

START CN Wednesday 11/17/21
HANDOUT #13: HYPOTHESES TESTING FOR MULTIPLE POPULATIONS

Tests Comparing K Population/Process Parameters

I. Tests about Differences in Population Means/Location Parameter

1. Normal distribution with $\sigma_1 = \sigma_2$:
 - Pooled t-Test
 - Robust or pooled t-Test
2. Normal distributions with $\sigma_1 \neq \sigma_2$:
 - Separate Variance t-Test
3. Nonnormal distributions with small/moderate n_i :
 - Wilcoxon Rank Sum Test
4. Paired Data:
 - Paired t-Test for Normally Distributed Differences
 - Wilcoxon Signed Rank Test for NonNormally Distributed Differences

II. Tests about Differences in Population Proportions

- Chi-squared test of Homogeneity of Proportions
- Chi-squared test of Independence
- Fisher's Exact Test (permutation test)
- Odds Ratio and Relative Risk
- Sensitivity and Specificity
- McNemar's Test for Matched Pairs
- Sets of Tables - Cochran-Mantel-Haenszel Tests
- Simpson's Paradox

III. Tests about Differences in Population Standard Deviations

- Sampling from Normal Populations - Hartley's Test
- General Procedure - Brown-Forsythe-Levene's Test

IV. Tests for Correlation in the Data

- For an AR(1) model with normal errors - von Neumann Test
- Distribution-free test - Runs Test

Supplemental Reading:

- Chapters 8, Chapter 9, and Section 14.2 in Tamhane/Dunlop book

Tests about Differences in Population Location Parameters

Suppose we have two populations/processes and we want to test hypotheses about differences in their means.

Our parameter of interest is

$$\theta = \mu_1 - \mu_2 \quad \text{with} \quad \hat{\theta} = \bar{X} - \bar{Y},$$

where we have X_1, \dots, X_n iid from Population #1 and

Y_1, \dots, Y_m iid from Population #2.

In order to make inferences based on $\hat{\theta}$ it is necessary to determine the sampling distribution of $\hat{\theta}$.

$$E[\hat{\theta}] = E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_1 - \mu_2$$

$$Var[\hat{\theta}] = Var[\bar{X} - \bar{Y}] = Var[\bar{X}] + Var[\bar{Y}] - 2Cov(\bar{X}, \bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m} - 2Cov(\bar{X}, \bar{Y})$$

We will also need to know the form of the cdf for $\hat{\theta}$.

Case 1 We will impose the following conditions:

- C1. Both populations have a normal distribution
- C2. $\sigma_1 = \sigma_2 = \sigma$ with σ unknown
- C3. X s and Y s are independent.

Because X s and Y s are independent

$$Cov(\bar{X}, \bar{Y}) = 0 \Rightarrow Var[\hat{\theta}] = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2 \left(\frac{1}{n} + \frac{1}{m} \right).$$

The distribution of $\hat{\theta} = \bar{X} - \bar{Y}$ is $N(\mu_1 - \mu_2, \sigma^2 (\frac{1}{n} + \frac{1}{m}))$.

Because σ is unknown we need to estimate it. We have two independent unbiased estimates of σ^2 :

S_1^2 and S_2^2 , therefore, we pool the two estimates and obtain:

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

S_p^2 is a weighted average of the sample variances from the two populations.

If $n = m$, then $S_p^2 = \frac{1}{2}(S_1^2 + S_2^2)$, the average of the two sample variances.

We can now just use our results about testing a single population mean and apply the t-test to our problem.

To test

$$H_0 : \mu_1 \leq \mu_2 + \theta_o \quad \text{vs} \quad H_1 : \mu_1 > \mu_2 + \theta_o$$

just convert the above hypotheses to a hypothesis about $\theta = \mu_1 - \mu_2$, that is, test

$$H_0 : \mu_1 - \mu_2 \leq \theta_o \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 > \theta_o$$

That is, test

$$H_0 : \theta \leq \theta_o \quad \text{vs} \quad H_1 : \theta > \theta_o$$

Our test statistic would be

$$T = \frac{\hat{\theta} - \theta_o}{\sqrt{\frac{S_P^2}{n} + \frac{S_P^2}{m}}} = \frac{(\bar{X} - \bar{Y}) - \theta_o}{S_P \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

We next need to determine the sampling distribution of T :

$$T = \frac{(\bar{X} - \bar{Y}) - \theta_o}{\sqrt{\frac{S_P^2}{n} + \frac{S_P^2}{m}}} = \frac{\frac{(\bar{X} - \bar{Y}) - \theta}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} + \frac{(\theta - \theta_o)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}}{\sqrt{\frac{(n+m-2)S_P^2/\sigma^2}{n+m-2}}}$$

$$\stackrel{D}{=} \frac{N(0, 1) + \Delta}{\sqrt{\frac{\chi_{n+m-2}^2}{n+m-2}}} \quad \begin{matrix} \Delta \text{ is our} \\ \text{D.C.F.} \end{matrix}$$

Therefore,

- T has a NonCentral t-distribution with
- $df = n + m - 2$
- noncentrality parameter (ncp)

$$\Delta = \frac{\theta - \theta_o}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

When $\theta = \mu_1 - \mu_2 = \theta_o$,

- t has a Central t distribution \leftarrow For p-value assume $\theta = \theta_o$
- determine the critical values and p-values using the Central t -distribution cdf with $df = n+m-2$
- In R, $t_{\alpha, df} = qt(1 - \alpha, df)$ and $p\text{-value} = 1 - pt(t_{cal}, df)$

For the power function we need to use the NonCentral t -distribution cdf with $df = n + m - 2$ and noncentrality parameter

$$\Delta = \frac{\theta - \theta_o}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

We will next consider the forms of the rejection region, p-value, and power function.

Case 1. $\mathbf{H}_o : \mu_1 - \mu_2 \leq \theta_o$ versus $\mathbf{H}_1 : \mu_1 - \mu_2 > \theta_o$

Let $t_{\alpha, n+m-2}$ be the upper α percentile from a central t -distribution with $df = n + m - 2$, which can be obtained using the R-function: $\text{t}_{\alpha, n+m-2} = \text{qt}(1 - \alpha, n + m - 2)$

- Reject H_o if $T > t_{\alpha, n+m-2}$.
- The power function for $\theta = \mu_1 - \mu_2$ is given by

$$\gamma(\theta) = P[\text{Reject } H_o \text{ when } \mu_1 - \mu_2 = \theta] = P_\theta(T > t_{\alpha, n+m-2}) = 1 - G(t_{\alpha, n+m-2}),$$

where G is the cdf of a noncentral t -distribution with $df = n + m - 2$ and noncentrality parameter

$$\Delta = \frac{\theta - \theta_o}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

- Use the following R-function to calculate power

$$\gamma(\theta) = 1 - \text{pt}(\text{t}_{\alpha, n+m-2}, \Delta) = 1 - \text{pt}(\text{qt}(1 - \alpha, n + m - 2), n + m - 2, \Delta)$$

- p-value is as follows:

$$p-value = P_{\theta_o}(T > T_{obs}) = 1 - G_o(T_{obs})$$

where T_{obs} is the value calculated from the data

G_o is the cdf of a central t -distribution with $df = n + m - 2$,

- Use the following R-function to obtain the p-value:

$$p-value = 1 - \text{pt}(\mathbf{T}_{obs}, n + m - 2)$$

Case 2. $\mathbf{H}_o : \mu_1 - \mu_2 \geq \theta_o$ versus $\mathbf{H}_1 : \mu_1 - \mu_2 < \theta_o$

- Reject H_o if $T < -t_{\alpha, n+m-2}$.
- The power function for $\theta = \mu_1 - \mu_2$ is given by

$$\gamma(\theta) = P_\theta(T < -t_{\alpha, n+m-2}) = G(-t_{\alpha, n+m-2}),$$

where G is the cdf of a noncentral t -distribution with $df = n + m - 2$ and noncentrality parameter

$$\Delta = \frac{\theta - \theta_o}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

- Use the following R-function to calculate power

$$\gamma(\theta) = \text{pt}(-\text{t}_{\alpha, n+m-2}, \Delta) = \text{pt}(-\text{qt}(1 - \alpha, n + m - 2), n + m - 2, \Delta)$$

- p-value is as follows:

$$p-value = P_{\theta_o}(T < T_{obs}) = G_o(T_{obs})$$

where T_{obs} is the value calculated from the data

G_o is the cdf of a central t -distribution with $df = n + m - 2$,

- Use the following R-function to obtain the p-value:

$$p-value = \text{pt}(\mathbf{T}_{obs}, n + m - 2)$$

For the two-sided hypotheses, we have

Case 3. $\mathbf{H}_0 : \mu_1 - \mu_2 = \theta_0$ versus $\mathbf{H}_1 : \mu_1 - \mu_2 \neq \theta_0$

Let $t_{\alpha/2, n+m-2}$ the upper $\alpha/2$ percentile from a central t -distribution with $df = n + m - 2$, which can be obtained using the R-function:

$$t_{\alpha/2, n+m-2} = \text{qt}(1 - \alpha/2, n + m - 2)$$

- Reject H_0 if $|T| > t_{\alpha/2, n+m-2}$
- The power function for $\theta = \mu_1 - \mu_2$ is given by

$$\begin{aligned}\gamma(\theta) &= P_\theta(|T| > t_{\alpha/2, n+m-2}) \\ &= P_\theta(T < -t_{\alpha/2, n+m-2}) + P_\theta(T > t_{\alpha/2, n+m-2}) \\ &= G(-t_{\alpha/2, n+m-2}) + 1 - G(t_{\alpha/2, n+m-2})\end{aligned}$$

where G is the cdf of a noncentral t -distribution with $df = n + m - 2$ and noncentrality parameter

$$\Delta = \frac{\theta - \theta_0}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

- Use the following R-function to calculate power:

$$\gamma(\theta) = \text{pt}(-\text{qt}(1 - \alpha/2, n + m - 2), n + m - 2, \Delta) + 1 - \text{pt}(\text{qt}(1 - \alpha/2, n + m - 2), n + m - 2, \Delta)$$

- The p-value is calculated as follows:

$$\begin{aligned}p-value &= P_{\theta_o}(|T| > |T_{obs}|) \\ &= 2P_{\theta_o}(T > |T_{obs}|) \\ &= 2(1 - G_o(|T_{obs}|)),\end{aligned}$$

where G_o is the cdf of a central t -distribution with

$df = n + m - 2$, and T_{obs} is the value of T calculated from the data.

- Use the following R-function to obtain the p-value:

$$p-value = 2(1 - \text{pt}(|T_{obs}|, n + m - 2))$$

Example Suppose we wanted to test

$$H_o : \mu_1 \leq \mu_2 + 2 \text{ vs } H_1 : \mu_1 > \mu_2 + 2 \Rightarrow H_o : \mu_1 - \mu_2 \leq 2 \text{ vs } H_1 : \mu_1 - \mu_2 > 2$$

using independent random samples of sizes $n = 28$ and $m = 31$. The data yields the following summary statistics:

$$\bar{X} = 93.2 \quad S_1 = 7.5 \quad \bar{Y} = 88.3 \quad S_2 = 7.8$$

Test H_o vs H_1 using a size .05 test. Compute the power function for your test.

Solution Based on normal probability plots and the Shapiro-Wilk test, the researcher was relatively certain that the data was from normally distributed populations. The sample standard deviations indicate that the two populations have the same standard deviation. Therefore, we will use the Pooled t-test with

$$S_p = \sqrt{\frac{(28-1)(7.5)^2 + (31-1)(7.8)^2}{28+31-2}} = 7.659 \text{ with } df = 28+31-2 = 57$$

Using the pooled t-test with $\alpha = .05$, we have

$$\text{Reject } H_o \text{ if } T > t_{.05, 57} = qt(.95, 57) = 1.672$$

$$T = \frac{(93.2 - 88.3) - 2}{7.659 \sqrt{\frac{1}{28} + \frac{1}{31}}} = 1.45 \Rightarrow T = 1.45 > 1.672 \text{ and}$$

$$p = \text{value} = P[T \geq 1.45] = 1 - pt(1.45, 57) = .076 > .05 = \alpha$$

Therefore, we would not reject H_o and conclude that there is not significant evidence ($p=\text{value}=.076$) that μ_1 is 2 units or larger than μ_2 .

To estimate the size of the difference in μ_1 and μ_2 , construct a 95% C.I. on $\mu_1 - \mu_2$:

$$\bar{X} - \bar{Y} \pm Z_{\alpha/2} S_p \sqrt{1/n_1 + 1/n_2} = 93.2 - 88.3 \pm (1.96)(7.659) \sqrt{1/28 + 1/31} = 4.9 \pm 3.9 = (1.0, 8.8)$$

What is the chance that we have committed a Type I error?

What is the chance that we have committed a Type II error?

Because we do not know the value of σ , the power function will be computed in terms of the parameter, Δ . That is,

$$\gamma(\Delta) = P[\text{Reject } H_o \text{ given } \Delta] = P[T > t_{\alpha, n+m-2} \text{ given } \Delta] = 1 - pt(qt(1 - \alpha, n+m-2, \Delta))$$

We have $df=n+m-2=28+31-2=57$ and $\gamma(\Delta) = 1 - pt(qt(1 - .05, 57, \Delta))$, for various values of Δ . Using the following R code, we obtain the table of values for the power:

```
d<- seq(-1.5,.25)
pow<- 1-pt(qt(.95,57),57,d)
pow<- round(pow,3)
output<- cbind(d,pow)
```

Δ	-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
$\gamma(\Delta)$	0.004	0.009	0.016	0.029	0.050	0.081	0.125	0.183	0.256	0.341	0.435	0.534	0.630
Δ	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00	4.25	4.50	4.75	5.00	
$\gamma(\Delta)$	0.718	0.795	0.858	0.906	0.941	0.965	0.980	0.989	0.995	0.997	0.999	1.000	

What is the chance the test committed a **Type II error**?

Robustness of t-based Procedures - 2 Populations

The t-based confidence intervals and test procedures for comparing two population means require three conditions to be satisfied.

C1. Both population distributions are normally distributed.

- Use Shapiro-Wilk test and normal reference distribution plots to determine if the two samples have been selected from normal populations.
- When the population distributions are either very heavy tailed or highly skewed, the coverage probability for confidence intervals and the level and power of the t test will differ greatly from the stated values. Nonparametric alternatives to the t test which do not require normality should be used.

C2. The population distributions have equal variances: $\sigma_1 = \sigma_2$

- Later in this handout, the BFL-test of equal variances will be described. It can be used to determine if this condition is valid.
- In order to illustrate the effect of unequal variances, a computer simulation was performed in which two independent random samples were generated from normal populations having the same means but unequal variances: $\sigma_1 = k\sigma_2$ with $k=.25,.5,1,2$ and 4. For each combination of sample sizes and standard deviations, 10,000 simulations were run. For each simulation, a level .05 test was conducted. The proportion of the 10,000 tests that incorrectly rejected H_0 are presented in Table 1. If the pooled t test was unaffected by the unequal variances we would expect the proportions to be close to .05, the intended level, in all cases.

Table 1: Effect of Unequal σ 's on Type I Error Rates (values in table should be .05)

		$\sigma_1=k\sigma_2$				
n_1	n_2	$k=.25$.50	1	2	4
10	10	.062	.055	.048	.052	.064
10	20	.012	.019	.051	.113	.161
10	40	.0008	.0057	.050	.181	.295

From the results in Table 1, we can observe that when the sample sizes are equal the proportion of type I errors remains close to .05 (ranged from .048 to .064). When the sample sizes are different, the proportion of type I errors deviated greatly from .05. The more serious case is when the smaller sample size is associated with the larger variance. In this case, the error rates were much larger than .05. For example, when $n_1=10$, $n_2=40$, $\sigma_1 = 4\sigma_2$, the error rate was .295. However, when $n_1=10$, $n_2=10$ and $\sigma_1 = 4\sigma_2$, the error rate was .064, much closer to .05. This is remarkable and provides a convincing argument to use equal sample sizes unless there are important reasons to assign greater resources to one treatment. The following expressions are an indication of the lack of robustness of the pooled t-test performs poorly when the sample sizes are unequal.

If $\sigma_1 > \sigma_2$ and $n_1 < n_2$ then

$$E[S_p^2] = \frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2} < \sigma_1^2 \Rightarrow E\left[\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}\right] < \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \Rightarrow$$

Thus, $t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$ will be larger than expected and hence will produce excessive number of rejections, i.e., an inflated Type I error rate.

C3. The third condition is that the two random samples are independent.

- At the end of this handout, a test of correlation will be described. It can be used to evaluate the independence condition.
- Practically, we mean that the two samples are randomly selected from two distinct populations and that the elements of one sample are statistically independent of those of the second sample. There are two types of dependencies (data is not independent) that commonly occur in experiments and studies.
 - The data may have a *cluster effect* which often results when the data have been collected in subgroups. For example, 50 children are selected from 5 different classrooms for an experiment to compare the effectiveness of two tutoring techniques. The children are randomly assigned to one of the two techniques. Since children from the same classroom have a common teacher and hence may tend to be more similar in their academic achievement than children from different classrooms, the condition of independence between participants in the study may be lacking.
 - A second type of dependence is the result of *serial or spatial correlation*. When measurements are taken over time, the observations which are closer together in time tend to be more similar than observations collected at greatly different times, serially correlated. A similar dependence occurs when the data is collected at different locations. For example, water samples taken at various locations in a lake in order to assess whether a chemical plant is discharging pollutants into the lake. Measurements which are physically closer to each other are more likely to be similar than measurements which are taken farther apart. This type of dependence is *spatial correlation*. When the data is dependent, the procedures based on the t distribution will produce confidence intervals having coverage probabilities different from the intended values and tests of hypotheses having type I error rates different from the stated values. There are appropriate statistical procedure for handling this type of data but they are of a more advanced nature. A book on longitudinal or repeated measures data analysis or the analysis of spatial data would provide the details for the analysis of dependent data.

Also, correlated data is covered in the courses STAT 626-Time Series, STAT 636-Multivariate Analysis, and STAT 647-Spatial Statistics.



Case 2: Both populations are normal, $\sigma_1 \neq \sigma_2$, Xs and Ys are independent.

Separate Variance t-Test - Welch/Satterthwaite/Miller Test

We will modify the **Pooled t-Test** by using separate estimators of σ_1 and σ_2 in the t-statistic:

$$t^* = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

Suppose the following conditions hold:

C1. Both X_i s and Y_i s have a normal distributions

C2. X_i s and Y_i s are independent.

Under the above conditions, Welch and Satterthwaite showed that

t^* had an **approximate t-distribution** with $df = \nu$ given by

$$\nu = \frac{(C+1)^2(n-1)(m-1)}{C^2(m-1) + (n-1)} \text{ where } C = \frac{S_1^2/n}{S_2^2/m}$$

where the degrees of freedom, ν , is a random variable depending on the observed data

Consider the following situations:

1. (Equal Sample Sizes, Equal Variances)

Suppose $n=m$ and $\sigma_1 = \sigma_2$, which implies $C \approx 1$, then $\nu = 2(n-1)$, the same value for df as in the pooled t-Test.

There would be no difference in the two tests.

2. (Unequal Sample Sizes, Equal Variances)

Suppose $m-1 = k(n-1)$, with $k > 1$ and $\sigma_1 = \sigma_2$, which implies $C \approx k$, then

$$\nu = \frac{(k+1)^2 k}{k^3 + 1} (n-1), \quad df_{pooled} = k(n-1) + (n-1) = (k+1)(n-1) \text{ and } \frac{(k+1)^2 k}{k^3 + 1} < k+1.$$

For example, consider the following samples sizes:

n	m	ν	df_{pooled}
10	10	18	18
10	19	18.3	27
10	28	15.9	36
10	37	14.2	45
10	46	13.2	54

From the above table, we can conclude

- When the population variances are equal, that is, the Pooled t-Test is appropriate, but the sample sizes are very unequal, using the Separate Variance t-Test would yield a test having lower power than the Pooled t-test. (The larger the value for df, the greater the power of the test.)

To illustrate that the separate-variance t test is less affected by unequal variances than is the pooled t test, the data from the computer simulation reported in Table 1 was analyzed using the separate-variance t test.

The proportion of the 10,000 tests that incorrectly rejected H_0 are presented in Table 1.

If the separate-variance t test was unaffected by the unequal variances we would expect the proportions to be close to .05, the intended level, in all cases.

Table 2 The Effect of Unequal Variances on the Type I Error Rates of the Separate-Variance t-test

		$\sigma_1 = k\sigma_2$				
n_1	n_2	k=.25	.50	1	2	4
10	10	.051	.051	.047	.047	.054
10	20	.051	.051	.052	.053	.053
10	40	.054	.048	.048	.052	.051

From the results in Table 2, we can observe that

1. the separate-variance t test's type I error rates were consistently close to .05 in all cases considered above.
2. Using the values in Table 1, we observe the pooled t test had type I error rates very different from .05 when the sample sizes were unequal and we sampled from populations having very different variances
3. How accurate are the estimated Type I error rates in Tables 1 and 2? That is, how close are the simulated values to the true values for $P[\text{Type I error}]$?

Let $p = P[\text{Type I error}]$. We have $m = 10,000$ Bernoulli trials in our simulation. Therefore, our estimate of p , \hat{p} , the proportion of Type I errors in the 10,000 trials, has standard error:

$$SE(\hat{p}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \approx \frac{\sqrt{.05(1-.05)}}{\sqrt{10,000}} = .0022$$

Thus, we have greater than two decimal place accuracy in our estimates.

Comparison of Pooled t-test, t and Separate-Variance t-test, t^*

1. The procedure for selecting the critical values and computing the p-values for the Separate Variance t-Test are identical to the procedures for the Pooled t-Test after making the change in degrees of freedom from $df_{pooled} = n + m - 2$ to $df_{sepvar} = \nu$.
2. The power function for the separate variance test is very complex because its degrees of freedom (df) depend on the data. Thus, the derivation of the sampling distribution of the separate variance t-Test under the alternative is only practically accomplished using computer simulation.
3. We developed pooled-variance t methods based on the requirement of independent random samples from normal populations with equal population variances.

When the variances are not equal, we introduced the separate-variance t^* statistic. Confidence intervals and hypothesis tests based on these procedures (t or t^*) need not give identical results.

4. Standard computer packages often report the results of both t and t^* tests. Which of these results should you use in your statistical analysis?

- If the sample sizes are equal and the population variances are equal, the separate-variance t test and the pooled t test give algebraically identical results, that is, the computed t equals the computed t' .

Thus, why not always use t^* in place of t when $n_1=n_2$?

- The reason we would select t over t^* is that the df for t are nearly always larger than the df for t^* and hence the power of the t test is greater than the power of the t^* test when the variances are equal.
- When the sample sizes and variances are very unequal, the results of the t and t^* procedures may differ greatly. The evidence in such cases indicates that the separate-variance methods are somewhat more reliable and more conservative than the results of the pooled t methods.
- However, if the populations have both different means and different variances, an examination of just the size of the difference in their means $\mu_1 - \mu_2$ would be an inadequate description of how the populations differ. We should always examine the size of the differences in both the means and the standard deviations of the populations being compared. The Brown-Forsythe-Levene test procedure can be used to examine the difference in the standard deviations of two populations.
- In many undergraduate textbooks, they do not bother to compute the correct df for the separate variance t-test. Instead they specify the degrees of freedom as $df = \min(n - 1, m - 1)$ which produces a test statistic having reduced power and hence greater chance of producing Type II errors. The test is referred to as a conservative test statistic. Not a good idea!
- Suppose $n = 12, m = 15, S_1 = 6, S_2 = 10 \Rightarrow C = \frac{36/12}{100/15} = .45, \nu = \frac{(1.45)^2(11)(14)}{(.45)^2(14)+11} = 23.4$
- $n + m - 2 = 25$ but $\min(n-1, m-1) = 11$. Thus, using the minimum of the sample sizes for the df would yield a considerable loss of power compared to using 23 or 25.

STOP CN ¹¹ Wednesday 11/17/21

START CN

Friday 11/19/24

Case 3: Both populations are nonnormal, but the two distributions belong to same loc-scale family; X s and Y s are independent.

Wilcoxon Rank Sum Test

Suppose we have two populations/processes distributions which are members of the same location-scale family with the same scale parameter but possibly different location parameters:

- Let X_1, \dots, X_n be iid with cdf $F_1(x) = F_o(\theta_1 + \nu_1 x)$ and
- Let Y_1, \dots, Y_n be iid with cdf $F_2(y) = F_o(\theta_2 + \nu_2 y)$,
- where $\nu_1 = \nu_2 = \nu$, with ν unknown and F_o being the standard member of the family of distributions.

The research proposal is to test hypotheses about $\theta_1 - \theta_2$.

The following conditions are placed on the population distributions:

C1. F_1 and F_2 are members of the same location-scale families with the same value for the scale parameter.

