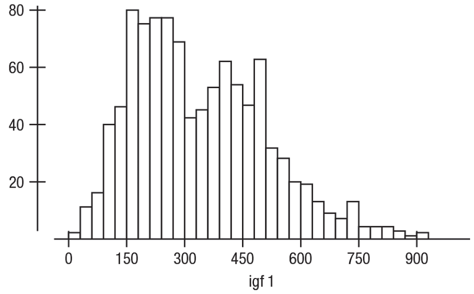
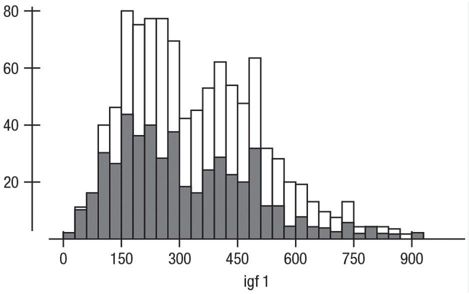
Cumulative Review Exercises

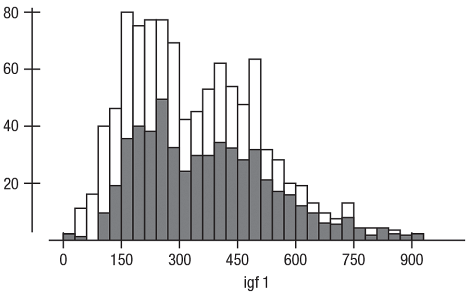
1. **Igf.**

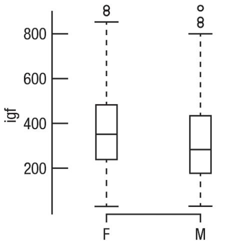
a) The distribution of insulin-like growth factor is bimodal and skewed to the right, with modes at approximately 200 and 400 $\mu g / l.$ The median is 316 $\mu g / l,$ with the middle 50% of observations between 203.5 and 464 $\mu g / l.$ The lowest observation was 25 $\mu g / l$ and the highest observation was 915 $\mu g / l.$

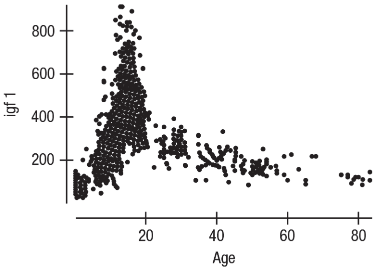


b) For both males and females, the distributions of insulin-like growth factor are very similar to the overall distribution. Both are bimodal and skewed to the right, with modes at approximately 200 and 400 $\mu g / l.$ The mean and median for males are 310.9 $\mu g / l$ and 280 $\mu g / l,$ respectively. Both the mean and median for females are higher, at 368.1 $\mu g / l$ and 352 $\mu g / l,$ respectively. The middle 50% of observations for males is between 176 and 430.75 $\mu g / l.$ For females, the middle 50% of the observations is generally higher, between 232.50 and 484 $\mu g / l.$ The extreme observations for males and females are approximately equal with minimum values of 29 and 25 $\mu g / l$ and maximum values of was 915 and 914 $\mu g / l,$ for males and females, respectively. Both groups have high outliers.







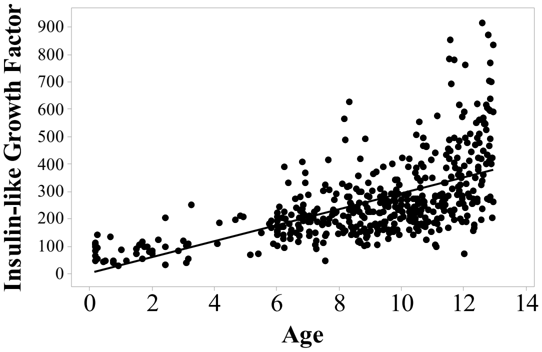


c) The scatterplot is shown above to the right. For ages younger than 20 years, the association between age and insulin-like growth factor is moderately strong, positive, and curved, with a very steep slope. For ages older than 20 years, there is a moderately strong, negative, linear association between age and insulin-like growth factor.

d) It is not appropriate to use linear regression to model the association between insulin-like growth factor and age. The relationship is neither linear nor consistent.

