Contents

1 Probability 2
2 Binomial distribution 5
3 Normal distribution 9
4 Sampling/study design 12
5 One-sample inference for means 14
6 Two-sample inference for means 18
7 Inference for proportions 25
8 Selected solutions 33
  8.1 Solutions – Exercises on probability 33
  8.2 Solutions – exercises on the binomial distribution 36
  8.3 Solutions – Exercises on the normal distribution 40
  8.4 Solutions — Sampling/study design 43
  8.5 Solutions — One-sample inference for means 44
  8.6 Solutions — Two-sample inference for means 45
  8.7 Solutions — Inference for proportions 48
1 Probability

1. For Mexican American infants born in Arizona in 1986 and 1987, the probability that a child’s gestational age is less than 37 weeks is 0.142 and the probability that his or her birth weight is less than 2500 grams is 0.051. Furthermore, the probability that these two events occur simultaneously is 0.031.

Let A be the event that an infant’s gestational age is less than 37 weeks and B the event that his or her birth weight is less than 2500 grams.

(a) Are A and B independent?
(b) For a randomly selected Mexican American newborn, what is the probability that A or B or both occur?
(c) What is the probability that event A occurs given that event B occurs?

2. Consider the following natality statistics for the U.S. population in 1992. According to these data, the probabilities that a randomly selected woman who gave birth in 1992 was in each of the following age groups are as follows:

<table>
<thead>
<tr>
<th>Age</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 15</td>
<td>0.003</td>
</tr>
<tr>
<td>15–19</td>
<td>0.124</td>
</tr>
<tr>
<td>20–24</td>
<td>0.263</td>
</tr>
<tr>
<td>25–29</td>
<td>0.290</td>
</tr>
<tr>
<td>30–34</td>
<td>0.220</td>
</tr>
<tr>
<td>35–39</td>
<td>0.085</td>
</tr>
<tr>
<td>40–44</td>
<td>0.014</td>
</tr>
<tr>
<td>45–49</td>
<td>0.001</td>
</tr>
<tr>
<td>Total</td>
<td>1.000</td>
</tr>
</tbody>
</table>

(a) What is the probability that a woman who gave birth in 1992 was 24 years of age or younger?
(b) What is the probability that she was 40 or older?
(c) Given that the mother of a particular child was under 30 years of age, what is the probability that she was not yet 20?
(d) Given that the mother was 35 years of age or older, what is the probability that she was under 40?

3. Looking at the United States population in 1993, the probability that an adult between the ages of 45 and 64 does not have health insurance coverage of any kind is 0.123.

(a) Suppose that you randomly select a 47-year-old woman and an unrelated 59-year-old man from this population. What is the probability that both individuals are uninsured?
(b) What is the probability that both adults have health insurance coverage?
(c) If five unrelated adults between the ages of 45 and 64 are chosen from the population, what is the probability that all five are uninsured?

4. One study has reported that the sensitivity of the mammogram as a screening test for detecting breast cancer is 0.85, while its specificity is 0.80.

(a) What is the probability of a false negative test result?
(b) What is the probability of a false positive result?
(c) In a population in which the probability that a woman has breast cancer is 0.0025, what is the probability that she has cancer given that her mammogram is positive?
5. The National Institute for Occupational Safety and Health has developed a case definition of carpal tunnel syndrome—an affliction of the wrist—that incorporates three criteria: symptoms of nerve involvement, a history of occupational risk factors, and the presence of physical exam findings. The sensitivity of this definition as a test for carpal tunnel syndrome is 0.67; its specificity is 0.58.

(a) In a population in which the prevalence of carpal tunnel syndrome is estimated to be 15%, what is the predictive value of a positive test result?

(b) How does this predictive value change if the prevalence is only 10%? If it is 5%?


**Pulmonary Disease**

Pulmonary embolism is a relatively common condition that necessitates hospitalization and also often occurs in patients hospitalized for other reasons. An oxygen tension (arterial \(P_{O_2}\) \(<\) 90 mm Hg is one of the important criteria used in diagnosing this condition. Suppose that the sensitivity of this test is 95%, the specificity is 75%, and the estimated prevalence is 20% (i.e., a doctor estimates that a patient has a 20% chance of pulmonary embolism before performing the test).

3.21 What is the predictive value positive of this test? What does it mean in words?

3.22 What is the predictive value negative of this test? What does it mean in words?

3.23 Answer Problem 3.21 if the estimated prevalence is 80%.

3.24 Answer Problem 3.22 if the estimated prevalence is 80%.
Genetics
Two healthy parents have a child with a severe autosomal recessive condition that cannot be identified by prenatal diagnosis. They realize that the risk of this condition for subsequent offspring is 1/4, but wish to embark on a second pregnancy. During the early stages of the pregnancy, an ultrasound test determines that there are twins.

3.45 Suppose that there are monozygotic, or MZ (identical) twins. What is the probability that both twins are affected? one twin affected? neither twin affected? Are the outcomes for the two MZ twins independent or dependent events?

3.46 Suppose that there are dizygotic, or DZ (fraternal) twins. What is the probability that both twins are affected? one twin affected? neither twin affected? Are the outcomes for the two DZ twins independent or dependent events?

3.47 Suppose there is a 1/3 probability of MZ twins and a 2/3 probability of DZ twins. What is the overall probability that both twins are affected? One twin affected? Neither affected?

3.48 Suppose we learn that both twins are affected but don’t know whether they are MZ or DZ twins. What is the probability that they are MZ twins given this additional information?


2 Binomial distribution

1. Suppose that you are interested in monitoring air pollution in Los Angeles, California, over a one-week period. Let $X$ be a random variable that represents the number of days out of the seven on which the concentration of carbon monoxide surpasses a specified level. Do you believe that $X$ has a binomial distribution? Explain.

2. Consider a group of seven individuals selected from the population of 65 to 74 year-olds in the United States. The number of persons in this sample who suffer from diabetes is a binomial random variable with parameters $n = 7$ and $p = 0.125$.
   
   (a) If you wish to make a list of the seven persons chosen, in how many ways can they be ordered?
   
   (b) Without regard to order, in how many ways can you select four individuals from this group of seven?
   
   (c) What is the probability that exactly two of the individuals in the sample suffer from diabetes?
   
   (d) What is the probability that four of them have diabetes?

3. It is known that when mice are injected with preparation L that 25% become infected. If 3 mice are injected independently, what is the probability that a) no mice, b) one mouse, c) two mice, d) all three mice become infected?

4. The binomial distribution with $n = 10$ and $p = 0.15$ describes the probability distribution for the number of positive outcomes (only positive or negative outcomes are possible) in 10 independent trials when the probability for a positive outcome in a single trial is 0.15. The binomial distribution for $n = 10$ and $p = 0.15$ is tabulated in Table 1.

   (a) If 15% of all pregnancies result in a miscarriage, what is the probability that
      i. exactly five out of ten randomly chosen pregnant women have a miscarriage?
      ii. more than five out of ten randomly chosen pregnant women have a miscarriage?
      iii. five or more out of ten randomly chosen pregnant women have a miscarriage?
   
   (b) Use the binomial formula to verify that $P(X = 1)=0.3474$ (Table 1).

Table 1: The binomial distribution for $n = 10$ and $p = 0.15$. $P(X = r)$ is the probability of observing $r$ successes in 10 trials.

<table>
<thead>
<tr>
<th></th>
<th>$P(X = r)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1969</td>
</tr>
<tr>
<td>1</td>
<td>0.3474</td>
</tr>
<tr>
<td>2</td>
<td>0.2759</td>
</tr>
<tr>
<td>3</td>
<td>0.1298</td>
</tr>
<tr>
<td>4</td>
<td>0.0401</td>
</tr>
<tr>
<td>5</td>
<td>0.0085</td>
</tr>
<tr>
<td>6</td>
<td>0.0012</td>
</tr>
<tr>
<td>7</td>
<td>0.0001</td>
</tr>
<tr>
<td>8</td>
<td>0.0000</td>
</tr>
<tr>
<td>9</td>
<td>0.0000</td>
</tr>
<tr>
<td>10</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
5. According to the Behavioral Risk Factor Surveillance System, 58% of all Americans adhere to a sedentary lifestyle.

(a) If you selected repeated samples of size twelve from the U.S. population, what would be the mean number of individuals per sample who do not exercise regularly? What would be the standard deviation?

(b) Suppose that you select a sample of twelve individuals and find that ten of them do not exercise regularly. Assuming that the Surveillance System is correct, what is the probability that you would have obtained results as bad as or worse than those you observed?

6. (a) Write down the probability distribution for a binomial random variable with \( n=6 \) and \( p = 0.15 \). That is, construct a table similar to Table 1 where \( n=6 \) and \( p = 0.15 \).

(b) A cruise ship sailing from Stockholm to Helsinki is fitted with 6 engines. Each of these engines has a probability of 0.15 of failing during any single voyage. The ship will, however, make it to Helsinki on time provided no more than 2 engines fail during a voyage. Assume that the total number of engines that fail during a voyage can be described by a binomial distribution.

i. What is the probability that the ship will make it to Helsinki on time during its next voyage?

ii. Over a large number of trips, what is the mean number of engines that fail per voyage?

iii. During a single voyage, what is the most likely value for the number of engines that fail?

iv. Do you think a binomial distribution is appropriate to describe this situation?

7. After spending an entire day studying for an exam, 5 students decide that one of them should go and buy pizza while the remainder continue studying. In order to choose who buys the pizza, they decide to play the game ‘odd man out’. The students each toss a coin and if all coins turn up the same, except for one, then the person tossing the minority coin must go and buy the pizza.

(a) What is the probability that it will only take one toss (i.e. each student tossing a coin once) to determine the identity of the pizza buyer?

(b) What is the probability that it will take exactly two rounds of tosses to determine the identity of the pizza buyer?

8. Each day, the stock exchange price of a new computer stock moves up or down one point with probabilities \( \frac{3}{4} \) and \( \frac{1}{4} \) respectively. What is the probability that after six days the stock will have returned to its original price?

**Pediatrics**
A hospital administrator wants to construct a special-care nursery for low-birthweight infants (≤ 2500 g) and wants to have some idea as to the number of beds she should allocate to the nursery. She is willing to assume that the recovery period of each baby is exactly 4 days and thus is interested in the expected number of premature births over the period.

4.19 If the number of premature births in any 4-day period is binomially distributed with parameters $n = 25$ and $p = .1$, then find the probability of $0, 1, 2, \ldots, 7$ premature births over this period.

4.20 The administrator wishes to allocate $x$ beds where the probability of having more than $x$ premature births over a 4-day period is less than 5%. What is the smallest value of $x$ that satisfies this criterion?


**Pulmonary Disease**
Each year approximately 4% of current smokers attempt to quit smoking, and 50% of those who try to quit are successful in the sense that they abstain from smoking for at least 1 year from the date they quit.

4.35 What is the probability that a current smoker will quit for at least 1 year?

4.36 What is the probability that among 100 current smokers, at least 5 will quit smoking for at least 1 year?

An educational program was conducted among smokers who attempt to quit to maximize the likelihood that such individuals would continue to abstain for the long term.

4.37 Suppose that of 20 people who enter the program when they first stop smoking, 15 still abstain from smoking 1 year later. Can the program be considered successful?

**Infectious Disease**

An outbreak of acute gastroenteritis occurred at a nursing home in Baltimore, Maryland, in December 1980 [3]. A total of 46 out of 98 residents of the nursing home became ill. People living in the nursing home shared rooms: 13 rooms contained 2 occupants, 4 rooms contained 3 occupants, and 15 rooms contained 4 occupants. One question that arises is whether or not a geographical clustering of disease occurred for persons living in the same room.

4.47 If the binomial distribution holds, what is the probability distribution of the number of affected people in rooms with 2 occupants? That is, what is the probability of finding zero affected people? One affected person? Two affected people?

4.48 Answer Problem 4.47 for the probability distribution of the number of affected people in rooms with 3 occupants.

4.49 Answer Problem 4.47 for the probability distribution of the number of affected people in rooms with 4 occupants.

A summary of the number of affected people and the total number of people in a room is given in Table 4.3.

(include Table 4.3 and questions 4.50–4.53)
3 Normal distribution

1. Consider the standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. Use the table provided at the end of this document to answer the following.
   
   (a) What is the probability that an outcome $z$ is greater than 2.60?
   
   (b) What is the probability that $z$ is less than 1.35?
   
   (c) What is the probability that $z$ is between $-1.70$ and $3.10$?
   
   (d) What value of $z$ cuts off the upper 15% of the standard normal distribution?
   
   (e) What value of $z$ marks off the lower 20% of the distribution?

