

University of Toronto Scarborough

STAB22 Final Examination

December 2010

For this examination, you are allowed two handwritten letter-sized sheets of notes (both sides) prepared by you, a non-programmable, non-communicating calculator, and writing implements.

This question paper has 31 numbered pages, with statistical tables at the back. Before you start, check to see that you have all the pages. You should also have a Scantron sheet on which to enter your answers. If any of this is missing, speak to an invigilator.

This examination is multiple choice. Each question has equal weight, and there is no penalty for guessing. To ensure that you receive credit for your work on the exam, fill in the bubbles on the Scantron sheet for your correct student number (under “Identification”), your last name, and as much of your first name as fits.

Mark in each case the best answer out of the alternatives given (which means the numerically closest answer if the answer is a number and the answer you obtained is not given.)

If you need paper for rough work, use the back of the sheets of this question paper.

Before you begin, two more things:

- Check that the colour printed on your Scantron sheet matches the colour of your question paper. If it does not, get a new Scantron from an invigilator.
- Complete the signature sheet, but *sign it only when the invigilator collects it*. The signature sheet shows that you were present at the exam.

At the end of the exam, you *must* hand in your Scantron sheet (or you will receive a mark of zero for the examination). You will be graded *only* on what appears on the Scantron sheet. You may take away the question paper after the exam, but whether you do or not, anything written on the question paper will *not* be considered in your grade.

1. A study was carried out to determine whether a new diet is effective in reducing cholesterol levels. Twenty subjects were recruited. For each subject, the cholesterol level was measured initially. Each subject was placed on the new diet for one month, and then that subject's cholesterol level was measured again. Which of the following methods of analysis is most appropriate?

- (a) * One-sided matched pairs t test.
- (b) Two-sided two-sample z-test
- (c) Two-sided two-sample t test.
- (d) One-sided two-sample t test.
- (e) Two-sided matched pair t test.

2. A medical treatment has a success probability of 0.7. Three patients will be treated with this treatment. Assuming the results are independent for the three patients, what is the probability that at least one of them will be successfully cured?

$$1 - 0.3^3 = 0.973 \text{ or use Table C } n=3, p=0.7$$

- (a) 0.70
- (b) 0.50
- (c) * 0.97
- (d) 0.99
- (e) 0.21

3. The distribution of the weight of chocolate bars produced by a certain machine has mean of 8.1 oz and a standard deviation of 0.1 oz. The quality control manager plans to take a simple random sample from the production line. How big should the sample size n be so that the sampling distribution of the sample mean has standard deviation 0.02 oz?

$$\sigma/\sqrt{n} = 0.02 = 0.1/\sqrt{n}, \text{ so } \sqrt{n} = 5 \text{ and } n = 25.$$

- (a) * 25
- (b) 100
- (c) 5
- (d) Cannot be determined unless we know the population follows a Normal distribution.
- (e) 10

4. A study was conducted to examine the quality of fish after seven days in ice storage. In the output below, y denotes a measurement of fish quality (on a 10-point scale) and x denotes the time (in hours) after being caught that the fish were placed in ice packing. The scatterplot of y versus x showed a linear relationship. Some output is shown below.

Regression Analysis: y versus x

The regression equation is

$$y = 8.43 - 0.144 x$$

Predictor	Coef	SE Coef	T	P
Constant	8.4318	0.1018	82.86	0.000
x	-0.14402	0.01391	-10.36	0.000

$S = 0.187052$ $R\text{-Sq} = 92.3\%$ $R\text{-Sq}(\text{adj}) = 91.4\%$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3.7524	3.7524	107.25	0.000
Residual Error	9	0.3149	0.0350		
Total	10	4.0673			

What is the value of the correlation between x and y ?

$\sqrt{0.923} = 0.96072889$ and the correlation is negative since the slope is negative.

- (a) -0.923
- (b) 0.923
- (c) 0.961
- (d) * -0.961
- (e) 9.61

5. One of the fish specimens in Question 4 above was placed in ice packing 3 hours after being caught and had a fish quality rating of 6. What is the residual for this fish specimen?

predicted quality is $8.43 - (0.144)(3) = 7.998$ (or 8 using more accurate figures). Residual is $6 - 8 = -2$.

- (a) * -2
 - (b) 8
 - (c) 2
 - (d) 6
 - (e) -8
6. In the analysis in Question 4 above, a plot was also made of residuals against x . What would you expect this residual plot to look like?

- (a) Has a downward non-linear trend.
 - (b) Has a “fanning-out” pattern.
 - (c) Has a linear trend.
 - (d) Has a curved pattern.
 - (e) * Has no pattern.
7. Two alloys, A and B, are used in the manufacture of steel bars. A study was carried out to compare the load capacities of bars made from the two alloys. (Load capacity is measured in tons.) The researchers collected data from simple random samples of 9 bars made from alloy A and 13 bars made from alloy B. Let μ_A denote the population mean load capacity of bars made from alloy A, and let μ_B denote the population mean load capacity of bars made from alloy B. You may assume that load capacities in the samples have normal distributions. Some output from an analysis is shown below. Sample 1 is alloy A, and sample 2 is alloy B.

Two-Sample T-Test and CI

Sample	N	Mean	StDev	SE Mean
1	9	28.50	2.49	0.83
2	13	26.20	1.80	0.50

Difference = mu (1) - mu (2)

Estimate for difference: 2.30000

95% CI for difference: (0.20135, 4.39865)

T-Test of difference = 0 (vs not =): T-Value = 2.37 P-Value = 0.034 DF = 13

Following are three statements based on this analysis. Each statement is either true or false.

