STAT 513 fa 2019 Lec 10

Contingency tables

Karl B. Gregory

An asymptotic test for equal proportions

- **Exercise:** Suppose Y_1 and Y_2 are independent random variables such that

$$Y_1 \sim \text{Binomial}(n_1, p_1)$$
$$Y_2 \sim \text{Binomial}(n_2, p_2).$$

Derive the asymptotic likelihood ratio test for testing the hypotheses

$$H_0\text{: } p_1 = p_2 \text{ versus } H_1\text{: } p_1 \neq p_2. \tag{1}$$

Answer: The likelihood function for the parameters p_1 and p_2 based on the data Y_1 and Y_2 is given by

$$L(p_1, p_2; Y_1, Y_2) = \binom{n_1}{Y_1} p_1^{Y_1}(1-p_1)^{n_1-Y_1} \binom{n_2}{Y_2} p_2^{Y_2}(1-p_2)^{n_2-Y_2}.$$

The maximum likelihood estimators of p_1 and p_2 are given by

$$\hat{p}_1 = Y_1/n_1 \quad \text{and} \quad \hat{p}_2 = Y_2/n_2,$$

respectively. Under the null hypothesis we have $p_1 = p_2 = p_0$ for some p_0, and our estimator of p_0 is given by

$$\hat{p}_0 = \underset{p\in[0,1]}{\text{argmax}} \quad L(p, p; Y_1, Y_2) = \frac{Y_1 + Y_2}{n_1 + n_2},$$

which we obtain by setting the derivative of $\ell(p, p; Y_1, Y_2)$ with respect to p equal to zero and solving for p. The likelihood ratio is thus

$$\begin{aligned}
\text{LR}(Y_1, Y_2) &= \frac{\text{L}(\hat{p}_0, \hat{p}_0; Y_1, Y_2)}{L(\hat{p}_1, \hat{p}_2; Y_1, Y_2)} \\
&= \frac{\binom{n_1}{Y_1}\hat{p}_0^{Y_1}(1-\hat{p}_0)^{n_1-Y_1}\binom{n_2}{Y_2}\hat{p}_0^{Y_2}(1-\hat{p}_0)^{n_2-Y_2}}{\binom{n_1}{Y_1}\hat{p}_1^{Y_1}(1-\hat{p}_1)^{n_1-Y_1}\binom{n_2}{Y_2}\hat{p}_2^{Y_2}(1-\hat{p}_2)^{n_2-Y_2}} \\
&= \frac{\hat{p}_0^{Y_1+Y_2}(1-\hat{p}_0)^{n_1+n_2-Y_1-Y_2}}{\hat{p}_1^{Y_1}(1-\hat{p}_1)^{n_1-Y_1}\hat{p}_2^{Y_2}(1-\hat{p}_2)^{n_2-Y_2}}.
\end{aligned}$$

The asymptotic likelihood ratio test is based on the test statistic

$$\begin{aligned}-2\log \mathrm{LR}(Y_1, Y_2) &= -2[(Y_1 + Y_2)\log \hat{p}_0 \\ &\quad + (n_1 + n_2 - Y_1 - Y_2)\log(1 - \hat{p}_0) \\ &\quad - Y_1 \log \hat{p}_1 - (n_1 - Y_1)\log(1 - \hat{p}_1) \\ &\quad - Y_2 \log \hat{p}_2 - (n_2 - Y_2)\log(1 - \hat{p}_2)] \\ &= 2[Y_1 \log(\hat{p}_1/\hat{p}_0) \\ &\quad + Y_2 \log(\hat{p}_2/\hat{p}_0) \\ &\quad + (n_1 - Y_1)\log((1 - \hat{p}_1)/(1 - \hat{p}_0)) \\ &\quad + (n_2 - Y_2)\log((1 - \hat{p}_2)/(1 - \hat{p}_0))].\end{aligned}$$

Under the null hypotheses, there is one parameter to estimate, and under the alternate hypothesis, there are two parameters to estimate, so under the null hypothesis, $-2\log \mathrm{LR}(Y_1, Y_2)$ converges in distribution to a random variable with the chi-squared distribution with $2 - 1 = 1$ degree of freedom. The size-α asymptotic likelihood ratio test is thus

$$\text{Reject } H_0 \text{ iff } -2\log \mathrm{LR}(Y_1, Y_2) > \chi^2_{1,\alpha}.$$

