Tests of Significance:

Chi-Square test and T-Test

Significance Testing

- Also called “hypothesis testing”
- Objective: to test a claim about parameter μ
- Procedure:
  A. State hypotheses H_0 and H_a
  B. Calculate test statistic
  C. Convert test statistic to P-value and interpret
  D. Consider significance level (optional)

Hypotheses

- ***H_0* (null hypothesis)** claims "no difference"
- ***H_a* (alternative hypothesis)** contradicts the null
- Example: We test whether a population gained weight on average…

  H_0: no average weight gain *in population*
  H_a: *H_0* is wrong (i.e., "weight gain")
- Next $\Rightarrow$ collect data $\Rightarrow$ quantify the extent to which the data provides evidence against *H_0*

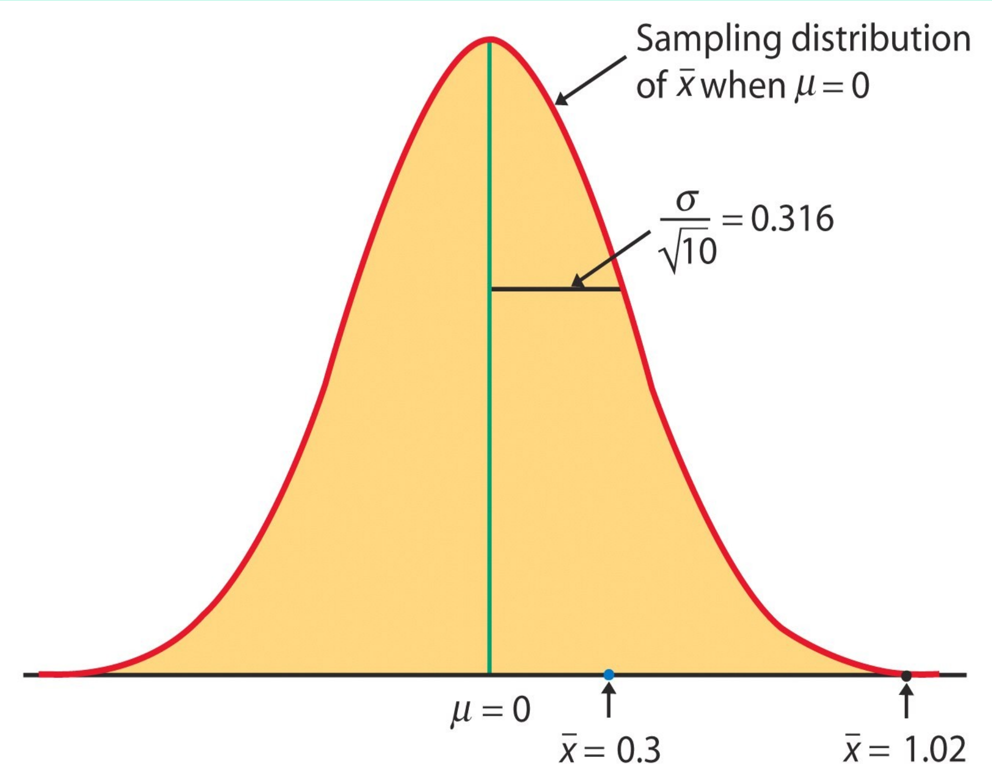
One-Sample Test of Mean

- To test a single mean, the null hypothesis is H_0: $\mu = \mu_0$, where μ_0 represents the "null value" (null value comes from the research question, not from data!)
- The alternative hypothesis can take these forms:
H_a: $\mu > \mu_0$ (one-sided to right) **or**
H_a: $\mu < \mu_0$ (one-side to left) **or**
H_a: $\mu \neq \mu_0$ (two-sided)
- For the weight gain illustrative example:
H_0: $\mu = 0$
H_a: $\mu > 0$ (one-sided) *or* H_a: $\mu \neq \mu_0$ (two-sided)
Note: $\mu_0 = 0$ in this example

Illustrative Example: Weight Gain

- Let X ≡ weight gain
- X ~N(μ, σ = 1), the value of μ unknown
- Under H_0, μ = 0
- Take SRS of n = 10
- $\sigma_{x\text{-bar}} = 1 / \sqrt{(10)} = 0.316$
- Thus, under H_0 x-bar~N(0, 0.316)

Figure: Two possible xbars when H_0 true



T-Tests

Independent Samples

T-Tests of Independence

- Used to test whether there is a significant difference between the means of two samples.

