Introduction to Statistical Methods
SOC-GA 2332

Lecture 2: Statistical Inference

Siwei Cheng

NEW YORK UNIVERSITY

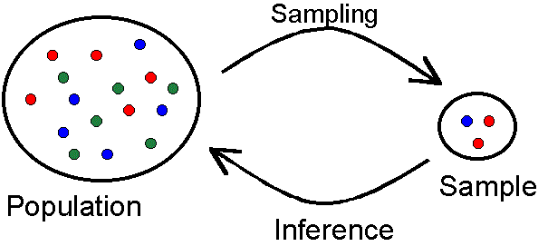
Lecture Outline

- Sampling distribution
- Estimation(Point estimate and confidence interval)
- Hypothesis Testing (Significance tests for a mean)

Key Concepts from Previous Lecture: Sample and Population

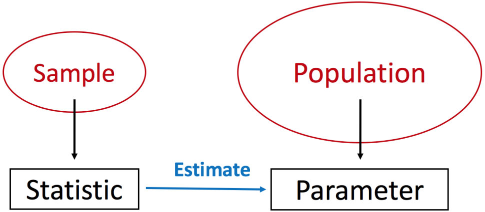
Key Concepts from Previous Lecture: Sample and Population

- Population: total set of subjects of interest in a study.
- Sample: subset of the population on which we collect data.
- Goal of quantitative data analysis: learn about the population using the sample data.



Key Concepts from Previous Lecture: Descriptive Statistics and Inferential Statistics

- A parameter is a numerical summary of the population.
- A statistic is a numerical summary of the sampled data.
- We make statistical inferences about the population based on properties of statistics from the sample.



Key Concepts from Previous Lecture

- Sample mean (a statistic):

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- Population mean (a parameter):

$$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Sample variance (a statistic):

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

- Population variance (a parameter):

$$\sigma^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{N}$$

Moving from descriptive to inferential statistics...

Sampling Distribution

- A sampling distribution describes the distribution of a statistic, such as a sample mean or variance, of that statistic was measured <u>over a range of different samples within a population</u>.
- Recall that a statistic is a random variable. The (imaginary) samples are drawn randomly.

Sampling Distribution

- A sampling distribution describes the distribution of a statistic, such as a sample mean or variance, of that statistic was measured over a range of different samples within a population.
- Recall that a statistic is a random variable. The (imaginary) samples are drawn randomly.
- For example, consider the sampling distribution of *the sample mean of earnings*...
- We can collect **many repeated samples** from the population, calculate the mean of earnings in each sample, and then describe the probability distribution of the sample mean of earnings over the range of these different samples.

Sampling Distribution: An Illustration

- Let's do an experiment!

Sampling Distribution: An Illustration

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.

Sampling Distribution: An Illustration

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.
- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.

Sampling Distribution: An Illustration

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.
- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.
- Now, I am going to collect our sample $\#1$...

Sampling Distribution: An Illustration

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.
- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.
- Now, I am going to collect our sample #1...
- And sample #2...

Sampling Distribution: An Illustration

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.
- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.
- Now, I am going to collect our sample $\#1$...
- And sample $\#2$...
- And sample $\#3$...

Sampling Distribution: An Illustration

- Let's do an experiment!

- Please randomly write down an integer from 1 to 10.

- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.

- Now, I am going to collect our sample #1...
- And sample #2...
- And sample #3...
- I can calculate the mean of every sample, call it $\bar{X}_k$, where k is the index for all the samples.

Sampling Distribution: An Illustration

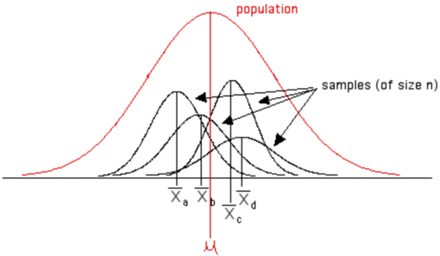
- Let's do an experiment!

- Please randomly write down an integer from 1 to 10.

- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.

