STAT 311: Hypothesis Testing for Two Way Tables

Y. Samuel Wang

Summer 2016

Logistics

- Final next Friday
- Practice final posted today
- Review practice final on Wednesday
- Thursday will be general review / Questions

Example: Ebola case outcomes

Consider the Ebola data from we've previously examined[1]

	Death	Survive	Total
Guinea	2536	1268	3804
Liberia	4806	5860	10666
Sierra Leone	3955	10167	14122
Total	11297	17295	28592

[1] Data available from World Health Organization: `http://apps.who.int/gho/data/view.ebola-sitrep.ebola-summary-latest?lang=en`. Up to date as of Dec 2015

Two Way Tables

	Col 1	Col 2	Col 3	Total
Row 1	n_{11}	n_{12}	n_{13}	n_{1+}
Row 2	n_{21}	n_{22}	n_{23}	n_{2+}
Row 3	n_{31}	n_{32}	n_{33}	n_{3+}
Total	n_{+1}	n_{+2}	n_{+3}	n_{++}

- Joint
- Marginal
- Conditional

Independence

Two events are independent if the conditional distribution is equal to the marginal distribution

$$P(A|B) = P(A)$$

which implies the joint is the product of the marginals

$$P(A \bigcap B) = P(A|B)P(B) = P(A)P(B)$$

Independence

Two events are independent if the conditional distribution is equal to the marginal distribution

$$P(A|B) = P(A)$$

which implies the joint is the product of the marginals

$$P(A \bigcap B) = P(A|B)P(B) = P(A)P(B)$$

In the two way table, the marginal distribution is

$$\frac{n_{i+}}{n_{++}}$$

or

$$\frac{n_{+i}}{n_{++}}$$

Independence

Two events are independent if the conditional distribution is equal to the marginal distribution

$$P(A|B) = P(A)$$

which implies the joint is the product of the marginals

$$P(A \bigcap B) = P(A|B)P(B) = P(A)P(B)$$

Independence

Two events are independent if the conditional distribution is equal to the marginal distribution

$$P(A|B) = P(A)$$

which implies the joint is the product of the marginals

$$P(A \bigcap B) = P(A|B)P(B) = P(A)P(B)$$

In the two way table, the marginal distribution is

$$\frac{n_{i+}}{n_{++}}$$

or

$$\frac{n_{+i}}{n_{++}}$$

Expected Counts under independence

Under the assumption of independence,

$$P(R = r_i \bigcap C = c_j) = P(R = r_i)P(C = c_j) = \frac{n_{i+}}{n_{++}}\frac{n_{+j}}{n_{++}}$$

so the expected count is

$$n_{++}\frac{n_{i+}}{n_{++}}\frac{n_{+j}}{n_{++}} = \frac{n_{i+}n_{+j}}{n_{++}}$$

Measure of deviation

Given that there are many cells in a table, how do we measure how "different" the counts are from what we would expect if the variables are independent? To test

$$H_0 : \text{No association between row and column variables}$$

$$H_A : \text{Association between row and column variables}$$

we can use the following test statistic

$$\chi = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed counts and E_{ij} is the expected counts

Under the null hypothesis, χ follows a χ^2 distribution with (Rows -1)(Columns - 1) degrees of freedom.

χ^2 Distribution

The χ^2 distribution has a single parameter k which is the degrees of freedom.

Given standard normal random variables Z_i, we can form a χ^2 variable with k degrees of freedom-

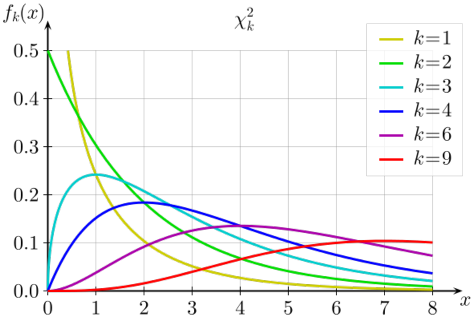
$$\chi = \sum_{i=1}^{k} Z_i^2$$

χ^2 Distribution

For $X \sim \chi^2_k$,

$$E(X) = k$$

$$Var(X) = 2k$$



Example: Ebola Case Outcomes

The observed counts are

	Death	Survive	Total
Guinea	2536	1268	3804
Liberia	4806	5860	10666
Sierra Leone	3955	10167	14122
Total	11297	17295	28592

The Expected counts are

	Death	Survive	Total
Guinea	1503.00	2301.00	3804
Liberia	4214.25	6451.75	10666
Sierra Leone	5579.75	8542.25	14122
Total	11297	17295	28592

Example: Ebola Case Outcomes

The deviations are

	Death	Survive	Total
Guinea	709.97	463.75	-
Liberia	83.09	54.27	-
Sierra Leone	473.10	309.037	-
Total	-	-	-

$$\chi = 2093$$

Under the null hypothesis, χ should be distributed as a $\chi^2(3)$
So we reject the null hypothesis that there is no association between the country and case outcome

Class Roadmap