Math 140
Introductory Statistics

Professor Silvia Fernández

Lectures 5 & 6
Based on the book *Statistics in Action*
by A. Watkins, R. Scheaffer, and G. Cobb.

Four different tables

Cumulative distributions reflect the total value *accumulated* from top to bottom (left to right on a plot) on the corresponding table. (More on this p. 78.)

Frequency table*

Weight	Frequency
2.99	1
3.01	4
3.03	4
3.05	4
3.07	7
3.09	17
3.11	24
3.13	17
3.15	13
3.17	6
3.19	2
3.21	1
Total	100

Cumulative frequency table

Weight	Frequency
2.99	1
3.01	5
3.03	9
3.05	13
3.07	20
3.09	37
3.11	61
3.13	78
3.15	91
3.17	97
3.19	99
3.21	100

Relative frequency table

Weight	Frequency
2.99	1/100=0.01
3.01	4/100=0.04
3.03	4/100=0.04
3.05	4/100=0.04
3.07	7/100=0.07
3.09	17/100=0.17
3.11	24/100=0.24
3.13	17/100=0.17
3.15	13/100=0.13
3.17	6/100=0.06
3.19	2/100=0.02
3.21	1/100=0.01
Total	100/100=1

Cumulative relative frequency table

Weight	Frequency
2.99	0.01
3.01	0.05
3.03	0.09
3.05	0.13
3.07	0.20
3.09	0.37
3.11	0.61
3.13	0.78
3.15	0.91
3.17	0.97
3.19	0.99
3.21	1

* This table shows the weights of the pennies in Display 2.3 on page 31.

Mean for a frequency table

Frequency table*

Weight	Frequency
2.99	1
3.01	4
3.03	4
3.05	4
3.07	7
3.09	17
3.11	24
3.13	17
3.15	13
3.17	6
3.19	2
3.21	1
Total	100

The data given by the frequency table is equivalent to

2.99,
3.01, 3.01, 3.01, 3.01,
3.03, 3.03, 3.03, 3.03,
3.05, 3.05, 3.05, 3.05,
3.07, 3.07, 3.07, 3.07, 3.07, 3.07, 3.07,
3.09, 3.09, 3.09, 3.09, 3.09, 3.09, 3.09, 3.09, 3.09, 3.09, 3.09,
3.09, 3.09, 3.09, 3.09, 3.09, 3.09, 3.09,
3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11,
3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11, 3.11,
3.11, 3.11, 3.11, 3.11,
3.13, 3.13, 3.13, 3.13, 3.13, 3.13, 3.13, 3.13, 3.13, 3.13, 3.13,
3.13, 3.13, 3.13, 3.13, 3.13, 3.13, 3.13,
3.15, 3.15, 3.15, 3.15, 3.15, 3.15, 3.15, 3.15, 3.15, 3.15, 3.15,
3.15, 3.15, 3.15,
3.17, 3.17, 3.17, 3.17, 3.17, 3.17,
3.19, 3.19,
3.21.

To compute the mean we need the total sum divided by 100. This is equivalent to

$$\bar{x} = (1(2..99)+4(3.01)+4(3.03)+4(3.05)+7(3.07)+17(3.09)+ 24(3.11)+ 17(3.13)+ 13(3.15)+ 6(3.17)+ 2(3.19)+1(3.21))/100=3.1078$$

* This table shows the weights of the pennies in Display 2.3 on page 31.

Standard deviation for a frequency table

Frequency table*

Weight	Frequency
2.99	1
3.01	4
3.03	4
3.05	4
3.07	7
3.09	17
3.11	24
3.13	17
3.15	13
3.17	6
3.19	2
3.21	1
Total	100

Given that $\bar{x} = 3.1078$ we can compute the standard deviation as follows

$$s_{n-1} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{1(2.99-3.1078)^2+4(3.01-3.1078)^2+\ldots+2(3.19-3.1078)^2+1(3.21-3.1078)^2}{99}}$$

$$= .0431$$

* This table shows the weights of the pennies in Display 2.3 on page 31.

Five-point summary using the cumulative frequency table

Frequency table*

Weight	Frequency
2.99	1
3.01	4
3.03	4
3.05	4
3.07	7
3.09	17
3.11	24
3.13	17
3.15	13
3.17	6
3.19	2
3.21	1
Total	100

Cumulative frequency table

Weight	Frequency
2.99	1
3.01	5
3.03	9
3.05	13
3.07	20
3.09	37
3.11	61
3.13	78
3.15	91
3.17	97
3.19	99
3.21	100

Since we have 100 values, then
lower quartile = (25th value + 26th value)/2
median = (50th value + 51st value)/2,
upper quartile = (75th value + 76th value)/2.

