

Frequency Data

- Chi-square distribution
- Inferences about proportions
- Multinomial chi-square tests
- Goodness-of-fit tests
- Contingency tables
 - 2×2 tables
 - $R \times C$ tables
- Log-linear analysis



Chi-square Statistics

Emphasis to this point has been on data from either continuous measurement variables or ranked variables.

Recall, attribute (or categorical) variables cannot be measured but must be expressed qualitatively (e.g., dead/alive, black/white, male/female).

These are enumeration (or count) data and usually rely on some form of chi-square analysis.

Chi-square Statistics

These statistics are used in several cases:

Goodness-of-fit tests; i.e., does sample data fit a probability model?

Homogeneity tests; i.e., do two samples come from the same [unknown] distribution?

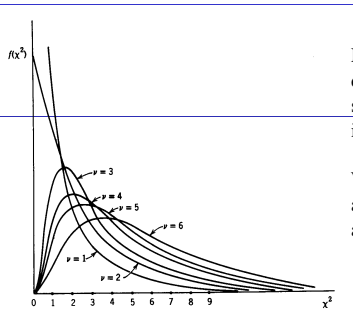
Tests of Association or Independence; i.e., is there a relationship between two or more variables?

Chi-square Distribution

The chi-square distribution is a theoretical probability distribution (analogous to normal, binomial, poisson, etc.).

However, it is usually not a direct model of population distribution (like normal), but it can be used to compare samples or variables.

Chi-square Distribution



Note that the χ^2 distribution is not symmetrical and is highly skewed.

When $df = 1$ then asymptotic to both axes!

Chi-square Distribution

If χ^2 is a random variable with a chi-square distribution:

χ^2 is a positive real number

The density function depends only on v (df)

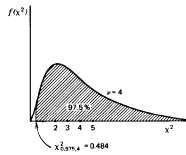
The expected value of $\chi^2 = v$

The variance of $\chi^2 = 2v$

The graph of $f(\chi^2)$ is not symmetrical

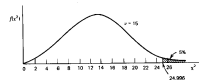
The graph of $f(\chi^2)$ approaches symmetry as $v = \infty$

Chi-square Distribution



Examples of values from the chi-square table (B.1, App12-16):

$$\chi^2_{0.975,4} = 0.4844$$

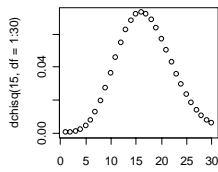
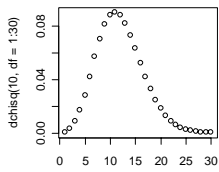
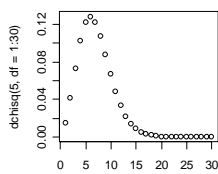
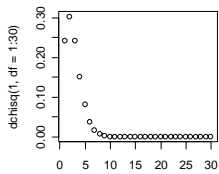


$$\chi^2_{0.05,15} = 24.9958$$

Chi-square Distribution

We can explore the properties of the C-S distribution through the use of R functions and graphics:

```
> par(mfrow=c(2,2),mar=c(3,4,3,3))
> layout.show(4)
> plot(dchisq(1,df=1:30))
> plot(dchisq(5,df=1:30))
> plot(dchisq(10,df=1:30))
> plot(dchisq(15,df=1:30))
```



Inferences About a Proportion

- Large Sample Size -

Just like the one sample z -test ($n \geq 30$) and t -test ($n < 30$) for continuous data and the sign test for ranked data, there exists a test for a single population described by a categorical variable.

As with most categorical variables, this can be viewed as a proportion (e.g., 30 of 300 flowers were pink = 0.10).

Inferences About a Proportion

- Large Sample Size -

$$Z = \frac{\hat{\pi} - \pi_0}{\hat{\sigma}_{\hat{\pi}}}$$

where SE is est. as :

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

and $100(1-\alpha)CI$ as :

$$\hat{\pi} \pm z(\alpha/2)\hat{\sigma}_{\hat{\pi}}$$

Pi-hat is the observed proportion of a response from a random sample of n (which is presumed to be large; i.e., ≥ 30)

Pi is the corresponding population proportion.

$$H_0: \pi = \pi_0$$

Inferences About a Proportion

-Example: Large Sample Size -

To assess the effectiveness of a new treatment for aphid infestation of tobacco plants, 100 treated plants were classified as 67 no infestation, 24 stem infestation only, 9 leaf infestation only.

