Coverage of test 2

1. Inference — confidence and prediction intervals
   (a) confidence interval for μ, known σ
   (b) confidence interval for μ, unknown σ
   (c) prediction interval for future observation, unknown σ
   (d) confidence interval for Binomial p
2. Inference — hypothesis testing
   (a) null, alternative hypotheses
   (b) Type I and II errors
   (c) one and two tailed tests
   (d) statistical significance versus practical importance
   (e) μ, known σ (with calculation of tail probabilities)
   (f) μ, unknown σ
   (g) paired samples, two–sample tests
   (h) Binomial p
   (i) two–sample Binomial comparison of probabilities
3. Regression
   (a) purposes of regression
   (b) linear model assumptions
   (c) least squares estimates; standard error of estimate; R^2
   (d) hypothesis testing — t, F tests
   (e) confidence intervals, prediction intervals
   (f) checking assumptions
   (g) multiple regression — t, F tests; interpretation of coefficients
   (h) transformations — log/log, semilog models

Note: You should not view the list above as exhaustive. Everything that we have discussed in class, or has been in a handout, is fair game to appear on the test.

Practice problems

1. The October 14, 1995, issue of the *New York Times* described a study of the conviction rates of people who are charged with killing their spouses. The study looked at cases in 1988 in 75 of the largest urban counties in the U.S., and found that 277 of 318 men charged with killing their wives were convicted or pleaded guilty, while 155 of 222 women charged with killing their husbands were convicted or pleaded guilty. Is there a difference between the probability that a randomly selected man charged with killing his wife is convicted or pleads guilty (p_M) and the probability that a randomly selected woman charged with killing her husband is convicted or pleads guilty (p_W)? Carefully state the hypotheses being tested here, and the test that you are using. Use $\alpha = .01$ for your test.

2. New York State law requires tobacco sellers not to sell to anyone under 18 years of age, and to seek proof of age from anyone looking 25 years of age or less. In May 1994, the Suffolk County Department of Health conducted an undercover operation

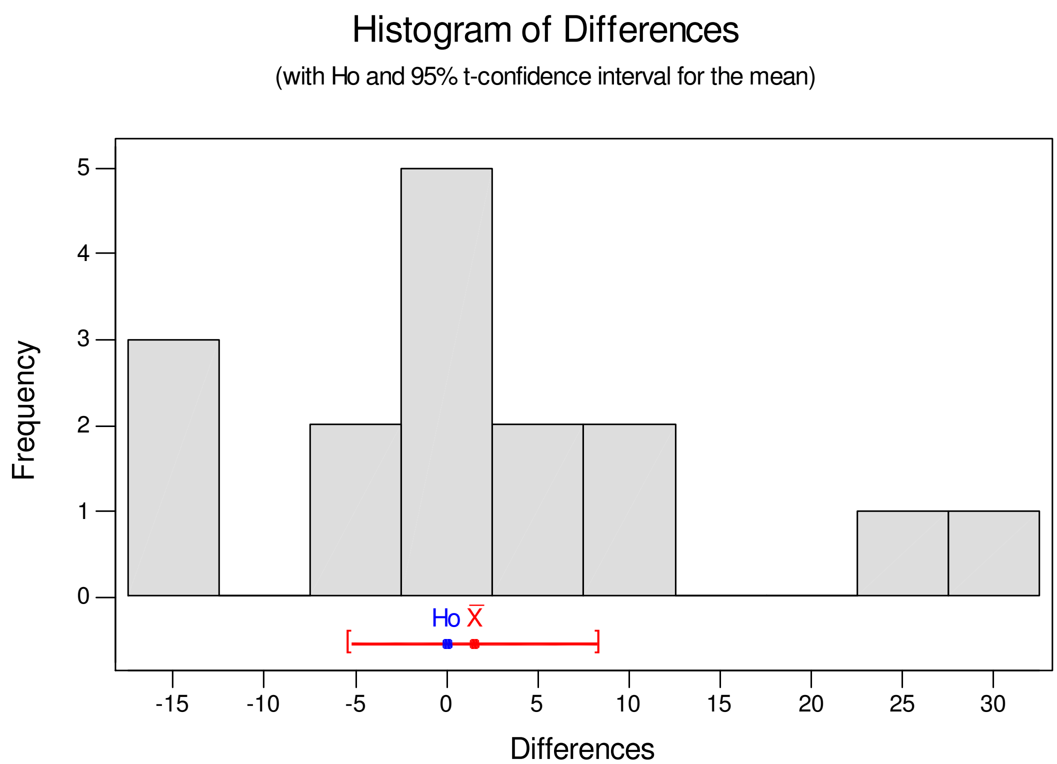
to see if sellers were complying with this law. Using operatives between 13 and 17 years of age, they visited 130 selling locations, sending in one operative at each to try to buy cigarettes. In 56 of the locations (i.e., 43.1%), the operative was successful in purchasing cigarettes.

