Prob and Stat for ECEs

Aidan Sharpe

Contents

Chapter 1

Probability Density and Cumulative Distribution Functions

Example 1.0.1

Suppose there are 30 resistors, 7 of them do not work. You randomly choose 3 of them. Let X be the number of defective resistors. Find the probability distribution of X.

$$X = [0, 3]$$

$$P(X = 0) = \frac{\binom{7}{0}\binom{23}{3}}{\binom{30}{3}} = 0.436$$

$$P(X = 1) = \frac{\binom{7}{1}\binom{23}{2}}{\binom{30}{3}} = 0.436$$

$$P(X = 2) = \frac{\binom{7}{2}\binom{23}{1}}{\binom{30}{3}} = 0.119$$

$$P(X = 3) = \frac{\binom{7}{3}\binom{23}{0}}{\binom{30}{3}} = 0.009$$

Probability distribution:

$$P(X = x) = \begin{cases} 0.436 & x = 0 \\ 0.436 & x = 1 \\ 0.119 & x = 2 \\ 0.009 & x = 3 \end{cases}$$

Definition 1.0.1: The Cumulative Distribution Function

The cumulative distribution function (CDF), $F(x)$, of a discrete random variable, x, with probability distribution, $f(x)$, is:

$$F(x) = P(X \leq x)$$

Find CDF for the example above:

$$F(0) = P(X \leq 0) = P(X = 0) = 0.436$$

$$F(1) = P(X \leq 1) = P((X = 0) \cup (X = 1)) = 0.872$$

$$F(2) = P(X \leq 2) = P((X = 0) \cup (X = 1) \cup (X = 2)) = 0.991$$

Since 3 is the largest possible value for x:

$$F(3) = P(X \leq 3) = 1$$

As a piecewise function:

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.436 & 0 \leq x < 1 \\ 0.872 & 1 \leq x < 2 \\ 0.991 & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

Example 1.0.2

Suppose that a days production of 850 manufactured parts contains 50 parts that to not conform to customer requirements. 2 parts are selected at random from the batch. Let X be the number of non-conforming parts.

a) Find the probability distribution for X:

$$P(X = 0) = \frac{\binom{50}{0}\binom{800}{2}}{\binom{850}{2}} = 0.8857$$

$$P(X = 1) = \frac{\binom{50}{1}\binom{800}{1}}{\binom{850}{2}} = 0.1109$$

$$P(X = 2) = \frac{\binom{50}{2}\binom{800}{0}}{\binom{850}{2}} = 0.0034$$

$$P(X = x) = \begin{cases} 0.8857 & x = 0 \\ 0.1109 & x = 1 \\ 0.0034 & x = 2 \end{cases}$$

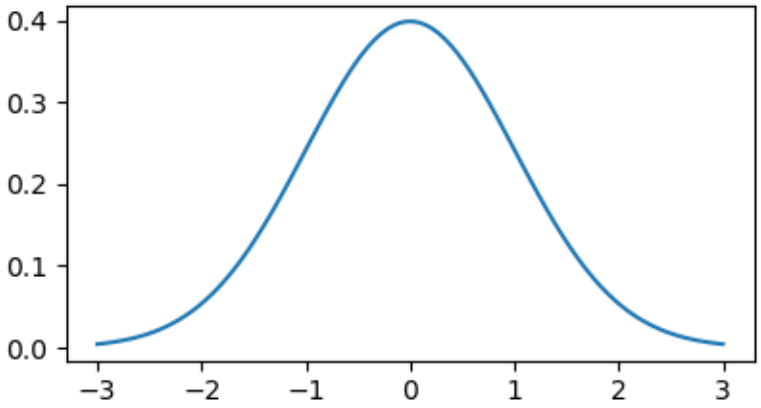
b) Find the CDF $F(x)$:

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.8857 & 0 \leq x < 1 \\ 0.9966 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

c) Plot $F(x)$:

1.1 Continuous Probability Distributions

A continuous random variable is a variable that can take on any value within a range. It takes on infinitely many possible value within the range.



