Question 1: Multiple choice: circle the best answer to each question:

(1)[2 points] A reader of the Indianapolis Star wrote in to express doubt that a poll based on a random sample of 600 voters is sufficient to estimate the proportion of 2.5 million voters favouring a certain candidate in an upcoming election. This reader's point is

(a) not valid: variability of sample statistics doesn't depend on population size as long as the population is at least 100 times larger than the sample.

(b) valid: a sample of size 600 is much too small.

(c) not valid: all election polls use this sample size.

(d) valid: samples of size less than 1000 are often yield biased results.

(e) not valid: samples of size at least 200 are always sufficient.

(2)[2 points] A study was conducted starting January 1 2007 to examine two treatments used for prostate cancer. The study used medical records of prostate cancer sufferers over the past ten years. The study found that patients treated with surgery tended to outlive those treated with a non-surgical procedure. This study is

(a) an experiment

(b) a completely randomized design.

(c) an observational study.

(d) a matched pair design.

(e) a block design.

(3)[2 points] The attached table of random digits beginning at line 125 is used to select a sample of size five from a population of size 140. Individuals are labeled $000, 001, \ldots$. The label of the second person selected is

(a) 001

(b) 002

(c) 121

(d) 119

(e) 033

(4)[2 points] A poll based on a survey 1396 randomly selected drivers concluded that 63% speed at least some of the time. The poll has a margin of error of approximately 2.5 percentage points nineteen times out of twenty. The margin of error divided by the standard deviation of the sampling distribution is approximately

(a) 1

(b) 2

(c) 2.5

(d) 19/20

(e) cannot be determined

(5)[2 points] Comparing treatments to control the effects of lurking variables is known as:

    (a) Simple Random Samples

    (b) Randomized Comparative Experiments

    (c) Statistical Significance

    (d) Comparing matched groupings

    (e) The Placebo effect.

(6)[2 points] The best method for testing causation would be:

    (a) Sample Surveys

    (b) Observational Studies

    (c) Experimentation

    (d) Census

    (e) Polling

(7)[2 points] Repeating a measurement five times and taking the average value

    (a) reduces bias and random error

    (b) reduces bias only

    (c) reduces random error only

    (d) reduces neither bias nor random error

    (e) none of the above

(8)[2 points] Based on data from 1998 to 2000, Columbia has the highest homicide rate in the world with approximately 61.8 homicides per year per $100,000$ residents. If the population of Columbia is approximately $41,008,227$, approximately how many deaths annually are due to murder?

    (a) $10,000$

    (b) $15,000$

    (c) $20,000$

    (d) $25,000$

    (e) $30,000$

(9)[2 points] Temperatures of four fish market freezers were taken and found to be 3.3, 3.7, 4.1 and 2.9 degrees Celsius. What is $\bar{x} - s$?

    (a) 2.0

    (b) -2.0

    (c) 0.75

    (d) 3.0

    (e) -3.0

(10)[2 points] With reference to the sample represented in the following stemplot

```
1 | 4  5
2 | 1  2  3  3
3 |
4 | 8  9
```

(a) the mean is 26.875 and the median is 22.5

(b) the first quartile is 21

(c) the median is 22

(d) the maximum is 88.5

(e) none of the above

(11)[2 points] The median of a distribution is larger than the mean when

(a) the distribution is skewed to the right.

(b) the distribution is skewed to the left.

(c) the distribution is symmetric.

(d) the distribution is represented using a stemplot.

(e) the distribution is normal.

(12)[2 points] The following boxplot summarizes data for populations A and B of equal size:
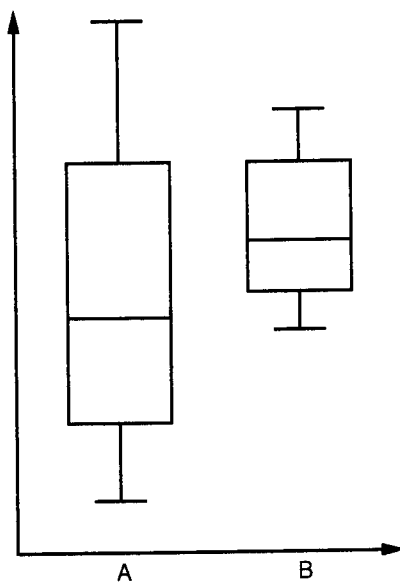
(a) the median of A is less than the median of B.