   When you return to your office, check that you can obtain the same answers using your spreadsheet program of choice (e.g. MS Excel) or a statistical package (e.g. Stata). Some hand calculators even have these functions.

2. Assuming that the height of adult males has a normal distribution, what proportion of males will be more than two standard deviations above the mean height?

3. Over a 25 year period the mean height of adult males increased from 175.8 cm to 179.1 cm, but the standard deviation stayed at 5.84 cm. The minimum height requirement for men to join the police force is 172 cm.
   
   (a) What proportion of men would be too short to join the police force at the beginning of the 25 year period?
   
   (b) What proportion of men would be too short to join the police force at the end of the 25 year period?

4. Assume that in order for an individual to be classified as hypertensive he/she needs to have an average systolic blood pressure of at least 90 mm Hg.
   
   (a) Compute the probability for misclassifying the status of each of the following patients on the basis of one single blood pressure measurement.
   
      i. Patient 1 has mean 85 and standard deviation 10
      ii. Patient 2 has mean 92 and standard deviation 5
   
   (b) What is the probability that a single measurement for patient 1 falls
      
      i. In the interval (80-90) ?
      ii. Outside the interval (80-90) ?

5. The distribution of weights for the population of males in the United States is approximately normal with mean $\mu = 172.2$ pounds and standard deviation $\sigma = 29.8$ pounds.
   
   (a) What is the probability that a randomly selected man weighs less than 130 pounds?
   
   (b) What is the probability that he weighs more than 210 pounds?
   
   (c) What is the probability that among five males selected at random from the population, at least one will have a weight outside the range 130 to 210 pounds?
6. In the Framingham Study, serum cholesterol levels were measured for a large number of healthy males. The population was then followed for 16 years. At the end of this time, the men were divided into two groups: those who had developed coronary heart disease and those who had not. The distributions of the initial serum cholesterol levels for each group were found to be approximately normal. Among individuals who eventually developed coronary heart disease, the mean serum cholesterol level was $\mu_d = 244 \text{ mg/100 ml}$ and the standard deviation was $\sigma_d = 51 \text{ mg/100 ml}$; for those who did not develop the disease, the mean serum cholesterol level was $\mu_{nd} = 219 \text{ mg/100 ml}$ and the standard deviation was $\sigma_{nd} = 41 \text{ mg/100ml}$.

(a) Suppose that an initial serum cholesterol level of 260 mg/100 ml or higher is used to predict coronary heart disease. What is the probability of correctly predicting heart disease for a man who will develop it?

(b) What is the probability of predicting heart disease for a man who will not develop it?

(c) What is the probability of failing to predict heart disease for a man who will develop it?

(d) What would happen to the probabilities of false positive and false negative errors if the cutoff point for predicting heart disease is lowered to 250 mg/100 ml?

(e) In this population, does initial serum cholesterol level appear to be useful for predicting coronary heart disease? Why or why not?


**Nutrition**

The distribution of serum levels of alpha tocopherol (serum vitamin E) is approximately normal with mean 860 $\mu g/dL$ and standard deviation 340 $\mu g/dL$.

**5.27** What percentage of people have serum alpha tocopherol levels between 400 and 1000 $\mu g/dL$?

**5.28** Suppose a person is identified as having toxic levels of alpha tocopherol if his or her serum level is $> 2000 \mu g/dL$. What percentage of people will be so identified?

**5.29** A study is undertaken for evidence of toxicity of 2000 people who regularly take vitamin-E supplements. The investigators found that 4 people have serum alpha tocopherol levels $> 2000 \mu g/dL$. Is this an unusual number of people with toxic levels of serum alpha tocopherol?
Hypertension
Blood-pressure measurements are known to be variable, and repeated measurements are essential to accurately characterize a person's BP status. Suppose a person is measured on \( n \) visits with \( k \) measurements per visit and the average of all \( nk \) diastolic blood pressure (DBP) measurements (\( \bar{x} \)) is used to classify a person as to BP status. Specifically, if \( \bar{x} \geq 95 \) mm Hg, then the person is classified as hypertensive; if \( \bar{x} < 90 \) mm Hg, then the person is classified as normotensive; and if \( \bar{x} \geq 90 \) mm Hg, and < 95 mm Hg, the person is classified as borderline. It is also assumed that a person's "true" blood pressure is \( \mu \), representing an average over a large number of visits with a large number of measurements per visit, and that \( \bar{x} \) is normally distributed with mean \( \mu \) and variance \( = 27.7/n + 7.9/(nk) \).

5.33 If a person's true diastolic blood pressure is 100 mm Hg, then what is the probability that the person will be classified accurately (as hypertensive) if a single measurement is taken at 1 visit?

5.34 Is the probability in Problem 5.33 a measure of sensitivity, specificity, or predictive value?

5.35 If a person's true blood pressure is 85 mm Hg, then what is the probability that the person will be accurately classified (as normotensive) if 3 measurements are taken at each of 2 visits?

5.36 Is the probability in Problem 5.35 a measure of sensitivity, specificity, or predictive value?

5.37 Suppose we decide to take 2 measurements per visit. How many visits are needed so that the sensitivity and specificity in Problems 5.33 and 5.35 would each be at least 95%?
4 Sampling/study design

1. Exercise 5.1 from p. 103 Altman. (copied below)

**EXERCISES**

5.1 In 1978–79 a random sample of 1007 residents (608 men and 399 women) of the Lothian region (around Edinburgh) had been asked precisely what alcohol they had drunk in the previous seven days. In March 1981 the combination of an increase in taxation and brewers’ prices meant that, for the first time in over 30 years, the price of alcoholic beverages increased faster than the retail price index. So in the autumn of 1981 the 676 respondents (484 men and 192 women) who had had at least one alcoholic drink in the seven days on which the original survey had been based – the so-called ‘regular drinkers’ – were reinterviewed.

The first survey was carried out between July 1978 and February 1979 and the second between September 1981 and March 1982. Over the three years, the cost of alcoholic beverages had risen by 61% while the retail price index had risen by 52%. Average earnings (and disposable income) had risen more than the retail price index, suggesting that those in regular employment were marginally better off than in 1981. Unemployment in the Edinburgh area, however, had risen steeply between 1978 and 1982 for both men and women.

(continues on next page)
The results of the second survey were reported as follows:

'Of the original 676 regular drinkers, 463 (69%) were successfully interviewed. Of the 213 who were not, 85 could not be traced, 48 were known to have left the region, 39 refused, and 23 were either dead or too ill to be interviewed. A disproportionate number of lost respondents were under the age of 30, unmarried, and not in regular employment. Nevertheless, the sex ratio and both male and female alcohol consumption at the time of the first survey of the 463 who were reinterviewed were representative of the original sample.' (Kendell et al., 1983)

(a) The authors were interested in reduction in alcohol intake, and so did not interview those subjects not reporting drinking in the first survey. Is this reasonable?

(b) What was the response rate to the second survey? How might non-respondents differ from respondents? What is the likely effect on the interpretation of the results of the survey?

(c) Does it matter that the two surveys were not carried out at exactly the same time of year?

(d) If the data showed a reduction in alcohol consumption among the 463 reinterviewed subjects, could the authors reasonably conclude that it was due to the rise in excise duty on alcohol?

The Discussion of the paper begins:

'The central finding of this before and after survey is that a representative population of 463 regular drinkers in the Lothian region reduced their alcohol consumption by 18% between 1978–9 and 1981–2 and simultaneously experienced a 16% reduction in adverse effects. The main cause of this fall in consumption was probably the rising cost of alcoholic beverage relative to the cost of living and average incomes during that three year period.'

(e) Were the 463 'regular drinkers' really a 'representative population'?

(f) Comment on the authors' interpretation of the results. Would your opinion be different if they had interviewed all 1007 subjects in the second survey?

In the final paragraph the authors wrote:

'The findings of this study indicate, therefore, that an increase in excise duty on alcoholic beverages can be an effective means of reducing the ill effects of excessive alcohol consumption.'

(g) Do these conclusions have any validity?
5 One-sample inference for means


**Bacteriology**
Suppose a group of mice are inoculated with a uniform dose of a specific type of bacteria and all die within 24 days, with the distribution of survival times given in Table 6.4.

**Table 6.4**  Survival time of mice after inoculation with a specific type of bacteria

<table>
<thead>
<tr>
<th>Survival time (days)</th>
<th>Number of mice</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>13</td>
<td>30</td>
</tr>
<tr>
<td>14</td>
<td>40</td>
</tr>
<tr>
<td>15</td>
<td>51</td>
</tr>
<tr>
<td>16</td>
<td>71</td>
</tr>
<tr>
<td>17</td>
<td>65</td>
</tr>
<tr>
<td>18</td>
<td>48</td>
</tr>
<tr>
<td>19</td>
<td>36</td>
</tr>
<tr>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
</tr>
</tbody>
</table>

6.27 Assume that the underlying distribution of survival times is normal. Estimate the probability $p$ that a mouse will survive for 20 or more days.

6.28 Suppose we are not willing to assume that the underlying distribution is normal. Estimate the probability $p$ that a mouse will survive for 20 or more days.

6.29 Compute 95% confidence limits for the parameter estimated in Problem 6.28.

6.30 Compute 99% confidence limits for the parameter estimated in Problem 6.28.

6.31 What is your best estimate of the effect of adopting a vegetarian diet on change in serum-cholesterol levels?

6.32 What is the standard error of the estimate given in Problem 6.31?

6.33 Provide a 95% CI for the effect of adopting the diet.

6.34 What can you conclude from your results in Problem 6.33?

Some physicians consider only changes of at least 10 mg/dL (the same units as in Table 2.1) to be clinically significant.

6.35 Among people with a clinically significant change in either direction, what is the best estimate of the proportion of subjects with a clinically significant decline in cholesterol?

6.36 Provide a 95% CI associated with the estimate in Problem 6.35.

6.37 What can you conclude from your results in Problem 6.36?

Serum-Cholesterol levels before and after adopting a vegetarian diet (mg/dL)

<table>
<thead>
<tr>
<th>Subject</th>
<th>Before</th>
<th>After</th>
<th>Before – After</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>195</td>
<td>146</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>145</td>
<td>155</td>
<td>-10</td>
</tr>
<tr>
<td>3</td>
<td>205</td>
<td>178</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>159</td>
<td>146</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>244</td>
<td>208</td>
<td>36</td>
</tr>
<tr>
<td>6</td>
<td>166</td>
<td>147</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>250</td>
<td>202</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>236</td>
<td>215</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>192</td>
<td>184</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>224</td>
<td>208</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>238</td>
<td>206</td>
<td>32</td>
</tr>
<tr>
<td>12</td>
<td>197</td>
<td>169</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>169</td>
<td>182</td>
<td>-13</td>
</tr>
<tr>
<td>14</td>
<td>158</td>
<td>127</td>
<td>31</td>
</tr>
<tr>
<td>15</td>
<td>151</td>
<td>149</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>197</td>
<td>178</td>
<td>19</td>
</tr>
<tr>
<td>17</td>
<td>180</td>
<td>161</td>
<td>19</td>
</tr>
<tr>
<td>18</td>
<td>222</td>
<td>187</td>
<td>35</td>
</tr>
<tr>
<td>19</td>
<td>168</td>
<td>176</td>
<td>-8</td>
</tr>
<tr>
<td>20</td>
<td>168</td>
<td>145</td>
<td>23</td>
</tr>
<tr>
<td>21</td>
<td>167</td>
<td>154</td>
<td>13</td>
</tr>
<tr>
<td>22</td>
<td>161</td>
<td>153</td>
<td>8</td>
</tr>
<tr>
<td>23</td>
<td>178</td>
<td>137</td>
<td>41</td>
</tr>
<tr>
<td>24</td>
<td>137</td>
<td>125</td>
<td>12</td>
</tr>
</tbody>
</table>

Mean 187.8  168.3  19.5
sd 33.2  26.8  16.8
n 24  24  24

**Nutrition**

As part of a dietary-instruction program, ten 25–34-year-old males adopted a vegetarian diet for 1 month. While on the diet, the average daily intake of linoleic acid was 13 g with standard deviation = 4 g.

7.1 If the average daily intake among 25–34-year-old males in the general population is 15 g, then, using a significance level of .05, test the hypothesis that the intake of linoleic acid in this group is lower than that in the general population.

7.2 Compute a \( p \)-value for the hypothesis test in Problem 7.1.

As part of the same program, eight 25–34-year-old females report an average daily intake of saturated fat of 11 g with standard deviation = 11 g while on a vegetarian diet.

7.3 If the average daily intake of saturated fat among 25–34-year-old females in the general population is 24 g, then, using a significance level of .01, test the hypothesis that the intake of saturated fat in this group is lower than that in the general population.

