(d) $B_s = 9$. Looking under $n_D = 9$ in Table 7, the P-value is 0.004.

8.4.2 (a) $P > 0.20$

(b) $P = 0.118$

(c) $P = 0.035$

(d) $P = 0.007$

(e) $P = 0.001$

(f) $P < 0.001$

8.4.3 (a) H_0: oral conjugated estrogen has no effect on PAI-1 level. H_A: oral conjugated estrogen does affect PAI-1 level.

(b) We reject H_0 because the P-value is smaller than 0.10. We have sufficient evidence to say that oral conjugated estrogen has an effect on PAI-1 level.

• **8.4.4** For the sign test, the hypotheses can be stated as

H_0: $p = 0.5$
H_A: $p > 0.5$

where p denotes the probability that the rat in the enriched environment will have the larger cortex. The hypotheses may be stated informally as

H_0: Weight of the cerebral cortex is not affected by environment
H_A: Environmental enrichment increases cortex weight

There were 12 pairs. Of these, there were 10 pairs in which the relative cortex weight was greater for the "enriched" rat than for his "impoverished" littermate; thus $N_+ = 10$ and $N_- = 2$. To check the directionality of the data, we note that

$N_+ > N_-$

Thus, the data so deviate from H_0 in the direction specified by H_A. The value of the test statistic is

B_s = larger of N_+ and N_-
$= 10$.

With $n_D = 12$ and $B_s = 10$ the nondirectional P-value is 0.039. We divide this in half to get $P = 0.0195$, which is less than $\alpha = 0.05$; thus, we reject H_0. There is sufficient evidence ($P = 0.0195$) to conclude that environmental enrichment increases cortex weight.

8.4.5 Let p denote the probability that a patient will have fewer minor seizures with valproate then with placebo.

H_0: Valproate is not effective against minor seizures ($p = 0.5$)
H_A: Valproate is effective against minor seizures ($p > 0.5$)

$N_+ = 14$, $N_- = 5$, $B_s = 14$; the data deviate from H_0 in the direction specified by H_A. Eliminating the pair with d = 0, we refer to Table 7 with $n_D = 19$. The entry of 14 shows a P-value of 0.064 (for a nondirectional alternative). We divide this in half to get $P = 0.032$, which is less than $\alpha = 0.05$; thus, we reject H_0. There is sufficient evidence ($P = 0.032$) to conclude that valproate is effective against minor seizures.

8.4.6 (a) $N_+ = 8$ and $N_- = 0$ so $B_S = 8$.

(b) We reject H_0 because the P-value is smaller than 0.01. We have sufficient evidence (P=0.008) to conclude that the Carolina subspecies tends to dominate the Northern.

8.4.7 (a) For $n_D = B_s = 4$, the four sample differences would be all positive or all negative. If H_0 is true ($p = 0.5$), obtaining a sample with all differences of the same sign would occur with probability $(2)(0.5^4) = 0.1250$. So, P-value = 0.1250.

(b) The most extreme possible result is to have all positive differences or all negative differences. The probability of this happening, if p=0.5, is $(2)(0.5^3) = 0.250$, so it is not possible to get a P-value as small as 0.20 for a non-directional test (or 0.10 for a directional test).

• **8.4.8** P-value = $(2)(0.5^{15}) = 0.000061$.

8.4.9 Let p denote the probability that hunger rating is higher when taking mCPP than when taking the placebo.

H_0: $p = 0.5$
H_A: $p \neq 0.5$

$N_+ = 3$, $N_- = 5$, $B_s = 5$. Looking under $n_D = 8$ in Table 7, we see that the leftmost column has an entry of 7, with a P-value of 0.070. This is under the 0.20 heading, so the next smallest possible B_s value (of 6) must have a P-value larger than 0.20. There is insufficient evidence ($P > 0.20$) to conclude that hunger ratings differ on the two treatments.

8.4.10 $P = (2)[(56)(0.5^5)(0.5^3) + (28)(0.5^6)(0.5^2) + (8)(0.5^7)(0.5^1) + (1)(0.5^8)] = 0.7265$.

• **8.4.11** If it is expected that one treatment "wins" in every pair, then B_s will equal n_D in this study. For this case, the P-value is computed as $(2)(0.5)^{n_D}$. In order for the sample to be large enough to reject H_0 at $\alpha = 0.05$, n_D must satisfy the equation $(2)(0.5)^{n_D} < 0.05$. The smallest value of n_D that satisfies this equation is $n_D = 6$. The P-value will be 0.03125.

• **8.5.1 (a)** $P > 0.20$

(b) P = 0.078

(c) P = 0.047

(d) P = 0.016

8.5.2 (a) P > 0.20

(b) P = 0.042

(c) P = 0.0093

(d) 0.0015 < P < 0.0093

• **8.5.3** H_0: Hunger rating is not affected by treatment (mCPP vs. placebo)
H_A: Treatment does affect hunger rating

The absolute values of the differences are 5, 7, 28, 47, 80, 7, 8, and 20.
The ranks of the absolute differences are 1, 2.5, 6, 7, 8, 2.5, 4, and 5.
The signed ranks are –1, 2.5, -6, -7, -8, 2.5, 4, and –5.
Thus, $W_+ = 2.5 + 2.5 + 4 = 9$ and $W_- = 1 + 6 + 7 + 8 + 5 = 27$.
$W_s = 27$ and $n_D = 8$; reading Table 8 we find P-value > 0.20 and H_0 is not rejected. There is insufficient evidence (P > 0.20) to conclude that treatment has an effect.

8.5.4 (a) H_0: Weight change is not affect by treatment (mCPP vs. placebo)

(b) The large P-value means that if the two groups really were the same then it would not be surprising to see the kinds of differences that arose in this experiment.