(a) Construct a 75% confidence interval for p, the probability that a randomly selected Suffolk County tobacco seller will sell cigarettes to an underage buyer. What assumptions are you making in the construction of this interval? Do they seem reasonable here?

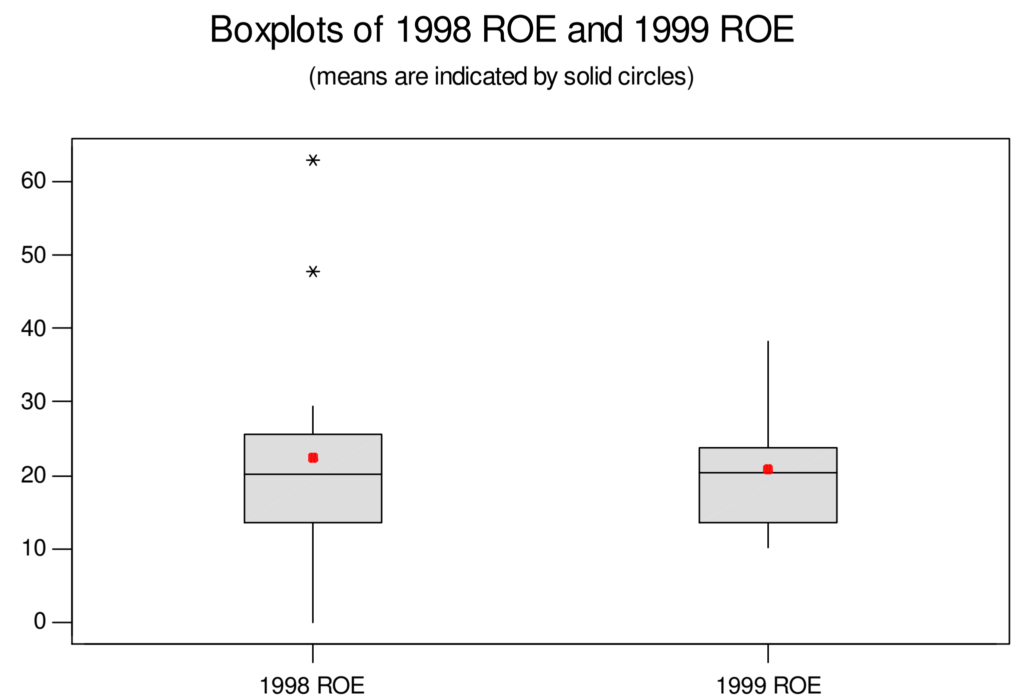
(b) A newspaper report states that "A good estimate of the probability that a randomly chosen typical underage teen will be able to successfully purchase cigarettes is .431." Comment.

