Exercises for the course on Statistical Inference

Serik Sagitov, Chalmers University of Technology and Gothenburg University

Abstract

This collection of solved exercises is produced for the undergraduate course on course MVE155 "Statistical Inference". These are the end-of-chapter exercises form the textbook: John Rice, Mathematical statistics and data analysis, 3rd edition. The exercise numbers are the same as in the textbook.

Do not read a solution before you tried to solve an exercise on your own with the help of the lecture notes that can be downloaded here:

http://www.math.chalmers.se/Stat/Grundutb/CTH/mve155/1819/Statistical Inference.pdf

Please send your corrections to serik@chalmers.se. *Last updated: January 4, 2019*

Contents

Abstract **1**

1 Survey sampling **2**

2 Estimation of parameters **6**

3 Testing hypotheses and assessing goodness of fit **14**

4 Introduction to Bayesian inference **21**

5 Summarising data **24**

6 Comparing two samples **30**

7 Analysis of variance **38**

8 Categorical data analysis **44**

9 Multiple regression **51**

1 Survey sampling

Problem 7.1

Consider a population consisting of five values

$$1, 2, 2, 4, 8.$$

Find the population mean and variance. Calculate the sampling distribution of the mean of a sample of size 2 by generating all possible such samples. From them, find the mean and variance of the sampling distribution, and compare the results to those obtained by the formulas from Section 1 of the Lecture Notes.

Solution 7.1

Here we consider sampling with replacement. For an answer in the case of sampling without replacement consult the book page A36.

Population distribution

Values	1	2	4	8
Probab.	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

Population mean and variance are computed in three steps

$$\begin{aligned} \mu &= 1 \cdot \tfrac{1}{5} + 2 \cdot \tfrac{2}{5} + 4 \cdot \tfrac{1}{5} + 8 \cdot \tfrac{1}{5} = 3.4 \\ \mathrm{E}(X^2) &= 1 \cdot \tfrac{1}{5} + 4 \cdot \tfrac{2}{5} + 16 \cdot \tfrac{1}{5} + 64 \cdot \tfrac{1}{5} = 17.8 \\ \sigma^2 &= 17.8 - \mu^2 = 6.24. \end{aligned}$$

The list of $\bar{X}$ values (and their probabilities in brackets) for $n = 2$ observations taken with replacement:

	1	2	4	8	Total prob.
1	1.0 (1/25)	1.5 (2/25)	2.5 (1/25)	4.5 (1/25)	1/5
2	1.5 (2/25)	2.0 (4/25)	3.0 (2/25)	5.0 (2/25)	2/5
4	2.5 (1/25)	3.0 (2/25)	4.0 (1/25)	6.0 (1/25)	1/5
8	4.5 (1/25)	5.0 (2/25)	6.0 (1/25)	8.0 (1/25)	1/5
Tot. prob.	1/5	2/5	1/5	1/5	1

This yields the followig sampling distribution of $\bar{X}$:

Values	1	1.5	2	2.5	3	4	4.5	5	6	8
Probab.	$\frac{1}{25}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

Using the same three steps we find

$$\begin{aligned} \mathrm{E}(\bar{X}) &= 1 \cdot \tfrac{1}{25} + 1.5 \cdot \tfrac{4}{25} + 2 \cdot \tfrac{4}{25} + 2.5 \cdot \tfrac{2}{25} + 3 \cdot \tfrac{4}{25} + 4 \cdot \tfrac{1}{25} + 4.5 \cdot \tfrac{2}{25} + 5 \cdot \tfrac{4}{25} + 6 \cdot \tfrac{2}{25} + 8 \cdot \tfrac{1}{25} = 3.4 \\ \mathrm{E}(\bar{X}^2) &= \tfrac{1}{25} + (1.5)^2 \cdot \tfrac{4}{25} + 4 \cdot \tfrac{4}{25} + (2.5)^2 \cdot \tfrac{2}{25} + 9 \cdot \tfrac{4}{25} + 16 \cdot \tfrac{1}{25} + (4.5)^2 \cdot \tfrac{2}{25} + 25 \cdot \tfrac{4}{25} + 36 \cdot \tfrac{2}{25} + 64 \cdot \tfrac{1}{25} = 14.68 \\ \mathrm{Var}(\bar{X}) &= 14.68 - (3.4)^2 = 3.12. \end{aligned}$$

We see that indeed,

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = 3.12 = \frac{\sigma^2}{n}.$$

Problem 7.9

In a simple random sample of 1500 voters, 55% said they planned to vote for a particular proposition, and 45% said they planned to vote against it. The estimated margin of victory for the proposition is thus 10%. What is the standard error of this estimated margin? What is an approximate 95% confidence interval for the margin?

Solution 7.9

Dichotomous data

$$n = 1500, \quad \hat{p} = 0.55, \quad 1 - \hat{p} = 0.45, \quad s_{\hat{p}} = \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{\tfrac{0.55\times 0.45}{1499}} = 0.013.$$

Population margin of victory

$$v = p - (1-p) = 2p - 1.$$

Estimated margin of victory

$$\hat{v} = \hat{p} - (1-\hat{p}) = 2\hat{p} - 1 = 0.1.$$

(a) Since

$$\mathrm{Var}(\hat{V}) = \mathrm{Var}(2\hat{P}),$$

the standard error of $\hat{v}$ is twice the standard error of $\hat{p}$

$$s_{\hat{v}} = 2s_{\hat{p}} = 0.026.$$

(b) Approximate 95% confidence interval for v is

$$I_v = \hat{v} \pm 1.96 s_{\hat{v}} = 0.10 \pm 0.05.$$

Problem 7.19

This problem introduces the concept of a one-sided confidence interval. Using the central limit theorem, how should the constant k_1 be chosen so that the interval

$$(-\infty, \bar{x} + k_1 s_{\bar{x}})$$

is a 90% confidence interval for μ? How should k_2 be chosen so that

$$(\bar{x} - k_2 s_{\bar{x}}, \infty)$$

is a 95% confidence interval for μ?

Solution 7.19

Normal approximation: $\frac{\bar{X}-\mu}{S_{\bar{X}}}$ is asymptotically N(0,1)-distributed. From

$$0.90 \approx \mathrm{P}(\tfrac{\bar{X}-\mu}{S_{\bar{X}}} > -1.28) = \mathrm{P}(-\infty < \mu < \bar{X} + 1.28 S_{\bar{X}}),$$
$$0.95 \approx \mathrm{P}(\tfrac{\bar{X}-\mu}{S_{\bar{X}}} < 1.645) = \mathrm{P}(\bar{X} - 1.645 S_{\bar{X}} < \mu < \infty).$$

we find

$$k_1 = 1.28, \quad k_2 = 1.645.$$

Problem 7.28

Warner (1965) introduced the method of randomised response to deal with surveys asking sensitive questions. Suppose we want to estimate the proportion q of illegal drug users among prison inmates. We are interested in the population as a whole - not in punishing particular individuals. Randomly chosen n inmates have responded yes/no to a randomised statement (after rolling a die):

> "I use heroin" (with probability 5/6)
> "I do not use heroin" (with probability 1/6).

Suggest a probability model for this experiment, find a method of moments estimate for q and its standard error.

Solution 7.28

Randomised response method. Consider

$$x = \text{ number of "yes" responses for } n \text{ inmates.}$$

Then X has Bin (n, p) distribution, where

$$p = \mathrm{P}(\text{a "yes" answer}) = \frac{5}{6} \cdot q + \frac{1}{6} \cdot (1-q) = \frac{1+4q}{6}.$$

Replacing p by $\hat{p} = \frac{x}{n}$ we get an equation

$$\hat{p} = \frac{1+4\tilde{q}}{6},$$

whose solution gives a method of moments estimate $\tilde{q}$ of the population proportion q

$$\tilde{q} = \frac{6\hat{p}-1}{4}.$$

The estimate is unbiased

$$\mathrm{E}(\tilde{Q}) = \frac{6p-1}{4} = q.$$

Its variance equals

$$\mathrm{Var}(\tilde{Q}) = \frac{9}{4} \cdot \mathrm{Var}(\hat{Q}) = \frac{9}{4} \cdot \frac{p(1-p)}{n} = \frac{(1+4q)(5-4q)}{16n}.$$

Take for example $n = 40$, $x = 8$. Then $\hat{p} = 0.2$ and

$$\tilde{q} = \frac{6\hat{p}-1}{4} = 0.05.$$

The estimated standard error

$$s_{\tilde{q}} = \sqrt{\frac{(1+4\tilde{q})(5-4\tilde{q})}{16n}} = 0.095.$$

The estimate is unreliable. We have to increase the sample size.

Problem 7.35

A simple random sample of a population size 2000 yields the following 25 values:

104	109	11	109	87
86	80	119	88	122
91	103	99	108	96
104	98	98	83	107
79	87	94	92	97

(a) Calculate an unbiased estimate of the population mean.
(b) Calculate an unbiased estimates of the population variance and $\mathrm{Var}(\bar{X})$.
(c) Give an approximate 95% confidence interval for the population mean and total.

Solution 7.35

Data summary

$$N = 2000, \quad n = 25, \quad \sum x_i = 2451, \quad \sum x_i^2 = 243505.$$

(a) Unbiased estimate of μ is

$$\bar{x} = \frac{2451}{25} = 98.04.$$

(b) Unbiased estimate of σ^2 is

$$\frac{N-1}{N} s^2 = \frac{1999}{2000} 133.71 = 133.64,$$

where

$$s^2 = \frac{n}{n-1}(\overline{x^2} - \bar{x}^2) = \frac{25}{24}\Big(\frac{243505}{25} - (98.04)^2\Big) = 133.71.$$

Unbiased estimate of $\mathrm{Var}(\bar{X})$ is

$$s_{\bar{x}}^2 = \frac{s^2}{n}\Big(1 - \frac{n}{N}\Big) = 5.28.$$

(b) An approximate 95% confidence interval for μ

$$I_\mu = \bar{x} \pm 1.96 s_{\bar{x}} = 98.04 \pm 1.96\sqrt{5.28} = 98.04 \pm 4.50.$$

Approximate 95% confidence interval for the total $\tau = N\mu$

$$I_\tau = N\bar{x} \pm 1.96 N s_{\bar{x}} = 196080 \pm 9008.$$

Problem 7.36

For a simple random sample, take $\bar{x}^2$ as a point estimate of μ^2. (This is an example of the method of moments estimate.) Compute the bias of this point estimate.

Solution 7.36

The bias is

$$\mathrm{E}(\bar{X}^2) - \mu^2 = \mathrm{E}(\bar{X}^2) - (\mathrm{E}\bar{X})^2 = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\Big(1 - \frac{n-1}{N-1}\Big).$$

For large n, the bias is small.

Problem 7.53

The following table (Cochran 1977) shows the stratification of all farms in a county by farm size and the mean and standard deviation of the number of acres of corn in each stratum.

Farm size	0-40	41-80	81-120	121-160	161-200	201-240	241+
Number of farms N_j	394	461	391	334	169	113	148
Stratum mean μ_j	5.4	16.3	24.3	34.5	42.1	50.1	63.8
Stratum standard deviation σ_j	8.3	13.3	15.1	19.8	24.5	26.0	35.2

(a) For a sample size of 100 farms, compute the sample sizes from each stratum for proportional and optimal allocation, and compare them.

(b) Calculated the variances of the sample mean for each allocation and compare them to each other and to the variance of an estimated formed from simple random sampling.

(c) What are the population mean and variance?

(d) Suppose that ten farms are sampled per stratum. What is $\mathrm{Var}(\bar{X}_s)$? How large a simple random sample would have to be taken to attain the same variance? Ignore the finite population correction.

(e) Repeat part (d) using proportional allocation of the 70 samples.

Solution 7.53

Stratified population of size $N = 2010$ with $k = 7$ strata.

(a) With $n = 100$, we get the following answers using the relevant formulas

Stratum number j	1	2	3	4	5	6	7	Weighted mean
Stratum proportion w_j	0.20	0.23	0.19	0.17	0.08	0.06	0.07	
Stratum mean μ_j	5.4	16.3	24.3	34.5	42.1	50.1	63.8	$\mu = 26.49$
Stratum standard deviation σ_j	8.3	13.3	15.1	19.8	24.5	26.0	35.2	$\bar{\sigma} = 17.04$
Optimal allocation $n\frac{w_j\sigma_l}{\bar{\sigma}_j}$	10	18	17	19	12	9	15	
Proportional allocation nw_j	20	23	19	17	8	6	7	

(b) $\mathrm{Var}(\bar{X}_{\mathrm{so}}) = \frac{\bar{\sigma}^2}{n} = 2.90$, $\mathrm{Var}(\bar{X}_{\mathrm{sp}}) = \frac{\overline{\sigma^2}}{n} = 3.44$, $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} = 6.21$, where σ^2 is computed in the next item.

(c) We have $\mu = 26.49$, and $\bar{\sigma}^2 = 275.33$. Moreover $\overline{\sigma^2} = 347.40$, and therfore

$$\sigma^2 = 347.40 + 275.33 = 622.73, \qquad \sigma = 24.95.$$

(d) If $n_1 = \ldots = n_7 = 10$ and $n = 70$, then $\mathrm{Var}(\bar{X}_{\mathrm{s}}) = 4.45$. The requested sample size $x = 140$ is found from the equation

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{x} = 4.45.$$

(e) If $n = 70$, then $\mathrm{Var}(\bar{X}_{\mathrm{sp}}) = 4.92$. Solving the equation

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{x} = 4.92,$$

we find that the the requested sample size is $x = 127$ which is smaller than that the answer for (d).

Problem 7.57

Consider stratifying the population of Problem 7.1 into two strata (1,2,2) and (4,8). Assuming that one observation is taken from each stratum, find the sampling distribution of the estimate of the population mean and the mean and standard deviation of the sampling distribution. Check the formulas of Section 1.4 in the Lecture Notes.

Solution 7.57

Stratified population with

$$N = 5, \quad k = 2, \quad w_1 = 0.6, \quad w_2 = 0.4, \quad \mu_1 = 1.67, \quad \mu_2 = 6, \quad \sigma_1^2 = 0.21, \quad \sigma_2^2 = 4.$$

Given $n_1 = n_2 = 1$ and $n = 2$, the sampling distribution of the stratified sample mean $\bar{x}_s = 0.6x_1 + 0.4x_2$ is

	$x_1 = 1$	$x_1 = 2$	Total prob.
$x_2 = 4$	2.2 (1/6)	2.8 (2/6)	1/2
$x_2 = 8$	3.8 (1/6)	4.4 (2/6)	1/2
Tot. prob.	1/3	2/3	1

We find that

$$\begin{aligned}
\mathrm{E}(\bar{X}_s) &= 2.2 \cdot \tfrac{1}{6} + 2.8 \cdot \tfrac{2}{6} + 3.8 \cdot \tfrac{1}{6} + 4.4 \cdot \tfrac{1}{6} = 3.4,\\
(\mathrm{E}(\bar{X}_s))^2 &= 11.56,\\
\mathrm{E}(\bar{X}_s^2) &= (2.2)^2 \cdot \tfrac{1}{6} + (2.8)^2 \cdot \tfrac{2}{6} + (3.8)^2 \cdot \tfrac{1}{6} + (4.4)^2 \cdot \tfrac{2}{6} = 12.28,\\
\mathrm{Var}(\bar{X}_s) &= 12.28 - 11.56 = 0.72.
\end{aligned}$$

These results are in agreement with the formulas

$$\mathrm{E}(\bar{X}_s) = \mu, \quad \mathrm{Var}(\bar{X}_s) = \tfrac{w_1^2\sigma_1^2}{n_1} + \ldots + \tfrac{w_k^2\sigma_k^2}{n_k} = 0.36\sigma_1^2 + 0.16\sigma_2^2.$$

2 Estimation of parameters

Problem 8.3

One of the earliest applications of the Poisson distribution was made by Student (1907) in studying errors made in counting yeast cells.In this study, yeast cells were killed and mixed with water and gelatin; the mixture was then spread on a glass and allowed to cool. Four different concentrations were used. Counts were made on 400 squares, and the data are summarised in the following table:

Number of cells	Concent. 1	Concent. 2	Concent. 3	Concent. 4
0	213	103	75	0
1	128	143	103	20
2	37	98	121	43
3	18	42	54	53
4	3	8	30	86
5	1	4	13	70
6	0	2	2	54
7	0	0	1	37
8	0	0	0	18
9	0	0	1	10
10	0	0	0	5
11	0	0	0	2
12	0	0	0	2

(a) Estimate the parameter λ for each of the four sets of data.
(b) Find an approximate 95% confidence interval for each estimate.
(c) Compare observed and expected counts.

Solution 8.3

Number X of yeast cells on a square. Test the Poisson model $X \sim \text{Pois}(\lambda)$.

Concentration 1.

$$\bar{x} = 0.6825, \quad \overline{x^2} = 1.2775, \quad s^2 = 0.8137, \quad s = 0.9021, \quad s_{\bar{x}} = 0.0451.$$

Approximate 95% confidence interval

$$I_\mu = 0.6825 \pm 0.0884.$$

Pearson's chi-square test based on $\hat{\lambda} = 0.6825$:

x	0	1	2	3	4+	Total
Observed	213	128	37	18	4	400
Expected	202.14	137.96	47.08	10.71	2.12	400

Observed test statistic $\chi^2 = 10.12$, df $= 5 - 1 - 1 = 3$, p-value < 0.025. Reject the model.

Concentration 2.

$$\bar{x} = 1.3225, \quad \overline{x^2} = 3.0325, \quad s = 1.1345, \quad s_{\bar{x}} = 0.0567.$$

Approximate 95% confidence interval

$$I_\mu = 1.3225 \pm 0.1112.$$

Pearson's chi-square test: observed test statistic $\chi^2 = 3.16$, df $= 4$, p-value > 0.10. Do not reject the model.

Concentration 3.

$$\bar{x} = 1.8000, \quad s = 1.1408, \quad s_{\bar{x}} = 0.0701.$$

Approximate 95% confidence interval for

$$I_\mu = 1.8000 \pm 0.1374.$$

Pearson's chi-square test: observed test statistic $\chi^2 = 7.79$, df $= 5$, p-value > 0.10. Do not reject the model.