(b) the median of A is less than the first quartile of B.

(c) The corresponding distributions are right skewed.

(d) (a) and (b) only.

(e) (a), (b) and (c).

(13)[2 points] Referring to the boxplot in (12),

    (a) the standard deviation of A is less than that of B.

    (b) the standard deviation of A is greater than that of B.

    (c) the standard deviation of A is about the same as that of B.

    (d) the standard deviations are the same since the third quartiles are equal.

    (e) none of the above.

(14)[2 points] In a normal distribution, the percentage of values that fall between one and two standard deviations from the mean is approximately

    (a) 95%

    (b) 99.7%

    (c) 68%

    (d) 27%

    (e) 54%

(15)[2 points] A large class has test scores which are normally distributed with a standard deviation of 5.5. The percentage of the class scoring at least one and a half standard deviations above the mean is approximately

    (a) 7%

    (b) 93%

    (c) 43%

    (d) 44%

    (e) 68%

(16)[2 points] The correlation for two variables is calculated to be $r = 0.72$. Which of the following values of r indicates a stronger linear relationship?

    (a) 1.2

    (b) 0.68

    (c) 0

    (d) -0.72

    (e) -0.83

(17)[2 points] Two researchers examine the same data set summarizing the trunk circumference (x, in metres) and location above sea level (y, in metres) of 50 spruce trees. Both researchers calculate the correlation, but the second researcher converts all measurements of circumference to centimetres first, and he instead uses x for location and y for circumference. If r_1 and r_2 are the correlations found by the first and second researchers, respectively, then

(a) $r_1 = 1/r_2$

(b) $r_1 = 100/r_2$

(c) $100r_1 = r_2$

(d) $r_1 = r_2$

(e) $100r_1 = 1/r_2$

(18)[2 points] Let x represent the height of male students at Malaspina, and assume the distribution of these heights is approximately normal with a certain mean μ and standard deviation σ. A student calculates a z-score (or standard score) of $z_1 = 2$ corresponding to a certain value of x, but then realizes the μ and σ he used were incorrect; the correct values of each are half as large. He calculates a second z-score z_2 using the correct values for mean and standard deviation. How will z_1 and z_2 compare?

(a) z_1 will be larger than z_2.

(b) z_1 will be smaller than z_2.

(c) z_1 will be the same as z_2.

(d) Because of the error x is no longer normal.

(e) none of the above.

(19)[2 points] In the distribution below
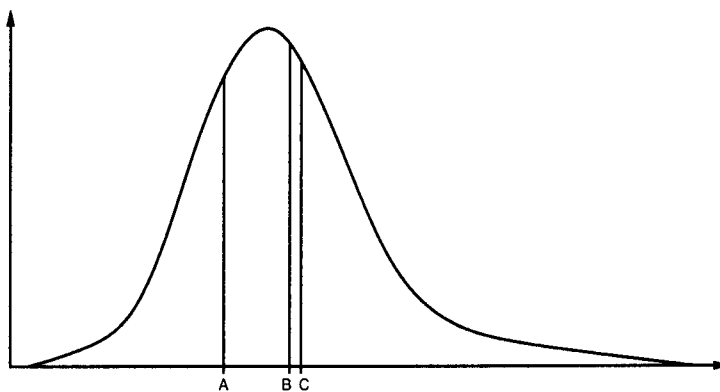
(a) the median is A and the mean is B.

(b) the median is B and the mean is C.

(c) the median is C and the mean is A.

(d) the median is C and the mean is B.

(e) the median is B and the mean is A.

