(S2016)

Section 2.1 – Describing Location in a Distribution

1. Measuring Position: Percentiles

Definition: The ***pth percentile*** of a distribution is the value with P % OF OBS. BELOW IT,

Example: The stemplot below shows the number of wins for each of the 30 Major League Baseball teams in 2009.

ROYALS/INDIANS

5 | 9
6 | 2455
7 | 00455589
8 | 0345667778
9 | 123557 — ROCKIES
10 | 3 — YANKEES.

Key: 5 | 9 represents a team with 59 wins.

Find the percentiles for the following teams: (a) The Colorado Rockies, who won 92 games; (b) The New York Yankees, who won 103 games; (c) the Kansas City Royals and the Cleveland Indians, who both won 65 games.

A) 92 → 24 BELOW $\frac{24}{30}$ = 80th %ile

B) 103 → 29 BELOW $\frac{29}{30}$ ≈ 97th %ile

C) 65 → 3 BELOW $\frac{3}{30}$ = 10th %ile

Note: some people define the *pth percentile* of a distribution as the value with *p* percent *less than or equal* to it. In this case it is possible for an individual to be at the 100th percentile.

---

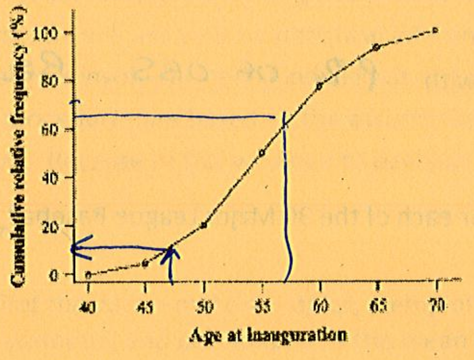
2. Cumulative Relative Frequency Graphs

When you are given a *frequency table* for a quantitative variable, it is possible to graphs that depict the *percentiles*. The table gives the inauguration ages of the first 44 US Presidents.

Age	Frequency	CUM FREQ	CUM. REL. FREQ
40-44	2	2	2/44 = 4.5%
45-49	7	9	9/44 = 20.5%
50-54	13	22	22/44 = 50%
55-59	12	34	34/44 = 77.3%
60-64	7	41	41/44 = 93.2%
65-69	3	44	44/44 = 100%

→

Interpreting Cumulative Relative Frequency graphs



(a) Was Barack Obama, at 47, unusually young?

YES ⇒ 11-12%ILE

(b) Estimate and interpret the 65th percentile of the distribution.

65TH %ILE → 58 YO.

65% OF PRES WERE YOUNGER THAN 58 YO @ TIME OF INAUG.

CHECK YOUR UNDERSTANDING

1. *Multiple choice: Select the best answer.* Mark receives a score report detailing his performance on a statewide test. On the math section, Mark earned a raw score of 39, which placed him at the 68th percentile. This means that

(a) Mark did better than about 39% of the students who took the test.

(b) Mark did worse than about 39% of the students who took the test.

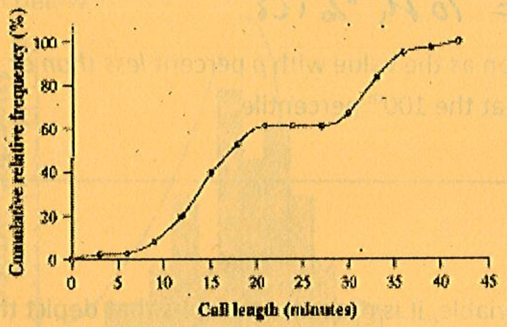
[✓ (c) Mark did better than about 68% of the students who took the test.]

(d) Mark did worse than about 68% of the students who took the test.

(e) Mark got fewer than half of the questions correct on this test.

2. Mrs. Munson is concerned about how her daughter's height and weight compare with those of other girls of the same age. She uses an online calculator to determine that her daughter is at the 87th percentile for weight and the 67th percentile for height. Explain to Mrs. Munson what this means.

Questions 3 and 4 relate to the following setting. The graph displays the cumulative relative frequency of the lengths of phone calls made from the mathematics department office at Gabalot High last month.



3. About what percent of calls lasted less than 30 minutes? 30 minutes or more?

4. Estimate Q_1, Q_3, and the IQR of the distribution.

(2) DAUGHTER WEIGHS MORE THAN 87% OF GIRLS HER AGE AND SHE IS TALLER THAN 67% OF GIRLS HER AGE.

