Lecture 2

Review of Basic Statistical Procedures

- Continuous response variables
- Categorical response variables

Basic Statistical Inference

Let $X_1, \cdots, X_n$ be a random sample with pdf $f(\theta)$, where either f or θ or both may be unknown.

Inference:
- Estimation
- Hypothesis Testing

Approaches:
- Frequentist
- Bayesian

Supervised Models
- Regression
- Decision Trees
- Neural Networks

Unsupervised Models
- Summarization
- Clustering
- Association (?)

Assumptions of Classical Statistical Inference

- Independence
- Normality (or other specified distribution)

Validation of Assumptions:

- Histograms, qqnorm, etc., for normality
- Box-plots, stem-and-leaf diagrams for outlier detection
- Simple time series plot for serial correlations: including trends and cycles.

When distribution of data skewed, confidence intervals tend to be large, on the average, and p-values may be inflated.

Nonparametric methods do not make explicit assumptions about underlying distributions.

Most commonly used nonparametric methods are based on rank transformations of the observations.

Example: Cedar-apple rust data:

- Disease that affects apple trees, with symptom, rust-colored spots on apple leaves. Study conducted to study cause.
- In the first year of experiment the number of affected leaves on a random sample of 8 trees was counted on Variety 1.
- Normally, the average number of affected leaves is at most 25.

Data:

Tree	Count
1	38
2	10
3	84
4	36
5	50
6	35
7	73
8	48

Does the data suggest the apple trees are affected?

Standard Tests for Location

One-Sample Problem

Let $X_1, \cdots, X_n$ be a random sample from a distribution with meand μ and variance σ^2. We wish to test the hypothesis

$$H_0 : \mu = \mu_o$$

vs.

$$H_1 : \mu > \mu_o$$

When the distribution is normal, and σ^2 unknown (the usual case), a procedure that is commonly used is the Student's t-test, given by

$$T = \frac{\sqrt{n}(\bar{X} - \mu_o)}{S}$$

which under the null has a t_{n-1} distribution.

Example: Cedar-apple rust data:

```
> variety1
[1] 38 10 84 36 50 35 73 48

> t.test(variety1,mu=25)

        One Sample t-test

data:  variety1
t = 2.651, df = 7, p-value = 0.03289

alternative hypothesis: true mean is not equal to 25

95 percent confidence interval:
 27.34963 66.15037
sample estimates:
mean of x:    46.75
```

If EDA shows that the data is non-normal, then the *Wilcoxon signed-rank* test may be used.

- Let $D_i = X_i - \mu_o$
- Rank $\mid D_1 \mid, \cdots, \mid D_n \mid$

- Let T_+ and T_- be the sums of the ranks assigned to the positive and negative $d_i's$, respectively. Clearly, if H_0 is true, $T_+ \approx T_-$.

Further, under H_o,

$$E[T_+] = \frac{n(n+1)}{4}$$

and

$$Var(T_+) = \frac{n(n+1)(2n+1)}{24}$$

- The Wilcoxon signed-rank test rejects for large values of T_+, or when

$$Z = \frac{T_+ - E[T_+ \mid H_o]}{\sqrt{Var(T_+ \mid H_o)}}$$

exceeds a standard normal critical point.

- The test requires the underlying distribution to be symmetric.

Example: Cedar-apple rust data (cont'd):

- Disease that affects apple trees, with symptom, rust-colored spots on apple leaves. Study conducted to study cause.
- In the first year of experiment the number of affected leaves on a random sample of 8 trees from Variety 1 and a random sample of 7 trees from Variety 2 counted.

Data:

Variety 1		Variety 2	
Tree	Count	Tree	Count
1	38	1	27
2	10	2	28
3	84	3	57
4	36	4	66
5	50	5	77
6	35	6	49
7	73	7	62
8	48		

Does the data suggest there is a difference between the two varieties in the mean number of affected leaves?

Two-Sample Problem

Let $X_1, \cdots X_n$, and $Y_1, \cdots Y_m$, be independent random samples from distributions with respective means, μ_1 and μ_2, and variances, σ_1^2 and σ_2^2.

We are interested in the hypothesis

$$H_o : \mu_1 - \mu_2 = \Delta_o$$

vs

$$H_1 : \mu_1 - \mu_2 > \Delta_o$$

When both distributions are normal, with unkown, but equal, variance σ^2, the two-sample t-test is given by

$$T = \frac{(\bar{X} - \bar{Y}) - \Delta_0)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}}$$

where the pooled variance S_p^2 is defined as

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

It is easy to show that T has a t_{n+m-2} distribution.

