Stat 104: Quantitative Methods

Class 28: ANOVA

One Way ANOVA

- The one-way analysis of variance method is used to test the claim that three or more population means are equal
- This is an extension of the two independent sample t-test
- Its called analysis of variance because it works by (non intuitively) comparing different sample variances.

Example ANOVA Situation

- Subjects: 25 patients with blisters
- Treatments: Treatment A, Treatment B, Placebo
- Measurement: # of days until blisters heal

- Data [and means]:
  - A: 5,6,6,7,7,8,9,10 [7.25]
  - B: 7,7,8,9,9,10,10,11 [8.875]
  - P: 7,9,9,10,10,10,11,12,13 [10.11]

- Are these differences significant?

The null and alternative hypothesis

- The null hypothesis is that the means are all equal

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

- What is the alternative hypothesis?

HARVARD LAW REVIEW

ARTICLE

FAIR DRIVING: GENDER AND RACE DISCRIMINATION IN RETAIL CAR NEGOTIATIONS

*Ian Ayres**

The struggle to eradicate discrimination on the basis of race and gender has a long history in American law. Based on the widely held belief that such discrimination will occur only in markets in which racial or gender animus distorts competition, regulatory efforts have been limited to areas in which interpersonal relations are significant and ongoing, such as housing and employment. In this Article, Professor Ayres offers empirical evidence that seriously challenges faith in the ability of competitive market forces to eliminate racial and gender discrimination in other markets. His Chicago-based research demonstrates that retail car dealerships systematically offered substantially better prices on identical cars to white men than they did to blacks and women. Professor Ayres details the nature and startling degree of the discrimination his testers encountered and evaluates various theoretical explanations for their disparate treatment. Based on his conclusions, Professor Ayres explores routes by which "fair driving" plaintiffs might bring suits against dealerships and mechanisms through which regulators might effectively rid the retail car market of such discrimination.

THE civil rights laws of the 1960s prohibit race and gender discrimination in the handful of markets — employment, housing, and public accommodations — in which discrimination was perceived to be particularly acute.[1] In recent years, lawsuits have increasingly

in each pair of testers. The white male results provide a bench-mark against which to measure the disparate treatment of the non-"white-male" tester. Three consumer pairs (black female and white male, black male and white male, and white female and white male) conducted approximately 180 tests at ninety Chicago dealerships.[19]

Each tester followed a bargaining script designed to frame the bargaining in purely distributional terms: the only issue to be negotiated was the price.[20] The script instructed the testers to focus quickly on buying a particular car,[21] and testers offered to provide

In order to minimize the possibility of non-uniform bargaining, particular attention was paid to issues of experimental control. A major goal of the study was to choose uniform testers and to train them to behave in a standardized manner. Testers were chosen to satisfy the following criteria for uniformity:

1. *Age*: All testers were twenty-four to twenty-eight years old.
2. *Education*: All testers had three or four years of college education.
3. *Dress*: All testers were dressed similarly during the negotiations. Testers wore casual "yuppie" sportswear: the men wore polo or button-down shirts, slacks, and loafers; the women wore straight skirts, blouses, minimal make-up, and flats.
4. *Economic Class*: Testers volunteered that they could finance the car themselves.
5. *Occupation*: If asked by a salesperson, each tester said that he or she was a young urban professional (for example, a systems analyst for First Chicago Bank).
6. *Address*: If asked by the salesperson, each tester gave a fake name and an address for an upper-class, Chicago neighborhood (Streeterville).
7. *Attractiveness*: Applicants were subjectively ranked for average attractiveness.

