Summary and Mathematical Tables : Statistics 2

1 Poisson Distribution

Poisson distribution is a suitable model for events which

- occur randomly in space or time,
- occur singly, that is events cannot occur simultaneously,
- occur independently, and
- occur at a constant rate, that is the mean number of events in a given time interval is proportional to the size of the interval.

If $X \sim \text{Po}(\lambda)$ follows Poisson distribution with mean λ then

- probability mass function : $\mathbb{P}(X = x) = e^{-\lambda}\dfrac{\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$
- expectation : $\mathbb{E}(X) = \mu = \lambda$
- variance : $\text{Var}(X) = \sigma^2 = \lambda$

If $\lambda > 15$ then $X \sim \text{Po}(\lambda)$ may reasonably be approximated by the normal distribution $Y \sim \text{N}(\lambda, \lambda)$. A continuity correction of $\pm\frac{1}{2}$ must be applied.

2 Linear Combinations of Random Variables

For any random variables X and Y and any real numbers a, b and c, the following holds

- $\mathbb{E}(aX + bY + c) = a\mathbb{E}(X) + b\mathbb{E}(Y) + c$
- if X and Y are independent, then $\text{Var}(aX + bY + c) = a^2\text{Var}(X) + b^2\text{Var}(Y)$

If $X \sim \text{N}(\mu_X, \sigma_X^2)$ and $Y \sim \text{N}(\mu_Y, \sigma_Y^2)$ are independent, then for any real numbers a, b, and c

$$aX + bY + c \sim \text{N}(a\mu_X + b\mu_Y + c, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

If $X \sim \text{Po}(\lambda_X)$ and $Y \sim \text{Po}(\lambda_Y)$ are independent, then $X + Y \sim \text{Po}(\lambda_X + \lambda_Y)$.

3 Continuous Random Variables

The probability density function (PDF) $f(x)$, of a continuous random variable X satisfies

- $f(x) \geq 0$ for all real number x
- $\displaystyle\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$
- $\mathbb{P}(a \leq X \leq b) = \displaystyle\int_a^b f(x)\mathrm{d}x$

The cumulative distribution function (CDF) $F(x)$, of a continuous random variable X satisfies

- $F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)\mathrm{d}u$
- $f(x) = \dfrac{\mathrm{d}}{\mathrm{d}x}F(x)$
- $\lim\limits_{x\to-\infty} F(x) = 0$ and $\lim\limits_{x\to+\infty} F(x) = 1$
- $F(x)$ is increasing everywhere.

The median of X is the value of M for which $\mathbb{P}(X \leq M) = \int_{-\infty}^{M} f(x)\mathrm{d}x = \dfrac{1}{2}$

Expectation : $\mathbb{E}(X) = \mu = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x$

Variance : $\mathrm{Var}(X) = \sigma^2 = \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x - \mu^2$

4 Sampling

A random sample of size n is a sample chosen in such a way that each possible group of size n which could be taken from the population has the same chance of being picked.

If a random sample consists of n observations $X_1, X_2, \ldots, X_n$ of a random variable X with $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$ and $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$, then

- $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.)
- $\mathbb{E}(\bar{X}) = \mu$
- $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$

The Central Limit Theorem (CLT). For any sequence of independent identically distributed random variables $X_1, X_2, \ldots, X_n$ with finite mean μ and non-zero variance σ^2, then, provided n is sufficiently large, $\bar{X}$ is distributed approximately as $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, where $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$.

- A rule of thumb is to apply the theorem when n is greater than about 30.
- If X is a discrete random variable, then a continuity correction of $\pm\frac{1}{2n}$ must be applied when computing the distribution of $\bar{X}$.

5 Estimation

If a random sample consists of n random observations $x_1, x_2, \ldots, x_n$, of X, then

$$\text{sample mean : } \bar{x} = \frac{1}{n}\sum x \qquad \text{sample variance : } s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)$$

An estimate $\hat{x}$, of a value x is a good estimator if

- the estimate is unbiased, that is $\mathbb{E}(\hat{x}) = x$

- the variance $\text{Var}(\hat{x})$ is minimised.

