Analysis of frequencies

We have primarily been concerned with quantitative variation up to now. Today we are going to talk about data that comes in the form of counts or frequencies. You have had some experience with these types of data, for example, we compared observed to expected counts in a Poisson distribution, though we did so only qualitatively.

Goodness of fit

Goodness of fit refers to how well a particular distribution fits another.
Typically, one of the distributions is a hypothesized distribution or more accurately, a distribution *expected* under a particular hypothesis. The other is observed data.

- When the expected distribution is simple and well-specified, the most accurate GOF test is to calculate the exact probability of observing the particular distribution
- When both distributions are empirical or the expected distribution is not well characterized mathematically, a test statistic can be calculated and compared to a (third) sampling distribution

Estimating goodness of fit: binomial

We've done something like this.
Imagine a monohybrid cross 'Aa' x 'Aa' of a locus that exhibits complete dominance. The expected distribution under this cross is 3:1 dominant:recessive phenotypes.
Now imagine that you saw 39 dominant and 10 recessive.
It's easy to calculate the probability of observing exactly this ratio:

```
> dbinom(39, 49, 0.75)

[1] 0.1050863
```

Sum probabilities

However the question is actually a one tailed test: what is the probability of this ratio or a more extreme ratio?

```
> svec <- c(dbinom(39, 49, 0.75), dbinom(40, 49, 0.75), dbinom(41,
+     49, 0.75), dbinom(42, 49, 0.75))
> svec <- c(svec, dbinom(43, 49, 0.75), dbinom(44, 49, 0.75), dbinom(45,
+     49, 0.75), dbinom(46, 49, 0.75))
> svec <- c(svec, dbinom(47, 49, 0.75), dbinom(48, 49, 0.75), dbinom(49,
+     49, 0.75))
> print(sum(svec))
```

```
[1] 0.2884739

> svec <- NULL
> for (i in 49:39) {
+     svec <- c(svec, dbinom(i, 49, 0.75))
+ }
> sum(svec)

[1] 0.2884739
```

R function

```
> binom.test(39, 49, 0.75, alt = "greater")

        Exact binomial test

data:  39 and 49
number of successes = 39, number of trials = 49, p-value = 0.2885
alternative hypothesis: true probability of success is greater than 0.75
95 percent confidence interval:
 0.6784577 1.0000000
sample estimates:
probability of success
             0.7959184
```

Test statistic approach

Test statistics usually depend upon expected *frequencies*
For the previous example, the expected number of dominant phenotypes is:

$$0.75 \times (39 + 10) = 36.75$$

The expected number of recessive phenotypes is $0.25 \times 49 = 12.25$ or:

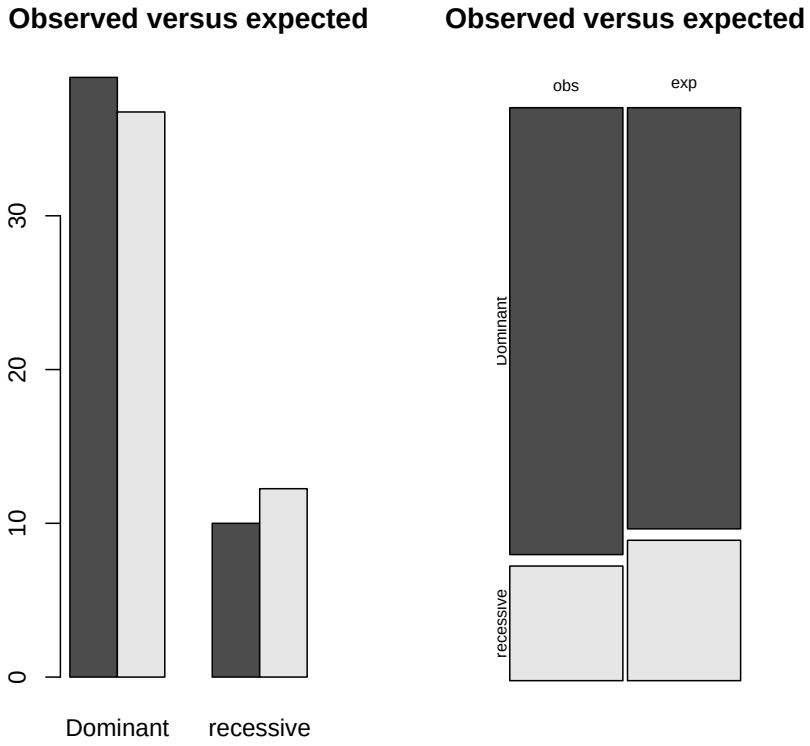
$$49 - 36.75 = 12.25$$

In general

$$E_i = p_i \times \Sigma_{i=1}^{k} O_i$$

Where k is the number of classes.

Graphical represenation



Graphical represenation: code

```
> tbl <- as.table(matrix(c(39, 36.75, 10, 12.25), 2, 2))
> colnames(tbl) <- c("Dominant", "recessive")
> rownames(tbl) <- c("obs", "exp")
> par(mfrow = c(1, 2))
> barplot(tbl, beside = T, main = "Observed versus expected")
> plot(tbl, color = T, main = "Observed versus expected")
```

Calculating statistic: X^2

The X^2 test statistic compares the observed to expected numbers like this:

$$X^2 = \Sigma_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

This test statistic is distributed as a χ^2 with $k - 1$ df.

```
> tbl
```

```
    Dominant recessive
obs    39.00     10.00
exp    36.75     12.25

> Chisq <- sum((tbl[1, ] - tbl[2, ])^2/tbl[2, ])
> Chisq

[1] 0.5510204
```

Statistical inference

As you might expect the χ^2 distribution in R is given by the `pchisq, dchisq`, etc.

```
> Chisq

[1] 0.5510204

> pchisq(Chisq, 1)

[1] 0.5420989
```

Watch out! Cell sizes that are small artificially inflate the type I error.

Better approach: G-test

The G-test is based upon the ratio of exact probabilities. It is less sensitive (not immune) to small cell sizes.
The binomial equation is:

$$\left(\begin{array}{c} n \\ f_1 \end{array} \right) p^{f_1}(1-p)^{f_2}$$

Probabilities

Based upon the binomial equation the probability of observing the data are:

$$\left(\begin{array}{c} 49 \\ 39 \end{array} \right) (\frac{39}{49})^{39}(\frac{10}{49})^{10} = 0.14$$

The probability of observing the data under the null is:

$$\left(\begin{array}{c} 49 \\ 39 \end{array} \right) 0.75^{39}0.25^{10} = 0.11$$

Likelihood ratio

If the probabilities are the same, the ratio is 1.0. If not the ratio for the observed proportions is greater. The ratio is often denoted as L.

$$L = p[obs]/p[exp] = 0.14/0.11 = 1.27$$

Take twice the natural log of this ratio. This is the G statistic.

$$G = 2lnL = 0.48$$

G is χ^2 distributed with $df = k - 1$

```
> pchisq(0.48, 1)

[1] 0.5115777
```

General formula for G

Because L is a ratio, the $\begin{pmatrix} n \\ f_1 \end{pmatrix}$ values are the same for both numerator and denominator and cancel. Ratios of fractions (the probabilities) result in the final form for L:

$$L = \left(\frac{f_1}{\hat{f}_1}\right)^{f_1} \left(\frac{f_2}{\hat{f}_2}\right)^{f_2}$$

