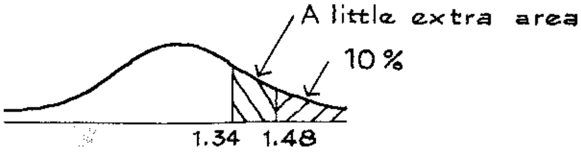


area to the right of 1.48 is 10%. And 1.34 is just to the left of 1.48. So the area to the right of 1.34 is a little more than 10%.

Exercise Set F

1. Find the area under Student's curve with 5 degrees of freedom:

   (a) to the right of 2.02
   (b) to the left of −2.02
   (c) between −2.02 and 2.02
   (d) to the left of 2.02.

2. The area to the right of 4.02 under Student's curve with 2 degrees of freedom is

   less than 1% between 1% and 5% more than 5%

   Choose one option, and explain.

3. True or false, and explain: to make a *t*-test with 4 measurements, use Student's curve with 4 degrees of freedom.

4. Each (hypothetical) data set below represents some readings on span gas. Assume the Gauss model, with errors following the normal curve. However, bias may be present. In each case, make a *t*-test to see whether the instrument is properly calibrated or not. In one case, this is impossible. Which one, and why?

   (a) 71, 68, 79
   (b) 71, 68, 79, 84, 78, 85, 69
   (c) 71
   (d) 71, 84

5. A new spectrophotometer is being calibrated. It is not clear whether the errors follow the normal curve, or even whether the Gauss model applies. In two cases, these assumptions should be rejected. Which two, and why? The numbers are replicate measurements on span gas.

   (a) 71, 70, 72, 69, 71, 68, 93, 75, 68, 61, 74, 67
   (b) 71, 73, 69, 74, 65, 67, 71, 69, 70, 75, 71, 68
   (c) 71, 69, 71, 69, 71, 69, 71, 69, 71, 69, 71, 69

6. A long series of measurements on a checkweight averages out to 253 micrograms above ten grams, and the SD is 7 micrograms. The Gauss model is believed to apply, with negligible bias. At this point, the balance has to be rebuilt, which may introduce bias as well as changing the SD of the error box. Ten measurements on the checkweight, using the rebuilt scale, show an average of 245 micrograms above ten grams, and the SD is 9 micrograms. Has bias been introduced? Or is this chance variation? (You may assume that the errors follow the normal curve.)

7. Several thousand measurements on a checkweight average out to 512 micrograms above a kilogram; the SD is 50 micrograms. Then, the weight is cleaned. The next 100 measurements average out to 508 micrograms above one kilogram; the SD is 52 micrograms. Apparently, the weight got 4 micrograms lighter. Or is this chance variation? (You may assume the Gauss model with no bias.)
   (a) Formulate the null and alternative hypotheses as statements about a box model.
   (b) Would you estimate the SD of the box as 50 or 52 micrograms?
   (c) Would you make a z-test or a t-test?
   (d) Did the weight get lighter? If so, by how much?

The answers to these exercises are on p. A94.

Technical notes. (i) The term "degrees of freedom" is slightly baroque; here is the idea behind the phrase. The SE for the average depends on the SD of the measurements, and that in turn depends on the deviations from the average. But the sum of the deviations has to be 0, so they cannot all vary freely. The constraint that the sum equals 0 eliminates one degree of freedom. For example, with 5 measurements, the sum of the 5 deviations is 0. If you know 4 of them, you can compute the 5th—so there are only 4 degrees of freedom.

(ii) Why use SD^+? Suppose we have some draws made at random with replacement from a box whose SD is unknown. If we knew the average of the box, the r.m.s. difference between the sample numbers and the average of the box could be used to estimate the SD of the box. However, we usually do not know the average of the box and must estimate that too, using the average of the draws. Now there is a little problem. The average of the draws follows the draws around; deviations from the average of the *draws* tend to be smaller than deviations from the average of the *box*. SD^+ corrects this problem.

7. REVIEW EXERCISES

Review exercises may cover material from previous chapters.

1. True or false, and explain:
   (a) The P-value of a test equals its observed significance level.
   (b) The alternative hypothesis is another way of explaining the results; it says the difference is due to chance.

