Stat 310 Homework 9 Key

Chapter 10, problem 16. Chapter 11, problems 1, 13 (see 12), 15, 18, 19, 23, 24, 34, 37. Due 12/7/99.

10.16 Suppose that F is $N(0,1)$ and G is $N(1,1)$. Sketch a $Q-Q$ plot. Repeat for G being $N(1,4)$.

11.1 A computer was used to generate four random numbers from a normal distribution with a set mean and variance: 1.1650, 0.6268, 0.0751, 0.3516. Five more random normal numbers with the same variance but perhaps a different mean were then generated (the mean may or may not actually be different): 0.3035, 2.6961, 1.0591, 2.7971, 1.2641.

a) What do you think the means of the random number generators were? What do you think the difference of means was? Well, our best guesses here will simply be the sample means and the difference of sample means. using X for the first sample and Y for the second, our guesses are:

$$\begin{aligned} \hat{\mu}_X &= \bar{x} = 0.5546, \\ \hat{\mu}_Y &= \bar{y} = 1.6240, \\ \widehat{\mu_X - \mu_Y} &= \bar{x} - \bar{y} = -1.0694. \end{aligned}$$

b) What do you think the variance of the random number generator was? In this case, we know that the variance was the same for both samples, and consequently we can combine the information about the variance across samples to get a better estimate than either one alone; this is the *pooled* variance estimate. The pooled estimate is

$$\begin{aligned} s_p^2 &= \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} \\ &= \frac{3(0.2163) + 4(1.1795)}{7} = 0.7667. \end{aligned}$$

c) What is the estimated standard error of your estimate of the difference of the means? Well, since we are using means we get to reduce the pooled variance by terms related to the sample size:

$$\text{est. std. err} = s_p\sqrt{\frac{1}{n}+\frac{1}{m}} = \sqrt{0.7667}\sqrt{\frac{1}{4}+\frac{1}{5}} = 0.5874.$$

d) Form a 90% confidence interval for the difference of means of the random number generators. Well, a 90% confidence interval for the true mean difference is given by

$$(\bar{x}-\bar{y}) \pm t_{n+m-2,0.95}\, s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$$

which, upon plugging in the numbers that we've found, gives

$$-1.0694 \pm 1.895 * 0.5874 = (-2.1824, 0.0437).$$

e) In this situation is it more appropriate to use a one-sided test or a two-sided test of the equality of the means? Here, a two-sided test is more appropriate since we only know that the means may be different; we don't have any *a priori* knowledge of which direction the difference is likely to be in.

f) What is the p-value of a two-sided test that the means were the same? Well, the test statistic is

$$\frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = -\frac{1.0694}{0.5874} = -1.8206,$$

and the null distribution is t_7. Using table 4, we see that 1.82 is between $t_{.90}$ and $t_{.95}$ for this distribution, so the p-value is somewhere between 0.1 and 0.2. Using matlab, the exact p-value is $2 * tcdf(-1.8206, 7) = 0.1115$.

g) Would the hypothesis that the means were the same versus a two-sided alternative be rejected at the significance level $\alpha = 0.1$? No. We can note this two ways - first, the p-value is the smallest value of α for which the null will be rejected. As the p-value is greater than 0.1, we will not reject. Second, the 90% confidence interval that we constructed in d) contains the value zero, and given the duality between confidence intervals and hypothesis tests, we would not reject any hypothesized value of the mean lying in this interval.

