What Should be the Sample Size for the Study?

Srilakshminarayana G.

Introduction

Sample size estimation is an important step in conducting the sample surveys effectively and efficiently. It is the number of sample responses needed to draw valid inferences about the key parameters of the study. It is directly linked to the permissible error that one allows in drawing the inferences as well as the confidence level at which the results will be presented. To ensure that the requirements are met, the researcher has to work out a plan to reach a balancing point using a scientific procedure. This scientific procedure should include the key parameter of the study, degree of variability, level of precision, confidence level, and, power of the testing procedure etc. Standard sampling designs that discuss sample size estimation can also be used to estimate the sample size. For example, simple random sampling, stratified random sampling etc., discuss estimation and allocation of sample size to different groups.

When one adopts a questionnaire for the survey, it is important to identify and estimate the key parameter of the study in order to estimate the sample size. Sometimes, it is very useful to apply multivariate techniques like principal component analysis (PCA), factor analysis (FA) etc. to group the variables that are associated or correlated (sometimes not noticed by the researcher) and then estimate the key parameter using the pilot sample.

sdmimd

Copyright © 2014 Shri Dharmasthala Manjunatheshwara Research Centre for Management Studies (SDMRCMS), SDMIMD, Mysore

The case writer(s) Srilakshminarayana.G, Assistant Professor - Quantitative Methods may be reached at lakshminarayana@sdmimd.ac.in Author(s) have prepared this case as the basis for class discussion rather than to illustrate either effective or ineffective handling of the situation. This case is fictionalized and any resemblance to actual person or entities is coincidental. This publication may not be digitized, photocopied, or otherwise reproduced, posted, or transmitted, without the permission of SDMRCMS, SDMIMD, Mysore. For Teaching Notes please contact sdmrcms@sdmimd.ac.in.

The case discussed is related to a researcher who wishes to estimate the sample size using the pilot survey and categorical principal component analysis (CATPCA).

The entire discussion is with respect to CATPCA using SPSS and hence the details are presented with respect to options in SPSS. Only few technical details are given and focus is on using CATPCA in estimation of sample size and further discussion is related to the same.

Situation

Mrs. AS, assistant professor is very busy verifying her research reports, she has obtained from her guide. She has to submit the revised version of the preliminary PhD reports that she has submitted to her guide. The guide asked her to work on estimating the sample size for her study. In order to estimate and understand the behaviour of the key parameters of her research, she has conducted a pilot survey. Now, she is in search of a statistician who can help her in understanding the statistical techniques that will be used to analyze the data and relate the same with the objectives of the study. After a long search, she could find Dr.SLN, who has agreed to help her in analysing the data. The first assignment he has agreed, is to estimate the sample size required to conduct the survey effectively and efficiently.

The data considered is in the area of Human resource management and the study is on work-life balance of women employees in private and public sector organizations.

In order to estimate the sample size, Dr.SLN had considered the pilot data of the study.

He had decided to use the following procedure to estimate the sample size.

Step-1: The key parameter considered is the average work life balance score of private and public sector organizations.

Step-2: To estimate the average work-life balance of private as well as public sector organizations, CATPCA is used.

Step-3: First, using CATPCA, different variables considered in the study are grouped based on their correlation structures, separately for private and public sectors. The grouped variables are called as components. Based on the variance accounted for by each component, the maximum number of components are chosen. Using the component loadings, object scores are computed and in turn used to measure the work-life balance among the women employees of private and public sector organizations separately (Detailed procedure is explained in later part of the discussion).

Step-4: Once the work-life balance is measured for private and publics organizations separately, the average work-life balance score is calculated.

Step-5: Using the average score and at 5% level of significance, the sample size is estimated.

In order to estimate the required parameters to estimate the sample size, a pilot survey was conducted.

Pilot Sample

The sample size considered for the pilot study was 120, out of which 93 are from public sector and 27 are from the private sector. The analysis was carried out separately and discussed in later sections. Only those variables that were used to measure the work-life balance were considered.

What is Categorical Principal Component Analysis (CATPCA)?

Categorical principal component analysis (CATPCA) is used to reduce the dimension of the variables under study, when the variables are measured on ordinal or nominal scale. This procedure quantifies the categorical variables. It is also used to find the relationships between the variables that have not been identified by the researcher (latent relationships amongst the variables under study). The goal is to reduce the original set of variables into a smaller set of uncorrelated components that represent most of the information found in the

original variables. The technique is most useful when a large number of variables prohibits effective interpretation of the relationships between objects (subjects and units). If the variables are measured on interval or ratio, the method will be same as standard principal component analysis (PCA).

Standard principal components analysis assumes linear relationships between numeric variables. On the other hand, the optimal-scaling approach allows variables to be scaled at different levels. Categorical variables are optimally quantified in the specified dimensionality. As a result, nonlinear relationships between variables can be modelled.

