

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Describing Data: Part II

John McGready
Johns Hopkins University

Lecture Topics

- ◆ The normal distribution
- ◆ Calculating measures of variability
- ◆ Variability in the normal distribution
- ◆ Calculating normal scores
- ◆ Sampling variability

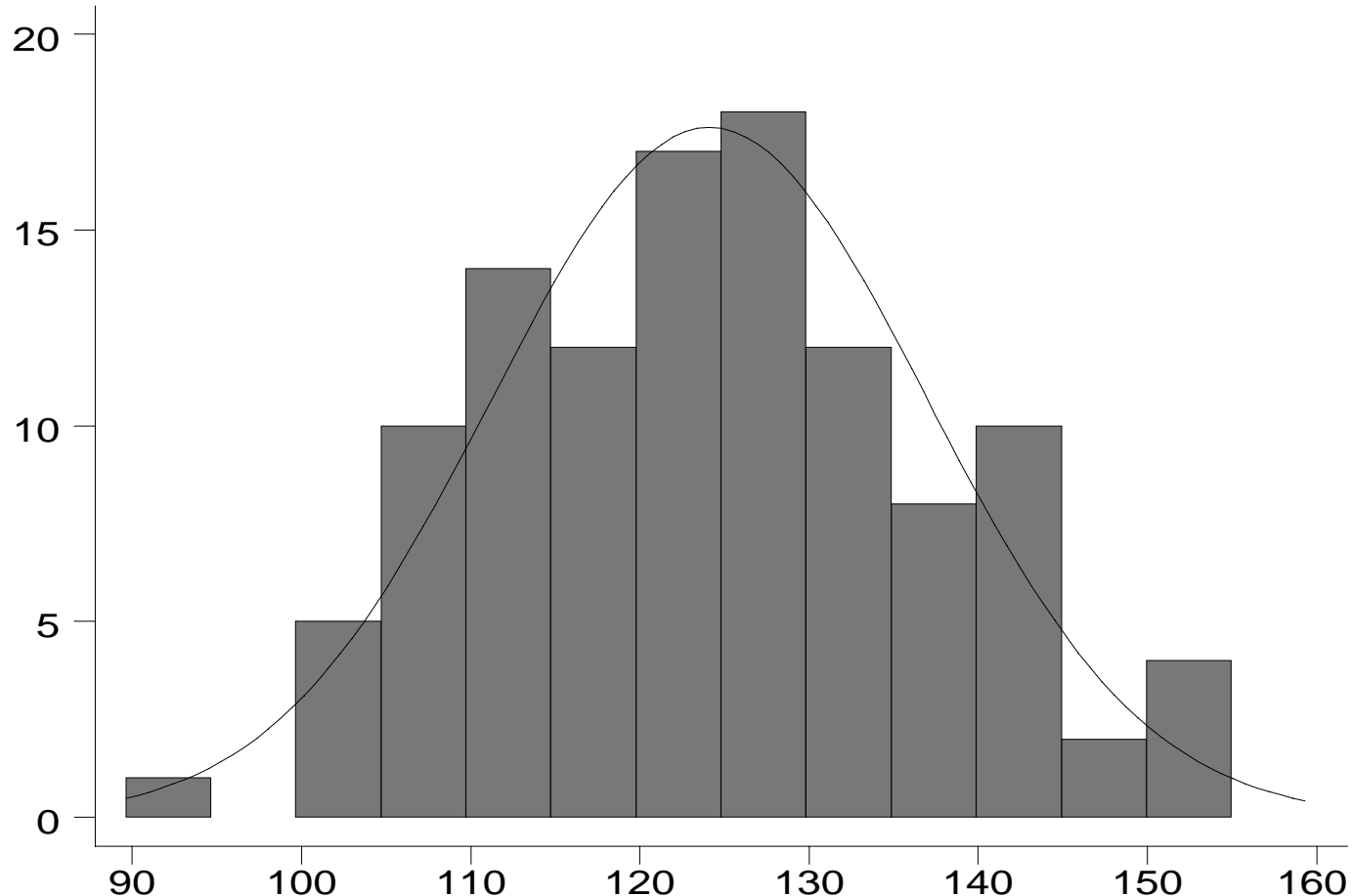


JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

*The Normal Distribution;
Calculating Measures of Variability*

Normal Distribution



The normal (Gaussian) distribution with the same mean and standard deviation (superimposed)

Normal Distribution

Q Is every variable normally distributed?

A Absolutely not

Normal Distribution

Q Then why do we spend so much time studying the normal distribution?

A Some variables are normally distributed; a bigger reason is the “Central Limit Theorem” (we will get to get that later)

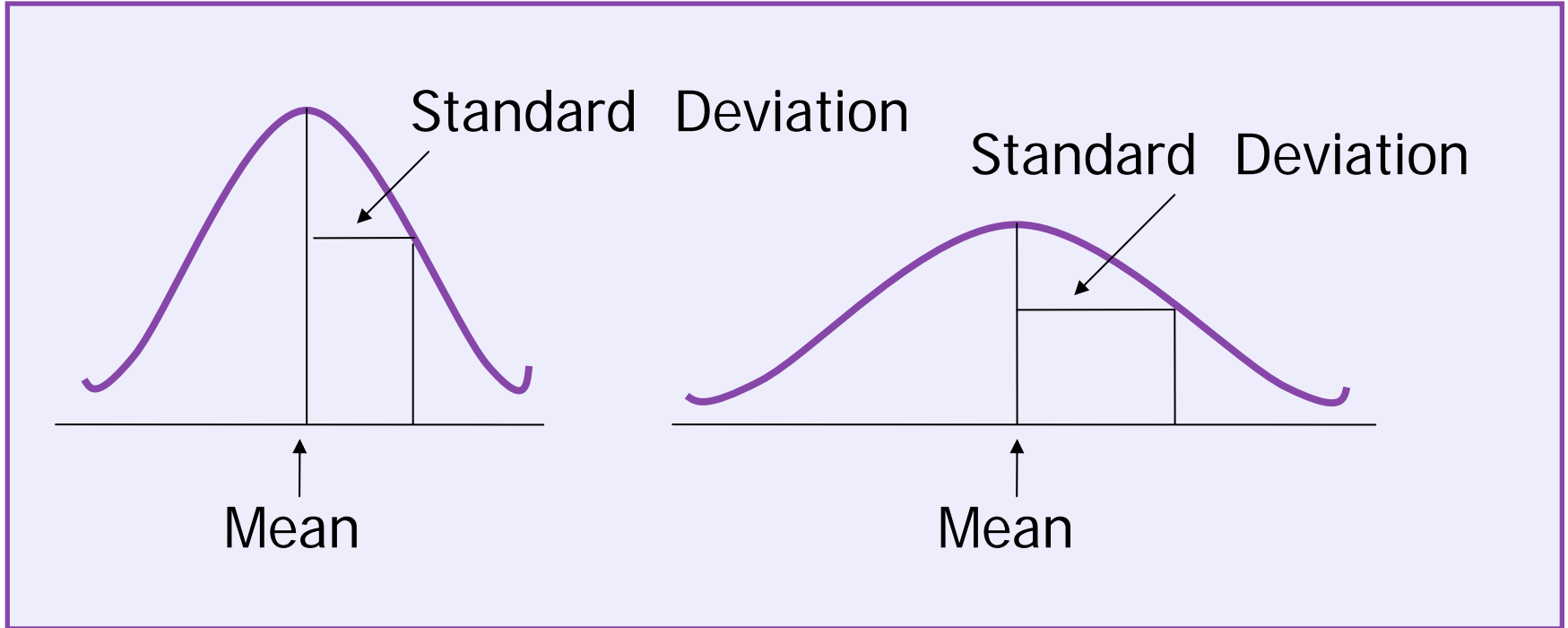
Normal Distribution

- ◆ There are lots of normal distributions!
 - Symmetric
 - Bell-shaped
 - Mean = Median

