Module 4A

I. Student Version

Title

Comparison of 12-Year-Old Baseball Pitching Speeds: Little League Baseball Regional Tournament vs Select Baseball USA Tournament

Project Summary

This project is designed to quickly introduce beginning kinesiology students to data analysis by comparing the pitching speeds, heights, weights, and years of playing experience of two samples of players during a little league allstar regional tournament versus a select baseball USA tournament. Actual data from a previous study will be compared using readily available descriptive and inferential statistical software. The purpose of this module is to introduce beginning students to the power and ease of using statistics to answer questions in kinesiology by inserting a two-class session analysis experience their freshman year. Furthermore, this analysis module is designed to be used by instructors who are not specialists in measurement and statistical analysis.

Instructions to Students

Background Information:

The Sample Data: The sample data for this module is from actual measures of 12-Year-old baseball players in a previous year. At the time this data was collected, there was much discussion about the abilities of good players in little league versus select league tournaments. It might be interesting to students that the Little League Regional Tournament sampled contained a team that was eventually runner-up national champion and third in the world in the Little League World Series. Additionally, tournaments held at the Baseball USA facility in Houston, Texas were, at the time, considered among the best in the nation in all youth age groups. It should be noted that good players from the select leagues rarely chose to also play Little League ball. Pitching speeds were measured with ®Jugs speed guns during a minimum of four innings pitched. Pitches considered "off-speed" were not included when averaging speeds. It should be noted that pitching conditions were somewhat different for the select leagues due to "holding" runners on base and pitching at slightly greater distances. Heights, weights, and years pitching were obtained in consultation with the coaches of the players selected to be timed. Players participating in the study were randomly selected from tournament teams and were not chosen based upon ability.

A Quick Guide to **Describing Data** with Statistics for this Project: A quick way to describe and compare a group of numbers such as pitching speeds and other selected player data, is to use one number in the middle of the data which best represents the group. These statistics are called measures of "*central tendency*," and include the *mean, median, and mode*. For this project the "***mean***" should be used because we are assuming the data is “normal” and the mean best represents all of the scores in the group. Using the ***descriptive statistics*** tool in the ***data analysis add-in*** for MS Excel® as described below under the *data collection* heading you can have the computer do the heavy calculating and provide you with the means for the performances of both groups you tested. An additional piece of information, the *variability* or “spread” of the performances in each group, is necessary to describe the data more completely. For this project, the “***standard deviation***” should be used to describe the variability in each group of performances. Once again, the ***descriptive statistics*** tool in the ***data analysis add-in*** for MS Excel® as described below under the *data collection* heading provides you with the needed statistic. So, you should use both statistics to help answer the question of which baseball league had faster pitching speeds, larger players, etc. It is important to note that the *mean* and *standard deviation* will also be used in the calculation of the inferential statistics used to determine the likelihood that the results from these two samples of players occurred by chance.

*A Quick Guide to the **Inferential Statistics** Used in This Module:*

The *t-test* (t) for Significant Differences: Quick Definition and Interpretation

A ***t-test*** is used to determine if there is a significant difference between two sample means. It is a difference that is greater than would be expected to occur due to chance alone. Every time you jump, for example, you would not expect to get the exact same height. If more than two groups are examined, then Analysis of Variance (ANOVA) must be used because a t-test cannot be repeated on multiple groups and still keep chance factors constant. This means if you run a bunch of t-tests you are essentially guaranteed to find a difference between two groups eventually, but that difference would reflect something that could happen just by chance.

A significant difference thus means that *the results of the experiment are reliable and there is only a small probability that the results occurred by chance.* Normally, researchers in the behavioral sciences are willing to accept a five-in-one-hundred (.05) probability that the results occurred by chance, using the notation **p ≤ .05** called the **alpha level**.

Similar to how your jump is never quite the same, the means of any two groups will never be exactly the same, even if they are samples drawn from the same population as in experimental research. The question the t-test answers is if the differences are reliable or how great is the probability that they occurred by chance.

