4/2/09 Math 152

9.41

$X_i \sim bin(n_i, p_i) \quad i = 1, \ldots, m$

H_0: $p_1 = p_2 = \ldots = p_m$

H_A: the p_i are not equal.

$$\Lambda = \frac{\prod_{i=1}^{m} \binom{n_i}{x_i} \hat{p}^{x_i} (1-\hat{p})^{n-x_i}}{\prod_{i=1}^{m} \binom{n_i}{x_i} \hat{p}_i^{x_i} (1-\hat{p}_i)^{n-x_i}}$$

where $\hat{p} = \frac{\sum_{i=1}^{m} x_i}{\sum_{i=1}^{m} n_i}$ $\qquad \hat{p}_i = \frac{x_i}{n_i}$

$$-2 \log \Lambda = -2 \sum_{i=1}^{m} x_i \log\left(\frac{\hat{p}}{\hat{p}_i}\right) + (n_i - x_i) \log\left(\frac{1-\hat{p}}{1-\hat{p}_i}\right)$$

$$= 2 \sum_{i=1}^{m} n_i \left[\hat{p}_i \log\left(\frac{\hat{p}_i}{\hat{p}}\right) + (1-\hat{p}_i) \log\left(\frac{1-\hat{p}_i}{1-\hat{p}}\right)\right]$$

$$\approx 2 \sum_{i=1}^{m} n_i \left[(p_i - \hat{p}) + \frac{1}{2} \frac{(\hat{p}_i - \hat{p})^2}{\hat{p}} + ((1-\hat{p}_i) - (1-\hat{p})) + \frac{1}{2}\left(\frac{((1-\hat{p}_i) - (1-\hat{p}))^2}{1-\hat{p}}\right)\right].$$

$$= \sum_{i=1}^{m} n_i (\hat{p}_i - \hat{p})^2 \left(\frac{1}{\hat{p}} + \frac{1}{1-\hat{p}}\right)$$

$$= \sum_{i=1}^{m} n_i \frac{(\hat{p}_i - \hat{p})^2}{\hat{p}(1-\hat{p})} = \sum_{i=1}^{m} \frac{(x_i - n_i \hat{p})^2}{n_i \hat{p}(1-\hat{p})}.$$

under the null hypothesis that all p_i are equal, this is asymptotically distributed like a χ^2_{m-1}.

This makes sense since

$$\frac{X_i - n_i \hat{p}}{n_i \hat{p}(1-\hat{p})}$$ is mean 0, variance 1

under the null.

Note that χ^2_1 has mean 1 and variance 2

so for large k

$$\chi^2_k \sim N(k, 2k)$$

9.42 c)

Breaks/Bar	Frequency
0	157
1	69
2	35
3	17
4	1
5	1

$$\hat{p} = \frac{157.0 + 69.1 + 35.2 + 3.17 + 4.1 + 5.1}{5(280)}$$

$$\simeq .142.$$

$$T = \sum_{i=1}^{280} \frac{(X_i - n_i \hat{p})^2}{n_i \hat{p}(1-\hat{p})}$$

$$= \frac{157\,(0-5\hat{p})^2}{5\hat{p}(1-\hat{p})} + \frac{69(1-5\hat{p})^2}{5\hat{p}(1-\hat{p})}$$

$$+ \frac{35\,(2-5\hat{p})^2}{5\hat{p}(1-\hat{p})} + \frac{17\,(3-5\hat{p})^2}{5\hat{p}(1-\hat{p})} + \frac{(4-5\hat{p})^2}{5\hat{p}(1-\hat{p})}$$

$$+ \frac{(5-5\hat{p})^2}{5\hat{p}(1-\hat{p})} \simeq 429$$

and T is $\simeq \chi^2_{279} \simeq N(279, 558)$.

$$\text{Prob}\{T \geq 429 \mid H_0\} = \text{Prob}\left\{\frac{T-279}{\sqrt{558}} \geq \frac{429-279}{\sqrt{558}}\right\}$$

$$\simeq 1 - \Phi\left(\frac{429-279}{\sqrt{558}}\right)$$

$$\simeq 1.077 \times 10^{-10}.$$

The value of p varies from bar to bar.

Qualitative Assessments of Goodness of Fit

Hanging Rootogram

A display of differences between observed and fitted values in a histogram

Plotting differences between $\sqrt{\text{observed}}$ and $\sqrt{\text{expected}}$ can be more useful in assessing magnitude of deviations from the fitted model since the square-root is (in many cases) variance-stabilizing

e.g. fitting a normal distribution

(μ, σ^2) estimated by $\bar{x}$, $\hat{\sigma}^2$

Intervals in our histogram are (x_{j-1}, x_j)

Then expected counts in (x_{j-1}, x_j) are given by:

$$\hat{n}_j = n \hat{p}_j$$

where

$$\hat{p}_j = \Phi\left(\frac{x_j - \bar{x}}{\hat{\sigma}}\right) - \Phi\left(\frac{x_{j-1} - \bar{x}}{\hat{\sigma}}\right).$$

We want to compare $\hat{n}_j$ with

n_j = observed counts in (x_{j-1}, x_j).

The problem is that

$\mathrm{Var}(n_j - \hat{n}_j)$ changes with j:

$$\mathrm{Var}(n_j - \hat{n}_j) \approx \mathrm{Var}(n_j) = n p_j (1 - p_j)$$

$$= n p_j - n p_j^2$$

$$\approx n p_j$$

when the p_j are small

(we have neglected $\mathrm{Var}(\hat{n}_j)$).

If the model is correct, we expect

larger fluctuations in the center and

smaller ones in the tails.

