Mathematical Statistics Recap for Computing

7 Limit Theorems

Motivation

For some new statistics, we may want to derive features of the distribution of the statistic.

When we can't do this analytically, we need to use statistical computing methods to *approximate* them.

We will return to some basic theory to motivate and evaluate the computational methods to follow.

7.1 Laws of Large Numbers

Limit theorems describe the behavior of sequences of random variables as the sample size increases ($n \to \infty$).

If $X_1, \dots, X_n \overset{iid}{\sim} f$

① What is the distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$? as $n \to \infty$?

② How big does n need to be for $\bar{X} \mathrel{\dot\sim}$ Normal?

Often we describe these limits in terms of how close the sequence is to the truth.

How far is $\bar{X}$ from μ?

$\bar{X}$: statistic calculated from dat

μ: true/population mean of dsn data generated from?

How do we measure this distance? e.g. $|\bar{X} - \mu|$, or $(\bar{X} - \mu)^2$ maybe

We can evaluate this distance in several ways.

Some modes of convergence – e.g.

- almost surely ($P(\lim_{n\to\infty} X_n = X) = 1$)
- in probability ("for all" $\forall \varepsilon > 0$, $\lim_{n \to \infty} P(|X_n - X| > \varepsilon) = 0$)
- in distribution ($\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$)

What happens to sequences of r.v.'s as n gets large

(gives us useful approximations!)

e.g. Laws of large numbers –

Weak LLN: sample mean $\bar{X}_n$ converges in probability to pop. mean μ.

Strong LLN: Sample mean $\bar{X}_n$ converges a.s. to pop mean μ.

7.2 Central Limit Theorem

Theorem 7.1 (Central Limit Theorem (CLT)) Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and finite variance $\sigma^2 > 0$, then the limiting distribution of $Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ is $N(0, 1)$. i.e. $\bar{X}_n \xrightarrow{d} X$, $X \sim N(\mu, \sigma^2/n)$ (converging in distribution).

Interpretation:

The sampling distribution of the sample mean approaches a Normal distribution as sample size increases.

Remember

Note that the CLT doesn't require the population distribution to be Normal.

8 Estimates and Estimators

Let $X_1, \ldots, X_n$ be a random sample from a population.

Let $T_n = T(X_1, \ldots, X_n)$ be a function of the sample.

Then T_n is a "statistic"

and pdf of T_n is called the "sampling distribution of T_n"

or pmf

functions of sample

Statistics estimate *parameters*.

from population.

Example 8.1

$\bar{X}_n$ estimates μ.

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ estimates σ^2 pop variance.

$S = \sqrt{S^2}$ estimates σ

Definition 8.1 An *estimator* is a rule for calculating an estimate of a given quantity.
Definition 8.2 An *estimate* is the result of applying an estimator to observed data samples in order to estimate a given quantity.

A statistic like $\bar{X}_n$ is a point estimator

A CI is an interval estimator

(If base on observations, they are estimates)

We need to be careful not to confuse the above ideas:

$\overline{X}_n$ function of r.v.'s $\longrightarrow$ estimator (statistic)

$\bar{x}_n$ function of observed data (an actual #) $\longrightarrow$ estimate (sample statistic).

μ fixed, but unknown quantity $\longrightarrow$ parameter.

We can make any number of estimators to estimate a given quantity. How do we know the "best" one?

What are some properties we can use to say an estimator is "better" than another one?

9 Evaluating Estimators

There are many ways we can describe how good or bad (evaluate) an estimator is.

