Hypothesis Testing 5

Day 15 (2/27/20)

4.7 Chi-square tests

Recall: If $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, then

1. $\frac{X_i - \mu}{\sigma} \sim N(0,1)$
2. $\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_1$ df
3. $\left(\frac{X_1-\mu}{\sigma}\right)^2 + \cdots + \left(\frac{X_n-\mu}{\sigma}\right)^2 \sim \chi^2_n$

since $W_1 + W_2 \sim \chi^2_{a+b}$ if $W_1 \sim \chi^2_a$ and $W_2 \sim \chi^2_b$, independent. The LHS of (3) is

$$\begin{aligned}
\sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 &= \sum_{i=1}^n \left(\frac{(X_i-\overline{X}) + (\overline{X}-\mu)}{\sigma}\right)^2 \\
&= \sum_{i=1}^n \left(\frac{(X_i-\overline{X})^2 + (\overline{X}-\mu)^2 + 2(X_i-\overline{X})(\overline{X}-\mu)}{\sigma^2}\right) \\
&= \sum_{i=1}^n \frac{(X_i-\overline{X})^2}{\sigma^2} + \frac{n(\overline{X}-\mu)^2}{\sigma^2} + 0
\end{aligned}$$

since $\sum(X_i - \overline{X}) = 0$. Now $\frac{n(\overline{X}-\mu)^2}{\sigma^2} = \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1$ and it follows that

$$\sum_{i=1}^n \frac{(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

There are two common ways the literature explains $n-1$ degrees of freedom.

- "Estimation of μ by $\overline{X}$ reduces df by 1"
- "Since $X_1 - \overline{X}, \ldots, X_n - \overline{X}$ sum to 0, there are only n-1 degrees of freedom"

Basically, the estimation of μ by $\overline{X}$ introduces a constraint.

Now, let $Z_1, Z_2, \ldots, Z_n$ be a random sample of Bernoulli(p) random variables,

$$Z_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Let $X_1 = \sum_{i=1}^n Z_i$ be the number of 1s (or successes) in the sample, and let $X_2 = n - X_1$ denote the number of 0s (or failures). Then both X_1 and X_2 are binomial

$$X_1 \sim \text{Bin}(n, p_1) \text{ and } X_2 \sim \text{Bin}(n, p_2)$$

where p_1 denotes the probability of success p and $p_2 = 1 - p$ is the probability of failure. From properties of the binomial random variable

$$E(X_1) = np_1, \qquad V(X_1) = np_1(1 - p_1)$$

$$E(X_2) = np_2, \qquad V(X_2) = np_2(1 - p_2)$$

By the CLT, a binomial random variable is approximately normal for large n so

$$\frac{X_1 - np_1}{\sqrt{np_1(1 - p_1)}} \text{ is approximately } N(0, 1)$$

$$Q = \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} \text{ is approximately } \chi_1^2$$

Now rewrite Q as follows

$$\begin{aligned} Q &= \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} [(1 - p_1) + p_1] = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{((n - X_2) - n(1 - p_2))^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \\ &= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \end{aligned}$$

where O_i and E_i denote observed and expected counts, respectively.

Theorem 4.2.1. *Let $(X_1, X_2, \dots, X_k)$ be a multinomial random variable with number of trials n and probability vector $(p_1, p_2, \dots, p_n)$. Then*

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \dot{\sim} \chi_{k-1}^2$$

Proof. Skip. □

Comments

- Often we write the χ^2 statistic as $Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ where $E_i = np_i$.
- This theorem provides a test for compatibility of observed frequencies with expected frequencies from a null model H_0. Reject H_0 if $Q > \chi^2_{.05,k-1}$.
- The χ^2 distribution is approximate. A usual rule of thumb is to require $E_i \geq 5, \ i = 1, \dots, k$.

4.3 χ^2 goodness-of-fit test

Example 4.7.1 Roll a six-sided die 60 times. Suppose that we observe the following outcome frequencies: (13, 19, 11, 8, 5, 4). Is the die fair, or does this provide evidence that the die is not balanced?

