Introduction to Econometrics

Review of Probability & Statistics

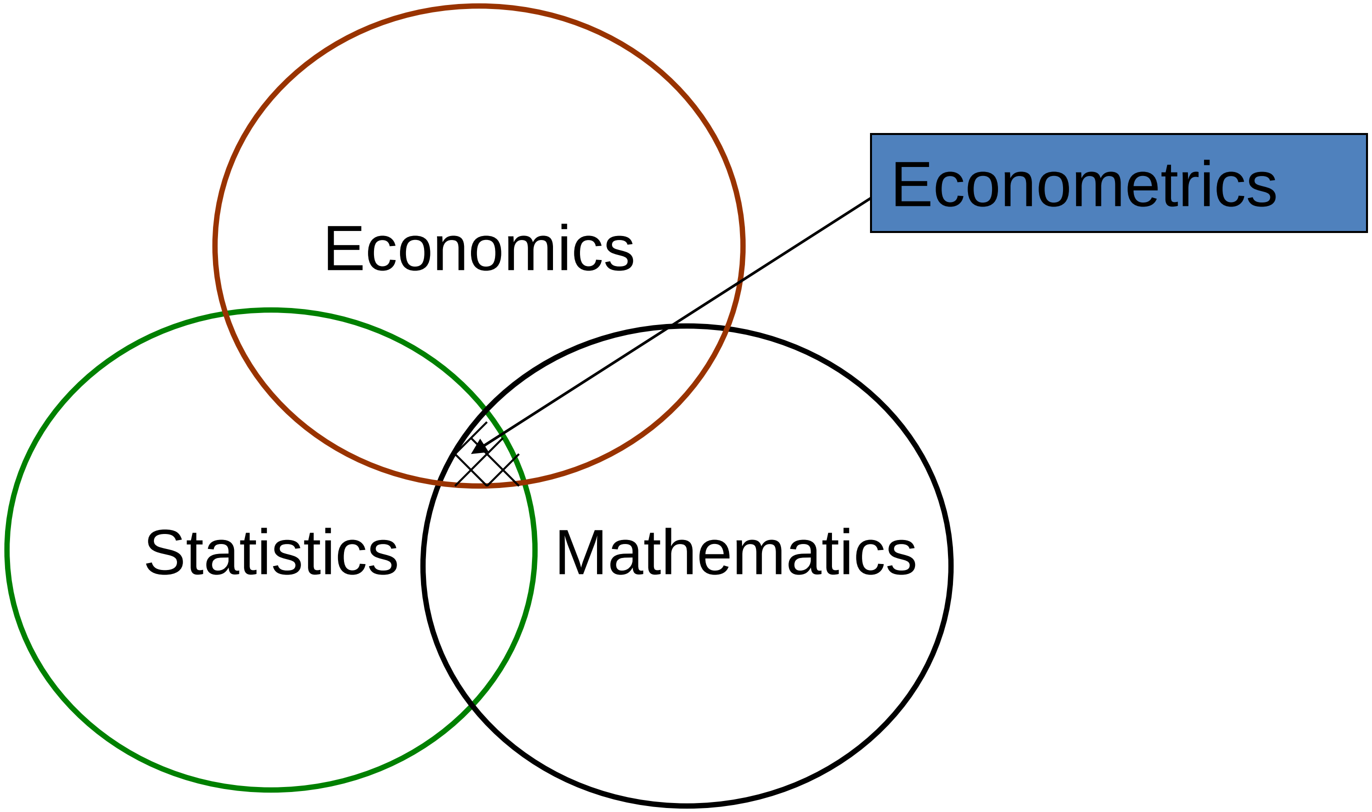
Peerapat Wongchaiwat, Ph.D.

wongchaiwat@hotmail.com

Introduction

- What is Econometrics?
- Econometrics consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results.
- Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.

What is Econometrics?



Why do we study econometrics?

- Rare in economics (and many other areas without labs!) to have experimental data
- Need to use nonexperimental, or observational data to make inferences
- Important to be able to apply economic theory to real world data

Why it is so important?

- An empirical analysis uses data to test a theory or to estimate a relationship

- A formal economic model can be tested

- Theory may be ambiguous as to the effect of some policy change – can use econometrics to evaluate the program

The Question of Causality

- Simply establishing a relationship between variables is rarely sufficient
- Want to get the effect to be considered causal
- If we’ve truly controlled for enough other variables, then the estimated effect can often be considered to be causal

Purpose of Econometrics

- Structural Analysis
- Policy Evaluation
- Economical Prediction
- Empirical Analysis

Methodology of Econometrics

- 1. Statement of theory or hypothesis.
- 2. Specification of the mathematical model of the theory.
- 3. Specification of the statistical, or econometric model.
- 4. Obtaining the data.
- 5. Estimation of the parameters of the econometric model.
- 6. Hypothesis testing.
- 7. Forecasting or prediction.

Example： Kynesian theory of consumption

- 1. Statement of theory or hypothesis.

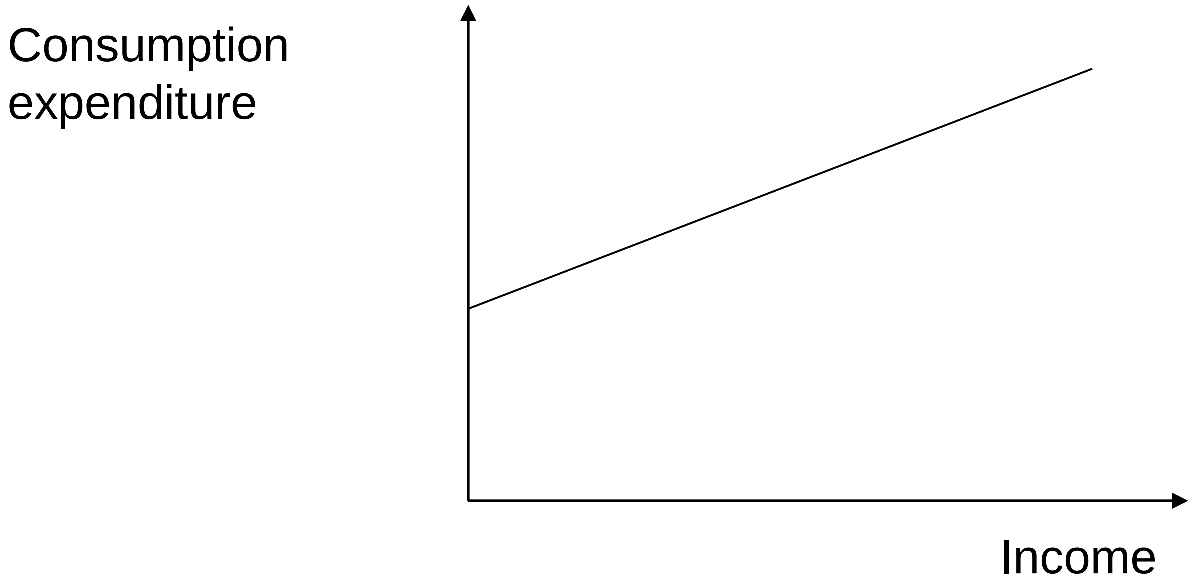
Keynes stated: The fundamental psychological law is that men/women are disposed, as a rule and on average, to increase their consumption as their income increases, but not as much as the increase in their income.

- In short, Keynes postulated that <u>the marginal propensity to consume (MPC),</u> the rate of change of consumption for a unit change in income, is greater than zero but less than 1

- 2.Specification of the mathematical model of the theory

- A mathematical economist might suggest the following form of the Keynesian consumption function:

$$Y = \beta_0 + \beta_1 X \qquad 0 < \beta_1 < 1$$



3. Specification of the statistical, or econometric model.

- To allow for the inexact relationships between economic variables, the econometrician would modify the deterministic consumption function as follows:

$$Y = \beta_0 + \beta_1 X + u$$

U, known as disturbance, or error term

- This is called an econometric model.

