ECMT 461 Lecture Slides for Exam 1 Material Spring 2024

Craig Schulman

The following slide decks cover material that will be included on Exam 1.

I reserve the right to amend/expand this material as needed.

WELCOME TO ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

A copy of the Course Syllabus is available in Howdy or at the course website (linked in Canvas)

http://people.tamu.edu/~cschulman/

1

HOWDY!

2

AGENDA FOR TODAY

- Syllabus
- Overview of the Course

3

SYLLABUS

- Please read the Syllabus carefully
- Copy available in Howdy
- At the class website: http://people.tamu.edu/~cschulman/
- And linked on the class Canvas page

4

THE CODE

- ***"An Aggie does not lie, cheat, or steal or tolerate those who do."***

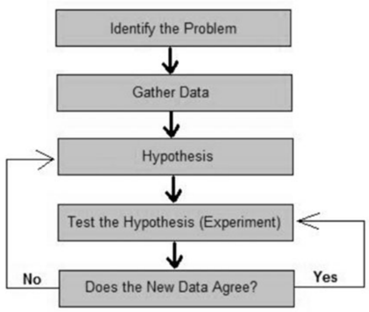
5

COURSE OVERVIEW

- Many Natural phenomena are inherently random
  - The roll of a pair of six-sided dice
- Other observed phenomena include, at least what we perceive to be, some degree of randomness. Two observationally equivalent states:
  - The phenomena in question have some inherent degree of randomness
  - The "data generating process" that results in what we observe is too complex to fully express. This results in what we observe as randomness.

6

THE SCIENTIFIC METHOD



Scientific Method: The scientific method is used in economics to study data, observe patterns, and predict results.

7

ENTER STATISTICAL ANALYSIS

- Statistical analysis allows us to make statements in probability regarding the size, relative position, and degree of association among variables that we perceive contain some degree of randomness.

8

SOME QUOTES ABOUT STATISTICS

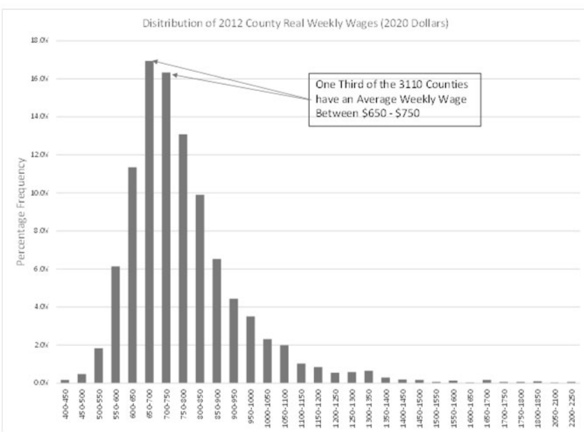
- "While it is easy to lie with statistics, it is even easier to lie without them." Frederick Mosteller
- "In God we trust. All others must bring data."
- "To understand God's thoughts, we must study statistics …" Florence Nightingale
- "Essentially, all models are wrong, but some are useful." George E. P. Box
- "…the statistician knows…that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world." George Box

9

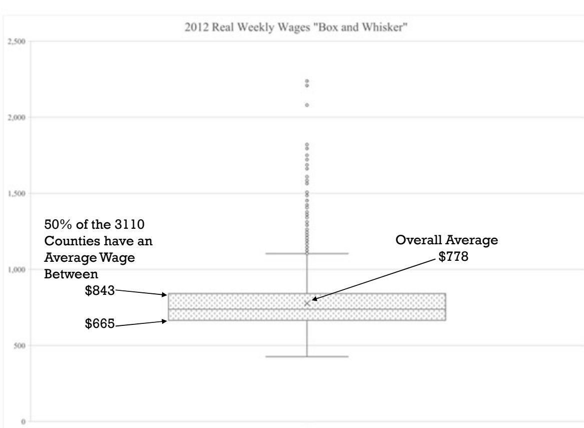
AN EXAMPLE

- 2012 Average Real Weekly Wages (measured in 2020 Dollars) among U.S. Counties
- 3110 Counties (excludes Alaska)
- Overall Average: $778
- Minimum: $426 (Keweenaw County, Michigan)
- Maximum: $2,238 (San Mateo County, California)

10



11

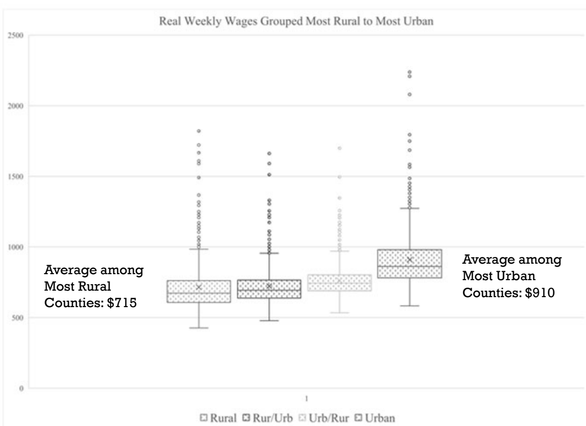


12

GROUPING COUNTIES INTO CATEGORIES

- Suppose we grouped the 3110 counties into 4 unique sub-samples based on County Population, from Most Rural, to Most Urban
- Do you expect Average Wages to be different among these sub-groups?
- How?

13

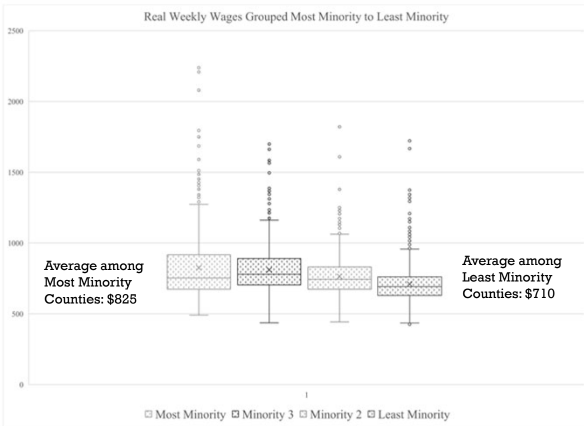


14

GROUPING COUNTIES INTO CATEGORIES

- Instead of grouping by County Population, suppose we group counties by the proportion of the population that is "minority" from Most Minority to Least Minority
- Do you expect Average Wages to be different among these sub-groups?
- How?

15

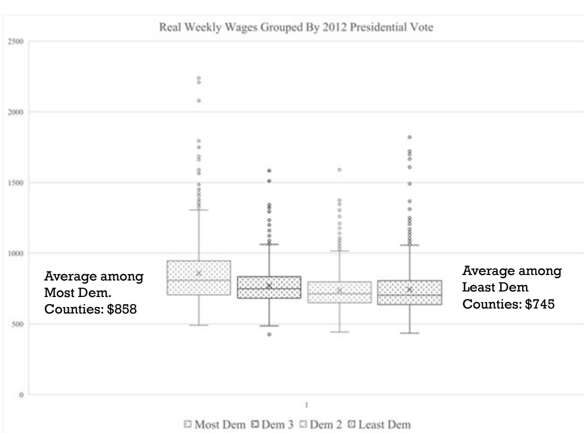


16

GROUPING COUNTIES INTO CATEGORIES

- 2012 was a Presidential Election Year
- Suppose we group counties by the proportion of the total county vote that voted for the Democratic Candidate, from Most Democratic Vote to Least Democratic
- Do you expect Average Wages to be different among these sub-groups?
- How?

17



18

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture #1

1

DATA

- Data: Information, especially information organized for analysis or used as the basis for decision making.
  - *Webster's II New Riverside University Dictionary*, Houghton Mifflin Company, 1984.
- Data in lists – most commonly in two-dimensional lists
  - Variables in columns, observations in rows, or
  - Variables rows, observations in columns

2

RANDOM VARIABLES, POPULATIONS AND RANDOM SAMPLES

- A ***random variable*** takes on different values with certain probabilities
- A ***population*** is the collection of all possible outcomes for that variable
- A ***sample*** is a subset of observations of a variable taken from the population
- A ***random sample*** is sample is such that
  - each observation in the population is equally likely of being selected, the selection of any one observation does not influence the selection of any other
  - every possible sample of a given size is equally likely of being selected

3

PARAMETERS VS. STATISTICS

- A ***parameter*** is a numerical measure that describes some aspect of a population – a population mean or variance, for example
  - Population parameters are usually unknown
- A ***statistic*** or ***sample statistic*** is a numerical measure that describes some aspect of a sample
  - Sample statistics are calculated from observations in a sample – if you have a sample of 10 different test scores, the sample mean (average) is calculated by adding all the scores (summing) and dividing by the sample size, 10.

4

SCALES OF MEASUREMENT – DATA TYPES

- Categorical variables
  - Identify mutually exclusive groups (subsets) of a population or sample
- Numerical variables
  - Ordinal Scales – information on rankings
  - Interval Scales – information on rank and difference from an arbitrary zero point
  - Ratio Scales – information on the rank and distance from a natural zero point

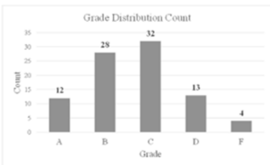
5

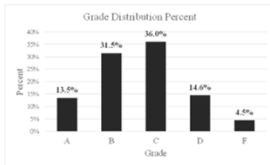
GRAPHS TO DESCRIBE DATA

Grade	Students	Percent
A	12	13.5%
B	28	31.5%
C	32	36.0%
D	13	14.6%
F	4	4.5%
Totals	89	100.0%

6

BAR CHARTS



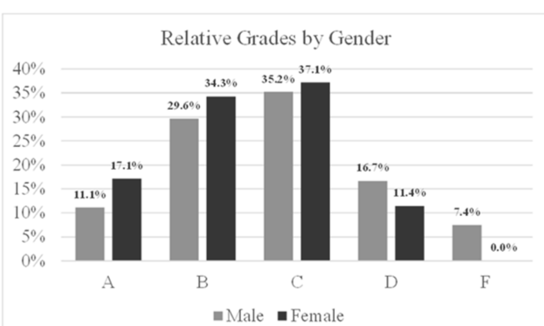


7

CONTINGENCY TABLES/CHARTS

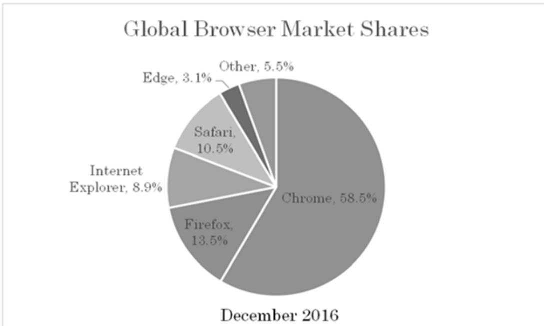
Grade	Male	Female	Total
A	6	6	12
B	16	12	28
C	19	13	32
D	9	4	13
F	4	0	4
Totals	54	35	89

