Chi-Square Tests

Testing for Goodness of Fit and Independence

Part I: Testing Goodness of Fit

- There is a chance model
- There are observed frequency counts
- Wish to see whether the counts are consistent with the chance model (whether it fits the data well)

Example: Counts of Suicides by Month in US in 1970

Jan	1867
Feb	1789
Mar	1944
Apr	2094
May	2097
Jun	1981
Jul	1887
Aug	2024
Sept	1928
Oct	2032
Nov	1978
Dec	1859
Total	23,480

Are all months equally likely? Compare observed frequencies to those expected from a box model:

- Tickets: labeled 1-365 for days of the year
- Draws: 23,480 with replacement
- Group the results into months

According to this chance model, a June ticket has a probability of 31/365. The expected number of June suicides is

$$23480 \times (31/365) = 1929.86$$

	Days	Observed	Expected
Jan	31	1867	1994.19
Feb	28	1789	1801.21
Mar	31	1944	1994.19
Apr	30	2094	1929.86
May	31	2097	1994.19
Jun	30	1981	1929.86
Jul	31	1887	1994.19
Aug	31	2024	1994.19
Sep	30	1928	1929.86
Oct	31	2032	1994.19
Nov	30	1978	1929.86
Dec	31	1859	1994.19

	Days	Observed	Expected	O - E	$(O - E)^2/E$
Jan	31	1867	1994.19	-127.19	8.11
Feb	28	1789	1801.21	-12.21	0.08
Mar	31	1944	1994.19	-50.19	1.26
Apr	30	2094	1929.86	164.14	13.96
May	31	2097	1994.19	102.81	5.30
Jun	30	1981	1929.86	51.14	1.36
Jul	31	1887	1994.19	-107.19	5.76
Aug	31	2024	1994.19	29.81	0.45
Sep	30	1928	1929.86	-1.86	0.00
Oct	31	2032	1994.19	37.81	0.72
Nov	30	1978	1929.86	48.14	1.20
Dec	31	1859	1994.19	-135.19	9.17

$$\text{Total} = \chi^2 = 47.37$$

“chi-square”

The chi-square statistic measures how closely the observed and expected counts agree.

Even if the chance model from which the expected counts are derived holds exactly, the two will not agree perfectly, just because of chance.

In order to judge how big is unusual, we need to know the probability law of the chi-square statistic when the chance model is true.

This is similar to the case of the z-statistic: it's numerator will generally be different from 0 even when the null hypothesis is true.

Null hypothesis: the chance model generated the data

Alternative hypothesis: it didn't, there is something else going on.

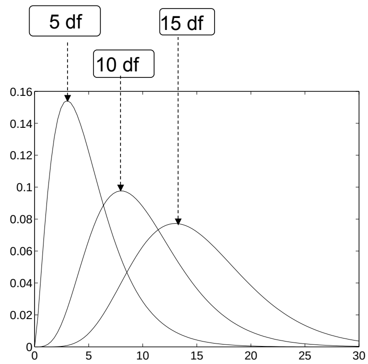
In our example:

Null hypothesis: suicides are equally likely on any day.

Alternative hypothesis: There is something else going on, like seasons have an effect.

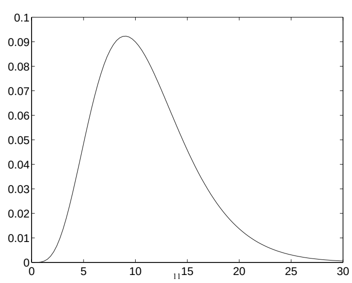
Chi-square distribution

If the null hypothesis is true, the probability histogram of the chi-square statistic is approximately equal to the chi-square distribution with "degrees of freedom" equal to the number of cells minus one.

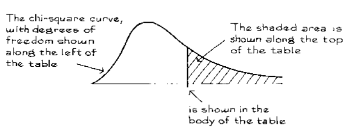


For the suicide data, there are 12 cells so df = 11. The chi-square statistic was 47.37.

