Introductory Applied Econometrics Analysis using Stata

November 14 – 18, 2016

Dushanbe, Tajikistan

Allen Park and Jarilkasin Ilyasov

What is econometrics?

Based on Chapter 1-2. Stock and Watson. “Introduction to Econometrics” 3rd Edition.

What is econometrics?

- Ask a half dozen econometricians what econometrics is, you might get half a dozen different answers:

- Econometrics is the science of testing economic theories;
- Econometrics is the set of tools for forecasting future values of economic variables (e.g. firm's sales; stock prices, etc.);
- Econometrics is the science and art of using historical data to make numerical/quantitative policy recommendations in government and business;
- …

- Econometrics is the science and art of using economic theory and statistical techniques to analyze economic data

- (Stock and Watson 2015 3rd edition of "Introduction to Econometrics")

Why econometrics?

Quantitative answers to quantitative questions:

Economics suggests important relationships, often with policy implications, but virtually never suggests quantitative magnitudes of causal effects.

- What is the *quantitative* effect of reducing class size on student achievement?
- How does another year of education change earnings?
- What is the price elasticity of cigarettes?
- What is the effect of applying a certain amount of fertilizer on the productivity of tomato, potato, etc.?

In this course you/we will learn:

- Sources and Types of data;
- Review of Probability and Statistics – why these are important?
  - Why randomness is of central importance in econometrics?
  - Three types of statistical methods: estimation, hypothesis testing and confidence intervals;
- Simplest econometric models – Ordinary Least Squares (OLS) with one and multiple variables;
- Learn to evaluate the regression analysis of others – this means you will be able to read/understand empirical economics papers (partially);
- Get some hands-on experience with simple regression analysis using Stata.

In more advanced course (e.g. WIUT in Tashkent) you will learn:

- More advanced methods for estimating causal effects using observational data;
- Some tools that can be used for other purposes; for example, forecasting using time series data;
- Focus on applications – theory is used only as needed to understand the whys of the methods;
- Learn to evaluate the regression analysis of others – this means you will be able to read/understand empirical economics papers;
- Get some hands-on experience with regression analysis in your problem sets.

Data sources:

- Ideally, we would like an **experiment**
  - What would be an experiment to estimate the effect of fertilizer application on potato and tomato yield?
  - **Randomized controlled experiment** → randomly assigned plots… AND control group (no fertilizer) vs treatment group (receives fertilizer)
- But almost always we only have **observational** (nonexperimental) data (e.g. surveys, administrative records, etc.)
  - returns to education
  - crop production and productivity
  - cigarette prices
- Most of the course in econometrics deals with difficulties arising from using observational to estimate causal effects
  - confounding effects (omitted factors)
  - “correlation does not imply causation”

Data Types:

1. Cross-Sectional Data
    - Data on different entities - e.g. farmers, consumers, cars, etc. - for a single time period
    - E.g "sysuse auto" from yesterday is a cross sectional data
    - Those data are for 1978 observations (***n*** = different cars) → **number of observations** usually denoted by ***n***
2. Time Series Data
    - Data on a single entity – e.g. a farm, a country, etc. – at multiple time periods
    - E.g. "sysuse gnp96.dta" from yesterday's homework is a time series data
    - 142 observations (***T*** = time periods) for two variables (i) date and (ii) gnp96
3. Panel Data (= longitudinal data)
    - Data on multiple entities in which each entity is observed at multiple time periods
    - Number of entities denoted by ***n*** and time periods denoted by ***T***
    - Number of observations = ***n*** x ***T***

Review of Probability and Statistics:

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?)

Review of Probability and Statistics:

The California Test Score Data Set

(use http://fmwww.bc.edu/ec-p/data/stockwatson/caschool)

Variables:

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

Review of Probability and Statistics:

Initial look at the data:

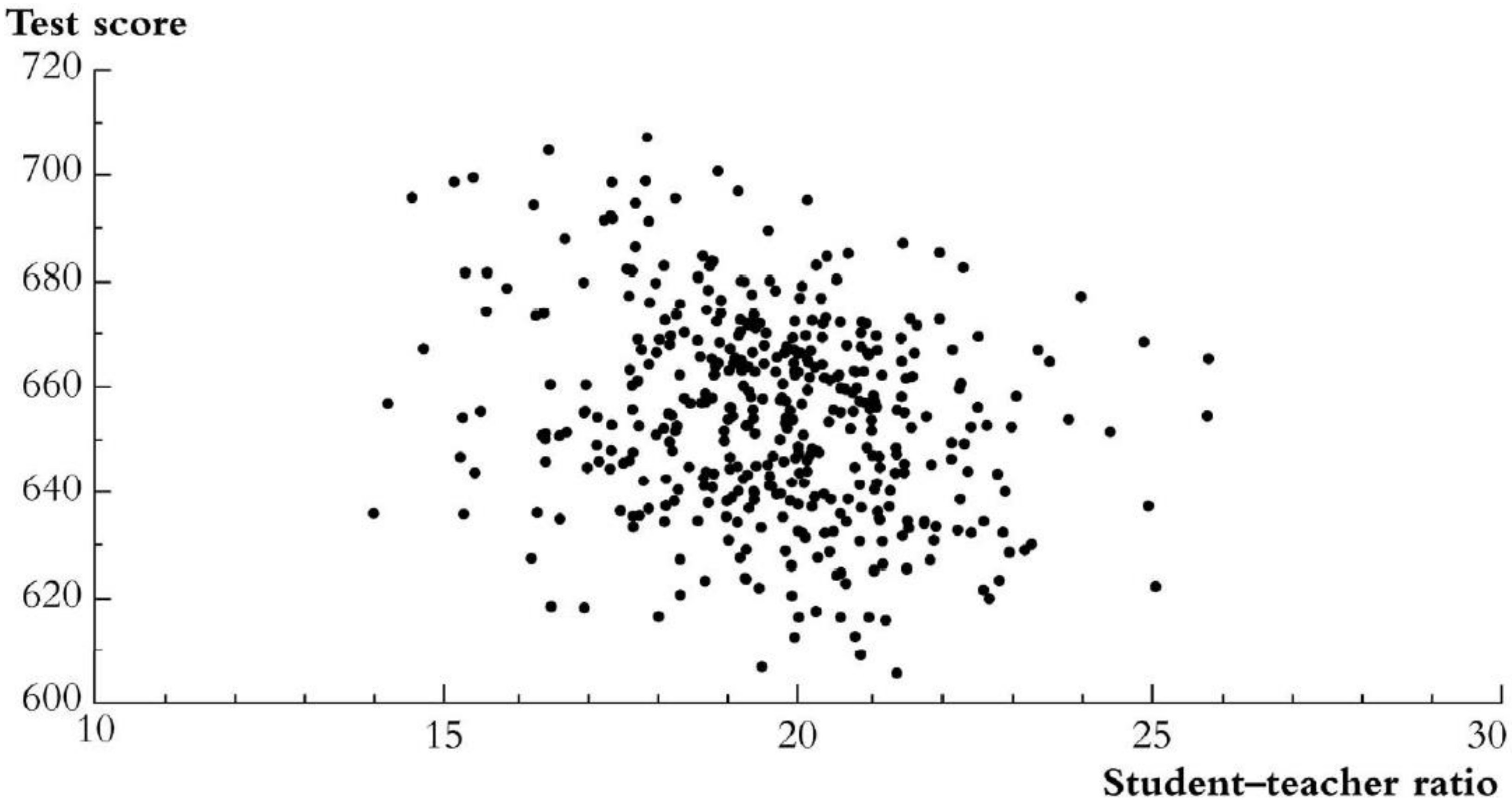
TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

			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

What does this table tell us about the relationship between test scores and the STR?

Review of Probability and Statistics:

Scatterplot of test score vs. STR



What does this figure show?

Review of Probability and Statistics:

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

1. **Estimation** - Compare average test scores in districts with low STRs to those with high STRs
2. **Hypothesis Testing** - 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
3. **Confidence Intervals** - Estimate an interval for the difference in the mean test scores, high v. low STR districts

These concepts extend directly to regression and its variants

Review of Probability:

Before that we review:

The probability framework for statistical inference

Element of **randomness** is everywhere → something that is not known until revealed

e.g. how many times we meet after this training?

how many times your PC will crash during the training course?

The theory of probability provides mathematical tools for quantifying and describing the randomness.