2. With a perfectly balanced roulette wheel, in the long run, red numbers should turn up 18 times in 38. To test its wheel, one casino records the results of 3,800 plays, finding 1,890 red numbers. Is that too many reds? Or chance variation?
   (a) Formulate the null and alternative hypotheses as statements about a box model.
   (b) The null says that the percentage of reds in the box is ______. The alternative says that the percentage of reds in the box is ______. Fill in the blanks.

(c) Compute z and P.
(d) Were there too many reds?

3. One kind of plant has only blue flowers and white flowers. According to a genetic model, the offsprings of a certain cross have a 75% chance to be blue-flowering, and a 25% chance to be white-flowering, independently of one another. Two hundred seeds of such a cross are raised, and 142 turn out to be blue-flowering. Are the data consistent with the model? Answer yes or no, and explain briefly.

4. One large course has 900 students, broken down into section meetings with 30 students each. The section meetings are led by teaching assistants. On the final, the class average is 63, and the SD is 20. However, in one section the average is only 55. The TA argues this way:

> If you took 30 students at random from the class, there is a pretty good chance they would average below 55 on the final. That's what happened to me—chance variation.

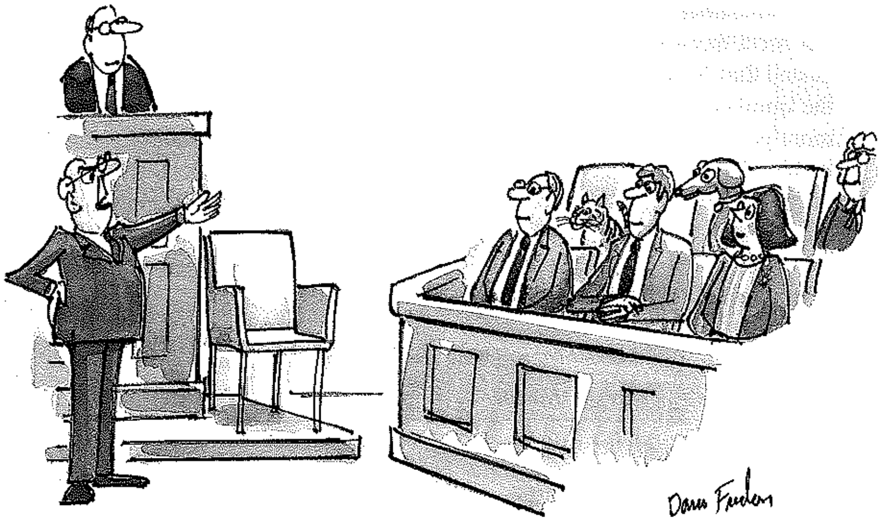
Is this a good defense? Answer yes or no, and explain briefly.

5. A newspaper article says that on the average, college freshmen spend 7.5 hours a week going to parties.[14] One administrator does not believe that these figures apply at her college, which has nearly 3,000 freshmen. She takes a simple random sample of 100 freshmen, and interviews them. On average, they report 6.6 hours a week going to parties, and the SD is 9 hours. Is the difference between 6.6 and 7.5 real?

(a) Formulate the null and alternative hypotheses in terms of a box model.
(b) Fill in the blanks. The null says that the average of the box is ______. The alternative says that average of the box is ______.
(c) Now answer the question: is the difference real?

6. In 1969, Dr. Spock came to trial before Judge Ford, in Boston's federal court house. The charge was conspiracy to violate the Military Service Act. "Of all defendants, Dr. Spock, who had given wise and welcome advice on child-rearing to millions of mothers, would have liked women on his jury."[15] The jury was drawn from a "venire," or panel, of 350 persons selected by the clerk. This venire included only 102 women, although a majority of the eligible jurors in the district were female. At the next stage in selecting the jury to hear the case, Judge Ford chose 100 potential jurors out of these 350 persons. His choices included 9 women.

(a) 350 people are chosen at random from a large population, which is over 50% female. How likely is it that the sample includes 102 women or fewer?
(b) 100 people are chosen at random (without replacement) from a group consisting of 102 women and 248 men. How likely is it that the sample includes 9 women or fewer?
(c) What do you conclude?



"YOUR HONOR, THE PROSECUTION OBJECTS TO THE COMPOSITION OF THIS JURY."

7. I. S. Wright and associates did a clinical trial on the effect of anticoagulant therapy for coronary heart disease.[16] Eligible patients who were admitted to participating hospitals on odd days of the month were given the therapy; eligible patients admitted on even days were the controls. In total, there were 580 patients in the therapy group and 442 controls. An observer says,

   > Since the odd-even assignment to treatment or control is objective and impartial, it is just as good as tossing a coin.

