Data Analytics

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA

School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

TA Information

- Nastaran Ghane <nghane@hawk.iit.edu>
- Her office hours
  Tuesday/Thursday
  1:00 pm - 2:00 pm
  Perlstein Hall, room 223

Schedule

- Course Structure
- Quick Reviews
- Use Sample to Estimate Population
- One-Sample Hypothesis Testing
- Two-Sample Hypothesis Testing

Schedule

- Course Structure
- Quick Reviews
- Use Sample to Estimate Population
- One-Sample Hypothesis Testing
- Two-Sample Hypothesis Testing

Course Structure

- Descriptive Statistics
  - Data Types
  - Descriptive Statistics for Nominal and Numerical vars
- Inferential Statistics
  - Use sample to estimate population
  - Hypothesis Testing
  - ANOVA
  - Predictive Models
    - Linear Regression
    - Classification

Schedule

- Quick Reviews
  - Statistical Applications
  - Data: Population and Sample
  - Data Types
  - Descriptive Statistics
    - For nominal variables
      - By metrics
      - By visualizations (note: be able to interpret plots)
    - For numerical variables
      - By metrics
      - By visualizations(note: be able to interpret plots)
    - Using R for descriptive statistics

Schedule

- Course Structure
- Quick Reviews
- Use Sample to Estimate Population
- One-Sample Hypothesis Testing
- Two-Sample Hypothesis Testing

Statistical Inference

- There are two ways for us to estimate or infer the population parameter, such as population mean:
  1) By estimating its value
     For example: estimate the age of students in IIT
  2) By testing hypothesis about its value
     For example:
     Method-1 is better than method 2.
     Students in 527(04) are better than 527(01).
     The average of working hours/day is no more than 8

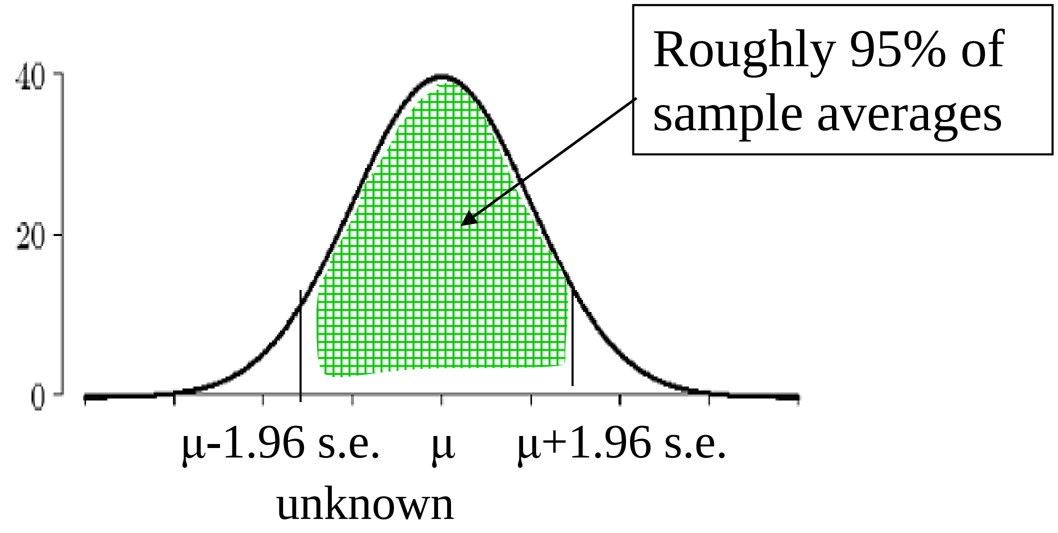
Statistical Inference by Estimating Population Mean

- You can follow these steps
  1) Collect sample, and calculate descriptive statistics
  2) The sample mean is assumed to be normal (n>=30) distributed and centered as population mean
  3) The standard error of the sampling distribution is expected to be as small as possible. Note: usually it becomes smaller if your n is larger
  4) Finally, make a conclusion by using statistical statements with confidence intervals

Statistical Inference by Estimating Population Mean

- Statistical statements with confidence intervals?
- In normal distribution

Roughly, there is 95% chance that the observed sample average will lie within 1.96 s.e.'s away from the center μ of the distribution



Statistical Inference by Estimating Population Mean

- Statistical statements with confidence intervals?
- So roughly 95% of the samples will capture the true population average μ in the interval
  sample average $\pm$ 1.96 * standard error
- This interval is called a 95% confidence interval. The confidence level (95% in this example) says how confident we are that the procedure will "catch" the true population average μ.

Statistical Inference by Estimating Population Mean

- In general a confidence interval has the form:
  sample estimate ± margin of error

90% Confidence Interval $1.64\dfrac{sd}{\sqrt{n}}$

95% Confidence Interval $1.96\dfrac{sd}{\sqrt{n}}$

99% Confidence Interval $2.57\dfrac{sd}{\sqrt{n}}$

Margin of error

$z_{\alpha/2}$

$\alpha = 1 -$confidence level

- How to calculate? [Example later]
  z value, TextBook section 1.8, Page 34

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

Statistical Inference by Estimating Population Mean

- Example: We'd like to estimate the average age of people in USA. A random sample of 200 people presents the average age is 32 and STD is 5. Estimate the average age of people in USA by the sample statistics using a 95% confidence interval.

Statistical Inference by Estimating Population Mean

- You can follow these steps
  1) Collect sample statistics, such as sample mean
  2) The sample mean is assumed to be normal ($n>=30$) and centered as population mean
  3) The standard error of the sampling distribution is expected to be as small as possible. Note: usually it becomes smaller if your n is larger
  4) Finally, make a conclusion by using statistical statements with confidence intervals

How about a smaller sample size? Such as $n<30$?