If $x_1, x_2, \ldots, x_n$ are random samples of a random variable X with $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$, then

- unbiased estimate for μ : $\hat{\mu} = \bar{x}$
- unbiased estimate for σ^2 : $\hat{\sigma}^2 = s^2$

where $\bar{x}$ is the sample mean and s^2 is the sample variance.

The $(1-\alpha)$ symmetric confidence interval for the population mean μ, for a sample of size n taken from a normal population with known variance σ^2 is given by

$$\left[\bar{x} - z_{1-\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}}\right]$$

where $\bar{x}$ is the sample mean and $z_{1-\frac{1}{2}\alpha}$ is the value such that $\Phi\left(z_{1-\frac{1}{2}\alpha}\right) = 1 - \frac{1}{2}\alpha$.

If the sample size is large ($n \geq 30$), then σ^2 the may reasonably be approximated by the sample variance s^2. The confidence interval is given by

$$\left[\bar{x} - z_{1-\frac{1}{2}\alpha}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\frac{1}{2}\alpha}\frac{s}{\sqrt{n}}\right]$$

If $X \sim \text{B}(n, p)$ is a sample of size n from a population in which a proportion of members, p, has a particular attribute, then the sample proportion $p_s = \dfrac{X}{n}$, is distributed approximately as $p_s \sim \text{N}\left(p, \dfrac{pq}{n}\right)$, provided n is large enough such that $np > 5$ and $n(1-p) > 5$.

The $(1-\alpha)$ confidence interval for p is given by

$$\left[p_s - z_{1-\frac{1}{2}\alpha}\sqrt{\frac{p_s(1-p_s)}{n}}, p_s + z_{1-\frac{1}{2}\alpha}\sqrt{\frac{p_s(1-p_s)}{n}}\right]$$

where $z_{1-\frac{1}{2}\alpha}$ is the value such that $\Phi\left(z_{1-\frac{1}{2}\alpha}\right) = 1 - \frac{1}{2}\alpha$.

6 Hypothesis Testing for Continuous Random Variables

Testing a population mean for a Normal distribution. Test if the mean μ of a normal random variable X with known variance σ^2 is equal to μ_0 at α significance level.

Null hypothesis : $H_0 : \mu = \mu_0$

Alternative hypothesis :

- $H_1 : \mu \neq \mu_0$, for two-tail test
- $H_1 : \mu > \mu_0$, for one-tail increase test
- $H_1 : \mu < \mu_0$, for one-tail decrease test

Test statistics : $\hat{z} = \dfrac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$, where n is the number of observations and $\bar{x}$ is the sample mean.

H_0 is accepted if

- $\hat{z} \leq z_{1-\frac{1}{2}\alpha}$, for two-tail test,
- $\hat{z} \leq z_{1-\alpha}$, for one-tail increase test, or

- $|\hat{z}| \leq |z_\alpha|$, for one-tail decrease test

otherwise H_0 is rejected and H_1 is accepted.

If the sample size is large ($n \geq 30$), then σ^2 the may reasonably be approximated by the sample variance s^2.

The p-value is the probability that the mean of the population would be more extreme (greater for one-tail increase case, lower for one-tail decrease case) than or equal to the actual observed results, on the condition that H_0 is true. This is the probability that the test statistic deviates from the population mean by the observed amount in either direction.

If $X_1, X_2, \ldots, X_n$ are random observations of the distribution $X \sim \mathrm{N}(\mu, \sigma^2)$ then

- p-value $= \mathbb{P}(|X - \mu| \geq |\bar{X} - \mu| | \mu = \mu_0) = 2\Phi\left(-\dfrac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}}\right) = 2\left(1 - \Phi\left(\dfrac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}}\right)\right)$, for two-tail test,
- p-value $= \mathbb{P}(X \geq \bar{X} | \mu = \mu_0) = \Phi\left(-\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)$, for one-tail increase test, or
- p-value $= \mathbb{P}(X \leq \bar{X} | \mu = \mu_0) = \Phi\left(\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)$, for one-tail decrease test

H_0 is accepted if p-value $\geq \alpha$. Otherwise H_0 is rejected and H_1 is accepted.