UNIVERSITÉ Concordia UNIVERSITY

COMP 233/2

Probability and Statistics for Computer Science

Week 6

Normal random variables
Chi-square distribution
t-distributions

Reading: Chap 6, 10, 14

Clicker Question 5.2

Jane and Adam play the following game. They roll a pair of fair dice. Jane is the winner if the sum is 6. Adam is the winner if the sum is 3 or 5. If the sum is none of the above, they continue playing until one of them wins. Assuming that successive rolls are independent, what is the probability that Jane eventually wins?

2/10/2020

Solution

Let E denote the event that Jane eventually wins.
Let A denote the event with outcome 6.
Let B denote the event with outcome 3 or 5.
Let C denote neither A nor B.
P(A) = 5/36 P(B) = 6/36 and P(C) = 25/36

P(E) = P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C)
= 1(5/36) + 0(6/36) + P(E)(25/36)

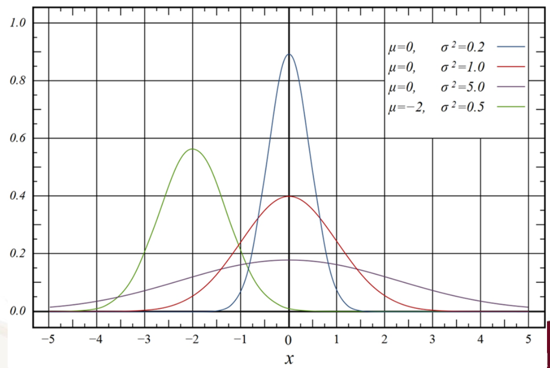
The Normal Density Function

- A continuous RV that is determined by the PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is said to be ***normally*** distributed ***with mean μ and standard deviation σ.***
- Also called a Gaussian distribution.

Normal Density Curves (various)



Normal Density Function Curve

- Convince yourself that
  - *f(x)* has a relative and absolute maximum at μ.
  - *f(x)* is symmetric about the vertical line x=μ.
  - *f(x)* has two inflection points at μ±σ.
- Think of how the curve changes if one varies the parameters μ and σ.

2/10/2020

CDF, Mean and Variance

$$f(\mathrm{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}}$$

$$P\{-\infty \le X \le \infty\} = \int_{-\infty}^{\infty} f(\mathrm{x})d\mathrm{x} = 1.$$

$$E[X] = \int_{-\infty}^{\infty} \mathrm{x}f(\mathrm{x})d\mathrm{x} = \mu,$$

$$E[X^2] = \int_{-\infty}^{\infty} \mathrm{x}^2 f(\mathrm{x})d\mathrm{x} = \mu^2 + \sigma^2,$$

$$Var(X) = E[X^2] - E[X]^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$$

- Hint: $z = \frac{\mathrm{x}-\mu}{\sigma}$ simplifies the expression

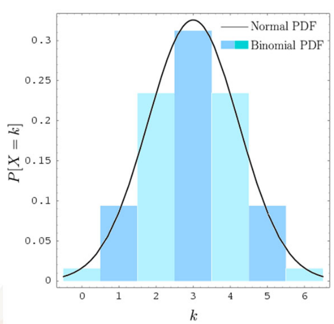
Why is the Normal Distribution Important?

- In practice, many random phenomena obey, at least approximately, a normal probability distribution. Some examples of this behaviour are
  - the *height* of a person,
  - the *velocity* in any direction of a molecule in gas, and
  - the *error* made in measuring a physical quantity.
- The sum of a large number of independent random variables is approximately normally distributed (this is the central limit theorem which we will study later).

Why is the Normal Distribution Important, cont.?

- The normal cumulative distribution function can be used as an approximation to some other cumulative distribution functions. For example:
  - A *binomial distribution* with n and p is approximately normal for large n and p not too close to 1 or 0. The approximating normal distribution has $\mu = np$, $\sigma^2 = np(1 - p)$.
  - A *Poisson distribution* with parameter λ is approximately normal for large λ. The approximating normal distribution has parameters $\mu = \sigma^2 = \lambda$.

