10.11 Goodness of Fit test (Karl Pearson 1900)

Suppose $X_1 \sim B(n, p_1)$. With $\begin{cases} np_1 \geq 5 \\ n(1-p_1) \geq 5 \end{cases}$

$$Z = \frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} \sim N(0,1) \quad \text{(Central limit theorem)}$$

and $$Z^2 = \frac{(X_1 - np_1)^2}{np_1(1-p_1)} \approx \chi^2(1)$$

Now $$\frac{a}{np_1} + \frac{a}{n(1-p_1)} = \frac{a(1-p_1) + ap_1}{np_1(1-p_1)} = \frac{a}{np_1(1-p_1)}$$

Therefore letting $a = (X_1 - np_1)^2$ we get.

$$Q = Z^2 = \frac{(X_1 - np_1)^2}{np_1(1-p_1)} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1-p_1)}$$

But $X_2 = n - X_1$ and $p_2 = 1 - p_1$

We have $(X_1 - np_1)^2 = \left(\overbrace{n - X_1}^{X_2} - n(1-p_1)\right)^2 = (X_2 - np_2)^2$

which gives.

$$Q = Z^2 = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \approx \chi^2(1)$$

Note that Q has the form $\sum_{i=1}^{2} \frac{(O_i - e_i)^2}{e_i}$ (observed O_i, expected e_i).

In a multinominal distribution the sample space consists of k (instead of two) mutually exclusive outcomes, and it can be shown that $Q = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \approx \chi^2(k-1)$

np_i must not be too small ($np_i \geq 5$)

Example. Tossing a die. Let X be the outcome

Want to find out. Is $P(X=x) = \frac{1}{6}$, $x = 1, 2, \ldots, 6$

For 120 tosses we observed

x:	1	2	3	4	5	6
Observed	20	22	17	18	19	24
Expected	20	20	20	20	20	20

The test statistic for a goodness-of-fit test is

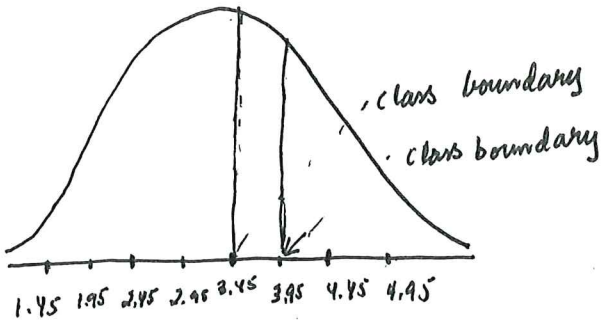
$$Q = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \approx \chi^2_{(k-1)}$$

$$Q_{obs} = \frac{(20-20)^2 + (22-20)^2 + (17-20)^2 + (18-20)^2 + (19-20)^2 + (24-20)^2}{20} \approx 1.7$$

$\chi^2_{0.05}(5) = 11.07$ i.e. no reason to reject H_0: $P(X=x) = \frac{1}{6}$, $x = 1, 2, \ldots, 6$.

Continuous distributions

Lifetime of batteries, $T \sim N(3.5, 0.7^2)$?



Class Boundaries	O_i		e_i		
1.45 – 1.95	2	} 7	0.5	} 8.5	0.2125
1.95 – 2.45	1		2.1		
2.45 – 2.95	4		5.9		
2.95 – 3.45	15		10.3		0.2775
3.45 – 3.95	10		10.7		0.2675
3.95 – 4.45	5	} 8	7.0	} 10.5	0.2625
4.45 – 4.95	3		3.5		

Expected number for the class (cell) 2.95 – 3.45

Probability to fall in the class

$$P(2.95 \le \bar{T} \le 3.45) = P\left(\frac{2.95-3.5}{0.7} \le \frac{\bar{T}-3.5}{0.7} \le \frac{3.45-3.5}{0.7}\right)$$

$$= \Phi(-0.07) - \Phi(-0.79) = 0.4721 - 0.2148 = 0.2573$$

Expected number is $40 \cdot 0.2573 = \underline{10.3}$

Cells with expected number less than 5 should be joined with neighbours. We are left with 4 cells.

1.45 – 2.95, 2.95 – 3.45, 3.45 – 3.95, 3.95 – 4.95

$$Q = \frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} = 3.05$$

$\chi^2_{0.05}(3) = 7.815 \Rightarrow$ No reason to reject H_0: $\bar{T} \sim N(3.5, 0.7^2)$

10.12. Test for Independence for categorical data.

Example: In order to find out if there are dependence between eye colour and haircolour for humans in a given population, 80 persons are randomly drawn from the population.
Eyecolour is classified as blue (B1) or brown (B2) and haircolour as light blond (H1), dark blond (H2) or dark (H3)

The observed data (frequencies) are presented in a contingency table.

Eye-colour \ Hair-colour	H1	H2	H3	Total
B1	29	12	7	48
B2	10	12	10	32
Total	39	24	17	80

This is a 2 x 3 contingency Table

Let us define $P(BI) = p_i$, $I = 1,2$ and $P(HJ) = p_J$, $J = 1,2,3$

Further let $p_{ij} = P(BI \cap HJ)$, $I = 1,2$, $J = 1,2,3$

Independence between eyecolour and haircolour can be defined as $p_{ij} = p_i \cdot p_J$, $\forall i,j$

H_0: $p_{ij} = p_i \cdot p_J$ $\forall i,j$, $i = 1,2$, $J = 1,2,3$.

The (estimated) expected cell numbers are $80 \cdot \hat{p}_{ij}$, $i = 1,2$, $J = 1,2,3$

Under H_0 this becomes $80 \cdot \hat{p}_i \cdot \hat{p}_J = 80 \cdot \frac{n_i}{80} \cdot \frac{c_J}{80} = \frac{n_i \cdot c_J}{80}$ (row total ↓, column total ↙)

or $\frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}$

The 2×3 contingency Table with observed and expected numbers

	H1	H2	H3	Total
B1	29 (23.4)	12 (14.4)	7 (10.2)	48
B2	10 (15.6)	12 (9.6)	10 (6.8)	32
Total	39	24	17	80

Ehs. $23.4 = \frac{48 \cdot 39}{80}$, $15.6 = \frac{32 \cdot 39}{80}$

$$Q = \frac{(29-23.4)^2}{23.4} + \frac{(12-14.4)^2}{14.4} + \frac{(7-10.2)^2}{10.2} + \frac{(10-15.6)^2}{15.6}$$

$$+ \frac{(12-9.6)^2}{24} + \frac{(10-6.8)^2}{6.8} = 6.86.$$

The degrees of freedom $\nu = (r-1)(c-1)$ (number of rows, number of columns)

$$\chi^2_{0.05}(2) = 5.9915$$

i.e. ~~the~~ Ho (independence) is rejected on a 5% level.

How to find the degrees of freedom. We have rc cells and we are estimating $r-1 + c-1$ parameters

Thereby $DF = rc - 1 - (r-1) - (c-1) = rc - r - (c-1)$

$= r(c-1) - (c-1) = \underline{(r-1)(c-1)}$