Module 2:

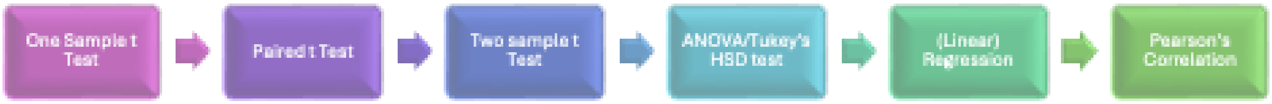
Inference for a Normal Population

Different flavours of t tests

Hypothesis testing for means using t tests Agenda

1. **Why do we use Student t-tests instead of Z scores?**
2. **What are the three types of t-tests**

- **One sample t tests**
  - ❑ Assumptions
  - ❑ When assumptions not met, use median and rank → **Signed test**
- **Paired t test**
  - ❑ Assumptions
- **Two sample t test**
  - ❑ Assumptions
  - ❑ When variances aren't equal → **Welch's approximate t test**
  - ❑ Other assumptions not met: median and rank → **Mann Whitney U test**



Independent Random Samples

Normally Distributed

How to test:

- Visual Assessment (Boxplot, histograms, qqplot, qqnorm)
- Shapiro Wilk test

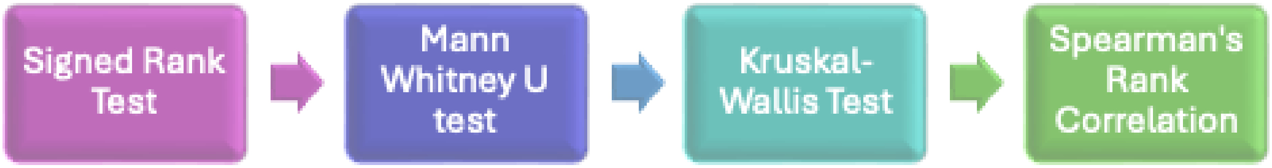
Homoscedasticity (variances are equal)

How to test

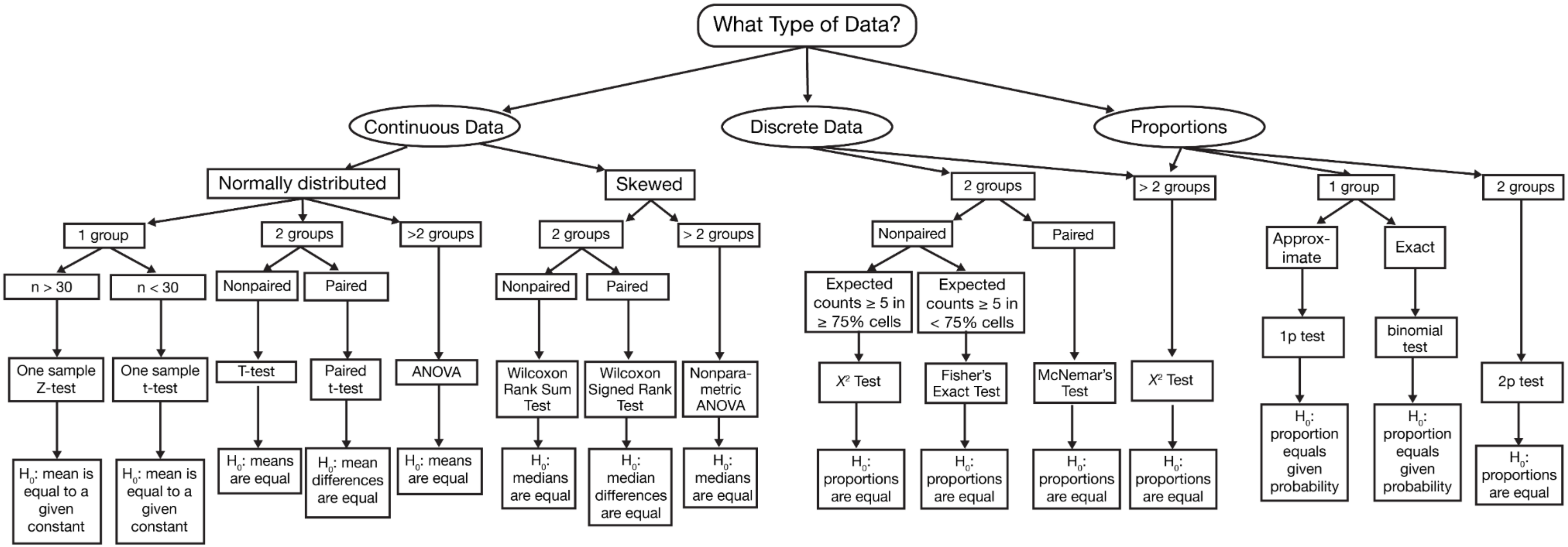
- Visual Assessment
- If normal; n=2 → var.test
- If normal; n>2 → Bartlett test
- If not norm → Levene's test
- If n=2 and variances are significantly different:

