Probability and Statistics

MAT 271E

Ahmet H. Kayran

Section 2

(STATISTICS)

April 2019

STATISTICS

1. Introduction

- Statistics deal with data. Therefore, Statistics is the science of collecting, organizing, analyzing and interpreting data. In other words, the goal of this course is to make inferences based on data.

Inference:

A conclusion reached on the basis of evidence and reasoning.
Deduction - Conclusion - Reasoning - Guess Speculation

- We can divide this process into three phases:
  - Collecting data: The design of an experiment is crucial to making sure the collected data is useful.
  - Describing data: Raw data often takes the form of massive list, array or database of labels and numbers.
  - An Analysing data: We want to draw inferences about the world. This is a statistical model for the random

---

* J. Orloff and J. Bloom, MIT, Class Notes, Spring 2014, "Introduction to Statistics".

process.

- To make sense of data, we can calculate summary statistics like the mean and range.
- We can also visualize the data using graphical devices like histograms.

Example 1:

In 1999, in Great Britain, Sally Clark was convicted by murdering her two children after each child died weeks after birth (The first in 1996, the second in 1998).

- Her conviction was largely based on a faulty use of statistics to rule out sudden infant dead syndrome.

Example 2:

We may model the result of two-candidate election by a Bernoulli (p) distribution, and we can use poll data to draw inferences about the value p.

First of all, we ask the following question: What is a statistics?

What is a Statistics?

Definition:

A statistics is anything that can be computed from the collected data.

What is a statistical question?

A statistical question is a question where you expect to get variety of answers, and you are interested in the distribution and tendency of these answers.

Example 1:

"What is the weight of a mouse?"

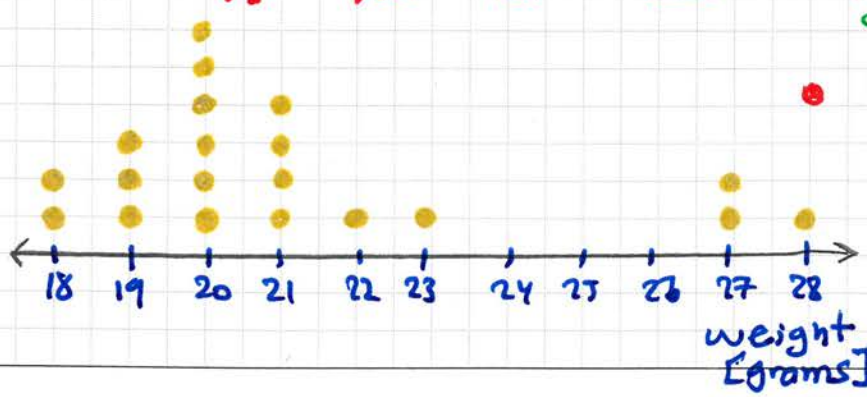
(a) Is this a statistical question? Explain.

Yes, because you would expect the weights of mice to vary.

Weights of mice [grams]

20	19	21	20
18	20	27	21
28	23	20	19
20	21	18	27
19	22	21	20

(b) Display the data in a dot plot. Identify any cluster, peaks or gabs.



- There is a cluster around 20.
- Peak at 20.
- Gap between 23 and 27.

(c) Use the distribution to answer "What is the weight of a mouse?

Most mice weight is about 20 grams.

Example 2:

- Consider the data of 1000 rolls of a die.

All of the following are statistics:

- The average of 1000 rolls.
- The number of times a 6 was rolled.
- The sum of the squares the rolls minus the number of even rolls.

- On the other hand, the probability of rolling a 6 is <u>not</u> a statistics whether or not the die is truly fair.
- However, this probability is a property of the die (and the way we roll it) which we can estimate using the data.
  - Such estimate is given by the statistics; "Proportion of the roll that were 6".

Example 3:

Suppose we treat a group of cancer patients with a new procedure and collect data on how long they survive post-treatment.

- From the data we can compute the average survival time of patients in the group.

- We might employ this statistics as an estimate of the average survival for future cancer patients following the procedure.
- However, the actual survival is <u>not</u> a statistic.

- In this course, we will use two types of statistics:

1. Point Statistics:

A single value computed from the data, such as the sample mean $\bar{x}_n$ or the sample standard deviation.

2. Interval Statistics:

An interval $[a, b]$ compiled from the data. This is really just a pair of point statistics and will often be presented in the form $\bar{x} \pm s$.

REVIEW OF BAYES' THEOREM

- Bayes' theorem allows us to "invert" conditional probabilities.
- If H and D are events, then Bayes' theorem says

$$Pr\{H|D\} = \frac{Pr\{D|H\}\,Pr\{H\}}{Pr\{D\}}.$$

- We start with hypothesis and collect data to test the hypothesis.
- Let H represent the event "our hypothesis is true" and let D be the collected data.
- In these words Bayes' theorem says

$$Pr\{\text{hypothesis is true} \mid \text{data}\} = \frac{Pr\{\text{data} \mid \text{hypothesis is true}\}\,Pr\{\text{hypothesis is true}\}}{Pr\{\text{data}\}}.$$

↑
The left hand term is the probability of our hypothesis is true given data we collected.

Remarks:

- If we know the exact values of all terms on the right, we can compute the probability on the left exactly.
- Unfortunately, in practice we rearly know the exact values of all terms on the right.

2. PARAMETER ESTIMATION

- Suppose we have data consisting of values $x_1, x_2, \ldots, x_n$ drawn from an exponential distribution. We want to know the parameter of this exponential distribution.

  - The exponential distribution, $X \sim \exp(\lambda)$, is a one-parameter distribution. Indeed, each value of λ defines a different distribution with pdf

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \in [0, \infty) \\ 0, & \text{otherwise.} \end{cases}$$

  - Similarly, a binomial distribution, $X \sim \text{Binom}(n, p)$ is determined by two parameters n and p.

  - A normal distribution, $N(\mu, \sigma^2)$ is determined by the two parameters μ and σ^2.

MAXIMUM LIKELIHOOD ESTIMATES (MLE)

- There are many methods for estimating unknown parameters from the data. First, we will study the MLE.

- MLE is a point estimate because it gives a single value for the unknown parameter.
- Actually this estimate answers the following question:

Question:

For which parameter value does the observed data have the biggest probability?

Example:

A coin is filipped 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability p of heads on a single toss.

For the binomial distribution

$$\Pr\{55 \text{ heads}\} = \binom{100}{55} p^{55} (1-p)^{45}.$$

Problem Formulation and Notation:

The probability of 55 heads depends on the value of p, so let's include p in by using the notation of conditional probability.

$$\Pr\{55 \text{ heads} \mid p\} = \binom{100}{55} p^{55} (1-p)^{45}.$$

$\Pr\{55 \text{ heads} \mid p\}$ should be read as:

"The probability of 55 heads given p" or more precisely as "the probability of 55 heads given that the probability of heads on a single toss is p".

