Contents

1 S-language

Problem set I

1. Write and test a command that creates absolute values; other than `abs()`.

2. Write and test a command that evaluates the standard error of the mean associated with a vector of n values labeled x without using `var(x)` where
$$standard\ error = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n(n-1)}}.$$

3. Write and test a set of commands that calculates the square root of T using the relationship $x_{i+1} = (x_i + T/x_i)/2$.

4. Compare the S-function `pnorm()` to the approximation $P(x)$ where
$$P(x) = \frac{1 + \sqrt{1 - e^{-2x^2/\pi}}}{2}$$
when x is between 0 and ∞.

Find the maximum difference between these two functions and the value at which the maximum difference occurs.

5. Pascal's triangle is

row 1	1
row 2	1 1
row 3	1 2 1
row 4	1 3 3 1
row 5	1 4 6 4 1
	...

Write and test an S-function thta generates the kth row. Verify thta the rows sum to 2^k.

6. The estimated standard deviation is

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}.$$

A correction for the bias of this estimate is

$$\alpha_n = \sqrt{\frac{2}{n}}\frac{\Gamma(n/2)}{\Gamma([n-1]/2)}$$

where $\Gamma(x)$ represents a gamma function evaluated at x. Calculate α_n for $n = 2, 3, 4, \cdots, 50$.

Show that

$$\alpha_n = \frac{3.5n - 3.62}{3.5n - 1}$$

is a better approximation.

7. Write and test S-commands that create the following two patterns of numbers:

123451234512345123451234512345
111112222233333444445555

Write and test S-commands thta will generate this pattern in general; that is, for any chosen integer.

8. Create an S-vector with 1000 values set to 1. Then add one to every second value, then add one to every third value, then add one to every fourth value, and so on 1000 times. Which values are odd and which are even? Justify the observed pattern.

9. Write and test an S-function that accepts two vectors of observations x and y as input and returns the Spearman rank correlation (the correlation coefficient calucalted using ranks of the observations).

   Verify the S-code using `cor.test(x, y, method=spearman)`.

10. A perfect shuffle of a deck of 52 playing cards occurs when the deck is split perfectly into halves (26/26) and the cards are exactly alternated. If the top card remains on the top after each perfect shuffle, how many perfect outside shuffles are necessary to restore the original order? If the top card becomes the second card on each shuffle, the shuffle is called a perfect inside shuffle. If the deck is ordered before the first shuffle, how many perfect inside shuffles are necessary to restore the original order?

11. Generate a vector (denotes `ybar`) containing $k = 20$ mean values each composed of $n = 50$ random observations from a normal distribution with mean = 2 and standard deviation = 2.

    Generate `ybar` using a "`for`-loop."

    Generate `ybar` without using a "`for`-loop."

    Increase k and n and note the difference in execution times.

12. Evaluate
$$f(x, y) = \frac{\sin(x)}{\sqrt{1 + \cos^2(y)}}$$
    over the range $-2\pi < x < 2\pi$ and $-2\pi < y < 2\pi$.

    Construct and test S-code using and not using a "`for`-loop." Note the difference in execution times.

13. Construct a three-column array where column 1 = `rnorm(100,2,2)`, column 2 = `rnorm(100,4,4)`, and column 3 = `rnorm(100,0,1)`. Use S-code to produce an array so that each column has exactly mean = 0 and exactly variance = 1.

    Use the command `scale(cbind(x1,x2,x3),center=T)` which does the same thing to verify your results.

    Verify that `cor(cbind(x1,x2,x3))` = `var(scale(cbind(x1,x2,x3)))`.

14. An approximation for $n!$ is
$$n! \approx \sqrt{2n\pi}(n/e)^n.$$

Write an S-code program to show that the difference between this approximation and $n!$ increases with increasing n but the relative error (i.e., $(absolute\ difference)/n$) decreases.

15. For the 100 observations in the following table, write an S-command that produces a vector labeles `test` with 100 observations with the same distribution of values (e.g., 1, 1, $\cdots$, 5, 5, 5, 5):

	1	2	3	4	5
count	20	35	25	10	10

To check: `table(test)` will reproduce the above table.

16. In a random matching of two equivalent decks of k cards the probability P_m of exactly m matches is

$$P_m = \frac{1}{m!}\left\{1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots \pm \frac{1}{(n-m)!}\right\}$$

where $m = 0, 1, 2, 3, \cdots, n-1$.

Write an S-code function to calculate P_m for a given value of $m < n$.

Show that for $m > 9$, $P_m \approx \frac{1}{m!}e^{-1}$.