Example: Cedar-apple rust data (cont'd):

```
> variety2  <- c(27,28,57,66,77,49,62)
```

```
> t.test(variety1,variety2,var.equal=T)

        Two Sample t-test
data:  variety1 and variety2
t = -0.501, df = 13, p-value = 0.6247
alternative hypothesis:
       true difference in means is not equal to 0
95 percent confidence interval:
 -29.40555  18.33412

sample estimates:
mean of x mean of y
 46.75000  52.28571
```

Example: Cedar-apple rust data (cont'd):

```
> var.test(variety1,variety2)

        F test to compare two variances

data:  variety1 and variety2
F = 1.499, num df = 7, denom df = 6,
p-value = 0.638
alternative hypothesis:
         true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2631926 7.6728062
sample estimates:
ratio of variances
          1.499006
```

When the variances are not equal, an approximate test is *Welch's modified two-sample t test*, given by

$$T = \frac{(\bar{X} - \bar{Y}) - \Delta_0)}{S_w}$$

where

$$S_w^2 = \frac{S_1^2}{n} + \frac{S_2^2}{m}$$

T has an approximate t_ν distribution, with

$$\nu \approx \left[\frac{c^2}{n-1} + \frac{(1-c)^2}{m-1} \right]^{-1}$$

where

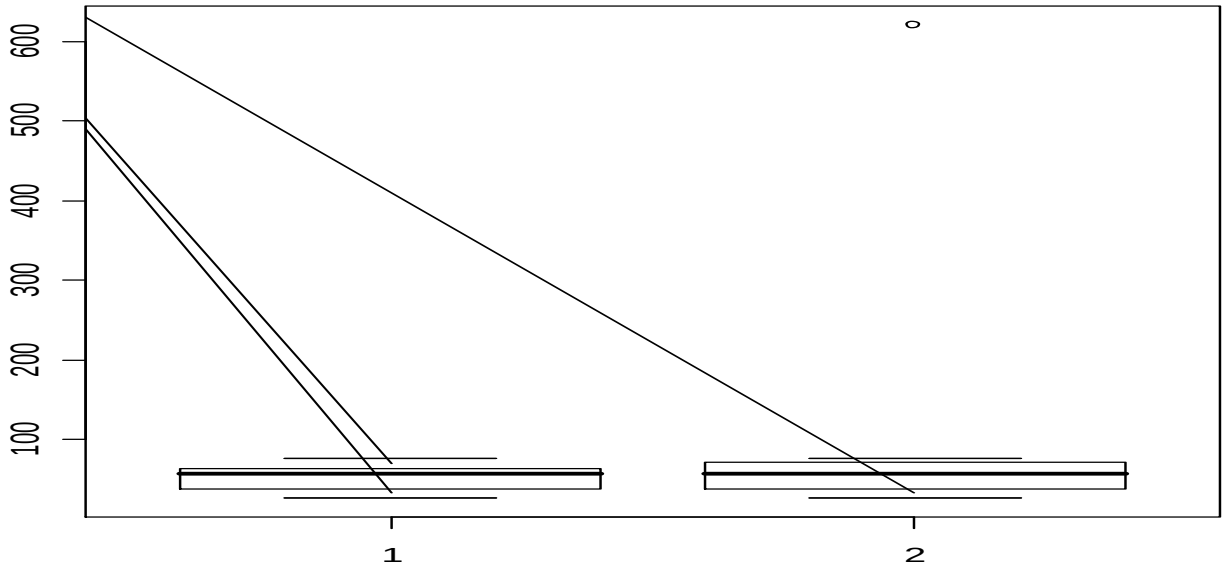
$$c = \frac{S_1^2}{nS_w^2}$$

Example: Cedar-apple rust data (cont'd):

```
> t.test(variety1,variety2)

        Welch Two Sample t-test

data:  variety1 and variety2
t = -0.5082, df = 12.956,
 p-value = 0.6198
alternative hypothesis: true difference in means is not equal
    to 0
95 percent confidence interval:
 -29.07417  18.00275
sample estimates:
mean of x mean of y
 46.75000  52.28571
```

Example: Cedar-apple rust data (cont'd):

```
> variety2  <- c(27,28,57,66,77,49,62)

>variety3  <- c(27,28,57,66,77,49,6200)

> boxplot(variety2,variety3)
```



