ASSIGNMENT - 2

2451-18-733-001
M. Ramani Priya

1. List out the differences between Probability Mass functions and Probability Density Functions.

→ Probability Mass Functions (PMF) depend on the values of any real number. PMF plays an important role in defining a discrete probability distribution and produces distinct outcomes.

The formula for PMF is $p(x) = P(X = x)$ i.e., the probability of x = the probability (X = one specific x).

→ Probability Density Function (PDF) predicts probability functions in terms of continuous random variable values.

It is also known as probability distribution function or a probability function ($f(x)$).

Probability density function (PDF)	Probability mass function (PMF)
(i) PDF is used when there is a need to find a solution in a range of continuous random values.	(i) PMF is used when there is a need to find a solution in a range of discrete random values.
(ii) PDF uses continuous random variables.	(ii) PMF uses discrete random variables.
(iii) $f(x) = P(a < x)$	(iii) $p(x) = P(X = x)$
(iv) The solution of PDF falls in the radius range of continuous random variables.	(iv) The solutions of PMF falls in the radius of between the numbers of discrete random variables.

(IV) Probability of range of outcomes.

(V) Probability of a certain outcome.

2. What is a hypothesis & how is it tested?

Hypothesis: A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true.

→ Hypothesis testing refers to the formal procedures used to accept or reject a hypothesis.

→ There are 2 types of statistical hypotheses,

1. Null hypothesis (H_0) is usually the hypothesis that sample observations result purely by chance.

2. Alternative hypothesis (H_a) or (H_1) is the hypothesis that sample observations are influenced by some non random cause.

Hypothesis tests:

Statisticians follow a formal process to determine whether to reject a null hypothesis, based on sample data. This process is called hypothesis testing, consists of 4 steps:

1

2451-18-733-001

1. State the hypothesis: This involves stating the null and alternative hypothesis. Thes hypotheses are stated in such a way that they are mutually exclusive.

2. Formulate an analysis plan: The analysis plan describes how to use sample data to evaluate the null hypothesis.

3. Analyze sample data: Find the value of the test statistic (mean score, proportion, t statistic, z-score etc) described in the analysis plan.

4. Interpret results: Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

3. How random variables are different from traditional variables used in algebra?

1. A random variable is different from the variable in algebra as it has whole set of values & it can take any of those randomly.

2. A variable is an unknown quantity that has an undetermined magnitude and random variables are used to represent events in a sample space related values as a dataset.

3. A variable can be defined in the domain as a set of real numbers or complex numbers while random variables can be either real numbers or some discrete entities in a set

4. A random variable can be used to denote an event related to some object - its purpose of a random variable is to introduce a mathematically manipulative value to that event.

5. Random variables are associated with probability & probability density function.

→ Algebraic operations performed on algebraic variables may not be valid for random variables.

6. Random variables are often associated designated by letters & can be classified as discrete, which are variables that have specific values or continuous, which are variables that can have any values within a continuous range.

4 List out the properties of probability mass & density functions?

Probability Mass function (PMF).

→ Also called a probability function (or) frequency function which characterizes the distribution of a discrete random variable.

→ Let X be a discrete random variable of a function, then the probability mass function of a random variable X is given by:

$$p(x) = P(X = x), \quad \forall x \in \text{range of } X.$$

→ It is noted that the probability function should fall on the condition:

$$P_X(x) \geq 0 \quad \&$$

$$\Sigma_{x \in \text{Range}(x)} P_x(x) = 1$$

Range (X) – countable set & can be written as $\{x_1, x_2, x_3, \ldots\}$.

→ This means that the random variable X takes $x_1, x_2, \ldots$

→ The PMF satisfies the following properties:

- $P(X = x) = f(x) \geq 0$; if $x \in$ Range of x that supports,
- $\sum_{x \in \text{Range of } x} f(x) = 1$
- $P(X \in A) = \sum_{x \in A} f(x)$

Probability Density Function (PDF).

→ A probability density function (PDF) is a function that describes the relative likelihood for this random variable to take on a given value.

→ It is given by the integral of the variable's density over that range.

→ It can be represented by the area under the density function that above the horizontal axis & between the lowest & greatest values of the range.

Properties:

(i) $f_X(x) \geq 0$

(ii) $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

(iii) $P[a \leq X \leq b] = \int_a^b f_X(x)\,dx$

(iv) It is the derivative of CDF of a continuous random variable.

5 What is the purpose sample statistics? Explain the properties of sample statistics.

→ The proliferation of data of varying quantity & relevance reinforces the need for sampling as a tool to work efficiently with a variety of data & to minimize bias.

→ Even in big data project, predictive models are typically developed & piloted with samples. Samples are also used in tests of various sorts.

→ A sample is a subset of data from a larger data set

→ The larger data set is called population (large defined set of data)

→ Random sampling is a process in which each available number of population being sampled has an equal chance of being chosen for sample at each draw. The sample that results is called simple random sample.

