

### Friendly Introductory Statistics Help (Version 3.0.3)

Presented at the Joint Statistical Meetings, August 2004, Toronto, Canada

Statisfaction Guaranteed (with 95% Confidence)

Copyright @ 2003 Gordon P. Brooks

Contact: brooksg@ohio.edu

Web Site URL: http://oak.cats.ohiou.edu/~brooksg/fish.htm

### Instructor's Manual

Prepared by Holly Raffle

### Table of Contents

Section 1 - Introduction and Features	2
Section 2 – Installation and Setup	3
Section 3 – Using the Instructor's Manual	4
Section 4 - Analyzing One Variable Using a FISH Data File	4
Calculate Measures of Central Tendency Calculate Deviation Scores Calculate Squared Deviation Scores Sum the Squared Deviation Scores Calculate Range, Variance, and Standard Deviation Calculate Z-score for Each Case Calculate Standard Error of the Mean Calculate Confidence Interval of the Mean Histogram Feature	5 7 8 9 11 13 13
Section 5 - Analyzing Two Variables Using a FISH Data File	16
Calculate Means for X and Y Calculate Standard Deviations Calculate Deviation Scores Calculate Cross Products Calculate Covariance Calculate Correlation Calculate Coefficient of Determination (r <sup>2</sup> ) Calculate Regression Slope Calculate Y-Intercept Scatterplot Feature	17 17 18 19 20 21 22 22 24
Section 6 – Using the Data Generator to Teach Statistical Concepts	25
Central Limit Theorem Distribution of the Slope	25 29
Section 7 – References	30
Section 8 – User Agreement	31

## 1 Introduction and Features

Friendly Introductory Statistics Help (FISH) is designed as a free, easy-to-use program for use by both students and instructors. The software's features make it an excellent teaching tool for secondary school, undergraduate and graduate level introductory statistics courses. For teaching purposes, FISH generates sample sets of data from a population defined by the user with a random data generator. This allows instructors to provide tangible demonstrations of abstract concepts such as the central limit theorem and sampling error. Because of FISH also allows users to manually enter data sets, it is appropriate for students to use as a self-paced tutorial package. The step-by-step presentation and the easily accessible "help" buttons allow students to use FISH as an independent study of basic statistical concepts.

FISH allows users to:

<u>Analyze One Variable.</u> This option calculates measures of central tendency, deviation scores, squared deviation scores, sum of the squared deviation scores, variance, range, standard deviation, z-scores, standard error of the mean, and the confidence interval of the mean.

Within this option, the user is able to toggle between the above calculations computed with (N - 1) for sample statistics and (N) for population parameters. Users are also able to toggle between several options for calculating confidence intervals. Using the z-distribution, the following confidence intervals can be calculated: 68% (z = 1.00), 90% (z = 1.645), 95% (z = 1.96), and 99% (z = 2.33). Using the t-distribution with n – 1 degrees of freedom, confidence intervals can be calculated at the following alpha levels:  $\alpha = .001$ ,  $\alpha = .01$ ,  $\alpha = .05$ , and  $\alpha = .10$ .

This option also allows the user to view a graphical representation of the results. Users are able to access frequency tables (with toggles allowing raw scores, z-scores, and both raw and z-scores to be shown) and grouped frequency tables.

When the user chooses to generate random data, the histogram feature is enhanced to allow the user to view the sampling distribution of the mean. From this screen, the user can generate additional samples in multiples of 1, 10, 100 or 10,000. These sample means are then added to the existing sampling distribution of the mean. <u>Analyze Two Variables.</u> This option calculates the means, standard deviations, deviation scores, and cross products for both variables. Additionally, the unstandardized correlation coefficient (covariance), correlation coefficient (Pearson's r), coefficient of determination (R<sup>2</sup>), regression slope and Y-intercept are calculated.

As in the Analyze One Variable option, the user is able to toggle between the above calculations computed with (N - 1) for sample statistics and (N) for population parameters.

This option also allows the user to view a graphical representation of the results. Users are able to access scatterplots (with toggles allowing raw scores, z-scores, and both raw and z-scores to be shown) and have the option to show or hide the mean lines and the regression line. The histogram option also shows the regression equation and the values for the correlation coefficient and the coefficient of determination. Additionally, the user has access to observed, predicted, and residual values, squared deviation scores, model sum of squares  $((\hat{Y} - \overline{Y})^2)$ , as well as calculations of the standard error of the estimate, standard error of the regression slope, and the 95% confidence interval for the regression slope.

When the user chooses to generate random data, the scatterplot feature is enhanced to allow users to view the sampling distribution of the regression coefficient. From this screen, the user can generate additional samples in multiples of 1, 10, 100 or 10,000. These sample regression coefficients are then added to the existing sampling distribution of the regression slope.

### 2 Installation and Setup

FISH operates in any Windows environment. To install FISH:

- 1. Insert the CD-ROM into the proper drive.
- 2. View the contents of the CD-ROM using Windows Explore.
- 3. Click on the FISH icon. The program will launch and you are ready to begin.

# **3** Using the Instructor's Manual

This Instructor's Manual is designed to complement the FISH User's Guide. The manual is intended to present learning activities and pedagogical pearls; it is not intended to be a comprehensive guide for using FISH. For software support and step-by-step instructions, please refer to the FISH User's Guide.

### 4 Analyzing One Variable Using a FISH Data File

To illustrate uses for FISH in the classroom setting, a data file has been included with this Instructor's Manual. The data (SBP FISH\_data) can be found on the CD-ROM. The data set was taken from Lemeshow, 2003 (Applied Regression Analysis Course Notes, 2003 Summer Program in Applied Statistical Methods, p. 4). The data includes observations on systolic blood pressure for a sample of 30 individuals.

Open the file "SBP FISH\_data" and import the data into FISH. The data will appear on the screen as follows:

😻 FISH: Friendly Introductor	y Statistics Help	
File Analysis Options Help		
SET NUMBER OF CASES		Do All Steps Lock Data Sort by Rank View Histogram
N= 30 • I Variable	Z Variables Reset	Case # Data (X)
ANALYSES	0750.0	1 144.000
Enter Data DONE	Calculate Central	2 220.000
or	Tendency	4 145.000
Generate Random Data	Help Do It	F 162.000
STEP 3:	STEP 4:	6 142.000
Calculate Deviation	Calculate Squared	7 170.000
Scores	Deviation Scores	8 124 000
Help Do It	Help Do It	9 158 000
STEP 5	STEP 6	10 154,000
Sum the Squared	Calculate Range	
Deviation Scores	and variance	COLUMN STATISTICS, using variance formulas with N-1=29:
Help Do It	Help Do It	
STEP 7	STEP 8	
Calculate the	Calculate Z-Scores	
Stanuaru Deviation	iur each case	
Help Do It	Help Do It	
STEP 9	STEP 10	
Calculate Standard Error of the Mean	Contidence Interval for the Mean	
Help Do It	Help Do It	