(c) We retain H_0. There is insufficient evidence (P=0.43) to conclude that treatment has an effect on weight change.

8.5.5 H_0: HL-A compatibility has no effect on graft survival time
H_A: Survival time tends to be greater when compatibility score is close

The differences tend to be positive, which is consistent with H_A.
The absolute values of the differences are 12, 6, 42+, 67, 5, 5, 6, 20, 11, 18+, and 1.
The ranks of the absolute differences are 7, 4.5, 10, 11, 2.5, 2.5, 4.5, 9, 6, 8, and 1.
The signed ranks are 7, 4.5, 10, 11, 2.5, 2.5, -4.5, 9, 6, 8, and -1.
Thus, $W_+ = 7 + 4.5 + 10 + 11 + 2.5 + 2.5 + 9 + 6 + 8 = 60.5$ and $W_- = 4.5 + 1 = 5.5$.
$W_s = 60.5$ and $n_D = 11$; reading Table 8 we find 0.0098/2 < P-value < 0.019/2 and H_0 is rejected. There is strong evidence (0.0049 < P-value < 0.0095) to conclude that survival time tends to be greater when compatibility score is close.

8.5.6 (a) $W_s = 61$

(b) We reject H_0 because the P-value is smaller than 0.01. We have strong evidence that average increase in plasma volume is greater for albumin than for polygelatin.

8.5.7 H_0: Caffeine has no effect on myocardial blood flow
H_A: Caffeine affects myocardial blood flow

The absolute values of the differences are 0.71, 0.14, 1.31, 0.49, 0.49, 0.04, 0.04, 0.02, 0.45, and 0.12.

The ranks of the absolute differences are 9, 5, 10, 7.5, 7.5, 2.5, 2.5, 1, 6, and 4.
The signed ranks are 9, 5, 10, 7.5, 7.5, -2.5, -2.5, -1, 6, and 4.
Thus, $W_+ = 9 + 5 + 10 + 7.5 + 7.5 + 6 + 4 = 49$ and $W_- = 2.5 + 2.5 + 1 = 6$.

$W_s = 49$ and $n_D = 10$; reading Table 8 we find that for Ws = 47, P-value = 0.049 and for Ws = 50, P-value = 0.020. Thus, 0.020 < P-value < 0.049 (a computer yields a P-value = 0.0321). There is significant evidence (0.020 < P-value < 0.049) to conclude that caffeine affects myocardial blood flow. This was an experiment, so drawing a cause-effect inference is appropriate.

8.6.1 Let 1 denote 5 weeks and 2 denote baseline.

(a) Let N denote no coffee.

H_0: Mean cholesterol does not change in the "no coffee" condition ($\mu_{N,1} = \mu_{N,2}$)
H_A: Mean cholesterol does change in the "no coffee" condition ($\mu_{N,1} \neq \mu_{N,2}$)

$SE = 27 / \sqrt{25} = 5.40$.

$t_s = -35/5.40 = -6.48$. With df = 24, Table 4 gives $t_{0.0005} = 3.745$. We reject H_0. There is sufficient evidence (P < 0.001) to conclude that mean cholesterol is reduced in the "no coffee" condition.

(b) Let U denote usual coffee.

H_0: Mean cholesterol does not change in the "usual coffee" condition ($\mu_{U,1} = \mu_{U,2}$)
H_A: Mean cholesterol does change in the "usual coffee" condition ($\mu_{U,1} \neq \mu_{U,2}$)

$SE = 56 / \sqrt{8} = 19.8$.

$t_s = 26/19.8 = 1.31$. With df = 7, Table 4 gives $t_{0.20} = 0.896$ and $t_{0.10} = 1.415$. We do not reject H_0. There is insufficient evidence (0.20 < P < 0.40) to conclude that mean cholesterol is changed in the "usual coffee" condition.

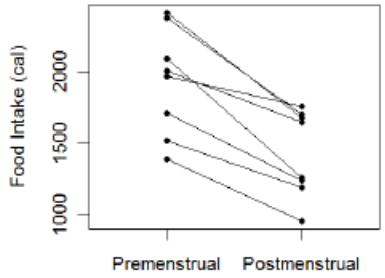
(c) Let N denote no coffee, U denote usual coffee, and d denote the change from baseline.

H_0: Mean cholesterol is not affected by discontinuing coffee ($\mu_{N,d} = \mu_{U,d}$)
H_A: Mean cholesterol is affected by discontinuing coffee ($\mu_{N,d} \neq \mu_{U,d}$)

$$SE = \sqrt{\frac{27^2}{25} + \frac{56^2}{8}} = 20.52.$$

t_s = (-35 - 26)/20.52 = -2.97. Formal (6.7.1) gives df = 8.1 and the conservative df value is min{24,7} = 7, whereas the liberal df value is $n_1 + n_2 - 2 = 31$. Using df = 8, Table 4 gives $t_{0.01} = 2.896$ and $t_{0.005} = 3.355$, which implies that $0.01 < P < 0.02$. (Using df = 30, the closest value to 31 in Table 4, we get $t_{0.005} = 2.750$ and $t_{0.0005} = 3.646$, which implies that $0.001 < P < 0.01$.) We reject H_0.

8.6.2



8.6.3 No. This result suggests that the population mean difference between the right eye and the left eye is less than 1.1 mg/dl, but positive and negative differences cancel each other out when such a mean is calculated. The result says nothing at all about the typical or average magnitude of the difference between the two eyes.

• **8.6.4** No. "Accurate" prediction would mean that the individual differences (d's) are small. To judge whether this is the case, one would need the individual values of the d's; using these, one could see whether most of the magnitudes (|d|'s) are small.

