Categorical Independent Variables

And multiple comparisons

One-way Analysis of variance

- Categorical IV
- Quantitative DV
- *p* categories (groups)
- H_0: All population means equal
- Normal conditional distributions
- Equal variances

Analysis means to split up

- With no IV, best predictor is the overall mean
- Variation to be explained is SSTO, sum of squared differences from the overall mean
- With an IV, best predictor is the group mean
- Variation still unexplained is SSW, sum of squared differences from the group means

$$SSTO = SSB + SSW$$

$$
\begin{aligned}
SSB &= \sum_{j=1}^{p} n_j(\overline{Y}_j - \overline{Y})^2 \\
SSW &= \sum_{j=1}^{p}\sum_{i=1}^{n_j}(Y_{i,j} - \overline{Y}_j)^2 \\
SSTO &= \sum_{j=1}^{p}\sum_{i=1}^{n_j}(Y_{i,j} - \overline{Y})^2.
\end{aligned}
$$

ANOVA Summary Table

Source	DF	Sum of Squares	Mean Square	F Value	Pr $>$ F
Model	$p-1$	SSB	$MSB = SSB/(p-1)$	MSB/MSW	p-value
Error	$n-p$	SSW	$MSW = SSW/(n-p)$		
Corrected Total	$n-1$	$SSTO$			

$$H_0 : \mu_1 = \ldots = \mu_p$$

R^2 is the proportion of variation explained by the independent variable

$$R^2 = \frac{SSB}{SSTO}$$

Contrasts

$$c = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$$

$$\widehat{c} = a_1\overline{Y}_1 + a_2\overline{Y}_2 + \cdots + a_p\overline{Y}_p$$

$$\text{where } a_1 + a_2 + \cdots + a_p = 0$$

Overall F-test is a test of p-1 contrasts

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

a_1	a_2	a_3	a_4
1	-1	0	0
0	1	-1	0
0	0	1	-1

$$c = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$$

Multiple Comparisons

- Most hypothesis tests are designed to be carried out in isolation
- But if you do a lot of tests and all the null hypotheses are true, the chance of rejecting at least one of them can be a lot more than α. This is **inflation of the Type I error rate**.
- Multiple comparisons (sometimes called follow-up tests, post hoc tests, probing) offer a solution.

Multiple comparisons

- Protect a *family* of tests against Type I error at some *joint* significance level α
- If all the null hypotheses are true, the probability of rejecting at least one is no more than α

Multiple comparisons of contrasts in a one-way design: Assume all means are equal in the population

- Bonferroni
- Tukey
- Scheffé

Bonferroni

- Based on Bonferroni's inequality

$$Pr\left\{\cup_{j=1}^{k} A_j\right\} \leq \sum_{j=1}^{k} Pr\{A_j\}$$

- Applies to *any* collection of k tests
- Assume all k null hypotheses are true
- Event A_j is that null hypothesis j is rejected.
- Do the tests as usual
- Reject each H_0 if $p < 0.05/k$
- Or, adjust the p-values. Multiply them by k, and reject if $pk < 0.05$

Bonferroni

- Advantage: Flexibility
- Advantage: Easy to do

- Disadvantage: Must know what all the tests are before seeing the data
- Disadvantage: A little conservative; the true joint significance level is *less* than α.

Tukey (HSD)

- Based on the distribution of the largest mean minus the smallest.
- Applies only to pairwise comparisons of means
- If sample sizes are equal, it’s most powerful, period
- If sample sizes are not equal, it’s a bit conservative

Scheffé

- Find the usual critical value for the initial test. Multiply by p-1. This is the Scheffé critical value.
- Family includes *all* contrasts: Infinitely many!
- You don't need to specify them in advance
- Based on the union-intersection principle – more details later, after F-tests.

Scheffé

- Follow-up tests *cannot* be significant if the initial overall test is not. Not quite true of Bonferroni and Tukey.
- If the initial test (of p-1 contrasts) is significant, there is a single contrast that is significant (not necessarily a pairwise comparison)
- Adjusted p-value is the tail area beyond F times (p-1)

Which method should you use?