- **Two-by-two contingency table notation:** Continuing the previous exercise, suppose we put the observed data in a table as follows:

	Successes	Failures	Total
Sample 1	Y_1	$n_1 - Y_1$	n_1
Sample 2	Y_2	$n_2 - Y_2$	n_2
Total	$Y_1 + Y_2$	$n_1 + n_2 - Y_1 - Y_2$	$n_1 + n_2$

In addition, suppose we make another table like the above, but we replace the observed counts in the center of the table with the expected counts under the null hypothesis, computing the expectations based on our estimator $\hat{p}_0$ of p_0. The table of expected values would be

	Successes	Failures	Total
Sample 1	$n_1\hat{p}_0$	$n_1(1 - \hat{p}_0)$	n_1
Sample 2	$n_2\hat{p}_0$	$n_2(1 - \hat{p}_0)$	n_2
Total	$Y_1 + Y_2$	$n_1 + n_2 - Y_1 - Y_2$	$n_1 + n_2$

If we denote by $O_{11}, O_{12}, O_{21}, O_{22}$ the observed counts in the first table and by $E_{11}, E_{12}, E_{21}, E_{22}$ the expected counts in the second table, then we may write

$$-2\log \mathrm{LR}(Y_1, Y_2) = 2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right).$$

The size-α asymptotic likelihood ratio test of H_0: $p_1 = p_2$ versus H_1: $p_1 \neq p_2$ can thus be formulated as

$$\text{Reject } H_0 \text{ iff } 2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right) > \chi^2_{1,\alpha}.$$

- **Exercise:** In a study of the efficacy of a surgery for treating migraine headaches, 75 subjects were randomly assigned to either a control or a treatment group. Those in the treatment group underwent a surgery, and those in the control group had only an incision made. For each subject it was recorded whether he or she experienced a reduction in migraine pain (a "success"). The results are tabulated here:

	Successes	Failures	Total
Treatment	41	8	49
Control	15	11	26
Total	56	19	75

Use the data to test the hypotheses

$$H_0\text{: } \textit{the treatment has no effect} \text{ versus } H_1\text{: } \textit{the treatment has an effect}$$

at the $\alpha = 0.05$ significance level.
Answer: The table of expected values under the null hypothesis is

	Successes	Failures	Total
Treatment	36.59	12.41	49
Control	19.41	6.59	26
Total	56	19	75

The likelihood ratio test statistic is

$$2\left[41\log\left(\frac{41}{36.59}\right) + 15\log\left(\frac{15}{19.41}\right) + 8\log\left(\frac{8}{12.41}\right) + 11\log\left(\frac{11}{6.59}\right)\right] = 5.845761.$$

Comparing this to the critical value $\chi^2_{1,0.05} = 3.841459$ leads us to reject the null hypothesis. The p-value, moreover, is 0.01561462, which is the area under the χ^2_1 pdf to the right of the test statistic value 5.845761.

- **Warning:** The test introduced in this section is only asymptotically size-α, so for small sample sizes it may lead to higher rates of Type I error than the desired α. A rule of thumb is to use this test only if all of the expected counts are greater than or equal to 5. The test presented in the following section is an alternative which can be used when the sample sizes are small.

Fisher's exact test

- Another test called *Fisher's exact test* takes a different approach. In the setup of the previous section, the row totals n_1 and n_2 were fixed and the column totals $Y_1 + Y_2$ and $n_1 + n_2 - Y_1 - Y_2$ were random, as were the four entries, Y_1, $n_1 - Y_1$, Y_2, and $n_2 - Y_2$, in the middle of the table.

- The approach of Fisher's exact test is to condition upon the row and column totals and regard only the values in the middle of the table as random variables. Let R_1 and R_2 be the row totals and C_1 and C_2 be the column totals and let N be the sum of all entries in the table. Then the observed 2×2 table has the form

