1 Contingency Tables

Measure association between categorical variables. Similar to Pearson's Correlation Coefficient for numerical variables.

1.1 Probability Structure

Distribution of ONE categorical variable X, with possible outcomes $x_1, \ldots, x_k$
The distribution is described by the outcomes and their probabilities: $\pi_i = P(X = x_i)$, with

$$\sum_i \pi_i = 1$$

The distribution table of the random variable X:

x_i	π_i
x_1	π_1
x_2	π_2
$\vdots$	$\vdots$
x_k	π_k
	1

Example 1
Random variable X=Use of phone while driving(Yes/No), therefore categorical

x	$P(X = x)$
Yes	0.6
No	0.4
	1

Joint distribution of two categorical random variables lead to probability tables.

Let X be a random variable with I categories,
Let Y be a random variable with J categories.
Then the joint distribution of X and Y is described by the joint probabilities

$$P(X = i, Y = j) = \pi_{ij}, \quad 1 \le i \le I, 1 \le j \le J, \quad \text{with} \quad \sum_{ij} \pi_{ij} = 1$$

These can be organized in a table:

	Y			
X	1	2	$\ldots$	J
1	π_{11}	π_{12}	$\ldots$	π_{1J}
2	π_{21}	π_{22}	$\ldots$	π_{2J}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
I	π_{I1}	π_{I2}	$\ldots$	π_{IJ}

Example 2
Let X= person uses phone while driving (yes, no), $I = 2$
Let Y= person received speeding ticket or was involved in a car accident, (yes, no), $J = 2$.

	Y	
X	yes	no
yes	π_{11}	π_{12}
no	π_{21}	π_{22}

$=$

	Y	
X	yes	no
yes	0.1	0.5
no	0.01	0.39

The table gives the joint distribution of X and Y, for example the probability that a person using the phone AND getting in an accident $P(X = yesANDY = yes) = 0.1$.
The *marginal distributions* are the row and the column totals of the joint probabilities and describe the distributions of X and Y, ignoring the other variable. They are denoted by π_{i+} and π_{+j}, the "+" replacing the index totalled.

	Y				
X	1	2	...	J	Total
1	π_{11}	π_{12}	...	π_{1J}	π_{1+}
2	π_{21}	π_{22}	...	π_{2J}	π_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
I	π_{I1}	π_{I2}	...	π_{IJ}	π_{I+}
Total	π_{+1}	π_{+2}	...	π_{+J}	$\pi_{++} = 1$

Example 3
Let X= person uses phone while driving (yes, no), $I = 2$
Let Y= person received speeding ticket or was involved in a car accident, (yes, no), $J = 2$.

	Y		
X	yes	no	Total
yes	0.1	0.5	0.6
no	0.01	0.39	0.4
Total	0.11	0.89	1

With the interpretation $P(X = \text{yes})$=0.6, $P(Y = \text{yes})$=0.11.
The marginal distributions of X and Y are, respectively:

x	$P(X = x)$
Yes	0.6
No	0.4
	1

y	$P(Y = y)$
Yes	0.11
No	0.89
	1

When interpreting one of the variables as a response variable (let's say Y), we are interested in the probabilities for the different outcomes of Y, given a certain outcome of the other variable (this must be then X).

$$P(Y = j|X = i) = \frac{\pi_{ij}}{\pi_{i+}}, \quad 1 \leq i \leq I, 1 \leq j \leq J$$

These probabilities make up the *conditional distribution of* Y *given* X.

Example 4

$$P(Y = yes|X = yes) = \frac{0.1}{0.6} = 0.166, \quad P(Y = yes|X = no) = \frac{0.01}{0.4} = 0.025$$

The conditional distribution of Y given X=yes and X =no are, respectively:

y	$P(Y = y\|X = yes)$
Yes	0.166
No	0.833
	1

y	$P(Y = y\|X = no)$
Yes	0.025
No	0.975
	1

Association between X and Y?

The same concepts can be applied to sample data, (replacing π by p), and the resulting numbers are then interpreted as estimates for the true joint, marginal, and conditional probabilities.

1.1.1 Sensitivity and Specificity

Sensitivity and specificity are certain conditional probabilities when discussing correct classifications using diagnostic tests.
Sensitivity = P(Test positive | Diagnosis yes), Specificity = P(Test negative | Diagnosis no)
The higher sensitivity and specificity the better the diagnostic test.

Positive Predictive Value = P(Diagnosis yes | Test positive),
Negative Predictive Value = P(Diagnosis no | Test negative)
Be aware that even when sensitivity and specificity are high the positive and negative predictive values do not have to be high.

Example 5
(From Statistics Notes(http://pubmedcentralcanada.ca/pmcc/articles/PMC2540489/pdf/bmj00444-0038.pdf))
Y = Pathology (abnormal, normal)
X = Liver scan (abnormal, normal)

	Pathology		
Liver Scan	Abnormal(a)	Normal(n)	Total
Abnormal(a)	231	32	263
Normal(n)	27	54	81
Total	258	86	344

Estimate for joint probability $p_{aa} = 231/344$
Estimate for marginal probability $p_{a+} = 263/344$
Estimate for marginal probability $p_{+a} = 258/344$
Estimate for conditional probability $\hat{P}(P = a|LS = a) = 231/263$
Estimate for sensitivity $\hat{P}(LS = a|P = a) = 231/258$
Estimate for specificity $\hat{P}(LS = n|P = n) = 54/86$
Estimate for positive predictive value $\hat{P}(P = a|LS = a) = 231/263$
Estimate for negative predictive value $\hat{P}(P = n|LS = n) = 54/81$

Definition 1
Two categorical random variables are statistically independent if the conditional distributions of Y are identical for all levels of X.

