Review

- **A test of significance** (calculating a **P-value**) is for comparing observed data with the null hypothesis (H_0 : $\mu_1 - \mu_2 = 0$) and is designed to quantify the strength of the evidence against the null hypothesis
- When **p-value ≤ α**, we say that the data are statistically significant at level α (i.e., we have significant evidence against the null hypothesis)
- **α** is the Type I error (chance of falsely rejecting the null hypothesis)
- Confidence Interval is a range of values that is likely to contain the value of an unknown population parameter

Two-sided Test and Confidence Intervals

A level α **two-sided** significance test rejects H_0: $\mu=\mu_0$ exactly when μ_0 falls outside a level $1-\alpha$ confidence interval for μ.

95% CI includes μ_0 $\Longleftrightarrow$ do not reject H_0 at $\alpha=0.05$

95% CI doesn't include μ_0 $\Longleftrightarrow$ reject H_0 at $\alpha=0.05$

Note: This is only true under a two-sided test

Two-Sample Inference

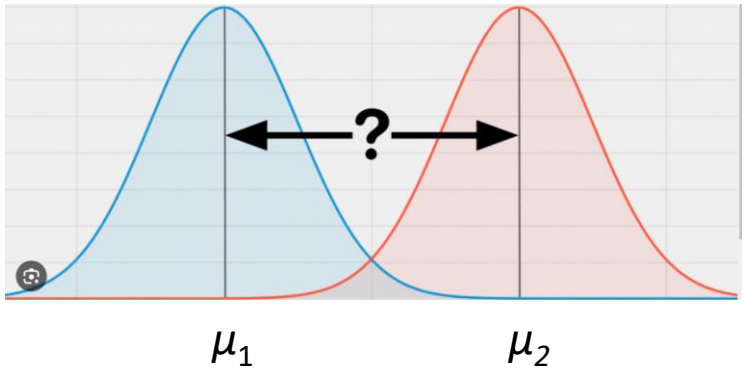
Independent Samples

- Two samples are said to be independent when the data points in one sample are unrelated to the data points in the second sample.

Example:

- Blood pressure in smokers vs non-smokers
- Pain Scores in Females vs Males
- Survey response rate in low- and high-income subjects

T-test for Independent Samples



Calcium Example

21 subjects were randomly assigned to two groups: 10 of them received a calcium supplement for 12 weeks, while the control group of 11 men received a placebo pill that looked identical. The response variable is the decrease in systolic (top number) blood pressure for a subject after 12 weeks, in millimeters of mercury. The **goal** is to find out whether the calcium has effect on the systolic blood pressure.

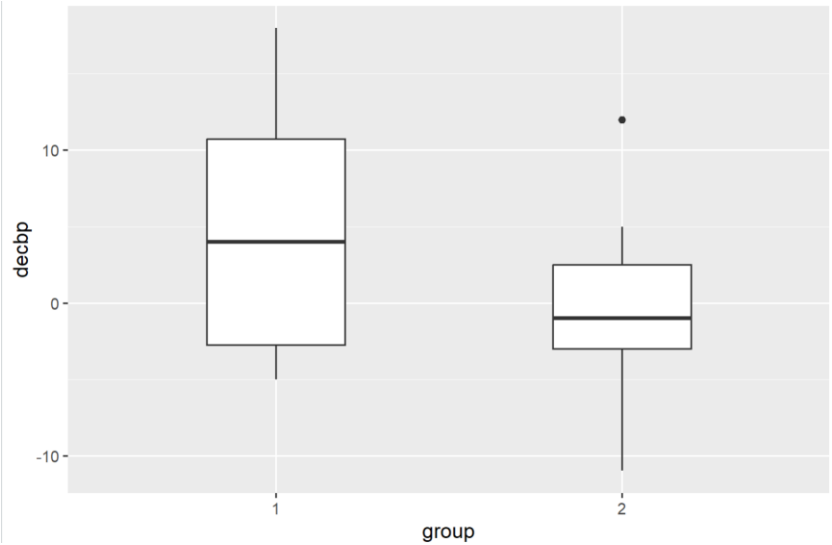
Data:

Group 1 (calcium):	7	−4	18	17	−3	−5	1	10	11	−2	
Group 2 (placebo):	−1	12	−1	−3	3	−5	5	2	−11	−1	−3

Make Visual Comparison

```
ggplot(data=cal, aes(x=group, y=decbp)) +
geom_boxplot(width=0.4)
```

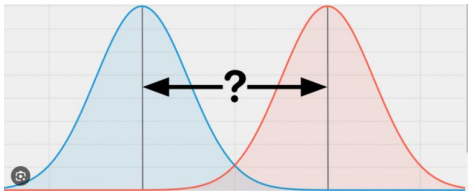
```
tapply(cal$decbp, cal$group, summary)
```




```
$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -5.00   -2.75    4.00    5.00   10.75   18.00

$`2`
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-11.0000  -3.0000  -1.0000  -0.2727   2.5000  12.0000
```

