STA 106 Notes - M. Pouokam
Dylan M Ang
April 11, 2022

Contents

1 Terminology

Subject: A person, place or thing from which we measure data.
Population: The collection of all subjects of interest.
Sample: A subset of the population, from which we collect data.
Response/Dependent Variable: The variable which is of primary interest.
Explanatory/Independent Variable: The variable which we believe helps explain some of the variance in the response variable.
For this class, the response variable is numerical, and the explanatory variable/s are categorical. This is often phrased as "how does this numerical variable differ by group?"
Random Variable: A variable whose outcome is random. These are typically denoted by capital letters.

2 Notation

Let Y = the random variable denoting all possible values of the response variable.
Let y = an observed value of Y (in other words, measured observations).
Let Y_{ij} = the rv denoting all possible values of the jth observation in group i.
Let y_{ij} = the ith observed value of the jth group in Y.
As an example, let $i = 1$ denote sex M, and $i = 2$ denote sex F.

- Y_{13} = All possible values of height for the 3rd male. The outcome is random and unknown.
- y_{13} = 72 inches. The specific observed value of the 3rd male once measured.

μ_i = The population mean for group i (a single value).
$\bar{Y}_i$ = All possible values of the sample mean for group i.
$\bar{y}_i$ = A specific, observed value of $\bar{Y}_i$. In other words, the mean of one given sample.
σ_i = The population standard deviation for group i.
S_i = All possible values of the sample standard deviation.
s_i = A specific, observed value of S_i. In other words, the standard deviation of one given sample.
The book does not make this distinction $\implies Y = y$ and $\bar{Y} = \bar{y}$

Parameter: The unknown population value of some statistic. For example μ_i, σ_i. These values are constant (non-random) if we could measure the population we would know the true value.
The goal of statistics 106 is to estimate parameters with sample values, and use the assumed distribution of those sample values to form Hypothesie Tests (HTs) and Confidence Intervals (CIs).

3 Mean and Variance of RVs

Let Y_i be drawn from a distribution with population mean μ_Y and population standard deviation σ_Y.
Let the mean of $Y_i = \mu_{Y_i} = E\{Y_i\} = \mu_Y$.
Let the standard deviation of $Y_i = \sigma_{Y_i} = \sigma\{Y_i\} = \sigma_Y$.

3.1 Linear Combinations of RVs

Let Y^* be a linear combination of Y where $a, b \in \mathbb{R}$, then

Combination	Mean	Variance
$Y^* = a + Y$	$\mu_{Y^*} = a + \mu_Y$	$\sigma^2_{Y^*} = \sigma^2_Y$
$Y^* = bY$	$\mu_{Y^*} = b\mu_Y$	$\sigma^2_{Y^*} = b^2\sigma^2_Y$
$Y^* = a + bY$	$\mu_{Y^*} = a + b\mu_Y$	$\sigma^2_{Y^*} = b^2\sigma^2_Y$

3.2 Summation Identities

Let $Y_1, Y_2, \ldots, Y_n$ be RVs

$$\begin{aligned} E\left\{\sum_{i=1}^{n} Y_i\right\} &= E\{Y_1 + Y_2 + \cdots + Y_n\} \\ &= E\{Y_1\} + E\{Y_2\} + \ldots E\{Y_n\} \\ &= \sum_{i=1}^{n} E\{Y_i\} \end{aligned}$$

If $Y_1, Y_2, \ldots, Y_n$ are independent RVs,

$$\sigma^2\left\{\sum_{i=1}^{n} Y_i\right\} = \sum_{i=1}^{n} \sigma^2\{Y_i\}$$

Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, Y_i is independent with mean μ_Y and std.dev. σ_Y.

$$\begin{aligned}
E\left\{\bar{Y}\right\} &= E\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i\right\} = E\left\{\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)\right\} \\
&= \frac{1}{n}E\left\{Y_1 + Y_2 + \cdots + Y_n\right\} \\
&= \frac{1}{n}\Big(E\left\{Y_1\right\} + E\left\{Y_2\right\} + \cdots + E\left\{Y_n\right\}\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left\{Y_i\right\} = \frac{1}{n}\sum_{i=1}^{n}\mu_Y \\
&= \frac{1}{n}(\mu_Y + \mu_Y + \cdots + \mu_Y) = \frac{1}{n}(n * \mu_Y) \\
E\left\{\bar{Y}\right\} &= \mu_Y \\
\sigma^2\left\{\bar{Y}\right\} &= \sigma^2\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i\right\} = \left(\frac{1}{n}\right)^2\sigma^2\left\{\sum_{i=1}^{n} Y_i\right\} \\
&= \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\sigma^2\left\{Y_i\right\}
\end{aligned}$$