   Do you agree or disagree? Explain briefly. Assume the trial was done in a month with 30 days.

8. Bookstores like education, one reason being that educated people are more likely to spend money on books. National data show the nationwide average educational level to be 13 years of schooling completed, with an SD of about 3 years, for persons age 18 and over.[17]

   A bookstore is doing a market survey in a certain county, and takes a simple random sample of 1,000 people age 18 and over. They find the average educational level to be 14 years, and the SD is 5 years. Can the difference in average educational level between the sample and the nation be explained by chance variation? If not, what other explanation can you give?

9. A computer is programmed to make 100 draws at random with replacement from the box $\boxed{\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{1}}$, and take their sum. It does this 144 times; the average of the 144 sums is 21.13. The program is working fine. Or is it?

   Working fine Something is wrong

   Choose one option, and explain your reason.

10. On November 9, 1965, the power went out in New York City, and stayed out for a day—the Great Blackout. Nine months later, the newspapers suggested that New York was experiencing a baby boom. The table below shows the number of babies born every day during a 25 day period, centered nine months and ten days after the Great Blackout.[18] These numbers average out to 436. This turns out not to be unusually high for New York. But there is an interesting twist to the data: the 3 Sundays only average 357. How likely is it that the average of 3 days chosen at random from the table will be 357 or less? Is chance a good explanation for the difference between Sundays and weekdays? If not, how would you explain the difference?

Number of births in New York, August 1–25, 1966

Date	*Day*	*Number*	*Date*	*Day*	*Number*
1	Mon.	451	15	Mon.	451
2	Tues.	468	16	Tues.	497
3	Wed.	429	17	Wed.	458
4	Thur.	448	18	Thur.	429
5	Fri.	466	19	Fri.	434
6	Sat.	377	20	Sat.	410
7	Sun.	344	21	Sun.	351
8	Mon.	448	22	Mon.	467
9	Tue.	438	23	Tues.	508
10	Wed.	455	24	Wed.	432
11	Thur.	468	25	Thur.	426
12	Fri.	462			
13	Sat.	405			
14	Sun.	377			

11. According to the census, the median household income in Atlanta (1.5 million households) was $52,000 in 1999.[19] In June 2003, a market research organization takes a simple random sample of 750 households in Atlanta; 56% of the sample households had incomes over $52,000. Did median household income in Atlanta increase over the period 1999 to 2003?

    (a) Formulate null and alternative hypotheses in terms of a box model.
    (b) Calculate the appropriate test statistic and P.
    (c) Did median family income go up?

12. (Hard.) Does the psychological environment affect the anatomy of the brain? This question was studied experimentally by Mark Rosenzweig and his associates.[20] The subjects for the study came from a genetically pure strain of rats. From each litter, one rat was selected at random for the treatment group, and one for the control group. Both groups got exactly the same kind of food and drink—as much as they wanted. But each animal in the treatment group lived with 11 others in a large cage, furnished with playthings which were changed daily. Animals in the control group lived in isolation, with no toys. After a month, the experimental animals were killed and dissected.

Cortex weights (in milligrams) for experimental animals. The treatment group (T) had an enriched environment. The control group (C) had a deprived environment.

Expt. #1		*Expt. #2*		*Expt. #3*		*Expt. #4*		*Expt. #5*	
T	*C*	*T*	*C*	*T*	*C*	*T*	*C*	*T*	*C*
689	657	707	669	690	668	700	662	640	641
656	623	740	650	701	667	718	705	655	589
668	652	745	651	685	647	679	656	624	603
660	654	652	627	751	693	742	652	682	642
679	658	649	656	647	635	728	578	687	612
663	646	676	642	647	644	677	678	653	603
664	600	699	698	720	665	696	670	653	593
647	640	696	648	718	689	711	647	660	672
694	605	712	676	718	642	670	632	668	612
633	635	708	657	696	673	651	661	679	678
653	642	749	692	658	675	711	670	638	593
		690	621	680	641	710	694	649	602

On the average, the control animals were heavier and had heavier brains, perhaps because they ate more and got less exercise. However, the treatment group had consistently heavier cortexes (the "grey matter," or thinking part of the brain). This experiment was repeated many times; results from the first 5 trials are shown in the table: "T" means treatment, and "C" is for control. Each line refers to one pair of animals. In the first pair, the animal in treatment had a cortex weighing 689 milligrams; the one in control had a lighter cortex, weighing only 657 milligrams. And so on.

