

# PSY30100-03 -- Assignment 1

---

## Chapter 1: Look at Data - Distributions

TA: Laura Lu

January 25, 2010

---

# Question 1.17 (p.23)

---

a) Add the counts and compute the percents for each age group.

Ans:  $10+97+70+36+14+5=232$

1-18	19-24	25-35	36-50	51-69	$\geq 70$
10/232	97/232	70/232	36/232	14/232	5/232
$\approx 4.31\%$	$\approx 41.48\%$	$\approx 30.17\%$	$\approx 15.52\%$	$\approx 6.03\%$	$\approx 2.16\%$

---

# Question 1.17 (p.23)

---

b) make a bar graph of the percent.

- Categorical variables
  - Bars do not touch
  - X-axis (Labels!)
  - Y-axis (Labels!)
-

# Question 1.17 (p.23)

---

c) Describe the distribution.

Ans: 1. right skewed (positive skewed)  
2. Unimodal

---

## Question 1.17 (p.23)

---

d) Explain why your bar graph is not a histogram.

Ans: The age group classes do not have equal width.

---

## Question 1.30 (p.26)

---

- a) Why do you think we choose to measure emissions per person rather than total  $CO_2$  emissions for each country?

Ans: Totals emissions would almost certainly be higher for very large countries. For example, we would expect that even with great attempts to control emissions, China would have higher total emissions than the smallest countries in the data set.

---

## Question 1.30 (p.26)

---

b) Display the data of Table 1.6 in a graph.

Ans: many ways.

1. modified box plot
  2. histogram
  3. stemplot
-

# Modified box plot

---

- ❑ To detect potential outliers
  - ❑ Interquartile range (IQR)
  - ❑ The  $1.5 \times \text{IQR}$  criterion for outliers
-



# Modified box plot

---

- Sort (total 48)

0.0 0.0 0.1 0.1 0.2 0.2 0.2 0.2 0.3 0.3 0.5 0.7 0.8 0.9 0.9 1.0 1.2 1.4  
1.7 1.8 2.3 2.5 2.5 2.8 3.6 3.7 3.8 3.9 3.9 4.6 4.8 5.1 6.1 6.8 7.3 7.6  
8.0 8.1 8.8 9.0 9.1 9.7 10 10.2 11 16.0 17.0 19.9

- Median= $(2.8+3.6)/2=3.2$ ,  
Q1= $(0.7+0.8)/2=0.75$   
Q3= $(7.6+8.0)/2=7.8$

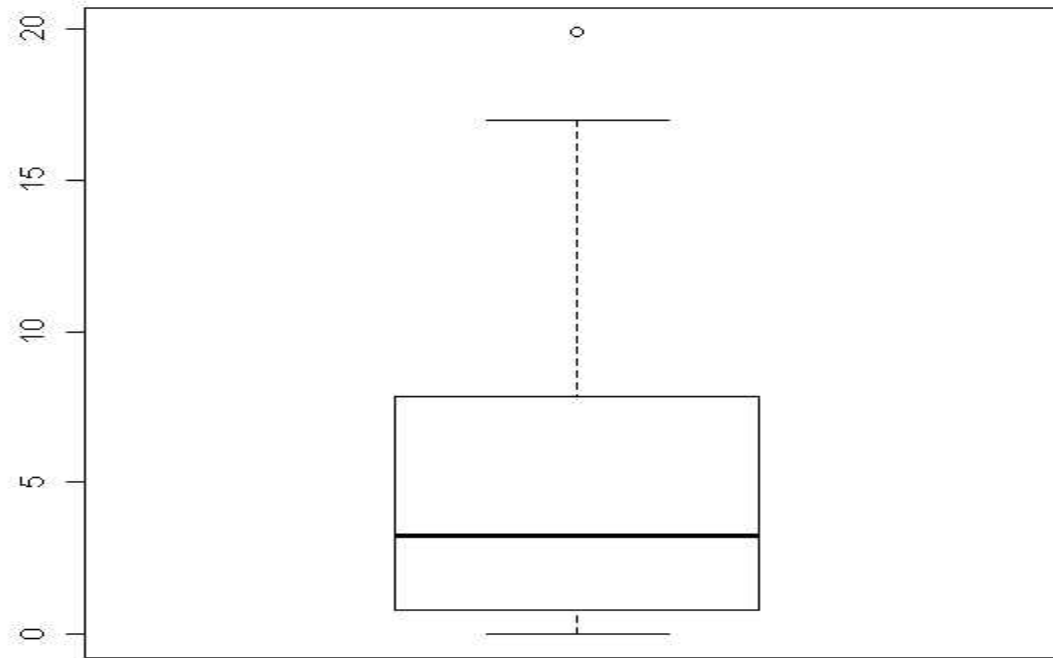
- Interquartile range (IQR)  
IQR=Q3-Q1= $7.8-0.75=7.05$

