STATS 101/108 WORKSHOP

TEST PREP 2: CHAPTERS 4 AND 6

WEDNESDAY 8 SEPTEMBER, 2010

get a handout from the back!

Students **MUST REGISTER** for all workshops with
The Student Learning Centre, 3rd Floor, Information Commons

Revision Notes

Chapter 4 - Probability

Look at blue pages for extra test/exam questions for practice

- A **probability** is a number between 0 and 1 that quantifies uncertainty.
- There are two main sources of probabilities that we will deal with.
  1. Probabilities using a model – some models that may involve equally likely outcomes are *tossing a coin* and *rolling a die*
  2. Probabilities from data
- A **random experiment** is an experiment where the outcome cannot be predicted.
- A **sample space** is the collection of all possible outcomes.
- An **event** is a collection of outcomes. An event **occurs** if any outcome making up that event occurs.
- If all **outcomes** are **equally likely:** $\text{pr}(A) = \dfrac{\text{no. of outcomes in } A}{\text{total no. of outcomes}}$
- The **complement** of an event A, denoted $\overline{A}$, occurs if A does not occur. A, and $\overline{A}$ are **mutually exclusive** events, ie they CANNOT occur at the same time.
- General probability rules:
  1. $\text{pr}(S) = 1$
  2. $\text{pr}(\overline{A}) = 1 - \text{pr}(A)$
- **Statistical Independence** – two events (A & B) are statistically independent if knowing whether B has occurred gives no new information about the chances of A occurring.

  i.e. $\text{pr}(A|B) = \text{pr}(A)$ and $\text{pr}(A \text{ and } B) = \text{pr}(A) \times \text{pr}(B)$
- **Two Types of Test/Exam Questions**
  1. Given a table of numbers/proportions, find the probability:
     - Easier question/s (can be between 1 and 3 of this type).
     - ~~May want to convert the table into table of probabilities first.~~

2. Given a short story with proportions, percentages and/or counts about two factors (qualitative variables), find the probability:

- ☞ Harder question/s (can be 1 or 2 of this type).
- ☞ Need to *interpret* the story first, and then construct a table.
- ☞ Use the table to find 1 or 2 probabilities.
- ☞ Steps to constructing a table:

  Step 1: highlight numbers

  Step 2: highlight factors

  Step 3: define factor levels

  Step 4: label table

  Step 5: enter appropriate table total

  Step 6: enter row/column totals from story

  Step 7: enter cell numbers from story

  Step 8: enter remaining numbers by +/-

- **Four Types of Probability Calculations**

1. Probability of <u>AN EVENT</u> (basic/simple)

- ☞ pr(A) → pr (an event)

example 1

rowls totalls ÷ GT

or columnls totalls ÷ GT

2. Probability of an event <u>AND</u> another event:

example 4

- ☞ pr(A and B) → pr(one event and another event)
- ☞ *Finding* pr(A) *<u>and</u>* pr(B) <u>(intersection)</u>

Use GT ***GRAND TOTAL (TABLE TOTAL)***

cells ÷ GT

3. Probability of an event <u>OR</u> another event:

example 3

- ☞ pr(A or B) → pr(one event or another event)
- ☞ *<u>Add</u>* pr(A) *to* pr(B), then *<u>subtract</u>* pr(A and B)

[rowls totalls + columnls totalls - cells] ÷ GT

4. ***CONDITIONAL* Probability:** example 2

☞ Harder to detect but will usually have one of the key words:

- "Given that..."
- "Of those..."
- "Among those..."

} Use ***ROW TOTAL/S*** *OR* ***COLUMN TOTAL/S***

- Look for language that restricts you to part of the table instead of the whole table.

may happen — must happen

$$\text{pr (one event | another event)} = \text{pr}(A|B) = \frac{\text{pr}(A \text{ and } B)}{\text{pr}(B)}$$

"given that"

The following table, taken from *Facts New Zealand*, 1993, is used in **Examples 1** to **4**. It shows household income according to family type (A = Solo parent with one or more children, B = Couple with one or more children, C = Couple with no children, D = Single person, E = Other).

