Methods Notes

Note: Finished Week 3

Introduction

Statistics Dictionary Definitions:

- Branch of mathematics dealing with the collection, analysis, interpretation, and presentation of data
- Art and science of drawing justifiable conclusions from data

Mathematically, the simple linear regression model in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

The matrix formulation has

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

The unknown parameters are

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{and} \quad \sigma^2.$$

We have the following results:

- **The least squares estimator**

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix},$$

is the minimum variance linear unbiased estimator for $\boldsymbol{\beta}$.

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{V} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

$$\mathbf{c}^\top\hat{\boldsymbol{\beta}} \sim N\left(\mathbf{c}^\top\boldsymbol{\beta},\ \mathbf{c}^\top\mathbf{V}\mathbf{c}\right).$$

- Test $H_0 : \mathbf{c}^\top\boldsymbol{\beta} = 0$ using

$$t = \frac{\mathbf{c}^\top \hat{\boldsymbol{\beta}} - 0}{\sqrt{\mathbf{c}^\top \mathbf{V}\mathbf{c}}}.$$

Statistics is the science of using information to make decisions and quantify uncertainty inherent to those decisions.

There are four basic steps in the statistical problem solving process (Deming):

1. Define the questions to be answered (Plan)
2. Gather appropriate data (Do)
3. Analyze the data (Study)
4. Interpret the results (Act)

Unit 1 Experiments

Terminology

Terminology

Experiment: an investigation in which the investigator applies (assigns) some treatments to experimental units and then observes the effect of the treatments on the experimental units by measuring one or more response variables.

Treatment: a condition or set of conditions applied to experimental units in an experiment.

Experimental Design: The assignment rule specifies which experimental units are to be observed under which treatments.

Experimental Unit: the physical entity to which a treatment is randomly assigned and independently applied. - the smallest division of material (e.g., land, plant, animal, etc.) to be studied

Response Variable: a characteristic of an experimental unit that is measured after treatment and analyzed to assess the effects of treatments on experimental units
(e.g., yield, gene expression level, etc.).

Observational Unit: the unit on which a response variable is measured. There is often a one-to-one correspondence between experimental units and observational units, but that is not always true.

Replication

- Applying a treatment independently to two or more experimental units
- Level of variability can be estimated for units that are treated alike.

Randomization

- Random assignment of treatments to experimental units
- Reduce or eliminate sources of bias (treatment groups are equivalent, *on average*, except for the assigned treatment)
- Cause and effect relationships can be demonstrated
- Create a probability distribution for a test statistic under the null hypothesis of no treatment effects

Blocking / Matching

- Group similar experimental units into blocks

- Apply each treatment to (the same number of) experimental units within each block (balance)
- Separate random assignment of units to treatments is done within each block (randomization)

Blinding

- Subjects do not know which treatment they received
- Researchers making measurements do not know the treatment assignments

Control of Extraneous Variables

- Control non-intervention factors
- Use homogeneous experimental units
- Accurate measurement of outcomes (responses)
- Tradeoff between accuracy and generalizability

Comparison to a Control Group

- Untreated (placebo) group
- Gold standard (best available treatment)

Scope

- Inferences are restricted to only those units used in the experiment
- Extending inferences beyond the units in the experiment
  - Were the units used in the experiment obtained from a **representative random sample** from some larger population?
    - Yes ⇒ can make inferences about the population
    - No ⇒ cannot make inferences about the population

Randomization Tests

Used for randomized experiments

Use the probability distribution imposed by the random assignment of units to treatment groups

- Under the null hypothesis
$$H_0: \text{ treatments have the same effect}$$
the response provided by any particular unit does not depend on the assigned treatment ($\Rightarrow \mu_1 = \mu_2$)
- Is the observed difference $\bar{y}_1 - \bar{y}_2$ inconsistent with H_0?
- Compare $\bar{y}_1 - \bar{y}_2$ with differences in sample means for all other possible random assignments of units to treatment groups
(What if H_0 is true?)

General Comments

- The randomization test is also called the permutation test
- The randomization test (permutation test) depends on identifying units to permute, which should be the units in the experiment that are **exchangeable under the null hypothesis**, determined by the design of the experiment and the factor(s) being tested.