- The 1.5 x IQR criterion for outliers  
 $1.5 \times \text{IQR} = 1.5 \times 7.05 = 10.575$   
 $Q3 + 1.5 \times \text{IQR} = 7.8 + 10.575 = 18.375 < 19.9$

- Largest non-outlier observation: 17.0  
Smallest non-outlier observation: 0.0
-

# Modified box plot

---



## Question 1.30 (p.26)

---

- b) shape: right skewed
    - center: mean=4.6; median=3.2
    - outliers: United States
    - spread: 0-19.9 (range=19.9, may not be the best measure of spread)
      - or 0-17 (apart from an outlier)
-

# Histogram

---

3 Steps (p.12) :

- 1. Divide the range of the table into classes of equal width.
  - 2. Make a frequency table.
  - 3. Draw a histogram.
-

# Histogram

---

- 1. Divide the range of the table into classes of equal width.

$$width \geq \frac{19.9 - 0.0}{10} = 1.99$$

- 2. Make a frequency table.

interval	count	interval	count
[0, 2)	20	[10, 12)	3
[2, 4)	9	[12, 14)	0
[4, 6)	3	[14, 16)	0
[6, 8)	4	[16, 18)	2
[8, 10)	6	[18, 20)	1

- 3. Draw a histogram (Labels!)
-

# Stemplot

---

3 Steps (p.9) :

- 1. Separate each observation into a stem consisting of all but the final (rightmost) digit and a leaf, the final digit. Stem may have as many digits as needed, but each leaf contains only a single digit.
  - 2. Write the stem in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
  - 3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.
-

# Stemplot

---

2 modifications (p.11):

- 1. Split each stem (when fewer than 5 stems).
  - 2. Trim (when too many stems).
-

# Stemplot

---

□ For our question:

1. original stemplot (20 stems)
  2. too many stems: trim ( $\rightarrow$  2 stems)
  3. too few stems: split ( $\rightarrow$  10 stems)
-



# Stemplot

---

b) Display the data of Table 1.6 in a stemplot

0	0000000000000000011111
0	222233333
0	445
0	6677
0	888999
1	001
1	
1	
1	67
1	9

---

# Question 1.37 (p.27)

---

- 1. gender in a large college class:

Ans: c

Usually we assume it is a normal class.

But you could also choose (b) as an answer for an unusual class.

---

# Question 1.37 (p.27)

---

2. handedness:

Ans: b.

Because we would expect that right-handed students should outnumber lefties substantially. (Round 10 to 15% of the population as a whole is left-handed)

But you could also choose (c) as an answer for an unusual class.

---

# Question 1.37 (p.27)

---

3. height:

Ans: d.

Because one would expect a fair amount of variation but no particular skewness to such a distribution.

---

# Question 1.37 (p.27)

---

- 4. study length: a

Because it is reasonable to expect this to be right-skewed (many students study little or not at all; a few study longer)

---

# Question

---

1. Calculate the sample means, sample medians, and sample standard deviations of the following two data sets. Please notice that each value of the second data set is obtained by multiplying the corresponding value of the first data set by  $-1$  and then adding  $3$ . Comment on the relationship between the two means, two medians the two standard deviations.
    - Data set 1: 1 3 6 9 10;
    - Data set 2: 2 0 -3 -6 -7
-

# Question

---

Data set 1: 1 3 6 9 10;

