Statistical Reasoning

1 Basic Definition

1.1 Random Variable

A random variable (RV) is a variable whose value is subject to variations due to chance (i.e. randomness, in a mathematical sense). This can be mathematically written as:

$$X \colon \Omega \to E$$

Where Ω is a probability space (set of possible outcomes) and E a measurable space which is often either $\mathbb{N}$ (discrete RV) or $\mathbb{R}$ (continuous RV).

An event $\omega \colon X(\omega)\ in I$ where I is an interval of E, is usually simplify by $X \in I$. Similarly, we write $\Pr(X \in I)$ instead of $\Pr(X^{-1}(I))$

1.2 Cumulative distribution function (CDF)

The cumulative distribution function (CDF) of a RV X (discrete or continious) is the probability that X will take a value less than or equal to x.

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R}$$

1.3 Probability density function (PDF)

A probability density function (PDF) of a **continuous** RV X is a function $f_X(x)$ that describes the relative likelihood for this random variable to take on a given value:

$$\Pr[a \leq X \leq b] = \int_a^b f_X(x)\,dx = F_X(b) - F_X(a)$$

Because F is derivable, we have $F'_X(x) = f(x)$

1.4 Characteristic function

The characteristic function of a RV is an alternative of CDF to describe the pdf, and define as the expected value of e^{itX}:

$$\varphi_X(t) = \mathrm{E}\left[e^{itX}\right] = \int_{\mathbb{R}} e^{itx}\,dF_X(x) = \int_{\mathbb{R}} e^{itx} f_X(x)\,dx$$

where $t \in \mathbb{R}$ is the argument of the characteristic function.

In order to better understand characteristic function we can write the equivalent for the cdf:

$$F_X(x) = \mathrm{E}\left[\mathbf{1}_{\{X \leq x\}(x)}\right]$$

where $\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$ is the indicator function.

The two function are equivalent and we can transform in the two direction easily.

If a random variable admits a pdf, then its characteristic function is the inverse Fourier transform of its pdf.

1.5 Gaussian Normal Distribution

A Random Variable (RV) following a Gaussian Normal Distribution is written $X \sim \mathcal{N}(\mu, \sigma^2)$ and has a corresponding probability distribution function $f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{\sigma^2}}$. Its expected value and variance are $\mathrm{E}[X] = \mu \quad \mathrm{Var}(X) = \sigma^2$

2 Theorem

2.1 Central Limit Theorem

Central Limit Theorem (CLT) states that the mean of an independent RV will be normally distributed, regardless of the underlying distribution

$$\lim_{n\to\infty} \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1) \quad \text{, with } \sigma^2 < \infty$$

Proof. Taylor expansion of e^{itx} in $t \to 0$:

$$e^{itx} = 1 + itx + (itx)^2/2 + O(t^2)$$

Using that in the definition of characteristic function:

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} f_X(x)\,dx = 1 + it\,\mathrm{E}[X] - \frac{t^2}{2}\,\mathrm{E}[X^2]$$

The characteristic function is basically a moment generating function, so that if we differentiate k time and evaluate it at zero we get the k-moment of X:

$$\mathrm{E}[X^k] = (-i)^k \varphi_X^{(k)}(0)$$

For any random variable, Y, with zero mean and a unit variance $Y \sim (0,1)$, using Tylor expansion metion above:

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + o(t^2)$$

to be continued...

3 Estimation

- The estimation of the mean of a RV $X \sim \mathcal{N}(\mu, \sigma^2)$ is written $\bar{X}$ and computed with the empirical mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The expected value and variance of $\bar{X}$ are found as followed:

$$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} n \mathrm{E}[X] = \mu$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n \mathrm{Var}(X) = \frac{\sigma^2}{n}$$

- Confidence interval of the mean (known variance) at the level $1-\alpha$ is:

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = \bar{X} \pm u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where the quantile u_p is defined by $P(\mathcal{N}(0,1) < u_p) = p$.

