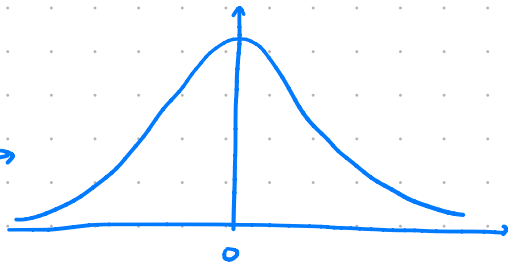
χ^2 CHI-SQUARED (Not "CHEE"!)

Distribution / Tests

Suppose $X_1, X_2, X_3 \ldots, X_n$ are independent and identically distributed (i.i.d.)

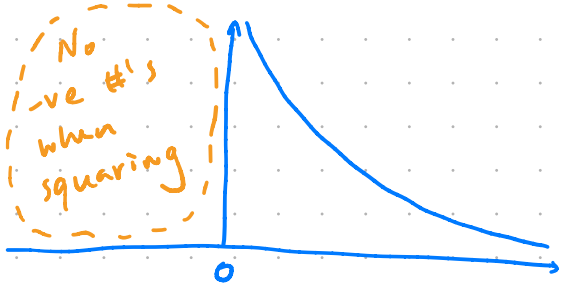
$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$

$Z_i = \frac{X_i - \mu}{\sigma} \overset{i.i.d.}{\sim} N(0, 1)$



↓ square each member

$Z_i^2 \sim \chi_1^2$ ← chi-squared dist^n with "1 degree of freedom"



χ "chi" (pronounced kye)

Just some parameter for now.

$$\sum_{i=1}^{n} Z_n^2 \sim \chi_n^2$$

Sum of n (normal distributions)^2 ↓ χ^2 with n degrees of freedom

As $n \to \infty$, χ_n^2 approaches a normal distribution

χ_n^2 dist^n is used to perform a Goodness-of-Fit Test

Is it reasonable to assume that the observed data has been generated by some statistical method?

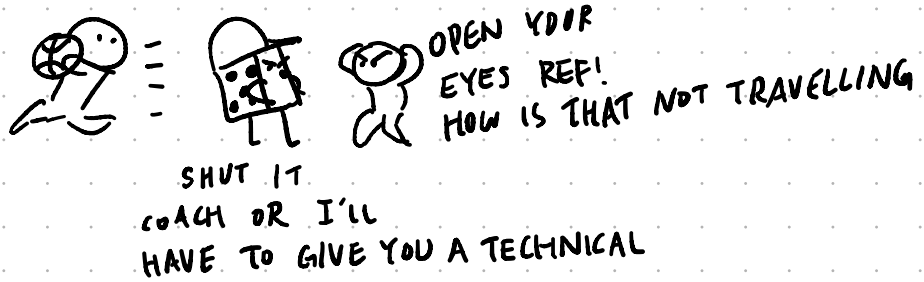
Goodness-of-Fit

How do we decide (objectively) whether a 6-sided die is fair?

1. Roll die many times
2. Compare set of observed outcomes against the set of expected outcomes
3. Develop some measure of similarity to assess the difference

Roll a die 60 times

$\begin{cases} H_0: \text{die is fair} \\ H_1: \text{die isn't fair} \end{cases}$



Category	i	1	2	3	4	5	6	
observed	O_i	12	7	14	10	8	9	(experiment)
expected	E_i	10	10	10	10	10	10	(assuming H_0)

If any E_i are less than 5, combine them!

How similar are O_i and E_i?

$\sum_{i=1}^{6}(O_i - E_i)$ Kind of makes sense. (Sum the differences)

Sometimes +ve, sometimes -ve, so differences would cancel.

$\sum_{i=1}^{6}|O_i - E_i|$ Now always +ve! (I guess...)

$\sum_{i=1}^{6}(O_i - E_i)^2$ Also would work, but not what we do ∴

$\sum_{i=1}^{6}\frac{(O_i - E_i)^2}{E_i} \approx \chi^2_r$ This is what we use!

