CM1103 PROBLEM SOLVING WITH PYTHON

STATISTICS

Dr Jing Wu
School of Computer Science & Informatics
Cardiff University

Overview

- Basic concepts: statistics, sample, population
- Two major types of statistics:
  - Descriptive statistics
  - Inferential statistics
- Two important measures:
  - Measures of centre
  - Measures of variation
- Statistics in Python
- Probability distribution
  - Discrete random variable
- Normal distribution

Statistics

Britain's Mr and Mrs average

A BREAKDOWN OF BRITAIN'S AVERAGE FAMILY

Woman's name - Susan

Man's name - David

Type of house - Semi-detached

Average number living in home - 2.7

Average waking hours spent there on a weekday - 8.1

Contents worth - £35,486

Buildings insured for - £177,790

Number of bedrooms - 3

Number of bathrooms - 1

Most wanted home improvement - New kitchen

Year home built in - 1930

Financed with - A mortgage

Average car driven - Ford Focus

Likelihood of being a smoker? - 8%

Most common house name - The Cottage

Statistics

Definition (statistics)
Statistics consists of a body of methods for collecting and analysing data.

Population and Sample

Definition (Population)
Population is the collection of all individuals or items under consideration in a statistical study.

Definition (Sample)
Sample is that part of the population from which information is collected.

Example	
A statistician is interested in the average height of British men. 1% are randomly sampled.	
What is the population?	
What is the sample?	

Sample: A window

↓ to

Population: Target ("the truth")

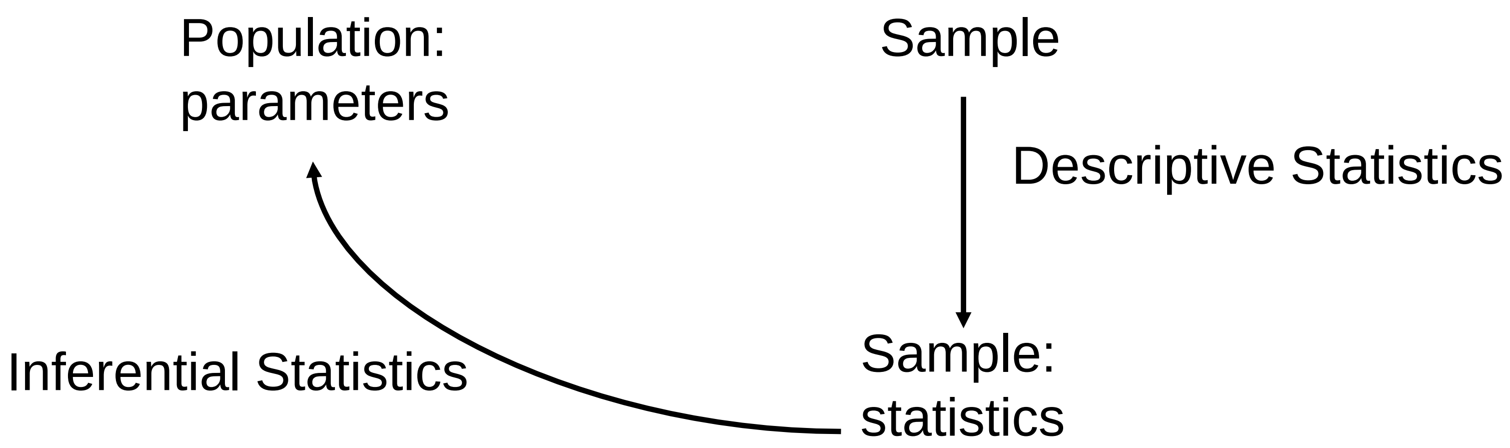
Descriptive and Inferential Statistics

Definition (Descriptive Statistics)

Descriptive statistics consist of methods for organising and summarising information from data, but not making conclusions

Definition (Inferential Statistics)

Inferential statistics consists of methods for drawing and measuring the reliability of conclusions about population based on information obtained from the descriptive statistics.



Measures of Centre

Definition (Mode)

Obtain the frequency of each observed value of the variable. The value occurs with the greatest frequency (2 or greater) is called a **sample mode** of the variable.