STANDARD TERMS:

- **Experiment:** Flip the coin 100 times and count the number of heads.
- **Data:** The data is the result of the experiment. In this case, it is a "55 heads".
- **Parameter(s) of interest:** We are interested in the unknown parameter p.
- **Likelihood or Likelihood Function:** This is

  $\Pr\{\text{data} \mid p\}$.

  - This is a function of both the data and the parameter p.
  - In this case the likelihood is

$$\Pr\{55 \text{ heads} \mid p\} = \binom{100}{55} p^{55} (1-p)^{45}.$$

Remarks:

- The likelihood $Pr\{\text{data} \mid p\}$ changes as the parameter p changes.
- The likelihood $Pr\{\text{data} \mid p\}$ should not be written as $Pr\{p \mid \text{data}\}$. We know that from Bayes' theorem, they are not equal!

Definition:

- Given data the maximum likelihood estimate (MLE) for the parameter p is the value of p that maximizes the likelihood $Pr\{\text{data} \mid p\}$.
- This means, the MLE is the value of p for which the data is most likely.

Answer:

$$Pr\{55 \text{ heads} \mid p\} = \binom{100}{55} p^{55} (1-p)^{45},$$

we take the derivative of the likelihood function and setting it to zero.

$$\frac{d}{dt} Pr\{55 \text{ heads} \mid p\} = \binom{100}{55} \left[55 p^{54} (1-p)^{45} - 45 p^{55} (1-p)^{44} \right] = 0$$

solving this equation for p we get,

$$55 p^{54} (1-p)^{45} = 45 p^{55} (1-p)^{44}$$

$$55(1-p) = 45p$$

$$55 = 100p \quad \Rightarrow \text{The MLE is } \hat{p} = 0.55.$$

Remarks:

1.) The MLE for p turned out to be exactly fraction of heads we saw in our data.

2.) The MLE is computed from the data is a statistic.

3.) The critical point is indeed a maximum. You can check this with the second derivative test.

LOG LIKELIHOOD:

It is often easier to work with the natural log of the likelihood function. Since $\ln(x)$ is increasing function, the maxima of the likelihood and log likelihood coincide.

Example:

Let's redo the previous example using log likelihood. The log lakelihood is

$$\ln\left(\Pr\{55 \text{ heads} | p\}\right) = \ln\left[\binom{100}{55}\right] + 55\ln(p) + 45\ln(1-p)$$

We maksimize log likelihood,

$$\frac{d}{dp}\left(\text{log likelihood}\right) = \frac{d}{dp}\left[\ln\left[\binom{100}{55}\right] + 55\ln(p) + 45\ln(1-p)\right]$$

$$= \frac{55}{p} - \frac{45}{1-p} = 0$$

$$\Rightarrow 55(1-p) = 45p \quad \text{and} \quad \hat{p} = 0.45.$$

MAXIMUM LIKELIHOOD FOR CONTNUOUS DISTRIBUTIONS

- We use the probability density function to define the likelihood.

Example:

- Suppose that the lifetime of light bulbs is modeled by an exponential distribution with unknown parameter λ.
- We test 5 bulbs and they have lifetimes of 2, 3, 1, 3 and 4 years, respectively.
- What is the MLE for λ?
- Let X_i be the life time of i^{th} bulb and let x_i be the value X_i takes. Then X_i has the following pdf

$$f_X(x_i) = \lambda e^{-\lambda x_i}, \quad \text{for } i = 1, \dots, 5.$$

- We also assume the lifetimes of the bulbs are independent, so the joint pdf is the product of individual densities;

$$f(x_1, x_2, x_3, x_4, x_5 \mid \lambda) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})(\lambda e^{-\lambda x_3})(\lambda e^{-\lambda x_4})(\lambda e^{-\lambda x_5})$$

$$= \lambda^5 e^{-\lambda (x_1 + x_2 + x_3 + x_4 + x_5)}.$$

- This conditional density is the likelihood function. The data is fixed and λ is variable. Our data had values,

$x_1 = 2$, $x_2 = 3$, $x_3 = 1$, $x_4 = 3$, $x_5 = 4$.

Therefore, the likelihood function with this data is

$$f(2,3,1,3,4 \mid \lambda) = \lambda^5 e^{-13\lambda}$$

and the log likelihood function

$$\ln\left[f(2,3,1,3,4 \mid \lambda)\right] = 5\ln(\lambda) - 13\lambda$$

and the MLE can be calculated as

$$\frac{d}{d\lambda}\left(\text{log likelihood}\right) = \frac{5}{\lambda} - 13 = 0$$

$$\Rightarrow \hat{\lambda} = \frac{5}{13}.$$

- Remark:
  - The MLE for λ turned out to be the reciprocal of the sample space mean $\bar{x}$, where

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{n} = \frac{1}{\hat{\lambda}}.$$

Example: Uniform Distributions.

- Suppose our data $x_1, x_2, \ldots, x_n$ are independently drawn from a uniform distribution $U(a, b)$. Let's find MLE for a and b.

Answer:

This example is different from the previous ones. We will not use calculus to find the MLE. The density for $U(a,b)$ is

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for } x \in [a,b] \\ 0, & \text{elsewhere.} \end{cases}$$

Therefore, our likelihood function is

$$f(x_1, x_2, \ldots, x_n \mid a, b) = \begin{cases} \left(\frac{1}{b-a}\right)^n, & \text{if all } x_i \text{ are in the interval } [a,b] \\ 0, & \text{otherwise.} \end{cases}$$

This is maximized by making $(b-a)$ as small as possible. The only restriction is that the interval $[a,b]$ must include all the data. Thus, the MLE for the pair (a,b) is

$\hat{a} = \min(x_1, \ldots, x_n)$ and $\hat{b} = \max(x_1, \ldots, x_n)$.

BAYESIAN UPDATING WITH DISCRETE PRIORS

We know that Bayes' theorem allows us to "invert" conditional probabilities. If H and D are events, then

$$Pr\{H|D\} = \frac{Pr\{D|H\}Pr\{H\}}{Pr\{D\}}.$$

Let us use a coin tossing problem to introduce terminology and a tabular format for Bayes' theorem. This will provide a simple, uncluttered example that shows our main points.

Example:

There are three types of coins which have different probabilities of landing head when tossed.

- Type A coins are fair, with probability 0.5 of heads.
- Type B coins are bent and have probability 0.6 of heads.
- Type C coins are bend and have probability 0.9 of heads.

Suppose we have a drawer contain 5 coins: 2 of typ A, 2 of type B, and 1 of type C. We reach into the drawer

and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it type A? Type B? Type C?