3. Warren Buffett, chairman of Berkshire Hathaway and arguably the world's most successful value investor, is notorious for his supposed lack of enthusiasm in investing in technology stocks. Actually, Buffett claims that he treats technology stocks no differently from any other stock — he just hasn't found technology stocks with the proper profile for investment. High on Buffett's list of desirable attributes for a stock is that the company exhibit consistent return on equity (ROE). As part of a story on Buffett in the December 2000 issue of *Bloomberg Personal Finance*, ROE values for 1998 and 1999 were given for a sample of technology stocks (that is, the 1998 and 1999 values were given for each stock). The following is output from `Minitab` related to these data. The plots given refer to the output given just before them.

```
Paired T-Test and CI: 1998 ROE, 1999 ROE

Paired T for 1998 ROE - 1999 ROE

                  N      Mean     StDev   SE Mean
1998 ROE         16     22.31     15.04      3.76
1999 ROE         16     20.77      8.02      2.00
Difference       16      1.54     12.67      3.17

95% CI for mean difference: (-5.21, 8.30)
T-Test of mean difference = 0 (vs not = 0): T-Value = 0.49
        P-Value = 0.633
```




```
Two-sample T for 1998 ROE vs 1999 ROE

           N      Mean     StDev   SE Mean
1998 ROE  16      22.3      15.0       3.8
1999 ROE  16     20.77      8.02       2.0

Difference = mu 1998 ROE - mu 1999 ROE
Estimate for difference:  1.54
95% CI for difference: (-7.16, 10.24)
T-Test of difference = 0 (vs not =): T-Value = 0.36
       P-Value = 0.720  DF = 30
Both use Pooled StDev = 12.0
```



(a) Is the average ROE in 1998 significantly different from that in 1999? Carefully state the hypotheses that you are testing and the test(s) that you are using, being sure to justify your choice(s).

(b) Is there any reason to doubt the results in part (a)? If so, what would you do to try to answer a question comparing 1998 and 1999 ROE values? You needn't do anything, but be specific as to what you would do.

4. A bank wishes to investigate how long it takes customers to do their business (waiting time and time to perform their transactions, which we call serving time). A consulting firm is hired by the bank, which observes midmorning operations for a while. They report back that during the survey, 20 customers were served, with the average serving time being 3 minutes, and sample standard deviation of serving time being 1.5 minutes. The bank wants to make sure that no individual customers have too long a service time. Construct a 95% interval estimate for the serving time for an individual customer.

5. The following is an excerpt from an article which appeared in *The Independent Magazine*, dated November 14, 1992:

> *A parapsychologist projected slides on to a screen in an adjacent sound-proofed room. The slides were either blank rectangles or of a powerful emotionally affecting image. I had to tell through clairvoyance which type of slide — blank or affecting — was being shown on the screen in the next room. My score was 13 out of 24 ... this was enough to make me believe I had clairvoyant powers.*

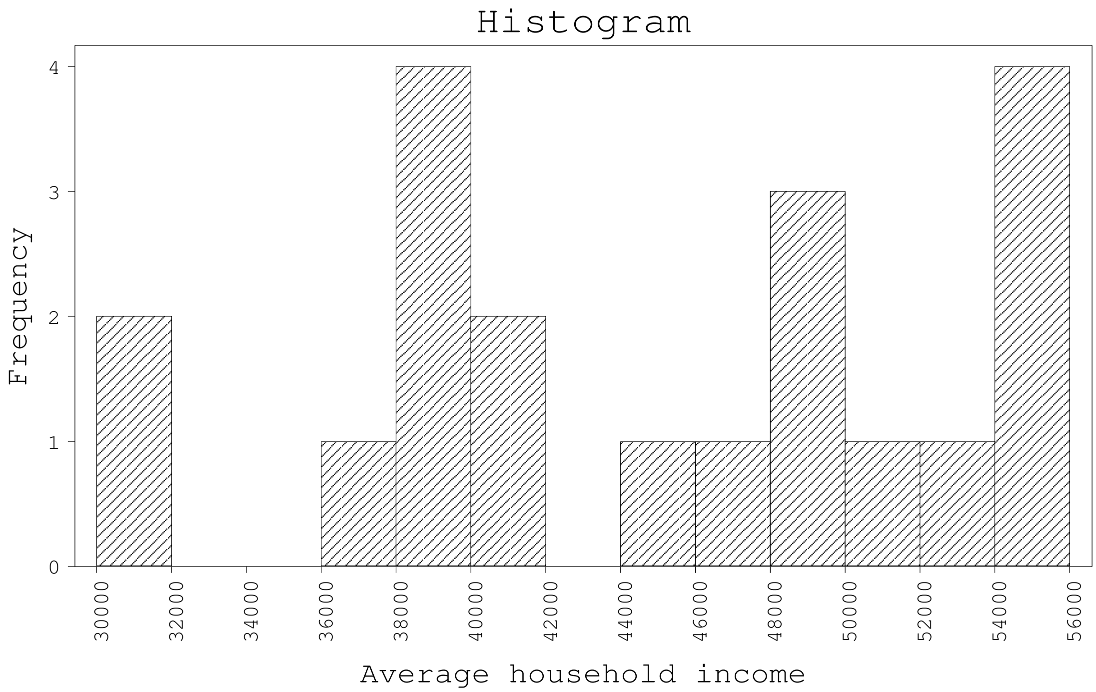
Do you agree with the author's conclusion? That is, does this result provide sufficient evidence to reject the hypothesis that it could have occurred simply by random chance? Carefully state the hypotheses you are testing, and the test that you are using. Use $\alpha = .05$.

