Chapter 9

Statistical Inference I

After taking a general look at the data, we are ready for more advanced and more informative statistical analysis.

In this chapter, we learn how

- *to estimate parameters* of the distribution. Methods of Chapter 8 mostly concern measure of location (mean, median, quantiles) and variability (variance, standard deviation, interquartile range). As we know, this does not cover all possible parameters, and thus, we still lack a general methodology of estimation.
- *to construct confidence intervals*. Any estimator, computed from a collected random sample instead of the whole population, is understood as only an approximation of the corresponding parameter. Instead of one estimator that is subject to a *sampling error*, it is often more reasonable to produce an interval that will contain the true population parameter with a certain known high probability.
- *to test hypotheses*. That is, we shall use the collected sample to verify statements and claims about the population. As a result of each test, a statement is either rejected on basis of the observed data or accepted (not rejected). Sampling error in this analysis results in a possibility of wrongfully accepting or rejecting the hypothesis; however, we can design tests to control the probability of such errors.

Results of such statistical analysis are used for making decisions under uncertainty, developing optimal strategies, forecasting, evaluating and controlling performance, and so on.

9.1 Parameter estimation

By now, we have learned a few elementary ways to determine the *family of distributions*. We take into account the nature of our data, basic description, and range; propose a suitable family of distributions; and support our conjecture by looking at a histogram.

In this section, we learn how to estimate parameters of distributions. As a result, a large family will be reduced to just one distribution that we can use for performance evaluation, forecasting, etc.

Example 9.1 (Poisson). For example, consider a sample of computer chips with a certain type of rare defects. The number of defects on each chip is recorded. This is the number of rare events, and thus, it should follow a Poisson distribution with *some* parameter λ.

We know that $\lambda = \mathbf{E}(X)$ is the expectation of a Poisson variable (Section 3.4.5). Then, should we estimate it with a sample mean $\overline{X}$? Or, should we use a sample variance s^2 because λ also equals $\text{Var}(X)$? ◊

Example 9.2 (Gamma). Suppose now that we deal with a Gamma(α, λ) family of distributions. Its parameters α and λ do not represent the mean, variance, standard deviation, or any other measures discussed in Chapter 8. What would the estimation algorithm be this time? ◊

Questions raised in these examples do not have unique answers. Statisticians developed a number of estimation techniques, each having certain optimal properties.

Two rather popular methods are discussed in this section:

– method of moments, and

– method of maximum likelihood.

Several other methods are introduced later: bootstrap in Section 10.3, Bayesian parameter estimation in Section 10.4, and least squares estimation in Chapter 11.

9.1.1 Method of moments

Moments

First, let us define the moments.

DEFINITION 9.1

The k-th **population moment** is defined as

$$\mu_k = \mathbf{E}(X^k).$$

The k-th **sample moment**

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

estimates μ_k from a sample $(X_1, \ldots, X_n)$.

The first sample moment is the sample mean $\overline{X}$.

Central moments are computed similarly, after centralizing the data, that is, subtracting the mean.

DEFINITION 9.2

For $k \geq 2$, the k-th **population central moment** is defined as

$$\mu'_k = \mathbf{E}(X - \mu_1)^k.$$

The k-th **sample central moment**

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^k$$

estimates μ_k from a sample $(X_1, \ldots, X_n)$.

Remark: The second population central moment is variance $\text{Var}(X)$. The second sample central moment is sample variance, although $(n-1)$ in its denominator is now replaced by n. We mentioned that estimation methods are not unique. For unbiased estimation of $\sigma^2 = \text{Var}(X)$, we use

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2;$$

however, method of moments and method of maximum likelihood produce a different version,

$$S^2 = m_2' = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

And this is not all! We'll see other estimates of σ^2 as well.

Estimation

Method of moments is based on a simple idea. Since our sample comes from a family of distributions $\{F(\theta)\}$, we choose such a member of this family whose properties are close to properties of our data. Namely, we shall match the *moments*.

To estimate k parameters, equate the first k population and sample moments,

$$\begin{cases} \mu_1 &= & m_1 \\ \dots & \dots & \dots \\ \mu_k &= & m_k \end{cases}$$

The left-hand sides of these equations depend on the distribution parameters. The right-hand sides can be computed from data. The **method of moments estimator** is the solution of this system of equations.

Example 9.3 (Poisson). To estimate parameter λ of Poisson(λ) distribution, we recall that

$$\mu_1 = \mathbf{E}(X) = \lambda.$$

There is only one unknown parameter, hence we write one equation,

$$\mu_1 = \lambda = m_1 = \overline{X}.$$

"Solving" it for λ, we obtain

$$\widehat{\lambda} = \overline{X},$$

the method of moments estimator of λ. ◊

This does not look difficult, does it? Simplicity is the main attractive feature of the method of moments.

If it is easier, one may opt to equate *central moments*.

Example 9.4 (Gamma distribution of CPU times). The histogram in Figure 8.6 suggested that CPU times have Gamma distribution with some parameters α and λ. To estimate them, we need two equations. From data on p. 217, we compute

$$m_1 = \overline{X} = 48.2333 \quad \text{and} \quad m_2' = S^2 = 679.7122.$$

and write two equations,

$$\begin{cases} \mu_1 = \mathbf{E}(X) = \alpha/\lambda = m_1 \\ \mu_2' = \mathrm{Var}(X) = \alpha/\lambda^2 = m_2'. \end{cases}$$

It is convenient to use the second *central* moment here because we already know the expression for the variance $m_2' = \mathrm{Var}(X)$ of a Gamma variable.

Solving this system in terms of α and λ, we get the method of moment estimates

$$\begin{cases} \widehat{\alpha} = m_1^2/m_2' = 3.4227 \\ \widehat{\lambda} = m_1/m_2' = 0.0710. \end{cases}$$

$\Diamond$

Of course, we solved these two examples so quickly because we already knew the moments of Poisson and Gamma distributions from Sections 3.4.5 and 4.2.3. When we see a new distribution for us, we'll have to compute its moments.

Consider, for example, *Pareto distribution* that plays an increasingly vital role in modern internet modeling due to very heavy internet traffic nowadays.

Example 9.5 (Pareto). A two-parameter *Pareto distribution* has a cdf

$$F(x) = 1 - \left(\frac{x}{\sigma}\right)^{-\theta} \quad \text{for } x > \sigma.$$

How should we compute method of moments estimators of σ and θ?

We have not seen Pareto distribution in this book so far, so we'll have to compute its first two moments.

We start with the density

$$f(x) = F'(x) = \frac{\theta}{\sigma}\left(\frac{x}{\sigma}\right)^{-\theta-1} = \theta\sigma^\theta x^{-\theta-1}$$

and use it to find the expectation

$$\begin{aligned} \mu_1 &= \mathbf{E}(X) = \int_\sigma^\infty x\, f(x)\, dx = \theta\sigma^\theta \int_\sigma^\infty x^{-\theta} dx \\ &= \theta\sigma^\theta \left.\frac{x^{-\theta+1}}{-\theta+1}\right|_{x=\sigma}^{x=\infty} = \frac{\theta\sigma}{\theta-1}, \quad \text{for } \theta > 1, \end{aligned}$$

and the second moment

$$\mu_2 = \mathbf{E}(X^2) = \int_\sigma^\infty x^2\, f(x)\, dx = \theta\sigma^\theta \int_\sigma^\infty x^{-\theta+1} dx = \frac{\theta\sigma^2}{\theta-2}, \quad \text{for } \theta > 2.$$

For $\theta \le 1$, a Pareto variable has an infinite expectation, and for $\theta \le 2$, it has an infinite second moment.

Then we solve the method of moments equations

$$\begin{cases} \mu_1 = \dfrac{\theta\sigma}{\theta-1} = m_1 \\ \mu_2 = \dfrac{\theta\sigma^2}{\theta-2} = m_2 \end{cases}$$

and find that

$$\widehat{\theta} = \sqrt{\frac{m_2}{m_2 - m_1^2}} + 1 \quad \text{and} \quad \widehat{\sigma} = \frac{m_1(\widehat{\theta} - 1)}{\widehat{\theta}}. \tag{9.1}$$

When we collect a sample from Pareto distribution, we can compute sample moments m_1 and m_2 and estimate parameters by (9.1). ◊

On rare occasions, when k equations are not enough to estimate k parameters, we'll consider higher moments.

Example 9.6 (NORMAL). Suppose we already know the mean μ of a Normal distribution and would like to estimate the variance σ^2. Only one parameter σ^2 is unknown; however, the first method of moments equation

$$\mu_1 = m_1$$

does not contain σ^2 and therefore does not produce its estimate. We then consider the second equation, say,

$$\mu_2' = \sigma^2 = m_2' = S^2,$$

which gives us the method of moments estimate immediately, $\widehat{\sigma}^2 = S^2$. ◊

Method of moments estimates are typically easy to compute. They can serve as a quick tool for estimating parameters of interest.

9.1.2 Method of maximum likelihood

Another interesting idea is behind the method of *maximum likelihood estimation*.

Since the sample $\boldsymbol{X} = (X_1, \dots, X_n)$ has already been observed, we find such parameters that maximize the probability (likelihood) for this to happen. In other words, we make the event that has already happened to be as likely as possible. This is yet another way to make the chosen distribution consistent with the observed data.

DEFINITION 9.3

Maximum likelihood estimator is the parameter value that maximizes the likelihood of the observed sample. For a discrete distribution, we maximize the joint pmf of data $P(X_1, \dots, X_n)$. For a continuous distribution, we maximize the joint density $f(X_1, \dots, X_n)$.

Both cases, discrete and continuous, are explained below.

Discrete case

For a discrete distribution, the probability of a given sample is the joint pmf of data,

$$\boldsymbol{P}\{\boldsymbol{X} = (X_1, \dots, X_n)\} = P(\boldsymbol{X}) = P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i),$$

because in a simple random sample, all observed X_i are independent.

To maximize this likelihood, we consider the critical points by taking derivatives with respect to all unknown parameters and equating them to 0. The maximum can only be attained at such parameter values θ where the derivative $\frac{\partial}{\partial\theta}P(\boldsymbol{X})$ equals 0, where it does not exist, or at the boundary of the set of possible values of θ (to review this, see Section 12.4.4).

A nice computational shortcut is to take logarithms first. Differentiating the sum

$$\ln\prod_{i=1}^{n}P(X_i)=\sum_{i=1}^{n}\ln P(X_i)$$

is easier than differentiating the product $\prod P(X_i)$. Besides, logarithm is an increasing function, so the likelihood $P(\boldsymbol{X})$ and the log-likelihood $\ln P(\boldsymbol{X})$ are maximized by exactly the same parameters.

Example 9.7 (Poisson). The pmf of Poisson distribution is

$$P(x)=e^{-\lambda}\frac{\lambda^x}{x!},$$

and its logarithm is

$$\ln P(x)=-\lambda+x\ln\lambda-\ln(x!).$$

Thus, we need to maximize

$$\ln P(\boldsymbol{X})=\sum_{i=1}^{n}(-\lambda+X_i\ln\lambda)+C=-n\lambda+\ln\lambda\sum_{i=1}^{n}X_i+C,$$

where $C=-\sum\ln(x!)$ is a constant that does not contain the unknown parameter λ.

Find the critical point(s) of this log-likelihood. Differentiating it and equating its derivative to 0, we get

$$\frac{\partial}{\partial\lambda}\ln P(\boldsymbol{X})=-n+\frac{1}{\lambda}\sum_{i=1}^{n}X_i=0.$$

This equation has only one solution

$$\widehat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}.$$

Since this is the only critical point, and since the likelihood vanishes (converges to 0) as $\lambda\downarrow 0$ or $\lambda\uparrow\infty$, we conclude that $\widehat{\lambda}$ is the maximizer. Therefore, it is the maximum likelihood estimator of λ.

For the Poisson distribution, the method of moments and the method of maximum likelihood returned the same estimator, $\widehat{\lambda}=\overline{X}$. $\lozenge$

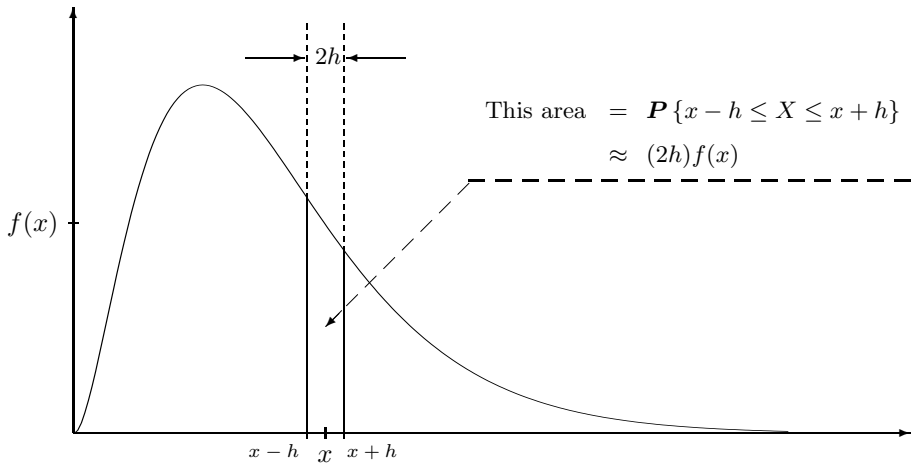


FIGURE 9.1: *Probability of observing "almost" $X = x$.*

Continuous case

In the continuous case, the probability to observe exactly the given number $X = x$ is 0, as we know from Chapter 4. Instead, the method of maximum likelihood will maximize the probability of observing "almost" the same number.

For a very small h,

$$\boldsymbol{P}\{x - h < X < x + h\} = \int_{x-h}^{x+h} f(y)dy \approx (2h)f(x).$$

That is, the probability of observing a value close to x is proportional to the density $f(x)$ (see Figure 9.1). Then, for a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, the maximum likelihood method will maximize the joint density $f(X_1, \ldots, X_n)$.

Example 9.8 (Exponential). The Exponential density is

$$f(x) = \lambda e^{-\lambda x},$$

so the log-likelihood of a sample can be written as

$$\ln f(\boldsymbol{X}) = \sum_{i=1}^{n} \ln\left(\lambda e^{-\lambda X_i}\right) = \sum_{i=1}^{n} (\ln\lambda - \lambda X_i) = n\ln\lambda - \lambda\sum_{i=1}^{n} X_i.$$

Taking its derivative with respect to the unknown parameter λ, equating it to 0, and solving for λ, we get

$$\frac{\partial}{\partial\lambda}\ln f(\boldsymbol{X}) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0,$$

resulting in

$$\widehat{\lambda} = \frac{n}{\sum X_i} = \frac{1}{\overline{X}}.$$

Again, this is the only critical point, and the likelihood $f(\boldsymbol{X})$ vanishes as $\lambda \downarrow 0$ or $\lambda \uparrow \infty$. Thus, $\widehat{\lambda} = \overline{X}$ is the maximum likelihood estimator of λ. This time, it also coincides with the method of moments estimator (Exercise 9.3b). $\Diamond$

Sometimes the likelihood has no critical points inside its domain, then it is maximized at the boundary.

Example 9.9 (UNIFORM). Based on a sample from Uniform$(0, b)$ distribution, how can we estimate the parameter b?

The Uniform$(0, b)$ density is

$$f(x) = \frac{1}{b} \quad \text{for } 0 \le x \le b.$$

It is decreasing in b, and therefore, it is maximized at the the smallest possible value of b, which is x.

For a sample $(X_1, \dots, X_n)$, the joint density

$$f(X_1, \dots, X_n) = \left(\frac{1}{b}\right)^n \quad \text{for } 0 \le X_1, \dots, X_n \le b$$

also attains its maximum at the smallest possible value of b which is now the largest observation. Indeed, $b \ge X_i$ for all i only if $b \ge \max(X_i)$. If $b < \max(X_i)$, then $f(\boldsymbol{X}) = 0$, and this cannot be the maximum value.

Therefore, the maximum likelihood estimator is $\widehat{b} = \max(X_i)$. ◊

When we estimate more than 1 parameter, all the partial derivatives should be equal 0 at the critical point. If no critical points exist, the likelihood is again maximized on the boundary.

Example 9.10 (PARETO). For the Pareto distribution in Example 9.5, the log-likelihood is

$$\ln f(\boldsymbol{X}) = \sum_{i=1}^{n} \ln\left(\theta \sigma^\theta X_i^{-\theta-1}\right) = n \ln \theta + n\theta \ln \sigma - (\theta + 1) \sum_{i=1}^{n} \ln X_i$$

for $X_1, \dots, X_n \ge \sigma$. Maximizing this function over both σ and θ, we notice that it always increases in σ. Thus, we estimate σ by its largest possible value, which is the smallest observation,

$$\widehat{\sigma} = \min(X_i).$$

We can substitute this value of σ into the log-likelihood and maximize with respect to θ,

$$\frac{\partial}{\partial \theta} \ln f(\boldsymbol{X}) = \frac{n}{\theta} + n \ln \widehat{\sigma} - \sum_{i=1}^{n} \ln X_i = 0;$$

$$\widehat{\theta} = \frac{n}{\sum \ln X_i - n \ln \widehat{\sigma}} = \frac{n}{\sum \ln (X_i / \widehat{\sigma})}.$$

The maximum likelihood estimates of σ and θ are

$$\widehat{\sigma} = \min(X_i) \quad \text{and} \quad \widehat{\theta} = \frac{n}{\sum \ln (X_i / \widehat{\sigma})}.$$

◊

Maximum likelihood estimators are rather popular because of their nice properties. Under mild conditions, these estimators are consistent, and for large samples, they have an approximately Normal distribution. Often in complicated problems, finding a good estimation scheme may be challenging whereas the maximum likelihood method always gives a reasonable solution.

9.1.3 Estimation of standard errors

How good are the estimators that we learned in Sections 9.1.1 and 9.1.2? Standard errors can serve as measures of their accuracy. To estimate them, we derive an expression for the standard error and estimate all the unknown parameters in it.

Example 9.11 (Estimation of the Poisson parameter). In Examples 9.3 and 9.7, we found the method of moments and maximum likelihood estimators of the Poisson parameter λ. Both estimators appear to be equal the sample mean $\widehat{\lambda} = \overline{X}$. Let us now estimate the *standard error* of $\widehat{\lambda}$.

Solution. There are at least two ways to do it.

On one hand, $\sigma = \sqrt{\lambda}$ for the Poisson(λ) distribution, so $\sigma(\widehat{\lambda}) = \sigma(\overline{X}) = \sigma/\sqrt{n} = \sqrt{\lambda/n}$, as we know from (8.2) on p. 219. Estimating λ by $\overline{X}$, we obtain

$$s_1(\widehat{\lambda}) = \sqrt{\frac{\overline{X}}{n}} = \frac{\sqrt{\sum X_i}}{n}.$$

On the other hand, we can use the sample standard deviation and estimate the standard error of the sample mean as in Example 8.17,

$$s_2(\widehat{\lambda}) = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n(n-1)}}.$$

Apparently, we can estimate the standard error of $\widehat{\lambda}$ by two good estimators, s_1 and s_2. $\Diamond$

Example 9.12 (Estimation of the Exponential parameter). Derive the standard error of the maximum likelihood estimator in Example 9.8 and estimate it, assuming a sample size $n \geq 3$.

Solution. This requires some integration work. Fortunately, we can take a shortcut because we know that the integral of any Gamma density is one, i.e.,

$$\int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\lambda x} dx = 1 \qquad \text{for any } \alpha > 0,\ \lambda > 0.$$

Now, notice that $\widehat{\lambda} = 1/\overline{X} = n/\sum X_i$, where $\sum X_i$ has Gamma (n, λ) distribution because each X_i is Exponential(λ).

