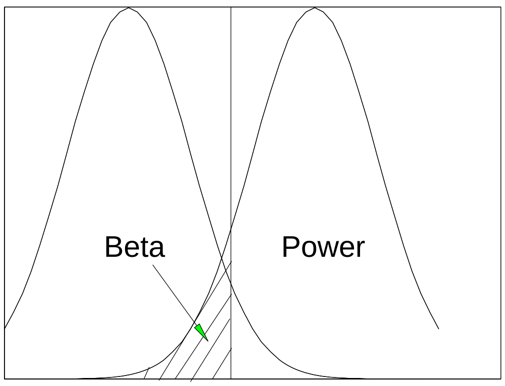
Lecture 16: Statistical Power

• 16. 1 Fundamentals and Factors Affecting Power of a Statistical Test

- ***Power*** $(1-\beta)$: The probability of correctly rejecting a false null hypothesis



- It is desirable to have a small alpha (few Type I errors) and large power (few Type II errors), but usually is a trade-off

- Power is affected by:

    - alpha (α)

    - the true H_1 (distance between μ_0 and μ_1)

        - the chances of finding a difference between μ_0 and μ_1 depend on how large the difference actually is

    - sample size and variability

        - the variance of the sampling distribution of the mean decreases as n increases and σ^2 decreases

        $$\sigma_N = \frac{\sigma}{\sqrt{n}}$$

        - it results in less variability without affecting the means

        - n is the most controllable factor

    - type of test (i.e., 1-tailed test vs. 2-tailed test)

• 16. 2 Effect Size

- Power depends on the overlap between the sampling distributions of H_0 and H_1

- This overlap is a function of the differences between $\mu_0 - \mu_1$ and the standard error (variability)

- One possible way to express the certainty of H_0 being false would be the distance $\mu_0 - \mu_1$ expressed in terms of standard errors

  - But this distance requires computing n (in the standard error), which is precisely a factor to be estimated when computing power

- An alternative way to compute the distance is **d** (effect size)

$$\mathbf{d} = \frac{\mu_1 - \mu_0}{\sigma}$$

where σ is the standard deviation of the parent population (thus, n is not required)

- Estimating **d**

  - Prior research: guess values for $\mu_0 - \mu_1$ and σ from sample means and variances in other studies

  - Desired differences: specific $\mu_0 - \mu_1$ determined by the researcher; estimate σ from other data (e.g., norming studies)

  - Convention: Cohen's effect sizes

Effect size	**d**	% Overlap
Small	.20	85
Medium	.50	67
Large	.80	53

- Estimating δ (delta): to combine the effect size with n

$$\delta = \mathbf{d}[f(n)]$$

where $[f(n)]$ is defined by the specific test, but δ is comparable across various tests

• 16. 3 Calculating Power for the One-Sample *t*

- The most basic test; $[f(n)] = \sqrt{n}$

$$\delta = \mathbf{d}\sqrt{n}$$

where δ can now be determined by power tables

- ***Example***: test the hypothesis of a precise difference (e.g., 5 points) between the mean IQ in the general population and the mean IQ in a specific population (e.g., clinically depressed) with a random sample of 25 individuals

$\mu_0 = 100$, $\mu_1 = 105$, $\sigma = 15$, $n = 25$

$$\mathbf{d} = \frac{\mu_1 - \mu_0}{\sigma} = \frac{105 - 100}{15} = 0.33$$

$$\delta = \mathbf{d}\sqrt{n} = 0.33\sqrt{25} = 1.65$$

Using $\alpha = .05$ and a two-tailed test, power ≈ .38 (.38 probability of detecting *true* differences…or .62 probability of making a Type II error)

- How can power be increased?

