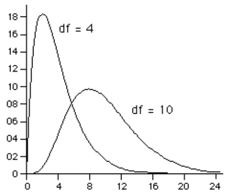
X^2 test



Developed 1900

Karl Pearson 1857-1936

X^2 (chi square) as goodness of fit test

- this term commonly used to refer to Pearson's chi-square, its also known as the goodness of fit test
- allows you to determine if what you observe in a distribution of frequencies would be what you would expect to occur by chance
- nominal data (categories)
- one-sample (1 dimension) and two-sample (2 dimensions)

1 sample test

- Research question: are wheat growing farms located with respect to soil type? That is, is wheat grown in particular soil-type areas?
- 1) take a random sample of 100 wheat farms and determine the soil types underlying the farms
- 2) there are 4 'classes' of soil type

Soil class					
	clay	sand	loam	limestone	
frequency of wheat farms	30	30	30	10	Σ=100

this is the 'observed' distribution of wheat farms

3) under a null hypothesis what would be our 'expected' distribution?
the rationale for the test is that you can compute what you would expect by chance

- you can do this by dividing the total number of occurrences by the number of classes
- so in this case 100/4 = 25 per class
- next we look at how different what we have versus what we expect

Soil class					
'Expected' land under soil type	clay	sand	loam	limestone	
	25	25	25	25	Σ=100

- formula to calculate chi-square
- where O_i = observed value in category I
- E_i= expected value in category I
- k= number of categories
- χ^2= chi square statistic

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

- Now that we have data let's do the test: there are 7 steps
- 1) state the null and alternate hypothesis
  - the null in this case is that there is no difference in the proportion of occurrences in each category: H_0: $P_1 = P_2 = P_3$
- here the percentages of the cases are equal but they needn't be as you'll see
- the number of categories can be as many as you want as long as the categories are mutually exclusive

- the alternate hypothesis is: H_1 $P^1 \neq P^2 \neq P^3$
- 2) set the level of significance (or type I error): α
- typically in geography α =.05 or α =.01
- 3) select the appropriate test statistic
  - any test between frequencies of mutually exclusive categories requires chi square

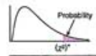
- 4) computation of the test statistic

category	O observed	E expected	D difference	$(O-E)^2$	$(O-E)^2/E$
clay	30	25	5	25	1
sand	30	25	5	25	1
loam	30	25	5	25	1
limestone	10	25	15	225	9
Total					12

- 5) determine the value needed for rejection of the null hypothesis
  - to do this we need the degrees of freedom: here its k-1 or 3
  - using this and the value you picked for α you go to the chi square table
  - with df=3 and α=.05 the critical value=7.815
  - be sure to practice finding values from the table on your own

- 6) compare the calculated value versus the critical value
- calculated = 12, critical = 7.815 so calculated is greater than the critical
- 7) decision time
  - if the calculated value is greater than the critical value then the null hypothesis can't be accepted
  - so what does $\chi^2(3)$ = 12, α=.05 mean?
  - χ^2 is the test statistic
  - 3 is the degrees of freedom
  - 12 is the calculated value

- α=.05 the probability is less than or equal to 5% on any one test of the null hypothesis that the frequency of farms is equally distributed across all categories



χ^2 Critical Values

(Table entry is the point χ^2 with given probability p lying above it.)

df	p .25	.20	..15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72

- there are cases where you might not want to use the number of occurrences/number of categories as you expected value
- if you have some other way of determining what the expected values might be you can use that

- for example: in our case we could use the proportions of different soil types in our study region as our expected values
- this is where the geography in a research question is important
- the distribution of land in each soil type is shown next

class					
	clay	sand	loam	limestone	
actual % of land under soil type	30	40	20	10	Σ=100

- 1) our null hypothesis is that:
- H_0: soil type has no influence on wheat farm location
- if H_0 was true, then we would expect the <u>observed</u> number of wheat farms to be roughly equal to/proportional to actual % of land under particular soil types

observed	30	40	20	10
Expected	30	40	20	10

what we found was

observed	30	30	30	10
Expected	30	40	20	10

- are these differences significant or could they have occurred due to random sampling differences?
- Our alternate is H_1: Soil type has an influence on wheat farm location
- 2) Set significance level at 95% confidence or α=0.05
- 3) use the chi square statistic since we have nominal data with frequencies or proportions

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

category	O observed	E expected	D difference	(O-E)²	(O-E)²/E
clay	30	30	0	0	0
sand	30	40	10	100	2.5
loam	30	20	10	100	5.0
limestone	10	10	0	0	0
Total					7.5

- df= k-1=4-1 = 3
  - df - means that given the total frequency, once the frequencies are known for all but one of the categories, the frequency in the final category is determined
- $X^2_{(3)}$ (α=0.05) = 7.815 same as in previous problem
- if O_i and E_i were equal then χ^2=0
- χ^2 increases as differences increase
- 7.815 defines value of χ^2 where top 5% distribution starts with df=3
- in this case χ^2=7.5
- 6) To reject H_0: calculated value must exceed the critical value

