Lecture 3: Sampling Distributions and Hypothesis Testing

• 3. 1 Populations, Parameters, and Statistics

- ***Population*** is the totality of potential units for observation

- ***Population distribution*** refers to the distribution of a random variable resulting from measuring and assigning values to particular observations from a population

    - The frequency distribution based on some large but finite number of cases (but assuming it's close enough to the population)

    - This distribution has a mathematical form with mean μ, variance σ^2, and all the features of any distribution

- Population values such as μ and σ^2 are ***parameters*** of the population (or ***true values***)

- Sample values such as $\bar{x}$ and S^2 are ***statistics*** of the sample

- We use statistics to make inferences about populations

• 3. 2 Sampling Distributions

- Typically, observations in a sample are drawn at the same time from the same population and a statistic is associated with the sample

- The goal is to determine the distribution of values of such statistic *across all possible samples* of N observations from the population

- Such a theoretical distribution is called a ***sampling distribution***

- Example:

    - Consider a population of 6,000 persons, and we are interested in IQ scores

    - We draw a sample of $N = 30$ and obtain a mean $\bar{x} = 118$

    - If we want to know how close is that sample – in, say, the mean – to the population we need to know the variability; that is, we need to draw more samples

    - How many? In theory, *all possible* combinations of samples of $N = 30$

    - If we do that, we create a sampling distribution of the parameter (in this case the mean)

    - Given that we now know the variability, we can make inferences

- A sampling distribution is a distribution of a statistic over all possible samples

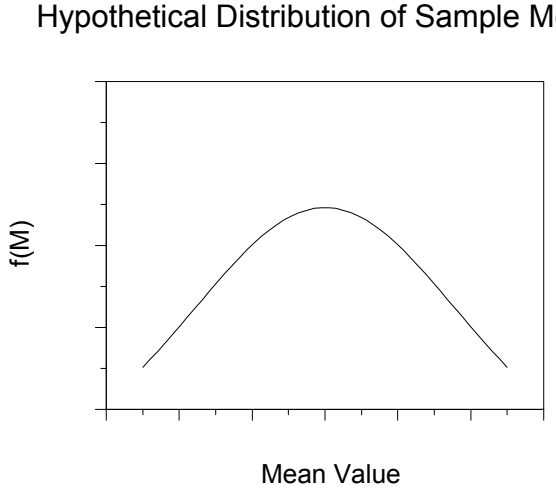
- To get a sampling distribution:

    - Take a sample of size *N* (a given number like 5, 10, or 1000) from a population

    - Compute the statistic (e.g., the mean) and record it

    - Repeat 1 and 2 a lot (infinitely)

    - Plot the resulting ***sampling distribution,*** a distribution of a statistic over repeated samples



- More formally, a ***sampling distribution*** is a theoretical probability distribution that shows the relation between the possible values of a given statistic and the probability associated with each value, for all possible samples of size *N* drawn from a particular population

- A sampling distribution implies that the frequency distribution is not of individual observations ($x_1, x_2, \ldots, x_n$) but of statistics ($\bar{x}, \bar{x}_2, \ldots, \bar{x}_n$) from an infinite number of random samples of size *N*

- When a sampling distribution is plotted, the *x*-axis represents the values of the sample statistic and the *y*-axis represents the probability density

- Let *G* be a sample statistic, if the sampling distribution of *G* is discrete, its expectation or mean over all values $G = g$ is

$$E(G) = \sum_g gp(g),$$

or the sum being taken over all values *g* that *G* can assume

- Similarly, the variance of the sampling distribution for any statistic *G* is (the expected value of the squared difference):

$\sigma^2_G = E(G - \mu_G)^2$, or

$\sigma^2_G = E(G^2) - [E(G)]^2$

- The standard deviation σ_G is the square root of σ^2_G. This σ_G represents the extent to which sample G values depart from the expectation and is called the ***standard error*** of the statistic G

• 3. 3 Sampling Statistics as Estimators

- A sample of cases drawn from a population contains information about the population distribution and its parameters

- Some of this information is reflected in the statistic computed from the sample

- The goal in inferential statistics is to use the value of the statistic to infer the value of the corresponding population parameter. This is called ***point estimation***

- Some statistics are better than others – contain more information – to infer population parameters