			Total
	X_{obs}		R_1
			R_2
Total	C_1	C_2	N,

where the value of X_{obs} determines all four entries in the table (since we have fixed row and column sums). Fisher's exact test is based on the idea that if there is no association between the row and column variable, then the value X_{obs} can be viewed as a realization of a Hypergeometric random variable: If we draw C_1 marbles without replacement from a bag of N marbles, R_1 of which are red, and let X equal the number of red marbles drawn, then we have

$$X \sim \text{Hypergeometric}(N, R_1, C_1).$$

The p-value of Fisher's exact test is computed as the sum of the probabilities of all 2×2 tables of which the probability under H_0 is less than or equal to that of the observed table. That is, we consider all tables which carry as much or more evidence against the null as the observed table and then sum together the probabilities of all these tables. Precisely, the p-value is computed as

$$p_{\text{Fisher}} = \sum_{x=\max\{0,C_1+R_1-N)\}}^{\min\{C_1,R_1\}} P(X = x) \cdot \mathbf{1}\left(P(X = x) \leq P(X = X_{\text{obs}})\right),$$

where, according to the pmf of the Hypergeometric(N, R_1, C_1) distribution, we have

$$P(X = x) = \frac{\binom{R_1}{x}\binom{N-R_1}{C_1-x}}{\binom{N}{C_1}}.$$

We reject the null hypothesis of no association if the p-value falls below the chosen significance level.

- Note that it does not matter how the table is oriented in terms of which are the rows and which are the columns. If we transpose the table, the p-value of Fisher's exact test does not change. This is because

$$P(X = x) = \frac{\binom{R_1}{x}\binom{N-R_1}{C_1-x}}{\binom{N}{C_1}} = \frac{\binom{C_1}{x}\binom{N-C_1}{R_1-x}}{\binom{N}{R_1}}$$

for all $x = \max\{0, C_1 + R_1 - N)\}, \ldots, \min\{C_1, R_1\}$.

- **Exercise:** Compute the p-value of Fisher's exact test for the data

	Successes	Failures	Total
Treatment	41	8	49
Control	15	11	26
Total	56	19	75

Answer: The following R code computes the p-value:

```
data <- matrix(c(41,8,15,11),2,2,byrow=TRUE)

Xobs <- data[1,1]
R1 <- data[1,1] + data[1,2]
C1 <- data[1,1] + data[2,1]
N <- sum(data)

obs.hyper.prob <- dhyper(Xobs, n = N - R1, m = R1, k = C1)
all.hyper.probs <- dhyper(max(0,C1-(N-R1)):min(C1,R1),n = N - R1, m = R1, k = C1)
pval <- sum(all.hyper.probs[all.hyper.probs<=obs.hyper.prob])
```

The p-value is $p_{\text{Fisher}} = 0.02409327$, which is close to the p-value of asymptotic likelihood ratio test. Note that R uses a different parameterization for the hypergeometric distribution; type `?dhyper` to learn more.

- The advantage of Fisher's exact test is that it is not based on any asymptotic result, so it can be used when the sample size is small; there is no requirement that the expected counts exceed some minimum value. This is where the "exact" part of its name comes from.
- It may not always be desirable to condition on the row and column totals; some unconditional tests have been developed in order to avoid this, but Fisher's test has become a default choice for many practitioners.

More than two proportions and/or samples

- **Example:** A study of the eye and hair color of 6800 people resulted in the following data:

	Brown	Black	Fair	Red	Total
Brown	438	288	115	16	857
Grey or Green	1387	746	946	53	3132
Blue	807	189	1768	47	2811
Total	2632	1223	2829	116	6800

Suppose we are interested in testing whether there is an association between eye color and hair color.

- **Multinoulli trial:** A statistical experiment with $M \geq 1$ possible outcomes which occur with the probabilities $p_1, \ldots, p_M$, where $\sum_{j=1}^{M} p_j = 1$, is sometimes called a *multinoulli trial*. This idea extends the *Bernoulli trial*, in which there are two outcomes which have the probabilities p and $1 - p$.

- **Multinoulli random vector:** Consider a Multinoulli trial with M possible outcomes with associated probabilities $p_1, \ldots, p_M$, and let the outcome be encoded in an $M \times 1$ random vector $X = (X_1, \ldots, X_M)^T$ such that

$$X_j = \begin{cases} 1 & \text{if outcome } j \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } j = 1, \ldots, M.$$

  That is, if outcome j occurs, the jth entry of X is equal to 1 and the remaining entries are equal to 0 for all $j = 1, \ldots, M$. Then the pmf of the random vector X is given by

$$P((X_1, \ldots, X_M)^T = (x_1, \ldots, x_M)^T) = p_1^{x_1} \cdot \cdots \cdot p_M^{x_M}$$

  for $(x_1, \ldots, x_M)^T \in \{0, 1\}^M$ such that $\sum_{j=1}^M x_j = 1$, and we write $X \sim \text{Multinoulli}(p_1, \ldots, p_M)$.