Welch's Approximate t test

Non-parametric analogs of the parametric tests above. Instead of testing population **means**, they test population **medians,** and they use **ranks.**

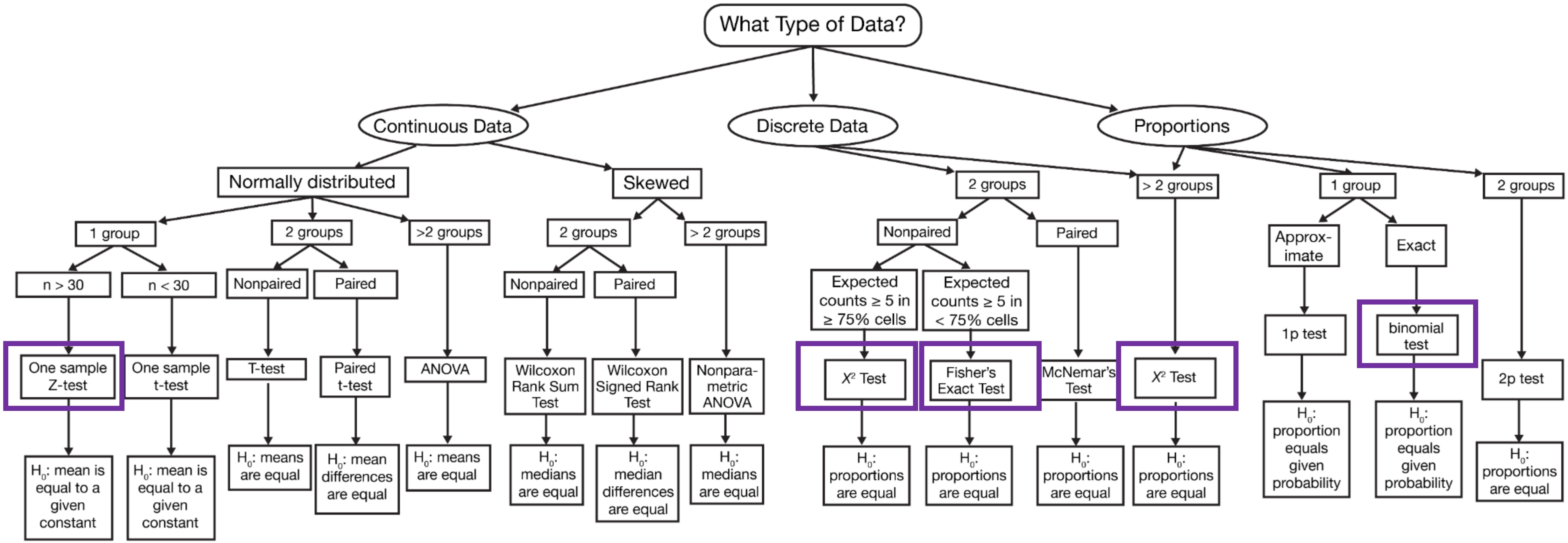


Flow chart: which test statistic should you use?