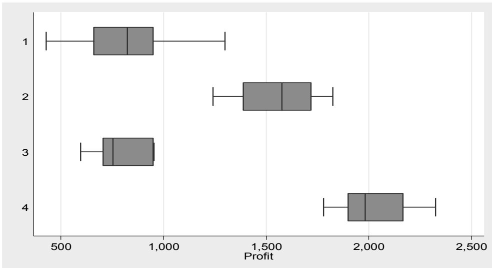
ANOVA popular research tool

- A famous Harvard Law Study

White Male		Black Male	White Female	Black Female
1300	853	1241	951	1899
646	727	1824	954	2053
951	559	1616	754	1943
794	429	1537	706	2168
661	1181		596	2325
824	853			1982
1038	877			1780
754				

Source: Ian Ayres. "Fair Driving: Gender and Race Discrimination in Retail Car Negotiations." *Harvard Law Review.* Vol. 104, No. 4, Feb. 1991.

The Data



Example

- A family doctor might wonder if the mean HDL (so-called good) levels of cholesterol of males in the age groups 20 to 29, 40 to 49, and 60 to 69 years old is the same or different.
- To test this, we assume that the mean HDL cholesterol of each age group is the same (the null hypothesis is always a statement of "no difference").
- If we call the 20- to 29-year-olds population 1, 40- to 49-year-olds population 2, and 60- to 69-year-olds population 3, our null hypothesis would be $H_0 : \mu_1 = \mu_2 = \mu_3$
- *Ha*: At least one population mean is different from the others

Why not do several 2 sample tests?

- At first glance, you might think that to compare the means of three or more samples, you can use the t test, comparing two means at a time. But there are several reasons why the t test should not be done.
- First, when you are comparing two means at a time, the rest of the means under study are ignored. With ANOVA, all the means are compared simultaneously.
- Second, the more means there are to compare, the more t tests are needed. For example, for the comparison of 3 means two at a time, 3 t tests are required. For the comparison of 5 means two at a time, 10 tests are required. And for the comparison of 10 means two at a time, 45 tests are required.

Why not several 2 sample tests?

- We would have to run 3 tests for the HDL example: $H_0 : \mu_1 = \mu_2$ $H_0 : \mu_1 = \mu_3$ $H_0 : \mu_2 = \mu_3$
  $H_a : \mu_1 \neq \mu_2$ $H_a : \mu_1 \neq \mu_3$ $H_a : \mu_2 \neq \mu_3$
- Each test would have a probability of a Type I error of .05 so each test would have a 95% probability of correctly not rejecting the null hypothesis. The probability that all three tests correctly do not reject the null hypothesis is $0.95^3 = 0.86$ (assuming that the tests are independent). There is a $1 - 0.95^3 = 0.14$, or 14%, probability that at least one test will lead to an incorrect rejection of *H*o which is too high.

Terminology

- The *response* variable is the variable you're comparing
- The *factor* variable is the categorical variable being used to define the groups
  - We will assume *k* samples (groups)
- The *one-way* is because each value is classified in exactly one way
  - Examples include comparisons by gender, race, political party, color, etc.
- There is two-way ANOVA taught in Stat 140

The Set Up

- The data set for the one-way analysis of variance consists of k independent univariate samples, each one with the same measurement units (e.g., dollars or miles per gallon).

	Sample 1	Sample 2	•••	Sample k
	$X_{1,1}$	$X_{2,1}$	•••	$X_{k,1}$
	$X_{1,2}$	$X_{2,2}$	•••	$X_{k,2}$
	•	•	•	•
	•	•	•	•
	X_{1,n_1}	X_{2,n_2}	•	X_{k,n_k}
Average	$\overline{X}_1$	$\overline{X}_2$	•••	$\overline{X}_k$
Standard deviation	S_1	S_2	•••	S_k
Sample size	n_1	n_2	•••	n_k

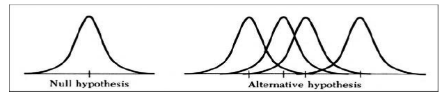
Conditions or Assumptions

- Conditions or Assumptions
  - The data are randomly sampled
  - **The variances of each sample are assumed equal**
  - The data are normally distributed
- These conditions can be limiting and can be tested.
- There are nonparametric procedures if these conditions are severely violated but ANOVA is pretty robust to violations of these assumptions.

Guideline: The largest sample standard deviation should be no more than three times the smallest.