8.6.5 **(a)** $t_s = \frac{31.1-0}{17.8/\sqrt{8}} = \frac{31.1}{6.29} = 4.94$. P-value = 0.002. There is strong, statistically significant ($\alpha = 0.10$) evidence that the mean coliform level at the fenced creek was lower after fencing.

(b) We cannot conclude that fencing was the reason for the drop in coliform count. There could be many other reasons coliform counts can change over time. It will be important to compare these results to the control creek.

(c) $t_s = \frac{(31.1-15.3)-0}{\sqrt{\frac{17.8^2}{8}+\frac{16.3^2}{7}}} = \frac{15.80}{9.826} = 1.79$. P-value = 0.098. There is weak, but statistically significant ($\alpha = 0.10$) evidence that the mean reduction in coliform is greater at the fenced creek than at the control creek.

(d) The answer to part (c) is consistent with the fencing being effective, since the reduction was greater for the fenced creek

8.S.1 **(a)** $\bar{y}_1 - \bar{y}_2 = \bar{d} = -1$, $s_d = 1.2$.

$SE_{(\bar{Y}_1-\bar{Y}_2)} = 1.2/\sqrt{15} = .3098$.

$-1 \pm (2.145)(0.3098)$ (df = 14)

(-1.66,-0.34) or $-1.66 < \mu_1 - \mu_2 < -0.34$.

(b) $SE_{(\bar{Y}_1-\bar{Y}_2)} = \sqrt{\frac{1.66^2}{15}+\frac{2.37^2}{15}} = 0.7471$.

$-1 \pm (2.145)(.7471)$ (using df = 14)

(-2.60,0.60) or $-2.60 < \mu_1 - \mu_2 < 0.60$. This interval is much wider than the one constructed in part (a).

8.S.2 H_0: The before and after means are the same ($\mu_1 = \mu_2$)
H_A: The before and after means are different ($\mu_1 \neq \mu_2$)

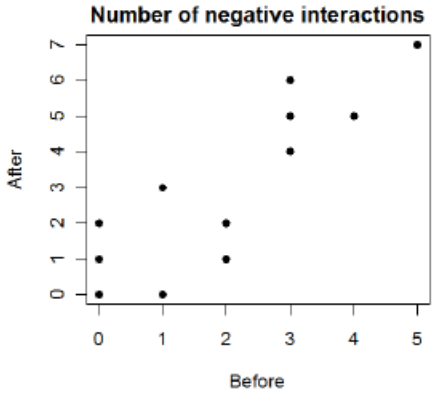
$SE_{(\bar{Y}_1-\bar{Y}_2)} = 1.2/\sqrt{15} = 0.3098$. $t_s = -1/0.3098 = -3.23$. With df = 14, Table 4 gives $t_{0.005} = 2.977$ and $t_{0.0005} = 4.140$; thus, $0.001 < P\text{-value} < 0.01$. We reject H_0: there is strong evidence ($0.001 < P < 0.01$) of a before and after difference.

8.S.3 Let p denote the probability that a before count is higher than the corresponding after count.

H_0: p = 0.5
H_A: p ≠ 0.5

$N_+ = 2$, $N_- = 10$, $B_s = 10$. Looking under $n_d = 12$ in Table 7, we see that P =0.039. There is sufficient evidence (P = 0.039) to conclude that the after count tends to be higher than the before count.

8.S.4 The scatterplot shows a positive relationship between before and after counts. The pairing removes the variability between cats from the analysis and is, therefore, effective.



8.S.5 Let 1 denote central and 2 denote top.

$\bar{y}_1 - \bar{y}_2 = \bar{d} = 2.533$, $s_d = 0.41312$.

$SE_{\bar{D}} = 0.41312 / \sqrt{6} = 0.1687$.

$2.533 \pm (2.015)(0.1687)$ (df = 5)

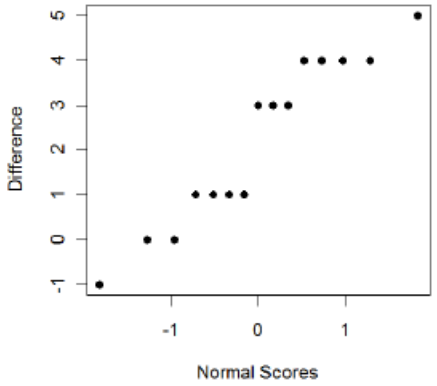
(2.19,2.87) or $2.19 < \mu_1 - \mu_2 < 2.87$ percent.

8.S.6 The standard error is $SE_{\bar{D}} = 1.86 / \sqrt{15} = 0.48$.

$2.2 \pm (1.761)(0.48)$ (df = 14)

(1.35,3.05) or $1.35 < \mu_1 - \mu_2 < 3.05$ species.

8.S.7 It must be reasonable to regard the 15 differences as a random sample from a normal population. We must trust the researchers that their sampling method was random. The normality condition can be verified with a normal probability plot of the differences. The plot below is fairly linear (although the plateaus show that there are several differences that have the same value), which supports the normality condition.



• **8.S.8** The null and alternative hypotheses are

H_0: The average number of species is the same in pools as in riffles ($\mu_1 = \mu_2$)
H_A: The average numbers of species in pools and in riffles differ ($\mu_1 \neq \mu_2$)

The standard error is

$$SE_{\bar{D}} = \frac{s_D}{\sqrt{n_D}} = \frac{1.86}{\sqrt{15}} = 0.48.$$

The test statistic is

$$t_s = \frac{\bar{d}}{SE_{\bar{D}}} = \frac{2.2}{0.48} = 4.58.$$

To bracket the P-value, we consult Table 4 with df = 15 - 1 = 14. Table 4 gives $t_{0.0005} = 4.140$. Thus, the P-value for the nondirectional test is bracketed as

$P < 0.001$.
At significance level $\alpha = 0.10$, we reject H_0 if $P < 0.10$. Since $P < 0.001$, we reject H_0. There is sufficient evidence ($P < 0.001$) to conclude that the average number of species in pools is greater than in riffles.