C2. The X is and Y is are independent

C3. X_1, \dots, X_n be iid with cdf $F_1(x) = F_o(\theta_1 + \nu x)$

C4. Y_1, \dots, Y_n be iid with cdf $F_2(y) = F_o(\theta_2 + \nu y)$

1. If F_o was a standard normal cdf then we would have the same conditions as Case 1 and could use the pooled t-Test.
2. The Wilcoxon Rank Sum test is a rank-transform procedure: We replace the data with their ranks in the combined sample and use the Wilcoxon Rank Sum test to evaluate differences in the location parameters: $\theta_1 - \theta_2$.

Implementation of Wilcoxon Rank Sum Test

1. Order the $N = n + m$ observations from smallest to largest
2. Replace the data values with their ranks: 1 to N (In case of ties, assign the average rank to the tied values)
3. Let W_1 be the sum of the ranks for the X is and let W_2 be the sum of the ranks for the Y is
4. $W_1 + W_2 = \sum_{i=1}^N i = N(N + 1)/2$
5. Given the size of the test, α , use Table A.11, Wilcoxon Rank Sum Table in eCampus, Lecture notes folder, or use the R functions `qwilcox` and `pwilcox` to set the critical value and to compute the p-value for W_1 or W_2

on next page

Note: qwilcox! pwilcox 12 are actually working w/
Mann-Whitney U-test (discussed below).

Table A.11 Upper-Tail Probabilities of the Null Distribution of the Wilcoxon-Mann-Whitney Statistic

$P(W \geq w_1)$					$P(W \geq w_1)$					$P(W \geq w_1)$				
$=$					$=$					$=$				
n_1	n_2	w_1	u_1	$P(U \geq u_1)$	n_1	n_2	w_1	u_1	$P(U \geq u_1)$	n_1	n_2	w_1	u_1	$P(U \geq u_1)$
		3	3	15 9 0.050	4	9	44 34	0.006		6	10	66 45	0.059	
		4	17	11 0.057		10	42 32	0.053			10	69 48	0.028	
		4	18	12 0.029		10	44 34	0.027			10	72 51	0.011	
		5	20	14 0.036		10	46 36	0.012			10	74 53	0.005	
		5	21	15 0.018		10	47 37	0.007		7	7	66 38	0.049	
		6	22	16 0.048	5	5	36 21	0.048		7	68 40	0.027		
		6	23	17 0.024		5	37 22	0.028		7	71 43	0.009		
		6	24	18 0.012		5	39 24	0.008		7	72 44	0.006		
		7	24	18 0.058		5	40 25	0.004		8	71 43	0.047		
		7	26	20 0.017		6	40 25	0.041		8	73 45	0.027		
		7	27	21 0.008		6	41 26	0.026		8	76 48	0.010		
		8	27	21 0.042		6	43 28	0.009		8	78 50	0.005		
		8	28	22 0.024		6	44 29	0.004		9	75 47	0.057		
		8	29	23 0.012		7	43 28	0.053		9	78 50	0.027		
		8	30	24 0.006		7	45 30	0.024		9	81 53	0.011		
		9	29	23 0.050		7	47 32	0.009		9	83 55	0.006		
		9	31	25 0.018		7	48 33	0.005		10	80 52	0.054		
		9	32	26 0.009		8	47 32	0.047		10	83 55	0.028		
		9	33	27 0.005		8	49 34	0.023		10	87 59	0.009		
		10	31	25 0.056		8	51 36	0.009		10	89 61	0.005		
		10	33	27 0.024		8	52 37	0.005		8	84 48	0.052		
		10	34	28 0.007		9	50 35	0.056		8	87 51	0.025		
		10	35	29 0.003		9	53 38	0.021		8	90 54	0.010		
	4	4	24	14 0.057		9	55 40	0.009		8	92 56	0.005		
	4	4	25	15 0.029		9	56 41	0.006		9	89 53	0.057		
	4	4	26	16 0.014		10	54 39	0.050		9	93 57	0.023		
	4	5	27	17 0.056		10	56 41	0.027		9	96 60	0.010		
	4	5	28	18 0.032		10	59 44	0.010		9	98 62	0.006		
	4	5	29	19 0.016		10	60 45	0.006		10	95 59	0.051		
	4	5	30	20 0.008	6	6	50 29	0.047		10	98 62	0.027		
	4	6	30	20 0.057		6	52 31	0.021		10	102 66	0.010		
	4	6	32	22 0.019		6	54 33	0.008		10	104 68	0.006		
	4	6	33	23 0.010		6	55 34	0.004		9	104 59	0.057		
	4	6	34	24 0.005		7	54 33	0.051		9	108 63	0.025		
	4	7	33	23 0.055		7	56 35	0.026		9	111 66	0.012		
	4	7	35	25 0.021		7	58 37	0.011		9	114 69	0.005		
	4	7	36	26 0.012		7	60 39	0.004		10	110 65	0.056		
	4	7	37	27 0.006		8	58 37	0.054		10	114 69	0.027		
	4	8	36	26 0.055		8	61 40	0.021		10	118 73	0.011		
	4	8	38	28 0.024		8	63 42	0.010		10	121 76	0.005		
	4	8	40	30 0.008		8	65 44	0.004		10	127 72	0.053		
	4	8	41	31 0.004		9	62 41	0.057		10	131 76	0.026		
	4	9	39	29 0.053		9	65 44	0.025		10	135 80	0.012		
	4	9	41	31 0.025		9	68 47	0.009		10	138 83	0.006		
	9	43	33	0.010		9	70 49	0.004						

SOURCE: Adapted from M. Hollander and D. A. Wolfe (1972), *Nonparametric Statistical Methods*, New York: Wiley, Table A.5.

An equivalent test was developed by Mann and Whitney at nearly the same time as the Wilcoxon Rank Sum test. Although the two tests appear to be different they are in fact identical.

Implementation of Mann-Whitney Test:

1. Compare each X_i with each Y_j . Let U_1 be the number of pairs (X_i, Y_j) in which $X_i > Y_j$: $U_1 = \sum_{i,j} I(X_i > Y_j)$
2. Compare each X_i with each Y_j . Let U_2 be the number of pairs (X_i, Y_j) in which $X_i < Y_j$: $U_2 = \sum_{i,j} I(X_i < Y_j)$
3. There are nm pairs and $U_1 + U_2 = nm$.
4. Given the size of the test α use Table A.11 to set the critical value (upper α - percentiles) or to compute the p-value

Note: The Wilcoxon Rank Sum Test and the Mann-Whitney U-test are equivalent because:

$$W_1 = U_1 + \frac{n(n+1)}{2} \quad \text{and} \quad W_2 = U_2 + \frac{m(m+1)}{2}$$

Thus, it does not matter which test we use to test hypotheses about $\theta_1 - \theta_2$.

H1. To test $H_o : \theta_1 \leq \theta_2$ versus $H_1 : \theta_1 > \theta_2$

- Reject H_o if $W_1 \geq W_{\alpha,n,m}$ and
- compute $p-value = 1 - G(W_1 - 1)$
where G is the cdf of W_1 when $\theta_1 = \theta_2$. Table 11 can be used to do both.
- R-function **wilcox** yields both critical value and p-value:

$$W_{\alpha,n,m} = qwilcox(1 - \alpha, n, m) + \frac{n(n+1)}{2}$$

$$G(W_1) = pwilcox(W_1 - \frac{n(n+1)}{2}, n, m)$$
- Alternatively, the R-function **wilcox.test(x,y,alternative="g",paired=F)** can be used to conduct the Wilcoxon Rank Sum Test, producing both critical value and p-value

H2. To test $H_o : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$

Reject H_o if $W_1 \geq W_{\alpha/2,n,m}$ or if $W_1 \leq n(n+m+1) - W_{\alpha/2,n,m}$

$$p-value = 2 * min(P[W_1 \geq w_1], P[W_1 \leq w_1]) = 2 * min(1 - G(w_1 - 1), G(w_1))$$

where w_1 is the observed value of w_1 and G is the cdf of W_1 when $\theta_1 = \theta_2$.

- R-function **wilcox** yields both critical value and p-value:

$$W_{\alpha/2,n,m} = qwilcox(1 - \alpha/2, n, m) + \frac{n(n+1)}{2}$$

$$G(W_1) = pwilcox(W_1 - \frac{n(n+1)}{2}, n, m)$$
- Alternatively, the R-function **wilcox.test(x,y,alternative="t",paired=F)** can be used to conduct the Wilcoxon Rank Sum Test, producing both critical value and p-value

Note that "g" for greater than was changed to "t" for two-sided test.

EXAMPLE

Many states are considering lowering the blood-alcohol level at which a driver is designated as driving under the influence (DUI) of alcohol. An investigator for a legislative committee designed the following test to study the effect of alcohol on reaction time. Ten subjects consumed a specified amount of alcohol. Another group of ten subjects consumed the same amount of a nonalcoholic drink, a placebo. The two groups did not know whether they were receiving alcohol or the placebo. The twenty subjects' average reaction times (in seconds) to a series of simulated driving situations are reported in the following table. Does it appear that alcohol consumption increases reaction time?

Placebo	0.90	0.37	1.63	0.83	0.95	0.78	0.86	0.61	0.38	1.97
Alcohol	1.46	1.45	1.76	1.44	1.11	3.07	0.98	1.27	2.56	1.32

We want to test the hypotheses:

H_0 : The distributions of reaction times for the placebo and alcohol populations are identical.

H_1 : The distribution of reaction times for the placebo consumption population is shifted to the left of the distribution for the alcohol population. (That is, larger reaction times are associated with the consumption of alcohol.)

→ not normal → no z-test

The Shapiro-Wilk test yields p-value = .084 and .022 for the Placebo and Alcohol data, respectively. Therefore, because of the small sample sizes and the non-normality of the data, the Wilcoxon rank sum test will be implemented.

Solution

The Wilcoxon rank sum test will be conducted to evaluate whether alcohol consumption increases reaction time. Let W_1 be the sum of the ranks for the Alcohol group.

Placebo	7	1	16	5	8	4	6	3	2	18
Alcohol	15	14	17	13	10	20	9	11	19	12

1. For $\alpha = .053$, reject H_0 if $W_1 \geq 127$,

using Table A11 in eCampus with $\alpha = .053$ one-tailed and $n = m = 10$.

2. W_1 is computed by summing the ranks from the Alcohol group,

$$W_1 = 15 + 14 + 17 + 13 + 10 + 20 + 9 + 11 + 19 + 12 = 140.$$

3. Since 140 is greater than 127, we reject H_0

$$p\text{-value} = 1 - G(W_1 - 1) = 1 - G(140 - 1) = 1 - G(139) < 1 - G(138) < .006, \text{ using Table 11.}$$

4. Using the R-functions:

$$W_{.05,10,10} = qwilcox(.948, 10, 10) + \frac{10(10+1)}{2} = 72 + 55 = 127$$

$$p\text{-value} = 1 - G(140 - 1) = 1 - pwilcox(139 - \frac{10(10+1)}{2}, 10, 10) = 1 - pwilcox(139 - 55, 10, 10) = .0034$$

The above calculations can be directly obtain using the following R program:

5. Conclude there is significant evidence, p-value=.0034, that the placebo population has shorter reaction times than the population of alcohol consumers.

Program and Output from R

Let x be the vector of values for the X_i s and y be the vector of values for the y_i s then

The R-function `wilcox.test(x,y,alternative="l",paired=F)` yields

```
x = c(.9,.37,1.63,.83,.95,.78,.86,.61,.38,1.97)
y = c(1.46,1.45,1.76,1.44,1.11,3.07,0.98,1.27,2.56,1.32)
wilcox.test(x,y,alternative="l",paired=F)

Wilcoxon rank sum test

data: x and y

W = 15, p-value = 0.003421

alternative hypothesis: true mu is less than 0
```

Note in the above output, R calculates the Mann-Whitney U not the Wilcoxon Rank Sum W :

$$W_2 = \frac{N(N+1)}{2} - W_1 = \frac{20(20+1)}{2} - 140 = 70$$

$$U_2 = W_2 - \frac{n(n+1)}{2} = 70 - 55 = 15$$

Approximation using Central Limit Theorem

~~Notation~~

For large n and m , the distribution of the Wilcoxon Rank Sum test can be approximated by a normal distribution: Let

$$\mu = \frac{n(n + m + 1)}{2}$$

$$\sigma = \sqrt{\frac{nm(n + m + 1)}{12}}$$

Then

$$Z = \frac{W_1 - \mu}{\sigma} \quad \text{is approx. distributed as } N(0, 1)$$

We can use the $N(0, 1)$ tables to set the critical values and compute p-values for W_1 .

In our example on the previous page, we would have

1. $W_1 = 140$

2. $\mu_W = \frac{n(n+m+1)}{2} = \frac{10(20+1)}{2} = 105$

3. $\sigma_W = \sqrt{\frac{nm(n+m+1)}{12}} = \sqrt{\frac{(10)(10)(10+10+1)}{12}} = 13.2288$

4. $W_{.05,10,10} \approx \mu_W + Z_{.05}\sigma_W = 105 + (1.645)(13.2288) = 126.76$

4. $Z = \frac{W_1 - \mu}{\sigma} = \frac{140 - 105}{13.2288} = 2.646 \Rightarrow p-value = P[Z > 2.646] = 1 - \Phi(2.646) = 0.0041$

5. Using the Wilcoxon Tables or R functions we obtained $W_{.053,10,10} = 127$ and p-value = 0.0034.

Comparison of Rank Sum Test to t-Test:

The Wilcoxon Rank Sum test is a distribution-free test in that the critical values and p-value are calculated without having to specify the functional form of F_o .

However, the power function of the Wilcoxon Rank Sum test depends on the size of difference in the location parameter and the population cdf F_o . When the sample sizes are large and the $N(0, 1)$ approximation is appropriate, the values of μ and σ when $\theta_1 - \theta_2 \neq 0$ will depend on F_o . This makes it nearly impossible to obtain an analytical form for the power function. Therefore, we will study power of the Wilcoxon Rank Sum test through the following simulation study:

The Wilcoxon rank sum test is an alternative to the two-sample t test with the rank sum test requiring somewhat weaker conditions on the populations from the data is selected than are required for using the t test.

1. The Wilcoxon's test does not require the two populations to have normal distributions, it only requires that the distributions are identical except possibly that one distribution is shifted from the other distribution.
2. When both distributions are normal, the t test will be more likely to detect an existing difference, that is, the t test has greater power than the rank sum test. This is logical since the t test uses the magnitudes of the observations rather than just their relative magnitudes (ranks) as is done in the rank sum test.
3. When the two distributions have long or heavy tails, the Wilcoxon rank sum test has greater power, that is, it is more likely to detect a shift in the population distributions than the t test.
4. Also, the level or probability of a type I error for the Wilcoxon rank sum test will be equal to nominal (stated) level for all population distributions. Whereas, the t test's *actual* level will deviate from its stated value when the population distributions are nonnormal. This is particularly true when nonnormality of the population distributions is present in the form of severe skewness or extreme outliers.
5. Randles and Wolfe (1979) investigated the effect of skewed and heavy tailed distributions on the power of the t test and the Wilcoxon rank sum test.
6. Table 3 contains a portion of the results of their simulation study.
 - a. For each set of distributions, sample sizes and shifts in the populations, 5000 samples were drawn and the proportion of times a level $\alpha = .05$ t test or Wilcoxon rank sum test rejected H_0 was recorded.
 - b. The distributions considered were Normal, Double Exponential (symmetric, heavy-tailed), Cauchy (symmetric, extremely heavy-tailed), and Weibull (skewed to the right).
 - c. Shifts of size 0, $.6\sigma$, and 1.2σ were considered, where σ denotes the standard deviation of the distribution, with the exception of the Cauchy distribution, where σ is a general scale parameter.

*curve
should have
all odds*

Table 3 Power of Pooled t-Test(t) and Wilcoxon Rank Sum Test(W) with $\alpha = .05$

		Distribution F_o												
		Normal			DoubleExp			Cauchy			Weibull			
Shift:		0	.6	1.2	0	.6	1.2	0	.6	1.2	0	.6	1.2	
n,m	Test	t	.044	.213	.523	.045	.255	.588	.024	.132	.288	.049	.221	.545
		W	.046	.208	.503	.049	.269	.589	.051	.218	.408	.049	.219	.537
5,15	t	t	.047	.303	.724	.046	.304	.733	.056	.137	.282	.041	.289	.723
		W	.048	.287	.694	.047	.351	.768	.046	.284	.576	.049	.290	.688
15,15	t	t	.052	.497	.947	.046	.507	.928	.030	.153	.333	.046	.488	.935
		W	.054	.479	.933	.046	.594	.962	.046	.484	.839	.046	.488	.927

t-test robust to slight skewness.

- When the distribution is normal, the t test is only slightly better, greater power values, than the Wilcoxon rank sum test.
- For the double exponential, the Wilcoxon test has greater power than the t test.
- For the Cauchy distribution, the level of the t test deviates significantly from .05 and its power is much lower than the Wilcoxon test.
- When the distribution was somewhat skewed, Weibull distribution, the tests had similar performance. Furthermore, the level and power of the t test were nearly identical to the values when the distribution was normal.
- The t test is quite robust to skewness, except when there are numerous outliers.

The accuracy of the simulated values for the power function are obtained by considering the power as a proportion, p , which we are estimating by running 5000 replications of a Bernoulli trial in which we record

$$Y = \begin{cases} 1 & \text{if Tests rejects } H_o \\ 0 & \text{if Tests fails to reject } H_o \end{cases}$$

Our estimate of the power is then $\hat{p} = \frac{m}{5000}$ where m is the number of times $Y = 1$.

The estimated standard error of \hat{p} is $\sqrt{\frac{\hat{p}(1-\hat{p})}{5000}}$ which yields the following values:

$$\text{SE(Estimated Power)} = \begin{cases} .0031 & \text{for } \hat{p} \leq .05, \hat{p} \geq .95 \\ .0057 & \text{for } \hat{p} \approx .20, p \approx .80 \\ .0067 & \text{for } \hat{p} \approx .35, \hat{p} \approx .65 \\ .0071 & \text{for } \hat{p} \approx .5 \end{cases}$$

where \hat{p} is the estimated power from the simulation.

Case 4: X's and Y's are Paired (~~there will be correlation between samples~~)

When the experimental units have a high degree of variability prior to being randomly assigned to one of two treatments, the experimental units are paired so as to reduce the variance between \bar{X} and \bar{Y} . The two units within a pair are more alike than units in different pairs. This type of design will result in a more precise evaluation of the differences in two populations.

Example 1: Common Attribute to Both Measurements

An experiment was conducted to compare the mean lengths of time required for the bodily absorption of two drugs, D_1 and D_2 , used in the treatment of epilepsy. A group of 20 epileptic patients were randomly assigned for inclusion in the study. The 20 patients had very different severity of the disease. The patients were paired so that the patients within a pair were diagnosed as having about the same level of severity of the disease. The two drugs were randomly assigned within each pair of patients. The length of time (in minutes) for the drug to reach a specified level in the blood was recorded for each of the patients. The researchers wanted to know if there was a difference in the performance of the two drugs.

	Disease Severity									
	1	2	3	4	5	6	7	8	9	10
Drug D_1	19.8	45.4	32.0	24.5	47.2	18.1	50.2	47.2	16.8	41.2
Drug D_2	23.0	37.3	11.4	60.6	72.1	41.4	42.6	43.8	42.8	65.3

Example 2: Before and After Study

After strip mining for coal, the mining company is required to restore the land to its condition prior to mining. One of the many factors that is considered is the pH of the soil, which is an important factor in determining what types of plants will survive in a given location. The strip mined area was divided into grids before the mining took place. Twelve grids were randomly selected and the soil pH was measured prior to the mining. When the mining was completed, the mining company attempted to restore the land to its unmined state. The soil pH readings were then taken on the same 12 grids. Is the distribution of pH readings different after mining when compared to the before mining readings?

	Grid Location											
	1	2	3	4	5	6	7	8	9	10	11	12
Before	10.02	10.16	9.96	10.01	9.87	10.05	10.07	10.08	10.05	10.04	10.09	9.92
After	10.21	10.16	10.11	10.10	10.07	10.13	10.08	10.30	10.17	10.10	10.06	10.24

A proper analysis of paired data needs to take into account the lack of independence between the two samples. The sampling distribution for the difference in the sample means, $(\bar{Y}_1 - \bar{Y}_2)$ will have mean and standard error

$$E[\bar{Y}_1 - \bar{Y}_2] = \mu_{\bar{Y}_1 - \bar{Y}_2} = \mu_1 - \mu_2 \quad - \text{Corr does not affect mean}$$

$$SE(\bar{y}_1 - \bar{y}_2) = \sigma_{\bar{y}_1 - \bar{y}_2} = \sqrt{Var(\bar{y}_1 - \bar{y}_2)} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho}{n}}, \quad - \text{Corr does not affect} \\ \text{Corr } \{ \text{SE}$$

where ρ measures the amount of correlation (dependence) between the two samples.

When the two samples produce similar measurements, ρ is positive and the standard error of $\bar{Y}_1 - \bar{Y}_2$ will be smaller than what would be obtained using two independent samples.

In this type of situation, the use of paired data will reduce the standard error of the difference in the sample means in comparison to using independent samples for the two populations. The reduction of the variance would lead to a test having greater power if a paired data experiment was conducted

rather than having two independent data sets. However, the reduction in variance is obtained only by a concurrent reduction in the df for the t -percentiles. That is,

$$t_{\frac{\alpha}{2}, n-1} > t_{\frac{\alpha}{2}, 2(n-1)}$$

Thus, in attempting to decide between using a paired analysis or two independent samples, the trade-off between reduced variability and reduced df must be taken into consideration.

Paired Data Analysis

*usually the reduced variability makes the paired test better
but it's possible reduced df makes the two-sample t-test better.*

In many cases the sample size, n , is too small to properly estimate the ρ . That is, suppose we have n independent pairs of data values

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$$

an estimator of ρ is given by

$$\hat{\rho} = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

For small n , the above estimator has a large standard error.

An analysis of paired data which eliminates the need to estimate ρ transforms the data to differences in the n pairs of measurements,

$$D_1 = X_1 - Y_1, \quad D_2 = X_2 - Y_2, \quad \dots, \quad D_n = X_n - Y_n$$

and obtains \bar{D} , S_D , the mean and standard deviations in the D_i 's.

Also, the hypotheses about μ_1 and μ_2 must be now formulated into hypotheses about the mean of the differences, $\mu_D = \mu_1 - \mu_2$.

With this formulation, it is then possible to apply standard 1-Population procedures to the differences, D_i 's.

The conditions required to develop t-distribution procedures for testing hypotheses and constructing confidence intervals for μ_x are

Paired Data: Normal Differences

1. The sampling distribution of the D_i 's is a normal distribution.
2. The D_i 's are independent, that is, the pairs of observations are independent.

The t-Test is then applied to hypotheses framed in terms of $\mu_D = \mu_1 - \mu_2$:

$$T = \frac{\sqrt{n}(\bar{D} - \mu_{Do})}{S_D}$$

The central t-distribution with $df = n - 1$ is then used to set critical values, compute p-values, and construct C.I.'s for the parameter: $\mu_D = \mu_1 - \mu_2$. Power calculations and sample size determinations are made using the noncentral t-distribution with

$$df = n - 1 \quad \text{and noncentrality parameter } \Delta = \frac{\sqrt{n}(\mu_D - \mu_{Do})}{\sigma_x}$$

Paired Data: Non-Normal Differences

1. The sampling distribution of the D_i 's is not a normal distribution.
2. The D_i 's are independent, that is, the pairs of observations are independent.

Case 1: If the sampling distribution of the D_i 's is symmetrically distributed with a continuous cdf, then we would simply apply the **Wilcoxon Signed-Rank** procedure to the differences:

$$D_1 = X_1 - Y_1, \quad D_2 = X_2 - Y_2, \quad \dots, \quad D_n = X_n - Y_n$$

Case 2: If the sampling distribution of the D_i 's is a continuous distribution which is not symmetric then the **sign test** could be used to test hypotheses about $\mu_D = \mu_1 - \mu_2$.

The R-functions **t.test** and **wilcox.test** can be used to conduct the tests for paired data by simply using **paired=T** in both functions.

Sample Size Calculations Based on Power Specifications

The determination of the appropriate samples sizes n and m can be formulated in the same manner as for the 1-sample problem covered in Handout 12:

Case 1: Two Independent Samples from Normal Populations with $\sigma_1 = \sigma_2 = \sigma$

Suppose we want to test $H_1 : \mu_1 - \mu_2 > 0$ (the results for $H_1 : \mu_1 - \mu_2 < 0$ are identical)

1. Suppose we require $n = m$.

Find the minimum sample size n such that a level α test has power of at least $1 - \beta$ to detect that $\mu_1 - \mu_2 > \delta$.

If our test statistic is

$$T = \frac{(\bar{X} - \bar{Y}) - \theta_o}{S_P \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{(\bar{X} - \bar{Y}) - \theta_o}{S_P \sqrt{\frac{2}{n}}}$$

Pooled T statistic

we would need to use a power function based on the non-central t-distribution to determine the appropriate value of n . Many software packages have sample size determination as an option for this type of testing situation.

