Hypothesis Testing

Sean Sall

May 18th, 2016

Objectives:

Today's objectives:

- Given a dataset, understand how to set up a hypothesis test
- Know **how** and **when** to use each of the following hypothesis tests, and how to interpret the results:
  - t-test (one/two-sample, paired t-test)
  - z-test (one/two-sample)
  - Chi-squared test
- Know **when** to use each of the following hypothesis tests:
  - F-test
  - Kolmogorov–Smirnov test (K-S test or KS test)
- Define Type I/II error, significance level, and power, and how they relate to each other
- Understand how and when to apply the Bonferonni correction

Agenda

1. General hypothesis testing framework
   - general steps of a hypothesis test
   - general assumptions
   - one-tailed versus two-tailed
2. Specific hypothesis tests (t-test, z-test, etc.)
   - general overview
   - examples
3. Hypothesis testing metrics
   - significance level
   - power
   - Type I/II error
4. Bonferonni correction

Why does this matter?

- Hypothesis testing allows us to quantify the likelihood of obtaining some parameter value(s)
  - Useful in business settings
  - Useful in scientific settings
  - Useful for constructing decision algorithms
  - Useful for determining if two data sets came from the same distribution (KS test), or a single data set follows a hypothesized distribution (KS test, chi-squared)

Hypothesis Testing Framework Part I

Hypothesis Testing Framework

- With **estimation**, the parameter value is unknown, and our goal is to find a point estimate for it, with confidence intervals holding likely values

- With **hypothesis testing**, a hypothesized value of a population parameter is tested, and our goal is to determine how reasonable/unreasonable that hypothesized value is (we might also test the difference, ratio, etc. between two population parameters)

General Steps of Hypothesis Testing

1. State the **null** (H_0) and **alternative** (H_a) hypotheses
   - The **null hypothesis** is typically a measure of the status quo, or that there is no effect
2. Choose the **significance level**, α
   - The **significance level** is the probability of rejecting the null hypothesis when it is in fact true (standard is $\alpha = 0.05$)
   - This is also the threshold that is used to determine whether to reject or fail to reject our null hypothesis given our p-value
3. Choose the **appropriate** statistical test, and calculate a **test-statistic**
4. Compute the **p-value** from the test-statistic calculated in 3, and either **reject** or **fail to reject** the null hypothesis

Central limit theorem

Recall that by the **central limit theorem**, we know that our sample mean is approximately normally distributed:

$$\bar{X} \sim Normal(\mu, \frac{\sigma^2}{n})$$

- From a high level, hypothesis testing is going to take some parameter (estimated by our sample(s)) that we **assume** follows a specified distribution (typically because of CLT), and then we ask how likely it is that we obtained our sample parameter value given the distribution it falls and our null hypothesis

Using sample as estimation of the population

- For our hypothesis tests, we are assuming that we can **approximate** those population parameters using **sample estimates**...
  - this is a huge assumption
  - if it's not met, our hypothesis tests are invalid **or** we have to add a lot of caveats to the interpretation of our results

Two-tailed versus one-tailed tests

- In a **two-tailed** test, we are interested in whether there is a difference in **either** direction (**either** greater than or less than our hypothesized value)
- In a **one-tailed** test, we are only interested in whether there is a difference in **one** direction, but not both