Example: Cedar-apple rust data (cont'd):

```
t.test(variety1,variety2,var.equal=T)

        Two Sample t-test
data:  variety1 and variety2
t = -0.501, df = 13,
p-value = 0.6247
alternative hypothesis:
       true difference in means is not
       equal to 0
95 percent confidence interval:
 -29.40555  18.33412

sample estimates:
mean of x mean of y
 46.75000  52.28571
```

```
> variety3  <- c(27,28,57,66,77,49,620)

> t.test(variety1,variety3,var.equal=T)

        Two Sample t-test

data:  variety1 and variety3
t = -1.1151, df = 13, p-value = 0.285
alternative hypothesis: true difference in
     means is not equal to 0
95 percent confidence interval:
 -250.4077   79.9077
sample estimates:
mean of x mean of y
    46.75    132.00
```

Wilcoxon Rank-Sum Test

- Combine the two samples, and rank the observations
- Let T_1 be the sum of the ranks for the observations in the 1st group, and T_2 for the second
- The Wilcoxon rank-sum test is then based on T_W, which has an asymptotic standard normal distribution,

$$T_W = \frac{T_1 - E(T_1 \mid H_o)}{\sqrt{Var(T_1 \mid H_o)}}$$

where

$$E[T_1 \mid H_o] = \frac{n(n+m+1)}{2}$$

and

$$Var(T_1 \mid H_o) = \frac{nm(n+m+1)}{12}$$

In some applications, the Mann-Whitney (U) form of the Wilcoxon rank-sum test is used, due to the relative ease of computing critical values using the latter. The two are related by the equation

$$U_1 = T_1 - \frac{n(n+1)}{2}$$

- The test is most appropriate when the populations have the same shape and differ only in location (e.g., same dispersion) Howeve, the distributions do not have to be symmetric.

Example: Cedar-apple rust data (cont'd):

```
> wilcox.test(variety1,variety2)

        Wilcoxon rank sum test

data:  variety1 and variety2
W = 24, p-value = 0.6943
alternative hypothesis: true mu
is not equal to 0
```

```
> variety3  <-
   c(27,28,57,66,77,49,620)

> wilcox.test(variety1,variety3)

        Wilcoxon rank sum test

data:  variety1 and variety3
W = 22, p-value = 0.5358
alternative hypothesis: true mu
    is not equal to 0
```

Example: Cedar-apple rust data (cont'd):

- Disease that affects apple trees, with symptom, rust-colored spots on apple leaves. Study conducted to study cause.
- In the first year of experiment the number of affected leaves on a random sample of 8 trees from Variety 1
- All red cedar trees within 100 yards of the orchard were then removed and the following year the same Variety 1 trees were examined for affected leaves.

Data:

	Year 1	Year 2
Tree	Count	Count
1	38	32
2	10	16
3	84	57
4	36	28
5	50	55
6	35	12
7	73	61
8	48	29

Does the data suggest there is a difference between the mean number of affected leaves in the two periods?

Example: Cedar-apple rust data (cont'd):

Apply 2-sample t-test?

```
> t.test(year1,year2,var.equal=T)

        Two Sample t-test

data:  variety1 and year2
t = 0.9899, df = 14, p-value = 0.3390
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.25070  33.25070
sample estimates:
mean of x mean of y
    46.75     36.25
```

Paired Samples

Consider a study comparing weight gains in rats before and after receiving a new dietary regime. Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be the respective weights of n rats. Clearly, there is dependence between obseravtions on the same rat, hence, the two-sample procedures discussed above will not be appropriate.

Let $D_i = X_i - Y_i,\ i = 1, \cdots, n.$ Then the problem may be handled based on the D_i's using the one-sample t-test (paired t-test) or the Wilcoxon signed-rank test.

Example: Cedar-apple rust data (cont'd):

```
> t.test(year1,year2,paired=T)

        Paired t-test

data:  variety1 and year2
t = 2.4342, df = 7,
p-value = 0.04514
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
  0.2999574 20.7000426
sample estimates:
mean of the differences
                  10.5
```

```
> wilcox.test(year1,year2,paired=T)

        Wilcoxon signed rank test with
continuity correction

data:  year1 and year2
V = 32.5, p-value = 0.04967
alternative hypothesis: true mu is not
equal to 0
```

Measures of Associations

Given a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ on n individuals, we are often interested in the degree of association or relationship between X and Y.

1. *Pearson's product-moment correlation coefficient*

The population correlation coefficient is defined as

$$\rho = \frac{E(X - \mu_x)(Y - \mu_y)}{\sigma_x \sigma_y}$$

and has the property $-1 \leq \rho \leq 1$.

The corresponding sample quantity is given by

$$r = \frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1) S_x S_y}$$

with similar properties as ρ.

