5. Comparing a continuous variable by group

Introduction to Medical Statistics

OUCRU, Ho Chi Minh City

March 23-27, 2026

Nguyen Lam Vuong

and the biostatistics crew

Learning Objectives

- Know statistical tests to compare a continuous variable by 2 independent groups:
  - Student's t-test vs. Wilcoxon rank-sum test
- Know statistical tests to compare a continuous variable by 2 dependent/paired groups
  - Paired t-test vs. Wilcoxon signed-rank test

Scenario

- Does a (numerical) population characteristic/variable differ by group
- Example: groups by
  - Sex
  - Type of treatment
  - TBM severity grade

This session: two groups

Split sample by group

– Student's t-test if we can assume that the sample mean per group has ≈ normal distribution if we would repeat the experiment, i.e. either
  1. population distribution per group not too skewed
  2. sample size per group is large

– Wilcoxon rank-sum test if sample small and distribution of variable looks too skewed

Relation between groups?

- Independent: groups composed of different individuals
- Dependent: two measurements from same individual

“Student” (1908)

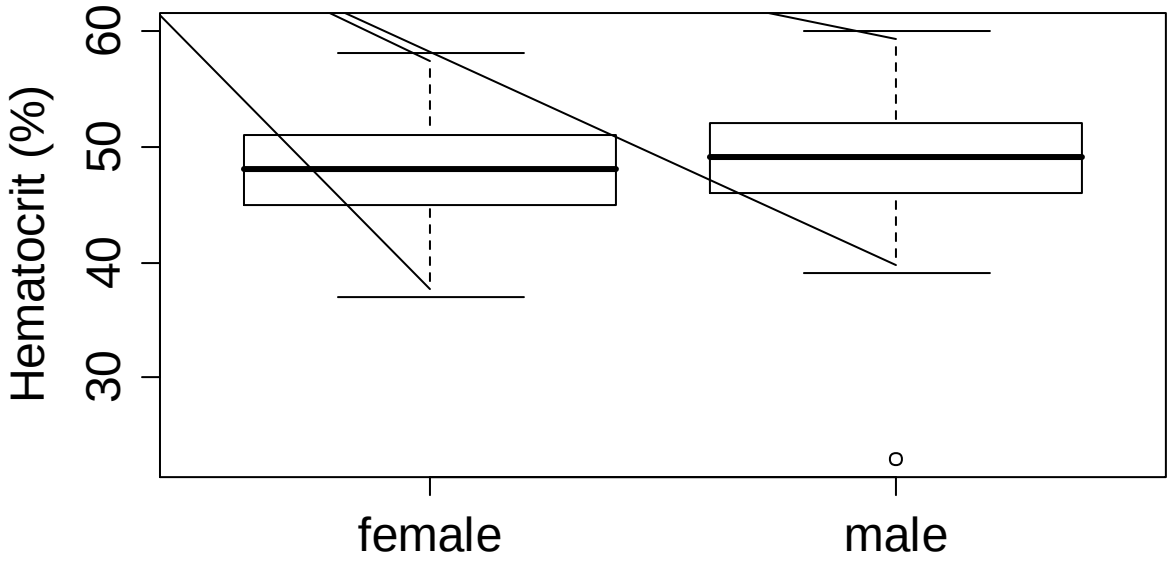
William Sealy Gosset, who developed the "t-statistic" and published it under the pseudonym of "Student"

Comparisons of two independent groups

Example

→ Population mean of hematocrit at presentation in children with dengue shock syndrome. Does it differ between boys and girls?

