Theory of Estimation

Course Taught at SUAD

Dr. Tanujit Chakraborty
Faculty @ Sorbonne
tanujitisi@gmail.com

Quote of the day..

Live as if you were to die tomorrow.
Learn as if you were to live forever

Mahatma Gandhi

Today's Topics…

- Principle of Statistical Inference (SI)
- Point Estimation
- Interval Estimation

Statistical inference

As a task of statistical inference, we usually follow the following steps:

- **Data collection**
  - Collect a sample from the population.
- **Statistics**
  - Compute a statistics from the sample.
- **Statistical inference**
  - From the statistics we made various statements concerning the values of population parameters.
    - For example, population mean from the sample mean, etc.

Basic terminologies

Some basic terminology which are closely associated to the above-mentioned tasks are reproduced below.

- **Population:** A population consists of the totality of the observation, with which we are concerned.
- **Sample:** A sample is a subset of a population.
- **Random variable:** A random variable is a function that associates a real number with each element in the sample.
- **Statistic:** Any function of the random variable constituting random sample is called a statistic.
- **Statistical inference:** It is an analysis basically concerned with generalization and prediction.

Statistical Inference

There are two facts, which are key to statistical inference.

1. Population parameters are fixed number whose values are usually unknown.
2. Sample statistics are known values for any given sample, but vary from sample to sample, even taken from the same population.

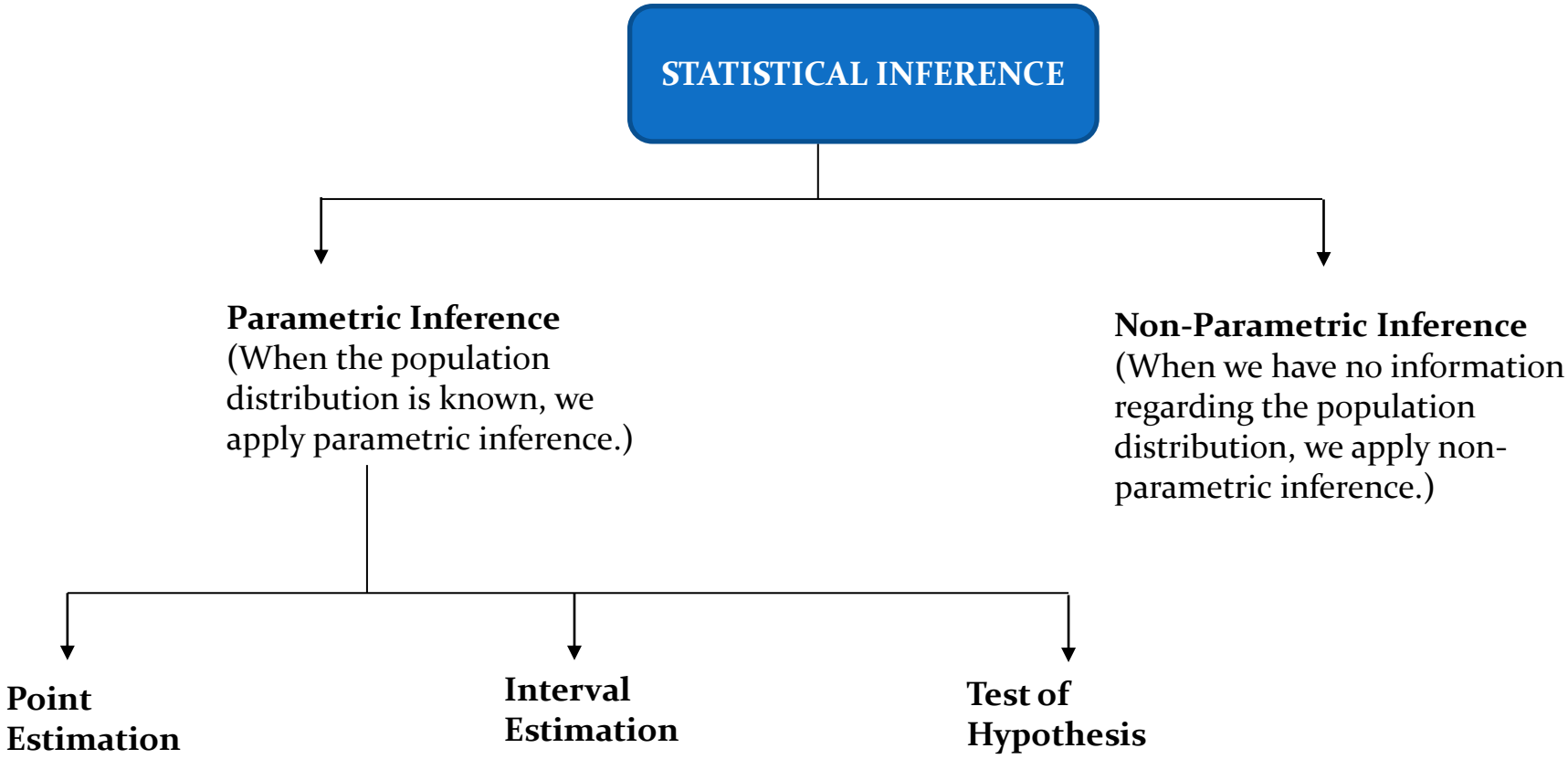
- In fact, it is unlikely for any two samples drawn independently, producing identical values of sample statistic.
- In other words, the variability of sample statistic is always present and must be accounted for in any inferential procedure.
- This variability is called sampling variation.

Note:

A sample statistic is random variable and like any other random variable, a sample statistic has a probability distribution.

Note: Probability distribution for random variable is not applicable to sample statistics.

Taxonomy of Statistical Inference



Introduction: Statistical Inference

- Descriptive analysis is valid only for the data set under consideration and cannot necessarily be generalized to other data.
- Statistical Inference allows us to infer from the sample data about the population of interest.
- It is not feasible to consider entire population for a analysis, hence we need to collect a representative sample.
- Whole discussion can be divided into two parts :
  - The first part deals with the properties of estimators.
  - The second part deals with the methods for finding estimators.