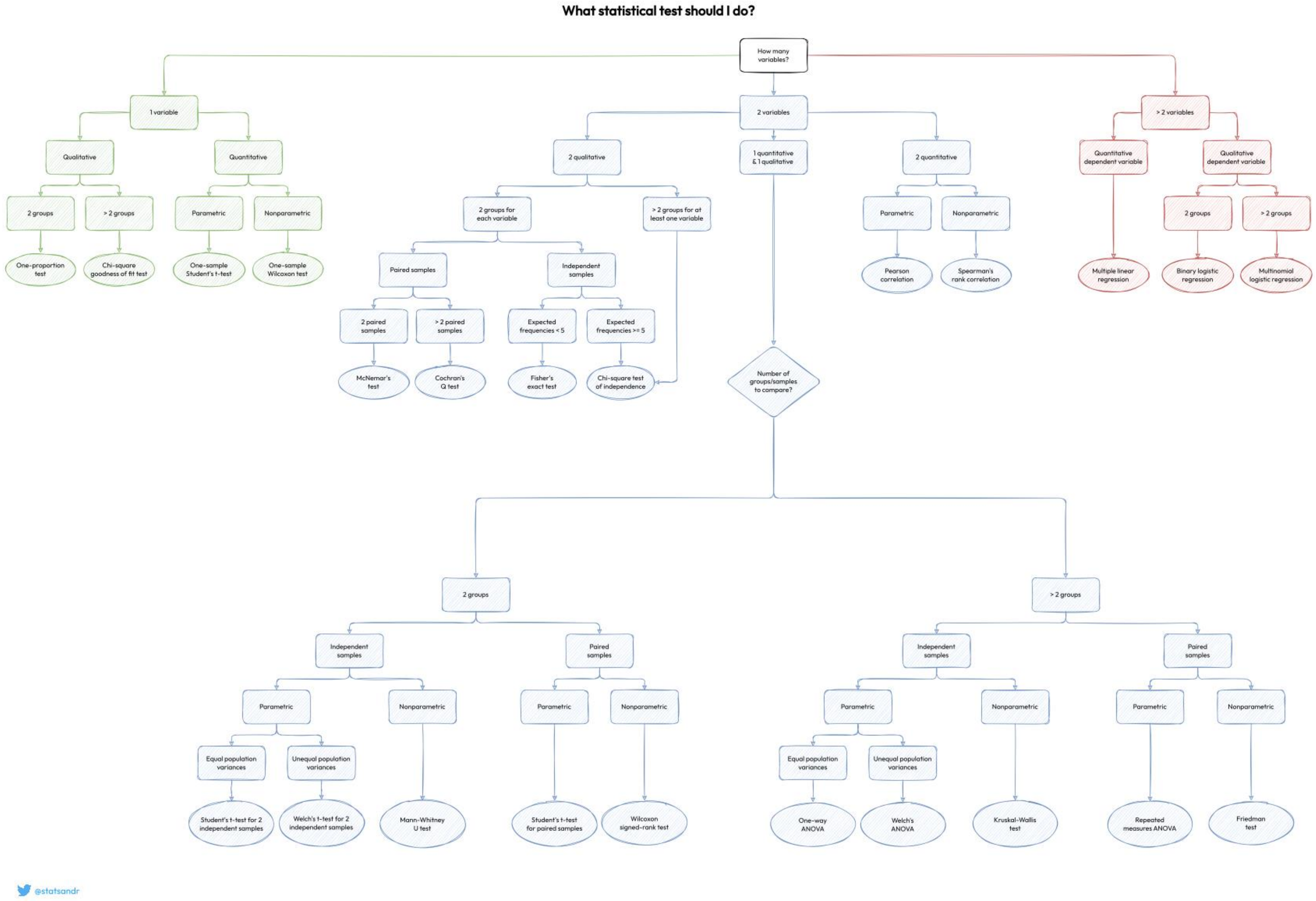


https://onishlab.colostate.edu/wp-content/uploads/2019/07/which_test_flowchart.png

Flow chart: which test statistic should you use?



https://onishlab.colostate.edu/wp-content/uploads/2019/07/which_test_flowchart.png



https://statsandr.com/blog/files/overview-statistical-tests-statsandr.pdf

Website with dataset, different types of data that lead to various tests, visualizations, and assumptions tested to run on it!

https://statisty.app/

Clear Table Export / Import Settings

nominal	metrisch	metrisch	nominal	metrisch	nominal	ordinal
Gender	**Salary**	**Age**	**Place**	**Weight**	**Company**	**Academic degree**
Female	1500	33	Chicago	80	BMW	Bachelor
Female	1200	33	Chicago	82.5	Ford	No
Male	2200	34	New York	100.8	BMW	Bachelor
Male	2100	42	New York	90	BMW	Master
Female	1500	29	Chicago	67	Ford	Master
Female	1700	19	Washington	60	Ford	Master
Male	3000	50	Washington	77	Ford	No
Male	3000	55	Washington	77	Ford	Bachelor
Female	2800	31	New York	87	Ford	Bachelor
Male	2900	46	New York	70	GM	Master
Female	2780	36	Washington	57	BMW	No
Male	2550	48	New York	64	GM	Master