8.S.9 (a) Let p denote the probability that there are more species in a pool than in its adjacent riffle.

H_0: The two habitats support equal levels of diversity ($p = 0.5$)
H_A: The two habitats do not support equal levels of diversity ($p \neq 0.5$)

$N_+ = 12$, $N_- = 1$, $B_s = 12$. Eliminating the two pairs with d = 0, we refer to Table 7 with $n_D = 13$. We see that the P-value when $B_s = 12$ is 0.003. There is sufficient evidence ($P = 0.003$) to conclude that species diversity is greater in pools than in riffles.

(b) $P = (2)[(13)(0.5^{12})(0.5^{1}) + 0.5^{13}] = 0.0034$.

8.S.10 H_0: Pools and riffles support equal levels of diversity
H_A: Pools and riffles support different levels of diversity

The absolute values of the differences are 3, 3, 4, 3, 4, 5, -1, 1, 1, 4, 1, 4, and 1.
The ranks of the absolute differences are 7, 7, 10.5, 7, 10.5, 13, 3, 3, 3, 10.5, 3, 10.5 and 3.
The signed ranks are 7, 7, 10.5, 7, 10.5, 13, -3, 3, 3, 10.5, 3, 10.5 and 3.
Thus, $W_+ = 7 + 7 + 10.5 + 7 + 10.5 + 13 + 3 + 3 + 3 + 10.5 + 3 + 10.5 + 3 = 88$ and $W_- = 3$.

$W_s = 88$ and $n_D = 13$; reading Table 8 we find $0.0007 < \text{P-value} < 0.0017$ and H_0 is rejected. There is strong evidence ($0.0007 < \text{P-value} < 0.0017$) to conclude that the diversity levels differ between pools and riffles.

8.S.11 There are several ties in the data, which means that the P-value from the Wilcoxon test is only approximate.

• **8.S.12** The null and alternative hypotheses are

H_0: Caffeine has no effect on RER ($\mu_1 = \mu_2$)
H_A: Caffeine has some effect on RER ($\mu_1 \neq \mu_2$)

We proceed to calculate the differences, the standard error of the mean difference, and the test statistic.

Subject	Placebo	Caffeine	Difference
1	105	96	9
2	119	99	20
3	92	89	3
4	97	95	2
5	96	88	8
6	101	95	6
7	94	88	6
8	95	93	2
9	98	88	10
Mean			7.33
SD			5.59

The standard error is

$$SE_{\bar{D}} = \frac{s_D}{\sqrt{n_D}} = \frac{5.59}{\sqrt{9}} = 1.86.$$

The test statistic is

$$t_s = \frac{\bar{d}}{SE_{\bar{D}}} = \frac{7.33}{1.86} = 3.94.$$

To bracket the P-value, we consult Table 4 with df = 9 - 1 = 8. Table 4 gives $t_{0.005} = 3.355$ and $t_{0.0005} = 5.041$. Thus, the P-value for the nondirectional test is bracketed as

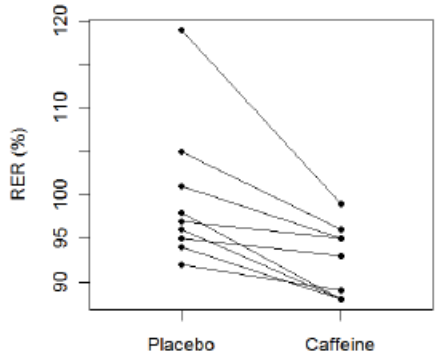
$0.001 < P < 0.01$.

At significance level $\alpha = .05$, we reject H_0 if $P < 0.05$. Since $P < 0.01$, we reject H_0. To determine the directionality of departure from H_0, we note that

$\bar{d} > 0$; that is, $\bar{y}_1 > \bar{y}_2$.

There is sufficient evidence ($0.001 < P < 0.01$) to conclude that caffeine tends to decrease RER under these conditions.

8.S.13



8.S.14 Let p denote the probability that RER for a subject is higher after taking placebo than after taking caffeine.

H_0: RER is not affected by caffeine ($p = 0.5$)
H_A: RER is affected by caffeine ($p \neq 0.5$)

$N_+ = 9$, $N_- = 0$, $B_s = 9$. Looking under $n_D = 9$ in Table 7, we see that P-value when $B_s = 9$ is 0.004. There is sufficient evidence (P = .04) to conclude that caffeine tends to decrease RER under these conditions.

8.S.15 (a) $t_s = \dfrac{4.64}{4.89/\sqrt{8}} = 2.69$

(b) H_0: Mean CP is the same in regenerating and in normal tissue ($\mu_1 = \mu_2$). H_A: Mean CP is different in regenerating and in normal tissue ($\mu_1 \neq \mu_2$).

(c) We reject H_0 because the P-value is smaller than 0.05. We have sufficient evidence (P=0.031) to conclude that mean CP is different in regenerating and in normal tissue.

(d) (0.55, 8.73) mg/100gm

(e) We are 90% confident that cutting the nerves reduces the mean CP by between 1.36 and 7.92 mg/100gm.

8.S.16 (a) $N_+ = 6$ and $N_- = 0$ so $B_S = 6$.

(b) We reject H_0 because the P-value is smaller than 0.10. We have sufficient evidence (P=0.031) to conclude that aldosterone differs (is higher) after Captopril treatment.

