

HYPOTHESIS TESTING

Two-Sample Inference

In a two-sample hypothesis test, we compare the underlying parameters of two different populations, neither of which values is assumed known.

This is one of the most frequently encountered situations in biological and medical research.

The Paired Sample Design

A **paired sample design** is typically used in a **longitudinal** or **follow-up study**.

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

Paired samples may represent two sets of measurements on the same individual. Alternatively, paired samples may also represent measurements on different individuals, chosen or matched on a basis so that using each member of the pair is very similar to the other.

For paired samples a **paired-t test** is used. Assume that μ_d = the mean difference within subject between follow-up and baseline and is constant for all individuals (pairs).

$H_0: \mu_d = 0$ vs. $H_1: \mu_d \neq 0$

Assume $d_i \sim N(\mu_d, \sigma_d^2)$.

Calculate $\bar{d} = (d_1 + d_2 + \dots + d_n)/n$, where n = the number of pairs and $2n$ = the number of individuals.

Calculate the test statistic:

$t_s = (\bar{d} - 0)/(s_d / \sqrt{n}) = \bar{d} / (s_d / \sqrt{n})$, where $s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^n [d_i - \bar{d}]^2}$ where n = number of matched pairs.

If $t_s > t_{1-\alpha/2, n-1}$ or $t_s < -t_{1-\alpha/2, n-1}$ Reject H_0 .

If t_s is between these two values, then do not reject H_0 .

Exact p-values are calculated by multiplying by 2 the area to the left (or right) of t_s under a $t_{[n-1]}$ distribution.

Confidence intervals for μ_d are obtained by using the formulae for the one-sample situation.

Sample size and power are obtained by using the formulae for the one-sample hypothesis test situation.

The Independent Sample Design

An **independent sample design** is used in the **cross-sectional design**.

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

If the assumption that the variances from the two samples are equal, then using Student's t-test or the two sample t-test for equal variances is valid.

Student's t-test

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

Alternatively, $H_0: \mu_1 - \mu_2 = 0$ vs. $H_1: \mu_1 - \mu_2 \neq 0$

$$\bar{X}_1 - \bar{X}_2 \sim N[\mu_1 - \mu_2, s_p^2((1/n_1) + (1/n_2))]$$

Standardizing, $(\bar{X}_1 - \bar{X}_2) / s_p \sqrt{(1/n_1) + (1/n_2)}$.

The pooled estimate of the variance from two independent samples is given by

$$s_p^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/d.f., \text{ where } d.f. = (n_1 + n_2 - 2).$$

If we can assume that the variances of the two samples are equal, then compute the test statistic:

$$t_s = (\bar{X}_1 - \bar{X}_2) / (s_p \sqrt{(1/n_1) + (1/n_2)}).$$

If $t_s > t_{1-\alpha/2, [n_1+n_2-2]}$ or $t_s < t_{\alpha/2, [n_1+n_2-2]}$ **Reject H_0 .**

If t_s is between these values, then do not reject H_0 .

Exact p-values are obtained by multiplying by 2 the area to the left or right of t_s under a $t_{[n_1+n_2-2]}$ distribution.

In general, d.f. is the d.f. associated with s . For the one sample or two sample paired situation, $d.f. = n - 1$. For the two sample (unpaired) situation, $d.f. = n_1 + n_2 - 2$.

Two-sided $(1 - \alpha) * 100\%$ C.I. for $\mu_1 - \mu_2$ is given by

$$\{((\bar{X}_1 - \bar{X}_2) - t_{1-\alpha/2, [n_1+n_2-2]} * s_p \sqrt{(1/n_1) + (1/n_2)}), ((\bar{X}_1 - \bar{X}_2) + t_{1-\alpha/2, [n_1+n_2-2]} * s_p \sqrt{(1/n_1) + (1/n_2)})\}.$$

One-sided upper or lower C.I. are computed in the usual way by substituting $\alpha/2$ for $\alpha/2$ in the appropriate equations.

Comparison of Two Variances

If the assumption about equal variances for the two samples is doubted, we can use an F-test, commonly called F', to determine the validity of this assumption.

Population 1: $X_1 \sim N(\mu_1, \sigma_1^2)$

Population 2: $X_2 \sim N(\mu_2, \sigma_2^2)$

$\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ are all unknown.

Use a two-sided test:

$H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_1: \sigma_1^2 \neq \sigma_2^2$

Sample 1: n_1, \bar{X}_1, s_1^2

Sample 2: n_2, \bar{X}_2, s_2^2

Under H_0 , $F_s = s_1^2 / s_2^2$ follows an F-distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ d.f.

For practical purposes label the populations (hence, the samples) such that $s_1^2 > s_2^2$, i.e., $F_s > 1.0$.

Then reject H_0 if $F_s > F_{1-\alpha/2}[\nu_1, \nu_2]$

This test is still conducted at the $\alpha * 100\%$ level of significance, but we use the upper-tail value in determining the F_s .

Because one tests with a one-sided test statistic, obtain F_s and compare to $F_{1-\alpha}[\nu_1, \nu_2]$.

Two Sample t-test for Unequal Independent Variances – the Behrens-Fisher Test

If the two variances are unequal, then we must use a two sample t-test for unequal independent variances, called the **Behrens-Fisher test**.

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

The appropriate d.f. must now be calculated based on s_1^2, s_2^2, n_1, n_2 .

$$d' = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]}$$

Traditionally, the d.f. for the t-distribution are d'' , which is d' rounded to the next smaller integer.

However, because we use SAS statistical software or other software to obtain p-values, we will use d' with all available decimal places.

P-values and C.I.'s are obtained in the usual way from a t_d -distribution.