Approximating the Binomial PDF



http://en.wikipedia.org/wiki/File:BinDistApprox_large.png

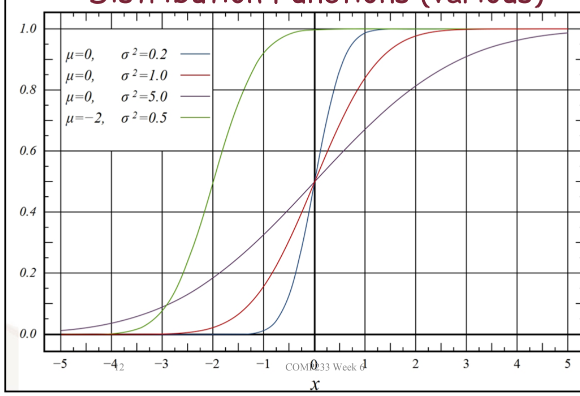
The Cumulative Distribution Function

- Probabilities involving the Normal Distribution may be computed using its Cumulative Distribution Function:

$$F(a) = P\{X \le a\} = \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$$

- Since this is awkward in general to compute for arbitrary mean μ and standard deviation σ, a standardized version of the distribution is frequently used instead.

Distribution Functions (various)



Observation

- By definition,

$$P\{a \le X \le b\} = \int_a^b f(\mathrm{x})dx = \frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}}d\mathrm{x}.$$

- We shall try simplifying the integrand.

Observation

- A simple substitution implies that

$$P\{a \le X \le b\} = \int_a^b f(\mathrm{x})dx = \frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}}dx$$

$$= \left| \begin{array}{c} \mathrm{z} = \frac{\mathrm{x} - \mu}{\sigma}, \quad dz = \frac{d\mathrm{x}}{\sigma} \\ a \to \frac{a - \mu}{\sigma} = z_a, b \to \frac{b - \mu}{\sigma} = z_b \end{array} \right| = \ldots$$

Observation

- A simple substitution implies that

$$P\{a \le X \le b\} = \int_a^b f(\mathrm{x})dx = \frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}}dx$$

$$= \left| \begin{array}{c} \mathrm{z} = \frac{\mathrm{x} - \mu}{\sigma}, \quad dz = \frac{d\mathrm{x}}{\sigma} \\ a \to z_a, b \to z_b \end{array} \right| =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{z_a}^{z_b} e^{-\frac{z^2}{2}}dz = P\{z_a \le Z \le z_b\}.$$

Conclusion

- A probability for a [non-standard] ND can be computed via a Standard ND (SND).

$$P\{a \le X \le b\} = P\{z_a \le Z \le z_b\},$$

$$\text{where } \frac{a - \mu}{\sigma} = z_a, \frac{b - \mu}{\sigma} = z_b.$$

- We only need to find a way to compute SND probabilities.

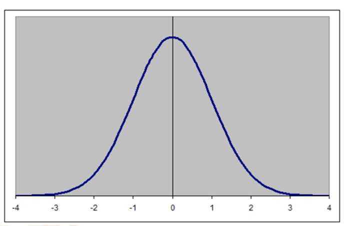
The Standard Normal Distribution (SND)

- A normally distributed RV is said to have ***standard ND*** if μ = 0 and σ = 1.
- It is traditionally denoted by random variable Z and its cumulative distribution function (CDF) is

$$\Phi(z) = P\{Z \le z\} = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy, \quad -\infty < z < \infty$$

SND Curve

- The curve of SND is bell-shaped, symmetric about Z=0.



"Bell Curve"

CDF of Standard ND

- It is known that it is impossible to express an anti-derivative ***Φ*** explicitly:

$$\Phi(z) = P\{Z \le z\} = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy, \quad -\infty < z < \infty$$

- However, there is a table (appendix Table 3 in A1 of text) with values of ***Φ***.
- Hence, $P\{a \le Z \le b\} = \Phi(z_b) - \Phi(z_a)$.

Example

- If X is a normal random variable (NRV) with mean $\mu = 3$ and variance $\sigma^2 = 16$, find
  - (a) P{X < 11};
  - (b) P{X > −1};
  - (c) P{2 < X < 7}.

