Chi-squared Tests

Part 2 - Tests of Independence

Ryan Miller

Introduction

In our previous lecture, we learned about the *Chi-squared test statistic*:

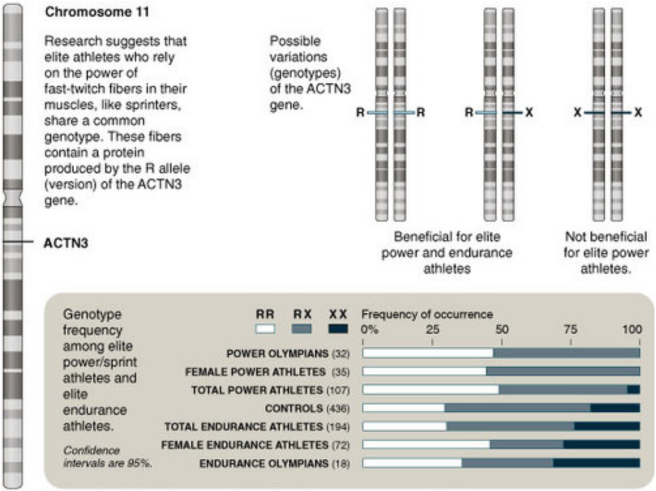
$$X^2 = \sum_{i=1}^{k} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

This test statistic gives us a single, standardized value summarizing how the observed frequencies of a nominal categorical variable differ from the frequencies that would be expected under a given null hypothesis.

Introduction (cont.)

- In our previous lecture we analyzed the frequencies of a *single categorical variable*
  - Frequencies were summarized using a *one-way frequency table*
  - This type of Chi-squared test is known as a **goodness of fit test**
- We can use the same test statistic formula to analyze the relationship between *two categorical variables*
  - Frequencies are summarized using a *two-way frequency table*
  - This type of Chi-squared test is known as a **test of independence**

Example - The Sprinter Gene



Sources: Stephen M. Roth, Ph.D., University of Maryland; American Journal of Human Genetics

Example - The Sprinter Gene

Researchers collected the genotypes of three groups, Olympic sprinters and endurance athletes, and controls who weren't elite athletes:

Group	RR	RX	XX	Total
Control	130	226	80	436
Sprint	53	48	6	107
Endurance	60	88	46	194
Total	243	362	132	737

If a person's genotype is independent of their success in sprint/endurance events, what distribution of RR, RX, and XX would you expect in each group?

Example - The Sprinter Gene

- The *null hypothesis* of independence implies the distribution of genotypes should be same for each row
  - Thus, since 243/737 (33.0%) individuals in the sample have the RR genotype, we'd expect 33% of each group to have the RR genotype under the null hypothesis
  - Similarly, we'd expect 362/737 (49.1%) of each group to be RX, and 132/737 (17.9%) of each group to be XX

Example - The Sprinter Gene

We call these *pooled proportions* as they pool all of the data together by ignoring one of the table's variables:

	RR	RX	XX	Total
Control	$p_{rr} = 0.33$	$p_{rx} = 0.49$	$p_{xx} = 0.18$	1
Sprint	$p_{rr} = 0.33$	$p_{rx} = 0.49$	$p_{xx} = 0.18$	1
Endurance	$p_{rr} = 0.33$	$p_{rx} = 0.49$	$p_{xx} = 0.18$	1

Example - The Sprinter Gene

The sample size of each group is multiplied by these pooled proportions to determine the counts that are expected under the null hypothesis:

	RR	RX	XX
Control	436*0.33 = 143.9	436*0.49 = 213.6	436*0.18 = 78.5
Sprint	107*0.33 = 35.3	107*0.49 = 52.5	107*0.18 = 19.3
Endurance	194*0.33 = 64.0	194*0.49 = 95.3	194*0.18 = 34.9

Note: this procedure is symmetric, so we could find the same expected counts using pooled proportions that ignore the table's column variable.

Example - The Sprinter Gene

Once we've determined the expected counts, the χ^2 test statistic is calculated in the usual manner:

$$\begin{aligned}\chi^2 &= \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \\ &= \frac{(130-143.9)^2}{143.9} + \frac{(226-213.6)^2}{213.6} + \frac{(80-78.5)^2}{78.5} \\ &+ \frac{(53-35.3)^2}{35.3} + \frac{(48-52.5)^2}{52.5} + \frac{(6-19.3)^2}{19.3} \\ &+ \frac{(60-64.0)^2}{64.0} + \frac{(88-95.3)^2}{95.3} + \frac{(46-34.9)^2}{34.9} = 24.8\end{aligned}$$

For an R by C two-way table, the degrees of freedom of the test are $(R-1)(C-1)$, so $df = 4$, and the p-value is 0.000055

Example - The Sprinter Gene

The Chi-squared test of independence allows use to evaluate whether are two variables are associated, but it doesn't tell us how they are related. For this we should use standardized residuals:

```
##                   RR         RX        XX
## Control   -2.192769  1.7756263  0.373375
## Sprint     3.941379 -0.9529769 -3.589783
## Endurance -0.705418 -1.2195484  2.454892
```

Thus, we conclude that the study provides strong evidence of an association between ACTN3 and sport ($p < 0.001$), with the RR genotype being over represented among elite sprinters.

Practical Considerations

- The Chi-squared test is built upon an assumption that with a large enough sample size the expected count in each cell of a frequency table will be approximately Normally distributed
  - In practice this assumption is assessed by checking whether each cell has an expected count of at least 5
  - When this assumption isn't met, **Fisher's Exact Test** should be used
- The Chi-squared test can also be unreliable and difficult to interpret when the data contain a large number of categories
  - Collapsing related categories together or restricting the eligibility criteria for the analysis are reasonable strategies