Handout 5

Two-way Contingency Tables.

A table with I rows and J columns has a total of IJ cells. When a table contains the counts for these IJ cells, it is called a two-way contingency table. Consider the following example from Handout 1.

Example 1. A random sample of 250 adults is taken in a state, and each person's political ideology and opinion on death penalty are noted. Summary of the counts is given below.

	Supports Death Penalty		Row Total
Political Ideology	Yes	No	
Liberal	37	76	113
Conservative	86	51	137
Column Total	123	127	250

Note that the table above is obtained by interchanging the rows with columns in Example 6 in Handout 1. This is a 2-way contingency table with $I = 2$ rows and $J = 2$ columns with a total of $IJ = 4$ cells. From this data, an estimate of the proportion of individuals in the state who support death penalty is $123/250 = 0.492$. An estimate of the proportion who are opposed to the death penalty and are conservatives is $51/250 = 0.204$.

Probabilistic Concepts

In order to facilitate further discussion, we need to recall a few probabilistic concepts which we will do using the following example.

Example 2. For the purpose of discussion here, assume that the population proportions in a certain state are known and they are given below.

	Supports Death Penalty		Row Total
Political Ideology	Yes	No	
Liberal	$\pi_{11} = 0.12$	$\pi_{12} = 0.28$	$\pi_{1+} = 0.4$
Conservative	$\pi_{21} = 0.36$	$\pi_{22} = 0.24$	$\pi_{2+} = 0.6$
Column Total	$\pi_{+1} = 0.48$	$\pi_{+2} = 0.52$	1

.Note that the proportion of the population who are liberal and support death penalty is π_{11}, the proportion of the population that are conservative and support death penalty is π_{21} etc. For a person selected at random, consider two categorical variables: ideology (X) and opinion on death penalty (Y). The levels for X are $1, \ldots, I$ and the levels for Y are $1, \ldots, J$. The joint probability that $X = i$ and $Y = j$ is π_{ij}, ie, $\pi_{ij} = P(X = i, Y = j)$. For instance, the probability that a person is liberal and supports death penalty is π_{11}, the probability that a person is conservative and supports death penalty is π_{21}.

We write the row totals as $\{\pi_{i+}\}$ and the column totals as $\{\pi_{+j}\}$. Therefore

$$P(X = 1) = P(X = 1, Y = 1) + P(X = 1, Y = 2) = \pi_{11} + \pi_{12} = \sum_j \pi_{1j} = \pi_{1+},$$

$$P(Y = 2) = P(X = 1, Y = 2) + P(X = 2, Y = 2) = \pi_{12} + \pi_{22} = \sum_i \pi_{i2} = \pi_{+2}.$$

In general

$$P(X = i) = \sum_{j=1}^{J} P(X = i, Y = j) = \sum_{j=1}^{J} \pi_{ij} = \pi_{i+}, \text{ and}$$
$$P(Y = j) = \sum_{i=1}^{I} P(X = i, Y = j) = \sum_{i=1}^{I} \pi_{ij} = \pi_{+j}.$$

Conditional probability

In Example 2, note that among the liberals, the proportion who support death penalty is $\pi_{11}/\pi_{1+} = \frac{0.12}{0.4} = 0.3$. Among the conservatives, the proportion which support death penalty is $\pi_{21}/\pi_{2+} = \frac{0..36}{0.6} = 0.6$. These can also be described as conditional probabilities. The conditional probabilities that $Y = 1$ given $X = 1$, and $Y = 1$ given $X = 2$ are

$$P(Y = 1|X = 1) = \frac{P(X = 1, Y = 1)}{P(X = 1)} = \frac{\pi_{11}}{\pi_{1+}} = \frac{0.12}{0.4} = 0.3,$$
$$P(Y = 1|X = 2) = \frac{P(X = 2, Y = 1)}{P(X = 2)} = \frac{\pi_{21}}{\pi_{2+}} = \frac{0.36}{0.6} = 0.6.$$

