

# Statistics

## Random Sample

The collection of random variables  $X_1, X_2, X_3, \dots, X_n$  is said to be a random sample of size  $n$  if they are independent and identically distributed (i.i.d.), i.e.,

1.  $X_1, X_2, X_3, \dots, X_n$  are independent random variables, and
2. they have the same distribution, i.e.,

$$F_{X_1}(x) = F_{X_2}(x) = \dots = F_{X_n}(x), \quad \text{for all } x \in \mathbb{R}.$$

## Properties of Random Samples

1. the  $X_i$  's are independent;
2.  $F_{X_1}(x) = F_{X_2}(x) = \dots = F_{X_n}(x) = F_X(x)$ ;
3.  $EX_i = EX = \mu < \infty$ ;
4.  $0 < \text{Var}(X_i) = \text{Var}(X) = \sigma^2 < \infty$ .

## Sample Mean

The sample mean is defined as

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Properties of the sample mean

1.  $E\bar{X} = \mu$ .
2.  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ .
3. Weak Law of Large Numbers (WLLN):

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0.$$

4. Central Limit Theorem: The random variable

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable as  $n$  goes to infinity, that is

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x), \quad \text{for all } x \in \mathbb{R}$$

where  $\Phi(x)$  is the standard normal CDF.

## Sample Variance and Standard Deviation

Let  $X_1, X_2, X_3, \dots, X_n$  be a random sample with mean  $EX_i = \mu < \infty$ , and variance  $0 < \text{Var}(X_i) = \sigma^2 < \infty$ . The sample variance of this random sample is defined as

$$S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2 = \frac{1}{n-1} \left( \sum_{k=1}^n X_k^2 - n\bar{X}^2 \right).$$

The sample variance is an unbiased estimator of  $\sigma^2$ . The sample standard deviation is defined as

$$S = \sqrt{S^2},$$

and is commonly used as an estimator for  $\sigma$ . Nevertheless,  $S$  is a biased estimator of  $\sigma$ .

## Confidence Intervals

Let  $X_1, X_2, X_3, \dots, X_n$  be a random sample from a distribution with a parameter  $\theta$  that is to be estimated. An interval estimator with confidence level  $1 - \alpha$  consists of two estimators  $\hat{\Theta}_l(X_1, X_2, \dots, X_n)$  and  $\hat{\Theta}_h(X_1, X_2, \dots, X_n)$  such that

$$P(\hat{\Theta}_l \leq \theta \leq \hat{\Theta}_h) \geq 1 - \alpha,$$

for every possible value of  $\theta$ . Equivalently, we say that  $[\hat{\Theta}_l, \hat{\Theta}_h]$  is a  $(1 - \alpha)100\%$  confidence interval for  $\theta$ .

## Finding Confidence Intervals

Let's review a simple fact from random variables and their distributions. Let  $X$  be a continuous random variable with CDF  $F_X(x) = P(X \leq x)$ . Suppose that we are interested in finding two values  $x_l$  and  $x_h$  such that

$$P(x_l \leq X \leq x_h) = 1 - \alpha.$$

One way to do this, is to choose  $x_l$  and  $x_h$  such that

$$P(X \leq x_l) = \frac{\alpha}{2}, \quad \text{and} \quad P(X \geq x_h) = \frac{\alpha}{2}.$$

Equivalently,

$$F_X(x_l) = \frac{\alpha}{2}, \quad \text{and} \quad F_X(x_h) = 1 - \frac{\alpha}{2}.$$

We can rewrite these equations by using the inverse function  $F_X^{-1}$  as

$$x_l = F_X^{-1}\left(\frac{\alpha}{2}\right), \quad \text{and} \quad x_h = F_X^{-1}\left(1 - \frac{\alpha}{2}\right).$$

We call the interval  $[x_l, x_h]$  a  $(1 - \alpha)$  interval for  $X$ . Figure 1 shows the values of  $x_l$  and  $x_h$  using the CDF of  $X$ , and also using the PDF of  $X$ .

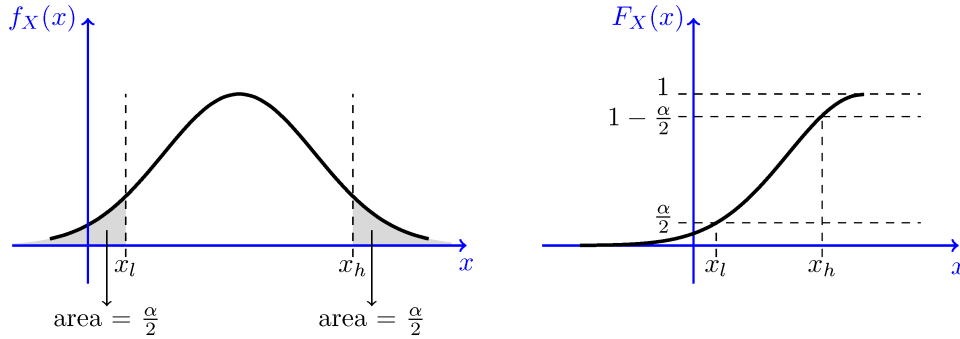


Figure 1:  $[x_l, x_h]$  is a  $(1 - \alpha)$  interval for  $X$ , that is,  $P(x_l \leq X \leq x_h) = 1 - \alpha$ .

## The U-test

The U-test, also known as the Mann-Whitney U test, is a non-parametric test used to determine whether there is a significant difference between the distributions of two independent samples. It is often used as an alternative to the t-test when the assumption of normality is not met.

### Key Concepts

- **Non-Parametric Test:** Unlike parametric tests, the U-test does not assume a specific distribution (e.g., normal distribution) for the data. This makes it suitable for data that do not meet the assumptions of parametric tests.
- **Independent Samples:** The test compares two independent groups. Independence means the samples are not related or paired.
- **Rank-Based:** The test works by ranking all the values from both groups together and then comparing the ranks between the groups.

### Steps to Perform the U-Test

1. **Combine and Rank Data:**
  - Combine the data from both samples.
  - Rank all the observations from the lowest to the highest, assigning ranks. In the case of ties (identical values), assign the average rank to the tied values.
2. **Sum of Ranks:**

- Calculate the sum of the ranks for each group. Let's denote these sums as  $R_1$  and  $R_2$  for the first and second group, respectively.

### 3. Calculate U Values:

- The U-test computes two U values, one for each group:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

Where  $n_1$  and  $n_2$  are the sample sizes of the two groups.

- The smaller of the two U values is used for the test statistic:

$$U = \min(U_1, U_2)$$

### 4. Determine Significance:

- Compare the calculated U value to a critical value from the Mann-Whitney U distribution table (which depends on the sample sizes and the chosen significance level).
- Alternatively, for larger samples, a normal approximation can be used:

$$Z = \frac{U - \mu_U}{\sigma_U}$$

Where  $\mu_U$  and  $\sigma_U$  are given by:

$$\mu_U = \frac{n_1 n_2}{2}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

- The Z value can then be compared to the standard normal distribution to determine the p-value.

### Interpretation

- **Null Hypothesis** ( $H_0$ ): The distributions of both groups are equal.
- **Alternative Hypothesis** ( $H_A$ ): The distributions of the two groups are not equal.

If the test statistic (U or Z) indicates a significant difference (p-value < significance level), you reject the null hypothesis and conclude that there is a significant difference between the two groups.