

Statistical Sampling

Live Session Unit 5

Mid-Term

- Saturday June 17, 9.00a.m CT
- Submit on Tuesday June 20, 11.00p.m. CT
- Text book, Live session notes, Labs, HWs, asynchronous videos (week1-week5).

Exercise!!

A population of 10 balls is made up of 4 white balls and 6 red balls as shown below:

R W W R W R W R R R

A systematic sample of size 2 is selected from the population and the proportion of red balls in the population is estimated from the sample using as the estimator

\hat{p} = the proportion of red balls in the sample

Find the sampling distribution of \hat{p} .

Exercise!!

R W W R W R W R R R
1 2 3 4 5 6 7 8 9 10

$$K=10/2=5$$

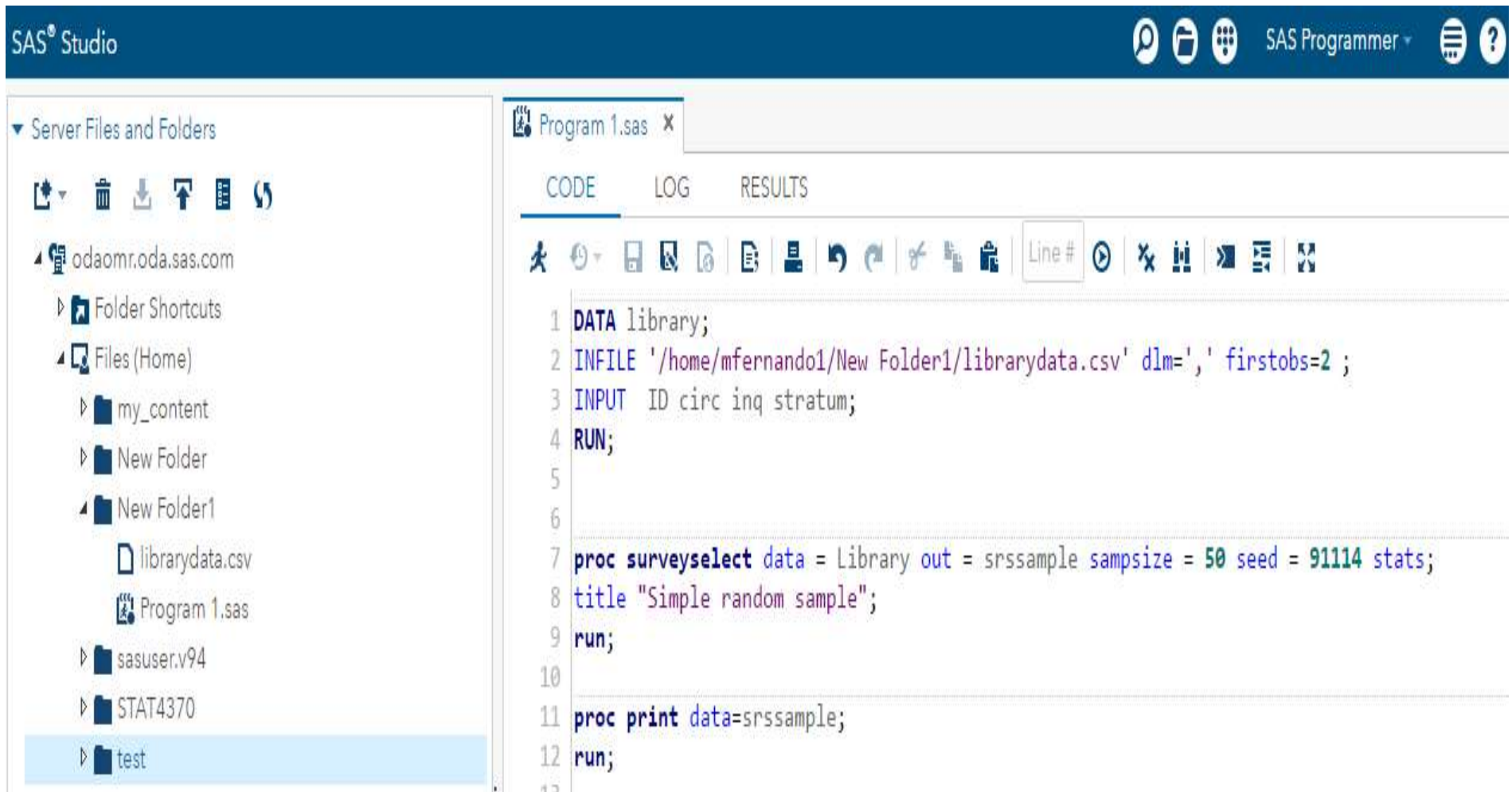
Random start between 1-5, then add k

Sample id	sample members	statistic value
1	1,6: R, R	1
2	2,7: W, W	0
3	3,8: W, R	0.5
4	4,9: R, R	1
5	5,10: W, R	0.5

statistic value	# of samples with this statistic value	proportion of samples with this statistic value
0	1	0.20
0.5	2	0.40
1	2	0.40

SAS on Demand

- <http://support.sas.com/software/products/on-demand-academics/#s1=2>



The screenshot displays the SAS Studio web interface. The top navigation bar includes the 'SAS Studio' logo, user profile icons, and the role 'SAS Programmer'. The left sidebar, titled 'Server Files and Folders', shows a tree structure for the server 'odaomr.oda.sas.com'. It includes 'Folder Shortcuts', 'Files (Home)' with folders like 'my_content', 'New Folder', and 'New Folder1', and a list of files including 'librarydata.csv' and 'Program 1.sas'. The main workspace is divided into a 'CODE' tab (active), 'LOG', and 'RESULTS'. The 'CODE' tab shows a SAS program with line numbers 1 through 13. The program code is as follows:

```
1 DATA library;
2 INFILE '/home/mfernando1/New Folder1/librarydata.csv' dlm=',' firstobs=2 ;
3 INPUT ID circ inq stratum;
4 RUN;
5
6
7 proc surveyselect data = Library out = srssample sampsize = 50 seed = 91114 stats;
8 title "Simple random sample";
9 run;
10
11 proc print data=srssample;
12 run;
13
```

R (Select a SRS)

- sample function
- `data1=read.csv(file.choose())` ← `LibrarySRS.csv`
- `head(data1)`
- `tail(data1)`
- `sampleSRS = data1[sample(1:nrow(data1), 50, replace=FALSE),]`

```
data1=read.csv(file.choose())  
  
head(data1)  
  
tail(data1)  
  
sampleSRS = data1[sample(1:nrow(data1), 50,  
                        replace=FALSE),]
```

R

- sampling package

<https://cran.r-project.org/web/packages/sampling/sampling.pdf>

- survey package

<https://cran.r-project.org/web/packages/survey/survey.pdf>

R (Select a SRS)

- `install.packages("sampling")`
- `data1=read.csv(file.choose())`
- `library(sampling)`
- `idnumber=data1$id`

- `samp=srswor(50,length(idnumber))`
- `#samp=srswor(50,369)`
- `sample1=subset(data1,samp==1)`

R (Select a Stratified sample)

- `data2=read.csv(file.choose())` ← `Library.csv`
- `nh=c(12,17,21)`
- `strata=data2$Stratum`
- `table(strata)`
- `inclusionprobastrata(strata,nh)`
- `st=strata(data2,stratanames=c("Stratum"),size=c(12,17,21), method="srswor")`
- `x=getdata(data2, st)`
- `x`

strata		
1	2	3
69	130	170

R- (Analyse SRS)

- `install.packages("survey")`
- `library(survey)`
- `data4=read.csv(file.choose())` ← sample3.csv
- `strat_design2 <- svydesign(ids=~1, probs=NULL, strata=NULL, variables=NULL, fpc=~FPC, weights=data4$SamplingWeight, data=data4)`
- `svymean(~inq,strat_design2)`