Answer:

Let A, B and C be the event that chosen coin was type A, type B, and type C, respectively. Let D be the event that the toss is heads. The problem asks to find

$$Pr\{A|D\}, \quad Pr\{B|D\}, \quad \text{and} \quad Pr\{C|D\}.$$

Some Terminology:

- **Experiment:** Pick a coin from the drawer at random, flip it, and record the result.
- **Data:** The result of the experiment. In this case, the event D = "heads".
- **Hypotheses:** We are testing three hypotheses: The coin type A, B or C.
- **Prior Probability:** The probability of each hypotheses prior to tossing the coin (collecting data). Since the drawer has 2 coins of type A, 2 coins of type B and 1 of type C we have

$$Pr\{A\} = 0.4, \quad Pr\{B\} = 0.4, \quad Pr\{C\} = 0.2.$$

Likelihood:

- This is the same likelihood we used for the MLE.
- The likelihood function is

$$Pr\{D|H\}.$$

The probability of the data assuming the hypothesis is true. For example, $Pr\{D|A\}$ = probability of heads if the coin is type A. In our case, the likelihoods are

$$Pr\{D|A\} = 0.5\,,\quad Pr\{D|B\} = 0.6\,,\quad Pr\{D|C\} = 0.9.$$

Warning !

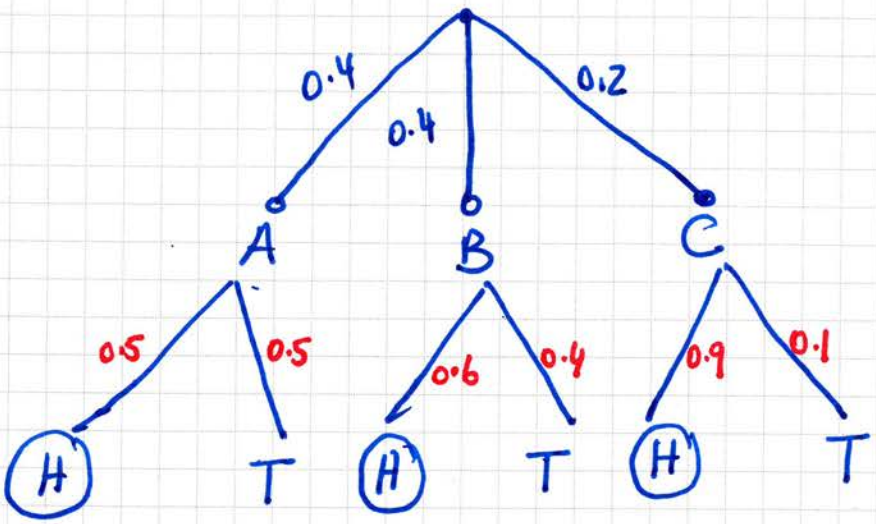
- In colloquial language likelihood and probability are synonymes. This leads to the likelihood function often confused with the probability of a hypothesis.
- Therefore, we will try to use "likelihood" carefully in order to minimize any confusion.

POSTERIOR PROBABILITY:

The probability of each hypothesis given data from tossing the coin: $Pr\{A|D\}$, $Pr\{B|D\}$, $Pr\{C|D\}$ These posterior probabilities are what the problem asks as to find.

PROBABILITY TREE:

We can organize the probabilities into a tree:



- From Bayes' theorem, we can write

$$Pr\{A|D\} = \frac{Pr\{D|A\}Pr\{A\}}{Pr\{D\}}.$$

- The denominator $Pr\{D\}$ is computed using the law of total probability.

$$Pr\{D\} = Pr\{D|A\}Pr\{A\} + Pr\{D|B\}Pr\{B\} + Pr\{D|C\}Pr\{C\}$$
$$= (0.5)(0.4) + (0.6)(0.4) + (0.9)(0.2) = 0.62.$$

- Then, we can compute each of the three posterior probabilities:

$$Pr\{A|D\} = \frac{Pr\{D|A\}Pr\{A\}}{Pr\{D\}} = \frac{(0.5)(0.4)}{0.62} = \frac{0.20}{0.62} = 0.3226$$

$$Pr\{B|D\} = \frac{Pr\{D|B\}\,Pr\{B\}}{Pr\{D\}} = \frac{(0.6)(0.4)}{0.62} = \frac{0.24}{0.62} = 0.3871$$

$$Pr\{C|D\} = \frac{Pr\{D|C\}\,Pr\{C\}}{Pr\{D\}} = \frac{(0.9)(0.2)}{0.62} = \frac{0.18}{0.62} = 0.2903.$$

Remarks:

(1) The total probability $Pr\{D\}$ is the same in each denominators and that is the sum of three numerators.

(2) We can organize all in a Bayesian update table:

Hypothesis	Prior	likelihood	Bayes Numerators	Posterior
H	$Pr\{H\}$	$Pr\{D\|H\}$	$Pr\{D\|H\}\,Pr\{H\}$	$Pr\{H\|D\}$
A	0.4	0.5	0.20	0.3226
B	0.4	0.6	0.24	0.3871
C	0.2	0.9	0.18	0.2903
Total	1		0.62	1

(3) The Bayes numerator is the product of prior and the likelihood.

(4) Each posterior probability is obtained by dividing the Bayes numerator by $Pr\{D\}=0.625$. According to the total probability, $Pr\{D\}$ is the sum of entries in the Bayes numerator column.

(5) The process of going from the prior probability $Pr\{H\}$ to the posterior $Pr\{H|D\}$ is called BAYESIAN UPDATING.

Indeed, Bayesian updating uses the data to alter our understanding the probability of each possible hypotheses.

OBSERVATIONS:

1.) There are two types of probabilities:

(a) The standard probability of data, for example, the probability of heads.

(b) The probability of the hypothesis, for example, the probability of the chosen coin is type A, B. or C.

- This type has prior (before the data) and posterior (after the data) values.

2.) After the data, the posterior probabilities for each hypothesis are in the last column. From this column,

we see that coin B is now the most probable. Its probability has decreased from a prior probability of 0.4 to a posterior probability of 0.39. On the other hand, the probability of type C has increased from 0.2 to 0.29.

3.) We can rescale the Bayes numerator column to compute the posterior probability column. As a result of this, the sum of posterior column is 1.

4.) If we are only interested in the most likely hypothesis, the Bayes numerator works as well as the normalized posterior.

5.) Since the likelihood function is not probability function, the likelihood column dos not sum to 1.

6.) When calculating the posterior, a large prior may be deflated by a small likelihood, and a small prior may be deflated by a large likelihood.

7.) The maximum likelihood estimate (MLE). for our example earlier is hypothesis C, with a likelihood $Pr\{D|C\}=0.9$. Hence the MLE is not the entire story; since B has the greatest posterior probability

COMMENTS ABOUT BAYES' THEOREM

$$Pr\{H|D\} = \frac{Pr\{D|H\}\, Pr\{H\}}{Pr\{D\}}$$

namely,

$$Pr\{\text{hypothesis}|\text{data}\} = \frac{Pr\{\text{data}|\text{hypothesis}\}\, Pr\{\text{hypothesis}\}}{Pr\{\text{data}\}}.$$

- With the fixed data, the denominator $Pr\{D\}$ just serves to normalize the total posterior probability to 1.
- Therefore, we can express Bayers' theorem as a statement about proportionality of two functions of H, namely, the last two columns of the table are proportional.

$$Pr\{\text{hypothesis}\} \sim Pr\{\text{data}|\text{hypothesis}\}\, Pr\{\text{hypothesis}\}$$

⇒ This leads Bayesian updating;

> Posterior ∝ likelihood × prior.

