Pre-Orientation Review Session

ENV710 Applied Data Analysis for Environmental Science

17 August 2017

Elizabeth A. Albright, Ph.D.
Assistant Professor of the Practice

Outline for Today

- Introductions
- Overview of diagnostic exam
- Review/Practice Problems

Overview of Diagnostic

- 20 questions
- One hour and 15 minutes
- No calculators
- No credit for work w/o correct answer
- Z-Distribution table will be supplied

Potential Topics

- Basic math and algebra
- Descriptive statistics
- Probability
- Sampling
- Inference
  - Confidence intervals
  - Comparison of means
  - Type I and Type II errors

The Statistics Review Website

http://sites.nicholas.duke.edu/statsreview

Basic Math

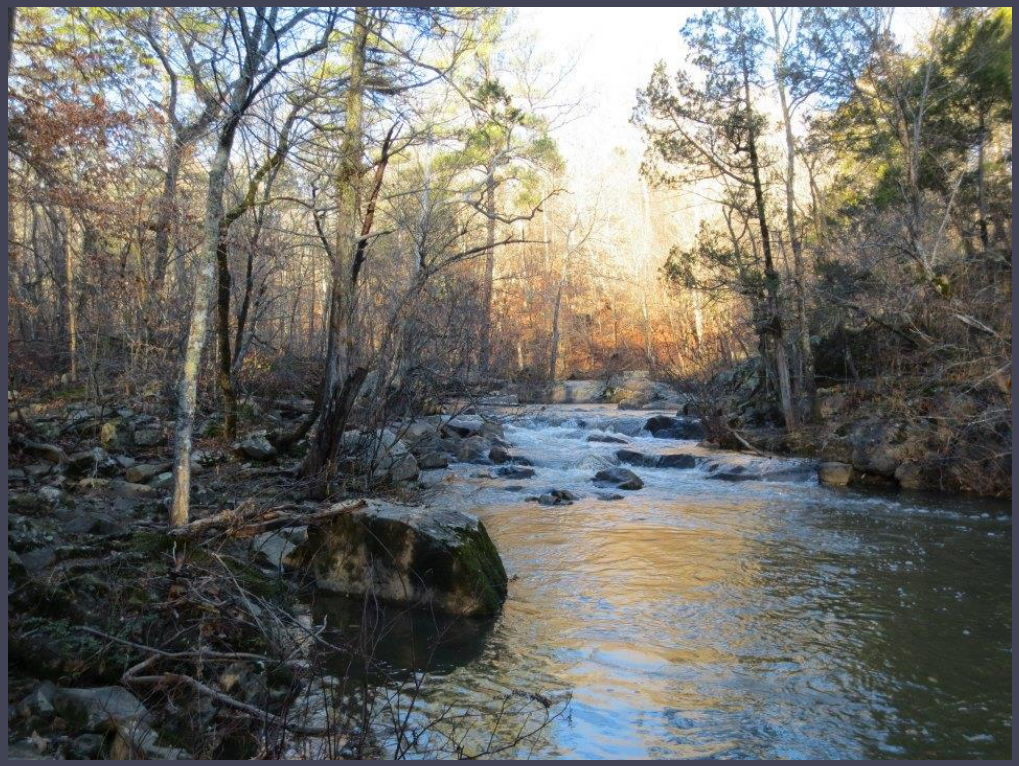
- Rounding/Significant digits
- Algebra
- Exponents and their rules
- Logarithms and their rules

Basic Math Practice Problems

- 0.306 contains how many significant digits?
- $3^6 * 3^2 = ?$
- $\log_{10}(8) - \log_{10}(2) = ?$
- Simplify: $(x^4x^{-2})^{-3}$
- Simplify: 6!/2!

Basic Math Solutions

- 0.306 contains three significant digits
- $3^6 * 3^2 = 3^8$
- $\log_{10}(8) - \log_{10}(2) = \log_{10}(4)$
- Simplify: $(x^4x^{-2})^{-3}=(x^2)^{-3} = x^{-6}$
- Simplify: $6!/2! = (6*5*4*3*2*1)/(2*1)=720/2=360$



Descriptive Statistics

Descriptive Statistics

- Measure of central tendency
  - Mean
  - Median
  - Mode
- Measure of spread
  - Standard deviation
  - Variance
  - IQR
  - Range
- Skewness
- Outliers

Question of Interest

Do Nicholas or Fuqua faculty members have larger transportation carbon footprints?