- Increase alpha

- Increase *n*, by how much? It depends on the power desired. For example, power = .80

If power = .80 and $\alpha = .05$, $\delta = 2.80$, thus

$$\delta = \mathbf{d}\sqrt{n}\ ;\ n = \left(\frac{\delta}{\mathbf{d}}\right)^2 = \left(\frac{2.80}{0.33}\right)^2 = 71.91$$, or 72 individuals

- ## 16. 4 Calculating Power for Differences Between Two Independent Means

- Assuming $\sigma^2_1 = \sigma^2_2 = \sigma^2$, under H_0, $\mu_1 - \mu_2 = 0$, so the difference to be expected under H_1 is

$$\mathbf{d} = \frac{(\mu_1 - \mu_2) - (\mu_1 - \mu_2)}{\sigma} = \frac{(\mu_1 - \mu_2) - (0)}{\sigma} = \frac{\mu_1 - \mu_2}{\sigma}$$

- If $n_1 = n_2 = n$, then $\delta = \mathbf{d}\sqrt{\frac{n}{2}}$

where n is the number of cases in any one sample

- If $n_1 \neq n_2$, when n_1 and n_2 are reasonably large and nearly equal, choosing the smallest n gives a conservative approximation

- If $n_1 \neq n_2$, when n_1 and n_2 are not reasonably large and not nearly equal, the harmonic mean of the sample sizes is used

$$\bar{n}_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{2n_1n_2}{n_1 + n_1}$$

- ***Example***: given two samples with $\overline{X}_1 = 9.58$ and $\overline{X}_2 = 6.55$, with $n_1 = 18$ and $n_2 = 12$, and a pooled standard deviation s = 3.10

$$\mathbf{d} = \frac{\mu_1 - \mu_2}{\sigma} = \frac{9.58 - 6.55}{3.10} = 0.98$$

$$\bar{n}_h = \frac{2n_1n_2}{n_1 + n_2} = \frac{2(18)(12)}{18 + 12} = \frac{432}{30} = 14.40$$

$$\delta \ = \ \mathbf{d}\sqrt{\frac{\bar{n}_h}{2}} = 0.98\sqrt{\frac{14.4}{2}} = 0.98\sqrt{7.2} = 2.63$$

For $\delta = 2.63$ and $\alpha = .05$ (two-tailed), power = 0.75

• 16. 5 Calculating Power for Differences Between Two Dependent Samples

- For differences between two matched samples, an additional parameter is needed

$$\mathbf{d} = \frac{\mu_1 - \mu_2}{\sigma_{x_1 - x_2}}$$

where σ_{x1-x2} is the standard deviation of the difference scores from the two populations, which is typically unknown

- Making a few assumptions, σ_{x1-x2} can be calculated. Based on the variance sum law

$$\sigma^2_{x_1 \pm x_2} = \sigma^2_{x_1} + \sigma^2_{x_2} \pm 2\rho\sigma_{x_1}\sigma_{x_2}$$

- If we assume homogeneity of variance $\sigma^2_{x_1} = \sigma^2_{x_2} = \sigma^2$

$$\sigma^2_{x_1 - x_2} = 2\sigma^2 - 2\rho\sigma^2 = 2\sigma^2(1-\rho) = \sigma\sqrt{2(1-\rho)}$$

where ρ is the correlation in the population between X_1 and X_2 (it is typically positive for all dependent samples, e.g., two siblings, mother-child)

$$\mathbf{d} = \frac{\mu_1 - \mu_2}{\sigma_{x_1 - x_2}} \text{ and } \delta = \mathbf{d}\sqrt{n}$$

- Power is positively related to ρ

- When $\rho = 0$, the problem is reduced to independent samples

- Sample sizes required for power = .80 and $\alpha = .05$

Effect size	**d**	One-Sample *t*	Two-Sample *t*
Small	.20	196	784
Medium	.50	32	126
Large	.80	13	49

• 16. 6 Calculating Power for Analysis of Variance

- It is a straightforward extension of the power analysis for *t*, but with different notation

$$\frac{E(MS_{treat})}{E(MS_{error})} = \frac{\sigma_e^2 + \frac{n\sum(\mu_j - \mu)^2}{k-1}}{\sigma_e^2} = \frac{\sigma_e^2 + \frac{n\sum \tau^2{}_j}{k-1}}{\sigma_e^2}$$

