STAT 135
9. Two-sample hypothesis tests

Spring 2022

Lecturer: Dr Rebecca Barter (*she/her*)
Office hours: Tu 9:30-10:30 (in person), Th 1:30-2:30 (virtual)
Office: Evans 339

Email: rebeccabarter@berkeley.edu
Twitter: @rlbarter
GitHub: rlbarter

Two-sample z-tests: variance known

Two-sample z-tests

So far we have only asked questions about whether a population parameter (e.g., the mean or a proportion) is equal to a particular value

In practice, it is more common to ask whether the mean/proportion for **two different** populations are equal to *each other*

$X_1, \ldots, X_n$ IID from a population with unknown mean μ_1

$Y_1, \ldots, Y_m$ IID from a population with unknown mean μ_2

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$
$$H_1 : \mu_1 < \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Two-sample z-tests

$X_1, \ldots, X_n$ IID from a population with *unknown* mean μ_1 and *known* variance σ_1^2

$Y_1, \ldots, Y_m$ IID from a population with *unknown* mean μ_2 and *known* variance σ_2^2

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$
$$H_1 : \mu_1 < \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Under H_0, we have $\mu_1 - \mu_2 = 0$

Let's use this to formulate a test statistic!

$$Z = \frac{\text{estimated value} - \text{null value}}{\text{SD of estimate}} = \frac{(\bar{x}_n - \bar{y}_m) - 0}{SD(\bar{X}_n - \bar{Y}_m)} = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \overset{H_0}{\sim} N(0,1)$$

Two-sample z-tests

$$Z = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$$

Under H_0

$H_1 : \mu_1 < \mu_0$

$$\text{P-value} = \Phi\left(\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}\right)$$

$H_1 : \mu_1 > \mu_0$

$$\text{P-value} = 1 - \Phi\left(\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}\right)$$

$H_1 : \mu_1 \neq \mu_0$

$$\text{P-value} = 2\left(1 - \Phi\left(\left|\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}\right|\right)\right)$$

Two-sample t-tests: variance unknown and not equal

Two-sample t-tests: unknown unequal variance

$X_1, \ldots, X_n$ are IID *Normal* with unknown mean μ_1 and **unknown variance**

$Y_1, \ldots, Y_m$ are IID *Normal* with unknown mean μ_2 and **unknown variance**

(It's ok to have non-Normal data if your sample size is big enough)

$$T = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}} \sim t_{df?}$$

$$df = \frac{\left(\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)^2}{\frac{\hat{\sigma}_1^4}{n^2(n-1)} + \frac{\hat{\sigma}_2^4}{m^2(m-1)}}$$

When the variance of each population is different, the df is a bit complicated

Where $\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$\hat{\sigma}_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$

This is often called **Welch's t-test**

Two-sample t-tests example

Suppose I am interested in comparing the average delivery times for two pizza companies. To test this hypothesis,
I ordered 7 pizzas from pizza company A, and recorded their delivery times:

20.4, 24.2, 15.4, 21.4, 20.2, 18.5, 21.5

And 5 pizzas from company B:

20.2, 16.9, 18.4, 17.3, 20.5

Let’s assume that the two companies have Normally distributed delivery times

Do I have enough evidence to conclude that the average delivery times for each company are different?

Two-sample t-tests example

$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$

x = 20.4, 24.2, 15.4, 21.4, 20.2, 18.5, 21.5

y = 20.3, 14.9, 18.5, 17.3, 20.5

$n = 7$ $m = 5$

$\bar{x} = 20.2, \quad \bar{y} = 18.7$

$\hat{\sigma}_1 = 2.74 \quad \hat{\sigma}_2 = 1.64$

$$df = \frac{\left(\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)^2}{\frac{\hat{\sigma}_1^4}{n^2(n-1)} + \frac{\hat{\sigma}_2^4}{m^2(m-1)}} = 9.8$$

$$t = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}} = \frac{20.2 - 18.7}{\sqrt{\frac{2.74^2}{7} + \frac{1.64^2}{5}}} = 1.24$$

p-value $= P(|T| \geq |t|)$ $T \sim t_{9.8}$

$= 2(1 - P(T \leq 1.24))$

$= 0.25$

So not enough evidence to reject that the two pizza places have the same average delivery times

Two-sample t-tests: variance unknown but equal

(pooled variance)

Two-sample t-tests: pooled variance

$X_1, \ldots, X_n$ are IID N*ormal* with unknown mean μ_1 and **unknown variance** σ^2

$Y_1, \ldots, Y_m$ are IID N*ormal* with unknown mean μ_2 and **unknown variance** σ^2

Assume that the two populations have **equal** (but unknown) variance

$$T = \frac{\bar{x}_n - \bar{y}_m}{\hat{\sigma}_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

Where we use the **pooled variance**

$$\hat{\sigma}_p^2 = \frac{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}{n+m-2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{n+m-2}$$

Two-sample t-tests: pooled variance

Unpooled test statistic

$$\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}}$$

Pooled test statistic

$$\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\hat{\sigma}_p^2}{n} + \frac{\hat{\sigma}_p^2}{m}}}$$

$$\hat{\sigma}_p^2 = \frac{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}{n+m-2}$$

When $n = m$,

$$\hat{\sigma}_p^2 = \frac{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)}{2}$$

The pooled variance involves using *all* of the data to estimate the variance, and so is typically a better estimate.

…But only when your assumption that the two populations have equal variance is true!

Two-sample t-tests (pooled var) example

Suppose I am interested in comparing the average delivery times for two pizza companies. To test this hypothesis,
I ordered 7 pizzas from pizza company A, and recorded their delivery times:

20.4, 24.2, 15.4, 21.4, 20.2, 18.5, 21.5

And 5 pizzas from company B:

20.2, 16.9, 18.4, 17.3, 20.5

Let's assume that the two companies have Normally distributed delivery times **with the *same variance*.**

Do I have enough evidence to conclude that the average delivery times for each company are different?

Two-sample t-tests (pooled var) example

$H_0 : \mu_1 = \mu_2$ $\quad$ $H_0 : \mu_1 \neq \mu_2$

x = 20.4, 24.2, 15.4, 21.4, 20.2, 18.5, 21.5

y = 20.3, 14.9, 18.5, 17.3, 20.5

$n = 7 \qquad m = 5$

$\bar{x} = 20.2, \quad \bar{y} = 18.7$

$\hat{\sigma}_1 = 2.74 \quad \hat{\sigma}_2 = 1.64$

$$\hat{\sigma}_p = \sqrt{\frac{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}{n+m-2}} = 2.36$$

$$t = \frac{\bar{x}_n - \bar{y}_m}{\hat{\sigma}_p\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{20.2 - 18.7}{2.36\sqrt{\frac{1}{7}+\frac{1}{5}}} = 1.13$$

$$\text{p-value} = P(|T| \geq |t|) \qquad T \sim t_{n+m-2} = t_{10}$$

$$= 2(1 - P(T \leq 1.13))$$

$$= 0.28$$

So not enough evidence to reject that the two pizza places have the same average delivery times

Two-sample t-test in R

See two_sample_test.R

Non-parametric two-sample test

Mann-whitney test

Mann-Whitney test

What if you don't want to assume your data is normal?