A very accurate approximation can be obtained using the test statistic:

$$Z = \frac{(\bar{X} - \bar{Y}) - \theta_o}{\sigma \sqrt{\frac{2}{n}}}$$

provided the sample sizes are not too small. Using this test statistic, we have the appropriate sample size is

$$n = 2 \left[\frac{\sigma(z_\alpha + z_\beta)}{\delta} \right]^2$$

if we pretend we know or we can use this.

where σ would be a guess at the standard deviation provided by the researcher.

For a two-sided research hypothesis: $H_1 : \mu_1 - \mu_2 \neq 0$, just replace z_α with $z_{\alpha/2}$.

2. For unequal sample sizes but with the sample sizes being multiples of each other, $m = kn$, then we have

$$n = \frac{k+1}{k} \left[\frac{\sigma(z_\alpha + z_\beta)}{\delta} \right]^2 \quad \text{and} \quad m = kn = (k+1) \left[\frac{\sigma(z_\alpha + z_\beta)}{\delta} \right]^2$$

Case 2: Paired Samples from Normal Populations

We can directly use our results from the 1-Population situation discussed in Handout 12.

Just consider the data as being $D_i = X_i - Y_i$ with mean $\mu_D = \mu_1 - \mu_2$ and standard deviation σ_D . We can then determine the appropriate sample size n , the number of pairs from either the Table A11 given in Handout 12 or from the formula:

$$n = \left[\frac{\sigma_D(z_\alpha + z_\beta)}{\delta} \right]^2$$

where the researcher would supply a guess at σ_D .

Example 1 Suppose we want to test $H_0 : \mu_1 \leq \mu_2$ vs $H_1 : \mu_1 > \mu_2$

based on independent random samples of size n and m .

The researcher wants to determine the values of $n = m$ such that

1. an $\alpha = .05$ test would have
2. a probability of at least 90% of rejecting H_0 whenever
3. μ_1 is at least 1.5 units larger than μ_2 .
4. The researcher is fairly certain that $\sigma_1 = \sigma_2 = 3$.

Solution

- $\alpha = .05$ and $\beta = 1 - .9 = .1$, the appropriate sample size would be

$$n = 2 \left[\frac{\sigma(z_\alpha + z_\beta)}{\delta} \right]^2 = 2 \left[\frac{(3)(1.645 + 1.282)}{1.5} \right]^2 = 68.5$$

- $n = m = 69$ observations would be needed from each of the populations to achieve the desired specifications.
- What is the actual power using the above approximation?

- With $n = m = 69$, $df = n + m - 2 = 136$, $\alpha = .05$, and

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{1.5}{3 \sqrt{\frac{1}{69} + \frac{1}{69}}} = 2.936835$$

we have the following power:

$$\gamma(\mu_1 - \mu_2) = \gamma(1.5) = 1 - pt(qt(.95, 136), 136, 2.936835) = .8992.$$

- The above approximation yielded a power value a bit lower than the specified value,

$$\gamma(2.936835) = 0.8992 < .9$$

- Using the sample size/power function in R,

```
power.t.test(n=,delta=1.5,sd=3,sig.level=.05,power=.9,type=c("two.sample"),alternative=c("one.sided"))
```

yields $n = 69.2$, use $n = 70$

```
power.t.test(n=70,delta=1.5,sd=3,sig.level=.05,power=,type=c("two.sample"),alternative=c("one.sided"))
```

yields power = 0.9030

Example 2 Suppose we want to test $H_0 : \mu_1 \leq \mu_2$ vs $H_1 : \mu_1 > \mu_2$ based on n independent pairs of observations.

The researcher wants to determine the values of n such that

1. an $\alpha = .01$ test would have
2. a probability of at least 80% of rejecting H_0
3. whenever μ_1 is at least 1.2 units larger than μ_2 .

Solution

- Suppose the researcher is fairly certain that $\sigma_D = 2$.
- With $\alpha = .01$ and $\beta = 1 - .8 = .2$, we would have that the appropriate sample sizes would be

1. Using Table A11 in Handout 12 with

$$\alpha = .01, \quad \beta = 1 - .8 = .2,$$

$$\phi = \frac{\mu_1 - \mu_2}{\sigma_D} = \frac{1.2}{2} = .6$$

We obtain, $n = 31$.

The power with $\Delta = \frac{\sqrt{n}\mu_D}{\sigma_D} = \frac{\sqrt{31}(1.2)}{2} = 3.34066$ is
 $\gamma(1.2) = 1 - pt(qt(.99, 30), 30, 3.34066) = .8056$

2. The quick and easy approximate formula, yields

$$n = \left[\frac{\sigma_D(z_\alpha + z_\beta)}{\delta} \right]^2 = \left[\frac{(2)(2.326 + .842)}{1.2} \right]^2 = 27.9$$

- Requires $n = 28$ pairs of observations from the populations to achieve the desired specifications.
- What is the actual power using the above approximation?
- With $n = 28, df = n - 1 = 27, \Delta = \frac{\sqrt{n}\mu_D}{\sigma_D} = \frac{\sqrt{28}(1.2)}{2} = 3.1749$,
- The following power of an $\alpha = .01$ test is

$$\gamma(\mu_D) = \gamma(1.2) = 1 - pt(qt(.99, 27), 27, 3.1749) = .7542 < .8.$$

- The above approximation yielded a power value considerably lower than the specified value.

3. From R:

```
power.t.test(n=,delta=1.2,sd=2,sig.level=.01,power=.8,type="paired",alternative=c("one.sided"))
we obtain n = 30.64 which implies we need 31 pairs of data values to achieve a power of 80% at
a difference of 1.2 using an  $\alpha = .01$  test of a 1-sided hypothesis.
```

START CN

Warday 11/22/21

Analysis of Multiple Population (Process) Proportions

MODEL I: Product Multinomial Sampling Model

Pearson Chi-square Test of Homogeneity of Proportions

The data in the following table show the incidence of injuries for 750 accidents involving breakaway light poles. Of interest to the project investigators is whether there are different proportions of injuries in different parts of the country. Let p_i be the proportion of accidents involving breakaway light poles that resulted in an injury. The researcher wanted to test the hypotheses

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 \text{ vs } H_1 : p_i \neq p_j \text{ for at least one pair } (i, j)$$

A random sample of 150 accidents involving breakaway light poles was taken in each of five geographic locations. The data is given in the following table:

	Geographic Location					Total
	1	2	3	4	5	
Injuries	35	36	46	20	24	161
No Injuries	115	114	104	130	126	589
Total	150	150	150	150	150	750

The product multinomial sampling model involves t populations from which independent random samples of size: n_1, \dots, n_t are selected. The n_i experimental units are classified into one of two categories: Type A or Type B. The p_i 's represent the proportion of units in population i that are of Type A. The sample data can be represented by a $2 \times t$ table:

	Populations					Total
	1	2	3	\dots	t	
Type A	O_{11}	O_{12}	O_{13}	\dots	O_{1t}	R_1
Type B	O_{21}	O_{22}	O_{23}	\dots	O_{2t}	R_2
Total	n_1	n_2	n_3	\dots	n_t	N

The hypotheses to be tested are

$$H_0 : p_1 = \dots = p_t = p_o \text{ vs } H_1 : p_i \neq p_j \text{ for at least one pair } (i, j)$$

where p_o is unknown.

we need a test statistic & we need to know the sampling distribution of the test statistic under the null hypothesis

- Under the null hypotheses of no difference in the t proportions, we would expect to observe
 - $E_{1j} = n_j p_o$ Type A units from the j th population
 - $E_{2j} = n_j(1 - p_o)$ Type B units from the j th population, where
 - p_o is the proportion of Type A units in each of the t populations.
- Since p_o is unknown, we estimate it with
 - $\hat{p}_o = \frac{R_1}{N}$ because
 - to get critical values of p-value
 - under the null hypothesis
 - a power for the test.
- all t populations have the same proportion under the null hypothesis.

- Under the null hypothesis $p_1 = \dots = p_t = p_o$, therefore, the *estimated* expected counts are
 1. $\hat{E}_{1j} = n_j \hat{p}_o = n_j R_1/N$
 2. $\hat{E}_{2j} = n_j(1 - \hat{p}_o) = n_j(1 - R_1/N) = n_j(N - R_1)/N = n_j R_2/N$
 3. The test statistic compares what was observed in the data, O_{ij} 's to what would be expected under the null hypothesis, E_{ij} :

$$T.S. = \chi^2 = \sum_{i=1}^2 \sum_{j=1}^t (O_{ij} - \hat{E}_{ij})^2 / \hat{E}_{ij}$$

4. Under H_o , the distribution of the chi-square statistic converges to a Chi-squared distribution with $df=t-1$.
5. Reject H_o when
 - $\chi^2 \geq \chi^2_{\alpha, t-1} = qchisq(1 - \alpha, t - 1)$ with
 - $p-value = 1 - F(\chi^2) = 1 - pchisq(\chi^2, t - 1)$, where
 - F is the cdf of a Chi-squared distribution with $df=t-1$.
6. If any $\hat{E}_{ij} < 1$ or more than 20% of the $\hat{E}_{ij} < 5$, then the asymptotic result does not hold and an exact permutation test is required.
7. Why are the degrees of freedom $t-1$? *- Fisher Exact Test*

With no restrictions, there are t unknown parameters: p_1, p_2, \dots, p_t

Under the null hypothesis, $H_o : p_1 = \dots = p_t = p_o$, there is only 1 unknown parameter, p_o

The asymptotic distribution of the Pearson Chi-squared statistic has degrees of freedom

$$df = (\# \text{ unknown parameters under } H_o \cup H_1) - (\# \text{ unknown parameters under } H_o) = t - 1$$

The following SAS program and Output will illustrate the above discussion using the accident data.

```

OPTIONS PS=55 LS=75 NOCENTER NODATE;
DATA CAT;
INPUT Injury $ Location $ CNT @@;
CARDS;
I 1 35 NI 1 115
I 2 36 NI 2 114
I 3 46 NI 3 104
I 4 20 NI 4 130
I 5 24 NI 5 126
RUN;
PROC FREQ; TABLES Injury*Location/EXPECTED CELLCHI2 CHISQ;
WEIGHT CNT;
EXACT CHISQ;
RUN;

```

OUTPUT:

	Injury	Frequency	Expected	Cell Chi-Square	Location	Percent	Row Pct	Col Pct	1	2	3	4	5	Total
I									35	36	46	20	24	161
									32.2	32.2	32.2	32.2	32.2	
									0.2435	0.4484	5.9143	4.6224	2.0882	
									4.67	4.80	6.13	2.67	3.20	21.47
									21.74	22.36	28.57	12.42	14.91	
									23.33	24.00	30.67	13.33	16.00	
NI									115	114	104	130	126	589
									117.8	117.8	117.8	117.8	117.8	
									0.0666	0.1226	1.6166	1.2635	0.5708	
									15.33	15.20	13.87	17.33	16.80	78.53
									19.52	19.35	17.66	22.07	21.39	
									76.67	76.00	69.33	86.67	84.00	
Total									150	150	150	150	150	750
									20.00	20.00	20.00	20.00	20.00	100.00

Statistic	DF	Value	Prob
Chi-Square	4	16.9568	0.0020
Likelihood Ratio Chi-Square	4	17.1802	0.0018
Mantel-Haenszel Chi-Square	1	5.7027	0.0169
Phi Coefficient		0.1504	
Contingency Coefficient		0.1487	
Cramer's V		0.1504	

Pearson Chi-Square Test

Chi-Square	16.9568
DF	4
Asymptotic Pr > ChiSq	0.0020
Exact Pr >= ChiSq	0.0019

Special Case $t=2$

When the number of population proportions being compare is two ($t = 2$) and the sample sizes n_1 and n_2 are large, we can test hypotheses about p_1 and p_2 using the following test statistic. This test statistic is equivalent to the Pearson chi-square statistic and will allow us to make power calculations and sample size determinations.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$. Using a Central Limit Theorem, with $p_1 - p_2 = 0$, the distribution of Z is approximately $N(0, 1)$ for large values of n_1 and n_2 .

Consider the following 3 cases:

H1: $H_o : p_1 \leq p_2$ vs $H_1 : p_1 > p_2$

- a. Reject H_o if $Z \geq z_\alpha$
- b. $p-value = 1 - \Phi(z) = 1 - pnorm(z)$, where z is the value of Z computed from the observed data and Φ is the $N(0, 1)$ cdf.
- c. The Power function is given by

$$\gamma(p_1, p_2) = P[Z \geq z_\alpha] = 1 - \Phi \left[\frac{z_\alpha \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} - (p_1 - p_2)}{\sigma} \right]$$

$$\text{where } \sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \text{ and } p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$$

H2: $H_o : p_1 \geq p_2$ vs $H_1 : p_1 < p_2$

- a. Reject H_o if $Z \leq -z_\alpha$
- b. $p-value = \Phi(z) = pnorm(z)$, where z is the value of Z computed from the observed data.
- c. The Power function is given by

$$\gamma(p_1, p_2) = P[Z \leq -z_\alpha] = \Phi \left[\frac{-z_\alpha \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} - (p_1 - p_2)}{\sigma} \right]$$

H3: $H_o : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$

- a. Reject H_o if $|Z| \geq z_{\alpha/2}$
- b. $p-value = P(|Z| \geq |z|) = 2(1 - \Phi(|z|)) = 2(1 - pnorm(|z|))$,
where z is the value of Z computed from the observed data.
- c. The Power function is given by

$$\gamma(p_1, p_2) = P[|Z| \geq z_\alpha] = 1 - \Phi \left[\frac{z_{\alpha/2} \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} - (p_1 - p_2)}{\sigma} \right] + \Phi \left[\frac{-z_{\alpha/2} \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} - (p_1 - p_2)}{\sigma} \right]$$

EXAMPLE The article “Adjuvant Radiotherapy and Chemotherapy in Node-Positive Premenopausal Women with Breast Cancer”, *New England Journal of Medicine*, 1997, pp. 956-962, reported on a study designed to evaluate two therapies:

Therapy 1: treating cancer patients with chemotherapy only and Therapy 2: treating cancer patients with chemotherapy and radiation.

There were 154 cancer patients who were treated with chemotherapy only. Of these patients, 76 survived at least 15 years after treatment.

There were 164 cancer patients who were treated with chemotherapy and radiation. Of these patients, 98 survived at least 15 years after treatment.

Is there sufficient evidence at an $\alpha = .01$ level that including radiation provided an increased survival rate over the chemotherapy only treatment?

$$H_o : p_1 \geq p_2 \text{ vs } H_1 : p_1 < p_2$$

a. Reject H_o if $Z \leq -z_{.01} = -2.326$

$$\hat{p}_1 = 76/154 = .494 \text{ and } \hat{p}_2 = 98/164 = .598 \Rightarrow \hat{p} = (76 + 98)/(154 + 164) = .547$$

$$Z = \frac{.494 - .598}{\sqrt{.547(1-.547)\left(\frac{1}{154} + \frac{1}{164}\right)}} = -1.86 > -2.326$$

Thus we fail to reject H_o and conclude there is not significant evidence (p-value=.0314) that including radiation in the treatment has increased the survival rate.

$$\text{b. } p - \text{value} = P[Z \leq -1.86] = \text{pnorm}(-1.86) = .0314 > .01.$$

c. A 99% C.I. on $p_1 - p_2$ is given by

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

$$.494 - .598 \pm (2.576) \sqrt{\frac{.494(1-.494)}{154} + \frac{.598(1-.598)}{164}} = -.104 \pm .143 = (-.247, .039)$$

Multiple Comparisons of t Proportions

Person
Ch. spec'd
test of equality
(above)

We can use the results for $t = 2$ to further investigate the accident data example. Based on the test of homogeneity of proportions, we found there was significant evidence that the proportion of accidents resulting in an injury differed over the four locations. To further investigate this data, we would want to determine which pairs of locations had significantly different proportions. There are $\binom{5}{2} = 10$ pairs of locations. It would be necessary to conduct 10 tests of $H_0 : p_k = p_m$ vs $H_1 : p_k \neq p_m$ for $k \neq m = 1, 2, 3, 4, 5$.

Just as in the case of simultaneous confidence intervals which was discussed in Handout 11, it is necessary to account for multiple testing of the five location proportions. Thus, each of the 10 tests of hypotheses will be conducted using a per comparison (PC) Type I error rate of $\alpha_{pc} = \alpha_{Exp}/M$ where α_{Exp} is the desired overall Type I error rate and M is the number tests conducted. In our accident example, if we want $\alpha_{Exp} = .05$ then is necessary to conduct each of the 10 tests at a level of $\alpha_{pc} = .05/10 = .005$. This may result in an inflated Type II error rate due to the very low level at which each of the 10 tests is conducted.

The proportion of accidents resulting in an injury are given below:

		Location				
		1	2	3	4	5
Proportion Injured		.233	.240	.307	.133	.160

Using the p-values from $Z_{jk} = \frac{\hat{p}_j - \hat{p}_k}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_j} + \frac{1}{n_k}\right)}}$, where \hat{p} is the pooled estimator of $p = .2167$.

conduct the 10 pairwise comparisons of the proportions (p_j vs p_k) we obtain the following results:

Comparison	Comparison									
	1 vs 2	1 vs 3	1 vs 4	1 vs 5	2 vs 3	2 vs 4	2 vs 5	3 vs 4	3 vs 5	4 vs 5
$p_j - p_k$	-.0067	-.0733	.1000	.0733	-.0667	.1067	.0800	.1733	.1467	-.0267
$Z_{j,k}$	-.1406	-1.5468	2.1092	1.5468	-1.4061	2.2498	1.6874	3.6560	3.0935	-.5625
p-value	.892	.153	.025	.110	.195	.018	.083	.0003	.0027	.514

Using a per comparison error rate of $.05/10 = .005$, there are two pairs having p-value $\leq .005$,

"Loc. 3 vs Loc. 4" and "Loc. 3 vs Loc. 5".

Thus, there was not significant evidence that the locations (1, 2, 3) had significantly different proportions of accidents resulting in injuries.

Also, there was not significant evidence that the locations (1, 2, 4, 5) had significantly different proportions of accidents resulting in injuries.

Thus, we have now gained considerably more insight about the relationship between proportion of accidents resulting in injuries and the type of location where the accident occurs. Using the test of homogeneity of proportions, it was determined that there was a difference in the five proportions but the test did not specify which locations were similar and which were different.

Sample Size Calculations

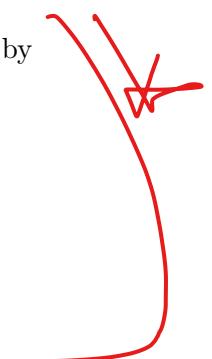
To determine the appropriate sample sizes, we will only consider the case $n_1 = n_2 = n$. 

Specifications:

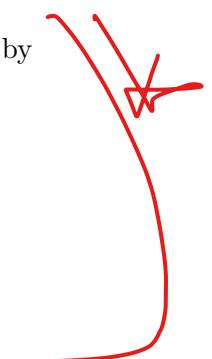
1. Level of Test: α
2. Type II error rate, β at a specified set of values for p_1 and p_2 .

That is, we want to determine the minimum sample size n , such that, an α level test will have a probability of Type II error of at most β if the true proportions are the given values of p_1 and p_2 .

3. Let $p_1 - p_2 = D$
4. The appropriate sample size for a 1-sided test ($H_1 : p_1 > p_2$ or $H_1 : p_1 < p_2$) is given by

$$n = \frac{\left[z_\alpha \sqrt{(p_1 + p_2)(2 - p_1 - p_2)/2} + z_\beta \sqrt{p_1(1 - p_1) + p_2(1 - p_2)} \right]^2}{D^2}$$


5. The appropriate sample size for a 2-sided test ($H_1 : p_1 \neq p_2$) is given by

$$n = \frac{\left[z_{\alpha/2} \sqrt{(p_1 + p_2)(2 - p_1 - p_2)/2} + z_\beta \sqrt{p_1(1 - p_1) + p_2(1 - p_2)} \right]^2}{D^2}$$


EXAMPLE A researcher wants to determine the minimum sample sizes $n_1 = n_2 = n$ such that an $\alpha = .05$ test of $H_1 : p_1 > p_2$ would have maximum probability of Type II error of at most $\beta = .1$ if the true values of p_1 and p_2 are $p_1 = .0003$ and $p_2 = .00015$.

Using the above formula we obtain

$$n = \frac{\left[z_{.05} \sqrt{(.0003 + .00015)(2 - .0003 - .00015)/2} + z_{.1} \sqrt{.0003(1 - .0003) + .00015(1 - .00015)} \right]^2}{(.0003 - .00015)^2} = 171072.1$$

Thus, we would need $n_1 = n_2 = 171072$ experimental units for each of the two sets of Bernoulli trials.

The power of the test at these values: $\alpha = .05$, $n_1 = n_2 = 171072$, $p_1 = .0003$ and $p_2 = .00015$ would be calculated as follows:

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = .000225 \Rightarrow \sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = .000036262$$

$$\gamma(p_1, p_2) = \gamma(.0003, .00015) = P[Z \geq z_{.05}] = P[Z \geq 1.645] = 1 - \Phi \left[\frac{z_{.05} \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} - (p_1 - p_2)}{\sigma} \right] \Rightarrow$$

$$\gamma(.0003, .00015) = 1 - \Phi \left[\frac{1.645 \sqrt{.000225(1 - .000225) \left(\frac{2}{171072} \right)} - (.0003 - .00015)}{.000036262} \right] \Rightarrow$$

$$\gamma(.0003, .00015) = 1 - pnorm(-1.279999) = .8997272 \approx .9 = 1 - .1$$

Using the R function

`power.prop.test(n=,p1=.0003,p2=.00015,sig.level=.05,power=.9,alternative=c("one.sided"))`

We obtain $n=171236.5$ and hence use $n = 171,237$



Fisher Exact Test - Permutation Test for Small Sample Sizes

When the condition for using the Chi-square approximation to the distribution of the Test Statistic is invalid, an inference will be derived which depends on an exact distribution rather than applying an asymptotic result that was required in the Chi-square test.

We will illustrate these methods using the comparison of two population proportions. This test is referred to as the *Fisher Exact Test*. The hypotheses to be tested are $H_0 : p_1 \leq p_2$ versus $H_1 : p_1 > p_2$, where p_i is the probability of "success" for population i , $i=1,2$. In developing a small-sample test of hypotheses, we need to develop the exact probability distribution for the cell counts in all 2×2 tables having the same row and column totals as the 2×2 table from the observed table:

Population	Outcome		Total
	Success	Failure	
1	X	$n_1 - X$	n_1
2	Y	$n_2 - Y$	n_2
Total	m	$n - m$	n

For tables having the same row and column totals: $n_1, n_2, m, n - m$, the value of X determines the counts for the remaining three cells.

Thus, we can use the hypergeometric distribution to express the distribution that applies to the sets of 2×2 tables having fixed row and column totals. When $p_1 = p_2$, the probability of observing a particular value for x , that is, the probability of a particular table being observed, is given by

$$P(X = x | X + Y = m) = \frac{\binom{n_1}{x} \binom{n_2}{m-x}}{\binom{n}{m}} \quad \text{⊗}$$

To test the difference in the two population proportions, the p-value of the test is the sum of hypergeometric probabilities for outcomes at least as in support of the alternative hypothesis as the observed table. For $H_1 : p_1 > p_2$, we need to determine which other possible 2×2 tables would provide stronger support of H_1 than the observed table. Given the marginal totals, $n_1, n_2, m, n - m$, tables having larger X values will have larger values for \hat{p}_1

and hence provide stronger evidence in favor of $p_1 > p_2$.

Let x be the observed value of X . Then, any integer greater than x would provide even greater evidence that $p_1 > p_2$. The possible values of x are $0, 1, \dots, \min(n_1, m)$. Therefore, we have that

$$\text{p-value} = P[X \geq x] = \sum_{k=x}^{\min(n_1, m)} \frac{\binom{n_1}{k} \binom{n_2}{m-k}}{\binom{n}{m}}$$

Using R we obtain: $p\text{-value} = 1 - \text{phyper}(x - 1, n_1, n_2, m)$ 

For the two sided alternative: $H_1 : p_1 \neq p_2$, the p-value is defined as the sum of the probabilities of tables no more likely than the observed table. Thus, the p-value is the sum of the probabilities of all values of $X = k$ for which $P(k) \leq P(x)$ where x is the observed value of X .

Note: The textbook has a formula for p-value but it is only valid if $n_1 = n_2$ and $m = n - m$.

The power function for the Fisher Exact test (with $n_1 = n_2$) is given by

$$\gamma(n, p_1, p_2) = \sum_{r=0}^{2n} \sum_{x \in C_r} \binom{n}{x} \binom{n}{r-x} p_1^x (1-p_1)^{n-x} p_2^{r-x} (1-p_2)^{n-r+x}$$

where r is the total number of observed successes in the two groups, x is the number of observed success in Group 1, and C_r is the critical region of the Fisher Exact Test.

