

## Biostatistics Lab Notes

### Lab 1: Measurement and Sampling

Because we used a chance mechanism to select our sample, each sample will differ. My data set (GerstmanB.sav), looks something like this:

ID	AGE	HIV	KAPOSISA	REPORTDA	OPPORTUN	SBP1	SBP2
35	21	Y	N	01/09/89	Y	120	126
37	42	Y	Y	09/01/89	Y	118	118
43	5	N	Y	01/12/90	Y	83	86
143	11	Y	N	02/17/89	Y	126	124
321	30	Y	Y	05/25/89	Y	87	82
329	50	Y	Y	12/29/89	N	114	118
337	28	N	N	08/19/89	Y	119	119
492	27	N	N	08/31/89	N	115	111
494	24	Y	Y	08/19/89	Y	127	129
546	52	Y	Y	10/13/89	Y	94	89

**Lab 2: Frequencies and Stem-and-Leaf Plots**

1. My 10 age values are: 21, 42, 05, 11, 30, 50, 28, 27, 24, 52 (Your data will differ). A stem-and-leaf plot of these data is shown below:

```

| 0 | 5
| 1 | 1
| 2 | 1478
| 3 | 0
| 4 | 2
| 5 | 02
(x10) years

```

Interpretation: The shape of this distribution leaves the impression of a “mound” with a mode in the “twenties.” (There may be another mound at the high end, but this is only based on two observations.) The center of the distribution is in the 20s (“location”), and values range from 5 to 52 (“spread”).

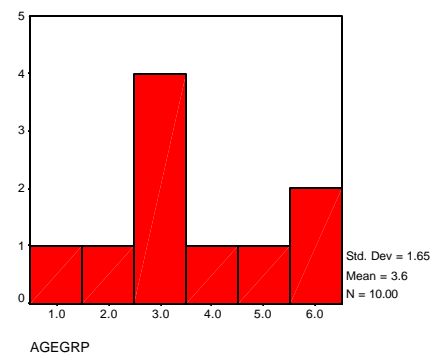
3. Frequency table of raw values:

Age	Frequency	RelFreq	CumFreq
5	1	10.0%	10.0%
11	1	10.0%	20.0%
21	1	10.0%	30.0%
24	1	10.0%	40.0%
27	1	10.0%	50.0%
28	1	10.0%	60.0%
30	1	10.0%	70.0%
42	1	10.0%	80.0%
50	1	10.0%	90.0%
52	1	10.0%	100.0%
Total	10	100.0%	

4. Frequency table with data in 10-year class-intervals:

Group Number	Age (years)	Freq	RelFreq	CumFreq
1	0 -9	1	10%	10%
2	10 - 19	1	10%	20%
3	20 - 29	4	40%	60%
4	30 - 39	1	10%	70%
5	40 - 49	1	10%	80%
6	50 - 59	2	20%	100%
Total		10	100.0	--

5. Histogram (see fig to right):



6. Frequency of HIV:

HIV+	Frequency	RelFreq	CumFreq
No	3	30%	30%
Yes	7	70%	100%
Total	10	100.0	

Seven (70%) of the 10 subjects are HIV positive.

**Lab 3: Summary Statistics**

Data set is GerstmanB.sav.

1. Summary Statistics: units of measure are “years”

$$n = 10; \Sigma x = 290; \bar{x} = 290 \text{ years} / 10 = 29.0 \text{ years}$$

$$SS = (21 - 29)^2 + (42 - 29)^2 + (5 - 29)^2 + (11 - 29)^2 + (30 - 29)^2 + (50 - 29)^2 + (28 - 29)^2 + (27 - 29)^2 + (24 - 29)^2 + (52 - 29)^2 = 2134 \text{ years}^2$$

$$s^2 = 2134 / (10 - 1) = 237.1111 \text{ years}^2$$

$$s = \sqrt{237.1111 \text{ years}^2} = 15.4 \text{ years}$$

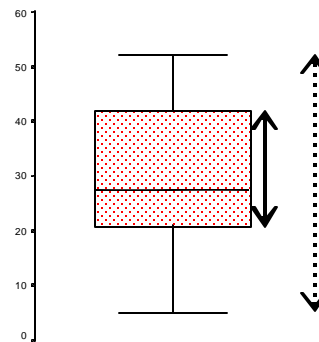
2. Ordered Array showing location of 5-point summary: 5-point summary is 5, 21, 27.5, 42, 52.