(20)[2 points] In the scatterplot below
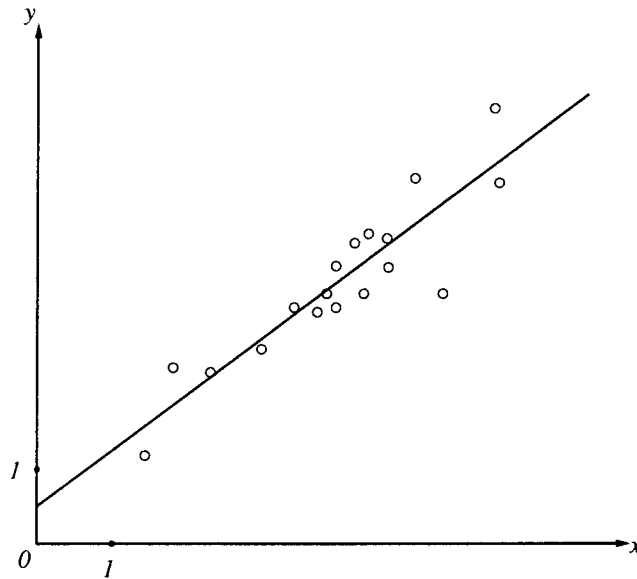
   (a) $r = 0.1$ and the regression line is $y = -0.5x + 1$.

   (b) $r = -0.1$ and the regression line is $y = 0.5x + 1$.

   (c) $r = 0.8$ and the regression line is $y = 0.75x + 0.5$.

   (d) $r = 0.8$ and the regression line is $y = 0.75x - 0.5$.

   (e) $r = 1.0$ and the regression line is $y = x$.

Question 2: A grocery store selling tomatoes must keep the inventory carefully balanced: too large an inventory results in tomatoes rotting in the storeroom or while out on display, while too small an inventory presents the risk of running out and hence losing customers. Reliable sales data shows that on average 207 kilograms of tomatoes are sold each week, while the standard deviation from this mean is 17 kilograms per week. Assume that the number of kilograms of tomatoes sold each week is normally distributed, and tomatoes are delivered to the store once per week.

(a)[2 points] The middle 99.7% of weekly demand values fall between what two extremes? Round to one decimal. You may use the "quick method" here if you like.

99.7 % of all values fall within 3 standard deviations of the mean, so the two extremes are

$207 - 3(17) = \boxed{156 \ kg}$ and

$207 + 3(17) = \boxed{258 \ kg.}$

(b)[4 points] How many kilograms of tomatoes should the grocer stock at the beginning of each week to ensure that demand is met 50 out of 52 weeks per year? Round to the nearest kilogram.

$\dfrac{50}{52} = 0.9615 = 96.15\%$ of weeks.

z-score corresponding to 96.15 percentile is $z \doteq 1.8$.

Let x be the required stock level. Then

$1.8 = \dfrac{x - 207}{17}$, so $x = 207 + 1.8(17) \doteq \boxed{238 \ kg}$

(c)[4 points] Suppose the grocer has a long standing weekly order for 224 kilograms per week, but he decides to increase his weekly order to 241 to ensure that demand will be met for a greater percentage of weeks. For how many additional weeks per year will demand now be met? Round to one decimal.
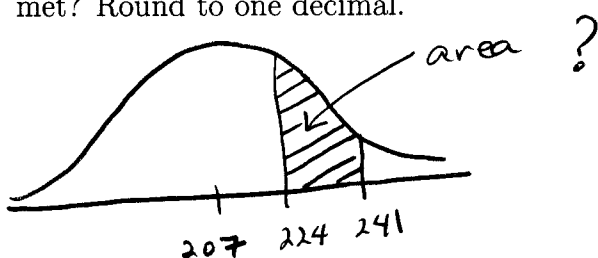

area ?

For 224 kg, $z = \dfrac{224 - 207}{17} = 1$

For 241 kg, $z = \dfrac{241 - 207}{17} = 2$

∴ Percentage represented by area is $97.73 - 84.13$

$= 13.6 \%$

∴ Demand will be met for $(0.136)(52) \doteq \boxed{7 \ additional \ weeks \ per \ year}$

