

The background of the slide is a blue-tinted photograph of the UCI Paul Merage School of Business building. The building is a modern, multi-story structure with a curved facade and many windows. A large blue arc is on the left side of the slide, and a yellow arc is at the bottom left.

UCI Paul Merage
School of Business

Leadership for a Digitally Driven World™

MFIN 290: **Financial Econometrics**

Lecture 1-1



This time

- Review of Course Logistics
- Review of Probability and Statistics



Who am I?

Purpose

- This course will teach advanced Econometric concepts relevant to a career in Finance:
- Least Squares Regression
- Endogeneity Problems/Violations of LS Assumptions
- Time Series Econometrics
- Logistic Regression
- Panel Data
- LASSO Methods / PCA?

Purpose

- Genesis of course/content



Purpose

- This is a newer course with a custom curriculum!
- We hope to have in person lectures. If necessary, we will hold lectures via zoom:
<https://ucimerage.zoom.us/my/kleinsp>
- I want your feedback on the material:
- Interesting? More examples? Too much math?
- We can be flexible (within reason). Last year, we discussed material that wasn't understood from other courses earlier in the year, current topics in the workplace, markets, etc.

Purpose

- This course will teach advanced Econometric concepts relevant to a career in Finance:
- While there will be some mathematics in this course (statistics, multivariable calculus, linear algebra), this is **not a math course and it is not a programming course**. I will give you the syntax for all commands needed as well as sample code.
- The goal is to convey the concepts and intuition behind the issues, increased familiarity and comfort with an array of econometric issues, an understanding of the right tools and tests you should think of given specific concerns relevant to a career in empirical finance.

Purpose

- At the end of the course, you should be able to acquire data, estimate relationships, think critically / test key assumptions, and test hypotheses.
- The applications are based on projects you are likely to see if you pursue a career in finance or financial research at a bank, asset management company, or other financial institution.
- By design, there is a lot of content in this course. We are looking for exposure and understanding sufficient to support applied work in your career.



Location/Office Hours

- The class meets once per week in the evenings:
- Monday
- 5-8PM
- I will try to finish early. I will almost always give us a meaningful break in the middle.
- My office hours will be held during the break, after class, or by appointment as needed. Please come with questions.

Course Website

- <https://canvas.eee.uci.edu/courses/55879>
- Under “Files”:
- Appendices to Greene’s 8th edition are posted, including probability and matrix algebra reviews.
- Lecture notes will be posted in advance. I recommend you to review them before lecture and come with questions and review again promptly after class.
- Problem sets will be posted before the material is covered. Consistent with course design, you usually don’t have to know everything to do well: you can pick and choose from certain sections for the final exam.

Textbook/Materials

- Textbook
- *Principles of Econometrics, 4th Edition*. Hill, Griffiths, Lim, William H. 2011. Wiley. (Recommended)
 - Hope is the 4th edition is a bit less expensive.
 - Can be a useful reference. I'll be using some of their sample datasets (posted to the course website) to demonstrate concepts.
 - I recommend this text as a reference, but I have done my best to make the lectures stand-alone during and after COVID. Though if you are the type who would use it, more than one reference can only be helpful.

Textbook/Materials

- I will also demonstrate tests and concepts in statistical software. I am agnostic on what software you use for problem sets and final projects, though I will demonstrate in Stata and MATLAB as their language is mostly “out of the way”

Sample Software:

- Stata (the right tool when learning econometrics)
[Knowledge - Stata \(service-now.com\)](https://www.stata.com/knowledge)
- MATLAB (the right tool when simulating/solving systems/models)
[MATLAB Total Headcount Site License — Office of Information Technology \(uci.edu\)](https://www.mathworks.com/education/licenses/matlab-total-headcount-site-license.html)
- R
- Python (the right tool for production/scale/flexibility)

MATLAB is free for UCI students, Python and R are free for all. Stata is heavily discounted and a **lifetime license** (I am still using mine!)

Textbook/Materials

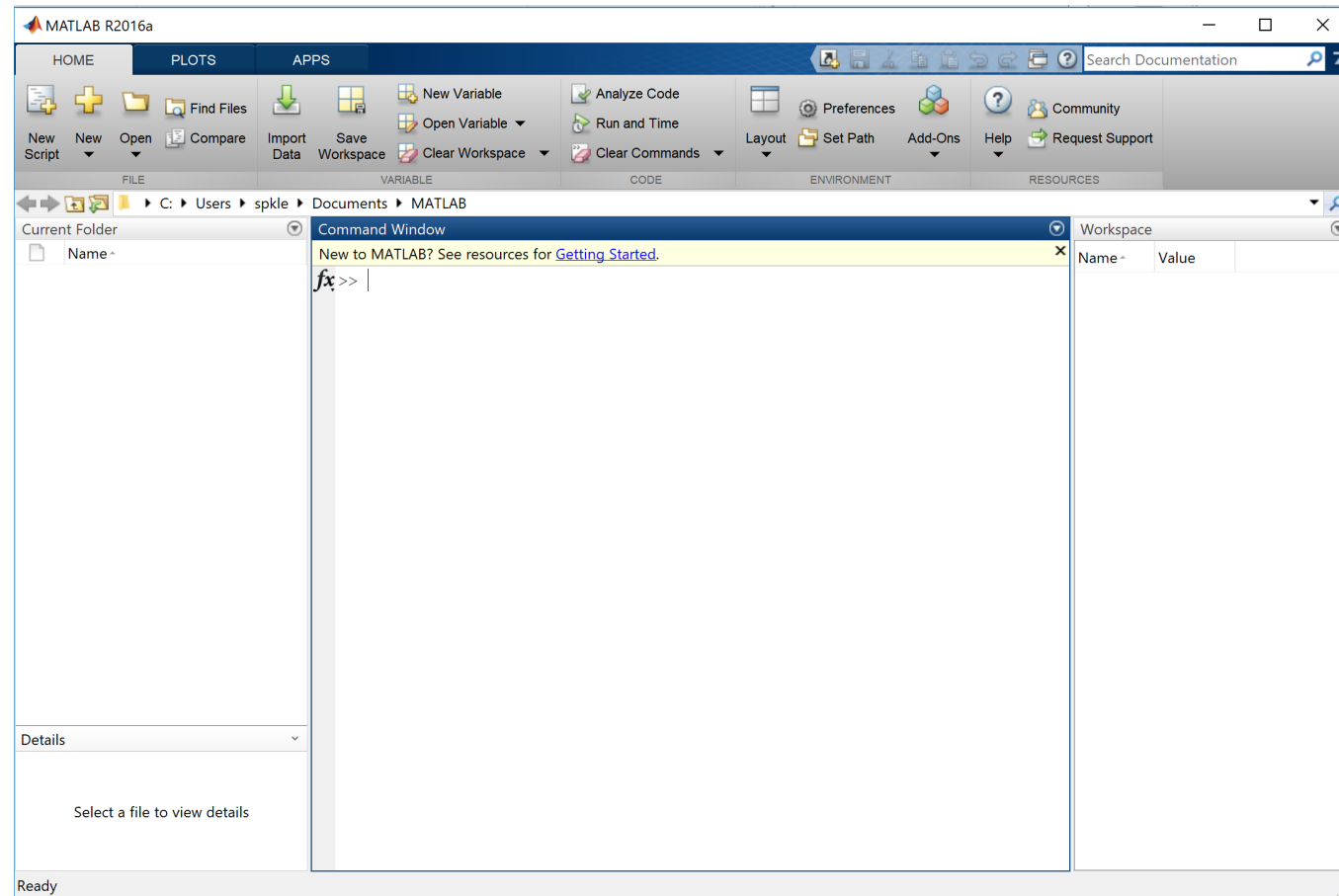
- Students can also use Stata and MATLAB for free through a virtual computer lab service setup by the group that supports computing for the entire campus, OIT. That can be accessed here:
<https://www.oit.uci.edu/labs/apporto/>.