$$\begin{aligned}
\mu_{(\bar{Y}_1 - \bar{Y}_2)} &= \mu_1 - \mu_2 \\
\mu_{(\bar{Y}_1 + \bar{Y}_2)} &= \mu_1 + \mu_2 \\
\sigma^2\left\{(\bar{Y}_1 - \bar{Y}_2)\right\} &= \sigma^2\left\{1\right\} + \sigma^2\left\{2\right\} \\
\sigma^2\left\{(\bar{Y}_1 + \bar{Y}_2)\right\} &= \sigma^2\left\{1\right\} + \sigma^2\left\{2\right\}
\end{aligned}$$

4 Normal RVs and χ^2 RV

4.1 Normal RVs

A normal RV follows a bell curve created by a probability density function (pdf).
If Y is normally distributed with mean μ_Y and std dev σ_Y, we say that $Y \sim N(\mu_Y, \sigma_Y)$
$Y \sim N(\mu_Y, \sigma_Y) \implies Y^* = a + bY \sim N(a + b\mu_Y, b\sigma_Y)$
From this we can get two more results,
If $Y_1, \ldots, Y_n$ independent and $Y_i \sim N(\mu_Y, \sigma_Y)$, then

1. $\bar{Y} \sim N(\mu_Y, \sigma_Y/\sqrt{n})$
2. $\sum Y_i \sim N(n\mu_Y, \sqrt{n\sigma_Y})$

The standard normal distribution is a specific linear combination of a general normal distribution, denoted Z. Let $Y \sim N(\mu_Y, \sigma_Y)$

$$\begin{aligned}
Z &= \frac{Y - \mu_Y}{\sigma_Y} = \frac{Y}{\sigma_Y} - \frac{\mu_Y}{\sigma_Y} \\
E\left\{Z\right\} &= \frac{-\mu_Y}{\sigma_Y} + \mu_Y(\frac{1}{\sigma_Y}) = 0 \\
\sigma_Z^2 &= \left(\frac{1}{\sigma_Y}\right)^2\sigma_Y^2 = 1
\end{aligned}$$

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

4.2 χ^2 Distribution

The χ^2 distribution (chi-squared) is a sum of independent squared Z distributions.
Let $Z_1, Z_2, \ldots, Z_n$ be independent RVs where $Z_i \sim N(0, 1)$

$$\begin{aligned}
X &= Z_1^2 + \cdots + Z_n^2 \sim \chi_r^2 \text{ with degrees of freedom} \\
r &= \text{ The number of summed and squared } Z_i^2 \\
E\left\{\chi_r^2\right\} &= r
\end{aligned}$$

5 Hypothesis Testing and Confidence Intervals

5.1 Testing for difference in means

Step 1: Declare Hypothesis

$$\begin{aligned}
H_0 &: \mu_1 = \mu_2 \; or \; \mu_1 \leq \mu_2 \; or \; \mu_1 \geq \mu_2 \\
H_A &: \mu_1 \neq \mu_2 \; or \; \mu_1 > \mu_2 \; or \; \mu_1 < \mu_2
\end{aligned}$$

Step 2: Calculate test-statistic

$$t_s = \frac{(\bar{y_1} - \bar{y_2}) - \Delta_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

If equal variances are assumed, the following test statistic formula can be used.

$$\begin{aligned}
t_s &= \frac{(\bar{y_1} - \bar{y_2}) - \Delta_0}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t, df = n_1 + n_2 - 2 \\
s_p^2 &= \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}
\end{aligned}$$

t_s = The number of estimated standard deviations our sample difference in means is from the null.
Step 3: Calculate the p-value
If $H_A \implies$ p-value

$H_A \implies$	p
$\mu_1 \neq \mu_2$	$2P\left\{t > \lvert t_s\rvert\right\}$
$\mu_1 < \mu_2$	$P\left\{t < t_s\right\}$
$\mu_1 > \mu_2$	$P\left\{t > t_s\right\}$

p-value $= P\left\{\text{our data or more extreme}|H_0 \; TRUE\right\}$
p-value = probability of observing our sample data or more extreme, if in reality the null hypothesis were true.
Step 4: State decision rule and conclusion

$$\begin{aligned}
&If \; p - value < \alpha, reject \; H_0 \\
&If \; p - value \geq \alpha, fail \; to \; reject \; H_0
\end{aligned}$$

Recall that

$$\alpha = P\left\{\text{Type I Error}\right\} = P\left\{reject \; H_0|H_0 true\right\}$$

