APMA 1655 Notes: Chapter 8:

Data

- data: random samples from a given population distribution $f_\theta(x)$
- form of $f_\theta(x)$ is assumed to be known (parametric statistics)
- parametric statistics: branch of statistics that assumes that sample data from a population comes from a population that follows a probability distribution based on a fixed set of parameters
- θ is an unknown but fixed quantity
- Goal: want to use these random samples to form estimates and make inference of θ, where θ is called the target parameter or population parameter

Estimator

- Let $\{X_1, X_2, ..., X_n\}$ be iid samples from distribution $f_\theta(x)$
- iid means identically and independently distributed
- estimator $\hat{\theta}$ is a function of the iid samples: $\hat{\theta} = T(X_1, X_2, ..., X_n)$
- estimator is used as an estimate for θ
- n is the sample size
- using random variables to estimate fixed number
- for single target parameter there can be many different estimators
- there may not be a best estimator

Consistency

- a estimator $\hat{\theta}$ is said to be consistent if $\hat{\theta} \to \theta$ as samples size n goes to infinity

Bias

- estimator $\hat{\theta}$ is said to be unbiased if E[$\hat{\theta}$]; otherwise it is biased
- bias: difference B($\hat{\theta}$) = E[$\hat{\theta}$] - θ

Mean Square Error (MSE)

- MSE of an estimate $\hat{\theta}$ is: $MSE[\hat{\theta}] = E(\hat{\theta} - \theta)^2$
- decomposition of MSE: $MSE[\hat{\theta}] = B^2[\hat{\theta}] + Var[\hat{\theta}] =$ (Bias)^2 + Variance

Unbiased Estimators

- Let $\{X_1, ..., X_n\}$ are iid samples from the population
- Let μ and σ^2 be the population mean and variance
- Unbiased estimator for Population mean: sample mean

$$\hat{\mu} = \bar{X} = \frac{1}{n}\Sigma_{i=1}^n X_i$$

- Unbiased estimator for population variance: sample variance

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\Sigma_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1}(\Sigma_{i=1}^n X_i - n\bar{X}^2)$$

- Unbiased estimators for Bernoulli Distributions:
  - Let $\{X_1, ..., X_n\}$ are iid samples from a Bernoulli distribution such that

    $$P(X_i = 1) = p = 1 - P(X_i = 0)$$

  - Unbiased estimator for p:

    $$\hat{p} = \bar{X} = \frac{1}{n}\Sigma_{i=1}^n X_i$$

  - Var[$\hat{p}$]=$p(1-p)/n$

Confidence Intervals

- Suppose the population distribution is $N(\mu, \sigma^2)$, where σ is known
- then a (1-α) confidence interval for μ is

$$[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

- $z_{\alpha/2}$ is determined by equation

$$\phi(-z_{\alpha/2}) = P(N(0,1) \geq z_{\alpha/2}) = \frac{\alpha}{2}$$

- 95% confidence interval:
  - 95% is called the confidence level or confidence coefficient; in general it is $(1-\alpha)$ with $0 < \alpha < 1$
  - interval [L, R] such that L and R are both functions of samples $\{X_1, ..., X_n\}$ and $P(\mu \in [L, R]) = 95\%$
  - suppose $\{X_1, ..., X_n\}$ are iid samples from population distribution $N(\mu, \sigma^2)$; assume σ^2 is known, want to estimate μ
  - unbiased estimator: sample mean $\bar{X} = (X_1 + ... + X_n)/n$
  - $\bar{X}$ is $N(\mu, \sigma^2/n)$; then the deviation $\bar{X} - \mu$ is $N(0, \sigma^2/n)$
  - distribution of the error $\bar{X} - \mu$:

    $$P(|\bar{X} - \mu| \leq b) = 95\%$$

Large-Sample Confidence Intervals

- Setup: Let $\{X_1, ..., X_n\}$ be iid samples from a population with mean μ, want to estimate μ; population variance σ^2 is either known or unknown; sample size n is large
- Estimate: sample mean $\hat{\mu} = \bar{X}$; standard deviation $\sigma_{\hat{\mu}} = \sigma/\sqrt{n}$
- if σ is unknown, it can be approximated by sample standard deviation s:

$$s^2 = \frac{1}{n-1}\sigma_{i=1}^n (X_i - \bar{X})^2$$

  when n is large

- normal approximation: distribution of $\bar{X}$ is approximately $N(\mu, \sigma_{\hat{\mu}})$ by central limit theorem
- Theorem: Large Sample Confidence Interval
  - If population variance σ^2 is unknown, the $(1-\alpha)$ confidence interval for μ is approximately

$$[\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}]$$

  - if σ^2 is known, replace s by σ
  - $Z = (\bar{X} - \mu)/\sigma_{\hat{\mu}}$ is approximately N(0, 1)
  - $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$