Metric: ☐ Salary ☐ Age ☐ Weight

Ordinal: ☐ Academic degree

Nominal: ☐ Gender ☐ Place ☐ Company

Descriptive Statistics

Hypothesis test

Regression

Charts

Part 1: **Why** t tests?

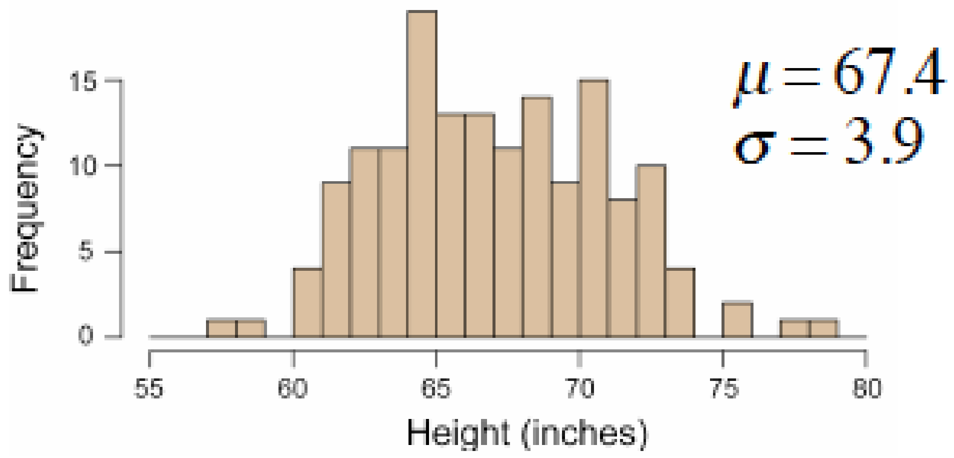
Inference about means:

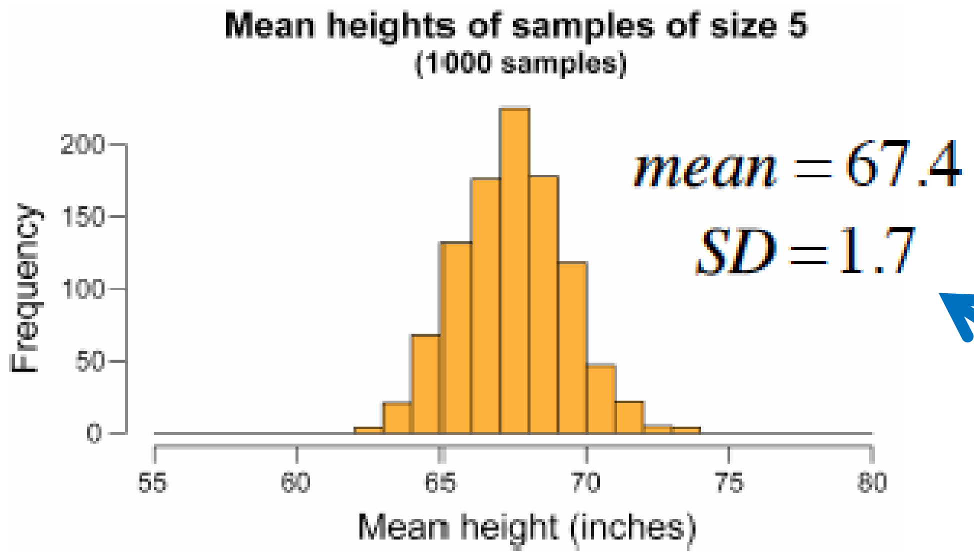
As a reminder from Learning Path I….