Statistical Inference by Estimating Population Mean

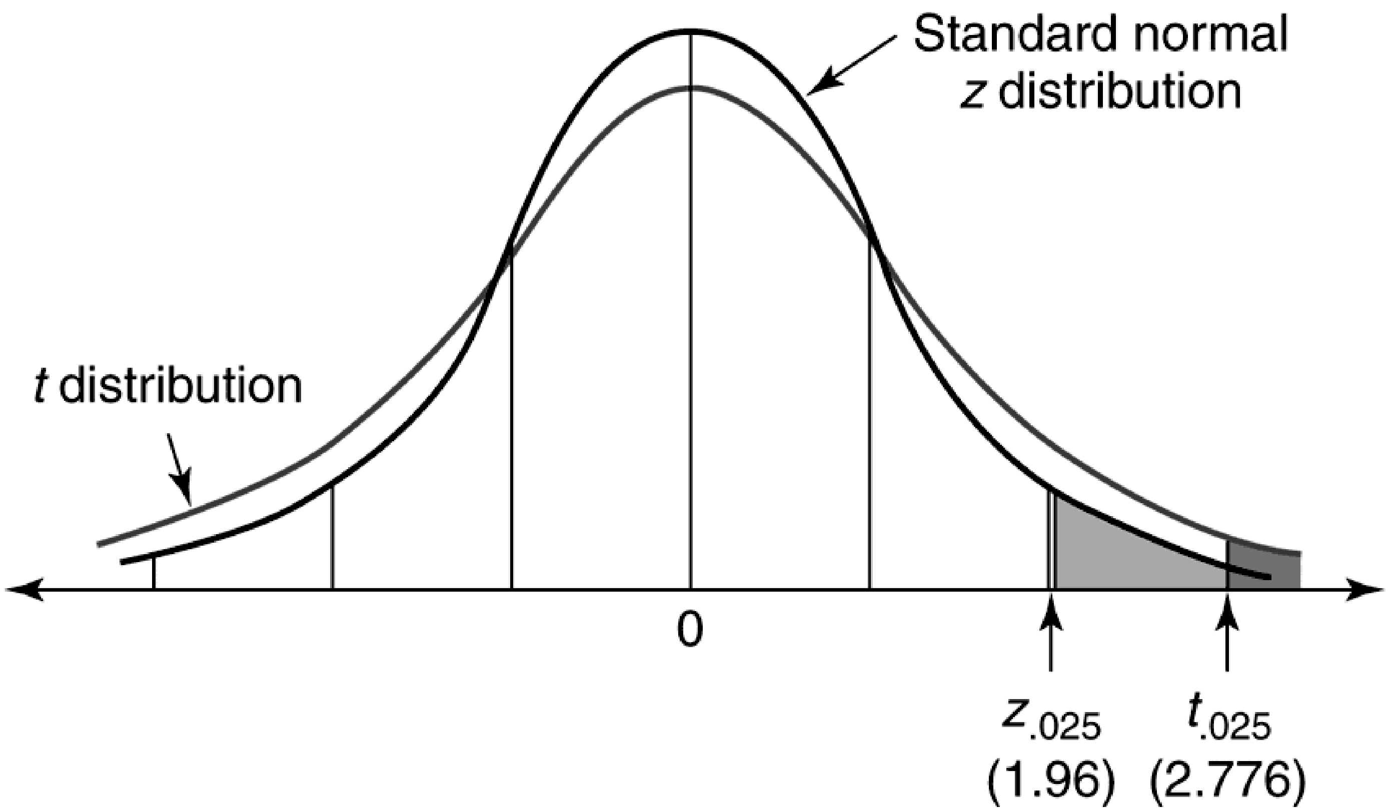
- Large vs Small sample size
  - When it comes to large sample size, we need to know either the population STD (note: usually we do not know it) or the sample is large enough so that we can use sample STD to estimate population STD
  - When it comes to smaller samples, we prefer to use t distribution rather than normal distribution

Statistical Inference by Estimating Population Mean

- When should we use t distribution
  - Sample size is small
  - We do not know population STD
- Difference between t and normal distribution
  - t distribution is similar to normal distribution
  - t distribution is applied when $n<30$
  - t distribution will be close to normal when n is increased
  - The only parameter in t distribution is the degree of freedom, df, $df = n-1$

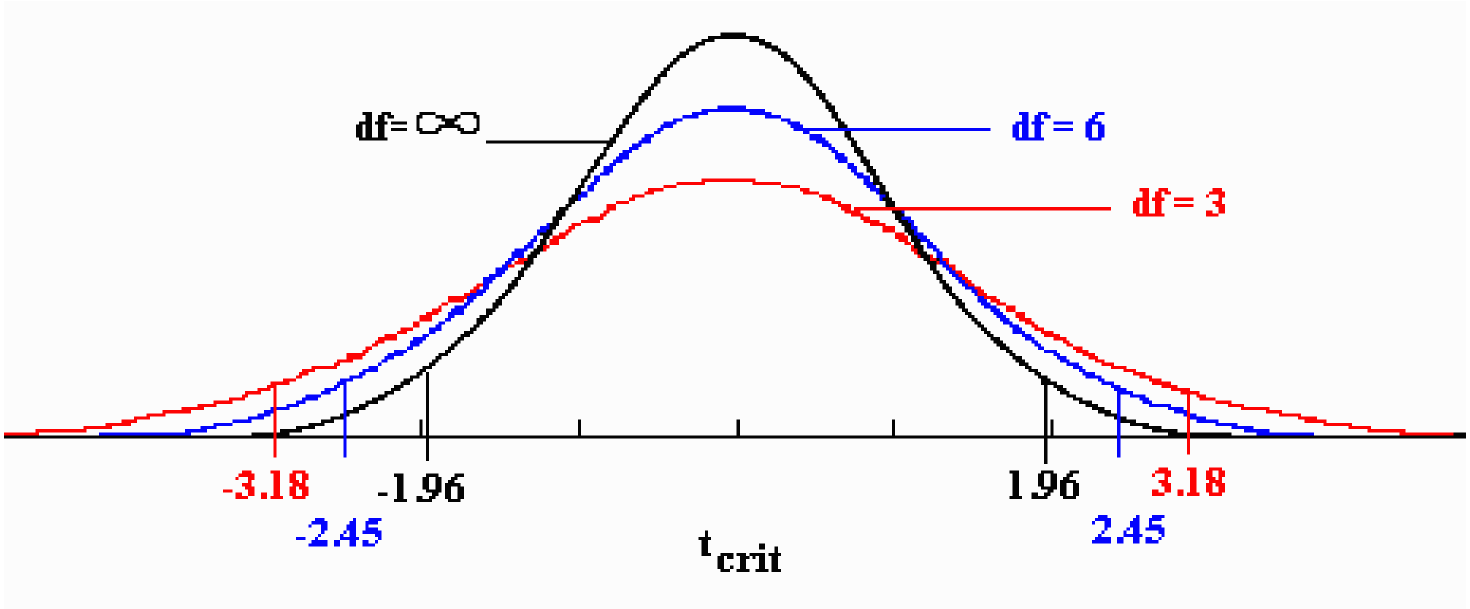
Statistical Inference by Estimating Population Mean

