Review of Part VII – Inference When Variables Are Related

1. Genetics.

H_0: The proportions of traits are as specified by the ratio 1:3:3:9.
H_A: The proportions of traits are not as specified.

Counted data condition: The data are counts.
Randomization condition: Assume that these students are representative of all people.
Expected cell frequency condition: The expected counts (shown in the table) are all greater than 5.

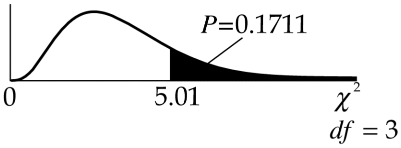
Under these conditions, the sampling distribution of the test statistic is χ^2 on 4 – 1 = 3 degrees of freedom. We will use a chi-square goodness-of-fit test.

Trait	Observed	Expected	Residual = $(Obs - Exp)$	$(Obs - Exp)^2$	Component = $\frac{(Obs - Exp)^2}{Exp}$
Attached, noncurling	10	7.625	2.375	5.6406	0.73975
Attached, curling	22	22.875	– 0.875	0.7656	0.03347
Free, noncurling	31	22.875	8.125	66.0156	2.8859
Free, curling	59	68.625	– 9.625	92.6406	1.35

$\sum \approx 5.01$

$\chi^2 = 5.01$. Since the *P*-value = 0.1711 is high, we fail to reject the null hypothesis.

There is no evidence that the proportions of traits are anything other than 1:3:3:9.



2. Tableware.

a) Since there are 57 degrees of freedom, there were 59 different products in the analysis.

b) 84.5% of the variation in retail price is explained by the polishing time.

c) Assuming the conditions have been met, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (59 – 2) = 57 degrees of freedom. We will use a regression slope *t*-interval. For 95% confidence, use $t_{57}^* \approx 2.0025$, or estimate from the table $t_{50}^* \approx 2.009$.

$b_1 \pm t_{n-2}^* \times SE(b_1) = 2.49244 \pm (2.0025) \times 0.1416 \approx (2.21,\ 2.78)$

d) We are 95% confident that the average price increases between $2.21 and $2.78 for each additional minute of polishing time.

3. Hard water.

a) H0: There is no linear relationship between calcium concentration in water and mortality rates for males. $(\beta_1 = 0)$
HA: There is a linear relationship between calcium concentration in water and mortality rates for males. $(\beta_1 \neq 0)$

b) Assuming the conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (61 – 2) = 59 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is: $\widehat{Mortality} = 1676 - 3.23(Calcium)$, where mortality is measured in deaths per 100,000, and calcium concentration is measured in parts per million.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$
$$t = \frac{-3.23 - 0}{0.48}$$
$$t \approx -6.73$$

The value of *t* = – 6.73. The *P*-value of less than 0.0001 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between calcium concentration and mortality. Towns with higher calcium concentrations tend to have lower mortality rates.

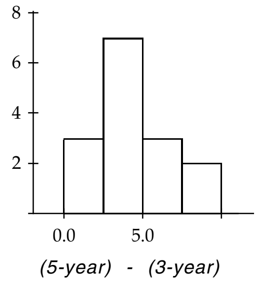
c) For 95% confidence, use $t^*_{59} \approx 2.001$, or estimate from the table $t^*_{50} \approx 2.009$.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = -3.23 \pm (2.001) \times 0.48 \approx (-4.19, -2.27)$$

d) We are 95% confident that the average mortality rate decreases by between 2.27 and 4.19 deaths per 100,000 for each additional part per million of calcium in drinking water.

4. Mutual funds.

a) **Paired data assumption:** These data are paired by mutual fund.
Randomization condition: Assume that these funds are representative of all large cap mutual funds.
10% condition: 15 mutual funds are less than 10% of all large cap mutual funds.
Nearly Normal condition: The histogram of differences is unimodal and symmetric.

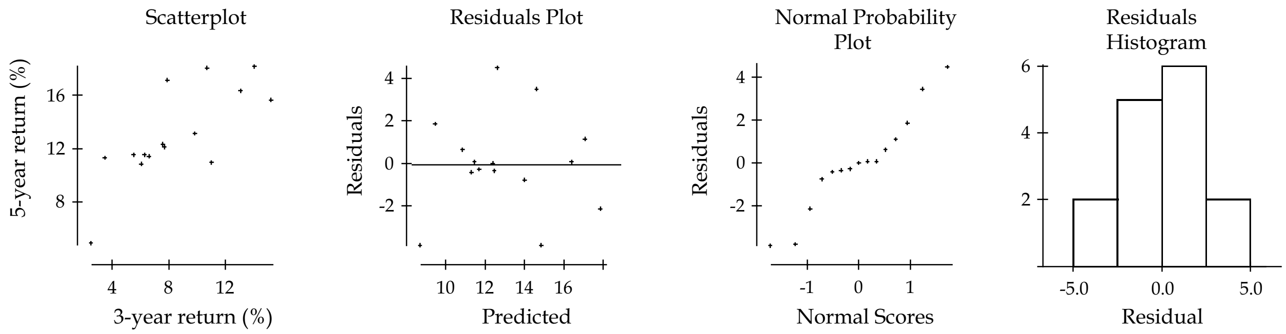


Since the conditions are satisfied, the sampling distribution of the difference can be modeled with a Student's *t*-model with 15 – 1 = 14 degrees of freedom. We will find a paired *t*-interval, with 95% confidence.

$$\bar{d} \pm t^*_{n-1}\left(\frac{s_d}{\sqrt{n}}\right) = 4.54 \pm t^*_{14}\left(\frac{2.50508}{\sqrt{15}}\right) \approx (3.15, 5.93)$$

Provided that these mutual funds are representative of all large cap mutual funds, we are 95% confident that, on average, 5-year yields are between 3.15% and 5.93% higher than 3-year yields.

b) H0: There is no linear relationship between 3-year and 5-year rates of return. $(\beta_1 = 0)$
HA: There is a linear relationship between 3-year and 5-year rates of return. $(\beta_1 \neq 0)$



Straight enough condition: The scatterplot is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The Normal probability plot of residuals isn't very straight. However, the histogram of residuals is unimodal and symmetric. With a sample size of 15, it is probably okay to proceed.

Since the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (15 – 2) = 13 degrees of freedom. We will use a regression slope *t*-test.

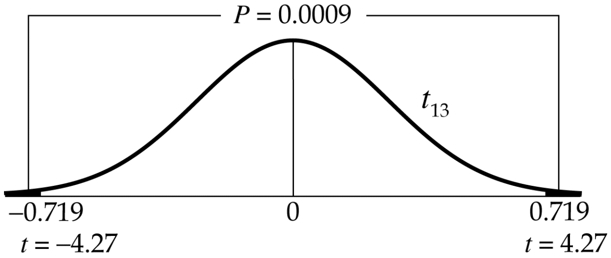
```
Dependent variable is:  5-year
No Selector
R squared = 58.4%     R squared (adjusted) = 55.2%
s =  2.360  with  15 - 2 = 13  degrees of freedom

Source        Sum of Squares    df    Mean Square    F-ratio
Regression    101.477            1     101.477        18.2
Residual      72.3804           13     5.56773

Variable    Coefficient   s.e. of Coeff   t-ratio    prob
Constant    6.92904       1.557           4.45       0.0007
3-year      0.719157      0.1685          4.27       0.0009
```

The equation of the line of best fit for these data points is: $(5\hat{year}) = 6.92904 + 0.719157(3year)$.

The value of $t \approx 4.27$. The *P*-value of 0.0009 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between the rates of return for 3-year and 5-year periods. Provided that these mutual funds are representative of all large cap mutual funds, mutual funds with higher 3-year returns tend to have higher 5-year returns.



5. Resume fraud.

In order to estimate the true percentage of people have misrepresented their backgrounds to within ± 5%, the company would have to perform about 406 random checks.

$$ME = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.05 = 2.326\sqrt{\frac{(0.25)(0.75)}{n}}$$

$$n = \frac{(2.326)^2(0.25)(0.75)}{(0.05)^2}$$

$$n \approx \text{ 406 random checks}$$

6. Paper airplanes.

a) It is reasonable to think that the flight distances are independent of one another. The histogram of flight distances (given) is unimodal and symmetric. Since the conditions are satisfied, the sampling distribution of the mean can be modeled by a Student's *t* model, with 11 – 1 = 10 degrees of freedom. We will use a one-sample *t* -interval with 95% confidence for the mean flight distance.

$$\bar{y} \pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right) = 48.3636 \pm t^*_{10}\left(\frac{18.0846}{\sqrt{11}}\right) \approx (36.21,\ 60.51)$$

We are 95% confident that the mean distance the airplane may fly is between 36.21 and 60.51 feet.

b) Since 40 feet is contained within our 95% confidence interval, it is plausible that the mean distance is 40 feet.

c) A 99% confidence interval would be wider. Intervals with greater confidence are less precise.

d) In order to cut the margin of error in half, she would need a sample size four times as large, or 44 flights.

