GE2262 Business Statistics

Topic 3
Discrete & Continuous Probability Distributions

Reference

Levine, D.M., Krehbiel, T.C. and Berenson, M.L., *Business Statistics: A First Course*, Pearson Education Ltd, Chapter 3 & 5 & 6

Outline

- Discrete Probability Distribution
- Binomial Distribution
- Continuous Probability Distribution
- Normal Distribution

Random Variables

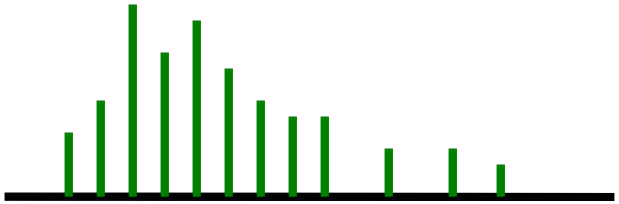
Random Variable

outcomes of an experiment with
probabilistic occurrence

Discrete Random Variable

produces outcomes that
come from a **counting process**

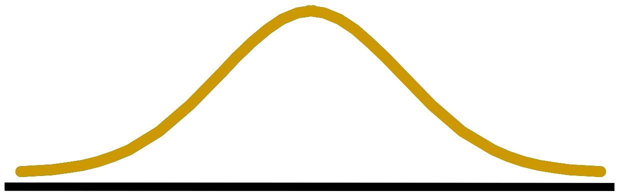
(e.g. number of courses you are
taking in this semester)



Continuous Random Variable

produce outcomes that
come from a **measurement**

(e.g. your annual salary,
or your weight)



Discrete Probability Distributions

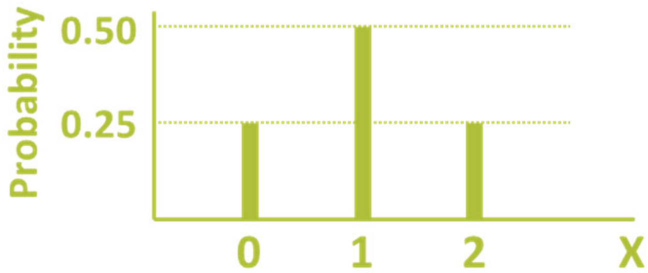
- A probability distribution for a discrete variable is a mutually exclusive listing of all possible numerical outcomes for that variable and a probability of occurrence associated with each outcome
  - The probabilities are obtained based on either priori knowledge (priori probability) or empirical approach (empirical probability)
    - Examples
      - Probability of selecting a black card from a deck of cards
      - Probability of a respondent who will purchase a HDTV
- The probability distribution of a variable forms a theoretical model which allows us to derive statistics and probabilities for the events related to the variable

Discrete Probability Distribution

Experiment: Toss 2 coins let X = No. of heads

Probability Distribution

X Value	Probability
0	1/4 = 0.25
1	2/4 = 0.50
2	1/4 = 0.25



Discrete Probability Distribution

Cont'd

Probability Distribution

X Value, x_i	Probability, $P(X = x_i)$
0	1/4 = 0.25
1	2/4 = 0.50
2	1/4 = 0.25
Total	**1**

- Mutually exclusive (Nothing in common)
- Collectively exhaustive (Nothing left out)

$$0 \leq P(X = x_i) \leq 1 \qquad \sum P(X = x_i) = 1$$

Discrete Random Variables – Measuring Center

Cont'd

- Expected value (Mean)
  - Weighted average of all possible values of X
  - Corresponding probability is treated as weight

$$\mu = E(X) = \sum_{i=1}^{N} x_i P(X = x_i)$$

X	P(X)
0	0.25
1	0.50
2	0.25

E.g. Toss 2 coins, count the number of heads

X = number of heads

Compute the expected value of X:

$$\mu = x_1 P(X = x_1) + x_2 P(X = x_2) + x_3 P(X = x_3)$$

Discrete Random Variables – Measuring Variation

Cont'd

- Variance
  - Weighted average squared deviation about the mean

$$\sigma^2 = \sum_{i=1}^{N}[x_i - E(X)]^2 P(X = x_i)$$

- Standard deviation
  - Square root of variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N}[x_i - E(X)]^2 P(P = x_i)}$$

E.g. Toss 2 coins, count the number of heads

X = number of heads

Compute the variance of X:

X	P(X)
0	0.25
1	0.50
2	0.25

$$\sigma^2 = (x_1 - \mu)^2 P(X = x_1) + (x_2 - \mu)^2 P(X = x_2) + (x_3 - \mu)^2 P(X = x_3)$$

Discrete Random Variables – Example

Cont'd

- Roll a fair die once. What is the expected value of the number rolled?

Number Rolled (x_i)	Probability $P(X = x_i)$	$x_iP(X = x_i)$
1	1/6	(1)(1/6) = 1/6
2	1/6	(2)(1/6) = 2/6
3	1/6	(3)(1/6) = 3/6
4	1/6	(4)(1/6) = 4/6
5	1/6	(5)(1/6) = 5/6
6	1/6	(6)(1/6) = 6/6
		$\mu = E(X) = 3.5$

- The expected value of the number rolled is 3.5

Discrete Random Variables – Example

Cont'd

- For the results of rolling a fair die, the expected value of the number rolled is 3.5
- Since you can never obtain a number of 3.5 by rolling a die, so what is the meaning of these statistics?
- How much money should we be willing to put up in order to have the opportunity of rolling a fair die if we were to be paid, in dollars, the amount on the face of the die?
  - On any particular roll, our payoff will be \$1.0, \$2.0, …, or \$6.0, but over many many rolls, the payoff can be expected to average out to \$3.5 per roll
  - If you pay less than \$3.5 for a roll, you are going to make a profit in long run
  - If you pay more than \$3.5 for a roll, you are going to loss in long run

Discrete Random Variables – Exercise

Cont'd

- Assume the following table shows the return per $1,000 for an investment under different economic conditions

Return in amount, Y_i	Economic Condition	P (Y_i)
-$200	Recession	0.2
+ 50	Stable Economy	0.5
+ 350	Expanding Economy	0.3

- Compute the expected return and standard deviation

Should you, from a statistical stand point, invest or not?

Calculating the Mean and Variance in Calculator (For Casio fx-50F)

Date Set:

X_j	20	30	40	50	60	75
$P(X_j)$	0.1	0.1	0.15	0.25	0.2	0.2

1. Change to "Lin" mode

   [MODE] [MODE] 5 1

2. Clear previous data

   [SHIFT] [CLR] 1 [EXE]

3. Input data

   20 [SHIFT] [,] 0.1 [M+] 30 [SHIFT] [,] 0.1 [M+]

   40 [SHIFT] [,] 0.15 [M+] 50 [SHIFT] [,] 0.25 [M+]

   60 [SHIFT] [,] 0.2 [M+] 75 [SHIFT] [,] 0.2 [M+]

4. Calculate descriptive statistics

   Mean: [SHIFT] 2 1 1 [EXE] = 50.5

   Population standard deviation: [SHIFT] 2 1 2 [EXE] = 17.02204453

Binomial Distribution

- A mathematical model is a mathematical expression representing some underlying phenomenon
- With such mathematical expressions are available, the exact probability of occurrence of any particular outcome of the random variable can be computed
- For discrete random variables, this mathematical expression is known as a probability distribution function
- One of the such probability distribution functions is called Binomial Distribution
  - A very important mathematical model used in many business situations

Binomial Distribution

Cont'd

What is the probability that, among 3 students, any 2 of them getting a pass in the test, with the probability of passing the test equals 0.7?

Student	A	B	C	Probability
Case 1	P	P	F	$0.7 \times 0.7 \times 0.3 = 0.147$
Case 2	P	F	P	$0.7 \times 0.3 \times 0.7 = 0.147$
Case 3	F	P	P	$0.3 \times 0.7 \times 0.7 = 0.147$

P(Any 2 getting pass)= 0.147 + 0.147 + 0.147 = 0.441

What is the probability that, among **30** students, any **20** of them getting a pass in the test, with the probability of passing the test equals 0.7?

Binomial Distribution – Conditions

Cont'd

- 'n' repetition of identical trials
  - E.g. totally 3 students
- 2 mutually exclusive outcomes (success and failure) in each trial
  - E.g. getting a pass or fail in the test
- Constant probability of success, π, in each trial
  - E.g. probability of getting a pass in the test for each student is 0.7, which is constant
- Trials are independent
  - E.g. the outcome of one student does not affect the outcome of the others

Binomial Distribution

Cont'd

- The binomial probability for a discrete random variable X is computed as

$$P(X = x) = \frac{n!}{x!\,(n-x)!}\pi^{x}(1-\pi)^{(n-x)}$$

Probability mass function

where $P(X = x)$ = probability that X = x events of interest (e.g. success)

n = number of observations

π = probability of an event of interest

X = number of events of interest in the sample (X = 0, 1, 2, ..., n)

$n! = n\,(n-1)\,(n-2)\,...\,(1)$; $0! = 1$

$\frac{n!}{x!(n-x)!}$ = no. of combinations of x successes out of n trials

π^{x} = total probability of having x successes

$(1-\pi)^{(n-x)}$ = total probability of having $(\mathrm{n}-x)$ failures

Binomial Distribution – Example

Cont'd

What is the probability that, among 3 students, any 2 of them getting a pass in the test, with the probability of passing the test equals 0.7?

