Stats 135, Fall 2019
Lecture 23, Friday, 10/25/2019

1 Pearson Chi Square T.S.

Today we'll be looking at the 'goodness of fit χ^2 test'. In chapter 11, we did 1 and 2 sample t tests for the mean of a normal random variable. Our box had a continuous number of tickets with mean μ and variance σ^2.

Now we have a categorical random variable having m outcomes having probabilities $p_1, \dots, p_m$. The chance a sample of size n for our box has a certain composition is given by the multinomial formula

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_m = x_m) = \binom{n}{x_1, \dots, x_m} p_1^X x_1 \cdots p_m^{x_m}.$$

Our goal is to test whether a model for the population distribution $(p_1, \dots, p_m)$ fits our data. We draw n times with replacement from our box and get observed counts $[x_1 \mid x_2 \mid \cdots \mid x_m]$. This is $x_1 + \cdots + x_m = n$. If the probability of tickets in the box (with respect to some parameter θ) is $\mathbb{P}_1(\theta), \dots, P_m(\theta)$, we expect to get binomial counts $[np_1(\theta \mid np_2(\theta) \mid \cdots \mid np_m(\theta))]$. We want to compare this expectation to the observed (above), and we do this via a Pearson Chi Square T.S. goodness of fit test. That is,

$$\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{m-1-k},$$

where k is the time dimension of θ. A goodness of fit test explores how good your probability model fits your data. If the p-value of our test is smaller, then we reject our null hypothesis.

2 Example

We looked at the arrivals of alpha particle emissions. We look at a box via a poisson process in our null box, given by the discrete Poisson random variable. We take 1207 draws with replacement into our observed box, where

$$[x_1 = 18 \mid x_2 = 28 \mid \cdots \mid x_{16} = 5].$$

The null is that our observed counts come from the multinomial distribution $MN(1207, P_1(\lambda), P_2(\lambda), \dots, P_{16}(\lambda))$, where $P_i(\lambda)$ is Poisson(λ).
Our alternative is that our observed counts come from some other multinomial distribution (not by this model).

We calculate the χ^2 statistic:

$$\sum_{\text{all rows}} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{\underbrace{16-1}_{m-1} - \underbrace{1}_{k}},$$

which is random because we have to estimate the Poisson distribution. Then we compare p value to $\alpha = 0.05$.

3 Topics Today

First we'll look at some more examples, and then we will develop some theory. We will see that

$$-2\log\Lambda \approx \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

4 Hardy Weinberg Equilibrium Model

This is a famous example. If the gene frequencies are in equilibrium, the genotypes AA, Aa, aa occur in the population with probability $(1-\theta)^2, 2\theta(1-\theta), \theta^2$, according the the Hardy Weinberg (HW) model.

Our null box is

$$[P_1(\theta) = (1-\theta)^2 \mid P_2(\theta) = 2\theta(1-\theta) \mid P_3(\theta) = \theta^2].$$

We observe the blood type

$$[M = AA = 342 = x_1 \mid MN = Aa = 500 = x_2 \mid N = aa = 187 = x_3],$$

where the total $n = 1029$.

4.1 Step 1

Find $\hat{\theta}_{ML}$. We take the likelihood:

$$\text{lik}(\theta) = \frac{n!}{x_1!x_2!x_3!}(1-\theta)^{2x_1}(2\theta(1-\theta))^{x_2}\theta^{2x_3}.$$

Then we take the log-likelihood:

$$\ell(\theta) = \log n! - \sum_{i=3}^{3} \log x_1! + x_1\log(1-\theta)^2 + x_2\log 2\theta(1-\theta) + x_3\log\theta^2,$$

and taking the derivative $\ell'(\theta) = 0$ implies:

$$\hat{\theta}_{ML} = \frac{2x_3 + x_2}{24} = \frac{2(187) + 500}{2(1029)} = \boxed{.4277}$$

Now we have the ML estimator, so we can write out the probabilities:

$$\mathbb{P}_1(\hat{\theta}) = (1 - .4277)^2 = .3200$$
$$\mathbb{P}_2(\hat{\theta}) = 2(.4277)(1 - .4277) = .4887$$
$$\mathbb{P}_3(\hat{\theta}) = (.4277)^2 = .1804.$$

So we have our null box.

