Chapter 10

Statistical Inference II

Statistical Inference journey continues. Methods covered in this chapter allow us to conduct new tests for independence and for the goodness of fit (sec. 10.1), test hypotheses without relying on a particular family of distributions (sec. 10.2), make full use of Monte Carlo methods for estimation and testing (sec. 10.3), and account for all the sources of information in addition to the real data (sec. 10.4).

10.1 Chi-square tests

Several important tests of statistical hypotheses are based on the *Chi-square distribution*. We have already used this distribution in Section 9.5 to study the population variance. This time, we will develop several tests based on the *counts* of our sampling units that fall in various categories. The general principle developed by Karl Pearson near year 1900 is to compare the *observed counts* against the *expected counts* via the *chi-square statistic*

Chi-square statistic

$$\boxed{\chi^2 = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)}.} \tag{10.1}$$

Here the sum goes over N categories or groups of data defined depending on our testing problem; $Obs(k)$ is the actually observed number of sampling units in category k, and $Exp(k) = \mathbf{E}\{Obs(k) \mid H_0\}$ is the expected number of sampling units in category k if the null hypothesis H_0 is true.

This is always a *one-sided, right-tail* test. That is because only the low values of χ^2 show that the observed counts are close to what we expect them to be under the null hypotheses, and therefore, the data support H_0. On the contrary, large χ^2 occurs when *Obs* are far from *Exp*, which shows inconsistency of the data and the null hypothesis and does not support H_0.

Therefore, a level α rejection region for this chi-square test is

$$R = [\chi^2_\alpha, +\infty),$$

and the P-value is always calculated as

$$P = \boldsymbol{P}\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\}.$$

Pearson showed that the null distribution of χ^2 converges to the Chi-square distribution with $(N-1)$ degrees of freedom, as the sample size increases to infinity. This follows from a suitable version of the Central Limit Theorem. To apply it, we need to make sure the sample size is large enough. The rule of thumb requires an *expected count of at least 5 in each category*,

$$Exp(k) \geq 5 \quad \text{for all } k = 1, \ldots, N.$$

If that is the case, then we can use the Chi-square distribution to construct rejection regions and compute P-values. If a count in some category is less than 5, then we should *merge* this category with another one, recalculate the χ^2 statistic, and then use the Chi-square distribution.

Here are several main applications of the chi-square test.

10.1.1 Testing a distribution

The first type of applications focuses on testing whether the data belong to a particular distribution. For example, we may want to test whether a sample comes from the Normal distribution, whether interarrival times are Exponential and counts are Poisson, whether a random number generator returns high quality Standard Uniform values, or whether a die is unbiased.

In general, we observe a sample $(X_1, \ldots, X_n)$ of size n from distribution F and test

$$H_0 : F = F_0 \qquad \text{vs} \qquad H_A : F \neq F_0 \tag{10.2}$$

for some given distribution F_0.

To conduct the test, we take all possible values of X under F_0, the *support* of F_0, and split them into N bins $B_1, \ldots, B_N$. A rule of thumb requires anywhere from 5 to 8 bins, which is quite enough to identify the distribution F_0 and at the same time have sufficiently high expected count in each bin, as it is required by the chi-square test ($Exp \geq 5$).

The observed count for the k-th bin is the number of X_i that fall into B_k,

$$Obs(k) = \#\{i = 1, \ldots, n \,:\, X_i \in B_k\}.$$

If H_0 is true and all X_i have the distribution F_0, then $Obs(k)$, the number of "successes" in n trials, has Binomial distribution with parameters n and $p_k = F_0(B_k) = \boldsymbol{P}\{X_i \in B_k \mid H_0\}$. Then, the corresponding expected count is the expected value of this Binomial distribution,

$$Exp(k) = np_k = nF_0(B_k).$$

After checking that all $Exp(k) \geq 5$, we compute the χ^2 statistic (10.1) and conduct the test.

Example 10.1 (Is the die unbiased?). Suppose that after losing a large amount of money, an unlucky gambler questions whether the game was fair and the die was really unbiased. The last 90 tosses of this die gave the following results,

Number of dots on the die	1	2	3	4	5	6
Number of times it occurred	20	15	12	17	9	17

Let us test $H_0 : F = F_0$ vs $H_A : F \neq F_0$, where F is the distribution of the number of dots on the die, and F_0 is the *Discrete Uniform distribution*, under which

$$P(X = x) = \frac{1}{6} \qquad \text{for} \qquad x = 1, 2, 3, 4, 5, 6.$$

Observed counts are $Obs = 20, 15, 12, 17, 9$, and 17. The corresponding expected counts are

$$Exp(k) = np_k = (90)(1/6) = 15 \quad \text{(all more than 5)}.$$

Compute the chi-square statistic

$$\chi^2_{\text{obs}} = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)}$$
$$= \frac{(20-15)^2}{15} + \frac{(15-15)^2}{15} + \frac{(12-15)^2}{15} + \frac{(17-15)^2}{15} + \frac{(9-15)^2}{15} + \frac{(17-15)^2}{15} = 5.2.$$

From Table A6 with $N - 1 = 5$ d.f., the P-value is

$$P = \boldsymbol{P}\left\{\chi^2 \geq 5.2\right\} = \text{ between 0.2 and 0.8 }.$$

It means that there is no significant evidence to reject H_0, and therefore, no evidence that the die was biased. ◊

10.1.2 Testing a family of distributions

One more step. The chi-square test can also be used to test the entire *model*. We are used to model the number of traffic accidents with Poisson, errors with Normal, and interarrival times with Exponential distribution. These are the models that are believed to fit the data well. Instead of just relying on this assumption, we can now test it and see if it is supported by the data.

We again suppose that the sample $(X_1, \dots, X_n)$ is observed from a distribution F. It is desired to test whether F belongs to some *family of distributions* $\mathfrak{F}$,

$$H_0 : F \in \mathfrak{F} \qquad \text{vs} \qquad H_A : F \notin \mathfrak{F}. \tag{10.3}$$

Unlike test (10.2), the parameter θ of the tested family $\mathfrak{F}$ is not given; it is unknown. So, we have to estimate it by a consistent estimator $\hat{\theta}$, to ensure $\hat{\theta} \to \theta$ and to preserve the chi-square distribution when $n \to \infty$. One can use the maximum likelihood estimator of θ.

Degrees of freedom of this chi-square distribution will be reduced by the number of estimated parameters. Indeed, if θ is d-dimensional, then its estimation involves a system of d equations. These are d constraints which reduce the number of degrees of freedom by d.

It is often called a *goodness of fit test* because it measures how well the chosen model fits the data. Summarizing its steps,

- we find the maximum likelihood estimator $\hat{\theta}$ and consider the distribution $F(\hat{\theta}) \in \mathfrak{F}$;
- partition the support of $F(\hat{\theta})$ into N bins $B_1, \dots, B_N$, preferably with $N \in [5, 8]$;
- compute probabilities $p_k = \boldsymbol{P}\{X \in B_k\}$ for $k = 1, \dots, N$ using $\hat{\theta}$ as the parameter value;
- compute $Obs(k)$ from the data, $Exp(k) = np_k$, and the chi-square statistic (10.1); if $np_k < 5$ for some k then merge B_k with another region;
- Compute the P-value or construct the rejection region using Chi-square distribution with $(N - d - 1)$ degrees of freedom, where d is the dimension of θ or the number of estimated parameters. State conclusions.

Example 10.2 (Transmission errors). The number of transmission errors in communication channels is typically modeled by a Poisson distribution. Let us test this assumption. Among 170 randomly selected channels, 44 channels recorded no transmission error during a 3-hour period, 52 channels recorded one error, 36 recorded two errors, 20 recorded three errors, 12 recorded four errors, 5 recorded five errors, and one channel had seven errors.

Solution. We are testing whether the unknown distribution F of the number of errors belongs to the Poisson family or not. That is,

$$H_0 : F = \text{Poisson}(\lambda) \text{ for some } \lambda \qquad \text{vs} \qquad H_A : F \neq \text{Poisson}(\lambda) \text{ for any } \lambda.$$

First, estimate parameter λ. Its maximum likelihood estimator equals

$$\hat{\lambda} = \bar{X} = \frac{(44)(0) + (52)(1) + (36)(2) + (20)(3) + (12)(4) + (5)(5) + (1)(7)}{170} = 1.55$$

(see Example 9.7 on p. 243).

Next, the support of Poisson distribution is the set of all non-negative integers. Partition it into bins, let's say,

$$B_0 = \{0\},\ B_1 = \{1\},\ B_2 = \{2\},\ B_3 = \{3\},\ B_4 = \{4\},\ B_5 = [5, \infty).$$

The observed counts for these bins are given: 44, 52, 36, 20, 12, and $5+1 = 6$. The expected counts are calculated from the Poisson pmf as

$$Exp(k) = np_k = ne^{-\hat{\lambda}}\frac{\hat{\lambda}^k}{k!} \qquad \text{for } k = 0, \dots, 4, \quad n = 170, \quad \text{and} \quad \hat{\lambda} = 1.55.$$

The last expected count $Exp(5) = 170 - Exp(0) - \ldots - Exp(4)$ because $\sum p_k = 1$, and therefore, $\sum np_k = n$. Thus, we have

k	0	1	2	3	4	5
p_k	0.21	0.33	0.26	0.13	0.05	0.02
np_k	36.0	55.9	43.4	22.5	8.7	3.6

We notice that the last group has a count below five. So, let's combine it with the previous group, $B_4 = [4, \infty)$. For the new groups, we have

k	0	1	2	3	4
$Exp(k)$	36.0	55.9	43.4	22.5	12.3
$Obs(k)$	44.0	52.0	36.0	20.0	18.0

Then compute the chi-square statistic

$$\chi^2_{\text{obs}} = \sum_k \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)} = 6.2$$

and compare it against the Chi-square distribution with $5 - 1 - 1 = 3$ d.f. in Table A6. With a P-value

$$P = \boldsymbol{P}\left\{\chi^2 \geq 6.2\right\} = \text{ between 0.1 and 0.2,}$$

we conclude that there is *no evidence against a Poisson distribution* of the number of transmission errors. ◊

Testing families of continuous distributions is rather similar.

Example 10.3 (NETWORK LOAD). The data in Exercise 8.2 on p. 234 shows the number of concurrent users of a network in $n = 50$ locations. For further modeling and hypothesis testing, can we assume an approximately Normal distribution for the number of concurrent users?

Solution. For the distribution F of the number of concurrent users, let us test

$H_0 : F = \text{Normal}(\mu, \sigma)$ for some μ and σ vs $H_A : F \neq \text{Normal}(\mu,\sigma)$ for any μ and σ.

Maximum likelihood estimates of μ and σ are (see Exercise 9.3 on p. 300)

$$\hat{\mu} = \bar{X} = 17.95 \quad \text{and} \quad \hat{\sigma} = \sqrt{\frac{(n-1)s^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}(X_i - \bar{X})^2} = 3.13.$$

Split the support $(-\infty, +\infty)$ (actually, only $[0, \infty)$ for the number of concurrent users) into bins, for example,

$$B_1 = (-\infty, 14),\ B_2 = [14, 16),\ B_3 = [16, 18),\ B_4 = [18, 20),\ B_5 = [20, 22),\ B_6 = [22, \infty)$$

(in thousands of users). While selecting these bins, we made sure that the expected counts in each of them will not be too low. Use Table A4 of Normal distribution to find the probabilities p_k,

$$p_1 = P(X \in B_1) = P(X \le 14) = P\left(Z \le \frac{14 - 17.95}{3.13}\right) = \Phi(-1.26) = 0.1038,$$

and similarly for $p_1, \dots, p_6$. Then calculate $Exp(k) = np_k$ and count $Obs(k)$ (check!),

k	1	2	3	4	5	6
B_k	$(-\infty, 14)$	$[14, 16)$	$[16, 18)$	$[18, 20)$	$[20, 22)$	$[22, \infty)$
p_k	0.10	0.16	0.24	0.24	0.16	0.10
$Exp(k)$	5	8	12	12	8	5
$Obs(k)$	6	8	13	11	6	6

From this table, the test statistic is

$$\chi^2_{\text{obs}} = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)} = 1.07,$$

which has a high P-value

$$P = \boldsymbol{P}\left\{\chi^2 \ge 1.07\right\} > 0.2.$$

Chi-square distribution was used with $6 - 1 - 2 = 3$ d.f., where we lost 2 d.f. due to 2 estimated parameters. Hence, there is *no evidence against a Normal distribution* of the number of concurrent users. ◊

10.1.3 Testing independence

Many practical applications require testing independence of two factors. If there is a significant association between two features, it helps to understand the cause-and-effect relationships. For example, is it true that smoking causes lung cancer? Do the data confirm that drinking and driving increases the chance of a traffic accident? Does customer satisfaction with their PC depend on the operating system? And does the graphical user interface (GUI) affect popularity of a software?

Apparently, chi-square statistics can help us test

H_0 : Factors A and B are independent vs H_A : Factors A and B are dependent.

It is understood that each factor partitions the whole population $\mathcal{P}$ into two or more categories, $A_1, \dots, A_k$ and $B_1, \dots, B_m$, where $A_i \cap A_j = \varnothing$, $B_i \cap B_j = \varnothing$, for any $i \ne j$, and $\cup A_i = \cup B_i = \mathcal{P}$.

Independence of factors is understood just like independence of random variables in Section 3.2.2. Factors A and B are independent if any randomly selected unit x of the population belongs to categories A_i and B_j independently of each other. In other words, we are testing

$$\begin{gathered} H_0 : \boldsymbol{P}\{x \in A_i \cap B_j\} = \boldsymbol{P}\{x \in A_i\}\, \boldsymbol{P}\{x \in B_j\} \text{ for all } i, j \\ \text{vs} \\ H_A : \boldsymbol{P}\{x \in A_i \cap B_j\} \neq \boldsymbol{P}\{x \in A_i\}\, \boldsymbol{P}\{x \in B_j\} \text{ for some } i, j. \end{gathered} \tag{10.4}$$

To test these hypotheses, we collect a sample of size n and count n_{ij} units that landed in the intersection of categories A_i and B_j. These are the *observed counts*, which can be nicely arranged in a *contingency table*,

	B_1	B_2	$\cdots$	B_m	row total
A_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{km}	$n_{k\cdot}$
column total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	$n_{\cdot\cdot} = n$

Notation $n_{i\cdot} = \sum_i n_{ij}$ and $n_{\cdot j} = \sum_j n_{ij}$ is quite common for the row totals and column totals.

Then we estimate all the probabilities in (10.4),

$$\widehat{P}\{x \in A_i \cap B_j\} = \frac{n_{ij}}{n}, \quad \widehat{P}\{x \in A_i\} = \sum_{j=1}^{m} \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n}, \quad \widehat{P}\{x \in B_j\} = \sum_{i=1}^{k} \frac{n_{ij}}{n} = \frac{n_{\cdot j}}{n}.$$

If H_0 is true, then we can also estimate the probabilities of intersection as

$$\widetilde{P}\{x \in A_i \cap B_j\} = \Big(\frac{n_{i\cdot}}{n}\Big)\Big(\frac{n_{\cdot j}}{n}\Big)$$

and estimate the *expected counts* as

$$\widehat{Exp}(i,j) = n\Big(\frac{n_{i\cdot}}{n}\Big)\Big(\frac{n_{\cdot j}}{n}\Big) = \frac{(n_{i\cdot})(n_{\cdot j})}{n}.$$

This is the case when expected counts $Exp(i,j) = \mathbf{E}\{Obs(i,j) \mid H_0\}$ are estimated under H_0. There is not enough information in the null hypothesis H_0 to compute them exactly.

After this preparation, we construct the usual *chi-square statistic* comparing the observed and the estimated expected counts over the entire contingency table,

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\Big\{Obs(i,j) - \widehat{Exp}(i,j)\Big\}^2}{\widehat{Exp}(i,j)}. \tag{10.5}$$

This χ^2_{obs} should now be compared against the chi-square table. How many *degrees of freedom* does it have here? Well, since the table has k rows and m columns, wouldn't the number of degrees of freedom equal $(k \cdot m)$?

It turns out that the differences

$$d_{ij} = Obs(i,j) - \widehat{Exp}(i,j) = n_{ij} - \frac{(n_{i\cdot})(n_{\cdot j})}{n}$$

in (10.5) have many constraints. For any $i = 1, \dots, k$, the sum $d_{i\cdot} = \sum_j d_{ij} = n_{i\cdot} - \frac{(n_{i\cdot})(n_{\cdot\cdot})}{n} = 0$, and similarly, for any $j = 1, \dots, m$, we have $d_{\cdot j} = \sum_i d_{ij} = n_{\cdot j} - \frac{(n_{\cdot\cdot})(n_{\cdot j})}{n} = 0$. So, do we lose $(k+m)$ degrees of freedom due to these $(k+m)$ constraints? Yes, but there is also one constraint among these constraints! Whatever d_{ij} are, the equality $\sum_i d_{i\cdot} = \sum_j d_{\cdot j}$ always holds. So, if all the $d_{i\cdot}$ and $d_{\cdot j}$ equal zero except the last one, $d_{\cdot m}$, then $d_{\cdot m} = 0$ *automatically* because $\sum_i d_{i\cdot} = 0 = \sum_j d_{\cdot j}$.

As a result, we have only $(k + m - 1)$ linearly independent constraints, and the overall number of degrees of freedom in χ^2_{obs} is

$$\text{d.f.} = km - (k + m - 1) = (k - 1)(m - 1).$$

Chi-square test for independence

Test statistic

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\Big\{ Obs(i,j) - \widehat{Exp}(i,j) \Big\}^2}{\widehat{Exp}(i,j)},$$

where

$Obs(i,j) = n_{ij}$ are observed counts,

$\widehat{Exp}(i,j) = \dfrac{(n_{i\cdot})(n_{\cdot j})}{n}$ are estimated expected counts,

and χ^2_{obs} has $(k-1)(m-1)$ d.f.

As always in this section, this test is one-sided and right-tail.

Example 10.4 (Spam and attachments). Modern email servers and anti-spam filters attempt to identify spam emails and direct them to a junk folder. There are various ways to detect spam, and research still continues. In this regard, an information security officer tries to confirm that the chance for an email to be spam depends on whether it contains images or not. The following data were collected on $n = 1000$ random email messages,

$Obs(i,j) = n_{ij}$	With images	No images	$n_{i\cdot}$
Spam	160	240	400
No spam	140	460	600
$n_{\cdot j}$	300	700	1000

Testing H_0: "being spam and containing images are independent factors" vs H_A: "these factors are dependent", calculate the estimated expected counts,

$$\widehat{Exp}(1,1) = \frac{(300)(400)}{1000} = 120, \qquad \widehat{Exp}(1,2) = \frac{(700)(400)}{1000} = 280,$$
$$\widehat{Exp}(2,1) = \frac{(300)(600)}{1000} = 180, \qquad \widehat{Exp}(2,2) = \frac{(700)(600)}{1000} = 420.$$

$\widehat{Exp}(i,j) = \frac{(n_{i\cdot})(n_{\cdot j})}{n}$	With images	No images	$n_{i\cdot}$
Spam	120	280	400
No spam	180	420	600
$n_{\cdot j}$	300	700	1000

You can always check that all the row totals, the column totals, and the whole table total are the same for the observed and the expected counts (so, if there is a mistake, catch it here).

$$\chi^2_{\text{obs}} = \frac{(160-120)^2}{120} + \frac{(240-280)^2}{280} + \frac{(140-180)^2}{180} + \frac{(460-420)^2}{420} = 31.75.$$

From Table A6 with $(2-1)(2-1) = 1$ d.f., we find that the P-value $P < 0.001$. We have a significant evidence that an email having an attachment is somehow related to being spam. Therefore, this piece of information can be used in anti-spam filters. ◊

Example 10.5 (Internet shopping on different days of the week). A web designer suspects that the chance for an internet shopper to make a purchase through her web site varies depending on the day of the week. To test this claim, she collects data during one week, when the web site recorded 3758 hits.