The Paul Merage School of Business



MATLAB

The Paul Merage School of Business



Leadership for a Digitally Driven World™

Course Requirements

- Problem Sets/Quiz
- Problem Sets will be assigned and turned in through Canvas.
- You are allowed and encouraged to work together, though everyone must hand in their own solutions and log files.
- There will also be one quiz during week six taking half of our time that covers material in the first half of the course
- They collectively contribute 35% to your final grade

Course Requirements

- Exam
- There will be one comprehensive final exam in this course, timing TBD.
- It contributes 30% to your final grade.
- It is likely, but not guaranteed, that this will be a take home exam. Regardless of the format, you **MUST** complete it on your own and it may be checked for plagiarism.



Course Requirements

- Final Project
- There will be a final research project in this course worth 35% of your grade.
- You can select your groups. I am looking for 4 -6 groups in total.
- It will consist of a written research paper and a presentation.
- Both will be completed as a group and presented to the class during the final meetings.
- Groups and Preliminary Topics must be proposed by July 10 (due on Canvas).
- Please leave time for questions and ensure everyone presents some of the results.

Course Requirements

- Final Project
- The grade will be determined by:
 - 1) Peer evaluation (30%)
 - 2) My evaluation (20%)
 - 3) Quality of presentation - evaluated collectively (20%)
 - 4) Quality of written materials - evaluated collectively (30%)

Course Requirements

- Final Project
- In the working world, projects are completed as a team. Can spread the workload to complete a more thorough analysis. Late submissions will be severely penalized.
- I am aware of free rider issues! Each student must evaluate every *other* member of the group. Failure to do so will result in a reduction in your grade.



Class Time

I will endeavor to provide some class time regularly for you to connect with your group and work on the final project.

I will be available for questions, suggestions, or general feedback during this time as well. This will help mitigate free rider issues and ensure all groups have an opportunity to work through their ideas with me without coordinating schedules.

Grading

- Participation.
- This is a smaller class in a relatively small program. As in life, you will get far more out of it if you are engaged.
- Ask questions if something is confusing.
- Answer each other's questions if you can. Bring examples up in the class of current events that bear on course content (I love this).
- This should not be onerous.

Grading

- Final Grades are based on the course requirements.
- Grading mistakes are rare, but they do happen. Any grading disputes must be made with me, in writing, within one week of the materials return to you.

Questions/Problems

- Sean Klein
- KLEINSP@UCI.EDU
- Ask questions in class!
- Ask questions in office hours!
- *I am teaching this course because I enjoy working with students. I want to familiarize you with topics and concepts enough to know about them and when they are necessary.*



Academic Integrity

- I will be available throughout the quiz and final exam: you may ask questions or otherwise communicate with me, but not with your classmates.
- You must write your own problem sets and exams. Plagiarism will not be tolerated in this course.
- UCI takes academic dishonesty extremely seriously. Cheating will not be tolerated and any incidents will be reported to the Dean of Students.

What is Econometrics

Uses statistical methods to analyze economic data.

Idea is to answer quantitative questions given observed realities

We want to determine the effect of one variable (X) on another variable (Y). Usually we want this to be **causal**.

Econometric Models

$$Q^d = \beta_0 + \beta_1 P + \beta_2 P^s + \beta_3 P^c + \beta_4 Inc + \epsilon$$

Estimating quantity demanded as a function of its price, P , the price of other goods, P^s and P^c , and Income.

Note that substitute goods can be plural \Rightarrow this (and the coefficient) can be *vectors*

Econometric Models

$$Q^d = \beta_0 + \beta_1 P + \beta_2 P^s + \beta_3 P^c + \beta_4 Inc + \epsilon$$

Dependent Variable: The outcome that is being modeled. It is dependent on the other variables.

Independent Variable: The variables that drive the dependent variable.

Both are random variables!

Random Variable: takes on different outcomes with certain probabilities.

How are Econometric Models Used?



$$Q^d = \beta_0 + \beta_1 P + \beta_2 P^s + \beta_3 P^c + \beta_4 Inc + \epsilon$$

We use the econometric model to conduct “statistical inference”

Making quantitative statements about the likely relationships in the world.
Involves three steps:

- 1) Estimating the Parameters (β 's here)
- 2) Predicting the Outcomes
- 3) Testing Hypotheses

Data Generating Process

$$Q^d = \beta_0 + \beta_1 P + \beta_2 P^s + \beta_3 P^c + \beta_4 Inc + \epsilon$$

Econometric models try to inform the *process* that generated the data that you have:

Given this functional form, and given this data, what are the “best” coefficients?

What is the right interpretation?

Even with a full population of data, we are still estimating relationships from that sample: there is some underlying data generating process that created the population...

Heavily impacts interpretation and use of hypothesis tests...



Equilibrium and Endogeneity

Key contribution of economists is the idea of equilibrium.

When people/agents/firms all optimize, the relationships in the data may be jointly determined, it may not be so simple to identify the effects we are interested in...

The economic models illuminate what is possible to estimate and make explicit the assumptions we make doing so.

This is a key advantage of Econometrics!

Equilibrium and Endogeneity

Estimating Demand and Equilibrium: Don't regress price on quantity...



Part I

- Probability/Statistics review
- Chapter P in your textbook



Review: Random Variables

- Random Variable (x): takes on different outcomes with certain probabilities.
- Outcomes (x_i): *mutually exclusive* potential results of a random process.
- Probability of an outcome: proportion of times that the outcome occurs in the long run.



Review: Random Variables

- Types of Random Variables:
 - Discrete Random Variable: takes on a finite number of values.
 - Example: coin toss; the number of times a computer crashes, number of students with an A in this course

Continuous Random Variable: takes on any value in a real interval

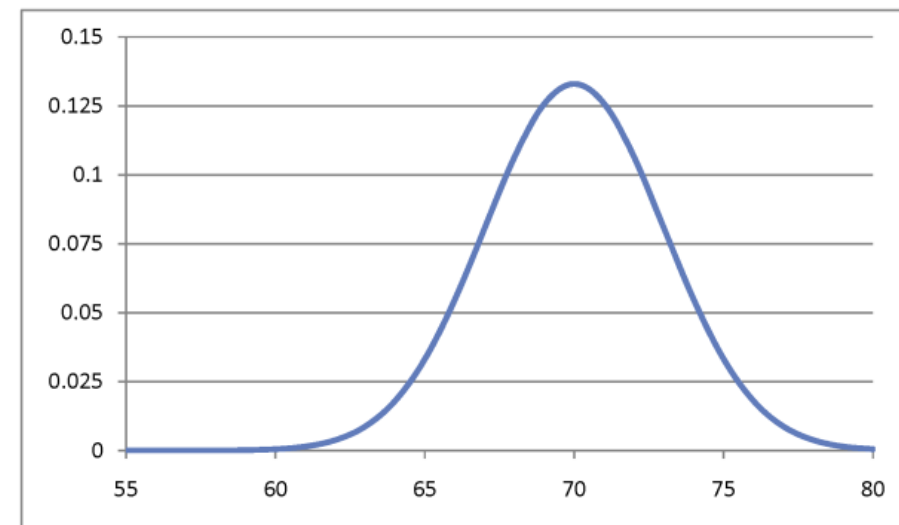
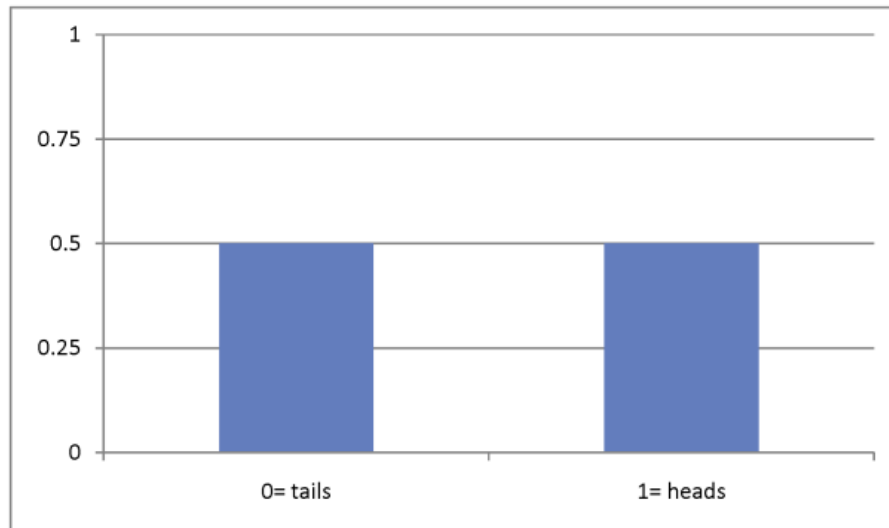
- Example: height of an individual; time it takes to commute to school (about that....).
- Each specific value has **zero probability**.

