

Estimation & Sampling

Overview

- Review
 - Expected Value, Variance
- Statistics
 - Parametric vs. Non-Parametric
- Inference
 - MOM, MLE, MAP
 - KDE
- Sampling
 - CLT
 - Population Inference
 - Confidence Intervals
 - Bootstrapping

Review - Expectation

- Discrete: Probability weighted average of all possible values

$$E(X) = x_1 * p_1 + x_2 * p_2 + \dots + x_k * p_k$$

- Continuous: Same idea, except replace Σ with *integral*, and replace *probabilities* with *probability densities*

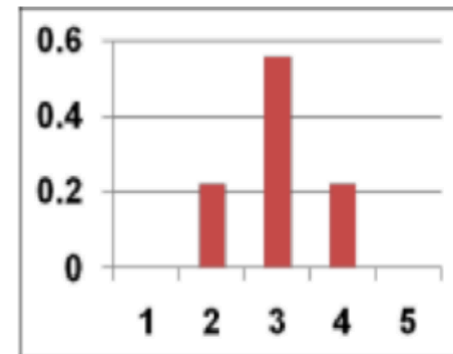
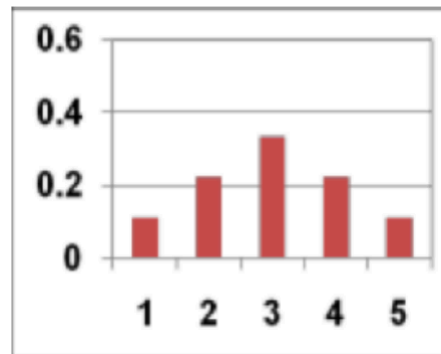
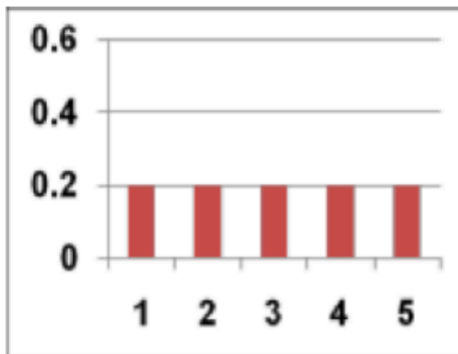
$$E(X) = \int_{-\infty}^{\infty} x * f(x) dx$$

Review - Variance

- Intuition

- Measures how much “spread” there is in a set of numbers
- The mean squared distance of random variable X from the mean, μ .

$$\begin{aligned}\text{Var}(X) &= \boxed{\text{E}[(X - \mu)^2]} & \text{Var}(X) &= \text{E}[(X - \text{E}[X])^2] \\ & & &= \text{E}[X^2 - 2X\text{E}[X] + (\text{E}[X])^2] \\ & & &= \text{E}[X^2] - 2\text{E}[X]\text{E}[X] + (\text{E}[X])^2 \\ & & &= \boxed{\text{E}[X^2] - (\text{E}[X])^2}\end{aligned}$$



Review - Variance

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

Suppose discrete r.v. X can take on k distinct values.

$$E(X) = \mu$$

$$Var(X) = E[(X - \mu)^2] = \sum_{i=1}^k p_i * (x_i - \mu)^2$$

- Continuous: Same idea, except replace Σ with *integral*, and replace *probabilities* p_i with *probability densities* $f(x)$

$$Var(X) = \sigma^2 = \int (x - \mu)^2 f(x) dx$$

Modeling – Parametric vs. Nonparametric

- Parametric

- Assumes data comes from a type of probability distribution and makes inferences about the parameters
- For example, $Normal(\mu, \sigma^2)$, $Poisson(\lambda)$
- May make use of some common sample statistics

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Non-parametric

- Unlike parametric case, non-parametric statistics make no assumptions about the probability distributions from which the variables arise

Inference – MOM and MLE

Method of Moments (MOM) and Method of Maximum Likelihood (MLE) are two **parameter estimation strategies**.

