Lecture 16
The Concept of Hypothesis Testing

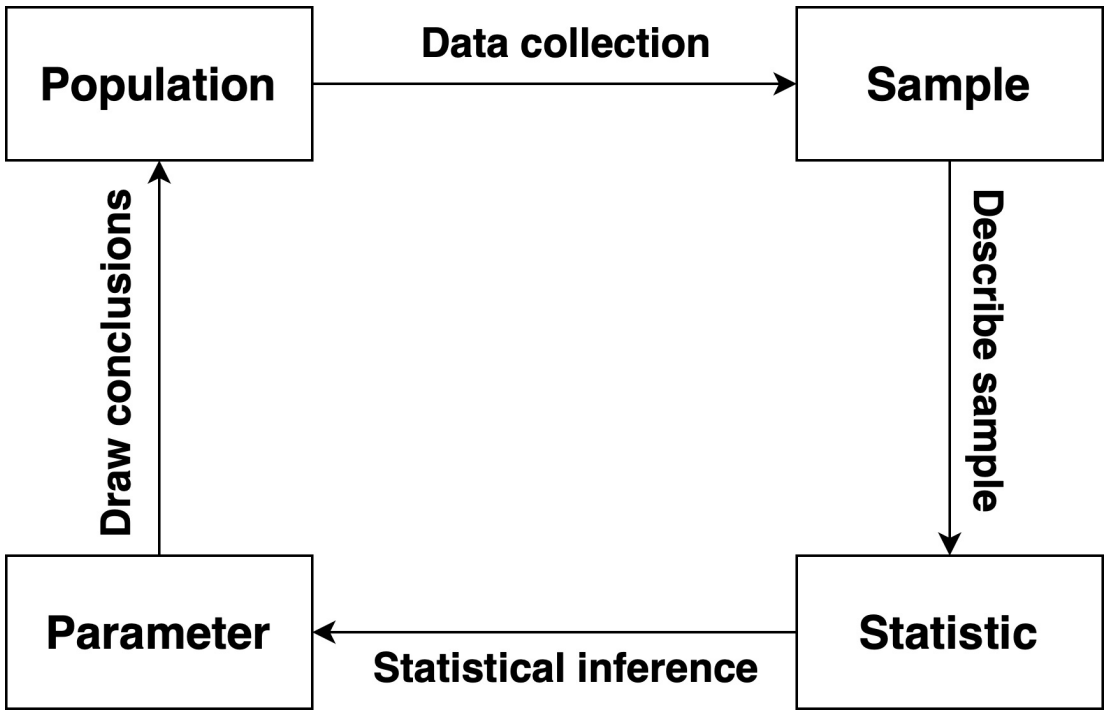
Chao Song

College of Ecology
Lanzhou University

November 18, 2024

What is hypothesis testing?

Parameter estimation and **hypothesis testing** are two essential tasks of statistics. Roughly speaking, hypothesis testing aims at reaching a decision about whether or not we reject a hypothesis about the value of the parameter.



Hypothesis testing: a motivating example

Example: A new manufacturing process was developed to make a ceramic ball precisely with 1 cm diameter. To test this new process, the factory made 10 ceramic balls with this new process. The mean diameter of the 10 balls was 1.45 cm and the standard deviation is 0.1 cm. Let's assume that the diameter of the ball follows a normal distribution. How would you decide whether this new manufacturing process attains its goal?

To address this problem, we need the following elements:

- A hypothesis about the diameter of the ball we aim at testing;
- Another hypothesis that include all scenarios except the one above;
- A statistic we can compute from the sample measurements;
- A decision rule based on the value of the statistics

Hypothesis testing: a motivating example

We form two competing hypotheses:

H_0: the mean diameter of the ball $\mu = 1$ cm.

