Review of Statistics

Osman DOĞAN

Outline: Review Statistics

- Review of Statistics:
  1. Estimation (of the population mean)
  2. Testing hypotheses (about the population mean)
  3. Confidence intervals (for the population mean)
  4. Comparing means from different populations
- Readings:
  1. Stock and Watson (2020, Chapter 3).
  2. Hanck et al. (2021, Chapter 3).

Estimation

- Three types of statistical methods are used throughout econometrics: estimation, hypothesis testing, and confidence intervals.
- Last week we saw that the sample mean $\bar{Y}$ can approximate the unknown population mean μ_Y with high probability in large samples.
- Therefore, $\bar{Y}$ said to be a good estimator for the unknown population mean μ_Y.

Definition 1

An estimator is a formula to estimate (or approximate) an unknown characteristic of the population distribution of random variable.

- An estimator is a function of sample data. For example, $\bar{Y}$ is a function of the sample data $Y_1, \ldots, Y_n$.

Definition 2

An estimate is the numerical value of the estimator when it is actually computed using data from a specific sample.

Estimation

- An estimator is a random variable because of randomness in selecting the sample, while an estimate is a nonrandom number.
- There may be many possible estimators for an unknown characteristic of the population distribution of random variable.
- For example, use the first observation, Y_1 as your estimator for μ_Y instead of $\bar{Y}$.
- So what makes one estimator better than another?

Estimation

- We need to define some measures to understand what we exactly mean by better.

Definition 3

Let $\hat{\mu}_Y$ be an estimator of μ_Y.

1. $\hat{\mu}_Y$ is called an unbiased estimator if $\mathrm{E}(\hat{\mu}_Y) = \mu_Y$.
2. $\hat{\mu}_Y$ is called a consistent estimator if $\hat{\mu}_Y \xrightarrow{p} \mu_Y$ as the sample size tends to infinity.
3. Let $\tilde{\mu}_Y$ be another estimator of μ_Y and suppose both $\hat{\mu}_Y$ and $\tilde{\mu}_Y$ are unbiased. Then, $\hat{\mu}_Y$ is said to be more efficient/precise than $\tilde{\mu}_Y$ if $\mathrm{Var}(\hat{\mu}_Y) < \mathrm{Var}(\tilde{\mu}_Y)$.

- Ideally, we would like to work with an estimator that is consistent, unbiased and very precise.
- If an estimator is the most precise within a class, it is said to be the best.
- For example, $\bar{Y}$ is the best, linear, unbiased estimator (BLUE) of μ_Y under simple random sampling.

Estimation

- Is $\bar{Y}$ unbiased under simple random sampling?

$$\mathrm{E}(\bar{Y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu_Y = \mu_Y.$$

- Is $\bar{Y}$ consistent under simple random sampling? One way to show this by showing $\mathrm{Var}(\bar{Y})$ approaches zero as the sample size grows without a bound.

$$\begin{aligned}
\mathrm{Var}(\bar{Y}) &= \mathrm{E}\left(\bar{Y} - \mu_Y\right)^2 = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu_Y\right)^2 \\
&= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mu_Y\right)\right)^2 = \mathrm{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(Y_i - \mu_Y\right)\left(Y_j - \mu_Y\right)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\left(Y_i - \mu_Y\right)^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathrm{E}\left(Y_i - \mu_Y\right)\left(Y_j - \mu_Y\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\left(Y_i - \mu_Y\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 = \frac{\sigma_Y^2}{n} \to 0 \text{ as } n \to \infty.
\end{aligned}$$

Estimation

- What does $\mathrm{E}(\bar{Y}) = \mu_Y$ mean?
- It simply means that the estimator $\bar{Y}$ attains the unknown population mean of Y on average.
- Notice hear that this is the case for any sample size, i.e., the sample size need not tend to infinity.
- What does $\bar{Y} \xrightarrow{p} \mu_Y$ mean?
- It simply means that the likelihood of $\bar{Y}$ approaching the unknown population mean of Y approaches to 1 as the sample size tends to infinity.
- Why is showing that $\mathrm{Var}(\bar{Y})$ approaches zero (as the sample size grows) sufficient for showing that $\bar{Y}$ is consistent?
- Since the mean of $\bar{Y}$ is μ_Y for any sample size and $\mathrm{Var}(\bar{Y})$ approaches zero as the sample size grows, the sampling distribution of $\bar{Y}$ collapses on μ_Y as the sample size grows.

