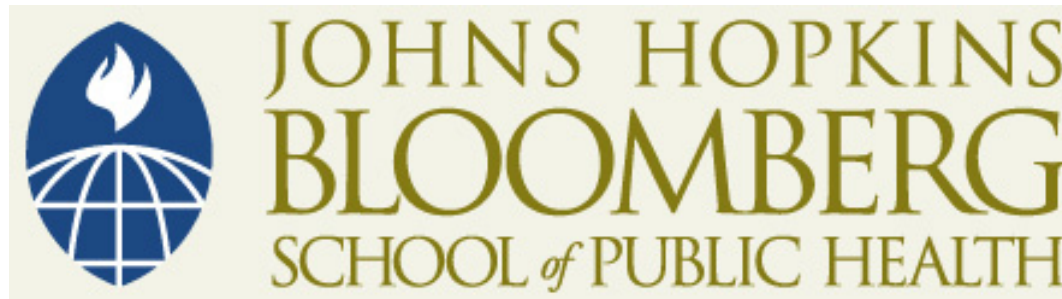


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and Karl W. Broman. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

Statistics for laboratory scientists II

Karl W Broman

Department of Biostatistics, Johns Hopkins University

Office: E3612 SPH; Email: kbroman@jhsph.edu

<http://www.biostat.jhsph.edu/~kbroman>

TA: Qing Li (qli@jhsph.edu, E3035)

Logistics

Lectures: MWF 10:30-11:30 (W2033 SPH)

Discussion/lab: W 1:30-3:30 (W3025 first half; W2033 second half)

Office hours: **Karl:** MF 1:30-2:30 (E3612 SPH)
Qing: by appointment (E3035 SPH)

Textbooks: Samuels & Witmer (2002) Statistics for the life sciences
Gonick & Smith (1993) The cartoon guide to statistics.
[recommended]
Dalgaard (2002) Introductory statistics with R statistics.
[recommended]

Grading

Grade based on:

- 3 Computer labs (67%)
- 1 Final project (33%)

Other work:

- Homework
- Reading assignments
- Deep and careful thought
- Discussions

Final project

- Obtain some real experimental data.
- Analyze the data
- Write a 4–8 page double-spaced paper describing the data, the goal, your analysis, and your results.
(Use the usual Introduction – Methods – Results – Discussion format.)

This term

- Goodness of fit
- Contingency tables
- Analysis of variance (Anova)
- More on multiple comparisons
- Linear regression
- More on design of experiments
- ...

Goodness of fit - 2 classes

A	B
78	22

Do these data correspond reasonably to the proportions 3:1?

We could use what we learned last term...

During the previous quarter we discussed several options for testing $p_A = 0.75$:

- Exact p-value
- Normal approximation
- Randomization test

Goodness of fit - 3 classes

AA	AB	BB
35	43	22

Do these data correspond reasonably to the proportions 1:2:1?

The likelihood-ratio test (LRT)

Back to the first example:

A	B
n_A	n_B

We want to test $H_0 : (p_A, p_B) = (\pi_A, \pi_B)$ versus $H_a : (p_A, p_B) \neq (\pi_A, \pi_B)$.

MLE under H_a : $\hat{p}_A = n_A/n$ where $n = n_A + n_B$.

Likelihood under H_a : $L_a = \Pr(n_A | p_A = \hat{p}_A) = \binom{n}{n_A} \times \hat{p}_A^{n_A} \times (1 - \hat{p}_A)^{n - n_A}$

Likelihood under H_0 : $L_0 = \Pr(n_A | p_A = \pi_A) = \binom{n}{n_A} \times \pi_A^{n_A} \times (1 - \pi_A)^{n - n_A}$

Likelihood ratio test statistic: $LRT = 2 \times \ln(L_a/L_0)$

If H_0 is true, then LRT follows a $\chi^2(df=1)$ distribution (approximately).

Likelihood-ratio test for the example

We observed $n_A = 78$ and $n_B = 22$.