4. Obtaining the data.

year	Y	X
1982	3081.5	4620.3
1983	3240.6	4803.7
1984	3407.6	5140.1
1985	3566.5	5323.5
1986	3708.7	5487.7
1987	3822.3	5649.5
1988	3972.7	5865.2
1989	4064.6	6062
1990	4132.2	6136.3
1991	4105.8	6079.4
1992	4219.8	6244.4
1993	4343.6	6389.6
1994	4486	6610.7
1995	4595.3	6742.1
1996	4714.1	6928.4

Source: Data on Y (Personal Consumption Expenditure) and X (Gross Domestic Product),1982-1996) all in 1992 billions of dollars

5. Estimation of the parameters of the econometric model.

- reg y x

```
      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) = 8144.59
       Model |  3351406.23     1  3351406.23           Prob > F      =  0.0000
    Residual |  5349.35306    13  411.488697           R-squared     =  0.9984
-------------+------------------------------           Adj R-squared =  0.9983
       Total |  3356755.58    14  239768.256           Root MSE      =  20.285

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    .706408   .0078275    90.25   0.000     .6894978    .7233182
       _cons |  -184.0779   46.26183    -3.98   0.002    -284.0205   -84.13525
------------------------------------------------------------------------------
```

6. Hypothesis testing.

- As noted earlier, Keynes expected the MPC to be positive but less than 1. In our example we found it is about 0.70.
- Then, is 0.70 statistically less than 1? If it is, it may support Keynes's theory.

Such confirmation or refutation of econometric theories on the basis of sample evidence is based on a branch of statistical theory know as statistical inference (hypothesis testing)

- 7.Forecasting or prediction.

- To illustrate, suppose we want to predict the mean consumption expenditure for 1997. The GDP value for 1997 was 7269.8 billion dollars. Putting this value on the right-hand of the model, we obtain 4951.3 billion dollars.
- But the actual value of the consumption expenditure reported in 1997 was 4913.5 billion dollars. The estimated model thus overpredicted.
- The forecast error is about 37.82 billion dollars.

Types of Data Sets

Cross-Sectional, Time Series, and Panel Data

- Cross-sectional data consist of multiple entities observed at a single time period.
- Time series data consist of a single entity observed at multiple time periods.
- Panel data (also known as longitudinal data) consist of multiple entities, where each entity is observed at two or more time periods.

TABLE 1.1 Selected Observations on Test Scores and Other Variables for California School Districts in 1998

Observation (district) number	District average test score (fifth grade)	Student-teacher ratio	Expenditure per pupil ($)	% of students learning English
1	690.8	17.89	$6,385	0.0%
2	661.2	21.52	5,099	4.6
3	643.6	18.70	5,502	30.0
4	647.7	17.36	7,102	0.0
5	640.8	18.67	5,236	13.9
⋮	⋮	⋮	⋮	⋮
418	645.0	21.89	4,403	24.3
419	672.2	20.20	4,776	3.0
420	655.8	19.04	5,993	5.0

Note: The California test score data set is described in Appendix 4.1.

TABLE 1.2 Selected Observations on the Rates of Consumer Price Index (CPI) Inflation and Unemployment in the United States: Quarterly Data, 1959–2000.

Obervation number	Date (Year:quarter)	CPI inflation rate (% per year at an annual rate)	Unemployment rate (%)
1	1959:II	0.7%	5.1%
2	1959:III	2.1	5.3
3	1959:IV	2.4	5.6
4	1960:I	0.4	5.1
5	1960:II	2.4	5.2
⋮	⋮	⋮	⋮
165	2000:II	3.0	4.0
166	2000:III	3.5	4.0
167	2000:IV	2.8	4.0

Note: The U.S. inflation and unemployment data set is described in Appendix 12.1.

TABLE 1.3 Selected Observations on Cigarette Sales, Prices, and Taxes, by State and Year for U.S. States, 1985–1995

Observation number	State	Year	Cigarette sales (packs per capita)	Average price per pack (including taxes)	Total taxes (cigarette excise tax + sales tax)
1	Alabama	1985	116.5	$1.022	$0.333
2	Arkansas	1985	128.5	1.015	0.370
3	Arizona	1985	104.5	1.086	0.362
⋮	⋮	⋮	⋮	⋮	⋮
47	West Virginia	1985	112.8	1.089	0.382
48	Wyoming	1985	129.4	0.935	0.240
49	Alabama	1986	117.2	1.080	0.334
⋮	⋮	⋮	⋮	⋮	⋮
96	Wyoming	1986	127.8	1.007	0.240
97	Alabama	1987	115.8	1.135	0.335
⋮	⋮	⋮	⋮	⋮	⋮
528	Wyoming	1995	112.2	1.585	0.360

Note: The cigarette consumption data set is described in Appendix 10.1.

Empirical problem: Class size and educational output

- Policy question: What is the effect on test scores (or some other outcome measure) of reducing class size by one student per class? By 8 students/class?
- We must use data to find out (is there any way to answer this *without* data?)

The California Test Score Data Set

All K-6 and K-8 California school districts (*n* = 420)

Variables:

- 5^{th} grade test scores (Stanford-9 achievement test, combined math and reading), district average
- Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers

Initial look at the data:

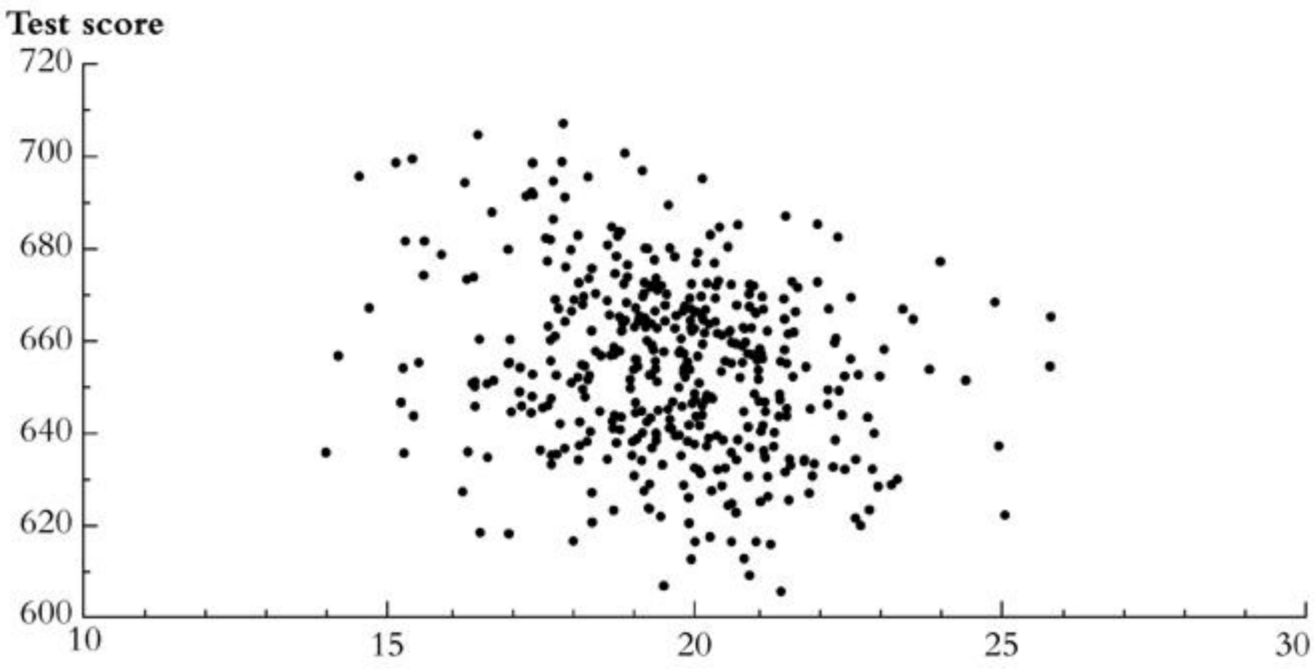
TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

- This table doesn't tell us anything about the relationship **between test scores and the *STR***.