- I: The margin of error of the 95 percent confidence interval for $\mu_A - \mu_B$ is greater than 2.2.
- II: The P-value for the test of the null hypothesis $\mu_A = \mu_B$ versus the alternative hypothesis $\mu_A > \mu_B$ is 0.034.
- III: The value 4.0 will be in the 99% confidence interval for $\mu_A - \mu_B$ based on this data.

Which of statements I, II and III is (are) true?

False ME = $(4.39865 - 0.20135)/2 = 2.09865$ not greater than 2.2.

False, it is $0.034/2$

Ans True 4.0 is in the 95% CI and so will be in 99% CI.

Ans : C) only statement III is true.

- (a) only statement I is true.
- (b) more than one of the three statements I, II and III is true.
- (c) * only statement III is true
- (d) none of the three statements I, II and III is true.
- (e) only statement II is true

8. Using the information in Question 7 above, what can we say about the P-value of the test of the null hypothesis $\mu_A = 26$ versus the alternative hypothesis $\mu_A > 26$?

$$t\text{-value} = (28.5-26)/(2.49/\sqrt{9}) = 3.012048193 \text{ with d.f. } 8.$$

p-value for the test is between 0.005 and 0.01

- (a) between 0.01 and 0.02
 - (b) between 0.0025 and 0.005
 - (c) less than 0.0025
 - (d) greater than 0.02
 - (e) * between 0.005 and 0.01
9. Every morning, John tosses a fair coin. If the coin comes up heads, he goes jogging that day, and if it comes up tails, he does not go jogging. What is the probability that he goes jogging on exactly two days in the next week (7 days)?

$$\text{bin}(7,0.5): \text{table C, } k=2: p=0.1641$$

- (a) 0.01
 - (b) 0.23
 - (c) 0.06
 - (d) * 0.16
 - (e) 0.50
10. The random variable X has a binomial distribution with mean 0.8. The standard deviation of X is also 0.8. What is the probability that $X \leq 2$?

mean = $np=0.8$, variance $np(1-p)=0.8^2=0.64$. Divide to get $(1-p)=0.64/0.8=0.8$, so $p=0.2$. Divide eqn for mean by this to get $n=4$. Use Table C with $n=4$, $p=0.2$: $\text{prob}=0.4096+0.4096+0.1536=0.9728$. normal approx no good, but gives $z=(2-0.8)/0.8=1.5$, $\text{prob}\sim 0.935$, correctly the wrong choice.

- (a) 0.50
- (b) 0.95
- (c) * 0.97
- (d) 0.90
- (e) 0.21

11. The scores on an exam have a Normal distribution with mean 65. The third quartile of the distribution of scores is 75. Which of the following numbers is closest to the standard deviation of the distribution of scores?

$$Q3 = \text{mean} + 0.67 \text{ SD} = 75 \text{ and so } \text{SD} = (75-65)/0.67 = 14.92537313$$

- (a) 19
(b) 17
(c) 21
(d) 12
(e) * 15
12. An advertisement for a cold relief claims an 80% success rate. Eight patients selected at random from a large population of patients are given this cold relief. Assume that the claimed success rate is correct. Find the probability that this treatment will be successful for seven or more of these eight patients.

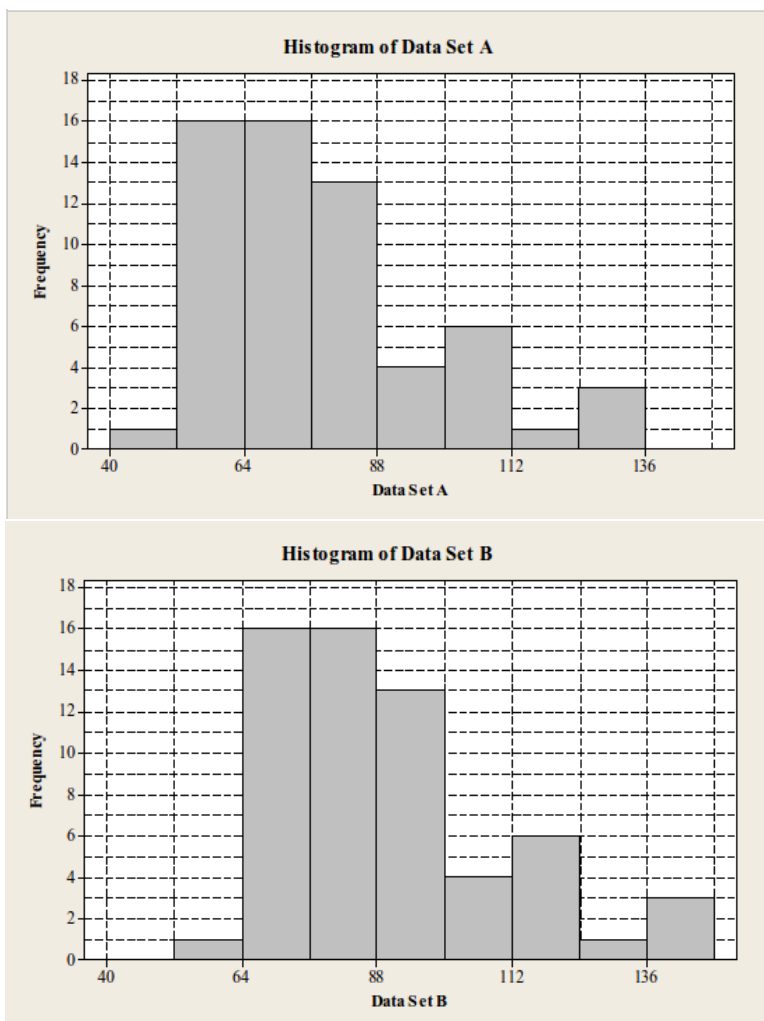
number of successes is $\text{Bin}(8,0.80)$. Can't use Table C, so convert to failures: p becomes $1-0.80=0.20$, # successes becomes $8-(7,8)=0,1$: $0.1678+0.3355=0.5033$.