Observational Studies

- In some cases, the treatments cannot be assigned to experimental units by some rule.
  - For example, study of the effects of smoking on cancer with humans as the experimental units
  - Neither ethical nor possible
- We can still gather data by observing some members of the target population as they naturally exist.
  - Census: Observe all members of population
  - Haphazard (convenience) sample
  - Representative random sample
- This type of study is called an observational study and is not an experiment.

Simple Random Sampling

Without Replacement: every subset of n unique units has the same probability of being selected (more typical)

With Replacement: on each draw every member of the population has the same chance of being selected and the selected unit is put back into the population before the next unit is selected (some units may be selected more than once)

Sampling Schemes

- Only consider simple random samples, but there are many other sampling schemes that produce representative samples (Stat 521: Survey Sampling)
- The sampling procedure dictates the method of analysis
- Can make predictions and inferences about associations
- Causal inferences are not justified

Model-based Inference Overview

The Normal

A random variable Y with density function

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}$$

is said to have a **normal (Gaussian)** distribution with

$$\text{Mean} \equiv E(Y) = \mu \qquad \text{and} \qquad \text{Variance} \equiv \text{Var}(Y) = \sigma^2.$$

The standard deviation is

$$\sigma = \sqrt{\text{Var}(Y)}.$$

We will use the notation

$$Y \sim N(\mu, \sigma^2).$$

The Standard Normal

Suppose Z is a random variable with a normal distribution where

$$E(Z) = 0 \quad \text{and} \quad \text{Var}(Z) = 1,$$

i.e.,

$$Z \sim N(0,1),$$

then Z has a **standard normal** distribution.

Linear Combinations

If Y_1 is a random variable with expectation μ_1 and variance σ_1^2 and Y_2 is a random variable with expectation μ_2 and variance σ_2^2, then

$$E(Y_1 + Y_2) = \mu_1 + \mu_2$$

$$E(aY_1 + bY_2 + c) = a\mu_1 + b\mu_2 + c$$

$$\mathrm{Var}(Y_1 + Y_2) = \sigma_1^2 + \sigma_2^2 \quad \text{if } Y_1 \text{ and } Y_2 \text{ are independent}$$

$$\mathrm{Var}(aY_1 + bY_2 + c) = a^2\sigma_1^2 + b^2\sigma_2^2 \quad \text{if } Y_1 \text{ and } Y_2 \text{ are independent}$$

$$\mathrm{Var}(Y_1 + Y_2) = \sigma_1^2 + \sigma_2^2 + 2\,\mathrm{Cov}(Y_1, Y_2)$$

$$\mathrm{Var}(aY_1 + bY_2 + c) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\ \mathrm{Cov}(Y_1, Y_2)$$

Useful Definitions

Variance:

$$\mathrm{Var}(Y_1) = \sigma_1^2 = E\left[(Y_1 - \mu_1)^2\right].$$

Covariance:

$$\mathrm{Cov}(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] = \rho_{12}\sigma_1\sigma_2,$$

where ρ_{12} is the correlation between Y_1 and Y_2.

The **correlation coefficient**

$$\rho_{12} = \frac{\mathrm{Cov}(Y_1, Y_2)}{\sigma_1\sigma_2}$$

measures the strength of the linear relationship between Y_1 and Y_2.

Distribution of a Sample Mean

- Assuming independent observations from a population with mean μ_k, the sample mean

  $$\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj}$$

  is the best linear unbiased estimator for μ_k.

- If $Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$ are i.i.d. $N(\mu_k, \sigma_k^2)$ random variables, i.e., a simple random sample from a normal population, then

  $$\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj} \sim N\left(\mu_k,\ \frac{\sigma_k^2}{n_k}\right).$$

- $\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj}$ is a random variable (an **estimator**).
  Use

  $$\bar{y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} y_{kj}$$

  to denote its **estimate** (observed value).

Distribution for Difference in Two Sample Means For independent simple random samples from two normal populations:

- $Y_{11}, \ldots, Y_{1n_1}$ are i.i.d. $N(\mu_1, \sigma_1^2)$,
- $Y_{21}, \ldots, Y_{2n_2}$ are i.i.d. $N(\mu_2, \sigma_2^2)$.

Then,

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

The Central Chi-Square Distribution Let Z_i, $i = 1, 2, \ldots, n$, be independent standard normal random variables.
The distribution of

$$W = \sum_{i=1}^{n} Z_i^2$$

is called the **central chi-square distribution** with n degrees of freedom.

We denote this by

$$W \sim \chi_\nu^2,$$

where ν is the number of degrees of freedom.