Household Income	**Family Type** A	B	C	D	E	**Total**
	No. of households (×1000)					
0-9999	29	13	5	57	2	86
10,000 - 19,999	43	26	61	92	8	230
20,000 - 39,999	36	107	93	63	40	339
40,000 - 74,999	13	153	87	17	47	317
75,000 +	1	61	25	3	19	109
Total	102	360	271	232	116	1081

Example 1: If one of the 1,081,000 families is chosen at random, the probability that the total family income is less than $20,000 is:

$$\text{pr}(\text{< \$20K}) = \frac{(86+230)}{1081}$$

$$= 0.2923 \quad \text{(4dp)}$$

Example 2: Given that the family is of type A [Solo parent with child(ren)], the probability that the total family income is less than $20,000 is:

$$\text{pr}(\text{< \$20K} \mid \text{A}) = \frac{(9+43)}{102}$$

$$= 0.5098 \quad \text{(4dp)}$$

Example 3: The probability that a randomly chosen family's total income is less than $20,000 or the family is of type A is:

$$\text{pr}(\text{< \$20K or A}) = \frac{(86+230+102-9-43)}{1081}$$

$$= 0.3386 \quad \text{(4dp)}$$

Example 4: The probability that a randomly chosen family is type A and their total income is less than $20,000 is:

$$\text{pr}(\text{A \& < \$20K}) = \frac{(9+43)}{1081}$$

$$= 0.0481 \quad \text{(4dp)}$$

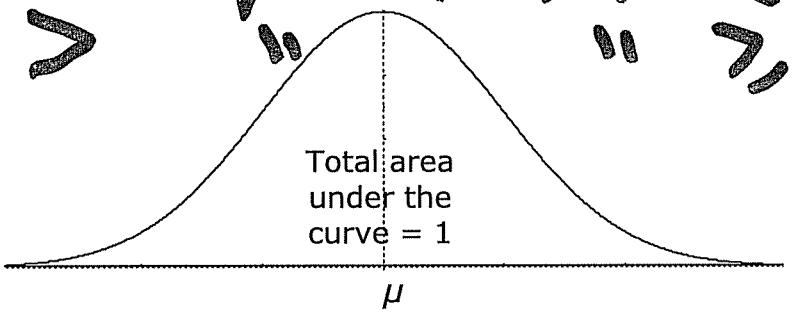
Chapter 6 – Continuous Random Variables

Look at blue pages for good notes and test/exam questions for practice

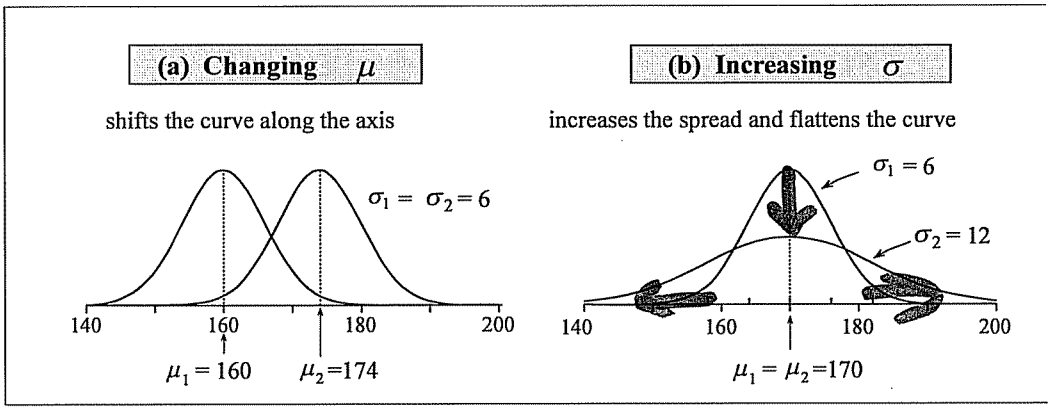
- A density curve is the probability distribution of a continuous random variable.
- There are no gaps between the values that a continuous random variable can take and therefore, when we calculate probabilities for a continuous random variable it does not matter whether **interval endpoints** are included or excluded

Normal Distribution

- The Normal Distribution has a probability density function curve, which is smooth, **bell-shaped**, and **symmetric**.



- The shape of the curve is solely determined by the parameters μ (mean) and σ (standard deviation).