The Steps

- Design the study
  - Random sampling
- Collect the data
- **Describe the data**
- Infer from the samples to the populations

CO2 Emissions (metric tons) from Transportation Sources for 10 Randomly Selected NSOE Faculty

7	1
2	4
2	8
7	15
2	2

Measure of Central Tendency

- Mean = 5 metric tons CO2
- Median = 3 metric tons CO2
- Mode = 2 metric tons CO2

The Mean (Expected Value)

$$\bar{x} = 1/n \sum_{i=1}^{n} x_i$$

Median

- If odd number of observations: middle value (50^{th} percentile)

- If even number of observations: halfway between the middle two values

Spread of a distribution

- **Range**: 15-1 = 14 metric tons CO2
  - Largest observation minus smallest observation
- **Variance** =
  - 18.9 metric tons ^{2}
- **Standard Deviation**
  - s=4.3 metric tons

VARIANCE

$$\text{variance}\,(s^2) = \frac{1}{n-1}\sum(\text{X} - \text{Xbar})^2$$

Probability

Random Variable

- A variable whose value is a function of a random process
  - Discrete
  - Continuous
- If X is a random variable, then p(X=x) is the probability that the the value x will occur

Which of the following is a discrete random variable?

I. The height of a randomly selected MEM student.

II. The annual number of lottery winners from Durham.

III. The number of presidential elections in the United States in the 20th century.

(A) I only (B) II only (C) III only (D) I and II (E) II and III

Properties of Probability

- The events A and B are mutually exclusive if they have no outcomes in common and so can never occur together.

- If A and B are **mutually exclusive** then

$$P(A \text{ or } B) = P(A) + P(B)$$

Example: Roll a die.
What's the probability of getting a 1 or a 2?

P(A or B)

What if events A and B are not mutually exclusive?

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Deck of Cards

P(A or B)

Example: What's the probability of pulling a black card or a ten from a deck of cards?

P(A or B)

Example: What's the probability of pulling a black card or a ten from a deck of cards?

P(black) = 26/52 P(10) = 4/52

Probability of a black card OR a ten =

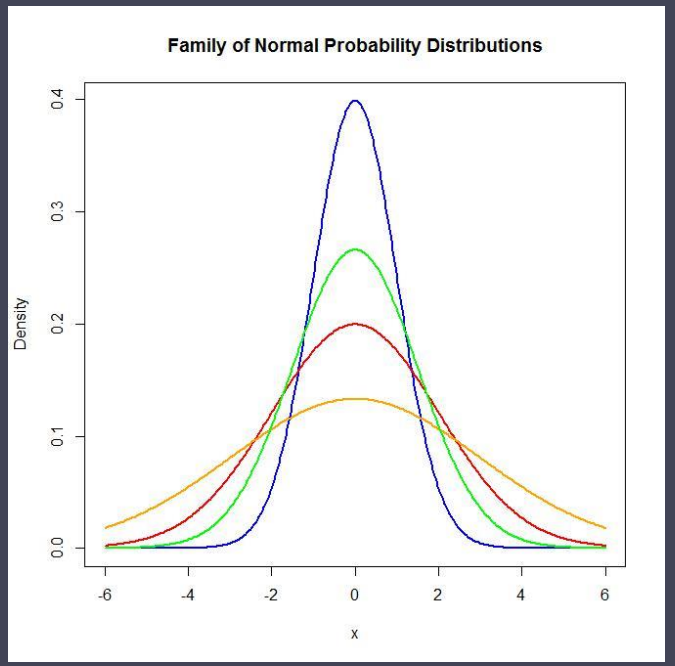
26/52 + 4/52 – 2/52 = 28/52

P(A AND B)

