Introductory Applied Econometrics Analysis using Stata

November 14 – 18, 2016

Dushanbe, Tajikistan

Allen Park and Jarilkasin Ilyasov

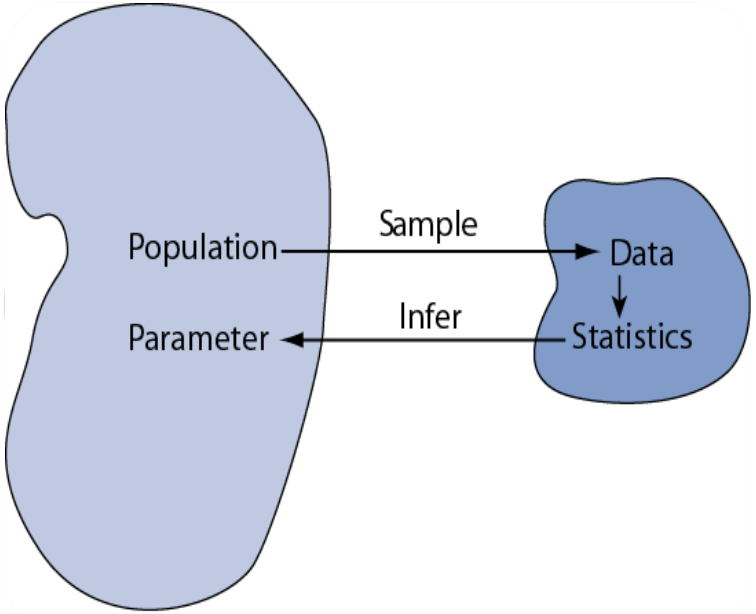
Review of Statistics

Based on Chapter 3. Stock and Watson. “Introduction to Econometrics” 3rd Edition.

Review of Statistics

- Statistics is the science of using data to learn about the world around us → helps us answers questions about unknown characteristics of distributions in populations of interest.
- Statistical inference is the act of generalizing from a sample to a population with calculated degree of certainty.

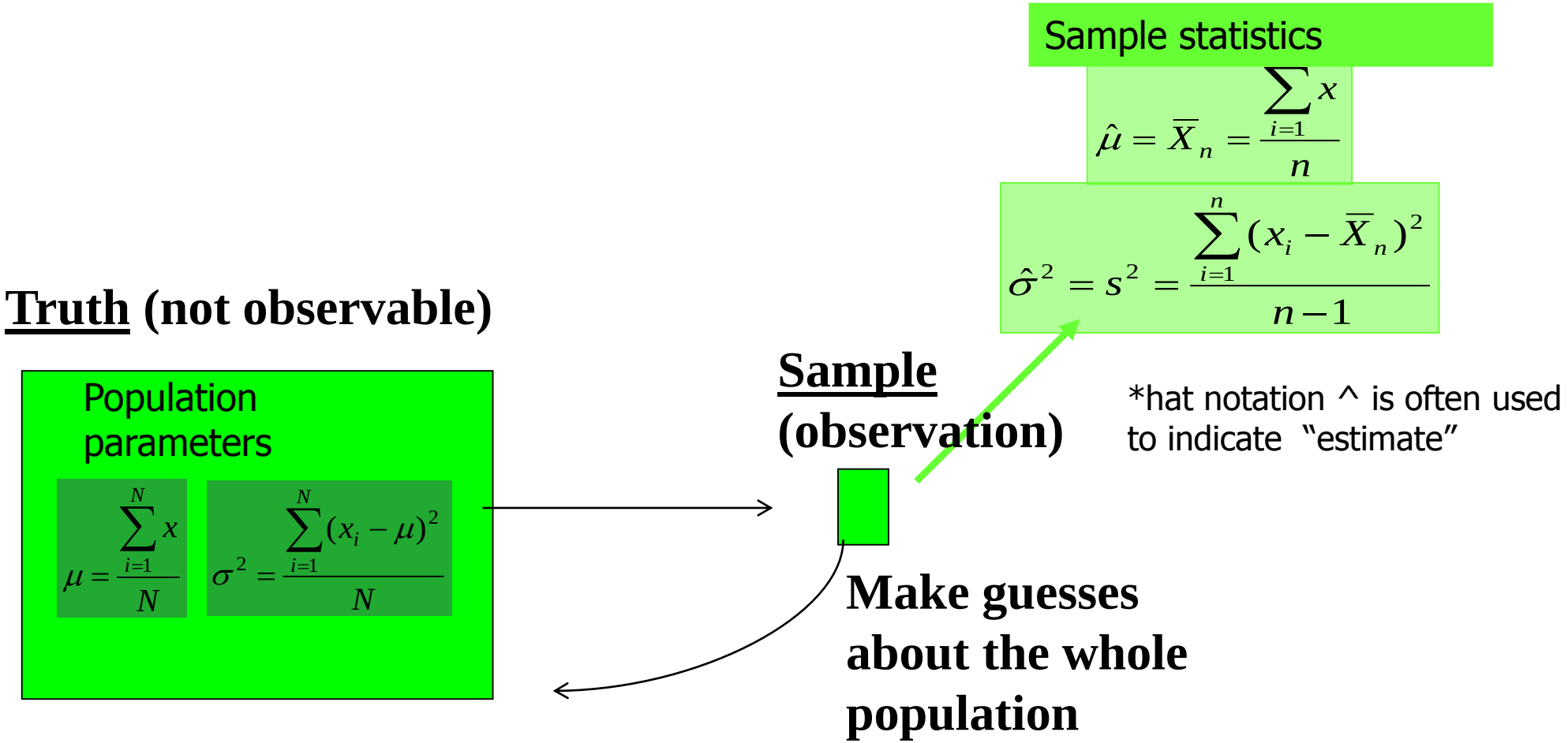
We want to learn about population *parameters* …