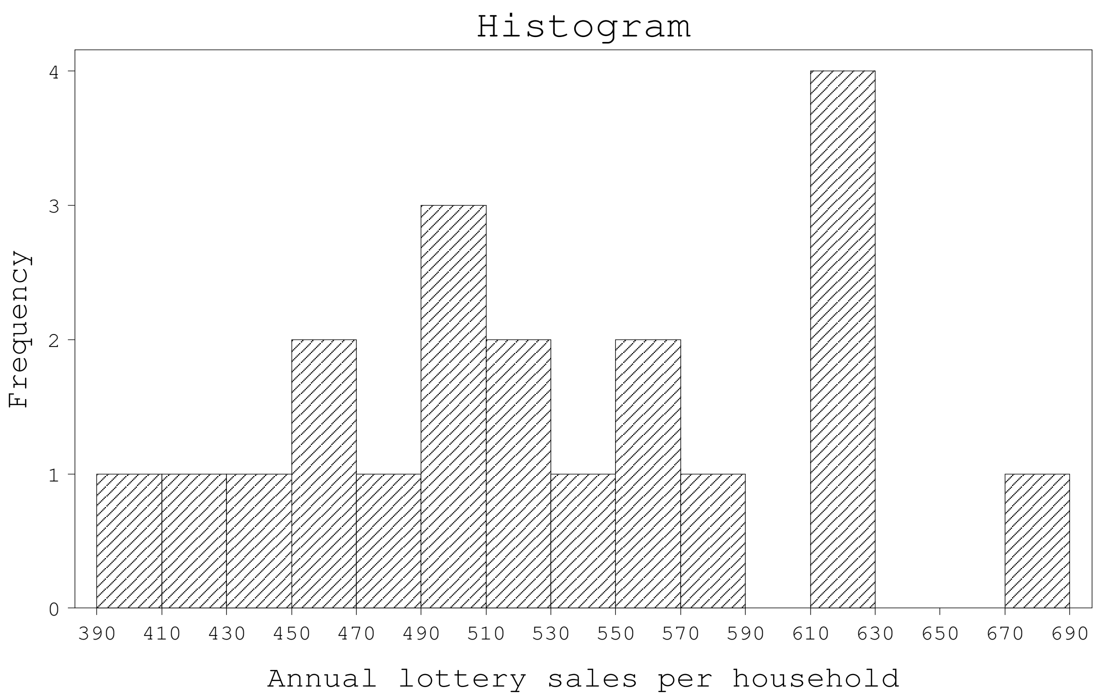
6. A personnel manager from a large financial services firm takes a random sample from the set of middle managers and determines their salaries. Here are the results (`SALARY` is the salary in thousands of dollars):

