CS 237: Probability in Computing

Prof. Alina Ene **Boston University**

Introduction to parameter estimation and hypothesis testing

Let us start with a motivating example. Suppose we have a coin that may be biased. The coin has some fixed probability p of coming up heads, but the value p is unknown to us. What procedure should we use to find a good estimate for p? How about if we simply want to decide whether the coin is fair or not?

These kind of questions arise in many other settings as well. For example, suppose we have a new medical treatment that we would like to compare with the existing default treatment or a placebo, and decide which treatment is more effective. Or we have a candidate that is running for office and we would like to estimate the fraction of voters that will vote for the candidate. Classical Statistics (also called *frequentist* Statistics) and probability theory provide us with methods for thinking quantitatively about questions like these. In the following, we give a brief introduction to some of the most common and widely used statistical methods. Using the coin example as a running example, we will explore two settings: the parameter estimation setting where we have a fixed but unknown quantity (such as the probability p of heads) that we would like to estimate, and the hypothesis testing setting where we have a hypothesis (such as the coin is fair) that we would like to test.

Parameter estimation by sampling

In parameter estimation, we have a fixed but unknown quantity, which we call a **parameter**. The goal is to construct an **estimate** of this parameter that is close to the true value in some probabilistic sense.

Consider the coin example, where the parameter that we want to estimate is the fixed but unknown probability p of heads. A natural approach for constructing an estimate is to flip the coin n times and estimate the probability using the fraction of heads among the n flips.

We can model the estimation procedure as a probabilistic experiment where we flip a coin independently n times. Let X_i be a random variable indicating whether the i-th flip is heads or not: $X_i = 1$ if the flip is heads and $X_i = 0$ otherwise. Thus we model our estimation procedure by assuming that we have n mutually independent indicator random variables $X_1, X_2, \ldots, X_n$, each with probability p of being equal to 1. Each of the indicator variables is called a **sample** from the Bernoulli(p) distribution. Let S_n be equal to the sum of the indicator variables:

$$S_n = \sum_{i=1}^{n} X_i.$$

The random variable $\overline{X}_n = \frac{S_n}{n}$ is called the **sample mean**. Intuitively (and correctly), we expect that $\overline{X}_n$ provides a useful approximation to the unknown probability p, so let us use it as our **statistical estimate** for p.

This probabilistic model allows us to reason quantitatively about the estimate. For example, we could ask: how many times should we flip the coin so that our estimate is within 0.1 of p with probability at least 0.99? That is, we want

$$\Pr\left(\left|\overline{X}_n - p\right| \leq 0.1\right) \geq 0.99 \quad \text{or equivalently} \quad \Pr\left(\left|\overline{X}_n - p\right| > 0.1\right) \leq 0.01.$$

As we know, the sum S_n is a Binomial(n, p) random variable, and it has expectation $\mathbf{Ex}(S_n) = np$ and variance $\mathbf{Var}(S_n) = np(1-p)$. Therefore

$$\begin{aligned} \mathbf{Ex}\left(\overline{X}_n\right) &= \frac{1}{n}\,\mathbf{Ex}(S_n) = p \\ \mathbf{Var}\left(\overline{X}_n\right) &= \frac{1}{n^2}\,\mathbf{Var}(S_n) = \frac{p(1-p)}{n}. \end{aligned}$$

Thus the expected value of our estimate is precisely the unknown value p; in Statistics, this is referred to as an **unbiased estimate**: an estimate is unbiased if its expectation is equal to the true value of the parameter. In the following, we state several theorems that are helpful for understanding how far away a random variable can be from its expectation[1].

Theorem 1 (Markov Inequality) *Let X be a* **non-negative** *random variable, that is, a random variable whose PDF f_X satisfies $f_X(x) = 0$ for all $x < 0$. For all $x > 0$, we have*

$$\Pr(X \geq x) \leq \frac{\mathbf{Ex}(X)}{x}.$$

We can equivalently restate the Markov inequality as follows.

