

# Chapter 2

---

## *Fundamental Concepts*

The wisest mind has something yet to learn.

—George Santayana

### Overview

- Review of basic concepts about:
  - ✓ Comparative studies
  - ✓ Sampling and estimation
  - ✓ General characteristics of statistical tests
  - ✓ Power estimation
  - ✓  $t$ ,  $F$ , and  $\chi^2$

# Concepts about comparative studies

---

- Basic distinction: Designs with independent samples vs. dependent samples
- Designs with dependent samples are also called correlated designs
- Balanced vs. unbalanced designs (i.e., equal or unequal group sizes):
  - ✓ Related to missing observations
  - ✓ Balanced design not always better
  - ✓ Affects generalizability and effect size estimation

# Concepts about comparative studies

---

- Types of correlated designs:

- ✓ Repeated measures designs:

- Intrinsically vs. non-intrinsically repeated-measures

- ✓ Matched groups designs

- Correlated designs *may* reduce error variance, but are more difficult (e.g., additional statistical assumptions, control for order effects)

# Concepts about comparative studies

---

- A covariate analysis is another way to reduce error variance
- Covariate is a variable that predicts outcome but is unrelated to the independent variable(s)
- The covariate is partialled out of the outcome variable, which controls for the covariate
- However, covariate analyses are generally best for experimental designs due in part to very restrictive assumptions

# Concepts about comparative studies

---

- Fixed-effects vs. random-effects factors:
  - ✓ Fixed: Levels selected systematically, results generalize only to those levels
  - ✓ Random: Levels selected randomly, results may generalize to other levels not studied
- Affects generalizability, statistical estimation of error variance, and estimation of certain kinds of effect sizes
- Most factors analyzed as fixed but are often interpreted as though they were random

# Sampling and estimation

---

- Types of samples:
  - ✓ Random
  - ✓ Systematic
  - ✓ Ad hoc (sample of convenience)
- Most samples in the behavioral sciences are ad hoc and not random
- Concept of random sampling does not generally apply to animal research
- However, the interpretation of results from statistical tests generally assumes random sampling
- There is thus a mismatch between design (i.e., what researchers do in practice) and analysis (i.e., the sampling model assumed by statistical tests)

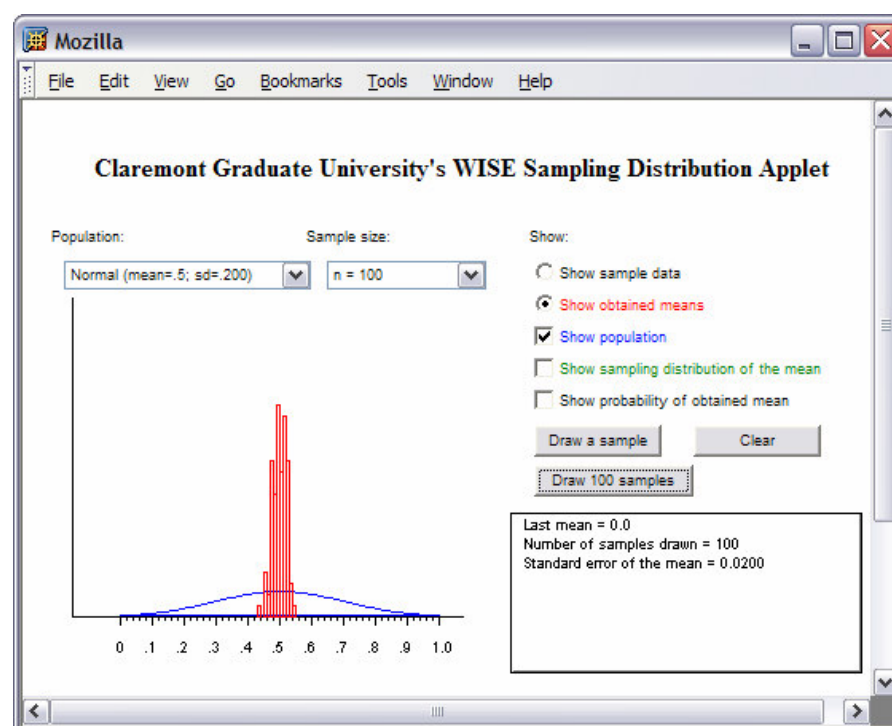
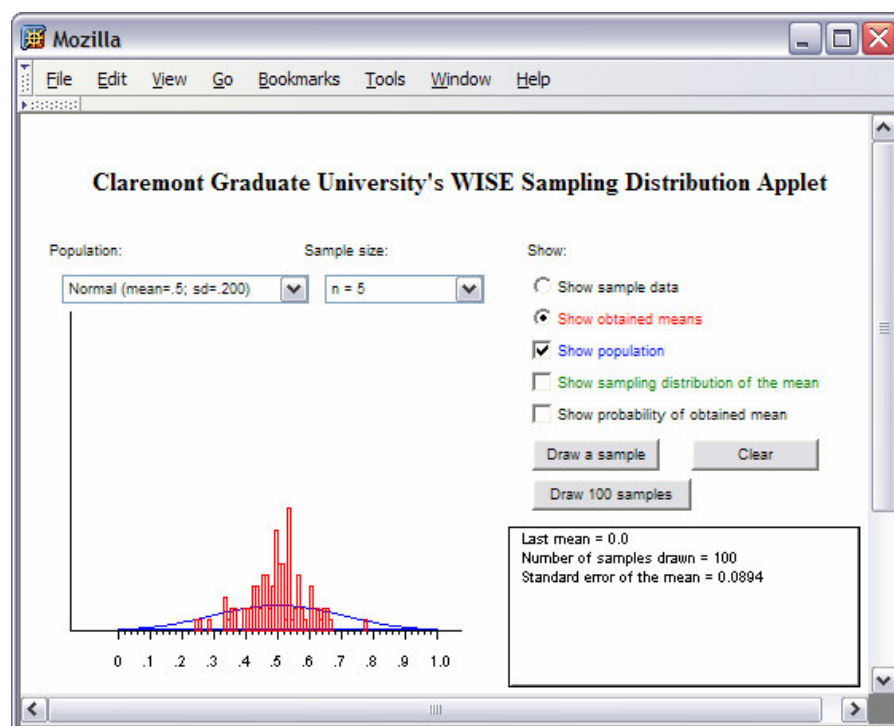
# Sampling and estimation