👹 FISH: Friendly Introductor	ry Statistics Help				
File Analysis Options Help					
SET NUMBER OF CASES	S AND VARIABLES	Do All Ste	ps Lock Data	Sort by Rank	View Histogram
N= 30 • 1 Variable	C 2 Variables Reset	Case #	Data (X)		Rank 🔼
ANALYSES		1	144.000		18.500
STEP 1	STEP 2	2	220.000		30.000
Enter Data DONE	Calculate Central	3	138.000		14.000
or		4	145.000		20.000
Generate Handom Data			162.000		26.500
STEP 3:	STEP 4:	6	142.000		16 500
Calculate Deviation	Calculate Squared	7	170.00 Step Two: C	alculate Central 1	Tendency
acores	Deviation acutes	8	124.00 This option	will calculate th	e mean,
Help Dolt	Help Do It	9	158.00 also rank t	he data and calcu	late the
STEP 5	STEP 6	10	154.00 me	dian rank score.	
Sum the Squared Deviation Scores	Calculate Range and Variance	COLUMNIS	TATISTICS using varia	ance formulas with N	M
Help Do It	Help Do It	Sum:	4276 000		
STEP 7	STEP 8	Mean (M):	142 533	2	
Calculate the	Calculate Z-Scores	Median:	141.000	· · · ·	15 500
Standard Deviation	for each Case	Mode:	120.000		
Help Do It	Help Do It				
STEP 9	STEP 10				
Calculate Standard Error of the Mean	Confidence Interval for the Mean				
Help Do It	Help Do It				

### Step Two: Calculate Central Tendency

#### Mean:

The mean is the most common measure of central tendency. It is also called the arithmetic average or the arithmetic mean. The population mean,  $\mu$ , is calculated:

$$\mu = \frac{\sum X}{N}$$

where N indicates the population size.

The sample mean, also called the estimate mean,  $\overline{X}$ , is calculated:

$$\overline{X} = \frac{\sum X}{n}$$

where n indicates the sample size.

Using the systolic blood pressure (SBP) data, students can confirm for themselves the calculation by using the sum score provided by FISH and dividing it by the sample size.

#### Median:

The median is the score that lies in the middle of a distribution when the scores are arranged in increasing order. The median is also called the 50<sup>th</sup> percentile or the score at or below which 50% of the scores in a distribution fall.

👹 FISH: Friendly Introductor	ry Statistics Help			
File Analysis Options Help				
SET NUMBER OF CASES	S AND VARIABLES	Do All Ste	eps Lock Data Sortby Rank	'iew Histogram
N= 30 • 1 Variable	C 2 Variables Reset	Case #		Rank 🔼
ANALYSES		1	144.000	18.500
STEP 1	STEP 2	2	220.000	30.000
Enter Data DONE	Calculate Central Tendency	3	138.000	14.000
Cananata Dandam Data	Hele Dell	4	145.000	20.000
Generale Random Dala	пер рок	5	162.000	26.500
STEP 3:	STEP 4:	6	142. the scores in order from the	16.500
Calculate Deviation Scores	Calculate Squared	7	170. smallest to largest. Note that it is	28.000
	Usis De la	8	124. a "toggle" button and will rank	6.500
		9	158. and un-rank the scores with the	23.500
STEP 5	STEP 6	10	154.000	<b>-1</b> 22.000
Sum the Squared Deviation Scores	and Variance			
Hole Dolt	Holp Dolt		STATISTICS, using variance formulas with IN-1	<u>.=29:</u>
Tierp Do it	neip Don	Sum:	4276.000	-
STEP 7	STEP 8	Mean (M):	142.533	
Standard Deviation	for each Case	Median:	141.000	15.500
Help Do It	Help Do It	Mode:	120.000	-
STEP 9 Calculate Standard Error of the Mean Help Do It	STEP 10 Confidence Interval for the Mean Help Do It			

To illustrate this concept using FISH, utilize the "Sort By Rank" Option.

### Mode:

The mode of a distribution of scores is the score that occurs most frequently. This is also easily illustrated by using the "Sort By Rank" Option. When the scores are ranked, it is easy to find the most frequent score.

Note: The current version of FISH (2.5.8) does not have the ability to determine if a distribution is bimodal or multimodal. It is important to inform students of this limitation, and to make corrections accordingly.

👹 FISH: Friendly Introductor	y Statistics Help						
File Analysis Options Help	R						
SET NUMBER OF CASES	SAND VARIABLES	Do All Ste	eps Lo	ck Data	Sort by Rank	View Histog	jram
N= 30 • 1 Variable (	C 2 Variables Reset	Case #	Data (X)	(X-M)		Rank	
ANALYSES		21	136.000	-6.533		13.000	
STEP 1	STEP 2	22	142.000	-0.533		16.500	
Enter Data DONE	Calculate Central	23	120.000	-22.533		4.500	
Concepto Dondom Dota	Hole Dolt	24	120.000	-22.533		4.500	
Generale Random Dala		25	160.000	17.467		25.000	
STEP 3:	STEP 4:	26	158.000	15.467		23.500	
Scores	Deviation Scores	27	144.000	1.467		18.500	
Holo A Dolt	Holp	28	130.000	-12.533		10.500	
Tierp Don		29	125.000	-17.533		8.000	
STEP 5	STEP 6	30	175.000	32.467		29.000	
Deviation Scores	and Variance	COLUMNS	STATISTIC	S, using vari	ance formulas with h	<u>N-1=29:</u>	
Help Do It	Help Do It	Sum:	4276.000	0.000			
STEP 7	STEP 8	Mean (M):	142.533	0.000			
Step Three: Calcu	late Deviation Scores	Median:	141.000			15.500	
This option will ca	alculate and sum the	Mode:	120.000				
average of the	deviation scores.						
STE	0151 10	5		0.0			
Calculate Standard	Confidence Interval						
Heip Dolt	Help Do It						-

Step Three: Calculate Deviation Scores

FISH makes it easy for students to visualize the following property of the mean: The sum of the deviation scores about the mean equals 0.

$$\sum \left( X - \overline{X} \right) = 0$$

Because the sum of the deviation scores equals 0, the mean of the deviation scores is clearly 0.

Students can confirm this property for themselves by verifying the calculations computed by FISH.