Therefore, the k-th moment of $\widehat{\lambda}$ equals

$$\begin{aligned}
\mathbf{E}(\widehat{\lambda}^k) &= \mathbf{E}\left(\frac{n}{\sum X_i}\right)^k = \int_0^\infty \left(\frac{n}{x}\right)^k \frac{\lambda^n}{\Gamma(n)}\, x^{n-1} e^{-\lambda x} dx = \frac{n^k \lambda^n}{\Gamma(n)} \int_0^\infty x^{n-k-1} e^{-\lambda x} dx \\
&= \frac{n^k \lambda^n}{\Gamma(n)} \frac{\Gamma(n-k)}{\lambda^{n-k}} \int_0^\infty \frac{\lambda^{n-k}}{\Gamma(n-k)} x^{n-k-1} e^{-\lambda x} dx \\
&= \frac{n^k \lambda^n}{\Gamma(n)} \frac{\Gamma(n-k)}{\lambda^{n-k}} \cdot 1 = \frac{n^k \lambda^k (n-k-1)!}{(n-1)!}.
\end{aligned}$$

Substituting $k = 1$, we get the first moment,

$$\mathbf{E}(\widehat{\lambda}) = \frac{n\lambda}{n-1}.$$

Substituting $k = 2$, we get the second moment,

$$\mathbf{E}(\widehat{\lambda}^2) = \frac{n^2\lambda^2}{(n-1)(n-2)}.$$

Then, the standard error of $\widehat{\lambda}$ is

$$\sigma(\widehat{\lambda}) = \sqrt{\mathrm{Var}(\widehat{\lambda})} = \sqrt{\mathbf{E}(\widehat{\lambda}^2) - \mathbf{E}^2(\widehat{\lambda})} = \sqrt{\frac{n^2\lambda^2}{(n-1)(n-2)} - \frac{n^2\lambda^2}{(n-1)^2}} = \frac{n\lambda}{(n-1)\sqrt{n-2}}.$$

We have just estimated λ by $\widehat{\lambda} = 1/\overline{X}$; therefore, we can estimate the standard error $\sigma(\widehat{\lambda})$ by

$$s(\widehat{\lambda}) = \frac{n}{\overline{X}(n-1)\sqrt{n-2}} \quad \text{or} \quad \frac{n^2}{\sum X_i (n-1)\sqrt{n-2}}.$$

$\Diamond$

Although this was not too long, estimation of standard errors can become much harder for just slightly more complex estimators. In some cases, a nice analytic formula for $\sigma(\widehat{\theta})$ may not exist. Then, a modern method of *bootstrap* will be used, and we discuss it in Section 10.3.

Computer notes

R package "MASS" (Modern Applied Statistics with S) includes maximum likelihood estimation for almost all the distributions discussed in this book. You just specify the observed variable and the family of distributions in R command `fitdistr`. For example, `fitdistr(X,'normal')`, `fitdistr(X,'Poisson')`, or `fitdistr(X,'geometric')`. Here is an R solution to Example 9.4.

— R —

```
install.packages("MASS")                # Invoke package "MASS"
library(MASS)
x<-c(70,36,43,69,82,48,34,              # Enter the data from Example 9.4
62,35,15,59,139,46,37,42,30,55,56,36,82,38,89,54,25,35,24,22,9,56,19);
fitdistr(x,'gamma')
```

As a result, R returns parameter estimates $\widehat{\alpha}$ (shape) and $\widehat{\lambda}$ (frequency or rate) along with their estimated standard errors in parentheses.

```
    shape          rate
 3.63007913    0.07526080
(0.89719441)  (0.01994748)
```

MATLAB has a similar tool...

— MATLAB ————

```
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);
fitdist(x,'gamma')
```

... with a slightly different output.

```
Gamma distribution
   a = 3.63007  [2.23591, 5.89356]
   b = 13.2871  [7.90162, 22.3433]
```

MATLAB understands the second parameter of Gamma distribution as a *scale parameter* β instead of a *frequency parameter* λ. They are directly related, $\beta = 1/\lambda$. Also, instead of standard errors, it attaches *confidence intervals* of both parameters. Confidence intervals? We learn them in the next section.

9.2 Confidence intervals

When we report an estimator $\widehat{\theta}$ of a population parameter θ, we know that most likely

$$\widehat{\theta} \neq \theta$$

due to a sampling error. We realize that we have estimated θ *up to some error*. Likewise, nobody understands the internet connection of 11 megabytes per second as exactly 11 megabytes going through the network every second, and nobody takes a meteorological forecast as the promise of exactly the predicted temperature.

Then how much can we trust the reported estimator? How far can it be from the actual parameter of interest? What is the probability that it will be reasonably close? And if we observed an estimator $\widehat{\theta}$, then what can the actual parameter θ be?

To answer these questions, statisticians use *confidence intervals*, which contain parameter values that deserve some confidence, given the observed data.

DEFINITION 9.4

An interval $[a, b]$ is a $(1-\alpha)100\%$ **confidence interval** for the parameter θ if it contains the parameter with probability $(1-\alpha)$,

$$\boldsymbol{P}\{a \leq \theta \leq b\} = 1 - \alpha.$$

The **coverage probability** $(1-\alpha)$ is also called a **confidence level**.

Let us take a moment to think about this definition. The probability of a random event $\{a \leq \theta \leq b\}$ has to be $(1-\alpha)$. What randomness is involved in this event?

The population parameter θ is not random. It is a population feature, independent of any random sampling procedure, and therefore, it remains constant. On the other hand,

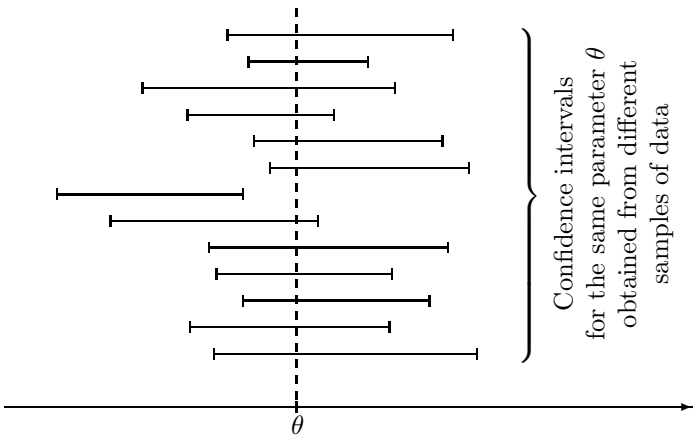


FIGURE 9.2: *Confidence intervals and coverage of parameter θ.*

the interval is computed from random data, and therefore, it is random. The *coverage probability* refers to the chance that our interval covers a constant parameter θ.

This is illustrated in Figure 9.2. Suppose that we collect many random samples and produce a confidence interval from each of them. If these are $(1-\alpha)100\%$ confidence intervals, then we expect $(1-\alpha)100\%$ of them to cover θ and $100\alpha\%$ of them to miss it. In Figure 9.2, we see one interval that does not cover θ. No mistake was made in data collection and construction of this interval. It missed the parameter only due to a *sampling error*.

It is therefore *wrong* to say, *"I computed a 90% confidence interval, it is* $[3, 6]$*. Parameter belongs to this interval with probability 90%."* The parameter is constant; it either belongs to the interval $[3, 6]$ (with probability 1) or does not. In this case, 90% refers to the proportion of confidence intervals that contain the unknown parameter in a long run.

9.2.1 Construction of confidence intervals: a general method

Given a sample of data and a desired confidence level $(1-\alpha)$, how can we construct a confidence interval $[a, b]$ that will satisfy the coverage condition

$$\boldsymbol{P}\{a \le \theta \le b\} = 1 - \alpha$$

in Definition 9.4?

We start by estimating parameter θ. *Assume there is an unbiased estimator $\widehat{\theta}$ that has a Normal distribution.* When we standardize it, we get a Standard Normal variable

$$Z = \frac{\widehat{\theta} - \mathbf{E}(\widehat{\theta})}{\sigma(\widehat{\theta})} = \frac{\widehat{\theta} - \theta}{\sigma(\widehat{\theta})}, \tag{9.2}$$

where $\mathbf{E}(\widehat{\theta}) = \theta$ because $\widehat{\theta}$ is unbiased, and $\sigma(\widehat{\theta}) = \sigma(\widehat{\theta})$ is its standard error.

This variable falls between the Standard Normal quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$, denoted by

$$\begin{aligned} -z_{\alpha/2} &= q_{\alpha/2} \\ z_{\alpha/2} &= q_{1-\alpha/2} \end{aligned}$$

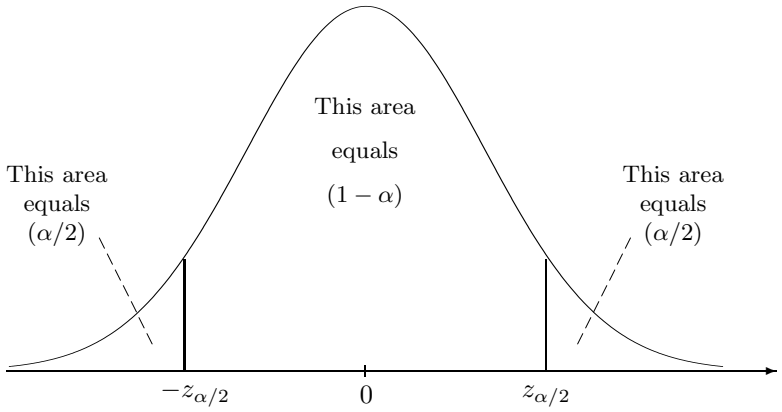


FIGURE 9.3: *Standard Normal quantiles $\pm z_{\alpha/2}$ and partition of the area under the density curve.*

with probability $(1-\alpha)$, as you can see in Figure 9.3.

Then,

$$\boldsymbol{P}\left\{-z_{\alpha/2} \le \frac{\widehat{\theta}-\theta}{\sigma(\widehat{\theta})} \le z_{\alpha/2}\right\} = 1-\alpha.$$

Solving the inequality inside $\{\ldots\}$ for θ, we get

$$\boldsymbol{P}\left\{\widehat{\theta} - z_{\alpha/2}\cdot\sigma(\widehat{\theta}) \le \theta \le \widehat{\theta} + z_{\alpha/2}\cdot\sigma(\widehat{\theta})\right\} = 1-\alpha.$$

The problem is solved! We have obtained two numbers

$$\begin{aligned} a &= \widehat{\theta} - z_{\alpha/2}\cdot\sigma(\widehat{\theta}) \\ b &= \widehat{\theta} + z_{\alpha/2}\cdot\sigma(\widehat{\theta}) \end{aligned}$$

such that

$$\boldsymbol{P}\{a \le \theta \le b\} = 1-\alpha.$$

Confidence interval, Normal distribution

If parameter θ has an unbiased, Normally distributed estimator $\widehat{\theta}$, then

$$\widehat{\theta} \pm z_{\alpha/2}\cdot\sigma(\widehat{\theta}) = \left[\widehat{\theta} - z_{\alpha/2}\cdot\sigma(\widehat{\theta}),\ \widehat{\theta} + z_{\alpha/2}\cdot\sigma(\widehat{\theta})\right]$$

is a $(1-\alpha)100\%$ confidence interval for θ.

If the distribution of $\widehat{\theta}$ is *approximately* Normal, we get an *approximately* $(1-\alpha)100\%$ confidence interval. (9.3)

In this formula, $\widehat{\theta}$ is the **center of the interval**, and $z_{\alpha/2} \cdot \sigma(\widehat{\theta})$ is the **margin**. The margin of error is often reported along with poll and survey results. In newspapers and press releases, it is usually computed for a 95% confidence interval.

We have seen quantiles $\pm z_{\alpha/2}$ in inverse problems (Example 4.12 on p. 91). Now, in confidence estimation, and also, in the next section on hypothesis testing, they will play a crucial role as we'll need to attain the desired confidence level α. The most commonly used values are

$$\begin{gathered} z_{0.10} = 1.282, \qquad z_{0.05} = 1.645, \qquad z_{0.025} = 1.960, \\ z_{0.01} = 2.326, \qquad z_{0.005} = 2.576. \end{gathered} \tag{9.4}$$

NOTATION $\left\| \begin{array}{l} z_\alpha = q_{1-\alpha} = \Phi^{-1}(1-\alpha) \text{ is the value of a Standard} \\ \text{Normal variable } Z \text{ that is exceeded with probability } \alpha \end{array} \right\|$

Several important applications of this general method are discussed below. In each problem, we

(a) find an unbiased estimator of θ,

(b) check if it has a Normal distribution,

(c) find its standard error $\sigma(\widehat{\theta}) = \text{Std}(\widehat{\theta})$,

(d) obtain quantiles $\pm z_{\alpha/2}$ from the table of Normal distribution (Table A4 in the Appendix), and finally,

(e) apply the rule (9.3).

9.2.2 Confidence interval for the population mean

Let us construct a confidence interval for the population mean

$$\theta = \mu = \mathbf{E}(X).$$

Start with an estimator,

$$\widehat{\theta} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The rule (9.3) is applicable in two cases.

1. If a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ comes from Normal distribution, then $\overline{X}$ is also Normal, and rule (9.3) can be applied.

2. If a sample comes from *any* distribution, but the sample size n is large, then $\overline{X}$ has an approximately Normal distribution according to the Central Limit Theorem on p. 93. Then rule (9.3) gives an approximately $(1-\alpha)100\%$ confidence interval.

In Section 8.2.1, we derived

$$\begin{aligned} \mathbf{E}(\overline{X}) &= \mu \qquad \text{(thus, it is an unbiased estimator);} \\ \sigma(\overline{X}) &= \sigma/\sqrt{n}. \end{aligned}$$

Then, (9.3) reduces to the following $(1-\alpha)100\%$ confidence interval for μ.

Confidence interval for the mean; σ is known

$$\boxed{\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}} \tag{9.5}$$

Example 9.13. Construct a 95% confidence interval for the population mean based on a sample of measurements

$$2.5,\ 7.4,\ 8.0,\ 4.5,\ 7.4,\ 9.2$$

if measurement errors have Normal distribution, and the measurement device guarantees a standard deviation of $\sigma = 2.2$.

Solution. This sample has size $n = 6$ and sample mean $\overline{X} = 6.50$. To attain a confidence level of

$$1 - \alpha = 0.95,$$

we need $\alpha = 0.05$ and $\alpha/2 = 0.025$. Hence, we are looking for quantiles

$$q_{0.025} = -z_{0.025} \qquad \text{and} \qquad q_{0.975} = z_{0.025}.$$

From (9.4) or Table A4, we find that $q_{0.975} = 1.960$. Substituting these values into (9.5), we obtain a 95% confidence interval for μ,

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 6.50 \pm (1.960)\frac{2.2}{\sqrt{6}} = \underline{6.50 \pm 1.76} \text{ or } \underline{[4.74,\ 8.26].}$$

$\Diamond$

The only situation when method (9.3) cannot be applied is when the sample size is small and the distribution of data is not Normal. Special methods for the given distribution of $\boldsymbol{X}$ are required in this case.

9.2.3 Confidence interval for the difference between two means

Under the same conditions as in the previous section,

– Normal distribution of data or
– sufficiently large sample size,

we can construct a confidence interval for the *difference* between two means.

This problem arises when we compare two populations. It may be a comparison of two materials, two suppliers, two service providers, two communication channels, two labs, etc. From each population, a sample is collected (Figure 9.4),

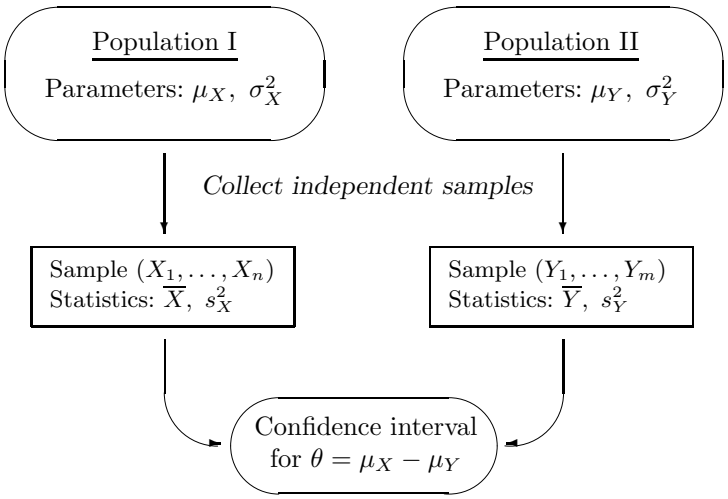


FIGURE 9.4: *Comparison of two populations.*

$\boldsymbol{X} = (X_1, \dots, X_n)$ from one population,
$\boldsymbol{Y} = (Y_1, \dots, Y_m)$ from the other population.

Suppose that the two samples are collected **independently** of each other.

To construct a confidence interval for the difference between population means

$$\theta = \mu_X - \mu_Y,$$

we complete the usual steps (a)–(e) below.

(a) Propose an estimator of θ,

$$\widehat{\theta} = \overline{X} - \overline{Y}.$$

It is natural to come up with this estimator because $\overline{X}$ estimates μ_X and $\overline{Y}$ estimates μ_Y.

(b) Check that $\widehat{\theta}$ is unbiased. Indeed,

$$\mathbf{E}(\widehat{\theta}) = \mathbf{E}\left(\overline{X} - \overline{Y}\right) = \mathbf{E}\left(\overline{X}\right) - \mathbf{E}\left(\overline{Y}\right) = \mu_X - \mu_Y = \theta.$$

(c) Check that $\widehat{\theta}$ has a Normal or approximately Normal distribution. This is true if the observations are Normal or *both* sample sizes m and n are large.

(d) Find the standard error of $\widehat{\theta}$ (using independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$),

$$\sigma(\widehat{\theta}) = \sqrt{\mathrm{Var}\left(\overline{X} - \overline{Y}\right)} = \sqrt{\mathrm{Var}\left(\overline{X}\right) + \mathrm{Var}\left(\overline{Y}\right)} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

(e) Find quantiles $\pm z_{\alpha/2}$ and compute the confidence interval according to (9.3). This results in the following formula.

Confidence interval for the difference of means; known standard deviations

$$\boxed{\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \tag{9.6}$$

Example 9.14 (EFFECT OF AN UPGRADE). A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 minutes before the upgrade, 7.2 minutes after it. Historically, the standard deviation has been 1.8 minutes, and presumably it has not changed. Construct a 90% confidence interval showing how much the mean running time reduced due to the hardware upgrade.

Solution. We have $n = m = 50$, $\sigma_X = \sigma_Y = 1.8$, $\overline{X} = 8.5$, and $\overline{Y} = 7.2$. Also, the confidence level $(1 - \alpha)$ equals 0.9, hence $\alpha/2 = 0.05$, and $z_{\alpha/2} = 1.645$.

The distribution of times may not be Normal; however, due to large sample sizes, the estimator

$$\widehat{\theta} = \overline{X} - \overline{Y}$$

is approximately Normal by the Central Limit Theorem. Thus, formula (9.6) is applicable, and a 90% confidence interval for the difference of means $(\mu_X - \mu_Y)$ is

$$8.5 - 7.2 \pm (1.645)\sqrt{1.8^2\left(\frac{1}{50} + \frac{1}{50}\right)} = \underline{1.3 \pm 0.6 \text{ or } [0.7, 1.9]}.$$

We can say that the hardware upgrade resulted in a 1.3-minute reduction of the mean running time, with a 90% confidence margin of 0.6 minutes. ◊

9.2.4 Selection of a sample size

Formula (9.3) describes a confidence interval as

$$\text{center} \ \pm \ \text{margin}$$

where

$$\begin{aligned} \text{center} &= \widehat{\theta}, \\ \text{margin} &= z_{\alpha/2} \cdot \sigma(\widehat{\theta}). \end{aligned}$$

We can revert the problem and ask a very practical question: *How large a sample should be collected to provide a certain desired precision of our estimator?*

In other words, what sample size n guarantees that the margin of a $(1-\alpha)100\%$ confidence interval does not exceed a specified limit Δ?