- Note that the Bernoulli(p) distribution is the same as the Multinoulli$(p, 1-p)$ distribution, for which $M = 2$.

- **Multinomial distribution:** Let $X_1, \ldots, X_n$ be independent Multinoulli$(p_1, \ldots, p_M)$ random vectors and define $Y = \sum_{i=1}^n X_i$. Then $Y = (Y_1, \ldots, Y_M)^T$ is called a *Multinomial random vector* and has the pmf

$$P((Y_1, \ldots, Y_M)^T = (y_1, \ldots, y_M)^T) = \left(\frac{n!}{y_1! \cdots y_M!} \right) p_1^{y_1} \cdots p_M^{y_M}$$

  for $(y_1, \ldots, y_M) \in \{0, 1, \ldots, n\}^M$ such that $\sum_{j=1}^M y_j = n$, and we write $Y \sim \text{Multinomial}(p_1, \ldots, p_M, n)$.

- Note that the Binomial(n, p) distribution is the same as the Multinomial$(p, 1-p, n)$ distribution, for which $M = 2$.

- **Exercise:** Find expressions for the maximum likelihood estimators of $p_1, \ldots, p_M$ based on $Y \sim \text{Multinomial}(p_1, \ldots, p_M, n)$.

  Answer: The likelihood function is

$$L(p_1, \ldots, p_M; Y_1, \ldots, Y_M) = \left(\frac{n!}{Y_1! \cdots Y_M!} \right) p_1^{Y_1} \cdots p_M^{Y_M},$$

  and the log-likelihood is

$$\ell(p_1, \ldots, p_M; Y_1, \ldots, Y_M) = \log \left(\frac{n!}{Y_1! \cdots Y_M!} \right) + Y_1 \log p_1 + \cdots + Y_M \log p_M.$$

  If we take the partial derivative of the log-likelihood with respect to any of $p_1, \ldots, p_K$ and set it equal to zero, we will find that we cannot solve for the maximum likelihood estimator. This is because we must take into account the constraint that $p_1 + \cdots + p_M = 1$. That is, we need to solve the constrained optimization problem

$$\begin{aligned} &\text{maximize} && \log \left(\frac{n!}{Y_1! \cdots Y_M!} \right) + Y_1 \log p_1 + \cdots + Y_M \log p_M \\ &\text{subject to} && p_1 + \cdots + p_M = 1. \end{aligned}$$

We can solve this constrained optimization problem by writing down the Lagrangian

$$\mathcal{L}(p_1, \ldots, p_M, \lambda) = \log\left(\frac{n!}{Y_1! \cdots Y_M!}\right) + Y_1 \log p_1 + \cdots + Y_M \log p_M + \lambda(1 - p_1 - \cdots - p_M)$$

and solving the system of $M + 1$ equations

$$\begin{aligned}
\frac{\partial}{\partial p_1}\mathcal{L}(p_1, \ldots, p_M, \lambda) &= \frac{Y_1}{p_1} - \lambda = 0 \\
&\vdots \\
\frac{\partial}{\partial p_M}\mathcal{L}(p_1, \ldots, p_M, \lambda) &= \frac{Y_M}{p_M} - \lambda = 0 \\
\frac{\partial}{\partial \lambda}\mathcal{L}(p_1, \ldots, p_M, \lambda) &= 1 - p_1 - \cdots - p_M = 0.
\end{aligned}$$

From the first M equations, p_j satisfies $p_j = Y_j/\lambda$, for each $j = 1, \ldots, M$. Plugging this into the last equation gives

$$1 - Y_1/\lambda - \cdots - Y_M/\lambda = 0 \iff \lambda = \sum_{j=1}^{M} Y_j = n.$$