Use sample to make inference on population



Make Inference Using R

```
R: t.test(decbp[group==1], decbp[group==2],
          alternative = "two.sided", var.equal = TRUE)
```

Assume equal variance for two groups

```
        Two Sample t-test

data:  decbp[group == 1] and decbp[group == 2]
t = 1.6341, df = 19, p-value = 0.1187
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.48077 12.02622
sample estimates:
 mean of x  mean of y
 5.0000000 -0.2727273
```

Conclusion: There is no statistically significant difference between the group with and without calcium in decreasing the systolic blood pressure (the mean difference is 5.27; 95% CI is from -1.48 to 12.03; p=0.12)

Based on this, how to design a new study to reach statistical significance?

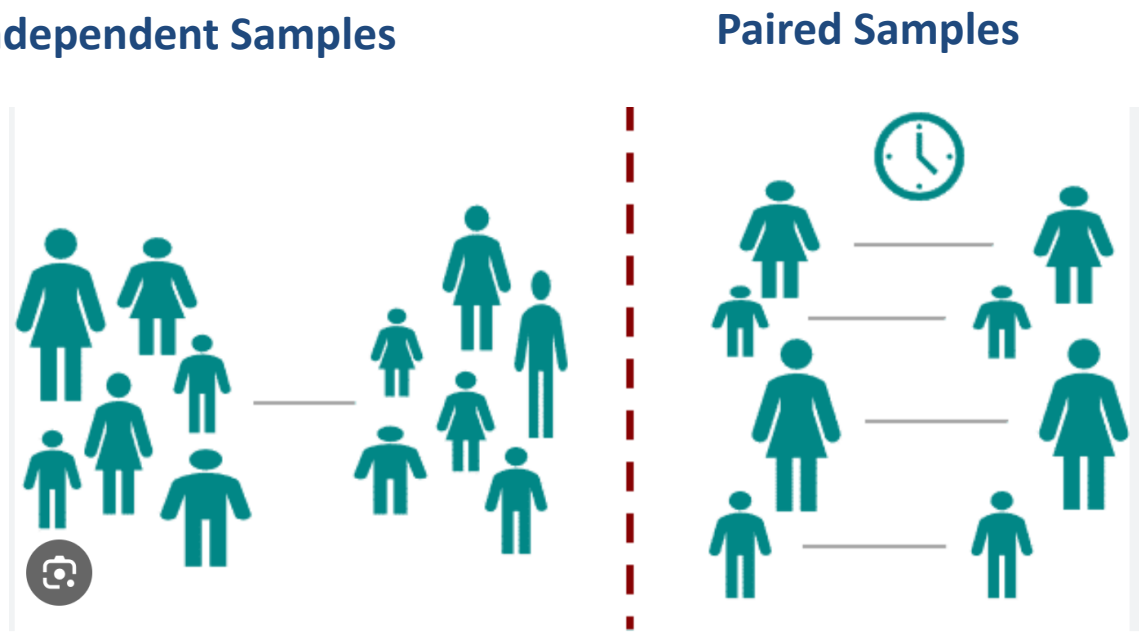
Paired Samples

- Two samples are said to be paired when each data point is matched and is related to a unique data point of the second sample.

Example:

- Measurements taken before and after treatment from the same subject. These measurements tend to be closer than across subjects
  - If baseline BP is high, chances are BP after the intervention will be high too
- Measurements taken from identical twins
- Matched subjects (matched by age…)

Independent vs. Paired Samples



Obtained from https://datatab.net/tutorial/paired-t-test

Aspirin Example

In a pediatric clinic a study is carried out to see how effective aspirin is in reducing temperature.

Twelve 5-year-old children suffering from influenza had their temperatures taken before and 1 hour after administration of aspirin.

Would like to test the hypothesis that aspirin significantly changes the temperature

Question: paired samples or independent samples?

Paired Samples

Goal: To test the difference (Δ) between two populations

Typical Data Structure

Subject ID	Before	After
1	102.4	99.6
2	103.2	100.1
…	…	…
n	101.4	100.2

Hypotheses: H_0: $\Delta = 0$ vs $\Delta \neq 0$

How to construct a test considering the paired nature of the data?