Question 3: A game consists of spinning the two wheels shown, first Wheel 1 (divided into three areas, the first twice the size of each of the other two), and second Wheel 2 (divided into three equally spaced areas.) In a single play of this game you receive an amount of money equal to the total of the numbers shown on the two wheels. For example, if you spin a 1 on Wheel 1 and a 0 on Wheel 2 then you receive $1 + 0 = 1$ dollar. A negative total means you have to pay. For example, a total of -2 mean you must pay $2.
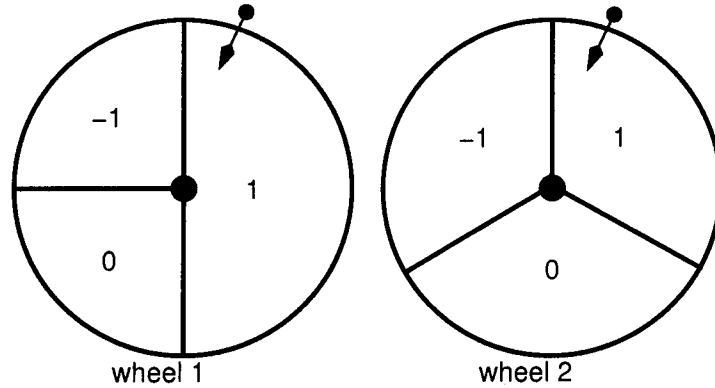
wheel 1   wheel 2

(a)[3 points] What is the probability of receiving less than $2 (including having to pay) in a single play of this game?

$2 is the maximum payout resulting from a 1 on wheel 1 & a 1 on wheel 2.

∴ $P(\text{less than } 2) = 1 - P(\text{total} = 2)$

$= 1 - P(1 \text{ on wheel } 1) P(1 \text{ on wheel } 2)$ by independence

$= 1 - (\frac{1}{2})(\frac{1}{3}) = \boxed{\frac{5}{6} \doteq 0.83}$

(b)[4 points] What is the probability of of getting a total of zero?

Break the event "total of zero" into disjoint events:

Wheel 1	Wheel 2	Probability
-1	1	$(\frac{1}{4})(\frac{1}{3}) = \frac{1}{12}$
0	0	$(\frac{1}{4})(\frac{1}{3}) = \frac{1}{12}$
1	-1	$(\frac{1}{2})(\frac{1}{3}) = \frac{1}{6}$

total : $\frac{1}{12} + \frac{1}{12} + \frac{2}{12} = \frac{1}{3}$

∴ $P(\text{total of zero}) = \frac{1}{3}$

(c)[3 points] What is the total of the expected outcomes of each of the wheels?

wheel 1: expected outcome $= (-1)(\frac{1}{4}) + (0)(\frac{1}{4}) + (1)(\frac{1}{2}) = \frac{1}{4}$

wheel 2: expected outcome $= (-1)(\frac{1}{3}) + (0)(\frac{1}{3}) + (1)(\frac{1}{3}) = 0$

∴ total of the expected outcomes is $\frac{1}{4} + 0 = \boxed{\frac{1}{4}}$

Question 4:

(a)[5 points] Of a random sample of 1764 seniors, 1135 felt that the current level of fully funded healthcare provided by the government was adequate, while 71% of them felt that more support was needed to allow seniors to stay in their own homes for as long as possible. Give a 99.9% confidence interval for the percentage of seniors who feel that healthcare levels are adequate. Round to one decimal.

$$n = 1764$$

$$\hat{p} = \frac{1135}{1764} \doteq 0.6434$$

$$z^* = 3.29$$

$$\therefore \quad \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 0.6434 \pm 3.29 \sqrt{\frac{0.6434(1-0.6434)}{1764}}$$

$$\doteq 0.6434 \pm 0.0375$$

$$\doteq \boxed{64.3\% \pm 3.8\%}$$

(b)[5 points] A 90% confidence interval for the mean lifetime of a certain brand of lightbulb was found to be 973 hours to 989 hours. The confidence interval was based on the sample mean and standard deviation of a sample of 784 randomly selected lightbulbs. What was the sample standard deviation? Round to the nearest hour.

$$n = 784$$

$$\bar{x} = \frac{989 + 973}{2} = 981$$

$$\therefore \bar{x} + z^* \frac{s}{\sqrt{n}} = 989$$

$$981 + 1.64 \frac{s}{\sqrt{784}} = 989$$

$$s = \left(\frac{989 - 981}{1.64}\right)\sqrt{784} \doteq \boxed{137 \text{ hrs}}$$

Question 5 [10 points]: In its 2006 annual report, BC Hydro states that 85% of customers give them a ranking of at least 7 on a 10 point scale of customer satisfaction. Suppose we conduct our own test and find that 40 of 175 randomly selected BC Hydro customers give a satisfaction rating of 6/10 or less. Is this result sufficient to conclude that the proportion of BC Hydro customers satisfied at the 7/10 (or better) level is less than what is claimed in the annual report? Test at a 0.04 level of significance. Clearly state the parameter you are testing as well as the null and alternative hypothesis. Clearly state your conclusion.