Plugging this value of λ back into each of the first M equations gives $p_j = Y_j/n$ for $j = 1, \ldots, M$, so the maximum likelihood estimators of $p_1, \ldots, p_M$ are given by

$$\hat{p}_j = Y_j/n, \quad \text{for} \quad j = 1, \ldots, M. \tag{2}$$

- **Dimension of multinomial parameter space:** Even though the Multinomial$(p_1, \ldots, p_M, n)$ distribution involves M parameters, $p_1, \ldots, p_M$, the dimension of the parameter space is not equal to M because of the requirement that the probabilities sum to 1. The parameter space is the set

$$\{p_1, \ldots, p_M : p_j \geq 0 \text{ for all } j = 1, \ldots, M \text{ and } p_1 + \cdots + p_M = 1\},$$

which actually has dimension equal to $M - 1$. To see this, note that if $p_1, \ldots, p_{M-1}$ are given, then one can compute p_M by

$$p_M = 1 - p_1 - \cdots - p_{M-1},$$

so that the set of parameters $p_1, \ldots, p_M$ is determined by $M - 1$ values.

- **Equal-probabilities hypotheses:** Consider observing a collection of Multinomial random variables $Y_1, \ldots, Y_K$, where

$$Y_k = (Y_{k1}, \ldots, Y_{kM})^T \sim \text{Multinomial}(p_{k1}, \ldots, p_{kM}, n_k), \quad \text{for } k = 1, \ldots, K,$$

and suppose we are interested in testing whether the outcome probabilities of the K Multinomial distributions are all the same. That is, suppose we wish to test

$$\begin{aligned}
&H_0\text{: } (p_{11}, \ldots, p_{1M}) = \cdots = (p_{K1}, \ldots, p_{KM}) \\
&\text{versus } H_1\text{: } (p_{j1}, \ldots, p_{jM}) \neq (p_{i1}, \ldots, p_{iM}) \text{ for some } i \neq j.
\end{aligned} \tag{3}$$

Note that for $K = 2$ and $M = 2$, this problem reduces to that in (1).

- **Exercise:** Derive the likelihood ratio test for the equal-probabilities hypotheses in (3).

  Answer: The likelihood function is

$$L(p_{11},\dots,p_{1M},\dots,p_{K1},\dots,p_{KM};Y_{11},\dots,Y_{1M},\dots,Y_{K1},\dots,Y_{KM}) = \prod_{k=1}^{K}\left(\frac{n_k!}{Y_{k1}!\cdots Y_{kM}!}\right)p_{k1}^{Y_{k1}}\cdots p_{kM}^{Y_{kM}},$$

  and the log-likelihood is

$$\ell(p_{11},\dots,p_{1M},\dots,p_{K1},\dots,p_{KM};Y_{11},\dots,Y_{1M},\dots,Y_{K1},\dots,Y_{KM}) = \sum_{k=1}^{K}\log\left(\frac{n_k!}{Y_{k1}!\cdots Y_{kM}!}\right)+\sum_{k=1}^{K}Y_{k1}\log p_{k1}+\cdots+\sum_{k=1}^{K}Y_{kM}\log p_{kM}.$$

  The maximum likelihood estimator of p_{kj} is given by $\hat{p}_{kj} = Y_{kj}/n_k$, for $k = 1,\dots,K$, $j = 1,\dots,M$. Under H_0, each of the K multinomial distributions has the same probabilities, which we may denote by $p_{01},\dots,p_{0M}$. Our estimators of $p_{01},\dots,p_{0M}$ are given by

$$\begin{aligned}(\hat{p}_{01},\dots,\hat{p}_{0M}) &= \underset{p_{01},\dots,p_{0M}}{\operatorname{argmax}} \quad \ell(p_{01},\dots,p_{0M},\dots,p_{01},\dots,p_{0M};Y_{11},\dots,Y_{1M},\dots,Y_{K1},\dots,Y_{KM})\\ &= \underset{p_{01},\dots,p_{0M}}{\operatorname{argmax}} \quad \sum_{k=1}^{K}\log\left(\frac{n_k!}{Y_{k1}!\cdots Y_{kM}!}\right)+\sum_{k=1}^{K}Y_{k1}\log p_{01}+\cdots+\sum_{k=1}^{K}Y_{kM}\log p_{0M}\\ &= \left(\sum_{k=1}^{K}Y_{k1}/n,\dots,\sum_{k=1}^{K}Y_{kM}/n\right),\end{aligned}$$

  where $n = n_1+\cdots+n_K$, which we can get by using the Lagrangian as we did to get the maximum likelihood estimators in (2).

