

1. Suppose that you were interested in estimating the variance of a population. What statistic would you compute? What would you do if you had four independent samples from a population? What term in an independent groups ANOVA source table represents an analogous situation? [5 pts]

You would gather a sample and compute s^2 to estimate σ^2 . If you had four samples, it would make sense to use the mean of the four sample variances as an estimate of σ^2 . (Doing so might make less sense if the sample variances were quite different, i.e., heterogeneity of variance.) The analogous situation in the independent groups ANOVA is MS_{within} (MS_{Error}).

2. An independent groups design and a repeated measures design are both important tools in the psychologist's experimental arsenal. Distinguish between the two designs in terms of setting up experiments, power, situations in which one design or the other might be inappropriate, etc. [10 pts]

Independent Groups Design	Repeated Measures Design
less efficient (need more Ss to generate the same amount of data)	more efficient (need fewer Ss for same amount of data)
each S is exposed to only one treatment and produces only one piece of data	Ss are exposed to more than one treatment and produce more than one piece of data
no need to counterbalance, but run in randomized replications	need to counterbalance (complete or incomplete?)
less powerful (condition means differ due to treatment, individual differences, and random)	more powerful (removing individual differences)

3. Recently Simone Schnall and her colleagues published a paper in *Psychological Science*:

Schnall, S., Roper, J., & Fessler, D. M. T. (2010). Elevation leads to altruistic behavior.

From their abstract:

Feelings of elevation, elicited by witnessing another person perform a good deed, have been hypothesized to motivate a desire to help others. However, despite growing interest in the determinants of prosocial behavior, there is only limited evidence that elevation leads to increases in altruistic behavior...Feelings of elevation, but not feelings of amusement or happiness, predicted the amount of helping. Together, these results provide evidence that witnessing another person's altruistic behavior elicits elevation, a discrete emotion that, in turn, leads to tangible increases in altruism.

From their Procedure section:

Participants were informed that they were taking part in a 1-hr experiment on episodic memory in which they would watch a film clip, write about it, and complete a 30-min computer task. Tested individually, participants were randomly assigned to watch the elevation film clip from the Oprah Winfrey Show (elevation condition), the control film clip (the first 7 min of "The Open Ocean," David Attenborough's (1984) nature documentary describing a journey through the deepest part of the ocean), or a clip from a British comedy ("Fawlty Towers") intended to induce mirth (mirth condition).

The experimenter then feigned three unsuccessful attempts to open the computer file that ostensibly needed to be completed by the participant. She then told the participant that, because it was impossible to complete the next part of the study, the participant was free to leave, but would still receive the full hour's worth of course credit. Following the procedure outlined in Bartlett and DeSteno (2006), when the participant got up to leave, the experimenter asked, apparently as an afterthought, whether she would be willing to complete another questionnaire, ostensibly from another study for which the experimenter needed to establish norms. The experimenter noted that the questionnaire was, unfortunately, rather boring, emphasizing that the participant was under no obligation, and was free to stop whenever she wanted, but that completing any number of the items would greatly assist the

experimenter. If the participant agreed to help, she was seated at a desk, reminded that she was free to stop whenever she wished, and given 85 elementary math problems. The participant's work on the problems was secretly timed (the dependent variable in the experiment, time spent on the task). The participant was then probed for suspicions regarding the purpose of the study and debriefed.

The results from the study were analyzed as illustrated in the incomplete source table below. First, complete the source table below.

Descriptives

Time

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Elevation	11	40.6364	17.06032	5.14388	29.1751	52.0976	18.00	63.00
Control	11	19.9091	8.36008	2.52066	14.2927	25.5255	10.00	32.00
Mirth	11	23.7273	14.06479	4.24069	14.2784	33.1761	8.00	48.00
Total	33	28.0909	16.07087	2.79758	22.3924	33.7894	8.00	63.00

Test of Homogeneity of Variances

Time

Levene Statistic	df1	df2	Sig.
4.915	2	30	.014

Multiple Comparisons

Time
Tukey HSD

(I) Type of Film	(J) Type of Film	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Elevation	Control	20.72727*	5.81932	.003	6.3811	35.0735
Elevation	Mirth	16.90909*	5.81932	.018	2.5629	31.2553
Control	Elevation	-20.72727	5.81932	.003	-35.0735	-6.3811
Control	Mirth	-3.81818	5.81932	.790	-18.1644	10.5280
Mirth	Elevation	16.90909	5.81932	.018	31.2553	2.5629
Mirth	Control	3.81818	5.81932	.790	10.5280	18.1644

*. The mean difference is significant at the 0.05 level.

