* Announcement

① HW5 due today

② HW6 ~ next Monday (10/5)

③ Midterm (10/9) : Ch1~Ch5 9PM Thursday (10/8) ~ 9PM Friday (10/9)

After today's lecture

→ HW6 Q1, Q2

STAT 88: Lecture 14

Contents

Section 5.4: Unbiased Estimators

Last time

Method of indicator to find $E(X)$

Step 1: Describe X. "random count"

Step 2: Find I_j (jth trial). (What is being counted?)

Step 3: Find $p = P(I_j = 1)$. (same for all indicators)

Step 4: Write X as a sum of indicators:

$$X = I_1 + I_2 + \cdots + I_n.$$

Step 5: Find $E(X)$.

If $X \sim \text{Binomial}(n, p)$, $E(X) = np$.

If $X \sim \text{HG}(N, G, n)$, $E(X) = n\frac{G}{N}$.

Warm up: A drawer contains s black socks and s white socks ($s > 0$). I pull two socks out at random without replacement and call that my first pair. Then I pull out two socks at random without replacement and call that my second pair. I proceed in this way until I have s pairs and the drawer is empty. Find the expected number of pairs in which two socks are of different colors.

Step 1 X = Number of pairs out of s that mismatch

Step 2 $I_j = \begin{cases} 1 & \text{if } s\text{th pair is mismatch} \\ 0 & \text{else} \end{cases}$, $j = 1, \cdots, B$

Step 3 $p = P(I_j = 1) = P(I_1 = 1) = \dfrac{\binom{s}{1}\binom{s}{1}}{\binom{2s}{2}}$ (Hypergeometric formula)

Symmetry

draw 1 black — draw 1 white

$= 2 * \dfrac{s}{2s} * \dfrac{s}{2s-1}$ Alternatively

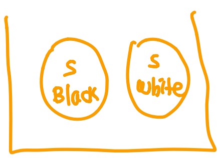
White - Black
Black - white

Step 4 $X = I_1 + I_2 + \cdots + I_s$

Step 5 $E(X) = s \cdot p$

Step3

pop



$p = P(\text{1st pair is mismatch})$

$= P(WB) + P(BW)$

$= P(W)P(B|W) + P(B)P(W|B)$

$= \dfrac{s}{2s} * \dfrac{s}{2s-1} + \dfrac{s}{2s} * \dfrac{s}{2s-1}$

P(1st trial is white sock) $= \dfrac{s}{2s}$

$= 2 * \dfrac{s}{2s} * \dfrac{s}{2s-1}$

Step5

$X = I_1 + I_2 + \cdots + I_s$

$E(X) = E(I_1 + I_2 + \cdots + I_s)$

$= E(I_1) + E(I_2) + \cdots + E(I_s)$

$= p + p + \cdots + p$

$= s \cdot p$

5.4. Unbiased Estimators

Preliminary: Linear Function Rule Let X be a random variable and let $Y = aX + b$. Then Y is a linear function of X. Then

$$E(Y) = E(aX+b) = \sum_{\text{all } x} (ax+b)P(X=x)$$

$$= a\underbrace{\sum_{\text{all } x} xP(X=x)}_{E(X)} + b\underbrace{\sum_{\text{all } x} P(X=x)}_{=1}$$

$$= aE(X) + b.$$

$E(g(X)) = \sum_{\text{all } x} g(x)P(X=x)$

with $g(x) = ax+b$

Terminology Data scientists often want to estimate a parameter of a population.

- A **parameter** is a fixed unknown number associated with the population.

- A **statistic** is a number based on the data in your sample.

- An **estimator** is a statistic used to approximate a parameter.

estimate

- An **unbiased estimator** of a parameter is an estimator whose expected value is equal to the parameter.

