Outline

Comparing Two Proportions

Association genotype - phenotype: cross 2 inbred lines of mice, one lean, one naturally obese. Backcross with the lean parent: F2 mice.

genotype at a given locus (one among thousands): LO or LL.

L=A,C,G or T, whatever the lean inbred parent line has.

O=A,C,G or T, whatever the obese inbred parent line has.

phenotype: either lean or obese.

$p_{LL} = p_1$: probability of obese phenotype among F2 backcrosses with genotype LL at the locus,

$p_{LO} = p_2$: probability of obese phenotype among F2 backcrosses with genotype LO at the locus.

Confidence interval for $p_1 - p_2$ (with 2×2 tables)

Data: $n_{LO} = 105$, $Y_{LO} = 71$ mice with genotype LO are obese,
$n_{LL} = 87$, $Y_{LL} = 45$ with genotype LL are obese.

success = obese = ☺

	LO	LL	total
☺	71	45	116
☹	34	42	76
total	105	87	192

	treatment 1	treatment 2
☺	y_1	y_2
☹	$n_1 - y_1$	$n_2 - y_2$
total	n_1	n_2

$p_1 = \mathbb{P}\{☺ |$ treatment 1 :LO$\}$ and $p_2 = \mathbb{P}\{☺ |$ tmt 2: LL$\}$

Next: confidence interval for $p_1 - p_2$,
chi-square test for H_0: $p_1 = p_2$.

textbook: CI for odds ratio. We cover CI for proportions instead.

Confidence interval for $p_1 - p_2$

Same trick we saw before:
add 4 fictitious individuals (mice),
one in each cell (no favorite cell!)

	drug 1	drug 2
☺	+1	+1
☹	+1	+1

	treatment 1	treatment 2
☺	$y_1 + 1$	$y_2 + 1$
☹	$n_1 - y_1 + 1$	$n_2 - y_2 + 1$
total	$n_1 + 2$	$n_2 + 2$

Estimates of p_1 and p_2 are $\tilde{p}_1 = \dfrac{y_1 + 1}{n_1 + 2}$ $\tilde{p}_2 = \dfrac{y_2 + 1}{n_2 + 2}$

Estimate of $p_1 - p_2$ is $\tilde{p}_1 - \tilde{p}_2$

Standard error of this estimate:

$$\mathrm{SE}_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{SE^2_{\tilde{p}_1} + SE^2_{\tilde{p}_2}} = \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2 + 2}}$$

Confidence interval for $p_1 - p_2$

95% confidence interval for $p_1 - p_2$ is

$$\tilde{p}_1 - \tilde{p}_2 \pm 1.96 * \text{ SE}_{\tilde{p}_1 - \tilde{p}_2}$$

Other confidence levels: use a z-multiplier from the Z-table.
90% confidence: $z_{.05} = 1.645$ and the interval becomes

$$\tilde{p}_1 - \tilde{p}_2 \pm 1.645 * \text{ SE}_{\tilde{p}_1 - \tilde{p}_2}$$

Mice: $\tilde{p}_1 = \dfrac{71+1}{105+2} = .67$, $\tilde{p}_2 = \dfrac{45+1}{87+2} = .52$, $\text{SE}_{\tilde{p}_1 - \tilde{p}_2} = 0.07$.

95% confidence interval: $.156 \pm 1.96 * .07$ i.e $(.019, .293)$.
90% confidence interval: $(.041, .271)$.

Conclusion

We are 95% confident that the mice with genotype LO have a probability of obesity between 0.02 and 0.29 higher (i.e. between 2 and 29 percentage points higher) than mice with genotype LL.

0 is not within the 95% interval for $p_{LO} - p_{LL}$, so $p_{LO} - p_{LL} = 0$ is not plausible. We would reject the null hypothesis that $p_{LO} = p_{LL}$ at the level $\alpha = 0.05$.

Genotype LO is associated with an increase of obesity rate (compared to LL)!