For the problem of testing $H_0 : \rho = 0$ vs. $H_1 : \rho > 0$, a suitable test statistic is given by

$$T = \sqrt{\frac{n-2}{1-r^2}}$$

which has a t_{n-2} null distribution.

The test of the more general hypothesis $H_0 : \rho = \rho_0$ vs. $H_1 : \rho > \rho_0$ involves Fisher's Z-transformation. Let

$$Q(r) = \frac{1}{2} ln \left(\frac{1+r}{1-r} \right)$$

Then a test statistic is given by

$$Z = \frac{Q(r) - Q(\rho)}{\sqrt{\frac{1}{n-3}}}$$

which has an apprximate standard normal null distribution.

For purposes of confidence interval construction, one often uses the following inverse hyperbolic tangent transformation. Let $H(r) = atanh(r)$. Then $H(r) - H(\rho)$ has an approximate $N(0, 1/(n-3))$ distribution. A

$100(1-\alpha)\%$ confidence interval for ρ is

$$tanh\left[(H(r) \pm Z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right].$$

Remarks

- The product moment correlation coefficient may not be appropriate for ordinal data
- The sample correlation coefficient r is extremely sensitive to outliers

2. *Spearman's rank correlation coefficient*

Spearman's rho is a measure of association based on ranks.

Given $(X_1, Y_1), \cdots, (X_n, Y_n)$, let d_i denote the difference between the ranks of X_i and Y_i, when the X's and Y's are ranked separately among themselves.

For tied observation, the mean of the ranks are taken.

When there are no ties, Spearman's rho r_s coincides with the Pearson product-moment correlation coefficient applied to the ranks, and is given by

$$r_s = 1 - \frac{6\,\Sigma_i\, d_i^2}{n(n^2 - 1)}$$

3. *Kendall's tau*

Kendall's tau is another rank-based measure of association. Let R_i and S_i denote the ranks of X_i and Y_i, respectively. Then

$$\tau = \frac{1}{\binom{n}{2}} \sum_{i<j} sgn(R_i - R_j)(S_i - S_j)$$

where

$$sgn(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

```
> cor.test(variety1,year2)

        Pearson's product-moment correlation

data:  variety1 and year2
t = 3.973, df = 6,
      p-value = 0.007342
alternative hypothesis:
  true correlation is not equal to 0
95 percent confidence interval:
 0.3662168 0.9725356

sample estimates:
     cor
0.851221
```

```
> cor.test(variety1,year2,method="sp")

        Spearman's rank correlation rho

data:  variety1 and year2

p-value = 0.002232
alternative hypothesis:
         true rho is not equal to 0

sample estimates:
      rho
0.9285714
```

The Kolmogorov-Smirnov Test

Let $X_1, \cdots, X_n$ be iid F. Then the one-sample K-S test is given by

$$T_{KS} = sup_x \mid F_n(x) - F(x) \mid$$

In the case of two samples, let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ be independent random samples from F and G, respectively. Then the K-S test for comparing the two distributions is given by:

$$T_{KS} = sup_x \mid F_n(x) - G_m(x) \mid$$

In *R*, the procedures are implemented using *ks.gof()*.

Bootstrap Tests

The bootstrap may be used to construct approximate confidence intervals and test critical points, when the associated distributions are not tractable.

Consider the problem of comparing two means μ_1 and μ_2, when sampling $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ from F and G, respectively.

To implement the bootstrap, one draws pseudo-random samples from F_n and G_m under H_0 : $\mu_1 = \mu_2$.

- Draw with replacement, B pseudo-samples: $X_1^*, \cdots, X_n^*$ from F_n and and $Y_1^*, \cdots, Y_m^*$ from $\tilde{G}_m$, where $\tilde{G}_m$ is the edf of $\tilde{Y}_1, \cdots, \tilde{Y}_m$, and

$$\tilde{Y}_i = Y_i + \bar{X} - \bar{Y}$$

For the b'th pseudo-sample, compute

$$Z_b^* = \frac{(\bar{X}^* - \bar{Y}^*)}{\sqrt{\frac{S_1^*}{n} + \frac{S_2^*}{m}}}$$

- Calculate the p-value as

$$p_B = \frac{\Sigma_b^B \, I[\mid Z_b^* \mid \geq z_{obs}]}{B}$$

where z_{obs} is the observed value of the test statistic based on the original sample.

Example. Suppose the manufacturer of a certain product claimed that less than 15%, the industry standard, of the items manufactured by his factory were defective.

