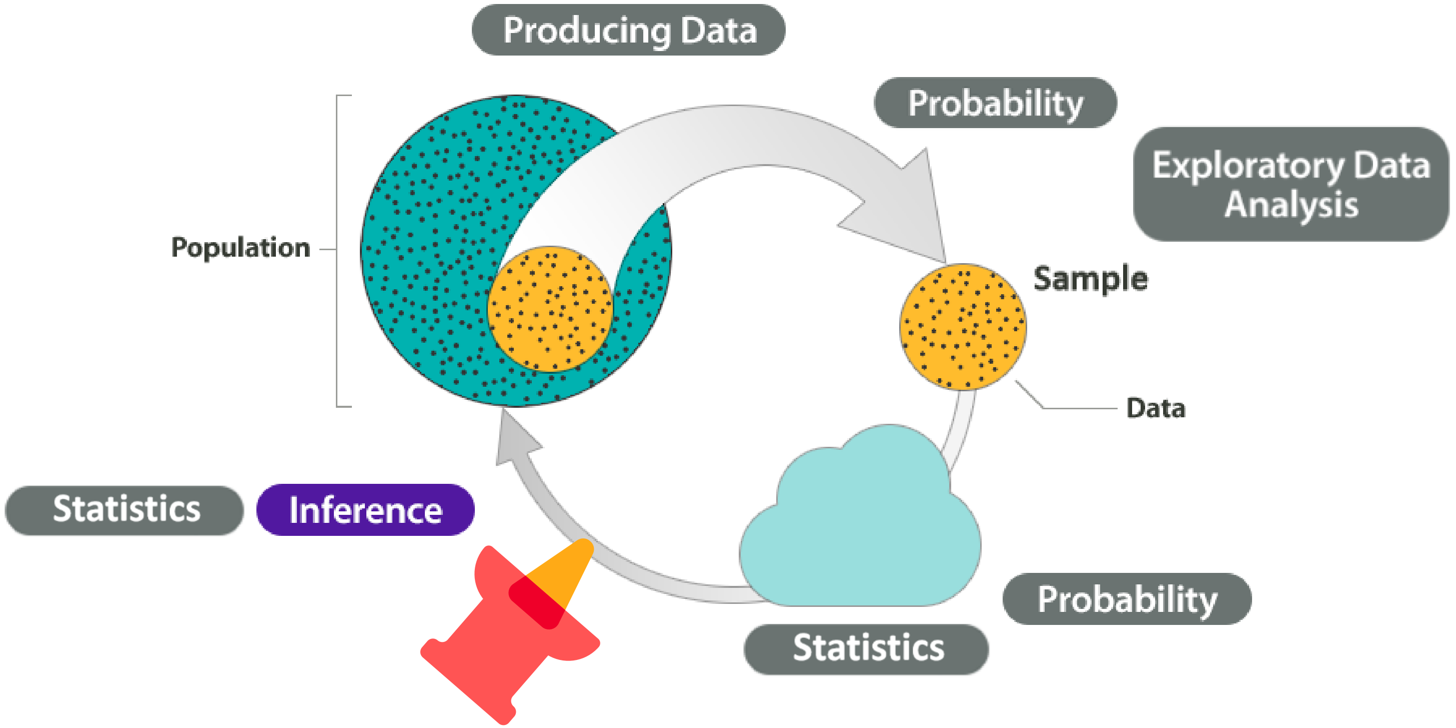
CDS 533
Statistics for Data Science

Instructor: Lisha Yu
Division of Artificial Intelligence
School of Data Science
Lingnan University

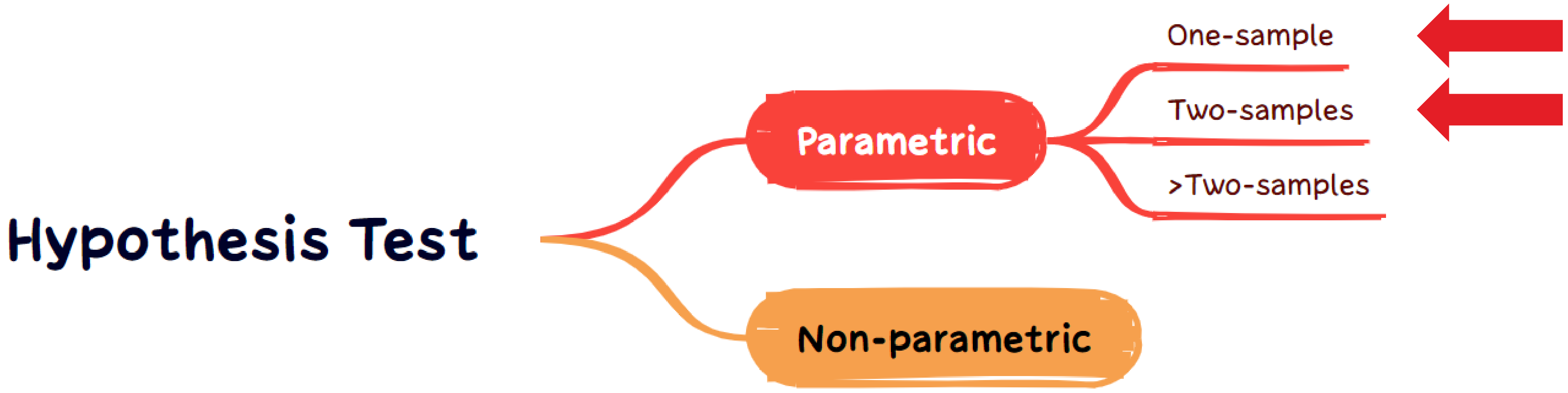
Fall 2024

Big Picture of Statistics



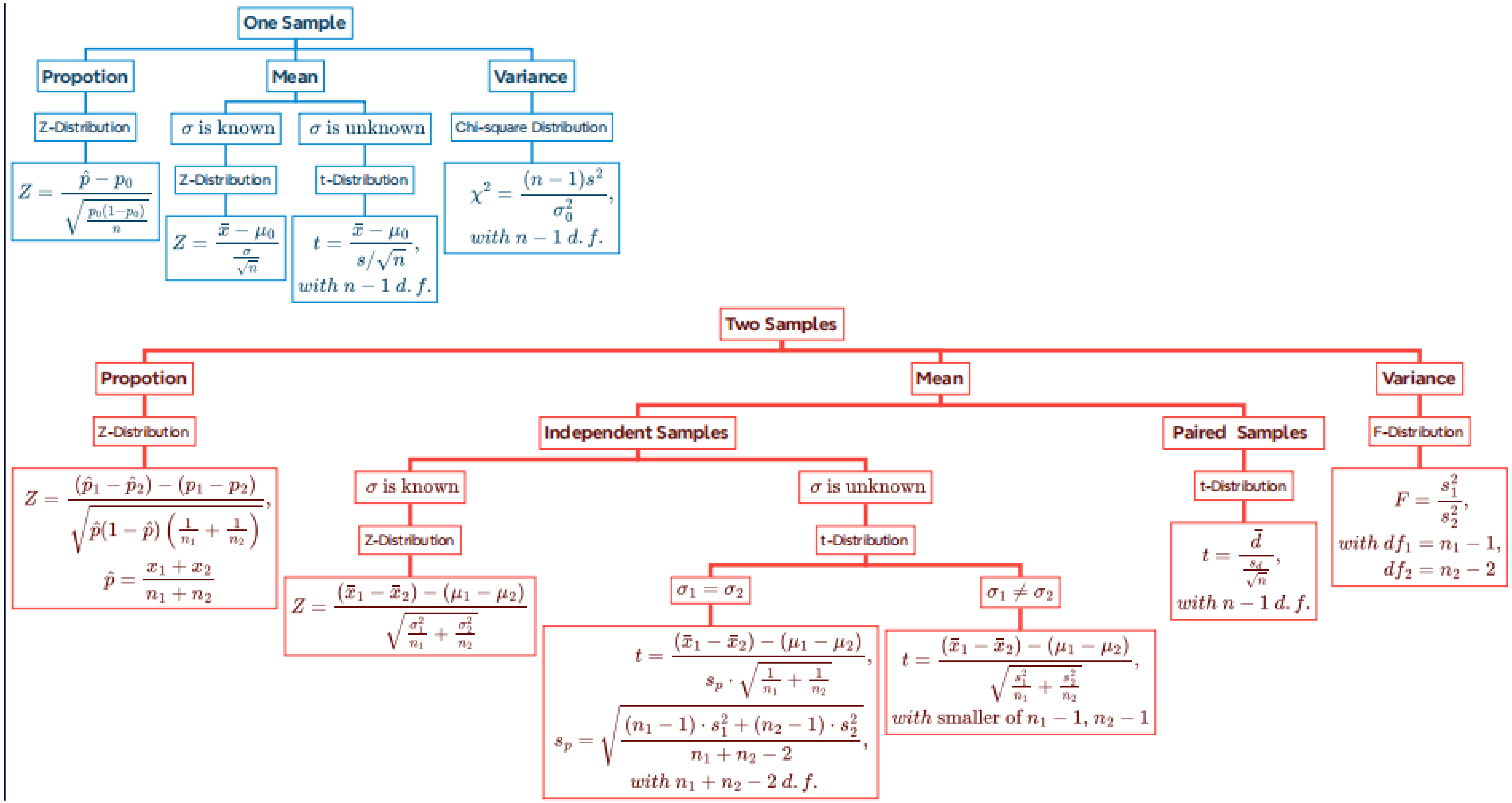
Statistical Inference (Hypothesis Testing)

Hypothesis Testing



Hypothesis Testing

Parametric



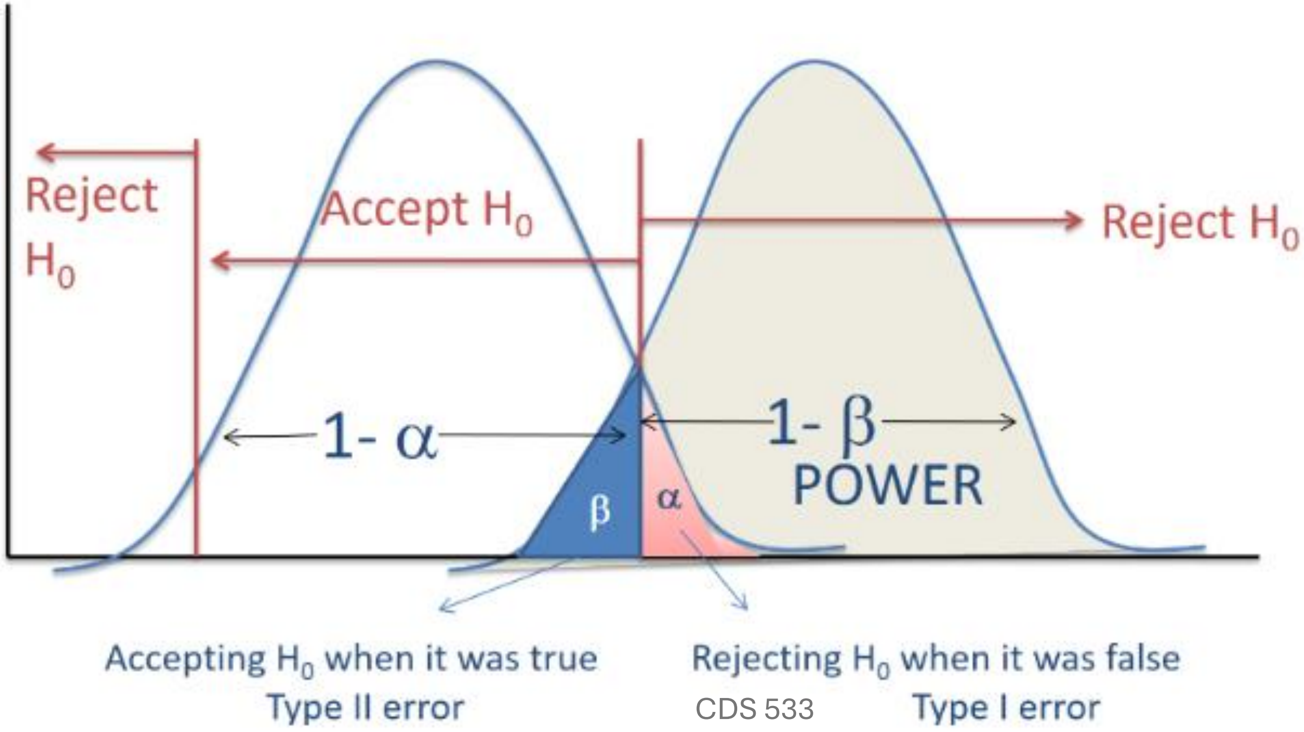
Power and Effect Size

Statistical Power

- The probability that if the research hypothesis is true, the results of your study will support the research hypothesis.
- The probability that you will not make a Type II error.
- The higher the power, the more sensitive the test is.