Concentration 4.

$$n = 410, \quad \bar{x} = 4.5659, \quad s^2 = 4.8820, \quad s_{\bar{x}} = 0.1091.$$

Approximate 95% confidence interval

$$I_\mu = 4.566 \pm 0.214.$$

Pearson's chi-square test: observed test statistic $\chi^2 = 13.17$, df $= 10$, p-value > 0.10. Do not reject the model.

Problem 8.4

Suppose that X is a discrete random variable with

$$\begin{aligned} \mathrm{P}(X=0) &= \tfrac{2}{3}\theta, \\ \mathrm{P}(X=1) &= \tfrac{1}{3}\theta, \\ \mathrm{P}(X=2) &= \tfrac{2}{3}(1-\theta), \\ \mathrm{P}(X=3) &= \tfrac{1}{3}(1-\theta), \end{aligned}$$

where $\theta \in [0,1]$ is parameter. The following 10 independent observations were taken from such a distribution:

$$(3, 0, 2, 1, 3, 2, 1, 0, 2, 1).$$

(a) Find the method of moments estimate of θ.
(b) Find an approximate standard error for your estimate.
(c) What is the maximum likelihood estimate of θ?
(d) What is an approximate standard error of the maximum likelihood estimate?

Solution 8.4

Population distribution: X takes values $0, 1, 2, 3$ with probabilities

$$p_0 = \frac{2}{3} \cdot \theta, \quad p_1 = \frac{1}{3} \cdot \theta, \quad p_2 = \frac{2}{3} \cdot (1-\theta), \quad p_3 = \frac{1}{3} \cdot (1-\theta),$$

so that

$$p_0 + p_1 = \theta, \quad p_2 + p_3 = 1 - \theta.$$

We are given an iid-sample with

$$n = 10, \quad \bar{x} = 1.5, \quad s = 1.08,$$

and observed counts

x	0	1	2	3	Total
O_x	2	3	3	2	10

(a) Method of moments. Using

$$\mu = \frac{1}{3} \cdot \theta + 2 \cdot \frac{2}{3} \cdot (1-\theta) + 3 \cdot \frac{1}{3} \cdot (1-\theta) = \frac{7}{3} - 2\theta,$$

derive an equation

$$\bar{x} = \frac{7}{3} - 2\tilde{\theta}.$$

It gives an unbiased estimate

$$\tilde{\theta} = \frac{7}{6} - \frac{\bar{x}}{2} = \frac{7}{6} - \frac{3}{4} = 0.417.$$

(b) To find $s_{\tilde{\theta}}$, observe that

$$\mathrm{Var}(\tilde{\Theta}) = \frac{1}{4}\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{40}.$$

Thus we need to find $s_{\tilde{\theta}}$, which estimates $\sigma_{\tilde{\theta}} = \frac{\sigma}{6.325}$. Next we estimate σ using two methods.

Method 1. From

$$\sigma^2 = \mathrm{E}(X^2) - \mu^2 = \frac{1}{3}\cdot\theta + 4\cdot\frac{2}{3}\cdot(1-\theta) + 9\cdot\frac{1}{3}\cdot(1-\theta) = \frac{7}{3} - 2\theta - \left(\frac{7}{3} - 2\theta\right)^2 = \frac{2}{9} + 4\theta - 4\theta^2,$$

we estimate σ as

$$\sqrt{\frac{2}{9} + 4\tilde\theta - 4\tilde\theta^2} = 1.093.$$

This gives

$$s_{\tilde\theta} = \frac{1.093}{6.325} = 0.173.$$

Method 2:

$$s_{\tilde\theta} = \frac{s}{6.325} = \frac{1.08}{6.325} = 0.171.$$

(c) Likelihood function is obtained using $(O_0, O_1, O_2, O_3) \sim \mathrm{Mn}(n, p_0, p_1, p_2, p_3)$

$$L(\theta) = \left(\tfrac{2}{3}\theta\right)^{O_0}\left(\tfrac{1}{3}\theta\right)^{O_1}\left(\tfrac{2}{3}(1-\theta)\right)^{O_2}\left(\tfrac{1}{3}(1-\theta)\right)^{O_3} = \text{const } \theta^t(1-\theta)^{n-t},$$

where $t = O_0 + O_1$ is a sufficient statistic. Notice that $T = O_0 + O_1$ has $\mathrm{Bin}(n,\theta)$ distribution.

Log-likelihood and its derivative

$$l(\theta) = \text{const } + t\ln\theta + (n-t)\ln(1-\theta),$$
$$l'(\theta) = \frac{t}{\theta} - \frac{n-t}{1-\theta}.$$

Setting the last expression to zero, we find

$$\frac{t}{\hat\theta} = \frac{n-t}{1-\hat\theta}, \quad \hat\theta = \frac{t}{n} = \frac{2+3}{10} = \frac{1}{2}.$$

The maximum likelihood estimate is the sample proportion, an unbiased estimate of the population proportion θ.

(d) We find $s_{\hat\theta}$ using the formula for the standard error of sample proportion

$$s_{\hat\theta} = \sqrt{\tfrac{\hat\theta(1-\hat\theta)}{n-1}} = 0.167.$$

A similar answer is obtained using the formula

$$s_{\hat\theta} = \sqrt{\tfrac{1}{nI(\hat\theta)}}, \quad I(\theta) = -\mathrm{E}(g(Y,\theta)), \quad g(y,\theta) = \tfrac{\partial^2}{\partial\theta^2}\ln f(y|\theta),$$

where $Y \sim \mathrm{Ber}(\theta)$. Since $f(1|\theta) = \theta$, $f(0|\theta) = 1-\theta$, we have

$$g(1,\theta) = \frac{\partial^2}{\partial\theta^2}\ln\theta = -\frac{1}{\theta^2}, \quad g(0,\theta) = \frac{\partial^2}{\partial\theta^2}\ln(1-\theta) = -\frac{1}{(1-\theta)^2},$$

we get

$$I(\theta) = -\mathrm{E}\left(g(Y,\theta)\right) = g(1,\theta)f(1|\theta) + g(0,\theta)f(0|\theta) = \frac{1}{\theta^2}\cdot\theta + \frac{1}{(1-\theta)^2}\cdot(1-\theta) = \frac{1}{\theta(1-\theta)}.$$

Problem 8.6

Suppose that $X \sim \mathrm{Bin}(n,p)$.

(a) Show that the maximum estimate of p is $\hat p = \frac{x}{n}$.

(b) Show that $\hat p = \frac{x}{n}$ attains the Cramer-Rao lower bound.

(c) If $n = 10$ and $X = 5$, plot the log-likelihood function.

Solution 8.6

Likelihood function of $X \sim \text{Bin}(n,p)$ for a given n and $X = x$ is

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x} \propto p^x (1-p)^{n-x}.$$

(a) To maximise $L(p)$ we minimise

$$\ln p^x (1-p)^{n-x}) = x \ln p + (n-x) \ln(1-p).$$

Since

$$\frac{\partial}{\partial p}(x \ln p + (n-x)\ln(1-p)) = \frac{x}{p} - \frac{n-x}{1-p},$$

we have to solve $\frac{x}{p} = \frac{n-x}{1-p}$, which brings the maximum likelihood estimate formula $\hat{p} = \frac{x}{n}$.

(b) We have $X = Y_1 + \ldots + Y_n$, where $(Y_1, \ldots, Y_n)$ are iid Bernoulli random variables with

$$f(y|p) = p^y (1-p)^{1-y}, \quad y = 0, 1.$$

By Cramer-Rao, if $\tilde{p}$ is an unbiased estimate of p, then

$$\text{Var}(\tilde{P}) \geq \frac{1}{nI(p)},$$

where

$$I(p) = -\text{E}\left(\frac{\partial^2}{\partial p^2} \ln f(Y|p)\right) = \frac{1}{p(1-p)},$$

see Solution 8.4 d. We conclude that the variance sample proportion $\hat{p}$ attains the Cramer-Rao lower bound since

$$\text{Var}(\hat{P}) = \tfrac{p(1-p)}{n}.$$

(c) Plot $L(p) = 252p^5(1-p)^5$. The top of the curve is in the middle $\hat{p} = 0.5$.

Problem 8.26

Capture-recapture method for estimating the number N of fish living in a lake:

1. capture and tag say $n = 100$ fish, then release them in the lake,
2. recapture say $k = 50$ fish and count the number of tagged fish.

Suppose $x = 20$ fish were tagged among the $k = 50$ fish. Find a maximum likelihood estimate N after suggesting a simple parametric model.

Solution 8.26

Statistical model: x is the number of black balls obtained by sampling k balls without replacement from an urn with N balls of which n balls are black. Hypergeometric distribution

$$\text{P}(X = 20) = \frac{\binom{n}{20}\binom{N-n}{30}}{\binom{N}{50}}.$$

The likelihood function

$$L(N) = \frac{\binom{100}{20}\binom{N-100}{30}}{\binom{N}{50}} = \text{const} \cdot \frac{(N-100)(N-101)\cdots(N-129)}{N(N-1)\cdots(N-49)}.$$

To find the value of $N = \hat{N}$ that maximises $L(N)$, consider the ratio

$$\frac{L(N)}{L(N-1)} = \frac{(N-100)(N-50)}{N(N-130)}.$$

If $N < \hat{N}$, then $\frac{L(N)}{L(N-1)} > 1$, and $N > \hat{N}$, then $\frac{L(N)}{L(N-1)} < 1$ Solving the equation

$$\frac{L(\hat{N})}{L(\hat{N}-1)} = 1 \Leftrightarrow (\hat{N}-100)(\hat{N}-50) = \hat{N}(\hat{N}-130),$$

we arrive at the maximum likelihood estimate estimate $\hat{N} = \frac{5000}{20} = 250$. The answer is very intuitive as we expect that

$$100 : N \approx 20 : 50.$$

Problem 8.32

The following 16 numbers came from normal random generator on a computer

5.3299	4.2537	3.1502	3.7032
1.6070	6.3923	3.1181	6.5941
3.5281	4.7433	0.1077	1.5977
5.4920	1.7220	4.1547	2.2799

(a) What would you guess the mean and the variance of the generating normal distribution were?
(b) Give 90%, 95%, and 99% confidence intervals for μ and σ^2.
(c) Give 90%, 95%, and 99% confidence intervals for σ.
(d) How much larger sample do you think you would need to halve the length of the confidence interval for μ?

Solution 8.32

An iid-sample of size $n = 16$ from a normal distribution.

(a) The summary statistics

$$\bar{x} = 3.6109, \quad s^2 = 3.4181, \quad s_{\bar{x}} = 0.4622$$

suggest an estimate for μ to be 3.6109, and an estimate for σ^2 to be 3.4181.

(b), (c) Exact confidence intervals

	90%	95%	99%
I_μ	3.61 ± 0.81	3.61 ± 0.98	3.61 ± 1.36
I_{σ^2}	(2.05; 7.06)	(1.87; 8.19)	(1.56; 11.15)
I_σ	(1.43; 2.66)	(1.37; 2.86)	(1.25; 3.34)

(d) To find sample size x that halves the confidence interval length we set up an equation using the exact confidence interval formula for the mean

$$t_{15}(\alpha/2) \cdot \frac{s}{\sqrt{16}} = 2 \cdot t_{x-1}(\alpha/2) \cdot \frac{s'}{\sqrt{x}},$$

where s' is the sample standard deviation for the sample of size x. A simplistic version of this equation $\frac{1}{4} = \frac{2}{\sqrt{x}}$ implies $x \approx (2 \cdot 4)^2 = 64$. Further adjustment for a 95% confidence interval is obtained using

$$t_{15}(\alpha/2) = 2.13, \quad t_{x-1}(\alpha/2) \approx 2,$$

yielding $x \approx (2 \cdot 4 \cdot \frac{2}{2.13})^2 = 56.4$. We conclude that going from a sample of size 16 to a sample of size 56 would halve the length of the confidence interval for μ.

Problem 8.53

Let $X_1, \ldots, X_n$ be i. i. d. uniform on $[0, \theta]$.

(a) Find the method of moments estimate of θ and its mean and variance.
(b) Find the maximum likelihood estimate of θ.
(c) Find the probability density of the maximum likelihood estimate and calculate its mean and variance. Compare the variance, the bias, and the mean square error to those of the method of moments estimate.
(d) Find a modification of the maximum likelihood estimate that renders it unbiased.

Solution 8.53

An IID sample $(X_1, \ldots, X_n)$ from the uniform distribution $\mathrm{U}(0, \theta)$ with density

$$f(x|\theta) = \tfrac{1}{\theta} 1_{\{0 \le x \le \theta\}}.$$

(a) Method of moments estimate $\tilde{\theta}$ is unbiased

$$\mu = \theta/2, \quad \tilde{\theta} = 2\bar{x}, \quad \mathrm{E}(\tilde{\Theta}) = \theta, \quad \mathrm{Var}(\tilde{\Theta}) = \frac{4\sigma^2}{n} = \frac{\theta^2}{3n}.$$

(b) Denote $x_{(n)} = \max(x_1, \ldots, x_n)$. Likelihood function takes the form

$$L(\theta) = f(x_1|\theta) \cdots f(x_n|\theta) = \tfrac{1}{\theta^n} 1_{\{\theta \ge x_1\}} \cdots 1_{\{\theta \ge x_n\}} = \tfrac{1}{\theta^n} 1_{\{\theta \ge x_{(n)}\}},$$

so that $x_{(n)}$ is a sufficient statistic. The maximum is achieved at $\hat{\theta} = x_{(n)}$.

(c) Sampling distribution of the maximum likelihood estimate $\hat{\theta} = x_{(n)}$:

$$\mathrm{P}(X_{(n)} \le x) = \mathrm{P}(X_1 \le x, \ldots, X_n \le x) = \mathrm{P}(X_1 \le x) \cdots \mathrm{P}(X_n \le x) = \left(\frac{x}{\theta}\right)^n$$

with pdf

$$f_{\hat{\Theta}}(x) = \frac{n}{\theta^n} \cdot x^{n-1}, \quad 0 \le x \le \theta.$$

The maximum likelihood estimate is biased

$$\mathrm{E}(\hat{\Theta}) = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{n+1}\theta, \quad \mathrm{E}(\hat{\Theta}^2) = \frac{n}{n+2}\theta^2, \quad \mathrm{Var}(\hat{\Theta}) = \frac{\theta^2}{(n+1)^2(n+2)},$$

but asymptotically unbiased. Compare two mean square errors:

$$\mathrm{MSE}(\hat{\Theta}) = \mathrm{E}(\hat{\Theta} - \theta)^2 = \left(-\frac{\theta}{n+1}\right)^2 + \frac{\theta^2}{(n+1)^2(n+2)} = \frac{n+3}{n+2} \cdot \frac{\theta^2}{(n+1)^2},$$

$$\mathrm{MSE}(\tilde{\Theta}) = \frac{\theta^2}{3n}.$$

(d) Corrected maximum likelihood estimate

$$\hat{\theta}_c = \frac{n+1}{n} \cdot x_{(n)}$$

becomes unbiased $\mathrm{E}(\hat{\Theta}_c) = \theta$ with $\mathrm{Var}(\hat{\Theta}_c) = \frac{\theta^2}{n^2(n+2)}$.

Problem 8.55

For two factors, starchy-or-sugary and green-or-white base leaf, the following counts for the progeny of self-firtilized heterozygotes were observed (Fisher 1958)

Type	Count
Starchy green	$x_1 = 1997$
Starchy white	$x_2 = 906$
Sugary green	$x_3 = 904$
Sugary white	$x_4 = 32$

According to the genetic theory the cell probabilities are

$$p_1 = \frac{2+\theta}{4}, \quad p_2 = \frac{1-\theta}{4}, \quad p_3 = \frac{1-\theta}{4}, \quad p_4 = \frac{\theta}{4},$$

where $0 < \theta < 1$. In particular, if $\theta = 0.25$, then the genes are unlinked and the genotype frequencies are

	Green	White	Total
Starchy	$9/16$	$3/16$	$3/4$
Sugary	$3/16$	$1/16$	$1/4$
Total	$3/4$	$1/4$	1

(a) Find the maximum likelihood estimate of θ and its asymptotic variance.

(b) For an approximate 95% confidence interval for θ based on part (a).

(c) Use the bootstrap to find the approximate standard deviation of the maximum likelihood estimate and compare to the result of part (a).

Solution 8.55

(a) Sample counts $(X_1, X_2, X_3, X_4) \sim \text{Mn}(n, p_1, p_2, p_3, p_4)$ with $n = 3839$. The likelihood function

$$L(\theta) = \binom{n}{x_1, x_2, x_3, x_4} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} = \binom{n}{x_1, x_2, x_3, x_4} (2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4} 4^{-n}$$

reveals two sufficient statistics (x_1, x_4), as

$$x_2 + x_3 = n - x_1 - x_4.$$

Putting

$$\frac{d}{d\theta} \ln L(\theta) = \frac{x_1}{2+\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta}$$

equal to zero, we arrive at the equation

$$\frac{x_1}{2+\theta} + \frac{x_4}{\theta} = \frac{x_2+x_3}{1-\theta}$$

or equivalently

$$\theta^2 n + \theta u - 2x_4 = 0,$$

where $u = 2x_2 + 2x_3 + x_4 - x_1$. We find the maximum likelihood estimate to be

$$\hat{\theta} = \frac{-u + \sqrt{u^2 + 8nx_4}}{2n} = 0.0357.$$

Asymptotic variance

$$\text{Var}(\hat{\Theta}) \approx \frac{1}{I(\theta)}, \quad I(\theta) = -\text{E}(g(X_1, X_2, X_3, X_4, \theta)).$$

where

$$g(x_1,x_2,x_3,x_4,\theta) = \frac{\partial^2}{\partial\theta^2}\ln f(x_1,x_2,x_3,x_4|\theta) = -\frac{x_1}{(2+\theta)^2} - \frac{x_2+x_3}{(1-\theta)^2} - \frac{x_4}{\theta^2}.$$

Since $\mathrm{E}(X_i) = np_i$, we find

$$I(\theta) = \frac{n}{4(2+\theta)} + \frac{2n}{4(1-\theta)} + \frac{n}{4\theta} = \frac{n(1+2\theta)}{2\theta(2+\theta)(1-\theta)},$$

and get $I(\hat\theta) = 29345.8$, so that $s_{\hat\theta} = 0.0058$.

