Chapter 11

Introduction to Estimation

So far, we have been learning about the language of probability models, and how to use such models to extract information about the random experiments they represent. For instance, we have discussed how to apply mathematical reasoning to a probability model in order to calculate the probability of a certain event, or to deduce information on the concentration or the long-run average of a random quantity. Mathematical reasoning on the basis of an already-chosen probability model is the subject of *probability theory*.

In practice, however, it is not always obvious which model is appropriate for a random phenomenon, and even if we are able to choose a reasonable parametric model, we may still need to tune the parameters to make the model compatible with the random experiment. A reasonable idea for bridging this gap is to use statistical evidence (in the form of random samples or other forms of data) to infer information about the model. For instance, we may use statistical data to estimate the parameters of the model, or to judge between competing models. This is the domain of *statistical inference*.

Starting from this chapter, we talk about two standard problems in statistical inference, namely estimation and hypothesis testing. Specifically, this chapter includes a teaser for the topic of estimation. Point and interval estimations will be studied more systematically in the following two chapters. We will discuss hypothesis testing afterwards. Our discussions will be limited to the so-called *frequentist* approach.

11.1 Teaser: measurement

Every measurement is subject to various sources of error. As a result, there is often a discrepancy (of unknown magnitude) between the result of the measurement and the quantity to be measured. We model this discrepancy using probabilities.

As an example, suppose we want to measure to voltage between two points A and B in an electronic circuit using a voltmeter (Figure 11.1). Every time we measure the voltage, we may get a slightly different reading

$$1.563\,\text{volt}\,, \qquad 1.559\,\text{volt}\,, \qquad 1.561\,\text{volt}\,, \qquad \dots$$

Ⓠ What is the true value of the voltage?

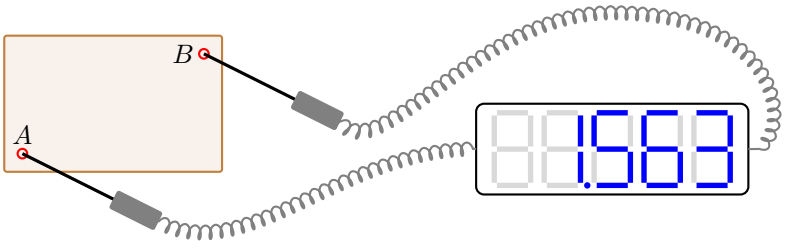

Figure 11.1: A voltmeter for measuring the voltage between A and B

11.1.1 A model for measurement

Let us denote the true value of the voltage between A and B by v_{AB}. This is a (non-random) number, which nevertheless is unknown to us. The *reading* on the voltmeter is not exactly v_{AB}, but

$$V := v_{AB} + R$$

where R indicates a *random error*, which we can model as a random variable. If the device is calibrated (i.e., well-tuned), we have $\mathbb{E}[R] = 0$. This implies that the reading on the voltmeter is *unbiased*, that is, $\mathbb{E}[V] = v_{AB}$. The variance of the error $\sigma_R^2 := \mathbb{V}\mathrm{ar}[R]$ is an indicator of how large the error can typically be. The standard deviation of the error σ_R is sometimes referred to as the *standard error* of the measurement.

Suppose we have a calibrated voltmeter with standard error $\sigma_R = 0.02\,\mathrm{volt}$, and we use that to measure the voltage between A and B. Thus $\mathbb{E}[V] = v_{AB}$ and $\mathbb{V}\mathrm{ar}[V] = \mathbb{V}\mathrm{ar}[R] = (0.02)^2\,\mathrm{volt}^2$.

Ⓠ Suppose the voltmeter shows $1.563\,\mathrm{volt}$. How can we interpret this reading?