Basic Inferential Approaches

Approach 1: Point and Confidence interval measurement

- We estimate one (or more) parameter(s) using sample statistics.
  - This estimation can be done in the form of a single estimated value (Point Estimation)
  - This estimation usually done in the form of an interval (Interval Estimation).
- Accuracy of the decision is expressed as the **level of confidence** we have in the interval.

Approach 2: Hypothesis testing

- We conduct **test on hypothesis**.
  - We hypothesize that one (or more) parameter(s) has (have) some specific value(s) or relationship.
- Make our decision about the parameter(s) based on one (or more) sample statistic(s)
- Accuracy of the decision is expressed as the probability that the decision is incorrect.

Estimation

- There are two types of estimations:
  - Single point estimate
    - For example, sample mean is a single point estimate.
    - This may vary from one sample to another.
    - This is called zero probability of being correct.
    - Not robust and reliable.
  - Interval estimated
    - Estimate with a range of values, for example, population mean is $20 \leq \mu \leq 22$
    - Reliable and robust with essentially non-zero probability of being correct.
    - An alternative method to statistical learning.
    - Popularly known as Confident Interval measurement.

Properties of Point Estimation

- The primary goal in statistical inference is to find a good estimate of population parameters.
- The parameters are associated with the Prob. Dist. Which is believed to characterize the population.
- If these parameters are known, then one can characterize the entire population.
- In practices, these parameters are unknown, so the objective is to estimate them.
- One can attempt to obtain them based on a function of the sample values.
- But what does this function look like; and if there is more than one such function, then which is the best?
- The answer is given by various statistical concepts such as bias, variability, consistency, efficiency, sufficiency and completeness of the estimates.

Discussion on Statistical Concept

- Assume $x = (x_1, x_2, \dots, x_n)$ are the observations of a random sample from a population of interest i.e., $x_1, x_2, \dots, x_n$ are the n observations collected on the random variable X.
- Any function of random variables is called a **Statistic**. It follows that a statistic is also a random variable.
- Consider a statistic $T(X)$ which is used to estimate a population parameter θ. We say $T(X)$ is an **Estimator** of θ. We write: $\hat{\theta} = T(X)$.
- When T is calculated from the sample values $x_1, x_2, \dots, x_n$, we write $T(x)$ and call it an **Estimate** of θ.
  - Example: $T(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an Estimator and $T(X)$ is a random variable, but $T(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the estimated value from the realized sample values $x_1, x_2, \dots, x_n$, and $T(x)$ is the observed value dependent on the actual sample.

Unbiasedness and Efficiency

- An estimator $T(X)$ is called an **unbiased estimator** of θ if
$$E_\theta(T(X)) = \theta.$$
- The bias of an estimator $T(X)$ is defined as
$$Bias_\theta(T(X)) = E_\theta(T(X)) - \theta.$$
- The variance of $T(X)$ is defined as
$$Var_\theta(T(X)) = E[(T(X) - E(T(X)))^2]$$
- Both bias and variance are measures which characterize the properties of an estimator.
- In statistical theory, we search for “good” estimators in the sense that the bias and the variance are as small as possible and therefore the accuracy is as high as possible.
- A measure which combines bias and variance into one measure is the mean squared error.

MSE, Bias and Variance

- In statistical theory, we search for “good” estimators in the sense that the bias and the variance are as small as possible and therefore the accuracy is as high as possible.

- A measure which combines bias and variance into one measure is the mean squared error (MSE).

- **MSE** = Bias^2 + Variance (Relevance in Statistics & ML)

Exercises:

- Let X follows Binomial (n,p) distribution where n is known and $0 \leq p \leq 1$. Find the unbiased estimator of p.

- $X_1, X_2, \ldots, X_n$ is a random sample from a normal population $N(\mu, \sigma^2)$. Show that S^2 is an unbiased estimator of the parameter σ^2.

- Find the unbiased estimator of μ^2, where $X_i \sim N(\mu, \sigma^2)$.

Consistency of Estimators

- For a good estimator, as the sample size increases, the values of the estimator should get closer to the parameter being estimated. This property of estimators is referred to as consistency.

- Definition: Let $T_1, T_2, \dots, T_n$ be a sequence of estimators for the parameter θ where $T_n = T_n(X_1, X_2, \dots, X_n)$ is a function of $X_1, X_2, \dots, X_n$. The sequence $\{T_n\}$ is a consistent sequence of estimators for θ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P[|T_n - \theta| < \epsilon] = 1$$

$$\text{or,} \quad \lim_{n\to\infty} P[|T_n - \theta| \geq \epsilon] = 0$$

Consistency of Estimators

- A Sufficient Condition for Consistency:

- Let $\{T_n\}$ be a sequence of estimators such that for all $\theta \in \Theta$,

$$(i) E_\theta(T_n) \to \gamma(\theta), n \to \infty \ \ and \ \ (ii)\ Var_\theta(T_n) \to 0, n \to \infty.$$

Then T_n is a consistent estimator of $\gamma(\theta)$.

Proof !

Exercises:

- Prove that in sampling from a $N(\mu, \sigma^2)$ population, the sample mean is a consistent estimator of μ.

- Let $X_1, X_2, \dots, X_n$ follows Poisson distribution with Parameter λ. Prove that the sample mean is a consistent estimator of λ.

Point Estimation

- In general case, properties such as unbiasedness and efficiency cannot be guaranteed for a finite sample.
- But often, the properties can be shown to hold asymptotically.
- Previously we have used several estimators without stating explicitly that they are estimators.
- Now question is that how to obtain a good statistic to estimate an unknown parameter i.e. for example how to determine that sample mean can be used to estimate μ.
- The Maximum Likelihood provides such an approach.

