Exercises in Introduction to Mathematical Statistics (Ch. 4)

Tomoki Okuno

October 29, 2022

Note

- Not all solutions are provided: Exercises that are too simple or not very important to me are skipped.
- Texts in red are just attentions to me. Please ignore them.

4 Some Elementary Statistical Inferences

4.1 Sampling and Statistics

4.1.1. Twenty motors were put on test under a high-temperature setting. The lifetimes in hours of the motors under these conditions are given below. Also, the data are in the file `lifetimemotor.rda` at the site listed in the Preface. Suppose we assume that the lifetime of a motor under these conditions, X, has a $\Gamma(1, \theta)$ distribution.

1	4	5	21	22	28	40	42	51	53
58	67	95	124	124	160	202	260	303	363

(a) Obtain a histogram of the data and overlay it with a density estimate, using the code `hist(x,pr=T); lines(density(x))` where the R vector `x` contains the data. Based on this plot, do you think that the $\Gamma(1, \theta)$ model is credible?

Solution.

```
load(".../lifetimemotor.rda")
hist(lifetimemotor$lifetime, pr = T)
lines(density(lifetimemotor$lifetime))}
```

We see that $\Gamma(1, \theta)$ model is credible.

(b) Assuming a $\Gamma(1, \theta)$ model, obtain the maximum likelihood estimate $\widehat{\theta}$ of θ and locate it on your histogram. Next overlay the pdf of a $\Gamma(1, \widehat{\theta})$ distribution on the histogram. Use the R function `dgamma(x,shape=1,scale=`$\widehat{\theta}$`)` to evaluate the pdf.

Solution.

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0 \quad \Rightarrow \quad \ell(\theta) = -n \log \theta - \frac{\sum_i x_i}{\theta}$$

$$\ell'(\theta) = -\frac{n}{\theta} + \frac{\sum_i x_i}{\theta^2} = \frac{n(\overline{x} - \theta)}{\theta^2}, \quad \ell''(\theta) = \frac{n}{\theta^2} - \frac{2\sum_i x_i}{\theta^3}.$$

Solving $\ell'(\theta) = 0$, we obtain $\widehat{\theta} = \overline{X} = 101.15$. Note that $\ell''(\widehat{\theta}) = \ell''(\overline{X}) < 0$. We can overlay the pdf using the following code:

```
theta.mle = mean(lifetimemotor$lifetime)
x = seq(0, 400)
```

```
hist(lifetimemotor$lifetime, pr = T)
lines(dgamma(x, shape = 1, scale = theta.mle))},
```

which implies that the pdf of the $\Gamma(1, \theta)$ has a good fit to the data (histogram).

(c) Obtain the sample median of the data, which is an estimate of the median lifetime of a motor. What parameter is it estimating (i.e., determine the median of X)?

Solution.

The median is `median(lifetimemotor$lifetime)` $= 55.5$. Since the cdf of X is given by

$$F_X(x) = \int_0^x \frac{1}{\theta} e^{-t/\theta} dt = 1 - e^{-x/\theta}, \quad x > 0,$$

the median of X is a solution of $F_X(x_m) = \frac{1}{2}$, that is, $x_m = \theta \log 2$.

(d) Based on the mle, what is another estimate of the median of X?

Solution. `theta.mle * log(2)` $= 70.11$.

4.1.3. Suppose the number of customers X that enter a store between the hours 9:00 a.m. and 10:00 a.m. follows a Poisson distribution with parameter θ. Suppose a random sample of the number of customers that enter the store between 9:00 a.m. and 10:00 a.m. for 10 days results in the values

$$9 \quad 7 \quad 9 \quad 15 \quad 10 \quad 13 \quad 11 \quad 7 \quad 2 \quad 12$$

(a) Determine the maximum likelihood estimate of θ. Show that it is an unbiased estimator.

Solution.

$$f(x) = \frac{e^{-\theta}\theta^x}{x!}, \quad x = 0, 1, 2, ... \quad \Rightarrow \quad \ell(\theta) = -n\theta - \sum_i x_i \log\theta + n \log x_i!$$

$$\ell'(\theta) = -n + \frac{\sum_i x_i}{\theta}, \quad \ell''(\theta) = -\frac{\sum_i x_i}{\theta^2} < 0.$$

Solving $\ell'(\theta) = 0$, we obtain $\widehat{\theta} = \overline{X}$. Since $E(\widehat{\theta}) = E(\overline{X}) = E(X) = \theta$, it is unbiased for θ.

(b) Based on these data, obtain the realization of your estimator in part (a). Explain the meaning of this estimate in terms of the number of customers.

Solution.

`mean(c(9, 7, 9, 15, 10, 13, 11, 7, 2, 12))` $= 9.5$, which is interpreted as an estimated number of customers that enter the store between 9.00 a.m. and 10:00 a.m based on these data.

4.1.6. Consider the estimator of the pmf in expression (4.1.10). In equation (4.1.11), we showed that this estimator is unbiased. Find the variance of the estimator and its mgf.

Solution.

Since the Bernoulli distribution

$$I_j(X_i) = \begin{cases} 1 & P(X_i = a_j) = p(a_j) \\ 0 & P(X_i \neq a_j) = 1 - p(a_j) \end{cases}$$

is independent of each other, the variance and the mgf are

$$\begin{aligned} \mathrm{Var}[\widehat{p}(a_j)] &= \frac{\sum_{i=1}^n \mathrm{Var}[I_j(X_i)]}{n^2} = \frac{p(a_j)[1 - p(a_j)]}{n} \\ M_{I_j}(t) &= E(E^{tI_j}) = e^t p(a_j) + [1 - p(a_j)]. \end{aligned}$$

4.1.7. The data set on Scottish schoolchildren discussed in Example 4.1.5 included the eye colors of the children also. The frequencies of their eye colors are

Blue	Light	Medium	Dark
2978	6697	7511	5175

Solution.

Since the sample size is $n = 22,361$, the estimate of the pmf are

	Blue	Light	Medium	Dark
Count	2978	6697	7511	5175
$\widehat{p}(a_j)$	0.133	0.299	0.336	0.231

4.1.8. Recall that for the parameter $\eta = g(\theta)$, the mle of η is $g(\widehat{\theta})$, where $\widehat{\theta}$ is the mle of θ. Assuming that the data in Example 4.1.6 were drawn from a Poisson distribution with mean λ, obtain the mle of λ and then use it to obtain the mle of the pmf. Compare the mle of the pmf to the nonparametric estimate. Note: For the domain value 6, obtain the mle of $P(X \geq 6)$

Solution.

By 4.1.3, mle of λ is $\widehat{\lambda} = \overline{X}$. Hence the mle of the pmf is

$$\widehat{p(x)} = \frac{e^{-\overline{x}}\overline{x}^x}{x!}, \quad x = 0, 1, 2, ...$$

In this dataset, we obtain $\overline{X} = 64/30$. So, $\widehat{P(X \geq 6)} =$ `1 - ppois(5, 64/30)` $= 0.0219$, which is not so different from the nonparametric estimate of the pmf at $X \geq 6$ (0.033).

4.2 Confidence Intervals

4.2.1. Let the observed value of the mean $\overline{X}$ and of the sample variance of a random sample of size 20 from a distribution that is $N(\mu, \sigma^2)$ be 81.2 and 26.5, respectively. Find respectively 90%, 95% and 99% confidence intervals for μ. Note how the lengths of the confidence intervals increase as the confidence increases.

Solution.

$n = 20 < 25$, which implies that CLT may not be applied. Since the upper critical values are $t_{0.05,19} = 1.73$, $t_{0.025,19} = 2.09$ and $t_{0.005,19} = 2.86$, respectively, the desired CIs are

$$\overline{X} \pm t_{\alpha/2,19}\sqrt{\frac{s^2}{n}} = 81.2 \pm t_{\alpha/2,19}\sqrt{\frac{26.5}{20}} = \begin{cases} (79.21,\ 83.19) & \alpha = 0.1 \\ (78.79,\ 83.61) & \alpha = 0.05 \\ (77.91,\ 84.49) & \alpha = 0.01 \end{cases}.$$

4.2.2.

Consider the data on the lifetimes of motors given in Exercise 4.1.1. Obtain a large sample 95% confidence interval for the mean lifetime of a motor.

Solution.

By the following code, we obtain (54.95 147.35).

```
n = length(lifetimemotor$lifetime)
mean = mean(lifetimemotor\$lifetime)
sd = sd(lifetimemotor$lifetime)
c(mean - 1.96*sd/sqrt(n), mean + 1.96*sd/sqrt(n))}
```

Note: the answer key of the textbook (51.82, 150,48) should be wrong. This 95% CI is not a large sample (approximate) CI but the exact CI using $t_{0.025,19}$.