2. Igf13.

a) The scatterplot of insulin-like growth factor and age, at the right, shows a curved association. We aren't going to be able to make a decent linear model for this association.



Model Summary

S	R-sq	R-sq(adj)
120.525	35.52%	35.38%

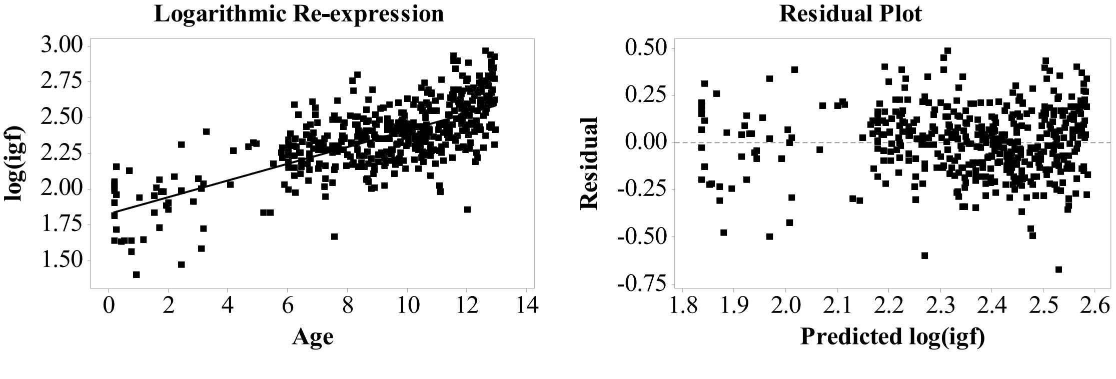
Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	2.35819	17.46	0.14	0.893
Age	29.0376	1.828	15.88	<0.0001

The equation of the least squares regression line is $\widehat{Igf} = 2.35819 + 29.0376age.$

b) $\widehat{Igf} = 2.35819 + 29.0376age = 2.35819 + 29.0376(20) \approx 583.11;$ According to the linear model, 20-year-olds are expected to have mean insulin-like growth factor levels of 583.11 $\mu g / l.$ This prediction is not likely to be accurate for two reasons. First, this is an extrapolation beyond the scope of our data. This model was made from a data set that only had subjects up to 13 years old. Second, the scatterplot shows a curved association. This linear model is not appropriate in this case. We can see from our scatterplot in Exercise 1 that 20-year-olds actually have insulin-like growth factor levels of about 300 to 400 $\mu g / l.$

c) The scatterplot of the logarithm of insulin-like growth factor and age, shown below, shows a much straighter association. The residual plot shows no pattern. The logarithmic re-expression allows us to fit an appropriate model.



2. (continued)

Model Summary

S	R-sq	R-sq(adj)
0.176630	51.10%	50.99%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	1.82574	0.0256	71.4	<0.0001
Age	0.058616	0.0027	21.9	<0.0001

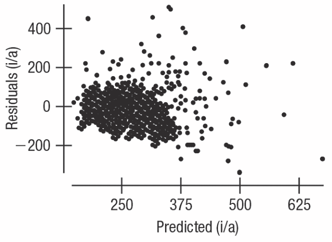
The equation of the least squares regression line is $\widehat{\log(Igf)} = 1.82574 + 0.058616age.$

d) No, using this model would be an extrapolation far beyond the data. This model was made from a data set that only had subjects up to 13 years old. We know from Exercise 1 that the relationship between *igf* and *age* changes drastically at age 20 or so.

e) $\widehat{\log(Igf)} = 1.82574 + 0.058616age = 1.82574 + 0.058616(20) \approx 2.998$, so $\widehat{igf} \approx 10^{2.998} \approx 995$; The linear regression model for the re-expressed data predicts that the mean insulin-like growth factor level of 20-year-olds is 995 $\mu g / l.$ This prediction is worse than the previous prediction. The relationship between *igf* and *age* changes before age 20.

