Math 110 Final Exam Review

YOU NEED TO KNOW

Population vs. Sample

A **population** is an entire collection of subjects that we want to know about. A **sample** is a subset of the population that we gather to make inferences about the population.

Parameter vs. Statistic

A **parameter** is a measurement from an entire population. A **statistic** is a measurement from a sample. Statistics are used to estimate parameters.

Experiments and Observational Studies

In an **experiment**, an **explanatory variable** is *manipulated by researchers*, and a **response variable** is observed to observe the effects. If no manipulation is occurring (where changes in the explanatory variable occur randomly), the experiment is **observational**.

Cause and Effect in Experiments

In an experiment, we can conclude that an explanatory variable is causing an effect if treatment and control groups are created to be similar. This is often done by random assignment of subjects into groups. Different treatment groups receive treatment at different levels. Control groups receive nothing or a placebo. The explanatory variable (the treatment level) is manipulated by assigning subjects into these groups where the variable is tested at different levels.

Blinding

If any group receives a placebo, no subject should know whether they are getting a placebo or a treatment. This is known as **blinding**. The best experiments are **double blind**, where neither the subjects nor those who implement the treatment know who receives placebo or who receives a treatment.

Generalizing to a Population

Results of an observational study can be generalized to a larger population if the sample is representative. Representative samples are often created by random selection, or by gathering a stratified sample.

Methods of Sampling

In a **random sample**, population members are chosen so that all members have the same chance of being selected. In a **stratified sample**, the population is divided into groups, then members are selected randomly from each group.

Types of Data

Data values are **categorical** if they are non-quantities. Quantities are **quantitative data** that are counted or measured.

Bias

Selection bias occurs when some population members are more likely to be selected than others. This occurs in **self-selecting** or **convenience samples**. **Response bias** occurs when the methods used to gather data influences the responses. **Non-response bias** occurs when many people refuse to be involved in a study, or they refuse to answer some questions.

TRY THESE

1. A medical doctor has forty patients with high cholesterol. She wants to compare the effectiveness of two treatments for high cholesterol. To make her comparison, she randomly assigns 20 patients to receive treatment A. She gives the remaining 20 patients treatment B. She gives the patients the treatment for 30 days and then measures their cholesterol level at the end.
   a. Is this an observational study or an experiment? Explain your answer.
   b. What is the explanatory variable? Is it categorical or quantitative?
   c. What is the response variable? Is it categorical or quantitative?
2. Researchers want to compare meditation to exercise and how they are related to stress levels. They randomly selected 100 people who regularly meditate, and 100 people who exercise regularly. Stress levels were measured, and the people who exercised regularly had lower stress levels, on average.
   a. Is this an experiment or an observational study? Explain your answer.
   b. Can we conclude that exercise causes people to have lower stress levels than meditation?
   c. What is the explanatory variable?
   d. What is the response variable?
3. You are designing a clinical trial to see whether adding calcium to the diets for middle-aged men, will reduce their blood pressure. 100 subjects have their blood pressure measured by a physician and then they are randomly assigned to two groups. One group receives calcium pills and the other group receives a placebo. After two weeks, each patient returns and their blood pressure is checked again.
   a. Which group is the control group?
   b. Describe how blinding is implemented in this experiment? What is the purpose of this?
   c. How could you make this experiment double blind?
   d. Suppose that the group which received a placebo had no change in average blood pressure, but the group that received the calcium pills reduced their average blood pressure significantly. Can we conclude that calcium lowers blood pressure?
4. City High School has 2000 students. The editor of the school newspaper wishes to determine whether students prefer having football games on Friday evenings (as they currently are at the school) or having them on Saturday evenings. The editor plans to conduct a survey of students in order to determine their preference.
   a. What is the population in the study?
   b. What is the sample?
   c. Since high school students take English all four years, the editor decided to randomly select 5 students from all English classes in each grade level (Freshman, Sophomore, Junior and Senior).
   d. What type of sample is this?
   e. Would this sample result in a representative sample of students at City High School?
   f. Suppose instead that the editor decided to take the survey by asking 100 students that eat near him in the Senior Quad at lunch. What type of sample is this?
   g. Does this method introduce bias? If so, what type? Explain your answer.
5. The administration at Mt. SAC wants to know how many hours a typical student at Mt. SAC spends on campus per week. To find this out, they take a list of all the students at Mt. SAC and randomly select 500 students and ask them how much time they spend on campus in a week.
   a. What type of sample is this?
   b. What is the population of interest?
   c. Will the sample be representative of the population from (b)? Explain your answer.
   d. Can the results be generalized to the population of interest (the one stated in (b))? Explain your answer.
   e. Can the administration use the results to generalize over all community college students in the USA? Explain your answer.