Smoking cessation

	no contact	group counseling	total
☺ quit smoking for 1 year	1	26	27
☹ resumed within a year	30	69	99
Total	31	95	126

$p_1 = \mathbb{P}\{\text{☺} | \text{ no contact}\}$ and $p_2 = \mathbb{P}\{\text{☺} | \text{ group counseling}\}$

$\hat{p}_1 = \frac{1}{31} = .032 \quad < \quad \hat{p}_2 = \frac{26}{95} = .274.$

But for a confidence interval for $p_2 - p_1$ we use
$\tilde{p}_1 = \frac{2}{33} = .061$ and $\tilde{p}_2 = \frac{27}{97} = .278$. We get $\tilde{p}_2 - \tilde{p}_1 =$**.218**,

$$\text{SE}_{\tilde{p}_2 - \tilde{p}_1} = \sqrt{\frac{.061 * .939}{33} + \frac{.278 * .722}{97}} = .062$$

95% confidence interval: $.218 \pm 1.960 * .062$ i.e **(0.097, 0.338)**.
90% confidence interval: $.218 \pm 1.645 * .062$ i.e **(0.116, 0.319)**.

Smoking cessation: conclusion

We are 90% confident that the increase in the probability of quitting smoking provided by group counseling (compared to no contact) is between 11.6% and 31.9 %

The 2 proportions can also be compared using the chi-square test of independence, which can be used for 2x2 or larger tables.

But before: `prop.test()` to do it in R.

prop.test() in R

```
> prop.test(c(71, 45), c(105, 87))

    2-sample test for equality of proportions with
    continuity correction

data:  c(71, 45) out of c(105, 87)
X-squared = 4.3837, df = 1, p-value = 0.03628
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01046848 0.30742971
sample estimates:
   prop 1    prop 2
0.6761905 0.5172414
```

by default: same CI we get by hand

prop.test() in R

```
> prop.test(c(71, 45), c(105, 87), correct=F)

    2-sample test for equality of proportions without
    continuity correction

data:  c(71, 45) out of c(105, 87)
X-squared = 5.0264, df = 1, p-value = 0.02496
alternative hypothesis: two.sided
95 percent confidence interval:
 0.02097751 0.29692068
sample estimates:
   prop 1    prop 2
0.6761905 0.5172414
```

with option `correct=F`: same X^2 test we get by hand.

Outline

Two categorical variables, 2 or more levels each

6,800 German men were sampled.

		Hair color				
		brown	black	fair	red	total
Eye color	brown	438	288	115	16	857
	gray/green	1387	746	946	53	3132
	blue	807	189	1768	47	2811
	total	2632	1223	2829	116	6800

H_0: Hair color and eye color are independent

H_A: Hair and eye color are not independent: are associated.

Null hypothesis of independence

H_0 can be stated in many ways:

The frequencies of **eye** colors do not depend on hair color:

$$\mathbb{P}\{\text{blue eyes}|\text{brown hair}\} = \mathbb{P}\{\text{blue eyes}|\text{black hair}\}$$
$$= \mathbb{P}\{\text{blue eyes}|\text{fair hair}\} = \mathbb{P}\{\text{blue eyes}|\text{red hair}\}$$

etc. with all other **eye** colors.

Or:

The frequencies of **hair** colors do not depend on eye color.

$$\mathbb{P}\{\text{red hair}|\text{brown eyes}\} = \mathbb{P}\{\text{red hair}|\text{gray/green eyes}\}$$
$$= \mathbb{P}\{\text{red hair}|\text{blue eyes}\}$$

etc. with all other **hair** colors.

H_A states that at least **one** of these equalities is not true.

The chi-square test of independence

1. H_0: the 2 categorical variables are independent.
   H_A: they are associated in some way.
2. Summary statistic:

$X^2 = \sum_{\text{cells}} \frac{(O-E)^2}{E}$ where expected values are:

$$E = \frac{\text{Row total * Column total}}{\text{Grand total}}$$

Expectation is H_0 is true: $X^2 \sim \chi^2_{\text{df}}$ distribution with

$$\text{df} = (\#\text{ columns} - 1)(\#\text{ rows} - 1).$$

3. p-value: $\mathbb{P}\{\chi^2_{\text{df}} \geq X^2\}$.
4. Conclusion: reject independence (H_0) and declare association if p-value$< \alpha$, fail to reject it p-value$> \alpha$.