Question: Do districts with smaller classes have higher test scores?

Scatterplot of test score v. student-teacher ratio



What does this figure show?

We need to get some numerical evidence on whether districts with low STRs have higher test scores – but how?

1. Compare average test scores in districts with low STRs to those with high STRs ("***estimation***")

2. Test the "null" hypothesis that the mean test scores in the two types of districts are the same, against the "alternative" hypothesis that they differ ("***hypothesis testing***")

3. Estimate an interval for the difference in the mean test scores, high v. low STR districts ("***confidence interval***")

Initial data analysis: Compare districts with "small" (STR < 20) and "large" (STR ≥ 20) class sizes:

Class Size	Average score ($\bar{Y}$)	Standard deviation (s_{Y})	n
Small	657.4	19.4	238
Large	650.0	17.9	182

1. ***Estimation*** of Δ = difference between group means
2. ***Test the hypothesis*** that Δ = 0
3. Construct a ***confidence interval*** for Δ

1. Estimation

$$\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = \frac{1}{n_{\text{small}}}\sum_{i=1}^{n_{\text{small}}} Y_i - \frac{1}{n_{\text{large}}}\sum_{i=1}^{n_{\text{large}}} Y_i$$

$$= 657.4 - 650.0$$

$$= 7.4$$

Is this a large difference in a real-world sense?

- Standard deviation across districts = 19.1
- Difference between 60$^{\text{th}}$ and 75$^{\text{th}}$ percentiles of test score distribution is 667.6 – 659.4 = 8.2
- This is a big enough difference to be important for school reform discussions, for parents, or for a school committee?

2. Hypothesis testing

Difference-in-means test: compute the *t*-statistic,

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

where $SE(\bar{Y}_s - \bar{Y}_l)$ is the "standard error" of $\bar{Y}_s - \bar{Y}_l$, the subscripts *s* and *l* refer to "small" and "large" STR districts, and

$$s_s^2 = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(Y_i - \bar{Y}_s)^2 \text{ (etc.)}$$

Compute the difference-of-means *t*-statistic:

Size	$\bar{Y}$	sBYB	n
small	657.4	19.4	238
large	650.0	17.9	182

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.4 - 650.0}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.4}{1.83} = 4.05$$

|t| > 1.96, so reject (at the 5% significance level) the null hypothesis that the two means are the same.

3. Confidence interval

A 95% confidence interval for the difference between the means is,

$$(\bar{Y}_s - \bar{Y}_l) \pm 1.96 \times SE(\bar{Y}_s - \bar{Y}_l)$$

$$= 7.4 \pm 1.96 \times 1.83 = (3.8,\ 11.0)$$

Two equivalent statements:

1. The 95% confidence interval for Δ doesn't include 0;
2. The hypothesis that $\Delta = 0$ is rejected at the 5% level.

What comes next…

- The mechanics of estimation, hypothesis testing, and confidence intervals should be familiar
- These concepts extend directly to regression and its variants
- Before turning to regression, however, we will review some of the underlying theory of estimation, hypothesis testing, and confidence intervals:
  - Why do these procedures work, and why use these rather than others?
  - So we will review the intellectual foundations of statistics and econometrics

Review of Statistical Theory

1. **The probability framework for statistical inference**
2. Estimation
3. Testing
4. Confidence Intervals

The probability framework for statistical inference

(a) Population, random variable, and distribution
(b) Moments of a distribution (mean, variance, standard deviation, covariance, correlation)
(c) Conditional distributions and conditional means
(d) Distribution of a sample of data drawn randomly from a population: $Y_1,\ldots, Y_n$

(a) Population, random variable, and distribution

Population

- The group or collection of all possible entities of interest (school districts)
- We will think of populations as infinitely large (N is an approximation to "very big")

Random variable Y

- Numerical summary of a random outcome (district average test score, district STR)

Population distribution of Y

- The probabilities of different values of Y that occur in the population, for ex. $\Pr[Y = 650]$ (when Y is discrete)
- or: The probabilities of sets of these values, for ex. $\Pr[640 < Y < 660]$ (when Y is continuous).

(b) Moments of a population distribution: mean, variance, standard deviation, covariance, correlation

mean = expected value (expectation) of Y

= $E(Y)$

= m_Y

= long-run average value of Y over repeated realizations of Y

variance = $E(Y - m_Y)^2$

= s_Y^2

= measure of the squared spread of the distribution

standard deviation = $\sqrt{\text{variance}} = s_Y$

Moments, ctd.

$$\textbf{\textit{skewness}} = \frac{E\left[\left(Y-\mu_Y\right)^3\right]}{\sigma_Y^3}$$

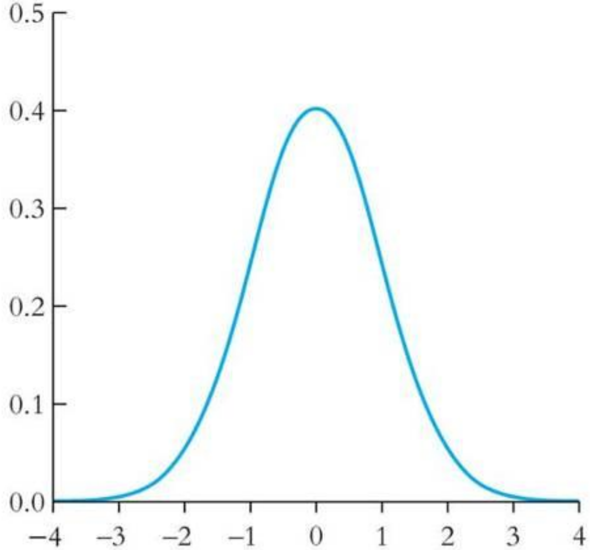
= measure of asymmetry of a distribution

- *skewness* = 0: distribution is symmetric
- *skewness* > (<) 0: distribution has long right (left) tail

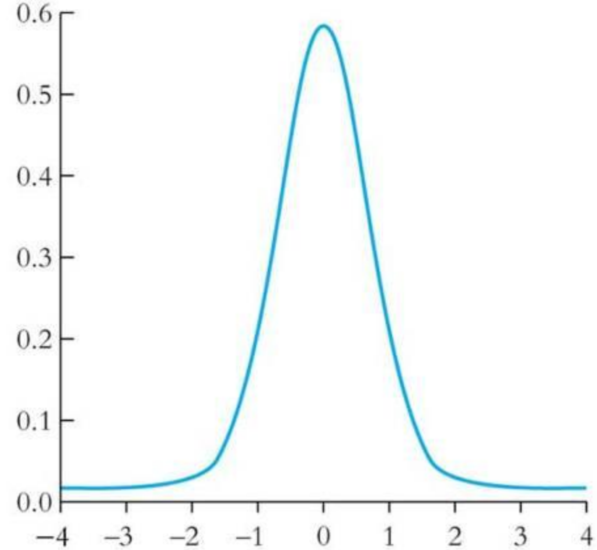
$$\textbf{\textit{kurtosis}} = \frac{E\left[\left(Y-\mu_Y\right)^4\right]}{\sigma_Y^4}$$