YOU NEED TO KNOW

To measure the center of a distribution of quantities, we use the sample mean, or median. The **sample mean** of a sample of quantitative values is: $\bar{x} = \frac{\sum x}{n}$. The **median** is the value in the middle of a list of sorted quantities. If there is an even number of quantities, the median is the mean of the two middle values.

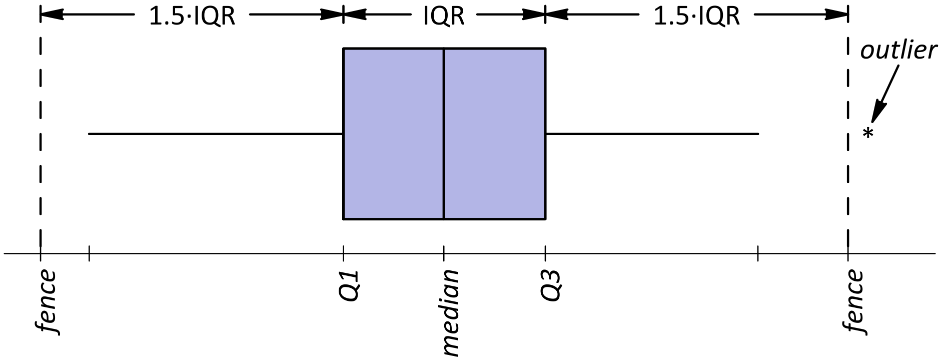
To measure the **spread** of a distribution of quantities, we use the range or the sample standard deviation. The **range** of a sample of quantities is (*maximum value*) – (*minimum value*). The **sample standard deviation** of a sample of quantitative values is:

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}.$$

The **lower quartile**, Q_1, is the median of the lower ½ of a data set. The **upper quartile**, Q_3, is the median of the upper ½ of a data set. The upper and lower halves exclude the middle value of a data set, if there is one. Sometimes we measure the spread of a distribution of values using the interquartile range (IQR). The interquartile range is:

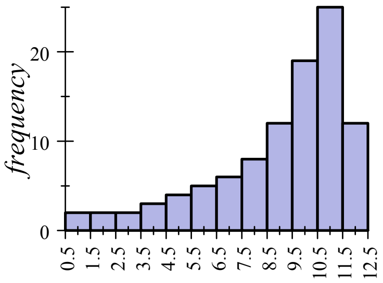
$$\text{IQR} = Q_3 - Q_1.$$

A **boxplot** is constructed using the values of the 5-point summary. If a value is less than the **lower fence** (Q1 – 1.5·IQR) or greater than **upper fence** (Q3 + 1.5·IQR), then it is an **outlier**. The lines extend on each side of the box to the farthest point that is not an outlier. Outliers are marked with asterisks (*).

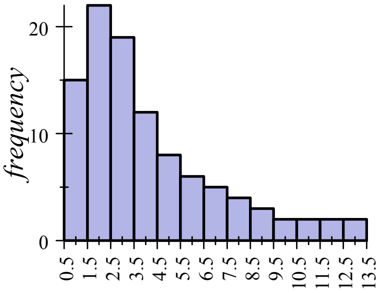


A **histogram** assigns a bar to each grouping of values in a frequency distribution. The height of each bar is the (possibly *relative*) **frequency** (the number of values) in the corresponding grouping.