7.4 Compute a \( p \)-value for the hypothesis test in Problem 7.3.

7.5 What is the relationship between your answers to Problems 7.3 and 7.4?


**Occupational Health**

7.25 Suppose it is known that the average life expectancy of a 50-year-old man in 1945 was 18.5 years. Twenty men aged 50 who have been working for at least 20 years in a potentially hazardous industry were ascertained in 1945. On follow-up in 1985 all the men have died, with an average lifetime of 16.2 years and a standard deviation of 7.3 years since 1945. Assuming that life expectancy of 50-year-old men is approximately normally distributed, test if the underlying life expectancy for workers in this industry is shorter than for comparably aged men in the general population.

**Epidemiology**

Height and weight are often used in epidemiological studies as possible predictors of disease outcomes. If the people in the study are assessed in a clinic, then heights and weights are usually measured directly. However, if the people are interviewed at home or by mail, then a person’s self-reported height and weight are often used instead. Suppose we conduct a study on 10 people to test the comparability of these two methods. The data for weight are given in Table 7.1.

### 7.33
Should a one-sided or two-sided test be used here?


**Hypertension**

Several studies have been performed relating urinary potassium excretion to blood-pressure level. These studies have tended to show an inverse relationship between these two variables, with the higher the level of potassium excretion, the lower the BP level. Therefore, a treatment trial is planned to look at the effect of potassium intake in the form of supplement capsules on changes in DBP level. Suppose that in a pilot study, 20 people are given potassium supplements for 1 month. The data are as follows:

<table>
<thead>
<tr>
<th>Mean change (1 month-baseline)</th>
<th>3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sd change</td>
<td>8.5</td>
</tr>
<tr>
<td>n</td>
<td>20</td>
</tr>
</tbody>
</table>

### 7.38
What test should be used to assess if potassium supplements have any effect on DBP level?

### 7.39
Perform the test in Problem 7.38 using a two-sided test and report the $p$-value.

### 7.40
Derive a 95% CI for the true mean change based on the preceding data. What is the relationship of your results here and in Problem 7.39?
6 Two-sample inference for means

1. A study was conducted of genetic and environmental influences on cholesterol levels. The dataset used for the study was obtained from a twin registry in Sweden (Heller et al., 1993). Specifically, four populations of adult twins were studied: (a) monozygotic (MZ) twins reared apart, (b) MZ twins reared together, (c) dizygotic (DZ) twins reared apart, and (d) DZ twins reared together. One issue is whether it is necessary to correct for sex before performing more complex analyses. The data in the table below were presented for total cholesterol levels for MZ twins reared apart by sex.

Table: Comparison of mean total cholesterol for MZ twins reared apart, by sex. n=number of individuals, i.e., for males there are 44 individuals, 22 pairs of twins.

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>253.3</td>
<td>271.0</td>
</tr>
<tr>
<td>sd</td>
<td>44.1</td>
<td>44.1</td>
</tr>
<tr>
<td>n</td>
<td>44</td>
<td>48</td>
</tr>
</tbody>
</table>

(a) If we assume that (i) serum cholesterol is normally distributed, (ii) the samples are independent, and (iii) the standard deviations for men and women are the same, then what is the name of the statistical procedure that can be used to compare the two groups?

(b) Suppose we wish to use the procedure in (a) using a two-sided test. State the hypothesis being tested and implement the method. Report a p-value.

(c) Are the assumptions in (a) likely to hold for these samples? Why or why not?
2. The following table appears in Corder et al. (1993) [1]:

![Table 1](image)

The authors make the following statement in the text (halfway down the middle column on page 922):

Age at onset in affected subjects and age at examination in unaffected subjects were similar in men and women (Table 1).

(a) Why do you think it was of interest to report this result?

(b) Without performing any calculations, do you think the authors’ statement is justified given the data in Table 1?

(c) If we were to test the authors’ claim using a statistical hypothesis test, what would the null and alternative hypotheses, the type of statistical test used, and the distribution of the test statistic under the null hypothesis?

(d) Use the data shown in the table to calculate the test statistics and p-values. Is the authors’ statement consistent with the data?

3. In a randomised placebo-controlled trial of the effects of a mouthwash on xerostomia (dry mouth) problems in terminally ill cancer patients there was a total of 16 patients on the active mouthwash and 16 on the placebo.

The improvement in xerostomia over a one week period for each patient was noted (positive values being better for the patients and negative values indicating a worsening in dryness of mouth). These are given below for the active and placebo mouthwashes.

Improvement scores:

**Active mouthwash**
-6 -4 -3 -2 -1 0 0
0 0 1 1 2 3 5

**Placebo mouthwash**
-3 -2 0 0 0 0 1 1
2 2 3 4 5 6 6
The results of an analysis of these data using Stata are given below.

```
ttest improvement, by(group)
```

Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>-.5</td>
<td>.7011895</td>
<td>2.804758</td>
<td>[-1.99455, .99455]</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>1.8125</td>
<td>.6721777</td>
<td>2.688711</td>
<td>[.3797871, 3.245213]</td>
</tr>
<tr>
<td>combined</td>
<td>32</td>
<td>.65625</td>
<td>.5209509</td>
<td>2.946943</td>
<td>[-.4062364, 1.718736]</td>
</tr>
<tr>
<td>diff</td>
<td>-2.3125</td>
<td>.9713339</td>
<td></td>
<td>-4.296229</td>
<td>-.3287715</td>
</tr>
</tbody>
</table>

Degrees of freedom: 30

Ho: mean(1) - mean(2) = diff = 0

<table>
<thead>
<tr>
<th>Ha: diff &lt; 0</th>
<th>Ha: diff != 0</th>
<th>Ha: diff &gt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = -2.3807</td>
<td>t = -2.3807</td>
<td>t = -2.3807</td>
</tr>
<tr>
<td>P &lt; t = 0.0119</td>
<td>P &gt;</td>
<td>t</td>
</tr>
</tbody>
</table>

(a) Considering only the results of the individuals randomised to using the ‘active’ mouthwash, is there evidence that using the mouthwash is associated with an improvement in xerostomia? Support your argument by quoting from the Stata output.

(b) Considering only the results of the individuals randomised to using the ‘placebo’ mouthwash, is there evidence that using the placebo is associated with an improvement in xerostomia? Support your argument by quoting from the Stata output.

(c) Is there evidence that the average improvement over the one-week period differs between the active and the placebo groups? Support your argument by quoting from the Stata output.

(d) Would you recommend the active mouthwash as a means of reducing dryness of the mouth?

(e) Comment on the assumptions involved in the 'Two-sample t test with equal variances’ reported above and whether they are valid for this study. (if more information or further analysis is needed to check these assumptions then state what you would do)
4. One method for assessing the effectiveness of a drug is to note its concentration in blood and/or urine samples at certain periods of time after giving the drug. Suppose we wish to compare the concentrations of two type of aspirin (types A and B) in urine specimens taken from the same person, 1 hour after he or she has taken the drug. Hence a specific dosage of either type A or type B aspirin is given at one time and the 1-hour urine concentration is measured. One week later, after the first aspirin has presumably been cleared from the system, the same dosage of the other aspirin is given to the same person and the 1-hour urine concentration is noted. Since the order of giving the drugs may affect the results, the choice of aspirin type to be given first is made randomly. This experiment is performed on 10 people. The results are:

Concentration of aspirin in urine samples:

<table>
<thead>
<tr>
<th>Person</th>
<th>Aspirin A 1-hour concentration (mg%)</th>
<th>Aspirin B 1-hour concentration (mg%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>36</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>mean</td>
<td>19.20</td>
<td>15.60</td>
</tr>
<tr>
<td>sd</td>
<td>8.63</td>
<td>7.78</td>
</tr>
</tbody>
</table>

Suppose we wish to test the hypothesis that the concentration of the two drugs are the same in urine specimens.

(a) What are the appropriate null and alternative hypotheses?
(b) What is the appropriate test to use? Conduct this test, by hand (i.e., with calculators)
(c) What is the best point estimate of the difference in concentrations between the two drugs?
(d) What is a 95% confidence interval for the mean difference?


**Ophthalmology**

In a study of the natural history of retinitis pigmentosa (RP), 94 RP patients were followed for 3 years [1]. Among 90 patients with complete follow-up, the mean ±1 se of ln (visual-field loss) over 1, 2, and 3 years was 0.02 ± 0.04, 0.08 ± 0.05, and 0.14 ± 0.07, respectively.

8.1 What test procedure can be used to test for changes in ln (visual field) over any given time period?

8.2 Implement the procedure in Problem 8.1 to test for significant changes in visual field over 1 year. Report a p-value.

8.3 Answer Problem 8.2 for changes over 2 years.

8.4 Answer Problem 8.2 for changes over 3 years.

**Hypertension**
Blood-pressure measurements taken on the left and right arms of a person are assumed to be comparable. To test this assumption, 10 volunteers are obtained and systolic blood-pressure readings are taken simultaneously on both arms by two different observers, Ms. Jones for the left arm and Mr. Smith for the right arm. The data are given in Table 8.2.

**Table 8.2** Effect of arm on level of blood pressure (mm Hg)

<table>
<thead>
<tr>
<th>Patient</th>
<th>Left arm</th>
<th>Right arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>130</td>
<td>126</td>
</tr>
<tr>
<td>2</td>
<td>120</td>
<td>124</td>
</tr>
<tr>
<td>3</td>
<td>135</td>
<td>127</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>95</td>
</tr>
<tr>
<td>5</td>
<td>98</td>
<td>102</td>
</tr>
<tr>
<td>6</td>
<td>110</td>
<td>109</td>
</tr>
<tr>
<td>7</td>
<td>123</td>
<td>124</td>
</tr>
<tr>
<td>8</td>
<td>136</td>
<td>132</td>
</tr>
<tr>
<td>9</td>
<td>140</td>
<td>137</td>
</tr>
<tr>
<td>10</td>
<td>155</td>
<td>156</td>
</tr>
</tbody>
</table>

8.11 Assuming that the two observers are comparable, test whether or not the two arms give comparable readings.

8.12 Suppose we do not assume that the two observers are comparable. Can the experiment, as it is defined, detect differences between the two arms? If not, can you suggest an alternative experimental design so as to achieve this aim?
Hypertension
A 1982 study by the Lipid Research Clinics looked at the relationship between alcohol consumption and level of systolic blood pressure in women not using oral contraceptives [3]. Alcohol consumption was categorized as follows: no alcohol use; \( \leq 10 \) oz/week alcohol consumption; \( > 10 \) oz/week alcohol consumption. The results for women 30–39 years of age are given in Table 8.4.

Suppose we wish to compare the levels of systolic blood pressure of groups A and B and we have no prior information regarding which group has higher blood pressure.

Table 8.4  Relationship between systolic blood pressure and alcohol consumption in 30–39-year-old women not using oral contraceptives

<table>
<thead>
<tr>
<th>Systolic blood pressure (mm Hg)</th>
<th>Mean</th>
<th>sd</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. No alcohol use</td>
<td>110.5</td>
<td>13.3</td>
<td>357</td>
</tr>
<tr>
<td>B. ( \leq 10 ) oz per week alcohol consumption</td>
<td>109.1</td>
<td>13.4</td>
<td>440</td>
</tr>
<tr>
<td>C. ( &gt; 10 ) oz per week alcohol consumption</td>
<td>114.5</td>
<td>14.9</td>
<td>23</td>
</tr>
</tbody>
</table>

Source: Reprinted with permission of the Journal of Chronic Diseases, 35(4), 251–257.

8.22 Should a one-sample or two-sample test be used here?

8.23 Should a one-sided or two-sided test be used here?

8.24 Which test procedure(s) should be used to test the preceding hypotheses?

8.25 Carry out the test in Problem 8.24 and report a \( p \)-value.

8.26 Compute a 95% CI for the mean difference in blood pressure between the two groups.

8.27 Answer Problem 8.24 for a comparison of groups A and C.

8.28 Answer Problem 8.25 for a comparison of groups A and C.

8.29 Answer Problem 8.26 for a comparison of groups A and C.

**Pulmonary Disease**
Forced expiratory volume (FEV) is a standard measure of pulmonary function representing the volume of air expelled in 1 second. Suppose we enroll 10 nonsmoking males ages 35–39, heights 68–72 inches in a longitudinal study and measure their FEV (L) initially (year 0) and 2 years later (year 2). The data in Table 8.5 are obtained.