- we cannot reject H_0: cannot say that we would expect the value 7.5 to occur > 95 out of 100
- if H_0 is correct, the probability of 7.5 occurring is > α=0.05
- therefore farming is not related to soil type

rules of thumb

- 1) if the number of categories is greater than 2, no more than 1/5 of the expected frequencies should be less than 5 and none should be 0
- 2) if the number of categories is 2, both the expected and observed frequencies should be 5 or larger
- if this isn't met there is Yate's correction that makes chi-square more conservative - that is more difficult to show significant difference

- also known as continuity correction
- these illustrate an important restriction on χ^2 in that for many categories there should not be small frequencies
- also the data must be in frequencies, χ^2 will give false results if used on proportions or percentages of occurrences in categories
- this last example illustrates a case where you can use external information for choosing your expected values

- this can be extended to cases where you can generate the expected values by referring to a distribution to obtain your expected values
- an example is using the poisson distribution to generate your expected values
- an alternative test for this purpose is the Kolmogorov-Smirnov test (k-s test)

Geographic examples

Observed Frequency Counts: Number of Interprovincial Migrants to British Columbia, Canada, 1991

Province of origin	Number of migrants to British Columbia, 1991
Newfoundland	11
Prince Edward Island	4
Nova Scotia	25
New Brunswick	14
Quebec	50
Ontario	236
Manitoba	73
Saskatchewan	65
Alberta	307
Yukon Territory	10
Northwest Territories	8
Total	803

Geographic models

- Population model
  - This model predicts that the expected number of migrants into British Columbia is directly proportional to the origin provinces' populations
    - Where E_{ij} = expected number of migrants into BC from province I to BC, province j
    - Pop_i= population of province i

$$E_{ij} = Pop_j$$

- Distance model
  - This model predicts that the expected number of migrants into BC is inversely proportional to the square of the distances between each origin province and BC
    - Where D_{ij}^2 = squared distance from origin I to BC

$$E_{ij} = \frac{1}{D_{ij}^{\,2}}$$

- Simplified Gravity model or composite model
  - This model predicts that the number of migrants is a function of both distance and population

$$E_{ij} = \frac{Pop_i}{D_{ij}^{\,2}}$$

Summary Table for Chi-square Goodness-of-Fit Proportional: Interprovincial Migration to British Columbia, Canada, 1991

Province of origin	Observed number of interprovincial migrants	Expected number of interprovincial migrants: "Population" model	"Distance" model	"Composite" model
Newfoundland	11	19	8	13
Prince Edward Island	4	4	12	8
Nova Scotia	25	30	12	21
New Brunswick	14	24	14	19
Quebec	50	231	18	125
Ontario	236	337	21	179
Manitoba	73	37	85	61
Saskatchewan	65	33	137	85
Alberta	307	85	340	213
Yukon Territory	10	1	58	29
Northwest Territories	8	2	98	50
Total	803	803	803	803

"Population" model $E_{ij} = Pop_i$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(11-19)^2}{19} + \frac{(4-4)^2}{4} + \frac{(25-30)^2}{30} + \frac{(14-24)^2}{24} + \frac{(50-231)^2}{231} + \frac{(236-337)^2}{337} + \frac{(73-37)^2}{37} + \frac{(65-33)^2}{33} + \frac{(307-85)^2}{85} + \frac{(10-1)^2}{1} + \frac{(8-2)^2}{2} = 925.33$$

"Distance" model $E_{ij} = \frac{1}{D_{ij}^2}$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(11-8)^2}{8} + \frac{(4-12)^2}{12} + \frac{(25-12)^2}{12} + \frac{(14-14)^2}{14} + \frac{(50-18)^2}{18} + \frac{(236-21)^2}{21} + \frac{(73-85)^2}{85} + \frac{(65-137)^2}{137} + \frac{(307-340)^2}{340} + \frac{(10-58)^2}{58} + \frac{(8-98)^2}{98} = 2443.73$$

"Composite" model $E_{ij} = \frac{Pop_i}{D_{ij}^2}$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(11-13)^2}{13} + \frac{(4-8)^2}{8} + \frac{(25-21)^2}{21} + \frac{(14-19)^2}{19} + \frac{(50-125)^2}{125} + \frac{(236-179)^2}{179} + \frac{(73-61)^2}{61} + \frac{(65-85)^2}{85} + \frac{(307-213)^2}{213} + \frac{(10-29)^2}{29} + \frac{(8-50)^2}{50} = 163.81$$

Summary of Chi-square Differences Not Explained by Spatial Interaction Model: Interprovincial Migration to British Columbia, 1991

Province of origin	Chi-square differences, not explained by model: "Population" model	"Distance" model	"Composite" model
Newfoundland	3.37	1.12	0.31
Prince Edward Island	0.00	5.33	2.00
Nova Scotia	0.83	14.08	0.76
New Brunswick	4.17	0.00	1.32
Quebec	141.82	56.89	45.00
Ontario	30.27	2201.19	18.15
Manitoba	35.03	1.69	2.36
Saskatchewan	31.03	37.84	4.71
Alberta	579.81	3.20	41.48
Yukon Territory	81.00	39.72	12.45
Northwest Territories	18.00	82.65	35.28
Total	925.33	2443.73	163.81

Weaknesses

- X^2 is an absolute measure, the expected values either are or are not statistically different
  - There is no measure of how well the model fits
  - To deal with this we need to use a different approach → PRE (Proportional reduction of error)
  - If the sample size is large we almost always reject H_0