

Introduction to Econometrics

Review of Statistics

Review of Statistics

- Statistics:
 - Use a random sample to answer questions about an unknown characteristic of a population distribution
- Estimation:
 - Compute a “best guess” numerical value for an unknown characteristic of a population distribution
- Hypothesis testing:
 - Evaluate a specific hypothesis about an unknown characteristic of a population distribution
- Confidence intervals:
 - Estimate a range of “likely” values for an unknown characteristic of a population distribution

Outline

Estimation of the Population Mean

Hypothesis Tests Concerning the Population Mean

Confidence Intervals for the Population Mean

Comparing Means from Different Populations

Scatterplots, the Sample Covariance and the Sample Correlation

Outline

Estimation of the Population Mean

Hypothesis Tests Concerning the Population Mean

Confidence Intervals for the Population Mean

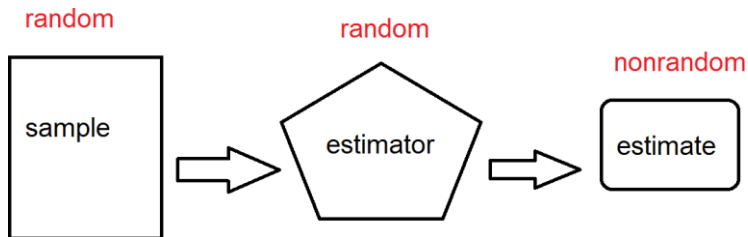
Comparing Means from Different Populations

Scatterplots, the Sample Covariance and the Sample Correlation

Estimator Versus Estimate

- Estimator: A function of a sample drawn randomly from the population
- Estimate: The numerical value of an estimator when it is computed using a specific sample
- Estimator is a random variable, estimate is a nonrandom number

Estimators Versus Estimate (cont'd)



Properties of Estimators

➤ Bias and unbiasedness:

➤ The bias of $\hat{\mu}_Y$ is $E[\hat{\mu}_Y] = \mu_Y$

➤ $\hat{\mu}_Y$ is unbiased if $E[\hat{\mu}_Y] = \mu_Y$

➤ Consistency:

➤ $\hat{\mu}_Y$ is consistent if $\hat{\mu}_Y \xrightarrow{p} \mu_Y$ (convergence in probability)

➤ $\hat{\mu}_Y \xrightarrow{p} \mu_Y$ means that $\lim_{n \rightarrow \infty} P_n[\mu_Y - \epsilon < \hat{\mu}_Y < \mu_Y + \epsilon] \rightarrow 1$ for all $\epsilon > 0$

➤ Variance and efficiency

➤ $\hat{\mu}_Y$ and $\bar{\mu}_Y$ are two unbiased estimators

➤ $\hat{\mu}_Y$ is more efficient than $\bar{\mu}_Y$ if $Var[\hat{\mu}_Y] < Var[\bar{\mu}_Y]$

Properties of \bar{Y}

- Assumptions:
 - Y_1, \dots, Y_n are i.i.d. draws (simple random sampling)
 - $E[Y_i] = \mu_Y$ and $Var[Y_i] = \sigma_Y^2 < \infty$
- \bar{Y} is an unbiased and consistent estimator of μ_Y
 - Sampling Distribution: $E[\bar{Y}] = \mu_Y$
 - Law of large numbers: $\bar{Y} \xrightarrow{p} \mu_Y$
- \bar{Y} is the Best Linear Unbiased Estimator (BLUE) of μ_Y
 - Let $\hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^n a_i Y_i$ where a_1, \dots, a_n are non random constants
 - If $E[\hat{\mu}_Y] = \mu_Y$, then $Var[\bar{Y}] < Var[\hat{\mu}_Y]$ unless $\hat{\mu}_Y = \bar{Y}$
- \bar{Y} is the least squares estimator of μ_Y
 - $m = \bar{Y}$ minimizes $\sum_{i=1}^n (Y_i - m)^2$

Outline

Estimation of the Population Mean

Hypothesis Tests Concerning the Population Mean

Confidence Intervals for the Population Mean

Comparing Means from Different Populations

Scatterplots, the Sample Covariance and the Sample Correlation

Null and Alternative Hypothesis

- Null hypothesis (H_0): The hypothesis to be tested

$$H_0 : \mu_Y = \mu_Y^{H_0}$$

- Alternative hypothesis (H_1): The hypothesis that holds if the null does not

- Two-sided: $H_1 : \mu_Y \neq \mu_Y^{H_0}$

- One-sided: $H_1 : \mu_Y > \mu_Y^{H_0}$ (or $H_1 : \mu_Y < \mu_Y^{H_0}$)