(Because we know it is distributed some way)

$$\sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}$$ is the goodness-of-fit statistic

What the f**k are degrees of freedom? (degrees of freedom = ν)

# of degrees of freedom = # of categories − # of restrictions

The restriction that's always there: $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i = N$

# of trials is the same for both observed and expected

Ex 6B

(7) $P(X > x) = 0.95 \qquad X \sim \chi^2_5$

Check the table of values ($\nu = 5$, $p = 0.95$)

$P(\chi^2_5 > 1.145) = 0.95 \qquad \therefore x = 1.145$

(8) $Y \sim \chi^2_{12}$ a) $P(Y < y) = 0.5 \quad P(Y \geq y) = 0.95$

$P(\chi^2_{12} > 5.226) = 0.95$

$\therefore y = 5.226$

These come from statistical tables

ν	0.995	0.990	0.975	0.950	0.900	..
1						
2						
3						
4						
5						
⋮						

BLAH BLAH BLAH

Goodness-of-Fit Testing

Expected count in category i = total count × P(fall in category i)

assuming H_0 is true

Ex 6C

① die rolled 72 times

#	1	2	3	4	5	6
observed O_i	16	11	13	15	8	9
expected E_i	12	12	12	12	12	12

$N=72$ H_0: die is fair (follows uniform distn)

H_1: die is unfair (follows non-uniform distn)

$E_i = \frac{1}{6}\times 72 = 12$ $\sum_{r=1}^{6} \frac{(O_i-E_i)^2}{E_i} = \frac{4^2}{12}+\frac{1^2}{12}+\frac{1^2}{12}+\frac{3^2}{12}+\frac{4^2}{12}+\frac{3^2}{12}$

$= 4.33 \leftarrow \chi^2$

N=12 is one restriction $\alpha = 0.05$ (5%)

$\nu = 6-1=5$

↑ 6 categories

from table $P(\chi^2_5 > 11.070) = 0.05$

$\chi^2_5(0.05) = 11.070 \not< 4.33$

∴ We fail to rej. H_0 at $\alpha = 0.05$ or 5%

∴ Insufficient evidence to suggest biased die

② LOTTERY! Tickets fall out of spinning thing (tombola) $N=120$

If ends in 0 or 5, win else, lose

In 10 numbers, 2 are winning

2/10 = 1/5 chance to win if tombola is fair $\alpha = 0.05$

i	Win	Lose
O_i	15	105
E_i	24	96

$\chi^2 = \frac{(15-24)^2}{24} + \frac{(105-96)^2}{96} = 4.21875$

H_0: tombola fair (1/5 chance to win)

H_1: tombola biased (not 1/5)

$\nu = 2-1 = 1$ degree of freedom

$\chi^2_{1,(0.05)} = 3.841 < 4.21875$

∴ Sufficient evidence to rej. H_0

Tombola is BIASED!

④ N=100 households, # of dogs

# of dogs	0	1	2	3	4	5	>5
O_i	45	19	11	8	7	6	4
E_i	55	20	10	7	4	3	1

Combine

	0	1	2	3	≥4
O_i	45	19	11	8	17
E_i	55	20	10	7	8

Combine categories when Any E_i are less than 5

$4+3+1=8$

$\sum \frac{(O_i-E_i)^2}{E_i} \sim \chi^2$ only if $E_i \geq 5$

$\nu = 5-1=4$ ← 5 categories

$\alpha = 5\%$

$\chi^2 = \sum \frac{(O_i-E_i)^2}{E_i}$

$= 12.236$

$\chi^2_{4(0.05)} = 9.488 < 12.236$ ∴ Sufficient disagreement

∴ # of dogs not distributed as such.

Why do we have to combine categories when $E_i < 5$?