Method of Moments

- Let $f(x;\theta_1,\theta_2,\dots,\theta_k)$ be the density function of the parent population with k parameters $\theta_1,\theta_2,\dots,\theta_k$.

- If μ'_r denotes the rth moment about origin, then

$$\mu'_r = \int_{-\infty}^{\infty} x^r f(x;\theta_1,\theta_2,\dots,\theta_k)\mathrm{dx}, \qquad \mathrm{r} = 1,2,\dots,\mathrm{k}.$$

- In general, $\mu'_1,\mu'_2,\dots,\mu'_k$ will be function of the parameters $\theta_1,\theta_2,\dots,\theta_k$.

- Let $X_i, i = 1,2,\dots,n$ be a random sample of size n from a given population.

- The method of moments consists in solving the k-equations for $\theta_1,\theta_2,\dots,\theta_k$ in terms of $\mu'_1,\mu'_2,\dots,\mu'_k$ and then replacing these moments μ'_r, $\mathrm{r} = 1,2,\dots,\mathrm{k}$ by the sample moments

$$\hat{\theta}_i = h_i(\hat{\mu}'_1,\hat{\mu}'_2,\dots,\hat{\mu}'_k) = h_i(\alpha_1,\alpha_2,\dots,\alpha_k), \quad i = 1,2,\dots,k,$$

where α_i is the ith moment about origin in the sample.

Exercises:

- Let $X_1, X_2, \ldots, X_n$ be a random sample from Poisson distribution with parameter λ. Find the method of moment estimator.

- Let $X \sim Bin(n, p)$ where n is known and $0 \leq p \leq 1$. Find the method of moment estimator.

Maximum Likelihood Estimation (MLE)

Objective:

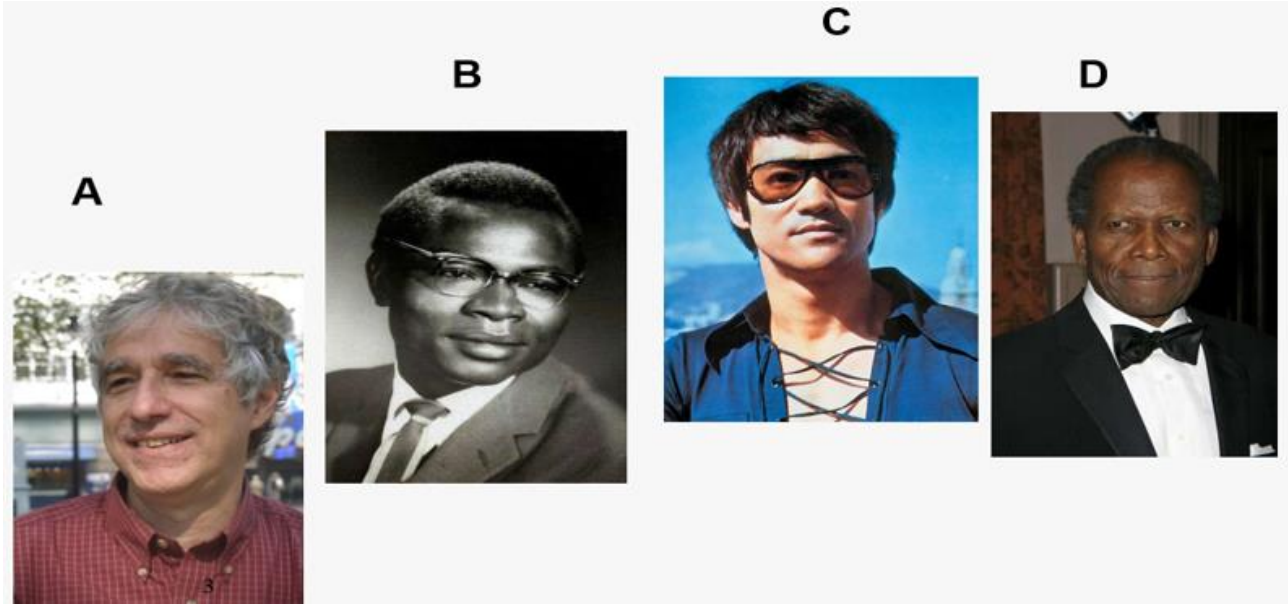
To introduce the idea of working out the most likely cause of an observed result by considering the likelihood of each of several possible causes and picking the cause with the highest likelihood. Start with a guessing game…

A guessing Game:

Here is a picture of a boy and his mother. Can you guess the daddy of this boy (next slide)?

Maximum Likelihood Estimation (MLE)

Who is the daddy of the child (out of the four men in the picture below)?



Maximum Likelihood Estimation (MLE)

- The child appears to be of mixed race parentage.
- His mother is white.
- Therefore, of the four possible daddies, daddy A is the least likely.
- Daddy C is the next least likely because of the child's appearance.
- This leaves B or D; but which one did you go for?
- If only we had more information, e.g., blood types or DNA or history of the two men, we could be more exact in our view of which of the two is more likely…
- NEVER be certain that which daddy is the real daddy! We simply accept the most likely choice. Every daddy is an MLE daddy! ☺

Did you guess…

- The child in the picture is President Barack Obama and his mother.
- Daddy D is the actor Sydney Poitier.
- C is the actor Bruce Lee.
- A is Professor Alan Agresti- Statistics Icon (writer of Foundation of Statistics for Data Scientists).
- B is Barack Obama, Sr.

An easier example…

- A box contains white and black balls mixed up in some unknown proportion. A ball is drawn from the box. Its colour is noted and then it is put back into the box. The experiment is repeated ten times.

- The results are as shown below:

- Which of the following populations do you think the balls were drawn from?

A: 10% White; 90% Black
B: 30% White; 70% Black
C: 50% White; 50% Black
D: 70% White; 30% Black
E: 90% White; 10% Black

Maximum Likelihood Method (MLE)

- In terms of color of balls, the result we have observed is WWWWBWWBBW

- Suppose the proportion of white balls from the box is some unknown quantity, say π, so that the proportion of black is $1-\pi$, then what is the probability of drawing this sequence of balls from the box?