X = no. of students passing the test out of 3 students
X follows Binomial distribution ($n = 3,\ X = 2,\ \pi = 0.7$)

$$P(X = 2) = \frac{n!}{x!\,(n-x)!}\pi^x(1-\pi)^{(n-x)}$$
$$= \frac{3!}{2!(3-2)!}0.7^2(1-0.7)^{(3-2)}$$
$$= 0.441$$

What is the probability that, among **30** students, any **20** of them getting a pass in the test, with the probability of passing the test equals 0.7?

Binomial Distribution

Cont'd

- Possible applications for the Binomial Distribution
  - A manufacturing plant labels items as either defective or acceptable
  - A firm bidding for contracts will either get a contract or not
  - A marketing research firm receives survey responses of "yes, I will buy" or "no, I will not"
  - New job applicants either accept the offer or reject it

Binomial Distribution – Exercise

Cont'd

An experiment about the interest of going to the cinema is conducted in a secondary school. Five students are selected randomly. Assume the probability of going to cinema within a week is 0.1.

X = no. of students going to cinema out of 5 students
X follows Binomial distribution $(n = 5, \pi = 0.1)$

The probability of 3 students going to the cinema out of these 5 students:

Binomial Distribution – Exercise

Cont'd

- What is the probability that there are 3 or more students going to the cinema within a week?

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)$$

- What is the probability that there are less than 3 students going to the cinema?

$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$$

Binomial Distribution Mean and Standard Deviation

Binomial Probability Distribution:

X_i	$P(X_i)$
0	0.59049
1	0.32805
2	0.0729
3	0.0081
4	0.00045
5	0.00001

$$\mu = \sum x_i P(X = x_i)$$

$$\sigma^2 = \sum (x_i - \mu)^2 P(X = x_i)$$

Binomial Distribution Mean and Standard Deviation

Cont'd

- If X follows a Binomial Distribution of size n and probability π, it can be shown that

$$\mu = n\pi$$
$$= (5)(0.1)$$
$$= 0.5$$

$$\sigma^2 = n\pi(1-\pi)$$
$$= (5)(0.1)(1\text{-}0.1)$$
$$= 0.45$$

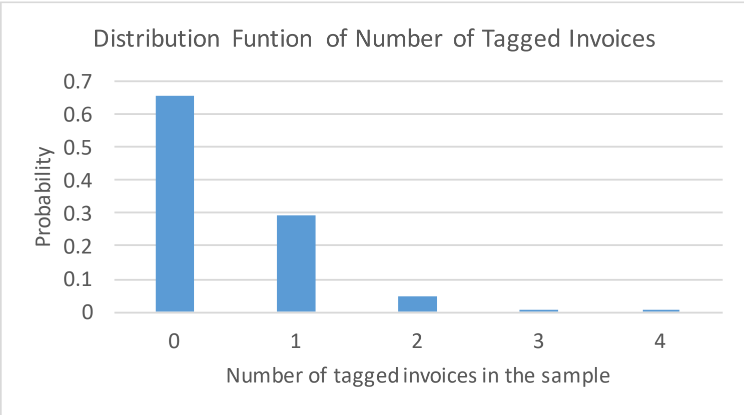
$$\sigma = \sqrt{n\pi(1-\pi)}$$
$$= \sqrt{(5)(0.1)(1-0.1)}$$
$$= 0.6708$$

Binomial Distribution

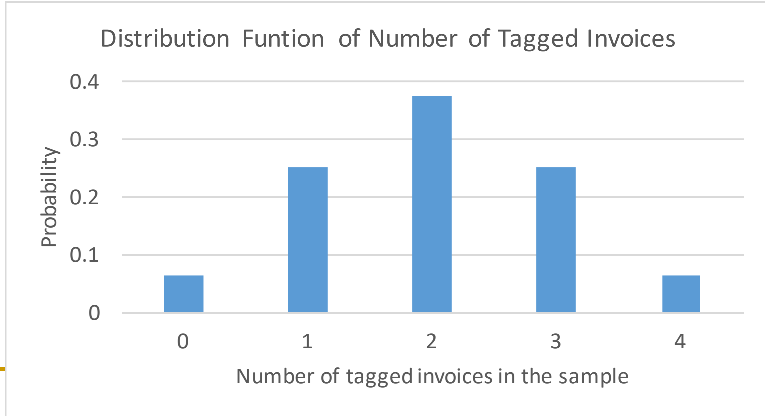
Cont'd

- The shape of a Binomial Distribution depends on the values of n and π
  - Whenever $\pi = 0.5$, the distribution is symmetrical, regardless of how large or small the value of n
  - Whenever $\pi \neq 0.5$, the distribution is skewed
    - $\pi < 0.5$, right-skewed; $\pi > 0.5$, left-skewed

n = 4, π = 0.1



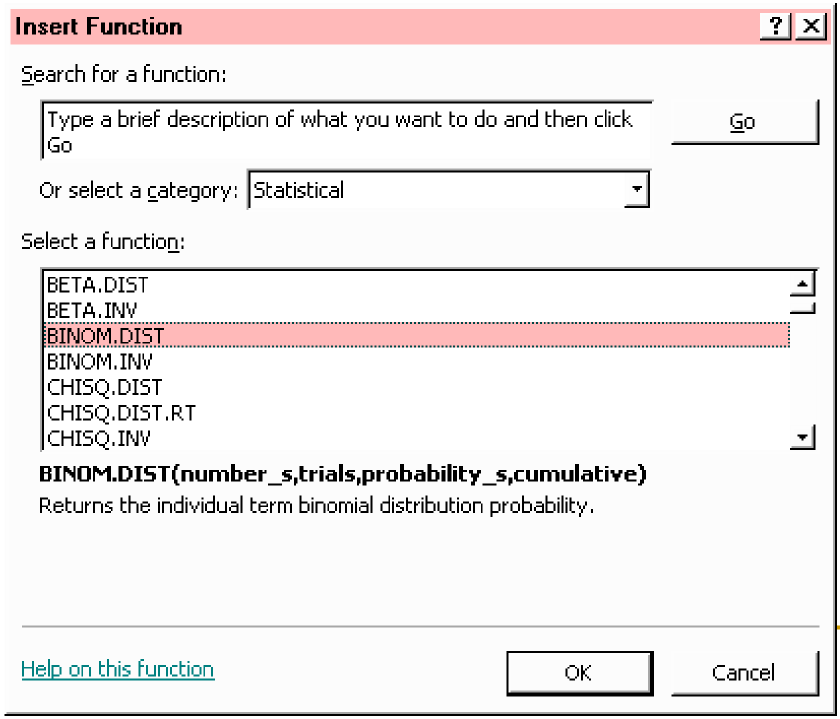
n = 4, π = 0.5

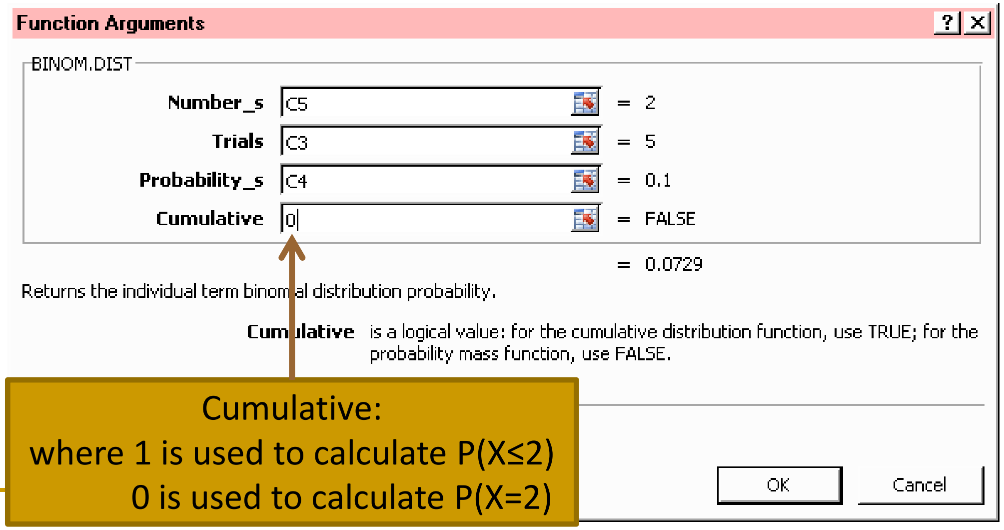


Binomial Distribution in Excel

- Using Excel to find out the probability that 2 students out of 5 going to cinema within 1 week
  - Type the required information (n, X, π)
  - Use BINOM.DIST function