- If the sample sizes are nearly equal and you are only interested in pairwise comparisons, use Tukey because it's most powerful

- If the sample sizes are not close to equal and you are only interested in pairwise comparisons, there is (amazingly) no harm in applying all three methods and picking the one that gives you the greatest number of significant results. (It’s okay because this choice could be determined in advance based on number of treatments, α and the sample sizes.)

- If you are interested in contrasts that go beyond pairwise comparisons and you can specify *all* of them before seeing the data, Bonferroni is almost always more powerful than Scheffé. (Tukey is out.)

- If you want lots of special contrasts but you don't know exactly what they all are, Scheffé is the only honest way to go, unless you have a separate replication data set.

Dummy Variables

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \epsilon_i$$

- X=1 means Drug, X=0 means Placebo

- Population mean is $$E[Y|X=x] = \beta_0 + \beta_1 x$$

- For patients getting the drug, mean response is
$$E[Y|X=1] = \beta_0 + \beta_1$$
- For patients getting the placebo, mean response is
$$E[Y|X=0] = \beta_0$$

Regression test of H_0: $\beta_1=0$

- Same as an independent t-test
- Same as a oneway ANOVA with 2 categories
- Same t, same F, same p-value.

Drug A, Drug B, Placebo

- x_1 = 1 if Drug A, Zero otherwise
- x_2 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

Regression coefficients are contrasts with the category that has no indicator - The **reference category**.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \;\Leftrightarrow\; \beta_1 = \beta_2 = 0$$

Indicator dummy variable coding with intercept

- Need p-1 indicators to represent a categorical IV with p categories
- If you use p dummy variables, trouble
- Regression coefficients are **contrasts** with the category that has no indicator
- Call this the **reference category**

Now add a quantitative variable (covariate)

- x_1 = Age
- x_2 = 1 if Drug A, Zero otherwise
- x_3 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Drug	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$(\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\beta_0 \quad + \beta_1 x_1$

Parallel slopes, ANCOVA

What do you report?

- x_1 = Age
- x_2 = 1 if Drug A, Zero otherwise
- x_3 = 1 if Drug B, Zero otherwise
- $\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$

Drug	x_2	x_3	$b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$
A	1	0	$(b_0 + b_2) + b_1 x_1$
B	0	1	$(b_0 + b_3) + b_1 x_1$
Placebo	0	0	$b_0 \quad + b_1 x_1$

Set all covariates to their sample mean values

- And compute Y-hat for each group
- Call it an "adjusted" mean, or something like "average university GPA adjusted for High School GPA."
- SAS calls it a **least squares mean** (`lsmeans`)

Test whether the average response to Drug A and Drug B is different from response to the placebo, controlling for age. What is the null hypothesis?

Drug	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$(\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\beta_0 \quad + \beta_1 x_1$

$$H_0 : \beta_2 + \beta_3 = 0$$

Show your work

$$
\begin{aligned}
& \tfrac{1}{2}\left[\,(\beta_0 + \beta_2 + \beta_1 x_1) + (\beta_0 + \beta_3 + \beta_1 x_1)\,\right] = \beta_0 + \beta_1 x_1 \\
\iff \quad & \beta_0 + \beta_2 + \beta_1 x_1 + \beta_0 + \beta_3 + \beta_1 x_1 = 2\beta_0 + 2\beta_1 x_1 \\
\iff \quad & 2\beta_0 + \beta_2 + \beta_3 + 2\beta_1 x_1 = 2\beta_0 + 2\beta_1 x_1 \\
\iff \quad & \beta_2 + \beta_3 = 0
\end{aligned}
$$

We want to avoid this kind of thing. It can get complicated.