5	11	21	24	27	28	30	42	50	52
min		Q1		m			Q3		max

**3 & 4. Boxplot.**

The boxplot in the figure to the right demonstrates the location and spread of the distribution. Whenever you draw a boxplot, you should check the central location of the data (summarized by location of the median and of the “box”) and the spread of the data (summarized by the distance between the hinges and the distance between the whiskers).

Location: In the boxplot to the right, median is a little less than 30, the lower hinge (quartile) is 21 and the upper hinge (quartile) is 42.



Spread: The middle 50% of the data lies between these two hinges (solid vertical line). The “whisker-spread”--which in this case is also the range-- lies between dotted line. These distances describe the spread of the data.

**Lab 4A (Binomial Distributions)**

1.a. Binomial distribution for  $X \sim b(n = 3, p = .25)$

$$\Pr(X = 0) = .4219$$

$$\Pr(X = 1) = .4219$$

$$\Pr(X = 2) = .1406$$

$$\Pr(X = 3) = .0156$$

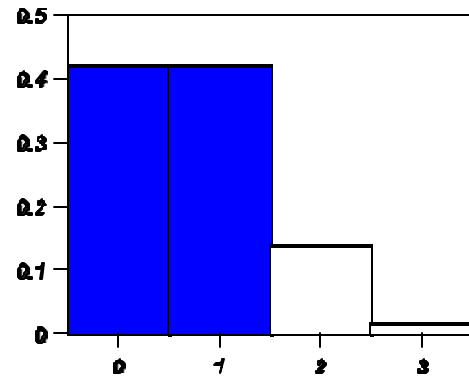
1.c. The above probability distribution is shown as a histogram in the figure to the right

1.d. The expected value and variance of this distribution can be calculated with short-cut formulas

$$\mu = np = (3)(.25) = .75.$$

$$\sigma^2 = npq = (3)(.25)(.75) = 0.5625.$$

Of course these describe the central location and spread of the distribution.



1. e. Cumulative probability function for the random variable  $X \sim b(n = 3, p = .25)$  is simply the values in the left tail of the distribution.

$$\Pr(X \leq 0) = 0.4219$$

$$\Pr(X \leq 1) = 0.8438$$

$$\Pr(X \leq 2) = 0.9843$$

$$\Pr(X \leq 3) = 1.0000$$

The probably histogram shown above shades the area corresponding to “less than or equal to one success.” This area sums to .8428. This shows  $\Pr(X \leq 1) = 0.8428$ , and is referred to the left tail of the distribution. Notice that the cumulative probability is the *left* tail of the distribution. The complement of this is the *right* tail of the distribution. The right tail of this distribution corresponds to  $\Pr(X \geq 2) = \Pr(X = 2) + \Pr(X = 3) = .1406 + .0156 = 0.1562$ . Notice that  $\Pr(X \leq 1) + \Pr(X \geq 2) = 1.0000$  for this distribution, because it includes all possibilities. Also notice that  $\Pr(X \geq 2) = 1 - \Pr(X \leq 1)$ , because of their complementary nature.

**Lab 4B (Normal Distributions)**

1. A standard normal variable ( $Z$ ) is a normal random variable with a mean of 0 and standard deviation of 1. You will be using this table often, so please become familiar with its use.

2. & 3. Normal probabilities are determined as areas under a standard normal curve. Here are some specific examples:

$$\Pr(Z < +1.96) = .975$$

$$\Pr(Z < -1.96) = .025$$

$$\Pr(Z < +1.5) = .9332$$

$$\Pr(Z > +1.5) = 1 - .9332 = .0668$$

$$\Pr(Z > -3.56) < .0001$$

I strongly urge you to *draw* the areas under the curve in each instance.