Example of Applying Fisher Exact Test

Example 9.6 in the textbook considers the results from a clinical trial for comparing two drug therapies for leukemia: P and PV. Twenty-one patients were assigned to drug P and forty-two patients to drug PV. The following table summarizes the success of both drugs:

Drug	Outcome		Total
	Success	Failure	
PV	38	4	42
P	14	7	21
Total	52	11	63

1. Compute the p-value for testing $H_0 : p_1 \geq p_2$ vs $H_1 : p_1 < p_2$,

where p_1, p_2 are the probabilities patient under Drug P, PV, respectively, goes into remission.

The probability of the observed table is

$$P(X = x) = \frac{\binom{42}{38} \binom{21}{14}}{\binom{63}{52}} = .0211 = dhyper(38, 42, 21, 52)$$

Thus, the one-sided p-value is the sum of the probabilities for all tables having 38 or more successes:

$$\begin{aligned} \text{p-value} = P[X \geq 38] &= \sum_{k=38}^{\min(42, 52)} \frac{\binom{42}{k} \binom{21}{52-k}}{\binom{63}{52}} \\ &= .02114 + .00379 + .00041 + .00002 + 0 = .02536 \end{aligned}$$

Using R: $p-value = 1 - phyper(x - 1, n_1, n_2, m) = 1 - phyper(38 - 1, 42, 21, 52) = .02536501$

2. Compute the p-value for testing $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$

The p-value for the two-sided alternative hypothesis is the sum of the probabilities for all tables having probability of occurring that are less than the probability of the observed table: $P(38) = .0211$:

k	0	1	.	30	31	32	33	34	35	36
P(k)	.000000	.000000	.	.000000	.006951	.050180	.152060	.254925	.262208	.173349
k	37	38	39	40	41	42	43	.	51	52
P(k)	.074962	.021136	.003794	.000411	.000024	.000001	.000000	.	.000000	.000000

Thus, the two-sided p-value is $.021136 + .006951 + .003794 + .000411 + .000024 + .000001 = .032317$.

Using the R code we obtain p-value for Fisher's Exact test:

```
M = matrix(c(38,14,4,7),nrow=2,byrow=F)
fisher.test(M,alternative="g")
.02537
fisher.test(M,alternative="two.sided")
.03232
```

SAS CODE to compute the Chi-square test and the value of the Fisher's Exact Test:

```
ods html;
ods graphics on;
OPTIONS PS=55 LS=75 NOCENTER NODATE;
TITLE 'TWO LEUKEMIA THERAPIES';
DATA LEU;
INPUT DRUG $ OUTCOME $ CNT @@;
CARDS;
PV   S   38
PV   F   4
P    S   14
P    F   7
RUN;
PROC FREQ ORDER=DATA;
TABLES DRUG*OUTCOME/CHISQ EXPECTED CELLCHI2  RISKDIFF MEASURES;
  WEIGHT CNT;
EXACT RISKDIFF ;
RUN;
ods graphics off;
ods html close;
```

The SAS output is given on the next page.

TWO LEUKEMIA THERAPIES

The FREQ Procedure

Frequency Expected Cell Chi-Square Percent Row Pct Col Pct	Table of DRUG by OUTCOME			
	DRUG	OUTCOME		
		S	F	Total
PV	38 34.667 0.3205 60.32 90.48 73.08	4 7.3333 1.5152 6.35 9.52 36.36		42 66.67
P	14 17.333 0.641 22.22 66.67 26.92	7 3.6667 3.0303 11.11 33.33 63.64		21 33.33
Total	52 82.54	11 17.46		63 100.00

Statistics for Table of DRUG by OUTCOME

Statistic	DF	Value	Prob
Chi-Square	1	5.5070	0.0189
Likelihood Ratio Chi-Square	1	5.2010	0.0226
Continuity Adj. Chi-Square	1	3.9788	0.0461
Mantel-Haenszel Chi-Square	1	5.4196	0.0199
Phi Coefficient		0.2957	
Contingency Coefficient		0.2835	
Cramer's V		0.2957	
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	38
Left-sided Pr <= F	0.9958
Right-sided Pr >= F	0.0254

Table Probability (P)	0.0211
Two-sided Pr <= P	0.0323

Column 1 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits	(Exact) 95% Confidence Limits		
Row 1	0.9048	0.0453	0.8160	0.9935	0.7738	0.9734
Row 2	0.6667	0.1029	0.4650	0.8683	0.4303	0.8541
Total	0.8254	0.0478	0.7317	0.9191	0.7090	0.9095
Difference	0.2381	0.1124	0.0178	0.4584	-0.0356	0.4891

Difference is (Row 1 - Row 2)

Column 2 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits	(Exact) 95% Confidence Limits		
Row 1	0.0952	0.0453	0.0065	0.1840	0.0266	0.2262
Row 2	0.3333	0.1029	0.1317	0.5350	0.1459	0.5697
Total	0.1746	0.0478	0.0809	0.2683	0.0905	0.2910
Difference	-0.2381	0.1124	-0.4584	-0.0178	-0.4891	0.0356

Difference is (Row 1 - Row 2)

Sample Size = 63

STOP

CN

Monday 11/22/21

Model II: Multinomial Sampling

Pearson Chi-squared Test of Independence

A social scientist wants to determine if there is an association between political affiliation and annual income level. A random sample of 592 registered voters is selected and each of the selected individuals was asked their political affiliation and their annual income. These values were then summarized in the following table:

	Income Level				Total
	(0, 30K)	[30K, 50K)	[50K, 100K)	[100K, ∞)	
Republican	20	84	17	94	215
Democrat	68	119	26	7	220
Independent	5	29	14	16	64
Other	15	54	14	10	93
Total	108	286	71	127	592

Note the difference between the research question posed above and the research question posed in the Homogeneity of Proportions research question. Also, there is a difference between the sampling schemes in the two settings.

In the test of Homogeneity of Proportions, there were t populations and a random sample of N_i units was selected from the i th population, $i = 1, 2, \dots, t$. These N_i units were then classified into r categories. For example, it has been claimed that the opinion on gun control is very different depending on the age of the person. To evaluate this claim, a random sample of 1000 members are selected from each of the three age groups: A_1 : 18-25, A_2 : 26-50, and A_3 : over 50; using registered voters lists. Within each age group, the 1000 people are asked their opinion on whether gun laws should be strengthened: Strongly Disagree, Disagree, Agree, Strongly Agree. The research question is then framed into

$$H_0 : (p_{11}, p_{12}, p_{13}, p_{14})_{A_1} = (p_{21}, p_{22}, p_{23}, p_{24})_{A_2} = (p_{31}, p_{32}, p_{33}, p_{34})_{A_3} \text{ versus}$$

$$H_1 : \text{At least two of } (p_{11}, p_{12}, p_{13}, p_{14})_{A_1}; (p_{21}, p_{22}, p_{23}, p_{24})_{A_2}; (p_{31}, p_{32}, p_{33}, p_{34})_{A_3} \text{ are not equal}$$

This is a **Product Multinomial Model**: The product of t multinomial distributions with r categories. In the example, we have $t = 3$ populations (age groups) with four categories of Opinion on gun control. There is a separate random sample selected from each of the t populations. There are t multinomial random variable each having r levels.

In the **Test of Independence** of two variables setting, a single random sample of N units is selected from a single population and the same two characteristics are measured on each unit. In the above example, it would be Party Affiliation and Income Level.

In the test of **Independence Problem**, we have two categorical random variables, X and Y having r and c levels, respectively.

Thus, we have a single multinomial random variable with rc levels.

Let $p_{ij} = P[X = X_i, Y = Y_j]$ be the probability that a randomly selected unit falls in the (i, j) cell of the $r \times c$ contingency table.

Let $p_{i\cdot} = P[X = X_i]$ and $p_{\cdot j} = P[Y = Y_j]$ be the **marginal** probabilities associated with X and Y , respectively.

The null hypothesis is that X and Y are independent, that is,

$$P[X = X_i, Y = Y_j] = P[X = X_i]P[Y = Y_j] \text{ that is, } p_{ij} = p_{i.}p_{.j} \text{ for all pairs } (i, j)$$

The alternative hypothesis is X and Y are not independent, :

$$H_o : p_{ij} = p_{i.}p_{.j} \text{ for all } (i, j) \text{ vs } H_1 : p_{ij} \neq p_{i.}p_{.j} \text{ at least one pair } (i, j)$$

Randomly select N units from the population and classify them into the $r \times c$ cells according to their values of (X, Y) , yielding the following contingency table:

	Y_1	Y_2	Y_3	\dots	Y_c	Total
X_1	O_{11}	O_{12}	O_{13}	\dots	O_{1c}	$n_{1.}$
X_2	O_{21}	O_{22}	O_{23}	\dots	O_{2c}	$n_{2.}$
X_3	O_{31}	O_{32}	O_{33}	\dots	O_{3c}	$n_{3.}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
X_r	O_{r1}	O_{r2}	O_{r3}	\dots	O_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	\dots	$n_{.c}$	N

The expected counts would need to be computed next in order to apply the Chi-square test:

- Under the null hypothesis that X_i and Y_j are independent, we have

$$p_{ij} = p_{i.}p_{.j}, \text{ for all } (i, j) \text{ pairs}$$

Therefore, the expected counts are computed under H_o as follows:

1. $E_{ij} = Np_{ij} = Np_{i.}p_{.j}$, which are unknown

2. Estimate $p_{i.}$ and $p_{.j}$ by

$$\hat{p}_{i.} = n_{i.}/N \text{ and } \hat{p}_{.j} = n_{.j}/N$$

3. Estimate E_{ij} by $\hat{E}_{ij} = N\hat{p}_{i.}\hat{p}_{.j} = N\frac{n_{i.}}{N}\frac{n_{.j}}{N} = \frac{n_{i.}n_{.j}}{N} = \frac{(\text{ith row total})(\text{jth column total})}{N}$

4. The test statistic compares what was observed in the data, O_{ij} 's to what would be expected under the null hypothesis, \hat{E}_{ij} :

$$T.S. = \chi^2 = \sum_{i=1}^r \sum_{j=1}^c (O_{ij} - \hat{E}_{ij})^2 / \hat{E}_{ij}$$

4. Under H_o , the distribution of the chi-square statistic converges to a Chi-squared distribution with $df = (r-1)(c-1)$

5. Reject H_o when

- $\chi^2 \geq \chi^2_{\alpha, df}$ with

- $p-value = 1 - F(\chi^2) = 1 - pchisq(\chi^2, (r-1)(c-1))$, where

- F is the cdf of a Chi-squared distribution with $df = (r-1)(c-1)$

6. If any $\hat{E}_{ij} < 1$ or more than 20% of the $\hat{E}_{ij} < 5$, then the asymptotic result does not hold and an exact permutation test is required.

7. Why are the degrees of freedom $(r-1)(c-1)$?
- With no restrictions, there are $rc - 1$ unknown parameters:
 $rc - 1$ unknown proportions: $p_{ij}, i = 1, \dots, r; j = 1, \dots, c$ because $\sum_{i=1}^r \sum_{j=1}^c p_{ij} = 1$
 - Under the null hypothesis, $H_o : p_{ij} = p_i p_{.j}$, and therefore there are $(r-1+c-1)$ unknown parameters because
 - (r-1) unknown p_i because $\sum_{i=1}^r p_i = 1$ and
 - (c-1) unknown $p_{.j}$ because $\sum_{j=1}^c p_{.j} = 1$
 - Therefore, $df = (\# \text{ parameters under } H_o \cup H_1) - (\# \text{ parameters under } H_o)$
 $= (rc - 1) - (r-1 + c-1) = (r-1)(c-1)$
 - Thus the asymptotic distribution of the Pearson Chi-squared statistic has degrees of freedom, $df = (r-1)(c-1)$

The calculations are very similar for the Pearson Chi-square Test of Homogeneity of Proportions and the Pearson Chi-square Test of Independence.

However, the data collection design and the hypotheses being tested are quite different.

- In the Test of Independence: A random sample of N units is selected from a population consisting of units which are classified according to two types of categories having r and c levels, respectively. p_{ij} is the proportion of the populations units which fall into the i th level of Category I and the j th level of Category II. The hypotheses being tested are

$$H_o : p_{ij} = p_i p_{.j} \text{ for all } (i, j) \text{ vs } H_1 : p_{ij} \neq p_i p_{.j} \text{ for some } (i, j)$$

- In the Test of Homogeneity of Proportions: Suppose we have r populations. In each population we have c types of units with proportions p_{ij} , where p_{ij} is the proportion of the i th population's units which are of Type j , $j = 1, 2, \dots, c$. A random sample of n_i units is selected from the i th population and the number of units of each type are counted, O_{ij} , $j = 1, 2, \dots, c$ where $\sum_{j=1}^c O_{ij} = n_i$. The hypotheses being tested are

$$H_o : (p_{11}, p_{12}, \dots, p_{1c}) = (p_{21}, p_{22}, \dots, p_{2c}) = \dots = (p_{r1}, p_{r2}, \dots, p_{rc}) \text{ vs}$$

$$H_1 : (p_{i1}, p_{i2}, \dots, p_{ic}) \neq (p_{k1}, p_{k2}, \dots, p_{kc}), \text{ for at least two sets of proportions } i, k.$$

- In the Test of Homogeneity of Proportions the degrees of freedom for the Chi-square test are the same, $df = (r - 1)(c - 1)$ but the justification is different:

- With no restrictions, there are $r(c - 1)$ unknown parameters:
 $c - 1$ unknown proportions: $p_{ij}, j = 1, \dots, c$; for each of the r populations because $\sum_{j=1}^c p_{ij} = 1$ for each value of $i = 1, \dots, r$
- Under the null hypothesis, H_o , all r vectors $(p_{i1}, p_{i2}, \dots, p_{ic})$ are equal to $(p_{o1}, p_{o2}, \dots, p_{oc})$, but the c values of p_{oj} are unknown. Therefore there are $c-1$ unknown parameters under H_o because $\sum_{j=1}^c p_{oj} = 1$
- Therefore, $df = (\# \text{ parameters under } H_o \cup H_1) - (\# \text{ parameters under } H_o)$
 $= r(c - 1) - (c-1) = (r-1)(c-1)$
- Thus the asymptotic distribution of the Pearson Chi-squared statistic has degrees of freedom, $df = (r-1)(c-1)$

The same SAS program that was used to analyze the accident data would be used for the Political Affiliation study:

```

OPTIONS PS=55 LS=75 NOCENTER NODATE;
DATA POL;
INPUT Party $ Income $ CNT @@;
CARDS;
Rep I1 20      Dem I1 68     Ind I1 5     Other I1 15
Rep I2 84      Dem I2 119    Ind I2 29    Other I2 54
Rep I3 17      Dem I3 26     Ind I3 14    Other I3 14
Rep I4 94      Dem I4 7      Ind I4 16    Other I4 10
RUN;
PROC FREQ; TABLES Party*Income/EXPECTED CELLCHI2 CHISQ;
WEIGHT CNT;
*EXACT CHISQ;
RUN;

```

SAS Output is given next:

The FREQ Procedure					
Table of Party by Income					
Party	Income				
Frequency	I1	I2	I3	I4	Total
Dem	68	119	26	7	220
	40.135	106.28	26.385	47.196	
	19.346	1.5214	0.0056	34.234	
	11.49	20.10	4.39	1.18	37.16
	30.91	54.09	11.82	3.18	
	62.96	41.61	36.62	5.51	
Ind	5	29	14	16	64
	11.676	30.919	7.6757	13.73	
	3.8169	0.1191	5.2109	0.3754	
	0.84	4.90	2.36	2.70	10.81
	7.81	45.31	21.88	25.00	
	4.63	10.14	19.72	12.60	
Other	15	54	14	10	93
	16.966	44.929	11.154	19.951	
	0.2279	1.8314	0.7263	4.9633	
	2.53	9.12	2.36	1.69	15.71
	16.13	58.06	15.05	10.75	
	13.89	18.88	19.72	7.87	
Rep	20	84	17	94	215
	39.223	103.87	25.785	46.123	
	9.4211	3.8005	2.9933	49.697	
	3.38	14.19	2.87	15.88	36.32
	9.30	39.07	7.91	43.72	
	18.52	29.37	23.94	74.02	
Total	108	286	71	127	592
	18.24	48.31	11.99	21.45	100.00

Statistics for Table of Party by Income

Statistic	DF	Value	Prob
Chi-Square	9	138.2898	<.0001
Likelihood Ratio Chi-Square	9	146.4436	<.0001
Mantel-Haenszel Chi-Square	1	89.4571	<.0001
Phi Coefficient		0.4833	
Contingency Coefficient		0.4352	
Cramer's V		0.2790	

Sample Size = 592

From the above results, we have that there is significant evidence (p-value<.0001) that Party Affiliation and Income level are not independent. Thus, the Income distribution for one party may be quite different from the Income distribution for a second party. In particular, Republicans have over 40% of their member in the upper two income levels, whereas Democrats have approximately 15% in the upper two income levels.

Odds Ratio and Relative Risk

The are numerous measures of association which are used to evaluate the strength of association in a contingency table. In particular, for 2×2 tables, the most widely used measure is the **Odds Ratio**, and a related measure of association, the **Relative Risk**.

Suppose we have two breeds of dogs and want to compare the frequency of occurrence of disease in the two breeds. Random sample of size $n_1.$ and $n_2.$ are selected from a registry of the two breeds. Suppose we obtain the following table:

	Has Disease		Total	Proportion with Disease
	Yes	No		
Breed 1	n_{11}	n_{12}	$n_{1.}$	$\hat{p}_1 = n_{11}/n_{1.}$
Breed 2	n_{21}	n_{22}	$n_{2.}$	$\hat{p}_2 = n_{21}/n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n	

Suppose that p_1 and p_2 are the population proportion of dogs having the disease from breed 1 and breed 2, respectively.

DEFINITION The **odds** compares the proportion of the population having a trait (the probability that an event occurs) to the proportion of the population not having the trait (probability that an event does not occur). That is, if p is the proportion of a population having a given trait ($P[\text{Event Occurs}]$) then

$$\text{Odds of Trait} = \frac{p}{1-p} = \frac{P[\text{Event Occurs}]}{P[\text{Event Does Not Occur}]}$$

In our example, we would estimate the odds of Breed 1 having the disease by

$$\widehat{\text{Odds}} = \frac{\hat{p}_1}{1-\hat{p}_1} = \frac{n_{11}/n_{1.}}{1-n_{11}/n_{1.}} = \frac{n_{11}/n_{1.}}{n_{12}/n_{1.}} = \frac{n_{11}}{n_{12}}$$

DEFINITION The **odds ratio** compares the odds of having the trait in one group to the odds of having the trait in a second group. That is, if p_1 and p_2 are the proportions in population 1 and population 2, respectively, having a given trait then

$$\text{Odds Ratio} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{\text{Odds in Population 1}}{\text{Odds in Population 2}}$$

In our example, we would estimate the odds ratio of having the disease for the two breeds as follows

$$\widehat{\text{OR}} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

- 1. If any of the $n_{ij} = 0$ then add .5 to all the n_{ij} prior to computing the OR. However, use the n_{ij} without .5 added to them in performing the Fisher Exact test.
- 2. Odds ratio range from 0 to ∞
- 3. When $\text{OR}=1$, then $p_1 = p_2$, that is, there is no difference between the two groups wrt to occurrence of trait.
- 4. When $\text{OR} < 1$, then $p_1 < p_2$, that is, Group 1 is less likely than Group 2 to have the trait.

$$\text{OR} < 1 \Rightarrow \frac{p_1}{1-p_1} < \frac{p_2}{1-p_2} \Rightarrow p_1 - p_1 p_2 < p_2 - p_1 p_2 \Rightarrow p_1 < p_2$$

An alternative representation of the Odds and Odds Ratio is used in logistic regression:

DEFINITION The *logit* for the proportion p is given by

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(\text{Odds})$$

The log of the Odds Ratio can be represented as the difference between $\text{logit}(p_1)$ and $\text{logit}(p_2)$:

$$\begin{aligned} L &= \log(\text{OR}) \\ &= \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right) \\ &= \log\left(\frac{p_1}{(1-p_1)}\right) - \log\left(\frac{p_2}{(1-p_2)}\right) \\ &= \text{logit}(p_1) - \text{logit}(p_2) \end{aligned}$$

In the courses STAT 659, STAT 608, and STAT 645, the logistic regression model will be developed in detail. This is a method to model categorical data as a function of explanatory variables. There are many models other than the logit model. See STAT 659 or the book, *Categorical Data Analysis*, by Alan Agresti for further details.

For example, suppose we wanted to study the proportion, p of dogs having contracted heart worms as a function the dog's breed, age, weight, and sex. We evaluate a randomly selected group of n dogs and record

$Y=1$ if dog has heart worms and $Y=0$ if dog does not have heart worms.

Also, the values of $X_1 = \text{breed}$, $X_2 = \text{age}$, $X_3 = \text{weight}$, and $X_4 = \text{sex}$ are recorded for each dog.

The data could be modeled using the logit model:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

or equivalently as

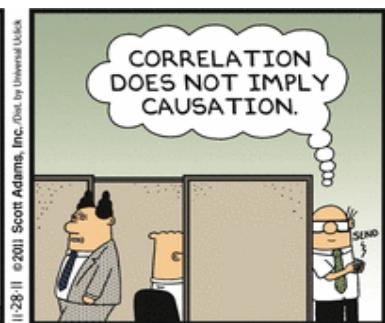
$$\frac{p}{1-p} = e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$$

which implies

$$p = \frac{e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}{1 + e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}$$

Using non-linear regression, estimates of the coefficients, β_i are obtained.

The results we obtain from observational data using statistical testing and estimation methodology should be interpreted with great care. See the cartoon on next page.



11-28-11 © 2011 Scott Adams, Inc. (Used by Universal Uclick)

From a Central Limit Theorem, we have that $\hat{L} = \log(\widehat{OR})$ has an asymptotic normal distribution with asymptotic mean and variance:

$$\hat{\mu}_L = \log(OR) \quad \text{and} \quad \text{AVAR}(L) = \hat{\sigma}_L^2 = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right)$$

That is, $\hat{L} = \log(\widehat{OR})$ has approximately a $N(\hat{\mu}_L, \hat{\sigma}_L^2)$ distribution for large n .

Thus, an approximate $100(1 - \alpha)\%$ C.I. for the Odds Ratio,

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

is obtained as follows:

$$\left[e^{\left(\hat{\mu}_L - z_{\alpha/2} \sqrt{\hat{\sigma}_L^2} \right)}, e^{\left(\hat{\mu}_L + z_{\alpha/2} \sqrt{\hat{\sigma}_L^2} \right)} \right]$$

where $\hat{L} = \text{logit}(\hat{p}_1) - \text{logit}(\hat{p}_2)$ and \hat{p}_1, \hat{p}_2 are the sample estimators of p_1 and p_2 , respectively.

For **retrospective or prospective studies**, often the type of study used in epidemiological studies, it is often more meaningful to use the parameter *Relative Risk*

$$RR = \frac{p_1}{p_2}$$

to compare the risk of developing a particular condition (often disease) for one group in comparison to a second group. The estimate of RR can be expressed as a function of the estimated value of the odds ratio:

$$\widehat{RR} = \widehat{OR} \times \frac{1 + n_{21}/n_{22}}{1 + n_{11}/n_{12}}$$

Note that OR and RR are nearly equal when the outcome (condition, disease) is a *rare outcome*. That is, when n_{11} and n_{21} are small relative to n_{12} and n_{22} , respectively.

When we are dealing with **cross-sectional** data, the RR is often referred to as the *prevalence ratio*. In this type of data, RR is not assessing the risk of a disease, because the disease and risk factor are assessed at the same time, but RR provides an assessment of the prevalence of a disease (or condition or opinion) in one group compared to a second group.

EXAMPLE In the article, “A multiple testing procedure for clinical trials,” *Biometrics*, **35**, pp. 549-556, the data for a randomized clinical trial for comparing Prednisone and Prednisone+VCR drug therapies for treating patients with leukemia is presented. The data is given in the following table:

Drug Type	Success	Failure	Total
Prednisone	13	8	21
Prednisone+VCR	38	4	42
Total	52	11	63

The following SAS code was used to analyze the data:

```
*The following program will calculate the ;
*Chi-square Homogeneity of Proportions Test and
*CI for difference
*for Example 9.6 from the Textbook;
*DrugTherapy.sas
```

```
ods html;
ods graphics on;
*The following program will calculate the;
*Chi-square Homogeneity of Proportions Test;
*for Example 9.6 from the Textbook;
OPTIONS PS=55 LS=75 NOCENTER NODATE;
TITLE 'TWO LEUKEMIA THERAPIES';

DATA LEU;
INPUT DRUG $ OUTCOME $ CNT @@;
CARDS;
PVCR S 38
PVCR F 4
P   S 13
P   F 8
RUN;

PROC FREQ ORDER=DATA;
TABLES DRUG*OUTCOME/CHISQ EXPECTED CELLCHI2 ALPHA=.01 RISKDIFF MEASURES;
WEIGHT CNT;

EXACT RISKDIFF ;

RUN;
ods graphics off;
ods html close;
```

The output from the above program is given on the next page.

