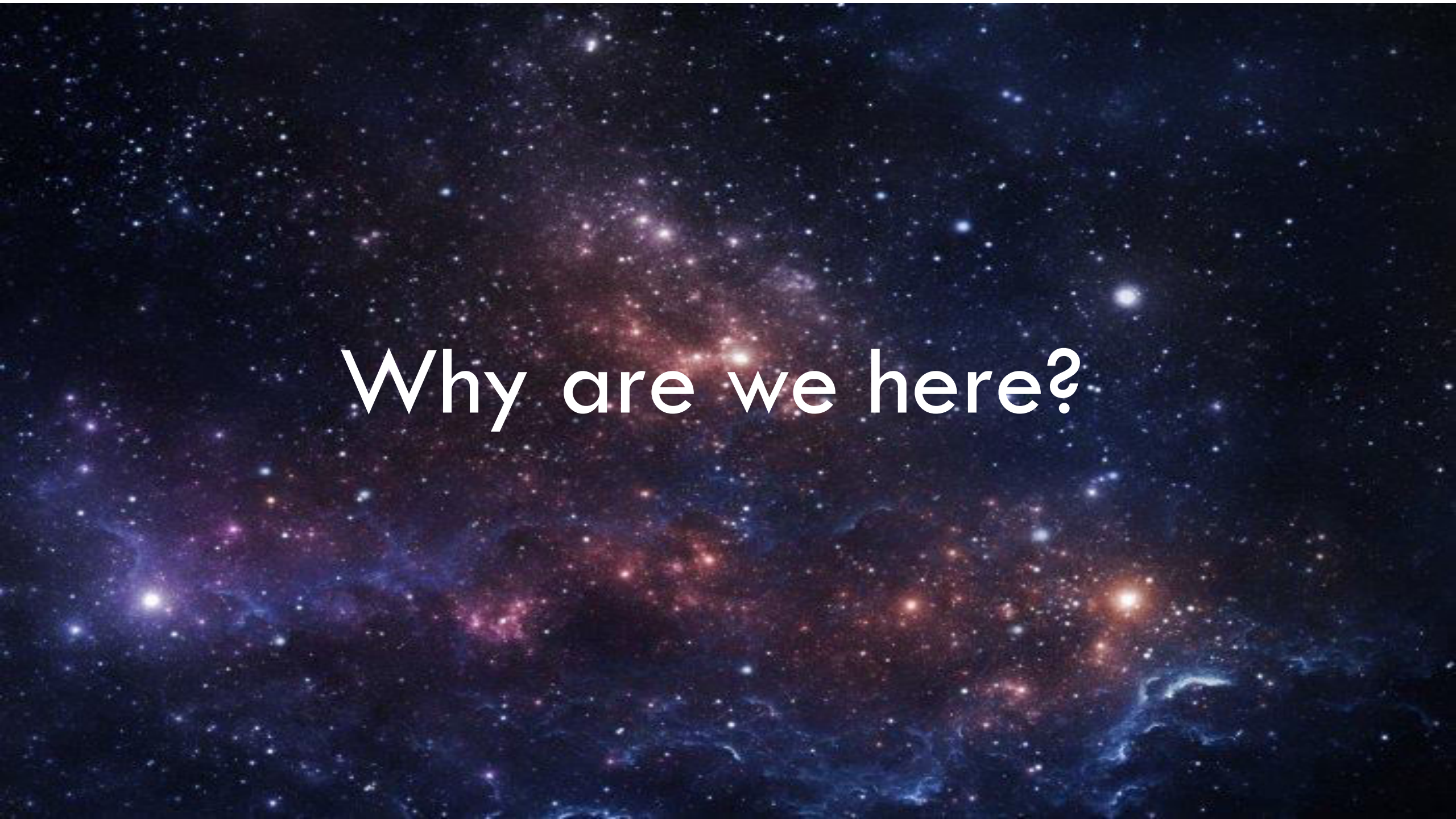
Advanced Data Analysis I
Course Introduction

PA 541 Week 1

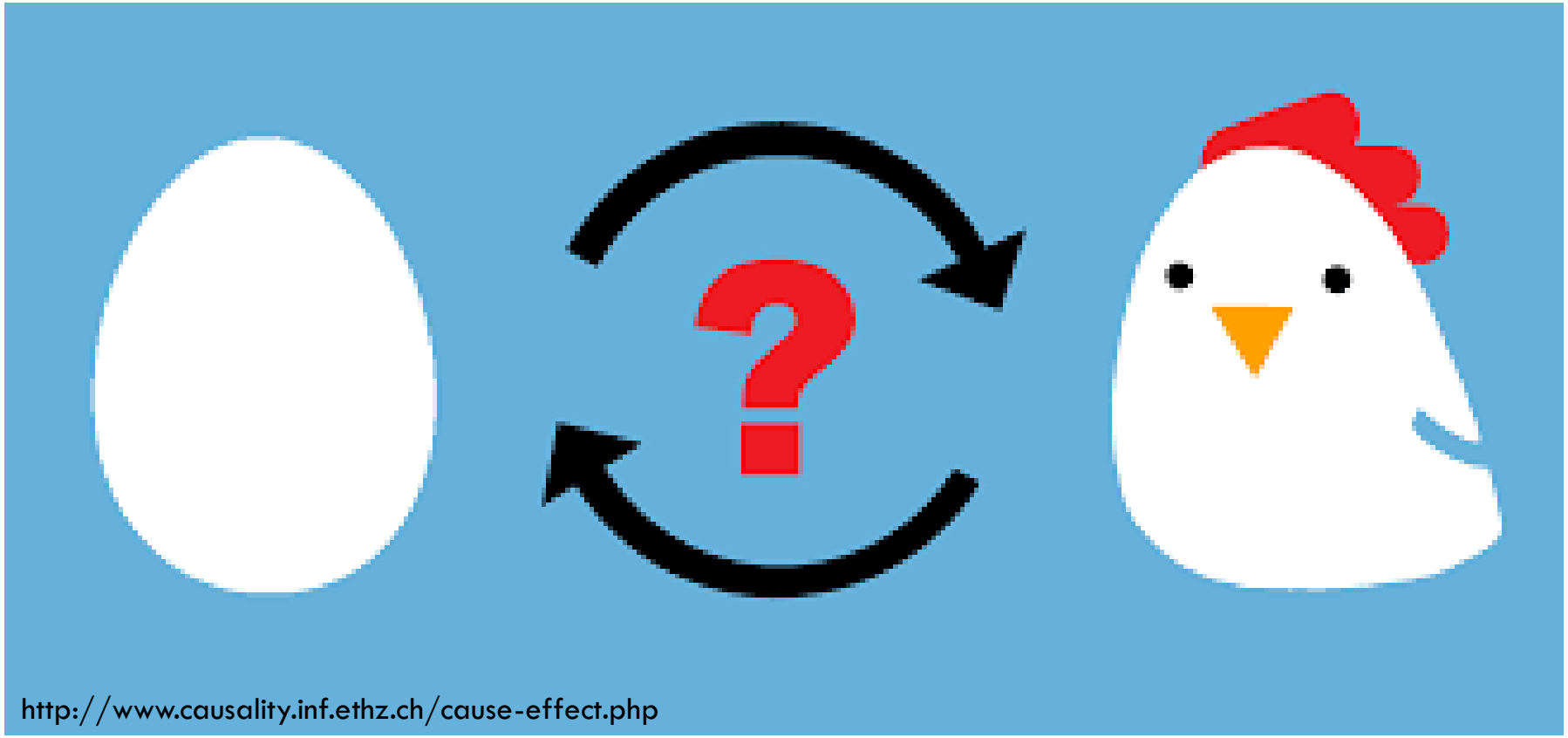
Michael D. Siciliano
Department of Public Administration
College of Urban Planning and Public Affairs


Why are we here?

Answer policy relevant questions

- Does going to college increase one's earnings?
- Does providing health insurance to people make them healthier?
- Does attending a charter school as opposed to a public school improve student performance?
- Does attending various police-community events influence an individual's perception of or trust in the police?