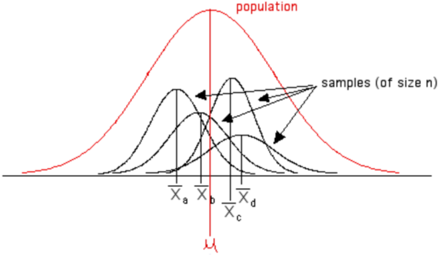
- Now, I am going to collect our sample #1...
- And sample #2...
- And sample #3...
- I can calculate the mean of every sample, call it $\bar{X}_k$, where k is the index for all the samples.
- Imagine that I have taken an infinite number of such samples.
- The sampling distribution of the populaiton mean is simply the distribution of the means across all my imaginary samples.

Sampling Distribution



- In other words, we can think of the mean of a given sample, $\bar{y}$, as a variable with a value that varies from sample to sample around the population mean μ.
- So the sampling distribution of $\bar{y}$ has mean μ.

Sampling Distribution



- In other words, we can think of the mean of a given sample, $\bar{y}$, as a variable with a value that varies from sample to sample around the population mean μ.
- So the sampling distribution of $\bar{y}$ has mean μ.
- But how about the standard deviation of the sampling distribution? We use a new concept called the standard error.

Sampling Distribution and Standard Error

- The standard deviation of a sampling distribution is called the standard error.
- For example, the sampling distribution of $\bar{y}$ in a sample of n observations has standard error:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

- But since we typically don't know the population standard deviation σ, we estimate it with the sample standard deviation s:

$$\hat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}}$$

- Over large samples, the sampling distributionof $\bar{y}$ is approximately normal (even though sample or population distribution may not be).

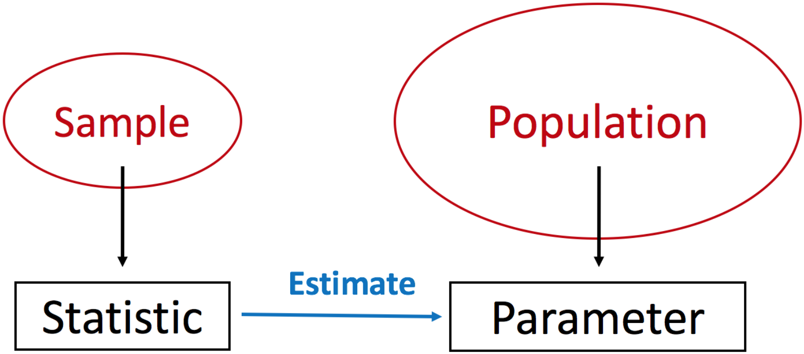
Comparing Three Types of Distributions

- Population distribution: described by parameters (usually unknown) such as mean (μ) and standard deviation (σ).
- Sample distribution: described by sample statistics such as sample mean ($\bar{y}$) and sample standard deviation (s).
- Sampling distribution: probability distribution of a sample statistic, such as sample mean. (We "imagine" the sampling distribution" in our heads.)
- Sampling distribution is important, because it determines the probability that a statistic falls within certain distance of population parameter.
- The sampling distribution of a sample mean equals population mean (μ), and the standard deviation of the sampling distribution (call STANDARD ERROR) is ($\hat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}}$).

Comparing Three Types of Distributions

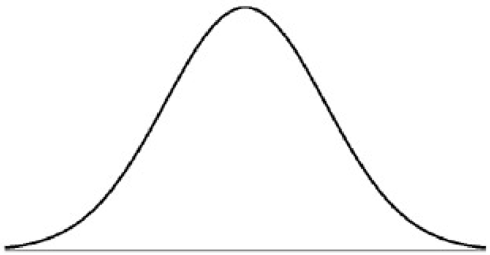
Statistical Inference I: Point Estimate and Confidence Interval

Estimation



Estimation

- Point estimate: a single number (the best guess)
- Interval estimate: an interval around the point estimate (confidence interval)

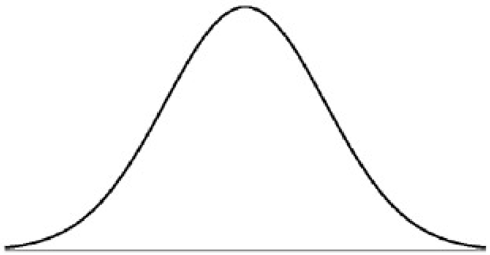


Point Estimator: Unbiasedness and Consistency

- In principle, any statistic computed from the sample can be used as and estimator.
- *For example, we can even define "the value of Y for fifth person in your sample" as an estimator for the population average of Y. But of course that's not a very good estimator!*
- So how do we choose estimators that are "good"?
- Now, let's consider some desirable properties of a point estimator: **unbiasedness** and **consistency**.