Types of Errors

Hypothesis Testing Outcomes		Reality: H_0 is True	Reality: H_0 is False
Research	Retain the H_0	Accurate $p = 1-\alpha$ ☺	Type II error $p = \beta$ ☹
Research	Reject the H_0	Type I error $p = \alpha$ ☹	Accurate $p = 1-\beta$ ☺ Power

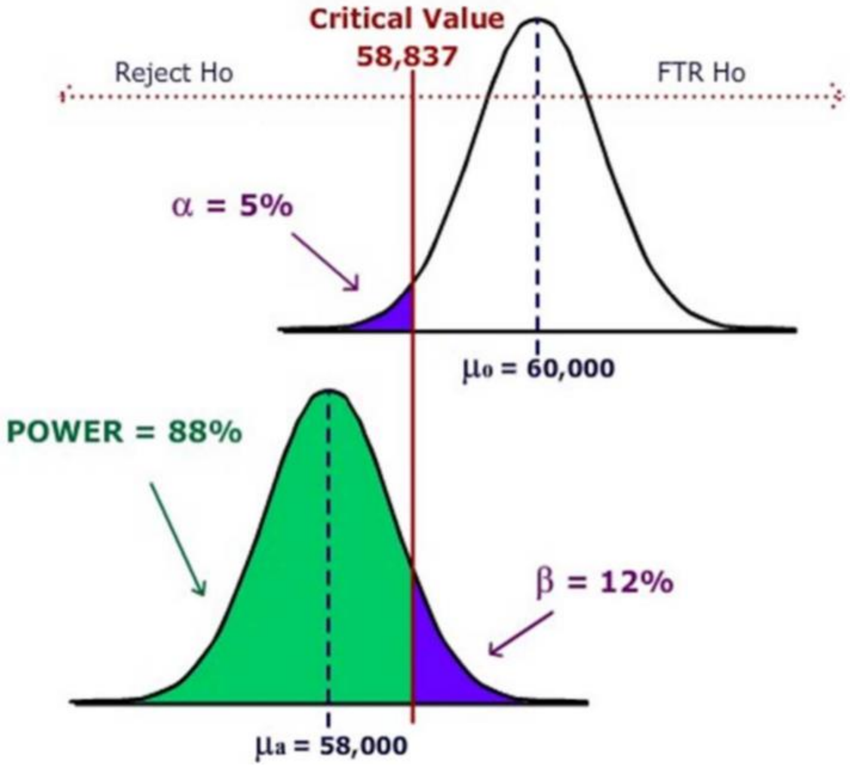


Power and Effect Size

Example

Bus brake pads are claimed to last on average at least 60,000 miles and the company wants to test this claim. The bus company considers a "practical" value for purposes of bus safety to be that the pads last at least 58,000 miles. If the standard deviation is 5,000 and the sample size is 50, find the probability of type II error and power of the test when the mean is really 58,000 miles. (Assume $\alpha = 0.05$)

$H_0: \mu \geq 60000$
$Ha: \mu < 60000$
$a = 0.05$



Power and Effect Size

What Affects Power

1. Sample size
2. Criterion for rejecting null (alpha, α)
3. Effect size

Feature of the Study	Increases Power	Decreases Power
Hypothesized difference between Population Means Pop 1 M – Pop 2 M	Large Differences	Small Differences
Population Standard Deviation	Small Pop SD	Large Pop SD
Sample Size (N)	Large N	Small N
Significance Level	Lenient (5% or 10%)	Stringent (1% or .1%)
One vs Two-tailed test	One-Tailed	Two-Tailed

https://melbapplets.ms.unimelb.edu.au/2021/07/12/power-of-a-hypothesis-test/

Power and Effect Size

Notes on Effect Size

- A hypothesis test does not evaluate the size of the effect of the treatment.
- An effect size provides a measurement of the magnitude of the effect.
- One common effect size you will see [**Group Difference Indices**] when means are compared is Cohen's d. Cohen's d is the difference between means in standard deviation units: [Independent Sample t-Test]

$$d = \frac{M_1 - M_2}{SD_{pooled}}$$

d= Cohen's d, meaning the difference between two means measured in units of the pooled standard deviation.

$M1$= mean of group 1

$M2$= mean of group 2

SD = Standard deviation depending on the test

Value of Cohen's d	Evaluation of Effect Size
$0 < d < 0.2$	Small effect
$0.2 < d < 0.8$	Medium effect
$d > 0.8$	Large effect

Power and Effect Size

Effect Size Example

Question: Are the treatment different?

Treatment	Size (n)	Mean (M)	Std Dev (SD)	Variance
A	29	85.7	8.35	69.7225
B	26	81.1	4.74	22.4676

Observed Difference in Means

$$\bar{D} = \bar{x_1} - \bar{x_2} = 85.7 - 81.1 = \mathbf{4.6\ pounds}$$

Pooled Standard Deviation

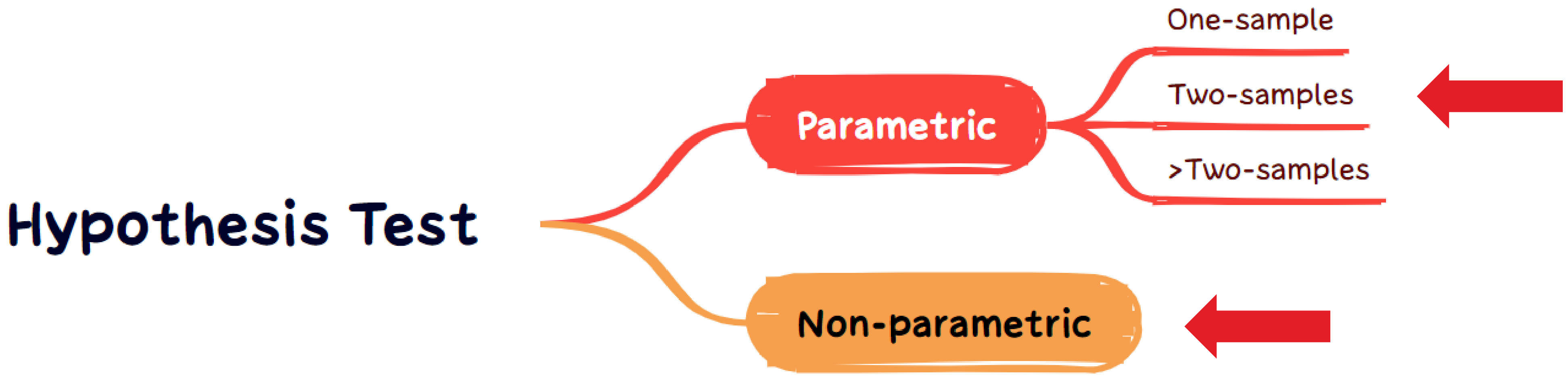
$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1-n_2-2}} = \sqrt{\frac{(28)69.7225+(25)22.4676}{53}} = \mathbf{6.8871}$$

Effect Size

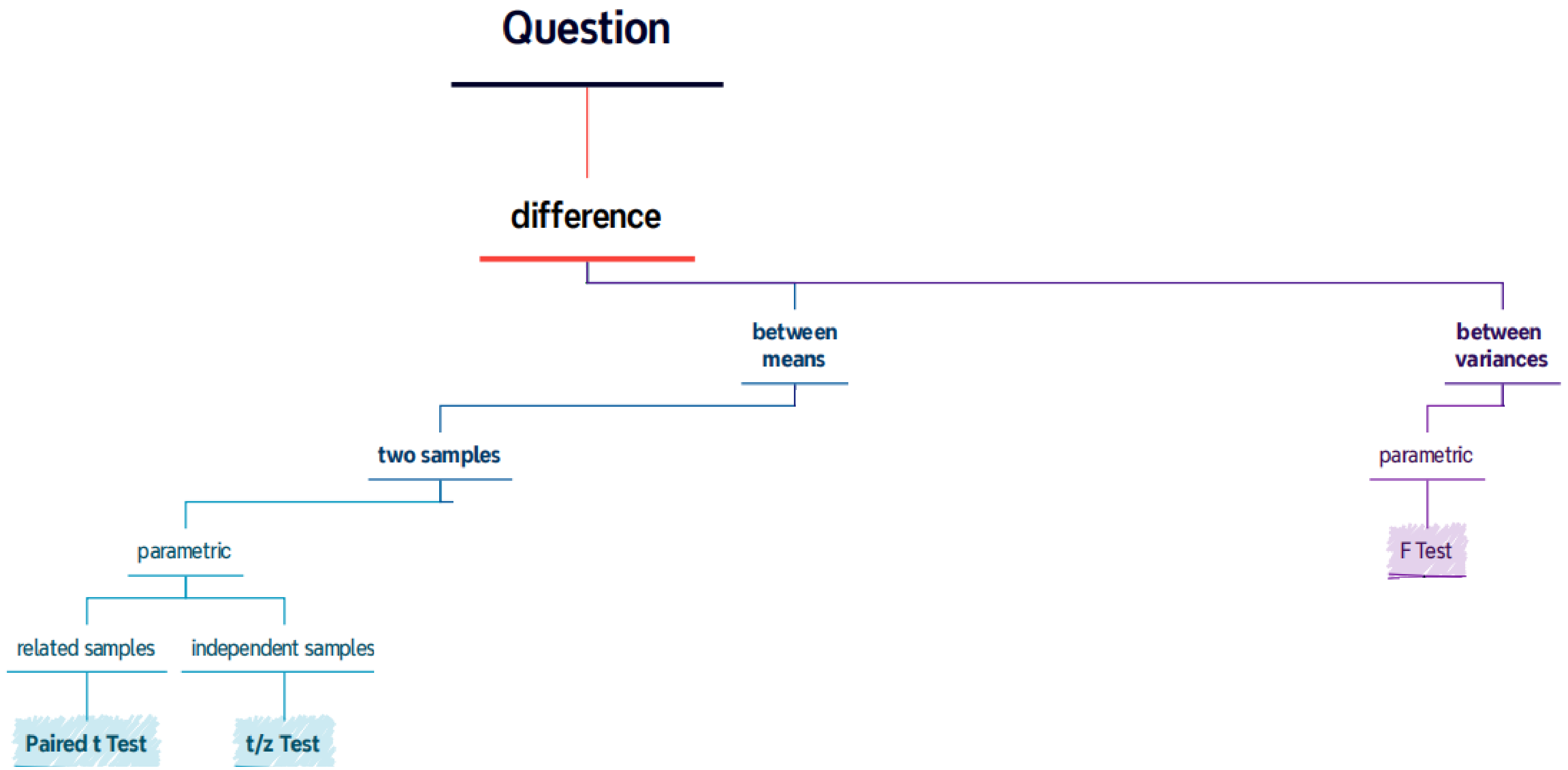
$$d_s = \frac{\mu_1-\mu_2}{\sigma} = 4.6/6.8871 = \mathbf{0.67}$$