$$p(A \text{ and } B) = p(A) * p(B)$$

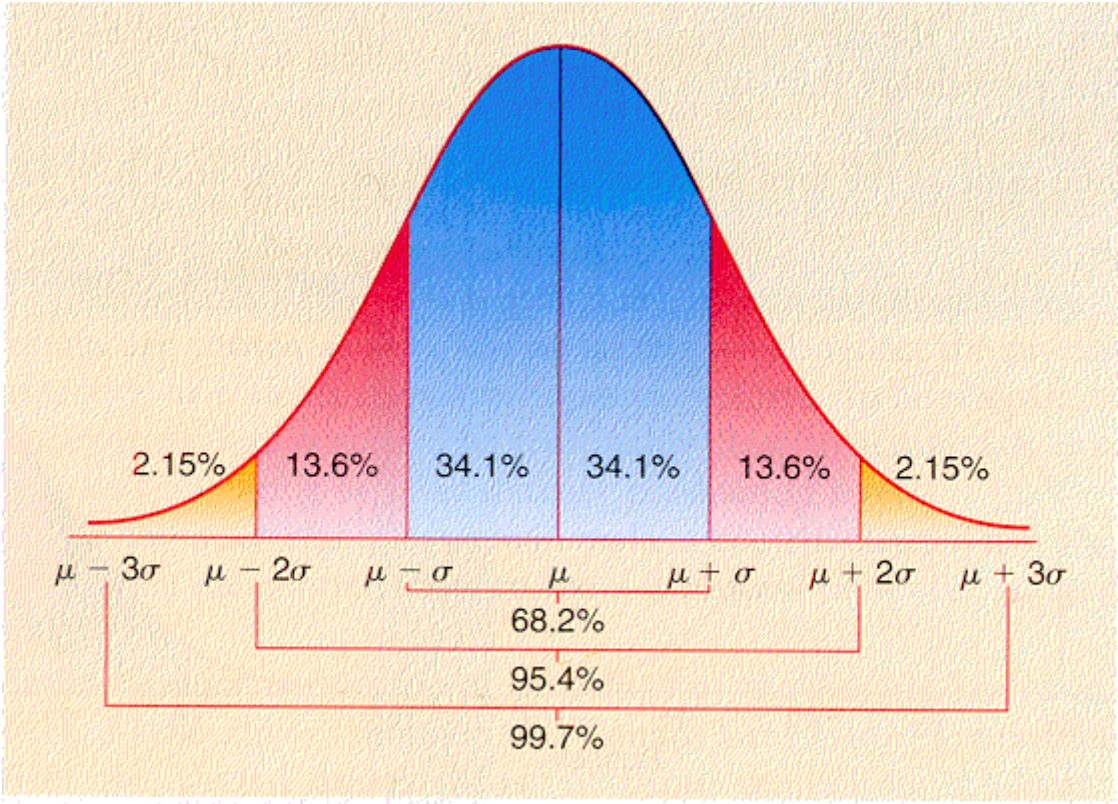
- Two consecutive flips of a coin, A and B
- A = [heads on first flip]
- B = [heads on second flip]

- p(A and B) = ???
- p(A and B) = ½ * ½ = 1/4



The Normal Distribution

The Normal Distribution



Normal Distribution (2012) Last accessed September, 2012 from http://www.comfsm.fm/~dleeling/statistics/notes06.html.

Table entry for z is the area under the standard normal curve to the left of z.

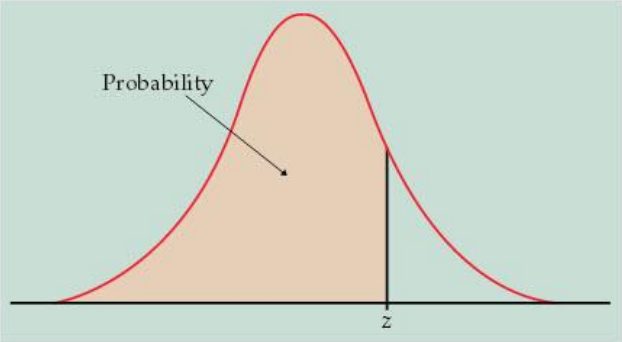


TABLE A Standard normal probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

Z Score

- How do you convert any normal curve to the standard normal curve?

$$Z = \frac{X - \mu}{\sigma}$$

Normal Distribution Calculations

If X is normally distributed around a mean of 32 and a standard deviation of 8, find:

a. $p(X>32)$
b. $p(X>48)$
c. $p(X<24)$
d. $p(40<X<48)$

Solutions

a. $p(X>32) = p(z>0) = 0.5$

b. $p(X>48) = p(z>2) = 0.0228$

c. $p(X<24) = p(z<-1) = 0.1587$

d. $p(40<X<48) = p(1<z<2) =$

$$0.1587 - 0.0228 = 0.136$$

Normal Distribution Practice Problem

- The crop yield is typically measured as the amount of the crop produced per acre. For example, cotton is measured in pounds per acre. It has been demonstrated that the normal distribution can be used to characterize crop yields.

- Historical data suggest that the probability distribution of next summer's cotton yield for a particular North Carolina farm can be characterized by a normal distribution with mean 1,500 pounds per acres and standard deviation 250. The farm in question will be profitable if it produces at least 1,600 pounds per acre.

