Department of Electrical Engineering
University of Arkansas

ELEG 3143 Probability & Stochastic Process

Ch. 5 Elements of Statistics

Dr. Jingxian Wu
wuj@uark.edu

OUTLINE

- Introduction: what is statistics?
- Sample mean and sample variance
- Confidence intervals
- Hypothesis testing

INTRODUCTION

- **Statistics**
  - A bridge between the probability theory and the real world
  - The science of gathering and analyzing data, and with the drawing of conclusions or inferences from the data.
    - Example
      - 1. collect data about the life span of all females in US
      - 2. analyze the data to find out:
        - » What is the average (mean) female life span in US?
        - » What is the variance of the female life span?
        - » What is the distribution of the female life span?
    - Example
      - 1. consider a communication system transmitting -1 and 1
      - 2. received signal is distorted by noise (e.g., Tx a 1, Rx a 0.2)
      - 3. Based on the noisy received information, find out what was transmitted (1 or -1)

INTRODUCTION

- **Classifications of statistics**
  - Sampling theory:
    - How to select samples from some collection of data that is too large to be examined completely.
  - Estimation theory:
    - Make some estimation or prediction based on the data that are available (e.g. estimate the average life span)
  - Hypothesis testing
    - Attempts to decide which of two or more hypotheses about the data are true (e.g. find out whether a 1 or -1 are transmitted in a communication system)

OUTLINE

- Introduction: what is statistics?
- **Sample mean and sample variance**
- Confidence interval
- Hypothesis testing

SAMPLE MEAN AND VARIANCE

- **Definition: population**
  - the collection of ALL objects or elements under study.
  - E.g. measure the female life span in US
    - member: the life span of one female in US (can be modeled as an RV: X)
    - Population: the life spans of ALL the females in US
  - Most of the time, it is extremely difficult and expensive, if not impossible, to get the data for the entire population.
  - We take a limited number of samples to represent the population
- **Definition: random sample (or, sample)**
  - A random sample is part of the population that has been selected at random.
    - E.g. Consider a population with N members. We can randomly pick $n << N$ members. The n members form a sample of the population.
    - All members of the population are equally likely being picked.
    - The picks are independent of each others.

SAMPLE MEAN AND VARIANCE

- **Random sample**
  - Consider a random sample with *n* members

$$\{X_1, X_2, \cdots, X_n\}$$

  - Each member is a random variable
  - The *n* random variables are mutually independent
  - The *n* random variables are identically distributed
  - With the random sample, we can infer (estimate) the properties of the population with *N* >> *n* members
    - E.g. mean, variance, distribution, etc.
  - Intuitively, the larger the value of *n*, the better the estimation.
    - We will prove this intuition.

SAMPLE MEAN AND VARIANCE

- **Sample mean**
  - The sample mean of $\{X_1, X_2, \cdots, X_n\}$ is

$$\hat{m}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - It is an estimation of the actual mean of the population $m_X = E(X)$
  - $\hat{m}_X$ is a random variable, because it is a function of *n* RVs $\{X_1, X_2, \cdots, X_n\}$

- **The first moment of the sample mean $\hat{m}_X$**
  - $E[\hat{m}_X] =$

- **Definition: unbiased estimator**
  - The mean of the estimation is the same as its true value
  - The sample mean is an unbiased estimation of the mean $E[\hat{m}_X] = m_X$

SAMPLE MEAN AND VARIANCE

- The 2nd central moment (variance) of the sample mean $\hat{m}_X$
  - $Var(\hat{m}_X) = E\left[(\hat{m}_X - m_X)^2\right] = E[\hat{m}_X^2] - E^2[\hat{m}_X]$

$$\boxed{Var(\hat{m}_X) = \frac{\sigma_X^2}{n}}$$

  - Recall: variance measures the deviation from the mean
    - The smaller the variance, the less the randomness
    - If variance is 0, then the RV is a constant
  - The larger the sample size *n*, the more accurate the estimation.