Value of Cohen's d	Evaluation of Effect Size
$0 < d < 0.2$	Small effect
$0.2 < d < 0.8$	Medium effect
$d > 0.8$	Large effect

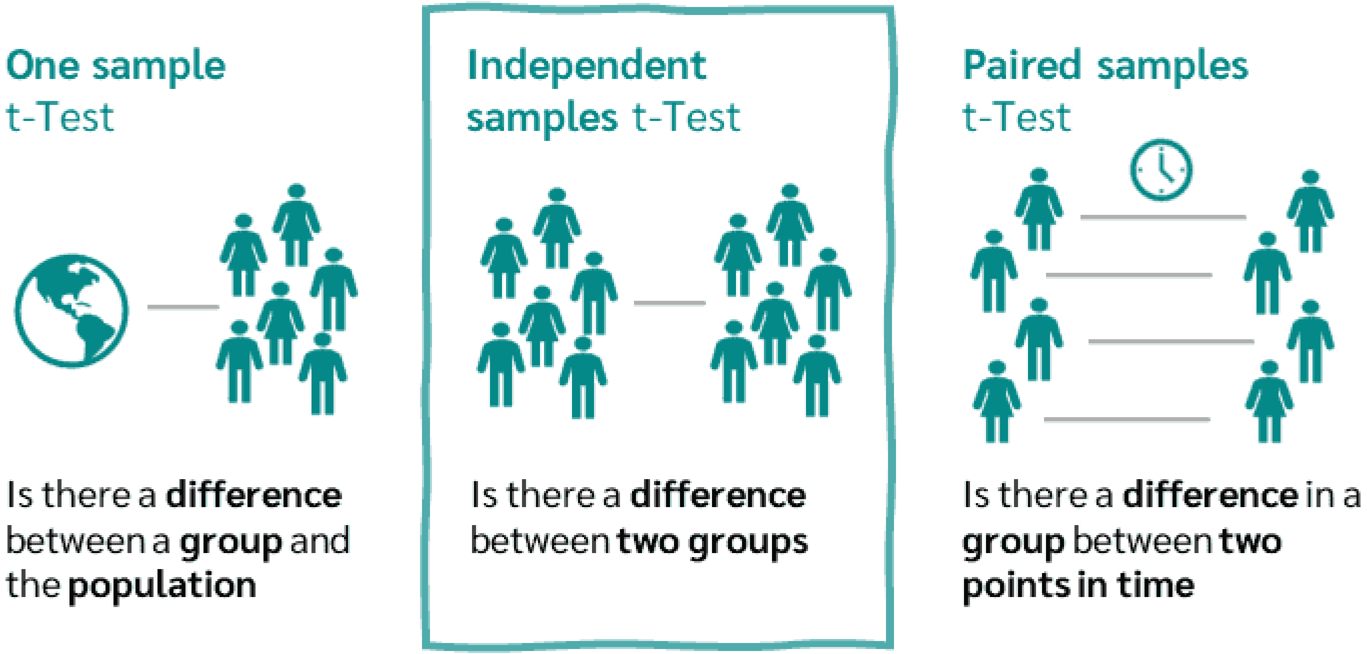
Hypothesis Testing



Review


Question
difference
between means
between variances
two samples
parametric
parametric
related samples
independent samples
Paired t Test
t/z Test
F Test

Sample t-test



Example

In the independent two-sample t-test, what are the **3 assumptions**?

Two Sample t-test

Equal Variance [Homogeneous]

- If n_1 is close to n_2, the test is moderately robust against unequal variance ($\sigma_1^2 \neq \sigma_2^2$)
- If n_1 and n_2 are quite different (e.g., differ by a ratio of ≥ 3), then the test is much less robust.

Motivation

When is testing the equal variance most commonly used?

It is most often used to **test assumptions for another hypothesis test.**

There is a difference between two medications in terms of perceived pain relief.

To test the **hypothesis** based on your **data**, you use a **hypothesis test**, such as a **t-test**.

Many **hypothesis tests** have the assumption that the **variances** in each group **are equal**.

Variance: P: F-test

Testing $H_0: \sigma^2_1 = \sigma^2_2$		
Alternative hypothesis	Decision rule: Reject H_0 if	Test statistic
$H_1: \sigma^2_1 \neq \sigma^2_2$	$F \geq F_{n1-1\,;\,n2-1\,;\,\alpha/2}$	$F = \frac{s_1^2}{s_2^2}$
$H_1: \sigma^2_1 > \sigma^2_2$	$F \geq F_{n1-1\,;\,n2-1\,;\,\alpha}$	
Assume population 1 has the larger variance. Thus always: $s^2_1 > s_2^2$		

The test is very sensitive to nonnormality.

Parametric or Non-parametric?

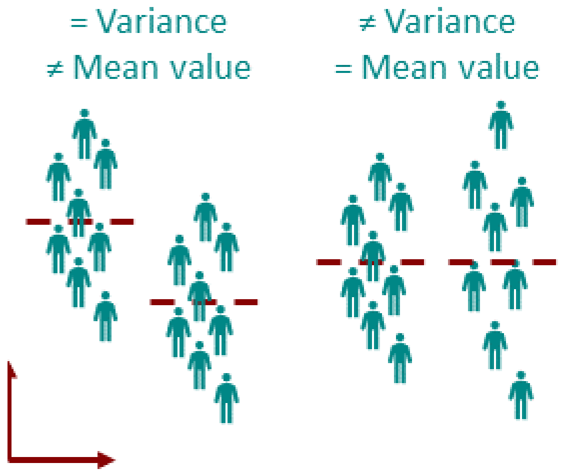
- Parametric tests are restricted to data that:
  1) show a normal distribution
  2) * are independent of one another
  3) * are on the same *continuous* scale of measurement

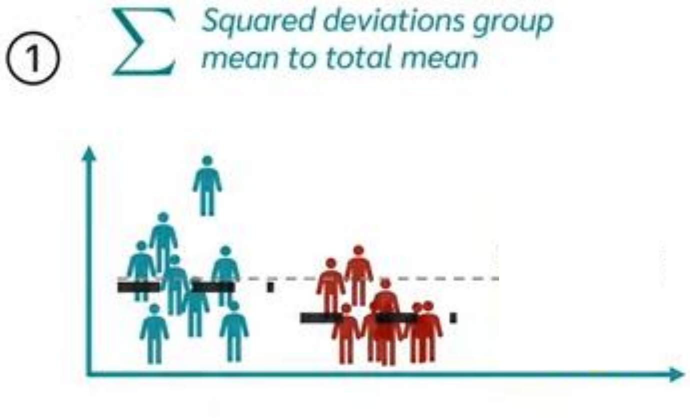
- Non-parametric tests are used on data that:
  1) show an other-than normal distribution
  2) are dependent or conditional on one another
  3) in general, do not have a continuous scale of measurement

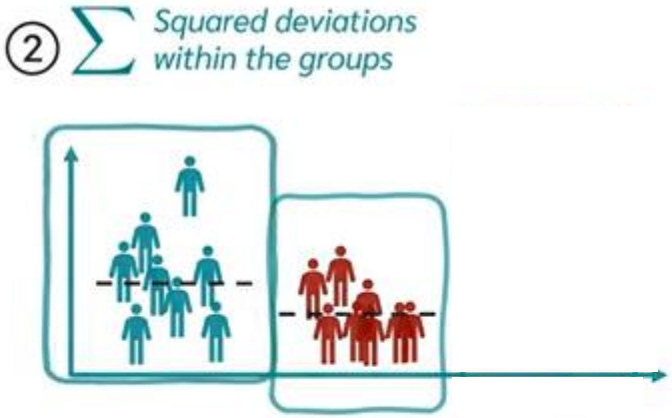
e.g., the length and weight of something vs. did the bacteria grow or not grow

Variance: NP: Levene's Test

Consider two groups Z_1 and Z_1:
Group 1: 4, 5, 6, 8, 10, 15, 17, 23
Group 2: 1, 2, 7, 8, 20, 22, 25, 35
Test $H_0: \sigma_1^2 = \sigma_2^2$ (Groups have equal variances)
$H_A: \sigma_1^2 \neq \sigma_2^2$ (Groups have different variances)







Variance: NP: Levene's Test

Group 1: 4, 5, 6, 8, 10, 15, 17, 23
Group 2: 1, 2, 7, 8, 20, 22, 25, 35
Test $H_0: \sigma_1^2 = \sigma_2^2$ (Groups have equal variances)
$H_A: \sigma_1^2 \neq \sigma_2^2$ (Groups have different variances)

L Statistic:

number of cases
number of cases in the i-th group
mean value of the i-th group
overall average

$$L = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2}$$

the respective value in the groups

F Distribution

The **degrees of freedom** result with:

Number of groups minus 1
$df_1 = k-1$

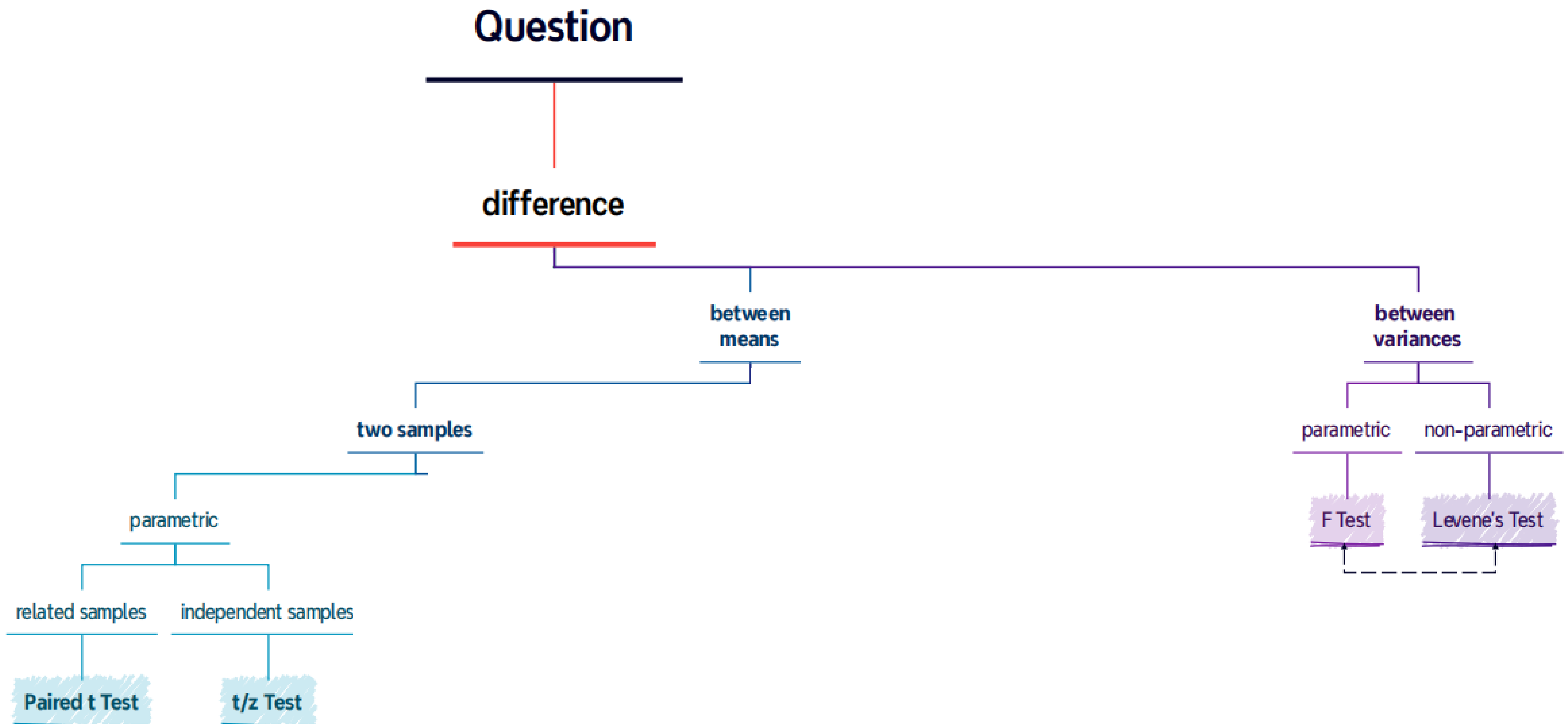
And number of cases minus number of groups.
$df_2 = N-k$