$$4. z_{.83} = 0.95$$

$$z_{.95} = 1.645$$

$$z_{.975} = 1.96$$

$$z_{.99} = 2.33$$

$$z_{.17} = -0.95$$

$$z_{.05} = -1.645$$

$$z_{.025} = -1.96$$

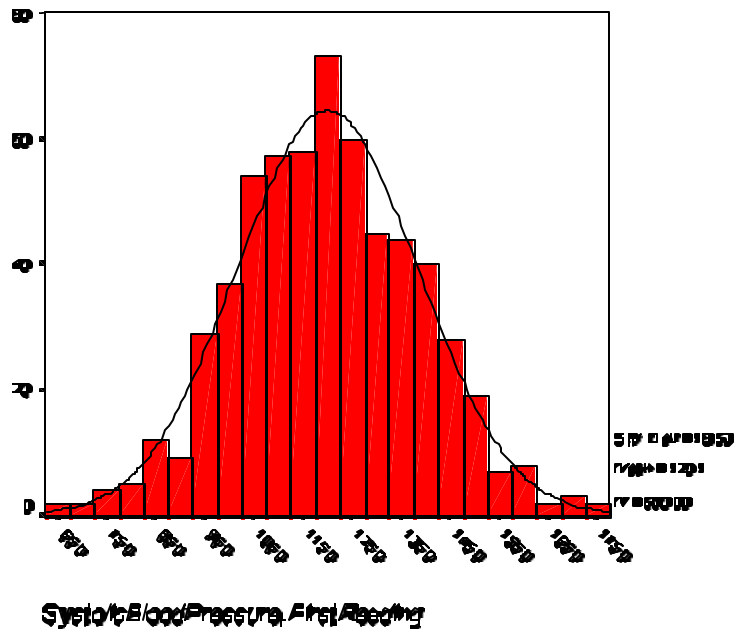
$$z_{.01} = -2.33$$

$$z_{.005} = -z_{.995} = -2.58$$

[Draw, draw, draw!]

6. Given  $X \sim N(200, 5)$ . Then,  $\Pr(X > 210) = \Pr(Z > 2) = .0227$ ;  $\Pr(X < 190) = \Pr(X < -2) = .0227$

7. This question addresses the distribution of the SBP1 variable. The figure below shows the distribution of this variable in the population. The distribution is approximately normal with a mean of 120.13 and standard deviation of 18.53. (The fit to the normal curve is not perfect, but is pretty good. In fact, no variable is every perfectly normal, but this is a pretty good fit for the purposes of statistical inference.) Notice how the area under the curve corresponds to the area of the histogram bars. Therefore, cumulative relative frequencies corresponds to “left tail areas” and hence cumulative probabilities. For example, the cumulative relative frequency of a systolic blood pressure in these data is 2.67%. (See your frequency table to confirm this) If we were to model the relative frequency (probability) with the normal curve, we would derive  $\Pr(X \leq 83) = \Pr(Z \leq (83 - 120.13) / 18.53) = \Pr(Z \leq -2.00) = .0228$ , which is close to the 2.67%. This shows that the normal model is a pretty good fit, at least at this point on the curve.



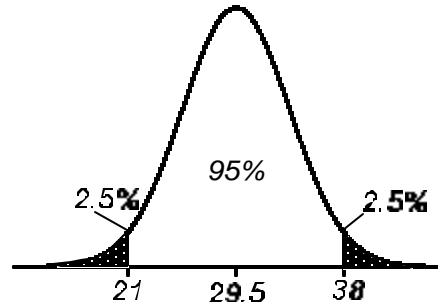
**Lab 5 (Estimating a Population Mean)**

Notes:

- $t_{\text{infinity}} = Z$
- It is helpful to draw the area under the curve when determining  $t$  percentiles.
- The  $t$  percentiles requested in this lab are as follows
  - $t_{12,.90} = 1.36$
  - $t_{12,.95} = 1.78$
  - $t_{12,.975} = 2.18$
  - $t_{12,.99} = 2.68$
  - $t_{12,.01} = -2.68$
  - $1.36 < t_{12,.93} < 1.78$  [ $t_{12,.93}$  is wedged between  $t_{12,.90}$  and  $t_{12,.95}$ ]
  - $-1.78 < t_{12,.07} < -1.36$  [ $t_{12,.07}$  is wedged between  $t_{12,.05}$  and  $t_{12,.10}$ ]
- There are numerous possible sample means. Therefore, the sample mean *is* a random variable. In contrast, there is only one population mean. The population mean *is* a constant.
- The variable MEANAGE in SampleMeans.SAV represents sample means from previous semesters. According to the sampling theory, the distribution of MEANAGEs will *tend* toward normality with a expected value of 29.5. The standard deviation of this distribution of means will tend toward  $\sigma / \sqrt{10} = 13.59 / \sqrt{10} = 4.3$ .
- When calculating a 95% CI for  $\mu$  using a population standard deviation, use the formula  $\bar{x} \pm (z_{1-\alpha/2})(s / \sqrt{n})$ .
- When calculating a 95% CI for  $\mu$  using the sample standard deviation, use the formula  $\bar{x} \pm (t_{n-1,1-\alpha/2})(s / \sqrt{n})$ .
- 5% of 95% confidence intervals for  $\mu$  will *fail* to capture the population mean.
- Optional: To attain margin of error  $d$  with 95% confidence when estimating  $\mu$ , use a sample size of  $n \approx \frac{4 \cdot s^2}{d^2}$ . For the variable AGE assume  $s = 13.59$ . For  $d = 5$ , the sample size,  $n$ , should be at least  $\frac{(4)(13.59^2)}{5^2} = 29.6 \approx 30$ .