- (a) 0.80
(b) * 0.50
(c) 0.20
(d) 0.35
(e) 0.10
13. Researchers are studying the yield of a crop in two locations. The researchers are going to compute independent 90% confidence intervals for the mean yield at each location. What is the probability that neither of the researchers' intervals will contain the true mean yields at their location?

$$\text{just } 0.10*0.10=0.01$$

- (a) 0.90
(b) * 0.01
(c) 0.99
(d) 0.05
(e) 0.10

14. The histograms of two data sets (labelled A and B) are shown below:



Descriptive statistics for data set A are also given below:

Descriptive Statistics: Data Set A

Variable	N	Mean	SE Mean	StDev	Q1	Q3
Data Set A	60	78.43	2.57	19.92	62.95	86.58

Which of the following could be the mean of data set B?

B is A shifted 12 units to the right, so mean is 12 bigger

- (a) 56.43
- (b) 58.43
- (c) 78.43
- (d) 98.43
- (e) * 90.43

15. In Data Set A described in Question 14, how many observations are outliers according to the “1.5 times IQR” rule?

1.5 IQR = $1.5(86.58-62.95)=35.445$; $Q1-35.445=27.905$ and $Q3+35.445=122.025$. There are no observations below 27.905, but there are $1+3=4$ observations in the last 2 bars (above 124), plus possibly some from the next bar down.

- (a) cannot tell from the information given
 - (b) no outliers
 - (c) exactly 2 outliers
 - (d) exactly 1 outlier
 - (e) * more than 2 outliers
16. A community is considering building a new skateboard park, and wants to survey people’s opinions. Younger people might have a different opinion about the skateboard park than older people, so, to ensure that each age group is properly represented, separate simple random samples are taken from the different age groups, and the results are combined. What is this sampling method called?

- (a) quota sampling
- (b) multistage sampling
- (c) systematic sampling
- (d) simple random sampling
- (e) * stratified sampling

17. Surveys conducted over the telephone often have problems with nonresponse. Why does this happen?

- (a) Not all households have phones, which causes bias.
- (b) Some people have unlisted numbers and so cannot be part of the survey.
- (c) * In many homes, when the survey company calls, no one is available to answer the phone.
- (d) This is a form of voluntary-response sampling, and randomization should be used instead.

18. A weather forecaster says that the probability of rain on Monday, Tuesday and Wednesday of next week is 0.2, 0.6 and 0.3 respectively. Assuming that the weather forecaster is correct, what is the probability that it rains on at least one of those three days?

$P(\text{no rain}) = (1-0.2) \times (1-0.6) \times (1-0.3) = 0.224$, at least one day is $1-0.224=0.776$

- (a) * 0.8
- (b) 0.6
- (c) 0.9
- (d) 0.7
- (e) 0.5

19. In a casino game, you have probability 0.1 of winning. If you play the game 20 times (independently), what is the probability that you win 2 times or fewer?

from Table C, $n=20$, $p=0.1$: $.1216+.2752+.2802=0.6770$

- (a) * 0.65
- (b) 0.30
- (c) 0.45
- (d) 0.05
- (e) 0.20

20. In the casino game described in Question 19 above, suppose that if you win, you win \$8, and if you lose, you lose \$1. You play the game 15 times (independently). What is the probability that you lose more money than you win?

win 1 or 0 times: table C, $n=15$, $p=0.1$: $.2059+.3432=0.5491$

- (a) * 0.55
(b) 0.80
(c) 0.40
(d) 0.25
(e) 0.10
21. A multiple-choice examination has two parts: Part A and Part B. Part A has 8 questions, each with five choices. Part B has 10 questions, each with four choices. In each question, only one of the choices is the correct answer.

A student has not prepared for the examination at all, so answers each question by picking a choice at random. Let X denote the total number of questions the student gets correct. What is the **standard deviation** of X ?

X is sum of independent $Y=B(8,1/5)$ and $Z=B(10,1/4)$. $\text{Var}(Y)=8(1/5)(4/5)=32/25=1.28$ and $\text{Var}(Z)=10(1/4)(3/4)=30/16=1.875$. So X has variance $1.28+1.875=3.155$ and SD approx 1.8.

- (a) 2.9
(b) * 1.8
(c) 1.4
(d) 2.2
(e) 1.0

22. Aldrin is a highly toxic organic compound that can cause various cancers. Ten water specimens were taken from random locations in Wolf River, downstream from a toxic waste site. The sample mean concentration was 5.109 nanograms per litre. Assuming that the population standard deviation of concentrations is 0.9 nanograms per litre, calculate a 95% confidence interval for the population mean concentration. What is the *lower* limit of this confidence interval?