- t distribution and estimates



Statistical Inference by Estimating Population Mean

- t distribution and estimates
  df is smaller, the spread will be greater
  df is large enough, it becomes normal distribution



Statistical Inference by Estimating Population Mean

- t distribution and estimates
  Confidence interval by t distribution

  $$\bar{y} \pm t_{\alpha/2} s_{\bar{y}} = \bar{y} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$, α = 1 – confidence level

  t value, refer to Textbook section 1.8, Page 37
- Difference between z and t values
  - z value is associated with α
  - t value is associated with α and df
  - Note: you do not need to know how to calculate z and t values, you can refer them to the z or t tables, or obtain the values from statistical software, such as R or SAS

Summary: Statistical Inference by Estimating Population Mean

- You can follow these steps
  1) Collect sample statistics, such as sample mean
  2) Sample is larger ($n>=30$), we assume sample mean follows normal distribution; otherwise, we assume it follows t distribution
  3) The standard error of the sampling distribution is expected to be as small as possible. Note: usually it becomes smaller if your n is larger
  4) Finally, make a conclusion by using statistical statements with confidence intervals
     sample estimate ± margin of error
     margin of error = z value or t value × standard error

Summary: Statistical Inference by Estimating Population Mean

- How to calculate z value or t value

1) If n >= 30, normal distribution, z value

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

2) Otherwise, t distribution, t value

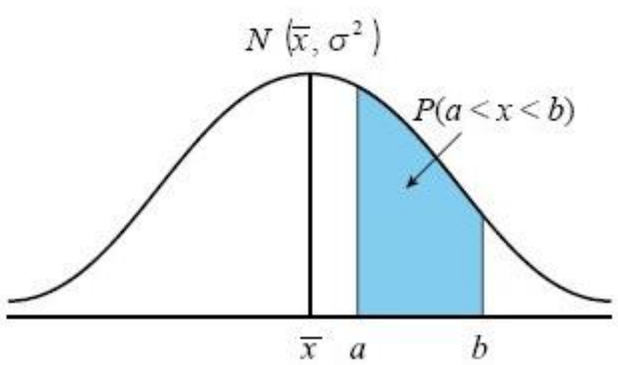
$$\bar{y} \pm t_{\alpha/2}s_{\bar{y}} = \bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

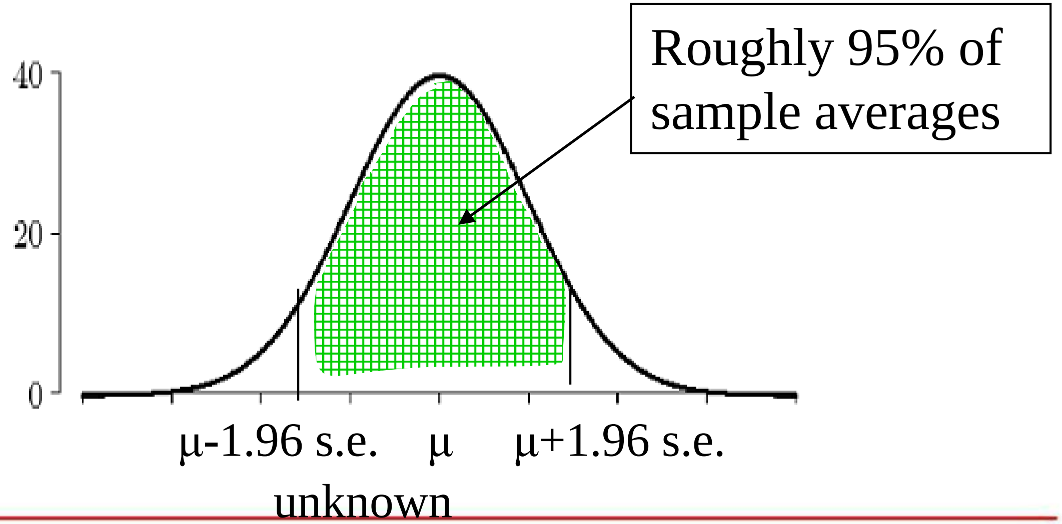
Summary: Statistical Inference by Estimating Population Mean

- How to calculate z value or t value
  1) If n >= 30, normal distribution, z value

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

Assume confidence level is 95%, α = 1-0.95=0.05





Summary: Statistical Inference by Estimating Population Mean

- How to calculate z value or t value

  1) If n >= 30, normal distribution, z value

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

Assume confidence level is 95%, α = 1-0.95=0.05

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

Summary: Statistical Inference by Estimating Population Mean

- How to calculate z value or t value
  1) If n >= 30, normal distribution, z value

     $\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$, α = 1 – confidence level

     Assume confidence level is 95%, α = 1-0.95=0.05

     Why we look for 0.975?

     the Z-table shows only the probability below a certain z-value, and you want the probability between two z-values, –z and z. If 95% of the values must lie between –z and z, you expand this idea to notice that a combined 5% of the values lie above z and below –z. So 2.5% of the values lie above z, and 2.5% of the values lie below –z.