Review of Probability:

1. Review probability distributions for a single random variable
2. Mathematical expectation, mean and variance of a single random variable
3. Basic elements of probability theory for two random variables
4. Four special probability distributions that play important role in statistics and econometrics (Normal, Chi-squared, student's t, and the F Distributions)
5. Random sampling and sampling distribution

Review of Probability:

Fundamental concepts

- ***Outcomes* -** the *mutually exclusive* potential results of a *random process*
  - e.g. your PC might never crash, crash once, twice, and so on.
- ***Probability* -** the probability of an outcome is the proportion of the time that the outcome occurs in the long-run
- **The sample space** - the set of all possible outcomes
- ***Event* -** a subset of **the sample space** with a set of one or more outcomes
  - e.g. The event "the computer will crash no more than once" consists of two outcomes: crash of 0 or 1

Review of Probability:

- **Random variable Y –** Numerical summary of a random outcome
  - e.g. the number of times your computer crashes, the number of girls in a family of three children

- *Types of random variables*
  - **Discrete** random variables – e.g. 0, 1, 2,...
  - **Continuous** random variables

Review of Probability:

- **Probability distribution** of a discrete random variable
  - is the list of all possible values of random variable with their respective probabilities
  - Cumulative probability distribution is the probability that random variable is less or equal to a value

	Outcome (number of crashes)				
	0	1	2	3	4
Probability distribution	0.8	0.1	0.06	0.03	0.01
Cumulative distribution	0.8	0.9	0.96	0.99	1

Review of Probability:

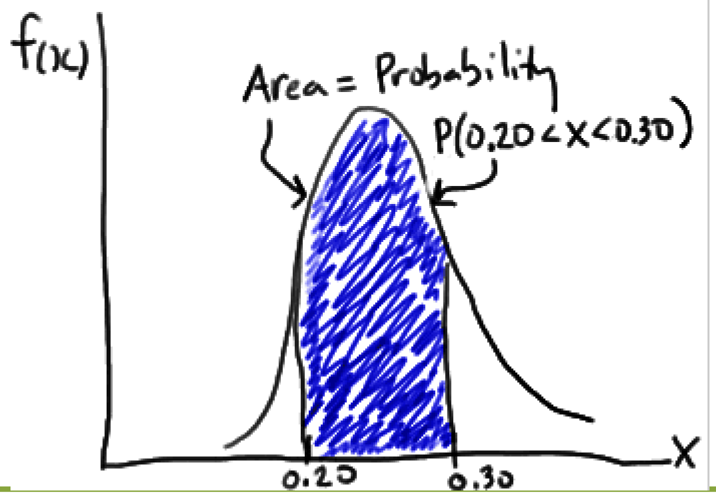
- **The Bernoulli distribution**
  - Special case of a discrete random variable → Random Variable *Y* is binary

e.g. If event G is the gender of the next person you meet, the Bernoulli distribution has the following notation

$$G = \begin{cases} 1\ (= girl) & with\ \ probability\ \ p \\ 0\ (= boy) & with\ probability\ 1 - p \end{cases}$$

Review of Probability:

- **Probability distribution** of a continuous random variable
  - is the list of all possible values of random variable with their respective probabilities
  - Cumulative probability distribution is continuum of possible values → list is not viable thus **probability density function,** also called **p.d.f.**, a **density function** or simply **density**



Review of Probability:

Expected Value, Mean and Variance

- **The expected value** of a random variable *Y*, denoted by ***E*(Y)**, is the long-run average value of the random variable.
- The expected value of *Y* is also called **mean** of *Y* and denoted by μ_Y
  - Suppose ***Y*** takes on ***k*** possible values, {y_i} for ***i=1,2,…,k*.** Then

$$E(Y) = \mu_Y = y_1 p_1 + y_2 p_2 + \ldots + y_k p_k = \sum_i^k y_i p_i$$

"the sum of $y_i p_i$ for *i* running from 1 to *k*."

Review of Probability:

Expected Value, Mean and Variance

- **The variance** and **deviation** measure the dispersion or the "spread" of a probability function and denoted by σ_Y^2
- "expected value of the square of the deviation of *Y* from its mean"

$$E\left[(Y-\mu_Y)^2\right]$$

– Suppose ***Y*** takes on ***k*** possible values, {$\boldsymbol{y_i}$} for ***i=1,2,…,k.*** Then

$$\sigma_Y^2 = \mathrm{var}(Y) = E\left[(Y-\mu_Y)^2\right] = \sum_i (y_i-\mu_Y)^2 p_i$$

Standard deviation??