We can "re-scale" the differences so

that the variability remains constant with j.

e.g. Suppose X is a random variable with mean μ and variance $\sigma^2(\mu)$ which depends on μ.

Let $Y = f(X)$ and seek a nice choice of f.

$$Y \approx f(\mu) + f'(\mu)(X-\mu) + \dots .$$

$$E(Y) \approx f(\mu)$$

and $$\mathrm{Var}(Y) \approx E\left((Y-f(\mu))^2\right) \approx f'(\mu)^2 \sigma^2(\mu).$$

So we aim to choose f s.t.

$f'(\mu)^2 \sigma^2(\mu)$ is constant.

In our example

$$E(n_j) = n p_j = \mu.$$

$$\mathrm{Var}(n_j) \approx n p_j = \sigma^2(\mu) = \mu.$$

so we choose f so that

$$\mu\,(f'(\mu))^2 = \text{Constant} \overset{(\text{e.g.})}{=} 1/4$$

then $$f'(\mu) = \frac{1}{2}\frac{1}{\sqrt{\mu}}$$

and $$f(\mu) = \sqrt{\mu}$$

is a variance stabilizer.

We get

$$E(\sqrt{n_j}) \approx \sqrt{n p_j}$$

$$\text{Var}(\sqrt{n_j}) \approx 1/4.$$

if the model is correct.

See the Serum. Poliovirus example.

An alternative is to plot

$$\frac{n_j - \hat{n}_j}{\sqrt{\hat{n}_j}}$$

(components of the Chi-square statistic).

$\text{Var}(n_j - \hat{n}_j) \approx \text{Var}(n_j) \approx n p_j = \hat{n}_j$ as before, so $\text{Var}\left(\frac{n_j - \hat{n}_j}{\sqrt{\hat{n}_j}}\right) \approx 1$

Probability Plots

Let $X_1, \ldots X_n$ i.i.d. uniform $[0,1]$.

and let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$

be the ordered sample values.

(order statistics)

then $E(X_{(j)}) = \frac{j}{n+1}$. (see next page).

So plotting $\left(\frac{j}{n+1}, X_{(j)}\right)$

should give us ~~the a plot~~ nearly linear picture.

If X is a continuous R.V with a strictly increasing C.D.F F_X then

$Y = F_X(X)$ is uniform on $[0,1]$.

$$\text{Prob}\{x \le X_{(j)} \le x + \Delta x\}$$

$\approx$ Prob $\{$ $j-1$ observations are $< x$ $\}$

- Prob $\{$ 1 observation is in $(x, x+\Delta x)$ $\}$

- Prob $\{$ $n-j$ observations are $> x+\Delta x$ $\}$.

$$\approx \binom{n}{(j-1)\ 1\ (n-j)} x^{j-1} \cdot (1-x)^{n-j} \Delta x$$

so

$$f_j(x) = \frac{n!}{(j-1)!\,(n-j)!} x^{j-1} (1-x)^{n-j}$$

$0 \le x \le 1$.

and we want.

$$\int_0^1 x\, f_j(x)\, dx = \frac{n!}{(j-1)(n-j)!} \int_0^1 x^j (1-x)^{n-j}\, dx.$$

but $$\int_0^1 x^j (1-x)^{n-j} dx = \int_0^1 x^{(j+1)-1} (1-x)^{(n+1)-(j+1)} dx$$

$$= \frac{j!\,(n-j)!}{(n+1)!}$$

so $E(X_{(j)}) = \frac{j}{n+1}$. (since $\int f_{j+1}(x)dx = 1$ with $n+1$ samples)

and

$$\text{Prob}\{Y \le x\} = \text{Prob}\{X \le F_X^{-1}(x)\}$$

$$= F_X(F_X^{-1}(x)) = x.$$

So if we plot

$$F(X_{(k)}) \quad \text{vs.} \quad \frac{k}{n+1}$$

we should see a nearly linear picture (if F is the correct model).

Equivalently,

$$X_{(k)} \quad \text{vs.} \quad F^{-1}\left(\frac{k}{n+1}\right)$$

should be nearly linear.

(k^{th} order statistics vs. $\frac{k}{n+1}$ quantiles).

IF $F(x) = G\left(\frac{x-\mu}{\sigma}\right)$

(μ and σ are location and scale parameters)

we expect

$$G\left(\frac{X_{(k)}-\mu}{\sigma}\right) \quad \text{vs} \quad \frac{k}{n+1}$$

to be nearly linear

or

$$\frac{X_{(k)}-\mu}{\sigma} \quad \text{vs} \quad G^{-1}\left(\frac{k}{n+1}\right)$$

to be linear

or just

$$X_{(k)} \quad \text{vs} \quad G^{-1}\left(\frac{k}{n+1}\right)$$

to be linear if the model is correct.

(and the correlation will be the same).

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\,\sqrt{\text{Var}(Y)}}$$

replacing X by $aX+b$

gives

$$\text{Corr}(aX+b,Y) = \frac{a\,\text{Cov}(X,Y)}{\sqrt{a^2}\,\sqrt{\text{Var}(X)}\,\sqrt{\text{Var}(Y)}}$$

$$= \text{Corr}(X,Y) \quad \text{if } a>0.$$

recall ~~Note~~ also that

$$\text{corr}(X, aX+b) = \frac{\text{cov}(X, aX+b)}{\sqrt{\text{Var}\,X}\,\sqrt{a^2\text{Var}\,X}}$$

$$= \frac{a\ \text{Var}(X)}{\sqrt{a^2}\ \text{Var}\,X}$$

$$= \frac{a}{\sqrt{a^2}} = \pm 1.$$