<table>
<thead>
<tr>
<th>Person</th>
<th>FEV year 0 (L)</th>
<th>FEV year 2 (L)</th>
<th>Person</th>
<th>FEV year 0 (L)</th>
<th>FEV year 2 (L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.22</td>
<td>2.95</td>
<td>6</td>
<td>3.25</td>
<td>3.20</td>
</tr>
<tr>
<td>2</td>
<td>4.06</td>
<td>3.75</td>
<td>7</td>
<td>4.20</td>
<td>3.90</td>
</tr>
<tr>
<td>3</td>
<td>3.85</td>
<td>4.00</td>
<td>8</td>
<td>3.05</td>
<td>2.76</td>
</tr>
<tr>
<td>4</td>
<td>3.50</td>
<td>3.42</td>
<td>9</td>
<td>2.86</td>
<td>2.75</td>
</tr>
<tr>
<td>5</td>
<td>2.80</td>
<td>2.77</td>
<td>10</td>
<td>3.50</td>
<td>3.32</td>
</tr>
</tbody>
</table>

Mean 3.43 3.28
sd 0.485 0.480

8.30 What are the appropriate null and alternative hypotheses in this case to test if mean pulmonary function has decreased over 2 years?

8.31 In words, what is the meaning of a type I and a type II error here?

8.32 Carry out the test in Problem 8.30. What are your conclusions?


**Endocrinology**
A study was performed comparing the rate of bone formation between black and white adults. The data in Table 8.16 were presented [8].

<table>
<thead>
<tr>
<th></th>
<th>Blacks (n = 12)</th>
<th>Whites (n = 13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ± se</td>
<td>0.033 ± 0.007</td>
<td>0.095 ± 0.012</td>
</tr>
</tbody>
</table>

8.70 What method can be used to compare mean bone-formation rate between blacks and whites?

8.71 Implement the method in Problem 8.70 and report a p-value.

8.72 Obtain a 95% CI for the mean difference in bone-formation rate between the two groups.
7 Inference for proportions

1. It is known that 30\% of all Swedish children have blond hair.
   
   (a) A Swedish family, selected at random, has 4 children.
   
   i. Use the binomial distribution to estimate the probability that less than 3 of the 4 children in the family are blond.
   
   ii. Is the binomial distribution appropriate for this problem? Explain your answer.
   
   (b) A randomly selected school class contains 30 children.
   
   i. Use the normal approximation to the binomial distribution to estimate the probability that at least 10 of the children in the class are blond.
   
   ii. Is the normal approximation to the binomial distribution appropriate for this problem? Explain your answer.
   
   (c) In random samples of children from other countries, the following numbers of children were found to be blonde.

Table 2: Number of blond children in random samples from various countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of blond children</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>43</td>
<td>230</td>
</tr>
<tr>
<td>Australia</td>
<td>31</td>
<td>165</td>
</tr>
<tr>
<td>China</td>
<td>12</td>
<td>378</td>
</tr>
</tbody>
</table>

   i. Give a point estimate and 95\% confidence interval for the proportion of children in Germany who are blond.
   
   ii. Test whether there is evidence of a statistically significant difference in the proportion of children who are blond between Germany and China.
   
   iii. Test whether there is evidence of a statistically significant difference in the proportion of children who are blond between Germany and Sweden.
2. Assume that it is known that 15% of all males over the age of 20 currently living in Sweden are left handed. This information can be obtained, for example, from the ‘rifle test’ conducted during military service.

In order to study whether the prevalence of left handedness differs between males and females, a researcher obtained a random sample of 200 females over the age of 20 currently living in Sweden and measured handedness.

(a) Let \( X \) be the number of females in the sample who are left handed. If the null hypothesis (no difference in the prevalence of left handedness between males and females) is true, what statistical distribution can be used to describe \( X \)? [State the name and the parameters of the distribution]

(b) In the early and mid 1900s, many natural left handers were forced to become right handers. What implications could this have for the study?

(c) Assuming that 22 females in the sample were left handed, give a point estimate and interval estimate for the true proportion of left handed females in the population. State any assumptions you make and comment on the appropriateness of the assumptions.

(d) Based on the aforementioned study, is there evidence of a statistically significant difference in the prevalence of left handedness between males and females?

3. The following data were obtained from a randomised experiment conducted in 1952 where 36 albino mice of both sexes were enclosed in chamber that was filled with the smoke of one cigarette every hour of a twelve hour day. A comparable number of mice were kept for the same period in a similar chamber without smoke. After one year, an autopsy was performed on those mice which had survived for at least the first two months to assess whether lung cancer had developed.

<table>
<thead>
<tr>
<th>Group</th>
<th>Tumour Present</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>treated</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td>control</td>
<td>19</td>
<td>32</td>
</tr>
</tbody>
</table>

(a) Give a point estimate and 95% confidence interval for the proportion of mice which develop lung cancer in the treatment group.

(b) Give a point estimate and 95% confidence interval for the difference in the proportion of mice which develop lung cancer between the treatment and control group.

(c) Is there evidence of a statistically significant treatment effect?

4. The management of restaurant Jöns Jacob believe that, among all students who attend the Solna campus of KI on any given Wednesday, 75% purchase lunch at Jöns Jacob. In order to test this, they randomly selected a sample of all students who visited KI on Wednesday November 28 2001 and asked them where they at lunch.

(a) If it is true that 75% of students eat lunch at Jöns Jacob, what is the probability that exactly 15 of the 20 students surveyed purchased lunch at Jöns Jacob that day?

(b) Of the 20 students surveyed, 11 responded that they purchased lunch at Jöns Jacob that day. Calculate a point estimate and confidence interval for the proportion of students who purchase lunch at Jöns Jacob on Wednesdays.

(c) Based on the results of part (b), what do you think of the claim that 75% of students purchase lunch at Jöns Jacob on Wednesdays?
5. As a part of the same survey, the restaurant also surveyed 20 randomly selected staff members of whom 6 responded that they purchased lunch at Jöns Jacob that day. In order to compare whether there was a difference between staff and students, the following analysis was performed in Stata.

\[ \text{. csi 11 6 9 14} \]

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>11</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>Noncases</td>
<td>9</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Risk</td>
<td>.55</td>
<td>.3</td>
<td>.425</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk difference</td>
<td>.25, -.0464344 to .5464344</td>
</tr>
<tr>
<td>Risk ratio</td>
<td>1.833333, .8420736 to 3.991469</td>
</tr>
</tbody>
</table>

\[ \chi^2(1) = 2.56 \quad \text{Pr}>\chi^2 = 0.1098 \]

(a) Interpret the estimated risk ratio (i.e. write a sentence explaining what it means).

(b) Is there evidence of a difference between the proportion of staff who purchase lunch at Jöns Jacob and the proportion of students who purchase lunch at Jöns Jacob?

(c) Do you think it is sufficient to survey 20 staff and 20 students on a single day?

6. A study registered the number of miscarriages among 70 women who had been pregnant 4 times. The following distribution was observed.

<table>
<thead>
<tr>
<th>Number of miscarriages</th>
<th>0 1 2 3 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of women</td>
<td>24 28 7 5 6</td>
</tr>
</tbody>
</table>

Of the total 280 pregnancies, 81 resulted in a miscarriage (28.9%). Based on the binomial distribution with \( n = 4 \) and \( p = 0.289 \) we can calculate the probability of observing 0, 1, 2, 3, or 4 miscarriages in 4 pregnancies. We then assume that the 4 pregnancies for each woman represent a sample from a binomial distribution. By multiplying the probabilities by 70 we obtain estimates of the expected distribution of the number miscarriages from 4 pregnancies under the assumption that a binomial distribution is appropriate.

<table>
<thead>
<tr>
<th>Number of miscarriages</th>
<th>0 1 2 3 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.256 0.415 0.253 0.068 0.007</td>
</tr>
<tr>
<td>Expected number of women</td>
<td>17.86 29.08 17.75 4.82 0.49</td>
</tr>
<tr>
<td>Observed number of women</td>
<td>24 28 7 5 6</td>
</tr>
</tbody>
</table>

(a) Comment on possible reasons for differences between the empirical and theoretical (binomial) distribution. That is, why do the observed numbers differ from the expected numbers.

(b) How could we formally test whether a binomial distribution with \( p = 0.289 \) is appropriate?
7. In a certain country, the probability of becoming infected with HIV via a blood transfusion is 1%.

(a) Assume that a person receives a blood transfusion on 30 occasions. What is the probability that the person becomes infected with HIV at some time by the blood transfusion?

(b) What is the corresponding probability for a person who has received a blood transfusion on 100 occasions?

(c) Consider a test used to detect the presence of performance enhancing drugs in athletes. The test returns a ‘false positive’ 1% of the time. If an athlete is tested, on average, 100 times per year would you be surprised if one test per year was positive?

8. The following results were obtained from a study of blood groups in south-west Scotland.

<table>
<thead>
<tr>
<th>Blood Group</th>
<th>Eskdale</th>
<th>Annandale</th>
<th>Nithsdale</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>33</td>
<td>54</td>
<td>98</td>
<td>185</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>14</td>
<td>35</td>
<td>55</td>
</tr>
<tr>
<td>O</td>
<td>56</td>
<td>52</td>
<td>115</td>
<td>223</td>
</tr>
<tr>
<td>AB</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>125</td>
<td>253</td>
<td>478</td>
</tr>
</tbody>
</table>

We have been asked whether there is evidence that the distribution of blood groups differs between the three regions. To test this hypothesis we calculated the expected numbers in each group under the hypothesis that there was no association between blood group and region.

<table>
<thead>
<tr>
<th>Blood Group</th>
<th>Eskdale</th>
<th>Annandale</th>
<th>Nithsdale</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>38.70</td>
<td>48.38</td>
<td>97.92</td>
<td>185.00</td>
</tr>
<tr>
<td>B</td>
<td>29.11</td>
<td>55.00</td>
<td></td>
<td>55.00</td>
</tr>
<tr>
<td>O</td>
<td>118.03</td>
<td>223.00</td>
<td></td>
<td>223.00</td>
</tr>
<tr>
<td>AB</td>
<td>3.14</td>
<td>3.92</td>
<td>7.94</td>
<td>15.00</td>
</tr>
<tr>
<td>Total</td>
<td>100.00</td>
<td>125.00</td>
<td>253.00</td>
<td>478.00</td>
</tr>
</tbody>
</table>

We then calculated \((O - E)^2 / E\) for each cell.

<table>
<thead>
<tr>
<th>Blood Group</th>
<th>Eskdale</th>
<th>Annandale</th>
<th>Nithsdale</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.84</td>
<td>0.65</td>
<td>0.00</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td>1.19</td>
</tr>
<tr>
<td>O</td>
<td></td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>AB</td>
<td>1.10</td>
<td>0.30</td>
<td>1.09</td>
</tr>
</tbody>
</table>

(a) Complete the missing cells in the tables above.

(b) Calculate a test statistic for the test of independence between blood group and region. You should report the value of the test statistic, its distribution under the null hypothesis, and a p-value or critical value.

(c) Write a sentence summarising your findings.
9. The association between level of physical activity and body weight was investigated amongst randomly selected men aged 35–54 years living in Stockholm. The following results were observed:

<table>
<thead>
<tr>
<th>Weight</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>404</td>
<td>422</td>
<td>502</td>
<td>1328</td>
</tr>
<tr>
<td>Overweight</td>
<td>357</td>
<td>244</td>
<td>164</td>
<td>765</td>
</tr>
<tr>
<td>Total</td>
<td>761</td>
<td>666</td>
<td>666</td>
<td>2093</td>
</tr>
</tbody>
</table>

(a) Calculate the proportion of overweight men with low, medium, and high levels of physical activity respectively.

(b) Compare the proportion overweight between men with a low level of physical activity and men with a medium level of physical activity by calculating a difference in proportions. Calculate a 95% CI for the difference in proportions. Is the difference statistically significant?

(c) Make the corresponding comparison between men with a high level of physical activity and men with a low level of physical activity. Calculate a 95% CI for the difference in proportions. Is the difference statistically significant?

(d) Test the null hypothesis that the proportion of overweight men is the same for each of the three groups.

10. In an investigation of the association between snuff consumption and heart disease mortality, 135,000 male construction workers were observed over a 20 year period. Out of 6,300 men who reported using snuff, 220 died of heart disease during the period of observation. Of 32,500 men who did not use any form of tobacco, 640 died of heart disease during the period of observation.

(a) Calculate the proportion of deaths from heart disease for each of the two groups.

(b) Calculate the relative risk (RR) of dying of heart disease among those who used snuff compared to those who did not use any form of tobacco. Calculate a 95% CI for the RR. Is there evidence of a statistically significant difference in heart disease mortality between the two groups?

(c) Calculate the difference in the proportion of men dying of heart disease between the two groups. Calculate a 95% CI for the difference in proportions. Is there evidence of a statistically significant difference in heart disease mortality between the two groups?