Simple random sample

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	50
Sum of Weights	369

Variable	N	Mean	Std Error of Mean
inq	50	75823	60871

```
> svymean(~inq, strat_design2)
```

```
      mean      SE
```

```
inq 75823 60871
```

R- (Analyse Stratified)

- `install.packages("survey")`
- `library(survey)`
- `data3=read.csv(file.choose())` ← samSRS.csv
- `strat_design <- svydesign(ids=~1, probs=NULL,
strata=~data3$stratum, variables=NULL,
fpc=~data3$FPC,
weights=data3$SamplingWeight, data=data3)`
- `svymean(~inq,strat_design)`

Proportional allocation

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	5
Number of Observations	50
Sum of Weights	369

Variable	Mean	Std Error of Mean
inq	33763	3589.379522

```
> svymean(~inq, strat_design)
      mean      SE
inq 33763 3589.4
```

Stratified Design

- Divide the sampling frame into groups based on stratifying variable(s)
- Select a simple random sample from each stratum
- Put the samples together to estimate

$$\bar{y}_{str} = \sum_{h=1}^H \left(\frac{N_h}{N} \right) \bar{y}_h$$

and standard error

$$\hat{\sigma}_{\bar{y}_{str}} = \sqrt{\sum_{h=1}^H \left(1 - \frac{n_h}{N_h} \right) \left(\frac{N_h}{N} \right)^2 \frac{s_h^2}{n_h}}$$

HW 05

Size	N_h	σ_h	$N_h * \sigma_h$
Under 50	2614	183	478362
50-99	1566	316	494856
100-199	1419	641	909579
200-299	683	1347	920001
300-499	679	2463	1672377
500+	609	7227	4401243
Total	7570		8876418

$$n_h = \left(\frac{\sigma_h N_h}{\sum \sigma_h N_h} \right) n = \left(\frac{478362}{8876418} \right) 1000 = 53.89 \approx 54$$

Table 5.2 Payroll Variance

$$h = 1$$

$$\left(1 - \frac{34}{1614}\right) * \left(\frac{1614}{6570}\right)^2 * \frac{183^2}{34} = 58.1906$$

$$h = 2$$

$$\left(1 - \frac{57}{1566}\right) * \left(\frac{1566}{6570}\right)^2 * \frac{316^2}{57} = 95.9068$$

$$h = 3$$

$$\left(1 - \frac{104}{1419}\right) * \left(\frac{1419}{6570}\right)^2 * \frac{641^2}{104} = 170.7891$$

$$h = 4$$

$$\left(1 - \frac{106}{683}\right) * \left(\frac{683}{6570}\right)^2 * \frac{1347^2}{106} = 156.277$$

$$h = 5$$

$$\left(1 - \frac{192}{679}\right) * \left(\frac{679}{6570}\right)^2 * \frac{2463^2}{192} = 242.0447$$

$$h = 6$$

$$\left(1 - \frac{507}{609}\right) * \left(\frac{609}{6570}\right)^2 * \frac{7227^2}{507} = 148.2500$$

$$\sigma_{\bar{x}}^2 = \sum_{h=1}^H \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}$$

Total = 871.4582

Stratified Design

- estimate proportion

$$\hat{p}_{str} = \sum_{h=1}^H \left(\frac{N_h}{N} \right) \hat{p}_h$$

and standard error

$$\hat{\sigma}_{\hat{p}_{str}} = \sqrt{\sum_{h=1}^H \left(1 - \frac{n_h}{N_h} \right) \left(\frac{N_h}{N} \right)^2 \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}}$$

Example (ACLS) (Source: Sharon Lohr)

- The American Council of Learned Societies (ACLS) used a stratified random sample of selected ACLS societies in seven disciplines to study publication patterns and computer and library use among scholars who belong to one of the member organizations of the ACLS. They want to estimate the percentage and number of respondents of the major societies in those seven disciplines that are female.