chi-square with df=11



A CHI-SQUARE TABLE



Degrees of freedom	*99%*	*95%*	*90%*	*70%*	*50%*	*30%*	*10%*	*5%*	*1%*
1	0.00016	0.0039	0.016	0.15	0.46	1.07	2.71	3.84	6.64
2	0.020	0.10	0.21	0.71	1.39	2.41	4.60	5.99	9.21
3	0.12	0.35	0.58	1.42	2.37	3.67	6.25	7.82	11.34
4	0.30	0.71	1.06	2.20	3.36	4.88	7.78	9.49	13.28
5	0.55	1.14	1.61	3.00	4.35	6.06	9.24	11.07	15.09
6	0.87	1.64	2.20	3.83	5.35	7.23	10.65	12.59	16.81
7	1.24	2.17	2.83	4.67	6.35	8.38	12.02	14.07	18.48
8	1.65	2.73	3.49	5.53	7.34	9.52	13.36	15.51	20.09
9	2.09	3.33	4.17	6.39	8.34	10.66	14.68	16.92	21.67
10	2.56	3.94	4.86	7.27	9.34	11.78	15.99	18.31	23.21
11	3.05	4.58	5.58	8.15	10.34	12.90	17.28	19.68	24.73
12	3.57	5.23	6.30	9.03	11.34	14.01	18.55	21.03	26.22

General Form of the Chi-Square Goodness of Fit Test

category	Observed count (O)	Theoretical probability	Expected count (E)	Contribution to chi-squared
1	O_1	P_1	$E_1 = NP_1$	$(O_1-E_1)^2 \div E_1$
2	O_2	P_2	$E_2 = NP_2$	$(O_2-E_2)^2 \div E_2$
etc	etc	etc	etc	etc
K	O_K	P_K	$E_K = NP_K$	$(O_K-E_K)^2 \div E_K$

N = SUM

Chi-square = SUM

DF = K-1

The chi-square test

- It is performed on *frequency counts -- not percents*.
- It depends on the number of degrees of freedom (df)
- The chi-square curve is an approximation which is good if the expected frequencies are all greater than 5.

Chi-Square Test & Z Test: How are they similar/different?

Both compare “observed” and “expected.”

Data:

•Z-test used for comparing averages of random samples

•Chi-square test used for comparing counts in categories

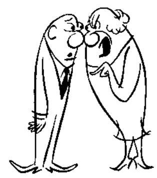
The forms of the test statistics are different.

The null distributions of the test statistics are different.

Does the example make sense?

Critic: This is total baloney! You have all the data from 1970. There is no sampling, no chance model. You're engaging in numerology.

Investigator: Well, it's true that there is no random sample. But my hypothesis is that there in no time effect, so that suicides occur totally randomly throughout the year. I want to see if the data are consistent with that model for how they came about.

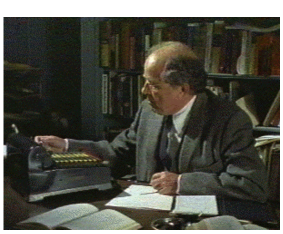


Critic: So there is no actual physical chance model. The chance model is all in your mind!

Investigator: Well, even a physical model is all in one's mind, after all. I don't see why I can't think about the number of suicides in a given month as random if you can think about your silly coin tosses as random.

Critic: Hrrumph... sophistry

Sometimes the fit is too good: the case of Cyril Burt



Sir Cyril Burt studied the relationship between IQ and socioeconomic status in British children. His studies were frequently cited as evidence that upper class children were smarter than working class children and should receive separate schooling. This logic was subsequently used to argue for separate educational institutions for different races. He argued heavily and published extensive data on the genetic basis of intelligence.

In 1946, he became the first psychologist to be knighted.

It was later revealed that he fabricated his data, and that his co-investigators didn't exist. His reports were not questioned because they were consistent with popular beliefs.

Famous study of intelligence of 40,000 fathers and sons. A goodness of fit test of their histograms to a normal distribution gave P-values $1\text{-}10^{-7}$ and $1\text{-}10^{-8}$

Part II: The Chi-Square Test of Independence

Testing independence of cross-classified categories

Example: Do military pilots father more girls than boys? Data were gathered to test this conventional wisdom.

Father's activity

	Flying Fighters	Flying Transports	Not Flying
Female Offspring	51	14	38
Male Offspring	38	16	46

Father's Activity

	Flying Fighters	Flying Transports	Not Flying
Female Offspring	57%	47%	45%
Male Offspring	43%	53%	55%

Is there something going on here, or could this be due to chance?