Normal Distribution

- ◆ You can tell which normal distribution you have by knowing the mean and standard deviation
 - The mean is the center
 - The standard deviation measures the spread (variability)

Two Different Normal Distributions



Describing Variability

- ◆ How can we describe the spread of the distribution?
- ◆ Minimum and maximum
range = Max – Min
- ◆ Sample standard deviation
(abbreviated *s* or *SD*)

Describing Variability

- ◆ Sample variance (s^2)
- ◆ Sample standard deviation (s or SD)
- ◆ The sample variance is the average of the square of the deviations about the sample mean

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

Describing Variability

- ◆ The sample standard deviation is the square root of s^2

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Describing Variability

- ◆ Example: $n = 5$ systolic blood pressures (mm Hg)

$$X_1 = 120$$

$$X_2 = 80$$

$$X_3 = 90$$

$$X_4 = 110$$

$$X_5 = 95$$

Describing Variability

- ◆ Example: $n = 5$ systolic blood pressures (mm Hg)
- ◆ Recall, from last lecture: $\bar{X} = 99$ mm HG
- ◆ Now:

$$\sum_{i=1}^5 (X_i - \bar{X})^2 = (120 - 99)^2 + (80 - 99)^2 + (90 - 99)^2 + (110 - 99)^2 + (95 - 99)^2$$

Describing Variability

- ◆ Example: $n = 5$ systolic blood pressures (mm Hg)

→
$$\sum_{i=1}^5 (X_i - \bar{X})^2 = (21)^2 + (-19)^2 + (-9)^2 + (11)^2 + (-4)^2$$

→
$$\sum_{i=1}^5 (X_i - \bar{X})^2 = (441) + (361) + (81) + (121) + (16)$$

→
$$\sum_{i=1}^5 (X_i - \bar{X})^2 = 1020$$

Describing Variability

- ◆ Sample variance

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{1020}{4} = 255$$

- ◆ Sample standard deviation (*SD*)

$$s = \sqrt{s^2} = \sqrt{255} = 15.97 \text{ (mm Hg)}$$

Notes on s

- ◆ The bigger s is, the more variability there is
- ◆ s measures the spread about the mean
- ◆ s can equal 0 only if there is no spread
 - All n observations have the same value

Notes on s

- ◆ The units of s are the same as the units of the data (for example, mm Hg)
- ◆ Often abbreviated SD
- ◆ s^2 is our best estimate of the population variance σ^2

Notes on s

◆ Interpretation

- “Most” of the population will be within about two standard deviations of the mean
- For a normally (**Gaussian**) distributed population, “most” is about 95%

Why Do We Divide by $n-1$ Instead of n ?

- ◆ We really want to replace \bar{X} with μ in the formula for s^2

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

- ◆ Since we don't know μ , we use \bar{X}

Why Do We Divide by $n-1$ Instead of n ?

- ◆ But generally, $(X_i - \bar{X})^2$ tends to be smaller than $(X_i - \mu)^2$
 - To compensate, we divide by a smaller number: $n-1$ instead of n

$n-1$

- ◆ $n-1$ is called the *degrees of freedom of the variance* or *SD*
- ◆ Why?
 - The sum of the deviations is zero
 - The last deviation can be found once we know the other $n-1$
 - Only $n-1$ of the squared deviations can vary freely

$$n-1$$

- ◆ The term *degrees of freedom* arises in other areas of statistics
- ◆ It is not always $n-1$, but it is in this case

Other Measures of Variation

- ◆ Standard deviation (SD or s)
- ◆ Minimum and maximum observation
- ◆ Range = maximum – minimum

Other Measures of Variation

- ◆ **What happens to these as sample size increases? Do they . . .**
 - Tend to increase?
 - Tend to decrease?
 - Remain about the same?

Other Measures of Variation

- ◆ **What happens to the max and min as sample size increases?**
 - Let's first tackle the maximum and minimum!
 - As it turns out, as sample size increases, the maximum tends to increase, and the minimum tends to decrease

Other Measures of Variation

- ♦ **What happens to the range as sample size increases?**
 - Extreme values are more likely with larger samples!
 - This will tend to increase the range

Other Measures of Variation

- ♦ **What happens to the mean as sample size increases?**
 - Because extreme values, both larger and smaller, are both more likely in larger samples they “balance each other out”

Other Measures of Variation

- ◆ **What happens to the mean as sample size increases?**
 - This “balancing” act tends to keep the mean in a “steady state” as sample size increases—it tends to be about the same

Other Measures of Variation

- ◆ **What happens to the SD as sample size increases?**
 - SD tends to stay the same as sample size increases by the same reasoning
 - Remember, it is a measure of variability about the mean, and the mean does not change by much with bigger samples

Other Measures of Variation

- ◆ **What happens to the SD as sample size increases?**
 - Because larger extremes are as likely as smaller extremes, the average squared deviation about the mean tends to stay the same, and therefore the SD stays about the same



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

Practice Problems

Practice Problems

- ◆ Let's revisit the data on annual income (on \$1000s of U.S. dollars) taken from a random sample of nine students in the Hopkins Internet-based MPH program

37 102 34 12 111 56 72 17 33

Practice Problems

37 102 34 12 111 56 72 17 33

1. Calculate the sample variance and standard deviation
2. What would happen to our estimate of standard deviation if the 111 were replaced with 132?



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

Practice Problem Solutions

Solutions

37 102 34 12 111 56 72 17 33

1. Calculate the sample variance and standard deviation
 - Recall, from last time: $\bar{X} = 52.7$

Solutions

- ◆ Recall:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

- ◆ For this data:

$$s^2 = \frac{\sum_{i=1}^9 (X_i - 52.7)^2}{8}$$

Solutions

- ◆ For this data:

$$s^2 = \frac{10,128}{8} = 1,266$$

(You should verify this for practice)

- ◆ So: $s = \sqrt{1266} = 35.6$. (thousands of \$)

Solutions

2. What would happen to our estimate of standard deviation if the 111 were replaced with 132?
 - Here, we are increasing the maximum in our data set while keeping sample size the same

Solutions

2. What would happen to our estimate of standard deviation if the 111 were replaced with 132?
 - We are not “balancing” this increase with a reduction in other values
 - This will cause our SD to increase—prove it to yourself!