- $Normal(\underline{\mu}, \underline{\sigma}^2), Poisson(\underline{\lambda})$
- May or may not result in the same estimate

For fixed set of data and underlying statistical model...

- MOM – Derive equations related to population moments

What's a moment? $E(X), E(X^2), E(X^3)...$
first moment second moment third moment

- MLE – Set values of parameters to maximize the likelihood $f(n)$

Inference – MOM

MOM – Derive equations related to population moments

What's a moment? $E(X), E(X^2), E(X^3)\dots$

$$X_i \stackrel{\text{iid}}{\sim} \text{Binomial}(N, \underline{p}), \quad i = 1, 2, \dots, n$$
$$\Rightarrow E(X_i) = Np$$

Assume data comes from some distribution

$$\bar{x} = Np$$

Compute first moment from **sample data**.

$$\hat{p} = \frac{\bar{x}}{N}$$

Estimate parameter p based on first moment

Inference – MOM

$$X_i \sim \text{Uniform}(-\theta, \theta), \quad i = 1, 2, \dots, n$$
$$\Rightarrow E(X_i) = 0$$

Again, assume data comes from some distribution

$E(X)$ does not depend on parameters, so first moment doesn't help, but...

$$\text{Var}(X_i) = \frac{\theta^2}{3}$$

Compute first and second moments from **sample data**.

Recall $\text{Var}(X) = E(X^2) - [E(X)]^2$

$$s^2 = \frac{\hat{\theta}^2}{3}$$

Estimate parameter Θ based on first moment and second moments.

Inference – MLE

MLE – Set values of parameters to maximize the likelihood $f(n)$

First off...what's a likelihood function?

- Since we assume x_1, x_2, \dots, x_n are i.i.d., we have the joint density function

$$f(x_1, x_2, \dots, x_n | \theta) = f(x_1 | \theta) * f(x_2 | \theta) * f(x_3 | \theta) * \dots * f(x_n | \theta)$$

- Just call this joint density the “Likelihood”, and the log of that joint density the “Log Likelihood” just to make calculus easier)

$$\mathcal{L}(\theta | x_1, \dots, x_n) = f(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta) \quad \text{“Likelihood”}$$

$$\log \mathcal{L}(\theta | x_1, \dots, x_n) = \sum_{i=1}^n \log[f(x_i | \theta)] \quad \text{“Log Likelihood”}$$

- Parameter estimate is simply the one that maximizes the likelihood $f(n)$

$$\hat{\theta}_{mle} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log \mathcal{L}(\theta | x_1, \dots, x_n)$$

Inference – MLE

MLE – Set values of parameters to maximize the likelihood $f(n)$.

$$X_i \sim \text{Binomial}(N, p), \quad i = 1, 2, \dots, n$$

As with MOM, assume data comes from some distribution

$$\Rightarrow f(x_i) = \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}$$

$$\Rightarrow \mathcal{L}(p|x) = \prod_{i=1}^n \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}$$

Define Likelihood

$$\Rightarrow \log \mathcal{L}(p|x) = \sum_{i=1}^n \log \binom{N}{x_i} + x_i \log p + (N - x_i) \log (1 - p)$$

Log Likelihood

$$\Rightarrow \frac{\partial \log \mathcal{L}(p|x)}{\partial p} = \sum_{i=1}^n \left[\frac{x_i}{\hat{p}} - \frac{N - x_i}{1 - \hat{p}} \right] = 0$$

$$\Rightarrow \hat{p} = \frac{\bar{x}}{N}$$

Estimate parameter using some calculus!

Inference – MLE

Logistic Regression

- For each data point, have feature vector, x_i , and observed response y_i

$$\mathcal{L}(\beta_0, \beta | x_1, \dots, x_n) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

Can think of each observation as a Bernoulli trial where probabilities are estimated by your Logistic Model

Pick coefficients that maximize the joint likelihood!