7. Back to Montana.

H_0: Political party is independent of income level in Montana.

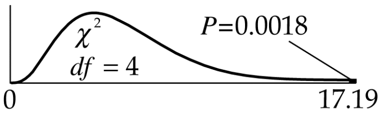
H_A: There is an association between political party and income level in Montana.

Counted data condition: The data are counts.
Randomization condition: Although not specifically stated, we will assume that the poll was conducted randomly.
Expected cell frequency condition: The expected counts are all greater than 5.

	Democrat ***(Obs/Exp)***	**Republican** ***(Obs/Exp)***	**Independent** ***(Obs/Exp)***
Low	30 / 24.119	16 / 22.396	12 / 14.485
Middle	28 / 30.772	24 / 28.574	22 / 14.653
High	26 / 29.109	38 / 27.03	6 / 13.861

Under these conditions, the sampling distribution of the test statistic is χ^2 on 4 degrees of freedom. We will use a chi-square test for independence.

$$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp} \approx 17.19$$



The *P*-value ≈ 0.0018

Since the *P*-value ≈ 0.0018 is low, we reject the null hypothesis. There is strong evidence of an association between income level and political party in Montana. An examination of the components shows that Democrats are more likely to have low incomes, Independents are more likely to have middle incomes, and Republicans are more likely to have high incomes.

8. Wild horses.

a) Since there are 36 degrees of freedom, 38 herds of wild horses were studied.

b) **Straight enough condition:** The scatterplot is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The histogram of residuals is unimodal and symmetric.

c) Since the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (38 – 2) = 36 degrees of freedom. We will use a regression slope *t*-interval, with 95% confidence. Use $t^*_{35} \approx 2.030$ as an estimate.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = 0.153969 \pm (2.030) \times 0.0114 \approx (0.131, 0.177)$$

d) We are 95% confident that the mean number of foals in a herd increases by between 0.131 and 0.177 foals for each additional adult horse.

e) The regression equation predicts that herds with 80 adults will have $-1.57835 + 0.153969(80) = 10.73917$ foals. The average size of the herds sampled is 110.237 adult horses. Use $t^*_{36} \approx 1.6883$, or use an estimate of $t^*_{35} \approx 1.690$, from the table.

$$\hat{y}_v \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$$

$$= 10.73917 \pm (1.6883)\sqrt{0.0114^2 \cdot (80 - 110.237)^2 + \frac{4.941^2}{38} + 4.941^2}$$

$$\approx (2.26, 19.21)$$

We are 95% confident that number of foals in a herd of 80 adult horses will be between 2.26 and 19.21. This prediction interval is too wide to be of much use.

9. Lefties and music.

H$_0$: The proportion of right-handed people who can match the tone is the same as the proportion of left-handed people who can match the tone. $(p_L = p_R \text{ or } p_L - p_R = 0)$

H$_A$: The proportion of right-handed people who can match the tone is different from the proportion of left-handed people who can match the tone. $(p_L \neq p_R \text{ or } p_L - p_R \neq 0)$

Random condition: Assume that the people tested are representative of all people.
10% condition: 76 and 53 are both less than 10% of all people.
Independent samples condition: The groups are not associated.
Success/Failure condition: $n\hat{p}$(right) = 38, $n\hat{q}$(right) = 38, $n\hat{p}$(left) = 33, and $n\hat{q}$(left) = 20 are all greater than 10, so the samples are both large enough.

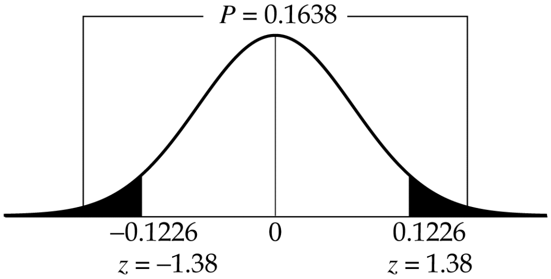
Since the conditions have been satisfied, we will model the sampling distribution of the difference in proportion with a Normal model with mean 0 and standard deviation estimated by $SE_{\text{pooled}}(\hat{p}_L - \hat{p}_R) = \sqrt{\frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_L} + \frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_R}} = \sqrt{\frac{\left(\frac{71}{129}\right)\left(\frac{58}{129}\right)}{53} + \frac{\left(\frac{71}{129}\right)\left(\frac{58}{129}\right)}{76}} \approx 0.089.$

The observed difference between the proportions is: 0.6226 – 0.5 = 0.1226.

Since the *P*-value = 0.1683 is high, we fail to reject the null hypothesis. There is no evidence that the proportion of people able to match the tone differs between right-handed and left-handed people.

$$z = \frac{0.1226 - 0}{0.089}$$

$$z \approx 1.38$$



10. AP Statistics scores.

a) H$_0$: The distribution of AP Statistics scores at Ithaca High School is the same as it is nationally.
H$_A$: The distribution of AP Statistics scores at Ithaca High School is different than it is nationally.

Counted data condition: The data are counts.
Randomization condition: Assume that this group of students is representative of all years at Ithaca High School.
Expected cell frequency condition: The expected counts (shown in the table) are all greater than 5.

Under these conditions, the sampling distribution of the test statistic is χ^2on 5 – 1 = 4 degrees of freedom. We will use a chi-square goodness-of-fit test.

Score	Observed	Expected	Residual = $(Obs - Exp)$	Standardized Residual = $\frac{(Obs - Exp)}{\sqrt{Exp}}$	Component = $\frac{(Obs - Exp)^2}{Exp}$
5	26	11.155	14.845	4.445	19.756
4	36	22.698	13.302	2.792	7.7955
3	19	24.153	– 5.153	– 1.049	1.0994
2	10	18.527	– 8.527	– 1.981	3.9245
1	6	20.467	– 14.47	– 3.198	10.226

$\sum \approx 42.801$

$\chi^2 \approx 42.801$. Since the *P*-value is essentially 0, we reject the null hypothesis.

There is strong evidence that the distribution of scores at Ithaca High School is different than the national distribution. Students at IHS get fewer scores of 2 and 1 than expected, and more scores of 4 and 5 than expected.

b) H$_0$: Gender and AP Statistics score are independent at Ithaca High School.

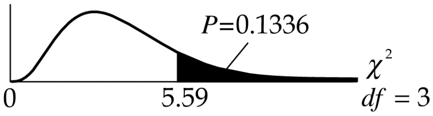
H$_A$: There is an association between gender and AP Statistics score at Ithaca High School.

Counted data condition: The data are counts.
Randomization condition: Assume this year's students are representative of all years.
Expected cell frequency condition: After combining the cells for scores of 2 and 1, the expected counts are all greater than 5.

	Boys ***(Obs/Exp)***	**Girls** ***(Obs/Exp)***
5	13 / 13.67	13 / 12.33
4	21 / 18.928	15 / 17.072
3	6 / 9.9897	13 / 9.0103
2 or 1	11 / 8.4124	5 / 7.5876

Under these conditions, the sampling distribution of the test statistic is χ^2 on 3 degrees of freedom. We will use a chi-square test for independence. (This is a test for independence, since we have one group that has been classified according to two variables, gender and score. However, if you said it was a test for homogeneity, since you were comparing two groups, no one would get terribly upset!)

With $\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp} \approx 5.59$, the *P*-value ≈ 0.1336.



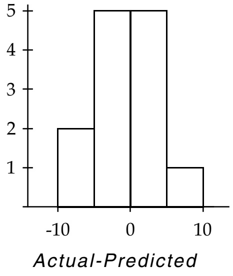
Since *P*-value ≈ 0.1336 is high, we fail to reject the null hypothesis. There is no evidence of an association between gender and score at Ithaca High School. The boys seem to do just as well as the girls.

11. Polling.

a) H$_0$: The mean difference in the number of predicted Democrats and the number of actual Democrats is zero. $(\mu_d = 0)$

H$_A$: The mean difference in the number of predicted Democrats and the number of actual Democrats is different than zero. $(\mu_d \neq 0)$

Paired data assumption: The data are paired by year.
Randomization condition: Assume these predictions are representative of other predictions.
10% condition: We are testing the predictions, not the years.
Nearly Normal condition: The histogram of differences between the predicted number of Democrats and the actual number of Democrats is roughly unimodal and symmetric. The year 1958 is an outlier, and was removed.