Among those who oppose the death penalty, the proportion who are liberal is given by $\frac{\pi_{12}}{\pi_{+2}} = \frac{0.28}{0.52} = 0.538$. So the conditional probability of $X = 1$ given $Y = 2$ is

$$P(X = 1|Y = 2) = \frac{P(X = 1, Y = 2)}{P(Y = 2)} = \frac{\pi_{12}}{\pi_{+2}} = \frac{0.28}{0.52} = 0.538.$$

Independence

Variables X and Y are said to be independent if $P(X = i, Y = j) = P(X = i)P(Y = j)$ for all i and j, ie, $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j. This is equivalent to the restatement: for each j, $P(Y = j|X = i) = \pi_{ij}/\pi_{i+}$ is the same for all i. If ideology and opinion on death penalty were independent, then the proportion of support for (or opposition to) death penalty would be the same for all ideological groups.

In Example 2, 30% of the liberals support death penalty, whereas 60% of the conservatives support it.Since these proportions are different, opinion on death penalty (X) and political ideology (Y) are not independent.

Estimation for joint multinomial (Example 1)

For the setup in Example 1, the vector of counts $(n_{11}, n_{12}, n_{21}, n_{22})$ is $multinomial(n; \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$. MLE for π_{ij}, π_{i+} and π_{+j} are

$$\hat{\pi}_{ij} = p_{ij} = n_{ij}/n, \ \hat{\pi}_{i+} = p_{i+} = n_{i+}/n, \text{ and } \ \hat{\pi}_{+j} = p_{+j} = n_{+j}/n.$$

Estimates for π_{ij}/π_{+j} and π_{ij}/π_{i+} are

$$\hat{\pi}_{ij}/\hat{\pi}_{+j} = n_{ij}/n_{+j}, \text{ and } \hat{\pi}_{ij}/\hat{\pi}_{i+} = n_{ij}/n_{i+}.$$

For the data given in Example 1, sample proportions are

	Supports Death Penalty		Row Total
Political Ideology	Yes	No	
Liberal	$\hat{\pi}_{11} = 0.148$	$\hat{\pi}_{12} = 0.304$	$\hat{\pi}_{1+} = 0.452$
Conservative	$\hat{\pi}_{21} = 0.344$	$\hat{\pi}_{22} = 0.204$	$\hat{\pi}_{2+} = 0.548$
Column Total	$\hat{\pi}_{+1} = 0.492$	$\hat{\pi}_{+2} = 0.508$	1

Proportion of the population who are conservative is $\hat{\pi}_{+2} = 0.548$. Proportion of the population who oppose the death penalty is $\hat{\pi}_{+2} = 508$. An estimate of the population proportion of 'No' among conservatives (π_{22}/π_{2+}) is

$$\hat{\pi}_{22}/\hat{\pi}_{+2} = n_{22}/n_{2+} = 51/137 = 0.3723..$$

Difference between joint multinomial sampling and independent multinomial samples.

In order to discuss the concepts consider the following example.

Example 3. Suppose a pollster took a random sample 125 liberals and independently took another random sample of 100 conservatives. Opinion of death penalty is recorded for all the sample individuals. The counts are given the table below.

	Supports Death Penalty		Total
Political Ideology	Yes	No	
Liberal	31	94	$n_{1+} = 125$
Conservative	65	35	$n_{2+} = 100$
Total	96	129	250

First note the difference in Examples 1 and 3. In Example 1, a random sample of 250 individuals are taken, and it was not known before taking the sample which individuals are liberal and which are conservative. Thus the row totals, ie, number of liberals n_{1+} and the number of conservatives n_{2+} can be considered random. In Example 3, the pollster decided to take two independent samples of sizes $n_{1+} = 125$ (liberals) and $n_{2+} = 100$ (conservatives). Thus the values of n_{1+} and n_{2+} are fixed in advance, and therefore they are not random.

In the setup of Example 1, the vector of counts $(n_{11}, n_{12}, n_{21}, n_{22})$ is $multinomial(n, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$.

