1. (a) To compute the rejection criterion, first use the standard normal distribution to determine what $z$ value cuts off the most extreme 5% (.05) of the distribution of possible sample means. The alternative hypothesis is direction, expecting a population mean that is below 8, so the entire region representing the 5% least likely outcomes is at the lower end of the distribution of sample means. The cutoff for the lower 5% (.05) is $z = -1.65$. To obtain the corresponding raw score value of the mean, we need to know the standard deviation of the distribution of sample means (i.e., the standard error of the mean). So we have: $\sigma_M = \frac{2}{\sqrt{25}} = 0.4$.

The raw score value of the sample mean that corresponds to that $z$ score is:

$$M = 8 + (0.4)(-1.65) = 7.34.$$ 

(b) Now if we assume that the true mean for the older adult population is 7, then the distribution of sample means and the established rejection criterion will look like this:

(c) The region in gray represents the sample means that are small enough (different enough from the null hypothesis value of $\mu$) to allow rejection of $H_0$. In this distribution, the proportion of sample means in that region is determined by the $z$ value that corresponds to $M = 7.34$. We have: $z = \frac{7.34 - 7}{0.4} = 0.85$. The area in the larger portion of the normal distribution for $z = 0.85$ is .8023. So the power of the $z$ test to reject the null hypothesis in this case is .80, if we assume that the true value of $\mu$ is 7.

(d) If we had used $\alpha = .01$ instead of .05, the critical $z$ value would have been $z = -2.33$, and the rejection criterion would be $M = 8 + (0.4)(-2.33) = 7.07$. If the real value of $\mu$ were 7, then the rejection criterion of 7.07 would correspond to a $z$ value of $z = \frac{7.07 - 7}{0.4} = 0.18$. The resulting area under the normal distribution below $z = .18$ is .5714, meaning that power would equal .57.
2. (a) The necessary $\delta$ value from Table E.5 (based on a target power level of .80 and a nondirectional test with $\alpha = .05$) is 2.80; $n = 2(2.80/0.5)^2 = 62.72$ or 63 per group.
(b) The necessary $\delta$ value (with $\alpha = .10$) from Table E.5 is 2.50; $n = 2(2.50/0.5)^2 = 50$ per group.