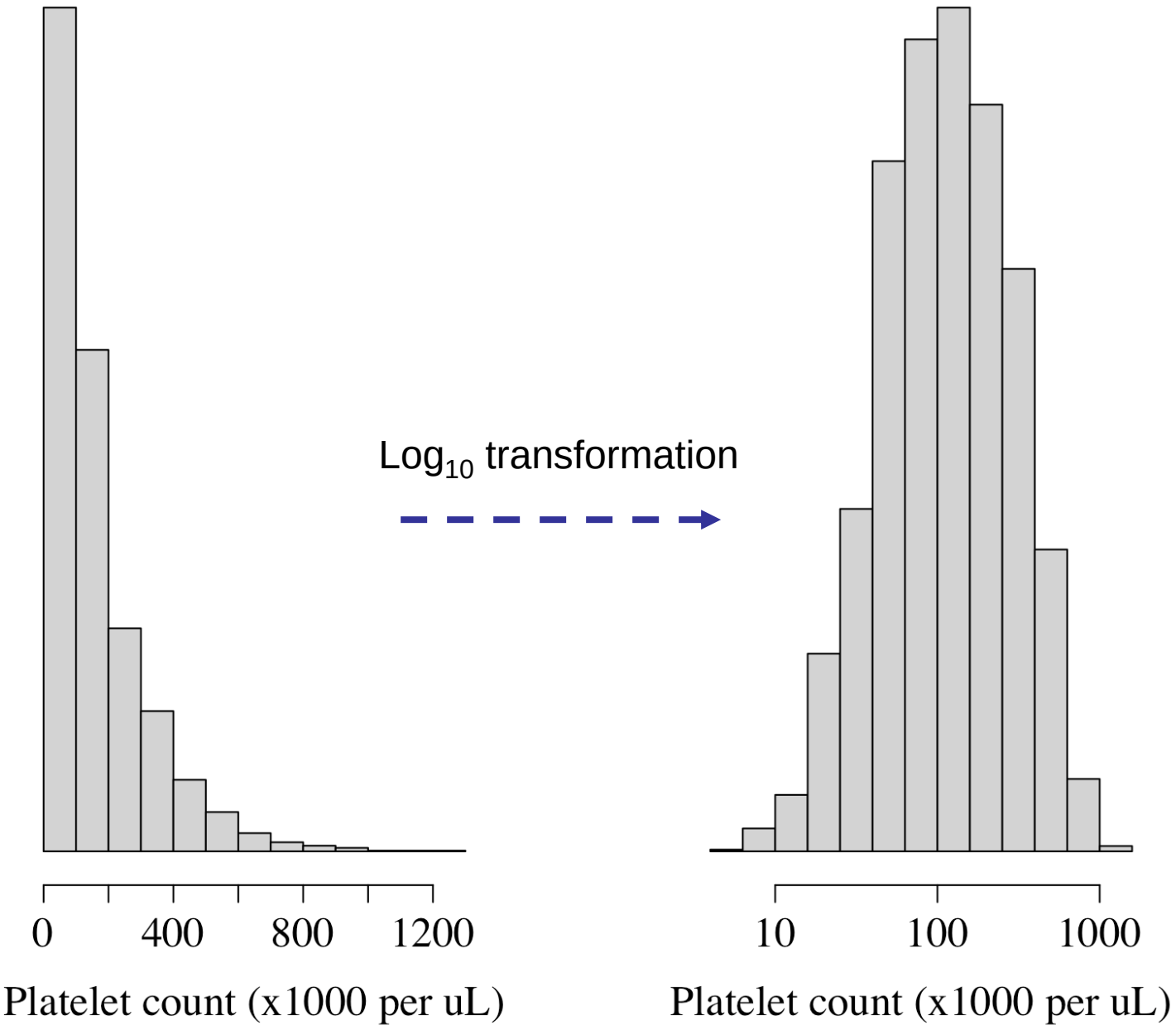
- Data: 510 hospitalized Vietnamese children



Comparison of numeric variable

- Can we assume that the distribution of sample mean per group is approximately normal?
  - Yes → t-test (variable not too skewed or large sample size)
  - No → Wilcoxon rank-sum test
- Decision based on
  - Prior knowledge/experience
  - Graphical examination of the data per group
    - Histogram, boxplot, normal QQ plots
- T-test is easier to interpret. Try to transform variable to more symmetric form if distribution too skewed

Platelet counts in 2600 severe malaria patients



Variable has ≈ normal distribution

- Setting:
  - Two groups
  - Subjects from groups sampled independently

  → Do the groups have a different population mean ("true" mean) of the variable?
- Procedure
  - Calculate difference in means in sample – estimate of population difference
  - Calculate confidence interval (CI) of the difference
  - Test the null hypothesis that the means in the two subpopulations are equal

Ingredients for a p-value

1. Null hypothesis
2. A test statistic
3. The distribution of the test statistic when the null hypothesis is true

Population and data summaries

- Population: means and standard deviations
  - Group 1: μ_1, σ_1
  - Group 2: μ_2, σ_2

- Sample: size, observed sample means and s.d.
  - Group 1: $n_1, \bar{x}_1, s_1$
  - Group 2: $n_2, \bar{x}_2, s_2$

Approximate comparison: two sample z-test

- Null hypothesis H_0: $\mu_1 = \mu_2$ (i.e. $\mu_1 - \mu_2 = 0$)
- Alternative hypothesis H_A: $\mu_1 \neq \mu_2$
- If the null-hypothesis is true, the difference in means has approximately a normal distribution with mean 0 and S.E. as on next slide
- Compute p-value based on standardized statistic
- Compute (95%) confidence interval