	A	B	C
1	**Binomial Distribution**		
2			
3	Number of trials	n	5
4	Probability	π	0.1
5	Total number of successes in n trial	X	2
6			
7	P (X=2)		



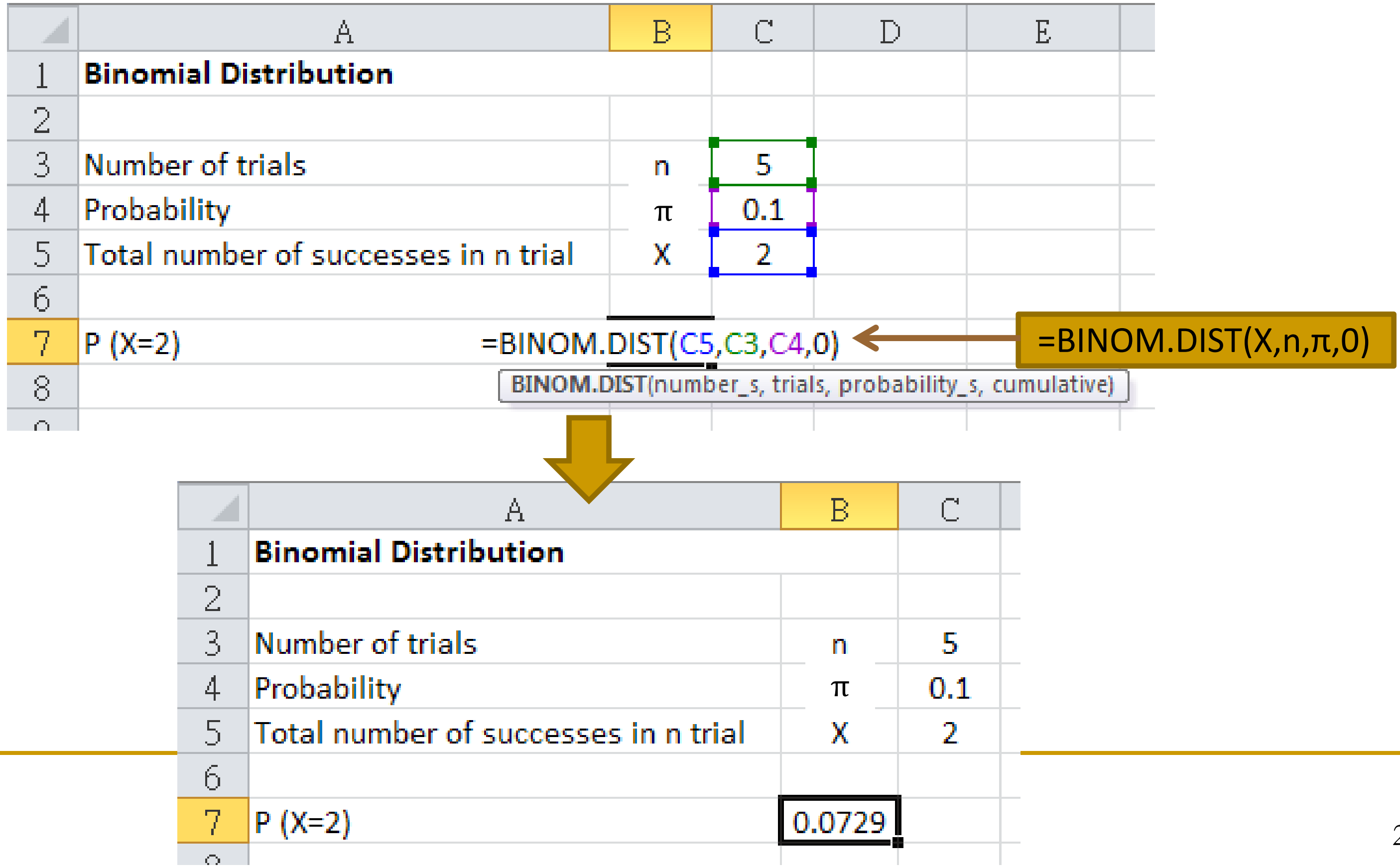


Cumulative:
where 1 is used to calculate P(X≤2)
0 is used to calculate P(X=2)

Binomial Distribution in Excel

Cont'd

- Using Excel to find out the probability that 2 students out of 5 going to cinema with 1 week



Continuous Probability Distributions

- A continuous variable is a variable that can be assume any value on a continuum (can assume an uncountable number of values)
- Examples
  - Time required to travel from home to campus
  - Temperature of a drink
  - Height, in inches
- These can potentially take on any value depending only on the ability to precisely and accurately measure
- In practice, a discrete numerical variable with large range of values is often considered as a continuous variable

Continuous Probability Distributions – Example

Cont'd

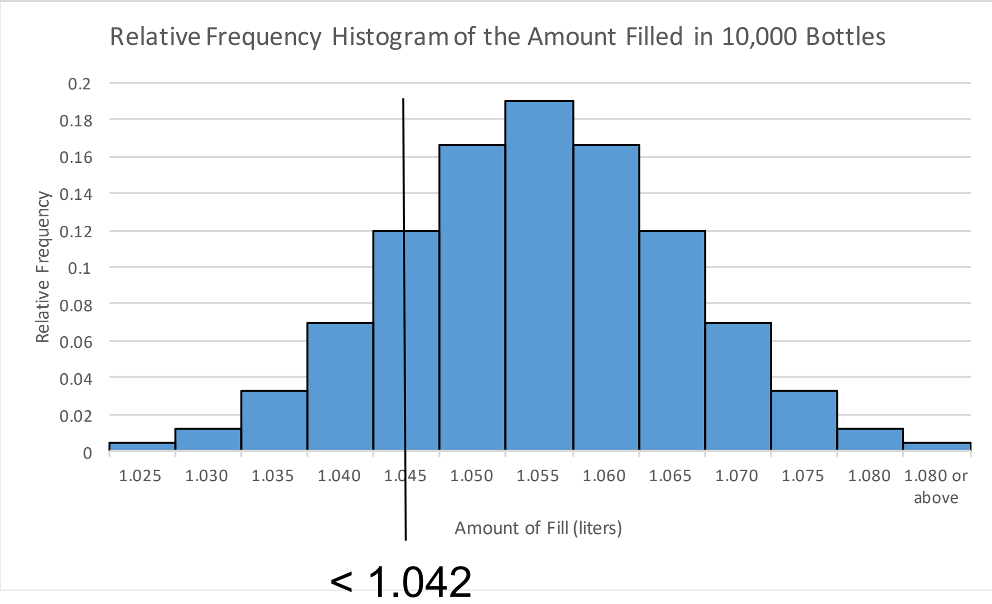
- The probability distribution as shown is obtained by categorizing the amount of soft drink (X) in 10,000 1-liter bottles filled on a recent day
  - $P(X < 1.025) = 0.0048$
  - $P(1.025 \leq X < 1.030) = 0.0122$
  - $P(X < 1.030) = 0.0048 + 0.0122$
    $= 0.0170$
  - $P(X < 1.042) = ?$
  - $P(1.032 \leq X < 1.042) = ?$
  - $P(X = 1.042) = ?$

Amount of Fill (liters)	Frequency	Relative Frequency
< 1.025	48	0.0048
1.025 < 1.030	122	0.0122
1.030 < 1.035	325	0.0325
1.035 < 1.040	695	0.0695
1.040 < 1.045	1198	0.1198
1.045 < 1.050	1664	0.1664
1.050 < 1.055	1896	0.1896
1.055 < 1.060	1664	0.1664
1.060 < 1.065	1198	0.1198
1.065 < 1.070	695	0.0695
1.070 < 1.075	325	0.0325
1.075 < 1.080	122	0.0122
1.080 or above	48	0.0048
		1.0000

Continuous Probability Distributions – Example

Cont'd

- Any required probabilities concerning the amount of soft drink filled can be obtained either from the raw data observed or from the relative frequency histogram