Two-tailed versus one-tailed tests: Visualized

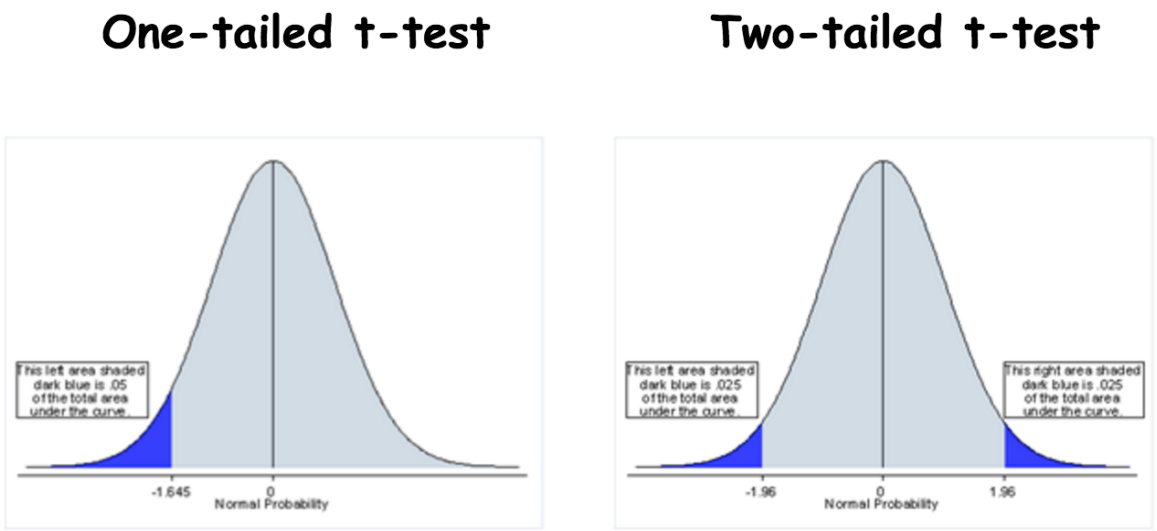

Figure 1:One and Two tails

- We need to be aware of whether we are performing a one-tailed or two-tailed test when calculating our p-values, remembering that two-tail tests look at both ends of the distribution (e.g. there is weight in both tails)

T-tests

When to use a t-test

- The t-test is a test used when we don't know the population variance, but instead have to estimate it (the t-test is built upon the t-distribution)
- Generally, when we are conducting a test where we want to compare the population mean to some hypothetical value, we don't know the population variance, which means we use a **t-test**
- One sample tests are used when we want to compare a single population mean to a hypothesized value, whereas two sample tests are used to compare two population means with each other.

One sample versus two sample t-tests

- The general form of the null and alternative hypothesis for a one sample t-test are:
  - **Null Hypothesis**: $H_0 : \mu = \mu_0$ (our population mean, μ, is equal to some hypothetical value, μ_0)
  - **Alternative Hypothesis**: $H_a : \mu \neq \mu_0$ (our population mean, μ, is not equal to some hypothetical value, μ_0)
- The general form of the null and alternative hypothesis for a two sample t-test are:
  - **Null Hypothesis**: $H_0 : \mu_1 - \mu_2 = d_0$ (the difference between population mean 1 (μ_1) and population mean two (μ_2) is d_0 (often 0))
  - **Alternative Hypothesis**: $H_a : \mu_1 - \mu_2 \neq d_0$ (the difference between population mean 1 (μ_1) and population mean two (μ_2) is not equal to d_0)
- Note: We aren't restricted to using equality in the null (we could use some form of inequality)

One sample t-test: Parts 1-2

Suppose I have a sample of kegs full of beer, and want to conduct a hypothesis test around the claim that on average, the population mean of gallons of beer in a keg is 15.5 gallons (like the breweries say).

1. State the **null** (H_0) and **alternative** (H_a) hypotheses

$$H_0: \mu_g = 15.5$$
$$H_a: \mu_g \neq 15.5$$

- *Note*: The value (e.g. 15.5) matches in the null and alternative hypotheses, and $=$ in the null is the converse of $\neq$ in the alternative. These two stipulations must be met for any null and corresponding alternative hypothesis.

2. Choose the **significance level**, α

Let's go with 0.05, since that's pretty common to use.