ANOVA

Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2677.091	2	1338.545	7.187	.003
Within Groups	5587.636	30	186.255		
Total	8264.727	32			

Next, analyze and interpret the results as completely as you can. [15 pts]

I've printed the SPSS post hoc analysis using Tukey's HSD. You would have had to compute HSD on your own:

$$HSD = 3.49 \sqrt{\frac{186.255}{11}} = 14.36$$

There was a significant effect of feelings on time spent on the math problem task, $F(2,30) = 7.187$, $MSE = 186.255$, $p = .003$, $\eta^2 = .324$. (You might note that the Levene test indicates that you should be concerned about heterogeneity of variance. However, the F -ratio would be significant even with a conservative $\alpha = .01$.) Post hoc tests using Tukey's HSD indicate that people in the Elevation condition spent significantly more time on the math problem task ($M = 40.636$) than people in the Control ($M = 19.909$) or Mirth ($M = 23.727$) conditions.

4. Distinguish between internal and external validity. Using evidence from your readings this semester (especially Mook), which type of validity would you argue is more important (i.e., less serious to violate). [10 pts]

Answer using resources from the class.

5. Another recent article in *Psychological Science* comes from Saul Miller and Jon Maner:

Miller, S. L., & Maner, J. K. (2010) Scent of a woman: Men's testosterone responses to olfactory ovulation cues.

Adapted from their abstract:

Adaptationist models of human mating provide a useful framework for identifying subtle, biologically based mechanisms influencing cross-gender social interaction. In line with this framework, the current studies examined the extent to which olfactory cues to female ovulation—scents of women at the peak of their reproductive fertility— influence endocrinological responses in men. Men in the current study smelled T-shirts worn by women near ovulation or far from ovulation. Men exposed to the scent of an ovulating woman subsequently displayed higher levels of testosterone than did men exposed to the scent of a nonovulating woman. Hence, olfactory cues signaling women's levels of reproductive fertility were associated with specific endocrinological responses in men— responses that have been linked to sexual behavior and the initiation of romantic courtship.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.136 ^a	.019	-.007	3.26161

a. Predictors: (Constant), Days from Ovulation

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.629	1	7.629	.717	.402 ^a
	Residual	404.249	38	10.638		
	Total	411.878	39			

a. Predictors: (Constant), Days from Ovulation

b. Dependent Variable: Testosterone level (

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.504	.535		15.886	.000
	Days from Ovulation	-.092	.108	-.136	-.847	.402

a. Dependent Variable: Testosterone level (

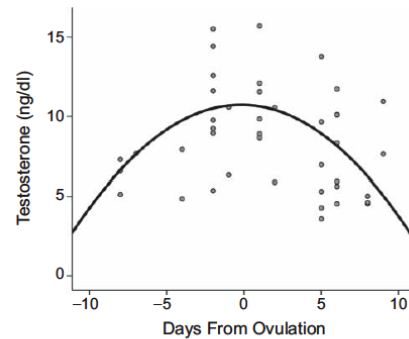


Fig. 2. Results from Study 2: postsmell testosterone levels (controlling for presmell testosterone levels) among men exposed to a woman's odor, as a function of the woman's estimated days from ovulation.

I've interpolated their data from the graph to produce the output above. The graph is from their paper (note the curve). Interpret their results as completely as you can. Given the output, how would you predict the man's testosterone level, when presented with a T-shirt from a woman who was ovulating (0 days from ovulation)? Judging from the abstract, would you describe this study as correlational? In other words, would you be comfortable making causal claims? [10 pts]

Although the curvilinear line of best fit seems a bit optimistic, it's clear that the relationship isn't linear. As a result, the linear relationship isn't significant, $r(38) = -.136, p = .402$. However, there may well be a significant curvilinear relationships between Days from Ovulation and Testosterone level. Given what you know about linear correlation and regression, you might well separately analyze the data for -10 to 0 days from ovulation and the data for 0 to +10 days from ovulation. In so doing, you might be able to predict the testosterone level with some accuracy...probably somewhere around 10 ng/dl. In this case, because the level of the IV given to each subject was randomly assigned, the design is actually experimental. Thus, were the results significant, one could make causal claims.

6. In the video at the beginning of the semester, two studies were described (the social psychological study on fear and affiliation and the perceptual development study on the effects of light deprivation). One of the studies would have benefited from a manipulation check and the other would not. Which one would, and why? Which one would not, and why not? Use either of these two studies to illustrate an operational definition by telling me how the researchers operationally defined one of their IV's or DV's. [5 pts]

It's difficult to imagine how one might do a manipulation check with kittens ☺, moreover, the manipulation there was easily measured (amount of time spent in darkness). On the other hand, for the social psychological study (fear and affiliation), it wasn't obvious that the attempt to manipulate fear was uniformly successful. That is, there's no simple way to determine if the manipulation actually induced fear. Thus, some manipulation check would be useful.

In the kitten study, the researchers operationally defined performance as pacing (number of times the animal changed direction). In the fear study, the researchers operationally defined affiliation as the self-report on a piece of paper as to whether or not the participant wanted to be alone or with other people...and how much.

