Stat 250 Gunderson Lecture Notes
1: Summarizing Data

> **" You must never tell a thing. You must illustrate it. We learn through the eye and not the noggin."**
>
> --*Will Rogers (1879 - 1935)*

"Simple summaries of data can tell an interesting story and are easier to digest than long lists." So we will begin by looking at some data.

Raw Data

Raw data correspond to numbers and category labels that have been collected or measured but have not yet been *processed* in any way. On the next page is a set of RAW DATA - information about a group of items in this case, individuals. The data set title is DEPRIVED and has information about a **sample size** of ***n*** = 86 college students. For each student we are provided with their answer to the question: "Do you feel that you are sleep deprived?" (yes or no), and their self reported typical amount of sleep per night (in hours). The information we have is organized into **variables**. In this case these 86 college students are a subset from a larger population of all college students, so we have **sample data**.

Definition:

A **variable** is a characteristic that differs from one individual to the next.

Sample data are collected from a subset of a larger population.

Population data are collected when all individuals in a population are measured.

A **statistic** is a summary measure of sample data.

A **parameter** is a summary measure of population data.

Types of Variables

We have 2 variables in our data set. Next we want to distinguish between the different types of ariables - different types of variables provide different kinds of information and the type will guide what kinds of summaries (graphs/numerical) are appropriate.

*University of Michigan Open Courses - Creative commons License

Think about it:

- Could you compute the "AVERAGE AMOUNT OF SLEEP" for these 86 students?
- Could you compute the "AVERAGE SLEEP DEPRIVED STATUS" for these 86 students?

SLEEP DEPRIVED STATUS is said to be a Categorical variable,

AMOUNT OF SLEEP is a quantitative variable.

> ***Definitions:***
>
> A **categorical** variable places an individual or item into one of several groups or categories. When the categories have an ordering or ranking, it is called an **ordinal** variable.
>
> A **quantitative** variable takes numerical values for which arithmetic operations such as adding and averaging make sense. Other names for quantitative variable are: **measurement** variable and **numerical** variable.

Try It! –

For each variable listed below, give its type as categorical or quantitative.

Activity 1A

- Age (years)
- Typical Classroom Seat Location (Front, Middle, Back)
- Number of songs on an iPod
- Time spent studying material for this class in the last 24-hour period (in hours)
- Soft Drink Size (small, medium, large, super-sized)
- The "And then ..." count recorded in a psychology study on children (details will be provided)

> ***Looking ahead:*** Later, when we talk about random variables, we will discuss whether a variable is modeled *discretely* (because its values are countable) or whether it would be modeled *continuously* (because it can take any value in an interval or collection of intervals). Go back through the list above and think about is it *discrete* or *continuous*?

Our data set is somewhat large, containing a lot of measurements in a long list. Presented as a table listing, we can view the record of a particular college student, but it is just a listing, and not easy to find the largest value for the amount of sleep or the number of students who felt they are sleep deprived. We would like to learn appropriate ways to summarize the data.

Summarizing Categorical Variables

Numerical Summaries

How would you go about summarizing the SLEEP DEPRIVED STATUS data? The first step is to simply count how many individuals/items fall into each category. Since percents are generally more meaningful than counts, the second step is to calculate the percent (or proportion) of individuals/items that fall into each category.

Sleep Deprived?	Count	Percent
Yes	51	59.3%
No	35	40.7%
Total		

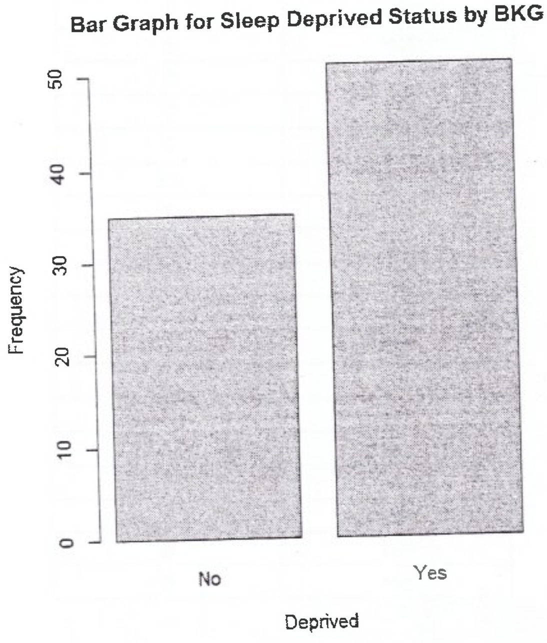
The table above provides both the **frequency distribution** and the **relative frequency distribution** for the variable SLEEP DEPRIVED STATUS.

Visual Summaries

There are two simple visual summaries for categorical data – a bar graph or a pie chart. Here is the table summary and bar graph made with R.

```
counts:
Deprived
No Yes
35  51

percentages:
Deprived
No  Yes
40.7 59.3
```



Aside: Does it matter whether the 'No' or 'Yes' bar is given first?