4.2.3. Suppose we assume that $X_1, X_2, \ldots, X_n$ is a random sample from a $\Gamma(1, \theta)$ distribution.

(a) Show that the random variable $(2/\theta)\sum_{i=1}^{n} X_i$ has a χ^2-distribution with $2n$ degrees of freedom.

Solution.

$$X \sim \Gamma(1,\theta) \Leftrightarrow M_X(t) = \frac{1}{1-\theta t},\ t < \frac{1}{\theta}.$$

Let $Y = (2/\theta)\sum_{i=1}^{n} X_i$, then

$$M_Y(t) = [M_X(2t/\theta)]^n = \frac{1}{(1-2t)^n},\ t < \frac{1}{2} \Leftrightarrow Y \sim \chi^2(2n).$$

(b) Using the random variable in part (a) as a pivot random variable, find a $(1-\alpha)100\%$ CI for θ.

Solution.

$$1-\alpha = P\left(\chi^2_{\alpha/2,2n} \le \frac{2}{\theta}\sum_{i=1}^{n} X_i \le \chi^2_{1-\alpha/2,2n}\right) = P\left(\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{1-\alpha/2,2n}} \le \theta \le \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{\alpha/2,2n}}\right).$$

Hence, a CI for θ is

$$\left[\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{1-\alpha/2,2n}},\ \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{\alpha/2,2n}}\right].$$

(c) Obtain the confidence interval in part (b) for the data of Exercise 4.1.1 and compare it with the interval you obtained in Exercise 4.2.2.

Solution.

Consider a 95% confidence interval to compare with the interval obtained in Exercise 4.2.2. The following code gives us (68.18, 165.60):

```
sum.x = sum(lifetimemotor$lifetime)
c(2*sum.x/qchisq(0.975, 2*n), 2*sum.x/qchisq(0.025, 2*n))
```

which has a wider length than (51.82, 150,48) we obtained Exercise 4.2.2. However, this result would be more reliable than the previous one because Y has exactly a $\chi^2(2n)$, while $\overline{X}$ has approximately a normal distribution.

4.2.6. Let X be the mean of a random sample of size n from a distribution that is $N(\mu, 9)$. Find n such that $P(\overline{X} - 1 < \mu < \overline{X} + 1) = 0.90$, approximately.

Solution. $n = (z_{0.05}\sigma)^2 = [3(1.645)]^2 = 24.35$. Thus, $n = 25$ (round-up).

4.2.7. Let a random sample of size 17 from the normal distribution $N(\mu,\sigma^2)$yield $\overline{x} = 4.7$ and $s^2 = 5.76$. Determine a 90% confidence interval for μ.

Solution.

$$\left[\overline{x} \pm t_{0.05,n-1}\frac{s}{\sqrt{n}}\right] = \left[4.7 \pm 1.746\frac{\sqrt{5.76}}{\sqrt{17}}\right] = (3.68, 5.72).$$

4.2.8. Let $\overline{X}$ denote the mean of a random sample of size n from a distribution that has mean μ and variance $\sigma^2 = 10$. Find n so that the probability is approximately 0.954 that the random interval $(\overline{X} - \frac{1}{2}, \overline{X} + \frac{1}{2})$ includes μ.

Solution.

$$0.954 = P\left(\overline{X} - \frac{1}{2} < \mu < \overline{X} + \frac{1}{2}\right) \Rightarrow 1 - 0.046 = P\left(-\frac{\sqrt{n}}{2\sigma} < \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} < \frac{\sqrt{n}}{2\sigma}\right)$$

Since $z_{0.046/2} = z_{0.023} = 1.995$, $1.995 = \sqrt{n}/(2\sqrt{10}) \Rightarrow n = 159.3$; take $n = 160$ (round-up).

4.2.10. Let $X_1, X_2, ..., X_n, X_{n+1}$ be a random sample of size $n+1$, $n > 1$, from a distribution that is $N(\mu, \sigma^2)$. Let $\overline{X} = \sum_1^n X_i/n$ and $S^2 = \sum_1^n (X_i - \overline{X})^2/(n-1)$. Find the constant c so that the statistic $c(\overline{X} - X_{n+1})/S$ has a t-distribution. If $n = 8$, determine k such that $P(\overline{X} - kS < X_9 < \overline{X} + kS) = 0.80$. The observe interval $\bar{x} - ks, \bar{x} + ks)$ is often called an 80% **prediction interval** for X_9.

Solution.

Since $\overline{X} \sim N(\mu, \sigma^2/n)$ and $\overline{X}$ and X_{n+1} are independent,

$$\overline{X} - X_{n+1} \sim N\left(0, \frac{\sigma^2}{n} + \sigma^2\right) \Rightarrow \sqrt{\frac{n}{n+1}}\left(\frac{\overline{X} - X_{n+1}}{\sigma}\right) \sim N(0,1).$$

Also, since we have $\overline{X}$ and S^2 are independent, $\overline{X} - X_{n+1}$ and S^2 are also independent:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Hence,

$$\frac{\sqrt{n/(n+1)}(\overline{X} - X_{n+1})/\sigma}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}} = \sqrt{\frac{n}{n+1}}\left(\frac{\overline{X} - X_{n+1}}{S}\right) \sim t_{n-1},$$

which gives us $c = \sqrt{n/(n+1)}$.

If $n = 8$, since $\alpha = 0.2$, $t_{n-1,\alpha/2} = t_{7,0.1} = 1.415$. Hence,

$$\begin{aligned} 0.80 &= P\left(-1.415 < \sqrt{\frac{8}{9}}\left(\frac{\overline{X} - X_9}{S}\right) < 1.415\right) \\ &= P\left(\overline{X} - 1.415\sqrt{\frac{9}{8}}S < X_9 < \overline{X} + 1.415\sqrt{\frac{9}{8}}S\right), \end{aligned}$$

so $k = 1.415(3/\sqrt{8}) = 1.50$.

4.2.14. Let $\overline{X}$ denote the mean of a random sample of size 25 from a gamma-type distribution with $\alpha = 4$ and $\beta > 0$. Use the Central Limit Theorem to find an approximate 0.954 confidence interval for μ, the mean of the gamma distribution.

Solution.

Use the CLT to obtain $\overline{X}$ is approximately normal distributed $N(4\beta, 4\beta^2/25)$. Since $z_{0.023} = 2$,

$$\begin{aligned} 0.954 &= P\left(-2 < \frac{\overline{X} - 4\beta}{2\beta/5} < 2\right) \\ &= P\left(-2 < \frac{5\overline{X}}{2\beta} - 10 < 2\right) \\ &= P\left(8 < \frac{5\overline{X}}{2\beta} < 12\right) \\ &= P\left(\frac{5\overline{X}}{6} < 4\beta < \frac{5\overline{X}}{4}\right). \end{aligned}$$

Hence, an approximate 0.954 confidence interval for $\mu = 4\beta$ is

$$\left(\frac{5\overline{X}}{6}, \frac{5\overline{X}}{4}\right).$$

Note that the answer key in the textbook is an approximate 0.954 CI for β, which is incorrect.

4.2.15. Let $\overline{x}$ be the observed mean of a random sample of size n from a distribution having mean μ and known variance σ^2. Find n so that $\overline{x} - \sigma/4$ to $\overline{x} + \sigma/4$ is an approximate 95% confidence interval for μ.

Solution. Since an approximate 95% CI for μ is $[\overline{x} \pm 1.96\sigma/\sqrt{n}]$, $n = [4(1.96)] = 61.5$; take $n = 62$.

4.2.16. Assume a binomial model for a certain random variable. If we desire a 90% confidence interval for p that is at most 0.02 in length, find n.

Solution.

Let $\widehat{p}$ denote the point estimate of p. Since $z_{0.05} = 1.645$, the length is

$$2(1.645)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq 3.29\sqrt{\frac{1}{4n}} = 0.02 \Rightarrow n = \left(\frac{3.29}{0.02}\right)^2 \frac{1}{4} \approx 6465.$$

4.2.17. It is known that a random variable X has a Poisson distribution with parameter μ. A sample of 200 observations from this distribution has a mean equal to 3.4. Construct an approximate 90% confidence interval for μ.