= measure of mass in tails
= measure of probability of large values

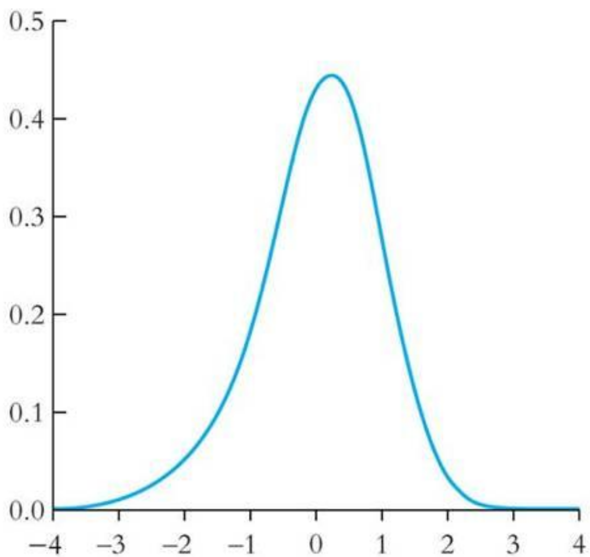
- *kurtosis* = 3: normal distribution
- *skewness* > 3: heavy tails ("***leptokurtotic***")



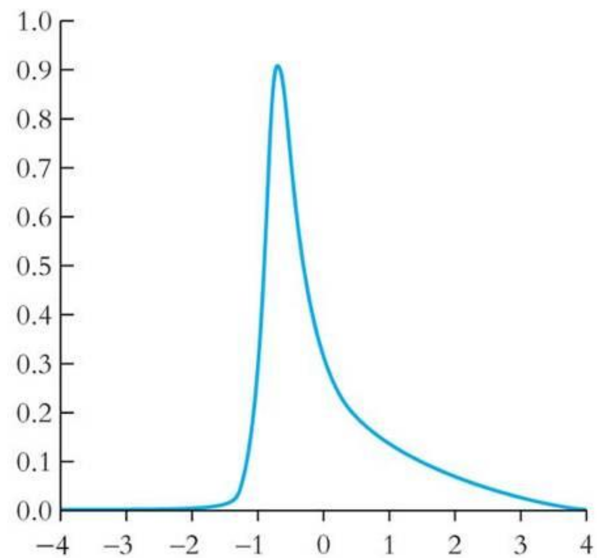
(a) Skewness = 0, kurtosis = 3



(b) Skewness = 0, kurtosis = 20



(c) Skewness = −0.1, kurtosis = 5



(d) Skewness = 0.6, kurtosis = 5

Random variables: joint distributions and covariance

- Random variables X and Z have a ***joint distribution***
- The ***covariance*** between X and Z is

$$\text{cov}(X,Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$$

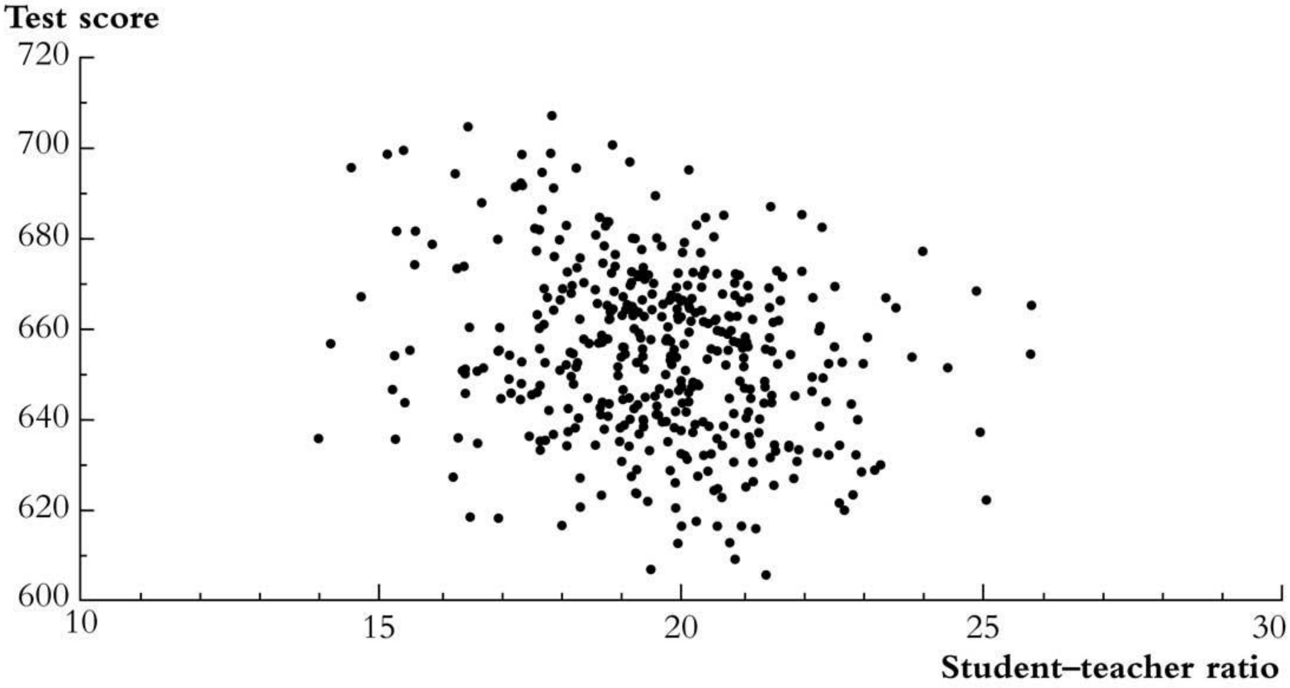
- The covariance is a measure of the linear association between X and Z; its units are units of X units of Z
- $\text{cov}(X,Z) > 0$ means a positive relation between X and Z
- If X and Z are independently distributed, then $\text{cov}(X,Z) = 0$ (but not vice versa!!)
- The covariance of a r.v. with itself is its variance:

$$\text{cov}(X,X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = \sigma_X^2$$

The covariance between Test Score and STR is negative:

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.