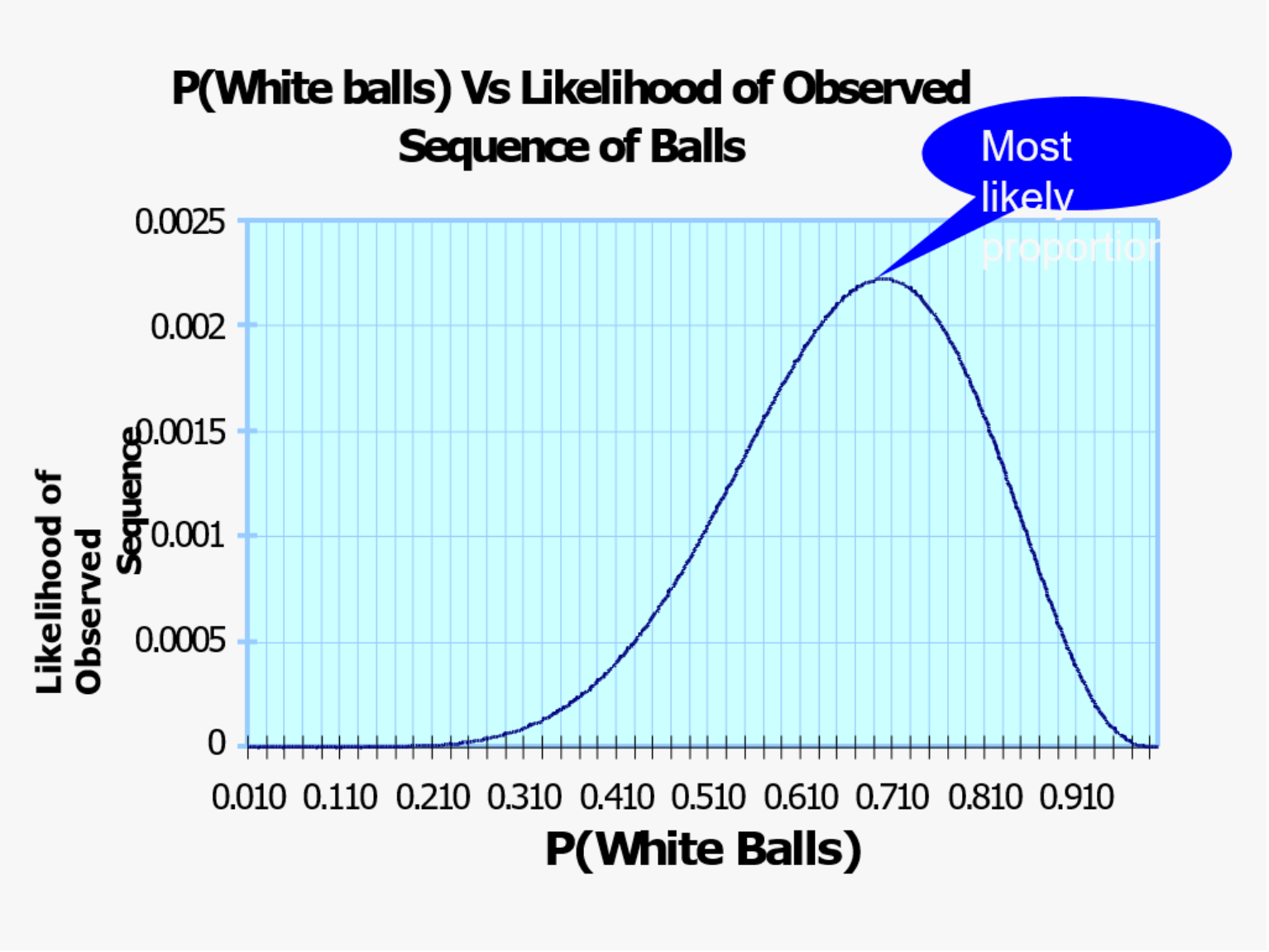
- The balls are selected independently from the box. Therefore, we can obtain the probability as
  $p = \pi*\pi*\pi*\pi*(1-\pi)*\pi*\pi*(1-\pi)*(1-\pi)*\pi = \pi 7(1-\pi)3$

- Now let's consider the probabilities we would find for the different values of π in the question above.

What is p, for different values of π?

- A: $\pi = 0.1$; p = 0.0000
- B: $\pi = 0.3$; p = 0.0001
- C: $\pi = 0.5$; p = 0.0010
- D: $\pi = 0.7$; p = 0.0022
- E: $\pi = 0.9$; p = 0.0005
- Therefore, it is most likely that a sequence of balls with colours WWWWBWWBBW will be drawn from a box where the proportion of black balls is 0.7.
- Probability (p) is about FUTURE events: When we looking back into the past, i.e., have already observed a result and we are trying to work out what could have caused it, we talk about likelihood (L).
- So, our Maximum Likelihood Estimate of $\pi = 0.7$

Visualization…



The concept of likelihood…

- If the probability of an event X dependent on model parameters *p* is written

$$P(X \mid p)$$

  - Then we would talk about the likelihood (the likelihood of the parameters given the data)

$$L(p \mid X)$$

- **The aim of Maximum Likelihood Estimator is to find the parameter value(s) that makes the observed data most likely…**
- In the case of *data analysis,* we have already observed all the data: once they have been observed they are fixed, there is no '**probabilistic**' part to them anymore (the word data comes from the Latin word meaning '**given**')
- We are much more interested in the likelihood of the model parameters that underlay the fixed data.

  - Probability

    Knowing parameters -> Prediction of outcome

  - Likelihood

    Observation of data -> Estimation of parameters

MLM: How it works?