…but we can only calculate *sample statistics*

Review of Statistics

Population and sample



Review of Statistics

Statistics vs. Parameters

Sample Statistic – any summary measure calculated from data; e.g., could be a mean, a difference in means or proportions, a correlation coefficient, etc.

Population parameter – the true value/true effect in the entire population of interest

Review of Statistics

Three types of statistical methods are used:

1. **Estimation** – entails computing a "best guess" numerical value for an unknown characteristic of a population distribution, such as its mean, from a sample data
2. **Hypothesis Testing** – entails formulating a specific hypothesis about the population, then use sample evidence to decide whether it is true
3. **Confidence Intervals** – uses set of data to estimate an interval or range for an unknown population characteristics

Review of Statistics

Three types of statistical methods are used:

- $\bar{Y}$ (sample average) is a natural way to estimate μ_Y (mean value of Y in a population)

  Why?? i.i.d. observations

Three specific characteristics:

- Unbiasedness → $\hat{\mu}_Y$ is unbiased estimator if $E(\hat{\mu}_Y) = \mu_Y$
- Consistency → $\hat{\mu}_Y$ is consistent estimator if $\hat{\mu}_Y \xrightarrow{p} \mu_Y$
- Variance and efficiency → $\hat{\mu}_Y$ is more efficient estimator than $\tilde{\mu}_Y$ if $\text{var}(\hat{\mu}_Y) < \text{var}(\tilde{\mu}_Y)$

Review of Statistics

- $\bar{Y}$ (sample average) is the least squares estimator of μ_Y
  - It provides the best fit to the data in the sense that average squared differences between the observations and $\bar{Y}$ are the smallest of all possible estimators.

Review of Statistics

Hypothesis Tests concerning the population mean

- Many hypotheses can be phrased as yes/no questions
  - e.g. Are mean earnings the same for women and men?
- Starting point in hypothesis testing is specifying the hypothesis be tested → **null hypothesis**

$$H_0 : E(Y) = \mu_{Y,0}$$

  – a hypothesis can usually be proven true or false. A null hypothesis is a hypothesis that researcher or experimenter will try to disprove or discredit.

- The **alternative hypothesis** specifies what is true if the null hypothesis is not. $H_1 : E(Y) \neq \mu_{Y,0}$ (two-sided alternative)

Review of Statistics

- **One sided and two-sided alternative hypotheses is also possible**
  - $H_0 : E(Y) = \mu_{Y,0}$ vs. $H_1 : E(Y) \rangle \mu_{Y,0}$ (1-sided, >)
  - $H_0 : E(Y) = \mu_{Y,0}$ vs. $H_1 : E(Y) \langle \mu_{Y,0}$ (1-sided, <)
  - $H_0 : E(Y) = \mu_{Y,0}$ vs. $H_1 : E(Y) \neq \mu_{Y,0}$ (2-sided)

Important: what if null hypothesis is "accepted"?

- **If "accepted" or null cannot be rejected, this doesn't mean that the statistician declares it to be true.**
- **Statistical hypothesis testing can be posed as either rejecting the null hypothesis or failing to do so**

Review of Statistics

Rarely $\bar{Y}$ exactly equals population mean $\mu_{Y,0}$

	Test result (H_0 True)	Test result (H_0 False)
True state (H_0 True)	Correct Decision	Type I Error
True state (H_0 False)	Type II Error	Correct Decision

$$\alpha = P(Type\ I\ Error) \quad \beta = P(Type\ II\ Error)$$

Goal: Keep α, β reasonably small

Review of Statistics

Some terminology for testing statistical hypotheses:

- ***p-value*** = probability of drawing a statistic (e.g. $\bar{Y}$) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.
- The ***significance level*** of a test is a pre-specified probability of incorrectly rejecting the null, when the null is true.

Calculating the p-value based on $\bar{Y}$:

$$\text{p-value} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$$

Where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed (nonrandom)

Review of Statistics

- Small p-values mean the null value is unlikely given our data.
- By convention, p-values of <.05 are often accepted as “statistically significant” in the medical literature; but this is an arbitrary cut-off.
- A cut-off of p<.05 means that in about 5 of 100 experiments, a result would appear significant just by chance (“Type I error”).