Sample size needed for comparing the means of two normally distributed samples of equal size using a two-sided test may be obtained by using the following formula:

$$n = (s_1^2 + s_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2 / d'^2$$

If the variances in the two groups are the same, the smallest total sample size involves equal sample sizes.

Sometimes it is not possible or practical to have equal sample sizes.

We can calculate the sample sizes needed for comparing the means of two normally distributed samples for a two-sided test using the following formulae:

$$n_1 = (s_1^2 + s_2^2/k)(z_{1-\alpha/2} + z_{1-\beta})^2 / d'^2$$

and

$$n_2 = (k s_1^2 + s_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2 / d'^2$$

where $k = n_2/n_1 =$ projected ratio of the two sample sizes.

For one-sided tests, substitute $1-\alpha$ for $1-\alpha/2$ in the formulae above.

Summary: Important Statistical Concepts Related to Estimation and Hypothesis Testing

1. The research question is stated in terms of what the investigator expects to find. Often this statement will suggest a directional result that the conclusion needs to address. In general, the research question and the conclusion should be consistent. However, the conclusion may state the direction of the finding even though the research question does not.
2. Using a one-sided hypothesis test or confidence interval requires extensive justification, e.g., previous results. Without justification, two-sided tests and confidence intervals should be used, regardless of whether the research question suggests directionality.

For example, a research problem, which explicitly states that there is no previous evidence, does not support the advantage of using a one-sided test or confidence interval. Thus, a two-sided test and confidence interval should be used.

3. Once the investigator decides on a one-sided or two-sided approach, the hypotheses, p-value associated with the test statistic, the critical value(s), and the confidence limit(s), as well as sample size and power calculations, must be consistent. The investigator should make a decision and stick with it throughout the analysis of results. It is better to make a single error of judgment and maintain it consistently than give the impression of indecision or potentially an indication that s/he is uncertain of what to do.
4. Statistics is never having to say you are certain! Therefore, a p-value, such as 4.0874×10^{-11} or 4.0874E-11, should be rounded up to $<.0001$. The probability of 1.0 should be rounded down to $>.9999$.
5. Although confidence intervals and hypothesis tests are linked mathematically, conceptually these represent very different approaches to analyzing data. Conclusions about one approach should not contain a reference to the other.
6. The conclusion must refer to the parameter being estimated (CIs) or tested under the null hypothesis. Thus, many z-tests, all t-tests, and the corresponding confidence intervals require conclusions that reference means. Many chi-square tests and F tests require conclusions that correspond to variances and standard deviations. However, some chi-square and z-tests, as well as exact poisson and binomial tests, have conclusions that refer to proportions or rates; these will be presented next semester.
7. Sample size calculations, resulting in small group sizes (e.g., $n < 100$), require adjustment. As sample size decreases below 20, this adjustment becomes increasingly more important. This rule applies to one sample, two sample, and multiple sample situations.

NONPARAMETRIC METHODS

If we do not wish to make assumptions about the shape of the distribution or if the central limit theorem is not applicable, then we must use **nonparametric statistical methods** to analyze data and make inferences.

A more descriptive term for these methods is **distribution-free methods**.

Often distribution-free methods are required because the data are **ordinal**.

Ordinal data are data that can be ordered but do not have specific numerical values.

In contrast, **cardinal data** are data that are measured on a scale where common arithmetic is meaningful.

Ordinal variables cannot be given numerical scales that make sense biologically or clinically.

Essentially, the ranks are arbitrarily assigned; these could be reversed and still retain the same meaning for the researcher.

Therefore, if data are ordinal, computation of means and standard deviations is absurd because there would be no universally accepted meaning (i.e., outside of a researcher's laboratory or clinic).

Medians and ranges are used instead.

In general, nonparametric methods are more flexible than parametric methods because nonparametric methods require fewer or no assumptions about the shape of the underlying distribution.

The Wilcoxon Sign Rank Test

This test is the nonparametric counterpart for the paired t-test.

This test based on the ranks of the observations rather than on their actual values.

This test is more powerful than the sign test because both sign and the magnitude of the differences based on rank are used in computing the test statistic.

$H_0: \mu = 0$ vs. $H_1: \mu \neq 0$,

where μ = the median difference in score between treatments A and B.

Ranking procedure:

1. Compute the differences, d_i , where $d_i = a_i - b_i$.

2. Arrange the differences in order of absolute value.
3. Count the number of differences with the same absolute value.
4. Ignore the observations with $d_i = 0$.
5. Rank the remaining observations from 1, for the observation with the lowest absolute value, up to n , for the observation with the highest absolute value.
6. If there is a group with several observations with the same absolute value, then find the lowest rank in the range = $1 + R$ and the highest rank in the range = $G + R$, where R = the highest rank used prior to considering this group and G = the number of differences in the range of ranks for the group.
7. Assign the average rank = $(\text{lowest rank in the range} + \text{highest rank in the range}) / 2$ as the rank for each difference in the group.

The Wilcoxon Sign Rank test is based on the sum of the ranks or rank sum (R_1) for the group of subjects with positive d_i 's.

A large rank sum indicates that differences favor treatment B, whereas a small rank sum indicates that the differences favor treatment A.

Under H_0 , $E(R_1) = n(n + 1)/4$

and

$\text{Var}(R_1) = n(n + 1)(2n + 1)/24$, where n = the number of nonzero d_i 's.

Usually we base the sign rank test on positive differences.

However, if we base the test on negative differences, we will always get the same test statistic and p-value.

Thus, we can arbitrarily compute the rank sum based on either positive or negative differences.

