Statistical Analysis

Johns Hopkins
Carey Business School

Some of the slides in this presentation are from Bowerman, B. L., O'Connell, R. T., & Murphree, E.S., (2010). *Business Statistics in Practice* (6th Ed.), Copyright © McGraw-Hill Education.

Hypothesis Testing

1. Null and Alternative Hypotheses Principles and Rules
2. Errors in Hypothesis Testing
3. z Tests about a Population Mean σ Known
4. t Tests about a Population Mean σ Unknown
5. z Tests about a Population Proportion
6. Two Population Tests

Principle of Hypothesis Testing

The null hypothesis is the hypothesis being tested. It is either rejected or not rejected on the basis of the sample information. The alternative hypothesis is specified as another choice if the null is rejected.

Rules of Hypothesis Testing

1. The null hypothesis always has the equal sign.

$$H_0: \mu =,\ H_0: \mu \leq,\ H_0: \mu \geq$$

2. The alternative is the compliment

$$H_a: \mu \neq,\ H_a: \mu >,\ H_a: \mu <$$

The Do Not Reject area (H_0) is $1-\alpha$.

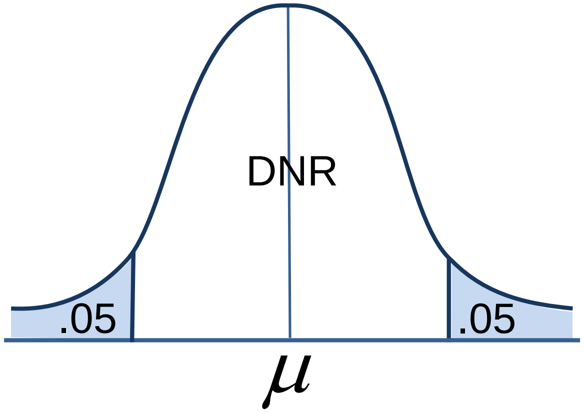
3. We *never accept* the H_0 because we cannot prove it. We have sample information that either supports or does not support H_0.

1 Tail and 2 Tail Problems

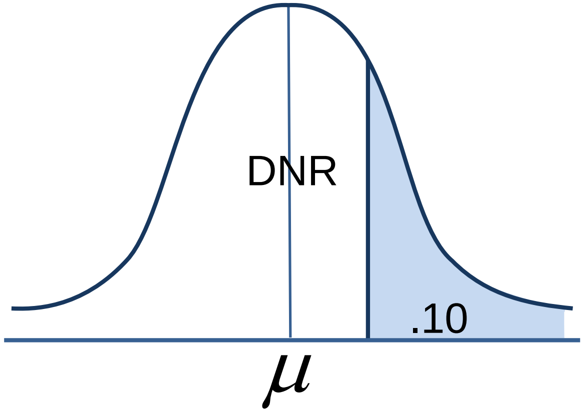
$$H_0: \mu =$$
$$H_a: \mu \neq$$
$$\alpha = .10$$

$$H_0: \mu \leq$$
$$H_a: \mu >$$
$$\alpha = .10$$

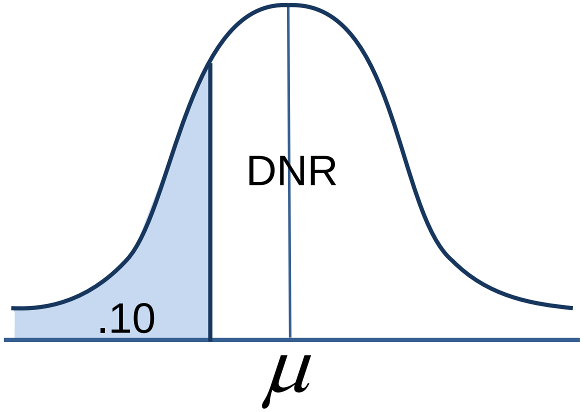
$$H_0: \mu \geq$$
$$H_a: \mu <$$
$$\alpha = .10$$



$$z\ value = \pm 1.65$$



$$z\ value = 1.28$$



$$z\ value = \text{-}1.28$$

Three Methods

1. Determine critical x values that define DNR (Do Not Reject)

$$x_c = \mu \pm z\sigma_{\bar{x}}$$

2. Determine number of standard deviations $\bar{x}$, sample mean, is from hypothesized mean and compare to test statistic z.