Review of Statistics

- To compute the *p*-value, you need the to know the sampling distribution of $\bar{Y}$, which is complicated if *n* is small.
- If *n* is large, you can use the normal approximation (CLT):

$$\text{p-value} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right|\right]$$

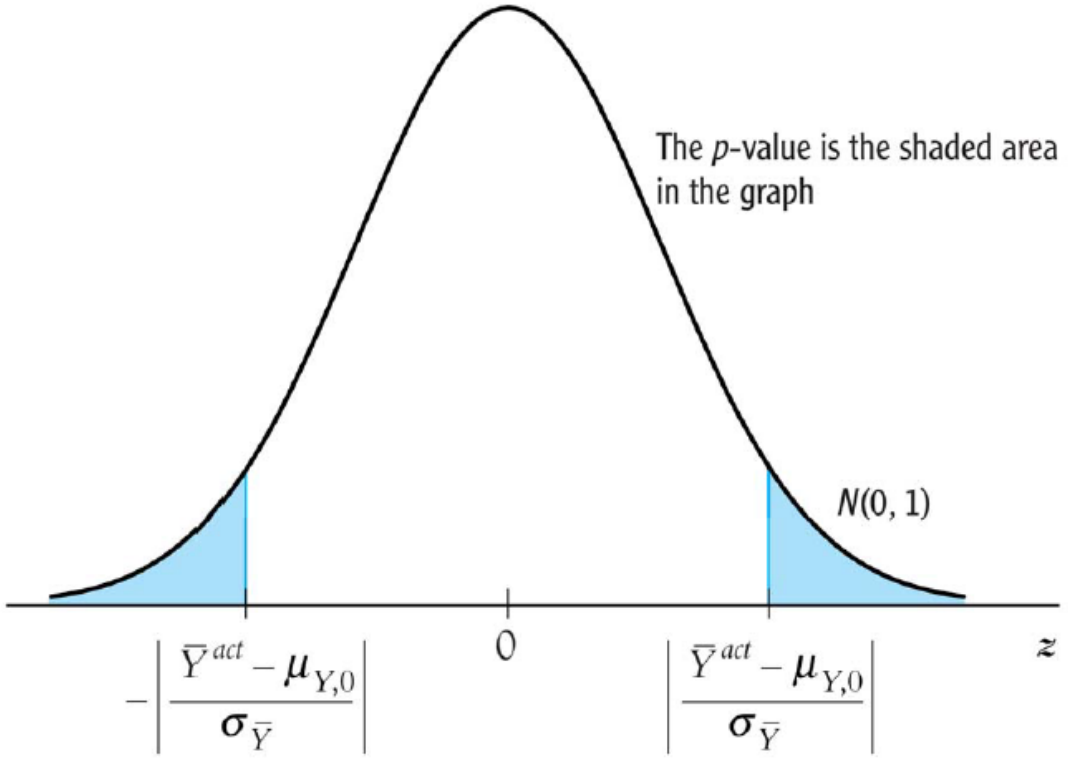
$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right|\right] = \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right]$$

$$\cong \text{probability under left + right N(0,1) tails}$$

where $\sigma_{\bar{Y}}$ (= std.dev. of the distribution of $\bar{Y}$) = $\sigma_Y / \sqrt{n}$

Review of Statistics

Calculating the p-value with σ_Y known:



For large *n*, p-value = the probability that a *N*(0,1) random variable falls outside

$$|(\bar{Y}^{act} - \mu_{Y,0}) / \sigma_{\bar{Y}}|$$

In practice, $\sigma_{\bar{Y}}$ is unknown – it must be estimated

Review of Statistics

***Estimator of the variance of Y*:**

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{“sample variance of } Y\text{”}$$

Fact:

$$\text{If } (Y_1,\ldots,Y_n) \text{ are i.i.d. and } E(Y^4) < \infty, \text{ then } s_Y^2 \xrightarrow{p} \sigma_Y^2$$

Why does the law of large numbers apply?

- Because s_Y^2 is a sample average; see Appendix 3.3
- Technical note: we assume $E(Y^4) < \infty$ because here the average is not of Y_i, but of its square; see App. 3.3

Review of Statistics

Computing the p-value with σ_Y^2 estimated:

$$p\text{-value} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|],$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right]$$

$$\cong \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right] \quad \text{(large } n\text{)}$$

Review of Statistics

Computing the p-value with σ_Y^2 estimated (continued):

So

$$\boxed{p\text{-value} = \Pr_{H_0}[|t| > |t^{act}|]} \quad (\sigma_Y^2 \textbf{ estimated})$$

$$\cong \text{probability under normal tails outside } |t^{act}|$$

$$\text{where } t = \frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}} \text{ (the usual } t\text{-statistic)}$$

$s_Y / \sqrt{n}$ = standard error of $\bar{Y}$

Review of Statistics

The Standard Error of $\bar{Y}$

- The standard error of $\bar{Y}$ is an estimator of the standard deviation of $\bar{Y}$. The standard error of $\bar{Y}$ is denoted by:

  $SE(\bar{Y})$ or $\hat{\sigma}_{\bar{Y}}$. When $Y_1,....,Y_n$ are i.i.d.,

  $$SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = s_Y / \sqrt{n}$$

Review of Statistics

What is the link between the *p*-value and the significance level?