Ⓐ The reading $1.563\,\mathrm{volt}$ is the observed value of the random variable V. For the random variable V (i.e., before performing the measurement), Chebyshev's inequality gives

$$\mathbb{P}\left(|V - v_{AB}| < 3 \times 0.02\right) \geq 1 - \frac{1}{3^2} \approx 88.9\% \,.$$

The event $|V - v_{AB}| < 3 \times 0.02$ can be equivalently expressed in either of the following forms:[1]

- ⟨the interval $v_{AB} \pm 0.06$ contains V⟩,
- ⟨the interval $V \pm 0.06$ contains v_{AB}⟩

(see Figure 11.2). Using the latter form, we obtain that

$$\mathbb{P}(V \pm 0.06 \text{ contains } v_{AB}) \geq 88.9\% \,. \qquad (\circledcirc)$$

Thinking of probabilities as idealized frequencies, we arrive at the following conclusion: If we repeat the measurement many many times, then in at least 88.9% of the measurements, the interval $V \pm 0.06$ will contain the true value of the voltage v_{AB}.

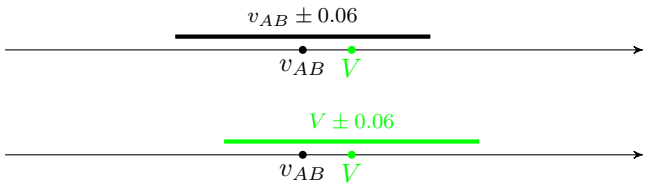

Figure 11.2: The intervals $v_{AB} \pm 0.06$ (black) and $V \pm 0.06$ (green). Note that V belongs to the black interval if and only if v_{AB} belongs to the green interval. Each of these conditions is equivalent to the condition that the distance between v_{AB} and V is less than $0.06\,\mathrm{volt}$. The black interval is a fixed interval which is unknown to us. The green interval is a random interval which can be computed on the basis of the observation.

Let us emphasize that the above interpretation concerns not with the specific value $1.563\,\mathrm{volt}$ but with the measurement procedure that has lead to that value.

In inequality ($\circledcirc$), the value 0.06 is an indicator of the *accuracy* of the measurement, and 88.9% is an indicator of the *level of confidence* we can have on the result. More generally, for every $a > 0$, Chebyshev's inequality gives

$$\mathbb{P}(V \pm 0.02a \text{ contains } v_{AB}) \geq 1 - 1/a^2 \,,$$

where again, $0.02a$ is an indicator of accuracy and $1 - 1/a^2$ is an indicator of the level of confidence. Note that there is a trade-off between accuracy and confidence:

- Choosing a to be larger, we get lower accuracy in our measurement but with higher confidence, while
- Choosing a to be smaller, we get higher accuracy in our measurement but with lower confidence.

Can we improve this? More specifically:

Ⓠ Can we use the same voltmeter to get more accuracy more confidently?

[1] The notation $a \pm b$ is a short and convenient way to refer to the interval $(a - b, a + b)$.

11.1.2 Repeated measurements

In order to achieve more accuracy with higher confidence, a natural idea is to repeat the measurement a number of times and take average.

Ⓠ What is the benefit of repeating the measurement?

🄰 We expect the average of n independent measurement to be typically closer to v_{AB} than the result of a single measurement.

To see why this is the case, let $V_1, V_2, \dots, V_n$ be the results of n independent measurements. Mathematically, $V_1, V_2, \dots, V_n$ are independent random variables with the same distribution as V. Let

$$\overline{V}_n \coloneqq \frac{V_1 + V_2 + \cdots + V_n}{n}$$

be the average of these n measurements. Note that

$$\mathbb{E}[\overline{V}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[V_i] = \frac{1}{n} \cdot n \cdot v_{AB} = v_{AB} \ ,$$
$$\mathbb{V}\mathrm{ar}[\overline{V}_n] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[V_i] = \frac{1}{n} \cdot n \cdot \sigma_R^2 = \frac{1}{n}\sigma_R^2 \ .$$

Thus, the distribution of $\overline{V}_n$ is more concentrated around v_{AB} than the distribution of a single measurement, which is to say, $\overline{V}_n$ is typically closer to v_{AB} than the result of a single measurement.

Recall that we are using a calibrated voltmeter with standard error $\sigma_R = 0.02\,\mathrm{volt}$. Suppose that we make $n = 5$ independent measurements, and use their average $\overline{V}_5$ to estimate v_{AB}.

Ⓠ How do $\overline{V}_5$ and a single measurement compare in terms of accuracy and confidence?