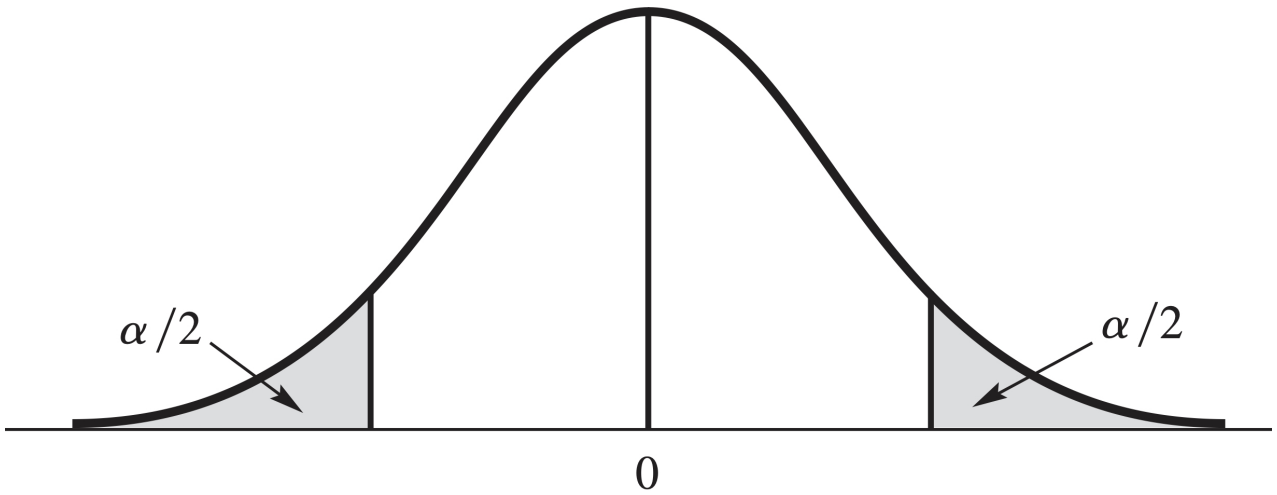
H_A: the mean diameter of the ball $\mu \neq 1$ cm.

If H_0 is true, $\overline{X}$ should not be too different from 1 cm. Given that X has a normal distribution, $\frac{\overline{X}-\mu}{s/\sqrt{n}}$ follows a t-distribution with $n-1$ degrees of freedom. Since $t_{0.025}(9) = 2.26$, we know that the statistic $T = \frac{\overline{X}-\mu}{s/\sqrt{n}}$ would be between -2.26 and 2.26 95% of times a random sample is drawn. We thus can devise a rule that we will reject H_0 if $T \leqslant -2.26$ or $T \geqslant 2.26$. Now $\frac{\overline{X}-\mu}{s/\sqrt{n}} = 14.23$, we thus reject H_0.

Hypothesis testing: a motivating example

In this motivating example,

- $T = \frac{\overline{X}-\mu}{s/\sqrt{n}}$ is referred to as the **test statistic**, whose distribution is known under the null hypothesis;
- $(-\infty, -2.26)$ and $(2.26, \infty)$ is called the **rejection region**.
- The probability that the test statistic falls in the rejection region under the null hypothesis is called the **significance level** and is often denoted α.



(Illustration of the rejection region)

Hypothesis testing: a motivating example

We make a decision about rejecting or not rejecting the null hypothesis based on whether the test statistic falls in the rejection region or not. We can also quantify how incompatible the sample measurements are with the null hypothesis.

How likely is it to obtain the value of the test statistic that deviates as much as the current sample or even more from the expectation of the null hypothesis?

$$P(T \leqslant -14.23 \text{ or } T \geqslant 14.23) = 1.78 \times 10^{-7}$$

This tail-end probability is called the **p-value**. It is the probability of obtaining as incompatible or more incompatible data if the null hypothesis is true. We make decision by comparing the p-value and the significance level α.

Elements of a hypothesis test

A hypothesis test contains the following elements:

- Null hypothesis, H_0
- Alternative hypothesis, H_A
- Test statistic
- Rejection region

In hypothesis testing, we compute a statistic from the sample measurements assuming the null hypothesis is true, then we examine if its value falls in a region that is compatible or incompatible with the null hypothesis. If it is incompatible, i.e., the value falls in the rejection region, we reject the null hypothesis. Otherwise, we do not reject the null hypothesis.

Hypothesis testing is based on the simple principle that “small chance event should not occur in reality”.

p-value

Small p-value implies that the true parameter is very likely different from the hypothesized value, i.e., the difference is statistically significant. However, it does not imply the magnitude of the difference. **Statistical significance** is not equivalent to **practical significance**.

- P-value typically decreases with sample size. Unless the true parameter value is exactly equal to the hypothesized value, we can detect any minor differences between the true value and the hypothesized one as long as sample size is large enough;
- In addition to reporting p-values, we should also report descriptive statistics of the sample.

Null and alternative hypothesis

Null hypothesis should be specific and usually specifies the values of the parameters of interest. This is necessary because we need to know the distribution of the statistic under the null hypothesis.

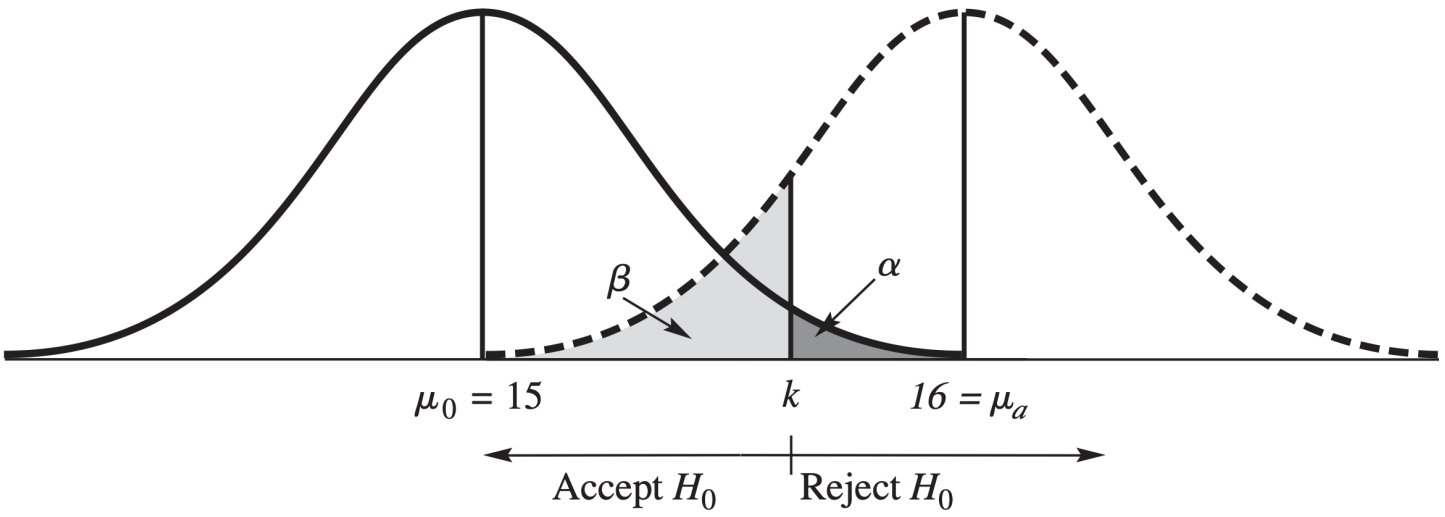
$\times$ H_0: the mean diameter of the ball $\mu < 1.5$ cm.
✓ H_0: the mean diameter of the ball $\mu = 1.5$ cm.

Alternative hypothesis typically include all possibilities except the one specified in the null hypothesis. Therefore, it is usually in the form of parameter values not equal to the values specified in the null hypothesis.