Observed	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Total
No purchase	399	261	284	263	393	531	502	2633
Single purchase	119	72	97	51	143	145	150	777
Multiple purchases	39	50	20	15	41	97	86	348
Total	557	383	401	329	577	773	738	3758

Testing independence (i.e., probability of making a purchase or multiple purchases is the same on any day of the week), we compute the estimated expected counts,

$$\widehat{Exp}(i,j) = \frac{(n_{i\cdot})(n_{\cdot j})}{n} \qquad \text{for} \qquad i = 1,\dots,7,\ j = 1,2,3.$$

Expected	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Total
No purchase	390.26	268.34	280.96	230.51	404.27	541.59	517.07	2633
Single purchase	115.16	79.19	82.91	68.02	119.30	159.82	152.59	777
Multiple purchases	51.58	35.47	37.13	30.47	53.43	71.58	68.34	348
Total	557	383	401	329	577	773	738	3758

Then, the test statistic is

$$\chi^2_{\text{obs}} = \frac{(399-390.26)^2}{390.26} + \ldots + \frac{(86-68.34)^2}{68.34} = 60.79,$$

and it has $(7-1)(3-1) = 12$ degrees of freedom. From Table A6, we find that the P-value is $P < 0.001$, so indeed, there is significant evidence that the probability of making a single purchase or multiple purchases varies during the week. ◊

MATLAB Demo. The χ^2 test for independence takes only a few lines of code in MATLAB. For example, here is the solution to Example 10.4.

```
X = [160 240; 140 460];                          % Matrix of observed counts
Row = sum(X')'; Col = sum(X); Tot = sum(Row);    % Row and column totals
k = length(Col); m = length(Row);                % Dimensions of the table
e = zeros(size(X));                              % Expected counts
for i=1:k; for j=1:m; e(i,j) = Row(i)*Col(j)/Tot; end; end;
chisq = (X-e).^2./e;                             % Chi-square terms
chistat = sum(sum(chisq));                       % Chi-square statistic
Pvalue = 1-chi2cdf(chistat,(k-1)*(m-1))          % P-value
```

Also, the new version of the Statistics Toolbox of Matlab has a special command `chisquarecont` for testing independence.

10.2 Nonparametric statistics

*Parametric*parametric methods statistical methods are always designed for a specific family of distributions such as Normal, Poisson, Gamma, etc. *Nonparametric statistics* does not assume any particular distribution. On one hand, nonparametric methods are less powerful because the less you assume about the data the less you can find out from it. On the other hand, having fewer requirements, they are applicable to wider applications. Here are three typical examples.

Example 10.6 (Unknown distribution). Our test in Example 9.29 on p. 276 requires the data, battery lives, to follow Normal distribution. Certainly, this t-test is *parametric*. However, looking back at this example, we are not sure if this assumption could be made. Samples of 12 and 18 data measurements are probably too small to find an evidence for or against the Normal distribution. Can we test the same hypothesis ($\mu_X = \mu_Y$) without relying on this assumption? ◊

Example 10.7 (Outliers). Sample sizes in Example 9.37 on p. 284 are large enough ($m = n = 50$), so we can refer to the Central Limit Theorem to justify the use of a Z-test there. However, these data on running times of some software may contain occasional outliers. Suppose that one computer accidentally got frozen during the experiment, and this added 40 minutes to its running time. Certainly, it is an outlier, one out of 100 observations, but it makes an impact on the result.

Indeed, if one observation Y_i increases by 40 minutes, then the new average of 50 running times after the upgrade becomes $\bar{Y} = 7.2 + 40/50 = 8.0$. The test statistic is now $Z = (8.5 - 8.0)/0.36 = 1.39$ instead of 3.61, and the P-value is $P = 0.0823$ instead of $P = 0.0002$. In Example 9.37, we concluded that the evidence of an efficient upgrade is overwhelming, but now we are not so sure!

We need a method that will not be so sensitive to a few outliers. ◊

Example 10.8 (Ordinal data). A software company conducts regular surveys where customers rank their level of satisfaction on a standard scale "strongly agree, agree, neutral, disagree, strongly disagree". Then, statisticians are asked to conduct tests comparing the customer satisfaction of different products.

These data are not even numerical! Yes, they contain important information but how can we conduct tests without any numbers? Can we assign numerical values to customers' responses, say, "strongly agree"=5, "agree"=4, ..., "strongly disagree"=1? This approach seems to claim some additional information that is not given in the data. For example, it implies that the difference between "agree" and "strongly agree" is the same as between "agree"

and "neutral". And further, the difference between "strongly agree" and "neutral" is the same as between "agree" and "disagree". We do have informative data, but this particular information is not given, so we cannot use it.

Are there methods in Statistics that allow us to work with the data that are not numbers but *ranks*, so they can be ordered from the lowest (like "strongly disagree") to the highest (like "strongly agree")? ◊

The situation in Example 10.8 is rather typical. Besides satisfaction surveys, one may want to analyze and compare the levels of education, high school to Ph.D.; military officer ranks, from a Second Lieutenant to a General; or chess skills, from Class J to Grandmaster. Also, many surveys ask for an interval instead of precise quantities, for example, income brackets - "is your annual income less than $20,000?" "between $20,000 and $30,000?" etc.

DEFINITION 10.1

Data that can be ordered from the lowest to the highest without any numeric values are called **ordinal data**.

Most nonparametric methods can deal with situations described in Examples 10.6–10.8. Here we discuss three very common nonparametric tests for one sample and two samples: the sign test, the signed rank test, and the rank sum test.

10.2.1 Sign test

Here is a simple test for one population. It refers to a population *median* M and tests

$$H_0 : M = m$$

against a one-sided or a two-sided alternative, $H_A : M < m$, $H_A : M > m$, or $H_A : M \neq m$. So, we are testing whether exactly a half of the population is below m and a half is above m.

To conduct the *sign test*, simply count how many observations are above m,

$$S_{\text{obs}} = S(X_1, \dots, X_n) = \text{number of } \{i : X_i > m\}.$$

Typically, it is assumed that the underlying distribution is continuous, so $\boldsymbol{P}\{X_i = m\} = 0$. In this case, if H_0 is true and m is the median, then each X_i is equally likely to be above m or below m, and S has *Binomial* distribution with parameters n and $p = 1/2$. Using Table A2 of Binomial distribution for small n or Table A4 of Normal distribution for large n (with a proper standardization and continuity correction, of course), we compute the P-value and arrive at a conclusion.

The right-tail alternative will be supported by large values of S_{obs}, so the P-value should be $P = \boldsymbol{P}\{S \geq S_{\text{obs}}\}$. Similarly, for the left-tail alternative, $P = \boldsymbol{P}\{S \leq S_{\text{obs}}\}$, and for the two-sided test, $P = 2\min\left(\boldsymbol{P}\{S \leq S_{\text{obs}}\}, \boldsymbol{P}\{S \geq S_{\text{obs}}\}\right)$.

The sign test

Test of the median, $H_0: M = m$

Test statistic $S =$ number of $X_i > m$

Null distribution $S \sim Binomial(n, 1/2)$
For large n, $S \approx Normal(n/2, \sqrt{n}/2)$
(if the distribution of X_i is continuous)

Example 10.9 (Unauthorized use of a computer account, continued). Example 9.39 on p. 285 shows very significant evidence that a computer account was used by an unauthorized person. This conclusion was based on the following times between keystrokes,

.24, .22, .26, .34, .35, .32, .33, .29, .19, .36, .30, .15, .17, .28, .38, .40, .37, .27 seconds,

whereas the account owner usually makes 0.2 sec between keystrokes. Our solution was based on a T-test and we had to assume the Normal distribution of data. However, the histogram on Fig. 10.1 does not confirm this assumption. Also, the sample size is too small to conduct a chi-square goodness-of-fit test (Exercise 10.5).

Let's apply *the sign test* because it does not require the Normal distribution. The test statistic for $H_0: M = 0.2$ vs $H_A: M \neq 0.2$ is $S_{\text{obs}} = 15$ because 15 of 18 recorded times exceed 0.2.

Then, from Table A2 with $n = 18$ and $p = 0.5$, we find the P-value,

$$\begin{aligned} P &= 2 \min \left(\boldsymbol{P}\{S \le S_{\text{obs}}\}, \boldsymbol{P}\{S \ge S_{\text{obs}}\} \right) \\ &= 2 \min (0.0038, 0.9993) = 0.0076. \end{aligned}$$

(A hint for using the table here: $\boldsymbol{P}\{S \ge 15\} = \boldsymbol{P}\{S \le 3\}$, because the Binomial(0.5) distribution is symmetric.) So, the sign test rejects H_0 at any $\alpha > 0.0076$), which is an evidence that the account was used by an unauthorized person.

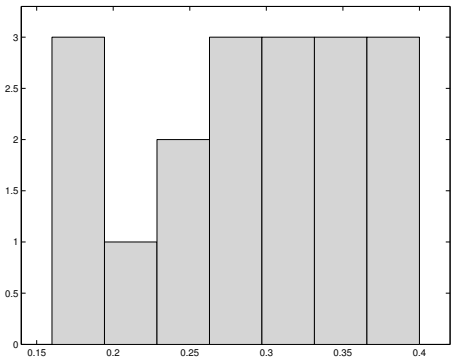

FIGURE 10.1: *The histogram of times between keystrokes does not support or deny a Normal distribution.*

Here is a large-sample version of the sign test.

Example 10.10 (Are the cars speeding?). When Eric learned to drive, he suspected that the median speed of cars on his way to school exceeds the speed limit, which is 30 mph. As an experiment, he drives to school with his friend Evanne, precisely at the speed of 30 mph, and Evanne counts the cars. At the end, Evanne reports that 56 cars were driving faster than Eric, and 44 cars were driving slower. Does this confirm Eric's suspicion?

Solution. This is a one-sided right-tail test of

$$H_0: M = 30 \text{ vs } H_A: M > 30,$$

because rejection of H_0 should imply that more than 50% of cars are speeding. The sign test statistic is $S = 56$, and the sample size is $n = 56 + 44 = 100$. The null distribution of S is approximately Normal with $\mu = n/2 = 50$ and $\sigma = \sqrt{n}/2 = 5$. Calculate the P-value (don't forget the continuity correction),

$$P = \boldsymbol{P}\{S \geq 56\} = \boldsymbol{P}\left\{Z \geq \frac{55.5 - 50}{5}\right\} = 1 - \Phi(1.1) = 1 - .8643 = 0.1357,$$

from Table A4. We can conclude that Eric and Evanne *did not find a significant evidence that the median speed of cars is above the speed limit.*

Notice that we used *ordinal data*. Knowing the exact speed of each car was not necessary for the sign test. ◊

Applying the *sign* test, statisticians used to mark the data above the tested value ($X_i > m$) with pluses and those below m with minuses... which fully explains the test's name.

10.2.2 Wilcoxon signed rank test

You certainly noticed how little information the sign test uses. All we had to know was how many observed values are above m and below m. But perhaps the unused information can be handy, and it should not be wasted. The next test takes magnitudes into account, in addition to signs.

To test the population *median* M, we again split the sample into two parts - below m and above m. But now we compare *how far* below m is the first group, collectively, and *how far* above m is the second group. This comparison is done in terms of statistics called *ranks*.

DEFINITION 10.2

Rank of any unit of a sample is its position when the sample is arranged in the increasing order. In a sample of size n, the smallest observation has rank 1, the second smallest has rank 2, and so on, and the largest observation has rank n. If two or more observations are equal, their ranks are typically replaced by their average rank.

NOTATION R_i = rank of the i-th observation

Having $R_i = r$ means that X_i is the r-th smallest observation in the sample.

Example 10.11. Consider a sample

$$3, 7, 5, 6, 5, 4.$$

The smallest observation, $X_1 = 3$, has rank 1. The second smallest is $X_6 = 4$; it has rank 2. The 3rd and 4th smallest are $X_3 = X_5 = 5$; their ranks 3 and 4 are averaged, and each gets rank 3.5. Then, $X_4 = 6$ and $X_2 = 7$ get ranks 5 and 6. So, we have the following ranks,

$$R_1 = 1,\ R_2 = 6,\ R_3 = 3.5,\ R_4 = 5, R_5 = 3.5, R_6 = 2.$$

◊

Wilcoxon signed rank test of $H_0 : M = m$ is conducted as follows.

1. Consider the *distances* between observations and the tested value, $d_i = |X_i - m|$.
2. Order these distances and compute their *ranks* R_i. *Notice! These are ranks of* d_i, *not* X_i.
3. Take only the ranks corresponding to observations X_i greater than m. Their sum is the test statistic W. Namely,
$$W = \sum_{i:\, X_i > m} R_i$$
(see Figure 10.3 on p. 320 to visualize distances d_i and their ranks R_i).
4. Large values of W suggest rejection of H_0 in favor of $H_A : M > m$; small values support $H_A : M < m$; and both support a two-sided alternative $H_A : M \neq m$.

This test was proposed by an Ireland-born American statistician Frank Wilcoxon (1892-1965), and it is often called *Wilcoxon test*.

For convenience, Wilcoxon proposed to use *signed ranks*, giving a "+" sign to a rank R_i if $X_i > m$ and a "−" sign if $X_i < m$. His statistic W then equals the sum of positive signed ranks. (Some people take the sum of negative ranks instead, or its simple transformation. These statistics are one-to-one functions of each other, so each of them can be used for the Wilcoxon test.)

We assume that the distribution of X is *symmetric* about its median M, which means

$$\boldsymbol{P}\{X \le M - a\} = \boldsymbol{P}\{X \ge M + a\}.$$

for any a (Figure 10.2). If the mean $\mu = \mathbf{E}(X)$ exists, it will also equal M.

Let's also assume that the distribution is continuous. So, there are no equal observations and no ranks need to be averaged.

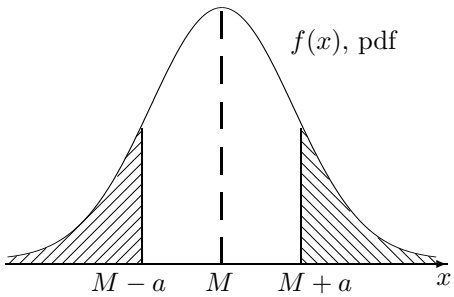

FIGURE 10.2: *Symmetric distribution. The shaded areas-probabilities are equal for any a.*

Under these assumptions, the null distribution of Wilcoxon test statistic W is derived later in this section on p. 321. For $n \le 30$, the critical values are given in Table A8. For $n \ge 15$, one can use a Normal approximation, with a continuity correction because the distribution of W is discrete.

Remark about acceptance and rejection regions: When we construct acceptance and rejection regions for some test statistic T, we always look for a critical value T_α such that $\boldsymbol{P}\{T \le T_\alpha\} = \alpha$ or $\boldsymbol{P}\{T \ge T_\alpha\} = \alpha$.

But in the case of Wilcoxon test, statistic W takes only a finite number of possible values for any n. As a result, there may be no such value w that makes probability $\boldsymbol{P}\{W \le w\}$ or $\boldsymbol{P}\{W \ge w\}$ exactly equal α. One can achieve only inequalities $\boldsymbol{P}\{W \le w\} \le \alpha$ and $\boldsymbol{P}\{W \ge w\} \le \alpha$.

Critical values in Table A8 make these inequalities as tight as possible (that is, as close as possible to equalities), still making sure that

$$\boldsymbol{P}\{\text{Type I error}\} \le \alpha.$$

Wilcoxon signed rank test

Test of the median, $H_0: M = m$.

Test statistic $W = \sum_{i:\, X_i > m} R_i$, where R_i is the rank of $d_i = |X_i - m|$.

Null distribution: Table A8 or recursive formula (10.6).

For $n \geq 15$, $W \approx Normal\left(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right)$

Assumptions: the distribution of X_i is continuous and symmetric

Example 10.12 (Supply and demand). Having sufficient supply to match the demand always requires the help of statisticians.

Suppose that you are managing a student computer lab. In particular, your duty is to ensure that the printers don't run out of paper. During the first six days of your work, the lab consumed

$$7, 5.5, 9.5, 6, 3.5, \text{ and } 9 \text{ cartons of paper.}$$

Does this imply significant evidence, at the 5% level of significance, that the median daily consumption of paper is more than 5 cartons? It is fair to assume that the amounts of used paper are independent on different days, and these six days are as good as a simple random sample.

