COMP 7036
APPLIED RESEARCH METHODS IN SOFTWARE DEVELOPMENT

Borna Noureddin, Ph.D.
British Columbia Institute of Technology

Statistics

OVERVIEW

- Nomenclature
- Simple statistics
- Correlation coefficient
- Confidence intervals
- Significance testing

SOME NOMENCLATURE

- A **population** includes each element from the set of observations that can be made.
- A **sample** consists only of observations drawn from the population.

- **Parameter**: measurable characteristic of population (e.g., mean or standard deviation)
- **Statistic**: measurable characteristic of sample

SOME NOMENCLATURE

- **Percentile:** values that divide a rank-ordered set of elements into 100 equal parts
- **Standard score** (**z-score**): how many standard deviations an element is from mean

$$z = (X - \mu) / \sigma$$

z: z-score

X: value of element

μ: mean of population

σ: standard deviation

SIMPLE STATISTICS

Mean: the average of a set of numbers

Median: The number found in the middle when looking at the set of numbers from smallest to largest

Mode: most commonly occurring value in a set of numbers

Variance: a measure of how data points differ from the mean

Standard Deviation: square root of variance

Confidence interval

CORRELATION COEFFICIENT

Measure strength of association between 2 variables

- Usually use Pearson product-moment correlation coefficient
- The value of a correlation coefficient ranges between -1 and 1.
- The greater the absolute value of a correlation coefficient, the stronger the linear relationship.
- The strongest linear relationship is indicated by a correlation coefficient of -1 or 1.

CORRELATION COEFFICIENT

Measure strength of association between 2 variables

- The weakest linear relationship is indicated by a correlation coefficient equal to 0.
- A positive correlation means that if one variable gets bigger, the other variable tends to get bigger.
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller.

CONFIDENCE INTERVALS

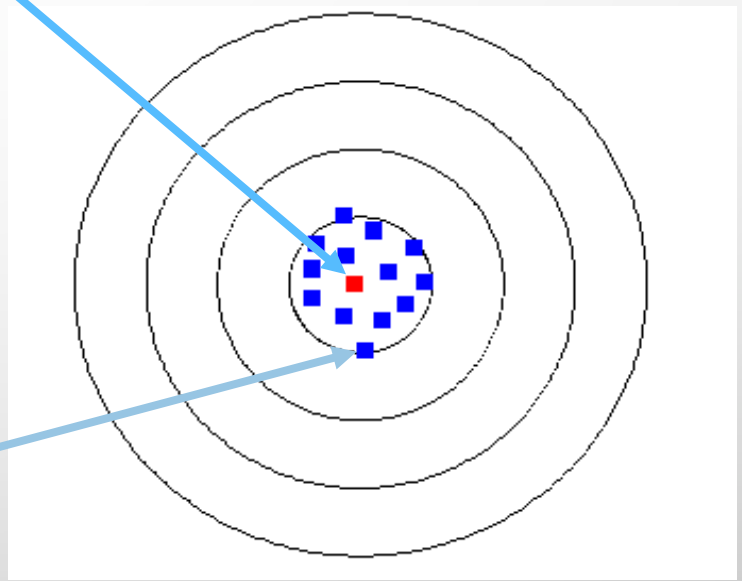
- Statistic estimates population parameter
  - Average salary
  - Mean latency
  - Average FPS
  - Average sales

- Unbiased estimate – on average statistic equals parameter of interest

CONFIDENCE INTERVALS

Unbiased and small variance.

- The center is the target (the parameter of interest).
- The blue points are statistics. Each one is based on a different sample. On average, they equal the target.