- What is the probability that the farm will lose money next summer?

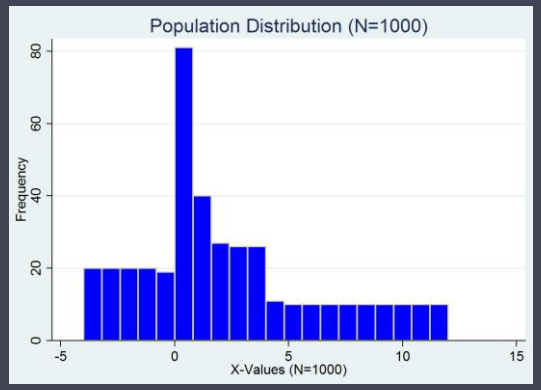
Normal Distribution Practice Problem

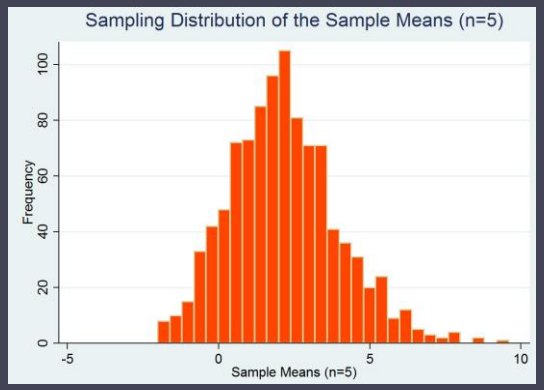
Historical data suggest that the probability distribution of next summer's cotton yield for a particular North Carolina farm can be characterized by a normal distribution with mean 1,500 pounds per acres and standard deviation 250. The farm in question will be profitable if it produces at least 1,600 pounds per acre.

- What is the probability that the farm will lose money next summer?

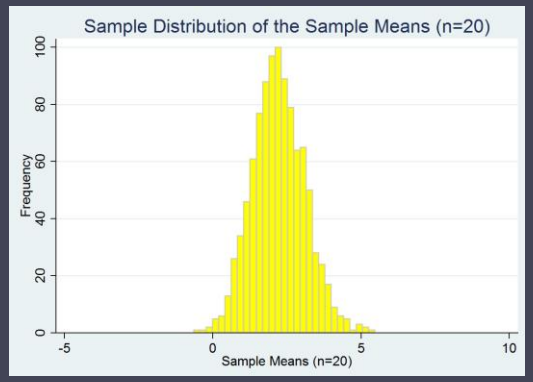
$$z = \frac{x - \mu}{\sigma} = \frac{1600 - 1500}{250} = \frac{100}{250} = 0.4$$

p(lose money) = p(z<0.4) = 0.655





Sampling and the Central Limit theorem

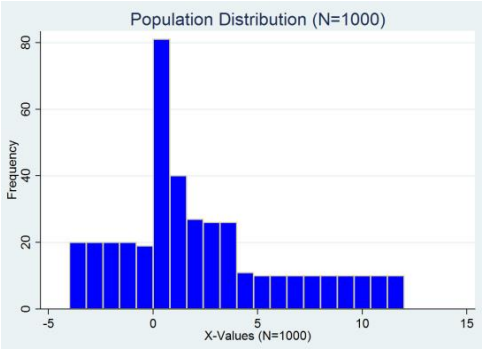


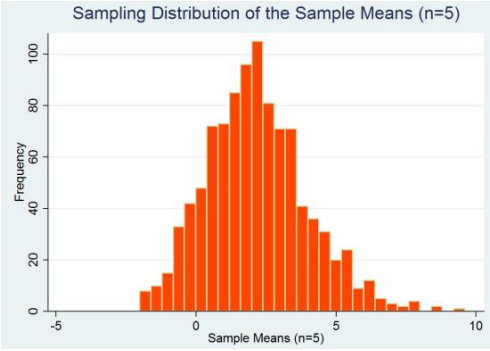
Sampling

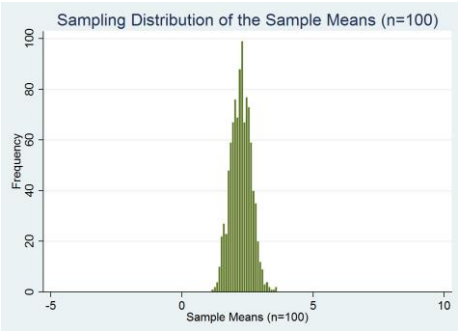
- Why do we sample?
- In simple random sampling every unit in the population has an equal probability of being sampled.
- Sampling error
  - Samples will vary because of the random process

