Chi-square tests: goodness of fit, independence and the chi-square distribution

Ten problems on chi-square tests: the chi-square distribution as a sum of squared standard normals with its mean, variance and the 5% cutoff for two degrees of freedom, a coin whose chi-square statistic is the square of its z score, a goodness-of-fit test of a die rolled 120 times, a classifier's predicted label distribution against a target mix, a Poisson fit to hourly error counts with an estimated rate and a merged tail cell, a 2 × 2 A/B table whose chi-square equals the two-proportion z squared, a 3 × 2 table of model versions and user ratings, its Pearson residuals and Cramér's V, how the statistic grows with the sample size when the proportions stay fixed, and the G-test as twice N times a KL divergence, with worked solutions and the mistakes of running the test on percentages instead of counts, forgetting to subtract a degree of freedom for an estimated parameter, using rc − 1 degrees of freedom for a contingency table, dividing by the observed counts instead of the expected ones and reading a large p-value as proof that the model fits.

BEFORE YOU START

A chi-square test asks whether counts in categories are close enough to the counts a model predicts. The model might say a die is fair, a classifier should predict labels in a target mix, server errors arrive as a Poisson process, or two variables in a table are independent. For each category, the gap between the observed count O and the expected count E is squared and divided by E, and the sum of these terms is compared with a chi-square distribution, the distribution of a sum of squared standard normals. The number of squared normals, the degrees of freedom, is the number of categories minus the constraints the expected counts were made to satisfy. These ten problems build the distribution, run goodness-of-fit and independence tests by hand, connect the 2×2 test to the z test of an A/B experiment and the G-test to the KL divergence, and show what the statistic does as the sample grows.

- **The distribution.** If $Z_1, \dots, Z_k$ are independent standard normals, $Z_1^2 + \cdots + Z_k^2$ has a chi-square distribution with k degrees of freedom, χ^2_k. Its mean is k and its variance is $2k$. For $k = 2$ the upper tail is exactly $\Pr(\chi^2_2 > x) = e^{-x/2}$.
- **Cutoffs at the 5% level.** 3.841 for 1 degree of freedom, 5.991 for 2, 7.815 for 3 and 11.07 for 5. The test is one-sided: only large values of the statistic count against H_0.
- **Pearson's statistic.** $X^2 = \sum (O - E)^2 / E$, summed over all categories, with O and E as **counts**, never percentages. The expected count of a category is N times its probability under H_0.

- **Goodness of fit.** With k categories, $\text{df} = k - 1 - m$, where m is the number of parameters estimated from the same data to compute the expected counts. The -1 is there because the expected counts must add up to N.
- **Independence.** For an $r \times c$ table, $E = (\text{row total})(\text{column total})/N$ and $\text{df} = (r-1)(c-1)$.
- **When the approximation holds.** The chi-square distribution is a large-sample approximation. A common rule asks for every expected count to be at least 5; merge sparse categories, chosen before looking at the data, to get there.
- **Effect size and residuals.** Cramér's $V = \sqrt{X^2/(N(\min(r,c)-1))}$ measures the strength of association on a 0 to 1 scale, whatever N is. The Pearson residual $(O-E)/\sqrt{E}$ shows which cells drive the statistic; their squares add up to X^2.
- **G-test.** $G = 2\sum O \ln(O/E)$ is the likelihood-ratio statistic. It is close to X^2 when the counts are large, and it has the same degrees of freedom.

Builds on: Hypothesis tests and p-values: z and t tests, power and multiple comparisons, Discrete random variables: expectation, variance and the binomial, geometric and Poisson

Problems

1. Let Z_1, Z_2, Z_3 be independent standard normals and $Q = Z_1^2 + Z_2^2 + Z_3^2$. Find $E[Q]$ and $\text{Var}(Q)$. Then, using $\Pr(\chi^2_2 > x) = e^{-x/2}$, find the 5% cutoff for 2 degrees of freedom. difficulty ·

2. A coin lands heads 60 times in 100 flips. Compute Pearson's statistic for $H_0 : p = \frac{1}{2}$ with the two categories heads and tails, compare it with the z statistic for the same test, and find the p-value. difficulty ·

3. A die is rolled 120 times, with these counts: difficulty ··

face	1	2	3	4	5	6
count	25	17	15	23	24	16

Test whether the die is fair at the 5% level.

