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 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$? $N(\mu, \frac{\sigma^2}{n})$. EX: VarX:

2. How big does n have to be for $\bar{X} \dot\sim$ Normal? 30

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

How close is $\bar{X}_n$ to μ? (how far away is it?).
statistic — true value we are estimating w/ that statistic

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

We can evaluate this distance in several ways. $\bar{X}_n$ is a random variable

Some modes of convergence – e.g.

- almost surely $(P(\lim_{n\to\infty} X_n = X) = 1)$.
- in probability $(\forall \varepsilon > 0, \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0)$
- 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

$$\forall \varepsilon > 0, \lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$$

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

$$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$$

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 $N(0, 1)$.

iid

(convergence in distribution), i.e. $\overline{X}_n \xrightarrow{d} X$, $X \sim N(\mu, \frac{\sigma^2}{n})$.

Interpretation:

The sampling distribution of the sample mean approaches a normal distribution as the 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. iid

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

Then T_n is a "statistic"

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

based on sample (data)

Statistics estimate parameters. → characterize the population.

Example 8.1

$\min_n X_n$ is a statistic that would estimate the min of the pop. values.

$\bar{X}_n$ sample mean estimates μ population mean.

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

$s = \sqrt{s^2}$ estimates σ pop. st. dev.

Definition 8.1 An *estimator* is a rule for calculating an estimate of a given quantity. function

Definition 8.2 An *estimate* is the result of applying an estimator to observed data samples in order to estimate a given quantity. an actual number based on data.

A statistic is a point estimator

A CI is an interval estimator

(if based on actual observed data they are estimates)

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

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

$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ function of observed data (an actual #) → estimate

μ – fixed but unknown quantity → 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

parameter (fixed but unknown) we want to estimate.

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

$$E[T(X_1,\ldots,X_n)] = \int_{\mathcal{X}} T(x_1,\ldots,x_n) f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_1 \ldots dx_n$$

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

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

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

Example 9.1

see HW

Rayleigh distribution has support $(0, \infty)$.

If you used Unif(0,1) as your envelope for Rayleigh dsn, you histogram of samples resulting from an accept-reject algorithm would be biased (too many small values, no large values - above 1).

Example 9.2 Let $X_1, \ldots, X_n$ be a random sample (iid) from a population w/ mean μ and variance $\sigma^2 < \infty$.

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

$$\Rightarrow bias(\bar{X}_n) = E\bar{X}_n - \mu = 0 \Rightarrow \bar{X}_n \text{ is unbiased estimator for pop. mean } \mu.$$

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 $ES^2 = \sigma^2$ but $\hat{\sigma}^2 = \frac{n-1}{n}\cdot S^2$, so

$$E\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2 \Rightarrow \hat{\sigma}^2 \text{ is a biased estimator.}$$

Note: for large n, $S^2 \approx \hat{\sigma}^2$.

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

$$MSE(\hat{\theta}_n) = E\left[(\theta - \hat{\theta}_n)^2\right]$$

can show

$$= Var(\hat{\theta}_n) + \left(bias(\hat{\theta}_n)\right)^2.$$

Generally, we want estimators with

(1) small bias

(2) small variance

often there is a bias - variance trade-off

we can't get both at the same time.

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:

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

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

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

see page 331 of Casella & Berger.

9.3 Standard Error

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

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

standar error = st. dev. of sampling distribution ~~of~~. $\hat{\theta}_n$.

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

Example 9.5

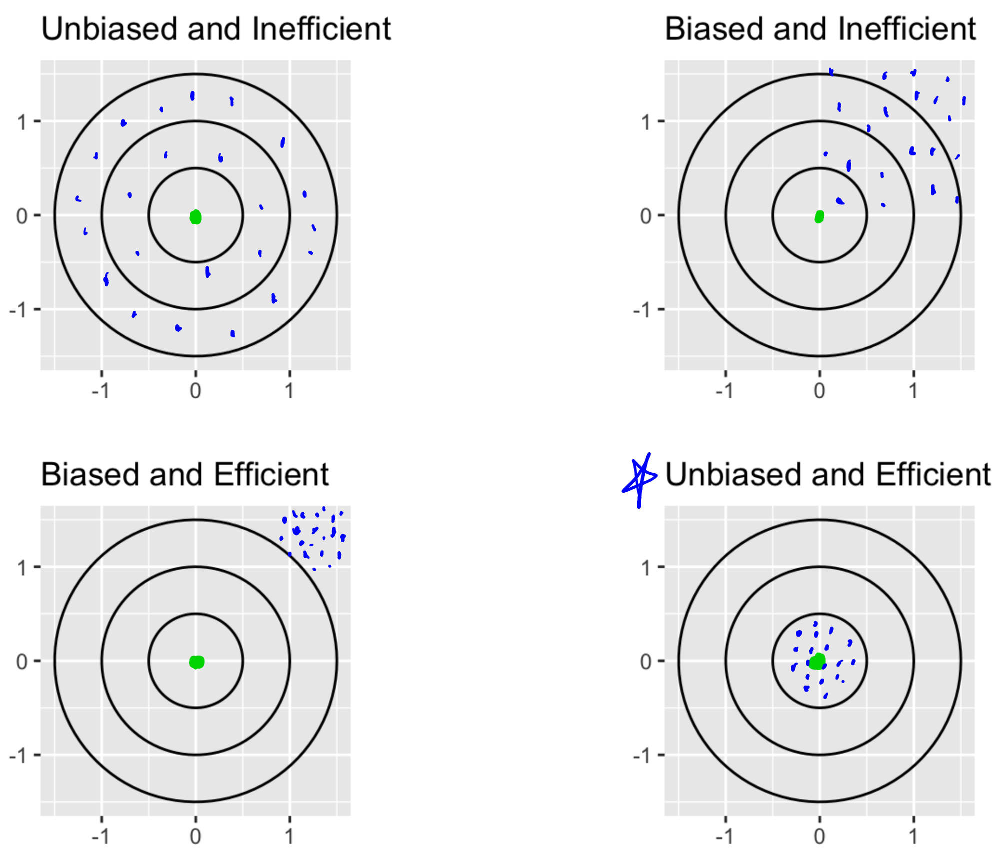
$$se(\bar{X}_n) = \sqrt{Var(\bar{X}_n)} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