8



9

PIE CHARTS



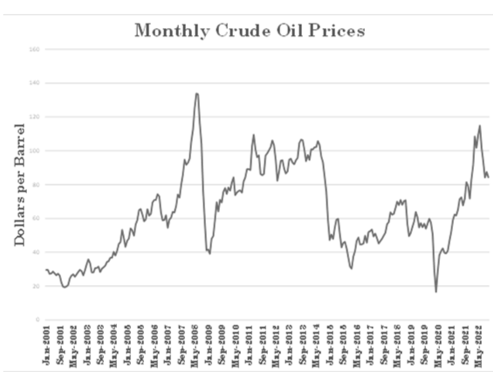
10

CROSS-SECTIONAL VS TIME-SERIES DATA

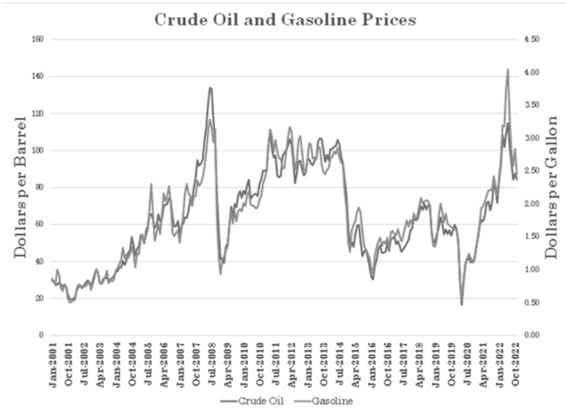
- Cross-Sectional data has no natural ordering either within a population or in any sample taken from a population
  - Most conducive to the concept of random samples
  - This type of data will be our main focus for most of the course
  - You will be dealing with cross-sectional data for your term projects
- Time-Series data is naturally ordered in time
  - Observed U.S. GDP in the 3rd quarter of 2019 naturally follows U.S. GDP in the 2nd quarter of 2019
  - Time-series data introduce a number of issues related to measurement and inference
  - We will briefly introduce time-series data at the end of the course.

11

GRAPHS FOR TIME-SERIES VARIABLES

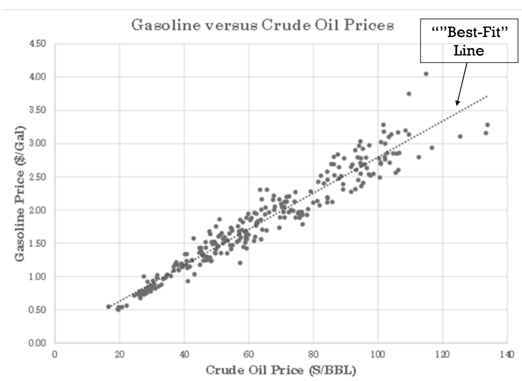


12



13

X-Y GRAPHS AND CO-MOVEMENT



14

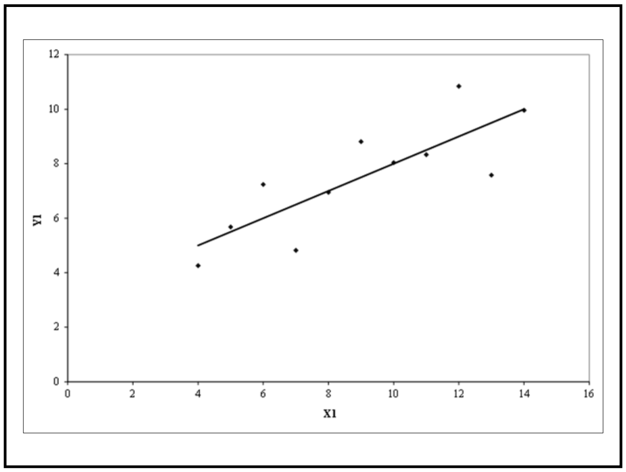
IMPORTANCE OF “LOOKING” AT YOUR DATA

- Four X-Y pairs with the same “Best-Fit” line and measured correlation

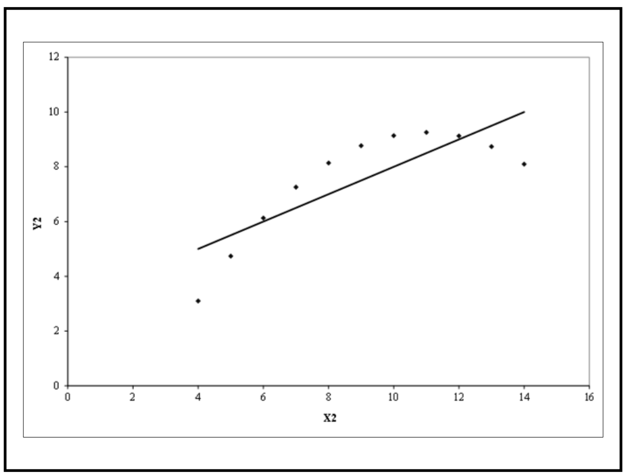
$$Y = 3.0 + 0.5X$$

$$Correlation = 0.82$$

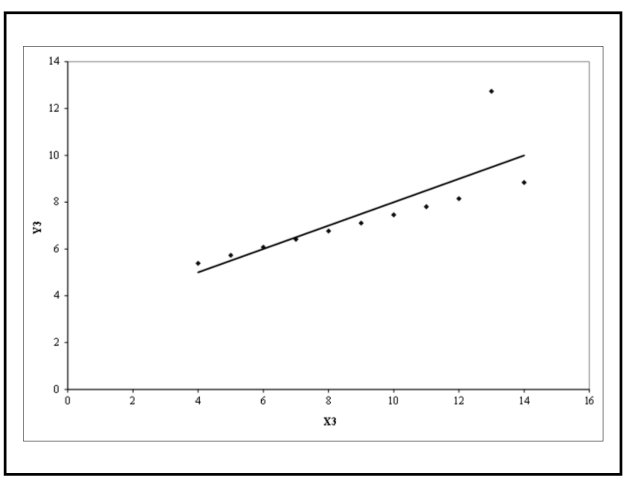
15



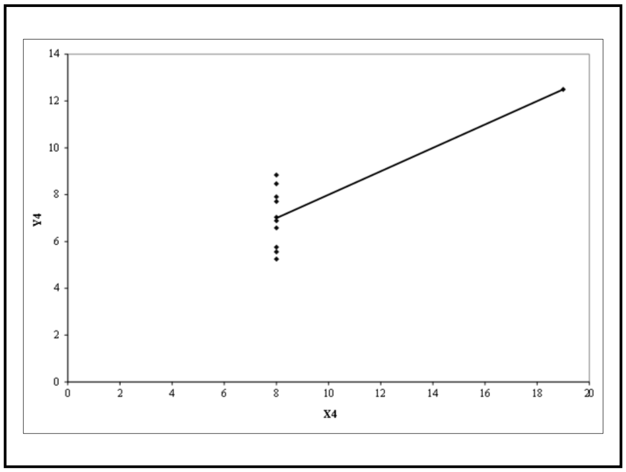
16



17



18



19

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture #2

1

TERM PROJECT FILES NOW AVAILABLE ON THE CLASS WEBSITE

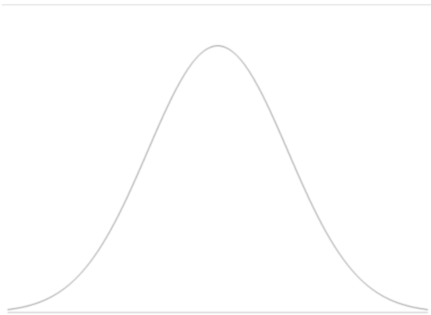
2

FREQUENCY DISTRIBUTIONS

- Frequency distributions are a tool to help visualize whether the observed values of a random variable tend to cluster around in a particular range.
- They also help to identify whether a distribution is *skewed* in a particular direction or exhibits more than one cluster range

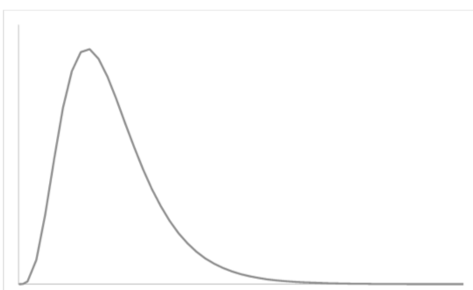
3

SYMMETRIC DISTRIBUTION



4

RIGHT SKEWED



5

LEFT SKEWED

6

BIMODAL DISTRIBUTION

7

NUMERIC RANDOM VARIABLES

- For a numeric random variable, constructing a frequency distribution is effectively a counting exercise to identify the number of observations that fall into given ranges (classes)
- Identify
  - Minimum: Smallest value in the sample
  - Maximum: Largest value in the sample
  - Range: Maximum minus the Minimum
  - Sample Size (usually denoted n): the number of individual observations in the sample

8

CLASSES AND CLASS RANGES

- For some variables there are natural classes to summarize the data
  - Grades – below 60, 60-70, 70-80, etc
  - Income – in $10k increments
- If there are no natural classes, choose the number of classes

Sample Size	Number of Classes
Fewer than 50	5 – 7
50 to 100	7 – 8
101 to 500	8 – 10
501 to 1,000	10 – 11
1,000 to 5,000	11 – 14
More than 5,000	14 – 20

9

CLASSES AND CLASS RANGES

- Next determine the class width
- $w = Class\ Width = \frac{Maximum - Minimum}{Number\ of\ Classes}$
- Round this up to capture the full range of data and set break points for your classes

Class	Value
1	Minimum + Class Width
2	Class 1 + Class Width
3	Class 2 + Class Width
…	…

- Next count the number of observations in the sample that fall into each class

10

CENTRAL TENDENCY

- For a given random variable, measures of Central Tendency provide a single value around which observations 'cluster'
- Mean
  - Arithmetic mean
  - Weighted Mean
  - Geometric Mean
- Median – the 'middle' observation
- Mode – the most frequent single value

11

SOME NOTATION

- Sumation
- $X_1 + X_2 + X_3 + \cdots + X_n = \sum_{i=1}^{n} X_i$
- Population mean is a parameter: Greek letter μ
- Sample (arithmetic) mean: $\overline{X}$
- Weighted mean: $\overline{X}_w$
- Geometric Mean : $\overline{X}_g$

12

SAMPLE MEAN $\overline{X}$

- Just the 'average' of the data
- Excel function =AVERAGE(range)
- Scale effects if a and b are constants and we define
- $Y = a + bX$
- Then the sample mean of Y is

$$\overline{X} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i$$

$$\overline{Y} = \sum \frac{(a + bX_i)}{n} = \frac{na}{n} + b\sum \frac{X_i}{n} = a + b\overline{X}$$

13

MODE

- The most frequent *single* value in a data series
- Often multiple modes
- In many economic variables, the mode is not defined

