L18

Basic Statistics for Economists

Spring 2020

Department of Statistics

Stockholm University

Example (made up)

- An ice cream comes in 4 flavors (C, V, S, R)
- It is <u>assumed</u> that the flavors are equally popular, i.e. if a buyer is drawn at random, the probability of choosing either is 0,25:

Flavor	C	V	S	R	Total
Probability	25%	25%	25%	25%	100%

- A sample is drawn and people were asked about their favorite:

Flavor	C	V	S	R	Total
Frequency	30 (30%)	18 (18%)	25 (25%)	27 (27%)	100

- Given these <u>observed</u> frequencies, should we still believe that the assumed probability distribution above is true?

The problem

- There is an assumption, a **null hypothesis**, on how popular the different flavors are, expressed as a **probability distribution**
- We collect and observe an **empirical frequency distribution**
- Student – teacher:
  - The assumed and the hypothetical distributions differ!
  - Yes they differ! In practice they will almost always differ! Why is that?
  - Well it's a sample and any of the observed differences may be due to randomness.
  - Of course! So the question is, are the observed differences **significant**? And can we **test** for it?

Stockholm University

Today: χ^2-tests, chi-square tests

Goodness-of-fit, (sv. *anpassningstest*)

- without parameter estimation (section 14.1)
- with parameter estimation (section 14.2) - overview
  - Jarque-Bera test for normality - skip

Test for independence / Contingency tables (section 14.3)

- (sv. *oberoendetest*)

Test for homogeneity / Contingency tables (section 14.3)

- slightly different outset but computationally identical to test for independence
- NCT doesn’t distinguish between the two

Definitions and notation

n iid observations on r.v. $Y_i,\ i = 1, \dots, n$ ← **Sample**

$$Y_i = \begin{cases} \text{category } 1 \\ \vdots \\ \text{category } j \\ \vdots \\ \text{category } K \end{cases}$$

- K categories.
- often categorical r.v. on a nominal- or ordinal scale
- may also be discrete count data (integers) or a categorized continuous r.v.

P_j = **probability** that Y_i = "category j"

$O_j = n_j$ = **observed freq.** "category j" ← **NOTE! Random variables**

NCT's notation for observed is "O, oh" not "0, zero"

Definitions, cont.

- **Empirical frequency distribution** across K categories:

Category j	1	2	…	K	Sum
Observed freq.	O_1	O_2	…	O_K	n

(empirical frequency distribution)

- **Probabilities** according to an assumption, a **null hypothesis**:

Probability	P_1	P_2	…	P_K	1

(hypothetical probability distribution)

- Given the P_j and n we calculate the **expected frequencies**:

Expected freq.	E_1	E_2	…	E_K	n

Magnitude of discrepancy from expectations

χ^2-tests are based on comparing the **observed** and the **expected frequencies** under the null hypothesis H_0

- Expected: $\boxed{E_j = nP_j}$

  Ex. If $n = 120$ and $P_j = 0{,}25$ then we would expect $E_j = 25\%$ of $120 = 40$

- Compare (diff): $O_j - E_j$
- Squared difference: $\left(O_j - E_j\right)^2$
- Relative difference: $\left(O_j - E_j\right)^2 / E_j$
- Sum over categories:

$$\boxed{\chi^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j}}$$

Stockholm University

Test variable

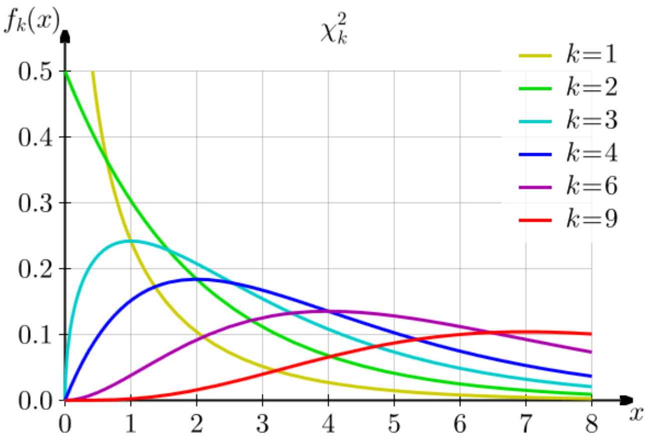
$$\chi^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j} = \sum_{j=1}^{K} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Properties of the test variable:

- The test variable χ^2 is **approximately χ^2-distributed** (NEW)
- χ^2-distributions are defined by the **degrees of freedom ν**
- The degrees of freedom are determined by the number of categories K: $\boxed{\nu = K - 1}$

χ^2-distribution

- Pronounced “kai square”, (sv. “*tji-kvadrat*”, “*tji-två*”)
- We write $\chi^2 \sim \chi^2_\nu$ or $\chi^2 \sim \chi^2(\nu)$
- The parameter ν (“nu”) denotes the **degrees-of-freedom (df)**
- Outcome space = $(\mathbf{0}, \infty)$
- $E(\chi^2) = \nu$
- $Var(\chi^2) = 2\nu$
- The distribution is not symmetric



Source: Wikipedia

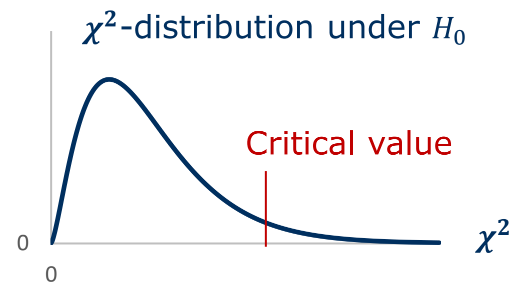
$k = \nu$

Stockholm University

Test variable, cont.

Imprint this in your mind:

$$\boxed{\chi^2 = \sum_{j=1}^{K} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}}$$



small deviations
⇔ small sum

large deviations
⇔ large sum

Properties:

- **Small discrepancies** yield a sum **closer to zero** i.e. a **good fit** to the P_j's i.e. observed ≈ expected and **support to H_0**

- **Large discrepancies** yield a **large sum** and **rejection of H_0**

- NOTE! χ^2**-tests are one-sided tests**. Why?

Exercise 1: Goodness-of-fit test (1a)

Claim: the **probability distribution** of the categorical r.v. $\boldsymbol{Y}$ is

Category $\boldsymbol{j}$	A	B	C	D	E	Σ
$\boldsymbol{P_j}$	0,35	0,25	0,20	0,10	0,10	1

From a sample we obtain the **empirical frequency distribution**, e.g. a sample with $\boldsymbol{n = 200}$ observations distributed as follows:

Category $\boldsymbol{j}$	A	B	C	D	E	Σ
$\boldsymbol{O_j}$	65	49	44	19	23	$\boldsymbol{n} = 200$

Is it reasonable to say that the data support the claim?

Stockholm University

Exercise 1, cont.

H_0: $P_1 = 0{,}35,\ P_2 = 0{,}25, \ldots,\ P_5 = 0{,}10$ (the assumed probability distribution)

H_1: the distribution of Y is not the above

Test variable:

$$\chi^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2_{K-1}$$

i.e. with $\nu = K - 1 = 4$ **degrees of freedom** (K = no. of categories)

- The first $K-1$ probabilities $P_1, \ldots, P_{K-1}$ can be defined (relatively) freely ($0 < P_j < 1$)
- The last one is fully determined by $P_K = 1 - \sum_{j=1}^{K-1} P_j$ **We loose 1 df.!**

Stockholm University

Exercise 1, cont.

Compare the **observed** to the **expected** frequencies under H_0:

Category j	A	B	C	D	E	Σ
O_j	65	49	44	19	23	n = 200
P_j	0,35	0,25	0,20	0,10	0,10	1
E_j	70	50	40	20	20	n = 200
$O_j - E_j$	-5	-1	4	-1	3	0
$(O_j - E_j)^2$	25	1	16	1	9	
$(O_j - E_j)^2/E_j$	25/70	1/50	16/40	1/20	9/20	**1,277**

$$nP_A = 200 \cdot 0,35 = 70$$

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = 1,277143$$

Stockholm University

Exercise 1, cont.

In this example with $\boldsymbol{K} = \mathbf{5}$ categories we have $\mathbf{5 - 1 = 4}$ **df.**

Significance level: say 5 %

Critical region/value: recejt $\boldsymbol{H_0}$ if

$$\chi^2_{obs} > \chi^2_{krit} = \chi^2_{4\,;0,05} = [\text{according to Table 4}] = \mathbf{9,488}$$

Conclusion: $\boldsymbol{H_0}$ is <u>not</u> rejected since $\chi^2_{obs} = \mathbf{1,277} < \chi^2_{krit} = \mathbf{9,488.}$
The claim that the distribution of $\boldsymbol{Y}$ is

Category j	A	B	C	D	E
P_j	0,35	0,25	0,20	0,10	0,10

cannot be rejected at the 5 % level.

Chi-square table

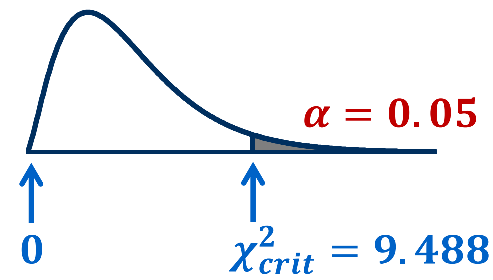


TABLE 4. χ^2-distribution

$Q \in \chi^2(\nu)$ where ν = degrees of freedom.

The value of q_α if $P(Q > q_\alpha) = \alpha$ where α is a given probability.

Significance level α

Degrees of freedom ν

ν	α = 0.999	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
1	0.000	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879	10.828
2	0.002	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597	13.816
3	0.024	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838	16.266
4	0.091	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860	18.467
5	0.210	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750	20.515
6	0.381	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548	22.458
7	0.598	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	24.322
8	0.857	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	26.124
9	1.152	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	27.877
10	1.479	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	29.588
11	1.834	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	31.264
12	2.214	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	32.909
13	2.617	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	34.528

Stockholm University

Rule of thumb

For the test to work the **expected frequencies** in each cell, i.e. all categories, must be **at least 5***, i.e.

$$\boxed{\boldsymbol{E_j = nP_j \geq 5}} \text{ for all } \boldsymbol{j = 1, \dots, K}$$

* Rule of thumb according to NCT (p. 605)
* Alternative rule: E_j on *average* ≥ 5 and at least ≥ 1 for all cells

If smaller than 5, we can collapse (merge) two or more categories.

Ex.

Category $\boldsymbol{j}$	1	2	3	4	5	$\Sigma = n$
$\boldsymbol{E_j = nP_j}$	20,9	51,6	15,2	8,2	4,1	100

< 5

Rule of thumb, cont.

… we **collapse** categories ”4” and ”5” to one category ”4&5”

8,2 + 4,1 = 12,3

Category j	1	2	3	4&5	$\Sigma = n$
$E_j = nP_j$	20,9	51,6	15,2	12,3	100

≥ 5

Stockholm University

Exercise 2: Goodness-of-fit test (1a)

- **Claim:** $\boldsymbol{X_i \sim Bin(4; 0,5)}$ i.e. $\boldsymbol{n = 4}$ and $\boldsymbol{P = 0,5}$
- We have $\boldsymbol{m = 100}$ observations (sample) of $\boldsymbol{X_i}$ distributed as:

$\boldsymbol{x}$	0	1	2	3	4	Σ
$\boldsymbol{O_x}$	14	32	36	16	2	$\boldsymbol{m} = 100$

$\boldsymbol{H_0}$: $\boldsymbol{X_i \sim Bin(4; 0,5)}$ i.e. $\boldsymbol{P = 0,5}$

$\boldsymbol{H_1}$: $\boldsymbol{X_i \nsim Bin(4; 0,5)}$ i.e. anything but <u>not</u> $\boldsymbol{Bin(4; 0,5)}$

Calculate the expected number of 0's, 1's, 2's, 3's and 4's under $\boldsymbol{H_0}$ i.e. under the assumption that $\boldsymbol{X_i \sim Bin(4; 0,5)}$

- $\boldsymbol{K = 5}$ Categories, $\boldsymbol{K - 1 = 4}$ df. $\boxed{\chi^2 \sim \chi^2_4}$

Stockholm University

Exercise 2, cont.

Reject H_0 if $\chi^2_{obs} > \chi^2_{K-1;\alpha} = \chi^2_{4;0,05} = 9,488$ **[Table 4]**

Calculations:

x	0	1	2	3	4	Σ
O_x	14	32	36	16	2	100
$P_x = \binom{4}{x} \cdot 0,5^4$	0,0625	0,2500	0,3750	0,2500	0,0625	1
$E_x = nP_x$	6,25	25,00	37,50	25,00	6,25	100
$O_x - E_x$	7,75	7,00	-1,50	-9,00	-4,25	0
$(O_x - E_x)^2/E_x$	9,61	1,96	0,06	3,24	2,89	$\boxed{\textbf{17,76}}$

Conclusion: $\chi^2_{obs} > \chi^2_{K-1;\alpha}$ so H_0 is rejected, the data provide evidence to suggest that the distribution of X_i is <u>not</u> $Bin(4; 0,5)$.

Note: $\chi^2_{obs} = 17,76$ yields a *p*-value $= 0,001375 < \alpha$ [see Table 4]

Exercise 3: with unknown parameter (1b)

- We believe that $\boldsymbol{X_i \sim Bin(4; P)}$ i.e. $\boldsymbol{n = 4}$ but $\boldsymbol{P =}$ **?**
- We have $\boldsymbol{m = 100}$ observations (sample) of $\boldsymbol{X_i}$ distributed as:

$\boldsymbol{x}$	0	1	2	3	4	Σ
$\boldsymbol{O_x}$	14	32	36	16	2	$\boldsymbol{m} = 100$

$\boldsymbol{H_0}$: $\boldsymbol{X_i \sim Bin(4; P)}$ i.e. $\boldsymbol{P}$ is **unknown/not specified**

$\boldsymbol{H_1}$: $\boldsymbol{X_i \nsim Bin(4; P)}$ i.e. anything else but not $\boldsymbol{Binomial}$ $(n = 4)$

Calculate the expected number of 0's, 1's, 2's, 3's and 4's under $\boldsymbol{H_0}$

However we first have to **estimate $\boldsymbol{P}$ with $\boldsymbol{\hat{p}}$**

- $\boldsymbol{K = 5}$ categories, $\boldsymbol{K - 1 - 1 = 3}$ df.

$\boxed{\chi^2 \sim \chi^2_3}$

Yet another degree of freedom is lost due to estimating one parameter!

Stockholm University

Exercise 3, cont.: Estimate *P*

- Total no. of Bernoulli-trials: $\boldsymbol{n \cdot m = 4 \cdot 100 = 400}$
- No. of successes:

x	0	1	2	3	4	Σ
O_x	14	32	36	16	2	100
$x \cdot O_x$	0	32	72	48	8	**160**

- Estimation of $\boldsymbol{P}$: $\hat{p} = \frac{no.\ of\ successes}{no.\ of\ trials} = \frac{160}{400} = 0,4$

- Estimated probability distribution: $\hat{p}_x = \binom{4}{x} 0,4^x (1 - 0,4)^{4-x}$

x	0	1	2	3	4	Σ
$\hat{p}_x$	0,1296	0,3456	0,3456	0,1536	0,0256	1

Stockholm University

Exercise 3, cont.

Reject $\boldsymbol{H_0}$ if $\chi^2_{obs} > \chi^2_{K-1-1;\alpha} = \chi^2_{3;0,05} = \mathbf{7,815}$ **[Table 4]**

Calculations:

$\boldsymbol{x}$	0	1	2	3	4	$\boldsymbol{\Sigma}$
$\boldsymbol{O_x}$	14	32	36	16	2	100
$\boldsymbol{\widehat{p}_x}$	0,1296	0,3456	0,3456	0,1536	0,0256	1
$\boldsymbol{E_x = nP_x}$	12,96	34,56	34,56	15,36	2,56	100
$\boldsymbol{O_x - E_x}$	1,04	-2,56	1,44	0,64	-0,56	0
$\boldsymbol{(O_x - E_x)^2 / E_x}$	0,0835	0,1896	0,0600	0,0267	0,1225	**0,4823**

Conclusion: $\boldsymbol{H_0}$ is <u>not</u> rejected, the data does <u>not</u> provide evidence to suggest that the distribution of $\boldsymbol{X_i}$ isn't $\boldsymbol{Binomial}$

Note: $\chi^2_{obs} = \mathbf{0,4823}$ yields a *p*-value $= \mathbf{0,9228} > \boldsymbol{\alpha}$

Stockholm University

Contingency tables / Frequency tables

Joint frequency distr.

No. of employees per income class and department		Annual salary, tkr			Total
		200-300	**300-400**	**400-**	
Departm.	**A**	18	4	2	**24**
	B	2	1	-	**3**
	Total	**20**	**5**	**2**	**27**

Marginal counts = Marginal distributions

Joint relative frequency distr.

% employees per income class and department		Annual salary, tkr			Total
		200-300	**300-400**	**400-**	
Departm.	**A**	**67 %**	**15 %**	**7 %**	**89 %**
	B	**7 %**	**4 %**	**-**	**11 %**
	Total	**74 %**	**19 %**	**7 %**	**100 %**

Joint prob. distr. for independent X and Y

Probabilities per income class and department		Annual salary, tkr			Total
		200-300	300-400	400-	
Departm.	A	0,659	0,169	0,062	0,89
	B	0,081	0,021	0,008	0,11
	Total	0,74	0,19	0,07	1

- Does salary **depend** on which department you belong to?
- If X and Y are statistically **independent** we know by now that

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$$

or using a simpler notation $P_{xy} = P_x \cdot P_y$

Stockholm University

Test for Independence, <u>two</u> variables

- Assume we have two **categorical** r.v. X and Y
  - typically nominal or ordinal but it'll work for count data and categorized interval and ratio scales as well

- **Hypotheses**:

  H_0 : X and Y are **independent**

  H_1: X and Y are <u>not</u> independent

Always assume independence in H_0

- The reason for assuming **independence** comes from the fact that we then can calculate the joint probabilities:

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$$

- And then we can calculate the **expected** frequencies E_{ij}

Stockholm University

Test for Independence, cont.

Assume the following two-way contingency table with **joint probabilities** and **marginal probabilities**:

Probability	$Y = 1$	2	3	Σ
$X =$ 1	P_{11}	P_{12}	P_{13}	$P_{1\bullet}$
2	P_{21}	P_{22}	P_{23}	$P_{2\bullet}$
3	P_{31}	P_{32}	P_{33}	$P_{3\bullet}$
4	P_{41}	P_{42}	P_{43}	$P_{4\bullet}$
Σ	$P_{\bullet 1}$	$P_{\bullet 2}$	$P_{\bullet 3}$	1

Q: Why have I shaded six of the cells darker?

Stockholm University

Test for Independence, cont.

Observed frequencies for X and Y

Obs. freq.	$Y = 1$	2	3	Σ
$X =$ 1	O_{11}	O_{12}	O_{13}	R_1
2	O_{21}	O_{22}	O_{23}	R_2
3	O_{31}	O_{32}	O_{33}	R_3
4	O_{41}	O_{42}	O_{43}	R_4
Σ	C_1	C_2	C_3	n

Row sums / frequencies

Column sums / frequencies

Stockholm University

Test for Independence, cont.

- Estimating the **marginal probabilities**:

Rows: $\hat{\boldsymbol{p}}_{\boldsymbol{i\bullet}} = \frac{\boldsymbol{R_i}}{\boldsymbol{n}}$ Columns: $\hat{\boldsymbol{p}}_{\boldsymbol{\bullet j}} = \frac{\boldsymbol{C_j}}{\boldsymbol{n}}$

Under the assumption that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **independent** (i.e. $\boldsymbol{H_0}$) the **expected** frequencies in each cell are calculated as

$$\boxed{\boldsymbol{E_{ij}}} = \boldsymbol{n} \cdot \hat{\boldsymbol{p}}_{\boldsymbol{ij}} = \boldsymbol{n} \cdot \hat{\boldsymbol{p}}_{\boldsymbol{i\bullet}} \cdot \hat{\boldsymbol{p}}_{\boldsymbol{\bullet j}} = n \cdot \frac{R_i}{n} \cdot \frac{C_j}{n} = \boxed{\frac{\boldsymbol{R_i C_j}}{\boldsymbol{n}}}$$

independence

Test for Independence, cont.

- Test variable: $$\boxed{\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}}$$

  i.e. it is a χ^2–distributed r.v. with $(\boldsymbol{r}-\mathbf{1})(\boldsymbol{c}-\mathbf{1})$ degrees of freedom where

  $\boldsymbol{r}$ = no. of rows and $\boldsymbol{c}$ = no. of columns

  (not counting the marginals)

- Reject $\boldsymbol{H_0}$ if $\chi^2_{\boldsymbol{obs}} > \chi^2_{(\boldsymbol{r}-\mathbf{1})(\boldsymbol{c}-\mathbf{1})\,;\,\boldsymbol{\alpha}}$ (Table 4)

Exercise 3: Test for Independence

Job satisfaction (rows $\boldsymbol{i}$) and Job performance (columns $\boldsymbol{j}$) among employees in a sample ($\boldsymbol{n = 190}$) in a certain industry sector is given in the following table:

Obs. freq		$j =$ Low	Medium	High	Σ
$i =$	Low	**46**	**61**	**53**	**160**
	High	**8**	**10**	**12**	**30**
	Σ	**54**	**71**	**65**	**190**

Assume that the $\boldsymbol{n = 190}$ observations are iid (sample)

$\boldsymbol{H_0}$: job satisfaction and performance are **independent**

$\boldsymbol{H_1}$: job satisfaction and performance are **not independent**

Stockholm University

Exercise 3, cont.

Test variable: $$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Distribution: χ^2 with $(2-1)(3-1) = 2$ df.

Significance level: $\alpha = 0,05$

Decision rule/critical value: Reject H_0 if $\chi^2_{\text{obs}} > \chi^2_{2\,;\,0{,}05} = 5,991$

Calculate the **expected frequencies** (E_{ij}) **under H_0** for each combination of i and j (each cell)

Stockholm University

Exercise 3, cont.

Note! All marginal counts and the grand total must be the same in both tables!

Obs. freq.		$j =$ Low	Medium	High	Σ
$i =$	Low	46	61	53	160
	High	8	10	12	30
	Σ	54	71	65	190

$$E_{\text{low,low}} = \frac{R_{\text{low}} C_{\text{low}}}{n} = \frac{54 \cdot 160}{190} \approx 45,474$$

Exp. freq.		$j =$ Low	Medium	High	Σ
$i =$	Low	45,474	59,789	54,737	160
	High	8,526	11,211	10,263	30
	Σ	54	71	65	190

Exercise 3, cont.

Calculations:

$$\underbrace{\boldsymbol{\chi^2_{obs}} = \frac{(46-45{,}474)^2}{45{,}474} + \frac{(8-8{,}526)^2}{8{,}526} + \cdots + \frac{(12-10{,}263)^2}{10{,}263}}_{\text{all six combinations } \boldsymbol{i,j}} = \mathbf{0,543}$$

Conclusion: Since $\boldsymbol{\chi^2_{obs}} = \mathbf{0,543}$ we cannot reject $\boldsymbol{H_0}$ at the 5 % level, the data does <u>not</u> provide evidence that job satisfaction and performance are dependent, we may assume that they are **independent**

Stockholm University

2. Homogeneity test

- Test if the **distributions for different groups are different**
  - e.g. test if the distribution of job satisfaction ($\boldsymbol{Y}$ ordinal) in different countries ($\boldsymbol{X}$ nominal) differ from each other or not

- What group a sampled individual belongs to is not random in the same sense as before
  - independent samples are drawn from each group with predetermined samples sizes
  - iid observations within samples

- Hypotheses: $\boldsymbol{H_0}$: distribution of $\boldsymbol{Y}$ is the same for all countries $\boldsymbol{X}$

  $\boldsymbol{H_1}$: at least one country differs from the others

2. Homogeneity test, cont.

- The null hypothesis "distribution of $\boldsymbol{Y}$ is the same for all groups" is basically the same as saying that it doesn't matter which country you're from, the probability distribution of $\boldsymbol{Y}$ is the same; $\boldsymbol{Y}$ is **independent** of group/country ($\boldsymbol{X}$)

- Apart from that, **same procedure as before**
  - summarize **observed** frequencies $\boldsymbol{O_{ij}}$ in a contingency table
  - calculate observed row and column sums (marginals)
  - calculate **expected** frequencies $\boldsymbol{E_{ij}}$ the same way as before
  - degrees of freedom same as before, $\mathrm{df.} = (\boldsymbol{r} - \mathbf{1})(\boldsymbol{c} - \mathbf{1})$

Stockholm University

Summary

Goodness-of-fit test (one variable):

$$\chi^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j} \qquad \chi^2_{\text{crit}} = \chi^2_{\nu;\,\alpha} \qquad \nu = K - 1$$

Independence & Homogeneity tests (two variables):

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \chi^2_{\text{crit}} = \chi^2_{\nu;\,\alpha} \qquad \nu = (r - 1)(c - 1)$$

Exercise 4: Homogeneity test

In a survey 152 persons from Sweden and 148 from Denmark were interviewed on health and living conditions. Smoking habits within the two samples are presented in the table below. Test if Sweden and Denmark exhibit different smoking habits (if they have different distributions).