3. More Igf13.

a) The association between insulin-like growth factor and weight is linear, positive, and weak. The association is straight enough to try linear regression.



Model Summary

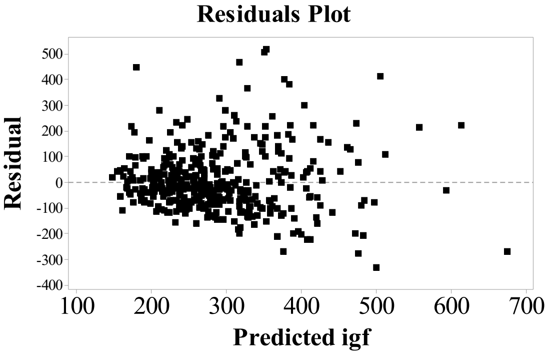
S	R-sq	R-sq(adj)
121.264	31.17%	31.00%

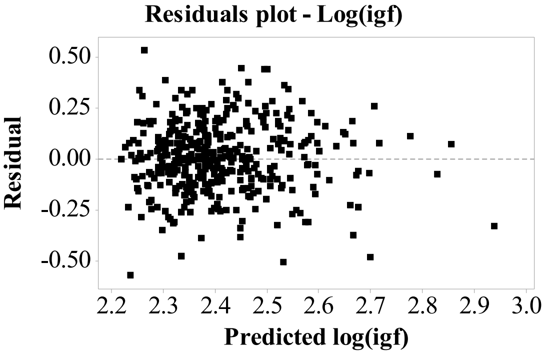
Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	2.94202	21.81	0.135	0.8928
Weight	8.06787	0.5972	13.5	<0.0001

3. (continued)

The equation of the least squares regression line is $\widehat{Igf} = 2.94202 + 8.06787 weight$. However, the residuals plot does not show a constant variance. We will have to try a re-expressed model.





Re-expressing the insulin-like growth factor with logarithms improves the model.

```
Model Summary
S              R-sq         R-sq(adj)
0.170465       29.83%       29.66%

Coefficients
Term           Coef         SE Coef        T-Value     P-Value
Constant       2.02218      0.0307         65.95       <0.0001
Weight         0.0109898    0.000840       13.09       <0.0001
```

The equation of the least squares regression line is $\widehat{\log(Igf)} = 2.02218 + 0.0109898 weight$.

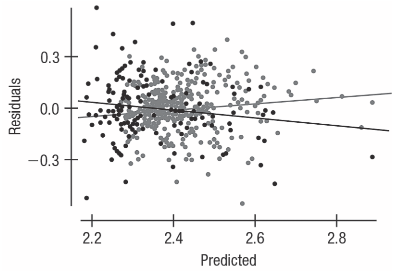
b) To add the variable *sex*, we must code it. Here we used male = 1, female = 0.

```
Model Summary
S          R-sq          R-sq(adj)
0.164999   34.42%        34.10%

Coefficients
Term           Coef         SE Coef        T-Value     P-Value
Constant       2.05963      0.0305         67.51       <0.0001
Weight         0.011009     0.000813       13.55       <0.0001
Sex            –0.087745    0.0165         –5.30       <0.0001
```

The interpretation of the coefficient of the logarithmic re-expression is difficult to interpret meaningfully. If we go back the linear model, and include *sex* as a predictor, the interpretation of the coefficient is that boys have, on average, an igf level that is 61.6 units lower than girls after allowing for the effects of weight.

A plot of the residuals shows that the slopes for boys and girls are not parallel for regression on log(igf). So, a careful student might conclude that this isn't an appropriate regression.