- Treat the relative frequency as the area of of each group

  $P(X < 1.042)$ = the area on the left of 1.042

  $=0.0048 + 0.0122 + 0.0325 + 0.0695 + (1.042-1.040)/0.005*0.1198$

  $=0.1190 + 0.0479 = 0.1669$

- The area under a single point is 0. Hence, $P(X = 1.042) = 0$

Continuous Probability Distributions

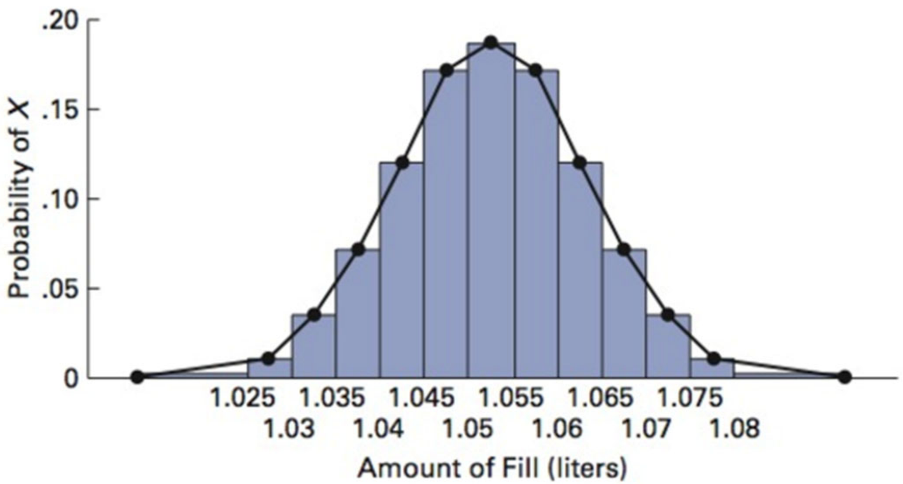
Cont'd

- Computing the probability of a continuous variable over a specified interval is like determining the area of the corresponding relative frequency histogram over the same interval
  - It is time consuming to construct such relative frequency histogram as many data values need to be collected
  - Sometimes it may not be possible to get the required data points
- Is there an alternative way to compute probabilities of a continuous variable without involving the relative frequency histogram?

Continuous Probability Distributions

Cont'd

- The figure below shows the relative frequency histogram and percentage polygon for the distribution of the amount filled in 10,000 bottles
  - Polygon is a graph made by joining the middle top points of each class interval of relative frequency histogram



- Determining the area of a relative frequency histogram over an interval is approximately equivalent to finding the area under the polygon over the same interval
- If we can assume that the polygon actually follows some known mathematical curve, then finding the area under such curve becomes very easy

Continuous Probability Distributions

Cont'd

- Probability Density Function
  - A probability density function, or density function of a continuous random variable is a function that describes the relative likelihood for this random variable to take on a given value
  - One may consider a density function is an approximation to the percentage polygon of its relative frequency distribution
  - A density function, *f* , for a random variable *X* has the following features:
    - $f(x) \geq 0$ for all x of X
    - The area bounded by the curve of $f(x)$ and the x-axis is equal to 1
  - The most important form of density function is called the Normal Density Function

Normal Distribution

Cont'd

- If the density function of a continuous random variable can be best described by normal density function, we say the random variable follows a Normal Distribution
- The Normal Density Function is defined as:

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

where x = any value that the continuous random variable X can take in the range of -∞ to + ∞

μ = mean of the population

σ = standard deviation of the population

e = the mathematical constant approximated by 2.71828…

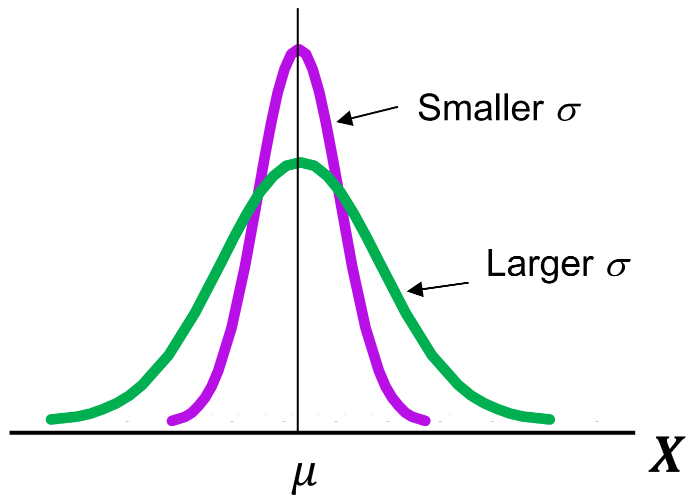
π = the mathematical constant approximated by 3.14159…

- Often denoted as $X \sim N(\mu, \sigma^2)$

Normal Distribution

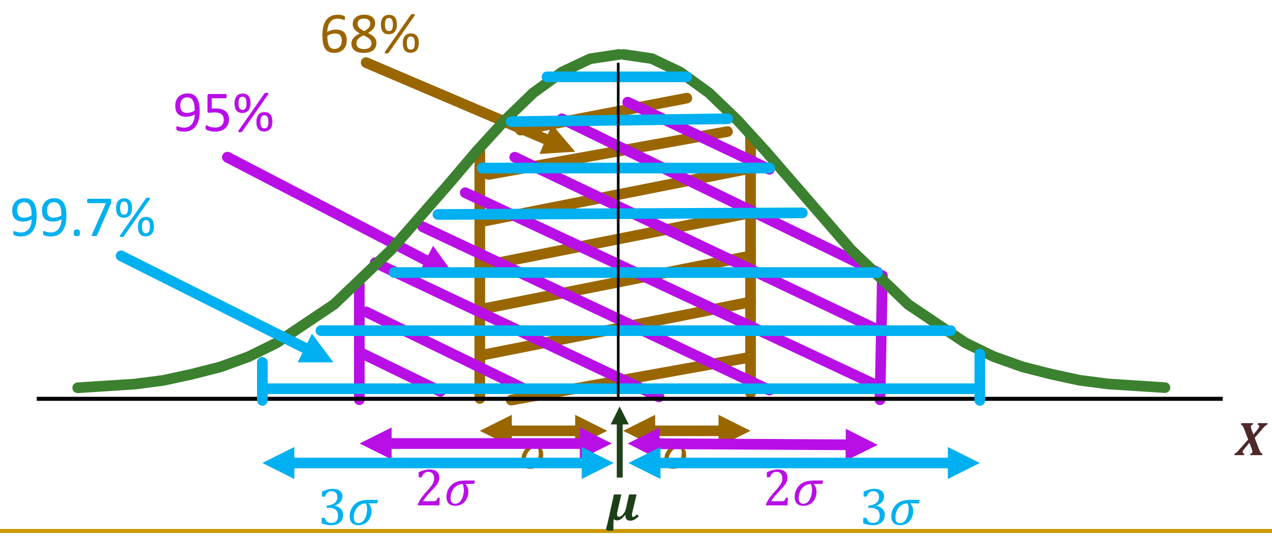
Cont'd

- For $X \sim N(\mu, \sigma^2)$
  - Has an infinite theoretical range, i.e. $-\infty$ to $+\infty$
  - Bell shaped
  - Symmetrical about $X = \mu$
  - Mean, median and mode are identical
  - The spread is determined by σ
    - For smaller σ, the X values are clustered more closely around μ
    - For larger σ, the X values are more spread out and away from μ
  - Follows the Empirical Rule



The Empirical Rule

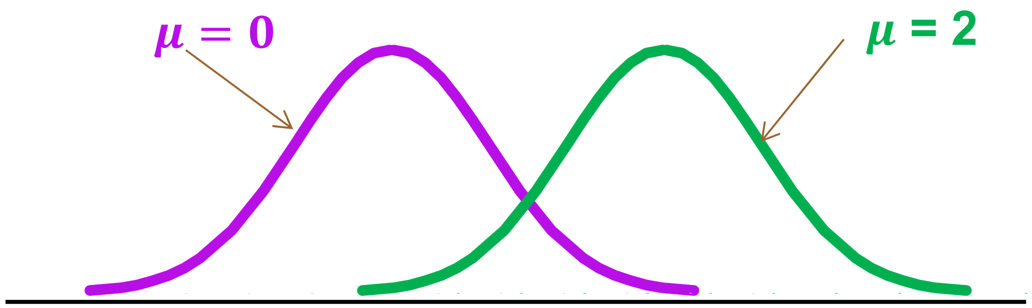
- The Empirical Rule said that
  - Area within $\mu \pm \sigma$ equals 68% approximately
  - Area within $\mu \pm 2\sigma$ equals 95% approximately
  - Area within $\mu \pm 3\sigma$ equals 99.7% approximately