2 Descriptive techniques

Problem set II

1. Plot the function $f(x) = x\log(x)$ for $0 < x < 1$. Use the `arrows()` S-command to point out the minimum of $f(x)$ and the value of x at which the minimum occurs (see `?arrows`).

2. Plot four distributions using the `density()` function on the same axes based on $n = 200$ observations sampled at random from each of four normal distributions with the same variance ($\sigma = 2$) but with mean values $= \mu_i = \{2, 4, 6, 8\}$.

   Also plot four distributions using the `density()` function on the same axes based on $n = 200$ observations sampled at random from each of four normal distributions with the same mean ($\mu = 2$) but with standard deviations $= \sigma_i = \{2, 2.5, 3, 3.5\}$.

   Repeat the two plots with n increased to $n = 2000$.

3. Suppose a needle with length l is tossed on a grid with horizontal lines spaced d units apart. Plot the 11 horizontal lines between 0 and 1, $d = 0.1$ units apart. Then, "drop" $n = 50$ random needles on the grid with $l = 0.05$.

   This plot is a display of Buffon's classic problem where he calculated the probability that a needle tossed on such a grid intersects a horizontal line (probability of an intersection $= l/[\pi d]$).

4. Plot the bivariate normal density with `persp()` and `contour()` where $\mu_x = 0$, $\sigma_x = 1$, $\mu_y = 2$, $\sigma_y = 3$ for $\rho_{xy} = \{0.1, 0.5, 0.9\}$ (see chapter for the mathematical expression of the density function).

5. The S-commands `x <- runif(n)` and `y <- runif(n)` generate a set of n random points on the unit square. Divide the x-axis into 10 intervals and calculate the median for values x and y in each interval. Generate three plots with $n = 100, 1000$ and 10,000 random points. Connect the median values to form a median trace (showing, if any, trends in the generated data). Plot the points and the smoothed median traces.

6. Generate a random sample from a normal distribution of $n = 10,000$ values. Test the boxplot definition, of an "outlier" which shoud produce about 0.7% or 70 generated values as "outliers".

7. The first smoothed observation is a combination of the next two smoothed observations. Justify[1] that $new - y_1 = 2y_2 - y_3$.

8. Plot the function
$$f(x, y) = \frac{\sin(x)}{\sqrt{1 + \cos^2(y)}}$$
   over the range $-2\pi < x < 2\pi$ and $-2\pi < y < 2\pi$ using `persp()` and `contour()`.

9. Consider the data describing protastic cancer by age and year of death (rates/100,000):

[1] Solve theoretically and not with a computer program.

age	1930	1935	1940	1945	1950	1955
45–49	3.5	4.9	6.6	7.2	4.7	4.7
50–54	5.4	11.4	16.0	15.7	23.3	16.3
55–59	18.8	22.4	32.1	41.4	39.0	44.4
60–64	24.0	45.9	50.2	60.4	77.7	84.2
65–69	35.1	60.2	72.3	74.1	74.1	141.1
70–74	60.7	66.5	90.1	126.0	148.0	168.5
75–79	47.5	90.6	151.4	130.0	219.2	234.4
80+	56.7	124.5	152.1	155.6	299.1	328.6

Separate each rate into an additive and a residual component using `sweep()`.

Assume the effects on the disease rate from year and age are additive. Calculate and inspect the residual pattern (e.g., plot the residuals).

Plot the additive "data" (each year = a line) and the observed rates on the same set of axes.

Repeat the same analysis, assuming the year and age influences have multiplicative influences on the rate of prostatic cancer.

3 Simulation: Random Values

Problem set III

1. Use the Kolmogorov test to assess the "randomness" of the values generated by $a = 2^{10} + 1$, $c = 0$, and $m = 2^{20}$ (`ran1` in the chapter).

2. Chuck-a-luck is a game where one bets on the numbers 1, 2, 3, 4, 5, and 6. Three dice are rolled and if a player's number appears 1, 2, or 3 times, the pay-off is respectively 1 or 2 or 3 times the origina bet (plus the player's bet). Simulate this game and estimate the player's expected gain or loss (ans: 7.9% loss).

3. If a chord is selected at random from a circle with a set radius, what proportion of lengths will be smaller than the radius of the circle? Write and test a simulation program to estimate the answer[2] to this question (ans: 1/3).

[2] *Note from the author*: Answer is not unique: One should also probably found 1/4 or 1/5 depending on how one defines what is meant by "drawing a chord at random".