$H_0 : (p_A, p_B) = (0.75, 0.25)$

$H_a : (p_A, p_B) \neq (0.75, 0.25)$

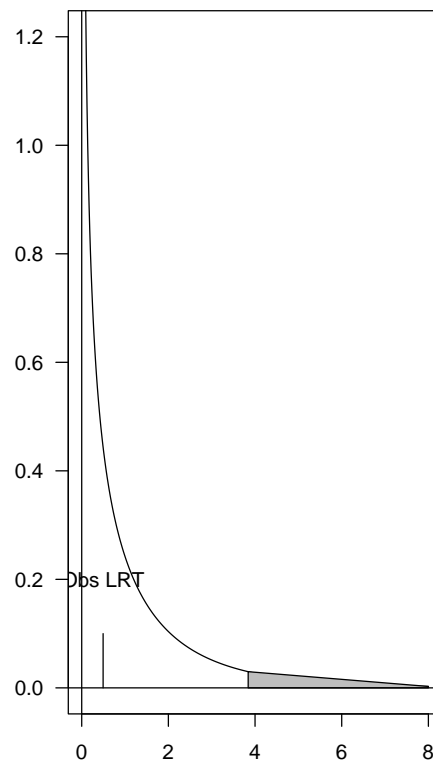
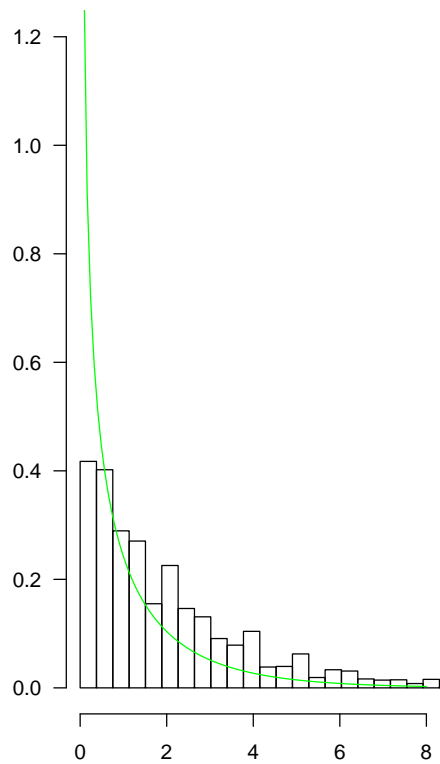
$L_a = \Pr(n_A=78 | p_A=0.78) = \binom{100}{78} \times 0.78^{78} \times 0.22^{22} = 0.096$.

$L_0 = \Pr(n_A=78 | p_A=0.75) = \binom{100}{78} \times 0.75^{78} \times 0.25^{22} = 0.075$.

$LRT = 2 \times \ln(L_a/L_0) = 0.49$. Using a $\chi^2(df=1)$ distribution, we get a p-value of 0.48.

In R: `p-value = 1 - pchisq(0.49, 1)`

We therefore have no evidence against the hypothesis $(p_A, p_B) = (0.75, 0.25)$.



A little math ...

$$n = n_A + n_B, \quad n_A^0 = E[n_A | H_0] = n \times \pi_A, \quad n_B^0 = E[n_B | H_0] = n \times \pi_B.$$

$$\text{Then } L_a/L_0 = \left(\frac{n_A}{n_A^0}\right)^{n_A} \times \left(\frac{n_B}{n_B^0}\right)^{n_B}.$$

$$\text{Or equivalently } \text{LRT} = 2 \times n_A \times \ln\left(\frac{n_A}{n_A^0}\right) + 2 \times n_B \times \ln\left(\frac{n_B}{n_B^0}\right).$$

Why do this?

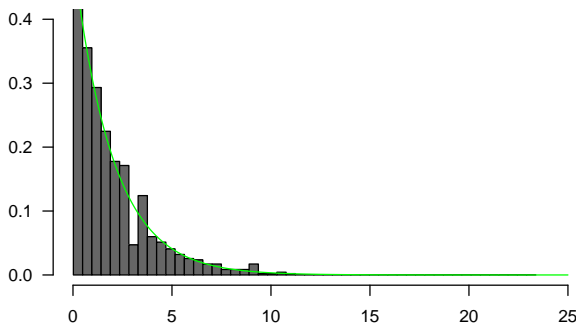
Generalization to more than two groups

If we have k groups, then the likelihood ratio test statistic is

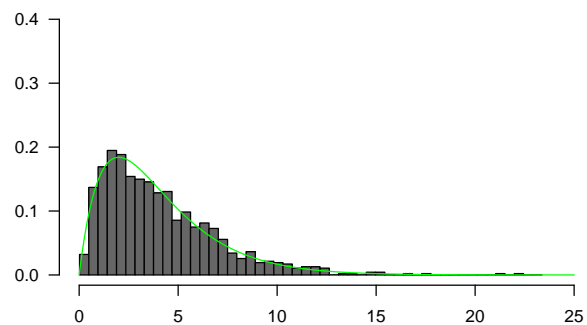
$$\text{LRT} = 2 \times \sum_{i=1}^k n_i \times \ln \left(\frac{n_i}{n_i^0} \right)$$

If H_0 is true, $\text{LRT} \sim \chi^2(\text{df}=k-1)$.