Use p-value to determine the test, i.e.,
p-value $> \alpha$ → NOT REJECT
p-value $< \alpha$ → REJECT

Level test of variance equality

F	df1	df2	Significance
2.016	2	27	0.153

Hypothesis Testing

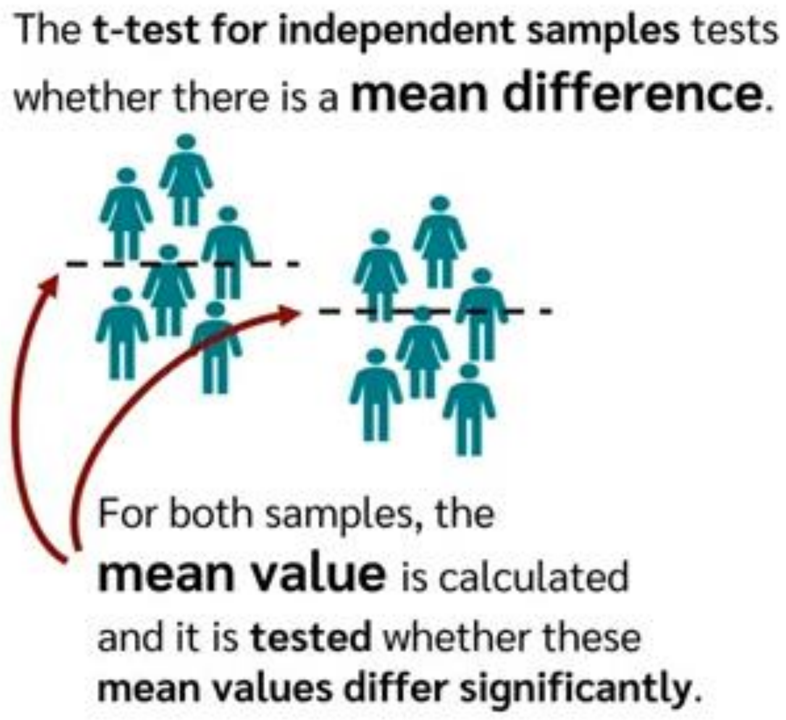


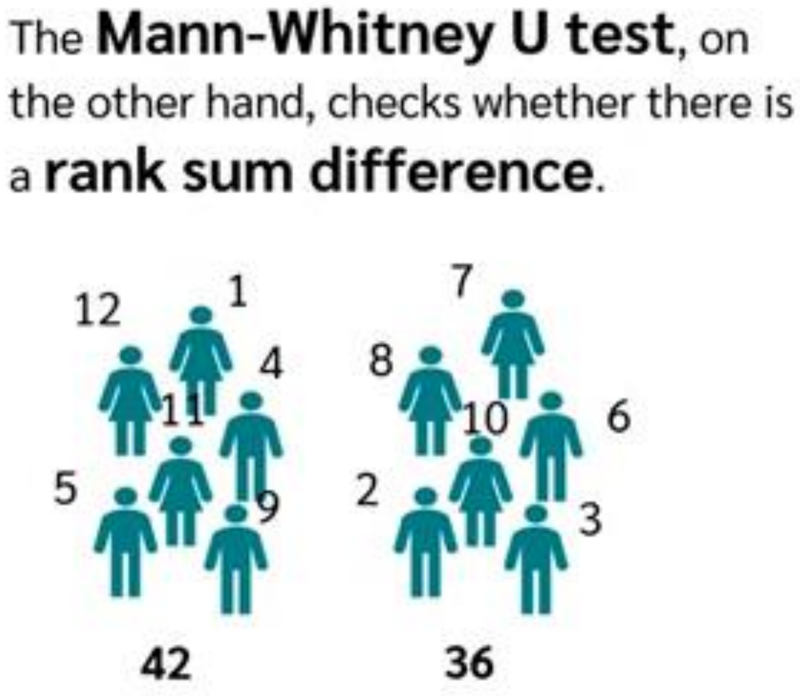
Independent Mean:NP: Mann-Whitney U-Test

The **Mann-Whitney U-Test** is **non-parametric counterpart** to the **t-test for independent samples**.

Independent Mean:NP: Mann-Whitney U-Test

But there is an **important difference** between the two tests.





Independent Mean:NP: Mann-Whitney U-Test

The **advantage** of taking the **rank sums** rather than the difference in means is that the data **need not be normally distributed**.

Assumptions of the Mann-Whitney U test:
Two **independent** samples with at least ordinal scaled characteristics need to be available

Independent Mean:NP: Mann-Whitney U-Test

H_0 : There is no difference (in terms of median) between the two groups in the population.
H_a : There is a difference (w.r.t. the median) between the two groups in the population.

U Statistic:

Gender	Reaction time	Rang
female	34	2
female	36	4
female	41	7
female	43	9
female	44	10
female	37	5
male	45	11
male	33	1
male	35	3
male	39	6
male	42	8

Female

Number of cases $n_1 = 6$ Rank sum $T_1 = 37$

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - T_1 = 6 \cdot 5 + \frac{6 \cdot (6+1)}{2} - 37 = 14$$

Male

Number of cases $n_2 = 5$ Rank sum $T_2 = 29$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - T_2 = 6 \cdot 5 + \frac{5 \cdot (5+1)}{2} - 29 = 16$$

U-Wert

$$U = min(U_1, U_2) = min(14, 16) = 14$$

Expected value of U

$$\mu_U = \frac{n_1 \cdot n_2}{2} = \frac{6 \cdot 5}{2} = 15$$

Standard error of U

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1+n_2+1)}{12}} = \sqrt{\frac{6 \cdot 5 \cdot (6+5+1)}{12}} = 5.4772$$

z-value

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{14-15}{5,4772} = -0.1825$$

p-value = 0.855

Depending on how large the sample is, the p-value is calculated in a different way.

- For up to 25 cases, the exact values are used, which can be read from a table.
- For larger samples, the normal distribution can be used as an approximation.

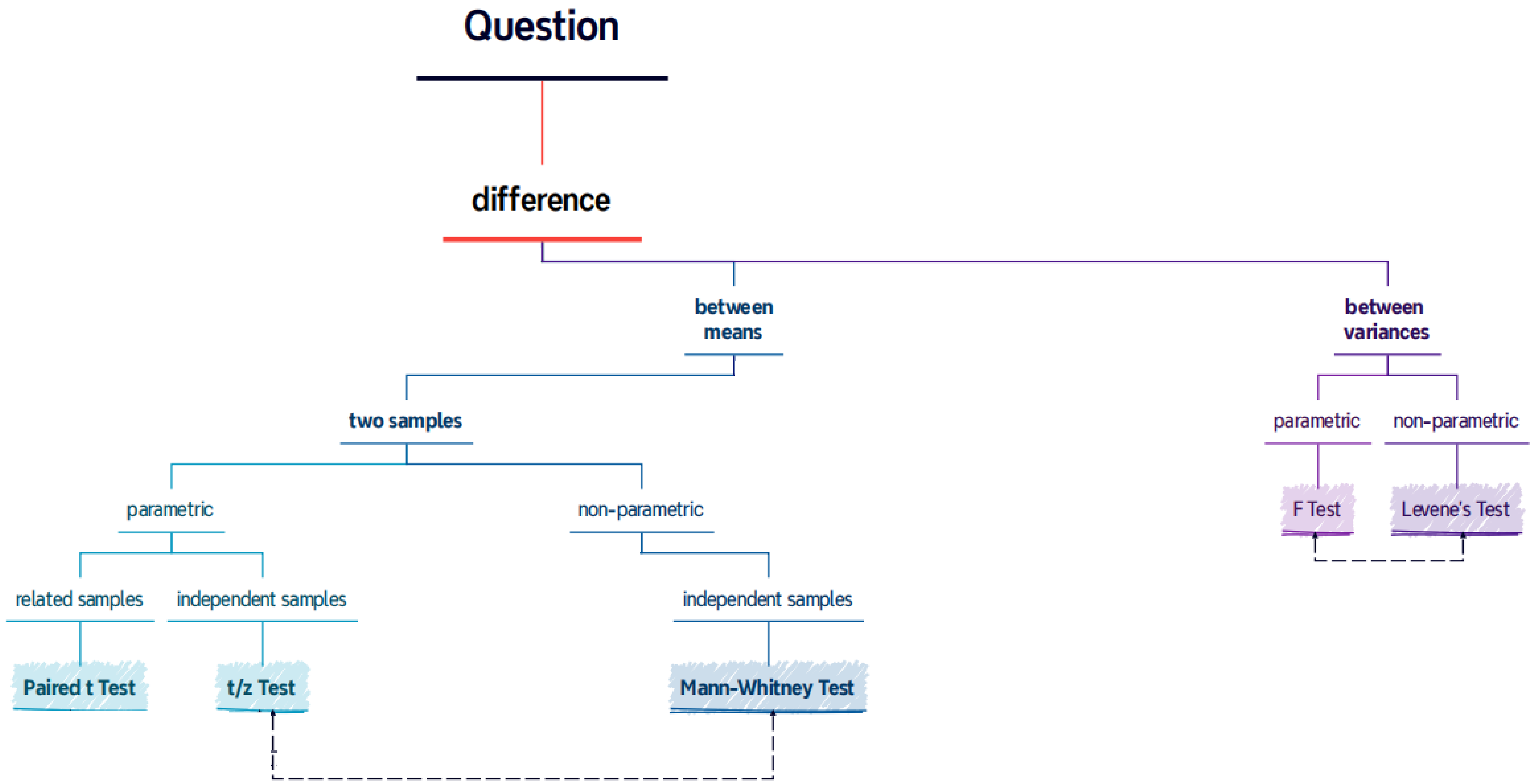
Independent Mean:NP: Mann-Whitney U-Test

Table 3 Critical values of U (5% significance).

n_1 \ n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

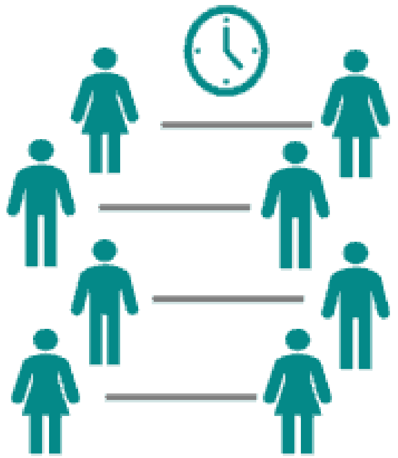
Reject H_0 if U $\leq$ critical value

Hypothesis Testing



Dependent Mean:NP: Wilcoxon Signed-rank Test

The Wilcoxon test (Wilcoxon signed-rank test) tests whether the mean values of two dependent groups differ significantly from each other.



The **Wilcoxon Test** is **non-parametric counterpart** to the **t-test for dependent samples**.