4. Which event is more likely: (i) $k = 1$ or more sixes in 6 tosses of a die or (ii) $k = 2$ or more sixes in 12 tosses of a die or (iii) $k = 3$ or more sixes in 18 tosses of a die?

   Write and test an S-program to simulate dice and answer the question: what is the probability of k or more sixes in $6k$ tosses of a die for selected values of k (ans: $k = 1, 2, 3$, and 4, then $p = 0.665, 0.619, 0.597$, and 0.584). Note: this problem is attributed to a question posed by Isaac Newton.

5. Robust linear regression: to achieve "robust" estimates of the intercept and slope of a straight line, the data are divided into three groups based on the ordered values of the independent variable x. Each group contains approximately one-third of the data (e.g., if the total number of observations is $n = 3k$, then the leftmost group has k members, the middle group has k members, and the rightmost group has k members—if the number of observations is not divisible evenly by three, then the observations are allocated as closely as possible to the ideal of $n/3$). Using these "thirds," a representative point is constructed based on the median of the x-values and the median of the y-values calculated separately from within each group. The pairs of median values

   $$(x_L, y_L), (x_M, y_M), \text{ and } (x_R, y_R)$$

   become the representative values of the left, middle, and right groups respectively.

   Estimates of the slope (b^*) and intercept (a^*) are then

   $$\hat{b}^* = \frac{y_R - y_L}{x_R - x_L} \text{ and } \hat{a}^* = \frac{1}{3}(y_L + y_M + y_R) - \hat{b}^* \frac{1}{3}(y_L + y_M + y_R).$$

   Since the estimate of the slope depends only on the median values from the left and right groups, it is almost certainly unaffected by extreme values, called a robust estimate.

   Simulate a set of "data" that conforms to the assumptions of simple linear regression—the dependent variable (y) is linearly related to the independent variable (x) and is sampled independently from a normal distribution with constant variance. Specifically,

   $$y_j = a + bx_j + e_j$$

   where e_j is one of series of independent and normally distributed values. Use 100 "data" sets to simulate the distribution of "robust"

estimator of the slope denoted $\hat{b}^*$. Also, compute 100 estimates of the slope b with `lsfit()`, denoted $\hat{b}$.

Compare the variances of the two estimated values $\hat{b}^*$ and $\hat{b}$.

6. For the positive triangular distribution, write and test four S-functions that produce the cumulative probabilities, quantiles, heights, and a random sample (e.g., `pptri()`, `qptri()`, `dptri()`, and `rptri()` functions).

   Show both theoretically and graphically that $x = \max(u_1, u_2)$ has a positive triangular distribution when u_1 and u_2 are two independent values from a uniform distribution $(0, 1)$.

7. Derive[3] the cumulative distribution and density function for a negative triangular distribution on the interval 0 to 1.

8. For the negative triangular distribution, write and test four S-functions that produce the cumulative probabilities, quantiles, heights, and a random sample (e.g., `pntri()`, `qntri()`, `dntri()`, and `rntri()` functions).

   Show both theoretically and graphically that $x = \min(u_1, u_2)$ has a positive triangular distribution when u_1 and u_2 are two independent values from a uniform distribution $(0, 1)$.

9. A left truncated standard normal distribution is given by the expression

$$f(z) = \frac{\frac{1}{\sqrt{2\pi}} e^{-0.5z^2}}{1 - P} \qquad \text{for } z > z_0$$

   where $P = P(Z < z_0)$.

   Simulate a random sample of $n = 1000$ values from this distribution when $z_0 = -1$ using two different methods.

10. Using S-code and Butler's method create an S-function that produces n random values from a chi-square distribution with *df* degrees of freedom. Compare the results to using `rchisq(n,df)`.

11. Demonstrate that for $n > 100$ and $p < 0.05$, the binomial and Poisson distributions produce similar probabilities.

[3] Solve theoretically and not with a computer program.

12. Demonstrate using S-functions such as `qqnorm()` that the Box-Muller transformation gives two independent standard normal variates from two independent and uniformally distributed random variables.

13. Plot a square with a circle inscribed within the boundaries. Generate random pairs (x, y) and determine whether these values are in the circle or not. Use the result to estimate $\pi = 3.1416$.

14. Demonstrate with an S-simulation program that the expected mean values resulting from sampling the same population with and without replacement are equal. Which sampling has the smallest variance?

15. Create and test an S-function that produces n random values from a Poisson distribution using the inverse transformation method for given value of λ. Compare the results to the values generated with `rpois()`.