Solution (a)

1. Find the appropriate Z value.

$$z = \frac{x-\mu}{\sigma} = \frac{11-3}{4} = 2.$$

1. Find the appropriate area:

$P\{X < 11\}$ = $P\{(X\text{-}3)/4 < (11\text{-}3)/4\}$
= $P\{Z < 2\}$ = $\Phi(2)$
= 0.9772.

2. Thus, $P\{X < 11\}$ = 0.9772.

Solution (b)

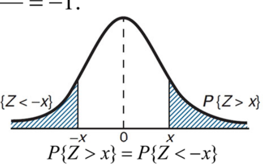
1. Find the appropriate Z value.

$$z = \frac{x-\mu}{\sigma} = \frac{-1-3}{4} = -1.$$

1. Find the appropriate area:

$P\{Z > -1\}$ = $P\{Z < 1\}$ = $\Phi(1)$
= 0.8413.

2. Thus, $P\{X < -1\}$ = 1- 0.8413.



Solution (c)

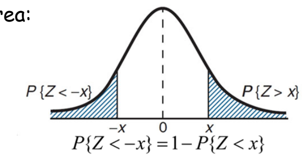
1. Find the appropriate Z values.

$$z_2 = \frac{x-\mu}{\sigma} = \frac{2-3}{4} = -\frac{1}{4} \text{ and } z_7 = \frac{x-\mu}{\sigma} = \frac{7-3}{4} = 1$$

1. Find the appropriate area:

$P\{-1/4 < Z < 1\}$
= $P\{Z < 1\}$ - $P\{Z < -1/4\}$
= $\Phi(1)$ – [1- $\Phi(1/4)$]
= 0.8413-[1-0.5987]
= 0.4400.

2. Thus, $P\{2 < X < 7\}$ = 0.4400.



Example

- A monthly amount of newspaper for garbage or recycling is normally distributed with a mean of 28 and a standard deviation of 2 pounds. If a household is selected at random, find the probability of its generating

(a) More than 30.2 pounds per month;

(b) Less than 90 pounds per three months.

Solution (a)

1. Find the appropriate Z value.

$$z = \frac{x-\mu}{\sigma} = \frac{30.2-28}{2} = 1.1.$$

1. Find the appropriate area:
   $P\{Z > 1.1\}$= 1 – $P\{Z < 1.1\}$ = 1 – $\Phi(1.1)$
   = 1 – 0.8643 = 0.1357.

2. Thus, $P\{X > 30.2\}$ = 0.1357.

Solution (b)

- We notice that the amount of newspaper is the SUM of the values for the first, the second and the third months, i.e.

$$X = X_1 + X_2 + X_3$$

- We know the Density Function of each term, but what is the Density Function of the whole sum?
- So we need to find out how the sums of random variables work!

2/10/2020

Reproductive Property of Normal Distribution

- Let X_k, k=$1,\ldots,n$, be *independent normally distributed* RVs. Then

$$E[X_1 + \ldots + X_n] = \sum_k E[X_k],$$

$$Var(X_1 + \ldots + X_n) = \sum_k Var(X_k),$$

$$X = X_1 + \ldots + X_n \text{ is Normally Distributed.}$$

The Normal Distribution (ND)

- That is the sum of independent NDs is a ND with the PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

$$\mu = \sum_k \mu_k, \quad \sigma = \sqrt{\sum_k \sigma_k^2}.$$

Reminder: Standard Normal Distribution

- A non-standard ND can be *standardized* by

$$Z = \frac{X-\mu}{\sigma}.$$

- That is, $P\{a \le X \le b\} = P\{z_a \le Z \le z_b\}$,

  where $\frac{a-\mu}{\sigma} = z_a, \frac{b-\mu}{\sigma} = z_b$.

Example (again)

- A monthly amount of newspaper for garbage or recycling is normally distributed with a mean of 28 and a standard deviation of 2 pounds. If a household is selected at random, find the probability of its generating

  (a) More than 30.2 pounds per month;

  (b) Less than 90 pounds per three months.

