

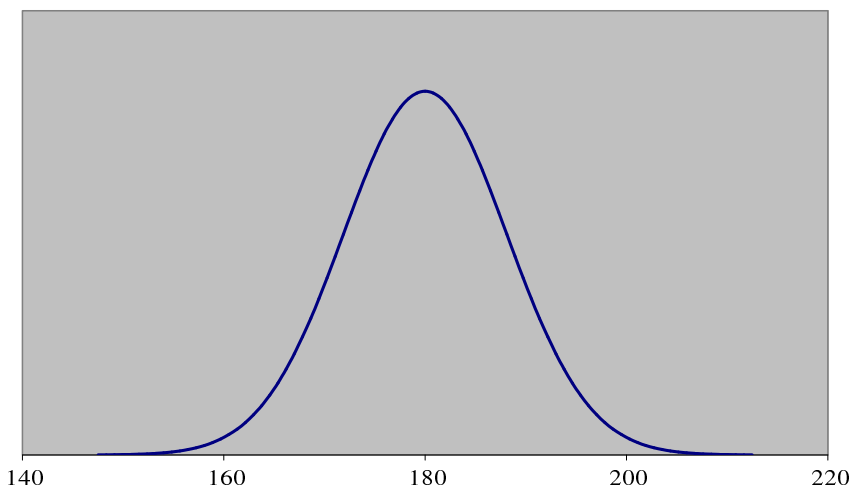
1 Power and sample-size calculations

Consider the following example. U.S. males ages 20-24 have mean cholesterol level $\mu = 180$ mg/ml. By comparison, we can assume that the mean cholesterol level in the overall population of males 20-74 is higher. Thus, we are carrying out the one-sided test of hypothesis:

$$\begin{aligned}H_o &: \mu \leq 180 \text{ mg/100 ml} \\H_a &: \mu > 180 \text{ mg/100 ml}\end{aligned}$$

If we collect a sample of $n = 25$ U.S. males ages 20-74 and measure their cholesterol level, we can use \bar{X}_n , the sample mean (mean cholesterol level in our sample) to make inferences about the (20-74 year-old male) population mean. The distribution of \bar{X}_n is normal, with mean $\mu = 180$ mg/100 ml (by the assumption implicit in the null hypothesis), and standard deviation $\sigma = \frac{46}{\sqrt{25}}$ mg/ml (by the central limit theorem). The situation is pictured in the following figure:

Figure 1: Distribution of $\bar{X}_n \sim N(180, 9.2)$

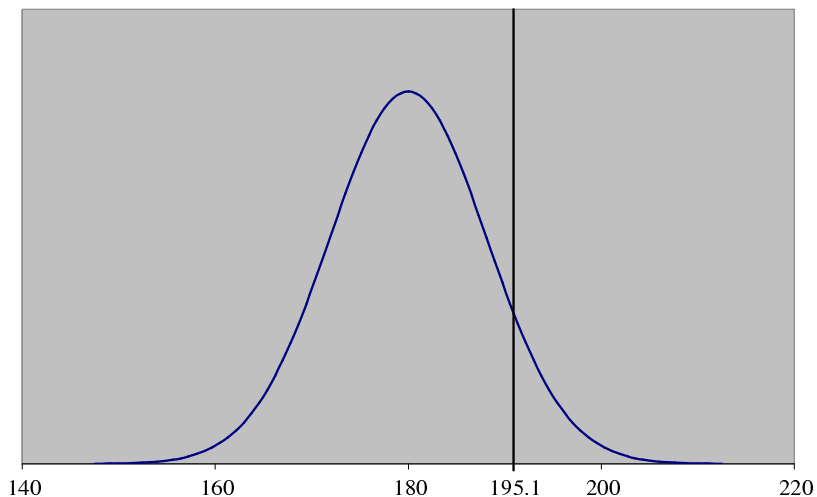


2 Types of error

In the one-sided case above, when determining the α level of a hypothesis test, you also determine a cutoff point beyond which you will reject the null hypothesis. This point can be found from test statistic $z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$, and the fact that we will reject the null hypothesis if $Z \geq z_\alpha$. In the cholesterol level example, the cutoff cholesterol level corresponding to a 5% α level is found as follows:

$$\begin{aligned}Z &\geq z_\alpha \\ \Leftrightarrow \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} &\geq z_\alpha \\ \Leftrightarrow \frac{\bar{X}_n - 180}{\frac{46}{\sqrt{25}}} &\geq 1.645 \\ \Leftrightarrow \bar{X}_n &\geq (1.645) \left(\frac{46}{25} \right) + 180 = 195.1\end{aligned}$$

