Regression II

COMP9417, 22T2

Stats, stats, stats . . .

Probability Distribution

A probability distribution represents the probability we see a value x in a sample X. We denote this as $P(X = x)$.

Definition

- Probability *mass* function applies to discrete X
- Probability *density* function applies to continous X

Expected Values

An expected value (denoted $\mathbb{E}$) represents the weighted average of the probability distribution. This is typically seen as the value the distribution will converge to over time if sampled randomly.

For a discrete random variable, the form for an expected value is as follows:

$$\mathbb{E}(X) = \sum_{x \in X} xP(X = x)$$

In the continuous case, where $f(x)$ is the probability density function:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Example

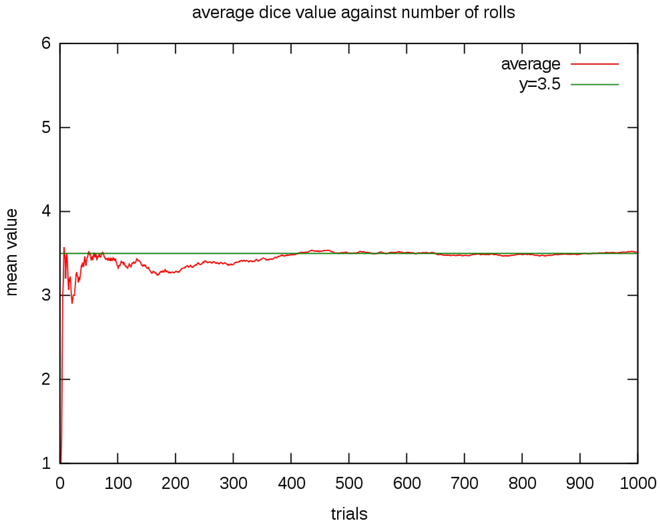
Problem: Model the probability mass function and find the expected value of the roll of a dice.

- $P(X = 1) = \frac{1}{6}$, $P(X = 2) = \frac{1}{6}$, ..., $P(X = 6) = \frac{1}{6}$

For the expected value:

$$\mathbb{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$$

This means that if we roll a dice, overtime the average dice value will converge to 3.5.


average dice value against number of rolls
average
y=3.5
mean value
trials
1
2
3
4
5
6
0
100
200
300
400
500
600
700
800
900
1000

We typically assume our samples are i.i.d (independent and identically distributed), helping us reduce the complexity of the problem and apply statistically supported conclusions.

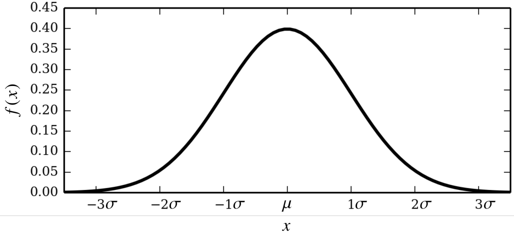
Gaussian Distribution

A standard probability distribution is the Gaussian, where:

$$\theta = (\mu, \sigma^2), \quad \mu \in \mathbb{R}, \sigma > 0$$

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

We typically write $X \sim \mathcal{N}(\mu, \sigma^2)$ to say X is normally distributed with mean μ and variance σ^2.



Maximum Likelihood Estimation

Maximum likelihood estimation is the process of estimating the parameters of a distribution of sample data by maximising the overall likelihood of the samples occuring in the distribution.

$$\begin{aligned}
\text{Prob of observing } X_1, \cdots, X_n &= \text{Prob of observing } X_1 \times \cdots \times \text{Prob of observing } X_n \\
&= p_\theta(X_1) \times \cdots \times p_\theta(X_n) \\
&= \prod_{i=1}^{n} p_\theta(X_i) \\
&=: L(\theta) \qquad \text{this is our likelihood function}
\end{aligned}$$

To make life easier, we typically work with the log of the likelihood function (log-likelihood). As log is a strictly increasing function, the maximisation of $L(\theta)$ and $\log(L(\theta))$ give us the same result.

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p_\theta(X_i) \\
\log(L(\theta)) &= \log \prod_{i=1}^{n} p_\theta(X_i) \\
&= \sum_{i=1}^{n} \log p_\theta(X_i)
\end{aligned}
$$

This makes differentiating, and therefore maximising much easier.

1a

1a

Problem: Given $X_1, \cdots, X_n \sim N(\mu, 1)$, find $\hat{\mu}_{\text{MLE}}$.

First, we define our likelihood function:

$$\begin{aligned}
\log L(\mu) &= \log\left(\prod_{i=1}^{n} p_\theta(X_i)\right) \\
&= \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i - \mu)^2\right)\right) \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2
\end{aligned}$$

Next, we differentiate with respect to our parameter μ,

$$\frac{\partial \log L(\mu)}{\partial \mu} = \sum_{i=1}^{n}(X_i - \mu)$$

$$= \sum_{i=1}^{n} X_i - n\mu$$

$$\frac{\partial \log L(\mu)}{\partial \hat{\mu}} = 0 \text{ at the minimum. So,}$$

$$\sum_{i=1}^{n} X_i - n\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\mu} = \bar{X}$$

1b

1b

Problem: Given $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$, find $\hat{p}_{\text{MLE}}$.

The Bernoulli distribution models processes with 2 outcomes (eg. a coin toss).

$$P(X = k) = p^k (1-p)^{1-k}, \qquad k = 0, 1 \quad p \in [0, 1]$$

First, we define our likelihood function:

$$\begin{aligned}
\log L(\mu) &= \log \left(\prod_{i=1}^{n} p^{X_i} (1-p)^{X_i} \right) \\
&= \sum_{i=1}^{n} \log p^{X_i} + \sum_{i=1}^{n} \log(1-p)^{X_i} \\
&= n\bar{X} \log p + n(1-\bar{X}) \log(1-p)
\end{aligned}$$

Next, we differentiate with respect to our parameter p,

$$
\begin{aligned}
\frac{\partial \log L(p)}{\partial p} &= \frac{n\bar{X}}{p} - \frac{n(1-\bar{X})}{1-p} \\
\frac{\partial \log L(p)}{\partial \hat{p}} &= 0 \text{ at the minimum. So,} \\
\frac{n\bar{X}}{\hat{p}} - \frac{n(1-\bar{X})}{1-\hat{p}} &= 0 \\
n\bar{X} - n\bar{X}\hat{p} &= n(1-\bar{X})\hat{p} \\
\hat{p}(n(1-\bar{X}) + n\bar{X}) &= n\bar{X} \\
\hat{p} &= \bar{X}
\end{aligned}
$$

Bias & Variance