- We are testing for independence, meaning the two samples are related or not.

- This is a one-time test, not over time with multiple observations.

T-Test of Independence

- Useful in experiments where people are assigned to two groups, when there should be no differences, and then introduce Independent variables (treatment) to see if groups have real differences, which would be attributable to introduced X variable. This implies the samples are from different populations (with different μ).
- This is the Completely Randomized Two-Group Design.

For example, we can take a random set of independent voters who have not made up their minds about who to vote for in the 2004 election. But we have another suspicion:

H_1: watching campaign commercials increases consumption of Twinkies (snackie cakes), or $\mu_1 \neq \mu_2$

Null is $\mu_1 = \mu_2$

After one group watches the commercials, but not the other, we measure Twinkie in-take. We find that indeed the group exposed to political commercials indeed ate more Twinkies. We thus conclude that political advertising leads to obesity.

Two Sample Difference of Means T-Test

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\left[\sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}}\left[\frac{n_1 + n_2}{n_1 n_2}\right]\right]}$$

Sp2 = $$\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}$$ Pooled variance of the two groups

$$\left[\frac{n_1 + n_2}{n_1 n_2}\right]$$ = common standard deviation of two groups

Two Sample Difference of Means T-Test

- The nominator of the equation captures difference in means, while the denominator captures the variation within and between each group.

- Important point: of interest is the difference between the sample means, not sample and population means. However, rejecting the null means that the two groups under analysis have different population means.

An example

- Test on GRE verbal test scores by gender:

Females: mean = 50.9, variance = 47.553, n=6

Males: mean=41.5, variance= 49.544, n=10

$$t = \frac{50.9 - 41.5}{\left[\sqrt{\frac{(6-1)47.553 + (10-1)49.544}{6+10-2}\left[\frac{6+10}{6(10)}\right]}\right]}$$

$$t = \left[\frac{9.4}{\sqrt{48.826(.26667)}}\right]$$

$$t = \frac{9.4}{\sqrt{13.02}}$$

$$t = \frac{9.4}{3.608} = 2.605$$

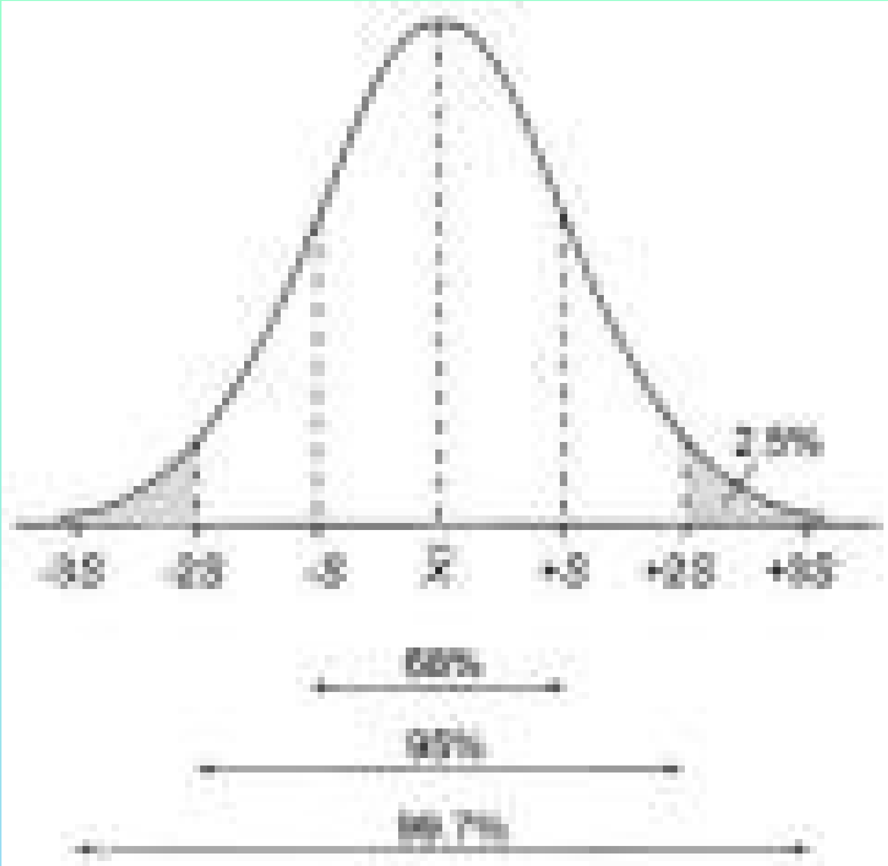
Now what do we do with this obtained value?