One sample t-test: Part 3

3. Choose the **appropriate** statistical test, and calculate a **test-statistic**.
   - We know we want to use a one sample t-test (since we don't know the population variance, and have a single sample), which means we'll need the **sample** mean, standard deviation, and size (e.g. `n`)
     - ★ Let's assume that we have a sample of 120 kegs, and that the **mean** value of gallons of beer among them is 15.4 gallons with a **standard deviation** of 0.2 gallons
   - Our test-statistic is calculated using the following formula:
     - ★ $\frac{\overline{x} - \mu_0}{s/\sqrt{n}}$
   - Which here, translates to:
     - ★ $\frac{15.4 - 15.5}{0.2/\sqrt{120}} =$ -5.48

One sample t-test: Part 4.1

4. Compute the **p-value** from the test-statistic calculated in 3, and either **reject** or **fail to reject** the null hypothesis

To compute the p-value, we use statistical software:

```
import scipy.stats as scs
import numpy as np
scs.t.sf(np.abs(-5.48), 119) * 2
```

This returns a `p-value` of `2.77e-07`.

Note: The multiplication by 2 above is because we are performing a two-tailed test.

One sample t-test: Part 4.2

- Our `p-value` of `2.77e-07` is **less than** our significance level of 0.05, which means that we **reject** the null hypothesis that the mean value of gallons of beer among kegs is equal to 15.5 gallons
  - *Interpretation*: Under the null hypothesis, there is a probability of `2.77e-07` of observing a result *as extreme* as we observed
- We can also do this by building confidence intervals around our sample mean and determining if the null value (15.5) lies within the confidence level given our significance level
  - Recall the CI formula: $(\bar{x} - t^{\nu}_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}, \bar{x} + t^{\nu}_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}})$ (ν is the *degrees of freedom*)
  - For us here, this results in: (15.364, 15.436)
  - Since 15.5 (our null value) doesn't lie in our 95% CI, we **reject** the null hypothesis
  - *Interpretation*: With confidence level 95%, μ lies in the interval (15.364, 15.436)

One sample t-test: Using scipy.stats

If we had an array with the number of gallons of each of our kegs in an array, we could simply calculate both the test-statistic and p-value using a `scipy.stats` 1 sample test of the mean.

```
import scipy.stats as scs
scs.ttest_1samp(kegs_array, 15.5)
```

This returns a `t-stat` of `-5.45` and a `p-value` of `2.71e-07`.

Note: `scs.ttest_1samp` performs a two-tailed test by default.

Two sample t-test: Parts 1-2

Suppose that I want to compare the average speed of tennis players and soccer players (during their respective matches), and I think that on average, soccer players have a higher speed (μ_s) than tennis players (μ_t) .

1. State the **null** (H_0) and **alternative** (H_a) hypotheses

$$H_0\text{: } \mu_s - \mu_t \geq 0$$
$$H_a\text{: } \mu_s - \mu_t < 0$$

2. Choose the **signifiance level**, α

Let's go with 0.05, since that's pretty common to use.

Two sample t-test: Part 3.1

3. Choose the **appropriate** statistical test, and calculate a **test-statistic**.
   - We know we want to use a two sample t-test (since we don't know the population variances, and have two samples), which means we'll need the **sample** means, standard deviations, and sizes
   - The actual test-statistic formula in a two sample t-test depends on the different samples sizes and variances, as well as if they are independent or not. There are variety of scenarios:
     - ★ Equal sample sizes, equal variance
     - ★ Equal or unequal sample sizes, equal variance
     - ★ Equal or unequal sample sizes, unequal variance

Two sample t-test: Part 3.2

- **Welchs t-test** is used when samples are of equal or unequal sample sizes, with unequal variance (this is sometimes referred to as an independent samples test using unpooled variance).
- The general formula for the test-statistic is: $t = \frac{\overline{x1} - \overline{x2} - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- For the degrees of freedom the formula is: $df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{\frac{s_1^2}{n_1}^2}{n_1 - 1} + \frac{\frac{s_2^2}{n_2}}{n_2 - 1}}$

