What to do today (Oct 19, 2020)?

1. *Introduction*

2. *Probability and Distribution (Chp 1-3)*

3. Essential Topics in Mathematical Statistics

3.1 Elementary Statistical Inferences (Chp 4)

- **3.1.1 Sampling and Statistics**
- **3.1.2 Confidence Interval**
- *3.1.3 Order Statistics*
- *3.1.4 Hypothesis Testing*
- *3.1.5 Statistical Simulation and Bootstrap*

3.2 Consistency and Limiting Distributions (Chp 5)

3.3 Maximum Likelihood Methods (Chp 6)

3.1 Elementary Statistical Inferences

3.1.1 Sampling and Statistics

In the information age, statistics are everywhere, since

- data are everywhere, and
- always resources are limited and our observation abilities are limited.

Various statistical methods.

- *to efficiently collect meaningful and sufficient information*: **Survey Sampling and Experimental Design**
- *to process the available information by tabulating/plotting the data*: **Descriptive Analysis**
- *to make inference about the target population, beyond what the information is directly on*: **Inferential Analysis**

Plus **Probability and Distribution**: *inferential reasoning with probability theory*

3.1.1 Sampling and Statistics

Consider rv $X \sim F(\cdot)$, the population distn:

- A sample on X with size n: rvs $X_1, \ldots, X_n$

  its observations $x_1, \ldots, x_n$ from a study are **realizations** of the sample.
- If $X_1, \ldots, X_n$ are *independent and identically distributed (iid)* with the same distn $F(\cdot)$, the sample $\{X_1, \ldots, X_n\}$ is a **random sample** on X with size n.
- A function of $X_1, \ldots, X_n$, say, $T = T(X_1, \ldots, X_n)$, is called a **statistic**.
- A statistic that is used to estimate a population parameter θ is called a **point estimator** of θ.

Definition. Let $X_1, \ldots, X_n$ be a sample on rv $X \sim F(x; \theta)$. A statistic $T = T(X_1, \ldots, X_n)$ is an **unbiased** estimator of θ if $E(T) = \theta$.

Is $\bar{X}$ a “good” (point) estimator of μ? How to obtain a “good” estimator for θ in general?

3.1.1 Sampling and Statistics: Two Commonly Used Point Estimation Procedures

A. Method of Moments Estimation (MME)

Thinking ... Recall sample mean $\bar{X}$ to estimate population mean μ.

Extend the idea to estimating kth population moment, with $k = 1, 2, \ldots$?

Point estimation of population moments:

Suppose $X \sim F(\cdot; \theta_1, \ldots, \theta_m)$ and iid observations $X_1, \ldots, X_n$.

- kth population moment of X: $\mu_k = E(X^k)$
- kth sample moment with $X_1, \ldots, X_n$:

$$\hat{\mu}_k = \frac{1}{n}(X_1^k + \ldots + X_n^k)$$

- **Use $\hat{\mu}_k$ to estimate μ_k!** (*unbaised estimator*)

eg, $\mu_2 = E(X^2)$ is estimated by

$$\hat{\mu}_2 = \frac{1}{n}(X_1^2 + \ldots + X_n^2).$$

Further, what if $X \sim F(\cdot; \theta_1, \ldots, \theta_m)$ with $\theta_1, \ldots, \theta_m$ not all population moments? For example, $X \sim N(\mu, \sigma^2): \quad \theta_1 = \mu; \theta_2 = \sigma^2$. *How to estimate μ and σ^2?*

Recall that

$$\mu_2 = E(X^2) = \sigma^2 + \mu^2 = \theta_2 + \theta_1^2$$

How about use the following?

$$\begin{cases} \widehat{\mu}_1 = \bar{X} \text{ to estimate } \mu_1 = \mu; \\ \widehat{\mu}_2 \text{ to estimate } \sigma^2 + \mu^2 \end{cases}$$

If so, then

$$\begin{cases} \widehat{\mu}_1 = \bar{X} \text{ as } \widehat{\mu}, \\ \widehat{\sigma}^2 = \widehat{\mu}_2 - \bar{X}^2 \text{ to estimate } \sigma^2 \end{cases}$$

3.1.1 Sampling and Statistics: Method of Moments Estimation (MME)

MM Estimation Procedure:

- $X_1, \ldots, X_n$ are iid observations from the population $X \sim F(\cdot; \theta_1, \ldots, \theta_m)$.
- Denote the kth population mean μ_k by $\mu_k = \mu_k(\theta_1, \ldots, \theta_m)$ with $k = 1, 2, \ldots$.
- The **MM estimators** $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are the solution to the equations jointly:

$$\widehat{\mu}_1 = \mu_1(\theta_1, \ldots, \theta_m); \ldots; \widehat{\mu}_m = \mu_m(\theta_1, \ldots, \theta_m)$$

Revisit to the example of estimating μ and σ^2 with $X \sim N(\mu, \sigma^2)$:

Solve $\begin{cases} \bar{X} = \mu, \\ \widehat{\mu}_2 = \sigma^2 + \mu^2 \end{cases}$, and obtain $\widehat{\mu} = \bar{X}, \quad \widehat{\sigma}^2 = \widehat{\mu}_2 - \bar{X}^2$.