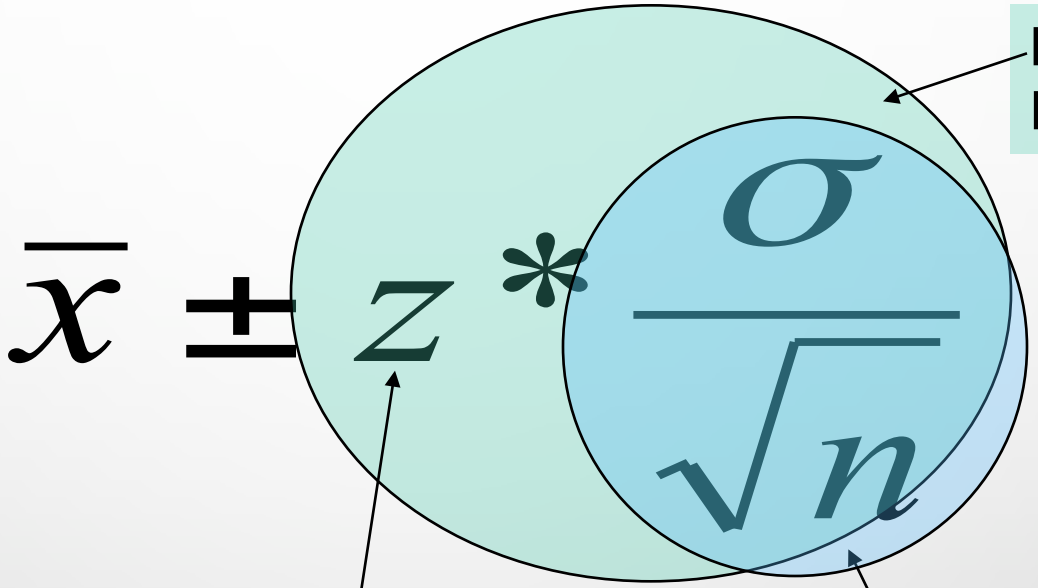
CONFIDENCE INTERVALS

- CONFIDENCE INTERVALS USE:
  - the statistic (mean) as a point estimate of the parameter.
  - the standard deviation of the distribution to help provide a sense of accuracy.
  - a measure of confidence in the effectiveness of the interval capturing the true mean (parameter).

CONFIDENCE INTERVALS

- CONSTRUCTING A CONFIDENCE INTERVAL
  - First find the mean from your sample.
  - Then, find your margin of error based upon your confidence level (z*) and sample size (n).
  - The population standard deviation, for now, will be assumed to be known.

CONFIDENCE INTERVALS



Margin of Error

An estimate for μ from your sample

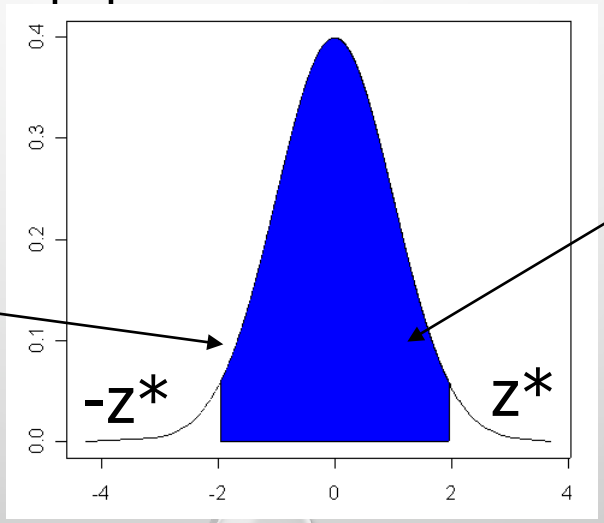
Based on level of confidence, how many standard deviations away you can tolerate

The standard deviation of the estimate.

CONFIDENCE INTERVALS

- z* (also called the critical value) is the score corresponding to the confidence level you want
- A 95% confidence level means that 95% of intervals of sample size n will capture (or contain) the population mean.
  - Find z* using the picture and a table or InvNorm on the calculator
- A 99% confidence level means that 99% of intervals of sample size n will capture (or contain) the population mean.

Standard Normal Distribution



The blue area is 95%

CONFIDENCE INTERVALS

- Confidence *Level* is how confident we want to be that the confidence interval WILL contain the parameter of interest.