Estimation

Example 1

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

(a) Show that $\hat{p}=\bar{Y}$.

(b) Show that $\hat{p}$ is an unbiased estimator of p.

(c) Show that $\text{Var}(\hat{p})=p(1-p)/n$.

Solution of Example 1.

Note that $\mathrm{E}(Y_i) = 1 \times p + 0 \times (1-p) = p$ and $\mathrm{Var}(Y_i) = \mathrm{E}(Y_i^2) - (\mathrm{E}(Y_i))^2 = p(1-p)$.

(a) Since Y_i takes either 0 or 1, we have $\hat{p} = \frac{\text{number of 1's}}{n} = \frac{\sum_{i=1}^n Y_i}{n} = \bar{Y}$.

(b) $\mathrm{E}(\hat{p}) = \mathrm{E}(\bar{Y}) = \mathrm{E}\left(\frac{\sum_{i=1}^n Y_i}{n}\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}(Y_i) = p.$

(c) $\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{\sum_{i=1}^n Y_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(Y_i) = p(1-p)/n.$

□

Hypothesis Testing

- What is a hypothesis in econometric analysis?
- It is a claim about some characteristic of a random variable, e.g., its mean, variance and so on.
- Here are some examples:
  - Do the mean hourly earnings of recent U.S. college graduates equal $\$20$ per hour?
  - Are mean earnings the same for male and female college graduates?
- These questions are some specific claims about the population distribution of earnings.
- The statistical challenge (the hypothesis testing problem) is to answer these questions based on evidence from a sample.

Hypothesis Testing

- The first thing to do in a hypothesis testing problem is to state the hypothesis to be tested and what happens if it does not hold.
- The hypothesis to be tested is called the null hypothesis, denoted by H_0.
- The scenario when the null does not hold is stated by the alternative hypothesis, denoted by H_1.
- Let Y denote the hourly earnings of recent college graduates.
- Then, the conjecture that, on average in the population, college graduates earn 20 per hour means, $H_0 : \mu_Y = 20$.
- The researcher needs to decide one of the alternative scenario:
    1. Two-sided alternative hypothesis: $H_1 : \mu_Y \neq 20$,
    2. One-sided alternative hypothesis: $H_1 : \mu_Y > 20$ or $H_1 : \mu_Y < 20$.

Hypothesis Testing

- If the null hypothesis cannot be rejected, it does not mean that it is true.
- It just means that there is not enough statistical evidence from the sample to reject the null hypothesis.
- The problem is how to decide whether to accept the null hypothesis or reject it using the evidence in a randomly selected sample of data. We will using the following steps:
    1. State the null H_0 and the alternative H_1,
    2. Choose a significance level for the test,
    3. Calculate a test statistics to test H_0,
    4. Decide whether to reject H_0 or not.
- Thus, we need to determine (i) a significance level, (ii) a test statistic and (iii) the rejection regions.

Hypothesis Testing

- Consider the errors you can make in a testing problem.
  - We reject a TRUE H_0, i.e., Type-I error.
  - We fail to reject a FALSE H_0, i.e., Type-II error.
- We cannot simultaneously minimize the likelihood of committing Type-I and Type-II errors.
- Hence, we choose a significance level, denoted by α, for committing the Type-I error, and then minimize the likelihood of committing Type-II error.
- You can think of the significance level of the test as your tolerance for rejecting a TRUE null.
- The conventional levels are 1%, 5% and 10%. If we set $\alpha = 0.05$, this means we are OK with rejecting a TRUE null 5 times out of 100 resamples.

