University of Toronto Scarborough
STAB22 Final Examination

December 2009

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 25 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. Heart problems can be examined via a small tube (called a catheter) threaded into the heart from a vein in the patient's leg. It is important that the company that manufactures the catheter maintains a diameter of 2 mm. Suppose μ denotes the population mean diameter. Each day, quality control personnel make measurements to test a null hypothesis that $\mu = 2.00$ against an alternative that $\mu \neq 2.00$, using $\alpha = 0.05$. If a problem is discovered, the manufacturing process is stopped until the problem is corrected.

   What, in this context, is a type II error?

   (a) Concluding that the mean catheter diameter is not 2 mm when it is actually less than 2 mm.

   (b) * Concluding that the mean catheter diameter is satisfactory when in fact it is either bigger or smaller than 2 mm.

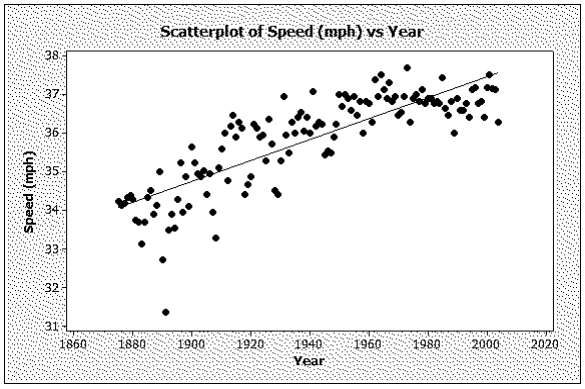
   (c) Using a sample size that is too small.

   (d) Concluding that the mean catheter diameter is 2 mm when it actually is 2 mm.

   (e) Concluding that the mean catheter diameter is unsatisfactory when in fact it is equal to 2 mm.

   A Type II error is failing to reject the null when it is in fact wrong. In this case, the mean is actually not 2, but we cannot reject the null hypothesis that it *is* 2. This is (b).

2. The Kentucky Derby is a famous horse race that has been run every year since the late 19th century. The scatterplot below shows the speed (in miles per hour) of the winning horse, plotted against the year. The regression line is shown on the plot.



   What kind of association do you see between speed and year?

   (a) no association

   (b) negative and non-linear

   (c) negative and linear

   (d) positive and linear

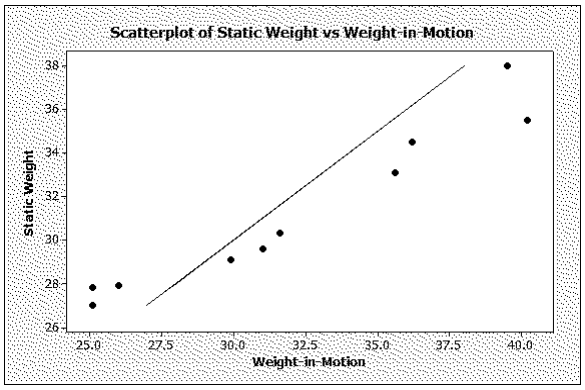
   (e) * positive and non-linear

   The pattern is not absolutely clear, but it does seem that on average the speed is higher when the year number is higher (that is, more recently). So the association is positive. Is it linear? Well, the majority of observations are below the line at the left and right ends, and above the line in the middle, so it looks as if a non-linear association would be a better description of what's going on. This is reasonable because you'd guess that there is a lot of "room for improvement" in the early years, but now, any breaking of the speed record is going to be only by a small amount.

3. Weighing large trucks is a slow business because the truck has to stop exactly on a scale. This is called the "static weight". The Minnesota Department of Transportation developed a new method, called "weight-in-motion", to weigh a truck as it drove over the scale without stopping. To test the

effectiveness of the "weight-in-motion" method, 10 trucks were each weighed using both methods. All weights are measured in thousands of pounds.

A scatterplot of the results is shown below. Superimposed on the scatterplot is the line that the data would follow if the weight-in-motion was always equal to the static weight. Use the scatterplot to answer this question and the one following.



How would you describe the positive association?

(a) not useful since the data are not close to the line
(b) * approximately linear
(c) definitely curved
(d) no association of note

The line on the plot is a bit of a distraction, because the points form a more or less linear pattern, just not around the line shown. (The trend in the points is not obviously a curve, at least).

4. Question 3 described some data on two methods of weighing trucks. The Minnesota Department of Transportation wants to predict the static weight of trucks from the weight-in-motion. Which of the following statements best describes what they can do?

(a) * taking the weights-in-motion and modifying them in some linear way would accurately predict the static weight.
(b) The weights-in-motion can be used to predict the static weights, but a non-linear transformation would have to be applied to do it.
(c) The static weight is accurately predicted by the weight-in-motion itself.
(d) There is no way to use the weights-in-motion to predict the static weight.

Since there is a linear association (just not of the form $y = x$), the static weight can be predicted reasonably well from the weight-in-motion, by multiplying the weight-in-motion by something and adding something else — that is, by using the linear regression equation.

5. A factory hiring people to work on an assembly line gives job applicants a test of manual agility. This test involves fitting strangely-shaped pegs into matching holes on a board. In the test, each job applicant has 60 seconds to fit as many pegs into their holes as possible. For one job application cycle, the results were as follows:

	Male applicants	Female applicants
Subjects	41	51
Mean pegs placed	17.9	19.4
SD of pegs placed	2.5	3.4

The factory wishes to see if there is evidence for a difference between males and females. Which is more appropriate, a matched-pairs t-test or a two-sample t-test? Using the more appropriate test, what P-value do you obtain?

(a) bigger than 0.05

(b) * between 0.01 and 0.02

(c) between 0.005 and 0.01

(d) less than 0.005

(e) between 0.02 and 0.05

This is a two-sample situation, because there is no way of matching up males and females (the giveaway being that there are different numbers of each). So, letting 1 be males and 2 females, we are testing $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$, since we are looking for *any* difference. This is a 2-sided test.

I do this kind of problem by first finding

$$d = \sqrt{\frac{2.5^2}{41} + \frac{3.4^2}{51}} = 0.6157$$

and then getting the test statistic as

$$t = \frac{17.9 - 19.4}{0.6157} = -2.436.$$

We use 40 df, and look up the t without the minus sign in Table D. (Equally good is to do the subtraction the other way around to get a positive result.) The P-value one-sided would be between 0.005 and 0.01; our test is two-sided so we have to double that, between 0.01 and 0.02.

6. A discrete random variable X has the distribution shown below:

Value	0	1	2
Probability	0.2	0.4	0.4

The random variable Y is defined as $Y = 2X + 10$. What is the mean (expected value) of Y?

(a) * 12.4

(b) 10

(c) 1.2

(d) 2.4

(e) there is not enough information

The mean of Y is two times the mean of X, plus 10. The table at the top of the question enables us to work out the mean of X as $0(0.2) + 1(0.4) + 2(0.4) = 1.2$, so the mean of Y is $2(1.2) + 10 = 12.4$.

