50008 - Probability and Statistics - Lecture 8

Oliver Killane

08/03/22

Goodness of Fit

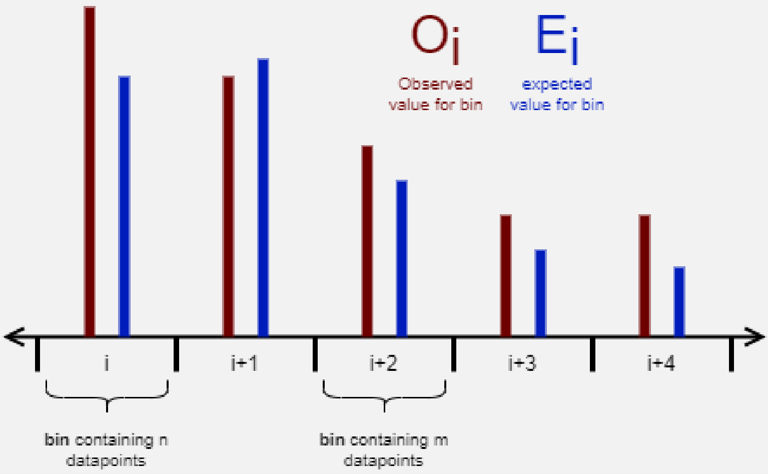
Lecture Recording

Lecture recording is available here

Definition: Binning

Given a distribution, we can partition it into several disjoint **bins**. Essentially we are creating a pesudo-**PMF** (potentially with ranges instead of just discrete values) describing how many datapoints/the frequency we would expect to find from a distribution.

As a result, we can directly compare the expected values E_i (from a distribution we are checking a sample against), with the observations O_i from a sample.



Definition: Goodness of Fit/Chi-Square Statistic

Denotes the difference between some expected values, and some observed.

For n bins we have:

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

Chi-Squared Test for Model Checking

Used to determine if an observed sample matches a given distribution to some significance.

1. Determine expected distribution (can use parameters estimated from the sample).
2. Create a hypotheses based some parameters θ:

$$H_0 \ : \ \theta = \theta_0 \ \text{ versus } \ H_1 \ : \ \theta \neq \theta_0$$

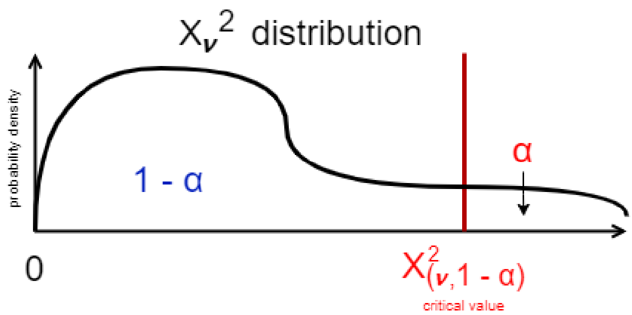
3. Bin the expected distribution (for comparison with the observed).
4. Calculate the **Goodness of Fit/Chi-Square Test Statistic** X^2.
5. Calculate the degrees of freedom as:

$$\nu = (\text{number of possible values } X \text{ can take}) - (\text{number of parameters being estimated}) - 1$$

6. Determine the critical value using the **Chi Squared Distribution** χ^2_ν and the significance α (typically using a table).
7. If $X^2 > \chi^2_{\nu,\ 1-\alpha}$ (test statistic larger than critical value)

Note that:

- All expected values must be larger than 5 for a good test. Hence some bins may have to be merged.
- The number of values X can take is typically the number of bins.



Example: Adverse Drug Effects

A study in the journal of the American Medical Association gives the causes of a sample of 95 adverse drug effects as:

Reason	**No. Adverse Effects**
Lack of Knowledge	29
Rule Violation	17
Faulty Dose Check	13
Slips	9
Other Cause	27

Test if the true percentages of causes of adverse effects are different at the 5% significance.

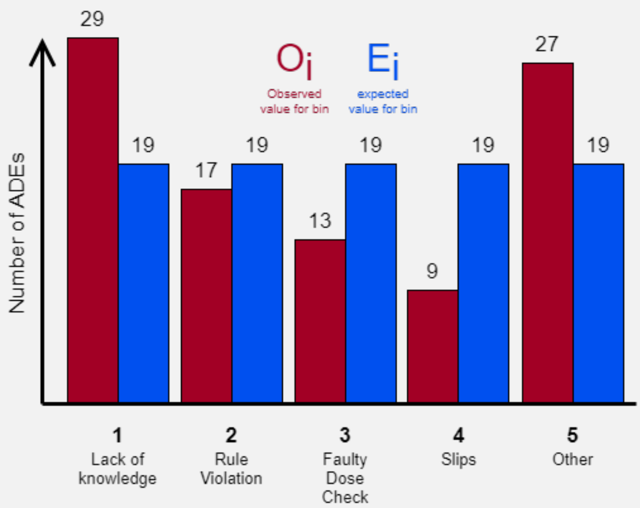
As we are checking the percentages are the same, we effectively have a discrete uniform distribution:

$$X \sim U(1, 5)$$

Hence we can calculate our **null and alternative hypotheses**:

$$H_0 \; : \; X \sim U(1, 5) \;\; \text{versus} \;\; H_1 \; : \; X \nsim U(1, 5)$$

Now we can bin the distribution, (no merging is required as all expected values are larger than 5):



It is now possible to compute goodness of fit.

$$\begin{aligned} X^2 &= \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(29-19)^2}{19} + \frac{(17-19)^2}{19} + \frac{(13-19)^2}{19} + \frac{(9-19)^2}{19} + \frac{(27-19)^2}{19} \\ &= 16 \end{aligned}$$

We have $\nu = 4$ as there are 5 possible values, and no parameters were estimated from the data.