Inference – MLE

Logistic Regression

- For each data point, have feature vector, x_i , and observed response y_i

$$\mathcal{L}(\beta_0, \beta | x_1, \dots, x_n) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

Can think of each observation as a Bernoulli trial where probabilities are estimated by your Logistic Model

Pick coefficients that maximize the joint likelihood!

Note: no closed form solution, but no matter!
We can use some numerical method, such as Gradient Descent.

MOM vs. MLE

- MOM introduced in 1894. Some say MLE has supplanted MOM.
- Still MOM has some good qualities
 - Fairly simple
 - Useful if MLE computationally intractable
 - Can be useful as stepping stone to solving MLE
 - First approximation to solutions of likelihood equations

Inference – MAP

- Maximum a posteriori (MAP) – mode of the posterior distribution

– For MLE, we have

$$\hat{\theta}_{mle} = \underset{\theta \in \Theta}{\operatorname{argmax}} f(x|\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \log \mathcal{L}(\theta|x_1, \dots, x_n)$$

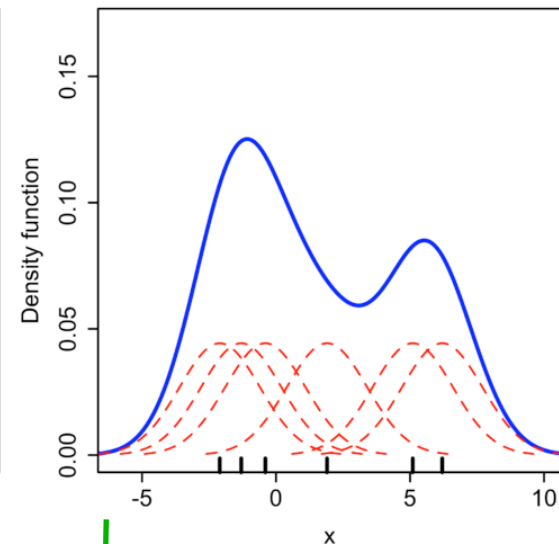
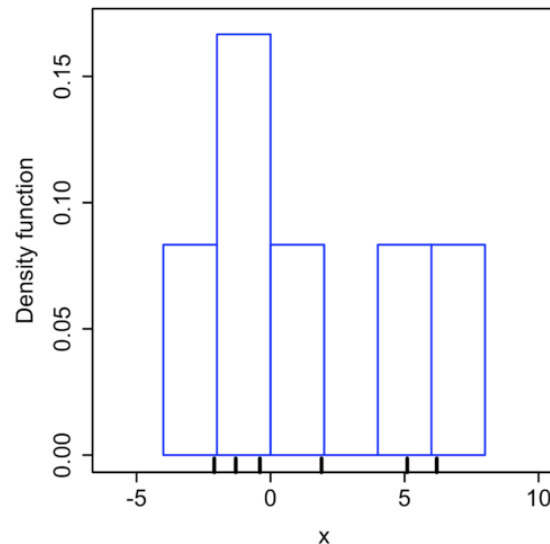
– For MAP, we assume a prior g over Θ , and go one step further to get the posterior.

$$\theta \mapsto f(\theta|x) = \frac{f(x|\theta) g(\theta)}{\int_{\vartheta \in \Theta} f(x|\vartheta) g(\vartheta) d\vartheta} \quad \leftarrow \text{Simply get Posterior using Bayes}$$

$$\hat{\theta}_{map} = \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{f(x|\theta) g(\theta)}{\int_{\vartheta} f(x|\vartheta) g(\vartheta) d\vartheta} = \underset{\theta \in \Theta}{\operatorname{argmax}} f(x|\theta) g(\theta).$$