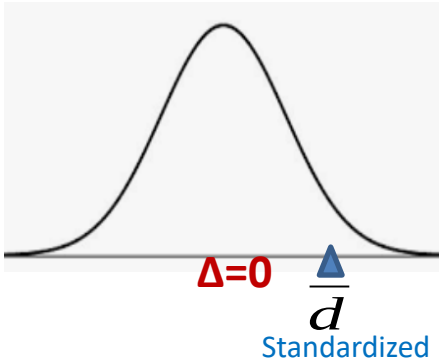
Testing Effect of Aspirin in a Paired Samples

Table. Body temperature (^{0}F) before and after taking aspirin

Patient	Before	After	d = After - Before
1	102.4	99.6	99.6 - 102.4 = -2.8
2	103.2	100.1	100.1 - 103.2 = -3.1
…	…	…	…
12	101.4	100.2	100.2 - 101.4 = -1.2

Reduced to one-sample, we can simply apply the one-sample test to the differences

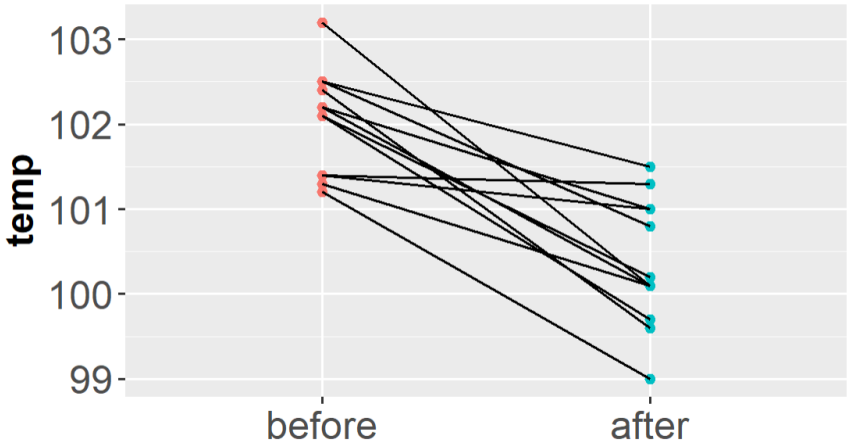
We use sample mean $\overline{d}$ to estimate population mean **Δ**



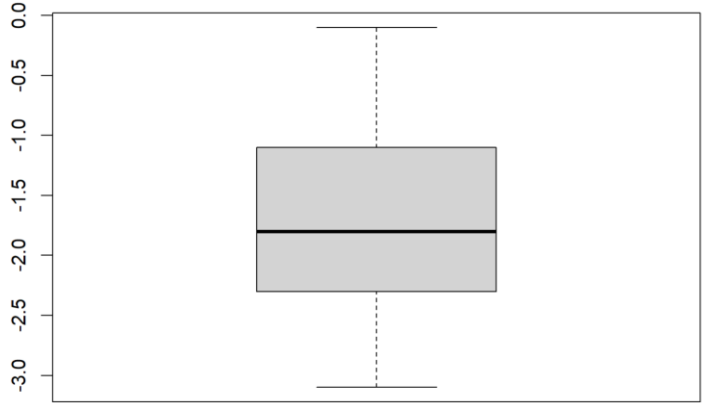
What graph(s) should be used to show the difference in temperature before and after taking aspirin?

Summarize the Change Data

Spaghetti plot to visualize the change



Boxplot to summarize the change




```
> summary(data$change)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -3.100  -2.250  -1.800  -1.675  -1.150  -0.100
```

Paired t-test

```
t.test(data$after, data$before, paired=TRUE)
Or,
t.test(data$change)
```

```
        One Sample t-test

data:  data$change
t = -6.2961, df = 11, p-value = 5.869e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.260547 -1.089453
sample estimates:
mean of x
   -1.675
```

p<0.05

95% CI doesn't include 0

Conclusion: Temperature is significantly reduced after taking the aspirin (mean difference is -1.68; 95% CI: -2.26 to -1.09; p <0.001)

Assumptions of t-tests

- Assumptions:
  - ✓ **Normality**
- t-test is robust to departures from a normal distribution
  - ✓ n ≥ 15 : t -test can be used except in the presence of outliers or strong skewness
  - ✓ n ≥ 40 : t -test can be used even for clearly skewed distributions

  log transformation helps right-skewed data
- Small samples with Non-Normality
  - **Non-parametric tests (free of distributions)**

Nonparametric Tests

Nonparametric Tests

Setting	Compare means Parametric Test	Compare medians Nonparametric Test
Matched pairs	Paired t-test (one sample t-test on the differences)	Wilcoxon signed rank test
Two independent samples	Two-sample t-test	Wilcoxon Rank Sum test
Several independent samples	One-way ANOVA	Kruskal-Wallis test

Will talk later

Two-sample Wilcoxon Test

- **Motivating Example**: The following table gives biceps skinfold measurements for **6** patients with Crohn's disease and **5** patients with Coeliac disease.
- The *objective* is to assess whether the distribution of the bicep measurements are the same in the two populations.