To answer this question, we only need to solve the inequality

$$\text{margin} \le \Delta \tag{9.7}$$

in terms of n. Typically, parameters are estimated more accurately based on larger samples, so that the standard error $\sigma(\widehat{\theta})$ and the margin are decreasing functions of sample size n. Then, (9.7) must be satisfied for sufficiently large n.

9.2.5 Estimating means with a given precision

When we estimate a population mean, the margin of error is

$$\text{margin} = z_{\alpha/2} \cdot \sigma/\sqrt{n}.$$

Solving inequality (9.7) for n results in the following rule.

Sample size for a given precision

> In order to attain a margin of error Δ for estimating a population mean with a confidence level $(1-\alpha)$,
>
> a sample of size $n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{\Delta}\right)^2$ is required. (9.8)

When we compute the expression in (9.8), it will most likely be a fraction. Notice that we can only *round it up* to the nearest integer sample size. If we round it down, our margin will exceed Δ.

Looking at (9.8), we see that a large sample will be necessary

- to attain a narrow margin (small Δ);
- to attain a high confidence level (small α); and
- to control the margin under high variability of data (large σ).

In particular, we need to quadruple the sample size in order to half the margin of the interval.

Example 9.15. In Example 9.13, we constructed a 95% confidence with the center 6.50 and margin 1.76 based on a sample of size 6. Now, that was too wide, right? How large a sample do we need to estimate the population mean with a margin of at most 0.4 units with 95% confidence?

Solution. We have $\Delta = 0.4$, $\alpha = 0.05$, and from Example 9.13, $\sigma = 2.2$. By (9.8), we need a sample of

$$n \geq \left(\frac{z_{0.05/2} \cdot \sigma}{\Delta}\right)^2 = \left(\frac{(1.960)(2.2)}{0.4}\right)^2 = 116.2.$$

Keeping in mind that this is the minimum sample size that satisfies Δ, and we are only allowed to round it up, we need a sample of at least 117 observations.

$\Diamond$

9.3 Unknown standard deviation

A rather heavy condition was assumed when we constructed all the confidence intervals. We assumed a *known standard deviation* σ and used it in all the derived formulas.

Sometimes this assumption is perfectly valid. We may know the variance from a large archive of historical data, or it may be given as precision of a measuring device.

Much more often, however, the population variance is unknown. We'll then estimate it from data and see if we can still apply methods of the previous section.

Two broad situations will be considered:

– large samples from any distribution,

– samples of any size from a Normal distribution.

In the only remaining case, a small non-Normal sample, a confidence interval will be constructed by special methods. A popular modern approach called *bootstrap* is discussed in Section 10.3.3.

9.3.1 Large samples

A large sample should produce a rather accurate estimator of a variance. We can then replace the true standard error $\sigma(\widehat{\theta})$ in (9.3) by its estimator $s(\widehat{\theta})$, and obtain an approximate confidence interval

$$\widehat{\theta} \pm z_{\alpha/2} \cdot s(\widehat{\theta}).$$

Example 9.16 (Delays at nodes). Internet connections are often slowed by delays at nodes. Let us determine if the delay time increases during heavy-volume times.

Five hundred packets are sent through the same network between 5 pm and 6 pm (sample $\boldsymbol{X}$), and three hundred packets are sent between 10 pm and 11 pm (sample $\boldsymbol{Y}$). The early sample has a mean delay time of 0.8 sec with a standard deviation of 0.1 sec whereas the second sample has a mean delay time of 0.5 sec with a standard deviation of 0.08 sec. Construct a 99.5% confidence interval for the difference between the mean delay times.

Solution. We have $n = 500$, $\overline{X} = 0.8$, $s_X = 0.1$; $m = 300$, $\overline{Y} = 0.5$, $s_Y = 0.08$. Large sample sizes allow us to replace unknown population standard deviations by their estimates and use an approximately Normal distribution of sample means.

For a confidence level of $1 - \alpha = 0.995$, we need

$$z_{\alpha/2} = z_{0.0025} = q_{0.9975}.$$

Look for the *probability* 0.9975 in the body of Table A4 and find the corresponding value of z,

$$z_{0.0025} = 2.81.$$

Then, a 99.5% confidence interval for the difference of mean execution times is

$$\begin{aligned} \overline{X}-\overline{Y} &\pm z_{0.0025}\sqrt{\frac{s_X^2}{n}+\frac{s_Y^2}{m}} = (0.8-0.5)\pm(2.81)\sqrt{\frac{(0.1)^2}{500}+\frac{(0.08)^2}{300}} \\ &= \underline{0.3\pm 0.018} \text{ or } \underline{[0.282,\, 0.318]}. \end{aligned}$$

$\Diamond$

9.3.2 Confidence intervals for proportions

In particular, we surely don't know the variance when we estimate a population proportion.

DEFINITION 9.5

We assume a subpopulation A of items that have a certain *attribute*. By the **population proportion** we mean the probability

$$p = \boldsymbol{P}\{i \in A\}$$

for a randomly selected item i to have this attribute.

A **sample proportion**

$$\widehat{p} = \frac{\text{number of sampled items from } A}{n}$$

is used to estimate p.

Let us use the *indicator* variables

$$X_i = \begin{cases} 1 & \text{if} \quad i \in A \\ 0 & \text{if} \quad i \notin A \end{cases}$$

Each X_i has Bernoulli distribution with parameter p. In particular,

$$\mathbf{E}(X_i) = p \quad \text{and} \quad \text{Var}(X_i) = p(1-p).$$

Also,

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is nothing but a sample mean of X_i.

Therefore,

$$\mathbf{E}(\widehat{p}) = p \quad \text{and} \quad \text{Var}(\widehat{p}) = \frac{p(1-p)}{n},$$

as we know from properties of sample means on p. 219.

We conclude that

1. a sample proportion $\widehat{p}$ is unbiased for the population proportion p;
2. it has approximately Normal distribution for large samples, because it has a form of a sample mean;

3. when we construct a confidence interval for p, we do not know the standard deviation $\text{Std}\,(\widehat{p})$.

Indeed, knowing the standard deviation is equivalent to knowing p, and if we know p, why would we need a confidence interval for it?

Thus, we estimate the unknown standard error

$$\sigma(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

by

$$s(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

and use it in the general formula

$$\widehat{p} \pm z_{\alpha/2} \cdot s(\widehat{p})$$

to construct an approximate $(1-\alpha)100\%$ confidence interval.

Confidence interval for a population proportion

$$\widehat{p} \pm z_{\alpha/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Similarly, we can construct a confidence interval for the *difference between two proportions*. In two populations, we have proportions p_1 and p_2 of items with an attribute. Independent samples of size n_1 and n_2 are collected, and both parameters are estimated by sample proportions $\widehat{p}_1$ and $\widehat{p}_2$.

Summarizing, we have

Parameter of interest:	$\theta = p_1 - p_2$
Estimated by:	$\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2$
Its standard error:	$\sigma(\widehat{\theta}) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Estimated by:	$s(\widehat{\theta}) = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$

Confidence interval for the difference of proportions

$$\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

Example 9.17 (PRE-ELECTION POLL). A candidate prepares for the local elections. During his campaign, 42 out of 70 randomly selected people in town A and 59 out of 100 randomly selected people in town B showed they would vote for this candidate. Estimate the difference in support that this candidate is getting in towns A and B with 95% confidence. Can we state affirmatively that the candidate gets a stronger support in town A?

Solution. We have $n_1 = 70$, $n_2 = 100$, $\widehat{p}_1 = 42/70 = 0.6$, and $\widehat{p}_2 = 59/100 = 0.59$. For the confidence interval, we have

$$\text{center} = \widehat{p}_1 - \widehat{p}_2 = 0.01,$$

and

$$\begin{aligned} \text{margin} &= z_{0.05/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \\ &= (1.960)\sqrt{\frac{(0.6)(0.4)}{70} + \frac{(0.59)(0.41)}{100}} = 0.15. \end{aligned}$$

Then

$$0.01 \pm 0.15 = \underline{[-0.14,\ 0.16]}$$

is a 95% confidence interval for the difference in support $(p_1 - p_2)$ in the two towns.

So, is the support stronger in town A? On one hand, the estimator $\widehat{p}_1 - \widehat{p}_2 = 0.01$ suggests that the support is 1% higher in town A than in town B. On the other hand, the difference could appear positive just because of a sampling error. As we see, the 95% confidence interval includes a large range of negative values too. Therefore, the obtained data does *not* indicate affirmatively that the support in town A is stronger.

In fact, we will test in Example 9.33 if there is any difference between the two towns and will conclude that there is no evidence for it or against it. A formal procedure for testing statements like that will be introduced in Section 9.4. ◊

9.3.3 Estimating proportions with a given precision

Our confidence interval for a population proportion has a margin

$$\text{margin} = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

A standard way of finding the sample size that provides the desired margin Δ is to solve the inequality

$$\text{margin} \le \Delta \quad \text{or} \quad n \ge \widehat{p}(1-\widehat{p})\left(\frac{z_{\alpha/2}}{\Delta}\right)^2.$$

However, this inequality includes $\widehat{p}$. To know $\widehat{p}$, we first need to collect a sample, but to know the sample size, we first need to know $\widehat{p}$!

A way out of this circle is shown in Figure 9.5. As we see, the function $\widehat{p}(1-\widehat{p})$ never exceeds 0.25. Therefore, we can replace the unknown value of $\widehat{p}(1-\widehat{p})$ by 0.25 and find a sample size n, perhaps larger than we actually need, that will ensure that we estimate $\widehat{p}$ with a margin not exceeding Δ. That is, choose a sample size

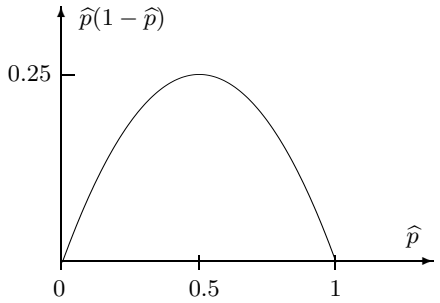


FIGURE 9.5: *Function $\widehat{p}(1-\widehat{p})$ attains its maximum at $\widehat{p}=0.5$.*

$$\boxed{n \geq 0.25 \left(\frac{z_{\alpha/2}}{\Delta}\right)^2.}$$

It will automatically be at least as large as the required $\widehat{p}(1-\widehat{p})(z_{\alpha/2}/\Delta)^2$, regardless of the unknown value of $\widehat{p}$.

Example 9.18. A sample of size

$$n \geq 0.25 \left(\frac{1.960}{0.1}\right)^2 = 96.04$$

(that is, at least 97 observations) always guarantees that a population proportion is estimated with an error of at most 0.1 with a 95% confidence. ◊

9.3.4 Small samples: Student's t distribution

Having a small sample, we can no longer pretend that a sample standard deviation s is an accurate estimator of the population standard deviation σ. Then, how should we adjust the confidence interval when we replace σ by s, or more generally, when we replace the standard error $\sigma(\widehat{\theta})$ by its estimator $s(\widehat{\theta})$?

A famous solution was proposed by *William Gosset* (1876–1937), known by his pseudonym *Student*. Working for the Irish brewery Guinness, he derived the T-distribution for the quality control problems in brewing.

Student followed the steps similar to our derivation of a confidence interval on p. 255. Then he replaced the true but unknown standard error of $\widehat{\theta}$ by its estimator $s(\widehat{\theta})$ and concluded that the **T-ratio**

$$t = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})},$$

the *ratio* of two random variables, no longer has a Normal distribution!

Student figured the distribution of a T-ratio. For the problem of estimating the mean based on n Normal observations $X_1, \ldots, X_n$, this was **T-distribution** with $(n-1)$ *degrees of*

freedom. Table A5 gives critical values t_α of the T-distribution that we'll use for confidence intervals.

So, using T-distribution instead of Standard Normal and estimated standard error instead of the unknown true one, we obtain the confidence interval for the population mean.

Confidence interval for the mean; σ is unknown

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is a critical value from T-distribution with $n-1$ degrees of freedom (9.9)

Example 9.19 (Unauthorized use of a computer account). If an unauthorized person accesses a computer account with the correct username and password (stolen or cracked), can this intrusion be detected? Recently, a number of methods have been proposed to detect such unauthorized use. The time between keystrokes, the time a key is depressed, the frequency of various keywords are measured and compared with those of the account owner. If there are significant differences, an intruder is detected.

The following times between keystrokes were recorded when a user typed the username and password (data set `Keystrokes`):

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

As the first step in detecting an intrusion, let's construct a 99% confidence interval for the mean time between keystrokes assuming Normal distribution of these times.

Solution. The sample size is $n = 18$, the sample mean time is $\overline{X} = 0.29$ sec, and the sample standard deviation is $s = 0.074$. The critical value of t distribution with $n - 1 = 17$ degrees of freedom is $t_{\alpha/2} = t_{0.005} = 2.898$. Then, the 99% confidence interval for the mean time is

$$0.29 \pm (2.898)\frac{0.074}{\sqrt{18}} = 0.29 \pm 0.05 = \underline{[0.24;\ 0.34]}$$

Example 9.28 on p. 283 will show whether this result signals an intrusion. ◊

The density of Student's T-distribution is a bell-shaped symmetric curve that can be easily confused with Normal. Comparing with the Normal density, its peak is lower and its tails are thicker. Therefore, a larger number t_α is generally needed to cut area α from the right tail. That is,

$$t_\alpha > z_\alpha$$

for small α. As a consequence, the confidence interval (9.9) is wider than the interval (9.5) for the case of known σ. This wider margin is the price paid for not knowing the standard deviation σ. When we lack a certain piece of information, we cannot get a more accurate estimator.

However, we see in Table A5 that

$$t_\alpha \to z_\alpha,$$

as the number of degrees of freedom ν tends to infinity. Indeed, having a large sample

(hence, large $\nu = n - 1$), we can count on a very accurate estimator of σ, and thus, the confidence interval is almost as narrow as if we knew σ in this case.

Degrees of freedom ν is the parameter of T-distribution controlling the shape of the T-density curve. Its meaning is the *dimension* of a vector used to estimate the variance. Here we estimate σ^2 by a sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

and thus, we use a vector

$$\boldsymbol{X}' = \left(X_1 - \overline{X}, \ldots, X_n - \overline{X} \right).$$

The initial vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ has dimension n; therefore, it has n degrees of freedom. However, when the sample mean $\overline{X}$ is subtracted from each observation, there appears a linear relation among the elements,

$$\sum_{i=1}^{n} (X_i - \overline{X}) = 0.$$

We lose 1 degree of freedom due to this constraint; the vector $\boldsymbol{X}'$ belongs to an $(n-1)$-dimensional hyperplane, and this is why we have only $\nu = n - 1$ degrees of freedom.

In many similar problems, degrees of freedom can be computed as

$$\begin{array}{c}\text{number of} \\ \text{degrees of freedom}\end{array} = \text{sample size} - \begin{array}{c}\text{number of estimated} \\ \text{location parameters}\end{array} \tag{9.10}$$

9.3.5 Comparison of two populations with unknown variances

We now construct a confidence interval for the difference of two means $\mu_X - \mu_Y$, comparing the population of X's with the population of Y's.

Again, independent random samples are collected,

$$\boldsymbol{X} = (X_1, \ldots, X_n) \quad \text{and} \quad \boldsymbol{Y} = (Y_1, \ldots, Y_m),$$

one from each population, as in Figure 9.4 on p. 259. This time, however, population variances σ_X^2 and σ_Y^2 are unknown to us, and we use their estimates.

Two important cases need to be considered here. In one case, there exists an exact and simple solution based on T-distribution. The other case suddenly appears to be a famous *Behrens–Fisher problem*, where no exact solution exists, and only approximations are available.

Case 1. Equal variances

Suppose there are reasons to assume that the two populations have equal variances,

$$\sigma_X^2 = \sigma_Y^2 = \sigma^2.$$

For example, two sets of data are collected with the same measurement device, thus, measurements have different means but the same precision.

In this case, there is only one variance σ^2 to estimate instead of two. We should use both samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ to estimate their common variance. This estimator of σ^2 is called a **pooled sample variance**, and it is computed as

$$s_p^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2}{n+m-2} = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}. \tag{9.11}$$

Substituting this variance estimator in (9.6) for σ_X^2 and σ_Y^2, we get the following confidence interval.

Confidence interval for the difference of means; equal, unknown standard deviations

$$\overline{X} - \overline{Y} \pm t_{\alpha/2}\, s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

where s_p is the *pooled standard deviation*, a root of the pooled variance in (9.11)

and $t_{\alpha/2}$ is a critical value from T-distribution with $(n+m-2)$ degrees of freedom

Example 9.20 (CD WRITER AND BATTERY LIFE). CD writing is energy consuming; therefore, it affects the battery lifetime on laptops. To estimate the effect of CD writing, 30 users are asked to work on their laptops until the "low battery" sign comes on.

Eighteen users without a CD writer worked an average of 5.3 hours with a standard deviation of 1.4 hours. The other twelve, who used their CD writer, worked an average of 4.8 hours with a standard deviation of 1.6 hours. Assuming Normal distributions with equal population variances ($\sigma_X^2 = \sigma_Y^2$), construct a 95% confidence interval for the battery life reduction caused by CD writing.

Solution. Effect of the CD writer is measured by the reduction of the mean battery life. We have $n = 12$, $\overline{X} = 4.8$, $s_X = 1.6$ for users with a CD writer and $m = 18$, $\overline{Y} = 5.3$, $s_Y = 1.4$ for users without it. The pooled standard deviation is

$$s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}} = \sqrt{\frac{(11)(1.6)^2 + (17)(1.4)^2}{28}} = 1.4818$$

(check: it has to be between s_X and s_Y). The critical value is $t_{0.025} = 2.048$ (use 28 d.f.). The 95% confidence interval for the difference between the mean battery lives is

$$(4.8 - 5.3) \pm (2.048)(1.4818)\sqrt{\frac{1}{18} + \frac{1}{12}} = -0.5 \pm 1.13 = \underline{[-1.63;\ 0.63]}.$$

◊

Remark: Let's discuss formula (9.11). First, notice that different sample means $\overline{X}$ and $\overline{Y}$ are used

for X-terms and Y-terms. Indeed, our two populations may have different means. As we know, variance of any variable measures its deviation from its mean. Thus, from each observation we subtract its own mean-estimate.

Second, we lose 2 degrees of freedom due to the estimation of two means. Two constraints,

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0 \quad \text{and} \quad \sum_{i=1}^{m}(Y_i - \overline{Y}) = 0,$$

show that the number of **degrees of freedom** is only $(n+m-2)$ instead of $(n+m)$. We see this coefficient in the denominator, and it makes s_p^2 an unbiased estimator of σ^2 (see Exercise 9.19).

Case 2. Unequal variances

The most difficult case is when both variances are unknown and unequal. Confidence estimation of $\mu_X - \mu_Y$ in this case is known as the *Behrens–Fisher problem*. Certainly, we can replace unknown variances σ_X^2, σ_Y^2 by their estimates s_X^2, s_Y^2 and form a T-ratio

$$t = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}.$$

However, it won't have a T-distribution.