Since the conditions are satisfied, the sampling distribution of the difference can be modeled with a Student's *t*-model with 13 – 1 = 12 degrees of freedom, $t_{12}\left(0, \frac{3.57878}{\sqrt{13}}\right)$.

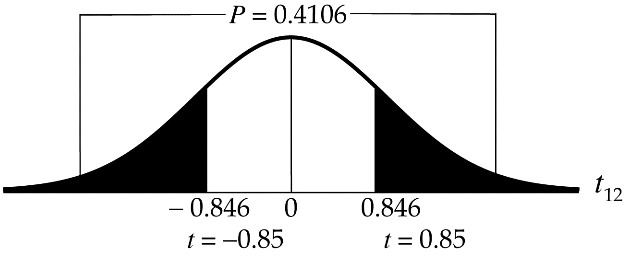
We will use a paired *t*-test, with $\bar{d} = -0.846$.

Since the *P*-value = 0.4106 is high, we fail to reject the null hypothesis. There is no evidence that the mean difference between the actual and predicted number of Democrats was anything other than 0.

$$t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

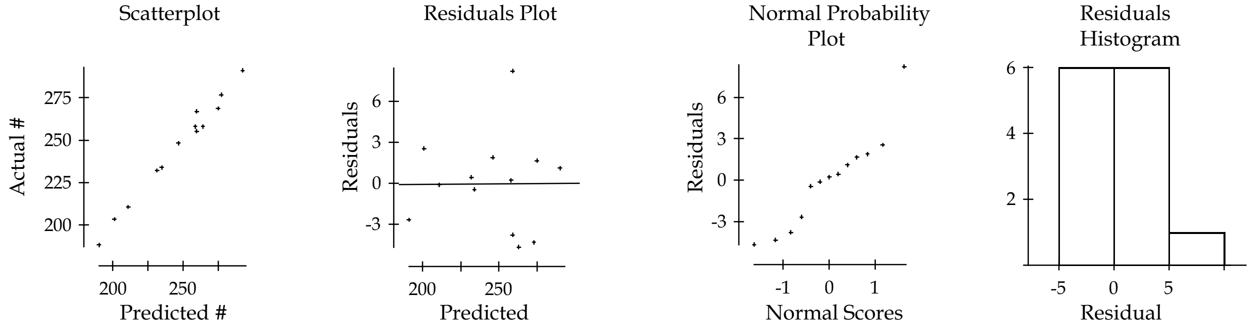
$$t = \frac{-0.846 - 0}{\frac{3.57878}{\sqrt{13}}}$$

$$t \approx -0.85$$



b) H_0: There is no linear relationship between Gallup's predictions and the actual number of Democrats. $(\beta_1 = 0)$

H_A: There is a linear relationship between Gallup's predictions and the actual number of Democrats. $(\beta_1 \neq 0)$



Straight enough condition: The scatterplot is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: After an outlier in 1958 is removed, the Normal probability plot of residuals still isn't very straight. However, the histogram of residuals is roughly unimodal and symmetric. With a sample size of 13, it is probably okay to proceed.

Since the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (13 – 2) = 11 degrees of freedom. We will use a regression slope *t*-test.

```
Dependent variable is:  Actual
No Selector
R squared = 98.7%     R squared (adjusted) = 98.6%
s =  3.628  with  13 - 2 = 11  degrees of freedom

Source         Sum of Squares    df    Mean Square    F-ratio
Regression     10874.4            1       10874.4       826
Residual       144.805           11       13.1641

Variable     Coefficient    s.e. of Coeff    t-ratio      prob
Constant     6.00180        8.395             0.715      0.4895
Predicted    0.972206       0.0338           28.7       ≤ 0.0001
```

The equation of the line of best fit for these data points is: $\widehat{Actual} = 6.00180 + 0.972206(Predicted)$.

The value of $t \approx 28.7$. The *P*-value of essentially 0 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between the number of Democrats predicted by Gallup and the number of Democrats actually in the House of Representatives. Years in which the predicted number was high tend to have high actual numbers also. The high value of $R^2 = 98.7\%$ indicates a very strong model. Gallup polls seem very accurate.

12. Twins.

H_0: There is no association between duration of pregnancy and level of prenatal care.

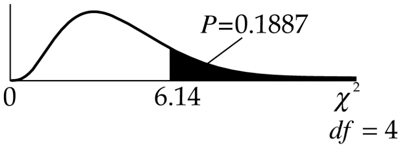
H_A: There is an association between duration of pregnancy and level of prenatal care.

Counted data condition: The data are counts.
Randomization condition: Assume that these pregnancies are representative of all twin births.
Expected cell frequency condition: The expected counts are all greater than 5.

	Preterm (induced or Cesarean) ***(Obs/Exp)***	**Preterm (without procedures)** ***(Obs/Exp)***	**Term or postterm** ***(Obs/Exp)***
Intensive	18 /16.676	15 / 15.579	28 / 28.745
Adequate	46 / 42.101	43 / 39.331	65 / 72.568
Inadequate	12 / 17.223	13 / 16.090	38 / 29.687

Under these conditions, the sampling distribution of the test statistic is χ^2on 4 degrees of freedom. We will use a chi-square test for independence.

$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp} \approx 6.14$, and the P-value ≈ 0.1887.



Since the P-value ≈ 0.1887 is high, we fail to reject the null hypothesis. There is no evidence of an association between duration of pregnancy and level of prenatal care in twin births.

13. Twins, again.

H0: The distributions of pregnancy durations are the same for the three years.

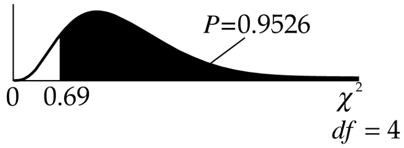
HA: The distributions of pregnancy durations are different for the three years.

Counted data condition: The data are counts.
Independence assumption: Assume that the durations of the pregnancies are mutually indpendent.
Expected cell frequency condition: The expected counts are all greater than 5.

	1990 ***(Obs/Exp)***	**1995** ***(Obs/Exp)***	**2000** ***(Obs/Exp)***
Preterm (induced or Cesarean)	11 /12.676	13 / 13.173	19 / 17.150
Preterm (without procedures)	13 / 13.266	14 / 13.786	18 / 17.948
Term or postterm	27 / 25.058	26 / 26.04	32 / 33.902

Under these conditions, the sampling distribution of the test statistic is χ^2on 4 degrees of freedom. We will use a chi-square test for homogeneity.

$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp} \approx 0.69$, and the P-value ≈ 0.9526.



Since the P-value ≈ 0.9526 is high, we fail to reject the null hypothesis. There is no evidence that the distributions of the durations of pregnancies are different for the three years. It does not appear that they way the hospital deals with twin pregnancies has changed.

14. Preemies.

H0: The proportion of "preemies" who are of "subnormal height" as adults is the same as the proportion of normal birth weight babies who are. $\left(p_P = p_N \text{ or } p_P - p_N = 0\right)$

HA : The proportion of "preemies" who are of "subnormal height" as adults is greater than the proportion of normal birth weight babies who are. $\left(p_P > p_N \text{ or } p_P - p_N > 0\right)$

Random condition: Assume that these children are representative of all children.
10% condition: 242 and 233 are both less than 10% of all children.
Independent samples condition: The groups are not associated.
Success/Failure condition: $n\hat{p}$(preemies) = 24, $n\hat{q}$(preemies) = 218, $n\hat{p}$(normal) = 12, and $n\hat{q}$(normal) = 221 are all greater than 10, so the samples are both large enough.

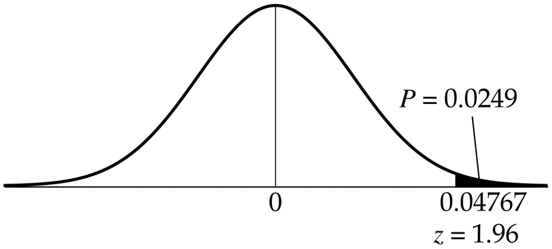
Since the conditions have been satisfied, we will model the sampling distribution of the difference in proportion with a Normal model with mean 0 and standard deviation estimated by $SE_{\text{pooled}}\left(\hat{p}_P - \hat{p}_N\right) = \sqrt{\dfrac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_P} + \dfrac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_N}} = \sqrt{\dfrac{\left(\frac{36}{475}\right)\left(\frac{439}{475}\right)}{242} + \dfrac{\left(\frac{36}{475}\right)\left(\frac{439}{475}\right)}{233}} \approx 0.02429.$

The observed difference between the proportions is:
0.09917 – 0.05150 = 0.04767.