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

3. Calculate p*-value* for sample mean and compare to the significance level, α.

Words and Mathematical Relationships

Changing words to mathematical relationships and determining H_0 or H_a

Words	Math Relationship	Hypothesis
is	=	H_0
more than	>	H_a
at least	$\geq$	H_0
greater than	>	H_a
mean of x or more	$\geq$	H_0
do not exceed	$\leq$	H_0

Interpretation

Meaning of significance value and *p-value*

If a significance value of .05 is used, we are saying that if the probability of the sample mean's value is less than 5 in 100, we reject the null hypothesis

If the sample mean has a *z-value* of 1.0, then we are saying that the probability of having that sample mean is 31.6% and therefore we do not reject the null. Therefore, the *p-value* is .316.

Error Probabilities

Type I Error: Rejecting H_0 when it is true

- α is the probability of making a Type I error
- $1 - \alpha$ is the probability of not making a Type I error

Type II Error: Failing to reject H_0 when it is false

- β is the probability of making a Type II error
- $1 - \beta$ is the probability of not making a Type II error

	State of Nature	
Conclusion	H_0 True	H_0 False
Reject H_0	**Type I Error (α)**	Correct Decision
Do not Reject H_0	Correct Decision	**Type II Error (β)**

Typical Values

Usually set α to a low value

1. So there is a small chance of rejecting a true H_0
2. Typically, $\alpha = 0.05$
   - Strong evidence is required to reject H_0
   - Usually choose α between 0.01 and 0.05

   α = 0.01 requires very strong evidence is to reject H_0
3. Tradeoff between α and β
   - For fixed sample size, the lower α, the higher β

   And the higher α, the lower β

z Tests about a Population Mean: σ Known

Test hypotheses about a population mean using the normal distribution

1. Called z tests
2. Require that the true value of the population standard deviation σ is known
   - In most real-world situations, σ is not known
     a. But often is estimated from s of a single sample
     b. When σ is unknown, test hypotheses about a population mean using the t distribution
   - Here, assume that we know σ

t Tests about a Population Mean: σ Unknown

- Assume the population being sampled is normally distributed
- The population standard deviation σ is unknown, as is the usual situation

  If the population standard deviation σ is unknown, then it will have to be estimated from a sample standard deviation
- Under these two conditions, use the t distribution to test the hypotheses

Steps in Hypothesis Testing

1. State the null and alternative hypotheses
2. Specify the significance level α
3. Select the test statistic
4. Determine the critical value rule for deciding whether or not to reject H_0
5. Collect the sample data and calculate the value of the test statistic
6. Decide whether to reject H_0 by using the test statistic and the rejection rule
7. Interpret the statistical results in managerial terms and assess their practical importance

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Hypothesis Testing *example*

You hypothesize that the IQ of the Hopkins Carey Business Student ***is*** 112. You random sample 81 students and the IQ score is 118. Test your hypothesis using a significance (α) of .05, and a population standard deviation of 15.

$$\sigma = 15, n = 81, \bar{x} = 118$$

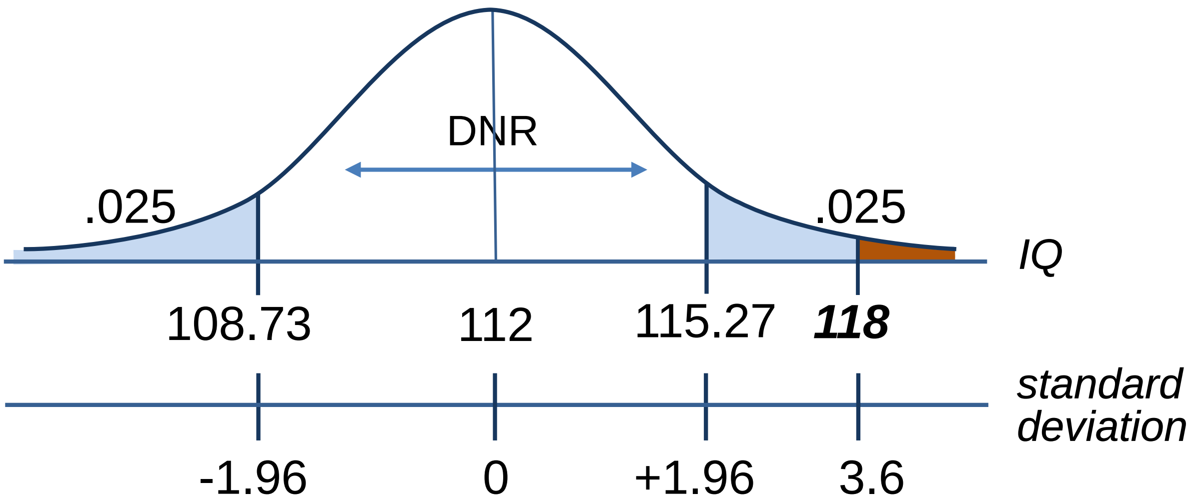
$$H_0: \quad \mu = 112$$

$$H_a: \quad \mu \neq 112$$

$$X_c = \mu_h \pm z(\sigma/\sqrt{n})$$

$$= 112 \pm 1.96 \ (15/\sqrt{81})$$

$$= 112 \pm 3.27$$



Interpretation: Reject the null, the students' IQ do not equal 112.

