Practical Probability for Data Science

Want to know how well data fits a model.

Three big metrics.

1. Test for independence
2. Goodness of fit tests
3. p-value

Main idea involves binary hypothesis testing.

Given n iid rvs $Y_1, \ldots, Y_n$ where $Y_k \in \{1, \ldots, L\}$.

How well does pmf $p_1, \ldots, p_L$ fit data?

Define null hypothesis

$$H_0: \Pr(Y = i) = p_i, \qquad i = 1, \ldots, L$$

Aim: given dataset, we either

- reject null hypothesis
- OR, based on dataset, "cannot reject null hypothesis"

Note: pmf p has $L-1$ independent parameters

Basic Template: hypothesis testing using "Pearson statistic"

Let H_0 be the hypothesis that something is true with m parameters.

Step 1. Assuming H_0, based on n iid data points, compute Pearson statistic

$$T = n \sum_{i=1}^{m} \frac{(\text{observed avg}_i(n) - \text{expected}_i)^2}{\text{expected}_i}$$

Let t = realized value of T.

Step 2. If t is large, H_0 is unlikely. Quantify this as follows: Compute

$$p\text{-value} = \Pr(T \geq t \mid H_0)$$

} probability of obtaining result at least as extreme as observed result ASSUMING H_0 is true

Clearly if t is large, then p-value is small. ($t = \infty \Rightarrow$ p-value $= 0$)

Typically if p-value < 0.05, then reject H_0

Step 3. Compute p-value. Need CDF of T under H_0.

$$T = Y_n' Y_n \sim \chi^2_m \quad \text{where} \quad Y_n = \sqrt{n}\,\Sigma^{-1/2}(\mu_n - \mu)$$

So for large n, CLT implies

$$p\text{-value} = P_r(T > t \mid H_0) = 1 - \chi^2_m(t)$$

where $\chi^2_m(t)$ is the chi-squared cdf w/ m-degrees of freedom evaluated at t.

Aside: χ^2 Distribution

Chi-squared with k degrees of freedom is

$$Z = \sum_{i=1}^{k} X_i^2 \quad , \quad X_i \sim N(0,1) \text{ iid}$$

We denote $Z \sim \chi^2_k$ with pdf

$$f_k(z) = \frac{z^{k/2-1}\, e^{-z/2}}{2^{k/2}\,\Gamma(k/2)} \quad , \quad z > 0$$

Notes

χ^2_1 is pdf of square of normal

χ^2_2 is exponential random variable

χ^2 Goodness of Fit for Discrete RVs

Given $\{Y_i\}_{i=1}^n$ iid rvs where $Y_k \in [L]$, how well does pmf $p_1, \ldots, p_L$ fit data?

Define null hypothesis

$$H_0: \Pr(Y=i) = p_i \quad , \quad i = 1, \ldots, L$$

let

$$N_i = \sum_{k=1}^{n} I(Y_k = i) \quad , \quad \hat{p}_i = \frac{N_i}{N}$$

By law of large numbers

$$\mathbb{E}[N_i] = np_i \qquad \text{(for large } n\text{)}$$

Define statistic

$$T = \sum_{i=1}^{L} \frac{(N_i - np_i)^2}{np_i} = n \sum_{i=1}^{L} \frac{(\hat{p}_i - p_i)^2}{p_i}$$

We reject H_0 if T is large. Note that T has $L-1$ degrees of freedom.

Supposing actual data yields value of t, p-value is

$$\text{p-value} = \Pr(T \geq t \mid H_0)$$

1. As $t \uparrow$, p-value $\downarrow$.
2. Reject H_0 if p-value $< \alpha$ where α is significance level specified by data analyst. Typically $\alpha = 0.05$

p-values. Some Intuition

p-value = $Pr(T \geq t \mid H_0)$

i.e the conditional probability of observing a deviation from H_0 which is greater than observed result t, given H_0 is true.

p-value = $Pr(\text{Obs difference} > t \mid \text{actual difference} = 0)$

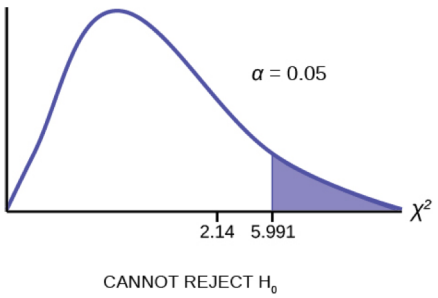
Typical jargon in hypothesis testing is

1. Type I error := reject true null hypothesis
2. Type-II error := accept false null hypothesis

In english,

1. false positive
2. false negative

If you reject H_0 when p-value $< \alpha$, then $Pr(\text{Type I error})$ will NEVER exceed α. α is called the significance level of test.



1. If $t = 2.14$, then cannot reject H_0 since p-value > 0.05.
2. If $t > 5.99$, reject H_0 since p-value < 0.05. The test is called *statistically significant.*

Caution

p-values widely misused.

$Pr(\text{data} \mid H_0)$ NOT $Pr(H_0 \mid \text{data})$.