Estimation of Variances For

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2), \qquad Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2),$$

- The sample variance

  $$S_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2$$

  is an unbiased estimator of σ_1^2.

- The sample variance

  $$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^2$$

  is an unbiased estimator of σ_2^2.

- If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (homogeneous variances), the pooled estimator is

  $$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Sum of Independent Chi-Squares The sum of two independent central chi-square random variables with ν_1 and ν_2 degrees of freedom has a central chi-square distribution with $\nu_1 + \nu_2$ degrees of freedom.

Consequently,

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

has a chi-square distribution with

$$(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$$

degrees of freedom.

The Student t-Distribution If

$$Z \sim N(0,1), \quad W \sim \chi^2_r,$$

and Z and W are independent random variables, then the random variable

$$T = \frac{Z}{\sqrt{W/r}}$$

has a **central Student t-distribution** with r degrees of freedom.

We denote this by

$$T \sim t_r.$$

Inference for Difference in Means with Equal Variances

Assumptions

- Two independent random samples:

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \quad \text{and} \quad Y_{21}, Y_{22}, \ldots, Y_{2n_2}$$

- Normality:

$$Y_{1i} \sim N(\mu_1, \sigma_1^2), \quad Y_{2j} \sim N(\mu_2, \sigma_2^2)$$

- Homogeneous population variances:

$$\sigma_1^2 = \sigma_2^2$$

Distribution for Inference

Let

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

Then

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

Hypothesis Testing

Hypotheses

$$H_0 : \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0)$$

$$H_a : \begin{cases} \mu_1 < \mu_2 & \text{(left-tailed)} \\ \mu_1 > \mu_2 & \text{(right-tailed)} \\ \mu_1 \neq \mu_2 & \text{(two-tailed)} \end{cases}$$

Test Statistic

The observed test statistic is

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}.$$

We assess whether this value is typical under H_0 or unlikely assuming H_0 is true.

Sampling Distribution

Assuming H_0 is true,

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

If H_0 is true, we expect T to be close to zero.
Large deviations from zero are unlikely under H_0.

p-Value

Definition:
The p-value is the probability of observing a test statistic at least as extreme as the one observed, assuming H_0 is true.

Interpretation: Scale-of-Evidence Framework

p-value range	Evidence for H_a
$p > 0.10$	little to no evidence
$0.05 < p \leq 0.10$	borderline / weak evidence
$0.025 < p \leq 0.05$	moderate evidence
$0.001 < p \leq 0.025$	strong evidence
$p \leq 0.001$	overwhelming evidence

Post-hoc Assessment: Errors

- If the p-value was small:
  - H_0 is true and we unluckily/randomly made an error
  - Type I error probability:
$$P(\text{reject } H_0 \mid H_0 \text{ true}) \leq \alpha$$
  - H_0 is false (no error committed)
- If the p-value was large:
  - H_a is true and we unluckily/randomly made an error
  - Type II error probability:
$$P(\text{fail to reject } H_0 \mid H_0 \text{ false}) = \beta$$
  - The power of a test is $1 - \beta$
  - H_0 is true (no error committed)

Confidence Intervals

The following is for estimating *differences in means*

Assumptions

- $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are i.i.d. $N(\mu_1, \sigma^2)$
- $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ are i.i.d. $N(\mu_2, \sigma^2)$
- Population variances are equal

- Y_{1i} and Y_{2j} are independent for all i and j

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{n_1+n_2-2,\, 1-\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where

$$S_p = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

Hypothesis Test Interpretation

A $100(1-\alpha)\%$ confidence interval can be constructed by including all values of δ such that the data does not provide sufficient evidence to reject the null hypothesis

$$H_0 : \ \mu_1 - \mu_2 = \delta$$

relative to the two-sided alternative

$$H_a : \ \mu_1 - \mu_2 \neq \delta$$

at the α significance level.

Interval Width

Confidence interval widths depend on:

- the confidence level (which is related to significance α),
- the value of σ,
- sample sizes n_1 and n_2.

Sample Size Considerations

Note: Sample size calculations refer to the experimental units to replicate, not the observational units (though they sometimes are one and the same!)