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section B

*Variability in the Normal
Distribution;
Calculating Normal Scores*

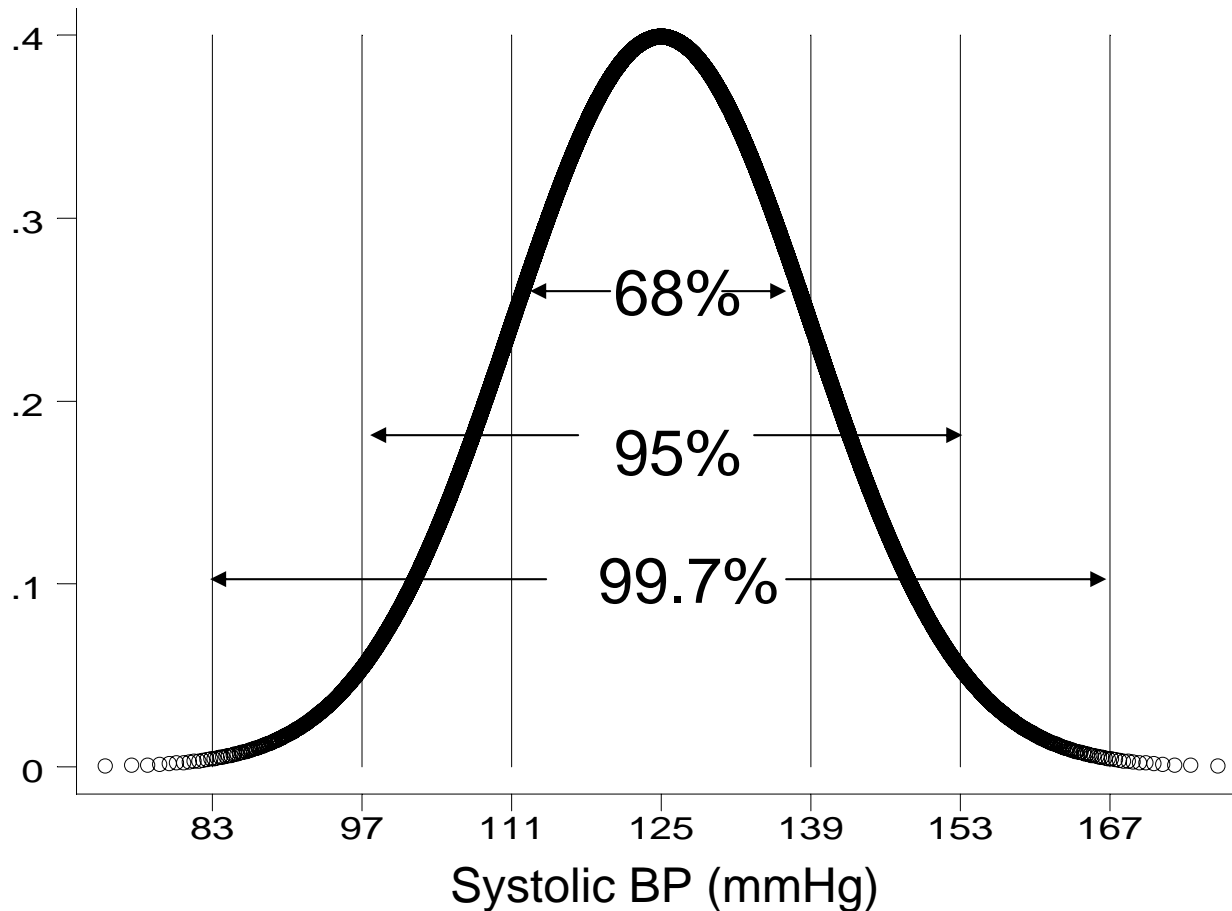
The 68-95-99.7 Rule for the Normal Distribution

- ◆ 68% of the observations fall within one standard deviation of the mean
- ◆ 95% of the observations fall within two standard deviations of the mean
- ◆ 99.7% of the observations fall within three standard deviations of the mean

Distributions of Blood Pressure

Approximately normal mean = 125 mmHG
Standard deviation = 14 mmHG

Distributions of Blood Pressure



The 68-95-99.7 rule applied to the distribution of systolic blood pressure in men.

Distributions of Blood Pressure

- ◆ The rule says that if a population is normally distributed, then approximately 68% of the population will be within 1 SD of \bar{x}
- ◆ It doesn't guarantee that exactly 68% of your sample of data will fall within 1 SD of \bar{x}

Standard Normal Scores

- ◆ How many standard deviations away from the mean are you?

$$\text{Standard Score (Z)} = \frac{\textit{Observation} - \textit{mean}}{\textit{Standard deviation}}$$

Standard Normal Scores

A standard score of . . .

- ◆ **Z = 1:** The observation lies one SD above the mean
- ◆ **Z = 2:** The observation is two SD above the mean

Standard Normal Scores

A standard score of . . .

- ◆ **Z = -1:** The observation lies 1 SD below the mean
- ◆ **Z = -2:** The observation lies 2 SD below the mean

Standard Normal Scores

- ◆ Example: Male Blood Pressure, mean = 125, s = 14 mmHg
 - BP = 167 mmHg

$$Z = \frac{167 - 125}{14} = +3.0$$

Standard Normal Scores

– BP = 97 mmHg

$$Z = \frac{97 - 125}{14} = -2.0$$

What is the Usefulness of a Standard Normal Score?

- ◆ It tells you how many SDs (s) an observation is from the mean
- ◆ Thus, it is a way of quickly assessing how “unusual” an observation is

What is the Usefulness of a Standard Normal Score?

- ◆ *Example:* Suppose the mean BP is 125 mmHg, and standard deviation = 14 mmHg
 - Is 167 mmHg an unusually high measure?
 - If we know $Z = 3.0$, does that help us?

Above and Below the SD

- ◆ The following table tells you what percent of the population lies above or below a number of standard deviations from the mean, assuming the population has a normal distribution

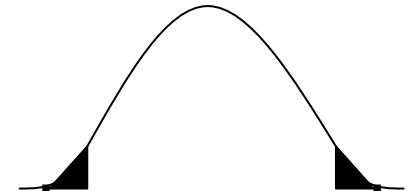
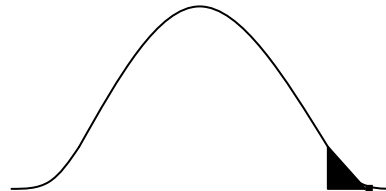
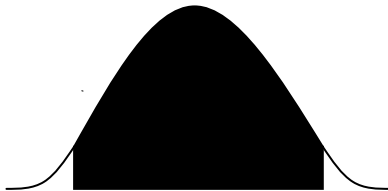
Fraction of Population

Within Z SDs
of the mean

More than Z
SDs above
the mean

More than Z
SDs above or
below the mean

Z



1.0	68.27%
2.0	95.45%
2.5	98.76%
3.0	99.73%

15.87%
2.28%
0.62 %
0.13%

31.73%
4.55%
1.24%
0.27%

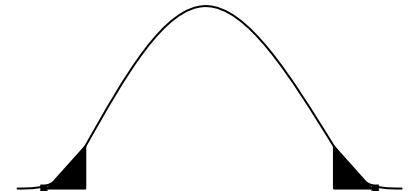
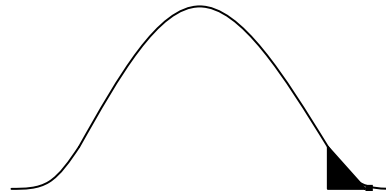
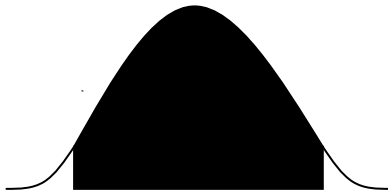
Fraction of Population