Summary: Statistical Inference by Estimating Population Mean

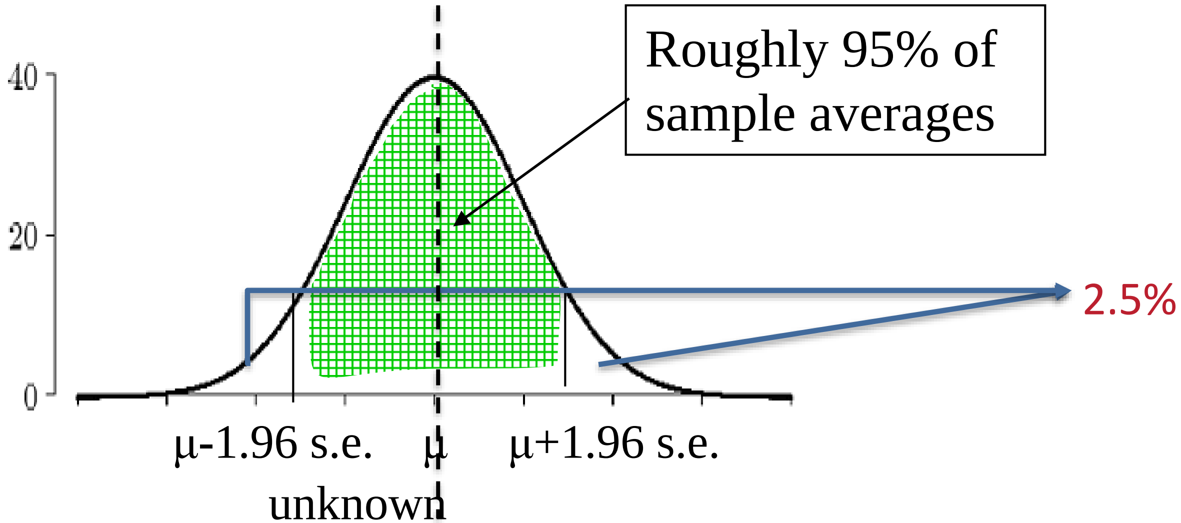
- How to calculate z value or t value
  1) If n >= 30, normal distribution, z value

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

, α = 1 – confidence level

Assume confidence level is 95%, α = 1-0.95=0.05

Why we look for 0.975?



Statistical Inference by Estimating Population Mean

- In general a confidence interval has the form:
  sample estimate ± margin of error

90% Confidence Interval $1.64 \frac{sd}{\sqrt{n}}$

95% Confidence Interval $1.96 \frac{sd}{\sqrt{n}}$

99% Confidence Interval $2.57 \frac{sd}{\sqrt{n}}$

Margin of error

- How to calculate?
  z value, TextBook section 1.8, Page 34

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

Schedule

- Course Structure
- Quick Reviews
- Use Sample to Estimate Population
- One-Sample Hypothesis Testing
- Two-Sample Hypothesis Testing

Hypothesis Testing

- What is a hypothesis and how useful it is
- What are statistical elements in hypothesis testing
- Types of hypothesis testing
- How to perform hypothesis testing

Hypothesis Testing

- What is a hypothesis and how useful it is
- What are statistical elements in hypothesis testing
- Types of hypothesis testing
- How to perform hypothesis testing

What is a hypothesis

- Hypothesis is a claim or assumption
- Example
  - Average age is 30
  - Average age is no more than 30
  - Average age in NYC is larger than the one in Chicago
- Hypothesis Testing is used to validate an hypothesis is true or false based on a confidence level

How useful the hypothesis it is

- Descriptive Statistics is used for you to briefly understand the data
- After that, you may have some initial concerns or questions which can be described by a hypothesis
- Let’s take the Case Study 1: Student grades for example
  - Student info: age, gender, nationality
  - Behaviors: # of hours in reading, assignments, games
  - Performance: exam, final grade, letter grade
- Do you have any concerns?

Hypothesis Testing

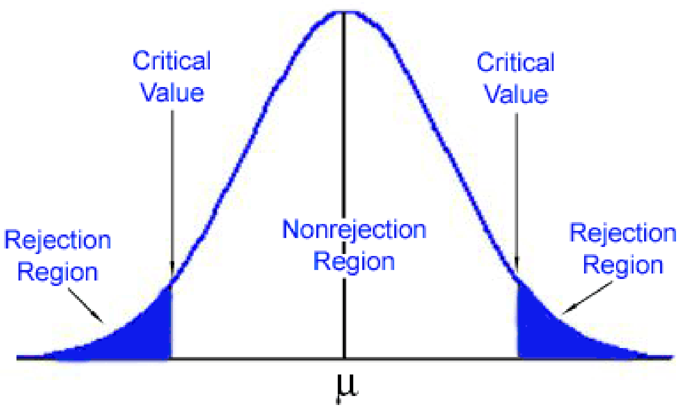
- What is a hypothesis and how useful it is
- What are statistical elements in hypothesis testing
- Types of hypothesis testing
- How to perform hypothesis testing

Elements in Hypothesis Testing

- Null Hypothesis, H_0
  This is the hypothesis we have doubts
- Alternative Hypothesis, H_a or H_1
  This is the hypothesis which is counter to the null hypothesis. Usually it is what we want to support
- Test Statistics
  It is used to make decisions
- Level of significance, α
  The probability of rejecting H_0 giving H_0 is true

Elements in Hypothesis Testing

- Rejection Region



If our test statistics faill into rejection region, we reject null hypothesis and accept the alternative hypothesis.

- P-value

It is a probability value between 0 and 1 as evidence to reject the null hypothesis.