16. The test-statistic $X^2 = (n-1)S_x^2/\bar{x}$ has an approximate chi-square distribution with $n-1$ degrees of freedom when the n values x_i are sampled from a Poisson distribution. Use a simulation program to verify that X^2 has an approximate chi-square distribution for $n = 100$ and $\lambda = 1.0$. Use this fact to test formally the fit of the Rutherford-Geiger data (Chapter 3) to a Poisson distribution.

4 General Linear Models

Problem set IV

1. A measure of influence on the estimated value $\hat{y}_i$ associated with the ith independent variable x_i can be defined as $\hat{b} - \hat{b}_{(i)}$ where $\hat{b}$ is estimated from all n observations and $\hat{b}_{(i)}$ is estimated from the same data set but with the ith observation removed. Compute all such values for the diastolic blood pressure data (Table 4.1) and determine the five most influential points. Locate the five points on a plot of the data.

2. Show[4] that if the coefficients a_i maximize the multivariate distance M^2, then the coefficients $ba_i + c$ also maximize M^2 where b and c are constants.

3. The goodness-of-fit S-code for the logistic model in the chapter requires the number of observations to be evenly divisible by 10. Write

[4]Solve theoretically and not with a computer program.

and test an S-program for the same goodness-of-fit test but make no assumptions about the total number of observations used in the logistic regression analysis.

4. The table showing case/control status and coeffe consumption (Table 4.7) is a summary of $n = 1010$ individuals where the frequencies in the table are the counts of observations with the identical values of the dependent and independent variables. For example, there are nine records (one for each person) where the outcome is a case who reports no coffee consumption ($x_1 = 0$) and is a male participant ($x_2 = 1$)—first cell in the summary table.

   Using Table 4.7 reconstruct the data so that tehre are 1010 individuals records where each record contains 0 or 1 for case/control status as well as the corresponding values of the independent variables (x_1 and x_2).

   Use these $n = 1010$ records and `glm()` and conduct a logistic regression analysis showing that the results are identical to the ones in the chapter where the analysis is performed directly on the tabular data using the cell frequencies as weights.

5. Consider the following data where birth weight and maternal age are recorded for three groups based on smoking exposure status:

non-smoker			quitters			smokers		
	bwt	age		bwt	age		bwt	age
1	9.1	35	1	7.2	32	1	6.7	24
2	8.9	29	2	7.7	30	2	6.5	24
3	8.5	34	3	6.8	26	3	7.2	28
4	7.4	32	4	7.0	33	4	6.5	26
5	7.5	28	5	7.4	28	5	6.5	26
6	7.3	28	6	6.2	29	6	7.1	26

   Using S-tools conduct a separate simple linear regression analysis for each smoking exposure group.

   Use the same data and the model $y = a + b_1x + b_2g_1 + b_3g_2 + b_4g_1x + b_5g_2x$ to conduct a linear regression analysis using all 18 observations simultaneously where y = birth weight (dependent variable) and x = age (independent variable). The design variable g is defined as $g_1 = g_2 = 0$ for non-smokers, $g_1 = 0$, $g_2 = 1$ for quitters and $g_1 = 1$, $g_2 = 0$ for smokers.

Demonstrate these two approaches are identical.

Formally, test the influence of the three-level smoking exposure categorical variable on birth weight.

6. The following data are deaths from lung cancer and person-years at risk, classified by age and exposure to radiation for workers at the Oak Ridge National Laboratory. data incl. in `oakRNL.txt`

age	mSv	deaths	p-years	age	mSv	deaths	p-years	age	mSv	deaths	p-years
1	1	0	29901	2	3	2	2423	3	5	0	476
2	1	1	6251	3	3	1	2281	4	5	0	387
3	1	4	5251	4	3	2	1918	5	5	0	225
4	1	3	4126	5	3	0	1322	6	5	1	164
5	1	3	2778	6	3	2	723	7	5	0	150
6	1	1	1607	7	3	3	538	1	6	0	779
7	1	3	1358	1	4	0	2341	2	6	0	296
1	2	1	71382	2	4	0	972	3	6	0	282
2	2	5	16705	3	4	1	958	4	6	1	251
3	2	4	13752	4	4	1	816	5	6	0	193
4	2	10	10439	5	4	0	578	6	6	0	125
5	2	11	7131	6	4	2	375	7	6	0	69
6	2	16	4133	7	4	3	303	1	7	0	520
7	2	11	3814	1	5	0	1363	2	7	0	188
1	3	0	6523	2	5	0	478	3	7	0	217
4	7	0	184	5	7	1	109	6	7	0	60
7	7	1	23	1	8	0	2104	2	8	0	1027
3	8	1	1029	4	8	3	827	5	8	1	555
6	8	2	297	7	8	2	153				

Recode age categories 1, 2, 3, 4, 5, 6, 7 into ages 45, 47.5, 52.5, 57.5, 62.5, 67.5, and 70 years. Recode exposure categories 1, 2, 3, 4, 5, 6, 7, 8 into exposures 0, 15, 30, 50, 70, 90, 110, and 120 mSv (milliseiverts).