14

WEIGHTED MEAN

- Simple mean can be inaccurate when there are substantially different sub-groups in a sample
- Especially true when looking at a "mean of means"
- Let w_i be a weight associated with the observation X_i
- Excel use SUMPRODUCT(w-array,x-array) and SUM(w-array)

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

15

GEOMETRIC MEAN

- Common with growth or asset return data
- Instead of sums, use products and fractional exponents
- Excel: GEOMEAN function

$$\overline{X}_g = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$$

16

MEDIAN, QUARTILES AND PERCENTILES

- The MEDIAN is the point in a data series below which 50% of the observations fall
  - If a series is symmetric, the mean and the median will be the same
  - If skewed right, the mean will be greater than the median
  - If skewed left, the mean will be less than the median
  - Excel MEDIAN(range)
- QUARTILES, break the sample into quarters
  - Q1 (25%), Q2 (50%), Q3 (75%)
  - Excel QUARTILE.EXC(range,k) k=1,2, or 3
- Percentiles into any percentage: 1% … 99%
  - Excel PERCENTILE.EXC(range,k) k=0.1, … 0.99

17

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture #3

1

TERM PROJECT OVERVIEW

- Criterion and Predictor Variables
- Sub-Groups

2

CRITERION VARIABLE

- This variable represents the focus of the analysis for your term project
- This MUST be a numeric variable
- In the introduction of your project report, you should clearly describe the variable to be analyzed, how it is measured, and the time period(s) over which it is measured

3

PREDICTOR VARIABLE

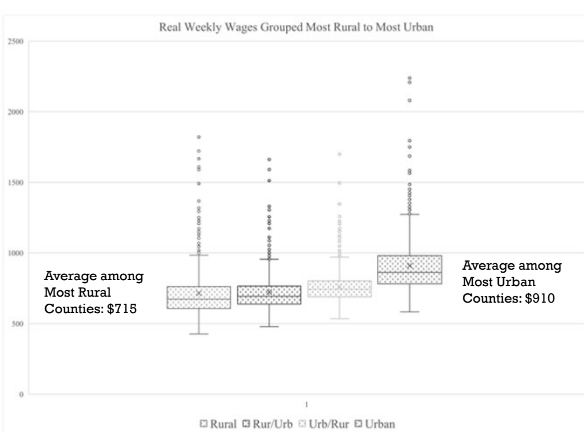
- This is a variable that you have reason to believe is Correlated with your Criterion Variable (think of the Predictor variable as the X-axis variable and the Criterion variable as the Y-axis variable)
- This MUST be a numeric variable
- In the introduction of your project report, you should clearly describe the variable, how it is measured, and the time period(s) over which it is measured

4

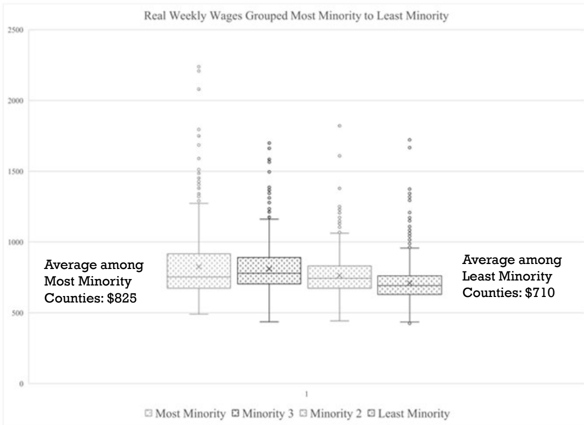
SUB-GROUPS

- You need to clearly describe a method for organizing your overall sample into **at least** 3 mutually exclusive sub-groups (more than 4 groups will get messy)
- This could be an existing categorical variable in one of the datasets, or created from one or two numeric variables
- Examples

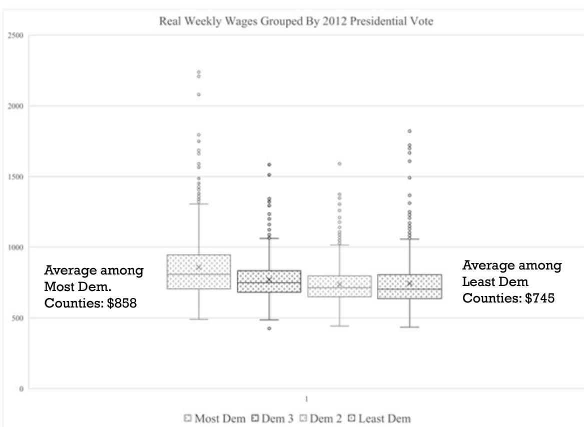
5



6



7



8

MEASURES OF VARIABILITY

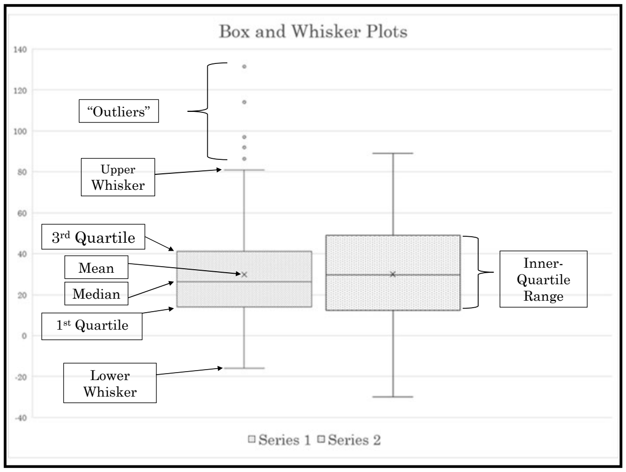
- Distance Measures
- Deviation Measures

9

DISTANCE MEASURES

- Range: Maximum – Minimum
- Inner Quartile Range: $Q_3 - Q_1$
- Box-and-Whisker Plots

10



11

BOX AND WHISKER PLOTS

- The "Box" for each of the two illustrated series, shows the 1st Quartile (bottom of the Box), the Median (solid line in the middle of the Box), and the 3rd Quartile (top of the Box). The Mean for each series is shown as an × on the chart.
- As shown on the Box for Series 2, the Inner-Quartile Range is illustrated by the height of the Box.
- The Lower Whisker shows one of two things – either the Minimum value of the series or the smallest value than is not less than $Q_1 - 1.5 \times Inner - Quartile\ Range$.
- The Upper Whisker is defined similar to the Lower Whisker – either the Maximum value of the series or the largest value that is not greater than $Q_3 + 1.5 \times Inner - Quartile\ Range$.
- Any data points outside the Lower of Upper Whiskers are defined to be "Outliers" or extreme values.

12

MEDIAN, MEAN, Q1, Q3, AND SKEW

- If the Median is close to the middle of the range between the 1st and 3rd quartiles, and the Mean is very close to the Median, (as with Series 2, above) the distribution will tend to be symmetric.
- If the Median it is closer to the first quartile, and the Mean is greater than the Median, (as with Series 1 above) the distribution will tend to be skewed right.
- If the Median it is closer to third quartile, and the Mean is less than the Median, the distribution will tend to be skewed left.

13

DEVIATION MEASURES

- Population Variance: σ^2
  - Excel VAR.P
- Sample Variance: s^2
  - Excel VAR.S
- Population Standard Deviation: $\sigma = \sqrt{\sigma^2}$
  - Excel STDEV.P
- Sample Standard Deviation: $s = \sqrt{s^2}$
  - Excel STDEV.S

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}$$

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

14

DEVIATION MEASURES

- Both Population Variance and Sample Variance measure average squared deviations from their respective mean
- We square the deviations so that observations below the mean (negative) don't offset observations above the mean
- Sample Variance has (n – 1) in the denominator as a ***degrees of freedom*** adjustment to account for estimated sample mean $\bar{X}$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}$$

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

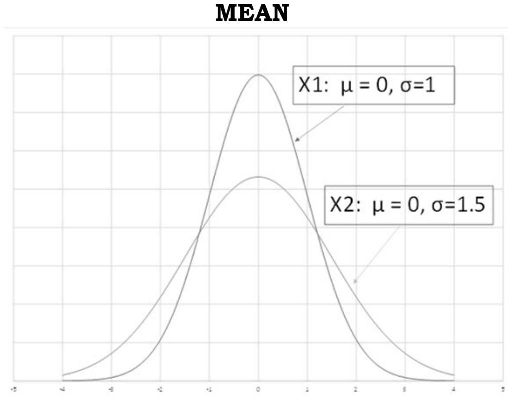
15

STANDARD DEVIATION

- The Population Standard Deviation: $\sigma = \sqrt{\sigma^2}$ and Sample Standard Deviation are central to the process of hypothesis testing
- The Empirical Rule:
- In Symmetric bell-shaped distributions, observations will fall within a range of plus or minus (±) two standard deviations around the mean about 96% of the time.

16

TWO DISTRIBUTIONS WITH THE SAME MEAN



17

CO-MOVEMENT

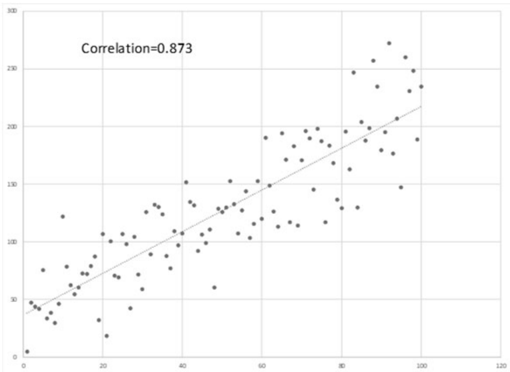
- Covariance
- Sample Correlation
  - Bound on (-1, 1)
  - Excel function CORREL(range1,range2)

$$S_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1}$$

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

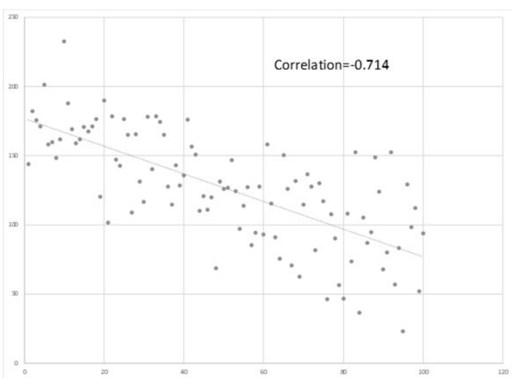
18

POSITIVE CORRELATION



19

NEGATIVE CORRELATION

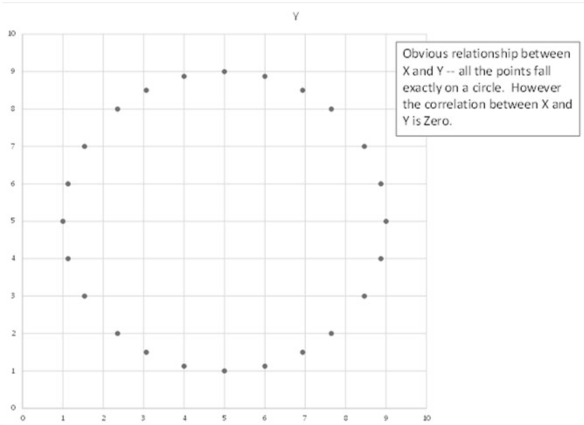


20

CORRELATION

- Note that Correlation is a measure of *statistical* association between two variables
- Correlation ***DOES NOT*** imply any type of causal association between variables
- IF there is a causal relation between two variables, there ***MAY*** be an observable correlation between them
- Stated formally:
- IF two variables are independent of one another, then their correlation is zero
- However, a correlation of zero DOES NOT imply independence