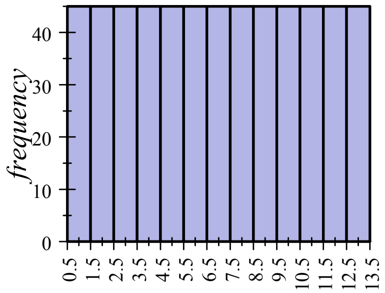
The most important features of a distribution of frequencies (represented by a histogram) are its **shape**, **center** (summarized by the mean or median), and **spread** (summarized by the standard deviation or range). The most important distribution shapes are summarized below.



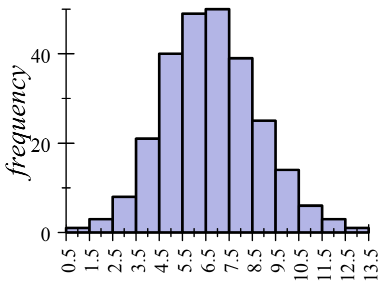
Skewed-Left



Skewed-Right



Uniform



Bell-Shaped

When a distribution is skewed left, the mean tends to be less than the median. When it is skewed right, the mean tends to be greater than the median. If the distribution is symmetric (not skewed) the mean and median are similar.

TRY THESE

6. The sorted data below represent the heights (in inches) of a random sample of 19 female Mt SAC students.

   47.7, 57.8, 57.9, 58.4, 59.4, 59.6, 61.7, 62.9, 62.9, 63.6, 64.1, 64.8, 64.9, 65.9, 66.9, 68.2, 68.7, 70.3, 70.7

   a. Compute the sample mean.
   b. Compute and interpret the sample standard deviation.
   c. Compute the median.
   d. Compare the mean and median. What information does this give about the shape of the values' distribution.
   e. Give the five point summary, and the inter-quartile range.
   f. What are the values of the upper and lower fences for outliers?
   g. Sketch a boxplot, noting outliers. Show your work.
   h. Construct a frequency table & histogram, with a first lower class limit of 47.7 and a class width of 4.7. Use 5 classes.

Class	Tally	Frequency
47.7 – 52.3		
52.4 – 57.0		
57.1 – 61.7		
61.8 – 66.4		
66.5 – 71.1		

   i. Describe the shape and spread of the distribution. Does the shape agree with your comparison in question (d)?

YOU NEED TO KNOW

In a correlation and regression problem, a scatterplot can reveal the relationship between two quantitative variables. If the points have a linear pattern that rises from lower left to upper right, the relationship is **positive** and *r* ≈ 1. If they have a linear pattern that falls from upper left to lower right, the relationship is **negative** and *r* ≈ –1. Stronger linear relationships have points closer to a line, and *r* is closer to ±1. If the relationship is weak very non-linear, then *r* ≈ 0.

The correlation coefficient is:

$$r = \frac{\sum Z_x Z_y}{n-1}.$$

Use your calculator's built-in statistics function to compute *r*.

When there is a moderate or strong linear relationship between the variables, the least-squares regression line can be used to predict values of *y* which correspond to chosen values of *x*. The equation of the line is

$$\hat{y} = slope \cdot x + yint.$$

The slope and *y*-intercept are given by the formulas below.

$$slope = \frac{r \cdot s_y}{s_x}$$
$$yint = \bar{y} - slope \cdot \bar{x}$$

TRY THESE

7. Below are amounts of sugar in a serving of 15 different breakfast cereals. With each sugar amount, the corresponding Consumer Reports rating is given (on a 1 to 100 scale).

Sugar (grams)	6	8	7	11	11	3	3	12	12	2	14	8	9	5	3
Rating	40	37	58	34	37	39	53	27	36	60	33	34	24	59	47

a. Sketch a scatterplot of these data points. Describe the visual appearance of the relationship between sugar and rating in terms of form, strength, and direction.
b. What is the correlation coefficient? Use this to argue whether your description above is accurate.
c. Is the relationship between sugar and rating weak, moderate, or strong?
d. Use your calculator to determine the values of $\bar{x}$, $\bar{y}$, s_x, s_y, and *r*.