(3) 65% LASTED LESS THAN 30 MIN.
35% LASTED > THAN 30 MIN.

(4) $Q_1 = 13$ MIN $Q_3 = 32$ MIN ⇒ IQR $= 32 - 13 = 19$ MIN

3. Measuring Position: Z-Scores

Another way of *measuring position* is to determine how many *standard deviations* above or below the mean an individual data point is. This is called computing a ***z-score***. This process is known as ***standardizing.***

Definition - **Standardized value (z-score)**:
If x is an observation from a distribution that has a known mean and standard deviation, the **standardized value** of x is

$$Z = \frac{X - \text{MEAN}}{\text{STD DEV}}$$

FORMULA SHEET

This measure tells how many standard deviations the given data point is from the mean.

Example: 2009 MLB Wins (revisited)

```
 5 | 9
 6 | 2455
 7 | 00455589
 8 | 0345667778
 9 | 123557
10 | 3
```

Key: 5 | 9 represents a team with 59 wins.

Mean: 81
Median: 83.5
StDev: 11.43
Minimum: 59
Maximum: 103
Q1: 74
Q3: 88

Use the information provided to find the standardized scores for the (a) Boston Red Sox with 95 wins; (b) Atlanta Braves with 86 wins; and (c) Washington Nationals with 59 wins.

(A) SOX 95 WINS: $z_{95} = \frac{95-81}{11.43} = 1.225$

(B) BRAVES 86 WINS: $z_{86} = \frac{86-81}{11.43} = 0.437$

(C) NAT'S 59 WINS: $z_{59} = \frac{59-81}{11.43} = -1.925$

DISCUSS RESULTS (COMPARING DATA SETS)

(1) $z_{65} = \frac{65-67}{4.29} = -0.466$

HT IS 0.466 SD'S BELOW MEAN HT OF CLASS.

(2) $z_{74} = \frac{74-67}{4.29} = 1.632$ ABOUT 1.632 SD'S ABOVE CLASS MEAN

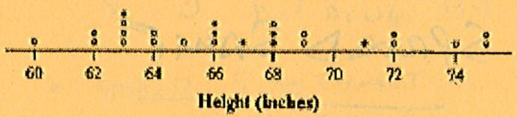
CHECK YOUR UNDERSTANDING

Mrs. Navard's statistics class has just completed the first three steps of the "Where Do I Stand?" Activity (page 84). The figure below shows a dotplot of the class's height distribution, along with summary statistics from computer output.

1. Lynette, a student in the class, is 65 inches tall. Find and interpret her z-score.

2. Another student in the class, Brent, is 74 inches tall. How tall is Brent compared with the rest of the class? Give appropriate numerical evidence to support your answer.

3. Brent is a member of the school's basketball team. The mean height of the players on the team is 76 inches. Brent's height translates to a z-score of −0.85 in the team's height distribution. What is the standard deviation of the team members' heights?



Variable	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max
Height	25	67	4.29	60	63	66	69	75

Homework: pp 100-101, 5-15 odd

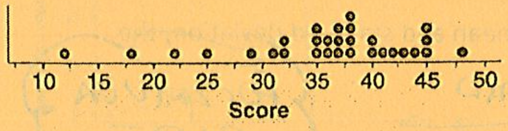
(3) $-0.85 = \frac{74-76}{s_x}$

$s_x = \frac{-2}{-0.85} = 2.35$ INCHES

Section 2.1 (continued)

3. Transforming Data

Example: Below is a graph and table of summary statistics for a sample of 30 test scores. The maximum possible score on the test was 50 points.



	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Score	30	35.8	8.17	12	32	37	41	48	9	36

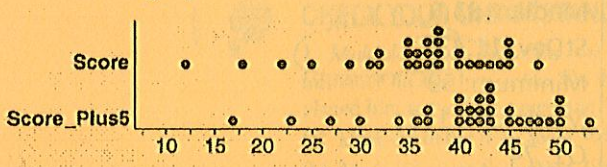
Suppose that the teacher was *nice* and added 5 points to each test score. How would this change the shape, center, and spread of the distribution?