Definition (Median)

Arrange the observed values of the variable in ascending order. The **sample median** is the middle value in the ordered list. If there are n observations,
If n is odd, the sample median is the observed value at position $(n+1)/2$;
If n is even, the sample median is the value halfway between the two middle observed values.

Definition (Mean)

The **sample mean** of the variable is the sum of observed values in a data divided by the number of observations. If there are n observations: $x_1, x_2, \ldots, x_n$, then the sample mean is:

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

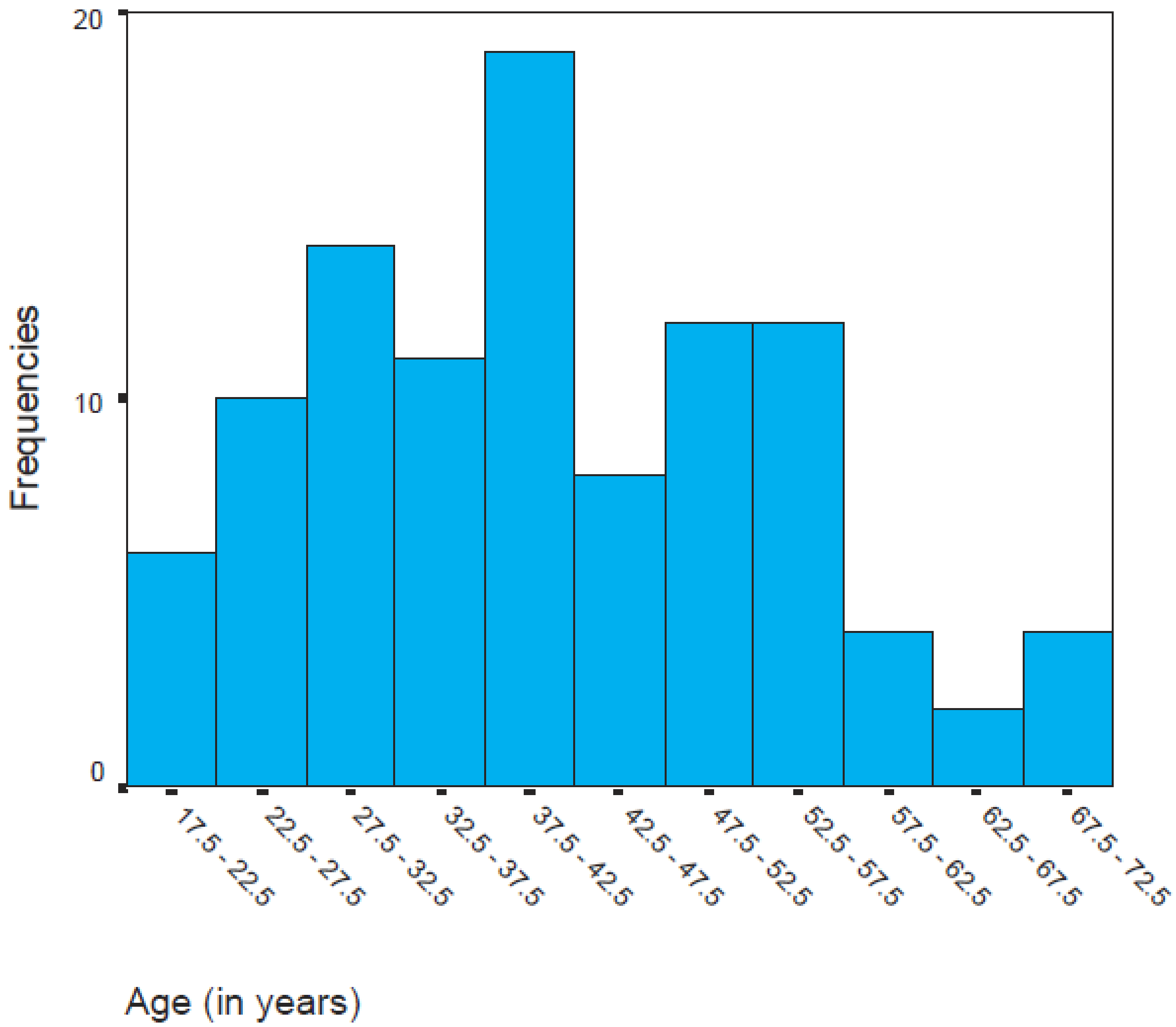
Measures of Centre

Example	
Collection of ages (in years) of 102 people 19, 20, 20, 21, 22, 22, 24, 24, 25, 25, 26, 26, 27, 27, 27, 28, 28, 28, 29, 29, 29, 29, 29, 30, 31, 31, 32, 32, 32, 33, 34, 34, 34, 35, 35, 36, 36, 36, 37, 37, 38, 38, 38, 38, 38, 39, 39, 39, 39, 39, 40, 40, 40, 40, 42, 42, 42, 42, 42, 43, 44, 44, 44, 45, 45, 46, 46, 46, 48, 48, 48, 48, 49, 49, 49, 50, 50, 51, 52, 53, 54, 54, 54, 55, 55, 56, 56, 56, 56, 56, 57, 57, 58, 60, 62, 62, 65, 67, 68, 68, 69, 72	
What’s the mode?	
What’s the median?	
What’s the mean?	

The mode

Frequency distribution of people's age

		Frequency	Percent	Cumulative Percent
Valid	18 - 22	6	5.9	5.9
	23 - 27	10	9.8	15.7
	28 - 32	14	13.7	29.4
	33 - 37	11	10.8	40.2
	38 - 42	19	18.6	58.8
	43 - 47	8	7.8	66.7
	48 - 52	12	11.8	78.4
	53 - 57	12	11.8	90.2
	58 - 62	4	3.9	94.1
	63 - 67	2	2.0	96.1
	68 - 72	4	3.9	100.0
	Total	102	100.0	



Histogram for people's age

Example

Participants in bike race had the following finishing times in minutes: 28, 22, 26, 29, 21, 23, 24.	
What is the median?	
What is the mean?	

Participants in bike race had the following finishing times in minutes: 28, 22, 26, 29, 21, 23, 24, 100.	
What is the median?	
What is the mean?	

Mean can be highly influenced by an observation that falls far from the rest of the data, called an **outlier**.

Measures of Variation

Definition (Range)