Dependent Mean:NP: Wilcoxon Signed-rank Test

Assumptions of the Wilcoxon test:

- Samples need not be normally distributed.
- Samples must be dependent.
    - **Repeat measurement:** A characteristic of a person, e.g. weight, was measured at two points in time
    - **Natural couples:** The values do not necessarily have to be from the same person but from people who belong together, for example lawyer/client, wife/husband and psychologist/patient. Of course, they do not have to be people either.
- Independence: The Wilcoxon sign test assumes independence, i.e., the paired observations are drawn randomly and independently.

*If the data are not available in pairs, the Mann-Whitney U test is used instead of the Wilcoxon test.

Dependent Mean:NP: Wilcoxon Signed-rank Test

H_0 : There is no difference (in terms of median) between the two groups in the population.
H_a : There is a difference (w.r.t. the median) between the two groups in the population.

W Statistic:

morning	evening
43	44
36	38
43	41
41	39
37	34
37	41
43	39
40	34

Dependent Mean:NP: Wilcoxon Signed-rank Test

H_0 : There is no difference (in terms of central tendency) between the two groups in the population.
H_a : There is a difference (w.r.t. the central tendency) between the two groups in the population.

W Statistic:

Total ranks

$$T^- = 1 + 3 + 6.5 = 10.5$$
$$T^+ = 3 + 3 + 5 + 6.5 + 8 = 25.5$$

Test statistics W

$$W = \min(T^-, T^+)$$
$$= \min(10.5, 25.5) = 10.5$$

Expected value of W

$$u_W = \frac{n \cdot (n+1)}{4} = \frac{8 \cdot (8+1)}{4} = 18$$

Standard deviation

$$\sigma_W = \sqrt{\frac{n \cdot (n+1) \cdot (2 \cdot n+1) - \boxed{\sum \frac{t_i^3 - t_i}{2}}}{24}}$$
$$= \sqrt{\frac{8 \cdot (8+1) \cdot (2 \cdot 8+1) - 15}{24}}$$
$$= 7.1$$

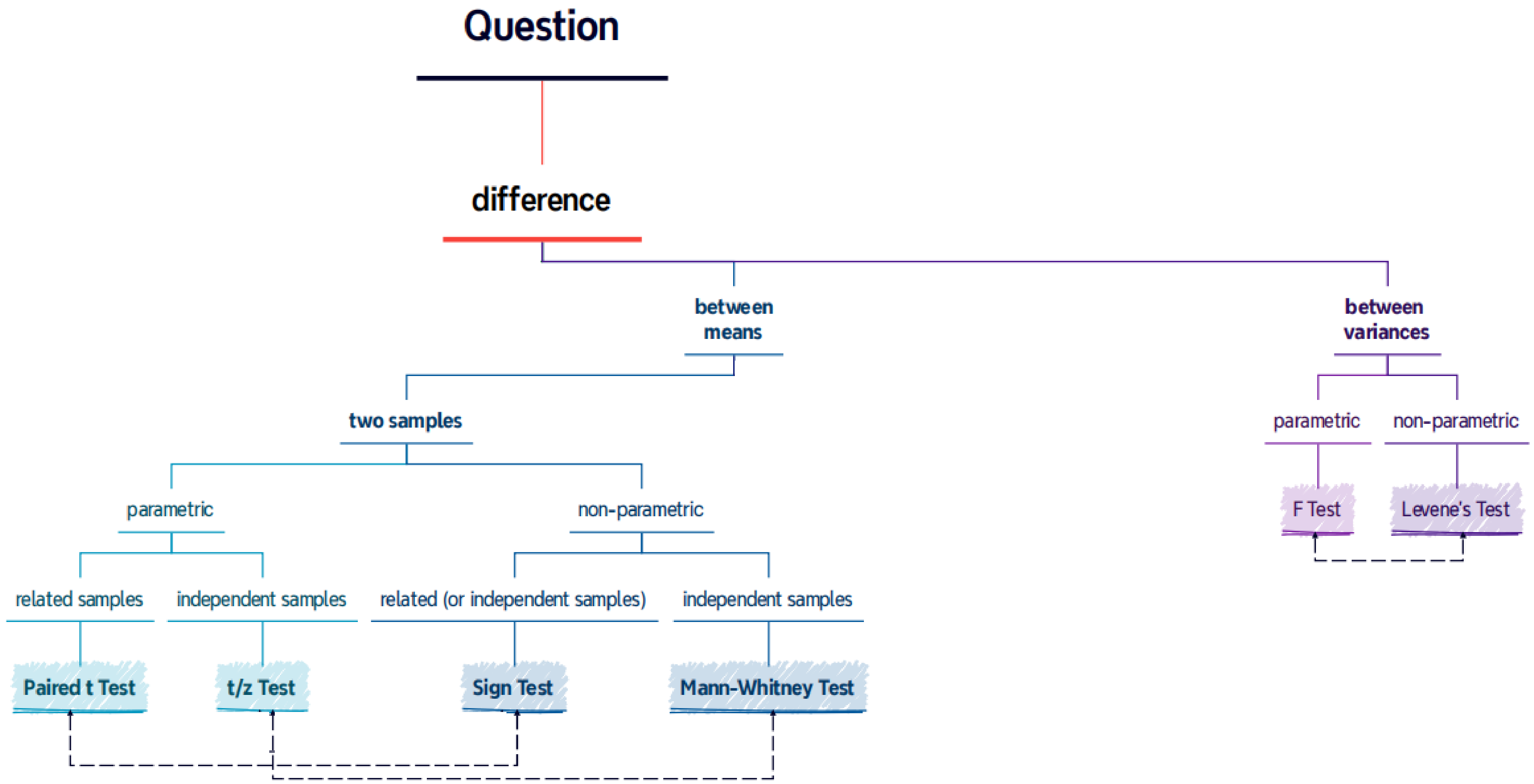
z-value

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{10,5 - 18}{7.1} = -1.06$$

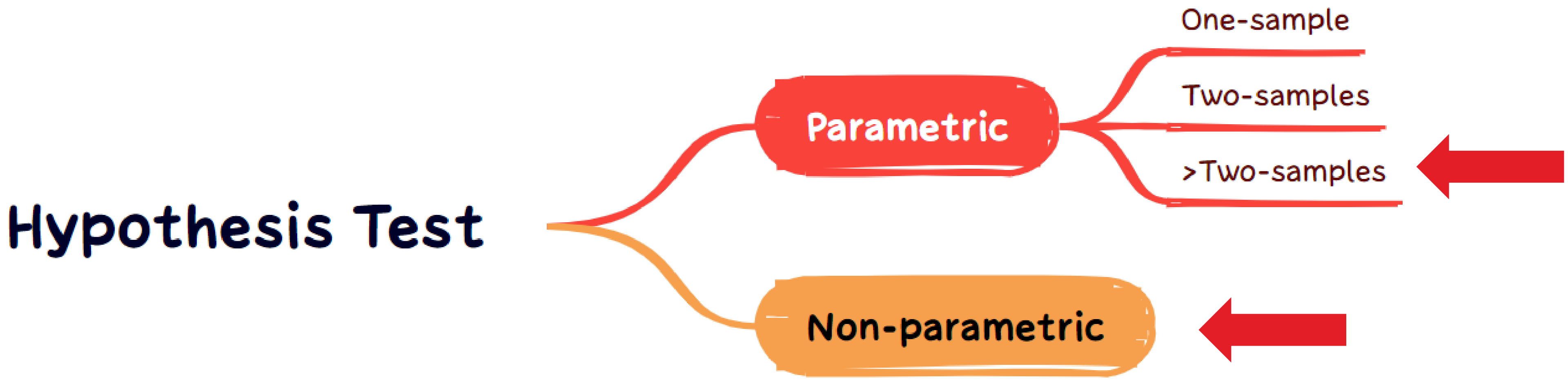
n	Two-Tailed Test		One-Tailed Test	
	α = .05	α = .01	α = .05	α = .01
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43

Reject H_0 if W $\leq$ critical value

Hypothesis Testing



Hypothesis Testing



Multiple Means:P: One-way ANOVA

ANOVA (Analysis of Variance)

- The one-way ANOVA is used to test the claim that three or more population means are equal.
- This is an extension of the two independent samples t-test.

Definition

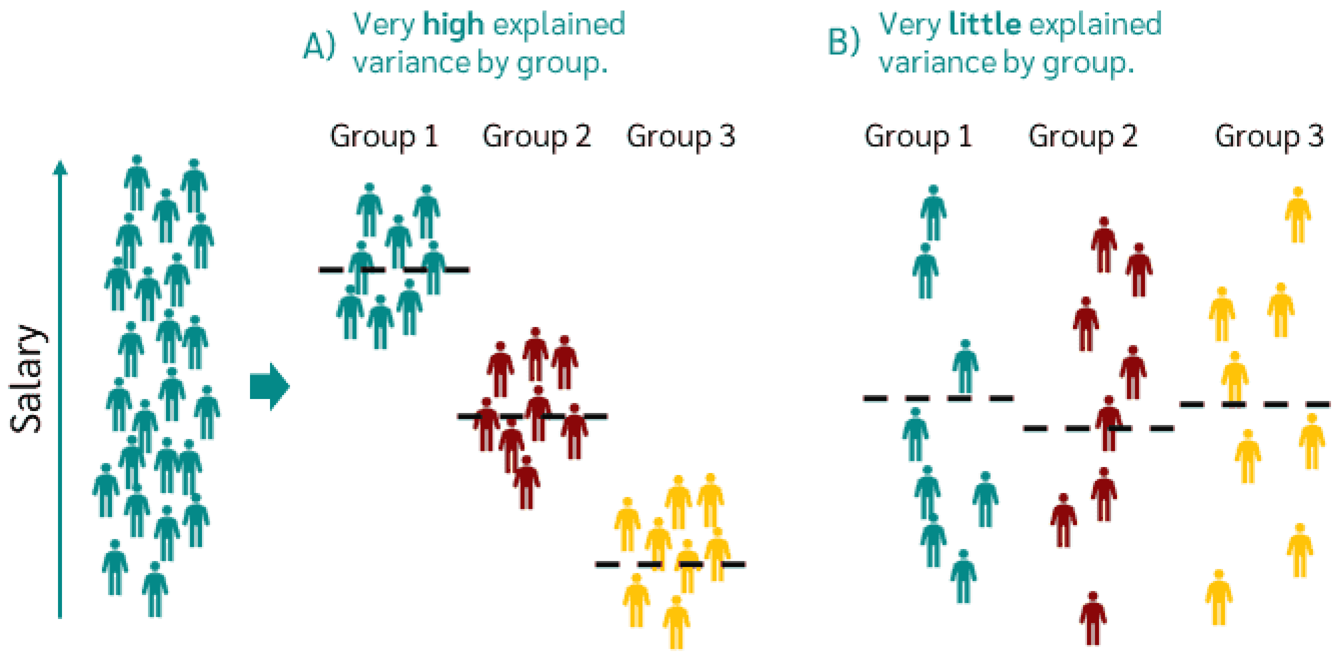
- Is there a difference in the population between the different groups of the independent variable with respect to the dependent variable?
  - The **dependent/*response* variable** is the variable you're comparing
  - The **independent/factor variable** is the categorical variable being used to define the groups
    - We will assume *k* samples (groups)

The aim of ANOVA is to explain as much variance as possible
in the dependent variable by dividing it into the groups.

Multiple Means:P: One-way ANOVA

Example

With the help of the independent variable, e.g. "highest educational qualification" with the three characteristics group 1, group 2 and group 3 should be explained as much variance of the dependent variable "salary" as possible.



	Variance within the groups	Variance between group means
Case A)	Small	Large
Case B)	Large	Small

Multiple Means:P: One-way ANOVA

Assumptions for One-way ANOVA

- Each sample are randomly sampled.
- Independent samples across groups.
- [Homogeneity] The variances of each group are assumed equal. [Levene's Test]
- [Normal Distribution] The sample within the groups should be normally distributed.