3. (a) $t = -1.23$, not significant [with $\alpha = .05$ and a two-tailed test, $t_{\text{critical}}(30) = 2.042$; no entry is available in the $t$ table for $df = 38$ so we next the next lowest $df$ value in the table, namely, 30]
\[s_p^2 = \frac{9.2^2(19)+8.8^2(19)}{38} = 81.04\]
\[s_{M-M} = \sqrt{\frac{81.04}{20} + \frac{81.04}{20}} = 2.85\]
\[t = \frac{22.8 - 26.3}{2.85} = -1.23\]
(b) $d = (22.8 - 26.3)/(9.00) = -.39$ [or .39--it is arbitrary which mean is subtracted from which; $s_p = \sqrt{81.04}$; note that you cannot simply average the two standard deviations to get the pooled standard deviation, even though the sample sizes are equal; instead, you must first square the standard deviations to get the sample variances, then average the two variances or convert the variances to SS values and go from there]
(c) This effect size is between small (.2) and medium (.5).
(d) For $d = .39$, $\delta = .39(\sqrt{20/2}) = 1.23$; from Table E.5, power is between .22 and .26. Note that the effect size is expressed as a positive value. The negative sign obtained in part (b) arises from an arbitrary choice of which mean to subtract from which. When computing $\delta$ values, only positive values of $d$ are used.
(e) One cannot conclude from this study that an effect size of .39 does not exist in the population. This study had very low power to detect such an effect, so even if the effect existed, a study such as this one would be very unlikely to produce a significant $t$ ratio. That is, this study's failure to find an effect is not inconsistent with the existence of an effect size of around .39.
(f) For $d = 1.0$, $\delta = 1.0(\sqrt{20/2}) = 3.16$; from Table E.5, power is about .88. This study had substantial power to detect an effect size of one standard deviation, yet no effect was found. Therefore, it is reasonable to conclude that an effect of that size does not exist in the population, because if such an effect existed, this study most likely would have detected it.
(g) The `pwr.t.test` function in R can be used to estimate the required sample size. Note that you will need to install the `pwr` package using the `install.packages` command if you have not done so previously, and that you will have to load this package for this session using the `library` command. The required parameters for this function are the effect size, significance level, desired level of power, type of $t$ test, and the nature of the alternative hypothesis:
> `pwr.t.test(d = .4, sig.level = .05, power = .8, type = "two.sample", alternative = "two.sided")`

```
Two-sample t test power calculation

 n = 99.08032
d = 0.4
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: $n$ is number in *each* group

4. (a) This is a nondirectional test with $\alpha = .05$, so the critical $z = \pm 1.96$; variability among possible sample means is computed as the standard error of the mean, which is $\sigma_M = 12/\sqrt{16} = 3$; the values of the
sample mean that are the rejection criteria (defining the boundaries of the regions of rejection) are \( M = 40 \pm 3(1.96) = 34.12 \) and 45.88. Under \( H_1 \), \( z \) for a mean of 34.12 = \((34.12 - 35)/3\) = -0.29, and \( z \) for a mean of 45.88 = \((45.88 - 35)/3\) = 3.63; for \( z = -0.29 \), area in smaller portion = .3859; for \( z = 3.63 \) area in smaller portion is about .0001; power = .3859 + .0001 = .3860 or .39.

(b) With \( \alpha = .01 \), the critical \( z = \pm 2.58 \). With \( \sigma_M = 3 \), the rejection criteria are \( M = 40 \pm 3(2.58) = 32.26 \) and 47.74. Under \( H_1 \), \( z \) for a mean of 32.26 = \((32.26 - 35)/3\) = -0.91, and \( z \) for a mean of 47.74 = \((47.74 - 35)/3\) = 4.25; for \( z = -0.91 \), area in smaller portion = .1814; for \( z = 4.25 \) area in smaller portion is virtually zero; power = .1814 or .18. Notice that power is much lower when the value of \( \alpha \) is reduced.

5. Using a significance level of .01 instead of .05 would reduce power (increase the probability of a type II error).

6. The necessary \( \delta \) value from Table E.5 is 2.80; \( n = (2.80/0.5)^2 + 1 = 32.36 \) or 33.

7. The necessary \( \delta \) value from Table E.5 is 3.25 (midway between 3.20 and 3.30); alternatively you could use either 3.20 or 3.30
\[ n = 2(3.25/0.5)^2 = 84.5 \] or 85 per group.

8. This is a very large sample, so it will have a high level of power to detect even small effects. For a correlation effect of .10, which means that chocolate consumption explains only 1% of the variability in getting the type of cancer being studied (\( r^2 = .10^2 = .01 \)), a sample size of 1,500 produces \( \delta = .10\sqrt{1500 - 1} = 3.87 \). From Howell's table E.5, with \( \alpha = .05 \) and a nondirectional test, we have power = .97. Thus, it is highly likely that if there is a correlation of only .10, this study will find it.

9. The relevant statistical effect size from Table E.5 is \( \delta = 2.80 \). With sample sizes of 16, we have \( 2.80 = d \sqrt{\frac{16}{2}} \). This reduces to \( 2.80 = d (2.828) \) and to \( d = 2.80/2.828 = 0.99 \). So the effect size would have to be large (nearly a whole standard deviation) for power to be .8.

10. With the correlation at .45, we have \( \delta = .5 \sqrt{\frac{20}{2(1-.45)}} = 2.13 \), and this yields a power estimate of .32 from Table E.5. With a correlation of .90, we instead would have \( \delta = .5 \sqrt{\frac{20}{2(1-.90)}} = 5.00 \), which leads to a power estimate of .99. With a stronger correlation between conditions, power is higher. If the correlation had been zero, we would have \( \delta = .5 \sqrt{\frac{20}{2(1-.0)}} = 1.58 \), and a power estimate of at best .17 (taken from \( \delta = 1.60 \)). If this were an independent-samples design with 20 subjects in each condition,