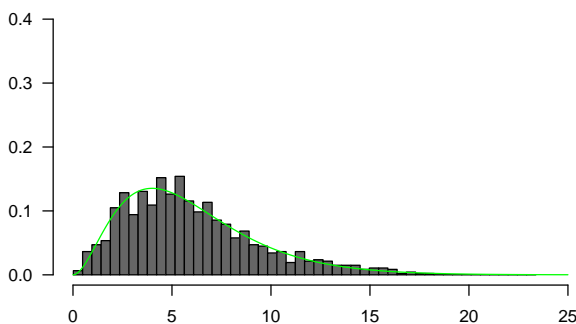
3 groups: χ^2 (df=2)



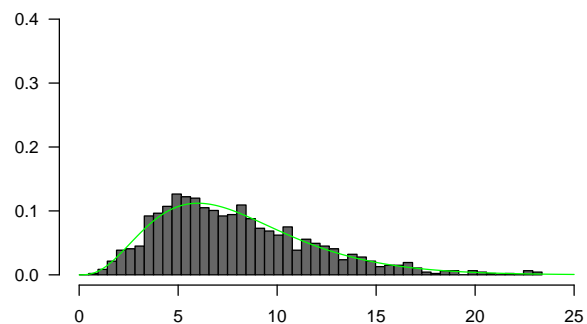
5 groups: χ^2 (df=4)



7 groups: χ^2 (df=6)



9 groups: χ^2 (df=8)

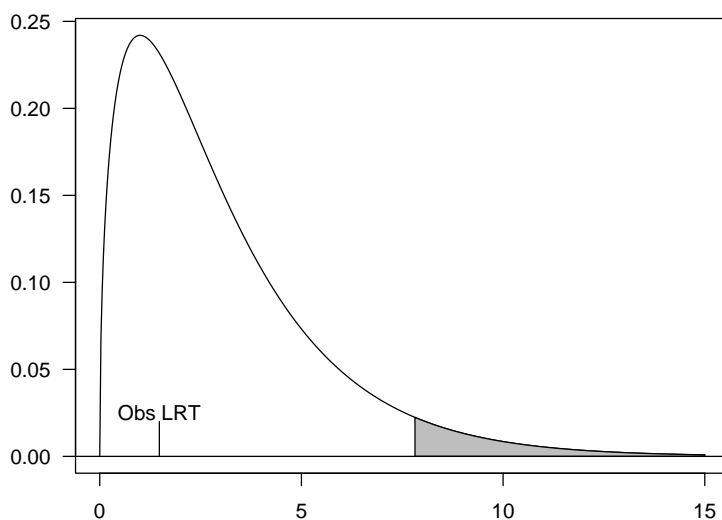


Example

In a dihybrid cross of tomatoes we expect the ratio of the phenotypes to be 9:3:3:1. In 1611 tomatoes, we observe the numbers 926, 288, 293, 104. Do these numbers support our hypothesis?

Phenotype	n_i	n_i^0	n_i/n_i^0	$n_i \times \ln(n_i/n_i^0)$
Tall, cut-leaf	926	906.2	1.02	20.03
Tall, potato-leaf	288	302.1	0.95	-13.73
Dwarf, cut-leaf	293	302.1	0.97	-8.93
Dwarf, potato-leaf	104	100.7	1.03	3.37
Sum	1611			0.74

Results



The test statistics LRT is 1.48. Using a $\chi^2(df=3)$ distribution, we get a p-value of 0.69. We therefore have no evidence against the hypothesis that the ratio of the phenotypes is 9:3:3:1.

The chi-square test

There is an alternative technique. The test is called the chi-square test, and has the greater tradition in the literature. For two groups, calculate the following:

$$X^2 = \frac{(n_A - n_A^0)^2}{n_A^0} + \frac{(n_B - n_B^0)^2}{n_B^0}$$

If H_0 is true, then X^2 is a draw from a $\chi^2(df=1)$ distribution (approximately).

Example

In the first example we observed $n_A = 78$ and $n_B = 22$. Under the null hypothesis we have $n_A^0 = 75$ and $n_B^0 = 25$. We therefore get

$$X^2 = \frac{(78-75)^2}{75} + \frac{(22-25)^2}{25} = 0.12 + 0.36 = 0.48.$$

This corresponds to a p-value of 0.49. We therefore have no evidence against the hypothesis $(p_A, p_B) = (0.75, 0.25)$.

Note: using the likelihood ratio test we got a p-value of 0.48.

Generalization to more than two groups

As with the likelihood ratio test, there is a generalization to more than just two groups.

If we have k groups, the chi-square test statistic we use is

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n_i^0)^2}{n_i^0} \sim \chi^2(\text{df}=k-1)$$

Tomato example

For the tomato example we get

$$\begin{aligned}\chi^2 &= \frac{(926-906.2)^2}{906.2} + \frac{(288-302.1)^2}{302.1} + \frac{(293-302.1)^2}{302.1} + \frac{(104-100.7)^2}{100.7} \\ &= 0.43 + 0.65 + 0.27 + 0.11 = 1.47\end{aligned}$$

Using a $\chi^2(\text{df}=3)$ distribution, we get a p-value of 0.69. We therefore have no evidence against the hypothesis that the ratio of the phenotypes is 9:3:3:1.

Note: using the likelihood ratio test we also got a p-value of 0.69.