Alternatively, the distribution of Y has the same probabilities as the distribution of X, but the possible values are $2(0) + 10 = 10$, $2(1) + 10 = 12$, $2(2) + 10 = 14$ instead of 0, 1 and 2. Then you can find the mean of Y directly as $10(0.2) + 12(0.4) + 14(0.4) = 12.4$. Longer, but it still works.

7. A consumer magazine tested 14 (randomly chosen) brands of vanilla yogurt and measured the number of calories per serving of each one. Some Minitab output from the analysis is shown below.

```
One-Sample T: Calories

Variable   N         90% CI
Calories  14   (136.676, 179.038)


One-Sample T: Calories
```

```
Variable   N         95% CI
Calories  14   (132.018, 183.696)
```

Suppose that a diet guide claims that there is an average of 130 calories per serving in vanilla yogurt. You wish to assess the evidence against this claim. From the information above, what can you say about the P-value of your test of significance?

(a) between 0.01 and 0.05

(b) less than 0.01

(c) * less than 0.05

(d) between 0.05 and 0.10

(e) greater than 0.10

The value 130 is inside neither confidence interval, so the P-value of a (two-sided) test of significance is less than 0.10 (from the 90% CI) and also less than 0.05 (from the 95% CI). We don't have the 99% CI so we can't say whether the P-value is less than 0.01 as well. So (c) is the best answer; one of (a) and (b) is certain to be correct, but we can't say which, so we have to eliminate those.

8. A simple random sample of 10 observations is taken from a normally-distributed population with mean 9 and standard deviation 2. Let $\bar{X}$ denote the mean of this sample. Independently of this, another simple random sample of 10 observations is taken from another normally-distributed population with mean 10 and standard deviation 1. Let $\bar{Y}$ denote the mean of this second sample.

   What is the probability that $\bar{X}$ is greater than $\bar{Y}$?

   (a) 0.00

   (b) 0.92

   (c) * 0.08

   (d) 0.52

   (e) 0.48

   The second sample is from a population with a higher mean, so that $\bar{X}$ will usually be less than $\bar{Y}$, and the probability of its being greater ought to be less than a half, eliminating (b) and (d).

   To find out how big the probability is, we have to pull together some things from Section 5.2, a fact about the normal distribution, and rules for means and SDs of differences of random variables. Here's how it goes:

   First, $\bar{X}$ has mean 9, SD $2/\sqrt{10} = 0.632$ and an *exactly* normal shape (we don't need the CLT because if you average a bunch of normally-distributed values, you get *exactly* a normal result. Thus, also, $\bar{Y}$ is also normally distributed with mean 10 and SD $1/\sqrt{10} = 0.316$. At this point you might guess that $\bar{X}$ and $\bar{Y}$ will be pretty close to their means (the latter especially), and so the answer should be pretty small.

   Next, since $\bar{X}$ and $\bar{Y}$ are both normal, so is $D = \bar{X} - \bar{Y}$. This is a crucial observation, since D will be positive if $\bar{X}$ is bigger and negative otherwise. So what we are actually looking for is $P(D > 0)$.

   The rules for means and variances of random variables give the mean of D as $9 - 10 = -1$, the variance of D as $(2/\sqrt{10})^2 + (1/\sqrt{10})^2 = 0.5$ (don't forget to *add!*), and the SD of D as $\sqrt{0.5} = 0.707$.

   Thus to find $P(D > 0)$, find $z = (0 - (-1))/0.707 = 1.41$. About 92% of the normal curve is less than this, so about 8% is greater, which is the answer.

   Tricky.

9. The probability that a US resident has visited Canada is 0.18, the probability that a US resident has visited Mexico is 0.09, and the probability that a US resident has visited both countries is 0.04. Consider the events "has visited Canada" and "has visited Mexico", as applied to a randomly-chosen US resident. Are these two events independent? Are they disjoint? What can you say about these events?

    (a) They are independent but not disjoint

    (b) They are disjoint but not independent

    (c) They are both independent and disjoint

    (d) * They are neither independent nor disjoint

    We can test for independence by seeing whether the multiplication rule works. Since $(0.18)(0.09) = 0.0162$, which is not equal to 0.04, the two events are not independent. To see whether they are disjoint, ask yourself whether both events could happen: that is, are there US residents that have visited both Canada and Mexico? There must be, so the two events are not disjoint either. (Another argument: is there anything stopping someone who has visited Mexico from visiting Canada later? No. So the events are not disjoint.)

10. A researcher expects a relationship between two variables, but finds that the correlation between them is close to zero. The researcher has plenty of data. What is a possible explanation?

    (a) * The relationship is a curve

    (b) The relationship is strongly linear but the correlation happened to come out close to zero

    (c) The relationship is a curved upward trend.

    (d) There cannot actually be a relationship between the variables if the correlation is close to zero.

    The point of this question is that a close-to-zero correlation only means that there is no *straight-line* relationship, but there could still be a curved one. (You might have seen an example of an x and y related by a nice parabola curve, and yet, if the numbers are done correctly, the correlation is *exactly* zero.) So (a) is possible. (d), in the light of what I just said, is not true. Since the researcher has lots of data, the correlation is very unlikely to come out close to zero just by chance, which rules out (b). If the relationship had been a curved upward trend, the correlation would have been positive (just not necessarily close to 1), which rules out (c). That leaves (a) as the only plausible explanation.

11. A shooter fires shots at a target. Each shot is independent, and each shot hits the bull's eye with probability 0.7. Use this information for this question and the next one.

    Suppose the shooter fires 5 shots. What is the probability that the shooter hits the bull's eye exactly 4 times?

    (a) 0.07

    (b) * 0.36

    (c) 0.80

    (d) 0.24

    (e) 0.69

    This is a binomial experiment (repeated independent shots, with the same chance of a bull's eye on each one). So use Table C with $n = 5$, $p = 0.7$ and $k = 4$. Except that Table C doesn't have $p = 0.7$, so use $p = 1 - 0.7 = 0.3$ and $k = 5 - 4 = 1$ instead. Table C gives 0.3601. (This logic is the same as: 4 bull's eyes is exactly 1 non bull's eye, and the probability of not hitting the bull's eye is $1 - 0.7 = 0.3$. You can subtract p from 1 and k from n without thinking too much, or you can go through the reasoning. Either way is good.)

12. Question 11 gave some information about a shooter. Suppose now the shooter fires 50 shots. What is the approximate probability that the shooter hits the bull's eye at least 40 times, using a suitable approximation? (Do not use a continuity correction.)