**Lab 6 (Null Hypothesis Testing a Mean)**

1. The sampling distribution of means looks like this:



2. The above sampling distribution represents the hypothetical frequency distribution of all possible of means based on  $n = 10$  taken from the population. Because of the central limit theorem, we know this distribution will tend toward normality. Because of the unbiasedness of the sample mean, we know it will be centered on 29.5 (i.e., the true population mean). We also know that the standard deviation (error) of this distribution will be equal to the standard deviation of the population divided by the square root of the sample size, or  $13.59 / \sqrt{10} = 4.30$  in this instance. The region defined by  $\bar{m} \pm (1.96)(s_{\bar{x}}) = 29.5 \pm (1.96)(4.30) = 29.5 \pm 8.4 @ = (21, 38)$ . This encompasses 95% of the distribution. Therefore, we expect 2.5% of sample means to fall below 21, and we expect 2.5% of sample means to fall above 38.

3. The test requested in part 3 is based on something we know is false, i.e.,  $\mu = 32$ . We know this population mean is incorrect, but test it anyway to demonstrate one of the fallacies of the  $p$ -value: it does *not* really provide an “objective” probability. Instead, it quantifies how well the data conform to a null hypothesis which may or may not be right. Given this background, most students will still retain the null hypothesis even though the null hypothesis is wrong. Is this a contradiction? No! Retention of the null hypothesis does not imply it is true. It merely implies that there is not enough evidence for its rejection.

The  $z$  test is used when using the standard deviation from the population and the  $t$  test is used when using the standard deviation from the sample. The  $t$  test adds a little more “wiggle room” for the uncertainty associated with the sample variance. You see, inferential statistics are mostly about quantifying (random) uncertainty.

SPSS’s one-sample  $t$  test will provide results identical to your hand calculations. Remember to enter the hypothesized mean ( $\mu_0$ ) as the “test value” in SPSS’s one-sample test dialogue box.

**Lab 7: Paired Samples and Their Differences (Notes)**

1. **Background:** The pairing must be maintained for during analysis. All data are systolic blood pressure measured in mm Hg;  $n = 10$  for all analyses.

2. **Paired Samples:** Means and standard deviations are calculated in the usual manner:

$$\bar{x}_1 = 110.30, s_1 = 16.14$$

$$\bar{x}_2 = 110.20, s_2 = 17.71$$

3. & 4. **Stem-and-leaf plot of DELTA** values in my sample:

```

| -0 | 6
| -0 | 234
| +0 | 0024
| +0 | 55
Difference in SBP (mm Hg)

```

Shape: Mound shaped and more-or-less symmetrical.

Central location: Around 0.

Spread: Values range from -6 to +5.

5. **Summary statistics for DELTA:**  $\bar{x}_d = -0.1, s_d = 3.87$

6. **Estimation:** The 95% confidence interval for  $\mu_d = 0.1 \pm (t_{9, .975})(3.87 / \sqrt{10}) = 0.1 \pm (2.26)(1.22) = (-2.67, +2.87)$ . This defines an interval that has a good (95%) chance of locating the true (population) mean difference.

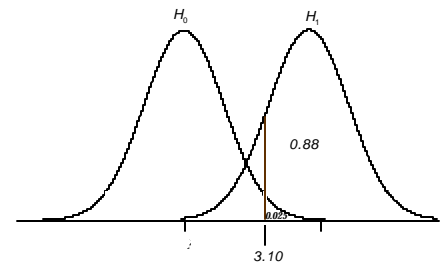
7. **Null hypothesis test:**  $H_0: \mu_d = 0$  vs.  $H_1: \mu_d \neq 0$  derives  $t_{\text{stat}} = (0.1 - 0) / 1.22 = 0.08$ ;  $df = 10 - 1 = 9$ ;  $p > .20$ ; retain  $H_0$ . (The difference is *not* significant.)