$\bar{x}$ = _________, $\bar{y}$ = _________, s_x = _________, s_y = _________, *r* = _________

e. Use the values above to compute the slope of the regression line. Show your work.
f. Use the values above to compute the *y*-intercept of the regression line.
g. Give the equation of the regression line. Compute a few points then graph the line on your scatterplot above.
h. If a cereal has 4 grams of sugar, what would you predict the consumer reports rating to be?

Retailers report that the use of coupons is increasing. The Scripps News Service reported that from a random sample of 800 U.S. households, 616 use coupons on a weekly basis.

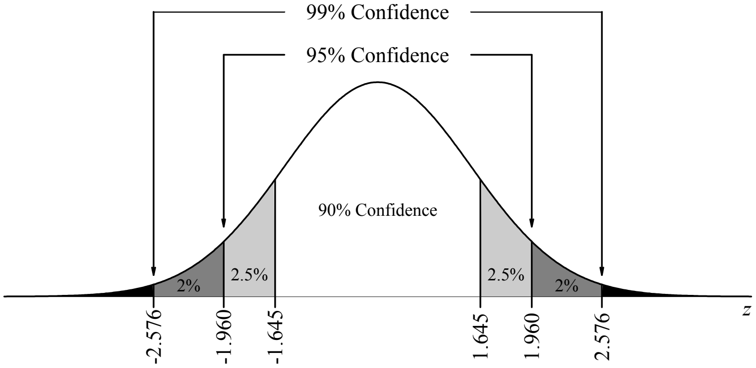
8. What variable was measured in this sample?
9. Is this variable quantitative or categorical?
10. With this type of variable, what summary statistic is most appropriate – a sample mean or sample proportion?
11. Compute the statistic you feel is most appropriate. Use the appropriate symbol to represent this value.
12. If we use this statistic to represent a population parameter, what would it be? Use the appropriate symbol.
13. Is the statistic you computed equal to the population parameter? If not, what could we do to account for any difference?

YOU NEED TO KNOW

To compute a confidence interval for a population proportion, *p*:

a. Verify that the criteria for approximate normality are met: at least 10 successes and 10 failures.

b. Choose a level of confidence, and this determines a positive *Z*-score that we denote $Z_{\alpha/2}$. We need some reference to find this value, such as the picture below. We usually use Table A1 (or A2) to find such values.



c. The standard error for a sample proportion is approximated by $\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$.

d. The margin of error for a sample proportion is $E = Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$.

e. The confidence interval is: $\hat{p} - E < p < \hat{p} + E$.

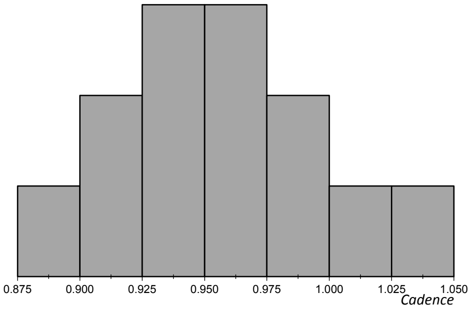
---

14. Follow steps *a* through *e* above to compute a 90% confidence interval for the population proportion.
15. Interpret your confidence interval.

NEXT STEPS

A study of the ability of individuals to walk in a straight line reported the accompanying data on cadence (strides per sec) for a sample of randomly selected healthy men. The histogram depicts these values.

0.95	0.85	0.92	0.95	0.90	0.98	1.00
0.91	0.96	1.01	0.93	1.04	0.96	0.96



16. What variable was measured in this sample?
17. Is this variable quantitative or categorical?
18. With this type of variable, what summary statistic is most appropriate – a sample mean or sample proportion?
19. Compute the most appropriate summary statistic. Use the appropriate symbol to represent this value.
20. If we use this statistic to estimate a parameter, what would it be? Use the appropriate symbol.
21. Is there error in this statistic?
22. Why can't we know the value of this error?
23. What do we compute in place of the true error in a statistic?