---

- Sample statistics estimate population parameters but are subject to sampling error
- Amount of sampling error is generally affected by:
  - ✓ Variability of population observations
  - ✓ Sample size
- Law of large numbers: Statistics in larger samples tend to be closer on average to the population parameter than in smaller samples
- But some researchers seem to forget this fact—for example, they tend to think in terms of the “law of small numbers” (chap. 3)

# Sampling and estimation

- Example: Claremont University WISE sampling distribution applet (<http://wise.cgu.edu/sdmmod/sdm.html>)
- Sampling distributions for 100 samples for  $N = 5$  (left) vs.  $N = 100$  (right) given  $\mu = .5$ ,  $\sigma = .2$ , and normality:



# Sampling and estimation

---

- Unbiased estimator: Average value across all random samples equals the population parameter—examples:
  - ✓  $M$  is an unbiased estimator of  $\mu$
  - ✓  $s^2 = SS/df$  is an unbiased estimator of  $\sigma^2$
- Biased estimator: Above property does not hold—examples:
  - ✓  $S^2 = SS/N$  is a negatively biased estimator of  $\sigma^2$
  - ✓  $s$  is a negatively biased estimator of  $\sigma$

# Sampling and estimation

---

- Corrections are available for some estimators—example:

$\hat{\sigma}$  approximates an unbiased estimator of  $\sigma$ :

$$\hat{\sigma} = \left( 1 + \frac{1}{4(N-1)} \right) s$$

- Similar corrections are available for some effect sizes
- Corrections are generally unnecessary in large samples

# Sampling and estimation

---

- Interval estimation involves the construction of a confidence interval based on a sample statistic
- Confidence interval explicitly estimates sampling error
- Formal definition (percentage-based):

A  $100(1 - \alpha)\%$  confidence interval for a parameter is a pair of statistics yielding an interval that, over repeated samples, includes the parameter  $100(1 - \alpha)\%$  of the time

# Sampling and estimation

---

- Traditional confidence interval: Width is usually the product of a central test statistic and an estimated standard error
- A *central test statistic* is from a distribution that assumes  $H_0$  is true
- A standard error is the standard deviation of a sampling distribution

# Sampling and estimation

---

- Exact formulas for standard errors are generally available for statistics with simple distributions, such as means
- Simple distribution: Statistic estimates just one parameter, and distribution shape is unaffected by the value of the parameter
- Approximate formulas amenable to hand calculation are available for some, but not all, statistics with complex distributions
- These approximate methods generally assume large samples (i.e., they are asymptotic standard errors)