8. **Power of the test:** The above null hypothesis was retained at  $\alpha = .05$  (two-sided). We might ask about the power of the test assuming, under  $H_1$ , the true mean difference was 5 mm Hg (i.e., under  $H_1, \Delta = \mu_d = 5$ ). We assume  $\sigma_d = 5$ , which is the actual standard deviation of the difference in the population. Using these assumptions,

$$1 - b = f\left(-1.96 + \frac{5|\sqrt{10}|}{5}\right) = \Phi(1.22) = .88. \text{ Therefore, the study had adequate power.}$$

To see how this works, imagine two sampling distributions of means. Under the null hypothesis,  $\mu_d = 0$ . Under the alternative hypothesis,  $\mu_d = 5$  (figure, right). Both distributions have a standard error of  $= \sigma_d / \sqrt{n} = 5 / \sqrt{10} = 1.58$ . If we find a sample mean that is at least 1.96 standard errors above 0, then the null hypothesis will be rejected (this is because  $\alpha = .05$ , two-sided). Therefore, the critical mean difference  $= (1.96)(1.58) = 3.10$ . This implies that a sample mean of 3.10 or greater will cause a rejection of the null hypothesis. Assuming the alternative hypothesis is correct (i.e.,  $\mu_d = 5$ ), the probability of observing a sample mean that is greater than 3.10 =

$$\Pr(\bar{x} \geq 3.10) = \Pr\left(z \geq \frac{3.10 - 5}{1.58}\right) = \Pr(z \geq -1.20) = 0.88.$$





**Lab 7: Paired Samples and Their Differences (Lab report format.)**

**Purpose:** To describe the difference in systolic blood pressure measurements in paired samples, to estimate the mean difference in the population, and to determine whether the difference was significant.

**Methods:** Data in *GerstmanB10.SAV* are used to describe the difference in paired blood pressure readings in individuals. The variables containing the data are SBP1 (systolic blood pressure measurement 1, in mm Hg) and SBP2 (systolic blood pressure measurement 2, also in mm Hg). Means and standard deviations were calculated using routine methods. Differences were plotted in the form of a stem-and-leaf plot. A 95% confidence interval for the mean difference was calculated with the formula:  $\bar{x}_d \pm (t_{n-1, .975})(se_d)$  where  $se_d$  represents the estimated standard error of the

mean difference  $se_d = \frac{s_d}{\sqrt{n}}$ . The difference was tested for significance with a paired  $t$  statistic:  $t_{stat} = \frac{\bar{x}_d - 0}{se_{\bar{x}_d}}$  with  $n -$

1  $df$ . A two-sided test was used. The power of the test to detect a difference of 5 mm Hg at  $\alpha = .05$  (two-sided) was complete by applying the formula presented in *StatPrimer*.

**Results:** Data are:

SBP1	SBP2	DELTA
120	126	-6
118	118	0
83	86	-3
126	124	2
87	82	5
114	118	-4
119	119	0
115	111	4
127	129	-2
94	89	5

The mean of SBP1 is 110.30 (SD = 16.14). The mean of SBP2 is 110.20 (SD = 17.71). A stem-and-leaf plot of DELTA values is:

```

| -0 | 6
| -0 | 234
| +0 | 0024
| +0 | 55
Difference in SBP (mm Hg)

```

Data are mound-shaped with values ranging from -6 to +5. The mean value of DELTA is 0.1 ( $s_d = 3.87$ ,  $n = 10$ ).

**Inference about the mean difference:**

The 95% confidence interval for  $\mu_d = 0.1 \pm (t_{9, .975})(3.87 / \sqrt{10}) = 0.1 \pm (2.26)(1.22) = (-2.67, +2.87)$ .

