# Lecture 9: Chi-square Tests (contd.)

## Contingency Tables

Suppose a random experiment results in an outcome that can be classified by two different attributes. Assume that the first attribute is assigned to one and only one of $k$ mutually exclusive and exhaustive events $A_1, A_2, \ldots, A_k$, and the second attribute falls into one and only one of $h$ mutually exclusive and exhaustive events $B_1, B_2, \ldots, B_h$.

Let the probability of $A_i \cap B_j$ be defined by $p_{ij} = P(A_i \cap B_j)$, $i = 1, 2, \ldots, k$
$j = 1, 2, \ldots, h$

Let the random event be repeated $n$ times and $Y_{ij}$ denote the frequency of the event $A_i \cap B_j$. Since there are $kh$ such events as $A_i \cap B_j$, the random variable

$$Q_{kh-1} = \sum_{j=1}^{h} \sum_{i=1}^{k} (Y_{ij} - np_{ij})^2 / np_{ij}$$

has an approximate $\chi^2$-distribution with $kh-1$ d.f., provided $n$ is large.

Now to test $H_0: P(A_i \cap B_j) = P(A_i)P(B_j)$

i.e $A$ and $B$ are independent, we define

$$p_{i\cdot} = \sum_{j=1}^{h} p_{ij} = P(A_i) \quad \text{and} \quad p_{\cdot j} = \sum_{i=1}^{k} p_{ij} = P(B_j)$$

$\Rightarrow H_0: p_{ij} = p_{i\cdot}\, p_{\cdot j}$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, h$.

Since $p_{i\cdot}$ and $p_{\cdot j}$ are usually unknown, we can estimate them by $\hat{p}_{i\cdot} = \frac{y_{i\cdot}}{n}$, $y_{i\cdot} = \sum_{j=1}^{h} y_{ij}$

and $\hat{p}_{\cdot j} = \frac{y_{\cdot j}}{n}$, $y_{\cdot j} = \sum_{i=1}^{k} y_{ij}$ (observed freq$^{y}$ for $B_j$)

But $\sum_{1}^{k} p_{i\cdot} = \sum_{1}^{h} p_{\cdot j} = 1$ and so we have $(k-1)+(h-1) = k+h-2$ parameters to estimate. Consequently, the r.v.

$$Q = \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{\left[Y_{ij} - n(Y_{i\cdot}/n)(Y_{\cdot j}/n)\right]^2}{n(Y_{i\cdot}/n)(Y_{\cdot j}/n)}$$

$\overset{\circ}{\approx} \chi^2(kh-1-(k+h-2) = (k-1)(h-1))$ provided $H_0$ is true. We therefore reject $H_0$ if $Q$ exceeds $\chi^2_\alpha((k-1)(h-1))$ at level $\alpha$.

Example 1.

A random sample of 400 undergraduate students at the University of Iowa were classified according to the college in which the students were enrolled and according to gender. The results are summarized in the table below:

| | College | | | | | |
|---|---|---|---|---|---|---|
| Gender | Bus | Eng | LibAr | Nur | Pharm | Total |
| Male | 21(16.625) | 16(9.5) | 145(152) | 2(7.125) | 6(4.75) | 190 |
| Female | 14(18.375) | 4(10.5) | 175(168) | 13(7.875) | 4(5.25) | 210 |
| Total | 35 | 20 | 320 | 15 | 10 | 400 |

We wish to test if gender and college are independent i.e. $H_0: p_{ij} = p_{i.}\, p_{.j}$, $i=1,2$, $j=1,2,3,4,5$.

(the college a student enrols is independent of the gender of the student).

Under $H_0$, $\hat{p}_{1\cdot} = 190/400$ and $\hat{p}_{2\cdot} = 210/400$
$= 0.475$ $= 0.525$

$\hat{p}_{\cdot 1} = 35/400 = 0.0875$, $\hat{p}_{\cdot 2} = 0.05$, $\hat{p}_{\cdot 3} = 0.8$,

$\hat{p}_{\cdot 4} = 0.0375$, $\hat{p}_{\cdot 5} = 0.025$

The expected frequencies are computed using the formula $n\left(y_{i\cdot}/n\right)\left(y_{\cdot j}/n\right)$ and as follows:

$$400\left(\frac{190}{400}\right)\left(\frac{35}{400}\right) = 400(0.475)(0.0875)$$
$$= 16.625$$

Etcetera.

The computed chi-square statistic is

$$q = (21-16.625)^2/16.625 + \cdots + (4-5.25)^2/5.25$$
$$= 18.93 > 13.28 = \chi^2_{0.01}(4) \text{ when } \alpha = 0.01$$

Note: Contribution from Pharm/female is small, so we can ignore the fact that expected freq is less than 5.

## Exercise

Each of two comparable classes of 15 students responded to two different methods of instruction, giving the following scores on a standardized test:

class U: 91 42 62 39 55 82 67 44
51 77 61 52 76 41 59

Class V: 80 71 55 67 61 93 49 78
57 88 79 81 63 51 75

Use a chi-square test with $\alpha = 0.05$ to test the equality of the distributions of test scores by dividing the combined sample into 3 equal parts (low, middle and high)

Hint: 1st tertile = minimum + 0.33 * range
2nd tertile = minimum + 0.66 * range.

# Lecture 10: Interval Estimation

Let $X \sim f(x|\theta)$, $\theta \in \Omega$ and is unknown.