From Minitab:

One-Sample Z: conc

The assumed standard deviation = 0.9

Variable	N	Mean	StDev	SE Mean	95% CI
conc	10	5.01900	1.10440	0.28460	(4.46118, 5.57682)

lower limit of corresponding t interval is 4.23

- (a) 4.2
- (b) * 4.5
- (c) 5.8
- (d) 5.1
- (e) 5.6

23. A simple random sample was taken of 11 students in a class. The sampled students reported the number of hours per week they studied. The sample mean was 10.26 hours and the sample standard deviation was 6.22 hours. Calculate a 99% confidence interval for the mean study time of all the students in the class. What is the upper limit of the confidence interval?

this should be t
Minitab gives (for t and z):

One-Sample T

N	Mean	StDev	SE Mean	99% CI
11	10.2600	6.2200	1.8754	(4.3163, 16.2037)

One-Sample Z

The assumed standard deviation = 6.22

N	Mean	SE Mean	99% CI
11	10.2600	1.8754	(5.4293, 15.0907)

- (a) 12.8
(b) 15.1
(c) 10.3
(d) * 16.2
(e) 4.3
24. When calculating a 95% confidence interval for a population mean, which is the most important situation where use of the t distribution would be better than use of the normal distribution?
- (a) When the sample size is large.
(b) When the sample standard deviation is unknown.
(c) When the population standard deviation is known.
(d) When the sample size is small.
(e) * When the population standard deviation is unknown.

25. Some communities have installed red-light cameras at intersections. Drivers who fail to stop at a red light are photographed and receive a ticket in the mail. Red-light cameras do cut down on the number of right-angle collisions, but there are more rear-end collisions as drivers brake suddenly to avoid a ticket. A highway agency collected data on rear-end collisions at intersections before and after red-light cameras were installed; in particular only those collisions where there was injury. In each case, the type of injury was recorded, as shown below.

Type of injury	Before camera	After camera
Death/disabling	61	27
Evident injury	210	136
Possible injury	1659	845
Total	1930	1008

What is the marginal proportion of evident injuries?

$$\text{Grand total} = 1930 + 1008 = 2938$$

$$\text{total of evident} = 346$$

$$346/2938 = 0.117$$

- (a) 0.34
 - (b) 0.03
 - (c) 0.66
 - (d) * 0.12
 - (e) 0.85
26. Look again at the data in Question 25. For those accidents that happened after the red-light cameras were installed, what is the conditional proportion of them that resulted in death or disabling injury?

$$27/1008 = 0.0267$$

- (a) * 0.029
- (b) 0.109
- (c) 0.135
- (d) 0.032
- (e) 0.268

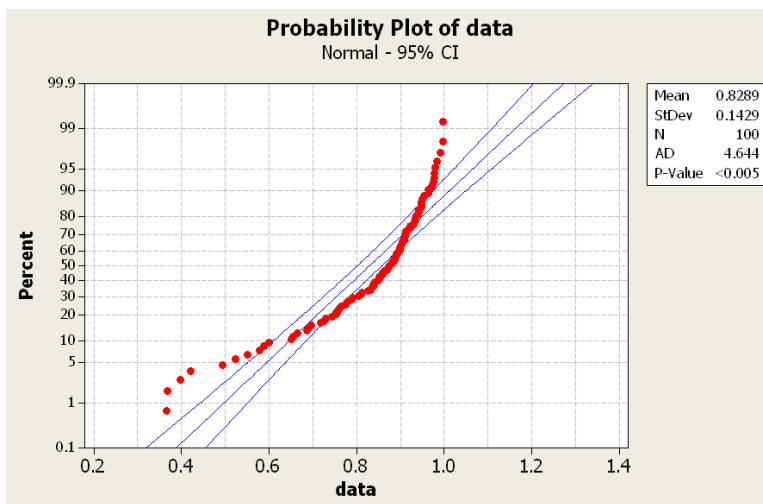
27. Refer again to the situation described in Question 25. The conditional distributions of injury types are very similar before and after the red-light cameras were installed. What would you conclude from this?

- (a) There was not much difference between the numbers of rear-end collisions before and after the red-light cameras were installed.
- (b) There was a difference in the distribution of types of injury suffered before and after the installation of red-light cameras.
- (c) Type of injury and before/after are positively associated.
- (d) Type of injury and before/after are negatively associated.
- (e) * Red-light cameras did not make much difference to the type of injury suffered in rear-end collisions.

28. The correlation coefficient is useful to decide whether:

- (a) A normal distribution adequately describes the data.
- (b) * There is a straight-line relationship between two variables.
- (c) There is any relationship between two variables, possibly a curved one.
- (d) One variable is the cause of another.
- (e) The intercept of the best-fitting straight line is large or small.

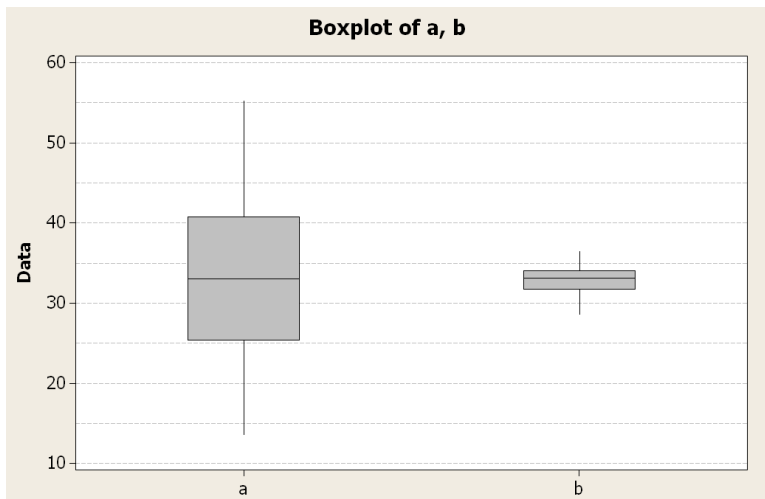
29. A normal probability plot for some data is shown below.