8.S.17 H_0: Blood aldosterone is not different after Captopril treatment
H_A: Blood aldosterone is different after Captopril treatment

The absolute values of the differences are 375, 169, 197, 180, 217, and 203
The ranks of the absolute differences are 6, 1, 3, 2, 5, and 4.
The signed ranks are 6, 1, 3, 2, 5, and 4.
Thus, $W_+ = 6 + 1 + 3 + 2 + 5 + 4 = 21$ and $W_- = 0$.

$W_s = 21$ and $n_D = 6$; reading Table 8 we find P-value = 0.031 and H_0 is rejected. There is moderate evidence (P-value = 0.031) to conclude aldosterone is different (higher)after Captopril treatment.

8.S.18 Without a control or placebo group it is impossible to determine if the observed effect is due to Captopril or for another reason. For example, aldosterone levels may vary regularly throughout the treatment period. A "before" measurement taken in the morning could systematically be higher (or lower) than an "after" measurement in the evening. A placebo/nocebo effect is also possible if the stress of receiving the treatment could alter aldosternone levels.

8.S.19 (a) Let 1 denote control and 2 denote benzamil.

H_0: Benzamil does not impair healing ($\mu_1 = \mu_2$)
H_A: Benzamil impairs healing ($\mu_1 > \mu_2$)

$\bar{y}_1 - \bar{y}_2 = \bar{d} = .09706$; $s_d = 0.14768$.

$SE_{\bar{D}} = 0.14768 / \sqrt{17} = 0.03582$.

$t_s = 0.09706/0.03582 = 2.71$. P = 0.0077, so we reject H_0. There is sufficient evidence (P = 0.0077) to conclude that benzamil impairs healing.

(b) Let p denote the probability that the control limb heals more than the benzamil limb.

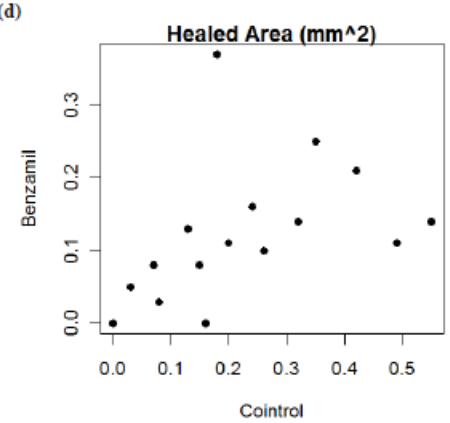
H_0: Benzamil does not impair healing (p = 0.5)
H_A: Benzamil impairs healing (p > 0.5)

$N_+ = 11$, $N_- = 4$, $B_s = 11$. The two animals with d = 0 are eliminated. Using $n_D = 15$ we see tht the nondirectional P-value when $B_s = 11$ is 0.118, so the directional P-value is 0.059 and we do not reject H_0. There is insufficient evidence (P = 0.059) to conclude that benzamil impairs healing.

[Remark: Unlike the t test in part (a), the sign test does not take account of the fact that the negative d's are smaller in magnitude than the positive d's. This illustrates the inferior power of the sign test.]

(c) (0.021,0.173) or $0.021 < \mu_1 - \mu_2 < 0.173$ mm^2.

(d)



Yes, the upward trend indicates that the pairing was effective.

8.S.20 Summary statistics are as follows:

	Experimental group			Control group		
	Rest (1)	Work (2)	Difference (3)	Rest (4)	Work (5)	Difference (6)
Mean	6.169	9.301	-3.133	5.291	4.973	0.319
SD	0.621	3.323	2.862	0.652	0.703	0.544

(a) Column (1) versus column (4)

H_0: Mean ventilation at rest is the same in the two conditions ($\mu_1 = \mu_4$)
H_A: Mean ventilation at rest is different in the two conditions ($\mu_1 \neq \mu_4$)

$t_s = 2.757$, df = 13.97, P = 0.015. We reject H_0. There is sufficient evidence (P = 0.015) to conclude that mean ventilation at rest is higher in the "to be hypnotized" condition than in the "control" condition.

(b) (i) Column (1) versus column (2):

H_0: Hypnotic suggestion does not change mean ventilation ($\mu_1 = \mu_2$)
H_A: Hypnotic suggestion increases mean ventilation ($\mu_1 < \mu_2$)

$t_s = -3.096$, df = 7, P = 0.0087. We reject H_0. There is sufficient evidence (P = 0.0087) to conclude that hypnotic suggestion increases mean ventilation.

(ii) Column (4) versus column (5):

H_0: Waking suggestion does not change mean ventilation ($\mu_4 = \mu_5$)
H_A: Waking suggestion increases mean ventilation ($\mu_4 < \mu_5$)

Because $\bar{y}_4 > \bar{y}_5$, the data do not deviate from H_0 in the direction specified by H_A. Thus, P > 0.50 and we do not reject H_0. There is no evidence that waking suggestion increases mean ventilation.

(iii) Column (3) versus column (6):

H_0: Hypnotic and waking suggestion produce the same mean change in ventilation ($\mu_3 = \mu_6$)
H_A: Hypnotic suggestion increases mean ventilation more than does waking suggestion ($\mu_3 < \mu_6$)

$t_s = -3.351$, df = 7.5, P = 0.0055. We reject H_0. There is sufficient evidence (P = 0.0055) to conclude that hypnotic suggestion increases mean ventilation more than does waking suggestion.

(c) (i) Sign test for column (1) versus column (2). Let p_1 denote the probability that a person's ventilation after hypnotic suggestion will be higher than that at rest.

H_0: Hypnotic suggestion does not change mean ventilation ($p_1 = 0.5$)
H_A: Hypnotic suggestion increases mean ventilation ($p_1 > 0.5$)

$B_s = 8$, P = 0.008/2 = 0.004. We reject H_0. There is sufficient evidence (P = 0.004) to conclude that hypnotic suggestion increases mean ventilation.