Corollary 2 (Equivalent form of the Markov Inequality) *Let X be a non-negative random variable. For all $c \geq 1$, we have*

$$\Pr(X \geq c \cdot \mathbf{Ex}(X)) \leq \frac{1}{c}.$$

The Markov Inequality is usually quite weak by itself, but it allows us to get useful bounds when we only have information about the expectation of a random variable. In our case, we also have information about the variance of the random variable. Using the Markov inequality, we can obtain other (more powerful) theorems. One such example is the Chebyshev Inequality.

[1] The proofs of these theorems can be found in LLM Chapter 20.

Theorem 3 (Chebyshev Inequality) *Let X be a random variable. For every real number $x > 0$, we have*

$$\Pr(|X - \mathbf{Ex}(X)| \geq x) \leq \frac{\mathbf{Var}(X)}{x^2}.$$

We can apply the Chebyshev Inequality to our parameter estimation setting as follows. Recall that $\mathbf{Ex}(\overline{X}_n) = p$ and $\mathbf{Var}(\overline{X}_n) = p(1-p)/n$. Note that the quantity $p(1-p)$ is maximized when $p = 1/2$ and thus $\mathbf{Var}(\overline{X}_n) \leq \frac{1}{4n}$. Therefore it follows from the Chebyshev Inequality that

$$\Pr\left(\left|\overline{X}_n - p\right| > 0.1\right) \leq \Pr\left(\left|\overline{X}_n - p\right| \geq 0.1\right) \leq \frac{\mathbf{Var}(\overline{X}_n)}{(0.1)^2} \leq \frac{1}{4n(0.1)^2} = \frac{25}{n}.$$

Therefore, in order to make our estimate be within 0.1 with probability at least 0.99, it suffices to have $n = 2500$.

Estimation by sampling. We can extend the coin example to the following more general parameter estimation approach. Suppose that there is a value μ that we would like to estimate; μ is often referred to as the **population mean**[2]. Also suppose that we have a sampling procedure that generates independent random variables $X_1, X_2, \ldots, X_n$ with the same expectation μ and variance σ^2. Let $S_n = \sum_{i=1}^n X_i$. A statistical estimator for μ is given by the sample mean $\overline{X}_n = \frac{S_n}{n}$.

Note that the random variables X_i are no longer Bernoulli random variables. Nevertheless, we can still use the Chebyshev Inequality to analyze the quality of our estimate. By linearity of expectation, we have

$$\mathbf{Ex}(S_n) = \sum_{i=1}^{n} \mathbf{Ex}(X_i) = n\mu,$$

and therefore

$$\mathbf{Ex}\left(\overline{X}_n\right) = \frac{1}{n}\,\mathbf{Ex}(S_n) = \mu.$$

Since the variables X_i are independent, we have

$$\mathbf{Var}(S_n) = \sum_{i=1}^{n} \mathbf{Var}(X_i) = n\sigma^2,$$

and therefore

$$\mathbf{Var}\left(\overline{X}_n\right) = \frac{1}{n^2}\,\mathbf{Var}(S_n) = \frac{\sigma^2}{n}.$$

Therefore the Chebyshev Inequality gives us the following result.

[2] The terms population mean and sample mean are best understood by analogy to a setting such as polling. We have a population of voters, and a μ fraction (the population mean) of the voters support a candidate. Suppose we sample n voters with replacement (a voter may be chosen more than once) and we let X_i be an indicator for whether the i-th sampled person supports the candidate. The sample mean $\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ is an estimate of the population mean μ.

Theorem 4 (Independent sampling) *Let $X_1, X_2, \ldots, X_n$ be independent random variables with the same expectation μ and variance σ^2. Let $\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$. For every real number $x > 0$,*

$$\Pr\left(\left|\overline{X}_n - \mu\right| \geq x\right) \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

The independent sampling theorem gives us a quantitative way to capture how the sample mean approaches the true mean. In particular, it proves the following result, known as the **weak law of large numbers**: if the sample size is large enough, the sample mean is arbitrarily close to the true mean, with probability close to 1.