- If H_0 is true, $\sum \tau^2{}_j = 0$ and $F = \frac{MS_{treat}}{MS_{error}}$ is distributed as the usual (central) *F* distribution

  - The mean of this *F* distribution $under\ H_0, E(F) = \frac{df_{error}}{df_{error} - 2}$, is close to 1 for large *n*

- If H_0 is false, $E(F) = \left(1 + \frac{n\sum \tau^2{}_j}{\sigma_e^2(k-1)}\right)\left(\frac{df_{error}}{df_{error} - 2}\right)$

  where $\frac{n\sum \tau^2{}_j}{\sigma_e^2(k-1)}$ is called a noncentrality parameter (***ncP***) and it displaces the *F* distribution up the scale away from 1 (as a function of the true differences among the population means)

- One way to obtain a standardized measure of effect size in the ANOVA context is

$$\phi' = \frac{\sigma_\tau}{\sigma_e} = \sqrt{\frac{\frac{\sum(\mu_j - \mu)^2}{k}}{\sigma_e^2}}$$

  where ϕ' is the equivalent to Cohen's measure of effect size (equal to **d** for $k = 2$), σ_τ is the standard deviation of group means, and *k* is the number of groups

- In order to incorporate sample size

  $\phi = \phi'\sqrt{n}$ and can be determined by the tables of the noncentral *F* distribution, given α, and df_s

- ***Example***: Let $\overline{X}_1 = 34.00$, $\overline{X}_2 = 50.80$, $\overline{X}_3 = 60.33$, $\overline{X}_4 = 48.50$, and $\overline{X}_5 = 38.10$ be five sample means of n = 10, with $\overline{X}_{..} = 46.346$ and σ_e^2 = 240.35 (MS_{error}, or average sample variance)

- Under a false H_0

$$E(F) = \left(1 + \frac{n\sum \tau^2{}_j}{\sigma_e{}^2(k-1)}\right)\left(\frac{df_{error}}{df_{error}-2}\right)$$

$$E(F) = \left(1 + \frac{10[(34-46.35)^2 + ... + (38.10-46.35)^2]}{(240.35)(5-1)}\right)\left(\frac{45}{45-2}\right) = (1+4.58)(1.046) = 5.838$$

$E(F)$ exceeds the critical value for $F_{4,45}$ = 2.58

$$\phi' = \frac{\sigma_\tau}{\sigma_e} = \sqrt{\frac{\frac{\sum(\mu_j-\mu)^2}{k}}{\sigma_e{}^2}} = \sqrt{\frac{\frac{[(34-46.35)^2 + ... + (38.10-46.35)^2]}{5}}{240.35}} = \sqrt{\frac{88.0901}{240.35}} = 0.6054$$

$$\phi = \phi'\sqrt{n} = 0.6054\sqrt{10} = 1.91$$

- For $\phi = 1.91$, df_t = 4, and df_e = 45, power can be determined by the tables of the noncentral F distribution, once we interpolate (or round off to the nearest value)

  - For $\phi = 1.8$, df_t = 4, df_e = 30, and α = .05, $F_{4,30,1.8}$ = .14 (this is β)

  - Power = $(1 - \beta)$ = .86, probability of detecting *true* differences among the means

- If we wanted to calculate the required sample sizes for a power = .80

  - If power = .80, β = .20; we need to find the corresponding value of $\phi = ?$

  - For df_t = 4, df_e = 30, α = .05, and β = .20, $\phi \approx 1.68$

  - Given $\phi = \phi'\sqrt{n}$, then $n = \frac{\phi^2}{\phi'^2} = \frac{(1.68)^2}{(.6054)^2} = 7.70$

  - We would need ≈ 8 subjects per group to have an 80% chance of rejecting H_0 if it is false