Let's test $H_0: M = 5$ vs $H_A: M > 5$. According to Table A8 for the right-tail test with $n = 6$ and $\alpha = 0.05$, we'll reject H_0 when the sum of positive ranks $T \geq 19$. To compute Wilcoxon test statistic T, compute distances $d_i = |X_i - 5|$ and rank them from the smallest to the largest.

i	X_i	$X_i - 5$	d_i	R_i	sign
1	7	2	2	4	+
2	5.5	0.5	0.5	1	+
3	9.5	4.5	4.5	6	+
4	6	1	1	2	+
5	3.5	−1.5	1.5	3	−
6	9	4	4	5	+

Then, compute T adding the "positive" ranks only, $T = 4 + 1 + 6 + 2 + 5 = 18$. It did not make it to the rejection region, and therefore, *at the 5% level, these data do not provide significance evidence that the median consumption of paper exceeds 5 cartons per day.* ◊

Example 10.13 (Unauthorized use of a computer account, continued). Applying Wilcoxon signed rank test to test $M = 0.2$ vs $M \neq 0.2$ in Example 10.9, we compute the distances $d_1 = |X_1 - m| = |0.24 - 0.2| = 0.04, \ldots, d_{18} = |0.27 - 0.2| = 0.07$ and rank them; see the ranks on Figure 10.3. Here we notice that the 9-th, 12-th, and 13-th observations are below the tested value $m = 0.2$ while all the others are above $m = 0.2$. Computing the

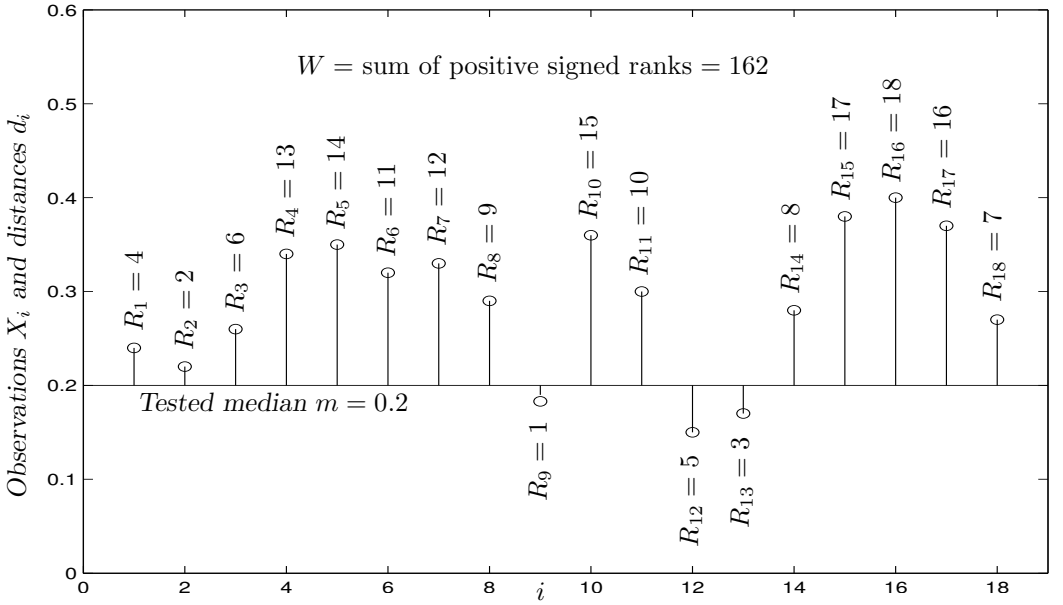

FIGURE 10.3: *Ranks of distances $d_i = |X_i - m|$ for the Wilcoxon signed rank test.*

sum of only positive signed ranks, without R_9, R_{12}, and R_{13}, we find the test statistic

$$W = \sum_{i:\, X_i > m} R_i = 162.$$

Can you find a shortcut in this computation? Look, the sum of all ranks is $1 + \ldots + 18 = (18)(19)/2 = 171$. From this, we have to deduct the ranks corresponding to small X_i, so $W_{\text{obs}} = 171 - R_9 - R_{12} - R_{13} = 171 - 1 - 5 - 3 = 162$.

Compute a P-value. This is a two-sided test, therefore,

$$P = 2 \min \left(\boldsymbol{P} \{W \le 162\}, \boldsymbol{P} \{S \ge 162\} \right) < 2 \cdot 0.001 = 0.002,$$

if we use Table A8 with $n = 18$.

For the sample size $n = 18$, we can also use the Normal approximation. The null distribution of W has

$$\mathbf{E}(W|H_0) = \frac{(18)(19)}{4} = 85.5 \quad \text{and} \quad \text{Std}(W|H_0) = \sqrt{\frac{(18)(19)(37)}{24}} = 23.0.$$

Making a *continuity correction* (because W is discrete), we find the P-value for this two-sided test,

$$P = 2\boldsymbol{P} \left\{ Z \ge \frac{161.5 - 85.5}{23.0} \right\} = 2(1 - \Phi(3.30)) = 2 \cdot 0.0005 = 0.001.$$

Wilcoxon test shows strong evidence that the account was used by an unauthorized person. $\diamondsuit$

As always, for a one-sided alternative, the P-value is $\boldsymbol{P}\{W \ge W_{\text{obs}}\}$ or $\boldsymbol{P}\{W \le W_{\text{obs}}\}$.

Comparing Examples 10.9 and 10.13, we can see that the evidence obtained by the Wilcoxon test is stronger. This should generally be the case. Wilcoxon signed rank test utilizes more information contained in the data, and therefore, it is more *powerful*, i.e., more sensitive to a violation of H_0. In other words, if H_0 is not true, the Wilcoxon test is more likely to show that.

MATLAB DEMO. The following code may be used to calculate the Wilcoxon signed rank test statistic, given the data in vector X and the tested median m.

```
x = []; y=[];                  % Split the data into X_i > m and X_i < m
for k = 1:length(X);
  if X(k)>m; x=[x X(k)];
  else y=[y X(k)]; end; end;
dx = abs(x-m); d = abs(X-m);   % Distances to the tested median
W=0; for k=1:length(x);        % Each rank is the number of distances d
  W=W+sum(d<dx(k))+1; end;     % not exceeding the given distance d_x(k)
W                              % W is the sum of positive ranks
```

Also, the Statistics Toolbox of MATLAB has a special command `signrank` for the Wilcoxon signed rank test. It is designed to test $H_0 : M = 0$ only, so subtract the tested value m from the data X before running this command: `signrank(X-m)`. This command returns a P-value for the two-sided test.

Null distribution of Wilcoxon test statistic

Under the assumption of a symmetric and continuous distribution of $X_1, \dots, X_n$, here we derive the null distribution of W. Elegant ideas are used in this derivation.

Exact distribution

The exact null distribution of test statistic W can be found *recursively*. That is, we start with the trivial case of a sample of size $n = 1$, then compute the distribution of W for $n = 2$ from it, use that for $n = 3$, and so on, similarly to the mathematical induction.

In a sample of size n, all ranks are different integers between 1 and n. Because of the symmetry of the distribution, each of them can be a "positive rank" or a "negative rank" with probabilities 0.5 and 0.5, if H_0 is true.

Let $p_n(w)$ denote the probability mass function of W for a sample of size n. Then, for $n = 1$, the distribution of W is

$$p_1(0) = \boldsymbol{P}\{W = 0 \mid n = 1\} = 0.5 \quad \text{and} \quad p_1(1) = \boldsymbol{P}\{W = 1 \mid n = 1\} = 0.5.$$

Now, make a transition from size $(n-1)$ to size n, under H_0. By symmetry, the most distant observation from M that produces the highest rank n can be above M or below M with probabilities 0.5 and 0.5. So, the sum of positive ranks W equals w in two cases:

(1) the sum of positive ranks without rank n equals w, and rank n is "negative"; or

(2) the sum of positive ranks without rank n equals $(w - n)$, and rank n is "positive", in which case it adds n to the sum making the sum equal w.

This gives us a recursive formula for calculating the pmf of Wilcoxon statistic W,

$$p_n(w) = 0.5p_{n-1}(w) + 0.5p_{n-1}(w - n). \tag{10.6}$$

Normal approximation

For large samples ($n \geq 15$ is a rule of thumb), the distribution of W is approximately Normal. Let us find its parameters $\mathbf{E}(W)$ and $\mathrm{Var}(W)$.

Introduce Bernoulli(0.5) variables $Y_1, \ldots, Y_n$, where $Y_i = 1$ if the i-th signed rank is positive. The test statistic can then be written as

$$W = \sum_{i=1}^{n} iY_i.$$

In this sum, negative signed ranks will be multiplied by 0, so only the positive ones will be counted. Recall that for Bernoulli(0.5), $\mathbf{E}(Y_i) = 0.5$ and $\mathrm{Var}(Y_i) = 0.25$. Using independence of Y_i, compute

$$\begin{aligned} \mathbf{E}(W|H_0) &= \sum_{i=1}^{n} i\,\mathbf{E}(Y_i) = \left(\sum_{i=1}^{n} i\right)(0.5) = \left(\frac{n(n+1)}{2}\right)(0.5) = \frac{n(n+1)}{4}, \\ \mathrm{Var}(W|H_0) &= \sum_{i=1}^{n} i^2\,\mathrm{Var}(Y_i) = \left(\frac{n(n+1)(2n+1)}{6}\right)\left(\frac{1}{4}\right) = \frac{n(n+1)(2n+1)}{24}. \end{aligned}$$

Applying this Normal approximation, do not forget to make a *continuity correction* because the distribution of W is discrete.

The following MATLAB code can be used to calculate the probability mass function of W under H_0 for $n = 1, \ldots, 15$.

```
N=15; S=N*(N+1)/2;                    % max value of W
p=zeros(N,S+1);                       % pmf; columns 1..S+1 for w=0..S
p(1,1)=0.5; p(1,2)=0.5;               % pmf for n=1
for n=2:N;                            % Sample sizes until N
  for w=0:n-1;                        % Possible values of W
    p(n,w+1) = 0.5*p(n-1,w+1); end;
  for w=n:S;                          % Possible values of W
    p(n,w+1) = 0.5*p(n-1,w+1) + 0.5*p(n-1,w-n+1); end;
end;
```

Taking partial sums `F(n,w)=sum(p(n,1:w))`, we can find the cdf. Table A8 is based on these calculations.

10.2.3 Mann-Whitney-Wilcoxon rank sum test

What about two-sample procedures? Suppose we have samples from two populations, and we need to compare their medians or just some location parameters to see how different they are.

Wilcoxon signed rank test can be extended to a two-sample problem as follows.

We are comparing two populations, the population of X and the population of Y. In terms of their cumulative distribution functions, we test

$$H_0 : F_X(t) = F_Y(t) \quad \text{for all } t.$$

Assume that under the alternative H_A, either Y is *stochastically larger* than X, and $F_X(t) > F_Y(t)$, or it is *stochastically smaller* than X, and $F_X(t) < F_Y(t)$. For example, it is the case when one distribution is obtained from the other by a simple shift, as on Figure 10.4.

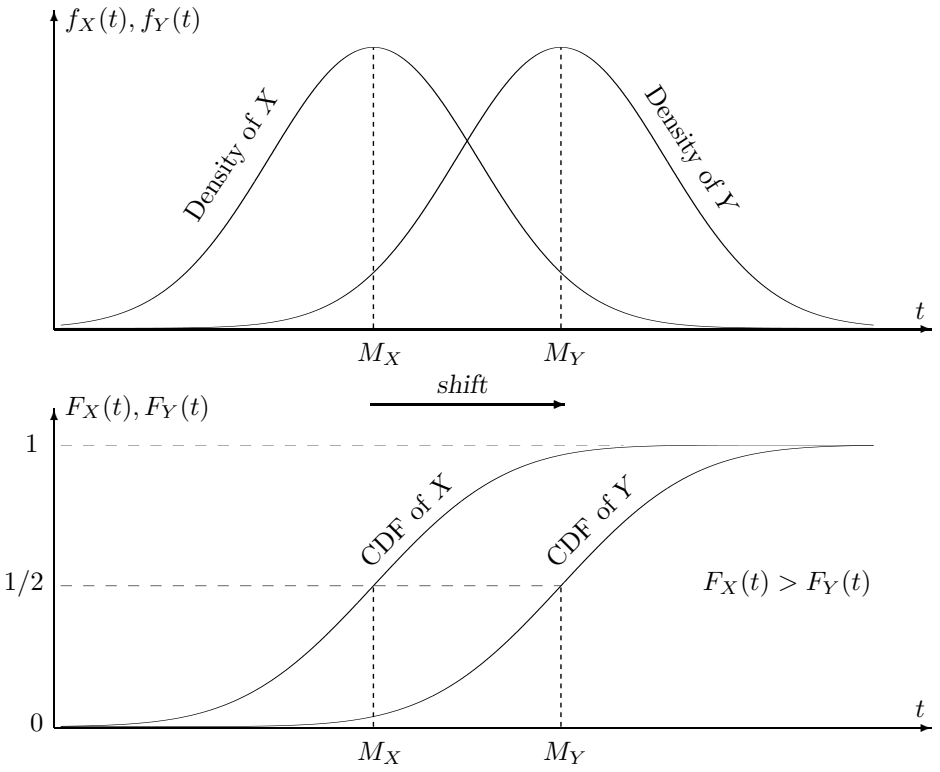

FIGURE 10.4: *Variable Y is stochastically larger than variable X. It has a larger median and a smaller cdf, $M_Y > M_X$ and $F_Y(t) < F_X(t)$.*

Observed are two independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$.

In the previous section, we compared observations with the fixed tested value m. Instead, now we'll compare one sample against the other! Here is how we conduct this test.

1. Combine all X_i and Y_j into one sample.
2. Rank observations in this combined sample. Ranks R_i are from 1 to $(n + m)$. Some of these ranks correspond to X-variables, others to Y-variables.
3. The test statistic U is the sum of all X-ranks.

When U is small, it means that X-variables have low ranks in the combined sample, so they are generally smaller than Y-variables. This implies that Y is stochastically larger than X, supporting the alternative $H_A : F_Y(t) < F_X(t)$ (Figure 10.4).

This test was proposed by Frank Wilcoxon for equal sample sizes $n = m$ and later generalized to any n and m by other statisticians, Henry Mann and Donald Whitney.

The null distribution of the Mann-Whitney-Wilcoxon test statistic U is in Table A9. For large sample sizes n and m (the rule of thumb is to have both $n > 10$ and $m > 10$), U is approximately Normal.

Mann-Whitney-Wilcoxon two-sample rank-sum test

Test of two populations, $H_0: F_X = F_Y$.

Test statistic $U = \sum_i R_i$, where R_i are ranks of X_i in the combined sample of X_i and Y_i.

Null distribution: Table A9 or recursive formula (10.7).

For $n, m \geq 10$, $U \approx Normal\left(\frac{n(n+m+1)}{2}, \sqrt{\frac{nm(n+m+1)}{12}}\right)$

Assumptions: the distributions of X_i and Y_i are continuous; $F_X(t) = F_Y(t)$ under H_0; $F_X(t) < F_Y(t)$ for all t or $F_X(t) > F_Y(t)$ for all t under H_A

Example 10.14 (On-line incentives). Managers of an internet shopping portal suspect that more customers participate in on-line shopping if they are offered some incentive, such as a discount or cash back. To verify this hypothesis, they chose 12 days at random, offered a 5% discount on 6 randomly selected days but did not offer any incentives on the other 6 days. The discounts were indicated on the links leading to this shopping portal.

With the discount, the portal received (rounded to 100s) 1200, 1700, 2600, 1500, 2400, and 2100 hits. Without the discount, 1400, 900, 1300, 1800, 700, and 1000 hits were registered. Does this support the managers' hypothesis?

Solution. Let F_X and F_Y be the cdf of the number of hits without the discount and with the discount, respectively. We need to test

$$H_0: F_X = F_Y \quad \text{vs} \quad H_A: X \text{ is stochastically smaller than } Y.$$

Indeed, the discount is suspected to boost the number of hits to the shopping portal, so Y should be stochastically larger than X.

To compute the test statistic, combine all the observations and order them,

$$\underline{700}, \underline{900}, \underline{1000}, 1200, \underline{1300}, \underline{1400}, 1500, 1700, \underline{1800}, 2100, 2400, 2600.$$

X-variables are underlined here. In the combined sample, their ranks are 1, 2, 3, 5, 6, and 9, and their sum is

$$U_{\text{obs}} = \sum_{i=1}^{6} R_i = \underline{26}.$$

From Table A9 with $n = m = 6$, we find that the one-sided left-tail P-value is $p \in (0.01, 0.025]$. Although it implies some evidence that discounts help increase the on-line shopping activity, this evidence is not overwhelming. Namely, we can conclude that the evidence supporting the managers' claim is significant at any significance level $\alpha > 0.025$. ◊

Example 10.15 (PINGS). Round-trip transit times (pings) at two locations are given in Example 8.19 on p. 229. Arranged in the increasing order, they are

Location I: 0.0156, 0.0210, 0.0215, 0.0280, 0.0308, 0.0327, 0.0335, 0.0350, 0.0355, 0.0396, 0.0419, 0.0437, 0.0480, 0.0483, 0.0543 seconds

Location II: 0.0039, 0.0045, 0.0109, 0.0167, 0.0198, 0.0298, 0.0387, 0.0467, 0.0661, 0.0674, 0.0712, 0.0787 seconds

Is there evidence that the median ping depends on the location?

Let us apply the Mann-Whitney-Wilcoxon test for

$$H_0 : F_X = F_Y \quad \text{vs} \quad H_A : F_X \neq F_Y,$$

where X and Y are pings at the two locations. We observed samples of sizes $n = 15$ and $m = 12$. Among them, X-pings have ranks 4, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 19, 21, 22, and 23, and their sum is

$$U_{\text{obs}} = \sum_{i=1}^{15} R_i = \underline{213}.$$

(It may be easier to calculate the Y-ranks and subtract their sum from $1 + \ldots + 27 = (27)(28)/2 = 378$.)

Preparing to use the Normal approximation, compute

$$\begin{aligned} \mathbf{E}(U \mid H_0) &= \frac{n(n+m+1)}{2} = 210, \\ \text{Var}(U \mid H_0) &= \frac{nm(n+m+1)}{12} = 420. \end{aligned}$$

Then compute the P-value for this two-sided test,

$$\begin{aligned} P &= 2\min\left(\boldsymbol{P}\{U \le 213\}, \boldsymbol{P}\{U \ge 213\}\right) = 2\min\left(\boldsymbol{P}\{U \le 213.5\}, \boldsymbol{P}\{U \ge 212.5\}\right) \\ &= 2\min\left(\boldsymbol{P}\left\{Z \le \frac{213.5-210}{\sqrt{420}}\right\}, \boldsymbol{P}\left\{Z \ge \frac{212.5-210}{\sqrt{420}}\right\}\right) \\ &= 2\boldsymbol{P}\left\{Z \ge \frac{212.5-210}{\sqrt{420}}\right\} = 2(1-\Phi(0.12)) = 2(0.4522) = \underline{0.9044} \end{aligned}$$

(from Table A4). There is no evidence that pings at the two locations have different distributions. ◊

Example 10.16 (COMPETITION). Two computer manufacturers, A and B, compete for a profitable and prestigious contract. In their rivalry, each claims that their computers are faster. How should the customer choose between them?

It was decided to start execution of the same program simultaneously on seven computers of each company and see which ones finish earlier. As a result, two computers produced by A finished first, followed by three computers produced by B, then five computers produced by A, and finally, four computers produced by B. The actual times have never been recorded.

Use these data to test whether computers produced by A are stochastically faster.

Solution. In the absence of numerical data, we should use nonparametric tests. Apply the Mann-Whitney Wilcoxon test for testing

$$H_0 : F_A = F_B \quad \text{vs} \quad H_A : X_A \text{ is stochastically smaller than } X_B,$$

where X_A and X_B are the program execution times spent by computers produced by A and by B, and F_A and F_B are there respective cdf.

From results of this competition,

$$A, A, B, B, B, A, A, A, A, A, B, B, B, B,$$

we see that A-computers have ranks 1, 2, 6, 7, 8, 9, and 10. Their sum is

$$U_{\text{obs}} = \sum_{i=1}^{7} R_i = \underline{43}.$$

From Table A9, for the left-tail test with $n_1 = n_2 = 7$, we find that the P-value is between 0.1 and 0.2. There is no significant evidence that computers produced by company A are (stochastically) faster. ◊

Mann-Whitney-Wilcoxon test in MATLAB

The Mann-Whitney-Wilcoxon test can be done in one line using the MATLAB Statistics toolbox. The general command `ranksum(x,y)` returns a P-value for the two-sided test. To see the test statistic, write `[P,H,stat] = ranksum(x,y)`. Then `stats` will contain the test statistic U (MATLAB calculates the smaller of the sum of X-ranks and the sum of Y-ranks in the combined sample), and its standardized value $Z = (U - \mathbf{E}(U))/\operatorname{Std}(U)$; `P` will be a two-sided P-value, and `H` will just equal 1 if H_0 is rejected at the 5% level and 0 otherwise.