Using the cumulative frequency table
1st value = 2.99
25th value = 26th value = 3.09
50th value = 51st value = 3.11
75th value = 76th value = 3.13
100th value = 3.21

Then the five-point summary is
min = 2.99
$Q_1 = 3.09$
$\bar{x} = 3.11$
$Q_3 = 3.13$
max = 3.21

* This table shows the weights of the pennies in Display 2.3 on page 31.

Using your calculator

Frequency table*

Weight	Frequency
2.99	1
3.01	4
3.03	4
3.05	4
3.07	7
3.09	17
3.11	24
3.13	17
3.15	13
3.17	6
3.19	2
3.21	1
Total	100

Store two lists, one for the weight (L_2) and one for the frequency (L_3). Enter

1-Var Stat L_2, L_3

into your calculator as follows:

Go to STAT
Move right to CALC
Choose 1:1-Var Stat
ENTER
L_2, L_3 ENTER

$\bar{x} = \text{mean}$
$\Sigma x = \text{sum}$
$\Sigma x^2 = \text{sum of squares}$
$Sx = \text{standard deviation } s_{n-1}$
$s x = \text{standard deviation } s_n$
$n = \text{number of values}$
$\min X = \text{minimum value}$
$Q_1 = \text{lower quartile}$
$\text{Med} = \text{median}$
$Q_3 = \text{upper quartile}$
$\max X = \text{maximum value}$

* This table shows the weights of the pennies in Display 2.3 on page 31.

Section 2.4 Recentering and Rescaling

- **Recentering** a data set

(adding the same number *c* to all the values in the set)

  - Shape or spread do not change.
  - It slides the entire distribution by the amount *c,* adding *c* to the median and the mean.

- **Rescaling** a data set

(multiplying all the values in the set by the same positive number *d)*

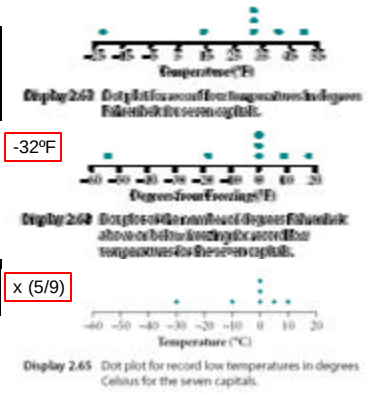
  - Basic shape doesn't change.
  - It stretches or shrinks the distribution, multiplying the spread (*IQR* or standard deviation) by *d* and multiplying the center (median or mean) by *d*.

Example

City	Country	Temperature (°F)
Addis Ababa	Ethiopia	32
Algiers	Algeria	32
Bangkok	Thailand	50
Madrid	Spain	14
Nairobi	Kenya	41
Brazilia	Brazil	32
Warsaw	Poland	-22

-32°F

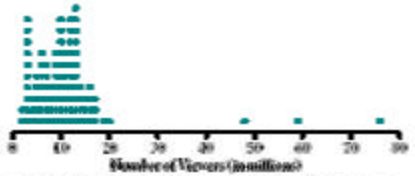
x (5/9)



Display 2.65 Dot plot for record low temperatures in degrees Celsius for the seven capitals.

The Influence of Outliers

- A summary statistic is
  - **resistant to outliers** if it does not change very much when an outlier is removed.
  - **sensitive to outliers** if the summary statistic is greatly affected by the removal of outliers.



Display 2.66 Number of viewers of prime-time television shows in a particular week.

Example

Variable	N	Mean	Median	StDev
Ratings	101	11.187	10.150	9.896
Variable	**Min**	**Max**	**Q1**	**Q3**
Ratings	2.320	76.260	6.160	12.855

Display 2.67 Printout of summary statistics for number of viewers.

Variable	N	Mean	Median	StDev
No Outs	98	9.666	10.145	4.250
Variable	**Min**	**Max**	**Q1**	**Q3**
No Outs	2.320	20.470	6.065	12.698

Display 2.68 Summary statistics for number of viewers without outliers.

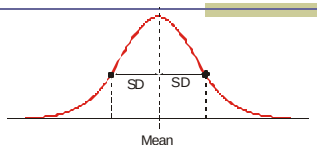
Display 2.66 Number of viewers of prime-time television shows in a particular week.

Percentiles and CRF plots

- You are responsible to read through this and understand the concepts of **percentile**, and **cumulative relative frequency plot**.

2.5 The Normal Distribution

- Shape



- Center: Mean

$$\bar{x} = \frac{\text{sum of values}}{\text{number of values}} = \frac{\sum x}{n}$$

- Spread: Standard Deviation

$$s_{n-1} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

- Variance

$$s_{n-1}^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

Applications of the Normal Distribution

- The normal distribution tells us how:
  - Variability in measures behaves.
  - Variability in population behaves.
  - Averages and some other summary statistics behave when you repeat a random process.
- Nice property: A normal distribution is determined by its mean and standard deviation!