To test whether his claim was true, a random sample of 100 items was taken, of which 13 turned out to be defective.

Test the relevant hypothesis.

Let p denote the probability of success

The hypothesis of interest is

$$H_0 : \ p = 0.15$$

vs

$$H_1 : \ p < 0.15$$

A reasonable test is one which rejects H_o for small values of X.

An *exact test* is obtained computing the p-value as $Pr[X \leq 13 \mid p = 0.15]$, or

$$= \sum_{j=1}^{13} \binom{100}{j} 0.15^j 0.85^{n-j}$$

```
binom.test(13, 100, 0.15, alt = "l")
        Exact binomial test

data:  13 out of 100
number of successes = 13, n = 100,
p-value = 0.3474
```

Let X be the number of successes in n trials.

For large n, such that $min(np, nq) \geq 5$,

$$Z = \frac{x - np - 0.5}{\sqrt{npq}}$$

approximately Standard Normal.

```
> prop.test(13,100,0.15,alt="l")
1-sample prop test with continuity correction

X-square = 0.1765, df = 1, p-value = 0.3372

95 percent confidence interval:
0.0000000 0.2009056
```

Denote $\hat{p} = X/n$.

An approximate $100(1-\alpha)\%$ confidence interval for p may be constructed based on the pivot

$$\frac{|\hat{p} - p| - \frac{1}{2n}}{\sqrt{pq/n}} \leq Z_{\alpha/2}$$

which gives (P_L, P_U), where

$$P_L = \frac{(2n\hat{p} + Z^2_{\alpha/2} - 1) - Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} - (2 + 1/n) + 4\hat{p}(n\hat{q} + 1)}}{2(n + Z^2_{\alpha/2})}$$

and

$$P_U = \frac{(2n\hat{p} + Z^2_{\alpha/2} + 1) + Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} + (2 + 1/n) + 4\hat{p}(n\hat{q} - 1)}}{2(n + Z^2_{\alpha/2})}$$

Example. Suppose the manufacturer of a certain product claimed that the percentage of defective items manufactured by his factory was less than that for the competitor.

To test whether his claim was true, random samples of 131 and 281 items were taken (from each company) of which 161 and 271, respectively, turned out to be non-defective.

Test the relevant hypothesis.

Let p_1 and p_2 denote the respective proportions of defective items.

The hypotheses of interest are

$$H_o: \ p_1 = p_2$$

vs.

$$H_1 : p_1 \neq p_2$$

A large sample test, with Yates' correction for continuity, is given by

$$Z = \frac{| \hat{p}_1 - \hat{p}_2 | - \frac{1}{2}(\frac{1}{n} + \frac{1}{m})}{\sqrt{\hat{p}_c \hat{q}_c (\frac{1}{n} + \frac{1}{m})}}$$

$$\hat{p}_c = \frac{n\hat{p}_1 + m\hat{p}_2}{n+m}$$

An approximate $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm \left(Z_{\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 / n + \hat{p}_2 \hat{q}_2 / m} + \frac{1}{2} \left(\frac{1}{n} + \frac{1}{m} \right) \right)$$

Example (cont'd):

2-sample test for equality of proportions with continuity correction

```
> x <- c(161,131)
> n <- c(271,281)

> prop.test(x,n)

X-square = 8.5518, df = 1, p-value = 0.0035

alternative hypothesis: two.sided

95 percent confidence interval:
0.04169418 0.21411336
```

Example: Suppose random samples of sizes n=100 and m=150 gave x=5 and y=7, respectively.

```
X<-c(5,7); n <- c(100,150)
prop.test(x, n)
```

```
X-square = 0.0146, df = 1, p-value = 0.9039
```

```
Warning messages:
 Expected counts < 5. Chi-square/normal
approximation may not be
appropriate. in: prop.test(x, n)
```

Denote $\hat{p} = X/n$.

An approximate $100(1-\alpha)\%$ confidence interval for p may be constructed based on the pivot

$$\frac{|\hat{p} - p| - \frac{1}{2n}}{\sqrt{pq/n}} \leq Z_{\alpha/2}$$

which gives (P_L, P_U), where

$$P_L = \frac{(2n\hat{p} + Z^2_{\alpha/2} - 1) - Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} - (2 + 1/n) + 4\hat{p}(n\hat{q} + 1)}}{2(n + Z^2_{\alpha/2})}$$

and

$$P_U = \frac{(2n\hat{p} + Z^2_{\alpha/2} + 1) + Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} + (2 + 1/n) + 4\hat{p}(n\hat{q} - 1)}}{2(n + Z^2_{\alpha/2})}$$

Example. Suppose the manufacturer of a certain product claimed that the percentage of defective items manufactured by his factory was less than that for the competitor.

