Review of Statistics

Dr. Patrick Toche

Textbook:

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.
Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.

The textbook comes with online resources and study guides. Other references will be given from time to time.

In this lesson you will learn ...

- estimators of the population mean, the sample mean, the sample variance, the sample standard deviation, the standard error, the sample covariance, the sample correlation
- the null hypothesis and alternatives
- the p-value, the t-statistic
- how to construct confidence intervals for the population mean
- how to compare means from different populations
- how to estimate the causal effect using differences of means

Estimating the Population Mean

- **Estimator:** A function of a sample of data drawn randomly from a population. An estimate is a numerical value using data from a specific sample. An estimator is a random variable, an estimate is not.
- The sample mean $\overline{Y}$ is a natural estimate of μ_Y.
- It is not the only possible estimator. There are many useful estimators of μ_Y.
- Can you think of other natural estimators?
- **Notation:** An estimator of μ_Y is typically denoted $\hat{\mu}_Y$. We might denote two estimators $\hat{\mu}_Y$ and $\tilde{\mu}_Y$. The sample mean $\overline{Y}$ is one such estimator. If we deal with a single random variable Y, we sometimes write just μ and $\hat{\mu}$. When our focus is on the sample size n, we often denote it for emphasis, e.g. $\overline{Y_n}$. We may denote a particular estimate with a lower-case letter, e.g. $\overline{y}$. We may denote a particular value of a parameter such as the population mean with a 0 index, e.g. μ_0.

Properties of Estimators

- What are desirable characteristics of the sampling distribution of an estimator?
  - Close to the unknown value ...in an "average" sense!
  - Stabilizes arbitrarily closely to the true value as the sample size is increased.
  - Uses sample data to avoid erratic fluctuations caused by small samples — estimates do not go "all over the place".
- Some of these properties may be traded off, in particular centrality and dispersion, e.g. bias or consistency vs efficiency.
- Some estimators may perform well for very large samples, but not so well for small samples.
- To fairly compare estimators, you must be clear about your objectives. For instance, an estimator that is equal to a constant, say 0, for any sample, could have a very large bias, but it is absolutely efficient!

Properties of Estimators

- **Desirable statistical properties of estimators:**
  1. **Unbiased:** The estimator $\hat{\mu}_Y$ is unbiased if:

     $$\mathrm{E}[\hat{\mu}_Y] = \mu_Y$$

     (The estimator is close on average)
  2. **Consistent:** The estimator is consistent if the event that the estimator $\hat{\mu}_Y$ is arbitrarily close to the true value μ_Y becomes certain as the sample size n increases:

     $$\Pr(|\hat{\mu}_Y - \mu_Y| < \epsilon) \xrightarrow{n \to \infty} 1$$

     A consistent estimator "converges in probability" to the true mean: $\hat{\mu}_Y \xrightarrow{p} \mu_Y$ (the uncertainty caused by random fluctuations in the sample becomes negligible).
  3. **Efficient:** $\hat{\mu}_Y$ is said to be more efficient than $\tilde{\mu}_Y$ if it has a smaller variance,

     $$\mathrm{var}(\hat{\mu}_Y) < \mathrm{var}(\tilde{\mu}_Y)$$

     (the more efficient estimator has the tighter sampling distribution).

BLUE

Best Linear Unbiased Estimator (BLUE)

Let $\hat{\mu}_Y$ be an estimator of μ_Y that is a weighted average of $Y_1, \ldots, Y_n$; that is, $\hat{\mu}_Y = (1/n)\sum_{i=1}^{n} a_i Y_i$, where $a_1, \ldots, a_n$ are nonrandom constants. If $\hat{\mu}_Y$ is unbiased, then $\mathrm{var}(\overline{Y}) \leq \mathrm{var}(\hat{\mu}_Y)$. Thus $\overline{Y}$ is the Best Linear Unbiased Estimator — the most efficient estimator of μ_Y among all unbiased estimators that are weigthed means of $Y_1, \ldots, Y_n$.

- **Least Squares Estimator:** The estimator m that minimizes the sum of the squared differences $Y_i - m$, a measure of the total squared differences between the estimator and the sample points.

  $$\sum_{i=1}^{n}(Y_i - m)^2$$

- Non-random sampling can lead to a bias in $\overline{Y}$.
- Think of examples of sampling procedures that produce bias.

Alf Landon Wins!