$X_1, \ldots, X_n$ are IID with unknown distribution F

$Y_1, \ldots, Y_m$ are IID with unknown distribution G

The Mann-Whitney test checks if there is a difference in the **ranks** of the two samples

It actually tests

$$H_0 : F = G \qquad H_1 : F \neq G$$

Rank sums

Note the rank *ignores which sample* the observation came from

Pizza place	Value	Rank
A	20.4	8
A	24.2	12
A	15.4	1
A	21.4	10
A	20.2	6.5
A	18.5	5
A	21.5	11
B	20.2	6.5
B	16.9	2
B	18.4	4
B	17.3	3
B	20.5	9

R_1 = sum of ranks from sample 1

$$= 8 + 12 + 1 + 10 + 6.5 + 5 + 11$$

$$= 53.5$$

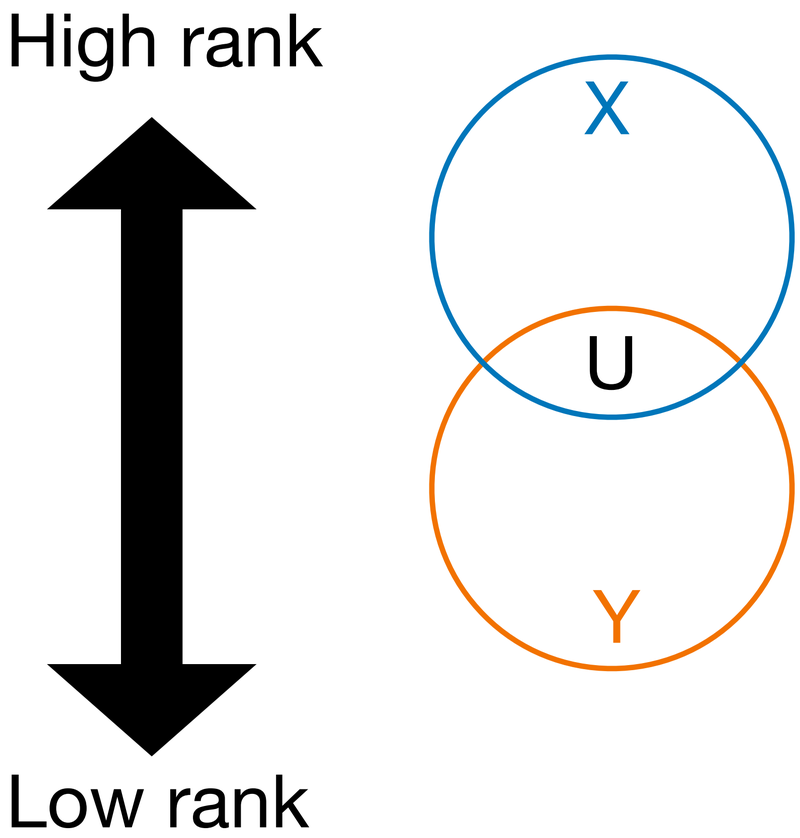
R_2 = sum of ranks from sample 2

$$= 6.5 + 2 + 4 + 3 + 9$$

$$= 24.5$$

Mann-Whitney U-statistic

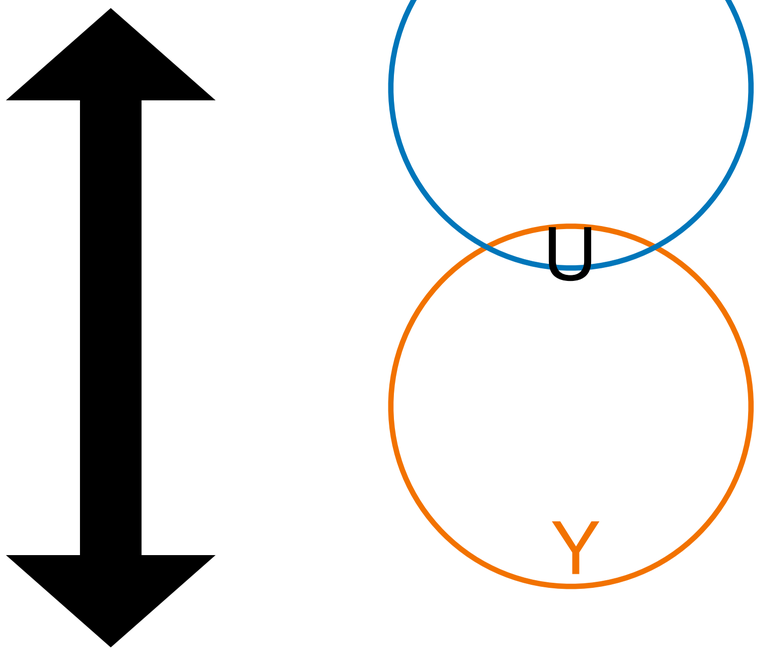
The U test statistic computes the amount of *overlap* in the ranks in each sample



Mann-Whitney U-statistic

The U test statistic computes the amount of *overlap* in the ranks in each sample

High rank



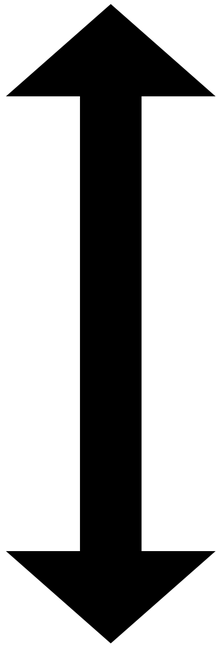
Low rank

Smaller U - Bigger difference between groups

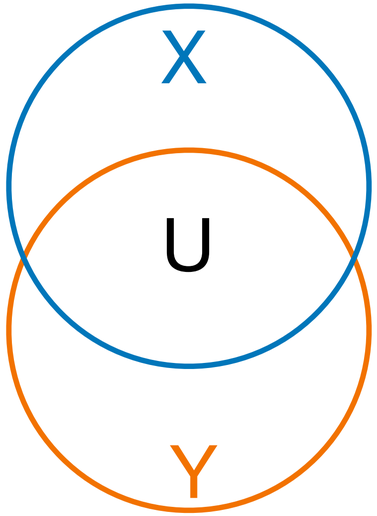
Mann-Whitney U-statistic

The U test statistic computes the amount of *overlap* in the ranks in each sample