T NUMBER OF CASE	S AND VARIABLES	Do Al St	eps Lo	ck Data	Sort by Rank	View Histo
• 1 Variable	C Z Variables Reset	Case #	Dota (K)	(05M)	(X-M0X-M)	Rank
IALYSES		21	136,000	-6.533	42.684	13.000
STEP 1	STEP 2	22	142.000	-0.533	0.284	16.500
nter Data DONE	Calculate Central	23	120.000	-22.533	507.751	4,500
Concession Property States	Note L Die N	24	128.000	-22.533	507.751	4.500
metheropation Hardin plant Thatair	neh net	25	160.000	17.467	305.084	25.000
STEP 3	STEP 4	26	158.000	15.467	239.218	23 500
Calculate Deviation	Calculate Squared	27	144,000	1.467	2.151	18.500
		28	130.000	-12.533	157.084	10,500
rieip Do a	nup	29	175 000	17522	367.419	e nan
STEP 5	STEP 6	944 9	tep Four:	Calculate	Squared Devia	tion Scores
Sum the Squared Deviation Scores	Calculate Range and Variance	10000	his option	will calcu	late the square scores.	d deviation
Sum the Squared Deviation Scores Help Do It	Calculate Range and Variance Help Dolt	COLU Bum:	4276.000	0.000	liate the square scores.	d deviation
Sum the Squared Deviation Scores Help Do It STEP 7	Calculate Range and Variance Help Do It STEP 8	Bum: Mean (M)	4278000 142.533	0.000	late the square scores.	d deviation
Sum the Squared Deviation Scores Help Do It STEP 7 Calculate the Standard Doviation	Calculate Range and Variance Help Dolt STEP 8 Calculate Z-Scores	Bum Mean (M) Median	4276000 142.533 141.000	0.000	late the square scores.	d deviation
Sum the Squared Deviation Scores Help Do It TEP 7 Calculate the Standard Deviation Help Do I	Calculate Range and Variance Help Dolt STEP 8 Calculate Z-Scores for each Case Help Dolt	Bum Mean (M) Median Mode	4276.000 142.533 141.000 120.000	0.000	late the square scores.	d deviation

Step Four: Calculate Squared Deviation Scores

Because the sum of the deviation scores is 0, it is not very interesting or useful. Therefore, the deviation scores are squared to provide better insight into the data.

Step Five: Sum the Squared Deviation Scores

<ul> <li>I Variable</li> </ul>	Z Variables Reset	Carat	Data 64	losan	ALLANDAR	Dack
MALVEED		an an	130.000	(C-TM)	(0-14)(0-14()	13.000
STEP 1	STEP 2	0.0	142.000	-0.000	0.004	10 000
nter Data DONE	Colculate Central	22	120.000	-0.030	607 761	4 600
(	Tendency	24	128.000	.22533	507.751	4 500
Senantin Flandom Data	Help Dolt	25	168.000	17 467	305.084	25,000
TEP 3:	STEP 4	26	158,000	15 467	239.218	23 500
Calculate Deviation	Calculate Squared	27	144,000	1.467	2151	18.500
Scores	Deviation Scores	28	130.000	-12,533	157.084	10,500
Heip Dolt	Huip Dolt	23	125.000	-17.533	307.418	8.000
TEP 5	STEP 6	30	175.000	32.467	1054.084	29.000
Sum the Squared Deviation Scores	Calculate Range and Variance	COLUMN	STATISTIC	S using ve	riance formulas wi	th P-1+29
Help Do It	Help Dolt	Sum:	4276.000	0.000	14787.5	
TEP 7	STEP 8	Mean (M):	142.533	D.O.B.		
Calculate the	Calculate Z-Scores	Median:	144.000	T		15.500
standard Deviation	or each case	Hodg:	120.000			
Heip Dall	Haip Dol					
Help Doll STEP 8 St Calculate Standar Thi Error of the Mean	ep Five: Sum the Square s option sums the square This sum is frequently o	d Deviation ed deviation alled the "f	n Scores in scores sum of			

The transition from the sum of the deviation scores to the sum of the squared deviation scores is a good way to introduce the concept of the *least-squares estimate* approach that is used in more advanced statistical procedures such as ANOVA.

The sum of the squared deviation scores or "sum of squares" is smaller than the sum of the squared deviation scores about any other number. The least-squares estimate is named such because the mean is used to minimize the squares of the deviation scores compared to the squares of the deviation of the scores from any other value.

👹 FISH: Friendly Introductor	ry Statistics Help					
File Analysis Options Help						
-SET NUMBER OF CASES	SAND VARIABLES	Do All Ste	eps Lo	ck Data	Sort by Rank	View Histogram
N= 30 • 1 variable	z vanables	Case #	Data (X)	(X-M)	(X-M)(X-M)	Rank 🔼
ANALYSES		21	136.000	-6.533	42.684	13.000
STEP 1	STEP 2	22	142.000	-0.533	0.284	16.500
Enter Data DONE	Calculate Central Tendency	23	120.000	-22.533	507.751	4.500
Concente Devilem Date		24	120.000	-22.533	507.751	4.500
Generale Handom Dala		25	160.000	17.467	305.084	25.000
STEP 3:	STEP 4:	26	158.000	15.467	239.218	23.500
Scores Help Do It	Deviation Scores Help Do It	28 Stan 29	dard Dev	variabilit iation) are and S	y (Range, Varia e calculated in Seven.	Steps Six
STEP 5 Sum the Squared Deviation Scores	Calculate Range and Variance		STATISTIC	32.467 : <mark>S, using va</mark>	1054.084 riance formulas w	29.000
Неір По іт	Help	Sum:	4276.000	0.000	14787.5	
STEP 7	STEP 8	Mean (M):	142.533	0.000		
Calculate the Calculate the Standard Deviation	for each Case	Median:	141.000			15.500
	Hole Delt	Mode:	120.000			
neih Do ii		Range:	110.000			
STEP 9	STEP 10	Variance:	509.913	or		
Error of the Mean Help Do It	for the Mean Help Do It	Std Dev:	22.581			

Steps Six and Seven: Calculate Range, Variance, and Standard Deviation

Measures of variability are useful to describe the dispersion of scores within a distribution.

### Range:

The range is simply the measure of the distance between the largest and smallest score within a distribution. FISH computes the range using this formula:

The *inclusive range* is sometimes reported (and called *range* at times). The inclusive range is calculated using this formula:

The "1" is added to indicate that the distance being measured is the upper limit of the largest score and the lower limit of the smallest score. If the scores are whole numbers, the largest score will be .5 below the upper limit and the smallest score will be .5 above the lower limit (.5 + .5 = 1).

Although the range and inclusive range are easy to calculate, they can be less than useful for two reasons: 1) The range is dependent on the two most extreme scores and 2) The range will likely increase as sample size increases.

Variance and Standard Deviation:

As discussed previously, deviation scores are the distance from each raw score to the mean. Recall that a property of the mean is that the sum of the deviation scores from the mean is 0. This property made the sum and the average of the deviation scores neither interesting nor useful. To compensate for this property of the mean, there are two obvious choices: 1) Take the absolute value of the deviation scores or 2) Square the deviation scores.

The mean absolute deviation is the sum of the absolute values of the deviations from the mean. This is seldom used in statistics. Instead, the terms are squared. The sum of squares is the sum of the squared deviation scores. The sum of squares is the preferred method and is used to calculate variance and standard deviation. The sum of squares is also used in many other statistical procedures.

The variance of a distribution is not a widely reported measure of variability. However, it is very useful in other analyses. The variance is the average of the sum of squares. When a researcher has measured all possible scores in the population, the population variance is calculated in the following manner:

$$\sigma^2 = \frac{\sum (X - m)^2}{N}$$

To provide an unbiased estimate of the population parameter  $\sigma^2$ , the following formula is used to calculate the sample variance:

$$S^{2} = \frac{\sum (X - \overline{X})^{2}}{N - 1}$$

Note that FISH allows users to "toggle" from calculations using N and N-1 at the touch of a button.

