2 - chi square test

10.1. Does the number of unsolicited (spam) emails follow a Poisson distribution? Here is the record of the number of spam emails received during 60 consecutive days.

$n=60$, Sum $=600$, mean $=10$, std $=8.858$

12	6	4	0	13	5	1	3	10	1	29	12	4	4	22
2	2	27	7	27	9	34	10	10	2	28	7	0	9	4
32	4	5	9	1	13	10	20	5	5	0	6	9	20	28
22	10	8	11	15	1	14	0	9	9	1	9	0	7	13

Choose suitable bins and conduct a goodness-of-fit test at the 1% level of significance.

10.2. Applying the theory of M/M/1 queuing systems, we assume that the service times follow Exponential distribution. The following service times, in minutes, have been observed during 24 hours of operation: $\lambda = \frac{1}{5} = 0.2$ 5 bins $[0,2), [2,4), [4,6), [6,8), [8,\infty)$

$n=64$, Sum $=320$, mean $=5$, $\sigma=3.55$

10.5	1.2	6.3	3.7	0.9	7.1	3.3	4.0	1.7	11.6	5.1	2.8	4.8	2.0	8.0	4.6
3.1	10.2	5.9	12.6	4.5	8.8	7.2	7.5	4.3	8.0	0.2	4.4	3.5	9.6	5.5	0.3
2.7	4.9	6.8	8.6	0.8	2.2	2.1	0.5	2.3	2.9	11.7	0.6	6.9	11.4	3.8	3.2
2.6	1.9	1.0	4.1	2.4	13.6	15.2	6.4	5.3	5.4	1.4	5.0	3.9	1.8	4.7	0.7

Is the assumption of Exponentiality supported by these data?

	$[0,2)$	$[2,4)$	$[4,6)$	$[6,8)$	$[8,\infty)$
p=	0.33	0.22	0.148	0.0993	0.202
ex=	21.12	14.08	9.472	6.355	12.93
obs=	13	16	15	7	13

→ at least 5 bins

the values must be at least 5

$\chi^2 = \frac{(21.12-13)^2}{21.12 \to \text{exp}} + \dots = 6.483$ $p = 0.09$ high → it can be exponential

$v = \underset{\text{bin size}}{n} - 1 - 1 = 3$

10.3. The following sample is collected to verify the accuracy of a new random number generator (it is already ordered for your convenience). $n=100$

-2.434	-2.336	-2.192	-2.010	-1.967	-1.707	-1.678	-1.563	-1.476	-1.388
-1.331	-1.269	-1.229	-1.227	-1.174	-1.136	-1.127	-1.124	-1.120	-1.073
-1.052	-1.051	-1.032	-0.938	-0.884	-0.847	-0.846	-0.716	-0.644	-0.625
-0.588	-0.584	-0.496	-0.489	-0.473	-0.453	-0.427	-0.395	-0.386	-0.386
-0.373	-0.344	-0.280	-0.246	-0.239	-0.211	-0.188	-0.155	-0.149	-0.112
-0.103	-0.101	-0.033	-0.011	0.033	0.110	0.139	0.143	0.218	0.218
0.251	0.261	0.308	0.343	0.357	0.463	0.477	0.482	0.489	0.545
0.590	0.638	0.652	0.656	0.673	0.772	0.775	0.776	0.787	0.969
0.978	1.005	1.013	1.039	1.072	1.168	1.185	1.263	1.269	1.297
1.360	1.370	1.681	1.721	1.735	1.779	1.792	1.881	1.903	2.009

(a) Apply the χ^2 goodness-of-fit test to check if this sample comes from the Standard Normal distribution.

(b) Test if this sample comes from the Uniform(-3,3) distribution.

(c) Is it theoretically possible to accept both null hypotheses in (a) and (b) although they are contradicting to each other? Why does it make sense? → they can fit both

(a)

$(-\infty,-2)$	$[-2,-1)$	$[-1,0)$	$[0,1)$	$[1,2)$	$[2,\infty)$
0.023	0.136	0.341	0.341	0.136	0.023

×100

ex = 0.23 < 5 choose bigger bin

bin sizes can be different

	$(-\infty,-1.5)$	$[-1.5,-0.5)$	$[-0.5,0)$	$[0,0.5)$	$[0.5,1.5)$	$[1.5,\infty)$
	0.067	0.242	0.191	0.191	0.242	0.067
exp=	6.7	24.2	19.1	19.1	24.2	6.7
obs=	8	24	22	15	23	8

$\chi^2 = 1.886$ $p = 0.864$ → high → it can be std distribution ✓

$v = \underset{\text{bin size}}{6} - 1 = 5$ (we did not estimate any parameter)

(b) $(-3,3)$ 6 bins $\frac{100}{6} = 16.67$

$[-3,-2)$	$[-2,-1)$	$[-1,0)$	$[0,1)$	$[1,2)$	$[2,3)$
4	19	31	27	18	1

no problem for obs value

$\chi^2 = 43.52$ $v=5$ $p < 0.000$ → low → it can't be X

10.4 In Example 10.3 on p. 309, we tested whether the number of concurrent users is approximately Normal. How does the result of the chi-square test depend on our choice of bins? For the same data, test the assumption of a Normal distribution using a different set of bins $B_1, \dots, B_N$.

$p > 0.2$ → confirm

$(-2.5, 2.5)$ → 5 bins $\frac{100}{5} = 20$ ex

$[-2.5,-1.5)$	$[-1.5,-0.5)$	$[-0.5,0.5)$	$[0.5,1.5)$	$[1.5,2.5]$
8	24	37	23	8

10.5 Show that the sample size is too small in Example 10.9 on p. 316 to conduct a χ^2 goodness-of-fit test of Normal distribution that involves estimation of its both parameters.

$\chi^2 = \frac{(8-20)^2}{20} + \frac{(24-20)^2}{20} + \dots = 7.2 + 0.8 + 14.45 + 0.05 + 7.2 = 29.7$ $v = 5-1 = 4$ $p < 0.00\dots$ → not standard

10.6. Two computer makers, A and B, compete for a certain market. Their users rank the quality of computers on a 4-point scale as "Not satisfied", "Satisfied", "Good quality", and "Excellent quality, will recommend to others". The following counts were observed,

Computer maker	"Not satisfied"	"Satisfied"	"Good quality"	"Excellent quality"	
A	20	40	70	20	150
B	10	30	40	20	100
	30	70	110	40	

Is there a significant difference in customer satisfaction of the computers produced by A and by B?

$\chi^2 = 3.0664$
$v = 3.1 = 3$
$p = 0.381 >$ high → they are independent, no significant difference

10.7. An AP test has been given in two schools. In the first school, 162 girls and 567 boys passed it whereas 69 girls and 378 boys failed. In the second school, 462 girls and 57 boys passed the test whereas 693 girls and 132 boys failed it.

(a) In the first school, are the results significantly different for girls and boys?

(b) In the second school, are the results significantly different for girls and boys?

(c) In both schools together, are the results significantly different for girls and boys?

For each school, construct a contingency table and apply the chi-square test.

Remark: This data set is an example of a strange phenomenon known as **Simpson's paradox**. Look, the girls performed better than the boys in *each* school; however, in both schools together, the boys did better!!!

Check for yourself... In the first school, 70% of girls and only 60% of boys passed the test. In the second school, 40% of girls and only 30% of boys passed. But in both schools together, 55% of boys and only 45% of girls passed the test. Wow!

first school

	pass	fail
Girls	162	69
Boys	567	378

(a) $\chi^2 = 8.084$, $v=1$, $p = 0.004 <$ low, yes it is different

second school

	pass	fail
G	462	693
B	57	132

$\chi^2 = 6.6363$, $v=1$, $p=0.0099$, yes, it is different

together

	P	F
G	624	762
B	624	510

$\chi^2 = 24.9743$, $v=1$, $p < 0.000$ → yes it is different

10.8. A computer manager decides to install the new antivirus software on all the company's computers. Three competing antivirus solutions (X, Y, and Z) are offered to her for a free 30-day trial. She installs each solution on 50 computers and records infections during the following 30 days. Results of her study are in the table.

Antivirus software	X	Y	Z	
Computers not infected	36	28	32	96
Computers infected once	12	16	14	42
Computers infected more than once	2	6	4	12
	50	50	50	

Does the computer manager have significant evidence that the three antivirus solutions are *not* of the same quality?

$ex = \frac{96 \cdot 50}{150} = 32$ $\frac{42 \cdot 50}{150} = 14$

exp: 32 32 32 / 14 14 14 / 4 4 4 → merge (less than 5)

exp

32	32	32
18	18	18

obs

36	28	32
14	22	18

$\chi^2 = 2.777$, $v=2$, $p = 0.2494 >$

no, they are not different

10.9. The Probability and Statistics course has three sections - S01, S02, and S03. Among 120 students in section S01, 40 got an A in the course, 50 got a B, 20 got a C, 2 got a D, and 8 got an F. Among 100 students in section S02, 20 got an A, 40 got a B, 25 got a C, 5 got a D, and 10 got an F. Finally, among 60 students in section S03, 20 got an A, 20 got a B, 15 got a C, 2 got a D, and 3 got an F. Do the three sections differ in their students' performance?

	A	B	C	D	F
S-1	40	50	20	2	8
S-2	20	40	25	5	10
S-3	20	20	15	2	3

ex

	A	B	C	D	E	
1						100
2						100
3				2.07	4.84	60
	60	110	60	9	21	t=260

low → merge

~~$\chi^2 = 10.28$ $v = 2.4 = 8$ $p = 0.2457$~~ p is too high so they are not different

answer → $0.1 < p < 0.2$

	A	B	C	D-E
1	40	50	20	10
2	20	40	25	15
3	20	20	15	5

$\chi^2 = 9.6162$
$v = 3.2 = 6$
$p = 0.1418$

testing a distribution

$$\chi^2 = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)} \rightarrow \text{rejection region } [\chi^2_\alpha, \infty)$$

k → counts

v → # bin - 1

Example 10.1 (Is the die unbiased?). Suppose that after losing a large amount of money, an unlucky gambler questions whether the game was fair and the die was really unbiased. The last 90 tosses of this die gave the following results,

$H_0: F = F_0$ (unbiased)

$H_A: F \neq F_0$ (biased)

$Ep = 90 \cdot \frac{1}{6} = 15$

$\chi^2_{obs} = \sum_{k=1}^{N} \frac{(obs(k) - Exp(k))^2}{exp(k)} = \frac{(20-15)^2}{15} + \frac{(15-15)^2}{15} + \frac{(12-15)^2}{15} + \frac{(17-15)^2}{15} + \frac{(9-15)^2}{15} + \frac{(17-15)^2}{15} = 5.2$

Number of dots on the die	1	2	3	4	5	6
Number of times it occurred	20	15	12	17	9	17

df → n-1 = 6-1 = 5 chi-square = 5.2 → p val = 0.382 → too high → dont reject → Ho → unbiased

if it was 0.01 reject Ho

Ör/ (Bezelyelerde kalıtımla ilgili ünlü bir deney) Mor çiçeklere ve uzun polen tohumlarına sahip bezelye türü, kırmızı çiçeklere ve yuvarlak polen tohumlarına sahip diğer bezelye türü ile çaprazlanıyor. İlk jenerasyon doğduktan sonra kendi aralarında bir kez daha çaprazlanıyor.
381 bitki için sonuçlar inceleniyor.

Mor/uzun	Mor/yuvarlak	Kırmızı/uzun	Kırmızı/yuvarlak
284	21	21	55

Gen bilimi genler aktarılıyorsa buradaki sonuçların 9:3:3:1 oranında olması gerektiğini söylüyor. Bu çalışma gen biliminin bu iddiası ile ilgili ne söylüyor?

Hipotez testini adımlarını yazarak uygulamaya çalışalım.

→ H_0: Fenotiplerin oranı 9:3:3:1 H_1: Fenotiplerin oranı 9:3:3:1 değil

→ Adım 1: H_0 doğru olsaydı gözlemlenen rakamlar ne olurdu? Hesapla

$\frac{9}{16} \times 381 = 214,3$ $\frac{3}{16} \times 381 = 71,4$ $\frac{1}{16} \times 381 = 23,8$

	Mor/uzun	Mor/yuvarlak	Kırmızı/uzun	Kırmızı/yuvarlak
Gözlenen	284	21	21	55
Beklenen	214,3	71,4	71,4	23,8

Beklenen 214,3 71,4 71,4

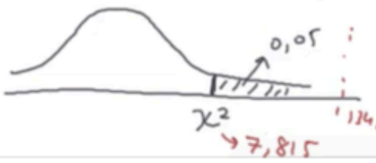
→ Adım 2: Test istatistiğini hesapla $\chi^2 = \sum \frac{(\text{Gözlenen} - \text{Beklenen})^2}{\text{Beklenen}}$

serb. der. = kategorilerin sayısı − 1

$$\chi^2 = \frac{(284-214,3)^2}{214,3} + \frac{(21-71,4)^2}{71,4} + \frac{(21-71,4)^2}{71,4} + \frac{(55-23,8)^2}{23,8} = 134,7$$

→ Adım 3
serb. der. = 4 − 1 = 3

Dikkat = Alanı iki tarafta almıyoruz. Gözlenen beklenen arasında fark olduğunda test istatistiği ancak büyür. (− değer almaz)



H_0 → ret
H_1 → kabul

p değeri çok küçük → reddediyoruz, Oran 9:3:3:1 değil

↳ reject H_0 when p val < 0.01 otherwise accept (it can be too high as well)
what is given in the question

Testing a family of distribution

$\downarrow$!! #bin > 5

Example 10.2 (Transmission errors). The number of transmission errors in communication channels is typically modeled by a Poisson distribution. Let us test this assumption. Among 170 randomly selected channels, 44 channels recorded no transmission error during a 3-hour period, 52 channels recorded one error, 36 recorded two errors, 20 recorded three errors, 12 recorded four errors, 5 recorded five errors, and one channel had seven errors.

H_0: F is poisson
H_a: F is not poisson

estimate parameter

↳ one parameter $\lambda = \bar{x} = \frac{44.0+52.1+36.2+20.3+12.4+5.5+1.7}{170} = 1.55$

portion it into bins

$B_0=\{0\}$ $B_1=\{1\}$ $B_2=\{2\}$ $B_3=\{3\}$ $B_4=\{4\}$ $B_5=\{5\}$ $B_4=\{4\}$

	$\lambda=1.55$, $x=0$	$x=1$	$x=2$	$x=3$	$x=4$	$x=[5,\infty)$!! for cont. distribution	$x=[4,\infty)$
$p \rightarrow$	0.21	0.33	0.26	0.13	0.05	0.02	0.07
out of $\overset{n}{170}$ times we expect them to occurred — their ex values	170.(0.21)2. = 36.0	170.(0.33).. = 55.9	170.(0.26) = 43.4	170.(0.13) = 22.5			170.(0.07) = 12.3
observed times	44	52	36	20			(12+5+1) = 18

virgül sonrası

too small can be combined

look in calculator $x \geq 4$

$\chi^2_{obs} = \sum_{k}^{N} \frac{(obs-ex)^2}{ex} = 6.2$

degrees of freedom

5 bins -1 -1

↳ for each estimated parameter ↳ we estimated only $1 \rightarrow \lambda$

$= v = 3$

$p = 0.1$

high: dont reject H_0 it can be poisson

testing independence

degrees of freedom = (row − 1).(column − 1)

Example 10.4 (Spam and attachments). Modern email servers and anti-spam filters attempt to identify spam emails and direct them to a junk folder. There are various ways to detect spam, and research still continues. In this regard, an information security officer tries to confirm that the chance for an email to be spam depends on whether it contains images or not. The following data were collected on $n = 1000$ random email messages,

H_0: they are independent
H_a: not independent

ex $\left(\frac{300 \cdot 400}{1000}\right)$

$Obs(i,j) = n_{ij}$	With images	No images	$n_{i\cdot}$
Spam	observed 160	240	400
No spam	140	460	600
$n_{\cdot j}$	300	700	1000

Results						
	Category 1	Category 2				Row Totals
Group 1	160 (120.00) [13.33]	240 (280.00) [5.71]				400
Group 2	140 (180.00) [8.89]	460 (420.00) [3.81]				600
Column Totals	300	700				1000 (Grand Total)

The chi-square statistic is 31.746. The *p*-value is < .00001. The result is significant at *p* < .01.

$\chi = 31.746$
$v = 2$

$p <$ too low $\rightarrow$ reject H_0, they are not independent
$\rightarrow$ there is a relationship btw them

$p >$ $\rightarrow$ high no significant difference btw data sets