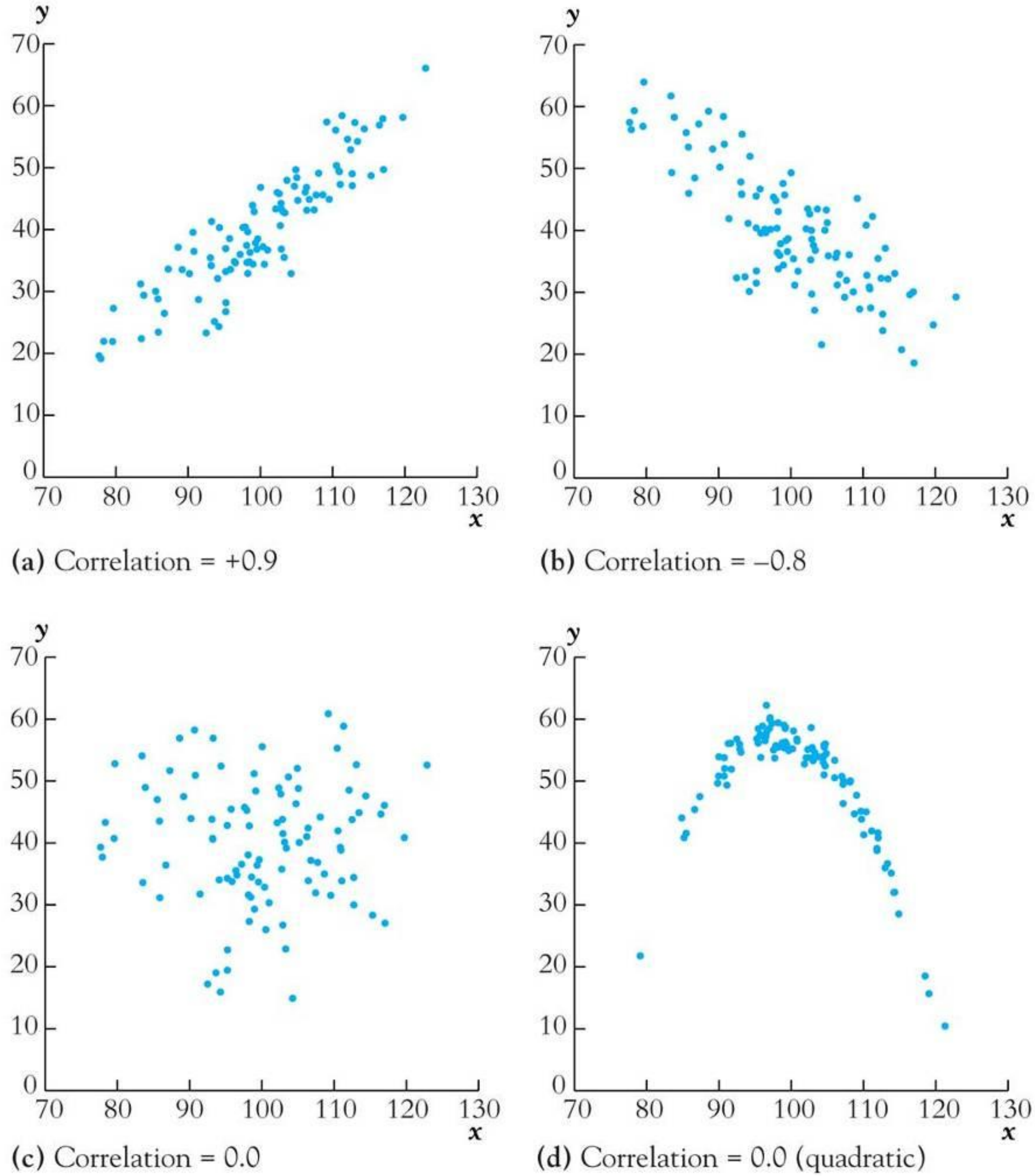
so is the ***correlation***…

The *correlation coefficient* is defined in terms of the covariance:

$$\operatorname{corr}(X,Z) = \frac{\operatorname{cov}(X,Z)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Z)}} = \frac{s_{XZ}}{s_X s_Z} = r_{XZ}$$

- $-1 < \operatorname{corr}(X,Z) < 1$
- $\operatorname{corr}(X,Z) = 1$ mean perfect positive linear association
- $\operatorname{corr}(X,Z) = -1$ means perfect negative linear association
- $\operatorname{corr}(X,Z) = 0$ means no linear association

The correlation coefficient measures <u>linear association</u>



(a) Correlation = +0.9

(b) Correlation = –0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

(c) Conditional distributions and conditional means

Conditional distributions

- The distribution of Y, given value(s) of some other random variable, X
- Ex: the distribution of test scores, given that STR < 20

Conditional expectations and conditional moments

- *conditional mean* = mean of conditional distribution
  $= E(Y|X = x)$ (***important concept and notation***)
- *conditional variance* = variance of conditional distribution
- *Example*: $E(Test\ scores|STR < 20)$ = the mean of test scores among districts with small class sizes

The difference in means is the difference between the means of two conditional distributions:

Conditional mean, ctd.

$\Delta = E(Test\ scores|STR < 20) - E(Test\ scores|STR \geq 20)$

Other examples of conditional means:

- Wages of all female workers (Y = wages, X = gender)
- Mortality rate of those given an experimental treatment (Y = live/die; X = treated/not treated)
- If $E(X|Z)$ = const, then corr$(X,Z) = 0$ (not necessarily vice versa however)

The conditional mean is a (possibly new) term for the familiar idea of the group mean

(d) Distribution of a sample of data drawn randomly from a population: $Y_1, \ldots, Y_n$

We will assume simple random sampling

- Choose and individual (district, entity) at random from the population

Randomness and data

- Prior to sample selection, the value of Y is random because the individual selected is random
- Once the individual is selected and the value of Y is observed, then Y is just a number – not random
- The data set is $(Y_1, Y_2, \ldots, Y_n)$, where Y_i = value of Y for the i^{th} individual (district, entity) sampled

Distribution of $Y_1,\ldots, Y_n$ under simple random sampling

- Because individuals #1 and #2 are selected at random, the value of Y_1 has no information content for Y_2. Thus:
  - Y_1 and Y_2 are ***independently distributed***
  - Y_1 and Y_2 come from the same distribution, that is, YB$_1$, Y_2 are ***identically distributed***
  - That is, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (***i.i.d.***).
  - More generally, under simple random sampling, $\{Y_i\}$, $i = 1,\ldots, n$, are i.i.d.

This framework allows rigorous statistical inferences about moments of population distributions using a sample of data from that population …

1. The probability framework for statistical inference
2. **Estimation**
3. Testing
4. Confidence Intervals

<u>Estimation</u>

$\bar{Y}$ is the natural estimator of the mean. But:

(a) What are the properties of $\bar{Y}$?

(b) Why should we use $\bar{Y}$ rather than some other estimator?

- Y_1 (the first observation)
- maybe unequal weights – not simple average
- median($Y_1, \ldots, Y_n$)

The starting point is the sampling distribution of $\bar{Y}$ …

(a) The sampling distribution of $\bar{Y}$

$\bar{Y}$ is a random variable, and its properties are determined by the ***sampling distribution*** of $\bar{Y}$

- The individuals in the sample are drawn at random.
- Thus the values of $(Y_1,\ldots, Y_n)$ are random
- Thus functions of $(Y_1,\ldots, Y_n)$, such as $\bar{Y}$, are random: had a different sample been drawn, they would have taken on a different value
- The distribution of $\bar{Y}$ over different possible samples of size n is called the ***sampling distribution*** of $\bar{Y}$.
- The mean and variance of $\bar{Y}$ are the mean and variance of its sampling distribution, $E(\bar{Y})$ and $\text{var}(\bar{Y})$.
- The concept of the sampling distribution underpins all of econometrics.

The sampling distribution of $\bar{Y}$ *ctd.*

Example: Suppose Y takes on 0 or 1 (a ***Bernoulli*** random variable) with the probability distribution,

$$\Pr[Y = 0] = 0.22,\ \Pr(Y = 1) = 0.78$$

Then

$$E(Y) = p \times 1 + (1 - p) \times 0 = p = 0.78$$

$$\sigma_Y^2 = E[Y - E(Y)]^2 = p(1 - p) \quad \textit{[remember this?]}$$

$$= 0.78(1\text{–}0.78) = 0.1716$$

The sampling distribution of $\bar{Y}$ depends on n.