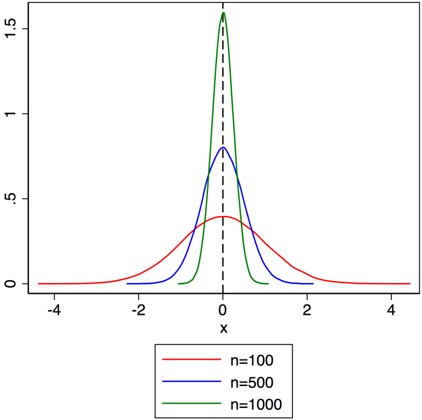
Point Estimator: Unbiasedness and Consistency

- Unbiased estimator: the sampling distribution of a statistic **centers around** the population parameter.
- Example: the sample mean is an unbiased estimator of population mean.



Point Estimator: Unbiasedness and Consistency

- Consistent estimator: The estimator **approaches** the population parameter as the sample size n goes to infinity.



Point Estimator: Unbiasedness and Consistency

Bottom line:

- An estimator is **unbiased** if the mean of this statistic over infinitely many repeated random samples approaches the true parameter.
- An estimator is **consistent** if when the sample size n approaches infinity, this estimator approaches the true parameter.

Point Estimator: Unbiasedness and Consistency

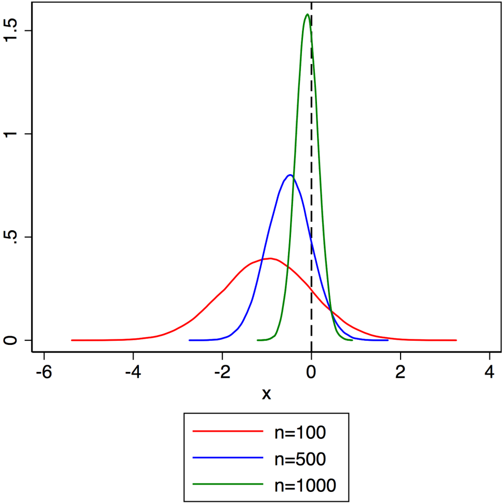
- Example 1: Sample mean is an unbiased estimator of population mean. Sample variance is an unbiased estimator of population variance.

Point Estimator: Unbiasedness and Consistency

- Example 1: Sample mean is an unbiased estimator of population mean. Sample variance is an unbiased estimator of population variance.
- Example 2: Consider a sample $x_1, x_2, x_3, ...x_n$. Suppose a research uses *the third observation in this sample*, x_3, as an estimator of the population mean.
- Is x_3 an unbiased estimator of the population mean?
- Is x_3 a consistent estimator of the population mean?

Point Estimator: Unbiasedness and Consistency

- Example 3: a consistent but not unbiased estimator:



(Weak) Law of Large Numbers

(Weak) Law of Large Numbers

Suppose $x_1, x_2, ..., x_n$ are i.i.d. sample drawn from a population distribution with mean μ, then: $\bar{x_n} \xrightarrow{p} \mu$

- **Intuition**: it becomes increasingly unlikely for $\bar{x_n}$ to be away from μ as n gets large.
- **Proof:**

  Assume that $\sigma < \infty$ (which is not necessary but this simplifies the proof), Using Chebyshev's inequality, we have:

  $P(|\bar{x_n} - \mu| > \epsilon) \leq \frac{Var(\bar{x_n})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$,

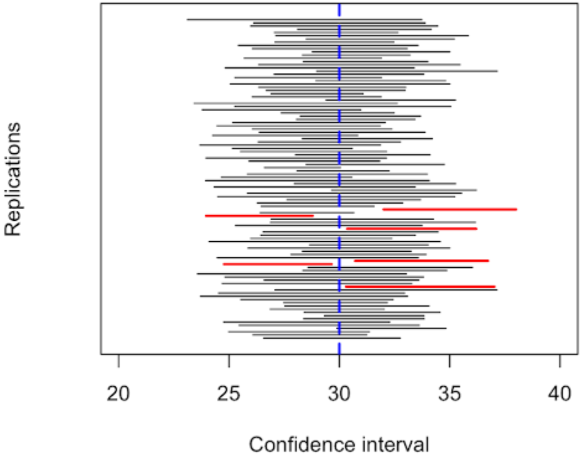
  which tends to 0 as $n \to \infty$.