Two methods of analyzing the data will be presented in the form of exercises. Both methods take into account the pairing, which is a crucial feature of the data. (The pairing comes from randomization within litter.)

(a) *First analysis.* How many pairs were there in all? In how many of these pairs did the treatment animal have a heavier cortex? Suppose treatment had no effect, so each animal of the pair had a 50–50 chance to have the heavier cortex, independently from pair to pair. Under this assumption, how likely is it that an investigator would get as many pairs as Rosenzweig did, or more, with the treatment animal having the heavier cortex? What do you infer?

(b) *Second analysis.* For each pair of animals, compute the difference in cortex weights "treatment − control." Find the average and SD of all these differences. The null hypothesis says that these differences are like draws made at random with replacement from a box whose average is 0—the treatment has no effect. Make a z-test of this hypothesis. What do you infer?

(c) To ensure the validity of the analysis, the following precaution was taken. "The brain dissection and analysis of each set of littermates was done in immediate succession but in a random order and identified only by code number so that the person doing the dissection does not know which cage the rat comes from." Comment briefly on the following: What was the point of this precaution? Was it a good idea?

8. SUMMARY

1. A *test of significance* gets at the question of whether an observed difference is real (the *alternative hypothesis*) or just a chance variation (the *null hypothesis*).

2. To make a test of significance, the null hypothesis has to be set up as a box model for the data. The alternative hypothesis is another statement about the box.

3. A *test statistic* measures the difference between the data and what is expected on the null hypothesis. The *z-test* uses the statistic

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

The expected value in the numerator is computed on the basis of the null hypothesis. If the null hypothesis determines the SD of the box, use this information when computing the SE in the denominator. Otherwise, you have to estimate the SD from the data.

4. The *observed significance level* (also called P, or the P-value) is the chance of getting a test statistic as extreme as or more extreme than the observed one. The chance is computed on the basis that the null hypothesis is right. Therefore, P does not give the chance of the null hypothesis being right.

5. Small values of P are evidence against the null hypothesis: they indicate something besides chance was operating to make the difference.

6. Suppose that a small number of tickets are drawn at random with replacement from a box whose contents follow the normal curve, with an average of 0 and an unknown SD. Each draw is added to an unknown constant to give a measurement. The null hypothesis says that this unknown constant equals some given value c. An alternative hypothesis says that the unknown constant is bigger than c. The SD of the box is estimated by the SD^+ of the data. Then the SE for the average of the draws is computed. The test statistic is

$$t = \frac{\text{average of draws} - c}{\text{SE}}$$

The observed significance level is obtained not from the normal curve but from one of the Student's curves, with

$$\text{degrees of freedom} = \text{number of measurements} - \text{one}.$$

This procedure is a *t-test*.

6. REVIEW EXERCISES

Review exercises may cover material from previous chapters.

1. Five hundred draws are made at random with replacement from a box of numbered tickets; 276 are positive. Someone tells you that 50% of the tickets in the box show positive numbers. Do you believe it? Answer yes or no, and explain.

2. One hundred draws are made at random with replacement from box A, and 250 are made at random with replacement from box B.

   (a) 50 of the draws from box A are positive, compared to 131 from box B: 50.0% versus 52.4%. Is this difference real, or due to chance?

   (b) The draws from box A average 1.4 and their SD is 15.3; the draws from box B average 6.3 and their SD is 16.1. Is the difference between the averages statistically significant?

3. The Gallup poll asks respondents how they would rate the honesty and ethical standards of people in different fields—very high, high, average, low, or very low.[22] The percentage who rated clergy "very high or high" dropped from 60% in 2000 to 54% in 2005. This may have been due to scandals involving sex abuse; or it may have been a chance variation. (You may assume that in each year, the results are based on independent simple random samples of 1,000 persons in each year.)

   (a) Should you make a one-sample z-test or a two-sample z-test? Why?

   (b) Formulate the null and alternative hypotheses in terms of a box model. Do you need one box or two? Why? How many tickets go into each box? How many draws? What do the tickets show? What do the null and alternative hypotheses say about the box(es)?