Example 6
The result of the liver scan would be independent from the pathology, if for abnormal and normal pathology the probability of getting a normal liver scan would be the same.

Remark:
Two variables are statistically independent iff the joint probabilities are the product of the marginal probabilities

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for} \quad 1 \leq i \leq I, 1 \leq j \leq J$$

1.1.2 Binomial and Multinomial Sampling

Watch out!!

- If X is a group variable (for example placebo versus treatment) and the sample sizes are fixed for each group (for example 50 for placebo and 50 for treatment), then it does not make sense to look at the joint probabilities.

  The conditional distributions of Y are then either binomial or multinomial for each level of X, depending on J ($J = 2 \rightarrow$ binomial, $J > 2 \rightarrow$ multinomial).

- The same is true when X is considered an explanatory variable, and Y the response variable. Here again it makes more sense to look at the conditional distributions of Y given X. With the same consequence for the distribution as above. (Example above: response = pathology, explanatory = liver scan)

- When both variables can be considered response variables (Example: ask randomly chosen students at MacEwan, which year of study they are in (1st, 2nd, 3rd, 4th), and if they use public transportation (yes/no)), then the joint distribution of X and Y is a multinomial distribution with the cell being the possible outcomes.

1.2 2×2 tables

1.2.1 Comparing two Proportions

Use

$$\pi_1 = P(Y = \text{success}|X = 1), \pi_2 = P(Y = \text{success}|X = 2)$$

and $n_1 = n_{1+}, n_2 = n_{2+}$.
Then the *Wald*-confidence interval for $\pi_1 - \pi_2$ is

$$(p_1 - p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Plus four method (Agresti-Cull): For small samples the confidence interval can be improved by adding two imaginary observations to each sample (one success and one failure).

An approximately normally distributed **test statistic** for comparing π_1 and π_2 is

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_p(1-p_p)}{n_1} + \frac{p_p(1-p_p)}{n_2}}}$$

with the pooled estimator

$$p_p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$$

Example 7
A genetic fingerprinting technique called polymerase chain reaction (PCR), can detect as few as 5 cancer cells in every 100,000 cells, which is much more sensitive than other methods for detecting such cells, like using a microscope.
Investigators examined 178 children diagnosed with acute lymphoblastic leukemia who appeared to be in remission using a standard criterion after undergoing chemotherapy (Cave et al, 1998). Using PCR traces of cancer were detected in 75 of the children. During 3 years of follow up 30 of the children suffered a relapse. Of the 103 children who did not show traces of cancer 8 suffered a relapse.
Do these data provide sufficient evidence that children are more likely to suffer a relapse, if the PCR is positive (detect cells)?

Data:

	Relapse		
PCR	yes	no	Total
positive	30	45	75
negative	8	95	103
Total	38	140	178

$p_1 = 30/75 = 0.4$, $p_2 = 8/103 = 0.07767$, then the 95% ci for $\pi_1 = \pi_2$

$$(p_1 - p_2) \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \rightarrow [0.200, 0.445]$$

Since 0 does not fall within the confidence interval at 95% confidence the data do provide sufficient evidence that the proportion of children who suffer a relapse differs for those with positive PCR and a negative PCR. The proportion of children who have a positive PCR is between 0.2 and 0.445 larger than of children with a negative PCR.

1.2.2 Relative Risk

The relative risk measures how much more likely a success is in one group than in another. It is a measure of association for the variables defining a 2×2 table, similar to how Pearson's correlation coefficient is measuring the association between two numerical variables.

Definition 2

$$\text{relative risk} = rr = \frac{\pi_1}{\pi_2}$$

- $rr = 1 \Leftrightarrow \pi_1 = \pi_2$ and the variables are independent.
- If $\pi_1 = 0.5001$ and $\pi_2 = 0.5$ then $\pi_1 - \pi_2 = 0.001$ and $rr = 1.0002$, but
- if $\pi_1 = 0.0002$ and $\pi_2 = 0.0001$ then still $\pi_1 - \pi_2 = 0.0001$ and now $rr = 2$ (the probability in group 1 is 2 times as high as in group 2).

  In the second case the difference between the probabilities is more consequential, and the relative risk the better description for the difference.

Confidence interval for the **log of the population rr**

$$ln(p_1/p_2) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}$$

Using the exponential function for upper and lower bound of this ci, results in a ci for the population rr.

Example 8
PCR and relapse:

Estimated rr:

$$\hat{rr} = \frac{0.4}{0.07767} = 5.15$$

The relative risk of a relapse for children with positive PCR versus negative PCR is estimated to be 5.15. The probability of a relapse 5.15 times larger for children with positive PCR.
The 95% ci for $\ln(rr)$:

$$\ln(5.14999) \pm 1.96\sqrt{0.6/30 + 0.92233/8} \rightarrow 1.63900 \pm 0.72092 \rightarrow [0.91807, 2.35992]$$

Then the 95% for the true relative risk is $\exp(0.91807), \exp(2.35992)] = [2.5044, 10.5901]$.
We are 95% confident that the true rr falls between 2.50 and 10.59.
Since 1 does not fall within the confidence interval at confidence level of 95% the data do provide sufficient evidence that the rr for a relapse is different from 1. $rr = 1$ would mean that the probabilities for a relapse is the same for children with positive and negative PCR. A relapse is between 2.5 and 10.6 times more likely for children with a positive PCR.