Hypothesis Testing

- Let Y denote the hourly earnings of recent college graduates and assume $Y \sim (\mu_Y, \sigma_Y^2)$.
- We conjecture that college graduates should earn 20 per hour, $H_0 : \mu_Y = 20$ against $H_1 : \mu_Y \neq 20$.
- Suppose we draw a random sample of n observations on Y using simple random sampling, $Y_1, Y_2, \ldots, Y_n$.
- We can use the sample average $\bar{Y}$ to estimate μ_Y (and to test hypotheses about μ_Y).
- We can use the sample variance $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ to estimate σ_Y^2, because $s_Y^2 \xrightarrow{p} \sigma_Y^2$ if $Y_1, Y_2, \ldots, Y_n$ are i.i.d. and $\mathrm{E}(Y^4)$ is finite.

Hypothesis Testing

- We use the following t statistic as our test statistic:

$$t = \frac{\bar{Y} - \mu_Y}{s_{\bar{Y}}} \tag{1}$$

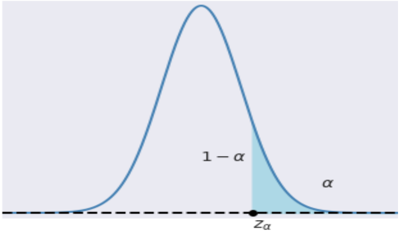
where $s_{\bar{Y}} = \sqrt{s_Y^2/n} = s_Y/\sqrt{n}$.

- It can be shown that (using a CLT for i.i.d. observations), under H_0, the sampling distribution of t statistic can be approximated by the standard normal distribution when n is large, i.e.,

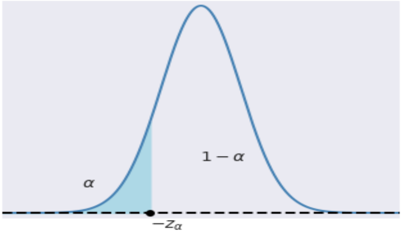
$$t = \frac{\bar{Y} - \mu_Y}{s_{\bar{Y}}} \overset{A}{\sim} N(0,1). \tag{2}$$

Hypothesis Testing

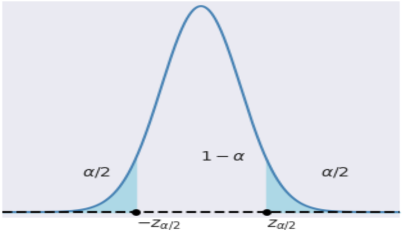
- We can now use the significance level and the distribution of the test to define the rejection regions.
- Figure 1 illustrates the rejection regions according to the type of H_1.
  1. In each case, the shaded blue area is the level of test.
  2. Depending on the type of H_1, z_α, $-z_\alpha$, $-z_{\alpha/2}$ and $z_{\alpha/2}$ are critical values.



(a) $H_0 : \mu_Y = 20$ versus $H_1 : \mu_Y > 20$



(b) $H_0 : \mu_Y = 20$ versus $H_1 : \mu_Y < 20$



(c) $H_0 : \mu_Y = 20$ versus $H_1 : \mu_Y \neq 20$

Figure 1: Rejection Regions

Hypothesis Testing

Example 2

Consider $H_1 : \mu_Y \neq 20$ in Figure 1c, and assume that $\alpha = 5\%$. Then, the critical value on the left tail is the 2.5th percentile of $N(0,1)$, which is $-z_{\alpha/2} = -1.96$, and the critical value on the right tail is the 97.5th percentile of $N(0,1)$, which is $z_{\alpha/2} = 1.96$. In R, we can use the following:

```
# Using R
qnorm(0.025,mean = 0,sd=1,lower.tail = TRUE) = -1.96

# Using Python
import scipy.stats as stats
stats.norm.ppf(0.025, loc=0, scale=1) = -1.96
```

Assume that we obtained $t = \frac{\bar{Y}-\mu_Y}{s_{\bar{Y}}} = 2$. Then, we will reject H_0 because the test statistic value is in the rejection region. i.e., $2 > z_{\alpha/2} = 1.96$.

Hypothesis Testing

- Alternatively, we can also calculate a p-value to decide between H_0 and H_1.
- You can think of it as restating the test statistic so that it becomes easier to decide. The $p-$value can be defined as

$$p\text{-value} = \begin{cases} P_{H_0}\left(t > |t_{\text{calc}}|\right) \text{ for } H_1 : \mu_Y > 20, \\ P_{H_0}\left(t < -|t_{\text{calc}}|\right) \text{ for } H_1 : \mu_Y < 20, \\ P_{H_0}\left(|t| > |t_{\text{calc}}|\right) \text{ for } H_1 : \mu_Y \neq 20. \end{cases}$$

  where

  1. P_{H_0} means this probability is calculated from the distribution of the test statistic under the null hypothesis, and
  2. t_{calc} is the value of the test statistic obtained from (1).