Review: PDF

- The function f that summarizes the information relating the possible outcomes of a random variable X and the corresponding probabilities is called the PDF, which stands for:
 - the Probability Distribution Function, for discrete variables; or
 - the Probability Density Function, for continuous functions.
- PDF's must satisfy:
 - $f(x_i) \geq 0$; and
 - $\sum_i f(x_i) = 1$ (if discrete) or $\int f(x) dx = 1$ (if continuous)
 - Probabilities are weakly positive
 - Probabilities sum to one

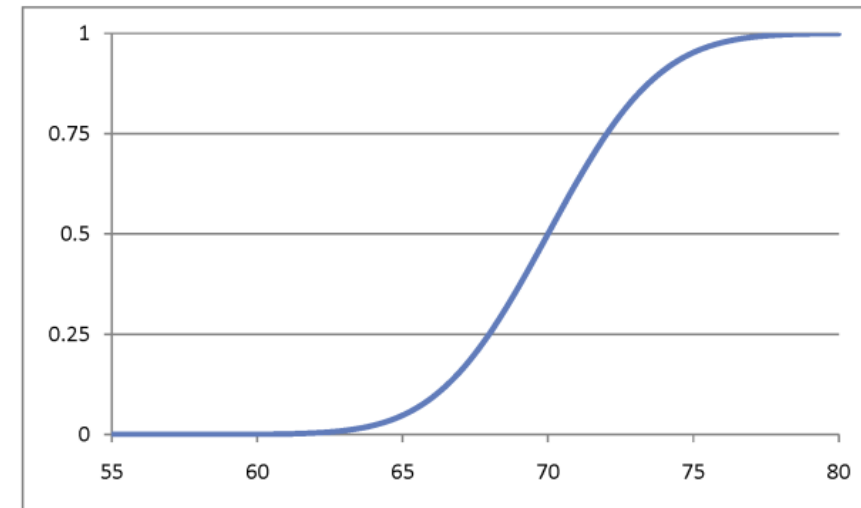
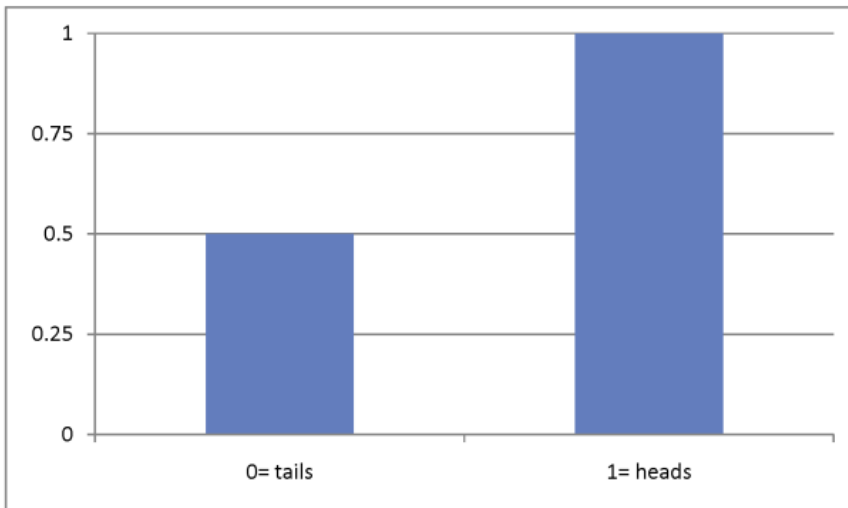
Review: PDF

- What do these look like?



Review: CDF

- Probability that x_i is less than or equal to a certain value
- Typically use $F(x_i)$. Note: $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow \infty} F(x) = 1$... why?



Review: Moments

- These are summary statistics of the distribution that take advantage of the PDF/CDFs.
- First one is the “Mean”/”Expectation”/”Expected Value”

$$\mu_X = E(X) = \begin{cases} \sum_i x_i f(x_i) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$


- Gives a sense of where the center of the distribution is...
- Weighted average of the values where the weights are the probabilities
- The sample average is the sample analog to this population concept.

Example in Stata

1. Generate 1000 random normal variables
2. Plot a histogram
3. Transform variable
4. Summarize sample

Help files are great! (see at right for drawnorm)

Create normal variables using “drawnorm”,
exponential transform using “generate”,
histogram with “hist”, information with
“summarize <>, detail”



The screenshot shows the Stata help window for the `drawnorm` command. The title bar reads "help drawnorm *". The window content includes the command description, syntax, and a table of options.

Title

[D] `drawnorm` — Draw sample from multivariate normal distribution

Syntax

```
drawnorm newvarlist [, options]
```

options	Description
Main	
<code>clear</code>	replace the current dataset
<code>double</code>	generate variable type as double ; default is float
<code>n(#)</code>	# of observations to be generated; default is current number
<code>sds(vector)</code>	standard deviations of generated variables
<code>corr(matrix vector)</code>	correlation matrix
<code>cov(matrix vector)</code>	covariance matrix
<code>cstorage(full)</code>	correlation/covariance structure is stored as a symmetric k*k matrix
<code>cstorage(lower)</code>	correlation/covariance structure is stored as a lower triangular matrix
<code>cstorage(upper)</code>	correlation/covariance structure is stored as an upper triangular matrix
<code>forcepsd</code>	force the covariance/correlation matrix to be positive semidefinite
<code>means(vector)</code>	means of generated variables; default is means(0)
Options	
<code>seed(#)</code>	seed for random-number generator