- An ***estimator*** is a formula or method for combining the values occurring in the data (e.g., random variable) in order to infer population parameters

- An ***estimate*** of the population parameter is the value that results from the application of the formula

• 3. 4 Properties of Estimators

- ***Properties of estimators*** are the criteria to judge how effective a statistic is in inferring population parameters

- A statistic G is said to be an ***unbiased*** estimator of θ when the expected value of G over all possible samples is exactly θ, or

  $E(G) = \theta$

  - The sample mean $\bar{x}$ is an unbiased estimator of the population mean μ

  - The sample variance S^2 is a biased estimator of the population mean σ^2

- A statistic G is ***consistent*** when the probability of G of approaching θ increases with N

  - The sample mean and the sample variance are consistent estimators

- A statistic G is ***efficient relative*** to a statistic H, both being unbiased estimators of θ, when

  σ^2_H / σ^2_G,

  where σ^2_H and σ^2_G are the variances of the sampling distributions of H and G, respectively

  - The more efficient estimator has the smaller sampling variance and, hence, smaller sampling error

  - Given a population with a normal distribution, the $\bar{x}$ is relatively more efficient than the Md, given the sample size N

- A statistic G is a ***sufficient*** estimator of the parameter θ if it contains all the information in the data about θ

  - When a statistic is a sufficient estimator, there is no additional information in the data (not included in G) that can improve our estimate of θ

- The sample mean and the sample variance (corrected) are examples of estimators that feature most (if not all) of these properties

• 3. 5 The Sample Distribution of the Mean

- In general, the sample mean is unbiased, consistent, and in many circumstances, efficient relative to other statistics and, together with σ^2, sufficient

  - The sample mean is an unbiased estimator of μ. The expected value of the sample mean is the population mean μ

    $E(\bar{x}) = \mu$

    For any N, the sample mean $\bar{x}$ is

    $\bar{x} = \frac{(x_1 + x_2 + ... + x_N)}{N}$, so that

    $E(x) = \frac{E(x_1 + x_2 + ... + x_N)}{N}$, or

    $E(x) = \frac{[E(x_1) + E(x_2) + ... + E(x_N)]}{N}$

    If for any given observation, $E(x) = \mu$, then

    $E(x) = \frac{(\mu + \mu + ... + \mu)}{N} = \frac{N\mu}{N} = \mu$

    The mean of the sampling distribution of means is the same as the population mean

- But the sampling distribution of the mean will not exactly match the population distribution

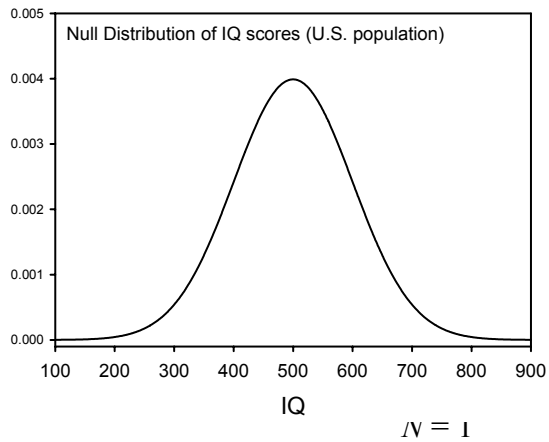
  - The variance of the sampling distribution of means σ^2_M will be smaller than σ^2 by a factor of $1/N$, or

    $\sigma^2_M = \frac{\sigma_2}{N}$

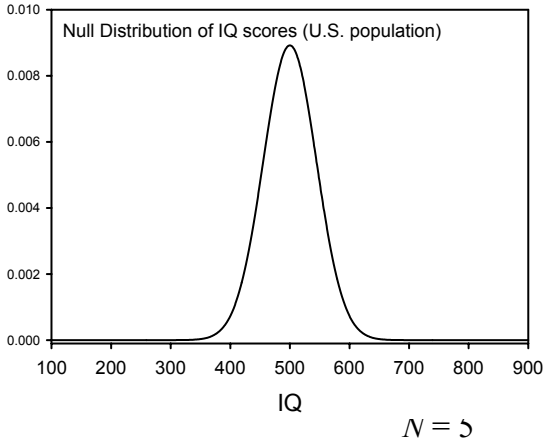
  - The larger the sample size, the smaller the variance of the sampling distribution of the means and, hence, the more probable it is that $\bar{x}$ matches μ