  The likelihood ratio is

$$\begin{aligned}\mathrm{LR}(Y_{11},\dots,Y_{1M},\dots,Y_{K1},\dots,Y_{KM}) &= \frac{\prod_{k=1}^{K}\left(\frac{n_k!}{Y_{k1}!\cdots Y_{kM}!}\right)\hat{p}_{01}^{Y_{k1}}\cdots\hat{p}_{0M}^{Y_{kM}}}{\prod_{k=1}^{K}\left(\frac{n_k!}{Y_{k1}!\cdots Y_{kM}!}\right)\hat{p}_{k1}^{Y_{k1}}\cdots\hat{p}_{kM}^{Y_{kM}}}\\ &= \frac{\hat{p}_{01}^{\sum_{k=1}^{K}Y_{k1}}\cdots\hat{p}_{0M}^{\sum_{k=1}^{K}Y_{kM}}}{\prod_{k=1}^{K}\hat{p}_{k1}^{Y_{k1}}\cdots\hat{p}_{kM}^{Y_{kM}}}.\end{aligned}$$

The asymptotic likelihood ratio test is thus based on the test statistic

$$\begin{aligned}
&-2\log \mathrm{LR}(Y_{11},\ldots,Y_{1M},\ldots,Y_{K1},\ldots,Y_{KM}) \\
&\quad = -2\left[\sum_{k=1}^{K} Y_{k1}\log\hat{p}_{01} + \cdots + \sum_{k=1}^{K} Y_{kM}\log\hat{p}_{0M} - \sum_{k=1}^{K} Y_{k1}\log\hat{p}_{k1} - \cdots - \sum_{k=1}^{K} Y_{kM}\log\hat{p}_{kM}\right] \\
&\quad = -2\left[\sum_{k=1}^{K} Y_{k1}\log\left(\frac{\hat{p}_{01}}{\hat{p}_{k1}}\right) + \cdots + \sum_{k=1}^{K} Y_{kM}\log\left(\frac{\hat{p}_{0M}}{\hat{p}_{kM}}\right)\right] \\
&\quad = 2\left[\sum_{k=1}^{K} Y_{k1}\log\left(\frac{\hat{p}_{k1}}{\hat{p}_{01}}\right) + \cdots + \sum_{k=1}^{K} Y_{kM}\log\left(\frac{\hat{p}_{kM}}{\hat{p}_{0M}}\right)\right] \\
&\quad = 2\sum_{j=1}^{M}\sum_{k=1}^{K} Y_{kj}\log\frac{\hat{p}_{kj}}{\hat{p}_{0j}} \\
&\quad = 2\sum_{j=1}^{M}\sum_{k=1}^{K} Y_{kj}\log\left(\frac{Y_{kj}}{n_k\hat{p}_{0j}}\right).
\end{aligned}$$

The null space

$$\{p_{01},\ldots,p_{0M} : p_{0j} \geq 0 \text{ for all } j = 1,\ldots,M \text{ and } p_{01}+\cdots+p_{0M} = 1\},$$

has dimension $M-1$ and the entire parameter space

$$\left\{p_{11},\ldots,p_{1M},\ldots,p_{K1},\ldots,p_{KM} : \begin{array}{l} p_{kj} \geq 0 \text{ for all } j = 1,\ldots,M, k = 1,\ldots,K, \text{ and} \\ p_{k1}+\cdots+p_{kM} = 1, \text{ for all } k = 1,\ldots,K \end{array}\right\},$$

has dimension $K(M-1)$. This tells us that under the null hypothesis, the asymptotic likelihood ratio test statistic converges in distribution to a chi-squared distribution with degrees of freedom equal to $K(M-1)-(M-1) = (K-1)(M-1)$. So the size-$\alpha$ asymptotic likelihood ratio test is

$$\text{Reject } H_0 \text{ iff} \quad 2\sum_{j=1}^{M}\sum_{k=1}^{K} Y_{kj}\log\left(\frac{Y_{kj}}{n_k\hat{p}_{0j}}\right) > \chi^2_{(K-1)(M-1),\alpha}.$$

- **Contingency table notation:** Continuing the previous exercise, suppose we put the observed data in a table as follows:

	Outcome 1	Outcome 2	...	Outcome M	Total
Sample 1	Y_{11}	Y_{12}	...	Y_{1M}	n_1
Sample 2	Y_{21}	Y_{22}	...	Y_{2M}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Sample K	Y_{K1}	Y_{K2}	...	Y_{KM}	n_K
Total	$Y_{\cdot 1}$	$Y_{\cdot 2}$	...	$Y_{\cdot M}$	n

In the above $Y_{\cdot j} = \sum_{k=1}^{K} Y_{kj}$, for $j = 1, \ldots, M$.