2/10/2020

Solution (b)

- We notice that the amount of newspaper is the <u>SUM</u> of the values for the first, the second and the third months, i.e.

$$X = X_1 + X_2 + X_3$$

- We know now that X is ND with a mean of

$$\mu = n\mu_i = 3(28) = 84 \text{ pounds}$$

  and a standard deviation of

$$\sigma = \sqrt{n\sigma_i^2} = \sqrt{3}(2) = 3.464 \text{ pounds.}$$

- It remains to use the SND table and calculate

Solution (b)

1. Find the appropriate Z value for 90.

$$z = \frac{x-\mu}{\sigma} = \frac{90-84}{3.464} = 1.73.$$

1. Find the appropriate area:

   $P\{Z < 1.73\}$ = $\Phi(1.73)$ = 0.9582.

2. Thus, $P\{X < 90\}$ = 0.9582.

Example

- The power W dissipated in a resistor is proportional to the square of the voltage V. That is,

$$W = rV^2$$

where r is a constant. If r = 3, and V can be assumed (to a very good approximation) to be a normal random variable with mean 6 and standard deviation 1, find

(a) E[W];

(b) P{W > 120}.

$$Var(V) = E[V^2] - (E[V])^2$$

Solution

(a) $E[W] = E[3V^2] = 3E[V^2]$

$= 3\left(Var(V) + (E[V])^2\right)$

$= 3(1+36) = 111$

(b) $P\{W > 120\} = P\{3V^2 > 120\} = P\left\{V > \sqrt{40}\right\}$

$= P\{V - 6 > \sqrt{40} - 6\}$

$= P\{Z > .3246\}$

$= 1 - \Phi(.3246) = 0.3727$

Food for Thought – 6.1

- The Natural χ Coffee Shop daily customer load follows a normal distribution with mean 45 and standard deviation 8.

a) Determine the probability that the number of customers tomorrow will be less than 42.

b) Determine the probability that the total number of customers over the next four days will be greater than 200.

Solution (a)

1. Find the appropriate Z value.

$$z = \frac{x-\mu}{\sigma} = \frac{42-45}{8} = -0.375$$

2. Find the probability:

$P\{Z < -0.375\}$

$= \Phi(-0.375)$

$= 1 - \Phi(0.375)$

$= 1 - 0.64617$

$= 0.35383.$

Solution (b)

1. Since we want the sum of four (n = 4) normal variables (one per day), we need the mean and standard deviation of the sum:
   $\mu = n\mu_i = 4(45) = 180$ customers
   $\sigma = \sqrt{n}\sigma_i = \sqrt{4}(8) = 16$ customers
2. Find the probability:
   $$P\left\{Z > \frac{x-\mu}{\sigma}\right\} = P\left\{Z > \frac{200-180}{16}\right\}$$
   $$= P\{Z > 1.25\} = 1 - P\{Z \le 1.25\}$$
   $$= 1 - \Phi(1.25) = 1 - (0.89435)$$
   $$= 0.10565$$

Food for Thought 6.2

- The Natural χ Coffee Shop daily customer load follows a normal distribution with mean 45 and standard deviation 8.
- Due to its remote location, coffee bean deliveries come once every four days. Any coffee left over after four days is thrown out (discarded) when the new delivery arrives.

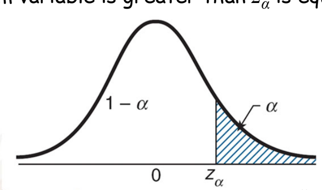
c) For how many customers should the owner plan for (i.e., to have just enough coffee) so that the shop runs out of coffee only 5% of the time?

2/10/2020

Solution (c)

Notation: z_α

- For $\alpha \in (0, 1)$, let z_α be such that
  $$P\{Z > z_\alpha\} = 1 - \Phi(z_\alpha) = \alpha$$
- That is, the probability that a standard normal random variable is greater than z_α is equal to α.



