Handout 4
Hypothesis Testing*

Instructor: Vira Semenova Note author: Danylo Tavrov, Vira Semenova

1 Definitions and Language

As you recall, an estimator is a random variable, and in each given sample, we observe only one of its realizations. Given this observed value, we are interested in answering the following question: is there enough evidence to deny some assertions about a population? This process is called *inference*: we want to infer something about the population parameters (of the unknown distribution of the data) from the sample estimates we observe.

Example. Assume we have an i.i.d. sample $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$, where F is some (unknown) distribution with parameter of interest $\mathbb{E}[X_i] \equiv \theta$. The statement about the population we want to answer is whether $\theta \in (-\infty, 0]$.

A *null hypothesis* H_0 is a statement about a population parameter we want to confirm or discard. For example, $H_0 : \theta < 2$, $H_0 : \sigma_X = \mu_X$. Observe that $H_0 : \overline{X} = 2$ is not a valid statement, because it equates a random variable, $\overline{X}$, and a constant 2, which is wrong for the absolute majority of random variables in this course. In general, the null hypothesis is formulated as $H_0 : \theta \in S$ for some set of values S.

An *alternative hypothesis*, or research hypothesis, H_1 is a complement to H_0. For the examples given above, the alternative hypotheses are formulated as $H_1 : \theta \geq 2$ and $H_1 : \sigma_X \neq \mu_X$, respectively.

Alternative hypotheses can be:

- *one-sided*: when $H_0 : \theta \geq \theta_0$, then $H_1 : \theta < \theta_0$ (or likewise when $H_0 : \theta \leq \theta_0$, then $H_1 : \theta > \theta_0$);
- *two-sided*: when $H_0 : \theta = \theta_0$, then $H_1 : \theta \neq \theta_0$.

These kinds of alternatives highlight which types of deviations from H_0 can be used as evidence against H_0 the researcher is trying to collect.

A *test* is a decision rule that maps potential realizations of the sample, $(x_1, \dots, x_n)$ to rejection or acceptance of the hypothesis. Typically, we construct a *test statistic* $T = T(X_1, \dots, X_n)$ as a function of available data. Then, a test is a decision rule mapping T into two values, accept and reject, more precisely, we reject H_0 if $T \in C$, where C is a *critical region*, and accept H_0 if $T \notin C$.

It is very important to note that accepting the hypothesis *does not mean* that we consider the hypothesis to be true. Many competing hypotheses might be "accepted" in this sense at the same time, but it does not mean that all of them are true. Therefore, we can either "reject" or "fail to reject" a hypothesis, but we *never* "accept" a hypothesis or claim it to be "true." Therefore, when we say that we "accept H_0," we are simply stating that there is no evidence in the data to reject it, nothing more.

By making decisions to "accept" or reject H_0, we will make mistakes from time to time. The mistakes can be classified into Type 1 or Type 2. A *Type 1* mistake (error) is the one of rejecting the correct null hypothesis, while a *Type 2* mistake (error) is the one of failing to reject the wrong null hypothesis. This is depicted in the following table:

	H_0 is true	H_1 is true
H_0 is accepted	Correct decision	Type 2 error
H_0 is rejected	Type 1 error	Correct decision

*We thank Prof. Anna Mikusheva and numerous former ECON 140 and 141 GSIs for their course notes.

The probability of committing a Type 1 error,

$$\text{size}(\theta_0) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta_0\right) = \mathbb{P}\left(T \in C \mid \theta_0\right) \ ,$$

is called the *size of the test*. Here, conditioning on θ_0 means that the probability is calculated under assumption that the test statistic T follows the distribution postulated by the null hypothesis. Then, the *significance level*, usually denoted by α, is the maximal acceptable (to the researcher) size of the test. In practice, popular choices of α are 0.1, 0.05, and 0.01, with 0.05 being the most ubiquitous.

The probability of correct rejection H_0,

$$\text{power}(\theta) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta\right) \ ,$$

for some $\theta \notin S$, is called the *power of the test*. Note that $\text{power}(\theta) = 1 - \mathbb{P}\left(\text{Accept } H_0 \mid \theta\right) = 1 - \mathbb{P}\left(\text{Type 2 error}\right)$.

2 Hypothesis Testing Process

In empirical applications, there is a trade-off between the two types of errors we can make. The current practice is to control for the probability of committing Type 1 errors, and then seek to use the test that would minimize the probability of committing Type 2 errors (maximize the power of the test). We will discuss several approaches to implement this principle.