To test whether his claim was true, random samples of 131 and 281 items were taken (from each company) of which 161 and 271, respectively, turned out to be non-defective.

Test the relevant hypothesis.

Let p_1 and p_2 denote the respective proportions of defective items.

The hypotheses of interest are

$$H_o: \ p_1 = p_2$$

vs.

$$H_1 : p_1 \neq p_2$$

A large sample test, with Yates' correction for continuity, is given by

$$Z = \frac{| \hat{p}_1 - \hat{p}_2 | - \frac{1}{2}(\frac{1}{n} + \frac{1}{m})}{\sqrt{\hat{p}_c \hat{q}_c (\frac{1}{n} + \frac{1}{m})}}$$

$$\hat{p}_c = \frac{n\hat{p}_1 + m\hat{p}_2}{n+m}$$

An approximate $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm \left(Z_{\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 / n + \hat{p}_2 \hat{q}_2 / m} + \frac{1}{2} \left(\frac{1}{n} + \frac{1}{m} \right) \right)$$

Example (cont'd):

2-sample test for equality of proportions with continuity correction

```
> x <- c(161,131)
> n <- c(271,281)

> prop.test(x,n)

X-square = 8.5518, df = 1, p-value = 0.0035

alternative hypothesis: two.sided

95 percent confidence interval:
0.04169418 0.21411336
```

Example: Suppose random samples of sizes n=100 and m=150 gave x=5 and y=7, respectively.

```
X<-c(5,7); n <- c(100,150)
prop.test(x, n)
```

```
X-square = 0.0146, df = 1, p-value = 0.9039
```

```
Warning messages:
 Expected counts < 5. Chi-square/normal
approximation may not be
appropriate. in: prop.test(x, n)
```

Fisher's Exact Test

Suppose an experiment on the effect of a certain chemical on the mood of subjects (e.g., Depressed/Not Depressed) gave the following data in 3 males and 4 females:

	Depressed	Not Depressed
Male	1	2
Female	3	1

Consider the 2×2 table

	Depressed	Not Depressed	
Male	n_{11}	n_{12}	n_{1+}
Female	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

Fix the marginal totals and compute the probably of observing the given cell frequencies:

$$p_o = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}}$$

Next compute p^* the probabilities for all tables having the same marginal totals. The p-value is computed as the

$$p = \sum_{p^*:p^* \leq p_o} p*$$

	Depressed	Not Depressed
Male	1	2
Female	3	1

```
depress.data <- matrix(c(1, 2, 3, 1), 2, 2)
fisher.test(depress.data)
```

p-value = 0.4857

Example. Suppose in a survey of public opinion about a certain political issue, a random sample of registered voters taken, n fixed. Sample included voters from each political group: Democrat, Republican, and Other, giving the data displayed below.

	Favor	Do Not Favor	Total
Democrat	198	202	$n_{1+} = 400$
Republican	140	210	$n_{2+} = 350$
Other	133	217	$n_{3+} = 350$
Totals	$n_{+1} = 471$	$n_{+2} = 629$	$n = 1100$

Hypothesis of interest:

H_o: No association between party Affiliation and Opinion

vs

H_1: There is association

Let p_{ij} denote the probability corresponding to the ij'th cell.
Then under H_0, $p_{ij} = p_{i+}p_{+j}$, and is estimated by

$$\hat{p}_{ij} = \frac{n_{i+}}{n}\frac{n_{+j}}{n}$$

The corresponding expected number is given by

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

A reasonable test is given by

$$X^2 = \Sigma \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

and has an approximate $\chi^2_{(I-1)(J-1)}$ distribution

The test is commonly referred to as Pearson's chi-square test.

For 2×2 tables, Yates' correction for continuity may be applied.

	Favor	Do Not Favor	Total
Democrat	198	202	$n_{1+} = 400$
Republican	140	210	$n_{2+} = 350$
Other	133	217	$n_{3+} = 350$
Totals	$n_{+1} = 471$	$n_{+2} = 629$	$n = 1100$

```
opinion.data <- matrix(c(198, 140, 133, 202,
210, 217), 3, 2, byrow = F)

chisq.test(opinion.data)
Pearson's chi-square test without Yates'
continuity correction

X-square = 11.7478, df = 2, p-value = 0.0028
```

Example. Consider the following artificial data on the relationship between lung cancer and passive smoking.