Alternative method: $z = \dfrac{x - \mu}{\sigma_{\bar{x}}}$

$$z = \frac{118 - 112}{15/\sqrt{81}} = \frac{6}{1.667} = 3.6 \quad 3.6 > 1.96, \text{ test z} \quad area = .0002$$

$$p = (.0002)(2) = .0004$$

The p-Value

- The p-value or the observed level of significance is the probability of obtaining the sample results if the null hypothesis H_0 is true

  The p-value is used to measure the weight of the evidence against the null hypothesis

- Sample results that are not likely if H_0 is true, have a low p -value and are evidence that H_0 is not true

  The p-value is the smallest value of α for which we can reject H_0

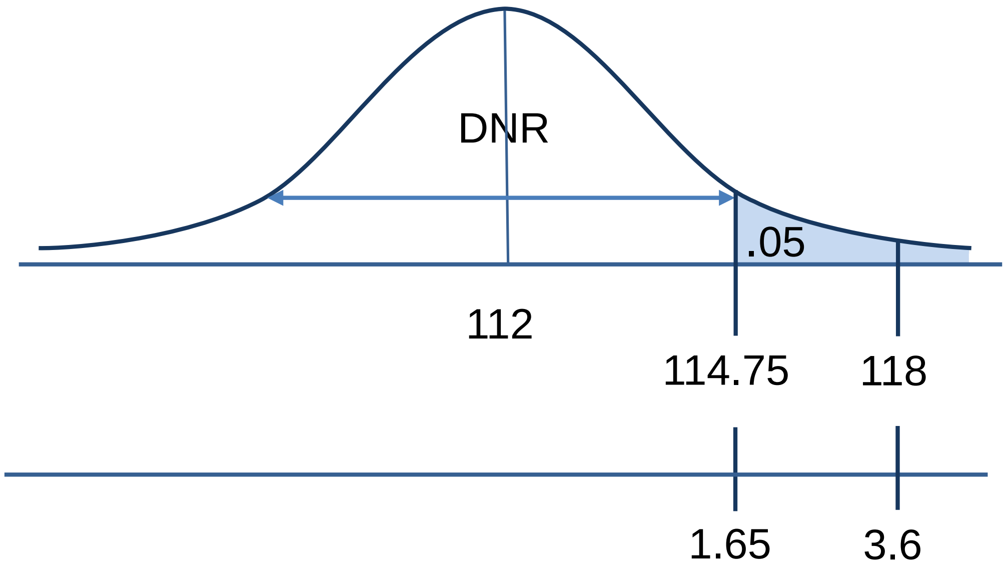
- The p-value is an alternative to testing with a z test statistic

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Hypothesis Testing *example*

Using the same example as before, test your hypothesis that the students have an IQ of ***more than*** 112

$$\sigma = 15, n = 81, \bar{x} = 118$$
$$H_0: \quad \mu \leq 112$$
$$H_a: \quad \mu > 112$$
$$X_c = \mu_h + z(\sigma/\sqrt{n})$$
$$= 112 + 1.65 \ (15/\sqrt{81})$$
$$= 112 + 2.75$$



Interpretation: We reject the null. However, our data supports that students have an IQ of more than 112 because we were testing for the alternative hypothesis

Alternative method: $z = 3.6 \text{ and } 3.6 > 1.65, \text{ test z}$

Hypothesis Testing *example*

Using the same number, what if we stated the students' IQ scores are ***at least*** 112.