ANOVA Guiding Principle

- ANOVA works by **assuming** all the data has the same variance. The only possible difference is with the means.



- It turns out there are two different ways to estimate this common variance; so we do that and compare the two values.

Ezample: Chips Dataset - Salt Content

BBQ	Cheese-Flavored	Olestra-Based	Baked
338	235	164	290
155	238	197	343
239	251	136	294
184	229	214	373
185	233	148	306
261	232	230	357

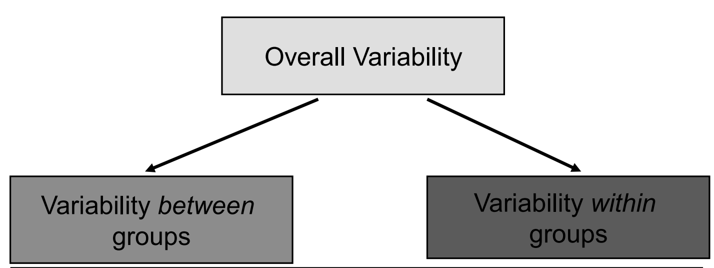
Question of Interest

- Is there a difference in the mean salt content (in mg of sodium) between the four brands of chips?
- How can we compare the mean salt content levels?

Dataset - Salt Content in Chips

- We notice that every observation in the dataset is different.
- What might contribute to this variability in salt content between the chips?
  - Different brands put in varying amounts of salt.
  - Even for same brand, manufacturing process creates variability between salt on each chip (temperature, humidity, machinery, human input)

Variability in Salt Content Between Chips



Variability in Salt Content

- Two main causes:

  (a) Variability between groups

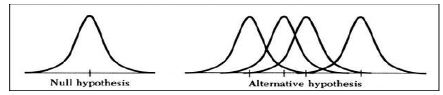
  We can assess this by comparing the mean salt content of each group to the overall mean salt content.

  (b) Variability within each group

  Can be assessed by comparing each individual observation to their brand mean.

ANOVA Guiding Principle

- ANOVA works by **assuming** all the data has the same variance. The only possible difference is with the means.



- It turns out there are two different ways to estimate this common variance; so we do that and compare the two values.

Estimating σ^2 Method 1

- We can estimate the variances of each sample, then combine them:

	Group 1	Group 2	Group 3	Group 4
	338.00	235.00	164.00	290.00
	155.00	238.00	197.00	343.00
	239.00	251.00	136.00	294.00
	184.00	229.00	214.00	373.00
	185.00	233.00	148.00	306.00
	261.00	232.00	230.00	357.00
mean	227.00	236.33	181.50	327.17
variance	4483.6	60.66666667	1429.5	1234.166667

Estimating σ^2 Method 1

- Since each sample variance is an unbiased estimate of σ^2 we can average them to create another (better) unbiased estimate of σ^2 .
- Define

$$s_W^2 = \frac{s_1^2 + s_2^2 + s_3^2 + s_4^2}{4} = 1801.98$$

- The subscript "W" stands for *within*; The only information used in obtaining this estimate is information contained ***within the separate samples.***
- This variability is the same whether or not the hypothesis is true (by assumption).

Estimating σ^2 Method 2

- This method only works if the null hypothesis is true.
- Recall from when we discussed the sample mean that

$$Var(\bar{X}) = \sigma^2 / n \quad so \quad \sigma^2 = nVar(\bar{X})$$

- We can then calculate the variance of the sample mean and get another estimate of σ^2

Estimating σ^2 Method 2

- Define what is called the grand mean (the mean of all the data)

$$\bar{\bar{x}} = \frac{388 + 155 + \cdots + 357}{24} = 243$$

- This is also equal to the mean of the sample means $\bar{\bar{x}} = \frac{227 + 236.33 + 181.5 + 327.17}{4} = 243$

