ORIGIN ≡ 1

G-test for Goodness of Fit

Log Liklihood ratios supply an alternative method for assessing goodness of fit based on a maximum liklihood approach. This test may be used for the same kinds of data as the $\chi 2$ test for Goodness of fit, sometimes with a different result. According to Zar 2010 p. 480, opinions differ as to which is the better test. This may be mostly a matter of religious adherence to a preferred methodology It's useful to run both and compare results.

Data Structure:

Observations are counts of individuals in k classes.

Goodness of Fit Table					
O_1	O_2	O_3	…	O_k	Total
E_1	E_2	E_3	…	E_k	

Assumptions:

- Observed values O_j are a random sample in k cells

Zar Example 22.1

seed classes				
yellow/smooth	*yellow/wrinkled*	*green smooth*	*green/wrinkled*	n
152	39	53	6	250

Model:

Let Expected Probabilities:
- E_j be specified as:
- internally specified model with g parameters estimated from the sample.
OR
- externally specified model

$k := 4$

^ number of classes

$g := 0$

^ number of internally specified parameters

Hypotheses:

H_0: P_j Probabilities are distributed according to the model

H_1: P_j Probabilities differ from the model < **Two sided test**

Construct Contingency Tables of Observed and Expected in each cell:

- Tabulate O_j for each cell
- Calculate Observed Row and Column Totals
- Calculate Expected for each cell:

$$O := \begin{pmatrix} 152 \\ 39 \\ 53 \\ 6 \end{pmatrix} \qquad E := \begin{bmatrix} \left(\frac{9}{16}\right)\cdot 250 \\ \left(\frac{3}{16}\right)\cdot 250 \\ \left(\frac{3}{16}\right)\cdot 250 \\ \left(\frac{1}{16}\right)\cdot 250 \end{bmatrix} \qquad E = \begin{pmatrix} 140.625 \\ 46.875 \\ 46.875 \\ 15.625 \end{pmatrix}$$

seed classes				
yellow/smooth	*yellow/wrinkled*	*green smooth*	*green/wrinkled*	n
152	39	53	6	250
140.625	46.875	46.875	15.625	

^ where: E_j are the expected probabilities of each cell based on theory in genetics. g = 0 in this case.

χ^2 Test Statistic:

$j := 1..k$

$$\chi_{sq} := \sum_j \frac{(O_j - E_j)^2}{E_j} \qquad \chi_{sq} = 8.9724$$

G Test Statistic:

$$G := 2 \cdot \sum_j O_j \cdot \ln\left(\frac{O_j}{E_j}\right) \qquad G = 10.83251 \qquad 2 \cdot \left(\sum_j O_j \cdot \ln(O_j) - \sum_j O_j \cdot \ln(E_j)\right) = 10.8325$$

^ alternate formula for G

Sampling Distribution:

If Assumptions hold and H_0 is true, then G ~$\chi_{(k\text{-}g\text{-}1)}$

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$df := k - g - 1$ $df = 3$ **< where: k = the number of cells,**
g = number of parameters of the *internally* specified model

$C := qchisq(1 - \alpha, df)$ $C = 7.814728$

Decision Rule:

If χ_{sq} > C THEN REJECT H_0,OTHERWISE ACCEPT H_0

IF G > C THEN REJECT H_0,OTHERWISE ACCEPT H_0

$\chi_{sq} = 8.9724$

$G = 10.8325$

Probability Value:

$P_\chi := (1 - pchisq(\chi_{sq}, df))$ $P_\chi = 0.02966$ **< for χ^2 test**

$P_G := (1 - pchisq(G, df))$ $P_G = 0.01267$ **< for G test**

Prototype in R:

```
#CHI-SQUARE TEST FOR GOODNESS OF FIT
#ZAR EXAMPLE 22.8
ZAR=read.table("c:/DATA/Biostatistics/ZarEX22.8R.txt")
ZAR
attach(ZAR)

chisq.test(observed,p=expected,rescale.p=TRUE)
```

```
> chisq.test(observed,p=expected,rescale.p=TRUE)

        Chi-squared test for given probabilities

data:  observed
X-squared = 8.9724, df = 3, p-value = 0.02966
```

```
#G-TEST FOR GOODNESS OF FIT
#APPROPRIATE TEST FUNCTION NOT YET FOUND IN R
#THEREFORE, I HAD TO DO THIS FROM SCRATCH:

#G STATISTIC:
G=2*sum(observed*log(observed/expected))
G

#CRITICAL VALUE:
alpha=0.05
k=4
g=0
df=k-g-1
C=qchisq(1-alpha,df)
C

#PROBABILITY:
P=(1-pchisq(G,df))
P
```

```
>#G STATISTIC:
> G
[1] 10.83251
>
> #CRITICAL VALUE:
> C
[1] 7.814728

> #PROBABILITY:
> P
[1] 0.01266689
```

Yates Correction when degrees of freedom = 1:

Correction is applied in a way that's analogous to what's done in the χ^2 case. Both are shown here for comparison.

$$O := \begin{pmatrix} 84 \\ 16 \end{pmatrix} \qquad E := \begin{bmatrix} \left(\frac{3}{4}\right) \cdot 100 \\ \left(\frac{1}{4}\right) \cdot 100 \end{bmatrix} \qquad E = \begin{pmatrix} 75 \\ 25 \end{pmatrix} \qquad k := 2 \qquad df := 1$$

$j := 1..k$

Zar Example 22.1:

Seed Color		
Yellow	*Green*	n
84	16	100
75	25	

χ^2 Test Statistic:

$j := 1..k$

$$\chi_{sq} := \sum_j \frac{(O_j - E_j)^2}{E_j} \qquad \chi_{sq} = 4.32$$

G Test Statistic:

$$G := 2 \cdot \sum_j O_j \cdot \ln\left(\frac{O_j}{E_j}\right) \qquad G = 4.758032$$

Yates Correction for the χ^2 Test Statistic:

In cases where number of classes = 2 (df =1) Yates correction is routinely employed to allow test statistic χ^2 to be destributed as $\chi^2_{(df=1)}$

$$\chi_{Csq} := \sum_j \frac{(|O_j - E_j| - 0.5)^2}{E_j} \qquad \chi_{Csq} = 3.8533$$

Yates Correction for the G Test Statistic:

$$O := \begin{pmatrix} 84 - 0.5 \\ 16 + 0.5 \end{pmatrix}$$ **< correction applied to observed values**

$$G_C := 2 \cdot \sum_j O_j \cdot \ln\left(\frac{O_j}{E_j}\right) \qquad G_C = 4.216863$$

Probability Value for the $\chi_C{}^2$ Test Statistic:

$$P_{\chi C} := \left(1 - \text{pchisq}\left(\chi_{Csq}, \text{df}\right)\right) \qquad P_{\chi C} = 0.049647$$

Probability Value for the G_C Test Statistic:

$$P_C := \left(1 - \text{pchisq}\left(G_C, \text{df}\right)\right) \qquad P_C = 0.04002409$$

Prototype in R:

```
#YATES CORRECTION:
ZAR2=read.table("c:/2010BiostatsData/ZarEX22.1R.txt")
ZAR2
attach(ZAR2)
df=1

#YATES CORRECTED G STATISTIC:
obs=c(84-0.5,16+0.5)
Gc=2*sum(obs*log(obs/expected))
Gc

#YATES CORRECTED PROBABILITY:
Pc=(1-pchisq(Gc,df))
Pc
```

```
> ZAR2
  observed expected  class
1       84       75 yellow
2       16       25  green

> #YATES CORRECTED G STATISTIC:
> Gc
[1] 4.216863
>
> #YATES CORRECTED PROBABILITY:
> Pc
[1] 0.04002409
```