(b) $I_\theta = 0.0357 \pm 1.96 \cdot 0.0058 = 0.0357 \pm 0.0114$.

(c) Parametric bootstrap using Matlab:

```
p1=0.5089, p2=0.2411, p3=0.2411, p4=0.0089,
n=3839; B=1000; b=ones(B,1);
x1=binornd(n,p1,B,1);
x2=binornd(n*b-x1,p2/(1-p1));
x3=binornd(n*b-x1-x2,p3/(1-p1-p2));
x4=n*b-x1-x2-x3;
u=2*x2+2*x3+x4-x1;
t=(-u+sqrt(u.ˆ2+8*n*x4))/(2*n);
std(t)
histfit(t)
```

gives std(t)=0.0058 similar to the answer in (a).

3 Testing hypotheses and assessing goodness of fit

Problem 9.3

Suppose that $X \sim \mathrm{Bin}(100, p)$. Consider a test

$$H_0: p = 1/2, \quad H_1: p \neq 1/2.$$

that rejects H_0 in favour of H_1 for $|x-50| > 10$. Use the normal approximation to the binomial distribution to answer the following:

(a) What is α?

(b) Graph the power as a function of p.

Solution 9.3

The z-score

$$Z = \frac{X-100p}{10\sqrt{p(1-p)}}$$

has a distribution that is approximated ny $\mathrm{N}(0,1)$.

(a) Under H_0 we have

$$Z = \frac{X-50}{10\sqrt{p(1-p)}},$$

and the significance level in question is

$$\alpha = \mathrm{P}(|X-50| > 10) = \mathrm{P}(|Z| > 2) \approx 2(1-\Phi(2)) = 2\cdot 0.0228 = 0.046.$$

(b) The power of the test is a function of the parameter value p

$$\begin{aligned}\mathrm{Pw}(p) &= \mathrm{P}(|X-50|>10) = \mathrm{P}(X<40)+\mathrm{P}(X>60)\\ &= \mathrm{P}\left(Z<\frac{40-100p}{10\sqrt{p(1-p)}}\right)+\mathrm{P}\left(Z>\frac{60-100p}{10\sqrt{p(1-p)}}\right)\\ &= \Phi\left(\frac{4-10p}{\sqrt{p(1-p)}}\right)+\Phi\left(\frac{10p-6}{\sqrt{p(1-p)}}\right).\end{aligned}$$

Putting $\delta = 1/2 - p$, we see that the power function

$$\mathrm{Pw}(p) = \Phi\left(\frac{10\delta-1}{\sqrt{1/4-\delta^2}}\right)+\Phi\left(-\frac{10\delta+1}{\sqrt{1/4-\delta^2}}\right)$$

is symmetric around $p = 1/2$

p	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
Pw(p)	0.986	0.853	0.500	0.159	0.046	0.159	0.500	0.853	0.986

Problem 9.4

Let X have one of the following two distributions

X-values	x_1	x_2	x_3	x_4
$\mathrm{P}(x\|H_0)$	0.2	0.3	0.3	0.2
$\mathrm{P}(x\|H_1)$	0.1	0.4	0.1	0.4

(a) Compare the likelihood ratio, Λ, for each x_i and order the x_i according to Λ.

(b) What is the likelihood ratio test of H_0 versus H_1 at level $\alpha = 0.2$? What is the test at level $\alpha = 0.5$?

Solution 9.4

Data: one observation of $X = x$. Likelihood ratio test: reject for small values of $\Lambda = \frac{\mathrm{P}(x|H_0)}{\mathrm{P}(x|H_1)}$.

(a) See the bottom line of the table:

X-values	x_4	x_2	x_1	x_3
$\mathrm{P}(x\|H_0)$	0.2	0.3	0.2	0.3
$\mathrm{P}(x\|H_1)$	0.4	0.4	0.1	0.1
Likelihood ratio $\Lambda = \frac{\mathrm{P}(x\|H_0)}{\mathrm{P}(x\|H_1)}$	0.5	0.75	2	3

(b) The null distribution of Λ

X-values	x_4	x_2	x_1	x_3
Likelihood ratio Λ	0.5	0.75	2	3
$\mathrm{P}(x\|H_0)$	0.2	0.3	0.2	0.3
Cumulative probab.	0.2	0.5	0.7	1

At $\alpha = 0.2$ we reject H_0 only if $\Lambda = 0.5$, that is when $X = x_4$.
At $\alpha = 0.5$ we reject H_0 for $\Lambda \leq 0.75$, that is when $X = x_4$ or x_2.

Problem 9.7

Let $(x_1, \ldots, x_n)$ be a sample from a Poisson distribution. Find the likelihood ratio for testing $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda = \lambda_1$, where $\lambda_1 > \lambda_0$. Use the fact that the sum of independent Poisson random variables follows a Poisson distribution to explain how to determine a rejection region for a test at level α.

Solution 9.7

Likelihood function

$$L(\lambda) = \prod_{i=1}^{n} \frac{1}{x_i!}\lambda^{x_i}e^{-\lambda} = e^{-\lambda n}\lambda^{y}\prod_{i=1}^{n}\frac{1}{x_i!}$$

where

$$y = x_1 + \ldots + x_n$$

is a sufficient statistic. Reject H_0 for small values of the likelihood ratio

$$\frac{L(\lambda_0)}{L(\lambda_1)} = e^{-n(\lambda_0-\lambda_1)}(\tfrac{\lambda_0}{\lambda_1})^{y}.$$

If $\lambda_1 > \lambda_0$, then we reject H_0 for large values of y. Test statistic Y has null distribution $\text{Pois}(n\lambda_0)$.

Problem 9.9

Let $(x_1, \ldots, x_{25})$ be a sample from a normal distribution having a variance of 100.

(a) Find the rejection region for a test at level $\alpha = 0.1$ of $H_0 : \mu = 0$ versus $H_1 : \mu = 1.5$.
(b) What is the power of the test?
(c) Repeat for $\alpha = 0.01$.

Solution 9.9

IID sample from $\text{N}(\mu, 100)$ of size $n = 25$. Two simple hypotheses

$$H_0 : \mu = 0, \quad H_1 : \mu = 1.5$$

Test statistic and its exact sampling distribution

$$\bar{X} \sim \text{N}(\mu, 4).$$

Its null and alternative distributions are

$$\bar{X} \overset{H_0}{\sim} \text{N}(0, 4), \quad \bar{X} \overset{H_1}{\sim} \text{N}(1.5, 4).$$

(a) The rejection region at $\alpha = 0.1$ is $\{\bar{x} > x\}$, where x is the solution of the equation

$$0.1 = \text{P}(\bar{X} > x|H_0) = 1 - \Phi(x/2).$$

From the normal distribution table we find $x/2 = 1.28$, so that $x = 2.56$ and the rejection region is

$$\mathcal{R} = \{\bar{x} > 2.56\}.$$

(b) The power of the test (a) is

$$\text{P}(\bar{X} > 2.56|H_1) = \text{P}(\tfrac{\bar{X}-1.5}{2} > 0.53|H_1) = 1 - \Phi(0.53) = 1 - 0.7019 = 0.298.$$

(c) For $\alpha = 0.01$, since $1 - \Phi(2.33) = 0.01$, the rejection region is

$$\mathcal{R} = \{\bar{x} > 4.66\}.$$

The power of this test is

$$\text{P}(\bar{X} > 4.66|H_1) = \text{P}(\tfrac{\bar{X}-1.5}{2} > 1.58|H_1) = 1 - \Phi(1.58) = 1 - 0.9429 = 0.057.$$

Problem 9.22

Let $(x_1, \ldots, x_{15})$ be a sample from a normal distribution.

(a) Find a 95% confidence interval for the population variance σ^2.
(b) Using the confidence interval method test $H_0 : \sigma = 1$ versus $H_0 : \sigma \neq 1$ at $\alpha = 0.05$.

Solution 9.22

Let s^2 be the sample variance computed from $(x_1, \ldots, x_{15})$.

(a) An exact 95% confidence interval for σ^2 if $n = 15$ is

$$I_{\sigma^2} = (0.536s^2; 2.487s^2).$$

(b) We reject $H_0 : \sigma = 1$ if the value 1 falls outside the confidence interval interval I_{σ^2}, so that

$$\mathcal{R} = \{1 \notin (0.536s^2; 2.487s^2)\} = \{s^2 > 1.866 \quad \text{or} \quad s^2 < 0.402\}.$$

Problem 9.23

An iid-sample from $\mathrm{N}(\mu, \sigma^2)$ gives a 99% confidence interval for μ to be $(-2, 3)$. Test

$$H_0 : \mu = -3 \quad \text{against} \quad H_1 : \mu \neq -3$$

at $\alpha = 0.01$.

Solution 9.23

Using the confidence interval-method of hypotheses testing we reject H_0 in favour of the two-sided alternative, since the value $\mu = -3$ is not covered by the two-sided confidence interval $(-2, 3)$.

Problem 9.24

Binomial model for the data value x:

$$X \sim \mathrm{Bin}(n, p).$$

(a) What is the generalised likelihood ratio for testing $H_0 : p = 0.5$ against $H_1 : p \neq 0.5$?
(b) Show that the test rejects for large values of $|x - \frac{n}{2}|$.
(c) How the significance level corresponding to the rejection region

$$\mathcal{R} = \{|x - \frac{n}{2}| > k\}$$

can be determined?
(d) If $n = 10$ and $k = 2$, what is the significance level of the test?
(e) Use the normal approximation to the binomial distribution to find the significance level if $n = 100$ and $k = 10$.

Solution 9.24

The analysis is the basis of the sign test.

(a) Generalised likelihood ratio

$$\Lambda = \frac{L(p_0)}{L(\hat{p})} = \frac{\binom{n}{x} p_0^x (1 - p_0)^{n-x}}{\binom{n}{x} \hat{p}^x (1 - \hat{p})^{n-x}} = \frac{(\frac{1}{2})^n}{(\frac{x}{n})^x (\frac{n-x}{n})^{n-x}} = \frac{(\frac{n}{2})^n}{x^x (n - x)^{n-x}}.$$

(b) The generalised likelihood ratio test rejects H_0 for small values of

$$\ln \Lambda = n \ln(n/2) - x \ln x - (n - x) \ln(n - x),$$

or equivalently, for large values of

$$x \ln x + (n - x) \ln(n - x),$$

or equivalently, for large values of

$$a(y) = (n/2 + y) \ln(n/2 + y) + (n/2 - y) \ln(n/2 - y),$$

where

$$y = |x - n/2|.$$

The function $a(y)$ is monotonely increasing over $y \in [0, n/2]$, since

$$a'(y) = \ln \frac{\frac{n}{2} + y}{\frac{n}{2} - y} > 0.$$

We conclude that the test rejects for large values of y.

(c) Compute the significance level for the rejection region $|x - \frac{n}{2}| > k$:

$$\alpha = \mathrm{P}(|X - \tfrac{n}{2}| > k|H_0) = 2 \sum_{i < \frac{n}{2} - k} \binom{n}{i} 2^{-n}.$$

(d) In particular, for $n = 10$ and $k = 2$ we get

$$\alpha = 2^{-9} \sum_{i=0}^{2} \binom{10}{i} = \frac{1 + 10 + 45}{512} = 0.11.$$

(d) Using the normal approximation for $n = 100$ and $k = 10$, we find

$$\alpha = \mathrm{P}(|X - \tfrac{n}{2}| > k|H_0) \approx 2(1 - \Phi(\tfrac{k}{\sqrt{n/4}})) = 2(1 - \Phi(2)) = 0.046.$$

Problem 9.28

Suppose that a test statistic Z has a standard normal distribution.

(a) If the test rejects for large values of $|z|$, what is the p-value corresponding to $z = 1.5$?
(b) Answer the same question if the test rejects for large values of z.

Solution 9.28

(a) Two-sided p-value = 0.134.

(b) One-sided p-value = 0.067.

Conclusion: choose H_1 before you see your data.

Problem 9.33

It has been suggested that dying people may be able to postpone their death until after an important occasion, such as a wedding or birthday. Phillips and King (1988) studied the patterns of death surrounding Passover, an important Jewish holiday.

(a) California data 1966-1984. They compared the number of deaths during the week before Passover to the number of deaths during the week after Passover for 1919 people who had Jewish surnames. Of these, 922 occurred in the week before and 997 in the week after Passover.
(b) For 852 males of Chinese and Japanese ancestry, 418 died in the week before and 434 died in the week after Passover.

Solution 9.33

We are supposed to test

H_0 : death cannot be postponed,
H_1 : death can be postponed until after an important date.

(a) Jewish data: $n = 1919$ death dates

$x = 922$ deaths during the week before Passover,
$n - x = 997$ deaths during the week after Passover.

Under the binomial model $X \sim \text{Bin}(n, p)$, we would like to test

$$H_0 : p = 0.5 \quad \text{against} \quad H_1 : p < 0.5.$$

We apply the large sample test for proportion. Observed test statistic

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{922 - 1919 \cdot 0.5}{\sqrt{1919} \cdot 0.5} = -1.712.$$

One-sided p-value of the test

$$\Phi(-1.712) = 1 - \Phi(1.712) = 1 - 0.9564 = 0.044.$$

Reject H_0 in favour of one-sided H_1 at the significance level 5%.

(b) To control for the seasonal effect the Chinese and Japanese data were studied

$$n = 852, \quad x = 418, \quad n - x = 434, \quad z = -0.548.$$

One-sided p-value is 29%, showing no significant effect.

(c) Overeating during the important occasion might be a contributing factor.

Problem 9.35

If gene frequencies are in equilibrium, the genotypes AA, Aa, and aa occur with probabilities

$$p_1 = (1-\theta)^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = \theta^2.$$

Plato et al. (1964) published the following data on haptoglobin type in a sample of 190 people

Genotype	Hp 1-1	Hp 1-2	Hp 2-2
Observed count x_i	10	68	112

Test the goodness of fit of the data to the equilibrium model.

Solution 9.35

Multinomial model

$$(X_1, X_2, X_3) \sim \text{Mn}(190, p_1, p_2, p_3).$$

Composite null hypothesis (Hardy-Weinberg Equilibrium)

$$H_0 : p_1 = (1-\theta)^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = \theta^2.$$

Likelihood function and maximum likelihood estimate

$$L(\theta) = \binom{190}{10, 68, 112} 2^{68} \theta^{292} (1-\theta)^{88}, \quad \hat{\theta} = \frac{88}{380} = 0.768.$$

Pearson's chi-square test:

cell	1	2	3	Total
observed	10	68	112	190
expected	10.23	67.71	112.07	190

Observed chi-square test statistic $\chi^2 = 0.0065$, df = 1, p-value $= 2(1 - \Phi(\sqrt{0.0065})) = 0.94$.

Conclusion: the Hardy-Weinberg Equilibrium model fits well the haptoglobin data.

Problem 9.36

US suicides in 1970. Check for the seasonal variation

Month	Number of suicides
Jan	1867
Feb	1789
Mar	1944
Apr	2094
May	2097
Jun	1981
Jul	1887
Aug	2024
Sep	1928
Oct	2032
Nov	1978
Dec	1859

Solution 9.36

Month	O_j	Days	E_j	$O_j - E_j$
Jan	1867	31	1994	-127
Feb	1789	28	1801	-12
Mar	1944	31	1994	-50
Apr	2094	30	1930	164
May	2097	31	1994	103
Jun	1981	30	1930	51
Jul	1887	31	1994	-107
Aug	2024	31	1994	30
Sep	1928	30	1930	-2
Oct	2032	31	1994	38
Nov	1978	30	1930	48
Dec	1859	31	1994	-135

Simple null hypothesis

$$H_0:\ p_1 = p_3 = p_5 = p_7 = p_8 = p_{10} = p_{12} = \frac{31}{365},\ p_2 = \frac{28}{365},\ p_4 = p_6 = p_9 = p_{11} = \frac{30}{365}.$$

The total number suicides $n = 23480$, so that the expected counts are

$$E_j = np_j^{(0)}, \quad j = 1, \ldots, 12.$$

The χ^2-test statistic

$$\chi^2 = \sum_j \frac{(O_j - E_j)^2}{E_j} = 47.4.$$

Since df $= 12 - 1 = 11$, and $\chi^2_{11}(0.005) = 26.8$, we reject H_0 of no seasonal variation. Merry Christmas!

Problem 9.43

In 1965, a newspaper carried a story about a high school student who reported getting 9207 heads and 8743 tails in 17950 coin tosses.

(a) Is this a significant discrepancy from the null hypothesis $H_0 : p = \frac{1}{2}$?a time (b) A statistician contacted the student and asked him exactly how he had performed the experiment (Youden 1974). To save time the student had tossed groups of five coins at a time, and a younger brother had recorded the results, shown in the table:

number of heads	0	1	2	3	4	5	Total
observed	100	524	1080	1126	655	105	3590

Are the data consistent with the hypothesis that all the coins were fair ($p = \frac{1}{2}$)?

(c) Are the data consistent with the hypothesis that all five coins had the same probability of heads but this probability was not necessarily $\frac{1}{2}$?

Solution 9.43

Number of heads

$$Y \sim \text{Bin}(n, p), \quad n = 17950.$$

(a) For $H_0 : p = 0.5$ the observed z-score

$$z = \tfrac{y - np_0}{\sqrt{np_0(1-p_0)}} = 3.46.$$

According to the three-sigma rule this is a significant result and we reject H_0.

(b) Pearson's chi-square test for the simple null hypothesis

$$H_0 : p_0 = (0.5)^5 = 0.031,\ p_1 = 5 \cdot (0.5)^5 = 0.156,\ p_2 = 10 \cdot (0.5)^5 = 0.313,$$
$$p_3 = 10 \cdot (0.5)^5 = 0.313,\ p_4 = 5 \cdot (0.5)^5 = 0.156,\ p_5 = (0.5)^5 = 0.031.$$

number of heads	0	1	2	3	4	5	Total
observed	100	524	1080	1126	655	105	3590
expected	112.2	560.9	1121.9	1121.9	560.9	112.2	3590

Observed $\chi^2 = 21.58$, df = 5, p-value = 0.001.