Based on Standard Error Difference in Means

- Difference in population means $(\mu_1 - \mu_2)$:

$$\text{s.e.}(\bar{Y}_1 - \bar{Y}_2) = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Assuming $n_1 = n_2 = n$, we have:

$$\text{s.e.}(\bar{Y}_1 - \bar{Y}_2) = S_p \sqrt{\frac{2}{n}}$$

- Specify an acceptable value for the standard error and solve for n:

$$\text{s.e.} = \frac{\sqrt{2}S_p}{\sqrt{n}} \quad \Rightarrow \quad n = \frac{2S_p^2}{(\text{s.e.})^2}$$

- Requires a value for S_p from:
  - a previous study
  - a pilot study
  - a guess

Based on Confidence Interval Difference in Means

- Width of the confidence interval (assuming $n_1 = n_2 = n$):

$$w = 2\, t_{2(n-1),\, 1-\alpha/2}\, S_p \sqrt{\frac{2}{n}}$$

- Find n to achieve specified width:

$$n = 8 \left(\frac{t_{2(n-1),\, 1-\alpha/2} S_p}{w} \right)^2$$

- One difficulty is that n enters twice (sample size and degrees of freedom for t):
  - Compute initial value using the normal approximation:

$$n_0 = 8 \left(\frac{z_{1-\alpha/2} S_p}{w} \right)^2$$

- Then improve using:

$$n = 8 \left(\frac{t_{2(n_0-1),\, 1-\alpha/2} S_p}{w} \right)^2$$

Recall: Four Possible Outcomes for Hypothesis Test

Decision	H_0 is true	H_0 is false
Reject H_0	Type I Error	Good Decision
Fail to reject H_0	Good Decision	Type II Error

Based on Hypothesis Test Difference in Means

For a t-test of
$H_0 : \mu_1 = \mu_2$
against
$H_a : \mu_1 \neq \mu_2$:

- Equal sample sizes: $n_1 = n_2 = n$
- Type I error rate: α
- Power: $1 - \beta$ for detecting $\delta = \mu_1 - \mu_2$
- Pooled estimate of population variance: S_p^2

The required sample size for each group is:

$$n = \frac{\left(t_{2(n-1),\,1-\alpha/2} + t_{2(n-1),\,1-\beta}\right)^2 (2S_p^2)}{\delta^2}$$

Based on Hypothesis Test (Two-Step Approach) Difference in Means

- As before, n enters twice. Use the same two-step approach.
- First compute:

$$n_0 = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (2S_p^2)}{\delta^2}$$

- Then update:

$$n = \frac{\left(t_{2(n_0-1),\,1-\alpha/2} + t_{2(n_0-1),\,1-\beta}\right)^2 (2S_p^2)}{\delta^2}$$

- Common to use power values of 80%, 90%, or 95%, just as arbitrary as using $\alpha = 5\%$.
- Can adapt to a one-sided alternative by replacing $\alpha/2$ with α in the formulas.

Inference Diagnostics

Assessing Equal Variances

Graphical Method

- Construct residual plots, histograms, or boxplots of values for each group/population
- Look for:
  - Outliers in each sample
  - Differences in IQR, range
  - Differences in shape of sample distributions

Summary Statistics

- Check the ratio of sample standard deviations

$$\frac{\max\{S_1, S_2\}}{\min\{S_1, S_2\}}$$

- Interpretation guidelines:
  - Between 1 and 2 — little impact
  - Between 2 and 3 — potential impact
  - Greater than 3 — likely impact

F-test

- Reject $H_0 : \sigma_1^2 = \sigma_2^2$ if

$$F_{\max} = \frac{\max\{S_1^2, S_2^2\}}{\min\{S_1^2, S_2^2\}} \geq F_{(a,b),\, 1-\alpha/2}$$

- where
  - $a = n_1 - 1,\ b = n_2 - 1$ if $S_1^2 > S_2^2$
  - $a = n_2 - 1,\ b = n_1 - 1$ if $S_2^2 > S_1^2$
- Notes:
  - Very sensitive to normal distribution assumption
  - Not recommended as the only check

Brown–Forsythe Test

- Conduct a two-sample t-test on the absolute deviations from the sample medians to assess homogeneous variability

Remedies to Unequal Variance Welch Approximation

- Very similar results to two-sample inference when sample sizes are nearly equal
- Better performance with unequal sample sizes **and** unequal variances

Transformation

- Replace Y_{ij} with $X_{ij} = h(Y_{ij})$
- Perform inference on the X_{ij}'s $\rightarrow$ e.g., compare $\bar{X}_1$ with $\bar{X}_2$
- Back-transform estimates to get conclusions on the Y scale
  - only approximate conclusions about population means on the Y scale

Transformation Cont.