21



22

Z-SCORES

- Standardized measure that shows the number of standard deviations a specific data value 'deviates' from the mean

$$z = \frac{X_i - \mu}{\sigma}$$

$$z_s = \frac{X_i - \bar{X}}{s}$$

23

PROBABILITY

- Can you count to ONE?
- If you answered Yes …
- You have what it takes to be a probability expert

24

RANDOM VARIABLES

- A random variable X takes on different values out of a defined set of possibilities
- Key measure of Central Tendency – the mean
- Key measure of variation – the standard deviation
- Need a way to link specific values (outcomes) of X to the probability of observing that outcome
- The ***PROBABILITY DISTRIBUTION FUNCTION (PDF)***: **f(X)**

25

THE PDF

- The PDF maps specific values of the random variable X to the probability of that value being observed
- For a specific value of X, X_0, the PDF $f(X_0)$ is such that

$$f(X_0) = Probability[X = X_0] \text{ or } P[X = X_0]$$

- All probabilities are bound on the interval [0,1] so $f(X_0)$ is also bound on [0,1]
- Random variables can be either ***discrete*** (countable) or ***continuous*** (take on any value in an interval)

26

DISCRETE RANDOM VARIABLES

- Outcomes are countable but could be a very large number of possibilities (the number of fire ants in a fire ant mound)
- Properties of the PDF
  - Probability of a specific outcome is $f(X_i)$
  - Bound between 0 and 1
  - Sum of probabilities over all possible outcomes is 1

$$f(X_i) = P[X = X_i]$$

$$0 \le f(X_i) \le 1$$

$$\sum_{X_i} f(X_i) = 1$$

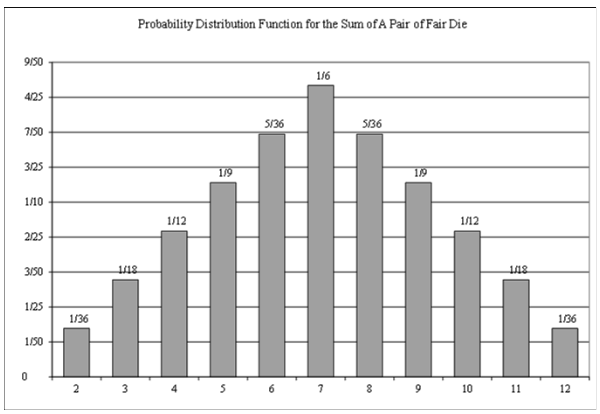
27

ROLL OF A PAIR OF DICE

- The 36 different combinations of the two dice result in one of 11 distinct outcomes from 2 to 12
- The probabilities over all possible outcomes sum to 1

Outcomes: X	Frequency	PDF: f(X)
2	1	1/36
3	2	2/36
4	3	3/36
5	4	4/36
6	5	5/36
7	6	6/36
8	5	5/36
9	4	4/36
10	3	3/36
11	2	2/36
12	1	1/36
Total	36	1

28



29

EXPECTED VALUES

- The population measures of the mean μ and variance σ^2 are defined using ***expected values*** as functions of the PDF

$$E(X) = \sum_{X_i} X_i f(X_i) = \mu$$

$$Var(X) = E[X - E(X)]^2 = E[X - \mu]^2 = \sum_{X_i} (X_i - \mu)^2 f(X_i) = \sigma^2$$

30

CUMULATIVE DISTRIBUTION FUNCTION – CDF

- the Cumulative Distribution Function, or CDF, is denoted F(X) and is the probability that X is less than or equal to a particular value
- Properties of the CDF

31

CDF PROPERTIES

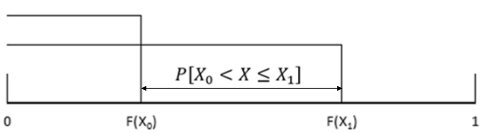
$$F(X_i) = P[X \leq X_i]$$

$$F(X_i) = \sum_{X \leq X_i} f(X)$$

$$0 \leq F(X_i) \leq 1$$

$$if\ X_0\ and\ X_1\ are\ such\ that\ X_0 < X_1\ then\ F(X_0) \leq F(X_1)$$

$$if\ X_0\ and\ X_1\ are\ such\ that\ X_0 < X_1\ then\ P[X_0 < X \leq X_1] = F(X_1) - F(X_0)$$



32

CDF PROPERTIES CONT.

$$Given\ P[X \leq X_i] = F(X_i)$$

$$Because\ f(X)\ must\ sum\ to\ 1, then$$

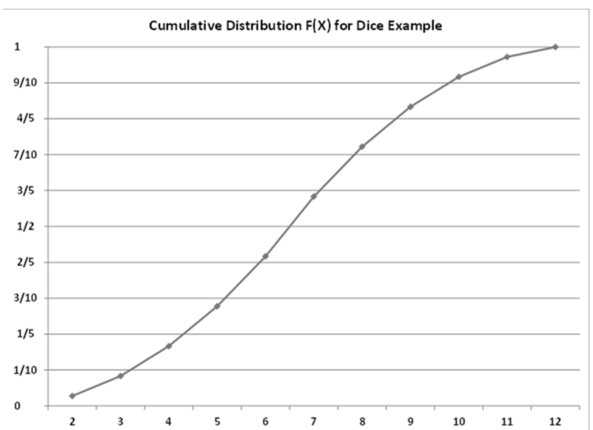
$$P(X > X_i) = 1 - F(X_i)$$

33

DICE EXAMPLE

Outcomes: X	Frequency	PDF: f(X)	CDF: F(x)
2	1	1/36	1/36
3	2	2/36	3/36
4	3	3/36	6/36
5	4	4/36	10/36
6	5	5/36	15/36
7	6	6/36	21/36
8	5	5/36	26/36
9	4	4/36	30/36
10	3	3/36	33/36
11	2	2/36	35/36
12	1	1/36	1
Total	36	1	

34



35

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture #4

1

PROBABILITY CONT.

- Binomial Distribution
- Poisson Distribution
- Exponential Distribution
- Normal Distribution
- Normal Approximation to the Binomial Distribution

2

BINOMIAL DISTRIBUTION

- A Binomial random variable can take on one of only two values: a "success" or a "failure."
- A fixed number of observations, n; for example, 13 tosses of a coin; 11 cell phones taken from a production line.
- Two mutually exclusive and collectively exhaustive categories; "Heads" or "tails" on the toss of a coin; "Defective" or "not defective" for a given cell phone;
- Constant probability of "success" for each observation;
- Observations are independent: the outcome of one observation does not affect the outcome of another.

3

BINOMIAL PDF AND CDF

- Derived from the number of sequences with x successes in n independent experiments

$$C_x^n = \frac{n!}{x!\,(n-x)!}$$

Where $n! = n(n-1)(n-2)\ldots, and\ 0! = 1$

- PDF

$$f(x) = \frac{n!}{x!\,(n-x)!}P^x(1-P)^{n-x}$$

- CDF

$$F(x_0) = P[x \le x_0] = \sum_{x=0}^{x_0}\frac{n!}{x!\,(n-x)!}P^x(1-P)^{n-x}$$

4

BINOMIAL DISTRIBUTION

- Let **n**; be the number of observations (the sample size or number of 'trials')
- Let X be the number of 'successes" out of the n observations
- Let P be the constant probability of "success" for each observation
- The expected (population mean) number of successes in a sample size of n is given by:

$$E(X) = \mu = nP$$

5

BINOMIAL EXAMPLE

- Based on Fall 2015 enrollment statistics, 59% of A&M students in the College of Liberal Arts are female.
- Suppose you select a random sample of 6 Liberal Arts students. What is the probability that 4 of those chosen are female?
- The probability of "success" is given as P = 0.59 and the sample size (or number of 'trials') is given as n = 6

$$P[x=4] = \frac{6!}{4!\,(6-4)!}(0.59)^4(1-0.59)^{(6-4)} \cong 0.3055$$

6

BINOMIAL EXAMPLE

- The formulas look very messy
- In practice, we can use the Excel function BINOM.DIST(x,n,P,0) to get the PDF and BINOM.DIST(x,n,P,1) for the CDF

P =	0.59
n =	6
X	f(x)
0	0.005
1	0.041
2	=BINOM.DIST(B7,C3,C2,0)
3	0.283
4	0.306
5	0.176
6	0.042

P =	0.59	
n =	6	
X	f(x)	F(x)
0	0.005	0.005
1	0.041	0.046
2	0.148	0.193
3	0.283	0.476
4	0.306	=BINOM.DIST(B9,C3,C2,1)
5	0.176	0.958
6	0.042	1.000

7

DISCRETE RANDOM VARIABLES AND INEQUALITIES

- Be careful with strict versus weak inequalities with discrete random variables
- "More than 3" ➔ $X \geq 4$ ➔ $1 - F(3)$
- "At least 3" ➔ $X \geq 3$ ➔ $1 - F(2)$
- "At least 2 but less than 5" ➔ $2 \leq X \leq 4$ ➔ $F(4) - F(1)$
- "More than 2 but 5 or less" ➔ $3 \leq X \leq 5$ ➔ $F(5) - F(2)$

8

POISSON DISTRIBUTION

- The Poisson probability distribution can be used to model the discrete number of occurrences (or successes) of a certain event in a given continuous interval such as time, spatial area, or length.
  - The number of trucks arriving at a warehouse in a given week.
  - The number of failures in a computer system in a given day.
  - The number of defects in a large roll of sheet metal.
  - The number of customers to arrive at a coffee bar in a given time interval.

9

POISSON ASSUMPTIONS

- Assume that an interval is divided into a large number of equal subintervals (e.g., milliseconds or nanometers) so that the probability of the occurrence of an event in any subinterval is small.
- The probability of the occurrence of an event is constant for all subintervals.
- There can be no more than one occurrence in each subinterval.
- Occurrences are independent – an occurrence in one subinterval does not have an effect on the probability of an occurrence in another subinterval.

10

POISSON PDF

- The PDF for the Poisson distribution is shown at the right, where:
- P(x) = the probability of x successes over a given time or space unit, given λ.
- The Greek letter lambda λ is the mean (or expected) number of successes per time or space unit, λ>0.
- e = 2.71828… (the base for natural logarithms).