Confidence Level	z^*
80%	1.282
90%	1.645
95%	1.96
98%	2.326
99%	2.576

EXAMPLE PROBLEM #1

- Based on a sample of 40 cars of a particular model, the fuel tank capacity is calculated for each. Based on this data, the sample mean is 18.92 gallons. The population standard deviation is to be 3.5 (remember this is the unrealistic part – we would not really know this).
- Construct a 95% confidence interval for the mean fuel capacity of this model of car.

EXAMPLE PROBLEM #1

- 95% of all possible confidence intervals for a sample size of 40 will contain the population mean
- We are 95% confident that the mean fuel capacity is between 17.8 and 20.0 gallons.

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

$$18.92 \pm 1.960 \frac{3.5}{\sqrt{40}}$$

$$18.92 \pm 1.0847$$

$$(17.8353, 20.0047)$$

EXAMPLE PROBLEM #1

- In a previous example, we stated that "We are 95% confident that the mean fuel capacity is between 17.8 and 20.0 gallons."
- Does this mean that

$$P(17.8 < \mu < 20.0) = 0.95?$$

The answer is NO.

EXAMPLE PROBLEM #1

- Suppose that the meteorologist forecasts that there is a 10% chance that it will rain this weekend.
- After Monday comes, does that mean there's a 10% chance that it *rained* that past weekend? No, either it rained or it didn't. So either there is a 0% or a 100% chance that it *rained* over the past weekend.

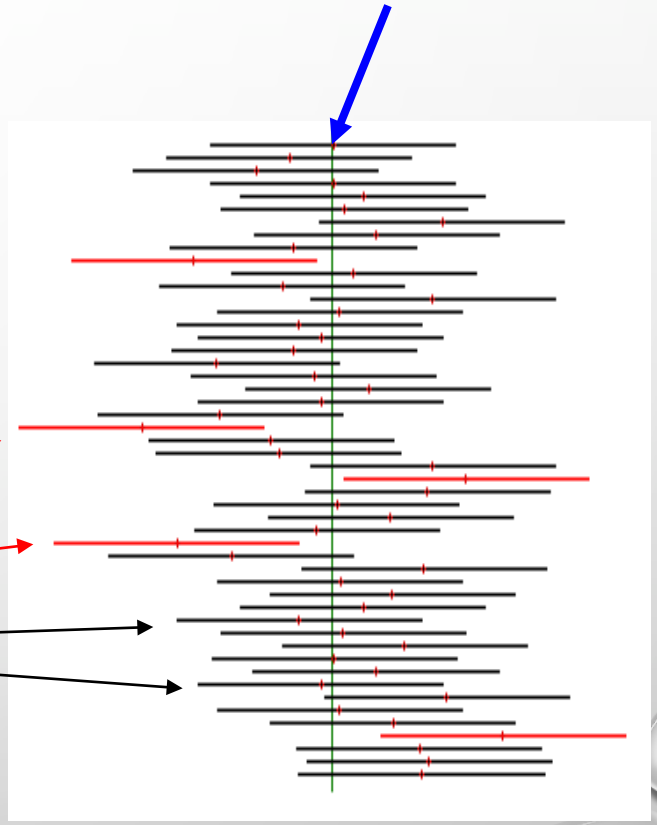
EXAMPLE PROBLEM #1

- BEFORE you collect the data, there's a 95% chance that μ will be in the interval but …
- AFTER you collect the data, either μ is in the interval or it isn't so the probability is either 0 or 1.
- The fact that we don't know what μ is doesn't change this.

EXAMPLE PROBLEM #1

The target (μ). If an interval crosses this vertical line, then μ is in the confidence interval.