4. A classifier should predict the labels cat, dog and bird in the mix 50%, 30% and 20%. On 1000 inputs it predicts cat 450 times, dog 350 times and bird 200 times. Test whether its predicted mix matches the target, at the 5% level. difficulty ··

5. A service logs its errors per hour for 200 hours: 80 hours had 0 errors, 64 had 1, 36 had 2, 16 had 3 and 4 had 4, with none higher. Test whether the counts follow a Poisson distribution, estimating its rate from the data, and merging the categories 3 and 4 or more so that every expected count is at least 5. difficulty ···

6. In an A/B test, 200 of 2000 users convert with the old checkout and 240 of 2000 with the new one. Arrange the data as a 2×2 table, run the chi-square test of independence and compare the statistic with the pooled two-proportion z statistic, $z \approx 2.021$. difficulty ··

7. Three model versions are each shown to 100 users, who give a thumbs up or a thumbs down: difficulty ··

	up	down
v1	60	40
v2	75	25
v3	45	55

Test whether the rating is independent of the version, at the 5% level.

8. For the table of Problem 7, find the Pearson residuals $(O - E)/\sqrt{E}$ of every cell and Cramér's V. Which cells drive the result? difficulty ··

9. Keep the classifier's predicted proportions of Problem 4, 45%, 35% and 20%, against the target 50%, 30%, 20%, but vary the number of inputs N. Write X^2 as a function of N, find the smallest N that gives significance at 5%, and find X^2 at $N = 100{,}000$. difficulty ··

10. For the classifier of Problem 4, compute the G-statistic $G = 2\sum O \ln(O/E)$, show that it equals $2N$ times the KL divergence from the target proportions to the observed ones, and compare it with Pearson's $X^2 \approx 13.33$. difficulty ···

Answers

1. $E[Q] = 3$, $\mathrm{Var}(Q) = 6$, and the 5% cutoff for χ^2_2 is $2 \ln 20 \approx 5.991$
2. $X^2 = 4 = z^2$, $\mathrm{df} = 1$, $p = \Pr(\chi^2_1 > 4) \approx 0.0455$
3. $X^2 = 5 < 11.07$ and $p \approx 0.416$: no evidence that the die is unfair
4. $X^2 \approx 13.33 > 5.991$, $p \approx 0.0013$: the predicted mix differs from the target
5. $X^2 \approx 2.79$ with $\mathrm{df} = 4 - 1 - 1 = 2$, so $p \approx 0.248$: the Poisson model is not rejected
6. $X^2 \approx 4.086$, $\mathrm{df} = 1$, $p \approx 0.043$; and $z^2 = 2.021^2 \approx 4.086$
7. $X^2 = 18.75$ with $\mathrm{df} = (3-1)(2-1) = 2$, so $p = e^{-9.375} \approx 8.5 \times 10^{-5}$: the rating depends on the version
8. Residuals $0, 0$; $1.94, -2.37$; $-1.94, 2.37$, and $V = \sqrt{18.75/300} = 0.25$
9. $X^2 = 0.01333N$; the smallest significant N is 450; at $N = 100{,}000$, $X^2 \approx 1333$
10. $G \approx 13.08 = 2000 \cdot 0.006541$, close to $X^2 \approx 13.33$, with $p = e^{-G/2} \approx 0.0014$

Worked solutions

Problem 1

> Let Z_1, Z_2, Z_3 be independent standard normals and $Q = Z_1^2 + Z_2^2 + Z_3^2$. Find $E[Q]$ and $\mathrm{Var}(Q)$. Then, using $\Pr(\chi^2_2 > x) = e^{-x/2}$, find the 5% cutoff for 2 degrees of freedom.

$E[Z^2] = \mathrm{Var}(Z) + (E[Z])^2 = 1$	A standard normal has mean 0 and variance 1.
$E[Z^4] = 3$, so $\mathrm{Var}(Z^2) = 3 - 1^2 = 2$	The fourth moment of a standard normal is 3.
$E[Q] = 3$ and $\mathrm{Var}(Q) = 3 \cdot 2 = 6$	Means add, and so do variances of independent terms.
$e^{-x/2} = 0.05$, so $x = 2\ln 20$	Set the upper tail equal to 0.05 and take logs.
$E[Q] = 3$, $\mathrm{Var}(Q) = 6$, **and the** 5% **cutoff for** χ^2_2 **is** $2\ln 20 \approx 5.991$	In general χ^2_k has mean k and variance $2k$, so a statistic close to its degrees of freedom is unremarkable.

Problem 2

> A coin lands heads 60 times in 100 flips. Compute Pearson's statistic for $H_0 : p = \frac{1}{2}$ with the two categories heads and tails, compare it with the z statistic for the same test, and find the p-value.