Review of Probability:

Expected Value, Mean and Variance

- **The variance** and **deviation** measure the dispersion or the "spread" of a probability function and denoted by σ_Y^2
- "expected value of the square of the deviation of *Y* from its mean"

$$E\left[(Y-\mu_Y)^2\right]$$

– Suppose ***Y*** takes on ***k*** possible values, {$\boldsymbol{y_i}$} for ***i=1,2,…,k.*** Then

$$\sigma_Y^2 = \mathrm{var}(Y) = E\left[(Y-\mu_Y)^2\right] = \sum_i (y_i-\mu_Y)^2 p_i$$

Standard deviation?? σ_Y = the square root of the variance

Review of Probability:

Expected Value, Mean and Variance

- For Bernoulli random variable *G*,

$$E(G) = 1xp+0x(1-p) = p$$

$$Var(G) = (0-p)^2x(1-p) + (1-p)^2xp = px(1-p)$$

Review of Probability:

Other measures the shape of a distribution

- **Skewness** = measure of asymmetry of a distribution

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

- skewness = 0: distribution is symmetric
- skewness > (<) 0: distribution has long right (left) tail
- Skewness mathematically describes how much a distribution deviates from symmetry

Review of Probability:

Other measures the shape of a distribution

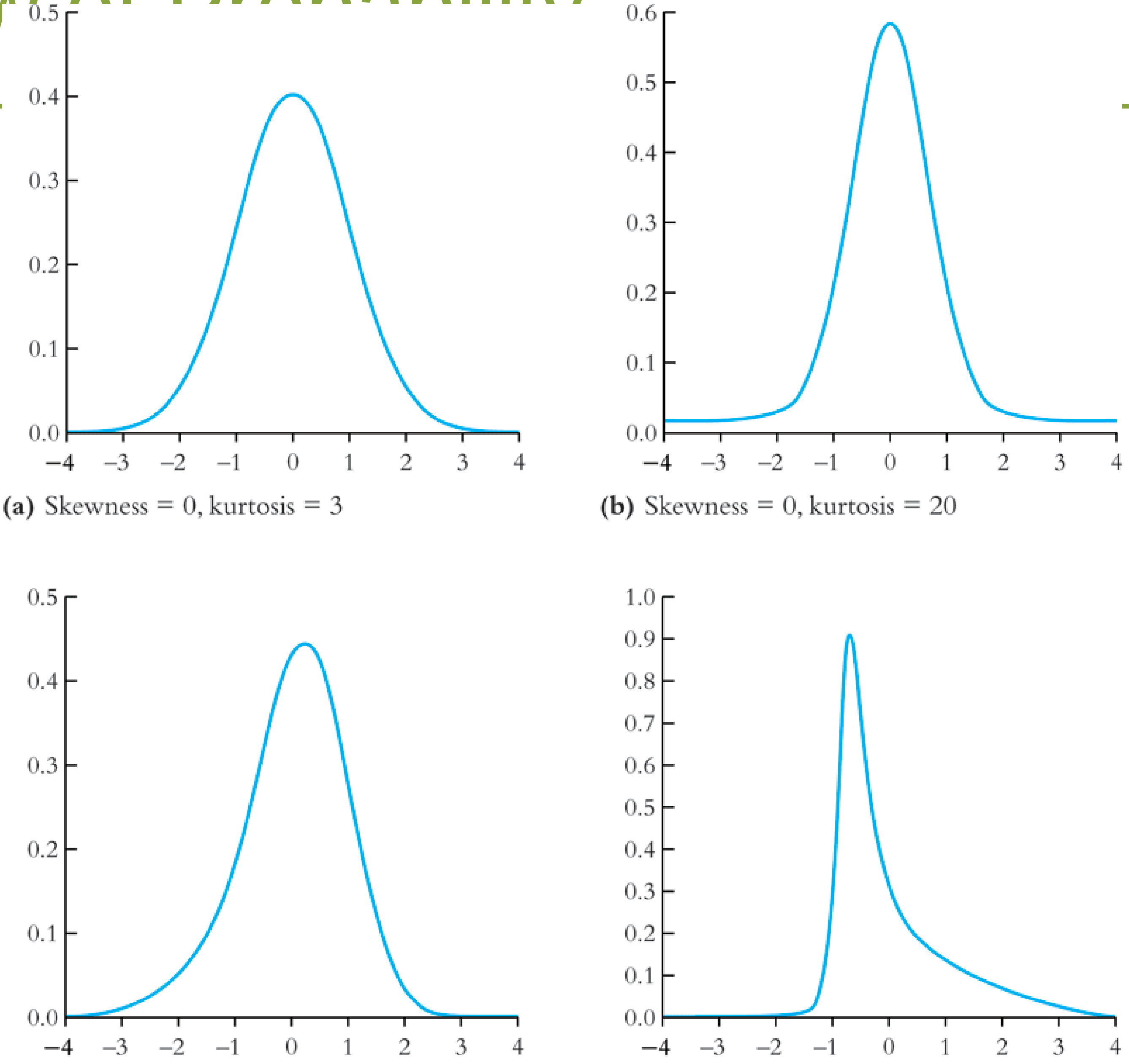
- ***kurtosis*** = $\dfrac{E\left[\left(Y-\mu_Y\right)^4\right]}{\sigma_Y^4}$

