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

finite sample → (1) What is the distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

limit → (2) What is the approximate dsn as $n \to \infty$? Normal.

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

How far is $\bar{X}$ from μ?

statistic; true value we are estimating.

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

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 $\forall \varepsilon > 0, \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$.
- 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 (give 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 $N(0,1)$. (converges in distribution)

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

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.

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"

Statistics estimate *parameters*.

from sample → from population

Example 8.1

$\bar{X}_n$ estimates μ

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

$S = \sqrt{S^2}$ estimate 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 is a point estimator.

A CI is an interval estimator.

(if based on 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).

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

μ fixed but unknown quantities ⟶ 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

a parameter we want to estimate.

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

random quantity.

joint dsn of $X_1, \ldots, X_n$.

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

$E[T(X_1,\ldots,X_n)] = \int_{\underline{x}} T(x_1,\ldots,x_n) f_{\underline{X}}(\underline{x})\, d\underline{x}$

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

$$bias(\hat{\theta}_n) = 0 \iff E[\hat{\theta}_n] = \theta$$

Example 9.1
If you had used $Unif(0,1)$ as your proposal dsn for this Rayleigh dsn, support $(0,\infty)$.
your histogram of samples would be biased.
(too many small values, no large values).

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

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

$\Rightarrow bias(\bar{X}) = E[\bar{X}] - \mu = 0 \Rightarrow$ sample mean is an unbiased estimator for μ.

Example 9.3 Compare 2 estimators for σ^2 for previous example.

Sample variance:

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

vs.

MLE of variance:

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

Can show $E s^2 = \sigma^2$

$\hat{\sigma}^2 = \frac{n-1}{n} \cdot s^2$ so

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

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

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 the bias-variance trade-off (can't get 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.

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

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

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 pg. 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

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

standard error = s.t. deviation of sampling dsn of $\hat{\theta}_n$.