Corollary 5 (Weak law of large numbers) *Let $X_1, X_2, \ldots, X_n$ be independent random variables with the same mean μ and the same finite standard deviation $\sigma < \infty$. Let*

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

Then for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|\overline{X}_n - \mu| \leq \epsilon) = 1.$$

As we have seen, the sample mean $\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ gives us an estimator for μ for which we can give quantitative guarantees on how close the estimator is to the true value using the Chebyshev Inequality.

A natural question is whether we can obtain better quantitative bounds. As we discuss in the following, the answer is yes. To this end, let us consider the following random variable:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

A quick calculation shows that Z_n has expectation 0 and variance 1.

$$\mathbf{Ex}(Z_n) = \frac{1}{\sigma\sqrt{n}}(\mathbf{Ex}(S_n) - n\mu) = 0$$
$$\mathbf{Var}(Z_n) = \frac{1}{\sigma^2 n}\mathbf{Var}(S_n) = 1$$

A surprisingly general and useful result, known as the **central limit theorem**, shows that the distribution of Z_n converges to the standard Normal$(0, 1)$ distribution in the following sense.

Theorem 6 (Central limit theorem) *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with the same finite mean μ and finite variance σ^2. Let*

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Then the CDF of Z_n converges to the standard Normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx,$$

in the sense that

$$\lim_{n\to\infty} \Pr(Z_n \leq z) = \Phi(z) \quad \textit{for every } z \in \mathbb{R}.$$

The central limit theorem is surprisingly general, and it applies to every kind of random variables (discrete, continuous, or mixed). The theorem is very powerful, both from a conceptual and an applications point of view.

Going back to our coin flipping example, let us see how to use the Central limit theorem to obtain better quantitative guarantees on the sample mean estimate $\overline{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$. Recall that the X_i variables are independent Bernoulli(p) random variables, with expectation $\mu = p$ and variance $\sigma^2 = p(1-p)$. As before, we are interested in upper bounding the probability $\Pr(|\overline{X}_n - p| > \epsilon)$, where ϵ is some desired accuracy (for example $\epsilon = 0.1$, the value we considered earlier). As in the statement of the Central limit theorem, we define

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{n\overline{X}_n - np}{\sqrt{p(1-p)n}}.$$

Note that

$$\overline{X}_n = p + \sqrt{\frac{p(1-p)}{n}} \cdot Z_n.$$

By the Central limit theorem, the CDF of Z_n is approximately Normal and we will treat it at such in our calculations. Since the Normal CDF is symmetric around the mean, we can approximate $\Pr(|\overline{X}_n - p| > \epsilon)$ as follows:

$$\begin{aligned}
\Pr(|\overline{X}_n - p| > \epsilon) &= \Pr\left(|Z_n| > \epsilon\sqrt{\frac{n}{p(1-p)}}\right) && \left(\text{Since } \overline{X}_n - p = \sqrt{\tfrac{p(1-p)}{n}} \cdot Z_n\right) \\
&\approx 2\Pr\left(Z_n > \epsilon\sqrt{\frac{n}{p(1-p)}}\right) && \text{(Normal CDF is symmetric)} \\
&= 2\left(1 - \Pr\left(Z_n \leq \epsilon\sqrt{\frac{n}{p(1-p)}}\right)\right) \\
&\approx 2\left(1 - \Phi\left(\epsilon\sqrt{\frac{n}{p(1-p)}}\right)\right)
\end{aligned}$$

Since we do not know p, we cannot look up $\Phi(\epsilon\sqrt{n/p(1-p)})$. Since we are content with an upper bound on the probability $\Pr(|\overline{X}_n - p| > \epsilon)$, it suffices to replace $\Phi(\epsilon\sqrt{n/p(1-p)})$ by its minimum over $p \in [0,1]$. $\Phi(z)$ is an increasing function of z and thus $\Phi(\epsilon\sqrt{n/p(1-p)})$ achieves its minimum at the value p that maximizes $p(1-p)$. As we saw earlier, $p(1-p)$ is maximized when $p = 1/2$. Thus $\Phi\left(\epsilon\sqrt{\frac{n}{p(1-p)}}\right) \geq \Phi(2\epsilon\sqrt{n})$ and

$$\Pr(|\overline{X}_n - p| > \epsilon) \lesssim 2(1 - \Phi(2\epsilon\sqrt{n})).$$