  = measure of mass in tails

  = measure of probability of large values

- *kurtosis* = 3: normal distribution
- *kurtosis* > 3: heavy tails ("***leptokurtotic***")
- The greater the kurtosis of a distribution, the more likely are outliers.

Review of Probability:



(a) Skewness = 0, kurtosis = 3

(b) Skewness = 0, kurtosis = 20

(c) Skewness = –0.1, kurtosis = 5

(d) Skewness = 0.6, kurtosis = 5

Review of Probability:

Two random variables

Most of the interesting questions in economics involve >2 variables

- Random variables X and Y
  - Together they have a ***joint distribution***
  - Each one has a ***marginal distribution***
  - Each one has a ***conditional distribution***

Review of Probability:

- ***Joint distribution*** of two discrete X and Y
  - Probability that X and Y simultaneously take on certain values, say x and y.
  - $Pr(X = x, Y = y)$ or $Pr(x, y)$ or $P(X = x, Y = y)$ or $P(x, y)$
  - NOTE lower case symbols x and y denote values and. . .
  - Upper case symbols X and Y denote random variables
  - Probabilities of all possible (x, y) combinations sum to 1

Review of Probability:

- After recording data for many commutes
  - *prob. of long, rainy commute = P(X=0,Y=0) = .15*
  - *prob. of long, clear commute = P(X=1,Y=0) = 0.07*
  - *prob. of short, rainy commute = P(X=0,Y=1) = ??*
  - *prob. of short, clear commute = P(X=1,Y=1) = ??*
  - *These four outcomes are mutually exclusive and exhaust all possibilities*

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

Review of Probability:

- **Marginal distribution** is P(X=x) or P(Y=y)
  - Sum of joint probabilities:
  - *prob. of long commute = P(X=0,Y=0) + P(X=1,Y=0) = .15 +.07 =.22*
  - *prob. of short commute = P(X=0,Y=1) + P(X=1,Y=1)= .15 +.63=.78*
  - *prob. of rainy commute = ??*
  - *prob. of clear commute = ??*

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

Review of Probability:

- ***Conditional distribution*** of X and Y
  - Probability that Y is some value conditional on (depending on or after) X taking on a specified value
- Examples: distribution of. . .
  - test scores, given that STR < 20
  - wages of all female workers (*Y* = wages, *X* = gender)
  - P(Y=y | X=x) or P(y | x)

$$P(Y=y \mid x=X) = \frac{P(X=x, Y=y)}{P(X=x)} \quad or \quad P(y \mid x) = \frac{P(x,y)}{P(x)}$$

Review of Probability:

- *prob. of long commute (Y=0) if you know it's raining (X=0)*

$$P(Y=0 \mid X=0) = \frac{P(X=0, Y=0)}{P(X=0)} = \frac{.15}{.30} = .50$$

- *prob. of short commute (Y=1) if you know it's raining (X=0)*

  = P(Y=1| X=0) = ??

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

Review of Probability:

- *prob. of short commute (Y=1) if you know it's raining (X=0)*

  = P(Y=1| X=0) = ??