Since the *P*-value = 0.0249 is low, we reject the null hypothesis. There is moderate evidence that "preemies" are more likely to be of "subnormal height" as adults than children of normal birth weight.

$$z = \frac{0.04767 - 0}{0.02429}$$

$$z \approx 1.96$$



15. LA rainfall.

a) **Independence assumption:** Annual rainfall is independent from year to year.
Nearly Normal condition: The histogram of the rainfall totals is skewed to the right, but the sample is fairly large, so it is safe to proceed.

The mean annual rainfall is 14.5165 inches, with a standard deviation 7.82044 inches. Since the conditions have been satisfied, construct a one-sample t-interval, with 22 – 1 = 21 degrees of freedom, at 90% confidence.

$$\bar{y} \pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right) = 14.5164 \pm t^*_{21}\left(\frac{7.82044}{\sqrt{22}}\right) \approx (11.65, 17.39)$$

We are 90% confident that the mean annual rainfall in LA is between 11.65 and 17.39 inches.

b) Start by making an estimate, either using $z^* = 1.645$ or $t^*_{21} = 1.721$ from above. Either way, your estimate is around 40 people. Make a better estimate using $t^*_{40} = 1.684$. You would need about 44 years' data to estimate the annual rainfall in LA to within 2 inches.

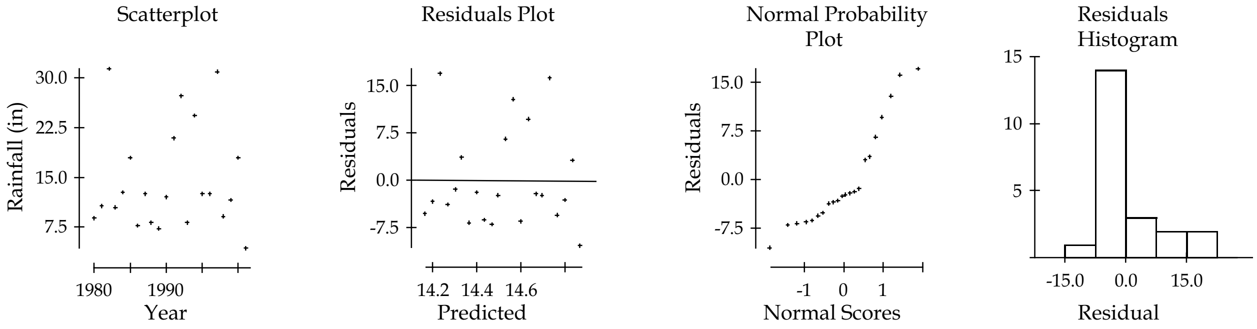
$$ME = t^*_{40}\left(\frac{s}{\sqrt{n}}\right)$$

$$2 = 1.684\left(\frac{7.82044}{\sqrt{n}}\right)$$

$$n = \frac{(2.064)^2(7.82044)^2}{(2)^2}$$

$$n \approx 44 \text{ years}$$

c) H_0: There is no linear relationship between year and annual LA rainfall. $(\beta_1 = 0)$
H_A: There is a linear relationship between year and annual LA rainfall. $(\beta_1 \neq 0)$



Straight enough condition: The scatterplot is straight enough to try linear regression, although there is no apparent pattern.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The Normal probability plot of residuals is not straight, and the histogram of the residuals is skewed to the right, but a sample of 22 years is large enough to proceed.

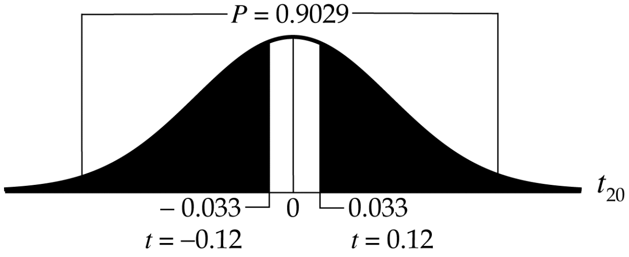
Since the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (22 – 2) = 20 degrees of freedom. We will use a regression slope *t*-test.

```
Dependent variable is:  Rain  (in.)
No Selector
R squared = 0.1%     R squared (adjusted) = -4.9%
s =  8.011   with  22 - 2 = 20   degrees of freedom

Source         Sum of Squares    df    Mean Square    F-ratio
Regression     0.979449           1     0.979449      0.015
Residual       1283.37           20      64.1684

Variable     Coefficient    s.e. of Coeff    t-ratio     prob
Constant     -51.6838       535.8            -0.096     0.9241
Year          0.033258        0.2692          0.124     0.9029
```

The equation of the line of best fit for these data points is: $\widehat{Rain} = -51.6838 + 0.033258(Year)$.



The value of $t \approx 0.124$. The *P*-value of 0.92029 means that the association we see in the data is quite likely to occur by chance. We fail to reject the null hypothesis, and conclude that there is no evidence of a linear relationship between the annual rainfall in LA and the year.

16. Age and party.

a) There is one sample, classified according to two different variables, so we will perform a chi-square test for independence.

b) H_0: There is no association between age and political party.

H_A: There is an association between age and political party.

Counted data condition: The data are counts.
Randomization condition: These data are from a representative phone survey.
Expected cell frequency condition: The expected counts are all greater than 5.

	Republican **(*Obs*/*Exp*)**	**Democrat** **(*Obs*/*Exp*)**	**Independent** **(*Obs*/*Exp*)**
18 – 29	241 / 275.39	351 /351.18	409 / 374.44
30 – 49	299 / 274.84	330 / 350.47	370 / 373.69
50 – 64	282 / 274.56	341 / 350.12	375 / 373.31
65 +	279 / 276.21	382 / 352.23	343 / 375.56

Under these conditions, the sampling distribution of the test statistic is χ^2on 6 degrees of freedom. We will use a chi-square test for independence.

$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp} \approx 16.66$, and the *P*-value ≈ 0.0106.

c) Since the *P*-value ≈ 0.0106 is low, we reject the null hypothesis. There is strong evidence of an association between age and political party.

The table of standardized residuals is useful for the analysis of the differences. Looking for the largest standardized residuals, we can see there are fewer Republicans and more Independents among those 18 – 29 than we expect, and fewer Independents than we expect among those 65 and older.

	Standardized Residuals		
	Republican	**Democrat**	**Independent**
18 – 29	–2.07	–0.01	1.79
30 – 49	1.46	–1.09	–0.19
50 – 64	0.45	–0.49	0.10
65 +	0.17	1.59	–1.68

17. Eye and hair color.

a) This is an attempt at linear regression. Regression inference is meaningless here, since eye and hair color are categorical variables.

b) This is an analysis based upon a chi-square test for independence.

H_0: Eye color and hair color are independent.

H_A: There is an association between eye color and hair color.

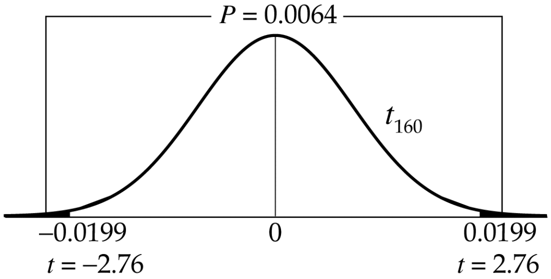
Since we have two categorical variables, this analysis seems appropriate. However, if you check the expected counts, you will find that 4 of them are less than 5. We would have to combine several cells in order to perform the analysis. (Always check the conditions!)

Since the value of chi-square is so high, it is likely that we would find an association between eye and hair color, even after the cells were combined. There are many cells of interest, but some of the most striking differences that would not be affected by cell combination involve people with fair hair. Blonds are likely to have blue eyes, and not likely to have brown eyes. Those with red hair are not likely to have brown eyes. Additionally, those with black hair are much more likely to have brown eyes than blue.

18. Depression and the Internet.

a) H0: There is no linear relationship between depression and Internet usage. $(\beta_1 = 0)$
HA: There is a linear relationship between depression and Internet usage. $(\beta_1 \neq 0)$

Since the conditions for inference are satisfied (given), the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (162 – 2) = 160 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is: $\widehat{DepressionAfter} = 0.565485 + 0.019948(InternetUsage)$.