Sample mean as an estimator of population mean

Ex Estimate the average annual income in California, μ.

parameter

$\to E(X_i) = \mu$ for $i = 1, \ldots, n$

Each X_i from the same distribution (population)

Suppose you draw a random sample of size n. $X_1, \ldots, X_n$ are sample incomes. The sample average is the statistic $\bar{X}$ defined as the function

$$\bar{X} = g(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- statistic

Important: μ is constant, unknown; $\bar{X}$ estimator, random variable

$\bar{X}$ is unbiased if $E(\bar{X}) = \mu$. In fact,

Pf

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

repeated samples →

Long run avg. value of $\bar{X}$, On avg, it is equal to μ

$E(aX) = aE(X)$ for any const a

Additivity $E(X+Y) = E(X) + E(Y)$

$n=3$ pop

Today $X_1 = 50K$, $X_2 = 60K$, $X_3 = 70K$ → $\bar{X} = 60K$

Tmrw $X_1 = 65K$, $X_2 = 50K$, $X_3 = 60K$ → $\bar{X} = \frac{50K + 60K + 65K}{3}$

⋮

Avg → μ

Which of these estimators of μ is unbiased?

(a) X_{15}.

(b) $(X_1 + X_{15})/15$.

(c) $(X_1 + 2X_{100})/3$.

Draw a sample X from your pop. whose mean $= \mu$.

Example $\quad E(X) = \mu$.

$X_1 \sim$ Binomial (n,p). $n \ast p$

$E(X_1) = np$

$X_2 \sim$ Binomial $(n,p) \rightsquigarrow E(X_2) = np$

(a) $E(X_{15}) = \mu$ Unbiased

(b) $E\left(\frac{X_1+X_{15}}{15}\right) = \frac{1}{15} E(X_1+X_{15}) = \frac{E(X_1)+E(X_{15})}{15} = \frac{2\mu}{15}$ biased

(c) $E\left(\frac{X_1+2X_{100}}{3}\right) = \frac{1}{3} E(X_1+2X_{100}) = \frac{1}{3} * (\mu+2\mu) = \mu$ unbiased

If we have a biased estimator how can we make it unbiased?

Let's make $\frac{X_1+X_{15}}{3}$ unbiased.

$$
\begin{aligned}
E\left(\frac{X_1+X_{15}}{3}\right) &= \frac{E(X_1)+E(X_{15})}{3} \\
&= \frac{\mu+\mu}{3} \\
&= \frac{2\mu}{3}.
\end{aligned}
$$

$\frac{3}{2} E\left(\frac{X_1+X_{15}}{3}\right) = \mu$

$E\left(\frac{3}{2} * \left(\frac{X_1+X_{15}}{3}\right)\right) = \mu$

$\parallel$

$E\left(\underbrace{\frac{X_1+X_{15}}{2}}_{\text{unbiased estimator}}\right)$

(a special case of sample mean when the pop. consists of 0s and 1s)

Sample proportion as an estimator of population proportion

When the population consists of zeros and ones, the population mean is the population proportion of ones.

Example

population

| 0 0 1 1 1 |

— population mean = p = 3/5 (= μ)

$X_1, \dots, X_n$ samples w/ replacement from pop.

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ = sample proportion, $E(\bar{X}) = p$

Example You roll a die 30 times and find the sample proportion of sixes. The population consists of $\{0, 0, 0, 0, 0, 1\}$. Repeat experiment 20,000 times and plot distribution of sample proportions.