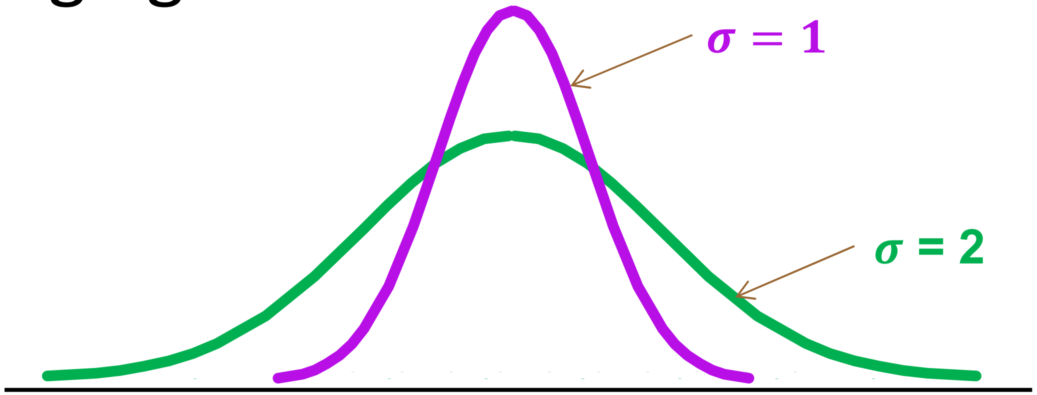
Normal Distribution

Cont'd

- Changing μ shifts the distribution left or right

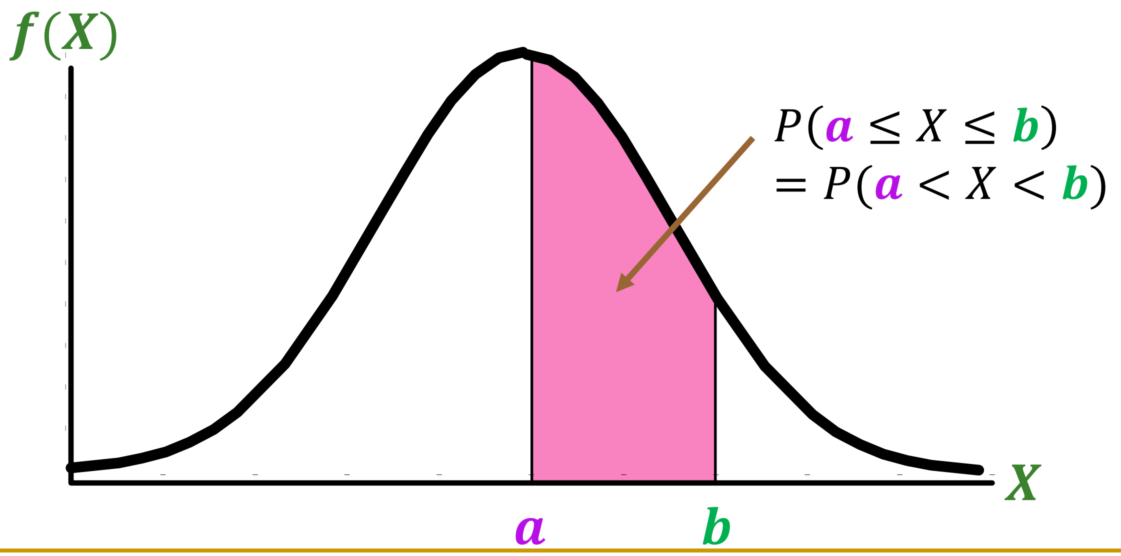


- Changing σ increases or decreases the spread



Computing Normal Probabilities

- The total area under the curve is 1
- Probability is measured by the area under the curve
- Note that the probability of any individual value is zero by definition, i.e. $P(X = a) = 0$

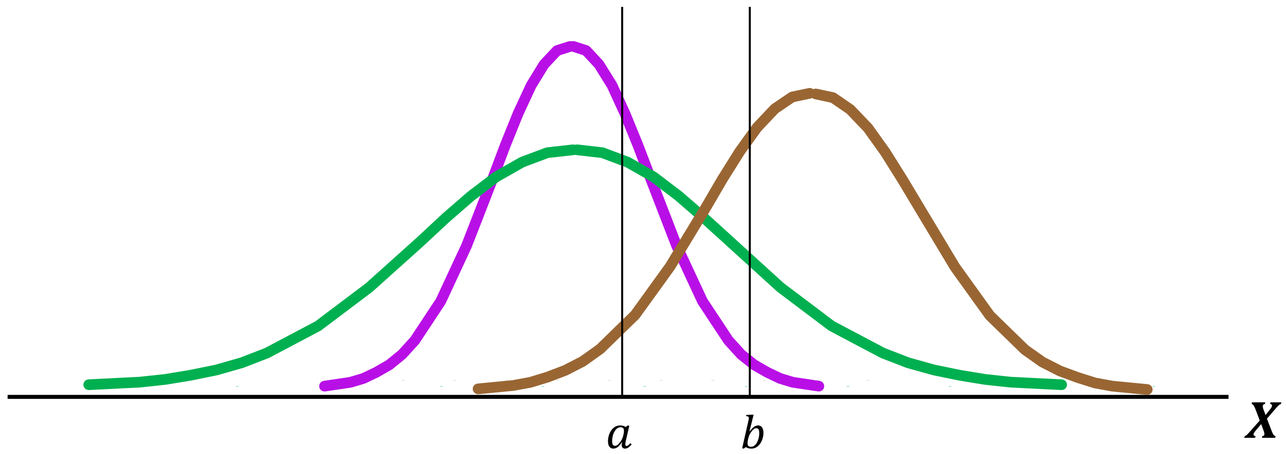


Computing Normal Probabilities

Cont'd

- Area under the curve is computed as

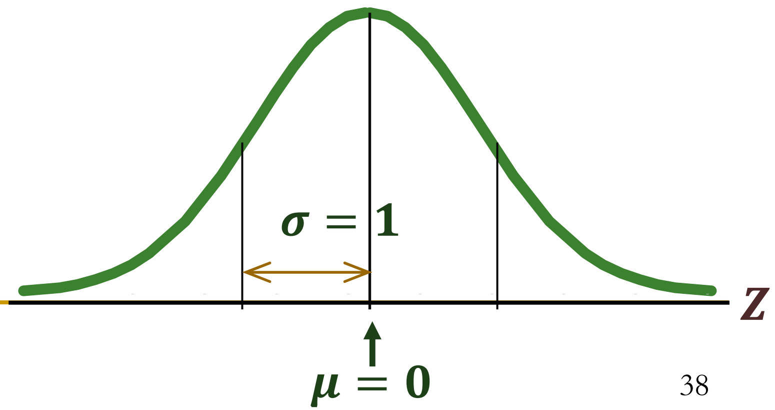
$$P(a \le X \le b) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-\left(\frac{1}{2}\right)\left[\frac{x-\mu}{\sigma}\right]^2} dx$$



- Varying the parameters μ and σ, we obtain different Normal Distributions

The Standardized Normal Distribution

- When a random variable Z follows a Normal Distribution with $\mu = 0$ and $\sigma = 1$, we say Z follows a Standard Normal Distribution
- Often denoted as $Z \sim N(0, 1^2)$
- An advantage of Z distribution is that the probabilities for Z are available on standard normal tables
  - A table gives the probability that Z is between the $-\infty$ and a desired value for Z



The Standardized Normal Distribution

Cont'd

- For any $X \sim N(\mu, \sigma^2)$, it can be standardized to $Z \sim N(0, 1^2)$ with the following formula