Example

Suppose $Y_k \in \{1,2,3,4,5\}$ for $k = 1, \ldots, 50$.

Aim: Test null hypothesis H_0 that $p_i = Pr(Y=i) = 1/5$

given following dataset of 50 samples:

$$N_1 = 12, \quad N_2 = 5, \quad N_3 = 19, \quad N_4 = 7, \quad N_5 = 7$$

Solution:

$$T = \sum_{i=1}^{5} \frac{(N_i - np_i)^2}{np_i} = 12.8$$

$\ell = 5 \Rightarrow 4$ dof

$$\Rightarrow \text{p-value} \approx 1 - \chi^2_4(12.8) = 0.0122$$

If $\alpha = 0.05$,

$$\text{p-value} < \alpha, \quad \text{reject } H_0$$

So, we have the Pearson statistic $T = \sum_{i=1}^{\ell} \frac{(N_i - np_i)^2}{np_i}$

and define a p-value

$$p = Pr(T \geq t \mid H_0) \approx 1 - \chi^2_{\ell-1}(t)$$

Why is this true?

Proof

Define $\hat{p}_i = \frac{N_i}{n}$. Then, under H_0

$$T = n \sum_{i=1}^{L} \frac{(\hat{p}_i - p_i)^2}{p_i}$$

$$= n \sum_{i=1}^{L-1} \frac{(\hat{p}_i - p_i)^2}{p_i} + n \frac{(\hat{p}_L - p_L)}{p_L}$$

$$= n \sum_{i=1}^{L-1} \frac{(\hat{p}_i - p_i)^2}{p_i} + n \frac{\left[\sum_{i=1}^{L-1} (\hat{p}_i - p_i)\right]^2}{p_L}$$

define $\hat{p} = [\hat{p}_1 \cdots \hat{p}_{L-1}]^T$, $p = [p_1 \cdots p_{L-1}]^T$

$$= n (\hat{p} - p)^T \Sigma^{-1} (\hat{p} - p)$$

$\Sigma = \mathbb{E}[(\hat{p} - p)(\hat{p} - p)^T]$

Then CLT implies, for sufficiently large n,

$$Y_n = \sqrt{n}\, \Sigma^{-1/2} (\hat{p} - p) \rightarrow N(0, I_{L-1})$$

Therefore

$$T = Y_n^T Y_n \xrightarrow{n} \chi^2_{L-1}$$

That is, sum of squares of $L-1$ iid Gaussian rvs has χ^2_{L-1} distribution

Test for Independence for Discrete RV

Two rvs X, Y are independent if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.

If X, Y independent, then

(i) $F_{X,Y}(x,y) = F_X(x) F_Y(y)$

(ii) $f_{X|Y}(x|y) = f_X(x)$

(iii) $F_{X|Y}(x|y) = F_X(x)$

Pearson's Test for Independence

Given n datapoints, let

$$N_{i,j} = \sum_{i=1}^{L} \sum_{j=1}^{M} I(X_k = i, Y_k = j)$$

$$\hat{p}_{i,j} = \frac{1}{n} N_{i,j}$$

$$\hat{p}_i = \sum_{j=1}^{M} \hat{p}_{i,j}, \qquad \hat{q}_j = \sum_{i=1}^{L} \hat{p}_{i,j}$$

1. Define $H_0: X \perp\!\!\!\perp Y$
2. Compute test statistic

$$T = \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$T = \sum_{i=1}^{L} \sum_{j=1}^{M} \frac{(N_{i,j} - n\hat{p}_i \hat{p}_j)^2}{\hat{p}_i \hat{p}_j} = n \sum_{i=1}^{L} \sum_{j=1}^{M} \frac{(\hat{p}_{i,j} - \hat{p}_i \hat{q}_j)^2}{\hat{p}_i \hat{q}_j}$$

3. For large n, $T \sim \chi^2_{(L-1)(M-1)}$ distribution.

A powerful tool for independence!

Example: Is $\overset{Y}{\text{age}} \perp\!\!\!\perp \overset{X}{\text{desire to ride a bicycle}}$

	$Y=[18,24]$	$Y=[25,34]$	$Y=[35,49]$	$Y=[50,64]$	Total
$X=$ yes	60	54	46	41	201
$X=$ no	40	44	53	57	194
Total	100	98	99	98	395

Null Hypothesis: $X \perp\!\!\!\perp Y$

From data, $t = 8.006$. So $L=2$, $M=4 \rightarrow 3$ dof

$$Pr(T \geq t \mid H_0) \approx 1 - \chi^2_3(8.006) < 0.05$$

At 0.05 significance level, reject null hypothesis.
Conclude that desire to ride bike depends on age.

Degrees of Freedom for Pearson Independence Test

1. $LM-1$ dof from joint PMF $P(X,Y)$
2. Also estimate $L-1+M-1$ parameters $\hat{p}_i$ and $\hat{q}_j$
   → gives $L-1+M-1$ additional constraints
3. So TOTAL dof
   $$= LM-1-(L-1+M-1) = (L-1)(M-1)$$