5.2 Confidence Interval for difference in means

The corresponding $(1 - \alpha)100\%$ CI for $(\mu_1 - \mu_2)$ is

$$(\bar{y_1} - \bar{y_2}) \pm t_{1-\alpha/2; n_1+n_2-2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$t_{1-\alpha/2;n_1+n_2-2}$ is the $(1-\alpha)100th$ percentile of a t distribution with $df = n_1 + n_2 - 2$

5.3 Assumptions

1. Random samples from both groups.
2. Groups are independent
3. $\sigma_1 = \sigma_2$ if using s_p formula.
4. $\bar{Y_1} - \bar{Y_2}$ is distributed normally, either because
   (a) Both populations are normal
   (b) n_1 and $n_2 \geq 30$ (Central limit theorem)

6 Experimental Design

6.1 Sampling

In ANOVA studies, the sampling scheme is very important. Typically, the categorical variable is seen as a treatment, and the goal is to see if it had an effect on the numerical variable.
In an experiment, subjects and randomly assigned a treatment, and the results are assessed to find a causal relationship between variables.
In an observational study, subjects are randomly sampled and may fall into natural treatment groups, but are not assigned one. The data is assessed to find correlations between variables.

6.2 Factors

Factors are the variables that experimenters control during an experiment in order to determine their effect on the response variable. A factor can take on only a small number of values, which are known as factor levels. Examples of factors are brand of equipment, where the factor levels are brand A, B, and C.
A treatment is a combination of factors that has been applied to a subject. Ex: A study with two factors - control vs drug group, and patient blood type.

bt/drug	A	B	AB	O
C	C,A	C,B	C, AB	C, O
D	D,A	D,B	D, AB	D, O

C,A and D,A are two possible treatments.

6.3 Crossed vs. Nested

When we have two factors, the design can be either crossed or nested.
A crossed design is where every possible treatment (combinations of factor levels) is present in the study.
A nested design is where not all possible treatments are present. For example, if we have 8 schools and two teaching methods, but not all schools teach both types.

	A	A	B	B	C	C	C	C
1	1,A	1,A	1,B	1,B				
2					2,C	2,C	2,C	2,C

Here, we would say that schools are nested within class format.

6.4 Blocking

Consider an experiment that is trying to determine if a new supplement increases vitamin C absorption.
Let Y response variable = vitamin C absorption
Let Factor A = group with levels "control" and "new".
There is a total variance in how subjects absorb vitamin C. If we can explain more of that variance, we are more likely to be able to tell if factor A had an effect.
Blocking is using another explanatory variable to further split the subjects. For example, perhaps gender affects how subjected absorb vitamin C. Then we could first block (separate) subjects by gender, then randomly assign them to factor A. This may reduce unexplained variance in Y.

6.5 ANOVA Designs

Most ANOVA models assume an underlying structure to the data,

$$Y = [overall\ constant] + [same\ things] \\ + [individual\ error]$$

For example, we may say that the height of a tree has some overall value which could be affected by the same things, and then also individual variance (error)
This is similar to a regression model
Depending on the design of the study, we use different models.
Completely Randomized Designs are where treatments are assigned to subjects randomly.
For example, say we assign a sample of trees randomly to 4 different fertilizers (A,B,C,D).
Then our model would be

$$height = [some\ constant] + [fertilizer\ effect] \\ + [ind\ error]$$

7 Single Factor ANOVA

Consider Y = numeric, and X = categorical with "a" categories total. The basic model we use is,

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \ldots, a, j = 1, \ldots, n_i$$

- Y_j = the jth value of Y for the ith group.
- μ_i = the unknown, true population mean for the ith group.
- ϵ_{ij} = the jth residual/error of Y for the ith group.

We assume

1. $Y_{ij}'s$ were randomly sampled (independent).
2. The ith group is independent $(i = 1, \ldots, a)$
3. $\epsilon_{ij} \sim N(o, \sigma_\epsilon^2)$ (errors are independent and normally distributed with mean 0, constant variance.)

