Stat 88: Probability & Mathematical Statistics in Data Science

THE AUTHOR OF THE WINDOWS FILE COPY DIALOG VISITS SOME FRIENDS.

https://xkcd.com/612/

Lecture 15: 2/24/2021

Wrap up method of indicators, unbiased estimators

Sections 5.3, 5.4

Missing classes

- We can use indicators to compute the chance that something *doesn't* occur.

- For example, say we have a box with balls that are red, white, or blue, with 35% being red, 30% being white, and 35% blue. If we draw n times with replacement from this box, what is the expected number of colors that *don't* appear in the sample?

X = # of colors that don't appear.

$I_{A_k} \Leftrightarrow I_k$

A_k = the event that the k^{th} color doesn't appear in n draws w/repl.

$k=1 \Leftrightarrow$ Red
$k=2 \Leftrightarrow$ White
$k=3 \Leftrightarrow$ Blue.

$P(A_1) = P(\text{no R in n draws})$
$= (0.65)^n = \mathbb{E}(\underbrace{I_1}_{I_{A_1}})$.

$$I_{A_k} = \begin{cases} 1 & \text{if } A_k \text{ is true} \\ 0 & \text{if not} \end{cases}$$

$X = I_1 + I_2 + I_3$

$$\mathbb{E}(X) = (0.65)^n + (0.7)^n + (0.65)^n$$

$\mathbb{E}(I_{A_k}) = P(A_k)$

Y = # of colors that DO appear, $Y = 3 - X$, $\mathbb{E}(Y) = 3 - \mathbb{E}(X)$

Examples

1. An instructor is trying to set up office hours during RRR week. On one day there are 8 available slots: 10-11, 11-noon, noon-1, 1-2, 2-3, 3-4, 4-5, and 5-6. There are 6 GSIs, each of whom picks one slot. Suppose the GSIs pick the slots at random, independently of each other. Find the expected number of slots that no GSI picks.

X = number of slots that no one picks

8 slots, Need to define $A_1, \ldots, A_8$

A_k = event that k^{th} slot was not chosen by any GSI

$$P(A_k) = \left(\frac{7}{8}\right)^6 = \mathbb{E}(I_{A_k}) \equiv \mathbb{E}(I_k)$$

$I_{A_k} \equiv I_k$

is equivalent to

$$X = I_1 + I_2 \cdots + I_8$$

$$\mathbb{E}(X) = \sum_{k=1}^{8} \mathbb{E}(I_k) = 8 \cdot \left(\frac{7}{8}\right)^6$$

2. A building has 10 floors above the basement. If 12 people get into an elevator at the basement, and each chooses a floor at random to get out, independently of the others, at how many floors do you expect the elevator to make a stop to let out one or more of these 12 people?

$I_k \Leftrightarrow I_{A_k}$

$X = I_{A_1} + I_{A_2} + \dots + I_{A_{10}} = I_1 + I_2 + \dots + I_k.$

A_k = event that the elevator stops on k^{th} floor (at least one person gets out)

$P(A_k) = 1 - P(\text{no one chooses floor } k)$

$= 1 - \left(\frac{9}{10}\right)^{12}$

$I_k = \begin{cases} 1, & A_k \text{ is true} \\ 0, & \text{otherwise} \end{cases}$

$\mathbb{E}(I_k) = P(A_k) = 1 - \left(\frac{9}{10}\right)^{12}$

$I_k \equiv I_{A_k}$

A_k

$\mathbb{E}(X) = \sum_{k=1}^{10} \mathbb{E}(I_k) = 10 \cdot \left(1 - \left(\frac{9}{10}\right)^{12}\right)$

5.4 Unbiased Estimators

Vocabulary slide.

- We showed the linearity of expectation earlier: $\boldsymbol{E(aX+b)=aE(X)+b}$
- We often want to estimate a ***population parameter***: some fixed number associated with the population, possibly unknown
- A statistic is any number that is computed from the data sample. Usually we use a *random sample.*
- Note that the parameter is *constant* and the statistic is a *random variable.*
- We will use a ***statistic*** to *estimate* (guess at the value of; approximate) the parameter. It is called an ***estimator*** of the parameter.
- If the expectation of the statistic is the parameter that it is estimating, we call the statistic an ***unbiased estimator of the parameter***.