High rank



Low rank



Smaller U - Bigger difference between groups

Bigger U - Smaller difference between groups

A smaller U test statistic is <u>more</u> significant

Mann-Whitney U-statistic

1. Compute the rank sum of each group

2. Identify which group has the smaller rank sum

3. For each data point in this group (with smaller rank sum), add up how many data points in the *other* group (with the larger sum rank) are smaller in rank (ties get 0.5)

4. Compare with the critical value

Rank sums

Pizza place	Value	Rank
A	20.4	8
A	24.2	12
A	15.4	1
A	21.4	10
A	20.2	6.5
A	18.5	5
A	21.5	11
B	20.2	6.5
B	16.9	2
B	18.4	4
B	17.3	3
B	20.5	9

$$R_1 = 53.5$$

$$R_2 = 24.5$$

Sample 2 (pizza place B) has the lower rank sum, so this will be our reference group

Rank sums

Pizza place	Value	Rank
A	20.4	8
A	24.2	12
A	15.4	1
A	21.4	10
A	20.2	6.5
A	18.5	5
A	21.5	11
B	20.2	6.5
B	16.9	2
B	18.4	4
B	17.3	3
B	20.5	9

$$U = 2.5+$$

Number of deliveries from pizza place A with shorter delivery times than this first delivery from pizza place B (ties are 0.5)

Rank sums

Pizza place	Value	Rank
A	20.4	8
A	24.2	12
A	15.4	1
A	21.4	10
A	20.2	6.5
A	18.5	5
A	21.5	11
B	20.2	6.5
B	16.9	2
B	18.4	4
B	17.3	3
B	20.5	9

$$U = 2.5 + 1 +$$

Rank sums

Pizza place	Value	Rank
A	20.4	8
A	24.2	12
A	15.4	1
A	21.4	10
A	20.2	6.5
A	18.5	5
A	21.5	11
B	20.2	6.5
B	16.9	2
B	18.4	4
B	17.3	3
B	20.5	9

$$U = 2.5 + 1 + 1 +$$

Rank sums

Pizza place	Value	Rank
A	20.4	8
A	24.2	12
A	15.4	1
A	21.4	10
A	20.2	6.5
A	18.5	5
A	21.5	11
B	20.2	6.5
B	16.9	2
B	18.4	4
B	17.3	3
B	20.5	9

$$U = 2.5 + 1 + 1 + 1 +$$

Rank sums

Pizza place	Value	Rank
A	20.4	8
A	24.2	12
A	15.4	1
A	21.4	10
A	20.2	6.5
A	18.5	5
A	21.5	11
B	20.2	6.5
B	16.9	2
B	18.4	4
B	17.3	3
B	20.5	9

$$U = 2.5 + 1 + 1 + 1 + 4 \quad = 9.5$$

Mann-Whitney U-statistic

Pizza place	Value	Rank
A	20.4	8
A	24.2	12
A	15.4	1
A	21.4	10
A	20.2	6.5
A	18.5	5
A	21.5	11
B	20.2	6.5
B	16.9	2
B	18.4	4
B	17.3	3
B	20.5	9

There is a formula for U!

$$n = 7 \quad m = 5 \qquad R_1 = 53.5 \qquad R_2 = 24.5$$

$$U_1 = mn + \frac{n(n+1)}{2} - R_1 = 9.5$$

$$U_2 = mn + \frac{m(m+1)}{2} - R_2 = 25.5$$

$$U = \min(U_1, U_2) = 9.5$$

Mann-Whitney test critical values

Nondirectional α=.05 (Directional α=.025)

	n_2														
n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	0	0	0	0	1	1	1	1
3	-	-	-	-	0	1	1	2	2	3	3	4	4	5	5
4	-	-	-	0	1	2	3	4	4	5	6	7	8	9	10
5	-	-	0	1	2	3	5	6	7	8	9	11	12	13	14
6	-	-	1	2	3	5	6	8	10	11	13	14	16	17	19
7	-	-	1	3	5	6	8	10	12	14	16	18	20	22	24
8	-	0	2	4	6	8	10	13	15	17	19	22	24	26	29
9	-	0	2	4	7	10	12	15	17	21	23	26	28	31	34
10	-	0	3	5	8	11	14	17	20	23	26	29	33	36	39
11	-	0	3	6	9	13	16	19	23	26	30	33	37	40	44
12	-	1	4	7	11	14	18	22	26	29	33	37	41	45	49
13	-	1	4	8	12	16	20	24	28	33	37	41	45	50	54
14	-	1	5	9	13	17	22	26	31	36	40	45	50	55	59
15	-	1	5	10	14	19	24	29	34	39	44	49	54	59	64

$U = 9.5$ is **not** <u>less than</u> 5

So we do not reject H_0 at the 0.05 level

Mann-Whitney U-statistic: normal approximation

$$U = \min(U_1, U_2)$$

Under H_0, and if there are **no ties**

$$E(U) = \frac{nm}{2} \qquad Var(U) = \frac{nm(n+m+1)}{12}$$

And $$\frac{U - E(U)}{\sqrt{Var(U)}} \sim N(0,1)$$

So a p-value can be approximated using can be approximated (depending on the format of the alternative hypothesis)

$$\Phi\left(\frac{u - E(u)}{\sqrt{Var(u)}}\right), \qquad 1 - \Phi\left(\frac{u - E(u)}{\sqrt{Var(u)}}\right), \qquad 2\left(1 - \Phi\left(\frac{u - E(u)}{\sqrt{Var(u)}}\right)\right)$$

Why would we want to do this?

So we don't have to use a table!

If interested in these derivations, see:
https://www.real-statistics.com/non-parametric-tests/mann-whitney-test/mann-whitney-test-advanced/

Mann-Whitney: normal approximation with ties

$$U = \min(U_1, U_2)$$

Under H_0, and if there are **<u>are</u> ties**

$$E(U) = \frac{nm}{2} \qquad Var(U) = \frac{nm(n+m+1)}{12} - \frac{nm\sum_{i=1}^{K}(t_k^3 - t_k)}{12(n+m)(n+m-1)}$$

Adjustment for ties
(t_k is the number of ties
for the kth rank)

Then $$\frac{U - E(U)}{\sqrt{Var(U)}} \sim N(0,1)$$

Mann-Whitney test in R

See two_sample_test.R

Two-sample test for proportions

Two-sample tests of proportions

$$X_1, \ldots, X_n \sim Bernoulli(p_1),$$

$$Y_1, \ldots, Y_m \sim Bernoulli(p_2)$$

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$

$$z = \frac{\text{estimate} - \text{null value}}{SD_{H_0}(\text{estimate})} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} \overset{\text{Under } H_0}{\sim} N(0,1)$$

Under H_0, $p := p_1 = p_2$, so we essentially use a **pooled** estimate of the variance:

$$\hat{p} := \frac{\sum_i X_i + \sum_i Y_i}{n + m}$$

Two-sample tests of proportions: example

A random sample of 800 adult Americans asked were “Should the federal tax on cigarettes be raised to pay for health care reform?”. We are interested in identifying whether smokers and non-smokers feel differently about cigarette taxes.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

	Non-smokers	**Smokers**
“Yes”	351	41
“No”	254	154
Total	605	195

Two-sample tests of proportions: example

A random sample of 800 adult Americans asked were “Should the federal tax on cigarettes be raised to pay for health care reform?”. We are interested in identifying whether smokers and non-smokers feel differently about cigarette taxes.