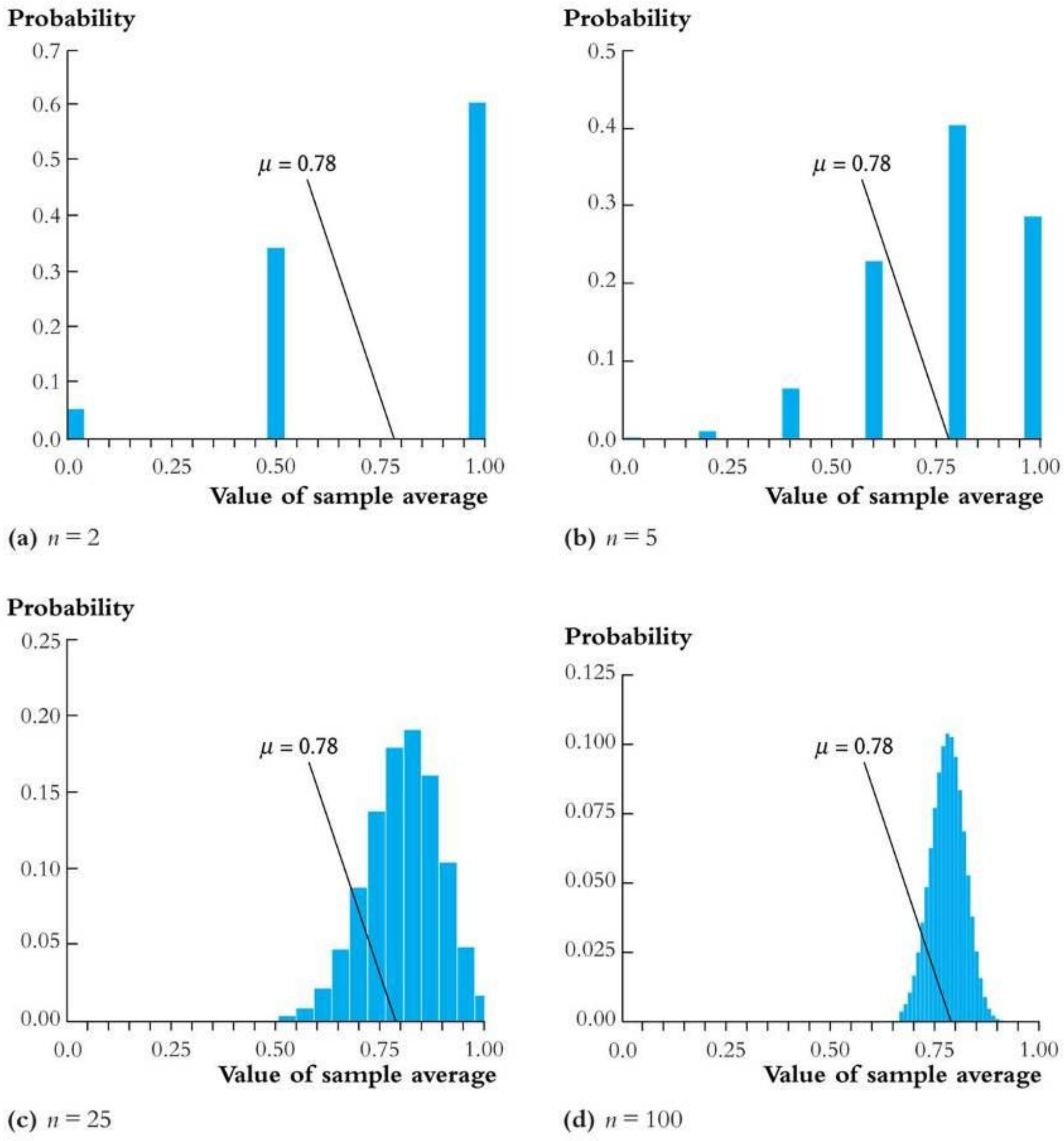
Consider $n = 2$. The sampling distribution of $\bar{Y}$ is,

$$\Pr(\bar{Y} = 0) = 0.22^2 = 0.0484$$

$$\Pr(\bar{Y} = \tfrac{1}{2}) = 2 \times 0.22 \times 0.78 = 0.3432$$

$$\Pr(\bar{Y} = 1) = 0.78^2 = 0.6084$$

The sampling distribution of $\bar{Y}$ when Y is Bernoulli (p = .78):



Things we want to know about the sampling distribution:

- What is the mean of $\bar{Y}$?
  - If $E(\bar{Y})$ = true $\mu = 0.78$, then $\bar{Y}$ is an ***unbiased*** estimator of μ
- What is the variance of $\bar{Y}$?
  - How does var($\bar{Y}$) depend on n (famous 1/n formula)
- Does $\bar{Y}$ become close to μ when n is large?
  - Law of large numbers: $\bar{Y}$ is a ***consistent*** estimator of μ
- $\bar{Y} - \mu$ appears bell shaped for n large…is this generally true?
  - In fact, $\bar{Y} - \mu$ is approximately normally distributed for n large (Central Limit Theorem)

The mean and variance of the sampling distribution of $\bar{Y}$

General case – that is, for Y_i i.i.d. from any distribution, not just Bernoulli:

mean: $E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y = \mu_Y$

Variance:

$$\begin{aligned}
\mathrm{var}(\bar{Y}) &= E[\bar{Y} - E(\bar{Y})]^2 \\
&= E[\bar{Y} - \mu_Y]^2 \\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - \mu_Y\right]^2 \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu_Y)\right]^2
\end{aligned}$$

so

$$\begin{aligned}
\operatorname{var}(\bar{Y}) &= E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_Y)\right]^2 \\
&= E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_Y)\right]\times\left[\frac{1}{n}\sum_{j=1}^{n}(Y_j-\mu_Y)\right]\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(Y_i-\mu_Y)(Y_j-\mu_Y)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(Y_i,Y_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 \\
&= \frac{\sigma_Y^2}{n}
\end{aligned}$$

Mean and variance of sampling distribution of $\bar{Y}$, ctd.

$$\boxed{E(\bar{Y}) = \mu_Y}$$

$$\boxed{\text{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}}$$

Implications:

1. $\bar{Y}$ is an *unbiased* estimator of μ_Y (that is, $E(\bar{Y}) = \mu_Y$)
2. var($\bar{Y}$) is inversely proportional to n
   - the spread of the sampling distribution is proportional to $1/\sqrt{n}$
   - Thus the sampling uncertainty associated with $\bar{Y}$ is proportional to $1/\sqrt{n}$ (larger samples, less uncertainty, but square-root law)

The sampling distribution of $\bar{Y}$ when *n* is large

For small sample sizes, the distribution of $\bar{Y}$ is complicated, but if *n* is large, the sampling distribution is simple!

1. As *n* increases, the distribution of $\bar{Y}$ becomes more tightly centered around μ_Y (the *Law of Large Numbers*)
2. Moreover, the distribution of $\bar{Y} - \mu_Y$ becomes normal (the *Central Limit Theorem*)

The *Law of Large Numbers*:

An estimator is ***consistent*** if the probability that its falls within an interval of the true population value tends to one as the sample size increases.

If $(Y_1,\ldots,Y_n)$ are i.i.d. and $\sigma_Y^2 < \infty$, then $\bar{Y}$ is a consistent estimator of μ_Y, that is,

$$\Pr[|\bar{Y} - \mu_Y| < \varepsilon] \to 1 \text{ as } n \to \infty$$

which can be written, $\bar{Y} \xrightarrow{p} \mu_Y$

("$\bar{Y} \xrightarrow{p} \mu_Y$" means "$\bar{Y}$ converges in probability to μ_Y").

(*the math*: as $n \to \infty$, $\text{var}(\bar{Y}) = \dfrac{\sigma_Y^2}{n} \to 0$, which implies that