If $\epsilon = 0.1$ and $n = 100$, we obtain the following upper bound:

$$\Pr(|\overline{X}_{100} - p| > 0.1) \lesssim 2(1 - \Phi(2)) = 0.046$$

Let us compare this bound with the one provided by the Chebyshev inequality. As we saw earlier, the Chebyshev inequality gives

$$\Pr(|\overline{X}_n - p| > \epsilon) \leq \frac{1}{4n\epsilon^2}$$

If $\epsilon = 0.1$ and $n = 100$, the upper bound provided by the Chebyshev inequality is $1/(4 \cdot 100 \cdot (0.1)^2) = 0.25$, which is much bigger than the upper bound provided by the Central limit theorem.

The Central limit theorem also provides us with a better upper bound on the number of samples needed to achieve an accuracy ϵ with probability at least $1 - \delta$, that is, $\Pr(|\overline{X}_n - p| \leq \epsilon) \geq 1 - \delta$, or equivalently $\Pr(|\overline{X}_n - p| > \epsilon) \leq \delta$. It suffices to choose a number n of samples such that

$$2(1 - \Phi(2\epsilon\sqrt{n})) \leq \delta.$$

As before, consider $\epsilon = 0.1$ and $1 - \delta = 0.99$, and thus $\delta = 0.01$. Then we need to choose n so that

$$2(1 - \Phi(0.2\sqrt{n})) \leq 0.01 \Rightarrow \Phi(0.2\sqrt{n}) \geq 0.995$$

From the Normal table[3], we see that $\Phi(2.58) = 0.9951$ and thus it suffices to have $n = 167$. This is significantly better than the bound of 2500 samples we found using the Chebyshev inequality.

Hypothesis testing

In hypothesis testing, we have several competing hypotheses, and we want to choose one of the hypotheses. For simplicity, we will consider the setting where we have two hypotheses.

Consider our coin example, and recall that the coin has an unknown probability p of coming up heads. Suppose we want to test whether the coin is fair or not. We formulate two hypotheses:

- Hypothesis H_0 is the hypothesis that the coin is fair ($p = 1/2$).
- Hypothesis H_1 (or H_A) is the hypothesis that the coin is biased ($p \neq 1/2$).

The hypothesis H_0 is called the **null hypothesis**, and the hypothesis H_1 is called the **alternate hypothesis**. Our goal is to design a procedure for deciding between the two hypotheses. We will follow a very similar approach to the one for parameter estimation, and along the way we introduce some of the terms that are commonly used in statistical data analysis. (These concepts arise often, so it is important to get familiar with them.)

[3] `http://www.normaltable.com/`

As in the parameter estimation setting, a natural approach is to flip the coin n times and obtain independent samples $X_1, X_2, \ldots, X_n$, where X_i is the indicator random variable of the i-th flip ($X_i = 1$ if the i-th flip is heads and $X_i = 0$ otherwise). As before, we let $\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ be the sample mean. If the coin is fair, we have $\mathbf{Ex}(\overline{X}_n) = 1/2$. Therefore a natural strategy is to reject the null hypothesis provided that $\overline{X}_n$ is sufficiently far from $1/2$:

$$\text{reject } H_0 \text{ if } \left|\overline{X}_n - \frac{1}{2}\right| > \epsilon$$

where ϵ is a suitable **critical value** to be determined. We will set the critical value ϵ so that the probability of falsely rejecting the null hypothesis is equal to a given value α, called the **significance level**. The significance level α is typically small, and we will use $\alpha = 0.05$ in this example.