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

Non-smokers	Smokers
$n = 605$	$m = 195$
$\sum_i X_i = 351$ said “yes”	$\sum_i Y_i = 41$ said “yes”
$\hat{p}_1 = \frac{351}{605} = 0.58$	$\hat{p}_2 = \frac{41}{195} = 0.21$

$$\hat{p} = \frac{351 + 41}{605 + 195} = 0.49$$

Two-sample tests of proportions: example

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

$$\hat{p} = 0.49$$

Non-smokers	Smokers
$n = 605$	$m = 195$
$\sum_i X_i = 351$ said "yes"	$\sum_i Y_i = 41$ said "yes"
$\hat{p}_1 = \frac{351}{605} = 0.58$	$\hat{p}_2 = \frac{41}{195} = 0.21$

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.58 - 0.21}{\sqrt{0.49 \times 0.51\left(\frac{1}{195} + \frac{1}{605}\right)}} = 8.99$$

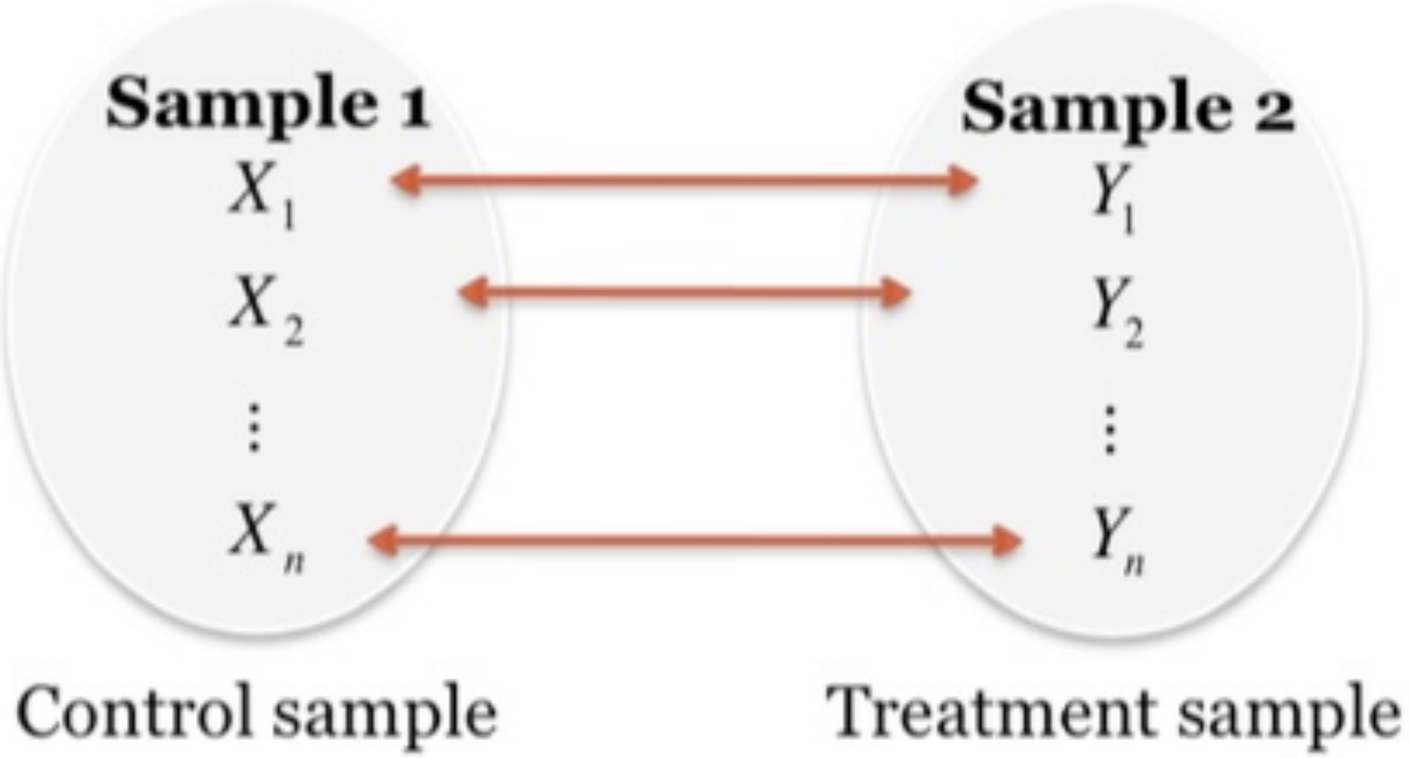
$$\text{P-value} = 2(1 - \Phi(8.99)) \approx 0$$

Note the sample size is bigger than 30 so the CLT should have kicked in!

Paired two-sample tests

Paired two-sample test

What if we are comparing *paired* observations from two groups?

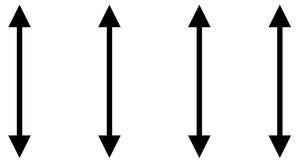


E.g., (X_i, Y_i) might be

- The salary of twins
- Weight before and after an intervention
- Midterm and final exam for the same set of students

Paired two-sample Z-test

$X_1, \ldots, X_n$ are IID



$Y_1, \ldots, Y_n$ are IID

A classic two-sample Z-test would test:

$$H_0 : \mu_X = \mu_Y$$

$$H_1 : \mu_X \neq \mu_Y$$

But if the data are paired, this can be **reduced to a single-sample test** that the mean *difference* equals 0!

Define $D_i = X_i - Y_i$

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

Paired two-sample Z-test

A **paired test can be reduced to a single-sample test** that the mean difference equals 0!

Define $D_i = X_i - Y_i$

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

Under H_0

$$Z = \frac{\bar{D} - 0}{\sigma_D / \sqrt{n}} \sim N(0,1)$$

P-value = $$P\left(|Z| \geq \left| \frac{\bar{d}}{\sigma_d / \sqrt{n}} \right| \right) = 2\left(1 - \Phi\left(\left| \frac{\bar{d}}{\sigma_d / \sqrt{n}} \right| \right) \right)$$

Paired two-sample t-test

If we don't know σ_d...