Two sample t-test: Part 3.3

- Assuming our sample of soccer players' speeds has size 120 with an average speed of 6.5 mph and standard deviation of 0.20 mph and our sample of tennis players' speeds has size 115 with a mean of 6.48 mph and standard deviation of 0.15 mph, we calculate our t-statistic as:

$$\frac{6.5 - 6.48 - 0}{\sqrt{\frac{0.20^2}{120} + \frac{0.15^2}{115}}} = 0.870$$

Two sample t-test: Part 4.1

4. Compute the **p-value** from the test-statistic calculated in 3, and either **reject** or **fail to reject** the null hypothesis

To compute the p-value here, we use statistical software:

```
import scipy.stats as scs
import numpy as np
scs.t.sf(np.abs(0.870), 220)
```

This returns a `p-value` of `0.193`.

Note: We do not multiply by 2 because we are conducting a one-tailed test.

Two sample t-test: Part 4.2

- Our `p-value` of `0.193` is **greater than** our significance level of 0.05, which means that we **fail to reject** the null hypothesis that the difference between the average speed of soccer players and tennis players is greater than 0
  - *Interpretation*: Under the null hypothesis, there is a probability of `0.193` of observing a result *as extreme* as we observed

Two sample t-test: Using scipy.stats

If we had an array with the speeds of our tennis and soccer players, we could simply calculate both the test-statistic and p-value using a `scipy.stats` 2 sample test of the mean.

```
import scipy.stats as scs
scs.ttest_ind(soccer_speeds, tennis_speeds, equal_var=False)
```

This returns a `t-stat` of `0.866` and a `p-value` of `0.387`.

Note: `scs.ttest_ind` performs a two-tailed test by default, so to get the p-value for a one-tailed test we have to divide the above 0.387 by 2 (giving 0.194).
Note: The `scs.ttest_ind` stands for *independent t-test*, to be used when your two samples are **independent**.

Two sample t-test: Paired t-tests

- **Paired t-tests** are used when we want to compare the sample means of two populations that are correlated (e.g. the samples are not independent)
  - Most often, these take the form of a before and after/control and treatment kind of study

Z-tests

When to use a z-test

- The z-test is a test used when we know the population variance (the z is built on the normal distribution)
- Generally, when we are conducting a test where we want to compare a population proportion to some hypothetical value, we use a z-test.
  - In this case, we are explicitly assuming that the population proportion is equal to some value p_0, and it naturally follows that if that is true, the variance is equal to $p_0(1 - p_0)$ ($\bar{x} \sim Bernoulli(p_0, \frac{p_0(1-p_0)}{n})$)

One sample versus two sample z-tests

- The general form of the null and alternative hypothesis for a one sample z-test are:
  - **Null Hypothesis**: $H_0 : p = p_0$ (our population proportion, p, is equal to some hypothetical value, p_0)
  - **Alternative Hypothesis**: $H_a : p \neq p_0$ (our population proportion, p, is not equal to some hypothetical value, p_0)
- The general form of the null and alternative hypothesis for a two sample z-test are:
  - **Null Hypothesis**: $H_0 : p_1 - p_2 = d_0$ (the difference between population proportion 1 (p_1) and population proportion 2 (p_2) is equal to d_0 (often 0))
  - **Alternative Hypothesis**: $H_0 : p_1 - p_2 \neq d_0$ (the difference between population proportion 1 (p_1) and population proportion 2 (p_2) is not equal to d_0)
- Note: We aren't restricted to using equality in the null (we could use some form of inequality)

One sample z-test: Parts 1-2

Suppose I want to conduct a hypothesis test for whether or not the proportion of people in Austin who like the Denver Broncos (p_b) is less than or equal to 50%. Let's assume that we have a sample of 135 people, and that the number of people who are broncos fans is 65.

1. State the **null** (H_0) and **alternative** (H_a) hypotheses

$$H_0: \ p_b \leq 0.50$$
$$H_a: \ p_b > 0.50$$

2. Choose the **signifiance level**, α

Let's go with 0.05, since that's pretty common to use.