Which results in

$$G = 2 \times \Sigma_{i=1}^{k} f_i ln \left(\frac{f_i}{\hat{f}_i}\right)$$

Carex data

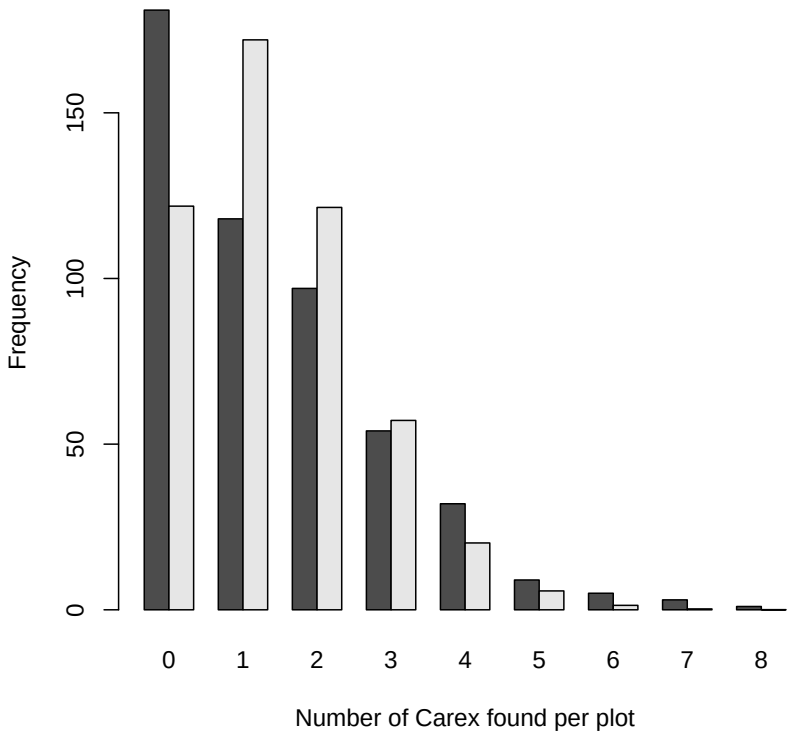
```
> number.per.quadrat <- 0:8
> obs.freq <- c(181, 118, 97, 54, 32, 9, 5, 3, 1)
> carex <- data.frame(number.per.quadrat, obs.freq)
> carex

  number.per.quadrat obs.freq
1                  0      181
2                  1      118
3                  2       97
4                  3       54
5                  4       32
6                  5        9
7                  6        5
8                  7        3
9                  8        1
```

Expected Distribution

We could test whether the *Carex* data exhibits a Poisson distribution. To generate the expected values use the approach from early in the semester.

```
> xbar <- sum(number.per.quadrat * obs.freq)/sum(obs.freq)
> xbar

[1] 1.412

> exp.freq <- sum(obs.freq) * dpois(0:8, xbar)
> exp.freq

[1] 121.82774227 172.02077209 121.44666510  57.16089704  20.17779665
[6]   5.69820978   1.34097870   0.27049456   0.04774229

> carex <- cbind(carex, exp.freq)
```

Carex distribution



The test

```
> G = 2 * sum(obs.freq * log(obs.freq/exp.freq))
> G

[1] 76.03028

> 1 - pchisq(G, length(obs.freq) - 1)

[1] 3.066436e-13
```

$r \times c$ tables: independence

It is also possible to use Goodness of fit tests to evaluate data of higher dimensions. Here we will look at two variables.
The typical question when looking at the joint frequency distribution of two variables is to ask if the frequencies vary independently or if they are associated.
The null hypothesis is: Counts of Var1 are independent of Counts of Var2.

- Two approaches exist: exact calculation of probability under independence
- Calculation of a test statistic (generally X^2 or G)

Example data: dioecy

Dioecy is uncommon in plants. It has also appears to be associated with fleshy fruits (Givnish; 1980).

Counts from the Carolinas

Here is a count of dioecious species and fleshy species in the Carolinas (from Radford et al. (1968)) presented in a 2×2 table (also called a contingency table).

	Fruit Type		
Sexuality	Fleshy	Dry	Total
Dioecious Obs.	65	82	147
Non-Dioe. Obs.	257	2874	3131
Total	322	2956	3278

Fisher's exact test

Fishers exact test is analogous to the binomial test from earlier. It is based on calculating the exact probability of obtaining the observed data. This is based upon a multinomial distribution. The final probability is:

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

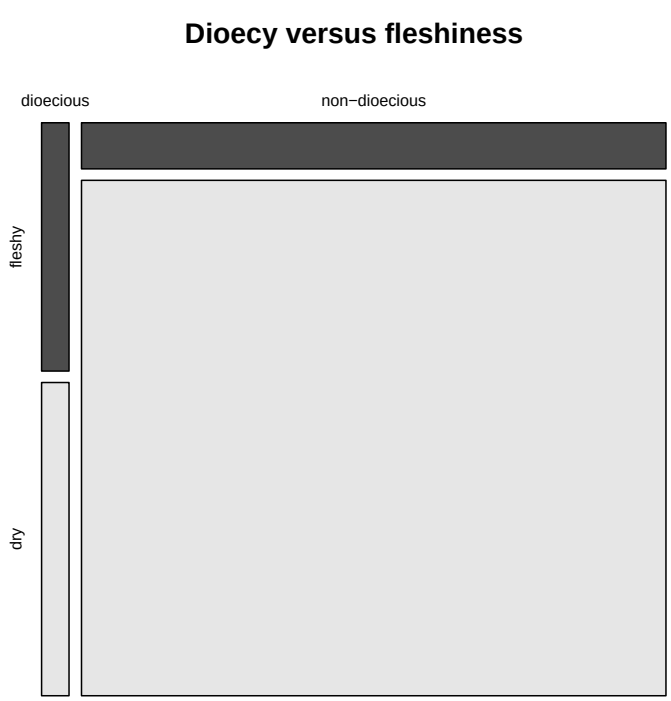
Where a, b, c, d, n are given below:

	a	b	$a+b$
	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=n$

R-version

```
> dioecy <- as.table(cbind(c(65, 257), c(82, 2874)))
> colnames(dioecy) <- c("fleshy", "dry")
> rownames(dioecy) <- c("dioecious", "non-dioecious")
> dioecy

              fleshy  dry
dioecious         65   82
non-dioecious    257 2874
```

Graphic



R-version-2

```
> fisher.test(dioecy)
```

```
        Fisher's Exact Test for Count Data

data:  dioecy
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  6.13509 12.74071
sample estimates:
odds ratio
  8.852807
```

Expected values

The expected values can be thought of as the total of each row proportionally divided into columns based upon relative column totals.
As a result, the expected values in the Dioecy row are:

$$147 \times \frac{322}{3278} = 14.4; 147 \times \frac{2956}{3278} = 132.6$$

Example data: expected values

	Fruit Type		
Sexuality	Fleshy	Dry	Total
Dioecious Obs.	65	82	147
Exp.	14.4	132.6	
Non-Dioe. Obs.	257	2874	3131
Exp.	307.5	2823.5	
Total	322	2956	3278