h) Suppose you know that the variance of the normal distribution was $\sigma^2 = 1$. How would your answers to the preceding questions change? Well, in this case we know the variance exactly so we can use a z-test instead of a t-test. Our best guesses as to the values of the means and the value of the mean difference (part a) do not change at all; these are not affected by the variance. Our answer to part b) is now 1; perfect knowledge has now been vouchsafed to us. The standard error associated with the difference in means (part c) is now $1 * sqrt(1/4 + 1/5) = 0.6708$, and this is exact given the sample sizes, as we don't need to estimate anything. The 90% confidence interval is now $-1.0694 \pm 1.645 * 0.6708 = (-2.1728, 0.0341)$, slightly narrower than what we had before as the increase in variance is offset by the smaller critical values needed for the normal distribution as opposed to the t distribution. It is still more appropriate to use a two-sided interval (part e) as the variance does not affect our lack of initial knowledge as to the direction of the deviation. As to the p-value (part f), our test statistic is now $-1.0694/0.6708 = -1.5942$, and this is from a standard normal distribution; from table 2 we see that the chance of a normal distrbution being less than this is $1 - 0.9441 = 0.0559$ and the p-value is twice this, or 0.1118. Finally, our answer to part g) remains the same, we would not reject, and the reasons for this are exactly as given above.

11.13 Let $X_1, \ldots, X_{25}$ be i.i.d. $N(.3, 1)$. Consider testing the null hypothesis $H_0 : \mu = 0$ versus $H_A : \mu > 0$ at significance level $\alpha = 0.05$. Compare the power of the sign test and the power of the test based on normal theory assuming that σ is known.

Well, first off, the sign test is a nonparametric test which does not invoke any knowledge of the underlying distribution. Thus, it sacrifices information, and here, where that information is available, we might expect that to cost us. In short, we expect going into this problem that the power of the sign test will be less than the power of the normal theory test. Let's look at the sign test first. Under the null hypothesis, the center of the distribution is at zero.

Thus, the chance of any given X falling on either side of zero is precisely $1/2$, and this change is the same for all of the X_i and the outcomes are independent. Thus, letting Y denote the number of X_i's that are greater than zero, the distribution of Y is Binomial$(25, 0.5)$. As the alternative is one-sided, we will reject only if Y is too large. Checking table 1, we see that rejecting when $Y \geq 17$ corresponds to $\alpha = 0.054$, which might be close enough; rejecting when $Y \geq 18$ corresponds to $\alpha = 0.022$. In this case, we'd probably take the first option. Now, given the true distribution (a specific setting of the alternative) the chance of any given X_i exceeding zero is again the same and independent for each, but the chance of "success" is now $P(X_i > 0) = P(Z > -0.3) = 0.6179$. Looking at table 1 again, we see values for $p = 0.6$, which is reasonably close; the power is $1 - 0.726 = 0.274$ for the first option and $1 - .846 = 0.154$ for the second option. Using matlab, the power using the first option is $1 - binocdf(16, 25, 0.6179) = 0.3378$ and $1 - binocdf(17, 25, 0.6179) = 0.2007$ using the second. Now, using normal theory we know that $\bar{X} \sim N(0, 1/25)$ under the null, so we would reject if $\bar{X}$ exceeded $1.645 * (1/5) = 0.3290$. Under the alternative, $\bar{X} \sim N(0.3, 1/25)$, and $P(\bar{X} > 0.3290) = P(Z > 0.0290/5) = P(Z > 0.0058) = 0.4977$. The z-test is quite a bit more powerful than the sign test! The sign test actually tosses away a lot of other information that the signed rank test, for example, makes use of - the rough magnitudes of the sizes of deviations above and below zero, for example. On the other hand, the sign test is *very* easy to use, and versions of it can be developed for any quantile whatsoever.

11.15 Suppose that n measurements are to be taken under a treatment condition and another n measurements are to be taken independently under a control condition. It is thought that the standard deviation of a single observation is about 10 under both conditions. How large should n be so that a 95% confidence interval for $\mu_X - \mu_Y$ has a width of 2? Use the normal distribution rather than the t distribution, since n will turn out to be rather large.

Well, in testing for the true mean difference our test statistic is $\bar{X} - \bar{Y}$, which has a normal distribution with parameters $N(\mu_X - \mu_Y, (\sigma_X^2 + \sigma_Y^2)/n)$. A 95% confidence interval for the true difference will thus be of the form

$$(\bar{X} - \bar{Y}) \pm z_{0.975}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}},$$

the width of which is simply

$$2 \pm z_{0.975}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}} = 2(1.96)\sqrt{\frac{100}{n} + \frac{100}{n}} = 39.2\sqrt{\frac{2}{n}}.$$

Setting this equal to 2, we get

$$\begin{aligned} \sqrt{n} &= 39.2/sqrt2 = 27.7186, \\ n &= 768.32 \approx 768, \end{aligned}$$

since n must be an integer.