Evaluate the exposure response in these data using a Poisson regression approach.

When the open-ended (last) interval coded at 120 mSv is recoded to 160 mSv, assess the impact on the exposure/risk relationship using a Poisson regression analysis.

Plot the impact on the dose-response relationship varying the definitions of the coded value for the last exposure group (e.g., 120, 130, 140, $\cdots$, 220).

7. Generate a sample of $n = 200$ random observations that are described by the logistic model

$$p_i = \frac{1}{1 + e^{-(a+bx_i)}}$$

where a and b are such that $or = 3.0$.

Using `glm()` and the simulated data, estimate the odds ratio.

5 Estimation

Problem set V

1. The state fish and game service requires salmon catches to be reported from any boat catching one or more fish. The boats that do not catch fish, do not report. The data are, therefore, truncated since the number of boats failing to catch fish are not recorded. An example of such data is

number of fish	1	2	3	4	5	6
boats	34	25	12	5	1	0

Assume that the number of salmon caught per boat are described by a Poisson distribution where

$$f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}/x_i!}{1 - e^{-\lambda}} \qquad i = 1, 2, 3, \cdots.$$

The symbol x_i represents the number of fish caught per boat. Use the S-function `ms()` to find the maximum likelihood estimate of λ.

Using the scoring technique to estimate parameters, find the maximum likelihood estimate of λ and an estimate of its variance.

Use the `uniroot()` S-function to find the maximum likelihood estimate of λ.

2. Consider the situation where the number of observations below c_0 is known but the actual values of the observations are not known (i.e., the distribution is left censored at c_0). Also the number of observations above c_0' is known but the actual values of the observations are not known (i.e., the distribution is also right censored at c_0'). Further assume that the sampled distribution is normally distributed with mean μ and standard deviation σ. Write and test an S-code program to

estimate μ and σ for this doubly censored normal distribution from n observations where n_0 are left censored, n_0' are right censored and $n - n_0 - n_0'$ are measured values.

3. Consider the following model constructed to estimate the proportion of dizygotic twins where

$$\begin{aligned}\text{probability of a like-sex twin pair} &= \text{P(like-sex twin)} = M + D/2 \text{ and} \\ \text{probability of an unlike-sex twin pairs} &= \text{P(unlike-sex twin)} = D/2\end{aligned}$$

where $D = 1 - M$ is the proportion of dizygotic (fraternal twins). A specific number of pairs of like-sex twins = 67 and unlike-sex twins = 42 are observed.

Find[5] the maximum likelihood estimate of D and its variance in closed form.

Use an S-code program to estimate D and its variance using scoring techniques.

Use a bootstrap procedure to estimate D and its variance.

4. A measure of skewness is

$$\hat{M} = \frac{\sum(x_i - \bar{x})^3}{n}$$

where $M = 0$ identifies a symmetric distribution using a sample of n observations. For the data $\{2, 5, 8, 2, 5, 9, 1, 4, 30\}$ estimate M and its standard error using a boostrap procedure (i.e., find $\hat{M}_{[\cdot]}$ and $\hat{se}(\hat{M}_{[\cdot]})$ estimates).

Find the same estimates using the jacknife procedure.

Estimate M and its standard error using $n = 100$ values sampled from a standard normal distribution using both estimation techniques.

5. Consider the following 15 observations:

$$0.28, -1.21, 0.60, 0.14, 0.51, 0.19, -0.27, 0.45, 0.29, 0.40, 0.04, 0.60, 1.11, 0.90.$$

Use a bootstrap strategy to assess the likelihood this sample arose from a population with a mean value of 0 ($\mu_0 = 0$?).

6. a sample of data yields the following 2×2 table:

[5] Solve theoretically and not with a computer program.

	disease	no disease	total
exposed	$a = 200$	$b = 120$	320
unexposed	$c = 80$	$d = 120$	200
total	280	240	520

The odds ratio measure of association between disease and exposure is estimated by $\hat{or} = ad/bc$. Construct and test an S-program to find the bootstrap estimate of the bias associated with this estimate. The logarithm of the estimated odds ratio ($\log[\hat{or}]$) is another measure of association. Use an S-program to estimate of the bias associated with this measure of association.