Testing procedure:

1. Compute the rank sum R_1 of the positive differences.
2. If there are no ties, compute

$$z_s = [|R_1 - (n(n + 1)/4)| - .5] / \sqrt{n(n + 1)(2n + 1)/24}$$

Note that this test statistic has the same form as z_s and is a z-score using the normal approximation to the binomial:

$$z_s = \frac{|R_1 - E(R_1)| - .5}{\text{VAR}(R_1)}$$

3. If there are ties, t_i refers to the number of differences with the same absolute value in the i -th tied group and g is the number of tied groups.

Compute the variance of R_1 as $\text{Var}(R_1) = [n(n+1)(2n+1)/24 - \sum_{i=1}^g (t_i^3 - t_i)]/48$

4. Compute the test statistic

$$z_s = [|R_1 - n(n+1)/4| - .5] / \sqrt{\text{Var}(R_1)}$$

5. If $z_s > z_{1-\alpha/2}$ then reject H_0 .
6. The p-value is computed in the usual way for a two-sided test:
 $p = (1 - \text{PROBNORM}(z_s))^*2$.
6. Restriction: This test should be used only if the number of nonzero differences ≤ 16 . Otherwise, exact tables must be used to determine critical values, e.g., Table 11 in Rosner's text.

The Wilcoxon sign rank test may be applied to cardinal data with an underlying continuous, symmetrical, but not necessarily normal, distribution.

If the distribution is normal, then this test has less power than the paired t-test; otherwise, it is the more powerful test.

The Wilcoxon Rank Sum Test or The Mann-Whitney U Test

This test is the nonparametric counterpart of the t-test for two independent samples.

This test is based on the ranks of the individual observations rather than on the actual values.

$H_0: M_1 = M_2$ vs. $H_1: M_1 \neq M_2$,
where M_1 and M_2 are the median responses of the two independent groups.

The test statistic is the sum of the ranks in the first sample (R_1).

If R_1 is large, then the first group is poorer.

Under H_0 , the expected value of the first group (designation is arbitrary) is the product of the sample size and the average rank in the combined sample.

$$E(R_1) = n_1(n_1 + n_2 + 1)/2.$$

$$\text{Var}(R_1) = n_1 n_2 (n_1 + n_2 + 1)/12.$$

Ranking procedure:

1. Combine the data from the two groups and order the values from the lowest to highest or from the best to the worst.
2. Assign ranks to the individual values with the best score having the lowest rank and the worst score having the highest rank.
3. If a group of observations has the same value, assign the average rank for each observation in the group, as described above for the Wilcoxon Sign Rank Test.
4. If the smaller of the two groups contains at least 10 observations, the normal approximation may be used. Otherwise, exact tables must be used to determine critical values, e.g., Table 12 in Rosner's text.

Testing procedure:

1. Compute the rank sum R_1 for one of the groups. The choice of the sample is arbitrary.
2. If there are no ties, compute
$$z_s = [|R_1 - (n_1(n_1 + n_2 + 1)/2)| - .5] / \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}.$$

Note that the test statistic has the form of a z_s :

$$z_s = \frac{|R_1 - E(R_1)| - .5}{\sqrt{\text{VAR}(R_1)}}$$

3. If there are ties, compute the variance of R_1 :

$$\text{Var}(R_1) = (n_1 n_2) * [n_1 + n_2 + 1 - \sum_{i=1}^g \{t_i(t_i^2 - 1) / (n_1 + n_2)(n_1 + n_2 - 1)\}] / 12$$

where t_i refers to the number of observations with the same value in the i -th tied group and g is the number of tied groups.

4. If $z_s > z_{1-\alpha/2}$ then reject H_0 .
5. Compute the p-value in the usual way for a two-sided test: $p = (1 - \text{PROBNORM}(z_s)) * 2$.
6. Restriction: This test should be used only if both n_1 and n_2 are ≥ 10 .

If either sample size is < 10 , we must use exact significance levels. Table 12 in Rosner's text gives upper and lower critical values for the rank sum ($T = z_s$) for two-sided tests with significance levels of 0.10, 0.05, 0.02 and 0.01, respectively. In general, the results are statistically significant at a particular significance level only if $T \leq T_L$ or $T \geq T_U$.

The **Mann-Whitney U Test** was developed for comparisons of medians, which come from underlying continuous distributions, whereas the Wilcoxon Rank Sum Test was developed for ordinal data.

Only small losses of power occur when this test is applied to data from discrete distributions.

General Comments

We can apply nonparametric tests to any cardinal data.

This application may be particularly appropriate when the assumption of normality appears to be grossly violated.

If the actual underlying distribution is in fact normal, then we pay a penalty because the nonparametric counterparts for parametric test statistics has less power.

Often data are not normally distributed even though a reasonable assumption has been made that the underlying (theoretical) distribution is normal.

Parametric methods are often robust to certain kinds of departures from normality.

A practice recommended by some statisticians is for analyzing continuous data by both parametric and nonparametric methods (if applicable).

If the results of both analyses are consistent, the researcher can feel assured that the results reported from parametric tests are not biased.

Results from these analyses may not be consistent, i.e., results from one analysis may be significant and results from the other very far from significance.

In this event results from parametric tests are probably biased.

After reviewing the data carefully, the researcher should (1) consider options for transforming the data so that parametric tests are valid or (2) report results from the nonparametric tests.

Whenever it is appropriate to use nonparametric methods, these are usually more powerful than parametric counterparts and the results of tests are unbiased.