SAMPLE MEAN AND VARIANCE

- **Example**
  - A population of 10 resistors is to be tested. The true standard deviation is 5 Ohm, and the true mean is 100 Ohm.
    - How large must be the sample size, if we want to obtain a sample mean, whose standard deviation is 2% of the true population mean?
    - If the sample size is 8, what is the standard deviation of the sample mean?

SAMPLE MEAN AND VARIANCE

- **Example**
  - The sample of a random time function follows a pdf given as follows:

$$f_X(x) = \frac{1}{\sqrt{10\pi}} \exp\left(-\frac{(x-3)^2}{10}\right)$$

The function is sampled so as to obtain independent sample values. How many sample values are required to obtain a sample mean, whose standard deviation is 1% from the true mean?

SAMPLE MEAN AND VARIANCE

- **Sample variance**
  - The sample variance of $\{X_1, X_2, \cdots, X_n\}$ is defined as

$$\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{m}_X)^2$$

  - The $\frac{1}{n-1}$ is used to make $\tilde{s}^2$ an unbiased estimator .
  - $\tilde{s}^2$ is a random variable.
- **First moment of the sample variance** $\tilde{s}^2$

$$E[\tilde{S}^2] = \sigma_X^2$$

  - The sample variance is an unbiased estimation of σ_X^2

SAMPLE MEAN AND VARIANCE

- **Example**
  - Consider 5 random numbers: 0.3, 0.2, 0.8, 0.7, 0.9
    - 1. Find the sample mean and sample variance
    - 2. If the 5 random numbers are randomly picked from a population of random numbers that are uniformly distributed in [0, 1], find the variance of the sample mean.

SAMPLE MEAN AND VARIANCE

- **Example**
  - Write a Matlab program to generate 100 independent random numbers. The random number are samples of uniformly distributed random variable in (0, 10). (1) Write functions to find the sample mean and sample variance. (2) What is the variance of the sample mean?

```
%-----------------------------
% main.m
clear all;
random_sample = 10*rand (1, 100); % generate 100 random numbers
sample_mean = find_sample_mean(random_sample);
sample_var = find_sample_var(random_sample);

%-------------------------------
% find_sample_mean.m
function     output = find_sample_mean(input)
n_sample = length(input);
output = sum(input)/n_sample;
```

SAMPLE MEAN AND VARIANCE

- **Example (Cont’d)**

```
%-------------------------------
% find_sample_var.m
function      output = find_sample_var(input)

% find out how many members are in the sample
n_sample = length(input);

% calculate the sample mean
sample_mean = find_sample_mean(input);

% calculate the sample variance
output = sum( (input-sample_mean).^2 )/(n_sample-1);
```

OUTLINE

- Introduction: what is statistics?
- Sample mean and sample variance
- **Confidence interval**
- Hypothesis testing

CONFIDENCE INTERVAL

- **Central Limit Theorem**
  - Let $\{X_1, X_2, \cdots, X_n\}$ be a random sample of size *n*. They are independent and identically distributed with mean m_X and variance σ_X^2. When $n \to \infty$, the sample mean, $\hat{m}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$, converges in distribution to a Gaussian

    distribution with mean m_X and variance σ_X^2.
  - When *n* is large (*n* > 30), $\hat{m}_X$ follows approximately to a Gaussian distribution, regardless of the distribution of X_i

CONFIDENCE INTERVAL

- **Confidence interval**
  - Example: estimate the mean, m_X , of a population of data.
    - The sample mean, $\hat{m}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$, is an RV
    - thus it could be quite different from the true mean.
  - Specify an interval that is highly likely to contain the true value of m_X
    - E.g. It is 99% likely that the true value of m_X is in the interval $[\hat{m}_X - a, \hat{m}_X + a]$

$$\Pr\left\{\hat{m}_X - a \le m_X \le \hat{m}_X + a\right\} = 0.99$$

    - $[\hat{m}_X - a, \hat{m}_X + a]$ is called the 99% confidence interval of the mean