The use of Categorical Principal Components Analysis is most appropriate when one looks at studying the patterns of variation in a single set of variables of mixed optimal scaling levels. This technique attempts to reduce the dimensionality of a set of variables while accounting for as much of the variation as possible. Scale values are assigned to each category of every variable so that these values are optimal with respect to the principal components solution.

In the present case, CATPCA is used to identify the relationships between the variables like job satisfaction, job demand, and supervision etc., measured using ordinal scale.

Critical Components of CATPCA Required for the Study

a. Data

The variables are measured on a five point Likert scale with 5 denoting strongly agree, 4- Agree, 3-Neutral, 2-Disagree, 1-Strongly agree. String variable values are always converted into positive integers by ascending alphanumeric order. User-defined missing values, system-missing values, and values less than 1 are considered missing; one can recode or add a constant to variables with values less than 1 to make them non-missing.

b. Assumption

The data considered for the study contains at least three valid cases. The analysis is based on positive integer data and measured on ordinal scale.

c. Scale and Weight in CATPCA

The original scale is ordinal and the same is preserved even in the scale options. The order of the categories of the observed variable is preserved in the optimally scaled variable. Category points will be on a straight line (vector) through the origin. The resulting transformation fits better than the spline ordinal transformation but is less smooth.

d. Discretization

The Discretization dialog box allows to select a method of recoding the variables. String variables are always converted into positive integers by assigning category indicators according to ascending alphanumeric order. Discretization for string variables applies to these integers. Other variables are left alone by default. The discretized variables are then used in the analysis.

The discretization method used in the case is ranking method.

e. Missing Values

The missing values/cases are excluded from the analysis.

f. Normalization Method

The normalization method used for analysis is the variable principal method.

This option optimizes the association between variables. The coordinates of the variables in the object space are the component loadings (correlations with principal components, such as dimensions and object scores). This is useful when we are primarily interested in the correlation between the variables.

g. Output

The Output dialog box allows to produce tables for object scores, component loadings, iteration history, correlations of original and transformed variables, the variance accounted for per variable and per dimension, category quantifications for selected variables, and descriptive statistics for selected variables.

h. The Save dialog box allows to save discretized data, object scores, transformed values, and approximations to an external IBM SPSS Statistics data file or dataset in the current session. One can also save transformed values, object scores, and approximations to the active dataset.

CATPCA using SPSS

The following give the output of the CATPCA using SPSS and further interpretation is provided accordingly. First, the analysis is presented for Private and then for Public sector.

Private Sector-Under this group, we have 27 cases and the analysis of the same is presented in the following tables.

Table-1

Case Processing Summary

Valid Active Cases	27
Active Cases with Missing Values	0
Supplementary Cases	0
Total	27
Cases Used in Analysis	27

Table-2
Model Summary

Dimension	Cronbach's Alpha	Variance Accounted For	
		Total (Eigenvalue)	% of Variance
1	.960	17.126	32.935
2	.953	15.329	29.479
3	.931	11.530	22.173
4	.797	4.580	8.808
Total	.999[a]	48.566	93.396

a. Total Cronbach's Alpha is based on the total Eigenvalue.

One can note that four components are extracted using CATPCA, which accounts for 93% of the variation and the value of the Cronbach alpha suggests that the model is reliable.

Table-3
Component Loadings

Variable -Code	Dimension			
	1	2	3	4
C1a	.920	-.225	.319	.027
C1b	.131	.748	.430	.259
C1c	-.878	.261	-.392	-.069
C1d	-.524	.616	-.163	.541
C1e	-.147	-.559	.423	.526
C2a	-.632	.069	.672	-.182
C2b	.324	.675	-.653	.096

C2c	.554	-.253	.427	-.341
C2d	.370	.677	-.621	-.071
C3a	.878	-.261	.392	.069
C3b	-.130	-.839	-.500	.154
C3c	.878	-.261	.392	.069
C4a	.878	-.261	.392	.069
C4b	.515	-.813	-.213	.143
C4c	.796	.469	-.364	.033
C5a	.324	.675	-.653	.097
C5b	.905	-.219	.359	.048
C5c	.218	.763	.369	.401
C5d	.734	.461	-.317	.108
C5e	.435	-.647	-.191	.476
C5f	.373	.676	-.625	.005
D1a	-.201	-.608	-.341	-.016
D1b	.101	.792	.581	-.128
D1c	.087	.804	.571	-.110
D1d	.237	.609	-.564	-.308
D1e	.880	-.258	.390	.068
D2a	-.121	-.778	-.593	.144
D2b	-.101	-.792	-.581	.128
D2c	-.834	-.022	-.548	-.021
D2d	.103	.794	.580	-.124

D2e	.158	.085	-.176	.945
E1	.423	-.078	-.665	-.490
E2	.498	-.728	-.254	-.033
E3	-.874	.261	-.401	-.071
E4	.446	-.749	-.155	.147
E5	.878	-.226	.411	.081
E6	.789	.502	-.333	.057
E7	.796	.469	-.364	.033
E8	.744	.470	-.320	.018
F1	.325	.674	-.652	.106
F2	.927	.302	-.181	.109
F3	.125	.083	-.155	.899
F4	-.878	.261	-.392	-.069
F5	-.548	.117	.593	.524
F6	.367	.627	-.677	.008
F7	.312	.606	.541	.033
F8	-.351	.027	.804	.405
F9	-.222	-.710	-.636	.193
F10	.648	-.080	-.692	-.014
F11	-.249	.716	.014	.548
F12	-.607	.266	-.456	.556
F13	.511	-.814	-.208	.162
Variable Principal Normalization.				