CTR ↑ 5 (MEAN + MEDIAN)

SHAPE SAME

SPREAD SAME (IQR, S_x, RG)

Here are the graphs and the summary statistics for the original scores and the +5 scores:



	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Score	30	35.8	8.17	12	32	37	41	48	9	36
Score + 5	30	40.8	8.17	17	37	42	46	53	9	36

MED : 37 → 42 (+5)

MIN : 12 → 17 ''

MAX : 48 → 53 ''

IQR
RG
S_x } SAME ⇒ SPREAD =

P. 94

Effect of Adding (or Subtracting) a Constant

Adding the same number *a* (either positive, zero, or negative) to each observation:

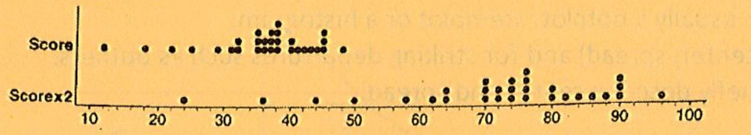
- Adds *a* to measures of center and location (mean, median, quartiles, percentiles), but
- Does not change the *shape* of the distribution or measures of spread (range, IQR, standard deviation.

Application: If 24 is added to every observation in a data set, the only one of the following that is *not* changed is:

(a) the mean (b) the 75th percentile (c) the median [✓ (d) the standard deviation] (e) the minimum

SPREAD SAME

Example (cont): Suppose that the teacher in the previous example wanted to convert the *original* test scores to percents. Since the test was out of 50 points, he should multiply each score by 2 to make them out of 100. Here are the graphs and summary statistics for the original scores and the doubled scores.



	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Score	30	35.8	8.17	12	32	37	41	48	9	36
Score × 2	60	71.6	16.34	24	64	74	82	96	18	72

What happened the measures of center, location and spread? DOUBLED

What happened to the shape? DID NOT CHANGE

P. 95

Effect of Multiplying (or Dividing) by a Constant

Multiplying (or dividing) each observation by the same number *b* (positive, negative or 0)

- Multiplies (divides) measures of *center, location* (mean, median, quartiles, percentiles) by *b*,
- Multiplies (divides) measures of *spread* (range, IQR, standard deviation) by |*b*|, but
- Does not change the *shape* of the distribution.

4. Transformations and Z-Scores

Example (continued). Suppose we wanted to standardize the original test scores. This would mean we would subtract each score from the mean of 35.8 and then divide by the standard deviation of 8.17.

	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Score	30	35.8	8.17	12	32	37	41	48	9	36

$$Z = \frac{x - 35.8}{8.17}$$

STANDARDIZED DIST: $\bar{x} = 0$ $S_x = 1$

What effect would these transformations have on:

- Shape? SHAPE STAYS SAME.
- Center? SUBTRACTING 35.8 WOULD REDUCE $\bar{x}$ BY 35.8 ⇒ $\bar{x} = 0$. DIVIDING BY 8.17 ⇒ $\bar{x} = 0$ STILL.
- Spread? SUBTRACTING 35.8 WOULD NOT CHANGE SPREAD BUT DIVIDING BY 8.17 WOULD CAUSE S_x TO BE 1.

Team Work: Complete Check Your Understanding on pp. 97-98

① SHAPE WILL NOT Δ. CTR/SPRD MULT BY 2.54

② SHAPE + SPREAD WILL NOT Δ. CENTER ↑ BY 6

③ SHAPE - NO. CTR = 0, STD DEV = 1

Homework: pp. ~~107-109~~, 19, 21, 23, 25-29
102-103

Section 2.2 – Density Curves & Normal Distributions

Density Curves

P.104

Exploring Quantitative Data

1. Always plot your data: make a graph, usually a dotplot, stemplot or a histogram.
2. Look for the overall pattern (shape, center, spread) and for striking departures such as outliers.
3. Calculate a numerical summary to briefly describe center and spread.

New step:

4. Sometimes the overall pattern of a *large* number of observations is so regular that we can describe it with a *smooth curve.* → LEADS TO INFO WE CAN USE LATER TO MAKE INFERENCES.

This type of *smooth curve* is called a **Density Curve.**

Definition: A **density curve** is a curve that

- Is always above the horizontal axis, and
- Has an area of exactly 1 underneath it

A density curve describes the overall pattern of a distribution. The area under the curve and above any interval of values on the horizontal axis is the proportion of all observations that fall in that interval.