Solution.

$$\left(\overline{X} - 1.645\sqrt{\frac{\overline{X}}{n}},\ \overline{X} + 1.645\sqrt{\frac{\overline{X}}{n}}\right) = \left(3.4 - 1.645\sqrt{\frac{3.4}{200}},\ 3.4 + 1.645\sqrt{\frac{3.4}{200}}\right) = (3.19, 3.61).$$

4.2.18. Skipped, but for (c), use the fact that

$$\sum_1^n \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_1^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n).$$

4.2.19. Let $X_1, X_2, ..., X_n$ be a random sample from a gamma distribution with known parameter $\alpha = 3$ and unknown $\beta > 0$. In Exercise 4.2.14, we obtained an approximate confidence interval for β (note: actually 4β) based on the Central Limit Theorem. In this exercise obtain an exact confidence interval by first obtaining the distribution of $2\sum_1^n X_i/\beta$.

Solution.

As with Exercise 4.2.14, we have $2\sum_1^n X_i/\beta \sim \chi^2(6n)$. Hence,

$$\begin{aligned} 0.95 &= P\left(\chi^2_{0.025,6n} < \frac{2}{\beta}\sum_1^n X_i < \chi^2_{0.975,6n}\right) \\ &= P\left(\frac{2\sum_1^n X_i}{\chi^2_{0.975,6n}} < \beta < \frac{2\sum_1^n X_i}{\chi^2_{0.025,6n}}\right), \end{aligned}$$

which means that the exact 95% CI for β is

$$\left[\frac{2\sum_1^n X_i}{\chi^2_{0.975,6n}},\ \frac{2\sum_1^n X_i}{\chi^2_{0.025,6n}}\right].$$

4.2.20. When 100 tacks were thrown on a table, 60 of them landed point up. Obtain a 95% confidence interval for the probability that a tack of this type lands point up. Assume independence.

Solution.

Let $\widehat{p}$ denote the point estimate of p,

$$\left(\widehat{p} \pm z_{0.975}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = \left(0.6 \pm 1.96\sqrt{\frac{0.6(0.4)}{100}}\right) = (0.504, 0.696).$$

4.2.21. Let two independent random samples, each of size 10, from two normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ yield $\bar{x} = 4.8$, $s_1^2 = 8.64$, $\bar{y} = 5.6$, $s_2^2 = 7.88$. Find a 95% confidence interval for $\mu_1 - \mu_2$.

Solution.

Since the two variances are equal but unknown, we use the pooled estimator of σ^2, which is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{9(8.64 + 7.88)}{18} = 8.26 \;\Rightarrow\; s_p = 2.874.$$

Thus a 95% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x} - \bar{y}) \pm t_{0.025,18}(s_p)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = -0.8 \pm 2.10(2.874)\sqrt{\frac{1}{5}} = (-3.50, 1.90).$$

Note: The answer in the textbook is incorrect, where $z_{0.025}$ seems to be used instead of $t_{0.025,18}$ by mistake.

4.2.22. Let two independent random variables, Y_1 and Y_2, with binomial distributions that have parameters $n_1 = n_2 = 100$, p_1, and p_2, respectively, be observed to be equal to $y_1 = 50$ and $y_2 = 40$. Determine an approximate 90% confidence interval for $p_1 - p_2$.

Solution.

Since the two point estimates are $\widehat{p_1} = 0.5$ and $\widehat{p_2} = 0.4$, the desired CI is

$$\widehat{p_1} - \widehat{p_2} \pm z_{0.05}\sqrt{\frac{\widehat{p_1}(1 - \widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1 - \widehat{p_2})}{n_2}} = 0.1 \pm 1.645\frac{0.7}{10} = (-0.015, 0.215).$$

4.2.23. Discuss the problem of finding a confidence interval for the difference $\mu_1 - \mu_2$ between the two means of two normal distributions if the variances σ_1^2 and σ_2^2 are known but not necessarily equal.

Solution.

When σ_1^2 and σ_2^2 are known, finding a confidence interval for the difference $\mu_1 - \mu_2$ is straightforward:

$$\begin{aligned} 1 - \alpha &= P\left(-z_{\alpha/2} < \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} < z_{\alpha/2}\right) \\ &= P\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right). \end{aligned}$$

4.2.24. Discuss Exercise 4.2.23 when it is assumed that the variances are unknown and unequal. This is a very difficult problem, and the discussion should point out exactly where the difficulty lies. If, however, the variances are unknown but their ratio σ_1^2/σ_2^2 is a known constant k, then a statistic that is a T random variable can again be used. Why?

Solution.

When it is assumed that the variances are unknown and unequal, we cannot eliminate their unknown variances in a T statistic. If we can assume $\sigma_1^2 = k\sigma_2^2$ instead of $\sigma_1^2 = \sigma_2^2$, however, then,

$$\overline{X} - \overline{Y} \sim N\left(0,\; \sigma_2^2\left(\frac{k}{n} + \frac{1}{m}\right)\right)$$

$$\frac{(n-1)S_1^2}{\sigma_1^2} + \frac{(m-1)S_2^2}{\sigma_2^2} = \frac{(n-1)S_1^2/k + (m-1)S_2^2}{\sigma_2^2} \sim \chi^2(n - m - 2),$$

which implies that we can eliminate σ_2^2 in a T statistic. Accordingly, the pooled estimator of σ_2^2 is given by

$$s_p^2 = \frac{(n-1)S_1^2/k + (m-1)S_2^2}{n - m - 2}.$$

4.2.26. Let X and Y be the means of two independent random samples, each of size n, from the respective distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, where the common variance is known. Find n such that

$$P(\bar{X} - \bar{Y} - \sigma/5 < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + \sigma/5) = 0.90.$$

Solution.

$$\begin{aligned} 0.90 &= P(\bar{X} - \bar{Y} - \sigma/5 < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + \sigma/5) \\ &= P(-\sigma/5 < (\bar{X} - \bar{Y}) - (\mu_1 - \mu_2) < \sigma/5) \\ &= P\left(-\frac{\sqrt{n}}{5\sqrt{2}} < \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{2/n}} < \frac{\sqrt{n}}{5\sqrt{2}}\right), \end{aligned}$$

which gives

$$\frac{\sqrt{n}}{5\sqrt{2}} = z_{0.05} = 1.645 \quad \Rightarrow \quad n = [5\sqrt{2}(1.645)]^2 = 135.30.$$

Thus, $n = 136$ suffices.

4.4 Order Statistics

4.4.5. Le $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size 4 from the distribution having pdf $f(x) = e^{-x}, 0 < x < \infty$, zero elsewhere. Find $P(Y_4 \geq 3)$.

Solution.

Since the cdf $F(x) = 1 - e^{-x}$, $x > 0$,

$$f_{Y_4}(y) = \frac{4!}{(4-1)!(4-4)!} F(x)^3 f(x) = \begin{cases} 4(1-e^{-y})^3 e^{-y}, & 0 < y < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

Thus,

$$P(Y_4 \geq 3) = \int_3^\infty f_{Y_4}(y)dy = \int_3^\infty 4(1-e^{-y})^3 e^{-y} dy = [(1-e^{-y})^4]_3^\infty = 1 - (1-e^{-3})^4.$$

4.4.7. Let $f(x) = \frac{1}{6}$, $x = 1, 2, 3, 4, 5, 6$, zero elsewhere, be the pmf of a distribution of the discrete type. Show that the pmf of the smallest observation of a random sample of size 5 from this distribution is

$$g_1(y_1) = \left(\frac{7 - y_1}{6}\right)^5 - \left(\frac{6 - y_1}{6}\right)^5, \quad y = 1, 2, 3, 4, 5, 6,$$

zero elsewhere. Note that in this exercise the random sample is from a distribution of the discrete type. All formulas in the text were derived under the assumption that the random sample is from a distribution of the continuous type and are not applicable. Why?

Solution.

Since the pmf of X is $F(x) = x/6$, $x = 1, 2, ..., 6$, the pmf of Y_1 is

$$\begin{aligned} g_1(y_1) &= P(Y_1 \geq y_1) - P(Y_1 \geq y_1 + 1) \\ &= P(X_i \geq y_1,\ i = 1, ..., 5) - P(X_i \geq y_1 + 1,\ i = 1, ..., 5) \\ &= [P(X \geq y_1)]^5 - [P(X \geq y_1 + 1)]^5 \\ &= [1 - P(X \leq y_1 - 1)]^5 - [1 - P(X \leq y_1)]^5 \\ &= \left(1 - \frac{y_1 - 1}{6}\right)^5 - \left(1 - \frac{y_1}{6}\right)^5 \\ &= \left(\frac{7 - y_1}{6}\right)^5 - \left(\frac{6 - y_1}{6}\right)^5 \quad y = 1, 2, ..., 6. \end{aligned}$$

4.4.8. Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ denote the order statistics of a random sample of size 5 from a distribution having pdf $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Show that $Z_1 = Y_2$ and $Z_2 = Y_4 - Y_2$ are independent.