S : statistic   Parameter θ $E(S)=\theta$

An example of an unbiased estimator: $\boldsymbol{E(\bar{X}) = \mu}$

- Let $X_1, X_2, \ldots, X_n$ be our random sample, and the sample mean is $\bar{X}$
- $\bar{X}$ is computed from the sample and will change depending on the sample values, so is a *random variable*.
- If $X_1, X_2, \ldots, X_n$ which are random draws from the population, all have expectation μ, what is the expectation of $\bar{X}$?

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}, \quad \text{Say } E(X_k) = \mu.$$

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{n} \underbrace{\sum_{i=1}^{n} \mu}_{\underbrace{\mu + \mu + \mu + \cdots + \mu}_{n \text{ times}}} = \frac{1}{n} \cdot n \cdot \mu = \mu.$$

Understanding unbiased parameters

- Let $X_1, X_2, \ldots, X_n$ be random draws from the population, all have expectation μ.
- If an estimator S is unbiased, then ***on average,*** it is equal to the number it is trying to estimate $\mathbb{E}(S) = \theta \leftarrow$ parameters.
- Which of the following are unbiased estimators of μ?

(a) X_{15}

(b) $\frac{X_1+X_{15}}{15}$

(c) $\frac{X_1+2X_{100}}{3}$

(d) How to make an biased estimator unbiased?

(e) If X_1 is unbiased, why bother taking the mean? Why not just use X_1? has to do with accuracy.

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

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

NOT UNBIASED

Understanding unbiased parameters

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

(d) let's consider $S = \frac{X_1 + X_{15}}{15}$ ← biased

b/c $E(S) = \frac{2\mu}{15}$

Look at $\frac{15S}{2}$ $\mathbb{E}\left(\frac{15S}{2}\right) = \mathbb{E}\left(\frac{\cancel{15}}{2} \cdot \frac{(X_1 + X_{15})}{\cancel{15}}\right)$

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

A special estimator: The sample proportion $\hat{p}$

estimator of p.

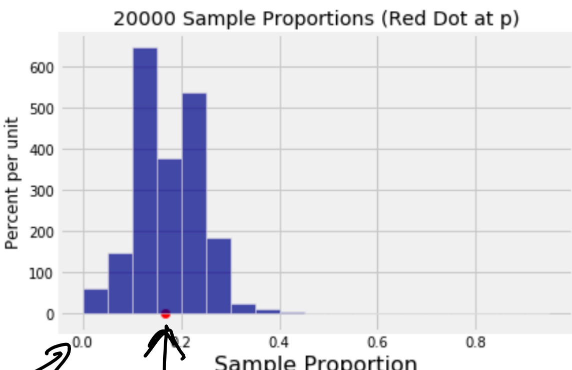
- Usual special case of population binary outcomes represented by 0 and 1
- Sum of draws = # of 1s that are in the sample (sample sum)
- Sample mean = proportion of 1s in sample

$\bar{X}$ = prop. of 1's in sample.

$\bar{X} \equiv \hat{p}$

- Consider a population of 0s and 1s, and draw n times from this population, **with** replacement: $X_1, X_2, \ldots, X_n$ are the draws, note that each of the X_k are Bernoulli or indicator random variables, with parameter p where p = proportion of 1s in the population. $\leftrightarrow P(\boxed{1}) = p$ $\quad \boxed{0}^{1-p} \ \boxed{1}^{p}$

- Note that the population mean $\mu = p$ and the sample mean $\bar{X} = \hat{p}$, and $\bar{X}$ is an unbiased estimator of p



$\mathbb{E}(\bar{X})$

0.1666667 ✓
0.2666667 ✓
0.2333333 ✓
0.2000000 ✓
0.1000000 ✓

$\bar{X} \approx 0.194$

$\underbrace{\{0, 0, 0, 0, 0, 1\}}_{\text{sampled 30 times \& looked fraction of times } \boxed{\cdot}}$