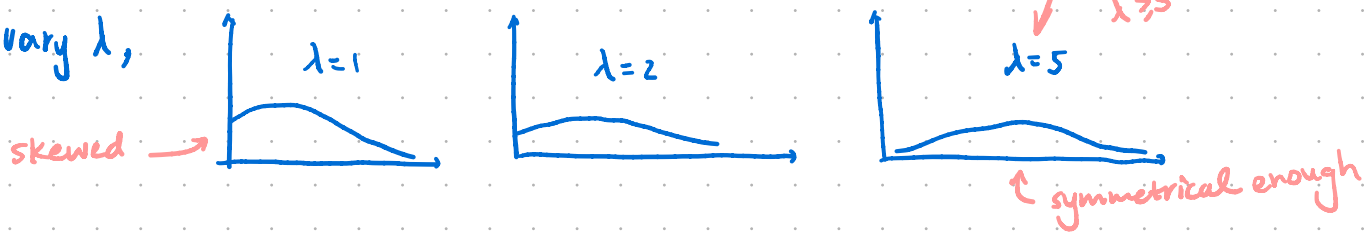
1. Recall that $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$
2. $Z_i = \frac{X_i-\mu}{\sigma} \overset{iid}{\sim} N(0,1)$
3. $Z_i^2 \overset{iid}{\sim} \chi^2_1$
4. $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$

Suppose observed counts: $O_i = 1, 2, 3, \ldots, n$

Assume $O_i \sim Po(\lambda) \rightarrow \lambda = E_i$ (λ is the expected occurrences)

$= E[O_i] = Var(O_i)$

When we vary λ,



If $E_i \geq 5$, $O_i \sim N(E_i, E_i)$

$Z_i = \frac{O_i-E_i}{\sqrt{E_i}} \sim N(0,1)$ $\quad Z_i^2 = \frac{(O_i-E_i)^2}{E_i} \sim \chi^2_1$

$\sum_{i=1}^{n} Z_i^2 = \sum_{E_i}^{n} \frac{(O_i-E_i)^2}{E_i} \sim \chi^2_{n-1}$

If $E_i < 5$, O_i no longer normal, Z_i^2 no longer χ^2 ∴

1 degree of freedom lost ∵ $\sum O_i = \sum E_i$

Ex 6D

(1) a)

x	0	1	2	3	4	5	>5
O_i	12	23	24	24	12	5	0
E_i	13.53	27.07	27.07	18.04	9.02	3.61	1.66

Just a list of $P(X=x) \cdot N$

From $X \sim Po(2)$

H_0: Data can be modelled by $Po(2)$

H_1: Data cannot be modelled by $Po(2)$

How well does this chosen distⁿ fit the data?

$$\chi^2 = \frac{(12-13.53)^2}{13.53} + \ldots + \frac{(5-3.61)^2}{3.61} = 4.101 \qquad \nu = 6-1=5$$

$$\chi^2_{5\,(0.05)} = 11.070 \not< 4.101$$

∴ Fail to reject H_0 at significance level of 0.05

b) Estimate λ (for a better model than $Po(2)$)

$\lambda = E[X]$ $\quad X \sim Po(\lambda)$

$\hat{\lambda}$ (estimate of λ) $= \bar{x} = \frac{12\times0+\ldots+5\times5}{12+23+\ldots+5} = \frac{216}{100} = 2.16$

new estimate of λ

estimates reduce degrees of freedom.

$df = \nu = 6 - 1 - 1 = 4$

another restriction!

$\sum O_i = \sum E_i$

(3) a)

# of accidents	0	1	2	3	4	>4
O_i (weeks)	15	13	9	13	0	0

$$\bar{x} = \frac{13+18+39}{15+13+9+13} = \frac{7}{5} = 1.4$$

b) Approximating using $Po(1.4)$, $(\hat{\lambda} = \bar{x})$

H_0: Data follows Po distⁿ

H_1: Data doesn't follow Po distⁿ

# of accidents	0	1	2	3	4	>4
p	0.2465	0.3452	0.2416	0.1127	0.0394	0.015
E_i	12.325	17.26	12.08	5.635	1.97	0.73

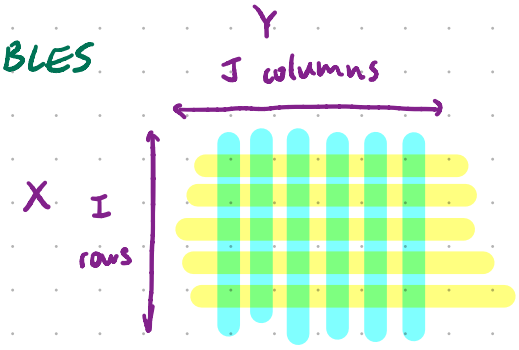
combine

$$\chi^2 = \frac{(15-12.325)^2}{12.325} + \ldots = 5.0282 \qquad df = 4-2=2$$

$\chi^2_{2\,(0.1)} = 4.605 < 5.0282$ ∴ sufficient evidence to rej. H_0

CONTINGENCY TABLES

GENERALISING!