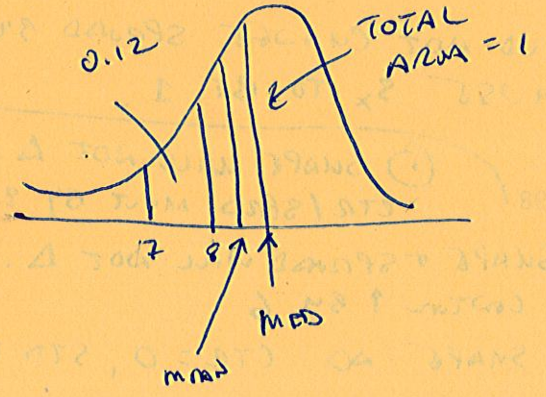
Note: *no set of real data is exactly described by a density curve. The curve is an approximation that is easy to use and accurate enough for practical use.*

Because the density curve represents a *population* of individuals, the mean is denoted by μ (the Greek letter mu) and the standard deviation is denoted by σ (the Greek letter sigma).

Distinguishing the Median and Mean of a Density Curve (Diagrams on p. ~~102)~~ 107 FIG 2.10

- The **median** of a density curve is the *equal-areas* point, the point that divides the area under the curve in half.
- The **mean** of a density curve is the *balance point*, the point at which the curve would balance if made of solid material.
- The median and mean are the same for a perfectly symmetric density curve. The both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

Team Work: Complete Check Your Understanding on p. 107.



① LEGIT: AREA = 1, POSITIVE

② BETWEEN 7 AND 8 12%

③ ④ } MEAN < MEDIAN (SKEWED LEFT)

Probably the most famous of all *density curves* are **Normal curves**. The distributions they describe are called **Normal distributions**. They play a very large part in statistics.

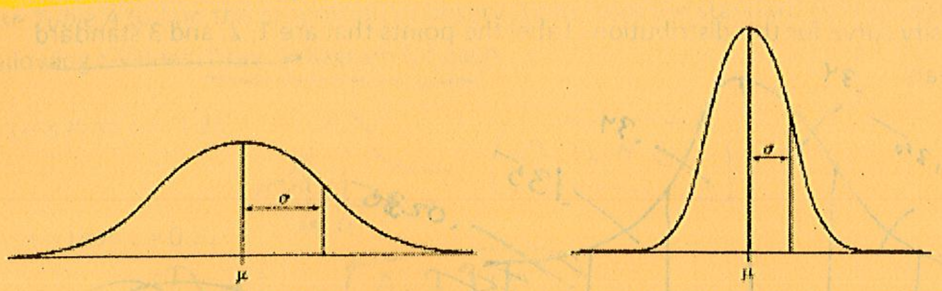


FIGURE 2.11 Two Normal curves, showing the mean μ and standard deviation σ.

Normal curves have several properties:

- All Normal curves have the same overall shape: symmetric, single-peaked, bell-shaped.
- Any specific Normal curve is completely described by its mean μ and standard deviation σ.
- The mean is located at the center and is equal to the median. Changing μ without changing σ moves the Normal curve along the horizontal axis without changing its shape.
- The standard deviation σ controls the spread of a Normal curve. Normal curves with larger standard deviations are more spread out.

The points at which the Normal curve changes from *concave down* to *concave up* occurs one standard deviation from the mean. Because of this, the standard deviation can be estimated by the graph.

(INFLECTION POINT)

Definition: A **Normal distribution** is described by a Normal density curve. Any particular Normal distribution is completely specified by its mean μ and standard deviation σ. The mean of a Normal distribution is at the center of the symmetric **Normal curve** and equals the median. The standard deviation is the distance from the center to the inflection points (where concavity changes) on either side.

Notation: We abbreviate the Normal distribution with mean μ and standard deviation σ as ***N(μ, σ)***.

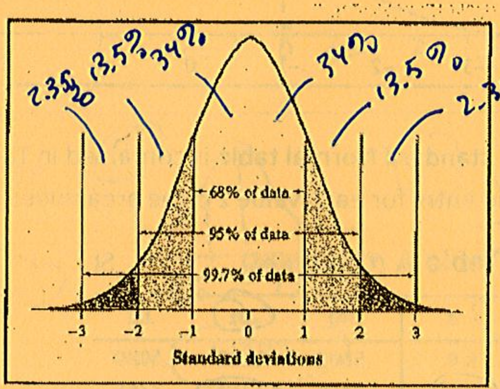
The 68-95-99.7 Rule

In a Normal distribution with mean μ and standard deviation σ:

- Approximately 68% of the observations fall within 1 σ of the mean μ.
- Approximately 95% of the observations fall within 2 σ's of the mean μ.
- Approximately 99.7% of the observations fall within 3 σ's of the mean μ.

(Note: this rule does not apply to any distribution – only the Normal. Common error on AP Exam.)