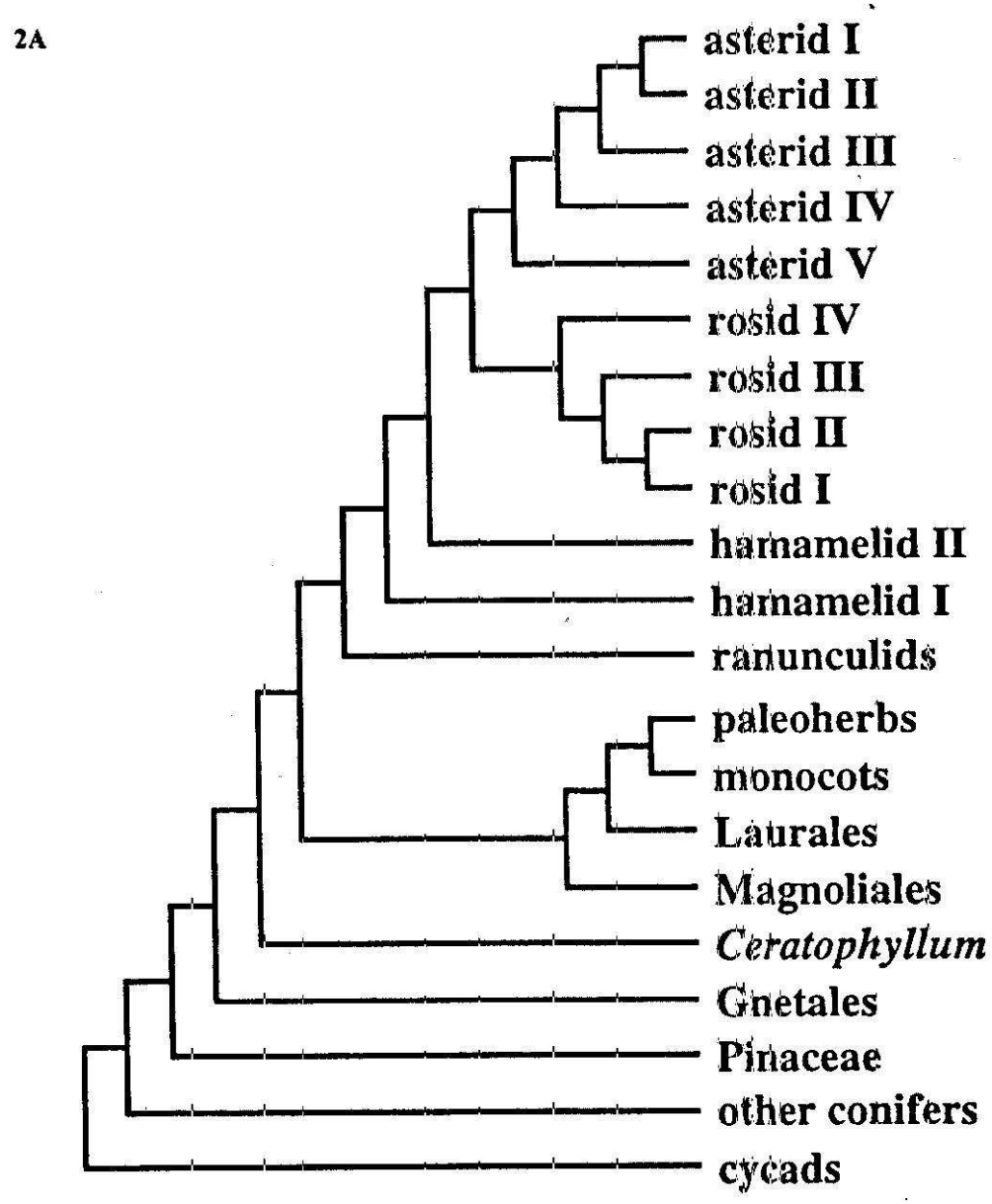
G-test

```
> exp.dioecy <- t(outer(colSums(dioecy), (rowSums(dioecy))/sum(dioecy)))
> exp.dioecy

                fleshy       dry
dioecious      14.43990  132.5601
non-dioecious 307.56010 2823.4399

> G = 2 * sum(dioecy * log(dioecy/exp.dioecy))
> G

[1] 126.5077
```