- To make statistical statements, we need to describe the **sampling distribution** of an estimator (for the null hypothesis probabilities).
  - The sampling distribution is the probability distribution of all values of an estimate that we might obtain when sampling a population
  - When the variable Y is normally distributed or n is large (if Y is not normally distributed) the sampling distribution for E(Y) is normal*

* thank-you Central Limit Theorem

Inference for a Normal Population





The central limit theorem has <u>two</u> constraints:

1. It depends on a large sample size (n > 30-ish)
2. To use it, we need to know σ^2 (i.i.d.), but we seldom do.

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{3.9}{\sqrt{5}} = 1.7$$

Inference about Means

- Because $\overline{Y}$ is normally distributed, we can convert the distribution to the **standard normal distribution:**

$$Z = \frac{\overline{Y} - \mu}{\sigma_{\overline{Y}}} = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}}$$

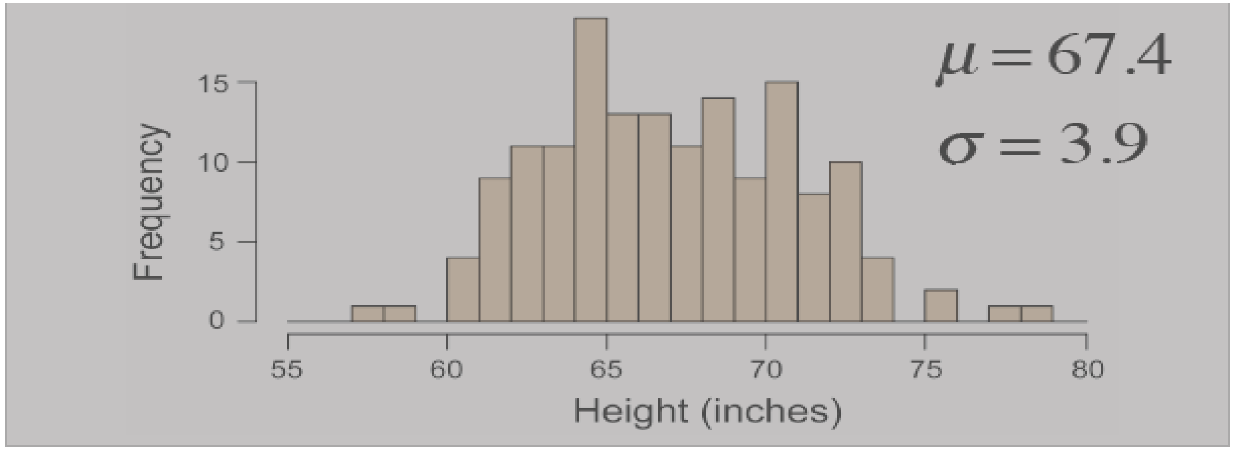
- This gives a probability distribution of <u>the difference</u> between **a sample** mean and **the population** mean

But we don't know σ!

Now what?

- we *do* know ***s***, the standard deviation of our sample, which estimates **σ**.
  - We can use ***s*** to get: $SE_{\overline{Y}} = \frac{s}{\sqrt{n}}$
  - This is used as an estimate of $\sigma_{\overline{Y}}$

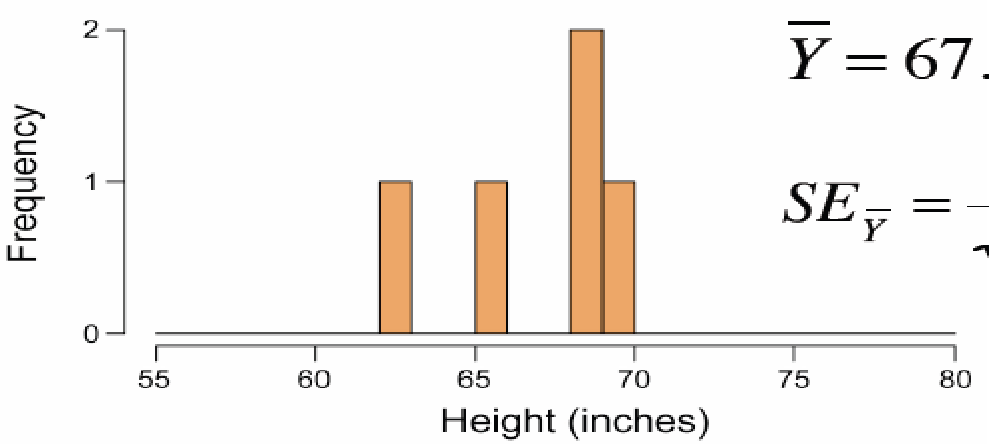
Inference for a Normal Population



In most cases, we don't know the real population distribution.

