Contents

1 Analysis of Categorical Data

The techniques presented in previous sections cover the analysis of numerical data and variables. Up to now it is left unanswered how to analyze categorical variables and data.

1.1 Goodness-of-fit χ^2 Test

In this section the χ^2 test for comparing the relative frequency distribution from a sample with a given probability distribution is introduced.

Example: A company filling grass seed bags wants to evaluate their filling machine. The following distribution is advertised on the bags:

kind of seeds	proportion
K1	0.5
K2	0.25
K3	0.15
K4	0.05
K5	0.05

The company wants to check if the seed distribution in the bags fits the advertised distribution. They take a sample of size 1000 and find the following summarized data:

kind of seeds	count	relative frequency
K1	480	0.480
K2	233	0.233
K3	160	0.160
K4	63	0.063
K5	64	0.064

Notation: for a given categorical random variable

k = number of categories

The true distribution of the categorical variable in the population:

p_1 = true probability to fall in category 1
p_2 = true probability to fall in category 2
$\vdots$
p_k = true probability to fall in category k

And the claimed distribution:
p_{01} = claimed probability to fall in category 1
p_{02} = claimed probability to fall in category 2
$\vdots$
p_{0k} = claimed probability to fall in category k

Together:

Category	true probability	claimed probability
C1	p_1	p_{01}
C2	p_2	p_{02}
	$\vdots$	
Ck	p_k	p_{0k}

In order to check if indeed the claimed probabilities match the true probabilities, a random sample is drawn from the population and the observed frequencies for the k categories are compared with the hypothetical or claimed distribution. For the comparison the χ^2 goodness–of–fit statistic is used.

If we examine an event that occurs with probability p, then in a sample of size n we would expect to see this event about $n \cdot p$ times. The χ^2 statistic is based on this fact.

The χ^2 goodness–of–fit statistic

For a given hypothesized distribution $p_{01}, \ldots, p_{0k}$ define:
$E_1 = n \cdot p_{01}$ = expected cell count for category 1
$E_2 = n \cdot p_{02}$ = expected cell count for category 2
$\vdots$
$E_k = n \cdot p_{0k}$ = expected cell count for category k

Let the numbers from the sample be:
O_1 = observed cell count for category 1
O_2 = observed cell count for category 2
$\vdots$
O_k = observed cell count for category k

The χ^2 statistic is then:

$$\chi^2 = \sum_{\text{all categories}} \frac{(O_i - E_i)^2}{E_i}$$

If $(p_{01}, \ldots, p_{0k})$ is the true distribution of the categorical variable, then the χ^2–statistic is χ^2–distributed with df $= k - 1$ degrees of freedom,

Continue Example:
With $p_{01} = 0.5, p_{02} = 0.25, p_{03} = 0.15, p_{04} = 0.05, p_{05} = 0.05$ we get:

kind of seeds	advertised proportion	observed count	expected count	(0-E)^2/E
K1	0.5	480	500	(480-500)^2/500=0.8
K2	0.25	233	250	(233-250)^2/250=1.156
K3	0.15	160	150	(160-150)^2/150=0.666
K4	0.05	63	50	(63-50)^2/50=3.38
K5	0.05	64	50	(64-50)^2/50=3.92

And $\chi^2 = 0.8 + 1.156 + 0.666 + 0.98 + 0.72 = 9.9226$. This number now has to be evaluated as coming from a χ^2 distribution with 5-1=4 df.

The χ^2 Goodness–of–fit Test
Given a distribution $p_{01}, \ldots p_{0k}$

1. **Hypotheses:** $H_0 : p_1 = p_{01}, \ldots, p_k = p_{0k}$ versus $H_a : H_0$ is not true.

   Choose α

2. **Assumption:** Random sample form the population. The sample size is *large*. The sample size is large enough for the χ^2 test to be appropriate as long as every count is at least 5.

3. **Test statistic:**
$$\chi_0^2 = \sum_{\text{all categories}} \frac{(O_i - E_i)^2}{E_i}$$
   and $df = k - 1$.

4. **P-value:** $P(\chi^2 > \chi_0^2)$ the uppertail area for a χ^2 distribution with $k - 1$ df, found in Appendix Table VII.

5. **Decision:** As usual

6. **Context:** As usual

Continue Example:

1. **Hypotheses:** $H_0 : p_1 = 0.5, p_2 = 0.25, p_3 = 0.15, p_4 = 0.05, p_5 = 0.05$ versus $H_a : H_0$ is not true.

   $\alpha = 0.05$

2. **Assumption:** Random sample, check. The sample size is *large* enough.

3. **Test statistic:**(see calculations above) $\chi_0^2 = 9.9226$, and $df = 4$.

4. **P-value:**
   To find the p-value use the row for $df = 4$ in table VII.

   Find the two values that enclose χ_0^2:

   $9.488 \leq \chi_0^2 \leq 11.143$.

   From the title of the columns where you found those values in the table get that
$$0.025 \leq p - value \leq 0.050$$

5. **Decision:** $P - value \leq \alpha = 0.05$ and reject H_0.