# Sampling and estimation

---

- Form of a traditional confidence interval for  $\mu$ :

$$M \pm s_M [t_{2\text{-tail}, \alpha} (N - 1)]$$

$$s_M = \frac{s}{\sqrt{N}}$$

$s_M$  is the standard error of the mean

$t_{2\text{-tail}, \alpha} (N - 1)$  is the two-tailed critical value at the  $\alpha$  level of statistical significance in a central  $t$  distribution with  $N - 1$  degrees of freedom

## Sampling and estimation

---

- Form of a traditional confidence interval for  $\mu_1 - \mu_2$  (independent means):

$$(M_1 - M_2) \pm s_{M_1 - M_2} [t_{2\text{-tail}, \alpha} (N - 2)]$$

$$s_{M_1 - M_2} = \sqrt{s_P^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$s_P^2 = \frac{SS_W}{df_W} = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}$$

$s_{M_1 - M_2}$  is the standard error of the mean difference

$s_P^2$ ,  $SS_W$ , and  $df_W$  are, respectively, the pooled within-groups variance, sum of squares, and degrees of freedom

## Sampling and estimation

---

- Form of a traditional confidence interval for  $\mu_D$  (dependent means):

$$M_D \pm s_{M_D} [t_{2\text{-tail}, \alpha} (n - 1)]$$

$$s_{M_D} = \frac{s_D}{\sqrt{n}}$$

$$s_D^2 = s_1^2 + s_2^2 - 2 \text{COV}_{12}$$

$$\text{COV}_{12} = r_{12} s_1 s_2$$

$s_{M_D}$  is the standard error of the mean difference

$s_D^2$  is the variance of the difference scores

$s_1^2$  and  $s_2^2$  are the within-groups variances

$\text{COV}_{12}$  and  $r_{12}$  are, respectively, the cross-conditions covariance and Pearson correlation

## Sampling and estimation

---

- Interpretation of a confidence interval—example:

96.28 to 103.72 is a 95% confidence interval for  $\mu$

- Generally *incorrect* (based on subjectivist view of probability):

There is a 95% chance that  $\mu$  is between 96.28 and 103.72

- Generally *correct* (based on frequentist view of probability):

If 95% confidence intervals are constructed around the means of all random samples, then 95/100 of them will include  $\mu$ , but 5/100 will not

- However, Bayesian statistics may permit a subjectivist interpretation in some circumstances (chap. 9)

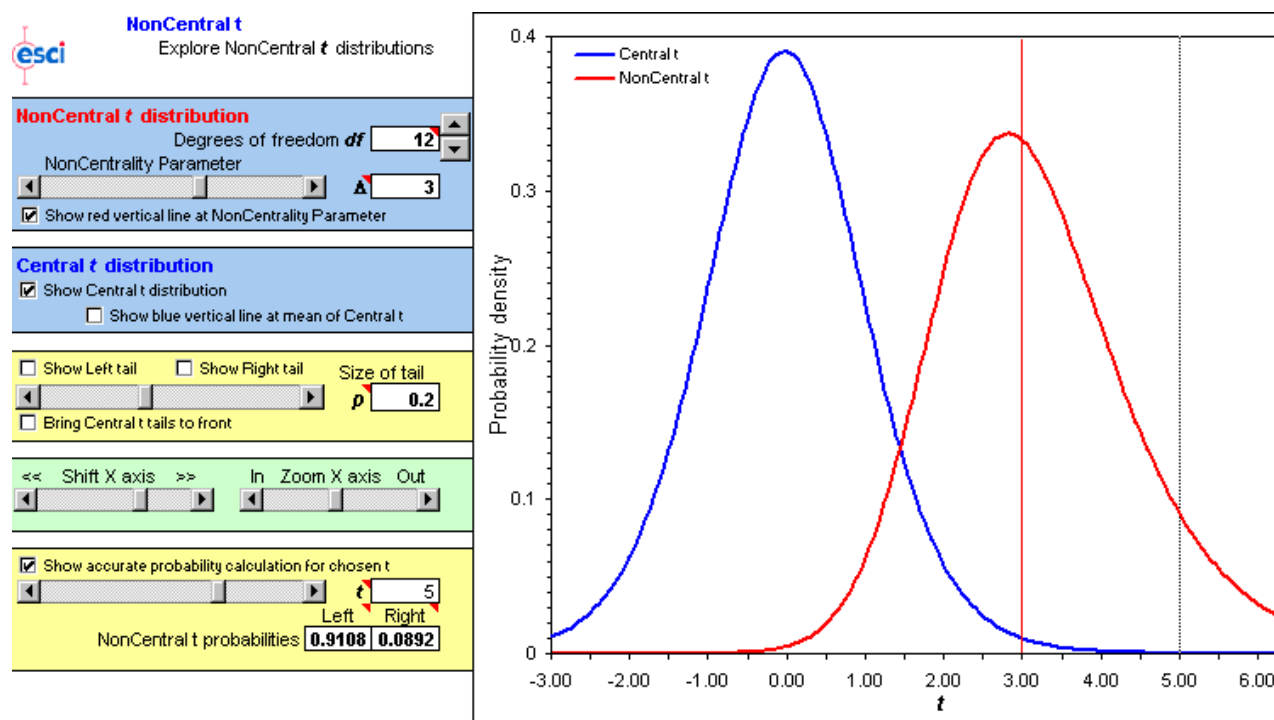
# Sampling and estimation

---

- A more exact method for constructing confidence intervals based on statistics with complex distributions is *noncentrality interval estimation*
- A *noncentral distribution* of test statistics does not assume a true  $H_0$
- Noncentral distributions have an additional parameter, the *noncentrality parameter*, which estimates the degree to which  $H_0$  is false
- All central distributions of test statistics are just special cases of noncentral distributions

# Sampling and estimation

- Example: This figure was created by ESCI (Cumming & Finch, 2001)
- It shows a central  $t$  distribution with 12 degrees of freedom ( $df$ ; blue, left), and a noncentral  $t$  distribution for the same  $df$  where the noncentrality parameter equals 3 (red, right):



# Sampling and estimation

---

- Estimation of the noncentrality parameter is usually required for a noncentral confidence interval
- This requires a computer tool, such as ESCI
- Effect sizes may have complex distributions, and the method of noncentrality interval estimation can be used for some of these statistics (chap. 4)

# General characteristics of statistical tests

---

- General logic of the “Intro Stats” method:
  1. Formulate two statistical hypotheses,  $H_0$  (null) and  $H_1$  (alternative)
  2. Set the level of  $\alpha$ , the probability of a Type I error
  3. Collect data, determine  $p$  under  $H_0$ , reject  $H_0$  if  $p < \alpha$

# General characteristics of statistical tests

---

- Two contexts for testing:

- ✓ Reject-support (most common):

Rejection of  $H_0$  supports the researcher's theory

- ✓ Accept-support:

Failure to reject  $H_0$  supports the researcher's theory

# General characteristics of statistical tests

---

- Two types of null hypotheses:

- ✓ Nil (most common):

Assumes zero population effect or difference (e.g.,  $H_0: \rho = 0$ ), which may be unrealistic

- ✓ Non-nil:

Allows a nonzero population effect (e.g.,  $H_0: \rho = .30$ ), which may be more realistic

- Rejection of an implausible null hypothesis is not impressive
- The probability of data under an implausible null hypothesis is expected to be low

# General characteristics of statistical tests

---

- Two types of alternative hypotheses:
  - ✓ Directional (e.g.,  $H_1: \mu_1 - \mu_2 > 0$ )
  - ✓ Nondirectional (e.g.,  $H_1: \mu_1 - \mu_2 \neq 0$ )
- Generally easier to reject  $H_0$  for a directional  $H_1$
- Switching from nondirectional to directional  $H_1$  in order to reject  $H_0$  may be considered “cheating”

## General characteristics of statistical tests

---

- The term  $\alpha$  is the conditional prior probability of rejecting  $H_0$  when it is true:

$$\alpha = p(\text{reject } H_0 \mid H_0 \text{ true})$$

- The term  $\alpha$  is a long-run, relative-frequency statement about the likelihood of a Type I error
- Switching  $\alpha$  (e.g., from .01 to .05) in order to reject  $H_0$  may be considered “cheating”

## General characteristics of statistical tests

---

- There are many statistical tests, but their  $p$  values generally estimate:

$$p(\text{Data} \mid H_0 \text{ true})$$

- The terms  $p$  and  $\alpha$  are defined in the same sampling distribution, but  $p$  is *not* the conditional prior probability of a Type I error

# General characteristics of statistical tests

---

- Most test statistics (and their  $p$  values) measure both effect size *and* sample size together with a single number
- This explains how a
  1. trivial effect size can be statistically significant in a large sample
  2. large effect size can fail to be statistically significant in a small sample
- Researchers need measures of effect size only—they already know their sample sizes!