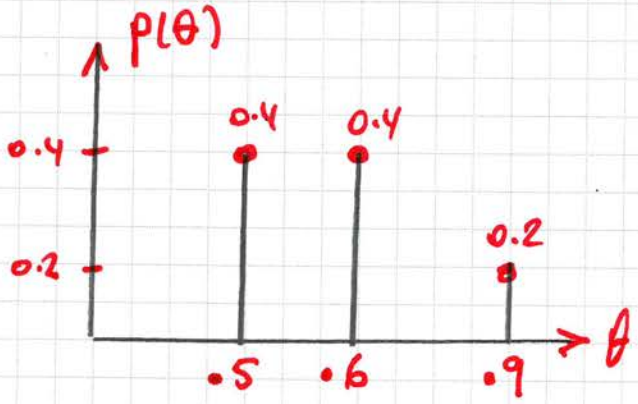
PRIOR and POSTERIOR Probability Mass Functions (pmf)

We can use random variables and probability mass functions as we have studied earlier in this course. Hence we assign values to events (head is 1 and tail is 0). We will do everything in the context of Bayesian updating.

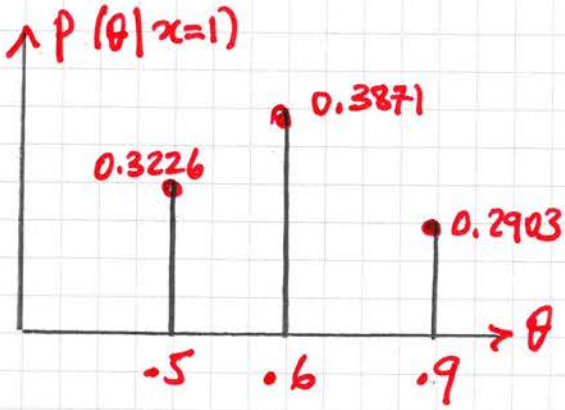
Standard Notations:

- θ is the value of the hypothesis.
- $p(\theta)$ is the prior pmf of the hypothesis.
- $p(\theta|D)$ is the posterior pmf of the hypothesis given data.
- $p(D|\theta)$ is the likelihood function. This is not a pmf.
  - In the given example, we can represent the three hypotheses A, B and C by $\theta = 0.5$, 0.6, and 0.9, respectively.
  - For the data, we will let $x=1$ mean heads and $x=0$ mean tails.
  - In the following table, the prior and posterior probabilities define the prior and posterior pmfs.

hypothesis	θ	Prior pmf $p(\theta)$	Posterior pmf $p(\theta \mid x=1)$
A	0.5	$Pr\{A\} = p(0.5) = 0.4$	$Pr\{A \mid D\} = p(0.5 \mid x=1) = 0.3226$
B	0.6	$Pr\{B\} = p(0.6) = 0.4$	$Pr\{B \mid D\} = p(0.6 \mid x=1) = 0.3871$
C	0.9	$Pr\{C\} = p(0.9) = 0.2$	$Pr\{C \mid D\} = p(0.9 \mid x=1) = 0.2903$



The prior pmf



The posterior pmf

Remark!

If the data is different than the likelihood column in the Bayesian update would be different In the coin example, there are two possibilities for the data: the toss is heads or the toss is tails. So the full likelihood table has two likelihood columns.

Likelihood Table

hypothesis	likelihood $p(x \mid \theta)$	
θ	$p(x=0 \mid \theta)$	$p(x=1 \mid \theta)$
0.5	0.5	0.5
0.6	0.4	0.6
0.9	0.1	0.9

Example:

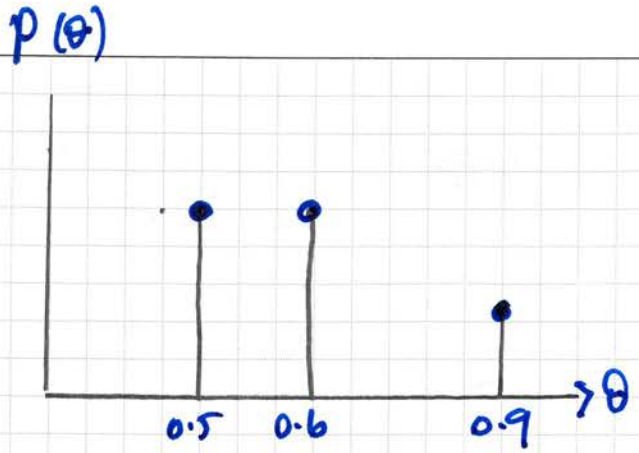
Using the notation $p(\theta)$, $p(x=0|\theta)$ and $p(\theta|x=0)$ we can redo the example assuming the flip was tails;

Answer:

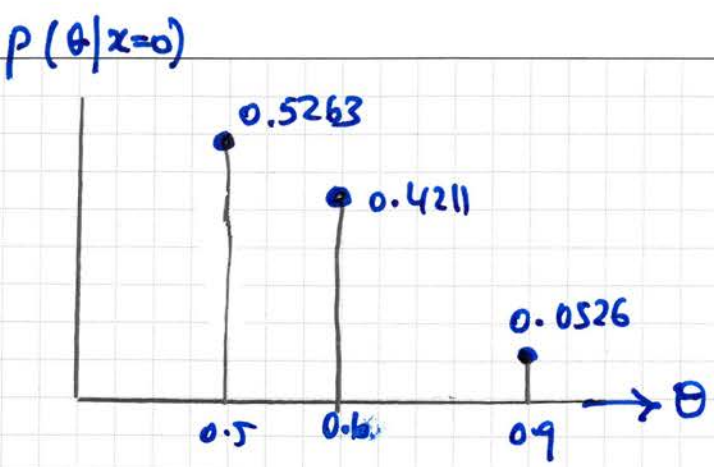
The likelihood column in the Bayesian update table is now $x=0$. Therefore, we take the $p(x=0|\theta)$ column from the likelihood table given earlier.

hypothesis	Prior	likelihood	Bayes Numerator	Posterior
θ	$p(\theta)$	$p(x=0\|\theta)$	$p(x=0\|\theta)p(\theta)$	$p(\theta\|x=0)$
0.5	0.4	0.5	0.20	0.5263
0.6	0.4	0.4	0.16	0.4211
0.9	0.2	0.1	0.02	0.0526
TOTAL	1		0.38	1

- For this new data [the flip was tails], the probability of type A increased from 0.4 to 0.5263, while the probability of type C has decreased from 0.2 to only 0.0526.
- Here we can see the prior and posterior probabilities.