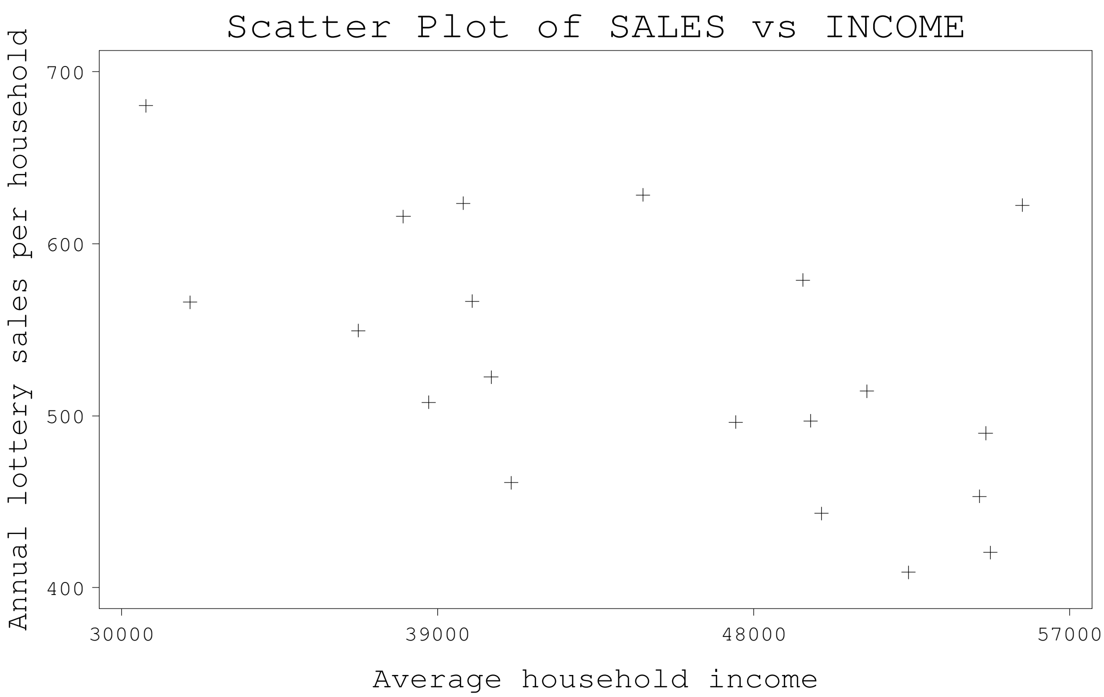
```
DESCRIPTIVE STATISTICS

                     SALARY
N                        18
MEAN                  86.92
SD                    36.29
MINIMUM               35.00
1ST QUARTILE          53.75
MEDIAN                82.50
3RD QUARTILE         113.30
MAXIMUM              142.00
```

   Construct an interval estimate (at a 90% confidence level) for the actual average salary of middle managers in this firm.

7. Billions of dollars are spent annually in the United States on state–run lottery games. One controversial aspect of these games is the potentially regressive nature of them as a form of taxation; that is, if poorer people spend more on the lottery than richer people do, this is effectively a tax imposed more heavily on the poor. One way to investigate this is to see if lottery sales per household are related to household income. The following analysis examines how the annual lottery sales per household in different ZIP code regions of Long Island is related to the average household income in that ZIP code (source: *Newsday*, December 5, 1995). First, here are histograms of the household incomes (`INCOME`) and lottery sales per household (`SALES`), as well as a scatter plot, for a sample of 20 ZIP code regions, followed by some regression output:


Histogram
Frequency
0
1
2
3
4
30000
32000
34000
36000
38000
40000
42000
44000
46000
48000
50000
52000
54000
56000
Average household income


Histogram
Frequency
0
1
2
3
4
390 410 430 450 470 490 510 530 550 570 590 610 630 650 670 690
Annual lottery sales per household




```
The regression equation is
Sales = 773.17 - 0.00534 Income

Predictor        Coef     SE Coef          T        P
Constant      773.171     87.9229       8.79    0.000
Income       -0.00534     0.00192      -2.78    0.012

S = 65.792      R-Sq = 30.0%     R-Sq(adj) = 26.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1     33436.8     33436.8      7.72    0.012
Residual Error    18     77915.1      4328.6
Total             19    111351.9
```

(a) What proportion of the variability in lottery sales per household is accounted for by average household income?

(b) The town of Albertson (ZIP code 11507) has an average household income of \$77,325. Give a 95% interval estimate for the true average annual lottery sales per household for ZIP code regions on Long Island with the same average household income as that of Albertson. Do you think this interval is likely to be useful in this case? The following output was obtained from Minitab using this income value:

```
Predicted Values for New Observations

New Obs      Fit     SE Fit         95.0% CI              95.0% PI
1         360.03     63.687   ( 226.23,  493.83)   ( 167.65,  552.41)

Values of Predictors for New Observations

New Obs    Income
1           77325
```

(c) The following output refers to a regression using the number of lottery outlets in the ZIP code region as well:

```
The regression equation is
Sales = 710.27 - 0.00453 Income + 0.89279 Outlets

Predictor        Coef      SE Coef          T        P
Constant      710.270      112.939       6.29    0.000
Income       -0.00453      0.00213      -2.13    0.049
Outlets       0.89279      0.99745       0.90    0.383

S = 66.159       R-Sq = 33.2%     R-Sq(adj) = 25.3%

Analysis of Variance

Source             DF          SS          MS          F        P
Regression          2     36943.5     18471.7       4.22    0.033
Residual Error     17     74408.4      4377.0
Total              19    111351.9
```

Which model do you prefer for these data? Why?