We only have a sample.

Heights of a sample of students (*n* = 5)



$\overline{Y} = 67.1 \qquad s = 3.1$

$$SE_{\overline{Y}} = \frac{s}{\sqrt{n}} = \frac{3.1}{\sqrt{5}} = 1.4$$

We use this as an estimate of $\sigma_{\overline{Y}}$

The Z score:

SIGNAL

$$Z = \frac{\boxed{Y - \mu}}{\boxed{\sigma_{\bar{Y}}}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

NOISE

The Student's t Distribution:

SIGNAL

$$t = \frac{\boxed{\bar{Y} - \mu}}{\boxed{SE_{\bar{Y}}}} = \frac{\bar{Y} - \mu}{s / \sqrt{n}}$$

NOISE

* Alias of William Gosset of the Guinness Brewing Company

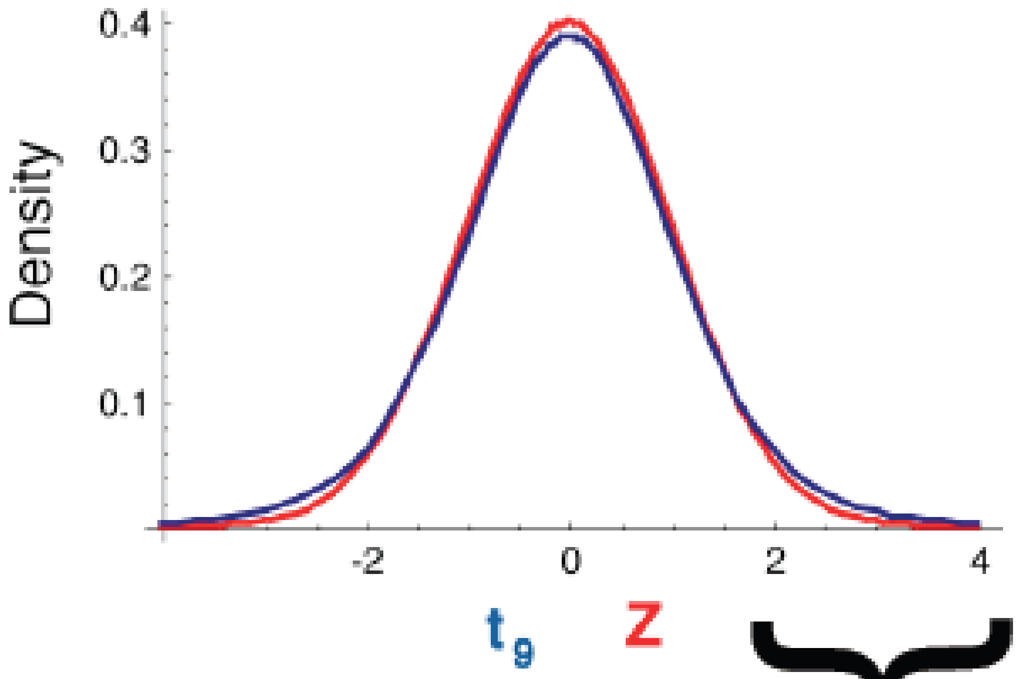
	t	**Z**
Stand. Error	$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$	$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$
dof	n - 1	n
Sampling Distribution	t-distribution	Normal Distribution