In the setup of Example 3, we have (n_{11}, n_{12}) and (n_{21}, n_{22}) are independent multinomials. More specifically, $(n_{11}, n_{12}) \sim multinomial(n_{1+}; \pi_{11}/\pi_{1+}, \pi_{12}/\pi_{1+})$ and $(n_{21}, n_{22}) \sim multinomial(n_{2+}; \pi_{21}/\pi_{2+}, \pi_{22}/\pi_{2+})$.

In example 3, the sample proportion of 'Yes' among liberals is $n_{11}/n_{1+} = 0.248$, and this an estimate of the population proportion (π_{11}/π_{1+}) of 'Yes' among the liberals. Similarly, the sample proportion of 'No' among conservatives is $35/100 = 0.35$, and this is an estimate of the population proportion (π_{22}/π_{2+}) of 'No' among the conservatives

A few observations are important.

Remark 1.

(a) From the data in Example 3, we can estimate the population proportions of 'Yes' and 'No' among liberals (conditional probabilities) $\pi_{11}/\pi_{1+}, \pi_{12}/\pi_{1+}$ as n_{11}/n_{1+}, n_{12}/n_{1+} respectively. Thus the estimates of $\pi_{11}/\pi_{1+}, \pi_{12}/\pi_{1+}$ are $31/125 = 0.248$ and $94/125 = 0.752$ respectively. Similarly, we can estimate proportions of 'Yes' and 'No' among conservatives (conditional probabilities) $\pi_{21}/\pi_{2+}, \pi_{22}/\pi_{2+}$ as $n_{21}/n_{2+}, n_{22}/n_{2+}$ respectively, and these estimates are $65/100 = 0.65$ and $35/100 = 0.35$.

(b) In Example 1, we could estimate $\pi_{11}, \pi_{12}, \pi_{21}$ and π_{22}. For instance an estimate of the proportion in the population who are liberal and support death penalty is $\hat{\pi}_{11} = 37/250 = 0.148$. However, from the data in Example 3, we cannot estimate π_{11}. **In general, we cannot estimate π_{ij}'s from the data in Example 3** (ie, from the data obtained using independent multinomial scheme).
(c) In Example 1, we estimate proportion of 'Yes' among liberals (π_{11}/π_{1+}) as $n_{11}/n_{1+} = 37/113 = 0.327$. From the data in Example 3, we can also estimate the proportion of 'Yes' among liberals as $n_{11}/n_{1+} = 0.248$. This tell us π_{1j}/π_{1+}, the proportion of people with opinion j on death penalty among the liberals, can be estimated the same way in Examples 1 and 3 (ie, the method of estimation is the same whether the sampling scheme is joint multinomial or independent multinomials). The same is true for π_{2j}/π_{2+}, the proportion of people with opinion j on death penalty among the conservatives,
(d) In example 1 we could estimate proportion of liberals and conservatives (π_{1+} and π_{2+}) in the state. However, from the data in Example 3, we cannot estimate these quantities. Similarly, we cannot estimate the proportion of 'Yes' and 'No' (π_{+1} and π_{+2}). From the data in Example 3, we cannot estimate the proportion of liberals (π_{11}/π_{+1}) among those who support the death penalty. In Example 3, we cannot estimate π_{i1}/π_{+1} nor can we estimate π_{i2}/π_{+2}.
(e) Summary of the observations in (a)-(d): based on the data in Example 3, we can only estimate the proportions of "yes' and 'No' among liberals $(\pi_{11}/\pi_{1+}, \pi_{12}/\pi_{1+})$ and among the conservatives $(\pi_{21}/\pi_{2+}, \pi_{22}/\pi_{2+})$.

Sensitivity and Specificity

Example 4. To investigate the effectiveness of a certain diagnostic method for breast cancer screening, 100 breast cancer patients and 100 cancer-free women were given the diagnostic test. The results are given the following table.