  (If you know the mean and SD you know everything)

The Two Main Problems.

- The distribution of the SAT scores for the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.

1. What percentage of scores were 920 or below?

   (Unknown percentage problem – given value, looking for a percentage)

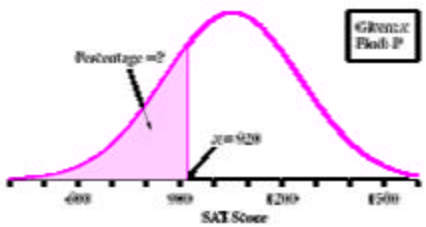
2. What SAT score separates the lowest 25% of the SAT scores from the rest?

   (Unknown value problem – given percentage, looking for a value, in this case values are SAT scores)

Unknown percentage problem.

- The distribution of the SAT scores for the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.

1. '

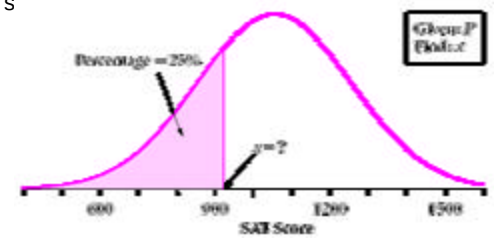


In other words, what percentage of the total area enclosed by the bell-shaped curved and the x-axis is represented by the shaded region?

Unknown value problem.

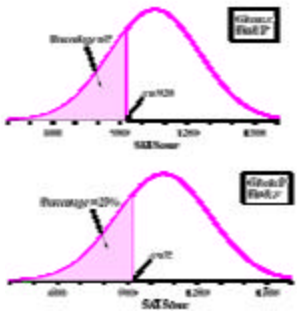
- The distribution of the SAT scores for the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.

2. What SAT score separates the lowest 25% of the SAT s



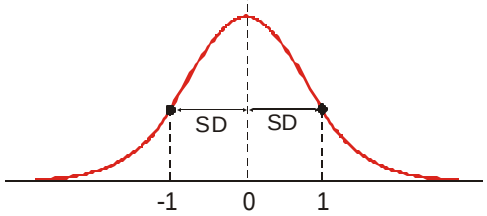
Which one is it?

1. Unknown percentage problem.
   Given x, find P.
2. Unknown value problem.
   Given P, find x.



The Standard Normal Distribution.

- It is the normal distribution with **Mean = 0**, and **standard deviation = 1**.
  The area under the curve equals 1 (or 100%)



The Standard Normal Distribution.

- It is the normal distribution with Mean = 0, and standard deviation = 1.
  The area under the curve equals 1 (or 100%)
- The Standard Normal Distribution is important because any normal distribution can be **recentered** and/or **rescaled** to the standard normal distribution. This process is called **standarizing** or **converting to standard units**.
- Also, the two main problems can be easily solved in the Standard Normal Distribution with the help of **tables** or a **calculator**.
- We use z instead of x to represent a value in the standard normal distribution. This is called a z-score.

The Two Main Problems in the Standard Normal Distribution.

Unknown Percentage. (Given z, find P)

- With Table A (end of the textbook)
  - Use the units and the first decimal to locate the row and the closest hundredths digits to locate the column. The number found is the percentage of the number of values below z.
- With Calculator
  - Enter `normalcdf`(-99999, z) to get the percentage of the number of values below z.

Example (given z find P)

Find the percentage, P, of values below $z = 1.23$.

Calculator
P = normalcdf(-99999,1.23)
= .8906513833
~ **89.07%**

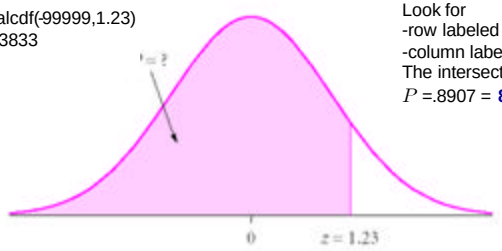


Table A
Look for
-row labeled 1.2
-column labeled .03
The intersection shows
P =.8907 = **89.07%**

Display 2.77 The percentage of values below $z = 1.23$.

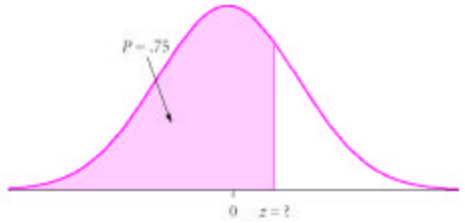
The Two Main Problems in the Standard Normal Distribution.