- $$P(Y=1|X=0)=\frac{P(X=0,Y=1)}{P(X=0)}=??$$

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

Review of Probability:

Independence

- Two random variables X and Y independent if
  - Knowing value of one tells you nothing about the value of the other
  - Conditional distribution of X & Y = marginal distribution of Y (or X)

    Pr(Y=y | X=x) = Pr(Y=y)

    Pr(X=x | Y=y) = Pr(X=x)

Pr(X=x , Y=y) = Pr(X=x)*Pr(Y=y), i.e. the joint distribution of two independent random variables is the product of their marginal distributions

Two random variables: covariance

- Measure of the extent to which two random variables move together
- The ***covariance*** between X and Y is

$$\text{cov}(X,Y) = E[(X - \mu_X)(Z - \mu_y)] = \sigma_{Xy}$$

- $\text{cov}(X,Y) > 0$ means a positive relation between X and Y
- If X and Y are independently distributed, then $\text{cov}(X,Y) = 0$ (but not vice versa!!)

Review of Probability:

Two random variables: Correlation

Alternative measure of dependence between two random variables

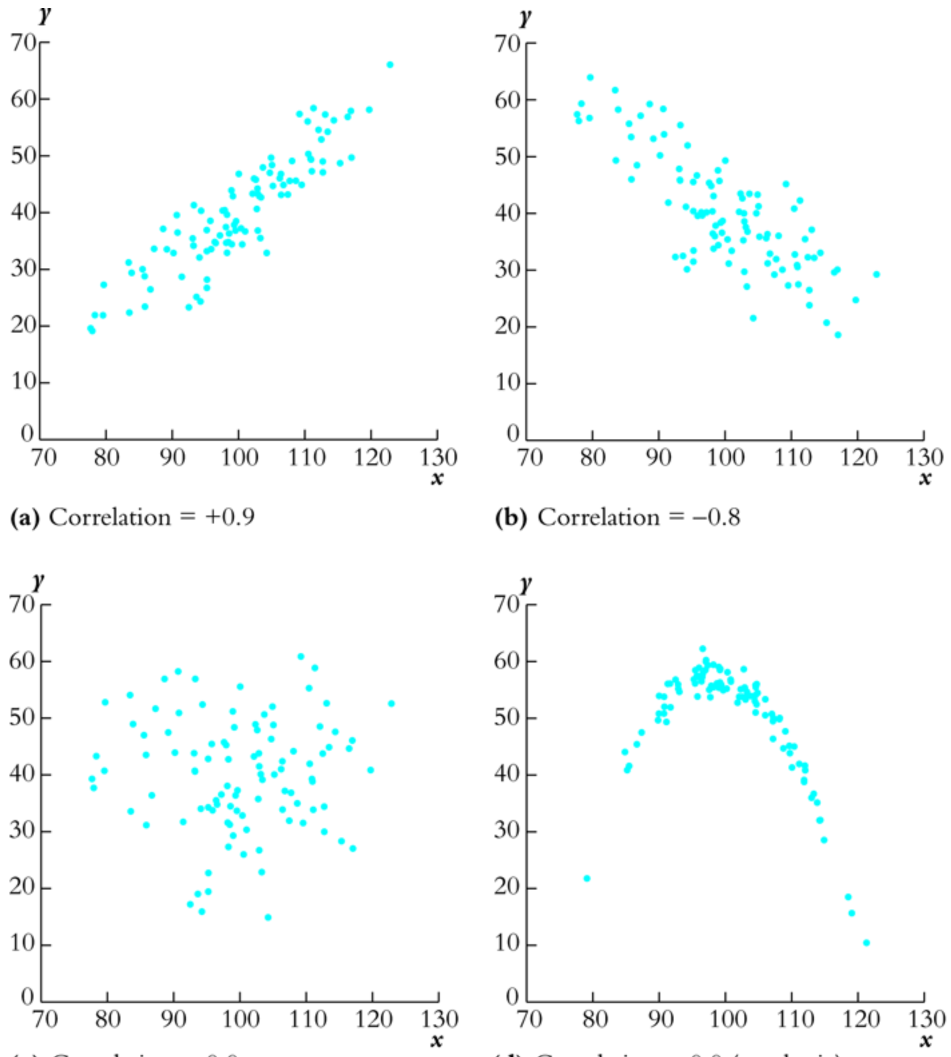
$$\text{corr}(X,Z) = \frac{\text{cov}(X,Z)}{\sqrt{\text{var}(X)\,\text{var}(Z)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = r_{XZ}$$