11.18 Two independent samples are to be compared to see if there is a difference in the population means. If a total of m subjects are available for the experiment, how should this

total be allocated between the two samples in order to (a) provide the shortest confidence interval for $\mu_X - \mu_Y$ and (b) make the test of $H_0 : \mu_X = \mu_Y$ as powerful as possible? Assume that the observations in the two samples are normally distributed with the same variance.

Well, we showed in class that the form of the most powerful test of $\mu_X = \mu_Y$ rejected when $(\bar{X} - \bar{Y})^2$ was too large. To decide what constitutes "too large", we need to consider the distribution of $\bar{X} - \bar{Y}$. Assuming that the two samples are normally distributed with the same variance, we have

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma^2}{n_X} + \frac{\sigma^2}{m - n_X}\right)$$

so for an α level confidence interval we would use

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n_X} + \frac{1}{m - n_X}}.$$

Now, the only effect of changing the allocation of sample units between X and Y is on the width of the confidence interval - it does not affect the center, and the shape of the interval is symmetric. Thus, making the confidence interval smaller corresponds to shrinking the width, thus increasing the size of the rejection region and consequently increasing the power of the test. Thus, the value of n_X that provides the shortest c.i. also makes the test as powerful as possible. To achieve this minimum, we want to choose n_X to minimize

$$\frac{1}{n_X} + \frac{1}{m - n_X};$$

differentiating with respect to n_X gives

$$-\frac{1}{n_X^2} + \frac{1}{m - n_X}^2$$

and setting this equal to zero gives $\hat{n}_X = m/2$, so we want to divide the sample as evenly as possible. A quick check shows that the second derivative is uniformly positive so that this is indeed a minimum. Note that this even splitting would not be optimal if we knew $\sigma_X^2 and \sigma_Y^2$, and the two were different. In that case, the optimal allocation would be the root of

$$-\frac{\sigma_X^2}{n_X^2} + \frac{\sigma_Y^2}{m - n_X}^2.$$

11.19 A study was done to comapre the performances of engine bearings made of different compounds (McCool 1979). Ten bearings of each type were tested. The following table gives the times until failure (in units of millions of cycles):

Type I	3.03	5.53	5.60	9.30	9.92	12.51	12.95	15.21	16.04	16.84
Type II	3.19	4.26	4.47	4.53	4.67	4.69	12.78	6.79	9.37	12.75

a) Use normal theory to test the hypothesis that there is no difference between the two types of bearings. Here, we would use a two-sample t-test. Letting X refer to type I and Y to type II,

$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_X^2}{n}+\frac{s_Y^2}{m}}} = \frac{10.69-6.75}{\sqrt{\frac{23.23}{10}+\frac{12.98}{10}}} = \frac{3.94}{\sqrt{3.62}} = 2.071.$$

Comparing this with a t-distribution with $n+m-2=18$ df, we see that this is between $t_{.95}=1.734$ and $t_{.975}=2.101$, so the p-value is somewhere between 0.1 and 0.05. Using matlab, we find the exact p-value to be $2*(1-tcdf(2.071,18))=0.053$.

b) Test the same hypothesis using a nonparametric method. Well, the obvious nonparametric method for comparing two independent samples is the Wilcoxon-Mann-Whitney rank-sum test. Note: there are no ties in the data. The ranks for the observations are