	Diagnosis		Total
Breast Cancer	Positive	Negative	
Yes	86	14	100
No	12	88	100
Total	98	102	200

First we note that this is case of two independent multinomial samples.
In the Health Sciences, following proportions are of interest:
Population proportion of 'Positive' (π_{11}/π_{1+}) among those whose have the cancer (sensitivity),
population proportion (π_{22}/π_{2+}) of 'Negative' among those who are cancer-free (specificity).
Ideally we would like both sensitivity and specificity to be as large as possible.
In Example 4, estimates of sensitivity and specificity are $86/100 = 0.86$ and $88/100 = 0.88$ respectively.

Difference in Proportions.

Consider Example 4, where we have independent samples of sizes $n_{1+} = 100$ and $n_{2+} = 100$ samples of cancer patients and cancer-free women. Denote the population proportions of 'Positive' among cancer patients (π_{11}/π_{1+}) and cancer-free women (π_{21}/π_{2+}) by π_1 and π_2 respectively. We would like to estimate the difference in proportions $\pi_1 - \pi_2$. Estimate of $\pi_1 - \pi_2$ is

$$\hat{\pi}_1 - \hat{\pi}_2 = n_{11}/n_{1+} - n_{21}/n_{2+} = 86/100 - 12/100 = 0.86 - 0.12 = 0.72.$$

An approximate $100(1-\alpha)$ confidence interval for $\pi_1 - \pi_2$ is given by $\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2} SE$, where area to the right of $z_{\alpha/2}$ under the standard normal curve is $\alpha/2$, and

$$SE = \sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_{1+} + \hat{\pi}_2(1-\hat{\pi}_2)/n_{2+}}.$$

From the Breast Cancer data, we have

$$SE = \sqrt{(0.86)(1-0.86)/100 + (0.12)(1-0.12)/100} = 0.0475.$$

Thus an approximate 95% confidence interval for $\pi_1 - \pi_2$ is

$$0.72 \pm (1.96)(0.0475), \text{ ie, } 0.72 \pm 0.0931, \text{ ie, } (0.627, 0.813).$$

Relative Risk

When both π_1 and π_2 are near 0, there is a useful descriptive measure known as the 'relative risk' and it is defined as π_1/π_2. Suppose that the proportion of incidence of lung cancer among smokers and non-smokers are $\pi_1 = 0.01$ and $\pi_2 = 0.001$ respectively. Note that π_1, π_2 and the difference $\pi_1 - \pi_2 = 0.009$ are all small. However, the relative risk $\pi_1/\pi_2 = 10$. This states that the incidence of lung cancer is 10 times higher for smokers than non-smokers.

Estimate of π_1/π_2 is given by $\hat{\pi}_1/\hat{\pi}_2$. For the Breast Cancer example estimate of the relative risk for 'Positive' is $0.86/0.12 = 7.167$.

If π_1 and π_2 are both close to 1, it may be useful to examine the ratio $(1-\pi_1)/(1-\pi_2)$.

Odd and Odds Ratio.

If we are tossing a fair coin with probability $\pi = 0.5$ of getting a tail, then the odds of getting a tail is $\pi/(1-\pi) = 1$. For a fair die, the chance of getting an ace is $\pi = 1/6$. Thus the odds of getting an ace is $\pi/(1-\pi) = 1/5 = 0.2$. It states that on the average we expect to get 5 times non-aces than aces.

Now consider Example 2 where population proportions are given. Let the population proportions of 'Yes' among liberals (π_{11}/π_{1+}) and conservatives (π_{21}/π_{2+}) be denoted by π_1 and π_2 respectively. Then the odds of 'Yes' among liberals and among conservatives are $\pi_1/(1-\pi_1)$ and $\pi_2/(1-\pi_2)$ respectively. The ratio of odds for 'Yes' for liberals to conservatives is defined to be

$$\theta = \frac{\text{odds for 'Yes' among liberals}}{\text{odds for 'Yes' among conservatives}} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}.$$

For Example 2, we have

$$\begin{aligned}
\pi_1 &= \pi_{11}/\pi_{1+} = 0.12/0.40 = 0.30, \\
\pi_2 &= \pi_{21}/\pi_{2+} = 0.36/0.60 = 0.60, \\
\theta &= \frac{\text{odds for 'Yes' among liberals}}{\text{odds for 'Yes' among conservatives}} \\
&= \frac{0.3/(1-0.3)}{(0.6)(1-0.6)} = \frac{0.4286}{1.5} = 0.2857.
\end{aligned}$$

Thus the odds for 'Yes' opinion for death penalty among liberals is about 0.29 times of the corresponding odds for conservatives.