Causality



http://www.causality.inf.ethz.ch/cause-effect.php

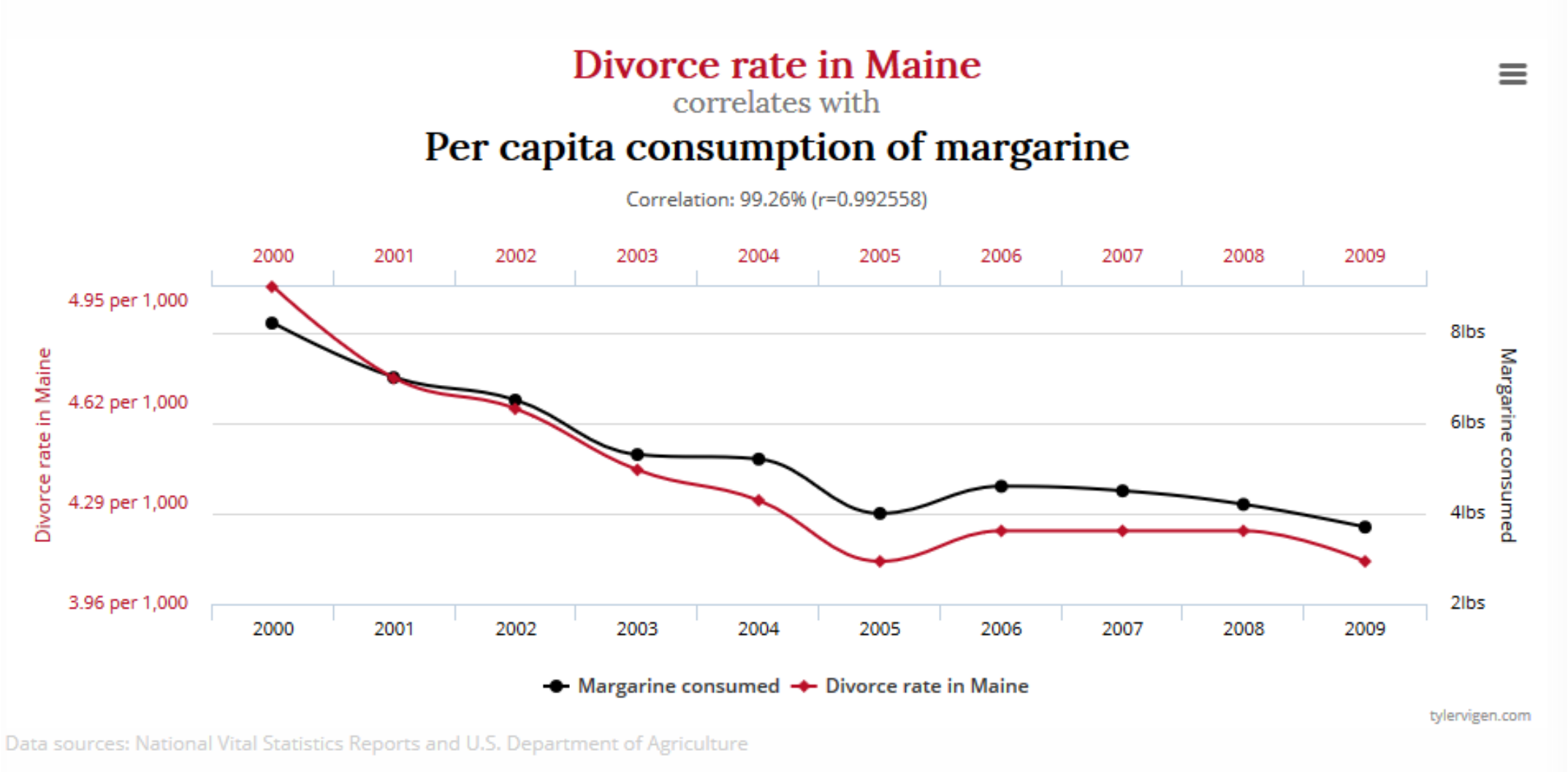
What is wrong with the following claim: Data show that individual income and marriage have a high positive correlation. Therefore, your earnings will increase if you get married (from Pearl et al. 2016).

Two major challenges to causality

- **Randomness** – is what we observe really just a coincidence?

- **Endogeneity** – is the relationship we observe really just due to another variable?

An example of randomness



An example of endogeneity

Ice cream sales predict homicide rate.

Course overview

- My main goal is to develop your skills and abilities to conduct empirical research using linear models in hopes of producing accurate inferences.
- The main focus will be on your ability to appropriately develop and interpret statistical models based on a strong conceptual understanding of modeling assumptions and limitations.
- Topics to be covered in this course include standard OLS regression, regression with qualitative predictors, model specification, data issues, limited dependent variable models, panel data methods, and data visualization.

Regression

- Two purposes of regression:
  - Prediction
    - Focus is on predicting the outcome value
  - Explanation
    - Focus is on identifying the effect of predictor variables on the outcome

When might we prefer one to the other?

What is econometrics

- Econometrics is the application of statistics and theory in order to test hypotheses.
- Theory will be used to help describe the relationship between the variables under study.
  - Theory provides a lens for our world; it helps us determine what we should be looking at; what variables we should measure.
- For example, the law of supply and demand tells us that there is a negative relationship between price and quantity demanded.
  - If you work for a city or state considering an increase in cigarette taxes (as Chicago has done in the past) you would want to know the magnitude of the relationship between price and quantity. In other words, by how much would demand fall if we increased price by $2.
  - To answer these questions, we can investigate the empirical relationship between cigarette prices and cigarette purchases.

Probability and Statistics

- In probability the parameters are known and they control the behavior of a random variable via a model.
  - We use the known odds or probabilities to estimate the likelihood of certain future events occurring.

- In statistics (probability in reverse) the random variables (or the data) are known, and they are used to estimate the unknown parameters that gave rise to them via a model.

Here is an example

- Model: suppose we play 100 independent win-or-lose games, each with a probability of X.
  - Then X is a parameter, and the number of wins is a statistic.

- **In probability:**
  - We fix X = 244/495 (i.e. the odds you win in a game of craps); go to the casino and see how you fair.

- **In statistics:**
  - I just got back from the casino and I won 51/100 games. What can I say about the value of X?
  - This approach is what we do in empirical research. We collect some data and try to find the processes that generated it.

What are statistical models?

- A statistical model is a formal representation of the process by which a social system produces output.
- Equivalent notation (King 1998)

  - Standard version:

$$Y_i = x_i\beta + \epsilon_i \qquad = \text{systematic} + \text{stochastic}$$
$$\epsilon_i \sim f_N(e_i|0, \sigma^2)$$

  - Alternative version:

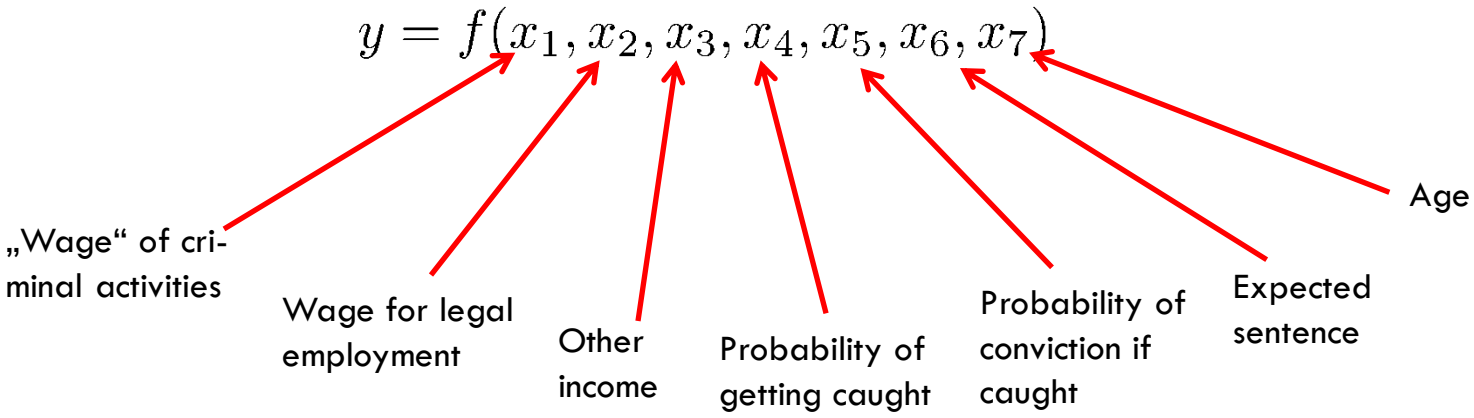
$$Y_i \sim f_N(y_i|\mu_i, \sigma^2) \qquad \text{stochastic}$$
$$\mu_i = x_i\beta \qquad \text{systematic}$$

Specifying statistical models

- Model specification has to deal with **two components** of our statistical model:
  - The functional form that relates measured variables to measured outcomes.
    - In other words, our belief in the relationship between the independent variables and the dependent variable.
    - Our goal is to identify the unknown model parameters.
      - One thing to keep in mind is that data do not produce a model…rather data reduce our uncertainty about the unknown model parameters.
  - Assumptions about the disturbances (i.e. the unmeasured variables influencing the observed outcomes – the error term).

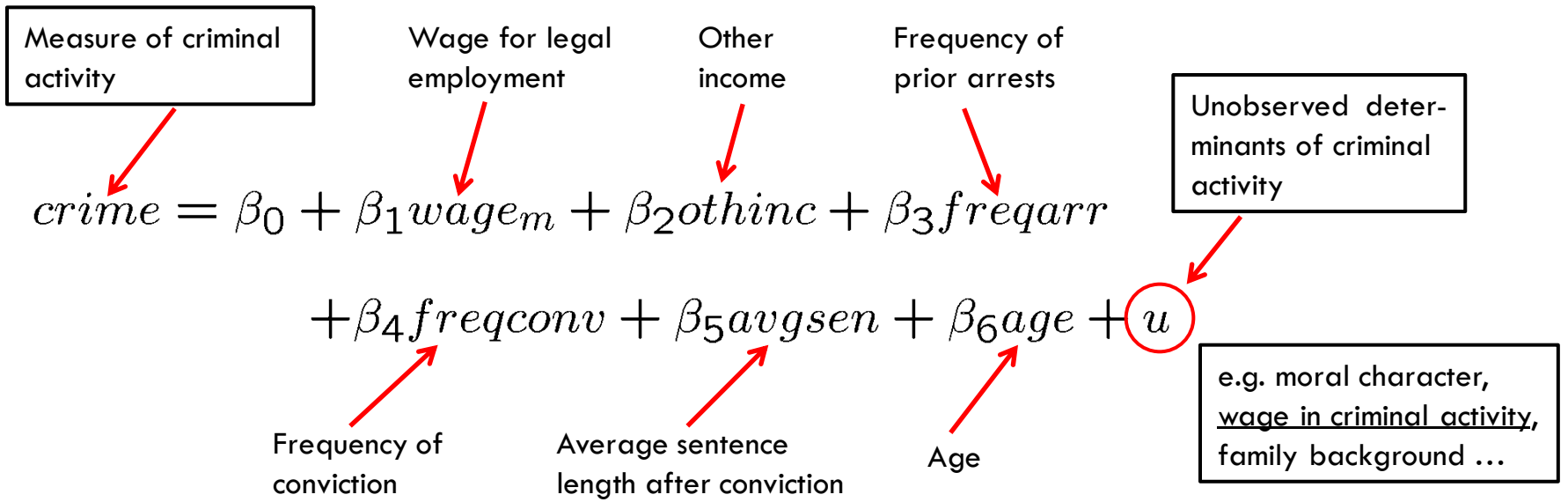
Model example

- **Economic model of crime (Becker (1968))**
  - Derives equation for criminal activity based on utility maximization



  - Functional form of relationship not specified
  - Should age have squared term? Should Criminal wages be interacted with probability of being caught?

