CS 536 : Hypothesis Testing and the Typicality of Data

16:198:536

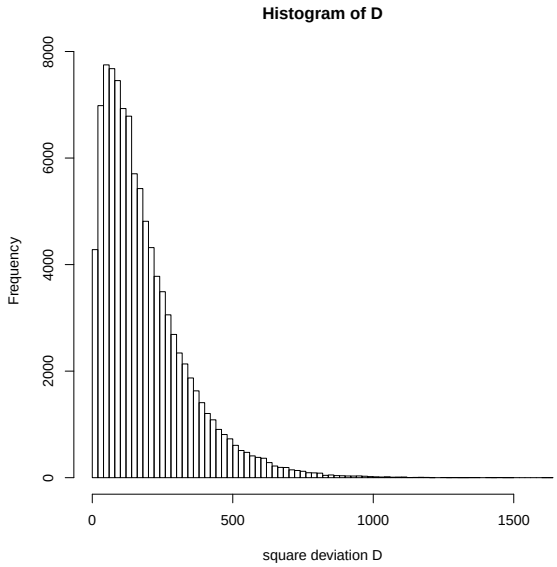
Imagine the following: you take two fair coins, and flip them both, recording the results. Because they're fair and independent, you expect Heads Heads (HH), Heads Tails (HT), Tails Heads (TH), and Tails Tails (TT) each with equal probability of 1/4. But if you were to repeat this experiment multiple times, say flipping both coins $n = 1000$ times, while the 'expected' number of each occurrence would be 250, you would not necessarily be surprised if you saw 231 instances of HH or 260 instances of HT. We naturally expect some variation because the data is random. Simulating this experiment with a random number generator yielded the following occurrence data:

$$\begin{aligned} O_{HH} &= 236 \\ O_{HT} &= 265 \\ O_{TH} &= 240 \\ O_{TT} &= 259. \end{aligned} \tag{1}$$

If we expect fluctuations around the expected value $E = 250$, can we say anything at all about what 'typical' data should look like? If we got for instance $O_{HH} = 1000$ and all the other occurrences were zero - technically, this is *possible*, but this kind of extreme fluctuation would certainly not be typical. So what do typical fluctuations look like? One way we could approach this problem is to simulate the experiment multiple times, and look at the distributions for this deviation. Let D be the average square deviation or

$$D = \frac{1}{4}\left[(O_{HH} - 250)^2 + (O_{HT} - 250)^2 + (O_{TH} - 250)^2 + (O_{TT} - 250)^2\right] \tag{2}$$

Repeating this experiment 100000 times generates the following histogram for D: From these results, we see that



'most' of the time, the deviation is less than 500. Over the simulations run, the average deviation was 188.036, and the median deviation was about 147.5. It is incredibly rare that the deviation exceeds 500 (about 4.7% of the time), and rarer still that it exceeds 1000 (about 0.12% of the time) .

This starts to justify why 1000 instances of HH would be so surprising - in this case, we would get $D = (1/4)((1000 - 250)^2 + 3 * 250^2) = 187500$. From our data, this would be an incredibly rare, extreme value of the deviation.

One way we can apply this observation is with the notion of **hypothesis testing**. Suppose that someone gives us data that they claim was generated with two independent, fair coins as above. If they are telling the truth, we would expect (over 95% of the time) that the deviation D for their data should be less than $D = 500$. If the data they gave us had $D = 600$ for instance, this is surprising - and we might start to think that their data was *not* generated by flipping two fair independent coins. The data they are giving us is *exceedingly unlikely* to have been generated in this way.

Can we quantify this more precisely? We would like to identify a threshold value D_0 such that if $D < D_0$, we conclude that the data is fairly typical for data generated by flipping two fair, independent coins, and if $D \geq D_0$, we conclude that the data is fairly atypical - and we would be tempted to reject the hypothesis that the coins were independent and fair. The typical standard for this kind of decision is to identify the point where $D \geq D_0$ only 5% of the time - this is the classic scientific threshold or p-value for 'statistical significance' (though in fact it is arbitrary, and you can set the threshold as you see fit). Estimating this from the data of the experiments gives a value of $D_0 \approx 492.5$. Note - it is entirely possible for fair, independent coins to generate deviations at 492.5 or even higher! However, these extreme deviations are *rare* and *surprising*, and therefore warrant a careful look at the underlying hypothesis.

Testing for Independence

For the purpose of this section, consider X and Y to be random variables with binary outcomes, indicated with 0 *and* 1.

Suppose we have data of the form of occurrences again - out of N instances, we have O_{xy} occurrences of $X = x, Y = y$. We can consider the following question: does the data support the hypothesis that X and Y are independent? This kind of question frequently arrises in problems like classification, we we want to determine whether a feature tells us anything about a property we would like to predict.

If we knew the probabilities associated with X and Y, for instance, $P(X = 1) = p, P(X = 0) = 1 - p$ and $P(Y = 1) = q, P(Y = 0) = 1 - q$ for known values of p and q, we could approach this question in much the same way as in the previous section:

- Simulate N instances of (X, Y) pairs
- Generate data on the 'typical' deviation between $O_{x,y}$ and the expected number of occurrences $E_{x,y} = P(X = x)P(Y = y) * N$
- Estimate how likely the actual or observed deviation was based on these simulations
- If the observed deviation D is 'extreme', we conclude that this data is very unlikely if X and Y are actually independent

However, if all we have is the data $\{O_{xy}\}$ and the hypothesis that X and Y are independent, we cannot implement this process directly because we don't know the underlying probabilities for X and Y - the values p and q are unknown. In this case, the best we can do is to estimate p and q from the data, in particular

$$\begin{aligned} \hat{p} &= \frac{O_{10} + O_{11}}{N} \\ \hat{q} &= \frac{O_{01} + O_{11}}{N}. \end{aligned} \tag{3}$$

With these estimates, we are in a position to be able to implement the simulation solution outlined above.

However, an alternative approach exists in the form of the χ^2-test (chi-squared test) for independence. In particular, instead of the average deviation outlined previously, consider a standardized or normalized deviation, defined as

$$T = \sum_x \sum_y \frac{(O_{xy} - E_{xy})^2}{E_{xy}}. \tag{4}$$

It can be shown that this test statistic has approximately a χ^2_1-distribution (chi-square with one degree of freedom). The more samples, the more accurate this approximation is. With an explicit approximation for the distribution of this test statistic, we do not need to perform any simulations, we can compute the probability of being greater than some threshold directly: in particular, we have that

$$\mathbb{P}\left(T \geq T_0\right) \approx \int_{T_0}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{t}} e^{-t/2} dt. \tag{5}$$

For a particular p-value such as 5%, you can compute from this integral that the normalized deviation T should be greater than about $T_0 = 3.841$ approximately 5% of the time. This again gives us a simple test for accepting or rejecting the hypothesis of independence at this significance level:

- Estimate the underlying probabilities from the frequencies
- Compute the normalized deviation T based on these estimates
- If $T \geq 3.841$, this magnitude of deviation would be rare if X and Y are independent - hence, we reject the hypothesis of independence
- If $T < 3.841$, this magnitude of deviation would be common if X and Y are independent - hence, we fail to reject the hypothesis of independence

Different thresholds of significance ($1\%, 0.5\%$, etc) will require different threshold values T_0, but all these can be worked out from the known density of the χ^2_1-distribution.

Note: It is worth emphasizing that this is an *approximation* - the true distribution of the test statistic will be different, depending on the number of samples, etc, but this serves as a good approximation, especially in the case of a large number of samples. This is not the only statistical test for independence - others exist, and more precise ones that work better with small sample sizes - but they generally will all have this same structure: a test statistic for the data is computed, compared to some known distribution, and the hypothesis of independence rejected or not depending on the magnitude of the test statistic and the significance level chosen.

Pearson's χ^2-test for Independence Suppose X has c many possible values, and Y has r many possible values. Let O_{xy} be the number of occurrences of $X = x, Y = y$ in N samples. We can approximate the underlying probabilities via the frequencies:

$$\begin{aligned} P(X = x) &= \frac{\sum_y O_{xy}}{N} \\ P(Y = y) &= \frac{\sum_x O_{xy}}{N}, \end{aligned} \tag{6}$$

and estimate the expected occurrences with $E_{xy} = P(X = x)P(Y = y)N$. The value of the test statistic is then

$$T = \sum_x \sum_y \frac{(O_{xy} - E_{xy})^2}{E_{xy}}, \tag{7}$$

which has an approximate distribution given by $\chi^2_{(r-1)(c-1)}$ - a chi-square distribution with $(r-1)(c-1)$ degrees of freedom. The value of the test statistic can then be compared to this distribution to determine if it represents typical or atypical results, at a specified significance level.