Figure 2: Implication of setting the α level of a test



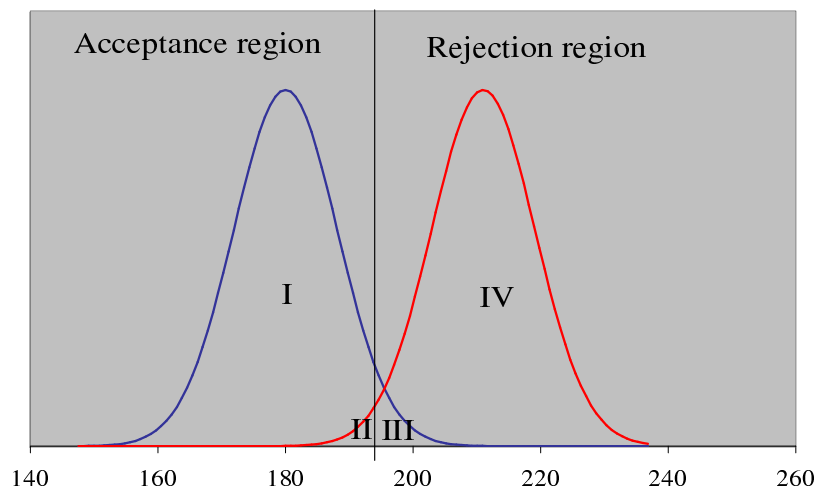
Thus, if the sample mean from a group of $n = 25$ males 20-74 years old is higher than 195.1 mg/dL, then we will reject the null hypothesis, and decide that the 20-74 population mean cholesterol level is higher than the 20-24 year old population. How often will such a sample mean be over 195.1 mg/dL even if the 20-74 year old males cholesterol level is the same as that of the 20-24 year olds?

This will happen $\alpha\%$ of the time. That is, $\alpha\%$ of the time we will be rejecting the null hypothesis even though it is true. This is called an *error of Type I*.

The alpha level of the test is the maximum allowed probability of type-I error

What if the "true" mean cholesterol of 20-74 year-olds is $\mu_1 = 211$ mg/dL? This situation is given in the following figure.

Figure 3: Normal distributions of \bar{X}_n with means $\mu_o = 180$ mg/dL and $\mu_1 = 211$ mg/dL and identical std. deviations $\sigma = 9.2$ mg/dL



There is a chance that even though the mean cholesterol level is truly $\mu_1 = 211$ mg/dL, that the sample mean will be to the left of the cutoff point (and in the acceptance region). In that case we would have to accept the null hypothesis (even though it would be false). That would be an *error of Type II*. The probability of a type-II error is symbolized by β .

What is this probability in this example? The probability of a type-II error is

$$\begin{aligned}\beta &= P(\bar{X}_n \leq 195.1 | \mu = \mu_1 = 211) \\ &= P\left(\frac{\bar{X}_n - 211}{\frac{46}{\sqrt{25}}} \leq \frac{195.1 - 211}{\frac{46}{\sqrt{25}}}\right) \\ &= P(Z \leq -1.73) = 0.042\end{aligned}$$

from the Appendix in the textbook.

There are four areas in the previous figure that are immediately identifiable:

- I. The distribution has mean $\mu = 180$ mg/dL and the null hypothesis is not rejected. In this case, the test made the correct decision.
- II. In this case, the distribution of cholesterol levels among 20-74 year-old males has a higher mean compared to that of 20-24 year-olds and the test has erroneously failed to reject the null hypothesis. This is an error of Type II.
- III. The mean cholesterol of the 20-74 year-olds is the same as that of 20-24 year-olds but the null hypothesis is rejected. This is an error of Type I.
- IV. The sample mean among 20-74 year-old individuals is truly higher than that of 20-24 year-olds and the test has correctly rejected the null hypothesis.