- What does the p-value tell us?
- The likelihood of obtaining a test statistic value that is more extreme than the actual one when the null is correct.
- Hence, the smaller p-value, the less likely that the null is correct.

Hypothesis Testing

- Since the asymptotic distribution of t statistic is $N(0,1)$, we have

$$p\text{-value} = \begin{cases} P_{H_0}(t > |t_{\text{calc}}|) = 1 - \Phi(|t_{\text{calc}}|) \text{ for } H_1 : \mu_Y > 20, \\ P_{H_0}(t < -|t_{\text{calc}}|) = \Phi(-|t_{\text{calc}}|) \text{ for } H_1 : \mu_Y < 20, \\ P_{H_0}(|t| > |t_{\text{calc}}|) = 2 \times \Phi(-|t_{\text{calc}}|) \text{ for } H_1 : \mu_Y \neq 20. \end{cases}$$

Example 3

Consider $H_1 : \mu_Y \neq 20$ in Figure 1c, and assume that $\alpha = 5\%$. Assume that we obtained $t = \frac{\bar{Y} - \mu_Y}{s_{\bar{Y}}} = 2$. Then,

$$\text{p-value} = P_{H_0}(|t| > 2) = P_{H_0}(t > 2) + P_{H_0}(t < -2) = 2 \times \Phi(-2).$$

In R, we can compute the p-value in the following way:

```
# Using R
2*pnorm(-2,mean = 0,sd=1,lower.tail = TRUE)= 0.045

# Using Python
from scipy import stats
2 * stats.norm.cdf(-2, loc=0, scale=1)= 0.045
```

Since $p-$value $= 0.045 < \alpha = 0.05$, we reject H_0.

Hypothesis Testing

- Recall that we used a CLT to figure out the distribution of the t test.
- In other words, it is approximately normal and approximation works well in large samples.
- What if we only have a few observations in the sample, say less than 30?
- If you are willing to assume that $Y \sim N(\mu_Y, \sigma_Y^2)$ and hence $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu_Y, \sigma_Y^2)$, then t statistic in (1) has the Student t distribution with $n-1$ degrees of freedom.
- This is exact. It is not an approximation.
- Recall that for $n > 30$, the t-distribution and $N(0,1)$ are very close (as n grows large, the t_{n-1} distribution converges to $N(0,1)$.
- According to Stock and Watson (2020), the t-distribution is an artifact from days when sample sizes were small and "computers" were people.

Confidence Intervals for the Population Mean

- Following the example from previous slides, Y denote the hourly earnings of recent college graduates.
- Although we do not know the mean earnings for recent college graduates, μ_Y, we can use data from a random sample to construct a set of values that contains the true population mean μ_Y with a certain pre-specified probability.
- Such a set is called a confidence set.
- The prespecified probability that μ_Y is contained in this set is called the confidence level.
- Since μ_Y is a scalar in our examples, the confidence set is an interval, called a confidence interval (CI).

Confidence Intervals for the Population Mean

- Suppose we choose the prespecified probability as 95%.

Definition 4

The 95% two-sided CI for μ_Y is an interval constructed so that it contains the true value of μ_Y in 95% of all possible random samples.

- Hence, it is equivalent to repeatedly testing a null hypothesis on μ_Y at the 5% significance level over all possible random samples.
- Then, the 95% CI for μ_Y by definition refers to the set of values for μ_Y such that $P\left(\left\{\mu_Y : \left|\frac{\bar{Y}-\mu_Y}{s_Y/\sqrt{n}}\right| \leq 1.96\right\}\right) = 0.95$.
- Hence,

