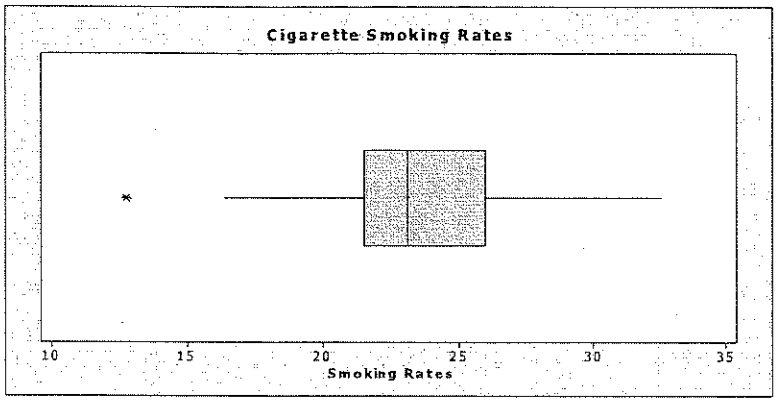
Answer Key for Practice Exam 2
Free Response

1.

(a) Determine if there are any outliers: $Q1 - 1.5IQR = 21.5 - 1.5(4.5) = 14.75$
$Q3 + 1.5IQR = 26 + 1.5(4.5) = 32.75$

According to the data, Utah's smoking rate of 12.7% is an outlier since it is less than the lower fence. The end of the lower whisker should extend to the last data point within the fence, which would be 16.4% (California). There are no upper outliers.



(b) Since the mean is slightly higher than the median, the smoking rate distribution is skewed to the right. There are a few states with fairly high smoking rates. Note that 50% of the states have rates between 23.1% and 32.6% (a difference of 9.5 percentage points) and the remaining states are between 12.7% and 23.1% (a difference of 10.4 percentage points). The outlier is Utah.

(c) Guam's smoking rate is quite high. This rate is more that 2.6 standard deviations above the mean $\left(z = \frac{32.1 - 23.35}{3.38}\right)$, putting it in near the top of the entire distribution of smoking rates. The smoking rate for Puerto Rico, on the other hand, was much more unusual. Its rate is actually an outlier (less than 14.75%) ranking it just ahead of Utah, the lowest state for smoking.

2.

(a) Both plans assume that the pigeons are representative and that you have a reasonable random sample of pigeons. In both plans, pigeons are randomly assigned to the two groups and that each pigeon's behavior is independent of other pigeons, i.e., each pigeon navigates its own way back to the coop and does not depend on other pigeons. Plan I considers the possibility that not all pigeons may return to the coop. By tallying how many pigeons returned to the coop within a day one would get the proportion of pigeons that returned under each treatment condition. Plan II, on the other hand, assumes that all pigeons will return and that their return time distribution would be approximately normal. If a few pigeons get lost and take most of the day to return or never return, while the rest arrive at the coop in a reasonably short period of time, the time distribution could be highly skewed with the possibility outliers. This would lead to a violation of the conditions necessary for a t test on two independent means. Plan I appears to be more optimal.

(b) For Plan I, the appropriate test is a two proportions z test.
Let p_1 = the proportion of pigeons from the group released on a sunny day that returned to the coop and p_2 = the proportion of pigeons from the group released on a cloudy day that returned to the coop.

$$H_o: p_1 = p_2 \text{ versus } H_a: p_1 \neq p_2$$

(c) Conditions are:

1. Each group is taken from a random sample. The problem states that a random sample of pigeons was used.
2. Each pigeon's behavior is independent of other pigeons. We will have to assume that this is true.
3. The random assignment of the pigeons creates two independent groups.
4. We have a large sample size. For each of the two pigeon groups, $n\hat{p} > 5$ and $n(1-\hat{p}) > 5$. For example, if only 40% of the pigeons returned from each group, then $n\hat{p} = (.4)(40) = 16 > 5$ and $n(1-\hat{p}) = .6(40) = 24 > 5$ would satisfy the condition of a large enough sample.

3.

(a) Proportion of teenage drivers who exhibit risky driving behavior when the passenger is male $= \dfrac{165+71}{944} = 0.25$

Proportion of teenage drivers who exhibit risky driving behavior when the passenger is female $= \dfrac{47+51}{944} = 0.104$

Teenage drivers are about 2.5 times as likely to exhibit risky driving behaviors when the passenger is male.

(b) Perform a chi-square test for the homogeneity of populations.

H_o: Males and females exhibit the same driving behaviors.
H_a: Males and females exhibit different driving behaviors.