Solution. The null model says that $(p_1, \ldots, p_6) = (1/6, \ldots, 1/6)$, so we test

$$H_0 : p_1 = \cdots = p_6 = 1/6 \quad \text{vs} \quad H_1 : \text{At least one inequality}$$

Under H_0, the expected values are $E_i = 60(1/6) = 10$ for $i = 1, \ldots, 6$ and

$$\begin{aligned} Q_5 &= \frac{(13-10)^2}{10} + \frac{(19-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(4-10)^2}{10} \\ &= .9 + 8.1 + .1 + .4 + .25 + 3.6 = 15.6 \end{aligned}$$

The 95th percentile of χ^2_5 is 11.1, so we reject H_0. The p-value is $P[\chi^2_5 > 15.6] = .0081$ ■

Example (Mendel's pea experiments, *https://www.ncbi.nlm.nih.gov/books/NBK22098/*) Mendel took two parental pure lines (one was yellow, wrinkled seeds, the other had green, round seeds). The cross between these two lines produced seeds which were all were round and yellow. Next, Mendel selfed the plants, allowing the pollen of each flower to fall on its own stigma. This time, wrinkled and green seeds appeared. The frequencies are reported below. Mendel's hypothesis of dominant and recessive traits predicted the four cells have frequencies (9/16, 3/16, 3/16, 1/16). Do the data agree?

	Yellow	Green
Round	315	108
Wrinkled	101	32

Solution. The observed counts are $(315, 108, 101, 32)$ for a total of 556 seeds. Under $H_0 : (9/16, 3/16, 3/16, 1/16)$ the expected counts are

$$n(p_1, p_2, p_3, p_4) = 556(9/16, 3/16, 3/16, 1/16) = (312.75, 104.25, 104.25, 34.75)$$

Then

$$Q_3 = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(101-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} = .4699$$

The 95th percentile of χ^2_3 is 7.81. Therefore, we do not reject the null. The data does not contradict Mendel's model. ■

Example 4.7.2 Suppose that the unit interval is partitioned into 4 segments

$$A_1 = (0, 1/4\,], A_2 = (1/4, 1/2\,], A_3 = (1/2, 3/4\,], A_4 = (3/4, 1)$$

A random sample of $n = 80$ observations yields the following frequencies falling into each interval: $(6, 18, 20, 36)$. Conduct a goodness-of-fit test for

$$H_0 : f(x) = 2x,\ 0 < x < 1 \quad \text{vs} \quad H_1 : \text{Not}$$

Under H_0, the probability vector of falling into each interval is $(1/16, 3/16, 5/16, 7/16)$. The expected counts for 80 observations are $(5, 15, 25, 35)$. Then

$$Q_3 = \frac{(6-5)^2}{5} + \frac{(18-15)^2}{15} + \frac{(20-25)^2}{25} + \frac{(36-35)^2}{35} = \frac{64}{35} = 1.8286$$

The 95th percentile of χ^2_3 is 7.81, and p-value=.6087. Therefore, we do not reject the null.

4.4 Nuisance parameters

Let $Y_1, \dots, Y_n$ be a random sample from $N(\mu, \sigma^2)$. Partition the real line into disjoint intervals $A_1, \dots, A_k$. We want to test

$$H_0 : N(\mu, \sigma^2) \quad \text{vs} \quad H_1 : \text{Not}$$

If we let $X_1, \dots, X_k$ be the frequency of $A_1, \dots, A_k$, then $Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$ but the $\{p_i\}$ cannot be computed because we do not know μ and σ^2. There are two options:

1. Replace μ and σ^2 with values that would minimize Q_{k-1}
2. Replace μ and σ^2 with their maximum likelihood estimates

Comments

- The values μ and σ^2 in case (1) are called *minimum chi-square estimates*. The resulting statistic Q is now smaller than it would have been if we had used the true values of μ and σ^2. In fact, it can be shown that the null distribution of Q is closer to χ^2 with $k-3$ degrees of freedom instead of $k-1$ (one df is lost for every nuisance parameter estimated). In general

$$Q = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \dot{\sim} \chi^2_{k-1-c}$$

  where c is the number of parameters that were replaced with minimum chi-square estimates.

- The statistic Q in case (2) is easier to calculate than in case (1). However, Q is also larger, so keep in mind that the probability of rejection, and hence the size of the test, may be inflated.

```
> # Goodness-of-fit test in R:
> x<-c(13,19,11,8,5,4)
> p0<-rep(1/6,6)              #### H0 values
> expect<-sum(x)*p0           #### H0 expected frequencies
> expect
[1] 10 10 10 10 10 10
> sum((x-expect)^2/expect)
[1] 15.6
>
> qchisq(.95,5)               #### Critical value
[1] 11.0705
> 1-pchisq(15.6,df=5)         #### P-value
[1] 0.008083914
>
> chisq.test(x,p=p0)          #### Using chisq.test()

Chi-squared test for given probabilities

data:  x
X-squared = 15.6, df = 5, p-value = 0.008084
```