🄰 The variance of $\overline{V}_5$ is $\mathbb{V}\mathrm{ar}[\overline{V}_5] = (0.02)^2/5\,\mathrm{volt}^2$, hence the standard error of $\overline{V}_5$ is $\mathbb{SD}[\overline{V}_5] = 0.02/\sqrt{5}\,\mathrm{volt} \approx 0.00895\,\mathrm{volt}$, which is smaller than the standard error $\sigma_R = 0.02\,\mathrm{volt}$ of a single measurement.

Applying Chebyshev's inequality, for every $a > 0$ we get

$$\mathbb{P}(\overline{V}_5 \pm 0.00895a \text{ contains } v_{AB}) \geq 1 - 1/a^2 \ .$$

For instance, choosing $a \coloneqq 3$, we get

$$\mathbb{P}(\overline{V}_5 \pm 0.0267 \text{ contains } v_{AB}) \geq 88.9\% \ . \qquad (5:☺)$$

Thus, by repeating the measurement 5 times, we achieve more than twice better accuracy (i.e., $0.0267\,\mathrm{volt}$ instead of 0.06, both in volts) with the same 88.9% level of confidence.

Alternatively, choosing $a \coloneqq 3\sqrt{5}$, we get

$$\mathbb{P}(\overline{V}_5 \pm 0.06 \text{ contains } v_{AB}) \geq 97.8\% \ .$$

Hence, repeating the measurement 5 times allows us to achieve the same accuracy of $0.06\,\mathrm{volt}$ with a much higher confidence level of 97.8% (compared to 88.9% for a single measurement).

Clearly, the larger the number of measurements, the better the estimate. More specifically, a larger number of measurements leads to a smaller standard error, hence we can obtain higher precision for a given level of confidence, or higher confidence for a given precision.

In the above discussion, we have been using Chebyshev's inequality to bound the level of confidence because we did not know the distribution of $\overline{V}_n$. If we happen to have more information on the distribution of $\overline{V}_n$, then we can identify the confidence level more accurately. We now consider different scenarios in which we do have more information on the distribution of $\overline{V}_n$.

11.1.3 Many repeated measurements

When the number of measurements n is large, the average $\overline{V}_n$ is, by the central limit theorem, approximately normally distributed.[2] In this case, in order to identify the level of confidence, we can use a normal approximation rather than Chebyshev's inequality.

[2] In most practical applications, $n = 50$ or $n = 100$ should be sufficient for the central limit theorem to provide a reasonable approximation, although for any fixed n, one can find a pathological distribution for which the approximation provided by the central limit theorem is poor.

Namely, the central limit theorem tells us that, when n is large, the average $\overline{V}_n$ of n independent measurements is approximately distributed according to the $\mathsf{N}(v_{AB}, \sigma_R^2/n)$ distribution. Thus, for every $a > 0$,

$$\begin{aligned}
\mathbb{P}\left(\overline{V}_n \pm \frac{\sigma_R}{\sqrt{n}}a \text{ contains } v_{AB}\right) &= \mathbb{P}\left(v_{AB} - \frac{\sigma_R}{\sqrt{n}}a < \overline{V}_n < v_{AB} + \frac{\sigma_R}{\sqrt{n}}a\right) \\
&= \mathbb{P}\left(-a < \frac{\overline{V}_n - v_{AB}}{\sigma_R/\sqrt{n}} < a\right) \qquad \text{approximately } \mathsf{N}(0,1) \\
&\approx \Phi(a) - \Phi(-a) \qquad = 1 - \Phi(a), \text{ by symmetry} \\
&= 2\Phi(a) - 1 \,,
\end{aligned}$$

where, as usual, Φ denotes the cdf of the standard normal distribution (see Figure 11.3).

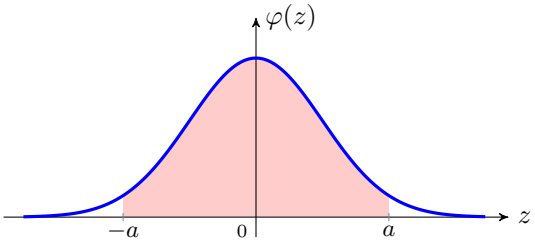

Figure 11.3: The pdf of the standard normal distribution. The area of the shaded region is $\Phi(a) - \Phi(-a) = 2\Phi(a) - 1$.