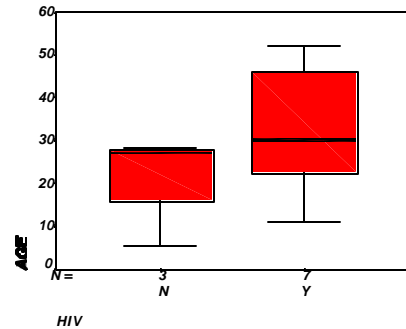
The test of  $H_0: \mu_d = 0$  vs.  $H_1: \mu_d \neq 0$  derives  $t_{stat} = (0.1 - 0) / 1.22 = 0.08$ ,  $df = 10 - 1 = 9$ . Using *Statable.exe*, the two-tailed  $p = .94$ . The null hypothesis is therefore retained; there is *not* a significant difference between the two samples.

The power of the test assuming  $\Delta = 5$ ,  $\sigma = 5$  and  $\alpha = .05$  is .88.

**Discussion:** Although individual differences ranged from -6 to +5, there was no significant difference in the two means of the two samples.

**Lab 8: Independent Samples and Their Differences (Notes)**

- Data:** Data will vary from sample to sample. In my sample, there are 7 HIV+ subjects and 3 HIV- subjects. Because this sample is small, it will be difficult to tell much from these data with any precision. (The data set is intentionally small, to allow for hand-calculations.)
- Summary Statistics by Group:** Let Group 1 be the HIV+ group. Let Group 2 be the HIV- group. See lab 1 for a listing of the data. The HIV+ group in my sample shows  $\bar{x}_1 = 32.86$ ,  $s_1 = 15.54$ ,  $n_1 = 7$ . The HIV- group shows  $\bar{x}_2 = 20.00$ ,  $s_2 = 13.00$ ,  $n_2 = 3$ .



- A **side-by-side boxplot** of the data is shown in the figure to the right. As noted above, the small sample sizes preclude detailed analysis. Nevertheless, the medians of the two groups seem to be similar. The interquartile range in the HIV+ group is greater than the IQR in the HIV- group.

- The pooled estimate of variance is the best estimate of within group variability. For my data:

$$s_p^2 = \frac{(6)(15.54)^2 + (2)(13.00)^2}{6 + 2} = 223.37 \text{ years}^2.$$

- The standard error of the mean difference is needed for inference about  $\mu_1 - \mu_2$ . For the current

$$\text{data, } se = \sqrt{223.37 \left( \frac{1}{7} + \frac{1}{3} \right)} = 10.31. \text{ Now, a 95\% CI for } \mu_1 - \mu_2 = (32.86 - 20.00) \pm (t_{8,975})(10.31) = 12.86 \pm$$

$(2.31)(10.31) = 12.86 \pm 23.82 = (-10.96, 36.64)$ . The lab states the actual value of  $\mu_1 - \mu_2$  is  $-0.64$  years. This was captured in the calculated interval (as will be true with 95% such intervals).

SPSS output for this procedure is shown below. Recall that we assumed equal variance in the two groups, so only the first row of the output table will be interpreted. The standard error of the difference and confidence interval calculated by SPSS (**highlighted in red**) are identical to our hand-calculated values.

**SPSS output: Independent Samples Test**

Levene's Test for Equality of Variances									
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	<b>95% Confidence Interval of the Difference</b>	
								<b>Lower</b>	<b>Upper</b>
Equal variances assumed	.485	.506	<b>1.247</b>	<b>8</b>	<b>.248</b>	12.86	10.31	<b>-10.92</b>	<b>36.64</b>
Equal variances not assumed			1.349	4.622	.240	12.86	9.53	-12.26	37.97

6. **Null hypothesis test:**  $H_0: \mu_1 - \mu_2 = 0$  vs.  $H_a: \mu_1 - \mu_2 \neq 0$ ;  $\alpha = .05$ ;  $t_{\text{stat}} = (32.86 - 20.00) / 10.31 = 1.25$ ;  $df = 6 + 2 = 8$ ;  $p > .2$ ; Retain  $H_0$ .

Retention of the null hypothesis does not imply it is correct. As a matter of fact, we have committed a type II error in this problem, since  $\mu_1 - \mu_2$  is actually not 0. (If you are not convince of this, you may download populati.sav and see for yourself.)

### 7. Sample Size Requirement

Assuming  $\sigma^2 = 184.69$ , the required per group sample size is  $n = [(16)(184.69)/10^2] + 1 = 30.5 \cong 31$ .

If  $\sigma^2$  were not known, we could make use of the fact that  $\sigma^2 \cong s_p^2 = 223.37$ . In this circumstance,  $n = [(16)(223.37)/10^2] + 1 = 36.7 \cong 37$ .