- The Normal distribution is important because it:
  - fits a lot of data particularly well
  - can be used to approximate other distributions
  - is very important in statistical inference
- If X is a continuous random variable from a Normal distribution then:
  - $E(X) = \mu$ and $sd(X) = \sigma$
  - Probability distribution function of X is written: $X \sim \text{Normal}(\mu, \sigma)$

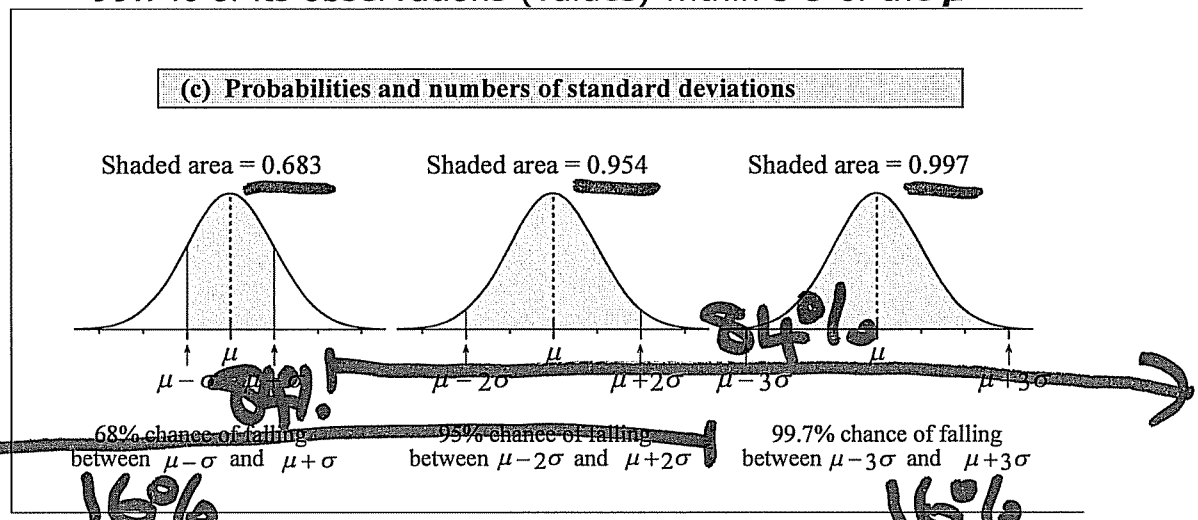
Chapter 6 test/exam questions

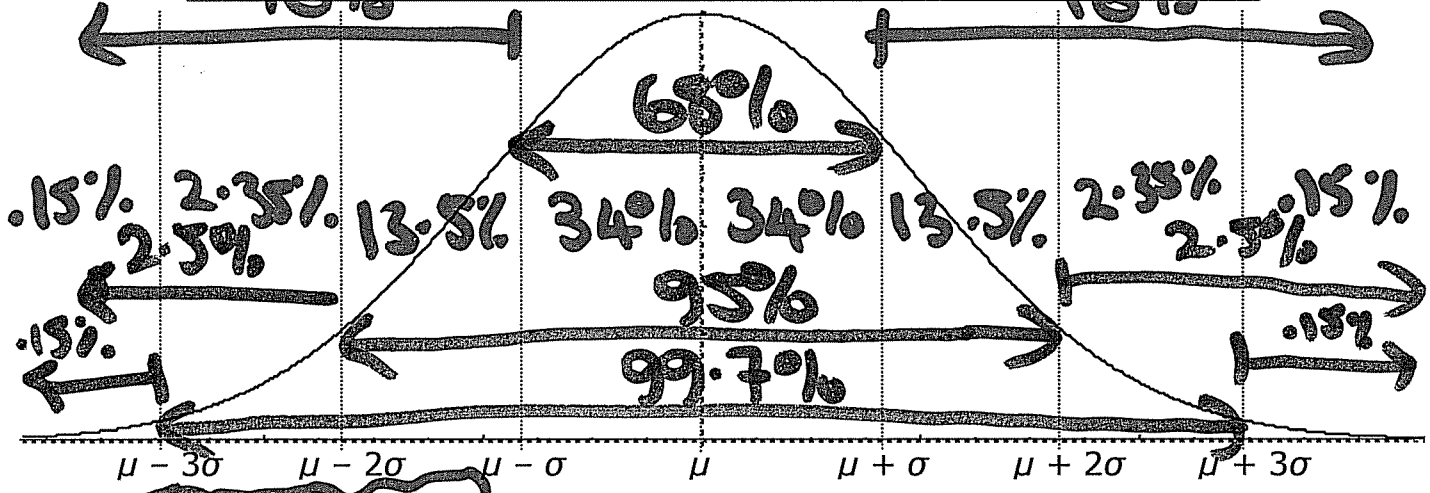
When doing Chapter 6 problems, it is sensible to draw a Normal curve and then mark on it what is known and what is unknown. There are **three (3)** types of Chapter 6 test/exam questions:

1. *True/False (Normal) Chapter 6 problem*

There will be five statements, each about one or the other or both of two different Normal distributions. Use the 68-95-99.7% rule or *z*-scores to determine whether four of the statements are true or false. The fifth statement will probably be comparing the means (centres/averages) and standard deviations (spread/variability) of the two distributions.

- **68-95-99.7% rule**: A population with a Normal distribution has:
  - ✓ 68% of its observations (values) within 1 σ of the μ
  - ✓ 95% of its observations (values) within 2 σ of the μ
  - ✓ 99.7% of its observations (values) within 3 σ of the μ



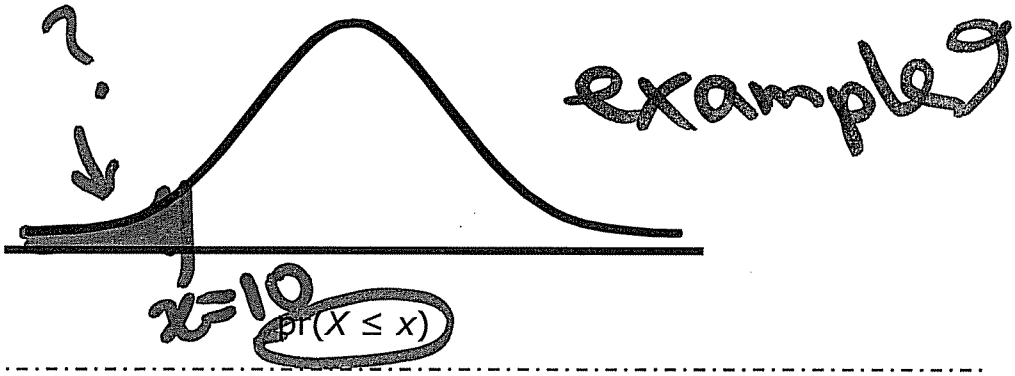


The **z-score**, $z = \frac{x - \mu}{\sigma}$, is a standardised number. It represents the number of standard deviations, σ, the value of x is away from the mean, μ. We can use *z*-scores to compare two or more different Normal distributions.

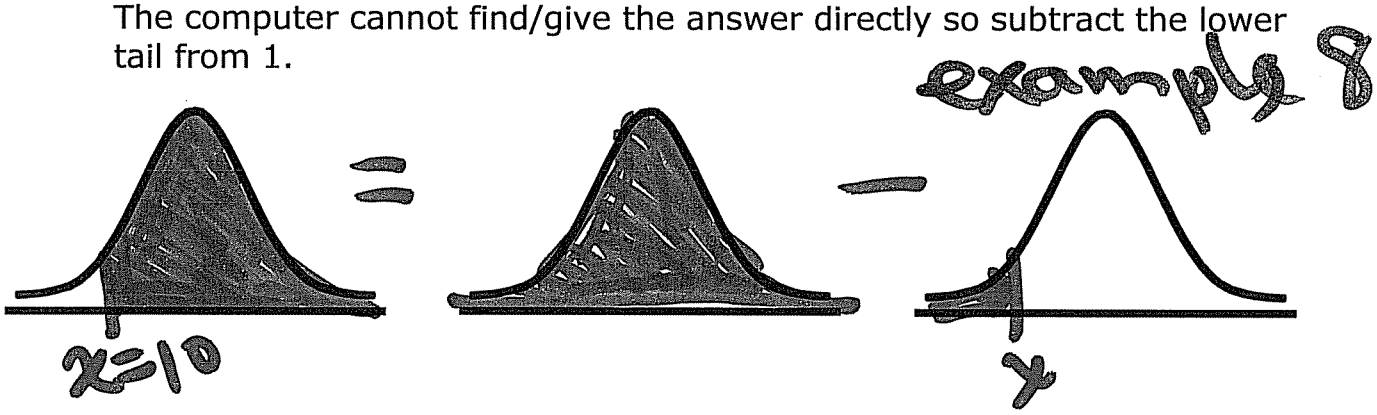
2. *Normal probability problem, i.e. find a probability associated with a number*