3. (continued)

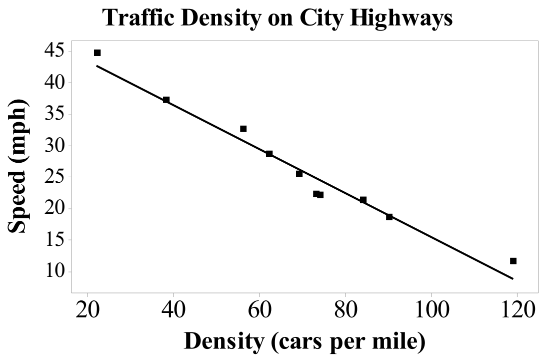
c) For a 100-pound female, $\widehat{\log(Igf)} = 2.05963 + 0.011(100) \approx 3.15963.$ In natural units, $10^{3.15963} \approx 1444.$ This is an extrapolation well beyond the data used to build the model.

d) No, all of the coefficients are not statistically significant. Height and weight are collinear (correlation 0.887).

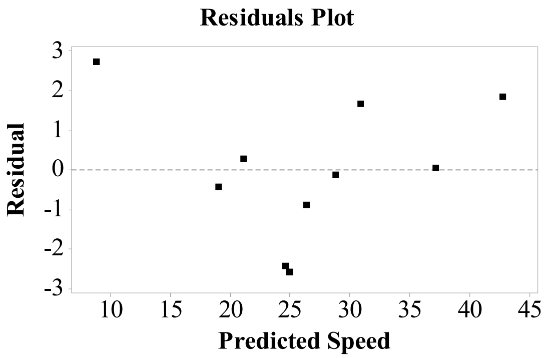
e) Use age as a predictor. It is odd to leave that out when we have ages from infants to teens and are trying to use height and weight as predictors.

4. **Speed and density.**

a) There is a strong, negative, linear association between *Speed* and *Density.* Higher traffic density is associated with lower speeds.



Traffic Density on City Highways



Residuals Plot

b) The association between *Speed* and *Density* is straight enough to fit a least squares regression model, and the residuals plot shows reasonable scatter. The equation of the model is $\widehat{Speed} = 50.47 - 0.3503 Density.$

c) $\widehat{Speed} = 50.47 - 0.3503(56) \approx 30.85;$ The predicted speed when the traffic density is 56 cars per mile is 30.85 mph. The residual is $32.5 - 30.85 \approx 1.65$ mph.

$\widehat{Speed} = 50.47 - 0.3503(20) \approx 43.46;$ The predicted speed when the traffic density is 20 cars per mile is 43.46 mph. The residual is $36.1 - 43.46 \approx -7.36$ mph.

d) The second residual is more unusual, since it is farther away from the predicted value.

e) A point at 125 cars per mile and 55 mph would be a very high speed at a very high traffic density. The point would be highly influential. The correlation and the slope would both decrease in magnitude. In other words, the association would appear weaker than it is currently, and the slope would be shallower. This wording is a bit tricky, since the shallower slope and weaker correlation are technically *higher* (closer to zero), since they are both negative to start with.

f) $\widehat{Speed} = 50.47 - 0.3503(200) \approx -19.6;$ The predicted speed when the traffic density is 200 cars per mile is –19.6 mph. This is not reasonable. First, the speed cannot be negative. This happens because 200 cars per mile is an extrapolation well beyond the scope of our data.

g) You can certainly convert all of the speeds and densities to *z*-scores, and find the regression equation. However, since the correlation between *Speed* and *Density* is $r = -0.98.$ the regression equation for the standardized variables is $\widehat{z_{Speed}} = -0.98 z_{Density}.$

h) Since the correlation is the same regardless of the direction of prediction, the equation that predicts standardized *Density* from standardized *Speed* is $\widehat{z_{Density}} = -0.98 z_{Speed}.$

5. **Hospital variables.**

   a) Two-sided one-proportion z-test for a proportion, $H_0: z = 0.50$.

   b) Two-sided two-proportion z-test for difference between proportions.

   c) One-sided two-sample t-test for difference between means.

   d) Two-sided two-sample t-test for difference between means.

   e) Two-sided linear regression t-test for slope of the line of best fit predicting 2018 appointments from 2017 appointments.