Denote $\boldsymbol{X}$ = country gender, $\boldsymbol{Y}$ = cigarettes per day.

	0	1-5	6-15	>15	Σ
Sweden	121	18	10	3	152
Denmark	123	3	17	5	148
Σ	244	21	27	8	300

If time permits, otherwise you should be able to DIY!

Exercise 4, cont.

Homogeneity test

- Assumptions: The two samples are independent of each other and the observations within each sample are *iid*
- **H_0: distributions of Y are the same in Sweden & Denmark**

  H_1: distributions of Y are not the same
- Test variable and its distribution: $$\chi^2 = \sum_{\text{country}} \sum_{\text{habit}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(\nu)$$
- Calculate expected frequencies under H_0: $E_{ij} = \frac{R_i C_j}{n}$

Conditional distribution of Y does not depend on $X \Leftrightarrow X$ and Y are independent

Stockholm University

Exercise 4, cont.

- Degrees of freedom for test variable's distribution:

  $(\mathbf{rows} - \mathbf{1})(\mathbf{columns} - \mathbf{1}) = (\mathbf{2} - \mathbf{1})(\mathbf{3} - \mathbf{1}) = \mathbf{2}$ df.

- Reject $\boldsymbol{H_0}$ if $\boldsymbol{\chi^2_{obs}} > \boldsymbol{\chi^2_{2\,;\,0,05}} = \mathbf{5,991}$ $\quad\quad \boldsymbol{\alpha} = \mathbf{0,05}$

Stockholm University

Exercise 4, cont.

Expected	0	1-5	6-15	>15	Σ
Sweden	123,63	10,64	13,68	**4,05**	152
Denmark	120,37	10,36	13,32	**3,95**	148
Σ	244	21	27	8	300

Ex. $\mathbf{244 \cdot 152/300 = 123,62667}$

- Expectations in cells in column "**>15**" are too low!
  - Rule of thumb: $\boldsymbol{E_{ij} \geq 5}$ for all combinations $\boldsymbol{i, j}$

- Collapse categories "6-15" and ">15"

Stockholm University

Exercise 4, cont.

Note! $\mathbf{13,68 + 4,05 = 17,73}$
$\mathbf{13,32 + 3,95 = 17,27}$

- New table:

	0	1-5	≥ 6	Σ
Sweden	121	18	13	152
Denmark	123	3	22	148
Σ	244	21	35	300

- New expected frequencies, still ***under H_0*** :

	0	1-5	≥ 6	Σ
Sweden	123,63	10,64	17,73	152
Denmark	120,37	10,36	17,27	148
Σ	244	21	35	300

Note that the marginal sums and grand totals are the same in both tables!

Stockholm University

Exercise 4, cont.

- $O_{ij} - E_{ij}$

	0	1-5	≥ 6	Σ
Sweden	-2,63	7,36	-4,73	0
Denmark	2,63	-7,36	4,73	0
Σ	0	0	0	**0**

Note! Margins and grand total always zero!

- $(O_{ij} - E_{ij})^2/E_{ij}$

	0	1-5	≥ 6	Σ
Sweden	0,0559	5,0911	1,2619	6,4089
Denmark	0,0575	5,2287	1,2955	6,5817
Σ	0,1134	10,3198	2,557	**12,991**

Exercise 4, cont.

- Calculation: $\chi^2_{obs} = \mathbf{12,991} > \mathbf{5,991}$

p-value: 0,00151 < 0,05

- **Conclusion**: We reject $\boldsymbol{H_0}$; the data provides strong evidence that the distribution in smoking habits differ between Sweden and Denmark, the result is significant at the 5 % level

Exercis 4: variation

- What if we don't collapse the categories?

Alternative rule of thumb: expected frequencies E_{ij} are at least 1 an on average at least 5.

$$\chi^2 = \sum_{\text{sex}} \sum_{\text{cig}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(3)$$

$$(\textbf{rows}-1)(\textbf{columns}-1) = (2-1)(4-1) = 3 \text{ df.}$$

- Reject H_0 if $\chi^2_{obs} > \chi^2_{3\,;\,0{,}05} = 7{,}815 \qquad \alpha = 0{,}05$
- Calculation: $\chi^2_{obs} = 12{,}99447 > 7{,}815$

p-value: 0,00465 < 0,005

Stockholm University

Next time

Time series

- Data recorded over time
- Components of a time series
- Seasonal adjustment