Pie Chart: Another graph for categorical data which helps us see what part of

DATA SET = DEPRIVED

Feel Sleep Deprived?	Amount Sleep per Night (hours)
No	9
No	7
No	8
Yes	7
Yes	7
Yes	8
Yes	7
Yes	8
No	10
No	8
No	9
No	8
Yes	8
Yes	4
Yes	6
No	8
No	10
No	4
Yes	7
Yes	8
No	9
No	9
No	7
Yes	8
No	9
No	9
No	8
No	6
No	9
Yes	7
Yes	11
Yes	7
No	9
Yes	7
No	8
Yes	7
Yes	7
Yes	9
Yes	1
Yes	7
Yes	6
No	8
Yes	6
No	8
No	7
No	9
Yes	7
Yes	7
Yes	7
Yes	7
Yes	6
No	8
Yes	6
No	9
No	8
Yes	7
Yes	8
No	8
No	8
Yes	7
Yes	7
Yes	7
No	7
Yes	7
Yes	8
No	7
Yes	7
Yes	7
Yes	7
Yes	8
Yes	6
Yes	6
Yes	8
No	9
No	7
Yes	8
Yes	6
Yes	7
Yes	8
Yes	5
Yes	6
No	7
No	8
Yes	8
Yes	7
Yes	6

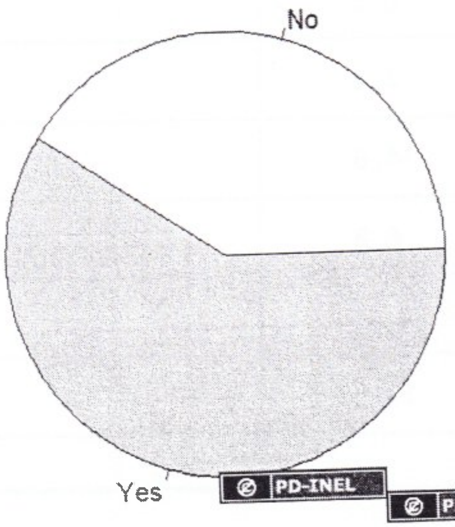
the whole each group forms.

Pie charts are not as easy to draw by hand. It is not as easy to compare sizes of pie pieces versus comparing heights of bars.

Thus we will prefer to use a bar graph for categorical data.

Recap: We have discussed that some variables are categorical and others are quantitative. We have seen that bar graphs and pie charts can be used to display data for categorical variables. We turn next to displaying the data for <u>**quantitative**</u> variables.

Pie Chart for Sleep Deprived Status by BKG



Summarizing Quantitative Data with Pictures

Recall our Sleep Deprived Data for n = 86 college students. We have data on two variables: sleep deprived status and hours of sleep per night. How would you go about summarizing the sleep hours data? These measurements do vary. How do they vary? What is the range of values? What is the pattern of variation?

- Find the smallest value = ___4___ and largest value = ___11___
- Take this overall range and break it up into intervals (of equal width).
  What might be reasonable here?
  Perhaps by 2's; but we need to watch the endpoints.

* Activity 1B

Summary Table:

Class	Frequency (or count)	Relative Frequency (or proportion)	Percent
0, 2			
2, 4			
4, 6			
6, 8			
8, 10			
10, 12			

Graph for quantitative data = Histogram:

Note: each bar represents a class, and the base of the bar covers the class.

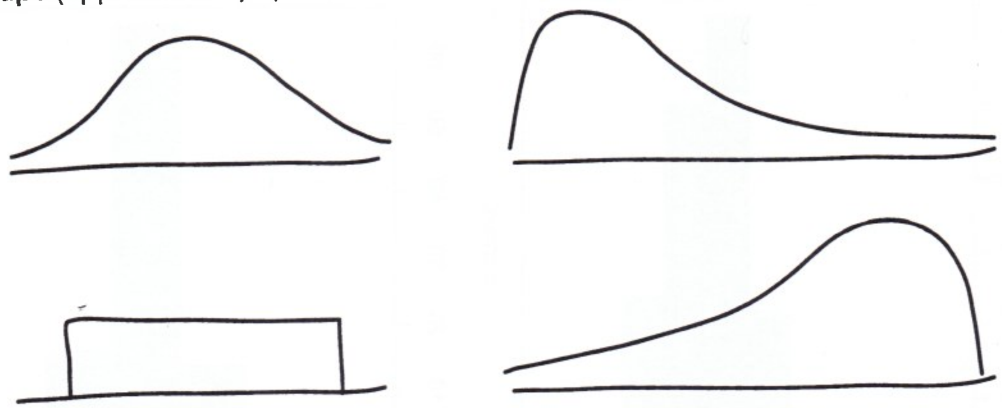
The above table and histogram show the **distribution of this quantitative variable** *SLEEP HOURS*, that is, the overall pattern of how often the possible values occur.

How to interpret?

- **Look for Overall Pattern**

  Three summary characteristics of the overall distribution of the data ...

  Shape (approximately symmetric, skewed, bell-shaped, uniform)



Location (center, average)

Spread (variability)