From Point Estimator to Confidence Interval

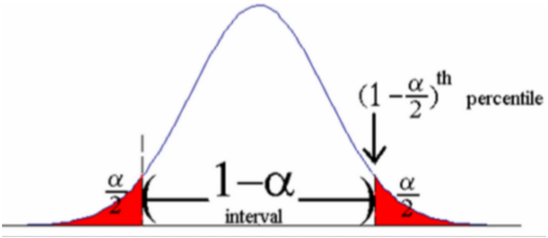
From Point Estimator to Confidence Interval

- Earlier, we have introduced the sampling distribution. So we know that there is some uncertainty on where the parameter falls.
- A confidence interval for a parameter gives us an interval that is believed to cover the parameter with a certain probability. This probability is called the confidence level.
- For example, a 95% confidence interval will cover the population parameter with probability 0.95.
- That is, there is a 5% chance that this interval “misses” the population parameter.

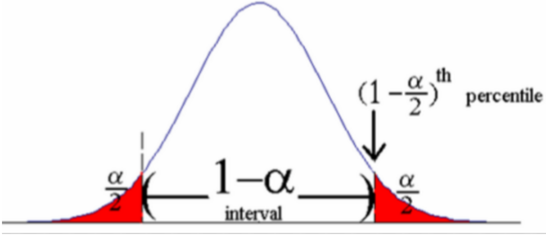
A confidence interval covers the true parameter with a certain probability



Constructing a Confidence Interval



Constructing a Confidence Interval



- Note that the law of large numbers says that the distribution of sample mean ($\bar{x}_n$) piles up near μ.
- But this isn't enough to help us approximate probability statement about $\bar{x}_n$, because we don't know the probability distribution of $\bar{x}_n$.
- So, we next introduce the **Central Limit Theorem**.

Central Limit Theorem

Central Limit Theorem

Suppose $x_1, x_2, ..., x_n$ are sample drawn from a population distribution with mean μ and variance σ^2. Let $\bar{x}$ be sample mean. Then:

$$\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

Or, equivalently,

$$\bar{x} \xrightarrow{d} N(\mu, \frac{\sigma^2}{n}).$$

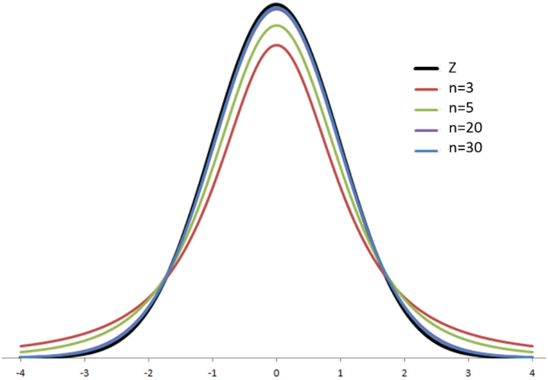
Central Limit Theorem

- CLT holds regardless of the underlying distribution of X in the population.
- CLT says that the probability distribution can be approximated using a Normal distribution.
- The CLT relies on large samples, what we refer to as asymptotic results.
- Note that it is the probability distribution that we are approximating, not the random variable itself.

The t-distribution

- We do not observe the population variance σ^2, and we can replace σ^2 with sample variance s^2.
- Statisticians found that a statistic called the *t-statistic*, computed as $\frac{\sqrt{n}(\bar{x}-\mu)}{s}$, follows a **t-distribution**.
- The shape of the t-distribution depends only on sample size n. Specifically, the *degree of freedom* of a t-distribution is equal to n - 1.
- As n goes to infinity, the t-distribution approaches the normal distribution.

The t-distribution



Constructing a Confidence Interval

- How to construct a confidence interval? Let's review the mean and standard error of the sampling distribution of sample mean, which we have learned earlier today (and in Week 1):
- The **sample mean** $\hat{\mu}$, and we know that if we take many many sample, the sample mean will center around population mean.
- The **standard error of sample mean** is: $\hat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}}$.