- ## 16. 7 Calculating Power for Factorial ANOVA

- It is a straightforward extension of the power analysis for ANOVA

$$\phi_\alpha' = \sqrt{\frac{\frac{\sum(\mu_j - \mu)^2}{j}}{\sigma_e^2}} \text{ and } \phi_\alpha = \phi'_\alpha\sqrt{nk}$$, for power of effect A

$$\phi_\beta' = \sqrt{\frac{\frac{\sum(\mu_k - \mu)^2}{k}}{\sigma_e^2}} \text{ and } \phi_\beta = \phi'_\beta\sqrt{nj}$$, for power of effect B, and

$$\phi_{\alpha\beta}' = \sqrt{\frac{\frac{\sum(\mu_{kj} - \mu)^2}{jk}}{\sigma_e^2}} \text{ and } \phi_{\alpha\beta} = \phi'_{\alpha\beta}\sqrt{n}$$, for power of the interaction

- ***Power*:** Probability a test rejects H_0 (depends on $\mu_1 - \mu_2$)

    - H_0 True: Power = P(Type I error) = α

    - H_0 False: Power = 1–P(Type II error) = 1– β

- So, once H_1 is specified, we can determine β (p of erroneously retaining H_0) and the probability of 1– β (p of correctly rejecting H_0).

- *Example:*

    - H_0: $\mu = 138$

    - H_A: $\mu = 142$

    - It is assumed that the population distribution in either situation has $\sigma = 20$

    - A sample $n = 100$ is drawn at random so $\sigma_M = 20/\sqrt{100} = 2$.

    - Decision Rule: Reject H_0 at $\alpha = 0.05$ significance level if the sample result falls among the highest 5% of means in a normal distribution; Otherwise retain H_0 (reject H_1).

    $\alpha = p(\text{reject } H_0 \mid \mu = 138)$ or $p(\text{reject } H_0 \mid H_0)$

    $\beta = p(\text{accept } H_0 \mid \mu = 142)$ or $p(\text{accept } H_0 \mid H_1)$

    - Rejection region must be bounded by a z_M such as

        $F(z_M) = .95$, or $1- F(z_M) = .05$.

    - From the tables, $z_M = 1.65$.

    - If $x = 142$, then

        $z_M = (142 -138)/2$

    - And the critical value of x forming the boundary of the rejection region is

        $x = 138 + 1.65\ \sigma_M$

        $= 138 + 3.30 = 141.30$

- The question is, To what z_M score would this critical value of 141.3 correspond if H_1 were true?

$$z_M = (141.3 - 142)/2 = -.35$$

- In a normal distribution, $F(-.35) = .36$, approximately, so we can determine $\beta = .36$. Thus, the two error probabilities are

$$\alpha = .05, \text{ and } \beta = .36$$

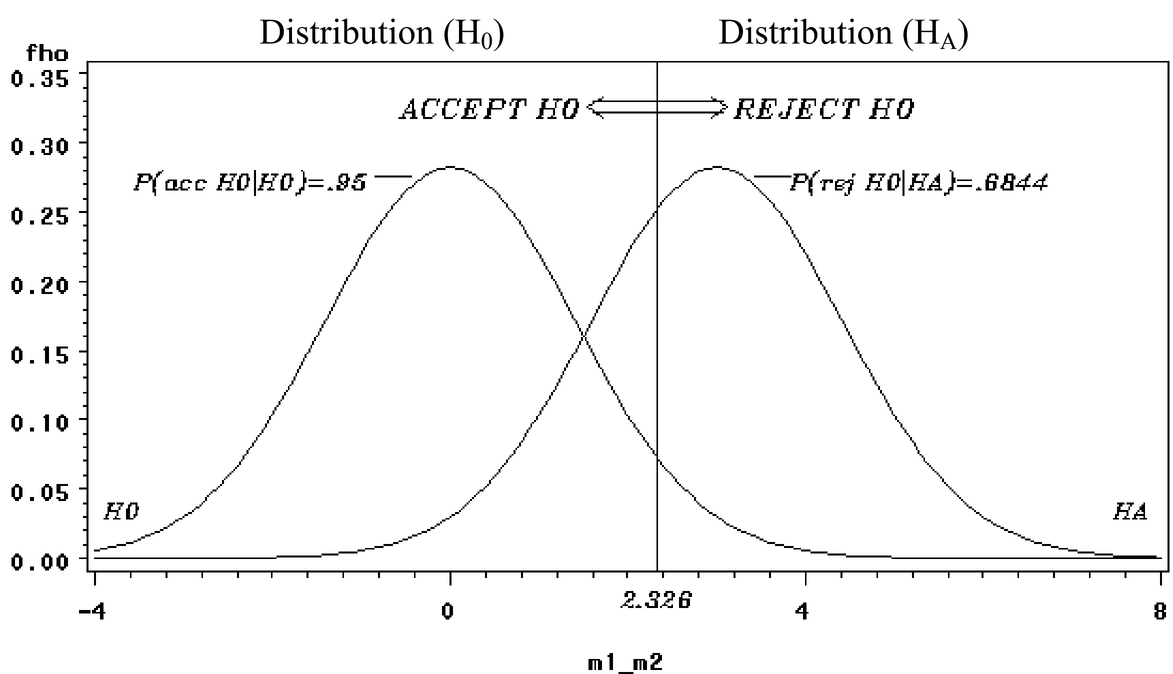
- Thus, the power of the test = $1 - \beta = .64$