95% confidence level, we reject H_0 if p-value<0.05

P-value = area under normal curve based on the test statistics

Elements in Hypothesis Testing

- Example
  Monthly cell bill is \$42
  I do not think this is true
- H0: $\mu = 42$
  Ha: $\mu \neq 42$

Hypothesis Testing

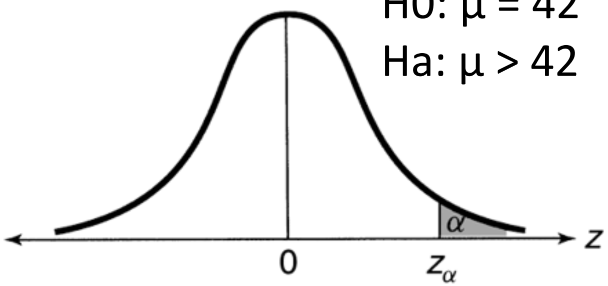
- What is a hypothesis and how useful it is
- What are statistical elements in hypothesis testing
- Types of hypothesis testing
- How to perform hypothesis testing

Types Hypothesis Testing: Based on Samples

- Hypothesis testing on one sample mean
  Monthly cell bill is $42
  I do not think this is true
- Hypothesis testing on two sample means
  Monthly cell bill by ATT and T-Mobile is the same
  ATT is more expensive than T-Mobile

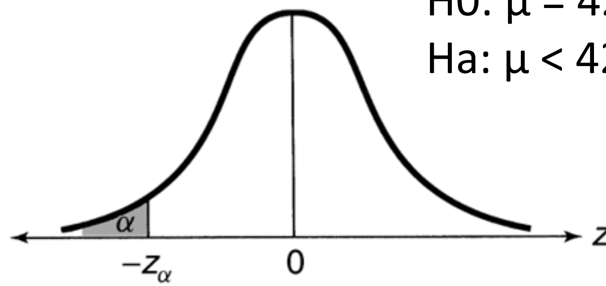
Elements in Hypothesis Testing: Based on Ha

- One-sided or one-tailed statistical test



H0: $\mu = 42$
Ha: $\mu > 42$

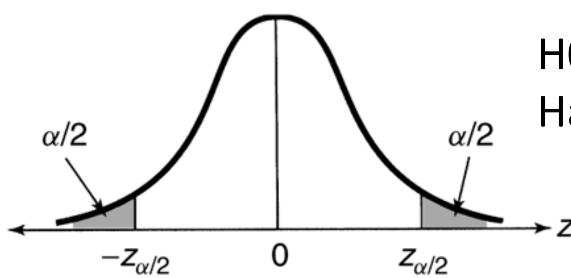
(a) $\mu > \mu_0$



H0: $\mu = 42$
Ha: $\mu < 42$

(b) $\mu < \mu_0$

- Two-sided or two-tailed statistical test



H0: $\mu = 42$
Ha: $\mu \neq 42$

(c) $\mu \neq \mu_0$

Hypothesis Testing

- What is a hypothesis and how useful it is
- What are statistical elements in hypothesis testing
- Types of hypothesis testing
- How to perform hypothesis testing

Steps in Hypothesis Testing

1. State the null hypothesis, H_0 and the alternative hypothesis, H_a
2. Based on Ha, decide it is one-tailed or two-tailed test
3. Choose the level of significance, α. Or, you can claim statistical confidence level, $\alpha = 1 -$ confidence level
4. Determine the appropriate test statistic and sampling distribution – depends on sample size
5. Determine the critical values that divide the rejection and non-rejection regions

Steps in Hypothesis Testing

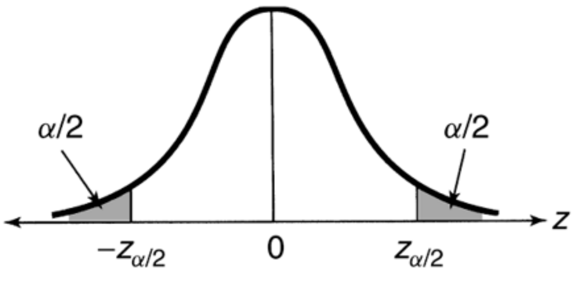
5. Make the statistical decision and state the managerial conclusion.
   - By using test statistics
     If it falls in the rejection area, we reject H0
   - By using p-value
     If the p-value < α, we reject Ho and accept Ha
   - By using confidence interval
     Note, the acceptance region is the confidence interval based on the confidence level

Hypothesis Testing

- We have three metrics to make decisions
  - You can use anyone of these three metrics
  - You will definitely get the same results
  - All of these three metrics are based on “reject region”
  - You need to fully understand rejection region in order to understand the three metrics.

Example: Two-tailed & large sample

- Example: two-tailed test for large sample (n>=30)