Using definitions:

a) sample mean

$$= \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{5} (1+3+6+9+10) = 5.8$$

b) sample median=6

---

# Question

---

c) sample standard deviation

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

$$= \sqrt{\frac{1}{5-1} [(1-5.8)^2 + (3-5.8)^2 + (6-5.8)^2 + (9-5.8)^2 + (10-5.8)^2]}$$

$$\approx 3.8$$

---



# Question

---

Data set 2: 2 0 -3 -6 -7

Two ways:

The first way: using definitions

- a) sample mean = -2.8
  - b) sample median = -3
  - c) sample standard deviation  $\approx 3.8$
-

# Question

---

The second way: using linear transformation!

Data set 1: 1 3 6 9 10 -> x

Data set 2: 2 0 -3 -6 -7 -> y

each y value is obtained by first multiplying the corresponding x value by -1 and then adding 3

Linear transformation (a=3, b=-1)

$$y = 3 + (-1)x$$

---

# Recap

---

## Properties of linear transformations

$$y = a + b x$$

$$\text{mean}(y) = a + b * \text{mean}(x)$$

$$\text{median}(y) = a + b * \text{median}(x)$$

$$\text{sd}(y) = |b| * \text{sd}(x)$$

---

# Question

---

Data set 1: 1 3 6 9 10;

Data set 2: 2 0 -3 -6 -7

	sample mean	sample median	sample standard deviation
Data set 1	5.8	6	$\approx 3.8$
Data set 2	$a+b*5.8$ $=3-1*5.8$ $=-2.8$	$a+b*6$ $=3-1*6$ $=-3$	$\approx  b *3.8$ $=1*3.8$ $=3.8$

---

# Question 1.111 (p.72)

---

A strategy to distinguish mean, mode, and median:

Step 1. Symmetric or not.

Step 2. If symmetric: three values take the same point.

If non-symmetric:

3 Steps:

(1) Find the highest peak: the mode.

(2) Find the middle point: the median.

(3) The rest one: the mean. (mean always moves away from the median towards the longer tail)

---

# Question 1.111 (p.72)

---

Ans:

- a) Non-symmetric: B: the median. C: the mean.
  - b) Symmetric: A: the median. A: the mean.
  - c) Non-symmetric: B: the median. A: the mean.
-

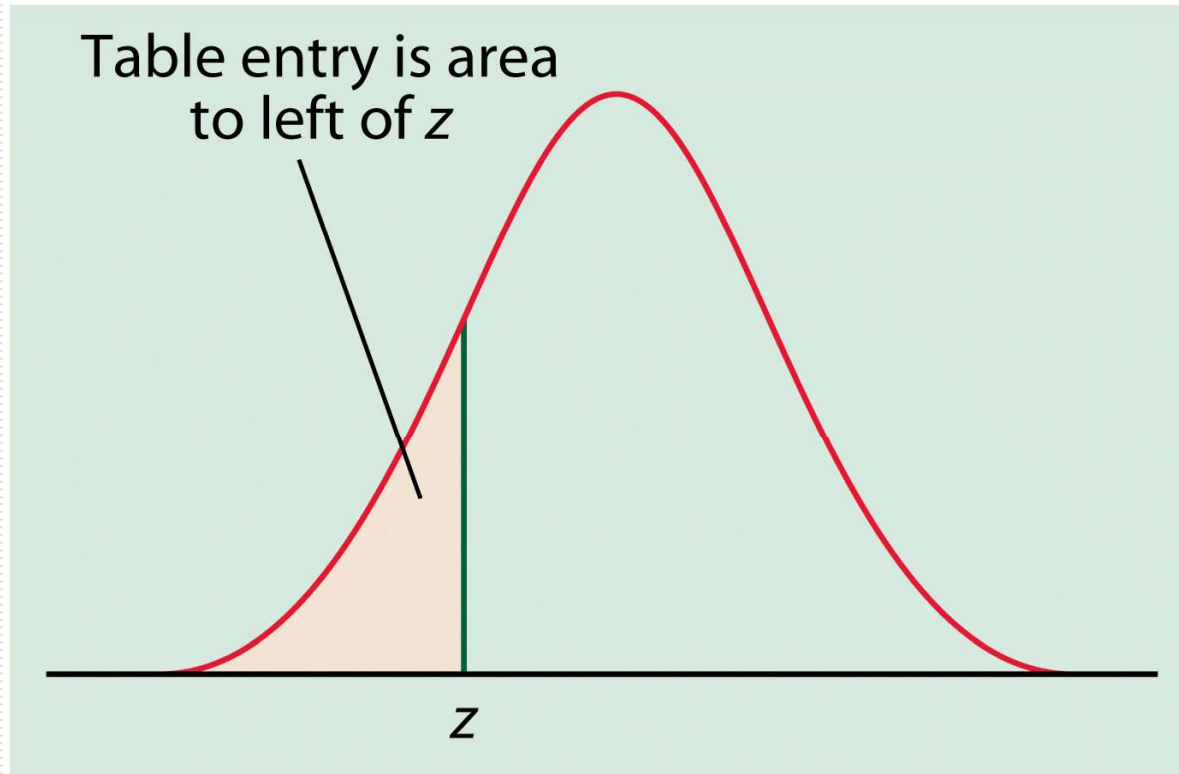
## Question 1.121 (p.74)