3 Power

The error associated with case II is called error of Type II. Just as we had defined the probability of a Type I error as α , and the probability of a Type II error as β , we have a special name for the probability associated with case IV, that is, the probability that the test will correctly reject the null hypothesis. This is called the *power* of the test. In other words, power is the chance that the test as defined will pick up true differences in the two populations.

$$\boxed{\text{Power} = 1 - \beta}$$

In the example above, the power is $1 - 0.042 = 0.958$ (or 95.8%).

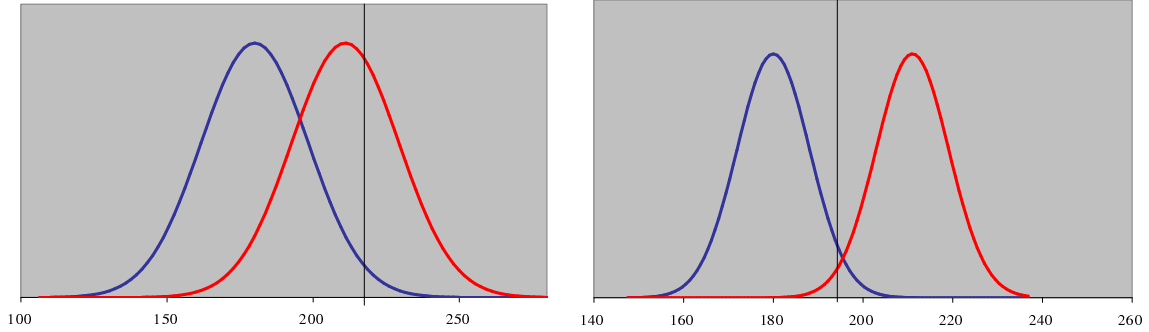
4 Sample size

In the beginning of a new experiment, we may ask ourselves how many subjects we should include. In other words, we want to determine the sample size of the study. All else being equal, the only way to increase the power of an experiment (i.e., increase its chance of detecting true differences) is by increasing the sample size. Consider the two cases: In the first case (left), the distributions are based on a sample size of $n = 5$ subjects versus $n = 25$ in the original situation (right).

To determine the sample size we need:

1. Null and alternative means

Figure 4: Separation of the null and alternative distributions with increasing sample size



2. Standard deviation (or estimate of variability)
3. Alpha level of the test
4. Desired power

Items 1 and 2 can sometimes be substituted by “standardized differences” $\delta = \frac{\mu_1 - \mu_2}{\sigma}$ where σ is the assumed common standard deviation.

Example: Cholesterol example (continued):

For example, if $\alpha = 1\%$, the desired power is $1 - \beta = 95\%$, and the two means are $\mu_0 = 180$ mg/dL and $\mu_1 = 211$ mg/dL respectively, then the cutoff point is

$$\bar{x} = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = 180 + (2.32) \left(\frac{46}{\sqrt{n}} \right)$$

Also, by virtue of the desired power,

$$\bar{x} = \mu_1 - z_\beta \frac{\sigma}{\sqrt{n}} = 211 - (1.645) \left(\frac{46}{\sqrt{n}} \right)$$

So,

$$180 + (2.32) \left(\frac{46}{\sqrt{n}} \right) = 211 - (1.645) \left(\frac{46}{\sqrt{n}} \right)$$

and thus,

$$n = \left[\frac{(2.32 - (-1.645))}{211 - 180} (46) \right]^2 = 34.6 \approx 35$$

In general,

$$n = \left[\frac{(z_\alpha + z_\beta)}{\mu_0 - \mu_1} \sigma \right]^2$$

To be assured of a 95% chance of detecting differences in cholesterol level between 20-74 and 20-24 year-old males (power) when carrying out the test at a 1% α level, we would need about 35 20-74 year-old males.

5 Computer implementation

Power and sample size calculations are performed in STATA via the command `sampsi`. The syntax is as follows (the underlined parts of the command are used to calculate the sample size and are omitted when calculating the power).

```
sampsi #1 #2 [,alpha(#) power(#) n1(#) n2(#) ratio(#) pre(#)
post(#) sd1(#) sd2(#) method(post|change|ancova|all) r0(#)
r1(#) r01(#) onsample onesided ]
```

In the two-sample case, when `n2` (and `ratio`) and/or `sd2` is omitted, they are assumed equal to `n1` and `sd1` respectively. You can use options `n1` and `ratio (=n2/n1)` to specify `n2`. The default is a two-sample comparison. In the one-sample case (population mean is known exactly) use option `onesample`. Options `pre(#)`, `post(#)`, `method(post|change|ancova|all)`, `r0(#)`, `r1(#)` refer to repeated-measures designs and are beyond the scope of this course.