For a continuous distribution, $f(x)$:

$$P(X = x) = 0$$

$$P(x_0 \leq X \leq x_1) = \int_{x_0}^{x_1} f(x)dx$$

$$P(X \geq x_0) = \int_{x_0}^{\infty} f(x)dx$$

Definition 1.1.1

The function, $f(x)$, is a probability density function fo the continuous random variable, X, defined over if:

1. $$f(x) \geq 0, \forall x \in$$

2. $$\int_{-\infty}^{\infty} f(x)dx = 1$$

3. $$\begin{aligned} P(x_0 \leq X \leq x_1) &= P(x_0 < X < x_1) \\ &= P(x_0 \leq X < x_1) \\ &= P(x_0 < X \leq x_1) \end{aligned}$$

Example 1.1.1

Suppose that the error in the reaction temperature in °C for a controlled lab experiment is a continuous random variable, X, having PDF:

$$f(x) = \begin{cases} \frac{x^2}{3} & -1 < x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

a) Verify that $f(x)$ is a PDF.

$$\int_{-1}^{2} \frac{x^2}{3} dx \stackrel{?}{=} 1$$

$$\frac{1}{3}\left[\frac{1}{3}x^3\Big|_{-1}^{2}\right] = \frac{1}{9}[8-(-1)] = 1$$

b) Find $P(0 < X < 0.5)$:

$$P(0 < X < 0.5) = \int_0^{0.5} \frac{x^2}{3}dx$$

$$\frac{1}{9}\left[x^3\Big|_0^{0.5}\right] = \frac{1}{9}[0.125] = 0.01389$$

Definition 1.1.2

The CDF, $F(x)$ of a continuous random variable, X, with probability density function $f(x)$ is:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Note:

1. $$P(a < X < b) = F(b) - F(a)$$

2. $$f(x) = \frac{d}{dx}F(x)$$

Example 1.1.2

Find the CDF of the previous example

$$f(x) = \begin{cases} \frac{x^2}{3} & -1 < x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

$$F(x) = \int_{-1}^{x} \frac{t^2}{3}dt$$

$$1/over9\left[t^3\Big|_{-1}^{x}\right] = \frac{1}{9}\left[x^3+1\right]$$

$$F(x) = \begin{cases} 0 & t < -1 \\ \frac{1}{9}\left[x^3+1\right] & -1 \leq x \leq 2 \\ 1 & \text{elsewhere} \end{cases}$$

Example 1.1.3

The proportion of the budget for a certain type of industrial company that is allotted to environmental and pollution control is coming under scrutiny. A data collection project determines that the distribution of these proportions is given by:

$$f(y) = \begin{cases} k(1-y)^4 & 0 \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find k that renders $f(y) a valid density function:

$$\int_0^1 k(1-y)^4 dy = 1$$

$$\frac{k}{5} = 1$$

$$\therefore k = 5$$

Chapter 2

Expected Value

Definition 2.0.1: Expected Value

Let X be a random variable with probability distribution $f(x)$. The mean, or expected value, of X is:
For a discrete distribution

$$E[X] = \sum_x xf(x)$$

For a continuous distribution:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Given $\{1, 2, 3, 3, 5\}$, the mean is:

$$\frac{1+2+3+3+5}{5} = 2.8$$

$$f(x) = \begin{cases} \frac{1}{5} & x = 1 \\ \frac{1}{5} & x = 2 \\ \frac{2}{5} & x = 3 \\ \frac{1}{5} & x = 5 \end{cases}$$

$$\sum_x xf(x) = \frac{1}{5}(1) + \frac{1}{5}(2) + \frac{1}{5}(3) + \frac{1}{5}(5) = 2.8$$

Example 2.0.1

The probability distribution of a discrete random variable X is:

$$f(x) = \binom{3}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{3-x}, x \in \{0, 1, 2, 3\}$$