4.2 Step 2

Now we can make our expected box:

$$[n\mathbb{P}_1(\hat{\theta}) = 340.57 \mid 502.83 \mid 185.60].$$

Then we can perform the test:

$$H_0: \text{ we have } MN(1029, .32, .49, .18)$$
$$H_A: \text{ we have some other } MN \text{ with } n = 1029.$$

Now with both boxes

$$\begin{aligned} O &: [342 \mid 500 \mid 187] \\ E &: [340.5 \mid 502.8 \mid 183.6] \end{aligned}$$

Our χ^2 test will have $3 - 1 - 1 = 1$ degree of freedom. The null space is 1 dimensional because p was determined by θ. So we can calculate

$$\begin{aligned} \chi_1^2 &= \frac{(342 - 340.6)^2}{340.6} + \cdots + \cdots \\ &= \boxed{0.0357} \end{aligned}$$

Then we can look at the χ_1^2 plot, and the p-value is the area under the curve to the right of 0.0357. This can be computed in R via

$$\begin{aligned} p - \text{value} &= \mathtt{1} - \mathtt{pchisq(0.0357)} \\ &= 0.85 > \alpha \end{aligned}$$

Hence we can accept our null hypothesis that our genes has a distribution described by Hardy Weinberg equilibrium.

5 Developing Theory

Take H_0: have a multinomial (MN) distribution of cell probabilites

$$\mathbb{P}(\theta) = (\mathbb{P}_1(\theta), \ldots, \mathbb{P}_m(\theta)),$$

where $\theta = (\theta_1, \ldots, \theta_k) \in \omega_0 \subseteq \mathbb{R}^k$. Our alternative is H_1: have a MN distribution with different cell probabilities than our null.
Here, the sample space Ω is the set of m nonnegative numbers that sum to 1. In our present example, this is $(p_1, \ldots, p_m)$ with $p_1 + p_2 + \cdots + p_m = 1$.
Recall that

$$\Lambda = \frac{\max_{\theta \in \omega_0} (\text{lik}(\mathbb{P}_1(\theta), \ldots, \mathbb{P}_m(\theta))}{\max_{(p_1, \ldots, p_m) \in \Omega} (\text{lik}(p_1, \ldots, p_m))}.$$

The numerator is equal to $\text{lik}(\mathbb{P}_1(\hat{\theta}), \ldots, \mathbb{P}_m(\hat{\theta}))$, and the denominator is equal to $\text{lik}(\hat{p}_1, \ldots, \hat{p}_m)$.
We intuitively have that (and can mathematically prove)

$$\hat{p}_i = \frac{x_i}{n},$$

which can be found in Rice §8.8.1 p.273, which uses Lagrance Multipliers.
So

$$\boxed{x_i = n\hat{p}_i}.$$

Recall that ML estimators are consistent. Then by consistency of the MLE,

$$\mathbb{P}_i(\hat{\theta}) \to \mathbb{P}_i(\theta),$$

whether or not $\mathbb{P}_i$ is a continuous function, or equivalently, $\hat{P}_i \to \mathbb{P}_i$ **if the null is true**.
Now let's rewrite Λ:

$$\begin{aligned} \Lambda &= \frac{\cancel{\frac{n!}{x_1! \cdots x_m!}} \mathbb{P}_1(\hat{\theta})^{x_1} \cdots \mathbb{P}_m(\hat{\theta})^m}{\cancel{\frac{n!}{x_1 \cdots x_m!}} \hat{p}_1^{x_1} \cdots \hat{p}_m^{x_m}} \\ &= \prod_{i=1}^m \left(\frac{\mathbb{P}_i(\hat{\theta})}{\hat{p}_i} \right)^{x_i}, \end{aligned}$$

where $x_i = n\hat{\mathbb{P}}_i$ from earlier. Recall that we're interested in $-2\log\Lambda$. This gives:

$$-2\log\Lambda = -2\sum_{i=1}^{m}(n\hat{p}_i)\log\left(\frac{n\mathbb{P}_i(\hat{\theta})}{n\hat{p}_i}\right),$$

where $O_i = n\hat{p}_i$ is our observed count and $E_i = n\mathbb{P}_i(\hat{\theta})$ is our expected count. Then we can rearrange to get:

$$\boxed{-2\log\Lambda = 2\sum_{i=1}^{m}O_i\log\left(\frac{O_i}{E_i}\right)},$$

which is something we can compute. This is not exactly the Pearson Chi Square statistic, but this is approximately that.

6 Taylor Series Argument

See p 342 of Rice for details.

$$2\sum_{i=1}^{m}O_i\log\left(\frac{O_i}{E_i}\right) \approx \sum_{i=1}^{m}\frac{(O_i-E_i)^2}{E_i},$$

where the RHS is the Pearson χ^2 test statistic. The dimension of our outcome space is $n-1$, because our outcome space is the set of all nonnegative numbers that add to 1 (this restriction of summing to 1 creates the 1 null space). That is,

$$\begin{aligned}\dim\Omega &= m-1\\ \omega_0 &= \{(\theta_1,\ldots,\theta_k),\end{aligned}$$

such that θ_i is in an open interval in $\mathbb{R}$. Then $\dim\omega_0 = k$, and this implies

$$-2\Lambda \approx \sum_{i=1}^{m}\frac{(O_i-E_i)^2}{E_i} \to \chi^2_{m-1-k}$$

Next time we'll look deeper into the assumptions that allowed for this theory, and we'll look into more examples in lab today.

Lecture ends here.