5.1 Power calculations

In the cholesterol example, we will use the STATA command `sampsi` to compute the power of the study ($n = 25$, and $\sigma = 46\text{mg/ml}$) as follows:

```
. sampsi 180 211, alpha(.05) sd1(46) n1(25) onsample onesided
```

```
Estimated power for one-sample comparison of mean
to hypothesized value
```

```
Test Ho: m = 180, where m is the mean in the population
```

```
Assumptions:
```

```
alpha = 0.0500 (one-sided)
alternative m = 211
sd = 46
sample size n = 25
```

```
Estimated power:
```

```
power = 0.9577
```

The power is $0.9577 \approx 0.958$ as we saw earlier.

5.2 Sample size calculations

The sample size under $\alpha = 0.01$ and power 95%, is calculated as follows:

```

. sampsi 180 211, alpha(.01) power(.95) sd(46) onsample onesided

Estimated sample size for one-sample comparison of mean
to hypothesized value

Test Ho: m = 180, where m is the mean in the population
Assumptions:

alpha = 0.0100 (one-sided)
power = 0.9500
alternative m = 211
sd = 46
Estimated required sample size:
n = 35

```

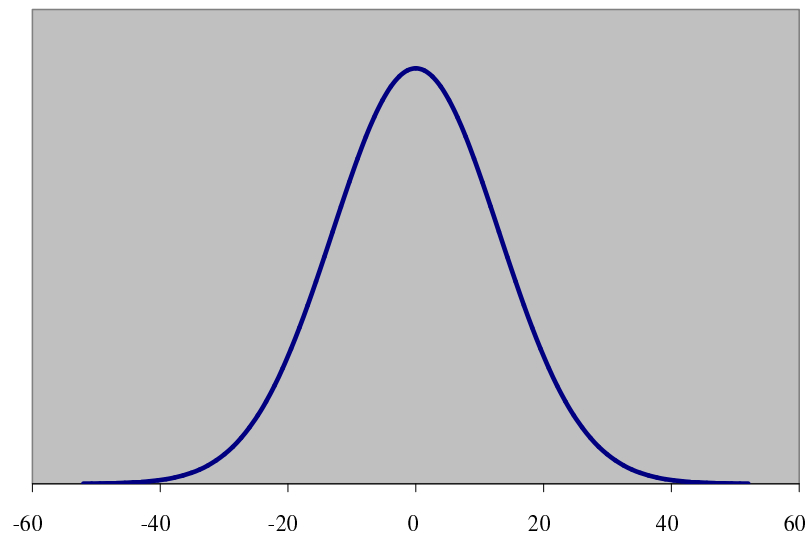
We use options `onesample` and `onesided` since we are working with a single sample ($\mu_0 = 180$ mg/ml is the population mean) and carry out a one-sided test.

6 The two (independent) sample case

In the two-sample case, recall that the null hypothesis is (usually) $H_0 : \mu_1 = \mu_2$. This is equivalent to $H_0 : \delta = \mu_1 - \mu_2 = 0$. Power and sample-size calculations are based on the distribution of the sample difference of the two means $\bar{d} = \bar{X}_1 - \bar{X}_2$ under some *a priori* assumptions.

The distribution of the sample difference of two means, assuming two equal-size $n_1 = n_2 = n$ (say) independent samples and known and equal variances ($\sigma_1 = \sigma_2 = s$) is $\bar{d} \sim N(\delta, \sigma_\delta)$, where $\sigma_\delta = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sigma \sqrt{\frac{2}{n}}$ if $n_1 = n_2 = n$.