The value of $t \approx 2.76$. The *P*-value of 0.0064 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between depression and Internet usage. Those with high levels of Internet usage tend to have high levels of depression. It should be noted, however, that although the evidence is strong, the association is quite weak, with $R^2 = 4.6\%$. The regression analysis only explains 4.6% of the variation in depression level.

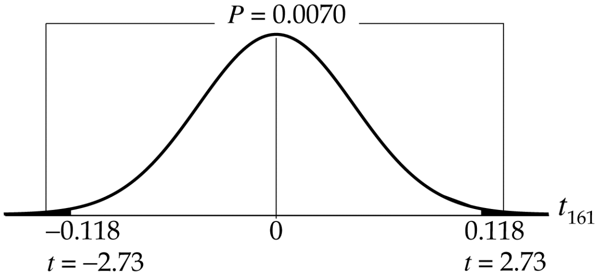
b) The study says nothing about causality, merely association. Furthermore, there are almost certainly other factors involved. In fact, if 4.6% of the variation in depression level is related to Internet usage, the other 95.4% of the variation must be related to something else!

c) H0: The mean difference in depression before and after the experiment is zero. $(\mu_d = 0)$

HA: The mean difference in depression before and after the experiment is different than zero. $(\mu_d \neq 0)$

Since the conditions are satisfied (given), the sampling distribution of the difference can be modeled with a Student's *t*-model with 162 – 1 = 161 degrees of freedom, $t_{161}\left(0, \frac{0.552417}{\sqrt{162}}\right)$.

We will use a paired *t*-test, with $\bar{d} = -0.118457$.



Since the *P*-value = 0.0070 is very low, we reject the null hypothesis. There is strong evidence that the mean depression level changed over the course of the experiment. These data suggest that depression levels actually decreased.

19. Pregnancy.

a) H0: The proportion of live births is the same for women under the age of 38 as it is for women over the age of 38. $(p_{<38} = p_{\geq 38} \text{ or } p_{<38} - p_{\geq 38} = 0)$

HA : The proportion of live births is different for women under the age of 38 than for women over the age of 38. $(p_{<38} \neq p_{\geq 38} \text{ or } p_{<38} - p_{\geq 38} \neq 0)$

Random condition: Assume that the women studied are representative of all women.
10% condition: 157 and 89 are both less than 10% of all women.
Independent samples condition: The groups are not associated.
Success/Failure condition: $n\hat{p}$(under 38) = 42, $n\hat{q}$(under 38) = 115, $n\hat{p}$(38 and over) = 7, and $n\hat{q}$(38 and over) = 82 are not all greater than 10, since the observed number of live births is only 7. However, if we check the pooled value, $n\hat{p}_{\text{pooled}}$(38 and over) = (89)(0.191) = 17. All of the samples are large enough.

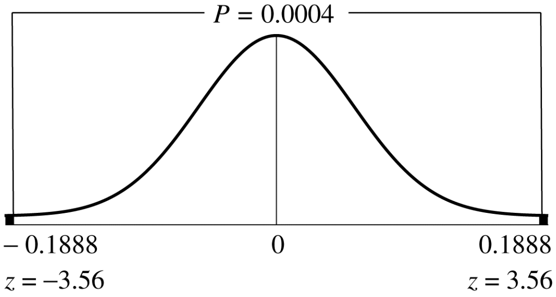
Since the conditions have been satisfied, we will model the sampling distribution of the difference in proportion with a Normal model with mean 0 and standard deviation estimated by $SE_{\text{pooled}}\left(\hat{p}_{<38}-\hat{p}_{\geq 38}\right)=\sqrt{\frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_{<38}}+\frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_{\geq 38}}}=\sqrt{\frac{\left(\frac{49}{246}\right)\left(\frac{197}{246}\right)}{157}+\frac{\left(\frac{49}{246}\right)\left(\frac{197}{246}\right)}{89}}\approx 0.0530.$

The observed difference between the proportions is:
0.2675 – 0.0787 = 0.1888.

Since the *P*-value = 0.0004 is low, we reject the null hypothesis. There is strong evidence to suggest a difference in the proportion of live births for women under 38 and women 38 and over at this clinic. In fact, the evidence suggests that women under 38 have a higher proportion of live births.

$z=\frac{0.1888-0}{0.0530}$

$z\approx 3.56$



b) H₀: Age and birth rate are independent.

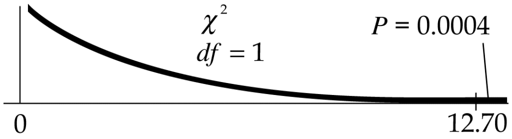
Hᴀ: There is an association between age and birth rate

Counted data condition: The data are counts.
Randomization condition: Assume that these women are representative of all women.
Expected cell frequency condition: The expected counts are all greater than 5.

	Live birth (*Obs/Exp*)	**No live birth (*Obs/Exp*)**
Under 38	42 / 31.272	115 / 125.73
38 and over	7 / 17.728	82 / 71.27

Under these conditions, the sampling distribution of the test statistic is χ^2 on 1 degree of freedom. We will use a chi-square test for independence.

$\chi^2=\sum_{\text{all cells}}\frac{(Obs-Exp)^2}{Exp}\approx 12.70$, and the *P*-value ≈ 0.0004.



Since the *P*-value ≈ 0.0004 is low, we reject the null hypothesis. There is strong evidence of an association between age and birth rate. Younger mothers tend to have higher birth rates.

c) A two-proportion z-test and a chi-square test for independence with 1 degree of freedom are equivalent. $z^2 = (3.563944)^2 = 12.70 = \chi^2$. The P-values are both the same.

20. Eating in front of the TV.

a) Displays may vary. Stacked bar charts or comparative pie charts are appropriate.

b) H_0: Response to the statement is independent of age.

H_A: There is an association between the response to the statement and age.

Counted data condition: The data are counts.
Randomization condition: These data are from a random sample.
Expected cell frequency condition: All of the expected counts are much greater than 5.

Under these conditions, the sampling distribution of the test statistic is χ^2 on 8 degrees of freedom. We will use a chi-square test for independence.

With $\chi^2 = \sum_{all\,cells} \frac{(Obs - Exp)^2}{Exp} \approx 190.96$, the P-value < 0.001.

Since the P-value is so small, reject the null hypothesis. There is strong evidence to suggest that response is not independent of age.

21. Old Faithful.

a) There is a moderate, linear, positive association between duration of the previous eruption and interval between eruptions for Old Faithful. Relatively long eruptions appear to be associated with relatively long intervals until the next eruption.

b) H_0: There is no linear relationship between duration of the eruption and interval until the next eruption. $(\beta_1 = 0)$
H_A: There is a linear relationship between duration of the eruption and interval until the next eruption. $(\beta_1 \neq 0)$

c) **Straight enough condition:** The scatterplot is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The histogram of residuals is unimodal and symmetric.

Since the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's t-model with (222 – 2) = 220 degrees of freedom. We will use a regression slope t-test. The equation of the line of best fit for these data points is: $\widehat{Interval} = 33.9668 + 10.3582(Duration)$.

d) The value of $t \approx 27.1$. The P-value of essentially 0 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between duration and interval. Relatively long eruptions tend to be associated with relatively long intervals until the next eruption.

e) The regression equation predicts that an eruption with duration of 2 minutes will have an interval until the next eruption of $33.9668 + 10.3582(2) = 54.6832$ minutes. ($t^*_{220} \approx 1.9708$)

$$\begin{aligned}&\hat{y}_v \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_v - \bar{x})^2 + \frac{s_e^2}{n}}\\&= 54.6832 \pm (1.9708)\sqrt{0.3822^2\cdot(2 - 3.57613)^2 + \frac{6.159^2}{222}}\\&\approx (53.24,\ 56.12)\end{aligned}$$

We are 95% confident that, after a 2-minute eruption, the mean length of time until the next eruption will be between 53.24 and 56.12 minutes.

f) The regression equation predicts that an eruption with duration of 4 minutes will have an interval until the next eruption of $33.9668 + 10.3582(4) = 75.3996$ minutes. ($t^*_{220} \approx 1.9708$)

$$\begin{aligned}&\hat{y}_v \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}\\&= 75.3996 \pm (1.9708)\sqrt{0.3822^2\cdot(4 - 3.57613)^2 + \frac{6.159^2}{222} + 6.159^2}\\&\approx (63.23,\ 87.57)\end{aligned}$$

We are 95%confident that the length of time until the next eruption will be between 63.23 and 87.57 minutes, following a 4-minute eruption.