To test the assertion that 15% of the population should have leaf only infestation, we test the $H_0: \pi \geq 0.15$ against $H_a: \pi < 0.15$

Inferences About a Proportion

-Example: Large Sample Size -

$$\hat{\pi} = 9/100 = 0.09$$

$n = 100$ (large)

$$z_c = \frac{0.09 - 0.15}{\sqrt{\frac{0.09(1-0.09)}{100}}} = -2.10$$

95% CI around proportion:

$$0.09 \pm 1.96 \sqrt{\frac{0.09(1-0.09)}{100}} = 0.09 \pm (1.96)(0.0286,$$

thus, the CI is: $0.034 \leq \pi \leq 0.146$

15% does NOT fall within the 95% CI of the expected proportion for leaves with infestation.



```
> prop.test(9,100,.15)
```

1-sample proportions test with continuity correction

```
data: 9 out of 100, null probability 0.15
X-squared = 2.3725, df = 1, p-value = 0.1235
alternative hypothesis: true p is not equal to 0.15
95 percent confidence interval:
 0.04455927 0.16833464
sample estimates:
 p
0.09
```

Inferences About a Proportion

- Small Sample Size -

When the sample size is small (i.e., $n < 30$), we can make inferences about π on the basis that the number of 1s in a random sample of size n from a 0-1 population can be regarded as the number of successes in a binomial experiment with n trials and probability of success π .

This procedure is referred to as a Binomial Test (see Zar Example 24.8, p.534).

Inferences About Two Proportions

- Large Sample Size -

$$Z = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}}$$

where SE is est. as :

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

and 100(1- α)CI as :

$$\hat{\theta} \pm z(\alpha/2)\hat{\sigma}_{\hat{\theta}}$$

Just like the one-sample *t*-test, there is a simple extension of the *z*-test to examine the difference between TWO proportions from independent samples (where both *n*₁ and *n*₂ are large; i.e., *n* ≥ 30).

Inferences About Two Proportions

- Small Sample Size -

For small independent samples, H₀: $\pi_1 = \pi_2$ can be tested using Fisher's Exact Test.

Whereas the Binomial Test for testing a hypothesis about a single proportion from a small sample uses the binomial distribution, Fisher's Exact Test uses the hypergeometric distribution.

Details of this procedure can be found in Zar Section 24.11.

Inferences About Paired Proportions

- Large Sample Size -

The McNemar Test is very commonly used to compare proportions based upon paired categorical data (see Zar sect. 9.4).

The test statistic has the same structure as the *z*-test for independent samples. The determination of *Z* is identical, but the estimated standard error differs.

```

> ## Zar Example 9.4

> Relief <- matrix(c(11,6,10,24),nrow=2,
dimnames=list("Lotion 1" = c("Relief", "No Relief"), "Lotion
2" = c("Relief", "No Relief")))

> Relief
      Lotion 2
Lotion 1  Relief No Relief
Relief      11      10
No Relief   6       24

> mcnemar.test(Relief)

      McNemar's Chi-squared test with continuity correction

data: Relief
McNemar's chi-squared = 0.5625, df = 1, p-value = 0.4533
    
```



Multinomial Chi-square

The multinomial test is useful in many areas of biology, particularly Mendelian genetics.

This is a specialized extension of the binomial distribution.

Permits one to test an hypothesis relating observed frequencies to expected frequencies.

Multinomial Chi-square

- Example -

A geneticist randomly samples a new variety of peas.

She scores them based on pea color (green or yellow) and quality (smooth or wrinkled).

Because these traits should follow a dihybrid inbreeding model, the prediction is a 9:3:3:1 ratio, or:

$$H_o: \frac{9}{16} \pi_1 = \frac{3}{16} \pi_2 = \frac{3}{16} \pi_3 = \frac{1}{16} \pi_4$$

$$H_a: \frac{9}{16} \pi_1 \neq \frac{3}{16} \pi_2 \neq \frac{3}{16} \pi_3 \neq \frac{1}{16} \pi_4$$

Multinomial Chi-square

- Example -

	9 SY	3 WY	3 SG	1 WG
O	315	101	108	32
E	313	104	104	35
O-E	2	3	4	3
(O-E) ²	4	9	16	9
(O-E) ² /E	0.013	0.087	0.154	0.257

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 0.013 + 0.087 + 0.154 + 0.257 = 0.511$$

$\chi^2_{0.05,3} = 7.815$ (i.e., splits distribution into 95% & 5%)
 $\chi^2_c < \chi^2_{\alpha}$, therefore, fail to reject H₀ of 9:3:3:1 ratio.