The following code can be used for Example 10.15.

```
x = [ 0.0156, 0.0210, 0.0215, 0.0280, 0.0308, 0.0327, 0.0335, 0.0350, ...
0.0355, 0.0396, 0.0419, 0.0437, 0.0480, 0.0483, 0.0543 ];
y = [0.0039, 0.0045, 0.0109, 0.0167, 0.0198, 0.0298, 0.0387, 0.0467, ...
0.0661, 0.0674, 0.0712, 0.0787 ];

[p,h,stats] = ranksum(x,y);
```

Variable `stats` contains the sum of Y-ranks, 165, because it is smaller than the sum of X-ranks, and $Z = 0.1220$. Also, `P` is 0.9029, slightly different from our $P = 0.9044$ because we rounded the value of Z to 0.12.

Without the Statistics Toolbox, the Mann-Whitney-Wilcoxon test statistic can be computed by simple commands

```
U = 0; XY = [X Y];                % Joint sample
for k = 1:length(X); U = U + sum( XY < X(k) ) + 1; end;
U                                 % Statistic U is the sum of X-ranks
```

Null distribution of Mann-Whitney-Wilcoxon test statistic

A recursive formula for the exact distribution of test statistic U follows the main ideas of the derivation of (10.6).

If H_0 is true, the combined sample of X and Y is a sample of identically distributed random variables of size $(n+m)$. Therefore, all $\binom{n+m}{n}$ allocations of X-variables in the ordered combined sample are equally likely. Among them, $\binom{n+m-1}{n}$ allocations have a Y-variable as the largest observation in the combined sample. In this case, deleting this largest observation from the sample will not affect statistic U, which is the sum of X-ranks. The other $\binom{n+m-1}{n-1}$ allocations have an X-variable as the largest observation, and it has a rank $(n+m)$. In this case, deleting it will reduce the sum of X-ranks by $(n+m)$.

Now, let $N_{n,m}(u)$ be the number of allocations of X-variables and Y-variables that results in $U=u$. Under H_0, we have the situation of equally likely outcomes, so the probability of $U=u$ equals

$$p_{n,m}(u) = \boldsymbol{P}\{U=u\} = \frac{N_{n,m}(u)}{\binom{n+m}{n}}.$$

From the previous paragraph,

$$\begin{aligned} N_{n,m}(u) &= N_{n,m-1}(u) + N_{n-1,m}(u-n-m) \\ &= \binom{n+m-1}{n} p_{n,m-1}(u) + \binom{n+m-1}{n-1} N_{n,m}(u-n-m), \end{aligned}$$

and therefore, dividing all parts of this equation by $\binom{n+m}{n}$, we get

$$p_{n,m}(u) = \frac{m}{n+m} p_{n,m-1}(u) + \frac{n}{n+m} p_{n-1,m}(u-n-m). \tag{10.7}$$

This is a recursive formula for the probability mass function of U. To start the recursions, let's notice that without any X-variables, $p_{0,m}=0$, and without any Y variables, $p_{n,0} = \sum_{i=1}^{n} i = n(n+1)/2$. These formulas are used to construct Table A9.

Normal approximation

When both $n \geq 10$ and $m \geq 10$, statistic U has approximately Normal distribution. Some work is needed to derive its parameters $\mathbf{E}(U)$ and $\mathrm{Var}(U)$, but here are the main steps.

We notice that each X-rank equals

$$R_i = 1 + \#\{j : X_j < X_i\} + \#\{k : Y_k < X_i\}$$

(each # denotes the number of elements in the corresponding set). Therefore,

$$U = \sum_{i=1}^{n} R_i = n + \binom{n}{2} + \sum_{i=1}^{n} \xi_i,$$

where $\xi_i = \sum_{k=1}^{m} I\{Y_k < X_i\}$ is the number of Y-variables that are smaller than X_i.

Under the hypothesis H_0, all X_i and Y_k are identically distributed, so $\boldsymbol{P}\{Y_k < X_i\} = 1/2$. Then, $\mathbf{E}(\xi_i) = m/2$, and

$$\mathbf{E}(U) = n + \binom{n}{2} + \frac{nm}{2} = \frac{n(n+m+1)}{2}.$$

For $\text{Var}(U) = \text{Var}\sum_i \xi_i = \sum_i \sum_j \text{Cov}(\xi_i, \xi_j)$, we have to derive $\text{Var}(\xi_i) = (m^2 + 2m)/12$ and $\text{Cov}(\xi_i, \xi_j) = m/12$. This is not trivial, but you can fill in all the steps, as a nice non-required exercise! If you decide to do that, notice that all X_i and Y_k have the same distribution, and therefore, for any $i \neq j$ and $k \neq l$,

$$\boldsymbol{P}\{Y_k < X_i \cap Y_l < X_i\} = \boldsymbol{P}\{X_i \text{ is the smallest of } X_i, Y_k, Y_l\} = \frac{1}{3},$$

$$\boldsymbol{P}\{Y_k < X_i \cap Y_l < X_j\} = \frac{6 \text{ favorable allocations of } X_i, X_j, Y_k, Y_l}{4! \text{ allocations of } X_i, X_j, Y_k, Y_l} = \frac{1}{4}.$$

Substituting into $\text{Var}(U)$ and simplifying, we obtain

$$\text{Var}(U) = \frac{nm(n+m+1)}{12}.$$

10.3 Bootstrap

Bootstrap is used to estimate population parameters by Monte Carlo simulations when it is too difficult to do it analytically. When computers became powerful enough to handle large-scale simulations, the bootstrap methods got very popular for their ability to evaluate properties of various estimates.

Consider, for example, the *standard errors*. From the previous chapters, we know the standard errors of a sample mean and a sample proportion,

$$s(\bar{X}) = \frac{\sigma}{\sqrt{n}} \text{ and } s(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}},$$

and they are quite simple. Well, try to derive standard errors of other statistics - sample median, sample variance, sample interquartile range, and so on. Each of these will require a substantial amount of work.

Many complicated estimates are being used nowadays in modern Statistics. How can one evaluate their performance, estimate their standard error, bias, etc.?

The difficulty is that we observe an estimator $\widehat{\theta}$ only once. That is, we have one sample $\mathcal{S} = (X_1, \ldots, X_n)$ from the population $\mathcal{P}$, and from this sample we compute $\widehat{\theta}$. We would very much like to observe many $\widehat{\theta}$'s and then compute their sample variance, for example, but we don't have this luxury. From one sample, we observe only one $\widehat{\theta}$, and this is just not enough to compute its sample variance!

10.3.1 Bootstrap distribution and all bootstrap samples

In 1970s, an American mathematician, Bradley Efron, Professor at Stanford University, proposed a rather simple approach. He called it **bootstrap** referring to the idiom *"to pull oneself up by one's bootstraps"*, which means to find a solution relying on your own sources and without any help from outside (a classical example is Baron Münchausen from the 18th century collection of tales by R. E. Raspe, who avoided drowning in a lake by pulling

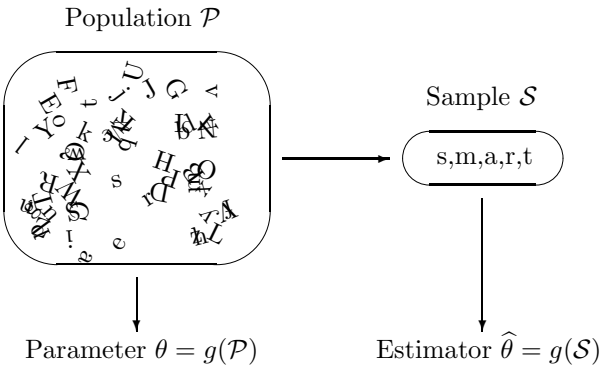

FIGURE 10.5: *Parameter θ and its estimator $\widehat{\theta}$ computed by the same mechanism g applied to the population and to the sample.*

himself from the water by his own hair!). In our situation, even though we really need several samples to explore the properties of an estimator $\widehat{\theta}$, we'll manage to do it with what we have, which is just one sample.

To start, let's notice that many commonly used statistics are constructed in the same way as the corresponding population parameters. We are used to estimating a population mean μ by a sample mean $\bar{X}$, a population variance σ^2 by a sample variance s^2, population quantiles q_p by sample quantiles $\widehat{q}_p$, and so on. To estimate a certain parameter θ, we collect a random sample $\mathcal{S}$ from $\mathcal{P}$ and essentially compute the estimator $\widehat{\theta}$ from this sample by the same mechanism as θ was computed from the whole population!

In other words, there is a function g that one can use to compute a parameter θ from the population $\mathcal{P}$ (Figure 10.5). Then

$$\theta = g(\mathcal{P}) \text{ and } \widehat{\theta} = g(\mathcal{S}).$$

Example 10.17 (Population mean and sample mean). Imagine a strange calculator that can only do one operation g – averaging. Give it 4 and 6, and it returns their average $g(4, 6) = 5$. Give it 3, 7, and 8, and it returns $g(3, 7, 8) = 6$.

Give it the whole population $\mathcal{P}$, and it returns the parameter $\theta = g(\mathcal{P}) = \mu$. Give it a sample $\mathcal{S} = (X_1, \ldots, X_n)$, and it returns the estimator $\widehat{\theta} = g(X_1, \ldots, X_n) = \bar{X}$. ◊

You can imagine similar calculators for the variances, medians, and other parameters and their estimates. Essentially, each estimator $\widehat{\theta}$ is a mechanism that copies the way a parameter θ is obtained from the population and then applies it to the sample.

Bradley Efron proposed one further step. Suppose some estimator is difficult to figure out. For example, we are interested in the variance of a sample median, $\eta = \text{Var}(\widehat{M}) = h(\mathcal{P})$. This mechanism h consists of taking all possible samples from the population, taking their sample medians $\widehat{M}$, and then calculating their variance,

$$\eta = h(\mathcal{P}) = \mathbf{E}(\widehat{M} - \mathbf{E}\widehat{M})^2.$$

Of course, we typically cannot observe all possible samples; that's why we cannot compute $h(\mathcal{P})$, and parameter η is unknown. How can we estimate it based on just one sample $\mathcal{S}$?

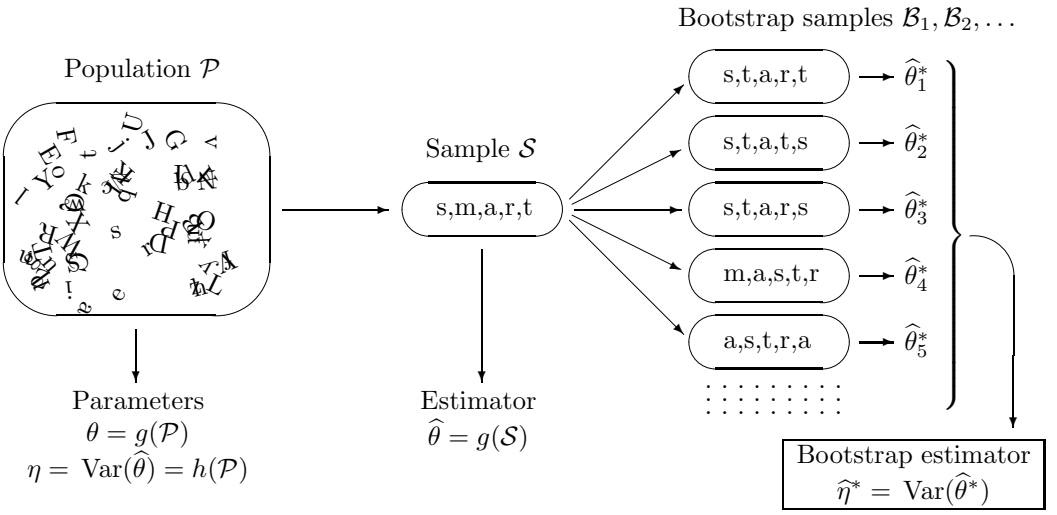

FIGURE 10.6: *Bootstrap procedure estimates* $\eta = \mathrm{Var}(\widehat{\theta})$ *by the variance of* $\widehat{\theta}^*_i$*'s, obtained from bootstrap samples.*

Efron's **bootstrap approach** is to apply the same mechanism to the sample $\mathcal{S}$. That is, take all possible samples from $\mathcal{S}$, compute their medians, and then compute the sample variance of those, as on Figure 10.6. It may sound strange, but yes, we are proposing to create more samples from one sample $\mathcal{S}$ given to us. After all, exactly the same algorithm was used to compute $\eta = h(\mathcal{P})$. The advantage is that we know the observed sample $\mathcal{S}$, and therefore, we can perform all the bootstrap steps and estimate η.

Sampling from the observed sample $\mathcal{S}$ is a statistical technique called *resampling*. Bootstrap is one of the resampling methods. The obtained samples from $\mathcal{S}$ are called *bootstrap samples*. Each bootstrap sample $\mathcal{B}_j = (X^*_{1j}, \dots, X^*_{nj})$ consists of values X^*_{ij} sampled from $\mathcal{S} = (X_1, \dots, X_n)$ independently and with equal probabilities. That is,

$$X^*_{ij} = \begin{cases} X_1 & \text{with probability } 1/n \\ X_2 & \text{with probability } 1/n \\ \dots & \dots \quad \dots \quad \dots \quad \dots \\ X_n & \text{with probability } 1/n \end{cases}$$

This is *sampling with replacement*, which means that the same observation X_i can be sampled more than once. An asterisk $(^*)$ is used to denote the contents of bootstrap samples.

DEFINITION 10.3

A **bootstrap sample** is a random sample drawn with replacement from the observed sample $\mathcal{S}$ of the same size as $\mathcal{S}$.

The distribution of a statistic across bootstrap samples is called a **bootstrap distribution**.

An estimator that is computed on basis of bootstrap samples is a **bootstrap estimator**.

i	$\mathcal{B}_i$	$\widehat{M}_i$
1	(2, 2, 2)	2
2	(2, 2, 5)	2
3	(2, 2, 7)	2
4	(2, 5, 2)	2
5	(2, 5, 5)	5
6	(2, 5, 7)	5
7	(2, 7, 2)	2
8	(2, 7, 5)	5
9	(2, 7, 7)	7

i	$\mathcal{B}_i$	$\widehat{M}_i$
10	(5, 2, 2)	2
11	(5, 2, 5)	5
12	(5, 2, 7)	5
13	(5, 5, 2)	5
14	(5, 5, 5)	5
15	(5, 5, 7)	5
16	(5, 7, 2)	5
17	(5, 7, 5)	5
18	(5, 7, 7)	7

i	$\mathcal{B}_i$	$\widehat{M}_i$
19	(7, 2, 2)	2
20	(7, 2, 5)	5
21	(7, 2, 7)	7
22	(7, 5, 2)	5
23	(7, 5, 5)	5
24	(7, 5, 7)	7
25	(7, 7, 2)	7
26	(7, 7, 5)	7
27	(7, 7, 7)	7

TABLE 10.1: *All bootstrap samples $\mathcal{B}_i$ drawn from $\mathcal{S}$ and the corresponding sample medians for Example 10.18.*

Example 10.18 (VARIANCE OF A SAMPLE MEDIAN). Suppose that we observed a small sample $\mathcal{S} = (2, 5, 7)$ and estimated the population median M with the sample median $\widehat{M} = 5$. How can we estimate its variance $\text{Var}(\widehat{M})$?

Solution. Table 10.1 lists all $3^3 = 27$ equally likely bootstrap samples that can be drawn from $\mathcal{S}$. Among these, 7 samples have $\widehat{M}_i^* = 2$, 13 samples have $\widehat{M}_i^* = 5$, and 7 samples have $\widehat{M}_i^* = 7$. So, the *bootstrap distribution* of a sample median is

$$P^*(2) = 7/27, \quad P^*(5) = 13/27, \quad P^*(7) = 7/27. \tag{10.8}$$

We use it to estimate $h(\mathcal{P}) = \text{Var}(\widehat{M})$ with the *bootstrap estimator*

$$\begin{aligned}
\widehat{\text{Var}}^*(\widehat{M}) &= h(\mathcal{S}) = \sum_x x^2 P^*(x) - \left(\sum_x x P^*(x) \right)^2 \\
&= (4)\left(\frac{7}{27}\right) + (25)\left(\frac{13}{27}\right) + (49)\left(\frac{7}{27}\right) - \left\{ (2)\left(\frac{7}{27}\right) + (5)\left(\frac{13}{27}\right) + (7)\left(\frac{7}{27}\right) \right\}^2 \\
&= \underline{3.303}. \qquad \Diamond
\end{aligned}$$

Here is a summary of the bootstrap method that we applied in this example.

Bootstrap (all bootstrap samples)

To estimate parameter η of the distribution of $\widehat{\theta}$:
1. Consider all possible bootstrap samples drawn with replacement from the given sample $\mathcal{S}$ and statistics $\widehat{\theta}^*$ computed from them.
2. Derive the bootstrap distribution of $\widehat{\theta}^*$.
3. Compute the parameter of this bootstrap distribution that has the same meaning as η.

All right, one may say, this certainly works for a "toy" sample of size three. But how about

bigger samples? We can list all n^n possible bootstrap samples for very small n. However, a rather modest sample of size $n = 60$ can produce $60^{60} \approx 4.9 \cdot 10^{106}$ different bootstrap samples, which is almost five million times larger than *the googol*!

Certainly, we are not going to list a googol of bootstrap samples. Instead, we'll discuss two alternative approaches. The first one proposes to compute the bootstrap distribution without listing all the bootstrap samples. This, however, is still feasible only in relatively simple situations. The second solution, by far the most popular one, uses Monte Carlo simulations to produce a large number b of bootstrap samples. Although this way we may not get *all* possible bootstrap samples, results will be very close for large b.

Bootstrap distribution

The only reason for considering all possible bootstrap samples in Example 10.18 was to find the distribution (10.8) and then to obtain the bootstrap estimator $\widehat{\eta}^* = \widehat{\text{Var}}(\widehat{M}^*)$ from it.

Sometimes it is possible to compute the bootstrap distribution without listing all bootstrap samples. Here is an example.

Example 10.19 (Bias of a sample median). A sample median may be biased or unbiased for estimating the population median. It depends on the underlying distribution of data. Suppose we observed a sample

$$\mathcal{S} = (3, 5, 8, 5, 5, 8, 5, 4, 2).$$

Find a bootstrap estimator of $\eta = \text{Bias}(\widehat{M})$, the bias of the sample median.

Solution. First, find the bootstrap distribution of a sample median $\widehat{M}^*$. Based on the given sample of size $n = 9$, the sample median of bootstrap samples can be equal to 2, 3, 4, 5, or 8. Let us compute the probability of each value.

Sampling from $\mathcal{S}$, values $X_{ij}^* = 2$, 3, and 4 appear with probability 1/9 because only one of each of them appears in the given sample. Next, $X_{ij}^* = 5$ with probability 4/9, and $X_{ij}^* = 8$ with probability 2/9.