22. Togetherness.

a) H$_0$: There is no linear relationship number of meals eaten as a family and grades. $(\beta_1 = 0)$
H$_A$: There is a linear relationship number of meals eaten as a family and grades. $(\beta_1 \neq 0)$

Since the conditions for inference are satisfied (given), the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (142 – 2) = 140 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is: $G\hat{P}A = 2.7288 + 0.1093(Meals\ /\ Week)$.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{0.1093 - 0}{0.0263}$$

$$t \approx 4.16$$

The value of $t \approx 4.16$. The *P*-value of less than 0.0001 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between grades and the number of meals eaten as a family. Students whose families eat together relatively frequently tend to have higher grades than those whose families don't eat together as frequently.

b) This relationship would not be particularly useful for predicting a student's grade point average. $R^2 = 11.0\%$, which means that only 11% of the variation in GPA can be explained by the number of meals eaten together per week.

c) These conclusions are not contradictory. There is strong evidence that the slope is not zero, and that means strong evidence of a linear relationship. This does not mean that the relationship itself is strong, or useful for predictions.

23. Learning math.

a) H0: The mean score of Accelerated Math students is the same as the mean score of traditional students. $\left(\mu_A = \mu_T \text{ or } \mu_A - \mu_T = 0\right)$

HA: The mean score of Accelerated Math students is different from the mean score of traditional students. $\left(\mu_A \neq \mu_T \text{ or } \mu_A - \mu_T \neq 0\right)$

Independent groups assumption: Scores of students from different classes should be independent.
Randomization condition: Although not specifically stated, classes in this experiment were probably randomly assigned to learn either Accelerated Math or traditional curricula.
10% condition: 231 and 245 are less than 10% of all students.
Nearly Normal condition: We don't have the actual data, so we can't check the distribution of the sample. However, the samples are large. The Central Limit Theorem allows us to proceed.

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 459.24 degrees of freedom (from the approximation formula).

We will perform a two-sample *t*-test. The sampling distribution model has mean 0, with standard error: $SE(\bar{y}_A - \bar{y}_T) = \sqrt{\dfrac{84.29^2}{231} + \dfrac{74.68^2}{245}} \approx 7.3158$.

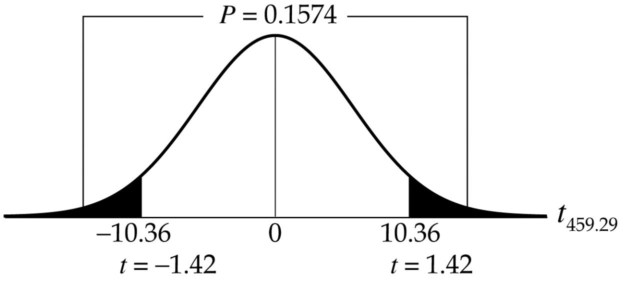
The observed difference between the mean scores is 560.01 – 549.65 = 10.36

Since the *P*-value = 0.1574, we fail to reject the null hypothesis. There is no evidence that the Accelerated Math students have a different mean score on the pretest than the traditional students.

$$t = \frac{(\bar{y}_A - \bar{y}_T) - (0)}{SE(\bar{y}_A - \bar{y}_T)}$$

$$t \approx \frac{10.36}{7.3158}$$

$$t \approx 1.42$$



b) H0: Accelerated Math students do not show significant improvement in test scores. The mean individual gain for Accelerated Math is zero. $\left(\mu_d = 0\right)$

HA: Accelerated Math students show significant improvement in test scores. The mean individual gain for Accelerated Math is greater than zero. $\left(\mu_d > 0\right)$

Paired data assumption: The data are paired by student.
Randomization condition: Although not specifically stated, classes in this experiment were probably randomly assigned to learn either Accelerated Math or traditional curricula.
10% condition: We are testing the Accelerated Math program, not the students.

Nearly Normal condition: We don't have the actual data, so we cannot look at a graphical display, but since the sample is large, it is safe to proceed.

The Accelerated Math students had a mean individual gain of $\bar{d} = 77.53$ points and a standard deviation of 78.01 points. Since the conditions for inference are satisfied, we can model the sampling distribution of the mean individual gain with a Student's *t* model, with 231 – 1 = 230 degrees of freedom, $t_{230}\left(0, \frac{78.01}{\sqrt{231}}\right)$. We will perform a paired *t*-test.

$$t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

$$t = \frac{77.53 - 0}{\frac{78.01}{\sqrt{231}}}$$

$$t \approx 15.11$$

Since the *P*-value is essentially 0, we reject the null hypothesis. There is strong evidence that the mean individual gain is greater than zero. The Accelerated Math students showed significant improvement.

c) H$_0$: Students taught using traditional methods do not show significant improvement in test scores. The mean individual gain for traditional methods is zero. $(\mu_d = 0)$

H$_A$: Students taught using traditional methods show significant improvement in test scores. The mean individual gain for traditional methods is greater than zero. $(\mu_d > 0)$

Paired data assumption: The data are paired by student.
Randomization condition: Although not specifically stated, classes in this experiment were probably randomly assigned to learn either Accelerated Math or traditional curricula.
10% condition: We are testing the program, not the students.
Nearly Normal condition: We don't have the actual data, so we cannot look at a graphical display, but since the sample is large, it is safe to proceed.

The students taught using traditional methods had a mean individual gain of $\bar{d} = 39.11$ points and a standard deviation of 66.25 points. Since the conditions for inference are satisfied, we can model the sampling distribution of the mean individual gain with a Student's *t* model, with 245 – 1 = 244 degrees of freedom, $t_{244}\left(0, \frac{66.25}{\sqrt{245}}\right)$.

We will perform a paired *t*-test.

Since the *P*-value is essentially 0, we reject the null hypothesis. There is strong evidence that the mean individual gain is greater than zero. The students taught using traditional methods showed significant improvement.

$$t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

$$t = \frac{39.11 - 0}{\frac{66.25}{\sqrt{245}}}$$

$$t \approx 9.24$$

d) H$_0$: The mean individual gain of Accelerated Math students is the same as the mean individual gain of traditional students. $(\mu_{dA} = \mu_{dT}$ or $\mu_{dA} - \mu_{dT} = 0)$

H$_A$: The mean individual gain of Accelerated Math students is greater than the mean individual gain of traditional students. $(\mu_{dA} > \mu_{dT}$ or $\mu_{dA} - \mu_{dT} > 0)$

Independent groups assumption: Individual gains of students from different classes should be independent.
Randomization condition: Although not specifically stated, classes in this experiment were probably randomly assigned to learn either Accelerated Math or traditional curricula.
10% condition: 231 and 245 are less than 10% of all students.
Nearly Normal condition: We don't have the actual data, so we can't check the distribution of the sample. However, the samples are large. The Central Limit Theorem allows us to proceed.

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 452.10 degrees of freedom (from the approximation formula).

We will perform a two-sample *t*-test. The sampling distribution model has mean 0, with standard error: $SE(\bar{d}_A - \bar{d}_T) = \sqrt{\frac{78.01^2}{231} + \frac{66.25^2}{245}} \approx 6.6527$.

The observed difference between the mean scores is 77.53 – 39.11 = 38.42

Since the *P*-value is less than 0.0001, we reject the null hypothesis. There is strong evidence that the Accelerated Math students have an individual gain that is significantly higher than the individual gain of the students taught using traditional methods.

$$t = \frac{(\bar{d}_A - \bar{d}_T) - 0}{SE(\bar{d}_A - \bar{d}_T)}$$

$$t = \frac{38.42 - 0}{6.6527}$$

$$t \approx 5.78$$

24. Pesticides.

H_0 : The percentage of males born to workers at the plant is 51.2%.(p = 0.512)
H_A : The percentage of males born to workers at the plant is less than 51.2%. (p < 0.512)

Independence assumption: It is reasonable to think that the births are independent.
Success/Failure Condition: np = (227)(0.512) = 116 and nq = (227)(0.488) = 111 are both greater than 10, so the sample is large enough.

The conditions have been satisfied, so a Normal model can be used to model the sampling distribution of the proportion, with $\mu_{\hat{p}} = p =$ 0.512 and $\sigma(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.512)(0.488)}{227}} \approx 0.0332$.

