Statistics Notes

Zac Garby

March 15, 2018

Contents

1 Probability

For any events A and B:

$$P(A') = 1 - P(A)$$

$$P(A) = P(A \cap B) + P(A \cap B')$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A' \cap B') = 1 - P(A \cup B)$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$P(A \cap B) + P(B|A)P(A)$$

When A and B are mutually exclusive events:

$$P(A \cap B) = 0$$

$$P(A \cup B) = P(A) + P(B)$$

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

When A and B are independent:

$$P(A \cap B) = P(A)P(B)$$

2 Statistical Distributions

A random variable can be distributed according to a distribution function. The two we need to know are the Binomial Distribution and the Normal Distribution.

2.1 Binomial Distribution

If a random variable X is distributed according to a Binomial Distribution, you can write:

$$X \sim B(n, p)$$

Where n is the number of trials and p is the probability of success for each one. A Binomial Distribution can only be used when all of the following conditions are true:

- A fixed number of trials
- Each trial ends in success or failure
- Trials are independent
- Probability of success is constant

$E(X)$ denotes the expected value of the random variable.

$$E(X) = np$$

$P(X = x)$ is the probability that the random variable X is equal to x.

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$P(X \leq x)$ is the probability that the random variable X is less than or equal to x. This can be calculated using an extension of the previous equation.

$$P(X \leq x) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

2.1.1 Normal Approximation

A Binomial Distribution can be approximated as a Normal one if either p is close to 0.5 or n is large. A general rule is if np and $n(1-p)$ are both greater than 5, a normal approximation can be used.

Due to the nature of the two distributions, a continuity correction must be used. Thus, for two random variables:

$$X \sim B(n, p)$$
$$Y \sim N(np, np(1-p))$$

A probability expression in terms of X can be transformed into one of Y:

$$P(X \geq 5) \approx P(Y > 4.5)$$
$$P(X \leq 10) \approx P(Y < 10.5)$$

2.2 Normal Distribution

If a random variable X is distributed according to a Normal Distribution, you can write:

$$X \sim N(\mu, \sigma^2)$$

Where μ is the mean, i.e. the value around which the distribution is symmetrical, and σ^2 is the variance. The square root of the variance, σ, known as the standard deviation, is used more often.

A Normal Distribution is used for continuous variables which are symmetrical and follow a bell-curve shape.

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

The probability of X being a particular value is so close to 0 that it actually is. This is because X is continuous, and therefore can take an infinite number of values. This also means:

$$P(X > x) = P(X \geq x) = 1 - P(X < x)$$

2.2.1 The Normal Normal Distribution

The random variable Z is defined such that:

$$Z \sim N(0, 1)$$

Other Normal Distributions can be converted to this distribution:

$$X \sim N(\mu, \sigma^2)$$
$$P(X < n) = P(Z < \frac{n - \mu}{\sigma})$$

2.2.2 The sample mean

The sample mean of a normal distribution $X \sim N(\mu, \sigma^2)$, denoted $\overline{X}$, is also distributed normally:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Where n is sample size.

2.3 Hypothesis Testing

A hypothesis test uses data from a sample to test whether or not a statement about a population is likely to be true.

The general idea is that you're given two statements. The first, called the null hypothesis, is always where p, or whatever variable you're testing, is equal to something. The other, the alternate hypothesis, is p being either greater than, less than, or not equal to the value which the null hypothesis claims.

If the alternative hypothesis says $p \neq v$, the test is two tailed, which means that at the end the significance level, α, is halved.