Computing z_α

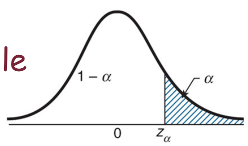
- The value of z_α can, for any α, be obtained from the SND Table (or program). For instance, since

  1 – Φ(1.645) = .05
  1 – Φ(1.96) = .025
  1 – Φ(2.33) = .01

  it follows that

  $z_{.05} = 1.645$, $z_{.025} = 1.96$, $z_{.01} = 2.33$

100(1 – α) percentile



- Since

$$P\{Z < z_\alpha\} = 1 - \alpha$$

  it follows that 100(1 – α) percent of the time a standard normal random variable will be less than z_α.
- As a result, we call z_α the 100(1 – α) percentile of the standard normal distribution.

2/10/2020

Distributions related to the Normal Distribution

Chi square distribution
Student t-distribution

Chi-Square Distribution

- Let $Z_1, \ldots, Z_n$ be independent standard normally-distributed random variables.
- The random variable

$$\chi_n^2 = Z_1^2 + \ldots + Z_n^2$$

  is a ***chi-square*** distribution with ***n degrees of freedom*** (d.f.).

  E[Z²] = 1 and Var(Z²) = 2. Apply the results on the sum of independent RV.

Parameters of Chi-Square Distribution

- The properties of *expectation* and *variance* imply that

$$E\left[\chi_n^2\right] = n, \quad Var\left(\chi_n^2\right) = 2n.$$

Chi-Square Density Function

- The PDF of a χ^2-distributed RV with n degrees of freedom is given by

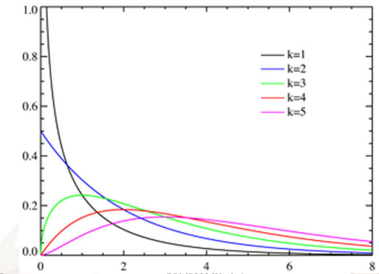
$$f_n(x) = a_n x^{(n-2)/2} e^{-x/2}, \; x > 0,$$

where a_n is a function of n.

2/10/2020

Density Function Curve

- Here are the curves of χ^2-density functions for some values of degrees of freedom.



Density Function of Chi-Square, even n

- If the number of degrees of freedom is even, then the PDF for a χ^2-distributed random variable is given by

$$f_n(x) = \frac{x^{(n-2)/2} e^{-x/2}}{2^{n/2}(n-1)!}, x > 0.$$

Elementary Property of Chi-Square Distribution

- The sum of two independent χ^2-distributions with *m* and *n* degrees of freedom, respectively, is again a χ^2-distributed variable with *m+n* degrees of freedom, i.e.

$$\chi^2_{n+m} = \chi^2_m + \chi^2_n$$

Computing Probabilities for χ^2-Distribution

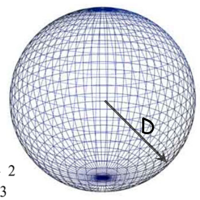
- In order to compute probabilities for a χ^2-distributed variable, we have to integrate the respective density.
- An explicit integration was not possible in the case of a ND. Neither it is for χ^2 (if *n* is odd).

2/10/2020

Example

- Suppose that we are attempting to locate a target in three-dimensional space, and that the three coordinate errors (in meters) of the point chosen are independent normal random variables with mean 0 and standard deviation 2.
- Find the probability that the distance between the point chosen and the target exceeds 3 meters.

Solution



- If D is the distance, then

$$D^2 = X_1^2 + X_2^2 + X_3^2$$

where X_i is the error in the i-th coordinate.
- Since X_i is a normal RV with μ=0, σ=2, the related standard normal RV Z_i is

$$z_i = \frac{x_i - \mu}{\sigma} = \frac{x_i - 0}{2} = \frac{x_i}{2}.$$

Solution

- So, it follows that

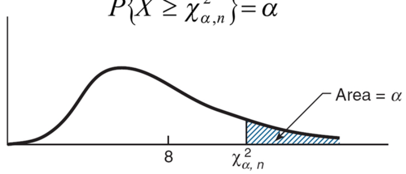
$$P\{D^2 > 9\} = P\{X_1^{\ 2} + X_2^{\ 2} + X_3^{\ 2} > 9\}$$
$$= P\left\{\left(\frac{X_1}{2}\right)^2 + \left(\frac{X_2}{2}\right)^2 + \left(\frac{X_3}{2}\right)^2 > \frac{9}{2^2}\right\}$$
$$= P\{\chi_3^{\ 2} > 9/4\} = 1 - P\{\chi_3^{\ 2} \le 9/4\}$$
$$\approx 1 - 0.4789 = 0.5211$$

where the final equality was obtained from the chi-square probability calculator.