Hence we get the critical value from the chi-squared table: $\chi^2_{4,\ 0.95} = 9.49$

As $16 > 9.49$ there is sufficient evidence at the 5% significance level to reject H_0, the percentages differ.

Lecture Recording

Lecture recording is available here

Example: Football Games

Given the total number of goals for 2608 football matches, determine if the number of goals scored in a match can be modelled by $X \sim Poisson(3.870)$ at the 5% significance.

Goals Scored (x)	0	1	2	3	4	5	6	7	8	9	≥ 10	
Matches (n_x)	57	203	383	525	532	408	273	139	139	45	27	16

Hence as we already have a distribution, we can create our hypotheses:

$$H_0 \; : \; X \sim Poisson(3.870) \;\; \text{versus} \;\; H_1 \; : \; X \not\sim Poisson(3.87)$$

We can then use the poisson distribution to calculate the expected for 2608 football matches, for the final (≥ 10) we use the cumulative to get the remaining probability.

Goals	0	1	2	3	4	5	6	7	8	9	≥ 10
O	57	203	383	525	532	408	273	139	45	27	16
E	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1
$\dfrac{(O-E)^2}{E}$	0.124	0.267	1.461	0.000	1.096	0.534	1.452	0.012	7.723	0.166	0.071

Hence we get our test statistic as: $X^2 = 12.906$.

As we did not estimate any parameters from the sample, the degrees of freedom are $\nu = 11 - 1 = 10$.

The critical value is: $\chi^2_{10,\ 0.95} = 16.91$.

Hence as $12.906 < 16.91$ we there is insufficient evidence as the 5% significance to reject H_0, the goals can be modelled as $Poisson(3.87)$.

Chi-Squared Test for Independence

Lecture Recording

Lecture recording is available here

Definition: Contingency Table

A table denoting the frequency of each combination of values for X and Y.

		Possible values of y				Marginal
		y_1	y_2	$\dots$	y_l	
Possible x	x_1	$n_{1,1}$	$n_{1,2}$	$\dots$	$n_{1,l}$	$n_{1,\bullet}$
	x_2	$n_{2,1}$	$n_{2,2}$	$\dots$	$n_{2,l}$	$n_{2,\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_k	$n_{k,1}$	$n_{k,2}$	$\dots$	$n_{k,l}$	$n_{k,\bullet}$
Marginal		$n_{\bullet,1}$	$n_{\bullet,2}$	$\dots$	$n_{\bullet,l}$	n

We can use the marginal values to determine the expected value, if the two distributions were independent.

Given a dataset of points $(x, y)_1, (x, y)_2, \dots, (x, y)_n$, we can consider it the joint distribution P_{XY} of the distributions P_X and P_Y.

To test if the distributions P_X and P_Y are independent from the sample (without knowing the actual distributions themselves) we can use a **contingency table**.

For the contingency table entry coordinates $0 < i \leq l$, $0 < j \leq k$:

$$O_{i,j} = n_{i,j} \quad \text{and} \quad E_{i,j} = \frac{n_{i,\bullet} \times n_{\bullet,j}}{n}$$

Hence we can now compute the X^2 (**Chi Squared test statistic**) using these observed and expected values.

We compute the degrees of freedom as $\nu = (rows - 1) \times (columns - 1)$ (each row and column alone has degrees of freedom $n - 1$ as they must sum to the row/column total), and can then do the **Chi-Squared Test** normally.

Example: Fitness and Stress

	Poor Fitness	Average Fitness	Good Fitness	
Stress	206	184	85	475
No Stress	36	28	10	74
	242	212	95	549

Determine at the 5% significance if there is a link between fitness and stress.

For this test the null hypothesis will be that fitness and stress are independent.

H_0 : Stress and fitness are independent versus H_1 : Stress and Fitness re not independent

Next we can calculate the expected values:

	Poor Fitness		Average Fitness		Good Fitness		
	O	E	O	E	O	E	
Stress	206	209.4	184	183.4	85	82.2	475
No Stress	36	32.6	28	28.6	10	12.8	74
	242		212		95		549

We can then calculate our test statistic to be $X^2 = 1.133$.

To compute the degrees of freedom $\nu = (2 - 1) \times (3 - 1) = 2$.

Hence we can get our critical value $\chi^2_{2,\ 0.95} = 5.99$.

As $5.99 > 1.133$, there is insufficient evidence to reject H_0 at the 5% significance level. Stress and fitness are not linked.