$$\sigma = 15,\ n = 81,\ \bar{x} = 118$$

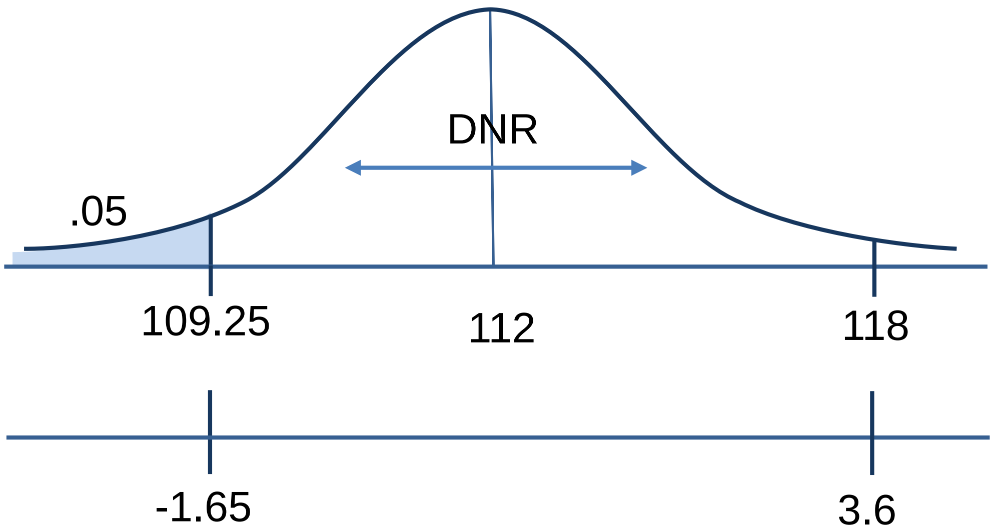
$$H_0:\ \mu \geq 112$$

$$H_a:\ \mu < 112$$

$$X_c = \mu - z(\sigma/\sqrt{n})$$

$$= 112 - 1.65\ (15/\sqrt{81})$$

$$= 112 - 2.75$$



Interpretation:
We do not reject H_0. Data supports students' IQ of at least 112.

Alternative method: $z = 3.6$ and $3.6 > -1.65$, test z

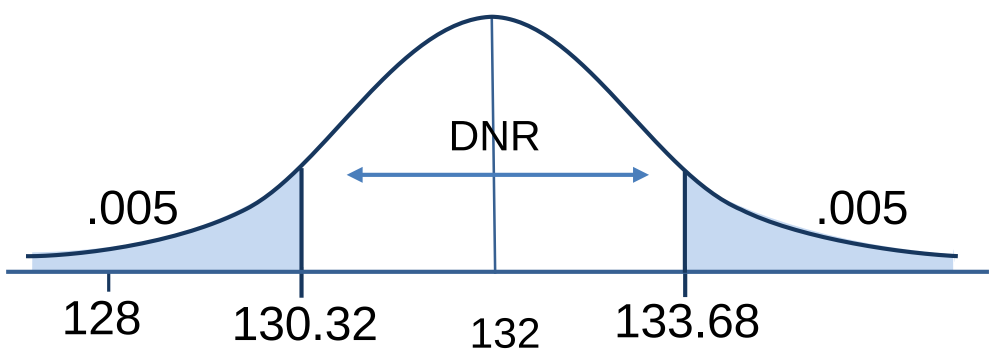
Hypothesis Testing *example*

Kohl's Department Store wants to determine the amount spent per child in back to school spending. They hypothesize that $132 will be spent. The random sample 25 orders and calculate average spending to be $128 with $s=3$. Test your hypothesis at $\alpha=.01$

$$n=25, \bar{x}=128, s=3, \alpha=.01$$

$$H_0: \quad \mu=132$$
$$H_a: \quad \mu\neq 132$$

$$X_c = \mu_h \pm t(s/\sqrt{n})$$
$$=132 \pm 2.797 \ (3/\sqrt{25})$$
$$=132 \pm 1.678$$



Interpretation: Reject H_0, our evidence does not support the null hypothesis that parents spend $132 per child

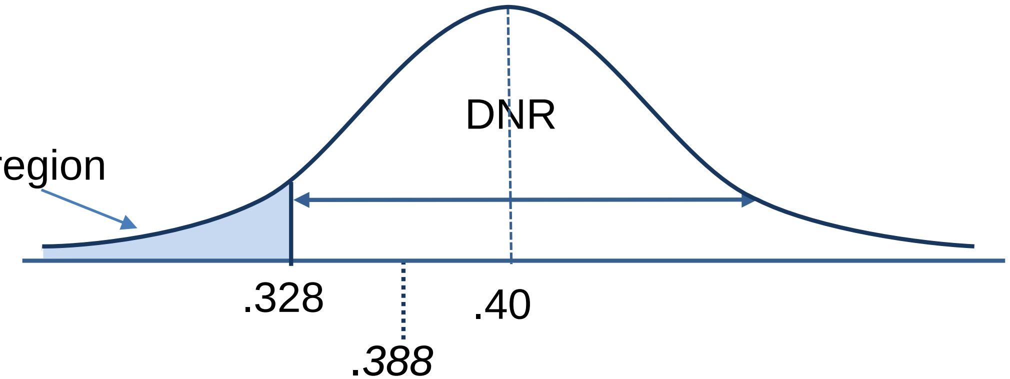
z Tests about a Population Proportion

$H_0:\ p_0 =$	$H_0:\ p_0 \geq$	$H_0:\ p_0 \leq$
$H_a:\ p_0 \neq$	$H_a:\ p_0 <$	$H_a:\ p_0 >$
$p_a = p \pm z\sigma_p$	$p_c = p - z\sigma_p$	$p_c = p_0 + z\sigma_p$
Where $\sigma_p = \sqrt{\dfrac{p_0(1-p_0)}{n}}$		
Where the test statistics is $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$		