The Literary Digest had correctly predicted the outcomes of the 1916, 1920, 1924, 1928, and 1932 elections by conducting polls. These polls were popular. In one of the largest surveys ever, the Literary Digest surveyed more than $2,000,000$ readers — a response rate of about 25% — and predicted Republican presidential candidate Alfred Landon would win 57 percent of the popular vote and 370 electoral votes. Instead, he lost the popular vote by more than 10 million votes and carried only two states, for a total of eight electoral votes to Roosevelt's 523. The problem was that the mailing had targeted people more like to vote Republican. By contrast, the American Institute of Public Opinion, founded by George Gallup, set quotas for the numbers of individuals needed for each type of respondent and by polling "only" $50,000$ people made an accurate prediction. Gallup became a household name.

Hypothesis Testing

- **Null Hypothesis:**

  $$H_0\colon \mathrm{E}[Y] = \mu_{Y,0}$$

  where $\mu_{Y,0}$ is a specific value.
- Example of null: On average in the population, college graduates earn $20 per hour.
- The alternative hypothesis specifies what is true if the null hypothesis is not.
- **Two-Sided Alternative:**

  $$H_1\colon \mathrm{E}[Y] \neq \mu_{Y,0}$$

- **One-Sided Alternative:**

  $$H_A\colon \mathrm{E}[Y] > \mu_{Y,0}$$

  (or in the other direction $\mathrm{E}[Y] < \mu_{Y,0}$)
- Statistical hypothesis testing: *either reject the null hypothesis or fail to reject it.*

The p-Value

- Differences between $\overline{Y}$ and $\mu_{Y,0}$ could arise ...
  - because the null hypothesis is false; or
  - because of random sampling.
- **It is impossible to distinguish between these two possibilities with certainty.**
- But it may be possible to distinguish between them with high confidence.
- ***p*-Value:** is the probability of drawing a statistic at least as adverse to the null hypothesis as the one computed from the sample, assuming the null hypothesis is correct. In other words, the probability of drawing $\overline{Y}$ at least *that* far in the tails of the H_0-distribution.
- The p-value is the area in the tails of the distribution of $\overline{Y}$ when the null is true.
- The p-value *aka* "significance probability."
- To compute the p-value, we must know the sampling distribution of $\overline{Y}$ under the null.
- We distinguish two cases:
  - σ_Y^2 known
  - σ_Y^2 unknown

p-Value for Known σ_Y^2

- Under the null hypothesis, the sampling distribution of $\overline{Y}$ if σ_Y^2 is (for some reason) known with certainty is

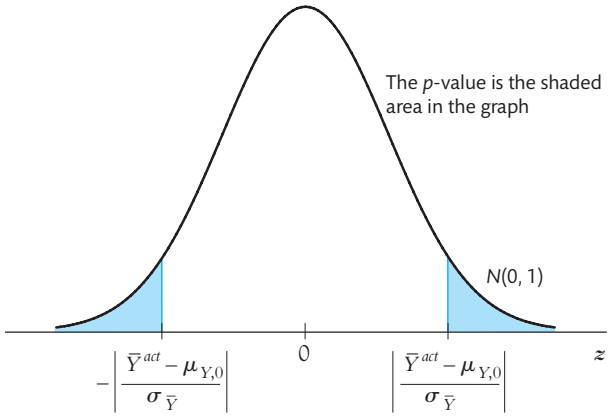
$$\overline{Y} \sim N(\mu_{Y,0}, \sigma_Y^2/n) \text{ for large } n$$

- Equivalently,

$$\frac{\overline{Y} - \mu_{Y,0}}{\sigma_Y} \sim N(0,1) \text{ as } n \to \infty$$

- The standard normal distribution $N(0,1)$ is symmetric around the mean, bell-shaped, and has relatively thin tails. We can take any interval of the real line and compute the corresponding probability that the population mean lies within this interval, e.g. compute the probability that the population mean lies within $[-1, 1]$. The probability is given by the area under the curve. We can similarly compute probabilities for half-intervals, e.g. $[0, \infty)$.
- The normal distribution is, mathematically speaking, defined all the way from $-\infty$ to $+\infty$, but the probability of drawing a value "far" from the mean quickly gets very small.
- Calculate the probability of drawing a value outside of $[-5, 5]$ from $N(0,1)$.

p-Value



The p-value is the probability of drawing a value of $\overline{Y}$ that differs from $\mu_{Y,0}$ by at least as much as the actual sample mean $\overline{Y}^{act}$. In large samples, $\overline{Y}$ is distributed $N(\mu_{Y,0}, \sigma_Y^2/n)$ under the null hypothesis. The p-value is the shaded standard normal tail probability.

