Chapter 1

Chi-square tests

> We have associations to things. We have, you know, we have associations to tables and to – and to dogs and to cats and to Harvard professors, and that's the way the mind works. It's an association machine.
>
> Daniel Kahneman

Chi-square (χ^2) test is used to compare the observed data with the data we expect to obtain according to a specific hypothesis. Through the comparison of observed versus expected data the above mentioned hypothesis is tested. Chi-square test is applied to *categorical variables*. Remind yourself that a categorical variable is the one which takes a finite number of substantively different values which can be ordered (like size characteristic: small, medium and big) or not ordered (like eyes' color: dark, blue, hazel, etc.).

Note that the word "Chi" is pronounced as "kai", just as the boy's name in "The Snow Queen" fairy tale.

There are two types of chi-square test: a goodness-of-fit test and a test of independence/association.

The **chi-square goodness-of-fit test** is used to check whether a variable is observed to fit some specific distribution. For example, one can test whether the evidence on birth statistics in Moscow support the hypothesis of equal sex distribution of newborn babies.

The **chi-square test of independence** (or test of association) allows to reveal whether there is an association between two variables in one population or to check whether distribution among different categories is the same in several populations (test for homogeneity). It answers questions like "Was surviving after the Titanic shipwreck independent of passenger class?" or "Do students and teachers have the same distribution of preferences about fizzy drinks (same set of proportions of those

who prefer Coke, Fanta and Sprite)?".

Goodness-of-fit test

As already mentioned, χ^2 goodness-of-fit test investigates whether an observed pattern of data *fits some given theoretical distribution*. Thus, the null hypothesis is that the categorical variable of interest takes values according to some specified rule. Alternative hypothesis just states the distribution is different, not equivalent to the one stated in H_0.

FACEBOOK VERSUS VKONTAKTE. Masha spends a lot of time in social networks. One day she noticed that she started using `facebook.com` more frequently and `vk.com` less frequently since the time she became an ICEF student. Masha decided to check whether ICEF students' preferences about social networks differ from the preferences shown by most young people in Russia.

According to a recent social research 65% of Russian youngsters prefer to use Vkontakte, 20% - Facebook, 10% - Odnoklassniki and the rest 5% prefer some other networks or do not use it at all. So, the distribution of network preferences in Russia is $p_{\text{vk}} = 0.65$, $p_{\text{fb}} = 0.2$, $p_{\text{ok}} = 0.1$ and $p_{\text{other}} = 0.05$ respectively.

To test whether ICEF students have the same distribution of preferences Masha interviewed 150 random students. She has got the following results: Vkontakte – 91 student, Facebook – 46 students, Odnoklassniki – 8 students, alternative/no use – 5 students.

So, can we say that ICEFers show the same preferences as other young people in Russia?

First, let's state the null and alternative hypotheses.

H_0: ICEF students show the following distribution of preferences for the social networks: 65% like Vkontakte most, 20% prefer Facebook, 10% choose Odnoklassniki and 5% have other preferences.

H_A: Social network preferences are *not* distributed this way (are distributed otherwise).

Second, we calculate expected frequencies under H_0 and check conditions.

As usually, we assume that H_0 is true. What should we then expect to observe? Number of students in a sample with a particular preference is a binomial random variable $\text{Binom}(n, p_i)$ where n is the sample size and p_i is the probability of this preference under H_0. For example, the number of students in the sample who likes Facebook most is distributed as Binom (150, 0.2) under H_0. As we've learned from Chapter 2 expectation of a binomial variable X is $E(X) = n \cdot p$. Therefore, the expected number of observations for each category under H_0 is $E_i = n \cdot p_i$.

This way we calculate all expected values: $E_{\text{vk}} = 150 \cdot 0.65 = 97.5$, $E_{\text{fb}} = 150 \cdot 0.2 = 30$, $E_{\text{ok}} = 150 \cdot 0.1 = 15$, $E_{\text{other}} = 150 \cdot 0.05 = 7.5$.

It is necessary that all expected values are greater than 5. This, as well as random sampling are necessary conditions for this test See explanation of that after the Step 3.

Note that expected number is not necessarily an integer. It can be 7.5 of students although we never truly expect to meet half of a student walking in university...

The next step is to calculate the statistic to compare Observed versus Expected – observed frequencies with what we expect to get under the true null hypothesis:

Favorite social network	`vk.com`	`facebook.com`	`ok.ru`	Other
Observed number of students (O_i)	91	46	8	5
Expected number of students (E_i)	97.5	30	15	7.5

Is the deviation of what we observe from what we expected too big, so that we would like to reject H_0?

To measure this χ^2_{st} statistic is used:

$$\chi^2(\text{df}) = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},\ \text{df} = k - 1$$

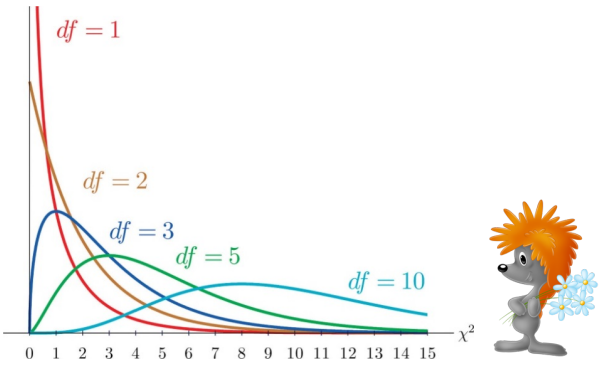
How does it work? First, we find the difference between observed and expected values $(O_i - E_i)$ for each category i. The larger are the overall deviations, the bigger is your desire to reject the null hypothesis. How to evaluate "overall deviations"? If we simply sum up all the deviations $(O_i - E_i)$ we'll get 0 (some deviations are positive and some are negative). So, to get the measure of overall differences between O_i and E_i we take squares of deviations $(O_i - E_i)^2$. The logic is the same as we've used in calculating variance of a discrete random variable. Taking squares not only allows to get a positive sum which increases with the size of deviations, but also "punishes" large deviations which are seen as especially undesirable. Now, we need to learn how large is the squared sum. Well, it depends on the mean amounts: deviation of 10 is much more significant if the mean amount is 5, than if it is 100...That's why each squared deviation is divided by the expected number of observations in the corresponding category: $\frac{(O_i - E_i)^2}{E_i}$.

If the number of observations is large enough the final sum $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ has **chi-square probability distribution** with $(k-1)$ degrees of freedom. k is the number of categories or the values the variable of interest can take. In our example $k = 4$.