6. **Context:**

   This means that at significance level of 5% the sample provides sufficient evidence that the filling machine is not working according to the advertised distribution.

1.2 χ^2 Test for Homogeneity and Independence in a Two-Way-Table

In this section the relationship/association of two **categorical** variables is studied. The best way of summarizing data on two categorical variables is a two-way-table.
Consider the following example.

Example:
A study has been conducted and resulted in the following (made up) data

		Smoker		
		yes	no	
heart disease	yes	23	15	38
	no	69	259	328
		92	274	366

Interpretation: E.g. 23 of the 366 people in this study were smoker and had heart disease.

The **joint distribution** of row and column variables gives the proportions of individuals that fall within the different cells created by row and column variable.

In the example:

		Smoker	
		yes	no
heart disease	yes	23/366	15/366
	no	69/366	259/366

The **marginal distributions** of the row and column variable in a two way table give the proportion of individuals that fall within the different categories of the row and the column variable.

In the example:
The marginal distribution of the variable "Smoker"

Smoker		
	yes	92/366=0.25
	no	274/366=0.75

The marginal distribution of the variable "Heart Disease"

Heart Disease		
	yes	38/366=0.10
	no	328/366=0.90

The study had been set up to study if smoking is a risk factor for heart disease. In order to study this question in the sample data, estimates for the following conditional probabilities are helpful:

$$\hat{P}(HD+|S+) = \frac{23}{92} = 0.25 \quad \text{and} \quad \hat{P}(HD+|S+) = \frac{15}{274} = 0.055$$

The proportion of smokers that had heart disease equals 0.25 versus a proportion of 0.055 of non smokers that did have heart disease.

The sample data seems to indicate that the risk for heart disease is much higher for smokers than for non-smokers.

In order to show that this is true for the entire population, a statistical test has to be conducted!

1. Comparing the distribution of a categorical variable between two or more populations.

Instead of postulating a distribution and comparing one sample with this distribution, as it was done for the goodness-of-fit test, here we will be looking at two or more samples from different populations and test if the distributions in both (or more) populations are the same. (In the example above the populations were smokers and non-smokers).

As an example suppose we are interested in the rate of sprouted seeds in two different kinds of water (rainwater, muddy water). The two categorical variables are Sprouted (yes or no) and water type (rainwater and muddy water).
Suppose 100 seeds were watered with rainwater and 100 seeds were watered with muddy water and then observed how many of those seeds sprouted.

The result could be the following table:

	sprouted		
	yes	no	total
rainwater	64	36	100
muddy water	74	26	100
Total	138	62	200

The two populations in this example are seeds watered with rainwater and seeds watered with muddy water. In this case we might be interested if the probability for sprouting in both populations is the same.

The null hypothesis to be tested in this kind of problem is that the distributions are homogeneous, they are equal for all populations.

H_0 : The probabilities for sprouting and nonsprouting is the same for both methods
H_a : The probabilities for sprouting and nonsprouting depend on the treatment

2. Testing two categorical variables for independence

When you study data that involves two variables, one important consideration is the relationship between the two variables. Does the proportion of measurements in the various categories for factor 1 depend on which category of factor 2 is being observed?

Example: 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 of 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

The question to be answered is, if the flu shot had an impact on the incidence of the flu.
In order to answer this question, a χ^2 test can be conducted, testing the following hypotheses:

H_0 : No relationship between treatment and incidence of flu
H_a : Incidence of flu depends on amount of flu treatment

The χ^2 Test for two way tables
The χ^2-test is for both sets of hypotheses the same.
The test statistic is similar to the one for the goodness-of-fit test, where the observed and the expected counts were compared.

The cell count for the sample will present the observed counts

So it remains to figure out how to calculate the expected count.
This is based on the definition of independence:

Two events, A and B are independent $\Leftrightarrow P(A \cap B) = P(A)P(B)$.

The expected cell count for the cell in row i and column j will be calculated by

$$E_{ij} = \frac{(row_i\ total)(column_j\ total)}{grand\ total}$$

Let

O_{ij} =the observed cell count in row i and column j

given by the sample.

The χ^2 Test for Homogeneity and Independence
Consider a row variable with r categories and a column variable with c categories.

1. **Hypotheses:**
   H_0 : The distribution of the variable of interest is homogeneous across the populations considered
   H_a : H_0 is not true
   or
   H_0 : The two categorical variables are independent
   H_a : H_0 is not true

2. **Assumption:** Random sample(s). The sample size is *large*. The sample size is considered large enough when every expected cell count is at least 5.

3. **Test statistic:**
   Compute for every cell of the two–way table

   $$E_{ij} = \frac{(row_i\ total)(column_j\ total)}{grand\ total}$$

   and then

   $$\chi_0^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

   df=$(r-1)(c-1)$

4. **P-value:** $P(\chi^2 > \chi_0^2)$ for a χ^2 distribution with $df = (k-1)(l-1)$ degrees of freedom, found in Table VII.