The **sample range** of the variable is the difference between its maximum and minimum values in a data set. $$Range = Max - Min$$

Example	
Participants in bike race had the following finishing times in minutes: 28, 22, 26, 29, 21, 23, 24.	
What is the range?	
If the finishing times are 28, 22, 26, 29, 21, 23, 24, 100,	
What is the range?	

Measures of Variation

Definition (Quartiles)

Let n denote the number of observations in a data set. Arrange the observed values of variable in a data in increasing order.

The **first quartile Q_1** is at position $\frac{n+1}{4}$ in the ordered list.

The **second quartile Q_2** (the median) is at position $\frac{n+1}{2}$ in the ordered list.

The **third quartile Q_3** is at position $\frac{3(n+1)}{4}$ in the ordered list.

If a position is not a whole number, linear interpolation is used.

Definition (Interquartile range)

The **sample interquartile range** of the variable, denoted IQR, is the difference between the first and third quartiles of the variable:

$$IQR = Q_3 - Q_1$$

Example

Example	
Participants in bike race had the following finishing times in minutes: 28, 22, 26, 29, 21, 23, 24.	
What are the first, second, third quartiles?	
What is the interquartile range?	
If the finishing times are 28, 22, 26, 29, 21, 23, 24, 100,	
What is the interquartile range?	

Measures of Variation

Definition (Standard deviation)

For a variable x, the sample standard deviation, denoted by S_x, is

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Instead of n
Bessel's correction

Example

Participants in bike race had the following finishing times in minutes: 28, 22, 26, 29, 21, 23, 24.

What is the standard deviation?

Statistics in Python

statistics module

https://docs.python.org/3/library/statistics.html

statistics.mean(x)

statistics.median(x)

statistics.mode(x)

statistics.stdev(x)

Probability Distributions

Definition (Probability)

The **probability** of a particular outcome is the proportion of times that outcome would occur in a long run of repeated observations.