    Example: $\sigma^2 = 10$

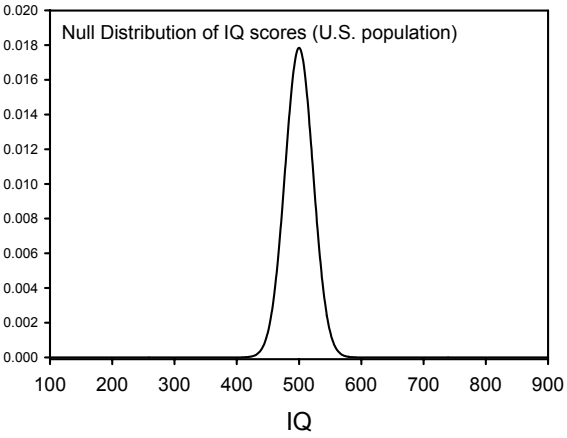
    If $N = 1$, $\sigma^2_M = 10$; If $N = 2$, $\sigma^2_M = 5$; If $N = 10$, $\sigma^2_M = 1$; If $N = 100$, $\sigma^2_M = .1$. As sample size increases, the sampling distribution of the mean becomes narrower, thus getting closer to the population mean



$N = 1$



$N = 5$



$N = 20$

As sample size increases, the variance (and standard error) decreases

- The standard deviation, or standard error, of the mean is

$$\sigma_M = \sqrt{\sigma_{2M}} = \frac{\sigma}{\sqrt{N}}$$

- This represents how far on average is the sample mean from the population mean. This is a value derived theoretically. We cannot know it jut by looking at a single sample

- What happens when a sample mean is put into standard form?

$$z_M = \frac{\bar{x} - \mu}{\sigma_M} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$$

- For large values of $\bar{x}$ (in absolute terms), in relation to μ and σ_M, the departure from μ is greater and, thus, the probability that $\bar{x}$ resembles μ is smaller

- Example: $\mu = ?$, $\sigma = 5$, $\bar{x} - \mu = 10$, and $N = 2$.

$\sigma_M = \frac{5}{\sqrt{2}} = 3.536$, and $z_M = \frac{10}{3.536} = 2.828$,

Is this result likely?

What about if $N = 200$,

$\sigma_M = \frac{5}{\sqrt{200}} = .3535$, so $z_M = \frac{10}{.3535} = 28.28$

- The sample mean is a minimum variance estimator. Any other statistic used to infer the population mean would produce greater variability in the sampling distribution of the mean (e.g., the standard deviation of the medians would be larger than the SD of the means)

• 3. 6 Confidence Intervals for the Mean

- Confidence intervals are areas of precision – or uncertainty – that accompany any point estimation

- Confidence intervals for the mean can be constructed because:

  - The sampling distribution of the mean is normal when the population is normal, and

  - When N is large, the sampling distribution of the mean is normal, irrespective of the population distribution

- In any normal distribution of sample means, with population μ and standard deviation σ_M, ***over all samples of size N, the probability is .95 for the event***

  $-1.96\sigma_M \leq \bar{x} - \mu \leq 1.96\sigma_M$.

  - approximately 95% of all possible sample means from the population must lie within 1.96 standard errors to either side of the true mean

- this can also be written as

$$\bar{x} - 1.96\sigma_M \leq \mu \leq \bar{x} + 1.96\sigma_M,$$

which states that, over all possible samples, the probability is about .95 that the true mean, μ is included between $\bar{x} - 1.96\sigma_M$ and $\bar{x} + 1.96\sigma_M$. This range of values is called the ***95% confidence interval for the mean***, whereas the boundaries are called ***95% confidence limits***

- Similarly, the 99% confidence interval for the mean is

$$\bar{x} - 2.58\sigma_M \leq \mu \leq \bar{x} + 2.58\sigma_M$$

- And the 100% confidence interval for the mean is

$$\bar{x} - \infty\sigma_M \leq \mu \leq \bar{x} + \infty\sigma_M$$

- When σ is unknown and the sample size N is reasonably large, confidence intervals can be found that are approximately correct by using $s/\sqrt{N}$ instead of σ_M. For example, for a large sample size N, a 95% confidence interval can be written as follows

$$\bar{x} - (1.96)(\, s/\sqrt{N}\,) \leq \mu \leq \bar{x} + (1.96)(\, s/\sqrt{N}\,)$$