1.3 The Odds Ratio

The odds ratio is another measure of association in a two way tables.

Let π be the probability of success

$$\text{odds} = \frac{\pi}{(1-\pi)}$$

If $\pi = 0.75$, then odds $= 0.75/0.25 = 3$, success is three times more likely than failure. For every failure we expect three successes.
If $\pi = 0.25$, then odds $= 0.25/0.75 = 1/3$, failure is three times more likely than success.

In reverse, finding the probability when knowing the odds:

$$\pi = \frac{\text{odds}}{\text{odds} + 1}$$

Example 9
Assume the odds for a randomly chosen student to get an A in Stat 151 are 1:10, therefore the probability, π of getting an A in Stat 151 is:

$$\pi = \frac{\text{odds}}{\text{odds}+1} = \frac{1/10}{1+1/10} = \frac{1/10}{(10+1)/10} = \frac{1}{10+1} = \frac{1}{11}$$

Definition 3
In a two way table, let $\text{odds}_1 = \pi_1/(1-\pi_1)$ the odds of success in row 1, and $\text{odds}_2 = \pi_2/(1-\pi_2)$ the odds of success in row 2.
Then

$$\theta = \frac{\text{odds}_1}{\text{odds}_2}$$

is the odds ratio.

Properties:

- $\theta \geq 0$
- When X and Y are independent, then $\theta = 1$ (the probabilities and odds for success are the same in both rows).
- $\theta > 1$ means the odds for success are higher in row 1 than in row 2. For example $\theta = 2$, means that the odds for success are two times higher in row 1 than in row 2. Individuals from row one are more likely to have a successes than those in row 2.
- $\theta < 1$ means the odds are smaller in row 1 than in row 2. Successes are less likely for individuals in row 1 than for those in row 2 ($\pi_1 < \pi_2$).
- θ_1 and θ_2 indicate the same strength in association if $\theta_1 = 1/\theta_2$. (0.25=1/4).
- The odds ratio does not change when the table is transposed, i.e. when the rows become the columns, and the columns the rows. (This is the advantage of the odds ratio over the relative risk.)
- When both variables are response variables then

  $$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

  and the sample odds ratio $\hat{\theta}$:

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

  which is also the ML estimator for θ (what does that mean again?).

Same as for the relative risk $\hat{rr}$, the distribution of $\hat{\theta}$ is very skewed. Therefore instead of finding a confidence interval for θ, we will find one for $\ln(\theta)$ and then transform the result to obtain a confidence interval for θ.

With

$$SE(ln(\hat{\theta})) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

A confidence interval for $\ln(\theta)$ is

$$\ln(\hat{\theta}) \pm z_{\alpha/2} SE(ln(\hat{\theta}))$$

Example 10
Relapse and PCR

$$\hat{\theta} = \frac{30 \times 95}{8 \times 45} = 7.92$$

The odds for a relapse are estimated to be almost 8 times higher for children with positive PCR than for children with a negative PCR.
A 95% confidence interval for $\ln(\theta)$

$$\ln(7.92) \pm 1.96\sqrt{\frac{1}{30} + \frac{1}{45} + \frac{1}{8} + \frac{1}{95}} \rightarrow 2.0694 \pm 0.8568 \rightarrow [1.213, 2.926]$$

Which gives the following confidence interval for θ: $[\exp(1.213), \exp(2.926)] = [3.36, 18.65]$. Since one does not fall within the interval at 95% confidence we do have sufficient evidence that a relapse is associated with PCR.

Example 11
Hepatitis B Vaccine and Risk of Multiple Sclerosis (Alberto Ascherio et al., N Engl J Med 2001; 344:327-332)

Does the vaccine cause MS?

Data from a large study was used. For each MS patient five matching healthy women were included for the analysis (relating to age and other factors).
Retrospective study, (was a women vaccinated, after finding if she has or has not MS).
This is a Case-control study!

Because this is a case-control study some limitations to what can be done apply.
Since the proportion of MS/Healthy patients has been chosen by the sampling design, we can not estimate probabilities based on π_{MS}, like $P(MS|Vaccine)$, but we can estimate $P(Vaccine|MS)$. (We can make comparisons between the two samples MS and non-MS, but not vaccine and non-vaccine)
Therefore we can not use this data for comparing the proportion of MS patients within the group of vaccinated and non vaccinated individuals. This data can also not be used for estimating the relative risk of MS in the two groups who were vaccinated or not.
In order to answer the question "Does the vaccine cause MS?" the only measure that is applicable is the odds ratio, because it is symmetric in X and Y (transposing the table has no effect on the odds ratio).

Data:

Vaccine	MS yes	no	Total
yes	32	84	116
no	158	450	608
Total	190	534	724

Two approaches can be taken to find the estimate for the odds ratio

1. $p_1 = 32/116 = 0.27586$, $p_2 = 158/608 = 0.25987$ (these numbers have no valuable interpretation as explained above and are only used for calculation purposes)
   (a) o_1: odds for vaccinated women: $o_1 = p_1/(1-p_1) = 32/84 = 0.38095$
   (b) o_2: odds for non vaccinated women: $o_2 = p_2/(1-p_2) = 158/450 = 0.35111$
   (c) $\hat{\theta}$: odds ratio $\hat{\theta} = o_1/o_2 = 1.0850$
2. $$\hat{\theta} = \frac{32 \times 450}{158 \times 84} = 1.0850$$

The odds of getting MS is a little higher for women who were vaccinated than for those who were not.