7. For the data

$$x = 5, 10, 15, 20, 25, 30, 35, 40, 45 \quad \text{and}$$
$$y = 0.08, 0.12, 0.22, 0.21, 0.27, 0.56, 0.70, 0.71, 0.84$$

use the model $y_i = [1 + e^{-(a+bx_i)}]^{-1}$ and `nls()`-function to estimate a and b. Hint: use initial value $a_0 = -3$ and $b_0 = 0.1$.

Apply a linearizing transformation to y and again estimate the parameters a and b usng ordinary least squares estimation.

8. Use a boostrap procedure to estimate θ, its standard error, and the bias for

$$\hat{\theta} = \frac{1}{n}\sum(x_i - \bar{x})^2$$

where $n = 15$ and $x = \{12, 13, 23, 31, 41, 22, 44, 37, 14, 18, 24, 36, 51, 11, 32\}$.

Use a boostrap procedure to estimate θ, its standard error, and the bias from a sample of $n = 100$ random values selected from a standard distribution.

9. For the two sets of $n = 15$ observations

$$x = \{1, 3, 2, 6, 8, 3, 8, 3, 9, 10, 15, 12, 18, 5, 2\}$$

and

$$y = \{10, 14, 15, 22, 28, 21, 14, 15, 12, 18, 33, 37, 33, 11, 12\}$$

write and test and S-program to assess the conjecture that x and y are samples of unrelated variables ($correlation = 0$) using a randomization strategy.

10. A simple Mendelian genetic system is represented by the following model:

AA-homozygote frequency:	p^2
Aa-heterozygote frequency:	$2pq$
aa-homozygote frequency:	q^2

where p represents the frequency of the A-gene and $q = 1-p$ represents the frequency of the a-gene.

Find[6] the maximum likelihood estimate of p and its variance when n_1, n_2, and n_3 represent the respective observed counts of AA, Aa, and aa genotypes where

$$\log(L) = n_1 \log(p^2) + n_2 \log(2pq) + n_3 \log(q^2).$$

If $n_1 = 250$, $n_2 = 441$, and $n_3 = 314$, find the maximum likelihood estimate of p using S-tools to verify the "closed-form" estimate and variance.

If the laboratory determination of the homozygotes is subject to misclassification, the log-likelihood function is then

$$\log(L) = n_1 \log(p^2 + e) + n_2 \log(2pq) + n_3 \log(q^2 - e)$$

where e represents the proportion misclassified homozygotic types.

If $n_1 = 250$, $n_2 = 441$, and $n_3 = 314$, find the maximum likelihood estimate of p and e using S-tools. Also estimate the variance/covariance array for the estimates of p and e.

11. Generate two sets of $n = 100$ random variables where x and y have independent standard normal distributions.

Use bootstrap tools to estimate the correlation between x and $2x + y$ (i.e., *correlation*(x,$2x + y$)). Also estimate the variance and bias associated with this estimate. Plot the histogram and the estimated density function of the estimated correlation coefficient.

6 Analysis of Tabular Data

Problem set VI

[6] Solve theoretically and not with a computer program.

1. For the data in the following table, work out the estimated values of $\hat{a}$, $\hat{b}_1$, $\hat{b}_2$, and $\hat{b}_3$ for the saturated model algebraically:

a_i	b_i	f_{ij}	F_i	data
0	0	f_{00}	F_1	23
1	0	f_{10}	F_2	12
0	1	f_{01}	F_3	45
1	1	f_{11}	F_4	122

   Verify the results using an S-program.

2. If n and p are parameters of a binomial distribution, then

$$z_0 = \frac{x - np}{\sqrt{np(1-p)}} \quad \text{and} \quad z_1 \frac{x - np \pm 0.5}{\sqrt{np(1-p)}}$$

   provide approximate binomial probabilities (i.e., $pnorm \approx pbinom$). Compare the maximum difference between the exact binomial and normal approximated probabilities for $n = 10$, $p = 0.5$; $n = 20$, $p = 0.2$; $n = 50$, $p = 0.1$; $n = 100$, $p = 0.05$.

3. Consider the 2×4 table:

	$X=0$	$X=1$	$X=2$	$X=3$	total
$Y=0$	1	7	15	40	63
$Y=1$	4	19	34	42	99
total	5	26	49	82	162

   Compute S_{xx}, S_{yy} and S_{xy}.