Are all MM estimators good? Is there any alternative estimation procedure?

3.1.1 Sampling and Statistics: B. Maximum Likelihood Estimation (MLE)

by R.A. Fisher (geneticist and statistician), 1920

Likelihood Function:

- Let the joint distribution (pmf, or pdf) of rvs $X_1, \ldots, X_n$ be $f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)$.
  When $x_1, \ldots, x_n$ are the observed values (realizations) of the rvs, the **likelihood function** of $\theta_1, \ldots, \theta_m$ given the data is

$$L(\theta_1, \ldots, \theta_m \mid \text{ data }) = f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)$$

- **interpretation**: a measure on how likely the observed sample is overall with the values of $\theta_1, \ldots, \theta_m$.
- Often $X_1, \ldots, X_n$ are iid observations (a random sample) from the population with distribution $f(x; \theta)$. If the observed values are $x_1, \ldots, x_n$, then the likelihood function is

$$L(\theta \mid \text{ data }) = \prod_{i=1}^{n} f(x_i; \theta) = f(x_1; \theta) \ldots f(x_n; \theta).$$

Maximum Likelihood Estimator (MLE):

- The **MLE** $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are the values of $\theta_1, \ldots, \theta_m$ that maximize the likelihood function:

$$L(\hat{\theta}_1, \ldots, \hat{\theta}_m \mid \text{ data }) = \max L(\theta_1, \ldots, \theta_m \mid \text{ data }).$$

- **interpretation**: The MLE $\hat{\theta}_1, \ldots, \hat{\theta}_m$ give the parameter values that agree most closely with the observed sample (the data).
- Often used **procedures**: (*Why?*)

(1) to maximize $\log L(\theta_1, \ldots, \theta_m)$
(2) to obtain the solution to

$$\begin{cases} \frac{\partial \ln L(\theta_1, \ldots, \theta_m)}{\partial \theta_1} = 0, \\ \ldots\ldots \\ \frac{\partial \ln L(\theta_1, \ldots, \theta_m)}{\partial \theta_m} = 0 \end{cases}$$

For example, iid $X_1, \ldots, X_{100} \sim N(\mu, \sigma^2)$ with observed values $x_1, \ldots, x_{100}$. The likelihood function of μ, σ^2 is

$$L(\mu, \sigma^2|\text{data}) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \Big\}$$

$$\log L(\mu, \sigma^2|\text{data}) = \sum_{i=1}^{100} \Big\{ \log\big(\frac{1}{\sqrt{2\pi}}\big) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \Big\}$$

$$\begin{cases} \frac{\partial \log L(\mu,\sigma^2)}{\partial \mu} = \sum_{i=1}^{100} \Big\{ \frac{2}{2\sigma^2}(x_i - \mu) \Big\} = 0 \\ \frac{\partial \log L(\mu,\sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{100} \Big\{ -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(x_i - \mu)^2 \Big\} = 0 \end{cases}$$

Thus the MLE of μ, σ^2 are $\widehat{\mu} = \bar{X}, \widehat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$.

Why MLE?

Large Sample Behavior of MLE $\hat{\theta}$:

With a random sample of size n, as $n \to \infty$

- $E(\hat{\theta}_n) \to \theta$: approximately unbiased
- $Var(\hat{\theta}_n) \to \sigma^{*2} = \min Var(\tilde{\theta})$ with unbiased $\tilde{\theta}$
- The distribution of $\hat{\theta}_n$ is approximately $N(\theta, \sigma^{*2})$

Remarks: MLE is widely used, because

- given the underlying population distribution, it is mechanically derived by calculus-based techniques
- is almost the best estimator that can be attained,
- is convenient to use to make statistical inferences.

3.1.2 Confidence Interval

Goal: Suppose $X \sim F(\cdot;\theta)$ and $X_1, \ldots, X_n$ iid observations from the population. To obtain a 'good' interval estimator of θ?

Definition. $\hat{\theta}_L$ and $\hat{\theta}_U$ are two statistics. The random interval $\big(\hat{\theta}_L, \hat{\theta}_U\big)$ is a $100(1-\alpha)\%$ **confidence interval (CI)** of θ is

$$P\big(\theta \in (\hat{\theta}_L, \hat{\theta}_U)\big) = 100(1-\alpha)\%.$$

Here, $(1-\alpha)$ is called the confidence level of the CI.