Define $D_i = X_i - Y_i$

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

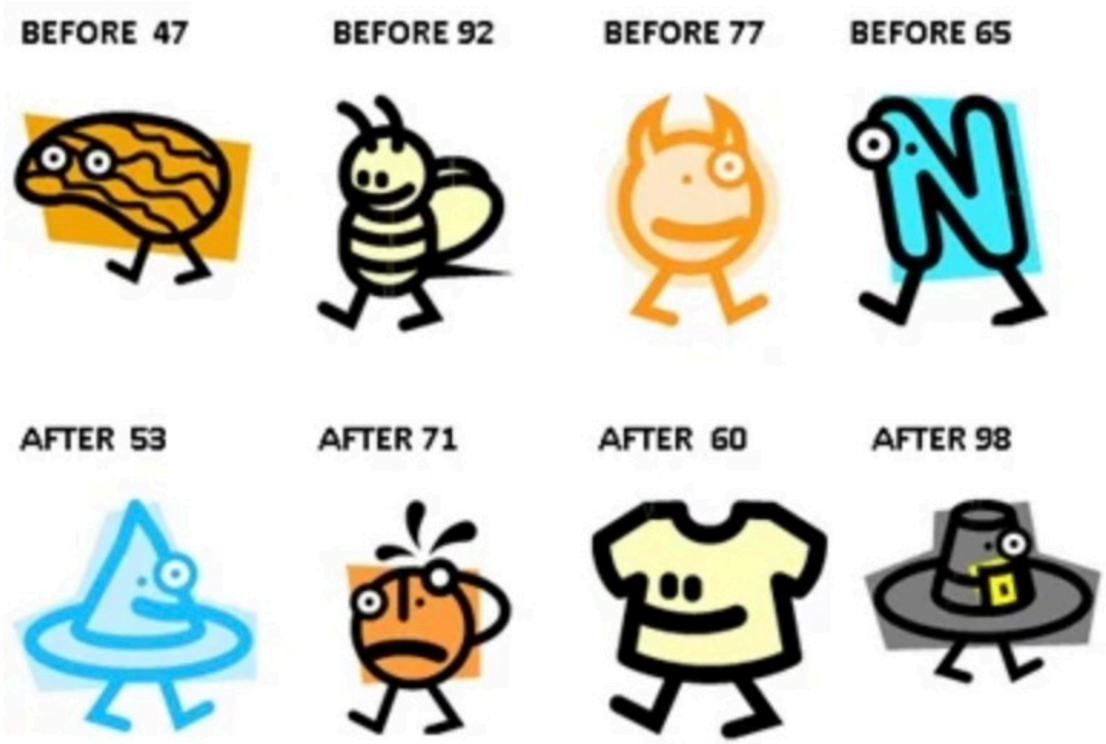
$$T = \frac{\bar{D} - 0}{\hat{\sigma}_D/\sqrt{n}} \overset{\text{Under } H_0}{\sim} t_{n-1}$$

$$\text{P-value} = P\left(|T| \geq \left| \frac{\bar{d}}{\hat{\sigma}_d/\sqrt{n}} \right| \right) = 2\left(1 - P\left(t_{n-1} \leq \left| \frac{\bar{d}}{\hat{\sigma}_d/\sqrt{n}} \right| \right) \right)$$

Comparing the unpaired and paired two-sample test

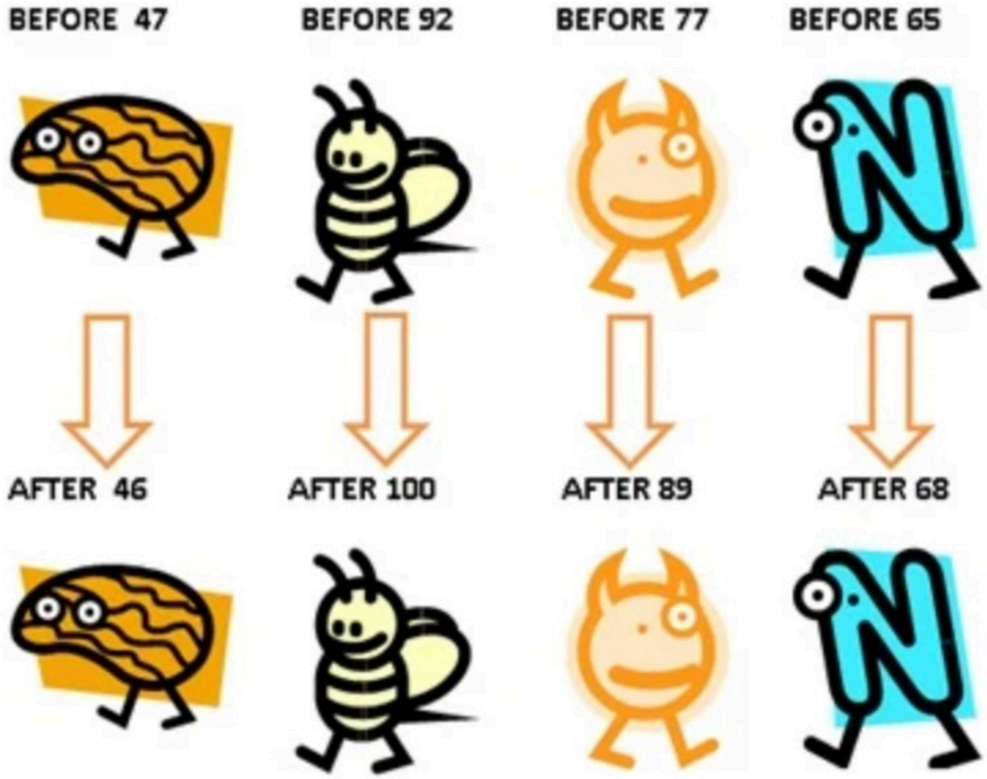
Unpaired two-sample test:

X_i, Y_is are independent



Paired two-sample test:

X_i, Y_is are *not* independent



Adorable images stolen from https://blog.minitab.com/en/statistics-and-quality-data-analysis/t-for-2-should-i-use-a-paired-t-or-a-2-sample-t

Comparing the unpaired and paired two-sample test

Unpaired two-sample test:

X_i, Y_is are independent

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{Var(\bar{X} - \bar{Y})}}$$

$$Var(\bar{X} - \bar{Y}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2)$$

Paired two-sample test:

X_i, Y_is are *not* independent

Define $D_i = X_i - Y_i$

$$T = \frac{\bar{D}}{\sqrt{Var(\bar{D})}} = \frac{\bar{X} - \bar{Y}}{\sqrt{Var(\bar{D})}}$$

$$Var(\bar{D}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y)$$

Covariance

Comparing the unpaired and paired two-sample test

Unpaired two-sample test:

$$Var(\bar{X} - \bar{Y}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2)$$

Paired two-sample test:

$$Var(\bar{D}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y)$$

Let's look at the case where $\sigma_X = \sigma_Y$

$$Var(\bar{X} - \bar{Y}) = \frac{2\sigma^2}{n}$$

$$Var(\bar{D}) = \frac{2\sigma^2(1-\rho)}{n}$$

If the correlation is $\rho = 0.5$, then a paired test with n pairs of subjects is equivalent to an unpaired design with $2n$ subjects per sample

Paired two-sample test example

A study is run to evaluate the effectiveness of an exercise program in reducing blood pressure in patients with pre-hypertension.