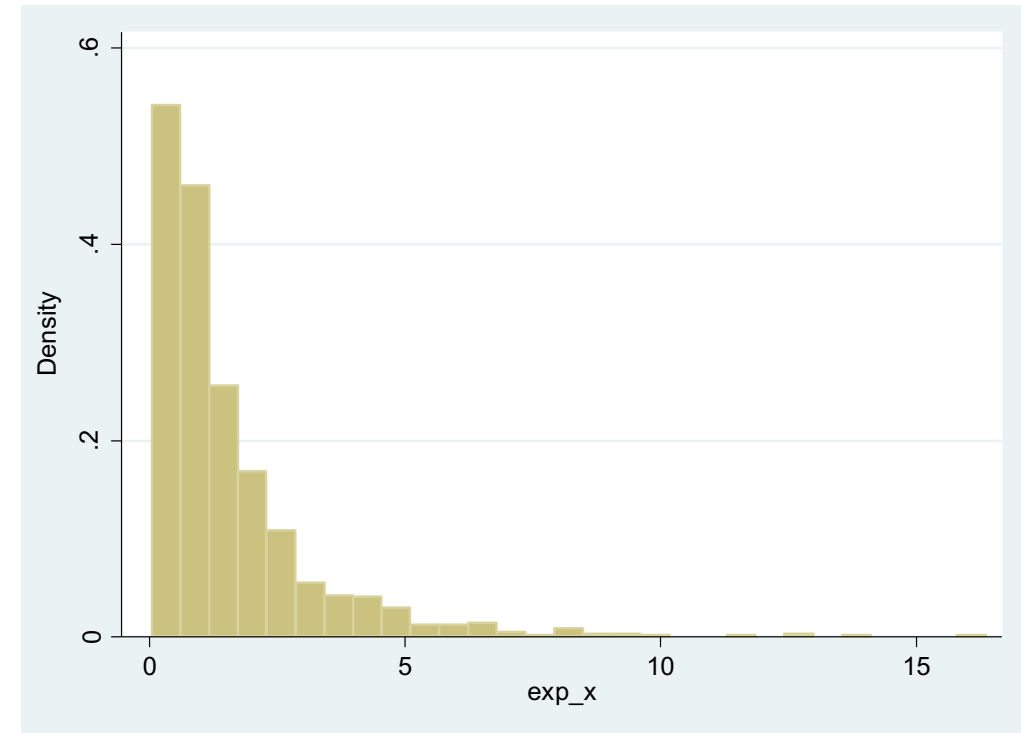
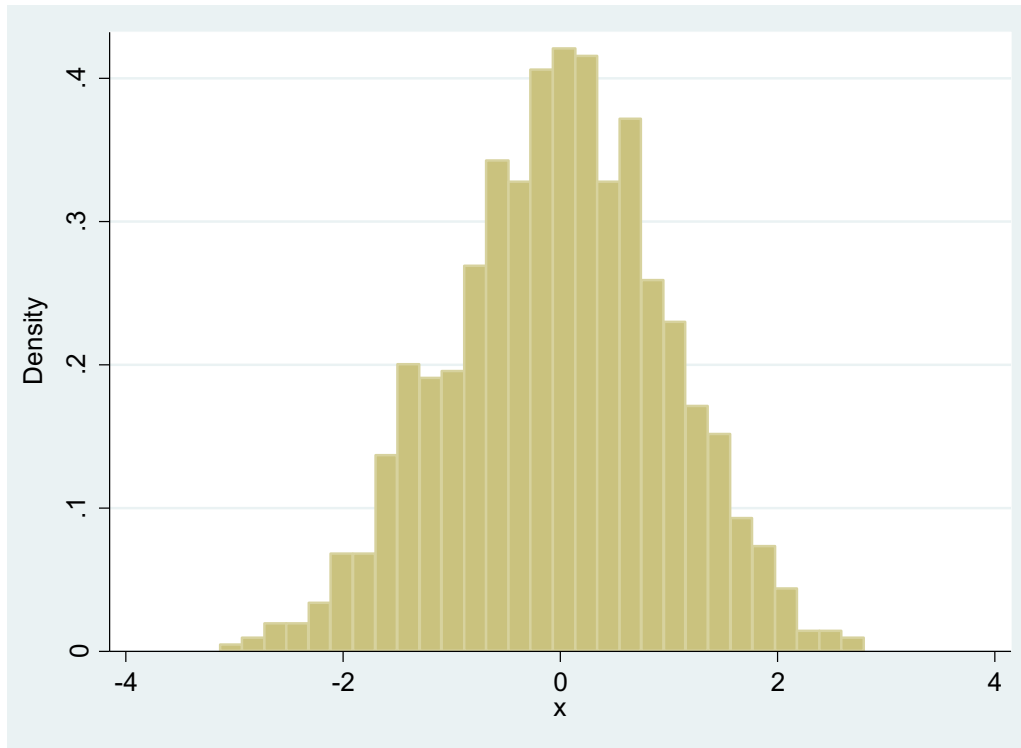
Example in Stata

1. Generate 1000 random normal variables
2. Plot a histogram
3. Transform variable
4. Summarize sample

Normal variables using “drawnorm”,
exponential transform using “generate”,
histogram with “hist”, information with
“summarize <>, detail”

```
Untitled.do* x
1  clear
2
3  set obs 1000
4  drawnorm x
5  hist x
6
7  gen exp_x = exp(x)
8  hist exp_x
9  |
```

Example in Stata



Normal variables using “drawnorm”, exponential transform using “generate”, histogram with “hist”, information with “summarize <>, detail”

Review: Moments

- Second (central) one is “Variance”

$$\begin{aligned}\sigma_X^2 &= \text{Var}(X) = E((X - \mu_X)^2) \\ &= \begin{cases} \sum_i (x_i - \mu_X)^2 f(x_i) & \text{if } X \text{ discrete} \\ \int (x - \mu_X)^2 f(x) dx & \text{if } X \text{ cont.} \end{cases}\end{aligned}$$

- You aren’t going to draw the mean (with probability one for continuous RVs). Given that you will be off, how far off will you be? Square it to cancel the positives and negatives...
- Gives a sense of the dispersion of the data. Square root of variance is called the “standard deviation”
- “summarize” in Stata.
- **Sample moments treat $f(x)$ as $\frac{1}{n}$ for each observation. Why does this make sense?**

Review: Moments

- Gives a sense of the dispersion of the data. Square root of variance is called the “standard deviation”
- “summarize” in Stata.

```
. summ x exp_x
```

Variable	Obs	Mean	Std. Dev.	Min	Max
x	1,000	-.0254268	.9922981	-3.134875	2.794861
exp_x	1,000	1.555559	1.731772	.0435052	16.36036

- Note that $E(e^X) = 1.55 \neq e^{E(x)} = e^0 = 1 \dots$

Review: Two Random Variables

- Joint Probability Density Function of X and Y : describes the probability that the random variables simultaneously take on certain values, x and y .

$$f_{X,Y}(x, y) = \Pr(X = x, Y = y)$$

- Marginal probability density: given the joint distribution, what is the probability that we see a given value for ONE of the variables?

$$f_Y(y) = \sum_{x_i} \Pr(X = x_i, Y = y) = \sum_x f_{X,Y}(x, y)$$

Review: Two Random Variables

- Conditional Probability Distribution
- Given a joint PDF, what is the PDF of Y given X takes on a certain value:

$$\begin{aligned}f_{Y|X}(y | x) &= \Pr(Y = y | X = x) \\&= \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} \\&= \frac{f_{X,Y}(x, y)}{f_X(x)}\end{aligned}$$

Covariance/Correlation

- How do variables move together? Informs “When X is above its mean, where do we think Y will be?”

$$\sigma_{XY} = \text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

- This is the “Covariance” of X and Y
- Remember, $E(\)$ applies probability weighting... this time of the joint PDF
- Issue: If the variance of X is higher, the covariance will be higher (in absolute value). This can be difficult to interpret.
- Solution?: Can rephrase as “When X is a certain number of Standard Deviations above its mean, how many standard deviations above its mean will Y be?” ...

Covariance/Correlation

- This is the Correlation between X and Y

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Between -1 and +1. If it is positive, then we expect Y to increase along with X, if it is negative, Y will decrease. If it is larger, there is a stronger association between the two variables.

Notes:

- $\rho_{X,X} = \rho_{X,aX} = 1$
- $\rho_{X,Y} = \rho_{Y,X}$
- This is DESCRIPTIVE, not CAUSAL.

“correlate” in Stata. Expressed in a (symmetric) matrix.