9.1 Bias

Definition 9.1 Let $X_1, \dots, X_n$ be a random sample from a population, θ a parameter of interest, and $\hat{\theta}_n = T(X_1, \dots, X_n)$ an estimator. Then the *bias* of $\hat{\theta}_n$ is defined as

iid

parameter we want to estimate

joint density of $X_1, \dots, X_n$

$E[T(X_1,\dots,X_n)] = \int_{\mathcal{X}} T(X_1,\dots,X_n) f_{\underline{X}}(\underline{x})\, d\underline{x}$

$$bias(\hat{\theta}_n) = E[\hat{\theta}_n] - \theta.$$

Definition 9.2 An *unbiased estimator* is defined to be an estimator $\hat{\theta}_n = T(X_1, \dots, X_n)$ where

$$bias(\hat{\theta}_n) = 0, \text{ i.e. } E(\hat{\theta}_n) = \theta$$

Example 9.1

If you used Unif(0,1) as your envelope for Rayleigh dsn, your histogram of values would be biased.

support $(0, \infty)$

(no large values)

Example 9.2

Let $X_1, \dots, X_n$ random sample from population w/ mean μ, variance $\sigma^2 < \infty$.

$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\cdot n \cdot \mu = \mu$

$\Rightarrow bias(\bar{X}_n) = E[\bar{X}_n] - \mu = 0 \Rightarrow \bar{X}_n$ is an <u>unbiased</u> estimator for μ.

Example 9.3 Compare 2 estimators of σ^2 for Ex. 9.2.

Sample variance

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

MLE of variance

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

Can show $E S^2 = \sigma^2$ but $\hat{\sigma}^2 = \frac{n-1}{n} S^2$, so

for large n $S^2 \approx \hat{\sigma}^2$

$E(\hat{\sigma}^2) = \frac{n-1}{n} E S^2 = \frac{n-1}{n}\sigma^2$

$\Rightarrow \hat{\sigma}^2$ is a biased estimator.

9.2 Mean Squared Error (MSE)

Definition 9.3 The *mean squared error (MSE)* of an estimator $\hat{\theta}_n$ for parameter θ is defined as

$$\begin{aligned} MSE(\hat{\theta}_n) &= E\left[(\theta - \hat{\theta}_n)^2\right] \\ &= Var(\hat{\theta}_n) + \left(bias(\hat{\theta}_n)\right)^2. \end{aligned}$$

can show

Generally, we want estimators with

① small bias

② small variance

often there is the bias - variance trade-off (can't have both)

Sometimes an unbiased estimator $\hat{\theta}_n$ can have a larger variance than a biased estimator $\tilde{\theta}_n$.

Example 9.4 Let's compare two estimators of σ^2.

sample variance $s^2 = \frac{1}{n-1}\sum(X_i - \overline{X}_n)^2$ MLE $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \overline{X}_n)^2$

$E(s^2) = \sigma^2$ $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$

but $Var(s^2) > Var(\hat{\sigma}^2)$!

Can show:

$MSE(s^2) = E((s^2 - \sigma^2)^2) = \frac{2}{n-1}\sigma^4$

$MSE(\hat{\sigma}^2) = E((\hat{\sigma}^2 - \sigma^2)^2) = \frac{2n-1}{n^2}\sigma^4$

$\Rightarrow MSE(s^2) > MSE(\hat{\sigma}^2)$.

See pg. 331 Casella & Berger

9.3 Standard Error

Definition 9.4 The *standard error* of an estimator $\hat{\theta}_n$ of θ is defined as

$$se(\hat{\theta}_n) = \sqrt{Var(\hat{\theta}_n)}.$$

← standard error = st. dev of sampling dsn of $\hat{\theta}_n$.

We seek estimators with small $se(\hat{\theta}_n)$.

Example 9.5

$$se(\bar{X}_n) = \sqrt{Var(\bar{X}_n)} = \sqrt{\frac{Var(X_1)}{n}} = \frac{\sigma_X}{\sqrt{n}}$$

10 Comparing Estimators

We typically compare statistical estimators based on the following basic properties:

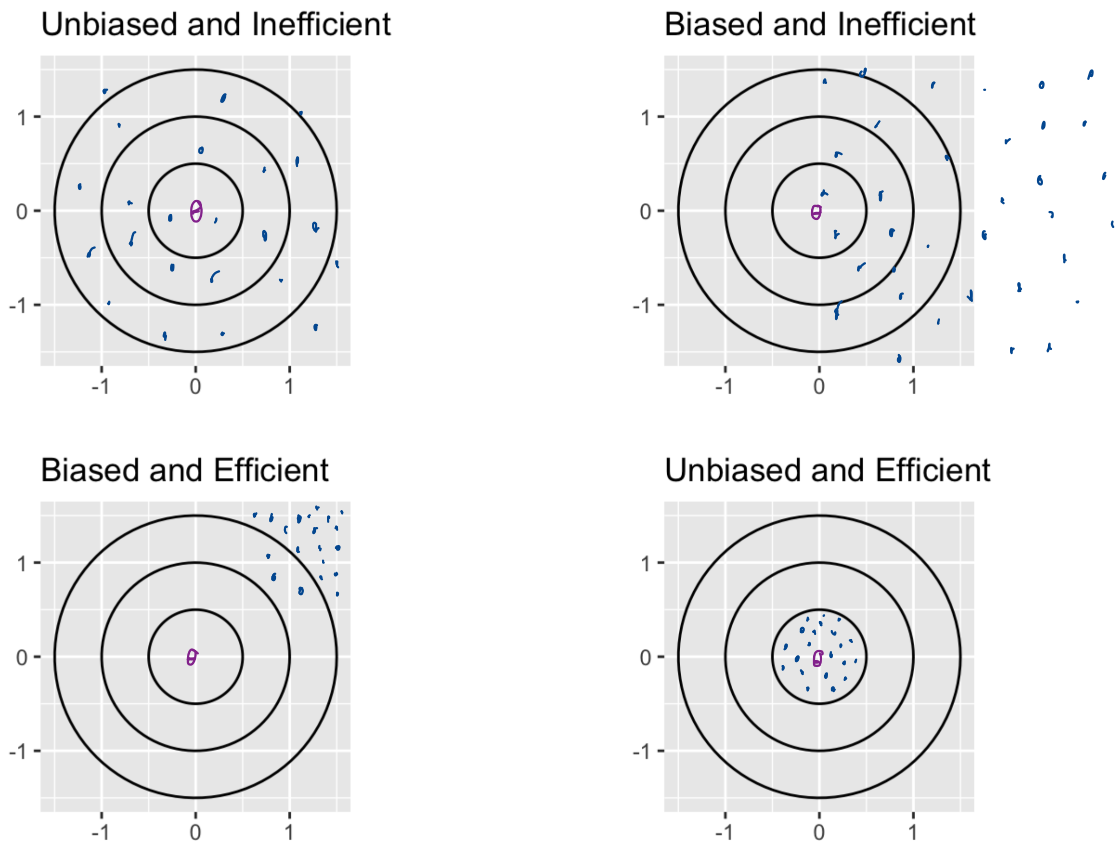
1. Consistency: as $n\uparrow$ does estimator converge to parameter its estimating? (convergence in probability).

2. Bias: Is the estimator unbiased? $E(\hat{\theta}_n) = \theta$

3. Efficiency: $\hat{\theta}_n$ is <u>more efficient</u> than $\tilde{\theta}_n$ if $Var(\hat{\theta}_n) < Var(\tilde{\theta}_n)$.

4. MSE: Compare $MSE(\hat{\theta}_n)$ to $MSE(\tilde{\theta}_n)$

   rember bias-variance tradeoff $MSE(\hat{\theta}_n) = Var(\hat{\theta}_n) + [Bias(\hat{\theta}_n)]^2$



Example 10.1 Let us consider the efficiency of estimates of the center of a distribution. A **measure of central tendency** estimates the central or typical value for a probability distribution.

Mean and median are two measures of central tendency. They are both **unbiased**, which is more efficient?