    (a) 0.24

    (b) * 0.06

    (c) 0.61

    (d) 0.47

    (e) 0.32

    We already concluded that this was a binomial experiment, but now $n = 50$ is too big for Table C, so we have to try a normal approximation. First, the rule of thumb: $np = 50(0.7) = 35$ and $n(1-p) = 50(0.3) = 15$, which are both greater than 10. (The two values just calculated are the mean number of bull's eyes and non-bull's-eyes in 50 shots.) Since 40 or more is a bit bigger than the mean, the shooter should be somewhat unlikely to achieve this, so we can rule out (c) and maybe (d) right away.

    For the normal approximation, the mean is $np = 35$ and the SD is $\sqrt{np(1-p)} = 3.240$. Thus for 40 bull's eyes,

    $$z = \frac{40 - 35}{3.240} = 1.54$$

    and table A gives probability 0.9382 of less and 0.0618 of more. So (b) is closest (which you might have guessed as soon as you saw how big z was).

13. A researcher studies children in school and finds a strong positive linear association between height and reading ability. What would the researcher's best conclusion be?

    (a) The observed association was an accident.

    (b) In any grade, taller children are better readers.

    (c) * There is a lurking variable that explains the correlation.

    (d) Height and reading ability are confounded.

    The obvious conclusion is that taller children are better readers. But that doesn't make sense on the face of it. You would expect *older* children to be better readers, and older children would also be taller. So maybe the cause-and-effect is really of *age* and reading ability, which is what you would guess. Age is a lurking variable.

    If you didn't see that, the fact that there is a strong correlation between height and reading ability overall doesn't mean that this also applies within a grade level (read (b) carefully). This, if you want to be precise, is the restricted-range problem: if you look only at children in, say, grade 2, their heights are not going to vary very much, and so the correlation between height and reading ability within grade 2 will be less than for the data as a whole. So (b) isn't true for this reason either.

14. Whitefish have lengths that vary according to a normal distribution. We want to see if there is any evidence that the population mean length μ is not 9 cm. A simple random sample of 10 whitefish has sample mean length 8.5 cm and sample standard deviation 1.2 cm. Let μ denote the mean length of all whitefish. Test $H_0 : \mu = 9$ against a suitable alternative. What do you conclude, using $\alpha = 0.05$?

    (a) accept the null hypothesis

    (b) reject the null hypothesis because the sample mean is not equal to 9

    (c) * fail to reject the null hypothesis

    (d) reject the null hypothesis

We don't know the population SD, only the sample SD, so we will have to use a t test. We are trying to prove that μ is not 9, so a two-sided alternative $H_a : \mu \neq 9$ is the right thing to do. The test statistic is

$$t = \frac{8.5 - 9}{1.2/\sqrt{10}} = -1.318.$$

With a two-sided alternative: take off the minus sign, look 1.318 up in the 9 df row of Table D, and double the result. So a one-sided P-value would be between 0.10 and 0.15, and the two-sided one is between 0.20 and 0.30. This is safely bigger than 0.05, so we don't reject the null, which is (c). Remember that we never *accept* the null, because we never can prove it to be correct, so (a) cannot be the right answer. Also, just because the sample mean isn't 9 doesn't mean that the null hypothesis should be rejected; sampling variability means that the sample mean could be something like 8.5 (as here) without being able to reject the null.

15. A waiter believes that the distribution of his tips is slightly skewed to the right, with a mean of \$9.60 and a standard deviation of \$5.40. The waiter is interested in the sample mean tip from his next 100 customers (which you can treat as a random sample of all possible customers). Calculate the probability that this sample mean is greater than \$10.00.

    (a) 0.53

    (b) 0.10

    (c) 0.77

    (d) 0.47

    (e) * 0.23

    This is all about a sample mean, so it requires the methods of section 5.2 (dividing by $\sqrt{n}$). So:

    $$z = \frac{10 - 9.60}{(5.40/\sqrt{100})} = 0.74$$

    and the probability of being greater than this is 0.23 (from Table A). Even though an individual tip could be almost anything, it is not *that* likely that the average of 100 tips is over \$10. As before, the answer is going to be at least a bit less than 0.50, which rules out some of the alternatives.

    It doesn't matter that the distribution of individual tips isn't quite normal, because we have the Central Limit Theorem to help us (and the sample of 100 tips is actually large, so the distribution of individual tips could be quite non-normal and we would still be OK with the above calculation).

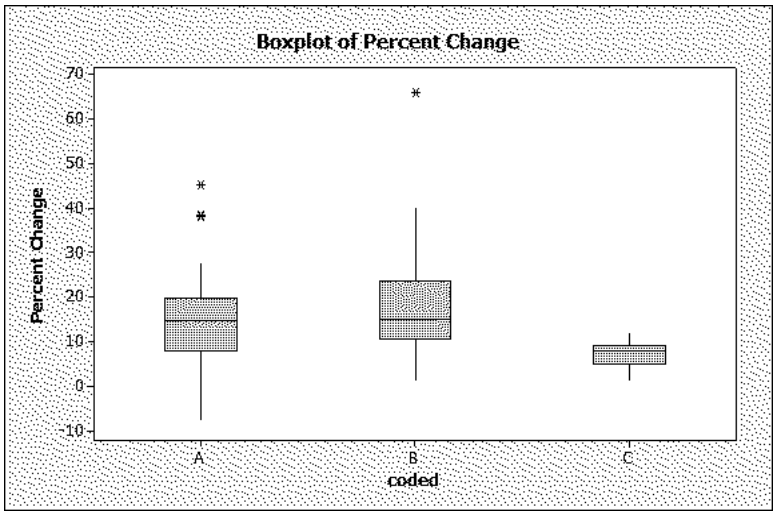
16. Data were collected on the percent change in populations for each US state (comparing the 1990 and 2000 census). The states were classified either as "south/west" (S/W) or "northeast/midwest" (NE/MW).

    Descriptive statistics for the population percent changes for each region are shown below:

Variable	region	N	Mean	SE Mean	StDev	Minimum	Q1	Median
Percent change	S/W	29	18.86	2.32	12.50	1.00	10.50	15.00
	NE/MW	19	7.158	0.677	2.949	1.000	5.000	8.000

Variable	region	Q3	Maximum
Percent change	S/W	23.50	66.00
	NE/MW	9.000	12.000

    Three side-by-side boxplots are shown below, labelled A, B, and C.



Which boxplot is which?

(a) *B is S/W and C is NE/MW

(b) A is NE/MW and B is S/W

(c) A is S/W and C is NE/MW

(d) A is NE/MW and C is S/W

(e) None of the other alternatives is correct.

The easiest way to tackle this one is to look at the *maximum* values, since they are very different for the three boxplots. Boxplot A has a maximum value near 45, B near 70 and C somewhere near 10. That means that S/W must be B and NE/MW must be C.

You can also identify C as NE/MW by looking at the median, which is a little less than 10 on the boxplot and 8 in the data. But looking at the median doesn't distinguish between A and B. For that, you can look at Q3, which is less than 20 in A and about 25 in B. This means that S/W must be B and not A.