Let us now choose the critical value ϵ so that the probability of falsely rejecting the null hypothesis is at most $\alpha = 0.05$. That is, we need to choose ϵ such that

$$\Pr\left(\left|\overline{X}_n - \frac{1}{2}\right| > \epsilon \;\middle|\; H_0\right) \leq \alpha = 0.05.$$

We already saw that we can use the Central limit theorem to analyze the probability above. We showed that, if n is large enough ($n \geq 30$ is a good rule of thumb), the Central limit theorem gives

$$\Pr\left(\left|\overline{X}_n - \frac{1}{2}\right| > \epsilon\right) \approx 2(1 - \Phi(2\epsilon\sqrt{n})),$$

where Φ is the Normal$(0,1)$ CDF. Thus it suffices to choose ϵ so that

$$2(1 - \Phi(2\epsilon\sqrt{n})) \leq 0.05 \Rightarrow \Phi(2\epsilon\sqrt{n}) \geq 0.975$$

By looking up the Normal table, we see that $\Phi(1.96) = 0.975$ and thus we can set $\epsilon = 1.96/(2\sqrt{n})$. For example, if $n = 1000$, a critical value of $\epsilon \approx 0.031$ suffices to obtain a significance level of 0.05. Note that in this case we reject the null hypothesis if $|\sum_{i=1}^n X_i - 500| \geq 31$.

In hypothesis testing, the guarantee that we obtained is sometimes stated in the following way: *the null hypothesis H_0 is not rejected at a significance level of* 0.05. It simply means that the probability of false rejection is at most 0.05. Note that the term used is *not rejected*, as opposed to *accepted*, since our analysis cannot firmly establish that the coin has probability of heads precisely equal to 1/2 (for instance, we cannot firmly distinguish between a probability of heads equal to 1/2 and 0.51).

We can summarize and generalize the salient points of the coin example to obtain a general methodology for hypothesis testing.

Significance testing methodology. The goal is to perform a statistical test of a hypothesis H_0 based on samples $X_1, X_2, \ldots, X_n$.

The following steps are performed before the samples are observed.

(a) Choose a **statistic** S, that is, a random variable that will summarize the data. Mathematically, we choose a function $h : \mathbb{R}^n \to \mathbb{R}$, which gives us the statistic $S = h(X_1, X_2, \ldots, X_n)$. In the coin example, $S = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(b) Specify the **rejection region**, that is, the set of values of S for which H_0 will be rejected as a function of the critical value ϵ (the critical value ϵ will be determined later). In the coin example, the rejection region is $\left\{s \in [0,1] \colon \left|s - \frac{1}{2}\right| > \epsilon\right\}$.

(c) Choose the **significance level**, that is, the desired probability α of a false rejection of H_0. In the coin example, we used $\alpha = 0.05$.

(d) Choose the **critical value** ϵ so that the probability of false rejection is approximately equal to α. In the coin example, we set $\epsilon = 1.96/(2\sqrt{n})$ to obtain a significance level $\alpha = 0.05$.

Once the values $x_1, x_2, \ldots, x_n$ of $X_1, X_2, \ldots, X_n$ are observed:

(i) Calculate the value $s = h(x_1, x_2, \ldots, x_n)$ of the statistic S.

(ii) Reject the hypothesis H_0 if s belongs to the rejection region.

Remarks on choosing the statistics. Note that there is no universal method for choosing a statistic. In some specific settings, such as the coin example, there is a natural choice that can be justified mathematically. In some other settings, we may not be so fortunate, and choosing a statistic is a bit of an art. In the following, we briefly discuss two test statistics for normally distributed samples. These are useful in many settings (in light of the Central limit theorem), and they are further explored in lab 5.

Hypothesis testing for samples with a Normal distribution

Let $X_1, X_2, \ldots, X_n$ be independent samples from a Normal(μ, σ^2) distribution. The mean μ is unknown. The variance σ^2 may be known or unknown, and we consider each of these cases separately. Our goal is to design decision procedures for hypothesis tests involving the unknown mean μ. We consider the following types of tests:

- Double-tailed test: $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$.
- Left-tailed test: $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$.
- Right-tailed test: $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$.