Discipline	N_h	n_h	Female (%)
Literature	9100	636	38
Classics	1950	451	27
Philosophy	5500	481	18
History	10850	611	19
Linguistics	2100	493	36
Political Science	5500	575	13
Sociology	9000	588	26
Totals	44000	3835	

Example (ACLS)

Discipline	N_h	n_h	Female (%)
Literature	9100	636	38
Classics	1950	451	27
Philosophy	5500	481	18
History	10850	611	19
Linguistics	2100	493	36
Political Science	5500	575	13
Sociology	9000	588	26
Totals	44000	3835	

estimate proportion

$$\hat{p}_{str} = \sum_{h=1}^H \left(\frac{N_h}{N} \right) \hat{p}_h = \frac{9100}{44000} 0.38 + \frac{1950}{44000} 0.27 + \dots + \frac{9000}{44000} 0.26$$

$$\hat{p}_{str} = 0.2465 \text{ (24.65\%)}$$

Example (ACLS)

Discipline	N_h	n_h	Female (%)
Literature	9100	636	38
Classics	1950	451	27
Philosophy	5500	481	18
History	10850	611	19
Linguistics	2100	493	36
Political Science	5500	575	13
Sociology	9000	588	26
Totals	44000	3835	

Standard error

$$\hat{\sigma}_{\hat{p}_{str}} = \sqrt{\sum_{h=1}^7 \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}} = \left(1 - \frac{636}{9100}\right) \left(\frac{9100}{44000}\right)^2 \frac{0.38(1-0.38)}{636-1} + \dots$$

$$\hat{\sigma}_{\hat{p}_{str}} = 0.0071$$

Example of Stratified design (Source: Sharon Lohr)

- The U.S. government conducts a Census of Agriculture every five years, collecting data on all farms in the 50 states. The census of Agriculture provides data on number of farms, the total acreage devoted to farms, farm size, yield of different crops, and a wide variety of other agricultural measures for each of the counties in United States.
- For this example, we use the four census regions of the United States as strata.

Stratum	N_h	n_h
Northeast	220	21
North Central	1054	103
South	1382	135
West	422	41
Total	3078	300

Example of Stratified design (Source: Sharon Lohr)

- Estimate the total number of acres devoted to farming in the United States in 1992.
- Estimate the proportion of counties with fewer than 200000 acres in farms in the United States in 1992.
- acres92: number of acres devoted to farms, 1992
- Lt200k is 1 if county with fewer than 200k acres in farms; 0 ($\geq 200k$)

The sample size problem

- How much data do I need to get a good answer to my problem?
- Depends on several factors:
 - What are you trying to estimate?
 - How close do you need to get?
 - How variable is your population?
 - What kind of sample design will you use?
- Sometimes a fixed amount of resources are available for data collection. Do I still need to go through this process?
 - Yes! You might as well save the money if the resources are not sufficient to provide an answer with the precision you need

An approach for determining sample size for SRS

The link between sample size and “good enough” is through length of confidence interval

- 95% confidence interval for mean is

$$\bar{y}_{srs} \pm 1.96 * \hat{\sigma}_{\bar{y}}$$

$$\bar{y}_{srs} \pm 1.96 * \underbrace{\frac{S}{\sqrt{n}} \sqrt{(1 - n/N)}}_{\text{margin of error (m.o.e.)}}$$

- Set this to an acceptable size (call it $I_{95\%}$) and solve for n

Solving for n

- $1.96 * \frac{S}{\sqrt{n}} \sqrt{(1 - n/N)} = I_{95\%}$
- This equation can be solved in a two-step process.
First ignore the fpc

$$n_{0,srs} = \frac{(1.96)^2 S^2}{(I_{95\%})^2}$$

- Then “adjust” for the fpc

$$n_{srs} = \frac{n_{0,srs}}{1 + \frac{n_{0,srs}}{N}}$$

But where do we get S?