TWO LEUKEMIA THERAPIES

The FREQ Procedure

Frequency Expected Cell Chi-Square Percent Row Pct Col Pct	Table of DRUG by OUTCOME				
	DRUG	OUTCOME			
		S	F	Total	
	PVCR	38	4	42	
		34	8		
		0.4706	2		
		60.32	6.35	66.67	
		90.48	9.52		
		74.51	33.33		
P	13	8	21		
	17	4			
	0.9412	4			
	20.63	12.70	33.33		
	61.90	38.10			
	25.49	66.67			
Total	51	12	63		
	80.95	19.05	100.00		

Statistics for Table of DRUG by OUTCOME

Statistic	DF	Value	Prob
Chi-Square	1	7.4118	0.0065
Likelihood Ratio Chi-Square	1	7.0235	0.0080
Continuity Adj. Chi-Square	1	5.6746	0.0172
Mantel-Haenszel Chi-Square	1	7.2941	0.0069
Phi Coefficient		0.3430	
Contingency Coefficient		0.3244	
Cramer's V		0.3430	47
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	38
Left-sided Pr <= F	0.9986
Right-sided Pr >= F	0.0099
Table Probability (P)	0.0085
Two-sided Pr <= P	0.0141

47(A)

Column 1 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits	(Exact) 95% Confidence Limits		
Row 1	0.9048	0.0453	0.8160	0.9935	0.7738	0.9734
Row 2	0.6190	0.1060	0.4113	0.8267	0.3844	0.8189
Total	0.8095	0.0495	0.7126	0.9065	0.6909	0.8975
Difference	0.2857	0.1152	0.0598	0.5116	0.0126	0.5316
Difference is (Row 1 - Row 2)						
Column 2 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits	(Exact) 95% Confidence Limits		
Row 1	0.0952	0.0453	0.0065	0.1840	0.0266	0.2262
Row 2	0.3810	0.1060	0.1733	0.5887	0.1811	0.6156
Total	0.1905	0.0495	0.0935	0.2874	0.1025	0.3091
Difference	-0.2857	0.1152	-0.5116	-0.0598	-0.5316	-0.0126
Difference is (Row 1 - Row 2)						
Estimates of the Relative Risk (Row1/Row2)						
Type of Study	Value	95% Confidence Limits				
Case-Control (Odds Ratio)	5.8462	1.5074	22.6734			
Cohort (Col1 Risk)	1.4615	1.0304	2.0731			
Cohort (Col2 Risk)	0.2500	0.0849	0.7361			
Sample Size = 63						

The Output from SAS has the following interpretation. The Pearson chi-squared p-value=0.0065 indicates there is substantial evidence of a difference in the success proportions for the two treatments. However, this result cannot be taken as valid because the small sample sizes, 42 and 21, have resulted in one of the four cells (25% of the cells) having expected count less than 5. Thus, the Fisher's exact test should be used. From the Fisher's exact test the p-value is .0141 which is considerably larger than the value from the Pearson chi-squared p-value but it is still less than .05 so we can still conclude that there is substantial evidence of a difference in the success proportions for the two treatments.

Next, we would investigate the size of the difference, the effect size. From the SAS output, 95% confidence intervals on the success probabilities are obtained:

(.774, .973) for p_1 , PVCR and (.384, .819) for p_2 , Prednisone.

Note that these two C.I.'s overlap which would seem to indicate that there is not significant evidence of a difference in p_1 and p_2 .

However, the 95% confidence interval on the difference, $p_1 - p_2$ is (.013, .532) which does not contain 0 which is consistent with the conclusion obtained from the Fisher's Exact test. That is, there is significant evidence of a difference in p_1 and p_2 .

This type of result often occurs. A test of the difference in two parameters using individual C.I.'s on the two parameters will often not detect a difference by overlapping. However, when a test of the difference in the two parameters is conducted or a C.I. is placed on the difference in the two parameters, significant evidence will be found of a difference in the two parameters. The individual C.I.'s are not as sensitive (powerful) in detecting differences in comparison to procedures which are based on the difference in the two parameters.

The Odds Ratio is given as 5.8462 with a 95% C.I. of (1.5074, 22.6734). We would interpret the OR as indicating that the odds of Success for the drug therapy Prednisone+VCR is much higher than the odds of Success using Prednisone. That is, Prednisone+VCR has odds of about 5.8 times the odds of Prednisone by itself.

The value 1.4615 is the ratio (RR) of the Success probability for Prednisone+VCR to the Success probability for Prednisone. This would indicate that Prednisone+VCR is providing a larger Success probability than the drug therapy consisting of just Prednisone. Note that the C.I. for RR, (1.0304, 2.0731), does not contain 1.0 and hence we would conclude that with $\alpha = .05$ there is substantial evidence that RR differs from 1.0.

You can alter the level of coverage for the C.I. for the individual probabilities and the OR by including the ALPHA = option in the TABLES statement.

STOP NOW

Sensitivity - Specificity - Bayes Theorem

Two measures often used when determining the efficacy of screening tests for pathogens, diseases, or dangerous materials, are sensitivity and specificity. Suppose we are evaluating a diagnostic test for detecting a particular event (disease, hazardous compound, pathogen). A positive result from the test (+) indicates that the event is present and a negative result from the test (-) indicates that the event is not present. Therefore, we have four possibilities:

(Test=+, Event Present), (Test=+, Event Absent), (Test=-, Event Present), (Test=-, Event Absent).

The terms Prevalence, Sensitivity, and Specificity are defined as follows:

DEFINITION

- The **Prevalence** of an Event is proportion of the population possessing the trait:

$$\text{Prevalence} = P[\text{Event}]$$

- The **Sensitivity** of a screening test is proportion of tests which result in a positive test result given the Event is present : Sensitivity = $P[\text{Test} = + | \text{EventPresent}]$
- The **Specificity** of a screening test is proportion of tests which result in a negative test result given the Event is not present : Specificity = $P[\text{Test} = - | \text{EventAbsent}]$

Suppose a screening test is newly developed to determine if *e. coli* is present in a meat sample. In order to determine the sensitivity and specificity of the test, n_1 meat samples have a specified amount of *e. coli* placed in them and n_2 meat samples are treated so that *e. coli* is absolutely not present in the meat samples. The screening test is applied to the $n = n_1 + n_2$ meat samples resulting in the following

<i>e.coli</i> Status	Test Result		Total
	Test +	Test -	
Present	n_{11}	n_{12}	n_1
Absent	n_{21}	n_{22}	n_2

Sensitivity of the test is estimated as

$$\text{Sensitivity} = P[\text{Test} = + | \text{e.coliPresent}] \approx \frac{n_{11}}{n_1}$$

Specificity of the test is estimated as

$$\text{Specificity} = P[\text{Test} = - | \text{e.coliAbsent}] \approx \frac{n_{22}}{n_2}$$

The sensitivity and specificity are important parameters for any screening test but a more important quantity is the probability that the Event is present given that the test result is positive. That is, what is the chance *e.coli* is in fact present in the meat sample when the screening test yields a positive test result. This probability can be computed using Bayes Theorem once the prevalence of the Event is known. Using Bayes theorem we have

$$\begin{aligned}
 P[\text{Present} | \text{Test}+] &= \frac{P[\text{Test}+ | \text{Present}]P[\text{Present}]}{P[\text{Test}+ | \text{Present}]P[\text{Present}] + P[\text{Test}+ | \text{Absent}]P[\text{Absent}]} \\
 &= \frac{(\text{Sensitivity})(\text{Prevalence})}{(\text{Sensitivity})(\text{Prevalence}) + (1-\text{Specificity})(1-\text{Prevalence})}
 \end{aligned}$$

When the prevalence is small and the disease is present, the probability of a positive test result is often **very small**. That is, a large proportion of positive test results are false positives. How can we improve the performance of the test? Should we concentrate on increasing the sensitivity or specificity of the test? The following table displays the performance of nine tests for detecting the presence of a pathogen which has a prevalence of .001.

Test	Sensitivity	Specificity	$Pr(\text{Pathogen} \text{Positive})$
1	.90	.90	.0089
2	.90	.95	.0177
3	.90	.999	.4739
4	.95	.90	.0094
5	.95	.95	.0187
6	.95	.999	.4874
7	.99	.90	.0098
8	.99	.95	.0194
9	.99	.999	.4977

When prevalence
is low, best to
~~Increase specificity~~

Based on the values given in the above table, which factor, sensitivity or specificity, is more crucial in obtaining an improved test?

Also, how can tests which seem to have relatively high values for Sensitivity and Specificity yield higher values for $Pr(\text{Pathogen}|\text{Positive})$?

One way is to run the tests multiple times on independent samples of meat from the same batch.

Declare the pathogen is present if k independent tests of the product ALL yield a positive result.

$$P[\text{pathogen present}|k \text{ independent tests are all +}] =$$

$$\frac{(\text{sensitivity})^k (\text{prevalence})}{(\text{sensitivity})^k (\text{prevalence}) + (1 - \text{specificity})^k (1 - \text{prevalence})}$$

For example, in the above table with Prevalence=.001, we have if Sensitivity = .99 and Specificity=.95, then

$$P[\text{pathogen present}|\text{test result is +}] = .0194$$

If we run k tests, how will this probability change:

Number of Tests(k)	Sensitivity	Specificity	$Pr(\text{Pathogen} k\text{Positives})$
1	.99	.95	.0194
2	.99	.95	.2818
3	.99	.95	.8860
4	.99	.95	.9935
5	.99	.95	.9997



McNemar's Test for Matched Pairs

In some situations the information in a 2×2 contingency table is collected from experimental units for which two related responses are obtained. There is no longer n independent responses categorized into the four cells but rather a pair of responses from related units. For example, responses from the same individual at two different times (before and after an intervention) or from two individuals who are physically related (husband-wife or twins) or from body parts of the same experimental unit (right hand-left hand or right eye-left eye).

The data from a study involving matched pairs has the same form as the 2×2 tables we discussed previously except now the response is recorded in such a manner that the pairing is identified. Consider the following table:

		Response 1		
Response 2		Yes	No	Total
Yes	Yes	n_{11}	n_{12}	$n_{1\cdot}$
	No	n_{21}	n_{22}	$n_{2\cdot}$
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	n

The interpretation of the data in the table is as follows: n_{11} is the number of pairs with Yes for both Responses, n_{21} is the number of pairs with Yes for Response 1 but No for Response 2, n_{12} is the number of pairs with No for Response 1 but Yes for Response 2, n_{22} is the number of pairs with No for both Responses.

Suppose the population of responses for all such pairs has proportions given in the following table:

Population Proportions

		Response 1		
Response 2		Yes	No	Total
Yes	Yes	p_{11}	p_{12}	$p_{1\cdot}$
	No	p_{21}	p_{22}	$p_{2\cdot}$
Total		$p_{\cdot 1}$	$p_{\cdot 2}$	1

The research question in this situation is whether the proportion of pairs responding Yes for Response 1 is the same as the proportion of pairs responding Yes for Response 2. The Pearson Chi-square test is not a valid test statistic because the cell counts may be correlated due to pairing of the two responses. We want to test the hypotheses:

$$H_0 : p_{1\cdot} = p_{\cdot 1} \text{ versus } H_1 : p_{1\cdot} \neq p_{\cdot 1},$$

or a corresponding 1-sided hypotheses: $H_0 : p_{1\cdot} \geq p_{\cdot 1}$ versus $H_1 : p_{1\cdot} < p_{\cdot 1}$.

First note that

$$p_{1\cdot} - p_{\cdot 1} = (p_{11} + p_{12}) - (p_{11} + p_{21}) = p_{12} - p_{21}$$

Therefore, a test of the marginal homogeneity for the matched pairs $H_0 : p_{1\cdot} = p_{\cdot 1}$ is equivalent to a test of $H_0 : p_{12} = p_{21}$.

That is, are the proportions of switches from Yes to No and from No to Yes equal.

When H_0 is true, the expected values for the counts n_{12} and n_{21} should be equal. Let $m = n_{12} + n_{21}$ be the total count in the off-diagonal cells in the 2×2 table. Under H_0 , the allocation of the m observations to the (1,2) and (2,1) cells is a binomial experiment with $\frac{1}{2}$ chance for each of the two cells. That is, under H_0 , n_{12} is Binomial(m , .5)

Let B have a $\text{Bin}(m,.5)$ distribution then the p-value for the test is given as follows:

1. For testing $H_1 : p_{1.} > p_{.1}$, p-value = $P[B \geq n_{12}]$, where B has a $\text{Bin}(m,.5)$ distribution.
2. For testing $H_1 : p_{1.} < p_{.1}$, p-value = $P[B \leq n_{12}]$, where B has a $\text{Bin}(m,.5)$ distribution.
3. For testing $H_1 : p_{1.} \neq p_{.1}$, p-value = $2 \cdot \min(P[B \geq n_{12}], P[B \leq n_{12}])$, where B has a $\text{Bin}(m,.5)$ distribution.

 In all three sets of hypotheses, we reject H_o when p-value $\leq \alpha$. Also, note the hypotheses are equivalent to a test comparing p_{12} to $\frac{1}{2}$.

Alternative Approach

Under H_o , n_{12} is Binomial(m, .5), thus for $m > 10$, we could use the normal approximation to the binomial distribution, and use the test statistic:

$$Z = \frac{\frac{n_{12}}{m} - .5}{\sqrt{\frac{(.5)(1-.5)}{m}}} = \frac{n_{12} - .5m}{\sqrt{(.5)(1-.5)m}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

McNemar in a 1947 article proposed

$$Q_{MN} = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \quad \text{under } H_o \quad Q_{MN} \text{ is approximately chi-square with df=1}$$

as the test statistic to evaluate these hypotheses. Note, that

$$Q_{NM} = Z^2 \quad \text{and } Z^2 \text{ is approximately distributed chi-square with df=1}$$

Thus McNemar's test is a large sample version of the exact binomial test but McNemar's test is applicable only to 2-sided alternatives.

 An approximate $100(1 - \alpha)\%$ confidence interval on $p_{1.} - p_{.1}$ is

$$(\hat{p}_{1.} - \hat{p}_{.1}) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_{1.}(1 - \hat{p}_{1.}) + \hat{p}_{.1}(1 - \hat{p}_{.1}) - 2(\hat{p}_{11}\hat{p}_{22} + \hat{p}_{12}\hat{p}_{21})}{n}}$$

which simplifies to

$$(\hat{\pi}_{1.} - \hat{\pi}_{.1}) \pm Z_{\alpha/2} \frac{1}{n} \sqrt{(n_{12} + n_{21}) + \frac{1}{n}(n_{12} - n_{21})^2}$$

EXAMPLE A case-control study was conducted in which the researchers were interested in determining if there was a relationship between diabetes and chronic circulatory problems. The 180 patients having chronic circulatory problems were matched by age, gender, occupation, and ethnicity with 180 patients without chronic circulatory problems. The 180 pairs of subjects were then asked whether they had been diagnosed as having diabetes. The data is given in the following table:

		With Circulatory Problems		
Without Circulatory Problems		Diabetes	No Diabetes	Total
	Diabetes	79	21	100
	No Diabetes	39	41	80
Total		118	62	180

The 180 pairs of subjects in the study consist of four groups:

Group	With Circulatory Problems	Without Circulatory Problems	Count
1	Diabetes Y	Diabetes Y	79
2	Diabetes Y	Diabetes N	39
3	Diabetes N	Diabetes Y	21
4	Diabetes N	Diabetes N	41

We want to test the research hypothesis that the proportion of Without Circulatory problems patients having diabetes is less than the proportion of With Circulatory problems patients having diabetes. That is, test $H_1 : p_{1.} < p_{.1}$ or equivalently, test $H_1 : p_{12} < p_{21}$

From the data we have $\hat{p}_{1.} = 100/180 = .556$ and $\hat{p}_{.1} = 118/180 = .656$

Therefore, $\hat{p}_{1.} - \hat{p}_{.1} = .556 - .656 = -.1$. The proportion of diabetic patients Without Circulatory problems is 10% less than the proportion of diabetic patients With Circulatory problems.

The p-value = $P[B \leq 21]$ where $m = 21 + 39 = 60$ and B is $\text{Bin}(60,.5)$.

Thus, p-value = $P[B \leq 21] = 0.01387 = \text{pbnom}(21,60,.5)$.

Hence, conclude there is substantial evidence that the proportion of Without Circulatory problems patients having diabetes is less than the proportion of With Circulatory problems patients having diabetes.

Using McNemar's Test, we have $Q_{NM} = \frac{(21-39)^2}{(21+39)} = 5.4$ with p-value = $P[\chi_1^2 \geq 5.4] = 0.020$. However, this p-value is for a two-sided alternative hypothesis. From the following SAS program, we obtain both the Asymptotic and Exact p-values. Once again, these values are computed for a two-sided alternative. The Exact p-value = $.0273 \approx 2(.01367)$, our 1-sided p-value.

The following is an approximate 95% C.I. on $p_{1.} - p_{.1}$

$$\left(\frac{100}{180} - \frac{118}{180} \right) \pm 1.96 \frac{1}{180} \sqrt{(21+39) + \frac{1}{180}(21-39)^2}$$

That is, $-.10 \pm .0856 = (-.186, -.014)$

The following SAS program will analyze the data for the above example:

```
*The following program will calculate the ;
*McNemar Test;
*for Example in Handout 13;

OPTIONS PS=55 LS=75 NOCENTER NODATE;
TITLE 'Circulatory Problem - McNemar Test';
DATA CIRC;
INPUT DIABWITHOUTCIR $ DIABWITHCIR $ CNT @@;
CARDS;
YES YES 79
YES NO 21
NO YES 39
NO NO 41
RUN;
PROC FREQ ORDER=DATA;
TABLES DIABWITHOUTCIR*DIABWITHCIR;
WEIGHT CNT;
EXACT MCNEM;
RUN;
```

From SAS, we obtain the following output. Note that the p-value from the EXACT option is for a two-sided alternative and is twice the value we obtained.

Circulatory Problem - McNemar Test

The FREQ Procedure

Table of DIABWITHOUTCIR by DIABWITHCIR

DIABWITHOUTCIR DIABWITHCIR

Frequency	YES	NO	Total
Percent			
Row Pct			
YES	79	21	100
	43.89	11.67	55.56
	79.00	21.00	
	66.95	33.87	
NO	39	41	80
	21.67	22.78	44.44
	48.75	51.25	
	33.05	66.13	
Total	118	62	180
	65.56	34.44	100.00

Statistics for Table of DIABWITHOUTCIR by DIABWITHCIR

McNemar's Test

Statistic (S) 5.4000
DF 1
Asymptotic Pr > S 0.0201
Exact Pr >= S 0.0273

Simple Kappa Coefficient

Kappa 0.3095
ASE 0.0704
95% Lower Conf Limit 0.1715
95% Upper Conf Limit 0.4474

Sample Size = 180

Design Study Without Pairing

Suppose there exists a pool of 360 subjects which are very similar with respect to age, occupation, gender, ethnicity, etc. How could we design the study to obtain a more effective evaluation of the association between the two diseases? We would not need to pair the subjects so the 360 subjects would be examined and the following table would be obtained:

	Diabetes		
Circulatory Problems	Yes	No	Total
Yes	n_{11}	n_{12}	$n_{1\cdot}$
No	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	360

We would then use the Pearson chi-square test to evaluate the degree association between subjects having circulatory problems and diabetes. This test would be a more sensitive test than the McNemar's test in determining associations provided the subjects are relatively homogeneous with respect to all characteristics other than the two characteristics for which association is being examined.

Pairing is used when the n subjects available for the study have major differences with respect to those characteristics which may impact the measured responses central to the researcher's goals. For example, age, gender, occupation, ethnicity, health condition.

SPRING Friday

12 13 [2]

Cochran-Mantel-Haenszel Test for Sets of Contingency Tables

The following example will illustrate the problems that may arise when we have sets of contingency tables which are often combined into a single table.

The results from a large survey are described in *American Statistician*, 50, pp. 340-341 (1996). The 1314 women in the survey were asked whether they smoked. Twenty years later, the women were again contacted and determined if they were alive or dead. The data is summarized as follows:

Smoking Status and Survival Status

Smoking Status	Survival Status		Total	Odds of Death
	Dead	Alive		
Yes	139	443	582	.314
No	230	502	732	.458
Total	369	945	1314	OR=.658

From the above table, $139/582=24\%$ of the smokers had died during the twenty years but even a larger percentage of the nonsmokers $230/732=31\%$ had died. Thus, smokers have a higher survival rate. The odds ratio of smokers versus nonsmokers for death is $OR = .658 < 1$, odds of death are considerably less for smokers than nonsmokers. Furthermore, the Pearson chi-square test comparing the survival rates for smokers vs nonsmokers has a p-value of 0.0025. Thus, there is substantial evidence that the survival rate for smokers is higher than for nonsmokers.

What could possibly explain this relationship? Another crucial variable is the age of the participant. It may be that older women are less likely to be smokers and this may help explain the differential in survival rate. The following table includes the explanatory variable, age of participant, when survey was initially conducted in 1975.

(want to adjust for age (see HW 8 # 9c))

Smoking Status and Survival Status, Separated by Women's Age

Smoking Status	Age Group							
	18-34		35-54		55-65		65+	
Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive	
Yes	5	174	41	198	51	64	42	7
No	6	213	19	180	40	81	165	28
OR	1.02		1.96		1.61		1.02	

Note that OR is the odds ratio of Smokers to NonSmokers for death and has a value greater than 1 for all four groups. That is, when considered separately within each of four age groups, the smokers are more likely to die than the non-smokers. Why is there a contradiction between the two tables?

The following table has the conditional probabilities of Death for Smokers and Nonsmokers, conditioning on Age:

Percentage of Deaths Conditioning on Women's Age

Smoking Status	Age Group				Overall
	18-34	35-54	55-65	65+	
Yes	2.8%	17.2%	44.3%	85.7%	23.9%
No	2.7%	9.5%	33.1%	85.5%	31.4%
OR	1.02	1.96	1.61	1.02	

From the above table, we can observe that

1. For the youngest and oldest age groups, the proportion of respondents who were dead twenty years later for Smokers is nearly equal to the proportion of deaths for nonsmokers.
2. For the middle two groups, the proportion of deaths is much higher for smokers.
3. For each of the four age groups, women who smoked had a higher death rate than the death rate for nonsmokers.
4. There appears to be a contradiction between the death rates in the overall table and the death rates observed conditional on the age of the respondent. This contradiction is often referred to as the **Simpson Paradox**.

An explanation of this apparent contradiction is as follows:

- The proportion of women who smoked tended to be higher for younger women:
45% of 18-34 smoked whereas, 20% of 65+ smoked.
- However, the proportion of 18-34 who were died 20 years later was 3% but
the proportion of 65+ women who were dead 20 years later was 86%.
- Thus, the difference in survival rates between smokers and nonsmokers in the overall table is most likely due to the fact that younger women are more likely to be smokers but they are also less likely to die twenty years after the survey was taken.
- The opposite is true for the oldest age group.
- Also, note that the four age groups are not equally represented in the overall table:

The percentage of respondents in the four age groups was 30.3%, 33.3%, 51.3%, and 18.4%.

This example illustrates how ignoring potential explanatory variables can often lead to very misleading conclusions in studies. The variable Age in this example in some sense explains why the conclusion that smokers have a lower death rate than nonsmokers is not true when other factors are taken into account. We will consider in STAT 642 a similar type of situation in trying to reach conclusions concerning design factors when there exists an interaction effect between two factors.

Conditional and Marginal Analysis

The analysis using the table pooled over Age is called a marginal analysis and the analysis taking into account Age is called a conditional analysis conditioning on the variable Age.

The following test referred to as the Cochran-Mantel-Haenszel (CMH) Test is a test of conditional independence given the values of a third variable.

In our example, CMH tests the conditional independence of the variables Smoking and Survival given the values of the variable Age.

In general, suppose we have K levels of a variable Z and at each level of Z we have two variables X and Y each with two levels, that is, we have K 2×2 tables.

The null hypothesis is that X and Y are independent given Z :

$$p_{ijk} = p_{i..}p_{jk} \quad \text{for all } k = 1, \dots, K$$

This hypothesis can be interpreted as the conditional odds ratio $\theta_{XY(k)}$ between X and Y equals 1 in each of the K 2×2 tables.

Given the row totals, $(n_{1..k}, n_{2..k})$, and column totals, $(n_{.1k}, n_{.2k})$ in the k th table, the count n_{11k} determines all other counts in this table.

Under the null hypothesis, the mean and variance of n_{11k} are given by

$$\mu_{11k} = E(n_{11k}) = \frac{n_{1..k}n_{.1k}}{n_{..k}} \quad \text{Var}(n_{11k}) = \frac{n_{1..k}n_{2..k}n_{.1k}n_{.2k}}{n_{..k}^2(n_{..k} - 1)}$$

When the true odds ratio θ_{XYk} exceeds 1.0 in the k th table, the expectation is that $(n_{11k} - \mu_{11k}) > 0$.