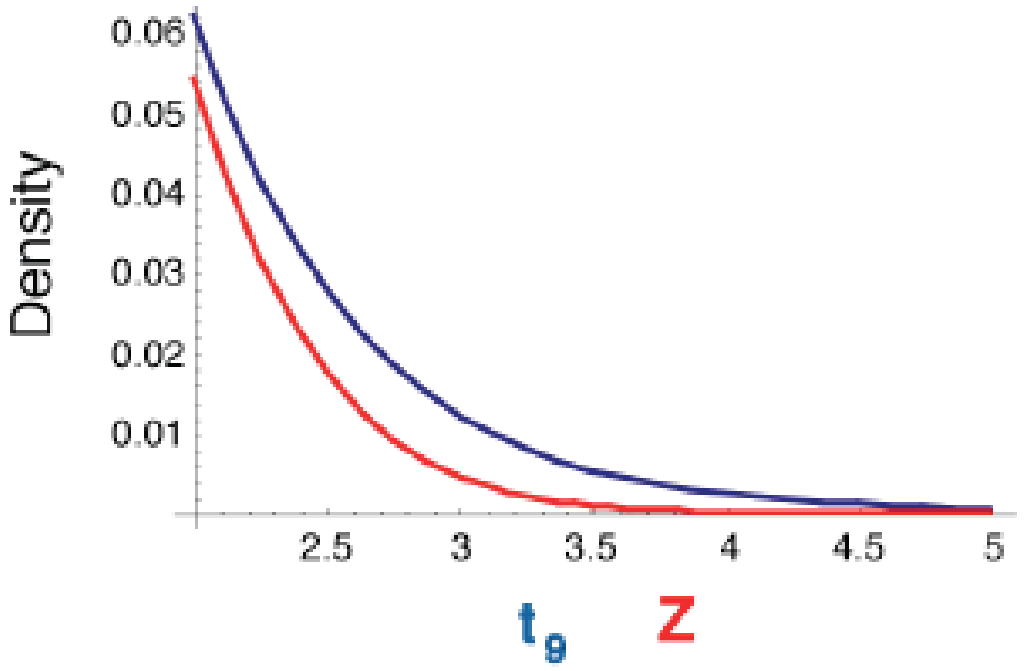
Why is degrees of freedom n-1 instead of n?

The consequences of using $SE_{\bar{Y}}$ instead of $\sigma_{\bar{Y}}$

- **The value of $SE_{\bar{Y}}$ is different for each sample; it doesn't have a constant value like $\sigma_{\bar{Y}}$**
  - Introduces some error
    - t-distribution is wider than the equivalent Normal distribution
      - therefore, it is not as precise
    - As sample size, *n*, increases the t-distribution narrows and approaches the Normal Distribution

- *dof = n - 1 because we have 'used up' one piece of information when we estimate $\sigma_{\bar{Y}}$ by using $SE_{\bar{Y}}$*





$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

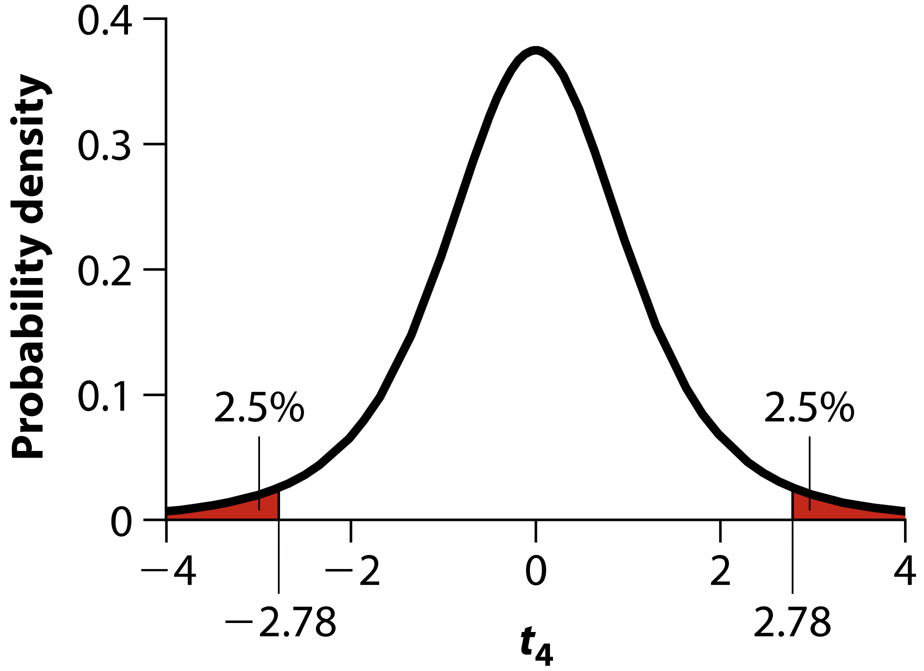
$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}}$$

Critical values for the student's t distribution:

https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

- $t_{0.05(2),df}$
  - *the 0.05 is the fraction of the area under the curve shared between the two tails of the distr 'n:*