   Calculate the chi-square statistics reflecting the total, linear, and non-linear influences. Demonstrate that the correlation between X and Y based on $n = 162$ pairs (X, Y), called the biserial correlation coefficient, is directly related to the chi-square statistic reflecting the linear association or

$$r_{XY} = \sqrt{\frac{linear\ chi\ square\ statistic}{n-1}}.$$

4. Show[7] algebraically that $or_1/or_2 = e^{\hat{b}_7}$ when a saturated loglinear model is applied to a $2 \times 2 \times 2$ table where or_1 is the odds ratio calculated from one 2 by 2 subtable and or_2 is the odds ratio calculated from the other subtable.

[7] Solve theoretically and not with a computer program.

5. An estimate of the variance of Yule's measure of association Y is

$$variance(Y) = \frac{1}{4}(1 - Y^2)^2 \left[\frac{1}{F_1} + \frac{1}{F_2} + \frac{1}{F_3} + \frac{1}{F_4}\right].$$

Create an S-function to calculate Y, the variance of Y and to construct an approximate 95% confidence interval for the expected value of Y. Show that Yule's Y is equivalent to the odds ratio measure $\hat{or}$ in a 2 by 2 table.

6. Demonstrate that the expected values calculated under the hypothesis of statistical independence (usual chi-square expected values—Chapter 1 example) are essentially the same as the estimates of the cell frequencies based on a loglinear additive model using the following data:

	b_1	b_2	b_3	b_4
a_1	12	8	22	4
a_2	5	3	11	2
a_3	26	17	44	10
a_4	53	38	82	18
a_5	108	75	167	44

7. Consider the following $2 \times 2 \times 2$ table:

x_1	x_2	x_3	F_i	count
1	1	0	F_1	$\hat{F}_1 = 11 + e$
1	0	0	F_2	$\hat{F}_2 = 8 - e$
0	1	0	F_3	$\hat{F}_3 = 12 - e$
0	0	0	F_4	$\hat{F}_4 = 37 + e$
1	1	1	F_5	$\hat{F}_5 = 22 - e$
1	0	1	F_6	$\hat{F}_6 = 5 + e$
0	1	1	F_7	$\hat{F}_7 = 3 + e$
0	0	1	F_8	$\hat{F}_8 = 7 - e$

Find the value e such that no interaction exists (i.e., $\hat{or}_1 = \hat{or}_2$). Display the "data" ($\hat{F}_i$) and show that the odds ratios measuring the association between any two variables at both levels of the third variable are the same. Use a loglinear model applied to the created "data" ($\hat{F}_i = F_i \pm e$ value) to show that the x1*x2*x3-term is zero (exact homogeneity).

7 Analysis of Variance and Some Othe S-Functions

Problem set VI

1. Use the cholesterol and behavior type data (Table 1.5) to show that a two-sample t-test, a simple linear regression analysis, and an analysis of variance give identical results. That is, all three approaches produce the same significance probability for comparing levels of cholesterol between behavior type A and B.

2. Use the `glm()` S-function to reproduce the results found in the chapter using the `aov()` S-command for the two-way table of lead level determination data (Table 7.2).

3. Construct and test an S-program to execute a Kruskal-Wallis rank test for independent samples. Compare the results with the S-function `kruskal.test()` for a set of simulated data with no identical (tied) values.
   Construct and test an S-program to execute a Wilcoxon signed rank test for a set of matched pairs data. Compare the results with the S-function `wilcox.test()` for a set of simulated data with no identical (tied) values.

4. Conduct a principal component analysis where $x_1 = \{1, 2, 3, 4, 5, 6, 7\}$ and $x_2 = \{7, 6, 5, 4, 3, 2, 1\}$. Calculate the variance of x_1, the variance of x_2, the variance of the first principal component, and the variance of the second principal component. Why are these variances the same?

5. Using the turtle data (Table 4.4) calculate the first two principal components. To identify differences by gender (clustering), plot the 48 values of each principal component (one against the other).

6. Use the following data to show that a canonical correlation and the multiple correlation coefficient are the same when one group consists of a single variable (y) and the other group (x) has $k = 2$ variables (i.e., compare results from `cancor()` with `lm()` S-functions).

	y	x_1	x_2
1.	4.8	1	0.2
2.	14.1	2	0.6
3.	10.7	3	0.2
4.	18.3	4	0.8
5.	12.7	5	0.2
6.	17.2	6	1.9
7.	16.0	7	1.3
8.	22.0	8	1.6
9.	22.0	9	1.7
10.	23.6	10	1.1

7. Using the weight gain matched pair data (Table 7.3), assess the association between maternal weight gain and low birth weight ignoring the paired structure (as if the two infants represent samples from separate and unrelated populations). In other words, does the paired pattern of the data collection improve the efficiency of the analysis or not?