Within Z SDs
of the mean

More than Z
SDs above
the mean

More than Z
SDs above or
below the mean

Z



1.0	68.27%
2.0	95.45%
2.5	98.76%
3.0	99.73%

15.87%
2.28%
0.62 %
0.13%

31.73%
4.55%
1.24%
0.27%

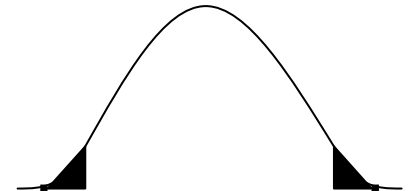
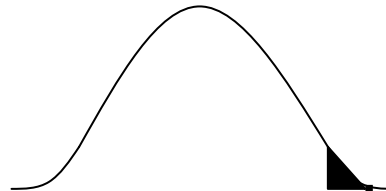
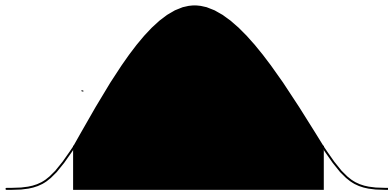
Fraction of Population

Within Z SDs
of the mean

More than Z
SDs above
the mean

More than Z
SDs above or
below the mean

Z



1.0	68.27%
2.0	95.45%
2.5	98.76%
3.0	99.73%

15.87%
2.28%
0.62 %
0.13%

31.73%
4.55%
1.24%
0.27%

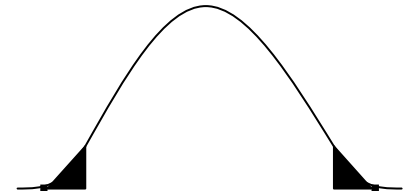
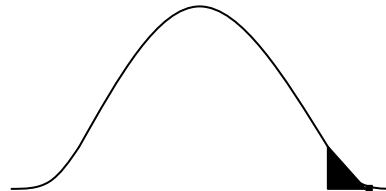
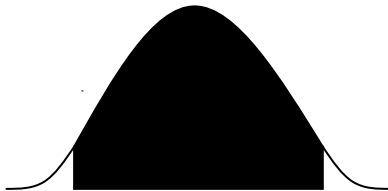
Fraction of Population

Within Z SDs
of the mean

More than Z
SDs above
the mean

More than Z
SDs above or
below the mean

Z



1.0	68.27%
2.0	95.45%
2.5	98.76%
3.0	99.73%

15.87%
2.28%
0.62 %
0.13%

31.73%
4.55%
1.24%
0.27%

Why Do We Like The Normal Distribution So Much?

- ◆ The truth is, there is nothing “special” about standard normal scores
 - These can be computed for observations from any sample/population of continuous data values
 - The score measures how far an observation is from its mean in standard units of statistical distance

Why Do We Like The Normal Distribution So Much?

- ◆ However, unless population/sample has a well known, “well behaved” distribution, we may not be able to use mean and standard deviation to create interpretable intervals, or measure “unusuality” of individual observations

Hospital Length of Stay Example

- ◆ Random sample of 1,000 patients
 - Mean length of stay—5.1 days
 - Median length of stay—3 days
 - Standard deviation—6.4 days

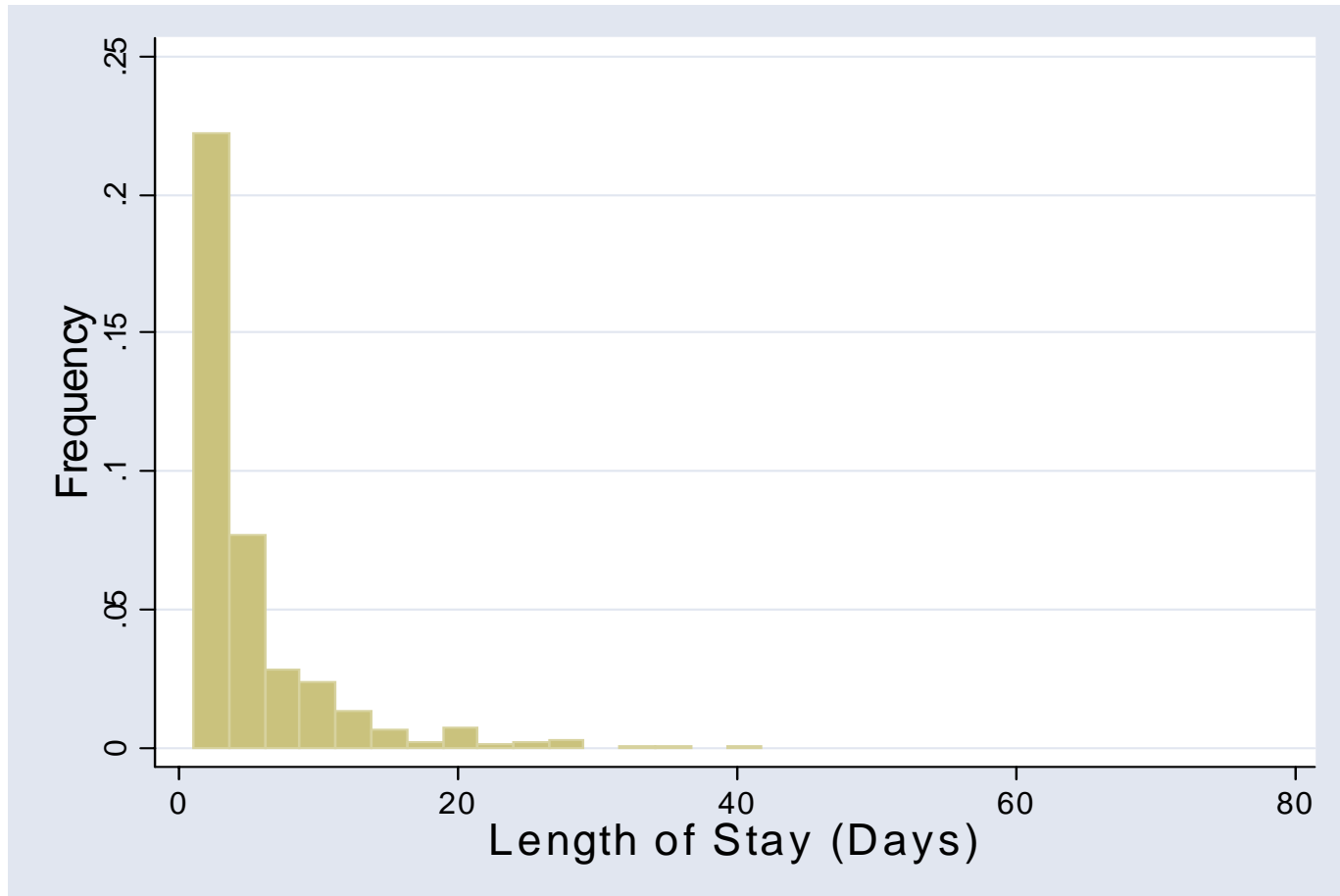
Hospital Length of Stay Example

- ◆ The data was entered into Stata

```
. list hospstay in 1/10
```

```
+-----+  
| hospstay |  
+-----+  
1.         3  
2.         2  
3.         4  
4.         3  
5.        11  
+-----+  
6.         2  
7.         4  
8.         1  
9.         1  
10.        6  
+-----+
```

Hospital Length of Stay Example



Constructing Intervals

- ◆ Suppose I wanted to estimate an interval containing roughly 95% of the values of hospital length of stay in the population
- ◆ Distribution right skewed—can not appeal to properties/methods of normal distribution!