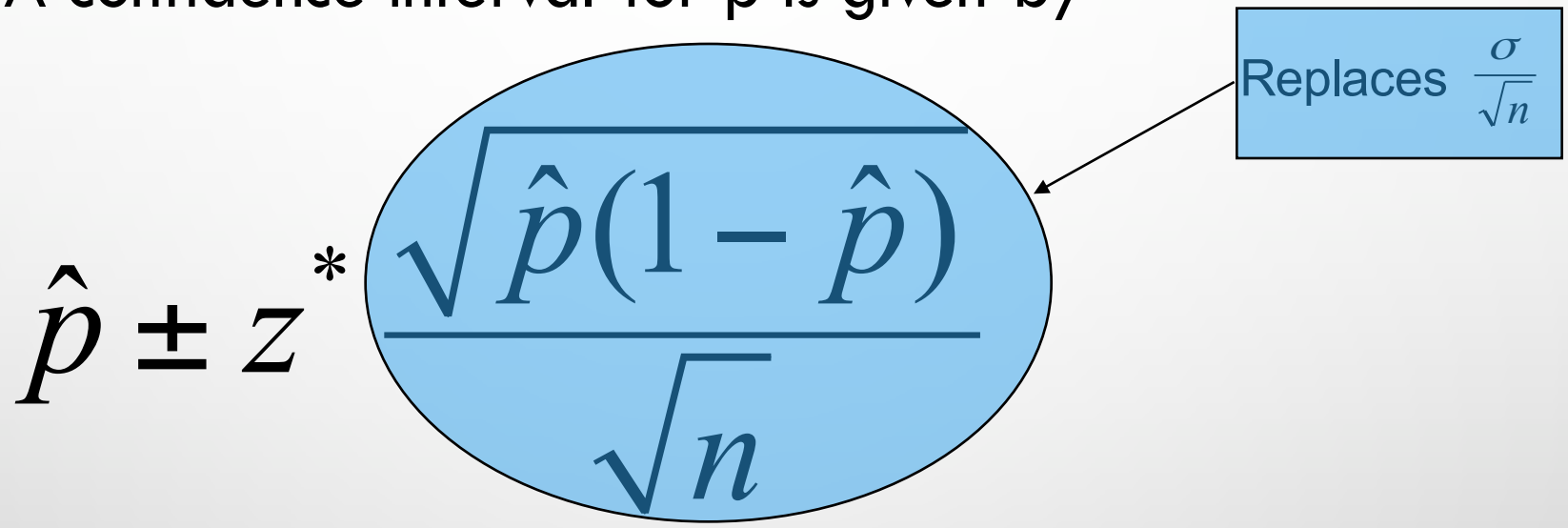
50 confidence intervals are obtained. Each one is a 90% confidence interval based on a sample of size n from the same population. We EXPECT approximately 90% of these to contain μ.



The red intervals don't contain μ but the black ones do.

CONFIDENCE INTERVALS FOR PROPORTIONS

- A confidence interval for p is given by



This is the same formula, just with the standard deviation for proportions in place of the standard deviation for means

EXAMPLE PROBLEM #2

- Suppose that a sample of 100 light bulbs that a plant produces is chosen at random. In the sample, there are 6 defective bulbs. Obtain a 98% confidence interval for the proportion of defective bulbs they produce.

EXAMPLE PROBLEM #2

- Remember that the confidence interval we're using is based on the normal distribution. We should check the rule of thumb to determine if we can use it or not. We can use it since

$$1.\ n\hat{p} = 100(6/100) = 6 \geq 5$$

$$2\ n(1-\hat{p}) = 100(1 - 6/100) = 94 \geq 5$$

EXAMPLE PROBLEM #2

- A 98% confidence interval for the proportion of defective bulbs produced is

$$\hat{p} \pm z^* \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

$$= 0.06 \pm 2.326 \frac{\sqrt{0.06(1-0.06)}}{\sqrt{100}}$$

$$= 0.06 \pm 0.0552$$

$$= (0.0047, 0.1152)$$

EXAMPLE PROBLEM #2

- Given the interval (0.0047,0.1152), we expect somewhere between 0.5% and 11.5% of the bulbs to be defective.

SIGNIFICANCE TESTING

A bottling company claims to put 12 fluid ounces of soda into every can with a standard deviation of $\sigma=0.18$. We suspect they are short changing us. So, we take a sample of 24 cans and measure the mean amount of soda in the cans. We find the mean amount of soda to be 11.92 fluid ounces. Is this enough evidence to suggest that the company is not being truthful?