Solving for n

- $Z_{\alpha/2} * \frac{S}{\sqrt{n}} \sqrt{(1 - n/N)} = I_{(1-\alpha)\%} (\text{margin of error})$
- $I_{(1-\alpha)\%} (\text{margin of error}) = MOE, ME$
- This equation can be solved in a two-step process.
First ignore the fpc

$$n_{0,srs} = \frac{(Z_{\alpha/2} S)^2}{(I_{(1-\alpha)\%})^2}$$

- Then “adjust” for the fpc

$$n_{srs} = \frac{n_{0,srs}}{1 + \frac{n_{0,srs}}{N}}$$

Example: Library inquiry data

Suppose we have no information that would let us calculate a value for S for inquiries prior to sampling. However, we do have circulation values for all 369 members of the population.

circ	
Mean	127885
Standard Error	24284.81
Median	27771
Mode	0
Standard Deviation	466496
Sample Variance	2.18E+11
Kurtosis	107.1163
Skewness	9.45128
Range	6384212
Minimum	0
Maximum	6384212
Sum	47189555
Count	369

Suppose we believe that the relative variability of circulation is similar to that of inquiries; i.e., the CV's are similar

$$\begin{aligned} \text{CV (Coefficient of Variation)} \\ = \frac{S}{\bar{Y}} = 466496/127885 = 3.65 \end{aligned}$$

Solving for n

- Step 1. Set $I_{95\%} = (0.20)\bar{Y}$. Then

$$n_{0,srs} = \frac{4S^2}{((0.20)\bar{Y})^2} = \frac{4 * CV^2}{0.04} = \frac{4 * 3.65^2}{0.04} = 1332!$$

- Step 2 Then “adjust” for the fpc

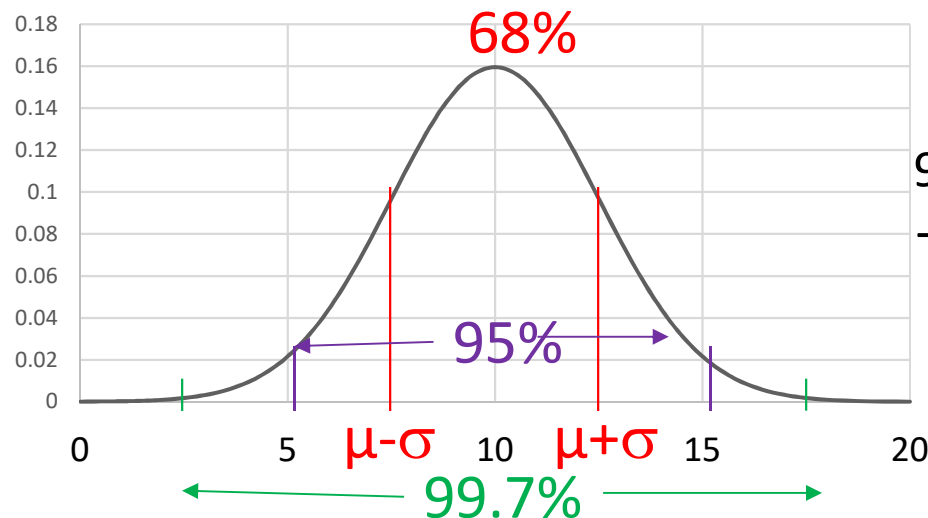
$$n_{srs} = \frac{1332}{1 + \frac{1332}{369}} = 289$$

Unfortunately requires a VERY large sample!!!

Where do we get info required for sample design planning? (con't)

- Next we consider estimation of means
 - We must have a preliminary estimate of S
- Suppose you have NO pilot data or proxy variable on which to base your estimate
 - One approach is to use the “empirical rule”