The standard deviation of a distribution is the square root of the average sum of squares. (Recall that the average sum of squares is the variance.) Squaring the deviation scores and then taking the square root preserves the units (in the sample SBP data, mmHg) of the original scores. Like the variance, there is a formula for the population standard deviation and the sample standard deviation. The population standard deviation is calculated with the following formula:

$$\sigma = \sqrt{\frac{\sum (X-m)^2}{N}}$$

The sample standard deviation is calculated:

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}}$$

Recall that FISH allows users to "toggle" from calculations using N and N-1 at the touch of a button

🙀 FISH: Friendly Introductor	y Statistics Help							
File Analysis Options Help	Ν							
SET NUMBER OF CASES		Do All Ste	eps Lo	ck Data	Sort by I	Rank	View Histog	ram
N= 30 • 1 Variable (	2 Variables Reset	Case #	Data (X)	(X-M)	(X-M)(X-M)	Z score	Rank	
ANALYSES		14	110.000	-32.533	1058.418	-1.441	1.000	
STEP 1	STEP 2	15	128.000	-14.533	211.218	-0.644	9.000	
Enter Data DONE	Calcute Control				57.084	-0.555	10.500	
	Step Eight: C	alculate Z-	score for	each Case	6.751	-0.334	12.000	
Generate Random Data	case Note that	t FISH also	e a z-scol	e for each	14.151	-1.264	2.000	
STEP 3:	STEP 4 and standa	rd deviatio	n of the a	-scores.	04.018	-1.175	3.000	
Calculate Deviation	Deviation Scores	-	1		43.484	-0.821	6.500	
Hala	Hele Delt	21	196,000	-6.533	42.684	-0.289	13.000	
Help Duit		2	142.000	-0.533	0.284	-0.024	16.500	
STEP 5	STEP 6	23	120.000	-22.533	507.751	-0.998	4.500	
Deviation Scores	and Variance	COLUMN S		S, using var	lance formu	las with N	<u>-1=29:</u>	
Help Do It	Help Do It	Sum:	4276.000	0.000	14787.5			
STEP 7	STEP 8	Mean (M):	142.533	0.000		0.000	-	
Calculate the	Calculate Z-Scores	Median:	141.000		12		15.500	
Standard Deviation	ior each case	Mode:	120.000					
Help Do It	Help Do It	Range:	110.000					
STEP 9	STEP 10	Variance:	509.913					
Calculate Standard	Confidence Interval	Std Dev:	22.581			1.000		
Help Do It	Help Do It							

Step Eight: Calculate Z-Score for Each Case

A Standard Score (z-score) relates a raw score to the mean and standard deviation its' distribution. A z-score represents the number of standard deviations a score is above (for a positive z) or below (for a negative z) the mean of a distribution. For a population, z-scores are calculated using this formula:

$$Z = \frac{X - \overline{X}}{S}$$

For a sample, the following formula is used:

$$Z = \frac{X - m}{s}$$

The z-distribution has a mean = 0 and a standard deviation = 1. Using the distribution of scores provided with the SBP data, this can be confirmed using FISH calculations by looking at the "column statistics" for the z-scores.

Why use z-scores?

1. Because the mean of a z-distribution is 0, the *sign* of the z-score is informative. By inspection, you can quickly determine if a score is above of below the mean. If the z-score is positive, it lies above the mean; if the z-score is negative, it lies below the mean.

2. Because the standard deviation of a z-distribution is 1, the *magnitude* of the z-score is also informative. When the distribution of scores is approximately normal, 2/3 of the scores fall between +1 and -1, and 95% of the scores fall between -2 and +2.

3. Z-scores permit comparisons across distributions with different means and standard deviations. This is only true if the distributions being compared are reasonably equivalent.

Steps Nine and Ten: Calculate Standard Error of the Mean and Confidence Interval for the Mean

🙀 FISH:		ry Statistics Help								X
File Analy	Confidence Interval	l options								
-SET N	Select the option	to be used when		Do All Ste	eps Lo	ck Data	Sort by	Rank	/iew Histog	ram
N= 30	calculating confid that SPSS uses th	ence intervals (note le t-distribution with	et	Case #	Data (X)	(X-M)	(X-M)(X-M)	Z score	Rank	
ANALYS	n-1 degrees of fre	edom).		14	110.000	-32.533	1058.418	-1.441	1.000	
Enter	C 68% CL (	7 = 1 ()()		15	1+00-000	14500	011.010	0.044		
or	C 90% CL (	7 = 1.645)		16 Sta	eps Nine	and Ten: F	ISH will ca	calculate t	he 500	
Gene	● 95% CL (	7 = 1.96)		1/ col	nfidence	interval fo	r the mea	n. Note t	hat	-
STEF	C 99% CL (	z = 2 33)		18 FI	SH gives	several op	otions for	calculati		-
Calc	3378 O.I. (i	2.00)	ł	20		onnuence	milervars	•	00	
1	🔿 Use t stati	istic (using n-1)		21	136.000	-6.533	42.684	-0.289	13.000	-
H			μ	22	142.000	-0.533	0.284	-0.024	16.500	
STEF				23	120.000	-22.533	507.751	-0.998	4.500	
Dev				COLUMNS	TATISTIC	S. using var	iance formu	las with N-	1=29:	
Н				Sum	4276:000	0.000	14787 5			
STEF			Ē	Mean (M):	142,533	0.000	1 Tror.5	0.000		
C	Use Default	СОК	s	Median:	141.000		2		15.500	
Stan				Mode:	120.000	1				_
He	elp Do It	Help Do It		Range:	110.000	1				
STEP	9	STEP 10		Variance:	509.913	1				
Erro	ulate Standard or of the Mean	for the Mean	al	Std Dev:	22.581			1.000		
He	ala Dolt	Help Do It		Std Error:	4.123					
	003	Theip DO R		95% CI (z):	(134.453,	150.614)				_

Up to this point, we have considered a distribution of raw scores, the SBP data. The standard error of the mean and the confidence interval of the mean are based upon a *theoretical* distribution: the sampling distribution. Note that a sampling distribution is different from a distribution of raw scores. A sampling distribution is based on the idea of repeated sampling: drawing an infinite number of samples from a population. The mean of all scores in a sampling distribution is called the *expected value* of the statistic (in our case the statistic is the mean) for that particular distribution. The standard deviation of a sampling distribution of a statistic is called the *standard error* of the statistic.

Because the statistic we are interested in is the mean, FISH is calculating the standard error of the mean. When the population standard deviation ( $\sigma$ ) is known, the standard error of the mean is calculated in the following manner:

$$s_{\overline{X}} = \frac{s}{\sqrt{N}}$$

When the population standard deviation is unknown, the standard error of the mean is calculated using this formula:

$$s_{\overline{X}} = \frac{s}{\sqrt{N}}$$

A confidence interval is a range of scores that provides a specific probability that the interval will contain the population parameter. In our case, we are estimating the confidence interval of the mean; therefore, the mean is the population parameter. If you had the ability to draw 100 samples from the population and calculate the 95% confidence interval for each sample (100 total intervals), approximately 95 of the confidence intervals would contain the population parameter (and 5 will not).