   (c) Can the difference between 60% and 54% be explained as a chance variation? Or was it the scandals? Or something else?

4. This continues exercise 3. In 2005, 65% of the respondents gave medical doctors a rating of "very high or high," compared to a 67% rating for druggists. Is the difference real, or a chance variation? Or do you need more information to decide? If the difference is real, how would you explain it? Discuss briefly. You may assume that the results are based on a simple random sample of 1,000 persons taken in 2005; each respondent rated clergy, medical doctors, druggists, and many other professions.[23]

5. One experiment involved 383 students at the University of British Columbia. 200 were chosen at random to get item A, and 92 of them answered "yes." The other 183 got item B, and 161 out of the second group answered "yes."[24]

   Item A) Imagine that you have decided to see a play and paid the admission price of $20 per ticket. As you enter the theatre, you discover that you have lost the ticket. The seat was not marked, and the ticket cannot be recovered. Would you pay $20 for another ticket?

from a box of
% of the tickets
yes or no, and

om box A, and

31 from box B:
hance?
5.3; the draws
erence between

esty and ethical
ge, low, or very
' dropped from
ndals involving
assume that in
lom samples of

z-test? Why?
of a box model.
ts go into each
/hat do the null

ned as a chance

ve medical doc-
for druggists. Is
ore information
Discuss briefly.
dom sample of
nedical doctors,

ritish Columbia.
answered "yes."
swered "yes."[24]

and paid the ad-
re, you discover
d, and the ticket
ket?

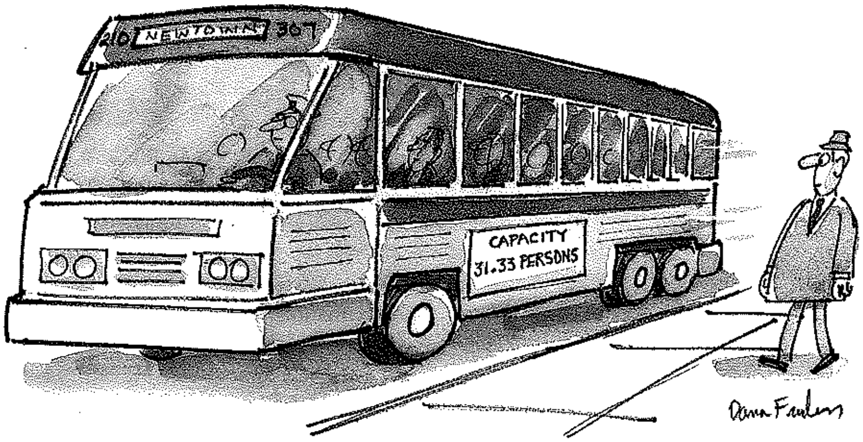
Item B) Imagine that you have decided to see a play where admission is \$20 per ticket. As you enter the theatre, you discover that you have lost a \$20 bill. Would you still pay \$20 for a ticket for the play? [In Canada, "theatre" is the right spelling.]

From the standpoint of economic theory, both items present the same facts and call for the same answer; any difference between them must be due to chance. From a psychological point of view, the framing of the question can be expected to influence the answer. What do the data say?

6. An experiment is performed to see whether calculators help students do word problems.[25] The subjects are a group of 500 thirteen-year-olds in a certain school district. All the subjects work the problem below. Half of them are chosen at random and allowed to use calculators; the others do the problem with pencil and paper. In the calculator group, 18 students get the right answer; in the pencil-and-paper group, 59 do. Can this difference be explained by chance? What do you conclude?

   The problem. An army bus holds 36 soldiers. If 1,128 soldiers are being bussed to their training site, how many buses are needed?

   Note. 1,128/36 = 31.33, so 32 buses are needed. However, 31.33 was a common answer, especially in the calculator group; 31 was another common answer.



7. When convicts are released from prison, they have no money, and there is a high rate of "recidivism:" the released prisoners return to crime and are arrested again. Would providing income support to ex-convicts during the first months after their release from prison reduce recidivism? The Department of Labor ran a randomized controlled experiment to find out.[26] The experiment was done on a selected group of convicts being released from certain prisons in Texas and Georgia. Income support was provided, like unemployment

insurance. There was a control group which received no payment, and four different treatment groups (differing slightly in the amounts paid).

The exercise is on the results for Georgia, and combines the four treatment groups into one. Assume that prisoners were randomized to treatment or control.