(d) Does this provide evidence that using snuff increases heart disease mortality?
11. A new drug treatment is tested against an existing treatment in a randomised trial involving laboratory animals. The criteria for a positive response to the drug is clearance of blood parasites within 36 hours. The test gave the following results.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Total number of cases</th>
<th>Number of cases with 36-hour clearance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing</td>
<td>184</td>
<td>129</td>
</tr>
<tr>
<td>New</td>
<td>103</td>
<td>80</td>
</tr>
</tbody>
</table>

(a) Give a point estimate and 95% confidence interval for the proportion responding in each group.

(b) Compare the efficacy (i.e. effectiveness) of the two drugs by estimating a point estimate and a 95% confidence interval.
   i. Using the risk difference (difference in proportions responding).
   ii. Using the risk ratio.
   iii. Interpret the results.

(c) Compare the efficacy (i.e. effectiveness) of the two drugs using a formal hypothesis test. Be sure to state the null hypothesis, the alternative hypothesis, and to interpret the results in words.

12. Mendel crossed peas that were heterozygotes for smooth/wrinkled, where smooth is dominant. He observed 423 smooth and 133 wrinkled peas in the offspring.

(a) What is the expected proportion of smooth peas (assuming Mendelian inheritance)?
(b) Perform a ‘one-sample test of proportions’ to test whether the observed proportion of smooth is consistence with the assumption of Mendelian inheritance.
(c) We can think of this problem a 2 x 1 contingency table where the observed cell counts are 423 and 133. Calculate the expected cell counts and perform a $\chi^2$ test of independence.
(d) Do the two different tests lead to the same conclusion? Would you expect them to? Is there a mathematical relation between the test statistics?

Cancer
A case-control study of the effectiveness of the Pap smear in preventing cervical cancer (by identifying precancerous lesions) was performed [1]. It was found that 28.1% of 153 cervical-cancer cases and 7.2% of 153 age-matched (within 5 years) controls had never had a Pap smear prior to the time of the case's diagnosis.

6.5 Provide a 95% CI for the percentage of cervical-cancer cases who never had a Pap test.

6.6 Provide a 95% CI for the percentage of controls who never had a Pap test.

6.7 Do you think the Pap test is helpful in preventing cervical cancer?


Cancer

7.26 An area of current interest in cancer epidemiology is the possible role of oral contraceptives (OC’s) in the development of breast cancer. Suppose that in a group of 1000 premenopausal women ages 40–49 who are current users of OC’s, 15 subsequently develop breast cancer over the next 5 years. If the expected 5-year incidence rate of breast cancer in this group is 1.2% based on national incidence rates, then test the hypothesis that there is an association between current OC use and the subsequent development of breast cancer.

Cancer
Radiotherapy is a common treatment for breast cancer, with about 25% of women receiving this form of treatment. Assume that the figure 25% is based on a very large sample and is known without error. One hypothesis is that radiotherapy applied to the contralateral breast may be a risk factor for development of breast cancer in the opposite breast 5 or more years after the initial cancer. Suppose that 655 women are identified who developed breast cancer in the opposite breast 5 or more years after the initial cancer [2].

7.46 If 206 of the women received radiotherapy after their initial diagnosis, then test the hypothesis that radiotherapy is associated with the development of breast cancer in the opposite breast. Please report a \( p \)-value.

7.47 Provide a 95% CI for the true proportion of women with contralateral breast cancer who received radiotherapy treatment.

7.48 Suppose that our \( p \)-value in Problem 7.46 = .03 (this is not necessarily the actual \( p \)-value in Problem 7.46). If we conduct a test using the critical-value method with \( \alpha = .05 \), then would we accept or reject \( H_0 \) and why? (Do not actually conduct the test.)


Cardiology
A group of patients who underwent coronary angiography between Jan. 1, 1972 and Dec. 31, 1986 in a particular hospital were identified [7]. 1493 cases with confirmed coronary-artery disease were compared with 707 controls with no plaque evidence at the time of angiography. Suppose it is found that 37% of cases and 30% of controls reported a diagnosis or treatment for hypertension at the time of angiography.

10.44 Are the proportions (37%, 30%) an example of prevalence, incidence, or neither?

10.45 What test can be used to compare the risk of hypertension between cases and controls?

10.46 Implement the test in Problem 10.45 and report a \( p \)-value.
8 Selected solutions

8.1 Solutions – Exercises on probability

1. (a) No. Since

\[ P(A) \times P(B) = 0.142 \times 0.051 \]
\[ = 0.0072 \]
\[ \neq P(A \cap B) \]
\[ = 0.031, \]

these two events are not independent.

(b) The probability that A or B or both occur is

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]
\[ = 0.142 + 0.051 - 0.031 \]
\[ = 0.162. \]

(c) The probability that A occurs given that B occurs is

\[ P(A \mid B) = \frac{P(A \cap B)}{P(B)} \]
\[ = \frac{0.031}{0.051} \]
\[ = 0.608. \]

2. (a) The probability that a woman who gave birth in 1992 was 24 years of age or younger is

\[ P(\leq 24) = P(< 15 \text{ or } 15–19 \text{ or } 20–24) \]
\[ = P(< 15) + P(15–19) + P(20–24) \]
\[ = 0.003 + 0.124 + 0.263 \]
\[ = 0.390. \]

(b) The probability that the woman was 40 years of age or older is

\[ P(\geq 40) = P(40–44 \text{ or } 45–49) \]
\[ = P(40–44) + P(45–49) \]
\[ = 0.014 + 0.001 \]
\[ = 0.015. \]

(c) Given that the woman was under 30 years of age, the probability that she was not yet 20 is

\[ P(< 20 \mid < 30) = \frac{P(< 20 \text{ and } < 30)}{P(< 30)} \]
\[ = \frac{P(< 20)}{P(< 30)} \]
\[ = \frac{0.003 + 0.124}{0.003 + 0.124 + 0.263 + 0.290} \]
\[ = 0.187. \]
Given that the woman was 35 years of age or older, the probability that she was under 40 is

\[
P(< 40 | \geq 35) = \frac{P(< 40 \text{ and } \geq 35)}{P(\geq 35)}
= \frac{P(35 - 39)}{P(\geq 35)}
= \frac{0.085}{0.085 + 0.014 + 0.001}
= 0.850.
\]

3. (a) Since these events are independent, the probability that both adults are uninsured is

\[
P(\text{both uninsured}) = P(\text{woman uninsured}) \times P(\text{man uninsured})
= 0.123 \times 0.123
= 0.015.
\]

(b) The probability that both adults are insured is

\[
P(\text{both insured}) = (1 - 0.123) \times (1 - 0.123)
= 0.877 \times 0.877
= 0.769.
\]

(c) The probability that all five adults are uninsured is

\[
0.123 \times 0.123 \times 0.123 \times 0.123 \times 0.123 = 0.000028.
\]

4. (a) The probability of a false negative result is

\[
P(\text{negative test } | \text{cancer}) = 1 - \text{sensitivity}
= 1 - 0.85
= 0.15.
\]

(b) The probability of a false positive result is

\[
P(\text{positive test } | \text{no cancer}) = 1 - \text{specificity}
= 1 - 0.80
= 0.20.
\]

(c) Since \(P(\text{cancer}) = 0.0025\) and \(P(\text{no cancer}) = 0.9975\), the probability that a woman has breast cancer given that her mammogram is positive is

\[
P(\text{cancer } | \text{+test}) = \frac{P(\text{cancer})P(\text{+test } | \text{cancer})}{P(\text{cancer})P(\text{+test } | \text{cancer}) + P(\text{no cancer})P(\text{+test } | \text{no cancer})}
= \frac{(0.0025)(0.85)}{(0.0025)(0.85) + (0.9975)(0.20)}
= 0.0105.
\]
5. (a) Note that

\[ P(\text{+ test } | \text{ cts}) = \text{sensitivity} = 0.67 \]

and

\[ P(\text{+ test } | \text{ no cts}) = 1 - \text{ specificity} = 1 - 0.58 = 0.42. \]

If the prevalence of carpal tunnel syndrome is 15\%, then \( P(\text{cts}) = 0.15 \) and \( P(\text{no cts}) = 0.85 \). The predictive value of a positive test result is

\[
P(\text{cts} | \text{+ test}) = \frac{P(\text{cts}) P(\text{+ test } | \text{ cts})}{P(\text{cts}) P(\text{+ test } | \text{ cts}) + P(\text{no cts}) P(\text{+ test } | \text{ no cts})} = \frac{0.15(0.67)}{0.15(0.67) + 0.85(0.42)} = 0.22.
\]

(b) If the prevalence is 10\%, then \( P(\text{cts}) = 0.10 \) and \( P(\text{no cts}) = 0.90 \). The predictive value of a positive test result is

\[
P(\text{cts} | \text{+ test}) = \frac{(0.10)(0.67)}{(0.10)(0.67) + (0.90)(0.42)} = 0.15.
\]

If the prevalence is 5\%, then \( P(\text{cts}) = 0.05 \) and \( P(\text{no cts}) = 0.95 \), and the predictive value of a positive test result is

\[
P(\text{cts} | \text{+ test}) = \frac{(0.05)(0.67)}{(0.05)(0.67) + (0.95)(0.42)} = 0.08.
\]

As the prevalence of carpal tunnel syndrome decreases, the predictive value of a positive test decreases as well.
8.2 Solutions – exercises on the binomial distribution

1. It is unlikely that $X$ has a binomial distribution. While there are a fixed number of trials ($n = 7$) that each result in one of two mutually exclusive outcomes, the outcomes of the trials are not independent. If the concentration of carbon monoxide is exceptionally high one day, the pollution will not all disappear over night; the concentration is more likely to be high the next day as well.

2. (a) The seven individuals can be ordered in $7! = 5040$ ways.

   (b) Four individuals can be selected from the group in
   
   $$\binom{7}{4} = \frac{7!}{4!(7-4)!} = 35$$
   
   different ways.

   (c) The probability that exactly two of the seven individuals suffer from diabetes is
   
   $$P(\text{two diabetics}) = \binom{7}{2} (0.125)^2 (0.875)^{7-2} = 0.168.$$  

   (d) The probability that exactly four of the seven persons suffer from diabetes is
   
   $$P(\text{four diabetics}) = \binom{7}{4} (0.125)^4 (0.875)^{7-4} = 0.006.$$  

3. This is a standard application of the binomial distribution.

<table>
<thead>
<tr>
<th>$r$</th>
<th>$\Pr(X = r)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.422</td>
</tr>
<tr>
<td>1</td>
<td>0.422</td>
</tr>
<tr>
<td>2</td>
<td>0.141</td>
</tr>
<tr>
<td>3</td>
<td>0.016</td>
</tr>
</tbody>
</table>

For example, the probability that 1 of the three mice is infected is given by

$$\Pr(X = r) = \frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}, \text{ where } n = 3, r = 1, \text{ and } p = 0.25.$$  

$$\Pr(X = 1) = \frac{3!}{1!(3-1)!}0.25^1(1-0.25)^{3-1} = \frac{3!}{1! \times 2!} \times 0.250.75^2 = \frac{3 \times 2 \times 1}{1 \times 2} \times 0.25 \times 0.75^2 = 0.422$$
4. (a) 
   i. 0.0085  
   ii. 0.0013  
   iii. 0.0098  

   (b) 
   \[ \Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, \]  
   where \( n = 10, r = 1, \) and \( p = 0.15. \) 

   \[ \Pr(X = 1) = \frac{10!}{1!(10-1)!} 0.15^1 (1-0.15)^{10-1} \]
   \[ = \frac{10!}{9!} 0.15 \times 0.85^9 \]
   \[ = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \times 0.15 \times 0.85^9 \]
   \[ = 0.3474 \]

5. (a) Since the number of individuals in the sample who do not exercise regularly follow a binomial distribution, the mean number per sample is \( np = 12(0.58) = 7.0 \) and the standard deviation is \( \sqrt{np(1-p)} = \sqrt{12(0.58)(0.42)} = 1.7. \)  

   (b) The probability of obtaining results as bad as or worse than those observed is 

   \[ P(\text{ten or more}) = P(\text{ten}) + P(\text{eleven}) + P(\text{twelve}) \]
   \[ = \binom{12}{10} (0.58)^{10} (0.42)^2 + \binom{12}{11} (0.58)^{10} (0.42)^1 + \binom{12}{12} (0.58)^{12} (0.42)^0 \]
   \[ = 0.064. \]

6. (a) \( X \) has a binomial distribution with parameters \( n = 6 \) and \( p = 0.15. \) This describes a situation where we have a sample of size 6 and \( X \) represents the total number of items in the sample with a specified characteristic. We know that 15% of the items in the population from which the sample was drawn have the specified characteristic. In the sample, we can observe \( r = 0, 1, 2, \ldots, 6 \) items with the specified characteristic. These outcomes occur with probabilities defined by the formula for the binomial distribution. 

   \[ \Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, \]  
   for \( r = 0, 1, 2, 3, 4, 5, 6. \)
For example, to calculate the probability that zero items in the sample have the specified characteristic (i.e. $X = 0$).

\[
\Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, \text{ for } r = 0, 1, 2, 3, 4, 5, 6.
\]

\[
\Pr(X = 0) = \frac{6!}{0!(6-0)!} 0.15^0 (1-0.15)^{6-0}
= \frac{6!}{1} 0.15^0 0.85^6
= 0.85^6
= 0.3771
\]

Note that $0! = 1$ by definition and that any number raised to the power 0 is equal to 1 (i.e. $0^0 = 1$).