17. A large city school system has 20 elementary schools. The school board is considering the adoption of a new policy that would require each student to pass a test in order to be promoted to the next grade. Before adopting this policy, the school board commissions a survey to determine whether parents agree with this plan. Use this information for this question and the 3 questions following.

Which of the following would be a multistage sample?

(a) Randomly select 100 parents of students from within the school board's schools.

(b) * Randomly select 4 of the elementary schools, and randomly select 25 parents of students within each of the schools chosen.

(c) Put a large advertisement in the local newspaper asking parents to visit the school board's website to fill out the survey.

(d) Randomly select 10 parents of students at each elementary school. Mail them a survey, and follow up with a phone call if the survey is not returned within a week.

Look for selection of one thing, and then the selection of something else *within* the first thing. (b) does that. (a) is a simple random sample, (c) is voluntary response. (d) is almost a multistage sample, but the sampling happens at *each* school rather than taking a sample of schools first and just looking at those.

18. In the situation of Question 17, two proposed questions to ask in seeking parents' opinions are these:

    I. Should elementary-school children have to pass high-stakes tests in order to remain with their classmates?

    II. Should schools and students be held accountable for meeting yearly learning goals by testing students before they advance to the next grade?

    Which question do you think will have a higher percentage of parents agreeing with it?

    (a) I

    (b) * II

    (c) the percentages will be about the same

    You can see that the two different questions come from commitees with different agendas. I. is trying to elicit a response like "we can't have our little children being obliged to do high-stakes tests", while II. gets at the idea of children learning what they should be learning, which would be much easier to agree with.

19. In the situation of Question 17, suppose the elementary schools are numbered 01, 02, ..., 20, and it is desired to select a simple random sample of 2 of them. Use the following excerpt from Table B to make the selection.

    `34508602020053161907483273 92`

    Which two schools did you choose?

    (a) 16 and 19

    (b) some other selection of schools not listed here

    (c) * 02 and 16

    (d) the excerpt from Table B is not long enough

    (e) 02 twice

    Pick the digits in twos (because 20 has 2 digits) and throw away two-digit numbers that are either (i) bigger than 20 or (ii) ones you've had before. That means: 34, 50, 86 (reject those), 02 (accept), 02 (no good since we just had it), 00 (no good), 53 (reject), 16 (accept). The two schools we chose are numbered 02 (2) and 16.

20. Suppose a simple random sample of 10 observations $X_1, X_2, \ldots, X_{10}$ is taken from a normally distributed population with mean 20 and standard deviation 2. Let $\bar{X}$ denote the mean of this sample. Independently of this, another simple random sample of 10 observations $Y_1, Y_2, \ldots, Y_{10}$ is taken from another normally distributed population with mean 10 and standard deviation 1. Let $\bar{Y}$ denote the mean of this sample. Use this information for this question and the question following.

    Consider the random variable $X_{10} - Y_1$ (the 10th value in the first sample minus the first value in the second sample). This random variable has a normal distribution. What are its mean and standard deviation (respectively)?

    (a) 0 and 1

    (b) 10 and 3

    (c) 0 and $\sqrt{5}$

    (d) * 10 and $\sqrt{5}$

    (e) 10 and 1

This one requires you not to get frightened. Any of the values in the first sample, such as X_{10}, has a normal distribution with mean 20 and SD 2, and any of the values in the second sample has a normal distribution with mean 10 and SD 1. So the difference has mean $20 - 10 = 10$ and variance $2^2 + 1^2 = 5$, so SD $\sqrt{5}$. Don't forget to *add* when you're finding the variance!

21. Using the information in Question 20, consider the random variable $\bar{X} - \bar{Y}$. What is the distribution of this random variable?

    (a) * normal with mean 10 and standard deviation 0.7

    (b) standard normal

    (c) a t distribution with 9 degrees of freedom

    (d) normal with mean 10 and standard deviation 0.3

    (e) some distribution not mentioned in the other alternatives

    This requires two things: the stuff in Section 5.2 about the distribution of the sample mean, and the rules for means and variances in Section 4.4. First, $\bar{X}$ has mean 20 and SD $2/\sqrt{10}$, and $\bar{Y}$ has mean 10 and SD $1/\sqrt{10}$. Second, the difference $\bar{X} - \bar{Y}$ has mean $20 - 10 = 10$ and variance $(2/\sqrt{10})^2 + (1/\sqrt{10})^2 = 5/10 = 0.5$, so SD $\sqrt{0.5} = 0.707$. The best answer is thus (a).

    The population SDs are known here, so there is no use of sample SDs and therefore no use of a t distribution.

22. In a study of the health risks of smoking, the cholesterol levels were measured for 43 smokers. A stemplot of the results is shown below.

```
Stem-and-leaf of Smokers  N  = 43
Leaf Unit = 10

 1   1  5
 1   1
 3   1  89
 18  2  000000011111111
(7)  2  2223333
 18  2  4444555
 11  2  67
 9   2  888888
 3   3  00
 1   3
 1   3  5
```

What is the interquartile range of the cholesterol levels?

(a) 260

(b) * 50

(c) 60

(d) 210

(e) 5

There are 43 values, so the median is the 22nd (from bottom or top). This means that Q1 is the median of the lowest 21 values (that is, the 11th one from the bottom), and Q3 is the median of the highest 21: the 11th from the top.

Looking at the stemplot, Q1 is the 8th value on the 4th line (210, since there are 7 200s), and Q3 is the 1st value on the 5th line from the end, 260. So the IQR is $260 - 210 = 50$.

Usually there is no shortcut in finding an IQR: you first have to find Q1 and Q3, and then subtract them. So the first reaction on being asked to find an IQR is to think "how do I find Q1 and Q3?".

23. A sample of 40 observations is taken from a population whose standard deviation is known to be 1. We are interested in testing the null hypothesis $H_0 : \mu = 12$ against the alternative $H_a : \mu < 12$. Suppose we decide to reject the null hypothesis if we observe a sample mean $\bar{x}$ that is less than 11.7. What is the probability of making a type I error with this test?

    (a) less than 0.02

    (b) between 0.03 and 0.04

    (c) between 0.04 and 0.06

    (d) greater than 0.06

    (e) * between 0.02 and 0.03

    The probability of making a type I error is just α, so we have to figure out what α would have been. If the sample mean was 11.7 exactly, the test statistic would have been $z = (11.7 - 12)/(1/\sqrt{40}) = -1.90$. This is the correct side for this one-sided alternative, so the P-value is the prob. of less, 0.0287. The question says that we reject when the sample mean is 11.7 or smaller, so we are rejecting when the P-value is 0.0287 or smaller, which is to say that α would have been 0.0287. (A strange choice of α, but that's by the way.)

24. When you are conducting a test about a population mean, in which one of the following situations would you use a t-test (based on the t distribution) instead of a z procedure (based on the normal distribution)?