- A result is observed, e.g., 7 white balls, 3 black
- Several hypotheses are proposed for what could have caused the observed results,
- e.g., $\pi = 0.1$, $\pi = 0.3$, etc.
- For each hypothesis, the likelihood that it could have caused the observed result is calculated.
- The hypothesis that has the maximum likelihood, e.g., $\pi = 0.7$ is taken as the best estimate for the observed result.

Maximum Likelihood Estimation (MLE)

- Likelihood Function:
  - Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a population with density function $f(x, \theta)$.
  - Then the likelihood function of the sample values $x_1, x_2, \dots, x_n$ usually denoted by $L = L(\theta)$ is their joint density function, given by:

$$L = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

  - For a given sample $x_1, x_2, \dots, x_n$, L becomes a function of the variable θ, the parameter.
- MLE Principle:
  - The principle of ML consists in finding an estimator for the unknown parameter $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, say which maximizes the likelihood function $L(\theta)$ for variations in parameter i.e. we wish to find $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ so that

$$L(\hat{\theta}) > L(\theta) \quad \forall \ \theta \in \Theta.$$

  - $\hat{\theta}$ is called Maximum Likelihood Estimator (MLE).

Exercises:

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli population with parameter $0 \leq p \leq 1$. Find MLE.

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $P(\lambda)$ population, where $\lambda > 0$. Find MLE of λ.

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $U(0, \theta)$ population, where $\theta > 0$. Find MLE.

Unbiasedness and Efficiency: MSE

- The mean squared error (MSE) of $T(X)$ is defined as

$$MSE_\theta\big(T(X)\big) = E[(T(X) - \theta)^2]$$

- MSE can be written as

$$MSE_\theta\big(T(X)\big) = Var_\theta\big(T(X)\big) + [Bias_\theta(T(X))]^2$$

- An estimator $T_1(X)$ is said to be MSE-better than another estimator $T_2(X)$ for estimating θ if

$$MSE_\theta\big(T_1(X)\big) < MSE_\theta\big(T_2(X)\big)$$

  where $\theta \in \Theta$ and Θ is the parameter space.

- For unbiased estimators, the MSE is equal to the variance of an estimator.
- More efficient unbiased estimator is used for comparing two unbiased estimators.

Unbiasedness and Efficiency: MEUE

- An unbiased estimator $T_1(X)$ is said to be more efficient than another unbiased estimator $T_2(X)$ for estimating θ if

$$Var_\theta\left(T_1(X)\right) \leq Var_\theta\left(T_2(X)\right), \forall \ \ \theta \in \Theta$$
$$\text{and } Var_\theta\left(T_1(X)\right) < Var_\theta\left(T_2(X)\right), \ \ for\ atleast\ one\ \theta \in \Theta$$

- For many problems, a best or most efficient estimate can be found. If such an estimator exists, it is said to be Uniformly Minimum Variance Unbiased Estimator (UMVUE).

- Uniformly means that it has the lowest variance among all other unbiased estimators for estimating the population parameter θ.

Exercise:

A random sample $(X_1, X_2, X_3, X_4, X_5)$ of size 5 is drawn from a normal population with unknown mean μ. Consider the following estimators to estimate μ :

i. $T_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$

ii. $T_2 = \frac{X_1 + X_2}{2} + X_3$

iii. $T_3 = \frac{2X_1 + X_2 + \lambda X_3}{3}$

where λ is such that T_3 is an unbiased estimator of μ. Find λ. Are T_1 and T_2 unbiased? State giving reasons, the estimator which is the best among T_1, T_2 *and* T_3.

Cramer Rao Inequality

If T is an unbiased estimator of $g(\theta)$, a function of parameter θ, then

$$Var(T) \geq \frac{(g'(\theta))^2}{E\left(\frac{\partial}{\partial\theta} Log\ L\right)^2},$$

where $I(\theta) = E\left(\frac{\partial}{\partial\theta} Log\ L\right)^2$ is the information on θ, supplied by the sample.

Remark 1: If T is an unbiased estimator of θ, i.e. $E(T) = \theta$, then

$$Var(T) \geq \frac{1}{E\left(\frac{\partial}{\partial\theta} Log\ L\right)^2}.$$

Remark 2: $I(\theta) = E\left(\frac{\partial}{\partial\theta} Log\ L\right)^2 = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2} Log\ L\right)$

Conditions for the equality sign in C-R Inequality

If the likelihood function L is expressible in the form

$$\frac{\partial}{\partial\theta} logL = \frac{T - g(\theta)}{\lambda(\theta)},$$

Then

i. T is an unbiased estimator of $g(\theta)$;

ii. Uniformly Minimum Variance Unbiased Estimator (UMVUE) T exists for $g(\theta)$;

iii. $\mathrm{Var}(T) = |g'(\theta)\lambda(\theta)|$

Exercises:

- Obtain UMVUE for μ in normal population $N(\mu, \sigma^2)$, where σ^2 is known.

- Show that $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, in random sample from

$$f(x,\theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & 0 < x < \infty \\ 0, & otherwise \end{cases}$$

Where $0 < \theta < \infty$ is an UMVUE of θ, and $Var(\bar{X}) = \frac{\theta^2}{n}$.

Sufficiency

- In the problem of statistical inference, the raw data collected from the field of enquiry is too numerous and hence too difficult to deal with and too costly to store.

- So, a statistician would like to condense the data by determining a function of the sample observation, i.e., by forming a statistic.

- Here the condensation should be done so that there is **'no loss of information'** regarding the population feature of interest.

- The statistic which exhaust all the relevant information about the labelling parameter, that contained in the sample are called "**sufficient statistic**".

- Clearly, sufficiency is an essential criteria for studying an inferential problem.

Sufficiency

- An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter.

- Let $X_1, X_2, \dots, X_n$ be a random sample from a pdf (or pmf) $f(x, \theta)$. A statistic T is said to be sufficient for θ if the conditional distribution of $X_1, X_2, \dots, X_n$ given $T = t$ is independent of θ.