The study has 15 patients, whose blood pressure is measured both before and after completing the exercise program

X = blood pressure measurements *before* program (assume IID normal)

(125, 132, 138, 120, 125, 127, 136, 139, 131, 132, 135, 136, 128, 127, 130)

Y = blood pressure measurements *after* program (assume IID normal)

(118, 134, 130, 124, 105, 130, 130, 132, 123, 128, 126, 140, 135, 126, 132)

If $D_i = X_i - Y_i$

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 > \mu_2 \end{aligned} \iff \begin{aligned} H_0 &: \mu_d = 0 \\ H_1 &: \mu_d > 0 \end{aligned}$$

Paired two-sample test example

Patient	BP before program (x)	BP after program (y)	Difference (d)
1	125	118	7
2	132	134	-2
3	138	130	8
4	120	124	-4
5	125	105	20
6	127	130	-3
7	136	130	6
8	139	132	7
9	131	123	8
10	132	128	4
11	135	126	9
12	136	140	-4
13	128	135	-7
14	127	126	1
15	130	132	-2

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d > 0$$

$$\bar{d} = 3.2 \qquad \hat{\sigma}_d = 7.1$$

$$\mathrm{T} = \frac{\bar{D} - 0}{\hat{\sigma}_D / \sqrt{n}} \text{ Under } H_0 \sim t_{n-1}$$

$$t = \frac{3.2}{7.1/\sqrt{(15)}} = 1.75$$

$$\text{p-value} = P(T \geq 1.75) = 0.05$$

Paired two-sample test in R

See paired_two_sample_test.R

Non-parametric paired two-sample test

The sign test

The wllcoxon signed rank test

Sign test example

Patient	BP before program (x)	BP after program (y)	Difference (d)	Sign
1	125	118	7	+
2	132	134	-2	-
3	138	130	8	+
4	120	124	-4	-
5	125	105	20	+
6	127	130	-3	-
7	136	130	6	+
8	139	132	7	+
9	131	123	8	+
10	132	128	4	+
11	135	126	9	+
12	136	140	-4	-
13	128	135	-7	-
14	127	126	1	+
15	130	132	-2	-

X IID from dist F

Y IID from dist G

$$H_0 : F = G \qquad H_1 : F > G$$

Under H_0, what trend would you expect to see for the *signs* of the differences?

You would expect an equal number of positive and negative differences

Sign test example

Patient	BP before program (x)	BP after program (y)	Difference (d)	Sign
1	125	118	7	+
2	132	134	-2	-
3	138	130	8	+
4	120	124	-4	-
5	125	105	20	+
6	127	130	-3	-
7	136	130	6	+
8	139	132	7	+
9	131	123	8	+
10	132	128	4	+
11	135	126	9	+
12	136	140	-4	-
13	128	135	-7	-
14	127	126	1	+
15	130	132	-2	-

$$H_0 : F = G \qquad H_1 : F > G$$

W = number of pairs with a positive difference ($x_i - y_i > 0$)

Under H_0, $W \sim Binom(n, 0.5)$

$$w = 9$$

This is because W has a discrete distribution!

P-value = $P(W \geq 9)$

$$= 1 - P(W \leq 8) = 0.304$$

Sign test normal approximation

$H_0 : F = G \qquad H_1 : F > G$ Under H_0, $W \sim Binom(n, 0.5)$

We can also use a normal approximation

$$E(W) = np = 7.5,$$
$$Var(W) = np(1-p) = 3.75$$

$$\frac{W - E(W)}{\sqrt{Var(W)}} \sim N(0,1)$$

$$\text{p-value} = P\left(Z \geq \frac{10 - 7.5}{\sqrt{3.75}}\right) = 1 - \Phi(1.29)$$

$$= 0.1$$

By only considering the sign of the difference, this test isn't using very much information from the data —> it will have low power!

Non-parametric paired two-sample test

The sign test

The wilcoxon signed rank test

Wilcoxon signed rank test

General procedure:

1. Remove observations with no difference ($D_i = 0$)
2. Compute R_i = rank of $|D_i|$
3. Calculate $W_i = R_i \times \text{sign}(D_i)$
4. Compute the test statistic: $$W_+ = \sum_{i:W_i>0} W_i$$

If H_0 is true, then

$$E(W_+) = \frac{n(n+1)}{4} \qquad Var(W_+) = \frac{n(n+1)(2n+1)}{24}$$

And we can use a **normal approximation** to compute a p-value

Wilcoxon signed rank test example

BP before program (x)	BP after program (y)	Difference (d)	Absolute difference |d|	**Rank of |d|**	**Signed rank**
125	118	7	7	10	**10**
132	134	-2	2	2.5	-2.5
138	130	8	8	12.5	**12.5**
120	124	-4	4	6	-6
125	105	20	20	15	**15**
127	130	-3	3	4	-4
136	130	6	6	8	**8**
139	132	7	7	10	**10**
131	123	8	8	12.5	**12.5**
132	128	4	4	6	**6**
135	126	9	9	14	**14**
136	140	-4	4	6	-6
128	135	-7	7	10	-10
127	126	1	1	1	**1**
130	132	-2	2	2.5	-2.5

