PSY 201: Statistics in Psychology

Lecture 35

Analysis of Variance

Ignoring (some) variability.

Greg Francis

Purdue University

Fall 2023

ANOVA TESTING

- 4 STEPS
  1. State the hypothesis. : $H_0 : \mu_1 = \mu_2 = ... = \mu_K$, $H_a : \mu_i \neq \mu_j$ for some i, j.
  2. Set the criterion: α
  3. Compute the test statistic: $F = MS_B / MS_W$, degrees of freedom, and p-value
  4. Interpret results.

ASSUMPTIONS

- to use ANOVA for independent means validly, the data must meet some restrictions
  - The observations are **random** and **independent** samples from the populations.
  - The distributions of the populations from which samples are selected are **normal**.
  - The variances of the distributions in the populations are equal. **Homogeneity of variance**.

ASSUMPTIONS

- it turns out that
- independence of samples is critical
- violations of normality have small effects on Type I error rates
- violations of homogeneity of variance have a big effect if the population sizes are different
  - similar to the standard t test
- means that ANOVA is robust as long as the sample sizes are the same across populations

t tests

- if we have only two groups

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

- we can use either ANOVA or the (standard) t-test discussed previously
- they give identical results!

t tests

- it turns out that the F distribution for $K-1, N-K$ $(1, N-2)$ degrees of freedom is simply the t distribution for $N-2$ df, squared.

$$t^2 = F$$

- so using either technique produces the same results (reject or not reject)

EXAMPLE

- A sociologist wants to determine whether sorority or dormitory women date more often. He randomly samples 12 women who live in sororities and 12 women who live in dormitories and determines the number of dates they each have during the ensuing month. The following are the results.

Sorority Women, X_1	Dormitory Women, X_2
8	9
5	7
6	3
4	4
12	4
7	8
9	7
10	5
5	8
3	6
7	3
5	5
$\overline{X}_1$=6.750	$\overline{X}_2$ =5.750

t TEST

- test with $\alpha = 0.05$, two-tailed

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

- we have equal numbers of subjects, so we do not need to worry about homogeneity of variance
- from data we calculate the pooled estimate of population variance

$$s^2 = 5.570$$

t TEST

- so standard error is

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$s_{\overline{X}_1-\overline{X}_2} = 0.963$$

- and

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1-\overline{X}_2}}$$

$$t = \frac{1.0}{0.963} = 1.038$$

$$df = n_1 + n_2 - 2 = 12 + 12 - 2 = 22$$

- From the t-distribution calculator, we find

$$p = 0.3105 > 0.05 = \alpha$$

- so do not reject H_0
- no evidence for a difference in number of dates

ANOVA

- The same hypotheses work for an ANOVA

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

- we can calculate

$$SS_B = 6.00$$

$$SS_W = 122.500$$

ANOVA

$$MS_B = \frac{SS_B}{K-1} = \frac{6.00}{1} = 6.00$$

$$MS_W = \frac{SS_W}{N-K} = \frac{122.500}{22} = 5.568$$

$$F = \frac{MS_B}{MS_W} = \frac{6.00}{5.568} = 1.078$$

- we have 1 *df* in the numerator and 22 *df* in the denominator, and we use the F-distribution calculator to find

$$p = 0.31042 > 0.05 = \alpha$$

- we do not reject H_0
- note:

$$F = 1.078 \approx 1.077 = (1.038)^2 = t^2$$

DEPENDENT MEASURES

- one way ANOVA deals with independent samples
- we want to consider a situation where all samples are "connected"
- e.g., tracking health patterns for a common set of patients across years; grades for a common set of students throughout school
- Often called a *within subjects* ANOVA or a *repeated measures* ANOVA
- there can be other kinds of dependencies
- e.g., IQ of first-born, second-born, and third-born siblings

SUM OF SQUARES

- scores for an "individual" are dependent
- scores for different "individuals" are independent

$$SS_T = SS_I + SS_O + SS_{Res}$$

- where
  - SS_T is the total sum of square
  - SS_I is the variation among individuals
  - SS_O is the variation among test occasions
  - SS_{Res} is any other type of variation

INDIVIDUALS

- the combined variation among individuals is

$$SS_I = \sum_i K \left(\overline{X}_i - \overline{X}\right)^2$$

- where

$$\overline{X}_i = \frac{\sum_k X_{ik}}{K}$$

- is the average for the ith individual across all observations
- SS_I deviation of individual means from overall mean
- does not correspond to SS_W or SS_B in the normal ANOVA
- we want to *ignore* this variability