When the true odds ratio θ_{XYk} is less than 1.0 in the k th table, the expectation is that $(n_{11k} - \mu_{11k}) < 0$.

The CMH test statistic sums these differences over the K tables:

$$CMH = \frac{\left[\sum_{k=1}^K (n_{11k} - \mu_{11k}) \right]^2}{\sum_{k=1}^K \text{Var}(n_{11k})}$$

 Under the null hypothesis, CMH asymptotically has a chi-squared distribution with $df = 1$.

The CMH statistic combines the information about the association between X and Y across the K tables. The CMH will be *large* when $(n_{11k} - \mu_{11k})$ is consistently positive or consistently negative for all K tables, rather than varying from positive in some tables and then negative for the remaining tables.

When an association exists between the variables X and Y and this association is relatively homogeneous across the K tables, it is informative to estimate this common value of the K true odds ratios.

Suppose $\theta_{XY(1)} = \theta_{XY(2)} = \dots = \theta_{XY(K)}$, the CMH estimator of the common value is given by a pooled estimate of the common odds ration:

$$\hat{\theta}_{CMH} = \frac{\sum_{k=1}^K (n_{11k}n_{22k}/n_{..k})}{\sum_{k=1}^K (n_{12k}n_{21k}/n_{..k})}$$

The standard error of the estimator is an extremely complex formula but can be obtained from SAS.

only use CMH test when no reason to believe we have a common odds ratio.

The CMH test is a powerful test for detecting association against the null hypothesis of conditional independence in the K - 2x2 tables. However, the CMH test loses its power to detect overall association between the variables X and Y when the associations vary from a positive association to a negative association across the K - 2x2 tables. Thus, the CMH test should be applied to situations in which $OR_1 = OR_2 = \dots = OR_k = OR_o$, where OR_o is the common odds ratio. Use Breslow-Day to evaluate this condition. The CMH procedure is then testing $H_o : OR_o = 1$ versus $H_o : OR_o \neq 1$.

Breslow-Day Test

A test of common odds ratio between X and Y for the K 2×2 tables:

$$H_o : \theta_{XY(1)} = \theta_{XY(2)} = \dots = \theta_{XY(K)} \quad \text{versus} \quad H_1 : \text{not all } \theta_{XY(k)} \text{ s are equal}$$

The Breslow-Day Test evaluates the above hypotheses. The computation of this statistic is relatively complex but can be obtained using SAS. When the Breslow-Day test fails to reject H_o , there is justification in summarizing the conditional association between X and Y by a single odds ratio, the CMH odds ratio, for all K tables. If the Breslow-Day test rejects H_o , then separate tests should be run for each of the K tables including a Bonferroni adjustment to the value of α to control for multiple testing's inflated Type I error rate.

- 1. First test $H_o : \theta_{XY(1)} = \theta_{XY(2)} = \dots = \theta_{XY(k)}$ using the Breslow-Day test.
- 2. If H_o is not rejected then test the Conditional Independence of (X, Y) using the Cochran-Mantel-Haenszel test.
- 3. If H_o is rejected then test the K 2x2 tables separately for each of the K levels of the confounding variable using $\alpha_{pc} = \frac{\alpha}{K}$, where α is the chance of a Type I error across all K tests.

Note that the Breslow-Day procedure tests the equality of the K odds ratios:

$$H_o : OR_1 = OR_2 = \dots = OR_K, \text{ whereas}$$

the C-M-H procedure tests

$$H_o : p_{ijk} = p_{i,k}p_{.jk} \text{ for } k = 1, 2, \dots, K$$

The following SAS code and output completes our analysis of the survival data of women smokers versus nonsmokers.

*The following program will calculate the ;
 *General Cochran-Mantel-Haenszel Statistic for the situation;
 *where there is a third variable over which the original;
 *counts were aggregated.;
 *Age refers to the age of the respondent in 1975;
 *Smoker refers to whether or not respondent was a smoker when survived;
 *Surv refers to the Survival Status of the respondent 20 years later in 1995;

```
OPTIONS PS=55 LS=75 NOCENTER NODATE;
```

```
DATA ByAge;
```

```
INPUT Age $ Smoker $ Surv $ CNT @@;
```

```
CARDS;
```

18-34 Y D 5	18-34 Y A 174
18-34 N D 6	18-34 N A 213
35-54 Y D 41	35-54 Y A 198
35-54 N D 19	35-54 N A 180
55-64 Y D 51	55-64 Y A 64
55-64 N D 40	55-64 N A 81
65+ Y D 42	65+ Y A 7
65+ N D 165	65+ N A 28

```
RUN;
```

```
PROC FREQ ORDER=DATA;
```

```
TABLES SMOKER*SURV/EXPECTED CELLCHI2 CHISQ;
```

```
WEIGHT CNT;
```

```
EXACT CHISQ;
```

```
RUN;
```

```
PROC FREQ ORDER=DATA;
```

```
TABLES AGE*SMOKER*SURV/CMH;
```

```
WEIGHT CNT;
```

```
RUN;
```

The FREQ Procedure

Table of Smoker by Surv

Smoker	Surv		
	Frequency	Expected	Cell Chi-Square
	Percent	Row Pct	Col Pct
	D	A	Total
Y	139	443	582
	163.44	418.56	
	3.6542	1.4269	
	10.58	33.71	44.29
	23.88	76.12	
	37.67	46.88	
N	230	502	732
	205.56	526.44	
	2.9054	1.1345	
	17.50	38.20	55.71
	31.42	68.58	
	62.33	53.12	
Total	369	945	1314
	28.08	71.92	100.00

Statistics for Table of Smoker by Surv

Statistic	DF	Value	Prob
Chi-Square	1	9.1209	0.0025
Likelihood Ratio Chi-Square	1	9.2003	0.0024
Continuity Adj. Chi-Square	1	8.7515	0.0031
Mantel-Haenszel Chi-Square	1	9.1140	0.0025
Phi Coefficient		-0.0833	
Contingency Coefficient		0.0830	
Cramer's V		-0.0833	

Pearson Chi-Square Test

Chi-Square	9.1209
DF	1
Asymptotic Pr > ChiSq	0.0025
Exact Pr >= ChiSq	0.0030

Statistics for Table of Smoker by Surv

Likelihood Ratio Chi-Square Test

Chi-Square	9.2003
DF	1
Asymptotic Pr > ChiSq	0.0024
Exact Pr >= ChiSq	0.0025

Mantel-Haenszel Chi-Square Test

Chi-Square	9.1140
DF	1
Asymptotic Pr > ChiSq	0.0025
Exact Pr >= ChiSq	0.0030

Fisher's Exact Test

Cell (1,1) Frequency (F)	139
Left-sided Pr <= F	0.0015
Right-sided Pr >= F	0.9990

Table Probability (P)	5.052E-04
Two-sided Pr <= P	0.0030

Sample Size = 1314

Table 1 of Smoker by Surv
Controlling for Age=18-34

Smoker	Surv		
Frequency			
Percent			
Row Pct			
Col Pct	D	A	Total
Y	5	174	179
	1.26	43.72	44.97
	2.79	97.21	
	45.45	44.96	
N	6	213	219
	1.51	53.52	55.03
	2.74	97.26	
	54.55	55.04	
Total	11	387	398
	2.76	97.24	100.00

Table 2 of Smoker by Surv
Controlling for Age=35-54

Smoker	Surv		
Frequency			
Percent			
Row Pct			
Col Pct	D	A	Total
Y	41	198	239
	9.36	45.21	54.57
	17.15	82.85	
	68.33	52.38	
N	19	180	199
	4.34	41.10	45.43
	9.55	90.45	
	31.67	47.62	
Total	60	378	438
	13.70	86.30	100.00

Table 3 of Smoker by Surv
Controlling for Age=55-64

Smoker	Surv		
	Frequency	Percent	
	Row Pct	Col Pct	Total
D	A		
Y	51	64	115
	21.61	27.12	48.73
	44.35	55.65	
	56.04	44.14	
N	40	81	121
	16.95	34.32	51.27
	33.06	66.94	
	43.96	55.86	
Total	91	145	236
	38.56	61.44	100.00

Table 4 of Smoker by Surv
Controlling for Age=65+

Smoker	Surv		
	Frequency	Percent	
	Row Pct	Col Pct	Total
D	A		
Y	42	7	49
	17.36	2.89	20.25
	85.71	14.29	
	20.29	20.00	
N	165	28	193
	68.18	11.57	79.75
	85.49	14.51	
	79.71	80.00	
Total	207	35	242
	85.54	14.46	100.00

Summary Statistics for Smoker by Surv
Controlling for Age

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.5896	0.0103
2	Row Mean Scores Differ	1	6.5896	0.0103
3	General Association	1	6.5896	0.0103

small p-value
 ⇒ Smoking and
 Survival are
 associated
 w/ each other

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	1.5611	1.1081	2.1994
	Logit	1.5556	1.1038	2.1922
Cohort (Col1 Risk)	Mantel-Haenszel	1.2282	1.0490	1.4381
	Logit	1.0718	0.9548	1.2032
Cohort (Col2 Risk)	Mantel-Haenszel	0.9370	0.8918	0.9845
	Logit	0.9811	0.9521	1.0109

95% CI of
 common odds
 ratio doesn't
 include 1

Breslow-Day Test for
Homogeneity of the Odds Ratios

Chi-Square	1.9862
DF	3
Pr > ChiSq	0.5753

Total Sample Size = 1314

fail to reject H₀
 all the odds ratios are
 the same ⇒ Do CMH test.

Because the Breslow-Day test fails to detect a difference in the four Age-Groups' odds-ratios, we can now estimate the common odds-ratio amongst the k groups:

$$\widehat{OR} = \frac{\sum_k n_{11k}n_{22k}/n_{..k}}{\sum_k n_{12k}n_{21k}/n_{..k}}$$

For the four Age-Groups in the Smoking example, we would obtain the following estimate of the common OR:

$$\widehat{OR} = \frac{(5)(213)/398 + (41)(180)/438 + (51)(81)/236 + (42)(28)/242}{(6)(174)/398 + (19)(198)/438 + (40)(64)/236 + (165)(7)/242} = 1.56 > 1$$

We would thus conclude that the odds of death during the 20 years is 56% higher for Smokers than for Non-Smokers.

~~W~~

The following R code will conduct both the Breslow-Day test and the CMH test for the Smoking Example

```
install.packages("lawstat")
install.packages("DescTools")

library(DescTools)
library(lawstat)

y = c(5, 6, 174, 213, 41, 19, 198, 180, 51, 40, 64, 81, 42, 165, 7, 28)

dim(y) = c(2,2,4)

BreslowDayTest(y, OR=NA, correct=FALSE)

cmh.test(y)
```

Output from R:

```
Breslow-Day test on Homogeneity of Odds Ratios

data: y
X-squared = 1.9862, df = 3, p-value = 0.5753

>
> cmh.test(y)

Cochran-Mantel-Haenszel Chi-square Test

data: y
CMH statistic = 6.589600, df = 1.000000, p-value = 0.010258, MH
Estimate = 1.561100, Pooled Odd Ratio = 0.684840, Odd Ratio of level 1
= 1.020100, Odd Ratio of level 2 = 1.961700, Odd Ratio of level 3 =
1.613700, Odd Ratio of level 4 = 1.018200
```

TESTS FOR COMPARING SEVERAL POPULATION VARIANCES

We have discussed a method for comparing variances from two normally distributed populations based on taking independent random samples from the populations. In many situations, we will need to compare more than two populations. For example, we may want to compare the variability in the level of nutrients of five different suppliers of a feed supplement or the variability in scores of the students using SAT preparatory materials from the three major publishers of those materials. Thus, we need to develop a statistical test which will allow us to compare $t > 2$ population variances. We will consider two procedures.

The first procedure, Hartley's test is very simple to apply but has the restriction that the population distributions must be normally distributed and the sample sizes equal.

The second procedure, Brown-Forsythe-Levene (BFL) test, is more complex in its computations but does not restrict the population distributions or the sample sizes. The BFL test can be obtained from many of the statistical software packages. For example, *SAS* and *Minitab* both use BFL test for comparing population variances. This test is available in the R package "lawstat" as the function `levene.test`.

Hartley F_{max} :

H. O. Hartley (1950) developed a test for the evaluating the hypotheses

$$H_o : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 \text{ vs } H_a : \sigma_i^2 \text{'s not all equal.}$$

*W.F.
Brown
Forsythe
Levene
Hartley*

The Hartley F_{max} requires that we have independent random samples of the same size n from t normally distributed populations. With the exception that we require $n_1 = n_2 = \cdots = n_t = n$, the Hartley test is a logical extension of the F test from the previous section for testing $t = 2$ variances. With s_i^2 denoting the sample variance computed from the i th sample: $s_1^2, s_2^2, \dots, s_t^2$

Let $s_{min}^2 = \min(s_1^2, s_2^2, \dots, s_t^2)$ and $s_{max}^2 = \max(s_1^2, s_2^2, \dots, s_t^2)$.

The Hartley F_{max} test statistic is

$$F_{max} = \frac{s_{max}^2}{s_{min}^2}.$$

The test procedure is summarized here.

Hartley's F_{max} Test for
Homogeneity of
Population Variances

$$H_o : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 \text{ homogeneity of variances}$$

$$H_a : \text{Population variances are not all equal}$$

$$\text{T.S.: } F_{max} = \frac{s_{max}^2}{s_{min}^2}$$

R.R. Reject H_o if $F_{max} \geq F_{max,\alpha,t,n-1}$

the value of $F_{max,\alpha,t,n-1}$ for specified values of α, t , and $df_2=n-1$, where n is the common sample size for the t random samples is given in the table on the next page. Note that this is not the standard F-tables.

TABLE IV
Percentage points of $F_{\max} = s_{\max}^2/s_{\min}^2$

		Upper 5% Points										
t	df_2	2	3	4	5	6	7	8	9	10	11	12
	2	39.0	87.5	142	202	266	333	403	475	550	626	704
	3	15.4	27.8	39.2	50.7	62.0	72.9	83.5	93.9	104	114	124
	4	9.60	15.5	20.6	25.2	29.5	33.6	37.5	41.1	44.6	48.0	51.4
	5	7.15	10.8	13.7	16.3	18.7	20.8	22.9	24.7	26.5	28.2	29.9
	6	5.82	8.38	10.4	12.1	13.7	15.0	16.3	17.5	18.6	19.7	20.7
	7	4.99	6.94	8.44	9.70	10.8	11.8	12.7	13.5	14.3	15.1	15.8
	8	4.43	6.00	7.18	8.12	9.03	9.78	10.5	11.1	11.7	12.2	12.7
	9	4.03	5.34	6.31	7.11	7.80	8.41	8.95	9.45	9.91	10.3	10.7
	10	3.72	4.85	5.67	6.34	6.92	7.42	7.87	8.28	8.66	9.01	9.34
	12	3.28	4.16	4.79	5.30	5.72	6.09	6.42	6.72	7.00	7.25	7.48
	15	2.86	3.54	4.01	4.37	4.68	4.95	5.19	5.40	5.59	5.77	5.93
	20	2.46	2.95	3.29	3.54	3.76	3.94	4.10	4.24	4.37	4.49	4.59
	30	2.07	2.40	2.61	2.78	2.91	3.02	3.12	3.21	3.29	3.36	3.39
	60	1.67	1.85	1.96	2.04	2.11	2.17	2.22	2.26	2.30	2.33	2.36
	∞	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		Upper 1% Points										
t	df_2	2	3	4	5	6	7	8	9	10	11	12
	2	199	448	729	1036	1362	1705	2063	2432	2813	3204	3605
	3	47.5	85	120	151	184	21(6)	24(9)	28(1)	31(0)	33(7)	36(1)
	4	23.2	37	49	59	69	79	89	97	106	113	120
	5	14.9	22	28	33	38	42	46	50	54	57	60
	6	11.1	15.5	19.1	22	25	27	30	32	34	36	37
	7	8.89	12.1	14.5	16.5	18.4	20	22	23	24	26	27
	8	7.50	9.9	11.7	13.2	14.5	15.8	16.6	17.9	18.9	19.8	21
	9	6.54	8.5	9.9	11.1	12.1	13.1	13.9	14.7	15.3	16.0	16.6
	10	5.85	7.4	8.6	9.6	10.4	11.1	11.8	12.4	12.9	13.4	13.9
	12	4.91	6.1	6.9	7.6	8.2	8.7	9.1	9.5	9.9	10.2	10.6
	15	4.07	4.9	5.5	6.0	6.4	6.7	7.1	7.3	7.5	7.8	8.0
	20	3.32	3.8	4.3	4.6	4.9	5.1	5.3	5.5	5.6	5.8	5.9
	30	2.63	3.0	3.3	3.4	3.6	3.7	3.8	3.9	4.0	4.1	4.2
	60	1.96	2.2	2.3	2.4	2.4	2.5	2.5	2.6	2.6	2.7	2.7
	∞	1.00	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Note: s_{\max}^2 is the largest and s_{\min}^2 the smallest in a set of t independent mean squares, each based on $df_2 = n - 1$ degrees of freedom. Values in the column $t = 2$ and in the rows $df_2 = 2$ and ∞ are exact. Elsewhere, the third digit may be in error by a few units for the 5% points and several units for the 1% points. The third-digit figures in parentheses for $df_2 = 3$ are the most uncertain.

Source: From *Biometrika Tables for Statisticians*, 3rd ed., Vol. 1; edited by E. S. Pearson and H. O. Hartley (New York: Cambridge University Press, 1966), Table, p. 202. Reproduced by permission of the *Biometrika Trustees*.

We will illustrate the application of the Hartley test with the following example.

EXAMPLE

Wludyka and Nelson(1997), *Technometrics*, Vol. 39, pp. 274-285 describe the following experiment. In the manufacture of soft contact lenses, a monomer is injected into a plastic frame, the monomer is subjected to ultraviolet light and heated (the time, temperature, and light intensity are varied), the frame is removed, and the lens is hydrated. It is thought that temperature can be manipulated to target the power (the strength of the lens), so interest is in comparing the variability in power. The data are coded deviations from target power using monomers from three different suppliers. We wish to test $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$.

Deviations from Target Power for Three Suppliers

Supplier	Sample									n	s_i^2
	1	2	3	4	5	6	7	8	9		
1	191.9	189.1	190.9	183.8	185.5	190.9	192.8	188.4	189.0	9	8.69
2	178.2	174.1	170.3	171.6	171.7	174.7	176.0	176.6	172.8	9	6.89
3	218.6	208.4	187.1	199.5	202.0	211.1	197.6	204.4	206.8	9	80.22

Solution Prior to conducting the Hartley test we must check the normality condition. The data are evaluated for normality using the normal probability plots and box plots. All three data sets appear to be from normally distributed populations. Thus, we will apply the Hartley F_{max} test to the data sets. From F-max Percentile Table, with $\alpha=.05$, $t=3$, and $df_2=9-1=8$, we have $F_{max,.05} = 6.00$. Thus, our rejection region will be

R.R.: Reject H_0 if $F_{max} \geq F_{max,.05,3,8} = 6.00$.

$$s_{min}^2 = \min(8.69, 6.89, 80.22) = 6.89 \quad \text{and} \quad s_{max}^2 = \max(8.69, 6.89, 80.22) = 80.22$$

Thus,

$$F_{max} = \frac{s_{max}^2}{s_{min}^2} = \frac{80.22}{6.89} = 11.64 > 6.00$$

Thus, we reject H_0 and conclude that the variances are not all equal.

If the sample sizes are not all equal, we can take $n = n_{max}$, where n_{max} is the largest sample size. The F_{max} no longer has an exact level α . In fact, the test is liberal in the sense that the probability of type I error is slightly more than the nominal value α . Thus, the test is more likely to falsely reject H_0 than the test having all n_i 's equal when sampling from normal populations with the variances all equal.

A simulation study was conducted to investigate the effect on the level of the F test of sampling from heavy tailed and skewed distributions rather than the required normal distribution.

For each pair of sample sizes $(n_1, n_2) = (10,10), (10,20)$ or $(20,20)$, random samples of the specified sizes were selected from one of the five distributions. A test of $H_o : \sigma_1^2 = \sigma_2^2$ vs $H_a : \sigma_1^2 \neq \sigma_2^2$ was conducted using an F test with $\alpha = .05$. This process was repeated 2500 times for each of the five distributions and three sets of sample sizes. The results are given in the following Table.

Proportion of Times $H_o : \sigma_1^2 = \sigma_2^2$ Was Rejected ($\alpha = .05$)

Sample Sizes	Distribution				
	Normal	Uniform	t (df=5)	Gamma (shape=1)	Gamma (shape=.1)
(10,10)	.054	.010	.121	.225	.693
(10,20)	.056	.0068	.140	.236	.671
(20,20)	.050	.0044	.150	.264	.673

The Hartley F_{max} test is quite sensitive to departures from normality. Thus, if the population distributions we are sampling from have a somewhat nonnormal distribution but the variances are equal, the F_{max} will reject H_o and declare the variances to be unequal. The test is detecting the nonnormality of the population distributions not the unequal variances. Thus, when the population distributions are nonnormal, the F_{max} is not recommended as a test of homogeneity of variances. An alternative approach which does not require the populations to have normal distributions is the Brown-Forsythe-Levene (BFL) test. However, the BFL test involves considerably more calculations than the Hartley test. Also, when the populations have a normal distribution, the Hartley test is more powerful than the BFL test. Conover, Johnson, and Johnson (1981), *Technometrics*, Vol. 23, pp. 351-361, conducted a simulation study of a variety of tests of homogeneity of variance including the Hartley and BFL test. They demonstrated the inflated α levels of the Hartley test when the populations have highly skewed distributions and recommended the BFL test as one of several alternative procedures.

Note: If there are only two populations being compared using samples of size n and m , then the Hartley F-max test has exactly an F-distribution with $df = n - 1, m - 1$.

STOP Friday

Hartley test were powerful only if pop variances are normal

Brown-Forsythe-Levene (BFL) Test for Homogeneity of Population Variances

When one or more of the population distributions is not normally distributed, a much more powerful procedure for evaluating the differences in the population variances is the Brown-Forsythe-Levene (BFL) test for homogeneity of variance.

The BFL test involves replacing the j th observation from sample i , y_{ij} , with the random variable $z_{ij} = |y_{ij} - \tilde{y}_i|$, where \tilde{y}_i is the sample median of the i th sample. We then compute the BFL test statistic on the z_{ij} 's.

$$H_o : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 \text{ homogeneity of variances}$$

$$H_a : \text{Population variances are not all equal}$$

$$\text{T.S.: } L = \frac{\sum_{i=1}^t n_i (\bar{z}_i - \bar{z}_{..})^2 / (t-1)}{\sum_{i=1}^t \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2 / (N-t)}$$

R.R. For a specified value of α , reject H_o if $L \geq F_{\alpha, df_1, df_2}$, where $df_1 = t-1$, $df_2 = N-t$, $N = \sum_{i=1}^t n_i$, and F_{α, df_1, df_2} is the upper α percentile from the F-distribution.

The following modification is suggested in the book *Theory of Rank Tests* by Hajek and Sidak:

When the sample sizes within the levels of the groups are odd, at least one value of $z_{ij} = |y_{ij} - \tilde{y}_i|$ will always be zero. This artificially dampens the variance estimate for that group. Thus, Hajek-Sidak recommend replacing the zero with the minimum non-zero value if there is only one z_{ij} which is zero. If more than one z_{ij} within a given group is zero, the zeros are kept.

We will illustrate the computations for the BFL test in the following example. The packages *R*, under "lawstat", *SAS* and *Minitab* have functions for conducting the BFL test.

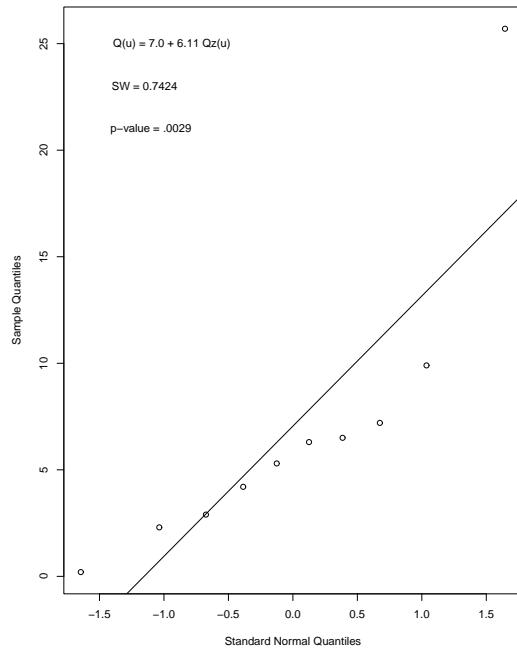
Percent Increase in MPG from Cars Driven Using Three Additives

Three different additives which are marketed for increasing the miles per gallon (mpg) for automobiles were evaluated by a consumer testing agency. Past studies have shown that an average increase of 8% in mpg for economy automobiles after using the product for 250 miles. The testing agency wants to evaluate the variability in the increase in mileage over a variety of brands of cars within the economy class. They randomly selected 30 economy cars of similar age, number of miles on their odometer, and overall condition of the power train to be used in the study. They then randomly assigned 10 cars to each additive. The percentage increase in mpg obtained by each car was recorded for a 250 miles test drive. The testing agency wanted to evaluate whether there was a difference between the three additives with respect to their variability in the increase in mpg. The data is give here along with the intermediate calculations needed to compute the BFL test statistic.