Covariance/Correlation

```
. matrix C = (1, .5 \ .5, 1)
```

```
. drawnorm x y, n(1000) corr(C)
(obs 1,000)
```

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
x	1,000	.023225	.9618351	-2.908514	2.847888
y	1,000	.0068477	1.003827	-3.276853	3.435619

```
. corr x y
(obs=1,000)
```

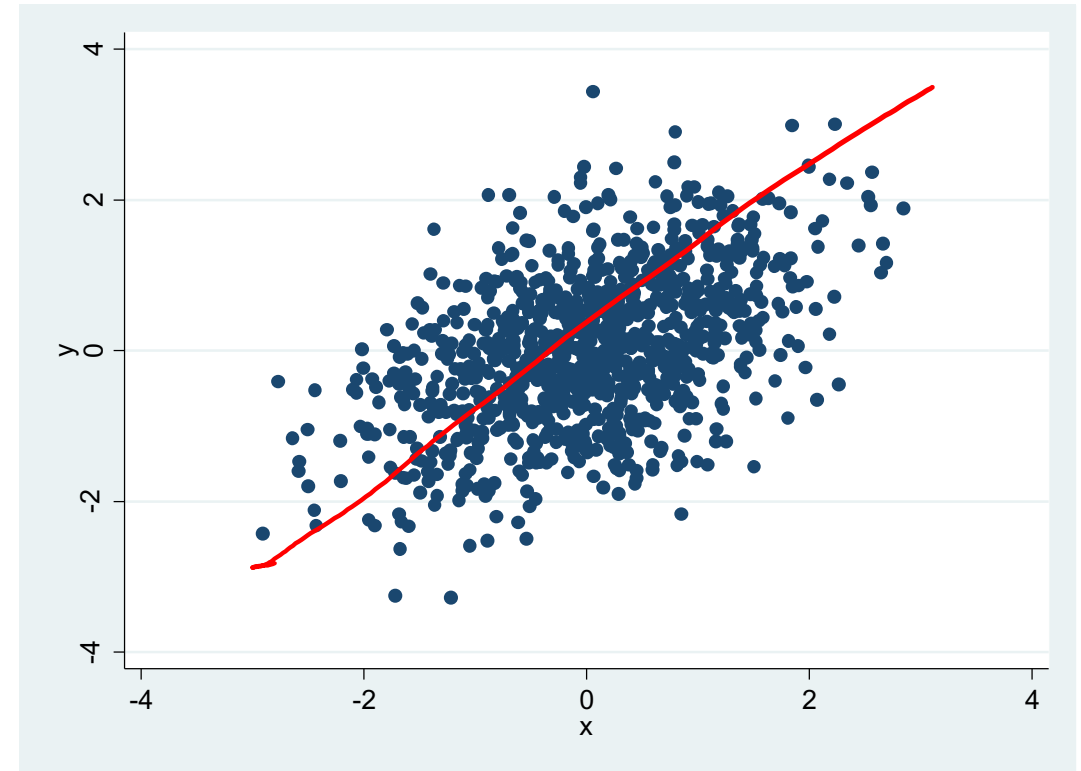
	x	y
x	1.0000	
y	0.4813	1.0000

Covariance/Correlation

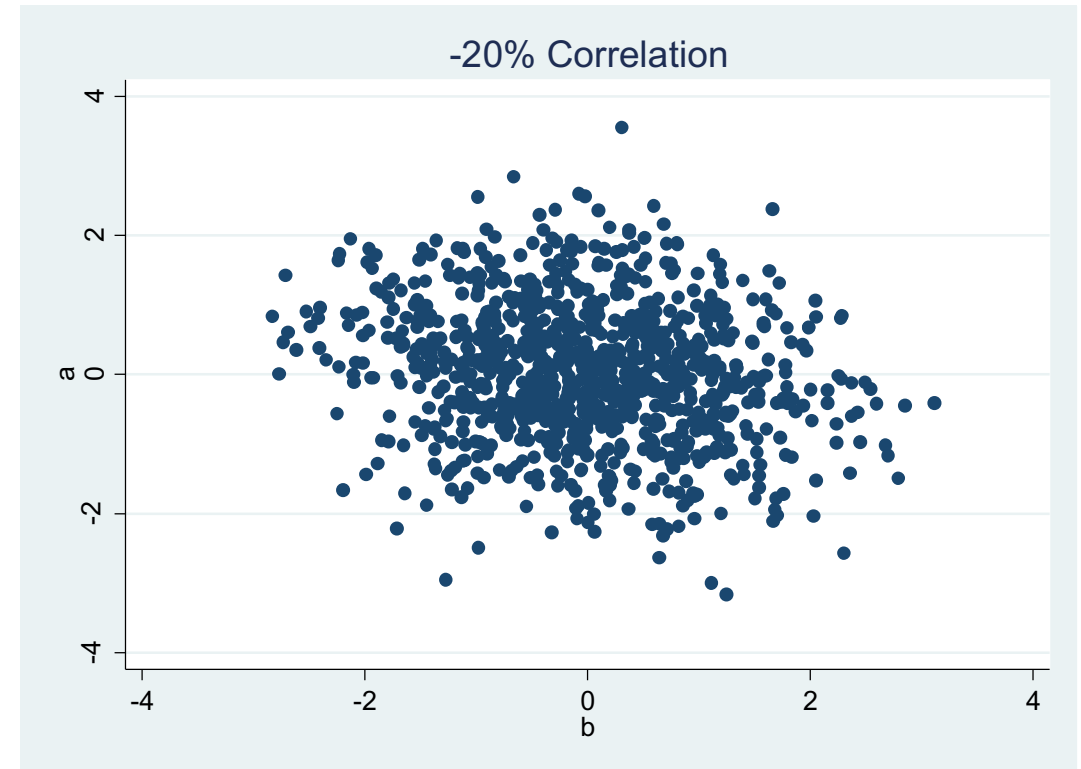
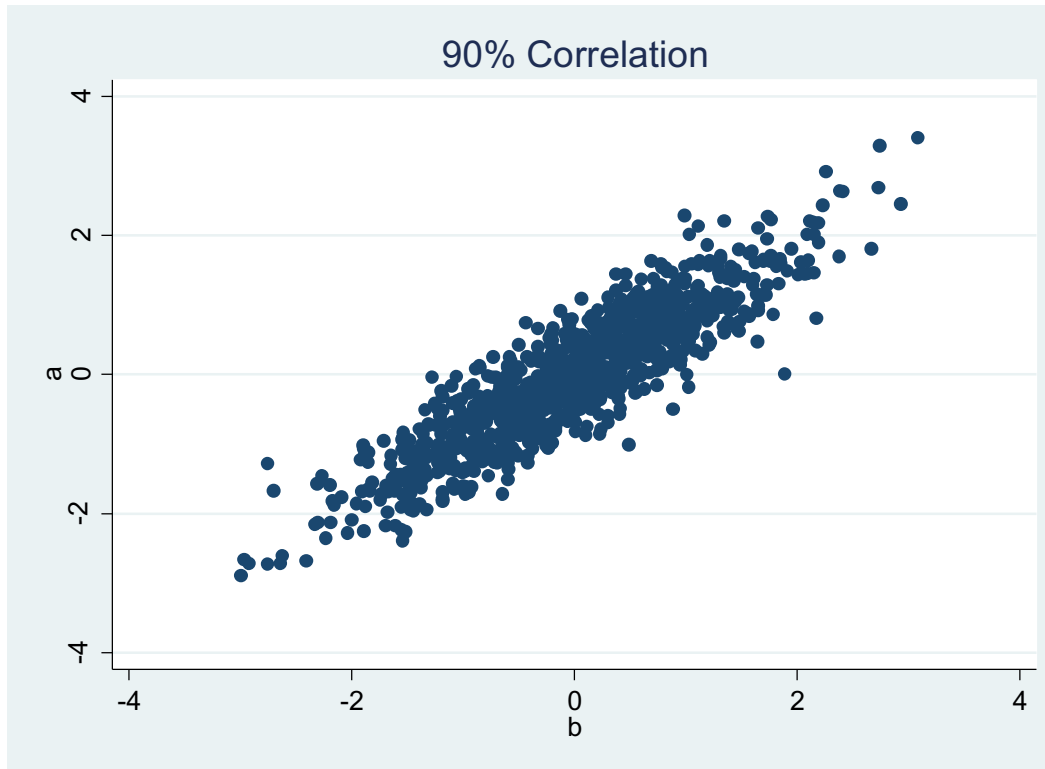
“two-way scatter y x ”

Graphs a scatterplot of x against y .