- The significance level is prespecified. For example, if the prespecified significance level is 5%,
  - you reject the null hypothesis if $|t| \geq 1.96$
  - equivalently, you reject if $p \leq 0.05$
  - The *p*-value is sometimes called the ***marginal significance level***.
  - Often, it is better to communicate the *p*-value than simply whether a test rejects or not – the *p*-value contains more information than the "yes/no" statement about whether the test rejects.

Review of Statistics

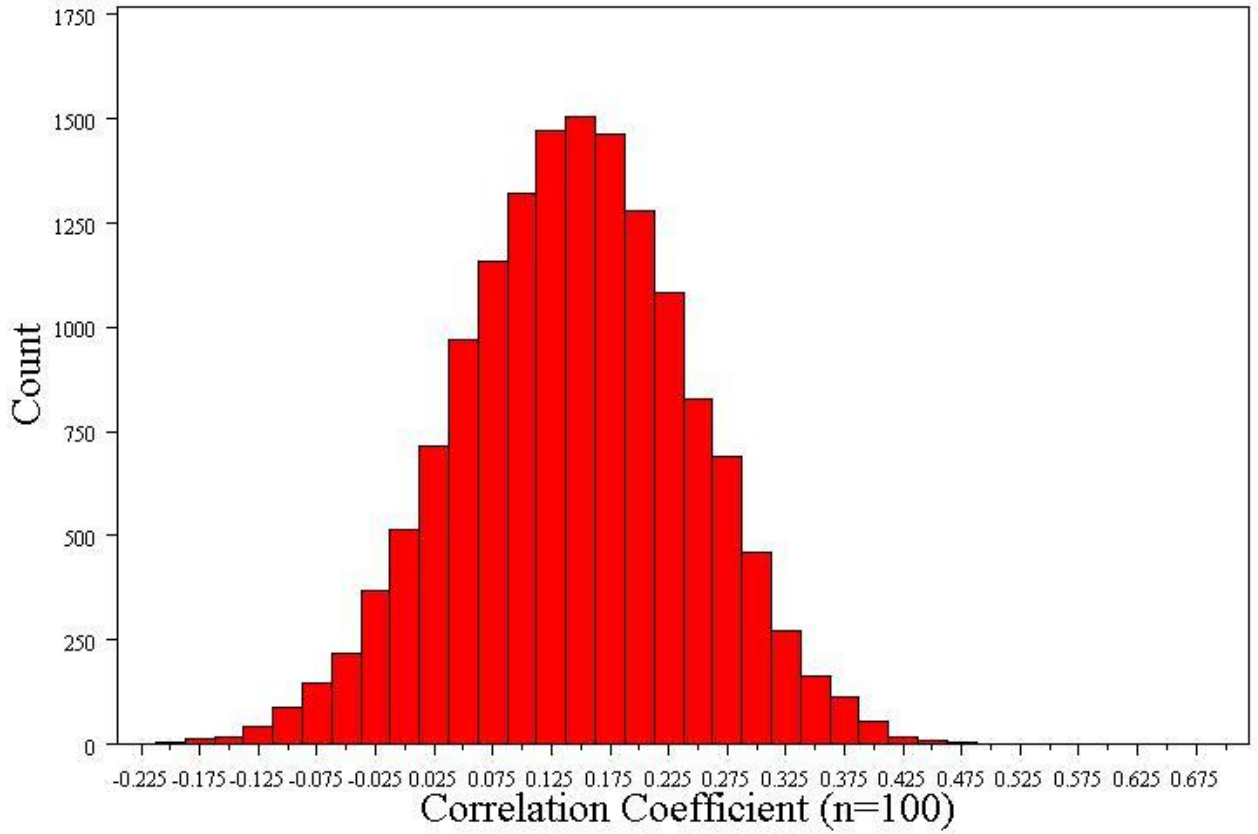
Digression: **What happened to distribution of a correlation coefficient?**

E.g.:

- 1. Specify the true correlation coefficient
  - Correlation coefficient = 0.15
- 2. Select a random sample of 100 virtual men from the population.
- 3. Calculate the correlation coefficient for the sample.
- 4. Repeat steps (2) and (3) 15,000 times
- 5. Explore the distribution of the 15,000 correlation coefficients.

Review of Statistics

Distribution of a correlation coefficient…



Normally distributed!

