STAT 251 / OBEE 216
Winter 2002
Prof. Ivo D. Dinov
Inference for population variances and proportions and intro to categorical data
Reading: Ch. 4.4, Ch. 6
.

Inference for the unknown variance σ^2 of a normal population

A marine biologist wishes to use male angelfish for an experiment and hopes their weights don't vary much. In fact, a previous random sample of $n = 16$ angelfish yielded the data below

$$\{y_1, \dots, y_n\} = \{5.1, 2.5, 2.8, 3.4, 6.3, 3.6, 3.9, 3.0, 2.7, 5.7, 3.5, 3.6, 5.3, 5.1, 3.5, 3.3\}$$

Sample statistics from these data include

$$\bar{y} = 3.96 \text{ lbs} \quad s^2 = 1.35 \text{ lbs}^2 \quad n = 16$$

Problem: obtain a $100(1-\alpha)\%$ confidence interval for σ^2.

Point Estimator for σ^2? How about S^2?

Sampling theory for S^2?

If a random sample $Y_1, \dots, Y_n$ is taken from a normal population with mean μ and variance σ^2, then

$$\frac{\sum(Y_j - \bar{Y})^2}{\sigma^2} \sim \chi^2(n-1)$$

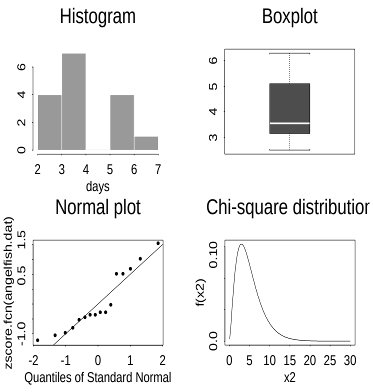
Critical values for the χ^2 distribution appear in Table C.3 on pp 813-814 of Rao. These values cover distributions with up to $\nu = 100$ degrees of freedom. This result can be used to obtain confidence intervals for the variance σ^2 of a normal population:

$$1 - \alpha = \Pr(\chi^2(n-1, 1-\alpha/2) \leq \frac{\sum(Y_j - \bar{Y})^2}{\sigma^2} \leq \chi^2(n-1, \alpha/2)).$$

The term in the middle is just $(n-1)S^2/\sigma^2$. The usual algebraic rearrangement yields a confidence interval of the form

$$\boxed{\frac{(n-1)S^2}{\chi^2(n-1,\alpha/2)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2(n-1,1-\alpha/2)}}.$$

Figure 1: Assessments of normality/sampling distribution of $(n-1)S^2/\sigma^2$:



For the angelfish data, first we might check for obvious departures from normality: To obtain a 95% confidence interval, the appropriate critical values are

$$\chi^2(15, 0.025) = 27.49 \quad \text{and } \chi^2(15, 0.975) = 6.26.$$

This yields the interval

$$\frac{(n-1)s^2}{\chi^2(n-1, \alpha/2)}, \ \frac{(n-1)s^2}{\chi^2(n-1, 1-\alpha/2)}$$

or

$$\frac{(16-1)1.35}{27.49}, \ \frac{(16-1)1.35}{6.26}$$

or

$$(0.74, 3.24)$$

The ratio of **two** population variances, σ_1^2/σ_2^2, from independent samples

Consider two independent random samples

$$Y_{1,1}, \dots, Y_{1,n_1}$$

$$Y_{2,1}, \dots, Y_{2,n_2}$$

from two **normal** populations with unknown variances σ_1^2 and σ_2^2, respectively. Questions:

- What is a good point estimator of σ_1^2/σ_2^2?
- Can this be used for a test of significance or confidence interval for σ_1^2/σ_2^2?