• 3. 7 Confidence Intervals and Inference

- Imagine that the average μ in a particular test over the years for all undergraduate UCD students equals 115. A sample of $n = 30$ is taken at random from the students to determine whether the current average has changed. The sample mean turns out to be 118, with a standard deviation S equal to 10. What can we infer from this sample?

- First, the standard error of the mean σ_M is unknown but it can be estimated

$$s/\sqrt{N} = 10/\sqrt{30} = 1.83$$

- Second, the limits of the 95% confidence interval can be obtained

$$118 - (1.96)(1.83) \text{ and } 118 + (1.96)(1.83), \text{ or}$$

$$(114.4 , 121.6)$$

- And the limits of the 99% confidence interval are

$$118 - (2.58)(1.83) \text{ and } 118 + (2.58)(1.83), \text{ or}$$

$$(113.3 , 122.7)$$

Here, we increase our confidence but lose precision

In any case, both intervals contain the population mean $\bar{x} = 115$. Because the probability is .95 (or .99) that the population value of the mean is covered in the intervals, the probability is only .05 (or .01) that the sample does not represent a population with a mean inside this range of values. Thus, one can conclude that the current average is not different from the historic values

• 3. 8 Hypothesis Testing: *Decide How to Decide*

- One has to make decisions even when it's unsure (e.g., school, marriage, doctor, jobs)

- Statistics provides an approach to decision making under uncertainty. Sort of decision making by choosing the same way you would bet. Maximize expected utility (subjective value)

- Comes from agronomy, where they were trying to decide what strain to plant

- Because of uncertainty (have to estimate things), we will be wrong sometimes. But, the idea is to be thoughtful about it; how many errors of what kinds? What are the consequences?

- Statistics allows us to calculate probabilities and to base our decisions on those. We choose (at least partially) the amount and kind of error

- Hypothesis testing is the branch of statistics that helps one determine whether or not your theory about something you observed can be rejected

- The process of comparing two hypotheses in the light of sample evidence is called a ***statistical test***. The competing hypotheses often are stated in terms of one or more parameters of the population distribution

• 3. 9 Hypothesis Testing

- Goal: Make statement(s) regarding unknown population parameter values based on sample data

- Elements of a hypothesis test:

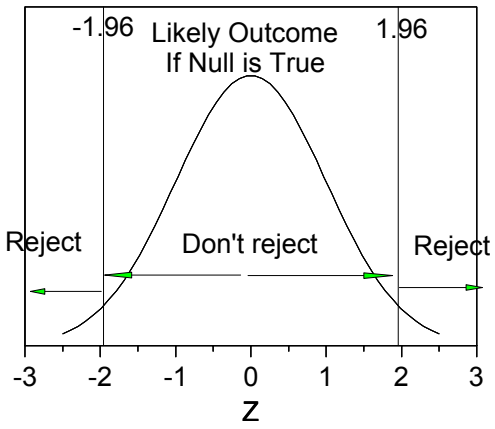
  - ***Null hypothesis*****:** Statement regarding the value(s) of unknown parameter(s). Typically will imply no association between explanatory and response variables in our applications (it typically contains an equality)

  - ***Alternative hypothesis****:* Statement contradictory to the null hypothesis (it typically contains an equality)

  - ***Test statistic:*** Quantity based on sample data and null hypothesis used to test between null and alternative hypotheses. Given some assumptions, the sampling distribution of the statistic is specified

  - ***Rejection region:*** Values of the test statistic for which we reject the null in favor of the alternative hypothesis (and associated p-value)

  - The ***sample*** (data) itself is obtained. If calculated value of statistic falls in rejection region, then reject H_0. If not, retain H_0



- ***Example***: The mean of a particular population μ is 75 with a $\sigma = 10$