Steps of Testing and Significance

1. Statement of null hypothesis: if there is not one then how can you be wrong?
2. Set Alpha Level of Risk: .10, .05, .01
3. Selection of appropriate test statistic: T-test, chi2, regression, etc.
4. Computation of statistical value: get obtained value.
5. Compare obtained value to critical value: done for you for most methods in most statistical packages.

Steps of Testing and Significance

6. Comparison of the obtained and critical values.
7. If obtained value is more extreme than critical value, you may reject the null hypothesis. In other words, you have significant results.
8. If point seven above is not true, obtained is lower than critical, then null is not rejected.



The critical values are set by moving toward the tails of the distribution. The higher the significance threshold, the more space under the tail.

Also, hypothesis testing can entail a one or two-tailed test, depending on if a hypothesis is directional (increase/decrease) in nature.

Steps of Testing and Significance

- The curve represents all of the possible outcomes for a given hypothesis.

- In this manner we move from talking about a distribution of data to a distribution of potential values for a sample of data.

Paired T-Tests

- **We use Paired T-Tests, test of dependence, to examine a single sample subjects/units under two conditions, such as pretest - posttest experiment.**

- For example, we can examine whether a group of students improves if they retake the GRE exam. The T-test examines if there is any significant difference between the two studies. If so, then possibly something like studying more made a difference.

$$\frac{\sum D}{\sqrt{\dfrac{n\sum D^2 - (\sum D^2)}{(n-1)}}}$$

ΣD = sum differences between groups, plus it is squared.

n = number of paired groups

Paired T-Tests

- Unlike a test for independence, this test requires that the two groups/samples being evaluated are dependent upon each other.

- For example, we can use a paired t-test to examine two sets of scores across time as long as they come from the same students.

- If you are doing more than two groups, use ANOVA.

The Chi-Square Test

Parametric and Nonparametric Tests (cont.)

- The term "non-parametric" refers to the fact that the chi-square tests do not require assumptions about population parameters nor do they test hypotheses about population parameters.
- Previous examples of hypothesis tests, such as the t tests and analysis of variance, are **parametric tests** and they do include assumptions about parameters and hypotheses about parameters.

Parametric and Nonparametric Tests (cont.)

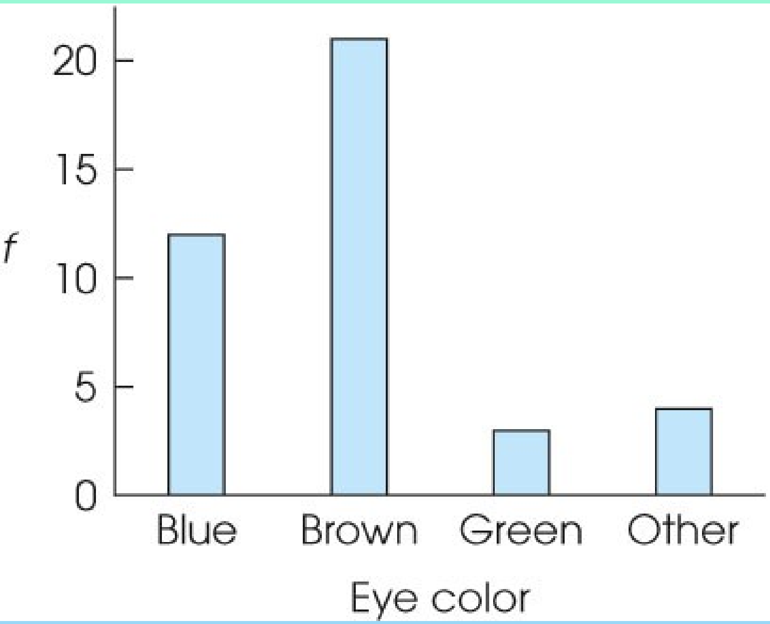
- The most obvious difference between the chi-square tests and the other hypothesis tests we have considered (t and ANOVA) is the nature of the data.
- For chi-square, the data are frequencies rather than numerical scores.