No matter what type of test we consider, we can use the following test statistics, depending on whether the variance σ^2 is known or unknown.

Case 1: the variance σ^2 is known. If the null hypothesis is true, we have $\mu = \mu_0$ and thus $\frac{n\overline{X}_n - n\mu_0}{\sigma\sqrt{n}}$ has the Normal$(0,1)$ distribution. This suggests using the following test statistic, called a

z-test (we will use Z_n instead of S to denote the statistic, to emphasize its connection to the Z_n random variable from the Central limit theorem):

$$Z_n = \frac{n\overline{X}_n - n\mu_0}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma}.$$

Note that Z_n behaves differently depending on whether $\mu = \mu_0$, $\mu < \mu_0$, or $\mu > \mu_0$. To see this, it is helpful to write Z_n as follows:

$$Z_n = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma} = \sqrt{n}\frac{\overline{X}_n - \mu + \mu - \mu_0}{\sigma} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} + \sqrt{n}\frac{\mu - \mu_0}{\sigma}$$

The first term $\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$ has a Normal$(0, 1)$ distribution. As $n \to \infty$, the second term is 0 if $\mu = \mu_0$, it diverges to $-\infty$ if $\mu < \mu_0$, and it diverges to $+\infty$ if $\mu > \mu_0$. As we discuss later, we can design a test procedure based on this behavior of Z_n.

Case 2: the variance σ^2 is unknown. Since we do not know the variance, we can no longer use the z-test. It is useful to consider the **sample variance**, which is defined as

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

We can show that $\mathbf{Ex}(S_n^2) = \sigma^2$, and thus S_n^2 is an unbiased estimator for the variance (we leave this as an exercise, see the practice problem sheets). A natural approach is to replace the unknown variance σ^2 by the sample variance S_n^2 in the z-test. The resulting statistic, denoted by T_n, is called a t-**test**:

$$T_n = \frac{n\overline{X}_n - n\mu_0}{S_n\sqrt{n}} = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}} = \sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n}$$

Recall that, if H_0 is true, Z_n has a Normal$(0, 1)$ distribution. This is no longer the case for T_n. If H_0 is true, T_n has a different distribution, called the Student t-distribution with $n-1$ degrees of freedom. There are calculators for the CDF for the t-distribution that one can use[4]; note that there is a different CDF for each value of n. The expectation of the t-distribution with $(n-1)$ degrees of freedom is equal to 0 if $n \geq 2$, and it is undefined otherwise. The variance is equal to $\frac{n-1}{n-2}$ if $n \geq 4$, ∞ if $n = 3$, and it is undefined otherwise.

Similarly to Z_n, T_n behaves as follows: if $\mu = \mu_0$, T_n has the t-distribution with $(n-1)$ degrees of freedom; if $\mu < \mu_0$, $T_n \to -\infty$; if $\mu > \mu_0$, $T_n \to +\infty$.

Significance testing. We can use the appropriate test statistic S ($S = Z_n$ if σ^2 is known, and $S = T_n$ if σ^2 is unknown) as part of the significance testing methodology we saw earlier.

Recall the behavior of the statistic: if $\mu = \mu_0$, S has a distribution with known CDF F that is symmetric, meaning that $\Pr(X \geq x) = \Pr(X \leq -x)$ for all $x \in \mathbb{R}$; if $\mu < \mu_0$, $S \to -\infty$; if $\mu > \mu_0$, $S \to +\infty$. We can design a test procedure based on this behavior as follows. We consider each type of tests in turn.

- **Double-tailed tests:** $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$.