2.5% > $t_{0.05(2),df}$ and **- 2.5% < $-t_{0.05(2),df}$**



Use the t-distribution to calculate the **confidence interval** for the mean of a normal distribution

$$\boxed{-t_{\alpha(2),df} < \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} < t_{\alpha(2),df}}$$

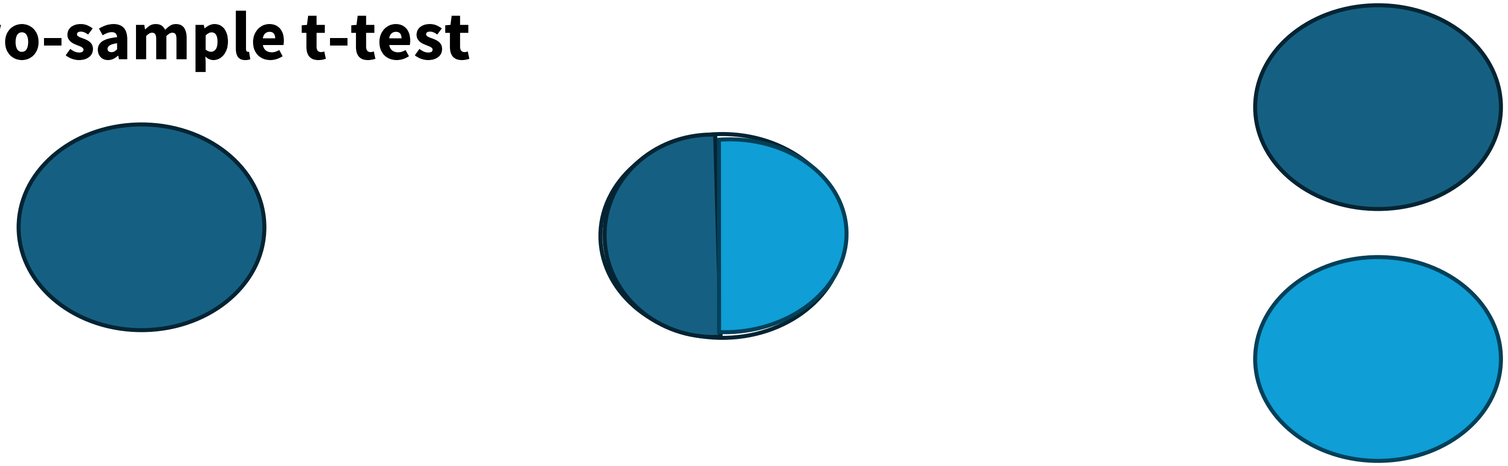
- This can be re-arranged:

$$\boxed{\bar{Y} - t_{\alpha(2),df} SE_{\bar{Y}} < \mu < \bar{Y} + t_{\alpha(2),df} SE_{\bar{Y}}}$$

Never this (why not?): $\cancel{\bar{Y} - t_{\alpha(2),df} SE_{\bar{Y}} < \bar{Y} < \bar{Y} - t_{\alpha(2),df} SE_{\bar{Y}}}$

Part 2: **What** t tests? We will look at the following t-tests:

1. Comparing one mean:
   a. **One-sample t-test**
2. Comparing two means:
   a. Paired t-test
   b. Two-sample t-test



one sample   paired   two sample

*Each of the above tests have **slightly different assumptions** which allow our conclusions to be supported. We will investigate what happens when these assumptions are violated and how robust our various t-tests are to violations.*

Researchers are studying the body weight of mice to understand the impact of a high-fat diet on genetically modified (GM) mice. They can collect the following data: Body Weight.

One-Sample t-test:

Two-Sample t-test:

Paired t-test:

Researchers are studying the body weight of mice to understand the impact of a high-fat diet on genetically modified (GM) mice. They can collect the following data: body weight.

One-Sample t-test: Does the **body weight of the GM mice** differ significantly from a **known population mean weight** of non-GM mice?

Two-Sample t-test: Does the body weight of GM mice on a **high-fat diet** differ from the body weight of a GM mice on a **standard diet**?

Paired t-test: They measure the body weights of a group of GM mice **before** and **after** they are switched from a normal diet to a high fat diet to see if there's a significant change in weight within the same group.