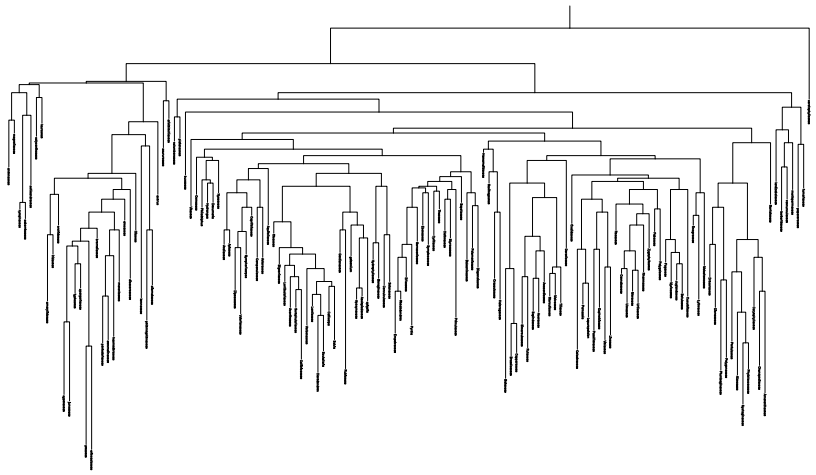
```
> 1 - pchisq(G, 1)

[1] 0
```

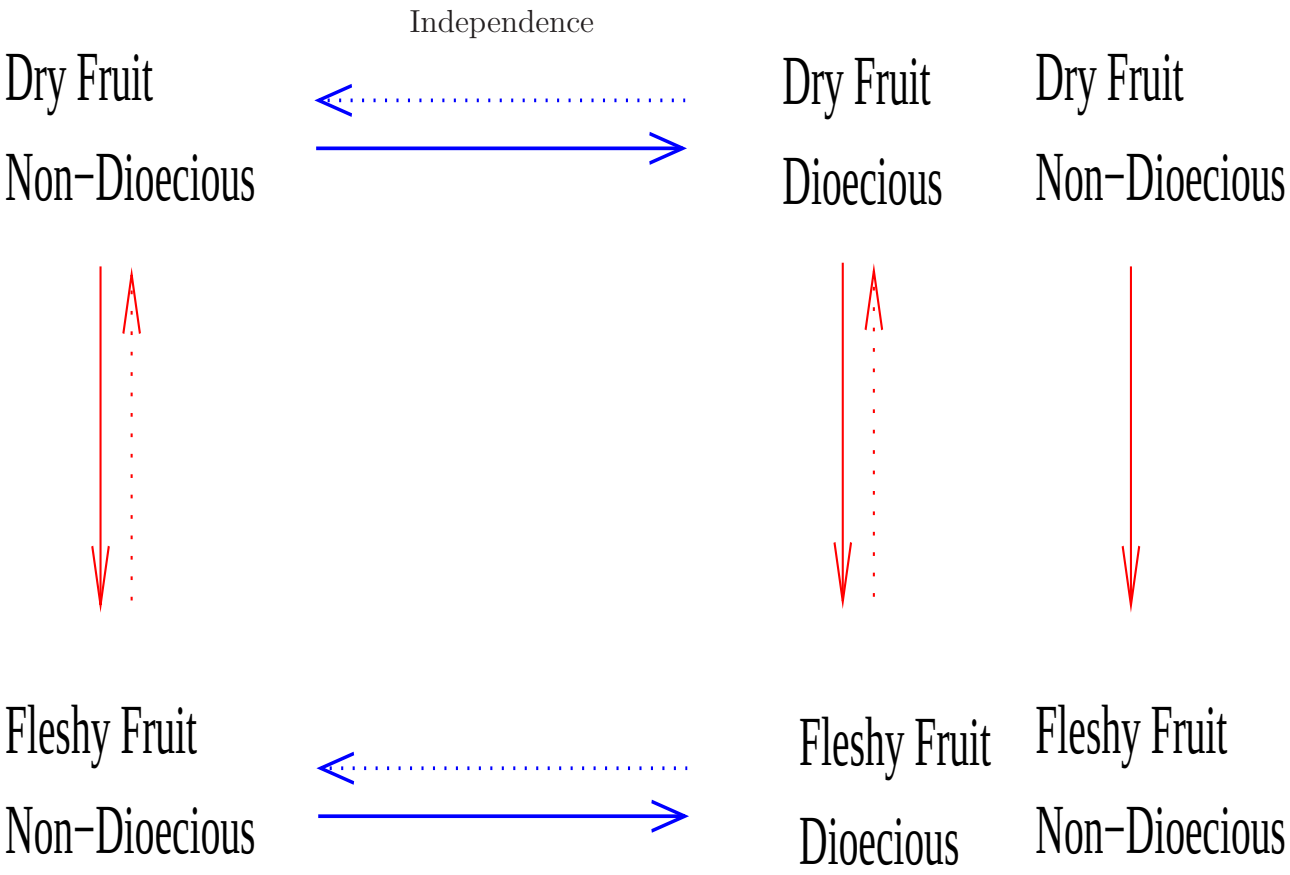
Assumption: uncorrelated samples

This analysis assumes that the only association is due to the effects of fruit-type, but:



Family phylogeny





Contrasting models using phylogeny

Model	Parameters	ln(L)	L/L_G	d.f.
Resource Allocation	3	-467.6	2.7 x 10^{41}	5
Independence	4	-376.8	99.5	4
General	8	-372.2		

References

Givnish, T. J. (1980). Ecological constraints on the evolution of breeding systems in seed plants: dioecy and dispersal in gymnosperms, *Evolution* **34**: 959–972.

Radford, A., Ahles, H. and Bell, C. (1968). *Manual of the vascular flora of the Carolinas*, University of North Carolina Press, Chapel Hill.