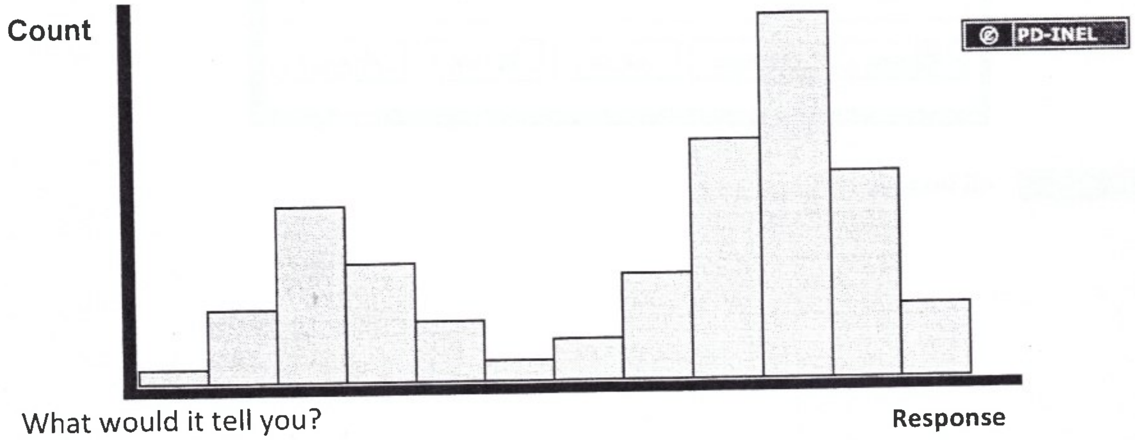
- **Look for deviations from Overall Pattern**

  Outliers = a data point that is not consistent with the bulk of the data.

  Outliers should not be discarded without justification.

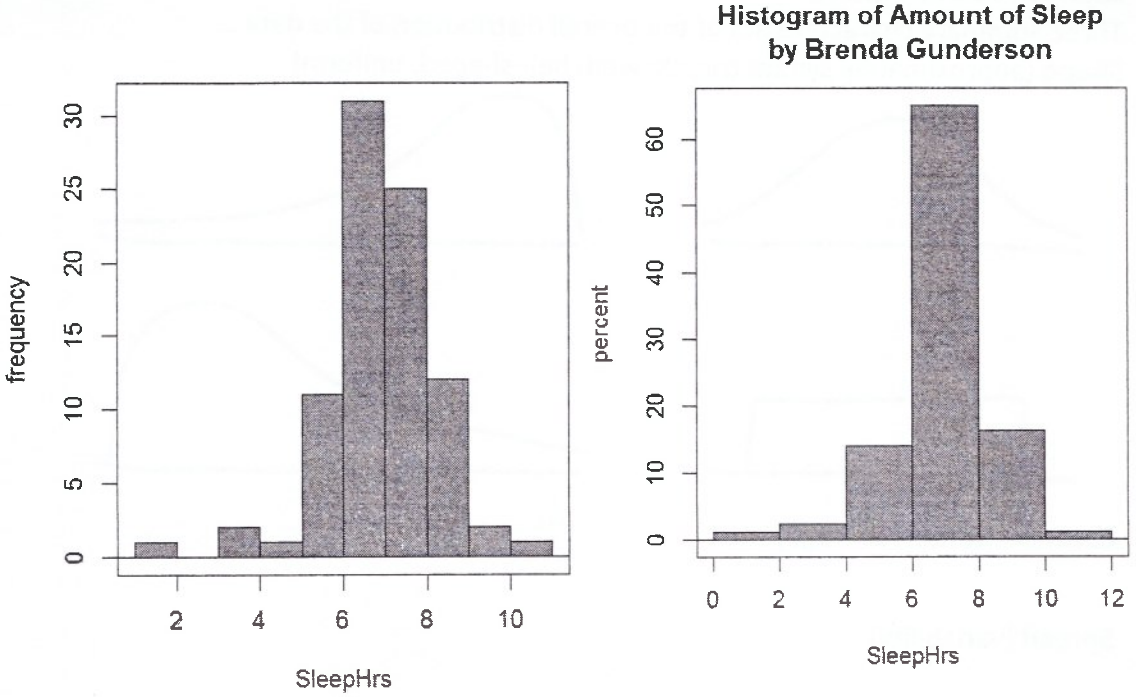
Describe the distribution for SLEEP HOURS:

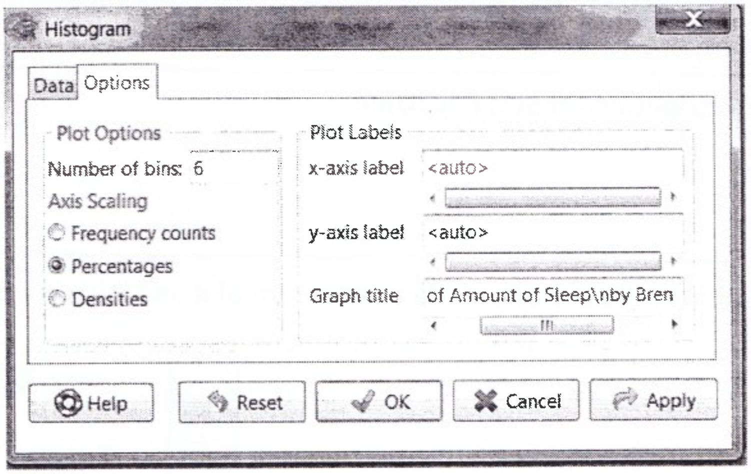
What if ... you had some data and you made a histogram of it and it looked like this...



What would it tell you?

R Histograms (default on the left and customized on the right):




Histogram

Data | Options

Plot Options

Number of bins: 6

Axis Scaling

○ Frequency counts

● Percentages

○ Densities

Plot Labels

x-axis label <auto>

y-axis label <auto>

Graph title of Amount of Sleep\nby Bren

Numerical Summaries of Quantitative Variables

We have discussed some interesting features of a quantitative data set and learned how to look for them in pictures (graphs). Section 2.5 focuses on numerical summaries of the center and the spread of the distribution (appropriate for quantitative data only).