Summary Statistics

- ◆ The summarize command
 - Syntax “summarize varname”

```
. summarize hospstay
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hospstay	1000	5.084	6.368792	1	75

Summary Statistics

- ◆ The summarize command
 - Syntax “summarize varname”

```
. summarize hospstay
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hospstay	1000	5.084	6.368792	1	75

Summary Statistics

- ◆ The summarize command
 - Syntax “summarize varname”

```
. summarize hospstay
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hospstay	1000	5.084	6.368792	1	75

Summary Statistics

- ◆ The summarize command
 - Syntax “summarize varname”

```
. summarize hospstay
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hospstay	1000	5.084	6.368792	1	75

Summary Statistics

- ◆ The summarize command with the “detail” option
 - Syntax “summarize varname, detail”

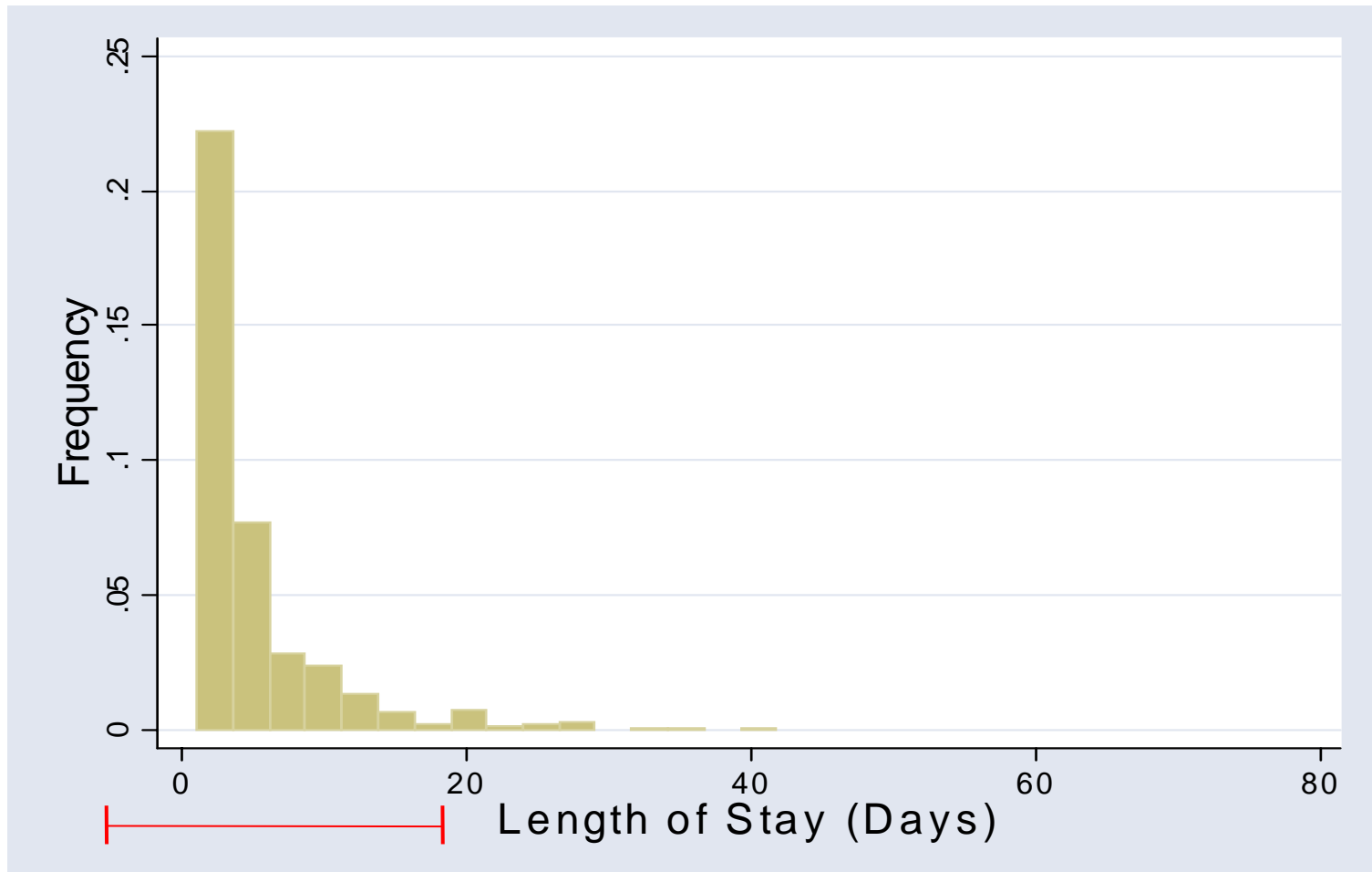
```
. summarize hospstay, detail
```

```
-----  
                    hospstay  
-----  
Percentiles      Smallest  
1%                1          1  
5%                1          1  
10%               1          1      Obs                1000  
25%               1          1      Sum of Wgt.         1000  
  
50%               3  
75%               6          40  
90%               12         42      Mean                5.084  
95%               17.5        47      Std. Dev.           6.368792  
99%               31         75      Variance            40.56151  
                        Variance            40.56151  
                        Skewness            3.672524  
                        Kurtosis            25.09711
```

Constructing Intervals

- ◆ Mean \pm 2SD's
 - $5.1 \pm 2 * 6.4$
 - This gives an interval from -7.7 to 17.9 days!