Unknown Value Problem. (Given P, find z)

- With Table A
  - Look for P in the body of the table. (or the number closest to it). Read back the row and column for that number. Use the row as the units and tenths of z, and the column as the hundredths digits of z. Note that P must be a percentage (written as a proportion, that is, a number between 0 and 1) of the number of values below a certain value z.
- With Calculator
  - Enter invNorm(P) to get the value z such that P equals the percentage of the number of values below z.

Example (given P find z-score)

Find the z-score that falls at the 75th percentile of the standard normal distribution; that is, the z-score that divides the bottom 75% of the values from the rest.



Calculator
z=invNorm(.75)
= .6744897495
~ **.67**

Table A
The value closest to .75 in the body of table A is .7486, which is in row .6 and column .07.
Then the z-score is **.67**

Standardizing (from x to z)

- When we standarize a value x it becomes z. We call z the z -score.
- Standard units = number of standard deviations that a given x value lies above or below the mean.
- As we said before, to standarize we just need to (re)center and (re)scale.

Standardizing (from x to z) cont.

- As we said before, to standarize we just need to (re)center and (re)scale.
  - Step1. Centering (This makes mean = 0)
    Q: How far and which way to the mean?
    A: $x - \overline{x}$
    Subtract the mean from all values.
  - Step 2. Rescaling (this makes SD = 1)
    Q: How many standard deviations is that?
    A: $\dfrac{x-\overline{x}}{SD}$
    Divide all values from Step 1 by the SD.

$$z = \frac{x-\overline{x}}{SD}$$

Unstandardizing (from z to x)

- Solve for x in the previous formula to get

$$x = \overline{x} + z(SD) = mean + z(SD)$$

where z is the z-score.

Example (p. 88 – given x find P)

Example

For groups of similar individuals, heights are often approximately normal in their distribution. For example, the heights of 18- to 24-year-old males in the United States are approximately normal, with mean 70.1 inches and standard deviation 2.7 inches. What percentage of these males are more than 74 inches tall?

Source: Statistical Abstract of the U.S. 1999.

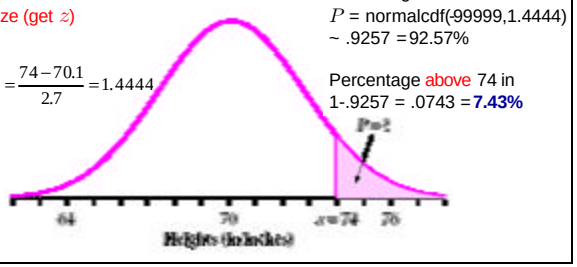
Standarize (get z)

$x = 74$

$z = \dfrac{x-\overline{x}}{SD} = \dfrac{74-70.1}{2.7} = 1.4444$

Percentage below 74 in
P = normalcdf(-99999,1.4444)
~ .9257 = 92.57%

Percentage above 74 in
1-.9257 = .0743 = **7.43%**



Example (p. 89 – given P find x)

Example

The heights of females in the United States who are between the ages of 18 and 24 are approximately normally distributed, with mean 64.8 inches and standard deviation 2.5 inches. What height separates the shortest 75% from the tallest 25%?

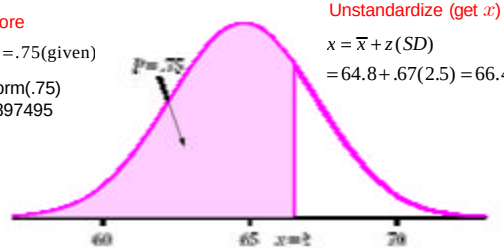
Get z-score

$P = 75\% = .75$(given)

z =invNorm(.75)
= .6744897495
~ **.67**

Unstandardize (get x)

$x = \overline{x} + z(SD)$
$= 64.8 + .67(2.5) = 66.475$ in



Example (p. 91 – given P find x)

- According to the table on page 87, the distribution of death rates from cancer per 100,000 residents by state is approximately normal*, with mean 196 and *SD* 31. The middle 90% of death rates are between what two numbers?

*Provided that Alaska and Utah, which are outliers because of their unusually young populations, are left out.

Example (p. 91 – given P find x) cont.

- According to the table on page 87, the distribution of death rates from cancer per 100,000 residents by state is approximately normal*, with mean 196 and *SD* 31. The middle 90% of death rates are between what two numbers?
- Get z-scores (middle 90% is betwee 5% and 95%)
  5% =.05 corresponds to z = -1.64485
  95%= .95 corresponds to z = 1.64485
- Unstandardize

$$x = \bar{x} + z(SD) = mean + z(SD) = 196 + (-1.64485)(31) = 145.00965$$

$$x = x + z(SD) = mean + z(SD) = 196 + (1.64485)(31) = 246.99035$$

- So the middle 90% of states have death rates between 145.009 and 246.99 deaths per 1000,000 residents.

*Provided that Alaska and Utah, which are outliers because of their unusually young populations, are left out.