Solution.

Since $F_X(x) = 1 - e^{-x}$, the joint pdf of Y_2 and Y_4 is

$$\begin{aligned} f_{Y_2,Y_4}(y_2, y_4) &= \frac{5!}{1!1!1!} F(y_2)[F(y_4) - F(y_2)][1 - F(y_4)] f(y_2) f(y_4) \\ &= 120(1 - e^{-y_2})(e^{-y_2} - e^{-y_4}) e^{-y_2} e^{-2y_4}. \end{aligned}$$

The inverse functions are $y_2 = z_1$ and $y_4 = z_1 + z_2$, so the $J = 1$. Hence, the joint pdf of Z_1 and Z_2 is

$$\begin{aligned} g_{Z_1,Z_2}(z_1, z_2) &= f_{Y_2,Y_4}(z_1, z_1 + z_2)|J| \\ &= 120(1 - e^{-z_1})(e^{-z_1} - e^{-z_1 - z_2}) e^{-z_1} e^{-2(z_1+z_2)} \\ &= 120(1 - e^{-z_1}) e^{-z_1} (1 - e^{-z_2}) e^{-3z_1} e^{-2z_2} \\ &= 120(1 - e^{-z_1}) e^{-4z_1} (1 - e^{-z_2}) e^{-2z_2}, \end{aligned}$$

which can be expressed as a product of a marginal function of Z_1 and a marginal function of Z_2. Thus, Z_1 and Z_2 are independent.

4.4.9. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from a distribution with pdf $f(x) = 1, 0 < x < 1$, zero elsewhere. Show that the kth order statistic Y_k has a beta pdf with parameters $\alpha = k$ and $\beta = n - k + 1$.

Solution.

$$\begin{aligned} f_{Y_k}(y) &= \frac{n!}{(k-1)!(n-k)!} y^{k-1}(1-y)^{n-k} \\ &= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} y^{k-1}(1-y)^{n-k}. \end{aligned}$$

which means $Y_k \sim \text{Beta}(k, n-k+1)$.

4.4.10. Let $Y + 1 < Y_2 < \cdots < Y_n$ be the order statistics from a Weibull distribution, Exercise 3.3.26. Find the distribution function and pdf of Y_1.

Solution.

We have a Weibull distribution: $f(x) = cx^b \exp[-cx^{b+1}/(b+1)],\ x > 0$. Since the cdf is

$$F_X(x) = \int_0^x ct^b \exp[-ct^{b+1}/(b+1)]dt = 1 - \exp[-cx^{b+1}/(b+1)],\ x > 0,$$

the pdf of Y_1 is

$$\begin{aligned} f_{Y_1}(y_1) &= \frac{n!}{0!(n-1)!}[1 - F_X(y_1)]^{n-1} f_X(y_1) \\ &= n \exp\left[-\frac{(n-1)cx^{b+1}}{b+1}\right] cx^b \exp\left[-\frac{cx^{b+1}}{b+1}\right] \\ &= ncx^b \exp\left[-\frac{ncx^{b+1}}{b+1}\right], \end{aligned}$$

which indicates that Y_1 also has a Weibull distribution.

4.4.11. Find the probability that the range of a random sample of size 4 from the uniform distribution having the pdf $f(x) = 1, 0 < x < 1$, zero elsewhere, is less than $1/2$.

Solution.

$$f_{Y_1,Y_4}(y_1,y_4) = 12(y_4-y_1)^2,\ 0<y_1<y_4<1$$

and zero elsewhere. Hence

$$\begin{aligned}
P(Y_4-Y_1<1/2) &= P(Y_4<Y_1+1/2)\\
&= \int_0^{1/2}\int_{y_1}^{y_1+1/2} 12(y_4-y_1)^2 dy_4 dy_1 + \int_{1/2}^{1}\int_{y_1}^{1} 12(y_4-y_1)^2 dy_4 dy_1\\
&= \int_0^{1/2} 4(y_4-y_1)^3\Big|_{y_1}^{y_1+1/2} dy_1 + \int_{1/2}^{1} 4(y_4-y_1)^3\Big|_{y_1}^{1} dy_1\\
&= \int_0^{1/2}\frac{1}{2}dy_1 + \int_{1/2}^{1} 4(1-y_1)^3 dy_1\\
&= \frac{1}{4}+\frac{1}{16}=\frac{5}{16}.
\end{aligned}$$

4.4.13. Suppose a random sample of size 2 is obtained from a distribution that has pdf $(x) = 2(1-x), 0 < x < 1$, zero elsewhere. Compute the probability that one sample observation is at least twice as large as the other.

Solution.

Let $Y_1 < Y_2$ be the order statistics of X_1, X_2.

$$f_{Y_1,Y_2}(y_1,y_2) = 8(1-y_1)(1-y_2).$$

Then

$$P(Y_2 \geq 2Y_1) = \int_0^1\int_0^{y_2/2} 8(1-y_1)(1-y_2)dy_1dy_2 = \cdots = \frac{7}{12}.$$

4.4.14. Let $Y_1 < Y_2 < Y_3$ denote the order statistics of a random sample of size 3 from a distribution with pdf $f(x) = 1, 0 < x < 1$, zero elsewhere. Let $Z = (Y_1 + Y_3)/2$ be the midrange of the sample. Find the pdf of Z.

Solution.

$$f_{Y_1,Y_3}(y_1,y_3) = 6(y_3-y_1),\ 0<y_1<y_3<1$$

Let $W = Y_1$ in addition to Z, which isone to one transformation. Then, since $y_1 = w, y_3 = 2z - w$, the Jacobian is 2. Thus

$$f_{z,w}(z,w) = f_{Y_1,Y_3}(w,2z-w)|J| = 24(z-w),\ 0<w<2z-w<1.$$

and zero elsewhere. Note that the support of Z, then the pdf of Z is

$$f_z(z) = \begin{cases} \int_0^z 24(z-w) = 12z^2 & 0<z<1/2\\ \int_{2z-1}^z 24(z-w) = 12(1-z)^2 & 1/2<z<1\\ 0 & \text{otherwise}\end{cases}.$$

4.4.15. Let $Y_1 < Y_2$ denote the order statistics of a random sample of size 2 from $N(0, \sigma^2)$.

(a) Show that $E(Y_1) = -\sigma/\sqrt{\pi}$.

Solution.

$$
\begin{aligned}
E(Y_1) = \int_{-\infty}^{\infty} y_1 f(y_1) dy_1 &= \int_{-\infty}^{\infty} y_1 \left[\int_{y_1}^{\infty} f(y_1, y_2) dy_2 \right] dy_1 \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{y_2} y_1 f(y_1, y_2) dy_1 \right] dy_2 \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{y_2} \frac{y_1}{\pi\sigma^2} e^{-\frac{y_1^2+y_2^2}{2\sigma^2}} dy_1 \right] dy_2 \\
&= \int_{-\infty}^{\infty} \left[-\frac{1}{\pi} e^{-\frac{y_1^2+y_2^2}{2\sigma^2}} \right]_{-\infty}^{y_2} dy_2 \\
&= -\int_{-\infty}^{\infty} \frac{1}{\pi} e^{-\frac{y_2^2}{\sigma^2}} dy_2 \\
&= -\frac{\sigma}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(\sigma^2/2)}} e^{-\frac{y_2^2}{\sigma^2}} dy_2 \\
&= -\frac{\sigma}{\sqrt{\pi}}.
\end{aligned}
$$

(b) Find the covariance of Y_1 and Y_2.