Mean = 0.15 (true correlation)

Standard error = 0.10

Review of Statistics

- 1. Shape of the distribution
  - Normally distributed for large samples
  - T-distribution for small samples (n<100)
- 2. Mean = true correlation coefficient (r)
- 3. Standard error $\approx \frac{1-r^2}{\sqrt{n}}$

Review of Statistics

Many statistics follow normal (or t-distributions)…

- Means/difference in means
  - T-distribution for small samples
- Proportions/difference in proportions
- Regression coefficients
  - T-distribution for small samples
- Natural log of the odds ratio

Review of Statistics

What happened to the *t*-table and the degrees of freedom?

Digression: the Student *t* distribution

- If $Y_i, i = 1, \ldots, n$ is i.i.d. $N(\mu_Y, \sigma_Y^2)$, then the *t*-statistic has the Student *t*-distribution with $n - 1$ degrees of freedom.
- The critical values of the Student *t*-distribution is tabulated in the back of all statistics books. Remember the recipe?
  - 1. Compute the *t*-statistic
  - 2. Compute the degrees of freedom, which is $n - 1$
  - 3. Look up the 5% critical value
  - 4. If the *t*-statistic exceeds (in absolute value) this critical value, reject the null hypothesis.

Review of Statistics

Comments on this recipe and the Student *t*-distribution

- 1. The theory of the *t*-distribution was one of the early triumphs of mathematical statistics. It is astounding, really: if *Y* is i.i.d. normal, then you can know the *exact*, *finite sample* distribution of the *t*-statistic – it is the Student *t*. So, you can construct confidence intervals (using the Student *t* critical value) that have *exactly* the right coverage rate, no matter what the sample size. This result was really useful in times when "computer" was a job title, data collection was expensive, and the number of observations was perhaps a dozen. It is also a conceptually beautiful result, and the math is beautiful too – which is probably why stats profs love to teach the *t*-distribution. But….

Review of Statistics

Comments on Student t distribution, ctd.

- 2. If the sample size is moderate (several dozen) or large (hundreds or more), the difference between the *t* distribution and N(0,1) critical values is negligible. Here are some 5% critical values for 2-sided tests:

degrees of freedom ($n-1$)	5% *t*-distribution critical value
10	2.23
20	2.09
30	2.04
60	2.00
∞	1.96

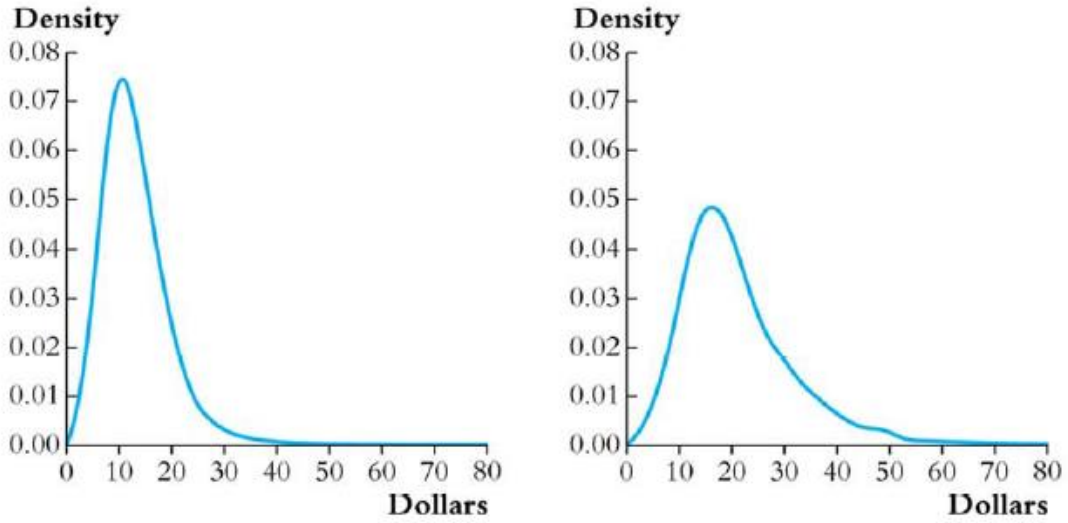
Review of Statistics

Comments on Student t distribution, ctd.

- 3. So, the Student-*t* distribution is only relevant when the sample size is very small; but in that case, for it to be correct, you must be sure that the population distribution of *Y* is normal. In economic data, the normality assumption is rarely credible. Here are the distributions of some economic data.
  - Do you think earnings are normally distributed?
  - Suppose you have a sample of $n = 10$ observations from one of these distributions – would you feel comfortable using the Student *t* distribution?