An approximate solution was proposed in the 1940s by *Franklin E. Satterthwaite*, who worked for General Electric Company at that time. Satterthwaite used the method of moments to estimate degrees of freedom ν of a T-distribution that is "closest" to this T-ratio. This number depends on unknown variances. Estimating them by sample variances, he obtained the formula that is now known as *Satterthwaite approximation*,

$$\nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}}. \tag{9.12}$$

This number of degrees of freedom often appears non-integer. There are T-distributions with non-integer ν, see Section 12.2.1. To use Table A5, just take the closest ν that is given in that table.

Formula (9.12) is widely used for t-intervals and t-tests.

Confidence interval for the difference of means; unequal, unknown standard deviations

$$\overline{X} - \overline{Y} \pm t_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

where $t_{\alpha/2}$ is a critical value from T-distribution with ν degrees of freedom given by formula (9.12)

Example 9.21 (COMPARISON OF TWO SERVERS). An account on server A is more expensive than an account on server B. However, server A is faster. To see if it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is. A certain computer algorithm is executed 30 times on server A and 20 times on server B with the following results,

	Server A	**Server B**
Sample mean	6.7 min	7.5 min
Sample standard deviation	0.6 min	1.2 min

Construct a 95% confidence interval for the difference $\mu_1 - \mu_2$ between the mean execution times on server A and server B, assuming that the observed times are approximately Normal.

Solution. We have $n = 30$, $m = 20$, $\overline{X} = 6.7$, $\overline{Y} = 7.5$, $s_X = 0.6$, and $s_Y = 1.2$. The second standard deviation is twice larger than the first one; therefore, equality of population variances can hardly be assumed. We use the method for unknown, unequal variances.

Using Satterthwaite approximation (9.12), we find degrees of freedom:

$$\nu = \frac{\left(\frac{(0.6)^2}{30} + \frac{(1.2)^2}{20}\right)^2}{\frac{(0.6)^4}{30^2(29)} + \frac{(1.2)^4}{20^2(19)}} = 25.4.$$

To use Table A5, we round this ν to 25 and find $t_{0.025} = 2.060$. Then, the confidence interval is

$$\begin{aligned} \overline{X} - \overline{Y} \pm t_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} &= 6.7 - 7.5 \pm (2.060)\sqrt{\frac{(0.6)^2}{30} + \frac{(1.2)^2}{20}} \\ &= \underline{-0.8 \pm 0.6} \quad \text{or} \quad \underline{[-1.4, -0.2]}. \end{aligned}$$

◊

9.4 Hypothesis testing

A vital role of Statistics is in verifying statements, claims, conjectures, and in general - *testing hypotheses*. Based on a random sample, we can use Statistics to verify whether

- a system has not been infected,
- a hardware upgrade was efficient,
- the average number of concurrent users increased by 2000 this year,
- the average connection speed is 54 Mbps, as claimed by the internet service provider,
- the proportion of defective products is at most 3%, as promised by the manufacturer,

– service times have Gamma distribution,

– the number of errors in software is independent of the manager's experience,

– etc.

Testing statistical hypotheses has wide applications far beyond Computer Science. These methods are used to prove efficiency of a new medical treatment, safety of a new automobile brand, innocence of a defendant, and authorship of a document; to establish cause-and-effect relationships; to identify factors that can significantly improve the response; to fit stochastic models; to detect information leaks; and so forth.

9.4.1 Hypothesis and alternative

To begin, we need to state exactly what we are testing. These are *hypothesis* and *alternative*.

NOTATION

$$\begin{array}{lcl} H_0 & = & \text{hypothesis (the null hypothesis)} \\ H_A & = & \text{alternative (the alternative hypothesis)} \end{array}$$

H_0 and H_A are simply two mutually exclusive statements. Each test results either in acceptance of H_0 or its rejection in favor of H_A.

A null hypothesis is always an equality, absence of an effect or relation, some "normal," usual statement that people have believed in for years. In order to overturn the common belief and to reject the hypothesis, we need *significant evidence*. Such evidence can only be provided by data. Only when such evidence is found, and when it strongly supports the alternative H_A, can the hypothesis H_0 be rejected in favor of H_A.

Based on a random sample, a statistician cannot tell whether the hypothesis is true or the alternative. We need to see the entire population to tell that. The purpose of each test is to determine whether the data provides sufficient evidence against H_0 in favor of H_A.

This is similar to a criminal trial. Typically, the jury cannot tell whether the defendant committed a crime or not. It is not their task. They are only required to determine if the presented evidence against the defendant is sufficient and convincing. By default, called *presumption of innocence*, insufficient evidence leads to acquittal.

Example 9.22. To verify that the the average connection speed is 54 Mbps, we test the hypothesis $H_0 : \mu = 54$ against the *two-sided alternative* $H_A : \mu \neq 54$, where μ is the average speed of all connections.

However, if we worry about a *low* connection speed only, we can conduct a one-sided test of

$$H_0 : \mu = 54 \quad \text{vs} \quad H_A : \mu < 54.$$

In this case, we only measure the amount of evidence supporting the *one-sided alternative* $H_A : \mu < 54$. In the absence of such evidence, we gladly accept the null hypothesis. $\Diamond$

DEFINITION 9.6

Alternative of the type $H_A : \mu \neq \mu_0$ covering regions on both sides of the hypothesis ($H_0 : \mu = \mu_0$) is a **two-sided alternative**.

Alternative $H_A : \mu < \mu_0$ covering the region to the left of H_0 is **one-sided, left-tail**.

Alternative $H_A : \mu > \mu_0$ covering the region to the right of H_0 is **one-sided, right-tail**.

Example 9.23. To verify whether the average number of concurrent users increased by 2000, we test

$$H_0 : \mu_2 - \mu_1 = 2000 \quad \text{vs} \quad H_A : \mu_2 - \mu_1 \neq 2000,$$

where μ_1 is the average number of concurrent users last year, and μ_2 is the average number of concurrent users this year. Depending on the situation, we may replace the *two-sided alternative* $H_A : \mu_2 - \mu_1 \neq 2000$ with a one-sided alternative $H_A^{(1)} : \mu_2 - \mu_1 < 2000$ or $H_A^{(2)} : \mu_2 - \mu_1 > 2000$. The test of H_0 against $H_A^{(1)}$ evaluates the amount of evidence that the mean number of concurrent users changed by fewer than 2000. Testing against $H_A^{(2)}$, we see if there is sufficient evidence to claim that this number increased by more than 2000. ◊

Example 9.24. To verify if the proportion of defective products is at most 3%, we test

$$H_0 : p = 0.03 \quad \text{vs} \quad H_A : p > 0.03,$$

where p is the proportion of defects in the whole shipment.

Why do we choose the *right-tail alternative* $H_A : p > 0.03$? That is because we reject the shipment only if significant evidence supporting this alternative is collected. If the data suggest that $p < 0.03$, the shipment will still be accepted. ◊

9.4.2 Type I and Type II errors: level of significance

When testing hypotheses, we realize that all we see is a random sample. Therefore, with all the best statistics skills, our decision to accept or to reject H_0 may still be wrong. That would be a *sampling error* (Section 8.1).

Four situations are possible,

	Result of the test	
	Reject H_0	**Accept** H_0
H_0 **is true**	*Type I error*	correct
H_0 **is false**	correct	*Type II error*

In two of the four cases, the test results in a *correct decision*. Either we accepted a true hypothesis, or we rejected a false hypothesis. The other two situations are sampling errors.

DEFINITION 9.7

A **type I error** occurs when we reject the true null hypothesis.

A **type II error** occurs when we accept the false null hypothesis.

Each error occurs with a certain probability that we hope to keep small. A good test results in an erroneous decision only if the observed data are somewhat extreme.

A type I error is often considered more dangerous and undesired than a type II error. Making a type I error can be compared with convicting an innocent defendant or sending a patient to a surgery when (s)he does not need one.

For this reason, we shall design tests that bound the probability of type I error by a pre-assigned small number α. Under this condition, we may want to minimize the probability of type II error.

DEFINITION 9.8

Probability of a type I error is the **significance level** of a test,

$$\alpha = \boldsymbol{P}\{\text{reject } H_0 \mid H_0 \text{ is true}\}.$$

Probability of rejecting a false hypothesis is the **power of the test**,

$$p(\theta) = \boldsymbol{P}\{\text{reject } H_0 \mid \theta;\ H_A \text{ is true}\}.$$

It is usually a function of the parameter θ because the alternative hypothesis includes a set of parameter values. Also, the power is the probability to avoid a Type II error.

Typically, hypotheses are tested at significance levels as small as 0.01, 0.05, or 0.10, although there are exceptions. Testing at a low level of significance means that only a large amount of evidence can force rejection of H_0. Rejecting a hypothesis at a very low level of significance is done with a lot of confidence that this decision is right.

9.4.3 Level α tests: general approach

A standard algorithm for a level α test of a hypothesis H_0 against an alternative H_A consists of 3 steps.

Step 1. Test statistic

Testing hypothesis is based on a **test statistic** T, a quantity computed from the data that has some known, tabulated distribution F_0 if the hypothesis H_0 is true.

Test statistics are used to discriminate between the hypothesis and the alternative. When

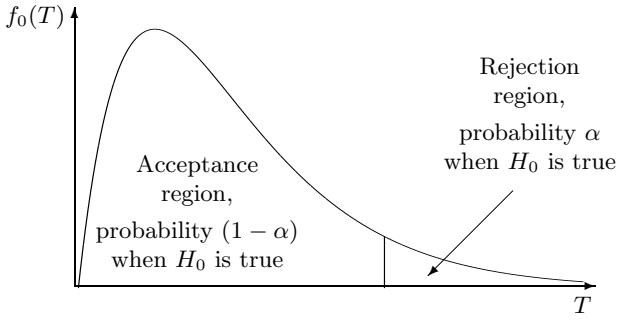


FIGURE 9.6: *Acceptance and rejection regions.*

we verify a hypothesis about some parameter θ, the test statistic is usually obtained by a suitable transformation of its estimator $\widehat{\theta}$.

Step 2. Acceptance region and rejection region

Next, we consider the **null distribution** F_0. This is the distribution of test statistic T when the hypothesis H_0 is true. If it has a density f_0, then the whole area under the density curve is 1, and we can always find a portion of it whose area is α, as shown in Figure 9.6. It is called **rejection region** ($\mathfrak{R}$).

The remaining part, the complement of the rejection region, is called **acceptance region** ($\mathfrak{A} = \overline{\mathfrak{R}}$). By the complement rule, its area is $(1-\alpha)$.

These regions are selected in such a way that the values of test statistic T in the rejection region provide a stronger support of H_A than the values $T \in \mathfrak{A}$. For example, suppose that T is expected to be large if H_A is true. Then the rejection region corresponds to the right tail of the null distribution F_0 (Figure 9.6).

As another example, look at Figure 9.3 on p. 256. If the null distribution of T is *Standard Normal*, then the area between $(-z_{\alpha/2})$ and $z_{\alpha/2}$ equals exactly $(1-\alpha)$. The interval

$$\mathfrak{A} = (-z_{\alpha/2}, z_{\alpha/2})$$

can serve as a level α acceptance region for a two-sided test of $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$. The remaining part consists of two symmetric tails,

$$\mathfrak{R} = \overline{\mathfrak{A}} = (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, +\infty);$$

this is the rejection region.

Areas under the density curve are probabilities, and we conclude that

$$\boldsymbol{P}\{T \in \text{ acceptance region } \mid H_0\} = 1-\alpha$$

and

$$\boldsymbol{P}\{T \in \text{ rejection region } \mid H_0\} = \alpha.$$

Step 3: Result and its interpretation

Accept the hypothesis H_0 if the test statistic T belongs to the acceptance region. Reject H_0 in favor of the alternative H_A if T belongs to the rejection region.

Our acceptance and rejection regions guarantee that the significance level of our test is

$$\begin{aligned} \text{Significance level} &= \boldsymbol{P}\{ \text{ Type I error } \} \\ &= \boldsymbol{P}\{ \text{ Reject } \mid H_0 \} \\ &= \boldsymbol{P}\{T \in \mathfrak{R} \mid H_0\} \\ &= \alpha. \end{aligned} \tag{9.13}$$

Therefore, indeed, we have a level α test!

The interesting part is to interpret our result correctly. Notice that conclusions like *"My level α test accepted the hypothesis. Therefore, the hypothesis is true with probability $(1-\alpha)$"* are *wrong*! Statements H_0 and H_A are about a non-random population, and thus, the hypothesis can either be true with probability 1 or false with probability 1.

If the test rejects the hypothesis, all we can state is that the data provides sufficient evidence against H_0 and in favor of H_A. It may either happen because H_0 is not true, or because our sample is too extreme. The latter, however, can only happen with probability α.

If the test accepts the hypothesis, it only means that the evidence obtained from the data is not sufficient to reject it. In the absence of sufficient evidence, by default, we accept the null hypothesis.

NOTATION

α	=	level of significance, probability of type I error
$p(\theta)$	=	power
T	=	test statistic
F_0, f_0	=	null distribution of T and its density
$\mathfrak{A}$	=	acceptance region
$\mathfrak{R}$	=	rejection region

9.4.4 Rejection regions and power

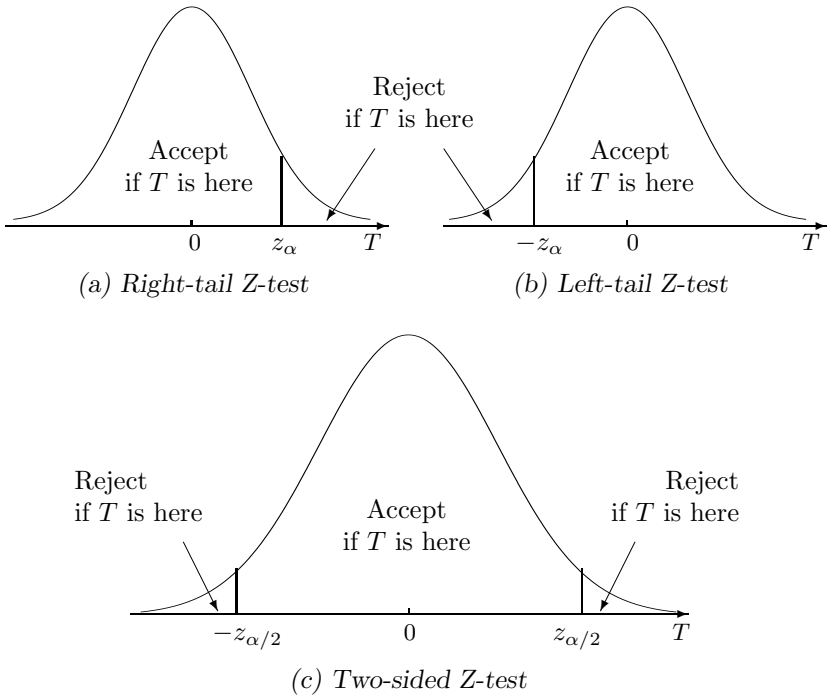
Our construction of the rejection region guaranteed the desired significance level α, as we proved in (9.13). However, one can choose many regions that will also have probability α (see Figure 9.7). Among them, which one is the best choice?

To avoid *type II errors*, we choose such a rejection region that will likely cover the test statistic T in case if the *alternative* H_A is true. This maximizes the *power* of our test because we'll rarely accept H_0 in this case.

Then, we look at our test statistic T under the alternative. Often

(a) a *right-tail alternative* forces T to be large,

(b) a *left-tail alternative* forces T to be small,



(a) Right-tail Z-test *(b) Left-tail Z-test*

(c) Two-sided Z-test

FIGURE 9.7: *Acceptance and rejection regions for a Z-test with (a) a one-sided right-tail alternative; (b) a one-sided left-tail alternative; (c) a two-sided alternative.*

(c) a *two-sided alternative* forces T to be either large or small

(although it certainly depends on how we choose T). If this is the case, it tells us exactly when we should reject the null hypothesis:

(a) For a **right-tail alternative**, the rejection region $\mathfrak{R}$ should consist of large values of T. Choose $\mathfrak{R}$ on the right, $\mathfrak{A}$ on the left (Figure 9.7a).

(b) For a **left-tail alternative**, the rejection region $\mathfrak{R}$ should consist of small values of T. Choose $\mathfrak{R}$ on the left, $\mathfrak{A}$ on the right (Figure 9.7b).

(c) For a **two-sided alternative**, the rejection region $\mathfrak{R}$ should consist of very small and very large values of T. Let $\mathfrak{R}$ consist of two extreme regions, while $\mathfrak{A}$ covers the middle (Figure 9.7c).

9.4.5 Standard Normal null distribution (Z-test)

An important case, in terms of a large number of applications, is when the null distribution of the test statistic is *Standard Normal*.

The test in this case is called a **Z-test**, and the test statistic is usually denoted by Z.

(a) A level α test with a **right-tail alternative** should

$$\begin{cases} \text{reject } H_0 & \text{if} \quad Z \geq z_\alpha \\ \text{accept } H_0 & \text{if} \quad Z < z_\alpha \end{cases} \tag{9.14}$$

The rejection region in this case consists of large values of Z only,

$$\mathfrak{R} = [z_\alpha, +\infty), \qquad \mathfrak{A} = (-\infty, z_\alpha)$$

(see Figure 9.7a).

Under the null hypothesis, Z belongs to $\mathfrak{A}$ and we reject the null hypothesis with probability

$$\boldsymbol{P}\{T \geq z_\alpha \mid H_0\} = 1 - \Phi(z_\alpha) = \alpha,$$

making the probability of false rejection (type I error) equal α .

For example, we use this acceptance region to test the population mean,

$$H_0 : \ \mu = \mu_0 \ \text{ vs } \ H_A : \mu > \mu_0.$$

(b) With a **left-tail alternative**, we should

$$\begin{cases} \text{reject } H_0 & \text{if} \quad Z \leq -z_\alpha \\ \text{accept } H_0 & \text{if} \quad Z > -z_\alpha \end{cases} \tag{9.15}$$

The rejection region consists of small values of Z only,

$$\mathfrak{R} = (-\infty, -z_\alpha], \qquad \mathfrak{A} = (-z_\alpha, +\infty).$$

Similarly, $\boldsymbol{P}\{Z \in \mathfrak{R}\} = \alpha$ under H_0; thus, the probability of type I error equals α.

For example, this is how we should test

$$H_0 : \ \mu = \mu_0 \ \text{ vs } \ H_A : \mu < \mu_0.$$

(c) With a **two-sided alternative**, we

$$\begin{cases} \text{reject } H_0 & \text{if} \quad |Z| \geq z_{\alpha/2} \\ \text{accept } H_0 & \text{if} \quad |Z| < z_{\alpha/2} \end{cases} \tag{9.16}$$

The rejection region consists of very small and very large values of Z,

$$\mathfrak{R} = (-\infty, z_{\alpha/2}] \cup [z_{\alpha/2}, +\infty), \qquad A = (-z_{\alpha/2}, z_{\alpha/2}).$$

Again, the probability of type I error equals α in this case.

For example, we use this test for

$$H_0 : \ \mu = \mu_0 \ \text{ vs } \ H_A : \mu \neq \mu_0.$$

This is easy to remember:

- for a two-sided test, divide α by two and use $z_{\alpha/2}$;
- for a one-sided test, use z_α keeping in mind that the rejection region consists of just one piece.

Now consider testing a hypothesis about a population parameter θ. Suppose that its estimator $\widehat{\theta}$ has Normal distribution, at least approximately, and we know $\mathbf{E}(\widehat{\theta})$ and $\mathrm{Var}(\widehat{\theta})$ if the hypothesis is true.

Then the test statistic

$$Z = \frac{\widehat{\theta} - \mathbf{E}(\widehat{\theta})}{\sqrt{\mathrm{Var}(\widehat{\theta})}} \tag{9.17}$$

has Standard Normal distribution, and we can use (9.14), (9.15), and (9.16) to construct acceptance and rejection regions for a level α test. We call Z a **Z-statistic**.