For instance, suppose that we repeat the measurement $n = 100$ times. As before, we assume that the voltmeter has a standard error of $\sigma_R = 0.02\,\text{volt}$.

Q How does $\overline{V}_{100}$ compare with a single measurement in terms of accuracy and confidence?

A With $n = 100$, the standard error is $\mathbb{SD}[\overline{V}_{100}] = \sigma_R/\sqrt{n} = 0.02/10 = 0.002\,\text{volt}$. For the sake of comparison with (⏦), let us choose $a := 3$. Using the statistical software R, we can find $\Phi(a) \approx 0.9986501$, hence $2\Phi(a) - 1 \approx 0.9973002$. Put together, we obtain

$$\mathbb{P}(\overline{V}_{100} \pm 0.006 \text{ contains } v_{AB}) \approx 99.7\% \,.$$

Thus, compared to the estimate provided by a single measurement, the average of 100 measurements achieves 10 times more accuracy (0.006 volt instead of 0.06 volt) with much higher confidence (99.7% instead of 88.9%).

While repeating the measurement many times improves both the accuracy and the confidence level, it is not always practical. In more realistic scenarios, each measurement has a cost (time, energy, money, ...), and we might not always be able to afford more than a few measurements.

11.1.4 When the error is normal

Suppose that, in its specification, the manufacturer of the voltmeter has provided the extra information that the error R is normally distributed. In other words, we know that $R \sim \mathsf{N}(0, \sigma_R^2)$, where $\sigma_R = 0.02\,\text{volt}$.

Q How can we use this extra information?

A In this case, even with a small sample size, we can achieve more accuracy/confidence than provided by Chebyshev's inequality.

Namely, in this case the measurements $V_1, V_2, \ldots, V_n$ are independent normally distributed random variables. Hence, even if n is small, by the stability of the normal distribution, their average $\overline{V}_n$ is also normally distributed. It follows that $\overline{V}_n \sim \mathsf{N}(v_{AB}, \sigma_R^2/n)$. Therefore, for every $a > 0$,

$$\begin{aligned}
\mathbb{P}\left(\overline{V}_n \pm \frac{\sigma_R}{\sqrt{n}}a \text{ contains } v_{AB}\right) &= \mathbb{P}\left(v_{AB} - \frac{\sigma_R}{\sqrt{n}}a < \overline{V}_n < v_{AB} + \frac{\sigma_R}{\sqrt{n}}a\right) \\
&= \mathbb{P}\left(-a < \frac{\overline{V}_n - v_{AB}}{\sigma_R/\sqrt{n}} < a\right) \\
&= \Phi(a) - \Phi(-a) \\
&= 2\Phi(a) - 1 \,,
\end{aligned}$$

without relying on the central limit theorem.

For instance, suppose that we repeat the measurement $n = 5$ times.

Q How does the normality of error affect the accuracy and confidence level of $\overline{V}_5$?

A With $n = 5$, we have the standard error $\mathbb{SD}[\overline{V}_5] = \sigma_R/\sqrt{n} = 0.02/\sqrt{5} \approx 0.0089$. For instance, choosing $a := 2$, we obtain

$$\mathbb{P}(\overline{V}_5 \pm 0.0178 \text{ contains } v_{AB}) = 2\Phi(2) - 1 \approx 95.4\% \,.$$

Thus, compared to (5:☹) for which we relied on Chebyshev's inequality, we achieve a higher level of accuracy (0.0178 volt instead of 0.0267 volt) and a higher level of confidence (95.4% instead of 88.9%) with the same number of measurements.

11.1.5 Normal error with unknown standard deviation

Sometimes, the assumption of the normality of the error makes sense but we do not know the true value of the standard deviation σ_R. For instance, the manufacturer of the voltmeter might have declared that the error is normal but have not provided us with the value of σ_R. Or we may have theoretical reasons to believe that the error must be normally distributed without having an estimate on the standard error.

Q Can we still use the information regarding the normality of the error to our benefit?

