Chi-squared Tests

Part 2 - Goodness of Fit Testing

Ryan Miller

Grinnell College
Statistics

Introduction

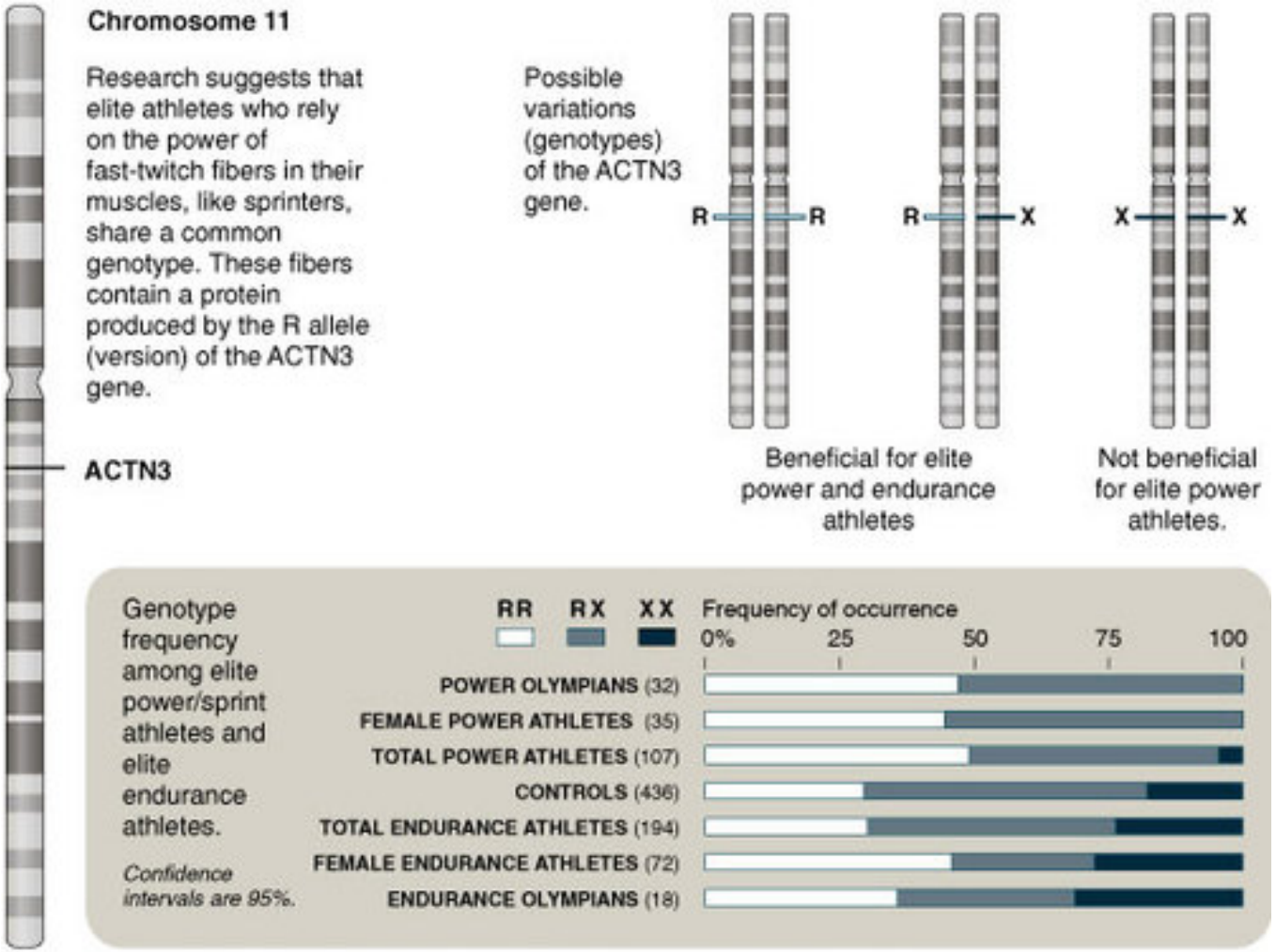
Chi-squared tests use a test statistic of the form:

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

- For a *one-way frequency table* (single categorical variable) this is called a **goodness of fit test**
- For a *two-way frequency table* (two categorical variables) this is called a **test of independence**

Today we'll focus on tests of independence, which are used to assess whether two categorical variables are associated.

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:

	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 we expect in each group?

Example - The Sprinter Gene

- A *null hypothesis* of independence implies the row proportions within each group are identical
  - 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%) to be XX

Example - The Sprinter Gene

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

	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 in 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 that this procedure is symmetric, so we calculated pooled proportions by ignoring 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

As always, we should follow-up our significant result by looking at 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, with the number of Olympic sprinters having the RR genotype being 3.9 standard deviations higher than expected under independence.

Practical Considerations

- The Chi-squared test is built upon the assumption that with a large enough sample size the counts in each cell of a frequency table will be approximately Normally distributed
  - This assumption has been translated into checking that each cell has an expected count of at least 5
  - When this assumption isn't met **Fisher's Exact Test**, which is described in this week's lab, provides an alternative
- 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