- **Econometric model of criminal activity**
  - The functional form has to be specified
  - Variables may have to be approximated by other quantities (some theoretically relevant variables we cannot obseve)



Models vs Estimation Procedures

(Hayashi 2000)

- A model is a formal representation of the process by which a social system produces output.
- An estimation procedure is a data-based protocol for choosing from the model the parameters most likely to have generated our data.

- For most of this class, we will be working with a linear regression model and using OLS as our estimation procedure.
  - For instance, you collect data on home ownership rates in different cities along with a range of predictor variables. You can estimate a linear regression via OLS to understand the size and significance of the predictor variables relationship with home ownership.
- When we move to logistic regression, we utilize a different probability distribution, and a new estimation procedure – maximum likelihood.

Class Content and Overview

Texts

1. Wooldridge, J. M. (2016). *Introductory econometrics: a modern approach*: Cengage Learning. [***Required: Earlier editions are fine***]

2. Long, J.D. & Teetor, Paul (2019). *R cookbook*: O'Reilly [***Optional***]

Weekly readings and class structure

- The goal of this course is to provide you with the fundamental tools needed to analyze relationships and investigate phenomena of interest to you.
- Each week we will:
  - explore different aspects of linear models,
  - witness the use of those models in scholarly articles, and
  - implement those models in R

- I will post my lecture slides before each class. All slides, readings, data sets, and R codes will be available in the 'Weekly Lectures and Readings' tab on blackboard.

Homework

- **Problem sets** will be due several times throughout the semester. I expect to assign 4 sets.
- I encourage you to work with your classmates on these assignments, but ultimately complete and write-up the answers by yourself.

Data Analysis Paper

- We will be putting our econometric skills into action through a data analysis paper (approximately 3,000 to 5,000 words).
  - You will work in groups of 3 or 4 on this paper.
- You will not need to collect original data. You will rely on existing data sources.
  - There are many online datasets and repositories such as such as ICPSR, World Values Survey, IES, Census Data, Pew Internet and American Life Project, Current Population Survey, FedView survey, Harvard Dataverse, open data portals, etc...
  - Alternatively, you can respectfully ask a faculty member at this or another institution if they would be willing to share their data.

Data analysis paper cont…

- You will be required to develop testable hypotheses using methods covered in this course and to situate your hypotheses within the current literature or policy debate regarding your topic of interest.
- The aim of this paper is to give you a more complete experience of the data analysis process. You will go from initial data screening and cleaning to final model output and interpretation.
- For this assignment you will write your paper as if preparing it for submission to a journal or as a report to a relevant agency, nonprofit or think tank.
- **Please find a dataset early on in the semester. You can then use this data set to work through the methods we are covering in class. Having data that is meaningful to you is one way to overcome a general distaste for anything quantitative.**

Grading

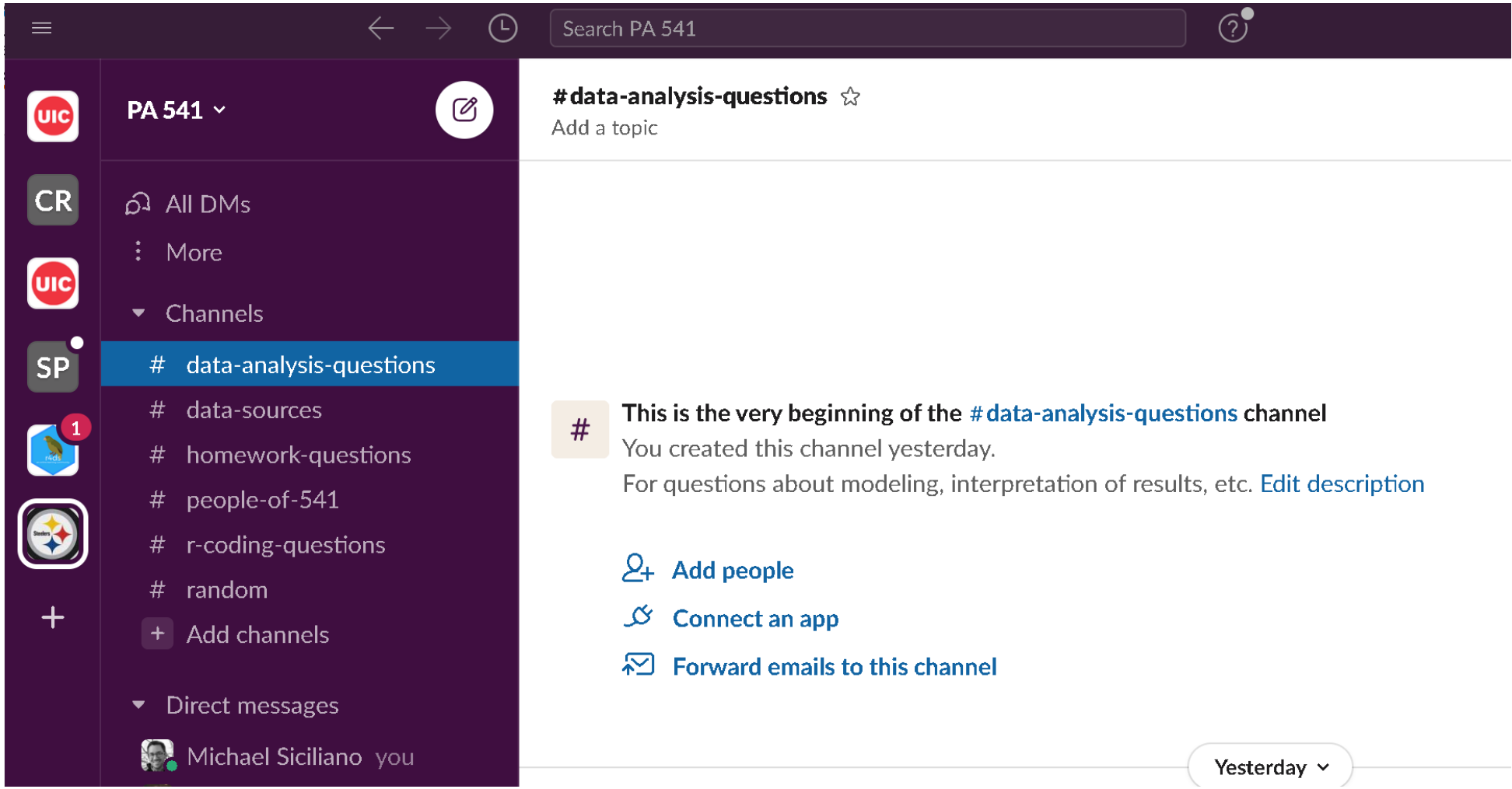
Updated Midterm Date as of 1/10

Your grade in this course will be calculated as follows:

Component	Percentage of overall grade	Due date
Midterm	25%	March 8
Final	25%	April 26
Problem Sets (4)	20%	Ongoing
Data Analysis Paper	20%	May 5
Class Participation	10%	Ongoing