Confidence intervals can be calculated using both the z-distribution and the t-distribution. When  $\sigma$  is known and the sample size is large (> 120), confidence intervals are calculated using the z-distribution:

68% Confidence Interval:	$\overline{X} \pm 1.00 s_{\overline{X}}$
90% Confidence Interval:	$\overline{X} \pm 1.645 s_{\overline{X}}$
95% Confidence Interval:	$\overline{X} \pm 1.96 s_{\overline{X}}$
99% Confidence Interval:	$\overline{X} \pm 2.33 s_{\overline{X}}$

When  $\sigma$  is unknown and the sample size is small (< 120), confidence intervals are calculated using the t-distribution:

$$X \pm \left[t_a \left(N-1\right)\right] S_{\overline{X}}$$

where  $t_a(N-1)$  is the tabulated *t* value with N-1 degrees of freedom at the desired  $\alpha$ .

It is important to note that standard errors and confidence intervals can be computed for other statistics such as the regression coefficient (slope) and the correlation coefficient.

Recall that FISH allows users to "toggle" between several confidence interval calculations with the click of a button.

### **Histogram Feature**

It is important to examine a distribution of data graphically. Using the histogram option, FISH allows users view the distribution of the data.



Using the "Now Showing" buttons, FISH allows users to "toggle" between the following options: Show Raw Scores, Show Z-Scores, and Show Both. The Show Both option offers a concrete way for students to verify that the distribution of raw scores and the distribution of z-scores have the same shape.



Please refer to the User's Guide for a more thorough examination of FISH's graphics capabilities.

### **5** Analyzing Two Variables Using a FISH Data File

To illustrate uses for FISH in the classroom setting, a data file has been included with this Instructor's Manual. The data (SBP & AGE FISH\_data) can be found on the CD-ROM. The data set was taken from Lemeshow, 2003 (Applied Regression Analysis Course Notes, 2003 Summer Program in Applied Statistical Methods, p. 4). The data includes observations on age and systolic blood pressure for a sample of 30 individuals.

Open the file "SBP & AGE FISH\_data" and import the data into FISH. The data will appear on the screen as follows:

ET NUMBER OF CASE	S AND VARIABLES	Do Al S	teps L.o	ck Data	Borthy Barik	View Spotlarph
- 30 C 1 Variable	• 2 Variables Reset	Case #	Deta (%)	Data (r)		
NALYSES		1	3/8 0.041	144.000		
Enter Deta	STEP 2 Calculate Means for	2	47.000	220.000		
M	X and Y	3	45.000	136.000		
Receiveder Flowinger Flore	Hain Dolt	4	47.000	145.000		
CIEVICIONE TROTROCIMIE/ORG	the p	6	65.000	162.000		
STEP 3:	STEP 4:	6	46.000	142.000		
Calculate Standard	Calculate Deviation	7	67.000	170.000		
Deviations	scores	8	42 000	124.000		
Help Doll	Help Doll	9	67.000	159.000		
TEP 5	STEP 6	10	56.000	154.000		-
Calculate Cross	Calculate	10	56.000	194.000		
Products	Covariance	COLUMN	STATISTIC	S using vari	ince formulas with	NH1-29
Help Dolt	Help Dolt		1			Indexed by
STEP 7	STEP 8	-				
Calculate Correlation	Coefficient of Determination (r*)					
Help Doll	Help Do II	<u> </u>				_
STEP 9 Calculate Regression Slope	STEP 10 Calculate Y-Intercept					

Steps Two, Three, and Four: Calculate Means for X and Y, Calculate Standard Deviations, and Calculate Deviation Scores

SET NOMBER OF CASES AND VARIABLES		Do All St	eps Lo	ck Dala	Epithy	Balk 19	ler Epation	
- Du A A A A A A A A A A A A A A A A A A	· 2 vanables Hose	Case #	Data (X)	Data (r)	(X-M)	(YHM)		
IALYSES	STED 2	1	39.000	1.44.000	-6.133	1.467		
nter Data DONE	Calculate Means for X and Y Help Do	3 Step	s Two and e mean, v deviation	d Three: F ariance, a of both v	ISH calcul nd standa variables.	ind 7		
TEP 3:	STEP	6	Lec one	1.42.000	0.002	67	_	
Calculate Standard	Calculate Deviation	7	67.000	170.000	21.867	27.467		
available in the little	there are the	8	42.000	124.000	+3.133	-18.533		
Help Bolt	neip 00.0	9	67.000	158.000	21.867	15.467		
STEP 5	STEP 6	10	56.000	154.000	10.867	11.467		
Products	Covariance	P COLUMN STATISTICS, using venence tomulas with NH-2						
Help Dolt	Help Doll	Sum	1354.000	4276.000				
TEP7	STEP 8	Mean (M):	45.133	142.533	0.000	0.000		
Calculate	Coefficient of	Variance:	233.913	509.913				
Help Doll	Help Doll	Std Dev:	15.294	22.581	15.294	22.581		
TEP 9 Calculate Regression Slope	STEP 10 Calculate Y-Intercept	Step Four: scores a deviatio	FISH calc and the m	ulates the ean and s deviation	deviation tandard scores.	n		

Please see pages 5-9 for information on calculating and interpreting measures of central tendency and variability.

Step Five: Calculate Cross Products

👹 FISH: Friendly Introductor	ry Statistics Help						
File Analysis Options Help							
-SET NUMBER OF CASES	AND VARIABLES	Do All Ste	eps Lo	ck Data	Sort by	/Rank	/iew Scatterplot
		Case #	Data (X)	Data (Y)	(X-M)	(Y-M)	(X-M)(Y-M)
ANALYSES		1	39.000	144.000	-6.133	1.467	8.996 📕
STEP 1	STEP 2	2	47.000	220.000	1.867	77.467	144.604
or DONE	Calculate Means for X and Y	3	45.000	138.000	-0.133	-4.525	0.604
Concrete Dendom Date	Hole Dolt	4	47.000	145.000	1.867	2.467	4.604
	Пер оок	5	65.000	162.000	19.867	19.467	386.738
STEP 3:	STEP 4: Step	5: Calculat	e Cross F	roducts	0.827	-0.533	-0.462
Deviations	Calculate Devil Cros	s products	are calcu	lated by	21.867	27.467	600.604
Holp	Holp	lying the d	eviation s	scores for	-3.133	-18.533	58.071
		13	107.000	130.000	21.867	15.467	338.204
STEP 5	STEP 6	10	56.000	154.000	10.867	11.467	124.604 👦
Products	Covariance COLUMN STATISTICS, using variance formulas with N-1=29;						
Help Do It	Help Do It	Sum:	1354.000	4276.000			
STEP 7	STEP 8	Mean (M):	45.133	142.533	0.000	0.000	
Calculate	Coefficient of Determination (r <sup>2</sup> )	Variance:	233.913	509.913			
		Std Dev:	15.294	22.581	15.294	22.581	
Help Do It	Help Do It						
STEP 9	STEP 10						
Calculate Begression Slope	Calculate Y-Intercent						
Help Do It	Help Do It		-				
				201			

Just as the deviation scores are important in the sum of squares, the cross products are important in calculating the sum of the cross products.