Why these expected values?

$$E = \frac{\text{Row total * Column total}}{\text{Grand total}}$$

	Hair color				
	brown	black	fair	red	total
brown	438	288	115	16	857
gray/green	1387	746	946	53	3132
blue	807	189	1768	47	2811
total	2632	1223	2829	116	6800

If H_0 is true: $p_{\text{brown eyes}}$, is the same for all hair colors, but we don't know this value. Best guess:

$$\hat{p}_{\text{brown eyes}} = \frac{\text{total \# brown eyes}}{\text{total \# men}} = \frac{857}{6800} = .126$$

Expected # brown eyes with brown hair:

$$2632 * \hat{p}_{\text{brown eyes}} = 2632 * \frac{857}{6800} = 331.71.$$

Expected # brown eyes with black hair:

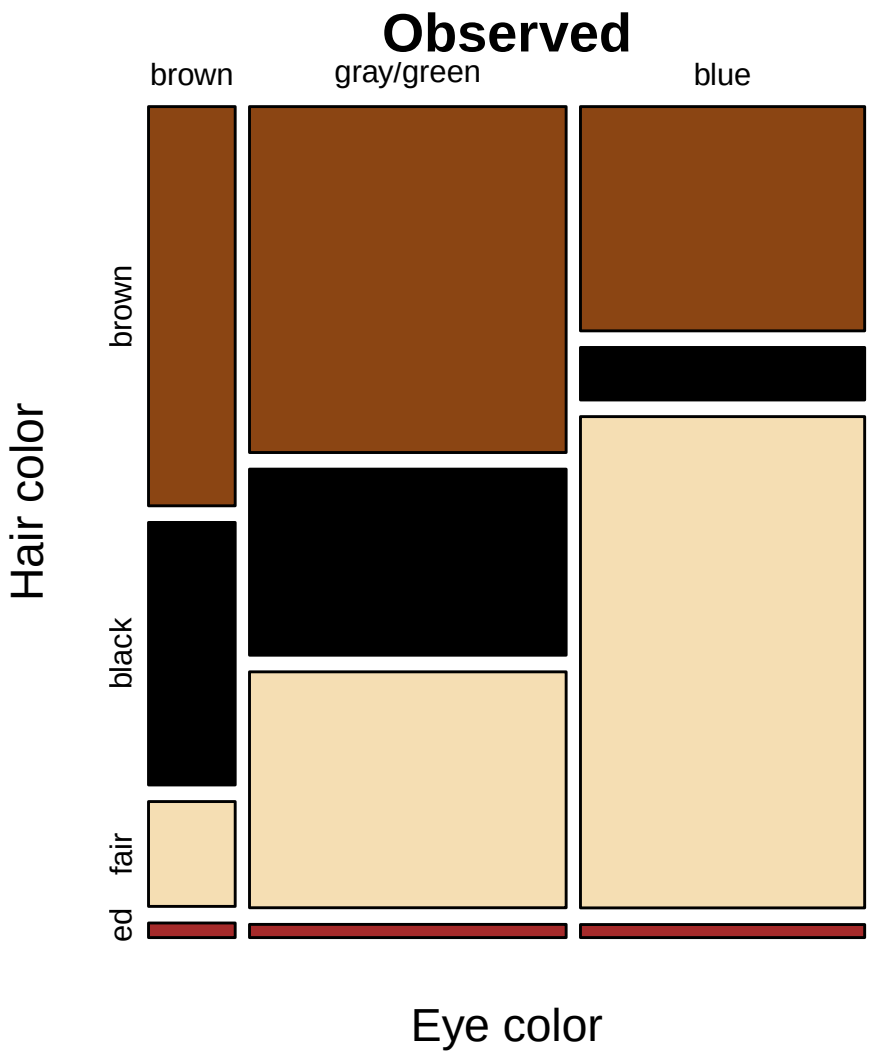
$$1223 * \hat{p}_{\text{brown eyes}} = 2632 * \frac{857}{6800} = 154.13.$$