Figure 5: Distribution of $\bar{d} \sim N(0, 13.01)$



Example: Cholesterol example (continued)

If the mean cholesterol among 20-24 year-olds were also unknown, while the standard deviation in both

groups were $\sigma = 46$ mg/ml, two samples would be collected. One from a group of 20-24 year-olds and one from a group of 20-74 year-olds. Under the null $H_o : \mu_1 = \mu_2 (= d = 0)$ the distribution of $\bar{d} \sim N(0, \sigma_{\bar{d}})$ with $\sigma_{\bar{d}} = \sigma \sqrt{\frac{2}{n}} = 13.01$ mg/dL, as depicted in Figure 5. If we wanted to carry out the cholesterol study in a manner identical to the previous discussion (with the exception of course that neither population mean is assumed known *a priori*) the procedure would be as follows:

1. $H_o: d = 0$
2. $H_a: d = 31$ mg/ml (corresponding to the situation where $\mu_1 = 180$ mg/ml and $\mu_2 = 211$ mg/ml)
3. $\alpha=0.01$
4. Power= $1 - \beta = 0.95$

To calculate the sample size for each group (n) we can use the previous one-sample formula, with the appropriate estimate of the variance of course. That is, each group will be comprised of individuals from each population,

$$nt = \left[\frac{(z_\alpha + z_\beta)}{\delta_\alpha} \sigma_d \right]^2 = \left[\frac{(z_\alpha + z_\beta)}{\delta_\alpha} \sigma \sqrt{2} \right]^2 = 2 \left[\frac{(z_\alpha + z_\beta)}{\delta_\alpha} \sigma \right]^2 = 2n$$

where n is the size of the identically defined one-sample case. That is, the sample size in the two-sample case will be roughly double that of the one-sample case. In this case, $nt = 2 \left[\frac{(2.32+1.645)}{31} (46) \right]^2 = 69.23$ The required sample size is at least 70 subjects per group (double the $n = 35$ subjects required in the identical one-sample study).

Note! The total required sample is 140 subjects, or *four times that of the single-sample study*. This is the penalty of ignorance of both means versus just one out of the two means.

7 Computer implementation

The computer implementation is given below.

```
. sampsi 180 211, alpha(.01) power(.95) sd1(46) onesided
Estimated sample size for two-sample comparison of means
```

```
Test Ho: m1 = m2, where m1 is the mean in population 1
and m2 is the mean in population 2
```

```
Assumptions:
```

```
alpha = 0.0100 (one-sided)
power = 0.9500
m1 = 211
m2 = 180
sd1 = 46
sd2 = 46
n2/n1 = 1.00
```

```
Estimated required sample sizes:
```

```
n1 = 70
n2 = 70
```

8 Power calculations when testing a single proportion

Power and sample size calculations can be carried out when the comparisons involve proportions. We discuss the one-sample case, as the two-sample case is a bit more complicated. When determining power or sample size for the comparison of a proportion (against a known population value), we make use of the normal approximation of the binomial distribution. That is, we use the fact that, at least for large n ,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

where p is the unknown population proportion, and $\hat{p} = \frac{x}{n}$ is the ratio of successes over the total number of trials (experiments).

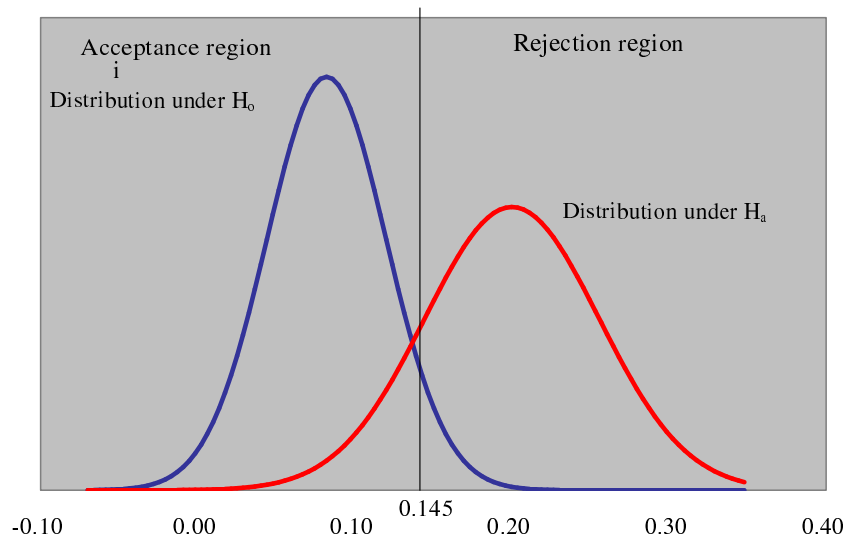
Example: Five-year survival rate of lung-cancer patients