Examples of Z-tests are in the next section.

9.4.6 Z-tests for means and proportions

As we already know,

- sample means have Normal distribution when the distribution of data is Normal;
- sample means have approximately Normal distribution when they are computed from large samples (the distribution of data can be arbitrary);
- sample proportions have approximately Normal distribution when they are computed from large samples;
- this extends to differences between means and between proportions

(see Sections 8.2.1 and 9.2.2–9.3.2).

For all these cases, we can use a Z-statistic (9.17) and rejection regions (9.14)–(9.16) to design powerful level α tests.

Z-tests are summarized in Table 9.1. You can certainly derive the test statistics without our help; see Exercise 9.6. The last row of Table 9.1 is explained in detail in Section 9.4.7.

Example 9.25 (Z-TEST ABOUT A POPULATION MEAN). The number of concurrent users for some internet service provider has always averaged 5000 with a standard deviation of 800. After an equipment upgrade, the average number of users at 100 randomly selected moments of time is 5200. Does it indicate, at a 5% level of significance, that the mean number of concurrent users has increased? Assume that the standard deviation of the number of concurrent users has not changed.

Solution. We test the null hypothesis $H_0 : \mu = 5000$ against a *one-sided right-tail alternative* $H_A : \mu > 5000$, because we are only interested to know if the mean number of users μ has increased.

Step 1: Test statistic. We are given: $\sigma = 800$, $n = 100$, $\alpha = 0.05$, $\mu_0 = 5000$, and from the sample, $\overline{X} = 5200$. The test statistic is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{5200 - 5000}{800/\sqrt{100}} = 2.5.$$

Null hypothesis	Parameter, estimator	If H_0 is true:		Test statistic
H_0	θ, $\widehat{\theta}$	$\mathbf{E}(\widehat{\theta})$	$\mathrm{Var}(\widehat{\theta})$	$Z = \dfrac{\widehat{\theta} - \theta_0}{\sqrt{\mathrm{Var}(\widehat{\theta})}}$
One-sample Z-tests for means and proportions, based on a sample of size n				
$\mu = \mu_0$	μ, $\overline{X}$	μ_0	$\dfrac{\sigma^2}{n}$	$\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$
$p = p_0$	p, $\widehat{p}$	p_0	$\dfrac{p_0(1-p_0)}{n}$	$\dfrac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
Two-sample Z-tests comparing means and proportions of two populations, based on independent samples of size n and m				
$\mu_X - \mu_Y = D$	$\mu_X - \mu_Y$, $\overline{X} - \overline{Y}$	D	$\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}$	$\dfrac{\overline{X} - \overline{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$
$p_1 - p_2 = D$	$p_1 - p_2$, $\widehat{p}_1 - \widehat{p}_2$	D	$\dfrac{p_1(1-p_1)}{n} + \dfrac{p_2(1-p_2)}{m}$	$\dfrac{\widehat{p}_1 - \widehat{p}_2 - D}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}}$
$p_1 = p_2$	$p_1 - p_2$, $\widehat{p}_1 - \widehat{p}_2$	0	$p(1-p)\left(\dfrac{1}{n} + \dfrac{1}{m}\right)$, where $p = p_1 = p_2$	$\dfrac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}$ where $\widehat{p} = \dfrac{n\widehat{p}_1 + m\widehat{p}_2}{n+m}$

TABLE 9.1: *Summary of Z-tests.*

Step 2: Acceptance and rejection regions. The critical value is

$$z_\alpha = z_{0.05} = 1.645$$

(don't divide α by 2 because it is a one-sided test). With the right-tail alternative, we

$$\begin{cases} \text{reject } H_0 & \text{if} \quad Z \geq 1.645 \\ \text{accept } H_0 & \text{if} \quad Z < 1.645 \end{cases}$$

Step 3: Result. Our test statistic $Z = 2.5$ belongs to the *rejection region*; therefore, we *reject the null hypothesis*. The data (5200 users, on the average, at 100 times) provided sufficient evidence in favor of the alternative hypothesis that the mean number of users has increased. $\Diamond$

Example 9.26 (Two-sample Z-test of proportions). A quality inspector finds 10 defective parts in a sample of 500 parts received from manufacturer A. Out of 400 parts from manufacturer B, she finds 12 defective ones. A computer-making company uses these parts in their computers and claims that the quality of parts produced by A and B is the same. At the 5% level of significance, do we have enough evidence to disprove this claim?

Solution. We test $H_0 : p_A = p_B$, or $H_0 : p_A - p_B = 0$, against $H_A : p_A \neq p_B$. This is a two-sided test because no direction of the alternative has been indicated. We only need to verify whether or not the proportions of defective parts are equal for manufacturers A and B.

Step 1: Test statistic. We are given: $\widehat{p}_A = 10/500 = 0.02$ from a sample of size $n = 500$; $\widehat{p}_B = 12/400 = 0.03$ from a sample of size $m = 400$. The tested value is $D = 0$.

As we know, for these Bernoulli data, the variance depends on the unknown parameters p_A and p_B which are estimated by the sample proportions $\widehat{p}_A$ and $\widehat{p}_B$.

The test statistic then equals

$$Z = \frac{\widehat{p}_A - \widehat{p}_B - D}{\sqrt{\dfrac{\widehat{p}_A(1-\widehat{p}_A)}{n} + \dfrac{\widehat{p}_B(1-\widehat{p}_B)}{m}}} = \frac{0.02 - 0.03}{\sqrt{\dfrac{(0.02)(0.98)}{500} + \dfrac{(0.03)(0.97)}{400}}} = -0.945.$$

Step 2: Acceptance and rejection regions. This is a two-sided test; thus we divide α by 2, find $z_{0.05/2} = z_{0.025} = 1.96$, and

$$\begin{cases} \text{reject } H_0 & \text{if} \quad |Z| \geq 1.96; \\ \text{accept } H_0 & \text{if} \quad |Z| < 1.96. \end{cases}$$

Step 3: Result. The evidence against H_0 is insufficient because $|Z| < 1.96$. Although *sample proportions* of defective parts are unequal, the difference between them appears too small to claim that *population proportions* are different. ◊

9.4.7 Pooled sample proportion

The test in Example 9.26 can be conducted differently and perhaps, more efficiently.

Indeed, we standardize the estimator $\widehat{\theta} = \widehat{p}_A - \widehat{p}_B$ using its expectation $\mathbf{E}(\widehat{\theta})$ and variance $\mathrm{Var}(\widehat{\theta})$ under the null distribution, i.e., when H_0 is true. However, under the null hypothesis $p_A = p_B$. Then, when we standardize $(\widehat{p}_A - \widehat{p}_B)$, instead of estimating two proportions in the denominator, we only need to estimate one.

First, we estimate the common population proportion by the overall proportion of defective parts,

$$\widehat{p}(\text{pooled}) = \frac{\text{number of defective parts}}{\text{total number of parts}} = \frac{n\widehat{p}_A + m\widehat{p}_B}{n+m}.$$

Then we estimate the common variance as

$$\widehat{\mathrm{Var}}(\widehat{p}_A - \widehat{p}_B) = \frac{\widehat{p}(1-\widehat{p})}{n} + \frac{\widehat{p}(1-\widehat{p})}{m} = \widehat{p}(1-\widehat{p})\left(\frac{1}{n} + \frac{1}{m}\right)$$

and use it for the Z-statistic,

$$Z = \frac{\widehat{p}_A - \widehat{p}_B}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}}.$$

Example 9.27 (EXAMPLE 9.26, CONTINUED). Here the pooled proportion equals

$$\widehat{p} = \frac{10+12}{500+400} = 0.0244,$$

so that

$$Z = \frac{0.02 - 0.03}{\sqrt{(0.0244)(0.9756)\left(\frac{1}{500}+\frac{1}{400}\right)}} = -0.966.$$

This does not affect our result. We obtained a different value of Z-statistic, but it also belongs to the acceptance region. We still don't have a significant evidence against the equality of two population proportions. ◊

9.4.8 Unknown σ: T-tests

As we decided in Section 9.3, when we don't know the population standard deviation, we estimate it. The resulting *T-statistic* has the form

$$t = \frac{\widehat{\theta} - \mathbf{E}(\widehat{\theta})}{s(\widehat{\theta})} = \frac{\widehat{\theta} - \mathbf{E}(\widehat{\theta})}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta})}}.$$

In the case *when the distribution of $\widehat{\theta}$ is Normal*, the test is based on *Student's T-distribution* with acceptance and rejection regions according to the direction of H_A:

(a) For a **right-tail alternative**,

$$\begin{cases} \text{reject } H_0 & \text{if} \quad t \geq t_\alpha \\ \text{accept } H_0 & \text{if} \quad t < t_\alpha \end{cases} \tag{9.18}$$

(b) For a **left-tail alternative**,

$$\begin{cases} \text{reject } H_0 & \text{if} \quad t \leq -t_\alpha \\ \text{accept } H_0 & \text{if} \quad t > -t_\alpha \end{cases} \tag{9.19}$$

(c) For a **two-sided alternative**,

$$\begin{cases} \text{reject } H_0 & \text{if} \quad |t| \geq t_{\alpha/2} \\ \text{accept } H_0 & \text{if} \quad |t| < t_{\alpha/2} \end{cases} \tag{9.20}$$

Quantiles t_α and $t_{\alpha/2}$ are given in Table A5. As in Section 9.3.4, the number of degrees of freedom depends on the problem and the sample size, see Table 9.2 and formula (9.10).

Hypothesis H_0	Conditions	Test statistic t	Degrees of freedom
$\mu = \mu_0$	Sample size n; unknown σ	$t = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$	$n - 1$
$\mu_X - \mu_Y = D$	Sample sizes n, m; unknown but equal standard deviations, $\sigma_X = \sigma_Y$	$t = \dfrac{\overline{X} - \overline{Y} - D}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$	$n + m - 2$
$\mu_X - \mu_Y = D$	Sample sizes n, m; unknown, unequal standard deviations, $\sigma_X \neq \sigma_Y$	$t = \dfrac{\overline{X} - \overline{Y} - D}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$	Satterthwaite approximation, formula (9.12)

TABLE 9.2: *Summary of T-tests.*

As in Section 9.3.4, the **pooled sample variance**

$$s_p^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2}{n + m - 2} = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n + m - 2}$$

is computed for the case of equal unknown variances. When variances are not equal, degrees of freedom are computed by Satterthwaite approximation (9.12).

Example 9.28 (Unauthorized use of a computer account, continued). A long-time authorized user of the account makes 0.2 seconds between keystrokes. One day, the data in Example 9.19 on p. 267 (data set `Keystrokes`) are recorded as someone typed the correct username and password. At a 5% level of significance, is this an evidence of an unauthorized attempt?

Let us test

$$H_0 : \mu = 0.2 \quad \text{vs} \quad H_A : \mu \neq 0.2$$

at a significance level $\alpha = 0.01$. From Example 9.19, we have sample statistics $n = 18$, $\overline{X} = 0.29$ and $s = 0.074$. Compute the T-statistic,

$$t = \frac{\overline{X} - 0.2}{s/\sqrt{n}} = \frac{0.29 - 0.2}{0.074/\sqrt{18}} = 5.16.$$

The rejection region is $\mathfrak{R} = (-\infty, -2.11] \cup [2.11, \infty)$, where we used T-distribution with $18 - 1 = 17$ degrees of freedom and $\alpha/2 = 0.025$ because of the two-sided alternative.

Since $t \in \mathfrak{R}$, we reject the null hypothesis and conclude that *there is a significant evidence of an unauthorized use of that account.* ◊

Example 9.29 (CD WRITER AND BATTERY LIFE). Does a CD writer consume extra energy, and therefore, does it reduce the battery life on a laptop?

Example 9.20 on p. 269 provides data on battery lives for laptops with a CD writer (sample $\boldsymbol{X}$) and without a CD writer (sample $\boldsymbol{Y}$):

$$n = 12,\ \overline{X} = 4.8,\ s_X = 1.6;\ m = 18,\ \overline{Y} = 5.3,\ s_Y = 1.4;\ s_p = 1.4818.$$

Testing

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_A : \mu_X < \mu_Y$$

at $\alpha = 0.05$, we obtain

$$t = \frac{\overline{X} - \overline{Y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{4.8 - 5.3}{(1.4818)\sqrt{\frac{1}{18} + \frac{1}{12}}} = -0.9054.$$

The rejection region for this left-tail test is $(-\infty, -z_\alpha] = (-\infty, -1.645]$. Since $t \notin \mathfrak{R}$, we accept H_0 concluding that there is *no evidence that laptops with a CD writer have a shorter battery life*. ◊

Example 9.30 (COMPARISON OF TWO SERVERS, CONTINUED). Is server A faster in Example 9.21 on p. 271? Formulate and test the hypothesis at a level $\alpha = 0.05$.

Solution. To see if server A is faster, we need to test

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_A : \mu_X < \mu_Y.$$

This is the case of unknown, unequal standard deviations. In Example 9.21, we used Satterthwaite approximation for the number of degrees of freedom and obtained $\nu = 25.4$. We should reject the null hypothesis if $t \le -1.708$. Since

$$t = \frac{6.7 - 7.5}{\sqrt{\frac{(0.6)^2}{30} + \frac{(1.2)^2}{20}}} = -2.7603 \in \mathfrak{R},$$

we reject H_0 and conclude that there is evidence that *server A is faster*. ◊

When the distribution of $\widehat{\theta}$ is not Normal, the Student's *T-distribution* cannot be used. The distribution of a T-statistic and all its probabilities will be different from Student's T, and as a result, our test may not have the desired significance level.

9.4.9 Duality: two-sided tests and two-sided confidence intervals

An interesting fact can be discovered if we look into our derivation of tests and confidence intervals. It turns out that we can conduct two-sided tests using nothing but the confidence intervals!

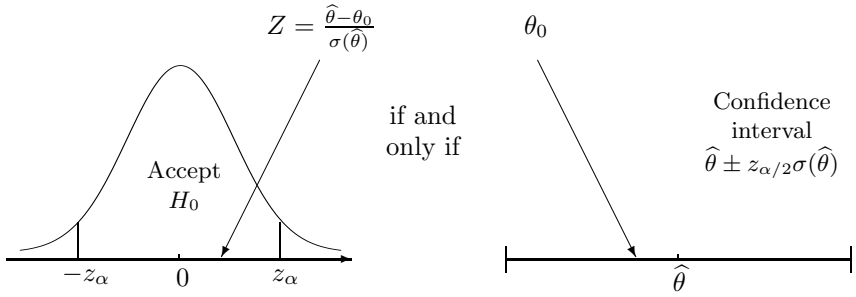


FIGURE 9.8: *Duality of tests and confidence intervals.*

A level α Z-test of $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$
accepts the null hypothesis

if and only if

a symmetric $(1-\alpha)100\%$ confidence Z-interval for θ contains θ_0. (9.21)

PROOF: The null hypothesis H_0 is accepted if and only if the Z-statistic belongs to the acceptance region, i.e.,

$$\left| \frac{\widehat{\theta} - \theta_0}{\sigma(\widehat{\theta})} \right| \leq z_{\alpha/2}.$$

This is equivalent to

$$\left| \widehat{\theta} - \theta_0 \right| \leq z_{\alpha/2}\sigma(\widehat{\theta}).$$

We see that the distance from θ_0 to the center of Z-interval $\widehat{\theta}$ does not exceed its margin, $z_{\alpha/2}\sigma(\widehat{\theta})$ (see (9.3) and Figure 9.8). In other words, θ_0 belongs to the Z-interval. □

In fact, any two-sided test can be conducted this way. Accept $H_0 : \theta = \theta_0$ whenever a $(1-\alpha)100\%$ confidence interval for θ covers θ_0. Under θ_0, this test will accept the null hypothesis as often as the interval will cover θ_0, i.e., with probability $(1-\alpha)$. Thus, we have a level α test.

Rule (9.21) applies *only* when

- we are testing against a two-sided alternative (notice that our confidence intervals are two-sided too);
- significance level α of the test matches confidence level $(1-\alpha)$ of the confidence interval. For example, a two-sided 3% level test can be conducted using a 97% confidence interval.

Example 9.31. A sample of 6 measurements

$$2.5, 7.4, 8.0, 4.5, 7.4, 9.2$$

is collected from a Normal distribution with mean μ and standard deviation $\sigma = 2.2$. Test whether $\mu = 6$ against a two-sided alternative $H_A : \mu \neq 6$ at the 5% level of significance.

Solution. Solving Example 9.13 on p. 258, we have already constructed a 95% confidence interval for μ,

$$[4.74,\ 8.26].$$

The value of $\mu_0 = 6$ belongs to it; therefore, at the 5% level, the null hypothesis is accepted. ◊

Example 9.32. Use data in Example 9.31 to test whether $\mu = 7$.

Solution. The interval $[4.74,\ 8.26]$ contains $\mu_0 = 7$ too; therefore, the hypothesis $H_0 : \mu = 7$ is accepted as well. ◊

In the last two examples, how could we possibly accept both hypotheses, $\mu = 6$ and $\mu = 7$? Obviously, μ cannot be equal 6 and 7 at the same time! This is true. By accepting both null hypotheses, we only acknowledge that sufficient evidence against either of them is not found in the given data.

Example 9.33 (Pre-election poll). In Example 9.17 on p. 265, we computed a 95% confidence interval for the difference of proportions supporting a candidate in towns A and B: $[-0.14,\ 0.16]$. This interval contains 0, therefore, the test of

$$H_0 : p_1 = p_2 \quad \text{vs} \quad H_A : p_1 \neq p_2$$

accepts the null hypothesis at the 5% level. Apparently, there is no evidence of unequal support of this candidate in the two towns. ◊

Example 9.34 (Hardware upgrade). In Example 9.14, we studied effectiveness of the hardware upgrade. We constructed a 90% confidence interval for the difference $(\mu_X - \mu_Y)$ in mean running times of a certain process: $[0.7, 1.9]$.

So, can we conclude that the upgrade was successful? Ineffective upgrade corresponds to a null hypothesis $H_0 : \ \mu_X = \mu_Y$, or $\mu_X - \mu_Y = 0$. Since the interval $[0.7, 1.9]$ does not contain 0, the no-effect hypothesis should be rejected at the 10% level of significance. ◊

Example 9.35 (Was the upgrade successful? One-sided test). Let's look at Example 9.34 again. On second thought, we can only use Rule (9.21) to test the **two-sided alternative** $H_A : \mu_X \neq \mu_Y$, right? At the same time, the hardware upgrade is successful only when the running time reduces, i.e., $\mu_X > \mu_Y$. Then, we should judge effectiveness of the upgrade by a **one-sided, right-tail test** of

$$H_0 : \ \mu_X = \mu_Y \quad \text{vs} \quad H_A : \mu_X > \mu_Y. \tag{9.22}$$

Let us try to use the interval $[0.7, 1.9]$ for this test too. The null hypothesis in Example 9.34 is rejected at the 10% level in favor of a two-sided alternative, thus

$$|Z| > z_{\alpha/2} = z_{0.05}.$$

Then, either $Z < -z_{0.05}$ or $Z > z_{0.05}$. The first case is ruled out because the interval $[0.7, 1.9]$ consists of positive numbers, hence it cannot possibly support a left-tail alternative.

We conclude that $Z > z_{0.05}$, hence the test (9.22) results in rejection of H_0 at the 5% level of significance.

Conclusion. Our 90% confidence interval for $(\mu_X - \mu_Y)$ shows significant evidence, at the 5% level of significance, that the hardware upgrade was successful. ◊

Similarly, for the case of unknown variance(s).