What are the "large enough" requirements? The threshold used here is the same as for proportion analytics: $n \cdot p_i$ should be at least 5 for all i, or equivalently all expected numbers must satisfy that: $E_i \geqslant 5$. One more requirement for the test is that the *sample should be taken randomly* from the population. As we've mentioned above, both conditions are satisfied.

Conditions for Chi-square test:
Random sample
$E_i \geqslant 5$ for all i

Below are the graphs of χ^2 probability density function with different degrees of freedom. The distribution is always right-skewed.



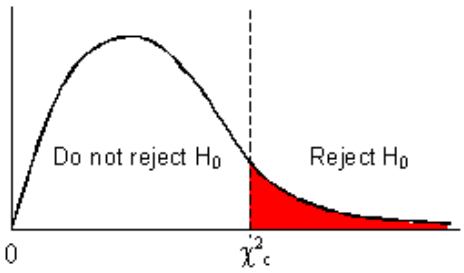
In our example $\chi^2_{\text{st}}(df) = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(3)$.

Why do we have $k-1$ degrees of freedom? Note that the observed frequencies are not independent as $n = O_1 + O_2 + O_3 + O_4$ or equivalently $O_4 = n - (O_1 + O_2 + O_3)$. Therefore, the last cell in the table does not provide any new information, which brings us to subtracting 1 from the number of categories, therefore d.f. $= (k-1)$.

Now, let's calculate the test statistic:

$$\chi^2_{\text{st}}(3) = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(91-97.5)^2}{97.5} + \frac{(46-30)^2}{30} + \frac{(8-15)^2}{15} + \frac{(5-7.5)^2}{7.5} \approx 13.067$$

Now we can test the hypothesis, calculate p-value and compare with α. Is 13.067 a typical value for $\chi^2(3)$ distribution or should it lead us to rejection of the hypothesis? To answer this question we should define the rejection area. First, we choose the significance level α. Let's take $\alpha = 0.05$. The more observations deviate from the null hypothesis, the larger is the sum $\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$, or the value of χ^2_{st}. Smaller χ^2 implies better fit, high value of χ^2 motivates to reject H_0. Therefore, rejection area is always located in the right tail of the distribution:



p-value is the probability of obtaining χ^2_{st} as extreme as the one obtained or even larger given H_0 is true.

p-value $= P(\chi^2(3) > \chi^2_{\text{st}}(3)) = P(\chi^2(3) > 13.067) \approx 0.0045$.

Since p-value < 0.05, the statistic is in the rejection area and we reject H_0 at 5% significance level.

Alternatively to using p-value, we can calculate the critical value of chi-square distribution with three degrees of freedom $\chi^2(3)$ and compare it with $\chi^2_{\text{st}}(3)$. Here $\alpha = 0.05$, therefore $\chi^2_{\text{critical}}(3)$ is such that: $P(P(\chi^2(3) > \chi^2_{\text{cr}}(3)) = 0.05$. This value could be found in calculator or taken from the table: $\chi^2_{\text{cr}} = \chi^2_{0.05}(3) \approx 7.815$. Again,

since $\chi^2_{\text{st}} > \chi^2_{\text{cr}}$, χ^2_{st} gets into the rejection area and H_0 is to be rejected in favor of H_A.

Finally, we state the conclusion. Let's return to the Masha's question. We've rejected the null hypothesis that ICEF students show the same network preferences as other young people in Russia. The conclusion is that ICEF students are in fact different from the others in this respect.

Thus, this type of problems is solved in five steps. Step 1 is to formulate null and alternative hypotheses. Step 2 is to calculate expected frequencies and check the conditions. At step 3 we write the formula and calculate chi-square statistic. At step 4 we calculate p-value and compare it with significance level. Finally, at step 5 we state the conclusion.

What is the Chi-Square random variable?
In fact it is the sum of $(k-1)$ independent squared standard normal variables $\sum_{i=1}^{k-1} z_i^2$. What does it mean?

Imagine a machine providing three random values each second, such that each value is independent of the two others and is produced according the standard normal law $N(0,1)$. If you square these values and sum them up each second – you'll get the sequence of values, which follow the chi-square distribution with 3 degrees of freedom.

For the proof that $\chi^2(k-1) = \sum_{i=1}^{k-1} z_i^2$ see the "Top secret section" at the end of this chapter.

Note that chi-square random variable fits the CLT! It is the sum of independent identically distributed variables. Then, given that k approaches infinity, chi-square distribution can be approximated by normal.

Chi-square test of independence

> If two things are independent they don't know of each other.
>
> Unknown author

Chi-square test of independence allows to test the hypothesis that two characteristics of population are independent or to check whether distribution among different categories is the same in several populations (test for homogeneity).

Chi-square test of independence can give answer to questions like: "Is getting an offer from McKinsey independent of subjects studied in university?", "Do ICEF students' grades depend on gender?".

The test procedure suggests the use of the so-called **contingency table**. It displays frequency distribution of the two variables. It differs from joint distribution because instead of giving true probabilities of intersections of values of X and Y, it shows their observed *frequencies* (how many times it happened that $X = x_i$ and $Y = y_i$).

BUILDING A CAREER. Masha dreams of working as a consultant for McKinsey. She had recently attended their lecture, where she heard an opinion that technical specialties graduates are more likely to get

an offer. Masha decided to investigate if there is any reason to worry (note that Economics is not a technical specialty, but belongs to humanities!). So, she decided to test the statement.

Masha has got the following data about acceptance and school major of 570 applicants to McKinsey:

	Economics	Engineering	Math/IT
Got an offer	14	12	6
Were rejected	316	128	94

Step 1. State the hypotheses.

H_0: The proficiency subject (school major) and getting an offer from McKinsey are independent.

H_A: The proficiency subject and McKinsey offer are not independent.

Step 2. Calculate the expected frequency (under H_0). Check conditions.

1. First, calculate the row and column totals:

	Economics	Engineering	Math/IT	Total
Got an offer	14	12	6	32
Was rejected	316	128	94	538
Total	330	140	100	570

2. Then, calculate the expected values as $\frac{n_i \cdot n_j}{n}$.Check conditions.

   Simple way to calculate expected value for each cell is to multiply the corresponding row and column totals and divide them by grand total. For example, the expected number of applicants who gets an offer and whose major was Economics is $\frac{n_{\text{offer}} \cdot n_{\text{economics}}}{n} = \frac{32 \cdot 330}{570} \approx 18.5$.