[4] `http://stattrek.com/online-calculator/t-distribution.aspx`

In this case, an indication that H_1 is true would be if $|S|$ becomes too large, i.e., $S \to \pm\infty$. Therefore we consider the decision rule:

$$H = \begin{cases} H_0 & \text{if } -\epsilon \leq S \leq \epsilon \\ H_1 & \text{if } |S| > \epsilon \end{cases}$$

The value ϵ (the critical value) depends on our desired significance level α. More precisely, we need to set ϵ so that

$$\Pr(H = H_1 \mid H_0) = \Pr(|S| \geq \epsilon \mid H_0) \leq \alpha.$$

Conditioned on H_0, S has a distribution with symmetric CDF F and thus

$$\Pr(|S| \geq \epsilon \mid H_0) = 2\Pr(S \geq \epsilon \mid H_0) = 2(1 - F(\epsilon)).$$

Thus it suffices to set ϵ so that $2(1 - F(\epsilon)) \leq \alpha$, or equivalently $F(\epsilon) \geq 1 - \frac{\alpha}{2}$. We can do so by looking up F values in the appropriate table.

Example. For example, suppose we want to use a t-test for $n = 10$ samples, and our goal is to have a confidence level $\alpha = 0.05$. Thus we need to set ϵ so that $F(\epsilon) = 0.975$. The degrees of freedom for the t-distribution are $n - 1 = 9$. Using the CDF calculator for the t-distribution with 9 degrees of freedom, we find that $\epsilon = 2.262$.

- **Left-tailed tests:** $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$.

  In this case, an indication that H_1 is true is that $S \to -\infty$. Therefore we consider the following decision rule:

  $$H = \begin{cases} H_0 & \text{if } S \geq \epsilon \\ H_1 & \text{if } S < \epsilon \end{cases}$$

  Similarly to the double-tailed case, the choice of ϵ is based on the condition:

  $$\Pr(H = H_1 \mid H_0) = \Pr(S < \epsilon \mid H_0) = F(\epsilon) \leq \alpha$$

- **Right-tailed tests:** $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$.

  In this case, an indication that H_1 is true is that $S \to +\infty$. Therefore we consider the following decision rule:

  $$H = \begin{cases} H_0 & \text{if } S \leq \epsilon \\ H_1 & \text{if } S > \epsilon \end{cases}$$

  The choice of ϵ is based on the condition:

  $$\Pr(H = H_1 \mid H_0) = \Pr(S > \epsilon \mid H_0) = 1 - F(\epsilon) \leq \alpha$$

The p-value of a test statistic. We can summarize the decision procedures discussed above using a concept called p-value. Let S be the statistic ($S = Z_n$ or $S = T_n$). Suppose that we observe the values $x_1, x_2, \ldots, x_n$ of the samples $X_1, X_2, \ldots, X_n$. Let $s = h(x_1, x_2, \ldots, x_n)$ be the value of our test statistic. The **p-value** of this observed statistic is the following probability:

- For a double-tailed test, the p-value is $2\Pr(S > |s| \mid H_0)$.

- For a left-tailed test, the p-value is $\Pr(S < s \mid H_0)$.
- For a right-tailed test, the p-value is $\Pr(S > s \mid H_0)$.

The p-value can be understood as a probability, given that H_0 is true, of observing an S-statistic value equally or less likely than the one we observed. If the p-value is small, the observed S-statistic is very unlikely under the null hypothesis, and thus we have strong evidence to reject H_0. This leads to the following decision procedure: **reject H_0 if and only if the p-value** $\leq \alpha$.

Example. Suppose we have $n = 64$ random samples from the Normal$(\mu, 10^2)$ distribution, and the observed sample mean is 73. Consider the right-tailed hypothesis test $H_0 : \mu = 70$, $H_1 : \mu > 70$. Since the variance is known, we can perform a z-test. The observed statistic is $z = (73 - 70)/(10/\sqrt{64}) = 2.4$. By looking up the Normal table, we obtain the p-value as follows: the p-value is $\Pr(Z_n > z) = 1 - \Pr(Z_n \leq z) = 1 - \Phi(2.4) = 1 - 0.9918 = 0.0082 < 0.05$. Thus we reject the null hypothesis for the significance level $\alpha = 0.05$.