(APPLET)

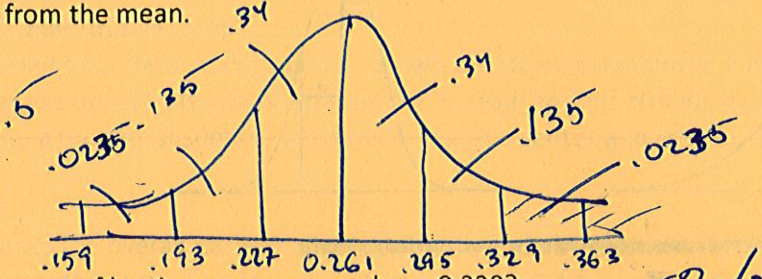


68 - 95 - 99.7

Example: The mean batting average for the 432 Major League Baseball players in 2009 was 0.261 with a standard deviation of 0.034. Suppose the distribution is exactly Normal with μ = 0.261 and σ = 0.034.

a. Sketch a Normal density curve for this distribution. Label the points that are 1, 2, and 3 standard deviations from the mean.

95
-68
27%/2 = 13.5

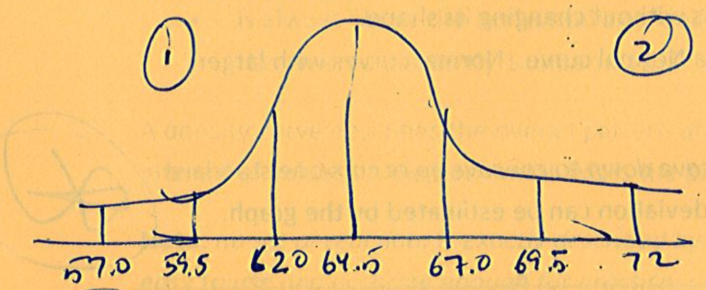


b. What percent of batting averages are above 0.329? 5%/2 = 2.5%

c. What percent of batting averages are between 0.193 and 0.295?

68% + 13.5% = 81.5%

Team Work: Complete Check Your Understanding on p. 112. 18-24 YO F's N(64.5, 2.5)



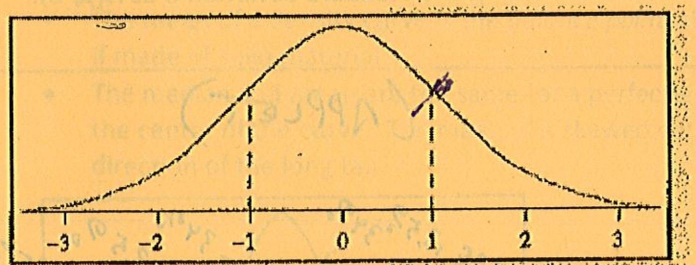
2 $P(x > 67) = \frac{100-68}{2} = 16\%$

3 $P(62 < x < 72) =$

$\frac{68}{2} + \frac{99.7}{2} = 84\%$

The Standard Normal Distribution

Definition: The **standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1. If a variable x has any Normal distribution *N(μ, σ)* with mean μ and standard deviation σ, then the standardized variable $z = \frac{x-\mu}{\sigma}$ has the standard Normal distribution.



68-95-99.7 Rule: For the standard Normal distribution

68% BETWEEN ± 1

95% BETWEEN ± 2

99.7 BETWEEN ± 3

The **standard Normal table** is contained in Table A. It is a table of areas under the Normal curve. The table entry for each value z is the area under the curve to left of z. This is also known as the *lower tail*.

$P(Z \leq 0.11) =$

SHOW TBL

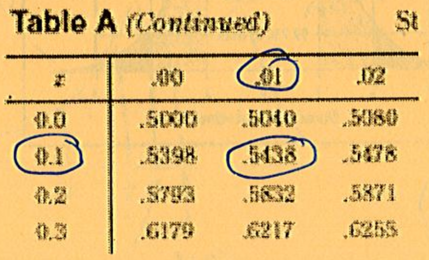
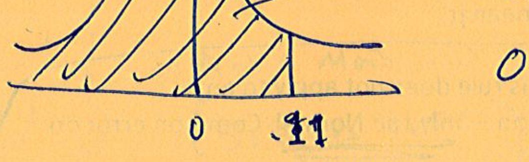


Table A *(Continued)* St

z	.00	.01	.02
0.0	.5000	.5040	.5080
0.1	.5398	.5438	.5478
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255



0.5438

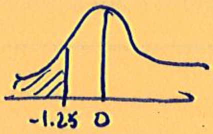
TBL 0.8 → 0.1 .7910

Example: Finding areas under the standard Normal curve. DRAW PICTURES!