A 95% ci for $\ln(\theta)$

$$\ln(1.0850) \pm 1.96\sqrt{\frac{1}{32} + \frac{1}{84} + \frac{1}{158} + \frac{1}{450}} \to 0.081580 \pm 0.44568 \to [-0.36410, 0.52726]$$

Which gives following ci for θ: $[\exp(-0.36410), \exp(0.52726)] = [0.69, 1.69]$. Since 1 falls within the interval at 95% confidence we do not have sufficient evidence that MS is associated with the vaccine.

Remark:

- An *odds ratio* of $\theta = 2$, means that the odds for group 1 are double the odds for group 2.
- A relative risk of $rr = 2$, means that the probability for success in group1 is double the success probability in group 2.
- $$\theta = rr \times \frac{1-\pi_2}{1-\pi_1}$$
  When both probabilities are close to zero the last fraction is close to one, and $\theta \approx rr$.

Remark:
Odds ratio as well as relative risk are measures of association in a 2×2 table and give how much higher the risk or odds are for one group versus another one. Usually the relative risk is easier to interpret because it compares probabilities instead of odds. Beside the difference in interpretation there is a difference when they are applicable. The odds ratio can be used in all types of studies,

but the relative risk can only be used, if the rows can be interpreted as random samples from a populations with and without the risk factor.
This means that in particular for case-control studies the relative risk is not applicable. A case-control design involves the selection of research subjects on the basis of the outcome measurement rather than on the basis of the exposure (see above, we chose MS patients and healthy subjects, instead of people who were vaccinated and not vaccinated).

1.3.1 Types of studies

- Case-control, for each case find matching individuals for the control group at a given ratio. The cases are defined to have the outcome to be studied (disease = yes)
- Cohort studies (individuals choose themselves which groups they are in, ex. person smokes (yes/no))
- Clinical studies (individuals are randomly assigned to treatment groups by experimenter)
- Cross-sectional studies (random samples are taken and individuals are assessed which group they fall within and their response, ex.: relapse, PCR study)
- observational studies: cohort, cross sectional, and case control (more practical)
- experimental studies: clinical (less pitfalls)

1.4 χ^2 tests for Independence

Imagine two categorical variables X with I levels and Y with J levels.
Then the joint distribution of X and Y is given by.

$$P(X = i, Y = j) = \pi_{ij} \quad 1 \leq i \leq I, 1 \leq j \leq J$$

For samples of size n from this joint distribution the *expected cell frequencies* are then

$$\mu_{ij} = n\pi_{ij} \quad 1 \leq i \leq I, 1 \leq j \leq J$$

(similar to the mean of a binomial distribution).

If we want to test the null hypothesis that the joint distribution of X and Y equals $\{\pi_{ij}\}$, then this can be done comparing the expected cell frequencies $\{\mu_{ij}\}$ for the claimed distribution with the observed cell counts $\{n_{ij}\}$.
The larger the differences $\{n_{ij} - \mu_{ij}\}$ the more likely the null hypothesis is wrong.

1.4.1 Pearson's statistic

(compare with Stat 151/141)

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

(studied by Pearson in 1900) Properties:

- $X^2 \geq 0$
- $X^2 = 0$ when $n_{ij} = \mu_{ij}$
- For large n: X^2 is approximately χ^2 distributed (diagram)

  The approximation is close enough when $\mu_{ij} \geq 5$ for all cells.
- $\mu_{X^2} = df$, $\sigma_{X^2} = \sqrt{2df}$
- X^2 is a score statistic, it is based on the distribution on the standard errors and covariances of the n_{ij} when H_0 is true.
- Large values of X^2 give evidence against H_0.

1.4.2 Likelihood Ratio Statistic

Reminder:

- The likelihood function l gives the probability of the data to occur depending on the parameter(s).
- Find the largest value of l if H_0 is true $\rightarrow$ call this l_0
- Find the largest value of l if H_0 is true or not $\rightarrow$ call this l_1
- $G^2 = -2\ln(l_0/l_1)$ is the test statistic.

We get the likelihood-ratio χ^2 statistic

$$G^2 = 2\sum_{ij} n_{ij} \ln \frac{n_{ij}}{\mu_{ij}}$$

Properties:

- $G^2 \geq 0$
- $G^2 = 0$ when $n_{ij} = \mu_{ij}$
- For large n: G^2 is χ^2 distributed

  The approximation is close enough when $\mu_{ij} \geq 5$ for all cells.
- $\mu_{G^2} = df$, $\sigma_{G^2} = \sqrt{2df}$
- Large values of G^2 give evidence against H_0.

1.4.3 Tests for independence

Both test statistics can be used for a test for independence. We only have to answer the question, how to find μ_{ij} when X and Y are independent.
When X and Y are independent, then

$$\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$$

which can be estimated by

$$\hat{\mu}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

The degree of freedom of the test statistic is the difference between the number of parameters under H_a and H_0.

- When H_a is true $\{\pi_{ij}\}$ are $I \times J$ parameters, but one of them can not be chosen freely, since the total has to be one, therefore there are $IJ - 1$ parameters under H_a
- When H_0 holds the parameters are the marginal distributions $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ with $I - 1$ and $J - 1$ parameters (subtract 1 because the total has to be 1), respectively, for a total of $(I - 1) + (J - 1) = I + J - 2$
- Then $df = (IJ - 1) - (I + J - 2) = (I - 1)(J - 1)$ (some algebra).