(ii) Sign test for column (4) versus column (5). Let p_2 denote the probability that a person's ventilation after waking suggestion will be higher than that at rest.

H_0: Waking suggestion does not change mean ventilation ($p_2 = 0.5$)
H_A: Waking suggestion increases mean ventilation ($p_2 > 0.5$)

$N_+ = 6$, $N_- = 2$. Thus, the data do not deviate from H_0 in the direction specified by H_A, so P > 0.50 and we do not reject H_0. There is no evidence that waking suggestion increases mean ventilation.

(iii) Wilcoxon-Mann-Whitney test for column (3) versus column (6):

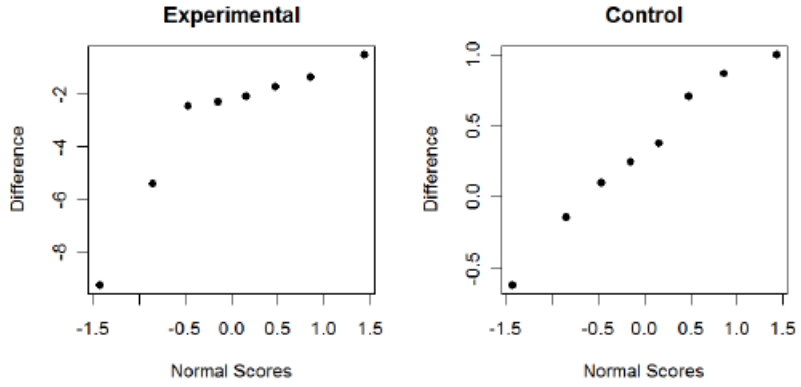
H_0: Hypnotic and waking suggestion produce the same mean change in ventilation

H_A: Hypnotic suggestion increases mean ventilation more than does waking suggestion

$U_s = 63$, P = 0.000155. We reject H_0. There is sufficient evidence (P = 0.000155) to conclude that hypnotic suggestion increases mean ventilation more than does waking suggestion.

(d) The following normal probability plot of rest – work differences summarized in column (3) shows that the data are quite skewed for the experimental group. This could account for two discrepancies: First, to compare column (1) to column (2), we used the differences in column (3); the t test gave P = 0.0087 whereas the sign test gave P = 0.004. Second, to compare column (3) to column (6), the t test gave P = 0.0055 whereas the Wilcoxon-Mann-Whitney test gave P = 0.000155. Both of the t tests rest on the questionable condition that the population distribution corresponding to column (3) is normal. The failure of this condition inflates the standard deviation and robs the t test of power, so that the nonparametric tests give stronger conclusions (smaller P-values).

A normal probability plot of the rest – work differences summarized in column (6) shows no evidence that the normality condition is violated for these data.



8.S.21 (a) By using matched pairs we eliminate the variability that is associated with the variables used to create the pairs (age, sex, etc.). This provides for greater precision and more power in the test.

(b) It may be that the pairing variables (age, sex, etc.) are unrelated to blood pressure. If this is the case, then the pairing accomplishes nothing, but it reduces the number of degrees of freedom, and therefore the power, of the test.

8.S.22 $N_+ = 10$, $N_- = 10$, $B_s = 10$. In this case, the data are as evenly balanced as possible, so P = 1. (Table 7 indicates that P > 0.20.) Thus, we do not reject H_0. There is no evidence that transdermal estradiol has an effect on PAI-1 level.

8.S.23 A normal probability plot of the data shows that the normality condition is not met. However, a sign test can be conducted. Let p denote the probability that urinary protein excretion will go down after plasmapheresis.

H_0: Plasmapheresis affects urinary protein excretion ($p = 0.5$)
H_A: Plasmapheresis does not affect urinary protein excretion ($p \neq 0.5$)

$N_+ = 6$, $N_- = 0$, $B_s = 6$. From Table 7, P = 0.031 (for a two-sided test). (Note: The P-value is $(2)(0.5^6) = 0.03125$.) Thus, there is evidence (P = 0.031) to conclude that urinary protein excretion tends to go down after plasmapheresis.

Note: Another approach would be to transform the data and then conduct a t test in the transformed scale. For example, taking the reciprocal of each difference yields a fairly symmetric distribution; a t test then gives $t_s = 5.4$ and P = 0.003.

UNIT II SUMMARY

II.1 The worst advice comes from Gloria. It is *not valid* to choose the direction for a one-sided alternative hypothesis by looking at the data. The direction of a one-sided hypothesis needs to be based on information available *before* the data were collected. (If you use the data to determine the direction of H_A and then do an $\alpha = 0.05$ test, you have a 0.10 chance of rejecting H_0 when H_0 is true.)

Rudd gives better advice. His procedure is valid, but he only gives the result of the test (reject H_0 or do not reject H_0 when $\alpha = 0.05$), not the strength of the evidence nor the size of the effect.

Steve's advice is even better. He tells us what the *P*-value is, so we know not only whether it is less than 0.05, but also how large or small it is. That is, Steve gives us everything that Rudd gives us and more. Nancy can choose to use $\alpha = 0.05$ if she likes, or she can choose a different level for α. Unlike Rudd, Steve would not restrict the inferences Nancy could make to the special case of $\alpha = 0.05$.

Linda has the best advice. From her we learn something about the size of the difference, if any, between μ_1 and μ_2. Thus, we are able to judge the magnitude of the effect. Moreover, we can examine her confidence interval to see whether or not it includes zero. If the confidence interval does not include zero, than we know that the difference in the means is statistically significant at the $\alpha = 0.05$ level. That is, Linda gives us everything that Rudd gives us and more.