Use *Table A* to find the proportion of observations from the standard Normal distribution given the following z-values. Draw a diagram for each.

PROBABILITIES

a. Less than z = -1.25



$\boxed{0.1056}$

NORMALCDF

2ND DISTR

2:NORMALCDF.

b. Less than z = 0.81

$\boxed{0.7910}$

c. Greater than z = 0.81

$1 - .7910 = \boxed{0.2090}$

d. Between z = -1.25 and z = 0.81

Example: Repeat the previous example using *technology*.

2ND DIST (VARS)

2: NORMALCDF(

LOWER

UPPER

μ

σ

a. Less than z = -1.25 ~~.4502~~ 0.1056

b. Less than z = 0.81 0.7910

c. Greater than z = 0.81 0.20897 → 0.2090

d. Between z = -1.25 and z = 0.81 0.6854

Z-SCORES

Example: Working backwards..... NEED LOWER TAIL!

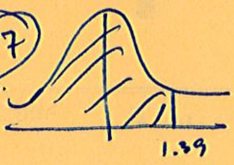
Find the 90th percentile of standard Normal distribution

a. Using *Table A*

b. Using *technology* INVNORM(.9 = 1.282 = Z

Team Work: Complete Check Your Understanding on p. 116.

① z < 1.39

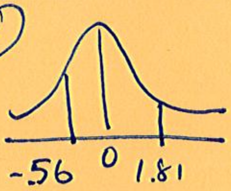


0.9177

② z > -2.15

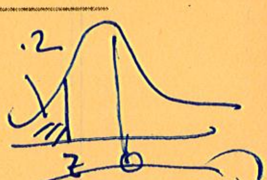
.9842

③



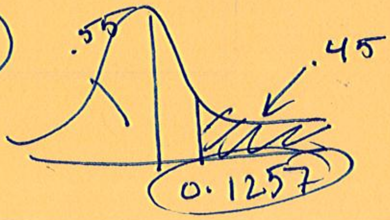
.6772

④



-0.84.6

⑤



0.1257

Homework: pp. 128 problems 35, 37, 41, 43, 45, 47, 49, 51

Section 2.2 Normal Distributions (continued)

******Normal Distribution Calculations******

We will use the previous procedures to answer questions about observations in *any* Normal distribution by *standardizing* and then using the standard Normal table.

4-Step Process

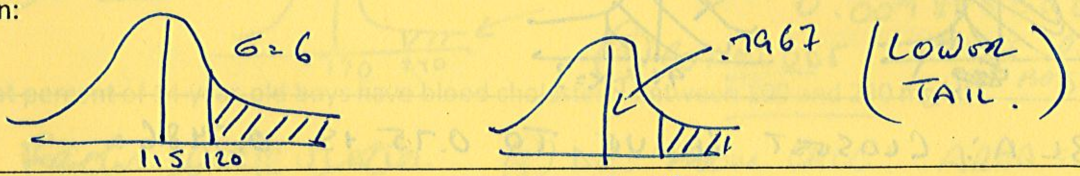
(P 120)

1. *State*: EXPRESS THE PROBLEM IN TERMS OF THE OBSERVED VARIABLE X.
2. *Plan*: DRAW A PICTURE OF DIST. + SHADE AREA OF INTEREST.
3. *Do*: PERFORM CALCULATIONS
   - STANDARDIZE X TO RESTATE PROBLEM IN TERMS OF Z
   - USE TABLE A AND THE FACT THAT TOTAL AREA UNDER CURVE IS 1 TO FIND REQUIRED AREA OF INTEREST.
4. *Conclude*: WRITE CONCLUSION IN CONTEXT OF PROBLEM.

Example: In the 2008 Wimbledon tennis tournament, Rafael Nadal averaged 115 miles per hour on his first serves. Assume that the distribution of his first serves is Normal with a mean of 115 mph and a standard deviation of 6 mph. About what proportion of his first serves would you expect to exceed 120 mph?

1. State: LET X = SPEED OF FIRST SERVE. THE VARIABLE IS N(115, 6). WE WANT THE PROPORTION OF SERVES ≥ 120.

2. Plan:



3. Do:

$$z = \frac{120-115}{6} = 0.83$$

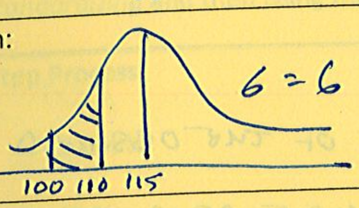
TABLE A $1 - 0.7967 = 0.2033$

4. Conclude: ABOUT 20% OF NADAL'S FIRST SERVES EXCEED 120 MPH.

Example continued: What percent of Rafael Nadal's first serves are between 100 and 110 mph?