Now, the sample median $\widehat{M}_i^*$ in any bootstrap sample $\mathcal{B}_i$ is the central or the 5th smallest observation. Thus, it equals 2 if at least 5 of 9 values in $\mathcal{B}_i$ equal 2. The probability of that is

$$P^*(2) = \boldsymbol{P}(Y \geq 5) = \sum_{y=5}^{9} \binom{9}{y} \left(\frac{1}{9}\right)^y \left(\frac{8}{9}\right)^{9-y} = 0.0014$$

for a Binomial$(n = 9, p = 1/9)$ variable Y.

Similarly, $\widehat{M}_i^* \leq 3$ if at least 5 of 9 values in $\mathcal{B}_i$ do not exceed 3. The probability of that is

$$F^*(3) = \boldsymbol{P}(Y \geq 5) = \sum_{y=5}^{9} \binom{9}{y} \left(\frac{2}{9}\right)^y \left(\frac{7}{9}\right)^{9-y} = 0.0304$$

for a Binomial$(n = 9, p)$ variable Y, where $p = 2/9$ is a probability to sample either $X_{ij}^* = 2$ or $X_{ij}^* = 3$.

Proceeding in a similar fashion, we get

$$\begin{aligned} F^*(4) &= \sum_{y=5}^{9} \binom{9}{y} \left(\frac{3}{9}\right)^y \left(\frac{6}{9}\right)^{9-y} = 0.1448, \\ F^*(5) &= \sum_{y=5}^{9} \binom{9}{y} \left(\frac{7}{9}\right)^y \left(\frac{2}{9}\right)^{9-y} = 0.9696, \text{ and} \\ F^*(8) &= 1. \end{aligned}$$

From this cdf, we can find the bootstrap probability mass function of $\widehat{M}_i^*$,

$$P^*(2) = 0.0014, \quad P^*(3) = 0.0304 - 0.0014 = 0.0290, \quad P^*(4) = 0.1448 - 0.0304 = 0.1144,$$

$$P^*(5) = 0.9696 - 0.1448 = 0.8248, \quad P^*(8) = 1 - 0.9696 = 0.0304.$$

From this, the bootstrap estimator of $\mathbf{E}(\widehat{M})$ is the expected value of the bootstrap distribution,

$$\mathbf{E}^*(\widehat{M}_i^*) = (2)(0.0014) + (3)(0.0290) + (4)(0.1144) + (5)(0.8248) + (8)(0.0304) = 4.9146.$$

Last step! The bias of $\widehat{M}$ is defined as $h(\mathcal{P}) = \text{Bias}(\widehat{M}) = \mathbf{E}(\widehat{M}) - M$. We have estimated the first term, $\mathbf{E}(\widehat{M})$. Following the bootstrap ideas, what should we use as an estimator of M, the second term of the bias? The answer is simple. We have agreed to estimate $g(\mathcal{P})$ with $g(\mathcal{S})$, so we just estimate the population median $g(\mathcal{P}) = M$ with the sample median $g(\mathcal{S}) = \widehat{M} = 5$ obtained from our original sample $\mathcal{S}$.

The bootstrap estimator of $\text{Bias}(\widehat{M}) = \mathbf{E}(\widehat{M}) - M$ (based on all possible bootstrap samples) is

$$\eta(\mathcal{S}) = \widehat{\text{Bias}}(\widehat{M}) = \mathbf{E}^*(\widehat{M}) - \widehat{M} = 4.9146 - 5 = \boxed{-0.0852}\,.$$

$\Diamond$

Although the sample in Example 10.19 is still rather small, the method presented can be extended to a sample of any size. Manual computations may be rather tedious here, but one can write a suitable computer code.

10.3.2 Computer generated bootstrap samples

Modern Statistics makes use of many complicated estimates. As the earliest examples, Efron used bootstrap to explore properties of the sample correlation coefficient, the trimmed mean[1], and the excess error[2], but certainly, there are more complex situations. Typically, samples will be too large to list all the bootstrap samples, as in Table 10.1, and the statistics will be too complicated to figure out their bootstrap distributions, as in Example 10.19.

This is where *Monte Carlo simulations* kick in. Instead of listing all possible bootstrap samples, we use a computer to generate a large number b of them. The rest follows our general scheme on Figure 10.6, p. 330.

[1]Trimmed mean is a version of a sample mean, where a certain portion of the smallest and the largest observations is dropped before computing the arithmetic average. Trimmed means are not sensitive to a few extreme observations, and they are used if extreme outliers may be suspected in the sample.

[2]This measures how well one can estimate the error of prediction in regression analysis. We study regression in the next chapter.

Bootstrap (generated bootstrap samples)

To estimate parameter η of the distribution of $\widehat{\theta}$:
1. Generate a large number b of bootstrap samples drawn with replacement from the given sample $\mathcal{S}$.
2. From each bootstrap sample $\mathcal{B}_i$, compute statistic $\widehat{\theta}_i^*$ the same way as $\hat{\theta}$ is computed from the original sample $\mathcal{S}$.
3. Estimate parameter η from the obtained values of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_b^*$.

This is a classical bootstrap method of evaluating properties of parameter estimates. Do you think it is less accurate than the one based on *all* the bootstrap samples? Notice that b, the number of generated bootstrap samples, can be very large. Increasing b gives more work to your computer, but it does not require a more advanced computer code or a larger original sample. And of course, as $b \to \infty$, our estimator of η becomes just as good as if we had a complete list of bootstrap samples. Typically, thousands or tens of thousands of bootstrap samples are being generated.

MATLAB Demo. The following MATLAB code generates b bootstrap samples from the given sample $\boldsymbol{X} = (X_1, \ldots, X_n)$.

```
n = length(X);            % Sample size
U = ceil(n*rand(b,n));    % A b × n matrix of random integers from 1 to n
B = X(U);                 % A matrix of bootstrap samples.
                          % The i-th bootstrap sample is in the i-th row.
```

For example, based on a sample $\boldsymbol{X} = (10, 20, 30, 40, 50)$, we can generate the following matrix U of random indices which will determine the following matrix B of bootstrap samples,

$$U = \begin{pmatrix} 1 & 4 & 5 & 5 & 1 \\ 3 & 2 & 3 & 1 & 5 \\ 3 & 4 & 5 & 2 & 3 \\ 1 & 4 & 1 & 2 & 3 \\ 2 & 4 & 3 & 5 & 1 \\ 1 & 3 & 1 & 3 & 5 \\ 4 & 1 & 5 & 5 & 4 \\ 2 & 2 & 1 & 1 & 2 \\ 3 & 5 & 4 & 2 & 3 \\ 1 & 1 & 5 & 1 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} \mathcal{B}_1 \\ \mathcal{B}_2 \\ \mathcal{B}_3 \\ \mathcal{B}_4 \\ \mathcal{B}_5 \\ \mathcal{B}_6 \\ \mathcal{B}_7 \\ \mathcal{B}_8 \\ \mathcal{B}_9 \\ \mathcal{B}_{10} \end{pmatrix} = \begin{pmatrix} 10 & 40 & 50 & 50 & 10 \\ 30 & 20 & 30 & 10 & 50 \\ 30 & 40 & 50 & 20 & 30 \\ 10 & 40 & 10 & 20 & 30 \\ 20 & 40 & 30 & 50 & 10 \\ 10 & 30 & 10 & 30 & 50 \\ 40 & 10 & 50 & 50 & 40 \\ 20 & 20 & 10 & 10 & 20 \\ 30 & 50 & 40 & 20 & 30 \\ 10 & 10 & 50 & 10 & 30 \end{pmatrix}$$

If b and n are so large that storing the entire matrices U and B requires too many computer resources, we can generate bootstrap samples in a do-loop, one at a time, and keep the statistics $\widehat{\theta}_i$ only.

In fact, MATLAB has a special command `bootstrp` for generating bootstrap samples and computing estimates from them. For example, the code

```
M = bootstrp(50000,@median,S);
```

takes the given sample $\mathcal{S}$, generates $b = 50,000$ bootstrap samples from it, computes medians from each of them, and stores these median in a vector M. After that, we can compute various sample statistics of M, the mean, standard deviation, etc., and use them to evaluate the properties of a sample median.

Example 10.20 (Bias of a sample median (continued)). Based on the sample

$$\mathcal{S} = (3,\ 5,\ 8,\ 5,\ 5,\ 8,\ 5,\ 4,\ 2)$$

from Example 10.19, we estimated the population median θ with the sample median $\hat{\theta} = 5$. Next, we investigate the properties of $\hat{\theta}$. The MATLAB code

```
S  =  [3 5 8 5 5 8 5 4 2];
b  =  100000;
M  =  bootstrp(b,@median,S);
biasM   =  mean(M) - median(S);
sterrM  =  std(M);
prob4   =  mean(M > 4);
```

is used to estimate the bias of $\hat{\theta}$, its standard error, and the probability of $\hat{\theta} > 4$. Vector M contains the bootstrap statistics $\hat{\theta}_1^*, \ldots, \hat{\theta}_b^*$. Based on $b = 100,000$ bootstrap samples, we obtained

$$\begin{aligned}
\widehat{\text{Bias}^*}(\hat{\theta}) &= \bar{M} - \hat{\theta} &&= -0.0858, \\
s^*(\hat{\theta}) &= s(M) &&= 0.7062, \qquad \text{and} \\
\widehat{\boldsymbol{P}}^*(\hat{\theta} > 4) &= \frac{\#\left\{i\,:\,\hat{\theta}_i^* > 4\right\}}{b} &&= 0.8558.
\end{aligned}$$

◊

10.3.3 Bootstrap confidence intervals

In the previous chapters, we learned how to construct confidence intervals for the population mean, proportion, variance, and also, difference of means, difference of proportions, and ratio of variances. Normal, T, χ^2, and F distributions were used in these cases. These methods required either a Normal distribution of the observed data or sufficiently large samples.

There are many situations, however, where these conditions will not hold, or the distribution of the test statistic is too difficult to obtain. Then the problem becomes nonparametric, and we can use *bootstrap* to construct approximately $(1 - \alpha)100\%$ confidence intervals for the population parameters.

Two methods of bootstrap confidence intervals are rather popular.

Parametric method, based on the bootstrap estimation of the standard error

This method is used when we need a confidence interval for a parameter θ, and its estimator $\hat{\theta}$ has approximately Normal distribution.

In this case, we compute $s^*(\hat{\theta})$, the bootstrap estimator of the standard error $\sigma(\hat{\theta})$, and use it to compute the approximately $(1 - \alpha)100\%$ confidence interval for θ,

Parametric bootstrap confidence interval

$$\boxed{\hat{\theta} \pm z_{\alpha/2} s^*(\hat{\theta})} \tag{10.9}$$

It is similar to our usual formula for the confidence interval in the case of Normal distribution (9.3), and $z_{\alpha/2} = q_{1-\alpha/2}$ in it, as always, denotes the $(1-\alpha/2)$-quantile from the Standard Normal distribution.

Example 10.21 (Confidence interval for a correlation coefficient). Example 8.20 on p. 231 contains the data on the number of times X the antivirus software was launched on 30 computers during 1 month and the number Y of detected worms,

X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0

X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

Scatter plots on Figure 8.11 showed existence of a negative correlation between X and Y, which means that in general, the number of worms reduces if the antivirus software is used more often.

Next, the computer manager asks for a 90% confidence interval for the correlation coefficient ρ between X and Y.

The following MATLAB code can be used to solve this problem.

```
alpha = 0.10;
X = [30 30 30 30 30 30 30 30 30 30 30 15 15 15 10 10 10 6 6 5 5 5 4 4 4 4 4 1 1 1];
Y = [0 0 1 0 0 0 1 1 0 0 0 0 1 1 0 0 2 0 4 1 2 0 2 1 0 1 0 6 3 1];
r = corrcoef(X,Y); % correlation coefficient from the given sample
r = r(2,1); % because corrcoef returns a matrix of correlation coefficients
b = 10000; n = length(X);
U = ceil(n*rand(b,n));
BootX = X(U); BootY = Y(U); % Values X and Y of generated bootstrap samples
BootR = zeros(b,1); % Initiate the vector of bootstrap corr. coefficients
for i=1:b;
BR = corrcoef(BootX(i,:),BootY (i,:));
BootR(i) = BR(2,1);
end;
s = std(BootR); % Bootstrap estimator of the standard error of r
CI = [ r + s*norminv(alpha/2,0,1), r + s*norminv(1-alpha/2,0,1) ];
disp(CI) % Bootstrap confidence interval
```

As a result of this code, we get the sample correlation coefficient $r = -0.4533$, and also, $b = 10,000$ bootstrap correlation coefficients $r_1^*, \ldots, r_b^*$ obtained from b generated bootstrap samples.

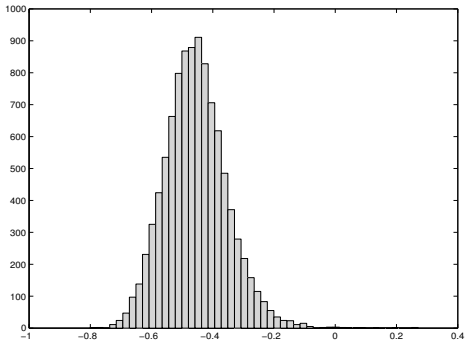

FIGURE 10.7 *The histogram of bootstrap correlation coefficients.*

Next, we notice that for the sample of size $n = 30$, r has approximately Normal distribution. For example, this can be confirmed by the histogram of bootstrap correlation coefficients on Figure 10.7.

Applying the parametric method, we compute $s^*(r) = 0.1028$, the standard error of $r_1^*, \ldots, r_b^*$, and use it to construct the 90% confidence interval

$$\begin{aligned} r \pm z_{\alpha/2} s^*(r) &= -0.4533 \pm (1.645)(0.1028) \\ &= \underline{[-0.2841, -0.6224]} \end{aligned}$$

◊

Nonparametric method, based on the bootstrap quantiles

Equation 10.9 simply estimates two quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ of the distribution of statistic $\hat{\theta}$, so that

$$\boldsymbol{P}\left\{q_{\alpha/2} \le \hat{\theta} \le q_{1-\alpha/2}\right\} = 1 - \alpha. \tag{10.10}$$

Since $\hat{\theta}$ estimates parameter θ, this becomes an approximately $(1-\alpha)100\%$ confidence interval for θ.

This method fails if the distribution of $\hat{\theta}$ is not Normal. The coverage probability in (10.10) may be rather different from $(1-\alpha)$ in this case. However, the idea to construct a confidence interval based on the quantiles of $\hat{\theta}$ is still valid.

The quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ of the distribution of $\hat{\theta}$ will then be estimated from the bootstrap samples. To do this, we generate b bootstrap samples, compute statistic $\hat{\theta}^*$ for each of them, and determine the sample quantiles $\hat{q}^*_{\alpha/2}$ and $\hat{q}^*_{1-\alpha/2}$ from them. These quantiles become the end points of the $(1-\alpha)100\%$ confidence interval for θ.

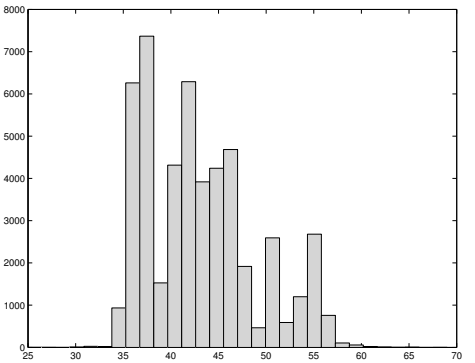

FIGURE 10.8 *The histogram of bootstrap medians.*

Nonparametric bootstrap confidence interval for parameter θ

$$\left[\hat{q}^*_{\alpha/2},\ \hat{q}^*_{1-\alpha/2}\right],$$

where $q^*_{\alpha/2}$ and $\hat{q}^*_{1-\alpha/2}$ are quantiles of the distribution of $\hat{\theta}$ estimated from bootstrap samples (10.11)

Example 10.22 (Confidence interval for the median CPU time). The population median was estimated in Example 8.12 on p. 217, based on the following observed CPU times:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

Let us now compute the 95% bootstrap confidence interval for the median CPU time.

```
alpha = 0.05;

X = [ 70 36 43 69 82 48 34 62 35 15 59 139 46 37 42 ...
      30 55 56 36 82 38 89 54 25 35 24 22 9 56 19]';

b = 50000; n = length(X);
U = ceil(n*rand(b,n));

BootX = X(U); BootM = zeros(b,1);
for i=1:b; BootM(i) = median(BootX(i,:)); end;

CI = [ quantile(BootM,alpha/2), quantile(BootM,1-alpha/2) ]
```

This MATLAB program generates $b = 50,000$ bootstrap samples and computes their sample medians $\hat{\theta}_1^*, \ldots, \hat{\theta}_b^*$ (variable `BootM`). Based on these sample medians, the 0.025- and 0.975-quantiles are calculated. The 95% confidence interval `CI` then stretches between these quantiles, from $\hat{q}^*_{0.025}$ to $\hat{q}^*_{0.975}$.

This program results in a confidence interval [35.5, 55.5].

By the way, the histogram of bootstrap medians $\hat{\theta}_1^*, \ldots, \hat{\theta}_b^*$ on Figure 10.8 shows a rather non-Normal distribution. We were essentially forced to use the nonparametric approach. ◊

MATLAB has a special command `bootci` for the construction of bootstrap confidence intervals. The problem in Example 10.22 can be solved by just one command

```
bootci(50000,{@median,X},'alpha',0.05,'type','percentile')
```

where 0.05 denotes the α-level, and `'percentile'` requests a confidence interval computed by the method of percentiles. This is precisely the nonparametric method that we have just discussed. Replace it with type `'normal'` to obtain a parametric bootstrap confidence interval that is based on the Normal distribution (10.9).

10.4 Bayesian inference

Interesting results and many new statistical methods can be obtained when we take a rather different look at statistical problems.

The difference is in our treatment of *uncertainty*.

So far, random samples were the only source of uncertainty in all the discussed statistical problems. The only distributions, expectations, and variances considered so far were distributions, expectations, and variances of data and various statistics computed from data. Population parameters were considered fixed. Statistical procedures were based on the distribution of data given these parameters,

$$f(\boldsymbol{x} \mid \theta) = f(X_1, \ldots, X_n \mid \theta).$$

This is the **frequentist approach**. According to it, all probabilities refer to random samples of data and possible long-run frequencies, and so do such concepts as unbiasedness, consistency, confidence level, and significance level:

- an estimator $\hat{\theta}$ is *unbiased* if in a long run of random samples, it averages to the parameter θ;
- a test has significance level α if in a long run of random samples, $(100\alpha)\%$ of times the true hypothesis is rejected;
- an interval has confidence level $(1-\alpha)$ if in a long run of random samples, $(1-\alpha)100\%$ of obtained confidence intervals contain the parameter, as shown in Figure 9.2, p. 248;
- and so on.