Hypothesis Test for Proportions

The manager of a local restaurant, has told her kitchen staff that at least 40 percent of her patrons purchase the daily special. If it appears that less than 40 percent of the patrons are selecting the special, the manger will remove the daily special from the menu. Of a sample of 250 patrons, 97 ordered the daily special. Set α equal to 0.01.

$$n = 250, r = 97, \alpha = .01$$



$$H_0: p \geq 0.40$$
$$H_a: p < 0.40$$

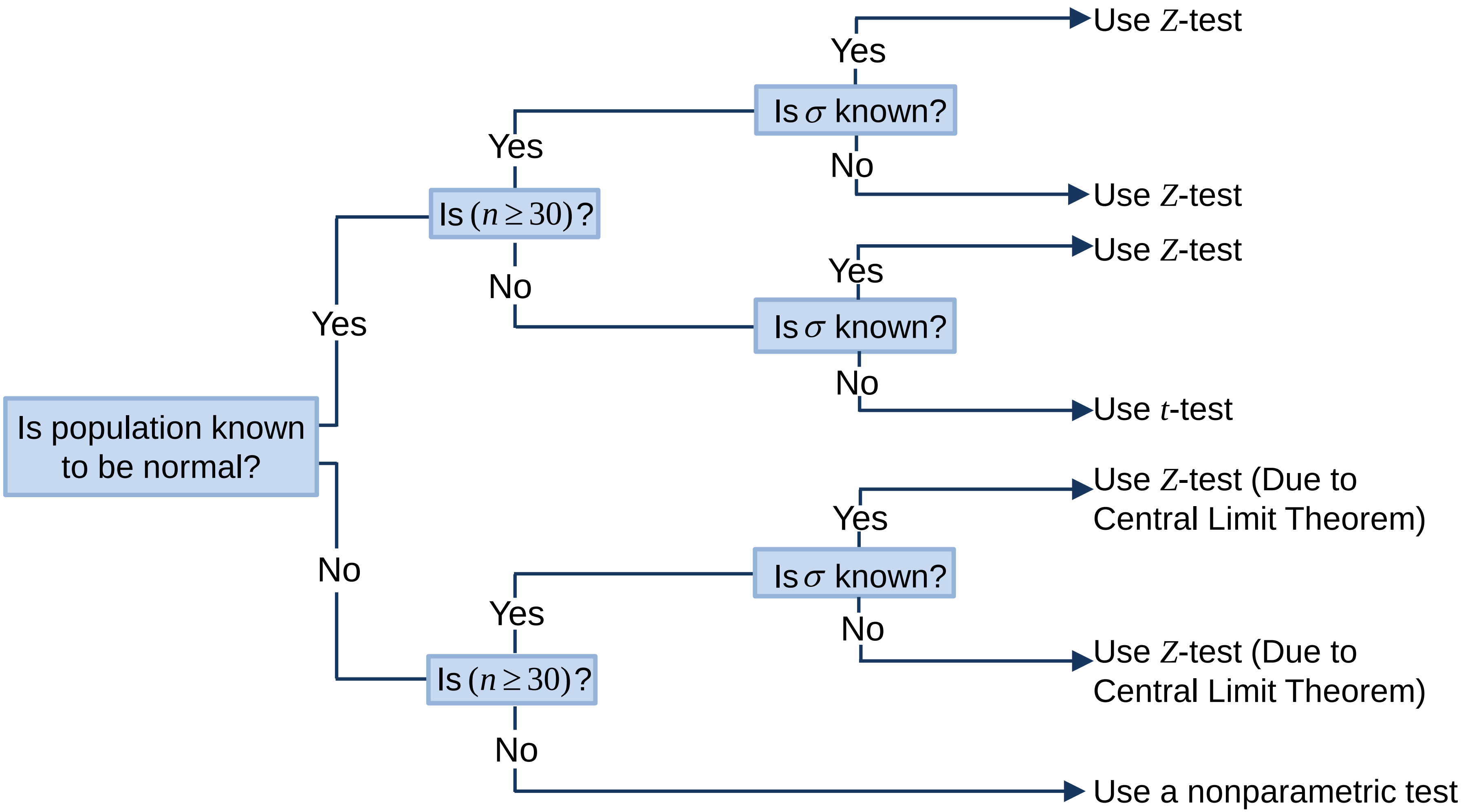
$$\hat{p} = \frac{r}{n} = \frac{97}{250} = .388$$

$$\sigma_p = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{.40(1-.40)}{250}} = 0.031$$

$$\begin{aligned} p_c &= p_0 - Z\sigma_p \\ &= 0.4 - (2.33)(0.031) \\ &= 0.4 - 0.072 \\ &= 0.328 \end{aligned}$$

