Probability and Distributions

Statistics deals with making inferences from experiments that involve uncertainties. For this, familiarity with **probability theory** is essential. The probability of occurrence of an event is the relative frequency of the event when the experiment is performed a sufficient number of times. If an event **R** can happen in **r** ways out of a total of **n** possible equally likely ways, then the probability of the occurrence of that event (called its **success**) can be denoted by **P(E) = r/n**.

The probability of non-occurrence of the event (called its **failure**) is denoted by

$P(\bar{E}) = (n - r)/n$

$P(E) + P(\bar{E}) = 1$

Sum of the probabilities in any **experiment** is 1.

P(E2|E1) is the **conditional probability** of E2 given that E1 has occurred.

If the occurrence or non-occurrence of E1 does not affect the probability of occurrence of E2, **P(E2|E1) = P(E2)**; E1 and E2 are said to be **independent events**.

Two or more events are said to be **mutually exclusive** if the occurrence of any one of them means the others will not occur.

A variable whose value is determined by the outcome of a random experiment is called a **random variable**.

When the set of assumed values of a random variable is countable, it is called a **discrete random variable**. If it is not countable, it is called a **continuous random variable**.

A table or formula that lists all possible values of a discrete variable along with associated probabilities is called a **discrete probability distribution**.

The function **f(x)** or **probability density function** for the **continuous random variable X** gives the total **area under the curve** bounded by the **x-axis** and is equal to 1. But the **AUC** between two ordinates **x = a** and **x = b** is the probability that **X** lies between **a** and **b**.

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \quad , \quad \int_{a}^{b} f(x)\,dx = P(a \le X \le b)$$

The mean **E(X)** or **μ** of the probability distribution is the expected value of **X** defined as:

$\Sigma\,(x_i \times P(x_i))$.

Variance of a discrete random variable **X** denoted by **V(X)** or **σ^2** is

$\Sigma[\{X - \mu\}^2 \times P(X)] = E(X^2) - [E(X)]^2$

Squareroot of **V(X)** is σ, the **standard deviation** of the probability distribution.

Mentioned below are different types of distributions; please look up formulae to calculate probabilities, mean and **σ** for each of them and read more about them.

Binomial distribution, Poisson distribution, normal (Gaussian) distribution

A **normal (Gaussian) curve** can be **standardized** so that its **μ** is 0 and **σ** is 1 unit. So all observations of any normal random variable **X** with mean **μ** and variance $\boldsymbol{\sigma^2}$ will be transformed to a new set of observations of another normal random variable **Z** with mean 0 and variance 1 by using the formula: $\mathbf{Z = (X - \mu)/\sigma}$

Both these graphs will have the same area. The new distribution of **Z** is called a **standard normal distribution**. Its probabilities are much easier to calculate. From this transformation comes the **z-Table**.

The **z-Table** gives the area to the right of the vertical center-line of the **z-curve** (standard normal curve) for different standard deviations. It can be used to find probabilities when the event in question follows a normal distribution.

Statistics seeks to make inferences from the data sample collected so that you can make inferences about the larger population from which you collected the sample. The samples need to be **independent**. **Confidence intervals (CI)** allow you to use the data from a sample to make inferences about the population. For instance, if you collected data for a small sample and you want to know how precisely you determined that value, a **95% CI** is a range of values that tells you that you can be 95% sure that it contains the true population value.

The Central Limit Theorem of statistics states that if your samples are large enough, the **distribution of means** will follow a **Gaussian (normal) distribution** even if the population is not Gaussian. So tests, such as the **t-test** and **ANOVA**, which deal with differences between means, will still work even when populations are not Gaussian. But samples have to be large (>10).

The ***P* value** is the probability (ranging from 0 to 1) of the difference between sample means arising because of chance. The smaller the ***P* value**, the less likely the difference between sample means is due to chance.

Hypothesis testing uses tests of significance to find out how likely a statement is likely to be true. So in any study, we formulate the hypothesis that we want to test, called the **alternative hypothesis ($\mathbf{H_a}$** or **$\mathbf{H_1}$)** that seeks to prove some underlying theory. We test this against the **null hypothesis ($\mathbf{H_0}$)** which states that there is no difference between two population or sample means and negates **$\mathbf{H_a}$**.