Prior pmf
$p(\theta)$



Posterior pmf
$p(\theta|x=0)$

UPDATING AGAIN AND AGAIN

- Throughout life, we are continually updating our beliefs with each new experience of the world.
- In Bayesian inference, after updating the prior to the posterior, we can take more data and update again.
  - For the second update, the posterior from the first data becomes the prior for the second.

Example:

- Suppose we have picked a coin as in our example earlier. We flip it once again get heads. Then you flip the same coin and get heads again.

What is the probability that the coin was type A? Type B? Type C?

Answer:

hypothesis θ	prior $p(\theta)$	likelihood 1 $p(x_1=1 \mid \theta)$	Bayes numerator 1 $p(x_1=1 \mid \theta)\, p(\theta)$	Posterior 1 $p(\theta \mid x_1=1)$
0.5	0.4	0.5	0.20	0.3226
0.6	0.4	0.6	0.24	0.3870
0.9	0.2	0.9	0.18	0.2904
TOTAL	1		0.62	1

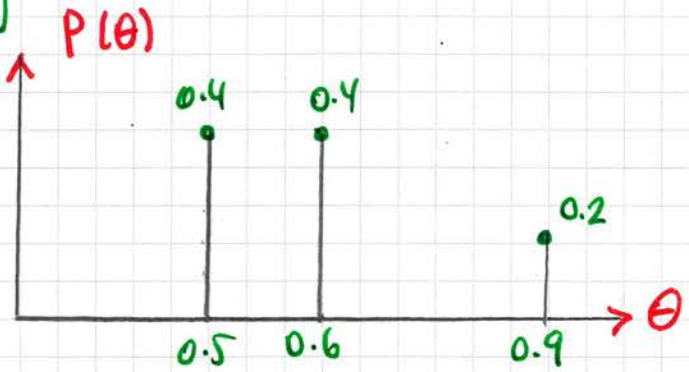
likelihood 2 $p(x_2=1 \mid \theta)$	Bayes numerator 2. $p(x_2=1 \mid \theta)\, p(x_1=1 \mid \theta)\, p(\theta)$	Posterior 2 $p(\theta \mid x_1=1, x_2=1)$
0.5	0.100	0.2463
0.6	0.144	0.3547
0.9	0.162	+ 0.3990
	0.402	1

Observations:

- Since we are intrested in the final posterior, there is no need to normalize until the last step.

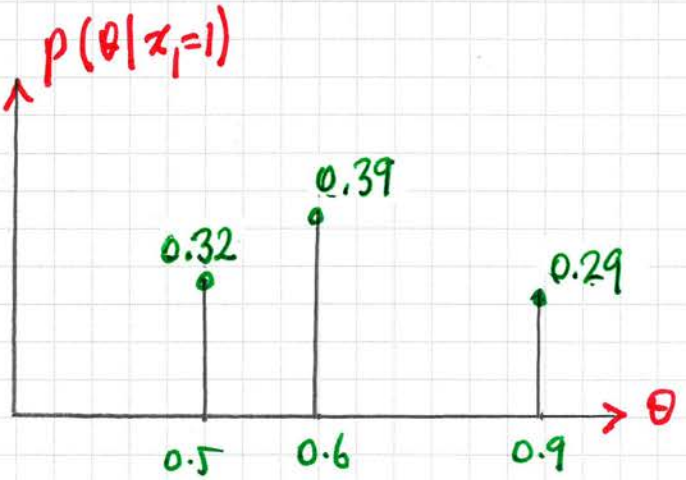
- The second Bayes numerator is computed by multiplying the first Bayes numerator and the second likelihood.
- After two heads, the type C hypothesis has finally taken the lead.

Prior $p(\theta)$



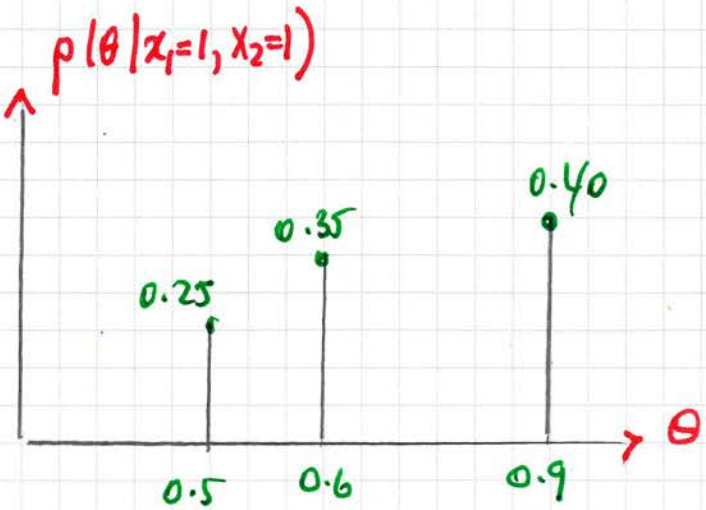
first Posterior

$p(\theta \mid x_1=1)$



Second Posterior

$p(\theta \mid x_1=1, x_2=1)$



THE BASE RATE FALLACY

Example:

- A screening test for a disease is usually positive when testing a person with disease and usually negative when testing someone without disease.
- Let us assume the true positive rate is 99%, the false positive rate is 2%.
- Suppose the prevalence (yaygınlığı) of the disease in general in population is 0.5 %.
- If a random person tests positive, what is the probability that they have the disease?

Notacion:

- Let $\mathcal{H}_+$ be the hypothesis (event) that the person has the disease and let $\mathcal{H}_-$ be the hypothesis (event) do not have disease.
- Let T_+ and T_- represent data of a positive and negative screening test, respectively.
- We are asked to compute:

$$\Pr\{\mathcal{H}_+ \mid T_+\} = ?$$

- We are given,

$$\Pr\{T_+ \mid \mathcal{H}_+\} = 0.99$$

$$\Pr\{T_+ \mid \mathcal{H}_-\} = 0.02$$

$$\Pr\{\mathcal{H}_+\} = 0.005$$

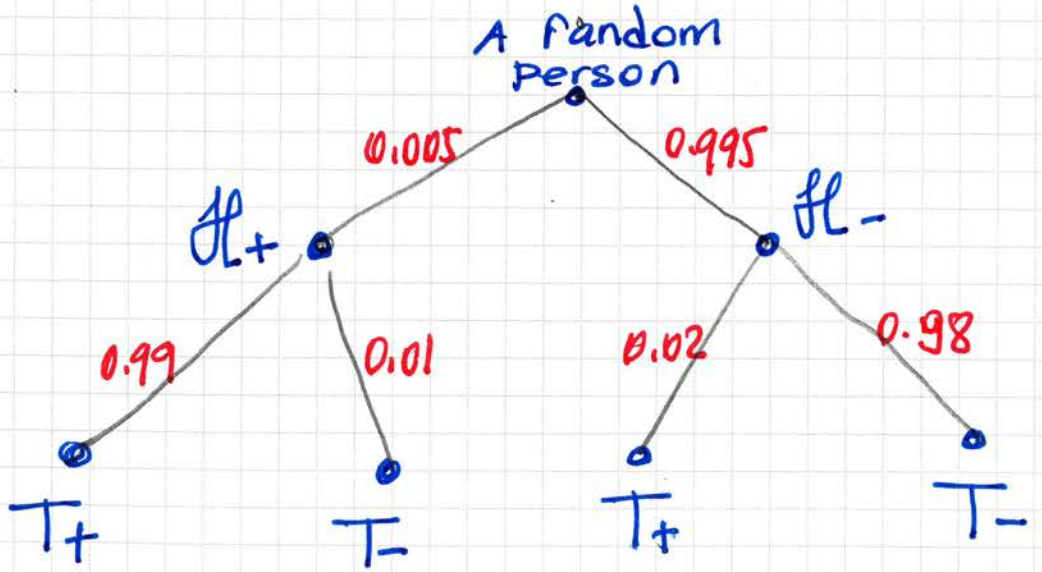
from these we can compute false negative and true negative values,

$$\Pr\{T_- \mid \mathcal{H}_+\} = 0.01$$

$$\Pr\{T_- \mid \mathcal{H}_-\} = 0.98 \quad \text{and}$$

$$\Pr\{\mathcal{H}_-\} = 0.995 .$$

- Now we can display all of these probabilities in a tree.



- We can obtain the same result using a Bayesian update table:

Hypothesis	Prior	likelihood	Bayes Numerator	Posterior
$\mathcal{H}$	$Pr\{\mathcal{H}\}$	$Pr\{T_+ \mid \mathcal{H}\}$	$Pr\{T_+ \mid \mathcal{H}\} Pr\{\mathcal{H}\}$	$Pr\{\mathcal{H} \mid T^+\}$
$\mathcal{H}^+$	0.005	0.99	0.00495	0.19920
$\mathcal{H}^-$	0.995	0.02	0.01990	0.80080
Total	1	NO SUM	0.02485	1

- This table also shows that the posterior probability $Pr\{\mathcal{H}_+ \mid T_+\}$ that a person with a positive test has the disease is about 20%. This is far less than the sensitivity of the test (99%)! However, it is much higher than the prevalence of the disease in the general population (0.5%).

SAMPLING THEORY *+

* Probability and Statistics, Murray R. Spiegel, McGraw-Hill, 1996.
+ Applied Statistics and Probability for Engineers, D.C. Montgomery, G.C. Runger

- The entire group of data is called the population. It is difficult or impossible to examine this population.
- Hence we can examine only a small part of this population, which is called sample.
- This is realized with a goal of inferring certain facts about the population from results found in the sample. The process of obtaining Samples from the population is called sampling.
- The word population is often used to denote the observations or measurements rather than individuals or objects. The Population size (number) is usually denoted by N (big letter).
- The population size, N, can be finite or infinite. Similarly, the number of sample size is shown by n (small letter). The sample size, n, is generally finite.

Example 1: We want to determine the fairness of a particular coin by tossing it repeatedly. The population consists of all possible tosses of the coin. A sample

could be obtained by observing, say, the first 100 tosses of the coin and noting the percentages heads and tails.

- In this example, the population size N is infinite and the sample size is $n=100$.

Example 2:

We may wish to draw conclusion about the heights (weights) of 20 000 university students by examining only 200 students (a sample) selected from this population.

- In this example, we can see that the population size is finite, $N = 20\,000$ and the sample size is $n=200$.

SAMPLING WITH OR WITHOUT REPLACEMENT

- If we draw an object from a box, we have the choice of replacing or not replacing the object into the box before a second drawing.

Sampling with Replacement:

- If we replace the particular object can come up again and again. Since each member of a population may be chosen more than once, this type of sampling is called **Sampling with Replacement**.

- A finite population which is sampled with replacement can theoretically be considered infinite.

Sampling Without Replacement

- In this case, we draw an object from the box and we don't replace it into the box before a second drawing. Therefore, a particular object can come up only once. In other words, each member of the population can not be chosen more than once. It is called sampling without replacement.

Random Samples

- In order to draw relaible conclusions, the sample must be properly chosen to represent the population sufficiently well. How can we choose a sample?
- For finite populations, we can make sure that each member of the populatton has the same chance of being in the sample. This is called a random sample.
- Random sampling can be accomplished for relatively small populations by drawing lots or, equivalently, by using a table of random numbers specially constructed for such purposes.

POPULATION PARAMETERS

- If we know the probability function, $f(x)$ of the associated random variable X, a population is considered to be known. For example, in Example 1, if X is a random variable whose values are the heights (or weights) of the 20 000 students then X has a probability distribution $f(x)$.
- If, for example, X is normally distributed than we say that, the population is normally distributed or we say that we have a normal population.
- Similarly, if X is binomially distributed, we say that the population is binomially distributed or that we have a binomial distribution.
- The probability function $f(x)$ has certain quantities which appear in its matematical model. For example, μ and σ appear in the case of normal distribution or p in the case of the binomial distribution. All such quantities are called the population parameters.
  - If the population is known, so that we know $f(x)$, then the papulation parameters are also known.

Remark:

- If the probability distribution function $f(x)$ is not known precisely, we may have some idea of the general behavior of $f(x)$. For example, we may have some reason to suppose that a particular population normally distributed. In this case, we would not know one or both of the values μ and σ^2 and so we might wish to estimate them.

SAMPLE STATISTIC

- A statatistic is any function of the observations in a random sample.
- Random samples are taken from the population and we use them to estimate the population parameters.
- A sample size n would be described by the values $x_1, x_2, \ldots, x_n$ of the random variables $X_1, X_2, \ldots, X_n$ would be independent, identically distributed (iid) random variables having probability distribution $f(x)$. Their joint distribution would be as follows:

$$Pr\{X_1=x_1, X_2=x_2, \ldots, X_n=x_n\} = f(x_1)f(x_2)\cdots f(x_n).$$

- Any quantity obtained from the sample for the purpose of estimating a population parameter is called a sample statistic, or briefly statistic.
- Mathematically, a sample statistic for a sample n can be defined as a function of random variables $X_1, X_2, \ldots, X_n$, namely, $g(X_1, \ldots, X_2)$.
  - It is interesting to note that the function $g(X_1, \ldots, X_n)$ is another random variable. Its values can be represented by $g(x_1, \ldots, x_n)$.
- Normally, we use Greek letters, such as, μ and σ, etc, for values of population parameters. However, we use Roman letters m, s, etc for values corresponding sample statistics.
- The probability distribution of a sample statistic is often called the sampling distribution of the statistic.