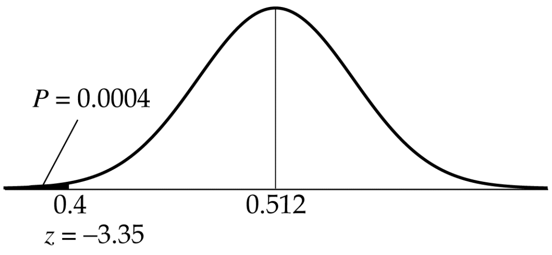
We can perform a one-proportion z-test. The observed proportion of males is $\hat{p} = 0.40$.

The value of $z \approx -3.35$, meaning that the observed proportion of males is over 3 standard deviations below the expected proportion. The *P*-value associated with this z score is approximately 0.0004.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{pq}{n}}}$$

$$z = \frac{0.4 - 0.512}{\sqrt{\frac{(0.512)(0.488)}{227}}}$$

$$z \approx -3.35$$



With a *P*-value this low, we reject the null hypothesis. There is strong evidence that the percentage of males born to workers is less than 51.2%. This provides evidence that human exposure to dioxin may result in the birth of more girls.

25. Dairy sales.

a) Since the CEO is interested in the association between cottage cheese sales and ice cream sales, the regression analysis is appropriate.

b) There is a moderate, linear, positive association between cottage cheese and ice cream sales. For each additional million pounds of cottage cheese sold, an average of 1.19 million pounds of ice cream are sold.

c) The regression will not help here. A paired *t*-test will tell us whether there is an average difference in sales.

d) There is evidence that the company sells more cottage cheese than ice cream, on average.

e) In part a, we are assuming that the relationship is linear, that errors are independent with constant variation, and that the distribution of errors is Normal.

In part c, we are assuming that the observations are independent and that the distribution of the differences is Normal. This may not be a valid assumption, since the histogram of differences looks bimodal.

f) The equation of the regression line is $\widehat{IceCream} = -26.5306 + 1.19334(CottageCheese)$. In a month in which 82 million pounds of ice cream are sold we expect to sell:

$\widehat{IceCream} = -26.5306 + 1.19334(82) = 71.32$ million pounds of ice cream.

g) Assuming the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (12 – 2) = 10 degrees of freedom. We will use a regression slope *t*-interval, with 95% confidence.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = 1.19334 \pm (2.228) \times 0.4936 \approx (0.09,\ 2.29)$$

h) We are 95% confident that the mean number of pounds of ice cream sold increases by between 0.09 and 2.29 pounds for each additional pound of cottage cheese sold.

26. Infliximab.

H_0: The remission rates are the same for the three groups.

H_A: The remission rates are different for the three groups.

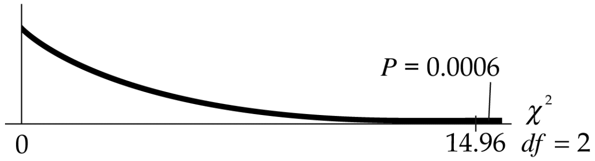
Counted data condition: The data are counts.
Randomization condition: Assume that these patients are representative of all patients.
Expected cell frequency condition: The expected counts are all greater than 5.

	Placebo ***(Obs/Exp)***	**5 mg** ***(Obs/Exp)***	**10 mg** ***(Obs/Exp)***
Remission	23 / 38.418	44 / 39.466	50 / 39.116
No Remission	87 / 71.582	69 / 73.534	62 / 72.884

Under these conditions, the sampling distribution of the test statistic is χ^2on 2 degrees of freedom. We will use a chi-square test for homogeneity.

$$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp} \approx 14.96,$$

and the P-value ≈ 0.0006.



Since the P-value ≈ 0.0006 is low, we reject the null hypothesis. There is strong evidence that the remission rates are different in the three groups. Patients receiving 10 mg of Infliximab have higher remission rates than the other groups. These data indicate that continued treatment with Infliximab is of value to Crohn's disease patients who exhibit a positive initial response to the drug.

27. Weight loss.

Randomization Condition: The respondents were randomly selected from among the clients of the weight loss clinic.
10% Condition: 20 people are less than 10% of all clients.
Nearly Normal Condition: The histogram of the number of pounds lost for each respondent is unimodal and symmetric, with no outliers.

The clients in the sample had a mean weight loss of 9.15 pounds, with a standard deviation of 1.94733 pounds. Since the conditions have been satisfied, construct a one-sample t-interval, with 20 – 1 = 19 degrees of freedom, at 95% confidence.

$$\bar{y} \pm t_{n-1}^{*}\left(\frac{s}{\sqrt{n}}\right) = 9.15 \pm t_{19}^{*}\left(\frac{1.94733}{\sqrt{20}}\right) \approx (8.24, 10.06)$$

We are 95% confident that the mean weight loss experienced by clients of this clinic is between 8.24 and 10.06 pounds. Since 10 pounds is contained within the interval, the claim that the program will allow clients to lose 10 pounds in a month is plausible. Answers may vary, depending on the chosen level of confidence.

28. Education vs. income.

a) **Straight enough condition:** The scatterplot is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The Normal probability plot is reasonably straight.

Since the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's t-model with (57 – 2) = 55 degrees of freedom. We will use a regression slope t-test. The equation of the line of best fit for these data points is: $\widehat{Income} = 5970.05 + 2444.79(Education)$.

b) The value of $t \approx 5.19$. The P-value of less than 0.0001 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between education level and income. Cities in which median education level is relatively high have relatively high median incomes.

c) If the data were plotted for individuals, the association would appear to be weaker. Individuals vary more than averages.

d) $b_1 \pm t^*_{n-2} \times SE(b_1) = 2444.79 \pm (2.004) \times 471.2 \approx (1500,\ 3389)$

We are 95% confident that each additional year of median education level in a city is associated with an increase of between \$1500 and \$3389 in median income.

e) The regression equation predicts that a city with a median education level of 11 years of school will have a median income of $5970.05 + 2444.79(11) = \$32862.74$ $(t^*_{55} \approx 1.6730)$

$$\hat{y}_v \pm t^*_{n-2}\sqrt{SE^2(b_1)\cdot(x_v - \bar{x})^2 + \frac{s_e^2}{n}}$$

$$= 32862.74 \pm (1.6730)\sqrt{471.2^2 \cdot (11 - 10.9509)^2 + \frac{2991^2}{57}}$$

$$\approx (32199,\ 33527)$$

We are 90% confident that cities with 11 years for median education level will have an average income of between \$32,199 and \$33,527.

29. Diet.

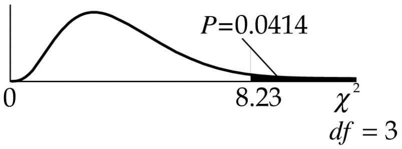
H_0: Cracker type and bloating are independent.

H_A: There is an association between cracker type and bloating.

Counted data condition: The data are counts.
Randomization condition: Assume that these women are representative of all women.
Expected cell frequency condition: The expected counts are all (almost!) greater than 5.

	Bloat	
	Little/None ***(Obs / Exp)***	**Moderate/Severe** ***(Obs / Exp)***
Bran	11 / 7.6471	2 / 5.3529
Gum Fiber	4 / 7.6471	9 / 5.3529
Combination	7 / 7.6471	6 / 5.3529
Control	8 / 7.0588	4 / 4.9412

Under these conditions, the sampling distribution of the test statistic is χ^2 on 3 degrees of freedom. We will use a chi-square test for independence.



$$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp} \approx 8.23, \text{ and the } P\text{-value} \approx 0.0414.$$

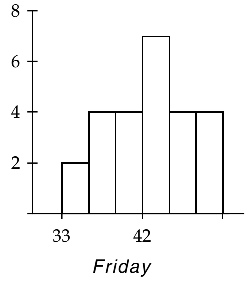
Since the *P*-value is low, we reject the null hypothesis. There is evidence of an association between cracker type and bloating. The gum fiber crackers had a higher rate of moderate/severe bloating than expected. The company should head back to research and development and address the problem before attempting to market the crackers.

30. Cramming.

a) H0: The mean score of week-long study group students is the same as the mean score of overnight cramming students. $(\mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0)$

HA: The mean score of week-long study group students is the same as the mean score of overnight cramming students. $(\mu_1 > \mu_2 \text{ or } \mu_1 - \mu_2 > 0)$

Independent Groups Assumption: Scores of students from different classes should be independent.
Randomization Condition: Assume that the students are assigned to each class in a representative fashion.
10% Condition: 45 and 25 are less than 10% of all students.
Nearly Normal Condition: The histogram of the crammers is unimodal and symmetric. We don't have the actual data for the study group, but the sample size is large enough that it should be safe to proceed.