YOU NEED TO KNOW

To compute a confidence interval for a population mean, μ:

a. The criteria for approximate normality for the sampling distribution require that the sample size is bigger than 30, or the population that the sample is drawn from is normal.

b. Choose a level of confidence, and this determines a *T*-score that we denote $T_{\alpha/2}$. Use Table A2 to find this. Remember, degrees of freedom are $n - 1$.

c. The standard error for a sample mean is approximated by $\frac{s}{\sqrt{n}}$.

d. The margin of error for a sample mean is $E = T_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$.

e. The confidence interval is $\bar{x} \pm E$, or $\bar{x} - E < \mu < \bar{x} + E$.

24. Follow steps *a* through *e* above to compute a 95% confidence interval for the population mean.
25. Interpret your confidence interval.

NEXT STEPS

Drug testing of job applicants is becoming increasingly common. The AP reported that, from a random sample of 600 California applicants, 73 tested positive.

26. What is the variable for this context, and is it quantitative or categorical?
27. We will test a claim with these data. Does the hypothesis involve a population proportion or a mean?
28. Does the context involve one population or two?

YOU NEED TO KNOW

To test a claim about a proportion from a single population, *p*:

a. The null hypothesis gives a hypothetical value for *p*: H_0: *p* = *value*. Use this to verify the criteria for approximate normality: $np \geq 10$ and $nq \geq 10$ (where $q = 1 - p$)

b. The test statistic is below. Use *p* from the null hypothesis, and $q = 1 - p$.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}.$$

c. Look up critical value(s) on the bottom row (*Z*) of Table A2.

d. Compute a *P*-value using Table A1. This is the *Z*-distribution *P*-value table.

e. If the *P*-value is less than the level of significance, reject the null hypothesis in favor of the alternative. Otherwise, fail to reject (but *never* support) the null hypothesis.

29. We want to test the claim at 5% significance that more than 10% of California applicants test positive in their drug screening. What are the null and alternative hypotheses for this test?
30. Are the criteria for approximate normality satisfied?
31. What is the standard error for the distribution of sample proportions?
32. What are the test statistic and critical value(s) for this test?
33. Give the *P*-value for this test.
34. What conclusion do you make regarding the null and alternative hypotheses?
35. Give a conclusion in the context of this problem.

NEXT STEPS

Next, we test a hypothesis using the data on cadence for adult men from Questions 16 through 25. We will use a 5% level of significance.

36. Would these data be used a claim about one proportion, two proportions, one mean, or two means?

YOU NEED TO KNOW

To test a claim about a proportion from a single mean, μ:

a. The null hypothesis gives a hypothetical value for μ: H_0: μ = *value*.
b. Verify the criteria for approximate normality for the sampling distribution of sample means. These require that $n > 30$, or the population that we're sampling from is normal.
c. The test statistic is below. This is a *T*-test because the population standard deviation is unknown.

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

d. Look up critical value(s) in Table A2 using $n - 1$ degrees of freedom.
e. Compute a *P*-value using Table A3. This is the *T*-distribution *P*-value table.
f. If the *P*-value is less than the level of significance, reject the null hypothesis in favor of the alternative. Otherwise, fail to reject (but *never* support) the null hypothesis.

37. We want to test the claim that the mean cadence for healthy mean is greater than 0.92 strides per second. What are the appropriate null and alternative hypotheses?
38. Are the criteria satisfied for the approximate normality of the sampling distribution of sample means? Explain your answer.
39. What are the test statistic and critical value(s) for this hypothesis test.
40. Compute the *P*-value for this hypothesis test.
41. Use the *P*-value to make a decision regarding the null and alternative hypotheses.
42. Write a conclusion in the context of the problem.

YOU NEED TO KNOW

A Type I Error occurs when we reject a true null hypothesis.
A Type II Error occurs when we fail to reject a false null hypothesis.

43. Suppose you reached the wrong conclusion in the test you just performed. Describe how this could happen, and what type of error you made.

