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

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

Compute the Chi-square statistic

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

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. $\Diamond$

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 $(X_1, \ldots, X_n)$ 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 $\widehat{\theta}$, to ensure $\widehat{\theta} \to \theta$ 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 $\widehat{\theta}$ and consider the distribution $F(\widehat{\theta}) \in \mathfrak{F}$;
- partition the support of $F(\widehat{\theta})$ into N bins $B_1, \ldots, B_N$, preferably with $N \in [5, 8]$;
- compute probabilities $p_k = \boldsymbol{P}\{X \in B_k\}$ for $k = 1, \ldots, N$ using $\widehat{\theta}$ 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.

<u>Solution</u>. 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

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

(see Example 9.7 on p. 249).

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

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

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. 240 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

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

Maximum likelihood estimates of μ and σ are (see Exercise 9.3 on p. 309)

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

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 $p_1, \ldots, p_6$. 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. ◊

R and MATLAB notes

For Chi-square tests, R has a command `chisq.test`. Here is how it solves Example 10.2.

— R ————————

```
counts = c(44,52,36,20,12,6)                    # Observed counts
prob = c( dpois(0:4,1.55), 1-ppois(4,1.55) )  # Poisson probabilities
                # of our chosen bins. These are P(0),...,P(4) and 1-F(4)
chisq.test(counts,p=prob)                       # Test for goodness of fit
```

As you notice, we test the observed counts against the Poisson distribution with parameter 1.55, the estimate derived in Example 10.2 by the method of maximum likelihood. It can also be computed in R using `fitdistr(x,'Poisson')`, the tool discussed in the end of Section 9.1.

A similar MATLAB command is `chi2gof`, a fancy abbreviation of a chi-square test for the goodness of fit. Let's apply it to Example 10.3.

— MATLAB —

```
X = [17.2 22.1 18.5 17.2 18.6 14.8 21.7 15.8 16.3 22.8 24.1 13.3...
16.2 17.5 19.0 23.9 14.8 22.2 21.7 20.7 13.5 15.8 13.1 16.1 21.9...
23.9 19.3 12.0 19.9 19.4 15.4 16.7 19.5 16.2 16.9 17.1 20.2 13.4...
19.8 17.7 19.7 18.7 17.6 15.9 15.2 17.1 15.0 18.8 21.6 11.9];
[ decision, pvalue, stat ] = chi2gof(X, 'CDF', fitdist(X,'Normal'))
```

This gives us three pieces of information. New variables `pvalue` and `decision` are a P-value of the test and a decision based on it (1 if H_0 is rejected in favor of H_A meaning an evidence against Normal distribution, 0 otherwise). Also, the new object `stat` contains all the steps of this test - edges of chosen bins, observed and expected counts, and finally, the χ^2 statistic.

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

$$H_0: \text{ Factors A and B are independent} \quad \text{vs} \quad H_A: \text{ Factors A and B are dependent.}$$

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

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

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),

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

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

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

and estimate the *expected counts* as

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

This is the case when expected counts $Exp(i,j) = \mathbf{E}\left\{Obs(i,j) \mid H_0\right\}$ 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,

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

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 $i = 1, \ldots, k$, 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 $j = 1, \ldots, m$, 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

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

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 e-mail servers and anti-spam filters attempt to identify spam e-mails 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 e-mail to be spam depends on whether it contains images or not. The following data were collected on $n = 1000$ random e-mail 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 e-mail having an attachment is somehow related to being spam. Therefore, this piece of information can be used in anti-spam filters. $\Diamond$

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,

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

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. $\Diamond$

R and MATLAB notes

The χ^2 test for independence is readily available in R and MATLAB. For example, here is a quick solution to Example 10.4 on page 323.

— R ————

```
counts = c(160,140,240,460)          # These counts are given in Example 10.4
T = matrix(counts,2,2)               # Create a contingency table T
chisq.test(T,correct=0)              # Chi-square test for independence
```

All components of this χ^2 test are saved in the data frame `chisq.test(T,correct=0)`, and you can see the list of them by typing `names(chisq.test(T,correct=0))`. For example, to get the expected counts, enter a line `chisq.test(T,correct=0)$expected`.

The option `correct=0` asks to perform the test precisely as in Example 10.4. Without this option, R uses a slightly different method that includes a continuity correction.

The same command can be applied to counts, as we just did, or to raw data. We can create a dataset with the same counts and then test the two factors for independence as follows.

— R ————

```
X1 = rbind(matrix("Spam",400,1),matrix("No spam",600,1))
X2 = rbind(matrix("Images",160,1),matrix("No images",240,1),
           matrix("Images",140,1),matrix("No images",460,1))
chisq.test(X1,X2,correct=0)
```

A similar tool for the χ^2 test exists in MATLAB.

— Matlab ————

```
X1 = [ones(400,1); zeros(600,1)];
X2 = [ones(160,1); zeros(240,1); ones(140,1); zeros(460,1)];
[ observed, statistic, pvalue ] = crosstab(X1,X2)
```

This way, we define variables `observed`, `statistic`, and `pvalue` that will contain the observed counts, the χ^2 statistic, and the P-value of the test allowing us to decide about independence of images and spam in e-mails.

Without special commands for χ^2 tests offered by R and MATLAB, users of other software languages can write a code along the following lines.

```
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
```

This code is actually executable in MATLAB. It computes the expected counts, the test statistic, and the P-value of the test step by step.

10.2 Nonparametric statistics

Parametric 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. 284 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. 292 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 $\overline{Y} = 7.2 + 40/50 = 8.0$. 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. ◊