8. Orley Ashenfelter, a professor of economics at Princeton, has caused controversy in the wine lover's world by using statistical methods to model and predict auction prices of fine wines. A major focus of this work is on French Bordeaux wines. The following output comes from a regression of the natural logarithm (base e) of the price per case (in British pounds Sterling) of different Bordeaux wine vintages (LOGPPC) on the age of the vintage in years (AGE), average temperature over the growing season in degrees Celsius (AVETEMP), rain in August and September in centimeters (RAINSEAS), average temperature in September in degrees Celsius (SEPTEMP) and rain in the months

preceding the vintage in centimeters (RAINPREC) (source: "Bordeaux wine vintage quality and the weather," by Orley Ashenfelter, David Ashmore and Robert Lalonde, *Chance*, Fall 1995, pages 7–14):

```
The regression equation is
LogPPC = 2.8163 + 0.02400 AGE + 0.60799 AVETEMP - 0.00380 RAINSEAS
         + 0.00765 SEPTEMP + 0.00115 RAINPREC

Predictor          Coef       SE Coef           T          P
Constant        2.81631       0.31251        9.01     0.0000
AGE             0.02400       0.00747        3.21     0.0039
AVETEMP         0.60799       0.11599        5.24     0.0000
RAINSEAS       -0.00380       0.00095        4.00     0.0006
SEPTEMP         0.00765       0.05650        0.14     0.8938
RAINPREC        0.00115       0.00051        2.28     0.0324

S = 0.2931       R-Sq = 82.8%     R-Sq(adj) = 79.1%

Analysis of Variance

Source            DF         SS          MS          F        P
Regression         5    9.49998     1.90000      22.14   0.0000
Residual Error    23    1.97341     0.08581
Total             28    11.4734
```

(a) Is the overall regression statistically significant? Clearly state the hypotheses being tested, and the test that you are using. Use a significance level of $\alpha = .05$.

(b) Do any individual variables provide significant predictive power for logged price per case given the other predictors? Clearly state the hypotheses being tested, and the test that you are using. Use a significance level of $\alpha = .05$.

(c) What does the coefficient of AVETEMP say about the relationship between the growing season's average temperature and the price per case of the wine? Carefully state the interpretation of this coefficient in terms of price per case.

(d) The *Wine Spectator* reacted very negatively to this work, printing the following editorial statement: "The theory depends for its persuasiveness on the match between vintage quality as predicted by climate data, and vintage price on the auction market. But the predictions come exactly true only 3 times in the 28 vintages since 1961 that he's calculated, even though the formula was specifically designed to fit price data that already existed. The predicted prices are both under and over the actual prices." Comment on the validity of this criticism of the usefulness of the regression model.

9. The U.S. Bureau of Labor Statistics makes projections of the job growth in different occupations over different time frames. The following regression output comes from a

regression relating the current average annual salary in thousands of dollars (`Salary`) to the projected job growth, correcting for overall population growth (`Job growth`), over the next ten years for ten different occupations. These job growth figures range from -0.23 to 0.65 for this sample. You can assume that residual plots and diagnostics have been checked, and that no problems have been found.

```
The regression equation is
Salary = 36.665 + 32.1453 Job growth

Predictor           Coef       SE Coef            T           P
Constant         36.6652       1.15346        31.79       0.000
Job growth       32.1453       2.73751        11.74       0.000

S = 3.1523       R-Sq = 94.5%     R-Sq(adj) = 93.8%


Analysis of Variance

Source            DF          SS          MS         F         P
Regression         1     1370.14     1370.14    137.89    0.0000
Residual Error     8     79.4934     9.93668
Total              9     1449.64
```

A Ph.D. student in labor economics is interested in "stable" occupations — that is, ones that will have no job growth or decline over the next ten years (correcting for overall population growth). She claims that the true average current salary of such occupations is greater than $35,000, a claim that is not consistent with generally accepted beliefs. Construct a test for her to see if there is enough evidence here to convince her advisor of her claims so that she can get her Ph.D.