(c) Composite null hypothesis

$$H_0 : p_i = \binom{5}{i} p^i (1-p)^{5-i}, \quad i = 0, 1, 2, 3, 4, 5.$$

Pearson's chi-square test based on the maximum likelihood estimate $\hat{p} = 0.5129$

number of heads	0	1	2	3	4	5	Total
observed	100	524	1080	1126	655	105	3590
expected	98.4	518.3	1091.5	1149.3	605.1	127.4	3590

Observed $\chi^2 = 8.74$, df $= 6 - 1 - 1 = 4$, p-value = 0.07. Do not reject H_0 at 5% level.

4 Introduction to Bayesian inference

Problem 8.4 (cont)

This is a continuation of the Problem 8.4 (a)-(d).

(e) Assume uniform prior $\Theta \sim \text{U}(0, 1)$ and find the posterior density. Plot it. What is the mode of the posterior?

Solution 8.4

Since

$$f(x|\theta) \propto \theta^5 (1-\theta)^5,$$

and the prior is flat, we get

$$h(\theta|x) \propto f(x|\theta) \propto \theta^5 (1-\theta)^5.$$

We conclude that the posterior distribution is Beta $(6, 6)$. This yields

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{PME}} = \frac{1}{2}.$$

Problem 8.8

In an ecological study of the feeding behaviour of birds, the number of hops between flights was counted for several birds.

Number of hops j	1	2	3	4	5	6	7	8	9	10	11	12	Tot
Observed frequency O_j	48	31	20	9	6	5	4	2	1	1	2	1	130

Assuming that the data were generated by a Geom(p) model and take a uniform prior for p. What is then the posterior distribution and what are the posterior mean and standard deviation?

Solution 8.8

Number of bird hops $X \sim \text{Geom}(p)$

$$f(x|p) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots.$$

Data in the table summarises an iid-sample

$$(x_1, \ldots, x_n), \quad n = 130.$$

(d) Using a uniform prior $P \sim \text{U}(0,1)$, we find the posterior to be

$$h(p|x_1, \ldots, x_n) \propto f(x_1|p) \cdots f(x_n|p) = (1-p)^{n\bar{x}-n}p^n, \quad n = 130, \quad n\bar{x} = 363.$$

It is a beta distribution

$$\text{Beta}(n+1, n\bar{x}-n+1) = \text{Beta}(131, 234).$$

Posterior mean

$$\mu = \frac{a}{a+b} = \frac{131}{131+234} = 0.36.$$

Observe that

$$\mu = \frac{1+\frac{1}{n}}{\bar{x}+\frac{2}{n}},$$

gets closer to the method of moments estimate of p as $n \to \infty$. The standard deviation of the posterior distribution

$$\sigma = \sqrt{\frac{\mu(1-\mu)}{a+b+1}} = \sqrt{\frac{0.36 \cdot 0.64}{366}} = 0.025.$$

Problem 8.61

Laplace's rule of succession. Laplace claimed that when an event happens n times in a row and never fails to happen, the probability that the event will occur the next time is $\frac{n+1}{n+2}$. Can you suggest a rationale for this claim?

Solution 8.61

We use the binomial model $X \sim \text{Bin}(n,p)$, with p being the probability that the event will occur at a given trial. Use an uninformative conjugate prior $p \sim \text{Beta}(1,1)$. Given $X = n$, the posterior becomes $P \sim \text{Beta}(n+1, 1)$. Since the posterior mean is $\frac{n+1}{n+2}$, we get

$$\hat{p}_{\text{PME}} = \frac{n+1}{n+2}.$$

Problem 9.4 (cont)

Let X have one of the following two distributions

X-values	x_1	x_2	x_3	x_4
$\text{P}(x|H_0)$	0.2	0.3	0.3	0.2
$\text{P}(x|H_1)$	0.1	0.4	0.1	0.4

(c) If the prior probabilities are $\text{P}(H_0) = \text{P}(H_1) = \frac{1}{2}$, which outcomes favour H_0?

(d) What prior probabilities correspond to the decision rules with $\alpha = 0.2$ and $\alpha = 0.5$?

Solution 9.4

Recall solutions of parts (a) and (b).

(c) By Bayes formula,

$$\mathrm{P}(H_0|x) = \frac{\mathrm{P}(x|H_0)\mathrm{P}(H_0)}{\mathrm{P}(x|H_0)\mathrm{P}(H_0) + \mathrm{P}(x|H_1)\mathrm{P}(H_1)} = \frac{\mathrm{P}(x|H_0)}{\mathrm{P}(x|H_0) + \mathrm{P}(x|H_1)}.$$

Thus the posterior odds ratio equals the likelihood ratio

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} = \Lambda,$$

and we conclude that outcomes x_1 and x_3 favour H_0 since with these outcomes $\Lambda > 1$.

(d) For the general prior

$$\mathrm{P}(H_0) = \pi_0, \quad \mathrm{P}(H_1) = \pi_1 = 1 - \pi_0,$$

we get

$$\mathrm{P}(H_i|x) = \frac{\mathrm{P}(x|H_i)\pi_i}{\mathrm{P}(x|H_0)\pi_0 + \mathrm{P}(x|H_1)\pi_1},$$

yielding a relation for the posterior odds ratio

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} = \frac{\mathrm{P}(x|H_0)\pi_0}{\mathrm{P}(x|H_1)\pi_1} = \Lambda \cdot \frac{\pi_0}{\pi_1}.$$

Assuming equal costs $c_0 = c_1$, the rejection rule is

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} < \frac{c_1}{c_0} = 1,$$

so that in terms of the likelihood ratio,

$$\Lambda < \frac{\pi_1}{\pi_0} = \frac{1}{\pi_0} - 1, \qquad \pi_0 < \frac{1}{1+\Lambda}.$$

If $x = x_4$, then $\Lambda = 0.5$, and we reject H_0, provided $\pi_0 < \frac{2}{3}$. This corresponds to the decision rules with $\alpha = 0.2$.

If $x = x_2$, then $\Lambda = 0.75$, and we reject H_0, provided $\pi_0 < \frac{4}{7}$. This corresponds to the decision rules with $\alpha = 0.5$.

Furthermore, if $x = x_1$, then $\Lambda = 2$, and we reject H_0, provided $\pi_0 < \frac{1}{3}$, and if $x = x_3$, then $\Lambda = 3$, and we reject H_0, provided $\pi_0 < \frac{1}{4}$.

Problem 9.14

Suppose that under H_0, a measurement X is $\mathrm{N}(0, \sigma^2)$, and under H_1, the measurement X is $\mathrm{N}(1, \sigma^2)$. Assume that the prior probabilities satisfy

$$\mathrm{P}(H_0) = 2\mathrm{P}(H_1).$$

The hypothesis H_0 will be chosen if $\mathrm{P}(H_0|x) > \mathrm{P}(H_1|x)$. For $\sigma^2 = 0.1, 0.5, 1.0, 5.0$:

(a) For what values of $X = x$ will H_0 be chosen?
(b) In the long run, what proportion of the time will H_0 be chosen if H_0 is true $\frac{2}{3}$ of the time?

Solution 9.14

For a single observation $X \sim \mathrm{N}(\mu, \sigma^2)$, where σ^2 is known, test $H_0 : \mu = 0$ vs $H_1 : \mu = 1$. Prior probabilities

$$\mathrm{P}(H_0) = \frac{2}{3}, \quad \mathrm{P}(H_1) = \frac{1}{3}.$$

(a) Likelihood ratio

$$\frac{f(x|0)}{f(x|1)} = \frac{e^{-\frac{x^2}{2\sigma^2}}}{e^{-\frac{(x-1)^2}{2\sigma^2}}} = e^{\frac{\frac{1}{2}-x}{\sigma^2}}.$$

Choose H_0 for x such that

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} = 2e^{\frac{\frac{1}{2}-x}{\sigma^2}} > 1, \qquad x < \frac{1}{2} + \sigma^2 \ln 2.$$

We conclude that

	$\sigma^2 = 0.1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 5$
Choose H_0 for	$x < 0.57$	$x < 0.85$	$x < 1.19$	$x < 3.97$

(b) In the long run, the proportion of the time H_0 will be chosen is

$$\mathrm{P}(X < \tfrac{1}{2}+\sigma^2 \ln 2) = \tfrac{2}{3}\mathrm{P}(X-\mu < \tfrac{1}{2}+\sigma^2 \ln 2)+\tfrac{1}{3}\mathrm{P}(X-\mu < \sigma^2 \ln 2-\tfrac{1}{2}) = \tfrac{2}{3}\Phi(\sigma \ln 2+\tfrac{1}{2\sigma})+\tfrac{1}{3}\Phi(\sigma \ln 2-\tfrac{1}{2\sigma}).$$

We conclude that

	$\sigma^2 = 0.1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 5$
Proportion of the time H_0 will be chosen	0.67	0.73	0.78	0.94

5 Summarising data

Problem 10.2

Suppose that $(X_1, \ldots, X_n)$ are independent uniform $\mathrm{U}(0, 1)$ random variables.

(a) Sketch the population distribution function $F(x)$ and the standard deviation of the empirical distribution function $\hat{F}(x)$.

(b) Generate many samples of size 16. For each sample plot $\hat{F}(x) - F(x)$ and relate what you see to your answer to (a).

Solution 10.2

Recall that for a fixed x, the empirical distribution function $\hat{F}(x)$ is the sample proportion estimate of $p = F(x) = x$.

(a) The variance of $\hat{F}(x)$ is

$$\sigma^2_{\hat{F}(x)} = \frac{p(1-p)}{n} = \frac{x(1-x)}{n},$$

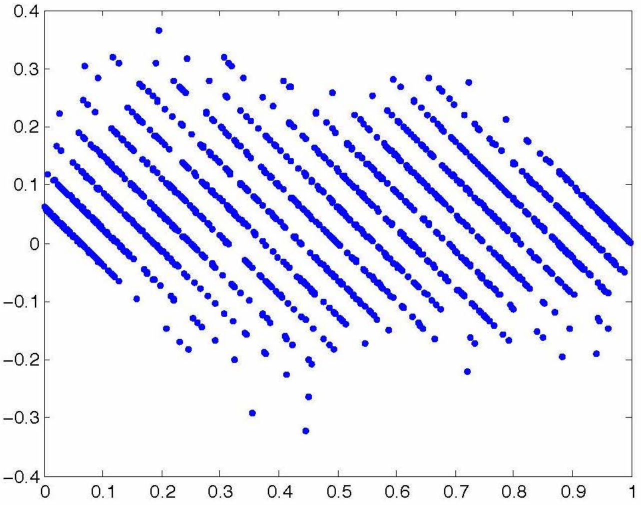
so that the standard deviation is

$$\sigma_{\hat{F}(x)} = \sqrt{\frac{x(1-x)}{n}}, \quad x \in [0, 1].$$

(b) Generate 100 samples of size $n = 16$. Matlab code

```
x=rand(16,100);
y=sort(x)’;
for k=1:100
plot(y(k,:),(1:16)/16-y(k,:),’.’)
hold on
end
```

See the figure



Problem 10.5

Let $(X_1, \dots, X_n)$ be independent random variables with the same distribution F, and let $\hat{F}$ denote the empirical distribution function. Show that for $u < v$,

$$\mathrm{Cov}(\hat{F}(u), \hat{F}(v)) = \tfrac{1}{n} F(u)(1 - F(v)).$$

If follows that $\hat{F}(u)$ and $\hat{F}(v)$ are positively correlated: if $\hat{F}(u)$ overshoots $F(u)$, then $\hat{F}(v)$ will tend to overshoot $F(v)$.

Solution 10.5

We have

$$\begin{aligned} \hat{F}(u) &= \frac{1_{\{X_1 \le u\}} + \ldots + 1_{\{X_n \le u\}}}{n}, \quad \mathrm{E}(\hat{F}(u)) = F(u), \\ \hat{F}(v) &= \frac{1_{\{X_1 \le v\}} + \ldots + 1_{\{X_n \le v\}}}{n}, \quad \mathrm{E}(\hat{F}(v)) = F(v). \end{aligned}$$

Assuming $u < v$, we get

$$\begin{aligned} \mathrm{E}(\hat{F}(u) \cdot \hat{F}(v)) &= \frac{1}{n^2} \Big[\sum_{i=1}^n \mathrm{E}(1_{\{X_i \le u\}} 1_{\{X_i \le v\}}) + \sum_{i=1}^n \sum_{j \ne i} \mathrm{E}(1_{\{X_i \le u\}} 1_{\{X_j \le v\}}) \Big] \\ &= \frac{1}{n^2} \Big[\sum_{i=1}^n F(u) + \sum_{i=1}^n \sum_{j \ne i} F(u)F(v) \Big] \\ &= \frac{1}{n} \Big[F(u) + (n-1)F(u)F(v) \Big]. \end{aligned}$$

Finish by using

$$\begin{aligned} \mathrm{Cov}(\hat{F}(u), \hat{F}(v)) &= \mathrm{E}(\hat{F}(u) \cdot \hat{F}(v)) - \mathrm{E}(\hat{F}(u)) \cdot \mathrm{E}(\hat{F}(v)) \\ &= \frac{1}{n} [F(u) + (n-1)F(u)F(v)] - F(u)F(v) \\ &= \tfrac{1}{n} F(u)(1 - F(v)). \end{aligned}$$

Problem 10.6

A random sample $x_1, \ldots, x_n$, $n = 59$:

14.27 14.80 12.28 17.09 15.10 12.92 15.56 15.38 15.15 13.98
14.90 15.91 14.52 15.63 13.83 13.66 13.98 14.47 14.65 14.73
15.18 14.49 14.56 15.03 15.40 14.68 13.33 14.41 14.19 15.21
14.75 14.41 14.04 13.68 15.31 14.32 13.64 14.77 14.30 14.62
14.10 15.47 13.73 13.65 15.02 14.01 14.92 15.47 13.75 14.87
15.28 14.43 13.96 14.57 15.49 15.13 14.23 14.44 14.57

are the percentages of hydrocarbons in each sample of beeswax.

(a) Plot the empirical distribution function, a histogram, and a normal probability plot. Find the 0.9, 0.75, 0.5, 0.25, and 0.1 quantiles. Does the distribution appear Gaussian?

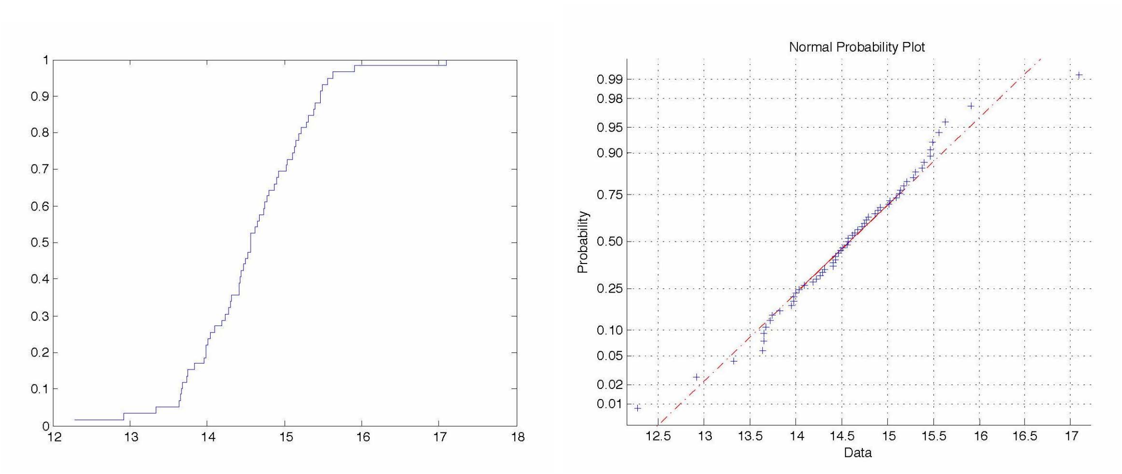
(b) The average percentage of hydrocarbons in a synthetic wax is 85%. Suppose that beeswax was diluted with 1% synthetic wax. Could this be detected? What about 3% and 5% dilution?

Solution 10.6

Ordered sample $x_{(1)}, \ldots, x_{(n)}$

12.28 12.92 13.33 13.64 13.65 13.66 13.68
13.73 13.75 13.83 13.96 13.98 13.98 14.01
14.04 25% quantile
14.10 14.19 14.23 14.27 14.30 14.32 14.41
14.41 14.43 14.44 14.47 14.49 14.52 14.56
14.57 50% quantile
14.57 14.62 14.65 14.68 14.73 14.75 14.77
14.80 14.87 14.90 14.92 15.02 15.03 15.10
15.13 75% quantile
15.15 15.18 15.21 15.28 15.31 15.38 15.40
15.47 15.47 15.49 15.56 15.63 15.91 17.09

(a) The figure shows the empirical distribution function and a normal probability plot.



Use Matlab commands

```
x=data vector;
stairs(sort(x),(1:length(x))/length(x))   % empirical cdf
hist(x)                                   % histogram, the same as hist(x,10)
normplot(x)                               % normal probability plot
prctile(x,90)                             % 0.90-quantile
```

The distribution appears to be rather close to normal. The 10% quantile

$$\frac{x_{(6)} + x_{(7)}}{2} = \frac{13.66 + 13.68}{2} = 13.67.$$

(b) Since $\bar{x} = 14.58$ and $s = 0.78$, the one-sided 99% of the population distribution for the natural wax is

$$(-\infty, 14.58 + 2.33 \cdot 0.78) = (-\infty, 16.40).$$

Expected means

1% dilution	$\mu_1 = 14.58 \cdot 0.99 + 85 \cdot 0.01 = 15.28$	can not be detected
3% dilution	$\mu_3 = 14.58 \cdot 0.97 + 85 \cdot 0.03 = 16.69$	can be detected
5% dilution	$\mu_5 = 14.58 \cdot 0.95 + 85 \cdot 0.05 = 18.10$	can be detected

Problem 10.11

Calculate the hazard function for the Weibull distribution (Waloddi Weibull was a Swedish engineer, scientist, and mathematician)

$$F(t) = 1 - e^{-\alpha t^{\beta}}, \quad t \geq 0,$$

where α and β are two positive parameters.