- Need to use the sample to decide whether to “accept” or reject H_0
 - We never really accept H_0 , we just fail to reject it

The p-Value

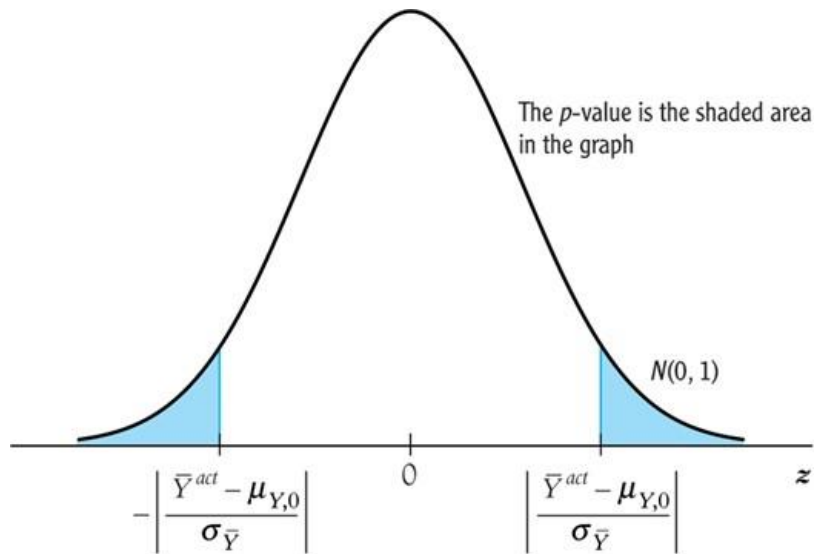
- \bar{Y} will rarely be exactly equal to $\mu_Y^{H_0}$ (H_0 false and/or sampling uncertainty)
 - ⇒ Need to decide how far is far enough for rejecting H_0
- P-value (significance probability): Probability of drawing a statistic at least as adverse to H_0 as the one actually observed in the sample, assuming H_0 is correct
- The higher the p-value the more likely it is that the deviation from H_0 you observed was due to random chance
- The lower the p-value the more likely it is that the deviation from H_0 you observed was due to H_0 being false
- To compute the p-value we need to know the sampling distribution of the statistic (e.g. \bar{Y}) under H_0

Calculating the p-Value When σ_Y Is Known

- Central limit theorem: Under $H_0, \bar{Y} \sim N\left(\mu_Y^{H_0}, \frac{\sigma_Y^2}{n}\right)$ for large n
- This means that we can compute the p-value as

$$\begin{aligned} p - \text{value} &= P_{H_0} \left[\left| \frac{\bar{Y} - \mu_0^{H_0}}{\frac{\sigma_Y}{\sqrt{n}}} \right| \geq \left| \frac{\bar{y}^{\text{actual}} - \mu_0^{H_0}}{\frac{\sigma_Y}{\sqrt{n}}} \right| \right] \\ &= 2 \times \Phi \left(- \left| \frac{\bar{y}^{\text{actual}} - \mu_0^{H_0}}{\frac{\sigma_Y}{\sqrt{n}}} \right| \right) \end{aligned}$$

Calculating the p-Value When σ_Y is Known (cont'd)



Sample Variance, Sample Standard Deviation, and Standard Error

- Sample variance

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n-1} \sum_{i=1}^n Y_i^2 + \frac{n}{n-1} \bar{Y}^2$$