To calculate the probability that two items in the sample have the specified characteristic (i.e. $X = 2$).

\[
\Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, \text{ for } r = 0, 1, 2, 3, 4, 5, 6.
\]

\[
\Pr(X = 2) = \frac{6!}{2!(6-2)!} 0.15^2 (1-0.15)^{6-2}
= \frac{6!}{2!4!} 0.15^2 0.85^4
= \frac{6 \times 5}{2} 0.15^2 0.85^4
= 15 \times 0.15^2 \times 0.85^4
= 0.1762
\]

\[
\begin{array}{c|c}
  r & \Pr(X = r) \\
  \hline
  0 & 0.3771 \\
  1 & 0.3993 \\
  2 & 0.1762 \\
  3 & 0.0415 \\
  4 & 0.0055 \\
  5 & 0.0004 \\
  6 & 0.0000 \\
\end{array}
\]

(b) i. $0.3771 + 0.3993 + 0.1762 = 0.9527$

ii. $E(X) = np = 6 \times 0.15 = 0.9$

iii. 1

iv. The binomial distribution is probably not strictly appropriate since the engine failures will not be independent. As each engine fails, the remaining engines are under a greater load so are more likely to fail.
7. Let $X$ be a random variable representing the total number of heads tossed by the 5 students in one round of tosses. Then $X$ can be described by a binomial distribution with parameters $n = 5$ and $p = 0.5$. Note that we could also let $X$ represent the number of tails.

(a) In order for the pizza buyer to be determined we require either one head and four tails (i.e. $X = 1$) or one tail and four heads (i.e. $X = 4$).

In general, $\Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$, for $r = 0, 1, 2, 3, 4, 5$.

For 1 head, $\Pr(X = 1) = \frac{5!}{1!(5-1)!} 0.5^1 (1-0.5)^{5-1} = 5 \times 0.5 \times 0.5^4 = 0.156$

$\Pr(X = 4)$ is identical to $\Pr(X = 1)$, due to the symmetry of the binomial distribution when $p = 0.5$. Another way of thinking of this is that $\Pr(X = 4)$ is the same as $\Pr(W = 1)$ where $W$ is the number of tails in the sample.

The probability that the pizza buyer is determined on the first toss is $\Pr(X = 1) + \Pr(X = 4) = 0.156 + 0.156 = 0.312$.

(b) In order for it to require exactly 2 rounds of tosses to determine the pizza buyer, we need to not decide the pizza buyer on the first toss ($X = 0, 2, 3, 5$) but then determine the pizza buyer on the second toss.

The probability that we do not decide the pizza buyer on the first toss is $1 - 0.312 = 0.688$.

The probability that the pizza buyer is determined during the second toss (given that the first toss did not determine the buyer) is 0.312 (from part (a)).

The probability that it requires exactly 2 rounds of tosses to determine the pizza buyer is the product of these two probabilities (since they are independent and they both must occur).

$\Pr(\text{exactly 2 rounds of tosses are required}) = 0.688 \times 0.312 = 0.215$.

8. This situation can be described by a binomial distribution with $n = 6$. Note that the probability that the stock rises on a given day is $\frac{3}{4}$ independent of what has happened previously. For the stock to be the same price after 6 days it must go up 3 days and down 3 days. That is, we require the probability that $r = 3$ for a binomial distribution with $n = 6$ and $p = 0.75$ (or, equivalently, $p = 0.25$).

$\Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$

$\Pr(X = 3) = \frac{6!}{3!(6-3)!} 0.75^3 0.25^3 = 0.132$

```
. tablesq B 6 3 0.25
```

B(6, 0.25) = 3
$\Pr(k == 3) = 0.1318$
$\Pr(k >= 3) = 0.1694$
$\Pr(k <= 3) = 0.9624$
8.3 Solutions – Exercises on the normal distribution

1. (a) The probability that $z$ is greater than 2.60 is 0.005.

(b) The probability that $z$ is less than 1.35 is $1 - 0.089 = 0.911$.

(c) The probability that $z$ is between $-1.70$ and $3.10$ is $1 - 0.045 - 0.001 = 0.954$.

(d) The value $z = 1.04$ cuts off the upper 15% (actually, 14.9%) of the standard normal distribution.

(e) The value $z = -0.84$ cuts off the lower 20% of the distribution.

2. The exact answer is $\Pr(Z > 2.0) = 0.023$. Since 2.0 is close to 1.96 and we know that $\Pr(Z > 1.96) = 0.025$ then an answer of 0.025 would also be acceptable.

3. (a)

$$\Pr(X < 172) = \Pr(Z < \frac{172 - 175.8}{5.84}) \text{ where } Z \sim N(0, 1)$$

$$= \Pr(Z < -0.65) = 0.258$$

25.8% of men are too short to join the police force.

(b)

$$\Pr(X < 172) = \Pr(Z < \frac{172 - 179.1}{5.84}) \text{ where } Z \sim N(0, 1)$$

$$= \Pr(Z < -1.22) = 0.111$$

11.1% of men are too short to join the police force.

4. (a) i. The blood pressure of patient 1 is represented by the random variable $X$, where $X \sim N(85, 10^2)$. That is, patient 1 is not hypertensive according to our criteria. We will misclassify patient 1 as hypertensive if the single blood pressure measurement is greater than 90.

[Technical aside: Classifying a patient based on a single classification is not in accordance with the criterion, which states that it is the average blood pressure we must study (although the criterion does not mention the crucial details of how many individual measurements should be taken and over what period).]

Recall that $\Pr(X \geq x) = \Pr(Z \geq \frac{x - \mu}{\sigma})$ where $X \sim N(\mu, \sigma^2)$ and $Z$ is a random variable having a standard normal distribution ($Z \sim N(0, 1)$). The probability of misclassifying patient 1 as hypertensive is given by:

$$\Pr(X \geq 90) = \Pr(Z \geq \frac{90 - 85}{10})$$

$$= \Pr(Z \geq 0.5)$$

$$= \Pr(Z \leq -0.5) = 0.3085$$

In Stata:

$. display normprob(-0.5)

.30853754

Consequently, the probability of misclassifying patient 1 as hypertensive, based on a single blood pressure measurement, is 0.3085.

[Technical aside: $\Pr(X = x) = 0$ for a continuous random variable $X$ and a real constant (measured exactly to an infinite number of decimal places) $x$. As such $\Pr(X \geq 90)$ is equivalent to $\Pr(X > 90)$. That is, $\Pr(X = 90) = 0$, the probability that the blood pressure measurement is exactly 90 mm Hg is zero. When we say, in practice, ‘the blood pressure measurement was 90 mm Hg’, we actually mean that the true measurement was between, say 89.5 and 90.5, for which there is a non-zero probability.]
ii. The blood pressure of patient 2 is represented by the random variable $X$, where $X \sim N(92, 5^2)$. That is, patient 1 is hypertensive according to our criteria. We will misclassify patient 1 as not hypertensive, based on a single measurement, if that measurement is less than 90 mm Hg.

$$\Pr(X < 90) = \Pr(Z < \frac{90 - 92}{5}) = \Pr(Z < -0.4) = 0.3446$$

In Stata:

```
. display normprob(-0.4)
.34457826
```

Consequently, the probability of misclassifying patient 2 as not hypertensive, based on a single blood pressure measurement, is 0.3446.

(b) i. The probability of being within 0.5 standard deviations from the mean is 0.383.

ii. The probability of being greater than 0.5 standard deviations from the mean is $1 - 0.383 = 0.617$.

5. (a) The probability that a randomly selected man weighs less than 130 pounds is

$$P(X < 130) = P \left( \frac{X - 172.2}{29.8} < \frac{130 - 172.2}{29.8} \right) = P(Z < -1.42) = 0.078.$$  

(b) The probability that he weighs more than 210 pounds is

$$P(X > 210) = P \left( \frac{X - 172.2}{29.8} > \frac{210 - 172.2}{29.8} \right) = P(Z > 1.27) = 0.102.$$  

(c) Among five males selected at random, the probability that at least one will have a weight outside the range 130 to 210 pounds is

$$P(\text{at least one } < 130 \text{ or } > 210) = 1 - P(\text{none } < 130 \text{ or } > 210) = 1 - P(\text{all between } 130 \text{ and } 210) = 1 - [P(130 \leq X \leq 210)]^5 = 1 - [1 - 0.78 - 0.102]^5 = 1 - (0.820)^5 = 0.629.$$
6. (a) The probability of correctly predicting coronary heart disease for a man who will develop it is

\[
P(X_d \geq 260) = P\left( \frac{X_d - 244}{51} \geq \frac{260 - 244}{51} \right) = P(Z \geq 0.31) = 0.378.
\]

(b) The probability of predicting heart disease for a man who will not develop it, or the probability of a false positive, is

\[
P(X_{nd} \geq 260) = P\left( \frac{X_{nd} - 219}{41} \geq \frac{260 - 219}{41} \right) = P(Z \geq 1.00) = 0.159.
\]

(c) The probability of failing to predict heart disease for a man who will develop it, or the probability of a false negative, is

\[
P(X_d < 260) = 1 - P(X_d \geq 260) = 1 - 0.378 = 0.622.
\]

(d) If the cutoff point is lowered to 250 mg/100 ml, the probability of a false positive result would increase while the probability of a false negative would decrease.

(e) Initial serum cholesterol level is not very useful for predicting coronary heart disease in this population. Because the normal curves for men who develop disease and those who do not have a great deal of overlap, the probabilities of false positive and false negative outcomes are both very high.
8.4 Solutions — Sampling/study design

1. (a) Nej, det är definitivt fel att bara intervjuva de som klassificerades som ”regular drinkers” vid första intervjun. Om man vill se om en grupp personer har förändrat sitt beteende (här alkoholkonsumtion) måste man se på förändringen i hela gruppen. Att bara studera en utvald del kan resultera i vilseledande resultat.

(b) Det var bara 69% av de tillfrågade som svarade i den andra studien. Det är mycket möjligt att de som inte svarade skiljer sig från de som svarade i avseende på deras alkoholkonsumtion. Kanske är det de som dricker mest alkohol som inte svarar, p.g.a. av de är sjuka (kan vara någon alkohol-relaterad sjukdom t.ex.) eller av någon annan anledning. Den höga andelen av personer som inte svarade kan ha påverkat resultaten.

(c) Ja, det hade varit mycket bättre att titta på alkoholkonsumtionen vid samma tid på året vid de två tillfällena eftersom alkoholvanorna ser olika ut för olika årstider.

(d) Nej. För det första kan vi inte dra en slutsats om orsakssamband utifrån att vi ser att det finns ett samband mellan två förändringar. För det andra kan vi förvänta oss att se en minskad alkoholkonsumtion eftersom vi endast valde att titta på de som klassificerades som ”regular drinkers” i den första studien. Hade man istället bara tittat på de som inte drack vid första studien hade man antagligen sett att alkoholkonsumtionen hade ökat, d.v.s. att några hade börjat dricka alkohol.

(e) Nej, av anledningarna beskrivna ovan är de inte ett representativt urval. Dessutom utgör de ett stickprov och inte en population.

(f) Tolkningen att det skedde en minskning av alkoholkonsumtionen, vilket orsakades av relativt högre kostnader på alkohol är dåligt underbyggd p.g.a. resonemanget i uppgift a)-d).

(g) Nej.
8.5 Solutions — One-sample inference for means

1.
8.6 Solutions — Two-sample inference for means

1. (a) Two sample t-test with equal variances
   
   (b) \( H_0: \mu_{\text{men}} = \mu_{\text{women}}, \) ie. mean cholesterol level is the same for males as for females
   
   \[ H_A: \mu_{\text{men}} \neq \mu_{\text{women}}, \] ie. mean cholesterol level is NOT the same for males as for females

   \[
   t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{-17.7}{\sqrt{44.1^2\left(\frac{1}{44} + \frac{1}{48}\right)}} = -1.92
   \]

   nb: \( s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}, \) but since in this case \( s_1^2 = s_2^2, \) then, simply, \( s_p^2 = 44.1^2 \)

   Compare to \( t_{90} \) (df=44+48-2=90).
   