Approximate confidence interval for the difference in population means

- Sampling distribution of the difference between two means is normally distributed with
  - mean $\mu_1 - \mu_2$
  - standard deviation (standard error):

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}$$

- An approximate 95% CI is calculated as

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96 \cdot se(\bar{x}_1 - \bar{x}_2)$$

where the s_1 and s_2 are used instead of σ_1 and σ_2 for the calculation of the standard error:

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{s_1^2 / n_1 + s_2^2 / n_2}$$

Approximate CI and z-test, example

- Data for hematocrit in children with dengue shock

$$\text{Females}: \quad n_1 = 256, \bar{x}_1 = 48.17, s_1 = 3.75$$

$$\text{Males}: \quad n_2 = 254, \bar{x}_2 = 48.59, s_2 = 4.46$$

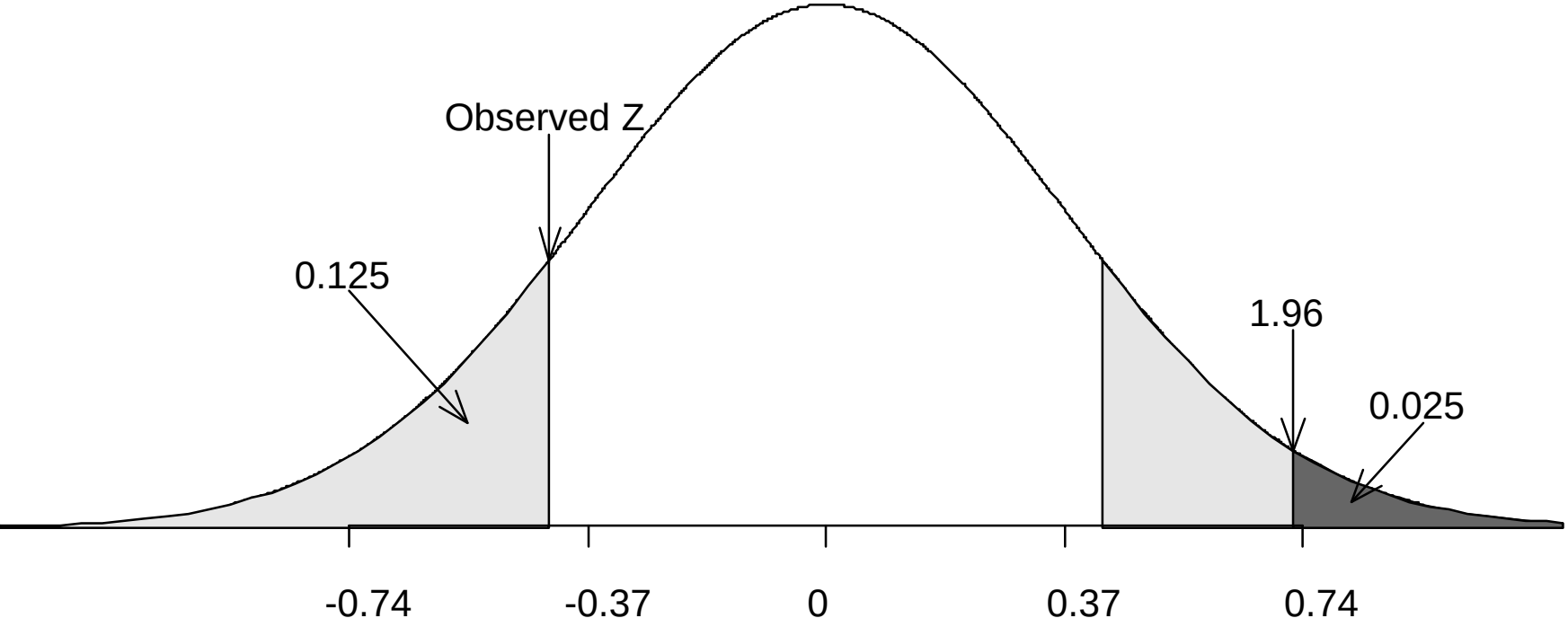
$$\rightarrow \quad \bar{x}_1 - \bar{x}_2 = -0.42, \quad se(\bar{x}_1 - \bar{x}_2) = 0.37$$

- 95% confidence interval for mean difference:
  -0.42 ± 1.96 * 0.37 = (-1.13, 0.29)

- p-value: P(diff < -0.42)+P(diff > 0.42)=2·0.125=0.25
  "Our data do not suggest that hematocrit at presentation differs between boys and girls (p=0.25)."