However, there is another approach: the **Bayesian approach**. According to it, uncertainty is also attributed to the unknown parameter θ. Some values of θ are more likely than others. Then, as long as we talk about the likelihood, we can define a whole distribution of values of θ. Let us call it a *prior distribution*. It reflects our ideas, beliefs, and past experiences about the parameter before we collect and use the data.

Example 10.23 (SALARIES). What do you think is the average starting annual salary of a Computer Science graduate? Is it \$20,000 per year? Unlikely, that's too low. Perhaps, \$200,000 per year? No, that's too high for a fresh graduate. Between \$40,000 and \$70,000 sounds like a reasonable range. We can certainly collect data on 100 recent graduates, compute their average salary, and use it as an estimate, but before that, we already have our beliefs on what the mean salary may be. We can express it as some distribution with the most likely range between \$40,000 and \$70,000 (Figure 10.9). ◊

Collected data may force us to change our initial idea about the unknown parameter. Probabilities of different values of θ may change. Then we'll have a *posterior distribution* of θ.

One benefit of this approach is that we no longer have to explain our results in terms of a "long run." Often we collect just one sample for our analysis and don't experience any long runs of samples. Instead, with the Bayesian approach, we can state the result in terms of the posterior distribution of θ. For example, we can clearly state the *posterior probability* for a parameter to belong to the obtained confidence interval, or the *posterior probability* that the hypothesis is true.

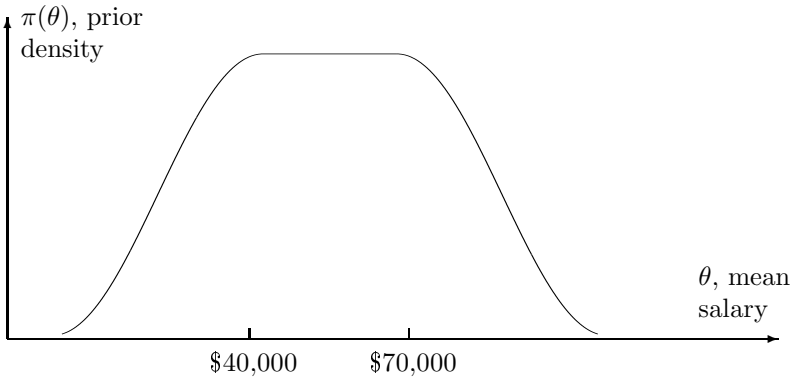

FIGURE 10.9: *Our prior distribution for the average starting salary.*

10.4.1 Prior and posterior

Now we have two sources of information to use in our Bayesian inference:

1. collected and observed data;
2. prior distribution of the parameter.

Here is how these two pieces are combined via the **Bayes formula** (see p. 29 and Figure 10.10).

Prior to the experiment, our knowledge about the parameter θ is expressed in terms of the **prior distribution** (prior pmf or pdf)

$$\pi(\theta).$$

The observed sample of data $\boldsymbol{X} = (X_1, \ldots, X_n)$ has distribution (pmf or pdf)

$$f(\boldsymbol{x}|\theta) = f(x_1, \ldots, x_n|\theta).$$

This distribution is conditional on θ. That is, different values of the parameter θ generate different distributions of data, and thus, conditional probabilities about $\boldsymbol{X}$ generally depend on the condition, θ.

Observed data add information about the parameter. The updated knowledge about θ can be expressed as the **posterior distribution**.

Posterior distribution

$$\boxed{\pi(\theta|\boldsymbol{x}) = \pi(\theta|\boldsymbol{X} = \boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})}.} \qquad (10.12)$$

Posterior distribution of the parameter θ is now conditioned on data $\boldsymbol{X} = \boldsymbol{x}$. Naturally, conditional distributions $f(\boldsymbol{x}|\theta)$ and $\pi(\theta|\boldsymbol{x})$ are related via the Bayes Rule (2.9).

According to the Bayes Rule, the denominator of (10.12), $m(\boldsymbol{x})$, represents the unconditional distribution of data $\boldsymbol{X}$. This is the **marginal distribution** (pmf or pdf) of the sample $\boldsymbol{X}$. Being unconditional means that it is constant for different values of the parameter θ. It can be computed by the *Law of Total Probability* (p. 31) or its continuous-case version below.

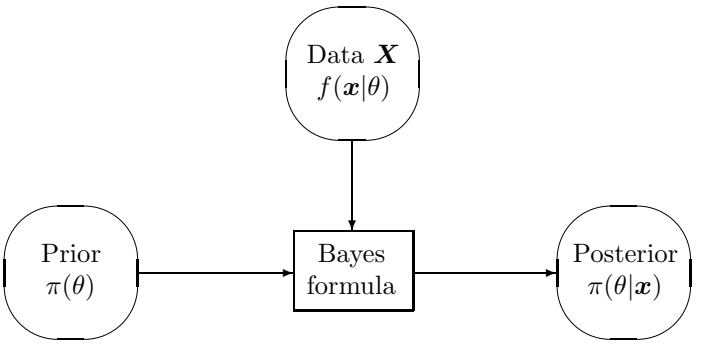

FIGURE 10.10: *Two sources of information about the parameter θ.*

Marginal distribution of data

$$m(\boldsymbol{x}) = \sum_{\theta} f(x|\theta)\pi(\theta)$$
for discrete prior distributions π

$$m(\boldsymbol{x}) = \int_{\theta} f(x|\theta)\pi(\theta)d\theta$$
for continuous prior distributions π (10.13)

Example 10.24 (QUALITY INSPECTION). A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 10%. We have to decide whether to accept or to reject the shipment based on θ, the proportion of defective parts.

Before we see the real data, let's assign a 50-50 chance to both suggested values of θ, i.e.,

$$\pi(0.05) = \pi(0.10) = 0.5.$$

A random sample of 20 parts has 3 defective ones. Calculate the posterior distribution of θ.

Solution. Apply the Bayes formula (10.12). Given θ, the distribution of the number of defective parts X is Binomial$(n = 20, \theta)$. For $x = 3$, Table A2 gives

$$f(x \mid \theta = 0.05) = F(3 \mid \theta = 0.05) - F(2 \mid \theta = 0.05) = 0.9841 - 0.9245 = 0.0596$$

and

$$f(x \mid \theta = 0.10) = F(3 \mid \theta = 0.10) - F(2 \mid \theta = 0.10) = 0.8670 - 0.6769 = 0.1901.$$

The marginal distribution of X (for $x = 3$) is

$$\begin{aligned} m(3) &= f(x \mid 0.05)\pi(0.05) + f(x \mid 0.10)\pi(0.10) \\ &= (0.0596)(0.5) + (0.1901)(0.5) = 0.12485. \end{aligned}$$

Posterior probabilities of $\theta = 0.05$ and $\theta = 0.10$ are now computed as

$$\pi(0.05 \mid X = 3) = \frac{f(x \mid 0.05)\pi(0.05)}{m(3)} = \frac{(0.0596)(0.5)}{0.1248} = 0.2387;$$

$$\pi(0.10 \mid X = 3) = \frac{f(x \mid 0.10)\pi(0.10)}{m(3)} = \frac{(0.1901)(0.5)}{0.1248} = 0.7613.$$

Conclusion. In the beginning, we had no preference between the two suggested values of θ. Then we observed a rather high proportion of defective parts, 3/20=15%. Taking this into account, $\theta = 0.10$ is now about three times as likely than $\theta = 0.05$. ◊

NOTATION			
	$\pi(\theta)$	=	prior distribution
	$\pi(\theta \mid \boldsymbol{x})$	=	posterior distribution
	$f(x\|\theta)$	=	distribution of data (model)
	$m(\boldsymbol{x})$	=	marginal distribution of data
	$\boldsymbol{X}$	=	$(X_1, \ldots, X_n)$, sample of data
	$\boldsymbol{x}$	=	$(x_1, \ldots, x_n)$, observed values of $X_1, \ldots, X_n$.

Conjugate distribution families

A suitably chosen prior distribution of θ may lead to a very tractable form of the posterior.

DEFINITION 10.4

A family of prior distributions π is **conjugate** to the model $f(\boldsymbol{x}|\theta)$ if the posterior distribution belongs to the same family.

Three classical examples of conjugate families are given below.

Gamma family is conjugate to the Poisson model

Let $(X_1, \ldots, X_n)$ be a sample from Poisson(θ) distribution with a Gamma(α, λ) prior distribution of θ.

Then

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \sim e^{-n\theta}\theta^{\sum x_i}. \tag{10.14}$$

Remark about dropping constant coefficients. In the end of (10.14), we dropped $(x_i!)$ and wrote that the result is "proportional" $(\sim)$ to $e^{-n\theta}\theta^{\sum x_i}$. Dropping terms that don't contain θ often simplifies the computation. The form of the posterior distribution can be obtained without the constant term, and if needed, we can eventually evaluate the normalizing constant in the end, making $\pi(\theta|\boldsymbol{x})$ a fine distribution with the total probability 1, for example, as we did in Example 4.1 on p. 77. In particular, the marginal distribution $m(\boldsymbol{x})$ can be

dropped because it is θ-free. But keep in mind that in this case we obtain the posterior distribution "up to a constant coefficient."

The Gamma prior distribution of θ has density

$$\pi(\theta) \sim \theta^{\alpha-1} e^{-\lambda\theta}.$$

As a function of θ, this prior density has the same form as the model $f(\boldsymbol{x}|\theta)$ – a power of θ multiplied by an exponential function. This is the general idea behind conjugate families.

Then, the posterior distribution of θ given $\boldsymbol{X} = \boldsymbol{x}$ is

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &\sim f(\boldsymbol{x}|\theta)\pi(\theta) \\ &\sim \left(e^{-n\theta}\theta^{\sum x_i}\right)\left(\theta^{\alpha-1}e^{-\lambda\theta}\right) \\ &\sim \theta^{\alpha+\sum x_i-1} e^{-(\lambda+n)\theta}. \end{aligned}$$

Comparing with the general form of a Gamma density (say, (4.7) on p. 85), we see that $\pi(\theta|\boldsymbol{x})$ is the Gamma distribution with new parameters,

$$\alpha_x = \alpha + \sum_{i=1}^{n} x_i \text{ and } \lambda_x = \lambda + n.$$

We can conclude that:

1. Gamma family of prior distributions is conjugate to Poisson models.
2. Having observed a Poisson sample $\boldsymbol{X} = \boldsymbol{x}$, we update the Gamma$(\alpha, \lambda)$ prior distribution of θ to the Gamma$(\alpha + \sum x_i, \lambda + n)$ posterior.

Gamma distribution family is rather rich; it has two parameters. There is often a good chance to find a member of this large family that suitably reflects our knowledge about θ.

Example 10.25 (NETWORK BLACKOUTS). The number of network blackouts each week has Poisson(θ) distribution. The weekly rate of blackouts θ is not known exactly, but according to the past experience with similar networks, it averages 4 blackouts with a standard deviation of 2.

There exists a Gamma distribution with the given mean $\mu = \alpha/\lambda = 4$ and standard deviation $\sigma = \sqrt{\alpha}/\lambda = 2$. Its parameters α and λ can be obtained by solving the system,

$$\left\{ \begin{array}{rcl} \alpha/\lambda & = & 4 \\ \sqrt{\alpha}/\lambda & = & 2 \end{array} \right. \Rightarrow \left\{ \begin{array}{rcccl} \alpha & = & (4/2)^2 & = & 4 \\ \lambda & = & 2^2/4 & = & 1 \end{array} \right.$$

Hence, we can assume the Gamma$(\alpha = 4,\ \lambda = 1)$ prior distribution θ. It is convenient to have a conjugate prior because the posterior will then belong to the Gamma family too.

Suppose there were $X_1 = 2$ blackouts this week. Given that, the posterior distribution of θ is Gamma with parameters

$$\alpha_x = \alpha + 2 = 6, \quad \lambda_x = \lambda + 1 = 2.$$

If no blackouts occur during the next week, the updated posterior parameters become

$$\alpha_x = \alpha + 2 + 0 = 6, \quad \lambda_x = \lambda + 2 = 3.$$

This posterior distribution has the average weekly rate of $6/3 = 2$ blackouts per week. Two weeks with very few blackouts reduced our estimate of the average rate from 4 to 2. $\Diamond$

Beta family is conjugate to the Binomial model

A sample from Binomial(k, θ) distribution (assume k is known) has the probability mass function

$$f(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} \binom{k}{x_i} \theta^{x_i}(1-\theta)^{k-x_i} \sim \theta^{\sum x_i}(1-\theta)^{nk-\sum x_i}.$$

Density of Beta(α, β) prior distribution has the same form, as a function of θ,

$$\pi(\theta) \sim \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } 0 < \theta < 1$$

(see Section 12.1.2 in the Appendix). Then, the posterior density of θ is

$$\pi(\theta \mid \boldsymbol{x}) \sim f(\boldsymbol{x} \mid \theta)\pi(\theta) \sim \theta^{\alpha+\sum_{i=1}^{n} x_i - 1}(1-\theta)^{\beta+nk-\sum_{i=1}^{n} x_i - 1},$$

and we recognize the Beta density with new parameters

$$\alpha_x = \alpha + \sum_{i=1}^{n} x_i \text{ and } \beta_x = \beta + nk - \sum_{i=1}^{n} x_i.$$

Hence,

1. Beta family of prior distributions is conjugate to the Binomial model.
2. Posterior parameters are $\alpha_x = \alpha + \sum x_i$ and $\beta_x = \beta + nk - \sum x_i$.

Normal family is conjugate to the Normal model

Consider now a sample from Normal distribution with an unknown mean θ and a known variance σ^2:

$$\begin{aligned} f(\boldsymbol{x} \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\} \sim \exp\left\{-\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2\sigma^2}\right\} \\ &\sim \exp\left\{\theta\frac{\sum x_i}{\sigma^2} - \theta^2\frac{n}{2\sigma^2}\right\} = \exp\left\{\left(\theta\bar{X} - \frac{\theta^2}{2}\right)\frac{n}{\sigma^2}\right\}. \end{aligned}$$

If the prior distribution of θ is also Normal, with prior mean μ and prior variance τ^2, then

$$\pi(\theta) \sim \exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\} \sim \exp\left\{\left(\theta\mu - \frac{\theta^2}{2}\right)\frac{1}{\tau^2}\right\},$$

and again, it has a similar form as $f(\boldsymbol{x}|\theta)$.

The posterior density of θ equals

$$\begin{aligned} \pi(\theta \mid \boldsymbol{x}) &\sim f(\boldsymbol{x} \mid \theta)\pi(\theta) \sim \exp\left\{\theta\left(\frac{n\bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}\right) - \frac{\theta^2}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\right\} \\ &\sim \exp\left\{-\frac{(\theta-\mu_x)^2}{2\tau_x^2}\right\}, \end{aligned}$$

where

$$\mu_x = \frac{n\bar{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} \text{ and } \tau_x^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}. \tag{10.15}$$

This posterior distribution is certainly Normal with parameters μ_x and τ_x.

We can conclude that:

1. Normal family of prior distributions is conjugate to the Normal model with unknown mean;

2. Posterior parameters are given by (10.15).

We see that the posterior mean μ_x is a weighted average of the prior mean μ and the sample mean $\bar{X}$. This is how the prior information and the observed data are combined in case of Normal distributions.

How will the posterior mean behave when it is computed from a large sample? As the sample size n increases, we get more information from the data, and as a result, the frequentist estimator will dominate. According to (10.15), the posterior mean converges to the sample mean $\bar{X}$ as $n \to \infty$.

Posterior mean will also converge to $\bar{X}$ when $\tau \to \infty$. Large τ means a lot of uncertainty in the prior distribution of θ; thus, naturally, we should rather use observed data as a more reliable source of information in this case.

On the other hand, large σ indicates a lot of uncertainty in the observed sample. If that is the case, the prior distribution is more reliable, and as we see in (10.15), $\mu_x \approx \mu$ for large σ.

Results of this section are summarized in Table 10.2. You will find more examples of conjugate prior distributions among the exercises.

Model $f(\boldsymbol{x}\|\theta)$	Prior $\pi(\theta)$	Posterior $\pi(\theta\|\boldsymbol{x})$
Poisson(θ)	Gamma$(\alpha,\ \lambda)$	Gamma$(\alpha + n\bar{X},\ \lambda + n)$
Binomial(k, θ)	Beta$(\alpha,\ \beta)$	Beta$(\alpha + n\bar{X},\ \beta + n(k - \bar{X}))$
Normal(θ, σ)	Normal(μ, τ)	Normal$\left(\dfrac{n\bar{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2},\ \dfrac{1}{\sqrt{n/\sigma^2 + 1/\tau^2}}\right)$

TABLE 10.2: *Three classical conjugate families.*

10.4.2 Bayesian estimation

We have already completed the most important step in Bayesian inference. We obtained the posterior distribution. All the knowledge about the unknown parameter is now included in the posterior, and that is what we'll use for further statistical analysis (Figure 10.11).

To estimate θ, we simply compute the **posterior mean**,

$$\hat{\theta}_{\mathrm{B}} = \mathbf{E}\left\{\theta|\boldsymbol{X} = \boldsymbol{x}\right\} = \begin{cases} \displaystyle\sum_{\theta} \theta\pi(\theta|\boldsymbol{x}) = \frac{\sum \theta f(\boldsymbol{x}|\theta)\pi(\theta)}{\sum f(\boldsymbol{x}|\theta)\pi(\theta)} & \text{if } \theta \text{ is discrete} \\ \displaystyle\int_{\theta} \theta\pi(\theta|\boldsymbol{x})d\theta = \frac{\int \theta f(\boldsymbol{x}|\theta)\pi(\theta)d\theta}{\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta} & \text{if } \theta \text{ is continuous} \end{cases}$$

The result is a conditional expectation of θ given data $\boldsymbol{X}$. In abstract terms, the **Bayes estimator** $\hat{\theta}_{\mathrm{B}}$ is what we "expect" θ to be, after we observed a sample.

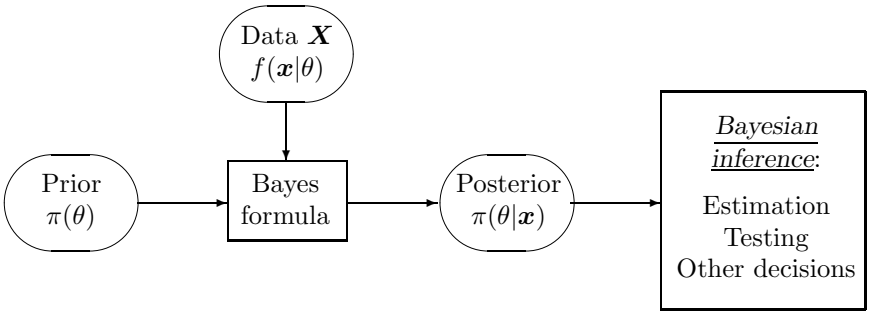

FIGURE 10.11: *Posterior distribution is the basis for Bayesian inference.*

How accurate is this estimator? Among all estimators, $\hat{\theta}_{\text{B}} = \mathbf{E}\{\theta|\boldsymbol{x}\}$ has the lowest squared-error **posterior risk**

$$\rho(\hat{\theta}) = \mathbf{E}\left\{(\hat{\theta} - \theta)^2 \mid \boldsymbol{X} = \boldsymbol{x}\right\}.$$

For the Bayes estimator $\hat{\theta}_{\text{B}} = \mathbf{E}\{\theta \mid \boldsymbol{x}\}$, posterior risk equals **posterior variance**,

$$\rho(\hat{\theta}) = \mathbf{E}\left\{(\mathbf{E}\{\theta|\boldsymbol{x}\} - \theta)^2 \mid \boldsymbol{x}\right\} = \mathbf{E}\left\{(\theta - \mathbf{E}\{\theta|\boldsymbol{x}\})^2 \mid \boldsymbol{x}\right\} = \text{Var}\{\theta|\boldsymbol{x}\},$$

which measures variability of θ around $\hat{\theta}_{\text{B}}$, according to the posterior distribution of θ.