	Passive	Not Passive
Cancer	281	235
No Cancer	210	279

Application of the Pearson chi-square test gives a p-value $= 0.0003$.

The effect of passive smoking on cancer is *confounded* with the smoking status of the individual.

	Smoker		Non-smoker	
	Passive	Not Passive	Passive	Not Passive
Cancer	261	118	20	117
No Cancer	130	124	80	155

Application of a chi-square test to each table is not a viable option.

First, it does not draw strength from the combined data, and hence may be less sensitive. The overall level of significance may be inflated, if each test is performed at the usual $\alpha = 0.05$.

Given K 2×2 tables, let n_{ijk} be the number of events in the ij'th cell of the k'th table.

The Mantel-Haenszel test is given by

$$X^2_{MH} = \frac{[\mid \Sigma_k(n_{11k} - E_{11k}) \mid -c]^2}{V_{11k}}$$

$$E_{11k} = E[N_{11k} \mid H_0],$$

$$E_{11k} = \frac{n_{1+k}n_{+1k}}{n_{++k}}$$

$$V_{11k} = \sum_k \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n^2_{++k}(n_{++k} - 1)}$$

For large sample, X^2_{MH} has an approximate χ^2_1 distribution.

	Smoker		Non-smoker	
	Passive	Not Passive	Passive	Not Passive
Cancer	261	118	20	117
No Cancer	130	124	80	155

```
passive.smoker.data <- array(c(261, 130, 118, 124,
20, 80, 117, 155), c(2, 2, 2))

mantelhaen.test(passive.smoker.data)
```

```
Mantel-Haenszel chi-square = 1.7258, df = 1,
p-value = 0.189
```

Remarks

- The approximation is reliable for large n. Even if the n_{ijk} are small, the marginal totals should be large.

- When the true association is similar in each cell, the test is more powerful than separate tests in each table.

- Interpretation of results when results are not consistent across tables

NB: p-value still valid, even when tables not homogenous.

Tests for homogeneity across tables.

- The Breslow-Day test, performs reliably when the sample size is large.
- For small samples, an alternative test may be Zelen's procedure.

Extensions to the case of $I \times J \times K$ tables, and when the rows and/or columns are ordinal.

1. *When Both Rows and Columns are Nominal*

   The generalized Cochran-Mantel-Haenszel test for general association concerns the hypothesis

   $$H_o : \textit{No association between } X \textit{ and } Y$$

   and the test statistic has an approximate $\chi^2_{(I-1)(J-1)}$ distribution.

2. *When the Row Variable is Nominal and Y Ordinal*

The hypothesis of interest is

Ho : No Diference among row mean scores

and the test statistic has an approximate χ^2_{I-1} distribution.

3. *When Both Row and Column Variables are Ordinal*

The hypothesis of zero correlation is based on

$$M^2 = (n-1)r^2$$

where r is the correlation coefficient between the scores of the row and column. The test statistic has an approximate χ^2_1 distribution.

4. *When X is Ordinal and the Column Variable is Nominal*

   When J=2 and K=1, this reduces to Cochran-Armitage test for trend.

The SAS procedure PROC FREQ implements most of the above situations.

Example. Suppose two eye treatments, A and B, are to be compared with respect to a binary outcome (cure/failure). One hundred eligible subjects, and one eye from each pair randomly assigned to either A or B. The results are given below.

	Cured	Not Cured
A	48	52
B	30	70

Let p_A and p_B be the proportions of cures for treatments A and B, respectively. The null hypothesis of interest is:

$$H_o : p_A = p_B$$

```
> prop.test(c(48,30),c(100,100))

        2-sample test for equality of proportions with
continuity correction

data:  c(48, 30) out of c(100, 100)
X-squared = 6.074, df = 1, p-value = 0.01372
alternative hypothesis: two.sided
95 percent confidence interval:
 0.03712658 0.32287342
sample estimates:
prop 1 prop 2
  0.48   0.30
```

Matched Samples

Due to dependence within each pair, the usual Pearson chi-square test is not applicable to the above table.

Instead, we need to consider the following table for the matched pairs:

		B	
		Cured	Not Cured
A	Cured	8	40
	Not Cured	22	30

To fix ideas, consider the following table:

		B	
		Cured	Not Cured
A	Cured	a	b
	Not Cured	c	d

The corresponding estimators for p_A and p_B are

$$\hat{p}_A = \frac{a+b}{n}$$

and

$$\hat{p}_B = \frac{a+c}{n}$$

Then

$$\hat{p}_A - \hat{p}_B = \frac{b-c}{n}$$

Under H_o, $b \approx c$, so that
$b \sim$ binomial (b+c, $\frac{1}{2}$).