Note the vertical axis comes first



Covariance/Correlation



Covariance/Correlation

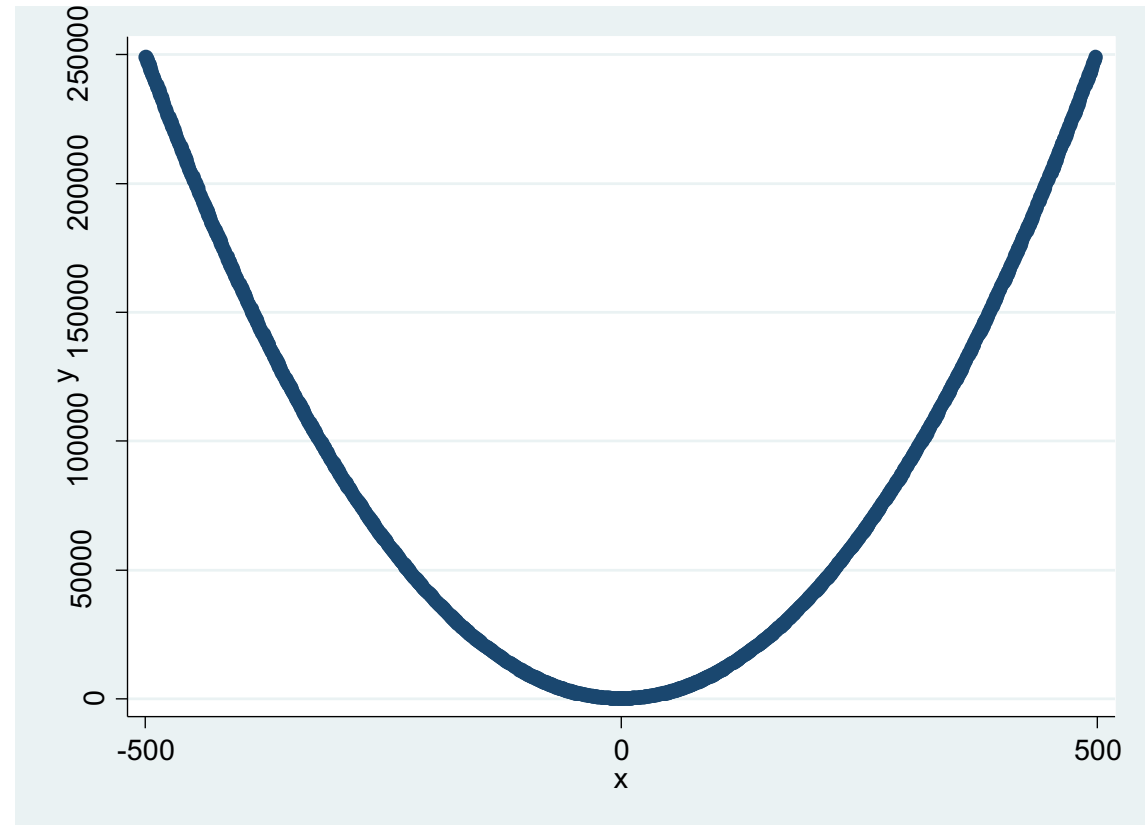
Correlation is a measure of LINEAR relationships.

Can perfectly know y given x , and still see a correlation of zero (or any other number!).

```
. gen y = x^2
```

```
. correl x y  
(obs=999)
```

	x	y
x	1.0000	
y	-0.0000	1.0000



Independence

- If two variables are independent, knowing one tells you nothing new about the other.
- To be precise: the conditional distribution is equal to the marginal...

$$f_{Y|X}(y | x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = f_Y(y)$$

- Often expressed as the joint is the product of the two marginals.

Properties of Expectations

$$E(a) = a$$

$$E(aX + b) = aE(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

$$\text{Var}(a) = 0$$

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

Properties of Expectations

$$E(a) = a$$

$$E(aX + b) = aE(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

$$\text{Var}(a) = 0$$

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

Properties of Expectations

$$E(a) = a$$

$$E(aX + b) = aE(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

$$\text{Var}(a) = 0$$

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$\int (aX - a\mu_x)^2 f(x) dx = \int a^2 (X - \mu_x)^2 f(x) dx$$

Properties of Expectations

$$\text{Var}(X) = \text{E}[(X - \text{E}[X])^2]$$

Expand this out and distribute the expectation....

$$= \text{E}[X^2] - \text{E}[X]^2$$

Properties of Expectations

$$\text{Var}(X) = \text{E}[(X - \text{E}[X])^2]$$

$$= \text{E}[X^2] - \text{E}[X]^2$$

Properties of Expectations

$$\text{cov}(X, Y) = \text{E}[(X - \text{E}[X]) (Y - \text{E}[Y])]$$

Expand this out and distribute the expectation.... Left as exercise

$$= \text{E}[XY] - \text{E}[X] \text{E}[Y].$$

$$\text{cov}(aX, bY) = ab \text{ cov}(X, Y)$$

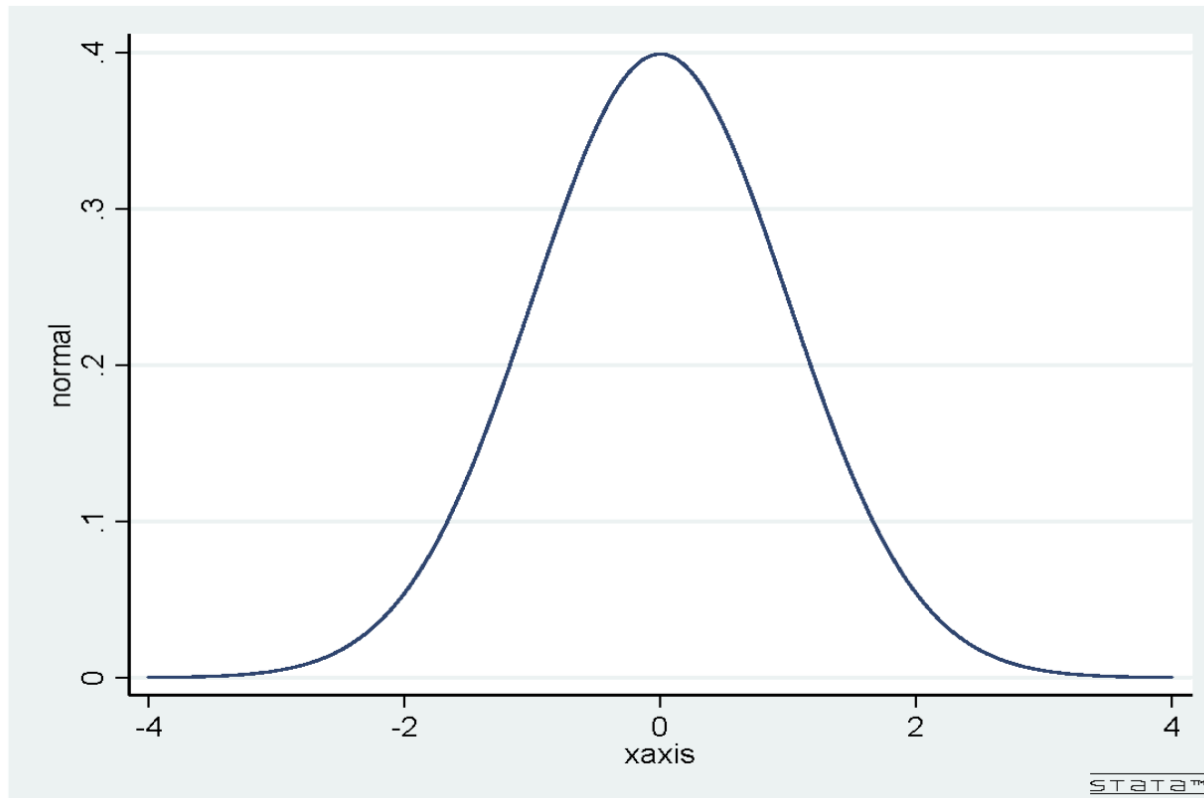
$$\text{cov}(X + a, Y + b) = \text{cov}(X, Y)$$