Neyman-Fisher Factorization Theorem (NFFT)

Let $X_1, X_2, \dots, X_n$ be a random sample from a pdf (or pmf) $f(x, \theta)$. A statistic $T = T(x_1, x_2, \dots, x_n)$ is said to be sufficient for θ iff the joint density function of $X_1, X_2, \dots, X_n$ can be factorized as

$$f(x_1, x_2, \dots, x_n; \theta) = g(t, \theta).h(x_1, x_2, \dots, x_n),$$

where $h(x_1, x_2, \dots, x_n)$ is nonnegative and does not involve θ; and $g(t, \theta)$ is a nonnegative function of θ which depends on $x_1, x_2, \dots, x_n$ only through t, which is a particular value of T.

Exercises:

- Let $X_1, X_2, \dots, X_n$ be a random sample from Bernoulli distribution with parameter p. Then Prove that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic.

- Let $X_1, X_2, \dots, X_n$ be a random sample from Poisson distribution with parameter λ. Then Prove that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic.

- Let $X_1, X_2, \dots, X_n$ be a random sample from $N(\mu, \sigma^2)$ population. Find sufficient estimators for μ and σ^2.

Interval Estimation

- A point estimate on its own does not take into account the accuracy of the estimate.
- The deviation between the point estimate and the true parameter (i.e. $|\bar{x} - \mu|$) can be considerable, especially when the sample size is small.
- To incorporate the information about the accuracy of an estimate in the estimated value, a **confidence interval** can be constructed.
- It is a **random interval** with **lower** and **upper bounds,** $\boldsymbol{I_l(X)}$ and $\boldsymbol{I_u(X)}$, such that the unknown parameter θ is covered by a prespecified probability of at least $1 - \alpha$:

$$\boldsymbol{P_\theta\left(I_l(X) \leq \theta \leq I_u(X)\right) \geq 1 - \alpha.}$$

- The probability $1 - \alpha$ is called the **confidence level.**
- $\boldsymbol{I_l(X)}$ is called the **lower confidence limit** and $\boldsymbol{I_u(X)}$ is called the **upper confidence limit.**
- Note that the **bounds** are **random** and the **parameter** is a **fixed value**, i.e. the true parameter is covered by the interval with probability $\boldsymbol{1 - \alpha}$.

Interval Estimation

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

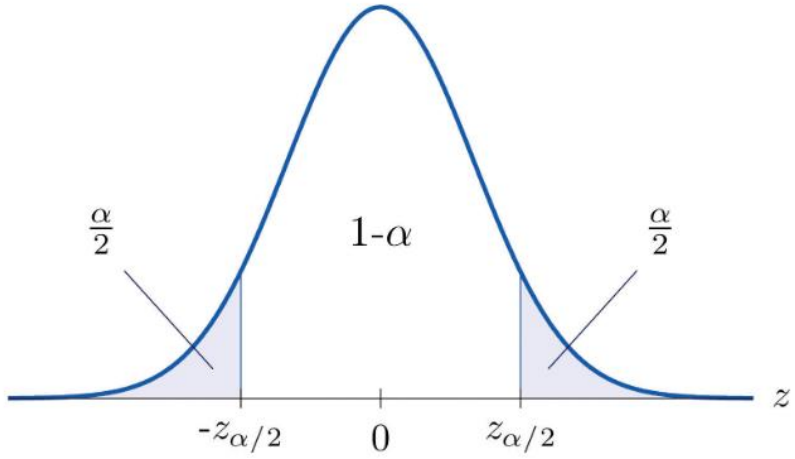
We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

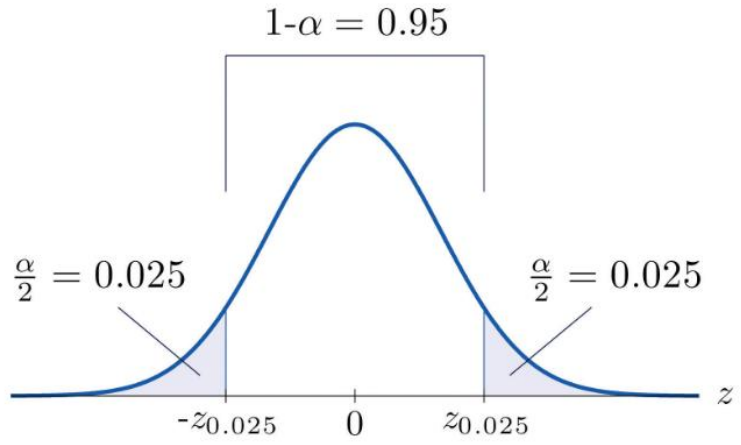
Interval Estimation

- It is standard practice to identify the level of confidence in terms of the area α in the two tails of the distribution of $\bar{X}$ when the middle part specified by the level of confidence is taken out.

- This is shown in Figure 1, drawn for the general situation, and in Figure 2, drawn for 95% confidence.



For 100 (1 − α)% confidence the area in each tail is α / 2.



For 95% confidence the area in each tail is α / 2 = 0.025.

Procedure Confidence Interval Measurement

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \Rightarrow \bar{X} - \mu = z \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow \bar{X} - z \cdot \frac{\sigma}{\sqrt{n}} = \mu$$

This implies that

$$P(Z > z_0) = P\left(\bar{X} - z \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Similarly,

$$P(Z < z_0) = P\left(\bar{X} + z \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Thus,

$$P\left(\bar{X} - z \cdot \frac{\sigma}{\sqrt{n}}\right) < \mu < P\left(\bar{X} + z \cdot \frac{\sigma}{\sqrt{n}}\right)$$ with probability 1-α

Therefore, the interval estimate of μ is customarily written as

$$\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Example: Confidence Interval Measurement

Suppose, a hypothesis testing for a population mean μ = 8.0 is as below.

