ORIGIN ≡ 1

χ^2 Test for Goodness of Fit

Goodness of Fit is the term applied to tests of nominal-scale data in which counts of observations are placed in separate blocks. These counts are then compared with an expectation for each block. The expectations may be derived either from an *intrinsic model* based on the data (such as using row and column totals) or from an *extrinsic model* based ideas completely unrelated to the data (such as the famous 9:3:3:1 ratio in genetics). The choice between intrinsic versus extrinsic model is reflected in a difference in degrees of freedom for the test.

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

Model:

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

$k := 2$

^ number of classes

Seed Color		
Yellow	*Green*	n
84	16	100

$g := 0$

^ 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**

Criterion for Normal Approximation:

- IF no more than 1/5 of the cells have expected values in each cell $E_j \le 5$ **AND no cell has expected value** $E_j < 1$ **THEN Approximation may be used**

Construct Contingency Tables of Observed and Expected in each cell:

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

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

$$O := \begin{pmatrix} 84 \\ 16 \end{pmatrix}$$

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

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

$g := 0$

χ^2 Test Statistic:

$$j := 1 .. k \qquad \chi_{sq} := \sum_j \frac{(O_j - E_j)^2}{E_j} \qquad \chi_{sq} = 4.32$$

Sampling Distribution:

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

Critical Value of the Test:

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

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

$C := \text{qchisq}(1 - \alpha, df)$ $C = 3.8415$

$\chi_{sq} = 4.32$

Decision Rule:

IF $\chi_{sq} > C$ THEN REJECT H_0,OTHERWISE ACCEPT H_0

Probability Value:

$P := (1 - \text{pchisq}(\chi_{sq}, df))$ $P = 0.03767$

Prototype in R:

```
#CHI-SQUARE TEST FOR GOODNESS OF FIT
#ZAR EXAMPLE 22.1
ZAR=read.table("c:/DATA/Biostatistics/ZarEX22.1R.txt")
ZAR
attach(ZAR)

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

Zar Example 22.1

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

        Chi-squared test for given probabilities

data:  observed
X-squared = 4.32, df = 1, p-value = 0.03767
```

^ Note the way to tell R that numbers are probabilities.
Note also the switch "rescale.p=TRUE" telling R to convert expected values into probabilities that sum to 1.

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_{sq} to be destributed as $\chi^2_{(df=1)}$

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

Critical Value & Decision Rule stays the same, but probability is modified.

Probability Value:

$P := (1 - \text{pchisq}(\chi_{Csq}, df))$ $P = 0.04964723$

Prototype in R:

```
#WITH YATES CORRECTION:

chisq.test(observed,p=expected,rescale.p=TRUE, correct=TRUE)
#NOTE THAT THIS APPARENTLY DOESN'T WORK IN R FOR NON 2x2 TABLES
#SO CALCULATING BY HAND:

O=observed
E=expected
df=1
YatesCHISQ=sum(((abs(O-E)-0.5)^2)/E)
YatesCHISQ

YatesProb=1-pchisq(YatesCHISQ,df)
YatesProb
```

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

        Chi-squared test for given probabilities

data:  observed
X-squared = 4.32, df = 1, p-value = 0.03767
```

^ this did not change when we might have expected that it should.

```
> YatesCHISQ
[1] 3.853333
>
> YatesProb
[1] 0.04964723
```

< these were calculated by hand in the script.