University of California, Los Angeles
Department of Statistics

Statistics 13 **Instructor: Nicolas Christou**

Contingency tables

Many experiment result in enumerative (count) data. For example patients suffering from a certain disease are classified based on the type of medication thy receive and the rate of recovery. We may want to test if recovery depends on the type of medication. So we want to test the dependency (contingency) between the two classification criteria.

Example 1:
A survey was conducted to evaluate the effectiveness of a new flu vaccine that was administered in a certain small community. The vaccine was provided in a two-shot sequence over a period of two weeks. Some people received the two-shot sequence, some they received only the first shot, and the others received neither. A survey of 1000 people gave the following results:

	No vaccine	One shot	Two shots	Total
Flu	24	9	13	46
No Flu	289	100	565	954
Total	313	109	578	1000

Do these data provide evidence to indicate a dependence between vaccine and flu occurrence?

First we calculate the expected frequencies using the row and column totals. For example, we would calculate $E(n_{11}) = \frac{r_1 c_1}{n} = \frac{(46)(313)}{1000} = 14.4$. Similarly we compute all the expected frequencies and complete the table below:

	No vaccine	One shot	Two shots	Total
Flu	14.4	5.0	26.6	46
No Flu	298.6	104.0	551.4	954
Total	313	109	578	1000

Test statistic: $Y = \frac{(24-14.4)^2}{14.4} + \frac{(289-298.6)^2}{298.6} + \ldots + \frac{(565-551.4)^2}{551.4} = 17.35$.

We reject the null hypothesis (no dependence) if $Y > \chi^2_{1-\alpha;df}$, where $df = (r-1)(c-1)$, with r and c the number of rows and columns respectively. For our example, $r = 2, c = 3$. Therefore, if we choose $\alpha = 0.05$ we get $\chi^2_{0.95;2} = 5.95$. Conclusion: We reject H_0.

Example 2:
Contingency table using Titanic data (surviving/class):

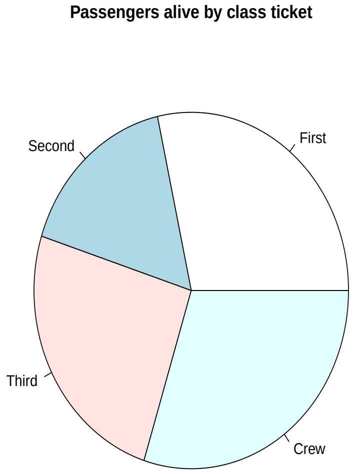
	First	Second	Third	Crew	Total
Alive	202	118	178	212	710
Dead	123	167	528	673	1491
Total	325	285	706	885	2201

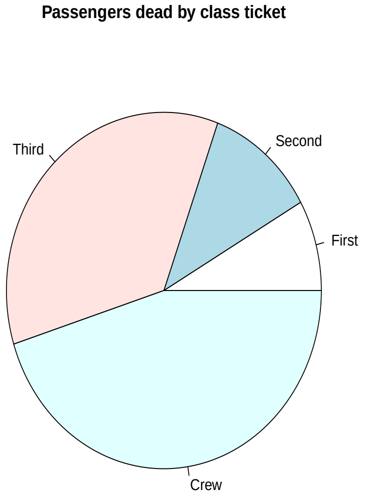
We want to investigate if class and surviving are independent.

```
x <- c(202,118,178,212)
y <- c(123,167,528,673)

names(x) <- c("First", "Second","Third", "Crew")
names(y) <- c("First", "Second","Third", "Crew")

pie(x, main="Passengers alive by class ticket")
pie(y, main="Passengers dead by class ticket")
```





Clearly there are differences. Are they statistically significant? Compute the test statistic:

```
y <- (202-104.8387)^2/104.8387 + (118-91.93548)^2/91.93548 +
     (178-227.7419)^2/227.7419+(212-285.4839)^2/285.4839   +

     (123-220.1613)^2/220.1613+(167-193.0645)^2/193.0645   +
     (528-478.2581)^2/478.2581+(673-455.9032)^2/455.9032.

y
[1] 282.1656
```

Highly significant!