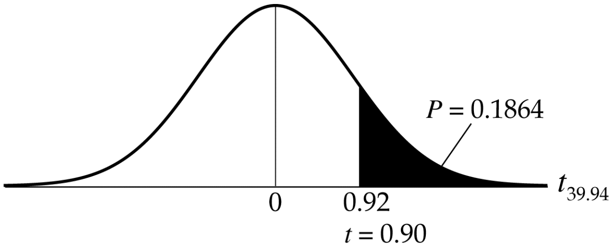
$$\bar{y}_1 = 43.2 \qquad \bar{y}_2 = 42.28$$
$$s_1 = 3.4 \qquad s_2 = 4.43020$$
$$n_1 = 45 \qquad n_2 = 25$$

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 39.94 degrees of freedom (from the approximation formula). We will perform a two-sample *t*-test. The sampling distribution model has mean 0, with standard error: $SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{3.4^2}{45} + \frac{4.43020^2}{25}} \approx 1.02076$.

The observed difference between the mean scores is 43.2 – 42.28 = 0.92.

Since the *P*-value = 0.1864 is high, we fail to reject the null hypothesis. There is no evidence that students with a week to study have a higher mean score than students who cram the night before.

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (0)}{SE(\bar{y}_1 - \bar{y}_2)}$$
$$t \approx \frac{0.92}{1.02076}$$
$$t \approx 0.90$$



b) H0: The proportion of study group students who will pass is the same as the proportion of crammers who will pass. $(p_1 = p_2 \text{ or } p_1 - p_2 = 0)$

HA : The proportion of study group students who will pass is different from the proportion of crammers who will pass. $(p_1 \neq p_2 \text{ or } p_1 - p_2 \neq 0)$

Random condition: Assume students are assigned to classes in a representative fashion.
10% condition: 45 and 25 are both less than 10% of all students.
Independent samples condition: The groups are not associated.
Success/Failure condition: $n_1\hat{p}_1$ = 15, $n_1\hat{q}_1$ = 30, $n_2\hat{p}_2$ = 18, and $n_2\hat{q}_2$ = 7 are not all greater than 10, since only 7 crammers didn't pass. However, if we check the pooled value, $n_2\hat{p}_{\text{pooled}}$ = (25)(0.471) = 11.775. All of the samples are large enough.

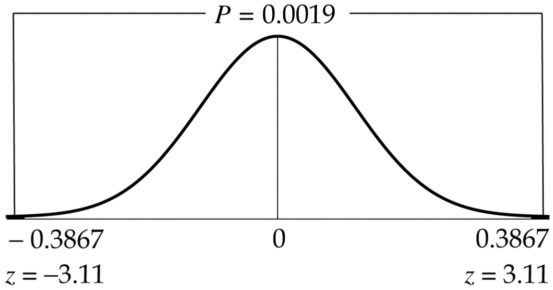
Since the conditions have been satisfied, we will model the sampling distribution of the difference in proportion with a Normal model with mean 0 and standard deviation estimated by $SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_1} + \frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_2}} = \sqrt{\frac{\left(\frac{33}{70}\right)\left(\frac{37}{70}\right)}{45} + \frac{\left(\frac{33}{70}\right)\left(\frac{37}{70}\right)}{25}} \approx 0.1245$.

The observed difference between the proportions is:
0.3333 – 0.72 = –0.3867.

Since the P-value = 0.0019 is low, we reject the null hypothesis. There is strong evidence to suggest a difference in the proportion of passing grades for study group participants and overnight crammers. The crammers generally did better.

$$z = \frac{-0.3867 - 0}{0.1245}$$

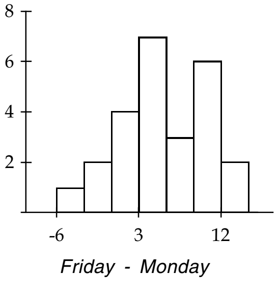
$$z \approx -3.11$$



c) H0: There is no mean difference in the scores of students who cram, after 3 days. $(\mu_d = 0)$

HA: The scores of students who cram decreases, on average, after 3 days. $(\mu_d > 0)$

Paired data assumption: The data are paired by student.
Randomization condition: Assume that students are assigned to classes in a representative fashion.
10% condition: 25 students are less than 10% of all students.
Nearly Normal condition: The histogram of differences is roughly unimodal and symmetric.



Since the conditions are satisfied, the sampling distribution of the difference can be modeled with a Student's t-model with 25 – 1 = 24 degrees of freedom, $t_{24}\left(0, \frac{4.8775}{\sqrt{25}}\right)$.

We will use a paired t-test, with $\bar{d} = 5.04$.

$$t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

$$t = \frac{5.04 - 0}{\frac{4.8775}{\sqrt{25}}}$$

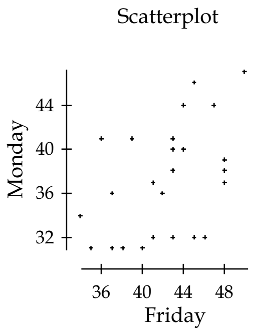
$$t \approx 5.17$$

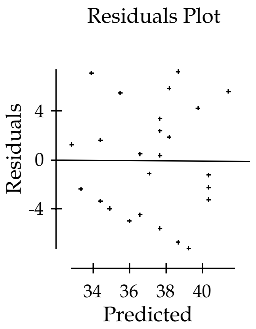
Since the P-value is less than 0.0001, we reject the null hypothesis. There is strong evidence that the mean difference is greater than zero. Students who cram seem to forget a significant amount after 3 days.

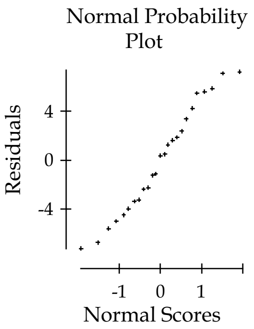
d) $\bar{d} \pm t^*_{n-1}\left(\frac{s_d}{\sqrt{n}}\right) = 5.04 \pm t^*_{24}\left(\frac{4.8775}{\sqrt{25}}\right) \approx (3.03, 7.05)$

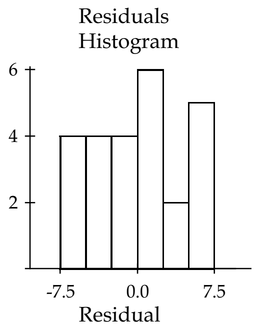
We are 95% confident that students who cram will forget an average of 3.03 to 7.05 words in 3 days.

e) H0: There is no linear relationship between Friday score and Monday score. $(\beta_1 = 0)$
HA: There is a linear relationship between Friday score and Monday score. $(\beta_1 \neq 0)$









Straight enough condition: The scatterplot is straight enough to try linear regression.
Independence assumption: The residuals plot shows no pattern.
Does the plot thicken? condition: The spread of the residuals is consistent.
Nearly Normal condition: The Normal probability plot of residuals is reasonably straight, and the histogram of the residuals is roughly unimodal and symmetric.

Since the conditions for inference are satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (25 – 2) = 23 degrees of freedom. We will use a regression slope *t*-test.

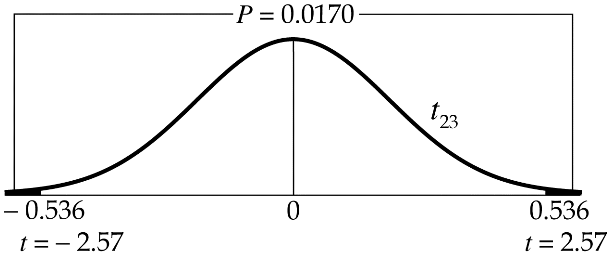
```
Dependent variable is:  Monday
No Selector
R squared = 22.4%     R squared (adjusted) = 19.0%
s =  4.518  with  25 - 2 = 23  degrees of freedom

Source        Sum of Squares   df   Mean Square   F-ratio
Regression    135.159           1      135.159     6.62
Residual      469.401          23      20.4087

Variable    Coefficient   s.e. of Coeff   t-ratio    prob
Constant    14.5921       8.847           1.65      0.1127
Friday      0.535666      0.2082          2.57      0.0170
```

The equation of the line of best fit for these data points is: $\widehat{Monday} = 14.5921 + 0.535666(Friday)$.

The value of $t \approx 2.57$. The *P*-value of 0.0170 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between Friday score and Monday score. Students who do better in the first place tend to do better after 3 days.


P = 0.0170
t23
– 0.536
t = – 2.57
0
0.536
t = 2.57