ANOVA Hypotheses

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$ (The mean value of all groups is the same.)
H_a : At least one of the means if different from the others.

Multiple Means:P: One-way ANOVA

ANOVA Sources of Variation

- Variation
  - the sum of the squares (SS) of the deviations between a value and the mean of the value
- Are all of the values identical?
  - there is some variation in the data
  - Total variation SS_{Total}
- Are all of the sample means identical?
  - there is some variation between the groups
  - Between group variation $SS_{Between}$/ SS_{Group}
- Are each of the values within each group identical?
  - there is some variation within the groups
  - Within group variation SS_{Within}/ SS_{Error}

Multiple Means:P: One-way ANOVA

ANOVA Sources of Variation

- Among Groups Variation SS_{Group}

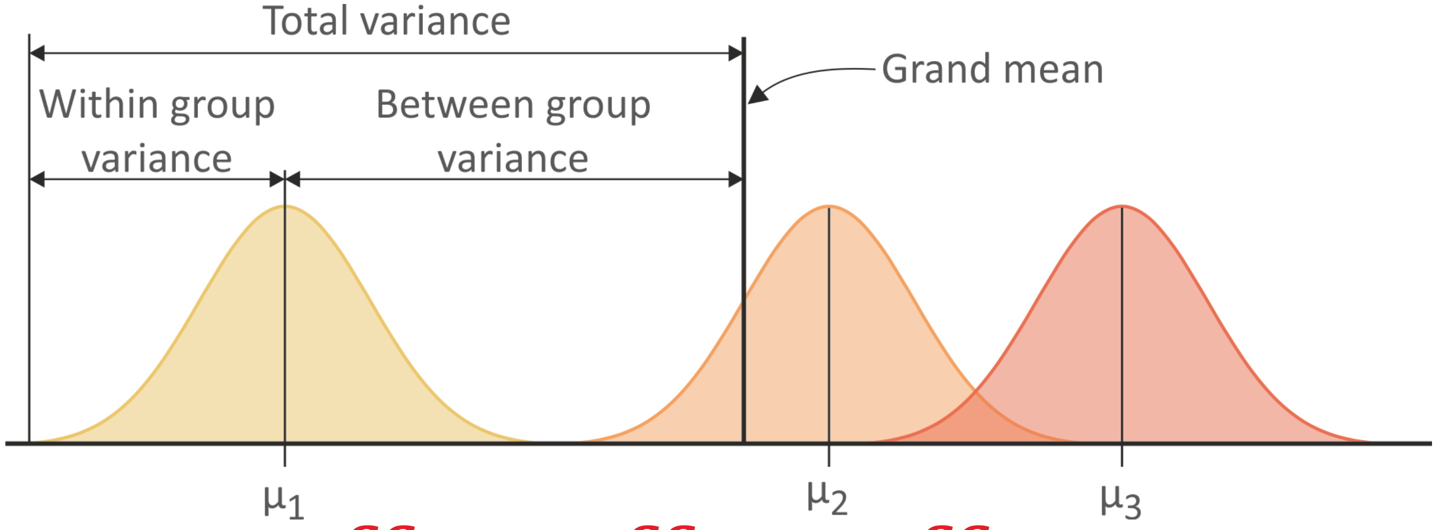
$$among - groups\ SS = \sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{X} \right)^2$$

$$among - groups\ DF = k - 1$$

- Within Group Variation SS_{Error}

$$within - groups\ SS = \sum_{i=1}^{k} \left[\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \right]$$

$$within - groups\ DF = \sum_{i=1}^{k} (n_i - 1) = N - k$$



$$SS_{Total} = SS_{Group} + SS_{Error}$$

Multiple Means:P: One-way ANOVA

ANOVA Test Statistic

To decide whether the variation associated with the among group differences are greater than the within group variation, we calculate ratios of the sums of squares. These are called Mean Squares (MS) .

- Mean Square for (among) groups

$$MS_{group} = \frac{among - groups\ SS}{among - groups\ DF}$$

- Mean Square for error

$$MS_{error} = \frac{within - groups\ SS}{within - groups\ DF}$$

- The ratio of the Mean Squares is called F, the test statistic for ANOVA.

$$F = \frac{MS_{Group}}{MS_{Error}}$$

Multiple Means:P: One-way ANOVA

Example

The Statistics classroom is divided into three rows: front, middle, and back. The instructor noticed that the further the students were from her, the more likely they were to miss class or use an instant messenger during the class. She wanted to see if the students further away did worse on the exams at a significance level of 0.01. A random sample of the students in each row was taken. The score for those students on the second exam (out of 20) was recorded.

Front: 12, 15, 17, 12

Middle: 16, 14, 21, 15, 19

Back: 14, 17, 20, 15

Multiple Means:P: One-way ANOVA

Example

Front	Middle	Back
12	16	14
15	14	17
17	21	20
12	15	15
	19	
$\bar{x}_1 = \frac{56}{4} = 14$	$\bar{x}_2 = \frac{85}{5} = 17$	$\bar{x}_3 = \frac{66}{4} = 16.5$
$s_1^2 = 6$	$s_2^2 = 8.5$	$s_3^2 = 7$

$k = 3$ (3 samples)
$N = n_1 + n_2 + n_3 = 4 + 5 + 4 = 13$ (sum of sample sizes)

- H_0: $\mu_1 = \mu_2 = \mu_3$
- H_a: At least one mean is different
- $\alpha = 0.01$
- $df_{Group} = 3 - 1 = 2$
- $df_{Error} = 13 - 3 = 10$

To find the test statistic, the following must be calculated.

$$\bar{\bar{x}} = \frac{\sum x}{N} = \frac{56+85+66}{13} \approx 15.92$$

$$MS_{Group} = \frac{SS_{Group}}{df_{Group}} = \frac{\sum n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$
$$= \frac{4(14-15.92)^2 + 5(17-15.92)^2 + 4(16.5-15.92)^2}{3-1}$$
$$\approx \frac{21.92}{2} = 10.96$$

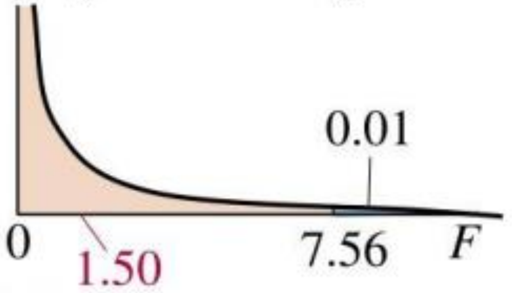
$$MS_{Error} = \frac{SS_{Error}}{df_{Error}} = \frac{\sum (n_i - 1)s_i^2}{N-k}$$
$$= \frac{(4-1)(6) + (5-1)(8.5) + (4-1)(7)}{13-3}$$
$$= \frac{73}{10} = 7.3$$

$$F = \frac{MS_{Group}}{MS_{Error}} = \frac{10.96}{7.3} \approx 1.50$$

$P(F_{2,10} > 1.5) > 0.01$

```
> 1 - pf(1.5, df1 = 2, df2 = 10)
[1] 0.2693291
```

- **Rejection Region:**



- **Decision:** Fail to Reject H_0

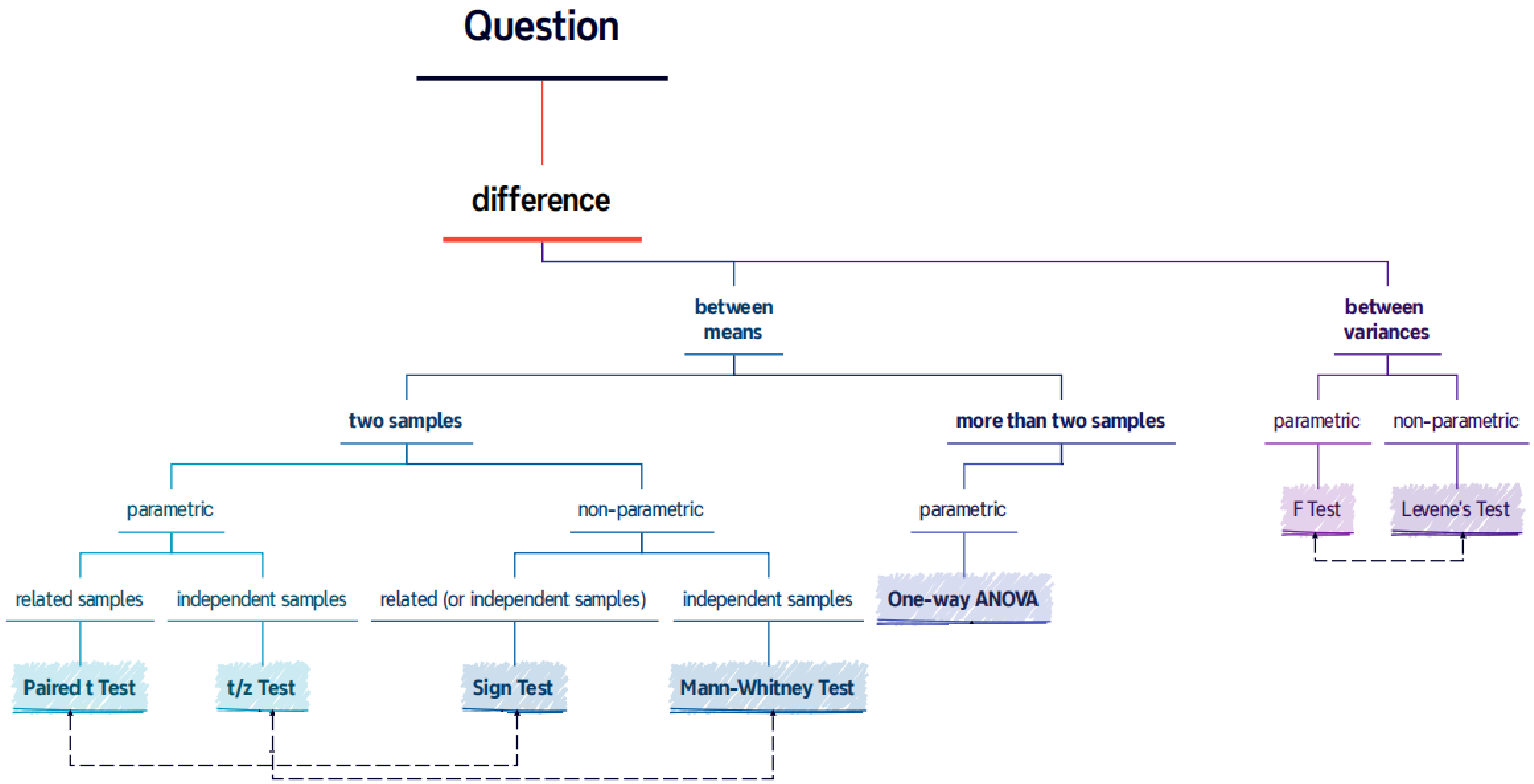
Multiple Means:P: One-way ANOVA

ANOVA Table

Source	SS	df	MS	F	p
Between/ Group					
Within/ Error					
Total					

```
> anova(lm(income ~ race, data=Income)) # analysis of variance comparing mean
Analysis of Variance Table              # income by group, for group = race
Response: income
          Df  Sum Sq  Mean Sq  F value   Pr(>F) # F = 4.24 = ratio of mean squares
race       2  3352.5  1676.23   4.2444  0.01784 # groups mean square = 1676.23
Residuals 77 30409.5   394.93                    # residual mean square = 394.93
```

Hypothesis Testing



Multiple Means:NP: Kruskal-Wallis Test

Kruskal-Wallis Test

The Kruskal-Wallis test is used to compare the medians of more than two groups, just like the one-way ANOVA.

- It is a nonparametric test that can be used to determine whether three or more independent samples were selected from populations having the same distribution.