Example 10.26 (Normal case). The Bayes estimator of the mean θ of Normal(θ, σ) distribution with a Normal(μ, τ) prior is

$$\hat{\theta}_{\text{B}} = \mu_x = \frac{n\bar{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2},$$

and its posterior risk is

$$\rho(\hat{\theta}_{\text{B}}) = \tau_x^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}$$

(Table 10.2). As we expect, this risk decreases to 0 as the sample size grows to infinity. ◊

Example 10.27 (Network blackouts, continued). After two weeks of data, the weekly rate of network blackouts, according to Example 10.25 on p. 343, has Gamma posterior distribution with parameters $\alpha_x = 6$ and $\lambda_x = 3$.

The Bayes estimator of the weekly rate θ is

$$\hat{\theta}_{\text{B}} = \mathbf{E}\{\theta|\boldsymbol{x}\} = \frac{\alpha_x}{\lambda_x} = 2 \text{ (blackouts per week)}$$

with a posterior risk

$$\rho(\hat{\theta}_{\text{B}}) = \text{Var}\{\theta|\boldsymbol{x}\} = \frac{\alpha_x}{\lambda_x^2} = \frac{2}{3}.$$

◊

Although conjugate priors simplify our statistics, Bayesian inference can certainly be done for other priors too.

Example 10.28 (QUALITY INSPECTION, CONTINUED). In Example 10.24 on p. 341, we computed posterior distribution of the proportion of defective parts θ. This was a discrete distribution,

$$\pi(0.05 \mid \boldsymbol{x}) = 0.2387; \quad \pi(0.10 \mid \boldsymbol{x}) = 0.7613.$$

Now, the Bayes estimator of θ is

$$\hat{\theta}_{\mathrm{B}} = \sum_{\theta} \theta \pi(\theta \mid \boldsymbol{x}) = (0.05)(0.2387) + (0.10)(0.7613) = 0.0881.$$

It does not agree with the manufacturer (who claims $\theta = 0.05$) or with the quality inspector (who feels that $\theta = 0.10$) but its value is much closer to the inspector's estimate.

The posterior risk of $\hat{\theta}_{\mathrm{B}}$ is

$$\begin{aligned} \mathrm{Var}\{\theta|\boldsymbol{x}\} &= \mathbf{E}\left\{\theta^2|\boldsymbol{x}\right\} - \mathbf{E}^2\{\theta|\boldsymbol{x}\} \\ &= (0.05)^2(0.2387) + (0.10)^2(0.7613) - (0.0881)^2 = 0.0004, \end{aligned}$$

which means a rather low posterior standard deviation of 0.02. ◊

10.4.3 Bayesian credible sets

Confidence intervals have a totally different meaning in Bayesian analysis. Having a posterior distribution of θ, we no longer have to explain the confidence level $(1-\alpha)$ in terms of a long run of samples. Instead, we can give an interval $[a, b]$ or a set C that has a posterior probability $(1-\alpha)$ and state that *the parameter θ belongs to this set with probability* $(1-\alpha)$. Such a statement was impossible before we considered prior and posterior distributions. This set is called a $(1-\alpha)100\%$ *credible set*.

DEFINITION 10.5

Set C is a $(1-\alpha)100\%$ **credible set** for the parameter θ if the posterior probability for θ to belong to C equals $(1-\alpha)$. That is,

$$\boldsymbol{P}\{\theta \in C \mid \boldsymbol{X} = \boldsymbol{x}\} = \int_C \pi(\theta|\boldsymbol{x}) d\theta = 1 - \alpha.$$

Such a set is not unique. Recall that for two-sided, left-tail, and right-tail hypothesis testing, we took different portions of the area under the Normal curve, all equal $(1-\alpha)$.

Minimizing the length of set C among all $(1-\alpha)100\%$ credible sets, we just have to include all the points θ with a high posterior density $\pi(\theta|\boldsymbol{x})$,

$$C = \{\theta : \ \pi(\theta|\boldsymbol{x}) \geq c\}$$

(see Figure 10.12). Such a set is called the **highest posterior density credible set**, or just the **HPD set**.

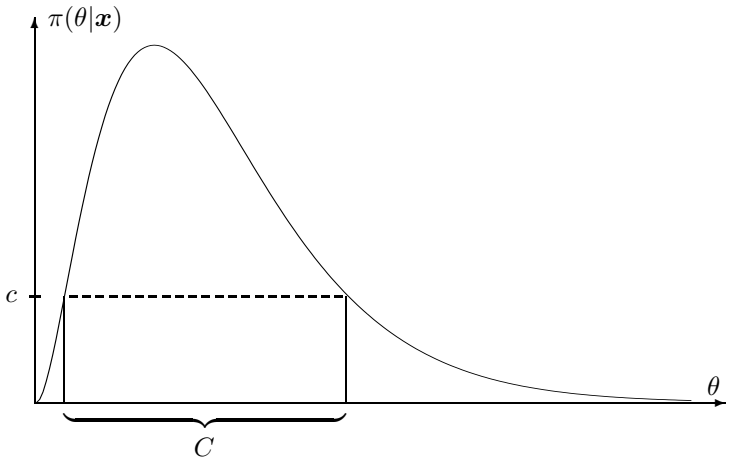

FIGURE 10.12: *The $(1-\alpha)100\%$ highest posterior density credible set.*

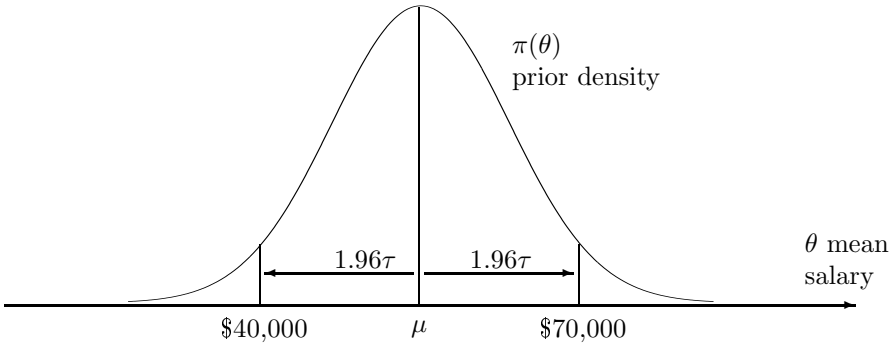

FIGURE 10.13: *Normal prior distribution and the 95% HPD credible set for the mean starting salary of Computer Science graduates (Example 10.29).*

For the Normal(μ_x, τ_x) posterior distribution of θ, the $(1-\alpha)100\%$ HPD set is

$$\mu_x \pm z_{\alpha/2}\tau_x = [\mu_x - z_{\alpha/2}\tau_x, \mu_x + z_{\alpha/2}\tau_x].$$

Example 10.29 (SALARIES, CONTINUED). In Example 10.23 on p. 339, we "decided" that the most likely range for the mean starting salary θ of Computer Science graduates is between \$40,000 and \$70,000. Expressing this in a form of a prior distribution, we let the prior mean be $\mu = (40,000 + 70,000)/2 = 55,000$. Further, if we feel that the range $[40,000; 70,000]$ is 95% likely, and we accept a Normal prior distribution for θ, then this range should be equal

$$[40,000; 70,000] = \mu \pm z_{0.025/2}\tau = \mu \pm 1.96\tau,$$

where τ is the prior standard deviation (Figure 10.13). We can now evaluate the prior standard deviation parameter τ from this information,

$$\tau = \frac{70,000 - 40,000}{2(1.96)} = 7,653.$$

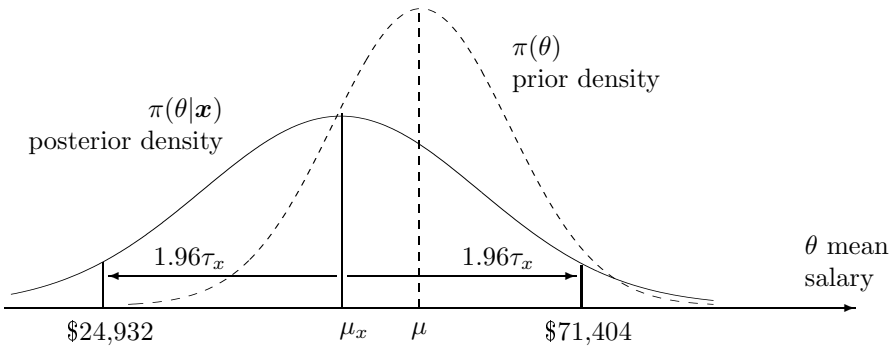

FIGURE 10.14: *Normal prior and posterior distributions for the mean starting salary (Example 10.29).*

This is the advantage of using a rich (two-parameter) family of prior distributions: we are likely to find a member of this family that reflects our prior beliefs adequately.

Then, *prior to collecting any data*, the 95% HPD credible set of the mean starting salary θ is

$$\mu \pm z_{0.025}\tau = \underline{[40{,}000;\ 70{,}000]}.$$

Suppose a random sample of 100 graduates has the mean starting salary $\bar{X} = 48,000$ with a sample standard deviation $s = 12,000$. From Table 10.2, we determine the posterior mean and standard deviation,

$$\begin{aligned}
\mu_x &= \frac{n\bar{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} = \frac{(100)(48,000)/(12,000)^2 + (55,000)/(7,653)^2}{100/(12,000)^2 + 1/(7653)^2} \\
&= 48,168; \\
\tau_x &= \frac{1}{\sqrt{n/\sigma^2 + 1/\tau^2}} = \frac{1}{\sqrt{100/(12,000)^2 + 1/(7653)^2}} = 11,855.
\end{aligned}$$

We used the sample standard deviation s in place of the population standard deviation σ assuming that a sample of size 100 estimates the latter rather accurately. Alternatively, we could put a prior distribution on unknown σ too and estimate it by Bayesian methods. Since the observed sample mean is smaller than our prior mean, the resulting posterior distribution is shifted to the left of the prior (Figure 10.14).

Conclusion. After seeing the data, the Bayes estimator for the mean starting salary of CS graduates is

$$\hat{\theta}_{\text{B}} = \mu_x = 48,168 \text{ dollars},$$

and the 95% HPD credible set for this mean salary is

$$\mu_x \pm z_{0.025}\tau_x = 48,168 \pm (1.96)(11,855) = 48,168 \pm 23,236 = \underline{[24{,}932;\ 71{,}404]}$$

Lower observed salaries than the ones predicted a priori extended the lower end of our credible interval. ◊

Example 10.30 (TELEPHONE COMPANY). A new telephone company predicts to handle an average of 1000 calls per hour. During 10 randomly selected hours of operation, it handled 7265 calls.

How should it update the initial estimate of the frequency of telephone calls? Construct a 95% HPD credible set. Telephone calls are placed according to a Poisson process. The hourly rate of calls has an Exponential prior distribution.

Solution. We need a Bayesian estimator of the frequency θ of telephone calls. The number of calls during 1 hour has Poisson(θ) distribution, where θ is unknown, with

$$\text{Exponential}(\lambda) = \text{Gamma}(1, \lambda)$$

prior distribution that has an expectation of

$$\mathbf{E}(\theta) = \frac{1}{\lambda} = 1000 \text{ calls.}$$

Hence, $\lambda = 0.001$. We observe a sample of size $n = 10$, totaling

$$\sum_{i=1}^{n} X_i = n\bar{X} = 7265 \text{ calls.}$$

As we know (see Table 10.2 on p. 345), the posterior distribution in this case is Gamma(α_x, λ_x) with

$$\begin{aligned} \alpha_x &= \alpha + n\bar{X} = 7266, \\ \lambda_x &= \lambda + n = 10.001. \end{aligned}$$

This distribution has mean

$$\mu_x = \alpha_x/\lambda_x = 726.53$$

and standard deviation

$$\tau_x = \alpha_x/\lambda_x^2 = 72.65.$$

The Bayes estimator of θ is

$$\mathbf{E}(\theta|\boldsymbol{X}) = \mu_x = \underline{726.53 \text{ calls per hour}}.$$

It almost coincides with the sample mean $\bar{X}$ showing the sample was informative enough to dominate over the prior information.

For the credible set, we notice that α_x is sufficiently large to make the Gamma posterior distribution approximately equal the Normal distribution with parameters μ_x and τ_x. The 95% HPD credible set is then

$$\mu_x \pm z_{0.05/2}\tau_x = 726.53 \pm (1.96)(72.65) = 726.53 \pm 142.39 = \underline{[584.14,\ 868.92]}$$

◊

10.4.4 Bayesian hypothesis testing

Bayesian hypothesis testing is very easy to interpret. We can compute prior and posterior probabilities for the hypothesis H_0 and alternative H_A to be true and decide from there which one to accept or to reject.

Computing such probabilities was not possible without prior and posterior distributions of the parameter θ. In non-Bayesian statistics, θ was not random, thus H_0 and H_A were either true (with probability 1) or false (with probability 1).

For Bayesian tests, in order for H_0 to have a meaningful, non-zero probability, it often represents a set of parameter values instead of just one θ_0, and we are testing

$$H_0: \ \theta \in \Theta_0 \ \text{ vs } \ H_A: \ \theta \in \Theta_1.$$

This actually makes sense because exact equality $\theta = \theta_0$ is unlikely to hold anyway, and in practice it is understood as $\theta \approx \theta_0$.

Comparing posterior probabilities of H_0 and H_A,

$$\boldsymbol{P}\{\Theta_0 \mid \boldsymbol{X} = \boldsymbol{x}\} \ \text{ and } \ \boldsymbol{P}\{\Theta_1 \mid \boldsymbol{X} = \boldsymbol{x}\},$$

we decide whether $\boldsymbol{P}\{\Theta_1 \mid \boldsymbol{X} = \boldsymbol{x}\}$ is large enough to present significant evidence and to reject the null hypothesis. One can again compare it with $(1-\alpha)$ such as 0.90, 0.95, 0.99, or state the result in terms of likelihood, "the null hypothesis is this much likely to be true."

Example 10.31 (Telephone company, continued). Let us test whether the telephone company in Example 10.30 can actually face a call rate of 1000 calls *or more* per hour. We are testing

$$H_0: \ \theta \geq 1000 \ \text{ vs } \ H_A: \ \theta < 1000,$$

where θ is the hourly rate of telephone calls.

According to the Gamma(α_x, λ_x) posterior distribution of θ and its Normal $(\mu_x = 726.53, \tau_x = 72.65)$ approximation,

$$\boldsymbol{P}\{H_0 \mid \boldsymbol{X} = \boldsymbol{x}\} = \boldsymbol{P}\left\{\frac{\theta - \mu_x}{\tau_x} \geq \frac{1000 - \mu_x}{\tau_x}\right\} = 1 - \Phi(3.76) = 0.0001.$$

By the complement rule, $\boldsymbol{P}\{H_A \mid \boldsymbol{X} = \boldsymbol{x}\} = 0.9999$, and this presents sufficient evidence against H_0.

We conclude that it's extremely unlikely for this company to face a frequency of 1000+ calls per hour. ◊

Loss and risk

Often one can anticipate the consequences of Type I and Type II errors in hypothesis testing and assign a **loss** $L(\theta, a)$ associated with each possible error. Here θ is the parameter, and a is our action, the decision on whether we accept or reject the null hypothesis.

Each decision then has its **posterior risk** $\rho(a)$, defined as the expected loss computed under the posterior distribution. The action with the lower posterior risk is our **Bayes action**.

Suppose that the Type I error causes the loss

$$w_0 = Loss(\text{Type I error}) = L(\Theta_0, \text{reject } H_0),$$

and the Type II error causes the loss

$$w_1 = Loss(\text{Type II error}) = L(\Theta_1, \text{accept } H_0).$$

Posterior risks of each possible action are then computed as

$$\begin{aligned} \rho(\text{reject } H_0) &= w_0 \pi(\Theta_0 \mid \boldsymbol{x}), \\ \rho(\text{accept } H_0) &= w_1 \pi(\Theta_1 \mid \boldsymbol{x}). \end{aligned}$$

Now we can determine the Bayes action. If $w_0\pi(\Theta_0 \mid \boldsymbol{x}) \le w_1\pi(\Theta_1 \mid \boldsymbol{x})$, the Bayes action is to accept H_0. If $w_0\pi(\Theta_0 \mid \boldsymbol{x}) \ge w_1\pi(\Theta_1 \mid \boldsymbol{x})$, the Bayes action is to reject H_0.

Example 10.32 (Quality inspection, continued). In Example 10.24 on p. 341, we are testing

$$H_0 : \ \theta = 0.05 \ \text{ vs } \ H_A : \ \theta = 0.10$$

for the proportion θ of defective parts. Suppose that the Type I error is three times as costly here as the Type II error. What is the Bayes action in this case?

Example 10.28 gives posterior probabilities

$$\pi(\Theta_0 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.2387 \ \text{ and } \ \pi(\Theta_1 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.7613.$$

Since $w_0 = 3w_1$, the posterior risks are

$$\begin{aligned} \rho(\text{reject } H_0) &= w_0 \pi(\Theta_0 \mid \boldsymbol{x}) = 3w_1(0.2387) = 0.7161 w_1, \\ \rho(\text{accept } H_0) &= w_1 \pi(\Theta_1 \mid \boldsymbol{x}) = 0.7613 w_1. \end{aligned}$$

Thus, rejecting H_0 has a lower posterior risk, and therefore, it is the Bayes action. Reject H_0. ◊

Summary and conclusions

A number of popular methods of Statistical Inference are presented in this chapter.

Chi-square tests represent a general technique based on counts. Comparing the observed counts with the counts expected under the null hypothesis through the chi-square statistic, one can test for the goodness of fit and for the independence of two factors. Contingency tables are widely used for the detection of significant relations between categorical variables.

Nonparametric statistical methods are not based on any particular distribution of data. So, they are often used when the distribution is unknown or complicated. They are also very handy when the sample may contain some outliers, and even when the data are not numerical.

A sign test and Wilcoxon signed rank test for one sample and a two-sample Mann-Whitney-Wilcoxon rank sum test are introduced in this chapter for testing and comparing distributions and their medians. Combinatorics helps to find the exact null distribution of each considered test statistic. For large samples, this distribution is approximately Normal.

