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.