To estimate $\theta$, we draw a random sample of size $n$ from $f(x|\theta)$ and obtain a statistic $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$, by the method of moments (MME) or the method of maximum likelihood (MLE) or some other estimation method (e.g. LSE, etc).

Because $X_1, X_2, \dots, X_n$ is a random sample, it is unlikely $\hat{\theta}$ is the true value of $\theta$. If $\hat{\theta}$ has a continuous distribution, then $P_\theta(\hat{\theta} = \theta) = 0$. So, there's an <u>error</u> in the <u>estimation</u> of $\theta$. We wish to quantify how much $\hat{\theta}$ differs from $\theta$.

Defn (Confidence Interval).

Let $X_1, X_2, \ldots, X_n$ be a r.s. from $f(x|\theta)$, $\theta \in \Omega$.

Let $L = L(X_1, X_2, \ldots, X_n)$ and $U = U(X_1, X_2, \ldots, X_n)$ be two statistics. Then the interval $(L, U)$ is a $(1-\alpha)100\%$ confidence interval for $\theta$ if

$$1-\alpha = P_\theta(\theta \in (L, U)).$$

That is, the probability that the interval includes $\theta$ is $1-\alpha$ (confidence coefficient or confidence level of the interval).

For a given sample, the realized value of the interval is $(l, u)$, which either includes $\theta$ or not, hence can be thought of as a Bernoulli trial with probability $1-\alpha$

For $M$ independent samples, one would expect to have $(1-\alpha)M$ successful confidence intervals that trap $\theta$ over time. Hence, one would feel $(1-\alpha)100\%$ confident that the true value of $\theta$ lies in the interval $(l, u)$.

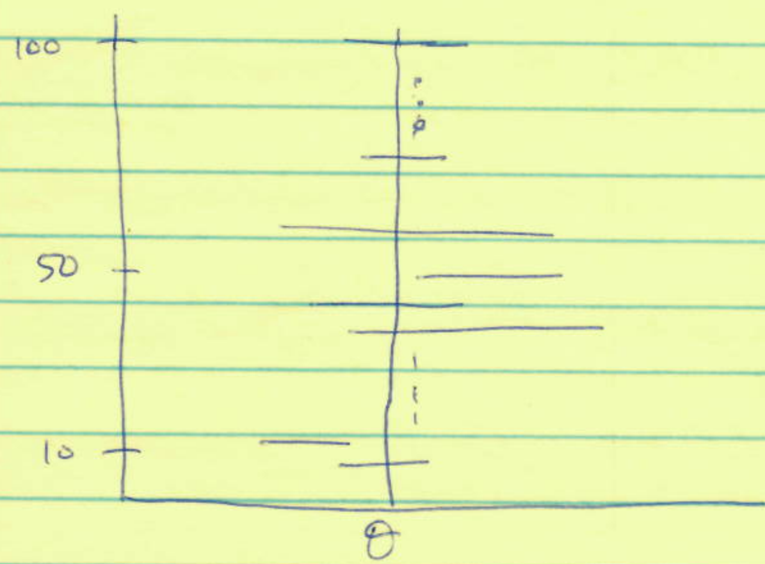
## Efficiency of a CI

Suppose $(L_1, U_1)$ and $(L_2, U_2)$ are two confidence intervals for $\theta$ with the same confidence coefficient. Then $(L_1, U_1)$ is more efficient than $(L_2, U_2)$ if

$$E_\theta(U_1 - L_1) \le E_\theta(U_2 - L_2) \text{ for all } \theta \in \Omega.$$

Here $U_i - L_i$, $i=1,2$, is the length of the interval.

Most common values of the confidence level are 90%, 95% and 99%.. So, the easiest way to visualize the 90% confidence level is by constructing 100 confidence intervals and counting the number of intervals that will trap $\theta$ (expected to be 90 out of 100)



Example 2 (Confidence Interval for $\mu$)

Suppose $X \sim N(\mu, \sigma^2)$. The MLE of $\mu$ and $\sigma^2$ are $\bar{X}$ and $S^2$, respectively, where

$$S^2 = \frac{1}{n} \sum (X_i - \bar{X})^2.$$

Let
$$T = (\bar{X} - \mu)/(S/\sqrt{n}) = \frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

Then $T \sim t(n-1)$.

We can "pivot" on $T$ to obtain a $(1-\alpha)100\%$ confidence interval for $\mu$:

$$1-\alpha = P(-t_{\alpha/2}(n-1) < T < t_{\alpha/2}(n-1)),$$

where $\frac{\alpha}{2} = P(T > t_{\alpha/2}(n-1))$.

$$1-\alpha = P\left(\bar{X} - t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}\right)$$

Thus, a $(1-\alpha)100\%$ CI for $\mu$ is
$$(\bar{x} \pm t_{\alpha/2}(n-1) \cdot s/\sqrt{n}).$$

The estimate of the standard deviation of $\bar{X}$, $s/\sqrt{n}$, is known as the standard error of $\bar{X}$ ( $SE(\bar{x})$).

Large sample CI for $\mu$:

Let $X_1, X_2, \ldots, X_n$ be a r.s. from $f(x|\theta)$, where $f(x|\theta)$ is not normal. Then, by the Central Limit Theorem,

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1), \text{ when } n \text{ is large.}$$

We can replace $\sigma$ by $S$ and keep the approximation valid. That is,

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx N(0,1).$$

Thus $1-\alpha \approx P_\mu\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right).$