Contingency Tables (Continued)

Coco Kusiak & Liz Austin, based on Agresti Chapters 2&3 and other related Bayesian content

9/12/2017

General Notation for Contingency Tables:

IJ Tables represent 2 categorical variables

	Y_1	Y_2	...	Y_J	
X_1	n_11	n_12	...	...	n_1+
X_2	n_21	...	...	...	n_2+
...	...	...	n_ij	...	...
X_I	...	...	...	n_IJ	...
	n_+1	n_+2	...	...	n

n_{ij} = counts with X = i and Y = j

π_{ij} = population parameter representing the true probability of being in the ij^{th} cell (when X = i and Y = j)

$$= \text{joint probability of X and Y}$$
$$= Pr(X = i, Y = j)$$
$$\text{with}\{\pi_{ij} : i = 1, ..., I; j = 1, ..., J\}$$

P_{ij} = sample/observed probabilities

$n = \sum_{i,j} n_{ij}$

π_{i+} = marginal distribution across rows = $\sum_j \pi_{ij}$

π_{+j} = marginal distribution across columns = $\sum_i \pi_{ij}$

$\pi_{j|i}$ = conditional probability of j given i = $\frac{\pi_{ij}}{\pi_{i+}} = Pr(Y = j|X = i)$

Note: This is equivalent to E[Y|X] in regression.

Sampling Types:

1. Poisson

- The overall **n** is not fixed
- There is generally a time interval implied
- Example: A prospective longitudinal cohort study about developing a disease

	Disease
X1	n_1
X2	n_2
X3	n_3

n_1 = total # of people in catergory X1 with the disease

	Number of Accidents	Number of Fatal Accidents
AM	n_11	n_12
PM	n_21	n_22

- Example: # of accidents at an intersection over a year

$$n_{12} = \text{total \# of fatal accidents which occurred in the morning}$$

2. Multinomial

a. with fixed **n**
  - Example: A cohort study with 3 categories of socioeconomic status and a binary outcome of illness (a fixed # of people are enrolled in the study)

	Sick	Not Sick	Total
SE_1	n_11	n_12	n_1+
SE_2	n_21	. . .	. . .
SE_3	. . .	. . .	. . .
Total	n_+1	. . .	2000

b. **row or column totals** are fixed
  - Example: A case-control study

	Case	Control
SE_1	. . .	. . .
SE_2	. . .	. . .
SE_3	. . .	. . .
Total	1000	1000

χ^2 Tests

$$\text{Test Statistic} = \frac{\sum (O-E)^2}{E} \text{over all cells/counts}$$

with E = expected # of counts under the null hypothesis

- We reject H_0 if the test statistics is "large"
  - A larger TS means larger deviations from expected counts
- Test is 2-sided by virtue of $(...)^2$
- Compare to a $(1-\alpha)$%ile of the χ^2_{df} distribution

Goodness of Fit χ^2 Test

$$H_0 : P_1 = P_{0_1}, P_2 = P_{0_2}, ..., P_k = P_{0_k}$$

This answers the question: Does a set of counts follow a specified distribution?

n = total # of observations
$E_i = P_{0_i} \cdot n$
df $= k - 1$

Note: $P_1 = P_2 = ... = P_k$ is a special case

Birth Order and Gender

We have data on 1,000 2-child families. It is typically thought that birth order/gender of two offspring from the same parents are i.i.d. Bernoulli(0.5). So, we can fill in the expected values as 250 for each group.

$$H_0 : P_1 = P_2 = P_3 = P_4 = 0.25$$

$$H_a : \text{at least 1 } P_i \neq 0.25$$

First Child	M		F		
Second Child	M	F	M	F	Totals
Count	218	227	278	277	1000
Expected Value	250	250	250	250	1000