Q&A and Communication Through Slack

slack



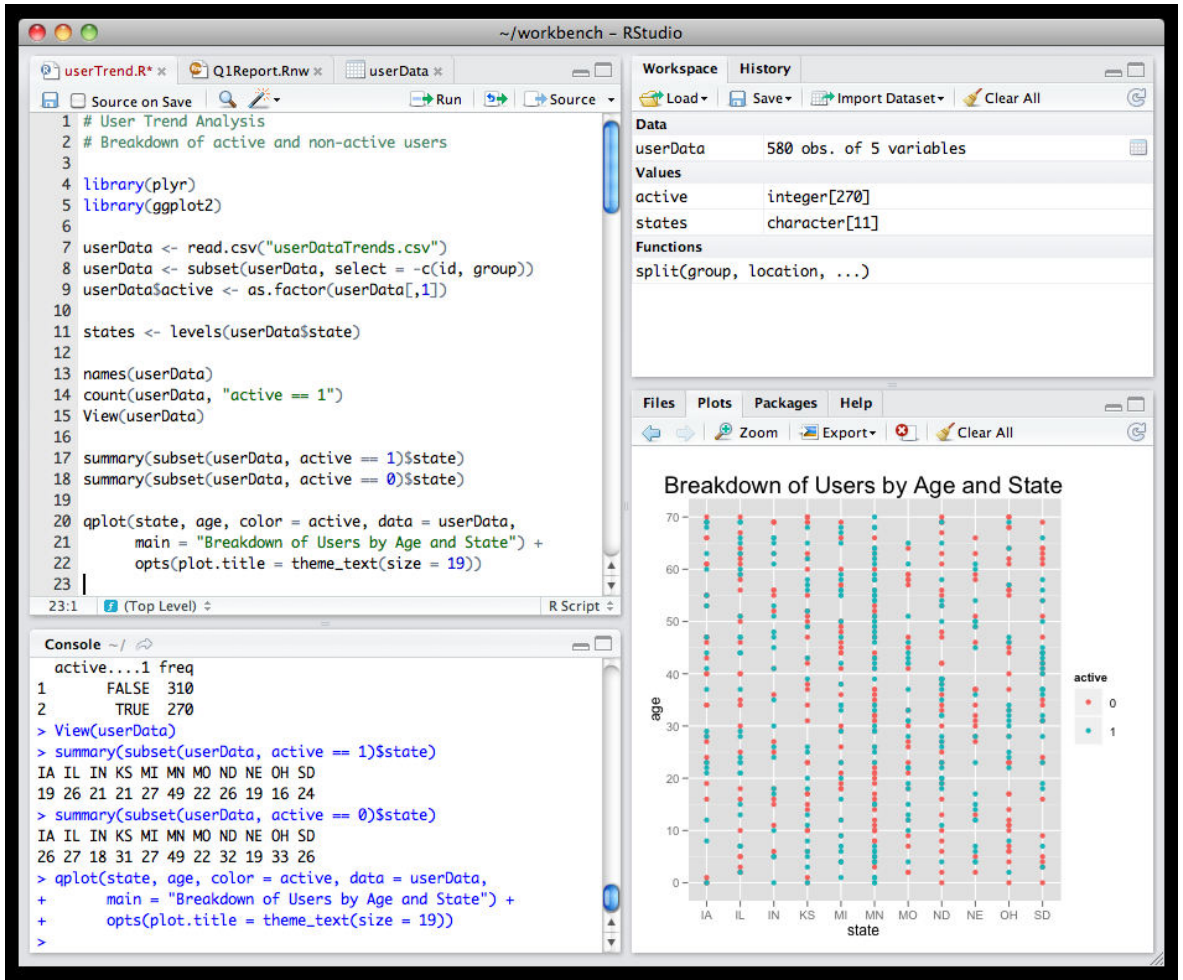
Why will be using R

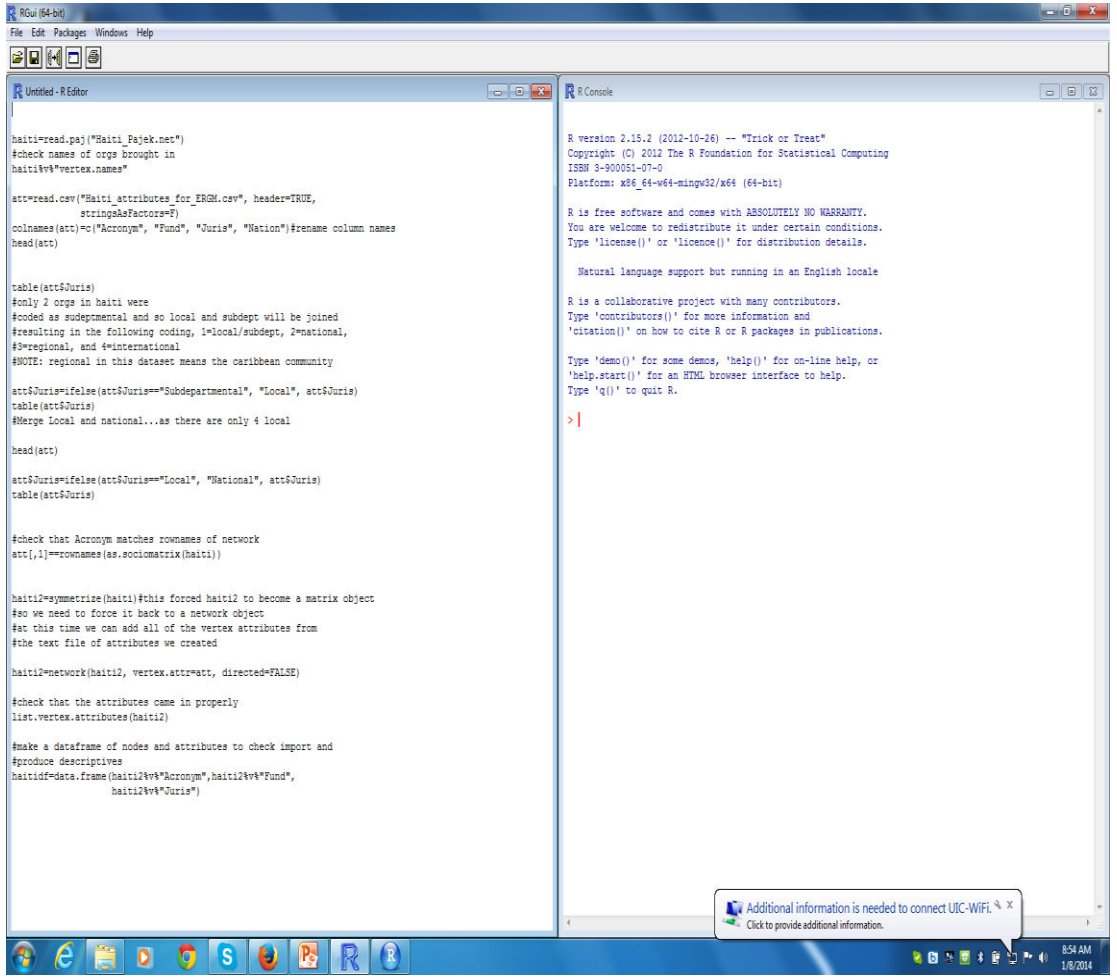
- Unrivalled coverage and availability of new, cutting-edge applications in the field (no need to wait for a new release of the software).
- Facilitate your understanding of the literature as more and more people are reporting their results in R.
- Quality of support available through listservs, Stack Exchange, etc…
- Ease with which you can write your own functions.
- The product is free!

- I have put some material in an R Resources folder on Blackboard.

R Studio

R Studio is an integrated development environment (IDE) for R.





Using Scripts

- Why use scripts?
  - Allows others to view and reproduce your work.
  - Allows you to go back to your work months later and see exactly what you did.
  - Facilitates small changes in your procedures without having to redo all of your other work.

  - I will provide you with my scripts for each class.

Optional Labs?

- Several times through the semester I or the TAs may hold optional labs to help students with homework, learning R, covering material in more detail, etc…
- These are *completely* optional, but have been really helpful to students in the past.
- I often hold them prior to the homework being due so you can ask questions about things you may be unsure about.

Questions on class structure, software, expectations, etc?

Intro to R and review of basic concepts

The R programming language

- R is a system for statistical computation and graphics.
- R uses object-oriented programming.
  - When you run a regression in SAS or SPSS the program dumps a bunch of output onto your screen.
  - In R, when you call the lm() regression function it returns an object containing all of the results. You can then pick and choose which aspects of the object you wan to extract (coefficients, standard errors, residuals, etc…).
  - This becomes extremely helpful as your data and analysis becomes more complex.
    - Ex. Analyzing data from several different sites and then wishing to compare the modeling results.

Modes of objects in R

- Modes are the data type of an object.
- Most basic modes for objects are:
  - Numeric
  - Logical
  - Character

Main object 'classes' in R

- The entities that R operates on are known as ***objects.***
  - Objects we will most commonly use are:
    - **Vectors –** Most basic data object in R. All elements of a vector must be of the same mode.
    - **Matrices –** Corresponds to the mathematical concept of the same name.
    - **Arrays** - Arrays are similar to matrices but can have more than two dimensions.
    - **Data frames** - A data frame is more general than a matrix, in that different columns can have different modes (numeric, character, factor, etc.). This is similar to SAS and SPSS datasets.
    - **Lists** - An ordered collection of objects (components). A list allows you to gather a variety of (possibly unrelated) objects under one name.
    - **Factors** - variables in R which take on a limited number of different values, i.e., categorical variables.

