Ch. 2 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 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$? $\text{Normal}\left(EX_1, \frac{\text{Var}X_1}{n}\right)$

② How big does n have to be for $\bar{X}_n \sim$ Normal? If $f \sim$ Normal, $n \geq 1$. If $f \not\sim$ Normal, $n = \infty$, but 30 is close enough...

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

How far is $\bar{X}_n$ from μ? true value we are estimating.

statistic (function of r.v.'s)

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

We can evaluate this distance in several ways.

Some modes of convergence –

- almost surely $P(\lim_{n\to\infty} X_n = X) = 1$ $X_n \xrightarrow{as} X$
- in probability $\forall \varepsilon > 0, \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$. $X_n \xrightarrow{p} X$
- in distribution $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ $X_n \xrightarrow{d} X$

Laws of large numbers –

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

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

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

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

7.2 Central Limit Theorem

iid

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 $N(0, 1)$. [convergence in distribution]

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

Interpretation:

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

distribution of the statistic we care about under some sample $X_1 \ldots X_n$

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

$s = \sqrt{s^2}$ 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 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 random variables. → estimator (statistic)

$\overline{x}_n$ function of observed data (an actual #) → estimate (sample statistic)

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 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

joint dsn of $X_1, \ldots, X_n$

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

$E(T(X_1, \ldots, X_n)) = \int_{\mathcal{X}} T(x_1, \ldots, x_n) f_X(x)\,dx$

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

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

support $(0, \infty)$

Example 9.1
If you used Unif(0,1) as your envelope for Rayleigh dsn your histogram of values 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 pop. w/ mean μ and variance $\sigma^2 < \infty$.

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

Example 9.3 Same setup as 9.2, Want to estimate σ^2.

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

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

Can show $E s^2 = \sigma^2$, but $\hat{\sigma}^2 = \frac{n-1}{n} s^2 \Rightarrow E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$

So, s^2 is unbiased and $\hat{\sigma}^2$ is biased.

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

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

can show

Generally, we want estimators with

① small bias

② small variance

often there is a 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.