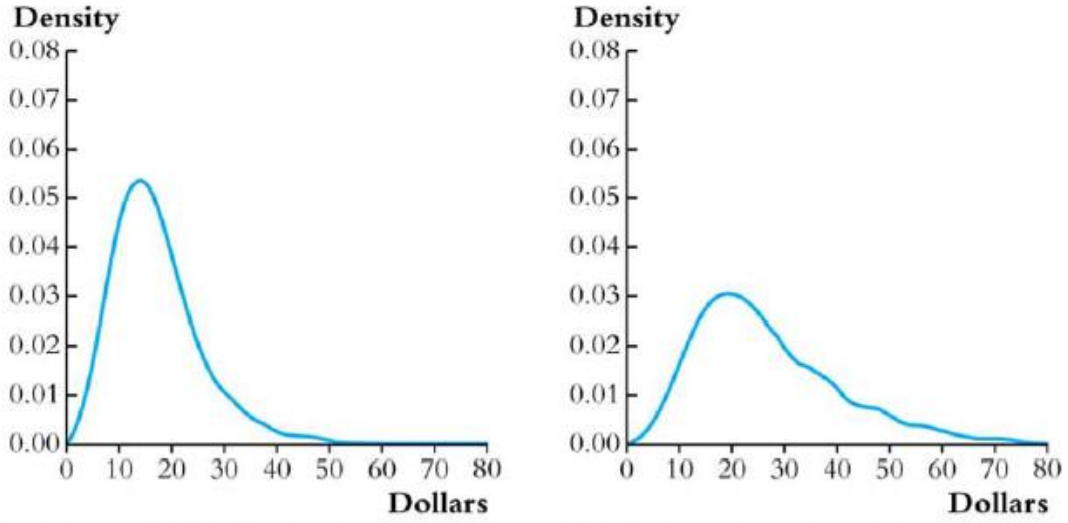
FIGURE 2.4 Conditional Distribution of Average Hourly Earnings of U.S. Full-Time Workers in 2004, Given Education Level and Gender

The four distributions of earnings are for women and men, for those with only a high school diploma (a and c) and those whose highest degree is from a four-year college (b and d).



(a) Women with a high school diploma

(b) Women with a college degree



(c) Men with a high school diploma

(d) Men with a college degree

Review of Statistics

Comments on Student t distribution, ctd.

- 4. You might not know this. Consider the *t*-statistic testing the hypothesis that two means (groups *s*, *l*) are equal:

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

- Even if the population distribution of *Y* in the two groups is normal, this statistic doesn't have a Student *t* distribution!
- There is a statistic testing this hypothesis that has a normal distribution, the "pooled variance" *t*-statistic – see SW (Section 3.6) – however the pooled variance *t*-statistic is only valid if the variances of the normal distributions are the same in the two groups. Would you expect this to be true, say, for men's v. women's wages?

Review of Statistics

The Student-t distribution – Summary

- The assumption that *Y* is distributed $N(\mu_Y, \sigma_Y^2)$ is rarely plausible in practice (Income? Number of children?)
- 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))
- The *t*-distribution is an artifact from days when sample sizes were small and "computers" were people
- For historical reasons, statistical software typically uses the *t*-distribution to compute *p*-values – but this is irrelevant when the sample size is moderate or large.
- For these reasons, focus on the large *n* approximation given by the CLT