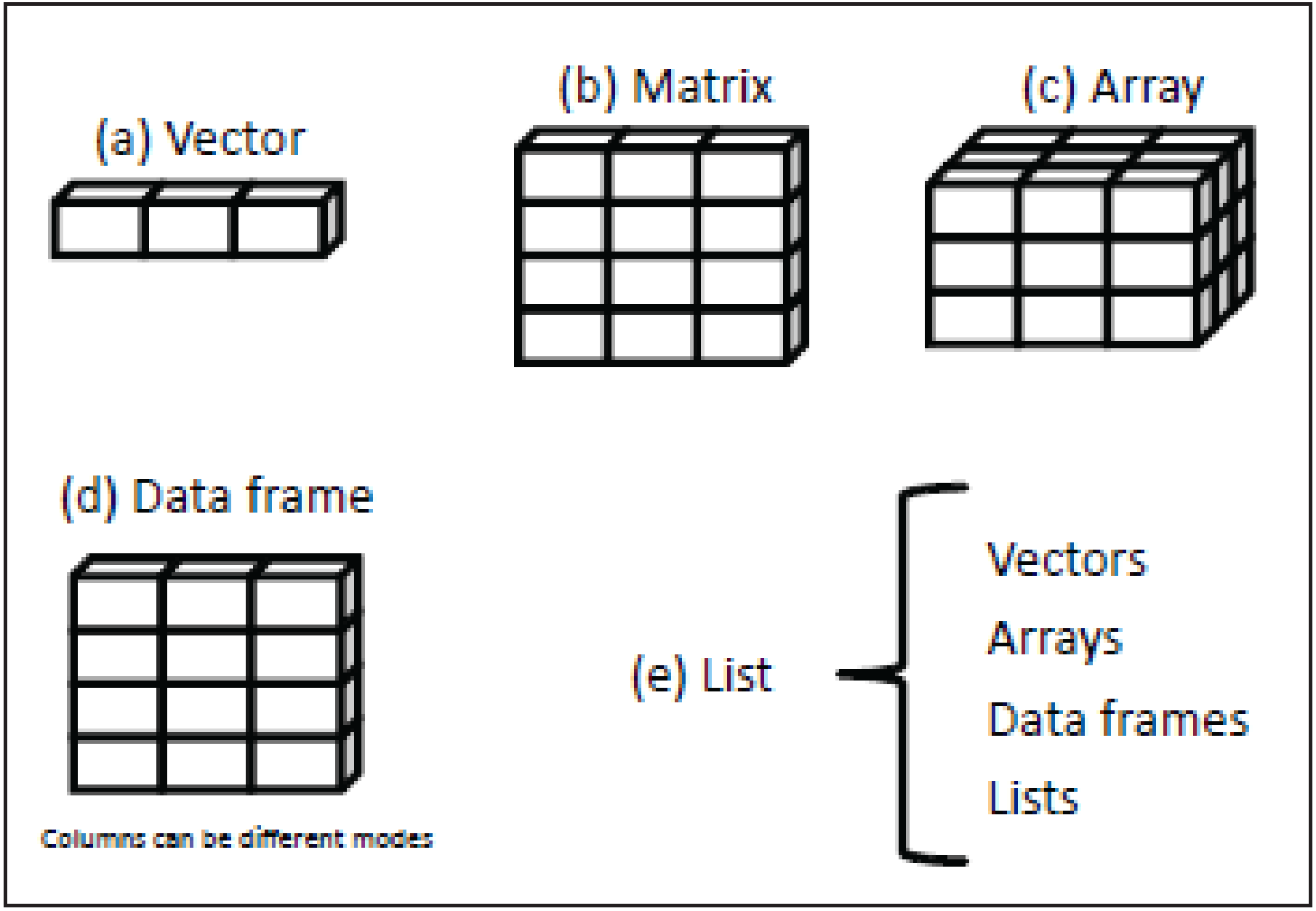
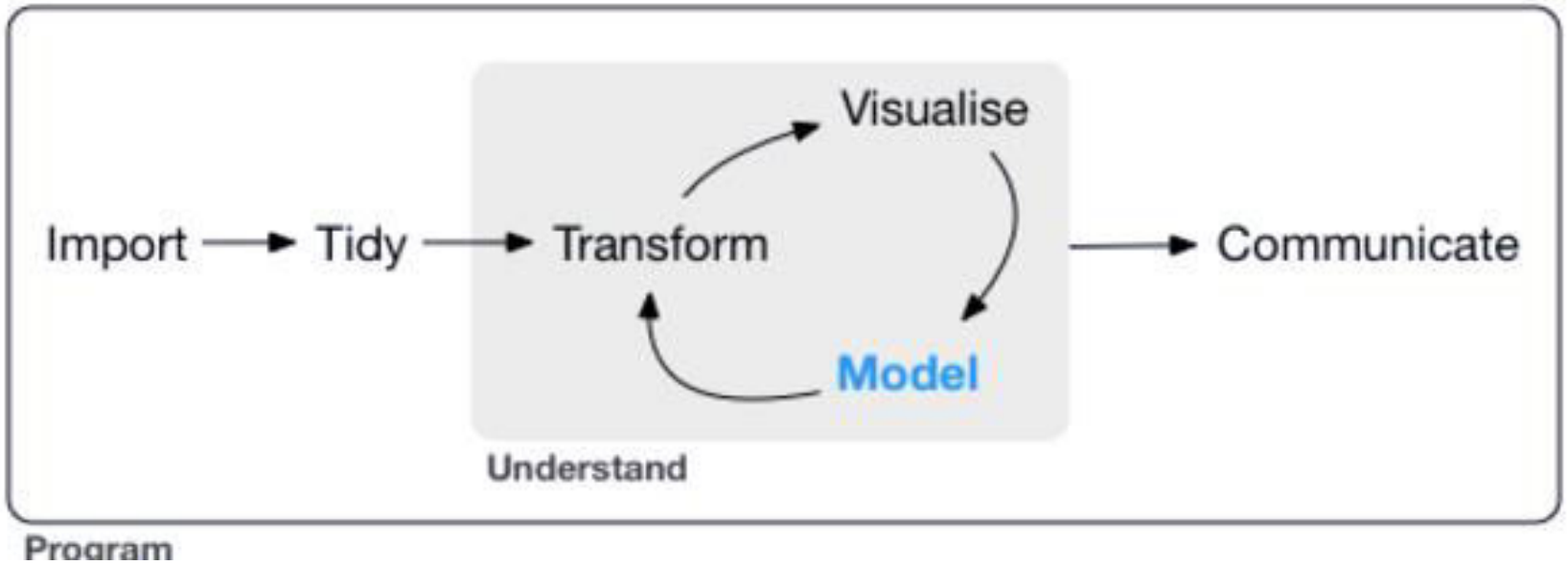


Figure 2.1 R data structures

Image from Kabacoff's R in Action (2011, p. 23)

Tidyverse


tidyverse
www.rstudio.com


Import
Tidy
Transform
Visualise
Model
Communicate
Understand
Program


dplyr
forcats
ggplot2
haven
lubridate
purrr
stringr
readxl
tibble
tidyr
tidyverse

- Switch to R Studio.
- The script I walk us through is posted on Blackboard and contains some sections for you to create code.

- To begin:
  - Save the R script on Blackboard into a local folder on your computer. I recommend starting a 541 folder, and then create subfolders for each week.
  - Open R Studio and then open the R script (file -> open file and then navigate to the appropriate folder.
  - We will walk through the script together.

In-class practice

- **See blackboard for script and data.**
- **Will work with the *Midwest* dataset and practice reading, examining, and wrangling data.**

Review of Statistical Concepts

For your own review; will not cover in class.

Sample characteristics

- In many instances, we will not know the true expected value or variance of a distribution of interest.
- For example, we may want to know about the degree of income inequality in society, and therefore want to know the expected value and variance of income distribution in the population.
  - If we took a census of all members of the intended population we could find the true values for the characteristics of the distribution.
  - However, such a process is extremely costly in terms of time and money.

- Instead, we often take a sample of the population and use the sample to generate our best guess as to what the true characteristics of the population distribution are.
  - Much of the data social scientists use come from sample surveys.
- A **random sample** is *n* observations that are drawn independently from the same population.

- Suppose we are interested in understanding the recidivism rate among inmates in the United States.
  - We would probably want to know the average rate and how much variation there is around that average.

Sample mean

- Let's say we go out and collect the following data on inmate recidivism from 5 randomly chosen inmates: 1, 2, 3, 3, 4
- Sample mean $= \bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} = (1+2+3+3+4)/5 = 2.6$
- We can think of the sample mean as an *estimator* of the true population mean, μ_x.

- Thus, estimators are a "guess" of the true value of a parameter of interest.

Sum of Squared Errors and Variance

- Sum of Squared Errors = $\sum (x_i - \bar{x})(x_i - \bar{x})$

  $= (-1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2 = 5.20$

  However, this value is dependent on the number of observations…

- Variance = $s^2 = \frac{\sum (x_i - \bar{x})^2}{N - 1}$

- Variance $= 5.20/4 = 1.3$

- In this example we can now say that the average error in our model is 1.3 squared. It makes little sense to talk about squared recidivism rates and so we often take the square root of the variance which is the standard deviation.

Standard Deviation

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N - 1}}$$

- Our standard deviation is the square root of 1.3 which is 1.14.
- Now our measure of average error is in the same units as the original measure.

Standard Deviation versus Standard Error

- When we take a sample, it is only one of many possible samples of our intended population.
- It is important to know how well a particular sample represents the overall population. To do so we need to consider the standard error.
- The standard error (s.e.) is based on a sampling distribution.
- Sampling Distribution – the frequency distribution of sample means from the same population.
  - Assume we gathered 100 different samples of size 50. Each of these samples will have a slightly different mean.

Standard Error

- Calculating the standard deviation of our sampling distribution would give us a measure of how much variability there was between the means of the different samples.
- The standard deviation of this sampling distribution is the standard error.
- In reality, we cannot (or should not) collect hundreds of samples to produce a sampling distribution. Therefore, we must rely on an estimate of the standard deviation of that distribution.
- The approximation is calculated as follows

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Probability distributions

- Discrete distributions – for variables that can take on only a finite number of values.
  - Many useful discrete distributions for statistics such as the binomial distribution and the Poisson distribution.
    - We will use these toward the end of the semester.
  - Example, the number of heads in two coin tosses is a discrete distribution as the only possible values are 0,1, or 2.

Probability distributions

- Continuous distributions
  - The normal distribution is a continuous probability distribution that has the well known bell-shaped curve.
  - If a random variable X, comes from a normal probability distribution, we write $X \sim N(\mu_x, \sigma_x^2)$
  - Many real world variables appear to follow a normal distribution, i.e., height.

Let's produce a probability distribution

Assume we flip two fair coins. What is the sample space of all possible outcomes?