---

- a)  $Z \leq -1.9$
  - b)  $Z \geq -1.9$
  - c)  $Z > 1.55$
  - d)  $-1.9 < Z < 1.55$
-

Table A gives the area under the **standard** Normal curve to the **left** of any  $z$  value: cumulative proportion.

---



---

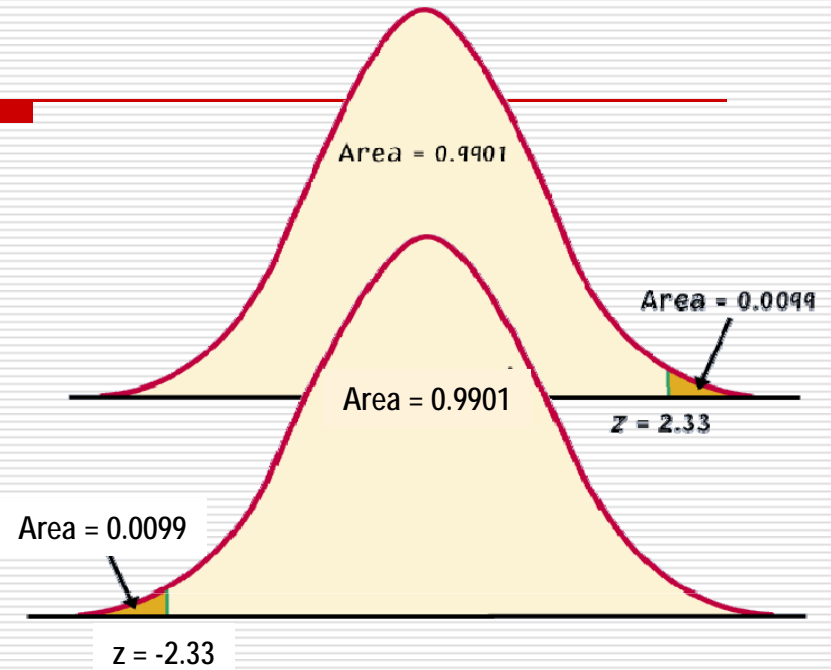
To use table A, we must have a variable with a normal distribution whose mean is 0 and whose standard deviation is 1.