Estimating σ^2 Method 2

- The sample averages allow us to compute an estimate of the variance of the sample mean.
- That is,

$$s_{\bar{x}}^2 = \frac{(227-243)^2 + (236.44-243)^2 + (181.5-243)^2 + (327.16-243)^2}{3} = 3722.24$$

- Since $\sigma^2 = nVar(\bar{X})$ our estimate of the sample variance is $s_B^2 = ns_{\bar{x}}^2 = 6(3722.24) = 22333.44$

The subscript ***B* stands for between, because it** is based on comparing averages ***between the various samples.***

Comparing Variances

- Since s_B^2 and s_W^2 are estimates of the same value σ^2 *if the null hypothesis is true,* we would expect them to be close together if the null is true.
- We define the *F* ratio to be

$$F = \frac{s_B^2}{s_W^2}$$

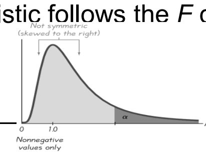
Fisher

Comparing Variances

- The *F* statistic is defined to be

$$F = \frac{s_B^2}{s_W^2}$$

- We reject the null that all the means are equal if this value is much larger than 1.
- This statistic follows the *F* distribution



We will use p-values instead of looking up cut-off values.

Back to our data

- For our potato chip data, the *F* statistic is

$$F = \frac{s_B^2}{s_W^2} = \frac{22333.44}{1801.33} = 12.39$$

- That is much larger than 1, but is it large enough? We go to R to compute the p-value.

Need Data to be Stacked

```
> foo
  Group.1 Group.2 Group.3 Group.4
1     338     235     164     290
2     155     238     197     343
3     239     251     136     294
4     184     229     214     373
5     185     233     148     306
6     261     232     230     357
> mydata=c(foo[,1],foo[,2],foo[,3],foo[,4])
> mydata
 [1] 338 155 239 184 185 261 235 238 251 229 233 232 164 197 136 214 148
230 290
[20] 343 294 373 306 357
```

R Output

```
> group=c(rep(1,6),rep(2,6),rep(3,6),rep(4,6))
> group=as.factor(group)
                                  s_B^2
> summary(aov(mydata~group))
            Df Sum Sq Mean Sq F value  Pr(>F)
group        3  67000   22333   12.39 8.37e-05 ***
Residuals   20  36040    1802
---                   s_W^2
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

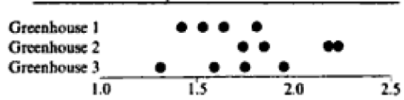

Conclusion

- The p-value is less than 0.05 so we reject the null hypothesis.
- There is enough evidence to conclude that the samples came from distributions with different means.

Example

- Kevin is growing tomatoes and has three green houses. He wonders if the yield is the same in each green house so he randomly selects 4 plants from each and measures the yield.

Greenhouse 1	1.42	1.64	1.81	1.53
Greenhouse 2	1.85	1.74	2.23	2.18
Greenhouse 3	1.75	1.31	1.95	1.59



Is there a difference in means? Possibly…

Questions

- Are these samples drawn from the same population, so that the differences among the observations are due entirely to sampling fluctuations from sample to sample?
- Or are the samples drawn from different populations, so that the differences among the observations are due to differences among the **mean tomato growth** in each greenhouse, as well as sampling fluctuations?

Aside: For R Need Stacked Data

- Our initial data looks like this

green1	green2	green3
1.42	1.85	1.75
1.64	1.74	1.31
1.81	2.23	1.95
1.53	2.18	1.59

- For R, we stack it

allgreen	house
1.42	1
1.64	1
1.81	1
1.53	1
1.85	2
1.74	2
2.23	2
2.18	2
1.75	3
1.31	3
1.95	3
1.59	3

R Output

```
>
allgreen=c(1.42,1.64,1.81,1.53,1.85,1.74,2.23,2.18,1.75,1.31,1.95,1.5
9)
> house=as.factor(c(rep(1,4),rep(2,4),rep(3,4)))
> summary(aov(allgreen~house))
            Df Sum Sq Mean Sq F value Pr(>F)
house        2 0.3800 0.19000    3.58 0.0718 .
Residuals    9 0.4776 0.05307
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

- Home Depot is thinking of offering install services for some popular items.
- In particular, they want to install water heaters.
- They set up a training system in four test markets.
- Are the install teams operating about the same?