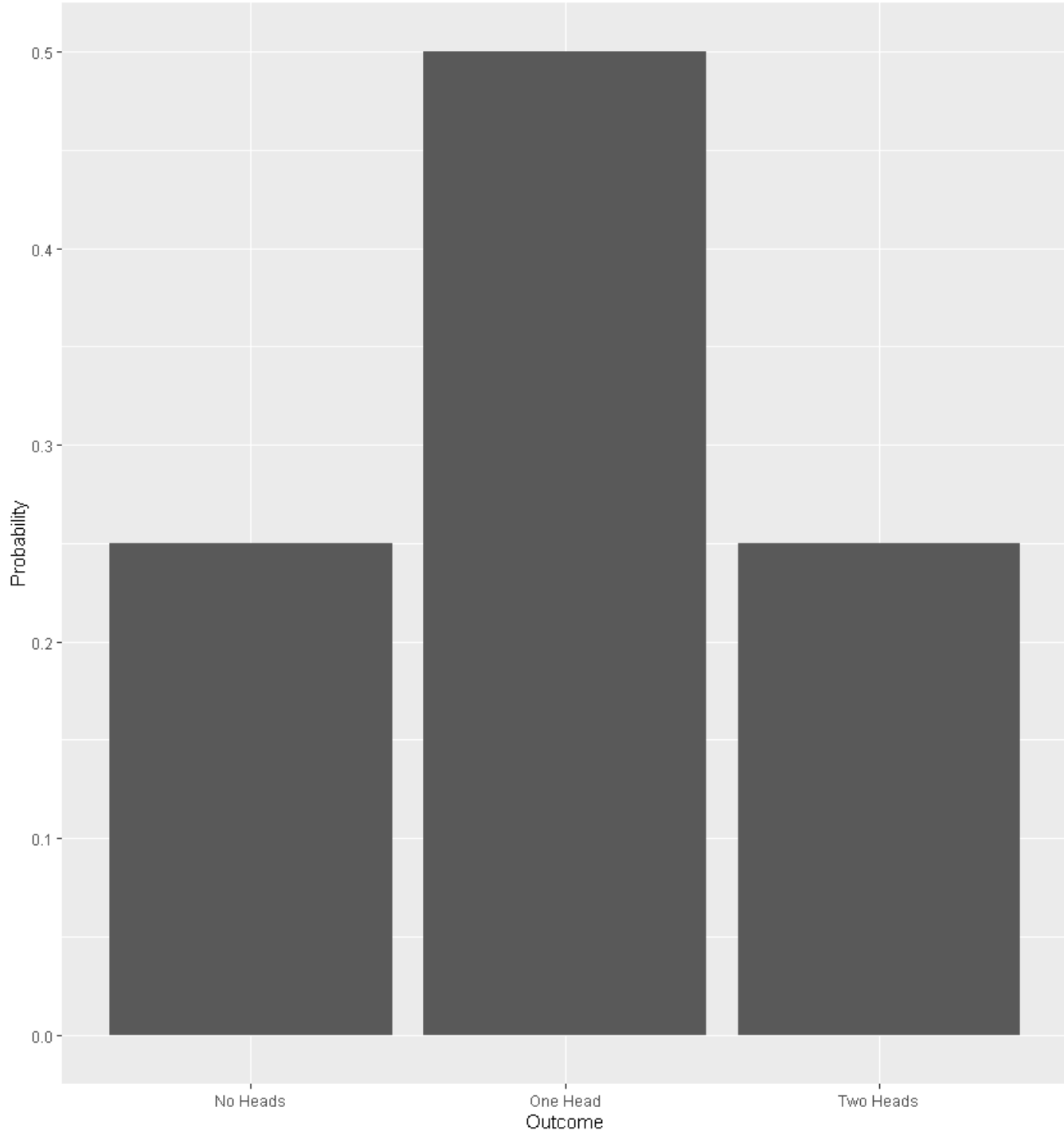
# of Heads	Sample Points	Probability

Let's produce a probability distribution

Assume we flip two fair coins. What is the sample space of all possible outcomes?

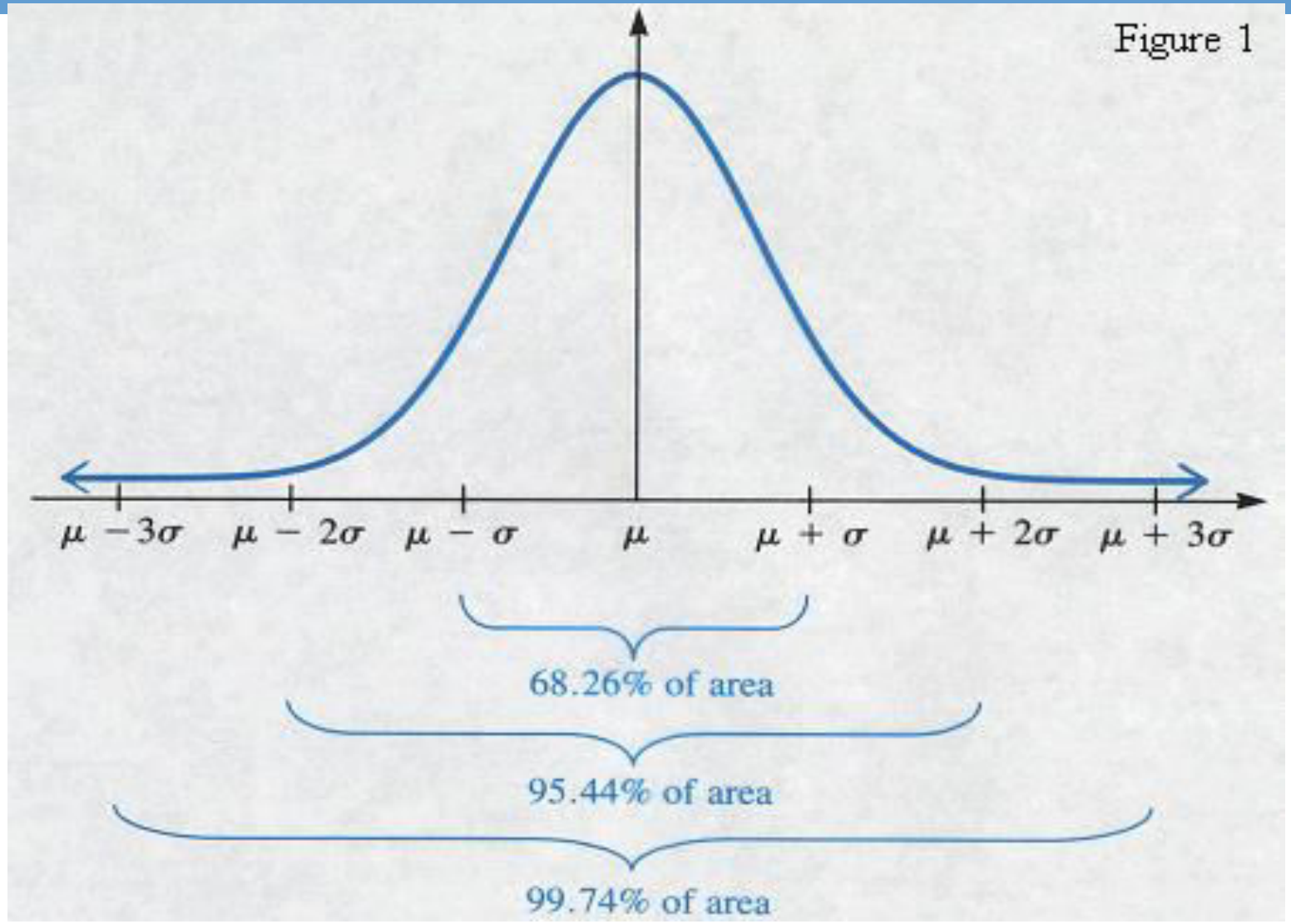
# of Heads	Sample Points	Probability
0	TT	¼
1	HT, TH	¼ + ¼ = ½
2	HH	¼

- This description of the probabilities for each event is called a Probability Distribution Function (PDF).
- The PDF shows the probability that a random variable will take on each outcome in a sample space.
- Let's draw the PDF.


0.5
0.4
0.3
0.2
0.1
0.0
Probability
No Heads
One Head
Two Heads
Outcome

What about for continuous variables?

Normal Distribution



Z-Scores

- To covert any normal distribution into a standard normal distribution with a mean of zero and standard deviation of 1, we simply need to:
  - Center the data at zero by subtracting the mean from all scores
  - Divide the resulting score by the standard deviation to ensure that the data have a standard deviation of 1.

$$z = \frac{X - \bar{X}}{s}$$

Example of z-score

- Assume the recidivism rate of a given population is normally distributed with mean of 2.4 and a standard deviation of 3.1.
- Thus, X ~ N(2.4, 3.1)
- What is the probability that there is a particular individual who has returned to jail 9 times. Based on the previous equation:

$$z = \frac{9 - 2.4}{3.1} = 2.12$$

- So our z-value is 2.12. We can use any z table to look up this value and find the probability of finding a value greater than this. In other words, what is the likelihood that a randomly sampled individual recidivated 9 times or more.

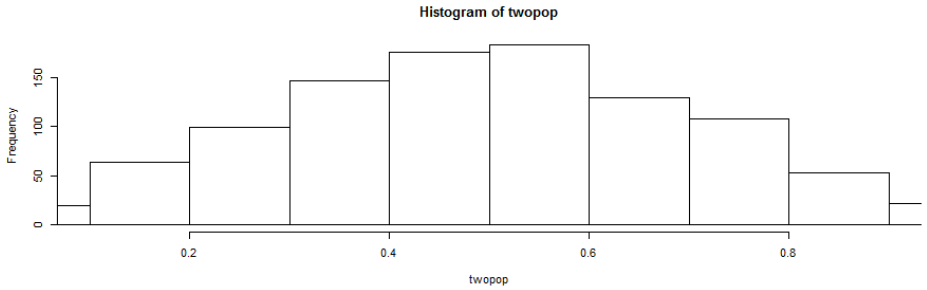
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

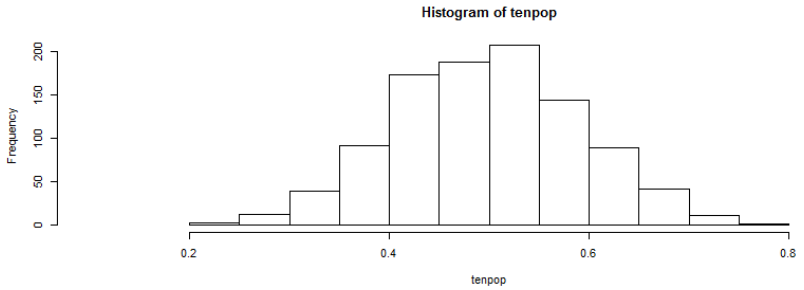
Looking at a table of z-scores (which indicate the area to the left of our z-value), we can see that at 2.12, 98.3 percent of all scores lie to the left of our value. And so there is only a 1.7% change of finding someone who has recidivated 9 times or more.