→ Sampling can be done with & without replacement

→ Data quality in data science involves completeness, consistency of format, cleanliness & accuracy of individual data points

statistics adds notions of representativeness.

Properties: of

→ Sampling distribution:
The probability distribution of a given statistic based on a random sample.

→ Sampling distributions are important for inferential statistics.

→ Standard error:
The standard deviation of the sampling distribution of a statistic is referred to as standard error of the quantity.

For the case, where the statistic is sample mean, & samples are uncorrelated,

Standard error $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

s - sample standard deviation.

n - size of sample.

If all sample means are very close to population mean, then standard error of the mean would be small.

If the sample means varied considerably, then standard error of mean would be large.

→ The overall shape of the distribution is symmetric & approximately normal.

→ There are no outliers or other important deviations from the overall pattern.

→ The center of distribution is very close to true population mean.

6. What is statistical hypothesis? Briefly describe the various test statistics.

Statistical hypothesis:

A statistical hypothesis is an assumption about a population population. This assumption may/may not be true.

As it is impractical to examine whole population for to determine if a statistical hypothesis is true, typically a random sample is examined from the population.

If sample data is not consistent with the statistical hypothesis, the hypothesis is rejected.

There are 2 types of statistical hypotheses.

1. Null hypothesis: The null hypothesis, denoted by H_0, is usually the hypothesis that sample observations result purely from chance.

2. Alternative hypothesis: The alternative hypothesis, denoted by H_1 or H_a, is the hypothesis that sample observations are influenced by some non-random cause.

Hypothesis tests:

A formal process is followed to determine whether to reject a null hypothesis, based on sample data. This process is called hypothesis testing, it consists of 4 steps,

1. State the hypothesis

2. Formulate an analysis plan.
3. Analyze sample & data
4. Interpret results.

Decision errors.

There are 2 types of errors.,

1. Type I error: It occurs when the researcher rejects a null hypothesis when it is true.
(significance error, α)
2. Type 2 error: It occurs when the researcher fails to reject a null hypothesis that is false.
(β) The probability of non committing type2 error is called power of test.

One tail

Decision rules:

The analysis plan includes decision rules for rejecting the null hypothesis.

decision rules are described in 2 ways -

1. In reference to p-value.
2. In reference to region of acceptance.

1. p-value:

Suppose the test statistic is equal to S.

P value is the probability of observing a test statistic as extreme as S, assuming H_0 is true.

If p-value is less than significance value (level), we reject the hypothesis.

2. Region of acceptance (ROA).

The region of acceptance is a range of values.

If the test statistic falls within the region of acceptance, the null hypothesis is not rejected.

The ROA is defined so that the chance of making type I error is equal to significance level.

The set of values outside ROA are called region of rejection (ROR)

One tailed & Two tailed test.

→ A test of a statistical hypothesis, where the ROR is on only one side of sampling distribution is called one tailed test.

→ A test of a statistical hypothesis, where the ROR is on both sides of sampling distribution is called two tailed test.

→ Test method: Typically the test method involves a test statistic & a sampling distribution.

Computed from sample data, a test statistic might be a mean score, proportion, difference between means, difference between proportions, z-score, t statistic, chi-square etc.

→ Given a test statistic & its sampling distribution, a researcher can assess probabilities associated with the test statistic.

If test statistic probability is less than the significance value (level), the null hypothesis is rejected.

→ Using sample data, perform computations called for in the analysis plan.

→ Test statistic, when the null hypothesis involves a mean or proportion, use either of the following equations to compute the test statistic

Test statistic = -(Statistic − Parameter)/(Standard deviation of statistic)

Test Statistic = (statistic − Parameter)/(Standard error of the statistic)

where parameter is the value appearing in the null hypothesis, & statistic is the point of estimate of the parameter.

Common test statistics.

1. z-tests are appropriate for comparing means under stringent conditions regarding normality & a known standard deviation.

One sample z-test $z = \dfrac{\bar{x} - \mu_0}{(\sigma/\sqrt{n})}$

Two sample z-test $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

2. t-test is appropriate for comparing means under relaxed conditions

One sample t test $t = \dfrac{\bar{x} - \mu_0}{(s/\sqrt{n})}$

Paired t-test $t = \dfrac{\bar{d} - d_0}{(s_d/\sqrt{n})}$ $df = n-1$

Two sample pooled t-tests, equal variances $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$

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

Two sample unpooled t-test, unequal variances

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

3. Chi-squared tests.

use the same calculations & the same probability distribution for different applications.

→ Chi-squared tests for variance are used to determine whether a normal population has a specified variance.

Chi squared test for variance $\chi^2 = (n-1)\frac{s^2}{\sigma_0^2}$

Chi squared test for goodness of fit $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$.

4. F-tests (analysis of variance, ANOVA) are commonly used when deciding whether groupings of data by category are meaningful.

Two sample F test for equality of variances $F = \frac{s_1^2}{s_2^2}$.