8. Write and test an S-program to execute a ranomization test for matched pair data. Conduct a matched pair randomization test using the paired data in Table 7.3. Compare the results to the tests used in the chapter to analyze the association between birth weight and maternal weight gain matched for pregnancy weight.

9. A total of $N = 11$ matched sets of data are collected (one case and two controls). For each infant with a birth defect (case) born in a rural area in France, two infants (control) were selected who were born at essentially the same time, in the same village and were the same sex. The matched data consist of the distances to electromagnetic radiation exposure (risk factor—measured in meters).

malformation	1150	100	2000	350	400	2700	1200	1800	10	250	350
control 1	300	100	2150	1350	800	1250	450	400	900	1950	1050
control 2	750	650	4050	450	700	2850	50	2300	150	300	1000

Using these 1:2 matched sets assess the association between electromagnetic radiation and birth defects (does the distances among cases differ from distances among controls?).
Ignoring the matched data collection design, again evaluate the association between electromagnetic radiation and birth defects. Does the

matched pattern of the data collection improve the efficiency of the analysis?

10. Simulate a data set of $n = 100$ pairs of matched observations. Demonstrate that the results using the binomial test (without a correction factor) and the Friedman test are identical.
    Simulate a data set of $n_1 = 100$ and $n_2 = 100$ observations from two independent populations. Demonstrate that the results using the Wilcoxon two-sample test (`pairs=F`) and the Kruskal-Wallis tests are identical.

8 Rates, Life Tables, and Survival

Problem set VIII

1. Show[8] the equivalence of the three expressions:

$$\begin{aligned} \text{person-years} &= \delta_x P_{x+\delta_x} + \tfrac{1}{2}\delta_x D_x, \\ \text{person-years} &= \delta_x P_x - \tfrac{1}{2}\delta_x D_x, \\ \text{person-years} &= \tfrac{1}{2}\delta_x (P_{x+\delta_x} + P_x). \end{aligned}$$

   Show[9] that for an exponential survival model

$$S(T > t_2 \mid T > t_1) = e^{-\lambda(t_2 - t1)}.$$

   Show[10] that when $\lambda_1(t)/\lambda_2(t) = c$, then $S_1(t) = [S_2(t)]^c$.

2. If the survival times from one group are $\{7.5, 12, 18, 33^+, 55.5, 61.5\}$ and for another group are $\{34.5, 60, 64.5, 76.5^+, 93\}$, show that the log-rank test (i.e., `surv.diff()` function) gives essentially the same results as the proportional hazards model (i.e., `coxreg()` function). The "+" indicates a censored survival time.

3. If d observations are complete (not censored) in a sample of n distinct survival times from exponentially distributed data, the likelihood function is

$$L = \prod_i \lambda e^{-\lambda t_i} \times \prod_j \lambda e^{-\lambda t'_j} \quad i = 1, 2, 3, \ldots, d \text{ and } j = 1, 2, 3, \ldots, n - d$$

[8] Solve theoretically and not with a computer program.
[9] Solve theoretically and not with a computer program.
[10] Solve theoretically and not with a computer program.

where t_i represents complete observations and t'_j represents censored observations. Find the maximum likelihood estimate[11] of λ. Verify this estimator using the survival data from problem 2 and the S-function `ms()`.

4. If the survival function is $S(t) = 1 - t/b$ where $0 \leq t \leq b =$ constant, find the hazard funciton $\lambda(t)$ and the cumulative hazard function $H(t)$. Plot these three curves on a single page. Derive an expression for the average rate ($R_t =$ deaths/person-years). Show[12] that $R_t \approx \lambda(t)$ for small time intervals.

5. Consider the $n = 11$ complete survival time $t = \{1, 4, 6, 8, 2, 12, 24, 23, 25, 27, 31\}$. Use an S-program to demonstrate that the Kaplan-Meier estimated mean survival time is the same as the usual mean value $\bar{t} = \sum t_i/n$ and the variance is $(n-1)\text{variance}(t)/n^2$.
   Show [13] algebrically that for complete survival data

$$\bar{t} = \frac{1}{n}\sum t_i = \sum P_{i-1}(t_i - t_{i-1})$$

   where $i = 1, 2, 3, \ldots, n$.

[11] Solve theoretically and not with a computer program.
[12] Solve theoretically and not with a computer program.
[13] Solve theoretically and not with a computer program.