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

and the sum of ranks for Type I is 130, and the sum for Type II is 80. Checking table 8, we see (using the smaller sum) that a value of 82 would correspond to $\alpha=0.1$ and a value of 78 would correspond to $\alpha=0.05$, so the p-value based on the Wilcoxon-Mann-Whitney test is between 0.1 and 0.05 as before. Using the normal approximation to the rank-sum distribution, we get

$$n_1\frac{n_1+n_2+1}{2} \pm z_{\alpha/2}\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$$

as a level α confidence interval - plugging in the numbers this gives

$$105 \pm 1.96 * sqrt175 = 105 \pm 25.92 \approx (79, 131).$$

Our results are near the bounds, but not outside them.

c) Which of the methods – that of part (a) or that of part (b) – do you think is better in this case? Well, the normal theory test relies upon the underlying assumption that the data from both samples is normally distributed. So, the latter test is better if we have reason to question these assumptions. We could test this visually by looking at Q-Q plots comparing X and Y to the normal distribution. The X sample looks pretty straight, and there is no reason to doubt normality there. The Y sample exhibits a bit of a dip below the ideal straight line, but whether it is big enough to be troublesome is a judgement call – I wouldn't be worried about it, but I'd go ahead and do both tests to see how different the results were. Doing that here shows that the results are not all that different, and both tests lead to the same conclusions.

d) Estimate π, the probability that a type I bearing will outlast a type II bearing. The easiest way to do this is simply to look at the proportion of the $n_1 * n_2$ paired comparisons of X_i with Y_j where $X_i > Y_j$. We can get this number from the sum of the ranks of the X_i fairly readily. If there were no Y_j present, then the ranks of the X_i would go from 1 to n_1, and the sum of the ranks would be $n_1(n_1+1)/2$. If we now add a single Y_j to the system, the total set of ranks now goes from 1 to n_1+1, and if the sum of the $X_i's$ goes up from $n_1(n_1+1)/2$, it goes up by the number of X_i's that exceed that particular Y_j. Continuing this process, we see that the total number of (X_i, Y_j) pairings in which X exceeds Y is given by the sum of the X ranks less $n_1(n_1+1)/2$. Here, that means that X_i exceeds Y_j

in $130 - 10 * 11/2 = 75$ of the paired comparisons; as there are $n_1 n_2 = 100$ total paired comparisons, this means that our best estimate of π is $\hat{pi} = 75/100 = 0.75$.

e) Use the bootstrap to estimate the sampling distribution of $\hat{pi}$ and its standard error. Ok, to do this, we need to estimate how the number of paired comparison outcomes would change under repeated sampling. So, we repeatedly sample. Here, though, we don't have the underlying distributions to work with, so we sample (with replacement!) from the data values that we have at hand. We come up with a new sample for X by choosing n_1 (10) values from the ones listed for X, and we come up with a new sample for Y by choosing n_2 (10) values from the ones listed for Y. We then compute the rank sums and estimated proportions as before, and we're off. As there are no ties between the X and Y values to begin with, there will also be none when we resample from these values, so any tied ranks will be wholly amongst the Xs or the Ys; if we have two tied X values at ranks 4 and 5 then it doesn't matter as far as the sum is concerned if we replace them both with 4.5 or not. Minor observation; it simplifies the coding somewat to be able to ignore ties. The code follows:

```
x = [3.03 5.53 5.60 9.30 9.92 12.51 12.95 15.21 16.04 16.84];
y = [3.19 4.26 4.47 4.53 4.67 4.69 12.78 6.79 9.37 12.75];

B = 1000;
n1 = length(x);
n2 = length(y);
thetavec = zeros(B,1);
basesum = n1*(n1+1)/2;
for(i = 1:B)

xnew = x(ceil(n1*rand(n1,1)));
ynew = y(ceil(n2*rand(n2,1)));

[a, b] = sort([xnew, ynew]);
[c, d] = sort(b); % These two lines generate the ranks for x and y

rankx = d(1:10);
thetavec(i) = sum(rankx) - basesum;
end

varest = var(thetavec)/(100*100);
```

on running this, we get an estimated variance of 0.0140 and an estimated standard deviation of 0.1184. A plot of the $\hat{pi}^*$ values looks roughly normal.

f) Use the bootstrap to find an approximate 90% confidence interval for π. Well, using the normal approximation and our answer to (g), we get

$$\hat{\pi} \pm z_{.95}\hat{se} = 0.75 \pm 1.645(0.1184) = (0.5552, 0.9448)$$

which is fairly broad. We can't make very precise statements here without more data. If we use the hybrid method, sorting the $\hat{pi}^*$ values we get

$$(2\hat{pi} - \hat{pi}^*_{(975)}, 2\hat{pi} - \hat{pi}^*_{(25)}) = (0.57, 0.95);$$

about the same thing.