Notice that Y_{ij} is a linear combination of the RV ϵ_{ij}, so

$$\begin{aligned} E\{Y_{ij}\} &= E\{\mu_i + \epsilon_{ij}\} \\ &= E\{\mu_i\} + E\{\epsilon_{ij}\} \\ &= \mu_i + 0 = \mu_i \end{aligned}$$

$$\begin{aligned}\sigma^2\{Y_{ij}\} &= \sigma^2\{\mu_i + \epsilon_{ij}\}\\ &= \sigma^2\{\mu_i\} + \sigma^2\{\epsilon_{ij}\}\\ &= 0 + \sigma^2_\epsilon = \sigma^2_\epsilon\end{aligned}$$

If we assume ϵ_{ij} is normally distributed $\implies$ Y_{ij} is normally distributed. Therefore,

$$Y_{ij} \sim N(\mu_i, \sigma^2_\epsilon)$$

7.1 Estimating μ_i and Notation

Estimate population mean with sample mean. Get μ_i from $\bar{y}_i$.
First, some notation. Let

1. $Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}$ = total for all n_i observations in group i.
2. $Y_{\bullet\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}$ = total for entire sample regardless of group.
3. $n_T = \sum_{i=1}^{a} n_i$ = overall sample size regardless of group.
4. $\bar{Y}_{i\bullet} = Y_{i\bullet}/n_i$ = sample mean for all observations in group i.
5. $\bar{Y}_{\bullet\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}/\sum_{i=1}^{a} n_i = \sum_{i=1}^{a}\bar{Y}_{i\bullet}n_i/n_T$ = overall sample mean.

Is $\bar{Y}_{i\bullet}$ a good estimator for μ?
Consider,

$$\begin{aligned}Q &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}\epsilon^2_{ij} = \text{ Sum of Squared Errors}\\ &= \sum_i\sum_j (Y_{ij}-\mu_i)^2\\ &= \sum_j (Y_{ij}-\mu_1)^2 + \sum_j (Y_{ij}-\mu_2)^2 + \cdots \sum_j (Y_{ij}-\mu_a)^2\end{aligned}$$

$$\begin{aligned}\frac{dQ}{d\mu_i} &= \frac{d}{d\mu_i}\left\{\sum_j (Y_{ij}-\mu_1)^2 + \cdots + \sum_j (Y_{ij}-\mu_i)^2 \ldots\right\}\\ &= \frac{d}{d\mu_i}\left\{\sum_j (Y_{ij}-\mu_i)^2\right\}\\ &= 2\sum_j (Y_{ij}-\mu_i)(-1) = -2\sum_j (Y_{ij}-\mu_i)\end{aligned}$$

then expand the sum and set to “0”.

$$\begin{aligned}-2\sum_j (Y_{ij}-\mu_i) = 0 \implies & \sum_j (Y_{ij}-\mu_i) = 0\\ &= \sum_j Y_{ij} - \sum_j \mu_i = 0\\ &= \sum_j \mu_i = \sum_j Y_{ij}\\ &= n_i\mu_i = \sum_j Y_{ij}\\ &= \hat{\mu_i} = \sum_j \frac{Y_{ij}}{n_i}\\ &= \bar{Y}_{i\bullet}\end{aligned}$$

Thus, $\hat{\mu_i} = \bar{Y_{i\bullet}}$ is the estimator of μ_i that minimizes the sum of squared errors.

Mean and Variance of $\hat{\mu_i}$

$$\begin{aligned}E\{\hat{\mu_i}\} &= E\left\{\sum_j Y_{ij}/n_i\right\} = \frac{1}{n_i}E\left\{\sum_j Y_{ij}\right\}\\ &= \frac{1}{n_i}\sum_j E\{Y_{ij}\} = \frac{1}{n_i}\sum_j \mu_i\\ &= \frac{1}{n_i}(n_i\mu_i) = \mu_i\\ \implies E\{\hat{\mu_i}\} &= \mu_i\end{aligned}$$

$$\begin{aligned}\sigma^2\{\hat{\mu_i}\} &= \sigma^2\left\{\sum_j Y_{ij}/n_i\right\} = (\frac{1}{n_i}^2)\sigma^2\left\{\sum_j Y_{ij}\right\}\\ &= \frac{1}{n_i}^2\sum_j \sigma^2\{Y_{ij}\} = \frac{1}{n_i^2}n_i\sigma^2\{\epsilon\}\\ \implies \sigma^2\{\hat{\mu_i}\} &= \frac{\sigma^2_\epsilon}{n_i}\end{aligned}$$

Since $\hat{\mu_i} = \sum_j Y_{ij}/n_i$ is a linear combination of normal RVs, it is normally distributed.
Thus,

$$\bar{Y}_{i\bullet} = \hat{\mu_i} \sim N(\mu_i, \sqrt{\sigma^2_\epsilon/n_i})$$

7.2 Residuals/Errors

The errors for a model are the actual values minus the estimated values, so

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

The estimated errors (residuals) are

$$e_{ij} = y_{ij} - \hat{\mu_i}\ (or\ \hat{\epsilon_{ij}} = y_{ij} - \hat{\mu_i})$$