    (a) * When the population standard deviation is not known

    (b) When the population mean is not known

    (c) When there is some doubt about whether the sampling distribution is normal

    (d) When the sample size is large

    (a) is the situation for which the t-test was derived.

25. Are more people generally admitted to emergency rooms for vehicular accidents on Friday 13th than on other Fridays? A study compared emergency room admissions for vehicular accidents on six different Friday 13th dates, and compared with Friday 6th *in the same months*. The results are as shown:

Year	Month	Friday 6th	Friday 13th
1989	October	9	13
1990	July	6	12
1991	September	11	14
1991	December	11	10
1992	March	3	4
1992	November	5	12

Some Minitab output for these data is as shown. The first part ("two-sample") is appropriate for two independent samples, and the second part ("paired") is appropriate for pairs of measurements which are in some way dependent. Column C1 contains the Friday 6th counts, and column C2 contains the Friday 13th counts.

```
Two-sample T for C1 vs C2

    N   Mean  StDev  SE Mean
C1  6   7.50   3.33      1.4
C2  6  10.83   3.60      1.5

Difference = mu (C1) - mu (C2)
Estimate for difference:  -3.33
95% CI for difference:  (-7.86, 1.20)
T-Test of difference = 0 (vs not =): T-Value = -1.66  P-Value = 0.130  DF = 9


Paired T-Test and CI: C1, C2

Paired T for C1 - C2

            N   Mean  StDev  SE Mean
C1          6   7.50   3.33     1.36
C2          6  10.83   3.60     1.47
Difference  6  -3.33   3.01     1.23


95% CI for mean difference: (-6.49, -0.17)
T-Test of mean difference = 0 (vs not = 0): T-Value = -2.71  P-Value = 0.042
```

Using the more appropriate one of these analyses, what would be an appropriate P-value for this study?

(a) *0.021

(b) the P-value is larger than 0.5 because the difference in sample means should be positive.

(c) 0.130

(d) 0.065

(e) 0.042

This one is paired (they are the 6th and 13th of the *same* months on each line), so the paired (dependent samples) analysis is the appropriate one. The question asked whether there were *more* people admitted to emergency rooms on the 13th (one-sided), but the output is given for a two-sided test, so we have to halve the P-value given there, having checked that we are indeed on the correct side.

There are more accidents in the sample on the 13th dates (the correct side), so we can get our P-value as half of 0.042, that is, 0.021.

26. In a certain population, 10% of the people are beautiful, 10% are intelligent but only 1% are beautiful and intelligent. Use this information for this question and the next one.

For a person picked at random from the population what is the probability that the person is not intelligent?

(a) 0.19

(b) * 0.90

(c) 0.11

(d) 0.01

Just $1 - 0.10 = 0.90$.

27. In the situation of Question 26, suppose a person is picked at random from the population. What is the probability that the person is either beautiful or intelligent? (Two of the alternatives below are reasonable answers to the question; if you mark either of them, you will get this question correct.)

    (a) * Cannot be done using the methods of this course

    (b) 0.99

    (c) * 0.19

    (d) 0.20

    (e) 0.50

    The events "beautiful" and "intelligent" are not disjoint, so for us (a) is the correct answer. If you go delving into Section 4.5, though, you'll find that the answer can be worked out as $0.10 + 0.10 - 0.01$.

28. A simple random sample was taken of 288 teachers in the state of Utah. A 95% confidence interval for the population mean μ is from \$37,500 to \$41,400. Which statement below correctly describes what the confidence interval tells us?

    (a) * If we took many random samples of Utah teachers, about 95% of them would produce a confidence interval that contained the mean salary of all Utah teachers.

    (b) If we took many random samples of Utah teachers, about 95% of them would produce this confidence interval.

    (c) About 95% of Utah teachers earn between \$37,500 and \$41,400.

    (d) We are 95% confident that the mean salary of all teachers in the United States is between \$37,500 and \$41,400.

    (e) About 95% of the teachers in the sample earn between \$37,500 and \$41,400.

    The "confidence" in "confidence interval" is about all possible samples from the population in question. The rather wordy (a) is the only one that gets it right. (d) would have also been correct if it had been talking about the correct population mean, but the population is "teachers in Utah" not "teachers in the whole US".

29. A researcher studied the times taken by mice to learn to run a simple maze. The times were measured in minutes. The researcher took samples of 6 white mice and 6 brown mice. Some output is given below, but unfortunately the important part of the output has been lost. What is the lower limit of a 95% confidence interval for the difference in maze learning times between white mice and brown mice, in that order?

```
Two-sample T for white vs brown

       N   Mean  StDev  SE Mean
white  6  17.00   4.56      1.9
brown  6  16.67   5.05      2.1
```

    (a) 5.9

    (b) * -5.9

    (c) 6.6

    (d) -6.6

    (e) too much of the output was lost: impossible to calculate.

All the information is there. Calculate $d = \sqrt{4.56^2/6 + 5.05^2/6} = 2.778$; then t^* using 5 df is 2.571, so the lower end of the interval is $17 - 16.67 - (2.571)(2.778) = -6.81$. Hmm. I can't remember what we did about this, but if this happens to you: (i) check your calculations, (ii) if you were right, mark the closest answer out of the alternatives listed, which would be (d).

30. "Shingles" is a painful, but not life-threatening, skin rash. A doctor has discovered a new ointment that he believes will be more effective in the treatment of shingles than the current medication. Eight patients are available to participate in the initial trials of this new ointment. Use this information for this question and the three questions following.

    What would be the best way to assess the doctor's belief? (You can assume that the available patients are a mixture of males and females and that there is no difference between males and females in the effectiveness of the current medication and the new ointment).

    (a) Allow the doctor to decide which four patients should receive the new ointment and which four should receive the current medication.

    (b) * Divide the eight patients at random into a treatment group and a control group, with the patients in the control group getting the current medication.

    (c) Give the new ointment to the four oldest patients, because they need it most, and give the current medication to the other patients.

    (d) There are only eight patients, so test the new ointment on them all.

    Two important things: (i) use randomization and (ii) protect yourself against a placebo effect by having a control group. This points to (b).

31. In the situation of Question 30, how could this experiment be made most nearly double-blind?

    (a) Employing a nurse to apply the new ointment and current medication, rather than the doctor.

    (b) * Giving the current medication in the form of an ointment.

    (c) Having a control group.

    (d) Using randomization.

    "Double-blind" means that the person administering the treatment/placebo doesn't know which is being administered. Having both medications be ointments would certainly help in this regard. The other things are desirable, but they are not related to making the study double-blind.

32. In the situation of Question 30, which of the following could be a response variable?

    (a) The gender of the patient.

    (b) The age of the patient.

    (c) * The time it takes for the skin rash to disappear.

    (d) Whether or not a patient receives the new ointment.

    (e) Whether or not the patient survives for a year.