The Chi-Square Test for Goodness-of-Fit

- The **chi-square test for goodness-of-fit** uses frequency data from a sample to test hypotheses about the shape or proportions of a population.
- Each individual in the sample is classified into one category on the scale of measurement.
- The data, called **observed frequencies**, simply count how many individuals from the sample are in each category.

The Chi-Square Test for Goodness-of-Fit (cont.)

- The null hypothesis specifies the proportion of the population that should be in each category.
- The proportions from the null hypothesis are used to compute **expected frequencies** that describe how the sample would appear if it were in perfect agreement with the null hypothesis.



Eye color (*X*)	*f*
Blue	12
Brown	21
Green	3
Other	4

Blue	Brown	Green	Other
12	21	3	4

The Chi-Square Test for Independence

- The second chi-square test, the **chi-square test for independence**, can be used and interpreted in two different ways:
  1. Testing hypotheses about the relationship between two variables in a population, or
  2. Testing hypotheses about differences between proportions for two or more populations.

The Chi-Square Test for Independence (cont.)

- Although the two versions of the test for independence appear to be different, they are equivalent and they are interchangeable.
- The first version of the test emphasizes the relationship between chi-square and a correlation, because both procedures examine the relationship between two variables.

The Chi-Square Test for Independence (cont.)

- The second version of the test emphasizes the relationship between chi-square and an independent-measures t test (or ANOVA) because both tests use data from two (or more) samples to test hypotheses about the difference between two (or more) populations.

The Chi-Square Test for Independence (cont.)

- The first version of the chi-square test for independence views the data as one sample in which each individual is classified on two different variables.
- The data are usually presented in a matrix with the categories for one variable defining the rows and the categories of the second variable defining the columns.

The Chi-Square Test for Independence (cont.)

- The data, called **observed frequencies**, simply show how many individuals from the sample are in each cell of the matrix.
- The null hypothesis for this test states that there is no relationship between the two variables; that is, the two variables are independent.

The Chi-Square Test for Independence (cont.)

- The second version of the test for independence views the data as two (or more) separate samples representing the different populations being compared.
- The same variable is measured for each sample by classifying individual subjects into categories of the variable.
- The data are presented in a matrix with the different samples defining the rows and the categories of the variable defining the columns..

The Chi-Square Test for Independence (cont.)

- The data, again called **observed frequencies**, show how many individuals are in each cell of the matrix.
- The null hypothesis for this test states that the proportions (the distribution across categories) are the same for all of the populations

The Chi-Square Test for Independence (cont.)

- Both chi-square tests use the same statistic. The calculation of the chi-square statistic requires two steps:

1. The null hypothesis is used to construct an idealized sample distribution of **expected frequencies** that describes how the sample would look if the data were in perfect agreement with the null hypothesis.

The Chi-Square Test for Independence (cont.)