Perceived shape of data



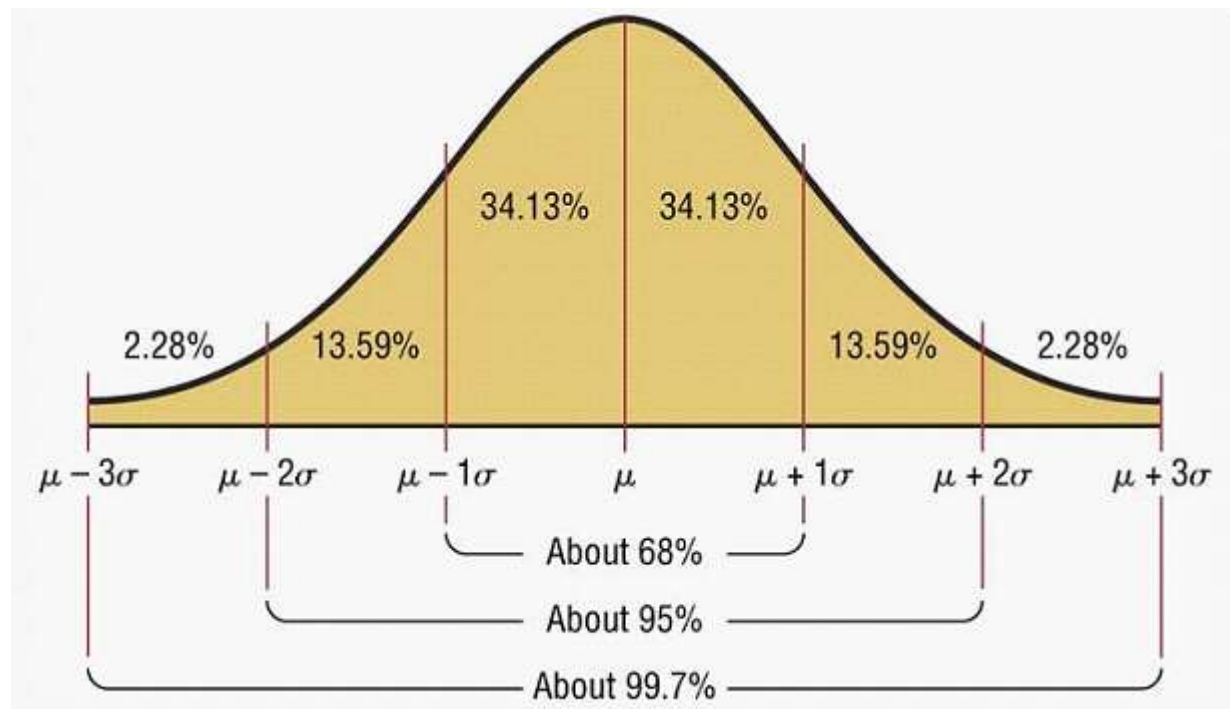
What is S ?

95% of data between 5 and 15
 $\rightarrow \sigma \approx (15-5)/4 = 2.5$

68-95-99.7 Rule (Empirical Rule)

Interpretation

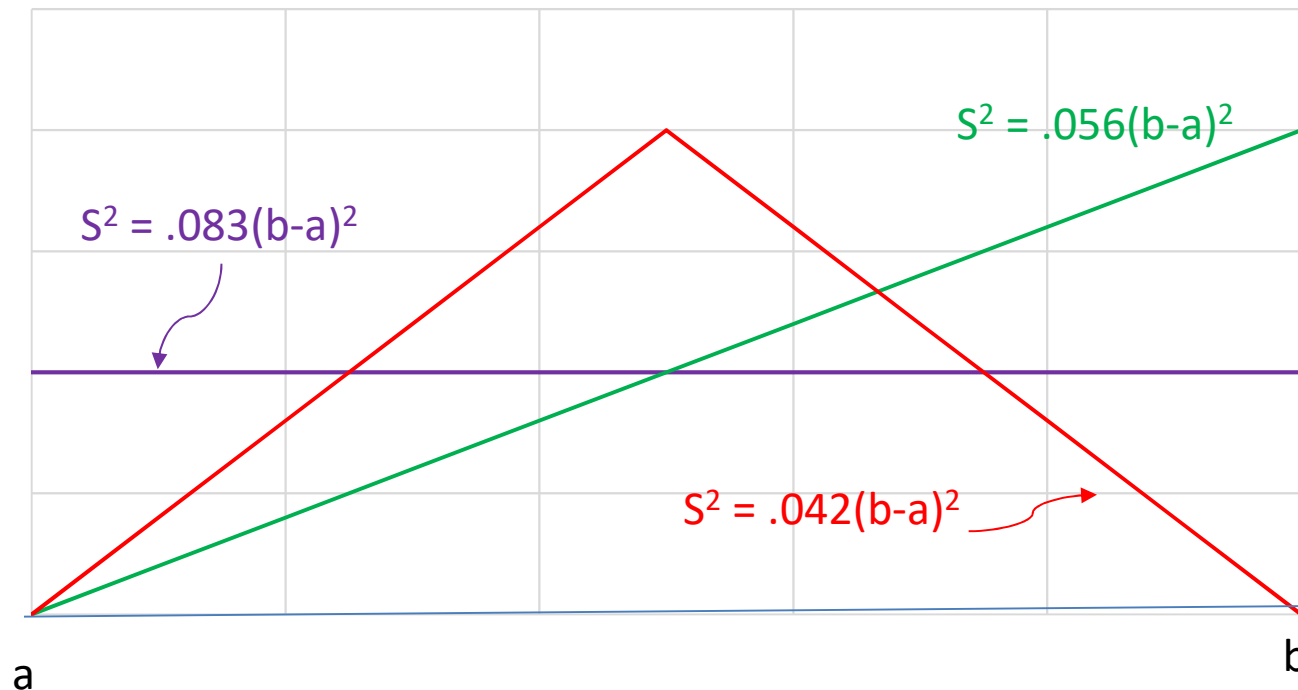
Each number represents the probability of getting an observation 1, 2, and 3 standard deviations away from the mean on a normal density curve.