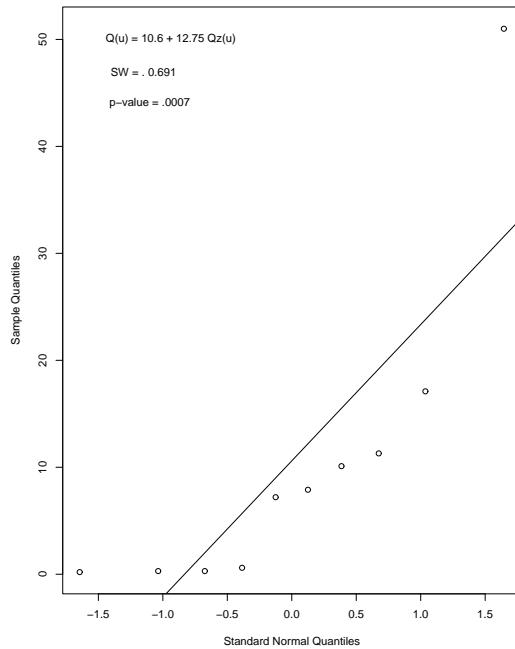
Additive	Percent Increase in MPG for Car i										n	s_i^2
	1	2	3	4	5	6	7	8	9	10		
1	4.2	2.9	0.2	25.7	6.3	7.2	2.3	9.9	5.3	6.5	10	50.525
2	0.2	11.3	0.3	17.1	51.0	10.1	0.3	0.6	7.9	7.2	10	234.927
3	7.2	6.4	9.9	3.5	10.6	10.8	10.6	8.4	6.0	11.9	10	7.220

Using the three normal reference distribution plots on the next page along with the values from the Shapiro-Wilks test, we can observe that the samples from Additive 1 and Additive 2 do not appear to be samples from normally distributed populations.

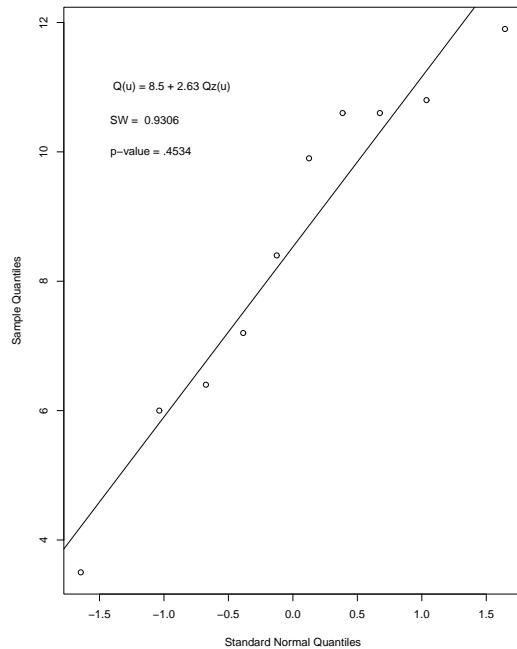
Normal Ref Plot for Additive 1



Normal Ref Plot for Additive 2



Normal Ref Plot for Additive 3



Therefore, we should not use the Hartley's F_{max} test for evaluating differences in the variances in this example.

The information in Table 5 on the next page will assist us in calculating the value of the BFL test statistic.

The medians of the percent increase in mileage, y_{ij} 's, for the three additives are 5.80, 7.55, and 9.15. We then calculate the absolute deviations of the data values about their respective medians, namely, $z_{1j} = |y_{1j} - 5.80|$, $z_{2j} = |y_{2j} - 7.55|$, and $z_{3j} = |y_{3j} - 9.15|$ for $j = 1, \dots, 10$.

These values are given in column 3 of Table 5.

Next, we calculate the three means of these values, $\bar{z}_{1.} = 4.07$, $\bar{z}_{2.} = 8.88$, and $\bar{z}_{3.} = 2.23$.

Next, we calculate the squared deviations of the z_{ij} 's about their respective means, namely, $(z_{ij} - \bar{z}_{i.})^2$, that is,

$$(z_{1j} - 4.07)^2, (z_{2j} - 8.88)^2, \text{ and } (z_{3j} - 2.23)^2.$$

These values are contained in column 6 of Table 5.

Then we calculate the squared deviations of the z_{ij} 's about the overall mean,

$$\bar{z}_{..} = 5.06, \text{ that is, } (z_{ij} - \bar{z}_{..})^2 = (z_{ij} - 5.06)^2.$$

The last column in Table 5 contains these values.

The final step is to sum columns 6 and 7 in Table 5 yielding

$$T_1 = \sum_{i=1}^3 \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i.})^2 = 1742.6 \quad \text{and} \quad T_2 = \sum_{i=1}^3 \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{..})^2 = 1978.4$$

The value of BFL test statistics, in an alternative form, is given by

$$L = \frac{(T_2 - T_1)/(t-1)}{T_1/(N-t)} = \frac{(1978.4 - 1742.6)/(3-1)}{1742.6/(30-3)} = 1.827$$

The rejection region for BFL test is reject H_o if $L \geq F_{\alpha, t-1, N-t} = F_{.05, 3-1, 30-3} = 3.35$.

Since $L = 1.827$, we fail to reject $H_o : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ and conclude that there is insufficient evidence of a difference in the population variances of the percent increase in mpg for the three additives.

Note that if we would have used the Hartley F_{max} test, our conclusion would have been changed.

$$s_1^2 = 50.525 \quad s_2^2 = 234.927 \quad s_3^2 = 7.220$$

$$F_{max} = \frac{234.927}{7.220} = 32.54 > 8.5 = F_{max, .01, 3, 9}$$

Using the Hartley Test, we would conclude there is significant evidence that the three population variances are different.

The Hartley test is distorted by the outliers in the data sets for Additive 1 and Additive 2.

Table 5. Steps in Calculating BFL Test

Additive	y_{1j}	\tilde{y}_1	$z_{1j} =$	$y_{1j} - 5.80$	$\bar{z}_{1.}$	$(z_{1j} - 4.07)^2$	$(z_{1j} - 5.06)^2$
1	4.2	5.80		1.60	4.07	6.1009	11.9716
1	2.9			2.90		1.3689	4.6656
1	0.2			5.60		2.3409	0.2916
1	25.7			19.90		250.5889	220.2256
1	6.3			0.50		12.7449	20.7936
1	7.2			1.40		7.1289	13.3956
1	2.3			3.50		0.3249	2.4336
1	9.9			4.10		0.0009	0.9216
1	5.3			0.50		12.7449	20.7936
1	6.5			0.70		11.3569	19.0096
Additive	y_{2j}	\tilde{y}_2	$z_{2j} =$	$y_{2j} - 7.55$	$\bar{z}_{2.}$	$(z_{2j} - 8.88)^2$	$(z_{2j} - 5.06)^2$
2	0.2	7.55		7.35	8.88	2.3409	5.2441
2	11.3			3.75		26.3169	1.7161
2	0.3			7.25		2.6569	4.7961
2	17.1			9.55		0.4489	20.1601
2	51.0			43.45		1195.0849	1473.7921
2	10.1			2.55		40.0689	6.3001
2	0.3			7.25		2.6569	4.7961
2	0.6			6.95		3.7249	3.5721
2	7.9			0.35		72.7609	22.1841
2	7.2			0.35		72.7609	22.1841
Additive	y_{3j}	\tilde{y}_3	$z_{3j} =$	$y_{3j} - 9.15$	$\bar{z}_{3.}$	$(z_{3j} - 2.23)^2$	$(z_{3j} - 5.06)^2$
3	7.2	9.15		1.95	2.23	0.0784	9.6721
3	6.4			2.75		0.2704	5.3361
3	9.9			0.75		2.1904	18.5761
3	3.5			5.65		11.6964	0.3481
3	10.6			1.45		0.6084	13.0321
3	10.8			1.65		0.3364	11.6281
3	10.6			1.45		0.6084	13.0321
3	8.4			0.75		2.1904	18.5761
3	6.0			3.15		0.8464	3.6481
3	11.9			2.75		0.2704	5.3361
Total				5.06	1742.6	1978.4	

A SAS program to compute BFL test is given next:

```

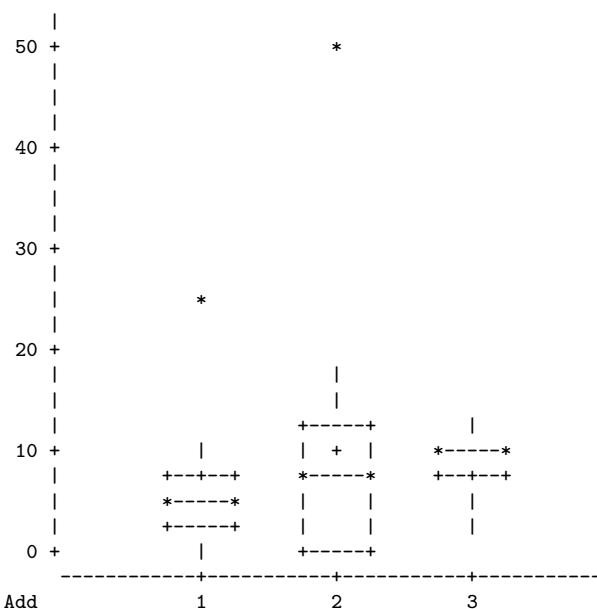
option ls=70 ps=50 nocenter nodate;
title 'BFL Test of Homogeneity of Variance';
data mpg;
infile '~/longneck/meth1/levene.dat';
input Add Y;
label Y = 'Percent Increase in MPG';

*generate Box Plots and Tests of Normality;
proc sort;by Add;
proc univariate def=5 plot normal; by Add;

*BFL Test of Equal Variances;
proc glm data=mpg;
class Add;
model Y = Add;
means Add/hovtest=bf;
run;

```

BFLTest of Homogeneity of Variance



BFL Test of Homogeneity of Variance

The GLM Procedure

Dependent Variable: Y Percent Increase in MPG

Source	DF	Sum of			F Value
		Squares	Mean Square	F Value	
Model	2	63.592667	31.796333	0.33	
Error	27	2634.046000	97.557259		
Corrected Total	29	2697.638667			

Brown and Forsythe's Test for Homogeneity of Y Variance ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of		F Value	Pr > F
		Squares	Mean Square		
Add	2	235.8	117.9	1.83	0.1803
Error	27	1742.6	64.5414		

The following R code will compute the Brown-Forsythe Levene test:

```
install.packages("car")
library(car)

y1=c(4.2,2.9,.2,25.7,6.3,7.2,2.3,9.9,5.3,6.5)
y2=c(.2,11.3,.3,17.1,51,10.1,.3,.6,7.9,7.2)
y3=c(7.2,6.4,9.9,3.5,10.6,10.8,10.6,8.4,6,11.9)
y=c(y1,y2,y3)
n1=length(y1)
n2=length(y2)
n3=length(y3)
grp=c(rep(1,n1),rep(2,n2),rep(3,n3))
grp=factor(grp)

leveneTest(y, grp)

Output from leveneTest:

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group  2  1.8268 0.1803
      27
```

The value of the p-value is identical to the value we obtained from SAS.

Longnecker RCode to compute BFL test for the Example:

BFL_H013Example.R

```
y1=c(4.2,2.9,.2,25.7,6.3,7.2,2.3,9.9,5.3,6.5)
y2=c(.2,11.3,.3,17.1,51,10.1,.3,.6,7.9,7.2)
y3=c(7.2,6.4,9.9,3.5,10.6,10.8,10.6,8.4,6,11.9)
n1=length(y1)
n2=length(y2)
n3=length(y3)
y=c(y1,y2,y3)
grp=c(rep(1,n1),rep(2,n2),rep(3,n3))
grp = as.factor(grp)

m1 = median(y1)
z1 = abs(y1-m1)
m2 = median(y2)
z2 = abs(y2-m2)
m3 = median(y3)
z3 = abs(y3-m3)
z = c(z1,z2,z3)

#BFL's Median Test on Raw Data

BFL = aov(z~grp)

summary(BFL)

  Df Sum Sq Mean Sq F value Pr(>F)
grp        2  235.8   117.91    1.827   0.18
Residuals 27 1742.6    64.54
```

von Neumann Test for Autocorrelation

If the data values have a normal distribution, it is often of interest to test if the correlation is of a specific form or pattern. A pattern which is often encountered is first-order autocorrelation, that is, the data values X_t , $t = 1, \dots, n$ have a time relationship: $X_t = \theta + \rho X_{t-1} + e_t$, where X_t is related to a constant θ , X_{t-1} , the previous value, ρ , the first-order autocorrelation coefficient, and e_t a random normal random variable which is independent of X_{t-1} . An estimate of ρ is given by

$$\hat{\rho} = \frac{\sum_{t=2}^n (X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

If $\rho = 0$ then the X_t 's are independent. Thus, we want a test of $H_0 : \rho = 0$ vs $H_a : \rho \neq 0$. The von Neumann test is given here:

1. Calculate $Q = \frac{\frac{1}{n-1} \sum_{t=2}^n (X_t - X_{t-1})^2}{\frac{1}{n} \sum_{t=1}^n (X_t - \bar{X})^2}$

- Q large implies Negative correlation and Q small indicates positive correlation ~~A~~

2. Determine $Q_{P,\alpha}$ and $Q_{N,\alpha}$ critical values for Q from the attached table for $n \leq 60$. For $n > 60$, we have the following asymptotic approximation:

$$Q_{P,\alpha} \approx \frac{2n}{n-1} - z_\alpha \frac{2}{\sqrt{n}} \text{ and } Q_{N,\alpha} \approx \frac{2n}{n-1} + z_\alpha \frac{2}{\sqrt{n}} \text{ and}$$

3. To test $H_0 : \rho = 0$ vs $H_a : \rho > 0$:

Reject H_0 if $Q < Q_{P,\alpha}$

4. To test $H_0 : \rho = 0$ vs $H_a : \rho < 0$:

Reject H_0 if $Q > Q_{N,\alpha}$

If the data X_1, X_2, \dots, X_n is positively correlated, then C.I.'s and tests of hypotheses constructed under the assumption that the data was independent will have the following flaws:

$$Var(\bar{X}) = \sigma^2/n + \Delta \text{ where } \Delta > 0 \text{ if } X_i \text{ are positively correlated}$$

$$\widehat{SE}(\bar{X}) = \sqrt{S^2/n + \Delta} > S/\sqrt{n}$$

1. The coverage probability will be less than the nominal value for the C.I., i.e.,

$$P[\theta \in C.I.] \leq 1 - \alpha$$

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} \text{ is narrower than } \bar{X} \pm t_{\alpha/2} \sqrt{S^2/n + \Delta}$$

The standard C.I. is too narrow to achieve the specified level of coverage

2. Tests of hypotheses will have inflated Type I error rates, i.e.,

$$\alpha_{Pos.Corr} \geq \alpha_{Independence}$$

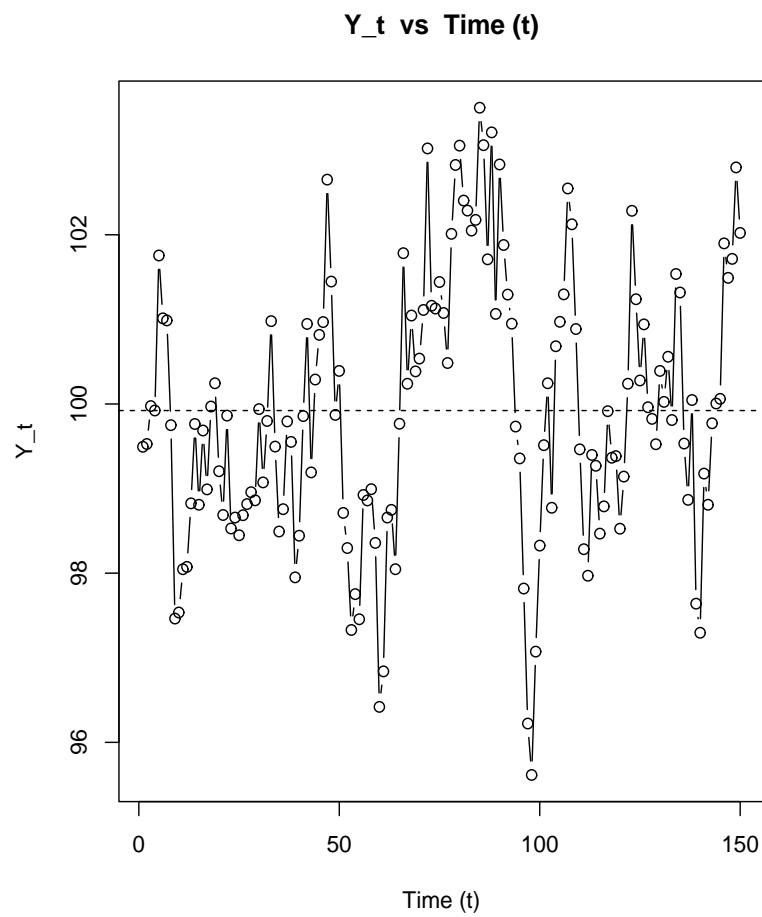
$$t = \frac{\bar{X} - \mu_o}{S/\sqrt{n}} > \frac{\bar{X} - \mu_o}{SE(\bar{X})}$$

When H_0 is true, the standard t-test will reject H_0 more often than the corrected t-test

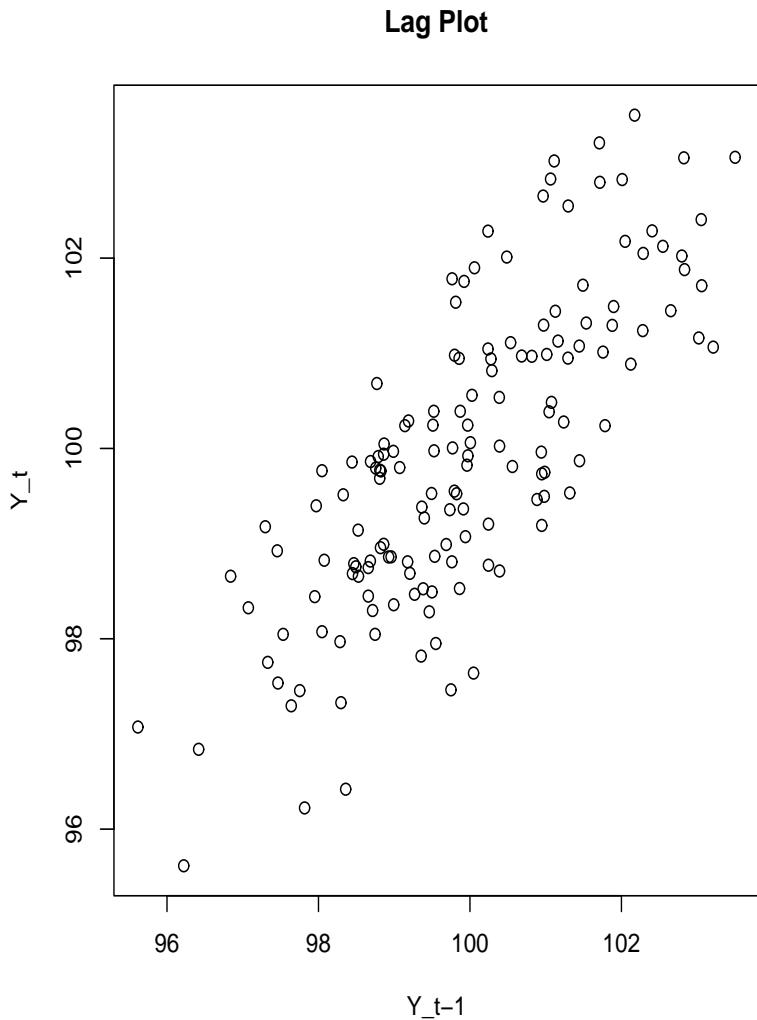
5%, 1%, and .1% Points of the Von Neumann Ratio^a

Number of Observations	5 %	1 %	.1 %	5 %	1 %	.1 %
	<i>One-tailed test against positive correlation</i>			<i>One-tailed test against negative correlation</i>		
4	1.0406	.8341	.7864	4.2927	4.4992	4.5469
5	1.0255	.6724	.5201	3.9745	4.3276	4.4799
6	1.0682	.6738	.4361	3.7318	4.1262	4.3639
7	1.0919	.7163	.4311	3.5748	3.9504	4.2356
8	1.1228	.7575	.4612	3.4486	3.8139	4.1102
9	1.1524	.7974	.4973	3.3476	3.7025	4.0027
10	1.1803	.8353	.5351	3.2642	3.6091	3.9093
11	1.2962	.8706	.5717	3.1938	3.5294	3.8283
12	1.2301	.9033	.6062	3.1335	3.4603	3.7574
13	1.2521	.9336	.6390	3.0812	3.3996	3.6944
14	1.2725	.9618	.6702	3.0352	3.3458	3.6375
15	1.2914	.9880	.6999	2.9943	3.2977	3.5858
16	1.3090	1.0124	.7281	2.9577	3.2543	3.5386
17	1.3253	1.0352	.7548	2.9247	3.2148	3.4952
18	1.3405	1.0566	.7801	2.8948	3.1787	3.4552
19	1.3547	1.0766	.8040	2.8675	3.1456	3.4182
20	1.3680	1.0954	.8265	2.8425	3.1151	3.3840
21	1.3805	1.1131	.8477	2.8195	3.0869	3.3523
22	1.3923	1.1298	.8677	2.7982	3.0607	3.3228
23	1.4035	1.1456	.8866	2.7784	3.0362	3.2953
24	1.4141	1.1606	.9045	2.7599	3.0133	3.2695
25	1.4241	1.1748	.9215	2.7426	2.9919	3.2452
26	1.4336	1.1883	.9378	2.7264	2.9718	3.2222
27	1.4426	1.2012	.9535	2.7112	2.9528	3.2003
28	1.4512	1.2135	.9687	2.6969	2.9348	3.1794
29	1.4594	1.2252	.9835	2.6834	2.9177	3.1594
30	1.4672	1.2363	.9978	2.6707	2.9016	3.1402
31	1.4746	1.2469	1.0115	2.6587	2.8864	3.1219
32	1.4817	1.2570	1.0245	2.6473	2.8720	3.1046
33	1.4885	1.2667	1.0369	2.6365	2.8583	3.0882
34	1.4951	1.2761	1.0488	2.6262	2.8451	3.0725
35	1.5014	1.2852	1.0603	2.6163	2.8324	3.0574
36	1.5075	1.2940	1.0714	2.6068	2.8202	3.0429
37	1.5135	1.3025	1.0822	2.5977	2.8085	3.0289
38	1.5193	1.3108	1.0927	2.5889	2.7973	3.0154
39	1.5249	1.3188	1.1029	2.5804	2.7865	3.0024
40	1.5304	1.3266	1.1128	2.5722	2.7760	2.9898
41	1.5357	1.3342	1.1224	2.5643	2.7658	2.9776
42	1.5408	1.3415	1.1317	2.5567	2.7560	2.9658
43	1.5458	1.3486	1.1407	2.5494	2.7466	2.9545
44	1.5506	1.3554	1.1494	2.5424	2.7376	2.9436
45	1.5552	1.3620	1.1577	2.5357	2.7289	2.9332
46	1.5596	1.3684	1.1657	2.5293	2.7205	2.9232
47	1.5638	1.3745	1.1734	2.5232	2.7125	2.9136
48	1.5678	1.3802	1.1807	2.5173	2.7049	2.9044
49	1.5716	1.3856	1.1877	2.5117	2.6977	2.8956
50	1.5752	1.3907	1.1944	2.5064	2.6908	2.8872
51	1.5787	1.3957	1.2010	2.5013	2.6842	2.8790
52	1.5822	1.4007	1.2075	2.4963	2.6777	2.8709
53	1.5856	1.4057	1.2139	2.4914	2.6712	2.8630
54	1.5890	1.4107	1.2202	2.4866	2.6648	2.8553
55	1.5923	1.4156	1.2264	2.4819	2.6585	2.8477
56	1.5955	1.4203	1.2324	2.4773	2.6524	2.8403
57	1.5987	1.4249	1.2383	2.4728	2.6465	2.8331
58	1.6019	1.4294	1.2442	2.4684	2.6407	2.8260
59	1.6051	1.4339	1.2500	2.4640	2.6350	2.8190
60	1.6082	1.4384	1.2558	2.4596	2.6294	2.8120

The following plot displays a time series of 150 data values. We want to determine if the data is correlated over time. From the plot, the data appears to be positively correlated.