Solution.

$$
\begin{aligned}
E(Y_2) = \int_{-\infty}^{\infty} y_2 f(y_2) dy_2 &= \int_{-\infty}^{\infty} y_2 \left[\int_{-\infty}^{y_2} f(y_1, y_2) dy_1 \right] dy_2 \\
&= \int_{-\infty}^{\infty} \left[\int_{y_1}^{\infty} y_2 f(y_1, y_2) dy_1 \right] dy_2 \\
&= \int_{-\infty}^{\infty} \left[\int_{y_1}^{\infty} \frac{y_2}{\pi\sigma^2} e^{-\frac{y_1^2+y_2^2}{2\sigma^2}} dy_1 \right] dy_2 \\
&= \int_{-\infty}^{\infty} \left[-\frac{1}{\pi} e^{-\frac{y_1^2+y_2^2}{2\sigma^2}} \right]_{y_1}^{\infty} dy_2 \\
&= \int_{-\infty}^{\infty} \frac{1}{\pi} e^{-\frac{y_1^2}{\sigma^2}} dy_2 \\
&= \frac{\sigma}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(\sigma^2/2)}} e^{-\frac{y_1^2}{\sigma^2}} dy_2 \\
&= \frac{\sigma}{\sqrt{\pi}}, \\
E(Y_1 Y_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{y_2} y_1 y_2 f(y_1, y_2) dy_1 dy_2 \\
&= \int_{-\infty}^{\infty} y_2 \int_{-\infty}^{y_2} \frac{y_1}{\pi\sigma^2} e^{-\frac{y_1^2+y_2^2}{2\sigma^2}} dy_1 dy_2 \\
&= -\int_{-\infty}^{\infty} \frac{y_2}{\pi} e^{-\frac{y_2^2}{\sigma^2}} dy_2 \\
&= 0.
\end{aligned}
$$

Hence, the covariance of Y_1 and Y_2 is

$$
\operatorname{Cov}(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2) = \frac{\sigma^2}{\pi}.
$$

4.4.17. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size $n = 4$ from a distribution with pdf $f(x) = 2x, 0 < x < 1$, zero elsewhere.

(a) Find the joint pdf of Y_3 and Y_4.

(b) Find the conditional pdf of Y_3, given $Y_4 = y_4$.

(c) Evaluate $E(Y_3|y_4)$.

Solution.

(a) $f_{Y_3,Y_4}(y_3, y_4) = (4!/2!)F(y_3)^2 f(y_3)f(y_4) = 12(y_3^2)^2(2y_3)(2y_4) = 48y_3^5 y_4,\ 0 < y_3 < y_4 < 1.$

(b) Since $f_{y_4}(y_4) = 4y_4^6(2y_4) = 8y_4^7$,

$$f_{Y_3|Y_4}(y_3|y_4) = \frac{f_{Y_3,Y_4}(y_3, y_4)}{f_{y_4}(y_4)} = \frac{6y_3^5}{y_4^6},\ 0 < y_3 < y_4.$$

(c)

$$E(Y_3|y_4) = \int_0^{y_4} y_3 \frac{6y_3^5}{y_4^6} dy_3 = \int_0^{y_4} \frac{6y_3^6}{y_4^6} dy_3 = \frac{6}{7} y_4.$$

4.4.18. Two numbers are selected at random from the interval (0, 1). If these values are uniformly and independently distributed, by cutting the interval at these numbers, compute the probability that the three resulting line segments can form a triangle.

Solution.

Let X_1 and X_2 denote the two numbers that are $U(0, 1)$ and $Y_1 < Y_2$ denote the order statistics. Then, the joint pdf of Y_1 and Y_2 is

$$f_{Y_1,Y_2}(y_1, y_2) = (2!/0!0!0!)f_X(y_1)f_X(y_2) = 2, \quad 0 < y_1 < y_2 < 1.$$

The conditions under which three resulting line segments can form a triangle are

$$y_1 < 1 - y_1 \Rightarrow y_1 < \frac{1}{2},$$
$$y_2 - y_1 < 1 - (y_2 - y_1) \Rightarrow y_2 - y_1 < \frac{1}{2},$$
$$y_2 > 1 - y_2 \Rightarrow y_2 > \frac{1}{2},$$

which is the support to compute the probability:

$$\int_0^{1/2} \int_{1/2}^{y_1+1/2} 2dy_2 dy_1 = \int_0^{1/2} 2y_1 dy_1 = \frac{1}{4}.$$

4.4.19. Let X and Y denote independent random variables with respective probability density functions $f(x) = 2x$, $0 < x < 1$, zero elsewhere, and $g(y) = 3y^2$, $0 < y < 1$, zero elsewhere. Let $U = \min(X, Y)$ and $V = \max(X, Y)$. Find the joint pdf of U and V.

Solution.

Since X and Y are independent, we have the joint pdf of X and Y:

$$f_{X,Y}(x, y) = f(x)g(y) = 6xy^2,\ 0 < x < 1,\ 0 < y < 1.$$

Note that the transformation is not one-to-one:

$$(1)\ x = u,\ y = v \quad \text{and} \quad (2)\ x = v,\ y = u.$$

Then, the Jacobians are $J_1 = 1$ and $J_2 = -1$, respectively. Thus, the joint pdf of U and V is

$$\begin{aligned} f_{U,V}(uv) &= f_{X,Y}(u,v)|J_1| + f_{X,Y}(v,u)|J_2| \\ &= 6uv^2(1) + 6vu^2(1) \\ &= 6uv(u+v), \quad 0 < u < v < 1. \end{aligned}$$

4.4.20. Let the joint pdf of X and Y be $f(x,y) = \frac{12}{7}x(x+y)$, $0 < x < 1, 0 < y < 1$, zero elsewhere. Let $U = \min(X,Y)$ and $V = \max(X,Y)$. Find the joint pdf of U and V.

Solution.

As with the previous exercise,

$$\begin{aligned} f_{U,V}(uv) &= f_{X,Y}(u,v)|J_1| + f_{X,Y}(v,u)|J_2| \\ &= \frac{12}{7}u(u+v) + \frac{12}{7}v(v+u) \\ &= \frac{12}{7}(u+v)^2, \quad 0 < u < v < 1. \end{aligned}$$

4.4.22. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from the exponential distribution with pdf $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere.

(a) Show that $Z_1 = nY_1$, $Z_2 = (n-1)(Y_2 - Y_1)$, $Z_3 = (n-2)(Y_3 - Y_2)$,..., $Z_n = Y_n - Y_{n-1}$ are independent and that each Z_i has the exponential distribution.

Solution.

The inverse transformation is

$$y_1 = \frac{z_1}{n},\ y_2 = \frac{z_1}{n} + \frac{z_2}{n-1},\ y_3 = \frac{z_1}{n} + \frac{z_2}{n-1} + \frac{z_3}{n-2},\ \ldots\ , y_n = \frac{z_1}{n} + \frac{z_2}{n-1} + \cdots + z_n,$$

which implits that $J = 1/n!$. By theorem 4.4.1, hence, the joint pdf of Z_i's is

$$\begin{aligned} f_{Z_1,\ldots,Z_n}(z_1,\ldots,z_n) &= f_{Y_1,\ldots,Y_n}\left(\frac{z_1}{n}, \frac{z_1}{n} + \frac{z_2}{n-1}, \ldots, \frac{z_1}{n} + \frac{z_2}{n-1} + \cdots + z_n\right)|J| \\ &= n! f_X\left(\frac{z_1}{n}\right) f_X\left(\frac{z_1}{n} + \frac{z_2}{n-1}\right) \cdots f_X\left(\frac{z_1}{n} + \frac{z_2}{n-1} + \cdots + z_n\right)\frac{1}{n!} \\ &= e^{-z_1 - z_2 - \cdots - z_n} \\ &= f_X(z_1)f_X(z_2)\cdots f_X(z_n), \end{aligned}$$

which is the desired result.

(b) Demonstrate that all linear functions of $Y_1, Y_2, ..., Y_n$, such as $\sum_1^n a_i Y_i$, can be expressed as linear functions of independent random variables.

Solution.

By part (a), we can transform Y_i's that are dependent to Z_i's that are independent each other.

$$\sum_{i=1}^n a_i Y_i = \sum_{i=1}^n a_i \sum_{j=1}^i \frac{Z_j}{n-j+1} = \sum_{i=1}^n \sum_{j=i}^n \frac{a_j}{n-j+1} Z_i \equiv \sum_{i=1}^n b_i Z_i,$$

which is a form of linear functions of independent random variables.

4.5 Introduction to Hypothesis Testing

4.5.1. Show that the approximate power function given in expression (4.5.12) of Example 4.5.3 is a strictly increasing function of μ. Show then that the test discussed in this example has approximate size α for testing

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0.$$

Solution.

Let $\phi(z)$ be a pdf of a standard normal random variable. The first derivative of $\gamma(\mu)$ with respect to μ,

$$\gamma'(\mu) = \phi\left(-z_\alpha - \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right)\frac{\sqrt{n}}{\sigma} > 0$$

because $\phi(x) > 0$, $n > 0$, and $\sigma > 0$. Hence, $\gamma(\mu)$ is strictly increasing function of μ.

Then, under $H_0 : \mu \leq \mu_0$,

$$\max_{\mu \leq \mu_0} \gamma(\mu) = \gamma(\mu_0) = \Phi(-z_\alpha) = \alpha,$$

which is the desired result.