ANALYSIS OF VARIANCE

Analysis of variance methods are applicable to structured data, i.e., data coming from different groups, populations or treatments for purposes of

- (1) partitioning the total variation due to different sources; or
- (2) detecting possible differences among groups.

i = group

j = each observation labeled separately, where $j = 1, 2, \dots, n_i$ (n_i 's are sample sizes).

We can obtain the variation within populations (groups) and among populations (groups).

In general,

	GROUP		
	1	2	a
Obs.	Y_{11}	Y_{21}	Y_{a1}
	Y_{12}	Y_{22}	Y_{a2}
	Y_{13}	Y_{23}	Y_{a3}
	Y_{1n}	Y_{2n}	Y_{an}
Total	$Y_{1.}$	$Y_{2.}$	$Y_{a.}$
Mean	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$	$\bar{Y}_{a.}$

Notation: The sum over the subscript is replaced by the dot:

$$Y_{i.} = \sum_{j=1}^n Y_{ij} \qquad \bar{Y}_{1.} = Y_{1.} / n$$

$$Y_{..} = \sum_{i=1}^a \sum_{j=1}^n Y_{ij} \qquad \bar{Y}_{..} = Y_{..} / an$$

Intuitively, for a sample

$$Y_{ij} = \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i.})$$

$$(Y_{ij} - \bar{Y}_{..}) = (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i.})$$

$$\sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^a \sum_{j=1}^n (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2$$

$$SS_{\text{Total}} = SS_{\text{Among}} + SS_{\text{Within}}$$

Let $\mu = \frac{1}{a} \sum_{i=1}^a \mu_i$, which is μ of the “super population”, then identically

$$\sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \mu)^2 = \sum_{i=1}^a (\mu_i - \mu)^2 + \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \mu_i)^2$$

or by dividing by an

$$1/an * \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \mu)^2 = 1/a * \sum_{i=1}^a (\mu_i - \mu)^2 + 1/an * \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \mu_i)^2$$

Total Variation Among Group Within Group
Variation Variation

(Variation of
population means)

Within groups variation = $1/a \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \mu_i)^2 / n$, which is similar to a variance.

We compute the variation within the i -th group and then average (or pool) these over all groups.

Therefore, $E[1/\{a(n-1)\} * \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2] = \sigma^2$, which is the $MS_{\text{Within groups}}$ and expresses the within group variability.

To obtain the $E(MS_{\text{Among groups}})$ we must pursue the following mathematical argument.

Recall that μ_i is unknown and estimated by $\bar{Y}_{i.}$; μ is also unknown and estimated by $\bar{Y}_{..}$.

Then, for a sample, $\sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^a (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2$

For the **linear statistical model** we can write

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where μ_i is a constant, $\bar{Y}_{i.} \sim N(\mu_i, \sigma^2)$ and ϵ_{ij} is distributed $N(0, \sigma^2)$.

Basic assumptions:

1. Y_{ij} is normally distributed
2. The a populations have the same variance, σ^2 .

Then,

$$\begin{aligned} n \sum_{i=1}^a (\bar{Y}_{i.} - \bar{Y}_{..})^2 &= n \sum_{i=1}^a [(\mu_i + \bar{\epsilon}_{i.}) - (\mu + \bar{\epsilon}_{..})]^2 \\ &= n \sum_{i=1}^a [(\mu_i - \mu) + (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})]^2 = n \sum_{i=1}^a (\mu_i - \mu)^2 + n \sum_{i=1}^a (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2 + 2n \sum_{i=1}^a (\mu_i - \mu)(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}) \end{aligned}$$

Therefore,

$$E\left[n \sum_{i=1}^a (\bar{Y}_{i.} - \bar{Y}_{..})^2\right] = n \sum_{i=1}^a (\mu_i - \mu)^2 + (a-1) \sigma^2, \text{ which is called the } SS_{\text{Among Groups}}.$$

$E\left[\frac{1}{(a-1)} * n \sum_{i=1}^a (\bar{Y}_{i.} - \bar{Y}_{..})^2\right] = \left[n \sum_{i=1}^a (\mu_i - \mu)^2\right] / (a-1) + \sigma^2$, which is called the $MS_{\text{Among groups}}$ and expresses the variation among groups.

$$\sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^a \sum_{j=1}^n (\epsilon_{ij} - \bar{\epsilon}_{i.})^2$$

$$E\left[\sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2\right] = a(n-1) \sigma^2, \text{ which is called the } SS_{\text{Within groups}}.$$

The ratio of these two $E(MS)$, i.e., $E(MS_{\text{Among groups}}) / E(MS_{\text{Within groups}})$, should be roughly equal to 1, if the H_0 is true.

$$F_s = \frac{\text{Variance between groups}}{\text{Variance within groups}} = \frac{n/(a-1) \sum_{i=1}^a (\mu_i - \mu)^2 + \sigma^2}{\sigma^2}$$

We want to test

$$H_0: \mu_1 = \mu_2 = \dots = \mu_a$$

H_1 : The μ 's are not all equal.

Or equivalently,

$$H_0: \left[\sum_{i=1}^a (\mu_i - \mu)^2 \right] / a-1 = 0$$

$$H_1: \left[\sum_{i=1}^a (\mu_i - \mu)^2 \right] / a-1 > 0$$

We will use a one-sided test.

ONE-WAY CLASSIFICATION ANOVA

For a balanced one-way ANOVA we have a groups and n observations/group.

Y_{ij} = the j -th observation in the i -th group, where $i = 1, \dots, a$ and $j = 1, \dots, n_i$

We wish to see if treatments have different effects or if the observations come from populations with different means.