What do you conclude about the shape of the data distribution from this plot?

- (a) approximately normal
- (b) * skewed to the left
- (c) symmetric but not normal
- (d) skewed to the right
- (e) not linear

30. Researchers called medical specialists' offices posing as new patients and requesting appointments for non-urgent problems. The waiting time, in days, was recorded, for each request. Boxplots for two different samples of requests, labelled A and B, are shown below.



The samples are the same size, and the distributions have symmetric shapes, so that the sample mean is very close to the sample median in each case. Which sample, A or B, offers more compelling evidence that the population mean waiting time exceeds 30 days?

- (a) It is impossible to tell from the boxplots.
- (b) The two samples offer about the same evidence
- (c) * Sample B
- (d) Sample A

31. The random variable Y has a mean of 5 and this probability distribution:

Value	4	8
Probability	0.75	0.25

What is the standard deviation of Y ?

mean is $4(0.75)+8(0.25)=3+2=5$

variance is $(4-5)^2(0.75)+(8-5)^2(0.25)=0.75+2.25=3$

sd is $\sqrt{3}=1.73$

or: $4 \cdot \text{Bernoulli}(0.25)+4$, so $\text{sd}=4 \cdot \sqrt{(0.25)(0.75)}=\sqrt{3}$

- (a) 3.0
 (b) 5.0
 (c) 2.2
 (d) *1.7
 (e) cannot be calculated from this information
32. Suppose we wish to calculate a 95% confidence interval for the average amount spent on books by freshmen in their first year at a major university. The interval is to have a margin of error of \$3. Assume that the amount spent on books by freshmen has a Normal distribution with a standard deviation of $\sigma = \$30$. What is the minimum sample size required to achieve this margin of error?

Solve $30=m=z \cdot \sigma / \sqrt{n}=1.96(30) / \sqrt{n}$ for n , or use formula
 $n=(z \cdot \sigma / m)^2 = ((1.96)(30)/3)^2=384.16 \Rightarrow 385$

- (a) * 385
 (b) 60
 (c) 609
 (d) 20
 (e) 3600

33. The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude, and study habits of college students. Scores follow a Normal distribution with mean 115 and standard deviation 25. You suspect that incoming freshmen at your school have a mean that is different from 115, because they may be excited yet anxious about entering college. Let μ be the mean SSHA score of all incoming freshmen at your school. You decide to test the hypotheses $H_0 : \mu = 115$ versus $H_a : \mu \neq 115$. The value of the Z -test statistic based on a random sample of 25 incoming freshmen was found to be $z = 0.24$ (assuming $\sigma = 25$). Calculate the upper limit of the 95% confidence interval for μ , again assuming $\sigma = 25$.

don't know sample mean \bar{x} , but know
 $z=0.24=(\bar{x}-115)/(25/\sqrt{25})=(\bar{x}-115)/5$, so $\bar{x}=116.2$. 95%
 ci uses $z^*=1.96$, so is $116.2 \pm 1.96(25/\sqrt{25})$; upper limit =
 126.

- (a) 125
 (b) * 126
 (c) 131
 (d) 124
 (e) 136
34. The weekly weight losses of all dieters on Diet I have a normal distribution with mean 1.3 pounds and standard deviation 0.4 pound. The weekly weight losses of all dieters on Diet II have a normal distribution with mean 1.5 pounds and standard deviation 0.6 pound. A random sample of 25 dieters on Diet I and another random sample of 36 dieters on Diet II are observed. What is the probability that the average weight loss (\bar{X}_1) for dieters on Diet I will be greater than the average weight loss (\bar{X}_2) for dieters on Diet II?

\bar{x}_1 is normal with mean 1.3, SD $0.4/\sqrt{25}=0.08$. \bar{x}_2 is
 normal with mean 1.5, SD $0.6/\sqrt{36}=0.10$. Difference is also
 normal, mean $1.3-1.5=-0.2$, variance $0.08^2+0.10^2=0.0164$, so SD
 $\sqrt{0.0164}=0.1281$. Prob that difference >0 is prob that
 $Z>(0-(-0.2))/0.1281=1.56)=1-.9406=0.0594$.

- (a) 0.90
 (b) 0.40
 (c) 0.10
 (d) * 0.06
 (e) 0.01

35. The breaking strength of yarn used in the manufacture of woven carpet material is Normally distributed with population standard deviation $\sigma = 2.4$ psi. A random sample of 16 specimens of yarn from a production run were measured for breaking strength, and based on the mean of the sample, \bar{x} , a confidence interval was found to be (128.7, 131.3). What is the confidence level, C, of this interval?

margin of error is $(131.3-128.7)/2=1.3$. This is $z^* \sigma / \sqrt{n} = z^* (2.4)/\sqrt{16}=0.6z^*$, so $z^*=2.167$. Confidence level is bigger than 95% (1.96) and smaller than 99% (2.576), so 97% must be the answer.

- (a) 90%
 - (b) 95%
 - (c) cannot determine from this information
 - (d) 99%
 - (e) * 97%
36. A population has mean 40 and standard deviation 5, and an approximately normal shape. A simple random sample of 10 values is taken from this population. What is the probability that the sample mean is less than 38? Use a suitable approximation.