Find $E[X]$:

$$f(x) = \begin{cases} 0 & x = 0 \\ 0.422 & x = 1 \\ 0.14 & x = 2 \\ \frac{1}{64} & x = 3 \end{cases}$$

$$E[X] = \sum_x x\binom{3}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{3-x}$$

$$E[X] = 0(0) + 0.422(1) + 0.14(2) + \frac{1}{64}(3) = 0.75$$

Example

Let X be the random variable that denotes the life in hours of a certain electronic device. The PDF is:

$$f(x) = \begin{cases} \frac{20000}{x^3} & x > 100 \\ 0 & elsewhere \end{cases}$$

Find the expected life of this type of device:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_{100}^{\infty} x \frac{20000}{x^3} dx = 200[\text{hrs}]$$

Note:

$$E[x^2] = \int_{\infty}^{\infty} x^2 f(x) dx$$

Properties of Expectations

$$E(b) = b$$

Where b is a constant

$$E(aX) = aE[X]$$

Where a is a constant

$$E(aX + b) aE[X] + b$$

$$E[X + Y] = E[X] + E[Y]$$

Where X and Y are random variables

Example

Given:

$$f(x) = \begin{cases} \frac{x^2}{3} & -1 < x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

Find the expected value of $Y = 4X + 3$:

$$E[Y] = E[4X + 3] = 4E[X] + 3$$

$$E[X] = \int_{-1}^{3} \frac{X^3}{3} dx = \frac{1}{12} X^4 \Big|_{-1}^{3} = \frac{5}{4}$$

Variance of a Random Variable

The expected value/mean is of special importance because it describes where the probability distribution is centered. However, we also need to characterize the variance of the distribution.

Definition

Let X be a random variable with probability distribution, $f(x)$, and mean, μ. The variance of X is given by:

$$\text{Var}[X] = E[(X - \mu)^2]$$

Which is the average squared distance away from the mean. This simplifies to:

$$\text{Var}[X] = E[X^2] - E[X]^2$$

Note: Generally,

$$E[X^2] \neq E[X]^2$$

The standard deviation, σ, is given by:

$$\sigma = \sqrt{\text{Var}[X]}$$

Note: The variance is a measure of uncertainty (spread) in the data.

Example

The weekly demand for a drinking water product in thousands of liters from a local chain of efficiency stores is a continuous random variable, X, having the probability density:

$$F(x) = \begin{cases} 2(x-1) & 1 < x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

Find the expected value:

$$E[X] = \int_1^2 x(2(x-1))dx = 2\int_1^2 (x^2 - x)dx$$

$$E[X] = 2\left[\frac{1}{3}x^3 - \frac{1}{2}x^2\Big|_1^2\right] = \frac{5}{3}$$

Find the variance:

$$\text{Var}[X] = E[X^2] - E[X]^2$$

$$E[X^2] = \int_1^2 2x^2(x-1)dx = 2\int_1^2 (x^3 - x^2)dx$$

$$E[X^2] = \frac{17}{6}$$

$$\text{Var}[X] = \frac{17}{6} - \left(\frac{5}{3}\right)^2 = \frac{1}{18}$$

Find the standard deviation:

$$\sigma = \sqrt{\text{Var}[X]} = \frac{1}{3\sqrt{2}} = \frac{\sqrt{2}}{6}$$

Example

The mean and variance are useful when comparing two or more distributions.

[]@lll@ () Plan 1 Plan 2
() Avg Score Improvement +17 +15
Standard deviation ±8 ±2
()

Theorem

If X has variance, $\text{Var}[X]$, then $\text{Var}[aX + b] = a^2\text{Var}[X]$.

Example

The length of time, in minutes, for an airplane to obtain clearance at a certain airport is a random variable, $Y = 3X - 2$, where X has the density:

$$F(x) = \begin{cases} \frac{1}{4}e^{x/4} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$E[X] = 4$$

$$\text{Var}[X] = 16$$

Find $E[Y]$:

$$E[Y] = E[3X - 2] = 3E[X] - 2 = 10$$
$$\text{Var}[Y] = 3^2\text{Var}[X] = 144$$
$$\sigma = \sqrt{\text{Var}[Y]} = 12$$