The following lag plot displays the positive correlation in the 150 data values.



The following R code plots the data and computes the von Neumann:

R CODE FOR DISPLAYING CORRELATION PLOTS FOR DATA:

```
postscript("u:/meth1/psfiles/corr,p1.ps",height=7,width=6,horizontal=F)

y  = matrix(0,150,1)
y  = scan("u:/meth1/Rfiles/corr.dat")

n  = length(y)
ytime1 = ts(y,start=1,frequency=1)
plot.ts(ytime1,type="b",ylab="Y_t",xlab="t",main="Y_t vs t")
ymean = mean(y)
abline(h=ymean,lty=2)

postscript("u:/meth1/psfiles/corr,p2.ps",height=7,width=6,horizontal=F)

yt  = y[2:150]
ytlag1 = y[1:149]
plot(ytlag1,yt,main="Lag Plot",ylab="Y_t",xlab="Y_t-1")

#Calculation of von Neumann Statistics
dif1  = (yt-ytlag1)^2
num1  = sum(dif1)
y2  = (y-ymean)^2
den1  = sum(y2)
Q  = (num1/(n-1))/(den1/n)
prd1  = (yt-ymean)*(ytlag1-ymean)
prdsum1 = sum(prd1)
rho1  = prdsum1/den1
rho1
Q

graphics.off()

#Output from Corrplot.s:
#[1] 0.7746074 • PNO
#[1] 0.441409 — Q
```

Runs Test for Correlation

When the data is nonnormal distributed, the von Neumann test is invalid. An alternative distribution-free procedure, the Runs Test, will be presented. Let X_1, X_2, \dots, X_T be T equally spaced observations on a random process. To test if the t observations are correlated:

1. Center the observations: $Y_t = X_t - \bar{X}$, where $\bar{X} = \frac{1}{T} \sum_{t=1}^T X_t$
2. Count the number of runs (R), where a run is defined as a sequence of observations of all positive values or all negative values
3. Count the number of positive Y_t s (n_1) and the number of negative Y_t s (n_2) *use rule*
4. If $n_1 \leq 20$ and $n_2 \leq 20$, then the data indicates that X_t is correlated if $R \leq R_L$ or $R \geq R_U$, where R_L and R_U are values given in the table given in *Annals of Mathematical Statistics*, **14**, pp. 66-87.
5. Large sample size critical values are obtained by declaring that the data indicates that X_t is correlated if $Z > Z_{\alpha/2}$, where

$$Z = \frac{|R - \mu| - 0.5}{\sigma}, \quad \mu = 1 + \frac{2n_1 n_2}{n_1 + n_2}, \quad \sigma^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the $N(0, 1)$ distribution.

Table A30(a) Lower Critical Values of r for the Runs Test* ($\alpha = 0.05$)

		Lower Critical Value																	
n_1	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2									2	2	2	2	2	2	2	2	2	2	2
3			2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3
4		2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4
5		2	2	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
7		2	2	3	3	3	4	4	5	5	5	5	6	6	6	6	6	6	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9		2	3	3	4	4	5	5	5	6	6	7	7	7	7	8	8	8	8
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12	2	2	3	4	4	5	6	6	7	8	8	9	9	9	9	9	9	10	10
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
14	2	2	3	4	5	5	6	7	7	8	8	8	8	10	10	10	10	11	11
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	12
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14

*Any value of r that is equal to or smaller than that shown in the body of this table for given values of n_1 and n_2 is significant at the 0.05 level. Tabled values are appropriate for one-tailed test at stated significance level or two-tailed test at twice the significance level.

Source: Adapted from Swed, F. S. and Eisenhart, C. (1943). "Tables for Testing Randomness of Grouping in a Sequence of Alternatives," *Annals of Mathematical Statistics*, 14, 66-87. Used by permission of the Institute of Mathematical Statistics.

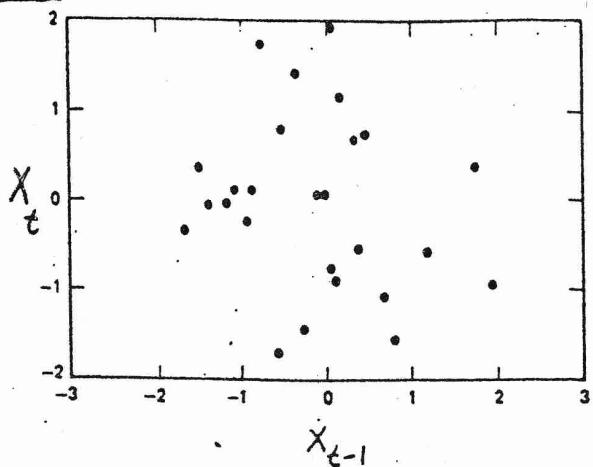
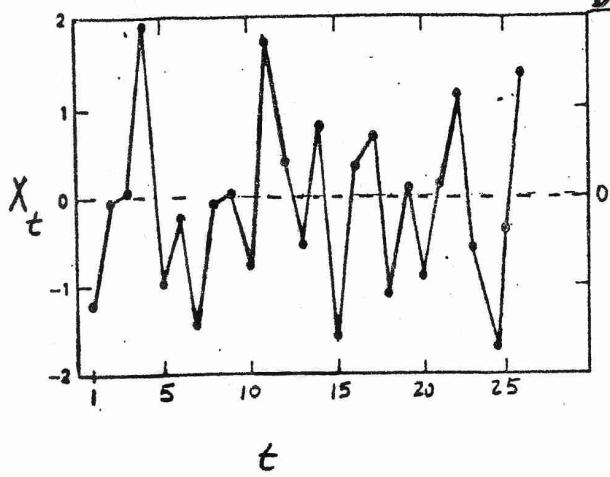
Table A30(b) Upper Critical Values of r for the Runs Test* ($\alpha = 0.05$)

		Upper Critical Value																	
n_1	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2																			
3				9	9														
4				9	9														
5			9	10	10	11	11												
6		9	10	11	12	12	13	13	13	13	13								
7			11	12	13	13	14	14	14	14	14	15	15	15					
8			11	12	13	14	14	15	15	15	16	16	16	16	17	17	17	17	17
9				13	14	14	15	16	16	16	17	17	18	18	18	18	18	18	18
10				13	14	15	16	16	17	17	18	18	18	19	19	19	20	20	20
11				13	14	15	16	17	17	18	19	19	19	20	20	20	21	21	21
12				13	14	16	16	17	18	19	19	19	20	20	21	21	21	22	22
13				15	16	17	18	19	19	20	20	21	21	22	22	22	23	23	23
14				15	16	17	18	19	20	20	21	21	22	22	23	23	23	24	24
15					15	16	18	18	19	20	21	21	22	22	23	23	24	24	25
16						17	18	19	20	21	21	22	23	23	24	24	25	25	25
17						17	18	19	20	21	22	23	23	24	25	25	26	26	26
18						17	18	19	20	21	22	23	24	25	25	26	26	27	27
19						17	18	20	21	22	23	23	24	25	26	26	27	27	27
20						17	18	20	21	22	23	24	25	25	26	27	27	28	

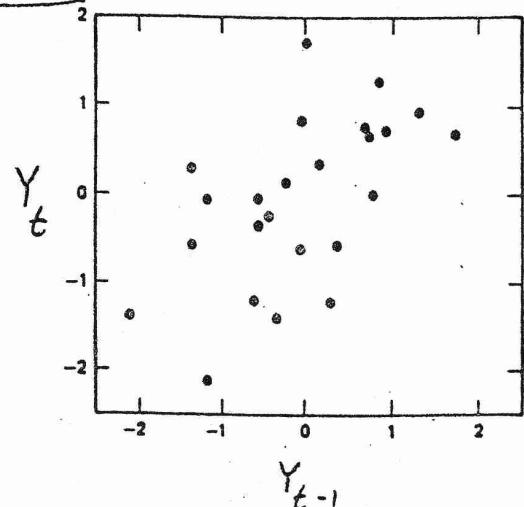
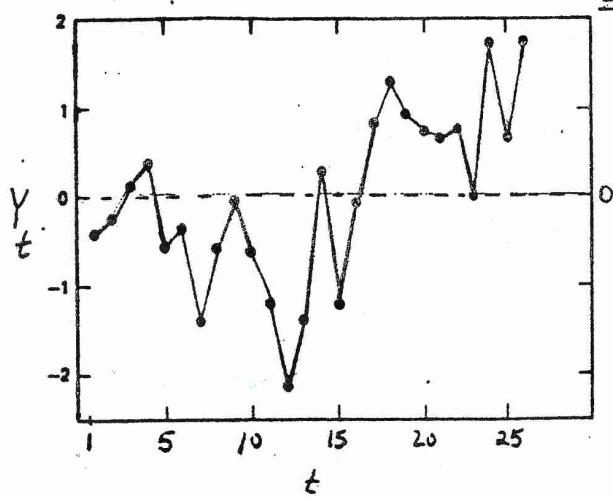
*Any value of r that is equal to or greater than that shown in the body of this table for given values of n_1 and n_2 is significant at the 0.05 level. Tabled values are appropriate for one-tailed test at stated significance level or two-tailed test at twice the significance level.

Source: Adapted from Swed, F. S. and Eisenhart, C. (1943). "Tables for Testing Randomness of Grouping in a Sequence of Alternatives," *Annals of Mathematical Statistics*, 14, 66-87. Used by permission of the Institute of Mathematical Statistics.

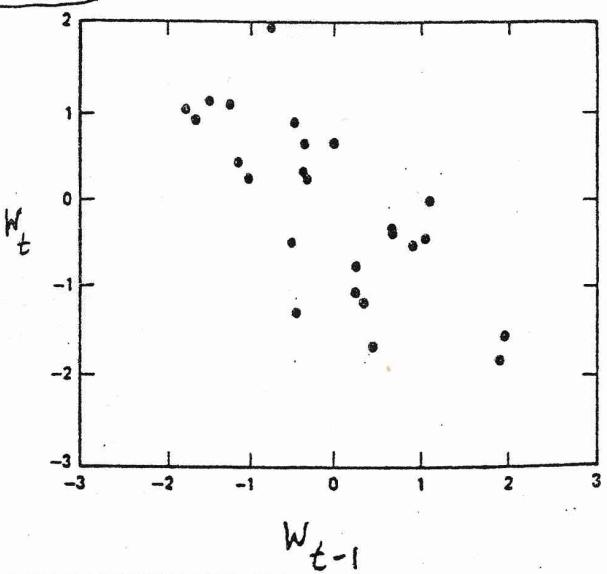
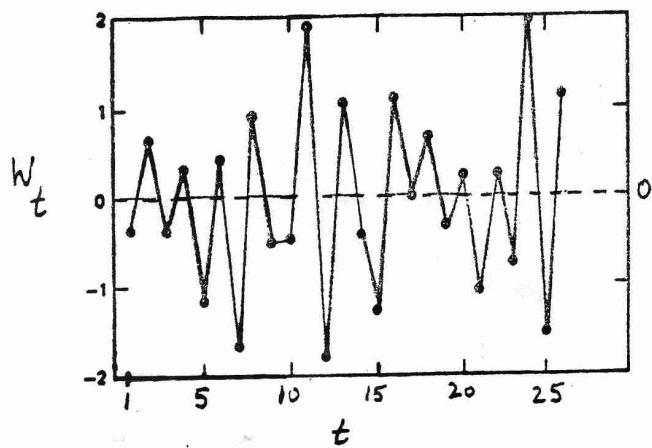
Data Set #1



Data Set #2



Data Set #3



Three Time Series from Applied Regression Analysis, 3rd Ed., Draper - Smith

Refer to the data plotted on pages 80-81 and conduct a runs test for correlated data using the following R code:

```

y = rep(0,150)
y = scan("u:/meth1/Rfiles/corr.dat")
n=length(y)
means = mean(y)
diff = y-means

n.neg = rep(0,n)
n.pos = rep(0,n)
n.neg = length(diff[diff<0])
n.pos = length(diff[diff>0])

numb.runs = 1
for (j in 2:n) {
  if (sign(diff[j]) != sign(diff[j-1])) {numb.runs = numb.runs + 1}
}

runs.result = as.data.frame(cbind(numb.runs, n.pos, n.neg))
names(runs.result) = c("No. runs", "N+", "N-")
runs.result

# No. runs  N+  N-
#    30      70   80

mu = 1 + (2*n.neg*n.pos)/(n.neg + n.pos)
sig2 = (2*n.neg*n.pos*(2*n.neg*n.pos - n.neg - n.pos))/((n.neg + n.pos)^2*(n.neg + n.pos-1))

z = (abs(numb.runs-mu)-.5)/sqrt(sig2)
pvalue = 2*(1-pnorm(abs(z)))

results = cbind(mu,sig2,z,pvalue)
results

#          mu        sig2         z        pvalue
#    75.66667  36.91573  7.433824  1.054712e-13

```

Using the formulas on page 83, we have

$$\begin{aligned}\mu &= 1 + \frac{2n_1n_2}{n_1 + n_2} = 1 + \frac{2(70)(80)}{70 + 80} = 75.66667 \\ \sigma^2 &= \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} = \frac{2(70)(80)(2(70)(80) - 70 - 80)}{(70 + 80)^2(70 + 80 - 1)} = 36.91573 \\ Z &= \frac{|R - \mu| - 0.5}{\sigma} = \frac{|30 - 75.66667| - 0.5}{\sqrt{36.91573}} = 7.433824\end{aligned}$$

$$\text{p-value} = 2P[Z > 7.433824] = 2(1 - pnorm(7.433824)) = 1.054712e-13$$

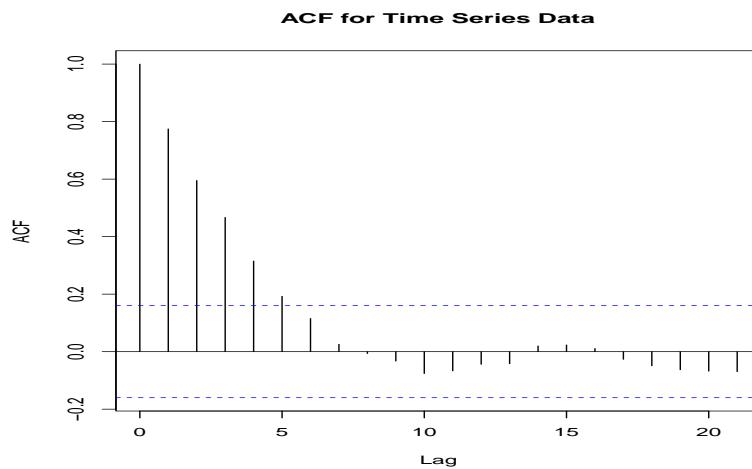
There is substantial evidence that the 150 data values are correlated.

Auto Correlation Function

Using the autocorrelation function from R:

```
acf(y,main="ACF for Time Series Data")
```

The following plot of the lag k correlations can be obtained with 95% C.I.'s:



Thus, we can determine that there is a strong positive correlation in the data because $\rho_1, \rho_2, \rho_3, \rho_4, \rho_5$ all have values outside the limits.

The estimated values of $\rho_k = \text{corr}(Y_t, Y_{t+k})$ are as follows:

Estimated lag k correlations:							
lag	1	2	3	4	5	6	7
$\hat{\rho}_k$	0.775	0.595	0.467	0.315	0.193	0.116	0.026

testing multiple pairs, i.e. is $\mu_j = \mu_i$
 for each i, j pair we accumulate a Type I error (α), type II error.
 The test adjusts for the accumulation of α .

A large percentage of the dietary energy in the bodies of infants is provided by lipids. Lipids are a class of hydrocarbon-containing organic compounds. The following data on total polyunsaturated fats(%) was reported for infants who were randomized to four different feeding regimens: BM (breast milk), CO corn-oil-based formula, SO (soy-oil-based formula), or SMO (soy-marine-oil based formula).

Regimen	$Y_1 - Y_n$	\bar{Y}_i	S_i	95% C.I. on μ_i
BM	41.8, 37.1, 40.7, 46.8, 44.3, 45.7, 41.5, 36.1, 44.0, 40.8, 45.4, 36.5, 41.4, 42.7, 44.4, 42.4, 45.6, 48.5	42.5	3.49	(39.06, 46.01)
CO	42.0, 43.7, 40.4, 40.0, 34.7, 41.3, 34.3, 36.2, 38.4, 39.8, 38.9, 45.6, 39.7	40.4	3.08	(37.06, 43.71)
SO	46.2, 44.2, 41.2, 46.1, 42.0, 47.8, 41.7, 39.8, 38.4, 46.7, 42.8, 46.4, 39.0, 43.0, 48.5, 48.1, 45.3	44.0	3.25	(40.61, 47.30)
SMO	46.5, 49.9, 43.7, 47.1, 49.5, 48.8, 54.7, 47.0, 42.6, 48.8, 50.2, 49.2, 46.6, 51.3	48.3	3.06	(45.00, 51.56)

The researcher wanted to determine if there was significant evidence at the $\alpha = .05$ level that the four Regimens produced different mean percentages of polyunsaturated fat in the infants.

The appropriate analysis would be to test for a difference in the six pairs of treatment means, $\mu_i - \mu_j$.

In general, if we had t populations, we would want to determine which pairs of treatment means are different. The most widely procedure is the Tukey Honest Significant Difference (HSD). HSD and a procedure by Scheffé are two of the very few multiple comparison procedures which have an exact value for the experimentwise error rate α_{Exp} , that is, $\alpha_{Exp} = \alpha_o$, whereas the Bonferroni procedure only places a bound, $\alpha_{Exp} \leq \alpha_o$.

Tukey's HSD procedure requires that the data have equal sample sizes $r = n_1 = n_2 = \dots = n_t$, equal standard deviations, $\sigma_1 = \sigma_2 = \dots = \sigma_t$, and the data are random samples from normal distributions.

Tukey's procedure is based on the studentized range statistic:

$$q(t, \nu) = \frac{\bar{y}_{max} - \bar{y}_{min}}{\hat{\sigma}_p \sqrt{\frac{1}{r}}}$$

$$\bar{y}_{min} = \min\{\bar{y}_1, \bar{y}_2, \dots, \bar{y}_t\}, \quad \bar{y}_{max} = \max\{\bar{y}_1, \bar{y}_2, \dots, \bar{y}_t\}, \quad \hat{\sigma}_p^2 = \frac{1}{\nu} \sum_{i=1}^t (n_i - 1) S_i^2.$$

Define

$$HSD(t, \nu) = q(\alpha_o, t, \nu) \sqrt{\frac{1}{2} \left(\widehat{SE}(\bar{y}_i - \bar{y}_j) \right)^2} = q(\alpha_o, t, \nu_2) \sqrt{\frac{1}{2} \left(\frac{2\hat{\sigma}_p^2}{r} \right)} \Rightarrow$$

$$HSD(t, \nu) = q(\alpha_o, t, \nu) \hat{\sigma}_p \sqrt{1/r}$$

where $\nu = t(r - 1)$, t is number of treatment means to be compared, r is the common sample size, α_o is the specified value for α_{Exp} , and $q(\alpha_o, t, \nu)$ is the solution to the equation:

$$P \left[\frac{|(\bar{y}_k - \mu_k) - (\bar{y}_h - \mu_h)|}{\hat{\sigma}_p \sqrt{1/r}} \leq q(\alpha_o, t, \nu_2) \text{ for all } k > h \right] = 1 - \alpha_o$$

Values for $q(\alpha_o, t, \nu)$ are given by the R function `qtukey(1 - alpha, t, nu)`.

For example, for $\alpha = .05$, $t = 5$ treatments, $r=4$, and $\nu = 5(4 - 1) = 15$,

`qtukey(.95, 5, 15) = 4.367.`

Tukey HSD Procedure:

- ~~1.~~ State there is significant evidence at level α_o that the treatment means μ_i and μ_j are different if

$$|\bar{y}_i - \bar{y}_j| \geq HSD(t, \nu)$$
2. Simultaneous $100(1 - \alpha_o)\%$ confidence intervals for the absolute difference in all pairs of treatment means $|\mu_i - \mu_j|$ are given by

$$|\bar{y}_i - \bar{y}_j| \pm HSD(t, \nu)$$
3. When the sample sizes are unequal, a procedure called the **Tukey-Kramer procedure** replaces
 ~~$\sqrt{1/r}$ with $\sqrt{\frac{1}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$ and $\nu = \sum_{i=1}^t (n_i - 1) = N - t$, with $N = \sum_{i=1}^t n_i$.~~
This yields a multiple comparison procedure in which $\alpha_{Exp} \leq \alpha_o$.

Dietary Energy Example

Determine if there is a difference in the mean amount of polyunsaturated fats in the four regimens. Using an experimentwise Type I error rate of .05, a pairwise comparison of the four regimens will be made using the Tukey HSD.

In this example, we have the following: $t = 4$, $n_1 = 18$, $n_2 = 13$, $n_3 = 17$, $n_4 = 14$. The Shapiro-Wilk test with p-values of .46, .96, .35, .81 and the four normal reference plots confirmed that a normal distribution is an excellent fit to all four populations. The BFL test had a p-value of .79 for the test of equal variances. Thus, the Tukey procedure is appropriate. However, because the sample sizes are unequal, the Tukey-Kramer modified HSD will be implemented.

$$t = 4, \quad \nu = (18 - 1) + (13 - 1) + (17 - 1) + (14 - 1) = 62 - 4 = 58, \quad qtukey(.95, 4, 58) = 3.741$$

$$\hat{\sigma}_p = \sqrt{\frac{1}{\nu} \sum_{i=1}^4 (n_i - 1) S^2} = \sqrt{611.6 / 58} = 3.247$$

$$HSD(t, \nu) = qtukey(.95, 4, 58) \hat{\sigma}_p \sqrt{\frac{1}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} = (3.741)(3.247) \sqrt{\frac{1}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Declare there is a significant difference in a pair of means μ_i and μ_j if $|\bar{X}_i - \bar{X}_j| \geq HSD$

Pair	(n_i, n_j)	HSD	$\bar{X}_i - \bar{X}_j$	Sim. C.I. on $\mu_i - \mu_j$	Sig. Evid. of Diff. in $\mu_i - \mu_j$?
CO-BM	(13, 18)	3.126	-2.1	(-5.2, 1.0)	No
SMO-BM	(14, 18)	3.061	5.8	(2.7, 8.9)	Yes
SO-BM	(17, 18)	2.905	1.5	(-1.4, 4.4)	No
SMO-CO	(14, 13)	3.308	7.9	(4.6, 11.2)	Yes
SO-CO	(17, 13)	3.165	3.6	(0.4, 6.8)	Yes
SO-SMO	(17, 14)	3.100	-4.3	(-7.4, -1.2)	Yes

Notice that 0 is contained in all the C.I.'s for those pairs of means for which there is not significantly evidence of a difference in the means based on HSD.

The next page contains the R code to perform the Tukey HSD procedure, **TukeyHSD.R**.

```

# TukeyHSD.R in eCampus R Files Folder:

y1 = c( 41.8,37.1,40.7,46.8,44.3,45.7,41.5,36.1,44.0,40.8,45.4,36.5,41.4,42.7,44.4,42.4,45.6,48.5)
y2 = c(42.0,43.7,40.4,40.0,34.7,41.3,44.3,36.2,38.4,39.8,38.9,45.6,39.7)
y3 = c(46.2,44.2,41.2,46.1,42.0,47.8,41.7,39.8,38.4,46.7,42.8,46.4,39.0,43.0,48.5,48.1,45.3)
y4 = c(46.5,49.9,43.7,47.1,49.5,48.8,54.7,47.0,42.6,48.8,50.2,49.2,46.6,51.3)

y = c(y1,y2,y3,y4)

R1 = rep("BM",18)
R2 = rep("CO",13)
R3 = rep("SO",17)
R4 = rep("SMO",14)
REGIMEN = c(R1,R2,R3,R4)
REGIMEN = as.factor(REGIMEN)
# Shapiro-Wilk Test of Normality

shapiro.test(y1)
shapiro.test(y2)
shapiro.test(y3)
shapiro.test(y4)
# BFL test of equal variance

z1<-y[1:18]
m1<-median(z1)
z1<-abs(z1-m1)
z2<-y[19:31]
m2<-median(z2)
z2<-abs(z2-m2)
z3<-y[32:48]
m3<-median(z3)
z3<-abs(z3-m3)
z4<-y[49:62]
m4<-median(z4)
z4<-abs(z4-m4)
z<-c(z1,z2,z3,z4)
lev1 <- aov(z~REGIMEN)
summary(lev1)
#Tukey-Kramer HSD

model = lm(y ~ REGIMEN)
summary(model)
AOV = aov(model)
summary(AOV)

TukeyHSD(AOV,"REGIMEN",ordered=FALSE,.95)
plot(TukeyHSD(AOV,"REGIMEN"))

```

95% family-wise confidence level

