AE 102: Data Analysis and Interpretation

January - April 2016
Department of Aerospace Engineering

Parameter Estimation

- Probability theory: you are given F
- Statistics: observed data $\rightarrow$ infer unknown parameters

Estimates

- Given that $X_1, \ldots, X_n$ from F_θ
- F_θ not fully specified, θ unknown
- Example:
  - Exponential distribution with unknown mean
  - Normal with unknown mean and variance.

Estimates/Estimators

- Point estimates
- Interval estimates
- Confidence
- Estimator: statistic to estimate unknown parameter θ

Maximum Likelyhood Estimators

- Assume unknown parameter θ
- Find joint PDF/PMF, $f(x_1, \ldots, x_n|\theta)$
- Maximize f w.r.t. $\theta \rightarrow \hat{\theta}$
- $f(x_1, \ldots, x_n|\theta)$, likelyhood function

- Provides a point estimate
- Note: f and $log(f)$ have same max location

MLE Example: Bernoulli Parameter

- n Bernoulli trials with p success probability
- What is the MLE of p?
- Data consist of valuex $X_1, \ldots, X_n$

Solution

$$P\{X_i = x\} = p^x(1-p)^{(1-x)}, \;\; x = 0, 1$$

$$f(x_1, \ldots, x_n | p) = p^{\sum_i x_i}(1-p)^{n - \sum_i x_i}$$

$$\text{maximize } \{\log f(x_1, \ldots, x_n | p)\}$$

Answer

$$\hat{p} = \frac{\sum_{i=1}^n x_i}{n}$$

MLE Example: Poisson Parameter

- n independent Poisson RVs with mean λ
- Find $\hat{\lambda}$

Solution

$$f(x_1, \ldots, x_n | \lambda) = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{x_1! \ldots x_n!}$$

$$\text{maximize } \{\log f(x_1, \ldots, x_n | \lambda)\}$$

Answer

$$\hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n}$$

MLE for Normal Population

- Self-study
- Same idea and approach
- Two parameters, so maximize w.r.t. each

MLE for a Uniform Distribution

- If $x \in (0, \theta)$
- θ should be small
- But large enough for largest X_i

Interval estimates

- Given that $X_1, ..., X_n$ from $\mathcal{N}(\mu, \sigma)$
- Unknown μ but **known** σ
- MLE $\hat{\mu} = \bar{X}$

- Is the MLE equal to actual μ??
- Can we provide an interval in which μ lies?

Interval estimates

- $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ is a standard normal
- So, for example:

$$P\left\{-1.96 < \sqrt{n}\frac{\bar{X}-\mu}{\sigma} < 1.96\right\} = 0.95$$

Interval estimates

- Can be modified to:

$$P\left\{\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

Example

- Given some $\bar{x}$, this means
  - With 95% confidence the mean lies within
  - $\pm 1.96 \frac{\sigma}{\sqrt{n}}$ of $\bar{x}$
- 95% percent confidence interval estimate of μ

Interpretation

- Whatever interval we obtain will contain the desired μ with 95% probability
- Once the interval is found, we only have a confidence of 95%

Example from textbook

Suppose that when a signal having value μ is transmitted from location A the value received at location B is normally distributed with mean μ and variance 4. That is, if μ is sent, then the value received is $\mu + N$ where N , representing noise, is normal with mean 0 and variance 4. To reduce error, suppose the same value is sent 9 times. If the successive values received are 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5, let us construct a 95 percent confidence interval for μ.

Two-sided vs one-sided

- With 95% confidence assert if μ is at least as large as the value

$$P\left\{\sqrt{n}\frac{\bar{X}-\mu}{\sigma} < 1.645\right\} = 0.95$$

$$P\left\{\bar{X} - 1.645\frac{\sigma}{\sqrt{n}} < \mu\right\} = 0.95$$

One-sided intervals

- One-sided upper CI for $\mu = \left(\bar{x} - 1.645\frac{\sigma}{\sqrt{n}}, \infty\right)$
- One-sided lower CI for $\mu = \left(-\infty, \bar{x} + 1.645\frac{\sigma}{\sqrt{n}}\right)$

Using the tables

- Recall $P\{Z > z_\alpha\} = \alpha$
- $P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - \alpha$

$$P\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

Finding suitable n

- Given desired interval size
- Find n to satisfy it

So what if variance is not known?

- Cannot assume $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ is Z
- We can find S^2

So what if variance is not known?

- $\sqrt{n}\frac{\bar{X}-\mu}{S}$ is t_{n-1}

$$P\left\{\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right\} = 1-\alpha$$

Non-normal populations

- Central limit theorem applies, so if n is "large enough" we should be good.

Confidence intervals for the variance

- Recall that $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$
- Homework.
- Note that χ^2 is not symmetric
- $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$

Example

The weights of 5 students was found to be 61, 65, 68, 58, and 70 Kgs. Determine a 95% confidence interval for their mean. Also determine a 95% lower confidence interval for this mean.

Difference in means

- $X_1, ..., X_n$ from $\mathcal{N}(\mu_\infty, \sigma_\infty)$
- $Y_1, ..., Y_m$ from $\mathcal{N}(\mu_\in, \sigma_\in)$
- CI for $\mu_1 - \mu_2$?
- Recall: distribution of two normally distributed RVs is normal

Difference in means

- MLE of $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y}$

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0,1)$$

When variances are not known?

- If $\sigma_1 \neq \sigma_2$ we have a problem
- If they are the same the same approach as before can be used

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathcal{N}(0,1)$$

Variances unknown

- $\bar{X}, S_1^2, \bar{Y}, S_2^2$ are independent
- If we consider

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

Variances unknown

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \sim t_{n+m-2}$$

Approximate CI for Bernoulli RV

- When n is large, $X \sim \mathcal{N}(np, np(1-p))$

Evaluating point estimators

- How good is an estimator, $d(X_1, \ldots, X_n)$?
- One measure is the mean-square error $E[(d(\mathbf{X}) - \theta)^2]$
- A desirable quality is unbiasedness

Unbiased estimators

- Bias is defined as: $b_\theta(d) = E[d(\mathbf{X})] - \theta$
- Unbiased if $b_\theta(d) = 0$
- If d is unbiased then $E[(d(\mathbf{X} - \theta)^2] = Var(d(\mathbf{X}))$

Bayes estimator

- **Prior** information on distribution of θ, i.e. $p(\theta)$
- Use data to find **posterior** density

$$\begin{aligned} f(\theta|x_1,\dots,x_n) &= \frac{f(\theta,x_1,\dots,x_n)}{f(x_1,\dots,x_n)} \\ &= \frac{p(\theta)f(x_1,\dots,x_n|\theta)}{\int f(x_1,\dots,x_n|\theta)p(\theta)d\theta} \end{aligned}$$

Bayes estimator

- Best estimate of θ is the mean of the posterior:

$$E[\theta|X_1 = x_1,\dots,X_n = x_n] = \int \theta f(\theta|x_1,\dots,x_n)d\theta$$

- See examples in the book

Contrived Example

- Lets say that the mean number of customers on a Sunday at Haiko between 2-3pm is either 20 or 40 (this is very questionable).
- Let us say that we feel that $P(\lambda = 20) = 0.7$ and $P(\lambda = 40) = 0.3$.
- Now let us say we observe that one day we find 40 people in this time.
- What should our new probability estimate be now? I.e. $P(\lambda = 20|X = 40)$?