10 Comparing Estimators

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

1. Consistency: as $n \uparrow \infty$ does the estimator converge in probability to parameter it is estimating?
2. Bias: Is the estimator unbiased? $E(\hat{\theta}_n) = \theta$?
3. Efficiency: $\hat{\theta}_n$ is more efficient than $\tilde{\theta}_n$ if $Var(\hat{\theta}_n) < Var(\tilde{\theta}_n)$.
4. MSE: Compare $MSE(\hat{\theta}_n)$ to $MSE(\tilde{\theta}_n)$ but remember bias/variance tradeoff:

$$MSE(\hat{\theta}_n) = Var(\hat{\theta}_n) + bias(\hat{\theta}_n)^2$$



variance

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?

↳ which has smaller variance?

```
set.seed(400)
```

→ reproducibility.

```
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))
```

do it 10,000 times.

$n = 100$ sized samples $X_1, \ldots, X_{100}$

Unif(0,1)

N(0,1)

store results. d.f. w/ "times" = rows, 2 columns, one for each statistic.

1 - 10,000 draws.

```
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)
}
```

← draw a sample from Unif(0,1).

← draw a sample from Norm(0,1).

store mean

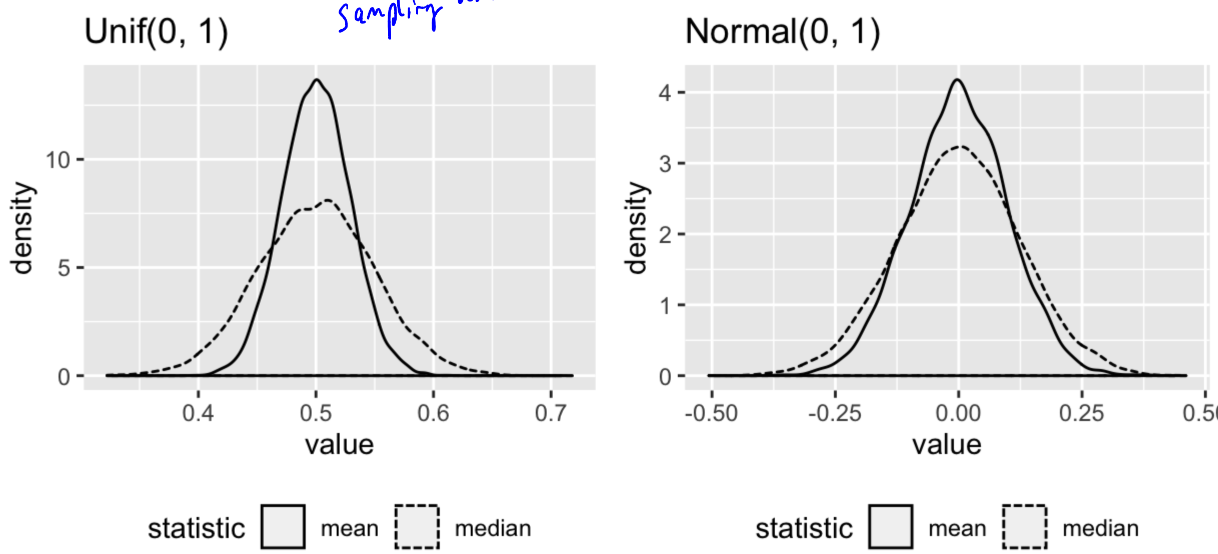
store median.

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

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

estimate density from sample. →

plotting the sampling distribution of each statistic $\bar{X}_n$ and $\text{median}(X_1, \ldots, X_n)$ for $X_1, \ldots, X_n \sim$ 2 dsns sampling dsns



	mean of sampling dsn $\hat{E}[\tilde{\theta}_n]$	$\hat{SE}[\tilde{\theta}_n]$
mean	.4999	0.029
median	.4999	0.0494

true mean = true median = 0.5.

	mean of sampling dsn $\hat{E}[\tilde{\theta}_n]$	$\hat{SE}[\tilde{\theta}_n]$
mean	0.0001	0.1
median	-0.0009	0.12

true mean = true median = 0

Ex: Cauchy distribution

NOTE: this is not the case for all distributions!

When a dsn is heavy tailed, median is more efficient than the mean. (robustness).

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

Bias: both mean and median unbiased

Efficiency: mean is more efficient $\hat{Var}(\text{mean}(X_1,\ldots,X_n)) < \hat{Var}(\text{median}(X_1,\ldots,X_n))$.

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

also efficiency!

$\int_{x} g(x) f_X(x)\,dx$

not always easy to evaluate analytically.