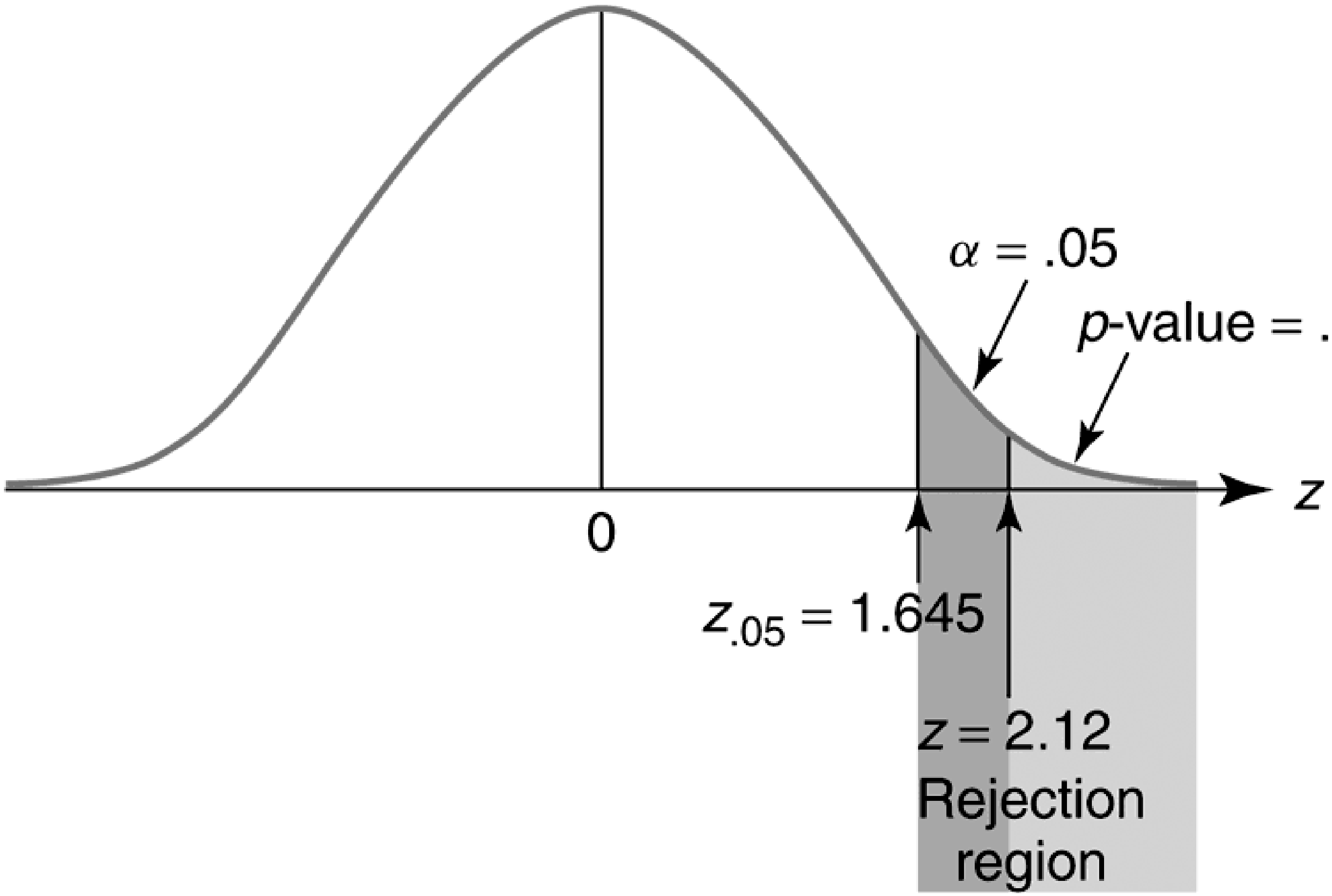
(c) $\mu \neq \mu_0$

H0: $\mu = 42$
Ha: $\mu \neq 42$

  - Confidence Interval [v1, v2], see μ falls in interval or not
  - Critical values as shown in Figure, to see whether the Z statistics falls in the non-rejection region or not
  - P-value is a similar way

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Example: One-tailed & large sample



Example: Diameter

We were told the average diameter of a brand new bolt is 30mm. We do not believe it! Assume we know STD is 0.8.

1. State the appropriate null and alternative hypotheses
   - H_0: μ = 30 H_1: μ ≠ 30 (This is a two-tail test)
2. Specify the desired level of significance and sample size
   - Suppose that $\alpha = 0.05$ (95% confidence to make the conclusions) and n = 100 are chosen for this test
3. Determine the appropriate technique
   - σ is assumed known and n is large, so this is a z test.

Hypothesis testing on one sample mean

Large-Sample ($n \geq 30$) Test of Hypothesis About μ

Test statistic: $z = (\bar{y} - \mu_0)/\sigma_{\bar{y}} \approx (\bar{y} - \mu_0)/(s/\sqrt{n})$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
	$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu \neq \mu_0$
Rejection region:	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$
p-*value*:	$\mathrm{P}(z < z_c)$	$\mathrm{P}(z > z_c)$	$2\mathrm{P}(z > z_c)$ if z_c is positve $2\mathrm{P}(z < z_c)$ if z_c is negative

Decision: Reject H_0 if $\alpha > p$-value, or if test statistic falls in rejection region

where $\mathrm{P}(z > z_\alpha) = \alpha$, $\mathrm{P}(z > z_{\alpha/2}) = \alpha/2$, z_c = calculated value of the test statistic, and $\alpha = \mathrm{P}(\text{Type I error}) = \mathrm{P}(\text{Reject } H_0 \mid H_0 \text{ true})$.

Example: Diameter

1. State the appropriate null and alternative hypotheses
   - H_0: $\mu = 30$ H_1: $\mu \neq 30$ (This is a two-tail test)
2. Specify the desired level of significance and sample size
   - Suppose that $\alpha = 0.05$ (95% confidence to make the conclusions) and $n = 100$ are chosen for this test
3. Determine the appropriate technique
   - σ is assumed known and n is large, so this is a z test.
4. Determine the critical values
   - For $\alpha = 0.05$ the critical Z values are ± 1.96

Example: Diameter

3. Determine the appropriate technique
   - σ is assumed known and n is large, so this is a z test.
4. Determine the critical values
   - For α = 0.05 the critical Z values are ±1.96
5. Collect the data and compute the test statistic
   - Suppose the sample results are