- Choosing the transformation
  - Trial and error: transform and check histogram
  - Rules of thumb:
    - Data are all positive — use $\log(Y)$
    - Data are proportions — use $\arcsin(\sqrt{Y})$
    - Data are counts — use $\sqrt{Y}$
  - Use transformation based on science
    (square root of area, cube root of volume)
  - Adjust for a variance–mean relationship
    (common for variance to increase with the mean)

Assessing Normality

Graphical Methods

- Histogram of values within each group/population
  - Look for symmetric, bell shape
- Normal probability plot within each group/population
  - Compare empirical cumulative distribution function (CDF) to CDF for theoretical normal distribution
  - Most commonly done using quantiles (Q–Q plot):
    plot empirical quantiles against expected quantiles from normal distribution

Normal Q–Q Plot

- Order residuals from smallest to largest
  (say $X_{(1)}, \ldots, X_{(n)}$)
- Compute expected quantiles $(q_{(1)}, \ldots, q_{(n)})$ from a standard normal distribution
  - Expected quantiles can be calculated with tables
  - General approximation:
    $$q_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$$
  - Blom approximation:
    $$q_i = \Phi^{-1}\left(\frac{i-.375}{n+.25}\right)$$
  - For $i = 5$, $n = 9$, $q_5 = \Phi^{-1}\left(\frac{5}{10}\right) = 0$
- Scatterplot of $X_{(i)}$ vs q_i should be close to a straight line with slope σ
- Curved patterns indicate non-normal distributions (or presence of outliers)

Numerical Summaries

- For any normal distribution:

  - Mean and median should be equal
  - Skewness = $E(Y-\mu)^3/\sigma^3 = 0$
    (Skewness measures the asymmetry)
  - Kurtosis = $E(Y-\mu)^4/\sigma^4 = 3$
  - Excess kurtosis = kurtosis -3
    (estimated by the *univariate* procedure in SAS)
  - The sample kurtosis measures the heaviness of the tails of the data distribution
  - Positive value: long-tail; negative value: short-tail

Tests

- Many proposed tests for normality
- Tests based on empirical CDFs: Kolmogorov–Smirnov, Anderson–Darling, etc.
- Tests based on skewness or kurtosis
- Chi-square goodness-of-fit tests
- Tests based on normal probability plots: Shapiro–Wilk, correlation tests
- Normality is almost always rejected for large sample sizes

Consequences of Non-Normality

- Large samples → few consequences (Central Limit Theorem)
- Small samples:
  - Sample distributions have same shape and
    - equal sample sizes → very little impact
    - different sample sizes → potential impact if distributions are skewed
  - Sample distributions have different shapes → impact

Remedy for Non-Normality

- Transformation (especially for skewness)
- Discussed earlier (under remedies for unequal variances)
- Detect and eliminate outliers
- Non-parametric tests

Non Parametric Tests Wilcoxon Rank–Sum Test

- Independence
- Null hypothesis: two populations have the same distribution
  - Distribution is not required to be normal
  - Implies equal medians, percentiles, means, and variances
- Can test against one- or two-sided alternative
- Can compute "exact" p-values based on the null distribution of the ranks

Wilcoxon Rank–Sum Test (Procedure)

- Order the combined $n_1 + n_2$ observations (small to large)

- Assign ranks
  - Smallest gets rank = 1, second smallest gets rank = 2, etc.
  - For tied observations, average the ranks
- Compute the sum of the ranks for one group (call it W)
- Assuming H_0 is true, compute:

$$E_0(W) = \frac{n_1(n_1 + n_2 + 1)}{2} \qquad \text{and} \qquad V_0(W) = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

- Large sample Z-test:

$$z = \frac{|W - E_0(W)| - 0.5}{\sqrt{V_0(W)}}$$

- Approximate p-value:

$$2 \times P(Z > |z|)$$

Unit 2 ANOVA

Motivation

- Do the populations or treatment groups have the same mean values for the variable?
- Two sources of variation:
  - Variability among observations within each treatment group (or within each population)
  - Variability among mean responses for treatments (or between populations)
- Question:
  - Are differences among group means large relative to variation within groups?
  - Do all populations have the same mean?