4.5.2. For the Darwin data tabled in Example 4.5.5, verify that the Student t-test statistic is 2.15.

Solution.

$$t = \frac{\overline{x} - 0}{s_x/\sqrt{n}} = \frac{2.62 - 0}{4.72/\sqrt{15}} = 2.149.$$

4.5.3. Let X have a pdf of the form $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, zero elsewhere, where $\theta \in \theta : \theta = 1, 2$. To test the simple hypothesis $H_0 : \theta = 1$ against the alternative simple hypothesis $H_1 : \theta = 2$, use a random sample X_1, X_2 of size $n = 2$ and define the critical region to be $C = \{(x_1, x_2) : \frac{3}{4} \leq x_1x_2\}$. Find the power function of the test.

Solution. Since X_1 and X_2 are independent, $f(x_1, x_2) = \theta^2(x_1x_2)^{\theta-1}$. Hence the power function is

$$\gamma_C(\theta) = P_\theta\left(X_1X_2 \geq \frac{3}{4}\right) = \int_{3/4}^{1}\int_{3/(4x_1)}^{1} \theta^2(x_1x_2)^{\theta-1}dx_2dx_1 = \cdots = 1 - \left(\frac{3}{4}\right)^\theta + \theta\left(\frac{3}{4}\right)^\theta \log\frac{3}{4},\ \theta = 1, 2.$$

4.5.4. Let X have a binomial distribution with the number of trials $n = 10$ and with p either 1/4 or 1/2. The simple hypothesis $_0 : p = \frac{1}{2}$ is rejected, and the alternative simple hypothesis $H_1 : p = \frac{1}{4}$ is accepted, if the observed value of X_1, a random sample of size 1, is less than or equal to 3. Find the significance level and the power of the test.

Solution.

$$\begin{aligned}\alpha = P_{p=1/2}(X \leq 3) &= \texttt{pbinom(3, 10, 0.5)} = 0.172,\\ 1 - \beta = P_{p=1/4}(X \leq 3) &= \texttt{pbinom(3, 10, 0.25)} = 0.776.\end{aligned}$$

4.5.5. Let X_1, X_2 be a random sample of size $n = 2$ from the distribution having pdf $f(x;\theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, zero elsewhere. We reject $H_0 : \theta = 2$ and accept $H_1 : \theta = 1$ if the observed values of X_1, X_2, say x_1, x_2, are such that

$$\frac{f(x_1;2)f(x_2;2)}{f(x_1;1)f(x_2;1)} \leq \frac{1}{2}.$$

Here $\Omega = \{\theta : \theta = 1, 2\}$. Find the significance level of the test and the power of the test when H_0 is false.

Solution.

$$\frac{f(x_1;2)f(x_2;2)}{f(x_1;1)f(x_2;1)} \le \frac{1}{2} \quad \Leftrightarrow \quad \frac{1}{4}e^{(x_1+x_2)/2} \le \frac{1}{2} \quad \Leftrightarrow \quad x_1 + x_2 \le 2\log 2.$$

Also, we have $X \sim \Gamma(1,\theta) \Rightarrow Y = X_1 + X_2 \sim \Gamma(2,\theta)$. Hence,

$$\begin{aligned} P_\theta(Y \le 2\log 2) &= \int_0^{2\log 2} \frac{1}{\theta^2} x e^{-x/\theta} dx \\ &= [-xe^{-x/\theta}/\theta]_0^{2\log 2} + \int_0^{2\log 2} \frac{1}{\theta} e^{-x/\theta} dx \\ &= -\frac{2\log 2}{\theta} e^{2\log 2/\theta} + 1 - e^{-2\log 2/\theta} \\ &= 1 - \left(1 + \frac{2\log 2}{\theta}\right) e^{-2\log 2/\theta}. \end{aligned}$$

Hence,

$$\begin{aligned} \alpha = P_2(Y \le 2\log 2) = 1 - (1+\log 2)/2 = (1-\log 2)/2 \approx 0.1534, \\ 1 - \beta = P_1(Y \le 2\log 2) = 1 - (1 + 2\log 2)/4 = (3 - 2\log 2)/4 \approx 0.403. \end{aligned}$$

4.5.8. Let us say the life of a tire in miles, say X, is normally distributed with mean θ and standard deviation 5000. Past experience indicates that $\theta = 30,000$. The manufacturer claims that the tires made by a new process have mean $\theta > 30,000$. It is possible that $\theta = 35,000$. Check his claim by testing $H_0 : \theta = 30,000$ against $H_1 : \theta > 30,000$. We observe n independent values of X, say $x_1, ..., x_n$, and we reject H_0 (thus accept H_1) if and only if $\overline{x} \ge c$. Determine n and c so that the power function $\gamma(\theta)$ of the test has the values $\gamma(30,000) = 0.01$ and $\gamma(35,000) = 0.98$.

Solution.

We have two equations:

$$\begin{aligned} \gamma(30,000) = 0.01 \ &\Rightarrow \ P\left(\frac{\overline{X} - 30000}{5000/\sqrt{n}} \ge \frac{c - 30000}{5000/\sqrt{n}}\right) = 0.01 \ \Rightarrow \ \frac{c - 30000}{5000/\sqrt{n}} = 2.326, \\ \gamma(35,000) = 0.98 \ &\Rightarrow \ P\left(\frac{\overline{X} - 35000}{5000/\sqrt{n}} \ge \frac{c - 35000}{5000/\sqrt{n}}\right) = 0.98 \ \Rightarrow \ \frac{c - 35000}{5000/\sqrt{n}} = -2.054, \end{aligned}$$

which gives us $n \approx 20$ and $c \approx 32661$.

4.5.11. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size $n = 4$ from a distribution with pdf $f(x;\theta) = 1/\theta$, $0 < x < \theta$, zero elsewhere, where $0 < \theta$. The hypothesis $H_0 : \theta = 1$ is rejected and $H_1 : \theta > 1$ is accepted if the observed $Y_4 \ge c$.

(a) Find the constant c so that the significance level is $\alpha = 0.05$.

Solution.

$$f_{Y_4}(y_4) = \frac{4!}{3!0!} F_X(y_4)^3 f_X(y_4) = \frac{4y_4^3}{\theta^4}, \quad 0 < y_4 < \theta.$$

Hence,

$$\alpha = 0.05 = P_{\theta=1}(Y_4 \ge c) = \int_c^1 4y_4^3 dy_4 = 1 - c^4 \ \Rightarrow \ c = (0.95)^{1/4} = 0.9872.$$

(b) Determine the power function of the test.

Solution.

$$\gamma(\theta) = P_\theta(Y_4 \ge c) = \int_c^\theta \frac{4y_4^3}{\theta^4} dy_4 = 1 - \frac{c^4}{\theta^4} = 1 - \frac{0.95}{\theta^4}.$$

4.5.12. Let $X_1, X_2, ..., X_8$ be a random sample of size $n = 8$ from a Poisson distribution with mean μ. Reject the simple null hypothesis $H_0 : \mu = 0.5$ and accept $H_1 : \mu > 0.5$ if the observed $\sum_{i=1}^{8} x_i \geq 8$.

(a) Show that the significance level is `1-ppois(7,8*.5)`.

Solution.

Since $Y = \sum_{i=1}^{8} X_i \sim \text{Poisson}(8\mu)$, $\alpha = P_{0.5}(Y \geq 8) = 1 - P_{0.5}(Y \leq 7) =$ `1-ppois(7,8*.5)` $= 0.051$.

(b) Use R to determine $\gamma(0.75)$, $\gamma(1)$, and $\gamma(1.25)$.

Solution.

$$\begin{aligned}\gamma(0.75) &= \texttt{1-ppois(7,8*.75)} = 0.256,\\ \gamma(1) &= \texttt{1-ppois(7,8)} = 0.547,\\ \gamma(1.25) &= \texttt{1-ppois(7,8*1.25)} = 0.780.\end{aligned}$$

(c) Modify the code in Exercise 4.5.9 to obtain a plot of the power function.

Solution. Skipped.

4.5.13. Let p denote the probability that, for a particular tennis player, the first serve is good. Since $p = 0.40$, this player decided to take lessons in order to increase p. When the lessons are completed, the hypothesis $H_0 : p = 0.40$ is tested against $H_1 : p > 0.40$ based on $n = 25$ trials. Let Y equal the number of first serves that are good, and let the critical region be defined by $C = \{Y : Y \geq 13\}$.

(a) Show that α is computed by $\alpha =$ `1 - pbinom(12 , 25, .4)`.

Solution. $\alpha = P_{p=0.40}(Y \geq 13) = 1 - P(Y < 12|p = 0.4) =$ `1 - pbinom(12 , 25, .4)` $= 0.154$.