Common Distributions

- Normal/Gaussian:
- $N(\mu, \sigma^2)$: “Mean μ ” and “Variance σ^2 ”. These two parameters tell us every point in the PDF/CDF
- “Standard Normal” has mean zero and variance 1
- $N(0,1)$, often written as Z
- CDF of standard normal given by Φ : $\Pr(Z \leq a) = \Phi(a)$
- If X is $N(\mu, \sigma^2)$, then $z = \frac{x-\mu}{\sigma}$ has variance 1 and mean 0, still normal => is “standard normal”.
- If X and Y are both normal, then the linear combination $aX + bY$ is normal

Common Distributions



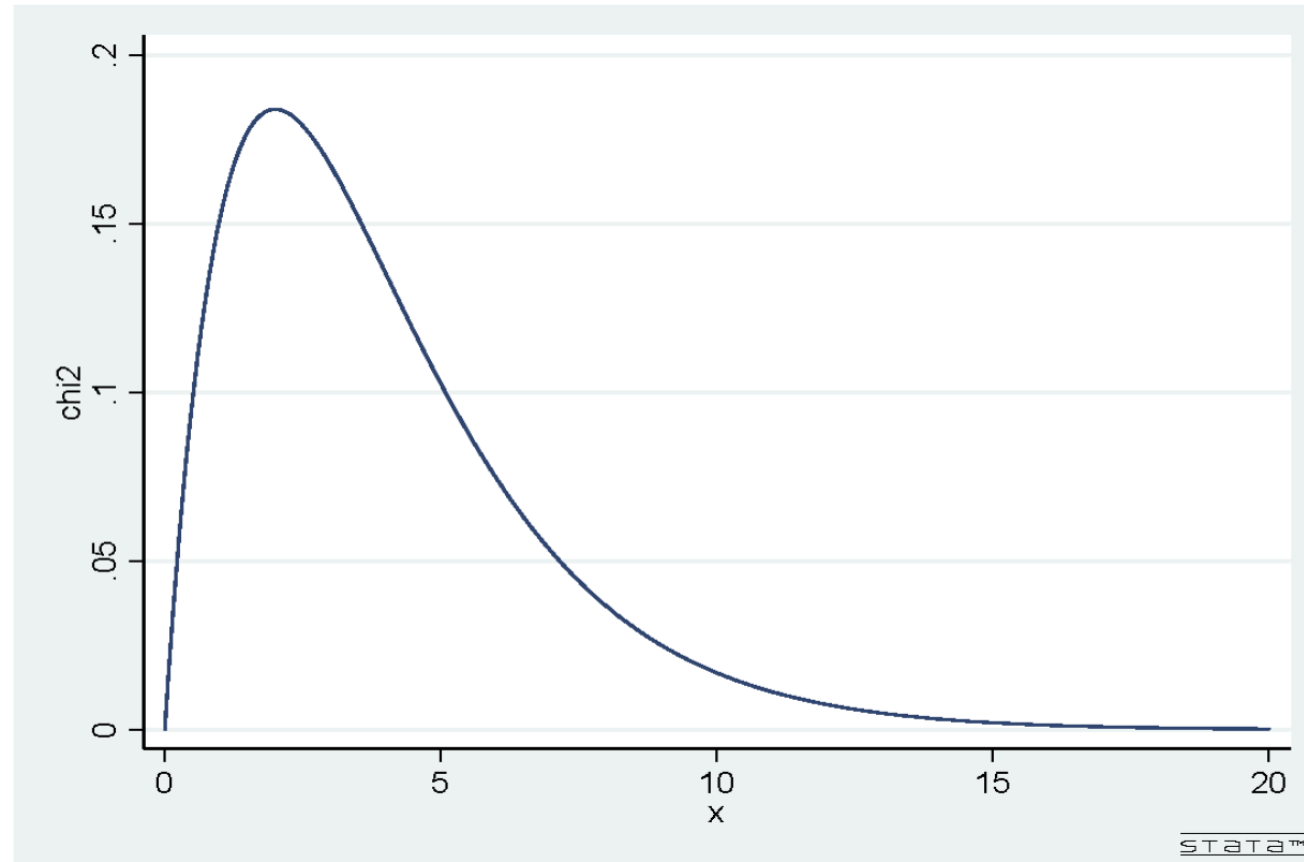
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$



Common Distributions

- Chi-Squared
- This distribution arises when we take N independent Standard Normal variables, square them, and add them up.
- $W = \sum_{i=1}^N z_i^2 \sim \chi_N^2 = \text{Chi-squared with } N \text{ degrees of freedom}$
- Will be skewed – a lot of small positive values, very few large positive values.
- Intuition:
- Mean should be $N = \text{df}$: expect to be off by 1 std for each of the N independent Normals...

Common Distributions



Common Distributions

- T-distribution
- Take a Z that is Standard Normal, W as a Chi Squared with m degrees of freedom (independent). If we “scale” Z by the standardized W, we get another distribution.

$$\frac{Z}{\sqrt{W/m}} \sim t_m$$

- T distribution with m degrees of freedom.
- This sounds strange, but occurs all the time: we want to express a variable in standard deviations above the mean, but don't know the right variance => we have to estimate it with an estimator that uses the sum of squares (like a variance should).
- “frequency distribution of (estimated) standard deviations of samples drawn from a normal population”

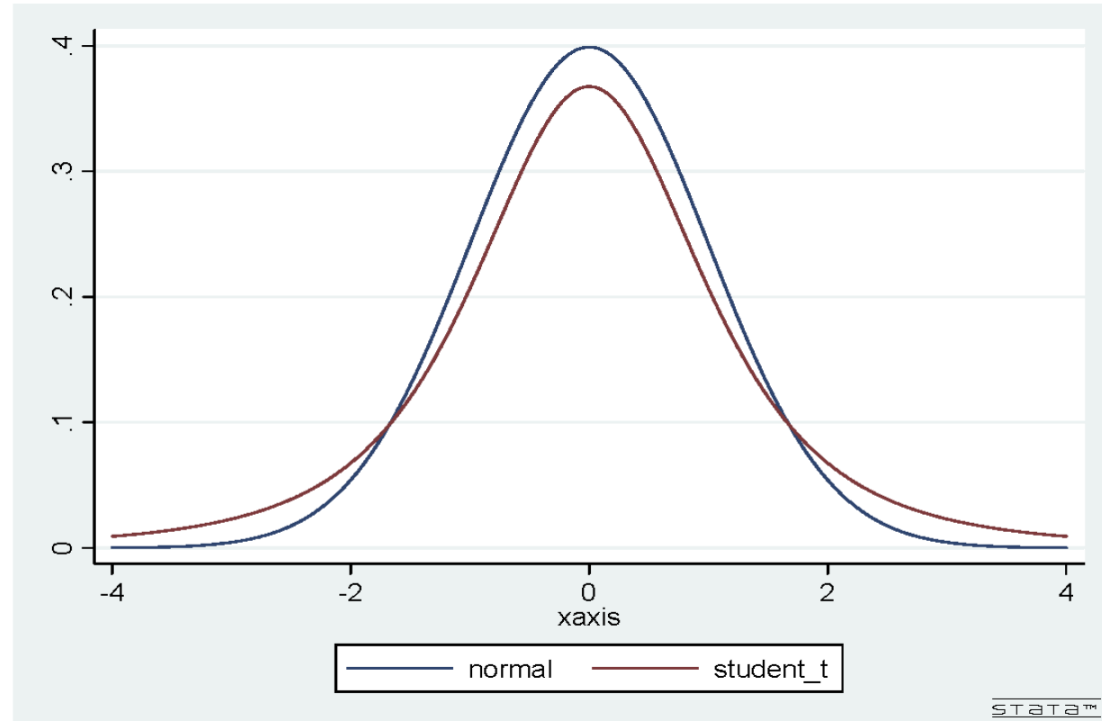


Common Distributions

- Where does this come from?
- Beer!
- Employee at the Guinness Factory in Dublin (William Gosset) was interested in small sample problems- how to test the quality of ingredients with few batches, for example.
- Published under a pseudonym (“Student”)

Common Distributions

- In the limit (as m increases) this heads to a normal distribution. For finite m , the tails will be fatter... why does this make sense?