6. **Lake source cooling.**

   a) IV

   b) II

   c) III

   d) III

   e) III

   f) IV

7. **Life expectancy and literacy.**

   a) The scatterplot looks straight enough to use the linear regression model, $\widehat{LifeExp} = 72.3949 - 1.2960 Illiteracy$. When the illiteracy rate is 0.70, the life expectancy is predicted to be $\widehat{LifeExp} = 72.3949 - 1.2960(0.70) \approx 71.49$ years.

   b) IV

   c) $r = \sqrt{R^2} = \sqrt{0.3462} \approx 0.588$

   d) IV

   e) I

8. **Hotel maids.**

   The correct answer is d.

9. **Hotel maids.**

   The correct answer is a.

10. **Hotel maids.**

    The correct answer is e.

11. **Hotel maids.**

    The correct answer is c.

12. **Hotel maids.**

    The correct answer is c.

13. **Inference methods.**

    a) A

    b) C

    c) E

    d) D

    e) B

14. **Olympic archery.**

    a) $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = (0.0814) \pm 1.960 \sqrt{\frac{(0.0814)(0.9186)}{4608}} = (0.0735, 0.089)$

    b) I am 95% confident that between 7.4% and 8.9% of arrows will hit in the X area.

15. Olympic archery.

a) $\bar{y} \pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right) = 624.41 \pm t^*_{63}\left(\frac{28.52}{\sqrt{64}}\right) \approx (617.28, 631.53)$; I am 95% confident that the mean women's archery seeding score is between 617.28 and 631.53.

b) This question does not relate to the confidence interval constructed in part (a). Confidence intervals are about means, not individual performances. With a mean score of 624.41 and a standard deviation of 28.52, a score of 660 is only about 1.25 standard deviations above the mean. Zhang Juanjuan's score is high, but not extraordinarily high.

16. Olympic archery.

To determine whether scores are consistent in the two halves, we will perform a paired *t*-test.

H$_0$: The mean difference between scores in the two halves is zero. $(\mu_{1st-2nd} = 0)$

H$_A$: The mean difference between scores in the two halves is different from zero. $(\mu_{1st-2nd} \neq 0)$

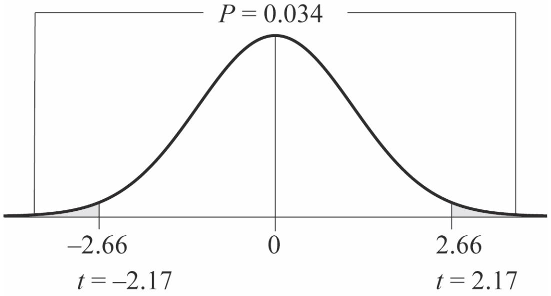
Paired data assumption: The data are paired by archer.
Randomization condition: We will assume that these archery scores are representative of the scores of female Olympic archers.
Normal population assumption: The histogram of differences between the first and second half scores is roughly unimodal and symmetric, but there are outliers, one on the low end, and one on the high end. The large sample size means that the Central Limit Theorem will allow us to continue. We will perform the test twice, once with the outliers, and once without, to see if we reach different conclusions.

Since the conditions are satisfied, the sampling distribution of the difference can be modeled with a Student's *t*-model with

64 – 1 = 63 degrees of freedom, $t_{63}\left(0, \frac{9.80}{\sqrt{64}}\right)$. We will use a paired *t*-test, with $\bar{d} = 2.66$.

$t = \frac{\bar{d} - 0}{\frac{s}{\sqrt{n}}} = \frac{2.66 - 0}{\frac{9.80}{\sqrt{64}}} \approx 2.17$; Since the P-value = 0.0338 is low, we reject the null hypothesis. There is strong of a difference in first and second half scores. First half scores average significantly higher.

Without the outliers, $t = 2.41$, and the P-value = 0.019. We would reach the same conclusion.