   \( 0.05 < p < 0.10 \) (p=0.058) ie. Insufficient evidence to reject \( H_0 \)

   (c) No, because measurements on twins are not independent!

2. (a) They would like to be able to pool the data for men and women to increase power. Cannot do this if there is evidence that disease etiology (or other factors of interest) differs between men and women.

   (b)

   (c) We would use a two-sample t-test (assuming equal variance) to test the hypotheses
   
   \( H_0: \mu_1 = \mu_2 \) vs \( H_A: \mu_1 \neq \mu_2. \) The test statistic would have a \( t \) distribution with \( 34 + 61 - 2 = 93 \) df for the affected individuals and \( 62 + 77 - 2 = 137 \) df for the unaffected individuals.

   (d) For the affected individuals:

   \[ . \texttt{ttesti 34 71.8 8.4 61 70.4 8.0} \]

   Two-sample t test with equal variances

   ---------------------------------------------------------------------------------------------------------------------
<table>
<thead>
<tr>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>34</td>
<td>71.8</td>
<td>1.440588</td>
<td>8.4</td>
</tr>
<tr>
<td>y</td>
<td>61</td>
<td>70.4</td>
<td>1.024295</td>
<td>8</td>
</tr>
<tr>
<td>combined</td>
<td>95</td>
<td>70.90105</td>
<td>8.33997</td>
<td>8.128797</td>
</tr>
<tr>
<td>diff</td>
<td>1.4</td>
<td>1.743031</td>
<td>-2.061313</td>
<td>4.861313</td>
</tr>
</tbody>
</table>
   ---------------------------------------------------------------------------------------------------------------------

   Degrees of freedom: 93

   Ho: mean(x) - mean(y) = diff = 0

   Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

   \[
   t = 0.8032 \quad t = 0.8032 \quad t = 0.8032
   \]

   \[
   p < t = 0.7680 \quad p > |t| = 0.4239 \quad p > t = 0.2120
   \]
For the unaffected individuals:

```
. ttesti 62 77.7 8 77 82.6 9.2
```

Two-sample t test with equal variances

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>62</td>
<td>77.7</td>
<td>1.016001</td>
<td>8</td>
<td>75.66838  79.73162</td>
</tr>
<tr>
<td>y</td>
<td>77</td>
<td>82.6</td>
<td>1.048437</td>
<td>9.2</td>
<td>80.51186  84.68814</td>
</tr>
<tr>
<td>combined</td>
<td>139</td>
<td>80.41439</td>
<td>.762796</td>
<td>8.993263</td>
<td>78.9061  81.92267</td>
</tr>
</tbody>
</table>

|       | diff | 1.482162 | -7.830873 | -1.969127 |

Degrees of freedom: 137

Ho: mean(x) - mean(y) = diff = 0

Ha: diff < 0

\[ t = -3.3060 \]
\[ P < t = 0.0006 \]

Ha: diff ≠ 0

\[ t = -3.3060 \]
\[ P > |t| = 0.0012 \]

Ha: diff > 0

\[ t = -3.3060 \]
\[ P > t = 0.9994 \]

There is very strong evidence that the mean ages differ between men and women for the unaffected individuals – the authors claim is suspect! (or there is a typo in the table)

3. The 95% confidence interval for the active mouthwash contains zero, the 95% confidence interval for the placebo does not. So, active mouthwash has no significant effect, but the placebo does - an unexpected result!

The two-sample t-test rejects \( H_0 : \mu_1 = \mu_2 \) in favour of \( H_A : \mu_1 \neq \mu_2 \) (\( p=0.0238 \)). conclude that the placebo has a more positive effect than the active mouthwash! (do we really believe this?)

4. (a) \( H_0 : \mu_d = 0 \), ie. no difference in concentration of the drug between aspirin types A & B.

\( H_A : \mu_d \neq 0 \), ie. a difference

(b) paired t-test

<table>
<thead>
<tr>
<th>Person</th>
<th>Aspirin A 1-hour concentration (mg%)</th>
<th>Aspirin B 1-hour concentration (mg%)</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>22</td>
<td>-2</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>36</td>
<td>30</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

| mean   | 19.20                               | 15.60                               | 3.60       |
| sd     | 8.63                                | 7.78                                | 3.098      |

\[ t = \frac{3.60}{3.098} = 3.67 \]

From tables: \( t_{0.025} = 2.262 \) & \( t_{0.005} = 3.250 \), so \( p < 0.01 \) since 3.67 > 3.25. We reject \( H_0 \) and conclude that there is a difference in concentrations (higher in A).
Paired t test

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AspirinA</td>
<td>10</td>
<td>19.2</td>
<td>2.727636</td>
<td>8.625543</td>
<td>13.02966 25.37034</td>
</tr>
<tr>
<td>AspirinB</td>
<td>10</td>
<td>15.6</td>
<td>2.459449</td>
<td>7.77746</td>
<td>10.03634 21.16366</td>
</tr>
<tr>
<td>diff</td>
<td>10</td>
<td>3.6</td>
<td>.9797959</td>
<td>3.098387</td>
<td>1.383548 5.816452</td>
</tr>
</tbody>
</table>

Ho: mean(AspirinA - AspirinB) = mean(diff) = 0

Ha: mean(diff) < 0    Ha: mean(diff) != 0    Ha: mean(diff) > 0

\[
t = 3.6742 \quad t = 3.6742 \quad t = 3.6742
\]

\[
P < t = 0.9974 \quad P > |t| = 0.0051 \quad P > t = 0.0026
\]

(c) 3.60

(d) 95% confidence interval is (1.38, 5.82), which does not include 0.
8.7 Solutions — Inference for proportions

1. (a) i.

<table>
<thead>
<tr>
<th>$r$</th>
<th>$Pr(X = r)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2401</td>
</tr>
<tr>
<td>1</td>
<td>0.4116</td>
</tr>
<tr>
<td>2</td>
<td>0.2646</td>
</tr>
<tr>
<td>3</td>
<td>0.0756</td>
</tr>
<tr>
<td>4</td>
<td>0.0081</td>
</tr>
</tbody>
</table>

$Pr(X < 3) = Pr(X = 0) + Pr(X = 1) + Pr(X = 2)$.
Alternatively, $Pr(X < 3) = 1 - Pr(X = 3) - Pr(X = 4)$.
The probability that less than 3 children are blond is 0.9163.

ii. The binomial distribution is not appropriate since the four children in a family are not independent with respect to hair colour. If both the mother and father have blond hair then each of the children will be more likely to have blond hair than if the parents had dark hair.

(b) i. Using the normal approximation to the binomial distribution, we assume that the number of children with blond hair, denoted by $X$, can be described by a normal distribution with $E(X) = np = 9$ and $Var(X) = np(1 - p) = 6.3$.
You can perform the calculations with or without a continuity correction. Without a continuity correction we have

$$Pr(X > 10) = Pr(Z > \frac{10 - 9}{\sqrt{6.3}}) \text{ where } Z \sim N(0,1)$$
$$= Pr(Z > 0.398)$$
$$= 0.3453$$

With a continuity correction we have

$$Pr(X > 9.5) = Pr(Z > \frac{9.5 - 9}{\sqrt{6.3}}) \text{ where } Z \sim N(0,1)$$
$$= Pr(Z > 0.199)$$
$$= 0.4207$$

ii. The normal approximation to the binomial distribution is appropriate for this problem since both $np$ and $n(1 - p)$ are greater than 5.

(c) i. A point estimate for the proportion of children in Germany who are blond is $43/230 = 0.187$. The variance is

$$\frac{p(1-p)}{n} = 0.00066.$$ 

A 95% confidence interval for the proportion of children in Germany who are blond is given by

$$0.187 \pm 1.96 \times \sqrt{0.00066} = (0.137, 0.237)$$

ii. The simplest approach to test whether there is evidence of a statistically significant difference in the proportion of children who are blond between Germany and China is to estimate a CI for the difference in proportions.
A point estimate for the difference is $43/230 - 12/378 = 0.155$. The variance is

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = 0.0007413$$

A 95% confidence interval for the difference in proportions is given by

$$0.155 \pm 1.96 \times \sqrt{0.0007413} = (0.102, 0.209)$$
Since the CI does not contain 0 we conclude that there is a statistically significant difference in the proportion of children who are blond between Germany and China.

Another approach is to formally test the null hypothesis, $H_0 : p_1 - p_0 = 0$ vs. the two-sided alternative $H_1 : p_1 - p_0 \neq 0$.

When formally testing a hypothesis, the variance should be estimated under the null hypothesis. The common value of the proportion under $H_0$, $\hat{p}_c$ is estimated as

$$\hat{p}_c = \frac{r_1 + r_0}{n_1 + n_0} = \frac{43 + 12}{230 + 378} = 0.0904$$

and the variance of $p_1 - p_0$ under $H_0$ is estimated by

$$\text{var}_{H_0}(p_1 - p_0) = \hat{p}_c(1 - \hat{p}_c) \frac{n_1}{n_1 + n_0} + \hat{p}_c(1 - \hat{p}_c) \frac{n_0}{n_1 + n_0} = 5.75 \times 10^{-4}$$ (1)

The standard error is $\sqrt{\text{var}_{H_0}(p_1 - p_0)} = \sqrt{5.75 \times 10^{-4}} = 0.02399$

The test statistic is then

$$\frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\text{var}_{H_0}(p_1 - p_0)}} = \frac{0.155}{0.02399} = 6.47$$

which will have a standard normal distribution under $H_0$. $\Pr(Z > 6.46)$ is less than 0.01 so we reject $H_0$ at the 1% level. We conclude that there is a statistically significant difference in the proportion of children who are blond between Germany and China.

In Stata, use the ‘Calculator / 2-sample test of proportions’ menu:

```
. prtesti 230 .1869565217391304 378 .0317460317460317, level(95)
```

Two-sample test of proportion

| Variable | Mean Std. Err. z  P>|z| [95% Conf. Interval] |
|----------|--------|--------|--------|------------------|
| x | .1869565 | .0257077 | 7.27239 | 0.0000 | .1365703 | .2373427 |
| y | .031746 | .0090177 | 3.52043 | 0.0004 | .0140718 | .0494203 |

| diff | .1552105 | .0272434 | .1018144 | .2086066 |

Ho: proportion(x) - proportion(y) = diff = 0

Ha: diff < 0 Ha: diff = 0 Ha: diff > 0
z = 6.471 z = 6.471 z = 6.471
P < z = 1.0000 P > |z| = 0.0000 P > z = 0.0000

49
We are now comparing the proportion in the sample with a known proportion (more correctly, assumed to be known). Note that the variance must be calculated under $H_0$ (i.e. where $p = 0.3$). In Stata, use the ‘Calculator / 1-sample test of proportions’ menu:

```
.bintest 230 43 0.3 , normal
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Proportion</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>230</td>
<td>0.1869565</td>
<td>0.0257077</td>
</tr>
</tbody>
</table>

$H_0: p = 0.3$  
$z = -3.74$  
$Pr > |z| = 0.0002$  
95% CI = (0.1366,0.2373)

2. (a) $X$ will have a binomial distribution with parameters $n = 200$ and $p = 0.15$.

(b) This means that, in the current population, younger people are more likely to be left handed than older people. The most important implication is that age (or, equivalently, birth cohort) will confound the association between gender and handedness. If gender is considered the exposure and left handedness the outcome then we know that age is associated with the outcome. We also know that age is associated with the exposure, since females live longer than men. Therefore, simply taking a random sample of females from the population will give a biased result. We have to either adjust for age in the analysis or use a matched design.

I don’t see any reason to believe that age would be an effect modifier. Age would be an effect modifier if, for example, left handed boys were forced to become right handed but left handed girls were not.

Another implication is that the binomial distribution is not strictly appropriate because the probability of being left handed is not the same for every individual in the population (it depends on age).

(c) Assume that the binomial distribution is appropriate. This is not strictly true (see the previous part) but for practical purposes the binomial distribution can be considered appropriate.

A point estimate for the true proportion of left handed females in the population is $\hat{p} = 22/200 = 0.11$.

A 95% confidence interval for $p$, assuming a normal approximation, is given by

$$\left(\hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

where $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0221246$.

A 95% confidence interval for $p$ is therefore (0.0666,0.1534).

The normal approximation to the binomial is appropriate because both $np$ and $n(1-p)$ are greater than 5.

(d) The confidence interval calculated in the previous part contains the null value (0.15), indicating that there is no evidence of a statistically significant difference at the 5% level. However, the null value is only just inside the CI, indicating that there is some evidence of a statistically significant difference.
3. This is another situation where we are comparing proportions in samples from 2 independent groups. In Stata, use the ‘Calculator / 2-sample test of proportions’ menu:

```
. prtesti 23 .913043782608695 32 .59375, level(95)
```

Two-sample test of proportion x: Number of obs = 23
y: Number of obs = 32

| Variable | Mean | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|----------|------|-----------|-------|-----|---------------------|
| x        | .9130435 | .0587534 | 15.5403 | 0.0000 | .797889 - 1.028198 |
| y        | .59375 | .0868207 | 6.8388 | 0.0000 | .4235845 - .7639155 |
| **diff** | .3192935 | .1048323 | 3.0778 | 0.0020 | .1079737 - .5306132 |

Ho: proportion(x) - proportion(y) = diff = 0

<table>
<thead>
<tr>
<th>Ha: diff &lt; 0</th>
<th>P &lt; z</th>
<th>0.9956</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ha: diff = 0</td>
<td>z = 2.623</td>
<td>0.0074</td>
</tr>
<tr>
<td>Ha: diff &gt; 0</td>
<td>z = 2.623</td>
<td>0.0074</td>
</tr>
</tbody>
</table>

(a) Point estimate is 0.913, 95% CI is (0.798, 1.028).
(b) Point estimate is 0.319, 95% CI is (0.114, 0.525).
(c) Yes, because the CI does not contain 0. As expected we also reject $H_0$ in a hypothesis test. Note that we should look at the results headed $Ha: \; diff \; = \; 0$ since this is a two-sided hypothesis test.

4. (a) 0.2023
(b) $\hat{p} = 0.55$, std error = 0.111, 95% CI (0.332, 0.768)
(c) The null value of 0.75 lies within the 95% CI (although only just). Therefore, we fail to reject the null hypothesis that the proportion is 0.75 at the 5% (two-sided) level. There is, however, weak evidence that the true proportion may be lower than 0.75.

5. (a) It is estimated that students are 83% more likely to eat lunch at Jöns Jacob than staff.
(b) There is no evidence that the difference is statistically significant. The CI for the differences in proportions contains the null value (0), the CI for the risk ratio contains the null value (1), and we fail to reject the hypothesis of independence.
(c) With these sample sizes we require a quite large difference in order to demonstrate a statistically significant result. That is, we have low power to detect differences of less than 25% percent units. If it is possible that the proportions vary on a week-to-week basis then it may be wise to take samples on several different Wednesdays.

6. (a) The binomial distribution is not strictly appropriate because the 4 pregnancies of a woman are not independent.
(b) We could sum $(O - E)^2/E$ and compare it to a $\chi^2$ distribution.
7. Let $X$ be a random variable describing the number of times HIV infection occurs in the $n$ transfusions.

(a) $\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - 0.99^{30} = 0.26$.
Therefore, the probability that the person becomes infected with HIV at some time by the blood transfusion is 0.26.

(b) $\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - 0.99^{100} = 0.63$.
Therefore, the probability that the person becomes infected with HIV at some time by the blood transfusion is 0.63.

(c) The probability of at least one positive test in a year is 0.63. Therefore it should not be a surprise if an athlete returns a single positive test.

8. (a) The complete tables are

<table>
<thead>
<tr>
<th></th>
<th>Eskdale</th>
<th>Annandale</th>
<th>Nithsdale</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>33</td>
<td>54</td>
<td>98</td>
<td>185</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>14</td>
<td>35</td>
<td>55</td>
</tr>
<tr>
<td>O</td>
<td>56</td>
<td>52</td>
<td>115</td>
<td>223</td>
</tr>
<tr>
<td>AB</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>125</td>
<td>253</td>
<td>478</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Eskdale</th>
<th>Annandale</th>
<th>Nithsdale</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>38.70</td>
<td>48.38</td>
<td>97.92</td>
<td>185.00</td>
</tr>
<tr>
<td>B</td>
<td>11.51</td>
<td>14.38</td>
<td>29.11</td>
<td>55.00</td>
</tr>
<tr>
<td>O</td>
<td>46.65</td>
<td>58.32</td>
<td>118.03</td>
<td>223.00</td>
</tr>
<tr>
<td>AB</td>
<td>3.14</td>
<td>3.92</td>
<td>7.94</td>
<td>15.00</td>
</tr>
<tr>
<td>Total</td>
<td>100.00</td>
<td>125.00</td>
<td>253.00</td>
<td>478.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Eskdale</th>
<th>Annandale</th>
<th>Nithsdale</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(O-E)^2/E$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.84</td>
<td>0.65</td>
<td>0.00</td>
<td>1.49</td>
</tr>
<tr>
<td>B</td>
<td>2.64</td>
<td>0.01</td>
<td>1.19</td>
<td>3.84</td>
</tr>
<tr>
<td>O</td>
<td>1.87</td>
<td>0.68</td>
<td>0.08</td>
<td>2.63</td>
</tr>
<tr>
<td>AB</td>
<td>1.10</td>
<td>0.30</td>
<td>1.09</td>
<td>2.49</td>
</tr>
<tr>
<td>Total</td>
<td>6.45</td>
<td>1.64</td>
<td>2.36</td>
<td>10.45</td>
</tr>
</tbody>
</table>

(b) The test statistic is $X^2 = 10.45$ which has a $\chi^2_6$ distribution under the null hypothesis (independence of blood group and region). The critical value at the 5% level is 12.6. Therefore, there is no evidence against the null hypothesis.

(c) Based on the observed data, there is no evidence of a dependency between region and blood type.
9. (a) The proportion overweight is 46.9%, 36.6%, and 24.6% for men with low, medium, and high levels of physical activity respectively.

(b) Two-sample test of proportion

| Variable | Mean   | Std. Err. | z      | P>|z|  | [95% Conf. Interval] |
|----------|--------|-----------|--------|------|---------------------|
| x        | 0.4691196 | 0.0180904 | 25.932 | 0.0000 | .4336631 -.5045761 \n| y        | 0.3663664 | 0.0186698 | 19.6235 | 0.0000 | .3297742 .4029585 \n| diff     | 0.1027532 | 0.0259966 | .0518008 | 0.1537056 |

A 95% CI for the difference in proportions is (0.052, 0.154). The difference in proportions is statistically significant, based on the formal hypothesis test (use the test labelled Ha: diff ≠ 0) or the fact that the 95% CI does not contain 0 (the null value).

(c) Two-sample test of proportion

| Variable | Mean   | Std. Err. | z      | P>|z|  | [95% Conf. Interval] |
|----------|--------|-----------|--------|------|---------------------|
| x        | 0.4691196 | 0.0180904 | 25.932 | 0.0000 | .4336631 -.5045761 \n| y        | 0.2462462 | 0.0166941 | 14.7505 | 0.0000 | .2135265 .278966 |
| diff     | 0.2228733 | 0.0246161 | .1746266 | 0.2711201 |

A 95% CI for the difference in proportions is (0.175, 0.271). The difference in proportions is statistically significant, based on the formal hypothesis test (use the test labelled Ha: diff ≠ 0) or the fact that the 95% CI does not contain 0 (the null value).

(d) `. tab weight activity [freq=n], chi`

<table>
<thead>
<tr>
<th>weight</th>
<th>High</th>
<th>Low</th>
<th>Medium</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>502</td>
<td>404</td>
<td>422</td>
<td>1328</td>
</tr>
<tr>
<td>Overweight</td>
<td>164</td>
<td>357</td>
<td>244</td>
<td>765</td>
</tr>
</tbody>
</table>

Total | 666   | 761   | 666    | 2093  |

Pearson chi²(2) = 76.0761 \ Pr = 0.000

The test statistic is 76.08 which has a \( \chi^2 \) distribution with two degrees of freedom under the null hypothesis. The critical value for a \( \chi^2 \) variate at the 5% level is 5.99. We reject \( H_0 \) and conclude that level of physical activity is associated with body weight.
(a) The proportion of deaths from heart disease was 3.5% among those who used snuff and 2.0% among those who did not use any tobacco products.

(b) The estimated relative risk (risk ratio) is 1.77. A 95% CI for the RR is (1.53–2.06). There is evidence of a statistically significant difference since the 95% CI for the RR does not contain 1.

(c) The estimated difference in proportions (risk difference) is 0.015. A 95% CI is (0.01–0.02). There is evidence of a statistically significant difference since the 95% CI does not contain 0.

(d) This does not provide conclusive evidence that using snuff increases heart disease mortality. Although there is an association, we have no evidence that it is causal. A bigger problem is that the association could be explained by confounding factors.

(a) The proportion of deaths from heart disease was 3.5% among those who used snuff and 2.0% among those who did not use any tobacco products.

(b) The estimated relative risk (risk ratio) is 1.77. A 95% CI for the RR is (1.53–2.06). There is evidence of a statistically significant difference since the 95% CI for the RR does not contain 1.

(c) The estimated difference in proportions (risk difference) is 0.015. A 95% CI is (0.01–0.02). There is evidence of a statistically significant difference since the 95% CI does not contain 0.

(d) This does not provide conclusive evidence that using snuff increases heart disease mortality. Although there is an association, we have no evidence that it is causal. A bigger problem is that the association could be explained by confounding factors.

(a) The proportion of deaths from heart disease was 3.5% among those who used snuff and 2.0% among those who did not use any tobacco products.

(b) The estimated relative risk (risk ratio) is 1.77. A 95% CI for the RR is (1.53–2.06). There is evidence of a statistically significant difference since the 95% CI for the RR does not contain 1.

(c) The estimated difference in proportions (risk difference) is 0.015. A 95% CI is (0.01–0.02). There is evidence of a statistically significant difference since the 95% CI does not contain 0.

(d) This does not provide conclusive evidence that using snuff increases heart disease mortality. Although there is an association, we have no evidence that it is causal. A bigger problem is that the association could be explained by confounding factors.
(b) . csi 80 129 23 55, level(95)

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>80</td>
<td>129</td>
<td>209</td>
</tr>
<tr>
<td>Noncases</td>
<td>23</td>
<td>55</td>
<td>78</td>
</tr>
<tr>
<td>Total</td>
<td>103</td>
<td>184</td>
<td>287</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk</th>
<th>.776699</th>
<th>.701087</th>
<th>.728223</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point estimate</td>
<td>[95% Conf. Interval]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk difference</td>
<td>.0756121</td>
<td>-.0285208 .1797449</td>
<td></td>
</tr>
<tr>
<td>Risk ratio</td>
<td>1.10785</td>
<td>.9630366 1.274439</td>
<td></td>
</tr>
</tbody>
</table>

chi²(1) = 1.91  Pr>chi² = 0.1672  

i. 0.075 (-0.0285, 0.1797)
This is identical to the CI shown in the Stata output for part (a) except the signs are reversed.

ii. 1.108 (0.963, 1.274)

iii. The new drug is slightly superior, although the difference is not statistically significant (based on the fact that the confidence intervals contain the null value).

(c) We will the null hypothesis, \( H_0 : p_1 - p_0 = 0 \) vs. the two-sided alternative \( H_1 : p_1 - p_0 \neq 0 \) where \( p_0 \) is the proportion with 36-hour clearance for the existing treatment and \( p_1 \) is the corresponding proportion for the new treatment.

When formally testing a hypothesis, the variance should be estimated under the null hypothesis. The common value of the proportion under \( H_0 \), \( \hat{p}_c \) is estimated as

\[
\hat{p}_c = \frac{r_1 + r_0}{n_1 + n_0} = \frac{129 + 80}{184 + 103} = 0.728
\]

and the variance of \( p_1 - p_0 \) under \( H_0 \) is estimated by

\[
\text{var}_{H_0}(p_1 - p_0) = \frac{\hat{p}_c(1 - \hat{p}_c)}{n_1} + \frac{\hat{p}_c(1 - \hat{p}_c)}{n_0} = 0.002997 \tag{2}
\]

The standard error is \( \hat{SE}_{H_0}(p_1 - p_0) = \sqrt{0.002997} = 0.0547 \). Note that in the output shown for part (a) above, this is given in the bottom row of the table (Std. Err. under \( H_0 \)). The test statistic is then

\[
\frac{\hat{p}_1 - \hat{p}_0}{\hat{SE}_{H_0}(p_1 - p_0)} = \frac{0.0756}{0.0547} = 1.381
\]

which will have a standard normal distribution under \( H_0 \). \( Pr(|Z| > 1.381) = 0.1672 \) so we fail to reject \( H_0 \). We conclude that there is no evidence of a statistically significant difference in the efficacy of the two treatments.

Note that the results of this hypothesis test are shown in the Stata output above for part (a). Stata tested the difference \( p_0 - p_1 \) so the test statistic was negative; the conclusion however is identical.

References