Analysis of Counts

When data are counts or classes
- dead or alive
- pregnant or not
- healthy or diseased

When continuous data is broken into classes
- income levels

Uses
1. Test hypothetical ratios
2. Determine whether characteristics are inter-related
3. Test whether samples are from different populations

Chi square test

$$\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

Chi-square test provides an approximation for a binomial distribution
Chi-square distribution is continuous and related to normal distribution

Best with:
- large sample size
- ratios close to 1:1
- expected classes all $\geq$ 5
- use Yates correction when df = 1 and row and column total are predetermined.

Yates Correction

$$\chi^2 = \sum \frac{(|\text{Obs} - \text{Exp}| - 0.5)^2}{\text{Exp}}$$

1. Test hypothetical ratios

Individuals classified in one way into 2 or more classes
Compare to hypothetical or expected ratio
Degrees of freedom = number of classes - 1
Example: frequency of observation of dominant genetic trait

2. Determine whether characteristics are inter-related

Individuals classified in two ways, in r and c classes
Test for independence between classification criteria
Degrees of freedom = (r - 1)(c - 1)

Example: Disease incidence in treated and untreated cattle

Null hypothesis:

no effect of treatment
disease incidence is the same in both groups
disease incidence and inoculation are independent
probability is product of probabilities of disease and treatment

Disease incidence in cattle

	Disease		
Treatment	Healthy	Diseased	Total
Treated	88	12	100
Expected	(77)	(23)	
Untreated	143	57	200
Expected	(154)	(46)	
Totals	231	69	300

$$\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

$$= \frac{(11)^2}{77} + \frac{(11)^2}{23} + \frac{(11)^2}{154} + \frac{(11)^2}{46} = 10.25$$

$$\chi^2_{.05,1} = 3.84$$

df = (2-1)(2-1) = 1

3. Test whether samples are from different populations

Calculate chi-square for each sample
Calculate chi-square for pooled samples
Difference is chi-square for heterogeneity
Only uncorrected chi-squares are additive
Example: normal and virescent marigolds in 8 progenies
Hypothetical ratio is 3:1 for expected values

Progeny	Normal	Virescent	Chi^2(3:1)	Chi^2(3160:854)
1	315	85	3.00	0.023
2	602	170	3.65	0.094
3	868	252	3.73	0.578
4	174	42	3.56	0.575
5	192	48	3.20	0.348
6	165	39	3.76	0.723
7	161	43	1.67	0.028

8	629	175	4.48	0.019
Totals			27.05	2.388
Pooled	3106	854	24.91	--
Heterogeneity			2.14	2.38

Confidence Intervals

	Outcome 1	Outcome 2	Total
Situation 1	r_1	n_1-r_1	n_1
Situation 2	r_2	n_2-r_2	n_2
Situation k	r_k	n_k-r_k	n_k
Total	Σr	Σn-Σr	Σn

Expected value for Outcome 1 and Situation k is $n_k(\Sigma r/\Sigma n)$
Expected value for Outcome 2 and Situation k is
$n_k(\Sigma n- \Sigma r)/\Sigma n$

Best estimate of Outcome 1 for Situation 1 is r_1/n_1

For sample mean = r/n
Distribution of means is approximately normal for large sample size

Probability of outcome r is p
For binomial distribution
expected mean value of r = np
variance is np(1-p)

For sample mean r/n
mean value of r/n = (np)/n = p
variance of r/n = p(1-p)/n
For large sample
95% confidence interval is $\frac{r}{n} \pm 1.96\sqrt{\frac{p(1-p)}{n}}$
Estimating p by r/n $\frac{r}{n} \pm 1.96\sqrt{\frac{r/n(1-r/n)}{n}}$

Example: Seed germination trial
r = 80 seeds out of n = 100 seeds germinate
r/n = 0.8 or 80% germination
variance is

$$s^2 = \frac{r/n(1-r/n)}{n} = \frac{(0.8\times0.2)}{100} = 0.0016$$

$$SD = \sqrt{s^2} = 0.04$$

Confidence Interval = $0.080 \pm 1.96\times0.04 = (0.72,0.88)$

Difference of two proportions

Population 1 proportion of successes = r_1/n_1
Population 2 proportion of successes = r_2/n_2
Difference = $r_1/n_1 - r_2/n_2$

Variance =

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

or

$$\frac{r_1/n_1(1-r_1/n_1)}{n_1} + \frac{r_2/n_2(1-r_2/n_2)}{n_2}$$

Confidence Interval

$$\left(\frac{r_1}{n_1} - \frac{r_2}{n_2}\right) \pm 1.96\sqrt{\frac{r_1/n_1(1-r_1/n_1)}{n_1} + \frac{r_2/n_2(1-r_2/n_2)}{n_2}}$$

Example continued: New process

r_2 = 175 seeds germinated out of n_2 = 200
Mean = r_2/n_2 = 175/200 = 0.875
Variance 2 = (0.875 x 0.125)/200 = 0.00055

Improvement in germination

0.875 - 0.8 = 0.075 or 7.5%

Confidence interval

$$0.075 \pm 1.96\sqrt{0.0016 + 0.00055}$$

or (−0.015 to 0.165)

Square of the ratio of the difference to the SD of the difference is equivalent to chi square

$(0.075/0.046)^2 = 2.66$

Sample size for estimating proportions

Need an estimate of p

If p = 0.85 or 85%
Variance = p(1-p)/n = (0.85)(0.15)/n = 0.1275/n

For n = 100, variance = 0.001275, SE = 0.0357

For SE = 0.01

$0.1275/n = (0.01)^2$
n = 1275

For p = 0.3 and desired confidence interval $\leq$ 0.1

$$\text{CIwidth} = 2(1.96)\sqrt{\frac{p(1-p)}{n}}$$

$$0.01 = 3.92\sqrt{\frac{(0.3)(0.7)}{n}}$$

$$n = \frac{(3.92^2)(0.21)}{0.1^2} = 323$$

To compare two percentages

for estimated n and p can calculate confidence interval
most precise with equal sample sizes, ie same n
null hypothesis is same p

If SE is less than 1/3 of difference desired to detect
power is 4 to 1 chance of detecting difference at 5% significance

For p = 0.6 and n = 250

If desired difference = 0.1 or 10%
SE should be 0.1/3 = 0.0333

$$\sqrt{\frac{2(0.6)(0.4)}{n}} = \sqrt{\frac{0.48}{n}} = 0.0333$$

$$n = \frac{0.48}{0.0333^2} = 432$$