One sample z-test: Part 3

3. Choose the **appropriate** statistical test, and calculate a **test-statistic**.
   - We know we want to use a one sample z-test (since we have a sample that contains observations drawn from a Bernoulli distribution), which means that we'll need the sample proportion and the number of observations
   - Our test-statistic is calculated using the following formula:
     - ★ $$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$
   - Which here, translates to:
     - ★ $$\frac{0.4814 - 0.5}{\sqrt{\frac{0.50(1-0.50)}{135}}} = \text{-0.430}$$

One sample z-test: Part 4.1

4. Compute the **p-value** from the test-statistic calculated in 3, and either **reject** or **fail to reject** the null hypothesis

To compute the p-value here, we use statistical software:

```
import scipy.stats as scs
import numpy as np
scs.norm.sf(np.abs(-0.430))
```

This returns a `p-value` of `0.33`.

Note: We don't multiply by 2 here because we are performing a one-tailed test.

One sample z-test: Part 4.2

- Our `p-value` of `0.33` is **greater than** our significance level of 0.05, which means that we **fail to reject** the null hypothesis that the proportion of Broncos fans is greater than or equal to 0.50
  - *Interpretation*: Under the null hypothesis, there is a probability of `0.33` of observing a result *as extreme* as we observed
- We can also do this by building confidence intervals around our sample proportion and determining if the null value (0.50) lies within the confidence level given our significance level
  - Recall the CI formula: $(\hat{p} - z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$
  - For us here, this results in: (0.398, 0.564)
  - Since 0.5 (our null value) lies in our 95% CI, we **fail to reject** the null hypothesis
  - *Interpretation*: With confidence level 95%, p lies in the interval (0.398, 0.564)

Two sample z-test: Parts 1-2

Suppose I now want to determine if the proportion of Broncos fans in Denver (p_d) is different from the proportion of Bronco fans in Austin (p_a). Let's say I have a sample from Austin of 95 people, 35 of which are Broncos fans. I also have a sample from Denver of 110 people, 65 of which are Broncos fans.

1. State the **null** (H_0) and **alternative** (H_a) hypotheses

Say that I think the proportion of fans is exactly equal.

$$H_0\text{: } p_d - p_a = 0$$
$$H_a\text{: } p_d - p_a \neq 0$$

2. Choose the **significance level**, α

Let's go with 0.05, since that's pretty common to use.

Two sample z-test: Part 3

3. Choose the **appropriate** statistical test, and calculate a **test-statistic**.
   - We know we want to use a two sample z-test (since we have samples that contain observations drawn from Bernoulli distributions), which means that we'll need the sample proportion and the number of observations for each sample
   - Our test-statistic is calculated using the following formula:
     - ★ $$\frac{\hat{p}_1 - \hat{p}_2 - d_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$
   - Which here, translates to:
     - ★ $$\frac{0.59 - 0.368 - 0}{\sqrt{\frac{0.59(1-0.59)}{110} + \frac{0.368(1-0.368)}{95}}} = 3.26$$

Two sample z-test: Part 4.1

4. Compute the **p-value** from the test-statistic calculated in 3, and either **reject** or **fail to reject** the null hypothesis

To compute the p-value here, we use statistical software:

```
import scipy.stats as scs
import numpy as np
scs.norm.sf(np.abs(3.26)) * 2
```

This returns a `p-value` of `0.0011`.

Note: We multiply by 2 here because we are performing a two-tailed test.

Two sample z-test: Part 4.2

- Our `p-value` of `0.0011` is **less than** our significance level of 0.05, which means that we **reject** the null hypothesis that the difference between the proportion of Broncos fans in Austin and Denver is 0.
  - *Interpretation*: Under the null hypothesis, there is a probability of `0.0011` of observing a result *as extreme* as we observed

Chi-Squared Tests

Chi-Squared Test

- A **chi-squared** test is used to:
  1. Test if a population variance is equal to a hypothesized value
  2. To assess goodness-of-fit
     - ★ For goodness-of-fit, we often want to know if some empirical data follows a given distribution (an expected distribution, uniform distribution, etc.)
  3. To determine whether or not categorical variables are independent.

