Review of 2-group 1-factor ANOVA

- E.g., effect of induced guilt vs. control on sharing (0 to $100).
- Quantitative DV, categorical IV
- Notation: k=2 groups; n subjects per group; n=k=N total subjects
- If subjects are randomly drawn from some population, the experiment is generalizable to that population, regardless of sample size, which sets external validity (narrow vs. broad). (Practically, subjects are representative of some larger group.)
- If treatment is randomly assigned and sample size is not too small, then the only subject characteristics with non-negligible average difference between groups is treatment (no confounding), and we can claim causality, i.e., good internal validity.
- Notation: μC, μG are population means of outcome ($) for the two treatment groups.
- We observe YC andYG, the sample means of outcome ($) for the two treatment groups.

Introduction

- Common mistaken impression: After seeing the p-value, and choosing “retain” vs. reject” H0 based on α=0.05, we know the chance that we have “made a mistake”.
- What the omniscient see:

Review of 2-group 1-factor ANOVA

- Goal: use sample means to make inference about the effects of changing IV levels on the population mean of the DV.
- H0: μC=μG, H1: μC≠μG
- Inference: Compare a statistic to its null sampling distribution.
- Statistic: F = MSbetween-groups / MSwithin-groups
- Null sampling distribution of the F-statistic; dfB=(k-1), dfW=k(n-1)
- F-statistic (calculation) → p-value (inferred from data and model)
- The only sampling distribution used is the null sampling distribution (not the alternative)
- Alpha (significance level) determines the Type 1 error (reject rate for true H0)
- “Critical” F value: above=reject, below=retain H0
**Components of an alternative scenario**

- Population means (see below, for more details)
- $n$ (sample size, per group; or $N=total$ sample size)
- $\sigma^2$ ($\sigma_e^2$) is the error variance
- Other: $x$-spacing for regression, etc.

**Life Experience Examples**

- Definitions:
  - A "positive" result for an experiment means finding $p \leq \alpha$. "Negative" means finding $p > \alpha$. **Neither needs omniscience.**
  - "True" means matching reality (i.e. reject $H_0$ when $H_0$ is really false or retain $H_0$ when $H_0$ is really true), and "false" means incorrect. **Both need omniscience!**
- Calculations (choosing $\alpha=0.05$):
  - Positive rate among null experiments: 5%
  - Positive rate for a specific alternative: "power" %

**Type 2 error and power**

Naomi Null studies the effects of various chants on blood sugar level. Every week she studies 15 controls and 15 people who chant a particular word from the dictionary for 5 minutes. After 1000 weeks (and 1000 words) what is her Type 1 error rate (positives among null experiments), type 2-error rate (negatives among non-null experiments) and power (positives among non-null experiments)? What percent of her positives are true? What percent of her negatives are true? [Assume chanting does not affect blood sugar.]
Life Experience Examples

Christine Cautious studies the change in glucose levels due to injecting cats with subcutaneous insulin in different locations. She divides the surface of a cat into 1000 zones and each week studies injection of 10 cats with water and 10 cats with insulin in a different zone. [Missing info:]

Life Experience Examples

Andrea Average works for a large pharmaceutical firm performing initial screening of potential new oral hypoglycemic drugs. Each week for 1000 weeks she gives 100 rats a placebo and 100 rats a new drug, then tests blood sugar. To increase power (at the expense of more false positives) she chooses alpha=.10. [Missing info:]

Life Experiences Conclusion

➢ For your career, you cannot know the chance that a negative result is an error or the chance that a positive result is an error.
➢ But you do know that when you study control vs. ineffective treatment (and your model assumptions are met) then you have only a 5% chance of incorrectly claiming the treatment is effective.
➢ And you know that the more you increase the power of an experiment, the better your chances are of detecting any truly effective treatment.