Expected	Economics	Engineering	Math/IT	Total
Got an offer	$\frac{32 \cdot 330}{570} \approx 18.5$	7.9	5.6	32
Was rejected	311.5	132.1	94.4	538
Total	330	140	100	570

We should always check that: $E_{ij} \geqslant 5$. In our example the requirement is held. We also assume that the sample was taken randomly.

Why do we use the formula $\frac{n_i \cdot n_j}{n}$? Again, it is calculated as expectation of a binomial random variable: by multiplying n by p_{ij}. Here p_{ij} is the probability of a joint event: $X = x_i$ and $Y = y_i$ given the true null hypothesis. Null hypothesis states that getting the offer (X) and having a specific major (Y) in university are independent. For independent variables X and Y the following holds: $P(X = x_i \cap Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$ for all pairs of i and j. Marginal probabilities $P(X = x_i)$

and $P(Y = y_j)$ are estimated by relative frequencies of events $X = x_i$ and $Y = y_j$ correspondingly: $P(X = x_i) = \frac{n_i}{n}$, $P(Y = y_j) = \frac{n_j}{n}$. Thus, under null hypothesis $P(X = x_i \cap Y = y_j) = \frac{n_i}{n} \cdot \frac{n_j}{n}$. Now, to get expected value we multiply this probability by n: $E_{ij} = \frac{n_i}{n} \cdot \frac{n_j}{n} \cdot n = \frac{n_i n_j}{n}$.

$$E_{ij} = \frac{n_i \cdot n_j}{n}$$

Step 3. Calculate chi-square statistic (compare Observed versus Expected).
Now we can calculate the chi-square statistic:

$$\chi^2_{\text{st}}(\text{df}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \ \text{df} = (r-1)(c-1)$$

Basically, the formula is the same as we've used for goodness-of-fit test: it is the sum of squared deviations divided by the expected values. However, since the sum includes the values in rows as well as in columns the formula involves double summation operator. The degrees of freedom here are $(r-1)(c-1)$, where r is the number of rows and c number of columns. Values in cells are dependent because marginal probabilities in cells and rows sum up to 1: $\sum_{i=1}^{r} p_i = 1$, $\sum_{j=1}^{c} p_j = 1$ (so are the numbers in cells: they should sum up to n). Therefore, once you know marginal probabilities in $(r-1)$ rows and $(c-1)$ columns you can calculate the last two marginal probabilities. Under H_0 (remember that statistics are calculated under null hypothesis) probabilities of intersections are determined as the product of corresponding marginal probabilities. Therefore, contents of $(r-1)(c-1)$ cells completely determine the numbers in the remaining cells, the latter are not free to vary once the former is given. So, we have $(r-1)(c-1)$ units of "data free to vary", and this number determines degrees of freedom of two-way chi-square statistic[1].

Therefore, $\text{df} = (r-1)(c-1) = (2-1)(3-1) = 2$.

$\chi^2_{\text{st}}(2) = \frac{(14-18.5)^2}{18.5} + \frac{(12-7.9)^2}{7.9} + \frac{(6-5.6)^2}{5.6} + \frac{(316-311.5)^2}{311.5} + \frac{(128-132.1)^2}{132.1} + \frac{(94-94.4)^2}{94.4} \approx 3.51$

Step 4. Calculate p-value and compare it with α (test the hypothesis).
Let's take $\alpha = 0.05$.
p-value $= P(\chi^2(2) > 3.51) \approx 0.17 > 0.05$.
Alternatively: critical value $\chi^2_{0.05}(2) \approx 5.99 > \chi^2_{\text{st}}(2) = 3.51$.

Step 5. State the conclusion.
The conclusion is that there is not enough evidence to reject H_0 that proficiency subject and McKinsey offer are independent. In other words, there is no evidence of dependence between success and university major. Thus, there is nothing for Masha to worry about, just study hard, crack the cases and apply!

Note that even if we've rejected the null hypothesis of independence that would not allow us to state any causal relationship. For example, having conducted chi-square test you can never resume: "Eating applies reduces dental caries". It is also not possible to conclude that: "Dental caries is significantly lower for those who eat apples". The test only answers the question of whether the two characteristics are independent. If they are *not*, it does not provide you with any information about

[1]For a funny explanation of what are degrees of freedom see: `http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-degrees-of-freedom-in-statistics`

type of the dependence revealed and about the significance of differences between categories. If you'd like to know what kind of dependence is there in place you need to apply other techniques, such as regression analysis based on experimental data (to make conclusions about dependence). In order to conclude about significance of revealed difference between categories you can use 2-sample z-tests (comparison of proportions among those who eat and do not eat apples).

Chi-square test for homogeneity has the same mechanism and in fact addresses the same question of independence but the question formulation is different.

Chi-square test for homogeneity allows to test whether a characteristic is equivalently distributed (has the same *proportions of categories*) in two or more populations. However, it is essentially equivalent to the test of independence. Asking whether a variable is homogenously distributed in different populations is the same as asking whether distribution of this variable is *independent* of belonging to specific population. So, if you meet questions like: "Are teachers of different subjects have the same distribution of levels of happiness?" or "Do boys and girls have the same preferences for vocation choices?" (homogeneity formulation) you can always reformulate them in terms of independency of two factors and apply the same method. In this case the questions would sound like "Is the level of happiness independent of the subject taught by a teacher?" and "Are vocation preferences independent of gender?" (independence formulation). If you define belonging to a population as another random variable you'd call the procedure the test of independence.

It is totally a question of what you define as a population. Therefore, one might call the test for homogeneity as the subtype of the test of independence.

ЗДЕСЬ НУЖНА КАРТИНКА ЗЛОЙ ХУДЕЮЩЕЙ МАШИ

BROCCOLI... OH, AGAIN!. A week ago Masha ordered broccoli in university canteen. She was really dissatisfied with its quality, because it contained a lot of oil, while Masha was on a diet. Being highly interested both in social problems in university and Statistics, Masha decided to test if the distribution of those who are satisfied with the canteen is the same among students, teachers and administrative stuff. She took random samples from each of the three groups and discussed the topic with them. The results were: 166 out of 200 students, 92 out of 100 teachers, 44 out of 50 administration members answered that food is good enough.

Step 1. State the hypotheses.

H_0: Students, teachers and administration have the same distribution of opinions about the food in canteen.

H_A: Students, teachers and administration do not have the same distribution of opinions about the food in canteen.