SIGNIFICANCE TESTING

- Assume company innocent until proven guilty
  - Assume there is 12 fluid ounces of soda in each can
  - Take a sample (24 cans) and measure (it came out with a mean of 11.92 fluid ounces)
  - Determine the probability of getting a mean of 11.92 fluid ounces provided that our assumption is true (company is telling the truth)
- This essentially amounts to following the scientific method

SIGNIFICANCE TESTING

1) Write hypotheses
   - Null hypothesis (Assume innocence)
     - The mean amount of soda in a can is 12 fluid ounces.
   - Alternate hypothesis (the crime they are charged with)
     - The mean amount of soda in a can is less than 12 fluid ounces

2) Calculations
   - Test Statistic
     - A measure of how many standard deviations away from the mean the sample was (a z-score)
   - P-value
     - The probability of the sample mean occurring given the null hypothesis is true

4) Conclusions
   - Statistical conclusion
     - Reject the null hypothesis (in favor of the alternate) or Fail to reject the null hypothesis
   - Contextual conclusion
     - State your conclusion in the context of the problem

SIGNIFICANCE TESTING

- The null hypothesis is denoted H_o.
- This is pronounced "H – not" or "h sub-o".
- The null hypothesis is a belief about the world, or population.
  - It is typically presuming innocence.
  - It is typically no change or things are as they should be.
  - It is typically what we are trying to show is false.

SIGNIFICANCE TESTING

- The alternative hypothesis is denoted H_a.
- This hypothesis is pronounced "H-sub a" or alternative hypothesis or alternate hypothesis.
  - This is typically your (or the study's) suspicion.
  - This is typically a belief about the population that you believe to be true.
- There are only 3 possible alternative hypotheses
  - $H_a : \mu < \mu_0$ *The mean is less than the null hypothesis*
  - $H_a : \mu > \mu_0$ *The mean is greater than the null hypothesis*
  - $H_a : \mu \neq \mu_0$ *The mean is different than the null hypothesis*

SIGNIFICANCE TESTING

- The test statistic is a convenient summary of the sample data that can be easily used to make decisions about the hypotheses.
- It is often some kind of transform of the data to a more convenient form, like Z-scores.
- Test statistic is calculated under assumption the H_o is true.
- It can be thought of as how many standard deviations away from the mean the sample was.

SIGNIFICANCE TESTING

- A p-value is a probability, so it is a number that is between 0 and 1.
- It is a measure of how consistent the sample data is with the null hypothesis.
- There are two definitions of p-value that will be useful to you:
  - The p-value is the probability of observing a value of the test statistic as extreme or more extreme than the one observed, if the H_o is true.
  - The p-value is the probability of observing data like ours if the null hypothesis is true.

SIGNIFICANCE TESTING

- Most of scientific world uses .05 as a standard p-value or cut-off point.
- P-values less than .05 are considered small chances and those above .05 are large.
- This gets ridiculous when p-values = .0501 or .049999. P-values are in shades of gray, not black and white.

SIGNIFICANCE TESTING

- The conclusion gives our final opinion about what the data tells us about our hypotheses.
  - Was the evidence of the data enough to convince us that the null hypothesis is not plausible?
  - Or…was there not enough evidence to refute the null hypothesis?

EXAMPLE PROBLEM #3

- For the null hypothesis, assume the company is innocent and really does put 12 fluid ounces in each can.
  - $H_o : \mu=12$
- For the alternative hypothesis, we think they are not giving us enough soda, so…
  - $H_a : \mu<12$

EXAMPLE PROBLEM #3

- For the calculations, we use the sampling distribution based upon the null hypothesis and a sample size of 24 cans.
  - Assuming normality (normal population, sample size and Central Limit Theorem), the sampling distribution should be N:

$$N\left(12, \frac{0.18}{\sqrt{24}}\right)$$

$$N(12, 0.0367)$$

EXAMPLE PROBLEM #3

- Our sample had a mean of 11.92 fluid ounces.
  - Find the test statistic (remember a z-score).