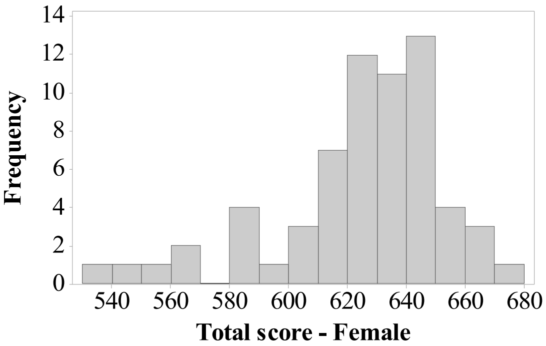


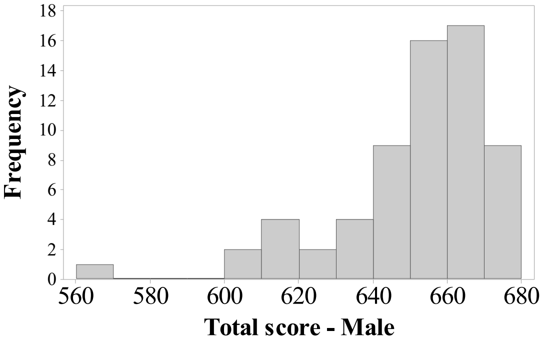
17. Olympic archery.

H$_0$: The mean seeding score is the same for females and males. $\left(\mu_F = \mu_M \text{ or } \mu_F - \mu_M = 0\right)$

H$_A$: The mean seeding score is different for females and males. $\left(\mu_F \neq \mu_M \text{ or } \mu_F - \mu_M \neq 0\right)$

Independent groups assumption: Female scores should not affect male scores.
Randomization condition: We will assume that these archery scores are representative of the scores of female and male Olympic archers.
Nearly Normal condition: The distributions of seeding scores are both skewed to the right, but with sample sizes of 64 for both groups, the Central Limit Theorem will allow us to continue.





Since the conditions are satisfied, it is appropriate to model the sampling distribution of the difference in means with a Student's *t*-model, with 115.77 degrees of freedom (from the approximation formula). We will perform a two-sample *t*-test. We have the following.

$\bar{y}_F = 624.40625$ $\bar{y}_M = 651.09375$

$s_F = 28.521626$ $s_M = 20.991093$

$n_F = 64$ $n_M = 64$

The sampling distribution model has mean 0, with standard error

$$SE(\bar{y}_F - \bar{y}_M) = \sqrt{\frac{28.521626^2}{64} + \frac{20.991093^2}{64}} \approx 4.4266754.$$

The observed difference between the mean scores is 624.40625 – 651.09375 ≈ –26.6875.

$$t = \frac{\left(\bar{y}_F - \bar{y}_M\right) - (0)}{SE\left(\bar{y}_F - \bar{y}_M\right)} \approx \frac{-26.6875}{4.426674} \approx -6.03;$$ Since the P-value < 0.0001 is very low, we reject the null hypothesis.

There is strong evidence that the mean seeding scores for males and females are different. These data suggest that males have significantly higher mean seeding scores.

18. Olympic medals.

a) H_0: The distribution of type of medal is the same for the three countries.

H_A: The distribution of type of medal is the different for at least one of the three countries.

We have three groups and one variable, type of medal, so we will perform a χ^2 test for homogeneity.

b) The P-value of 0.497 is very close to our typical level of significance of $\alpha = 0.05$. Formally, we could reject the null hypothesis, but with a P-value that close to α, we should be cautious about making too strong a claim. These data provide some evidence of a difference in medal distribution among the three countries.

c) The residuals indicate that the biggest difference is France's deficit in gold medals. Great Britain has more gold and fewer bronze medals than might be expected. Since none of the standardized residuals have magnitude greater than 2, there doesn't appear to be any great differences in the distribution of medals from these countries.

19. Belmont stakes.

a) We are 95% confident that the mean clockwise speed is between 1.76 and 2.54 miles per hour slower than the mean counterclockwise speed. Since 0 is not in the interval, these data provide evidence that the clockwise speeds are statistically significantly slower, on average.

b) Oddly, horses appear to run faster in longer races, with the 1.625-mile races being an exception. But that doesn't make sense, so there may be a lurking variable.

c) I expect the confidence interval to be narrower. There is a bigger effect and less variation.