The **t statistic** is compared to a **critical value** that it must exceed in order for the difference to be significant.

Excel determines the exact ***p* value** in decimals and it will be ≤ .05 if it meets or exceeds the critical value. Use the critical value for a **one-tailed test** when you have hypothesized a difference in a certain direction (greater or less than the control group); use the critical value for a **two-tailed test** when you have hypothesized that the difference could be in any direction (no prediction of greater or less).

If your results are *not significant*, i.e., the p value is greater than .05, you cannot make claims about the differences between means in your discussion because they are unreliable, with too great a likelihood they occurred by chance.

If your results *are significant*, you can talk about the differences between means because your results are reliable, with a very small probability that they occurred by chance. This does not necessarily mean that you results are "important." You have to determine that by other means not discussed in this unit on t-testing.

The *Pearson Product-Moment Correlation Coefficient (r)* Quick Definition and Interpretation

The **Pearson Product-Moment Correlation Coefficient (*r*)** is the statistic calculated to look at the linear relationship (correlation) between two variables. If examining the relationship between more than two variables, one uses **Multiple Correlation (*R*) or Regression,** which is using correlation to predict some outcome. This course will only focus upon correlation because multiple correlation and regression are extensions of it.

When *r* is calculated, it is interpreted by describing the (1) ***strength*** and (2) ***direction*** of the relationship (positive or negative). The **correlation coefficient (*r*)** is a number that ranges between -1 and +1. The closer the number is to either -1 or +1, the greater the strength of the relationship. The closer to 0 the weaker the relationship. Many statisticians describe a correlation (positive or negative) of .40 to .60 as a **weak relationship**, .60 to .80 as a **moderately strong relationship**, and .80 to 1.0 as a **strong relationship.** Scatterplot charts that have data points that are clustered tightly together around a straight line indicate a stronger relationship between the two variables. Note that this holds true for negative correlations as well.

A positive number means that the two variables vary together in the same direction (as one variable goes up, the other variable goes up or as one variable goes down, the other variable goes down) and a negative number means the two variables vary together in opposite directions (as one variable goes up, the other variable goes down or as one variable goes down, the other variable goes up. For example, if the relationship between intensity of exercise and heart rate is $r = .92$, this would be a *strong, positive relationship*, describing both the *strength* and *direction* of the relationship. Scatterplot charts have data points clustered around a line going from bottom left to upper right for positive relationships and from top left to bottom right for negative relationships.

It is important to understand that correlations do not represent "cause-and-effect" relationships, although strong correlations often lead to experimental research designs that can examine causality directly. This is because of the possibility of unmeasured variables that influence the correlated variables so that they appear to be casually related.

r^2 is a statistic called the ***Coefficient of Determination*** and indicates the amount of change in one variable that is explained by the other variable. Many correlations are designed to answer these types of questions and this is why good correlational research includes reporting both r and r^2 **.**

Another characteristic of correlation coefficients is that in studies with very large numbers of subjects, even very small correlations may be statistically significant. However, just because a correlation is statistically significant, it does not mean that the two variables are strongly related to each other. It remains important to examine the magnitude (i.e., as noted previously, a correlation of .60 to .80 is considered a moderately strong relationship).

If your data follow a curved line on the scatterplot, you have "curvilinear" data and cannot use Pearson Product-Moment Correlation Coefficient (Pearson's r) as it examines linear relationships. This is why it is good practice to always examine scatterplots. A correlation close to "0" could be found with Pearson's r for two variables that have a very strong curvilinear relationship.

Individual or Group Project:

Students will perform the statistical analyses and produce a concise final report individually or in groups according to the course instructor's wishes. If working in groups, make certain each member participates in all aspects of the analysis fully and is able to answer questions posed by the class when the projects are discussed.

Data Collection:

The data set for this project will be posted as a spreadsheet by the instructor. Pitching speeds were measured with ®Jugs speed guns during a minimum of four innings pitched. Pitches considered "off-speed" were not included when averaging speeds. It should be noted that pitching conditions were somewhat different for the select leagues due to "holding" runners on base and pitching at slightly greater distances. Heights, weights, and years pitching were obtained in consultation with the coaches of the players selected to be timed. Players participating in the study were randomly selected from tournament teams and were not chosen based upon ability.