Where do we get info required for sample design planning? (con't)

- Another method that sometimes works for getting a preliminary estimate of S is to model the shape and range of the data

Three data shapes



Estimating a proportion

- The sample size needed varies according to the parameter of interest
- Let $y_i = 1$ or 0 according to whether the i th unit has some attribute or not.
- For this type of variable, $S^2 = p(1 - p)$, and a 95% confidence interval for p is

$$\hat{p}_{srs} \pm 2 * \underbrace{\sqrt{\frac{p(1-p)}{n} (1 - n/N)}}_{\text{m.o.e.}}$$

- To determine a sample size, we must specify a desired m.o.e. and solve for n

Library Example

- Some libraries have no staff to handle inquiries. Suppose we are interested in estimating what fraction are in this category from a sample.
- Suppose our belief is that p is somewhere near .20 and that we are satisfied with a m. o. e. of $I_{95\%} = 0.10$.

$$\text{Step 1 } n_{0,srs} = \frac{4*(0.2)(0.8)}{(0.10)^2} = 64$$

Step 2 Then “adjust” for the fpc

$$n_{srs} = \frac{64}{1 + \frac{64}{369}} = 55$$

Solving for n for estimating sample proportion

- This equation can be solved in a two-step process.
First ignore the fpc

$$n_{0,srs} = \frac{4p(1-p)}{(I_{95\%})^2}$$

- Then “adjust” for the fpc

$$n_{srs} = \frac{n_{0,srs}}{1 + \frac{n_{0,srs}}{N}}$$

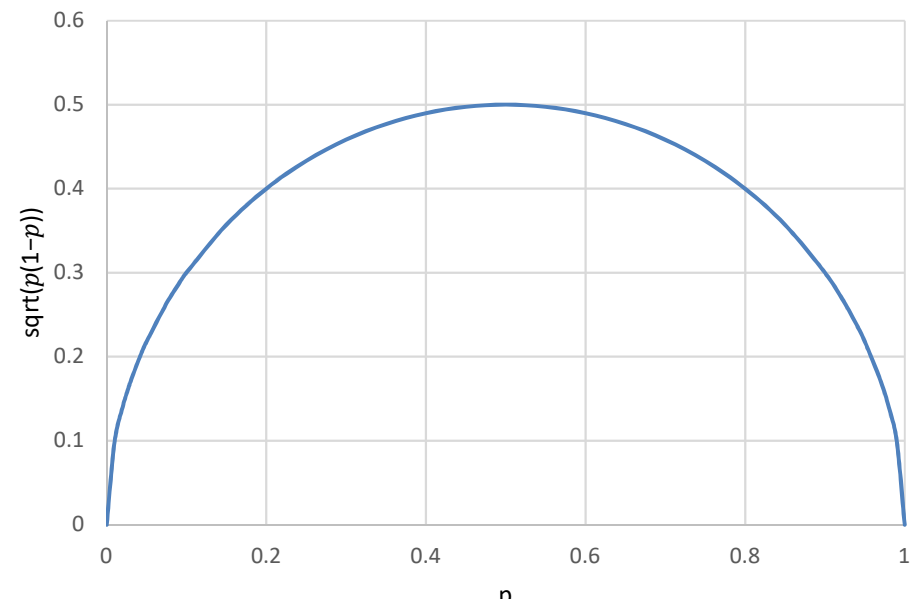
Note that to use this formula, we must have a preliminary estimate of the very parameter we are trying to estimate!