- Sample standard deviation:

$$s_Y = \sqrt{s_Y^2}$$

- s^2 is an unbiased and consistent estimator of σ_Y^2 :

$$E[s_Y^2] = \sigma_Y^2$$

$$s_Y^2 \xrightarrow{p} \sigma_Y^2$$

- Standard error of \bar{Y} : Estimator of the standard deviation of \bar{Y}

$$SE[\bar{Y}] = \frac{s_Y}{\sqrt{n}}$$

Calculating the p-Value When σ_Y Is Unknown

- In practice we compute the p-value by replacing the unknown $\frac{\sigma_Y}{\sqrt{n}}$ with its estimator $SE[\bar{Y}]$

$$\begin{aligned} p - value &= P_{H_0} \left[\left| \frac{\bar{Y} - \mu_0^{H_0}}{SE[\bar{Y}]} \right| \geq \left| \frac{\bar{y}^{actual} - \mu_0^{H_0}}{SE[\bar{Y}]} \right| \right] \\ &= 2 \times \Phi \left(- \left| \frac{\bar{y}^{actual} - \mu_0^{H_0}}{SE[\bar{Y}]} \right| \right) \end{aligned}$$

- This works because $s_Y^2 \xrightarrow{p} \sigma_Y^2$ (we are close enough to the true value in large samples)

The t-Statistic

- t-statistic: Standardized sample average (or other estimator of interest)

$$t = \frac{\bar{Y} - \mu_0^{H_0}}{SE[\bar{Y}]}$$

- This is an example of a test statistic, a statistic used to perform a hypothesis test
- We can compute the p-value also using the t-statistic

$$p - value = 2 \times \Phi(-|t^{actual}|)$$

Where

$$t^{actual} = \frac{\bar{y}^{actual} - \mu_0^{H_0}}{SE[\bar{Y}]}$$

Hypothesis Testing with a Prespecified Significance Level

- Hypothesis tests can be performed without computing p-value by specifying in advance the probability you tolerate of incorrectly rejecting H_0
- For instance, reject H_0 whenever p-value is less than 5%
- Because the area under the tails of $N(0, 1)$ outside of ± 1.96 is 5%, this gives a simple rule: Reject H_0 if $|t^{actual}| > 1.96$
- Most commonly we use 5%, but you may want to use a more conservative approach in some cases
- When choosing the significance level, you face a tradeoff between size and power of the test

The Terminology of Hypothesis Testing

- Type I error: H_0 is rejected when it is true
- Type II error: H_0 is not rejected when it is false
- Significance level: Prespecified probability of type I error
- Critical value: Value of the test statistic for which H_0 is just rejected
- Rejection region: Values of the test statistic for which H_0 is rejected
- Acceptance region: Values of the test statistic for which H_0 is not rejected
- Size of the test: Probability of incorrectly rejecting H_0
- Power of the test: Probability of correctly rejecting H_0

One-Sided Alternatives

- One-sided alternative hypothesis: $H_1 : \mu_Y > \mu_Y^{H_0}$
 - Only large positive values of the t-statistic reject H_0
$$p - value = 1 - \Phi(t^{actual})$$
 - Critical value at the 5% significance level 1.64
- One-sided alternative hypothesis: $H_1 : \mu_Y < \mu_Y^{H_0}$
 - Only large negative values of the t-statistic reject H_0
$$p - value = \Phi(t^{actual})$$
 - Critical value at the 5% significance level -1.64

Example

- Y = hourly earnings of recent college graduates in the US
- $H_0 = \mu_Y = 20, H_1: \mu_Y \neq 20$
- $n = 200, \bar{Y} = 22.64, s_Y = 18.14$
- $SE(\bar{Y}) = \frac{18.14}{\sqrt{200}} = 1.28$
- $t = \frac{22.64 - 20}{1.28} = 2.06$
- $p = 2 \times \phi(-2.06) = 0.039$

Outline

Estimation of the Population Mean

Hypothesis Tests Concerning the Population Mean

Confidence Intervals for the Population Mean

Comparing Means from Different Populations

Scatterplots, the Sample Covariance and the Sample Correlation

Confidence Intervals for the Population Mean

- Confidence interval: set of values that contain μ_Y with a prespecified probability (called confidence level)
- Coverage probability: Probability that the confidence set contains μ_Y
- $(1 - \alpha)\%$ confidence interval for μ_Y is the collation of all $\mu_0^{H_0}$ that are not rejected at the $\alpha\%$ significance level
- 95% confidence interval for μ_Y :

$$[\bar{Y} - 1.96 \times SE[\bar{Y}], \bar{Y} + 1.96 \times SE[\bar{Y}]]$$

$$P[\bar{Y} - 1.96 \times SE[\bar{Y}] \leq \mu_Y \leq \bar{Y} + 1.96 \times SE[\bar{Y}]] = 0.95$$