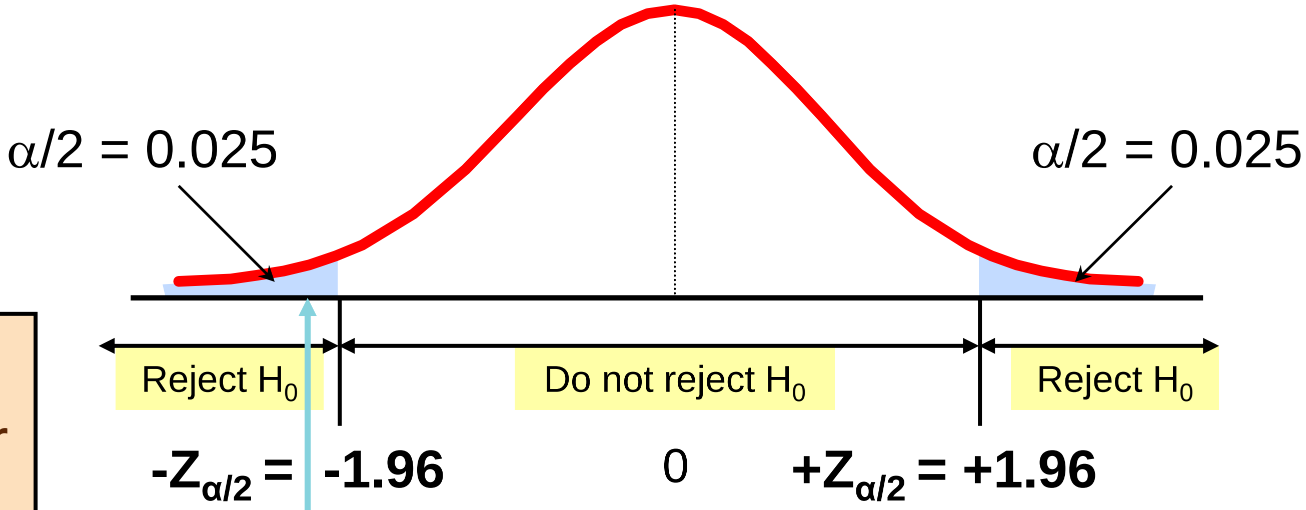
     n = 100, $\bar{x}$ = 29.84 (σ = 0.8 is assumed known)

So the test statistic is:

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{29.84 - 30}{\frac{0.8}{\sqrt{100}}} = \frac{-0.16}{0.08} = -2.0$$

Example: Diameter

6. Is the test statistic in the rejection region? [**by z statistics**]



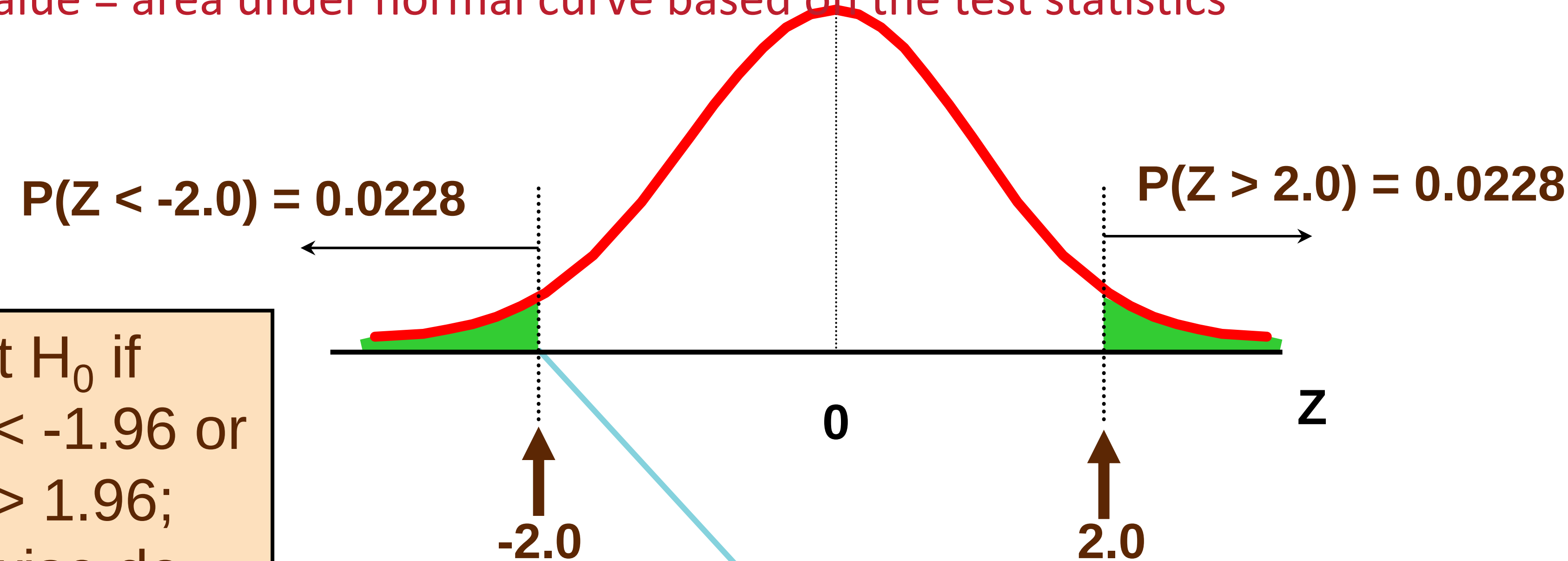
Reject H_0 if $Z_{STAT} < -1.96$ or $Z_{STAT} > 1.96$; otherwise do not reject H_0

Here, $Z_{STAT} = -2.0 < -1.96$, so the test statistic is in the rejection region

Example: Diameter

6. Is the test statistic in the rejection region? [**by p-value**]

P-value = area under normal curve based on the test statistics



Reject H_0 if $Z_{STAT} < -1.96$ or $Z_{STAT} > 1.96$; otherwise do not reject H_0