Graphical illustration of p-value for the example

Approximate distribution of difference under the null hypothesis



Student's t-test and confidence interval

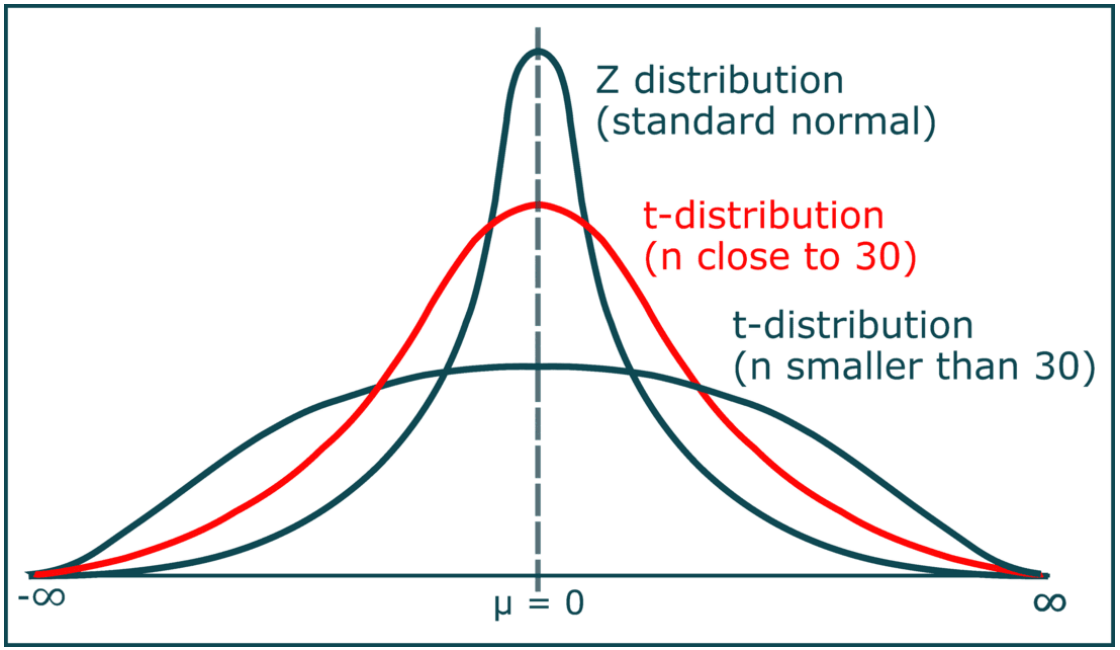
- Take into account that the standard deviation is estimated from the data and not known
  - Minor modifications to the test statistic
  - Based on the t-distribution instead of the normal distribution
- This method (t-test) is commonly used, and given in the statistical packages
- Approximation assuming a normal distribution gives very similar results if sample size in both groups >20

t distribution vs. normal distribution

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$



R: t-test function

Note: `t.test` makes use of the "formula" notation `hct1~sex`

(Left hand side: variable; right hand side: group)

```
> t.test(hct1 ~ sex, data = dengue)

        Welch Two Sample t-test

data:  hct1 by sex
t = -1.1571, df = 492.105, p-value = 0.2478
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1401995  0.2949731
sample estimates:
mean in group female   mean in group male
            48.17188             48.59449
```

Relation between confidence intervals and p-values

- Note for the example:
  - 95% confidence interval includes 0 (i.e. it is plausible given the data that there is no difference between the groups)
  - p-value>0.05
- This is generally true:
  - 95% confidence interval includes 0 ↔ p-value>0.05
  - 95% confidence interval does not include 0 ↔ p-value≤0.05
  - Also holds for other %CI and corresponding p-value

Alternative: Wilcoxon rank-sum test

- If we don't trust the assumption that the difference has a normal distribution
- Null hypothesis H_0
  - Variable has the same distribution in both groups

  (Not necessarily a normal distribution)
- Alternative hypothesis H_A

  Distribution of variable differs between both groups
- Test depends on the ranks (ordering) of the observations, not on the actual measurements
- Also called Mann-Whitney U test

Alternative: Wilcoxon rank-sum test

1. All observations are combined and ranked
   - Smallest observation (in combined sample) gets rank 1, second smallest gets rank 2 etc.