Sample Variance and Standard Deviation

- **Sample Variance:** An estimator of the population variance σ_Y^2, denoted s_Y^2 or $\hat{\sigma}_Y^2$.
- **Sample Standard Deviation:** An estimator of the population stdev σ_Y, denoted s_Y or $\hat{\sigma}_Y$.
- The sample standard deviation is the square root of the sample variance.
- **Population Variance:**

$$\sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)^2$$

- **Sample Variance:**

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

- **Degrees of freedom adjustment:** Division by $n-1$ instead of n.
- **Consistency:** The sample variance is a consistent estimator of the population variance:

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

Standard Error

- **Standard Error:** An estimator of the standard deviation of the sampling distribution of $\overline{Y}$, denoted $\mathsf{SE}(\overline{Y})$.
- The standard deviation of the sampling distribution of $\overline{Y}$ is $\sigma_Y/\sqrt{n}$, so it is natural to use $s_Y/\sqrt{n}$ as an estimator of $\sigma_{\overline{Y}}$. Indeed, $s_Y/\sqrt{n}$ is a particular estimator $\hat{\sigma}_{\overline{Y}}$, and the most common estimator used in practice.
- If $Y_1, \ldots, Y_n$ are i.i.d., the standard error is an unbiased and consistent estimator of σ_Y.

$$\mathsf{SE}(\overline{Y}) = s_Y/\sqrt{n}$$

- If $Y_1, \ldots, Y_n$ are i.i.d. draws from a Bernoulli distribution with success probability p, the formula for the variance $\overline{Y}$ simplifies to $p(1-p)/n$; and the standard error is

$$\mathsf{SE}(\overline{Y}) = \sqrt{\overline{Y}(1-\overline{Y})/n}$$

p-Value for Unknown σ_Y^2

- Because s_Y^2 is a consistent estimator of σ_Y^2, the p-value can be computed by replacing $\sigma_{\overline{Y}}$ by the standard error $\hat{\sigma}_{\overline{Y}}$.
- If $Y_1, \ldots, Y_n$ are i.i.d., the p-value is calculated from

$$p\text{-value} = 2\Phi\left(-\left|\frac{\overline{Y}^{\mathsf{act}} - \mu_{Y,0}}{\mathsf{SE}(\overline{Y})}\right|\right)$$

- ***t*-statistic:** For large sample size n, the estimator s_Y^2 is close to σ_Y^2 with high probability. The distribution of the t-statistic is approximately the same as the distribution of $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$, which is well approximated by the standard normal distribution.

$$t \sim N(0,1) \text{ for large } n$$

- The p-value can be written in terms of the t-statistic:

$$p\text{-value} = 2\Phi\left(-|t^{\mathsf{act}}|\right)$$

t-Statistic

- Example: A sample of size $n = 200$ recent college graduates is used to test the null hypothesis that the mean wage $\mathrm{E}[Y]$ is \$20 per hour. The sample average is $\overline{Y}^{\mathsf{act}} = \$22.64$ and the sample standard deviation is $s_Y = \$18.14$.
- The standard error of $\overline{Y}$ is

$$\frac{s_Y}{\sqrt{n}} = \frac{18.14}{\sqrt{200}} = 1.28$$

- The t-statistic is:

$$t^{\mathsf{act}} = \frac{22.64 - 20}{1.28} = 2.06$$

- The corresponding p-value is:

$$p\text{-value} = 2\Phi(-2.06) = 0.039$$

- Under the null hypothesis, the probability of drawing a sample mean no less different from the null as the one actually computed is 3.9%.

One-Sided Alternative

- **One-Sided Alternative:**

$$H_A\colon \mathrm{E}[Y] > \mu_{Y,0}$$
$$\implies p\text{-value} = \Pr(Z > t^{\mathsf{act}} | H_0) = 1 - \Phi(t^{\mathsf{act}})$$

- Because the distribution is symmetric, one-sided alternatives can be found with the same critical value in absolute value — only the sign of the critical value differs between left-sided and right-sided alternatives.
- The critical value for the rejection region in a test against a one-sided alternative is $t_{\mathsf{crit}} = 1.64$ — smaller than the critical value in the case of a two-sided alternative $t_{\mathsf{crit}} = 1.96$. This is because the tail probability is located exclusively on one-side, rather than divided evenly.