- eg, $\alpha = 0.05$, a $100(1-\alpha)\%$ CI of θ is a CI with confidence level of 95%.

3.1.2 Confidence Interval

- **Interpretation**. (frequentist)
  With 100 experiments' outcomes, there're at least $100(1-\alpha)$ out of the 100 CI realizations containing the true value of θ.

  Bayesian interpretation: different!

- **Confidence Level, Precision, and Sample Size**:
  - $100(1-\alpha)\%$ CI $(\hat{\theta}_L, \hat{\theta}_U)$: the confidence level is $1-\alpha$.

    $$P\big(\theta \in (\hat{\theta}_L, \hat{\theta}_U)\big) = 1-\alpha$$

  - Length (Width) of CI: $\hat{\theta}_U - \hat{\theta}_L$, about CI's **precision/accuracy**.
  - Often to determine the sample size n such that a $1-\alpha$ CI has a desired precision $\Rightarrow$ **Study Design**

Example 3.1

- *Study:* To determine the true average response time of a new operating system. What sample size is necessary to ensure the resulting 95% CI has a width of (at most) 10? Assume $\sigma = 25$.
- *Stats formulation:* Assuming a response time $X \sim N(\mu, \sigma^2)$ with $\sigma = 25$. To obtain a 95% CI of μ with length ≤ 10
- *Interval estimator:* $\left(\bar{X} - 1.96\frac{25}{\sqrt{n}}, \bar{X} + 1.96\frac{25}{\sqrt{n}}\right)$.
- *Sample size determination:* The length $2(1.96)(25/\sqrt{n})$ is to be at most 10:

  $$2(1.96)(25/\sqrt{n}) \leq 10.$$

  Thus $\sqrt{n} \geq 2(1.96)(25)/10 = 9.80$. So, n should be at least 97 ($9.80^2 = 96.04$).

Deriving a CI: a general procedure
to find $\hat{\theta}_L = l(X_1, \ldots, X_n)$ and $\hat{\theta}_U = u(X_1, \ldots, X_n)$ to satisfy

$$P\big(l(X_1, \ldots, X_n) < \theta < u(X_1, \ldots, X_n)\big) = 1 - \alpha$$

How? not easy! See a few examples. ...

3.1.2 Confidence Interval: to estimate μ

Consider rv X with $\mu = E(X)$, and a random sample $\{X_1, \ldots, X_n\}$ from the population.

Setting 1. $X \sim N(\mu, \sigma^2)$ with σ^2 *known*.

- **Point Estimator.** $\hat{\theta} = \bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$: with the following "good" properties
  - $E(\bar{X}) = \mu$, *unbiased*
  - $V(\bar{X}) = \frac{\sigma^2}{n}$, converging to zero as $n \to \infty$
  - $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

- **Confidence Interval.** $(\hat{\theta}_L, \hat{\theta}_U)$ with

$$\hat{\theta}_L = \bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \hat{\theta}_U = \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

  - $P\big((\hat{\theta}_L, \hat{\theta}_U) \ni \mu\big) = 95\%$,
    since $P\big(\hat{\theta}_L \geq \mu\big) = 2.5\%$ and $P\big(\hat{\theta}_U \geq \mu\big) = 97.5\%$.
  - for a general α?

3.1.2 Confidence Interval: to estimate μ

Consider rv X with $\mu = E(X)$, and a random sample $\{X_1, \ldots, X_n\}$ from the population.

Setting 2. $X \sim N(\mu, \sigma^2)$ with σ^2 *unknown*.

- **Point Estimator.** $\hat{\mu} = \bar{X}$ with "Good" properties:
  - $E(\bar{X}) = \mu$; $V(\bar{X}) = \frac{\sigma^2}{n}$; $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

How about the unknown σ^2?

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

- $E(S^2) = \sigma^2$; distn of S^2?

Proposition. (1) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. (2) S^2 and $\bar{X}$ are indpt. (3) $T = \frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} \sim t(n-1)$.

- **Confidence Interval.** $(\hat{\theta}_L, \hat{\theta}_U)$ with

$$\hat{\theta}_L = \bar{X} - \left(t_{1-\frac{\alpha}{2}}(n-1)\right) \frac{\hat{\sigma}}{\sqrt{n}}, \quad \hat{\theta}_U = \bar{X} + \left(t_{1-\frac{\alpha}{2}}(n-1)\right) \frac{\hat{\sigma}}{\sqrt{n}}$$

  - $P\big((\hat{\theta}_L, \hat{\theta}_U) \ni \mu\big) = 1 - \alpha$, since $P\big(\hat{\theta}_L \geq \mu\big) = \alpha/2$ and $P\big(\hat{\theta}_U \geq \mu\big) = 1 - \alpha/2$.