Error associated with hypothesis testing

Rejection region is a range of values that are **improbable but not impossible** if the null hypothesis is true. Thus, the decision about rejecting or not rejecting H_0 always has the chance of committing an error.

	Not reject H_0	Reject H_0
H_0 is true	Correct	Type I error (α)
H_0 is false	Type II error (β)	Correct



Error associated with hypothesis testing

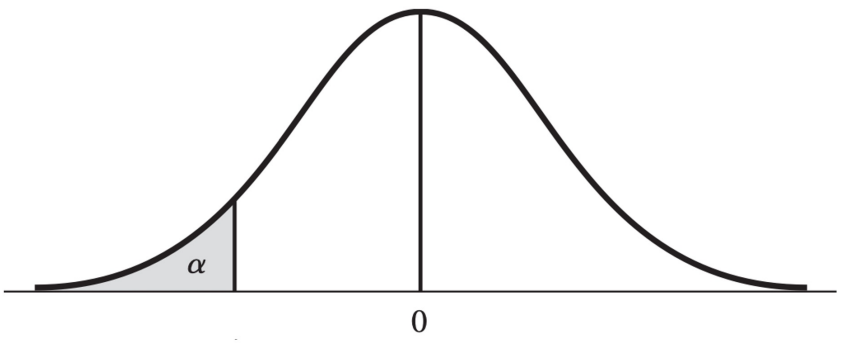
Comments on errors associated with hypothesis testing:

- Type I and II errors are both conditional probabilities;
- P-value is equal to type I error. We can control the probability of committing type I error by setting α.
- Type II error is typically unknown. Thus, conventionally, Instead of accepting H_0, we do not reject it or fail to reject it.
- The probability of rejecting a false H_0, $1 - \beta$, is called the statistical power of the test. Power depends on the type of statistical test, the significance level, the sample size, and the true value of the parameters.
- For a given data set and a given test, α and β are inversely related. We want to choose a test that minimizes β when α is fixed at some level.

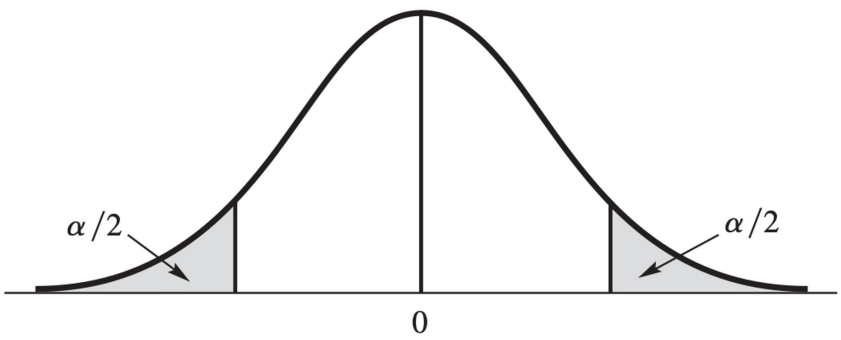
One-tailed vs two-tailed test

For H_0: $\mu = \mu_0$, H_A most commonly includes all possibilities except H_0, i.e., H_A: $\mu \neq \mu_0$. We reject H_0 when the test statistic is either much smaller or much greater than expected under H_0. We call such a test a **two-tailed test**.

If we are only interested in detecting $\mu < \mu_0$, we only reject H_0 when the test statistic is much smaller than expected under H_0. Similarly, if we only want to detect $\mu > \mu_0$, we reject H_0 when the test statistic is much greater than expected under H_0. We call such a test a **one-tailed test**.



One-tailed test



Two-tailed test

One-tailed vs two-tailed test

Comments on one-tailed and two-tailed tests:

- Given the same data, the p-value from the one-tailed test is smaller than the two-tailed the test;
- The choice between one-tailed and two-tailed tests depends on the alternative value of the parameter that the experimenter is trying to detect. Never chose a test based on what the p-value is.
- Unless you have a compelling reason, we use a two-tailed test in practice for most of the time.