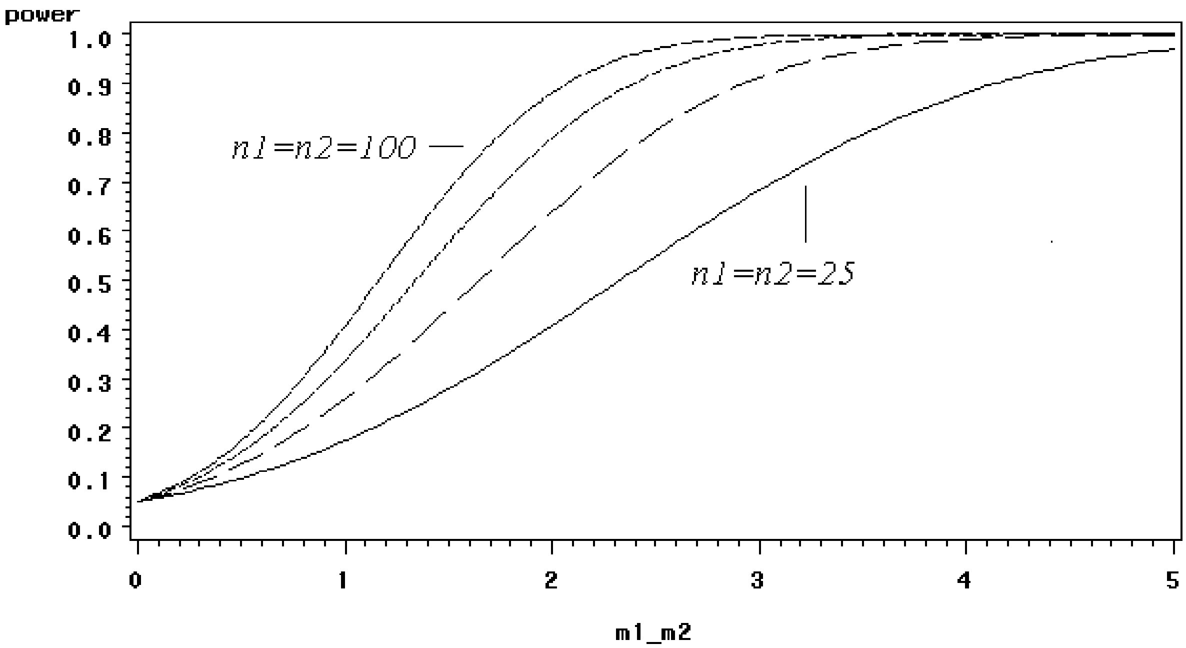
- All else being equal (*Ceteris Paribus*):

  - As sample sizes increase, power increases

  - As population variances decrease, power increases

  - As the true mean difference increases, power increases





- Power Curves for group sample sizes of 25, 50, 75, and 100, and varying true values $\mu_1 - \mu_2$ with $\sigma_1 = \sigma_2 = 5$.