Notation

- If X is a chi-square random variable with n degrees of freedom, then for any $\alpha \in (0, 1)$, the quantity $\chi^2_{\alpha,n}$ is defined to be such that

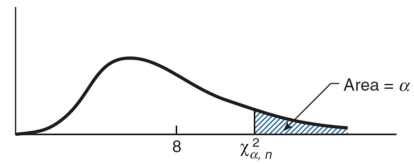
$$P\{X \ge \chi^2_{\alpha,n}\} = \alpha$$



2/10/2020

Computing Probabilities for χ^2-Distribution

- 100(1 – α) percentage points, for select popular choices of α, for the χ^2-distribution, which are important for applications, can be taken from a table (Table 5 in A1 of text).



Student's t-distribution

- Suppose the RV Z and χ^2 are independent.
- A random variable T_n is said to be ***t*-distributed** with ***n* degrees of freedom** if

$$T_n = \frac{Z}{\sqrt{\dfrac{\chi_n^2}{n}}}.$$

Density of t-Distribution

- The PDF for a t-distributed random variable with n d.f. is given by

$$g_n(t) = \frac{c_n}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}$$

where $c_n = \dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}$ which depends on n.

Example PDFs of t-Distribution

- The PDF for a t-distributed random variable with $n=1$ d.f. is given by

$$g_1(t) = \frac{1}{\pi(1+x^2)}$$

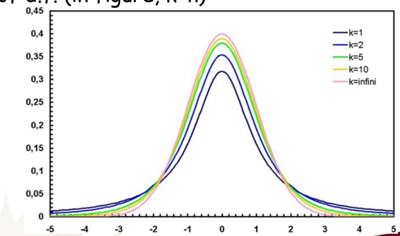
- And with $n=2$ d.f. is given by

$$g_2(t) = \frac{1}{(2+x^2)^{3/2}}$$

2/10/2020

Density Function Curve

- Here are the curves of t-densities for some values n of d.f. (in figure, k=n)



http://en.wikipedia.org/wiki/File:Student_densite_best.JPG

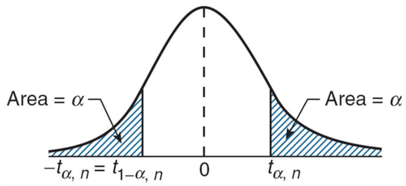
The t-distribution vs. the Standard Normal Distribution

1. The t-distribution is bell-shaped and symmetric about the mean.
2. The curve never touches the x-axis.
3. $E[T_n] = 0,\ n > 1.$
4. $Var(T_n) = n/(n-2),\ n > 2$, i.e. always > 1.
5. As the number of degrees of freedom increases, the *t-distribution* approaches a *standard normal distribution.*

Notation

- For α, 0 < α < 1, let $t_{\alpha,n}$ be such that

$$P\{T_n \geq t_{\alpha,n}\} = \alpha$$



The t-distribution

- To compute probabilities for a t-distributed variable, we have to integrate the respective density.
- It is natural to expect that explicit integration will be difficult.
- Some percentage points for t-distributed random variables for various α and n are collected in a table (Table 4 in A1 of text).

2/10/2020

Future Plans

- After the midterm we are going to see some applications of probabilistic techniques in *Statistics*.
- The first step is to get familiarized with the fundamental concepts of the statistical analysis.
- So the next lecture in Week 8 is on

  INTRODUCTION TO STATISTICS.

References/Resources Used

- Lecture Slides for MATH 401 of Dr. Oleksiy Us, Department of Mathematics, German University of Cairo. [PPT]