X has I categories

Y has J categories

O_{ij} → observed count in row i column j

Sum of row 1: $\sum_{j=1}^{J} O_{1j} = O_{11} + O_{12} + O_{13} + \dots + O_{1,J-1} + O_{1,J} = O_{1\cdot}$

Sum of column 1: $\sum_{i=1}^{I} O_{i1} = O_{11} + O_{21} + O_{31} + \dots + O_{I-1,1} + O_{I,1} = O_{\cdot 1}$

we have sums!

Sum of rows

X \ Y						
	O_{11}	O_{12}	O_{13}	$\dots$	$O_{1,J-1}$	$O_{1,J}$	$O_{1\cdot}$
	O_{21}	O_{22}	O_{23}	$\dots$	$O_{2,J-1}$	$O_{2,J}$	$O_{2\cdot}$
	O_{31}	O_{32}	O_{33}	$\dots$	$O_{3,J-1}$	$O_{3,J}$	$O_{3\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
	$O_{I-1,1}$	$O_{I-1,2}$	$O_{I-1,3}$	$\dots$	$O_{I-1,J-1}$	$O_{I-1,J}$	$O_{I-1\cdot}$
	$O_{I,1}$	$O_{I,2}$	$O_{I,3}$	$\dots$	$O_{I,J-1}$	$O_{I,J}$	$O_{I\cdot}$
Sum of columns →	$O_{\cdot 1}$	$O_{\cdot 2}$	$O_{\cdot 3}$	$\dots$	$O_{\cdot J-1}$	$O_{\cdot J}$	N ← grand total

$\hat{P}(X=1) = \frac{O_{1\cdot}}{N}$ $\quad \hat{P}(X=i) = \frac{O_{i\cdot}}{N}$

$\hat{P}(Y=1) = \frac{O_{\cdot 1}}{N}$ $\quad \hat{P}(Y=j) = \frac{O_{\cdot j}}{N}$

Assuming Independent

$\hat{P}(X=1 \cap Y=1) = P(X=1) \cdot P(Y=1) = \frac{O_{i\cdot} \times O_{\cdot j}}{N^2}$

$$\hat{E}_{ij} = N \times P(X=1 \cap Y=1) = \frac{O_{i\cdot} \times O_{\cdot j}}{N}$$

use the data to estimate the probability of each cell, that's all.

# of degrees of freedom = # of cells that are free to vary

= total # of cells − # of cells that aren't free to vary

In every row/column, the last cell cannot vary since they have to sum to a number

Total # of cells = IJ cells that can't vary = margins of the table = (I−1)+(J−1)+1

$$\text{\# of degrees of freedom} = IJ - \left((I-1)+(J-1)+1\right) = IJ - I - J + 1 = (I-1)(J-1)$$

Criterion 1 : Grade
Criterion 2 : Teacher

45 pupils cross-classified to 2 categorical variables.

OBSERVED	A*	A	≤B	
Dr Green	0	2	13	15
Mr ?	4	6	5	15
Monkey	13	2	0	15
	17	10	18	45

H_0: grade independent of teacher

H_1: grade dependent of teacher

To test H_0, compare table of observed counts to table of expected counts calculated assuming H_0 is true

EXPECTED	A*	A	≤B	
Dr Green	5.66	3.33	6	15
Mr ?	5.66	3.33	6	15
Monkey	5.66	3.33	6	15
	17	10	18	45

all rows same since independent

Now we test the observed table against expected table (Goodness of fit)

15 people distributed accordly to the column sums

Goodness of fit: $\sum \frac{(O_i - E_i)^2}{E_i}$ for every cell and do the test!

$G = 33.18$ $\quad df = (3-1)(3-1) = 4$

$\chi^2_{4(0.05)} = 9.488 < 33.18 = G$ $\quad \therefore$ sufficient evidence to rej. H_0!

$\therefore$ Grades dependent of teacher