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

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

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

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

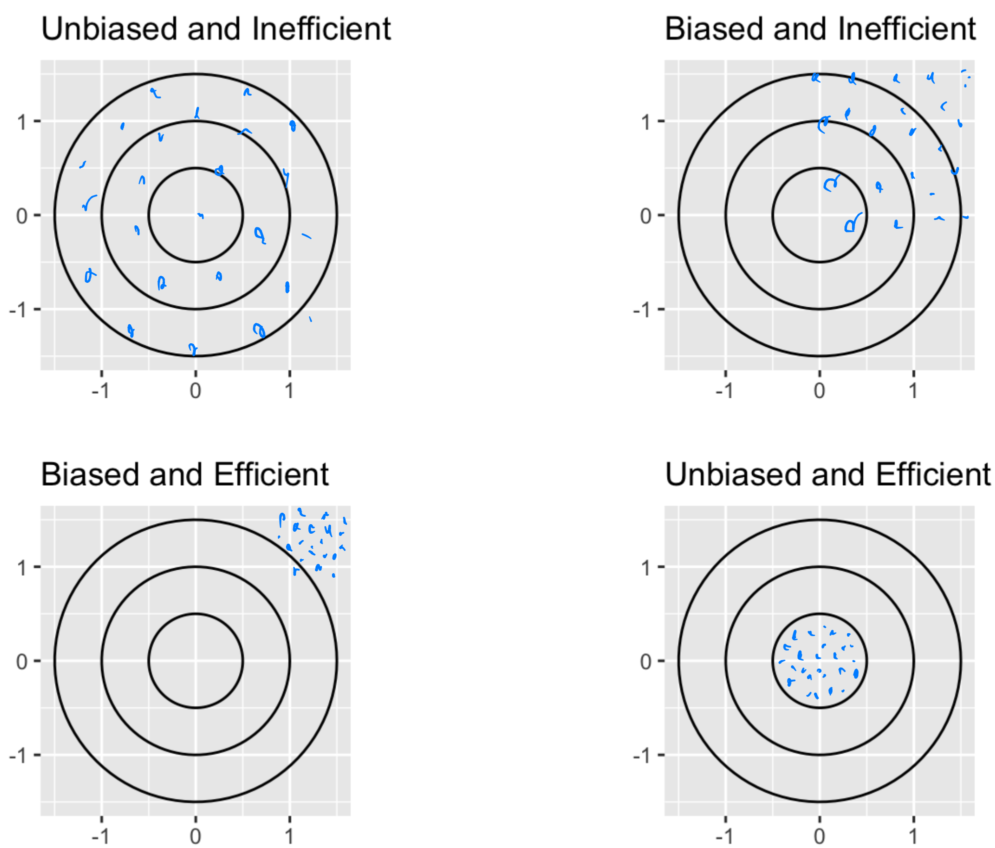
Example 9.5

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

10 Comparing Estimators

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

1. Consistency: as $n \uparrow$ does the estimator converge to the parameter it's estimating? (convergence in probability)
2. Bias: Is the estimator unbiased? $E(\hat{\theta}_n) = \theta$.
3. Efficiency: $\hat{\theta}_n$ is more efficient $\tilde{\theta}_n$ if $Var(\hat{\theta}_n) < Var(\tilde{\theta}_n)$.
4. MSE: Compare $MSE(\hat{\theta}_n)$ to $MSE(\tilde{\theta}_n)$ (want the smallest one), but remember 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?

→ which has smaller variance?

Unif(0,1)

Normal(0,1).

need samples from sampling dsn of mean & median.

each estimator created from sample of size 100

store results

for 1 to 10,000 draw from the sampling dsn.

```
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))

for(i in 1:times) {
  x <- runif(n) ← draw a unif sample
  y <- rnorm(n) ← draw a normal sample
  uniform_results[i, "mean"] <- mean(x) ← store mean
  uniform_results[i, "median"] <- median(x) ← store median.
  normal_results[i, "mean"] <- mean(y)
  normal_results[i, "median"] <- median(y)
}
```

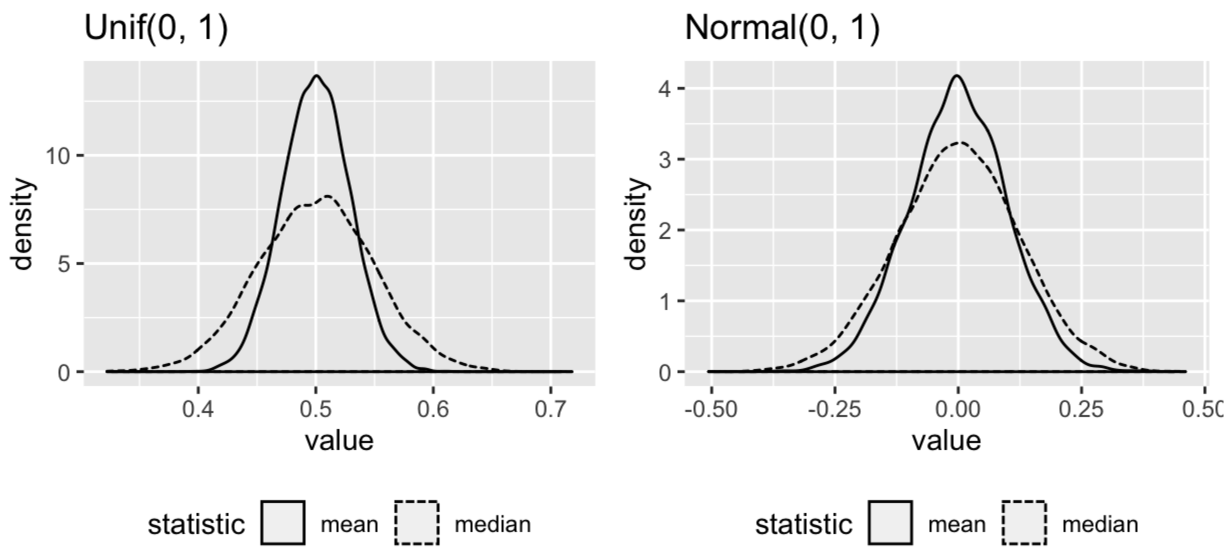
estimate density from samples.

plot samples.

```
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")
```

Sampling dsns.



Sampling dsn of	$\bar{X}$	S
mean	.4999	.029
median	.4999	.0499

true mean = true median = 0.5

Sampling dsn of	$\bar{X}$	S
mean	0.0001	0.10
median	-.0009	0.12

true mean = true median = 0

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

Bias: Both median and mean are unbiased

NOTE: This is not the case for all distributions. When a dsn is heavy tailed, median is more efficient than mean.

EFFICIENCY: mean is more efficient $\hat{\mathrm{var}}(\mathrm{mean}(X_1, \ldots, X_{100})) < \hat{V}(\mathrm{median}(X_1, \ldots, X_{100}))$

Next Up In ~~Ch. 5,~~ Ch. 6 we'll look at a method that produces *unbiased* estimators of $E(g(X))$!