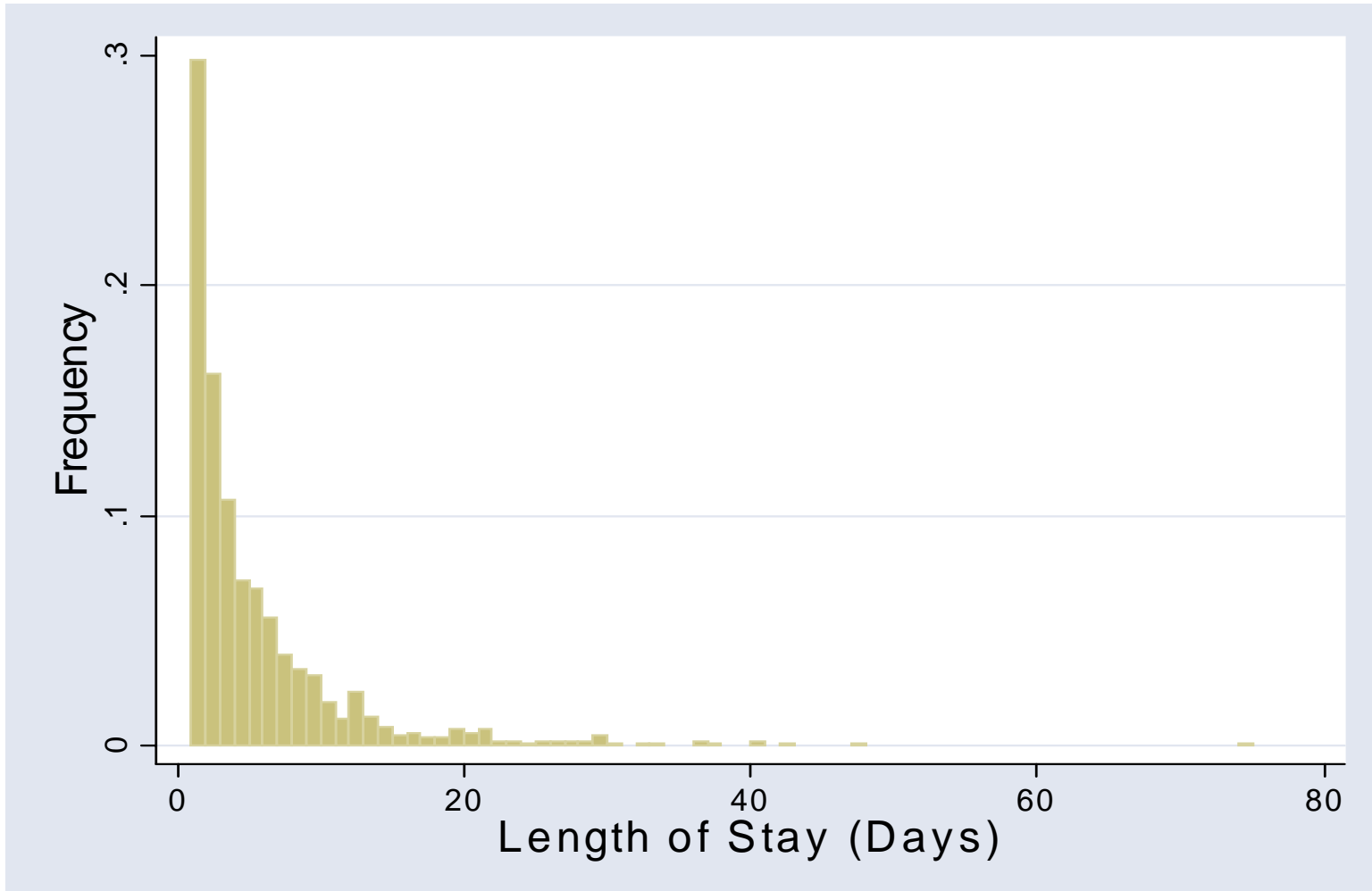
Constructing Intervals



Constructing Intervals

- ◆ We would need to estimate this interval from the histogram and/or by finding sample percentiles

Constructing Intervals



Constructing Intervals

- ◆ Using percentiles
 - Syntax "centile varname, c(#1, #2, . . .)

```
. centile hospstay, c(2.5, 97.5)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf. Interval]	
hospstay	1000	2.5	1	1	1
		97.5	23	20.35954	28

Constructing Intervals

◆ Using percentiles

```
. centile hospstay, c(2.5, 97.5)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf. Interval]	
hospstay	1000	2.5	1	1	1
		97.5	23	20.35954	28