Bootstrap is a popular modern resampling technique that is widely used nowadays for studying properties and evaluating performance of various statistics. A simple idea behind the bootstrap allows us to analyze complicated statistics using only the power of our computer. This chapter showed the most basic applications of bootstrap to the estimation of standard errors and biases of parameter estimates and construction of parametric and nonparametric confidence intervals.

Bayesian inference combines the information contained in the data and in the prior distribution of the unknown parameter. It is based on the posterior distribution, which is the conditional distribution of the unknown parameter, given the data.

The most commonly used Bayesian parameter estimator is the posterior mean. It minimizes the squared-error posterior risk among all the estimates.

Bayesian $(1-\alpha)100\%$ *credible sets* also contain the parameter θ with probability $(1-\alpha)$, but this time the probability refers to the distribution of θ. Explaining a $(1-\alpha)100\%$ credible set, we can say that given the observed data, θ belongs to the obtained set with probability $(1-\alpha)$.

For *Bayesian hypothesis testing*, we compute posterior probabilities of H_0 and H_A and decide if the former is sufficiently smaller than the latter to suggest rejection of H_0. We can also take a Bayesian action that minimizes the posterior risk.

Exercises

10.1. Does the number of unsolicited (spam) emails follow a Poisson distribution? Here is the record of the number of spam emails received during 60 consecutive days.

12	6	4	0	13	5	1	3	10	1	29	12	4	4	22
2	2	27	7	27	9	34	10	10	2	28	7	0	9	4
32	4	5	9	1	13	10	20	5	5	0	6	9	20	28
22	10	8	11	15	1	14	0	9	9	1	9	0	7	13

Choose suitable bins and conduct a goodness-of-fit test at the 1% level of significance.

10.2. Applying the theory of M/M/1 queuing systems, we assume that the service times follow Exponential distribution. The following service times, in minutes, have been observed during 24 hours of operation:

10.5	1.2	6.3	3.7	0.9	7.1	3.3	4.0	1.7	11.6	5.1	2.8	4.8	2.0	8.0	4.6
3.1	10.2	5.9	12.6	4.5	8.8	7.2	7.5	4.3	8.0	0.2	4.4	3.5	9.6	5.5	0.3
2.7	4.9	6.8	8.6	0.8	2.2	2.1	0.5	2.3	2.9	11.7	0.6	6.9	11.4	3.8	3.2
2.6	1.9	1.0	4.1	2.4	13.6	15.2	6.4	5.3	5.4	1.4	5.0	3.9	1.8	4.7	0.7

Is the assumption of Exponentiality supported by these data?

10.3. The following sample is collected to verify the accuracy of a new random number generator (it is already ordered for your convenience).

-2.434	-2.336	-2.192	-2.010	-1.967	-1.707	-1.678	-1.563	-1.476	-1.388
-1.331	-1.269	-1.229	-1.227	-1.174	-1.136	-1.127	-1.124	-1.120	-1.073
-1.052	-1.051	-1.032	-0.938	-0.884	-0.847	-0.846	-0.716	-0.644	-0.625
-0.588	-0.584	-0.496	-0.489	-0.473	-0.453	-0.427	-0.395	-0.386	-0.386
-0.373	-0.344	-0.280	-0.246	-0.239	-0.211	-0.188	-0.155	-0.149	-0.112
-0.103	-0.101	-0.033	-0.011	0.033	0.110	0.139	0.143	0.218	0.218
0.251	0.261	0.308	0.343	0.357	0.463	0.477	0.482	0.489	0.545
0.590	0.638	0.652	0.656	0.673	0.772	0.775	0.776	0.787	0.969
0.978	1.005	1.013	1.039	1.072	1.168	1.185	1.263	1.269	1.297
1.360	1.370	1.681	1.721	1.735	1.779	1.792	1.881	1.903	2.009

(a) Apply the χ^2 goodness-of-fit test to check if this sample comes from the Standard Normal distribution.

(b) Test if this sample comes from the Uniform(-3,3) distribution.

(c) Is it theoretically possible to accept both null hypotheses in (a) and (b) although they are contradicting to each other? Why does it make sense?

10.4. In Example 10.3 on p. 309, we tested whether the number of concurrent users is approximately Normal. How does the result of the chi-square test depend on our choice of bins? For the same data, test the assumption of a Normal distribution using a different set of bins $B_1, \ldots, B_N$.

10.5. Show that the sample size is too small in Example 10.9 on p. 316 to conduct a χ^2 goodness-of-fit test of Normal distribution that involves estimation of its both parameters.

10.6. Two computer makers, A and B, compete for a certain market. Their users rank the quality of computers on a 4-point scale as "Not satisfied", "Satisfied", "Good quality", and "Excellent quality, will recommend to others". The following counts were observed,

Computer maker	"Not satisfied"	"Satisfied"	"Good quality"	"Excellent quality"
A	20	40	70	20
B	10	30	40	20

Is there a significant difference in customer satisfaction of the computers produced by A and by B?

10.7. An AP test has been given in two schools. In the first school, 162 girls and 567 boys passed it whereas 69 girls and 378 boys failed. In the second school, 462 girls and 57 boys passed the test whereas 693 girls and 132 boys failed it.

(a) In the first school, are the results significantly different for girls and boys?

(b) In the second school, are the results significantly different for girls and boys?

(c) In both schools together, are the results significantly different for girls and boys?

For each school, construct a contingency table and apply the chi-square test.

Remark: This data set is an example of a strange phenomenon known as **Simpson's paradox**. Look, the girls performed better than the boys in *each* school; however, in both schools together, the boys did better!!!

Check for yourself... In the first school, 70% of girls and only 60% of boys passed the test. In the second school, 40% of girls and only 30% of boys passed. But in both schools together, 55% of boys and only 45% of girls passed the test. Wow!

10.8. A computer manager decides to install the new antivirus software on all the company's computers. Three competing antivirus solutions (X, Y, and Z) are offered to her for a free 30-day trial. She installs each solution on 50 computers and records infections during the following 30 days. Results of her study are in the table.

Antivirus software	X	Y	Z
Computers not infected	36	28	32
Computers infected once	12	16	14
Computers infected more than once	2	6	4

Does the computer manager have significant evidence that the three antivirus solutions are *not* of the same quality?

10.9. The Probability and Statistics course has three sections - S01, S02, and S03. Among 120 students in section S01, 40 got an A in the course, 50 got a B, 20 got a C, 2 got a D, and 8 got an F. Among 100 students in section S02, 20 got an A, 40 got a B, 25 got a C, 5 got a D, and 10 got an F. Finally, among 60 students in section S03, 20 got an A, 20 got a B, 15 got a C, 2 got a D, and 3 got an F. Do the three sections differ in their students' performance?

10.10. Among 25 jobs sent to the printer at random times, 6 jobs were printed in less than 20 seconds each, and 19 jobs took more than 20 seconds each. Is there evidence that the median response time for this printer exceeds 20 sec? Apply the sign test.

10.11. At a computer factory, the median of certain measurements of some computer parts should equal m. If it is found to be less than m or greater than m, the process is recalibrated. Every day, a quality control technician takes measurements from a sample of 20 parts. According to a 5% level sign test, how many measurements on either side of m justify recalibration? (In other words, what is the rejection region?)

10.12. When a computer chip is manufactured, its certain crucial layer should have the median thickness of 45 nm (nanometers; one nanometer is one billionth of a metre). Measurements are made on a sample of 60 produced chips, and the measured thickness is recorded as

34.9	35.9	38.9	39.4	39.9	41.3	41.5	41.7	42.0	42.1	42.5	43.5	43.7	43.9	44.2
44.4	44.6	45.3	45.7	45.9	46.0	46.2	46.4	46.6	46.8	47.2	47.6	47.7	47.8	48.8
49.1	49.2	49.4	49.5	49.8	49.9	50.0	50.2	50.5	50.7	50.9	51.0	51.3	51.4	51.5
51.6	51.8	52.0	52.5	52.6	52.8	52.9	53.1	53.7	53.8	54.3	56.8	57.1	57.8	58.9

(This data set is already ordered, for your convenience.)

Will a 1% level sign test conclude that the median thickness slipped from 45 nm?

10.13. Refer to Exercise 10.12. It is also important to verify that the first quartile Q_1 of the layer thickness does not exceed 43 nm. Apply the idea of the sign test to the first quartile instead of the median and test

$$H_0 : Q_1 = 43 \quad \text{vs} \quad H_A : Q_1 > 43.$$

The test statistic will be the number of measurements that exceed 43 nm. Find the null distribution of this test statistic, compute the P-value of the test, and state the conclusion about the first quartile.

10.14. The median of certain measurements on the produced computer parts should never exceed 5 inches. To verify that the process is conforming, engineers decided to measure 100 parts, one by one. To everyone's surprise, after 63 of the first 75 measurements exceeded 5 in, one engineer suggests to halt measurements and fix the process. She claims that whatever the remaining 25 measurements are, the median will be inevitably found significantly greater than 5 inches after all 100 measurements are made.

Do you agree with this statement? Why or why not? Certainly, if the remaining 25 measurements make no impact on the test then it should be a good resource-saving decision to stop testing early and fix the process.

10.15. Professor has stated that the median score on the last test was 84. Eric asked 12 of his classmates and recorded their scores as

$$76, 96, 74, 88, 79, 95, 75, 82, 90, 60, 77, 56.$$

Assuming that he picked the classmates at random, can he treat these data as evidence that the class median was less than 84? Can he get stronger evidence by using the sign test or the Wilcoxon signed rank test?

10.16. The starting salaries of eleven software developers are

$$47, 52, 68, 72, 55, 44, 58, 63, 54, 59, 77 \text{ thousand of dollars.}$$

Does the 5% level Wilcoxon signed rank test provide significant evidence that the median starting salary of software developers is above \$50,000?

10.17. Refer to Exercise 10.12. Does the Wilcoxon signed rank test confirm that the median thickness no longer equals 45 nm?

10.18. Refer to Exercise 10.2. Do these data provide significant evidence that the median service time is less than 5 min 40 sec? Conduct the Wilcoxon signed rank test at the 5% level of significance. What assumption of this test may not be fully satisfied by these data?

10.19. Use the recursive formula (10.6) to calculate the null distribution of Wilcoxon test statistic W for sample sizes $n = 2$, $n = 3$, and $n = 4$.

10.20. Apply the Mann-Whitney-Wilcoxon test to the quiz grades in Exercise 9.23 on p. 304 to see if Anthony's median grade is significantly higher than Eric's. What is the P-value?

10.21. Two internet service providers claim that they offer the fastest internet in the area. A local company requires the download speed of at least 20 Megabytes per second (Mbps), for its normal operation. It decides to conduct a fair contest by sending 10 packets of equal size through each network and recording their download speed.

For the 1st internet service provider, the download speed is recorded as 26.7, 19.0, 26.5, 29.1, 26.2, 27.6, 26.8, 24.2, 25.7, 23.0 Mbps. For the 2nd provider, the download speed is recorded as 19.3, 22.1, 23.4, 24.8, 25.9, 22.2, 18.3, 20.1, 19.2, 27.9 Mbps.

(a) According to the sign test, is there significant evidence that the median download speed for the 1st provider is at least 20 Mbps? What about the 2nd provider? Calculate each P-value.

(b) Repeat (a) using the Wilcoxon signed rank test. Do the P-values show that this test is more sensitive than the sign test?

(c) At the 1% level, is there significant evidence that the median download speed for the 1st provider exceeds the median download speed for the 2nd provider? Use the suitable test.

10.22. Fifteen email attachments were classified as benign and malicious. Seven benign attachments were 0.4, 2.1, 3.6, 0.6, 0.8, 2.4, and 4.0 Mbytes in size. Eight malicious attachments had sizes 1.2, 0.2, 0.3, 3.3, 2.0, 0.9, 1.1, and 1.5 Mbytes. Does the Mann-Whitney-Wilcoxon test detect a significant difference in the distribution of sizes of benign and malicious attachments? (If so, the size could help classify email attachments and warn about possible malicious codes.)

10.23. During freshman year, Eric's textbooks cost \$89, \$99, \$119, \$139, \$189, \$199, and \$229. During his senior year, he had to pay \$109, \$159, \$179, \$209, \$219, \$259, \$279, \$299, and \$309 for his textbooks. Is this significant evidence that the median cost of textbooks is rising, according to the Mann-Whitney-Wilcoxon test?

10.24. Two teams, six students each, competed at a programming contest. The judges gave the overall 1st, 3rd, 6th, 7th, 9th, and 10th places to members of Team A. Now the captain of Team A claims the overall victory over Team B, according to a one-sided Mann-Whitney-Wilcoxon test? Do you concur with his conclusion? What hypothesis are you testing?

10.25. Refer to Exercise 10.2. After the first 32 service times were recorded (the first two rows of data), the server was substantially modified. Conduct the Mann-Whitney-Wilcoxon test at the 10% level to see if this modification led to a reduction of the median service time.

10.26. On five days of the week, Anthony spends 2, 2, 3, 3, and 5 hours doing his homework.

(a) List all possible bootstrap samples and find the probability of each of them.

(b) Use your list to find the bootstrap distribution of the sample median.

(c) Use this bootstrap distribution to estimate the standard error and the bias of a sample median.

10.27. In Exercise 10.16, we tested the median starting salary of software developers. We can actually estimate this starting salary by the sample median, which for these data equals $\hat{M} = \$58,000$.

(a) How many different bootstrap samples can be generated? Find the number of all possible ordered and unordered samples.

(b) Find the bootstrap distribution of the sample median. Do not list all the bootstrap samples!

(c) Use this distribution to estimate the standard error of $\hat{M}$.

(d) Construct an 88% bootstrap confidence interval for the population median salary M.

(e) Use the bootstrap distribution to estimate the probability that 11 randomly selected software developers have their median starting salary above $50,000.

Exercises 10.28–10.30 require the use of a computer.

10.28. Refer to Exercise 10.15. Eric estimates the class median score by the sample median, which is any number between 77 and 79 for these data (Eric takes the middle value 78). Generate 10,000 bootstrap samples and use them for the following inference.

(a) Estimate the standard error of the sample median used by Eric.

(b) Is this sample median a biased estimator for the population median? Estimate its bias.

(c) Construct a 95% bootstrap confidence interval for the population median. Can it be used to test whether the class median score is 84?

10.29. Refer to Exercise 10.13 where the first quartile of layer thickness is being tested. This quartile is estimated by the first sample quartile $\hat{Q}_1$ (in case of a whole interval of sample quartiles, its middle is taken). Use the bootstrap method to estimate the standard error and the bias of this estimator.

10.30. Is there a correlation between the Eric's and Anthony's grades in Exercise 9.23 on p. 304? Construct an 80% nonparametric bootstrap confidence interval for the correlation coefficient between their quiz grades (a small sample size does not allow to assume the Normal distribution of the sample correlation coefficient).

10.31. A new section of a highway is opened, and $X = 4$ accidents occurred there during one month. The number of accidents has Poisson(θ) distribution, where θ is the expected number of accidents during one month. Experience from the other sections of this highway suggests that the prior distribution of θ is Gamma(5,1). Find the Bayes estimator of θ under the squared-error loss and find its posterior risk.

10.32. The data set consists of a sample $\boldsymbol{X} = (2, 3, 5, 8, 2)$ from the Geometric distribution with unknown parameter θ.

(a) Show that the Beta family of prior distributions is conjugate.

(b) Taking the Beta(3,3) distribution as the prior, compute the Bayes estimator of θ.

10.33. Service times of a queuing system, in minutes, follow Exponential distribution with an unknown parameter θ. The prior distribution of θ is Gamma(3,1). The service times of five random jobs are 4 min, 3 min, 2 min, 5 min, and 5 min.

(a) Show that the Gamma family of prior distributions is conjugate.

(b) Compute the Bayes estimator of parameter θ and its posterior risk.

(c) Compute the posterior probability that $\theta \geq 0.5$, i.e., at least one job every two minutes can be served.

(d) If the Type I error and the Type II error cause approximately the same loss, test the hypothesis $H_0 : \theta \geq 0.5$ versus $H_A : \theta < 0.5$.

10.34. An internet service provider studies the distribution of the number of concurrent users of the network. This number has Normal distribution with mean θ and standard deviation 4,000 people. The prior distribution of θ is Normal with mean 14,000 and standard deviation 2,000.

The data on the number of concurrent users are collected; see Exercise 8.2 on p. 234.

(a) Give the Bayes estimator for the mean number of concurrent users θ.

(b) Construct the highest posterior density 90% credible set for θ and interpret it.

(c) Is there significant evidence that the mean number of concurrent users exceeds 16,000?

10.35. Continue Exercise 10.34. Another statistician conducts a non-Bayesian analysis of the data in Exercise 8.2 on p. 234 about concurrent users.

(a) Give the non-Bayesian estimator for the mean number of concurrent users θ.

(b) Construct a 90% confidence interval for θ and interpret it.

(c) Is there significant evidence that the mean number of concurrent users exceeds 16,000?

(d) How do your results differ from the previous exercise?

10.36. In Example 9.13 on p. 251, we constructed a confidence interval for the population mean μ based on the observed Normally distributed measurements. Suppose that prior to the experiment we thought this mean should be between 5.0 and 6.0 with probability 0.95.

(a) Find a conjugate prior distribution that fully reflects your prior beliefs.

(b) Derive the posterior distribution and find the Bayes estimator of μ. Compute its posterior risk.

(c) Compute a 95% HPD credible set for μ. Is it different from the 95% confidence interval? What causes the differences?

10.37. If ten coin tosses result in ten straight heads, can this coin still be fair and unbiased?

By looking at a coin, you believe that it is fair (a 50-50 chance of each side) with probability 0.99. This is your prior probability. With probability 0.01, you allow the coin to be biased, one way or another, so its probability of heads is Uniformly distributed between 0 and 1. Then you toss the coin ten times, and each time it turns up heads. Compute the posterior probability that it is a fair coin.

10.38. Observed is a sample from Uniform$(0, \theta)$ distribution.

(a) Find a conjugate family of prior distributions (you can find it in our inventory in Section 12.1).

(b) Assuming a prior distribution from this family, derive a form of the Bayes estimator and its posterior risk.

10.39. A sample $X_1, \ldots, X_n$ is observed from $Beta\,(\theta, 1)$ distribution. Derive the general form of a Bayes estimator of θ under the Gamma(α, λ) prior distribution.

10.40. Anton played five chess games and won all of them. Let us estimate θ, his probability of winning the next game. Suppose that parameter θ has Beta(4,1) prior distribution and that the game results are independent of each other.

(a) Compute the Bayes estimator of θ and its posterior risk.

(b) Construct a 90% HPD credible set for θ.

(c) Is there significant evidence that Anton's winning probability is more than 0.7? To answer this question, find the posterior probabilities of $\theta \le 0.7$ and $\theta > 0.7$ and use them to test $H_0 : \theta \le 0.7$ vs $H_A : \theta > 0.7$.

Notice that the standard frequentist estimator of θ is the sample proportion $\hat{p} = 1$, which is rather unrealistic because it gives Anton no chance to lose!