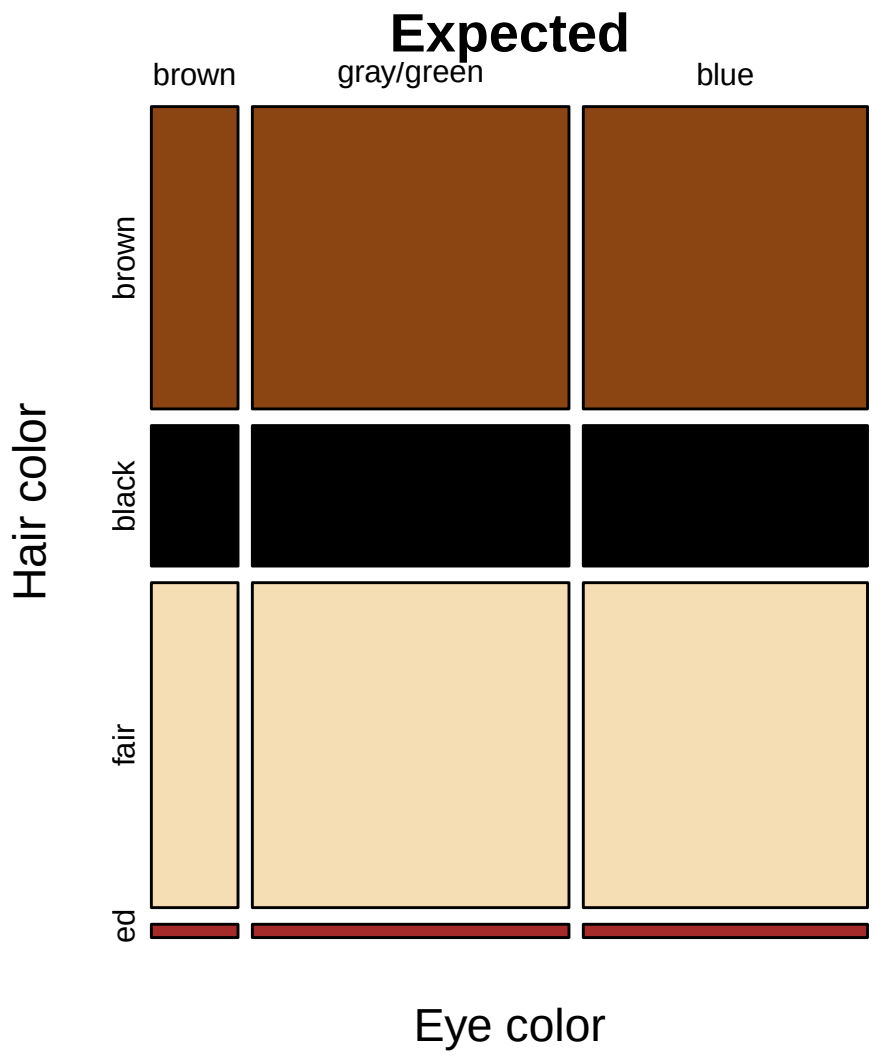
Expected values

	Hair color				
	brown	black	fair	red	total
brown	438	288	115	16	857
gray/green	1387	746	946	53	3132
blue	807	189	1768	47	2811
total	2632	1223	2829	116	6800

	Hair color				
	brown	black	fair	red	total
brown	**331.71**	**154.13**	**356.54**		857
gray/green					3132
blue					2811
total	2632	1223	2829	116	6800

Mosaic plots





Mosaic plots: R commands

```
> mat = matrix(c(438,1387,807,288,746,189,115,946,1768,16,53,47),3,4)
> rownames(mat) = c("brown","gray/green","blue")
> colnames(mat) = c("brown","black","fair","red")
> names(dimnames(mat)) = c("Eye color","Hair color")
> mat
           Hair color
Eye color    brown black fair red
  brown        438   288  115  16
  gray/green  1387   746  946  53
  blue         807   189 1768  47

> expected = apply(mat,1,sum) %*% t(apply(mat,2,sum)) / sum(mat)
> rownames(expected) = c("brown","gray/green","blue")
> names(dimnames(expected)) = c("Eye color","Hair color")
> expected
           Hair color
Eye color        brown    black      fair      red
  brown       331.7094 154.1340  356.5372 14.61941
  gray/green 1212.2682 563.2994 1303.0041 53.42824
  blue       1088.0224 505.5666 1169.4587 47.95235

> mosaicplot(mat,      col=c("chocolate4","black","wheat","brown"))
> mosaicplot(expected, col=c("chocolate4","black","wheat","brown"))
```