Chi-Squared Test: Population Variance Test

- In the case of a population variance test, our null and alternative hypotheses are something like the following:
  - **Null hypothesis**: $H_0 : \sigma^2 = \sigma_0^2$ (our population variance, σ^2, is equal to a hypothesized value, σ_0^2)
  - **Alternative hypothesis**: $H_a : \sigma^2 \neq \sigma_0^2$ (our population variance, σ^2, is not equal to a hypothesized value, σ_0^2)
- We calculate the **test-statistic** using the following formula, and then compare it to the chi-squared distribution with `n-1` degrees of freedom:

$$\chi^2 = (n-1)\frac{s^2}{\sigma_0^2}$$

Chi-Squared Test: Goodness of Fit or Independence of Categorical Variables

- In the case of goodness-of-fit, our null and alternative hypotheses are something like the following:
  - **Null hypothesis**: H_0: The data are consistent with a specified discrete distribution.
  - **Alternative hypothesis**: H_a: The data are not consistent with a specified discrete distribution.
- The testing for **independence of categorical variables** can be thought of as a special case where we are testing for a goodness-of-fit to the uniform distribution.
- We calculate the **test-statistic** using the following formula, and then compare it to the chi-squared distribution with our calculated degrees of freedom:

$$\chi^2 = \sum_{Cells} \frac{(Observed - Expected)^2}{Expected}$$

Chi-Squared Test: Goodness of Fit Example Parts 1-2

Continuum claims that 90% of their employees are software engineers, 7% are data scientists, and 3% are other (marketing, management, etc.), and we want to conduct a hypothesis test to determine if that's true. Suppose we get a sample of 100 Continuum employees, of which 6 are other, 12 are data scientists, and 82 are software engineering.

1. State the **null** (H_0) and **alternative** (H_a) hypotheses

H_0: The data are consistent with what Continuum says.
H_a: The data are not consistent with what Continuum says.

2. Choose the **signifiance level**, α

Let's go with 0.05, since that's pretty common to use.

Chi-Squared Test: Goodness of Fit Example Part 3

3. Choose the **appropriate** statistical test, and calculate a **test-statistic**.

- Since we are comparing a goodness-of-fit, we want to use a chi-squared test
- There are three total cells (e.g. software engineer, data scientist, and other), and we need to calculate the expected number for each one
  - This is calculated given the percentages that Continuum claims (90% software engineers, 7% data scientists, 3% other)
  - Given these percentages and a sample of 100, we would **expect** that 90 are software engineers, 7 are data scientists, and 3 are other
- We can calculate the **test-statistic** as :

$$\frac{(82-90)^2}{90} + \frac{(12-7)^2}{7} + \frac{(6-3)^2}{3} = 7.28$$

Chi-Squared Test: Goodness of Fit Example Part 4.1

4. Compute the **p-value** from the test-statistic calculated in 3, and either **reject** or **fail to reject** the null hypothesis

   - To do this, we need to compare our chi-squared test-statistic with our degrees of freedom (here `k-1`, where `k` is the # of levels of our categorical variable)
   - We can do this using `scipy.stats.chi2`...

```
import scipy.stats as scs
scs.chi2.sf(7.28, 2)
```

This returns a p-value of `0.026`.

Note: We did not multiply by two even though it was is a two-sided test because the chi-squared is strictly positive.

Chi-Squared Test: Goodness of Fit Example Part 4.2

- Our `p-value` of `0.026` is **less than** our significance level of 0.05, which means that we **reject** the null hypothesis that the data are consistent with what Continuum says
  - *Interpretation*: Under the null hypothesis, there is a probability of `0.026` of observing a result *as extreme* as we observed

Chi-Squared Test: Goodness of Fit with scipy.stats

- We could have performed the entire test by passing in the *observed* and *expected* numbers into `scipy.stats.chisquare`...

```
import scipy.stats as scs
scs.chisquare([82, 12, 6], [90, 7, 3]) # (Observed, Expected)
```

This returns a *chi-squared test-statistic* of `7.28`, and a `p-value` of `0.026`.

