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 B black socks and B white socks ($B > 0$). 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 B pairs and the drawer is empty. Find the expected number of pairs in which two socks are of different colors.

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

$$
\begin{aligned}
E(Y) = E(aX+b) &= \sum_{\text{all } x} (ax+b)P(X=x) \\
&= a\sum_{\text{all } x} xP(X=x) + b\sum_{\text{all } x} P(X=x) \\
&= aE(X) + b.
\end{aligned}
$$

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

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.

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

<u>Sample mean as an estimator of population mean</u>

<u>Ex</u> Estimate the average annual income in California, μ.

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

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

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

Important:

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

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

$E(aX) = aE(X)$

Which of these estimators of μ is unbiased?

(a) X_{15}.

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

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

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

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

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

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

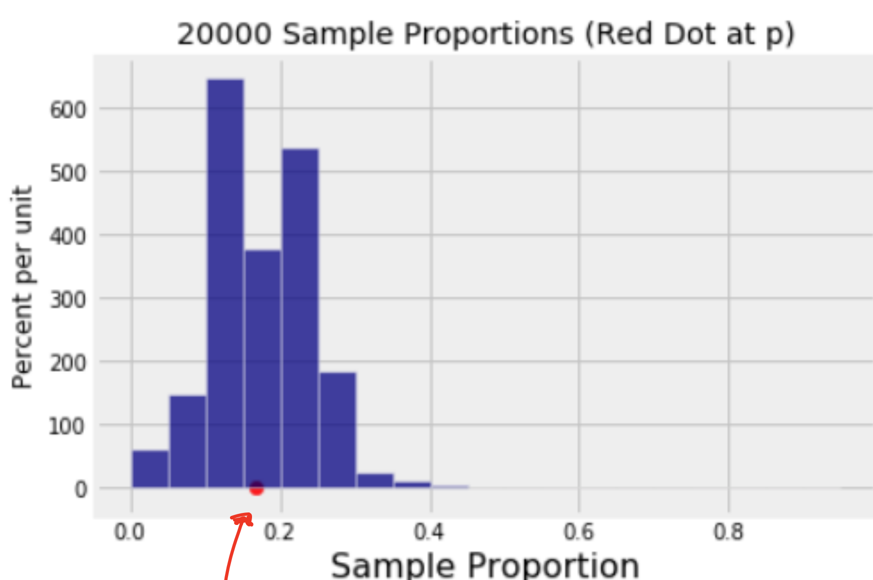
population

$\boxed{0\ 0\ 1\ 1\ 1}$ — population mean $= p = 3/5$

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

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



p since E(sample proportion) = p

Estimating the largest possible value

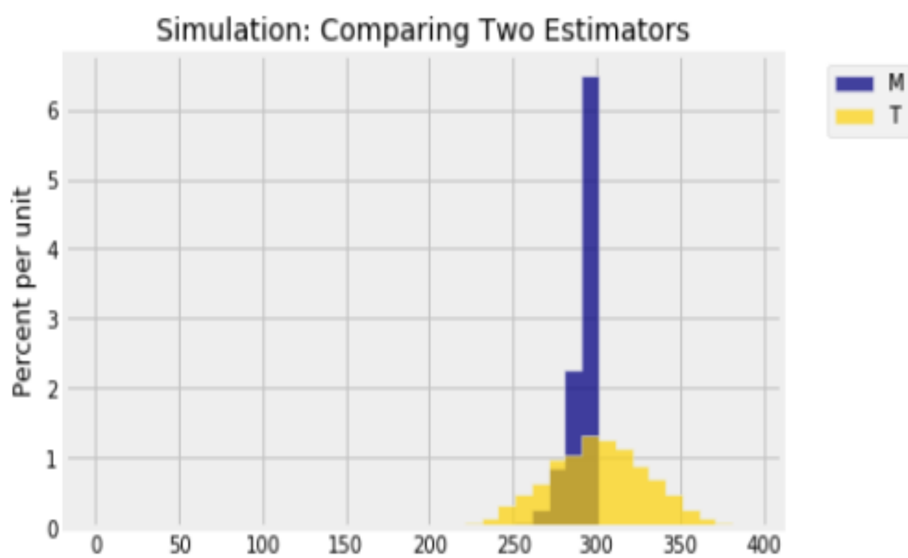
Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim}$ Uniform$\{1, 2, \ldots, N\}$ for some fixed but unknown N. To estimate N, you may think $M = \max\{X_1, \ldots, X_n\}$ and this is an estimator but we want an unbiased estimator.

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