Now suppose we make a second table like the one above but with the observed counts replaced by the corresponding expected counts under the null hypothesis, basing the expectations on the estimated probabilities $\hat{p}_{01}, \ldots, \hat{p}_{0M}$. This would look like

	Outcome 1	Outcome 2	$\ldots$	Outcome M	Total
Sample 1	$n_1\hat{p}_{01}$	$n_1\hat{p}_{02}$	$\ldots$	$n_1\hat{p}_{0M}$	n_1
Sample 2	$n_2\hat{p}_{01}$	$n_2\hat{p}_{02}$	$\ldots$	$n_2\hat{p}_{0M}$	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Sample K	$n_K\hat{p}_{01}$	$n_K\hat{p}_{02}$	$\ldots$	$n_K\hat{p}_{0M}$	n_K
Total	$Y_{\cdot 1}$	$Y_{\cdot 2}$	$\ldots$	$Y_{\cdot M}$	n

If we denote the observed counts in the first table by O_{kj} and the corresponding expected counts in the second table by E_{kj}, for $k = 1, \ldots, K$ and $j = 1, \ldots, M$, then we can express the asymptotic likelihood ratio test statistic for testing the hypotheses in (3) as

$$-2 \log \mathrm{LR}(Y_{11}, \ldots, Y_{1M}, \ldots, Y_{K1}, \ldots, Y_{KM}) = 2 \sum_{j=1}^{M} \sum_{k=1}^{K} O_{kj} \log\left(\frac{O_{kj}}{E_{kj}}\right).$$

The size-α asymptotic likelihood ratio test for testing the hypotheses in (3) is thus

$$\text{Reject } H_0 \text{ iff} \quad 2 \sum_{j=1}^{M} \sum_{k=1}^{K} O_{kj} \log\left(\frac{O_{kj}}{E_{kj}}\right) > \chi^2_{(K-1)(M-1),\alpha}. \tag{4}$$

- **Exercise:** Use the asymptotic likelihood ratio test to test whether there is an association between hair and eye color at the $\alpha = 0.01$ significance level based on the data in the following table:

	Brown	Black	Fair	Red	Total
Brown	438	288	115	16	857
Grey or Green	1387	746	946	53	3132
Blue	807	189	1768	47	2811
Total	2632	1223	2829	116	6800

Answer: The table of expected counts under the null hypothesis of no association is

	Brown	Black	Fair	Red	Total
Brown	331.7094	154.1340	356.5372	14.61941	857
Grey or Green	1212.2682	563.2994	1303.0041	53.42824	3132
Blue	1088.0224	505.5666	1169.4587	47.95235	2811
Total	2632	1223	2829	116	6800

And the value of the test statistic for the asymptotic likelihood ratio test can be computed in R as follows:

```
O <- matrix(c(438,288,115,16,1387,746,946,53,807,189,1768,47),3,4,byrow=TRUE)
E <- apply(O,1,sum) %*% t(apply(O,2,sum))/sum(O)
T.stat <- 2*sum(O*log(O/E))
```

The value is 1137.606. The degrees of freedom of the limiting chi-squared distribution of the test statistic under the null hypothesis is $(3-1)(4-1) = 6$, so that the critical value is $\chi^2_{6,0.01} = 16.81189$. We therefore reject the null hypothesis.

- There exist modifications to Fisher's exact test for tables with dimensions greater than 2×2, but we will not discuss them in this course.

Connection to Pearson's chi-squared test

- **Pearson's chi-squared test:** A classical test of the hypotheses in (3) is Pearson's chi-squared test, which is given by

$$\text{Reject } H_0 \text{ iff } \quad \sum_{j=1}^{M}\sum_{k=1}^{K} \frac{(O_{kj} - E_{kj})^2}{E_{kj}} > \chi^2_{(K-1)(M-1),\alpha}.$$

The invention of this test predates the development of the likelihood ratio approach to building tests of hypotheses, so we don't really have a principled way to derive it. One can show, however, that Pearson's test statistic is very close to that of the asymptotic likelihood ratio test (note moreover that the rejection region is the same).