(b) Find $\beta = P(Y < 13)$ when $p = 0.60$; that is, $\beta = P(Y \leq 12;\ p = 0.60)$ so that $1 - \beta$ is the power at $p = 0.60$.

Solution. $\beta = P_{p=0.6}(Y < 13) =$ `pbinom(12 , 25, .6)` $= 0.154 \Rightarrow 1 - \beta = 0.846$.

4.6 Additional Comments About Statistical Tests

4.6.2. Consider the power function $\gamma(\mu)$ and its derivative $\gamma(\mu)$ given by (4.6.5) and (4.6.6). Show that $\gamma(\mu)$ is strictly negative for $\mu < \mu_0$ and strictly positive for $mu > \mu_0$.

Solution.

Given (4.6.6):

$$\begin{aligned}\gamma'(\mu) &= \frac{\sqrt{n}}{\sigma}\left[\phi\left(\frac{\sqrt{n}(\mu_0-\mu)}{\sigma} + z_{\alpha/2}\right) - \phi\left(\frac{\sqrt{n}(\mu_0-\mu)}{\sigma} - z_{\alpha/2}\right)\right]\\ &= \frac{\sqrt{n}}{\sigma}\left[\phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu_0-\mu)}{\sigma}\right) - \phi\left(z_{\alpha/2} - \frac{\sqrt{n}(\mu_0-\mu)}{\sigma}\right)\right]\end{aligned}$$

because of $\phi(z) = \phi(-z)$, or $\phi(z)$'s symmetry. Also we have the further from the origin z, the smaller $\phi(z)$ is. So, if $\mu < \mu_0$,

$$\begin{aligned}&z_{\alpha/2} + \frac{\sqrt{n}(\mu_0-\mu)}{\sigma} > z_{\alpha/2} - \frac{\sqrt{n}(\mu_0-\mu)}{\sigma}\\ &\Rightarrow \phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu_0-\mu)}{\sigma}\right) < \phi\left(z_{\alpha/2} - \frac{\sqrt{n}(\mu_0-\mu)}{\sigma}\right)\\ &\Rightarrow \gamma'(\mu) < 0,\end{aligned}$$

indicating $\gamma(\mu)$ is strictly decreasing and vice versa.

4.6.3. Show that the test defined by 4.6.9 has exact size α for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_0$.

Solution.

Since $\sqrt{n}(\overline{X} - \mu_0)/S \sim t_{n-1}$, if $\overline{X} > \mu_0$,

$$P\left(\left|\frac{\sqrt{n}(\overline{X} - \mu_0)}{S}\right| \geq t_{\alpha/2,n-1}\right) = 2P\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{S} \geq t_{\alpha/2,n-1}\right) = 2(\alpha/2) = \alpha.$$

If $\overline{X} < \mu_0$,

$$P\left(\left|\frac{\sqrt{n}(\overline{X} - \mu_0)}{S}\right| \geq t_{\alpha/2,n-1}\right) = 2P\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{S} \leq -t_{\alpha/2,n-1}\right) = 2(\alpha/2) = \alpha.$$

4.6.8. Let p equal the proportion of drivers who use a seat belt in a country that does not have a mandatory seat belt law. It was claimed that $p = 0.14$. An advertising campaign was conducted to increase this proportion. Two months after the campaign, $y = 104$ out of a random sample of $n = 590$ drivers were wearing their seat belts. Was the campaign successful?

(a) Define the null and alternative hypotheses.

Solution. $H_0 : p = 0.14 \quad \text{versus} \quad H_A : p > 0.14.$

(b) Define a critical region with an $\alpha = 0.01$ significance level.

Solution.

Let $\widehat{p} = 104/590 = 0.176$. Then a critical region is

$$Z = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} > z_\alpha = 2.326.$$

(c) Determine the approximate p-value and state your conclusion.

Solution.

$$p = P\left(Z > \frac{0.176 - 0.14}{\sqrt{0.14(0.86)/590}}\right) = 1 - \Phi(2.52) = 0.00587.$$

Since $p < \alpha = 0.01$ (or $Z = 2.52 > 2.326$), H_0 is rejected; there is sufficient evidence to show that the campaign was successful.

4.6.9. In Exercise 4.2.18 we found a confidence interval for the variance σ^2 using the variance S2 of a random sample of size n arising from $N(\mu, \sigma^2)$, where the mean μ is unknown. In testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 > \sigma_0^2$, use the critical region defined by $(n-1)S^2/\sigma_0^2 \geq c$. That is, reject H_0 and accept H_1 if $S^2 \geq c\sigma_0^2/(n-1)$. If $n = 13$ and the significance level $\alpha = 0.025$, determine c.

Solution. Since $(n-1)S^2/\sigma_0^2 \sim \chi^2_{n-1} = \chi^2_{12}$, $c = \texttt{qchisq(0.975, 12)} = 23.337$.

4.6.10. In Exercise 4.2.27, in finding a confidence interval for the ratio of the variances of two normal distributions, we used a statistic S_1^2/S_2^2, which has an F-distribution when those two variances are equal. If we denote that statistic by F, we can test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$ using the critical region $F \geq c$. If $n = 13$, $m = 11$, and $\alpha = 0.05$, find c.

Solution. Since $F \sim F_{12,10}$, $c = \texttt{qf(0.95, 12, 10)} = 2.913$.

4.7 Chi-Square Tests

4.7.1. Consider Example 4.7.2. Suppose the observed frequencies of $A_1, ..., A_4$ are 20, 30, 92, and 105, respectively. Modify the R code given in the example to calculate the test for these new frequencies. Report the p-value.

Solution.

Use the following R code to obtain $p = 0.01837$:

```
x = c(20, 30, 92, 105); ps = c(1, 3, 5, 7)/16; chisq.test(x, p = ps)
```

4.7.2. A number is to be selected from the interval $\{x : 0 < x < 2\}$ by a random process. Let $A_i = \{x : (i-1)/2 < x \leq i/2\}$, $i = 1, 2, 3$, and let $A_4 = \{x : \frac{3}{2} < x < 2\}$. For $i = 1, 2, 3, 4$, suppose a certain hypothesis assigns probabilities p_{i0} to these sets in accordance with $p_{i0} = \int_{Ai}(\frac{1}{2})(2-x)dx$, $i = 1, 2, 3, 4$. This hypothesis (concerning the multinomial pdf with $k = 4$) is to be tested at the 5% level of significance by a chi-square test. If the observed frequencies of the sets $A_i, i = 1, 2, 3, 4$, are respectively, 30, 30, 10, 10, would H_0 be accepted at the (approximate) 5% level of significance? Use R code similar to that of Example 4.7.2 for the computation.

Solution.

Since

$$p_{i0} = \int_{(i-1)/2}^{i/2} \left(\frac{1}{2}\right)(2-x)dx = x - \frac{x^2}{4}\Big|_{(i-1)/2}^{i/2} = \frac{9}{16} - \frac{i}{8},$$

$p_{10} = 7/16$, $p_{20} = 5/16$, $p_{30} = 3/16$, $p_{40} = 1/16$. Hence, use the following R code:

```
x = c(30, 30, 10, 10); ps = c(7, 5, 3, 1)/16; chisq.test(x, p = ps)
```

to obtain χ^2 statistic $= 8.38$ and $p = 0.03816 < 0.05$; H_0 is rejected, which means that the observations have a lack of fit to the assignsed probabilities.

4.7.3. Define the sets $A_1 = \{x : -\infty < x \leq 0\}$, $A_i = \{x : i-2 < x \leq i-1\}$, $i = 2, ..., 7$, and $A_8 = \{x : 6 < x < \infty\}$. A certain hypothesis assigns probabilities p_{i0} to these sets A_i in accordance with

$$p_{i0} = \int_{A_i} \frac{1}{2\sqrt{2\pi}} \exp\left[-\frac{(x-3)^2}{2(4)}\right] dx, \quad i = 1, 2, ..., 7, 8.$$

This hypothesis (concerning the multinomial pdf with $k = 8$) is to be tested, at the 5% level of significance, by a chi-square test. If the observed frequencies of the sets $A_i, i = 1, 2, ..., 8$, are, respectively, 60, 96, 140, 210, 172, 160, 88, and 74, would H_0 be accepted at the (approximate) 5% level of significance? Use R code similar to that discussed in Example 4.7.2. The probabilities are easily computed in R; for example, $p_{30} =$ `pnorm(2,3,2) - pnorm(1,3,2)`.

Solution.