Therefore, we want to test

$$H_0: \mu_1 = \mu_2 = \dots = \mu_a$$

H_1 : The μ 's are not all equal.

or equivalently

$$H_0: \left[\sum_{i=1}^a (\mu_i - \mu)^2 \right] / a-1 = 0$$

$$H_1: \left[\sum_{i=1}^a (\mu_i - \mu)^2 \right] / a-1 > 0$$

For $\left[\sum_{i=1}^a (\mu_i - \mu)^2 \right] / a-1$ to be equal to zero all μ_i 's must be equal.

Because μ equals the average of the μ_i 's, the numerator cannot be negative.

Therefore, we will use a one-sided test.

If H_1 is true, then not all of the treatments have the same effect.

Under H_0 ,

$$F_s = \frac{\sum_{i=1}^a (\bar{Y}_i - \bar{Y}_{..})^2 / (a-1)}{\sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2 / a(n-1)}$$

and follows an F-distribution with $\nu_1 = a-1$ and $\nu_2 = a(n-1)$ degrees of freedom.

ν_2 : For each sample we have $n-1$ d.f. for computing s^2 .

We compute this a times (once for each group).

Therefore, we have $a(n-1)$ d.f.

Thus, we compute s^2 for each sample and average it over the different groups.

ν_1 : We get this from the number of groups that we are averaging to compute the s^2 .

If $F_s > F_{1-\alpha, (a-1), a(n-1)}$, then we reject H_0 , i.e., conclude that the population group means are different or that the treatments have different effects (at the $1-\alpha$ *100% level of significance).

Whenever we use ANOVA procedures, we are assuming that (1) all observations are taken from a normal distribution and (2) because each group has the same variance, the only possible difference among groups is attributable to differences in the means.

Model I ANOVA (Fixed Effects Model):

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

$$= \mu + \alpha_i + \epsilon_{ij}$$

where $\alpha_i = \mu_i - \mu$, and

where $\mu = 1/a \sum_{i=1}^a \mu_i$ and $\sum_{i=1}^a \alpha_i = 0$.

μ_i , μ , and α_i are constants and ϵ_{ij} is a random variable that is distributed $N(0, \sigma^2)$.

$$H_0: \sum_{i=1}^a (\mu_i - \mu)^2 / a - 1 = 0$$

$$\text{or } \sum_{i=1}^a \alpha_i^2 / (a-1) = 0$$

$$H_1: \sum_{i=1}^a (\mu_i - \mu)^2 / a - 1 > 0$$

$$\text{or } \sum_{i=1}^a \alpha_i^2 / (a-1) > 0$$

We will use Model I when we have certain treatment groups (a finite number) and we are interested in these and only these.

If we have the same number of samples/treatment, it is much easier for us to compute statistics and interpret results.

Equal sample sizes or balanced data are not required.

We can accommodate easily small differences in sample size.

However, large differences and missing cells in more complex models can create major problems and make interpretation very difficult.

ANOVA Table

Source	d.f.	SS	MS	F-ratio (F_s)	EMS
Among					
group	a-1	SS_A	$MS_A = SS_A/a-1$	MS_A/MS_W	$[n \sum_{i=1}^a (\mu_i - \mu)^2]/(a - 1) + \sigma^2$
Within					
group	a(n-1)	SS_W	$MS_W = SS_W/a(n-1)$		σ^2
Total	an-1	SS_T			

Estimation of σ^2 :

We wish to estimate the common or pooled variance of each group.

$$\hat{\sigma}^2 = MS_W$$

Kruskal-Wallis Test:

If we are using Model I (fixed effects) for a one-way ANOVA, then we can test the hypothesis that the medians of the groups are the same (versus the alternative that not all are the same) using the Kruskal-Wallis test, which does not require the assumption of normality.

Model II ANOVA (Random Effects Model):

$Y_{ij} = \mu + \alpha_i + \beta_{ij}$, where μ = a constant and α_i and β_{ij} are random variables.

$$\alpha_i \sim N(0, \sigma^2)$$

$$\beta_{ij} \sim N(0, \sigma^2)$$

$$\sigma_Y^2 = \sigma^2 + \sigma^2$$

σ^2 indicates the extent to which individuals vary.

σ^2 indicates the extent to which groups vary.

We wish to ask questions about σ^2 , i.e., do all observations come from the same population or do they come from different populations?

We use this model when we are interested in more than just the groups tested or when we are trying to make an inference about a total population.

In general, Model II is used to make a broader statement than Model I.

$$H_0: \sigma^2 = 0$$

$$H_1: \sigma^2 > 0$$

ANOVA Table

Source	d.f.	SS	MS	F-ratio (F_s)	EMS
Among group	a-1	SS_A	$MS_A = SS_A/a-1$	MS_A/MS_W	$n\sigma^2 + \sigma^2$
Within group	a(n-1)	SS_W	$MS_W = SS_W/a(n-1)$		σ^2
Total	an-1	SS_T			

Example for Model II:

a = 10 individuals (randomly selected from a population)

n = 3 determinations per individual of DNA from liver cells

μ_i = effect due to the i-th individual (among individuals)

μ_{ij} = effect due to the j-th determination in the i-th individual (among determination or within individual)

Generally, we are trying to find sources of variation in an attempt to control the total variation.

Thus, we are breaking down the variability to see where it is ultimately coming from.

$$H_0: \sigma^2 = 0 \text{ vs. } H_1: \sigma^2 > 0$$

This is a one-sided test.

σ^2 expresses the variation among individuals.

$$E(MS_A) = \sigma^2 + n\sigma^2$$

$$E(MS_W) = \sigma^2$$

Test Statistic:

$$F_s = MS_A / MS_E \sim F_{a-1, a(n-1)}$$

If $F_s > F_{1-\alpha, a-1, a(n-1)}$, then reject H_0 and conclude that $\sigma^2 > 0$.