Solution 10.11

Taking the derivative of

$$1 - F(t) = e^{-\alpha t^{\beta}},$$

we find the density

$$f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}},$$

and dividing the latter by the former we obtain the hazard function

$$h(t) = \alpha\beta t^{\beta-1}.$$

Problem 10.13 and 10.14

Give an example of a distribution with an increasing failure rate. Give an example of a distribution with a decreasing failure rate.

Solution 10.13 and 10.14

Take the Weibull distribution with parameters α and β.

- If $\beta = 1$, then $h(t) = \alpha$ is constant and the distribution is memoryless.
- If $\beta > 1$, then $h(t)$ increases with t meaning that the older individuals die more often than the younger.
- If $0 < \beta < 1$, then $h(t)$ decreases with t meaning that the longer you live the healthier you become.

Problem 10.29

Of the 26 measurements of the heat of sublimation of platinum, 5 are outliers.

136.3 136.6 135.8 135.4 134.7 135 134.1 143.3 147.8 148.8
134.8 135.2 134.9 146.5 141.2 135.4 134.8 135.8 135 133.7
134.4 134.9 134.8 134.5 134.3 135.2

Stem and leaf display for $n = 26$ observations including $k = 5$ outliers:

133:7
134:134
134:5788899
135:002244
135:88
136:36
High: 141.2, 143.3, 146.5, 147.8, 148.8

Let N be the number of outliers in a non-parametric bootstrap sample.

(a) Explain why the distribution of N is binomial.
(b) Find $\mathrm{P}(N \geq 10)$.
(c) In 1000 bootstrap samples, how many would you expect to contain 10 or more of theses outliers.
(d) What is the probability that a bootstrap sample is composed entirely of these outliers?

Solution 10.29

(a) Due to sampling with replacement we have $N \sim \text{Bin}\ (26, \frac{5}{26})$.

(b) Using the binomial distribution command we find

$$\mathrm{P}(N \leq 9) = \text{binocdf}(9, 26, 5/26) = 0.9821,$$
$$\mathrm{P}(N \geq 10) = 1 - 0.9821 = 0.018.$$

(c) In $B = 1000$ bootstrap samples, we expect

$$B \cdot \mathrm{P}(N \geq 10) = 18$$

samples to contain 10 or more of outliers.

(d) The probability that a bootstrap sample is composed entirely of these outliers is negligibly small

$$\mathrm{P}(N = 25) = (5/26)^{26} = 2.4 \cdot 10^{-19}.$$

Problem 10.37

For the data in Problem 10.6.

(a) Find the mean, median, and 10% and 20% trimmed means.
(b) Find an approximate 90% confidence interval for the mean.
(c) Find a confidence interval with coverage near 90% for the median.
(d) Use the bootstrap to find a pproximate standard errors of the trimmed means.
(f) Find and compare the standard deviation of the measurements, the interquartile range, and the MAD.
(g) Use the bootstrap to find the approximate sampling distribution and standard error of the upper quartile.

Solution 10.37

(a) The Matlab commands

```
trimmean(x,10)
trimmean(x,20)
```

give $\bar{X}_{0.1} = 14.586$ and $\bar{X}_{0.2} = 14.605$.

> m = trimmean(X,percent) calculates the trimmed mean of the values in X. For a vector input, m is the mean of X, excluding the highest and lowest k data values, where k=n*(percent/100)/2 and where n is the number of values in X.

(b) An approximate 90% confidence interval for the mean is

$$I_\mu = 14.58 \pm 1.645 \cdot \frac{0.78}{\sqrt{59}} = 14.58 \pm 0.17 = (14.41; 14.75)$$

(c) Nonparametric 90% confidence interval for the population median M is $(X_{(k)}, X_{(60-k)})$, where $\mathrm{P}(Y < k) = 0.05$ and $Y \sim$ Bin $(59, 0.5)$. Applying the normal approximation for Bin (n,p) with continuity correction

$$\mathrm{P}(Y < k) = \mathrm{P}(Y \leq k-1) \approx \Phi\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}}\right),$$

we arrive at equation

$$\frac{k - 0.5 - \frac{59}{2}}{\sqrt{\frac{59}{4}}} = -1.645,$$

that gives $k = 24$. This yields

$$I_m = (x_{(k)}, x_{(60-k)}) = (x_{(24)}, x_{(36)}) = (14.43; 14.75).$$

(d) The Matlab commands for the non-parametric bootstrap

```
n=59; B=1000;
z=x(random('unid',n,n,B));        % ('unid',n) - uniform discrete [1, n], 1000 samples of size n
t1=trimmean(z,10);
t2=trimmean(z,20);
std(t1)
std(t2)
```

give the standard errors 0.1034 and 0.1004 for $\bar{x}_{0.1}$ and $\bar{x}_{0.2}$ respectively.

(f) Matlab commands

```
iqr(x)
median(abs(x-median(x)))
```

Warning: mad(x) in Matlab stands for the mean abs. dev.

(g) Matlab commands (vector z comes from the (d) part)

```
q=prctile(z,75);
hist(q)
std(q)
```

give the standard error 0.1332 of the upper quartile.

Problem 10.40

Olson, Simpson, and Eden (1975) discuss the analysis of data obtained from a cloud seeding experiment. The following data present the rainfall from 26 seeded and 26 control clouds.

> Seeded clouds
> 129.6, 31.4, 2745.6, 489.1, 430, 302.8, 119, 4.1, 92.4, 17.5,
> 200.7, 274.7, 274.7, 7.7, 1656, 978, 198.6, 703.4, 1697.8, 334.1,
> 118.3, 255, 115.3, 242.5, 32.7, 40.6
> Control clouds
> 26.1, 26.3, 87, 95, 372.4, .01, 17.3, 24.4, 11.5, 321.2,
> 68.5, 81.5, 47.3, 28.6, 830.1, 345.5, 1202.6, 36.6, 4.9, 4.9,
> 41.1, 29, 163, 244.3, 147.8, 21.7

Make a QQ-plot for rainfall versus rainfall and log rainfall versus log rainfall. What do these plots suggest about the effect, if any, of seeding?

Solution 10.40

Matlab command (x = control and y = seeded data)

```
qqplot(x,y)
```

produces a QQ-plot that fits the line $y = 2.5x$ claiming 2.5 times more rainfall from seeded clouds. On the other hand, Matlab command

```
qqplot(log(x),log(y))
```

produces a QQ-plot that fits the line

$$\ln y = 2 + 0.8 \ln x$$

meaning a decreasing slope in the relationship $y = 7.4x^{0.8}$.

6 Comparing two samples

Problem 11.1

Four random numbers generated from a normal distribution

$$x_1 = 1.1650, \quad x_2 = 0.6268, \quad x_3 = 0.0751, \quad x_4 = 0.3516,$$

along with five random numbers with the same variance σ^2 but perhaps a different mean

$$y_1 = 0.3035, \quad y_2 = 2.6961, \quad y_3 = 1.0591, \quad y_4 = 2.7971, \quad y_5 = 1.2641.$$

(a) What do you think the means of the random normal number generators were? What do you think the difference of the means was?

(b) What do you think the variance of the random number generator was?

(c) What is the estimated standard error of your estimate of the difference of the means?

(d) Form a 90% confidence interval for the difference of the means.

(e) In this situation, is it more appropriate to use a one-sided test or a two-sided test of the equality of the means?

(f) What is the p-value of a two-sided test of the null hypothesis of equal means?

(g) Would the hypothesis that the means were the same versus a two-sided alternative be rejected at the significance level $\alpha = 0.1$?

(h) Suppose you know that the variance of the normal distribution was $\sigma^2 = 1$. How would your answers to the preceding questions change?

Solution 11.1

(a) $\bar{x} = 0.5546$, $\bar{y} = 1.6240$, $\bar{y} - \bar{x} = 1.0694$

(b) We have $s_x^2 = 0.2163$, $s_y^2 = 1.1795$, $s_{\text{p}}^2 = 0.7667$. The latter is an unbiased estimate of σ^2.

(c) $s_{\bar{y}-\bar{x}} = 0.5874$

(d) Based on t_7-distribution, an exact 90% confidence interval for mean difference is

$$I_{\mu_y - \mu_x} = 1.0694 \pm 1.1128.$$

(e) More appropriate to use a two-sided test.

(f) From the observed test statistic value $t = 1.8206$, we find the two-sided p $= 0.1115$ using the Matlab command 2*tcdf(-1.8206,7).

(g) No, because the obtained p-value is larger than 0.1.

(h) Given $\sigma^2 = 1$, we answer differently to some of the the above questions:

b: $\sigma^2 = 1$,
c: $s_{\bar{y}-\bar{x}} = 0.0.6708$,
d: $I_{\mu_y - \mu_x} = 1.0694 \pm 1.1035$,
f: $z = 1.5942$ two-sided p-value = 0.11.

Problem 11.3

In the "two independent samples" setting we have two ways of estimating the variance of $\bar{X} - \bar{Y}$:

(a) $s_{\text{p}}^2(\frac{1}{n} + \frac{1}{m})$, if $\sigma_x = \sigma_y$,
(b) $\frac{s_x^2}{n} + \frac{s_y^2}{m}$ without the assumption of equal variances.

Show that if $m = n$, then these two estimates are identical.

Solution 11.3

If $m = n$, then

$$s_{\text{p}}^2\left(\frac{1}{n} + \frac{1}{m}\right) = \frac{2}{n} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{2n-2} = \frac{s_x^2 + s_y^2}{n} = \frac{s_x^2}{n} + \frac{s_y^2}{m}.$$

Problem 11.8

An experiment of the efficacy of a drug for reducing high blood pressure is performed using four subjects in the following way:

> two of the subjects are chosen at random for the control group and two for the treatment group.

During the course of a treatment with the drug, the blood pressure of each of the subjects in the teatment group is measured for ten consecutive days as is the blood pressure of each og the subjects in the control group.

(a) In order to test whether the treatment has an effect, do you think it is appropriate to use the two-sample t test with $n = m = 20$?
(b) Do you think it is appropriate to use the rank sum test?

Solution 11.8

Test the null hypothesis of no drug effect

$H_0 : \mu_1 = \mu_2$, the drug is not effective for reducing high blood pressure.

Suggested measurement design: during the same $n = 10$ days take blood pressure measurements on 4 people, two on the treatment

$$x_{11}, \ldots, x_{1n},$$
$$x_{21}, \ldots, x_{2n},$$

and two controls

$$x_{31}, \ldots, x_{3n},$$
$$x_{41}, \ldots, x_{4n}.$$

Dependencies across the days and the people make inappropriate both two-sample t test and rank sum test. Proper design for 40 measurements is that of two independent samples: 20 people on the treatment and 20 controls:

$$x_1, \ldots, x_{20},$$
$$y_1, \ldots, y_{20}.$$

Problem 11.13

Let $x_1, \ldots, x_{25}$ be an iid-sample drawn from $\mathrm{N}(0.3, 1)$. Consider testing at $\alpha = 0.05$

$$H_0 : \mu = 0, \qquad H_1 : \mu > 0.$$

Compare
(a) the power of the sign test , and
(b) the power of the test based on the normal theory assuming that σ is known.

Solution 11.13

(a) The sign test statistic

$$t = \text{number of positive } x_i, \qquad T \overset{H_0}{\sim} \mathrm{Bin}(25, \tfrac{1}{2}) \approx \mathrm{N}(\tfrac{25}{2}, \tfrac{25}{4}).$$

Reject H_0 for $t \geq k$, where k is found from

$$0.05 = \mathrm{P}(T \geq k|H_0) = \mathrm{P}(T > k-1|H_0) \approx 1 - \Phi\left(\frac{k - 0.5 - 12.5}{5/2}\right) = 1 - \Phi\left(\frac{k-13}{2.5}\right),$$

which gives

$$\frac{k-13}{2.5} = 1.645, \qquad k = 17.$$

We know the true population distribution is $\mathrm{N}(0.3, 1)$. Since

$$\mathrm{P}(X > 0|\mathrm{N}(0.3, 1)) = 1 - \Phi(-0.3) = \Phi(0.3) = 0.62,$$

we can use

$$T \sim \mathrm{Bin}(25, 0.62) \approx \mathrm{N}(15.5, 5.89)$$

to find the power of the sign test by

$$1 - \beta = \mathrm{P}(T \geq 17) \approx 1 - \Phi\left(\frac{17 - 0.5 - 15.5}{2.4}\right) = 1 - \Phi(0.41) = 0.34.$$

(b) Normal distribution model $X \sim \mathrm{N}(\mu, 1)$. Since $\frac{\bar{X}-\mu}{1/5} \sim \mathrm{N}(0,1)$, we reject H_0 for

$$5\bar{x} > 1.645, \quad \text{that is for } \ \bar{x} > 0.33.$$

The power of the test

$$1 - \beta = \mathrm{P}(\bar{X} > 0.33 | \mu = 0.3) = 1 - \Phi\left(\frac{0.33 - 0.3}{1/5}\right) = 1 - \Phi(0.15) = 0.44$$

is higher than the power of the sign test.

Problem 11.15

Suppose that n measurements are to be taken under a treatment condition and another n measurements are to be taken independently under a control condition. It is thought that the standard deviation of a single observation is about 10 under both conditions. How large should n be so that a 95% confidence interval for the mean difference has a width of 2? Use the normal distribution rather than the t-distribution, since n will turn out to be rather large.

Solution 11.15

Two independent samples

$$x_1, \ldots, x_n, \quad y_1, \ldots, y_n,$$

are taken from two population distributions with equal standard deviation $\sigma = 10$. Approximate 95% confidence interval

$$I_{\mu_1 - \mu_2} = \bar{x} - \bar{y} \pm 1.96 \cdot 10 \cdot \sqrt{\tfrac{2}{n}} = \bar{x} - \bar{y} \pm \tfrac{27.72}{\sqrt{n}}.$$

If the confidence interval has width 2, then

$$\tfrac{27.72}{\sqrt{n}} = 1,$$

implying $n \approx 768$.

Problem 11.21

Data: millions of cycles until failure for two types of engine bearings

Type I	Type II
3.03	3.19
5.53	4.26
5.60	4.47
9.30	4.53
9.92	4.67
12.51	4.69
12.95	6.79
15.21	9.37
16.04	12.75
16.84	12.78

(a) Use normal theory to test the null hypothesis of no difference against the two-sided alternative

$$H_0 : \mu_x = \mu_y, \qquad H_1 : \mu_x \neq \mu_y.$$

(b) Test the hypothesis that there is no difference between the two types of bearing using a nonparametric method.

(c) Which of the methods (a) or (b) do you think is better in this case?

(d) Estimate π, the probability that a type I bearing will outlast a type II bearing.

(e) Use the bootstrap to estimate the sampling distribution of $\hat{\pi}$ and its standard error.

Solution 11.21

	Rank	Type I	Type II	Rank
	1	3.03	3.19	2
	8	5.53	4.26	3
	9	5.60	4.47	4
	11	9.30	4.53	5
	13	9.92	4.67	6
	14	12.51	4.69	7
	17	12.95	6.79	10
	18	15.21	9.37	12
	19	16.04	12.75	15
	20	16.84	12.78	16
Rank sum	130			80

(a) Two-sample t-test

$$\bar{x} = 10.693, \quad \bar{y} = 6.750, \quad s_x^2 = 23.226, \quad s_y^2 = 12.978, \quad s_{\bar{x}-\bar{y}} = \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2} = 1.903.$$

Assume equal variances. The observed test statistic

$$t = \frac{10.693 - 6.750}{1.903} = 2.072.$$

With df = 18, the two-sided p-value = 0.053 is found using the Matlab command 2*tcdf(-2.072,18).

(b) Wilcoxon rank sum test statistics $R_x = 130$, $R_y = 80$. From the table on page A22 we find that the two-sided p-value is between $0.05 <$ p-value < 0.10.

(c) The non-parametric test in (b) is more relevant, since both normplot(x) and normplot(y) show non-normality of the data distribution.

(d) To estimate the probability π, that a type I bearing will outlast a type II bearing, we turn to the ordered pooled sample

X-YYYYYY-XX-Y-X-Y-XX-YY-XXXX.

Pick a pair (X, Y) at random, then by the division rule of probability

$$\mathrm{P}(X < Y) = \frac{\text{number of } (x_i < y_j)}{\text{total number of pairs } (x_i, y_j)} = \frac{10 + 4 + 4 + 3 + 2 + 2}{100} = 0.25.$$

This implies a point estimate $\hat{\pi} = 0.75$.

(e) The matlab commands

```
u=x(random('unid',10,10,1000));
v=y(random('unid',10,10,1000));
N=zeros(1,1000);
for k=1:1000 for i=1:10 for j=1:10
N(k)=N(k)+(u(i,k)>v(j,k));
end,end,end
P=N/100;
hist(P,20)
std(P)
```

estimate the sampling distribution of $\hat{\pi}$ with $s_{\hat{\pi}} = 0.1187$.

Problem 11.27

Find the exact null distribution for the test statistic of the signed rank test with $n = 4$.

Solution 11.27

Model: iid-sample of the differences $d_1, \ldots, d_n$ whose population distribution is symmetric around the unknown median m. Test the null hypothesis of no difference $H_0 : m = 0$ using the signed ranks defined as follows:

step 1: remove signs $|d_1|, \ldots, |d_n|$,
step 2: assign ranks $1, \ldots, n$ to $|d_1|, \ldots, |d_n|$,
step 3: attach accordingly the original signs to the ranks $1, \ldots, n$,
step 4: compute w_+ as the sum of the positive ranks.

Under $H_0 : m = 0$, on the step 4, the signs ± are assigned symmetrically at random. There are 16 equally likely outcomes

1	2	3	4	w_+
−	−	−	−	0
+	−	−	−	1
−	+	−	−	2
+	+	−	−	3
−	−	+	−	3
+	−	+	−	4
−	+	+	−	5
+	+	+	−	6
−	−	−	+	4
+	−	−	+	5
−	+	−	+	6
+	+	−	+	7
−	−	+	+	7
+	−	+	+	8
−	+	+	+	9
+	+	+	+	10

Thus the null distribution of W_+ is given by the table

k	0	1	2	3	4	5	6	7	8	9	10
p_k	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

The smallest one-sided p-value is $\frac{1}{16} = 0.06$ which is higher than 5%. Thus $n = 4$ is too small sample size.