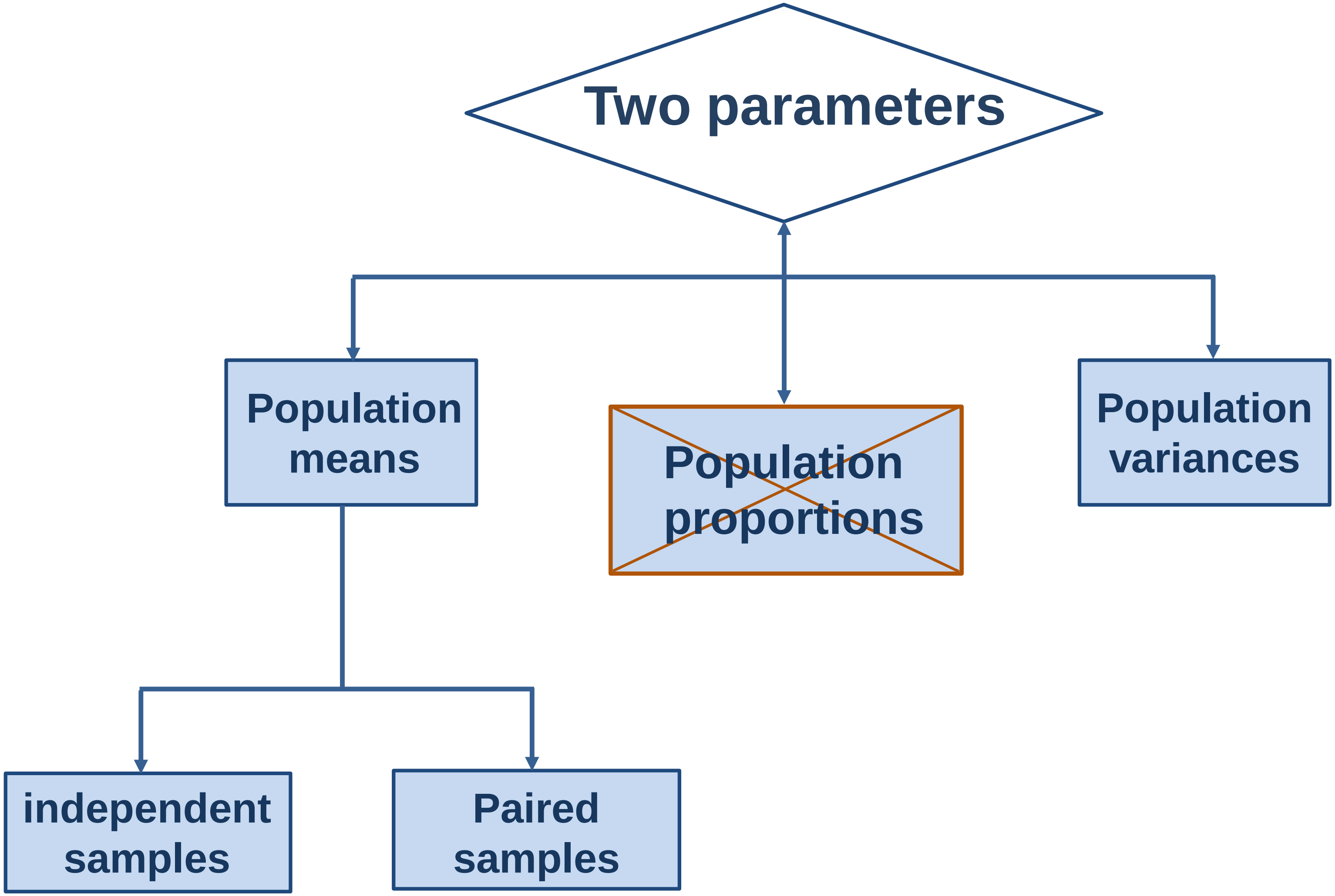
Interpretation: We do not reject the null. The daily special stays on the menu.