= covariance between X and Z divided by their standard deviation

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

Review of Probability:

The correlation coefficient measures linear association



(a) Correlation = +0.9

(b) Correlation = –0.8

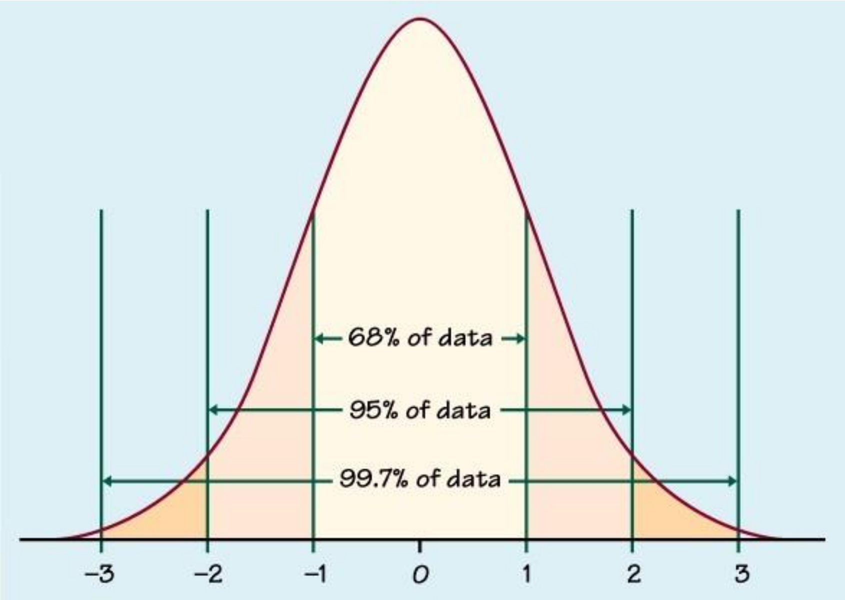
(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

Review of Probability:

Four Distributions: normal, chi-squared, Student t, F

- Normal Distribution
  - "bell-shaped" probability density
  - $X \sim N(\mu, \sigma^2)$
  - Standard normal
    - $Z \sim N(0, 1)$

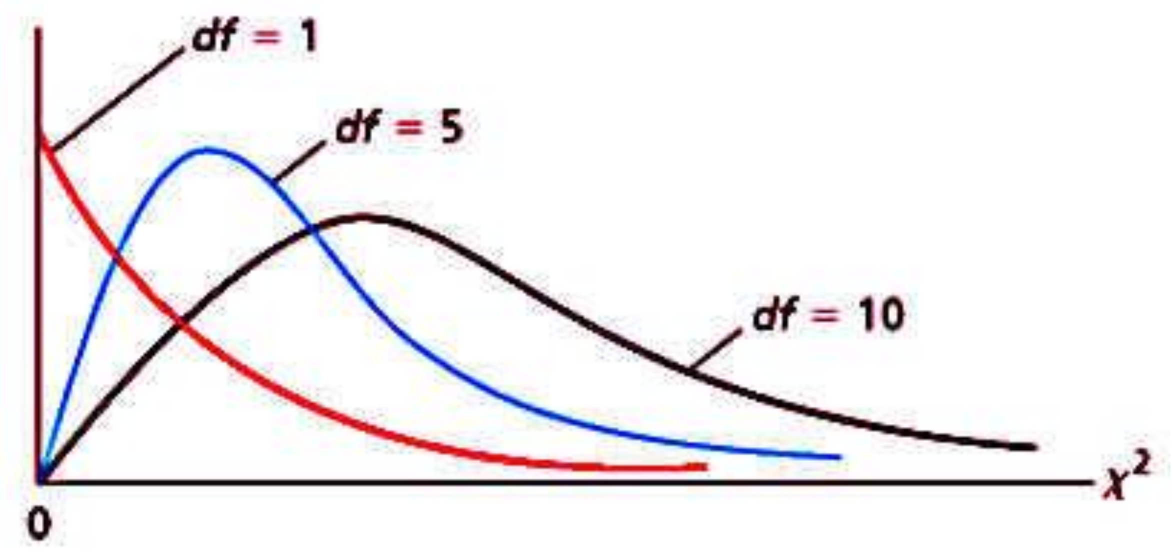


  - Standardizing a normal random variable (z score)
    - Used for finding probabilities about $X \sim N(\mu, \sigma^2)$