Notation for a generic raw set of data:

$x_1, x_2, x_3, \ldots, x_n$ **where n = # items in the data set or sample size**

Describing the Location or Center of a Data Set

Two basic measures of location or center:

- **Mean** -- the numerical average value
  We represent the ***mean of a sample*** (called a statistic) by ...

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

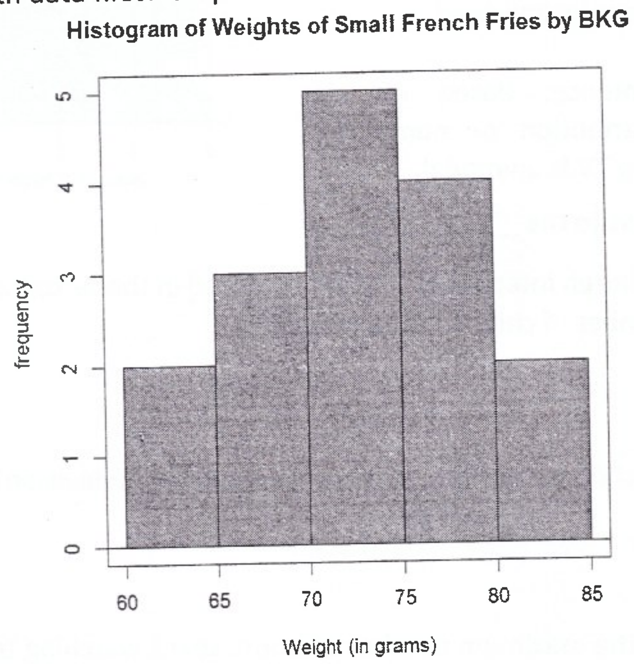
- **Median** -- the middle value when data arranged from smallest to largest.

Try It! French Fries

Weight measurements for 16 small orders of French fries (in grams).

78	72	69	81	63	67	65	75
79	74	71	83	71	79	80	69

What should we do with data first? Graph it!



Based on our histogram, the distribution of weight is unimodal and approximately symmetric, so computing numerical summaries is reasonable. The weights (in grams) range from the 60's to the lower 80's, centered around the lower 70's, with no apparent outliers.

Other comments –

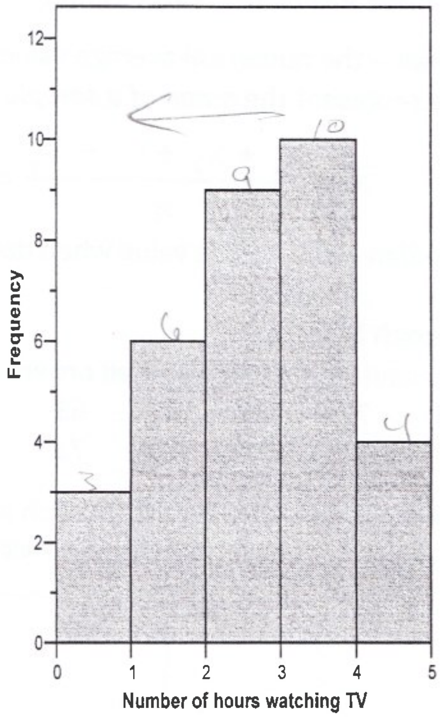
- NO SPACE BETWEEN BARS! Unless there are no observations in that interval.
- How Many Classes? Use your judgment: generally somewhere between 6 and 15 intervals.
- Better to use relative frequencies on the y axis when comparing two or more sets of observations.
- Software has defaults and many options to modify settings.

One More Example:

A study was conducted in Detroit, Michigan to find out the number of hours children aged 8 to 12 years spent watching television on a typical day.

A listing of all households in a certain housing area having children aged 8 to 12 years was first constructed. Out of the 100 households in this listing, a random sample of 20 households was selected and all children aged 8 to 12 years in the selected households were interviewed.

The following histogram was obtained for all the children aged 8 to 12 years interviewed.



Activity 1C

a. Complete the sentence: Based on this histogram, the distribution of number of hours spent watching TV is unimodal,

with a slight skewness to the ___________.

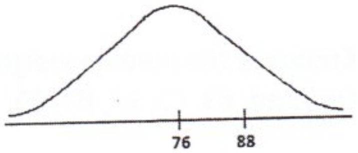
b. Assuming that all children interviewed are represented in the histogram, what is the total number of children interviewed?

c. What proportion of children spent 2 hours or less watching television?

d. Can you determine the maximum number of hours spent watching television by one of the interviewed children? If so, report it. If not, explain why not.

Describing Spread: Range and Interquartile Range

Midterms are returned and the "average"
was reported as 76 out of 100.
You received a score of 88.
How should you feel?



Often what is missing when the "average" of something is reported, is a corresponding measure of spread or variability. Here we discuss various measures of variation, each useful in some situations, each with some limitations.

Range: Measures the spread over 100% of the data.
Range = High value – Low value = Maximum – Minimum

Percentiles: The p^{th} percentile is the value such that p% of the observations fall at or below that value.

Some Common percentiles:

Median:	50th percentile
First quartile:	25th percentile
Third quartile:	75th percentile

Five Number Summary:

	Variable Name and Units **(*n* = number of observations)**		
Median		M	
Quartiles	Q1		Q3
Extremes	Min		Max

Provides a quick overview of the data values and information about the center and spread. Divides the data set into approximate quarters.