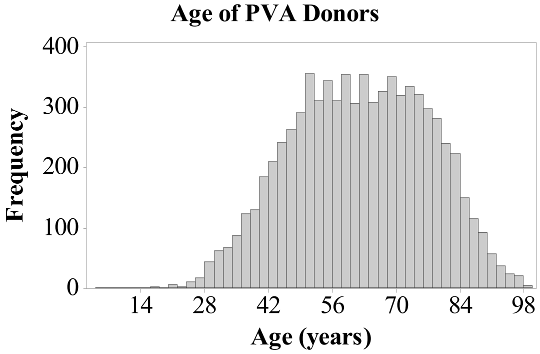
d) Year is the lurking variable. In fact, horses have gotten faster, so more recent races (run counterclockwise) were faster. There is no evidence that horses care which way they run.

20. Body fat.

Answers will vary.

21. PVA.

a) The distribution of ages of PVA donors is unimodal and symmetric, with a mean donor age of about 61.5 years, and a standard deviation on donor age of about 15.0 years. There are several questionable ages. It seems unlikely that there were donors of ages 4, 13, and 15 years. Perhaps someone donated to the organization on their behalf.



21. (continued)

We have taken a random sample of the entire donor database, and the distribution of ages is roughly unimodal and symmetric, so we can use a one-sample *t*-interval to estimate the mean age of all donors to PVA.

$$\bar{y} \pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right) = 61.524 \pm t^*_{7599}\left(\frac{15.030}{\sqrt{7600}}\right) \approx (61.186,\ 61.862)$$

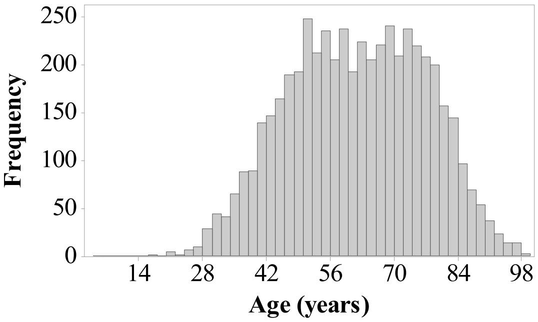
We are 95% confident that the mean age of all PVA donors is between 61.186 and 61.862 years.

b) H$_0$: The mean age of homeowners among PVA donors is the same as the mean age of non-homeowners among PVA donors. $\left(\mu_H = \mu_N \text{ or } \mu_H - \mu_N = 0\right)$

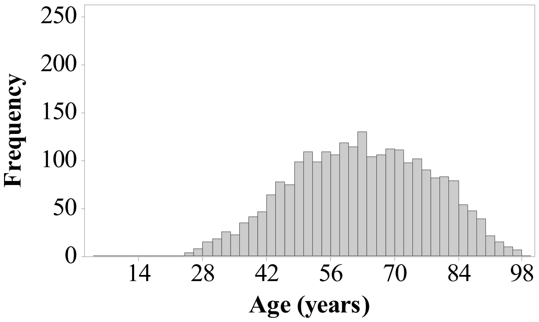
H$_A$: The mean age of homeowners among PVA donors higher than the mean age of non-homeowners among PVA donors. $\left(\mu_H \neq \mu_N \text{ or } \mu_H - \mu_N \neq 0\right)$

Independent groups assumption: Homeowners ages should not affect non-homeowners ages.
Randomization condition: This data set is a random sample of all donors in the PVA database.
Nearly Normal condition: The distributions of ages for both groups are roughly unimodal and symmetric.

Homeowners



Non-homeowners



Since the conditions are satisfied, it is appropriate to model the sampling distribution of the difference in means with a Student's *t*-model, with 4896 degrees of freedom (from the approximation formula). We will perform a two-sample *t*-test. We are have the following.