Problem 11.28

Turn to the two-sided signed rank test. For $n = 10, 20, 25$ and $\alpha = 0.05, 0.01$, compare the critical values from the table and using the normal approximation of the null distribution.

Solution 11.28

Using

$$W_{0.05}(n) = \frac{n(n+1)}{4} - 1.96 \cdot \sqrt{\frac{n(n+1)(2n+1)}{24}},$$
$$W_{0.01}(n) = \frac{n(n+1)}{4} - 2.58 \cdot \sqrt{\frac{n(n+1)(2n+1)}{24}},$$

we find (table/normal approximation)

	$n=10$	$n=20$	$n=25$
$\frac{n(n+1)}{4}$	27.5	105	162.5
$\sqrt{\frac{n(n+1)(2n+1)}{24}}$	9.81	26.79	37.17
$\alpha=0.05$	8/8.3	52/53.5	89/89.65
$\alpha=0.01$	3/2.2	38/36.0	68/67.6

Problem 11.34

Two population distributions with $\sigma_x = \sigma_y = 10$. Two samples of sizes $n = 25$ can be taken in two ways

(a) paired with $\mathrm{Cov}(X_i, Y_i) = 50$, $i = 1, \ldots, 25$,
(b) unpaired $x_1, \ldots, x_{25}$ and $y_1, \ldots, y_{25}$.

Compare the power curves for testing

$$H_0 : \mu_x = \mu_y, \qquad H_1 : \mu_x > \mu_y, \qquad \alpha = 0.05.$$

Solution 11.34

(a) The variance of a difference

$$\mathrm{Var}(D) = \mathrm{Var}(X - Y) = \sigma_x^2 + \sigma_y^2 - 2\mathrm{Cov}(X, Y) = 100 + 100 - 100 = 100.$$

Using the normal approximation we get

$$\bar{D} = \bar{X} - \bar{Y} \approx \mathrm{N}(\mu_x - \mu_y, \tfrac{100}{25}) = \mathrm{N}(\delta, 4).$$

The rejection region becomes

$$\mathcal{R} = \{\tfrac{\bar{d}}{2} > 1.645\} = \{\bar{d} > 3.29\}.$$

The power function

$$\mathrm{Pw}(\delta) = \mathrm{P}(\bar{D} > 3.29|\mathrm{N}(\delta, 4)) \approx 1 - \Phi(\tfrac{3.29-\delta}{2}).$$

(b) Two independent samples

$$\bar{D} \approx \mathrm{N}(\mu_x - \mu_y, \tfrac{100}{25} + \tfrac{100}{25}) = \mathrm{N}(\delta, 8).$$

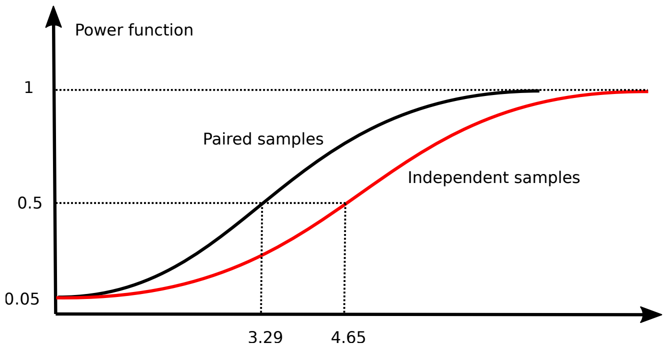
The rejection region

$$\mathcal{R} = \{\tfrac{\bar{d}}{\sqrt{8}} > 1.645\} = \{\bar{d} > 4.65\}.$$

The power function

$$\mathrm{Pw}(\delta) = \mathrm{P}(\bar{D} > 4.65|\mathrm{N}(\delta, 8)) \approx 1 - \Phi(\tfrac{4.65-\delta}{2.83}).$$

The two power functions are compared graphically on the next figure.



Problem 11.36

Lin, Sutton, and Qurashi (1979) compared microbiological and hydroxylamine methods for the analysis of ampicillin dosages. In one series of experiments, pairs of tablets were analysed by the two methods. The data in the table give the percentages of claimed amount of ampicillin found by the two methods in several pairs of tablets.

Microbiological method	Hydroxylamine method
97.2	97.2
105.8	97.8
99.5	96.2
100	101.8
93.8	88
79.2	74
72	75
72	67.5
69.5	65.8
20.5	21.2
95.2	94.8
90.8	95.8
96.2	98
96.2	99
91	100.2

What are $\bar{x} - \bar{y}$ and $s_{\bar{x}-\bar{y}}$? If the pairing had been erroneously ignored and it had been assumed that the two samples were independent, what would have been the estimate of the standard deviation of $\bar{X} - \bar{Y}$? Analyse the data to determine if there is a systematic difference between the two methods.

Solution 11.36

Paired samples

$$\begin{aligned}
\bar{x} &= 85.26, \quad s_x = 21.20, \quad s_{\bar{x}} = 5.47, \quad n_x = 15,\\
\bar{y} &= 84.82, \quad s_y = 21.55, \quad s_{\bar{y}} = 5.57, \quad n_y = 15,\\
\bar{d} &= \bar{x} - \bar{y} = 0.44,\\
s_d &= 4.63, \quad s_{\bar{x}-\bar{y}} = 1.20.
\end{aligned}$$

If the pairing had been erroneously ignored, then the two independent samples formula would give 6 times larger standard error

$$s_{\bar{x}-\bar{y}} = 7.81.$$

To test $H_0 : \mu_x = \mu_y$ against $H_1 : \mu_x \neq \mu_y$ assume $D \sim \mathrm{N}(\mu, \sigma^2)$ and apply one-sample t-test

$$t = \frac{\bar{d}}{s_{\bar{d}}} = 0.368.$$

With df = 14, two-sided p-value = 0.718, we can not reject H_0.

Without normality assumption we apply the signed rank test. Matlab command

```
signrank(x,y)
```

computes the two-sided p-value = 0.604. We can not reject H_0.

Problem 11.52

The media often present short reports of the results of experiments. To the critical reader, such reports often raise more questions than they answer. Comment on the following pitfalls in the interpretation of each of the following.

(a) It is reported that patients whose hospital rooms have a window recover faster than those whose rooms do not.

(b) Nonsmoking wives whose husbands whose husbands smoke have a cancer rate twice that of wives whose husbands do not smoke.

(c) A two-year study in North Carolina found that 75% of all industrial accidents in the state happened to workers who had skipped breakfast.

(d) A school integration program involved busing children from minority schools to majority (primarily white) schools. Participation in the program was voluntary. It was found that the students who were bused scored lower on standardised tests than did their peers who chose not to be bused.

(e) When a group of students were asked to match pictures of newborns and with pictures of their mothers, they were correct 36% of the time.

(f) A survey found that that those who drank a moderate amount of beer were healthier than those who totally abstained from alcohol.

(g) A 15-year study of more than 45 000 Swedish soldiers revealed that heavy users of marijuana were six times more likely than nonusers to develop schizophrenia.

(h) A University of Wisconsin study showed that within 10 years of wedding, 38% of those who had lived together before marriage had split up, compared to 27% of those who had married without a ”trial period”.

(i) A study of nearly 4000 elderly North Carolinians has found that those who attended religious services every week were 46% less likely to die over a six-year period than people who attended less often or not at all.

Solution 11.52

Possible explanations

(a) room with a window ← $\boxed{\text{rich patient}}$ → recovers faster,

(b) besides passive smoking: smoker ← $\boxed{\text{the man is a bad husband}}$ → wife gets cancer,

(c) no breakfast ← $\boxed{\text{more stress}}$ → accident,

(d) choose to change the school and to be bused ← $\boxed{\text{lower grades before}}$ → lower grades after,

(e) match two babies with two mothers (or even 3 babies with 3 mothers) then it is pure chance,

(f) abstain from alcohol ← poor health,

(g) marijuana ← schizophrenia,

(h) total time together = time before wedding + time after wedding,

(i) being part of a community can have a positive effect on mental health and emotional wellbeing.

7 Analysis of variance

Useful Matlab commands:

```
boxplot(x)
anova1(x)
anova2(x)
```

where x is the data matrix.

Problem 12.3

For a one-way analysis of variance with two treatment groups, show that the F statistic is t^2, where t is the test statistic for a two-sample t-test.

Solution 12.3

Consider one-way ANOVA test statistic

$$F = \frac{MS_A}{MS_E} = \frac{\frac{J}{I-1}\sum_{i=1}^{I}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\frac{1}{I(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{i\cdot})^2}$$

For $I = 2$ and $J = n$, put

$$\bar{y}_{1\cdot} = \bar{x}, \quad \bar{y}_{2\cdot} = \bar{y}, \quad \bar{y}_{\cdot\cdot} = \frac{\bar{x}+\bar{y}}{2}.$$

In this two-sample setting, the F-test statistic becomes

$$F = \frac{n[(\bar{x} - \frac{\bar{x}+\bar{y}}{2})^2 + (\bar{y} - \frac{\bar{x}+\bar{y}}{2})^2]}{\frac{1}{2(n-1)}[\sum_{j=1}^n (x_j - \bar{x})^2 + \sum_{j=1}^n (y_j - \bar{y})^2]} = \frac{2n(\frac{\bar{x}-\bar{y}}{2})^2}{s_{\mathrm{p}}^2} = \Big(\frac{\bar{x}-\bar{y}}{s_{\mathrm{p}}\sqrt{\frac{2}{n}}}\Big)^2.$$

This equals t^2, where $t = \frac{\bar{x}-\bar{y}}{s_{\mathrm{p}}\sqrt{\frac{2}{n}}}$ is the two-sample t-test statistic.

Problem 12.5

Derive the likelihood ratio test for the null hypothesis of the one-way layout, and show that it is equivalent to the F-test.

Solution 12.5

The null hypothesis says that the data (y_{ij}) comes from a single normal distribution

$$H_0 : \mu_1 = \ldots = \mu_I = \mu$$

described by two parameters μ and σ^2, so that dim $\Omega_0 = 2$, while

$$\dim \Omega = I + 1$$

since the general setting is described by parameters $\mu_1, \ldots, \mu_I$ and σ^2. The likelihood ratio

$$\Lambda = \frac{L_0(\hat{\mu}, \hat{\sigma}_0^2)}{L(\hat{\mu}_1, \ldots, \hat{\mu}_I, \hat{\sigma}^2)},$$

is expressed in terms of two likelihood functions

$$L(\mu_1, \ldots, \mu_I, \sigma^2) = \prod_{i=1}^{I}\prod_{j=1}^{J} \tfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}} \propto \sigma^{-n} \exp\{-\sum_i \sum_j \tfrac{(y_{ij}-\mu_i)^2}{2\sigma^2}\},$$

$$L_0(\mu, \sigma^2) = L(\mu, \ldots, \mu, \sigma^2) \propto \sigma^{-n} \exp\{-\sum_i \sum_j \tfrac{(y_{ij}-\mu)^2}{2\sigma^2}\}.$$

where $n = IJ$. We find the maximum likelihood estimates to be

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \quad \hat{\sigma}_0^2 = \frac{SS_{\mathrm{T}}}{n}, \quad \hat{\mu}_i = \bar{y}_{i\cdot}, \quad \hat{\sigma}^2 = \frac{SS_{\mathrm{E}}}{n},$$

which yields

$$\Lambda = \frac{\hat{\sigma}_0^{-n} \exp\{-\sum\sum \frac{(y_{ij}-\hat{\mu})^2}{2\hat{\sigma}_0^2}\}}{\hat{\sigma}^{-n} \exp\{-\sum\sum \frac{(y_{ij}-\hat{\mu}_i)^2}{2\hat{\sigma}^2}\}} = \Big(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\Big)^{-n/2}.$$

The likelihood ratio test rejects the null hypothesis for small values of Λ or equivalently for large values of

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = \tfrac{SS_{\mathrm{T}}}{SS_{\mathrm{E}}} = 1 + \tfrac{SS_{\mathrm{A}}}{SS_{\mathrm{E}}} = 1 + \tfrac{J(I-1)MS_{\mathrm{A}}}{I(J-1)MS_{\mathrm{E}}} = 1 + \tfrac{J(I-1)}{I(J-1)} \cdot F$$

that is for large values of F-test statistics. This leads to an asymptotic approximation of the $F_{J(I-1),I(J-1)}$ in terms of the chi-square distribution with df $= I - 1$.

Problem 12.10

Suppose in a one-way layout there are 10 treatments and seven observations under each treatment. What is the ratio of the length of a simultaneous confidence interval for the difference of two means formed by Tukey's method to that of one formed by the Bonferroni method? How do both of these compare in length to an interval based on the t-distribution that does not take account of multiple comparisons?

Solution 12.10

One-way layout with $I = 10$, $J = 7$,

$$Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2).$$

Pooled sample variance

$$s_\mathrm{p}^2 = MS_\mathrm{E} = \frac{1}{I(J-1)} \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2$$

uses $\mathrm{df} = I(J-1) = 60$.

(a) A 95% confidence interval for a single difference $\mu_u - \mu_v$

$$I_{\mu_u - \mu_v} = \bar{y}_{u\cdot} - \bar{y}_{v\cdot} \pm t_{60}(0.025) s_\mathrm{p} \sqrt{\tfrac{2}{J}}$$

has the half-width of

$$2.82 \cdot \tfrac{s_\mathrm{p}}{\sqrt{J}}.$$

(b) Bonferroni simultaneous 95% confidence interval for $\binom{10}{2} = 45$ differences $\mu_u - \mu_v$

$$B_{\mu_u - \mu_v} = \bar{y}_{u\cdot} - \bar{y}_{v\cdot} \pm t_{60}(\tfrac{0.025}{45}) s_\mathrm{p} \sqrt{\tfrac{2}{J}}$$

has the half-width of

$$4.79 \cdot \tfrac{s_\mathrm{p}}{\sqrt{J}},$$

giving the ratio

$$\tfrac{4.79}{2.82} = 1.7.$$

(c) Tukey simultaneous 95% confidence interval for differences $\mu_u - \mu_v$

$$T_{\mu_u - \mu_v} = \bar{y}_{u\cdot} - \bar{y}_{v\cdot} \pm q_{10,60}(0.05) \tfrac{s_\mathrm{p}}{\sqrt{J}}$$

has the half-width of

$$4.65 \cdot \tfrac{s_\mathrm{p}}{\sqrt{J}},$$

giving the ratio

$$\tfrac{\text{Bonferroni}}{\text{Tukey}} = \tfrac{4.79}{4.65} = 1.03.$$

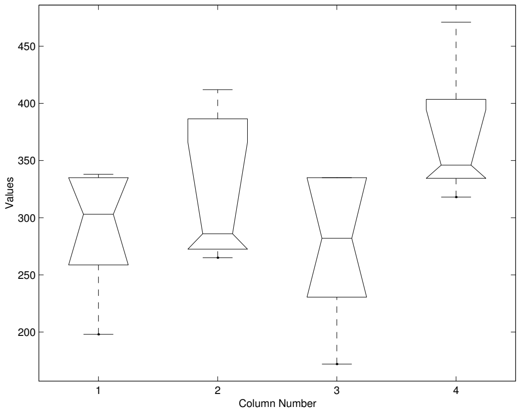
Problem 12.21

Duringeach of four experiments on the use of carbon tetrachloride as a worm killer, ten rats were infested with larvae (Armitage 1983). Eight days later, five rats were treated with carbon tetrachloride; theother five were kept as controls. After two more days, all the rats were killed and the numbers of worms were counted. The table below gives the counts of worms for the four control groups.

Group I	Group II	Group III	Group IV
279	378	172	381
338	275	335	346
334	412	335	340
198	265	282	471
303	286	250	318

Significant differences among the control groups, although not expected, might be attributable to changes in the experimental conditions. A finding of significance differences could result in more carefully controlled experimentation and thus greater precision in later work.

Use both graphical techniques and the F-test to test whether there are significant differences among the four groups. Use a nonparametric technique as well.



Solution 12.21

For $I = 4$ control groups of $J = 5$ mice each, test H_0: no systematic differences between groups. One way ANOVA table

Source	SS	df	MS	F	P
Columns	27230	3	9078	2.271	0.12
Error	63950	16	3997		
Total	91190	19			

Do not reject H_0 at 10% significance level. Boxplots show non-normality. The largest difference is between the third and the fourth boxplots. Control question: why the third boxplot has no upper whisker?

Kruskal-Wallis test. Pooled sample ranks

Group I	2	6	9	11	14	$\bar{r}_{1.} = 8.4$
Group II	4	5	8	17	19	$\bar{r}_{2.} = 10.6$
Group III	1	3	7	12.5	12.5	$\bar{r}_{3.} = 7.2$
Group IV	10	15	16	18	20	$\bar{r}_{4.} = 15.8$

Kruskal-Wallis test statistic

$$K = \frac{12 \cdot 5}{20 \cdot 21} \left((8.4 - 10.5)^2 + (10.6 - 10.5)^2 + (7.2 - 10.5)^2 + (15.8 - 10.5)^2 \right) = 6.20.$$

Since $\chi_3^2(0.1) = 6.25$, we do not reject H_0 at 10% significance level.

Problem 12.26

The concentrations (in nanogram per milimiter) of plasma epinephrine were measured for 10 dogs under isofluorane, halothane, and cyclopropane anesthesia. The measurements are given in the following table (Perry et al. 1974).

	Dog 1	Dog 2	Dog 3	Dog 4	Dog 5	Dog 6	Dog 7	Dog 8	Dog 9	Dog 10
Isofluorane	0.28	0.51	1.00	0.39	0.29	0.36	0.32	0.69	0.17	0.33
Halothane	0.30	0.39	0.63	0.68	0.38	0.21	0.88	0.39	0.51	0.32
Cyclopropane	1.07	1.35	0.69	0.28	1.24	1.53	0.49	0.56	1.02	0.30

Is there a difference in treatment effects? Use a parametric and a nonparametric analysis.

Solution 12.26

Two-way layout with $I = 3$ treatments on $J = 10$ subjects with $K = 1$ observations per cell. ANOVA table

Source	SS	df	MS	F	P
Columns (blocks)	0.517	9	0.0574	0.4683	0.8772
Rows (treatments)	1.081	2	0.5404	4.406	0.0277
Error	2.208	18	0.1227		
Total	3.806	29			

Reject

$$H_0\text{: no treatment effects}$$

at 5% significance level. (Interestingly, no significant differences among the blocks.)

Friedman's test. Ranking within blocks:

	Dog 1	Dog 2	Dog 3	Dog 4	Dog 5	Dog 6	Dog 7	Dog 8	Dog 9	Dog 10	$\bar{r}_{i.}$
Isof	1	2	3	2	1	2	1	3	1	3	1.9
Halo	2	1	1	3	2	1	3	1	2	2	1.8
Cycl	3	3	2	1	3	3	2	2	3	1	2.3

The observed value of the Friedman test statistic

$$Q = \frac{12 \cdot 10}{3 \cdot 4}\left((1.8-2)^2 + (1.9-2)^2 + (2.3-2)^2\right) = 1.4.$$

Since $\chi_2^2(0.1) = 4.61$, we can not reject H_0 even at 10% significance level.

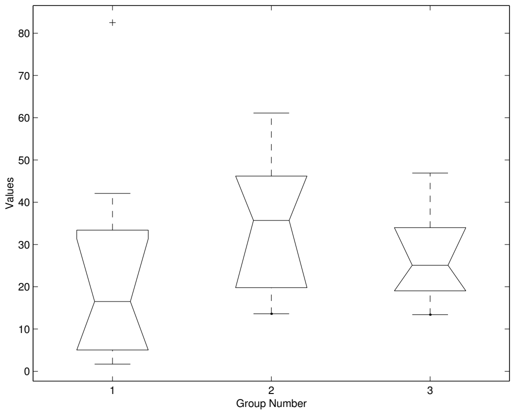
Problem 12.28

Samples of each of three types of stopwatches were tested. The following table gives thousands of cycles (on-off-restart) survived until some part of the mechanism failed (Natrella 1963).

Type I	1.7	1.9	6.1	12.5	16.5	25.1	30.5	42.1	82.5
Type II	13.6	19.8	25.2	46.2	46.2	61.1			
Type III	13.4	20.9	25.1	29.7	46.9				

Test whether there is a significant difference among the types. Use both a parametric and a nonparametric technique.

Solution 12.28



One way layout with $I = 3$ types of stopwatches, different sample sizes. ANOVA table

Source	SS	df	MS	F	P
Columns	446.6	2	223.3	0.4974	0.6167
Error	7632	17	449		
Total	8079	19			

gives the p-value of 0.6167. We do not reject H_0: no systematic differences between groups.

Kruskal-Wallis test. Pooled sample ranks

Type I	1	2	3	4	7	10.5	14	15	20	$\bar{r}_{1.} = 8.5$
Type II	6	8	12	16.5	16.5	19				$\bar{r}_{2.} = 13.0$
Type III	5	9	10.5	13	18					$\bar{r}_{3.} = 11.1$

The observed value of the test statistic

$$K = \frac{12}{20 \cdot 21}\big(9 \cdot (8.5 - 10.5)^2 + 6 \cdot (13.0 - 10.5)^2 + 5 \cdot (11.1 - 10.5)^2\big) = 2.15.$$

Since $\chi^2_2(0.1) = 4.61$, we do not reject H_0 even at 10% significance level.

Problem 12.34

The following table gives the survival times (in hours) for animals in an experiment whose design consisted of three poisons, four treatments, and four observations per cell.

	Treatment A		Treatment B		Treatment C		Treatment D	
Poison I	3.1	4.5	8.2	11.0	4.3	4.5	4.5	7.1
	4.6	4.3	8.8	7.2	6.3	7.6	6.6	6.2
Poison II	3.6	2.9	9.2	6.1	4.4	3.5	5.6	10.0
	4.0	2.3	4.9	12.4	3.1	4.0	7.1	3.8
Poison III	2.2	2.1	3.0	3.7	2.3	2.5	3.0	3.6
	1.8	2.3	3.8	2.9	2.4	2.2	3.1	3.3

(a) Conduct a two-way analysis of variance to test the effects of the two main factors and their interaction.

(b) Box and Cox (1964) analysed the reciprocals of the data, pointing out that the reciprocal of a survival time can be interpreted as the rate of death. Conduct a two-way analysis of variance, and compare to the results of part (a). Comment on how well the standard two-way ANOVA model fits and on the interaction in both analyses.

Solution 12.34

Forty eight survival times: $I = 3$ poisons and $J = 4$ treatments with $K = 4$ observations per cell. Cell means for the survival times

	A	B	C	D
I	4.125	8.800	5.675	6.100
II	3.200	8.150	3.750	6.625
III	2.100	3.350	2.350	3.250

Draw three profiles: I and II cross each other, and profile III is more flat. Three null hypotheses of interest

H_A: no poison effect,
H_B: no treatment effect,
H_{AB}: no interaction.

(a) Survival in hours x data matrix. Results of anova2(x,4)

Source	SS	df	MS	F	P
Columns (treatments)	91.9	3	30.63	14.01	0.0000
Rows (poisons)	103	2	51.52	23.57	0.0000
Intercation	24.75	6	4.124	1.887	0.1100
Error	78.69	36	2.186		
Total	298.4	47			

Reject H_A and H_B at 1% significance level, we can not reject H_{AB} even at 10% significance level:

3 poisons act differently,
4 treatments act differently,
some indication of interaction.

Analysis of the residuals

normal probability plot reveals non-normality,
skewness = 0.59,
kurtosis = 4.1.

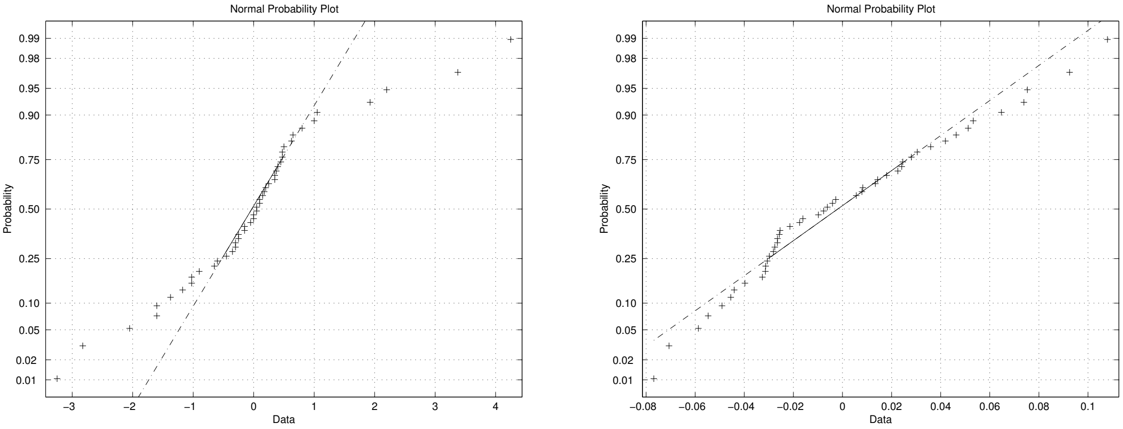


Figure 1: Left panel: survival times. Right panel: death rates.

(b) Transformed data: death rate = 1/survival time. Cell means for the death rates

	A	B	C	D
I	0.249	0.116	0.186	0.169
II	0.327	0.139	0.271	0.171
III	0.480	0.303	0.427	0.309

Draw three profiles: they look more parallel.

New data matrix y=x.^(-1). Results of anova2(y,4):

Source	SS	df	MS	F	P
Columns (treatments)	0.204	3	0.068	28.41	0.0000
Rows (poisons)	0.349	2	0.174	72.84	0.0000
Intercation	0.01157	6	0.0026	1.091	0.3864
Error	0.086	36	0.0024		
Total	0.6544	47			

Reject H_A and H_B at 1% significance level. Do not reject H_{AB}. Conclusions

3 poisons act differently,
4 treatments act differently,
no interaction,
the normal probability plot of residuals reveals a closer fit to normality assumption.

8 Categorical data analysis

Warning: in some of the contingency tables the expected counts are rounded. If you then will compute the chi-square test statistic χ^2 from the table, you will often get a somewhat different value.

Problem 13.1

Adult-onset diabetes is known to be highly genetically determined. A study was done comparing frequencies of a particular allele in a sample of such diabetics and a sample of nondiabetics. The data is shown in the following table:

	Diabetic	Normal	Total
Bb or *bb*	12	4	16
BB	39	49	88
Total	51	53	104

Are the relative frequencies of the alleles significantly different in the two groups?

Solution 13.1

Test

H_0: same genotype frequencies for diabetics and normal
using the chi-square test of homogeneity.

	Diabetic	Normal	Total
Bb or *bb*	12 (7.85)	4 (8.15)	16
BB	39 (43.15)	49 (44.85)	88
Total	51	53	104

Observed χ^2=5.10, df=1, p-value = 0.024. Reject H_0. Diabetics have genotype BB less often.

The exact Fisher test uses Hg(104,51,$\frac{16}{104}$) as the null distribution of the test statistic $N_{11} = 12$

one-sided P-value: 1-hygecdf(11,104,16,51)=0.0225,
two-sided P-value $P = 0.045$.

Normal approximation of the null distribution

$$\text{Hg}(104, 51, \tfrac{16}{104}) \approx \text{ N}(7.85, 3.41).$$

Since $z_{\text{obs}} = \frac{12-7.85}{\sqrt{3.41}}$=2.245, the approximate two-sided p-value = 0.025.

Problem 13.3

Overfield and Klauber (1980) published the following data on the incidence of tuberculosis in relation to blood groups in a sample of Eskimos. Is there any association of the disease and blood group within the ABO system or within the MN system?

	O	A	AB	B
Moderate	7	5	3	13
Minimal	27	32	8	18
Not present	55	50	7	24

	MM	MN	NN
Moderate	21	6	1
Minimal	54	27	5
Not present	74	51	11

Solution 13.3

(a) H_0: no association of the disease and the ABO blood group:

	O	A	AB	B	Total
Moderate	7 (10.4)	5 (9.8)	3 (2.0)	13 (6.2)	28
Minimal	27 (30.4)	32 (29.7)	8 (6.1)	18 (18.8)	85
Not present	55 (48.6)	50 (47.5)	7 (9.8)	24 (30.0)	136
Total	89	87	18	55	249

Observed χ^2=15.37, df=6, p-value = 0.018. Reject H_0.

(b) H_0: no association of the disease and the MN blood group:

	MM	MN	NN	Total
Moderate	21 (16.7)	6 (9.4)	1 (1.9)	28
Minimal	54 (51.3)	27 (28.9)	5 (5.8)	86
Not present	74 (81.1)	51 (45.7)	11 (9.2)	136
Total	149	84	17	250

Observed χ^2=4.73, df=4, p-value = 0.42. Can not reject H_0.

Problem 13.6

It is conventional wisdom in military squadron that pilots tend to father more girls than boys. Snyder (1961) gathered data for military fighter pilots. The sex of the pilots' offspring were tabulated for three kinds of flight duty during the month of conception, as shown in the following table.

	Girl	Boy
Flying fighter	51	38
Flying transport	14	16
Not flying	38	46

(a) Is there any significant difference between the three groups?

(b) In the United States in 1950, 105.37 males were born for every 100 females. Are the data consistent with this sex ratio?

Solution 13.6

(a) Apply the chi-square test of homogeneity:

	Girl	Boy	Total
Flying fighter	51 (45.16)	38 (43.84)	89
Flying transport	14 (15.22)	16 (14.78)	30
Not flying	38 (42.62)	46 (41.38)	84
Total	103	100	203

Observed χ^2=2.75, df=2, p-value = 0.25. Can not reject H_0.

(b) Goodness of fit chi-square test for the same sex ratio for three father's activities

$$H_0\text{: boys proportions } p_{12} = p_{22} = p_{32} = 0.513.$$

Here 0.513 is obtained as

$$\frac{105.37}{105.37 + 100} = 0.513.$$

Observed and expected counts

	Girl	Boy	Total
Flying fighter	51 (45.66)	38 (43.34)	89
Flying transport	14 (15.39)	16 (14.61)	30
Not flying	38 (43.09)	46 (40.91)	84
Total	103	100	203

Observed χ^2=2.775, df=3, p-value = 0.43. Can not reject H_0.

Problem 13.8

A randomized double-blind experiment compared the effectiveness of several drugs in ameliorating postoperative nausea. All patients were anesthetized with nitrous oxide and ether. The following table shows the incidence of nausea during the first four hours for each of several drugs and a placebo (Beecher 1959).

	Number of patients	Incidence of nausea
Placebo	165	95
Chlorpromazine	152	52
Dimenhydrinate	85	52
Pentobarbital (100 mg)	67	35
Pentobarbital (150 mg)	85	37

Compare the drugs to each other and to the placebo.

Solution 13.8

We use the chi-square test for homogeneity

	No nausea	Incidence of nausea	Total
Placebo	70 (84)	95 (81)	165
Chlorpromazine	100 (78)	52 (74)	152
Dimenhydrinate	33 (43)	52 (42)	85
Pentobarbital (100 mg)	32 (34)	35 (33)	67
Pentobarbital (150 mg)	48 (43)	37 (42)	85
Total (150 mg)	283	271	554

The observed test statistic $\chi^2 = 35.8$ according to the χ^2_4-distribution table gives p-value $= 3 \cdot 10^{-7}$. Comparing the observed and expected counts we conclude that Placebo and Dimenhydrinate are most effective in ameliorating postoperative nausea.

Problem 13.17

In a study of the relation of blood type to various diseases, the following data were gathered in London and Manchester (Woolf 1955).

London	Control	Peptic Ulcer
Group A	4219	579
Group O	4578	911

Manchester	Control	Peptic Ulcer
Group A	3775	246
Group O	4532	361

First, consider the two tables separately. Id there a relationship between blood type and propensity to peptic ulcer? If so, evaluate the strength of the relationship. Are the data from London and Manchenster comparable?

Solution 13.17

(a) H_0: no relation between blood group and disease in London:

	Control	Peptic Ulcer	Total
Group A	4219 (4103.0)	579 (695.0)	4798
Group O	4578 (4694.0)	911 (795.0)	5489
Total	8797	1490	10287

Observed χ^2=42.40, df=1, p-value = 0.000. Reject H_0. Odds ratio $\hat{\Delta} = 1.45$.

(b) H_0: no relation between blood group and disease in Manchester:

	Control	Peptic Ulcer	Total
Group A	3775 (3747.2)	246 (273.8)	4021
Group O	4532 (4559.8)	361 (333.2)	4893
Total	8307	607	8914

Observed χ^2=5.52, df=1, p-value = 0.019. Reject H_0. Odds ratio $\hat{\Delta} = 1.22$.

(c) H_0: London Group A and Manchester Group A have the same propensity to Peptic Ulcer:

	C and A	PU and A	Total
London	4219 (4349.2)	579 (448.8)	4798
Manchester	3775 (3644.8)	246 (376.2)	4021
Total	7994	825	8819

Observed χ^2=91.3, df=1, p-value = 0.000. Reject H_0.

H_0: London Group O and Manchester Group O have the same propensity to Peptic Ulcer:

	C and O	PU and O	Total
London	4578 (4816.5)	911 (672.5)	5489
Manchester	4532 (4293.5)	361 (599.5)	4893
Total	9110	1272	10382

Observed χ^2=204.5, df=1, p-value = 0.000. Reject H_0.

Problem 13.18

Record of 317 patients at least 48 years old who were diagnosed as having endometrial carcinoma were obtained from two hospitals (Smith et al. 1975). Matched controls for each case were obtained from the two institutions: the controls had cervical cancer, ovarian cancer, or carcinoma of the vulva. Each control was matched by age at diagnosis (within four years) and year of diagnosis (within two years) to a corresponding case of endometrial carcinoma.

The following table gives the number of cases and controls who had taken estrogen for at least 6 months prior to the diagnosis of cancer.

	Controls: estrogen used	Controls: estrogen not used	Total
Cases: estrogen used	39	113	152
Cases: estrogen not used	15	150	165
Total	54	263	317

(a) Is there a significant relationship between estrogen use and endometrial cancer?

(b) This sort of of design, called a retrospective case-control study, is frequently used in medical investigations where a randomised experiment is not possible. Do you see any possible weak points in a retrospective case-control design?

Solution 13.18

D = endometrical carcinoma, X = estrogen taken at least 6 months prior to the diagnosis of cancer.

(a) Matched controls, retrospective case-control study

	$\bar{D}X$	$\bar{D}\bar{X}$	Total
DX	39	113	152
$D\bar{X}$	15	150	165
Total	54	263	317

Apply McNemar test for

$$H_0 : \pi_{1\cdot} = \pi_{\cdot 1} \quad \text{vs} \quad H_1 : \pi_{1\cdot} \neq \pi_{\cdot 1}.$$

Observed value of the test statistic

$$\chi^2 = \tfrac{(113-15)^2}{113+15} = 75$$

is highly significant as $\sqrt{75} = 8.7$ and the corresponding two-sided P-value obtained from N(0,1) table is very small.

(b) Possible weak points in a retrospective case-control design

- selection bias: some patients have died prior the study,
- information bias: have to rely on other sources of information.

Problem 13.19

A psychological experiment was done to investigate the effect of anxiety on a person's desire to be alone or in company (Lehman 1975). A group of 30 subjects was randomly divided into two groups of sizes 13 and 17. The subjects were told that they would be subjected to some electric shocks, but one group (high-anxiety) was told that the shocks would be quite painful and the other group (low-anxiety) was told that they would be mild and painless. Both groups were told that there would be a 10-min wait before the experiment began, and each subject was given the choice of waiting alone or with the other subjects. The following are the results:

	Wait Together	Wait Alone	Total
High-Anxiety	12	5	17
Low-Anxiety	4	9	13
Total	16	14	30

Use Fisher's exact test to test whether there is a significant difference between the high- and low-anxiety groups. What is a reasonable one-sided alternative?

Solution 13.19

(a) The exact Fisher test uses Hg(30,17,$\frac{16}{30}$) as the null distribution of the test statistic $n_{11} = 12$. It gives

one-sided P-value: 1-hygecdf$(11, 30, 16, 17) = 0.036$,
two-sided P-value $P = 0.071$.