Constructing Intervals

◆ Using percentiles

```
. centile hospstay, c(2.5, 97.5)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf. Interval]	
hospstay	1000	2.5	1	1	1
		97.5	23	20.35954	28



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section B

Practice Problems

Practice Problems

- ◆ Suppose a population is normally distributed (and you are a member of the population)
- 1. If you have a standard score of $Z = 2$, what percentage of the population has a score *greater* than your score?

Practice Problems

2. If you have a standard score of $Z = -2$, what percentage of the population has a score *greater* than your score?
3. If you have a standard score of $Z = 1$, what percentage of the population has a score *less* than your score?

Practice Problems

4. Suppose the distribution of grades in your statistics class is normal, with mean = 83.4, $s = 7$. There is a total of 120 students in the class. If you score a 97.4 in the class, roughly how many people have scores higher than your score?

Practice Problems

5. Suppose we call unusual observations those that are either at least 2 SD above the mean or about 2 SD below the mean. What percent is unusual? In other words, what percent of the observation will have a standard score either $Z > + 2.0$ or $Z < - 2.0$? What percentage would have $|Z| > 2$?

Practice Problems

- ◆ The results of these exercises will turn out to be very important later in our discussion of p-values!



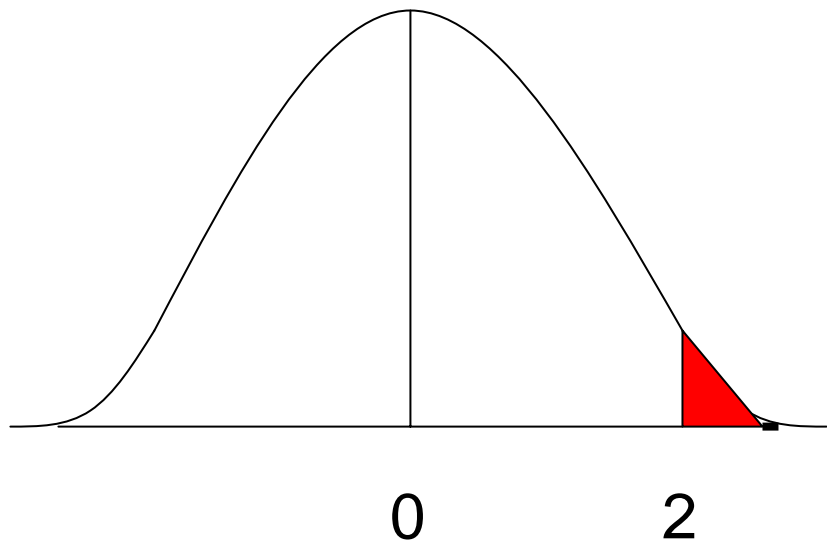
JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section B

Practice Problem Solutions

Solutions

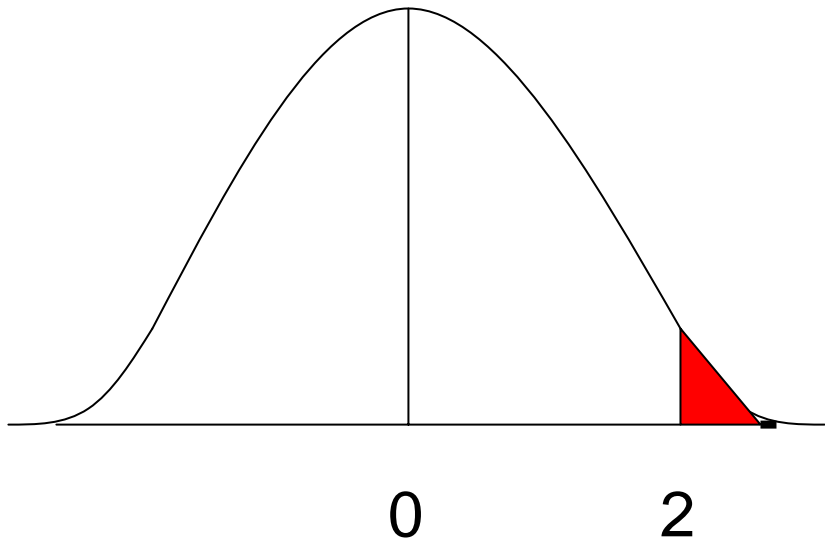
1. If you have a standard score of $Z = 2$, what percentage of the population has a score greater than your score?



The area of the red portion is our answer!

Solutions

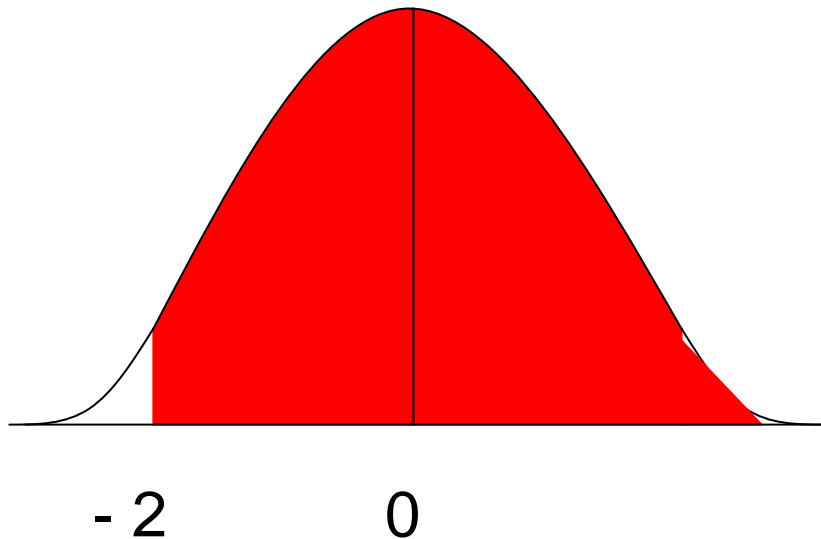
1. If you have a standard score of $Z = 2$, what percentage of the population has a score greater than your score?



From the table, we know this is 2.28% of the population.

Solutions

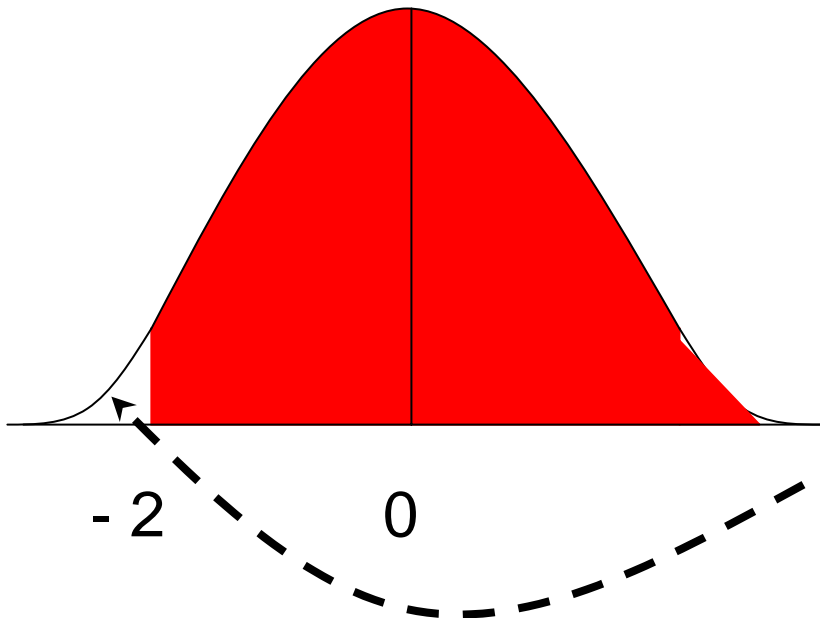
2. If you have a standard score of $Z = -2$, what percentage of the population has a score greater than your score?



Again, the area of the red portion is our answer!

Solutions

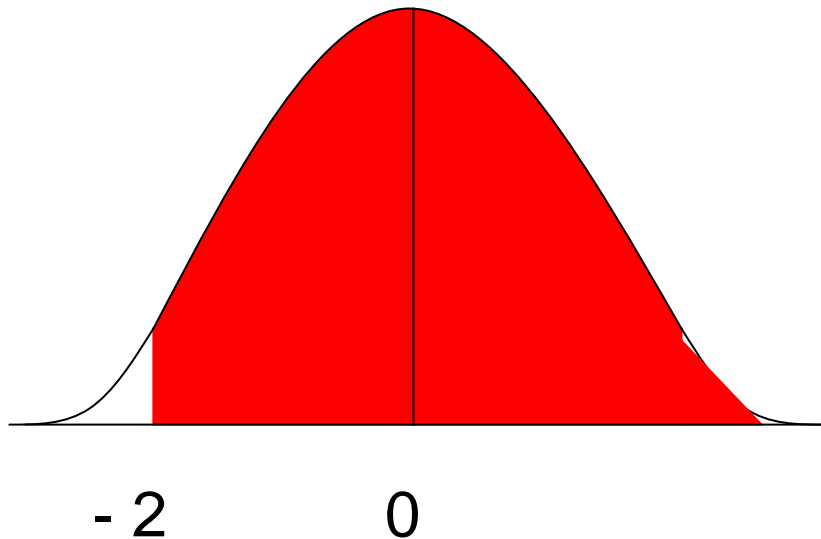
2. If you have a standard score of $Z = -2$, what percentage of the population has a score greater than your score?



From our table, we can only get the area of the white portion on the left side (2.28%).

Solutions

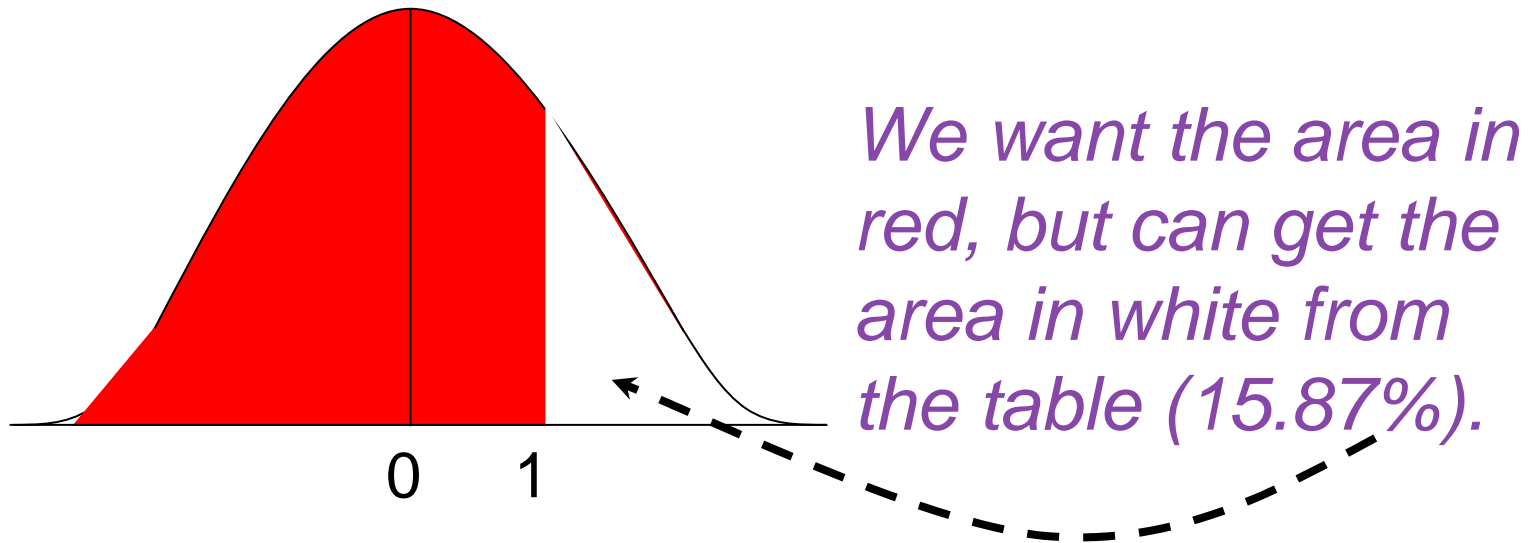
2. If you have a standard score of $Z = -2$, what percentage of the population has a score greater than your score?



Because the total area under the curve is 100%, the area in red is just 100%—2.28% = 97.72%.

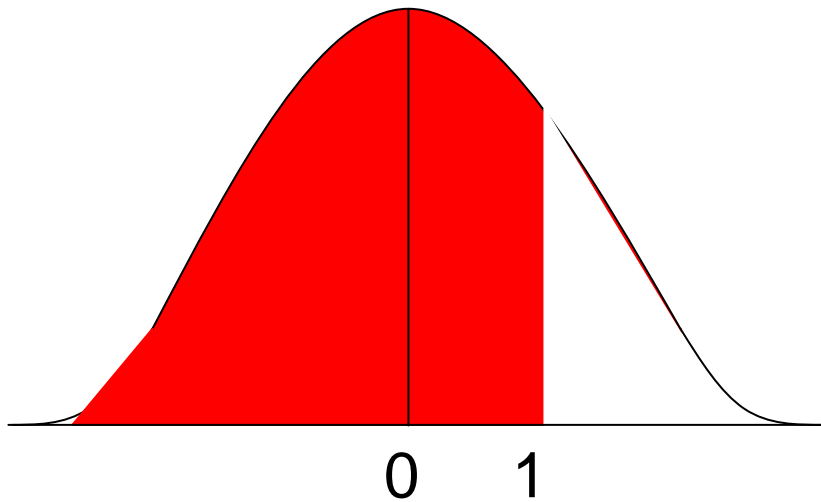
Solutions

3. If you have a standard score of $Z = 1$, what percentage of the population has a score less than your score?



Solutions

3. If you have a standard score of $Z = 1$, what percentage of the population has a score less than your score?



*The area in red is 100%:
15.87% = 84.13%.
Approximately 84% of the
population have scores
less than you.*

Solutions

4. Suppose the distribution of grades in your statistics class is normal, with mean = 83.4, $s = 7$. There is a total of 120 students in the class. If you score a 97.4 in the class, roughly how many people have scores higher than your score?

$$Z = \frac{\text{Observed} - \text{Mean}}{\text{sd}} = \frac{97.4 - 83.4}{7} = \frac{14}{7} = 2$$

Solutions

4. (Continued)

- If you have a standard score of 2, we know that 2.3% of the population has a score greater than your score (and therefore a higher exam score)
- There are 120 people in the class, so about $(.023) * (120) = 2.76 \approx 3$ people have higher scores. Good job!

Solutions

5. Suppose we call unusual observations those that are either at least 2 SD above the mean or about 2 SD below the mean. What percent is unusual? In other words, what percent of the observation will have a standard score either $Z > + 2.0$ or $Z < - 2.0$? What percent would have $|Z| > 2$?

Solutions

5. (Continued)

- We know from the table that $4.55\% \approx 5\%$ of the observations are “unusual” by this definition
- We will revisit this idea *many* times in our upcoming discussion of p-values