For example:

- Group 1 ($n_1 = 6$): 77 78 70 72 65 74
- Group 2 ($n_2 = 7$): 60 62 70 76 68 72 70

60	62	65	68	70	70	70	72	72	74	76	77	78
1	2	3	4	5	6	7	8	9	10	11	12	13
1	**2**	**3**	**4**	**6**	**6**	**6**	**8.5**	**8.5**	**10**	**11**	**12**	**13**

Alternative: Wilcoxon rank-sum test

2. Test statistic is sum of ranks in the first group
   - Test is based on the idea that if the two groups do not differ, i.e. if the null hypothesis is true, the rank sum in the two groups should be approximately the same (if $n_1=n_2$)

For example:

Sum of ranks: $R_1 = 3 + 6 + 8.5 + 10 + 12 + 13 = \mathbf{52.5}$

$R_2 = 1 + 2 + 4 + 6 + 6 + 8.5 + 11 = \mathbf{38.5}$

Test statistics: $U_1 = 52.5 - 6 \times (6 + 1)/2 = \mathbf{31.5}$

$U_2 = 38.5 - 7 \times (7 + 1)/2 = \mathbf{10.5}$

The Mann-Whitney U test statistic is selected as the smallest of the two calculated U values

Alternative: Wilcoxon rank-sum test

3. Calculate p-value, i.e. probability of getting a rank sum as extreme as the observed one or more extreme by chance

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

U = 10.5 > 6

→ p > 0.05

Wilcoxon rank-sum test for the example

```
v1 <- c(rep(1,6), rep(2,7))
v2 <- c(77,78,70,72,65,74,60,62,70,76,68,72,70)
wilcox.test(v2 ~ v1)

   Wilcoxon rank sum test with continuity correction

data:  v2 by v1
W = 31.5, p-value = 0.1503
alternative hypothesis: true location shift is not equal
   to 0
```

t-test vs. Wilcoxon test

- t-test
  - Optimal if variables normally distributed
  - Easy to get confidence intervals for the population difference in mean
  - Approximately valid for non-normal distribution of variable
    - Do not use for strongly skewed variable (try to transform!) or in presence of large outliers, especially if group size is fairly small

t-test vs. Wilcoxon test

- Wilcoxon test
  - No distributional assumptions, robust against outliers
  - Use if group size is small and distribution of variable is far from normal and cannot be transformed to more normal one
  - Also for ordered categorical data with many categories
  - Not a test to compare medians
  - Confidence interval can be calculated, but difficult to communicate (does not quantify difference in medians)
- Simple rule
  - If no confidence intervals are desired, the Wilcoxon test is safer
  - Always valid, but gives some loss in power if variable has ≈ normal distribution in each group

Comparison of two dependent/paired groups – numeric variables

Paired data

- Previous section
  - Measurements come from two independent groups (separate individuals, e.g. males and females)
- In some circumstances, data consists of pairs of outcome measurements
  - The same individuals are studied twice, usually in different circumstances
  - Two different groups of subjects are studied where each person in one group is individually matched with a member of the other group (e.g. by gender and age)

  →Statistical tests need to be adapted to the dependence within pairs

Comparison of numeric paired data

- Interest in average difference between values in each pair and variability of these differences
- Two paired sample problem is reduced to one sample problem: difference for each pair
- **Example:** Mean dietary intake (kJ) over 10 pre-menstrual and 10 post-menstrual days in 11 women

Subject	Pre-menstral	Post-menstrual	Difference
1	5260	3910	1350
2	5470	4220	1250
3	5640	3885	1755
4	6180	5160	1020
5	6390	5645	745
6	6515	4680	1835
7	6805	5265	1540
8	7515	5975	1540
9	7515	6790	725
10	8230	6900	1330
11	8770	7335	1435

Paired t-test

1. H_0: mean population difference = 0
   H_A: mean population difference ≠ 0
2. Assume difference has a normal distribution
3. Calculate individual differences d for each pair
4. Calculate mean and standard deviation of the difference: $\bar{d},\ s_d \rightarrow se(\bar{d}) = s_d / \sqrt{n}$
5. Perform the one-sample t-test on the difference
6. Compute confidence interval and p-value

Paired-sample t-test for the example

```
> t.test(pre, post, paired = TRUE)

        Paired t-test

data:  pre and post
t = 11.9414, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1074.072 1566.838
sample estimates:
mean of the differences
                1320.455
```

→ “Pre-menopausal dietary intake higher than post-menopausal intake ($p<0.0001$).”