Chi-Squared Test: Independence of Categorical Variables Example Parts 1-2

Say I want to conduct a hypothesis test that whether or not you are a dog or cat "person" is independent of whether or not you are male or female. To do so, I gather a sample of 99 males and 103 females, where 43 of the males are dog people and 39 of the females are dog people.

1. State the **null** (H_0) and **alternative** (H_a) hypotheses

H_0: Being a dog/cat person is independent of being a male/female.
H_a: Being a dog/cat person is **not** independent of being a male/female.

2. Choose the **significance level**, α

Let's go with 0.05, since that's pretty common to use.

Chi-Squared Test: Goodness of Fit Example Part 3.1

3. Choose the **appropriate** statistical test, and calculate a **test-statistic**.

- Since we are determining whether or not two categorical variables are independent of each other, we want to use a chi-squared test
- There are four total cells (e.g. male dog person, female dog person, male cat person, female cat person), and we need to calculate the expected number for each one
  - To do this, we take the $\frac{(\text{row total})(\text{column total})}{\text{total number of obs.}}$ for each cell

Chi-Squared Test: Goodness of Fit Example Part 3.2

- Observed Data

	Male	Female	Row Total
Dog	43	39	82
Cat	56	64	120
Column Total	99	103	202

- Expected Data ($\frac{\text{(row total)(column total)}}{\text{total number of obs.}}$ for each cell)

	Male	Female	Row Total
Dog	40.19	41.81	82
Cat	58.81	61.19	120
Column Total	99	103	202

Chi-Squared Test: Goodness of Fit Example Part 3.3

To determine our **chi-squared test-statistic**, we use the formula:

$$\chi^2 = \sum_{Cells} \frac{(Observed - Expected)^2}{Expected}$$

For us, this is:

$$\frac{(43-40.18)^2}{40.18} + \frac{(56-58.81)^2}{58.81} + \frac{(39-41.81)^2}{41.81} + \frac{(64-61.19)^2}{61.19} = 0.650$$

Chi-Squared Test: Goodness of Fit Example Part 4.1

4. Compute the **p-value** from the test-statistic calculated in 3, and either **reject** or **fail to reject** the null hypothesis

- To do this, we need to compare our chi-squared test-statistic to the chi-squared distribution with the appropriate degrees of freedom (here `(r-1) * (c-1)`, where `r` is the # of rows in our table and `c` is the # of columns)
- We can do this using `scipy.stats.chi2`...

```
import scipy.stats as scs
scs.chi2.sf(0.650, 1)
```

This returns a p-value of `0.420`.

Note: We did not multiply by two even though it was is a two-sided test because the chi-squared is strictly positive.

Chi-Squared Test: Goodness of Fit Example Part 4.2

- Our `p-value` of `0.420` is **greater than** our significance level of 0.05, which means that we **fail to reject** the null hypothesis that being a dog/cat person is independent of being male/female
  - *Interpretation*: Under the null hypothesis, there is a probability of `0.420` of observing a result *as extreme* as we observed

Note: We could have also used a two-sample proportion test for this, with the null hypothesis that the proportion of dog people are equal among males and females.

Chi-Squared Test: Goodness of Fit with scipy.stats

- We could have performed the entire test by passing in the *observed* numbers into `scipy.stats.contingency`...

```
import scipy.stats as scs
scs.chi2_contingency([[43, 56], [39, 64]])
```

This returns a *chi-squared test-statistic* of `0.650`, and a `p-value` of `0.420`.

Other tests

F-Test for Ratio of Variance

- The **f-test** can be used to conduct a test for the ratio of two population variances, or in a regression setting when looking at an ANOVA table (we'll talk about it next week during regression).

F-test for Ratio of Variance

- For the **population variance case**, or our general null and alternative hypotheses take the form:
  - **Null hypothesis**: $H_0 : \frac{s_1^2}{s_2^2} = r_0$
    - ★ the ratio of population variance 1 (s_1^2) to population variance two (s_2^2) is equal to some hypothesized value, r_0 (often 1)
  - **Alternative hypothesis**: $H_a : \frac{s_1^2}{s_2^2} \neq 1$
    - ★ the ratio of population variance 1 (s_1^2) to population variance two (s_2^2) is not equal to some hypothesized value, r_0
- We'll calculate the **f-statistic** using the formula below, and then compare it to the F-distribution with our given degrees of freedom $(n_1 - 1, n_2 - 1)$:

$$F = \frac{s_1^2}{s_2^2}$$

Kolmogorov-Smirnov Test

- A **Kolmogorov-Smirnov** test (KS-test) can be used to compare goodness-of-fit for a single sample to a **continuos** distribution, or two samples to each other

Cheat-Sheets

Cheat-Sheet Part 1

Hypothesis Test	Test-Statistic
1 samp. t	$\frac{\overline{x}-\mu_0}{s/\sqrt{n}}$
Welch's t	$\frac{\overline{x1}-\overline{x2}-d_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
1 samp. z	$\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
2 samp. z	$\frac{\hat{p_1}-\hat{p_2}-d_0}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$

Cheat-Sheet Part 2

Hypothesis Test	Test-Statistic
Chi-Squared Population Variance	$(n-1)\frac{s^2}{\sigma_0^2}$ (n-1 degrees of freedom)
Chi-Squared Goodness of Fit	$\sum_{Cells} \frac{(Observed-Expected)^2}{Expected}$ (k-1 degrees of freedom)
Chi-Squared Independence of Categoricals	$\sum_{Cells} \frac{(Observed-Expected)^2}{Expected}$ ((n-1)(c-1) degrees of freedom)
F-test Ratio of Variances	$F = \frac{s_1^2}{s_2^2}$ ((n_1 - 1, n_2 - 1) degrees of freedom)

Cheat-Sheet Part 3

The wiki for Statistical_hypothesis_testing#Common_test_statistics is also pretty nice.

Hypothesis Testing Framework Part II

Hypothesis Testing Metrics - Definitions

- **significance level** - the probability of rejecting the null hypothesis when it is in fact true (α)
- **Type I error** - rejecting the null hypothesis when it is in fact true (happens with probability α)

- **power** - the probability of rejecting the null hypothesis when it is in fact false ($1 - \beta$)
- **Type II error** - failing to reject the null hypothesis when it is in fact false (happens with probability β)

Hypothesis Testing Metrics - Visualized

Decision Made	True State of Nature: H_0 is true	True State of Nature: H_a is true
Accept H_0	**Correct decision** Probability $= 1 - \alpha$	**Type II error** Probability $= \beta$
Reject H_0	**Type I error** Probability $= \alpha$ (significance level)	**Correct decision** Probability $= 1 - \beta$ (power)

Figure 2:Hypothesis Testing!

Multiple Comparisons Problem - Motivation

- Say we want to run a hypothesis test with an α of 0.05
  - The 1st time we run a test, there is a 95% chance of not getting a false positive (e.g. 5% chance of getting type I error)
  - The 2nd time we run a test, there is a 95% chance of not getting a false positive
    - ★ The probability of not getting a false positive over both tests is $0.95^2 = 0.9025$
  - ...
  - ...
  - ...
  - The probability of not getting a false positive over n tests is 0.95^n
    - ★ If we run 100 tests, we can expect 5 to show statistically significant results even if the null hypothesis that there is no effect is true

This is known as the **multiple comparisons** problem

Multiple Comparisons Problem - A Fix

- While there are many ways to account for this, the **Bonferonni correction** is a common one
- To use it, we simply take α/m, where $\mathtt{m}$ corresponds to the number of hypothesis tests we will be conducting (or have already conducted), and use this as our new α