Terminology

- **Type-I Errors:** Incorrectly reject the null hypothesis even though it is true.
- **Type-II Errors:** Incorrectly fail to reject the null hypothesis even though it is false.
- **Significance level:** If you set a threshold probability for your tolerance to Type-I error (rejecting a correct null), say a threshold of 5%, then you will reject the null hypothesis if, and only if, the p-value is less than the significance level 0.05. Against a two-sided alternative, the simple rule follows:

$$\text{reject } H_0 \text{ if } |t^{\text{act}}| > 1.96$$

  where 1.96 is the **critical value** for a **two-sided test**.
- **Rejection region:** The set of values of the test statistic for which the test rejects the null hypothesis.
- **Acceptance region:** The set of values of the test statistic for which it does not reject the null hypothesis.
- **Size of the test:** The probability that the test actually incorrectly rejects the null hypothesis when it is true.
- **Power of the test:** The probability that the test correctly rejects the null hypothesis when the alternative is true.

Confidence Intervals

- **Confidence set:** A set of values that contains the true population mean μ_Y with a certain pre-specified probability.
- **Confidence level:** The pre-specified probability that μ_Y is contained in the confidence set.
- **Confidence interval:** The confidence set contains all the possible values of the mean between a lower and an upper limit — it is an interval.
- **Coverage probability:** The coverage probability of a confidence interval for the population mean is the probability, computed over all possible random samples, that it contains the true population mean.
- A 95% two-sided confidence interval for μ_Y is an interval constructed so that it contains the true value of μ_Y in 95% of all possible random samples.

$$90\% \text{ confidence interval for } \mu_Y : \{\overline{Y} \pm 1.64\, \mathsf{SE}(\overline{Y})\}$$
$$95\% \text{ confidence interval for } \mu_Y : \{\overline{Y} \pm 1.96\, \mathsf{SE}(\overline{Y})\}$$
$$99\% \text{ confidence interval for } \mu_Y : \{\overline{Y} \pm 2.58\, \mathsf{SE}(\overline{Y})\}$$

Mean Differences

- **Test for Difference Between Two Means:** Consider the null that two groups w and m differ by a quantity d_0:

$$H_0 : \mu_m - \mu_w = d_0$$
$$H_1 : \mu_m - \mu_w \neq d_0$$

- Example: Test the hypothesis that the earnings of women and men are no different on average, that is $d_0 = 0$:

$$H_0 : \mu_{\text{men}} = \mu_{\text{women}}$$
$$H_1 : \mu_{\text{men}} \neq \mu_{\text{women}}$$

- To test the hypothesis, we collect two samples of size n_w and n_m and compute the actual sample means $\overline{Y}_w$ and $\overline{Y}_m$. A natural estimator for this test is $\overline{Y}_m - \overline{Y}_w$.

Mean Differences

- What is the sample distribution of the estimator $\overline{Y}_m - \overline{Y}_w$?
- We have

$$\overline{Y}_w \sim N(\mu_w, \sigma_w^2/n_w), \quad \overline{Y}_m \sim N(\mu_m, \sigma_m^2/n_m), \quad \overline{Y}_m \perp\!\!\!\perp \overline{Y}_w$$

- The distribution follows:

$$\overline{Y}_m - \overline{Y}_w \sim N\left(\mu_m - \mu_w, \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}\right)$$

- The standard error:

$$\mathsf{SE}(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}}$$

- The t-statistic for comparing two means:

$$t = \frac{(\overline{Y}_m - \overline{Y}_w) - d_0}{\mathsf{SE}(\overline{Y}_m - \overline{Y}_w)}$$

Summary

1. The sample mean $\overline{Y}$ is an estimator of the population mean μ_Y. If the sample observations $Y_1, \ldots, Y_n$ are i.i.d.,
   a. The sampling distribution of $\overline{Y}$ has mean μ_Y and variance $\sigma^2_{\overline{Y}}/n$.
   b. The sample mean $\overline{Y}$ is unbiased.
   c. *Law of Large Numbers (LLN):* As the sample size n is increased, the sample $\overline{Y}$ tends to the population mean μ_Y — The sample mean is a consistent estimator of the population mean.
   d. *Central Limit Theorem (CLT):* For large sample sizes n, the sample mean $\overline{Y}$ has an approximately normal sampling distribution.
2. The t-statistic is used to test the null hypothesis that the population mean takes on a particular value. If n is large, the t-statistic has a standard normal sampling distribution when the null hypothesis is true.