Sampling distributions of S_1^2 and S_2^2 from normal populations

Recall the t-test comparing air pollution in homes of smokers and non-smokers. The common variances procedure was ruled out because of the large difference in sample variances:

$$\begin{array}{ll} S_1^2 = 26.0 & (n_1 = 11) \\ S_2^2 = 195.4 & (n_2 = 9) \end{array}$$

Suppose we want to formally test the hypothesis that the population variances are equal. Consider a test of the form

$$H_0 : \sigma_1^2 = \sigma_2^2 \ \text{ vs. } \ H_1 : \sigma_1^2 \neq \sigma_2^2$$

which can also be written

$$H_0 : \theta = \frac{\sigma_1^2}{\sigma_2^2} = 1 \ \text{ vs. } \ H_1 : \theta = \frac{\sigma_1^2}{\sigma_2^2} \neq 1.$$

How about

$$\hat{\theta} = \frac{S_1^2}{S_2^2}?$$

Where
S_1^2 is the sample variance from $Y_{1,1}, \dots, Y_{1,n_1}$ and
S_2^2 is the sample variance from $Y_{2,1}, \dots, Y_{1,n_2}$:

$$\begin{aligned} S_1^2 &= \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1,i} - \bar{Y}_1)^2 \\ S_2^2 &= \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_{2,i} - \bar{Y}_2)^2 \end{aligned}$$

($\hat{\theta}$ is sometimes called an F-ratio.)

To test H_0, the hypothesis of equality population variances, we use the following result:

$$\frac{\hat{\theta}}{\theta} \sim F_{n_1-1,n_2-1}$$

which can also be written

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

This yields a probability statement of the form

$$1-\alpha = \Pr(F(n_1-1, n_2-1, 1-\alpha/2) \leq \frac{S_1^2}{S_2^2}\frac{1}{\theta} \leq F(n_1-1, n_2-1, \alpha/2); H_0 \text{ is true }) \tag{1}$$

Values of the F-ratio which are far from one constitute evidence against the null hypothesis. Formally, the critical region with level α calls for rejection of H_0 whenever

$$\hat{\theta} < F_{n_1-1,n_2-1}(1-\alpha/2) \qquad \text{or} \qquad \hat{\theta} > F_{n_1-1,n_2-1}(\alpha/2).$$

Manipulation of (1) leads to the following $100(1-\alpha)\%$ confidence interval for $\theta = \sigma_1^2/\sigma_2^2$:

$$\boxed{\left(\frac{S_1^2}{S_2^2}\frac{1}{F(n_1-1,n_2-1,\alpha/2)},\quad \frac{S_1^2}{S_2^2}\frac{1}{F(n_1-1,n_2-1,1-\alpha/2)}\right)}$$

For a 95% confidence interval for σ_1^2/σ_2^2 in the smoking data, we need

$$F(10, 8, 0.975) = 0.259 \quad F(10, 8, 0.025) = 4.295$$

which yields the interval

$$\left(\frac{26.0}{195.4 \times 4.295}, \quad \frac{26.0}{195.4 \times 0.259}\right)$$

or

$$(0.031, 0.512)$$

which clearly doesn't contain 1, so that $H_0 : \sigma_1^2 = \sigma_2^2$ is rejected at level $\alpha = 0.05$.

The p-value for such a test can be obtained from the F distribution and the observed test statistic:

$$F_{obs} = \hat{\theta}_{obs} = 26/195.4 = 0.133$$

However, recall that Table C.4 only gives upper critical values.

Therefore, to obtain a p-value, take as the test statistic

$$\max\{\hat{\theta}, 1/\hat{\theta}\}$$

and multiply the right-tail probability from the F-distribution by 2. Use the following numerator (df_1) and denominator (df_2) degrees of freedom:

$$\begin{aligned} df_1 &= df \text{ from bigger of } \{s_1^2, s_2^2\} \\ df_2 &= df \text{ from smaller of } \{s_1^2, s_2^2\} \end{aligned}$$