A In this case, we can still achieve better confidence level for the same accuracy, compared to what is provided by Chebyshev's inequality.

Namely, suppose that we perform n independent measurements $V_1, V_2, \ldots, V_n$, and use their average $\overline{V}_n$ to estimate v_{AB}. By the stability of the normal distribution, the average $\overline{V}_n$ is still normally distributed, that is, $\mathsf{N}(v_{AB}, \sigma_R^2/n)$. However, we do not know the variance σ_R^2.

In order to circumvent this problem, a natural idea is to use the same measurements $V_1, V_2, \ldots, V_n$ to estimate σ_R^2. A reasonable estimate for σ_R^2 is given by the *sample variance*

$$\widehat{\sigma}_R^2 := \frac{1}{n-1}\sum_{k=1}^{n}(V_k - \overline{V}_n)^2$$

which we discussed in Chapter 2.[3,4]

Q How does replacing the true variance σ_R^2 with its estimate $\widehat{\sigma}_R^2$ affect the accuracy and confidence level?

Let us consider two different case, based on whether n is large or small.

A1 (when n is large)
When n is large, the sample variance $\widehat{\sigma}_R^2$ should provide a good approximation for the true variance σ_R^2. Therefore,

$$\frac{\overline{V}_n - v_{AB}}{\widehat{\sigma}_R/\sqrt{n}} \approx \frac{\overline{V}_n - v_{AB}}{\sigma_R/\sqrt{n}} \sim \mathsf{N}(0,1) \,.$$

Hence, in this case, for every $a > 0$,

$$\begin{aligned}
\mathbb{P}\left(\overline{V}_n \pm \frac{\widehat{\sigma}_R}{\sqrt{n}}a \text{ contains } v_{AB}\right) &= \mathbb{P}\left(-a < \frac{\overline{V}_n - v_{AB}}{\widehat{\sigma}_R/\sqrt{n}} < a\right) \\
&\approx \mathbb{P}\left(-a < \frac{\overline{V}_n - v_{AB}}{\sigma_R/\sqrt{n}} < a\right) \\
&= \Phi(a) - \Phi(-a) \\
&= 2\Phi(a) - 1 \,,
\end{aligned}$$

as in the case in which σ_R^2 is known to us. Thus, in this case, replacing the true variance σ_R^2 with its estimate $\widehat{\sigma}_R^2$ changes nothing but to make our identification of accuracy and confidence less reliable due to the approximation.

[3] In the following chapter, we will talk more about the sample variance. In particular, we will explain the reason for dividing the sum by $n - 1$ rather than n.

[4] We use the "hat" in $\widehat{\sigma}_R$ to indicate the distinction with the true value σ_R.

A2 (when n is small)

When n is small, the sample variance $\widehat{\sigma}_R^2$ could be very far off from the true variance σ_R^2. Nevertheless, we can exploit another remarkable property of the normal distribution.

Namely, let $T := (\overline{V}_n - v_{AB})/(\widehat{\sigma}_R/\sqrt{n})$ (i.e., the mean measurement standardized using the sample variance rather than the true variance). Although T does not have the $\mathsf{N}(0,1)$ distribution, its distribution turns out to be still independent of σ_R^2. The distribution of T is called *Student's t-distribution* with $n-1$ *degrees of freedom* (see below), and its pdf and cdf can be calculated using standard statistical software such as R.

Therefore, for every $a > 0$,

$$\mathbb{P}\left(\overline{V}_n \pm \frac{\widehat{\sigma}_R}{\sqrt{n}} a \text{ contains } v_{AB}\right) = \mathbb{P}\left(-a < \frac{\overline{V}_n - v_{AB}}{\widehat{\sigma}_R/\sqrt{n}} < a\right) \quad \text{has } \mathsf{T}(n-1) \text{ distribution}$$

$$= F_{\mathsf{T}(n-1)}(a) - F_{\mathsf{T}(n-1)}(-a) \quad = 1 - F_{\mathsf{T}(n-1)}(a) \text{ by symmetry}$$

$$= 2F_{\mathsf{T}(n-1)}(a) - 1 \,,$$

where $F_{\mathsf{T}(n-1)}$ denotes the cdf of Student's t-distribution with $n-1$ degrees of freedom.