Constructing a Confidence Interval

- How to construct a confidence interval? Let's review the mean and standard error of the sampling distribution of sample mean, which we have learned earlier today (and in Week 1):
- The **sample mean** $\hat{\mu}$, and we know that if we take many many sample, the sample mean will center around population mean.
- The **standard error of sample mean** is: $\hat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}}$.
- The **upper bound** of 95% CI: $\hat{\mu} + t_{0.025,n-1} \cdot \hat{\sigma}_{\bar{y}}$
- The **lower bound** of 95% CI: $\hat{\mu} - t_{0.025,n-1} \cdot \hat{\sigma}_{\bar{y}}$
- The critical value $t_{0.025,n-1}$ is chosen such that the probability that a t-statistics with degree of freedom $n-1$ is greater than $t_{0.025,n-1}$ equals 0.025 (half of 0.05). n-1 is the degree of freedom.

Constructing a Confidence Interval

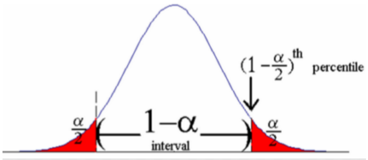
- In a sample of n observations, the confidence interval that *covers the population mean with probability 95%* is given by:

$$\hat{\mu} \pm t_{0.025,n-1} \cdot se$$

- i.e. Point estimate $\pm$ Margin of error

- Or, we can say that:

$$P(\hat{\mu} - t_{0.025,n-1} \cdot \hat{\sigma}_{\bar{y}} \leq \mu \leq \hat{\mu} + t_{0.025,n-1} \cdot \hat{\sigma}_{\bar{y}}) = 0.95$$



Constructing a Confidence Interval - Example 1

- A researcher collected information on BMI (Body Mass Index) sample of 100 women. The sample mean is 22 and sample variance is 4. ($t_{0.025,99} = 1.98$)
- What is the point estimate of the population mean?
- What is the 95% confidence interval of the population mean?

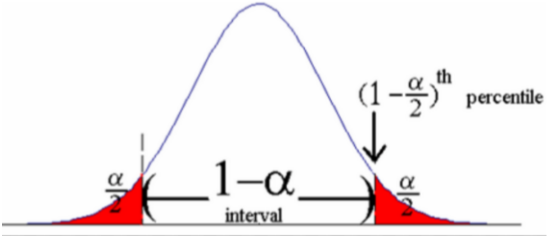
Constructing a Confidence Interval - Example 1

- A researcher collected information on BMI (Body Mass Index) sample of 100 women. The sample mean is 22 and sample variance is 4. ($t_{0.025,99} = 1.98$)
- What is the point estimate of the population mean?
- What is the 95% confidence interval of the population mean?

1. $se = \hat{\sigma}_{\bar{y}} = \frac{\hat{\sigma}}{\sqrt{n}} =$
2. $\hat{\mu} \pm t_{0.025,n-1} \cdot se =$

Two Comments on Confidence Interval

- Greater confidence level requires wider CI.
- Larger n produces narrower CI.



Constructing a Confidence Interval

- Given the CLT, when sample size n is large, the 95% confidence interval can also be written as:

$$\hat{\mu} \pm z_{0.025} \cdot se$$

- Since $z_{0.025} = 1.96$, so the confidence interval of sample mean in large sample can be written as:

$$\hat{\mu} \pm 1.96 \cdot se$$

Constructing a Confidence Interval - Example 2

- A researcher collected information on BMI (Body Mass Index) sample of 10,000 women. The sample mean is 22 and sample variance is 4.
- What is the 95% confidence interval of the population mean, given that we only know the critical value for the normal distribution $z_{0.025} = 1.96$?

Constructing a Confidence Interval - Example 2

- A researcher collected information on BMI (Body Mass Index) sample of 10,000 women. The sample mean is 22 and sample variance is 4.
- What is the 95% confidence interval of the population mean, given that we only know the critical value for the normal distribution $z_{0.025} = 1.96$?

1. $\hat{\mu} \pm z_{0.025} \cdot se =$

Extension:
What if we don't know the limiting distribution of the sample statistic?

1. The delta method.
2. Bootstrapping.

The Delta Method

- Single variable case:
- Suppose that a random variable Y_n has a limiting Normal distribution, that is: $\frac{\sqrt{n}(Y_n-\mu)}{\sigma} \xrightarrow{d} N(0,1)$,

  or equivalently: $Y_n \approx N(\mu, \frac{\sigma^2}{n})$.
- Also suppose that $g(.)$ is a differentiable function such that $g'(\mu) \neq 0$.
- Then:

  $g(Y_n) \approx N(g(\mu), (g'(\mu))^2 \frac{\sigma^2}{n}))$.
- There is also a multivariate delta method, in which we can construct a new variable as a function of several random vectors.