11.23 Referring to Example A in Section 11.2.1, (a) if the smallest observation for method B (79.94) is made arbitrarily small, will the t-test still reject? (b) If the largest observation for method B (80.03) is made arbitrarily large, will the t-test still reject? (c) Answer the same questions for the Mann-Whitney test.

a) To answer this, we consider what happens as the smallest value is made arbitrarily small. First off, the mean decreases, which increases the separation $\bar{x} - \bar{y}$, so this would appear to make it more likely that the t-test will still reject. Up to a point this is so, but decreasing the smallest value also *increases* the variance of the results from method B. Eventually, the effect of this variance increase outstrips the effect of the mean decrease; while the means of the two samples are different, they are not different enough according to the standard deviation measure. Empirically, the value of the t-statistic is about 3.5 to begin with (there is a significant difference) but it drops to 1.65 if we replace 79.94 with 78.94 and continues to drop from there.

b) If we increase the largest observation for method B, the same logic as for part A tells us that we will not reject for arbitrarily large values, but here we have recourse to the simpler observation that increasing the largest observation increases the mean, and by increasing it enough we can make $\bar{x} - \bar{y}$ go to zero. Hence, it will not reject for arbitrarily large increases. The t-test can be sensitive to a single extreme value, or outlier.

c) When we are using the Mann-Whitney test, all of our results are based on the ranks of the data alone, not their values directly. In the case of the smallest value for Method B, that value is the smallest value in either sample, so it has rank 1. Reducing the value of the observation does not affect the rank at all; it stays 1, and hence the value of the Mann-Whitney test (and the conclusion that we draw from it) remains the same. When we increase the maximum value of the Method B observations, things do change slightly - this value initially had rank 11.5 (see p.405), and as we keep increasing things it will eventually attain rank 21. Once there, however, any further increase will have no effect. To begin with, the rank sum for the Method B results was 51, and this increases to 60.5. The critical value for a level $\alpha = 0.05$ test is 60, so our conclusion may chang if we're using this level initially, but not if we started with $\alpha = 0.1$. Bascially, the amount of influence that a single observation can have on the outcome of any rank-based test is bounded, but that does not hold for a test based on means and variances.

11.24 Let $X_1, \ldots, X_n$ be a sample from an $N(0,1)$ distribution and let $Y_1, \ldots, Y_n$ be an independent sample from an $N(1,1)$ distribution.

a) Determine the expected rank sum of the X_i's. This problem is an exercise in understanding the derivation of the expected rank sum in any given case - we revert back to a consideration of the paired comparisons. If we pick an X at random, and a Y at random, what is the chance that X will exceed Y; what is $P(X > Y)$? To compute this, we proceed as follows:

$$P(X > Y) \quad = \quad P(X - Y > 0)$$

$$= \quad P\left(\frac{(X-Y)+1}{\sqrt{2}} > \frac{1}{\sqrt{2}}\right) = P(Z > 1/\sqrt{2}) = 0.2398,$$

where we have used the fact that the difference of two normal rv's is again a normal rv. Ok, with this in mind, assume that we have n Y observations and just one X observation. What is the expected rank of the X value? Well, we would expect the X to be larger than 0.2398 of the Y observations, so the expected rank would be $0.2398n$ (the number of Y's the X value would exceed) plus 1 (the X value itself). Checking to make sure that this makes sense, we plug in $n = 1$ and note that the expected rank of X is 1.2398; this makes sense. The expected rank of Y is 1.7602 so the expected sum is always 3. What happens if we increase the number of X's to 2? Well, the expected number of Y values that the X's would exceed would be the same for both, so the expected sum would be $0.2398n + 0.2398n + 1 + 2 = 0.4796n + 3$. As we expand the number of X's until there are also n of those, the expected rank sum is thus