X^2, degree of freedom and p-value

$$X^2 = \sum_{\text{all cells}} \frac{(\text{obs} - \text{exp})^2}{\text{exp}} = \frac{(438 - 331.71)^2}{331.71} + \cdots + \frac{(47 - 47.95)^2}{47.95}$$

$$= 34.1 + 116.3 + 163.6 + 0.1 + 25.2 + 59.3 + 97.8 + 0.004$$
$$+72.6 + 198.2 + 306.3 + 0.02 = \mathbf{1073.5}$$

Degree of freedom: # pieces of information (cells) needed to fill in entire table. Marginals (totals in the margins) are known.

$$\text{df} =$$

Here $df = 6$.

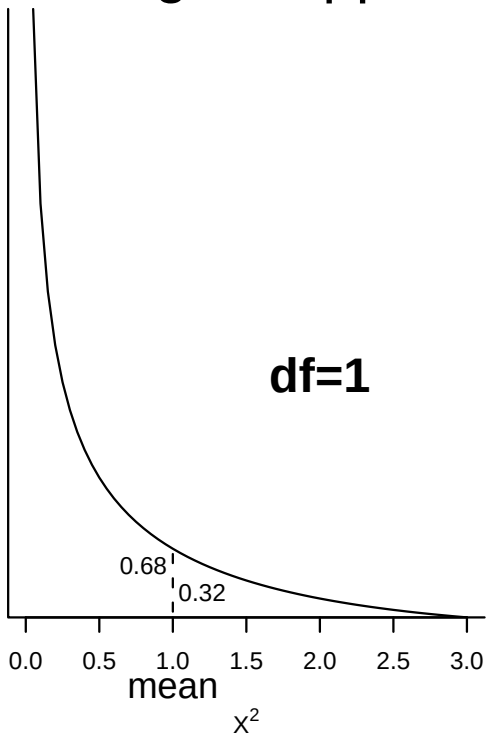
p-value: $\mathbb{P}\{\chi^2_6 \geq 1073.5\}$. Table A gives p-value $< .0001$. Overwhelming evidence that hair and eye color are not independent. They are associated.