Multinomial Chi-square

- Caveats -

As described, these are 2-tailed tests. A 1-tailed test becomes much more involved.

Chi-square test of df = 1 can be done, but their strength is very weak (recall distribution).

Chi-square accuracy destabilizes when E values fall below 5 (you may need to combine adjacent cells).

When $n < 25$, you should add Yates' Continuity

Correction:
$$\chi^2_c = \sum_{i=1}^k \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Goodness-of-fit Tests

The multinomial chi-square test that we just did is actually a subset of a larger group of statistics known as goodness-of-fit tests.

G-O-F tests use the same approach and test stats as the multinomial χ^2 but can be used to address hypotheses such as:

- H₀: This sample is from a poisson distribution
- H_a: This sample is not from a poisson distribution

Contingency Tables

Often times, we have no *a priori* assumption about the distribution of our sample.

We may be interested in:

1. Do two or more samples come from the same population? If so, use a chi-square test of homogeneity.
2. Is there a relationship between two or more variables? If so, use a chi-square test of independence.

Contingency Tables

Homogeneity test
 Emphasis on how 2 or more populations differ WRT one character

Independence test
 Emphasis on how a single sample differs WRT two or more characters
 (2×2 or $R \times C$ tests)



Contingency Tables

- Example: 2×2 -




Suppose a field botanist finds two different morphs of a rare species (pink & blue flowers).


Suppose also that the botanist suspects that morphs are related to soil type (normal & calcareous).

A stratified random sample is then collected of 100, 50 of each flower color, and the soil type noted.

- H_o : Flower color is independent of soil type
- H_a : Flower color is dependent on soil type



Contingency Tables




- Example -

The observed frequency (count) data are:


	B	P	Total
Normal soil	34	12	46
Calcareous soil	16	38	54
Total	50	50	100

First calculate expected values for each cell element: $(\text{Row-total})(\text{Col-total})/100$

So, for N/B, $E = (46)(50)/100 = 23$



Contingency Tables



- Example -

Class	O	E	O-E	(O-E) ²	(O-E) ² /E
N, B	34	23	11	121	5.261
C, B	16	27	-11	121	4.481
N, P	12	23	-11	121	5.261
C, P	38	27	11	121	4.481
					19.485

$v = (r-1)(c-1)$; where r = rows & c = columns

Thus, in this example, $v = 1$

$\chi^2_{\text{calc}} = 19.485 > \chi^2_{0.05,1} = 3.841$

Therefore, reject H_0 , flower color is soil dependent

```

> Flowers<-matrix(c(34,16,12,38), nrow=2, dimnames =
list("Soil" = c("Normal", "Calcareous"), "Flowers" =
c("Blue", "Pink")))

> Flowers
      Flowers
Soil   Blue Pink
Normal  34  12
Calcareous 16  38

> fisher.test(Flowers)

Fisher's Exact Test for Count Data

data: Flowers
p-value = 1.885e-05
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 2.575011 17.903778
sample estimates:
odds ratio
 6.58276
    
```

Contingency Tables

- R × C -

It should now be obvious that this general technique can be expanded to accommodate any number of R rows and C columns (R × C).

For example, we could have had 4 color morphs (white, pink, lavender, blue) and two soil types (normal, calcareous) and tested the same hypothesis.



```
> Bears<-matrix(c(32,55,43,65,9,16),
ncol = 3, dimnames=list("Sex" =
c("Male", "Female"), "Hair color" =
c("Black", "Brown", "Red")))
```

```
> chisq.test(Bears)
```

Pearson's
Chi-squared test



```
data: Bears
X-squared = 0.2447, df = 2, p-value =
0.8848
```

```
> chisq.test(Bears)$observed
Hair color
```

Sex	Black	Brown	Red
Male	32	43	9
Female	55	65	16



```
> chisq.test(Bears)$expected
```

Sex	Black	Brown	Red
Male	33.21818	41.23636	9.545455
Female	53.78182	66.76364	15.454545

Chisq.test has some extended functions that allow us to examine the observed and expected values separately. We can then calculate chi-square manually to see what the specific contributions are for each $R \times C$ combination:

```
> E <- chisq.test(Bears)$expected
> O <- chisq.test(Bears)$observed
> (O-E)^2/E
```



Sex	Hair color		
	Black	Brown	Red
Male	0.04467333	0.07542889	0.03116883
Female	0.02759235	0.04658843	0.01925134

Log-Linear Analysis

It is quite common in biological sciences to want to analyze the relationship (independence) of three or more variables simultaneously. This is simply an expansion of the $R \times C$ contingency table into additional dimensions (e.g., $R \times C \times D$)

Let's look at a classic data set which examines hair color, eye color, and sex using a log-linear approach and the R function LOGLIN.