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

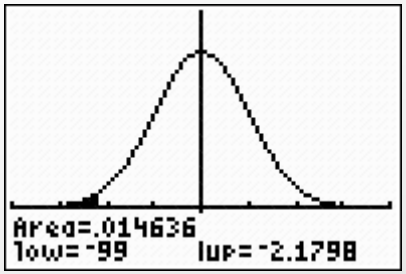
$$z = \frac{11.92 - 12}{0.0367}$$

$$z = -2.1798$$

EXAMPLE PROBLEM #3

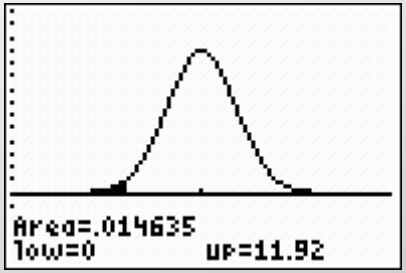
- Our sample had a mean of 11.92 fluid ounces.
  - Find the P-value (the probability of this happening given the null hypothesis is true).

$$p = normalcdf(-99,-2.1798) = 0.0146$$ ⟶



or

$$p = normalcdf(0,11.92,12,.0367) = 0.0146$$ ⟶



EXAMPLE PROBLEM #3

- Interpret the P-value
  - There is only a 1.46% chance of us getting the sample we did given that company was telling the truth.
- Generally speaking, if $p<.05$ or $p<5\%$, it is considered enough evidence to reject the null hypothesis.

EXAMPLE PROBLEM #3

- Since .0146<.05, we have enough evidence!
- Statistical conclusion
  - Reject the null hypothesis in favor of the alternative hypothesis
- Contextual conclusion
  - There is sufficient evidence to suggest that company is putting less than 12 fluid ounces of soda into their cans.

EXAMPLE PROBLEM #4

Another bottling company claims claims that one of their energy drinks has 40 mg of sodium per can. We work for a consumer organization that tests such claims. We take a random sample of 50 cans and find that the mean amount of sodium in the sample is 41.4 mg. The standard deviation in all cans is 7.2 mg. We suspect that there is more than 40 mg of sodium per can.

EXAMPLE PROBLEM #4

- For the null hypothesis, assume the company is innocent and the soda really does contain 40 mg of sodium in each can.
  - $H_o : \mu=40$
- For the alternative hypothesis, we think there may be more than 40 mg of sodium, so…
  - $H_a : \mu>40$

EXAMPLE PROBLEM #4

- For the calculations, we use the sampling distribution based upon the null hypothesis and a sample size of 50 cans.
  - Assuming normality (normal population, sample size and Central Limit Theorem), the sampling distribution should be N:

$$N\left(40, \frac{7.2}{\sqrt{50}}\right)$$

$$N(40, 1.0182)$$

EXAMPLE PROBLEM #4

- Our sample had a mean of 42.6 mg of sodium.
  - Find the test statistic (remember a z-score).

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

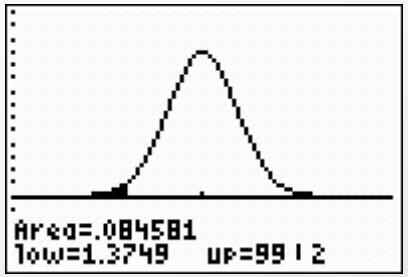
$$z = \frac{41.4 - 40}{1.0182}$$

$$z = 1.3749$$

EXAMPLE PROBLEM #4

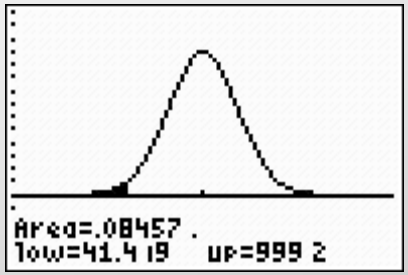
- Our sample had a mean of 42.6 mg of sodium.
  - Find the P-value (the probability of this happening given the null hypothesis is true).

