Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Hypothesis Testing

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

OCTOBER 11, 2016

χ^2 Example

Random sample of 500 U.S. adults: political affiliation and opinion on a tax reform. Dependent at a 5% level of significance?

Observed

	Favor	Indifferent	Oppose
Dem	138	83	64
Rep	64	67	84

Expected

	Favor	Indifferent	Oppose
Dem			
Rep			

χ^2 Example

Random sample of 500 U.S. adults: political affiliation and opinion on a tax reform. Dependent at a 5% level of significance?

Observed

	Favor	Indifferent	Oppose
Dem	138	83	64
Rep	64	67	84

Expected

	Favor	Indifferent	Oppose
Dem	115.14		
Rep			

χ^2 Example

Random sample of 500 U.S. adults: political affiliation and opinion on a tax reform. Dependent at a 5% level of significance?

Observed

	Favor	Indifferent	Oppose
Dem	138	83	64
Rep	64	67	84

Expected

	Favor	Indifferent	Oppose
Dem	115.14	85.50	
Rep			

χ^2 Example

Random sample of 500 U.S. adults: political affiliation and opinion on a tax reform. Dependent at a 5% level of significance?

Observed

	Favor	Indifferent	Oppose
Dem	138	83	64
Rep	64	67	84

Expected

	Favor	Indifferent	Oppose
Dem	115.14	85.50	84.36
Rep			

χ^2 Example

Random sample of 500 U.S. adults: political affiliation and opinion on a tax reform. Dependent at a 5% level of significance?

Observed

	Favor	Indifferent	Oppose
Dem	138	83	64
Rep	64	67	84

Expected

	Favor	Indifferent	Oppose
Dem	115.14	85.50	84.36
Rep	86.86		

χ^2 Example

Random sample of 500 U.S. adults: political affiliation and opinion on a tax reform. Dependent at a 5% level of significance?

Observed

	Favor	Indifferent	Oppose
Dem	138	83	64
Rep	64	67	84

Expected

	Favor	Indifferent	Oppose
Dem	115.14	85.50	84.36
Rep	86.86	64.50	

χ^2 Example

Random sample of 500 U.S. adults: political affiliation and opinion on a tax reform. Dependent at a 5% level of significance?

Observed

	Favor	Indifferent	Oppose
Dem	138	83	64
Rep	64	67	84

Expected

	Favor	Indifferent	Oppose
Dem	115.14	85.50	84.36
Rep	86.86	64.50	63.64

χ^2 Example

Random sample of 500 U.S. adults: political affiliation and opinion on a tax reform. Dependent at a 5% level of significance?

Observed

	Favor	Indifferent	Oppose
Dem	138	83	64
Rep	64	67	84

Expected

	Favor	Indifferent	Oppose
Dem	115.14	85.50	84.36
Rep	86.86	64.50	63.64

$$4.539 + 0.073 + 4.914 + 6.016 + 0.097 + 6.514 = 22.152 \tag{1}$$

Running test: df, *p*-Value

- Degrees of Freedom?

Running test: df, *p*-Value

- Degrees of Freedom? $(r-1)(c-1) = 1 \cdot 2 = 2$
- *p*-value

Running test: df, *p*-Value

- Degrees of Freedom? $(r-1)(c-1) = 1 \cdot 2 = 2$
- *p*-value

```
>>> from scipy.stats.distributions import chi2
>>> 1 - chi2.cdf(22.15, 2)
1.5494894118783797e-05
>>> from scipy.stats import chisquare
>>> chisquare([138, 83, 64, 64, 67, 84],
...           [115.14, 85.5, 84.36, 86.86, 64.5, 63.64],
...           3)
Power_divergenceResult(statistic=22.152468645918482,
                                 pvalue=1.5475780213
```

Cow tails

A herd of 1,500 steer was fed a special highâĂŘprotein grain for a month. A random sample of 29 were weighed and had gained an average of 6.7 pounds. If the standard deviation of weight gain for the entire herd is 7.1, test the hypothesis that the average weight gain per steer for the month was more than 5 pounds.

We Need: What test? What distribution? What's the null?

Setup

- Test?

Setup

- Test? z-test
- Distribution?

Setup

- Test? z-test
- Distribution? Normal with mean 5, s.d. 7.1
- Null?

Setup

- Test? z-test
- Distribution? Normal with mean 5, s.d. 7.1
- Null? $H_0 : \mu_0 = 5$
- α?

Setup

- Test? z-test
- Distribution? Normal with mean 5, s.d. 7.1
- Null? $H_0 : \mu_0 = 5$
- α? Let's say 0.05

Cow tails

A herd of 1,500 steer was fed a special highâĂŘprotein grain for a month. A random sample of 29 were weighed and had gained an average of 6.7 pounds. If the standard deviation of weight gain for the entire herd is 7.1, test the hypothesis that the average weight gain per steer for the month was more than 5 pounds.

Test Statistic:

Cow tails

A herd of 1,500 steer was fed a special highâĂŘprotein grain for a month. A random sample of 29 were weighed and had gained an average of 6.7 pounds. If the standard deviation of weight gain for the entire herd is 7.1, test the hypothesis that the average weight gain per steer for the month was more than 5 pounds.

Test Statistic: $Z = \frac{6.7-5}{\frac{7.1}{\sqrt{29}}} = \frac{1.7}{1.318} = 1.289$

p-value

```
>>> from scipy.stats import norm
>>> 1.0 - norm.cdf(1.28)
0.10027256795444206
```

US vs. Japanese Mileage

Read in Data

```
>>> import pandas as pd
>>> mpg = pd.read_csv("jp-us-mpg.dat", delim_whitespace=Tru
>>> mpg.head()
   US  Japan
0  18   24.0
1  15   27.0
2  18   27.0
3  16   25.0
4  17   31.0
```

Is the average car in the US as efficient as the average car in Japan?

Two-Tailed Two-Sample t-test

- Compute means

Two-Tailed Two-Sample t-test

- Compute means

```
>>> from numpy import mean
>>> mean(mpg["Japan"].dropna())
30.481012658227847
>>> mean(mpg["US"].dropna())
20.144578313253011
```

- Compute sample variances

Two-Tailed Two-Sample *t*-test

- Compute means

```
>>> from numpy import mean
>>> mean(mpg["Japan"].dropna())
30.481012658227847
>>> mean(mpg["US"].dropna())
20.14457831325301
```

- Compute sample variances

```
>>> from numpy import var
>>> us = mpg["US"].dropna()
>>> jp = mpg["Japan"].dropna()
>>> jp_var = var(jp) * len(jp) / float(len(jp) - 1)
>>> us_var = var(us) * len(us) / float(len(us) - 1)
```

Degrees of Freedom

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}} \tag{2}$$

Degrees of Freedom

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}} \tag{2}$$

$\nu = 136.8750$

t-Statistic

$$T = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{3}$$

t-Statistic

$$T = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{3}$$

$T = 12.94$

p-value

p-value

```
>>> 2*(1.0 - t.cdf(abs(12.946), 136.8750))
0.0
```