$E = 50$ for heads and 50 for tails	Np and $N(1-p)$.
$X^2 = \dfrac{10^2}{50} + \dfrac{(-10)^2}{50} = 4$	Both gaps are 10, in opposite directions.
$z = \dfrac{60 - 50}{\sqrt{100 \cdot 0.25}} = 2$	The binomial has standard deviation $\sqrt{Np(1-p)} = 5$.
$X^2 = z^2$, with $\mathrm{df} = 2 - 1 = 1$	With two categories the two gaps are tied, so there is one squared normal, not two.

$X^2 = 4 = z^2$, $\mathrm{df} = 1$, $p = \Pr(\chi^2_1 > 4) \approx 0.0455$	The same p-value as the two-sided z test, and the cutoff 3.841 is 1.96^2.

Problem 3

> A die is rolled 120 times, with these counts:
>
> | face | 1 | 2 | 3 | 4 | 5 | 6 |
> |---|---|---|---|---|---|---|
> | count | 25 | 17 | 15 | 23 | 24 | 16 |
>
> Test whether the die is fair at the 5% level.

$E = 120/6 = 20$ for every face	Under H_0 each face has probability $\frac{1}{6}$.
$O - E = 5, -3, -5, 3, 4, -4$	The gaps add up to 0, as they always must.
$\sum (O - E)^2 = 25 + 9 + 25 + 9 + 16 + 16 = 100$	Square every gap.
$X^2 = 100/20 = 5$, with $\mathrm{df} = 6 - 1 = 5$	All the expected counts are equal, so divide once.
$X^2 = 5 < 11.07$ **and** $p \approx 0.416$: **no evidence that the die is unfair**	The statistic is right at its mean of 5. That is not evidence that the die is fair, only that 120 rolls cannot tell (Mistake 5).

Problem 4

> A classifier should predict the labels cat, dog and bird in the mix 50%, 30% and 20%. On 1000 inputs it predicts cat 450 times, dog 350 times and bird 200 times. Test whether its predicted mix matches the target, at the 5% level.

$E = 500, 300, 200$	1000 times each target proportion.
$\frac{(-50)^2}{500} = 5$ and $\frac{50^2}{300} \approx 8.333$	Cat and dog; bird matches exactly and adds 0.

$X^2 = 5 + 8.333 + 0 \approx 13.33$	The same gap of 50 counts for more in the smaller category.
$\mathrm{df} = 3 - 1 = 2$, and $p = e^{-13.33/2} \approx 0.0013$	The exact tail for 2 degrees of freedom.
$X^2 \approx 13.33 > 5.991$, $p \approx 0.0013$: **the predicted mix differs from the target**	The classifier over-predicts dog. Running the test on the percentages 45, 35, 20 instead gives $X^2 \approx 1.33$ and the opposite conclusion (Mistake 1).

Problem 5

> A service logs its errors per hour for 200 hours: 80 hours had 0 errors, 64 had 1, 36 had 2, 16 had 3 and 4 had 4, with none higher. Test whether the counts follow a Poisson distribution, estimating its rate from the data, and merging the categories 3 and 4 or more so that every expected count is at least 5.

$\hat{\lambda} = (64 + 72 + 48 + 16)/200 = 1$	The maximum-likelihood rate is the sample mean: total errors over hours.
$E = 200e^{-1}\left(1, 1, \tfrac{1}{2}\right) \approx 73.58,\ 73.58,\ 36.79$	$200\,\Pr(K = k)$ for $k = 0, 1, 2$, with $\Pr(K = k) = e^{-\lambda}\lambda^k/k!$.
$E_{3+} = 200 - 183.94 \approx 16.06$, with $O_{3+} = 20$	The cell 4 or more alone would expect only 3.80, so it is merged with 3.
$X^2 \approx 0.561 + 1.246 + 0.017 + 0.966 \approx 2.79$	The four terms $(O - E)^2/E$.
$X^2 \approx 2.79$ **with** $\mathrm{df} = 4 - 1 - 1 = 2$, **so** $p \approx 0.248$: **the Poisson model is not rejected**	One degree of freedom goes to the total and one to the estimated λ. Forgetting the second gives $p \approx 0.425$ (Mistake 2).

Problem 6

In an A/B test, 200 of 2000 users convert with the old checkout and 240 of 2000 with the new one. Arrange the data as a 2×2 table, run the chi-square test of independence and compare the statistic with the pooled two-proportion z statistic, $z \approx 2.021$.