NEXT STEPS

The article, "Truth and DARE", compared drug use for 238 randomly selected high school seniors exposed to a drug education program (DARE) and 335 high school seniors who were not exposed to such a program. Data for marijuana use is given in the accompanying table.

	n	Number who use marijuana
Exposed to DARE	288	141
Not exposed to DARE	335	181

44. What variable is being measured in each sample? Is this quantitative or categorical?
45. We will to test a claim about these data at 5% significance. Would this be a claim about one mean, two means, one proportion, or two proportions?

YOU NEED TO KNOW

To test a claim comparing two population proportions p_1 *and* p_2:

a. The null hypothesis is H_0: $p_1 = p_2$.
b. The criteria for approximate normality for the sampling distribution require at least 10 successes and failures in each sample.
c. The test statistic is below. This uses $\bar{p} = \frac{x_1+x_2}{n_1+n_2}$, and $\bar{q} = 1 - \bar{p}$.

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\cdot\bar{q}}{n_1} + \frac{\bar{p}\cdot\bar{q}}{n_2}}}.$$

d. Look up *Z* critical values(s) on the bottom row (*Z*) of Table A2.
e. Compute a *P*-value using Table A1 (this is for *Z* distribution *P*-values).
f. If the *P*-value is less than the level of significance, reject the null hypothesis in favor of the alternative. Otherwise, fail to reject (but *never* support) the null hypothesis.

46. Are the criteria for the approximate normality of the sampling distribution satisfied for this example?
47. We want to test the claim that the proportion of children who smoke marijuana is lower for those who attended the DARE program. What are the null and alternative hypotheses for this example?
48. What are the sample proportions, and the pooled proportion?
49. What are the test statistic and critical value(s)?
50. Sketch the test statistic on an appropriate probability distribution. Shade the region that represents the *P*-value.
51. What is the *P*-value?
52. What conclusion do you make regarding the null and alternative hypotheses?
53. Give a conclusion in the context of this problem.

NEXT STEPS

The article “Workaholism in Organizations: Gender Differences” (*Sex Roles*, 1999) gave the following data on 1996 income for random samples of female and male MBA graduates from a particular Canadian business school.

Gender	n	$\bar{x}$	s
Female	233	\$105,156	\$98,525
Male	258	\$133,442	\$131,090

We will test the claim at 6% significance that the mean salaries for female and male MBAs from this college are different.

54. Is this a claim about one mean, one proportion, two means, or two proportions?

YOU NEED TO KNOW

To test a claim comparing two population means μ_1 *and* μ_2:

a. The null hypothesis is H_0: $\mu_1 = \mu_2$.
b. The criteria for approximate normality for the sampling distribution require that the sample sizes are greater than 30, or both populations are normal.

c. The test statistic is below.

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

d. The degrees of freedom for this test are complicated. These will be provided for you. Use this to find the critical value(s) in Table A2.

e. Look up the *P*-value using Table A3. This is the *T*-distribution *P*-value table.

f. If the *P*-value is less than the level of significance, reject the null hypothesis in favor of the alternative. Otherwise, fail to reject (but *never* support) the null hypothesis.

---

55. What are the appropriate null and alternative hypotheses for this test?
56. Are the criteria for approximate normality of the sampling distribution satisfied?
57. The degrees of freedom for this test are 473. What are the test statistic and critical value(s) for this test?
58. Sketch the test statistic on its probability distribution. Shade the area of the tail it occupies.
59. What is the *P*-value for the test?
60. What conclusion do you make regarding the null and alternative hypotheses?
61. Give a conclusion in the context of this problem.

NEXT STEPS

The article “Factors Associated with Sexual Risk Taking Among Adolescents” (*Marriage and Family*, 1994) examined the relationship between gender and contraceptive use by sexually active teens. Each person in a random sample of sexually active teens was classified according to gender and contraceptive use.

Contraceptive Use	Female	Male
Rarely/Never	210	350
Sometimes/Most Times	190	320
Always	400	530

We will test the claim that contraceptive use is independent of gender at 5% significance.