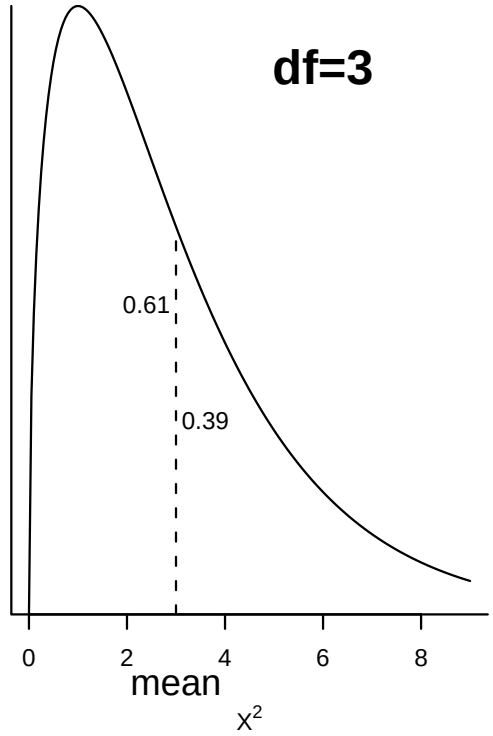
The χ^2 distribution

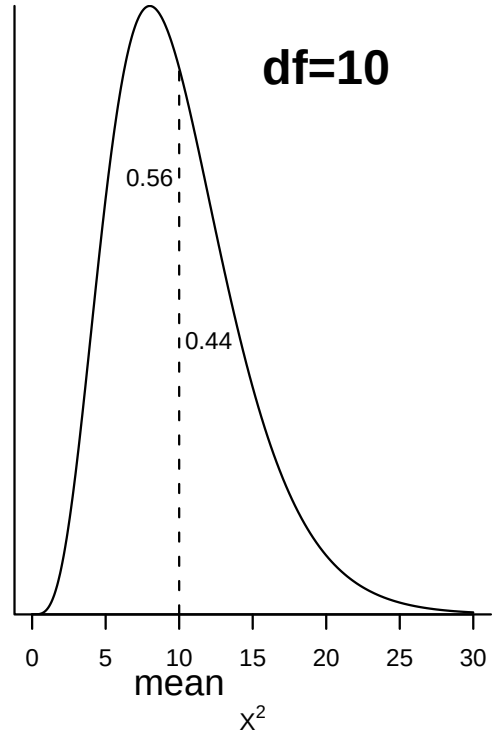
$X^2 \geq 0$ always

$X^2 = 0$ means observed = expected counts: data in perfect agreement with the claim. X^2 close to 0: supports H_0.

X^2 large: supports H_A.







benchmark: $X^2 \leq$ df supports H_0.

Interpretation

We can now look at the largest contributions to X^2 and see where the association is the strongest.
$(O-E)^2/E$ values (sum=$X^2 = 1073.5$):

	Hair color			
	brown	black	fair	red
brown	34.1	116.3	**163.6**	0.1
gray/green	25.2	59.3	97.8	.004
blue	72.6	**198.2**	**306.3**	0.02

Blue eyes/fair hair are associated: blue-eyed people tend to have fair hair more frequently than non blue-eyed people.

On the opposite, blue-eyed people tend to have black hair less frequently than non blue-eyed people, and people with fair hair tend to have brown eyes less frequently than non-fair hair people.

Assumptions

Independence of observations

Expected counts ≥ 5, for the χ^2 distribution to be a good approximation.

If some cells have small counts, what can be done?

Fisher's test (2×2 tables only): but won't cover this.

Group cells together.

Ex: eye/hair colors with 10-fold decrease in sample size.

Grouping cells

	Hair color				
	brown	black	fair	red	total
brown	44 (32.9)	29	11	1 (1.37)	85
gray/green	138 (121.0)	75	95	5 (5.06)	313
blue	81 (109.1)	19	177	5 (4.56)	282
total	263	123	283	11	680

	Hair color			
	brown	black	fair/red	total
brown	44 (32.9)	29	12 (36.8)	85
gray/green	138 (121.0)	75	100 (135.3)	313
blue	81 (109.1)	19	182 (121.9)	282
total	263	123	294	680

Now expected counts are ≥ 5 in all cells. We get df$=$,
$X^2 = 107$ and $p < .0001$.

Mice: genotypes LO and LL and phenotype

Are phenotype and genotype at a given locus independent? associated?

	LO	LL	total
obese	71	45	116
lean	34	42	76
total	105	87	192

He want to test H_0: $p_{LO} = p_{LL}$ against H_A: $p_{LO} \neq p_{LL}$.

Equivalently:
H_0: genotype and obesity phenotype are **independent**.
H_A: genotype at the locus and phenotype are not independent: one genotype tends to be associated with one phenotype.

Test of independence

Observed counts:

	LO	LL	total
☺	71	45	116
☹	34	42	76
total	105	87	192

Expected counts if independent phenotype & genotype:

	LO	LL	total
☺	63.44	52.56	116
☹	41.56	34.44	76
total	105	87	192

$$X^2 = \sum_{\text{all cells}} \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$

$$= \qquad = 5.026$$

Validity: Are all expected counts ≥ 5?