This now leads to
Tests for Independence

1. Hyp.: $H_0 : X$ and Y are independent versus $H_a : H_0$ is not true, choose α
2. Assumptions: random samples, $\{\hat{\mu}_{ij} \geq 5\}$.
3. Test Statistic: X_0^2 or G_0^2, $df = (I - 1)(J - 1)$
4. P-value: P-value=$P(\chi^2 > X^2)$ or P-value=$P(\chi^2 > G^2)$

Example 12

A survey was conducted to evaluate the effectiveness of a new flu vaccine that had been administered in a small community. It consists of a two–shot sequence in two weeks.
A survey of 1000 residents the following spring provided the following information:

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

1. Hyp.: H_0 : Getting the flu and the vaccine are independent versus $H_a : H_0$ is not true, $\alpha = 0.05$
2. Assumptions: random samples, $\{\hat{\mu}_{ij} \geq 5\}$ (check later).

3. Test Statistic: X_0^2 or G_0^2, $df = (I-1)(J-1)$ First find the expected cell counts:

	No vaccine	One Shot	Two Shots	Total
Flu	24	9	13	46
	14.40	5.01	26.59	
No Flu	289	100	565	954
	298.60	103.99	551.41	
Total	313	109	578	1000

Then $df = (3-1)(2-1) = 2$ and

$$X_0^2 = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = 6.404 + 3.169 + 6.944 + 0.309 + 0.153 + 0.335 = 17.31$$

and

$$G_0^2 = 2\sum_{ij} n_{ij} \ln \frac{n_{ij}}{\hat{\mu}_{ij}} = 2\left(24 \ln \frac{24}{14.40} + \cdots + 565 \ln \frac{565}{551.41}\right) = 17.26$$

4. P-value: P-value=$P(\chi^2 > X_0^2) < 0.005$ or P-value=$P(\chi^2 > G_0^2) < 0.005$ (table VII).
5. Reject H_0, since the P-value is smaller than $\alpha = 0.05$.
6. At significance level of 5% the data provide sufficient evidence that there is an association between the vaccination status and getting the flu.

To study the nature of the association we could, for each cell, compare the observed and the expected cell count. But since for larger estimated expected cell counts the standard error increases we need to standardize:

$$Z_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}$$

is the standardized cell residual (called adjusted residual by SPSS) for the cell in row i and column j.
If H_0 would be true these would be approximately standard normally distributed. Therefore standardized cell residuals which do not fall between -1.96 and 1.96 indicate that the particular cell gives evidence that H_0 is violated.
Should we do a multiple comparison?

Example 13

The standardized cell residuals for the flu example:

	No vaccine	One Shot	Two Shots	Total
Flu	24	9	13	46
	3.1	1.9	-4.2	
No Flu	289	100	565	954
	-3.1	-1.9	4.2	
Total	313	109	578	1000

The significant positive residual for people, who did not get vaccinated and did get the flu, indicates that these cases occurred more often than the we would predict under independence. The significant negative residual for people, who were vaccinated twice and did not get the flu, indicates that these cases occurred less often than we would predict under independence.
Together this means that non-vaccinated people did get the flu more often than expected and vaccinated did got the flu less often, therefore we can conclude that the vaccine helps preventing people from getting the flu.

Since we have to do 3 tests (one for each column, since we only have two rows and the standardized cell cell residuals add up to zero for each row and column), it might be more appropriate to do a multiple comparison, using a significance level of $\alpha^* = \alpha/3$ for each test.
The sample odds ratio for no vaccine and 2 shots is

$$\hat{\theta} = \frac{24 \times 565}{289 \times 13} = 3.6$$

The estimated odds for getting the flu are 3.6 times higher for people who did not get vaccinated compared to those who got 2 shots. Or we can say, that the estimated odds for getting the flu are 260% higher for people who did not get vaccinated compared to those who got 2 shots.
Yep, the vaccine works.

1.5 Testing Independence of Ordinal Variables

The trick for studying the association between two categorical variables is to assign numerical scores to the categories which reflect the distance between the categories and then use tools which we would usually use for numerical variables, like Pearson's correlation coefficient.

1.5.1 Linear relationship between categorical variables

We are going to introduce a test to test if two ordinal variables are linearly related (if there is a linear trend association).
– diagram
Assume scores have been assigned (how to do this we will look at later). The test statistic based on the covariance (correlation) is then sensitive to a linear trend.
Let $u_1 \leq u_2 \leq \cdots \leq \ldots u_I$ denote the scores for the row categories, and $v_1 \leq v_2 \leq \cdots \leq v_J$ denote scores for the column categories (the ordering of the scores should reflect the ordering of the categories) with greater distances between categories which are considered farther apart.
The mean score for the rows and columns are then, respectively:

$$\bar{u} = \sum_i \frac{n_{i+}}{n} u_i = \sum p_{i+} u_i, \bar{v} = \sum_j \frac{n_{+j}}{n} v_j = \sum p_{+j} v_j$$

The sample covariance for X and Y is

$$\widehat{Cov}(X, Y) = \sum_{ij} p_{ij}(u_i - \bar{u})(v_j - \bar{v})$$

the sample standard deviations for X and Y, respectively:

$$s_X = \sqrt{\sum_i p_{i+}(u_i - \bar{u})^2}, s_Y = \sqrt{\sum_j p_{+j}(v_j - \bar{v})^2}$$

and the sample correlation

$$r = \frac{\widehat{Cov}(X,Y)}{s_x s_y}$$

Properties:

- $-1 \leq r \leq 1$
- $r < 1$ indicates a negative relationship (the larger X the smaller Y).
- $r > 1$ indicates a positive relationship (the larger X the larger Y).
- $r \approx 0$ no relationship.