Summary

3. The t-statistic can be used to calculate the p-value associated with the null hypothesis. The p-value is the probability of drawing a statistic at least as adverse to the null hypothesis as the one you actually computed in your sample, assuming the null hypothesis is correct. A small p-value is evidence that the null hypothesis is false.
4. A 95% confidence interval for $\overline{Y}$ is an interval constructed so that it contains the true value of $\overline{Y}$ in 95% of all possible samples.
5. Hypothesis tests and confidence intervals for the difference in the means of two populations are conceptually similar to tests and intervals for the mean of a single population.
6. The sample correlation coefficient is an estimator of the population correlation coefficient and measures the linear relationship between two variables – that is, how well their scatterplot is approximated by a straight line.

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 3, Exercise 1.

In a population, $\mu_Y = 100$ and $\sigma^2_Y = 43$. Use the central limit theorem to answer the following questions:

1. In a random sample of size $n = 100$, find $\Pr(\overline{Y} < 101)$.
2. In a random sample of size $n = 64$, find $\Pr(101 < \overline{Y} < 103)$.
3. In a random sample of size $n = 165$, find $\Pr(\overline{Y} > 98)$.

Stock & Watson, Introduction (4th), Chapter 3, Exercise 2.

Let Y be a Bernoulli random variable with success probability $\Pr(Y = 1) = p$, and let $Y_1, \ldots, Y_n$ be i.i.d. draws from this distribution. Let $\hat{p}$ be the fraction of successes (1s) in this sample.

1. Show that $\hat{p} = \overline{Y}$.
2. Show that $\hat{p}$ is an unbiased estimator of p.
3. Show that $\text{var}(\hat{p}) = p(1-p)/n$.

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 3, Exercise 3.

In a survey of 400 likely voters, 215 responded that they would vote for the incumbent, and 185 responded that they would vote for the challenger. Let p denote the fraction of all likely voters who preferred the incumbent at the time of the survey, and let $\hat{p}$ be the fraction of survey respondents who preferred the incumbent.

1. Use the survey results to estimate p.
2. Use the estimator of the variance of $\hat{p}$, $\hat{p}(1-\hat{p})$, to calculate the standard error of your estimator.
3. What is the p-value for the test of $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$?
4. What is the p-value for the test of $H_0: p = 0.5$ vs. $H_1: p > 0.5$?
5. Why do the results from (c) and (d) differ?
6. Did the survey contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey? Explain.

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 3, Exercise 6.

Let $Y_1, \ldots, Y_n$ be i.i.d. draws from a distribution with mean μ. A test of H_0: $\mu = 5$ vs. H_1: $\mu \neq 5$ using the usual t-statistic yields a p-value of 0.03.

1. Does the 95% confidence interval contain $\mu = 5$? Explain.
2. Can you determine if $\mu = 6$ is contained in the 95% confidence interval? Explain.

Stock & Watson, Introduction (4th), Chapter 3, Exercise 7.

In a given population, 11% of the likely voters are African American. A survey using a simple random sample of 600 landline telephone numbers finds 8% African Americans. Is there evidence that the survey is biased? Explain.

Keywords

estimator estimate bias consistency efficiency BLUE (Best Linear Unbiased Estimator) least squares estimator hypothesis tests null hypothesis alternative hypothesis two-sided alternative hypothesis p-value (significance probability) sample variance sample standard deviation degrees of freedom standard error t-statistic t-ratio test statistic type I error type II error significance level critical value

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 3, Exercise 13.

Data on fifth-grade test scores (reading and mathematics) for 420 school districts in California yield average score $\overline{Y} = 654.2$ and standard deviation $s_Y = 19.1$.

1. Construct a 95% confidence interval for the mean test score in the population.
2. When the districts were divided into those with small classes (< 20 students per teacher) and those with large classes (≥ 20 students per teacher), the following results were found:

Class Size	Average Score ($\overline{Y}$)	Standard Deviation (s_Y)	n
Small	657.4	19.4	238
Large	650.0	17.9	182

Is there statistically significant evidence that the districts with smaller classes have higher average test scores? Explain.