When finding a probability, shade the desired area under the curve and then devise a way to obtain it using lower tail probabilities which is all the computer can give. There are three types of Normal probability problems:

- *Find a lower tail probability (area)*
  The computer can find/give the answer directly.

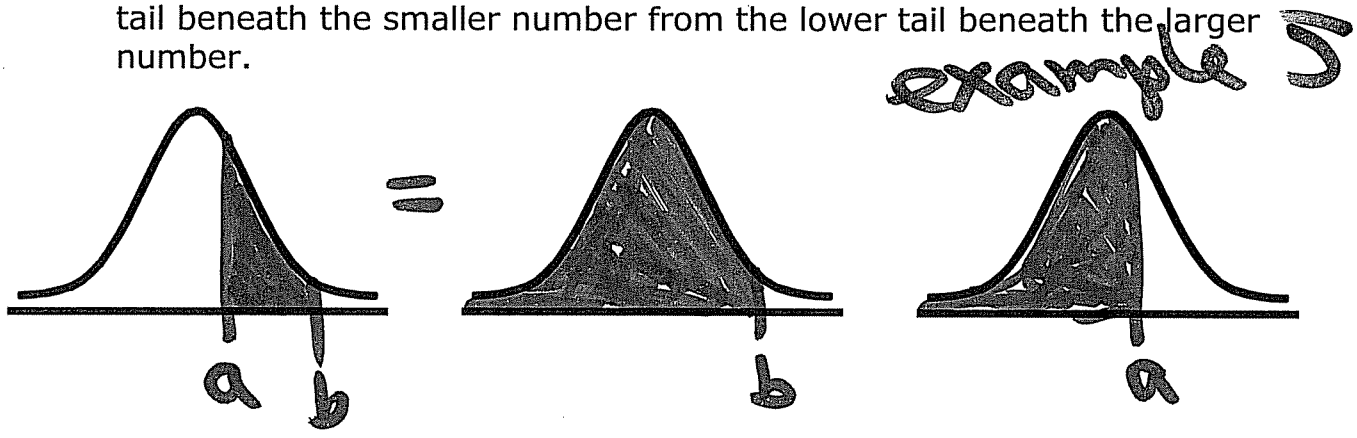


- *Find an upper tail probability (area)*
  The computer cannot find/give the answer directly so subtract the lower tail from 1.



pr($X \geq x$) = 1 − pr($X \leq x$)

- *Find a probability (area) between two numbers*
  The computer cannot find/give the answer directly so subtract the lower tail beneath the smaller number from the lower tail beneath the larger number.



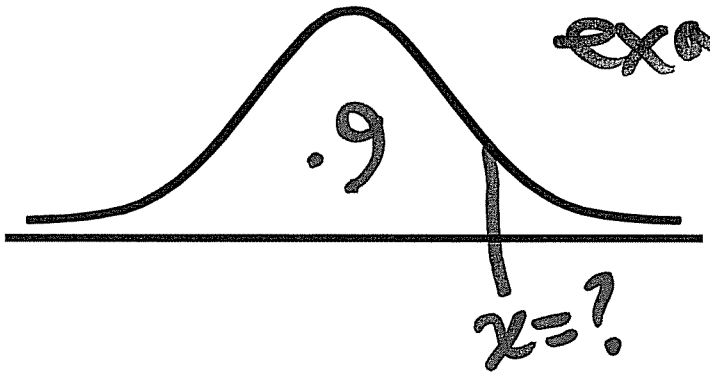
pr($a \leq X \leq b$) = pr($X \leq b$) − pr($X \leq a$)

3. *Inverse Normal problem, i.e. find a number associated with a probability*

This type of problem occurs when we know the probability (e.g. the highest 10% in the class) and we need to find out the number associated with it, x (e.g. the mark). There are three types of inverse Normal problems:

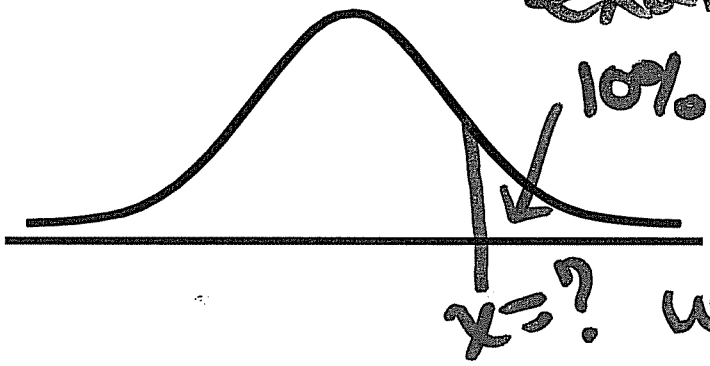
- *Given a lower tail probability, find the number associated with it*
  The computer can find/give the answer directly.

example 7



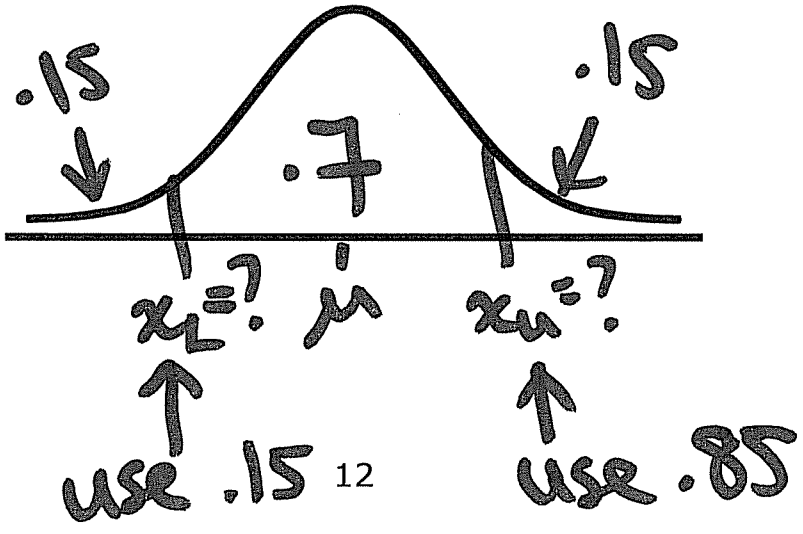
- *Given an upper tail probability, find the number associated with it*
  The computer cannot find/give the answer directly so subtract the upper tail probability from 1 & use the lower tail probability to find the answer.

example 10



- *Given a central area/probability, find the two numbers associated with it (the lower limit and the upper limit)*
  The computer can give the two limits as long as you use the lower tails/areas beneath each of them.

example 8



Examples 5 to **10** are about the following information.

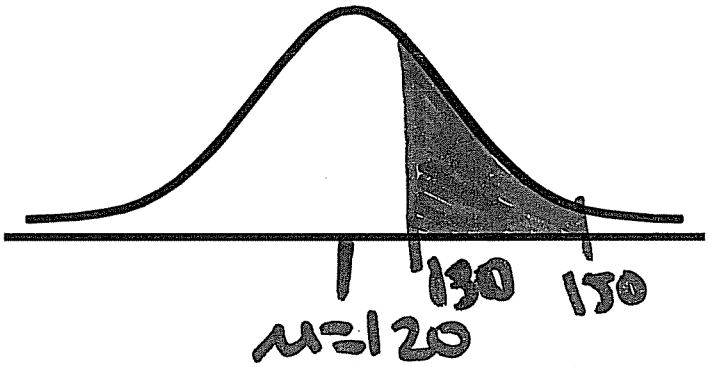
The systolic blood pressure of 18-year-old women is Normally distributed with a mean of 120mm Hg and a standard deviation of 12mm Hg.

Normal with mean = 120.000 and standard deviation = 12.0000

x	Pr(X ≤ x)
110	0.202
120	0.500
125	0.662
130	0.798
150	0.994

Pr(X ≤ x)	x
0.050	100
0.250	112
0.350	115
0.500	120
0.650	125
0.750	128
0.950	140

Example 5: The proportion of 18-year-old women with a systolic blood pressure between 130 and 150 is: [The table given is for a Normal(120, 12) distribution].