Kruskal-Wallis Test Hypotheses

H_0 : the k distributions are identical
H_a : at least one distribution is different

Multiple Means:NP: Kruskal-Wallis Test

Kruskal-Wallis Test

- Rank the total measurements in all k samples from 1 to n. Tied observations are assigned average of the ranks they would have gotten if not tied.
- Calculate

T_i = rank sum for the i^{th} sample $\quad i = 1, 2, \dots, k$

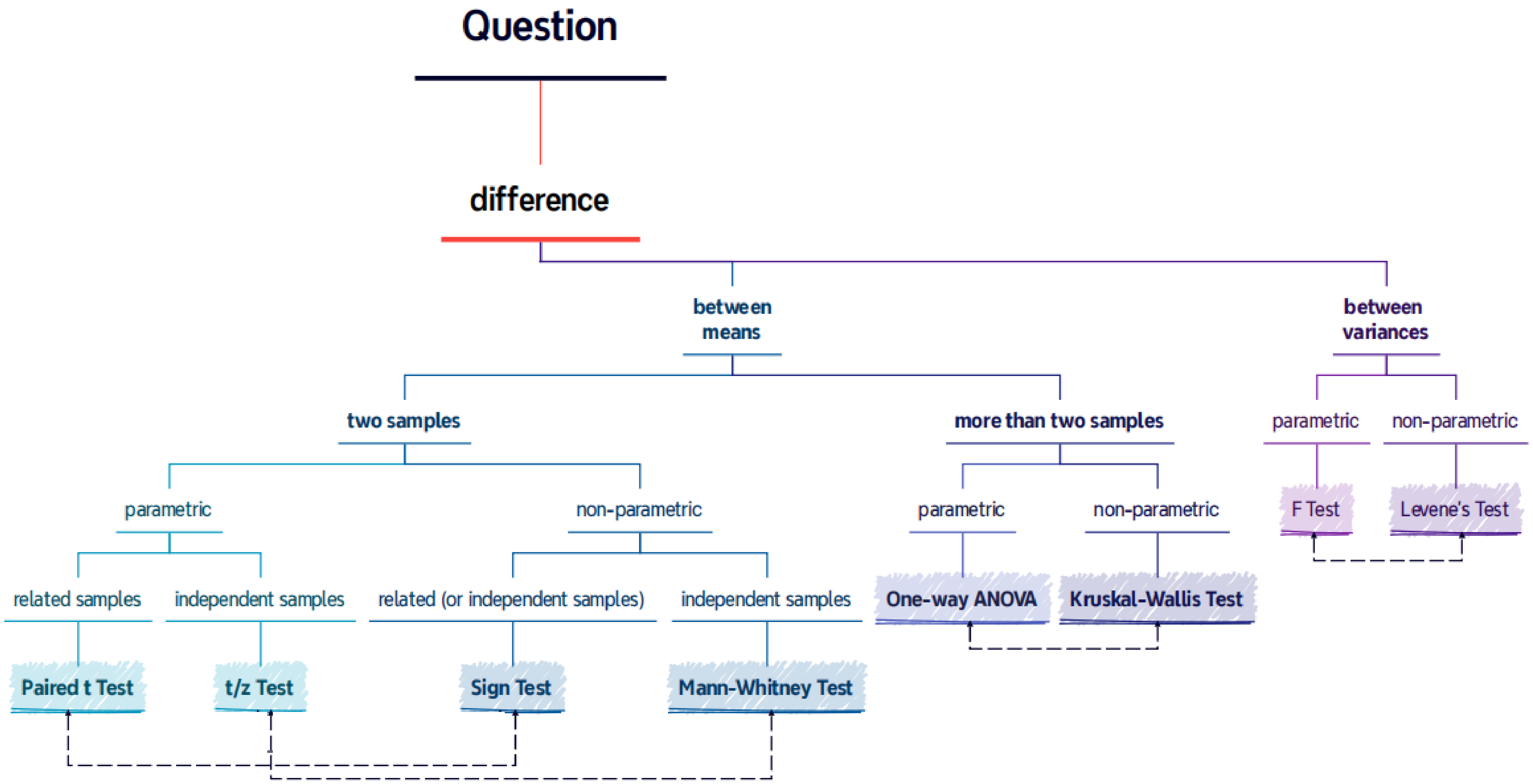
	1	2	3	4
	65 (3)	75 (7)	59 (1)	94 (16)
	87 (13)	69 (5)	78 (8)	89 (15)
	73 (6)	83 (12)	67 (4)	80 (10)
	79 (9)	81 (11)	62 (2)	88 (14)
T_i	31	35	15	55

Test Statistic

$$H = \frac{12}{n(n+1)} \sum \frac{T_i^2}{n_i} - 3(n+1)$$

When H_0 is true, the test statistic H has an approximate chi-square distribution with $df = k - 1$.

Hypothesis Testing



Hypothesis Testing

A brief review

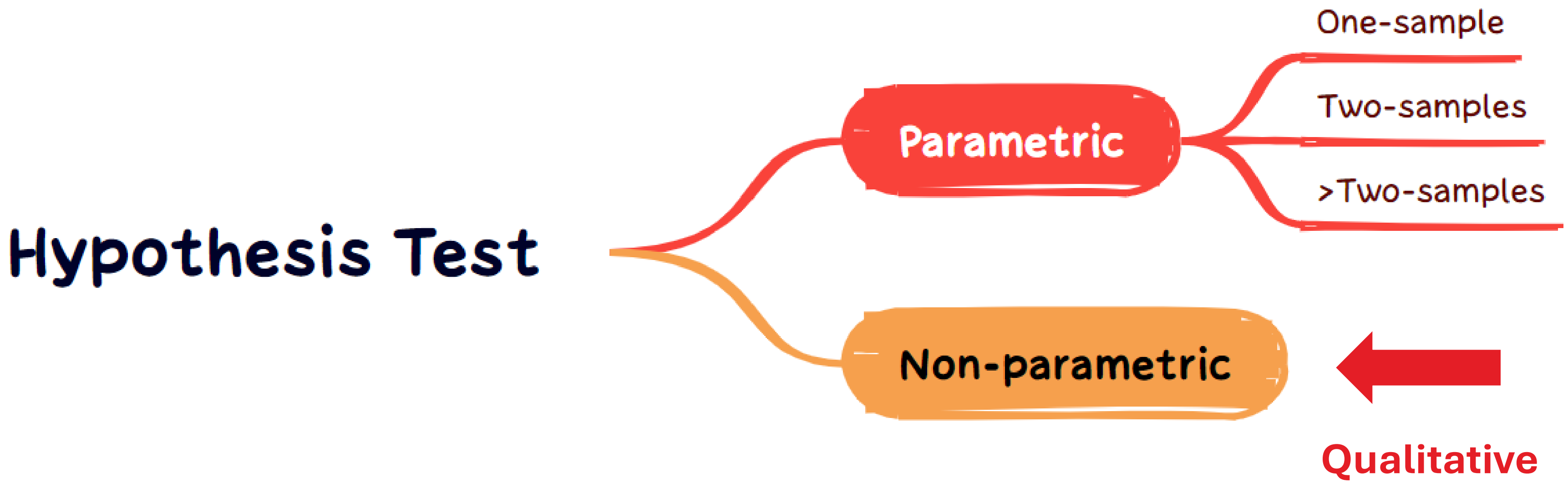
Dist'n	One-Sample Inference	Two-Sample Inference
Normal	$H_0 : \mu = \mu_0$	Paired $H_0 : \mu_D = 0$, CI for μ_D (Z or T_{n-1})
	CI for μ	2 ind samples $H_0 : \mu_1 = \mu_2$, CI for $\mu_1 - \mu_2$ ($T_{n_1+n_2-2}$)
	σ^2 is known (Z) or unknown (T_{n-1})	k ind samples $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ ($F_{k-1,N-k}$)
	$H_0 : \sigma^2 = \sigma_0^2$, CI for σ^2 (χ^2_{n-1})	$H_0 : \sigma^2 = \sigma_0^2$, CI for σ^2 (χ^2_{N-k})
Arbitrary	$H_0 : \mu = \mu_0$, CI for μ (CLT Z)	Paired $H_0 : \mu_D = 0$ (Signed rank)
		2 ind samples $H_0 : \mu_1 = \mu_2$ (Mann-Whitney)
		2 ind samples $H_0 : \sigma_1^2 = \sigma_2^2$ (Levene's)
		k ind samples $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ (Kruskal-Wallis)

Non-Parametric Test and Test Power

Now of course the question may come, why don't I just always use the non-parametric test instead of the parametric test? Then I don't need to test for normality, homogenous, ect.

Parametric tests are usually more powerful!

Hypothesis Testing



Analysis of Two-Way Tables

$r \times c$ Contingency Table

- Shows the observed frequencies for two variables.
- The observed frequencies are arranged in r rows and c columns.
- The intersection of a row and a column is called a cell.

Example:

The contingency table shows the results of a random sample of 550 company CEOs classified by age and size of company. *(Adapted from Grant Thornton LLP,The Segal Company)*

	Age				
Company size	39 and under	40 - 49	50 - 59	60 - 69	70 and over
Small / Midsize	42	69	108	60	21
Large	5	18	85	120	22

Finding the Expected Frequency

- Assuming the two variables are **independent**, you can use the contingency table to find the expected frequency for each cell.
- The expected frequency for a cell $E_{r,c}$ a contingency table is

$$\text{Expected frequency } E_{r,c} = \frac{(\text{Sum of row } r) \times (\text{Sum of column } c)}{\text{Sample size}}$$

	Age					
Company size	39 and under	40 - 49	50 - 59	60 - 69	70 and over	Total
Small / Midsize	42	69	108	60	21	300
Large	5	18	85	120	22	250
Total	47	87	193	180	43	550

marginal totals

Finding the Expected Frequency

$$\text{Expected frequency } E_{r,c} = \frac{(\text{Sum of row } r) \times (\text{Sum of column } c)}{\text{Sample size}}$$

	Age					
Company size	39 and under	40 - 49	50 - 59	60 - 69	70 and over	Total
Small / Midsize	42	69	108	60	21	300
Large	5	18	85	120	22	250
Total	47	87	193	180	43	550

$E_{1,2} = \frac{300 \cdot 87}{550} \approx 47.45$ $E_{1,3} = \frac{300 \cdot 193}{550} \approx 105.27$ $E_{2,1} = \frac{250 \cdot 47}{550} \approx 21.36$ $E_{2,2} = \frac{250 \cdot 87}{550} \approx 39.55$ $E_{2,3} = \frac{250 \cdot 193}{550} \approx 87.73$

$E_{1,4} = \frac{300 \cdot 180}{550} \approx 98.18$ $E_{1,5} = \frac{300 \cdot 43}{550} \approx 23.45$ $E_{2,4} = \frac{250 \cdot 180}{550} \approx 81.82$ $E_{2,5} = \frac{250 \cdot 43}{550} \approx 19.55$

Chi-Square Independence Test

- Used to test the independence of two variables.
- Can determine whether the occurrence of one variable affects the probability of the occurrence of the other variable.

Requirements:

- The observed frequencies must be obtained by using a random sample.
- Each expected frequency must be greater than or equal to 5.

Chi-Square Independence Test

Test Statistic

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

- Follows a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.
- O represents the observed frequencies.
- E represents the expected frequencies.