The Data

- Time in Minutes to Install a Water Heater

Pittsburgh	*Richmond*	*Spokane*	*Topeka*
143	159	154	165
187	152	164	188
156	172	162	157
167	177	200	145
213	145	179	167
202	161	155	114

Questions

- Are these samples drawn from the same population, so that the differences among the observations are due entirely to sampling fluctuations from sample to sample?
- Or are the samples drawn from different populations, so that the differences among the observations are due to differences among the **mean install times** in the cities, as well as sampling fluctuations?

R Output


```
> pitt=c(143,187,156,167,213,202)
> rich=c(159,152,172,177,145,161)
> spokane=c(154,164,162,200,179,155)
> topeka=c(165,188,157,145,167,114)
> install.time=c(pitt,rich,spokane,topeka)
> group=c(rep(1,6),rep(2,6),rep(3,6),rep(4,6))
➢  group=as.factor(group)

> summary(aov(install.time~group))
            Df  Sum Sq  Mean Sq  F value  Pr(>F)
group        3   1668    556.0    1.222   0.328
Residuals   20   9098    454.9
```

Btw, there is a test for variances


```
> bartlett.test(install.time~group)

        Bartlett test of homogeneity of variances

data:  install.time by group
Bartlett's K-squared = 3.3663, df = 3, p-value = 0.3385
```

Conclusion

- The p-value is 0.3276 which is not below 0.05 so we fail to reject the null hypothesis.
- There is not enough evidence to conclude that the samples came from distributions with different means.

Example

- An apple juice manufacturer is planning to develop a new product -a liquid concentrate.
- The marketing manager has to decide how to market the new product.
- Three strategies are considered
  - Emphasize convenience of using the product.
  - Emphasize the quality of the product.
  - Emphasize the product's low price.

Example

- An experiment was conducted as follows:
  - In three cities an advertisement campaign was launched
  - In each city only one of the three characteristics (convenience, quality, and price) was emphasized.
  - The weekly sales were recorded for twenty weeks following the beginning of the campaigns.

The Data

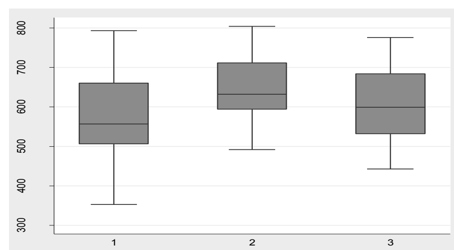
Weekly sales

Weekly sales

Weekly sales

Convnce	Quality	Price
529	804	672
658	630	531
793	774	443
514	717	596
663	679	602
719	604	502
711	620	659
[illegible]	697	689
[illegible]	706	675
[illegible]	615	512
[illegible]	492	691
[illegible]	719	733
604	787	698
495	[illegible]	776
485	[illegible]	561
557	[illegible]	572
353	[illegible]	469
557	[illegible]	581
542	580	679
614	624	532

The Boxplots



The ANOVA output in R

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

```
. > summary(aov(sales~group))
            Df Sum Sq Mean Sq F value Pr(>F)
group        2  57512   28756   3.233 0.0468 *
Residuals   57 506984    8894
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> bartlett.test(sales~group)

        Bartlett test of homogeneity of variances

data:  sales by group
Bartlett's K-squared = 0.73887, df = 2, p-value = 0.6911
```

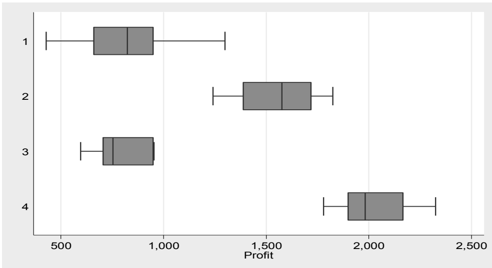
ANOVA popular research tool

- A famous Harvard Law Study

White Male		Black Male	White Female	Black Female
1300	853	1241	951	1899
646	727	1824	954	2053
951	559	1616	754	1943
794	429	1537	706	2168
661	1181		596	2325
824	853			1982
1038	877			1780
754				

Source: Ian Ayres. "Fair Driving: Gender and Race Discrimination in Retail Car Negotiations." *Harvard Law Review.* Vol. 104, No. 4, Feb. 1991.