$$TS = \frac{\sum_{k=1}^{4} (n_{obs} - n_{exp})^2}{n_{exp}}$$

$$n = 1000,\ k = 4, df = \text{k-1} = 3$$

$$\alpha = 0.05$$

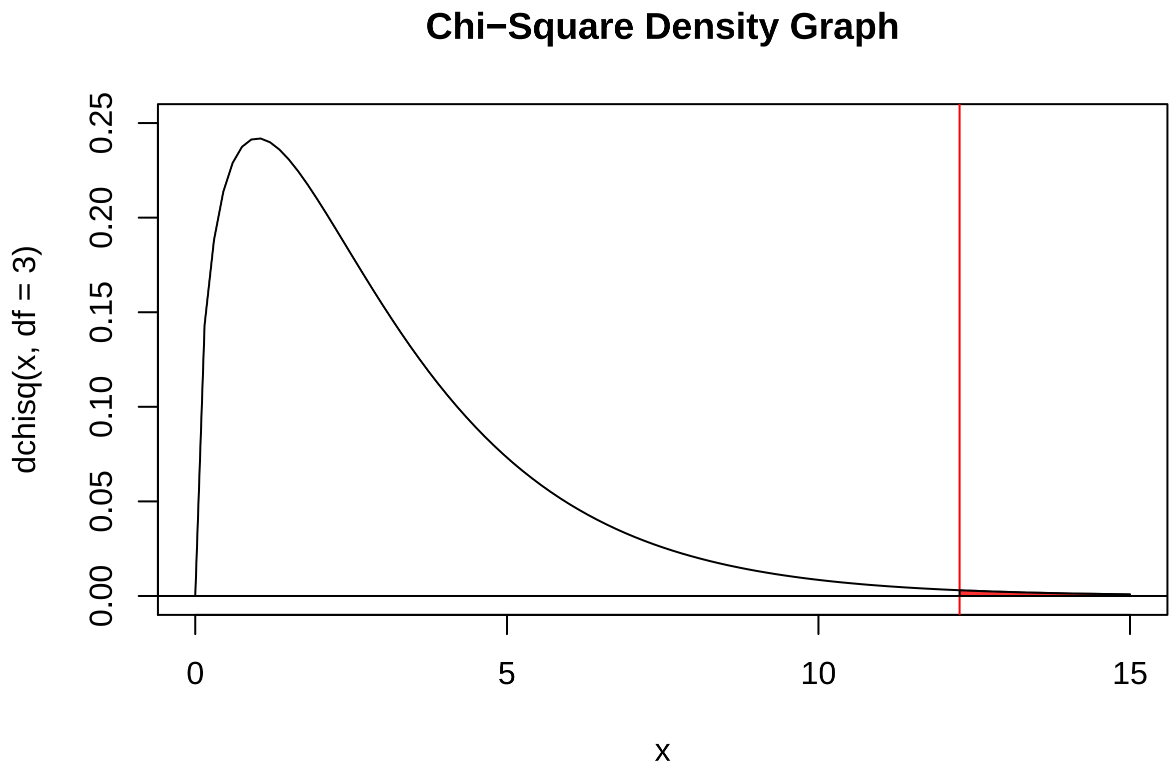
```
n_obs <- c(218, 227, 278, 277)
n_exp <- c(250, 250, 250, 250)
ts <- sum((n_obs-n_exp)^2/n_exp)
ts
```

```
## [1] 12.264
```

```
pval <- 1 - pchisq(ts, df = 3)
pval
```

```
## [1] 0.006531413
```

With a p-value of 0.0065, we reject the null hypothesis at a significance level of 0.05. We have evidence to suggest that at least 1 $P_k \neq 0.25$.



Test of Independence

Definition: Two variables are independent if $\forall i \in \{1, ..., I\}, j \in \{1, ..., J\}$, $\pi_{ij} = \pi_{i+}\pi_{+j}$.

This is equivalent to: $Pr(X = i, Y = j) = Pr(X = i) \cdot Pr(Y = j)$.

If true, this implies $\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j}$.

For χ^2 Test:

$$H_0 : \pi_{ij} = \pi_{i+} \cdot \pi_{+j} \ \forall_{i,j}$$

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n},\ \hat{\pi}_{+j} = \frac{n_{+j}}{n}$$

$$E_{ij} = \hat{\mu}_{ij} = n \cdot \hat{\pi}_{i+} \cdot \hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}$$

$$df = (\#rows - 1)(\#columns - 1)$$

Book suggestion: Expand a small table and include the n_{ij}, E_{ij}, and the $(O - E)_{ij}$ for each cell. This may allow you to see patterns and information in the data beyond just the p-value.

Bayesian Multinomial

The likelihood of Y is a vector of counts with the # of observations for each category/outcome j.

$$P(Y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_i} \text{ where } \sum_{j=1}^{k} \theta_j = 1$$

In this case, the conjugate prior distribution is a multivariate generalization of Beta called the Dirichlet Distribution.

$$P(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \text{ with } \theta_j \geq 0 \text{ with } \sum_{j=1}^{k} \theta_j = 1$$

This implies the posterior follows the Dirichlet Distribution:

$$P(\theta|Y) \sim Dirichlet(\alpha_1 + y_1,\, \alpha_2 + y_2 ...,\, \alpha_k + y_k)$$

Choices for priors:

1. $\alpha_j = 1 \; \forall \; j$
   - this distribution assigns equal density to any vector θ such that $\sum \theta_j = 1$
2. $\alpha_j = 0 \; \forall \; j$
   - this is an improper prior, but it is uniform on $log(\theta_j)$
   - as long as $y_j > 0 \; \forall \; j$, then there are no problems with your posterior