Estimation

Clayton W. Schupp

Galvanize

Mathematical Expectation

- If X is a discrete random variable and $P(X = x)$ the value of its probability mass function at x, then for any function $g(x)$, the expected value is

$$E[g(X)] = \sum_{x \in S} g(x) \cdot P(X = x)$$

- If X is a continuous random variable and $f(x)$ the value its probability density function at x, then for any function $g(x)$, the expected value is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) dx$$

Useful Properties of $E(\cdot)$

- if a is a constant

$$E(a) = a$$

- If X is a random variable and a is a constant

$$E(aX) = aE(X)$$

- If X and Y are random variables and a and b are constants

$$E(aX + bY) = aE(X) + bE(Y)$$

1^{st} Moment: Expected Value of X $\longrightarrow$ Mean

- Discrete: Probability weighted average of all possible k values

$$E(X) = \mu = \sum_{i=1}^{k} x_i \cdot p_i$$

where $p_i = P(X = x_i)$

- Continuous: Same idea, except replace the summation with an integral, and replace probabilities with probability densities

$$E(X) = \mu = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Variance of a Random Variable

$$Var(X) = \sigma^2 = E[(X-\mu)^2] = \cdots = E[X^2] - \mu^2$$

- Discrete: Probability weighted average of all possible k squared deviations from mean

$$Var(X) = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot p_i$$

- Continuous: Same idea, except replace the summation with an integral, and replace probabilities with probability densities

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x) dx$$

Useful Properties of $Var(\cdot)$

- if a is a constant

$$V(a) = 0$$

- If X is a random variable and a is a constant

$$Var(aX) = a^2 Var(X)$$

- If X and Y are random variables and a and b are constants

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

Parametric vs. Nonparametric

Parametric and nonparametric procedures are two broad classifications of statistical methods

- Parametric
  - Based on assumptions about the distribution of the underlying population and its parameters from which the sample was taken
  - If the data deviates strongly from the assumptions, using a parametric procedure could lead to incorrect conclusions
- Nonparametric
  - Do not rely on assumptions about the shape or parameters of the underlying population distribution
  - Generally have less power than the corresponding parametric procedure
  - Interpretation can also be more difficult than parametric method

Method of Moments (MOM)

Derive equations related to the population moments:

$$E(X), E(X^2), E(X^3), \ldots$$

Method

1. Equate the first sample moment about the origin $M_1 = \frac{1}{n}\sum X_i$ to the first theoretical moment $E(X) = \mu$
2. Equate the second sample moment about the origin $M_2 = \frac{1}{n}\sum X_i^2$ to the second theoretical moment $E(X^2) = \sigma^2 + \mu^2$
3. Continue until you have as many equations as you have parameters
4. Solve for parameters

Method of Moments (MOM)

Example: Estimate probability of success in binomial distribution

$$X_i \overset{iid}{\sim} Bin(n, p) \qquad\qquad i = 1, 2, \ldots, n$$

$$E(X) = np \;\longrightarrow\; \bar{x} = np$$

$$\longrightarrow\; \hat{p}_{MOM} = \frac{\bar{x}}{n}$$

Method of Moments (MOM)

Example: Estimate lower and upper bound of a symmetric random uniform

$$X_i \overset{iid}{\sim} Unif(-\theta, \theta) \qquad i = 1, 2, \ldots, n$$

$$E(X) = \mu = 0 \longrightarrow \text{No help}$$

$$E(X^2) = \sigma^2 + \mu^2 \longrightarrow \frac{1}{n}\sum X_i^2 = \frac{1}{3}\theta^2$$

$$\longrightarrow \widehat{\theta}_{MOM} = \sqrt{\frac{3}{n}\sum X_i^2}$$

Maximum Likelihood Estimators

Set values of parameters to values that will maximize the likelihood function

- Assume $X_1, X_2, \ldots, X_n$ are *iid*, then the likelihood function is the joint density function

$$\mathcal{L}(\theta) = f(x_1, x_2, \ldots, x_3|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- Maximizing the likelihood function is the same as maximizing the log likelihood function which simplifies calculations

$$log\,\mathcal{L}(\theta) = \sum_{i=1}^{n} log[f(x_i|\theta)]$$

$$\widehat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, log\,\mathcal{L}(\theta)$$

MLE Example

Maximize the likelihood function by differentiating with respect to the parameter, setting equal to zero to solve, and setting that as the MLE estimate of the parameter

Example:

$$X_i \overset{iid}{\sim} Bin(n,p) \qquad i = 1, 2, \ldots, n \qquad f(x_i|p) = \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}$$

$$\begin{aligned}
log\mathcal{L}(p) &= \sum_{i=1}^{n}\left[log\binom{n}{x_i} + x_i logp + (n-x_i)log(1-p)\right] \\
\frac{\partial log\mathcal{L}(p)}{\partial p} &= \sum_{i=1}^{n}\left[\frac{x_i}{p} - \frac{n-x_i}{1-p}\right] = 0 \\
\hat{p}_{MLE} &= \frac{\bar{x}}{n}
\end{aligned}$$

Maximum a Posteriori (MAP)

- Mode of the posterior distribution
- We assume a prior distribution g over Θ and go one step further to calculate the posterior distribution

$$f(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_{\theta\in\Theta} f(x|\theta)g(\theta)d\theta} \propto f(x|\theta)g(\theta)$$

$$\widehat{\theta}_{MAP} = \operatorname*{argmax}_{\theta\in\Theta} f(x|\theta)g(\theta)$$

Kernel Density Estimation (KDE)

KDE is used to estimate the pdf of a random variable and is essentially a data smoothing problem

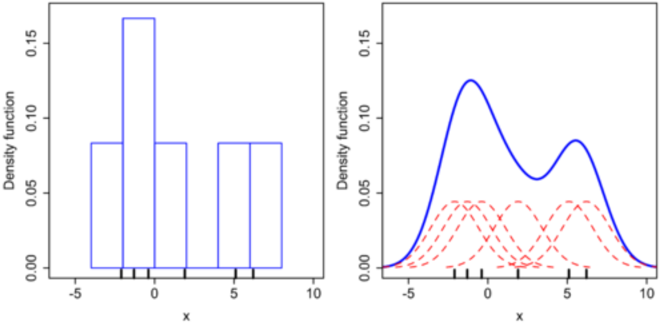
Let $(x_1, x_2, \ldots, x_n)$ be $i.i.d$ sample drawn from some distribution with unknown density f, we are interested in estimating the shape of the is function. Its kernel density estimator is

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where $K(\cdot)$ is the *kernel*: a non-negative function that integrates to one and has mean zero; and $h > 0$ is a smoothing parameter called the *bandwidth*

Kernel Density Estimation (KDE) Example

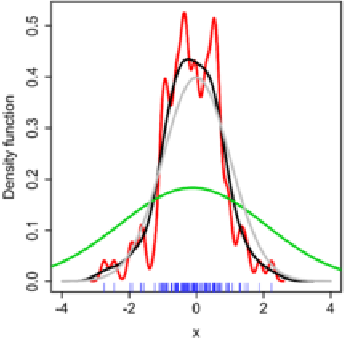
Closely related to histograms, but can be made smooth by using a suitable kernel.



Instead of binning boxes for a histogram, we are summing kernels

Bandwidth Selection

In figure below, the grey line is the standard normal distribution and the KDE are based on a random sample of 100 points



- A free parameter which exhibits a strong influence on the resulting estimate
- The most common optimality criterion used to select the parameter is the mean integrated squared error

$$MISE(h) = E\int(\hat{f}_h(x) - f(x))^2 dx$$