Bootstrapping

- **The CLT** enables us to derive direct standard error formulas when the key statistic takes various forms of averaging (e.g. sample mean or OLS estimator), but for hardly anything else.
- **The delta method** introduced above extends the formulas to smooth functions of averages, but it can still rely on tedious Taylor series expansions. In addition, the transformation function $g(.)$ needs to be differentiable (smooth), which is not always the case.
- The jackknife (1957) and the bootstrap (1979) move us towards a computation-based, non-formulaic approach to standard errors and sampling distributions.
- The idea is quite straightforward: The true probability distribution F is unknown, but we can create an empirical distribution as our estimate for F, termed $\hat{F}$. The empirical distribution is obtained by resampling the sample data.

Bootstrapping

- The bootstrap estimate of standard error for a statistic $\hat{\theta} = s(x)$ computed from a random sample $x = (x_1, x_2, ...x_n)$ begins with the notion of a bootstrap sample: $x^* = (x_1^*, x_2^*, ..., x_n^*)$, where each x_i^* is drawn randomly with equal probability and replacement from the original sample.
- Each bootstrap sample provides a bootstrap replication of the statistic of interest: $\hat{\theta}^* = s(x^*)$.
- We created a total of B bootstrap samples (e.g. B=200 is usually sufficient if your statistics is something like an "average"), and we use $\hat{\theta}^{*b} = s(x^{*b})$ to denote the statistic in the corresponding bootstrap replication b $(b = 1, 2, ...B)$.

Bootstrapping

- Suppose the true data generating process is:

  $F \xrightarrow{i.i.d.} x \xrightarrow{s} \hat{\theta}$.

- The **true probability distribution** F is, however, unobserved.
- We can describe the bootstrap process (or the process that generates bootstrap replications) as starting from the **empirical probability distribution**:

  $\hat{F} \xrightarrow{i.i.d.} x^* \xrightarrow{s} \hat{\theta}^*$.

Bootstrapping

- The resulting bootstrap estimate of standard error for the statistic of interest $\hat{\theta}$ is just the empirical standard deviation of the $\hat{\theta}^{*b}$ values across these B bootstrap replications:

  $\hat{se}_{boot} = [\Sigma_{b=1}^{B} \frac{(\hat{\theta}^{*b} - \hat{\theta}^{**})^2}{B-1}]^{1/2}$, where $\hat{\theta}^{**} = \Sigma_1^B \theta^{*b}/B$.
- The above describes the one-sample nonparametric bootstrap. There are also methods for conducting parametric and multisample versions of bootstrap.
- The bootstrap has been described as "computer-intensive statistics."

Statistical Inference II: Hypothesis Testing

Hypothesis and Significance Tests

- A hypothesis is a statement about a population.
- A significance test uses data in the sample to summarize the evidence about a hypothesis.
- Like the point estimate and confidence interval, hypothesis testing also relies on the sampling distribution of a statistic.

Hypothesis and Significance Tests

- Each significance test tests two hypotheses:
- Null hypothesis (H_0) states that the parameter takes a particular value.
- Alternative hypothesis (H_a) states that the parameter differs from that value, or falls in some alternative range of values.

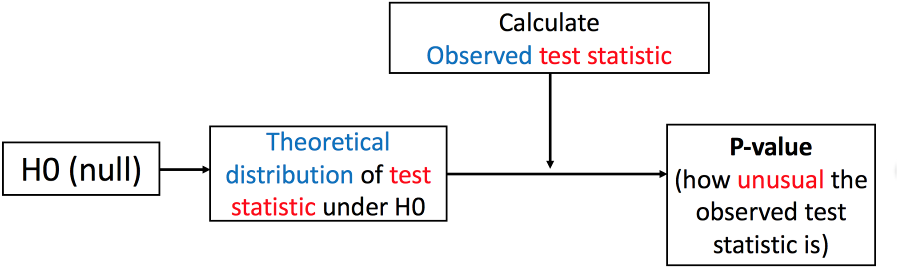
- Example:

  H_0: The mean weekly work hours among PhD students equals 35.

  H_a: The mean weekly work hours among PhD students does not equal 35.