5. **Decision:** As usual

6. **Context:** As usual

Continue Flu Example:

1. **Hypotheses:**

   H_0 : Getting the flue and the number of shots are independent
   H_a : H_0 is not true

   $\alpha = 0.05$.

2. **Assumption:** Random sample(s), check. To see if the sample is large enough we have to wait for the calculation of the expected cell counts.

3. **Test statistic:** Calculate the expected cell count:

Flu/no vaccine:	(46·313)/1000=14.40
Flu/one shot:	(46·109)/1000=5.01
Flu/two shots:	(46·578)/1000=26.59
No Flu/no vaccine:	(954·313)/1000=298.60
No Flu/one shot:	(954·109)/1000=103.99
No Flu/two shots:	(954·578)/1000=551.41

   Observe that all expected counts are greater than 5, and the assumptions are met.

   Put observed and expected cell counts into the table:

	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

   Now add for all six cells $(O-E)^2/E$:
   $\chi^2 = 6.404 + 3.169 + 6.944 + 0.309 + 0.153 + 0.335 = 17.313$ and df=(3-1)(2-1)=2

4. **P-value:** $P - value = P(\chi^2 > 17.313)$ for a χ^2 distribution with $df = 2$ degrees of freedom, found in Table VII.

   $\chi_0^2 > \chi_{0.005}^2$, so that the $p - value < 0.005$.

5. **Decision:** Since $p - value < 0.005 < 0.05 = \alpha$, reject the null hypothesis

6. **Context:** At significance level of 5% the data provide sufficient evidence that the probability of getting the flu is associated with the number of shots an individual received.

At this point all we know is that there is evidence that the variables are not independent, but we do not know yet, if the vaccine increases or decreases the likelihood of getting the flu. To find out use the following estimated conditional probabilities:

$$\hat{P}(\text{flu} \mid \text{no vaccine})=24/313=0.077$$
$$\hat{P}(\text{flu} \mid \text{one-shot})=9/109=0.083$$
$$\hat{P}(\text{flu} \mid \text{two-shots})=13/578=0.022$$

These estimates show that the people with no vaccine and one shot have a comparable incidence of the flu, but the incidence decreases for people who received two shots.

Continue Heart Disease Example:
Conduct a χ^2test for homogeneity at a significance level of $\alpha = 0.05$.

1. Hypotheses: H_0 : The probability for heart disease is the same for smokers and non-smokers
   H_a : H_0 is not true.

   Choose α=0.05

2. Assumptions: random sample, check. Calculate the expected sell counts to decide if the assumptions are met.

3. Test Statistic:

   Calculate the expected cell counts:

   HD+/smoker+: (38·92)/366=9.55
   HD+/smoker-: (38·274)/366=28.45
   HD-/smoker+: (328·92)/366=82.45
   HD-/smoker-: (328·274)/366=245.55

   All expected cell counts are greater than 5, the assumptions are met.

   Put the observed and expected cell counts into the table:

	Smoker +	Smoker -
HD+	23	15
	9.55	28.45
HD-	69	259
	82.45	245.55

   For every cell calculate now

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

   and then add the results:
   $\chi_0^2 = 18.94 + 6.36 + 2.19 + 0.74 = 28.23$ and df=(2-1)(2-1)=1

4. Table VII shows that $\chi_0^2 > \chi_{0.0005}^2$, so that the p-value is less than 0.0005.

5. Since the P-value is less than $\alpha = 0.05$ reject the null hypothesis.

6. At significance level 0.05 we conclude that the probability for heart disease is different for smokers and non smokers.

The estimates for the conditional probabilities $P(HD+|S+)$ and $P(HD+|S-)$, show that smoking increases the risk for heart disease.

Example:
Some time ago there was a report in the news that an AIDS vaccine tested in Thailand did not show any effect. Let's redo the analysis.
The data quoted in the news is presented in the two–way table below (including the expected cell counts):

	Placebo	Vaccine	Total
HIV+	105	106	211
	105.5	105.5	
HIV-	1168	1167	2335
	1167.5	1167.5	
Total	1273	1273	2546

1. **Hypotheses:**

   H_0 : The probability for being tested HIV-positive is the same for those receiving a placebo and tho
   H_a : H_0 is not true

   $\alpha = 0.05$

2. **Assumption:** Random sample(s)? We don't know. All expected cell counts exceed 5, assumptions are met.

3. **Test statistic:**

   Calculate $(O-E)^2/E$ for all cells and add: $\chi^2 = 0.00237+0.00237+0.000214+0.000214 = 0.005168$ and df=(2-1)(2-1)=1

4. **P-value:** $P - value = P(\chi^2 > 0.005168)$ with $df = 1$

   According to table VII 0.005168 falls between $\chi^2_{0.975}$ and $\chi^2_{0.950}$, so that 0.950< p-value< 0.975.

5. **Decision:** $P - value > \alpha = 0.05$, do not reject the null hypothesis.

6. **Context:** The data do not provide sufficient evidence that the HIV infection–rate was impacted by the vaccine.