Remark 2 (Properties of the odds ratio θ)
(a) $0 < \theta < \infty$.
(b) If $\theta = 1$, the categorical variables X and Y are independent. In Example 2, political ideology (X) and opinion on death penalty (Y) are not independent since $\theta \neq 1$.
(c) Further is the value of θ from 1, stronger is the dependence between X and Y.
(d) If one odds ratio is the reciprocal of the other, then the strength of the association is the same.
(e) Often one looks at the logarithm of the odds ratio $\log(\theta)$ instead of θ. Note that $\log(1) = 0$ and $\log(\theta) = -\log(1/\theta)$.
(f) The value of θ does not change if the rows are swapped with columns.
(g) Note that $\theta = (\text{relative risk})\frac{1-\pi_2}{1-\pi_1}$. Thus if π_1 and π_2 are close to zero, then the odds ratio is close to the relative risk.

Estimation of odds ratio

Consider the Breast Cancer data in Example 4. Estimates of $\pi_1 = \pi_{11}/\pi_{1+}$ and $\pi_2 = \pi_{21}/\pi_{2+}$ are $\hat{\pi}_1 - n_{11}/n_{1+}$ and $\hat{\pi}_2 = n_{21}/n_{2+}$ respectively. Therefore an estimate of the odds ratio os

$$\hat{\theta} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{(n_{11}/n_{1+})/(1-n_{11}/n_{1+})}{(n_{21}/n_{2+})/(1-n_{21}/n_{2+})} = \frac{n_{11}/(n_{1+}-n_{11})}{n_{21}/(n_{2+}-n_{21})}$$

In a 2×2 table, this formula become simple

$$\hat{\theta} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

since $n_{1+} = n_{11} = n_{12}$ and $n_{2+} - n_{21} = n_{22}$.

Thus for the Breast Cancer data in Example 4,

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{(86)(88)}{(12)(14)} = 45.0476, \;\; \log(\hat{\theta}) = 3.8077.$$

Thus the estimated odds for a positive diagnosis for cancer patients is about 45 times higher than the corresponding odds for cancer-free women.

Confidence intervals for odds ratio can be constructed using the normal approximation. Theoretical arguments show that the normal approximation is more appropriate for $\log(\hat{\theta})$ than for $\hat{\theta}$. For this reason it is customary to construct a confidence interval for $\log(\theta)$ first, and exponentiate it to get a confidence interval for θ. As discussed in Handout 4 in the context of Pearson's chi-square and LR tests, there is a rule of thumb for the normal approximation to be reasonable: all the counts $\{n_{ij}\}$ should be 5 or larger.

An approximate $100(1-\alpha)\%$ confidence interval for $\log(\theta)$ is given by $\log(\hat{\theta}) \pm z_{\alpha/2}SE$, where area to the right of $z_{\alpha/2}$ is $\alpha/2$ and

$$SE = \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}.$$

For the Breast Cancer data,

$$SE = \sqrt{1/86 + 1/14 + 1/12 + 1/88} = 0.4216.$$

So an approximate 95% confidence interval for $\log(\theta)$ is

$$\log(\hat{\theta}) \pm 1.96SE, \text{ ie, } 3.8077 \pm (1.96)(0.4216),$$
$$\text{ie, } 3.8077 \pm 0.8263, \text{ ie, } (2.9814, 4.6340)$$

Thus an approximate 95% confidence interval for θ is $(e^{2.9814}, e^{4.6340}) = (19.7154, 102.9249)$. Note that the confidence interval for θ does not include 1. Thus the the null hypothesis of independence $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$ can be rejected at a level of significance $\alpha = 0.05$