$$0.2398n^2 + \frac{n(n+1)}{2}.$$

In general, if we let $Z_{ij} = I(X_i > Y_j)$, then the rank sum of the X's is

$$\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij} + \sum_{i=1}^{n} i = \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij} + \frac{n(n+1)}{2}$$

and the expected rank sum is

$$E\left(\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij} + \sum_{i=1}^{n} i\right) = mnE(Z_{ij}) + \frac{n(n+1)}{2}.$$

b) Determine the variance of the rank sum of the X_i's. Having done the setup in a), this follows fairly directly.

$$V\left(\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij} + \sum_{i=1}^{n} i\right) = V\left(\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij}\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{m} Cov(Z_{ij}Z_{kl}).$$

As in class, there are four cases to consider here: If $i = k, j = l$, then $E(Z_{ij}Z_{ij}) = E(Z_{ij}) = 0.2398$, so $Cov(Z_{ij}) = 0.2398 - 0.2398^2 = 0.1823$. If $i \neq k, j \neq l$, then $E(Z_{ij}Z_{kl}) = E(Z_{ij})E(Z_{kl})$ and $Cov(Z_{ij}, Z_{kl}) = 0$. The remaining two cases are a bit more tricky. Consider first the case where $i = k, j \neq l$. In this case, $Z_{ij}Z_{il} = 1$ if and only if $X_i > Y_j$ and $X_i > Y_l$. The difficulty lies with the fact that the outcomes of these two paired comparisons are not independent. We can see this by redefining the problem in terms of $U_{ij} = X_i - Y_j$ and $U_{il} = X_i - Y_l$. Both U_{ij} and U_{il} have normal distributions with mean -1 and variance 2, and their joint distribution is bivariate normal, but the two are not uncorrelated! Since the U's share a contributing X term which contributes half of the variability of the U, the correlation between the two is $\rho = 0.5$. The chance that the X will be the smallest value of the three is thus

$$\int_0^\infty \int_0^\infty \frac{1}{2\pi\sqrt{2^2}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(u_1+1)^2}{2} - 2\rho\frac{(u_1+1)(u_2+1)}{2} + \frac{(u_2+1)^2}{2}\right)\right] du_1 du_2.$$

This integral is unfortunately difficult to solve in closed form, so we resort to simulation:

```
x = randn(10000,3);
x(:,2:3) = x(:,2:3) + 1;
y = (x(:,1) > x(:,2)) & (x(:,1) > x(:,2));
propest = sum(y)/10000;
```

running this gives us a value for $E(Z_{ij}Z_{il})$ of 0.1139. The associated covariance is $0.1139 - 0.2398^2 = 0.0563$. Note that this answer is not perfect as we are estimating a proportion of successes; the variance of a proportion is approximately $\hat{p}(1-\hat{p})/n = 0.1139 * (1 - .1139)/10000 = 1.0093e-05$ and the associated standard deviation is 0.0032 so a 95% confidence interval for the true value would be $0.1139 \pm 1.96 * 0.0032 = (0.1077, 0.1201)$. Still, this gives us a way of attacking the problem. Numerical integration would also work.

A similar problem arises when we consider the case $i \neq k, j = l$; In this case we are asking for the chance that $X_i > Y_j$ and $X_k > Y_j$. Jumping straight to simulating,

```
x = randn(10000,3);
x(:,3) = x(:,3) + 1;
y = (x(:,1) > x(:,3)) & (x(:,2) > x(:,3));
propest = sum(y)/10000;
```

this is actually the same as the integral that we just approximated above! In the first case, we were asking for the chance that one value from a small distribution would exceed two values from a large distribution, and here we are asking for the chance that two values from a small distribution exceed one value from a large distribution. As the shapes of the distributions are the same, so are the chances. Actually, at this point I got a bit frustrated at not knowing things too well, so I generated one million samples of size 3 to get the standard deviation down by a factor of 10 from the results using just 10000 samples. The point estimate is now 0.1133 and the standard error is $[.1133(1 - .1133)/1000000]^0.5 = 0.00032$, so we know the first 3 significant figures. The Covariance is thus $0.1133 - 0.2398^2 = 0.0558$.