> A level α T-test of $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$
> accepts the null hypothesis
>
> if and only if
>
> a symmetric $(1 - \alpha)100\%$ confidence T-interval for θ contains θ_0.

Example 9.36 (Unauthorized use of a computer account, continued). A 99% confidence interval for the mean time between keystrokes is

$$[0.24;\ 0.34]$$

(Example 9.19 on p. 267 and data set `Keystrokes`). Example 9.28 on p. 283 tests whether the mean time is 0.2 seconds, which would be consistent with the speed of the account owner. The interval does not contain 0.2. Therefore, at a 1% level of significance, we have significant evidence that *the account was used by a different person.* ◊

9.4.10 P-value

How do we choose α?

So far, we were testing hypotheses by means of acceptance and rejection regions. In the last section, we learned how to use confidence intervals for two-sided tests. Either way, we need to know the *significance level* α in order to conduct a test. Results of our test depend on it.

How do we choose α, the probability of making type I sampling error, rejecting the true

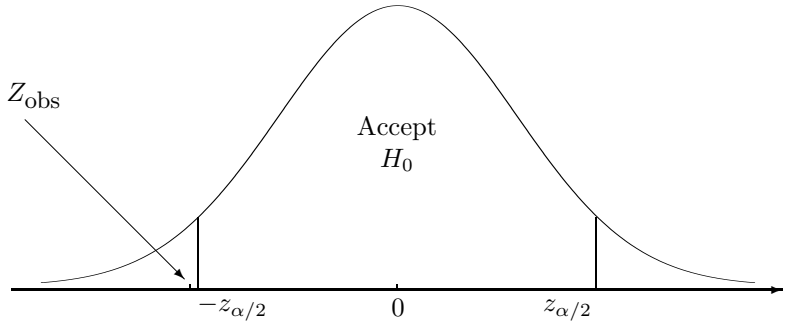


FIGURE 9.9: *This test is "too close to call": formally we reject the null hypothesis although the Z-statistic is almost at the boundary.*

hypothesis? Of course, when it seems too dangerous to reject true H_0, we choose a low significance level. How low? Should we choose $\alpha = 0.01$? Perhaps, 0.001? Or even 0.0001?

Also, if our *observed* test statistic $Z = Z_{\text{obs}}$ belongs to a rejection region but it is "too close to call" (see, for example, Figure 9.9), then how do we report the result? Formally, we should reject the null hypothesis, but practically, we realize that a slightly different significance level α could have expanded the acceptance region just enough to cover Z_{obs} and force us to accept H_0.

Suppose that the result of our test is crucially important. For example, the choice of a business strategy for the next ten years depends on it. In this case, can we rely so heavily on the choice of α? And if we rejected the true hypothesis just because we chose $\alpha = 0.05$ instead of $\alpha = 0.04$, then how do we explain to the chief executive officer that the situation was marginal? What is the statistical term for "too close to call"?

P-value

Using a P-value approach, we try not to rely on the level of significance. In fact, let us try to test a hypothesis using *all levels of significance*!

Considering all levels of significance (between 0 and 1 because α is a probability of Type I error), we notice:

Case 1. If a level of significance is *very low*, we *accept* the null hypothesis (see Figure 9.10a). A low value of

$$\alpha = \boldsymbol{P}\{ \text{ reject the null hypothesis when it is true } \}$$

makes it very unlikely to reject the hypothesis because it yields a very small rejection region. The right-tail area above the rejection region equals α.

Case 2. On the other extreme end, a *high significance level* α makes it likely to reject the null hypothesis and corresponds to a large rejection region. A sufficiently large α will produce such a large rejection region that will cover our test statistic, forcing us to *reject* H_0 (see Figure 9.10b).

Conclusion: there exists a boundary value between α-to-accept (case 1) and α-to-reject (case 2). This number is a *P-value* (Figure 9.11).

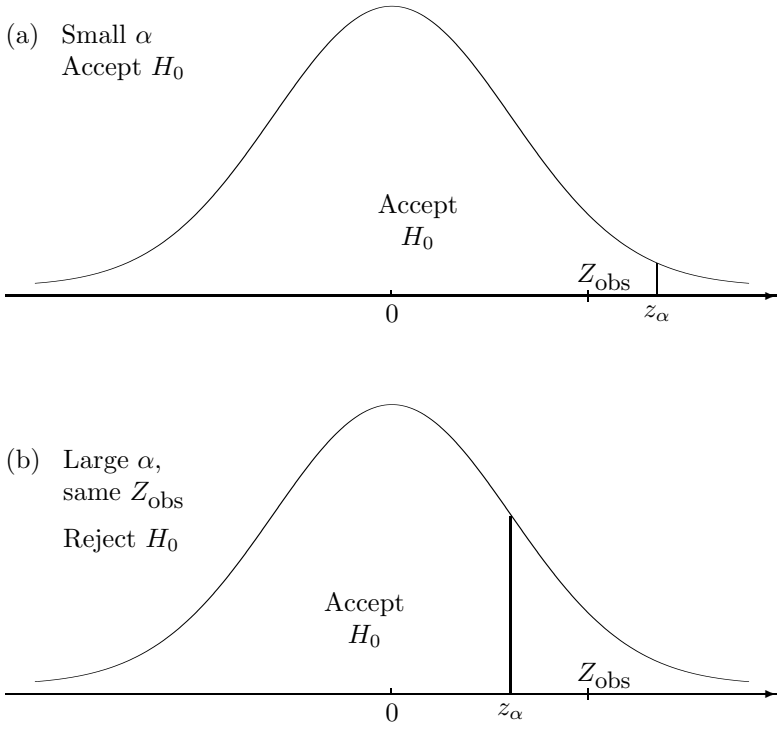


FIGURE 9.10: *(a) Under a low level of significance α, we accept the null hypothesis. (b) Under a high level of significance, we reject it.*

DEFINITION 9.9

P-value is the lowest significance level α that forces rejection of the null hypothesis.

P-value is also the highest significance level α that forces acceptance of the null hypothesis.

Testing hypotheses with a P-value

Once we know a P-value, we can indeed test hypotheses at *all* significance levels. Figure 9.11 clearly shows that for all $\alpha < P$ we accept the null hypothesis, and for all $\alpha > P$, we reject it.

Usual significance levels α lie in the interval [0.01, 0.1] (although there are exceptions). Then, a P-value greater than 0.1 exceeds all natural significance levels, and the null hypothesis should be accepted. Conversely, if a P-value is less than 0.01, then it is smaller than all natural significance levels, and the null hypothesis should be rejected. Notice that we did not even have to specify the level α for these tests!

Only if the P-value happens to fall between 0.01 and 0.1, we really have to think about the level of significance. This is the "marginal case," "too close to call." When we report

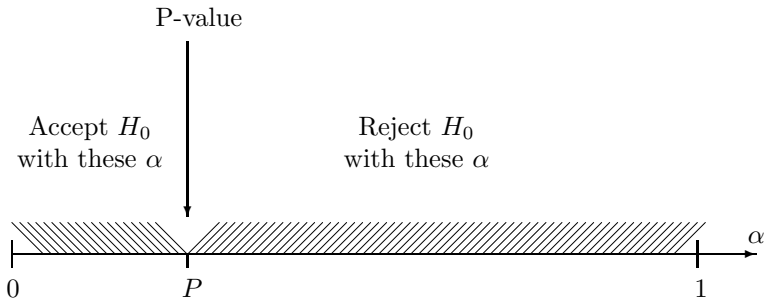


FIGURE 9.11: *P-value separates α-to-accept and α-to-reject.*

the conclusion, accepting or rejecting the hypothesis, we should always remember that with a slightly different α, the decision could have been reverted. When the matter is crucially important, a good decision is to collect more data until a more definitive answer can be obtained.

Testing H_0 with a P-value

For	$\alpha < P$,	accept H_0
For	$\alpha > P$,	reject H_0
Practically,		
If	$P < 0.01$,	reject H_0
If	$P > 0.1$,	accept H_0

Computing P-values

Here is how a P-value can be computed from data.

Let us look at Figure 9.10 again. Start from Figure 9.10a, gradually increase α, and keep your eye at the vertical bar separating the acceptance and rejection region. It will move to the left until it hits the observed test statistic Z_{obs}. At this point, our decision changes, and we switch from case 1 (Figure 9.10a) to case 2 (Figure 9.10b). Increasing α further, we pass the Z-statistic and start accepting the null hypothesis.

What happens at the border of α-to-accept and α-to-reject? Definition 9.9 says that this borderline α is the **P-value**,

$$P = \alpha.$$

Also, at this border our observed Z-statistic coincides with the critical value z_α,

$$Z_{\text{obs}} = z_\alpha,$$

and thus,

$$P = \alpha = \boldsymbol{P}\{Z \geq z_\alpha\} = \boldsymbol{P}\{Z \geq Z_{\text{obs}}\}.$$

In this formula, Z is any Standard Normal random variable, and Z_{obs} is our observed test

statistic, which is a concrete number, computed from data. First, we compute Z_{obs}, then use Table A4 to calculate

$$\boldsymbol{P}\{Z \geq Z_{\text{obs}}\} = 1 - \Phi(Z_{\text{obs}}).$$

P-values for the left-tail and for the two-sided alternatives are computed similarly, as given in Table 9.3.

This table applies to all the Z-tests in this chapter. It can be directly extended to the case of unknown standard deviations and T-tests (Table 9.4).

Understanding P-values

Looking at Tables 9.3 and 9.4, we see that *P-value* is the probability of observing a test statistic *at least as extreme as* Z_{obs} or t_{obs}. Being "extreme" is determined by the alternative. For a right-tail alternative, large numbers are extreme; for a left-tail alternative, small numbers are extreme; and for a two-sided alternative, both large and small numbers are extreme. In general, the more extreme test statistic we observe, the stronger support of the alternative it provides.

This creates another interesting definition of a P-value.

DEFINITION 9.10

P-value is the probability of observing a test statistic that is as extreme as or more extreme than the test statistic computed from a given sample.

The following philosophy can be used when we test hypotheses by means of a P-value.

We are deciding between the null hypothesis H_0 and the alternative H_A. Observed is a test statistic Z_{obs}. If H_0 were true, how likely would it be to observe such a statistic? In other words, are the observed data consistent with H_0?

A high P-value tells that this or even more extreme value of Z_{obs} is quite possible under H_0, and therefore, we see no contradiction with H_0. The null hypothesis is not rejected.

Conversely, a low P-value signals that such an extreme test statistic is unlikely if H_0 is true.

Hypothesis H_0	Alternative H_A	P-value	Computation
$\theta = \theta_0$	right-tail $\theta > \theta_0$	$\boldsymbol{P}\{Z \geq Z_{\text{obs}}\}$	$1 - \Phi(Z_{\text{obs}})$
	left-tail $\theta < \theta_0$	$\boldsymbol{P}\{Z \leq Z_{\text{obs}}\}$	$\Phi(Z_{\text{obs}})$
	two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\lvert Z\rvert \geq \lvert Z_{\text{obs}}\rvert\}$	$2(1 - \Phi(\lvert Z_{\text{obs}}\rvert))$

TABLE 9.3: *P-values for Z-tests.*

Hypothesis H_0	Alternative H_A	P-value	Computation
$\theta = \theta_0$	right-tail $\theta > \theta_0$	$\boldsymbol{P}\{t \geq t_{\text{obs}}\}$	$1 - F_\nu(t_{\text{obs}})$
	left-tail $\theta < \theta_0$	$\boldsymbol{P}\{t \leq t_{\text{obs}}\}$	$F_\nu(t_{\text{obs}})$
	two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\|t\| \geq \|t_{\text{obs}}\|\}$	$2(1 - F_\nu(\|t_{\text{obs}}\|))$

TABLE 9.4: *P-values for T-tests (F_ν is the cdf of T-distribution with the suitable number ν of degrees of freedom).*

However, we really observed it. Then, our data are not consistent with the hypothesis, and we should reject H_0.

For example, if $P = 0.0001$, there is only 1 chance in 10,000 to observe what we really observed. The evidence supporting the alternative is highly significant in this case.

Example 9.37 (How significant was the upgrade?). Refer to Examples 9.14 and 9.34. At the 5% level of significance, we know that the hardware upgrade was successful. Was it marginally successful or very highly successful? Let us compute the P-value.

Start with computing a Z-statistic,

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{8.5 - 7.2}{\sqrt{\frac{1.8^2}{50} + \frac{1.8^2}{50}}} = 3.61.$$

From Table A4, we find that the P-value for the right-tail alternative is

$$P = \boldsymbol{P}\{Z \geq Z_{\text{obs}}\} = \boldsymbol{P}\{Z \geq 3.61\} = 1 - \Phi(3.61) = 0.0002.$$

The P-value is very low; therefore, we can reject the null hypothesis not only at the 5%, but also at the 1% and even 0.05% level of significance! We see now that the hardware upgrade was extremely successful. ◊

Example 9.38 (Quality inspection). In Example 9.26, we compared the quality of parts produced by two manufacturers by a two-sided test. We obtained a test statistic

$$Z_{\text{obs}} = -0.94.$$

The P-value for this test equals

$$P = \boldsymbol{P}\{|Z| \geq |-0.94|\} = 2(1 - \Phi(0.94)) = 2(1 - 0.8264) = 0.3472.$$

This is a rather high P-value (greater than 0.1), and the null hypothesis is not rejected. Given H_0, there is a 34% chance of observing what we really observed. No contradiction with H_0, and therefore, no evidence that the quality of parts is not the same. ◊

Table A5 is not as detailed as Table A4. Often we can only use it to bound the P-value from below and from above. Typically, it suffices for hypothesis testing.

Example 9.39 (Unauthorized use of a computer account, continued). How significant is the evidence in Examples 9.28 and 9.36 on pp. 283, 287 that the account was used by an unauthorized person?

Under the null hypothesis, our T-statistic has T-distribution with 17 degrees of freedom. In the previous examples, we rejected H_0 first at the 5% level, then at the 1% level. Now, comparing $t = 5.16$ from Example 9.28 with the entire row 17 of Table A5, we find that it exceeds all the critical values given in the table until $t_{0.0001}$. Therefore, a two-sided test rejects the null hypothesis at a very low level $\alpha = 0.0002$, and the P-value is $P < 0.0002$. *The evidence of an unauthorized use is very strong!*

$\Diamond$

9.5 Inference about variances

In this section, we'll derive confidence intervals and tests for the population variance $\sigma^2 = \text{Var}(X)$ and for the comparison of two variances $\sigma_X^2 = \text{Var}(X)$ and $\sigma_Y^2 = \text{Var}(Y)$. This will be *a new type of inference* for us because
(a) variance is a scale and not a location parameter,
(b) the distribution of its estimator, the sample variance, is not symmetric.

Variance often needs to be estimated or tested for quality control, in order to assess stability and accuracy, evaluate various risks, and also, for tests and confidence intervals for the population means when variance is unknown.

Recall that comparing two means in Section 9.3.5, we had to distinguish between the cases of equal and unequal variances. We no longer have to guess! In this section, we'll see how to test the null hypothesis $H_0 : \sigma_X^2 = \sigma_Y^2$ against the alternative $H_A : \sigma_X^2 \neq \sigma_Y^2$ and decide whether we should use the pooled variance (9.11) or the Satterthwaite approximation (9.12).

9.5.1 Variance estimator and Chi-square distribution

We start by estimating the population variance $\sigma^2 = \text{Var}(X)$ from an observed sample $\boldsymbol{X} = (X_1, \dots, X_n)$. Recall from Section 8.2.4 that σ^2 is estimated *unbiasedly* and *consistently* by the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 .$$

The summands $\left(X_i - \overline{X}\right)^2$ are not quite independent, as the Central Limit Theorem on p. 93 requires, because they all depend on $\overline{X}$. Nevertheless, the distribution of s^2 is approximately Normal, under mild conditions, when the sample is large.

For small to moderate samples, the distribution of s^2 is not Normal at all. It is not even symmetric. Indeed, why should it be symmetric if s^2 is always non-negative!

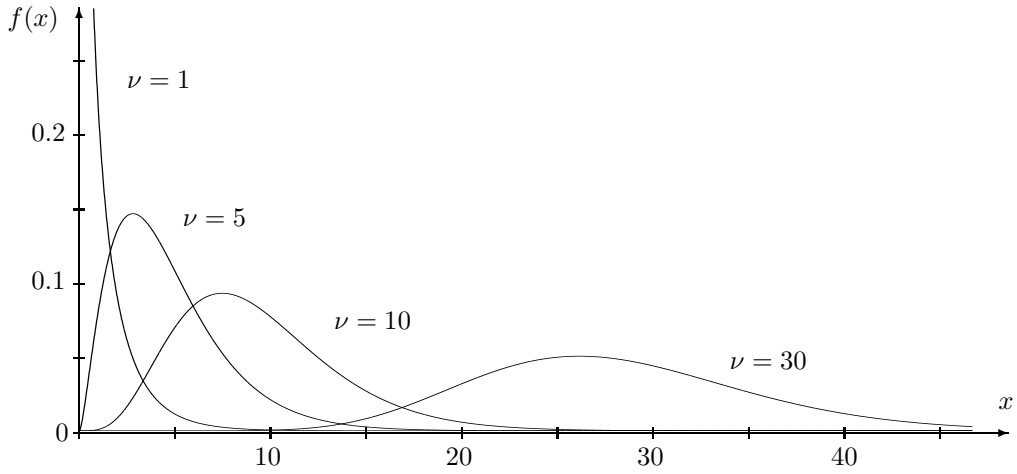


FIGURE 9.12: *Chi-square densities with ν = 1, 5, 10, and 30 degrees of freedom. Each distribution is right-skewed. For large ν, it is approximately Normal.*

Distribution of the sample variance

When observations $X_1, \dots, X_n$ are independent and Normal with $\text{Var}(X_i) = \sigma^2$, the distribution of

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2$$

is *Chi-square with $(n-1)$ degrees of freedom*

Chi-square distribution, or χ^2, is a continuous distribution with density

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\nu/2-1} e^{-x/2}, \quad x > 0,$$

where $\nu > 0$ is a parameter that is called *degrees of freedom* and has the same meaning as for the Student's T-distribution (Figure 9.12).

Comparing this density with (4.7) on p. 85, we see that Chi-square distribution is a special case of Gamma,

$$\text{Chi-square}(\nu) = \text{Gamma}(\nu/2, 1/2),$$

and in particular, the Chi-square distribution with $\nu = 2$ degrees of freedom is Exponential(1/2).

We already know that Gamma(α, λ) distribution has expectation $\mathbf{E}(X) = \alpha/\lambda$ and $\text{Var}(X) = \alpha/\lambda^2$. Substituting $\alpha = \nu/2$ and $\lambda = 1/2$, we get the Chi-square moments,

$$\mathbf{E}(X) = \nu \quad \text{and} \quad \text{Var}(X) = 2\nu.$$

Chi-square distribution (χ^2)

$$\boxed{\begin{array}{lcl} \nu & = & \text{degrees of freedom} \\ f(x) & = & \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\nu/2-1}e^{-x/2}, \quad x>0 \\ \mathbf{E}(X) & = & \nu \\ \mathrm{Var}(X) & = & 2\nu \end{array}} \tag{9.23}$$

Chi-square distribution was introduced around 1900 by a famous English mathematician *Karl Pearson* (1857-1936) who is regarded as a founder of the entire field of *Mathematical Statistics*. By the way, Pearson was a teacher and collaborator of William Gosset, which is why Student was Gosset's pseudonym.

Table A6 in the Appendix contains critical values of the Chi-square distribution.

9.5.2 Confidence interval for the population variance

Let us construct a $(1-\alpha)100\%$ confidence interval for the population variance σ^2, based on a sample of size n.