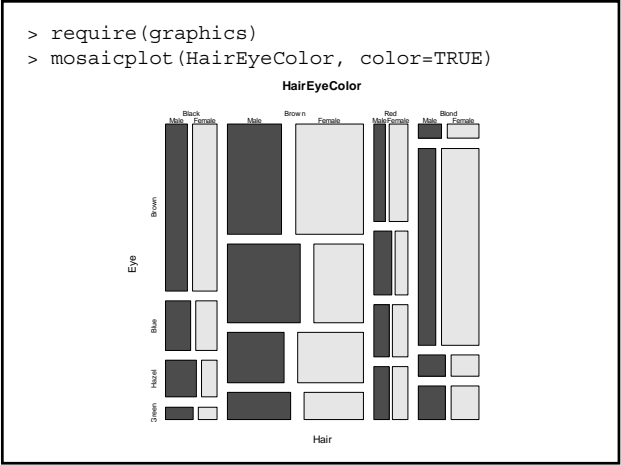
Data set is based on 592 observations (people) that are scored for 3 characters:

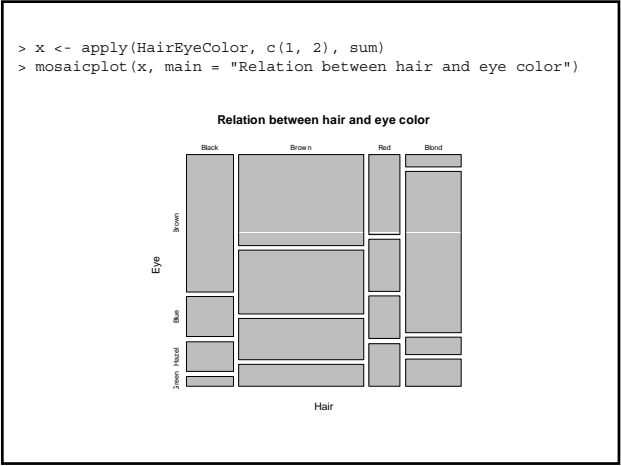
- Hair = Black, Brown, Red, Blond
- Eye = Brown, Blue, Hazel, Green
- Sex = Male, Female

```
title 'Hair - Eye color data';
data haireye;
length hair $8 eye $6 sex $6;
drop c i black brown red blond;
array h{*} black brown red blond;
c='Black Brown Red Blond';
input sex $ eye $ black brown red blond;
do i=1 to dim(h);
  count = h(i); hair=scan(c,i);
  output;
end;
datalines;
M Brown 32 53 10 3
M Blue 11 50 10 30
M Hazel 10 25 7 5
M Green 3 15 7 8
F Brown 36 66 16 4
F Blue 9 34 7 64
F Hazel 5 29 7 5
F Green 2 14 7 8
;
```



<http://euclid.psych.yorku.ca/ftp/sas/vcd/catdata/haireye.sas>





```

> fm <- loglin(HairEyeColor, list(c(1, 2), c(1, 3), c(2, 3)))
5 iterations: deviation 0.04093795

> fm
$lrt          ## the Likelihood ratio Statistic
[1] 6.761258

$pearson      ## the Pearson chi-sq test statistic
[1] 6.868292

$df           ## degrees of freedom
[1] 9

$margin      ## List of margins fit to model
$margin[[1]]
[1] "Hair" "Eye"

$margin[[2]]
[1] "Hair" "Sex"

$margin[[3]]
[1] "Eye" "Sex"

```

To determine probability of model:

```
> 1 - pchisq(fm$lrt, fm$df)
[1] 0.66196
```

There is no 3-way interaction. A model with two factors will likely fit best. Now, look at all 2-way interactions separately.

The End

This completes our discussion of methods used to handle one variable and two or more samples.

Next we will turn our attention to methods used to relate two variables from one sample.