This is an estimate for the population correlation ρ.
If $\rho = 0$

$$M^2 = (n-1)r^2, df = 1$$

is approximately χ^2-distributed with $df = 1$. Based on this we can develop a test for linear trend for two ordinal variables.

A test about the correlation (for linear trend):

1. Hyp.: $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$, choose α
2. random samples
3. Test Statistic:
$$M_0^2 = (n-1)r^2, df = 1$$
4. P-value: $P = P(\chi^2 > M_0^2)$

Example 14
Effect of smoking on the development of tartar. (Toutenberg: Statistical Analysis of designed experiments (1992))

	Tartar			
Smoking	none	middle	heavy	Total
no	284	236	48	568
middle	606	983	209	1798
heavy	1028	1871	425	3324
Total	1918	3090	682	5690

Use coding 1,2,3 for tartar and smoking, then $r = 0.095$, which seems to be quite low, this is due to the small number of possible values. Use coding 1,2,3 for tartar, and for smoking 0, 2, 3,(indicate a larger difference between no smoking and some, than between some and heavy) then $r = 0.103$, which is a little larger.
To test if there is a linear trend between these two variables

1. Hyp.: $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$, choose $\alpha = 0.05$

2. random samples

3. Test Statistic:
$$M_0^2 = (n-1)r^2 = (5689)0.103^2 = 60.35, df = 1$$

4. P-value: $P = P(\chi^2 > M_0^2) \approx 0$

5. Decision: Reject H_0

6. Context: At significance level of 5% the data provide sufficient evidence that the more a person smokes the more extreme the tartar build up.

Advantages of test for linear trend over test for independence:

- gain in power (power is the probability of a correct decision, when not rejecting H_0). i.e. if the test is not significant then it is more likely that there truly is no (positive or negative) association between the ordinal variables.
- smaller sample size required to detect association (test is more sensitive).
- smaller samples are required for the approximation of the distribution to be close.

1.5.2 Assigning Scores to Ordinal Variables

Agresti "For most data sets, the choice of scores has little effect on the results. Different choices of ordered scores usually give similar results."
The biggest problems occur when the data is unbalanced, i.e. if some categories have many more observations than other. (For example when looking at the evaluation of an excellent instructor, then categories "excellent and "very good" get high frequencies, but "poor" gets none or very little). Usually it is reasonable to use evenly spaced scores for the categories, like $1, 2, 3, \ldots$ or $10, 20, \ldots$ for X. Another solution is to use midranks.
For example for the data in the previous table this would give scores

categories	frequencies	midrank
no	568	(568+1)/2=284.5
middle	1798	((568+1) + (568+1798))/2=1467.5
heavy	3324	((568+1798+1)+(568+1798+3324))/2=4028.5

In the case when using midranks then r is called Spearman's ρ (or should it be called r) or Spearman's correlation.
For the example Spearman's correlation=0.086 (M^2=42.08, $df = 1$, $P < 0.001$). This is consistent with the results using Pearson's correlation, but the better choice because this statistic takes into account that the data is ordinal and not numerical, and the scores are not dropping out of the sky, but are based on the data.

1.5.3 Kendall's τ

Kendall's $\tau - b$ (tau) is an alternative measure for the relationship between two ordinal measures with the same number of outcomes (the contingency table is square). It is based on the number of concordant and discordant observations in the sample.
A pair of observations, $(x_i, y_i), (x_j, y_j)$, is called concordant if both $x_i < x_j$ and $y_i < y_j$, or both $x_i > x_j$ and $y_i > y_j$.
A pair of observations, $(x_i, y_i), (x_j, y_j)$, is called discordant if both $x_i < x_j$ but $y_i > y_j$, or both $x_i > x_j$ but $y_i < y_j$.
Let n_c be the number of concordant pairs, and n_d the number of discordant pairs.
Then Kendall's τ is

$$\hat{\tau} = 2\frac{n_c - n_d}{n(n-1)}$$

Example 15

	Tartar			
Smoking	none	middle	heavy	Total
no	284	236	48	568
middle	606	983	209	1798
heavy	1028	1871	425	3324
Total	1918	3090	682	5690

Then

$$n_c = 284(983 + 209 + 1871 + 425) + 236(209 + 425) + 606(1871 + 425) + 983(425) = 2,949,367$$

and

$$n_d = 48(606 + 983 + 1028 + 1871) + 236(606 + 1028) + 209(1028 + 1871) + 983(1028) = 2,217,463$$

therefore

$$\hat{\tau} = 0.045$$

The problem with τ is that it ignores the presence of ties. The measure Kendall's $\tau - b$ is a modification of Kendall's τ accounting for ties applicable in square tables only, and Kendall's $\tau - c$ generalizes Kendall's $\tau - b$ for non square tables. The interpretation remains the same.

For this data Kendall's $\tau - b$ is 0.080 (t=6.436, df=1, p <0.001).

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	.080	.012	6.437	.000
	Spearman Correlation	.086	.013	6.545	.000[c]
Interval by Interval	Pearson's R	.095	.013	7.188	.000[c]
N of Valid Cases		5690			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

1.5.4 $2 \times J$ tables

In this case the trend test, tests if the median scores differ for the two groups (given by X), when using the midrank scores this test is equivalent to the Wilcoxon Rank Sum Test (also called Mann-Whitney Test) In Stat 252 we said this test is only applicable if there are not too many ties. One has to be careful with the interpretation.