2.1 Comparing Test Statistic to Critical Values

The first approach states that we specify level α, and then find the *critical value* c such that the test has desired size. In particular:

- for a one-sided alternative of the form $H_1 : \theta < \theta_0$, we select c that satisfies the following equation:

  $$\mathbb{P}\left(T < c \mid \theta_0\right) = \alpha \ .$$

  Clearly, in this case, c is an α-quantile of the distribution of T. The test rejects H_0 if the value of T, calculated for a given sample, is less than c;

- for a one-sided alternative of the form $H_1 : \theta > \theta_0$, we select c that satisfies the following equation:

  $$\mathbb{P}\left(T > c \mid \theta_0\right) = \alpha \ .$$

  Clearly, in this case, c is a $(1-\alpha)$-quantile of the distribution of T. The test rejects H_0 if the value of T, calculated for a given sample, is greater than $c_{1-\alpha}$;

- for a two-sided alternative, we select two critical values, c_1 and c_2, such that

  $$\mathbb{P}\left(T < c_1 \mid \theta_0\right) + \mathbb{P}\left(T > c_2 \mid \theta_0\right) = \alpha \ .$$

  In this case, we typically take c_1 to be the $\frac{\alpha}{2}$-quantile of the distribution of T and c_2 to be the $\left(1 - \frac{\alpha}{2}\right)$-quantile of the distribution of T. The test rejects H_0 if the value of T, calculated for a given sample, is less than c_1 or greater than c_2.

Example. Consider an election with two running candidates. Suppose candidate A received 42% of votes, and candidate B received 58% of votes. Candidate A became convinced that the election was rigged, so they hired an agency that randomly sampled 100 voters, 53 of whom said that they voted for A. Should A conclude that election fraud occurred?

Naive (and incorrect) answer would be to answer in the affirmative, because 53 in 100 is much larger than 42% candidate A got officially. However, this number was calculated only using one sample of 100 randomly chosen voters, and we want to infer something about the population probabilities of voting for each candidate.

Note that each vote, X_i, $i = 1, \ldots, n$, is an i.i.d. draw from a Bernoulli distribution with probability of voting for candidate A equal to (unknown) π. We want to test the following hypothesis: $H_0 : \pi \leq 0.42$ vs. $H_1 : \pi > 0.42$.

Note that we typically state the null hypothesis in such terms that we will be able to make some definitive conclusion. By rejecting the null as stated, we will be able to state (at some level α) that the election was indeed rigged, in the sense that the random sample data come from a Bernoulli distribution with $\pi > 0.42$, which is of course more favorable for candidate A. Had we instead posed $H_0 : \pi \geq 0.42$, we would be interested in accepting this hypothesis. But, as mentioned earlier, that would not mean that H_0 is in any sense true, because with the same success we could fail to reject other hypotheses, which we did not test for, thus our conclusions would not be definitive.

Note also that we use a one-sided hypothesis here because if we reject the null as stated, we will be able to conclude (at some level α) that the true proportion of votes candidate A got was higher that officially stated, which would prove the election was rigged in the sense that is desirable for candidate A. Testing $H_0 : \pi = 0.42$ vs. $H_1 = \pi \neq 0.42$ and rejecting H_0 would only conclude that the true population vote share is not 0.42, but this could mean that it is possibly also less than 0.42, which is perhaps even less desirable for candidate A, and thus would not corroborate their suspicions in any meaningful way.

To test the hypothesis stated above, we use the following test statistic $T(X_1, \ldots, X_n) = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \equiv Z_n$. Note that Z_n is indeed a statistic under the null, because under the null μ_X is assumed to equal some fixed and known value (specifically, $\pi = 0.42$, because $\mu_X = \pi_X$ in this case), and for Bernoulli distribution $\sigma_X^2 = \mu_X(1 - \mu_X)$.

From the central limit theorem, we know that, under the null,

$$Z_n \xrightarrow{d} N(0, 1) \, .$$

According to the above discussion, we need to select the critical value such that

$$\mathbb{P}\left(Z_n > q_{1-\alpha} \mid \pi = 0.42\right) = 1 - \Phi(q_{1-\alpha}) = \alpha \, .$$

For $\alpha = 0.05$, the critical value is the 0.95-quantile of the standard normal distribution, which is approximately equal to $q_{0.95} \approx 1.65$. Therefore, our test will reject if $Z_n > 1.65$. We have:

$$Z_n = \frac{0.53 - 0.42}{\sqrt{0.42 \cdot 0.58 / 100}} \approx 2.229 > 1.65 \, ,$$

thus we reject the null hypothesis and conclude that election was indeed rigged.