Calculating the frequencies we would expect on the basis of chance alone:

	Flying Fighters	Flying Transports	Not Flying	Total	%
Female Offspring	51	14	38	103	50.7
Male Offspring	38	16	46	100	49.3
Total	89	30	84	203	

Of the 89 children born to fighter pilots, how many females would be expected? 89 x .507 = 45.1

Of the 30 children born to transport pilots, how many females would be expected? 30 x .507 = 15.2

Observed Frequencies

	Flying Fighters	Flying Transports	Not Flying	Total	%
Female Offspring	51	14	38	103	50.7
Male Offspring	38	16	46	100	49.3
Total	89	30	84	203	

Expected Frequencies

	Flying Fighters	Flying Transports	Not Flying	Total	%
Female Offspring	45.1	15.2	42.6	103	50.7
Male Offspring	43.9	14.8	41.4	100	49.3
Total	89	30	84	203	

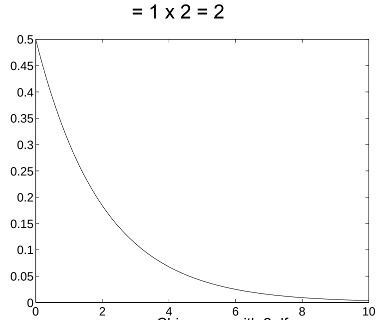
$$\chi^2 = \text{Sum of } \frac{(\text{observed freq - expected freq})^2}{\text{expected freq}}$$

	Flying Fighters	Flying Transports	Not Flying	Total	%
Female Offspring	.77	.09	.50	103	50.7
Male Offspring	.79	.10	.51	100	49.3
Total	89	30	84	203	

$$\chi^2 = 2.76$$

degrees of freedom = (# rows - 1) x (# cols - 1)

= 1 x 2 = 2



From the table the p-value for 2.76 is greater than 10% and a little smaller than 30% (check it)

Why (# rows - 1) x (# columns - 1)?

How many “degrees of freedom” are there in a 2x2 table?

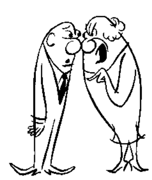
	?	?	100
	?	?	200
fixed col totals	125	175	

fixed row totals

Does hypothesis testing in this example make any sense?

Critic: Your calculation of P-values is silly. You don’t have any chance model. You just went out and got a bunch of records of military pilots, not a random sample from any population by any stretch of the imagination.

Investigator: Well, you’re right that there wasn’t any random sample, but I do think there was chance at work.



Critic: “Chance at work,” huh. You are going to have to give me a real model, not just a vague statement like that.

Investigator: OK, my chance model is that sexes of all the children were like tosses of a coin, independent of what kind of airplane the father was flying. It may not be quite a fair coin, so I estimate the chance of a boy or girl from all the births. Certainly, the gender of a particular birth is as random as one of your silly coin tosses! Now I use a hypothesis test to see if the data are consistent with this model.

Critic: Well, you’re a little more convincing than last time. Just remember for the future that I’m watching every move you make when you do those hypothesis tests you do.

Observed Frequencies

	Flying Fighters	Flying Transports	Not Flying	Total	%
Female Offspring	51	14	38	103	50.7
Male Offspring	38	16	46	100	49.3
Total	89	30	84	203	

Expected Frequencies

	Flying Fighters	Flying Transports	Not Flying	Total	%
Female Offspring	45.1	15.2	42.6	103	50.7
Male Offspring	43.9	14.8	41.4	100	49.3
Total	89	30	84	203	

Note: the expected frequency in a cell can be found by multiplying the row and column totals corresponding to that cell and then dividing by the grand total!

For example, 42.6 = (103 x 84)/203

31

Summary

Both chi-squared tests operate on counts in tables.

Both use the chi-square tables. Expected counts should be greater than 5 for the table to give a good approximation.

Goodness of fit: are the counts consistent with an hypothesized probability law? The expected frequencies are based on given theoretical probabilities. DF = #cells -1

Independence test: are the counts consistent with the row and column categories being independent of each other? The expected counts are based on probabilities that are estimated from the observed counts. DF = (#rows - 1)x(#cols - 1)

32

? In 1991 a study was done to assess the possible effects of a new Virginia law requiring the use of seat belts. Historical data for the treatment of drivers in accidents were as follows

Treatment	None	Treated and released	Admitted to hospital	Died
percentage	50%	40%	8%	2%

A random sample of 500 accidents was taken the year after the seat belt law went into effect, with the following results:

Treatment	None	Treated and released	Admitted to hospital	Died
Number	300	165	30	5

Is there a statistically significant change relative to historical percentages?

33

? What if the study had compared a SRS in 1990 to one in 1991?

	None	Treated and Released	Admitted to Hospital	Died
1990	250	200	40	10
1991	300	165	30	5