THE SAMPLE MEAN

- Let $X_1, X_2, \ldots, X_n$ denote the random variables for a sample size n. The mean of the sample or sample mean is a random variable defined by

$$\hat{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

Sample Range

- If $x_1, x_2, \ldots, x_n$ denote values obtained in a particular sample of size n, the mean for that sample is

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n).$$

SAMPLING DISTRIBUTION OF MEANS

- The probability distribution of a statistic is called a sampling distribution.
- Suppose that we draw a sample size n from some given population with a probability distribution $f(x)$. The probability distribution of the sample statistic $\bar{x}$ is called the sampling distribution of means. Then we can state the following theorems:

Theorem 1:

The expected value of the sample mean is the population mean,

$$E[\hat{X}] = \mu_{\bar{X}} = \mu \,, \qquad (1)$$

where μ is the mean of the population.

Theorem 2:

If the population is infinite or if sampling is with replacement, then the variance of the sampling distribution of means, namely, $\sigma_{\bar{X}}^2$, is given by

$$E\left[(\bar{X}-\mu)^2\right] = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \qquad (2)$$

where σ^2 is the variance of the population.

Theorem 3:

If the population size is N, if sampling is without replacement, and if the sample size is $n \leq N$, then (2) is replaced by,

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \qquad (3)$$

while $\mu_{\bar{X}}$ is still given by (1).

- Note that (3) is reduced to (2) as $N \to \infty$.

Theorem 4:

If the population from which samples are taken is normally distributed with mean μ and variance σ^2 then the sample mean is normally distributed with mean m and variance σ^2/n, namely,

If $X \sim N(\mu, \sigma^2)$,

Then $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$.

Theorem 5:

Suppose that the population from which samples are taken is <u>not</u> a normal distribution. However, its probability distribution has a mean μ and variance σ^2. Then the standard random variable associated with $\bar{X}$ is given by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}. \qquad (4)$$

Z is asymptotically normal, namely,

$$\lim_{n\to\infty} \Pr\{Z \le z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\, du. \qquad (5)$$

Remarks:

- It is assumed here that the population is infinite or that sampling is with replacement.
- If the population is finite, the above is correct if we replace $\sigma/\sqrt{n}$ in (4) by $\sigma_{\bar{X}}^2$ as given by (3).
- Theorem 5 is a consequence of the central limit theorem.

Example 1:

Let's consider the eight observations collected. These are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$ and $x_8 = 13.1$. The sample mean is

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{1}{8}(12.6 + 12.9 + \cdots + 13.6) = \frac{104}{8} = 13.0 .$$

The sample variance has the value,

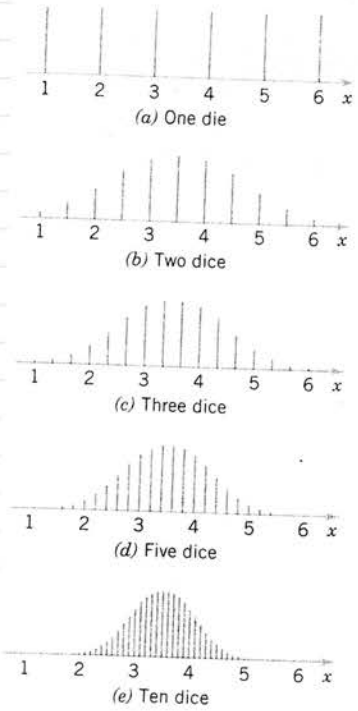
$$s^2 = \frac{1}{n}\left[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right]$$

$$= \frac{1}{8}\left[(12.6 - 13.0)^2 + \cdots + (13.1 - 13.0)^2\right] = \frac{160}{8} = 0.2.$$

The unbised estimate of the sample variance is given by

$$\hat{S}^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right]$$

$$= \frac{1}{8-1}\left[(12.6 - 13.0)^2 + \cdots + (13.1 - 13)^2\right] = \frac{16.0}{7} = 0.2286 .$$



- The normal approximation for $\bar{X}$ depends on the sample size n.
- Figure (a) shows the distribution obtained for throws of single, six-sided true die. We can see that the probabilities are equal (1/6) for all values obtained, 1, 2, 3, 4, 5, or 6.
- Figure (b) shows the distribution of the average score obtained when tossing two dice.
- Figure (c), (d), and (e) show the distributions of average scores obtained when tossing three, five and ten dice.

OBSERVATIONS:

- When the population one die, the distribution is relatively far from the normal.
- However, the central limit theorem works well for small samples ($n = 4, 5$) in most cases.
- In practice, if $n \geq 30$, the normal approximation will be satisfactory regardless of the shape of the population.
- The dice throw distributions are discrete, however, while the normal is continuous.

Example 2:

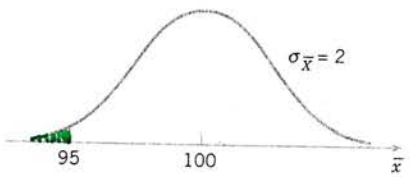
A company manufactures resistors that have a mean resistance of 100 ohms and the standard deviation of 10 ohms. Find the probability that a random sample

of $n=25$ resistors will have an average resistance less than 95 ohms.

The sampling distribution of $\bar{X}$ is normal, with mean $\mu_{\bar{X}} = 100$ ohms and a standard deviation of

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2.$$

Therefore, the desired probability corresponds to the shaded are in the following figure,



- Standardizing the point $\bar{X} = 95$, we find

$$z = \frac{95-100}{2} = -2.5$$

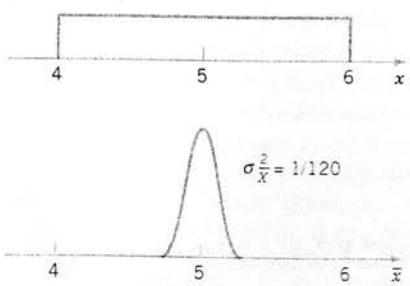
and from the table in the textbook,

$$Pr\{\bar{X} < 95\} = Pr\{Z \leq -2.5\} = 0.0062.$$

Example 3: Suppose the random variable X has a continuous uniform distribution,

$$f(x) = \begin{cases} \frac{1}{2}, & 4 \leq x \leq 6 \\ 0, & \text{otherwise.} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n=40$.



- The mean and variance of X are $\mu = 5$ and $\sigma^2 = \frac{(6-4)^2}{12} = \frac{1}{3}$.
- The central limit theorem indicates that the distribution of $\bar{X}$ is approximately normal with mean $\mu_{\bar{X}} = 5$ and $\sigma^2_{\bar{X}} = \frac{\sigma^2}{40} = \frac{1/3}{40} = \frac{1}{120}$. The distribution of X and $\bar{X}$ are shown in the figure.

THE SAMPLE VARIANCE

If $X_1, X_2, \ldots, X_n$ denote random variables for a sample of n, then the variance of the sample or the sample variance is defined as

$$S^2 = \frac{1}{n}\left[(X_1-\bar{X})^2+(X_2-\bar{X})^2+\cdots+(X_n-\bar{X})^2\right] \quad (6)$$

- In theorem 5, we found that $E[\bar{X}]=\mu$ and it would be nice if we could also have $E[S^2]=\sigma^2$.
- However, it turns out, that

$$E[S^2] = \mu_{S^2} = \frac{n-1}{n}\sigma^2$$

for large values of n (say $n \geq 30$), it is very nearly σ^2.