Common Distributions

- F distribution
- Imagine that we have two Chi Squared distributions:
- W with N degrees of freedom and V with M degrees of freedom. If we scale each by their degrees of freedom and take the ratio, that has a known distribution.
- $\frac{W/N}{V/M} \sim F_{n,m}$
- This comes up if we want to compare two estimated relationships: can look at the “errors” and compare them.
- $nF_{n,\infty} = \chi_m^2$ (intuition: we end up dividing by a single value)
 $F_{1,m} = t_m^2$



Recovering Probabilities

- We can use these distributions to compute the probability of any event occurring.
- EX: Assume that adult male heights are normally distributed with mean 70 inches and standard deviation 4 inches
- Let X denote the height of men. $X \sim N(70, 16)$.
- Note: Variance = 16 \Rightarrow Standard Deviation = 4.

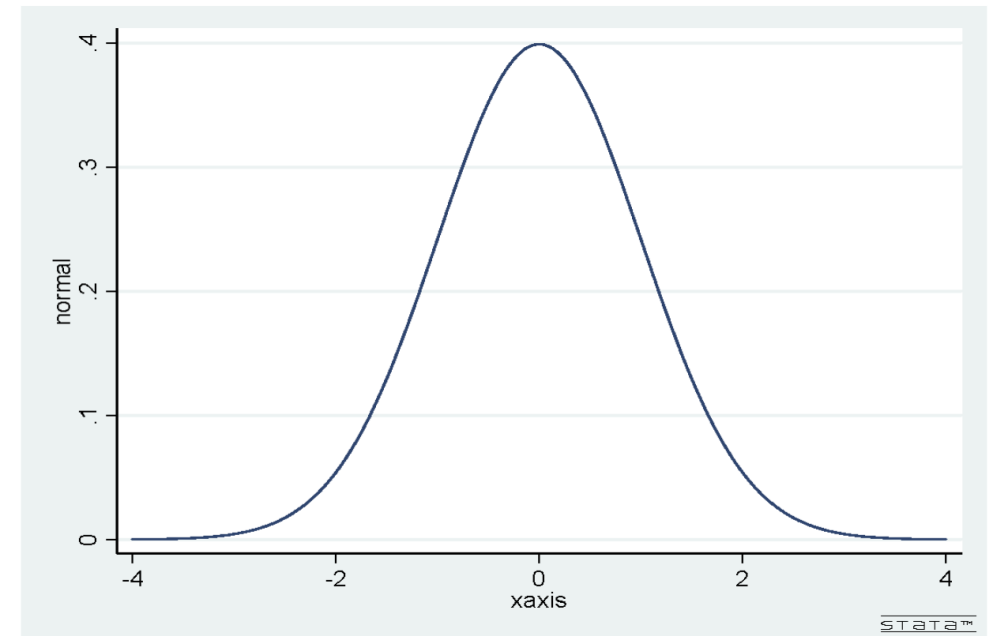
Recovering Probabilities

- If you are 65 inches tall, what percentage of men are shorter than you?
- 1) Standardize X :
- 2) Look up critical value in normal table/in your software

Recovering Probabilities

- If you are 65 inches tall, what percentage of men are shorter than you?
- 1) Standardize X :
- $Z = (X - 70)/4 = -5/4$
- Now we have $Z \sim N(0,1)$, and we know
- 2) $\phi(-5/4) = 10.6\%$

```
. di normal(-1.25)  
.10564977
```



When in doubt, Simulate!

```
. drawnorm x,n(1000) sd(4) mean(70)
(obs 1,000)

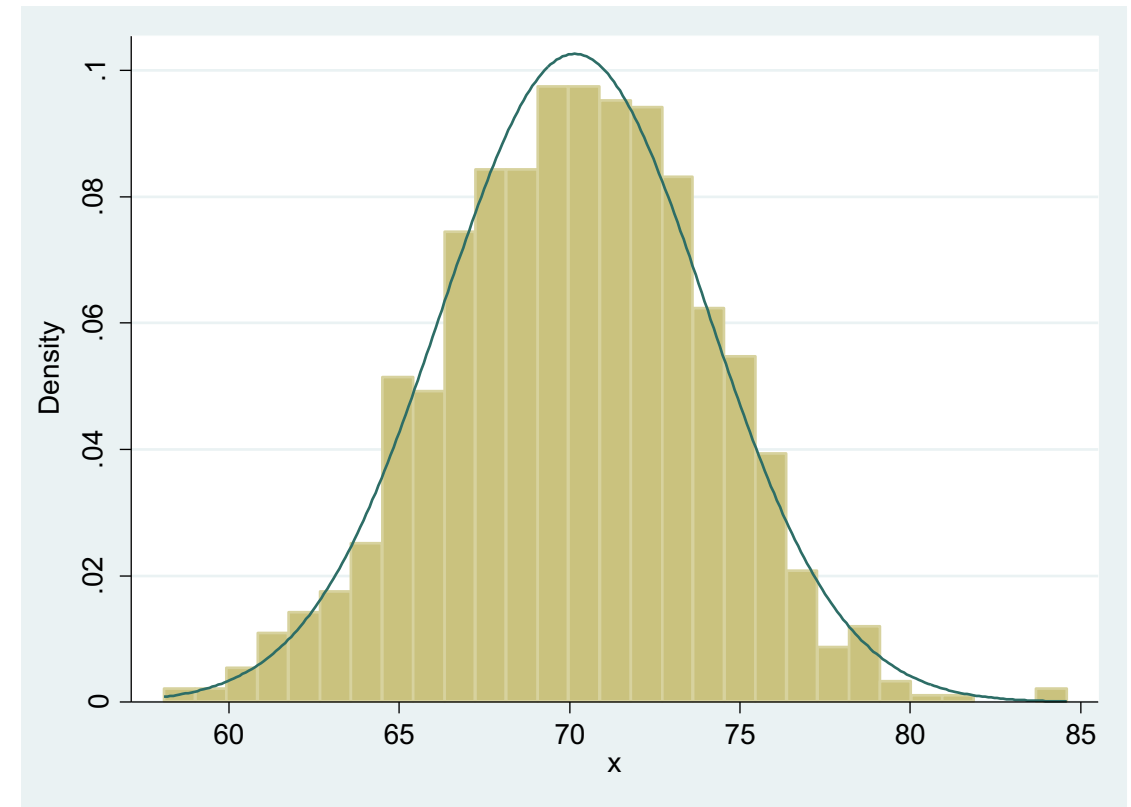
. hist x, normal
(bin=29, start=58.101913, width=.91355422)
```

```
. count if x <=65
91
```

Not quite 106... but “close”, right?

```
. drawnorm x,n(1000000) sd(4) mean(70)
(obs 1,000,000)

. count if x <=65
106,045
```



Recovering Probabilities

- Example 2: Payoff to an investment is normally distributed with a mean of \$2.5M and standard deviation of \$300,000. What is the probability the payoff is greater than \$3M?
- 0) Draw a picture
- 1) Standardize X :
 - $Z = (\$3M - \$2.5M)/\$0.3M = 5/3$
 - Now we have $Z \sim N(0,1)$, and we know this distribution
 - 2) $1 - \phi(5/3) = 5\%$

```
. di 1-normal(1.67)  
.04745968
```