Inference – KDE

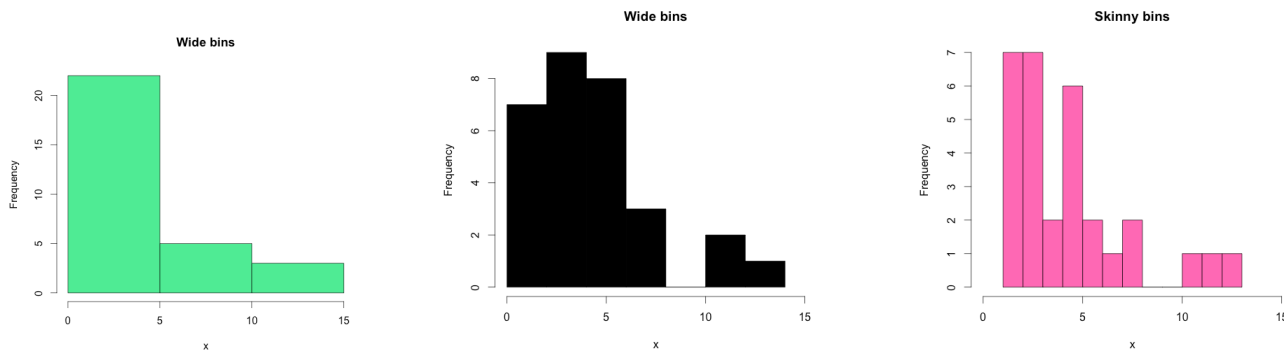
- Kernel Density Estimation (KDE)
 - Non-parametric way to estimate pdf of a random variable
 - Really, a data smoothing problem
 - Very similar to histograms
 - *Data:* $x_1 = -2.1, x_2 = -1.3, x_3 = -0.4, x_4 = 1.9, x_5 = 5.1, x_6 = 6.2$



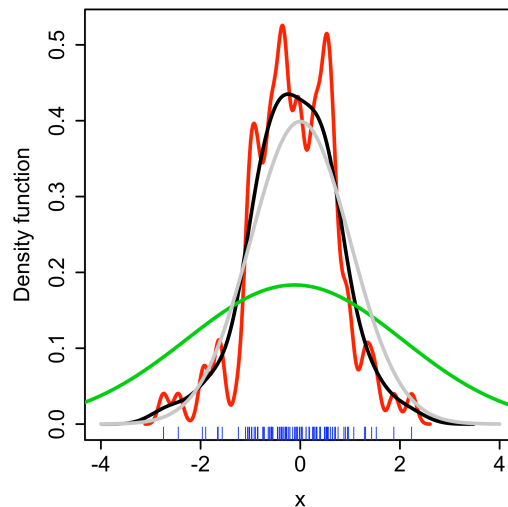
Instead of summing rectangles,
can sum gaussian curves

Inference – KDE

- Kernel Density Estimation (KDE)
 - Varying Bandwidth (for histograms)



- Instead, can use Gaussian kernels



Which bandwidth seems to be overfitting? Underfitting?

Questions

- MOM vs. MLE
 - What do they solve for?
 - How does each approach tackle the problem?
- How about MAP?
 - How does it relate to the MLE?
- What's bootstrapping?
 - When might I think of using it?
 - What are the steps to setting up a bootstrap estimate?

Questions

- MOM vs. MLE
 - What do they solve for? [Parameter Estimation](#)
 - How does each approach tackle the problem?
 - Both assume a specific distribution already.
 - MOM uses moment matching to get at parameters
 - MLE asks what parameter would maximize the the likelihood of the resulting data
- How about MAP?
 - How does it relate to the MLE? [Similar to MLE, but need to account for Prior](#)
- What's bootstrapping? [Random sampling w/ replacement technique](#)
 - When might I think of using it? [Want sense of accuracy of some sample estimate](#)
 - What are the steps to setting up a bootstrap estimate?
 - [Sample w/ replacement, B times → Compute B estimates from B samples → Get Standard Errors, Confidence Intervals, etc.](#)