Definition (Random variable)

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

Example: flip a coin

Value of X	0 (tail)	1 (head)
Probability	1/2	1/2

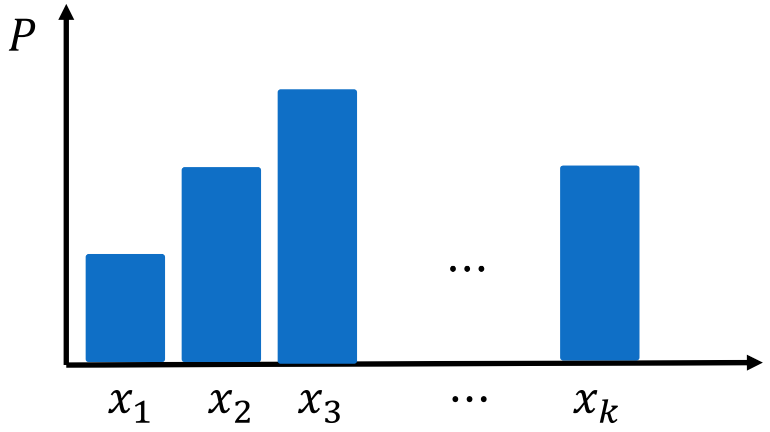
Probability Distributions

Definition (Probability distribution)

Probability distribution of a random variable X lists all the possible outcomes together with their probabilities.

Discrete random variable

Value of X	x_1	x_2	x_3	$\cdots$	x_k
Probability	$P(x_1)$	$P(x_2)$	$P(x_3)$	$\cdots$	$P(x_k)$



$P(x_i)$ must satisfy

1. $0 \leq P(x_i) \leq 1$
2. $P(x_1) + P(x_2) + \cdots + P(x_k) = 1$

Probability Distributions

Value of X	x_1	x_2	x_3	$\cdots$	x_k
Probability	$P(x_1)$	$P(x_2)$	$P(x_3)$	$\cdots$	$P(x_k)$

Mean of the discrete random variable X (expected value of $X, E(X)$)

$$\mu = x_1P(x_1) + x_2P(x_2) + x_3P(x_3) + \cdots + x_kP(x_k)$$
$$= \sum_{i=1}^{k} x_iP(x_i)$$

Variance of the discrete random variable X:

$$\sigma^2 = (x_1 - \mu)^2P(x_1) + (x_2 - \mu)^2P(x_2) + \cdots + (x_k - \mu)^2P(x_k)$$
$$= \sum_{i=1}^{k} (x_i - \mu)^2P(x_i)$$

Probability Distribution

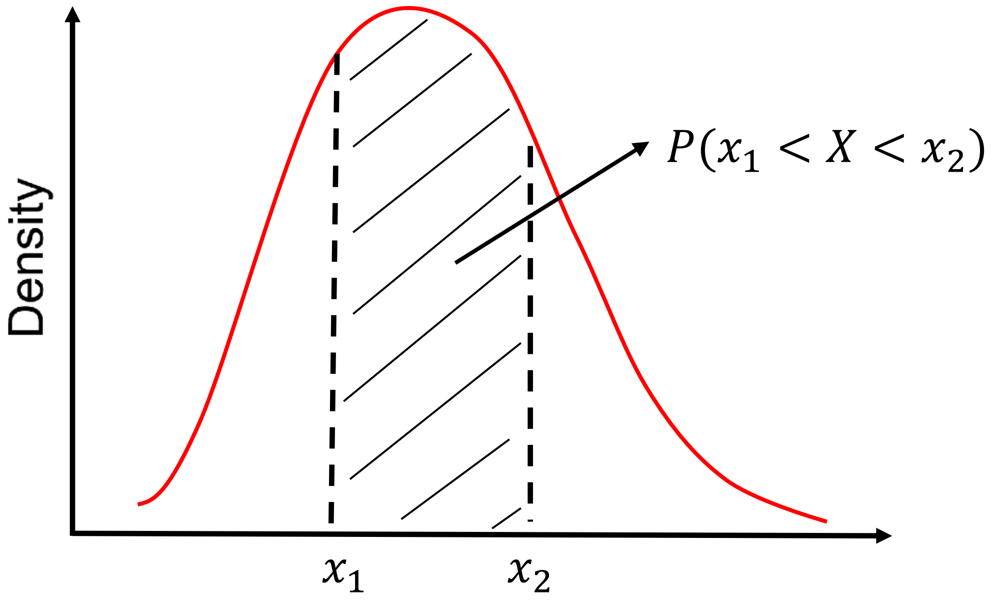
Continuous random variable $X \in [a, b]$