Estimation of Variance Components:

The σ^2 and σ^2 are called **variance components**.

$$\hat{\sigma}_Y^2 = \hat{\sigma}^2 + \hat{\sigma}^2$$

$$\hat{\sigma}^2 = MS_W$$

$$\hat{\sigma}^2 = (MS_A - MS_W) / n$$

If the numerator of $\hat{\sigma}^2$ is negative, then we assume that $\hat{\sigma}^2 = 0$.

To estimate or compare $\hat{\sigma}^2$, one should have a large sample size.

We can also calculate the percent of the total variation, which is attributable to various sources.

$$\hat{\sigma}^2 / (\hat{\sigma}^2 + \hat{\sigma}^2) * 100\% \text{ and}$$

$$\hat{\sigma}^2 / (\hat{\sigma}^2 + \hat{\sigma}^2) * 100\%$$

These quantities are frequently called **reliability** or **repeatability** depending on the interpretation of the data.

Estimating the Intraclass Correlation Coefficient:

If we take our Model II but define the EMS somewhat differently, we have an **alternative model** for describing the situation in which members of the same class tend to act alike.

Suppose that the observations Y_{ij} are all distributed about the same mean μ with the same variance σ^2 , but that any two members of the same class (where i is constant) have a common correlation coefficient ρ_1 , called the **intraclass correlation coefficient**.

<u>Source</u>	<u>MS</u>	<u>EMS</u>
Between classes	s_b^2	$s^2\{1 + (n - 1)\rho_1\}$
Within classes	s_w^2	$s^2(1 - \rho_1)$

This model is useful in applications in which it is natural to think of members of the same class as correlated, e.g., studies of twins where a highly positive correlation generally occurs.

The model is more general than the components of variance model (Model II).

If ρ_1 is negative, s_b^2 has a smaller expected value than s_w^2 .

With Model II this situation cannot occur.

However, this situation can be a reflection of real phenomenon, and not be an accident of sampling.

For example, caged rats competing for a limited food supply may get unequal amounts of food, thereby making s_b^2 less than s_w^2 .

One restriction on this model is that $\hat{\rho}_I$ cannot be less than $-1/(n - 1)$, because $E[s_b^2] \geq 0$.

$$\hat{\rho}_I = \frac{[(a - 1)s_b^2] - a * s_w^2}{[(a - 1)s_b^2] + [a(n-1)s_w^2]}$$

In contrast, if we use Model II,

$$\hat{\rho}_I = \frac{s_b^2}{s_b^2 + s_w^2}$$

ANOVA with Unequal Sample Sizes:

Computational formulas for Model I and II ANOVA with n_i observations in the i -th sample and n , total number of observations.

For Model I ANOVA,

$$E(MS_A) = \sigma^2 + \left[\sum_{i=1}^a n_i(\mu_i - \mu)^2 \right] / (a - 1), \text{ where}$$

$$\left[\sum_{i=1}^a n_i(\mu_i - \mu)^2 \right] / (a - 1) = 0 \mid H_0 \text{ and } \mu = \left(\sum_{i=1}^a n_i \mu_i \right) / n.$$

For Model II ANOVA,

$$E(MS_A) = \sigma^2 + n_0 \sigma_A^2$$

where $n_0 = 1 / (a - 1) * [n - \left(\sum_{i=1}^a n_i^2 \right) / n]$

Thus, n_0 is the **harmonic mean** and represents the **average number** or **effective number** of observations in each group.

We need n_0 to determine $\hat{\sigma}^2$ and $\hat{\sigma}_A^2$.

The effect of unequal subclass numbers is more important in Model II cases than in Model I cases.

COMPARISONS AMONG POPULATION (GROUP) MEANS WITHIN THE CONTEXT OF ANOVA

Comparisons of the Means and Variances of Two Populations or Groups:

This is a special case of the Model I ANOVA. We have discussed this situation as a **two independent sample t-test** with equal variances. Alternatively, this comparison may be handled by use of ANOVA procedures in which $F_s = t_s$.

Comparisons Among the Means of More than Two Populations or Groups:

A frequent error made by researchers in performing comparisons among means is that they use in a pair-wise fashion only the sample variances for the groups that they wish to compare rather than the pooled estimate of σ^2 based on all samples.

If we use the sample variances from only the samples we are comparing, we will get a different estimate of the σ^2 for each comparison we choose to make.

This procedure is flawed because, before we started the experiment, we assumed or determined that all groups had the same underlying variance.

Furthermore, because the estimate of the pooled variance is based on more information (larger n), this estimate of σ^2 (obtained from all of the groups) will be more accurate than any of those from selected subgroups.

Increased accuracy of the estimate of σ^2 is the principle advantage of performing t-tests within the framework of the ANOVA.

Considering each pair of groups or each collection of groups separately and performing individual t-tests – as if the experiments were separate – is costly in terms of efficiency and can produce results that are misleading.

Because 1- α or power may be very low for individual tests, we are likely to make Type II errors. In addition, we may not be maintaining our experiment-wise Type I error rate at α , even though the comparison-wise error rate is α .

By using ANOVA, we may design specific comparisons to test for differences between each pair of means or to test for general differences among collections of means.

We use the overall F_s to test for general differences among the means.

If we decide in favor of H_0 , then we do only the pre-planned tests.

If we decide in favor to H_1 , then we can do either pre-planned tests or all possible (unplanned) tests.

A Priori Tests – Planned or pre-planned comparisons:

Other names for this type of comparison are **linear contrast** and **orthogonal contrast**.