Step Six: Calculate Covariance

The covariance is the average sum of the cross products. It is also referred to as the average sum of the combined differences. Covariance is calculated with this formula:

$$\operatorname{cov}_{(x,y)} = \frac{\sum (X - \overline{X}) - (Y - \overline{Y})}{N - 1}$$

Covariance is also known as the unstandardized correlation coefficient. Covariance can be used to assess whether two variables are related. A *positive covariance* indicates that as one variable deviates from the mean, the other variable deviates in the same direction (e.g.: as X increases, Y also increases). A *negative covariance* indicates that as one variable deviates from the mean, the other variable deviates from the mean in the opposite direction (e.g.: as X increases, Y decreases).

Because covariance is not standardized, it is scale dependent. That is, you can not compare the covariance of two distributions in an objective fashion unless both distributions are measured in the same units.

#### Step Seven: Calculate Correlation

To standardize the covariance term, the standard deviations of X and Y are used. The *standardized* correlation coefficient is known as the Pearson Product-Moment Correlation Coefficient or Pearson's *r*.

👹 FISH: Friendly Introducto	ry Statistics Help						
File Analysis Options Help							
SET NUMBER OF CASE	S AND VARIABLES	Do All St	eps Lo	ck Data	Sort by	/Rank	/iew Scatterplot
	• 2 variables Reset	Case #	Data (X)	Data (Y)	(X-M)	(Y-M)	(X-M)(Y-M)
ANALYSES		1	39.000	144.000	-6.133	1.467	-8.996
STEP 1	STEP 2	2	47.000	220.000	1.867	77.467	144.604
Enter Data DONE	Calculate Means for X and Y	3	45.000	138.000	-0.133	-4.533	0.604
Comparis Dandam Data		4	47.000	145.000	1.867	2.467	4.604
Generale Bandom Dala		5	65.000	162.000	19.867	19.467	386.738
STEP 3:	STEP 4:	6	46.000	142.000	0.867	-0.533	-0.462
Calculate Standard	Calculate Deviation	7	67.000	170.000	21 867	27 /67	<u> </u>
	Itels De K	8	42.000	124.01 FL	tep 7: Cal SH calcul	iculate Co	rrelation
пер оп		9	67.000	158.0	betwee	n X and Y	using
STEP 5	STEP 6	10	56.000	154.0	Pe	earson's r	
Calculate Cross Products	Calculate Covariance	COLUMN	1 STATISTIC	S, using va	rian e form	ulas with N-	1=29:
Help Do It	Help Do It	Sum:	1354.000	4276.000	/		6585,867
STEP 7	STEP 8	Mean (M):	45.133	142.523	0.000	0.000	
Calculate	Coefficient of	Variance:	233.913	501,913	Constraints	107.5.5	
Correlation	Determination (r <sup>2</sup> )	Std Dev:	15.294	22.581	15.294	22.581	
Help Do It	Help Do It	Covarianc	227.09				
STEP 9	STEP 10	Pearson r	0.658				
Calculate Begression Slope	Calculate Y-Intercent						
Help Do It	Help Do It						

Pearson's r is calculated conceptually with the following formula:

$$r = \frac{\text{cov}_{xy}}{s_x * s_y}$$

Computationally, the formula for Pearson's r is:

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{(N - 1)s_x * s_y}$$

By standardizing the covariance, values for r range from +1 to -1. When r = +1, X and Y are perfectly, positively correlated. As one variable increases, the other variable increases proportionally. When r = -1, X and Y are perfectly, negatively correlated. When r = 0, there is no linear relationship between X and Y.

Note that FISH will calculate *r* using either raw scores or z-scores. Users may toggle between the two options. This will allow users verify two items: 1) When using standard scores, the covariance term = correlation coefficient and 2) The correlation coefficient is the same when calculated with z-scores.

👹 FISH: Friendly Introductor	ry Statistics Help							
File Analysis Options Help								
SET NUMBER OF CASES	AND VARIABLES	Do All St	eps Lo	ck Data	Sort by	y Rank	√iew Scatter	plot
		Case #	Data (X)	Data (Y)	Zx	Zy	ZxZy	
ANALYSES		1	39.000	144.000	-0.401	0.065	-0.026	
STEP 1	STEP 2	2	47.000	220.000	0.122	3.431	0.419	Ĵ.
Enter Data DONE	Calculate Means for	2	45.000	132.000	-0.009	-0.201	0.002	
U	Note that EISH wi			45.000	0.122	0.109	0.013	
Generate Handom Data	the correlation be	etween X a	and Y	162.000	1.299	0.862	1.120	
STEP 3:	ST using Z s	scores.		142.000	0.057	-0.024	-0.001	
Calculate Standard	Ca	13		170.000	1.430	1.216	1.739	
Hale De B	Hale Dalk	8	42.000	124.000	-0.205	-0.821	0.168	
Help Do It		9	67.000	158.000	1.430	0.685	0.979	
STEP 5	STEP 6	10	56.000	154.000	0.711	0.508	0.361	
Products	Covariance		STATISTIC	S, using val	riance form	nulas with N-	<u>-1=29:</u>	
Help Do It	Help Do It	Sun	1354.000	4276.000			19.069	
STEP 7	STEP 8	Mean (M):	45.133	142.533	0.000	0.000		
Calculate	Coefficient of	Variance:	233.913	509.913	1		1	
Correlation	Determination (r-)	Std Dev.	15.294	22.581	1.000	1.000		
Help Do It	Help Do It	Covarian	0.658				1	
STEP 9	STEP 10	Pearson r	0.658					
Calculate Regression Slope	Calculate Y-Intercept							
Help Do It	Help Do It							

### Step Eight: Coefficient of Determination (r<sup>2</sup>)

SET NUMBER OF CASES AND VARIABLES		Do Al St	Do Al Steps Look Bata		-Epin p	Eartby Plank Vie		ew Scatterplat	
		Cose #	Deta.00	Date (r)	Zx	Zy	2x2y	1	
ALYSES		1	39.000	144,000	-0.401	0.065	-0.026	_	
STEP 1	STEP 2	2	47.000	220.000	0.122	3.431	0.419		
nter Data. DONE	Calculate Means for	3	45.000	138.000	-0.009	-0.201	0.002		
en la companya de la		4	47.008	145.000	0.122	0.109	0.013		
The service is a service of the serv	neip 110 li	5	65.000	162,000	1.299	0.862	1.120		
STEP 3.	STEP 4	8	48.000	142,000	0.057	-0.024	-0.001		
Colculate Standard	d Calculate Z-Scores	7	67.000	Step 8: The coefficient of determination (R2) is calculated.					
Logoranoes		B	42.000						
Help Dolt	Holp Do It	9	67.000						
STEP 5	EP 5 STEP 6		56.000			100			
Calculate Cross Products	Colculate Covariance	COLUMN	TATISTIC	S. using ve	ienos te n	nuice with	NH-29		
Help Duit	Help Dull	Sum	1354.000	4276.000	1		19.069		
STEP 7	STEP 8	Mean (M)	45.133	142.533	0.000	0.000			
Calculate	Coefficient of	Variance.	233,913	509.91					
Correlation	Determination (r*)	Std Dev.	15.294	22.55	1.000	1.000			
Hidge Dolla	:Help: Dectu	Coverienc	0.658	/					
TEP 5	STEP 10	Pearsonn	0.658	/					
	Calculate							-	