(b) Using normal approximation

$$\text{Hg}(30, 17, \tfrac{16}{30}) \approx \text{N}(9.1, (1.4)^2)$$

and continuity correction, we find the one-sided p-value to be

$$\text{P}(N_{11} \geq 12|H_0) = \text{P}(N_{11} > 11|H_0) \approx 1 - \Phi(\tfrac{11.5-9.1}{1.4}) = 1 - \Phi(1.71) = 0.044.$$

(c) Approximate chi-square test yields: observed χ^2=4.69, df=1, two-sided p-value

$$2(1 - \Phi(\sqrt{4.69})) = 2(1 - \Phi(2.16)) = 0.03.$$

Problem 13.24

Hill and Barton (2005): red against blue outfits - does it matter in combat sports? Although other colors are also present in animal displays, it is specifically the presence and intensity of red coloration that correlates with male dominance and testosterone levels. Increased redness during aggressive interactions may reflect relative dominance.

In the 2004 Olympic Games, contestants in four combat sports were randomly assigned red and blue outfits. The winner counts in different sports

	Red	Biue	Total
Boxing	148	120	268
Freestyle wrestling	27	24	51
Greco-Roman wrestling	25	23	48
Tae Kwon Do	45	35	80
Total	245	202	447

(a) Let π_R denote the probability that the contestant wearing red wins. Test the null hypothesis that $\pi_R = 0.5$ versus the alternative hypothesis that π_R is the same in each sport, but $\pi_R \neq 0.5$.

(b) Test the null hypothesis that $\pi_R = 0.5$ versus the alternative hypothesis that allows π_R to be different in different sports, but not equal to 0.5.

(c) Are these hypothesis tests equivalent to that which would test the null hypothesis $\pi_R = 0.5$ versus the alternative hypothesis $\pi_R \neq 0.5$, using as data the total numbers of wins summed over all the sports?

(d) Is there any evidence that wearing red is more favourable in some of the sports than others?

Solution 13.24

Denote

$\pi_1 =$ probability that red wins in boxing,
$\pi_2 =$ probability that red wins in freestyle wrestling,
$\pi_3 =$ probability that red wins in Greco-Roman wrestling,
$\pi_4 =$ probability that red wins in Tae Kwon Do.

(a, c) Assuming

$$H_{eq}: \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi,$$

we test

$$H_0: \pi = \tfrac{1}{2} \quad \text{vs} \quad H_1: \pi \neq \tfrac{1}{2}.$$

We use the large sample test for proportion based on the statistic $X = 245$ whose null distribution is $\text{Bin}(n, \frac{1}{2})$, $n = 447$. The two-sided P-value is approximated by

$$2(1 - \Phi(\tfrac{245 - \frac{447}{2}}{\sqrt{447} \cdot \frac{1}{2}}) = 2(1 - \Phi(2.034) = 0.042.$$

At 5% level we reject the $H_0: \pi = \frac{1}{2}$. The maximum likelihood estimate is $\hat{\pi} = \frac{245}{447} = 0.55$.

(d) Is there evidence that wearing red is more favourable in some of the sports than others? We test

$$H_{eq}: \pi_1 = \pi_2 = \pi_3 = \pi_4 \quad \text{vs} \quad H_{ineq}: \pi_i \neq \pi_j \quad \text{for some } i \neq j$$

using the chi-square test of homogeneity. From

	Red	Biue	Total
Boxing	148 (147)	120 (121)	268
Freestyle wrestling	27 (28)	24 (23)	51
Greco-Roman wrestling	25 (26)	23 (22)	48
Tae Kwon Do	45 (44)	35 (36)	80
Total	245	202	447
Marginal proportions	0.55	0.45	1.00

we find that the test statistic $\boldsymbol{\chi}^2 = 0.3$ is not significant. We can not reject H_{eq}, which according to (a) leads to $\hat{\pi} = 0.55$.

(b) Now we state the hypotheses of interest directly: consider

$$H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4 = \tfrac{1}{2} \quad \text{vs} \quad H_1: (\pi_1, \pi_2, \pi_3, \pi_4) \neq (\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}).$$

Here we need a new chi-square test, a chi-square test for k proportions with $k = 4$ (see below). Given four observed counts $x_1 = 148$, $x_2 = 27$, $x_3 = 25$, $x_4 = 45$, we obtain

	Red	Biue	Total
Boxing	148 (134)	120 (134)	268
Freestyle wrestling	27 (25.5)	24 (25.5)	51
Greco-Roman wrestling	25 (24)	23 (24)	48
Tae Kwon Do	45 (40)	35 (40)	80
H_0 proportions	0.5	0.5	1.00

we find $\boldsymbol{\chi}^2_{\text{obs}} = 4.4$. Since $\chi^2_4(0.1) = 7.8$, we do not reject $H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4 = \frac{1}{2}$.

Chi-square test for k proportions

We derive it using the likelihood ratio approach.

the likelihood function based on the binomial model has the form

$$L(\pi_1,\pi_2,\pi_3,\pi_4)=\prod_{i=1}^{4}\binom{n_i}{x_i}\pi_i^{x_i}(1-\pi_i)^{n_i-x_i}.$$

Using $\hat\pi_i=\frac{x_i}{n_i}$, we compute the likelihood ratio as

$$\Lambda=\frac{L(\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2})}{L(\hat\pi_1,\hat\pi_2,\hat\pi_3,\hat\pi_4)}=\frac{(\frac{1}{2})^n}{\prod_{i=1}^{4}(\frac{x_i}{n_i})^{x_i}(\frac{n_i-x_i}{n_i})^{n_i-x_i}}.$$

Turning to the logarithms,

$$\Delta=-\ln\Lambda=\sum_{i=1}^{4}x_i\ln\tfrac{2x_i}{n_i}+(n_i-x_i)\ln\tfrac{2(n_i-x_i)}{n_i},$$

we take 2Δ as the test statistic.

Next we show that the null distribution of 2Δ is approximately χ^2_4. Under H_0 we have $\frac{2x_i}{n_i}\approx 1$, and using a Taylor expansion we find that

$$2\Delta\approx\boldsymbol{\chi}^2=\sum_{i=1}^{4}\tfrac{(x_i-\frac{n_i}{2})^2}{n_i/4},$$

where $Z_i=\frac{X_i-\frac{n_i}{2}}{\sqrt{n_i/4}}$ are independent and approximately N(0,1) distributed, provided $X_i\sim\text{Bin}(n_i,\frac{1}{2})$.

9 Multiple regression

Problem 14.2

Ten pairs

x	0.34	1.38	-0.65	0.68	1.40	-0.88	-0.30	-1.18	0.50	-1.75
y	0.27	1.34	-0.53	0.35	1.28	-0.98	-0.72	-0.81	0.64	-1.59

Draw a scatter plot.

(a) Fit a straight line $y=a+bx$ by the method of least squares, and sketch it on the plot.

(b) Fit a straight line $x=c+dy$ by the method of least squares, and sketch it on the plot.

(c) Are the lines on (a) and (b) the same? If not, why not?

Solution 14.2

We have after ordering

x	-1.75	-1.18	-0.88	-0.65	-0.30	0.34	0.50	0.68	1.38	1.40
y	-1.59	-0.81	-0.98	-0.53	-0.72	0.27	0.64	0.35	1.34	1.28

and

$$\bar x=-0.046,\quad \bar y=-0.075,\quad s_x=1.076,\quad s_y=0.996,\quad r=0.98.$$

(a) Simple linear regression model

$$Y=\beta_0+\beta_1X+\epsilon,\quad \epsilon\sim\text{N}(0,\sigma^2).$$

Fitting a straight line using

$$y-\bar y=r\cdot\tfrac{s_y}{s_x}(x-\bar x)$$

we get the predicted response

$$\hat{y} = -0.033 + 0.904 \cdot x.$$

Estimated σ^2

$$s^2 = \tfrac{n-1}{n-2} s_y^2 (1 - r^2) = 0.05.$$

(b) Simple linear regression model

$$X = \beta_0 + \beta_1 y + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Fitting a straight line using

$$x - \bar{x} = r \cdot \tfrac{s_x}{s_y} (y - \bar{y})$$

we get the predicted response

$$\hat{x} = 0.033 + 1.055 \cdot y.$$

Estimated σ^2

$$s^2 = \tfrac{n-1}{n-2} s_x^2 (1 - r^2) = 0.06.$$

(c) First fitted line

$$y = -0.033 + 0.904 \cdot x$$

is different from the second

$$y = -0.031 + 0.948 \cdot x.$$

They are different since in (a) we minimise the vertical residuals while in (b) - horizontal.

Problem 14.4

Two consecutive grades

X = the high school GPA (grade point average),
Y = the freshman GPA.

Allow two different intercepts for females

$$Y = \beta_F + \beta_1 X + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

and for males

$$Y = \beta_M + \beta_1 X + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Give the form of the design matrix for such a model.

Solution 14.4

Using an extra explanatory variable f which equal 1 for females and 0 for males, we rewrite this model in the form of a multiple regression

$$Y = f\beta_F + (1 - f)\beta_F + \beta_1 X + \epsilon = \beta_0 + \beta_1 X + \beta_2 f + \epsilon,$$

where

$$\beta_0 = \beta_M, \quad \beta_2 = \beta_F - \beta_M.$$

Here $p = 3$ and the design matrix is

$$\mathbb{X} = \begin{pmatrix} 1 & x_1 & f_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & f_n \end{pmatrix}.$$

After $\beta_0, \beta_1, \beta_2$ are estimated, we compute

$$\beta_M = \beta_0, \quad \beta_F = \beta_0 + \beta_2.$$

A null hypothesis of interest $\beta_2 = 0$.

Problem 14.14

Simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Using n pairs of (x_i, y_i) we fit a regression line by

$$y = b_0 + b_1 x, \quad \mathrm{Var}(b_0) = \tfrac{\sigma^2 \overline{x^2}}{(n-1)s_x^2}, \quad \mathrm{Var}(b_1) = \tfrac{\sigma^2}{(n-1)s_x^2}, \quad \mathrm{Cov}(b_0, b_1) = -\tfrac{\sigma^2 \bar{x}}{(n-1)s_x^2}.$$

For a given $x = x_0$, we wish to predict the value of a new observation

$$y_0 = \beta_0 + \beta_1 x_0 + e_0$$

by

$$\hat{y}_0 = b_0 + b_1 x_0.$$

(a) Find an expression for the variance of $\hat{Y}_0 - Y_0$, and compare it to the variance of $\hat{Y}_0$. Find C_n, the standard deviation of $\frac{\hat{Y}_0 - Y_0}{\sigma}$.

(b) Derive the formula for 95% prediction interval such that

$$\mathrm{P}(Y_0 \in I) = 0.95$$

using

$$\frac{Y_0 - \hat{Y}_0}{sC_n} \sim t_{n-2}.$$

Solution 14.14

(a) The predicted value $\hat{y}_0$ and actual observation Y_0 are independent random variables, therefore

$$\mathrm{Var}(Y_0 - \hat{Y}_0) = \mathrm{Var}(Y_0) + \mathrm{Var}(\hat{y}_0) = \sigma^2 + \mathrm{Var}(b_0 + b_1 x_0) = \sigma^2 C_n^2,$$

where

$$C_n^2 = 1 + \tfrac{\mathrm{Var}(b_0) + \mathrm{Var}(b_1) x_0^2 - 2x_0 \mathrm{Cov}(b_0, b_1)}{\sigma^2} = 1 + \tfrac{\overline{x^2} + x_0^2 - 2\bar{x}x_0}{(n-1)s_x^2} = 1 + \tfrac{\overline{x^2} - \bar{x}^2 + (x_0 - \bar{x})^2}{(n-1)s_x^2} = 1 + \tfrac{1}{n} + \tfrac{(x_0 - \bar{x})^2}{(n-1)s_x^2}.$$

(b) A 95% prediction interval I for the new observation Y_0 is obtained from

$$\tfrac{Y_0 - \hat{Y}_0}{sC_n} \sim t_{n-2}.$$

Since

$$0.95 = \mathrm{P}(|Y_0 - \hat{Y}_0| \le t_{n-2}(0.025) \cdot sC_n) = \mathrm{P}(Y_0 \in \hat{Y}_0 \pm t_{n-2}(0.025) \cdot sC_n),$$

we conclude that a 95% prediction interval for the new observation Y_0 is given by

$$I = b_0 + b_1 x_0 \pm t_{n-2}(0.025) \cdot s\sqrt{1 + \tfrac{1}{n} + \tfrac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}.$$

The further from $\bar{x}$ lies x_0, the more uncertain becomes the prediction.

Problem 14.23

Data collected for

$x =$ midterm grade,
$y =$ final grade,

gave

$$r = 0.5, \quad \bar{x} = \bar{y} = 75, \quad s_x = s_y = 10.$$

(a) Given $x = 95$, predict the final score.

(b) Given $y = 85$ and not knowing the midterm score, predict the midterm score.

Solution 14.23

(a) Given $x = 95$, we predict the final score by

$$\hat{y} = 75 + 0.5(95 - 75) = 85.$$

Regression to mediocracy.

(b) Given $y = 85$ and we do not know the midterm score, we predict the midterm score by

$$\hat{x} = 75 + 0.5(85 - 75) = 80.$$

Problem 14.33

Let

$$Y = X + \beta Z,$$

where $X \in \mathrm{N}(0, 1)$ and $Z \in \mathrm{N}(0, 1)$ are independent.

(a) Show that the correlation coefficient for X and Y is

$$\rho = \tfrac{1}{\sqrt{1+\beta^2}}.$$

(b) Use the result of part (a) to generate bivariate samples (x_i, y_i) of size 20 with population correlation coefficients -0.9, -0.5, 0, 0.5, and 0.9. Compute the sample correlation coefficients.

Solution 14.33

(a) Find the correlation coefficient ρ for (X, Y). Since $\mathrm{E}X = 0$, we have

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) = \mathrm{E}(X^2 + \beta XZ) = 1, \quad \mathrm{Var}Y = \mathrm{Var}X + \mathrm{Var}Z = 1 + \beta^2,$$

and we see that the correlation coefficient is always positive

$$\rho = \tfrac{1}{\sqrt{1+\beta^2}}.$$

(b) Use (a) to generate five samples

$$(x_1, y_1), \ldots, (x_{20}, y_{20})$$

with different

$$\rho = -0.9, \quad -0.5, \quad 0, \quad 0.5, \quad 0.9,$$

and compute the sample correlation coefficients.

From $\rho = \tfrac{1}{\sqrt{1+\beta^2}}$, we get $\beta = \sqrt{\rho^{-2} - 1}$ so that

$$\rho = 0.5 \Rightarrow \beta = 1.73, \qquad \rho = 0.9 \Rightarrow \beta = 0.48.$$

How to generate a sample with $\rho = -0.9$ using Matlab:

```
X=randn(20,1);
Z=randn(20,1);
Y=-X+0.48*Z;
r=corrcoeff(X,Y)
```

How to generate a sample with $\rho = 0$ using Matlab:

```
X=randn(20,1);
Y=randn(20,1);
r=corrcoeff(X,Y)
```

Simulation results

ρ	-0.9	-0.5	0	0.5	0.9
r	-0.92	-0.45	-0.20	0.32	0.92

Problem 14.42

The stopping distance of an automobile on a certain road was studied as a function of velocity (Brownee 1960)

velocity of a car x (mi/h)	20.5	20.5	30.5	40.5	48.8	57.8
stopping distance y (ft)	15.4	13.3	33.9	73.1	113.0	142.6

Fit y and $\sqrt{y}$ as linear functions of velocity, and examine the residuals in each case. Which fit is better? Can you suggest any physical reason that explains why?

Solution 14.42

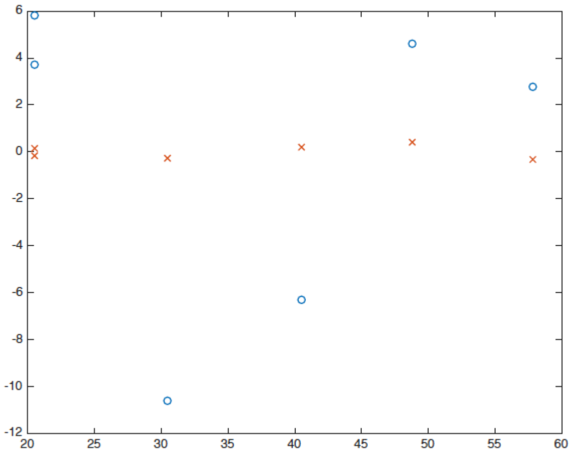
Matlab commands (x and y are columns)

```
[b,bint,res,rint,stats]=regress(y,[ones(6,1),x])
[b,bint,res,rint,stats]=regress(sqrt(y),[ones(6,1),x])
```

give two sets of residuals - see the plot. Two simple linear regression models

$$y = -62.05 + 3.49 \cdot x, \quad r^2 = 0.984,$$
$$\sqrt{y} = -0.88 + 0.2 \cdot x, \quad r^2 = 0.993.$$

Kinetic energy formula explains why the second model is better.



Problem A.1

Suppose we agiven a two-dimensional iid-sample

$$(x_1, y_1), \ldots, (x_n, y_n).$$

Verify that the sample covariance is an unbiased estimate of the population covariance.

Solution A.1

Recall that the sample covariance and the population covariance are

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad \mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

It is enough to check that

$$\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})\right)=(n-1)\mathrm{E}(XY)-(n-1)\mathrm{E}(X)\mathrm{E}(Y).$$

To do this, observe that

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+n\bar{x}\bar{y}=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y},$$

and

$$n^2\bar{x}\bar{y}=\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}x_iy_i+\sum_{i\neq j}\sum_{j=1}^{n}x_iy_j,$$

so that

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{n-1}{n}\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\sum_{i\neq j}\sum_{j=1}^{n}x_iy_j.$$

It remains to see that

$$\mathrm{E}\left(\sum_{i=1}^{n}X_iY_i\right)=n\mathrm{E}(XY),\qquad \mathrm{E}\left(\sum_{i\neq j}\sum_{j=1}^{n}X_iY_j\right)=n(n-1)\mathrm{E}(X)\mathrm{E}(Y).$$