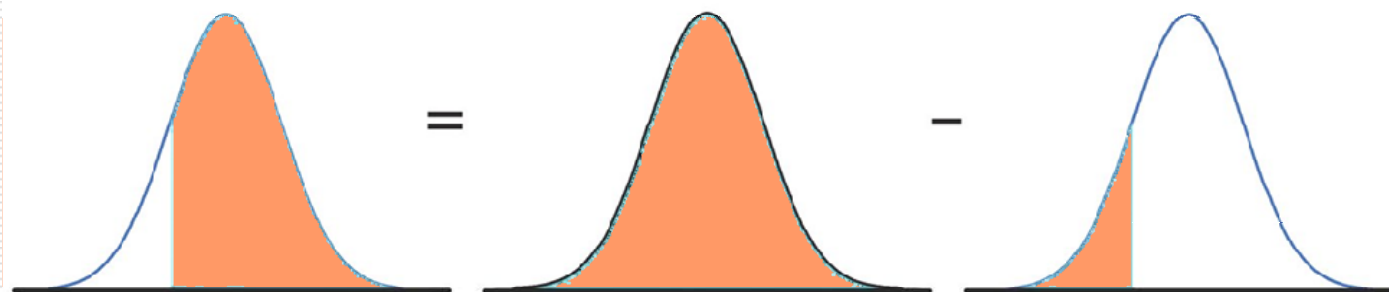
To calculate the area under the standard Normal curve to the right of a z value: 2 ways

1. symmetrical Normal distribution

2. The area to the right of z  
= 1 - the area to the left of z



area right of z = area left of -z



area right of z = 1 - area left of z

TABLE A Standard normal probabilities

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641



**TABLE A Standard normal probabilities**

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0046	.0045	.0044	.0043	.0042	.0041	.0040	.0039	.0038
-2.5	.0062	.0061	.0060	.0059	.0058	.0057	.0056	.0055	.0054	.0053
-2.4	.0082	.0081	.0080	.0079	.0078	.0077	.0076	.0075	.0074	.0073
-2.3	.0107	.0106	.0105	.0104	.0103	.0102	.0101	.0100	.0099	.0098
-2.2	.0139	.0138	.0137	.0136	.0135	.0134	.0133	.0132	.0131	.0130
-2.1	.0179	.0178	.0177	.0176	.0175	.0174	.0173	.0172	.0171	.0170
-2.0	.0228	.0227	.0226	.0225	.0224	.0223	.0222	.0221	.0220	.0219
-1.9	.0287	.0286	.0285	.0284	.0283	.0282	.0281	.0280	.0279	.0278
-1.8	.0359	.0358	.0357	.0356	.0355	.0354	.0353	.0352	.0351	.0350
-1.7	.0446	.0445	.0444	.0443	.0442	.0441	.0440	.0439	.0438	.0437
-1.6	.0548	.0547	.0546	.0545	.0544	.0543	.0542	.0541	.0540	.0539
-1.5	.0668	.0667	.0666	.0665	.0664	.0663	.0662	.0661	.0660	.0659
-1.4	.0808	.0807	.0806	.0805	.0804	.0803	.0802	.0801	.0800	.0799

.0606 is the area under  $N(0,1)$  left of  $z = -1.55$

.0287 is the area under  $N(0,1)$  left of  $z = -1.9$

## Question 1.121 (p.74)

---

a)  $Z \leq -1.9$   $\rightarrow$  check table  $A = .0287$

b)  $Z \geq -1.9$   $\rightarrow 1 - \text{area}(Z < -1.9) = 1 - .0287 = .9713$

c)  $Z > 1.55$   $\rightarrow \text{area}(Z < -1.55) = .0606$

d)  $-1.9 < Z < 1.55$

$\rightarrow \text{area}(Z < 1.55) - \text{area}(Z < -1.9)$

$= [1 - \text{area}(Z < -1.55)] - \text{area}(Z < -1.9)$

$= .9394 - .0287 = .9107$

---

# Question 1.124 & 1.125 (p.74)

---

Usually, 2 steps:

- 1) Standardize  $x$  to a Z-score using the following formula

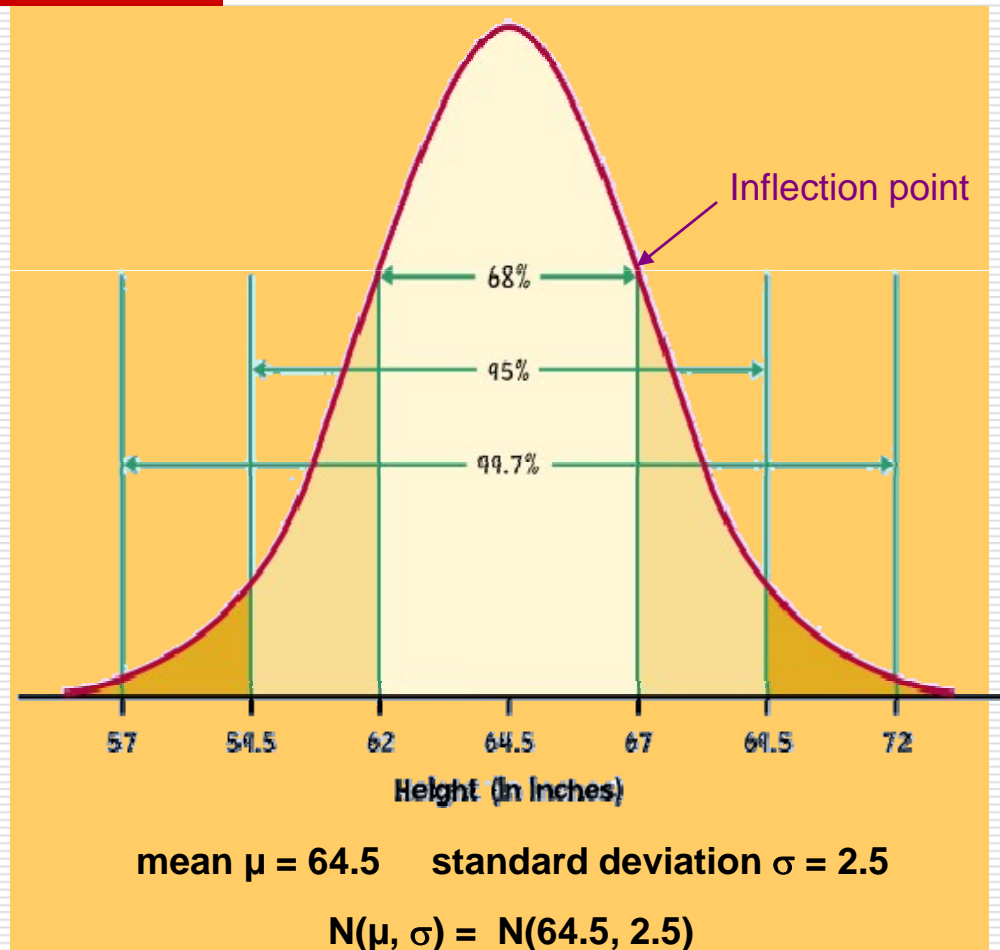
$$z = \frac{x - \mu}{\sigma}$$

- 2) Check Table A for the percentage  
[But for some special cases, we can use "The 68-95-99.7 Rule".]
-

# Recap: The 68-95-99.7 Rule

All Normal curves  $N(\mu, \sigma)$  share the same properties

- ▣ About 68% of all observations are within 1 standard deviation ( $\sigma$ ) of the mean ( $\mu$ ).
- ▣ About 95% of all observations are within 2  $\sigma$  of the mean  $\mu$ .
- ▣ Almost all (99.7%) observations are within 3  $\sigma$  of the mean.



## Question 1.124 & 1.125 (p.74)

---

1.124 Ans:

$$z = \frac{x - \mu}{\sigma} = \frac{70 - 100}{15} = -2$$

using "The 68-95-99.7 Rule", the answer is 2.5%.

1.125 Ans:

$$z = \frac{x - \mu}{\sigma} = \frac{130 - 100}{15} = 2$$

using "The 68-95-99.7 Rule", the answer is 2.5%.



# Question 1.126 & 1.129 (p.74)

---

1.126 Ans:

Tonya: 
$$z = \frac{x - \mu}{\sigma} = \frac{1320 - 1026}{209} \approx 1.4067$$

Jermaine: 
$$z = \frac{x - \mu}{\sigma} = \frac{28 - 20.8}{4.8} \approx 1.5$$

1.129 Ans:

Maria: 
$$z = \frac{x - \mu}{\sigma} = \frac{29 - 20.8}{4.8} \approx 1.7083$$

$$z = \frac{x - \mu}{\sigma} \Leftrightarrow x = z\sigma + \mu \approx 1.7083 \times 209 + 1026 = 1383$$

---