As always, we start with the estimator, the sample variance s^2. However, since the distribution of s^2 is not symmetric, our confidence interval won't have the form "estimator $\pm$ margin" as before.

Instead, we use Table A6 to find *the critical values* $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ of the Chi-square distribution with $\nu = n-1$ degrees of freedom. These critical values chop the areas of $(\alpha/2)$ on the right and on the left sides of the region under the Chi-square density curve, as on Figure 9.13. This is similar to $\pm z_{\alpha/2}$ and $\pm t_{\alpha/2}$ in the previous sections, although these Chi-square quantiles are no longer symmetric. Recall that $\chi^2_{\alpha/2}$ denotes the $(1-\alpha/2)$-quantile, $q_{1-\alpha/2}$.

Then, the area between these two values is $(1-\alpha)$.

A rescaled sample variance $(n-1)s^2/\sigma^2$ has χ^2 density like the one on Figure 9.13, so

$$\boldsymbol{P}\left\{\chi^2_{1-\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2}\right\} = 1-\alpha.$$

Solving the inequality for the unknown parameter σ^2, we get

$$\boldsymbol{P}\left\{\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right\} = 1-\alpha.$$

A $(1-\alpha)100\%$ confidence interval for the population variance is obtained!

Confidence interval for the variance

$$\boxed{\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}},\ \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]} \tag{9.24}$$

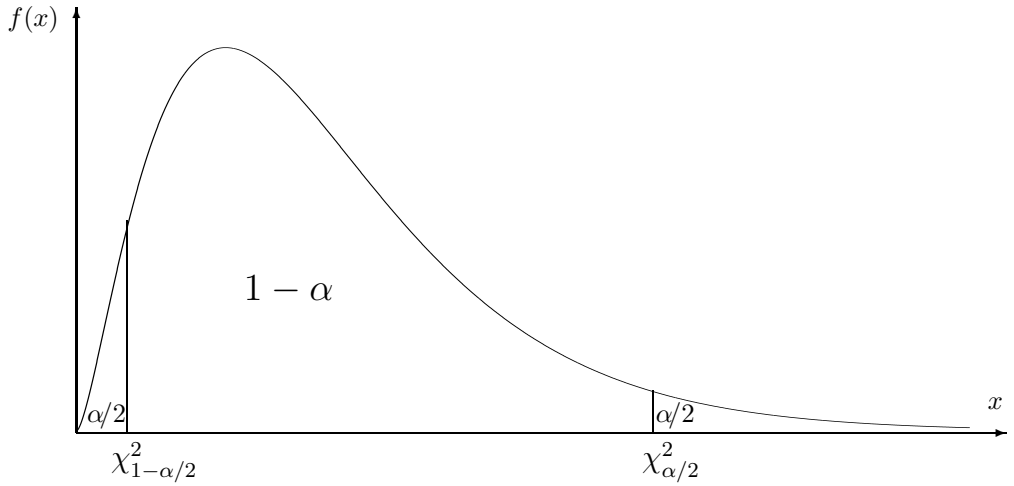


FIGURE 9.13: *Critical values of the Chi-square distribution.*

A confidence interval for the population standard deviation $\sigma = \sqrt{\sigma^2}$ is just one step away (Exercise 9.21).

Confidence interval for the standard deviation

$$\boxed{\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}},\ \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}\right]} \tag{9.25}$$

Example 9.40. In Example 9.31 on p. 285, we relied on the reported parameters of the measurement device and assumed the known standard deviation $\sigma = 2.2$. Let us now rely on the data only and construct a 90% confidence interval for the standard deviation. The sample contained $n = 6$ measurements, 2.5, 7.4, 8.0, 4.5, 7.4, and 9.2.

Solution. Compute the sample mean and then the sample variance,

$$\overline{X} = \frac{1}{6}(2.5 + \ldots + 9.2) = 6.5;$$

$$s^2 = \frac{1}{6-1}\left\{(2.5-6.5)^2 + \ldots + (9.2-6.5)^2\right\} = \frac{31.16}{5} = 6.232.$$

(actually, we only need $(n-1)s^2 = 31.16$).

From Table A6 of Chi-square distribution with $\nu = n - 1 = 5$ degrees of freedom, we find the critical values $\chi^2_{1-\alpha/2} = \chi^2_{0.95} = 1.15$ and $\chi^2_{\alpha/2} = \chi^2_{0.05} = 11.1$. Then,

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}},\ \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}\right] = \left[\sqrt{\frac{31.16}{11.1}},\ \sqrt{\frac{31.16}{1.15}}\right] = \underline{[1.68,\ 5.21]}.$$

is a 90% confidence interval for the population standard deviation (and by the way, $[1.68^2, 5.21^2] = [2.82, 27.14]$ is a 90% confidence interval for the variance). ◊

9.5.3 Testing variance

Suppose now that we need *to test* the population variance, for example, to make sure that the actual variability, uncertainty, volatility, or risk does not exceed the promised value. We'll derive a level α test based on the Chi-square distribution of the rescaled sample variance.

Level α test

Let $X_1, \dots, X_n$ be a sample from the Normal distribution with the unknown population variance σ^2. For testing the null hypothesis

$$H_0 : \sigma^2 = \sigma_0^2,$$

compute the χ^2-*statistic*

$$\chi^2_{\text{obs}} = \frac{(n-1)s^2}{\sigma_0^2}.$$

As we know, it follows the χ^2 distribution with $(n-1)$ degrees of freedom if H_0 is true and σ_0^2 is indeed the correct population variance. Thus, it only remains to compare χ^2_{obs} with the critical values from Table A6 of χ^2 distribution, using $\nu = n-1$.

Testing against the *right-tail* alternative $H_A : \sigma^2 > \sigma_0^2$, reject H_0 if $\chi^2_{\text{obs}} \geq \chi^2_\alpha$.

Testing against the *left-tail* alternative $H_A : \sigma^2 < \sigma_0^2$, reject H_0 if $\chi^2_{\text{obs}} \leq \chi^2_{1-\alpha}$.

Testing against the *two-sided* alternative $H_A : \sigma^2 \neq \sigma_0^2$, reject H_0 if either $\chi^2_{\text{obs}} \geq \chi^2_{\alpha/2}$ or $\chi^2_{\text{obs}} \leq \chi^2_{1-\alpha/2}$.

As an exercise, please verify that in each case, the probability of type I error is exactly α.

P-value

For one-sided χ^2-tests, the P-value is computed the same way as in Z-tests and T-tests. It is always *the probability of the same or more extreme value of the test statistic than the one that was actually observed.* That is,

$$\begin{aligned}
\text{P-value} &= \boldsymbol{P}\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\} = 1 - F(\chi^2_{\text{obs}}) \quad \text{for a right-tail test,} \\
\text{P-value} &= \boldsymbol{P}\left\{\chi^2 \leq \chi^2_{\text{obs}}\right\} = F(\chi^2_{\text{obs}}) \quad \text{for a left-tail test,}
\end{aligned}$$

where F is a cdf of χ^2 distribution with $\nu = n-1$ degrees of freedom.

But how to compute the P-value for the *two-sided* alternative? Which values of χ^2 are considered "more extreme"? For example, do you think $\chi^2 = 3$ is more extreme than $\chi^2 = 1/3$?

We can no longer claim that the value further away from 0 is more extreme, as we did earlier for Z- and T-tests! Indeed, the χ^2-statistic is always positive, and in two-sided tests, its very small or very large values should be considered extreme. It should be fair to say that smaller values of χ^2 are more extreme than the observed one if χ^2_{obs} itself is small, and larger values are more extreme if χ^2_{obs} is large.

To make this idea rigorous, let us recall (from Section 9.4.10) that P-value equals the

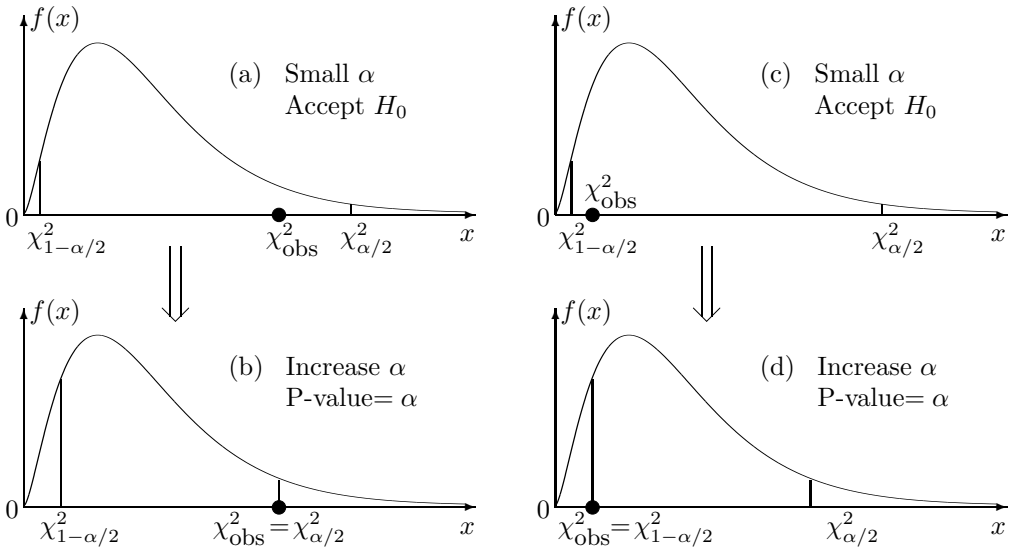


FIGURE 9.14: *P-value for a Chi-square test against a two-sided alternative.*

highest significance level α that yields acceptance of H_0. Start with a very small α as on Figure 9.14a. The null hypothesis H_0 is still accepted because $\chi^2_{\text{obs}} \in [\chi^2_{1-\alpha/2}, \chi^2_{\alpha/2}]$. Slowly increase α until the boundary between acceptance and rejection regions hits the observed test statistic χ^2_{obs}. At this moment, α equals to the P-value (Figure 9.14b), hence

$$P = 2\left(\frac{\alpha}{2}\right) = 2\boldsymbol{P}\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\} = 2\left\{1 - F(\chi^2_{\text{obs}})\right\}. \tag{9.26}$$

It can also happen that the lower rejection boundary hits χ^2_{obs} first, as on Figure 9.14cd. In this case,

$$P = 2\left(\frac{\alpha}{2}\right) = 2\boldsymbol{P}\left\{\chi^2 \leq \chi^2_{\text{obs}}\right\} = 2F(\chi^2_{\text{obs}}). \tag{9.27}$$

So, the P-value is either given by (9.26) or by (9.27), depending on which one of them is smaller and which boundary hits χ^2_{obs} first. We can write this in one equation as

$$P = 2\min\left(\boldsymbol{P}\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\}, \boldsymbol{P}\left\{\chi^2 \leq \chi^2_{\text{obs}}\right\}\right) = 2\min\left\{F(\chi^2_{\text{obs}}),\ 1 - F(\chi^2_{\text{obs}})\right\},$$

where F is the cdf of χ^2 distribution with $\nu = n - 1$ degrees of freedom.

Testing procedures that we have just derived are summarized in Table 9.5. The same tests can also be used for the *standard deviation* because testing $\sigma^2 = \sigma_0^2$ is equivalent to testing $\sigma = \sigma_0$.

Example 9.41. Refer to Example 9.40 on p. 296. The 90% confidence interval constructed there contains the suggested value of $\sigma = 2.2$. Then, by *duality* between confidence intervals and tests, there should be no evidence against this value of σ. Measure the amount of evidence against it by computing the suitable P-value.

Solution. The null hypothesis $H_0 : \sigma = 2.2$ is tested against $H_A : \sigma^2 \neq 2.2$. This is a two-sided test because we only need to know whether the standard deviation equals $\sigma_0 = 2.2$ or

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection region	P-value
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\dfrac{(n-1)s^2}{\sigma_0^2}$	$\chi^2_{\text{obs}} > \chi^2_{\alpha}$	$\boldsymbol{P}\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\}$
	$\sigma^2 < \sigma_0^2$		$\chi^2_{\text{obs}} < \chi^2_{\alpha}$	$\boldsymbol{P}\left\{\chi^2 \leq \chi^2_{\text{obs}}\right\}$
	$\sigma^2 \neq \sigma_0^2$		$\chi^2_{\text{obs}} \geq \chi^2_{\alpha/2}$ or $\chi^2_{\text{obs}} \leq \chi^2_{1-\alpha/2}$	$2\min\left(\boldsymbol{P}\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\}, \boldsymbol{P}\left\{\chi^2 \leq \chi^2_{\text{obs}}\right\}\right)$

TABLE 9.5: χ^2*-tests for the population variance.*

not.

Compute the test statistic from the data in Example 9.40,

$$\chi^2_{\text{obs}} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(5)(6.232)}{2.2^2} = 6.438.$$

Using Table A6 with $\nu = n - 1 = 5$ degrees of freedom, we see that $\chi^2_{0.80} < \chi^2_{\text{obs}} < \chi^2_{0.20}$. Therefore,

$$\boldsymbol{P}\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\} > 0.2 \text{ and } \boldsymbol{P}\left\{\chi^2 \leq \chi^2_{\text{obs}}\right\} > 0.2,$$

hence,

$$P = 2\min\left(\boldsymbol{P}\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\}, \boldsymbol{P}\left\{\chi^2 \leq \chi^2_{\text{obs}}\right\}\right) \underline{\geq 0.4}.$$

The evidence against $\sigma = 2.2$ is very weak; at all typical significance levels, $H_0 : \sigma = 2.2$ should be accepted. ◊

Example 9.42 (Remaining battery life). Creators of a new software claim that it measures the remaining notebook battery life with a standard deviation as low as 5 minutes. To test this claim, a fully charged notebook was disconnected from the power supply and continued on its battery. The experiment was repeated 50 times, and every time the predicted battery life was recorded. The sample standard deviation of these 50 normally distributed measurements was equal 5.6 minutes. At the 1% level of significance, do these data provide evidence that the actual standard deviation is greater than 5 min?

Solution. Test $H_0 : \sigma = 5$ against a right-tail $H_A : \sigma > 5$. From Table A6 with $\nu = n-1 = 49$ degrees of freedom, $\chi^2_{0.01} = 74.9$. The null hypothesis will be rejected if $\chi^2_{\text{obs}} > 74.9$.

Compute the test statistic,

$$\chi^2_{\text{obs}} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(49)(5.6)^2}{5^2} = 61.5.$$

Accept $H_0 : \sigma = 5$. The evidence against it is not significant at the 1% level of significance. ◊

9.5.4 Comparison of two variances. F-distribution.

In this section, we deal with two populations whose variances need to be compared. Such inference is used for the comparison of accuracy, stability, uncertainty, or risks arising in two populations.

Example 9.43 (Efficient upgrade). A data channel has the average speed of 180 Megabytes per second. A hardware upgrade is supposed to improve *stability* of the data transfer while maintaining the same average speed. Stable data transfer rate implies low standard deviation. How can we estimate the relative change in the standard deviation of the transfer rate with 90% confidence? ◊

Example 9.44 (Conservative investment). Two mutual funds promise the same expected return; however, one of them recorded a 10% higher *volatility* over the last 15 days. Is this a significant evidence for a conservative investor to prefer the other mutual fund? (Volatility is essentially the standard deviation of returns.) ◊

Example 9.45 (Which method to use?). For marketing purposes, a survey of users of two operating systems is conducted. Twenty users of operating system ABC record the average level of satisfaction of 77 on a 100-point scale, with a sample variance of 220. Thirty users of operating system DEF have the average satisfaction level 70 with a sample variance of 155. We already know from Section 9.4.8 how to compare the mean satisfaction levels. But what method should we choose? Should we assume equality of population variances, $\sigma_X^2 = \sigma_Y^2$ and use the pooled variance? Or we should allow for $\sigma_X^2 \neq \sigma_Y^2$ and use Satterthwaite approximation? ◊

To compare variances or standard deviations, two independent samples $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_m)$ are collected, one from each population, as on Figure 9.4 on p. 259. Unlike population means or proportions, variances are scale factors, and they are compared through their *ratio*

$$\theta = \frac{\sigma_X^2}{\sigma_Y^2}.$$

A natural estimator for the ratio of *population* variances $\theta = \sigma_X^2/\sigma_Y^2$ is the ratio of *sample* variances

$$\widehat{\theta} = \frac{s_X^2}{s_Y^2} = \frac{\sum(X_i - \overline{X})/(n-1)}{\sum(Y_i - \overline{Y})/(m-1)}. \tag{9.28}$$

The distribution of this statistic was obtained in 1918 by a famous English statistician and biologist *Sir Ronald Fisher* (1890-1962) and developed and formalized in 1934 by an American mathematician *George Snedecor* (1881-1974). Its standard form, after we divide each sample variance in formula (9.28) by the corresponding population variance, is therefore called the *Fisher–Snedecor distribution* or simply *F-distribution* with $(n-1)$ and $(m-1)$ degrees of freedom.

Distribution of the ratio of sample variances

For independent samples $X_1, \dots, X_n$ from Normal (μ_X, σ_X) and $Y_1, \dots, Y_m$ from Normal (μ_Y, σ_Y), the standardized ratio of variances

$$F = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} = \frac{\sum(X_i - \overline{X})^2/\sigma_X^2/(n-1)}{\sum(Y_i - \overline{Y})^2/\sigma_Y^2/(m-1)}$$

has *F-distribution with* $(n-1)$ and $(m-1)$ *degrees of freedom.*

(9.29)

We know from Section 9.5.1 that for the Normal data, both s_X^2/σ_X^2 and s_Y^2/σ_Y^2 follow χ^2-distributions. We can now conclude that *the ratio of two independent χ^2 variables, each divided by its degrees of freedom, has F-distribution.* A ratio of two non-negative continuous random variables, any F-distributed variable is also non-negative and continuous.

F-distribution has two parameters, the *numerator degrees of freedom* and the *denominator degrees of freedom*. These are degrees of freedom of the sample variances in the numerator and denominator of the F-ratio (9.29).

Critical values of F-distribution are in Table A7, and we'll use them to construct confidence intervals and test hypotheses comparing two variances.

One question though... Comparing two variances, σ_X^2 and σ_Y^2, should we divide s_X^2 by s_Y^2 or s_Y^2 by s_X^2? Of course, both ratios are ok to use, but we have to keep in mind that in the first case we deal with $\mathrm{F}(n-1, m-1)$ distribution, and in the second case with $\mathrm{F}(m-1, n-1)$. This leads us to an important general conclusion –

If F has $F(\nu_1, \nu_2)$ distribution, then the distribution of $\dfrac{1}{F}$ is $F(\nu_2, \nu_1)$. (9.30)

9.5.5 Confidence interval for the ratio of population variances

Here we construct a $(1-\alpha)100\%$ confidence interval for the parameter $\theta = \sigma_X^2/\sigma_Y^2$. This is about the sixth time we derive a formula for a confidence interval, so we are well familiar with the method, aren't we?