$z=(38-40)/(5/\sqrt{10})=-1.26$, prob is 0.1038

- (a) * 0.1
- (b) 0.7
- (c) 0.3
- (d) 0.9
- (e) 0.5

37. When individuals are being fitted for hearing aids, it is standard practice to play a tape on which 50 words are pronounced clearly, and for the individual to repeat the words as they are heard. Several different tapes of words are in common use, and it is intended that all the tapes are of equal difficulty.

In an experiment, two such tapes are compared, using a sample of 24 people with normal hearing. Each person hears both tapes (in randomized order), in the presence of background noise, and the number of words correctly identified by each person on each tape is recorded. Some summary data for the experiment is shown below.

	N	Mean	StDev
tape1	24	32.7500	7.4089
tape2	24	29.6667	8.0578
Difference	24	3.08333	6.95274

To determine whether there is any evidence of a difference in difficulty between these two tapes, carry out a suitable test using the information above. What can you say about your P-value?

```
paired t not two-sample (I changed output above to "tape1/2" from
"list 1/2")
```

```
output from both below
```

```
Paired T-Test and CI: list1, list2
```

```
Paired T for list1 - list2
```

	N	Mean	StDev	SE Mean
list1	24	32.7500	7.4089	1.5123
list2	24	29.6667	8.0578	1.6448
Difference	24	3.08333	6.95274	1.41922

```
95% CI for mean difference: (0.14745, 6.01922)
```

```
T-Test of mean difference = 0 (vs not = 0): T-Value = 2.17 P-Value = 0.040
```

```
Two-Sample T-Test and CI: list1, list2
```

```
Two-sample T for list1 vs list2
```

	N	Mean	StDev	SE Mean
list1	24	32.75	7.41	1.5
list2	24	29.67	8.06	1.6

Difference = μ (list1) - μ (list2)
Estimate for difference: 3.08333
95% CI for difference: (-1.41694, 7.58360)
T-Test of difference = 0 (vs not =): T-Value = 1.38 P-Value = 0.174 DF = 45

t-value with 23 df giving P-value 0.02 (one-sided) is 2.177
using 23 df on two-sample test gives P-value for 1.38 of 0.05-0.10
one-sided, 0.10-0.20 two-sided

- (a) between 0.05 and 0.10
 - (b) less than 0.01
 - (c) between 0.10 and 0.20
 - (d) * close to 0.04
 - (e) close to 0.02
38. A polling agency is going to carry out a survey of 1000 people in a large city (chosen as a simple random sample). The survey will ask “who do you intend to vote for as Mayor: Candidate A or Candidate B?”. These are the only two candidates for Mayor. Suppose that 61% of all the city’s residents intend to vote for Candidate A. What is the probability that between 58% and 62% of the people *in the sample* will intend to vote for Candidate A? Use a suitable approximation.

normal approx to binomial
mean of sample proportion 0.61, SD =
 $\sqrt{(0.61)(0.39)/1000}=0.01542$
for 58\%: $z=(0.58-0.61)/0.01542=-1.95$
for 62\%: $z=(0.62-0.61)/0.01542=0.65$
table A: $0.7422-0.0256=0.7166$

- (a) * 0.7
- (b) 0.3
- (c) 0.6
- (d) 0.4
- (e) 0.5

39. A population has 50 members, numbered 01 to 50. We want to select a simple random sample of 4 individuals from this population. An excerpt from a table of random digits is given below.

615672081132087200163299871307

Using the above table of random digits, what is the number of the **4th** individual selected in the sample? (Note that you do not need Table B for this question.)

- (a) 07
 - (b) 08
 - (c) The table of random digits is not long enough.
 - (d) 20
 - (e) * 16
40. At a large university, 35% of all students live on campus, and the remainder live off-campus. Which of the four statements below is the best application of the law of large numbers?
- (a) * If a simple random sample of 1000 students is taken, close to 35% of those students will live on campus.
 - (b) The sampling distribution of the sample proportion, for simple random samples of 1000 students, will have a shape like a normal distribution.
 - (c) A simple random sample is random, so we cannot conclude anything about what the sample proportion might be.
 - (d) If a simple random sample of 100 students is taken, close to 35% of those students will live on campus.

41. A magazine reports that most of the readers who have written to the editor about same-sex marriage are opposed to the idea. This survey is unlikely to be a fair representation of all Canadians' opinions. Why?

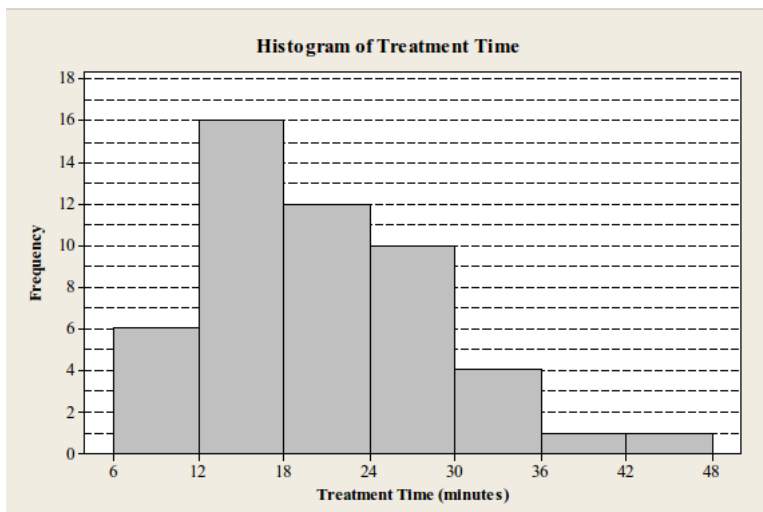
- (a) the magazine has a wide readership among all Canadians
- (b) * the people in the survey were not chosen using randomization
- (c) the question was poorly worded
- (d) there was nonresponse in the survey

42. The discrete random variable X can take only three possible values: 0, 2 and 5. $P(X = 5) = 0.4$, and the mean of X is 3. What is $P(X = 0)$?

mean is $0 \cdot P(X=0) + 2 \cdot P(X=2) + 5 \cdot P(X=5) = 2 \cdot P(X=2) + 5(0.4) = 3$ (given), so $2 \cdot P(X=2) = 1$ and $P(X=2) = 0.5$. The three probabilities have to add up to 1, so $P(X=0) + 0.5 + 0.4 = 1$, and $P(X=0) = 0.1$.

- (a) 0.2
- (b) * 0.1
- (c) 0.4
- (d) 0.5
- (e) 0.3

43. 50 patients visited a health clinic. Their treatment times are shown in the histogram below.



Consider the following three statements, each of which is either true or false:

- I: The median treatment time is between 24 and 30 minutes.
- II: The largest possible value for the mean treatment time is less than 24 minutes.
- III: The smallest possible value for the mean treatment time is greater than 17 minutes.

Which of these statements is (are) true?

- (a) all of the statements are true
 - (b) * statements II and III only
 - (c) statements I and III only
 - (d) statements I and II only
 - (e) one or fewer of the statements is true
44. Which of the following describes the drawing of a *stratified* sample from a population of businesses in a city?
- (a) Use a table of random digits to select the businesses to be in the sample, regardless of where in the city they are.
 - (b) Advertise that businesses are needed for a survey, and have the sample consist of those businesses that respond.
 - (c) * Divide the city into a west part and an east part. Take separate simple random samples from each of the two parts of the city.
 - (d) Divide the city into 20 neighbourhoods. Take a simple random sample of neighbourhoods, and within each sampled neighbourhood, take a further simple random sample of businesses.

45. In a large city, 20% of adults favour a new, stricter recycling program. A simple random sample of 625 adults in the city is to be taken. What is the probability that more than 24% of adults *in the sample* will be in favour of the stricter recycling program? Use a suitable approximation.

sample proportion approx normal with mean 0.2 and sd
 $\sqrt{0.2*0.8/625}=0.016$.
 $z=(0.24-0.2)/0.016=2.5$; prob of greater is $1-0.9938=0.0062$

- (a) 0.003
(b) 0.994
(c) 0.947
(d) * 0.006
(e) 0.001
46. Each professor who drives to UTSC has to tell parking officials about the model and year of their car. It is therefore possible to calculate the mean age of the cars owned by all UTSC professors who drive to campus; this is 3.7 years. A simple random sample is taken from the UTSC professors who drive to campus; for the 100 professors in the sample, the mean age of their cars turns out to be 4.1 years. In this scenario, what is the value of the parameter?

- (a) 4.1 years
(b) unknown
(c) 100
(d) * 3.7 years

47. In the same scenario as Question 46, what is the value of the statistic?

- (a) 3.7 years
(b) 100
(c) * 4.1 years
(d) unknown

48. From a random sample of 10 batteries, a 99% confidence interval was calculated for the mean capacity μ of all batteries. The confidence interval was found to be from 147.44 to 154.56 ampere hours. This interval was calculated using a t distribution. What is the t -test statistic for testing $H_0 : \mu = 156$ versus $H_a : \mu < 156$?

sample mean is middle of interval: $(147.44+154.56)/2=151$. Margin of error is $154.56-151=3.56=t* s/\sqrt{n}$. From Table D, $t^*=3.25$, so $3.56=3.25s/\sqrt{10}$ or $s/\sqrt{10}=1.095$. Test statistic is $t=(\bar{x}-\mu)/(s/\sqrt{n})=(151-156)/(1.095)=-4.565$, closest to -4.5.

- (a) * -4.5
(b) -4.0
(c) -5.0
(d) 3.0
(e) -3.5
49. A sample is to be taken to estimate the mean of a population. Which is the most important reason a multistage sample might be used?

- (a) The sample mean is more likely to be close to the population mean, compared to other types of sampling.
(b) Multistage sampling handles the case where the population is naturally split into different subgroups.
(c) * It is usually more convenient than other types of sampling method.
(d) Other types of sampling do not use randomization.

50. High school dropouts make up 12.1% of all Americans aged 18 to 24. A vocational school that wants to attract dropouts mails an advertising flyer to 25,000 persons between the ages of 18 and 24. If the mailing list can be considered a simple random sample of the population, what is the probability that at least 3100 high school dropouts will receive the flyer?

normal approx to binomial, $n=25000$, $p=0.121$, mean 3025, sd 51.565,
 $z=(3100-3025)/51.565=1.45$, prob of greater $1-0.9265=0.0735$

- (a) 0.35
(b) * 0.07
(c) 0.21
(d) 0.14
(e) 0.28
51. Suppose we are testing a null hypothesis that a population has mean 40, using a sample which has mean 57. How should the alternative hypothesis be written?

- (a) $H_a : \mu \neq 57$
(b) $H_a : \bar{x} \neq 57$
(c) none of the other alternatives is plausible.
(d) * $H_a : \mu \neq 40$
(e) $H_a : \bar{x} \neq 40$

52. The random variable X has the following probability distribution:

Value	1	2	3	4
Probability	0.1	0.5	0.3	0.1

What is the mean of X ?

$$1(0.1)+2(0.5)+3(0.3)+4(0.1)=2.4$$

- (a) 2.5
(b) 2
(c) * 2.4
(d) 3
(e) cannot tell from this information

53. Sarah's grade in her statistics class is determined by three term tests and one final exam. Each term test has a weight of 20 percent and the final exam has a weight of 40 percent on the course grade. The grades in all tests and in the exam are given out of 100. The average of her three term test grades is 78. She wants to get a course grade of at least 80 in order to qualify for an A-. What is the minimum score she needs on the final exam to obtain an A- grade?

If x_1, x_2, x_3 are the term test grades and x_4 the final exam score, the course grade is $0.2x_1+0.2x_2+0.2x_3+0.4x_4 = 80$ i.e. $78*3*0.2+0.4*x_4 = 80$
 $0.4*x_4 = 80 - 78*3*0.2$ And $x_4 = (80 - 78*3*0.2)/0.4 = 83$

or do it backwards: try the possible answers and see which ones give an overall average bigger than 80

Ken says: want $0.6*78+0.4*x=80 \Rightarrow x = (80-46.8)/0.4=83$

- (a) 80
 - (b) 82
 - (c) 84
 - (d) 81
 - (e) * 83
54. Suppose we intend to calculate a 99% confidence interval for a population mean. Which of the statements below best describes the meaning of the "99%"?
- (a) If you draw a sample from this population, there is a 99% probability that the population mean will be between the highest and lowest sample values.
 - (b) * Out of all the samples we could possibly draw from this population, only 1% of them will produce a 99% confidence interval that does not contain the population mean.
 - (c) 99% of the values in the population will be close to the population mean.
 - (d) The probability that the population mean will lie between the endpoints of a 99% confidence interval calculated from one sample is 99%.

55. A statistics class wanted to estimate the weight of a large bag of McDonalds French fries. They bought 30 bags at randomly chosen times at the same McDonald's restaurant. The sample mean was 163.07 grams, with a sample standard deviation of 12.28 grams. The target weight for a large bag of fries is 171 grams. To determine whether this sample offers evidence that this McDonald's restaurant is under-filling its bags of large fries, what can you say about the P-value of a suitable test?

from Minitab
One-Sample T

Test of mu = 171 vs < 171

				95%		
				Upper		
N	Mean	StDev	SE Mean	Bound	T	P
30	163.070	12.280	2.242	166.879	-3.54	0.001

Excerpt from table:

0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01
		0.005	0.0025	0.001	0.0005		
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150
		2.756	3.038	3.396	3.659		

between 0.0005 and 0.001 one-sided. (Normal would give less than 0.0003.)

- (a) less than 0.0006, but cannot be more precise
- (b) greater than 0.002
- (c) between 0.001 and 0.002
- (d) less than 0.0005
- (e) * between 0.0005 and 0.001

56. In the manufacture of glass for airplane windows, samples from the process are taken from time to time, and the strengths of the glass specimens measured. A sample of 16 specimens had a sample mean strength of 30.81 ksi (thousands of pounds per square inch). The process is known to have a standard deviation of 7 ksi. It is desired to assess whether this sample offers evidence that the population mean strength differs from 32 ksi. Using a suitable test, calculate the P-value as accurately as your tables permit. What do you get?

should be a z, two-sided ; Minitab t output included below.

One-Sample Z

Test of mu = 32 vs not = 32
The assumed standard deviation = 7

N	Mean	SE Mean	99% CI	Z	P
16	30.8100	1.7500	(26.3023, 35.3177)	-0.68	0.497

One-Sample T

Test of mu = 32 vs not = 32

N	Mean	StDev	SE Mean	95% CI	T	P
16	30.8100	7.0000	1.7500	(27.0800, 34.5400)	-0.68	0.507

answers below figured out from tables assuming t/z=-0.68 to 2 dp

- (a) * 0.4966
- (b) 0.2483
- (c) greater than 0.25
- (d) greater than 0.50
- (e) 0.7517

57. Farmer Zybicki grows watermelons. When they are ripe, his watermelons have weights that vary according to a normal distribution with mean 21 pounds and standard deviation 2.4 pounds. A supermarket will only buy those melons that weigh between 18 and 23 pounds. What proportion of Farmer Zybicki's watermelons will the supermarket buy?

For 18, $z=(18-21)/2.4=-1.25$, proportion less is 0.1056.
For 23, $z=(23-21)/2.4=2/2.4=0.83$, proportion less is 0.7967,
so proportion between is $0.7967-0.1056=0.6911$.
or: less than 18 is 0.1056, greater than 23 is 0.2033,
between is $1-0.1056-0.2033=0.6911$.

- (a) 0.2
(b) * 0.7
(c) 0.3
(d) 0.1
(e) 0.8
58. A statistics student constructs a large number of confidence intervals for a population mean by computing $\bar{x} \pm 1.96(s/\sqrt{n})$, where \bar{x} denotes the sample mean, s denotes the sample standard deviation and n the sample size. She does this for each of the sample sizes $n = 30, 60, 95, 150$. For which of these sample sizes will the proportion of her confidence intervals that contain the population mean be closest to 0.95?

largest sample size is best because should really be using t^* ,
and largest sample size has t^* closest to 1.96.

- (a) $n = 60$
(b) the sample size does not matter
(c) * $n = 150$
(d) $n = 95$
(e) $n = 30$