Selecting an Appropriate Test Statistic



Webster, Allen L., (1992). Applied Statistics for Business and Economics (2nd Ed.), Copyright © Irwin

Hypothesis Testing: Two Parameters



Two Parameters from Two Populations

One sided	Two sided	One sided
$H_0: \mu_1 \geq \mu_2$	$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 \leq \mu_2$
$H_a: \mu_1 < \mu_2$	$H_a: \mu_1 \neq \mu_2$	$H_a: \mu_1 > \mu_2$

We can write $H_0: \mu_1 = \mu_2$ as $H_0: \mu_1 - \mu_2 = 0$. The right hand side has null value, hence the name null hypothesis. Even for the one sided tests we can use $H_0: \mu_1 - \mu_2 = 0$.

Sometimes we test $H_0: \mu_1 - \mu_2 = D0$ and $H_a: \mu_1 - \mu_2 \neq D0$

One sided	Two sided	One sided
$H_0: \sigma_1^2 \geq \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$	$H_0: \sigma_1^2 \leq \sigma_2^2$
$H_a: \sigma_1^2 < \sigma_2^2$	$H_a: \sigma_1^2 \neq \sigma_2^2$	$H_a: \sigma_1^2 > \sigma_2^2$

More Sampling Distributions

- Suppose there are two populations, $[\mu_1, \sigma_1]$ and $[\mu_2, \sigma_2]$, and we take samples from both (assume independence).
- Then the sample distribution of the difference of two sample means [i.e. distribution of $(\bar{x}_1 - \bar{x}_2)$] has the following :

$$\mu_{\bar{x}_1 - \bar{x}_2} = (\mu_1 - \mu_2) \text{ and}$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{[(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)]}$$

- When the original distributions are normal (or sample sizes are large) the distribution of $(\bar{x}_1 - \bar{x}_2)$ is also normal.
- When the original distributions are normal with unknown variances, we use "t" for the distribution of $(\bar{x}_1 - \bar{x}_2)$ and replace σ with s. Formulas vary for different cases.

z Test for (μ_1, μ_2). σ_1, σ_2 Known

When σ_1 and σ_2 are known and both populations are normal or both sample sizes are at least 30, the test statistic is a z-value…

One sided	Two sided	One sided
$H_0: \mu_1 - \mu_2 \geq D_0$	$H_0: \mu_1 - \mu_2 = D_0$	$H_0: \mu_1 - \mu_2 \leq D_0$
$H_a: \mu_1 - \mu_2 < D_0$	$H_a: \mu_1 - \mu_2 \neq D_0$	$H_a: \mu_1 - \mu_2 > D_0$

If $D_0 = 0$, we can move μ_2 to the right hand side.

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

In the above formula we replace $\mu_1 - \mu_2$ with D_0.

Example

A teacher claims that her students will score higher on a standardized test than her colleague's students. The mean score in her class is 22.1 and the std. deviation is 4.8 with 49 students. Values in the colleague's class are 19.8 and 5.4 (with 44 students). At $\alpha = .10$, can the teacher's claim be supported?

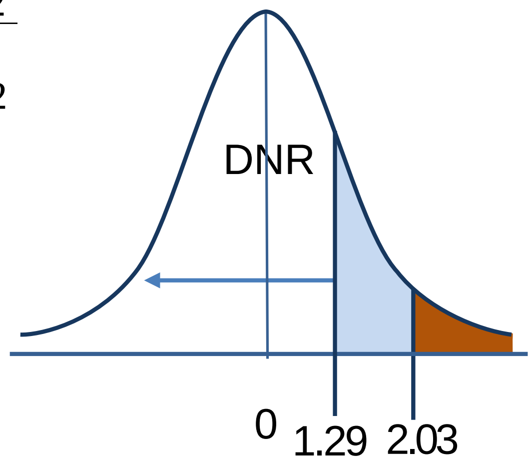
For large samples, we will use s values as estimates of σ.

$$\alpha = .10$$
$$H_0: \ \mu_1 \le \mu_2$$
$$H_a: \ \mu_1 > \mu_2$$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{4.8^2}{49} + \frac{5.4^2}{44}} \approx 1.06$$

$$z = \frac{(22.1 - 19.8) - 0}{1.06} = 2.035$$



Reject H_0

t Test for (μ_1, μ_2), σ_1, σ_2 Unknown

We assume normal populations. There are two cases.

Assume population variances to be equal. We then calculate *pooled standard deviation (s_p)* from the two samples and use it in the test statistics t.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{d.f.} = (n_1 + \text{n}_2 - 2)$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If we *cannot assume population variances to be equal*, we use a *different formula* in the test statistics t with *different degrees of freedom.*

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\text{d.f.} = \min\,(n_1 - 1, \text{n}_2 - 1)$$

Example

You're a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	NASDAQ
Sample Size	21	25
Sample Mean	3.27	2.53
Sample Std Deviation	1.30	1.16

Assuming equal variances, and normality, is there a difference in average yield $\alpha = .05$?

