18.05 Lecture 34
May 6, 2005

Contingency tables, test of independence.

	Feature 2 = 1	F2 = 2	F2 = 3	...	F2 = b	row total
Feature 1 = 1	N_{11}	...	...	...	N_{1b}	N_{1+}
F1 = 2	...	...	...	...	...	...
F1 = 3	...	...	...	...	...	...
...	...	...	...	...	...	...
F1 = a	N_{a1}	...	...	...	N_{ab}	N_{a+}
col. total	N_{+a}	...	...	...	N_{+b}	n

$X_i^1 \in \{1, ..., a\}$
$X_i^2 \in \{1, ..., b\}$

Random Sample:
$X_1 = (X_1^1, X_1^2), ..., X_n = (X_n^1, X_n^2)$

Question: Are X^1, X^2 independent?
Example: When asked if your finances are better, worse, or the same as last year, see if the answer depends on income range:

	Worse	Same	Better
$\leq$ 20K	20	15	12
20K - 30K	24	27	32
$\geq$ 30K	14	22	23

Check if the differences and subtle trend are significant or random.

$\theta_{ij} = \mathbb{P}(i, j) = \mathbb{P}(i) \times \mathbb{P}(j)$ if independent, for all cells ij

Independence hypothesis can be written as:
$H_1 : \theta_{ij} = p_i q_j$ where $p_1 + ... + p_a = 1, q_1 + ... + q_b = 1$
H_2 : otherwise.
r = number of categories = ab
s = dimension of parameter set = $a + b - 2$
The MLE p_i^*, q_j^* needs to be found $\rightarrow$

$$\mathcal{T} = \sum_{i,j} \frac{(N_{ij} - np_i^* q_j^*)^2}{np_i^* q_j^*} \sim \chi^2_{r-s-1=ab-(a+b-2)-1=(a-1)(b-1)}$$

Distribution has (a - 1)(b - 1) degrees of freedom.

Likelihood:

$$\psi(\overrightarrow{p}, \overrightarrow{q}) = \prod_{i,j} (p_i q_j)^{N_{ij}} = \prod_i p_i^{N_{i+}} \times \prod_j q_j^{N_{+j}}$$

Note: $N_{i+} = \sum_j N_{ij}$ and $N_{+j} = \sum_i N_{ij}$
Maximize each factor to maximize the product.

$\sum_i N_{i+} \log p_i \to \max,\ p_1 + ... + p_a = 1$

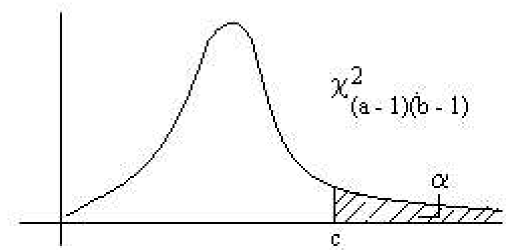
Use Lagrange multipliers to solve the constrained maximization:
$\sum_i N_{i+} \log p_i - \lambda(\sum_i p_i - 1) \to \max_p \min_\lambda$

$$\frac{\partial}{\partial p_i} = \frac{N_{i+}}{p_i} - \lambda = 0 \to p_i = \frac{N_{i+}}{\lambda}$$

$$\sum_i p_i = \frac{n}{\lambda} = 1 \to \lambda = n \to p_i^* = \frac{N_{i+}}{n}$$

$$p_i^* = \frac{N_{i+}}{n}, q_j^* = \frac{N_{+j}}{n}$$

$$\mathcal{T} = \sum_{i,j} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n} \sim \chi^2_{(a-1)(b-1)}$$

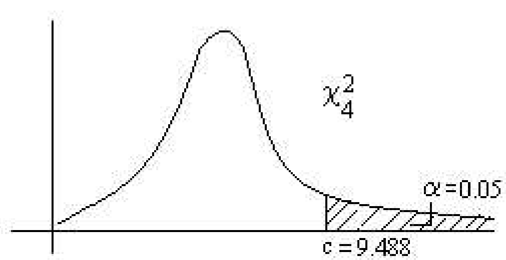


Decision Rule:
$\delta = \{H_1 : T \le c; H_2 : T > c\}$
Choose c from the chi-square distribution, (a - 1)(b - 1) d.o.f., at a level of significance α = area.

From the above example:
$N_{1+} = 47, N_{2+} = 83, N_{3+} = 59$
$N_{+1} = 58, N_{+2} = 64, N_{+3} = 67$
n = 189
For each cell, the component of the T statistic adds as follows:

$$\mathcal{T} = \frac{(20 - 58(47)/189)^2}{58(47)/189} + ... = 5.210$$

Is T too large?
$T \sim \chi^2_{(3-1)(3-1)} = \chi^2_4$



For this distribution, c = 9.488
According to the decision rule, accept H_1, because $5.210 \le 9.488$

Test of Homogeniety - very similar to independence test.

	Category 1	...	Category b
Group 1	N_{11}	...	N_{1b}
...	...	...	...
Group a	N_{a1}	...	N_{ab}

1. Sample from entire population.
2. Sample from each group separately, independently between the groups.

Question: $\mathbb{P}$(category j | group i) = $\mathbb{P}$(category j)
This is the same as independence testing!

$\mathbb{P}$(category j, group i) = $\mathbb{P}$(category j)$\mathbb{P}$(group i)

$$\rightarrow \mathbb{P}(C_j|G_i) = \frac{\mathbb{P}(C_jG_i)}{\mathbb{P}(G_i)} = \frac{\mathbb{P}(C_j)\mathbb{P}(G_i)}{\mathbb{P}(G_i)} = \mathbb{P}(C_j)$$

Consider a situation where group 1 is 99% of the population, and group 2 is 1%.
You would be better off sampling separately and independently.
Say you sample 100 of each, just need to renormalize within the population.
The test now becomes a test of independence.

Example: pg. 560
100 people were asked if service by a fire station was satisfactory or not.
Then, after a fire occured, the people were asked again.
See if the opinion changed in the same people.

Before Fire	80	20
After Fire	72	28
	satisfied	unsatisfied

But, you can't use this if you are asking the same people! Not independent!
Better way to arrange:

Originally Satisfied	70	10
Originally Unsatisfied	2	18
	After, Satisfied	After, Not Satisfied

If taken from the entire population, this is ok. Otherwise you are taking from a dependent population.

** End of Lecture 34