Answers to practice problems

1. We are testing

$$H_0 : p_M = p_W$$

versus

$$H_a : p_M \neq p_W,$$

and use a z–statistic

$$\begin{aligned} z &= \frac{\overline{p}_M - \overline{p}_W}{\sqrt{\overline{p}_M(1-\overline{p}_M)/n_M + \overline{p}_W(1-\overline{p}_W)/n_W}} \\ &= \frac{.871 - .698}{\sqrt{(.871)(.129)/318 + (.698)(.302)/222}} \\ &= 4.79, \end{aligned}$$

which has a tail probability less than .00001. So, there is very strong evidence that men are convicted or plead guilty at a higher rate than women are.

2. Our analysis is based on a Binomial process.

(a) The confidence interval for a Binomial probability p has the form

$$\bar{p} \pm z_{\alpha/2}\sqrt{\bar{p}\,(1-\bar{p}\,)/n}.$$

For a 75% confidence interval, $\alpha = .25$, so $z_{\alpha/2} = 1.15$. Thus, the interval has the form

$$\begin{aligned} .431 &\pm (1.15)\sqrt{(.431)(.569)/130} \\ &= .431 \pm .05 \\ &= (.381, .481). \end{aligned}$$

We are assuming the usual assumptions of a Binomial process; that is, (1) the purchasing result at each store must be independent of the result at any other store, and (2) there is a constant probability of successful purchase at each store. Also, (3) we are assuming that the sample size is large enough for the Central Limit Theorem to be operating. Assumptions (1) and (3) are probably reasonable, but assumption (2) probably isn't — we would expect that the probability of successfully purchasing cigarettes would be different for a 13 year old operative than for a 17 year old operative.

(b) This is not a reasonable statement. The frequency estimate .431 refers to a proportion of *sellers*, not a proportion of *buyers* (teenagers). These might seem like the same thing, but they're not. Here's a way to think about it. Say there are 1000 tobacco sellers in the county, and 450 do sell tobacco to underage buyers while 550 don't (that is, the true p is .45). The proportion of underage buyers who successfully purchase tobacco products isn't necessarily p, because it also depends on the probability that the underage buyer goes to a particular seller (which depends on the geographic distribution of buyers and sellers, for example), and how often they do so.

3. (a) This is a paired sample problem, so the two–sample t–test output is irrelevant (we have *one* set of companies, and we're comparing *two* variables). We're testing

$$H_0 : \mu_{1998} = \mu_{1999}$$

versus

$$H_a : \mu_{1998} \neq \mu_{1999}.$$

The test ($t = .49$) has $p = .633$, so we don't reject the null. That is, the average ROE values are not significantly different.

   (b) We would wonder about normality of the ROE difference. The histogram doesn't look too bad, so we're probably okay.

4. The desired prediction interval has the form

$$\bar{X} \pm t^{n-1}_{\alpha/2} s \sqrt{1 + \frac{1}{n}},$$

or

$$3 \pm (2.09)(1.5)\sqrt{1.05} = 3 \pm 3.212 = (-.212, 6.212).$$

Note that the interval includes (impossible) negative values. This is because the service time distribution is probably not normally distributed (it is more likely to be exponentially distributed, which is long right–tailed).

5. The hypotheses: let p be the probability that the author will match the type of picture correctly. The hypotheses being tested are then

$$H_0 : p = \frac{1}{2}$$

versus

$$H_a : p \neq \frac{1}{2}.$$

The test to use is a z–test, which has the form

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/N}},$$

where $\bar{p} = s/N$, s is the number of successes in N Binomial trials, and p_0 is the hypothesized value of p under the null hypothesis. Here $\bar{p} = 13/24 = .542$, so

$$z = \frac{.542 - .5}{\sqrt{(.5)(.5)/24}} = .41$$

To calculate the tail probability for this test, we need to calculate $P(|Z| > .41)$, where Z follows a standard normal distribution. The normal table provides the result of .682. So, we do not reject the null hypothesis at any reasonable α level; that is, the result does **not** provide any evidence that the author actually has clairvoyant powers. We're assuming here that the normal approximation to the binomial is valid, which might be questionable given the relatively small number of trials. The exact test results from `Minitab` show, however, that the conclusion stays the same even without the Central Limit Theorem assumption:

```
Test and Confidence Interval for One Proportion

Test of p = 0.5 vs p not = 0.5

                                                          Exact
Sample      X      N  Sample p         95.0 % CI        P-Value
1          13     24  0.541667  (0.328208, 0.744470)     0.839
```