62. How many variables are observed in this data set, and what are they?
63. Are these variables quantitative or categorical?

YOU NEED TO KNOW

The test for independence of two categorical variables is a right-tailed χ² test.

a. The null hypothesis states that the explanatory variable is independent of the response variable. The alternative hypothesis states that they are dependent.

b. For each observed frequency, *O*, in the table, an expected frequency, *E*, must be computed.

$$E = \frac{(row\ total) \cdot (column\ total)}{(grand\ total)}$$

c. The χ² test statistic is: $\chi^2 = \sum \frac{(O-E)^2}{E}$.

d. The statistic above is distributed, approximately, according to the χ² distribution as long as each expected frequency is at least 5.

e. The degrees of freedom for this right-tailed test are *df* = (*rows* – 1)·(*columns* – 1).

f. This test is always right tailed. Look up the critical value on Table A4.

g. Reject the null hypothesis and support the alternative if the test statistic is greater than the critical value.

64. What is the explanatory variable for this study?
65. What is the response variable?
66. Add a total row and column to the two-way table. Compute the expected frequencies and enter them below the observed frequencies in the table.
67. Are the requirements satisfied for the statistic to be distributed, approximately, according to the χ^2 distribution?
68. Enter the observed and expected frequencies below. Compute the contribution of each pair to the χ^2 statistic, and total these values.

O	E	$\frac{(O-E)^2}{E}$
	Total:	

69. What is the χ^2 statistic?
70. What are the degrees of freedom for this test?
71. Sketch a χ^2 distribution, plot the test statistic, and shade the tail that represents the *P*-value for this test.
72. What is the critical value?
73. What conclusion do you make regarding the null and alternative hypotheses?
74. Give a conclusion in the context of this problem.

NEXT STEPS

The article "Compression of Single Wall Corrugated Shipping Containers using Fixed and Floating Text Platens" (*Testing and Evaluation*, 1992) describes and experiment in which several different types of boxes were compared with respect to compression strength. The results are below for four types of boxes. We assume that the samples were drawn from normal populations with equal variances.

Type of Box	Sample Mean	Sample SD	Sample Size
A	713.00	46.55	6
B	756.93	40.34	6
C	698.07	37.20	6
D	562.02	39.87	6

We want to test the claim, at 1% significance that the mean compression strengths are all equal.

75. Do the statistics above summarize quantitative or categorical data?
76. Are we comparing means or proportions for these four samples?

YOU NEED TO KNOW

To compare means from $k \geq 3$ samples, use Analysis of Variance (ANOVA).

a. The test requires that the samples are drawn from normal populations with equal variances.
b. The null hypothesis states that the means from the populations are all equal. The alternative hypothesis states that at least one mean is different from the others.
c. Variance is equal to (*standard deviation*)^2.

d. $s_{\bar{x}}^2$ is the variance of the sample means.
e. s_p^2 is the mean of the sample variances.
f. The simplified version of the test assumes each sample is of the same size, *n*.
g. The test statistic is: $F = \frac{n \cdot s_{\bar{x}}^2}{s_p^2}$.
h. The *F* test is a right-tailed test.
i. The degrees of freedom for the numerator are $df_1 = k - 1$.
j. The degrees of freedom for the denominator are $df_2 = k(n - 1)$.
k. Look up the critical value on Table A5. This is always a right tailed test.
l. Reject the null hypothesis if the test statistic is greater than the critical value.

---

77. What are the null and alternative hypotheses for this test?
78. What is $s_{\bar{x}}^2$?
79. What is s_p^2?
80. What is the value of the *F* test statistic?
81. What are the degrees of freedom for the variance in the numerator?
82. What are the degrees of freedom for the variance in the denominator?
83. Sketch an *F*-distribution, and plot the test statistic. Shade the area of the region that represents the *P*-value.
84. What is the critical value for this test?
85. What conclusion do you make regarding the null and alternative hypotheses?
86. Give a conclusion in the context of this problem.