3.1.2 Confidence Interval: to estimate μ

Consider rv X with $\mu = E(X)$, and a random sample $\{X_1, \ldots, X_n\}$ from the population.

Setting 3. $X \sim F(x; \theta)$ with $\theta = \mu$, the population mean. To estimate $\theta = \mu$ when $n >> 1$.

- **Point Estimator.** $\hat{\mu} = \bar{X}$ with "good" properties:
  - $E(\bar{X}) = \mu$; $V(\bar{X}) = \frac{\sigma^2}{n}$;
  - By the CLT, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ approximately.
- $1 - \alpha$ **Confidence Interval.**
  - an approximate CI of $(1 - \alpha)$ level when σ^2 is known:

    $$\bar{X} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}}$$

    becasue $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$ approximately.
  - an approximate CI of $(1 - \alpha)$ level when σ^2 is unknown:

    $$\bar{X} \pm t_{1-\frac{\alpha}{2}} \sqrt{\frac{S^2}{n}} \approx \bar{X} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{S^2}{n}}$$

    becasue $\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t(n-1)$ approximately, close to $N(0, 1)$ if

Example. r.v. $X \sim Bernoulli(p)$: $X = \begin{cases} 1 \\ 0, \end{cases}$ with $P(X=1) = p$.
To estimate p with a random sample $\{X_1, \ldots, X_n\}$ when $n >> 1$.

- Firstly, $\mu = E(X) = p$ and $\sigma^2 = V(X) = p(1-p)$.
- Thus, a point estimator of p: $\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- Because $n >> 1$, an approximate $1-\alpha$ CI of p:
  $\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{S^2}{n}}$,
  similar to $\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$,

since $S^2 = \frac{1}{n-1}\big(\sum X_i^2 - n\bar{X}^2\big) = \frac{n}{n-1}\hat{p}(1-\hat{p}) \approx \hat{p}(1-\hat{p})$

Example 3.2 From a sample of 1250 BC voters, 420 of them indicate that they support the NDP. Obtain an approximate 95% CI for the proportion of BC voters who support the NDP. [(.310, .362)]

- Population: r.v. X=1 or 0 to indicate a vote for NDP in BC. $X \sim (1, p)$.
- Random sample: iid r.v.s $X_1, X_2, \ldots, X_{1250}$ (votes from BC). with $\bar{X}_{obs} = \bar{x} = 420/1250$ ($\hat{p}$).
- 95% Confidence Interval of $\theta = \mu = E(X) = p$:

$$\hat{\theta}_L = \bar{x} - 1.96\frac{s}{\sqrt{1250}}, \quad \hat{\theta}_U = \bar{x} + 1.96\frac{s}{\sqrt{1250}}$$

$s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}) \approx \hat{p}(1-\hat{p}) = 0.223 \Longrightarrow$ an approximate 95% CI: $(.310, .362)$.

Remarks:

- The result is usually reported in *news* as "33.5% $\pm$ 2.6% support for NDP"

- An alternative solution:
  - Y=# of NDP supporters out of 1250 BC voters $\sim Bin(1250, p)$, approximately $N(1250p, 1250p(1-p))$.
  - $Z = \frac{Y - 1250p}{\sqrt{1250p(1-p)}} = \frac{\frac{Y}{1250} - p}{\sqrt{p(1-p)/1250}} \sim N(0,1)$ approximately.

    So, an approximate 95% CI of p is
    $\frac{Y}{1250} \pm (1.96)\sqrt{\hat{p}(1-\hat{p})/1250} = .336 \pm .026.$
    $\hat{p} = \frac{Y}{1250} = \bar{X}$

3.1.2 Confidence Interval: to estimate other population parameter

Consider the following topics:

- How to estimate $\sigma^2 = Var(X)$?
- How to estimate $\mu_X - \mu_Y$ for the two populations, say, X, Y?
- How about to estimate θ when $X \sim F(x; \theta)$ in general?

What will we study next?

1. Introduction

2. Probability and Distribution (Chp 1-3)

3. Important Topics in Mathematical Statistics (Chp 4-6)

- **3.1 Elementary Statistical Inferences**
  - *3.1.1 Sampling and Statistics*
  - *3.1.2 Confidence Interval*
  - **3.1.3 Order Statistics**
  - **3.1.4 Hypothesis Testing**
  - *3.1.5 Statistical Simulation and Bootstrap*
- *3.2 Consistency and Limiting Distributions*
- *3.3 Maximum Likelihood Methods*

4. Further Topics, Selected from Chp 7-11