    A response is an outcome, one that would enable us to discover whether the new ointment is effective. (c) fits the bill. (a), (b) and (d) are explanatory variables, and (e) is not relevant because shingles is not life-threatening (so anyone who doesn't survive died from something else).

33. In the experiment of Question 30, it turns out that the new ointment does appear to be effective, but that the new ointment is more effective for younger patients. A second study is planned, and many more patients are available. Based on the knowledge gained from the first study, what would you do in the second study?

(a) Use age as the response variable.

(b) * Use treatment and control groups and also ensure that the patients in the two groups are similar in terms of age.

(c) Decide which patients receive the new ointment using a multistage sample.

(d) Design the second study in the same way as the first.

Designing the second study the same way as the first is not bad, because using randomization will probably give treatment and control groups of about the same average age, but doing some kind of matching on age would be better. (You might, for example, arrange the subjects into pairs of people the same age, then flip a coin to decide which subject gets the treatment and which the placebo.)

34. You roll a (fair, 6-sided) die. If you get a 6 on the first roll, you win $100. If not, you roll a second time, and if you get a 6 on the second roll, you win $50. Otherwise, you win nothing. What are your mean winnings from this game?

(a) $10.50

(b) $30.00

(c) $75.00

(d) $25.00

(e) * $23.50

This is not binomial, since you don't always roll twice, so you have to tackle it from first principles. To win on the first roll (and thus win $100), you must roll a 6 first time, which has probability 1/6. To win $50, you must roll a 6 on the second roll, *having failed to roll a 6 on the first roll* (otherwise you would have won $100). This has probability $(5/6)(1/6) = 5/36$, since the two rolls of the die are independent. The probability of failing to roll any 6's is 1 minus the sum of those two, but you don't need to work it out since the winnings associated with it is $0.

So the mean winnings is

$$100(1/6) + 50(5/6)(1/6) = 23.61$$

and you mark (e).

It is tempting to think of this as binomial: the probability of rolling at least one six in two rolls, but then you have to associate the right winnings with the places where the successes come.

35. A smelt is a type of small fish. Suppose that smelt lengths have a standard deviation of 2 cm. A simple random sample of 30 smelts has a sample mean length of 7.5 cm. Use this information for this question and the next one.

What is the upper limit of a 95% confidence interval for the mean length of all smelts?

(a) * 8.22 cm

(b) 8.50 cm

(c) 7.50 cm

(d) 8.36 cm

(e) 6.78 cm

A standard one-sample z since σ is known. The upper limit is

$$7.5 + (1.96)(2/\sqrt{30}) = 8.22.$$

36. In the situation of Question 35, and using the same data, suppose now that you are testing the null hypothesis $H_0 : \mu = 8.5$ cm against $H_a : \mu < 8.5$ cm, where μ is the mean length of all smelts. What do you conclude from your test? Use $\alpha = 0.01$.

    (a) * reject the null hypothesis

    (b) fail to reject the null hypothesis

    (c) cannot do test without knowing what μ is

    (d) need to use a larger value of α

    (e) accept the null hypothesis

    You can't use the answer to the previous question since the confidence level (95%) and $\alpha = 0.01$ do not correspond. Further, the test is one-sided. So we have to actually do this one.

    $$z = \frac{7.5 - 8.5}{(2/\sqrt{30})} = -2.74.$$

    So z is correctly negative, and the P-value is the small probability of being less than -2.74, which is 0.0031. Even using $\alpha = 0.01$, this is small enough to reject the null. The mean length of all smelts is, we conclude, less than 8.5 cm.

37. Out of all American workers, 56% have a workplace retirement plan, 68% have health insurance, and 49% have both benefits. For a randomly sampled worker, are "has a workplace retirement plan" and "has health insurance" independent events?

    (a) * No, more workers have both benefits than you would expect if the two events were independent.

    (b) No, fewer workers have both benefits than you would expect if the two events were independent.

    (c) There is not enough information to decide.

    (d) Yes.

    To test the independence, see whether the multiplication rule works. If it does, the proportion having both benefits should be $0.56 \times 0.68 = 0.3808$, which is not equal to 0.49. So the answer is one of (a) and (b). If the independence were true, only about 38% of workers would have both benefits, but it's actually more than that; there's a kind of positive relationship in that a worker who has one benefit is more likely to have the other one as well.

    Another way to tackle this is as a two-way table, in the spirit of Section 2.5. Assuming there were 100 people, the table would look like this:

		Health		
		Yes	No	Total
Retirement	Yes	49		56
	No			
	Total	68		

    Filling in the missing values would give

		Health		
		Yes	No	Total
Retirement	Yes	49	7	56
	No	19	25	44
	Total	68	32	100

    and you can see rather clearly that people with one benefit are likely to have the other one as well, and people without one probably don't have the other either. (Compare the conditional distributions to be precise.)

38. A dogfish is a kind of shark. A simple random sample of 10 of one type of dogfish produced a sample mean length of 1.2m and a sample standard deviation of 0.1m. Dogfish lengths are believed to have a shape close to a normal distribution. A 95% confidence interval for the mean length of all dogfish of this type has what lower limit?

    (a) 1.099
    (b) need to know the population mean
    (c) 1.138
    (d) * 1.128
    (e) need to know the population standard deviation

    There were originally a lot of questions about smelts, so I edited this one to be about dogfish. The population SD is not known, so it will be a t rather than a z; the bit about "close to normal" indicates that t will be at least reasonably accurate here. With 9 df, $t^* = 2.262$, so the lower limit of the interval is

    $$1.2 - 2.262(0.1/\sqrt{10}) = 1.128.$$

39. A and B are disjoint events with $P(A) = 0.2$ and $P(B) = 0.7$. Use this information for this question and the next one.

    What is the probability that A and B both happen?

    (a) there is not enough information
    (b) * 0.00
    (c) 0.90
    (d) 0.14

    Don't even think about multiplying here. Two disjoint events *cannot* both happen, so the answer is zero.

40. In the situation of Question 39, what is the probability that exactly one of the two events occurs?

    (a) * 0.62
    (b) there is not enough information
    (c) 0.76
    (d) 0.90

    Since the events are disjoint, the probability that either A happens or B happens is $0.2+0.7 = 0.9$. The two events cannot both happen, so either-or is the same as "exactly one". Hmm.

41. A continuous random variable X has a probability density function $f(x)$ that is equal to 1 for $0 < x < 1$. What is $P(3/4 < X < 1)$?

    (a) 3/4
    (b) 1
    (c) * 1/4
    (d) 0

    X has a uniform distribution on the interval from 0 to 1, so the area under the density function between 3/4 and 1 is the base $(1 - 3/4 = 1/4)$ times the height (1). Or, simpler, the density function is a square, and the piece of it between 3/4 and 1 is the rightmost quarter of it.

42. A discrete random variable X has this distribution:

Value	2	3	4
Probability	0.1	0.5	0.4

What is the mean (expected value) of X?