Review of Statistics

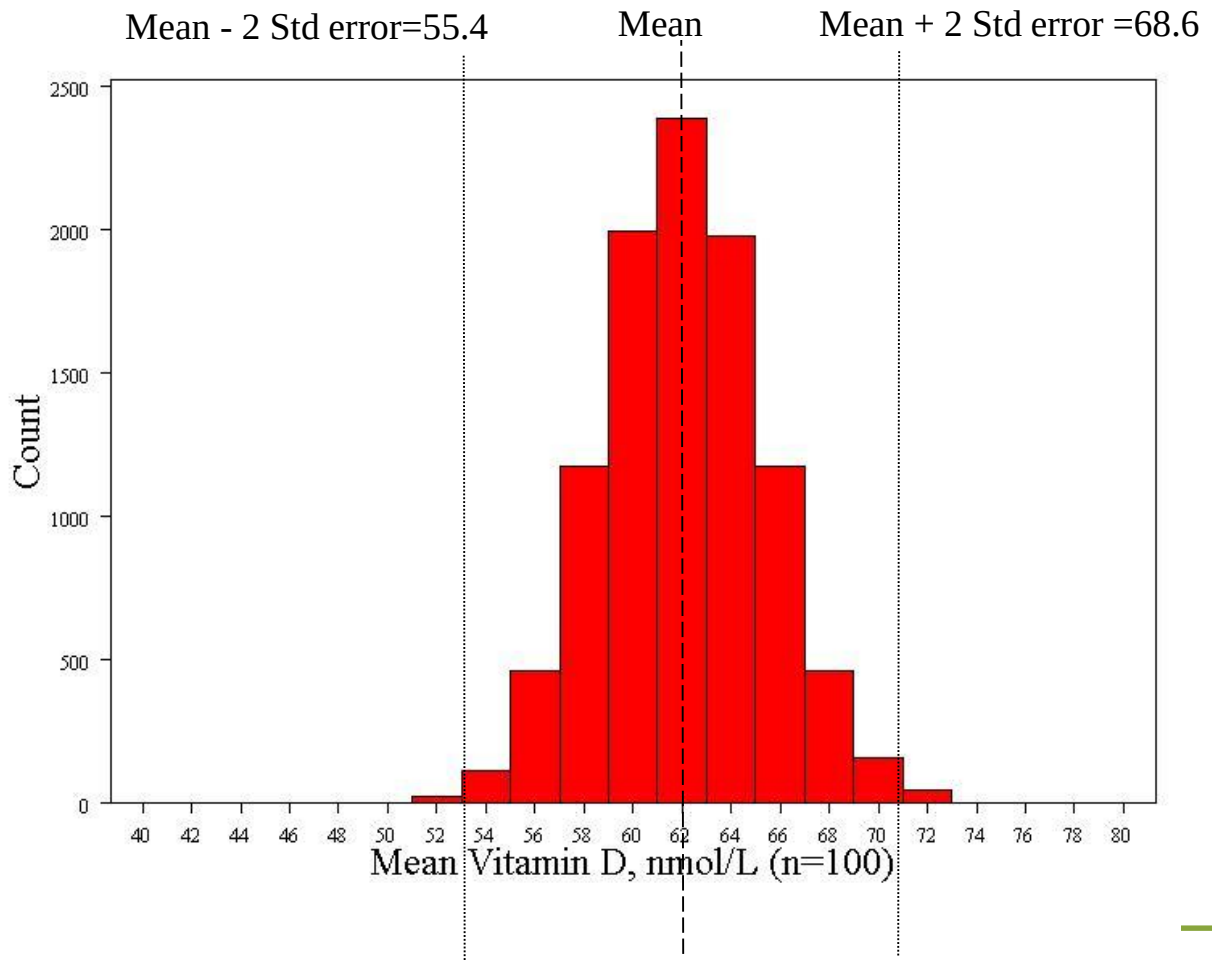
Confidence Intervals

- A 95% ***confidence interval*** for μ_Y is an interval that contains the true value of μ_Y in 95% of repeated samples.

- *Digression*: What is random here? The values of *Y*1,…,*Yn* and thus any functions of them – including the confidence interval. The confidence interval will differ from one sample to the next. The population parameter, μ_Y, is not random; we just don't know it.

Review of Statistics

Recall: 68-95-99.7 rule for normal distributions! These is a 95% chance that the sample mean will fall within two standard errors of the true mean



To be precise, 95% of observations fall between Z=-1.96 and Z= +1.96 (so the "2" is a rounded number)…

Review of Statistics

Confidence intervals, ctd.

- A 95% confidence interval can always be constructed as the set of values of μ_Y not rejected by a hypothesis test with a 5% significance level.

$$\{\mu_Y: \left|\frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}}\right| \leq 1.96\} = \{\mu_Y: -1.96 \leq \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \leq 1.96\}$$

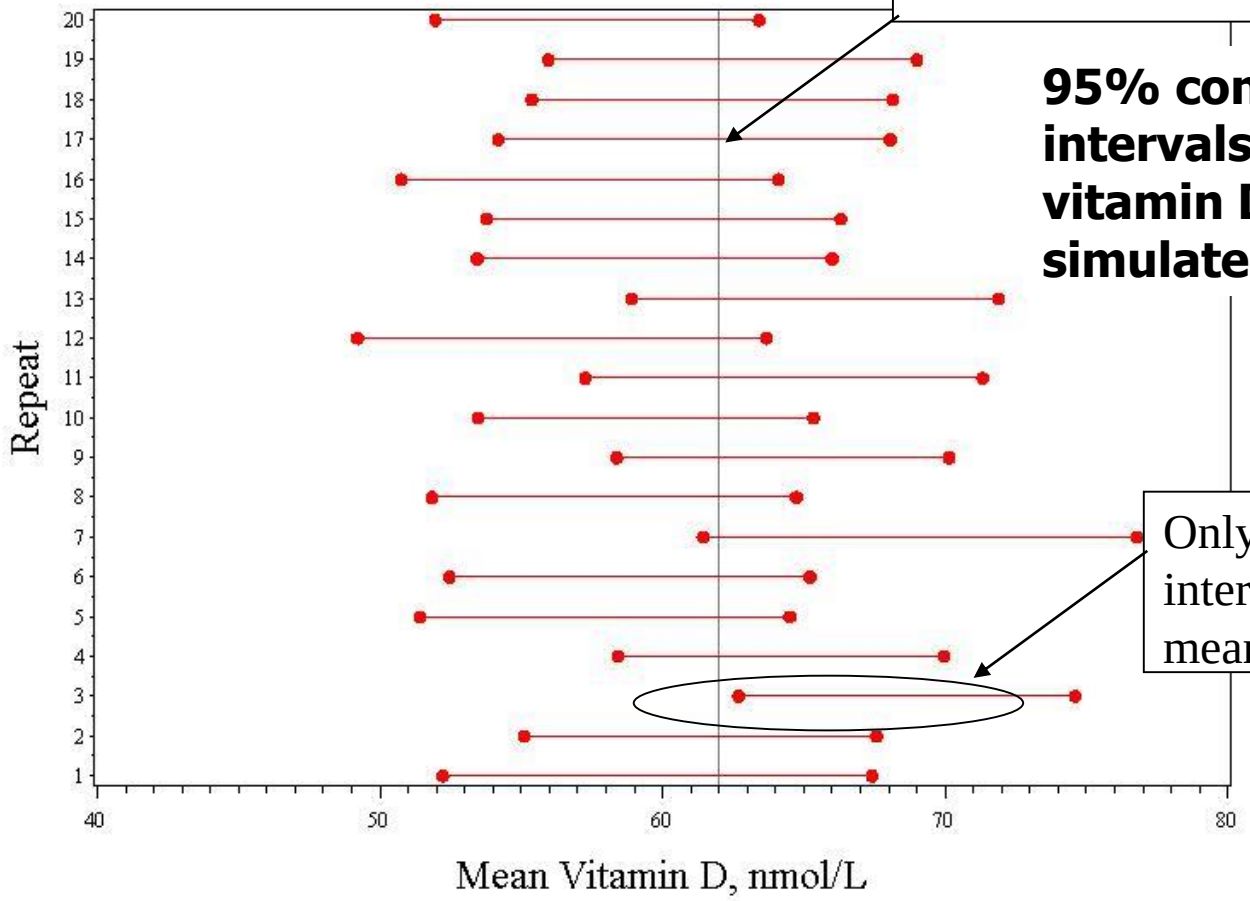
$$= \{\mu_Y: -1.96\frac{s_Y}{\sqrt{n}} \leq \bar{Y} - \mu_Y \leq 1.96\frac{s_Y}{\sqrt{n}}\}$$