Example

- Y = hourly earnings of recent college graduates in the US
- $n = 200, \bar{Y} = 22.64, s_Y = 18.14$
- $SE(\bar{Y}) = \frac{18.14}{\sqrt{200}} = 1.28$
- 95% $CI : 22.64 \pm 1.96 \times 1.28 = [20.13, 25.15]$

Outline

Estimation of the Population Mean

Hypothesis Tests Concerning the Population Mean

Confidence Intervals for the Population Mean

Comparing Means from Different Populations

Scatterplots, the Sample Covariance and the Sample Correlation

Hypothesis Tests and Confidence Intervals for the Difference Between Two Means

- $H_0 : \mu_m - \mu_w = d_0$ vs $H_1 : \mu_m - \mu_w \neq d_0$
- Sample: $n_m, n_w, \bar{Y}_m, \bar{Y}_w, s_m, s_w$
- Estimator: $\bar{Y}_m - \bar{Y}_w$
- Standard error:

$$SE[\bar{Y}_m - \bar{Y}_w] = \sqrt{\widehat{Var}[\bar{Y}_m] + \widehat{Var}[\bar{Y}_w]} = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

- T-statistic: $t = \frac{\bar{Y}_m - \bar{Y}_w - d_0}{SE[\bar{Y}_m - \bar{Y}_w]} \sim N(0,1)$ for large n_m and n_w under H_0
- 95% confidence interval: $\bar{Y}_m - \bar{Y}_w \pm 1.96 \times SE[\bar{Y}_m - \bar{Y}_w]$

Example

- Y = hourly earnings of recent college graduates in the US
- $H_0: \mu_m - \mu_{mw} = 0, H_1: \mu_m - \mu_{mw} \neq 0$
- $n_m = 1838, \bar{Y}_m = 24.98, s_m = 11.78$
- $n_w = 1871, \bar{Y}_w = 20.87, s_w = 9.66$
- $\bar{Y}_m - \bar{Y}_w = 4.11$
- $SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{11.78^2}{1838} + \frac{9.66^2}{1871}} = 0.35$
- $t = \frac{4.11 - 0}{0.35} = 11.47$
- $p = 2 \times \Phi(-11.47) \approx 0$
- 95% CI : $4.11 \pm 1.96 \times 0.35 = [3.41, 4.80]$

Outline

Estimation of the Population Mean

Hypothesis Tests Concerning the Population Mean

Confidence Intervals for the Population Mean

Comparing Means from Different Populations

Scatterplots, the Sample Covariance and the Sample Correlation

Scatterplot, Sample Covariance, and Sample Correlation

- Scatterplot: Plot of n observations on X_i and Y_i in which each observation is represented by the point (X_i, Y_i)
- Sample covariance:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} \sum_{i=1}^n X_i Y_i - \frac{n}{n-1} \bar{X} \bar{Y}$$

- Sample correlation:

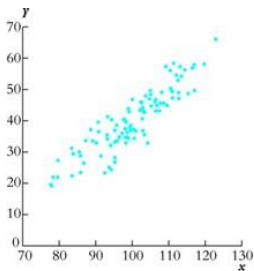
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

- Sample covariance and correlation are consistent estimators:

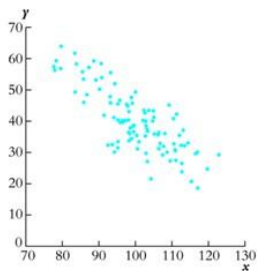
$$s_{XY} \xrightarrow{p} \sigma_{XY}$$

$$r_{XY} \xrightarrow{p} \rho_{XY}$$

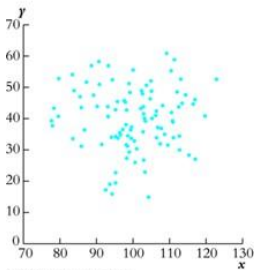
Scatterplots for Four Hypothetical Data Sets



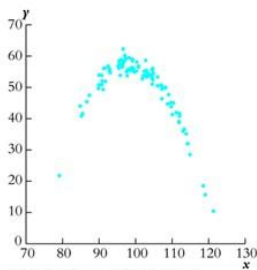
(a) Correlation = +0.9



(b) Correlation = -0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)