Start with the estimator, $\widehat{\theta} = s_X^2/s_Y^2$. Standardizing it to

$$F = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} = \frac{s_X^2/s_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{\widehat{\theta}}{\theta},$$

we get an F-variable with $(n-1)$ and $(m-1)$ degrees of freedom. Therefore,

$$\boldsymbol{P}\left\{ F_{1-\alpha/2}(n-1, m-1) \le \frac{\widehat{\theta}}{\theta} \le F_{\alpha/2}(n-1, m-1) \right\} = 1 - \alpha,$$

as on Fig 9.15. Solving the double inequality for the unknown parameter θ, we get

$$\boldsymbol{P}\left\{ \frac{\widehat{\theta}}{F_{\alpha/2}(n-1, m-1)} \le \theta \le \frac{\widehat{\theta}}{F_{1-\alpha/2}(n-1, m-1)} \right\} = 1 - \alpha.$$

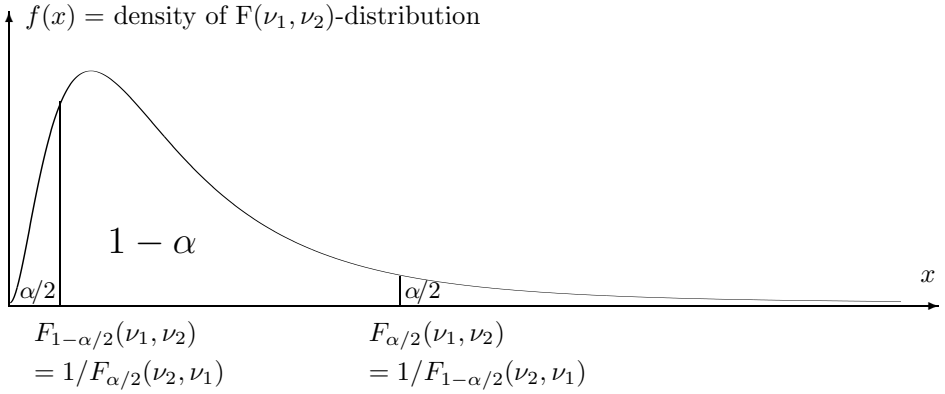


FIGURE 9.15: *Critical values of the F-distribution and their reciprocal property.*

Therefore,

$$\left[\frac{\widehat{\theta}}{F_{\alpha/2}(n-1,m-1)},\ \frac{\widehat{\theta}}{F_{1-\alpha/2}(n-1,m-1)}\right] = \left[\frac{s_X^2/s_Y^2}{F_{\alpha/2}(n-1,m-1)},\ \frac{s_X^2/s_Y^2}{F_{1-\alpha/2}(n-1,m-1)}\right] \tag{9.31}$$

is a $(1-\alpha)100\%$ confidence interval for $\theta = \sigma_X^2/\sigma_Y^2$.

The critical values $F_{1-\alpha/2}(n-1,m-1)$ and $F_{\alpha/2}(n-1,m-1)$ come from F-distribution with $(n-1)$ and $(m-1)$ degrees of freedom. However, our Table A7 has only small values of α. What can we do about $F_{1-\alpha/2}(n-1,m-1)$, a critical value with a large area on the right?

We can easily compute $F_{1-\alpha/2}(n-1,m-1)$ by making use of statement (9.30).

Let $F(\nu_1,\nu_2)$ have F-distribution with ν_1 and ν_2 degrees of freedom, then its reciprocal $F(\nu_2,\nu_1) = 1/F(\nu_1,\nu_2)$ has ν_1 and ν_2 degrees of freedom. According to (9.30),

$$\alpha = \boldsymbol{P}\left\{F(\nu_1,\nu_2) \le F_{1-\alpha}(\nu_1,\nu_2)\right\} = \boldsymbol{P}\left\{F(\nu_2,\nu_1) \ge \frac{1}{F_{1-\alpha}(\nu_1,\nu_2)}\right\}$$

We see from here that $1/F_{1-\alpha}(\nu_1,\nu_2)$ is actually the α-critical value from $F(\nu_2,\nu_1)$ distribution because it cuts area α on the right; see Fig 9.15. We conclude that

Reciprocal property of F-distribution

> The critical values of $F(\nu_1,\nu_2)$ and $F(\nu_2,\nu_1)$ distributions are related as follows,
>
> $$F_{1-\alpha}(\nu_1,\nu_2) = \frac{1}{F_{\alpha}(\nu_2,\nu_1)}$$

(9.32)

We can now obtain the critical values from Table A7 and formula (9.32), plug them into (9.31), and the confidence interval is ready.

Confidence interval for the ratio of variances

$$\left[\frac{s_X^2}{s_Y^2 F_{\alpha/2}(n-1,m-1)},\ \frac{s_X^2 F_{\alpha/2}(m-1,n-1)}{s_Y^2}\right] \tag{9.33}$$

Example 9.46 (EFFICIENT UPGRADE, CONTINUED). Refer to Example 9.43. After the upgrade, the instantaneous speed of data transfer, measured at 16 random instants, yields a standard deviation of 14 Mbps. Records show that the standard deviation was 22 Mbps before the upgrade, based on 27 measurements at random times. We are asked to construct a 90% confidence interval for the relative change in the standard deviation (assume Normal distribution of the speed).

Solution. From the data, $s_X = 14$, $s_Y = 22$, $n = 16$, and $m = 27$. For a 90% confidence interval, use $\alpha = 0.10$, $\alpha/2 = 0.05$. Find $F_{0.05}(15, 26) \approx 2.07$ and $F_{0.05}(26, 15) \approx 2.27$ from Table A7. Or, alternatively, use functions `qf(0.95,15,26)`, `qf(0.95,26,15)` in R or `finv(0.95,15,26)`, `finv(0.95,26,15)` in MATLAB to get the exact values, 2.0716 and 2.2722. Then, the 90% confidence interval for the *ratio of variances* $\theta = \sigma_X^2/\sigma_Y^2$ is

$$\left[\frac{14^2}{22^2 \cdot 2.07},\ \frac{14^2 \cdot 2.27}{22^2}\right] = [0.20,\ 0.92].$$

For the *ratio of standard deviations* $\sigma_X/\sigma_Y = \sqrt{\theta}$, a 90% confidence interval is obtained by simply taking square roots,

$$[\sqrt{0.20},\ \sqrt{0.92}] = \underline{[0.44,\ 0.96]}.$$

Thus, we can assert with a 90% confidence that the new standard deviation is between 44% and 96% of the old standard deviation. With this confidence level, *the relative reduction in the standard deviation of the data transfer rate (and therefore, the relative increase of stability) is between 4% and 56%* because this relative reduction is $(\sigma_Y - \sigma_X)/\sigma_Y = 1 - \sqrt{\theta}$. ♢

Example 9.47 (EFFICIENT UPGRADE, CONTINUED AGAIN). Refer again to Examples 9.43 and 9.46. Can we infer that the channel became twice as stable as it was, if increase of stability is measured by the proportional reduction of standard deviation?

Solution. The 90% confidence interval obtained in Example 9.46 contains 0.5. Therefore, at the 10% level of significance, there is no evidence against $H_0 : \sigma_X/\sigma_Y = 0.5$, which is a two-fold reduction of standard deviation (recall Section 9.4.9 about the duality between confidence intervals and tests). This is all we can state - there is no evidence against the claim of a two-fold increase of stability. There is no "proof" that it actually happened. ♢

Testing hypotheses about the ratio of variances or standard deviations is in the next section.

Null Hypothesis $H_0: \frac{\sigma_X^2}{\sigma_Y^2} = \theta_0$		Test statistic $F_{\text{obs}} = \frac{s_X^2}{s_Y^2}/\theta_0$
Alternative Hypothesis	Rejection region	P-value Use $\text{F}(n-1, m-1)$ distribution
$\frac{\sigma_X^2}{\sigma_Y^2} > \theta_0$	$F_{\text{obs}} \geq F_\alpha(n-1, m-1)$	$\boldsymbol{P}\{F \geq F_{\text{obs}}\}$
$\frac{\sigma_X^2}{\sigma_Y^2} < \theta_0$	$F_{\text{obs}} \leq F_\alpha(n-1, m-1)$	$\boldsymbol{P}\{F \leq F_{\text{obs}}\}$
$\frac{\sigma_X^2}{\sigma_Y^2} \neq \theta_0$	$F_{\text{obs}} \geq F_{\alpha/2}(n-1, m-1)$ or $F_{\text{obs}} < 1/F_{\alpha/2}(m-1, n-1)$	$2\min\left(\boldsymbol{P}\{F \geq F_{\text{obs}}\}, \boldsymbol{P}\{F \leq F_{\text{obs}}\}\right)$

TABLE 9.6: *Summary of F-tests for the ratio of population variances.*

9.5.6 F-tests comparing two variances

In this section, we test the null hypothesis about a *ratio of variances*

$$H_0: \frac{\sigma_X^2}{\sigma_Y^2} = \theta_0 \tag{9.34}$$

against a one-sided or a two-sided alternative. Often we only need to know if two variances are equal, then we choose $\theta_0 = 1$. F-distribution is used to compare variances, so this test is called the **F-test**.

The test statistic for (9.34) is

$$F = \frac{s_X^2}{s_Y^2}/\theta_0,$$

which under the null hypothesis equals

$$F = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}.$$

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are samples from Normal distributions, this *F-statistic* has F-distribution with $(n-1)$ and $(m-1)$ degrees of freedom.

Just like χ^2, F-statistic is also non-negative, with a non-symmetric right-skewed distribution. Level α tests and P-values are then developed similarly to χ^2, see Table 9.6.

Example 9.48 (Which method to use? Continued). In Example 9.45 on p. 300, $n = 20$, $\overline{X} = 77$, $s_X^2 = 220$; $m = 30$, $\overline{Y} = 70$, and $s_Y^2 = 155$. To compare the population means by a suitable method, we have to test whether the two population variances are equal or not.

Solution. Test $H_0: \sigma_X^2 = \sigma_Y^2$ vs $H_A: \sigma_X^2 \neq \sigma_Y^2$ with the test statistic

$$F_{\text{obs}} = \frac{s_X^2}{s_Y^2} = 1.42.$$

For testing equality of variances, we let the tested ratio $\theta_0 = 1$. This is a two-sided test, so the P-value is

$$P = 2\min\left(\boldsymbol{P}\{F \geq 1.42\}, \boldsymbol{P}\{F \leq 1.42\}\right) = \ldots ?$$

How to compute these probabilities for the F-distribution with $n - 1 = 19$ and $m - 1 = 29$ degrees of freedom? R and MATLAB, as always, can give us the exact answer. Typing `1-pf(1.42,19,29)` in R or `1-fcdf(1.42,19,29)` in MATLAB, we obtain $\boldsymbol{P}\{F \geq 1.42\} = 0.1926$. Then,

$$P = 2\min(0.1926,\, 1 - 0.1926) = \underline{0.3852}.$$

Table A7 can also be used, for an approximate but a completely satisfactory solution. This table does not have exactly 19 and 29 degrees of freedom and does not have a value $F_\alpha = 1.42$. However, looking at 15 and 20 d.f. for the numerator and 25 and 30 d.f. for the denominator, we see that 1.42 is always between $F_{0.25}$ and $F_{0.1}$. This will do it for us. It implies that $\boldsymbol{P}\{F \geq .42\} \in (0.1, 0.25)$, $\boldsymbol{P}\{F \leq 1.42\} \in (0.75, 0.9)$, and therefore, the P-value is

$$P = 2\boldsymbol{P}\{F \geq 1.42\} \underline{\in (0.2,\, 0.5)}\,.$$

This is a high P-value showing no evidence of different variances. It should be ok to use the exact two-sample T-test with a pooled variance (according to which there is a mild evidence at a 4% level that the first operating system is better, $t = 1.80$, $P = 0.0388$). ◊

Example 9.49 (Are all the conditions met?). In Example 9.44 on p. 300, we are asked to compare volatilities of two mutual funds and decide if one of them is more risky than the other. So, this is a one-sided test of

$$H_0 : \sigma_X = \sigma_Y \quad \text{vs} \quad H_A : \sigma_X > \sigma_Y.$$

The data collected over the period of 30 days show a 10% higher volatility of the first mutual fund, i.e., $s_X/s_Y = 1.1$. So, this is a standard F-test, right? A careless statistician would immediately proceed to the test statistic $F_{\text{obs}} = s_X^2/s_Y^2 = 1.21$ and the P-value $P = \boldsymbol{P}\{F \geq F_{\text{obs}}\} \underline{\geq 0.25}$ from Table A7 with $n - 1 = 29$ and $m - 1 = 29$ d.f., and jump to a conclusion that there is *no evidence that the first mutual fund carries a higher risk.*

Indeed, why not? Well, every statistical procedure has its assumptions, conditions under which our conclusions are valid. A careful statistician always checks the assumptions before reporting any results.

If we conduct an F-test and refer to the F-distribution, what conditions are required? We find the answer in (9.29) on p. 301. Apparently, for the F-statistic to have F-distribution under H_0, each of our two samples has to consist of independent and identically distributed Normal random variables, and the two samples have to be independent of each other.

Are these assumptions plausible, at the very least?

1. Normal distribution - may be. Returns on investments are typically not Normal but log-returns are.

2. Independent and identically distributed data within each sample - unlikely. Typically, there are economic trends, ups and downs, and returns on two days in a row should be dependent.

3. Independence between the two samples - it depends. If our mutual funds contain stocks from the same industry, their returns are surely dependent.

Actually, conditions 1-3 can be tested statistically, and for this we need to have the entire samples of data instead of the summary statistics.

The F-test is quite *robust*. It means that a mild departure from the assumptions 1-3 will not affect our conclusions severely, and we can treat our result as *approximate*. However, if the assumptions are not met even approximately, for example, the distribution of our data is asymmetric and far from Normal, then the P-value computed above is simply wrong. ◊

Discussion in Example 9.49 leads us to a very important practical conclusion.

> *Every statistical procedure is valid under certain assumptions.*
> *When they are not satisfied, the obtained results may be wrong and misleading.*
> *Therefore, unless there are reasons to believe that all the conditions are met,*
> *they have to be tested statistically.*

R notes

R has tools for tests and confidence intervals discussed in this chapter.

Testing means. To do one-sample T-test of $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ or to obtain a confidence interval for μ, type `t.test(X,mu=mu0)`. The output will include the sample mean, the test statistic, degrees of freedom, the p-value of this two-sided test, as well as the 95% confidence interval.

For a one-sided test, add an `alternative` option with values `"less"` or `"greater"`. However, notice that the confidence interval will also be one-sided in this case. The default confidence level is 95%. For other levels, add an option called `conf.level` option. For example, to get a 99% confidence interval for μ, type `t.test(X, conf.level=0.99)`, and for a one-sided, right-tail test of $H_0 : \mu = 10$ vs $H_A : \mu > 10$, use `t.test(X, mu=10, alternative="greater")`.

Two-sample tests and confidence intervals. The same command `t.test` can be employed for two-sample tests and confidence intervals. You only need to enter two variables instead of one. So, to test $H_0 : \mu_X - \mu_Y = 3$ vs $H_A : \mu_X - \mu_Y \neq 3$ and to construct a 90% confidence interval for the difference of means $(\mu_X - \mu_Y)$, enter `t.test(X, Y, mu=3, conf.level=0.90)`. Satterthwaite approximation will be applied by default. To conduct the equal-variances test with a pooled variance, give the option `var.equal = TRUE`. For a paired T-test, use `paired = TRUE`.

Testing variances. A two-sample F-test for variances requires a command `var.test`, whose syntax is rather similar to `t.test`. Say, to test $H_0 : \sigma_X^2 = \sigma_Y^2$ vs $H_A : \sigma_X^2 \neq \sigma_Y^2$, type `var.test(X, Y, ratio=1)`. This tool also has options `alternative` and `conf.level`, to let you choose a one-sided alternative and the desired confidence level.

Testing proportions. A special command `prop.test` is for the inference about population proportions. Variable X in this case is the number of successes, and it has to be followed by the number of trials n and the null proportion p_0 that we are testing. For example, `prop.test(60,100,0.5)` is for testing $H_0 : p = 0.5$ vs $H_A : p \neq 0.5$ based on a

sample of $X = 60$ successes in $n = 100$ Bernoulli trials. This command also allows options `alternative` and `conf.level`.

For a two-sample test comparing two population proportions, a pair of counts and a pair of sample sizes should be written in place of X and n. For example, suppose that we need to conduct a two-sample test of proportions $H_0 : p_1 = p_2$ vs $H_A : p_1 < p_2$ based on samples of sizes $n_1 = 100$ and $n_2 = 120$. Thirty-three successes in the first sample and forty-eight in the second sample are observed. Command `prop.test( c(33,48), c(100,120), alternative="less" )` returns a p-value of 0.176, so the difference between proportions is not found statistically significant.

By the way, R applies a χ^2-test (Chi-square test) here although we know that it can also be done with a Z-test. The method used by R is more general. Doing it with a χ^2-test, R can handle more than two proportions with this command... as we'll see in the next chapter!

MATLAB notes

Some hypothesis tests are built in MATLAB. For a Z-test of $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$, write `ztest(X,mu0,sigma,alpha)`, with the given tested value of μ_0, known standard deviation σ, and significance level α. For a one-sided test, add `'left'` or `'right'` after α. The test will return 0 if the null hypothesis is not rejected and 1 if it has to be rejected at level α. To see a P-value for this test, write `[h,p] = ztest(X,mu0,sigma)`.

T-tests can be done similarly, with a command `[h,p] = ttest(X,mu0)`. Of course, the standard deviation is not entered here; it is unknown.

Similar procedures are available for two-sample T-tests, as well as testing variances and ratios of variances. The corresponding commands are `ttest2(X,Y,alpha)`, `vartest(X,sigma0)`, and `vartest2(X,Y,alpha)`.

Confidence intervals for the mean, the difference of means, the variance, and the ratio of variances can be obtained simply by adding `CI` in `[h,p,CI]`. For example, `[h,p,CI] = vartest2(X,Y,0.01)` returns a 99% confidence interval for the ratio of variances σ_X^2/σ_Y^2, just as we derived it in Section 9.5.5.

Summary and conclusions

After taking a general look at the data by methods of Chapter 8, we proceeded with the more advanced and informative statistical inference described in this chapter.

There is a number of methods for estimating the unknown population parameters. Each method provides estimates with certain good and desired properties. We learned two general methods of **parameter estimation**.

Method of moments is based on matching the population and sample moments. It is relatively simple to implement, and it makes the estimated distribution "close" to the actual distribution of data in terms of their moments.

Maximum likelihood estimators maximize the probability of observing the sample that was

really observed, thus making the actually occurring situation as likely as possible under the estimated distribution.

In addition to parameter estimates, **confidence intervals** show the margin of error and attach a confidence level to the result. A $(1-\alpha)100\%$ confidence interval contains the unknown parameter with probability $(1-\alpha)$. It means that $(1-\alpha)100\%$ of all confidence intervals constructed from a large number of samples should contain the parameter, and only $100\alpha\%$ may miss it.

We can use data to verify statements and **test hypotheses**. Essentially, we measure the evidence provided by the data against the *null hypothesis* H_0. Then we decide whether it is sufficient for rejecting H_0.

Given *significance level* α, we can construct acceptance and rejection regions, compute a suitable *test statistic*, and make a decision depending on which region it belongs to. Alternatively, we may compute a P-value of the test. It shows how significant the evidence against H_0 is. Low P-values suggest rejection of the null hypothesis. P-value is the boundary between levels α-to reject and α-to-accept. It also represents the probability of observing the same or more extreme sample than the one that was actually observed.

Depending on what we are testing against, the *alternative hypothesis* may be two-sided or one-sided. We account for the direction of the alternative when we construct acceptance and rejection regions and when we compute a P-value.

We developed confidence intervals and tests of hypotheses for a number of parameters – population means, proportions, and variances, as well as the difference of means, the difference of proportions, and the ratio of variances. This covers the majority of practical problems.

Each developed statistical procedure is valid under certain assumptions. Verifying them is an important step in statistical inference.

Exercises

9.1. Estimate the unknown parameter θ from a sample

$$3, 3, 3, 3, 3, 7, 7, 7$$

drawn from a discrete distribution with the probability mass function

$$\begin{cases} P(3) &= \theta \\ P(7) &= 1-\theta \end{cases}.$$

Compute two estimators of θ:

(a) the method of moments estimator;

(b) the maximum likelihood estimator.

Also,