X IID from dist F,
Y IID from dist G

$$H_0 : F = G$$

$$H_1 : F > G$$

$$W_+ = \sum_{i:W_i>0} W_i$$

$$w_+ = 10 + 12.5 + \ldots + 1$$

$$= 89$$

Wilcoxon signed rank test: example

X IID from dist F,
Y IID from dist G

$$H_0 : F = G \qquad H_1 : F > G$$

Some result exists that says that if H_0 is true

$$\frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}} \sim N(0,1)$$

where

$$E(W_+) = n(n+1)/4$$

$$Var(W_+) = n(n+1)(2n+1)/24$$

$$w_+ = 89 \qquad E(W_+) = 60$$

$$n = 15 \qquad Var(W_+) = 310$$

$$\frac{w_+ - E(W_+)}{\sqrt{Var(W_+)}} = \frac{89 - 60}{\sqrt{310}} = 1.647$$

p-value = $1 - \Phi\left(Z \geq 1.647\right) = 0.0498$

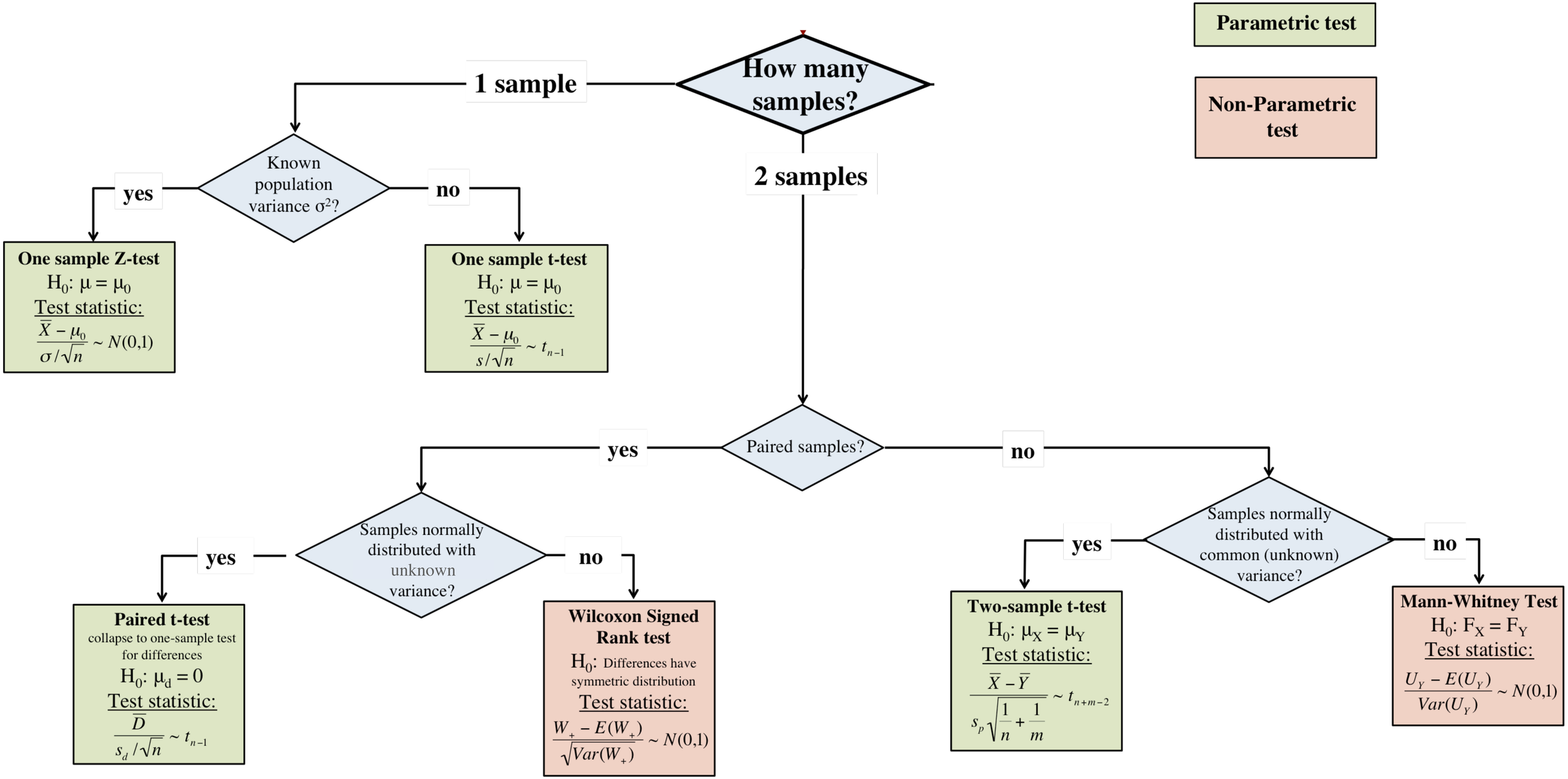
Which is borderline for rejecting H_0
but is technically less than 0.05

Wilcoxon signed rank test example

See paired_two_sample_test.R

A summary of the tests we've learned so far

Testing flow chart



Multiple testing and the reproducibility crisis

Multiple testing

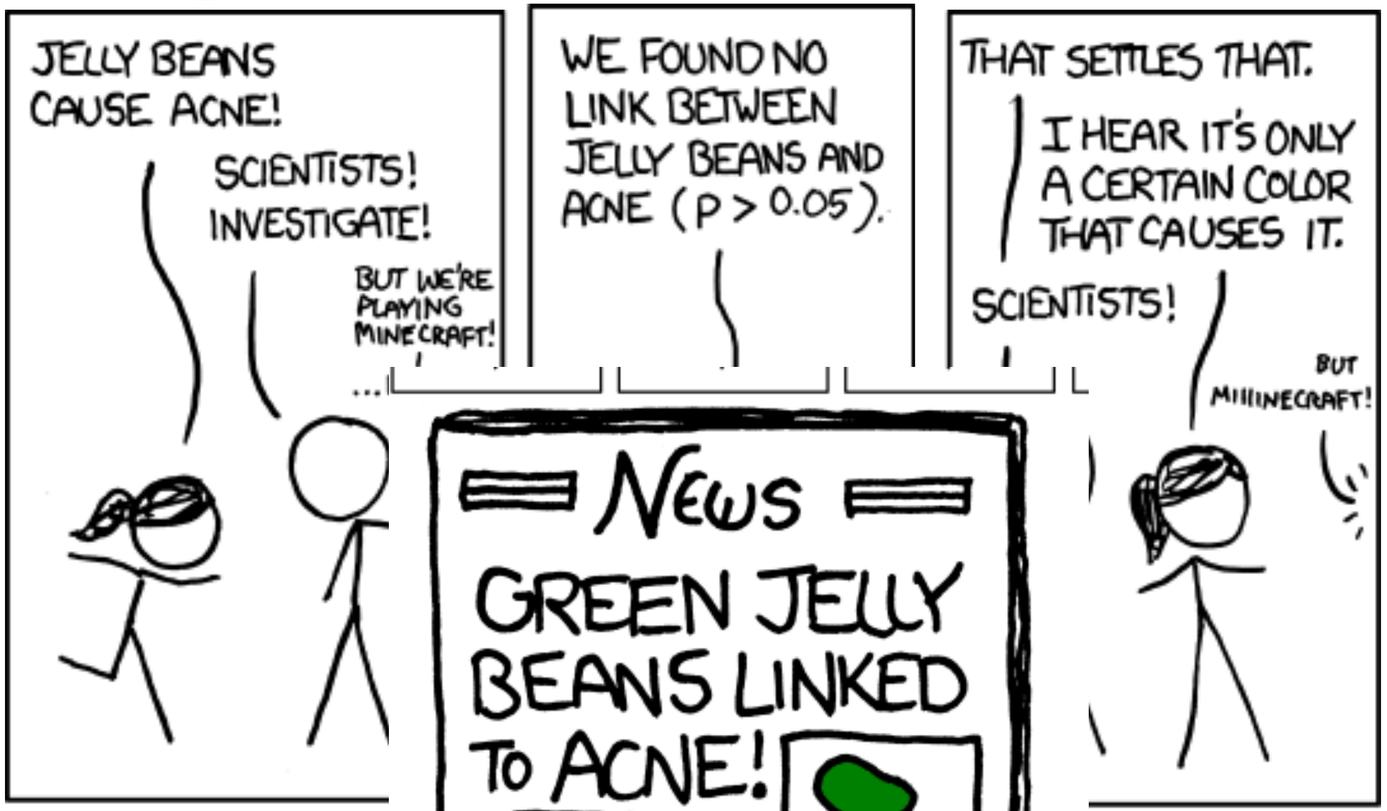
If you have set your **critical value to** $\alpha = 0.05$, and you conduct **100 hypothesis tests**.