Step 2. Calculate expected values. Check conditions

Let's assume the null hypothesis is true. Then, proportions of satisfied students, teachers and stuff are the same and equal to some p. So, expected number of those satisfied in each category i is $p \cdot n_i$. p is the marginal probability of satisfied among all people $\frac{n_{\text{satisfied}}}{n}$. So, the expected number of satisfied students is $\frac{n_{\text{satisfied}}}{n} \cdot n_{\text{students}}$.

Thus, the formula for expected values is the same: $E_{\text{ij}} = \frac{n_i \cdot n_j}{n}$.

This way we get the following results (values in brackets indicate expected values):

	Satisfied	Not satisfied	
Students	166 (172.6)	34 (27.4)	200
Teachers	92 (86.3)	8 (13.7)	100
Administration	44 (43.1)	6 (6.9)	50
	302	48	350

After calculating expected values we should check that each value is no less than 5: $E_{ij} \geqslant 5$ for all i, j. In our example the requirement holds. We are also told that the three samples were random.

Step 3. Calculate chi-square statistic (compare Observed versus Expected).

Compute the test statistic $\chi^2_{\text{st}}(\text{df}) = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(O_{\text{ij}} - E_{\text{ij}})^2}{E_{\text{ij}}}$, $\text{df} = (3-1)(2-1) = 2$

$$\chi^2_{\text{st}}(2) = \frac{(166 - 172.6)^2}{172.6} + \frac{(34 - 27.4)^2}{27.4} + \ldots + \frac{(6 - 6.9)^2}{6.9} \approx 4.708$$

Step 4. Calculate p-value and compare it with α (test the hypothesis).

Let's take $\alpha = 0.05$.

p-value $= P(\chi^2(\text{df}) > \chi^2_{\text{st}}) = P(\chi^2(2) > 4.708) \approx 0.095 > 0.05$

Equivalently, $\chi^2_{\text{st}} = 4.7$ is smaller than critical value $\chi^2_{0.05}(2) = 5.99$.

Step 5. State the conclusion.

So, there is not enough evidence to reject the null hypothesis at 5% significance level. We conclude that students, teachers and administration have the same distribution of opinions about the canteen.

You must be able to reproduce even being half-awake

- Goodness-of-fit test checks *whether population fits some* assumed distribution.

$$\chi^2_{\text{st}}\,(\text{df}) = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \text{df} = k - 1$$

- Independence test checks for association between two variables in one population. It can be reformulated as the Homogeneity test, which compares distributions of a categorical variable in two or more populations.

$$\chi^2_{\text{st}}(\text{df}) = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},\ \text{df} = (r{-}1)(c{-}1),\ E_{ij} = \frac{n_i \cdot n_j}{n}$$

- In any Chi-square test you should check that: sample is random, $E_i \geqslant 5$ for all i.

Full score strategy

- Step 0. Identify and name the test.
- Step 1. State the hypotheses.
- Step 2. Calculate expected values. Check conditions.
- Step 3. Write down the formula and calculate chi-square statistic.
- Step 4. Calculate p-value and compare it with α
- Step 5. State the conclusion.

Calculator BOX

1. Goodness-of-fit test:

   - Put observed values and expected values into the 2 lists in the Stat section
   - `TEST` → `CHI` → `GOF` →

     `Observed`: indicate the list with O_i

     `Expected`: indicate the list with E_i

     `df`: enter the relevant degrees of freedom $(k-1)$

     `CNTRB`: indicate the list where you'd like to find the values of contribution of each category to the value of statistic: $(O_i - E_i)^2/E_i$
   - → `EXE`

2. Test of independence:

   - `TEST` → `CHI` → `2WAY` →

     Observed: indicate the matrix (table) where you will enter the observed values

     Expected: indicate the matrix (table) where you'd like to find expected values calculated
   - → `F2`, matrix menu → `DIM`, indicate matrix dimension →

     `m`: number of rows

     `n`: number of columns
   - → `EXE` (Enter the observed values into the corresponding matrix) → `Exit` → `Exit` → `EXE`

   If you now come back to the matrix of expected values you'll find the values E_{ij} calculated.

Top secret information

We have stated that chi-square statistic with $(k-1)$ degrees of freedom is the sum of $(k-1)$ squared standard normal variables: $\chi^2(k-1) = \sum_{i=1}^{k-1} z_i^2$

Let's prove this for the case of just 2 categories in X, $k=2$. Let X be the number of students who passed the AP exam in a sample of n students. Let's test the null hypothesis that probability of a randomly chosen student to pass is p:

H_0: ICEF students show the following results on AP: proportion p passes the exam, and proportion $(1-p)$ fails the exam.

H_A: ICEF students show different distribution of passes.

Under H_0 expected number of students who passed the exam is $E(X) = np$, expected number of fails is $E(n-X) = n(1-p)$. Recall that when sample is large $(np > 5, n(1-p) > 5)$, binomial variable is approximately normally distributed by CLT, so that $X \sim N(np, np(1-p))$ approximately. So, subtracting mean and dividing by standard deviation we get standard normal variable z:

$$\frac{X - np}{\sqrt{np(1-p)}} \approx z \sim N(0,1)$$

Now let's calculate the test statistic:

$$\text{(1.1)} \quad \chi^2_{\text{st}}(\text{df}) = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} =$$

$$= \frac{(X-np)^2}{np} + \frac{((n-X) - n(1-p))^2}{n(1-p)} = \frac{(1-p)(X-np)^2 + p((n-X) - n(1-p))^2}{np(1-p)} =$$

$$= \frac{(1-p)(X-np)^2 + p(n-X-n+np)^2}{np(1-p)} = \frac{(1-p)(X-np)^2 + p(X-np)^2}{np(1-p)} =$$

$$= \frac{(X-np)^2(1-p+p)}{np(1-p)} = \frac{(X-np)^2}{np(1-p)} = \left(\frac{X-np}{\sqrt{np(1-p)}}\right)^2 \approx z^2$$

Thus, we've shown that for $k=2$ random variable $\chi^2_{\text{st}}(k-1)$ is "the sum" of $k-1=1$ squared standard normal variable. The proof for larger number of categories is much more complicated.

Sample AP problems with solutions

Problem 1. AP 2013 №4

You'll have about 13 minutes to solve this problem. It will bring you 15% of score for Free Response section.

The Behavioral Risk Factor Surveillance System is an ongoing health survey system that tracks health conditions and risk behaviors in the United States. In one of their studies, a random sample of 8,866 adults answered the question "Do you consume five or more servings of fruits and vegetables per day?" The data are summarized by response and by age-group in the frequency table below.