Central Limit Theorem

As the size of a sampling distribution increases, the sampling distribution of X_{bar} concentrates more and more around μ. The shape of the distribution also gets closer and closer to normal.







population n=5 n=100

Profundity of Central Limit Theorem

- As sample size gets larger, even if you start with a non-normal distribution, the sampling distribution approaches a normal distribution

Sampling Distribution of the Sample Means

- Mean of the sample means
- Standard Error
  - Standard deviation of the sampling distribution of sample means

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

SE vs. SD

- What is the difference between standard deviation and standard error?
  - SD is the typical deviation from the average. SD does not depend on random sampling.
  - SE is the typical deviation from the expected value in a random sample. SE results from random sampling.

INFERENCE….

Inference

- We infer from a sample to a population.
- Need to take into account sampling error.

- Confidence intervals
- Comparison of means tests

Confidence Interval with KNOWN standard deviation

- Let's construct a 95% confidence interval

$$(X_{bar}\text{-}1.96*SE < \mu < X_{bar} + 1.96*SE)$$

- Where did I get the 1.96 (the multiplier)?
- Very important!!! *It is the confidence interval that varies, not the population mean.*

CI Practice Problem

We want to construct a 95% confidence interval around the mean number of hours that Nicholas MEM students (who are enrolled in statistics) spend studying statistics each week. We randomly sample 36 students and find that the average study time is eight hours. The standard deviation of study time of the population of all students in statistics is 2 hours.

Calculate the 95% confidence interval of the mean study time.

How do you interpret the confidence interval?

Confidence Interval Solution

- (X_{bar}-1.96*SE < μ <X_{bar} + 1.96*SE)
- Xbar = 8 hours
- σ = 2 hours
- SE = 2/sqrt(36) = 2/6 = 0.333
- (8 – 1.96*0.333 < μ < 8 + 1.96 * 0.333)
- (7.35 hours < μ < 8.65 hours)

We are 95% confident that the interval (7.35 hrs, 8.65 hrs) covers the true average number of hours MEM students spend studying statistics.

Comparison of Means Tests

- One sample
  - Is the average dissolved oxygen concentration less than 5mg/L?
- Two independent samples
  - Do residents of North Carolina spend more on organic food than residents of South Carolina?
- Matched/Pairs/Repeated samples
  - Are individuals' left hands larger than their right hands?

One-Sample Hypothesis Testing Approach

- Set up a 'null hypothesis', (typically hypothesizing there is no difference between the population mean and a given value)
- Establish an alternative hypothesis (that there is a difference between the population mean and a given value)
- Calculate sample mean, standard deviation, standard error
- Calculate a the test statistic and a p-value
- The smaller the p-value, the more statistically significant results
- Interpret results

Test Statistic

- z vs. t test statistic
  - Z: known population standard deviation or large sample size
  - t: used when estimating standard deviation of population with the standard deviation of the sample

P-Values

- P-value = the probability of getting the sample statistic as least as extreme as what was observed, assuming that the null hypothesis is true.

- The smaller the p-value, the more evidence there is AGAINST the null hypothesis.

Are these new light bulbs better?

A standard manufacturing process has produced millions of light bulbs, with a mean life of 1200 hours. A new process, recommended by the USEPA, produces a sample of 25 bulbs, with an average of 1265 hours (standard deviation of the population of light bulbs is 300 hours). Although this sample makes the new process look better, is this just a sampling fluke? Is it possible that the new process is really no better than the old?

Solution

Set up hypotheses (μ_o=1200 hours)

Null Hypothesis: $\mu \leq 1200$ hours

Alternative Hypothesis: $\mu > 1200$ hours

Solution Continued

$$z = \frac{\bar{x} - \mu_o}{\sigma_{\bar{x}}}$$

$$z = \frac{1265 - 1200}{300/\sqrt{25}} = \frac{65}{60} = 1.08$$

Solution

- Now we need to calculate a p-value from our z-statistic.
- $P(Z>1.08) = 0.14$. This is our p-value.
- Assuming that our null hypothesis is true, there is 0.14 probability of getting a test statistic as extreme or more extreme than we observed.
- A p-value of 0.14 does NOT provide strong evidence against the null. We can NOT conclude that the new bulbs last longer than the old bulbs.

QUESTIONS?