Procedures/Instructions for Analyzing the Data:

The following steps should be followed for analyzing the data:

Descriptive Statistical Analysis software: The statistical analysis tool "**Descriptive Statistics**" included in the ***Data Analysis*** add-in for *MS Excel®* will be used for this project. This is important as the mathematical formulas necessary for computing the specific statistics are found there. Students should find the data analysis tools in *MS Excel* easy to access on and off campus on PC's and most Apple computer versions of *MS Office®*. *MS Excel* does not require institutional technical help or specialized classrooms for student use and students are encouraged to notify the instructor if they anticipate any problems with computing while taking this course. The '*Data Analysis'* add-in may already be activated and found in the far-right hand side of the top ribbon and category buttons as *'Data Analysis'* after opening *Excel* and clicking the *'Data'* button. Because computers do not automatically activate the add-ins, the following steps may be required:

Open Excel | File | Options | Add-Ins (from left list) | Analysis Toolpak | Go | OK

Find the Data Analysis add-in by:

Open Excel | Data | Data Analysis (on right hand side of tools)

Computers that do not allow permanent changes require activation of the *Data Analysis* Add-In every time the computer is turned on (most university computers).

The following instructions should be followed when using the **descriptive statistics** tool in the data analysis add-in to produce the print-out used under the "examples" heading in this document:

Using the **Descriptive Statistics** data analysis tool:

1. Select **Data | Data Analysis** (to open the Data Analysis dialog box)
2. Choose ***Descriptive Statistics*** (to open the Descriptive Statistics dialog box)
3. **In the *Descriptive Statistics* dialog box, enter the appropriate values by highlighting the cells** containing the data plus the first line above the data if you want to automatically use the first line above the data as a label (recommended, and the reason each of these lines is a unique set of letters and numbers on the spreadsheet). Be careful not to include empty cells and cells with symbols other than numbers.
4. **Click the *Columns* Radio Button** to indicate that the data are organized into columns (the standard way to enter data on a spreadsheet)
5. **Check the *Labels in First Row* checkbox** so that Excel does not recognize the cell as a number if you used the first line above the data as a label in step 3 above.
6. **Click the *New Worksheet Ply* radio button** to create a new tabbed sheet at the bottom within the current worksheet, and to send the results to the newly created worksheet (recommended).
7. **Click the *Summary Statistics* checkbox** (and leave the others unchecked)
8. **Click *OK* to close the dialog box.**

Students should use a measure of (1) central tendency and (2) variability to describe differences in pitching speeds and other measures between the leagues. The highlighted numbers in the example are using the best measures and the ones students should use for this project, the *Mean* and *Standard Deviation*. These statistics should be all the student needs to answer the research question (without using the inferential statistics in Module 2). A short description of these two statistics is included in the background information above. Further examples/guidance may be provided by the instructor as needed.

Examples of a data spreadsheet and descriptive statistics printout (not based upon the data in this project):

					Counter-	Isometric
					Movement	Preload
			height	weight	Jump	Jump
	(years)		(inches)	(lbs)	(inches)	(inches)
SUBJ	**AGE**	**GENDER**	**HT**	**WT**	**CMJ**	**IPJ**
1	18	F	66	136	18	16
2	18	M	70	188	26	22

3	20	F	68	120	24	21
4	19	F	66	110	20	20
5	18	M	72	192	24	20
6	19	M	68	168	28	26
7	20	M	69	178	19	20
8	21	F	64	138	17	16
9	19	M	71	210	23	20
10	18	F	71	160	21	19
11	18	M	74	218	22	20
12	18	M	70	191	32	28

CMJ	
Mean	21.00
Standard Error	0.78
Median	21.00
Mode	24.00
Standard Deviation	4.44
Sample Variance	19.68
Kurtosis	0.78
Skewness	0.04
Range	22.00
Minimum	10.00
Maximum	32.00
Sum	672.00
Count	32.00