$$p = normalcdf(1.3749, 99) = 0.0846$$



or

$$p = normalcdf(41.4, 999, 40, 1.0182) = 0.0846$$



EXAMPLE PROBLEM #4

- Interpret the P-value
  - There is an 8.46% chance of us getting the sample we did given that the company was telling the truth.
- Generally speaking, if $p<.05$ or $p<5\%$, it is considered enough evidence to reject the null hypothesis.

EXAMPLE PROBLEM #4

- Since .0846>.05, we do not have enough evidence!
- Statistical conclusion
  - Fail to reject the null hypothesis in favor of the alternative hypothesis
- Contextual conclusion
  - There is not sufficient evidence to suggest that the energy drink has more than 40 mg of sodium in a can of soda.

EXAMPLE PROBLEM #5

Let's suppose that a computer shop typically sell 3400 items of your company's products per week (with a standard deviation of 500). You then begin a new marketing campaign complete with a new slogan and mass marketing strategy. You are curious how the new advertising will affect sales (to see if sales increased or decreased). You sample 40 random stores for a week and find that the mean number of items the stores sell is 3550.

EXAMPLE PROBLEM #5

- For the null hypothesis, assume sales do not change and the stores still sell 3400 items per week.
  - H_o : $\mu=3400$
- For the alternative hypothesis, we want to see if there is a change (for better or worse), so…
  - H_a : $\mu\neq3400$

EXAMPLE PROBLEM #5

- For the calculations, we use the sampling distribution based upon the null hypothesis and a sample size of 40 stores.
  - Assuming normality (normal population, sample size and Central Limit Theorem), the sampling distribution should be N:

$$N\left(3400, \frac{500}{\sqrt{40}}\right)$$

$$N(3400, 79.0569)$$

EXAMPLE PROBLEM #5

- Our sample had a mean of 3550 items.
  - Find the test statistic (remember a z-score).

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

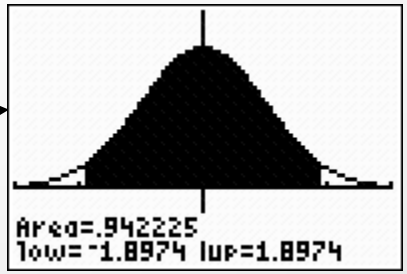
$$z = \frac{3550 - 3400}{79.0569}$$

$$z = 1.8974$$

EXAMPLE PROBLEM #5

- Our sample had a mean of 3550 items.
  - Find the P-value (the probability of this happening given the null hypothesis is true).

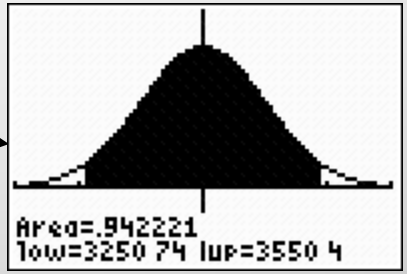
$$p = 2normalcdf(1.8974,99) = 0.0578$$



It is the un-shaded area that we want

or

$$p = 2normalcdf(3550,99999,3400,79.0569) = 0.0578$$



EXAMPLE PROBLEM #5

- Interpret the P-value
  - There is an 5.78% chance of us getting the sample we did given that the stores typically sell 3400 items per week.
- Generally speaking, if $p<.05$ or $p<5\%$, it is considered enough evidence to reject the null hypothesis.

EXAMPLE PROBLEM #5

- Since .0578>.05, we do not have enough evidence!
- Statistical conclusion
  - Fail to reject the null hypothesis in favor of the alternative hypothesis
- Contextual conclusion
  - There is not sufficient evidence to suggest that the marketing campaign had an impact on the sales of your products.

SUMMARY

- Nomenclature
- Simple statistics
- Correlation coefficient
- Confidence intervals
- Significance testing