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section C

Sampling Variability

Population versus Sample

- ◆ The *population* of interest could be
 - All women between ages 30–40
 - All patients with a particular disease
- ◆ The sample is a small number of individuals from the population
 - The sample is a subset of the population

Population versus Sample

- ◆ *Sample mean* (\bar{X}) versus *population mean* (μ)
 - For example, mean blood pressures
 - We know the sample mean \bar{X}
 - We don't know the population mean μ , but we would like to

Population versus Sample

- ◆ *Sample proportion* versus *population proportion*
 - For example, proportion of individuals with health insurance
 - We know the sample proportion (for example, 80%)
 - We don't know the population proportion

Population versus Sample

- ◆ **Key Question:**

- How close is the *sample mean* (or proportion) to the *population mean* (or proportion)?

The Population and the Sample

- ◆ A parameter is a number that describes the population
 - A parameter is a fixed number, but in practice we do not know its value
 - Example: Population mean
Population proportion

The Population and the Sample

- ◆ A statistic is a number that describes a sample of data
 - A statistic can be calculated
 - We often use a statistic to estimate an unknown parameter
 - Example: **Sample mean**
Sample proportion

How accurate are the sample statistics for estimating the population parameter?

Sources of Error

- ◆ **Errors from biased sampling**
 - The study systematically favors certain outcomes
 - Voluntary response
 - Non-response
 - Convenience sampling
 - Solution: Random sampling

Sources of Error

- ◆ **Errors from (random) sampling**
 - Caused by chance occurrence
 - Get a “bad” sample because of bad luck (by “bad” we mean not representative)
 - Can be controlled by taking a larger sample

Sources of Error

- ◆ Using mathematical statistics, we can figure out how much potential error there is from random sampling (standard error)

Potentially Biased Sampling

- ◆ **Example:** Blood pressure study of population of women age 30–40
 - *Volunteer*
 - Non-random; selection bias
 - *Family members*
 - Non-random; not independent
 - *Telephone survey; random-digit dial*
 - Random or non-random sample?

Potentially Biased Sampling

- ◆ **Example:** Clinic population, 100 consecutive patients
 - Random or non-random sample?
 - Convenience samples are sometimes assumed to be random

Potentially Biased Sampling

- ◆ **Example:** 1936 *Literary Digest* poll of presidential election—Landon vs Roosevelt
 - *Election result:* 62% voted for Roosevelt
 - *Digest prediction:* 43% voted for Roosevelt

Sampling Bias

◆ Selection bias

- Mail questionnaire to 10 million people
- Sources: Telephone books, clubs
- Poor people are unlikely to have telephone (only 25% had telephones)

Sampling Bias

◆ Non-response bias

- Only about 20% responded (2.4 million)
- Responders different than non-responders

Bottom Line

- ◆ When a selection procedure is biased, taking a larger sample does not help
 - This just repeats the mistake on a larger scale
- ◆ Non-respondents can be very different from respondents
 - When there is a high non-response rate, look out for non-response bias

Random Sample

- ◆ When a sample is randomly selected from a population, it is called a *random sample*
- ◆ In a simple random sample, each individual in the population has an equal chance of being chosen for the sample

Random Sample

- ◆ Random sampling helps control systematic bias
- ◆ But even with random sampling, there is still *sampling variability* or *error*

Sampling Variability

- ◆ If we repeatedly choose samples from the same population, a statistic will take different values in different samples

Idea

- ◆ If you repeat the study and the statistic does not change much (you get the same answer each time), then it is fairly reliable (not a lot of variability)

Example

- ◆ Estimate the proportion of persons in a population who have health insurance
- ◆ Choose a sample of size $N = 978$
- ◆ Sample 1

$$n = 978 \quad P = \frac{812}{978} = .8302$$

Example

- ◆ Is the sample proportion reliable?
 - If we took another sample of another 978 persons, would the answer bounce around a lot?

Example

- ◆ This tells us how “close” the sample statistic should be to the population parameter