7.3 Total Variance Partitioning

The total or overall variance of a data set is widely accepted to be the sum of squared distance from the mean. It is often denoted SSTO and defined as

$$SSTO = \sum_i\sum_j (Y_{ij} - \bar{Y}_{\bullet\bullet})^2$$

We can decompose $(Y_{ij} - \bar{Y}_{\bullet\bullet})$

$$\begin{aligned}(Y_{ij} - \bar{Y}_{\bullet\bullet}) &= Y_{ij} - \bar{Y}_{\bullet\bullet} + \bar{Y}_{i\bullet} - \bar{Y}_{i\bullet}\\ &= (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) + (Y_{ij} - \bar{Y}_{i\bullet})\\ &= (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) + e_{ij}\end{aligned}$$

Square both sides and expand,

$$(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + (e_{ij}^2) + 2(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})e_{ij}$$

$$\sum_i \sum_j (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_i \sum_j (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_i \sum_j (e_{ij}^2) + 0$$

In other words, the total variance can be partitioned into 2 parts:

1. $SSA = \sum_i \sum_j (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_i n_i (\bar{Y}_{i\bullet - \bar{Y}_{\bullet\bullet}})^2 =$ Factor A Sum of Squares.
2. $SSE = \sum_i \sum_j e_{ij}^2 =$ Sum of Squares Error.

In general, we want SSE to be small (low error) so then SSA is large. This means that much of the variance in Y is due to a difference in the overall mean vs. group mean, and not due to error. This is why it is called analysis of variance.

7.4 SS Properties

Notice SSTO, SSA, and SSE are sums of squared values, so they keep growing larger as i or j increase. They never converge.
Due to this, we stabilize or standardize the SS values to obtain "Mean Square" values. The value we stabilize them with is the df (degrees of freedom).

To find $df\{SSTO\} = MSTO$
$df\{SSTO\} =$ number of observations that can freely vary, subject to our constraint.

Our constraint

$$\implies \sum_i \sum_j (Y_{ij} - \bar{Y}_{\bullet\bullet}) = 0$$
$$\implies \bar{Y}_{\bullet\bullet} = \sum_i \sum_j Y_{ij}/n_T$$

So, we may allow $n_T - 1$ values of Y_{ij} to vary freely, but the last one cannot.
So, $df\{SSTO\} = n_T - 1$. Thus,

$$MSTO = SSTO/df\{SSTO\} = \frac{SSTO}{n_T - 1}$$

To find $df\{SSA\} = MSA$

$$SSA = \sum_i n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$$

We have "a" values of $\bar{Y}_{i\bullet}$ that we are summing over, and the constraint is $\sum_i n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) = 0$
Thus, $df\{SSA\} = a - 1$

$$MSA = SSA/df\{SSA\}$$
$$E\{MSA\} = \sigma_\epsilon^2 + \left(\sum_i n_i(\mu_i - \mu)^2\right)/(a-1)$$

To find $df\{SSE\} = MSE$
$SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\bullet})^2$ which uses n_T values of Y_{ij} and has "a" constraints: $\sum_j (Y_{ij} - \bar{Y}_{i\bullet}) = 0$
Thus $df\{SSE\} = n_T - a$

$$MSE = SSE/df\{SSE\}$$
$$E\{MSE\} = \sigma_\epsilon^2$$

7.5 F test for equal means

The first main question of single factor ANOVA is "is the statistical evidence to suggest the group means are equal for all groups?"
To answer that question, we will compare MSE to MSA, and use the fact that $E\{MSE\} = E\{MSA\}$ if and only if $\mu_i = \mu$ for all i (i.e. the group means equal the overall mean).

Step 1: Declare Hypotheses
$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$
H_A : At least one $\mu_i \neq$ to another.
Assume H_0 is true.

Step 2: Find test-statistic
Let our test-statistic be

$$F_s = \frac{MSA}{MSE}$$

Notice if H_0 were exactly true, $F^* = 1$
We can show that, if H_0 is true, F_s is the ratio of two independent χ^2 variables, and is F distributed with $df\{num\} = a - 1, df\{denom\} = n_T - a$

Step 3: Calculate p-value

$$p = P\{F > F_s\} \text{ where}$$
$$df\{num\} = a - 1, df\{denom\} = n_T - a$$

Step 4: Decision Rule and Conclusion
If p-value $< \alpha$, reject H_0
If p-value $\geq \alpha$, fail to reject H_0
Note: SSE can be written as :

$$SSE = \sum_i s_i^2(n_i - 1)$$

$s_i^2 =$ sample variance for the ith group