1.5.5 $I \times 2$ tables

In this case the response has only two outcomes (success/failures), and the question is if the ordinal variable has a positive or negative effect on the success rate of the response. In this case the test is called the Cochran-Armitage trend test.

Warning: Do not use this test if one of the variables is nominal and has more than 2 categories. In this case the test for trend is inappropriate.

1.6 Exact Tests (in Lab?)

1.7 Three Way Tables

For three categorical variables X, Y, and Z, where
Y - response variable
X - explanatory variable
Z - confounding variable (to be adjusted for)

Example 16
Y - response variable (cancer(yes, no))
X - explanatory variable (second hand smoking (none, some, lots))
Z - confounding variable (age(age groups))

One way of controlling for Z is by looking at tables for X and Y for each level of Z, called partial tables.

Example 17
Data from IMDb (January 11, 2011)
Y= rating for movies X = type of movie, horror(Alien) or animated movie (Wall-e), controlling for Z=sex (age group).
Male

	Rating					
Movie	1-2	3-4	5-6	7-8	9-10	Total
Alien	1502	1404	7090	49158	77097	136251
Wall-e	5654	2261	8199	43116	94079	153309
Total	7156	3665	15289	92274	171176	289560

Female

	Rating					
Movie	1-2	3-4	5-6	7-8	9-10	Total
Alien	430	356	1313	5075	6777	13951
Wall-e	1104	441	1274	5998	19106	27923
Total	1534	797	2587	11073	25883	41874

The partial tables describe the conditional joint distribution of X and Y (conditional on the different levels of Z).
Given that the rater was female the probability that a rater chose to rate Alien and rated it 9 or 10 equals 6777/41874.

movie * rating * sex Crosstabulation

sex				rating					Total
				12	34	56	78	910	
f	movie	Alien	Count	430	356	1313	5075	6777	13951
			% within movie	3.1%	2.6%	9.4%	36.4%	48.6%	100.0%
			% within rating	28.0%	44.7%	50.8%	45.8%	26.2%	33.3%
		Wall-e	Count	1104	441	1274	5998	19106	27923
			% within movie	4.0%	1.6%	4.6%	21.5%	68.4%	100.0%
			% within rating	72.0%	55.3%	49.2%	54.2%	73.8%	66.7%
	Total		Count	1534	797	2587	11073	25883	41874
			% within movie	3.7%	1.9%	6.2%	26.4%	61.8%	100.0%
			% within rating	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
m	movie	Alien	Count	1502	1404	7090	49158	77097	136251
			% within movie	1.1%	1.0%	5.2%	36.1%	56.6%	100.0%
			% within rating	21.0%	38.3%	46.4%	53.3%	45.0%	47.1%
		Wall-e	Count	5654	2261	8199	43116	94079	153309
			% within movie	3.7%	1.5%	5.3%	28.1%	61.4%	100.0%
			% within rating	79.0%	61.7%	53.6%	46.7%	55.0%	52.9%
	Total		Count	7156	3665	15289	92274	171176	289560
			% within movie	2.5%	1.3%	5.3%	31.9%	59.1%	100.0%
			% within rating	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Ignoring Z entirely results in the marginal table for X and Y. To obtain the marginal table just add the equivalent table entries of the partial tables.

Example 18

```
Rows: movie     Columns: rating

              12        34        56        78       910       All

 Alien      1932      1760      8403     54233     83874    150202
 Wall-e     6758      2702      9473     49114    113185    181232
 All        8690      4462     17876    103347    197059    331434

  Cell Contents -- Count
```

We can now use these tables to test for conditional independence and marginal independence.

Example 19
From SPSS:

Test	χ^2	df	P
conditional independence (female)	1793.102	4	< 0.001
conditional independence (male)	3778.475	4	< 0.001
marginal independence	4692.375	4	< 0.001

At significance level of 5% the data provide sufficient evidence for conditional association for men and women, and marginal association of movies and rating.

We can also do conditional and marginal trend tests (I used scores 1,2,3,4,5 for the rating): Use SPSS (1 = Alien, 2 = Wall-e) to find the correlations, and $M_0^2 = (r^2(n-1)$:

Test	r	M_0^2	df	P
trend (conditional female)	0.188	1480.0	1	< 0.001
trend (conditional male)	-0.024	166.79	1	< 0.001
marginal trend	-0.006	11.93	1	< 0.001

At significance level of 5% the data provide sufficient evidence that the median ratings for Alien and Wall-e conditional for women and men and combined are not the same.
The sign of r indicates that the median rating by women for Wall-e is higher than for Alien, and Alien is higher rated by men.
The marginal result indicates that the median rating for Alien is higher overall than for Wall-e. (We are missing an important fact here. This result is due to the fact that men are much more likely to give a rating on IMDb). These findings indicate that we have an interaction effect between sex and movie on rating, an appropriate model should give us more insight.

In the case that the marginal and conditional distributions are 2×2 tables we also can look at conditional and marginal odds ratios etc. (compare text book).

1.7.1 Simpson's Paradox

In a paper by Charig et al. (British Medical Journal, Clinical Research Ed., March 1986, 292(6524): 879-882) two different treatments for kidney stones are compared. Call the treatments A and B.
It is reported that treatment A is successful for 78% of patients and treatment B is successful for 83% of patients. So one would assume that treatment B is "better" than treatment A.
BUT they also report that for small kidney stones treatment A is successful for 93% of all cases, but B only in 87% of all cases
and that for large kidney stones treatment A is successful for 73% of all cases, but B only in 69% of all cases.
So treatment A is "better" than treatment B?

