Point Estimators for Mean and Variance

- We know by now that the sample mean $(\hat{\mu}_n)$ is an unbiased estimator for the mean and its MSE is $\frac{\sigma^2}{n}$. It is also consistent.
- What about sample variance ? How can it be defined ?
- Since $\sigma^2 = E[(X - \mu)^2]$, we can define sample variance estimator as $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$.
- Problem with this estimator is that it needs the true mean which will not be available!
- What if we replace true mean by sample mean in the above formula?

Point Estimators for Mean and Variance

- Let $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$.
- HW Exercise: Is S^2 an unbiased estimator ? If no, find $B(\bar{S}^2)$.
- You will see that $E[S^2] = \frac{(n-1)\sigma^2}{n}$ and therefore $B(S^2) = \frac{-\sigma^2}{n}$.
- Can you think of an unbiased estimator of the variance ?
- How about $\bar{S}^2 = \frac{nS^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$?

> The sample variance defined by $\bar{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$ is an unbiased estimator of the variance.

- Is $\sqrt{\bar{S}^2}$ and unbiased estimator for the standard deviation σ.

Maximum likelihood estimation

- We have seen point estimators for mean and variance. What if we want to estimate other parameter in general like shape, scale, rate?
- Let $X_1, \ldots, X_n$ be i.i.d samples from a distribution with a parameter θ^*. Let $\mathcal{D} = \{X_1 = x_1, \ldots, X_n = x_n\}$.
- If X_i's are discrete, then the likelihood function is defined

$$L(x_1, x_2, \ldots, x_n; \theta) = p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta)$$

- $L(x_1, \ldots, x_n; \theta) = f_{X_1, \ldots, X_n}(x_1, \ldots, x_n; \theta)$ ($X_i's$ continuous)
- When samples are i.i.d, this is just the product of the densities/pmf's with parameter θ
- In such cases, it is easier to work with the log likelihood function given by $\ln L(x_1, x_2, \ldots, x_n; \theta)$
- Find the likelihood when $\mathcal{D}$ are samples from $exp(\theta)$, $\mathcal{N}(\theta, 1)$, $Binom(\theta, p)$, $Binom(n, \theta)$ etc.

Maximum likelihood estimation

- $L(x_1, \ldots, x_n; \theta) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta)$
- You want to find the best θ that represents the data!

Given $\mathcal{D} = \{x_1, \ldots, x_n\}$, the estimate $\hat{\Theta}_{ML}$ is given by

$$\begin{aligned} \hat{\Theta}_{ML} &= \arg\max_{\theta} L(x_1, \ldots, x_n; \theta) \\ &= \arg\max_{\theta} logL(x_1, \ldots, x_n; \theta) \end{aligned}$$

- We can generalize this to setting where more than one parameters say $(\theta_1^*, \ldots, \theta_k^*)$ are unknown.
- Note that differentiating w.r.t θ and equating to zero may not help if the parameter we are estimating is known to be an integer.

Properties of MLEs (without proof)

Let $X_1, \ldots, X_n$ be a i.i.d sample from a distribution with parameter θ^*. Then, under some mild regularity conditions,

1. $\hat{\Theta}_{ML}$ is asymptotically consistent, i.e., $\lim_{n\to\infty} P(|\hat{\Theta}_{ML} - \theta^*| > \epsilon) = 0$
2. $\hat{\Theta}_{ML}$ is asymptotically unbiased, i.e., $\lim_{n\to\infty} E[\hat{\Theta}_{ML}] = \theta^*$

Bayesian Inference with posterior distribution

- In Bayesian Inference we aim to extract information about unknown quantity θ^* based on observing a collection $X = (x_1, x_2, \ldots x_n)$ using Bayes rule.
- If Θ (model for θ^*) and X are discrete, Bayesian inference assumes a prior distribution $p_\Theta(\theta)$ on the unknown parameter Θ and uses the likelihood $p_{X|\Theta}(x|\theta)$ for observing data x to obtain the posterior $p_{\Theta|X}(\theta|x)$.
- If Θ and X are continuous, Bayesian inference assumes a prior distribution $f_\Theta(\theta)$ on the unknown parameter Θ and uses the likelihood $f_{X|\Theta}(x|\theta)$ for observing data x to obtain the posterior $f_{\Theta|X}(\theta|x)$.
- Θ being continuous and X discrete, and vice versa case are analogously obtained.