H_0: $\mu = 75$

H_1: $\mu \neq 75$

- A sample of $n = 25$ independent observations is collected

- Given n, H_0, and assumptions (i.e., normality), the sampling distribution of the mean $\bar{x}$ is known: a normal distribution with $\mu = 75$ and a standard error $\sigma_M = 10/\sqrt{25} = 2$

- The researcher decides to reject H_0 if $\bar{x}$ departs extremely from the expected population value $\mu = 75$ in either direction, and defines the rejection regions to contain only 5% of all possible sample results when the hypothesis H_0 is actually true ($z = \geq \left| 1.96 \right|$)

- These z values give a total probability of .025 + .025 = .050 for the combined intervals

- Critical values are calculated (a value of the sample mean that lies exactly on the boundary of a rejection region)

$-1.96\sigma_M + \mu = -1.96(2) + 75 = 71.08$

$+1.96\sigma_M + \mu = +1.96(2) + 75 = 78.92$

- A sample is drawn with $\bar{x} = 79$. Its z-value is calculated as

$z_M = (79 - \mu)/\ \sigma_M = (79\text{-}75)/2 = 2$

- Because 2 > 1.96 (it falls within the rejection region), the experimenter says that the sample result is ***significant beyond the 5% level***. That is, less than 5% of the samples should show results as different from expectation if H_0 is actually true. This is not tenable – under the experimenter's decision rules – and, hence, H_0 is rejected

- ## 3. 10 Type I and Type II Error

- ***Type I Error:*** What if one is wrong? Generally, one wants to make sure not to reject the Null Hypothesis when it is true. This is called a ***Type I error*** (α)

    - Traditionally α = *P*(Type I error) = 0.05

- *Type II error:* The reverse, however, is also possible in that one does not reject the Null hypothesis when it's false. This is called a ***Type II error***

    - Traditionally an important difference (Δ) is assigned and sample sizes chosen so that:

$\beta = P(\text{Type II error} \mid \mu_1 - \mu_2 = \Delta) \leq .20$

		True State	
		H_0 True	H_0 False
Decision	Retain H_0	Correct	Type II Error
	Reject H_0	Type I Error	Correct

$\alpha = P(\text{Type I Error})$

$\beta = P(\text{Type II Error})$

Goal: Keep α and β reasonably small

• 3. 11 Power of a Statistical Test

- ***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)

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

  $\beta = p$(accept $H_0 \mid \mu = 142$) or p(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 $\bar{x} = 142$, then

    $z_M = (142 - 138)/2$

  - And the critical value of $\bar{x}$ forming the boundary of the rejection region is

    $\bar{x} = 138 + 1.65\ \sigma_M$

    $= 138 + 3.30 = 141.30$

- 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 = $1 - \beta = .64$

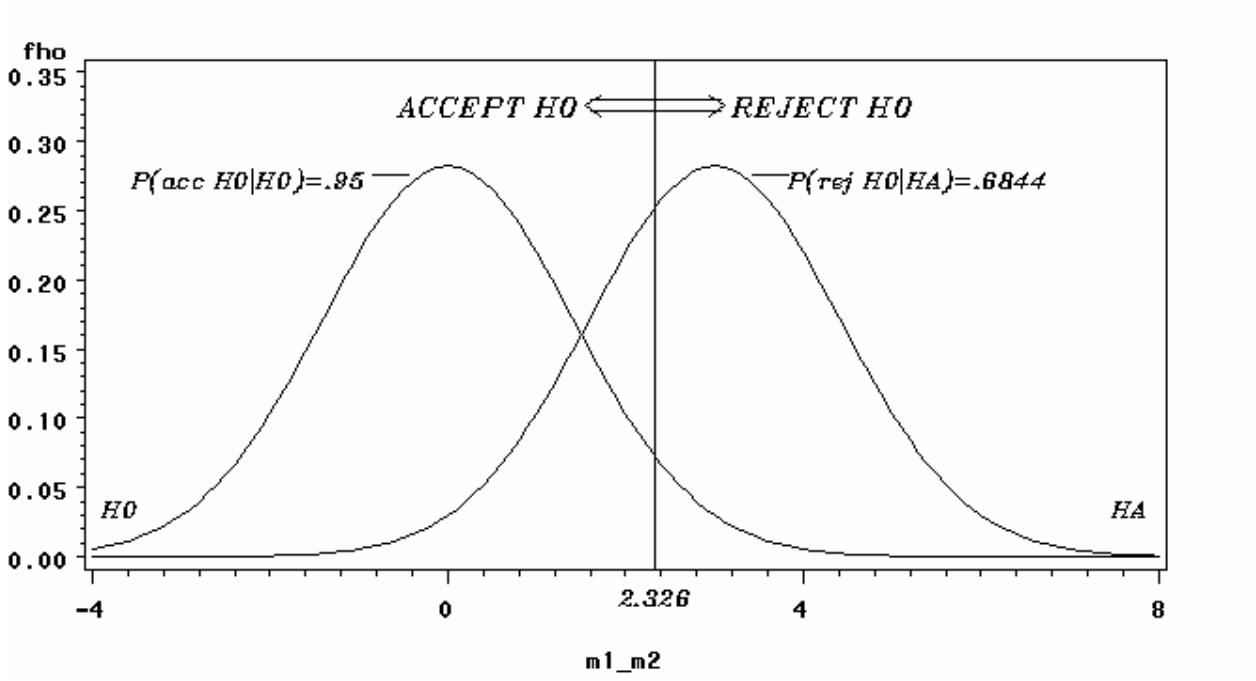
- All else being equal (*Ceteris Paribus*):