Conditions check:

- We have count data and they are bivariate in nature.
- We randomly selected five suburban high schools in a large metropolitan area.
- The expected cell counts are all greater than five. $\begin{bmatrix} 122.25 & 113.75 \\ 50.76 & 47.24 \\ 50.76 & 47.24 \\ 265.22 & 246.78 \end{bmatrix}$

Mechanics:

Performing the test of homogeneity yields a test statistic of $\chi^2 = 43.046$.

Conclusion:

Since the P-value ≈ 0, we reject H_o. There is sufficient evidence to conclude that teenage males and females at the suburban schools in this metropolitan area exhibit significantly different driving behaviors.

Based on the work in part (b), the driving behaviors of male and females are significantly different. Part (a) indicates that much riskier behavior occurs when the passenger is male rather than female. A closer look at the data shows that male drivers have a much larger number of recorded risky behaviors (165) when the passenger is male. This is much higher than any other of the risky behavior cell counts.

Below is an alternative argument that could be presented for part (b).

We can calculate the conditional probabilities as follows:

$$\text{P(risky behavior with a male passenger} \mid \text{male driver)} = \frac{165}{489} = 0.337$$

$$\text{P(risky behavior with a female passenger} \mid \text{male driver)} = \frac{47}{489} = 0.096$$

$$\text{P(risky behavior with a male passenger} \mid \text{female driver)} = \frac{71}{455} = 0.156$$

$$\text{P(risky behavior with a female passenger} \mid \text{female driver)} = \frac{51}{455} = 0.112$$

Females have about the same rate of risky driving behaviors whether or not they have a male or a female passenger. Furthermore, this appears to be pretty close to what they exhibit if they have no passengers at all. Males, on the other hand, are most reckless when they have a male passenger on board. They are more than twice as likely to engage in risky driving behavior than are female drivers who have a male passenger. At the same time, males seem to apply the brakes and are far more cautious when they are driving a female passenger. When escorting a female, they are only one-fourth as likely to exhibit risky driving behavior as when they have a male passenger.

(c) We cannot generalize these results to all teenagers since we sampled only one metropolitan area and looked only at suburban schools. The scope of inference would be to suburban high schools in this metropolitan area. It may be that urban students in this large metropolitan area have access to public transportation and therefore do less driving. Furthermore, students in other metropolitan areas and rural areas may exhibit different risky driving behaviors.

4.

(a) This was an experiment, since a treatment (new method) was imposed on the workers.

(b) Since the number of figurines that can be painted may vary from day to day, counting the number of figurines painted over the course of a week would give a better estimate for daily productivity.

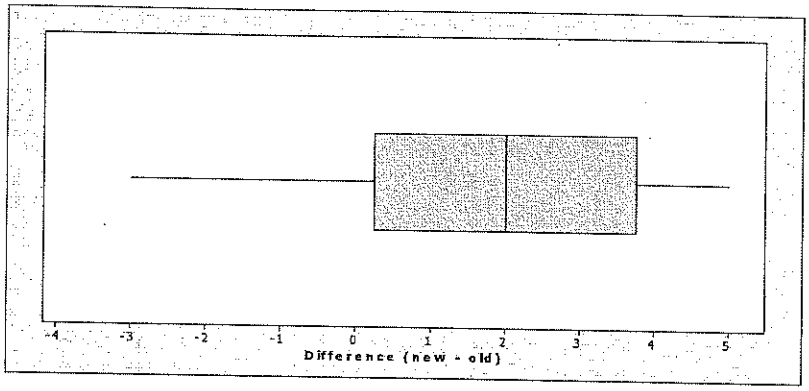
(c) Using only experienced painters reduces the variability that would exist if the group contained painters who possessed various levels of experience. In addition, there might be confounding if most of the inexperienced painters ended up in one group (the new method) and the experienced painters ended up in the other (the old method). If no significant difference in number of figurines painted were found, the company would not know whether it was due to the new method being ineffective or the lack of experience of the painters.

(d) This is a matched pair t test.

H_o: $\mu_d = 0$ versus H_a: $\mu_d > 0$, where μ_d = the true mean difference of new method – old method.

Conditions to conduct the test:

1. Data are from an SRS. This is stated in the problem.
2. There are no extreme departures from normality. The boxplot indicates no oultiers.