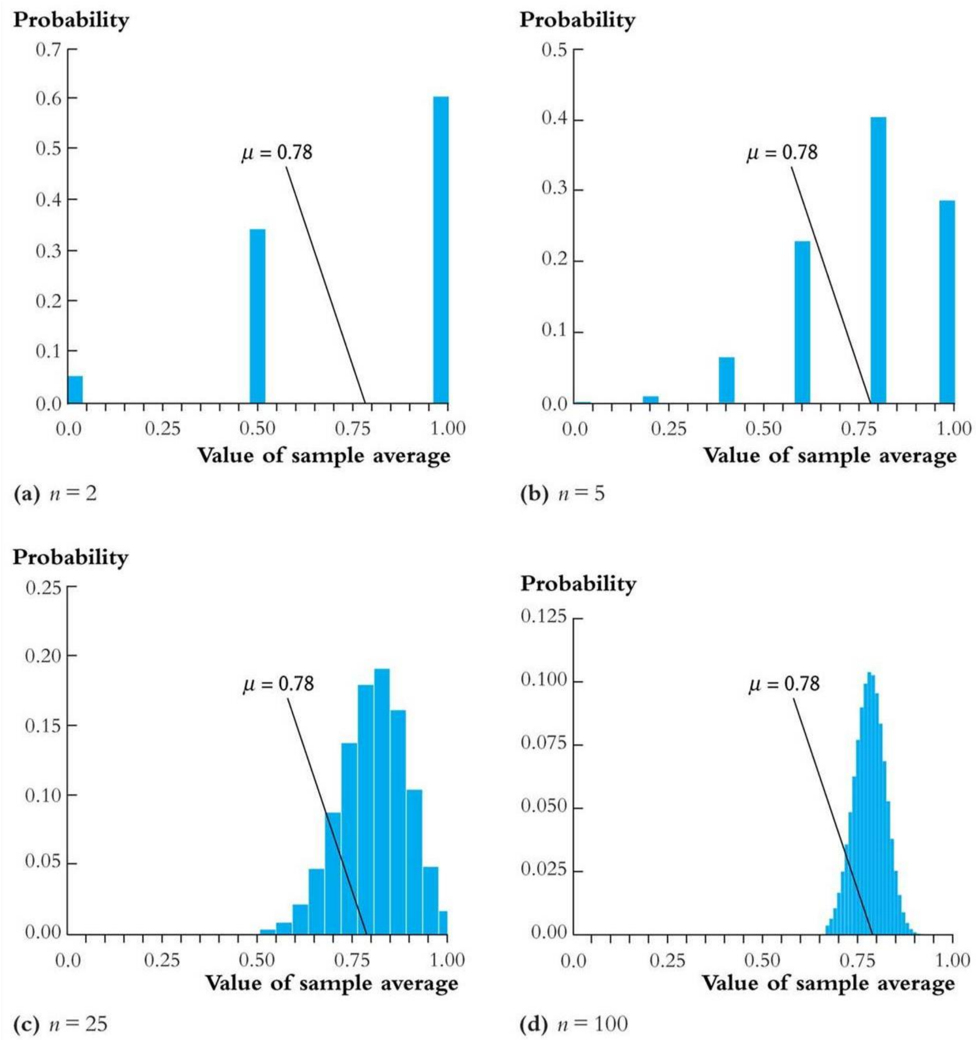
$\Pr[|\bar{Y} - \mu_Y| < \varepsilon] \to 1$.)

The *Central Limit Theorem* (CLT):

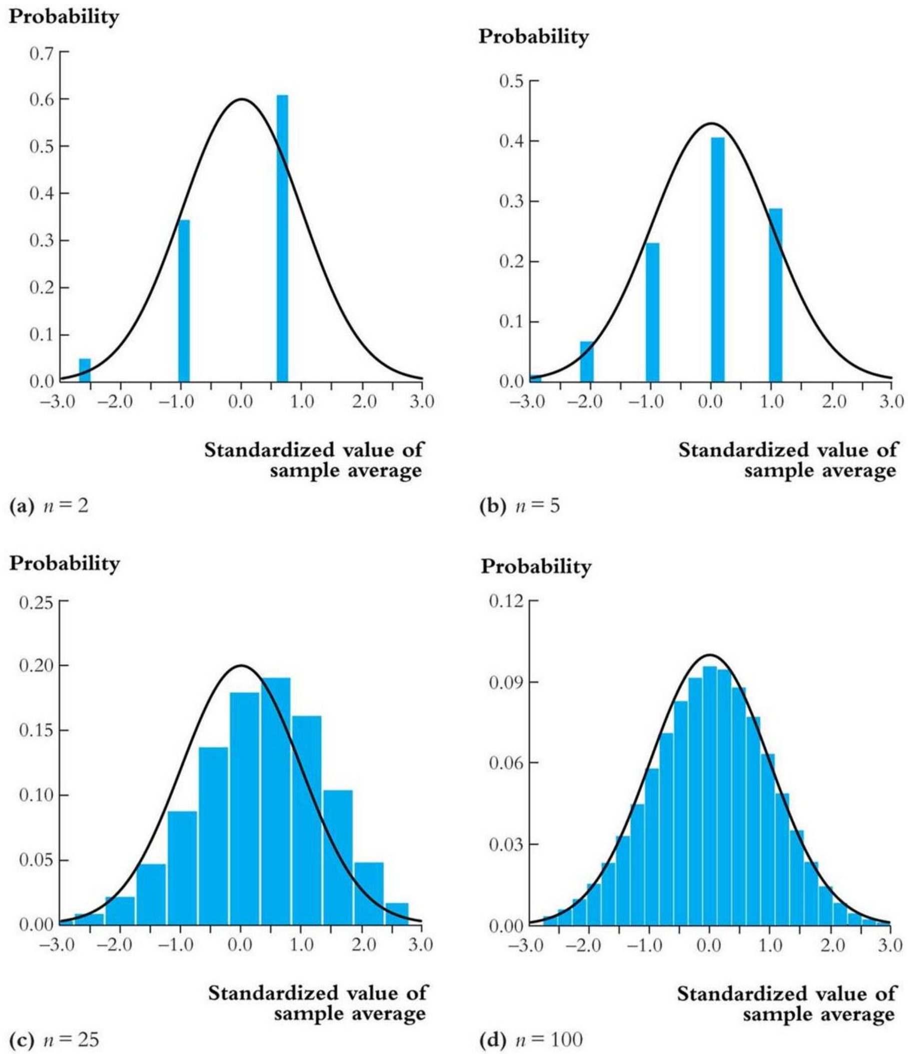
If $(Y_1,\ldots,Y_n)$ are i.i.d. and $0 < \sigma_Y^2 < \infty$, then when n is large the distribution of $\bar{Y}$ is well <u>approximated by a normal distribution</u>.

- $\bar{Y}$ is approximately distributed $N(\mu_Y, \frac{\sigma_Y^2}{n})$ ("normal distribution with mean μ_Y and variance σ_Y^2/n")
- $\sqrt{n}(\bar{Y} - \mu_Y)/\sigma_Y$ is approximately distributed $N(0,1)$ (standard normal)
- **That is, "standardized" $\bar{Y}$ = $\frac{\bar{Y}-E(\bar{Y})}{\sqrt{\operatorname{var}(\bar{Y})}} = \frac{\bar{Y}-\mu_Y}{\sigma_Y/\sqrt{n}}$ is approximately distributed as $N(0,1)$**
- **The larger is n, the better is the approximation.**

Sampling distribution of $\bar{Y}$ when Y is Bernoulli, p = 0.78:



Same example: sampling distribution of $\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\mathrm{var}(\bar{Y})}}$:



Summary: The Sampling Distribution of $\bar{Y}$

For $Y_1,\ldots,Y_n$ i.i.d. with $0 < \sigma_Y^2 < \infty$,

- The exact (finite sample) sampling distribution of $\bar{Y}$ has mean μ_Y ("$\bar{Y}$ is an unbiased estimator of μ_Y") and variance σ_Y^2/n
- Other than its mean and variance, the exact distribution of $\bar{Y}$ is complicated and depends on the distribution of Y (the population distribution)
- When n is large, the sampling distribution simplifies:
  - $\boxed{\bar{Y} \xrightarrow{p} \mu_Y}$ (Law of large numbers)
  - $\boxed{\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\mathrm{var}(\bar{Y})}} \text{ is approximately } N(0,1)}$ (CLT)

(b) Why Use $\bar{Y}$ To Estimate μ_Y?

- $\bar{Y}$ is unbiased: $E(\bar{Y}) = \mu_Y$
- $\bar{Y}$ is consistent: $\bar{Y} \xrightarrow{p} \mu_Y$
- $\bar{Y}$ is the "least squares" estimator of μ_Y; $\bar{Y}$ solves,

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

so, $\bar{Y}$ minimizes the sum of squared "residuals"

optional derivation

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

Set derivative to zero and denote optimal value of m by $\hat{m}$:

$$\sum_{i=1}^{n} Y = \sum_{i=1}^{n}\hat{m} = n\hat{m} \text{ or } \hat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Why Use $\bar{Y}$ To Estimate μ_Y?, ctd.

- $\bar{Y}$ has a smaller variance than all other **linear unbiased** estimators: consider the estimator, $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, where $\{a_i\}$ are such that $\hat{\mu}_Y$ is unbiased; then $\text{var}(\bar{Y}) \leq \text{var}(\hat{\mu}_Y)$

- $\bar{Y}$ isn't the only estimator of μ_Y – can you think of a time you might want to use the median instead?