Hence, a test statistic with continuity correction is

$$X^2_{McN} = \frac{[\,|\,b - c\,|\,-1]^2}{b + c}$$

which has an approximate χ^2_1 distribution under H_o.

The test is known as McNemar's test.

		B	
		Cured	Not Cured
A	Cured	8	40
	Not Cured	22	30

```
paired.data <- cbind(c(8, 22), c(40, 30))
mcnemar.test(paired.data)

McNemar's chi-square test with
continuity correction

data:  paired.data
McNemar's chi-square = 4.6613, df = 1,
 p-value = 0.0308
```

Measuring Degree of Association

Consider the following two hypothetical cases.

In one study $p_{11} = 0.01$ and $p_{21} = 0.001$, giving a difference $\Delta = 0.009$.

In the second case, assume $p_{11} = 0.410$ and $p_{21} = 0.401$, giving the same difference $\Delta = 0.009$

However, the relative value $\frac{p_{11}}{p_{21}}$ for the first case is 10, while it is approximately 1 in the second case.

Relative Risk

$$RR = \frac{\hat{p}_{11}}{\hat{p}_{21}}$$

95% Confidence Interval for true RR:

$$ln\left(\frac{\hat{p}_{11}}{\hat{p}_{21}}\right) \pm Z_{\alpha/2}\sqrt{\frac{1-\hat{p}_{11}}{n_1+\hat{p}_{11}} + \frac{(1-\hat{p}_{21}}{n_2+\hat{p}_{21}}}$$

Odds Ratio

Recall that $\frac{p_{11}}{1-p_{11}}$ and $\frac{p_{21}}{1-p_{21}}$ correspond to the odds of having the disease, for X=1 and X=0, respectively. Hence the odds ratio of having the disease for X=1 relative to X=0 is given by

$$\psi = \frac{p_{11}}{1-p_{11}} \div \frac{p_{21}}{1-p_{21}}$$

An estimator of ψ is

$$\hat{\psi} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

A large sample confidence interval for $ln(\psi)$ is

$$ln(\hat{\psi}) \pm Z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

The Breslow-Day test for homogeneity of the odds ratios:

$$T_{BD} = \sum_k \frac{(n_{11k} - \hat{\mu}_{11k})^2}{\hat{\mu}_{11k}}$$

which under H_o has a χ^2_{K-1} approximate distribution. The approximation is reliable provided $\hat{\mu}_{ijk} > 5$ for at least 80% of the cells.

using the Woolf test for interaction:

```
woolf <- function(x) {
  x <- x + 1 / 2
  k <- dim(x)[3]
  or <- apply(x, 3, function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
  w <-  apply(x, 3, function(x) 1 / sum(1 / x))
  1 - pchisq(sum(w * (log(or) - weighted.mean(log(or), w)) ^ 2), k - 1)
}
woolf(UCBAdmissions)
## => p = 0.003, indicating that there is significant heterogeneity.
## (And hence the Mantel-Haenszel test cannot be used.)
```

Problem Set 2

Reading Assignment:

Chapter 2,3,4. The Statistical Sleuth: A Course in Methods of Data Analysis. Ramsey & Schafer

Consider the data on **Chicken Weights by Feed Type (**chickwts**) in** R library MASS

1. Determine whether there is a significant difference in the mean weights of chicks fed soybean vs. those fed casein using each of the following procedures:
   a) A parametric procedure
   b) A non-parametric procedure
   c) A re-sampling procedure

   Discuss the assumption underlying each of the analyses, their validity, and any remedial measures to be taken
2. Using the data for chicks fed casein and those on sunflower, compute the following, based on a suitable bootstrap method:
   a) A 95% confidence interval for the difference in median weight for the two groups
   b) A 95% bootstrap confidence interval for the ratio of the variances of soybean fed to sunflower fed chicks
   c) A 95% confidence interval for the ratio of the variances of soybean fed to sunflower fed chicks

   Discuss the assumption underlying each of the analyses, their validity, and any remedial measures to be taken
3. Assume that if the weight of a chick is below 256, that chick is classified under “LOW WEIGHT”. For chicks fed meatmeal vs. those fed soybean,
   a) Determine whether there is a significant difference in the proportions of the chicks classified under “LOW WEIGHT”.
   b) Construct a 95% confidence interval for the difference in the proportions of the chicks classified under “LOW WEIGHT”.

   Discuss the assumption underlying each of the analyses, their validity, and any remedial measures to be taken