The coefficient of determination  $(r^2)$  is a measure of the strength of the relationship between two variables. The coefficient of determination is sometimes referred to as the proportion of variance on one variable predictable from or attributable to the other variable. The coefficient of determination is calculated by simply squaring the correlation coefficient:

$$r = \left(\frac{\sum (X - \overline{X})(Y - \overline{Y})}{(N - 1)s_x * s_y}\right)^2$$

If  $r^2$  is multiplied by 100, it can be thought of as a measure of the percentage of variance explained by correlation between the two variables. In the example FISH data set,  $r^2 = .432$ . Therefore, the correlation with age (X) accounts for 43.2% of the variance in systolic blood pressure (Y).

Steps Nine and Ten: Calculate Regression Slope and Calculate Y-Intercept

Regression is a statistical procedure which fits a predictive model to the data and uses that model to predict values of the dependent variable from one or more independent variables. FISH provides an example of *simple regression*, the prediction of an outcome (dependent variable) from a single predictor (independent variable).

Regression models are a linear representation of the distribution of scores. Regression models attempt to summarize a distribution of scores with a straight line. There are many lines which can summarize the general trend of a distribution. Regression employs the *method of least squares* (see page 9), a mathematical technique to establish the line that BEST fits the distribution of scores.

ET NUMBER OF CASES	AND VARIABLES		we Die	es Dahr	1 inthe	- 15	view Scotter	
(+).00 (C: I Variable -	2 Variables Paret	Cause	Date pa	Date (V)	DCM)	0146	16440/-W	01-
BALVIES		1	1281	144088	-6130	1.457	-8.996	1
STEP 1	STEP 2	12	47.088	220.068	1.117	77.467	141804	
Enter Date _ 200NFD	Galcolote Means Inc	0	46.000	136.086	-0.133	-4822	8.804	
A second s	A send Y	4	42.000	145.086	1.867	2.467	4804	
Approximation of the second second	THIN DATE	5	05.000	102.000	19.057	11.467	385.735	
STEP 3	STEP 4:	6	45,044	A US DEEL	1.101	44.20	142	
Culculate Standard Dovistines	Calculate Deviation Scores	7	Step	s 9 and 10 regression	FISH call	ulates th id the	· 100.004	
Help Dult 1	Distant Strengt	-	1	¥-1	nercept.		210.001	
TEP 5 Colculate Creas	STEP 6 Colcalate	10	26.088	154,000	1.857	12.467	124884	1
Products	Covariance	COLUMNS	TATIETS	S. ISTRA	dance torm	ARE WERE	1523	
Help: Dit H	Thip Dan	San	B.C.F.	4276 8 4			1925-1927	ī
TEP 7	STEPS	Nech ME	610	14251	8,390	0.000		
Calculoto	Coefficient of	Verience.	233 383	505 13				
Completion	Constantine (s.)	StdDer	16.294	22 61	18,294	22.881		
1840 00.8	Help Dorff.	Covenent	227 889	1				
TEPS	STEP IN	Peariet (	0.628	1				
Calculato	Calculate	r-squered	0.432	1				
understory stops	1. insucohe	Slope	0.971					
- Thrue - Thrue	Theta The R.	COMPANY OF	08718					

Review of Straight Lines:

Any line can be drawn if two things are known: 1) slope (gradient) and 2) point at which the line crosses the y-axis of the graph (intercept). The equation for straight lines is represented by:

$$Y = \boldsymbol{b}_o + \boldsymbol{b}_1 \boldsymbol{X}_i + \boldsymbol{e}_i$$

Where Y = Outcome variable of interest (dependent variable)  $\beta_0 = Y$ -intercept  $\beta_1 = Slope$   $X_i = i^{th}$  subject's score on the predictor (independent variable)  $\epsilon_i = Residual \text{ or Error term } (\hat{X}_i - X_i)$ 

<u>Slope:</u>

Conceptually, the formula for slope is:

$$\frac{\operatorname{cov}_{(x,y)}}{S_x}$$

Notice that the covariance is the numerator for both the slope and the correlation coefficients (page 20). Recall that the correlation coefficient provides a measure of how two random variables are associated in a sample. It is also a measure of the strength of the linear relationship between X and Y. Intuitively, the slope is an indicator of the relationship between two variables.

Therefore, if the slope is positive, then the correlation coefficient will be positive and vice versa. The correlation coefficient will always have the same sign as the slope.

Y-Intercept:

The formula for the Y-intercept is:

$$\hat{b}_{o} = \overline{Y} - \hat{b}_{1} * \overline{X}$$

**Regression Equation:** 

The least squares line is generally represented as:

$$\hat{Y} = \hat{b}_o + \hat{b}_1 * X$$

Note that FISH provides the regression equation under the "View Scatterplot" option.

### Scatterplot Feature

It is important to examine a distribution of data graphically. Using the scatterplot option, FISH allows users view the distribution of the data.



By clicking "View Predicted Values", FISH will calculate the observed FISH also has the capability to calculate observed, predicted, and residual values, squared deviation scores, model sum of squares  $((\hat{Y} - \overline{Y})^2)$ , as well as calculations of the standard error of the estimate, standard error of the regression slope, and the 95% confidence interval for the regression slope.

Please refer to the User's Guide for a more thorough examination of FISH's graphics capabilities.

# 6 Using the Data Generator to Teach Statistical Concepts

FISH allows users to generate random data. Data is generated by defining a population and then sampling from that population. FISH's graphics capabilities complement the data generator and allow for abstract concepts such as sampling distributions to be concretely explained to students.

With the one variable analysis, the histogram feature is enhanced to allow the user to view the sampling distribution of the mean. From this screen, the user can generate additional samples in multiples of 1, 10, 100 or 10,000. These sample means are then added to the existing sampling distribution of the mean.

With the two variable analysis, the scatterplot feature is enhanced to allow users to view the sampling distribution of the regression coefficient. From this screen, the user can generate additional samples in multiples of 1, 10, 100 or 10,000. These sample regression coefficients are then added to the existing sampling distribution of the regression slope.

It is important to note that as each sample is drawn from the population, the data are entered into the spreadsheet on the main FISH page. It can be helpful to "toggle" back and forth from the graphics to show students that as samples are drawn, new statistics are calculated for each sample.