Thus, from worst to best we have Gloria (worst), then Rudd, then Steve, and finally Linda (best).

One could argue that learning the strength of the difference (by finding the *P*-value) is more important than learning the size of the effect (by finding a confidence interval). So one could defend the order Gloria (worst), then Rudd, then Linda, and finally Steve (best).

• **II.2 (a)** The SE for $\bar{Y}_1 - \bar{Y}_2$ is $\sqrt{\frac{3.89^2}{27} + \frac{1.06^2}{14}} = 0.80$. The test statistics is $t_s = \frac{6.59 - 3.96}{0.80} = 3.29$.

Using 32 degrees of freedom, we have $0.001 < P\text{-value} < 0.01$. Since the *P*-value is less than the α level of 0.02, we reject H_0.

(b) There is strong evidence that the average ecological footprint for the women differs from that for the men. (Indeed, the data suggest that the average for women is greater than the average for men.) The difference in sample averages cannot easily be explained by chance. (*Note:* The *sample* means are statistically significantly different, so we infer that the population means are *different* [not "significantly different"].)

(c) Using 32 degrees of freedom and calculations from part (a) above, a 95% confidence interval is

$$6.59 - 3.96 \pm t_{0.025} \times 0.80$$

$$6.59 - 3.96 \pm 2.042 \times 0.80$$

$$2.63 \pm 1.634$$

(1.00,4.46) hectares

Since the interval is entirely above 0.7 hectares, the difference is "ecologically important."

II.3 **(a)** For the women, $\bar{y} - 2*(\text{SD})$ is less than zero, but an ecological footprint cannot be negative, so the distribution cannot be normal and must be skewed to the right, which concerns Norman since a *t* test depends on normality.

(b) Rebecca could use a Wilcoxon-Mann-Whitney test, or a randomization test, which does not require normality.

(c) The sample size of women is $n = 27$ and the *t* test is robust against departures from normality, particularly with moderate to large samples. Thus, the sampling distribution of the sample mean for women is probably normal despite the underlying distribution being skewed. (For men, there is less evidence of skewness, so the smaller sample size of 14 is probably adequate, making things okay for the men as well.)

• **II.4** The confidence interval excludes zero, so we would reject H_0; this means that the *P*-value is less than 0.05. Thus the *P*-value is less than 0.10, so we would reject H_0.

II.5 **(a)** False. The confidence interval describes the difference in population *means*, not differences for individual infants in the population.

(b) False. Although the confidence interval does cross zero, we cannot rule out nonzero values for the difference in population means. Thus, the results are inconclusive. The difference in population mean hospitalization length under the two conditions (nitric oxide vs. control) may be zero, or it may be as large as 16.1 in magnitude!

(c) True. The largest we estimate the difference to be, with 95% confidence, is 16.1 days.

II.6 **(a)** False. The *P*-value is the probability of data at least this extreme *if* H_0 is true. (Either the distributions are the same, or they are not. We don't talk in terms of the probability that they are the same.)

(b) True. This is what a *P*-value is. (See part (a).)

(c) False. This probability depends on (1) the choice of a and (2) whether or not H_0 is true—which is not known.

(d) True. The *P*-value would be cut in half to 0.03, so it would be less than 0.05 and we would reject H_0.

II.7 **(a)** $t_s = \dfrac{2.1-1.9}{\sqrt{\dfrac{0.7^2}{12}+\dfrac{0.7^2}{12}}} = \dfrac{0.2}{0.286} = 0.70$. With df = 22, we have $P > 2(0.20)$. Thus, $P > 0.40 > \alpha$, so we retain H_0. There is insufficient evidence to conclude that the mean Andro level differs between women recently in love and other women.

(b) It might be that the population difference is 0.4 or greater, but it might be less than 0.4. Thus, we cannot say whether or not the results are medically important. (See Table 7.6.3.)

• **II.8** **(a)** Power goes up as *n* goes up. This is because the SE goes down as *n* goes up. A larger sample size gives a smaller standard error; hence, more accuracy in estimating the difference in means. Thus, if there is a true difference, we are more likely to detect it when $n = 18$ than when $n = 12$.

(b) These normal curves have a common SD of 1 unit—the distance from the peak of a curve to the point of inflection; or note that ±3 units covers essentially all of a distribution. The peaks of the normal curves are separated by 1.5 units, so the effect size is = 1.5.

• **II.9** **(a)** There are three differences ≥ 31 and two differences ≤ –31. Thus *P*-value = 5/28 ≈ 0.1563.

(b) There is no statistically significant evidence that men and women differ with respect to the mean of variable *Y*. The *P*-value = $0.1536 \not< \alpha = 0.10$.

(c) This test is a directional test. Here we only consider differences ≥ 31. Thus, the directional *P*-value is 3/28 ≈ 0.107.

II.10 **(a)** (I) This is pairing of similar units. (II) True.

(b) (I) This is pairing by time. (II) True.

(c) (I) This is pairing of units after the fact, using the response variable. Here a paired *t* test is *not* valid. (II) False.

II.11 The 90% confidence interval includes zero, so a test of H_0: $\mu_1 = \mu_2$ against H_A: $\mu_1 \neq \mu_2$ has a *P*-value greater than 0.10. For testing H_0: $\mu_1 = \mu_2$ against H_A: $\mu_1 > \mu_2$ we note that the data deviate in the right direction (the bulk of the confidence interval is to the positive side of zero), so the directional *P*-value is half of the nondirectional *P*-value. $P > 0.05$, so we retain H_0.

II.12 **(a)** The sign test is less likely to reject H_0, and provide statistical evidence for H_A, than is a *t* test.

(b) A sign test can be conducted in many cases, even with censored data. Also, a sign test does not require normality.