For instance, suppose that we repeat the measurement $n = 5$ times.

Q Assuming the normality of the error, how does not knowing σ_R^2 affect the accuracy and confidence level of $\overline{V}_5$?

A Choosing $a := 3$, we can use the computer software R to find $F_{\mathsf{T}(4)}(3) \approx 0.980029$, and $2F_{\mathsf{T}(4)}(3) - 1 \approx 0.960058$. Hence,

$$\mathbb{P}\left(\overline{V}_5 \pm \frac{\widehat{\sigma}_R}{\sqrt{5}} \times 3 \text{ contains } v_{AB}\right) \approx 96.0\% \,.$$

Note that in this case, the accuracy $\frac{\widehat{\sigma}_R}{\sqrt{5}} \times 3$ of the estimate is random. The confidence level 96.0% is somewhat lower than the value 99.7% suggested by the standard normal distribution.

Student's t-distribution. Let $X_1, X_2, \ldots, X_n$ be independent random variables with distribution $\mathsf{N}(\mu, \sigma^2)$, and let $\overline{X}_n := (X_1 + X_2 + \cdots + X_n)/n$ be their mean. By the stability of the normal distribution, $\overline{X}_n$ has distribution $\mathsf{N}(\mu, \sigma^2/n)$. Thus, its standardized version

$$Z := \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution. If instead of the true variance σ^2, we use the *sample variance*

$$\widehat{\sigma}^2 := \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 \,,$$

to standardize $\overline{X}_n$, we obtain a random variable

$$T := \frac{\overline{X}_n - \mu}{\widehat{\sigma}/\sqrt{n}} \,.$$

Remarkably, the distribution of T still does not depend on the parameters μ and σ, but it depends on n. The distribution of T is called *Student's t-distribution*[5] with $n-1$ *degrees of freedom*, and is denoted by $\mathsf{T}(df = n-1)$.[6]

The t-distribution is a symmetric unimodal distribution similar to the standard normal distribution (see Figure 11.4). It has a mean of 0 and a standard deviation which is slightly larger than 1.[7] The larger the parameter df, the closer is the distribution $\mathsf{T}(df)$ to the distribution $\mathsf{N}(0,1)$. However, for small values of df, the two distributions $\mathsf{T}(df)$ and $\mathsf{N}(0,1)$ are considerably different from one another.[8]

[5] Named after statistician William Sealy Gosset (1876–1937), who used Student as his pen name.

[6] Standard software such as R and Python have routines for computing the pdf and the cdf of the t-distribution.

[7] To be specific, the variance of $\mathsf{T}(df)$ is $df/(df-2)$ when $df > 2$. When $df \le 2$, the variance does not exist.

[8] More specifically, compared to a $\mathsf{N}(0,1)$ random variable, a $\mathsf{T}(df)$ random variable typically has a larger absolute value. This is consistent with the fact that the variance of $\mathsf{T}(df)$ is larger than the variance of $\mathsf{N}(0,1)$.

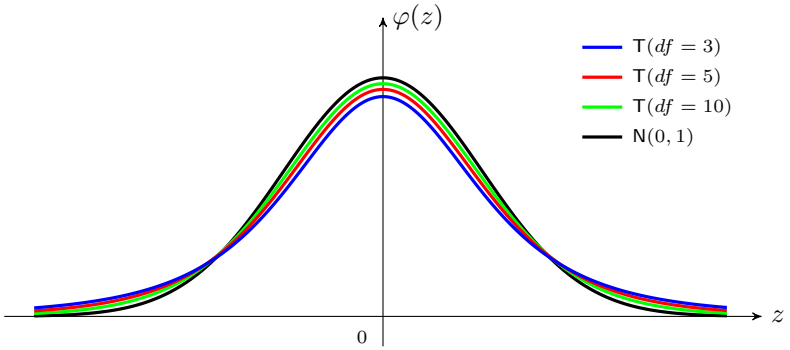

Figure 11.4: The pdf of the $\mathsf{T}(df)$ distribution for a few values of the parameter df, compared with the pdf of the standard normal distribution.