Analysis of Variance (ANOVA)

- Calculate three variations based on observations Y_{ij}:
  - Variation due to group means
  - Variation due to residuals
  - Total variation
- These are called the **sums of squares (SS)**

Cell Means Model

Linear Model Form

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

- Each observation Y_{ij} can be described by two components:
  - Fixed mean value μ_i
  - Random error term ε_{ij}
- Gives an equation for each of the
$$N = \sum_{i=1}^{r} n_i$$
observations

Matrix Form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- The vector $\mathbf{Y}$ is length N and is the vector of observations.
- The matrix $\mathbf{X}$ is size $N \times r$ and is called the design matrix.
  It relates the observations to the parameters according to the model.
  It is fixed (non-random).
- The vector $\boldsymbol{\beta}$ is length r and is the vector of parameter values.
- The vector $\boldsymbol{\varepsilon}$ is length N and is the vector of random error terms.

Basic ANOVA

Variation due to Group Means

$$SS_{\text{among groups}} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{r} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

- Also called SS_{model}
- If the population means are the same (different), this value should be small (large)

Variation due to Residuals

$$SS_{\text{within groups}} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

$$= \sum_{i=1}^{r}(n_i - 1)S_i^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i} e_{ij}^2$$

- Also called SS_{error} or $SS_{\text{residuals}}$

Total Variation

$$SS_{\text{total}} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = SS_{\text{model}} + SS_{\text{error}}$$

ANOVA Table

Source of variation	Degrees of freedom	Sums of squares	Mean square	F
Model	$r-1$	SS_{model}	$MS_{\text{model}} = \frac{SS_{\text{model}}}{r-1}$	$\frac{MS_{\text{model}}}{MS_{\text{error}}}$
Error	$N-r$	SS_{error}	$MS_{\text{error}} = \frac{SS_{\text{error}}}{N-r}$	
Total	$N-1$	SS_{total}		

Note:

$$MS_{\text{error}} = S_p^2$$

‘

Model Assumptions

- Assumptions on random error terms:
  - ε_{ij} are i.i.d. from a normal distribution with mean 0 and variance σ^2
  - $\boldsymbol{\varepsilon}$ is multivariate normal with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$
- This implies that:
  - Y_{ij} are i.i.d. from a normal distribution with mean μ_i and variance σ^2
  - $\mathbf{Y}$ is multivariate normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance $\sigma^2\mathbf{I}$
- In addition, we assume groups are independent of each other

ANOVA F-test

- Null hypothesis:
$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$$
- Alternative hypothesis:
$$H_a : \text{at least one } \mu_i \text{ is different for } i = 1, \ldots, r$$
- Test statistic:
$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$
- P-value:
$$P(F_{r-1,\, N-r} > F)$$

Effects Model

Linear Effects Model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- Each observation Y_{ij} can be described by two components:
  - **Fixed mean value:** $\mu_i = \mu + \alpha_i$
    - Overall mean value: μ
    - Treatment effects compared with overall mean: α_i
    - Goal: find which α_i's are different from 0
  - **Random error term:** ε_{ij}

Identifiability Issues

- Model has too many parameters: estimates r means with $r + 1$ parameters
- Design matrix $\mathbf{X}$ is not full column rank
- The usual inverse $(\mathbf{X}^\top \mathbf{X})^{-1}$ does not exist
- There are infinitely many least squares estimators

Solution: impose constraints on the parameters - Set $\alpha_r = 0$ (baseline constraint), or - Set

$$\sum_{i=1}^{r} \alpha_i = 0$$

(sum-to-zero constraint)

Least Squares Estimator of $\boldsymbol{\beta}$

When

$$\sum_{i=1}^{r} \alpha_i = 0,$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} \frac{1}{r} \sum_{i=1}^{r} \bar{Y}_{i\cdot} \\ \bar{Y}_{1\cdot} - \frac{1}{r} \sum_{i=1}^{r} \bar{Y}_{i\cdot} \\ \bar{Y}_{2\cdot} - \frac{1}{r} \sum_{i=1}^{r} \bar{Y}_{i\cdot} \\ \vdots \\ \bar{Y}_{(r-1)\cdot} - \frac{1}{r} \sum_{i=1}^{r} \bar{Y}_{i\cdot} \end{pmatrix} = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{r-1} \end{pmatrix}.$$

Cautions

- The above two types of constraints are not the only ways to model the means
- The choice of constraint affects the least squares estimator $\hat{\boldsymbol{\beta}}$
- You must determine which constraint was applied before interpreting parameter estimates
- The interpretation of parameters (elements of $\boldsymbol{\beta}$) depends on the parametrization