#### Central Limit Theorem:

The central limit theorem is an important, yet difficult concept for beginning (and sometimes intermediate!) statistics students. FISH's random data generator allows a sampling distribution for the mean to be created, and the program's graphics capabilities help to make the concept more tangible for students.

Using the one variable analysis, FISH creates a sampling distribution of the *mean*. A sampling distribution of the mean is a distribution of sample means when repeated samples are taken from a population. (It is important to note that sampling distributions can be created for any statistic, not just the mean.)

Recall that the central limit theorem is based on the characteristics of a sampling distribution when a reasonable large (N = 30) sample is drawn. The central limit theorem states that as sample size (N) gets infinitely large, the shape of the sampling distribution of the mean approaches the normal distribution with:

standard deviation = 
$$\frac{s}{\sqrt{N}}$$

Recall that the standard deviation of a sampling distribution is called the standard error of the mean (see page 14). Therefore, as  $N \rightarrow \infty$ , the graph of the sampling distribution of the mean approaches normal with an

expected value =  $\mu$  and a standard error of  $\frac{s}{\sqrt{N}}$ . Because the sample size is in the denominator of the formula for standard error, as N increases, the

standard error of the mean decreases. Intuitively, a large N will result in sample means that are closer in value to the population mean.

To simulate the central limit theorem using FISH, data must be generated using the data generator. See the User's Guide for the data generation procedure.



As you set the population parameters, it is helpful to explain to students that this is a contrived example. In reality, we rarely know the population parameters. If we knew the population parameters, we would not need to estimate them using statistics.

The goal of this learning activity is to concretely explain the central limit theorem. Therefore, it is helpful to create a scenario. An example would be: Ohio Stadium seats 105,000 people (the population N). We, as researchers, decide to conduct a survey of the people in the stadium to determine how much money the average Ohio State fan has in his/her wallet on game day. To complete our study, everyone in class is going to survey 30 people (the sample n). For some reason (and this NEVER happens, because it is a contrived example) Brutus the Buckeye mascot whispers in our ear "The population mean is \$75 ( $\mu$  = 75) and the population standard deviation is \$10 ( $\sigma$  = 10)." Brutus then informs us that the data are normally distributed.

After students are given the contrived scenario, FISH can be used to create a sample (n = 30) for each member of the class. Students can record their sample means and standard deviations and watch as their mean is added to the distribution of sample means.



An example of how the activity would proceed follows:





At this point, it should be pointed out that our population of 105,000 in the example is not large enough to draw 10,000 samples of size 30. But since this is a *contrived* example, students typically get the picture.

The next step in the learning activity is to explore the central limit theorem further. Notice that the central limit theorem did NOT say that the population has to be distributed normally. To confirm this, repeat the exercise using a non-normally distributed population. Further, you could highlight that as long as the sample size is reasonably large (n > 30) the distribution of sample means will be approximately normal. This can be confirmed by adjusting the sample size. Using FISH, students can clearly see that when n is reasonably large, a distribution of sample means will approximate the normal distribution even if the means come from a population that is non-normal.

It should be noted that the central limit theorem applies to all populations and can be used to estimate the mean from any population when reasonably large samples are drawn. The central limit theorem was named such because the idea is CENTRAL to statistics. The central limit theorem allows us to assess how likely it is that a random sample will yield a value for the sample mean ( $\overline{X}$ ) that is any given distance from the population mean ( $\mu$ ). This concept will be used throughout any statistics course.

#### Distribution of the Slope

COMING SOON – WATCH THE WEB FOR UPDATES! http://oak.cats.ohiou.edu/~brooksg/fish.htm

### 7 References

The following statistics references were consulted in the writing of this manual:

Field, A. (2000). Discovering statistics using SPSS for windows. Thousand Oaks, CA: Sage.

Harris, M. B. (1998). Basic statistics for behavioral science research(2<sup>nd</sup> ed.). Needham Heights, MA: Allyn & Bacon.

Howell, D.C. (2004). Fundamental statistics for the behavioral sciences(5<sup>th</sup> ed.). Belmont, CA: Brooks/Cole – Thompson Learning.

Lemeshow, S. (2003, June). Applied Regression Analysis. In 2003 Summer Program in Applied Statistical Methods, conducted at The Ohio State University, Columbus, OH.

### 8 User Agreement

Use of the FISH software program provided with this Agreement constitutes acceptance of these terms and conditions of use. If you do not agree to the terms of this agreement, do not use the FISH software program.

#### LICENSE

FISH is a copyrighted program and is NOT public domain. The user is granted license, not ownership, to use the FISH software program on any computer subject to the restrictions described in the User Agreement and Disclaimer.

FISH is Freeware. The user is licensed to make an unlimited number of exact copies of the FISH software program, to give these exact copies to any other person for their personal use, and to distribute the FISH software program in its unmodified form only via disk, email, or local area network. If these methods of distribution are unavailable, any person wanting to use the FISH software program should be directed either to contact the author or to visit the author's internet web site (the URL is provided below and may be posted on any web site).

If you find the program useful, if you copy it for others, if you find problems or bugs in the program, or if you use the program for teaching, educational, or consulting purposes, you are requested to inform the authors:

Author:	Gordon P. Brooks, Ph.D.
Address:	McCracken Hall, Ohio University, Athens, OH 45701
Telephone:	740-593-0880
Fax:	740-593-0477
Email:	brooksg@ohiou.edu
Web Page:	http://oak.cats.ohiou.edu/~brooksg/fish.htm

#### TERMS OF USE

The FISH software program may be used and copied for personal use subject to the following The FISH software program may be used and copied for personal use subject to the following license restrictions:

- the FISH software program shall be copied and supplied in its original, unmodified form;
- the FISH software program shall not be sold or used for profit, nor may any amount or fee be charged for use, rental, lease, or distribution of the program;
- the FISH software program shall not be included or bundled with any other goods or services;
- the FISH software program may not be decompiled, disassembled, or otherwise modified; any such unauthorized use without expressed, written permission granted by the author shall result in immediate and automatic termination of this license.

#### DISCLAIMER OF WARRANTY

Every effort has been made to ensure the accuracy of the FISH program, the algorithms and subroutines used, and the results produced by the FISH software program, both on screen and printed.

However, no warranty is expressed or implied concerning the function or fitness of the FISH software program, subroutines, or results provided by the program. That is, the FISH software program is provided on an "as is" basis without warranty of any kind. The author shall have neither liability nor responsibility to any person or entity with respect to any liability, loss, or damage directly or indirectly arising from the use of or inability to use the FISH software program or the results of the analyses provided by the FISH software program, even if the author has been advised of the possibility of such damages or claims. In no event shall any liability exceed the license fee paid to the author of the FISH software program. In the event of invalidity of any provision of this license, the user agrees that such invalidity shall not affect the validity of the remaining portions of this license.

All rights not expressly granted here are reserved to the author of the FISH software program.

User's Guide and Instructor's Manual Prepared by Holly Raffle.