- **Confidence Interval**
  - Confidence interval v.s. sample mean
    - Sample mean attempts to use a single number, $\hat{m}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$, to represent the sample mean
      - The sample mean is an RV, thus it could be quite different from the true mean.
      - How much can we trust this single number estimation?
    - Confidence interval, instead, attempts to specify an interval that is highly likely to contain the true value of the estimation.
  - The $q \times 100\ \%$ confidence interval for the estimation of the parameter *m* is defined as $[\hat{m}_X - a, \hat{m}_X + a]$, such that

$$\Pr\left\{\hat{m}_X - a \le m_X \le \hat{m}_X + a\right\} = q$$

CONFIDENCE INTERVAL

- **Confidence Interval (Cont'd)**

$$\Pr\left\{\hat{m}_X - a \le m_X \le \hat{m}_X + a\right\} = \Pr\left\{m_X - a \le \hat{m}_X \le m_X - a\right\} = \int_{m_X - a}^{m_X + a} f_{\hat{m}_X}(x)\,dx$$

– Based on the central limit theorem, $\hat{m}_X$ is Gaussian distributed with mean m_X and variance $\frac{\sigma_X^2}{n}$

$$\int_{m_X - a}^{m_X + a} f_{\hat{m}_X}(x)\,dx = \int_{-\frac{a\sqrt{n}}{\sigma_X}}^{\frac{a\sqrt{n}}{\sigma_X}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,dz = 1 - 2Q\left(\frac{a\sqrt{n}}{\sigma_X}\right)$$

$$\Pr\left\{\hat{m}_X - a \le m_X \le \hat{m}_X + a\right\} = 1 - 2Q\left(\frac{a\sqrt{n}}{\sigma_X}\right)$$

$$\Pr\left\{m_X - a \le \hat{m}_X \le m_X + a\right\} = 1 - 2Q\left(\frac{a\sqrt{n}}{\sigma_X}\right)$$

CONFIDENCE INTERVAL

- **Example**
  - A very large population of resistor values has a true mean of 100 Ohm and a standard deviation of 5 Ohm. Find the 95% confidence interval (confidence limits) on the sample mean if the sample size is 100.

    Q(1.96) = 0.025

$$\Pr\left\{m_X - a \le \hat{m}_X \le m_X + a\right\} = 1 - 2Q\left(\frac{a\sqrt{n}}{\sigma_X}\right)$$

OUTLINE

- Introduction: what is statistics?
- Sample mean and sample variance
- Confidence interval
- **Hypothesis testing**

HYPOTHESIS TESTING

- **Hypothesis testing**
  - Testing an assertion about a population based on a random sample.
  - Example:
    - Hypothesis: a given coin is fair
    - Test: flip the coin 100 times, count the number of heads
      - If the coin is fair, we expect approximately 50 heads.
      - E.g. if the number of heads is in [47, 53], the hypothesis is true. The hypothesis is false otherwise.
      - The interval [47, 53] is chosen arbitrarily. How to systematically choose the interval?
  - Example:
    - Hypothesis: the light bulb from a certain manufacture can last 1000 hrs.
    - Test: take 50 light bulbs, measure their life, and find the sample mean
      - If the sample mean is greater than t hrs, the hypothesis is true.
      - How do we determine the value t? 900 hrs? 950 hrs? 999 hrs?

HYPOTHESIS TESTING

- **Hypothesis testing:**
  - Null hypothesis H_0 :
    - The hypothesis to be tested
  - Alternative hypothesis H_1 :
    - The complement (opposite) of the null hypothesis
  - Example:
    - Test the hypothesis that a given coin is fair.

H_0 :the coin is fair

H_1 :the coin is not fair

    - Test the hypothesis that a lightbulb can last 1,000 hours

H_0 : the lightbulb can last 1,000 hours

H_1 : the lightbulb cannot last 1,000 hours

    - Test the hypothesis that a certain patient does not have cancer

H_0 : the patient has no cancer (negative)

H_1 : the patient has cancer (positive)

HYPOTHESIS TESTING

- **Errors**
  - Type I error: Reject H_0 when H_0 is true
    - False positive, false alarm
  - Type II error: Accept H_0 when H_0 is false
    - False negative