The observed test statistic becomes

$$F_{obs} = 1/\hat{\theta} = \frac{s_2^2}{s_1^2} = \frac{195.4}{26.0} = 7.53$$

and since

$$F(8, 10, 0.01) = 5.057$$

the area to the right of $F_{obs} = 7.53$ under the $F_{8,10}$ distribution is less than 0.01, which corresponds to a two-sided p-value less than 0.02. Note that the degrees of freedom must be switched when $S_2^2 > S_1^2$.

```
options ls=75 nodate;

data one;
   infile "datasets/smokers.dat";
   input y smoke;
   label y="suspended particulate matter";
run;

proc ttest;
   class smoke;
   var y;
run;
```

```
                              The SAS System                              1

                              TTEST PROCEDURE

Variable: Y            suspended particulate matter

SMOKE       N               Mean             Std Dev             Std Error
--------------------------------------------------------------------------
    0       9        92.77777778         13.98014465            4.66004822
    1      11       133.18181818          5.09545252            1.53633674

Variances        T       DF    Prob>|T|
---------------------------------------
Unequal    -8.2343      9.7      0.0001
Equal      -8.9320     18.0      0.0000

For H0: Variances are equal, F' = 7.53    DF = (8,10)    Prob>F' = 0.0045
```

Large sample interval estimation for a population proportion

Out of a random sample of $n = 330$ triathletes, 167 indicated that they had suffered a training-related injury during the past year. Using these data, give a point estimate, standard error and confidence interval for

p : the proportion among ALL triathletes who suffered an injury

Let

$$\hat{p} := \text{sample proportion of injured triathletes}$$

We know from the CLT for proportions that the sampling distribution of $\hat{p}$ is approximately normal. This yields the following approximate probability statement:

$$\begin{aligned} 0.95 &\approx \Pr\left(-1.96 < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < 1.96\right) \\ &\vdots \\ &= \Pr\left(\hat{p} - 1.96\sqrt{p(1-p)/n} < p < \hat{p} + 1.96\sqrt{p(1-p)/n}\right) \\ &\approx \Pr\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \end{aligned}$$

The endpoints for a 95% confidence interval for an unknown population proportion p based on a random sample of size n with sample proportion $\hat{p}$ are then given by

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ and } \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

which is commonly written

$$\boxed{\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.}$$

For the triathlete data, a 95% confidence interval for p based on the sample proportion of $\hat{p} = 167/330 = 0.506$ is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

or

$$0.506 \pm 1.96(0.028)$$

or

$$0.506 \pm 0.053$$

Sample size computations for confidence intervals

Case 2: Estimation of a population proportion p.
The sample size necessary to obtain a 95% confidence interval of the form

$$\hat{p} \pm B$$

for an unknown population proportion p based on a random sample can be solved for similarly, yielding the equation

$$n = \left(\frac{1.96\sqrt{\hat{p}(1-\hat{p})}}{B} \right)^2 . \tag{2}$$

Upon inspection of (2), it can be seen that the term on the right is bounded above by

$$\left(\frac{1.96}{B} \right)^2 * (1/4)$$

so that a conservative sample size, which will ensure a 95% confidence interval of length 2*B is given by

$$n = \left(\frac{1.96}{B} \right)^2 * (1/4).$$

Exercise: Suppose you want to estimate the proportion p of trees that will survive to a certain lifetime under some treatment of interest. In particular, you'd like a 95% confidence interval of the form

$$\hat{p} \pm 0.02.$$

How large does your sample size n need to be ...

- without knowing anything about p?
- with the knowledge that the least p could reasonably be is $p = 0.9$?

