Probability and Mathematical Statistics in Data Science

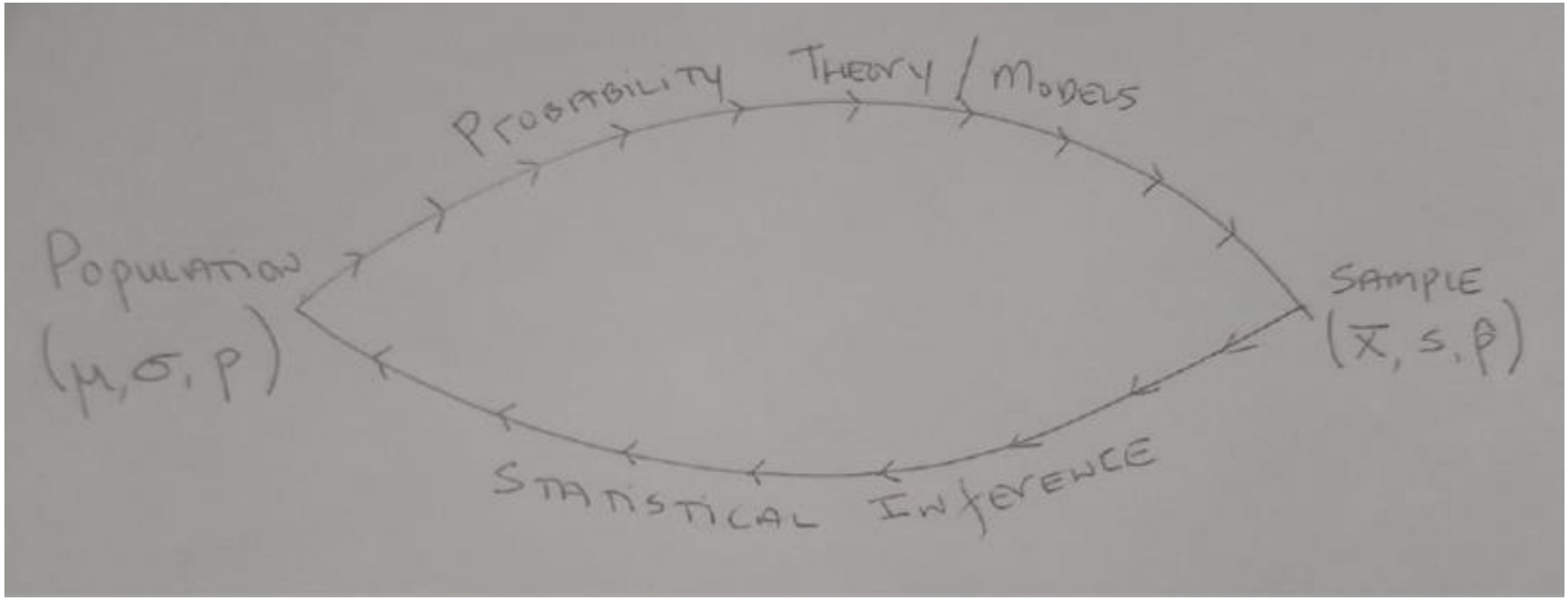
Lecture 14: 5.3: Methods of Indicators cont. 5.4: Unbiased Estimators

Example

- 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?

Probability and Statistical Inference



Point Estimate

A **point estimate** of a parameter θ is a single number that can be regarded as a sensible value for θ. A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the **point estimator** of θ.

Unbiased Estimators

- We often want to estimate a *population parameter*: some fixed number associated with the population
- 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*** 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.

Unbiased Estimators

A point estimator $\hat{\theta}$ is said to be an **unbiased estimator** of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ. If $\hat{\theta}$ is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the **bias** of $\hat{\theta}$.

Point Estimates : Example

- Consider the following values of a normally distributed random variable (lifetimes of a battery):

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

a. Estimator $= \overline{X}$, estimate $= \overline{x} = \Sigma x_i / n = 555.86/20 = 27.793$

b. Estimator $= \widetilde{X}$, estimate $= \widetilde{x} = (27.94 + 27.98)/2 = 27.960$

c. Estimator $= [\min(X_i) + \max(X_i)]/2 =$ the average of the two extreme lifetimes, estimate $= [\min(x_i) + \max(x_i)]/2 = (24.46 + 30.88)/2 = 27.670$

- Which of the estimators are closet to the population mean?

Measure of Good Estimators

In principle, there are many estimators for a given parameters. Two properties of estimators are desired.

Unbiasedness This is about how faithful the estimator is. A point estimator $\hat{\theta}$ is an unbiased estimator of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ. If $\hat{\theta}$ is not unbiased, then $E(\hat{\theta}) - \theta$ is called the bias of $\hat{\theta}$.

Small Variance This is about how stable the estimator is. Unbiased estimators are faithful in the long run, but might have large fluctuation when sample size is small.

Principles of Selecting Estimators

- First, choose the estimators that are unbiased.
- Then, among the unbiased estimators, choose the one with the smallest variance.

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 $\bar{X}$?**

Understanding Unbiased Estimators

- 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
- 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) If X_1 is unbiased, why bother taking the mean? Why not just use X_1?

Point Estimate: The Sample Proportion

- An automobile manufacturer has developed a new type of bumper, which is supposed to absorb impacts with less damage than previous bumpers. The manufacturer has used this bumper in a sequence of 25 controlled crashes against a wall, each at 10 mph, using one of its compact car models. Let X = the number of crashes that result in no visible damage to the automobile.

- *Let* p = P(no damage in a single crash). If X is observed to be x =15, the most reasonable estimator and estimate are:

$$\text{estimator } \hat{p} = \frac{X}{n} \qquad \text{estimate} = \frac{x}{n} = \frac{15}{25} = .60$$

Point Estimate: The Sample Proportion

The sample proportion X/n is used as an estimator of p, where X, the number of sample successes, had a binomial distribution with parameters n and p. Thus

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$$

When X is a binomial rv with parameters n and p, the sample proportion $\hat{p} = X/n$ is an unbiased estimator of p.

- No matter what the true value of p is, the distribution of the estimator will be centered at the true value.

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 $X_1, X_2, \ldots X_n$ 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 entire sample to construct an unbiased estimator of θ.