Review of Probability:

The Chi-squared Distribution

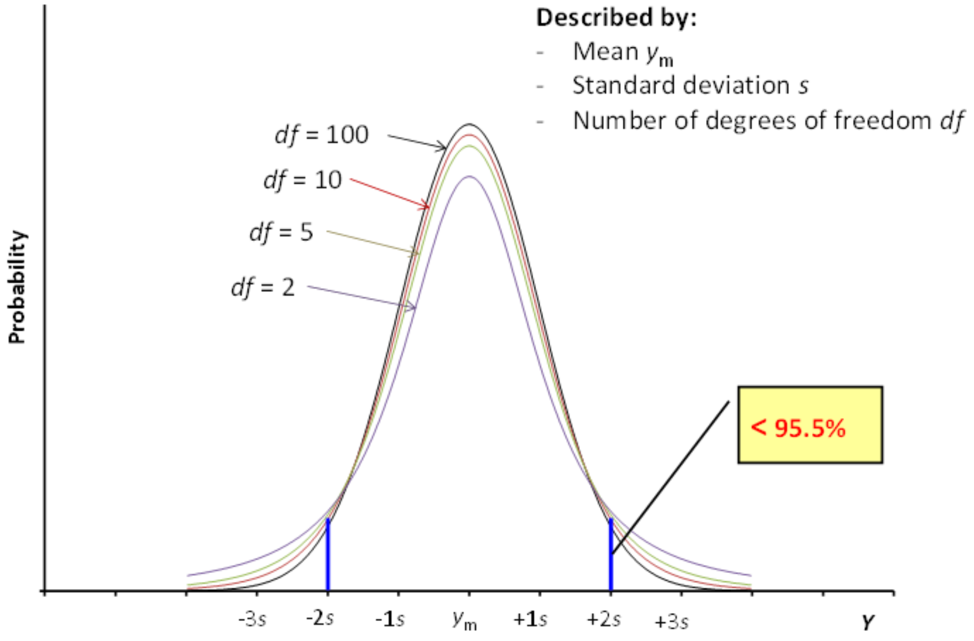
- Usually written as χ^2_m used in testing certain types of hypothesis in statistics and econometrics
- Shape depends on degrees of freedom ***m*** (



Review of Probability:

The Student t Distribution

- Always lower case "t"
- Shape of distribution
  - Symmetric like normal distribution
- Shape depends on degrees of freedom *m*
  - $m < 20$: fatter tails than normal distribution
  - $m > 30$: shape close to normal distribution
  - $m \rightarrow \infty$: exactly like normal distribution



Review of Probability:

The F Distribution, $F_{m,n}$

- Ratio of two independent chi-squared variables with degrees of freedom ***m*** and ***n***

  (W/m) / (V/n) where ***W*** and ***V*** are independently distributed chi-squared random variables

Shape depends on <u>two</u> degrees of freedom

- – Numerator d.f. ***m***
- – Denominator d.f. ***n***

Review of Probability:

Simple random sampling drawn from a population: $Y_1, \ldots, Y_n$

We will assume simple random sampling

- Choose an 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,…, Y_n), where Y_i = value of Y for the i^{th} individual (district, entity) sampled

Review of Probability:

Simple random sampling drawn from a population: $Y_1,\ldots, Y_n$

- 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 population (distribution). That is, Y_1, Y_2 are ***identically distributed***
  - So, 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.

Review of Probability:

Sampling distribution of the sample average

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

- The individuals in the sample are drawn at random.
- Thus the values of (Y_1, …, Y_n) are random
- Thus functions of (Y_1, …, Y_n), such as $\bar{Y}$, are random: had a different sample been drawn, they would have taken on a different value

Review of Probability:

Mean and variance of sampling distribution of Y

$$E(\bar{Y}) = \mu_Y \qquad \mathrm{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

Implications:

Y is an unbiased estimator of μ_Y

var(*Y*) is inversely proportional to *n*

- For small sample sizes, the distribution of *Y* is complicated, but if *n* is large, the sampling distribution is simple!
- As *n* increases, the distribution of *Y* becomes more tightly centered around μ_Y (the *Law of Large Numbers*)

How large is "large enough"?