Probability is assigned to any interval $[x_1, x_2]$, where $x_1, x_2 \in [a, b]$.

It is required that

1. $0 \leq P(x_1 \leq X \leq x_2) \leq 1$ for any interval $[x_1, x_2]$.
2. $P(a \leq X \leq b) = 1$

Density curve



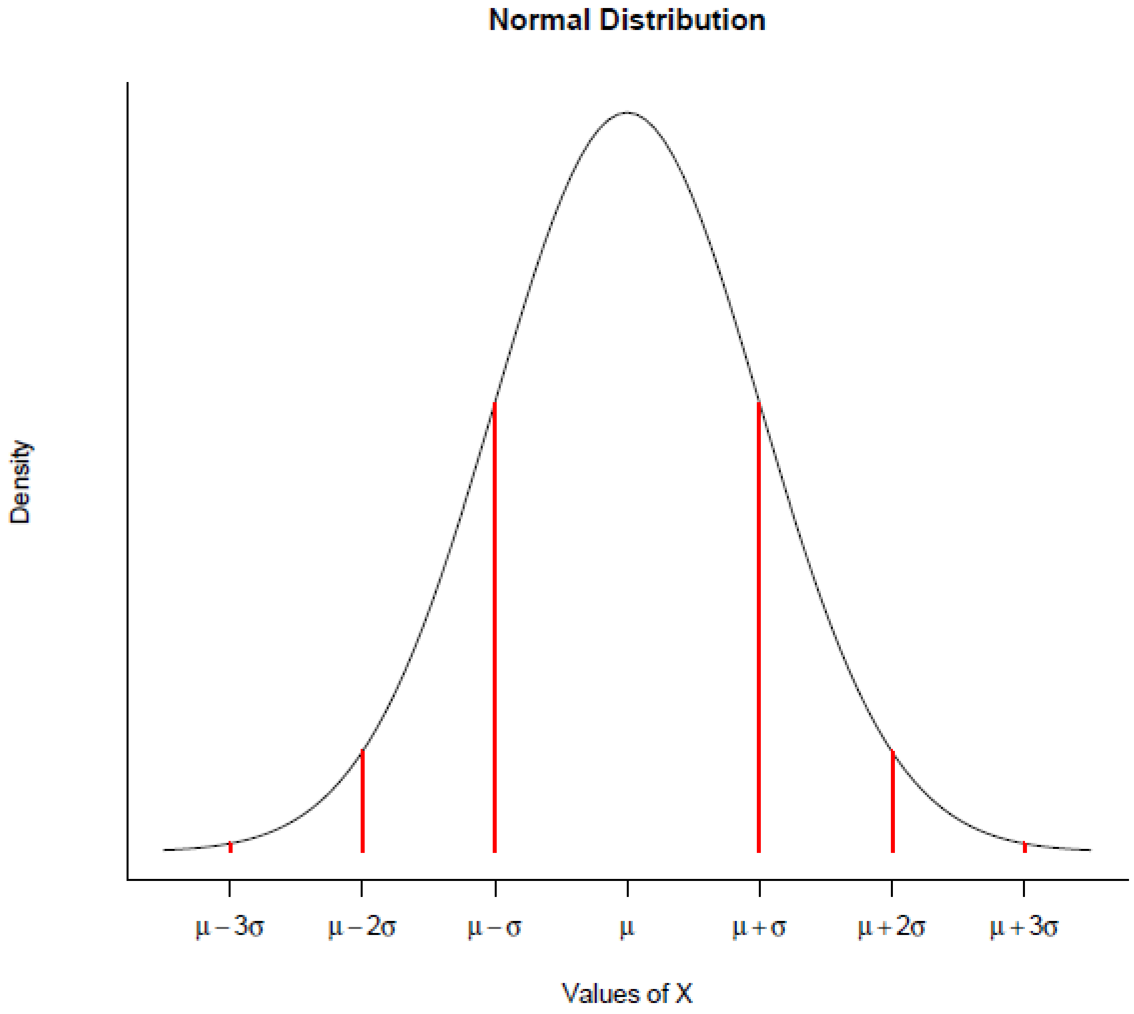
Normal Distribution

Definition (Normal distribution)

A continuous random variable X is said to be normally distributed or to have a **normal distribution** if its density curve is a symmetric, bell-shaped curve with the density function:

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

A random variable X following normal distribution with a mean of μ and standard deviation of σ is denoted by $X \sim N(\mu, \sigma)$.



Normal Distribution

Property:

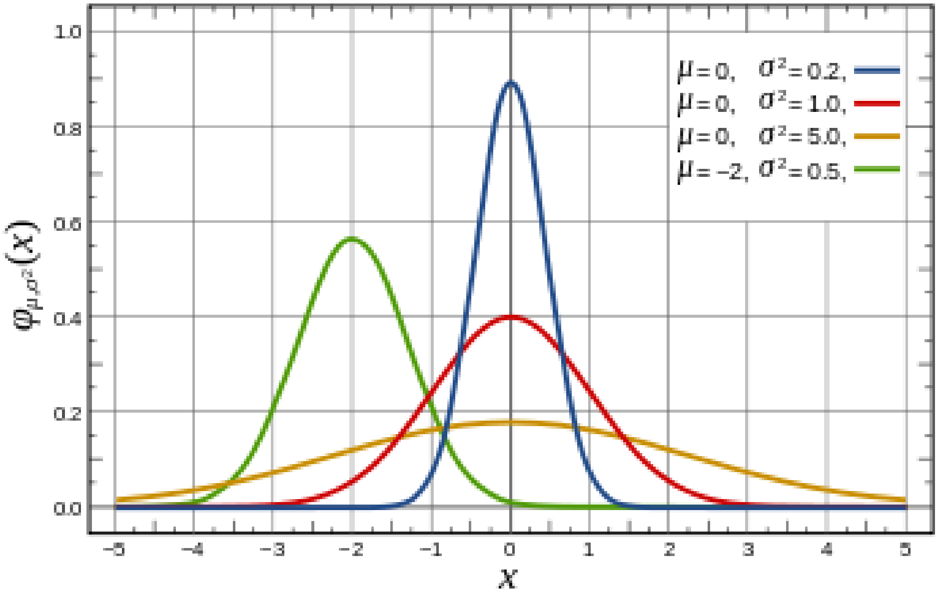
For each fixed number z, the probability within interval $[\mu - z\sigma, \mu + z\sigma]$ is the same for all normal distributions.

Particularly:

$$P(\mu - \sigma < X < \mu + \sigma) = 0.683$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$



Normal Distribution

Definition (Standard normal distribution)

A continuous random variable Z is said to have a **standard normal distribution** if Z is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$, i.e. $Z \sim N(0,1)$.

Standard normal table

https://en.wikipedia.org/wiki/Standard_normal_table

If the random variable X is distributed as $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma}$ is the standardized variable.

Z has the standard normal distribution, i.e. $Z \sim N(0, 1)$, and

$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$.

Example

Example:
The number of Calories in a salad on the lunch menu is normally distributed with mean $\mu = 200$ and standard deviation $\sigma = 5$. Find the probability that the salad you select will contain:
(a) More than 208 calories
(b) Between 190 and 200 calories

Summary

You should

- Understand the relationship between sample statistics and population parameters
- Be able to find the mode/median/mean/standard deviation from a set of numerical observed values, both by hand and using Python
- Understand probability distribution
- Know how to visualise probability distribution for discrete/continuous random variables
- Be able to use the standard normal table to calculate probabilities concerning normally distributed random variables.