1. Usually few ($a-1$)
2. Suggested by subject matter considerations, i.e., the investigator has thought it out.
3. Not determined by data from the present study.

Example: Sections grown in tissue culture with $a = 5$ treatments (groups)

Treatments

- 1 = Control
- 2 = 2% glucose
- 3 = 2% fructose
- 4 = 1% glucose + 1% fructose
- 5 = 2% sucrose

$a - 1 = 4$ or 4 planned or orthogonal comparisons are allowed based on the d.f.

Comparison I: Control vs. Sugars

$$H_0: \mu_1 - 1/4 (\mu_2 + \mu_3 + \mu_4 + \mu_5) = 0$$

Comparison II: Mixed vs. Pure Sugars

$$H_0: \mu_4 - 1/3 (\mu_2 + \mu_3 + \mu_5) = 0$$

Comparison III: Sucrose vs (Glucose and Fructose)

$$H_0: \mu_5 - 1/2 (\mu_2 + \mu_3) = 0$$

Comparison IV: Glucose vs. Fructose

$$H_0: \mu_2 - \mu_3 = 0$$

In general, any comparison or contrast can be written as

$$C = c_1\mu_1 + c_2\mu_2 + \dots + c_a\mu_a \text{ with } c_1 + c_2 + \dots + c_a = 0$$

Comparison	Coefficients for				
	μ_1	μ_2	μ_3	μ_4	μ_5
1	1	-1/4	-1/4	-1/4	-1/4
2	0	-1/3	-1/3	1	-1/3
3	0	-1/2	-1/2	0	1
4	0	1	-1	0	0
or alternatively					
1	4	-1	-1	-1	-1
2	0	-1	-1	3	-1
3	0	-1	-1	2	0
4	0	1	-1	0	0

$H_0: C = 0$ vs. $H_1: C \neq 0$ for two-tailed tests or $H_0: C > 0$ or $H_1: C < 0$ for one-tailed tests.

C is estimated by $\hat{C} = c_1 \bar{Y}_1 + c_2 \bar{Y}_2 + \dots + c_a \bar{Y}_a$
with $\text{Var}(\hat{C}) = (c_1^2/n_1 + c_2^2/n_2 + \dots + c_a^2/n_a) MS_W$,
which is estimated by $\text{Var}(\hat{C}) = (c_1^2/n_1 + c_2^2/n_2 + \dots + c_a^2/n_a) MS_W$

Then, $t_s = \hat{C} / \sqrt{\text{Var}(\hat{C})}$ follows a t-distribution with $\nu = (n - a)$ d.f.

We reject H_0 if $|t_s| > t_{1-\alpha/2, n-a}$

Alternatively, $F_s = (\hat{C})^2 / \text{Var}(\hat{C}) = t_s^2$ follows an F-distribution with $\nu_1 = 1$ and $\nu_2 = n - a$ d.f.

We reject H_0 if $F_s > F_{1-\alpha, 1, n-a}$

In the context of the ANOVA table, $F_s = MS_C / MS_W = SS_C / MS_W$, where $MS_C = SS_C$ is the sum of squares with 1 d.f. associated with the contrast.

When we use *a priori* orthogonal contrasts, the SS_C should add up to the $SS_{\text{Among Treatments}}$.

Thus, we have described the variation among treatments by a set of meaningful (in the context of the study) **orthogonal contrasts**.

For SAS analyses, be sure to use the type III SS for contrasts. The type I SS can be incorrect.

Orthogonal contrasts may not always be meaningful.

We can specify a “reasonable” number of **pre-planned, but non-orthogonal contrasts**.

These can be used to detect significant differences among treatment means and determine significance levels, even when the overall F-test for treatments is not significant.

Reasonable implies that we are not making “all possible” comparisons, but only those that we have thought out.

The **overall test** gives an “average” significance level.

Sometimes one or a few treatment(s) may be significantly different, but this difference may be obscured because several other treatments are so similar.

Another useful application of linear contrasts can be made, if the different groups represent different levels of a particular treatment (e.g., doses).

The levels must be equidistant and there need to be equal number of observations in each level.

Generally, we are not interested in the overall test of differences between groups but rather in the characterization of the trend across groups.

We choose the coefficients of the linear contrasts to reflect the particular treatment-response relationship, i.e., linear only; linear and quadratic; etc.

This method is analogous to the orthogonal contrasts introduced in regression analysis.

Other “reasonable” comparisons include comparing each treatment to the control (Dunnett’s test) as well as special sets, determined by the objectives of the study.

For example, A to B, B to C, C to D, and D to E.

These special contrasts may be used when the sample sizes of the groups are small and/or when the overall F-test is not significant.

A Posteriori or Post Hoc Tests

These tests are designed to keep us from making incorrect decisions.

1. Usually many – $a(a - 1)/2$
2. All possible simple comparisons between means.
3. Methods do not allow for grouped comparisons.
4. Suggested by the data from the present study.

Because there are a large number of comparisons, we are likely to find some differences by chance.

Therefore, we need to apply methods that will protect against our making this error.

To ensure that we do not identify too many falsely significant differences, these procedures maintain the overall probability of declaring any difference between all possible pairs of groups to be significant at some fixed level, α .

Without taking these precautions, the overall error rate can be very high – possibly as high as .8.

The multiple comparison procedures are stricter than the ordinary t-tests.

There will be comparisons between means for which individual t-tests would declare a significant difference and the multiple comparison procedure would not.

If the experimentwise error rate is set at 0.05, then the comparisonwise rate needs to be set somewhat lower.

This is the price that we pay for trying to fix the α level of any significant difference among pairs by using multiple comparisons rather than having an “unknown” α level for all possible t-tests.