The Data



R Output

```
> wm=c(1300,646,951,794,661,824,1038,754,853,727,559,429,1181,853,877)
> bm=c(1241,1824,1616,1537)
> wf=c(951,954,754,706,596)
> bf=c(1899,2053,1943,2168,2325,1982,1780)
> group=c(rep(1,length(wm)),rep(2,4),rep(3,5),rep(4,7))
> group=as.factor(group)
> alldata=c(wm,bm,wf,bf)
➢ summary(aov(alldata~group))

            Df  Sum Sq Mean Sq F value   Pr(>F)
group        3 8107321 2702440   61.37 3.51e-12 ***
Residuals   27 1189048   44039
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>

Bartlett's test for equal variances:  chi2(3) =   1.0207  Prob>chi2 = 0.796

.
```


Testing Equality of Variances

```
> bartlett.test(alldata~group)

        Bartlett test of homogeneity of variances

data:  alldata by group
Bartlett's K-squared = 1.0207, df = 3, p-value = 0.7962
```


Things you should know

- ANOVA is a useful procedure for testing the equality of 3 or more means.
- It assumes the population variances are all the same; if this isn’t the case there are nonparametric methods that can be used.
- If you reject the null hypothesis, it doesn’t imply all the means are unequal; just that there is some difference between them.

Example

The purpose of statistical inference is to provide information about the

a) sample based upon information contained in the population
b) population based upon information contained in the sample
c) population based upon information contained in the population
d) mean of the sample based upon the mean of the population

Example

Analysis of variance is used to test:

a) Whether *k* population variances are all equal.
b) Whether *k* population standard deviations are all equal.
c) Whether k population means are all equal.
d) Whether *k* sample means are all equal.

Example

The Central Limit Theorem is important in statistics because

a) For a large n, it says the population is approximately normally distributed.
b) For any population, it says the sampling distribution of the sample mean is approximately normal, regardless of the sample size.
c) For a large n, it says the sampling distribution of the sample mean is approximately normal, regardless of the shape of the population.
d) For any sized sample, it says the sampling distribution of the sample mean is approximately normal.

Example

In testing the hypotheses *H*o: μ = 50 vs. *H*a: μ > 50, the following information is known: $n = 64$, $\bar{x} = 53.5$, and s = 10. The test statistic equals:

a) 1.96
b) -2.80
c) 2.80
d) -1.96

Example

A survey claims that 9 out of 10 doctors recommend aspirin for their patients with headaches. To test this claim against the alternative that the actual proportion of doctors who recommend aspirin is less than 0.90, a random sample of 100 doctors' results in 83 who indicate that they recommend aspirin. The value of the test statistic in this problem is approximately equal to:

a) −1.67
b) −2.33
c) −1.96
d) −0.14

Example

A 95% confidence interval for the true proportion, p, is (0.23, 0.53). What is the best interpretation of the confidence interval?

a) We are 95% confident that the true proportion is 38%.
b) 95% of the time, the true proportion will fall between 23% and 53%.
c) In repeated sampling from this same population and calculating 95% confidence intervals, about 95% of these intervals will contain 38%.
d) In repeated sampling from this same population and calculating 95% confidence intervals, about 95% of these intervals will contain the true proportion.