$$
\begin{aligned}
&\left\{\mu_Y : \left|\frac{\bar{Y}-\mu_Y}{s_Y/\sqrt{n}}\right| \leq 1.96\right\} = \left\{\mu_Y : -1.96 \leq \frac{\bar{Y}-\mu_Y}{s_Y/\sqrt{n}} \leq 1.96\right\} \\
&= \left\{\mu_Y : -\bar{Y} - 1.96 \times s_Y/\sqrt{n} \leq -\mu_Y \leq -\bar{Y} + 1.96 \times s_Y/\sqrt{n}\right\} \\
&= \left\{\mu_Y : \bar{Y} - 1.96 \times s_Y/\sqrt{n} \leq \mu_Y \leq \bar{Y} + 1.96 \times s_Y/\sqrt{n}\right\} \\
&= \left\{\mu_Y : \bar{Y} \pm 1.96 \times \text{SE}(\bar{Y})\right\}
\end{aligned}
$$

Confidence Intervals for the Population Mean

- If you choose another prespecified probability, say 90% or 99%, you only need to change the critical value in the formula, i.e.,

$$\left\{\mu_Y : \bar{Y} \pm 1.64 \times \mathsf{SE}(\bar{Y})\right\} \quad \text{or} \quad \left\{\mu_Y : \bar{Y} \pm 2.58 \times \mathsf{SE}(\bar{Y})\right\}.$$

- How do we use the CI to test a null hypothesis on μ_Y?
- After calculating the CI for μ_Y, say 95% CI,
  - reject the null hypothesis if the null value of μ_Y is not contained in the CI (at the 5% level),
  - fail to reject the null hypothesis if the null value of μ_Y is contained in the CI (at the 5% level).

Confidence Intervals for the Population Mean

Example 4

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.

(a) What is the $p-$value for the test of $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$?

(b) Construct a 95% confidence interval for p.

(c) Construct a 99% confidence interval for p.

(d) Why is the interval in (c) wider than the interval in (b)?

(e) Did the survey contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey? Explain.

Solution to Example 4.

(a) Note that we have $\hat{p} = \frac{215}{400} = 0.5375$. Also from Example 1, we have $\mathrm{Var}(\hat{p}) = p(1-p)/n$. Thus, $\widehat{\mathrm{Var}}(\hat{p}) = \hat{p}(1-\hat{p})/n = 6.2148 \times 10^{-4}$. Then,

$$t_{\text{calc}} = \frac{\hat{p} - 0.5}{SE(\hat{p})} = \frac{0.5375 - 0.5}{\sqrt{6.2148 \times 10^{-4}}} = 1.506.$$

Then, the $p-$value is $P_{H_0}(t > |t_{\text{calc}}|) = 1 - \Phi(|t_{\text{calc}}|) = 1 - \Phi(1.506)$, which can be computed as

```
p-value=1-pnorm(1.506,mean=0,sd=1,lower.tail =T)=0.066.
```

(b) The 95% confidence interval for p is

$$\hat{p} \pm 1.96 \times SE(\hat{p}) = 0.5375 \pm 1.96 \times 0.0249 = (0.4887,\ 0.5863).$$

(c) The 99% confidence interval for p is

$$\hat{p} \pm 2.57 \times SE(\hat{p}) = 0.5375 \pm 2.57 \times 0.0249 = (0.4735,\ 0.6015).$$

(d) Mechanically, the interval in (c) is wider because of a larger critical value (2.57 versus 1.96). Substantively, the 99% confidence interval is wider than the 95% confidence because the 99% confidence interval must contain the true value of p in 99% of all possible samples, while a 95% confidence interval must contain the true value of p in only 95% of all possible samples.

(e) According to (a), (b) and (c), we fail to reject H_0. Thus, the survey did not contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey.

□

Hypothesis Tests for the Difference Between Two Means

- Suppose you would like to test for the difference between mean earnings of female college graduates and mean earnings of male college graduates.

  1. Let μ_w be the mean hourly earnings in the population of women recently graduated from college.
  2. Let μ_m be the mean hourly earnings in the population of men recently graduated from college.

- Consider the null hypothesis that mean earnings for these two populations differ by a certain amount, say d_0.

- The null hypothesis and the two-sided alternative hypothesis are

$$H_0 : \mu_m - \mu_w = d_0 \quad \text{vs.} \quad H_1 : \mu_m - \mu_w \neq d_0.$$

- Because these population means are unknown, they must be estimated from samples of men and women.