(a) 592 prisoners were assigned to the treatment group, and of them 48.3% were rearrested within a year of release. 154 were assigned to the control group, and of them 49.4% were rearrested within a year of release. Did income support reduce recidivism? Answer yes or no, and explain briefly.

(b) In the first year after their release from prison, those assigned to the treatment group averaged 16.8 weeks of paid work; the SD was 15.9 weeks. For those assigned to the control group, the average was 24.3 weeks; the SD was 17.3 weeks. Did income support reduce the amount that the ex-convicts worked? Answer yes or no, and explain briefly.

8. One experiment contrasted responses to "prediction-request" and to "request-only" treatments, in order to answer two research questions.[27]

(i) Can people predict how well they will behave?
(ii) Do their predictions influence their behavior?

In the prediction-request group, subjects were first asked to predict whether they would agree to do some volunteer work. Then they were requested to do the work. In the request-only group, the subjects were requested to do the work; they were not asked to make predictions beforehand. In parts (a-b-c), a two-sample z-test may or may not be legitimate. If it is legitimate, make it. If not, why not?

(a) 46 residents of Bloomington, Indiana were chosen at random for the "prediction-request" treatment. They were called and asked to predict "whether they would agree to spend 3 hours collecting for the American Cancer Society if contacted over the telephone with such a request." 22 out of the 46 said that they would. Another 46 residents of that town were chosen at random for the "request-only" treatment. They were requested to spend the 3 hours collecting for the American Cancer Society. Only 2 out of 46 agreed to do it. Can the difference between 22/46 and 2/46 be due to chance? What do the data say about the research questions (i) and (ii)?

(b) Three days later, the prediction-request group was called again, and requested to spend 3 hours collecting for the American Cancer Society: 14 out of 46 agreed to do so. Can the difference between 14/46 and 2/46 be due to chance? What do the data say about the research questions (i) and (ii)?

(c) Can the difference between 22/46 and 14/46 be due to chance? What do the data say about the research questions (i) and (ii)?

ent, and four
l).

our treatment
tment or con-

f them 48.3%
ed to the con-
ear of release.
), and explain

signed to the
SD was 15.9
rage was 24.3
ce the amount
lain briefly.

d to “request-

redict whether
e requested to
ested to do the
parts (a-b-c), a
ate, make it. If

random for the
sked to predict
for the Amer-
vith such a re-
er 46 residents
nly” treatment.
r the American
n the difference
e data say about

lled again, and
n Cancer Soci-
between 14/46
out the research

o chance? What
)?

9. A researcher wants to see if the editors of journals in the field of social work are biased. He makes up two versions of an article, “in which an asthmatic child was temporarily separated from its parents in an effort to relieve the symptoms of an illness that is often psychosomatic.” In one version, the separation has a positive effect; in another, negative.[28] The article is submitted to a group of 107 journals; 53 are chosen at random to get the positive version, and 54 get the negative one. The results are as follows:

	Positive	*Negative*
Accept	28	8
Reject	25	46

The first column of the table says that 28 of the journals getting the positive version accepted it for publication, and 25 rejected it. The second column gives the results for the journals that got the negative version. Is chance a good explanation for the results? If not, what can be concluded about journal publication policy?

10. An investigator wants to show that first-born children score higher on IQ tests than second-borns. He takes a simple random sample of 400 two-child families in a school district, both children being enrolled in elementary school. He gives these children the WISC vocabulary test (described in exercise 7 on pp. 507–508), with the following results.

- The 400 first-borns average 29 and their SD is 10.
- The 400 second-borns average 28 and their SD is 10.

(Scores are corrected for age differences.) He makes a two-sample z-test:

SE for first-born average ≈ 0.5

SE for second-born average ≈ 0.5

SE for difference $= \sqrt{0.5^2 + 0.5^2} \approx 0.7$

$z = 1/0.7 \approx 1.4, \quad P \approx 8\%$

Comment briefly on the use of statistical tests.

11. (Hard.) The logic of the two-sample z-test in section 27.2 relies on two mathematical facts: (i) the expected value of a difference equals the difference of the expected values, and (ii) the expected value of the sample average equals the population average. Explain briefly, with reference to the NAEP reading scores.

7. SUMMARY

1. The expected value for the difference of two quantities equals the difference of the expected values. (Independence is not required here.)