Fixed vs. Random Effects

Fixed Effects

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- The r treatments (or groups) examined in the study are the only ones under consideration
- Research questions concern treatment means or differences in means
  - e.g., two drugs, four pesticides

Random Effects

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- The r treatments (or groups) are a random sample from a larger population of possible treatments (or groups)
- Research questions concern variability among sets of treatments (or groups) that could be selected for different studies
- Additional assumptions:
  $$\alpha_i \sim N(0, \sigma_\alpha^2),$$
  and α_i is independent of ε_{ij}

ANOVA Diagnostics and Remedies

ANOVA Assumptions

- ε_{ij} are i.i.d. $N(0, \sigma^2)$
- Independence of groups and observations
- Homogeneous (equal) variance:
  $$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2$$
- Normal distribution:
  - Random error terms are normally distributed

Model Diagnostics

- Many results from two-sample model diagnostics apply:
  - Independence: critical aspect
  - Equal variances: important
  - Normality: only a concern for small sample sizes or very skewed distributions
  - Outliers: results not robust
- Use residuals to assess model assumptions:
  $$e_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$$

Independence Assumption

Data collection: - Random sample(s) from multiple populations - Observations from multiple independent groups

- Study designed to produce independent responses

Equal Variance Assumption (Graphical Checks)

- Construct histograms of residuals for each group
- Construct boxplots of residuals for each group
- Plot residuals versus predicted values (there should be no trend)
  - Beware of interpretation if n_i's are very unequal
  - Expect larger range of e_{ij} if n_i is larger
- Study ratio of sample standard deviations:

$$\frac{\max\{S_i\}}{\min\{S_i\}}$$

Equal Variance Assumption (Formal Tests)

- Tests for equality of variances:
  - Brown–Forsythe test
  - Levene's test
  - etc.
- Consequences of unequal variances on the F-test:
  - Minor if sample sizes are the same
  - Large distortion of α level if sample sizes are very unequal
  - Decreased power

Normality Assumption

- Histogram of residuals
- Normal probability plot of residuals
- Numerical summaries:
  - Skewness
  - Kurtosis
- Tests for normality:
  - Shapiro–Wilk
  - Kolmogorov–Smirnov
  - Cramér–von Mises
  - Anderson–Darling

Non-Parametric

Kruskal–Wallis Test

- Combine the data into a single data set
- Order the N observations from smallest to largest
- Assign ranks R_{ij}:
  - Smallest observation gets rank 1, second smallest gets rank 2, etc.

  - For tied observations, average the ranks
- Calculate $\bar{R}_{i\cdot}$ = mean rank of observations in group i
- Test statistic:
$$H = (N-1)\frac{\sum_{i=1}^{r} n_i(\bar{R}_{i\cdot} - \bar{R})^2}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(R_{ij} - \bar{R})^2}$$
where
$$\bar{R} = \frac{N+1}{2}$$
- If H_0 is true, H has an approximate χ^2 distribution with $r-1$ degrees of freedom
- Approximation is best when $n_i \geq 5$ for all i
- p-value:
$$P(\chi^2_{r-1} > H)$$

ANOVA Contrasts

Motivation

- Inference for a single population mean
- Linear combinations of means, including contrasts
- Pairwise comparisons

Inference for Single Population Mean

- $100(1-\alpha)\%$ confidence interval for a single group mean:
$$\bar{Y}_{i\cdot} \pm t_{N-r,\,1-\alpha/2}\sqrt{\frac{MS_{\text{error}}}{n_i}}$$
- Notes:
  - MS_{error} is the estimate of the population variance σ^2
  - Degrees of freedom for the t distribution: $N-r$
  - Valid for inference on a *single* population mean (not used for comparison between means)

Contrast

- A **contrast** is a linear combination of the population means with $\sum_{i=1}^{r} c_i = 0$:
$$\gamma = \sum_{i=1}^{r} c_i \mu_i$$

Orthogonal Contrasts

- Two contrasts
$$\gamma_1 = \sum_i c_i \mu_i, \qquad \gamma_2 = \sum_i b_i \mu_i$$
are **orthogonal** if
$$\sum_i \frac{b_i c_i}{n_i} = 0$$

- If γ_1 and γ_2 are orthogonal:
  - They represent statistically unrelated pieces of information
  - One contrast conveys no information about the other
  - Estimates $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are uncorrelated
  - Hypothesis tests for γ_1 and γ_2 are independent (i.e., results of one test do not affect results of the other)
  - Confidence intervals for γ_1 and γ_2 are independent

Why Are Orthogonal Contrasts Useful?