Use a right tailed one-sample t-test with 19 degrees of freedom

$$t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{1.85 - 0}{\frac{2.277}{\sqrt{20}}} = 3.63 \qquad P\text{-value} = P(t > 3.63) = 0.00089$$

Decision:

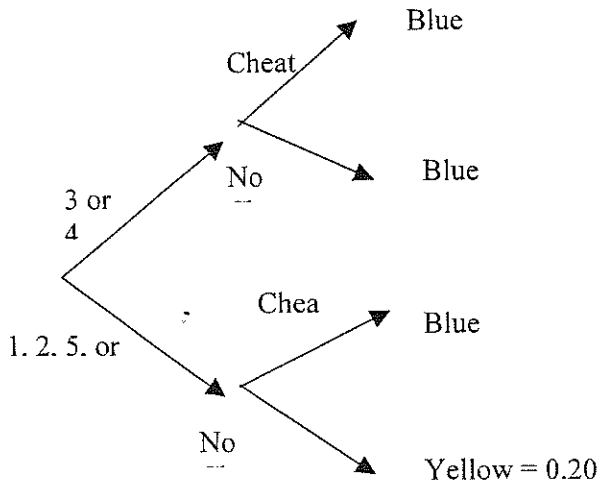
Since the P-value is so small, we reject H_o.

Conclusion:

There is sufficient evidence to support the claim that there was an increase in the mean number of figurines painted by each worker. The new method significantly increases the productivity of the painters.

5.

(a) Use a tree diagram to determine possible cases



P(3 or 4) = $\frac{1}{3}$ and P(1, 2, 5, or 6) = $\frac{2}{3}$.

P(did not cheat) = X. So P (1, 2, 5, or 6) = $\frac{2}{3}X = 0.20 \Rightarrow X = 0.30$

P(Cheat) = 1 − 0.30 = <u>0.70</u>

(b) p = true proportion of people who have been dishonest on their income tax return

<u>Conditions check:</u>

- We are told that a random sample was taken.
- $n\hat{p} = (90)(.7) = 63 > 10$ *and* $n(1-\hat{p}) = (90)(.3) = 27 > 10$. Therefore the sample is large enough to achieve normality.

<u>Mechanics:</u>

$$C.I. = \hat{p} \pm z^*_{0.96}\sigma_{\hat{p}} = 0.70 \pm 2.054\left(\sqrt{\frac{(0.70)(0.30)}{90}}\right) = 0.70 \pm 0.099 \Rightarrow 0.601 \le p \le 0.799$$

<u>Interpretation:</u>

We are 96% confident that the true proportion of people who have been dishonest on their income tax returns is between 0.601 and 0.799.

6.

(a) It appears that Model B is somewhat better in determining the relationship between median household income and median housing prices. The r-square value is somewhat higher (.62 versus .52) and the scatterplot is more linear. In addition, the residual plot appears to be a bit more random. While there is still some nonconstant variance in the logarithmic model it seems less variable than the residuals from the original linear model.

(b) Hypotheses and identification of test:

β = true slope of the regression line relating ln(house value) and median household income.

$H_0 : \beta = 0 \quad versus \quad H_a : \beta \neq 0$

Use the t test for linear regression.

Conditions check: The problem states that the conditions are reasonably met.

Mechanics:

$$t = \frac{b - 0}{\frac{s_b}{\sqrt{n}}} = \frac{.00004895}{0.00000553} = 8.85$$

With df = 48, P-value = $P(t > 8.85) = 0.000$.

Alternatively, look at the computer output and find T = 8.85 and P-value ≈ 0.

Conclusion:

Since the P-value is so small, we reject H_0 and conclude that a linear relationship exists between ln (house value) and median household income.

(c) If median household income is \$40,000 then
ln(median house value) = 9.7517 + 0.000049(40,000) = 11.7117.
Therefore, predicted median house value = $e^{11.7117}$ = \$121,990.69.

(d) There does not appear to be a relationship between median household income and the rate of homeownership. No matter the income level, the percent of people who own their own homes in the majority of states tends to mostly fall in the 70% to 80% range. It may be that the rate of home ownership is less (below 65%) in those states that are heavily urbanized, such as the Northeast, where more people might tend to rent rather than own or in the various parts of the country where house prices are unusually high.

(e) From the four scatterplots it is clear that homeownership declines substantially in the West and the Northeast as the median house values increase. Homeownership in the Midwest and South tended to increase slightly as the median house value increased, perhaps due to the more moderate prices of houses. By examining the horizontal axes of the four graphs, the regional price differences can be readily seen. In the Midwest and South the median prices do not go above $200,000, while those in the Northeast and West reach a median value between $350,000 and $400,000, making these latter areas much more difficult to buy a house in.

(f) The computer output supports the conclusions in part (e). The mean affordability ratio for the Northeast is 4.408 and for the West it is 4.222. These are both higher than the suggested maximum affordability value of 3. This would indicate house prices in these two areas are far outpacing the ability of people to buy them, which would decrease homeownership rates. Furthermore, the two confidence intervals for the Northeast and West are also significantly higher than those for the South and Midwest. This might explain the low ownership rates from the scatterplot in part (d). The Midwest and South, on the other hand, have affordability values less than 3, making homeownership more accessible.