Interquartile Range: Measures the spread over the middle 50% of the data. **IQR = Q3 – Q1**

Try it! French Fries Data
Ordered: 63, 65, 67, 69, 69, 71, 71, 72, 74, 75, 78, 79, 79, 80, 81, 83

Activity 1E

Find the five-number summary:

	Weight of Fries (in grams) **(*n* = 16 orders)**		
Median			
Quartiles			
Extremes			

Range: **IQR:**

Activity 1D

1. **Compute the mean weight.**

2. **Compute the median weight.**
   Ordered: 63, 65, 67, 69, 69, 71, 71, 72, 74, 75, 78, 79, 79, 80, 81, 83

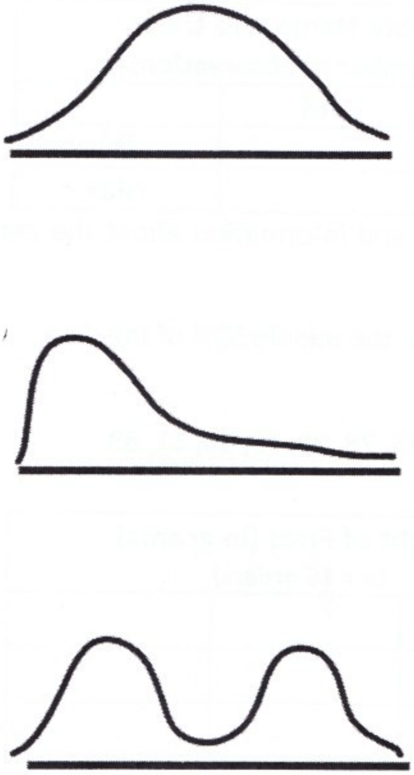
3. **What if the smallest weight was incorrectly entered as 3 grams instead of 63 grams?**

Note: The **mean** is Sensitive to extreme observations.

The **median** is insensitive to extreme observations.

Most graphical displays would have detected such an outlying value.

Some Pictures: Mean versus Median



And confirming these values using R we have:

```
> numSummary(FrenchFries[,"Weight"], statistics=c("mean", "sd", "IQR",
+    "quantiles"), quantiles=c(0,.25,.5,.75,1))
 mean       sd IQR 0% 25% 50% 75% 100%  n
 73.5 6.0663  10 63  69  73  79   83 16
```

Activity 1F

Example: Test Scores

The five-number summary for the distribution of test scores for a very large math class is provided below:

	Test Score (points) (*n* = 1200 students)		
Median		58	
Quartiles	46		78
Extremes	34		95

1. What is the test score interval containing approximately the lowest ¼ of the students?

2. Suppose you scored a 46 on the test. What can you say about the percentage of students who scored higher than you?

3. Suppose you scored a 50 on the test. What can you say about the percentage of students who scored higher than you?

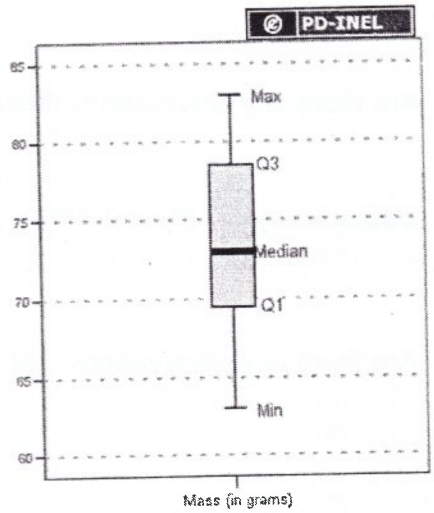
4. If the top 25% of the students received an A on the test, based on this summary, what was the minimum score needed to get an A on the test?

Boxplots

A boxplot is a graphical representation of the five-number summary.

Steps:

- **Label an axis with values to cover the minimum and maximum of the data.**
- **Make a box with ends at the quartiles Q1 and Q3.**
- **Draw a line in the box at the median *M*.**
- **Check for possible outliers using the 1.5*IQR rule and if any, plot them individually.**
- **Extend lines from end of box to smallest and largest observations that are not possible outliers.**



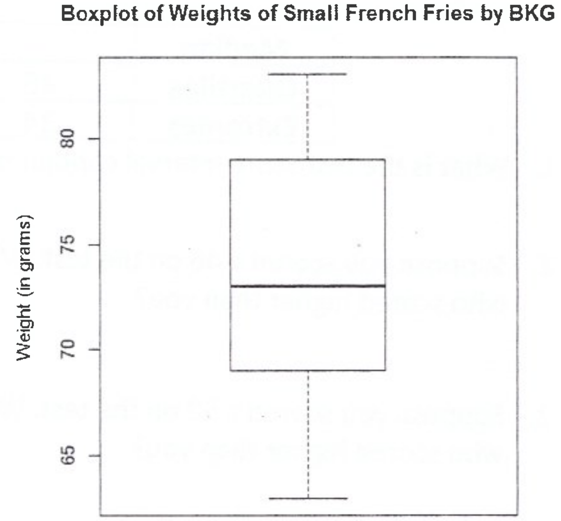
Note: Possible outliers are observations that are more than 1.5*IQR outside the quartiles, that is, observations that are below Q1 - 1.5*IQR or observations that are above Q3 + 1.5*IQR.