Use the R code below:

```
    x = c(60, 96, 140, 210, 172, 160, 88, 74)
    p1 = pnorm(0,3,2)
    p2 = pnorm(1,3,2) - pnorm(0,3,2)
    p3 = pnorm(2,3,2) - pnorm(1,3,2)
    p4 = pnorm(3,3,2) - pnorm(2,3,2)
    p5 = pnorm(4,3,2) - pnorm(3,3,2)
    p6 = pnorm(5,3,2) - pnorm(4,3,2)
    p7 = pnorm(6,3,2) - pnorm(5,3,2)
    p8 = 1 - pnorm(6,3,2)
    ps = c(p1, p2, p3, p4, p5, p6, p7, p8)
    chisq.test(x, p = ps)}
```

to obtain $p = 0.4368 > 0.05$; H_0 would be accepted at 5% significance level.

4.7.4. A die was cast n = 120 independent times and the following data resulted:

Spot Up	1	2	3	4	5	6
Frequency	b	20	20	20	20	$40-b$

If we use a chi-square test, for what values of b would the hypothesis that the die is unbiased be rejected at the 0.025 significance level?

Solution.

Under the null hypothesis that a die is unbiased, $p_{i0} = 1/6$ for $\forall i$, so $np_{i0} = 120(1/6) = 20$. Hence the test statistic is

$$\frac{(b-20)^2}{20} + \frac{[(40-b)-20]^2}{20} = \frac{(b-20)^2}{10}.$$

Since $\chi^2_{0.025,5} =$ `qchisq(0.975, 5)` $= 12.83$, the null is rejected if

$$\frac{(b-20)^2}{10} > 12.83 \Rightarrow b - 20 > 11.33 \text{ or } b - 20 < -11.33 \Rightarrow b > 31.33 \text{ or } b < 8.67.$$

If b is an integer, then $b \leq 8$ or $b \geq 32$.

4.7.5. Consider the problem from genetics of crossing two types of peas. The Mendelian theory states that the probabilities of the classifications (a) round and yellow, (b) wrinkled and yellow, (c) round and green, and (d) wrinkled and green are $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$, and $\frac{1}{16}$, respectively. If, from 160 independent observations, the observed frequencies of these respective classifications are 86, 35, 26, and 13, are these data consistent with the Mendelian theory? That is, test, with $\alpha = 0.01$, the hypothesis that the respective probabilities are $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$, and $\frac{1}{16}$.

Solution.

This is a table to compute chi-square test statistic:

	(a)	(b)	(c)	(d)
Observed	86	35	26	13
Expected	90	30	30	10

The test statistic is

$$X^2 = \frac{4^2}{90} + \frac{5^2}{30} + \frac{4^2}{30} + \frac{3^2}{10} = 2.44.$$

Since $X^2 < \chi^2_{0.01,3} =$ `qchisq(0.99, 3)` $= 11.34$, the null is not rejected; these data would be consistent with the Mendelian theory, as there is insufficient evidence to show that they are different from the theory.

4.7.6. Two different teaching procedures were used on two different groups of students. Each group contained 100 students of about the same ability. At the end of the term, an evaluating team assigned a letter grade to each student. The results were tabulated as follows.

Group	A	B	C	D	F	Total
I	15	25	32	17	11	100
II	9	18	29	28	16	100

If we consider these data to be independent observations from two respective multinomial distributions with $k = 5$, test at the 5% significance level the hypothesis that the two distributions are the same (and hence the two teaching procedures are equally effective). For computation in R, use `r1=c(15,25,32,17,11); r2=c(9,18,29,28,16); mat=rbind(r1,r2); chisq.test(mat)`

Solution.

This is a χ^2 test for independence. The R code shows $p = 0.1711 > 0.05$; the hypothesis is not rejected; there is insufficient evidence to show that the two teaching procedures are differently effective.

4.7.8. Let the result of a random experiment be classified as one of the mutually exclusive and exhaustive ways A_1, A_2, A_3 and also as one of the mutually exhaustive ways B_1, B_2, B_3, B_4. Say that 180 independent trials of the experiment result in the following frequencies:

	B_1	B_2	B_3	B_4	Total
A_1	$15-3k$	$15-k$	$15+k$	$15+3k$	60
A_1	15	15	15	15	60
A_3	$15+3k$	$15+k$	$15-k$	$15-3k$	60
Total	45	45	45	45	180

where k is one of the integers 0, 1, 2, 3, 4, 5. What is the smallest value of k that leads to the rejection of the independence of the A attribute and the B attribute at the $\alpha = 0.05$ significance level?

Solution.

The expected values are all $45(60)/80 = 15$. Hence the chi-square statistic is

$$\left[\frac{(-3k)^2}{15} + \frac{(-k)^2}{15} + \frac{k^2)}{15} + \frac{(3k)^2}{15}\right] \times 2 = \frac{8}{3}k^2.$$

In this case, the degrees of freedom is $(3-1)(4-1) = 6$. Since `qchisq(0.95, 6)` $= 12.6$, the null hypothesis of the independence A and B is rejected if

$$\frac{8}{3}k^2 > 12.6 \;\Rightarrow\; k > 2.17,$$

which gives us the smallest value of integer $k = 3$.

4.7.9. It is proposed to fit the Poisson distribution to the following data:

x	0	1	2	3	$3 < x$
Frequency	20	40	16	18	6

(a) Compute the corresponding chi-square goodness-of-fit statistic.

Solution.

The mean of the Poisson distribution is computed as

$$\frac{0(20) + 1(40) + 2(16) + 3(18) + 4(6)}{20 + 40 + 16 + 18 + 6} = 1.5.$$

Hence,

$$\begin{aligned}
p_{00} &= P(X = 0) = \texttt{dpois(0, 1.5)} = 0.223\\
p_{10} &= P(X = 1) = \texttt{dpois(1, 1.5)} = 0.335\\
p_{20} &= P(X = 2) = \texttt{dpois(2, 1.5)} = 0.251\\
p_{30} &= P(X = 3) = \texttt{dpois(3, 1.5)} = 0.126\\
p_{40} &= P(X > 3) = \texttt{1 - ppois(3, 1.5)} = 0.066.
\end{aligned}$$

Then the following code shows the chi-square goodness-of-fit statistic is 7.23:

```
x = c(20, 40, 16, 18, 6)
ps = c(dpois(0, 1.5), dpois(1, 1.5), dpois(2, 1.5), dpois(3, 1.5), 1-ppois(3, 1.5))
chisq.test(x, p = ps)
```

(b) How many degrees of freedom are associated with this chi-square?

Solution. Since this χ^2 test is a hi-square goodness-of-fit, the degrees of freedom is $5 - 1 = 4$.

(c) Do these data result in the rejection of the Poisson model at the $\alpha = 0.05$ significance level?

Solution.

The critical values is `qchisq(0.95, 4)` $= 9.49$. The statistic obtained in part (a) is less than 9.49. Thus, the Poisson model is not rejected at 5% level; we were not able to show that the observed data do not fit the Poisson distribution.

4.9 Bootstrap Procedures

4.9.8. Consider the data of Example 4.9.2. The two-sample t-test of Example 4.6.2 can be used to test these hypotheses. The test is not exact here (why?), but it is an approximate test. Show that the value of the test statistic is $t = 0.93$, with an approximate p-value of 0.18.

Solution.

The reason that the test is not exact here is that the distribution is contaminated. By the Example 4.9.2, we have the two sample variances $s_x^2 = 20.407^2$ and $s_y^2 = 18.585^2$. Hence, the pooled variance estimator of σ^2 is

$$s_p^2 = \frac{(15-1)s_x^2 + (15-1)s_y^2}{30-2} = \frac{14(20.41^2 + 18.59^2)}{28} = 380.924 \;\Rightarrow\; s_p = 19.52,$$

which gives

$$\begin{aligned} t &= \frac{\overline{y} - \overline{x}}{s_p\sqrt{2/n}} = \frac{6.63}{19.52\sqrt{2/15}} = 0.93, \\ p &= P(t_{28} > 0.93) = \texttt{1 - pt(0.93, 28)} = 0.18. \end{aligned}$$

4.9.12. For the situation described in Example 4.9.3, show that the value of the one-sample t-test is $t = 0.84$ and its associated p-value is 0.20.

Solution.

Conduct one-sided one-sample t-test for testing $H_0 : \mu = 90$ versus $H_A : \mu > 90$.

```
X <- c(119.7, 104.1, 92.8, 85.4, 108.6, 93.4, 67.1, 88.4,   101.0, 97.2,
95.4, 77.2, 100.0, 114.2, 150.3, 102.3, 105.8, 107.5, 0.9, 94.1)
t.test(X, mu = 90, alternative = ’greater’)
```

which shows `t = 0.84475, df = 19, p-value = 0.2044`.