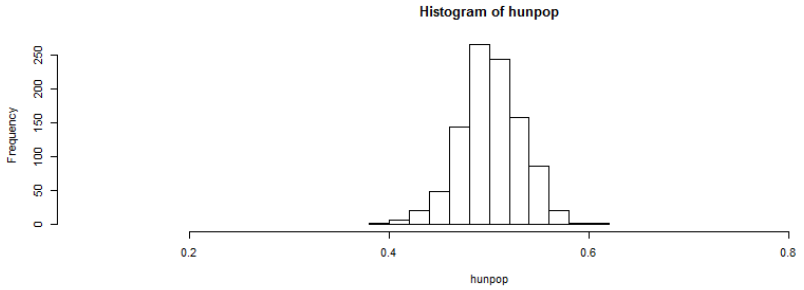
Central Limit Theorem

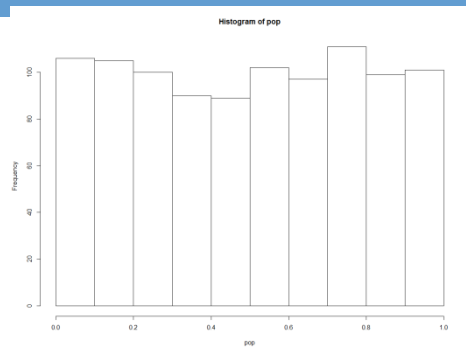
- Regardless of the distribution of the original variable *x* from which the statistic was generated, the asymptotic sampling distribution of the statistic will be normal.
- Again,
  - Consider that when we take a random sample of size *n*, this sample is only one of many possible samples of size n that could be drawn from the population.
  - Obviously, if we took a second sample of size *n*, we would not end up with the same respondents and so any statistic calculated (like the mean) would be different from the first sample.
  - The CLT, however, says that if we were to take all possible random samples of a given size from the population and compute the value of the statistic of interest for each one, the distribution of these statistics – the sampling distribution – would be normal (assuming the sample size is large enough).
    - The standard deviation of this sampling distribution is our standard error.

Demonstrating the CLT










```
pop=runif(1000, 0,1)
hist(pop)

twopop=vector()
for (i in 1:1000) twopop[i]= mean(runif(2,
0,1))

tenpop=vector()
for (i in 1:1000) tenpop[i]= mean(runif(10,
0,1))

hunpop=vector()
for (i in 1:1000) hunpop[i]= mean(runif(100,
0,1))

par(mfrow=c(3,1))
hist(twopop, xlim=range(.1,.9))
hist(tenpop, xlim=range(.1,.9))
hist(hunpop, xlim=range(.1,.9))
```

From 1 variable to 2

Covariance and Correlation

- A correlation is a measure of a linear relationship between two variables.
- Two measures are particularly important for statistically expressing the relationship between two variables:
  - Covariance
  - Correlation Coefficient

Covariance

- Recall our formula for variance: $s^2 = \frac{\sum(x_i - \bar{x})^2}{N - 1}$
- Which is the equivalent of: $s^2 = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{N - 1}$
- Covariance is a measure of the average relationship between <u>two</u> variables. It is the average cross-product deviation.

Covariance Cont…

- Covariance = cov(x,y) = $\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N-1)}$

- When we multiply the differences in one variable by the corresponding differences in the other we produce what is known as the cross-product deviations. We then divide this value by N-1 to produce the average sum of differences, known as the covariance.

Example of Covariance

Country	1	2	3	4	5	Mean	s
GNP	5	4	4	6	8	5.4	1.67
Crime	8	9	10	13	15	11.0	2.92

- cov(x,y) = $\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N-1)}$
- = [(-0.4)(-3)+ (-1.4)(-2) + (-1.4)(-1) + (-0.6)(2) + (2.6)(4)] / 4 = 17/4 = 4.25
- A positive covariance indicates that when one variable deviates from the mean the other variable does so in the same direction.
- One Problem: Covariance is dependent on the unit of measurement and does not allow for objective comparison across datasets.
  - For instance, if I multiple GNP by 10, then the covariance would also increase by a factor of 10.

Standardizing the Covariance

- To overcome the issue of dependence on measurement scale we often convert the covariance into a standard set of units.
- In the same manner as calculating z-scores, we standardize covariances by dividing by the standard deviation.
- We have actually have two standard deviations (because we have two variables whose relationship we are interested in).

Pearson Correlation Coefficient

- The standardized covariance is known as a correlation coefficient.

- Pearson Correlation Coefficient:

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y} = \frac{cov_{xy}}{s_x s_y} = r$$

- By standardizing the covariance we force our value to be between -1 and +1. A measure of $\pm$ 1 indicates perfect correlation between two variables.

Correlation as a measure of effect size

- Correlation coefficient can be uses as a standard measure of effect size
  - ± .1 is a small effect
  - ± .3 is a medium effect
  - ± .5 is a large effect

Type I and Type II Errors

- The American Court System is designed to limit Type I errors; the burden of proof rests with the plaintiff and thus the null hypothesis is that one is innocent until proven guilty.
- Type I error (false positive) – we reject the null when in fact it was true. We reject the null of innocence and find someone guilty when in reality they committed no crime. So we place an innocent man in jail.
- Type II error (false negative) – we accept the null when it should have been rejected. We find someone innocent when in reality they were guilty of the crime.

- In more statistical terms:
  - **Type I Error**: we find a significant relationship between two variables or a difference between two means when in reality there is no relationship or difference.
  - **Type II Error**: We find no relationship between two variables or no difference between two means when in reality there was a relationship or difference.

Confidence Intervals

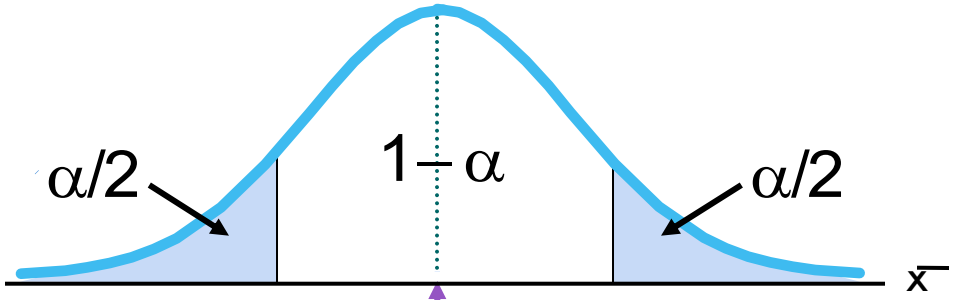
Estimation

- There are two general methods of communicating results of an analysis:
  - Point estimation: guess a single value for the true value (e.g., we use the sample mean to estimate the true population mean).
  - Interval estimation: use a sampling distribution to guess an interval for the true value.
    - As we discussed earlier, a sample mean is usually not equal to the population mean; generally there is some sampling error.
    - Thus, we often want to accompany any point estimate with information that indicates the accuracy of that estimate.
      - These intervals are known as confidence intervals.

Confidence Level, $(1-\alpha)$

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = 0.95$
- A relative frequency interpretation:
  - From repeated samples, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter.
- Any specific interval either will contain or will not contain the true parameter
  - No probability involved in a specific interval

Sampling Distribution of the Mean

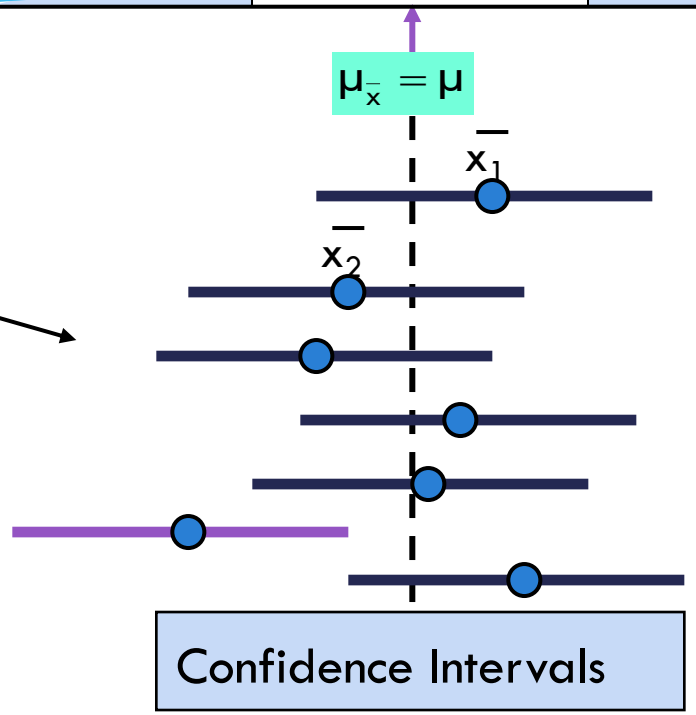


Intervals extend from

$$\bar{x} - z\frac{\sigma}{\sqrt{n}}$$

to

$$\bar{x} + z\frac{\sigma}{\sqrt{n}}$$



100(1-α)% of intervals constructed contain μ;

100(α)% do not.