With normality, sample sizes under 30 and assumption of equal variances, we can use t test with pooled variance.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_p = \sqrt{[(21-1)(1.3)^2 + (25-1)(1.16)^2]/(21+25-2)} = 1.2256$$

$$\text{d.f.} = 21+25\text{-}2=44$$

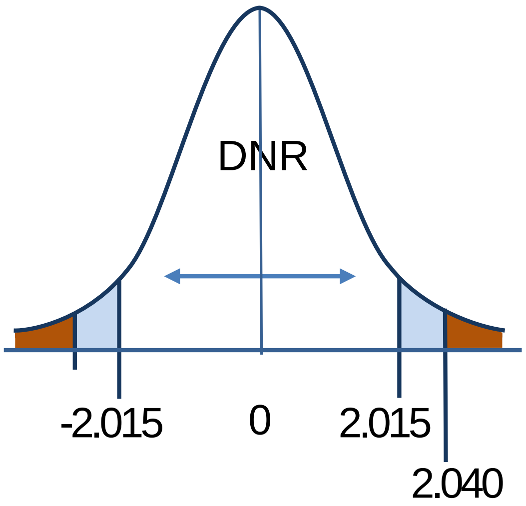
Example *continued*

	NYSE	NASDAQ
Sample Size	21	25
Sample Mean	3.27	2.53
Sample Std Deviation	1.30	1.16

$$\text{d.f.} = (n_1 + n_2 - 2)$$

$$\alpha = 0.05$$
$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



$$t = \frac{(3.27 - 2.53) - (0)}{1.2256\sqrt{(1/21) + (1/25)}} = 2.040$$

Excel formula: TINV(0.05,44) = 2.015

Reject H_0

Example

A random sample of 18 police officers in city A has a mean annual income of $46,500 and s = $3,800. In city B, a random sample of 22 officers has a mean annual income of $44,900 and s = $4,400. Test the claim at α=0.05 that the mean annual incomes in the two cities are not the same.

$$\alpha = 0.05$$
$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

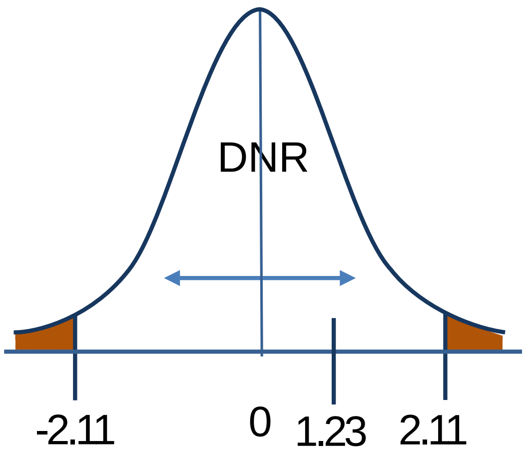
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\text{d.f.} = \min(n_1 - 1, n_2 - 1)$$

t formula, variances unequal?

$$\text{df} = \min(18-1, 22-1) = 17$$
$$t_{0.025,17} = 2.110$$

$$\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$$
$$= \sqrt{(3800^2/18) + (4400^2/22)} \approx 1297$$
$$t = [(46500 - 44900) - (0)]/1297 = 1.23$$



Do not reject H_0

Determining z or t?

Are both sample sizes at least 30?

Yes → Use the z-test.

No ↓

Are both populations normally distributed?

No → You cannot use the z-test or the t-test.

Yes ↓

Are both population standard deviations known?

Yes ↓

Use the z-test.

No → Are the population variances equal?

Yes → Use the t-test with

$$\sigma_{\bar{x}_1-\bar{x}_2} = \hat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$and\ d.f. = n_1 + n_2 - 2$$

No ↓

Use the t-test with

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$$and\ d.f. = \text{smaller of } n_1 - 1\ or\ n_2 - 1$$

Paired t Test

In many experiments, when we are testing hypotheses between two population means, we want to remove the source of difference caused by the observations themselves.

- 15 cars involved in accidents are sent to two repair shops to compare estimates
- Identical twins are used to test the effect of two drugs
- Same students are used to check performance before and after a new lesson is taught
- New drug efficacy (pain before and after, weight before and after, etc.)
- Logistics – change in the mean time-in-transit from supplier to customer (change in route, trucker rest times, etc.)

$$t = (\bar{d} - \mu_d)/(s_d/\sqrt{n})$$

With n pairs, we use n-1 df.