- **Closeness of test statistics of LRT and Pearson's test:** We can show that the test statistic of Pearson's test is close to that of the likelihood ratio test by using a Taylor expansion of the natural logarithm:

$$\log(1 + x) \approx x - \frac{1}{2}x^2.$$

Additionally noting that

$$O_{kj} = E_{kj}\left(1 + \frac{O_{kj} - E_{kj}}{E_{kj}}\right) \quad \text{and} \quad \frac{O_{kj}}{E_{kj}} = 1 + \frac{O_{kj} - E_{kj}}{E_{kj}},$$

we write

$$
\begin{aligned}
2\sum_{j=1}^{M}\sum_{k=1}^{K} O_{kj}\log\left(\frac{O_{kj}}{E_{kj}}\right) &= 2\sum_{j=1}^{M}\sum_{k=1}^{K} E_{kj}\left(1+\frac{O_{kj}-E_{kj}}{E_{kj}}\right)\log\left(1+\frac{O_{kj}-E_{kj}}{E_{kj}}\right) \\
&\approx 2\sum_{j=1}^{M}\sum_{k=1}^{K} E_{kj}\left(1+\frac{O_{kj}-E_{kj}}{E_{kj}}\right)\left[\frac{O_{kj}-E_{kj}}{E_{kj}}-\frac{1}{2}\left(\frac{O_{kj}-E_{kj}}{E_{kj}}\right)^2\right] \\
&= 2\sum_{j=1}^{M}\sum_{k=1}^{K} E_{kj}\left[\frac{O_{kj}-E_{kj}}{E_{kj}}+\frac{1}{2}\left(\frac{O_{kj}-E_{kj}}{E_{kj}}\right)^2\underbrace{-\frac{1}{2}\left(\frac{O_{kj}-E_{kj}}{E_{kj}}\right)^3}_{\text{discard}}\right] \\
&\approx \underbrace{2\sum_{j=1}^{M}\sum_{k=1}^{K} E_{kj}\left(\frac{O_{kj}-E_{kj}}{E_{kj}}\right)}_{=0}+\sum_{j=1}^{M}\sum_{k=1}^{K} E_{kj}\left(\frac{O_{kj}-E_{kj}}{E_{kj}}\right)^2 \\
&= \sum_{j=1}^{M}\sum_{k=1}^{K}\frac{(O_{kj}-E_{kj})^2}{E_{kj}},
\end{aligned}
$$

where to obtain the last equality we have used the fact that

$$
\sum_{j=1}^{M}\sum_{k=1}^{K}(O_{kj}-E_{kj}) = \sum_{j=1}^{M}\sum_{k=1}^{K}(Y_{kj}-n_k\hat{p}_{0j}) = \sum_{j=1}^{M}\sum_{k=1}^{K}(Y_{kj}-n_k(Y_{\cdot j}/n)) = n-n=0.
$$

- **Exercise:** Compute the test statistic for Pearson's test on the migraine surgery data:

	Successes	Failures	Total
Treatment	41	8	49
Control	15	11	26
Total	56	19	75

Answer: Pearson's test statistic is

$$
\frac{(41-36.59)^2}{36.59}+\frac{(15-19.41)^2}{19.41}+\frac{(8-12.41)^2}{12.41}+\frac{(11-6.59)^2}{6.59} = 6.051762.
$$

This is fairly close to the value of the likelihood ratio test statistic, which was 5.845761. For larger sample sizes Pearson's test statistic and the likelihood ratio test statistic will be closer.

- **Equivalence of Pearson's test to other classical test in 2-by-2 case:** In the two-sample setup with

$$Y_1 \sim \text{Binomial}(n_1, p_1)$$
$$Y_2 \sim \text{Binomial}(n_2, p_2),$$

an asymptotic size-α test of H_0: $p_1 = p_2$ versus H_1: $p_1 \neq p_2$ is

$$\text{Reject } H_0 \text{ iff } \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)(1/n_1 + 1/n_2)}} > z_{\alpha/2},$$

where $\hat{p}_1 = Y_1/n_1$, $\hat{p}_2 = Y_2/n_2$ and $\hat{p}_0 = (Y_1 + Y_2)/(n_1 + n_2)$. It turns out that the square of the test statistic in the above test is equal to the test statistic of Pearson's test; therefore the p-values associated with the tests will be equal.