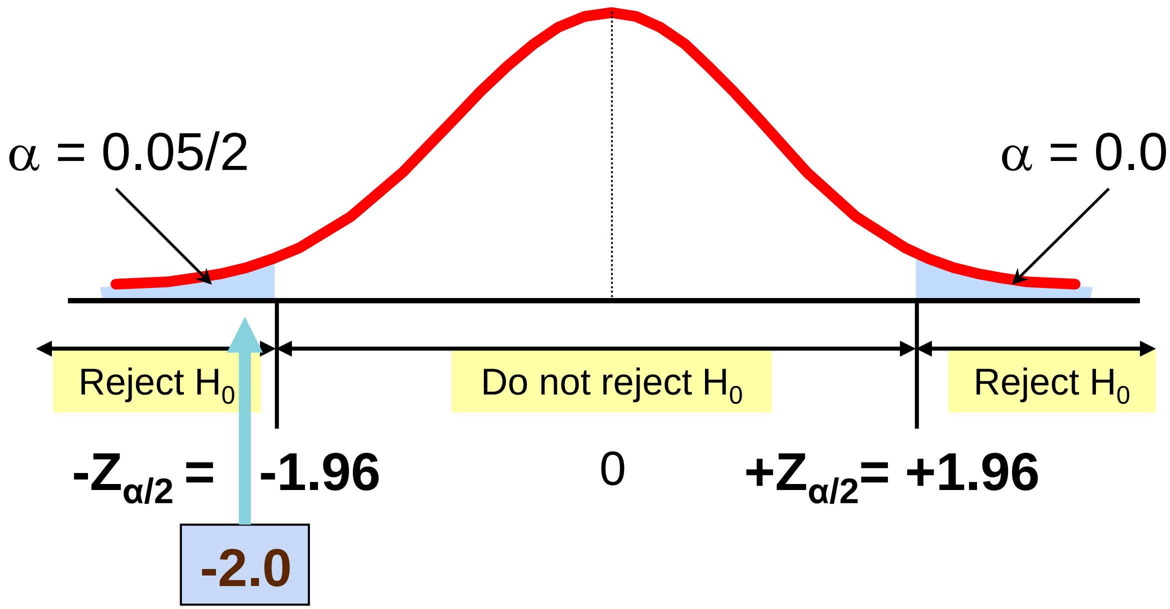
Here, $Z_{STAT} = -2.0$

In two-sided test, p-value = 2 * Pr(z<-2.0) = 2*Pr(z>2.0) = 2*.0228 = 0.0456 < 0.05!!!!

We use 95% confidence level, we reject H0 if p-value<0.05

Example: Diameter

6 (continued). Reach a decision and interpret the result



Since $Z_{STAT} = -2.0 < -1.96$ or p-value < 0.05, reject the null hypothesis and conclude there is sufficient evidence that the mean diameter of a manufactured bolt is not equal to 30

Example: Diameter

6 (continued). Reach a decision and interpret the result
Or, we can use the confidence interval to make a decision

- For $\overline{X}$ = 29.84, σ = 0.8 and n = 100, the 95% confidence interval is:

$$\boxed{29.84 - (1.96)\frac{0.8}{\sqrt{100}} \quad \text{to} \quad 29.84 + (1.96)\frac{0.8}{\sqrt{100}}}$$

$$\boxed{29.6832 \le \mu \le 29.9968}$$

- Since this interval does not contain the hypothesized mean (30), we reject the null hypothesis at $\alpha = 0.05$

Hypothesis testing on one sample mean

Hypothesis Tests for μ

σ Known, n>=30

(z test)

The test statistic is:

$$Z_{STAT} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

σ Unknown, n<30

(t test)

The test statistic is:

$$t_{STAT} = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Hypothesis testing on one sample mean

Large-Sample ($n \geq 30$) Test of Hypothesis About μ

Test statistic: $z = (\bar{y} - \mu_0)/\sigma_{\bar{y}} \approx (\bar{y} - \mu_0)/(s/\sqrt{n})$

	ONE-TAILED TESTS		TWO-TAILED TEST
	H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$
	H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$	H_a: $\mu \neq \mu_0$
Rejection region:	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z\rvert > z_{\alpha/2}$
p-*value*:	$P(z < z_c)$	$P(z > z_c)$	$2P(z > z_c)$ if z_c is positve $2P(z < z_c)$ if z_c is negative

Decision: Reject H_0 if $\alpha >$ p-value, or if test statistic falls in rejection region

where $P(z > z_\alpha) = \alpha$, $P(z > z_{\alpha/2}) = \alpha/2$, z_c = calculated value of the test statistic, and α = P(Type I error) = P(Reject H_0| H_0 true).

Hypothesis testing on one sample mean

Small-Sample Test of Hypothesis About μ

Test statistic: $t = (\bar{y} - \mu_0)/(s/\sqrt{n})$

	ONE-TAILED TESTS		TWO-TAILED TEST
	H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$
	H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$	H_a: $\mu \neq \mu_0$
Rejection region:	$t < -t_\alpha$	$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
p-*value*:	$P(t < t_c)$	$P(t > t_c)$	$2P(t > t_c)$ if t_c is positve $2P(t < t_c)$ if t_c is negative

Decision: Reject H_0 if $\alpha > p$-value, or if test statistic falls in rejection region

where $P(t > t_\alpha) = \alpha$, $P(t > t_{\alpha/2}) = \alpha/2$, t_c = calculated value of the test statistic, and α = P(Type I error) = P(Reject $H_0 | H_0$ true).

Assumption: The population from which the random sample is drawn is approximately normal.

Steps in Hypothesis Testing

1. State the null hypothesis, H_0 and the alternative hypothesis, H_a
2. Based on Ha, decide it is one-tailed or two-tailed test
3. Choose the level of significance, α. Or, you can claim statistical confidence level, α = 1 – confidence level
4. Determine the appropriate test statistic and sampling distribution – depends on sample size
5. Determine the critical values that divide the rejection and non-rejection regions

Steps in Hypothesis Testing

5. Make the statistical decision and state the managerial conclusion.
   - By using test statistics
     If it falls in the rejection area, we reject H0
   - By using p-value
     If the p-value < α, we reject Ho and accept Ha
   - By using confidence interval
     Note, the acceptance region is the confidence interval based on the confidence level