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \; for \; x = 0, 1, 2, \ldots$$

$$e = \sum_{n=1}^{\infty} 1 + \frac{1}{n!} = 1 + \frac{1}{1} + \frac{1}{1 \times 2} + \frac{1}{1 \times 2 \times 3} + \cdots$$

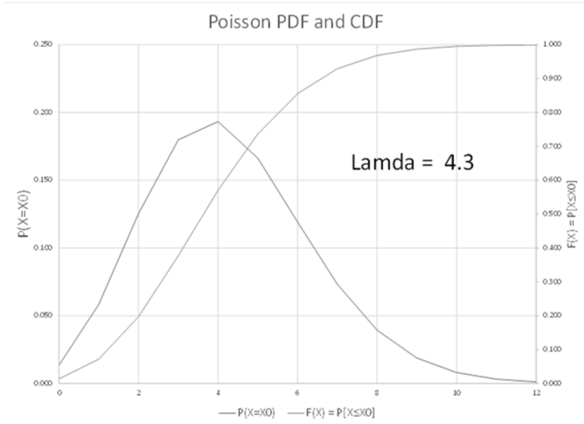
11

POISSON CDF

- Recall that the CDF is defined as the probability of X being less than or equal to some specific value, X_0
- We get the CDF by summing up the PDF for all values of X≤ X_0

$$P[X \leq X_0] = F(X) = \sum_{X_i=0}^{X_0} P(X_i)$$

12



13

POISSON EXAMPLE

- Along a particular stretch of Highway 290 in Houston there are, on average, 4.3 auto accidents on weekday mornings. For the following questions, assume that each weekday morning is independent and that accident occurrences follow the Poisson probability distribution.
  - What is the probability that on a randomly chosen weekday morning, there will be 4 auto accidents?
  - What is the probability that on a randomly chosen weekday morning, there will be 2 or fewer auto accidents?
  - What is the probability that on a randomly chosen weekday morning, there will more than 3 auto accidents?

14

POISSON EXAMPLE

- We let Excel handle the calculations using
  - =POISSON.DIST(X,λ,0) for the PDF P(X) and
  - =POISSON.DIST(X,λ,1) for the CDF F(X)

Lamda =	4.3	
X	P(X)	F(X)
0	0.014	0.014
1	0.058	0.072
2	0.125	0.197
3	0.180	0.377
4	=POISSON.DIST($A8,$B$1,0)	
5	0.166	0.737
6	0.119	0.856

15

POISSON EXAMPLE

- What is the probability that on a randomly chosen weekday morning, there will be 4 auto accidents?
  - Need P[X=4] = 0.193
- What is the probability that on a randomly chosen weekday morning, there will be 2 or fewer auto accidents?
  - Need P[X≤2] = 0.197
- What is the probability that on a randomly chosen weekday morning, there will more than 3 auto accidents?
  - Need 1 – P[X≤3] = 1 – 0.377 = 0.623

Lamda =	4.3	
X	P(X)	F(X)
0	0.014	0.014
1	0.058	0.072
2	0.125	0.197
3	0.180	0.377
4	0.193	0.570
5	0.166	0.737
6	0.119	0.856

16

EXPONENTIAL DISTRIBUTION

- A continuous distribution in that outcomes can take on any value greater than zero
- Used to model the length of time between occurrences of an event
  - Time between trucks arriving at a warehouse.
  - Time between customers calling a help-line.
- PDF of the Exponential distribution: $f(t) = \lambda e^{-\lambda t}$
- Where:
  - λ is the mean number of **occurrences per unit** t
  - The mean time between occurrences is $^1/_\lambda$
  - t is the number of units (time or space)
  - e is the nature number = 2.71828 …

17

EXPONENTIAL CDF

- Because time t is measured continuously, ***the probability of t equal to a specific value is technically zero*** – only ranges of time are meaningful in terms of probability
- As a result, with continuous distributions, there is no real distinction between a weak inequality (≤) and a strict inequality (<)

$$F(t_0) = P[t \leq t_0] = P[t < t_0] = 1 - e^{-\lambda t_0}$$

$$t > 0$$

18

EXPONENTIAL EXAMPLE

- Along a particular stretch of Highway 290 in Houston, there are, on average, 2.1 auto accidents per hour during the busiest portion of the weekday morning commute. For the following questions, assume accidents are independent and that the timing of accidents follows the exponential probability distribution.
- On a randomly chosen weekday morning, what is the probability of an accident occurring within 15 minutes (15 minutes or less)?
- What is the probability of an accident occurring between 20 and 40 minutes?
- What is the probability that it will be more than 1 hour before an accident occurs?

19

EXPONENTIAL EXAMPLE

- In Excel, we can calculate the CDF directly using the exponential function EXP or use the EXPON.DIST function
- On a randomly chosen weekday morning, what is the probability of an accident occurring within 15 minutes (15 minutes or less)?

$$P\left(t \leq \frac{15}{60}\right) = 1 - e^{-2.1\times\frac{15}{60}} = 0.4084$$

Lambda =	2.1	per hour ==> 60 minutes		
Using EXPON.DIST function			Using EXP Function	
t in Minutes	F(t)		t in Minutes	F(t)
15	=EXPON.DIST(A4/60,B1,1)			0.408
20	0.503		20	0.503
40	0.753		40	0.753
60	0.878		60	0.878

Lambda =	2.1	per hour ==> 60 minutes		
Using EXPON.DIST function			Using EXP Function	
t in Minutes	F(t)		t in Minutes	F(t)
15	0.408		15	=1-EXP(-B1*D4/60)
20	0.503		20	0.503
40	0.753		40	0.753
60	0.878		60	0.878

20

EXPONENTIAL EXAMPLE

- What is the probability of an accident occurring between 20 and 40 minutes?
  - P[20/60 ≤ t ≤ 40/60] = F(40/60) – F(20/60)

Lambda =	2.1	per hour ==> 60 minutes	
X in Minutes	F(t)		
15	0.408		
20	0.503		F(40/60) - F(20/60)
40	0.753		0.250
60	0.878		1 - F(60/60)
			0.122

- What is the probability that it will be more than 1 hour before an accident occurs?
  - P[t > 1] = P[t > 60/60] = 1 – F(1)

21

EXPONENTIAL EXAMPLE

- What is the probability of an accident occurring between 20 and 40 minutes? Solving this problem with a calculator:

$$P\left[\frac{20}{60} \le t \le \frac{40}{60}\right] = F\left(\frac{40}{60}\right) - F\left(\frac{20}{60}\right)$$

$$= \left(1 - e^{-2.1\times\frac{40}{60}}\right) - \left(1 - e^{-2.1\times\frac{20}{60}}\right)$$

$$= e^{-2.1\times\frac{20}{60}} - e^{-2.1\times\frac{40}{60}}$$

22

NORMAL DISTRIBUTION

- A symmetric bell-shaped distribution that closely approximates many phenomena observed in nature and economics
- A continuous random variable X with mean (expected value) μ and variance σ^2 follows the normal probability distribution if the PDF of X is

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

- And CDF

$$F(X) = \int_{-\infty}^{X} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} dx$$

23

NORMAL DISTRIBUTION

- The CDF of the Normal cannot be 'solved' – we must use numerical algorithms to get a solution
- However, any Normal random variable X, denoted **X~N(μ, σ²)**, can be transformed into a Standard Normal random variable by creating a Z-score

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- And …

24

STANDARD NORMAL

$$P[X \le X_0] = P[Z \le Z_0] = F(Z_0)$$

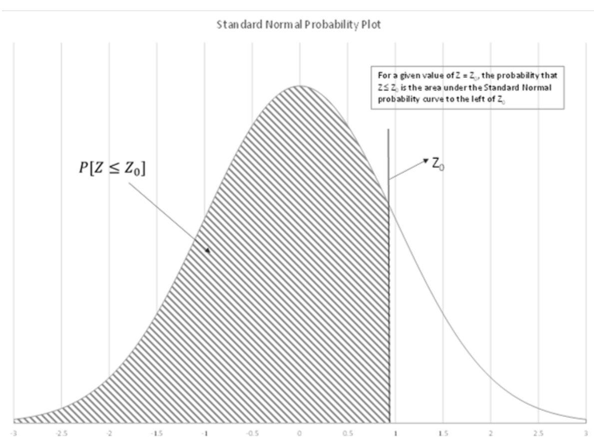
$$P[X_0 \le X \le X_1] = P[Z_0 \le Z \le Z_1] = F(Z_1) - F(Z_0)$$

- In Excel use the =NORM.S.DIST(Z,1)

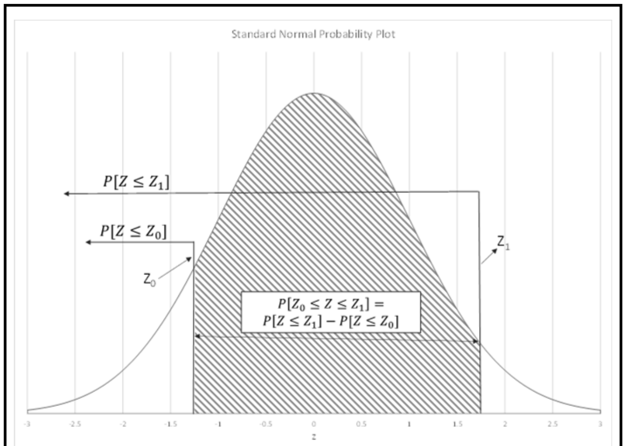
Standard Normal		
Mean μ =	12	
Sigma σ =	4	
X	Z	F(X)
5	=(A21-B17)/B18	
21	2.25	0.988

Standard Normal		
Mean μ =	12	
Sigma σ =	4	
X	Z	F(X)
5	-1.75	=NORM.S.DIST(B21,1)
21	2.25	0.988

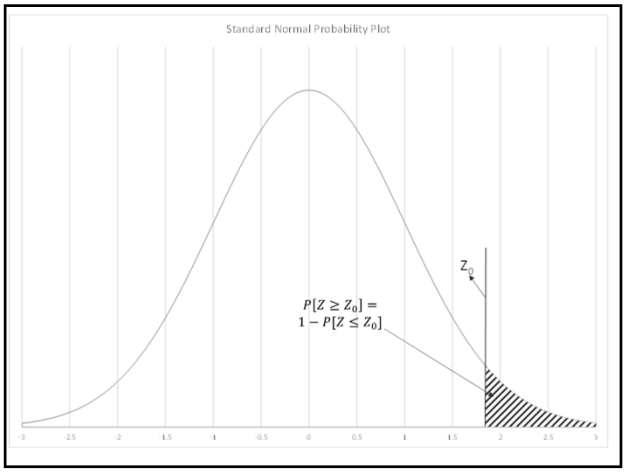
25



26



27



28

NORMAL EXAMPLE

- The random variable X is distributed as Normal with a mean μ=12 and a standard deviation σ=4
  - What is the probability that
  - $X \leq 5$?
  - $X \leq 21$?
  - $X \geq 21$?
  - $5 \leq X \leq 21$?