2.2 Working With p-Values

Working with critical values necessitates attaching oneself to a specific significance level before performing the test. An alternative approach lies in calculating the probability of observing a value of the test statistic T as large as its sample realization t, which is called the *p-value* and denoted by p. Depending on the kind of the alternative, we proceed as follows:

- for a one-sided alternative of the form $H_1 : \theta < \theta_0$:
$$p = \mathbb{P}\left(T < t \mid \theta_0\right) \, ,$$
- for a one-sided alternative of the form $H_1 : \theta > \theta_0$:
$$p = \mathbb{P}\left(T > t \mid \theta_0\right) \, ,$$
- for a two-sided alternative:
$$p = \mathbb{P}\left(|T| > |t| \mid \theta_0\right) \, .$$

When the p-value is relatively large, we do not have strong confidence to reject the null hypothesis. Otherwise, when the p-value is quite small (for example, less than 5%), we reject the null hypothesis. The threshold for p-value is set arbitrarily, depending on which probability is considered "unlikely" in any given context.

Example. For the election example above, we calculate the p-value as follows:

$$p = \mathbb{P}\left(Z_n > 2.229 \mid \pi = 0.42\right) = 1 - \Phi(2.229) \approx 0.011 \, .$$

In other words, we reject H_0 at any level $\alpha \geq 0.011$, including 0.05.

2.3 Confidence Intervals

Another important concept is that of a *confidence set* at a given level $1-\alpha$, which is the set of such parameter values that cannot be rejected at level α. In this course, the confidence set takes form of a *confidence interval*:

$$\mathbb{P}\left(\mathrm{CI}_{1-\alpha}\left(\hat{\theta}\right) \ni \theta_0\right) = 1-\alpha\ . \tag{2.1}$$

The confidence set characterizes the uncertainty about a parameter value. We say that a confidence interval *covers* θ_0 if $\mathrm{CI}_{1-\alpha}\left(\hat{\theta}\right) \ni \theta_0$.

Note that $\mathrm{CI}_{1-\alpha}\left(\hat{\theta}\right)$ is *random*. Two common misinterpretations of the concept of a confidence interval are as follows:

- the true parameter θ_0 falls into $\mathrm{CI}_{1-\alpha}\left(\hat{\theta}\right)$ with probability $1-\alpha$. This in incorrect because the true parameter is fixed and therefore any probabilistic statements about θ_0 are meaningless;
- any *given* confidence interval, e.g. $[-1;1]$, covers the true parameter θ_0 with probability $1-\alpha$. This is incorrect because any specific interval covers any fixed value with probability 1 if this value belongs to the interval, and probability 0 otherwise.

A correct way to interpret confidence intervals is to recognize that they are random variables, and therefore a confidence interval, *as a random interval*, covers the true parameter with probability $1-\alpha$, meaning that if we sample the data many times and compute $\mathrm{CI}_{1-\alpha}\left(\hat{\theta}\right)$ for each sample, a fraction of $1-\alpha$ of these intervals will contain the true parameter θ_0.

Example. Suppose a new standardized test is given to 100 randomly selected third-grade students in New Jersey. The sample average score $\overline{Y}$ on the test is 58 points and the sample standard deviation s_Y is 8 points. We would like to construct a 95% confidence interval for μ_Y using the statistic $\overline{Y}$.

Recall that

$$\overline{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)\ ,$$

so that the t-statistic

$$t = \frac{\overline{Y}-\mu_Y}{\mathrm{se}\left(\overline{Y}\right)} = \frac{\overline{Y}-\mu_Y}{s_Y/\sqrt{n}} \xrightarrow{d} N(0,1)\ .$$

Note that t is indeed a statistic, because, under the null, μ_Y is a fixed and known value, and s_Y is itself a statistic used to estimate an unknown standard deviation of Y_i.

The $1-\alpha$ confidence interval for t equals to $\left[-q_{1-\frac{\alpha}{2}}; q_{1-\frac{\alpha}{2}}\right]$, where q_α is an α-quantile of the standard normal distribution, because

$$\begin{aligned}\mathbb{P}\left(t \in \left[-q_{1-\frac{\alpha}{2}}; q_{1-\frac{\alpha}{2}}\right]\right) &= \mathbb{P}\left(-q_{1-\frac{\alpha}{2}} \le t \text{ and } t \le q_{1-\frac{\alpha}{2}}\right)\\ &= \Phi\left(q_{1-\frac{\alpha}{2}}\right) - \Phi\left(-q_{1-\frac{\alpha}{2}}\right)\\ &= 1-\frac{\alpha}{2}-\frac{\alpha}{2} = 1-\alpha\ ,\end{aligned}$$

where we have used the properties of the standard normal distribution, in particular, that $-q_\alpha = q_{1-\alpha}$.