Age	Yes	No	Total
18–34	231	741	972
35–54	669	2,242	2,911
55 or older	1,291	3,692	4,983
Total	2,191	6,675	8,866

Do the data provide convincing statistical evidence that there is an association between age-group and whether or not a person consumes five or more servings of fruits and vegetables per day for adults in the United States?

Solution

Step 0. Identify the test. The appropriate test is a chi-square test of independence.

Step 1. State the hypothesis.

H_0: Fruit and vegetable consumption is independent of age group for the population of adults in the United States.

H_A: Fruit and vegetable consumption is not independent of age group for the population of adults in the United States. OR

$H_0 : p_1 = p_2 = p_3$, where p_i is the population proportion of adults who eat at least 5 servings of fruits and vegetables among people in '18-24', '35-54' and '55+' age categories correspondingly.

H_A: At least one of the population proportions p_1, p_2, p_3, differs from the other two.

Step 2. Calculate expected values E_i. Check conditions

We calculate expected values as $E_{\text{ij}} = \frac{n_i \cdot n_j}{n}$.

Results are provided in the table below (expected values are in brackets):

Age	Five or more servings of fruit and vegetables	Four or fewer servings of fruit and vegetables	Total
18–34	231 (240.2)	741 (731.8)	972
35–54	669 (719.4)	2,242 (2191.6)	2,911
55 or older	1,291 (1231.4)	3,692 (3751.5)	4,983
Total	2,191	6,675	8,866

The conditions for this test are satisfied because:

1. It is stated that the sample was randomly selected.
2. $E_{\text{ij}} = \frac{n_i \cdot n_j}{n} \geq 5$ for all i, j.

Step 3. Write down the formula and calculate the chi-square statistic
$\chi^2\,(\text{df}) = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 8.983$, $\text{df} = (3-1)(2-1) = 2$.
Step 4. Calculate p-value and compare it with α.
Let's choose $\alpha = 0.05$.
p-value $= P(\chi^2(2) \geqslant 8.983) \approx 0.011 < 0.05$
Step 5. State a conclusion.
Since p-value $< \alpha$, we reject the null hypothesis at 5% significance level.
Thus, the data provides strong statistical evidence of association between age group and consumption of fruits and vegetables for adults in the United States.

Problem 2. AP 2008 №5

You'll have about 13 minutes to solve this problem. It will bring you 25% of score for Free Response section.

A study was conducted to determine where moose are found in a region containing a large burned area. A map of the study area was partitioned into the following four habitat types.

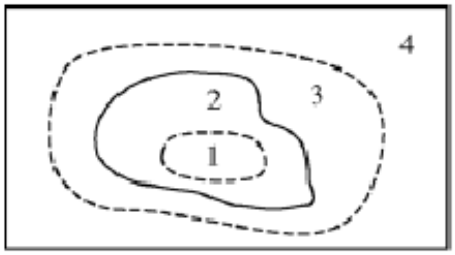
(1) Inside the burned area, not near the edge of the burned area,

(2) Inside the burned area, near the edge,

(3) Outside the burned area, near the edge, and

(4) Outside the burned area, not near the edge.

The figure below shows these four habitat types.



Note: Figure not drawn to scale.

The proportion of total acreage in each of the habitat types was determined for the study area. Using an aerial survey, moose locations were observed and classified into one of the four habitat types. The results are given in the table below.

Habitat type	Proportion of Total Acreage	Number of Moose Observed
1	0.340	25
2	0.101	22
3	0.104	30
4	0.455	40
Total	1.000	117

(a) The researchers who are conducting the study expect the number of moose observed in a habitat type to be proportional to the amount of acreage of that type of habitat. Are the data consistent with this expectation?

Conduct an appropriate statistical test to support your conclusion. Assume the conditions for inference are met.

(b) Relative to the proportion of total acreage, which habitat types did the moose seem to prefer? Explain.

Solution

(a) **Step 0. Identify the test.** The appropriate test is a chi-square goodness-of-fit test.

Step 1. State the hypothesis.

H_0: $p_1 = 0.34$, $p_2 = 0.101$, $p_3 = 0.104$, $p_1 = 0.455$ are the proportions of moose in the four habitats.

H_A: moose have some other preferences with respect to habitat hypes.

Step 2. Calculate expected values E_i. Check conditions

We calculate expected values as $E_i = np_i$, $n = 117$.

Results are provided in the table below:

Habitat type	Number of Moose Observed	Number of Moose Expected
1	25	39.78
2	22	11.817
3	30	12.168
4	40	53.235

The conditions for this test are said to be met.

Step 3. Write down the formula and calculate the chi-square statistic.

$$\chi^2\,(\text{df}) = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(25 - 39.78)^2}{39.78} + \ldots + \frac{(40 - 53.325)^2}{53.325} \approx$$
$$\approx 5.49 + 8.77 + 26.13 + 3.29 \approx 43.689,\ \text{df} = (4-1) = 3$$

Step 4. Calculate p-value and compare it with α.

Let's choose $\alpha = 0.01$.

p-value $= P(\chi^2(3) \geqslant 43.689) \approx 1.75 \cdot 10^{-9} < 0.01$.

Step 5. State a conclusion.

Since p-value $< \alpha$, we reject the null hypothesis at 1% significance level.

Thus, the data is not consistent with the claim that the number of moose observed in a habitat type is proportional to the amount of acreage of that type of habitat.

(b) As can be seen from calculation of chi-square statistic, the largest contribution is done by the 3rd area (26.13), where observed number is much higher than expected ($30 > 13.168$). Moose also seem to prefer 2nd habitat more than others. Again, observed number exceeds expected in that habitat.

Problem 3. AP 2006 №6

You'll have about 25 minutes to solve this problem. It will bring you 25% of score for Free Response section.

A manufacturer of thermostats is concerned that the readings of its thermostats have become less reliable (more variable). In the past, the variance has been 1.52 degrees Fahrenheit (F) squared. A random sample of 10 thermostats are given in the table below.

Thermostat	1	2	3	4	5	6	7	8	9	10
Temperature (F)	66.8	67.8	70.6	69.3	65.9	66.2	68.1	68.6	67.9	67.2

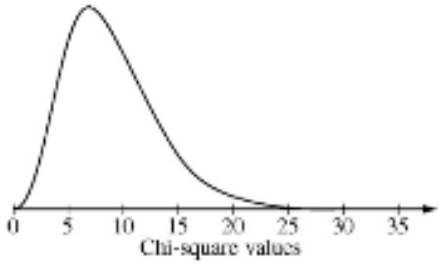
(a) State the null and alternative hypothesis that the manufacturer is interested in testing.