7. In the light deprivation study from that video, we discussed the option of using a different dependent variable than depicted in the study. What was that DV, and would you agree that it would be more sensitive? Of course, greater sensitivity is another way of describing more power. First, define power and then describe general strategies for increasing the power in a study. [5 pts]

Instead of pacing, several of us thought that a simple "time to complete the maze" would make more sense as the dependent variable. Doing so would remove the need for multiple raters (and an assessment of concordance among the raters), and may well yield a more precise measure of performance. Power, of course, is the probability of correctly rejecting H_0 . The strategies for increasing power include: increasing n , increasing the treatment level (or making the treatment levels more different), decreasing the individual differences (e.g., using all males, or all females), and decreasing random variability (e.g., making instructions clearer).

8. A repeated measures ANOVA will typically yield a higher F -ratio than an independent groups ANOVA on the same data. Under which circumstances will that *not* be true? [2 pts]

The F -ratio for the repeated measures analysis will be smaller than that for the independent groups ANOVA when the individual differences are relatively small. That is, when you "remove" the SS_{Subject} from the SS_{Within} to yield the SS_{Error} , if the SS_{Subject} is proportionately small, then your MS_{Error} will be larger for a repeated measures ANOVA, which means that your F -ratio will be smaller.

9. Dr. Jones decides to test the effectiveness of two different experimental methodology textbooks. He gets two of his colleagues to agree to use the texts and to give the same exams throughout the term. At the end of the term, he finds that there was no difference in mean performance between the two classes (Mean = 98% and 96% for Class A and Class B, respectively). He concludes that there is no difference between the two texts. Would you agree? Why or why not? [3 pts]

First of all, you might be concerned about a confound. That is, there is a confound between text and professor. Were a significant difference to emerge, you couldn't be sure if it's due to the text or the professor. However, the question really is based on the performance,

which suggests a ceiling effect. The mean scores for both classes are so high that it's difficult to know if the texts played a role. Beyond that point, however, keep in mind that we never want to accept H_0 (absolutely no difference between the two conditions), because with more power a modest difference may become statistically significant.

10. For Lab 2, we're studying the extent to which faces in a photo array (a six pack) are similar to one another. Knowing the basic description provided by an eyewitness (or eyewitnesses), police then create an array that contains the photo of their suspect, as well as five other photos that should be similar to the description (and similar to the suspect). Thus, ideally the faces in an array would be so similar that a person who was not an eyewitness should be choosing a face based on the description in a fairly random fashion. In other words, when rating the extent to which each face in the photo array matched the eyewitness description, a participant should rate each face as an equally good match to the description. With $H_0: \mu_{\text{Face1}} = \mu_{\text{Face2}} = \mu_{\text{Face3}} = \mu_{\text{Face4}} = \mu_{\text{Face5}} = \mu_{\text{Face6}}$, we're asserting that people should be rating the faces as equally good matches. In other words, if the array is unbiased, the data should be consistent, leading to a non-significant F ratio. Complete the analysis below (for the first photo array) and interpret the results as completely as you can. (A1F1 means Array 1 Face 1.) [15 pts]

Descriptive Statistics

	Mean	Std. Deviation	N
A1F1	3.4719	1.32365	89
A1F2	4.5955	1.39577	89
A1F3	3.9326	1.49086	89
A1F4	2.3933	1.18339	89
A1F5	5.7865	1.42599	89
A1F6	3.6854	1.29330	89

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^a
face	Sphericity Assumed	579.146	5	115.829	71.999	.000	.450	359.996	1.000
	Greenhouse-Geisser	579.146	4.308	134.436	71.999	.000	.450	310.171	1.000
	Huynh-Feldt	579.146	4.557	127.081	71.999	.000	.450	328.122	1.000
	Lower-bound	579.146	1.000	579.146	71.999	.000	.450	71.999	1.000
Error(face)	Sphericity Assumed	707.854	440	1.609					
	Greenhouse-Geisser	707.854	379.102	1.867					
	Huynh-Feldt	707.854	401.043	1.765					
	Lower-bound	707.854	88.000	8.044					

a. Computed using alpha =

Given the significant results, you'd need to compute a post hoc test:

$$HSD = 4.03 \sqrt{\frac{1.609}{89}} = .542$$

There was a significant effect of face on ratings of match to the provided description, $F(5,440) = 71.999$, $MSE = 1.609$, $p < .001$, $\eta^2 = .450$. Post hoc tests using Tukey's HSD indicate that Face 5 in the array was judged as a significantly better fit to the description ($M = 5.787$) than Face 2 ($M = 4.595$), Face 3 ($M = 3.933$), Face 6 ($M = 3.685$), Face 1 ($M = 3.472$), and Face 4 ($M = 2.393$). Subjects judged Face 2 to be a significantly better fit than Face 3, Face 6, Face 1, and Face 4. Although ratings of Face 3, Face 6, and Face 1 did not differ, all three were rated as better fits to the description than Face 4. Thus, the photo array was biased, with some of the faces providing a much better fit to the eyewitness descriptions than other faces.