HYPOTHESIS TESTING

- **Hypothesis testing: significance testing**
  - Test a hypothesis H_0 about a parameter *p* of a random variable *X*
    - Example: test whether a coin is fair
      - *X*: Bernoulli RV with parameter *p: P(X=1)=p, P(X=0)=1-p*
      - H_0 *: p = 0.5 (*the coin is fair*).*
  - Objective: accept or reject the hypothesis based on a random sample $\{X_1, X_2, \cdots, X_n\}$

HYPOTHESIS TESTING

- **Example**
  - A certain coin is claimed to be fair with a 95% confidence interval. To test the hypothesis, we flip the coin 100 times, and find the sample mean, $\hat{m}_X$. If $\hat{m}_X = 0.43$, can we accept the claim?

    Sol: Hypothesis: this is a fair coin → $m_X = p = 0.5$, $\sigma_x^2 = p - p^2 = 0.25$

    find out if $\hat{m}_X = 0.43$ is in the 95% confidence interval

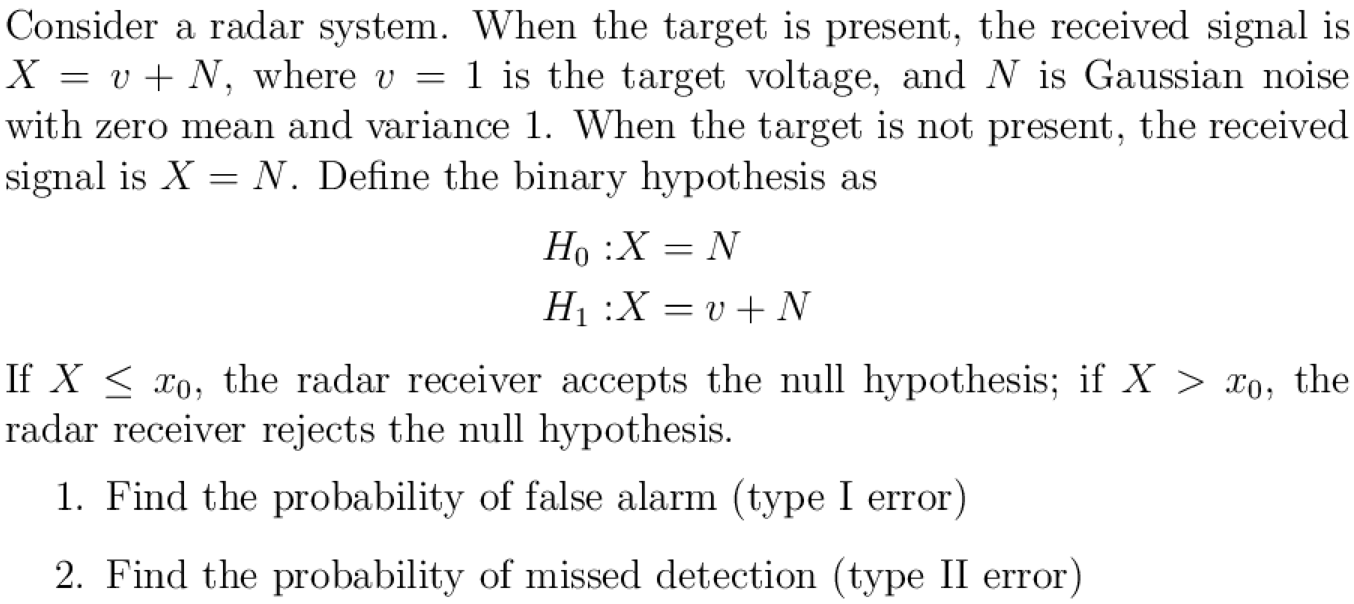
$$\Pr\{m_X - a \le \hat{m}_X \le m_X + a\} = 1 - 2Q\left(\frac{a\sqrt{n}}{\sigma_X}\right)$$

HYPOTHESIS TESTING

- **Example:**
  - A resistor manufacture is testing the quality of a batch of resistors with nominal value 1K Ohm, with a 95% confidence level. A sample of 100 resistors are tested, and the sample mean is 1040 Ohm, and the sample standard deviation is 100 Ohm. Do the resistors pass the quality check?