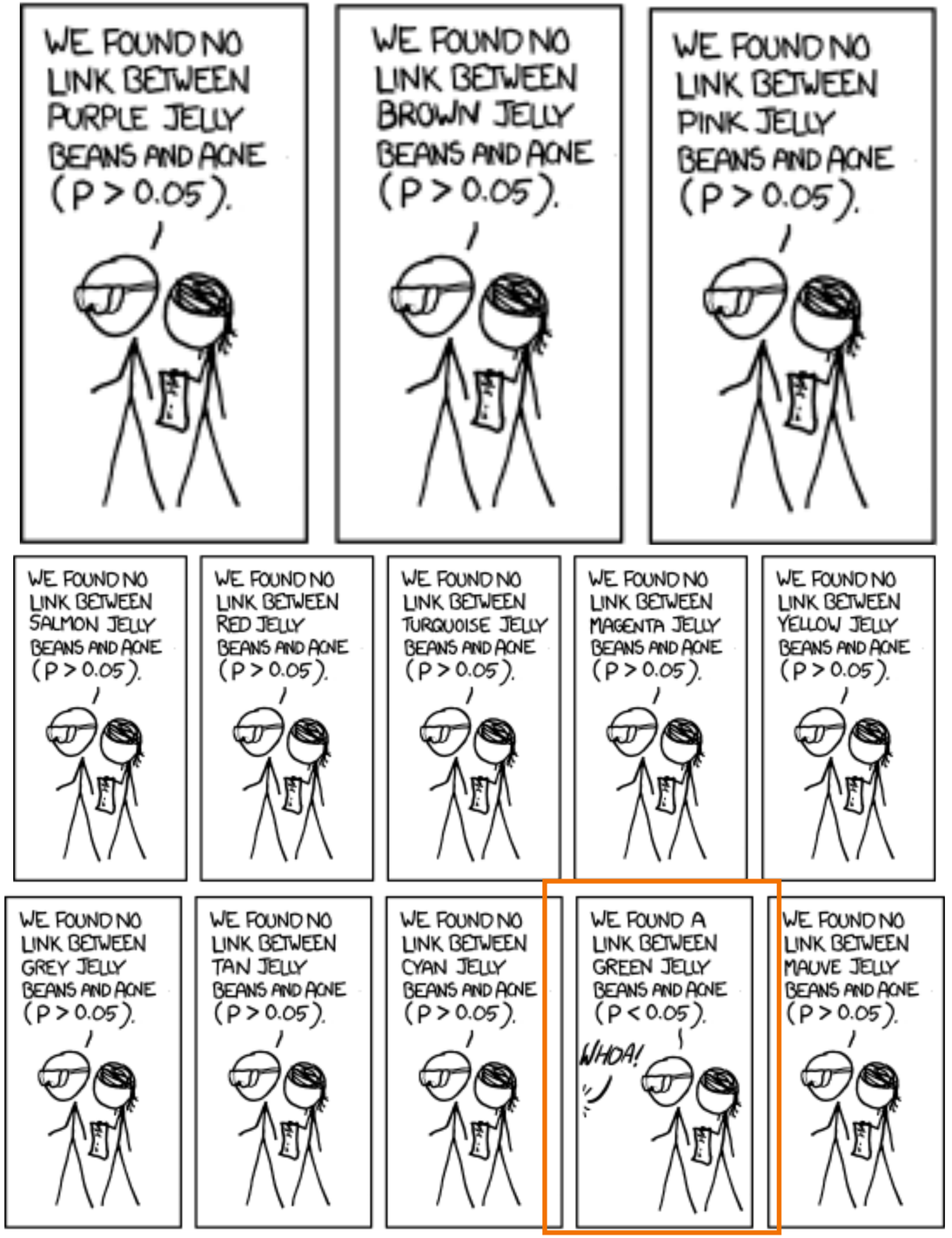
Assume the **null hypothesis is true** for all of the tests you conducted.

For how many of the tests do you expect to reject the null hypothesis?

5

Jellybeans cause acne!





Reproducibility crisis

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on

Research findings in a scientific field are less likely to be true,

1. the smaller the studies conducted.
2. the smaller the effect sizes.
3. the greater the number and the lesser the selection of tested relationships.

The Bonferroni correction for multiple testing

Multiple testing

Imagine that we conduct K hypothesis tests

Example: we want to see if a drug had any benefits at all.

So we tested to see whether patients taking the drug had an affect on:

1. Weight
2. Number of headaches
3. Blood pressure
4. Incidence of diabetes
5. …

Multiple testing

We are conducting K hypothesis tests, each at the $\alpha = 0.05$ level

$H_0 : \mu_X^{(1)} = \mu_Y^{(1)}$ Test for symptom 1

$H_0 : \mu_X^{(1)} = \mu_Y^{(1)}$ Test for symptom 2

$H_0 : \mu_X^{(3)} = \mu_Y^{(3)}$ Test for symptom 3

... ...

$H_0 : \mu_X^{(K)} = \mu_Y^{(K)}$ Test for symptom K

Assume the null is **true** in all cases.

The probability of rejecting any individual test is $5\,\%$

What is the probability of ***at least one incorrect rejection of the null?***

Multiple testing

We are conducting $K = 100$ hypothesis tests, each at the $\alpha = 0.05$ level

Assume the null is **true** in all cases.

What is the probability of ***at least one incorrect rejection of the null?***

$$P(\text{at least one } H_0 \text{ rejected} \,|\, H_0) = 1 - P(\text{no } H_0 \text{ rejected} \,|\, H_0)$$

$$= 1 - 0.95^{100}$$

$$= 0.994$$

I.e. there is a 99.4% chance that we will incorrectly reject at least one of our 100 tests!

The Bonferroni correction

The **family-wise error rate (FWER)** is the probability of at least one incorrect rejection among K tests

Our goal is to set the FWER to be 0.05. Currently it is **99.4%** if K = 100

Idea: maybe we can have a stricter (lower) significance level for each individual test

Bonferroni correction: If you have K hypothesis tests, reject H_0 if the p-value is less than α/K, rather than α

If $\alpha = 0.05$, and $K = 100$, this means you reject H_0 when you get a p-value less than $0.05/100 = 0.0005$