Try it! French Fries Data

Ordered: 63, 65, 67, 69, 69, 71, 71, 72, 74, 75, 78, 79, 79, 80, 81, 83

The five-number summary:

	Weight of Fries (in grams) (*n* = 16 orders)		
Median		73	
Quartiles	69		79
Extremes	63		83



Boxplot of Weights of Small French Fries by BKG

From the boxplot shown, we see there are no points plotted separately, so there are no outliers by the 1.5(IQR) rule.

Verify there are no outliers using this rule.

IQR = 79 – 69 = 10 grams

1.5*IQR = 1.5 (10) = 15 grams

*Activin 16

Lower boundary (fence) = Q1 - 1.5*IQR =

Are there any observations that fall below this lower boundary?

Upper boundary (fence) = Q3 + 1.5*IQR =

Are there any observations that fall above this upper boundary?

What if ... the largest weight of 83 grams was actually 95 grams?

Ordered: 63, 65, 67, 69, 69, 71, 71, 72, 74, 75, 78, 79, 79, 80, 81, <u>95</u>

Then the five number summary would be:

	Weight of Fries (in grams) (*n* = 12 orders)		
Median		73	
Quartiles	69		79
Extremes	63		

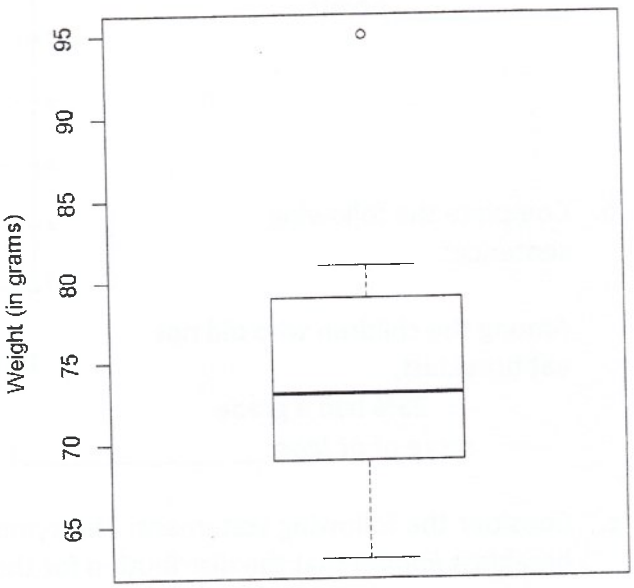
The IQR and 1.5*IQR would be the same, so the "boundaries" for checking for possible outliers are again 54 and 94.

Now we would have one potential high outlier, the maximum value of 95.

The modified boxplot when we have this one outlier is shown.

Why is the line extending out on the top side now drawn out to just 81?

Boxplog of Weights (with one outlier) by BKG



Notes on Boxplots:

- **Side-by-side boxplots are good for ...**

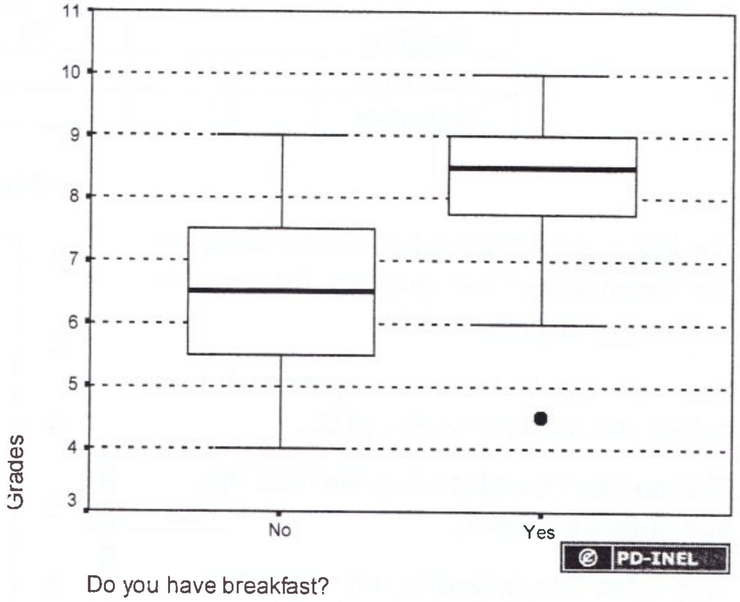
- **Watch out - points plotted individually are ...**

- **Can't confirm ....**

- **When reading values from a graph show what you are doing (so appropriate credit can be given on exam/quiz).**

* Activity 1H *

Try It: Side-by-side boxplots
A random sample of 100 parents of grade-school children were recently interviewed regarding the breakfast habits in their family. One question asked was if their children take the time to eat a breakfast (recorded as breakfast status – Yes or No). The grades of the children in some core classes (e.g. reading, writing, math) were also recorded and a standardized grade score (on a 10-point scale) was computed for each child. Side-by-side boxplots of the children's standardized grade scores are provided.



a. What is (approx) the lowest grade scored by a child who **does** have breakfast?

__________ points

b. Complete the following sentence:

Among the children who did **not** eat breakfast,
25% had a grade score of ***at least*** ____________ points.

c. Consider the following statement: The symmetry in the boxplot for the children **not** eating breakfast *implies* that the distribution for the grade scores of such students is bell-shaped.

True or False?