$E(\bar{X}) = E[\tilde{X}] = \mu$

i.e. which has smaller variance?

```
set.seed(400)

times <- 10000 # number of times to make a sample
n <- 100 # size of the sample
uniform_results <- data.frame(mean = numeric(times), median =
  numeric(times))
normal_results <- data.frame(mean = numeric(times), median =
  numeric(times))
```

# of draws from sampling dsn of statistic

store results.

```
for(i in 1:times) {
  x <- runif(n)
  y <- rnorm(n)
  uniform_results[i, "mean"] <- mean(x)
  uniform_results[i, "median"] <- median(x)
  normal_results[i, "mean"] <- mean(y)
  normal_results[i, "median"] <- median(y)
}
```

unif[0,1] sample of size n=100

Normal(0,1) sample of size n=100

store mean of unif samples

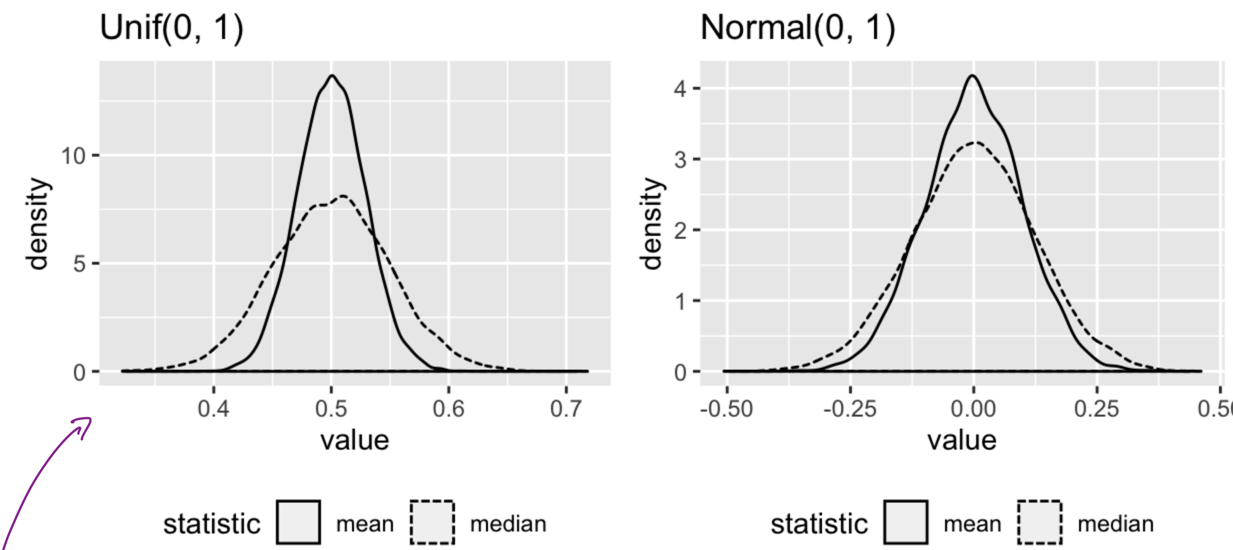
store median of uniform samples

```
uniform_results %>%
  gather(statistic, value, everything()) %>%
  ggplot() +
  geom_density(aes(value, lty = statistic)) +
  ggtitle("Unif(0, 1)") +
  theme(legend.position = "bottom")
```

estimate and plot density of statistic from samples

plot results.

```
normal_results %>%
  gather(statistic, value, everything()) %>%
  ggplot() +
  geom_density(aes(value, lty = statistic)) +
  ggtitle("Normal(0, 1)") +
  theme(legend.position = "bottom")
```



sampling dsns

sampling dsn of	$\bar{X}$	S
mean	.4999	.029
median	.4999	.0494

true mean = true median = 0.5

sampling dsn of	$\bar{X}$	S
mean	.0001	0.10
median	-.0009	0.12

true mean = true median = 0.

For both data scenarios, Unif(0,1) and N(0,1),

BIAS: both mean and median are unbiased.

EFFICIENCY: mean more efficient than median

$\hat{Var}(\text{mean}(X_1, \dots, X_n)) < \hat{Var}(\text{median}(X_1, \dots, X_n))$

This is not the case for all dsns. Ex: When a dsn is heavy tailed, median can be more efficient than the mean.

Next Up In Ch. 5, we'll look at a method that produces *unbiased* estimators of $E(g(X))$!