(a) * 3.3

(b) less than 2

(c) 3.6

(d) 3.0

(e) 2.5

Just $2(0.1) + 3(0.5) + 4(0.4) = 3.3$.

43. The residents in a certain street are concerned that vehicles are driving too fast along their street, where the speed limit is 40 km/h. The residents collect a large random sample of vehicles driving along their street, and obtained a sample mean of 42 km/h (2 km/h over the speed limit). Let μ denote the mean speed of all vehicles driving along the street. For a test of $H_0 : \mu = 40$ against $H_a : \mu > 40$, the P-value is less than 0.05.

What would be the most appropriate reaction to these results?

(a) Because the P-value is small, there is evidence that cars are travelling dangerously fast on the street.

(b) * An average of 2 km/h over the speed limit is not of practical importancein this case.

(c) The P-value is small, so there is no evidence that vehicles are travelling too fast on average.

(d) The test says that the average of 2 km/h over the speed limit must be of practical importance.

(c) is false, and (d) is also false, because a test says nothing about practical importantance. (a) is almost true, but "dangerously fast" is not a conclusion that can be drawn from the small P-value (we can only say that the mean is greater than 40, not how much greater. (b) seems the best answer, since 2 km/h hardly seems something to get excited about.

44. A random survey of cars parked in a university's parking lot revealed the following information about the country of origin of the car and whether its driver was a student or a staff member:

	Driver	
Origin	Student	Staff
America	107	105
Europe	33	12
Asia	55	47

Use this information for this question and the two following.

In the joint distribution, what proportion of cars are from Europe and driven by a student?

(a) 0.17

(b) 0.25

(c) 0.03

(d) 0.73

(e) * 0.09

33 out of the grand total, which is 359. $33/359 = 0.0919$.

45. Using the information from Question 44, in the marginal distribution of car origins, what proportion of cars are from America?

(a) 0.55

(b) 0.64

(c) 0.75

(d) 0.72

(e) * 0.59

Total up the cars from America: $107 + 105 = 212$, out of the grand total 359, is 0.59.

46. Look again at the information in Question 44. Given that a driver is a member of staff, what proportion of these drivers drive a car from Asia?

(a) 0.35

(b) 0.28

(c) * 0.29

(d) 0.54

(e) 0.46

This is a conditional distribution question. Only consider the staff (ignore the students); 47 of those drive Asian cars, out of 164 total, which is 0.29.

47. Suppose two fair 6-sided dice are rolled, and the numbers of spots on the uppermost faces are recorded. Use this information for this question and the next one.

What is the probability that the total number of spots is exactly 5?

(a) * 4/36

(b) 2/36

(c) 1/36

(d) 4/12

There are $6 \times 6 = 36$ equally likely ways the dice could come up. The ways of these that give a total of 5 are 1 on the first die and 4 on the second, 2 and 3, 3 and 2 or 4 and 1. This is 4 equally likely ways out of 36.

48. In Question 47, two fair dice were rolled. What is the probability of getting at least one 5? (That is, what is the probability that at least one of the dice shows 5 spots?)

(a) 10/36

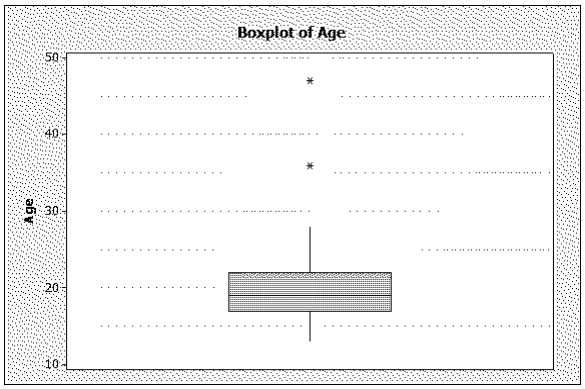
(b) * 11/36

(c) 2/6

(d) 5/6

Do your standard "at least one" thing here: the probability that neither die shows a 5 is $(5/6)(5/6) = 25/36$, so the probability of at least 1 5 is $1 - 25/36 = 11/36$. (This looks, and is, binomial with $n = 2$ and $p = 1/6$, but the value of p is not in Table C.)

49. A company monitors accidents at rock concerts. When someone is injured due to "crowd crush", the company collects information about the victim. A boxplot of the victims' ages is shown below, along with some other information about the ages. Use these to answer this question and the next one.



Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Age	66	0	20.136	0.628	5.099	13.000	17.000	19.000	22.000

Variable	Maximum
Age	47.000

How would you describe the shape of the distribution?

(a) Like a normal distribution.

(b) * Skewed to the right.

(c) Skewed to the left.

(d) Approximately symmetric.

Q3 is further above the median than Q1 is below, and the maximum is further above Q3 than the minimum is below Q1. Also, the mean is bigger than the median. This all points to a right-skewed shape.

50. In Question 49, some information was given about the ages of victims of "crowd crush". According to the usual rule, how high an age would be considered an outlier?

(a) 28.2

(b) * 29.5

(c) 25.5

(d) 22.0

(e) 19.0

$IQR = 22 - 17 = 5$. Go 1.5 times this above Q3, which is $22 + 7.5 = 29.5$. Anything above this is an outlier.

51. A simple random sample of 17 cows were given a special feed supplement to see if it will promote weight gain. The sample mean weight gain was 55.5 pounds, and the population standard deviation was known to be 10 pounds. The population distribution of weight gain is very close to normal. Calculate a 95% confidence interval for the population mean weight gain for all cows. What is the lower limit of this confidence interval, to the nearest pound?

(a) 56

(b) 50

(c) 60

(d) 61

(e) * 51

Since the population SD is known, use z. Even though the sample size isn't very big, the central limit theorem should be OK (the population distribution is close to normal).
So: $z^* = 1.96$, and the margin of error is $1.96(10/\sqrt{17}) = 4.75$. The lower confidence limit is 55.5-4.75=50.75, which is closest to 51 pounds.

52. Question 51 discussed an investigation of the effect of a special feed for cows on weight gain. Use the information in that question, except suppose now that the figure of 10 pounds is the *sample* standard deviation. What is the lower limit of a 95% confidence interval for the population mean, to the nearest pound?

(a) 60
(b) * 50
(c) 56
(d) 61
(e) 51

This time σ is not known, so we have to use s and t^*. With 16 df, t^* for a 95% CI is 2.120, and the margin of error is $2.120(10/\sqrt{17}) = 5.14$, so the lower limit is $55.5 - 5.14 = 50.36$ which is closest to 50. You might have guessed that not knowing σ would make the interval a bit wider and the lower limit a bit smaller.

53. A psychology researcher is testing how long it takes for rats to run through a certain maze (to reach the food at the other end). The researcher tests a maze with 10 rats, obtaining a sample mean of 48 seconds. Other researchers using similar mazes have found a mean time of 55 seconds, so the first researcher carries out a test of $H_0 : \mu = 55$ against $H_a : \mu \neq 55$ for his data, and obtains a P-value of 0.08, which is not smaller than the $\alpha = 0.05$ chosen.