Why is an assumption of $p = \frac{1}{2}$ conservative?

- First we will consider estimation of proportions
 - Common approach: assume $p = \frac{1}{2}$.

- If p is extreme, this is too conservative

Relationship between p and the SD of a 0/1 variable Y



Let's return to the library data

We projected a sample size of 289 (out of the 369 libraries state-wide) would be needed to obtain an adequate estimate of mean inquiries

But we also saw that the precision can be improved tremendously by stratification.

Sample design	Estimate	Sterr
SRS	27,978,768	10,961,257
STR, proportionate	12,457,169	1,324,500
STR, Neyman	9,235,739	282,512

Suppose I decide to use a proportionately allocated stratified sample. What sample size do I need for that?

Design effect

- The design effect (deff) is a measure of the quality of a sample design; i.e., how does it compare to a SRS?
- It is defined as

$$deff_{complex} = \frac{V(\bar{y}_{complex})}{V(\bar{y}_{SRS})}$$

for same sample size.

- Example: Library data

$$deff_{str, prop} = \frac{V(\bar{y}_{complex})}{V(\bar{y}_{SRS})} = \left(\frac{1,324,500}{10,961,257} \right)^2 = .1208$$

The purpose of deff

- deff is an adjustment that can be used for determining sample size.
- If n_{srs} is the sample size needed to achieve a targeted m.o.e. with a SRS, then the sample size needed to achieve that same m.o.e. for a complex design is approximately

$$n_{complex} = deff_{complex} * n_{SRS}$$

- Survey organizations calculate deff's for samples from their commonly used designs so they can use them for planning in the future.
- The formula above needs to be adjusted if the population size is small relative to the sample size, since the fpc enters differently in the two designs

Steps in calculating sample size for complex design when deff is available:

Step 1. Calculate sample size for srs *without fpc*
(called $n_{0,srs}$)

Step 2. Calculate $n_{0,complex} = n_{0,srs} * deff_{complex}$

Step 3. Correct for fpc by calculating

$$n_{complex} = n_{0,complex} / (1 + \frac{n_{0,complex}}{N})$$

Example: Library data

- Step 1. Find $n_{complex}$ needed for estimating mean inquiries using proportionately allocated stratified design

- Calculate n for SRS $n_{0,srs} = \frac{4S^2}{((0.20)\bar{Y})^2} = \frac{4*3.65^2}{0.04} = 1332$

- Step 2. Apply design effect

- $n_{0,complex} = 1332 * .1208 = 161$

- Step 3. Correct for fpc

- $n_{complex} = \frac{161}{\left[1 + \frac{161}{369}\right]} = 112$

Much better than the sample size of 289 that was needed for a SRS!

Summary

- Determining sufficient sample sizes for SRS is straightforward
- It is NOT so straightforward for complex designs
- deff is a sample size adjustment that can be used for obtaining an appropriate sample size
- If your organization collects many samples, it should maintain data allowing calculation of deff's.
- Some survey software will provide estimates of deff as part of the analysis.
 - SAS does this in two of its survey analysis products (SURVEYFREQ and SURVEYREG), but not its other routines.