> It can be shown that of the population of thermostat temperatures are normally distributed, the sampling distribution of $\frac{(n-1)s^2}{1.52}$ follows a chi-square distribution with $(n-1)$ degrees of freedom.

(b) Calculate the value of $\frac{(n-1)s^2}{1.52}$ for these data.

(c) Assume that the population of thermostat temperatures follows a normal distribution. Use the test statistic $\frac{(n-1)s^2}{1.52}$ from part (b) and the chi-square distribution to test the hypothesis in part (a).

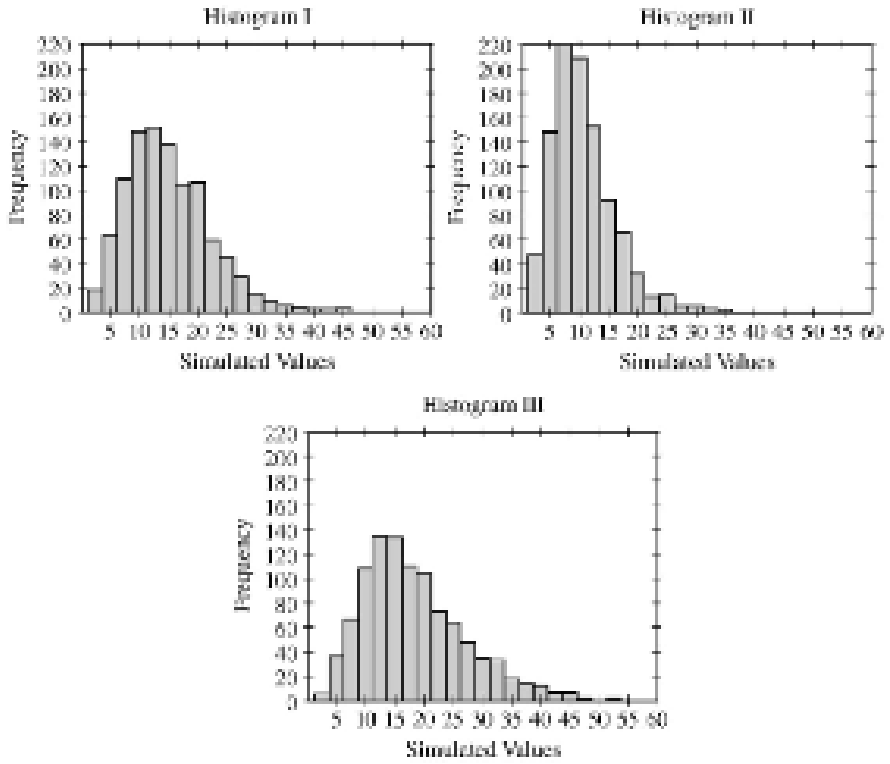
(d) For the test conducted in part (c), what is the smallest value of the test statistic that would have led to the rejection of the null hypothesis at the 5 percent significance level?



Mark this value of the test statistic on the graph of the chi-square distribution below. Indicate the region that contains all of the values that would have led to the rejection of the null hypothesis.

(e) Using simulation, 1,000 samples, each of size 10, were randomly generated from 3 populations with different variances. Each population was normally distributed with mean 68 and variance greater than 1.52. The histograms below show the simulated sampling distribution of $\frac{(n-1)s^2}{1.52}$ for each population.

Mark the region identified in part (d) on each of the histograms below.



(f) Based on the regions that you marked on part (e), identify the simulated sampling distribution that corresponds to the population with the largest variance. Then

identify the simulated sampling distribution that corresponds to the population with the smallest variance. Justify your choices.

Solution

(a) Based on the information stated in the conditions, these are the hypotheses that the manufacturer is interested in testing.

H_0: variance of the thermostat's readings is 1.52 degrees Fahrenheit (F) squared

H_A: variance of the thermostat's readings is more than 1.52 degrees Fahrenheit (F) squared

(b) Based on the given data the sample variance is $s^2 = \frac{\sum_{i=1}^{10}(x_i-\overline{X})^2}{10-1} \approx 2.038$. Thus, $\frac{(n-1)s^2}{1.52} \approx \frac{9 \cdot 2.038}{1.52} \approx 12.068$.

(c) Since we are given that under condition that temperatures are normally distributed, the sampling distribution of $\frac{(n-1)s^2}{1.52}$ follows a chi-square distribution with $n-1$ degrees of freedom, then, we can calculate the test-statistic as: $\chi^2(9) = \frac{(n-1)s^2}{1.52} \approx 12.068$. Let's choose $\alpha = 0.05$.

p-value can be calculated as: p-value $= P(\chi^2(9) \geqslant 12.068) \approx 0.210 > 0.05$. Since p-value $> \alpha$, we do not reject the null hypothesis at 10% significance level.

Thus, the data does not provide enough statistical evidence to claim that thermostats have become more variable.

(d) The value of the chi-distribution with 9 degree of freedom which cuts off 5% of highest values is $\chi^2_{0.05}(9) \approx 16.919$. For any value of chi-square statistic above it p-value would be smaller than 0.05 and thus, H_0 would be rejected. Therefore 16.919 is the smallest value of the test statistic that would have led to the rejection of the null hypothesis at the 5 percent significance level. The region under the pdf curve to the right of 16.919 must be indicated on the graph.

(e) Indicate approximately the regions to the right of 16.92 on each of the histograms.

(f) As can be seen from the pictures in part (e) the largest rejection area is on histogram III, and the smallest one – on histogram II. Thus, we would be most likely to reject H_0 and conclude that variance is above 1.52 with the results of simulation III, and least likely to do that with the results of simulation II. Therefore, we conclude that simulation II and simulation III produce distributions with the smallest and largest values of variance, correspondingly.

Problem 4. AP 2002 Form B №6

You'll have about 17 minutes to solve this problem. It will bring you 20% of score for Free Response section.

In September 1990, each student in a random sample of 200 biology majors at a large university was asked how many lab classes he or she was enrolled in. The sample results are shown below.

Number of lab classes	Number of students
0	28
1	62
2	58
3	28
4	16
5	8
(Total)	200

$$\bar{x} = 1.83,\ s = 1.29$$

To determine whether the distribution has changed over the past 10 years, a similar survey was conducted in September 2000 by selecting a random sample of 200 biology majors. Results from the year 2000 sample are shown below.