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

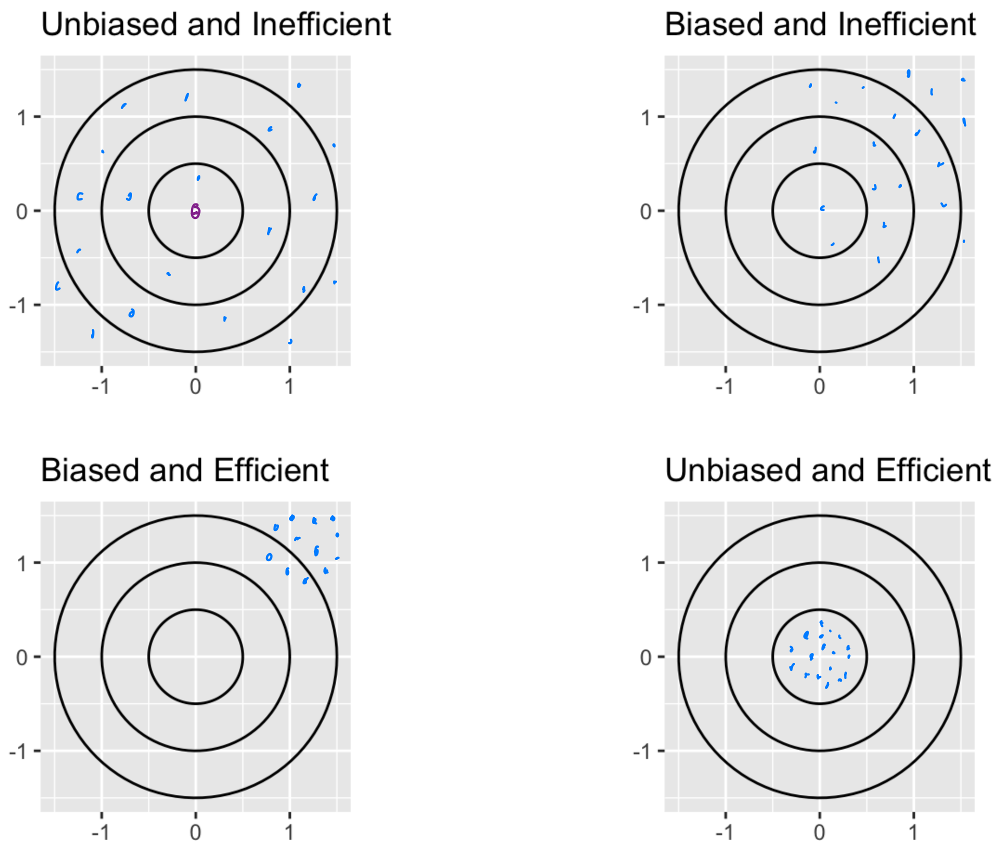
Example 9.5

$$Se(\bar{X}) = \sqrt{Var(\bar{X})} = \sqrt{\frac{Var(X_1)}{n}} = \frac{\sigma}{\sqrt{n}}.$$

10 Comparing Estimators

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

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

↳ 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))
```

number of draws from the sampling dsn.

store results.

for 1 – 10,000 draw from sampling dsn of mean median.

```
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 sample of size 100 from unif

draw sample of size 100 from N(0,1).

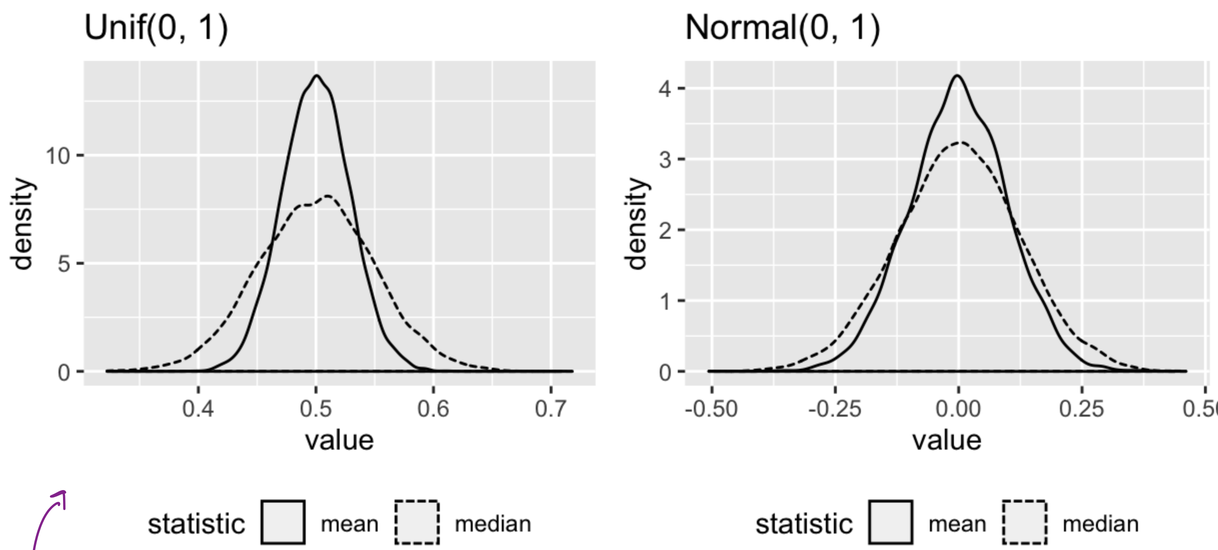
```
uniform_results %>%
  gather(statistic, value, everything()) %>%
  ggplot() +
  geom_density(aes(value, lty = statistic)) +
  ggtitle("Unif(0, 1)") +
  theme(legend.position = "bottom")
```

line type.

estimate the density from the samples.

plot results.

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



estimated sampling dsn of

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

estimated sampling dsn of

	$\bar{X}$	S
mean	.0001	0.1
median	-.0009	0.12

true mean = true median = 0.5

true mean = true median = 0

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

Bias: both mean and median unbiased.

Note: this is not always true for all dsns! when a distribution is heavy tailed. median is more efficient than mean.

Efficiency: mean is more efficient $\hat{Var}(mean(X_1,\dots,X_{100})) < \hat{Var}(median(X_1,\dots,X_{100}))$.

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