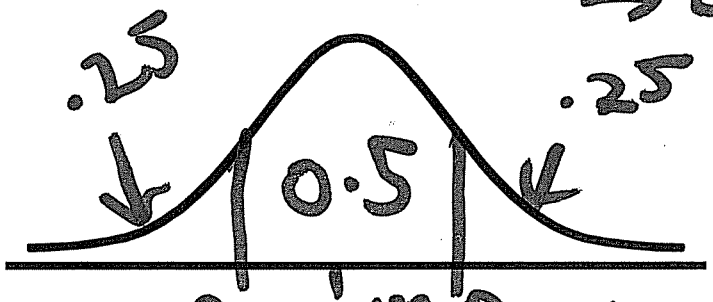


.994 − .798

= .196

Example 6: The interquartile range of systolic blood pressure of 18-year-old women is:

↳ central 50%!



use .75 → 128

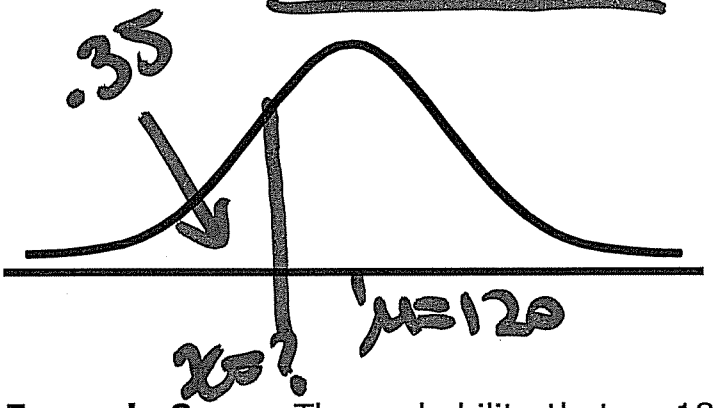
use .25 → 112

IQR = $Q_3 - Q_1$

= 128 − 112

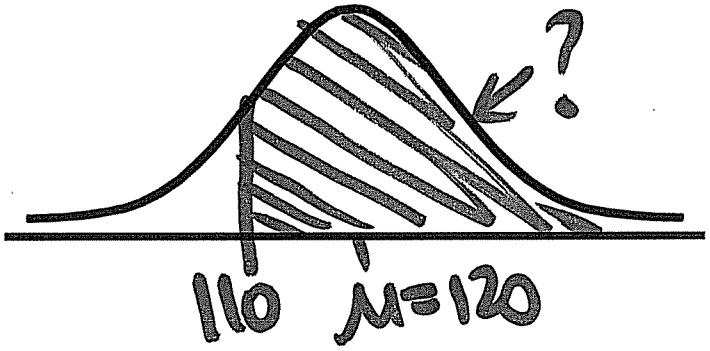
= 16

Example 7: 35% of the population of 18-year-old women would have a systolic blood pressure that is no more than:



$x = 115$

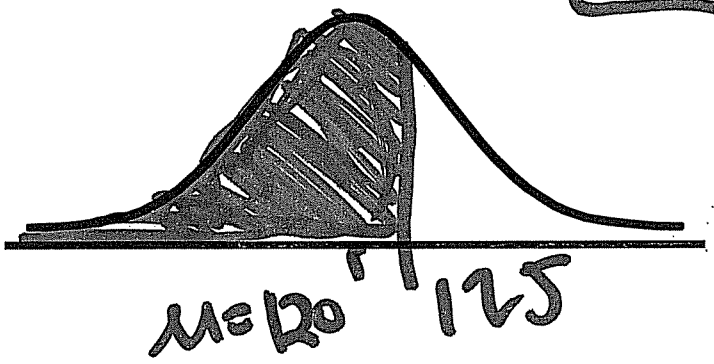
Example 8: The probability that an 18-year-old women's systolic blood pressure exceeds 110 is:



$1 - .202$

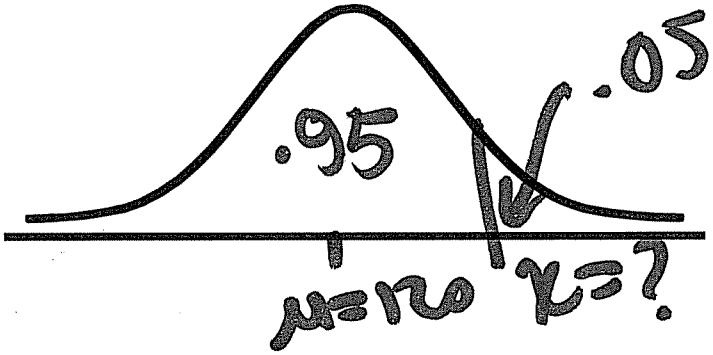
$= .798$

Example 9: The probability that an 18-year-old women with a systolic blood pressure of at most 125 is:



.662

Example 10: The systolic blood pressure that 5% of the population of 18-year-old women would exceed is:



$\mu = .95$

$x = 140$

8. In a STATS 10x/191 assignment, students were asked 'Do you have access to a computer at home?'. In total 922 students responded to this question, 305 of whom were **108** students, and 143 of whom were **191** students. A total of 708 students responded 'Yes', while 116 of the **101** students and 69 of the **108** students responded 'No'. The probability of a randomly chosen student responding 'Yes', given they are in **191**, is

(1) 0.161
(2) 0.124
(3) 0.797
(4) not possible to calculate.
(5) 0.767

	101	108	191	
Yes	358	236	114	708
No	116	69	29	214
	474	305	143	922

$Pr(Y|191) = \frac{114}{143} = .797$

9. In a *Listener* (Feb. 4, 1995) article, it was reported from a survey of people 15 years and over that 83% approved of abortion if the mother's health was at risk. Thirty-six percent of people surveyed were in the age group 25-39 years. Of these, 85% approved of abortion if the mother's health was at risk.

What is the probability that a person selected at random from the age group 15 and over approved of abortion if the mother's health was at risk or the person was in the 25-39 age group?

(1) 0.340
(2) 0.884
(3) 0.485
(4) 0.299
(5) 0.891

	A	$\bar{A}$	
25-39	$\frac{85}{100} \times 3{,}600 = 3060$	540	$\frac{36}{100} \times 10{,}000 = 3{,}600$
15-24 & 40+	5240	1160	6,400
	$\frac{83}{100} \times 10{,}000 = 8{,}300$	1700	10,000

$Pr(A \text{ or } 25\text{-}39) = \frac{(830 + 3600 - 3060)}{10{,}000} = .884$

10. Auditors developing systems to check the accuracy of regular tax returns for such taxes as GST, look at the changes in a firm's returns between tax periods. If the change is greater than some threshold, the firm's return is subjected to a rigorous audit. Such systems designed to detect cases of tax evasion must face the problem of false positives, that is, that the system indicates that the return is suspicious when, in fact, the change represents a real alteration in business conditions rather than tax evasions.

Let E be the event that the firm is really attempting to evade tax, and T be the event that the system indicates possible tax evasion. Experience indicates that the incidence of tax evasion is 1 in 100 firms, while 90% of all cases of tax evasion are detected. Of those firms that are not really attempting to evade tax, the system indicates that 5% are possible tax evaders. The probability that a firm has actually evaded tax given that the system indicates evasion is:

(1) 0.009
(2) 0.050
(3) 0.154
(4) 0.900
(5) 1.000

	E	$\bar{E}$	
T	$\frac{90}{100} \times 100 = 90$	$\frac{5}{100} \times 9900 = 495$	585
$\bar{T}$	10	9405	9415
	$\frac{1}{100} \times 10000 = 100$	9900	10,000

$Pr(E|T) = \frac{90}{585} = .154$

11. The probability of having a positive ELISA test given that you have HIV is 0.95. The probability of having a positive ELISA test given that you don't have HIV is 0.05. The probability of having HIV is 0.004. The probability of having HIV given a positive ELISA test is:

(1) 0.076
(2) 0.886
(3) 0.071
(4) 0.05
(5) 0.95

	HIV+	HIV−	
+ve test	$.95 \times 40 = 38$	$.05 \times 9960 = 498$	536
−ve	2	9462	9464
	$.004 \times 10,000 = 40$	9960	10,000

$Pr(HIV+|+ve) = 38/536 = .071$