Number of lab classes	Number of students
0	20
1	72
2	60
3	10
4	26
5	12
(Total)	200

$$\bar{x} = 1.93,\ s = 1.37$$

(a) See chapter 11.

(b) Does the test in (a) address the question of whether the distribution of number of lab classes was different in 2000 than it was in 1990? If so, explain your reasoning. If not, carry out an appropriate statistical test using $\alpha = 0.10$ to answer this question.

(c) Use the results of your analyses in (a) and (b) to write a few sentences that summarize how the distribution of the number of lab classes did or did not differ. Use appropriate graphs to help communicate your message. This summary should be understandable to someone who has not studied statistics.

Solution

(a) No, the tests only address the difference in average number of classes, not the distribution of these numbers. In order to check whether the distribution of number of lab classes was different in 2000 than it was in 1990 we need a chi-square homogeneity test, or, equivalently a test of independence of number of classes of the year of studies.

Step 1. State the hypothesis

H_0: Number of classes in which a student enrolled was equally distributed in 1990 and 2000

H_A: Otherwise

Step 2. Calculate expected values E_i. Check conditions

We calculate expected values as $E_{ij} = \frac{n_i \cdot n_j}{n}$.

Results are provided in the table below (expected values are in brackets):

Number of lab classes	Year 1990	Year 2000
0	24	24
1	67	67
2	59	59
3	19	19
4	21	21
5	10	10
(Total)	200	200

The conditions for this test are satisfied because:

(a) It is stated that the sample was randomly selected.

(b) $E_{ij} = \frac{n_i \cdot n_j}{n} \geqslant 5$ for all i, j.

Step 3. Write down the formula and calculate the chi-square statistic

$\chi^2\,(\text{df}) = \sum_{i=1}^{6} \sum_{j=1}^{2} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \approx 13.821$, $\text{df} = (6-1)(2-1) = 5$.

Step 4. Calculate p-value and compare it with α

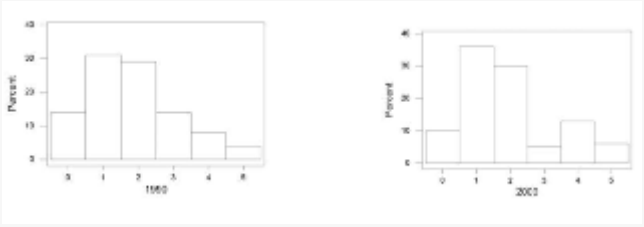
p-value $= P(\chi^2(5) \geqslant 13.821) \approx 0.017 < \alpha = 0.05$.

Step 5. State a conclusion.

Since p-value $< \alpha$, we reject the null hypothesis at 5% significance level.

Thus, the data provides strong statistical evidence that the distribution of number of lab classes in 2000 and in 1990 was different.

(b) Below are the histograms of distributions of number of classes in 1990 and 2000.



It can be seen that, although they seem to be centered at the same number, shapes of distributions differ. Distribution of 2000 has much less students enrolled in 3 classes than students in 1990.

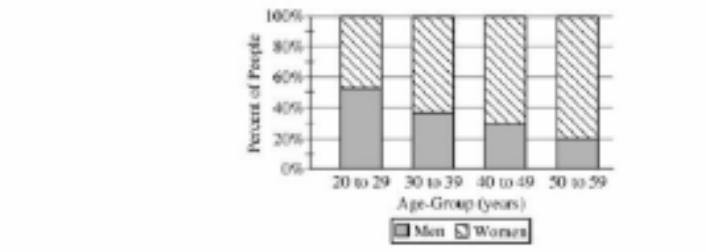
Practice AP problems

Problem 1. AP 2017 №5

The table and the bar chart below summarize the age at diagnosis, in years, for a random sample of 207 men and women currently being treated for schizophrenia.

Age-Group (years)

	20 to 29	30 to 39	40 to 49	50 to 59	Total
Women	46	40	21	12	119
Men	53	23	9	3	88
Total	99	63	30	15	207



Do the data provide convincing statistical evidence of an association between age-group and gender in the diagnosis of schizophrenia?

Problem 2. AP 2016 №2

You'll have about 13 minutes to solve this problem. It will bring you 15% of score for Free Response section.

Product advertisers studied the effects of television ads on children's choices for two new snacks. The advertisers used two 30-second television ads in an experiment. One ad was for a new sugary snack called Choco-Zuties, and the other ad was for a new healthy snack called Apple-Zuties.

For the experiment, 75 children were randomly assigned to one of three groups, A, B, or C. Each child individually watched a 30-minute television program that was interrupted for 5 minutes of advertising. The advertising was the same for each group with the following exceptions.

- The advertising for group A included the Choco-Zuties ad but not the Apple-Zuties ad.
- The advertising for group B included the Apple-Zuties ad but not the Choco-Zuties ad.
- The advertising for group C included neither the Choco-Zuties ad nor the Apple-Zuties ad.

After the program, the children were offered a choice between the two snacks. The table below summarizes their choices.

Group	Type of Ad	Number Who Chose Choco-Zuties	Number Who Chose Apple-Zuties
A	Choco-Zuties only	21	4
B	Apple-Zuties only	13	12
C	Neither	22	3

(a) Do the data provide convincing statistical evidence that there is an association between type of ad and children's choice of snack among all children similar to those who participated in the experiment?

(b) Write a few sentences describing the effect of each ad on children's choice of snack.

Problem 3. AP 2010 Form B №5

You'll have about 3 minutes to solve this problem. It will bring you 3% of score for Free Response section.

An advertising agency in a large city is conducting a survey of adults to investigate whether there is an association between highest level of educational achievement and primary source for news. The company takes a random sample of 2,500 adults in the city. The results are shown in the table below.

	HIGHEST LEVEL OF EDUCATIONAL ACHIEVEMENT			
Primary Source for News	Not High School Graduate	High School Graduate But Not College	College Graduate	Total
Newspapers	49	205	188	442
Local television	90	170	75	335
Cable television	113	496	147	756
Internet	41	401	245	687
None	77	165	38	280
Total	370	1437	693	2500

(a) See Chapter 3.

(b) See Chapter 3.

(c) See Chapter 3.

(d) The company wants to conduct a statistical test to investigate whether there is an association between educational achievement and primary source for news for adults in the city. What is the name of the statistical test that should be used?

What are the appropriate degrees of freedom for this test?