Example

A recent study of 750 internet users in Europe found that 35% of internet users were women. What is the 95% confidence interval of the true proportion of women in Europe who use the internet?

a) (.349,.351)
b) (.321,.379)
c) (.316,.384)
d) (.309,.391)

Example

The National Center for Education would like to estimate the proportion of students who defaulted on their student loans for the state of Arizona. The total sample size needed to construct a 95% confidence interval for the proportion of student loans in default with a margin of error equal to 4% is ________.

a) 336
b) 416
c) 455
d) 601

Example

) The U.S. Department of Labor and Statistics wanted to compare the results of an unemployment program for the past two months in the U.S. Suppose the proportion of the unemployed two months ago is p2 and the proportion of the unemployed one month ago is p1. A study found a 95% confidence interval for p2–p1 is (-0.0012, 0.003). Give an interpretation of this confidence interval.

a) We are 95% confident that the proportion of the unemployed one month ago is between 0.12% less and 0.3% more than the proportion of the unemployed two months ago.
b) We are 95% confident that the proportion of the unemployed two months ago is between 0.12% less and 0.3% more than the proportion of the unemployed one month ago.
c) We know that 95% of the unemployed two months ago is between 0.12% less and 0.3% more than the unemployed one month ago.
d) We know that 95% of all random samples done on the population will show that the proportion of the unemployed two months ago is between 0.12% less and 0.3% more than the proportion of the unemployed one month ago.
e) We know that 95% of the unemployed one month ago is between 0.12% less and 0.3% more than the unemployed two months ago.

Example

According to research company comScore, Facebook users spent an average of 405 minutes on the site during the month of January 2012. Assume that this population has a standard deviation of 135 minutes. A random sample of 32 users was selected from this population. What is the probability that the average number of minutes on the site in January was less than 390 minutes?

a) 0.2643
b) 0.3669
c) 0.4801
d) 0.5398

Example

Suppose you want to test Ho: μ = 30 versus Ha: μ < 30. Which of the following possible sample results based on a sample size of 36 gives the strongest evidence to reject Ho in favor of Ha?

a) $\bar{x} = 28, s = 6$
b) $\bar{x} = 27, s = 4$
c) $\bar{x} = 32, s = 2$
d) $\bar{x} = 26, s = 9$

Example

You are looking at your calculator before an exam and are faced with the following hypothesis test:

Ho: The amount of power remaining in the battery is equal to a level which is high enough to last the length of the exam.
Ha: The amount of power remaining in the battery is less than the level which is high enough to last the length of the exam.

Which statement best describes a Type I error you could make with these hypotheses?

a) I assume the battery has enough power, when in fact it does not, and the battery dies during the exam.
b) I assume the battery has enough power, when in fact is does, and lasts throughout the exam.
c) I assume the battery will die, when in fact it does have sufficient power, and I replace a good battery with a new one.
d) I assume the battery will die, when in fact it does not have sufficient power, and I replace a bad battery with a new one.
e) I assume the battery has enough power, when in fact it does, but I still bring five extra calculators just in case.

Example

Three instructors teach different sections of an introductory-level economics class during the fall semester. The number of students in each section is:

Mankiw	Miron	Sumners
100	125	75

The null hypothesis is that all three instructors are equally popular. What is the value of the Chi-square goodness of fit statistic for this data?

Example

Suppose a 95% confidence interval for the proportion of Americans who exercise regularly is 0.29 to 0.37. Which one of the following statements is FALSE?

A. It is reasonable to say that more than 25% of Americans exercise regularly.
B. It is reasonable to say that more than 40% of Americans exercise regularly.
C. The hypothesis that 33% of Americans exercise regularly cannot be rejected.
D. It is reasonable to say that fewer than 40% of Americans exercise regularly.

Example

A hypothesis test is done in which the alternative hypothesis is that more than 10% of a population is left-handed. The p-value for the test is calculated to be 0.25. Which statement is correct?

A. We can conclude that more than 10% of the population is left-handed.
B. We can conclude that more than 25% of the population is left-handed.
C. We can conclude that exactly 25% of the population is left-handed.
D. We cannot conclude that more than 10% of the population is left-handed.