Therefore, the $1-\alpha$ confidence interval for $\overline{Y}$ is given by

$$\mathrm{CI}_{1-\alpha}\left(\overline{Y}\right) = \left[\overline{Y} - q_{1-\frac{\alpha}{2}}\cdot \mathrm{se}\left(\overline{Y}\right); \overline{Y} + q_{1-\frac{\alpha}{2}}\cdot \mathrm{se}\left(\overline{Y}\right)\right]\ .$$

For a 95% confidence interval, we have:

$$\mathrm{CI}_{0.95}\left(\overline{Y}\right) = \left[\overline{Y} - 1.96\cdot \mathrm{se}\left(\overline{Y}\right); \overline{Y} + 1.96\cdot \mathrm{se}\left(\overline{Y}\right)\right] = \left[58 - 1.96\cdot\frac{8}{10}; 58 + 1.96\cdot\frac{8}{10}\right] = [56.342; 59.568]\ .$$

In other words, if we formulated a null hypothesis with any value in $[56.342; 59.568]$ against a two-sided alternative, we would not be able to reject it.

3 Monte Carlo Simulations

To illustrate the concepts introduced above, we will perform a Monte Carlo exercise as follows. Suppose we have a sample of cities that have the status of an enterprise zone, and we are interested in answering the question, whether there are any effects of having this status on investment.

Let Y_i denote the percent change in investment from the year before the status was granted to the year after. The hypothesis of interest is then $H_0 : \mu_Y = 0$ vs. $H_1 : \mu_Y \neq 0$. We will simulate $Y_i \sim N(2, 1)$.

We will fix significance level at $\alpha = 0.05$ and perform $T = 10{,}000$ simulations. For each simulation, we will generate a random sample Y_i, $i = 1, \ldots, n$, and test the above hypothesis using the t-statistic $t = \frac{\overline{Y}-2}{s_Y/\sqrt{n}}$. We expect the percentage of incorrectly rejected null hypotheses to be close to α. Also, for each sample, we will compute a $(1-\alpha)$-confidence interval and determine whether it covers $\mu_Y = 2$. We expect the percentage of samples, for which this is true, to be close to $1 - \alpha$.

For $n = 10$, we obtain the following result:

- H_0 is rejected 6.37% of the time;
- confidence intervals cover the true mean 93.63% of the time.

These numbers are reasonably close to the expected figures, even though we used the sample standard deviation to estimate the population standard deviation.

Increasing the sample size to $n = 1000$, we have:

- H_0 is rejected 5.20% of the time;
- confidence intervals cover the true mean 94.8% of the time.

These values are even closer to the theoretically expected.

The R code for this exercise is given in Listing 1.

Listing 1: R code used to perform Monte Carlo simulations

```
# Function performing a Monte Carlo simulation
# for a data sample of size n with T repetitions
monte.carlo <- function(n, T, mu, sd, alpha){
  print(paste0("Sample size: ", n))

  # allocate memory
  reject.H0 <- rep(0, T)
  cover.CI <- rep(0, T)

  # get critical value
  t.critical <- qnorm(alpha/2)

  for (i in 1:max.T){
    # generate a normal sample with mean and variance
    X <- rnorm(n, mu, sd)

    # compute sample mean and standard deviation
    X.mean <- mean(X)
    X.sd <- sd(X)
    X.se <- sqrt(X.sd / n)

    # compute test statistics for sample mean
    t <- (X.mean - mu) / X.se

    # perform hypothesis test
    reject.H0[i] <- 1*(abs(t) > abs(t.critical))

    # compute confidence interval
    CI.a <- X.mean - abs(t.critical)*X.se
```

```
    CI.b <- X.mean + abs(t.critical)*X.se

    # determine if the CI covers the true mean
    cover.CI[i] <- 1*(mu >= CI.a && mu <= CI.b)
  }

  # report percentage of (incorrectly) rejected null hypotheses
  print(
    sprintf(
      "Percent of (incorrectly) rejected null hypotheses: %0.4f", mean(reject.H0)
    )
  )

  # report percentage of intervals that covered the true mean
  print(
    sprintf(
      "Percent of confidence intervals that cover the true mean: %0.4f", mean(cover.CI)
    )
  )
}


# set random number generator seed for reproducibility
set.seed(100)

# set working directory to the current file
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

# set constants
max.T <- 10000
alpha <- 0.05
# DGP:
mu <- 2
sd <- 1

# run Monte Carlo simulations with different sample sizes
monte.carlo(
  n = 10, T = max.T,
  mu = mu, sd = sd,
  alpha = alpha
)
monte.carlo(
  n = 1000, T = max.T,
  mu = mu, sd = sd,
  alpha = alpha
)
```