A measure of effect size for ANOVA

Example: $\mu_1=5$, $\mu_2=15$, $\mu_3=40$

Using SPSS “descriptive statistics“:

$\sigma_A = SD[\text{treatment}]=18.0$

Key observation: A larger difference between population means increases $\sigma_A$. Only the spacing matters.

E.g., sd(5,15,40) = sd(6,16,41) = sd(0,25,35)
**Expected Mean Square (EMS)**

- Let $\sigma_e^2$ be the true error variance (including subject-to-subject, treatment application, environmental, and measurement variability) for each group. As usual, $n$ is the number of subjects per group.
- Here is the EMS table for one-way (between subjects) ANOVA for any mean spacing.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor A</td>
<td>$M_{S_A}$</td>
<td>$\sigma_e^2 + n \sigma_A^2$</td>
</tr>
<tr>
<td>Error (residual)</td>
<td>$M_{S_{error}} = M_{S_{within groups}}$</td>
<td>$\sigma_e^2$</td>
</tr>
</tbody>
</table>

**EMS, F statistic, and power**

- $E(F) = E(MS_A/MS_{error}) = \frac{\sigma_e^2 + n \sigma_A^2}{\sigma_e^2}$.
- E.g., $n \sigma_A^2 = 10$, $\sigma_e = 10$ vs. 1

**Power Calculation**

- Here we focus on the simple case: power in a one-way between-subjects design. Two-way ANOVA without interaction is demonstrated in lab. Two-way with interaction and linear regression are shown in the textbook (§12.84, §12.85, optional).

- **Sine qua non**: Beyond $k$ and alpha ($\alpha$), power depends on sample size, an estimate of experimental error (variance or s.d.), and one or more target effect sizes (or their spacing).

**Power Calculation, cont.**

- Technical note: Alternative F sampling distributions are non-central F distributions, with a 3rd index call the non-centrality parameter, which equals zero for $H_0$.
- We need to specify particular alternative hypotheses (target effect sizes): (§12.6)
  - reasonably likely to occur
  - or minimally interesting
  - or minimum effect size that will change your behavior
Power Calculation, cont.

- Obtaining an estimate of $\sigma^2$ (§12.5)
  - Statistical analysis of previous experiments (MSE, $MS_{within}$ or $MS_{residual}$) with similar error variance.
  - Pilot experiment: variance of the outcome measurement for a number of subjects exposed to the same (any) treatment.
  - Expert knowledge: guesstimate the 95% range ($\pm 2$ s.d.) of, say, control subjects. Assuming normality, $\sigma$ is estimated as the 95% range divided by 4.
- Conventionally, "acceptable" power is 80%

The calculation: Lenth Power applet

- Let alpha=0.10 and $n=11$ per cell. In a similar experiment MSE=36. What is the power for the alternative hypothesis $\mu_1=10$, $\mu_2=12$, $\mu_3=14$, $\mu_4=16$?
- Under the null hypothesis $F$ will follow the [central] $F$ distribution with $k-1=3$ and $k(n-1)=40$ df. The applet (silently) finds that $F_{critical}=2.23$.

Power Applet, cont.

- Find $sd(10,12,14,16) = 2.58$
- In the applet enter $SD[\text{treatment}]= 2.58$
- The power is the area under the particular [non-central] $F$ curve corresponding to your alternative scenario and which is higher than $F_{critical}=2.23$. The applet finds that this area is 0.62. This indicates that we have a 62% chance of rejecting the null hypothesis if the given alternate hypothesis is true. So the power is 62%.

Power Calculation, cont.

- You should know that the power is
  - bigger than what we calculated (62%, here) if
    - the true error variance is smaller than what we used for $\sigma^2$
    - the true population means are more spread out than for what we calculated
    - more than $kn$ subjects are studied
  - and vice versa.
Conclusion

Although there is a bit of educated guesswork in calculating (estimating) power, it is strongly advised to make some power calculations before running an experiment to find out if you have enough power to make running the experiment worthwhile.