Chi-Square Independence Test

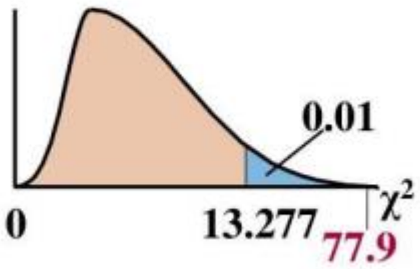
Example

Company size	Age					Total
	39 and under	40 - 49	50 - 59	60 - 69	70 and over	
Small / Midsize	42 (25.64)	69 (47.45)	108 (105.27)	60 (98.18)	21 (23.45)	300
Large	5 (21.36)	18 (39.55)	85 (87.73)	120 (81.82)	22 (19.55)	250
Total	47	87	193	180	43	550

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$
$$= \frac{(42-25.64)^2}{25.64} + \frac{(69-47.45)^2}{47.45} + \frac{(108-105.27)^2}{105.27} + \frac{(60-98.18)^2}{98.18} + \frac{(21-23.45)^2}{23.45}$$
$$+ \frac{(5-21.36)^2}{21.36} + \frac{(18-39.55)^2}{39.55} + \frac{(85-87.73)^2}{87.73} + \frac{(120-81.82)^2}{81.82} + \frac{(22-19.55)^2}{19.55}$$
$$\approx 77.9$$

- H_0: **CEOs' ages are independent of company size**
- H_a: **CEOs' ages are dependent on company size**
- α = **0.01**
- **d.f. = (2 – 1)(5 – 1) = 4**

- **Rejection Region**



- **Test Statistic:**
  $\chi^2 = 77.9$

- **Decision: Reject H_0**

There is enough evidence to conclude CEOs' ages are dependent on company size.

Chi-Square Independence Test

- There is a *distinctly different* situation in which the chi-square test may also be used.

Suppose we have **c independent random samples from c different populations**, and we classify the individuals within each sample according to a categorical variable that takes on r values.

- We now have an $\boldsymbol{r} \times \boldsymbol{c}$ table that looks the same as before but is obtained via a different sampling scheme.
- In this second situation, we may use the chi-square test to test the hypothesis:

H_0: the distribution of the categorical variable is the same in each of the *c* populations

Chi-Square Independence Test

In summary, the chi-square test can be used when sampling is done in either of the following two ways:

- Independent SRSs from two or more populations, with each individual classified according to one categorical variable
- A single SRS, with each individual classified according to both of two categorical variables

In either scenario, in order for the chi-square test to be valid, the following should be true:

- **The average of the expected cell counts should be at least 5.**
- **All individual cell counts should be at least 1.**
- **In a 2 × 2 table, all four expected cell counts should be at least 5.**

Chi-Square Goodness-of-fit

Motivation Example

A radio station claims that the distribution of music preferences for listeners in the broadcast region is as shown below.

Distribution of music Preferences			
Classical	4%	Oldies	2%
Country	36%	Pop	18%
Gospel	11%	Rock	29%

Each outcome is classified into **categories**.

The probability for each possible outcome is fixed.

Chi-Square Goodness-of-fit

Chi-Square Goodness-of-Fit Test

- Used to test whether a frequency distribution fits an expected distribution.
- The null hypothesis states that the frequency distribution fits the specified distribution.
- The alternative hypothesis states that the frequency distribution does not fit the specified distribution.

Requirements:

- The observed frequencies must be obtained by using a random sample.
- Each expected frequency must be greater than or equal to 5.

Chi-Square Goodness-of-fit

Distribution of music Preferences			
Classical	4%	Oldies	2%
Country	36%	Pop	18%
Gospel	11%	Rock	29%

Example

To test the radio station's claim at a significance level of 0.01, the executive can perform a chi-square goodness-of-fit test using the following hypotheses.

H_0 : The distribution of music preferences in the broadcast region is 4% classical, 36% country, 11% gospel, 2% oldies,18% pop, and 29%rock. (claim)

H_a : The distribution of music preferences **differs from** the claimed or expected distribution.

Chi-Square Goodness-of-fit

Example

Distribution of music Preferences			
Classical	4%	Oldies	2%
Country	36%	Pop	18%
Gospel	11%	Rock	29%

To calculate the test statistic for the chi-square goodness-of-fit test, the observed frequencies and the expected frequencies are used.

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

- The **observed frequency** $\boldsymbol{O}$ of a category is the frequency for the category observed in the sample data.

- The **expected frequency** $\boldsymbol{E}$ of a category is the calculated frequency for the category.

$$E_i = np_i$$

Chi-Square Goodness-of-fit

Distribution of music Preferences			
Classical	4%	Oldies	2%
Country	36%	Pop	18%
Gospel	11%	Rock	29%

Expected Frequency: $E_i = np_i$

Survey results (n = 500)	
Classical	8
Country	210
Gospel	72
Oldies	10
Pop	75
Rock	125

observed frequency

Type of music	% of listeners	Observed frequency	Expected frequency
Classical	4%	8	500(0.04) = 20
Country	36%	210	500(0.36) = 180
Gospel	11%	72	500(0.11) = 55
Oldies	2%	10	500(0.02) = 10
Pop	18%	75	500(0.18) = 90
Rock	29%	125	500(0.29) = 145

$n = 500$

Chi-Square Goodness-of-fit

- H_0: music preference is 4% classical, 36% country, 11% gospel, 2% oldies, 18% pop, and 29% rock
- H_a: music preference differs from the claimed or expected distribution
- $\alpha = 0.01$
- d.f. = 6 – 1 = 5

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

$$= \frac{(8-20)^2}{20} + \frac{(210-180)^2}{180} + \frac{(72-55)^2}{55} + \frac{(10-10)^2}{10} + \frac{(75-90)^2}{90} + \frac{(125-145)^2}{145}$$

$$\approx 22.713$$

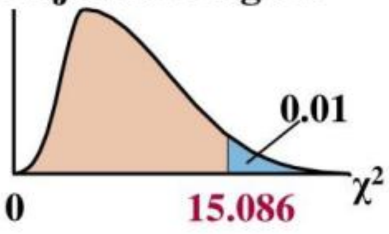
- **Test Statistic:**

  $\chi^2 = 22.713$

- **Decision:** Reject H_0

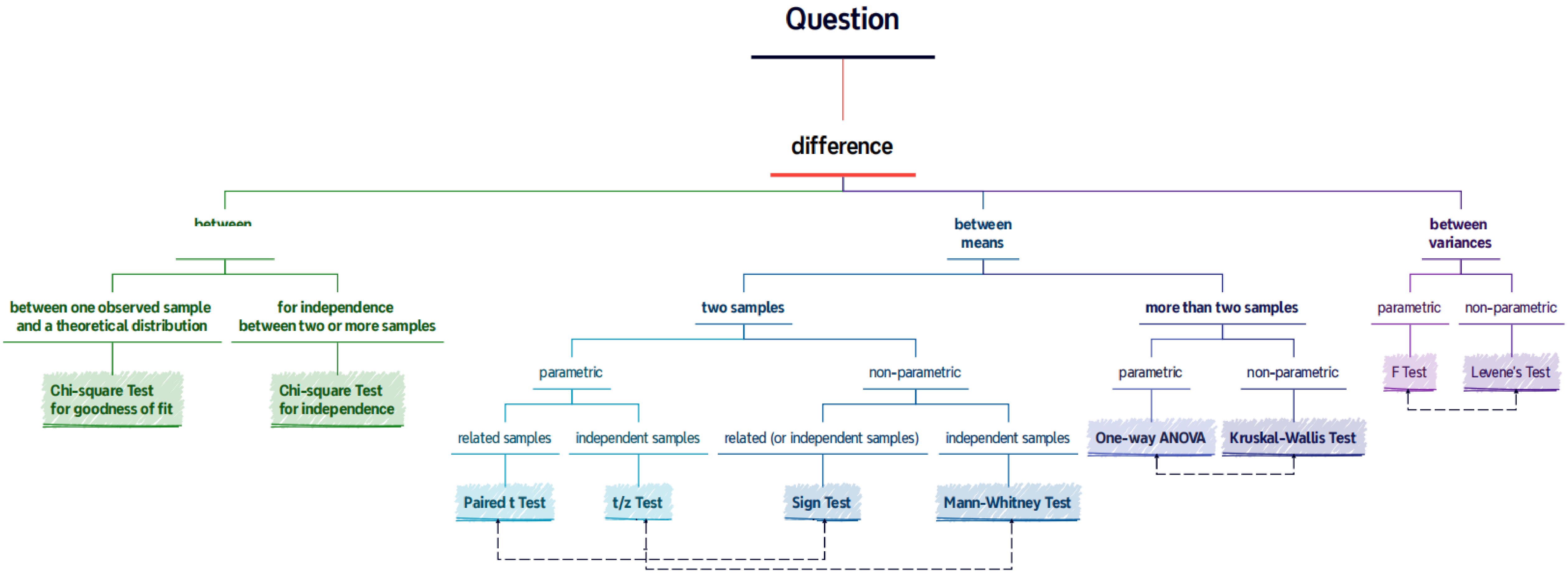
There is enough evidence to conclude that the distribution of music preferences differs from the claimed distribution.

- **Rejection Region**



Type of music	Observed frequency	Expected frequency
Classical	8	20
Country	210	180
Gospel	72	55
Oldies	10	10
Pop	75	90
Rock	125	145

Hypothesis Testing



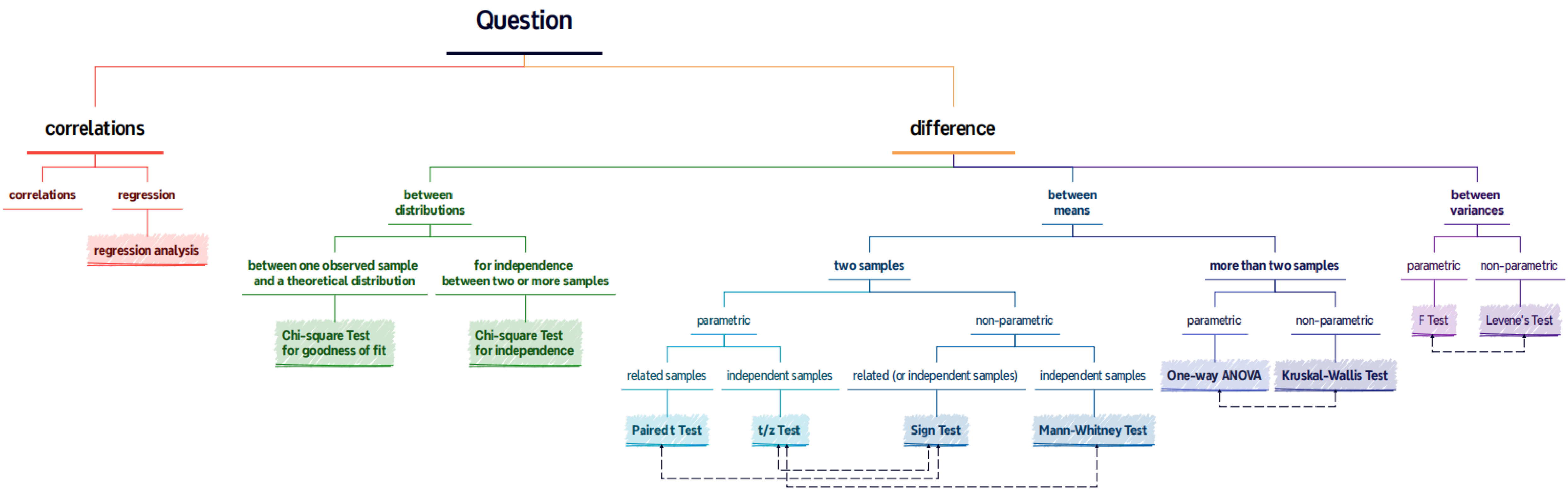
The First Question

After examining your data, ask: **does what you're testing seem to be a question of relatedness or a question of difference?**

- If **difference**, you will be testing for independence between distributions, means or variances. Different tests will be employed if your data show parametric or non parametric properties.

- If **relatedness**, you will be using tests for correlation (positive or negative) or regression.

The First Question



Khoot Time!

Lab Time