- The desired unbiased estimator is defined by

$$\hat{S}^2 = \frac{n}{n-1}S^2$$

$$= \frac{1}{n-1}\left[(X_1-\bar{X})^2+\cdots+(X_n-\bar{X})^2\right] \quad (7)$$

so that

$$E[\hat{S}^2] = \frac{n}{n-1}E[S^2]$$

$$= \frac{n}{n-1}\,\frac{n-1}{n}\sigma^2 = \sigma^2.$$

Remarks:

- Because of this, some statisticians choose to define the sample variance by $\hat{S}^2$ rather than S^2 and they replace n by $n-1$ in the denominator of (6).
- Another way of thinking about this is to consider the sample variance S^2 as being based on $(n-1)$ degrees of freedom.
  - The terms "degrees of freedom" results from the fact that n deviations $x_1-\bar{x}, x_2-\bar{x}, \ldots, x_n-\bar{x}$ always sum zero. Another words, any $(n-1)$ of these quantities automatically determines the remaining ones. Therefore, only $n-1$ of the n deviations, $x_i-\bar{x}$, are freely determined.
- If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample range is

$$r = \max(x_i) - \min(x_i).$$

- Since the deviations $x_i-\bar{x}$ always sum to zero, we must use a measure of variability that changes the negative deviations to nonnegative quantities. Squaring the deviations is the approach used in sample variance.

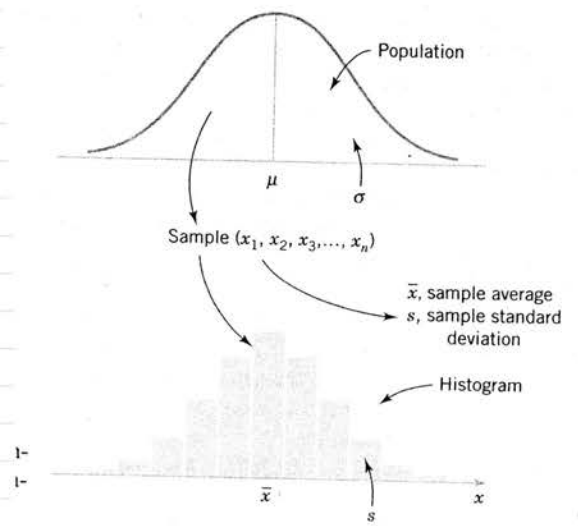
- The following table shows the calculation of terms for the sample variance and sample standard definition..

Table 6-1 Calculation of Terms for the Sample Variance and Sample Standard Deviation

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	−0.4	0.16
2	12.9	−0.1	0.01
3	13.4	0.4	0.16
4	12.3	−0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	−0.4	0.16
8	13.1	0.1	0.01
	104.0	0.0	1.60

Relationship Between a Population and a Sample:

- In most statistics problems, we work with a sample of observations selected from the population that we are interested in studying.



- This figure illustrates the relationship between the population and the sample.

ESTIMATION THEORY

(definitions)

Unbiased Estimates:

- A statistic is called an unbiased estimator of a population parameter if the mean or expectation of the statistic is equal to parameter.
- The corresponding value of the statistic is then called an unbiased estimate of the parameter.
- For example, the mean $\bar{X}$ and variance $\hat{S}^2$ are defined as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X)^2.$$

since $E[\bar{X}] = \mu$ and $E[\hat{S}^2] = \sigma^2$, $\bar{X}$ and $\hat{S}^2$ are unbiased estimators of the population mean μ and the variance σ^2.

Efficient Estimate:

- If the sampling distribution of two statistics have the same mean, the statistic with smaller variance is called a more efficient estimator of the mean.

- The corresponding value of the efficient statistic called an efficient estimate.

- For example, the sampling distribution of the mean and median both have the same mean, namely, the population mean. However, the variance of sampling distribution of means is smaller than that of the sampling distributions of medians. Thus the mean provides more efficient estimate than the median.

Definition: Median

The median of a set of data is that divides the data into two equal halves. When the number of observations is even, say $2n$, it is customary to define the median as the average of the nth and $(n+1)$st rank-ordered values. The median is also defined for a random variable. For example, in the case of continuous random variable X, the median M can be defined as

$$\int_{-\infty}^{M} f(x)dx = \int_{M}^{\infty} f(x)dx = \frac{1}{2}.$$

Point Estimates:

An estimate of a population parameter given by a single number is called a **point estimate** of the parameter.

Interval Estimates:

An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an **interval estimate** of the parameter.

- For example, if we say that a distance is 6.25 meters, we are giving a point estimate. On the other hand, if we say that the distance is 6.25 ± 0.05 meters, namely, the distance lies between 6.20 and 6.30 meters, we are giving an interval estimate.

Relaibility:

A statement of the error or precision of an estimate is often called its **relaibility**.

Confidence Interval Estimates of Population Parameters

- Let μ_S and σ_S be the mean and standard deviation of the sampling distribution of a statistic S.

- If the sampling distribution of S' is approximately normal ($n \geq 30$), we can expect to find S' lying in the following intervals:
- If $\mu_{S'} - \sigma_{S'} \leq S' \leq \mu_{S'} + \sigma_{S'}$, then we can be confident about 68.27%.
- If $\mu_{S'} - 2\sigma_{S'} \leq S' \leq \mu_{S'} + 2\sigma_{S'}$, then we can be confident about 95.45%.
- If $\mu_{S'} - 3\sigma_{S'} \leq S' \leq \mu_{S'} + 3\sigma_{S'}$, then we can be confident about 99.73%
- Therefore, we call these respective intervals the 68.27%, 95.45% and 99.73% confidence intervals for estimating the population parameter, μ, in the case of an unbiased S'.

Confidence Limits:

- The end numbers of these intervals, $S \pm \sigma_{S'}$, $S \pm 2\sigma_{S'}$, $S \pm 3\sigma_{S'}$ are then called the 68.27%, 95.45% and 99.73% confidence limits

- The percentage confidence is often called the Confidence level. The numbers 1.96, 2.58 etc., in the confidence are called Confidence coefficients or critical values.
- The following table shows values of z_c (critical values) corresponding levels used in practice.

Confidence Level	99.73%	99%	98%	96%	95%	90.%	80%	50%
z_c	3.0	2.58	2.33	2.05	1.96	1.645	1.28	0.6745

HYPOTHESIS TESTING

- In many engineering require that we decide whether to accept or reject a statement about some parameter. The statement is called a hypothesis, and the decision-making precedure about the hypothesis is called hypothesis testing.
- All these decision-making problems, tests, or experiments in the engineering would can be formulated as hypothesis testing problem.
- There is a very close connection between hypothesis testing and confidence intervals.