Rows old, new; columns converted, not: $200, 1800$ and $240, 1760$	Row totals 2000 and 2000; column totals 440 and 3560.
$E = 2000 \cdot 440/4000 = 220$ converted per row, 1780 not	Row total times column total over N.
$\lvert O - E \rvert = 20$ in every cell	In a 2×2 table all four gaps have the same size.
$X^2 = 400\left(\frac{2}{220} + \frac{2}{1780}\right) \approx 3.636 + 0.449$	Two cells divide by 220 and two by 1780.
$X^2 \approx 4.086$, df $= 1$, $p \approx 0.043$; **and** $z^2 = 2.021^2 \approx 4.086$	The chi-square test of a 2×2 table is exactly the two-sided pooled z test, squared, and gives the same p-value. Using $rc - 1 = 3$ degrees of freedom gives $p \approx 0.25$ (Mistake 3).

Problem 7

Three model versions are each shown to 100 users, who give a thumbs up or a thumbs down:

	up	down
v1	60	40
v2	75	25
v3	45	55

Test whether the rating is independent of the version, at the 5% level.

Column totals 180 up and 120 down, $N = 300$	Each row total is 100.
$E = 100 \cdot 180/300 = 60$ up and 40 down in every row	Under independence each version gets the overall 60% rate.

v1 adds 0; v2 and v3 each add $\frac{15^2}{60} + \frac{15^2}{40}$	v2 is 15 above in up, v3 15 below.
$X^2 = 2(3.75 + 5.625) = 18.75$	$225/60 = 3.75$ and $225/40 = 5.625$.
$X^2 = 18.75$ **with** $\mathrm{df} = (3-1)(2-1) = 2$, **so** $p = e^{-9.375} \approx 8.5 \times 10^{-5}$**: the rating depends on the version**	Far beyond the cutoff 5.991.

Problem 8

> For the table of Problem 7, find the Pearson residuals $(O - E)/\sqrt{E}$ of every cell and Cramér's V. Which cells drive the result?

v1: 0 and 0	The first version matches the overall rate exactly.
v2: $15/\sqrt{60} \approx 1.94$ up, $-15/\sqrt{40} \approx -2.37$ down	Positive residuals mark more counts than independence predicts.
v3: -1.94 up, $+2.37$ down	The mirror image of v2.
$1.94^2 \cdot 2 + 2.37^2 \cdot 2 = 7.5 + 11.25 = 18.75$	The squared residuals add up to X^2.
Residuals $0, 0$; $1.94, -2.37$; $-1.94, 2.37$, **and** $V = \sqrt{18.75/300} = 0.25$	The $\min(r, c) - 1$ is 1 here. v2 is liked and v3 is disliked more than chance allows, and the down cells weigh more because their expected counts are smaller.

Problem 9

> Keep the classifier's predicted proportions of Problem 4, 45%, 35% and 20%, against the target 50%, 30%, 20%, but vary the number of inputs N. Write X^2 as a function of N, find the smallest N that gives significance at 5%, and find X^2 at $N = 100{,}000$.

$X^2 = \sum \dfrac{(N\hat{p} - Np)^2}{Np} = N \sum \dfrac{(\hat{p} - p)^2}{p}$	Every count is N times a proportion.
$\sum \dfrac{(\hat{p} - p)^2}{p} = \dfrac{0.0025}{0.5} + \dfrac{0.0025}{0.3} \approx 0.01333$	The bird term is 0.
$0.01333N > 5.991$ when $N > 449.3$	The cutoff for 2 degrees of freedom.
$N = 450$ gives $p \approx 0.0498$; $N = 449$ gives $p \approx 0.0501$	The same mismatch flips from not significant to significant.
$X^2 = 0.01333N$; **the smallest significant** N **is** 450; **at** $N = 100{,}000$, $X^2 \approx 1333$	X^2 grows like N while the effect size $w = \sqrt{X^2/N} \approx 0.115$ stays put. With enough data any mismatch, however small, is significant, so report the size of the effect, not only the p-value.

Problem 10

> For the classifier of Problem 4, compute the G-statistic $G = 2 \sum O \ln(O/E)$, show that it equals $2N$ times the KL divergence from the target proportions to the observed ones, and compare it with Pearson's $X^2 \approx 13.33$.