A common error

- Categorical IV with *p* categories
- *p* dummy variables (rather than *p-1*)
- And an intercept

- There are *p* population means represented by *p+1* regression coefficients - not unique

But suppose you leave off the intercept

- Now there are *p* regression coefficients and *p* population means
- The correspondence is unique, and the model can be handy -- less algebra
- Called **cell means coding**

Cell means coding: *p* indicators and no intercept

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Drug	x_1	x_2	x_3	$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	0	$\mu_1 = \beta_1$
B	0	1	0	$\mu_2 = \beta_2$
Placebo	0	0	1	$\mu_3 = \beta_3$

Add a covariate: x_4

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Drug	x_1	x_2	x_3	$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
A	1	0	0	$\beta_1 + \beta_4 x_4$
B	0	1	0	$\beta_2 + \beta_4 x_4$
Placebo	0	0	1	$\beta_3 + \beta_4 x_4$

Effect coding

- *p-1* dummy variables for *p* categories
- Include an intercept
- Last category gets -1 instead of zero
- What do the regression coefficients mean?

Group	x_1	x_2	$E[Y\|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

Meaning of the regression coefficients

Group	x_1	x_2	$E[Y\|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

$$\mu = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \beta_0$$

With effect coding

- Intercept is the *Grand Mean*
- Regression coefficients are deviations of group means from the grand mean
- Equal population means is equivalent to zero coefficients for all the dummy variables
- Last category is not a reference category

Group	x_1	x_2	$E[Y\|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

Sometimes speak of the "main effect" of a categorical variable

- More than one categorical IV (factor)
- Marginal means are average group mean, averaging across the other factors
- This is loose speech: There are actually p main effects for a variable, not one
- Blends the "effect" of an experimental variable with the technical statistical meaning of effect.
- It's harmless

Add a covariate: Age = x_1

Group	x_2	x_3	$E[Y\|\boldsymbol{X}=\boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = \beta_0 + \beta_2 \quad + \beta_1 x_1$
B	0	1	$\mu_2 = \beta_0 + \beta_3 \quad + \beta_1 x_1$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_2 - \beta_3 + \beta_1 x_1$

Regression coefficients are deviations from the average conditional population mean (conditional on x_1).

So if the regression coefficients for all the dummy variables equal zero, the categorical IV is unrelated to the DV, controlling for the covariates.

We will see later that effect coding is very useful when there is more than one categorical independent variable and we are interested in *interactions* --- ways in which the relationship of an independent variable with the dependent variable depends on the value of another independent variable.

What dummy variable coding scheme should you use?

- Whichever is most convenient
- They are all equivalent, if done correctly
- Same test statistics, same conclusions

Interactions

- Interaction between independent variables means “It depends.”
- Relationship between one IV and the DV *depends* on the value of another IV.
- Can have
  - Quantitative by quantitative
  - Quantitative by categorical
  - Categorical by categorical

Quantitative by Quantitative

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \\ E(Y|\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \end{aligned}$$

For fixed x_2

$$E(Y|\mathbf{x}) = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

Both slope and intercept depend on value of x_2

And for fixed x_1, slope and intercept relating x_2 to E(Y) depend on the value of x_1

Quantitative by Categorical

- Interaction means slopes are not parallel
- Form a product of quantitative variable by each dummy variable for the categorical variable
- For example, three treatments and one covariate: x_1 is the covariate and x_2, x_3 are dummy variables

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &\quad + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon \end{aligned}$$

General principle

- Interaction between A and B means
  - Relationship of A to Y depends on value of B
  - Relationship of B to Y depends on value of A
- The two statements are formally equivalent

$$E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
2	0	1	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
2	0	1	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

What null hypothesis would you test for

- Parallel slopes
- Compare slopes for group one vs three
- Compare slopes for group one vs two
- Equal regressions
- Interaction between group and x_1

What to do if H_0: $\beta_4=\beta_5=0$ is rejected

- How do you test Group "controlling" for x_1?
- A good choice is to set x_1 to its sample mean, and compare treatments at that point.

- How about setting x_1 to sample mean of the group (3 different values)?
- With random assignment to Group, all three means just estimate $E(X_1)$, and the mean of all the x_1 values is a better estimate.

Categorical by Categorical

- Soon
- But first, an example