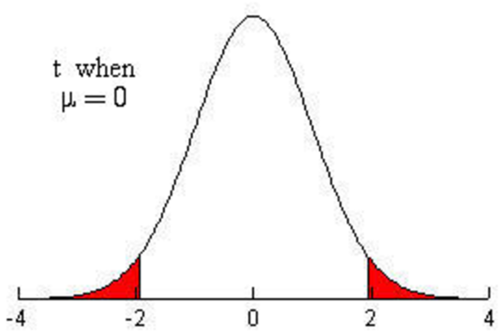
  Or, H_a can be: The mean weekly work hours among PhD students is greater than 35.

The Logic of Significance Test



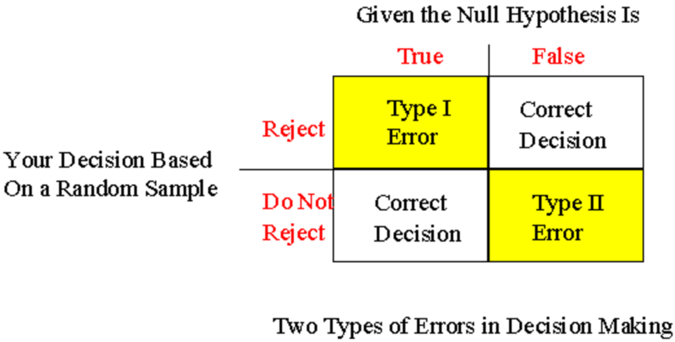
If p-value is sufficiently small, we reject H0

The Logic of Significance Test

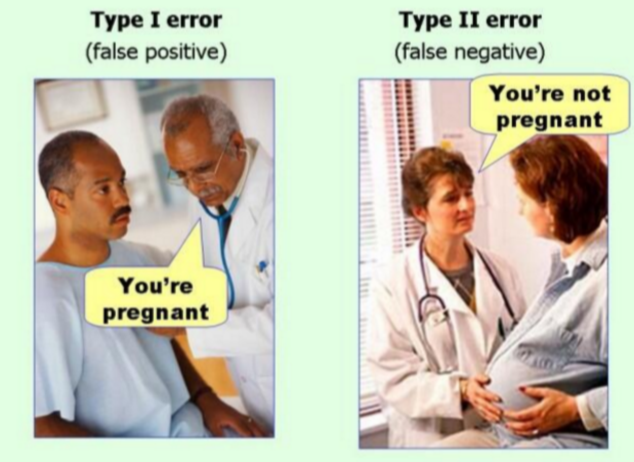


Note that in large samples, the test statistic distribution can be approximated by the normal distribution.

Type I and Type II Error

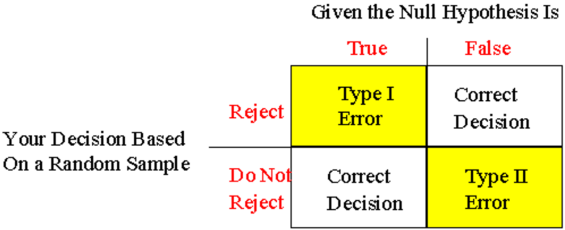


Type I and Type II Error



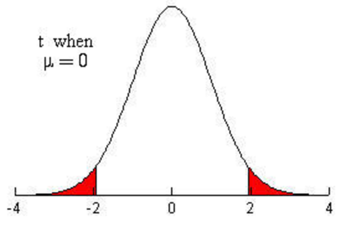
Significance Level

- In significance tests, the **significance level** α is the maximum level of **Type I error rate** (i.e. false discovery rate) allowed.
- *"If there was actually no effect, how unlikely would it be to get estimates such as we obtained?"*