At the end of the day, then, the variance of the rank sum of the X_i's is

$$0.1823 * n^2 + 0.0558 * n^2(n-1) + 0.0558 * n^2(n-1) = n^2(0.1823 + (2 * 0.0558)(n-1)).$$

11.34 Lin, Sutton, and Qurashi (1979) compared microbiological and hydroxlamine methods for the analysis of ampicillin dosages. In one series of experiments, pairs of tablest were analyzed by the two methods. The data in the following table give the percentages of claimed amount of ampicillin found by the two methods in several pairs of tablets. What are $\bar{X} - \bar{Y}$ and $s_{\bar{X}-\bar{Y}}$? If the pairing had been erroneously ignored and it had been assumed that the two samples were independent, what would have been the estimate of the standard deviation of $\bar{X} - \bar{Y}$? Analyze the data to see if there is a systematic difference between the two methods.

Microbiological Method	Hydroxylamine Method	Difference
97.2	97.2	0.0
105.8	97.8	8.0
99.5	96.2	3.3
100.0	101.8	-1.8
93.8	88.0	5.8
79.2	74.0	5.2
72.0	75.0	-3.0
72.0	67.5	4.5
69.5	65.8	3.7
20.5	21.2	-0.7
95.2	94.8	0.4
90.8	95.8	-5.0
96.2	98.0	-1.8
96.2	99.0	-2.8
91.0	100.2	-9.2
$\bar{x} = 85.26$	$\bar{y} = 84.82$	$d = 0.44$
$s_X^2 = 449.28$	$s_Y^2 = 464.57$	$s_D^2 = 21.44$

In this example, the standard deviation of the difference in means, using the pairings, is

$$s_{\bar{X}-\bar{Y}} = \sqrt{s_D^2/n} = \sqrt{21.44/15} = 1.1955.$$

Using the data to test for a difference, our test statistic would be

$$\frac{\bar{d}}{\sqrt{s_D^2/n}} = \frac{0.44}{1.1955} = 0.3680;$$

we can check this against a t-distribution with 14 degrees of freedom, but there is no indication of a significant difference between the two methods.

If we ignored the fact that the data were paired, then our estimate of the standard deviation of the difference in means would be

$$s_{\bar{X}-\bar{Y}} = \sqrt{\frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2}} * \sqrt{\frac{1}{n_X}+\frac{1}{n_Y}} = \sqrt{\frac{449.28+464.57}{2}} * \sqrt{\frac{2}{15}} = 7.8053.$$

Clearly, the standard deviation is much higher when we ignore the pairing! Pairing let us focus in on the effect of interest much more precisely. Using the data to test for a difference, our test statistic ignoring pairing would be

$$\frac{\bar{x}\,\bar{y}}{s_p\sqrt{\frac{1}{n}+\frac{1}{n}}} = \frac{0.44}{7.8053} = 0.0564;$$

we can check this against a t-distribution with 28 degrees of freedom, but there is no indication of a significant difference between the two methods based upon this test. Of course, the variance in the sizes of the tablets used means that the difference would have had to be quite large for us to detect it using a two-sample t-test!