Crohn's disease	Coeliac disease
1.8, 2.8, 4.2, 5.2, 2.2, 3.2	1.8, 2.0, 2.2, 2.0, 3.0

Null hypothesis: The distributions in the two populations are the same (same medians)

Alternative hypothesis: The distributions in the two populations are NOT the same (different medians)

Wilcoxon Rank Sum test

- Analogue of two-sample (independent) t-test is the Wilcoxon Rank Sum test

- Also called the **Mann-Whitney test**
  - Sample of size *n* from population 1
  - Sample of size *m* from population 2

Wilcoxon Rank Sum test

Step 1: Rank (n+m) subjects regardless of the populations
Step 2: Sum the ranks of subjects in sample 1 and call it ***W*** (test statistic)
Step 3: Reject null hypothesis when ***W*** is "far" from its expected value under the null

Data
1.8
1.8
2
2
2.2
2.2
2.8
3
3.2
4.2
5.2

Data	Rank
1.8	1.5
1.8	1.5
2	3.5
2	3.5
2.2	5.5
2.2	5.5
2.8	7
3	8
3.2	9
4.2	10
5.2	11

Data	Rank	Group
1.8	1.5	Coeliac
1.8	1.5	Crohn
2	3.5	Coeliac
2	3.5	Coeliac
2.2	5.5	Coeliac
2.2	5.5	Crohn
2.8	7	Crohn
3	8	Coeliac
3.2	9	Crohn
4.2	10	Crohn
5.2	11	Crohn

Wilcoxon Rank Sum Test – Calcium Example

R: wilcox.test (decbp[group==1], decbp[group==2])

```
        Wilcoxon rank sum test with continuity correction

data:  decbp[group == 1] and decbp[group == 2]
W = 69.5, p-value = 0.3228
alternative hypothesis: true location shift is not equal to 0
```

Conclusion: There is no statistically significant difference between the group with and without calcium in decreasing the systolic blood pressure (median reduction is 4 in calcium vs -1 in control; p=0.32)

Wilcoxon Signed Rank Test- Aspirin Example

R: wilcox.test(data$after, data$before, **paired**=TRUE)

```
        Wilcoxon signed rank test with continuity correction

data:  data$after and data$before
V = 0, p-value = 0.002516
alternative hypothesis: true location shift is not equal to 0
```

Conclusion: Temperature is significantly reduced after taking the aspirin (median reduction is -1.8; p =0.0025)

Rank Tests vs t Tests

- Rank methods focus on significance tests, not confidence intervals.
- Inference based on ranks is restricted to simple settings, whereas *t* procedures extend to more complicated situations, such as experimental design and multiple regression.
- *t* test is robust, consider it when n>25 per group.
- When data is ordinal (e.g., low, median, high; "=0", "<2" or "≥ 2"), rank methods are preferred.

Note: For small samples and non-normal data, Wilcoxon test is more powerful than two-sample t-test. If data truly are normally distributed, t-test is more powerful.

Exercise

-Which test(s) should be used?

Study I: Mineral

The total mineral content of 8 people were measured by each of 2 technicians. We want to know if the results from the two technicians are significantly different.

Subject	Tech_1	Tech_2
1	1.328	1.323
2	1.342	1.322
3	1.075	1.073
4	1.228	1.233
5	0.939	0.934
6	1.004	1.019
7	1.178	1.184
8	1.286	1.304

Study II: LOS

- Compare the length of hospital stay (LOS) between two groups (Control vs Experimental)

```
> head(dd)
      los         group
1 150.100       Control
2 195.350 Experimental
3  66.367       Control
4   9.500 Experimental
5  59.650 Experimental
6  96.350       Control
```

Study III– Ear, Nose & Throat

Study Cohort (Rhinoplasty=9, Non-Rhinoplasty=15): patients who had previously received nasal surgery (non-rhinoplasty or rhinoplasty) but are still experiencing moderate-to-extreme nasal obstruction.

Study Goal:
(1) Test whether radiofrequency (RF) treatment to the internal nasal valves is effective, by comparing the Nasal Obstruction Symptom Evaluation (NOSE) Score before and after the RF treatment
(2) Compare the effect RF treatment of between two types of previous surgery, non-rhinoplasty or rhinoplasty

Outcome: NOSE Score is a numeric score ranges from 0 - 100

Study IV- OBGYN

Study Cohort: 30 patients who underwent surgery for vaginal prolapse.

Study Goal:
Test whether patients experienced improvement in symptoms by comparing the outcomes from validated questionnaires given before and after treatment.

Outcomes:
(1)Answer for a question 'How much does experiencing frequent urination bother you?' which are on a Likert scale (1, 2, 3, 4).
(2)Summary scores such as the MESA urinary incontinence score which is a numeric score ranging from 0 to 27.

Illustration of advantage of a paired t test