$\bar{y}_H = 61.234$	$\bar{y}_N = 62.121$
$s_H = 615.016$	$s_N = 15.044$
$n_H = 5120$	$n_N = 2480$

The sampling distribution model has mean 0, with standard error

$$SE(\bar{y}_H - \bar{y}_N) = \sqrt{\frac{15.016^2}{5120} + \frac{15.044^2}{2480}} \approx 0.368.$$

The observed difference between the mean scores is 61.234 – 62.121 ≈ –0.887.

$$t = \frac{(\bar{y}_H - \bar{y}_N) - (0)}{SE(\bar{y}_H - \bar{y}_N)} \approx \frac{-0.887}{0.368} \approx -2.41;$$ Since the P-value = 0.992 is high, we fail to reject the null hypothesis. There is no evidence that the mean age of homeowners is higher than the mean age of non-homeowners among PVA donors. (Did you make it this far without realizing that the mean age of homeowners is actually *lower* in this sample?)

Since the mean age of donors overall is quite high, this discrepancy could be caused by many of the donors having retired. These retirees may no longer own homes.

21. (continued)

c) H$_0$: Among PVA donors, the proportion of males who own homes is the same as the proportion of females who own homes. $\left(p_M = p_F \text{ or } p_M - p_F = 0\right)$

H$_A$: Among PVA donors, the proportion of males who own homes is greater than the proportion of females who own homes. $\left(p_M > p_F \text{ or } p_M - p_F > 0\right)$

Randomization condition: This data set is a random sample of PVA donors.
Independent groups condition: The groups are not associated.
Success/Failure condition: $n\hat{p}$ (male) =2287, $n\hat{q}$ (male) = 1026, $n\hat{p}$ (female) =2833, and $n\hat{q}$ (female) = 1454 are all greater than 10, so the samples are both large enough.

Since the conditions have been satisfied, we will model the sampling distribution of the difference in proportion with a Normal model with mean 0 and standard deviation estimated by

$$SE\left(\hat{p}_{None}-\hat{p}_{Dep}\right)=\sqrt{\frac{\hat{p}_M\hat{q}_M}{n_M}+\frac{\hat{p}_F\hat{q}_F}{n_F}}=\sqrt{\frac{\left(\frac{2287}{3313}\right)\left(\frac{1026}{3313}\right)}{3313}+\frac{\left(\frac{2833}{4287}\right)\left(\frac{1454}{4287}\right)}{4287}}\approx 0.010808.$$

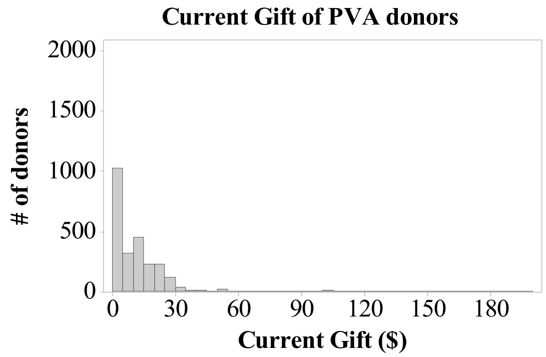
(Note: The pooled standard error could have used as well.)

The observed difference between the proportions is 0.690311 – 0.660835 = 0.0294758.

$z=\dfrac{\hat{p}_M-\hat{p}_F}{SE(\hat{p}_M-\hat{p}_F)}=\dfrac{0.0294758}{0.010808}\approx 2.72$; Since the P-value = 0.006 is low, we reject the null hypothesis. There is evidence to suggest that, among PVA donors, the proportion of males who own homes is higher than the proportion of females who own homes.

We could have used a χ^2 test for independence as well (χ^2 = 7.386, P-Value = 0.0066). Recall that the test for independence is two-sided by nature, so we would have to test a hypothesis of a difference in homeownership rates, rather than the hypothesis that males have a higher homeownership rate. Nevertheless, our conclusions would have been the same.

d) The distribution of current gift of PVA donors is unimodal and extremely skewed to the right. The median gift is \$8, but gifts are highly variable. The lowest current gift is \$0 (with over a thousand current donors giving nothing this year), and the highest current gift is \$200. There are many high outliers.


Current Gift of PVA donors
# of donors
Current Gift ($)