With either estimate, it is clear that the current sample size ( $n_1 = 7, n_2 = 3$ ) needs to be expanded in order to achieve adequate power for the stated purpose.

**Lab 9: Inference About a Proportion (Notes)**

1. Samples will vary. The data in my file show  $x = 7$  and  $n = 10$ . Therefore,  $\hat{p} = 7 / 10 = .7$ .
2. In my particular sample,  $n\hat{p}\hat{q} = (10)(.7)(.3) = 2.1$ . Therefore, methods based on normal approximations should be avoided.
3. Using the Web-calculator and my data, an exact 95% CI for  $p = (.35, .93)$ . This allows us to infer that the population proportion (prevalence) is between .35 and .93. This particular confidence interval is very wide, but does indeed capture the population proportion.
4. Obviously, the above confidence interval is very broad. Our estimate, therefore, is not precise. The margin of error ( $d$ ) is, by definition, half the confidence interval width. The confidence interval width for the current illustration  $= 0.93 - 0.35 = 0.58$ . The margin of error is half this:  $d = .58 / 2 = .29$ . We want to achieve  $d = 0.10$ . Since  $p$  is not known accurately, let us temporarily assume  $p = 0.5$ . (This will maximize the sample size.) Thereby,  $n = (1.96)^2(.5)(.5)/.1^2 \cong 96$ .
5. For GerstmanSampleBig.sav,  $x = 68$  and  $n = 96$ . Therefore,  $\hat{p} = .708$ . In this case,  $n\hat{p}\hat{q} = (96)(.708)(1-.708) = 19.8$ . Therefore, normal approximation based methods can be used with immunity.
6. A 95% CI for  $p$  based on the data in GerstmanSampleBig.sav  $= .708 \pm (1.96)(\text{sqrt}[(.708)(1-.708)/96]) = .708 \pm (1.96)(.0464) = .708 \pm .091 = (.617, .799)$ .
7. Null hypothesis test. First note  $SE_{\hat{p}} = \sqrt{\frac{(.5)(.5)}{96}} = 0.0510$ . The reason we use .5 as the value of  $p$  is that the null hypothesis is assumed to be true under the testing model. The steps of the test are:
  - $H_0: p = .5$  vs.  $H_1: p \text{ not } = .5$
  - $\alpha = .05$
  - $z_{\text{stat}} = (.708 - .5) / 0.0510 = 4.08, p < .002$
  - Since  $p < \alpha$ , reject  $H_0$

**Lab 10: Comparison of Independent Proportions (Notes)**

## 1. Cross-tabulation of Sex and HIV status

		HIV		Total
		1	2	
SEX	1	46	22	68
	2	20	4	24
Total		66	26	92

Prevalence in males ( $\hat{p}_1$ ) =  $46 / 68 = .6765$ ; Prevalence in females ( $\hat{p}_2$ ) =  $20 / 24 = .8333$ . Notice that the prevalence is slightly higher in females.

## 2. Expected frequencies

		HIV		Total
		1	2	
SEX	1	48.8	19.2	68
	2	17.2	6.8	24
Total		66	26	92

## 3. Print chi-square table.

## 4. Null hypothesis test

- $H_0: p_1 - p_2 = 0$  vs.  $H_a: p_1 - p_2 \neq 0$
- Let  $\alpha = .05$
- $\chi^2 = (46 - 48.8)^2/48.8 + (22 - 19.2)^2/19.2 + (20 - 17.2)^2/17.2 + (4 - 6.8)^2/6.8 = 0.16 + 0.41 + 0.46 + 1.15 = 2.18$ ;  $df = (2-1)(2-1) = 1$ ;  $p > .1$
- Retain  $H_0$ . ("No significant difference in prevalence.")

Notice that in the population, there is a slight difference in prevalence ( $p_1 = .230$ ;  $p_2 = .220$ ). Therefore, a type II error has been committed. (Although, only a minor one!)

## 5. Output from SPSS Chi-square procedure

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.153	1	.142		
Continuity Correction	1.449	1	.229		
Likelihood Ratio	2.315	1	.128		
Fisher's Exact Test				.190	.112
Linear-by-Linear Association	2.130	1	.144		
N of Valid Cases	92				

a Computed only for a 2x2 table

b 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.78.

SPSS provides the important message 0 cells (.0%) have expected count less than 5. This lets you know a chi-square method can be used.