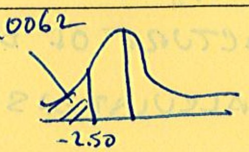
1. State: LET X = SPEED OF FIRST SERVE. X IS $N(115, 6)$.
WE WANT PROPORTION OF FIRST SERVES
WITH $100 < X < 110$

2. Plan:

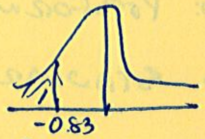


$\sigma = 6$

3. Do:



$$z_{100} = \frac{100-115}{6} = -2.50$$



$$z_{110} = \frac{110-115}{6} = -0.83$$

TABLE A: ~~x~~ $X < 100 \Rightarrow z < -2.50$ 0.0062
$X < 110 \Rightarrow z < -0.83$ 0.2033

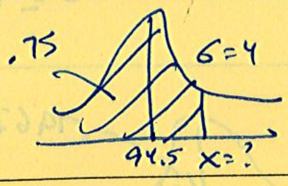
$$\begin{array}{r} 0.2033 \\ -\,0.0062 \\ \hline 0.1971 \end{array}$$

4. Conclude: ABOUT 20% OF NADAL'S FIRST SERVES ARE BETWEEN 100 AND 110 MPH.

Example: According to the Centers for Disease Control (CDC), the heights of three-year-old females are approximately Normally distributed with a mean of 94.5 cm and a standard deviation of 4 cm. What is the third quartile of this distribution?

1. State: LET X = HT OF RANDOMLY SELECTED 3YO FEMALE
X IS $N(94.5, 4)$. WHAT IS 3RD QUARTILE OF DIST?

2. Plan:



3. Do: TBL A: CLOSEST VALUE TO 0.75 IS 0.7486.
CORRESPONDS TO Z-SCORE OF 0.67.

UNSTANDARDIZE $0.67 = \frac{x - 94.5}{4}$

$$\boxed{x = 97.18\text{cm}}$$

4. Conclude: THE 3RD QUARTILE OF HTS IS 97.18cm.

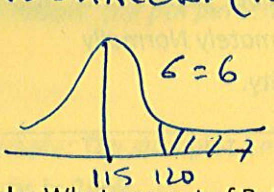
(DISCUSS CALCULATOR SPEAK) (P. 123) (DISTR) 2ND VARS

****Normal Distribution Calculations with Technology****

Example: Nadal *N(115, 6)*. Find the proportion of first serves we expect to exceed 120 mph.

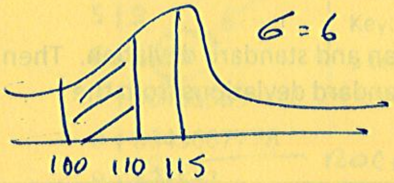
NORMALCDF(120, E99, 115, 6)

2: NORMALCDF(LB, UB, M, σ)



0.2023283246 ⟹ 20% OF TIME.

Example: What percent of Rafael Nadal's first serves are between 100 and 110 mph?

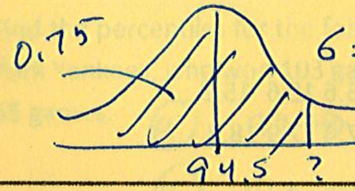


NORMALCDF(100, 110, 115, 6)

0.1961186447 ⟹ 20% OF TIME

Example: 3-year-olds *N(94.5, 4)*. What is the third quartile of this distribution?

3: INVNORM(PROB, M, σ)



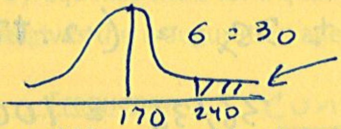
97.197959 cm (DISCUSS DIFF FROM TBL)

Check Your Understanding. Use the 4-Step Process for each of these. Include a properly labeled diagram.

1. Cholesterol levels in 14-year-old boys is approximately Normally distributed with a mean of 170 mg/dl of blood and standard deviation 30 mg/dl. What percent of 14-year-old boys have more than 240 mg/dl of cholesterol?

① X = CHOLEST. LEVELS IN 14 YO BOYS. X IS N(170, 30). WHAT % HAVE GREATER THAN 240mg/dl?