1. The probability framework for statistical inference
2. Estimation
3. **Hypothesis Testing**
4. Confidence intervals

<u>Hypothesis Testing</u>

The ***hypothesis testing*** problem (for the mean): make a provisional decision, based on the evidence at hand, whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) > \mu_{Y,0}$ (1-sided, >)

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) < \mu_{Y,0}$ (1-sided, <)

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) \neq \mu_{Y,0}$ (2-sided)

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 $\bar{Y}$:

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

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

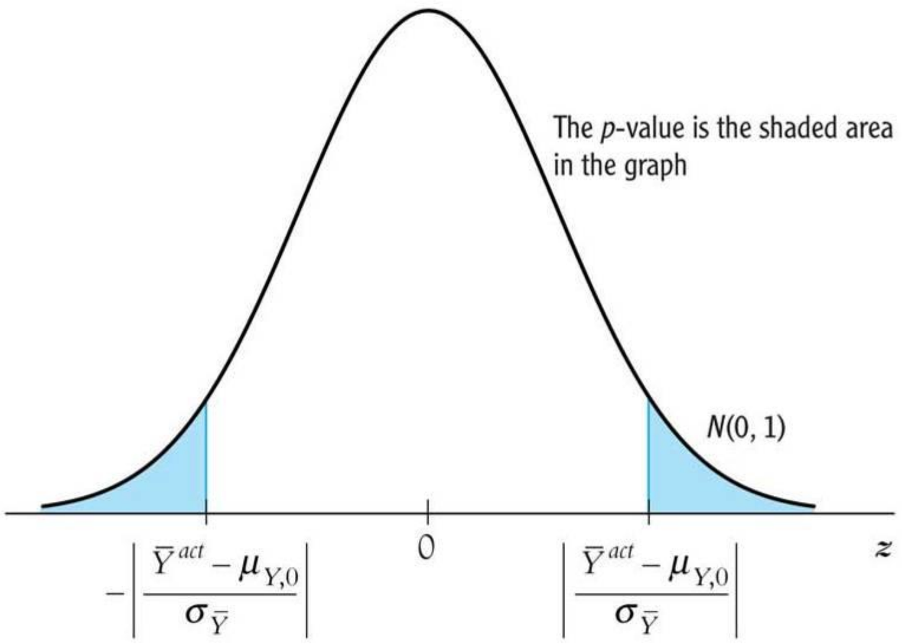
Calculating the p-value, ctd.

- 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):

$$
\begin{aligned}
p\text{-value} &= \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|], \\
&= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}|] \\
&= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}|] \\
&\cong \text{probability under left+right } N(0,1) \text{ tails}
\end{aligned}
$$

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

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

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:

If $(Y_1,\ldots,Y_n)$ are i.i.d. and $E(Y^4) < \infty$, 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

Computing the p-value with σ_Y^2 estimated:

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

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

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

so

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

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

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

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.

At this point, you might be wondering,...

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

<u>Digression: the Student *t* distribution</u>

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.

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….

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 are negligible. Here are some 5% critical values for 2-sided tests:

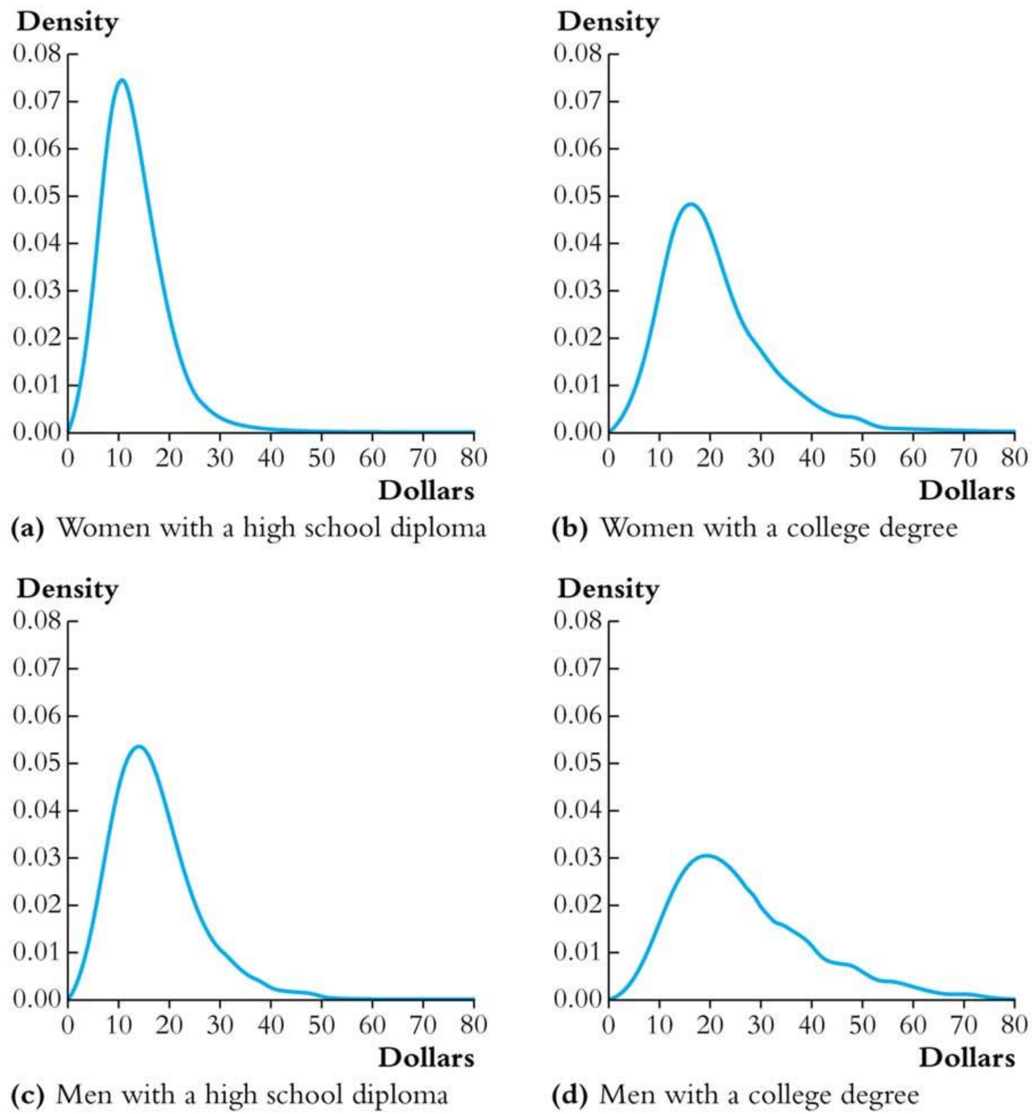
degrees of freedom (n – 1)	5% *t*-distribution critical value
10	2.23
20	2.09
30	2.04
60	2.00
∞	1.96

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).



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 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?

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, in this class we will focus on the large-n approximation given by the CLT

1. The probability framework for statistical inference
2. Estimation
3. Testing
4. **Confidence intervals**

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 $Y_1,\ldots,Y_n$ and thus any functions of them – including the confidence interval. The confidence interval it will differ from one sample to the next. The population parameter, μ_Y, is not random, we just don't know it.

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.

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

$$= \{\mu_Y\colon -1.96\frac{s_Y}{\sqrt{n}} \le \bar{Y} - \mu_Y \le 1.96\frac{s_Y}{\sqrt{n}}\}$$

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

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

Summary:

From the two assumptions of:

(1) simple random sampling of a population, that is, $\{Y_i, i = 1,\ldots,n\}$ 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 test statistic)

Are assumptions (1) & (2) plausible in practice? **Yes**