$\mathbb{E}(\hat{p}) = p$

$\mathbb{E}(\bar{X}) = p$

roll a fair die 30 times & count # of times $\boxed{\cdot}$ Ace

$P(\text{success}) = \frac{1}{6}$

Repeating 20,000 times, avg fraction times $\boxed{0} \approx 0.1667 \approx \frac{1}{6}$

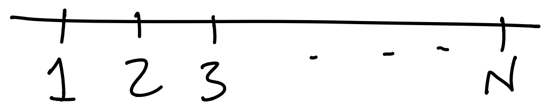
Estimating the largest possible value

- $X_1, X_2, \dots, X_n$ are drawn at random with replacement from $\{1, 2, \dots, N\}$. That is, they are independent and identically distributed random variables with the discrete uniform distribution on 1, 2, … N.

- We want to estimate N using an unbiased estimator. Does the sample mean work?

$$\mathbb{E}(\bar{X}) = \mu = \frac{N+1}{2} \leftarrow \text{Not unbiased}$$

$$\mathbb{E}(\bar{X}) \neq N$$



What would be an unbiased estimator?

$$\boxed{T = 2\bar{X} - 1}$$

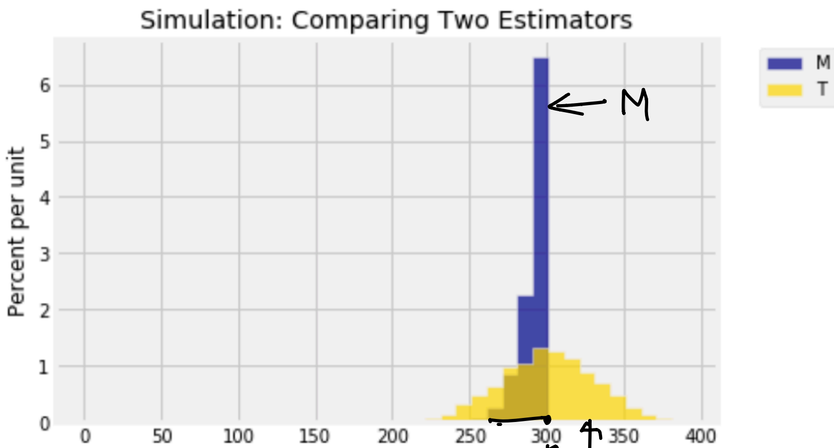
unbiased estimator of μ

$$\mathbb{E}(T) = N.$$

Comparing two estimators: T and M (max sample value)

iid Unif{1,2,..N}

- Let $X_1, X_2, \dots, X_n$ be as earlier, and let $M = \max\{X_1, X_2, \dots, X_n\}$. Below are histograms for M and $T = 2\,\bar{X} - 1$, from simulations assuming that N=300 and that the sample size is 30 (5,000 repetitions, computing T, M each time).



N T

1, 7, 8

M = 8 = max value sample

$T = 2\bar{X} - 1$

$E(T) = N$ (300 in this example)

$E(M) \neq N$

- pros & cons for M : M is not unbiased, (con) less variation (pro)
- pros & cons for T :

pro : unbiased

con : higher variation

Bias-variance trade off.

Example: (5.7.11)

A data scientist believes that a randomly picked student at his school is twice as likely not to own a car as to own one car. He knows that no student has three cars, though some students do have two cars. He therefore models the probability distribution for the number of cars owned by a random student as follows. The model involves an unknown positive parameter θ.

# of cars	0	1	2
Probability	2θ	θ	$1-3\theta$

(a) Find $E(X_k)$

(b) Let $X_1, X_2, \dots X_n$ be the numbers of cars owned by n random students picked independently of each other. Assuming that the data scientist's model is good, use the $\underbrace{\text{entire sample to}}_{\text{hint use } \bar{X}}$ construct an unbiased estimator of θ.

Example: (5.7.11)