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

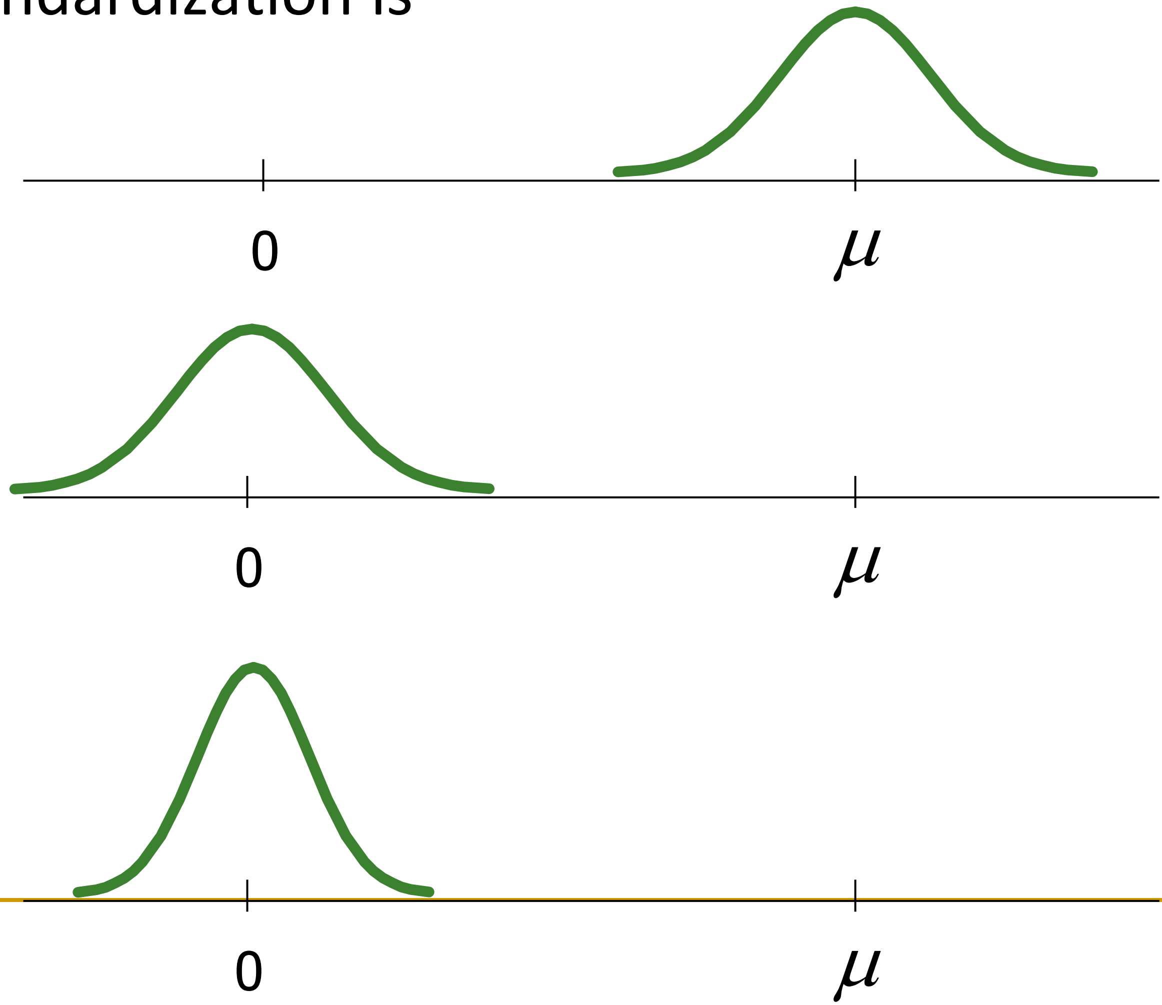
Standardization of Normal Distributions

- The idea of standardization is

Step 1: X

Step 2: $X - \mu$

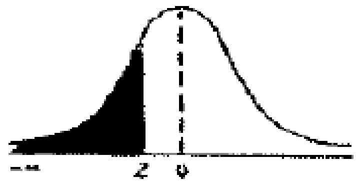
Step 3: $\mathrm{Z} = \frac{X-\mu}{\sigma}$



The Standardized Normal Table

The column gives the value of Z to the second decimal point

The Cumulative Standardized Normal Distribution
Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-6.0	0.000000001									
-5.5	0.000000019									
-5.0	0.000000287									
-4.5	0.000003398									
-4.0	0.000031671									
-3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
-3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
-3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
-3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
-3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
-3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024

The row shows the value of Z to the first decimal point

The value within the table gives the probability from $Z = -\infty$ up to the desired Z value
$P(Z < -3.45) = 0.00028$

Computing Normal Probabilities – Example

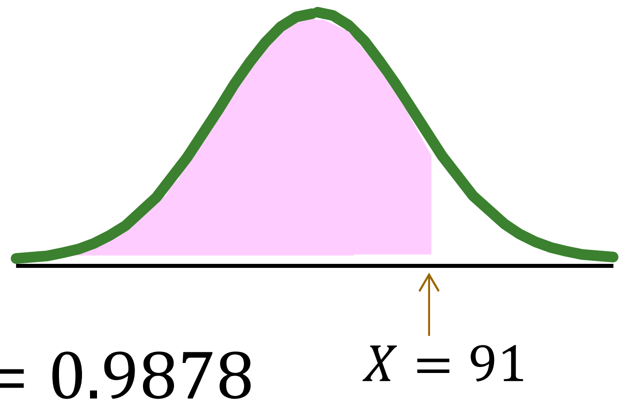
- A set of final exam scores was normally distributed with a population mean 73 and population standard deviation 8

1. What is the probability of getting a score not higher than 91 on this exam?

Let the score be X, and $X \sim N(73, 8^2)$

$$P(X \leq 91)$$

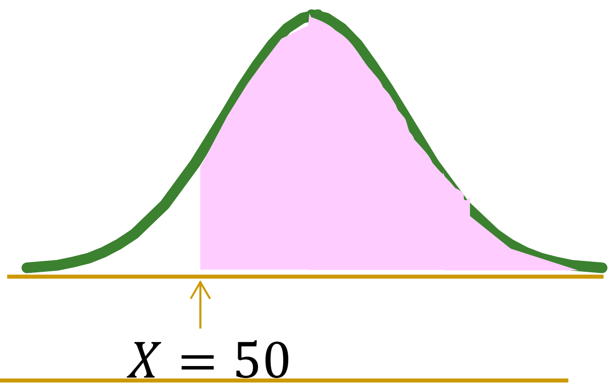
$$= P\left(Z \leq \frac{91-73}{8}\right) = P(Z \leq 2.25) = 0.9878$$



Computing Normal Probabilities – Exercise

Cont'd

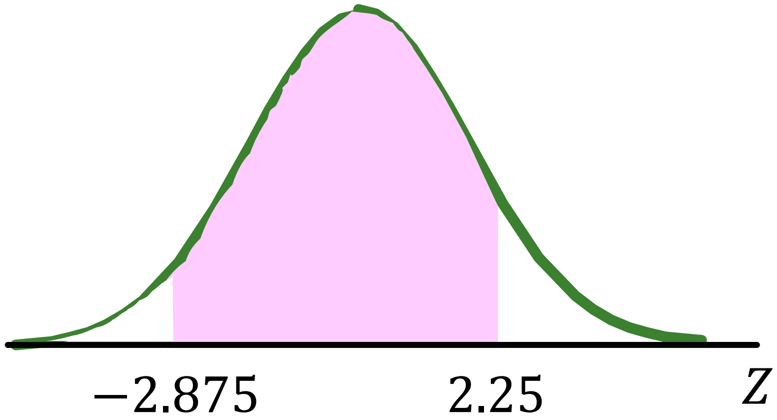
2. If the passing score is 50, what is the chance that a student can pass the exam?



Computing Normal Probabilities – Exercise

Cont'd

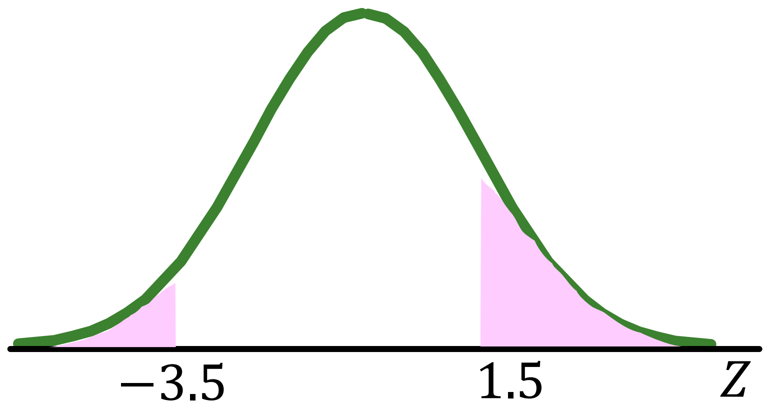
3. What percentage of students scored between 50 and 91?