Testing with dichotomous data

Example: There is a theory that the anticipation of a birthday can prolong a person's life. In a study, it was found that only $x = 60$ out of a random sample of $n = 747$ people whose obituaries were published in Salt Lake City in 1975 died in the three-month period preceding their birthday (Newsweek, 1978). Let p denote the proportion of all deaths which fall in the three-month period preceding a birthday. Consider the following test

$$H_0 : p = 0.25 \ (= p_0) \quad \text{vs} \quad H_1 : p < 0.25$$

The test statistic for this problem takes the usual form

$$Z = \frac{\text{est} - \text{null}}{\text{SE(est)}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/747}}$$

Note that the standard error term in the denominator does not need to be estimated (by $\hat{p}$) since it is specified under H_0. The left-tailed test with level α rejects H_0 if $Z < -z(\alpha)$. Similarly for right-tailed and two-tailed tests:

Alternative	Critical region
$H_1 : p < p_0$	$Z < -z(\alpha)$
$H_1 : p > p_0$	$Z > z(\alpha)$
$H_1 : p \neq p_0$	$\lvert Z \rvert > z(\alpha/2)$

For the Newsweek obituary data,

$$z_{obs} = \frac{60/747 - 0.25}{\sqrt{0.25(1-0.25)/747}} = \frac{0.08 - 0.25}{0.0099} = -17$$

So we reject H_0 with a p-value less than 0.001.

Some “categorical” datasets:

Dataset #1: Tomato plants.

Phenotype	Frequency
Tall, cut	926
Tall, potato	288
Dwarf, cut	293
Dwarf, potato	104

Dataset #2: Yeast cells

The distribution of yeast cells observed over $n = 400$ squares of a haemacytometer:

y	0	1	2	3	4	5	≥ 6
$f(y)$	213	128	37	18	3	1	0

Dataset #3: Colds among skiers taking vitamin C and placebo

	Cold	No Cold	Total
Placebo	31	109	140
Vitamin C	17	122	139

Dataset #4: Presidential candidates

		after debate		
		G	B	
before	G	63	21	84
debate	B	4	12	16

Dataset #5: Handedness and gender

Handedness	Men	Women	Total
Right	934	1070	2004
Left	113	92	205
Ambidextrous	20	8	28
Total	1067	1170	2237

The multinomial probability distribution

The **multinomial distribution** is a generalization of the binomial distribution arising from independent, identically distributed trials, each of which can be categorized as one and only one of $C \geq 2$ possible categories, with probabilities $\pi_1, \pi_2, \ldots, \pi_C$. If n such i.i.d. trials are observed, each with probabilities $(\pi_1, \ldots, \pi_C)$ then the probability of obtaining exactly

- y_1 trials categorized as type 1
- y_2 trials categorized as type 2
- $\vdots$
- y_C trials categorized as type C

is given by

$$\frac{n!}{y_1!y_2! \times \cdots \times y_C!}\pi_1^{y_1}\pi_2^{y_2} \times \cdots \times \pi_C^{y_C}$$

For example, if tomato plants are grown in such a way that they are classified as one of the four phenotypes in Dataset #1 with probabilities

$$\pi_1 = 0.56, \quad \pi_2 = 0.19, \quad \pi_3 = 0.19, \quad \pi_4 = 0.063$$

and $n = 10$ plants are grown, then the chance of getting, say exactly

$y_1 = 5$	Tall,cut
$y_2 = 2$	Tall,potato
$y_3 = 2$	Dwarf/cut
$y_4 = 1$	Dwarf/potato

is given by

$$\frac{10!}{5!2!2!1!}0.56^5 0.19^2 0.19^2 0.063^1 = 0.033$$

Note: Results for the multinomial distribution underlie many of the techniques for categorical data analysis we'll study.

The χ^2 goodness-of-fit tests for categorical data with completely specified cell probabilities

The χ^2 goodness-of-fit test can be used for inference about these $C-1$ parameters. (Since $\pi_1 + \pi_2 + \cdots + \ldots \pi_C = 1$ there are really only $C-1$ parameters.) In particular, it can be used to test hypotheses of the form

$$H_0 : \pi_1 = \pi_{10}, \pi_2 = \pi_{20}, \ldots, \pi_C = \pi_{C0}$$

versus

$$H_1 : \pi_j \neq \pi_{j0} \text{ for at least one } j$$

Suppose that n i.i.d. trials are observed, each with probability of being classified (uniquely) as category j given by π_j. Let the RV representing the number of trials classified as category j be denoted by O_j:

$$O_j = \text{ \# trials classified as type} j.$$

Using properties of this multinomial distribution, it can be shown that when H_0 holds, the χ^2 test statistic below has (approximately) the χ^2 distribution with $C-1$ degrees of freedom:

$$\chi^2 = \sum_{j=1}^{j=C} \frac{(O_j - n\pi_{j0})^2}{n\pi_{j0}}$$

This test statistic is a bit easier to remember in the following form

$$\chi^2 = \sum_{j=1}^{j=C} \frac{(O_j - E_j)^2}{E_j}$$

where O_j denotes the observed count in the j^{th} category and E_j is the expected count under H_0:

$$E_j = E(O_j; H_0) = n\pi_{j0}$$

A critical region for χ^2 is the set of values bigger than $\chi^2(C-1, \alpha)$. That is,

$$\text{reject } H_0 \text{ if } \chi^2 \geq \chi^2(C-1, \alpha).$$

The p-value is just the area to the right of the observed value of the test statistic under the χ^2 curve with $C-1$ degrees of freedom.

Example: Two traits that have been widely studied in tomato plants are *height* ("tall" vs "dwarf") and *leaf type* ("cut" vs "potato"). "Tall" and "cut" are dominant. When a homozygous "tall,cut" is crossed with a "dwarf,potato" the resulting progeny is called a dihybrid. When dihybrids are crossed, the following proportions of phenotypes should appear in the offspring provided the alleles governing the two traits segregate independently (this is a $9:3:3:1$ ratio:)

Phenotype	Relative Frequency
Tall, cut	0.5625
Tall, potato	0.1875
Dwarf, cut	0.1875
Dwarf, potato	0.0625

In one experiment done with these two traits a total of 1611 progeny of dihybrid crosses were categorized by phenotype. The data are summarized in the table below:

Phenotype	Frequency
Tall, cut	926
Tall, potato	288
Dwarf, cut	293
Dwarf, potato	104

Specify the null and alternative hypothesis for this experiment:

$$H_0 :?$$

$$H_1 :?$$

How about

$$H_0 : \pi_1 = 0.5625,\ \ \pi_2 = 0.1875,\ \ \pi_3 = 0.1875,\ \ \pi_4 = 0.0625$$

vs

$$H_1 : \text{at least one } \ \pi_j \neq \pi_{j0}?$$

To test these hypotheses, the χ^2 test statistic becomes

$$\begin{aligned}\chi^2 &= \sum_1^4 \frac{(O_j - E_j)^2}{E_j} \\ &= \frac{(926 - 906.2)^2}{906.2} + \frac{(288 - 302.1.2)^2}{302.1} + \frac{(293 - 302.1)^2}{302.1906} + \frac{(104 - 100.7)^2}{100.7} \\ &= 1.47\end{aligned}$$

Is this statistically significant?

The distributional result on page 1 implies that when H_0 holds, the test statistic should have a χ^2 sampling distribution with $4 - 1 = 3$ degrees of freedom. The 95^{th} percentile for this distribution, found in Rao, is given by

$$\chi^2(0.05, 3) = 7.8147$$

The observed test statistic is therefore not statistically significant using $\alpha = 0.05$. The p-value, obtained using statistical software is given by

$$p - \text{value} = \Pr(\chi^2 \geq 1.47; H_0) = 0.69.$$

Conclusion ?:

Rule of thumb to check validity of χ^2 approximation

- at least 75% of the cells have $E_j \geq 5$ (Expected counts are not too small) AND
- no Expected 0's ($E_j \neq 0$)

Another test for categorical data: partially specified probabilities

Often, the category probability parameters are not completely specified, but rather are specified up to some unknown parameter. Examples include fitting a well-known discrete probability model, such as the poisson or binomial models to data, or making a continuous model into a discrete model by grouping observations in bins.

Example: a poisson probability model. The distribution of yeast cells observed over $n = 400$ squares of a haemacytometer is given below:

y	0	1	2	3	4	5	≥ 6
$f(y)$	213	128	37	18	3	1	0

To test the hypothesis that these data are a random sample from a Poisson distribution, we could write

$$H_0 : \Pr(y \text{ yeast cells in a square}) = e^{-\lambda}\lambda^y/y! \text{ for } y = 0, 1, 2, \ldots$$

$$H_1 : \Pr(y \text{ yeast cells in a square}) \neq e^{-\lambda}\lambda^y/y! \text{ for } \text{ for some } y$$

however, there would be many zeroes and many small cell counts, so we can bin the data a bit differently to avoid this problem.

Category j	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
y	0	1	2	3	≥ 4
$f(y)$	213	128	37	18	4

Then we can test

$$\begin{aligned} H_0 : \quad & \pi_1 = e^{-\lambda}, \\ & \pi_2 = e^{-\lambda}\lambda, \\ & \pi_3 = e^{-\lambda}\lambda^2/2!, \\ & \pi_4 = e^{-\lambda}\lambda^3/3!, \\ & \pi_5 = 1 - \textstyle\sum_1^4 \pi_j \\ H_1 : \quad & \text{any other probabilities} \end{aligned}$$

only we don't know λ and must estimate it from the data. This is what is meant by partially specified probabilities. The resulting test statistic below has an approximate χ^2 distribution with $C - 1 - p$ degrees of freedom where C denotes the number of categories or bins and p denotes the number of parameters used to specify the category probabilities. For the poisson model, $p = 1$.

(For a normal model, where μ and σ must be estimated by $\bar{Y}$ and S, the number of parameters would be $p = 2$.)

The mean of the sample, $\bar{Y}$ can be used to estimate λ, the mean of the poisson distribution:

$$\hat{\lambda} = \bar{y} = \frac{\sum y_j}{n} = \frac{273}{400} = 0.6825.$$

Substituting $\hat{\lambda}$ into the poisson model for the category probabilities yields the following expected cell counts:

y	0	1	2	3	≥ 4
j	1	2	3	4	5
O_j	213	128	37	18	4
$\hat{\pi}_j$	0.505	0.345	0.118	0.028	0.005
E_j	202.1	138.0	47.1	10.7	2.1

The $j = 3$ cell probability $\hat{\pi}_3$, for example, comes from

$$\hat{\pi}_3 = \frac{e^{-\hat{\lambda}}\hat{\lambda}^2}{2!} = 0.118$$

and the expected cell counts are just

$$E_j = n\hat{\pi}_j \text{ for } j = 1, \ldots, 5$$

The $\alpha = 0.05$ critical value for the χ^2 test statistic can be obtained from the χ^2 distribution with $C - 1 - p = 3$ degrees of freedom from Table C.3, (p. 814) of Rao:

$$\chi^2(3, 0.05) = 7.8147$$

The observed value of the test statistic is

$$\chi^2 = \sum \frac{(O_j - E_j)^2}{E_j} = \frac{(213 - 202.1)^2}{202.1} + \cdots + \frac{(4 - 2.1)^2}{2.1} = 10.12$$

Q: Does the test statistic fall in the $\alpha = 0.05$ critical region?

Q: Do the poisson model fit these data?

Q: What is the p-value for the test statistic χ^2 under H_0?

$$\chi^2(3, 0.025) = 9.3 \text{ and } \chi^2(3, 0.01) = 11.3$$

Large sample comparison of population proportions, π_1, π_2 based on independent random samples

Example: In a review of the evidence regarding the therapeutic value of vitamin C for prevention of the common cold, Pauling (1971) describes a 1961 French study involving 279 skiers during two periods of 5-7 days. One group of 140 subjects received a placebo while the remaining 139 received 1 gram of vitamin C per day. Of interest is the relative occurrence of colds for the two groups. The data are shown below. Let p_1 denote the proportion among a population of people who take the treatment who would catch a cold. Let p_2 denote the proportion among a population of people who take the placebo who would catch a cold.

	Cold	No Cold	Total
Placebo	31	109	140
Vitamin C	17	122	139

1. Formulate a test of hypotheses to investigate whether or not the catching of colds differs by vitamin C intake.
2. Calculate the $p-$value for your test from these data. If you use an approximation to obtain this $p-$value, verify that it is appropriate.
3. Obtain a 95% confidence interval for the quantity $p_1 - p_2$.
4. Let q_1 be defined by $q_1 = 1 - p_1$. Suppose that you are particularly interested in the quantity $\theta = q_1 - p_1$. Propose a point estimator of this quantity.
5. Construct a 95% confidence interval for θ.

We've seen from the CLT for proportions that if $\hat{p}_1$ denotes a sample proportion (of some 0-1 trait of interest) from a random sample of size n_1 taken from a population with proportion p_1 then (approximately)

$$\frac{\hat{p}_1 - p_1}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}} \sim N(0,1)$$

Similarly, if another sample proportion $\hat{p}_2$ is obtained from a random sample of size n_2 taken independently from another population with proportion p_2, then (approximately)

$$\frac{\hat{p}_2 - p_2}{\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$

We also know that a sum or difference of two independent, normally distributed random variables also has a normal distribution. This implies that

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$

The following probability statement is a consequence of this normality:

$$1 - \alpha \approx \Pr\left(-z(\alpha/2) \le \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \le z(\alpha/2)\right)$$

The usual rearrangement yields a 95% confidence interval for $p_1 - p_2$ of the form

$$\boxed{\hat{p}_1 - \hat{p}_2 \pm z(\alpha/2)\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

For tests like $H_0 : p_1 - p_2 = D_0$ versus $H_1 : p_1 - p_2 \neq D_0$, the following test statistic can be used:

$$Z_1 = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

For the (most common) case where $D_0 = 0$ is of interest, a better test is one based on the statistic

$$Z_2 = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where

$$\hat{p} = \frac{n_1}{n_1 + n_2}\hat{p}_1 + \frac{n_2}{n_1 + n_2}\hat{p}_2.$$

Critical regions for one-sided and two-sided alternatives are formed in the usual manner. It can be shown that Z^2 and the χ^2 statistic for independence in a 2 × 2 table are the same (see pp. 20-21.)

For the vitamin C data, a 95% confidence interval for $p_p - p_C$ is given by $(0.004, 0.194)$. The 2nd test statistic works out to $z_{obs} = 2.19$ and a two-sided p-value of 0.0283:

```
data one;
   input cold trt $ frq;
   cards;
   1 p 31
   0 p 109
   1 C 17
   0 C 122
;
run;

proc freq;
   weight frq;
   tables cold*trt/chisq;
run;
```

```
                              The SAS System                                 1
                           TABLE OF TRT BY COLD

                    TRT       COLD

                    Frequency|
                    Row Pct  |       0|       1|  Total
                    ---------+--------+--------+
                    C        |    122 |     17 |    139
                             |  87.77 |  12.23 |
                    ---------+--------+--------+
                    p        |    109 |     31 |    140
                             |  77.86 |  22.14 |
                    ---------+--------+--------+
                    Total         231       48      279

                    STATISTICS FOR TABLE OF TRT BY COLD

          Statistic                     DF      Value        Prob
          ------------------------------------------------------
          Chi-Square                     1      4.811       0.028
          Fisher's Exact Test (Left)                        0.991
                              (Right)                       0.021
                              (2-Tail)                      0.038
```

McNemar's test for significance of changes

McNemar's test can be used to test for a difference of proportions in paired categorical data. That is, two 0-1 measurements are made on each experimental unit. Consider hypothetical data representing preferences among democratic voters for a presidential candidate, G or B, before and after a debate. Here, there are two measurements made on each experimental unit (democratic voters).