- Estimation of the variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\mathrm{E}[S^2] = \sigma^2$$

- **Chi-square distribution** is the distribution of a sum of the squares of k independent standard normal random variables:

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

We can show that:

$$(n-1)\frac{S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{(n-1)}$$

As $n \to \infty$, $\chi_n^2 \to \mathcal{N}(n, \propto n)$

- Confidence interval of the variance at the level $1-\alpha$

$$[S_{inf}, S_{sup}] = [S^2(n-1)/x^{n-1}_{1-\alpha/2}, S^2(n-1)/x^{n-1}_{\alpha/2}]$$

where the quantile x_p is defined by $P(\chi^{n-1} < x_p) = p$.

- **Student t distribution** For $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi_n^2$ idpdt:

$$\frac{X}{Y/\sqrt{n}} \sim t_n$$

- **Fisher F distribution** For $X \sim \chi^2_{n_X}$ and $Y \sim \chi^2_{n_Y}$ idpdt:

$$\frac{X/\sqrt{n_X}}{Y/\sqrt{n_Y}} \sim F^{n_Y}_{n_X}$$

- Confidence interval of the mean (unknown variance)

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = \bar{X} \pm t_{1-\alpha/2}\frac{S}{\sqrt{n}}$$

4 Testing

- **Trail example:** A statistical test procedure is comparable to a criminal trial; a defendant is considered not guilty as long as his or her guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough charging evidence the defendant is convicted.

  In the start of the procedure, there are two hypotheses H_0: "the defendant is not guilty", and H_1: "the defendant is guilty". The first one is called null hypothesis, and is for the time being accepted. The second one is called alternative (hypothesis). It is the hypothesis one hopes to support.

  The hypothesis of innocence is only rejected when an error is very unlikely, because one doesn't want to convict an innocent defendant. Such an error is called error of the first kind (i.e., the conviction of an innocent person), and the occurrence of this error is controlled to be rare. As a consequence of this asymmetric behaviour, the error of the second kind (acquitting a person who committed the crime), is often rather large.

	Do not reject H_0	Reject H_0
H_0 true	Correct	Type I error
H_1 true	Type II error	Correct

- **Level and Power** The **level** α is equal to the probability of type I error, that is $1 - P(\text{Not rejecting } H_0 \mid H_0 \text{ is true})$

  The **power** β is the complement to one of the probability of a non-error of type II, that is $1 - P(\text{Reject } H_0 \mid H_1 \text{ is true})$

- **P-value** The P-value is the probability of obtaining an equal or more extreme test statistics than what was actually observed, assuming H_0 is true. A large value of p indicates weak evidence against the null hypothesis (and small value, a strong evidence)

5 Simple regression

- **Linear Correlation** measure the linear relationship between two variables X and Y:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \in [-1,1]$$

where $\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$ Correlation does not mean causality because of spurious (outliers, compositional data) or dependence.

It can be shown that the score z tends to a normal distribution

$$z = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)$$

$$\lim_{n\to\infty} z = \sim \mathcal{N}\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$$

The Interval of confidence:

$$z \in [z_{inf} = z - u_{1-\alpha/2}/\sqrt{n-3}, z_{sup} = z + u_{1-\alpha/2}/\sqrt{n-3}]$$

$$\rho \in [(e^{2z_{inf}} - 1)/(e^{2z_{inf}} + 1), (e^{2z_{sup}} - 1)/(e^{2z_{sup}} + 1)]$$

- Probability setting of the linear regression

$$Y = a + bx + \epsilon$$

Because $\epsilon \sim \mathcal{N}(0, \sigma^2)$, Y is a random variable $Y \sim \mathcal{N}(a + bx, \sigma^2)$

- **Least-squares method**

$$\arg\min \sum (Y_i - \hat{a} - \hat{b}x_i)^2$$

- **Maximum likelihood**

6 Linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

7 ANOVA

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among group means and their associated variances.