OBSERVATIONS

- the combined variation across observations is

$$SS_O = \sum_k n \left(\overline{X}_k - \overline{X} \right)^2$$

- where

$$\overline{X}_k = \frac{\sum_i X_{ik}}{n}$$

- is the average for the kth observation across all subjects
- SS_O deviation of observation mean from overall mean
- similar to SS_B in the independent ANOVA

RESIDUAL

- we need a term that corresponds to SS_W
- we can directly calculate the total sum of squares

$$SS_T = \sum_k \sum_i \left(X_{ik} - \overline{X}\right)^2$$

- if there is variation beyond SS_I and SS_O, we can calculate it as

$$SS_{Res} = SS_T - SS_I - SS_O$$

- this is similar to SS_W
  - factors out variation due to individuals and variation due to observations

VARIANCE ESTIMATES

- $SS_{Res} > 0$ due to random sampling (choice of individuals)

$$MS_{Res} = \frac{SS_{Res}}{(K-1)(n-1)}$$

- estimates the variance of the population distribution
- the degrees of freedom associated with this estimate is

$$(K-1)(n-1)$$

VARIANCE ESTIMATES

- SS_O can vary due to random sampling, or due to differences across observations
- if H_0 is true

$$H_0 : \mu_1 = \mu_2 = ... = \mu_K$$

- then there are no population differences across observations, so all variation must be due to random sampling. So,

$$MS_O = \frac{SS_O}{K - 1}$$

- estimates the variance of the population distribution **if H_0 is true**
  - otherwise it overestimates it

F RATIO

- as before we compare these estimates with the F statistic

$$F = \frac{MS_O}{MS_{Res}}$$

- if H_0 is true

$$F \approx 1.0$$

- if H_0 is not true

$$F > 1.0$$

- look up p value using $(K-1)$ and $(K-1)(n-1)$ degrees of freedom
- everything else is the same as before

EXAMPLE

- A school principal traces reading comprehension scores on a standardized test for a random sample of dyslexic students across three years. The data are given below. Complete the ANOVA using $\alpha = 0.05$.

Student	Third Grade	Fourth Grade	Fifth Grade
1	2.8	3.2	4.5
2	2.6	4.0	5.1
3	3.1	4.3	5.0
4	3.8	4.9	5.7
5	2.5	3.1	4.4
6	2.4	3.1	3.9
7	3.2	3.8	4.3
8	3.0	3.6	4.4

(1) HYPOTHESIS

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

$$H_a : \mu_i \neq \mu_k \text{for some } i \text{ and } k$$

- use $\alpha = 0.05$

(2) TEST STATISTIC

- It turns out that
$$SS_T = 18.66$$
and
$$SS_I = 5.67$$
and
$$SS_O = 12.0858$$
- so any remaining variation is residual
$$SS_{Res} = SS_T - SS_I - SS_O$$
$$SS_{Res} = 18.66 - 5.67 - 12.09 = 0.9075$$
- this cannot be negative!

(2) TEST STATISTIC

- now calculate

$$MS_O = \frac{SS_O}{K-1} = \frac{12.0858}{2} = 6.0429$$

- and

$$MS_{Res} = \frac{SS_{Res}}{(K-1)(n-1)} = \frac{0.9075}{14} = 0.0648$$

- and get the F statistic

$$F = \frac{MS_O}{MS_{Res}} = \frac{6.0429}{0.0648} = 93.22$$

(3) P-VALUE

- for the numerator (observation sum of squares) we have

$$df = K - 1 = 3 - 1 = 2$$

- for the denominator (residual sum of squares) we have

$$df = (K - 1)(n - 1) = (3 - 1)(8 - 1) = 14$$

- so from the F-distribution calculator, we find the $F = 93.22$ corresponds to

$$p \approx 0.000 < 0.05 = \alpha$$

(4) DECISION

- we reject H_0.
- there is evidence that the reading scores for these subjects are different across the years

CALCULATORS

- No one does these computations by hand. Computer programs do it for you. Your text provides a Dependent ANOVA One-Way calculator.
- You have to format the data correctly

5	ThirdGrade	2.5
6	ThirdGrade	2.4
7	ThirdGrade	3.2
8	ThirdGrade	3
1	FourthGrade	3.2
2	FourthGrade	4
3	FourthGrade	4.3
4	FourthGrade	4.9
5	FourthGrade	3.1
6	FourthGrade	3.1
7	FourthGrade	3.8
8	FourthGrade	3.6
1	FifthGrade	4.5
2	FifthGrade	5.1
3	FifthGrade	5
4	FifthGrade	5.7
5	FifthGrade	4.4
6	FifthGrade	3.9
7	FifthGrade	4.3
8	FifthGrade	4.4

formatted with one score
subject (e.g., name or ot
independent variable (le
(score). The variables m
comma, or a tab. The sc
for the subject and indep
contiguous text (no spac
for each level. For exam
happy for each of three s
PaulAtreides then your

subject1	level1	7
Greg	happy	6
subject1	happy	12
PaulAtreides	happy	8
PaulAtreides	level1	3
Greg	level1	9