Standard Normal		
Mean μ =	12	
Sigma σ =	4	
X	Z	F(X)
5	-1.75	0.040
21	2.25	0.988
P[X ≥ 21] = 1 - P[X ≤ 21]		
	0.012	
P[5 ≤ X ≤ 21] = P[-1.75 ≤ Z ≤ 2.25] = F(2.25) - F(-1.75)		
	0.948	

29

NORMAL APPROX. TO BINOMIAL

- With a binomial random variable, let X be the number of 'successes" out of the n observations
- Let P be the constant probability of "success" for each observation, then

$$E(X) = \mu = nP$$

$$Var(X) = \sigma^2 = nP(1 - P)$$

30

NORMAL APPROX. TO BINOMIAL

- If n is 'large' (when $n > 5 \times P(1-P)$), we can approximate the binomial probabilities by calculating a Z-score using nP for μ and nP(1 – P) for σ^2

$$Z = \frac{X_0 - nP}{\sqrt{nP(1-P)}}$$

- Similarly …

31

NORMAL APPROX. TO BINOMIAL

- To make probability statements about a sample ***proportion*** X/n

$$E\left(\frac{X}{n}\right) = \mu = P$$

$$Var\left(\frac{X}{n}\right) = \sigma^2 = \frac{P(1-P)}{n}$$

32

NORMAL APPROX. TO BINOMIAL

- So to get the probability $P[P \leq P_0]$ we can calculate a Z-Score

$$Z = \frac{P_0 - P}{\sqrt{\frac{P(1-P)}{n}}}$$

33

NORMAL APPROX. TO BINOMIAL

- At a major metropolitan hospital, 48% of patients admitted to the hospital have medical coverage through Medicare/Medicaid. On a typical day, there are 155 new patients admitted to the hospital.
  - What is the mean number of Medicare/Medicaid covered patients?
  - What is the variance of the number of Medicare/Medicaid covered patients?
  - What is the mean of the sample proportion of Medicare/Medicaid covered patients?
  - What is the variance of the sample proportion of Medicare/Medicaid covered patients?
  - Using the Normal Approximation to the Binomial Distribution, what is the probability that the number of Medicare/Medicaid patients is greater than 75?
  - Using the Normal Approximation to the Binomial Distribution, what is the probability that the sample proportion is greater than 51% (or 0.51)?

34

NORMAL APPROX. TO BINOMIAL

Normal Approximation to the Binomial Distribution

P =	0.48	
n =	155	
E(X) = μ =	74.4	<== Mean number of Medicare/Medicaid Patients
Var(X) = σ^2 =	38.688	<== Variance of number of M/M Patients
E(X/n) = P =	0.48	<== Mean proportion of M/M patients
Var(X/n) =	0.00161	<== Variance of proportion of M/M patients

X	Z	F(Z)
60	-2.3151	0.0103
65	-1.5113	0.0654
70	-0.7074	0.2397
75	0.0965	0.5384
80	0.9003	0.8160

X/n	Z	F(Z)
0.47	-0.249	0.4016
0.49	0.249	0.5984
0.51	0.748	0.7726
0.53	1.246	0.8936

35

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture #5

1

EXAM 1

- Exam 1 will be administered online through Canvas on Thursday, February 15th
- You will need to download and install the LockDown browser and have a webcam equipped PC/Mac
  - See the LockDown browser instructions on the class website
- I have set up an unscored "Trial Exam" in Canvas for you to test out how the LockDown browser works
- Exam Study Guide and Example Problems are available on the class website
- Exam review and discussion of procedure on Tuesday, 2/13

2

NORMAL EXAMPLE

- The random variable X is distributed as Normal with a mean μ=12 and a standard deviation σ=4
  - The probability is 0.25 that X is greater than what number
  - $P[X > ?] = 0.25$
- With the info given $P\left[Z > \frac{X-12}{4}\right] = 0.25 \;\; so \;\; P\left[Z < \frac{X-12}{4}\right] = 0.75$
- Use NORM.S.INV(0.75) to get $\frac{X-12}{4} \cong 0.6745$
- Solve to get: $\text{X} \cong 4 * 0.6745 + 12 \cong 14.698$

3

Z-SCORES AGAIN

- For a random variable X with mean equal to Mean(X), or E(X), and variance equal to Var(X), we constructed a Z-Score using:

$$Z = \frac{X - E(X)}{\sqrt{Var(X)}}$$

- How we get E(X) and Var(X) can change depending on the nature of the random variable
- It could be given as $E(X) = \mu$ and $Var(X) = \sigma^2$, or it could relate to other parameters or sample statistics

4

NORMAL APPROX. TO BINOMIAL

- With a binomial random variable, let X be the number of 'successes" out of the **n** observations
- Let P be the constant probability of "success" for each observation, then

$$E(X) = \mu = nP$$

$$Var(X) = \sigma^2 = nP(1 - P)$$

5

NORMAL APPROX. TO BINOMIAL

- If n is 'large' (when $n > 5 \times P(1 - P)$), we can approximate the binomial probabilities by calculating a Z-score using nP for μ and nP(1 – P) for σ^2

$$Z = \frac{X_0 - nP}{\sqrt{nP(1 - P)}}$$

- Note: When using the Normal Approximation to the Binomial, treat X as a continuous variable, so $X < 73$ is treated the same as $X \leq 73$
- Similarly …

6

NORMAL APPROX. TO BINOMIAL

- To make probability statements about a sample ***proportion*** X/n

$$E\left(\frac{X}{n}\right) = \mu = P$$

$$Var\left(\frac{X}{n}\right) = \sigma^2 = \frac{P(1-P)}{n}$$

7

NORMAL APPROX. TO BINOMIAL

- To get the probability $P[P \leq P_0]$ we can calculate a Z-Score

$$Z = \frac{P_0 - P}{\sqrt{\frac{P(1-P)}{n}}}$$

- Again, when using the Normal Approximation to the Binomial, treat P_0 as a continuous variable, so $P_0 < 0.4$ is the same as $P_0 \leq 0.4$

8

NORMAL APPROX. TO BINOMIAL

- At a major metropolitan hospital, 48% of patients admitted to the hospital have medical coverage through Medicare/Medicaid. On a typical day, there are 155 new patients admitted to the hospital.
  - What is the mean number of Medicare/Medicaid covered patients?
  - What is the variance of the number of Medicare/Medicaid covered patients?
  - What is the mean of the sample proportion of Medicare/Medicaid covered patients?
  - What is the variance of the sample proportion of Medicare/Medicaid covered patients?
  - Using the Normal Approximation to the Binomial Distribution, what is the probability that the number of Medicare/Medicaid patients is greater than 75?
  - Using the Normal Approximation to the Binomial Distribution, what is the probability that the sample proportion is greater than 51% (or 0.51)?

9

NORMAL APPROX. TO BINOMIAL

Normal Approximation to the Binomial Distribution		
P =	0.48	
n =	155	
E(X) = μ =	74.4	<== Mean number of Medicare/Medicaid Patients
Var(X) = σ² =	38.688	<== Variance of number of M/M Patients
E(X/n) = P =	0.48	<== Mean proportion of M/M patients
Var(X/n) =	0.00161	<== Variance of proportion of M/M patients
X	Z	F(Z)
60	-2.3151	0.0103
65	-1.5113	0.0654
70	-0.7074	0.2397
75	0.0965	0.5384
80	0.9003	0.8160
X/n	Z	F(Z)
0.47	-0.249	0.4016
0.49	0.249	0.5984
0.51	0.748	0.7726
0.53	1.246	0.8936

10

SAMPLING DISTRIBUTIONS

- The previous discussion dealt with probabilities for specific values of a random variable
- We want to be able to make probability statements regarding various ***sample statistics*** that we calculate from sample data such as the sample mean $\overline{X}$, the sample variance S^2, or a sample proportion $\hat{P}$
- If a random variable X has a constant population mean μ and variance σ^2 it is denoted $X \sim (\mu, \sigma^2)$
- Given a random sample of size n for X, the **Sampling Distribution** is the probability distribution function (PDF) of the various sample statistics

11

SAMPLING DISTRIBUTION OF THE SAMPLE MEAN $\overline{X}$

- For a random variable $X \sim (\mu, \sigma^2)$ – a random variable with a constant mean and constant variance, the sample mean $\overline{X}$ has the distribution properties:

$$\bar{X} = \frac{\sum X_i}{n}$$

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim (0, 1)$$

12

EXPECTED VALUES AGAIN

- For a random variable $X \sim (\mu, \sigma^2)$, with a random sample $\{X_1, \ldots, X_n\}$ and constants $\{a_1, \ldots, a_n\}$

$$E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n) = \mu(a_1 + \cdots + a_n)$$

$$Var(a_1X_1 + \cdots + a_nX_n) = a_1^2Var(X_1) + \cdots + a_n^2Var(X_n) = \sigma^2(a_1^2 + \cdots + a_n^2)$$

$$E(\overline{X}) = E\left(\frac{1}{n}\right)(X_1 + \cdots + X_n) = \left(\frac{1}{n}\right)[E(X_1) + \cdots + E(X_n)] = \left(\frac{1}{n}\right)[n\mu] = \mu$$

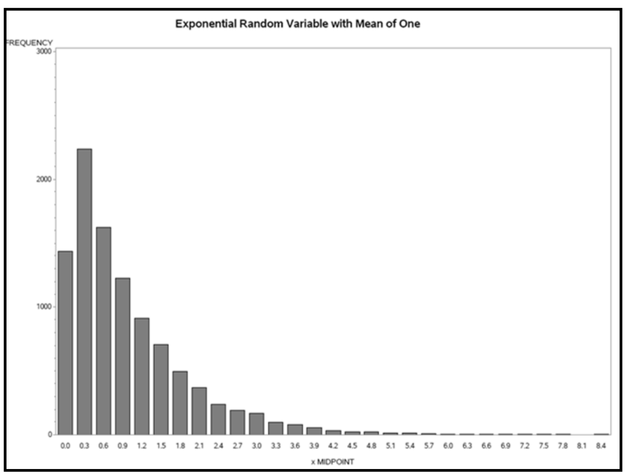
$$Var(\overline{X}) = \left(\frac{1}{n}\right)^2 [Var(X_1) + \cdots + Var(X_n)] = \left(\frac{1}{n}\right)^2 [n\sigma^2] = \frac{\sigma^2}{n}$$