Inferential and Correlation Procedures

t-test: Running in MS Excel

Use when your sample groups are Independent (separate)

1. Add-in the **Data Analysis** component if it is missing
2. In **Data Analysis** choose **t-Test: Two Sample Assuming Equal Variances**
3. In the **Variable 1 Range Box**, enter data for one sample
4. In the **Variable 2 Range Box**, enter data for the other sample
5. If the cell ranges include column headings, **check** the **Labels** box
6. Type "0" in the **Hypothesized Mean Difference Box**
7. Leave the **.05** in the **Alpha Box**
8. In the **Output Options** Section choose **New Worksheet Ply** and name your sheet **t-test**

9. Click **Okay**

Use when your sample groups are Paired (same subjects)

1. Add-in the **Data Analysis** component if it is missing
2. In **Data Analysis** choose **t-Test: Paired Two Sample for Means**
3. In the **Variable 1 Range Box**, enter data for one sample
4. In the **Variable 2 Range Box**, enter data for the other sample
5. If the cell ranges include column headings, **check** the **Labels** box
6. Type "0" in the **Hypothesized Mean Difference Box**
7. Leave the **.05** in the **Alpha Box**
8. In the **Output Options** Section choose **New Worksheet Ply** and name your sheet **t-test**
9. Click **Okay**

Example print-out using data from another study:

t-Test: Two-Sample Assuming Equal Variances

	D1 Ht	*D2 Ht*
Mean	74.11	72.17
Variance	3.02	5.11
Observations	36.00	36.00
Pooled Variance	4.07	
Hypothesized Mean Difference	0.00	
df	70.00	
t Stat	4.09	
P(T<=t) one-tail	0.0001	
t Critical one-tail	1.67	
P(T<=t) two-tail	0.00	
t Critical two-tail	1.99	

Correlation Using Pearson's *r*: Running in MS Excel

1. Add-in the **Data Analysis** component if it is missing
2. In **Data Analysis** choose **Correlation**
3. In the **Input Range Box,** enter (highlight) the cell range that contains the data. Note: the two variables you are correlating need to be side-by-side in columns
4. *In the **Grouped By**, choose **Columns**
5. If the cell ranges include column headings, **check** the **Labels** box
6. In the **Output Options** Section choose **New Worksheet Ply** and name your sheet **correlation**
7. Click **Okay**

*You will probably need to copy the data columns for the variables you are analyzing in ANOVA on the spreadsheet at a different location so that the columns are adjacent. The software knows these columns are the two variables you are trying to correlate. Do not delete them from where they originated but delete the copies from the spreadsheet once you have finished running the analysis.

In Excel, the ***Regression option is often better than the *Correlation* option because it gives you **r** (change R to r on the printout), **r^2** (change R^2 to r^2), and will create a scatterplot with a **line-of-best-fit** for use in presenting the data.

	D1 VRJMP	*D1 SQT*
D1 VRJMP	1	
D1 SQT	-0.33	1

r = -0.33

r^2 =0.11

Duration:

This project can be completed in two class sessions—one session for locating the sample data and running the *descriptive and inferential statistics* procedure in Excel, and a second session involving individual or group presentations based upon the analysis.

Deliverables and Evaluation:

Each individual student or group will prepare a short written report and deliver a short presentation based upon the report

(1) comparing pitching speeds, heights, weights, and years of experience (descriptive statistics)

(2) probability the results occurred by chance (t-test)

(3) the relationship between height and pitching speed and weight and pitching speed in both leagues combined (correlation)

The report should be no more than 1 page in length and accompanied by a data print-out using the results of the (1) descriptive statistics, (2) t-test, and (3) correlation analysis tools in Excel. The evaluation will be based upon (1) correct measurement values, (2) correct analysis using the results of the descriptive and inferential statistics tools in Excel (print-out), and (3) answers to the research question using logical conclusions drawn from the data.