2.0.1 The Exponential Distribution

The continuous random variable, X, has an exponential distribution with parameter β if its density function is given by:

$$F(x) = \begin{cases} \frac{1}{\beta}e^{-x/\beta} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Where $\beta > 0$.

$$E[X] = \beta$$
$$E[X] = \int_0^\infty x\frac{1}{\beta}e^{-x/\beta}dx$$

Aside:

$$\Gamma(Z) = \int_0^\infty x^{Z-1}e^{-x}dx$$

Where $\Gamma(Z) = (Z-1)!$

$$E[X] = \beta\int_0^\infty \left(\frac{x}{\beta}\right)^{(2-1)} e^{-x/\beta}\left(\frac{dx}{\beta}\right) = \beta\Gamma(2)$$
$$E[X] = \beta(2-1)! = \beta$$

$$\text{Var}[X] = E[X^2] - E[X]^2$$
$$E[X^2] = \int_0^\infty x^2\frac{1}{\beta}e^{-x/\beta}dx = \beta^2\int_0^\infty \left(\frac{x}{\beta}\right)^{(2-1)} e^{-x/\beta}\left(\frac{dx}{\beta}\right)$$
$$E[X^2] = \beta^2\Gamma(3) = 2\beta^2$$
$$\text{Var}[X] = 2\beta^2 - \beta^2 = \beta^2$$

Application Reliability analysis: the time to failure of a certain electronic component can be modeled by an exponential distribution.

Example

Let T be the random variable which measures the time to failure of a certain electronic component. Suppose T has an exponential distribution with $\beta = 5$.

$$F(x) = \begin{cases} \frac{1}{5}e^{-x/5} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

If 6 of these components are in use, what is the probability that exactly 3 components are still functioning at the end of 8 years?

What is the probability that an individual component is still functioning after 8 years?

$$P(T > 8) = \int_8^\infty \frac{1}{5} e^{-x/5} dx \approx 0.2$$

$$\binom{6}{3}(0.2)^3(0.8)^3 = 0.08192$$

[] >>> from math import comb >>> comb(6,3) * 0.2**3 * 0.8**3 0.08192000000000003

2.0.2 The Normal Distribution

The most important continuous probability distribution in the field of statistics is the normal distribution. It is characterized by 2 parameters, the mean, μ, and the variance, σ^2.

$$\text{mean} = \text{median} = \text{mode}$$

$$F(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{\left(\frac{1}{2\sigma^2}(x-\mu)^2\right)}$$

$$E[X] = \mu$$

$$\text{Var}[X] = \sigma^2$$

For a normal curve:

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} F(x)dx$$

Definition

The distribution of a normal variable with mean 0 and variance 1 is called a standard normal distribution.

The transformation of any random variable, X into a standard normal variable, Z:

$$Z = \frac{X - \mu}{\sigma}$$

Example

Given a normal distribution with mean $\mu = 30$ and standard deviation, $\sigma = 6$, find the normal curve area to the right of $x = 17$.

Transform to standard normal.

$$Z = \frac{17 - 30}{6} = -2.16$$

That is, $x = 17$ on a normal distribution with $\mu = 30$ and $\sigma = 6$ is equivalent to $Z = -2.16$ on a normal distribution with $\mu = 0$ and $\sigma = 1$.

$$P(X > 17) = P(Z > -2.16)$$

$$P(Z > -2.16) = 1 - P(Z \leq -2.16) = 0.9846$$

[] >>> from scipy.stats import norm >>> norm.cdf(-2.16) 0.015386334783925445

Example

The finished inside diameter of a piston ring is normally distributed with mean, $\mu = 10$[cm], and standard deviation, $\sigma = 0.03$[cm].

What is the probability that a piston ring will have inside diameter between 9.97[cm] and 10.03[cm]?

$$Z_1 = \frac{9.97 - 10}{0.03} = -1$$

$$Z_2 = \frac{10.03 - 10}{3} = 1$$

$$P(9.97 < x < 10.03) = 0.68$$

[] >>> from scipy.stats import norm >>> norm.cdf(1) - norm.cdf(-1) 0.6826894921370859

Chapter 3

Hypothesis Testing

There is a 1 to 1 relationship between the test of a hypothesis about any parameter, say, θ, and the confidence interval for θ.