HYPOTHESIS TESTING

- **Binary hypothesis testing: Example**



Consider a radar system. When the target is present, the received signal is $X = v + N$, where $v = 1$ is the target voltage, and N is Gaussian noise with zero mean and variance 1. When the target is not present, the received signal is $X = N$. Define the binary hypothesis as

$$\begin{aligned} H_0 &: X = N \\ H_1 &: X = v + N \end{aligned}$$

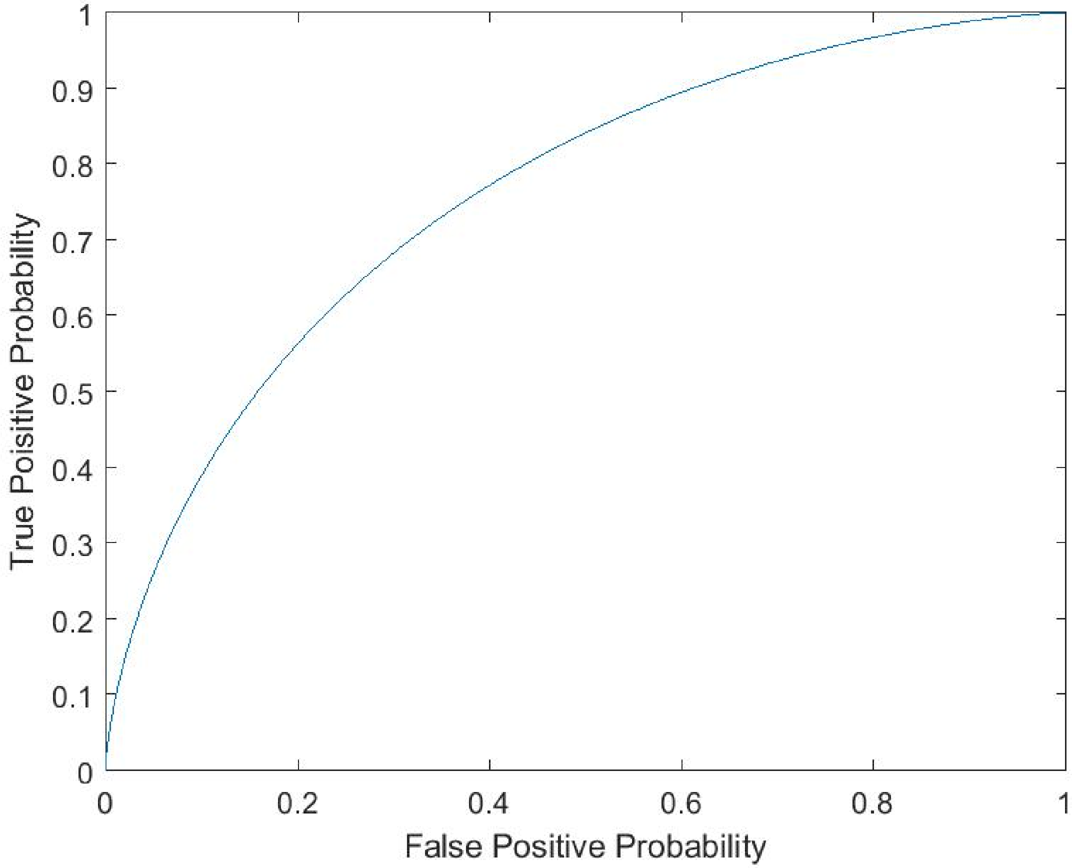
If $X \leq x_0$, the radar receiver accepts the null hypothesis; if $X > x_0$, the radar receiver rejects the null hypothesis.

1. Find the probability of false alarm (type I error)

2. Find the probability of missed detection (type II error)

HYPOTHESIS

- **Receiver operating characteristics (ROC) curve**
  - Plot true positive probability (complement of type II error) as a function of false positive probability (type I error)


True Poisitive Probability
False Positive Probability