What would be the best conclusion from these data?

(a) Because the P-value is greater than α, we have proved that the population mean is equal to 55 seconds.
(b) * It might still be true that the mean time is not equal to 55 seconds, but the test does not have enough power to prove this.
(c) The researcher should use a smaller sample size.
(d) The P-value is small enough to conclude that the mean is not 55 seconds.

The P-value is not small enough to reject the null and conclude that $\mu \neq 55$. So (d) is wrong. (a) is also no good because we never prove that a null hypothesis is correct. A smaller sample size would, for the same sample mean, give a *larger* P-value (more information in larger sample sizes), so (c) is no good. What about (b)? That pretty much sums up what happened, whether or not you like the word "power". The sample size ought to be bigger rather than smaller.

54. Suppose a simple random sample of size 10 is drawn from a normal population with mean $\mu = 20$ and standard deviation $\sigma = 2$. Denote the 10 sampled values $X_1, X_2, \ldots, X_{10}$, let $\bar{X}$ denote the sample mean, and let S denote the sample standard deviation. Use this information for this question and the two following questions.

The random variable $(X_1 - 20)/2$ has a normal distribution. What are its mean and standard deviation (respectively)?

(a) 0 and 2
(b) * 0 and 1

(c) 0 and $1/\sqrt{10}$

(d) 20 and 2

(e) 20 and $2/\sqrt{10}$

X_1 is an individual value drawn from this population. It has been standardized, so its mean and SD are 0 and 1. (Or use the rules of mean and variance to get the same thing.)

55. Using the information in Question 54, the random variable $\bar{X} - 10$ has a normal distribution. What are its mean and standard deviation (respectively)?

(a) * 10 and $2/\sqrt{10}$

(b) 0 and 1

(c) 0 and $2/\sqrt{10}$

(d) 10 and 2

(e) 0 and 2

$\bar{X}$ has mean 20 and SD $2/\sqrt{10}$. Subtracting 10 reduces the mean to $20 - 10 = 10$ but doesn't change the SD.

56. Using the information in Question 54, consider the random variable $(\bar{X} - 20)/(S/\sqrt{10})$. What distribution does this have?

(a) a standard normal distribution

(b) some distribution not given in the other alternatives

(c) a t-distribution with 19 degrees of freedom

(d) a normal distribution, but not a standard normal distribution

(e) * a t-distribution with 9 degrees of freedom

This is exactly how you get a t test statistic. This one is based on a sample of size 10, so it has 9 df.

57. The company that makes a well-known brand of chocolate-chip cookies advertises that the 500g bags contain an average of "at least 1100 chocolate chips". To test this claim, a dedicated group of students purchased (a random sample of) 16 500g bags of these cookies and counted the number of chocolate chips in each bag. The students found a sample mean of 1188 chocolate chips per bag. The population standard deviation is known to be 100, from knowledge of the manufacturing process.

Calculate the P-value for the test of $H_0 : \mu = 1100$ against $H_a : \mu \neq 1100$, where μ is the mean number of chocolate chips in all 500g bags of these cookies. What do you get?

(a) between 0.001 and 0.0025

(b) * less than 0.0006

(c) between 0.002 and 0.005

(d) greater than 0.005

(e) less than 0.0003

You might reasonably claim that the alternative should be "less than", since that's what the students are trying to prove, but anyway, let's answer the question as asked.

We already have the hypotheses, so we proceed immediately to the test statistic, which is a z, the population SD being known to be 100:

$$z = \frac{1188 - 1100}{(100/\sqrt{16})} = 3.52.$$

This is off the end of Table A, so the probability of being less than this is 0.9997 or bigger, so the probability of being greater is 0.0003 or less. The P-value is twice that: "less than 0.0006". We can't be more accurate than that with the table we have. A more accurate answer, from software, for the P-value is 0.000432. The population mean number of chocolate chips per bag is not 1100. (From the data, it looks as if it's more than 1100. You could find a confidence interval to assess that.)

Since the sample mean is bigger than 1100, this sample offers no evidence that the population mean is *less than 1100*: we are on the "wrong side" for that.

58. According to real estate data, 21% of homes for sale have swimming pools. A simple random sample of 5 homes for sale is taken. What is the probability that at least one of the homes in the sample has a swimming pool?

    (a) 0.31
    (b) 0.21
    (c) * 0.69
    (d) 0.79
    (e) 0.50

    The simplest way is to recognize that "at least one" is "not zero". The homes in the simple random sample have swimming pools or not independently of each other, so the probability that none of them have a swimming pool is $(1 - 0.21)^5 = 0.31$, so the probability that at least one does is $1 - 0.31 = 0.69$.

    Alternatively, the number of homes in the sample that have a swimming pool has a binomial distribution with $n = 5$ and $p = 0.21$. This value of p isn't in Table C (which should alert you that there is another way to do it), but if you pretend that $p = 0.20$, you get 0.3277 as the probability that none of the houses have a swimming pool and $1 - 0.3277 = 0.6723$ (or add up the entries for 1, 2, 3, 4, 5) as the probability that at least one does. Since the true probability of a house having a swimming pool is a little higher than this, the answer for "at least one" should be a little higher too, pointing you at 0.69.

59. What does the Central Limit Theorem say?

    (a) * The sampling distribution of the sample mean is approximately normal for large samples.
    (b) If the sample is large, the population from which the sample comes is approximately normal.
    (c) The sampling distribution of the sample standard deviation is approximately normal for large samples.
    (d) The sample mean is likely to be very close to the population mean if the sample is large.

    The Central Limit Theorem tells you that the sample mean will have an approximate normal distribution if the sample size is large, regardless of the shape of the population. (The randomness here comes from the sampling variability: a different sample will have a different sample mean, but the way in which the sample means vary will have approximately a normal shape.)

    The CLT doesn't say anything about the shape of the population, or about the standard deviation. (d) is the Law of Large Numbers.

60. A 2004 survey of the world's countries found a strong positive correlation between the percentage of the country's population regularly using cellphones and life expectancy at birth (in years). What can you conclude from this?

    (a) Cellphone use and life expectancy are confounded

(b) In countries where cellphone use is low, cellphone use should be encouraged in order to increase life expectancy

(c) Using cellphones is good for your health

(d) * Some other variable is the cause of the high correlation

This correlation doesn't make any sense on the face of it, so go looking for another explanation. One might be standard of living: a richer country is likely to have more cellphone use and better health care. (It doesn't matter what you come up with: anything will direct you towards (d).) (b) and (c) could be true if there really is a cause and effect. Confounding happens with two *explanatory* variables: because they are highly correlated with each other, you can't tell which one is the cause of the response. So that doesn't apply here.