Paired non-normal data: sign test

- H_0: median population difference = 0
- Test statistic
  - k: the number of differences > 0
  - If the null hypothesis is true
    - k will not be “too different” from n/2
    - More precisely: k has a binomial distribution B(n, p=1/2)

  → Reduced to hypothesis test for single proportion

Paired non-normal data: sign test

- Dietary example:

```
pre <- c(5260, 5470, 5640, 6180, 6390, 6515, 6805, 7515, 7515, 8230, 8770)
post <- c(3910, 4220, 3885, 5160, 5645, 4680, 5265, 5975, 6790, 6900, 7335)
data <- pre - post
BSDA::SIGN.test(data, md = 0)

        One-sample Sign-Test

data:  data
s = 11, p-value = 0.0009766
alternative hypothesis: true median is not equal to 0
95 percent confidence interval:
  941.000 1601.764
sample estimates:
median of x
      1350
```

Paired non-normal data: Wilcoxon signed-rank test

1. Compute paired differences, rank by absolute value
2. Calculate sum of ranks of all differences > 0
   - If the true median population difference is 0 and the distribution is symmetric, this should be “not too different” from the sum of ranks of all differences < 0

```
> wilcox.test(pre, post, paired = TRUE)

  Wilcoxon signed rank test with continuity correction

data:  pre and post

p-value = 0.00384
```

t-test vs. sign test vs. Wilcoxon signed-rank test for comparing paired groups

- Advantages/disadvantages of t-test compared to nonparametric tests (sign test and Wilcoxon signed-rank test) similar to the independent group case
  - Don’t do t-test with highly non-normally distributed differences, especially if sample size is small
- Wilcoxon signed-rank test is based on the assumption that the distribution of the differences is symmetric
  - Result of this test can be affected by transforming the data
  - Differences symmetrically distributed → Wilcoxon is preferable
  - Differences not symmetrically distributed → use sign test

Interpretation of p-values and confidence intervals – An example

Example

- 3 new drugs (A, B, C) to lower cholesterol in middle-aged persons at high risk of heart attack
  - Drugs A and B are cheap
  - Drug C is expensive
- 5 randomized trials of the 3 drugs vs. control group (placebo)
- Primary outcome measure
  - Cholesterol at one year
  - Clinical interpretation of mean reduction in cholesterol (vs. placebo)
    - 40 mg/dl or more → substantial protection against subsequent heart disease
    - 20-40 mg/dl → moderate protection

Results of the trials - Interpretation?

Trial	Drug	Cost	No. of pt. per group	Mean cholesterol at one year (mg/dl)		Reduction due to drug		
				Drug	Placebo	Esti-mate	95% CI	p-value
1	A	Cheap	30	140	180	-40	(-118,+38)	0.32
2	A	Cheap	3000	140	180	-40	(-48,-32)	<0.001
3	B	Cheap	40	160	180	-20	(-85,+45)	0.54
4	B	Cheap	4000	178	180	-2	(-8.5,+4.5)	0.54
5	C	Expensive	5000	175	180	-5	(-8.9,-1.1)	0.01

Key issues

- Large p-value does not mean that we have shown that the null hypothesis is true
  - Small study → true effects may not reach statistical significance
- Evidence for difference (small p-value) is not same as clinical relevance
  - Huge study → tiny population effects still have low p-value
- Confidence interval often more informative than p-value
  - Shows range of values that are plausible given the data
  - Narrow confidence interval that covers zero ("no effect") allows to rule out large effects
  - → General advice: report an estimate of the effect with both a 95% CI and a p-value

Statistical hypothesis testing

The p-value is the probability that the null hypothesis is wrong.

Statistical hypothesis testing

The p-value is the probability that the null hypothesis is wrong.

NO!

The p-value is the probability of seeing current data or data that is more extreme (based on the test statistic) assuming the null hypothesis is true.

It is a measure of **surprise** relative to the null hypothesis

Statistical hypothesis testing

Small p-values indicate large effects.

Statistical hypothesis testing

Small p-values indicate large effects.

WRONG!

p-values tell you next to nothing about the size of an effect.

Summary today

- Characteristic (variable) in population
  - Distribution ≈ normal or skewed?
  - Distribution differs between groups? Mostly characterized as difference in mean
- Inference based on sample:
  - Sample mean estimates population mean
  - Distribution of sample mean approximately normal if
    - distribution of variable ≈ normal or
    - sample size large

    → **Use t-test**
  - Otherwise, we can use another test (**Wilcoxon**) to compare groups (but CI are difficult to obtain and interpret)