Features of Bell-Shaped Distributions

We have already described the distribution of our french fries weight data as being unimodal and somewhat symmetric. If we were to draw a curve to smooth out the tops of the bars of the histogram, it would resemble the shape of a bell, and thus could be called **bell-shaped.**

One fairly common distribution of measurements with this shape has a special name, called a **normal distribution** or **normal curve**. We will see normal curves in more detail when we study random variables. When a distribution is somewhat bell-shaped (unimodal, indicating a fairly homogeneous set of measurements), a useful measure of spread is called the **standard deviation.** In fact, the mean and the standard deviation are two summary measures that completely specify a normal curve.

Describing Spread with Standard Deviation

When the mean is used to measure center, the most common measure of spread is the standard deviation. The standard deviation is a ***measure of the spread of the observations from the mean***. We will refer to it as a kind of "average distance" of the observations from the mean. But it actually is the square root of the average of the squared deviations of the observations from the mean. Since that is a bit cumbersome, we like to **think of the standard deviation as "roughly, the average distance the observations fall from the mean."** Here is a quick look at the formula for the stand deviation when the data are a sample from a larger population:

s = sample standard deviation = $\sqrt{\dfrac{\Sigma (x_i - \bar{x})^2}{n-1}}$

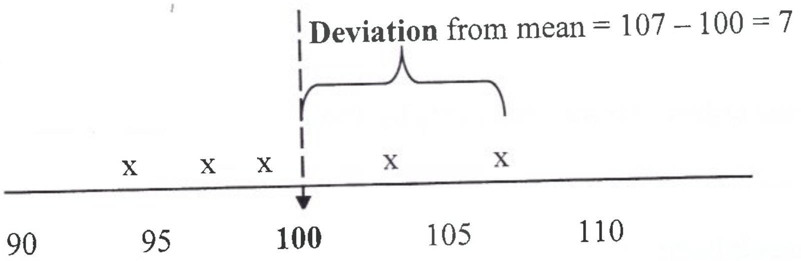
x_i = observation i

$\bar{X}$ = sample mean

n = num of observations

Note: The squared standard deviation, denoted by **s^2**, is called the **variance**. We emphasize the standard deviation since it is in the original units.

Example: Consider this sample of *n* = 5 scores: 94, 97, 99, 103, 107.
The sample mean is 100 points. Let's measure their *spread* by considering how much each score ***deviates* from the mean**. Then consider the "average of these deviations".



How the calculations are done:

x	x – 100	$(x – 100)^2$	**Calculations**
94	-6	36	1. $\bar{x}$ = 500/5 = 100 (used in columns 2 & 3) 2. Variance: s^2 = 104/(5-1) = 26 3. **Standard deviation: s = √(26) = 5.1** Note # of *deserved* decimals used for s.
97	-3	9	
99	-1	1	
103	3	9	
107	7	49	
500	**0**	**104**	← Sums (or totals) of the columns

Activity II

Try it! French Fries Data

Weight measurements for 16 small orders of French fries (in grams).

78	72	69	81	63	67	65	75
79	74	71	83	71	79	80	69

The mean was computed earlier to be 73.5. **Find the standard deviation** for this data.

s =

Not much fun to do it by hand, but not too bad for a small number of observations. In general, we will have a calculator or computer do it for us.

Interpretation:

The weights of small orders of french fries are *roughly*

________________away from their mean weight of __________, *on average*.

OR

On average, the weights of small orders of french fries vary by about ________________

from their mean weight of ____________________.

Notes about the standard deviation:

- ***s* = 0 means ...**

- **Like the mean, *s* is ...**

- **So use the mean and standard deviation for____.**

  The five-number summary is better for ____.

18

- **Technical Note about difference between population and sample:** Datasets are commonly treated as if they represent a sample from a larger population. A numerical summary based on a sample is called a statistic. The sample mean and sample standard deviation are two such statistics. However, if you have all measurements for an entire population, then a numerical summary would be referred to as a parameter.

  The symbols for the mean and standard deviation for a population are different, and the formula for the standard deviation is also slightly different. A population mean is represented by the Greek letter μ ("mu"), and a population standard deviation is represented by the Greek letter σ ("sigma"). The formula for the population standard deviation is below.

  You will see more about the distinction between statistics and parameters in the next chapter and beyond.

Population standard deviation: $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$

where *N* is the size of the population.

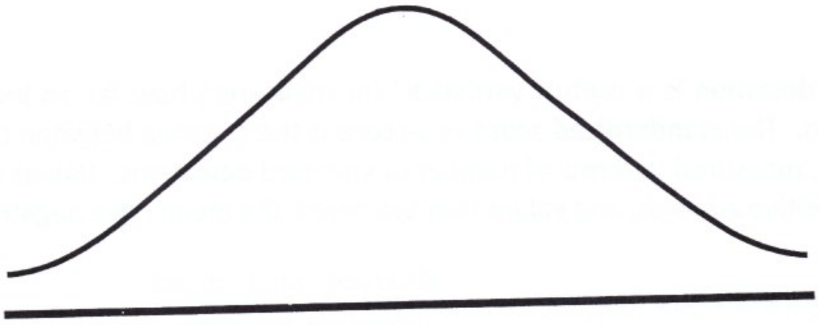
Empirical Rule

For bell-shaped curves, approximately…

68% of the values fall within 1 standard deviation of the mean in either direction.

95% of the values fall within 2 standard deviations of the mean in either direction.

99.7% of the values fall within 3 standard deviations of the mean in either direction.



