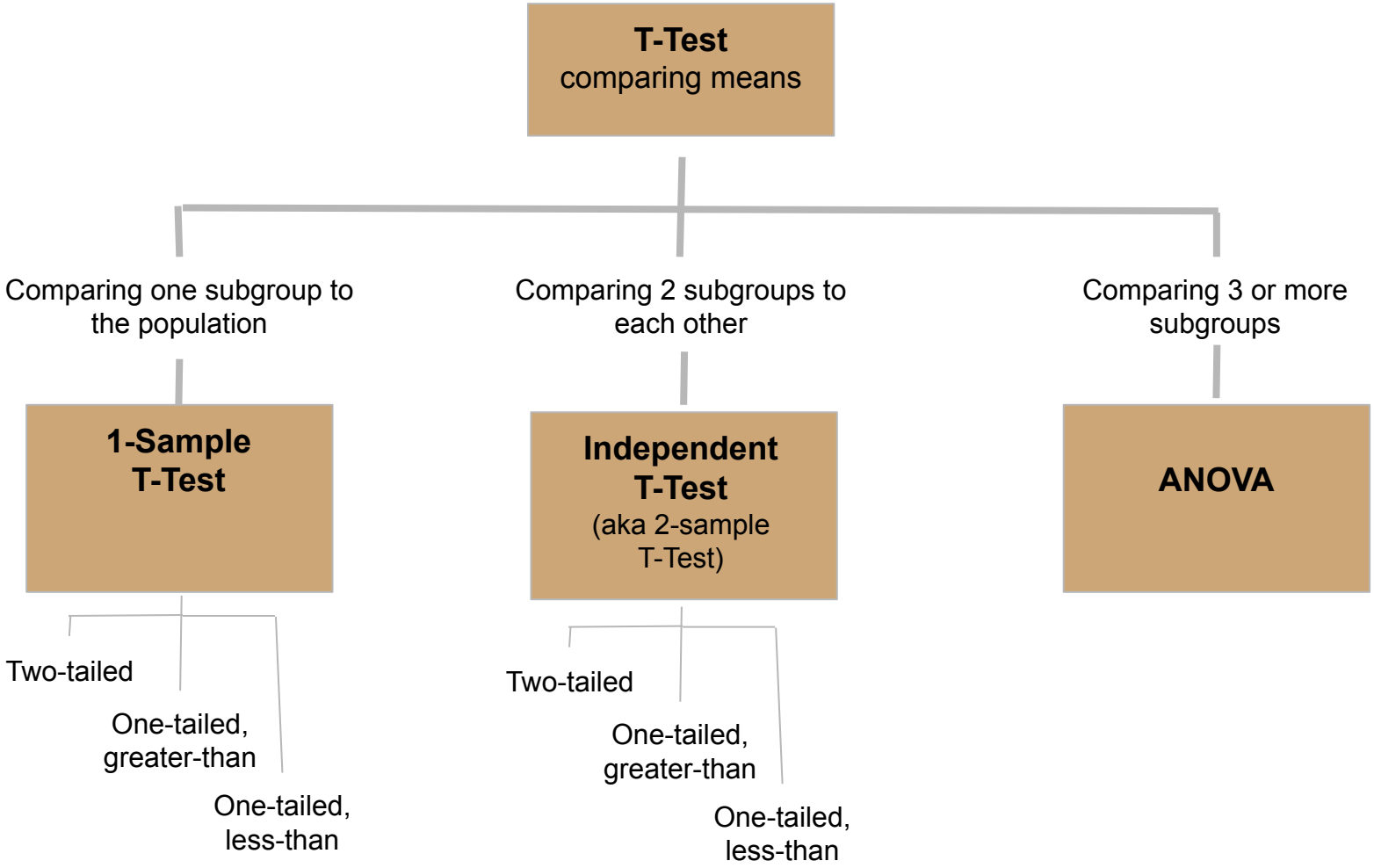


Correlation Test



guess the correlation coefficients

Chi-squared

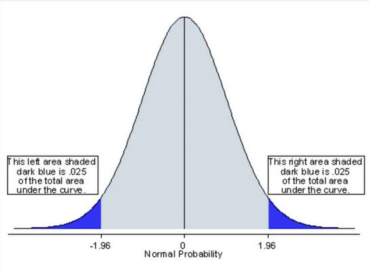

T-Test
comparing means
Comparing one subgroup to the population
Comparing 2 subgroups to each other
Comparing 3 or more subgroups
1-Sample T-Test
Independent T-Test
(aka 2-sample T-Test)
ANOVA
Two-tailed
One-tailed, greater-than
One-tailed, less-than
Two-tailed
One-tailed, greater-than
One-tailed, less-than

1-Sample T-Test
comparing 1 subgroup to the population

Independent T-Test
comparing 2 subgroups

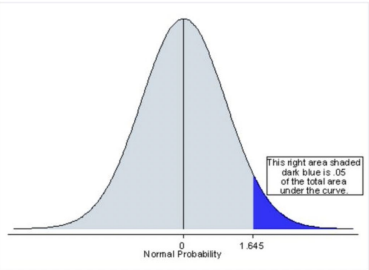
Is there ANY difference?

Two-Tailed



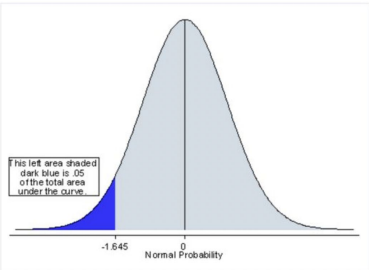
Is Group A
GREATER THAN
Group B (or the population)

One-tailed, greater than



Is Group A
LESS THAN
Group B (or the population)

One-tailed, less than



How to Perform Hypothesis Tests

How to conduct a hypothesis test

General Steps

1. Get the data
2. Establish Hypotheses
3. Visualize the data
4. Set alpha level
5. Verify necessary assumptions (for T-Tests)
6. Run the test (compute test statistic (*r* or t) and p-value)
7. Evaluate and conclude

Comparing Tests that Compare Means

Goal	H_0	Data Needed	Parametric Test	Assumptions*	Non-parametric Test
Compare observed mean to theoretical one	$\mu_{obs} = \mu_{th}$	array-like of observed values & float of theoretical	One sample t-test: scipy.stats.ttest_1samp	Normally Distributed**	One sample Wilcoxon signed rank test
Compare two observed means (independent samples)	$\mu_a = \mu_b$	2 array-like samples	Independent t-test (or 2-sample): scipy.stats.ttest_ind	Independent, Normally Distributed**, Equal Variances***	Mann-Whitney's test
Compare several observed means (independent samples)	$\mu_a = \mu_b = \mu_n$	n array-like samples	ANOVA: scipy.stats.f_oneway	Independent, Normally Distributed**, Equal Variances	Kruskal-Wallis test

Appendix - Useful Code or Resources

More Binomial Examples

Probability of Success & Number of Trials

- 5% chance student shows up late, with class of 20, what is likelihood everyone is on time? Stats.binom
  - **stats.binom(20 .05).pmf(0)**
- Multiple choice test with 30 questions, each question has 4 possible answers, choosing one at random, what is probability you get 11 or more correct?
  - **stats.binom(30, .25).sf(10)**
- Probability visitor will make a purchase when browsing website is 1.5% you expect 350 visitors.
  - **stats.binom(350, .015).isf(x)**
- A marketing website has an average click-through rate of 2%. One day they observe 4326 visitors and 97 click-throughs. How likely is it that this many people or more click through?
  - **stats.binom(4326, .02).sf(96)**