Bayes rule revisited .... revisited

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{f_X(x)} \quad (X, \Theta \text{ continuous})$$

$$p_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)} \quad (X, \Theta \text{ discrete})$$

$$p_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{f_X(x)} \quad (X \text{ cont}, \Theta \text{ discrete})$$

$$f_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{p_X(x)} \quad \Theta \text{ cont}, X \text{ discrete})$$

Example 1: Beta prior & Posterior, Binomial likelihood

- Suppose I toss a biased coin with θ^* as the true probability of head which you want to estimate based on data $\mathcal{D}_n$ from n tosses. Let X denote the number of heads in $\mathcal{D}_n$. Suppose we assume a $Beta(\alpha, \beta)$ prior on θ^*, i.e.,

$$f_\Theta(\theta) = \begin{cases} \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \text{ for } 0 < \theta < 1 \\ 0 \text{ otherwise.} \end{cases}$$

Then show that the posterior distribution $f_{\Theta|X}(\theta|k)$ has Beta distribution with parameters $\alpha' = \alpha + k$ and $\beta' = n - k + \beta$.

Example 1: Beta prior & Posterior, Binomial likelihood

- First note that the mean and variance for $Beta(\alpha, \beta)$ is given by $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- Also verify that when $\alpha = \beta = 1$, it corresponds top a uniform distribution.
- Now note that if we start with a uniform prior (or $Beta(1, 1)$), then the mean of the posterior distribution is given by $\frac{k+1}{n+2}$ and $\frac{(k+1)(n+1)}{(k+n+2)^2(k+n+2)}$.
- What happens as $n \to \infty$? The mean goes to θ^* almost surely using SLLN and the variance goes to zero.
- The posterior distribution therefore becomes a dirac-delta at θ^*.

Example 2: Gaussain Pior, Likelihood & Posterior

- Suppose we observe realisation $x = (x_1, \ldots, x_n)$ of $X = (X_1, \ldots, X_n)$ where X_i are i.i.d with true mean θ^* and true variance σ^2. Suppose we know σ^2 but not θ^* and also know that X_i is Gaussian. How do we infer θ^*?
- Lets model $\theta*$ by a Gaussian random variable $\Theta \sim \mathcal{N}(\mu_0, \sigma^2)$.
- Since X_i are i.i.d, the likelihood are given by

$$f_{X|\Theta}(x|\theta) = \prod_{i=1}^{n} f_{X_i|\Theta}(x_i|\theta)$$

- Now show that $f_{\Theta|X}(\theta|x)$ is Gaussian with mean $\frac{\sum_{i=1}^{n} x_i . + \mu_0}{n}$ and variance $\frac{\sigma^2}{n+1}$.
- What happens as $n \to \infty$?

Conjugate Priors

- Clearly, there are occasions where the prior and posterior are of the same family of distributions.
- The prior and posterior are called conjugate distributions and the prior is called conjugate prior.
- This makes it very convenient as now you only need to keep track of the parameters of the distribution than the distribution itself.
- `https://en.wikipedia.org/wiki/Conjugate_prior`

Point Estimation : Maximum aposteriori probability (MAP)

- Suppose θ^* is an unknown quantity. Let $f_\Theta(\theta)$ denote its prior.
- Let X denote random variable that is observable and is dependent on θ^*.
- Suppose we observe $X = x$. From this can we have a point estimate (a single value) for θ^*?

The MAP estimate $\hat{\theta}_{MAP}$ of θ^* given observation $X = x$ is the value of θ that maximizes $f_{\Theta|X}(\theta|x)$ (resp. $p_{\Theta|X}(\theta|x)$) when X is continuous (resp. discrete) random variable.

Maximum aposteriori probability (MAP)

The MAP estimate $\hat{\theta}_{MAP}$ of θ^* given observation $X = x$ is the value of θ that maximizes $f_{\Theta|X}(\theta|x)$ (resp. $p_{\Theta|X}(\theta|x)$) when X is continuous (resp. discrete) random variable.

- From Bayes rule this is same as maximizing $f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)$ (ignoring the dinominator since it is independent of θ).
- How do you optimize this to obtain $\hat{\theta}_{MAP}$?
- $\hat{\theta}_{MAP} \in \left\{\theta : \frac{d}{d\theta}\left(f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)\right) = 0\right\}$
- Compare this with MLE

$$\hat{\theta}_{ML} = argmax_{\theta} f_{X|\Theta}(x|\theta)$$

MAP for Example 2

- Recall Example 2 where we saw that given Gaussian samples $(x_1, \ldots, x_n)$ but with unknown mean μ, we model the unknwon mean as a random variable Θ with a Gaussian prior.
- We then get a Gaussian posterior $f_{\Theta|X}(\theta|x)$ with mean $\frac{\sum_{i=1}^{n} x_i . + \mu_0}{n}$ and variance $\frac{\sigma^2}{n+1}$.
- What is $\hat{\theta}_{MAP}$?
- Gaussian is a unimodal function and hence $\hat{\theta}_{MAP} = \frac{\sum_{i=1}^{n} x_i . + \mu_0}{n}$
- Is it same as MLE? HW!

Conditional Expectation Estimator

- Yet another estimator for the unknown θ^* is the conditional expectation estimator given by

$$\theta_{CE} = E[\Theta|X = x] = \int_{\theta} \theta f_{\Theta|X}(\theta|x) d\theta$$

.

- Find θ_{CE} for all the previous examples.