Test of independence

3. **Calculate the p-value.** If there is independence (success does not depend on drug) then X^2 has a χ^2 distribution with df$=1$ here.
   p-value=$\mathbb{P}\{\chi^2_{1\,df} \geq 5.026\}$ Table A: $.025 <$ p-value $< .05$.
4. **Conclusion:** moderate evidence that the phenotype is associated with the genotype. Genotypes have different obsesity rates ($p = 0.025$, chi-square test of independence).
   Furthermore, we had $\hat{p}_{LO} = .68 > \hat{p}_{LL} = .52$. There is evidence that genotype LO has higher obesity rate.

chisq.test() with R: data already in table

```
> mice = matrix( c(71,34,45,42),  2,2)
> mice
     [,1] [,2]
[1,]   71   45
[2,]   34   42

> chisq.test(mice)

        Pearson’s Chi-squared test with
        Yates’ continuity correction
data:  mice
X-squared = 4.3837, df = 1, p-value = 0.03628


> chisq.test(mice, correct=FALSE)

         Pearson’s Chi-squared test
data:  mice
X-squared = 5.0264, df = 1, p-value = 0.02496
```

`chisq.test`: full data in columns

```
> mice
    phenotype genotype
1       obese       LO
2       obese       LO
...
71      obese       LO
72       lean       LO
73       lean       LO
...
105      lean       LO
106     obese       LL
107     obese       LL
...
191      lean       LL
192      lean       LL
```

```
> table(mice$phenotype, mice$genotype)
       LO LL
  obese 71 45
  lean  34 42

> with(mice, table(phenotype,genotype))
          genotype
phenotype LO LL
    obese 71 45
    lean  34 42
```

`chisq.test`: full data in columns

```
> chisq.test( table(mice$phenotype, mice$genotype) )

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(mice$phenotype, mice$genotype)
X-squared = 4.3837, df = 1, p-value = 0.03628


> with(mice, chisq.test(table(phenotype,genotype)) )

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(phenotype, genotype)
X-squared = 4.3837, df = 1, p-value = 0.03628
```

chisq.test on eye/hair color

```
> mat = matrix(c(438,1387,807,288,746,189,115,946,
                 1768, 16, 53, 47),
               3,4)
> mat
     [,1] [,2] [,3] [,4]
[1,]  438  288  115   16
[2,] 1387  746  946   53
[3,]  807  189 1768   47

> chisq.test(mat)

        Pearson's Chi-squared test
data:  mat
X-squared = 1073.508, df = 6, p-value < 2.2e-16
```

chisq.test: warning if some E's < 5

```
> smallmat = matrix(c(44,138,81,29,75, 19,11,95,177,
                        1, 5, 5),
                      3,4)
> smallmat
     [,1] [,2] [,3] [,4]
[1,]   44   29   11    1
[2,]  138   75   95    5
[3,]   81   19  177    5

> chisq.test(smallmat)

        Pearson's Chi-squared test

data:  smallmat
X-squared = 108.2808, df = 6, p-value < 2.2e-16

Warning message:
In chisq.test(smallmat) : Chi-squared approximation may
                          be incorrect
```

Chi-square: goodness or fit vs. test of independence

Two “chi-square” tests

Analogies:

- Both for categorical data
- Same definition for the X^2 value (after getting E values)
- Same chi-square distribution to obtain the p-value

Different calculations of **expected** values:

- goodness-of-fit: $E_i = \text{Row total} * p_i$
- test of independence: $E_{ij} = \dfrac{\text{Row total} * \text{Column total}}{\text{Grand total}}$

Different degrees of freedom:

- $\#\text{cells} - 1$ vs. $(\#\text{rows} - 1)(\#\text{columns} - 1)$

Chi-square: goodness or fit vs. test of independence

Different numbers of variables:

goodness-of-fit: 1 categorical variable

test of independence: 2 categorical variables

Different questions, i.e. different H_0's:

goodness-of-fit: compare proportion in the sample with proportions from a claim.
ex: do seals swim clockwise with $p = 0.50$?

test of independence: compare proportions for 1 variable across the categories of the other variable.
ex: do sheep survive with more often when vaccinated than not? i.e. $p_{\text{survive}|\text{vaccine}} > p_{\text{survive}|\text{control}}$?
ex: are $p_{\text{brown hair}}$'s all the same in all groups of eye color?