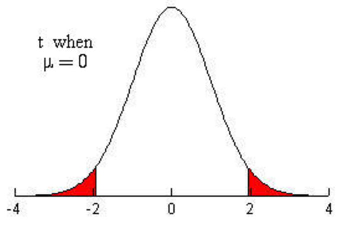
Two Types of Errors in Decision Making

P-value and Rejection Region



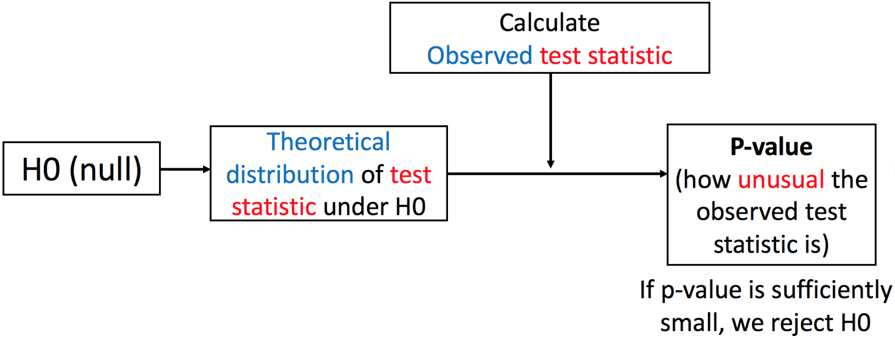
- The P-value is the **probability** that the test statistic equals the observed value or a value even more extreme in the direction predicted by H_a, assuming that H_0 is true.
- Intuition: the smaller the P-value, the more "extreme" the observed value is, and thus the stronger the evidence against H_0.
- If the P-value is smaller than the significance level (α), then the test statistic is in the rejection region (i.e. evidence against H_0).

P-value and Rejection Region



- **Warning 1:** p-value is NOT the probability that the null hypothesis is true! (i.e. it is NOT $Pr(H_0|Data)$).
- **Warning 2:** A small p-value can be seen as strong evidence against H_0, but a large p-value is NOT strong evidence in favor of H_0!

The Logic of Significance Test




Captured by Prathibha

Significance Test for the Mean - 5 Steps

1. Assumptions (randomization, quantitative variable, population distribution)
2. State the null and alternative hypotheses
3. Test statistic
4. P-value (smaller P-value is stronger evidence)
5. Conclusion

Significance Test for the Mean - 1. Assumption

- random sample
- quantitative variable
- population distribution is normal distribution (but in large sample this is not required)

Significance Test for the Mean - 2. Null and Alternative Hypotheses

Two-sided test:

- $H_0 : \mu = \mu_0$
- $H_a : \mu \neq \mu_0$

One-sided test:

- $H_0 : \mu = \mu_0$
- $H_a : \mu > \mu_0$
- Or: $H_a : \mu < \mu_0$

Significance Test for the Mean - Test Statistic and P-Value

- Test statistics

$$t = \frac{\bar{y} - \mu_0}{se} \quad where \quad se = s/\sqrt{n}$$

- The test statistics tell us how far the data fall from μ_0.
- Next, we need the P-value to tell us how unlikely this can happen if H_0 is true .
- For two-sided tests, P-value is 2 times the tail probability.
- For one-sided tests, P-value is the tail probability.
- We are focusing on two-sided tests in this course.

Significance Test for the Mean - Conclusion

- We compare the P-value to the pre-specified significance level α.
- If $P \leq \alpha$, we reject the null hypothesis; If $P > \alpha$, we do not reject the null hypothesis.

Equivalence between result of significant test and confident interval

- When $P \leq \alpha$ in two-sided test, the $(1-\alpha)$ CI for μ does not contain μ_0.
- When $P > \alpha$ in two-sided test, the $(1-\alpha)$ CI for μ contains μ_0

Effect of sample size on tests

- With large n, assumption of normal population distribution is not very important because of Central Limit Theorem.
- For a given observed sample mean and standard deviation, the larger the sample size n, the larger the test statistic, and the smaller the P-value. (i.e. We have more evidence with more data!)
- We'are more likely to reject a false H_0 when we have a larger sample size (the test has more power).
- Statistical significance does not necessarily imply practical significance!

Extension: Some More Thoughts on Samples

- In the above discussions, we are assuming that the sample is a **probabilistic sample** and the sample units drawn via **random selection** from a population.
- But that's rarely the case in real social science research!
- **Examples**:
  - response bias/non-response bias;
  - attrition bias in panel data;
  - convenience sample;
  - unobserved outcomes (e.g. potential wage levels among current inmates).

Extension: Statistical Power and Type II Error

Statistical Decision	Reality: H_0	Reality: H_A
H_0	$1 - \alpha$	β
H_A	α	$1 - \beta$

Also: α = probability of a Type I error
β = probability of a Type II error

- The power of any test of statistical significance is defined as the probability that it will **reject a false null hypothesis**.
- Statistical power is inversely related to β (the probability of making a Type II error). That is, power $= 1 - \beta$.

Extension: A Power Calculation Example