Definition 3.0.1

If (LB, UB) is a 100(1-α)% confidence interval for the parameter, θ, the test of size α of the hypothesis:

H_0: $\theta = \theta_0$

H_a: $\theta \neq \theta_0$

will lead to a rejection of H_0 if and only if θ_0 is not in the 100(1-α)% confidence interval.

Example 3.0.1

Consider homework 6, problem 1. We had $\bar{x} = 664$ and $s = 500$. Test the hypothesis:

H_0: $\mu = 634$

H_a: $\mu \neq 634$

α: 0.05

The 95% confidence interval for μ:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$$t^* = 1.961$$

$$664 \pm 1.961 \frac{500}{\sqrt{1700}}$$

$$(640.22, 687.78)$$

Since the value indicated by H_0 (634) is not within the 95% confidence interval, it is not a plausible value for μ, and thus we reject H_0 at $\alpha = 0.05$.

Sample evidence suggests that the average number of social ties for a cell phone user is significantly different from 634.

3.0.1 Confidence Interval for the Difference of Two Means

If $\bar{x}_1$ and $\bar{x}_2$ are the means of independent random samples of size n_1 and n_2 from approximately normal populations with unknown but equal variances, a 100(1-α)% confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Where:

$t_{\frac{\alpha}{2}}$ is the t-value with $n_1 + n_2 - 2$ degrees of freedom

Example 3.0.2

Homework 7, problem 3: burn time for fuses.

Supplier A	Supplier B
$n_1 = 30$	$n_2 = 30$
$\bar{x}_1 = 30.62$	$\bar{x}_2 = 31.37$
$s_1^2 = 0.384$	$s_2^2 = 0.185$

Does the sample suggest that the mean burn time for supplier A is different than that for supplier B? Use $\alpha = 0.05$.

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

The 95% confidence interval for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p^2 = \frac{(30 - 1)0.384 + (30 - 1)0.185}{30 + 30 - 2} = 0.2845$$

$$\therefore s_p = 0.5334$$

$$t_{\frac{\alpha}{2}} = 2.002$$

$$(30.62 - 31.37) \pm 2.002(0.5534)\sqrt{\frac{1}{30} + \frac{1}{30}}$$

$$(-1.02, -0.47)$$

We are 95% confident that the difference of mean burn time between supplier A and supplier B is somewhere between -1.02 and -0.47.

Since 0 is not in the confidence interval, we reject H_0. Sample evidence suggests that the mean burn time for supplier A is different than that for supplier B.

Note:

Since the entire confidence interval is below zero, we conclude with 95% confidence that $\mu_1 - \mu_2 < 0$ and by extension, $\mu_1 < \mu_2$.

The usefulness of using confidence intervals for significance testing:

- A confidence interval provides information about the magnitude and direction of the difference between μ_1 and μ_2.
- However, a hypothesis test does not provide such information. It only provides information about significance.

3.0.2 Paired t-test

Comparing two treatments where observations occur in pairs or are related to each other.

Example 3.0.3

Two teaching methods to be compared by using 50 students divided into two equal classes.

Method 1:
Randomly assign 25 students to each class and compare average scores when experiment is concluded.

What if one group gets better students? It would no longer be a fair comparrison of the two methods. In this case, there would be two sources of variation:

1. Due to teaching method
2. Due to differences between students

This inflates the variance and leads to lower power.

Possible solution:
Pair students according to preference/ability. This would have mainly variance due to teaching method.

Example 3.0.4

10 adult males between the ages of 35 and 50 participated in a study to evaluate the effect of diet and exercise on blood cholesterol levels.
The total cholesterol was measured in each subject initially and three months after.