Run One-Way Dependent ANOVA

Source	df	SS	MS	F	p-value
Individuals	7	5.6662	0.8095		
Occasions	2	12.0858	6.0429	93.2241	0.00000
Residual	14	0.9075	0.0648		
Total	23	18.6596			

CALCULATORS

- Extra information is important for interpreting the results
- means, correlations
  - Not always reported, but should be

Summary table

Condition	Mean	Standard deviation	Sample size
ThirdGrade	2.9250	0.4559	8
FourthGrade	3.7500	0.6392	8
FifthGrade	4.6625	0.5680	8

Correlation table

	ThirdGrade	FourthGrade	FifthGrade
ThirdGrade	1.0000	0.8383	0.6826
FourthGrade	0.8383	1.0000	0.8912
FifthGrade	0.6826	0.8912	1.0000

CONTRASTS

- We set up *contrast weights*, c_i, for each sample
- Our null hypothesis will be

$$H_0 : \sum_{i=1}^{K} (c_i \mu_i) = 0$$

- and we require that the contrast weights sum to 0:

$$\sum_{i=1}^{K} c_i = 0$$

- Our alternative hypothesis is

$$H_a : \sum_{i=1}^{K} (c_i \mu_i) \neq 0$$

- (one-tailed tests are also possible)

TEST STATISTIC

- We compute the weighted sum of means

$$L = \sum_{i=1}^{K} (c_i \overline{X}_i)$$

- which has a standard error of:

$$s_L = \sqrt{MS_{\text{Res}} \sum_{i=1}^{K} \frac{c_i^2}{n}}$$

- and our test statistic is

$$t = \frac{L}{s_L}$$

- which follows a t distribution with

$$df = (K-1)(n-1)$$

- where N is the sum of sample sizes across all groups and K is the number of groups

CALCULATORS

- A one-tailed contrast to compare scores in Third Grade against scores in Fourth Grade

Specify hypotheses:

H_0: -1 $\mu_{ThirdGrade}$ + 1 $\mu_{FourthGrade}$ + 0 $\mu_{FifthGrade} = 0$

H_a: Positive one-tail

α 0.05

Run Contrast

Contrast test summary	
Null hypothesis	H_0: $(-1)\mu_{ThirdGrade} + (1)\mu_{FourthGrade} + (0)\mu_{FifthGrade} = 0$
Alternative hypothesis	H_a: $(-1)\mu_{ThirdGrade} + (1)\mu_{FourthGrade} + (0)\mu_{FifthGrade} > 0$
Type I error rate	α=0.05
Weighted sum of sample means	L= 0.8250
Standard error	s_L =0.1273
Test statistic	t=6.4807
Degrees of freedom	df=14
p value	p=0.00001
Decision	Reject the null hypothesis

CALCULATORS

- A one-tailed contrast to compare scores in Fourth Grade against scores in Fifth Grade

Specify hypotheses:

H_0: 0 $\mu_{ThirdGrade}$ + -1 $\mu_{FourthGrade}$ + 1 $\mu_{FifthGrade} = 0$

H_a: Positive one-tail

α 0.05

Run Contrast

Contrast test summary

Null hypothesis	H_0: $(0)\mu_{ThirdGrade} + (-1)\mu_{FourthGrade} + (1)\mu_{FifthGrade} = 0$
Alternative hypothesis	H_a: $(0)\mu_{ThirdGrade} + (-1)\mu_{FourthGrade} + (1)\mu_{FifthGrade} > 0$
Type I error rate	α=0.05
Weighted sum of sample means	L= 0.9125
Standard error	s_L =0.1273
Test statistic	t=7.1681
Degrees of freedom	df=14
p value	p=0.00000
Decision	Reject the null hypothesis

ASSUMPTIONS

- ANOVA for dependent measures depends on four assumptions
  - The sample was randomly selected for a population.
  - The dependent variable (e.g., reading scores) is normally distributed in the population.
    - deviations tend to not cause serious problems
  - The population variances for the test occasions are equal. (homogeneity of variance)
    - Can be compensated for sometimes
  - The population correlation coefficients between pairs of test occasion scores are equal.
    - Can be compensated for sometimes

CONCLUSIONS

- assumptions of one-way independent ANOVA
- ANOVA for dependent measures
- contrasts for dependent ANOVA
- assumptions of dependent ANOVA

NEXT TIME

- power for dependent ANOVA

Leverage relationships.