$$= \{\mu_Y \in (\bar{Y} - 1.96\frac{s_Y}{\sqrt{n}}, \bar{Y} + 1.96\frac{s_Y}{\sqrt{n}})\}$$

- *This confidence interval relies on the large-n results that* $\bar{Y}$ *is approximately normally distributed and* $s_Y^2 \xrightarrow{p} \sigma_Y^2$

Review of Statistics

Simulation of 20 studies of 100 men…



Vertical line indicates the true mean (62)

95% confidence intervals for the mean vitamin D for each of the simulated studies.

Only 1 confidence interval missed the true mean.

Review of Statistics

Summary:

- From the two assumptions of:
  - (1) simple random sampling of a population, that is, {*Yi*, *i* =1,…,*n*} are i.i.d.
  - (2) $0 < E(Y^4) < \infty$

we developed, for large samples (large *n*):

- Theory of estimation (sampling distribution of $\bar{Y}$)
- Theory of hypothesis testing (large-*n* distribution of *t* statistic and computation of the *p*-value)
- Theory of confidence intervals (constructed by inverting the test statistic)
- Are assumptions (1) & (2) plausible in practice? **Yes**

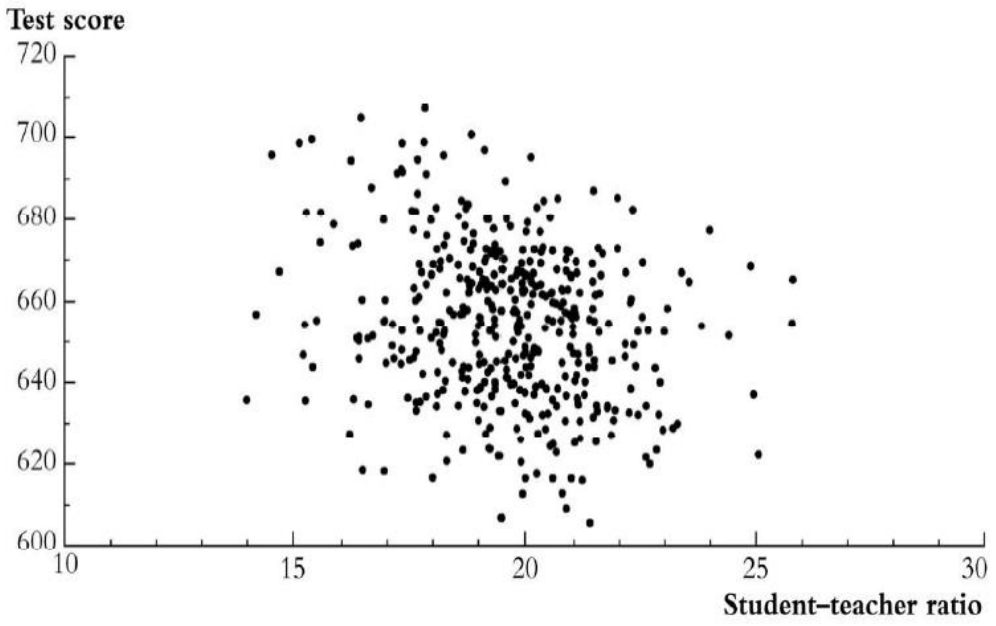
Review of Statistics

Let's go back to the original policy question:

- What is the effect on test scores of reducing STR by one student/class?
- *Have we answered this question?*

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



There is a weak negative relationship between the STR and the test scores:
The sample correlation is -0.23