11.37 An experiment was done to test a method for reducing faults on telephone lines (Welch 1987). Fourteen matched pairs of areas were used. The following table shows the fault rates for the control areas and for the test areas:

Test	Control	Difference	Signed Rank
676	88	588	11
206	570	-364	-6
230	605	-375	-8
256	617	-361	-5
280	653	-373	-7
433	2913	-2480	-14
337	924	-587	-10
466	286	180	4
497	1098	-601	-12
512	982	-470	-9
794	2346	-1552	-13
428	321	107	2
452	615	-163	-3
512	519	-7	-1
$\bar{x} = 434.21$	$\bar{y} = 895.5$	$\bar{d} = -461.29$	$W_+ = 17$
$s_X^2 = 2.8007e+04$	$s_Y^2 = 6.2495e+05$	$s_D^2 = 5.7427e+05$	

a) Plot the differences versus the control rate and summarize what you see. When the differences are plotted versus the control rate, the data exhibit an almost linear pattern, with a slope near -1. What's going on is that the test method has not only reduced the number of errors on average, it has greatly increased stability by reducing the variance. The control observations are so widely variable that this variance gets carried over into the differences. If we plot Test, Control and Difference all versus index on the same plot, it is apparent that while the Test regions may have error occurrences that follow a normal distribution, the results in the control regions are quite variable and skewed; there is a long upper tail. Given results like this, a normal theory test (such as the paired t-test) will have problems; a nonparametric test such as the Wilcoxon signed-rank test should fare much better.

b) Calculate the mean difference, its standard deviation, and a confidence interval. Looking at the table above, the mean difference is $\bar{d} = -461.29$, its standard deviation is $\sqrt{s_D^2/n} = \sqrt{5.7427e+05/14} = 202.53$, and a 95% confidence interval for the true difference is

$$\bar{d} \pm t_{13,0.975}\sqrt{s_D^2/n} = -461.29 \pm 2.16 * 202.53 = (-898.83, -23.75).$$

Here, the difference was sufficiently dramatic that it was picked up even though there were large outliers. The t statistic is $-461.29/202.53 = -2.2776$, for which the one-sided p-value is 0.0201 (Remember that we are looking for an *improvement* so the one-sided test is the appropriate one).

c) Calculate the median difference and a confidence interval and compare to the previous result. If we sort the differences, the median is taken to be the average of the two middle differences as the number of observations is even - here, the sorted differences are

$$-2480, -1552, -601, -587, -470, -375, -373, -364, -361, -163, -7, 107, 180, 588$$

the two middle values are -373 and -364, and the median is -368.5. To form a confidence interval for the median, we use the sorted values; our interval will be of the form $(X_{(k)}, X_{(n-k+1)})$ (see Ch.10.4.2). The bounds are chosen based on a binomial distribution - each time we take a measurement, the chance that we will be less than the median is $1/2$, by definition. Hence, the number of observations falling below the median value is a binomial random variable with parameters $(n, p = 0.5)$. The lower limit for the confidence interval is thus the sorted value $X_{(k)}$ such that

$$\frac{1}{2^n} \sum_{j=0}^{k-1} \binom{n}{j} < \alpha/2.$$

In our case, with $n = 14$, choosing $k = 4$ yields $\alpha = 0.0574$, which is pretty close to 0.05. So, an approximate confidence interval is $(X_{(4)}, X_{(11)}) = (-587, -7)$. The lower limit is a good deal closer to zero here than it was in the case of the t-test; the skewness of the data affects this test a bit less.

d) Do you think it is more appropriate to use a t test or a nonparametric method to test whether the apparent difference between test and control could be due to chance? Why? Carry out both tests and compare. Well, as discussed above, the skewness in the data, both in the Control values and subsequently in the Difference values (seen in a) strongly suggest that we should use a nonparametric method rather than a t test. We performed a t-test above (implicitly, by computing the confidence interval in b; we also found the test stat and p-value there though it was not asked for there). So, what remains is to perform the corresponding nonparametric test; the Wilcoxon signed-rank test. In this case, we have listed the signed ranks in the table above. There are many more negative terms than positive ones, so we elect to work with the sum of the positive ranks, which in this case is $W_+ = 17$. Checking table 9 for $n = 14$, we see that a value of 21 corresponds to an α of 0.025 for a one-sided test, and a value of 16 corresponds to an α of 0.01. Consequently, the one-sided p-value is between these two values, and a good deal closer to the latter. There does appear to be a significant improvement after the new method is implemented.