		after debate		
		G	B	
before	G	$a = 63$	$b = 21$	$N_1 = 84$
debate	B	$c = 4$	$d = 12$	$N_2 = 16$
Total		$M_1 = 67$	$M_2 = 33$	$N = 100$

The difference in the proportion of people who support Gore before (π_1) and after (π_2) the debate, $\theta = \pi_1 - \pi_2$ can be estimated using

$$\hat{\theta} = \hat{\pi}_1 - \hat{\pi}_2 = \frac{N_1}{N} - \frac{M_1}{N}$$

For these data, this works out to

$$\hat{\theta} = \frac{84}{100} - \frac{67}{100} = \frac{63 + 21 - (63 + 4)}{100} = \frac{21 - 4}{100}$$

In general (a, b, c, d) this estimator works out to

$$\hat{\theta} = \frac{b - c}{N}$$

It can be shown that the standard error can be estimated by of $\hat{\theta}$ is given by

$$SE(\hat{\theta}) = \frac{\sqrt{b + c}}{N}$$

yielding a test statistic for $H_0 : \pi_1 - \pi_2 = \theta_0$ of the form

$$Z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}.$$

When $\theta_0 = 0$, this becomes

$$Z = \frac{b - c}{\sqrt{b + c}}.$$

In large samples, $Z \sim N(0, 1)$ and confidence intervals and tests can be constructed as usual.

For these hypothetical data, the test statistic becomes

$$Z_{obs} = \frac{21 - 4}{\sqrt{21 + 4}} = 3.4$$

which differs significantly from 0, indicating that Bush won the debate.

χ^2 test for independence

The χ^2 test for independence can be used to detect independence among two categorical variables.

Example: A random sample of $n_{++} = 2237$ adults was conducted and there gender and handedness were observed and are tabulated below:

Handedness	Men	Women	Total
Right	934	1070	2004
Left	113	92	205
Ambidextrous	20	8	28
Total	1067	1170	2237

Define a RV O_{ij} to model the observed counts for the cell in the i^{th} row and j^{th} column. Note the notational difference between rows and columns. Let the expected value for these RVs be denoted by E_{ij} respectively.

	Observed		Expected		
Handedness	Men	Women	Men	Women	Totals
Right	O_{11}	O_{12}	E_{11}	E_{12}	n_{1+}
Left	O_{21}	O_{22}	E_{21}	E_{22}	n_{2+}
Ambidextrous	O_{31}	O_{32}	E_{31}	E_{32}	n_{3+}
Totals	n_{+1}	n_{+2}			n_{++}

Under the (null) hypotheses that handedness and gender are independent,

$$\Pr(\text{Left-handed} \cap \text{man}) = \Pr(\text{Left-handed}) \times \Pr(\text{man})$$

and so on for each gender and each category of handedness. So, an estimate for the number of left-handed men in the sample under this hypothesis is just the fraction of left-handers times the fraction of men times the sample size, or for general cell (i, j):

$$E_{ij} = n_{++} \times \frac{n_{i+}}{n_{++}} \times \frac{n_{+j}}{n_{++}} = n_{++} \times n_{i+} \times n_{+j}$$

	Observed		Expected	
Handedness	Men	Women	Men	Women
Right	934	1070	955.9	1048.1
Left	113	92	97.8	107.2
Ambidextrous	20	8	13.4	14.6

Then the χ^2 test for independence in an $I \times J$ contingency table is based upon the test stastistic

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

which has a χ^2 distribution with degrees of freedom given by $(I - 1) \times (J - 1)$ under the null hypothesis of independence. In our example,

$$\chi^2 = \left[\frac{(934 - 955.9)^2}{955.9} + \frac{(1070 - 1048.1)^2}{1048.1} + \cdots + \frac{(8 - 14.6)^2}{14.6} \right] \approx 12$$

The critical value for this test statistic is $\chi^2(2, 0.05) = 5.99$.

H_0 and H_1?

Conclusion: ?

p-value: ?