Computing Normal Probabilities – Exercise

Cont'd

4. What percentage of students scored below 45 or above 85?



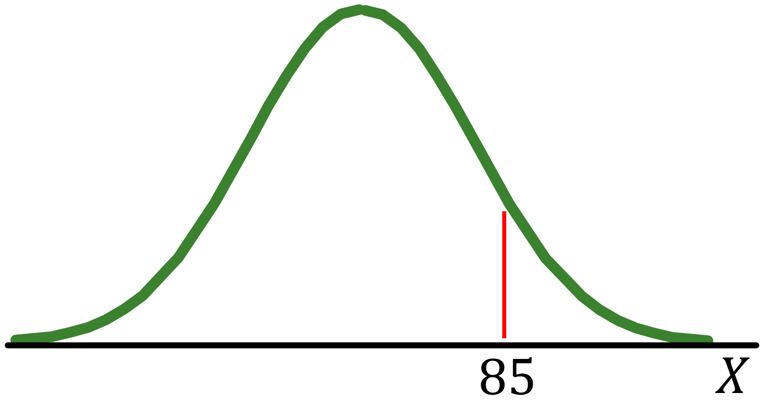
Computing Normal Probabilities – Exercise

Cont'd

5. What is the probability for a student to score exactly 85?

Not an area,

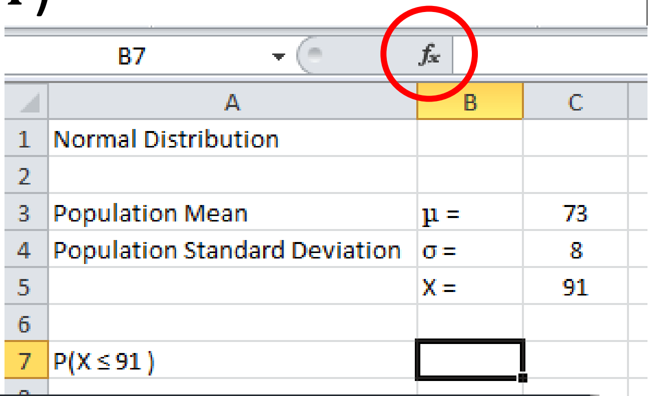
but just a line!!!

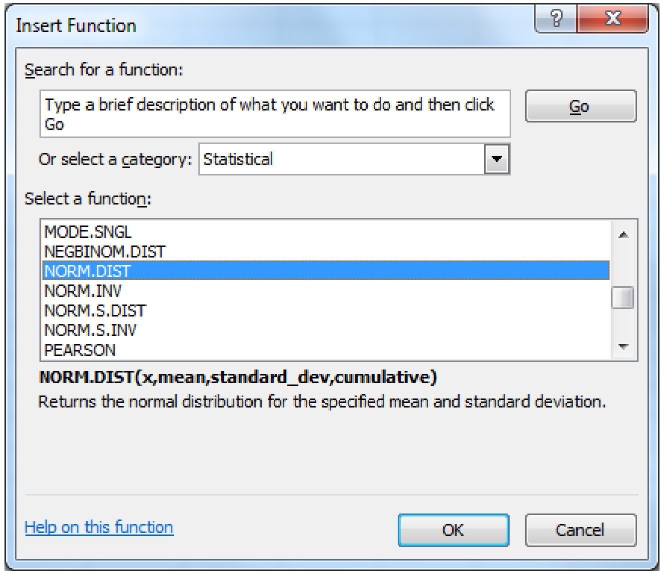


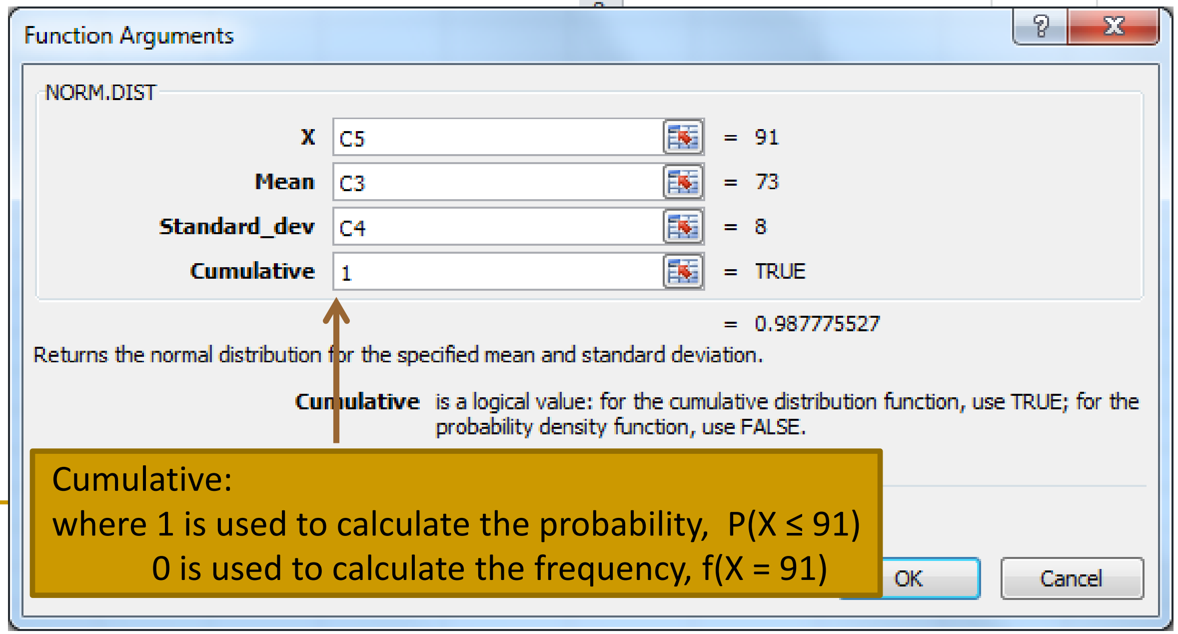
Computing Normal Probabilities in Excel

Step 1: Type the given information (μ, σ, X)

Step 2: Insert the "NORM.DIST" function



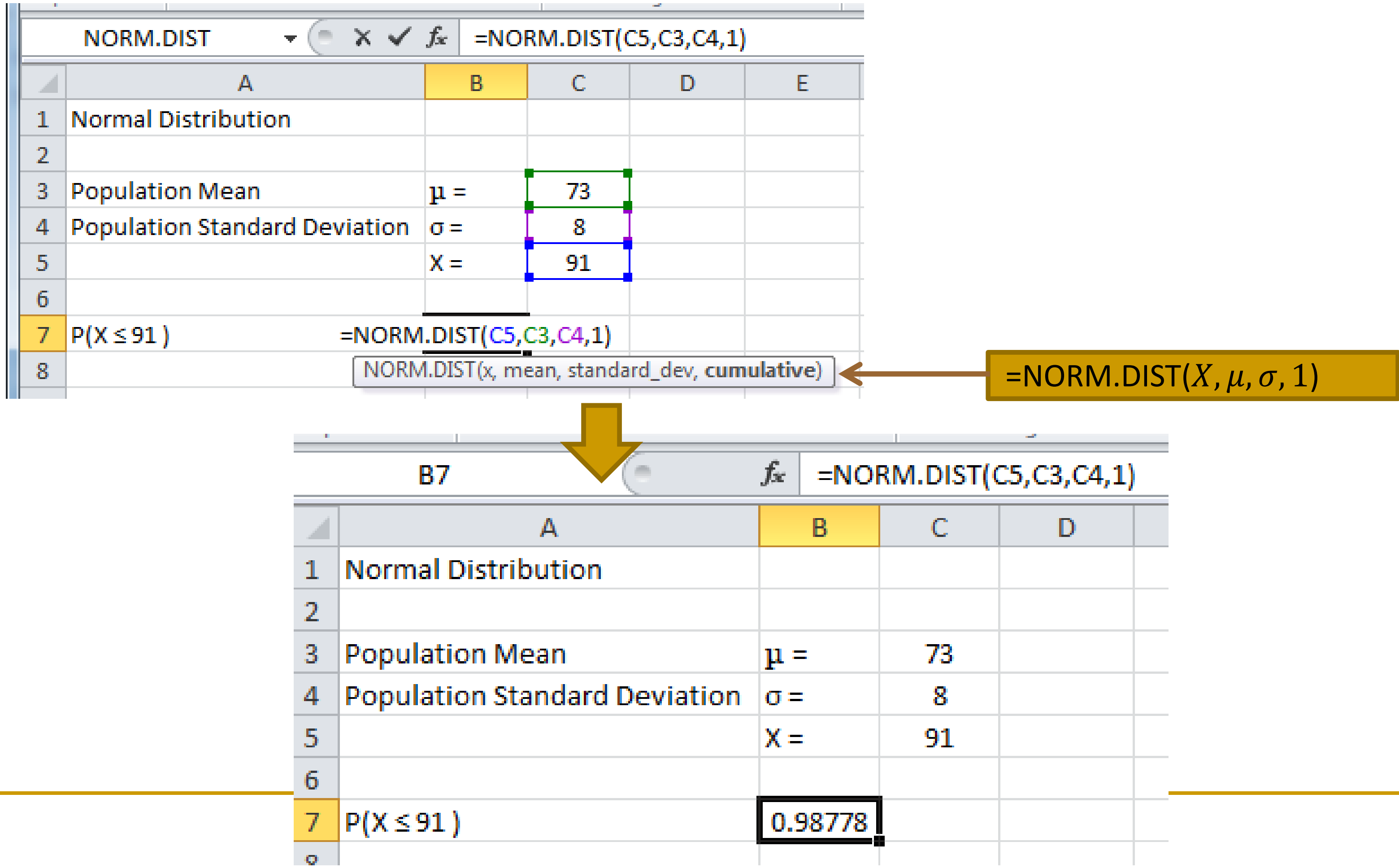




Cumulative:
where 1 is used to calculate the probability, $P(X \leq 91)$
0 is used to calculate the frequency, $f(X = 91)$