Treatment A

	Success	No Success	Total
small	81	6	87
large	192	71	263
Total	273	77	350

Treatment B

	Success	No Success	Total
small	234	36	270
large	55	25	80
Total	289	61	350

Treatment A is better than treatment B, because looking at the conditional distributions for treatment and success, given the size, treatment A shows a higher success rate for small as well as large kidney stones, i.e. 93% instead of 87%, and 73% instead of 69%.
How come, that overall treatment B seems to be the better treatment?
Both treatments are much more successful for small kidney stones than for large ones. Find that the sample for treatment B mostly consists of small kidney stones, but the sample for treatment A mostly consists of patients treated for large kidney stones, therefore showing a favourable overall success rate for treatment B.
As a result the overall success rate for treatment B is closer to the success rate for small kidney stone (looks good), but the overall success rate for treatment A is closer to the success rate for large kidney stones (looks not as good).

1.7.2 Homogeneous association

We speak of an homogeneous association between X and Y controlling for Z, when the association between X and Y is the same for all levels of Z. In ANOVA and multiple regression we said that X and Z do not interact in their effect on Y.
For categorical data means that the odds ratios (for each combination of 2 levels from X and two levels of Y) are not affected by Z.
How to test for homogeneous association we see later after building models.

1.8 Cochran-Mantel-Haenzel Test

Assume three variables, X, Y, Z, all being binary. (This test can be done for Z non-binary)
Let $\vartheta_{XY(k)}$ denote the conditional odds ratio between X and Y, given $Z = k$.

Then a homogeneous relationship between X, Y and Z implies that

$$\vartheta_{XY(1)} = \vartheta_{XY(2)} = \cdots = \vartheta_{XY(K)}$$

We speak of conditional independence between X and Y, given Z, iff

$$\vartheta_{XY(1)} = \vartheta_{XY(2)} = \cdots = \vartheta_{XY(K)} = 1$$

Cochran-Mantel-Haenzel Test for conditional independence

1. $H_0 : \vartheta_{XY(1)} = \vartheta_{XY(2)} = \cdots = \vartheta_{XY(K)} = 1$
   H_a : at least one conditional OR is different from 1
2. large random sample
3. $$\chi_0^2 = \frac{(\sum_k n_{11k} - \hat{\mu}_{11k})^2}{\sum_k Var(n_{11k})}, \quad df = 1$$
   with
   $$Var(n_{11k}) = \frac{n_{1+k} n_{2+k} n_{+1k} n_{+2k}}{n_{++k}^2 (n_{++k} - 1)}$$

4. P-value= $P(\chi^2 > \chi_0^2)$

Example 20
(Modified from example on Handbook of Biological Statistics)
McDonald and Siebenaller (1989) surveyed allele frequencies at the Lap locus in the mussel Mytilus trossulus on the Oregon coast. At two estuaries (estuary = the tidal mouth of a large river, where the tide meets the stream), they collected mussels from inside the estuary and from a marine habitat outside the estuary. There were three common alleles and a couple of rare alleles; based on previous results, the biologically interesting question was whether the Lap94 allele was less common inside estuaries, so all the other alleles were pooled into "non-94" class.
There are three nominal variables: allele (94 or non-94), habitat (marine or estuarine), and area (Tillamook, Yaquina, Alsea, or Umpqua). The null hypothesis is that at each area, there is no difference in the proportion of Lap94 alleles between the marine and estuarine habitats.
This table shows the number of 94 and non-94 alleles at each location. There is a smaller proportion of 94 alleles in the estuarine location of each estuary when compared with the marine location; we wanted to know whether this difference is significant.

Location	**Allele**	**Marine**	**Estuarine**
Tillamook	94	56	69
	non-94	40	77
Yaquina	94	61	257
	non-94	57	301

To test wether the Lap94 allele was less common inside estuaries will test for conditional independence between inside/outside estuary and Lap94 allele. If the test is significant, it indicates that the proportion of Lap 94 is different inside and outside of estuaries, and since in the samples data they are less common inside the estuaries, it would support the research hypothesis.

1. $H_0 : \vartheta_{XY(T)} = \vartheta_{XY(Y)} = 1$
   H_a : at least one conditional OR is different from 1
   $\alpha = 0.05$

2. Random samples have been collected and the samples are large enough (rule of thumb: at least 5 in each cell)

3. $$Var(n_{11T}) = \frac{125 \times 117 \times 96 \times 146}{241^2(242-1)} = 14.52$$
   $$Var(n_{11Y}) = \frac{218 \times 358 \times 118 \times 558}{667^2(667-1)} = 17.34$$
   $$\chi^2 = \frac{(\sum_k n_{11k} - \hat{\mu}_{11k})^2}{\sum_k Var(n_{11k})} = \frac{(56 - (96 \times 125)/242)^2 + (61 - (118 \times 318)/667)^2}{14.52 + 17.34} = \frac{41.13 + 22.49}{31.86} = 1.997$$
   $df = 1$

4. P-value= $P(\chi^2 > 1.997^2) = 0.16$

5. The null hypothesis can not be rejected

6. At significance level of 5% the data do not provide sufficient evidence that the proportion of Lap94 alleles is different in the marine and estuarine locations. (Comment: The original study included more locations, and together the data provided sufficient evidence that the proportions are different)