13

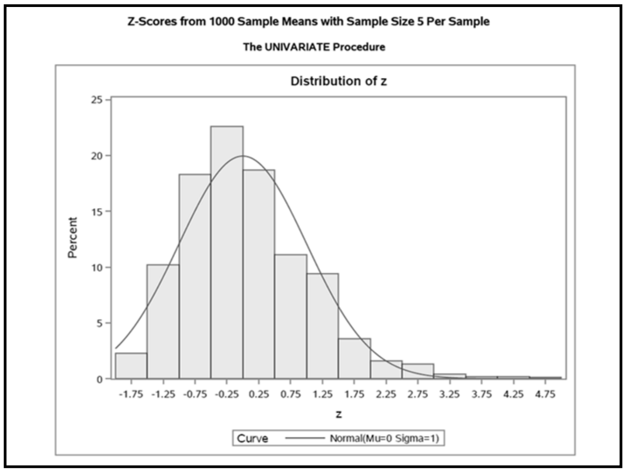
CENTRAL LIMIT THEOREM

- For a random variable $X \sim (\mu, \sigma^2)$, as the sample size n becomes “large” then Z as defined above is approximately Normally distributed.
- Thus, for *any* random variable with a constant mean and variance, we can make statements in probability about the sample mean based on the Normal probability distribution when we have a sufficiently large sample
- Usually, 20 or more observations is a sufficiently large sample to invoke the Central Limit Theorem

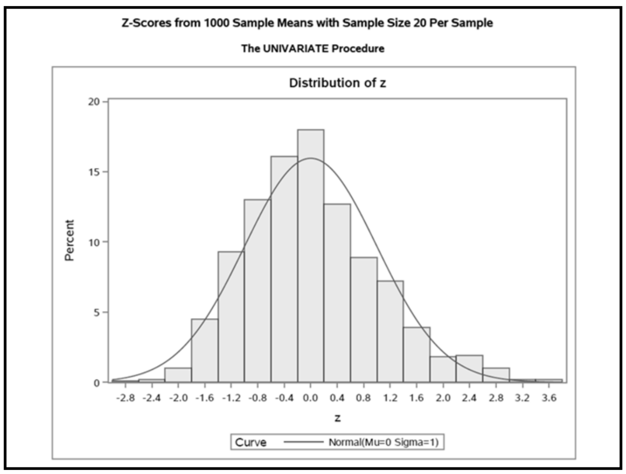
14



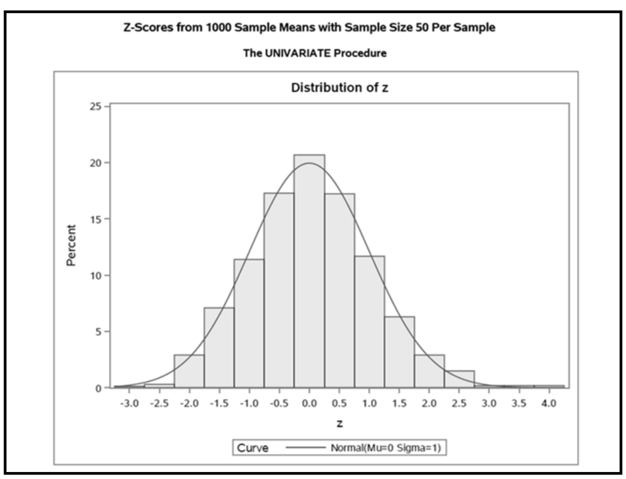
15


Z-Scores from 1000 Sample Means with Sample Size 5 Per Sample
The UNIVARIATE Procedure
Distribution of z
Percent
z
Curve
Normal(Mu=0 Sigma=1)

16


Z-Scores from 1000 Sample Means with Sample Size 20 Per Sample
The UNIVARIATE Procedure
Distribution of z
Percent
z
Curve
Normal(Mu=0 Sigma=1)

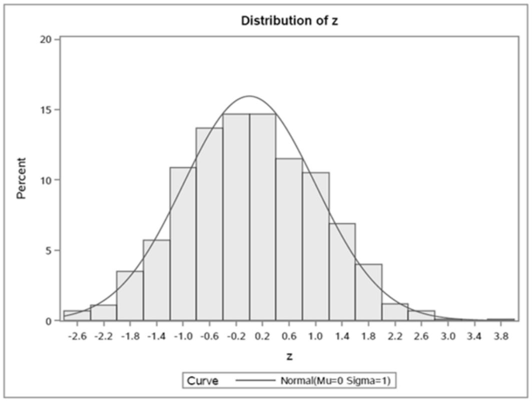
17


Z-Scores from 1000 Sample Means with Sample Size 50 Per Sample
The UNIVARIATE Procedure
Distribution of z
Percent
z
Curve
Normal(Mu=0 Sigma=1)

18

Z-Scores from 1000 Sample Means with Sample Size 100 Per Sample

The UNIVARIATE Procedure



19

EXAMPLE

- A random variable X is distributed with a constant mean μ = 23 and variance $\sigma^2 = 400$
- A random sample of size n = 36 is obtained. What is the probability of finding a sample mean $\overline{X} \leq 17$?
- Find the Z-Score for $\overline{X} = 17$

$$Z_0 = \frac{\overline{X}_0 - \mu}{\sqrt{\sigma^2/n}} = \frac{17 - 23}{\sqrt{400/36}} = -1.8$$

- In Excel, use =NORM.S.DIST(-1.8,1) to get $P[Z \leq Z_0] \cong 0.036$

20

EXAMPLE CONT.

- What is the probability of finding a sample mean $\overline{X} \geq 27$?
- Find the Z-Score for $\overline{X} = 27$

$$Z_0 = \frac{27 - 23}{\sqrt{400/36}} = 1.2$$

- In Excel, use =1-NORM.S.DIST(1.2,1) to get $P[Z \geq Z_0] \cong 0.115$
- To find $P[17 \leq \overline{X} \leq 27]$ use
  - =NORM.DIST(1.2,1) – NORM.DIST(-1.8,1) = 0.849

21

NORMAL APPROX. TO BINOMIAL

- Let X be a binomial random variable that equals one (a "success") with probability P and is zero with probability (1 – P)
- The sampling distribution of the sample ***proportion*** $\hat{P}$ is
- n is "large" if: $n\hat{P}(1-\hat{P}) > 5$

$$\hat{P} = \frac{\sum X_i}{n}$$

$$E(\hat{P}) = P$$

$$Var(\hat{P}) = \frac{P(1-P)}{n}$$

$$Z = \frac{\hat{P} - P}{\sqrt{P(1-P)/n}} \sim N(0,1) \text{ if } n \text{ is "large"}$$

22

EXAMPLE

- In Major League Baseball, the probability that a pitcher throws strictly left-handed is 26% (P = 0.26)
- In a random sample of n = 100 MLB pitchers, what is the probability of finding a sample proportion $\hat{P} \geq 0.33$?

$$Z_0 = \frac{\hat{P} - P}{\sqrt{P(1-P)/n}} = \frac{0.33 - 0.26}{\sqrt{0.26(1-0.26)/100}} \cong 1.596$$

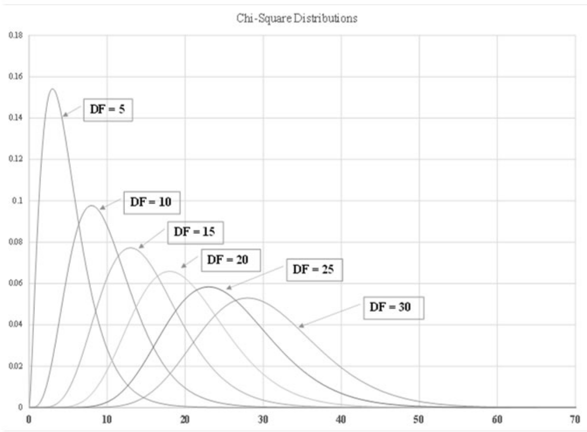
$$P\left[P \geq \hat{P}\right] = P[Z \geq Z_0] = 1 - P[Z \leq Z_0] \cong 0.055$$

23

SAMPLING DIST. OF THE SAMPLE VARIANCE S^2

- If Z~N(0, 1) then Z^2 is distributed as a Chi-Squared random variable with 1 degree of freedom. This is denoted $\chi^2_{(1)}$.
- If $Z_1, Z_2, \ldots, Z_n$ are independently distributed N(0, 1) then $\sum Z_i^2 \sim \chi^2_{(n)}$ (Chi-Squared with n degrees of freedom).
- Note that the Chi-Squared distribution is not symmetric – its shape depends on the degrees of freedom

24



25

SAMPLING DIST. OF THE SAMPLE VARIANCE S^2

- The sampling distribution of the sample variance S^2 is then

$$E(S^2) = \sigma^2$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

26

EXAMPLE

- The random variable X is known to follow the Normal distribution with ***standard deviation σ = 16***
- In a random sample of n = 90, what is the probability of finding a sample variance $S^2 \leq 200$
- Calculate $\chi^2_{calc} = \frac{(n-1)S^2}{\sigma^2} = \frac{(90-1)200}{16^2} \cong 69.531$
- In Excel, use =CHISQ.DIST(chi2_calc,n-1,1) to get

$$P[S^2 \leq 200] = P\left[\chi^2_{(n-1)} \leq \chi^2_{calc}\right] \cong 0.0629$$

- To get the probability $S^2 \geq 200$, use
  - =1-CHISQ.DIST(chi2_calc,n-1,1) or
  - =CHISQ.DIST.RT(chi2_calc,n-1)

27

CHI-SQUARE DISTRIBUTION

28

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture 6: Exam 1 Review

1

EXAM 1

- Exam 1 will be administered online through Canvas on Thursday, February 15th
- You will need to download and install the LockDown browser and have a webcam equipped PC/Mac
  - See the LockDown browser instructions on the class website
- I have set up an unscored "Trial Exam" in Canvas for you to test out how the LockDown browser works
- Exam Study Guide and Example Problems are available on the class website

2

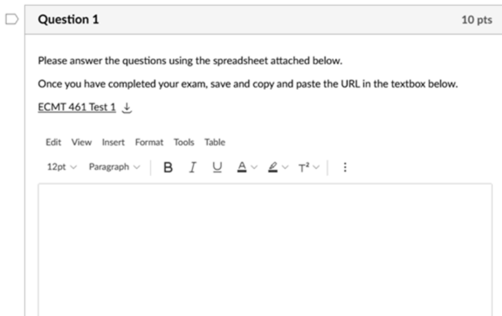
EXAM 1

- Exam 1 will be available in Canvas starting at 12:01 am on Thursday morning, February 15th (just after midnight on Wednesday) and remain open until midnight Thursday night
- Pick any 2-hour window to take the exam
- You need to have the LockDown browser installed on a webcam equipped PC/Mac
- No regular lecture on Monday

3

EXAM 1 STUFF

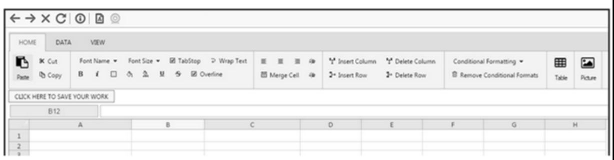
When you open the exam, you will see a single question with a link to the Excel based worksheet. Click the download arrow



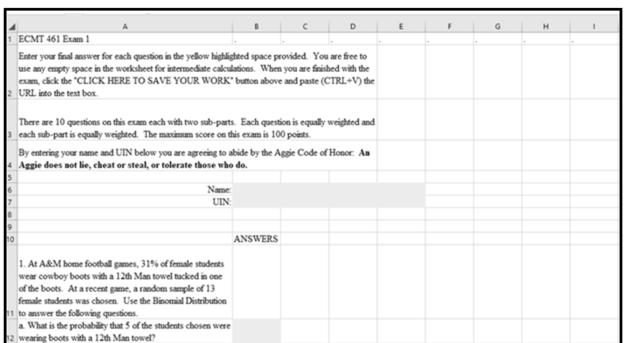
4

EXAM 1 STUFF

The exam will open in a worksheet app that looks like the following



5



Fill in your name, UIN and final answers for each question in the highlighted areas. You can use any empty spaces in columns C – I for intermediate calculations. For questions requiring calculations, cells are formatted to 4 decimal places. If you simply fill in a number for a final answer rather than a formula, we cannot give partial credit.