Computing Normal Probabilities in Excel

Cont'd



Recovering X Values from Known Probabilities

- With a given (cumulative) probability, we can use the Z table to recover the Z value
- With μ and σ of the X variable, we can recover the X value

Recovering X Values – Example

- Given that the exam scores, X, follow normal distribution with mean 73 and standard deviation 8, i.e. $X \sim N(73, 8^2)$

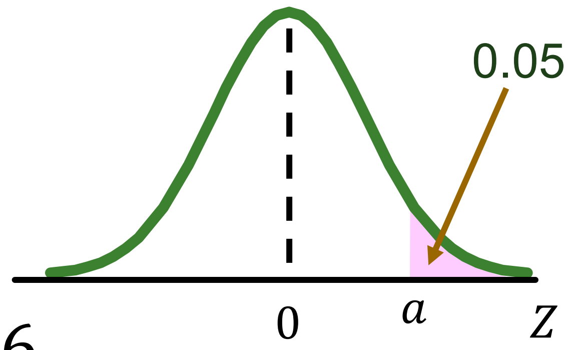
1. What is the minimum score a student needs in order to be in the top 5% of the class?

For $P(Z \geq a) = 0.05$, $a = 1.645$

As $Z = \frac{X-\mu}{\sigma}$,

hence $X = \mu + Z\sigma$

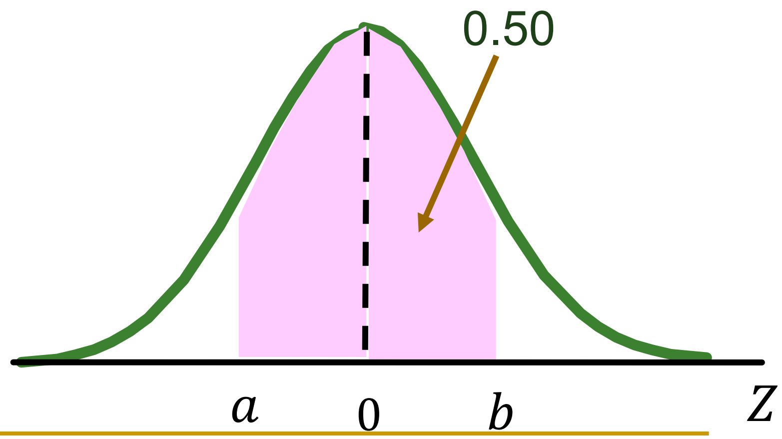
$= 73 + 1.645 \times 8 = 86.16$



Recovering X Values – Exercise

Cont'd

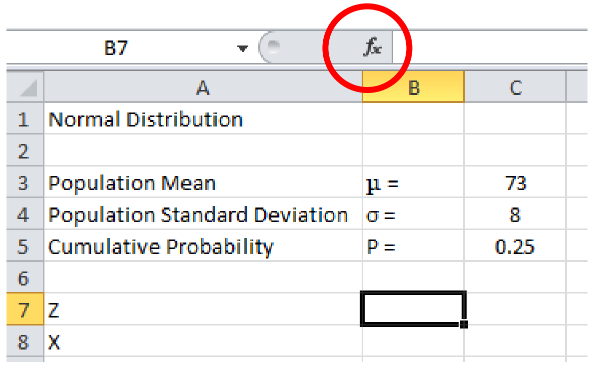
2. The middle 50% of the students scored between what two scores?

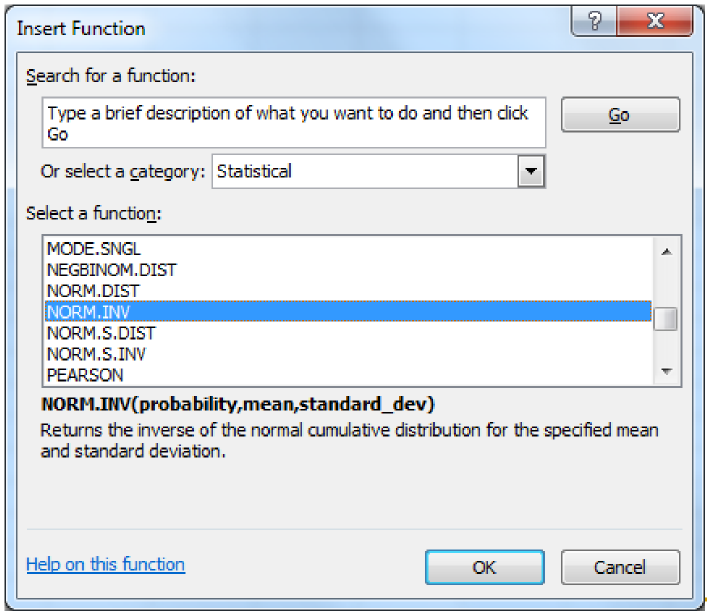


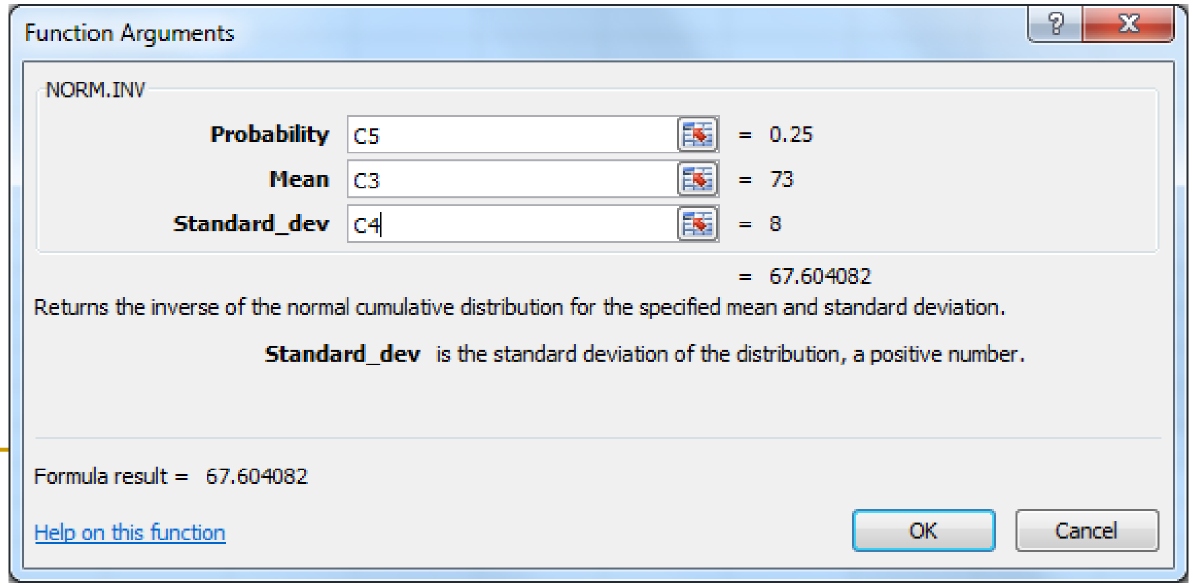
Recovering X Values in Excel

Step 1: Type the given information (μ, σ, P)

Step 2: Insert the "NORM.S.INV" function for recovering Z value; "NORM.INV" for recovering X value







Recovering X Values in Excel

Cont'd

B8 f_x =NORM.INV(C5,C3,C4)

	A	B	C	D
1	Normal Distribution			
2				
3	Population Mean	µ =	73	
4	Population Standard Deviation	σ =	8	
5	Cumulative Probability	P =	0.25	
6				
7	Z	-0.67449		
8	X	67.60408		

=NORM.S.INV(P)

=NORM.INV(P, μ, σ)

Importance of Normal Distribution

- Most common continuous distribution used in statistics
- Provides the basis for statistical inference because of its relationship to the Central Limit Theorem
  - To be discussed in Topic 4
- Can be used to approximate various discrete probability distributions, such as Binomial distribution, for large sample size, therefore, simplifying computations
  - To be discussed in Topic 7

You Sometimes Get More Than You Pay For

- According to McDonald's "fact sheet", their ice cream cones weigh 3.7 ounces and contain 170 calories
- Do the ice cream cones really weigh exactly 3.7 ounces?
- To get 3.7 ounces for every cone would require a very fine-tuned machine, or an employee with a very good skill and sense of timing
- Thus, we expect some natural variation in the weight of these cones

You Sometimes Get More Than You Pay For

Cont'd

- If you buy one ice cream cone everyday through out the week, you may be surprised that they all weighed more than 3.7 ounces
- How likely this would happened?