  - For given $\mu_1 - \mu_2$, power increases with sample size

  - For given sample size, power increases with $\mu_1 - \mu_2$

- ***Sample Size Calculations for Fixed Power***:

    - **Goal:** Choose sample sizes to have a favorable chance of detecting a *clinically meaning difference*

    - **Step 1:** Define an important difference in means:

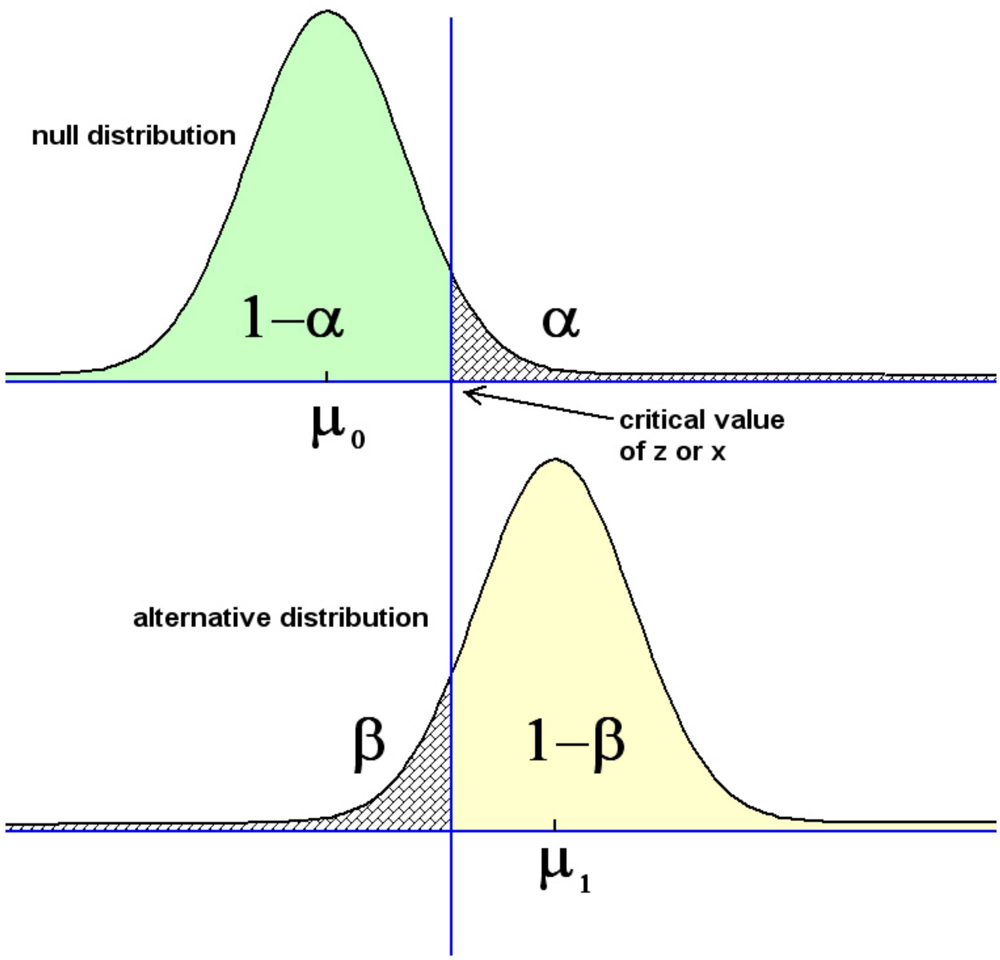
        - **Case 1:** σ approximated from prior experience or pilot study - difference can be stated in units of the data

        - **Case 2:** σ unknown - difference must be stated in units of standard deviations of the data

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

- **Step** 2: Choose the desired power to detect the meaningful difference (1- β, typically at least .80 in clinical trials). For 2-sided test:

$$n_1 = n_2 = \frac{2\left(z_{\alpha/2} + z_\beta\right)^2}{\delta^2}$$


null distribution
1−α
α
μ0
critical value
of z or x
alternative distribution
β
1−β
μ1

- Power to detect a main effect (ES = .20 - .50)