we would have exactly the same situation as a related-samples design with no correlation between conditions (which implies independence between the conditions): \( \delta = 0.5 \frac{20}{\sqrt{2}} = 1.58 \). Notice that the correlation between conditions plays a big role in determining the power of the related-samples \( t \) test, all the way up from independence (no correlation at all) and low power, to a high correlation and high power.

11. When one cannot assume equal variance for the two populations involved in an independent-samples \( t \) test, the degrees of freedom used to determine the critical \( t \) ratio is reduced (in the most extreme case) from \( df_1 + df_2 \) to the \( df \) for the smaller of the two samples. With a smaller value for \( df \), the critical \( t \) ratio will be larger and so a larger effect will be needed to achieve a significant difference between means.

This can be seen in the two example cases. The obtained \( t \) ratio is the same in both cases, because even though the sample variances are different from one another in the second case, the difference between means \((30.0 - 27.4 = 2.6)\) and even the standard error of difference \((1.265)\) is the same in the two cases. The resulting \( t \) ratio in each case is 2.06. With \( df = 38 \), which is based on the equal variance assumption that holds in Case 1, the critical \( t \) ratio from Table E.6 is between 2.042 (from \( df = 30 \)) and 2.021 (from \( df = 40 \)), making the \( t \) ratio significant. For Case 2, the strict modification of the \( df \) value in response to heterogeneity of variance produces \( df = 19 \), and a critical \( t \) ratio of 2.093, meaning that the observed \( t \) ratio fails to reach significance. This result means that with heterogeneity of variance, the researcher has less power than when the population variances can be assumed to be equal. In this case, the difference in power is not great, as you can see that even in Case 2 the test nearly came out significant. With a more modest adjustment to \( df \), as might be done in a statistical package such as R, even Case 2 might have proved to be a significant effect. There are factors, especially sample size and effect size, that have much bigger influences on power.

12. (a) The critical \( t \) ratio with 15 \( df \) is 2.131, so the obtained \( t \) ratio (2.58) is significant.

(b) For the first set of data, the pooled estimate of population standard deviation is obtained by pooling the two sample variances, then taking the square root: 
\[
sp = \sqrt{\frac{(15)(67.9^2)+(15)(79.3^2)}{15+15}} = 73.82
\] 
So we have \( d = \frac{925.2-90.63}{73.82} = 0.26 \)

(c) Power to detect an effect of this size with a related-samples \( t \) test is based on
\[
\delta = 0.26 \frac{\sqrt{16}}{2(1-0.93)} = 2.78.
\] 
From Howell's table, this value of \( \delta \) is close to 2.80, which yields a power estimate (for \( \alpha = .05 \) and a nondirectional test) of .80.

(d) For the second set of data, the \( t \) ratio is not significant.

(e) The pooled standard deviation is only slightly different: 
\[
sp = \sqrt{\frac{(15)(67.9^2)+(15)(79.7^2)}{15+15}} = 74.04, \text{ so } d = \frac{925.2-914.1}{74.04} = 0.15.
\] 
The power estimate is based on \( \delta = 0.15 \frac{\sqrt{16}}{2(1-0.94)} = 1.73 \)

From Howell's table, the power value for \( \delta = 1.70 \) is .40, so power here is slightly larger than .40.

Notice that when using the observed effect size, if the hypothesis test turns out to be significant (as it was in the first case), power is quite high, but when it is nonsignificant (as it was in the second case), power is low. This is a general principle and it indicates that there is not much point in using the
observed effect size when computing power because one will get little new information beyond what the outcome of the significance test provided.

(f) Treating the first set of data as an independent-samples $t$ test yields an estimate of the standard error of difference as follows: 

$$s_{M_1-M_2} = \sqrt{\frac{5449.4}{16} + \frac{5449.4}{16}} = 26.10$$

(note that $s_p^2 = 73.82^2 = 5449.4$)

The resulting $t$ ratio is 

$$t(30) = \frac{925.2-906.3}{26.10} = 0.72$$

This $t$ ratio is clearly not significant. Notice that the independent-samples $t$ test, even though it is based on twice as many subjects as the related-samples $t$ test, is much less powerful. The power of the independent-samples $t$ test to detect an observed effect size of 0.26 (the same effect size as in part (b)) is based on $\delta = 0.26 \sqrt{\frac{16}{2}} = 0.73$. This value of $\delta$ is so small that it does not appear in Howell's table, but we can see from the table that with a nondirectional test and $\alpha = .05$, power is less than .17 – quite dismal.

13. (a) The study failed to find a significant effect, but if the null hypothesis actually is false, then a Type II decision error has occurred (the study failed to find an effect that exists).

(b) To compute the power of a one-sample $t$ test to detect an effect of .8 with $n = 20$, $\alpha = .05$, and a directional alternative hypothesis, we have $\delta = .8\sqrt{20} = 3.58$, which yields power of about .97. So this study had plenty of power to detect a large effect, yet it found no significant effect. Therefore, if a precognition effect actually exists, it is unlikely to be a large effect -- if it were, then this study should have found it.

(c) As for a small effect, this study's power to detect such an effect was as follows: $\delta = .2\sqrt{20} = 0.89$, which yields power of less than .26. With such low power, a true small effect might exist, but this study would be unlikely to detect it. So we cannot tell whether there is no effect at all or just a small effect, and that is why a null effect such as this does not allow us to accept the null hypothesis.

14. As in question 2, ensure that the pwr package is loaded into your library.

(a) Specify the sample size as 20, the effect size as .5, etc.

> pwr.t.test(n = 20, d = .5, sig.level = .05, type = "two.sample", alternative = "two.sided")