* Activity 1J *

Try It! Amount of Sleep

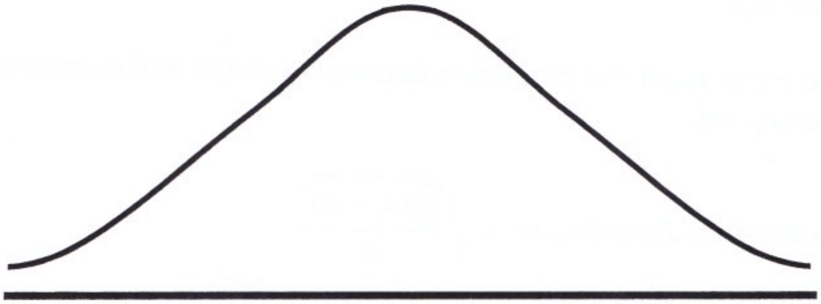
The typical amount of sleep per night for college students has a bell-shaped distribution with a mean of 7 hours and a standard deviation of 1.7 hours.

About 68% of college students typically sleep between ________ and _______ hours per night.

Verify the values below that complete the sentences.

- About 95% of college students typically sleep between **3.6** and **10.4** hours per night.
- About 99.7% of college students typically sleep between **1.9** and **12.1** hours per night.

Draw a picture of the distribution showing the mean and intervals based on the empirical rule.



Suppose last night you slept 11 hours.
How many standard deviations from the mean are you?

Suppose last night you slept only 5 hours.
How many standard deviations from the mean are you?

The standard deviation is a useful "**yardstick**" for measuring how far an individual value falls from the mean. The **standardized score** or ***z*-score** is the distance between the observed value and the mean, measured in terms of number of standard deviations. Values that are above the mean have positive *z*-scores, and values that are below the mean have negative *z*-scores.

Standardized score or *z*-score: $z = \dfrac{\text{observed value - mean}}{\text{standard deviation}}$

* Activity 1k *

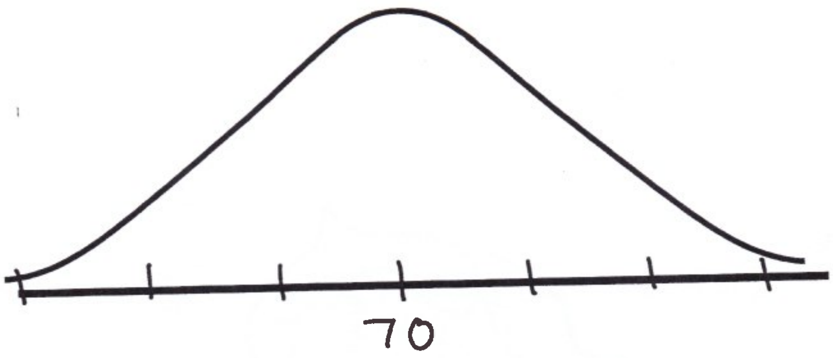
Try It! Scores on a Final Exam

Scores on the final in a course have approximately a bell-shaped distribution.
The mean score was 70 points and the standard deviation was 10 points.

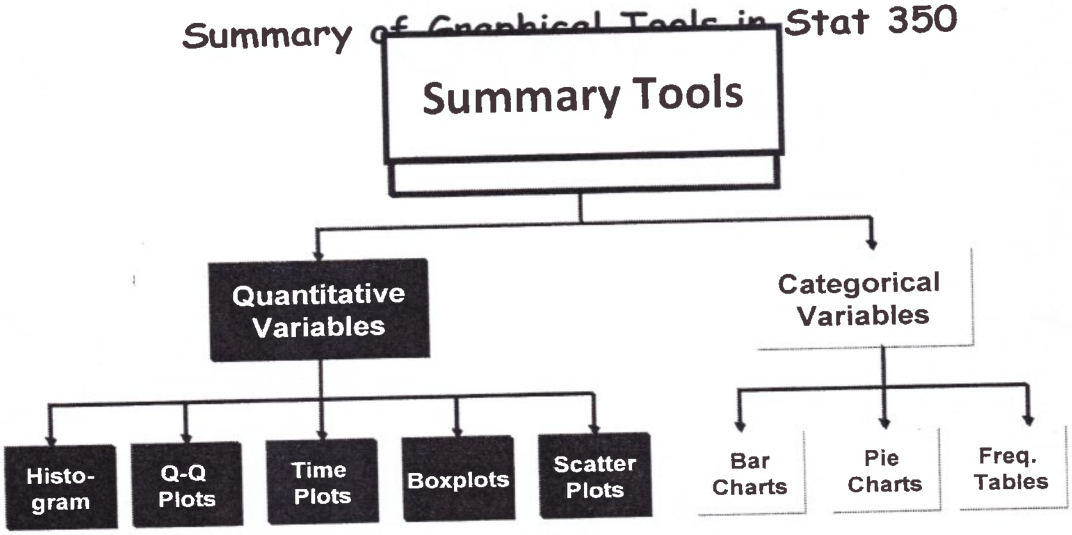
Suppose Rob, one of the students, had a score that was 2 standard deviations above the mean. What was Rob's score?

What can you say about the proportion of students who scored higher than Rob?

Sketching a picture may help.



Summary of Graphical Tools in Stat 350



Summary Tools

Quantitative Variables

Histo-gram

Q-Q Plots

Time Plots

Boxplots

Scatter Plots

Categorical Variables

Bar Charts

Pie Charts

Freq. Tables

Only Cash side get Journal entries

outstanding ch's – from Bank balance