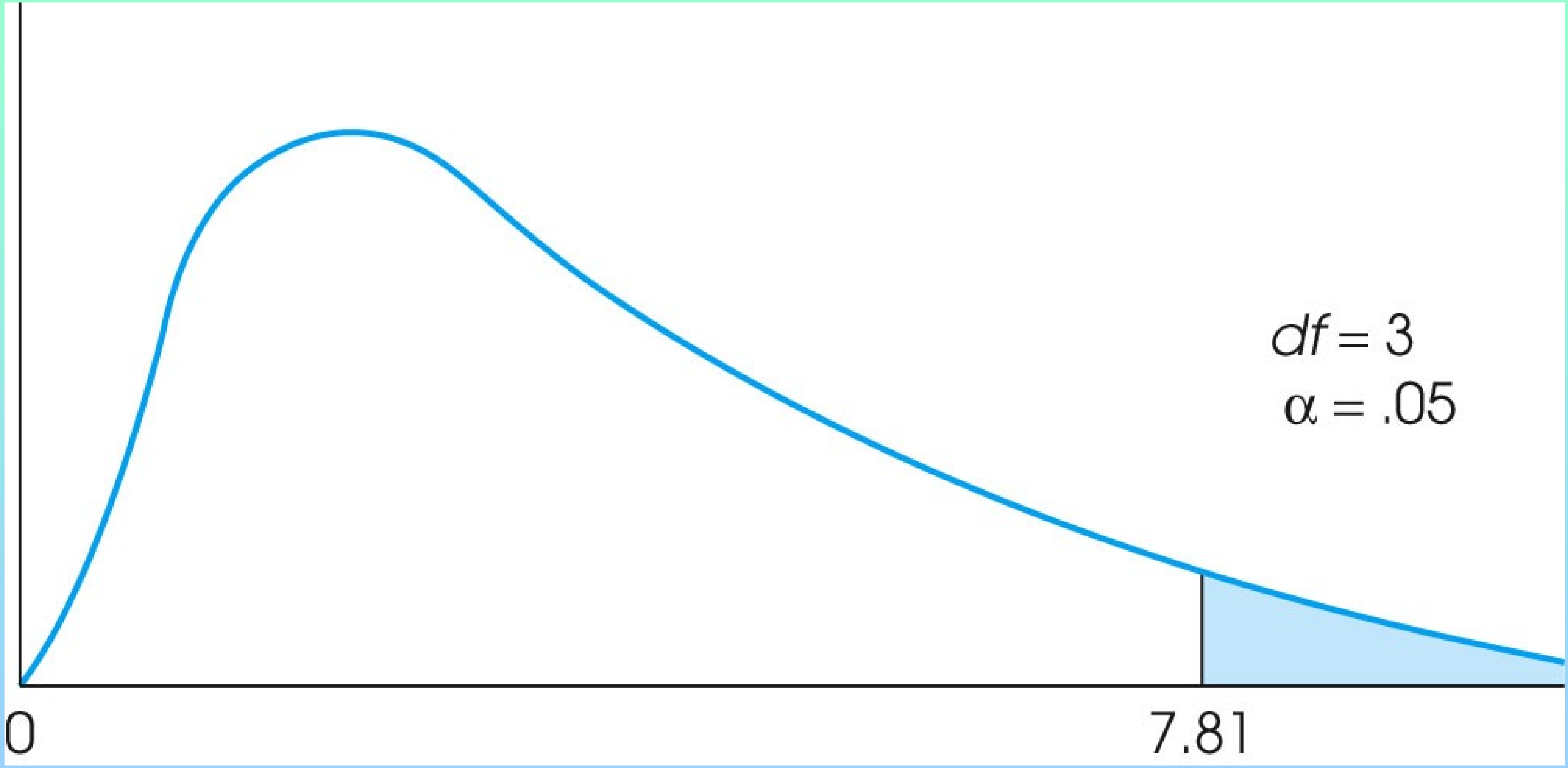
For the goodness of fit test, the expected frequency for each category is obtained by

$$\text{expected frequency} = f_e = pn$$

(p is the proportion from the null hypothesis and n is the size of the sample)

For the test for independence, the expected frequency for each cell in the matrix is obtained by

$$\text{expected frequency} = f_e = \frac{(\text{row total})(\text{column total})}{n}$$


df = 3
α = .05
0
7.81

The Chi-Square Test for Independence (cont.)

2. A chi-square statistic is computed to measure the amount of discrepancy between the ideal sample (expected frequencies from H_0) and the actual sample data (the observed frequencies = f_o).

   A large discrepancy results in a large value for chi-square and indicates that the data do not fit the null hypothesis and the hypothesis should be rejected.

The Chi-Square Test for Independence (cont.)

The calculation of chi-square is the same for all chi-square tests:

$$\text{chi-square} = \chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

The fact that chi-square tests do not require scores from an interval or ratio scale makes these tests a valuable alternative to the t tests, ANOVA, or correlation, because they can be used with data measured on a nominal or an ordinal scale.

Measuring Effect Size for the Chi-Square Test for Independence

- When both variables in the chi-square test for independence consist of exactly two categories (the data form a 2x2 matrix), it is possible to re-code the categories as 0 and 1 for each variable and then compute a correlation known as a **phi-coefficient** that measures the strength of the relationship.

Measuring Effect Size for the Chi-Square Test for Independence (cont.)

- The value of the phi-coefficient, or the squared value which is equivalent to an r^2, is used to measure the effect size.
- When there are more than two categories for one (or both) of the variables, then you can measure effect size using a modified version of the phi-coefficient known as Cramér=s V.
- The value of V is evaluated much the same as a correlation.

TABLE 18.10

Standards for interpreting Cramér's V as proposed by Cohen (1988).

	Small Effect	Medium Effect	Large Effect
For $df^* = 1$	0.10	0.30	0.50
For $df^* = 2$	0.07	0.21	0.35
For $df^* = 3$	0.06	0.17	0.29