Subject	Before	After	Difference
1	265	229	36
2	240	231	9
3	258	227	31
4	295	240	55
5	251	238	13
6	245	241	4
7	287	234	53
8	314	256	58
9	260	247	13
10	279	239	40

Run a one-sample t-test on the differences:

H_0: $\mu_d = 0$

H_a: $\mu_d \neq 0$

Test statistic:

$$t^* = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

Do the data support the claim that the diet and exercise are of value in production of a mean reduction in blood cholesterol levels using $\alpha = 0.05$?

$$t^* = \frac{31.2}{\frac{20.43}{\sqrt{n}}} = 4.829$$

The associated p-value is very close to 0. Since the p-value less than α, we reject H_0. Sample evidence strongly suggests that diet and exercise are of value in producing an effect in blood cholesterol levels.

Note:

With the pairing approach, we have our degrees of freedom compared to the two sample approach.

However, if paired observations are highly similar or related, the reduction in variance more than compensates for the loss of degrees of freedom.

3.1 Analysis of Variance

Consider the problem of deciding whether observed differences among more than two sample means can be attributed to chance, or whether there are real differences among the populations being sampled.

H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$

H_a: at least one mean differs.

Example 3.1.1

Consider the following observations:

Group 1	Group 2	Group 3
77	72	76
81	58	85
71	84	82
76	66	80
80	70	88

Suppose our data can be written as:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Where:

ε_{ij} deviation from the group means

μ_u i^{th} group mean

Y_{ij} j^{th} observation from i^{th} group.

To infer if at least one μ_i differs from the others, compare the variance within the groups against the variance between the groups.

The sum of the observations in the i^{th} group:

$$y_{i\cdot} = \sum_{j=1}^{r} y_{ij}$$

The overall sum:

$$y_{\cdot\cdot} = \sum_{i=1}^{t} \sum_{j=1}^{r} y_{ij}$$

The mean of observations in the i^{th} group:

$$\bar{y}_{i\cdot} = \frac{1}{r} \sum_{j=1}^{r} y_{ij}$$

The overall mean:

$$\bar{y}_{\cdot\cdot} = \frac{1}{rt} \sum_{i=1}^{t} \sum_{j=1}^{r} y_{ij}$$

Decompose the observation:

$$y_{ij} - \bar{y}_{\cdot\cdot} = y_{ij} + \bar{y}_{i\cdot} - \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} = (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot})$$

The deviation of an observation from the grand mean is the same as the sum of the deviation of the treatment mean from the grand mean and the deviation of the observation from its treatment mean.

Theorem 3.1.1

$$\sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{..})^2 = r\sum_{i=1}^{t}(\bar{y}_{c}-\bar{y}_{..})^2 + \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{i.})^2$$

The variability of observations about the grand mean is the sum of the variability of treatment means about the grand mean and the variability of observations about their treatment means.
SSTotal = SSTreatment + SSError

The Anova Table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares
Treatment (between groups)	SSTreat	t - 1	$\frac{\text{SSTreat}}{t-1}$
Error (within groups)	SSE	$t(r-1)$	$\frac{\text{SSE}}{t(r-1)}$
Total	SSTotal	$tr-1$	

Note:

1. dfTotal = dfTreatment + dfError
2. σ^2 is estimated by MSE = $\frac{\text{SSE}}{t(r-1)}$ which is a pooled estimate of σ^2 from all the data in the experiment.

3.1.1 Tests of Hypotheses

H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$

H_a: at least one mean differs

Assumptions:

1. $E[\varepsilon_{ij}] = 0$
2. $\text{Var}[\varepsilon_{ij}] = \sigma^2$
3. $\text{Cov}[\varepsilon_{ij}, \varepsilon_{i'j'}] = 0$; $i \neq i'$, $j \neq j'$
4. ε_{ij} has a normal distribution

Test statistic:

$$F = \frac{\text{MSTreat}}{\text{MSE}} \sim F_{t-1,t(r-1)}$$

If H_0 is true:

$$F = \frac{\text{MSTreat}}{\text{MSE}} \approx 1$$

If H_a is true, $F_{\text{obs}} > 1$ and increases as treatment differences increase. P-value:

$$P(F_{t-1,t(r-1)} \geq F_{\text{obs}})$$

Example 3.1.2

Three types of signals wer utilized in a study to investigate traffic delay. Three types of traffic signals wer utilized in the study:

1. pretimesd
2. semi-actuated
3. fully actuated

Five intersections were used for each type of signal. The measure of traffic delay used in the study was the average stopped time per vehicle at each intersection. The data are given by:

Pretimed	Semi-Actuated	Fully Actuated
36.6	17.5	15.0
39.2	20.6	10.4
30.4	18.7	18.9
37.1	25.7	10.5
34.1	22.0	15.2

Compute the analysis of variance:

$$\bar{y}_{\cdot\cdot} = \frac{1}{15}(36.6 + 39.2 + \cdots + 15.2) = 23.46$$

$$\bar{y}_{1\cdot} = \frac{1}{5}(36.6 + 39.2 + 30.4 + 37.1 + 34.1) = 35.48$$

$$\bar{y}_{2\cdot} = 20.9$$

$$\bar{y}_{3\cdot} = 14$$

$$\text{SSTotal} = \sum_{i=1}^{3}\sum_{j=1}^{5}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = 1340.456$$

$$\text{SSTreat} = 5\sum_{i=1}^{3}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = 1202.626$$

$$\text{SSE} = \text{SSTotal} - \text{SSTreat} = 137.83$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean of Squares
Treatments	1202.626	3 - 1 = 2	601.313
Error	137.83	3(5-1) = 12	11.4858
Total	1340.456	14	

$$F_{\text{obs}} = \frac{\text{MSTreat}}{\text{MSE}} = \frac{601.313}{11.48} = 52.35$$

For such a high observed F, the p-value is very close to zero. Sample evidence suggests that the mean delays of the three types of traffic signals differ.

3.2 Non-Parametric Testing

Most of the hypothesis testing and confidence interval procedures have been based on the assumption that the samples are random from normally distributed populations. These are called parametric methods. Non-parametric or distribution-free methods make no assumptions about the distribution of the underlying population.

3.2.1 The Signed Rank Test

The only assumption is that the data is continuous and comes from a symmetric distribution.

1. Compute the differences $X_i - \mu_0, i = 1 \cdots n$.
2. Compute the absolute differences $|X_i - \mu_0|$ in ascending order
3. Compute w^+, the sum of the positive ranks, and w^-, the sum of the absulute negative ranks
4. The test statistic is given by:
$$w^{\text{observed}} = \min(w^-, w^+)$$

5. Use lookup table and reject H_0 if $w^{\text{observed}} \leq w^*_\alpha$.

For a one-sided test, if the alternative hypothesis is $\mu > \mu_0$, then $w^{\text{observed}} = w^-$. If the alternative hypothesis is $\mu < \mu_0$, then $w^{\text{observed}} = w^+$.

Example 3.2.1

A report on a study in which a rocket motor is formed by binding an igniter propellant and a sustainer propellant together inside a metal housing. The shear strength of the bond between the two types of propellant types is an important characteristic. The results of testing 10 randomly selected motors are shown below. Do the data suggest that the mean shear strength is different from 2000 psi using $\alpha = 0.05$.

Observation	X_i	$X_i - \mu_0$	Rank
1	2158.7	157.7	2
2	1678.15	-321.85	8
3	2316.00	316.00	7
4	2016.00	16.00	1
5	2207.5	207.5	3
6	1708.3	-291.70	6
7	1784.7	-215.3	4
8	2575.10	575.10	10
9	2357.9	357.90	8
10	2256.7	256.7	5

For this two-sided test:
H_0: $\mu = 2000$
H_a: $\mu \neq 2000$
The sum of the positive ranks:

$$w^+ = 2 + 7 + 1 + 3 + 10 + 9 + 5 = 37$$

The sum of the negative ranks:

$$w^- = 8 + 6 + 4 = 18$$

$$w^{\text{observed}} = \min(w^-, w^+) = 18$$

The critical value, w^*_α, is found in a lookup table. In this case it is 8. Since the observed test statistic is greater than this critical value, we fail to reject H_0. Sample evidence does not suggest that the mean shear strength is different from 2000 psi.

Where does w^*_α come from?