$450 \ln \frac{450}{500} \approx -47.41$	Cat: $\ln 0.9 \approx -0.1054$.
$350 \ln \frac{350}{300} \approx 53.95$	Dog: $\ln \frac{7}{6} \approx 0.1542$. Bird adds $200 \ln 1 = 0$.
$G = 2(53.95 - 47.41) \approx 13.08$	The likelihood-ratio statistic.
$\mathrm{KL}(\hat{p} \,\|\, p) = \sum \hat{p} \ln(\hat{p}/p) \approx 0.006541$	Each term of G divided by $2N$, since $O = N\hat{p}$ and $O/E = \hat{p}/p$.

$G \approx 13.08 = 2000 \cdot 0.006541$, **close to** $X^2 \approx 13.33$, **with** $p = e^{-G/2} \approx 0.0014$	$G = 2N\,\mathrm{KL}(\hat{p} \,\|\, p)$ exactly, and X^2 is its second-order approximation, so the two agree when the observed and expected proportions are close.

Where this goes wrong

1. Percentages instead of counts

The predicted mix is reported as percentages, and percentages look like the natural thing to compare.

Observed $45, 35, 20$ and expected $50, 30, 20$, in percent	Right so far as a description of Problem 4.
"$X^2 = \dfrac{25}{50} + \dfrac{25}{30} + 0$."	The statistic needs counts. Rescaling every count by c rescales X^2 by c, so percentages behave as if N were 100.
$X^2 \approx 1.33$, $p \approx 0.51$: the classifier matches the target ✗	The counts out of 1000 give $X^2 \approx 13.33$ and $p \approx 0.0013$. The percentages threw away a factor of 10 in the sample size.

2. Not subtracting a degree of freedom for an estimated parameter

The Poisson test of Problem 5 has four categories, and the rule "categories minus one" is remembered on its own.

$X^2 \approx 2.79$ over 4 categories	Right so far: Problem 5.
"$\mathrm{df} = 4 - 1 = 3$."	The rate $\hat{\lambda} = 1$ was estimated from the same counts, which pulls the expected counts toward the observed ones and removes one more degree of freedom.
$p = \Pr(\chi^2_3 > 2.79) \approx 0.425$ ✗	The right reference is χ^2_2, with $p \approx 0.248$ and cutoff 5.991 instead of 7.815. Too many degrees of freedom makes every fit look better than it is.

3. rc − 1 degrees of freedom for a contingency table

A 2×2 table has four cells, so four minus one looks like the count of free cells.

$X^2 \approx 4.086$ for the A/B table	Right so far: Problem 6.
"$\text{df} = 4 - 1 = 3$."	Under independence the expected counts must match both the row totals and the column totals. Once one cell of a 2×2 table is fixed with all totals known, the other three follow, so $\text{df} = (2 - 1)(2 - 1) = 1$.
$p = \Pr(\chi^2_3 > 4.086) \approx 0.25$, not significant ✗	Against the cutoff 7.815 the result vanishes. With 1 degree of freedom the cutoff is 3.841 and $p \approx 0.043$, the same as the z test.

4. Dividing by the observed counts

The formula is remembered as "squared difference over the count", and the observed count is the one on the page.

$O - E = 5, -3, -5, 3, 4, -4$ for the die	Right so far: Problem 3.
"$X^2 = \sum (O - E)^2 / O$."	Pearson's statistic divides by E: under H_0 a count has variance about E, so $(O - E)/\sqrt{E}$ is roughly a standard normal. Dividing by O gives a different statistic, Neyman's, which only agrees with Pearson's when every count is large, and is infinite whenever a category has no observations.
$X^2 = \frac{25}{25} + \frac{9}{17} + \frac{25}{15} + \frac{9}{23} + \frac{16}{24} + \frac{16}{16} \approx 5.25$ ✗	The right value is $100/20 = 5$. The gap is small here because the counts are near 20; with a sparse category it can be huge.

5. A large p-value read as proof that the model fits

The die of Problem 3 passed its test easily, so it is declared fair.

$X^2 = 5$, $p \approx 0.416$	Right so far: Problem 3.
“The test shows the die is fair.”	A test can only fail to find a difference; whether it would have found one depends on its power.
The die is fair ✗	A die that shows 1 a fifth of the time, and each other face 16% of the time, gives 120 rolls a noncentrality of only $120 \sum (p_1 - p_0)^2 / p_0 = 0.96$, and the test rejects it only about 10% of the time. Ninety percent of the time such a loaded die passes. Keeping H_0 means "not shown to be wrong", not "shown to be right".

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 21 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/chi-square-tests-goodness-of-fit-and-independence