```
Two-sample t test power calculation

     n = 20
    d = 0.5
sig.level = 0.05
      power = 0.337939
alternative = two.sided
```

NOTE: n is number in *each* group

Power is about .34, which is quite low.

(b) To have power of .8 with an effect size of .5, here is how many subjects are needed in each group:

> pwr.t.test(d = .5, sig.level = .05, power = .8, type = "two.sample", alternative = "two.sided")

```
Two-sample t test power calculation

    n = 13.7661
```

Power is about 0.8, which is quite high.
Two-sample t test power calculation

\[ n = 63.76561 \]
\[ d = 0.5 \]
\[ \text{sig.level} = 0.05 \]
\[ \text{power} = 0.8 \]
\[ \text{alternative} = \text{two.sided} \]

NOTE: \text{n is number in *each* group}

Each group would have to have 64 subjects, not just 20 as in the original study.

(c) Now we see how much power a study with 64 subjects per condition would have to detect an effect of \( d = 0.2 \).

\[ > \text{pwr.t.test}(n = 64, \ d = 0.2, \ \text{sig.level} = 0.05, \ \text{type} = \text{"two.sample"}, \ \text{alternative} = \text{"two.sided"}) \]

Two-sample t test power calculation

\[ n = 64 \]
\[ d = 0.2 \]
\[ \text{sig.level} = 0.05 \]
\[ \text{power} = 0.2022645 \]
\[ \text{alternative} = \text{two.sided} \]

NOTE: \text{n is number in *each* group}

Hmm… not very encouraging – power is only .20!

(d) To push power up to .80, we need this many subjects:

\[ > \text{pwr.t.test}(d = 0.2, \ \text{sig.level} = 0.05, \ \text{power} = 0.8, \ \text{type} = \text{"two.sample"}, \ \text{alternative} = \text{"two.sided"}) \]

Two-sample t test power calculation

\[ n = 393.4057 \]
\[ d = 0.2 \]
\[ \text{sig.level} = 0.05 \]
\[ \text{power} = 0.8 \]
\[ \text{alternative} = \text{two.sided} \]

NOTE: \text{n is number in *each* group}

Ugh! That would be a lot of work: 393 subjects in each condition! But at least we would have a very good chance of finding an effect of .2.

(e) Pushing the assumed effect size to a lower value renders even a sample size of 393 subjects per condition inadequate:

\[ > \text{pwr.t.test}(n = 393, \ d = 0.1, \ \text{sig.level} = 0.05, \ \text{type} = \text{"two.sample"}, \ \text{alternative} = \text{"two.sided"}) \]

Two-sample t test power calculation
n = 393
d = 0.1
sig.level = 0.05
power = 0.2881647
alternative = two.sided

NOTE: n is number in *each* group

Yikes, power is only .29! So pump up the sample size again to bring power up to .80:
> pwr.t.test(d = .1, sig.level = .05, power = .8, type = "two.sample", alternative = "two.sided")

Two-sample t test power calculation

 n = 1570.733
d = 0.1
sig.level = 0.05
power = 0.8
alternative = two.sided

NOTE: n is number in *each* group

Ahh, that's more like it. Only 1,571 subjects per condition. Good luck with that.

Notice that we can keep shrinking the assumed effect size, but as long as we have a large enough sample, we will have adequate power to detect it. Similarly, when we fail to detect an effect in a particular study, we cannot say that the null hypothesis is true (i.e., conclude that there is no effect), because the true effect size might just be too small for our study to detect it (primarily due to an insufficient sample size).