Problem 4. AP 2009 №1

You'll have about 5 minutes to solve this problem. It will bring you 5% of score for Free Response section.

A simple random sample of 100 high school seniors was selected from a large school district. The gender of each student was recorded, and each student was asked the following questions.

1. Have you ever had a part-time job?
2. If you answered yes to the previous question, was your part-time job in the summer only?

The responses are summarized in the table below.

Job Experience	Gender		
	Male	Female	Total
Never had a part-time job	21	31	52
Had a part-time job during summer only	15	13	28
Had a part-time job but not only during summer	12	8	20
Total	48	52	100

(a) See Chapter 6.

(b) See Chapter 6.

(c) Which test of significance should be used to test if there is an association between gender and job experience for the population of high school seniors in the district?

State the null and alternative hypotheses for the test, but do not perform the test.

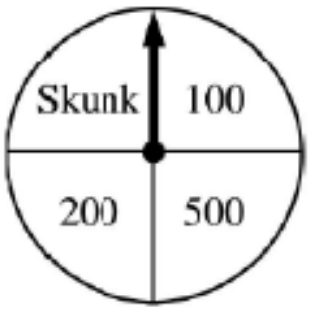
Problem 5. AP 2004 №5 (a)
You'll have about 8 minutes to solve this problem. It will bring you 8% of score for Free Response section.

A rural county hospital offers several health services. The hospital administrators conducted a poll to determine whether the residents' satisfaction with the available services depends on their gender. A random sample of 1,000 adult county residents was selected. The gender of each respondent was recorded and each was asked whether he or she was satisfied with the services offered by the hospital. The resulting data are shown in the table below.

	Male	Female	Total
Satisfied	384	416	800
Not Satisfied	80	120	200
Total	464	536	1,000

(a) Using a significance level of 0.05, conduct an appropriate test to determine if, for adult residents of this county, there is an association between gender and whether or not they were satisfied with services offered by the hospital.

Problem 6. AP 2003 Form B №5
You'll have about 5 minutes to solve this problem. It will bring you 5% of score for Free Response section.



Contestants on a game show spin a wheel like the one shown in the figure below. Each of the four outcomes on this wheel is equally likely and outcomes are independent from one spin to the next.

- The contestant spins the wheel.
- If the result is a skunk, no money is won and the contestant's turn is finished.
- If the result is a number, the corresponding amount in dollars is won. The contestant can then stop with these winnings or can choose the spin again, and his or her turn continues.
- If the contestant spins again and the result is a skunk, all of the money earned on that turn is lost and the turn ends.
- The contestant may continue adding to his or her winnings until he or she chooses to stop or until a spin results in a skunk.

(a) See chapter 2

(b) See chapter 2

(c) A contestant who lost at this game alleges that the wheel is not fair. In order to check on the fairness of the wheel, the data in the table below were collected for 100 spins of this wheel.

Result	Skunk	$100	$200	$500
Frequency	33	21	20	26

Based on these data, can you conclude that the four outcomes on this wheel are not equally likely? Give appropriate statistical evidence to support your answer.

Problem 7. AP 2003 №5

You'll have about 13 minutes to solve this problem. It will bring you 15% of score for Free Response section.

A random sample of 200 students was selected from a large college in the United States. Each selected student was asked to give his or her opinion about the following statement.

"The most important quality of a person who aspires to be the President of the United States is a knowledge of foreign affairs."

Each response was recorded in one of five categories. The gender of each selected student was noted.

The data are summarized in the table below.

	Response Category				
	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Male	10	15	15	25	25
Female	20	25	25	25	15

Is there sufficient evidence to indicate that the response is dependent on gender? Provide statistical evidence to support your conclusion.

Problem 8. AP 1999 №2

You'll have about 13 minutes to solve this problem. It will bring you 15% of score for Free Response section.

The Colorado Rocky Mounting Rescue wishes to study the behavior of lost hikers. If more were known about the direction in which lost hikers tend to walk, then more effective search strategies could be devised. Two hundred hikers selected at random from those applying for hikers permits are asked whether they would head uphill, downhill, or remain in the same place if they became lost while hiking. Each hiker in the sample was also classified according to whether he or she was an experienced or novice hiker. The resulting data are summarized in the following table.

	Direction		
	Uphill	Downhill	Remain in Same Place
Novice	20	50	50
Experienced	10	30	40

Do these data provide convincing evidence of an association between the level of hiking expertise and the direction the hiker would head if lost?

Give appropriate statistical evidence to support your conclusion.

Answers to practice problems

Problem 1: $\chi^2(3) \approx 10.884$, p-value $= 0.012$. We conclude there is an association

Problem 2: (a)$\chi^2(2) \approx 10.291$, p-value ≈ 0.006. We conclude ad and choosing a snack are not independent.

(b) When watching no ad 88% (22 out of 25) chose Choco-Zuites, which is very similar to the same proportion with watching Choco-Zuites ad, 84% (21 out of 25). However, with an Apple-Zuties ad 52% (13 of 25) chose Choco-Zuites. So, the latter ad seems to significantly influence the choice of kids.

Problem 3: (d) chi-square test of independence, df=8.

Problem 4: (c) Chi-square test of independence.

H_0: gender and job experience are independent in the population of high school seniors in the district.

H_A: there is an association between gender and job experience in the population of high school seniors in the district.

Problem 5: $\chi^2(1) \approx 4.117$, p-value $= 0.042$. We conclude that there is evidence of an association between gender and satisfaction with health services offered by the hospital for adult residents of this county.

Problem 6: (c) $\chi^2(3) \approx 4.24$, p-value ≈ 0.237 Thus, the data does not provide enough statistical evidence to claim that the four results of spinning ate not equally likely.

Problem 7: $\chi^2(4) \approx 8.921$, p-value $= 0.063$. Conclusion depends on the chosen level of significance.

Problem 8: $\chi^2(2) \approx 1.5046$, p-value $= 0.471$. There is no evidence of association.

Доп идеи:

Пример задачи: что делать, если $E < 5$. Можно попробовать объединить родственные категории в в более крупные, в которых наблюдаемые и ожидаемые значения будут выше.

Подумать над смыслом степеней свободы.

Essentially degrees of freedom is the number of pieces of data ("observations") that are free to vary, so that we can use them to estimate parameter. Recall $(n-1)$ degrees of freedom for t-distribution used to estimate mu based of sample mean. Initially n observations can take any n values. However, once you estimate the sample mean

+Agresti p.38