Confidence Intervals

General Formula

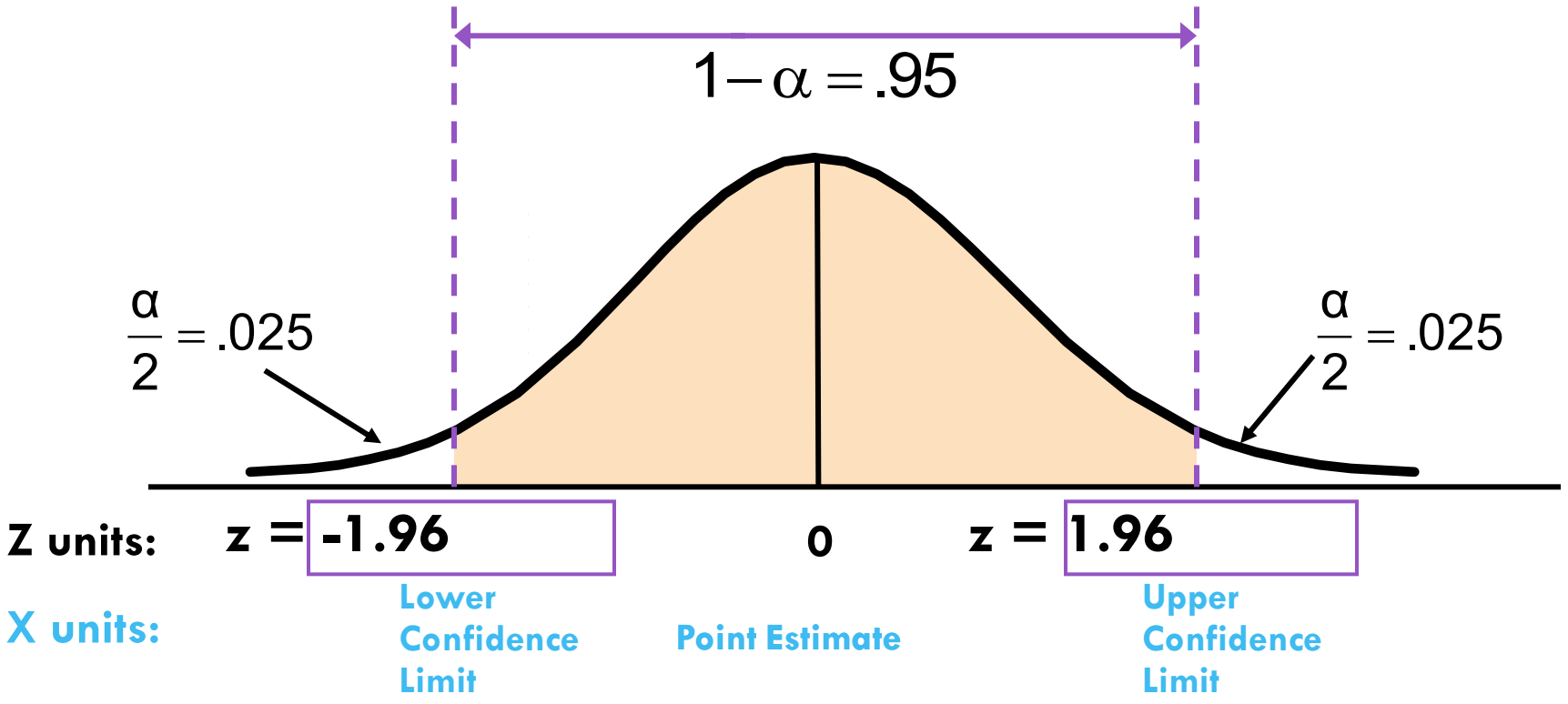
The general formula for all confidence intervals is:

Point Estimate ± (Reliability Factor)(Standard Error)

- The value of the reliability factor depends on the desired level of confidence

Finding the Reliability Factor, $z_{\alpha/2}$

- Consider a 95% confidence interval:



Common Levels of Confidence

Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	Confidence Coefficient, $1-\alpha$	$Z_{\alpha/2}$ value
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27

Example

- A sample of 11 adult inmates from a large population has a mean recidivism rate of 2.20 times. Assume we know from past surveys that the population standard deviation is 0.35.

- Determine a 95% confidence interval for the true mean recidivism rate of the population.

- A sample of 11 adult inmates from a large normal population has a mean recidivism rate of 2.20 times. Assume we know from past surveys that the population standard deviation is 0.35.
- Solution:

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96\,(.35/\sqrt{11})$$

$$= 2.20 \pm .2068$$

$$1.9932 < \mu < 2.4068$$

Interpretation

- We are 95% confident that the true mean recidivism rate is between 1.9932 and 2.4068 times per inmate.
  - More appropriately, although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean.

Confidence Interval for μ (σ^2 Unknown)

- If the population standard deviation σ is unknown, we can substitute the sample standard deviation, s
- This introduces extra uncertainty, since s is variable from sample to sample
- So we use the t distribution instead of the normal distribution

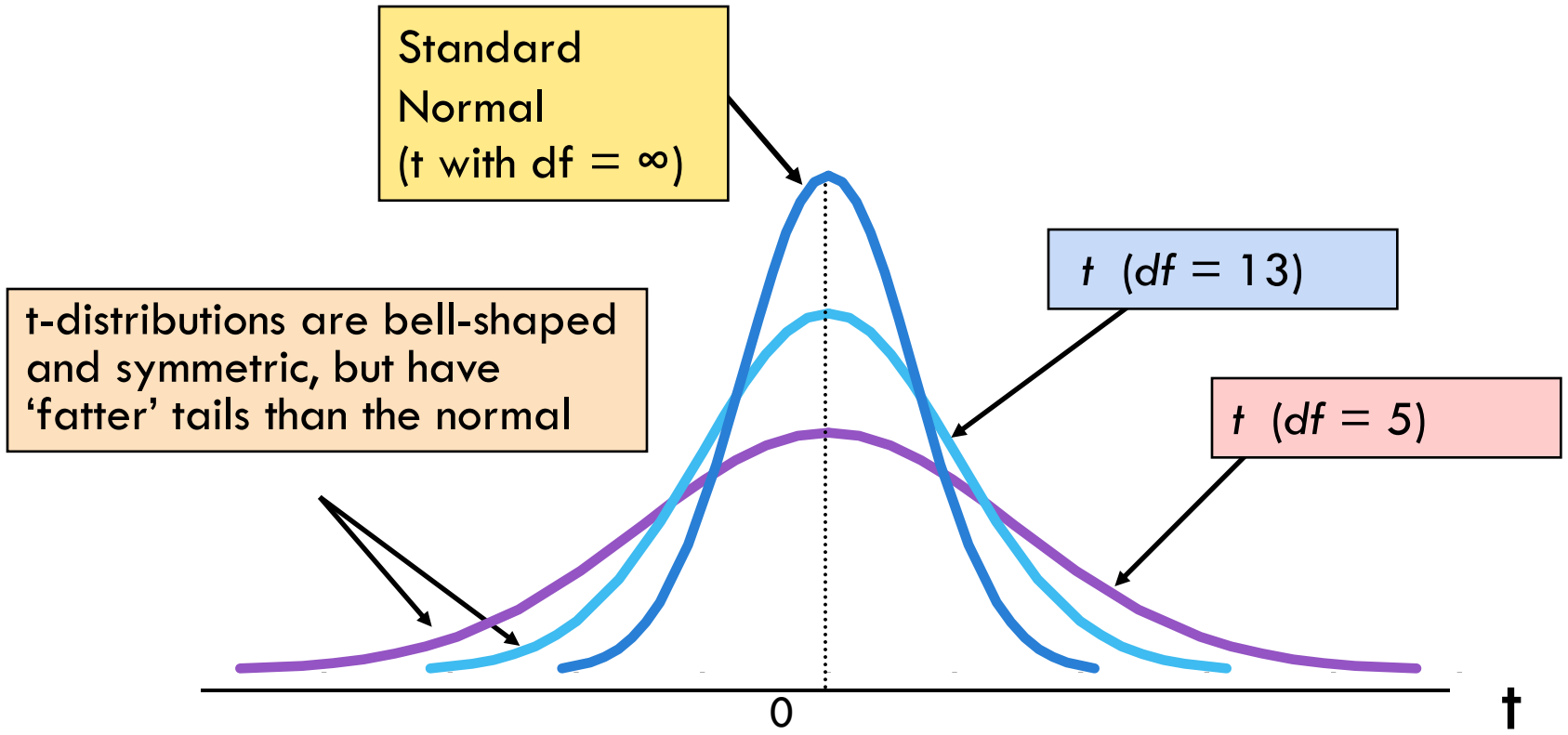
Student's t Distribution

- The t is a family of distributions
- The t value depends on degrees of freedom (d.f.)
  - Number of observations that are free to vary after sample mean has been calculated

$$d.f. = n - 1$$

Student's t Distribution

Note: t ⟶ Z as n increases

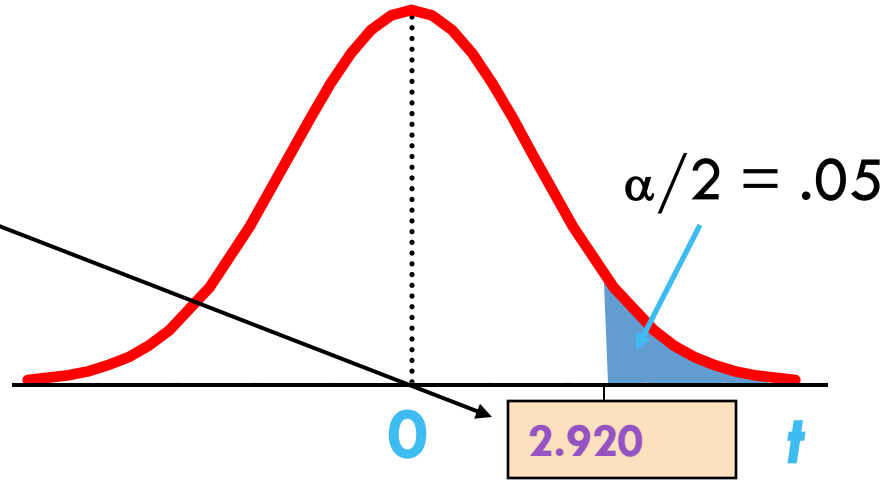


Student's t Table

	Upper Tail Area		
df	.10	**.05**	.025
1	3.078	6.314	12.706
2	1.886	**2.920**	4.303
3	1.638	2.353	3.182

The body of the table contains t values, not probabilities

Let: n = 3
df = n - 1 = 2
$\alpha = .10$
$\alpha/2 = .05$



t distribution values

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	Z
.80	1.372	1.325	1.310	1.282
.90	1.812	1.725	1.697	1.645
.95	2.228	2.086	2.042	1.960
.99	3.169	2.845	2.750	2.576

Note: t $\longrightarrow$ Z as n increases

Example

A random sample of n = 25 has $\bar{x} = 50$ and s = 8. Form a 95% confidence interval for μ

- d.f. = n – 1 = 24, so $t_{n-1,\alpha/2} = t_{24,.025} = 2.0639$

The confidence interval is

$$\bar{x} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

$$50 - (2.0639)\frac{8}{\sqrt{25}} < \mu < 50 + (2.0639)\frac{8}{\sqrt{25}}$$

$$46.698 < \mu < 53.302$$