1 2 3 4 5 6

↓ draw w/ replacement

Exp 1: $X_1, X_2, \dots, X_{30} \to \bar{X}^{(1)}$

Exp 2: $X_1, X_2, \dots, X_{30} \to \bar{X}^{(2)}$

⋮

Exp 20,000: $X_1, \dots X_{30} \to \bar{X}^{(20000)}$

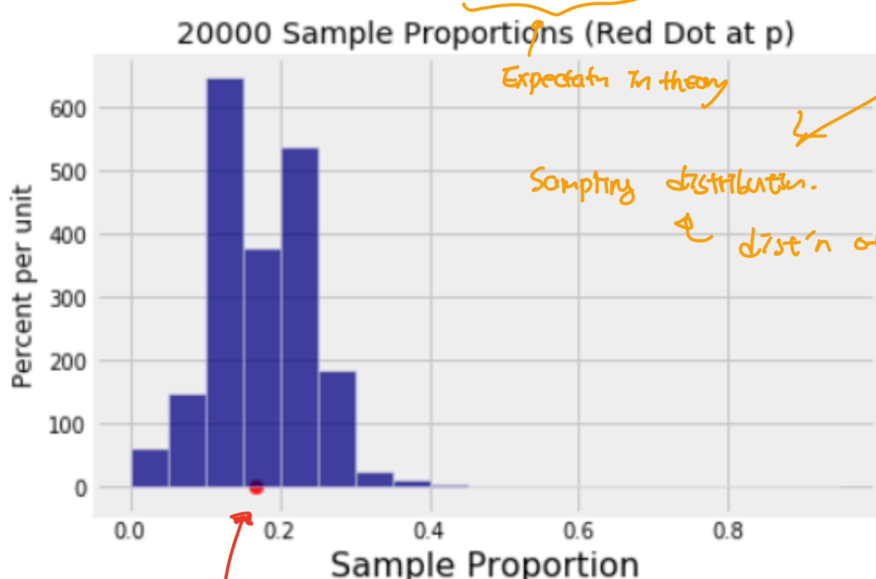
} 20,000 different observed values of $\bar{X}$

```
n = 30
p = 0.1667
Average of observed sample proportions = 0.1664
```

$\frac{1}{6} =$

$\underbrace{0.1667 = p}_{\text{Expectation in theory}} = E(\bar{X}) \approx \frac{1}{20000}\sum_{i=1}^{20000} \bar{X}^{(i)} = \underbrace{0.1664}_{\text{Expectation in numerical exp.}}$

histogram



Sampling distribution.

← dist'n of your estimator (= $\bar{X}$)

p since E(sample proportion) = p

Estimating the largest possible value

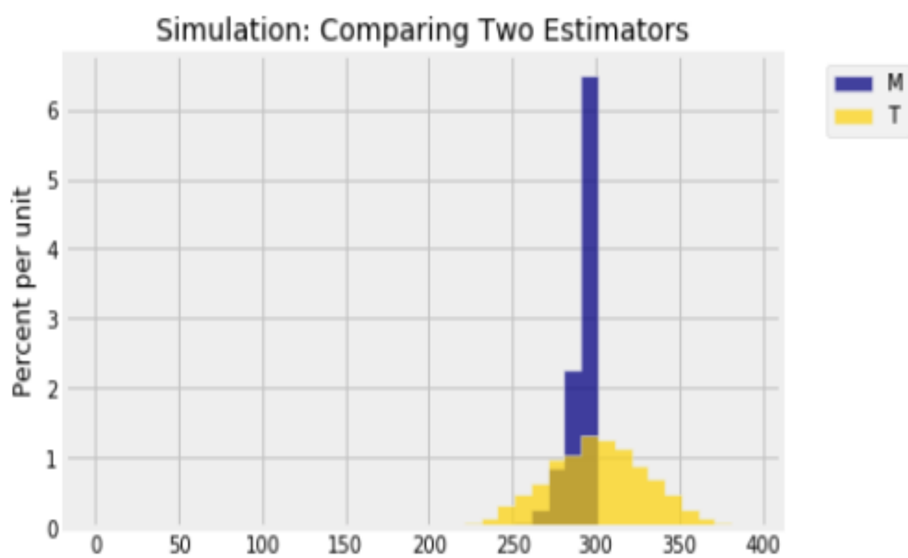
Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim}$ Uniform$\{1, 2, \ldots, N\}$ for some fixed but unknown N. To estimate N, you may think $M = \max\{X_1, \ldots, X_n\}$ and this is an estimator but we want an unbiased estimator.

The population mean is $\mu = (N+1)/2$ and $E(\bar{X}) = (N+1)/2$ since it is unbiased. What is an estimator such that

$$E(\text{estimator}) = N?$$

Lets look at sampling distribution of (1) $T = 2\bar{X} - 1$ and (2) $M = \max(X_1, \ldots, X_n)$.

```
N = 300
n = 30
5000 repetitions
```



The histograms show that both estimators have pros and cons.

M - Pros: small spread of values; Cons: biased.

T - Pros: unbiased; Cons: big spread of values.

Unbiasedness is a good property, but so is low variability. Bias-variance tradeoff