II.13 **(a)** $\text{SE}_{\text{diff}}\ SE_{\bar{Y}_1-\bar{Y}_2} = \sqrt{1.3^2/50 + 1.4^2/40} = 0.29$. The CI is (11.4 - 5.0) ± 1.984(0.29) or (5.82,6.98).

(b) True. The mean ± 2(SD) covers about 95% of the population. Using our sample mean ± 2(sample SD) should cover about 95% of the population, assuming that decrease in blood pressure has a distribution that is reasonably symmetric and bell shaped.

II.14 **(a)** By using matched pairs, we control for variability that is due to age and sex, which means that precision is improved and we have more power when conducting a test.

(b) By pairing we sacrifice degrees of freedom (in effect, we reduce the sample size) and thus we lose power unless age and sex contribute a great deal to variability in blood pressure.

II.15 **(a)** H_0: Training has no effect on stress. The population mean cortisol level is the same for cats who are trained and untrained.
H_A: Training reduces stress. The population mean cortisol level is lower for trained cats than for untrained cats.

(b) This study is an experiment since the cats have been assigned to the two conditions studied (trained vs. untrained) by the researchers. Thus, a cause–effect conclusion (training *lowers* stress) can follow from this study if the data supports it.

(c) $H_0: \mu_{\text{trained}} = \mu_{\text{untrained}}$. $H_A: \mu_{\text{trained}} < \mu_{\text{untrained}}$.
$t_s = \frac{0.541 - 2.324}{\sqrt{0.215^2 + 0.239^2}} = \frac{1.783}{0.322} = -5.543$. Using df = 28, the one-tailed (lower tail) P-value is P-value < 0.0005.

There is very strong, statistically significant evidence that training cats reduces stress during blood draws, as measured by lower mean cortisol compared to a control group of untrained cats ($t_s = -5.543$, df = 28, P-value < 0.0005).

(d) A 95% confidence interval for the difference in mean cortisol levels for trained and untrained cats is

$$0.541 - 2.324 \pm 2.048 \times 0.322$$

$$1.784 \pm 0.659$$

$$(1.12, 2.44)\mu g/dL$$

The interval indicates that the smallest we estimate the difference in mean cortisol to be is 1.12 µg/dL, which is larger than 1.0 µg/dL, the threshold for "medical significance." Thus, we can regard the results as medically practical.

II.16 **(a)** No; there is evidence that the trained population is not normal.

(b) Samples will always reflect the population, so a larger sample will not produce a plot that looks "more normal" if the population is not normal to begin with.

(c) Because the sample size is small, it is especially important that the data come from a normal population for the *t* test and confidence interval to be valid. However, the data from the trained cats appears to be coming from a nonnormal population, as evidenced by right skew and small Shapiro-Wilk normality test P-value. Thus, a randomization based test, or a nonparametric test such as the Wilcoxon-Mann-Whitney test, should be used.

(d) Using the Wilcoxon-Mann-Whitney test when it is not needed results in lower power. That is, it is less able to detect H_A when H_A is true.

II.17 **(a)** False. This is a tricky question. If there is a difference to detect (if H_A is true), then increasing the sample size will increase the chance of detection (i.e., increase power). However, if there is no difference, the sample size will have no effect on the chance of detecting difference. In fact, if there is no difference, the chance of detecting a difference is α, the probability of a Type I error.

(b) False. The P-value is determined by the data and is then only compared to α, which is fixed at the outset of the study.

(c) False. Choosing a small α means that a small P-value is needed to declare statistical significance. Data exhibiting stronger evidence for H_A yields smaller P-values than weaker data.

(d) True. If there is an effect (i.e., a difference in population means), a larger sample will have a better chance of detecting the effect than a smaller sample.

(e) False. A Type I error occurs when we wrongly reject the null hypothesis. Since P-value = $0.022 \not< \alpha = 0.01$ we would not reject the null hypothesis. Thus, we cannot be making a Type I error. (A Type II error, however, could have been made.)

II.18 **(a)** False. Since the 95% confidence interval does not contain zero, we know that the *t* test would reject H_0 with $\alpha = 0.05$. Thus, the P-value would be less than $\alpha = 0.05$.

(b) True. Since the 95% confidence interval does not contain zero, we know that the *t* test would reject H_0 with $\alpha = 0.05$.

(c) Cannot be determined. Because the interval is a 95% interval and it does not contain zero, we know that the P-value is less than 0.05, but we cannot determine how much less without further computations.

(d) True. Similar to the above, since the P-value is known to be less than 0.05, it is also less than 0.10.

II.19 We are 95% confident that trained cats attempt between 0.17 and 1.29 fewer escapes than untrained cats, on average.

• **II.20** Since the authors are trying to describe the variability of the actual amount of hydrogen cyanide in the seed source and not variability in the sample mean, the ± 8.3 mg/kg seed would refer to the standard deviation and not the standard error.

II.21 **(a)** A paired analysis would me most appropriate. Two measurements are taken at each location yielding 15 differences for analysis.

(b) An independent samples analysis would be most appropriate. As described, there is no linking or connection between the eight observations taken within the marine protected zone and those outside of the zone.

(c) An independent samples analysis would be most appropriate. Although differences in weight will be analyzed (i.e., weight gain), the children receiving the 10% milk supplement are presumably independent of those receiving the 25% supplement. The weight gains for the groups are independent.

II.22 **(a)** There are several ways a paired analysis could be conducted. Eight days could be chosen, and on each of the eight days catch per unit effort could be measured at a location inside and outside the protected area. The data would be matched by day. Another method of pairing could be based on habitat characteristics such as depth. Eight locations in the protected habitat could be chosen first and then matched to similar locations (based on depth and perhaps other environmental factors) outside the protected habitat