② PLAN:



③ DO: FROM CALCULATOR, AREA = 0.009815 3068

④ CONCLUDE: ABOUT 0.9% OF 14 YO'S BOYS HAVE CHOL > 240 mg/dl

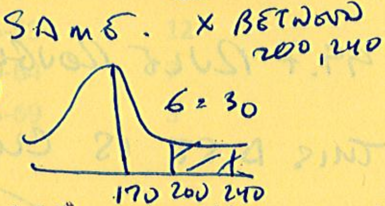
2. What percent of 14-year-old boys have blood cholesterol between 200 and 240 mg/dl?

~~FROM CALCULATO.~~ ③ DO: FROM CALC: AREA = 0.1488

① SAME. X BETWEEN 200, 240

④ CONCLUDE: ABOUT 15% OF 14 YO BOYS HAVE CHOL. BETWEEN 200 AND 240 mg/dl.

②

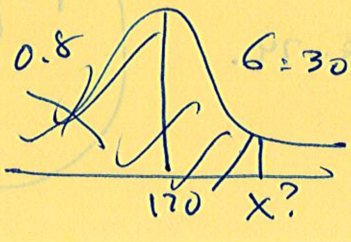


3. What level of cholesterol would represent the 80th percentile?

① SAME. WHAT LEVEL IS 80%ILE?

③ FROM CALC X = 195.249

② 0.8



④ 80%ILE IS ≈ 195 mg/dl.
80% OF 14 YO BOYS HAVE CHOL LEVELS BELOW 195 mg/dl.

(STOPPED HERE)

Assessing Normality

The Normal distributions provide good models for some distributions of real data. In the latter part of this course, we will use various statistical inference procedures to try to answer questions important to us. These tests involve sampling individuals and analyzing data to gain insights about populations. Many of these procedures are based on the assumption that the population is *approximately Normally distributed*. Because of this we need to develop a strategy for assessing Normality.

Procedure.

Step 1: *Plot the data* – make a dotplot, stemplot, or histogram. See if the graph is approximately symmetric and bell-shaped. Is the mean close to the median?

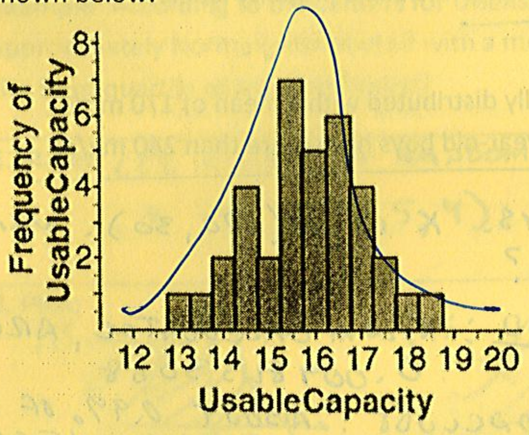
Step 2: *Check whether the data follow the 68-95-99.7 rule.* Find the mean and standard deviation. Then count the number of observations within one, two, and three standard deviations from the mean and compute these to percents.

Example. The measurements listed below describe the usable capacity (in cubic feet) of 36 side-by-side refrigerators. Are the data close to Normal?

12.9 13.7 14.1 14.2 14.5 14.5 14.6 14.7 15.1 15.2 15.3 15.3 15.3 15.3 15.5 15.6 15.6 15.8
16.0 16.0 16.2 16.2 16.3 16.4 16.5 16.6 16.6 16.6 16.8 17.0 17.0 17.2 17.4 17.4 17.9 18.4

The mean and standard deviation of these data are 15.825 and 1.217 cubic feet. The histogram is shown below.

NORMAL PROB. PLOT # P.123



$\bar{x} \pm 1s_x = (14.608, 17.042)$

$24/36 = 66.7\%$

$\bar{x} \pm 2s_x = (13.391, 18.259)$

$34/36 = 94.4\%$

$\bar{x} \pm 3s_x = (12.174, 19.467)$

$36/36 = 100\%$

GRAPH: ROUGHLY SYMMETRIC
%'S FOLLOW 68-95-99.7 RULE ROUGHLY.
∴ GOOD EVIDENCE THAT THIS DIST IS CLOSE TO NORMAL.

53, 55, 59 →

Homewor~~k: pp 132-135 – 53-59 odd, 63a, 63b, 68-74~~

PP 130 [illegible]: 53, 56, 59, 63, 68-74.

NORMAL PROB PLOT P.123