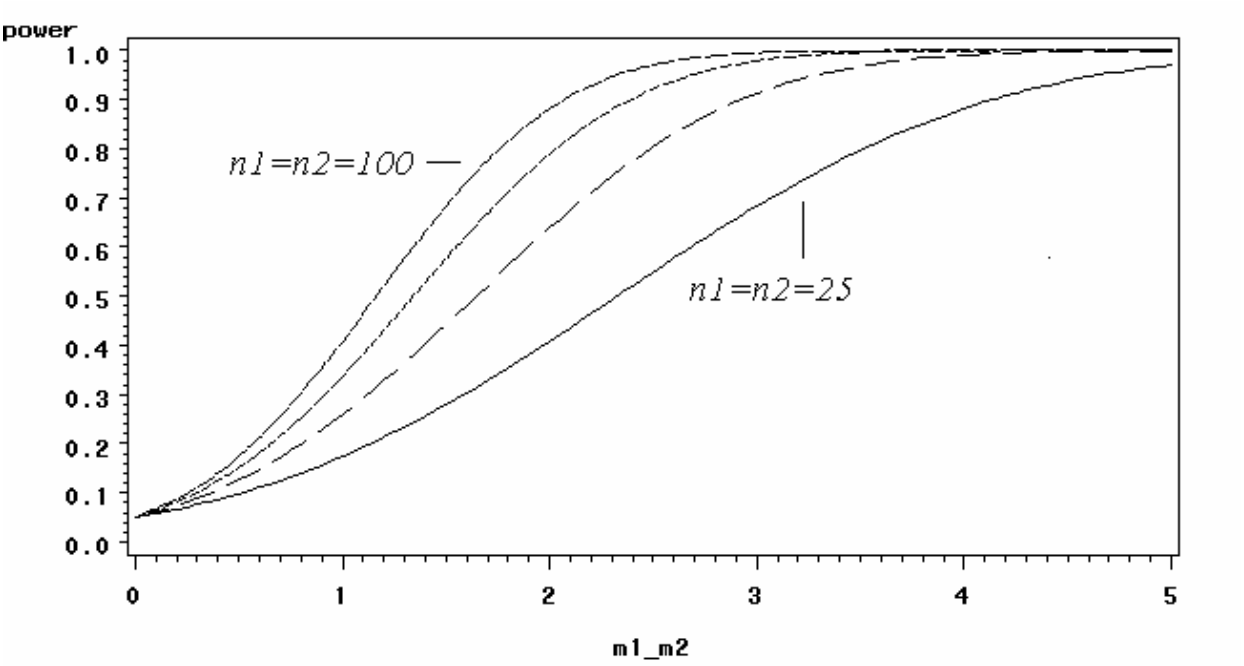
  - As sample sizes increase, power increases

  - As population variances decrease, power increases

  - As the true mean difference increases, power increases

Distribution (H_0) Distribution (H_A)





- 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 ($1-\beta$, typically at least .80 in clinical trials). For 2-sided test:

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