- The F-test from the ANOVA table:
  - Tests whether all groups have the same mean
  - We do not always care about the omnibus F-test
- Contrasts:
  - Focus attention on specific scientific questions
  - Require the researcher to explicitly specify those questions
- Orthogonality implies:
  - Independence of test results
  - Tests for contrasts can be interpreted individually
  - A natural partitioning of sums of squares into *"interesting"* components and *"everything else"*

ANOVA Multiple Comparisons

Pairwise Comparisons

Each pairwise comparison has Type I error level α, or confidence level $100(1-\alpha)\%$.

When there are r groups, we perform
$$\binom{r}{2}$$
pairwise comparisons.

If r is large, some significant differences are expected by chance even if all of the population means are the same.

Comparison-wise Type I Error Rate

The comparison-wise Type I error rate is defined as
$$P\big(\text{reject } H_0 \text{ for one test} \mid H_0 \text{ is true for that test}\big).$$

Experiment-wise Type I Error Rate

The experiment-wise Type I error rate is defined as
$$P\big(\text{reject at least one of the } H_0\text{'s} \mid \text{all } H_0\text{'s are true}\big).$$

Multiple Comparisons Adjustment

Multiple comparisons adjustments are used to avoid too many false significant findings. The goal is to make the experiment-wise Type I error rate reasonably small.

These adjustments are equivalent to constructing simultaneous confidence intervals; that is, all confidence intervals in a set include their individual targets with a specified probability.

Basic Approach

Adjust the critical value $t_{N-r,1-\alpha/2}$ used in individual $100(1-\alpha)\%$ confidence intervals or individual α-level t-tests.

The cost of this approach is lower power, meaning it is less likely to detect a non-zero effect.

The benefit is that the experiment-wise Type I error rate is no larger than the specified α.

Least Significant Difference (LSD)

(Note: This is a Comparison-wise adjustment)

First conduct the overall F-test of

$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_r$$

at the α level.

If H_0 is not rejected, then declare all means the same. In this case, the chance of any false declaration of a significant difference is less than α.

If H_0 is rejected, then calculate confidence intervals or conduct hypothesis tests for pairwise comparisons.

This method is commonly used, but there can be substantial loss of power when only a few groups have different means.

(Note: What follows are examples of Experiment-wise adjustments)

Scheffé's Method

Scheffé's method works for any number of linear contrasts, including all possible linear contrasts.

It is the most conservative multiple comparison procedure, but it is relatively easy to apply.

In place of $t_{N-r,1-\alpha/2}$, use

$$\sqrt{(r-1)F_{r-1,N-r,1-\alpha}}.$$

Declare a significant difference between groups i and j if

$$\left|\bar{Y}_i - \bar{Y}_j\right| \geq \sqrt{(r-1)F_{r-1,N-r,1-\alpha}}\sqrt{MS_{\text{error}}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}.$$

Tukey–Kramer Honest Significant Difference (HSD)

The Tukey–Kramer procedure is based on the distribution of the studentized range.

The studentized range statistic is

$$q_{(r,N-r)} = \frac{\max_i \bar{Y}_i - \min_i \bar{Y}_i}{S_p/\sqrt{n}}.$$

For confidence intervals, use the critical value

$$\frac{1}{\sqrt{2}} q_{(r,N-r,1-\alpha)}.$$

For hypothesis tests, declare a significant difference if

$$\left|\bar{Y}_i - \bar{Y}_j\right| \geq \frac{1}{\sqrt{2}} q_{(r,N-r,1-\alpha)} \sqrt{MS_{\text{error}} \left(\frac{1}{n} + \frac{1}{n}\right)}.$$

Bonferroni Method

If we conduct m tests (or confidence intervals), replace α with α/m for each test (or confidence interval).

This method is easy to implement.

Declare a significant difference if

$$\left|\bar{Y}_i - \bar{Y}_j\right| \geq t_{N-r,1-\alpha/(2m)} \sqrt{MS_{\text{error}} \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

The Bonferroni method is conservative, especially when m is large and the tests are not independent, resulting in an experiment-wise Type I error rate less than α.

The number of comparisons m must be specified in advance.