A probability level **α** is set, which is the **significance level,** and is the probability that we reject **$\mathbf{H_0}$** when it is in fact true. In other words, **α** is the probability that we would say that there is a difference between sample means when there is none. So tests of hypotheses will determine a critical region of size **α** using the sampling distribution of an appropriate test

statistic, determine the value of the test statistic from the sample data, and check whether the value of the test statistic falls within the critical region. If it does, we reject $\mathbf{H_0}$ and accept $\mathbf{H_a}$. If not, we fail to reject $\mathbf{H_0}$. $\boldsymbol{\alpha}$ is usually set at 0.05 which means that a result is statistically significant if it occurs < 5% of the time when populations are identical. Thus, finding a statistically significant result when the populations are identical gives you a **Type I error** while finding a "not significant" result when there is in fact a difference between your populations gives you a **Type II error**.

When you know the population variance $\boldsymbol{\sigma^2}$, the population is normally distributed and/or the sample size is large and want to test the mean of a distribution, we can use the **z-Test**.

But when the sample size is not large ($\mathbf{n}$<30) and you don't know the variance, you need to use the **t-Test** with **n-1 degrees of freedom (df)**. The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

You will need a **paired t-test** when your data are matched or paired, *e.g*., measurements before and after an intervention in the same subjects; treatment of one set of a matched pair with the other remaining untreated; and measurement of a variable in pairs of data points. The **Pearson correlation coefficient r** allows you to calculate a ***P*** value. If the pairing was effective, **r** will be positive and ***P*** will be small. This means that the two groups are significantly correlated and it makes sense to choose a paired test.

Nonparametric tests make fewer assumptions about the distribution of data but are less powerful than **parametric tests** that assume Gaussian distributions. So ***P*** values from nonparametric tests are higher, making it harder to detect real differences. Hence, it is better to have large samples when using nonparametric tests.

The **unpaired t-test** compares means of two groups, assuming that the data are sampled from Gaussian populations. The ***P*** value and the **confidence interval** are very important.

Analysis of Variance (ANOVA) assumes that your populations have a Gaussian distribution. **One-way ANOVA** compares ≥ 3 groups when the data are categorized in one way, *e.g*., a control group compared to two treated groups. If data are categorized in two ways (*e.g*., control versus two treated groups in males and females), use **two-way ANOVA**. **Post-tests** such as **Bonferroni, Tukey, Newman-Keuls** and **Dunnett's post-tests** are performed after **ANOVA**. These **post-tests** are performed when comparing three or more groups.

The (**one-way**) **ANOVA table** gives you the ***F* ratio** and ***P*** value. The ***F* ratio** is the ratio of two mean square values. If $\mathbf{H_0}$ is true, ***F*** should be close to 1 most of the time. A large ***F* ratio** means that variation among group means is more than what you'd expect to see by chance. When $\mathbf{H_0}$ is wrong and also when random sampling gives large values in some groups and small values in others, you will get a large ***F* ratio**.

One-way ANOVA gives you ***P*** value and the **post tests**. If the overall ***P*** value is large, you do not have convincing evidence to say that the means are different.

If the overall ***P*** value is small, it is not likely that the differences you see are due to random sampling. At least one mean differs from the other means. To find where the differences are, you need to look at the results of the **post tests**.

Look at which differences between column means are statistically significant (***P*** value). If the ***P*** value for a **post test** is small, you can reject the $\mathbf{H_0}$ that those two populations have identical means. The **95% CI** for the difference between all or selected pairs of means will tell you that this interval contains the true difference between the two means. Only scientific judgment can tell you whether the ends of the CI make the difference scientifically important or not.

Two-way ANOVA tells you how a response is affected by two factors: does either factor systematically affect the results, do the two factors interact? Each factor has to be **categorical**.

Interaction: $\mathbf{H_0}$ says that any systematic differences between columns are the same for each row and *vice versa*. The ***P*** value tells you the probability of randomly sampling subjects and ending up with as much (or more) interaction than what is seen if $\mathbf{H_0}$ is true.

Column factor: ***P*** value tells you the probability of randomly obtaining column means as different (or more) than what is seen if $\mathbf{H_0}$ is true.

Row factor: ***P*** value tells you the probability of randomly obtaining row means as different (or more) than what is seen if $\mathbf{H_0}$ is true.

The **95% CI** for the difference between two means tells you whether you can be 95% certain that all intervals contain the true difference between means. Two data sets (columns) will mean that **post tests** will compare two means from each row. For each row, you get a ***P*** value and **95% CI**.

If ***P*** for a **post test** is small, it is unlikely that the difference seen is because of random sampling. The two populations do not have identical means.

If ***P*** from a **post test** is large, then the data do not give you any reason to conclude that means of these two groups are different. The **95% CI** tells you that you can be 95% sure that it contains the true difference between the two means.

Finally, just remember that **statistical calculations lend strength to your conclusions but cannot and should not replace scientific judgement and common sense**.