$\bar{X}$ = 7.89, n = 16, = 0.2, σ = 0.2 and α = 0.05

For this testing, we have

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \Longrightarrow 1.96$$

Thus,

$$z_{\alpha/2} = 1.96$$

Hence,

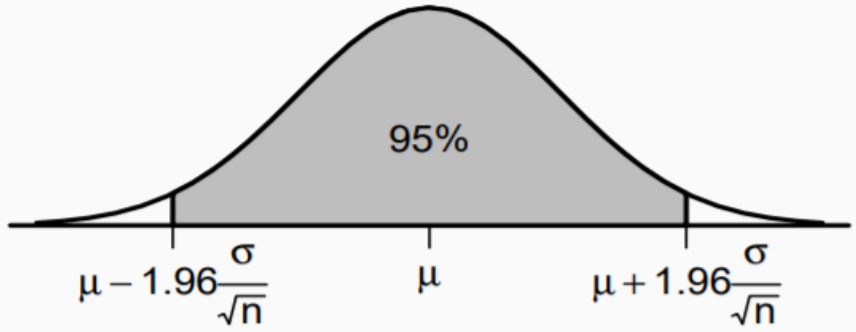
Confidence interval is $7.89 \pm 1.96(0.2)/\sqrt{16}$

- This is the interval estimate with 95% confidence (i.e., accuracy)
  - We are 95% confident that the true mean is between 6.91 to 8.87
- Here, the term E = $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ is called maximum error (also called error margin)
- Alternatively,

CI estimate is $\bar{X} \pm E$

Example:

Recall that CLT says, for large n, $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. For a normal curve, 95% of its area is within 1.96 SDs from the center. That means, **for 95% of the time, $\overline{X}$ will be within $1.96\frac{\sigma}{\sqrt{n}}$ from μ.**



Alternatively, we can also say, **for 95% of the time, μ will be within $1.96\frac{\sigma}{\sqrt{n}}$ from $\overline{X}$.**

Hence, we call the interval

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}} = \left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

a **95% confidence interval for μ.**

Case: σ is unknown

When the population SD σ is unknown, we replace it with our best guess - the sample SDs. So, an approximate 95% confidence interval for μ is

$$\bar{X} \pm 1.96 \frac{s}{\sqrt{n}}$$

However, this replacement is hazardous because

- s is a poor estimate of σ if the sample size n is small
- s is very sensitive to outliers

So, we require n ≥ 30 and sample shouldn't have any outlier nor be too skewed implies need to check histogram of the data.

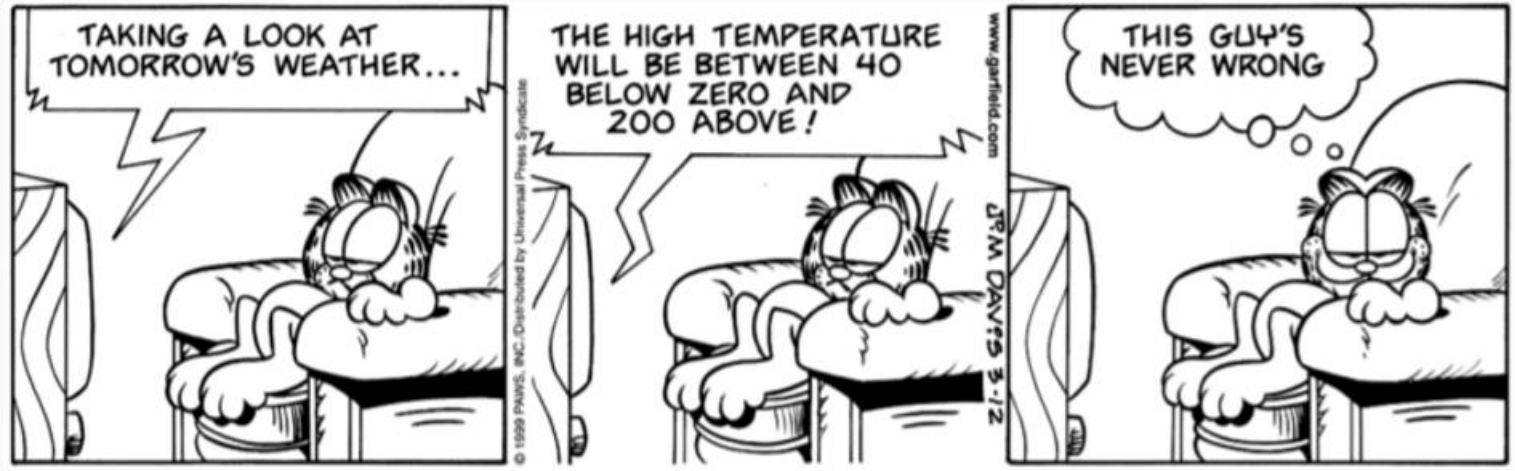
What does “95% confidence” mean?

What is the thing that has a 95% chance to happen?

- It is the procedure to construct the 95% interval.
- About 95% of the intervals constructed following the procedure (taking SRS and then calculating $\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$) will cover the true population mean μ.
- After taking the sample and an interval is constructed, the constructed interval either covers μ or it doesn’t.
- Just like lottery, before you pick the numbers and buy a lottery ticket, you have some chance to win the prize. After you get the ticket, you either win or lose.

What about the width of an interval?

- If we want to be more certain that we capture the population parameter, i.e., increase our confidence level, should we use a wider interval or smaller interval?
- A wider interval.
- Can you see any drawbacks to using a wider interval?



- If the interval is too wide, it may not be very informative.

Confidence Interval for the Mean of Normal Distribution

Confidence Interval for μ When $\sigma^2 = \sigma_0^2$ is Known.

- Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a $N(\mu, \sigma^2)$ distribution where σ_0^2 is assumed to be known.
- We use the point estimate $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate μ and construct a confidence interval around the mean μ.
- Using CLT, it follows that $\bar{X} \sim N(\mu, \sigma_0^2/n)$. Therefore, $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} \sim N(0,1)$, and it follows that

$$P_\mu\left(\left|\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}\right| \leq z_{\alpha/2}\right) = 1-\alpha, \qquad (2)$$

$z_{\alpha/2}$ denote the $(\alpha/2)$ quantile of the $N(0,1)$.

- From (2) we find the confidence interval as follows:

$$P_\mu\left[\bar{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right] = 1-\alpha. \qquad (3)$$

- This interval is known as $100(1-\alpha)\%$ confidence interval for μ.

Example:

A random sample of 120 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university.

Solution:

For confidence level 90%, $\alpha = 1 - 0.90 = 0.10$, so $z_{\alpha/2} = z_{0.05}$. From

we read directly that $z_{0.05} = 1.645$. Since n = 120, $\overline{x} = 2.71$, and s = 0.51,

$$\overline{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 2.71 \pm 1.645\left(\frac{0.51}{\sqrt{120}}\right) = 2.71 \pm 0.0766$$

One may be 90% confident that the true average GPA of all students at the university is contained in the interval $(2.71 - 0.08, 2.71 + 0.08) = (2.63, 2.79)$.

Exercise:

Suppose a random sample of size $n = 20$ of the weight of 10-year-old children in a particular city is drawn. Let us assume that the children's weight in the population follows $N(\mu, 6^2)$ distribution. The sample provides the following values of weights (in Kg):

40.2 32.8 38.2 43.5 47.6 36.6 38.4 45.5 44.4 40.3

34.6 55.6 50.9 38.9 37.8 46.8 43.6 39.5 49.9 34.2

Find the upper and lower limit of a 95% confidence interval.

Confidence Interval for the Mean of Normal Distribution

Confidence Interval for μ When σ^2 is Unknown.

- Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a $N(\mu, \sigma^2)$ distribution where σ^2 is assumed to be unknown and is being estimated by the sample variance S_X^2.
- Note 1: $\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}$
- Note 2: $\frac{\sqrt{n}(\bar{X}-\mu)}{S_X} \sim t_{n-1}$.
- We can determine the confidence interval for μ as

$$P_\mu\left[\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S_X}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S_X}{\sqrt{n}}\right] = 1-\alpha.$$

- This interval is known as $100(1-\alpha)\%$ confidence interval for μ.

Exercise:

Suppose a random sample of size $n = 20$ of the weight of 10-year-old children in a particular city is drawn. Let us assume that the children's weight in the population follows $N(\mu, \sigma^2)$ distribution. The sample provides the following values of weights (in Kg):

40.2	32.8	38.2	43.5	47.6	36.6	38.4	45.5	44.4	40.3
34.6	55.6	50.9	38.9	37.8	46.8	43.6	39.5	49.9	34.2

Find the upper and lower limit of a 95% and 99% confidence interval, respectively.

Estimating for Variance

- If s^2 is the variance of a random sample of size n from a normal population, a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are χ^2 −value with (n-1) degrees of freedom, leaving areas of $\alpha/2$ and $1-\alpha/2$, respectively to the right.

- An approximate $100(1-\alpha)\%$ confidence interval for σ is obtained by taking the square root of each endpoint of the interval for σ^2.

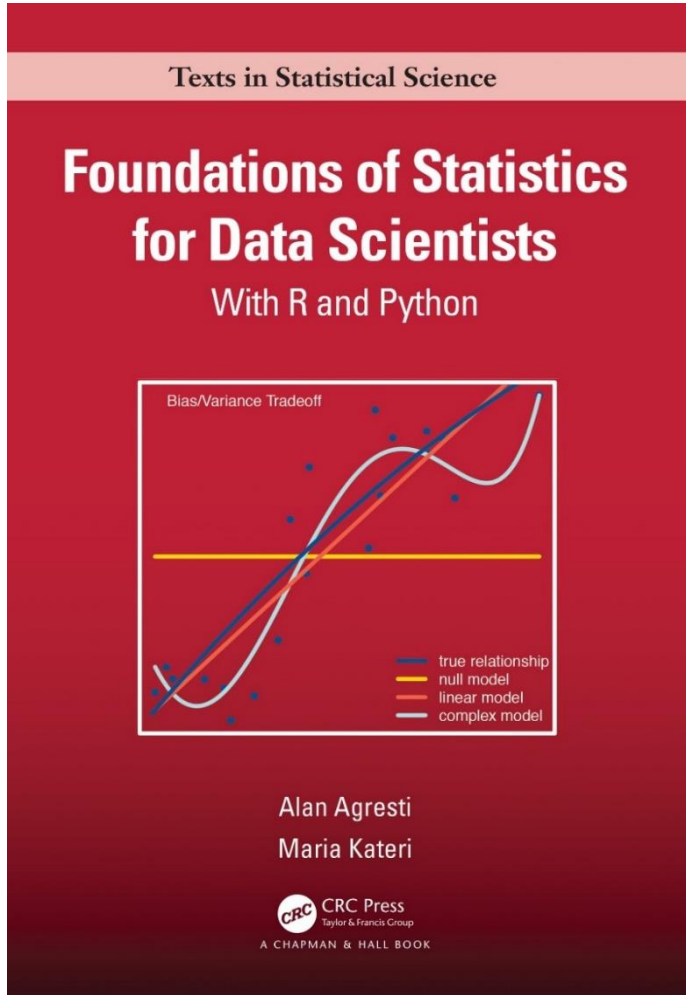
Exercise:

The following are the weights, in decagrams, of 10 packages of grass seed distributed by a certain company:

$$46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, 46.0.$$

Find a 95% confidence interval for the variance of the weights of all such packages of grass seed distributed by this company, assuming a normal population.

Reference:


Texts in Statistical Science
Foundations of Statistics for Data Scientists
With R and Python
Bias/Variance Tradeoff
true relationship
null model
linear model
complex model
Alan Agresti
Maria Kateri
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK