single **categorical** variable: **1-sample chi-square test** (aka **goodness-of-fit test**)

OVERVIEW

TECHNIQUES USED

- method: 1 variable: *categorical - 1 sample chi square - goodness-of-fit*
- tasks:
  - Descriptive statistics:
    - Describe frequency/proportions of eye colour in the dataset.
  - Inferential statistics:
    - Is eye colour evenly distributed amongst the subjects in the dataset?
    - Is eye colour in sample data distributed the same as in the general population?
- datasets: datasets::HairEyeColor
- functions: str() , **margin.table()** , **chisq.test()** , browseURL() , rm() , ls()

Methodology Summary: basis of 1-sample (goodness-of-fit chi square test)

- quick test
- can specify
  - equal proportions by default
  - own particular proportions
- → gives you flexibility in the test
- **X-Square** - Greek Chi - takes place of Greek C
  - lower case chi, looks like X
  - upper care chi, looks a lot like X

SCRIPTS SUMMARY

LOAD DATA

```
?HairEyeColor
str(HairEyeColor)
HairEyeColor
```

PREPARE DATA

```
margin.table(HairEyeColor, 2)
eyes <- margin.table(HairEyeColor, 2)
eyes
round(prop.table(eyes), 2)
```

PEARSON'S CHI-SQUARED TEST

```
chi1 <- chisq.test(eyes)
chi1
browseURL("http://www.statisticbrain.com/eye-color-distribution-percentages/")
chi2 <- chisq.test(eyes, p = c(.41, .32, .15, .12))
chi2
```

CLEAN UP

```
rm(list = ls())
```

SCRIPTS & NOTES

IF categorical variable - where people fall into different groups
THEN test with 1-sample chi-test (aka goodness of fit test)

LOAD DATA

```
?HairEyeColor
```

- 592 statistics students who recorded their `hair colour` | `eye colour` | `sex`

```
str(HairEyeColor)
```

```
 'table' num [1:4, 1:4, 1:2] 32 53 10 3 11 50 10 30 10 25 ...
 - attr(*, "dimnames")=List of 3
  ..$ Hair: chr [1:4] "Black" "Brown" "Red" "Blond"
  ..$ Eye : chr [1:4] "Brown" "Blue" "Hazel" "Green"
  ..$ Sex : chr [1:2] "Male" "Female"
```

- structure is **table**
- several values: ex. `eye colour`: brown, blue, hazel, green

```
HairEyeColor
```

- see entire table distribution

```
, , Sex = Male

       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8

, , Sex = Female

       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8
```

PREPARE DATA

get marginal frequencies for eye color

- **margin.table** function (**dataset**=HairEyeColor, **2nd variable** listed in structure)
- eye color is 2nd variable in table

```
margin.table(HairEyeColor, 2)
```

```
Eye
Brown  Blue Hazel Green
  220   215    93    64
```

save eye color to data frame - by feeding into object called 'eyes'

```
eyes <- margin.table(HairEyeColor, 2)
```

- → workspace: values eyes table[4]

```
eyes
```

Values	
eyes	'table' num [1:4(1d)] 220 215 93 64

```
Eye
Brown  Blue Hazel Green
  220   215    93    64
```

- eyes - same as result above - except now we're calling it from one object

compare proportions

```
round(prop.table(eyes), 2)        # show as proportions w/2 digits
```

- **prop.table** function (**data frame object**) → wrapped with **round** function

```
Eye
Brown  Blue Hazel Green
 0.37  0.36  0.16  0.11
```

Conclusion: In our sample of statistics students, 30% have brown eyes, 36% blue, 16% hazel, 11% green

PEARSON'S CHI-SQUARED TEST

Pearson's chi-squared test to analyse this data

- need one-dimensional goodness-of-fit test

default test

- default test (assume equal distribution) - whether eye colors evenly distributed across the 4 categories
- in this case nobody expects to have as many people with green eyes as we have with brown or blue eyes (but in many situations - equal distribution is appropriate hypothesis)

```
chi1 <- chisq.test(eyes)
```

- **chisq.test** function
- feed test into "`chi1`" object
- → default results: workspace: values chi1 htest [9] (list of 9)

```
chi1   # check results - look at what's inside chi1
```

```
> chi1

        Chi-squared test for given probabilities

data:  eyes
X-squared = 133.47, df = 3, p-value < 2.2e-16
```

- we have a **chi-square test** for given probabilities
- **data** is `eyes`
- **chi-square test** = 133
- **df** 3 (bc 4 categories)
- **p-value** extremely small → tf scientific notation bc many zeros

Conclusion: This test tells us that our sample significantly deviates from the null, normal distribution of equal number of people in each of the 4 categories. We conclude that there are not equal numbers of each eye colour amongst students in this dataset.
→ not surprising: dn expect that in the 1st place

Compare to population distribution

- What we should be comparing these people against: population proportions

Population data from:

```
browseURL("http://www.statisticbrain.com/eye-color-distribution-percentages/")
```

- → found population statistics about approximate eye colors in the population
- approximate proportions:
  - brown: .41 (Combining Brown Irises with Specks & Dark Brown Irises)
  - blue: .32 (Blue / Grey Irises)
  - hazel: .15 (Blue / Grey / Green Irises with Brown / Yellow Specks)
  - green .12 (Green / Light Brown Irises with Minimal Specks)
- ⇒ `p = c(.41, .32, .15, .12)`
  - the population proportions - in same order that appear in sample data
  - → tf will compare our sample - against population of 41, 32, 15, & 12% respectively for each of the categories

```
chi2 <- chisq.test(eyes, p = c(.41, .32, .15, .12))
```

- 1st parts same: calculate chi square test | tell it what dataset
- → now providing explicit proportion values to compare example against
  - instead of assuming 25, 25, 25, 25 → values need to add up to 100
- → saved in new object: `chi2`
- → workspace: values chi2 htest[9] - or list of 9

```
chi2   #look at what's inside chi2
```

```
        Chi-squared test for given probabilities

data:  eyes
X-squared = 6.4717, df = 3, p-value = 0.09079
```

- **chi square test for given probabilities** (same)
- **data** - eyes (same)
- **x-square value**: `6.47` (was 133)
- **df** `3` (same)
- **p-value** (probability value | probability of getting observed findings in null hypothesis true) is `0.09`
- **standard cutoff** in social sciences is **0.05**

⇒ **conclude**: this group of statistics students do not differ significantly from the general population in terms of eye colour.

CLEAN UP

```
rm(list = ls())
```