# Power estimation

---

- Power is the conditional prior probability of making the correct decision to reject  $H_0$  when it is actually false:

$$\text{Power} = p(\text{reject } H_0 \mid H_0 \text{ false})$$

$$p(\text{Type II error}) = \beta = 1 - \text{power}$$

- Power is affected by  $N$ ,  $\alpha$ , directionality of  $H_1$ , study design, score reliability, the type of test statistic, and magnitude of the true effect
- Power should be estimated *before* the data are collected, not *after*
- Power analysis should be like a diagnostic procedure, not an autopsy

# Power estimation

---

- Ideally, power should be close to 1.00
- However, typical power in behavioral research is only about .50
- When power = .50, we can replace the study with a coin toss and still have the same probability of detecting a true effect

# Power estimation

---

- However, increasing power may require unrealistically large sample sizes when effects of interest are not large
- When power is not high, the meaning of the failure to reject the null hypothesis is ambiguous
- Power is irrelevant if statistical tests are not used

## $t$ , $F$ , and $\chi^2$

---

- Perhaps the most widely used statistical tests in the behavioral sciences
- Results of each test are also widely misinterpreted
- $t$  and  $F$  have specific distributional assumptions that may be untenable in perhaps most behavioral studies, especially in correlated designs

## $t$ , $F$ , and $\chi^2$

---

- Contrary to what many researchers believe, assumptions of  $t$  and  $F$  for designs with independent samples cannot be violated with impunity—examples:
  - ✓ Glass, Peckham, and Sanders (1972)
  - ✓ Wilcox (1987, 1998)
  - ✓ Winer, Brown, and Michels, (1991)
- Too many researchers do not bother to check distributional assumptions (e.g., Lix, L. Keselman, & H. Keselman, 1996)
- However, it is useful to know about test statistics because in many cases effect sizes can be computed from them

## $t$ , $F$ , and $\chi^2$

---

- Independent-samples  $t$  for a nil  $H_0$ :

$$t(df_W) = \frac{M_1 - M_2}{S_{M_1 - M_2}}$$

- Dependent-samples  $t$  for a nil  $H_0$ :

$$t(n - 1) = \frac{M_D}{S_{M_D}}$$

- Both express a mean contrast as the proportion of its standard error
- Standard error metric of  $t$  is affected by sample size (see Table 2.2)

## $t$ , $F$ , and $\chi^2$

---

- Independent-samples  $F$  for the simultaneous comparison of  $a \geq 3$  means (i.e., the omnibus effect of factor  $A$  with  $df_A = a - 1$  degrees of freedom):

$$F(df_A, df_W) = \frac{MS_A}{MS_W}$$

$$MS_A = \frac{SS_A}{df_A}, \quad SS_A = \sum_{i=1}^a n_i (M_i - M_T)^2$$

$$MS_W = \frac{SS_W}{df_W}, \quad SS_W = \sum_{i=1}^a df_i (s_i^2)$$

$M_i$  is the mean of the  $i$ th group,  $M_T$  is the grand mean

- Assumes normality, independence, and homogeneity of variance
- Numerator of  $F$  affected by sample size (see Table 2.6)

## $t$ , $F$ , and $\chi^2$

---

- Dependent-samples  $F$  (nonadditive model):

$$F(df_A, df_{A \times S}) = \frac{MS_A}{MS_{A \times S}}$$

$$MS_{A \times S} = MS_W - M_{COV} = \frac{SS_W - SS_S}{df_W - df_S}$$

$A \times S$  refers to a person  $\times$  treatment interaction

$S$  refers to the subjects effect

$M_{COV}$  is the average cross-conditions covariance

## $t$ , $F$ , and $\chi^2$

---

- The dependent-samples  $F$  assumes normality, independence, and homogeneity of variance and covariance (i.e., sphericity, circularity)
- Violation of sphericity may result in positive bias ( $H_0$  rejected too often)
- This assumption may be untenable in perhaps most correlated designs

## $t$ , $F$ , and $\chi^2$

---

- Chi-square test of association in two-way contingency tables:

$$\chi^2((r-1) \times (c-1)) = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{o_{ij}} - f_{e_{ij}})^2}{f_{e_{ij}}}$$

$r$  is the number of rows,  $c$  is the number of columns

- Assumes independence and “reasonable” minimum values of  $f_{e_{ij}}$
- Like most statistical tests, also sensitive to sample size (see Table 2.9)

# References

---

- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-288.
- Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumptions violations revisited: A quantitative review of alternatives to the one-way analysis of variance  $F$  test. *Review of Educational Research, 66*, 579-620.
- Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology, 38*, 29-60.
- Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*, 300-314.
- Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). Boston: McGraw-Hill.