

Monday, Apr 17

### Genetic Linkage (Bateson, Saunders, and Punnett)

Genetic linkage is the tendency for DNA sequences that are close on a chromosome to be inherited together. An early demonstration of linkage used two traits of sweet peas: flower color (purple or red) and pollen grain shape (long or round). If there is *no linkage* then the probabilities of each combination of traits are as shown in the table below.

Traits	Probability	Count	
		observed	expected
Purple and Long	9/16	284	
Purple and Round	3/16	21	
Red and Long	3/16	21	
Red and Round	1/16	55	

How would we conduct a goodness-of-fit test for no linkage?

## Tests of Independence

Two categorical variables are said to be **independent** if the distribution of one variable does not depend on the value of the other variable.

**Example:** Consider the following data where a sample of 1398 children were classified with respect to tonsil size and carrier status of *Streptococcus pyogenes*.<sup>1</sup>

Size	Carrier		Total
	yes	no	
<b>small</b>	19	497	516
<b>medium</b>	29	560	589
<b>large</b>	24	269	293
<b>Total</b>	72	1326	1398

These are the *observed* counts. The table below shows the estimated *expected* counts under the assumption that tonsil size and carrier status are *independent*. How are they computed?

Size	Carrier		Total
	yes	no	
<b>small</b>	26.58	489.42	516
<b>medium</b>	30.33	558.67	589
<b>large</b>	15.09	277.91	293
<b>Total</b>	72	1326	1398

---

<sup>1</sup>Holmes, M. C. & Willaims, R. E. O. (1954). The distribution of carriers of *Streptococcus pyogenes* among 2413 healthy children. *Journal of Hygiene*, 52, 165–179.

## Steps of a Test of Independence

1. State hypotheses in terms of independence of the variables.
2. Check assumptions (all expected counts should be at least five).
3. Compute the  $X^2$  test statistic. Estimate the expected counts using the formula

$$\frac{R \times C}{T}$$

where  $R$  and  $C$  are the the sum of the observed counts in the corresponding row and column, respectively, and  $T$  is the total of all the observed counts.

4. Compute the  $p$ -value using  $(r - 1)(c - 1)$  as the degrees of freedom, where  $r$  and  $c$  are the number of rows and columns of observed counts in the table, respectively.
5. Make a decision/conclusion.

## The Two-Sample Test of Proportions

Recall the study of the influence of applicant's sex on personnel decisions.<sup>2</sup>

Applicant	Promotion		Total
	yes	no	
<b>male</b>	21	3	24
<b>female</b>	14	10	24
<b>Total</b>	35	13	48

We could investigate the relationship between applicant sex and promotion decision by a test of the hypotheses  $H_0: p_m - p_f = 0$  versus  $H_a: p_m - p_f \neq 0$  using the test statistic

$$z = \frac{\hat{p}_m - \hat{p}_f}{\sqrt{\hat{p}(1 - \hat{p})(1/n_m + 1/n_f)}},$$

which yields a test statistic of  $z \approx 2.27$  and a p-value of about 0.02. How is this test related to the test of independence using the  $X^2$  test statistic? How is the  $z$  test statistic limited?

---

<sup>2</sup>Rosen, B. & Jerdee, J. (1974). Influence of sex role stereotypes on personnel decisions. *Journal of Applied Psychology*, 59, 9–14.

## Comparison of Chemotherapy Treatment Strategies

Consider the following data from a randomized experiment comparing two strategies for chemotherapy.<sup>3</sup>

Strategy	Tumor Response				Total
	progressive disease	no change	partial remission	complete remission	
<b>sequential</b>	32	57	34	28	151
<b>alternating</b>	53	51	23	21	148
<b>Total</b>	85	108	57	49	299

---

<sup>3</sup>Holtbrugge, W. & Schumacher, M. (1991). A comparison of regression models for the analysis of ordered categorical data. *Applied Statistics*, 40, 249–259.