6. What is required here is a confidence interval for the mean salary. This has the form

$$\begin{aligned}\bar{X} \pm t_{\alpha/2}^{n-1} \frac{s}{\sqrt{n}} &= 86.92 \pm (1.74) \frac{36.29}{\sqrt{18}} \\ &= 86.92 \pm 14.88 \\ &= (\$72040, \$101800)\end{aligned}$$

7. (a) This is estimated by the R^2, which is 30% here.
   (b) What is desired here is a confidence interval for the true average annual sales (the "fitted" bounds). This is $(226.23, 493.83)$. This is unlikely to be very useful, however, since we are *extrapolating* (the household income for Albertson is much higher than that of any of the observations in the sample), which is always a dangerous thing to do.
   (c) Assuming regression plots and diagnostics look okay, we prefer the one–variable model, since `OUTLETS` does not add any significant predictive power to the model.
8. (a) We are testing the hypotheses

$$H_0 : \beta_{AGE} = \beta_{AVETEMP} = \beta_{RAINSEAS} = \beta_{SEPTEMP} = \beta_{RAINPREC} = 0$$

   versus

$$H_a : \text{at least one of these coefficients} \neq 0.$$

   This is tested using the F–test, which has a tail probability less than .0001. Thus, the overall regression relationship is highly statistically significant.
   (b) We are testing the hypotheses

$$H_0 : \beta_j = 0$$

   versus

$$H_a : \beta_j \neq 0,$$

   where j refers to the five predicting variables `AGE`, `AVETEMP`, `RAINSEAS`, `SEPTEMP` and `RAINPREC`. These are tested using t–statistics, as follows:

AGE	AVETEMP	RAINSEAS	SEPTEMP	RAINPREC
$t = 3.21$	$t = 5.24$	$t = 4.00$	$t = 0.14$	$t = 2.28$
$p = .0039$	$p < .0001$	$p = .0006$	$p = .8938$	$p = .0324$
Reject H_0	Reject H_0	Reject H_0	Don't reject H_0	Reject H_0

Thus, AGE, AVETEMP, RAINSEAS and RAINPREC add significantly to the model's predictive power at a .05 level.

(c) The coefficient is an estimate of the average change in log price per case associated with a one degree increase in AVETEMP, holding all of the other variables in the model fixed. Here, the estimated average change is an increase of .608; since the target variable is in the log scale, that translates into an estimate of multiplying price per case by $e^{.608} = 1.837$, or an increase of about 84% (recall that the target variable is natural log, not log base 10).

(d) This comment is pretty funny, as it reveals a complete lack of understanding of statistical modeling. The worth of a model is not measured by the number of *exact* predictions, but rather by the general level of accuracy of the predictions. The writer of the editorial clearly does not understand random variability. Also, the predictions being both above and below the actual values is a *good* thing, since if they were consistently too high or too low, that would show that there was a bias in the model somewhere.

9. This question relates to the constant term β_0 in the regression. Specifically, the student's question tests the hypotheses

$$H_0 : \beta_0 = 35$$

versus

$$H_a : \beta_0 > 35.$$

These hypotheses are tested using a t–test:

$$\begin{aligned} t &= \frac{\hat{\beta}_0 - 35}{\text{s.e.}(\hat{\beta}_0)} \\ &= \frac{36.665 - 35}{1.153} \\ &= 1.44 \end{aligned}$$

The critical value for a one–tailed $\alpha = .05$ test is 1.86, so this is not significant at a .05 level. The critical value at a .10 level is 1.40, so it is significant at that level (in fact, the tail probability is .094). So, there is weak evidence in favor of her hypothesis, but probably not enough for her to get her doctorate.