Suppose that we are testing the hypothesis that the 5-year survival of lung-cancer patients under 40 is 8.2%, equal to 5-year survival rates of lung-cancer patients over 40 years-old. That is, we test $H_o : p_o \leq 0.082$ versus $H_a : p_o > 0.082$. If the 5-year survival among younger patients is as high as 20% (i.e., $p_a = 0.200$), and the study sample size is $n = 52$, then $\sigma_{\hat{p}} = \sqrt{\frac{p_o(1-p_o)}{n}} = \sqrt{\frac{0.082(1-0.082)}{52}} = 0.038$ under the null hypothesis, and $\sigma_{\hat{p}} = \sqrt{\frac{p_a(1-p_a)}{n}} = \sqrt{\frac{0.200(1-0.200)}{52}} = 0.055$, under the alternative. The null hypothesis will be rejected if $Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} > z_\alpha = 1.645$, that is, if

$$\hat{p} > p_o + z_\alpha \sqrt{\frac{p_o(1-p_o)}{n}} = 0.082 + 1.645 \sqrt{\frac{0.082(1-0.082)}{52}} \approx 0.145$$

This situation is depicted in the following figure:

Figure 6: Distribution of \hat{p} under the null hypothesis $\hat{p} \sim N(0.082, 0.038)$ (blue) and alternative hypothesis $\hat{p} \sim N(0.200, 0.055)$ (red)



Example: Five-year survival of lung-cancer patients (continued):

First of all,

$$\begin{aligned}
 z_\beta &= \frac{(p_a - p_0) - z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}}{\sqrt{\frac{p_a(1-p_a)}{n}}} \\
 &= \frac{(0.200 - 0.082) 1.645 \sqrt{\frac{0.082(1-0.082)}{52}}}{\sqrt{\frac{0.200(1-0.200)}{52}}} \approx 1.00
 \end{aligned}$$

Thus, the probability of a type-II error is $\beta = P(Z > z_\beta) = 0.159$ and thus, the power of a test for a single proportion based on $n = 52$ subjects is $1 - \beta = 0.841$ or about 84%.

If the power is 95% and the alpha level of the test is 1%, then the required sample size is

$$\begin{aligned}
 n &= \left[\frac{z_\beta \sqrt{p_a(1-p_a)} + z_\alpha \sqrt{p_0(1-p_0)}}{(p_a - p_0)} \right]^2 \\
 &= \left[\frac{1.645 \sqrt{0.200(1-0.200)} + 2.32 \sqrt{0.082(1-0.082)}}{(0.200 - 0.082)} \right]^2 = 120.4
 \end{aligned}$$

That is, about 121 lung cancer patients under 40 years old will be necessary to be followed, and their 5-year survival status determined in order to ensure power of 95% when carrying out the test at the 1% alpha level.

8.1 Computer implementation of power calculations for proportions

The power of a study testing $H_o : p = 0.082$ (known) versus $H_a : p=0.200$, involving $n = 52$ subjects, with a statistical test performed at the 5% alpha level, is computed with STATA as follows:

```

. sampsi .082 .200, n(52) alpha(.05) onsample onesided

Estimated power for one-sample comparison of proportion
to hypothesized value

Test Ho: p = 0.0820, where p is the proportion in the population

Assumptions:

      alpha =    0.0500  (one-sided)
alternative p =    0.2000
sample size n =         52

Estimated power:

power =    0.8411

```

Notice that omission of estimates for the standard deviation (sd1 and/or sd2) produced power calculations for proportions.

8.2 Computer implementation for sample-size calculations for proportions

In the case where the same hypotheses as above are tested, assuming a power of 95% and alpha level of 1%, the required sample size will be computed as follows:

```
. sampsi .082 .200, alpha(.01) power(0.95) onsample onesided

Estimated sample size for one-sample comparison of proportion
to hypothesized value

Test Ho: p = 0.0820, where p is the proportion in the population

Assumptions:

      alpha = 0.0100 (one-sided)
      power = 0.9500
  alternative p = 0.2000

Estimated required sample size:

n =      121
```

Thus, $n = 121$ subjects will be necessary to be involved in the study and followed for 5-year survival.