6

BUGS IN THE WORKSHEET APP

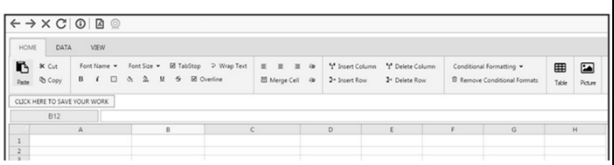
- There are bugs in the LockDown browser worksheet app related to the BINOM.DIST and POISSON.DIST function that give non-sensical answers.
- To deal with this, for problems involving the Binomial or Poisson distribution, you will be asked how you would calculate the probability. – simply type in PDF(x_0) or CDF(x_0) for your answer
- For Exponential, Normal and Normal Approximation to the Binomial problems, you will be asked to calculate the (numeric) probability.
- Examples

7

BUGS IN THE WORKSHEET APP

- For example, for a Binomial problem with n=6 and P=0.39, to get the probability that X=3, write
  - PDF(3)
- To get the probability that X<3, write
  - CDF(2)
- To get the probability that X>3, write
  - 1 – CDF(3)
- For a Poisson problem with mean=4.5, to get the probability that X=5, write
  - PDF(5)
- To get the probability that X≤4, write
  - CDF(4)
- Etc.

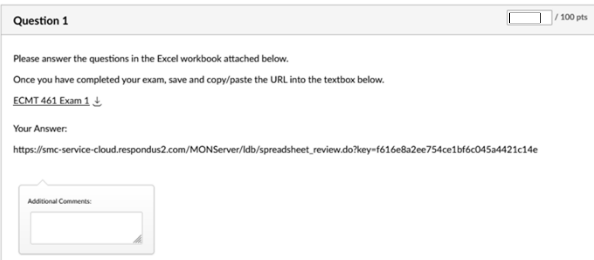
8



After you have completed all the questions in the workbook, click the "CLICK HERE TO SAVE YOUR WORK" button at the top, then go back to the "essay" question and paste (CTRL-V) the URL into the text box.

9

After you have pasted the URL into the textbox, it should look like the following:



10

BINOMIAL EXAMPLE

- Major league baseball player Ronald Acuna, Jr. leads the MLB with a batting average of 0.355. Assume Mr. Acuna's at-bats can be treated as an independent Binomial random variable with a constant probability of a hit of 35.5%. For a random sample of n=23 of Mr. Acuna's at-bats:
  - What is the probability that Mr. Acuna gets 11 hits?
    - *PDF(11)*
  - What is the probability that Mr. Acuna gets less than 9 hits?
    - *CDF(8)*
  - What is the probability that Mr. Acuna at least 7 hits?
    - $\mathbf{1 - \mathit{CDF}(6)}$
  - What is the probability that Mr. Acuna gets more than 5 hits but less than 13 hits?
    - $\mathbf{\mathit{CDF}(12) - \mathit{CDF}(5)}$

11

POISSON EXAMPLE

- During the 2021 academic year, the Texas A&M Police dealt with an average of 3.7 liquor law violations per week. For the following questions, assume liquor law violations are independent and follow the Poisson Probability Distribution. In a randomly chosen week:
  - What is the probability that A&M Police dealt with 5 liquor law violations?
    - *PDF(5)*
  - What is the probability that A&M Police dealt with less than 5 liquor law violations?
    - *CDF(4)*
  - What is the probability that A&M Police dealt with more than 5 liquor law violations?
    - $\mathbf{1 - \mathit{CDF}(5)}$

12

EXPONENTIAL EXAMPLE

- Assume the Texas A&M Police deal with an average of 3.6 liquor law violations per week. For the following questions, assume the ***timing*** liquor law violations are independent and follow the Exponential Probability Distribution. =EXPON.DIST(EXPON.DIST(x, lambda, cumulative)
  - What is the probability that it will take less than four days for A&M Police to deal with a liquor law violation?
    0.8722 =EXPON.DIST(4/7,3.6,1)
  - What is the probability that it will be more than 2 days before A&M Police deal with a liquor law violations?
    0.3575 =1-EXPON.DIST(2/7,3.6,1)
  - What is the probability that it will be between 3 and 5 days for A&M Police to deal with a liquor law violation?
    0.1373 =EXPON.DIST(5/7,3.6,1)-EXPON.DIST(3/7,3.6,1)

13

NORMAL EXAMPLE

- In 2021, the annual percentage change in real weekly wages for U.S. Counties had a mean $\mu = -8.6$ and a standard deviation $\sigma = 3.7$ For the following questions, assume the annual percentage change in real weekly wages follow the Normal Probability Distribution. For a randomly chosen County:
  - What is the probability the annual percentage change in real weekly wages was less than -7?

14

$$\mu = -8.6 \quad \sigma = 3.7 \quad P[X < -7]?$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

mu	sigma
-8.6	3.7
Calculate Z	
0.4324	=(-7-(-8.6))/3.7

mu	sigma
-8.6	3.7
Calculate Z	
0.4324	=(-7+8.6)/3.7

Calculate Z	
0.4324	
Probability Less than Z	
0.6673	=NORM.S.DIST(B7,1)

15

$$\mu = -8.6 \quad \sigma = 3.7 \quad P[X > 0]?$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

mu	sigma
-8.6	3.7

X	Z
0	=(E7-E4)/F4

Probability greater than Z	
0.0101	=1-NORM.S.DIST(F7,1)

16

$$\mu = -8.6 \quad \sigma^2 = 13.69 \quad P[X > 0]?$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

mu	sigma^2
-8.6	13.69

X	Z
0	=(E7-E4)/SQRT(F4)

Probability greater than Z	
0.0101	=1-NORM.S.DIST(F7,1)

17

NORMAL APPROXIMATION TO BINOMIAL EXAMPLE

- In major league baseball, 3.9% of all at-bats include a batter breaking their bat. Assume broken bats are independent and follow the Binomial Probability Distribution. Us the Normal Approximation to the Binomial Distribution to answer the following questions. For a random sample of 500 at-bats:
  - What is the probability that the number of broken bats in the sample is 15 or less?

18

$\mu = ? \quad \sigma = ? \quad P[X \leq 15]?$

$$E(X) = \mu = nP$$

$$Var(X) = \sigma^2 = nP(1-P)$$

	A	B
1	P	0.039
2	n	500
3	mus for X	=B1*B2
4	sigma^2 for X	=B2*B1*(1-B1)
5	sigma for X	=SQRT(B4)

	A	B
1	P	0.0390
2	n	500
3	mus for X	19.5
4	sigma^2 for X	18.7395
5	sigma for X	4.3289

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

	A	B	C	D	E
1	P	0.0390			
2	n	500		X	Z
3	mus for X	19.5		15	=(D3-B3)/B5
4	sigma^2 for X	18.7395			
5	sigma for X	4.3289		Probability Less Than Z	
6					0.1493

19

NORMAL APPROXIMATION TO BINOMIAL EXAMPLE

- In major league baseball, 3.9% of all at-bats include a batter breaking their bat. Assume broken bats are independent and follow the Binomial Probability Distribution. Us the Normal Approximation to the Binomial Distribution to answer the following questions. For a random sample of 500 at-bats:
  - What is the probability that the *proportion* of broken bats in the sample is greater than 5%?

20

$\mu = ? \quad \sigma = ? \quad P\left[\frac{X}{n} > 0.05\right]?$

$$E\left(\frac{X}{n}\right) = \mu = P$$

$$Var\left(\frac{X}{n}\right) = \sigma^2 = \frac{P(1-P)}{n}$$

	A	B
1	P	0.039
2	n	500
3	mus for X/n	=B1
4	sigma^2 for X/n	=B1*(1-B1)/B2
5	sigma for X/n	=SQRT(B4)

	A	B
1	P	0.0390
2	n	500
3	mus for X/n	0.0390
4	sigma^2 for X/n	0.000075
5	sigma for X/n	0.0087

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

	A	B	C	D	E
1	P	0.0390			
2	n	500		X	Z
3	mus for X/n	0.0390		0.05	=(D3-B3)/B5
4	sigma^2 for X/n	0.000075			
5	sigma for X/n	0.0087		Probability Greater Than Z	
6					0.1019

21

SAMPLING DISTRIBUTIONS

- A random variable X is distributed with a constant mean μ = 23 and variance $\sigma^2 = 400$
- A random sample of size n = 36 is obtained. What is the probability of finding a sample mean $\overline{X} \leq 17$?
- Find the Z-Score for $\overline{X} = 17$

$$Z_0 = \frac{\overline{X}_0 - \mu}{\sqrt{\sigma^2/n}} = \frac{17 - 23}{\sqrt{400/36}} = -1.8$$

- In Excel, use =NORM.S.DIST(-1.8,1) to get $P[Z \leq Z_0] \cong 0.036$

22

SAMPLING DIST. OF THE SAMPLE VARIANCE S^2

- The sampling distribution of the sample variance S^2 is then

$$E(S^2) = \sigma^2$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

23

EXAMPLE

- The random variable X is known to follow the Normal distribution with ***standard deviation σ = 16***
- In a random sample of n = 90, what is the probability of finding a sample variance $S^2 \leq 200$
- Calculate $$\chi^2_{calc} = \frac{(n-1)S^2}{\sigma^2} = \frac{(90-1)200}{16^2} \cong 69.531$$
- In Excel, use =CHISQ.DIST(chi2_calc,n-1,1) to get

$$P[S^2 \leq 200] = P\left[\chi^2_{(n-1)} \leq \chi^2_{calc}\right] \cong 0.0629$$

- To get the probability $S^2 \geq 200$, use
  - =1-CHISQ.DIST(chi2_calc,n-1,1) or
  - =CHISQ.DIST.RT(chi2_calc,n-1)

24