Let p = proportion of customers satisfied at the $\frac{7}{10}$ or better level.

$$H_0 : p = 0.85$$

$$H_a : p < 0.85$$

$$\hat{p} = \frac{175-40}{175} \doteq 0.7714 \quad \Big\} \text{ note: } 40 \text{ gave scores of } 6/10 \text{ or less, so } 175-40 \text{ gave scores of } 7/10 \text{ or more.}$$

All such $\hat{p}$ values are approx. normally distributed with $\mu_{\hat{p}} = p = 0.85$, and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.8(1-0.8)}{175}} \doteq 0.03024$

$$\therefore z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.7714 - 0.85}{0.03024} \doteq -2.6$$

$$P(z \leq -2.6) = \frac{0.47}{100} = 0.0047 .$$

Since the P-value of 0.0047 is less than $\alpha = 0.04$, we reject H_0 in favour of H_a. There is sufficient evidence to conclude that the proportion of customers satisfied at the $\frac{7}{10}$ (or better) level is less than 85%.

Question 6 [10 points]: A car dealer recommends that new cars be brought in after the first 5000 km for tuning. A random sample of the records of 121 new vehicle customers shows that the average mileage reading on the vehicles brought in for the initial tuning is 5037. The sample had a standard deviation of 194 km. Is this sufficient evidence to conclude that, on average, customers drive more than the recommended 5000 km before bringing their vehicle in for the initial tuning? Test at the 0.05 level of significance. Clearly state the parameter you are testing as well as the null and alternative hypothesis. Clearly state your conclusion.

Let μ = mean distance driven prior to bringing car in for first tuning.

$$H_0 : \mu = 5000$$

$$H_a : \mu > 5000$$

$$\bar{x} = 5037$$
$$n = 121$$
$$s = 194$$

Distribution of all such $\bar{x}$ is approx. normal with $\mu_{\bar{x}} = \mu = 5000$, $\sigma_{\bar{x}} = \dfrac{s}{\sqrt{n}} = \dfrac{194}{\sqrt{121}} \doteq 17.64$

$\therefore z = \dfrac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \dfrac{5037 - 5000}{17.64} \doteq 2.1$

$P-value : P(z \geq 2.1) = \dfrac{1.79}{100} = 0.0179$

Since 0.0179 is less than $\alpha = 0.05$, we reject H_0 in favour of H_a. There is sufficient evidence to conclude that, on average, customers drive more than the recommended 5000 km before bringing their vehicles in for initial tuning.

Question 7 [10 points]: Some people are less willing to participate in surveys than others. Does age have anything to do with this willingness to participate? Data from a survey on bank preference is available: the following table gives data on individuals broken down by age group, and within each age group, individuals were categorized as participating or not participating in the survey.

	18-39	40-59	60 and over	Total
participated	573 (554.5)	274 (276.0)	202 (218.5)	1049
did not	64 (82.5)	43 (41.0)	49 (32.5)	156
Total	637	317	251	1205

Is there sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that there is an association between age and willingness to participate in the survey?

H_0: no association between age & willingness to participate in surveys

H_a: there is an association.

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

$$= \frac{(573 - 554.5)^2}{554.5} + \cdots + \frac{(49 - 32.5)^2}{32.5}$$

$$\overset{\circ}{=} 14.5$$

$$df = (r-1)(c-1) = (2-1)(3-1) = 2$$

Critical value in table 24.1 corresponding to $\alpha = 0.05$ and $df = 2$ is 5.99.

Since $\chi^2 = 14.5 > 5.99$ we reject H_0 in favour of H_a: There does appear to be an association between age and willingness to participate in surveys.