Hypothesis Tests for the Difference Between Two Means

- Suppose we have samples of n_m men and n_w women drawn at random from their populations.
- Let the sample average annual earnings be $\bar{Y}_m$ for men and $\bar{Y}_w$ for women.
- An estimator of $\mu_m - \mu_w$ is $\bar{Y}_m - \bar{Y}_w$.
- Due to simple random sampling we can invoke the CLT to write $\bar{Y}_m \overset{A}{\sim} N(\mu_m, \frac{\sigma_m^2}{n_m})$ and $\bar{Y}_w \overset{A}{\sim} N(\mu_w, \frac{\sigma_w^2}{n_w})$.
- Recall that a weighted average of two normal random variables is itself normally distributed.
- Because $\bar{Y}_m$ and $\bar{Y}_w$ are constructed from different randomly selected samples, they are independent random variables, and

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

Hypothesis Tests for the Difference Between Two Means

- We cannot use this result yet to test for the null since σ_m^2 and σ_w^2 are unknown.

- They must be estimated.

- Replace them with their sample counterparts, $s^2_{Y_m}$ and $s^2_{Y_w}$.

- The test statistic is the following t statistic:

$$t = \frac{(\bar{Y}_m - \bar{Y}_w) - d_0}{\sqrt{\frac{s^2_{Y_m}}{n_m} + \frac{s^2_{Y_w}}{n_w}}} \overset{A}{\sim} N(0,1).$$

- You can also easily calculate a confidence interval for $d = \mu_m - \mu_w$, say the 95% CI for

$$\{d : (\bar{Y}_m - \bar{Y}_w) \pm 1.96 \times \mathsf{SE}(\bar{Y}_m - \bar{Y}_w)\},$$

where $\mathsf{SE}(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s^2_{Y_m}}{n_m} + \frac{s^2_{Y_w}}{n_w}}$ is the standard error of $(\bar{Y}_m - \bar{Y}_w)$.

The Gender Gap in Earnings of College Graduates in the U.S.

- Table 1 includes the average earnings of male and female college graduates over the period 1992-2008.

Table 1: The gender gap in earnings

	Men			Women			Difference		
year	$\bar{Y}_m$	s_{Y_m}	n_m	$\bar{Y}_w$	s_{Y_w}	n_w	$(\bar{Y}_m - \bar{Y}_w)$	SE$(\bar{Y}_m - \bar{Y}_w)$	95% CI
1992	23.27	10.17	1594	20.05	7.87	1368	3.22***	0.33	[2.58,3.88]
1996	22.48	10.10	1379	18.98	7.95	1230	3.50***	0.35	[2.80,4.19]
2000	24.88	11.60	1303	20.74	9.36	1181	4.14***	0.42	[3.32,4.97]
2004	25.12	12.01	1894	21.02	9.36	1735	4.10***	0.36	[3.40,4.80]
2008	24.98	11.78	1838	20.87	9.66	1871	4.11***	0.35	[3.41,4.80]

- The estimates in the table are computed using data on full-time workers aged 25-34 surveyed in the CPS conducted in March of the specified year.

- *** implies the difference is significantly different from zero at the 1% significance level.

- What are the results we can derive from the table?

The Gender Gap in Earnings of College Graduates in the U.S.

- The gender gap estimates are large: an hourly gap of $\$4.11$ adds up to $\$8220$ assuming 40-hour work week and 50 paid weeks per year.
- We also see an increase in the gender gap estimates over time (in real terms).
- The gap is also large if it is measured in percentage terms. In 2008, women earned $\$4.11/\$24.98 = 16\%$ less per hour than men did.
- It seems the gap in hourly earnings is large and has been stable over the recent past.
- The analysis does not tell us why the gap exists.
- Does it arise from the gender discrimination in the labor market?
- Does it reflect differences in skills, experience, or education between men and women?
- Does it reflect differences in choice of jobs?

Bibliography I

Hanck, Christoph et al. (2021). *Introduction to Econometrics with R.* URL: https://www.econometrics-with-r.org/index.html.

Stock, James H. and Mark W. Watson (2020). *Introduction to Econometrics.* Fourth. Pearson.