If we are comparing only two means, then the p-values from both procedures will be identical.

When t-tests and multiple comparison procedures should be applied is controversial (even among statisticians).

Least Significant Difference (LSD) method

This method is only applicable when the F-test in the ANOVA is significant.

Recall: $\mu_1 \neq \mu_2$
if $|\bar{Y}_1 - \bar{Y}_2| / \sqrt{2/n * MS_W} > t_{1-\alpha/2, (n-1)}$
or
 $|\bar{Y}_1 - \bar{Y}_2| > t_{1-\alpha/2, (n-1)} * \sqrt{2/n * MS_W}$

The **right hand side (RHS)** is the same for any comparison.

Hence, any difference between sample means, $|\bar{Y}_1 - \bar{Y}_2|$ must be at least as large as the **least significant difference** or **LSD** ($t_{1-\alpha/2, (n-1)} * \sqrt{2/n * MS_W}$) before the sample means \bar{Y}_1 and \bar{Y}_2 . (or the corresponding population means, μ_1 and μ_2) are significantly different.

Even though the overall F-test is significant, it is possible that none of the comparisons is significant.

LSD Test for Unequal Sample Sizes

All computations must be done individually for each comparison.

$LSD(i, i') = t_{1-\alpha/2, (n-1)} * \sqrt{(1/n_i + 1/n_{i'}) * MS_W}$

Multiple Range Tests

These are based on the range of the sample means and the distribution of the sample means. Some of these tests may be applicable when the F_s from the ANOVA is not significant; rules depend on the specific test.

Examples

- a) **Duncan's multiple range test**
- b) **(Student-) Newman-Keuls (SNK) Test**

Difference between the SNK and LSD is that SNK is applicable when F_s from the ANOVA is not significant and LSD can only be used when F_s is significant.

The SNK test, as well as other multiple range tests, is based on the distribution of the ranges starting with the largest range (i.e., the difference between the largest and smallest sample means).

Successively smaller ranges are then tested (if necessary).

Whenever a difference is declared non-significant, then all means included in that range are also non-significantly different.

All differences can be difficult to sort out.

Results from these tests are not transitive.

A sequential procedure must be used.

The SNK Test

For comparing a range involving k means ($2 \leq k \leq a$), we compute the **least significant range (LSR)**.

$$LSR(k) = Q_{\alpha/2, k, 2} / \sqrt{2} * \sqrt{2/n} * MS_W$$

$$= Q_{\alpha/2, k, 2} * \sqrt{MS_W/n}$$

where 2 is the d.f. associated with MS_W .

If $|\bar{Y}_k - \bar{Y}_1| \geq LSR(k)$, we conclude that $\mu_k > \mu_1$.

$$\bar{Y}_1, \quad \bar{Y}_2, \quad \bar{Y}_3, \quad \bar{Y}_4.$$

SNK Test for Unequal Sample Sizes

If sample sizes are unequal, the SNK test is difficult and time-consuming to compute.

Making more than one comparison is complicated.

Generally, one has to make all possible comparisons.

1. LSR(k) is replaced by LSR(k, i, i') where k = the number of means included in the range and i, i' denote the means to be compared.

$$LSR(k, i, i') = Q_{\alpha/2, k, 2} / 1/2 * (1/n_1 + 1/n_2) * MS_W$$

2. The test is only approximate.
3. Even if a test including k means in the range is not significant, one may have to go to smaller ranges including $k-1, k-2$, etc. means.

Example

Trt	1	2	3	4
Mean	10	11	18	20
n_k	2	30	30	2

Because of small sample sizes, treatments 1 and 4 are compared with low precision (relative to the other comparisons) and, even if judged to be not significant, one should make the other comparisons.

Scheffe Procedure (SS-STP)

This method for unplanned comparisons is the most general procedure. For any contrast $C = c_1\mu_1 + c_2\mu_2 + \dots + c_a\mu_a$ obtain $SS(\hat{C})$ in the same manner as explained for the *a priori* comparisons and reject $H_0: C = 0$ if $SS(\hat{C}) > (a - 1)F_{1 - \alpha, 1, n - a} * MS_W$. The difference between this procedure and the *a priori* test is that this procedure yields a larger critical value. Therefore, it is harder to reject H_0 . We are penalized because we have advance knowledge from the ANOVA.

Example

Trt	1	2	3	4	5
Mean	70	59	58	67	64
n	10	6	4	9	10
c_i	4	-1	-1	-1	-1

$$SS(\hat{C}) = 794.73$$

$$4 * F_{.95}[4,34] * MS_W = 4 * 2.62 * 6.07 = 63.61$$

$$794.73 \gggg 63.61 \Rightarrow \text{Reject } H_0$$

Kruskal-Wallis Test

If we are using Model I (fixed effects) for a one-way ANOVA, then we can compare the groups using the Kruskal-Wallis test without making the assumption of normality. If we wish to compare specific groups or make orthogonal contrasts, we must rank the outcome variable and then proceed with the usual one-way ANOVA and make the appropriate comparisons among the means (which are now medians) of the ranks.

Using SAS, we might write the following SAS commands:

```
PROC NPAR1WAY WILCOXON;  
CLASS GROUP;  
VAR OUTCOME;  
  
PROC RANK;  
RANKS ROUTCOME;  
VAR OUTCOME;  
  
PROC GLM;  
CLASSES GROUP;  
MODEL OUTCOME ROUTCOME= GROUP;  
CONTRAST '1 V 2,3&4'  
GROUP 3 -1 -1 -1;  
LSMEANS GROUP/STDERR PDIFF;  
MEANS GROUP/SNK;
```