Based on the component loadings of the respective variables in the study, the variables are grouped. For example, all those variables whose component loadings are at least 0.5 are grouped to form individual components.

Table-4

Object Scores

Respondent	Dimension			
	1	2	3	4
1	.671	-.171	-.736	-.690
2	.475	-.013	-.403	-.952
3	.671	-.171	-.736	-.690
4	.671	-.171	-.736	-.690
5	.671	-.171	-.736	-.690
6	.671	-.171	-.736	-.690
7	.671	-.171	-.736	-.690
8	.671	-.171	-.736	-.690
9	.671	-.171	-.736	-.690
10	.671	-.171	-.736	-.690
11	.430	-.020	-.174	2.026
31	-3.104	.922	-1.385	-.244
32	-3.104	.922	-1.385	-.244
33	.403	.451	-.802	3.571
34	.260	.137	-.127	1.685
35	.491	.131	-.521	.191
39	.345	-.158	-.418	-.070

41	.514	.070	-.350	1.250
71	-.601	-1.272	1.209	.014
72	-.601	-1.272	1.209	.014
73	.286	2.241	1.643	-.363
94	-.601	-1.272	1.209	.014
95	-.601	-1.272	1.209	.014
96	.286	2.241	1.643	-.363
117	-.601	-1.272	1.209	.014
118	-.601	-1.272	1.209	.014
119	.286	2.241	1.643	-.363

Note that the above table gives the object scores for the 27 cases and the same are used to measure the work-life balance, which will be discussed in the next section in detail.

Public Sector-Under this group, we have 93 cases and the analysis is presented in the following tables.

Table-5

Case Processing Summary

Valid Active Cases	93
Active Cases with Missing Values	0
Supplementary Cases	0
Total	93
Cases Used in Analysis	93

Table-6

Model Summary

Component	Cronbach's Alpha	Variance Accounted For	
		Total (Eigenvalue)	% of Variance
1	.972	21.306	40.973
2	.902	8.667	16.667
3	.849	5.960	11.462
4	.769	4.074	7.834
5	.704	3.233	6.217
Total	.996[a]	43.240	83.155
a. Total Cronbach's Alpha is based on the total Eigenvalue.			

Table-7

Component Loadings

Variable-Code	Dimension				
	1	2	3	4	5
C1a	.570	-.343	-.174	-.121	-.539
C1b	-.128	.402	-.033	-.624	.378
C1c	.517	-.481	-.301	.222	-.397
C1d	.158	-.399	.032	.650	-.443
C1e	-.789	.063	-.020	.313	-.314
C2a	-.869	.246	-.417	-.001	-.061
C2b	.870	-.166	.066	-.238	.283

C2c	.533	-.375	.323	-.538	.057
C2d	.907	-.331	.059	-.055	.064
C3a	.316	-.054	.083	-.409	-.224
C3b	-.216	.086	.941	.037	-.181
C3c	-.715	.274	-.378	-.315	-.110
C4a	-.643	.656	-.355	.048	.081
C4b	.279	.642	.317	.117	-.309
C4c	.756	.075	-.017	.428	-.378
C5a	.868	-.114	.044	-.251	.244
C5b	.897	-.237	-.105	-.075	-.079
C5c	.837	.226	-.331	.187	-.107
C5d	.943	-.291	.066	.096	.049
C5e	-.305	.533	.361	-.271	-.399
C5f	.901	-.186	.026	-.033	-.155
D1a	-.817	.096	-.041	.281	-.378
D1b	.175	-.115	-.952	.017	.157
D1c	.256	-.161	-.929	.042	.156
D1d	.378	-.615	-.193	.417	.135
D1e	-.448	-.788	-.112	-.317	-.184
D2a	-.925	.307	-.011	-.070	.012
D2b	-.284	-.078	.670	.291	.386
D2c	-.175	-.155	.752	.139	.375
D2d	-.757	.083	-.260	-.127	.370

D2e	.113	-.358	.204	.099	.582
E1	-.527	-.660	-.095	-.184	-.020
E2	-.490	-.453	.221	.515	-.095
E3	-.533	-.564	-.135	-.099	-.048
E4	.626	.498	.199	.177	-.154
E5	.848	.151	-.133	-.327	.171
E6	.850	.226	-.275	-.069	-.203
E7	.678	.410	-.373	-.106	-.064
E8	.627	.613	.090	.219	.102
F1	.919	-.294	.065	.030	.033
F2	.943	-.232	.072	.064	.056
F3	.813	-.226	.362	-.053	-.135
F4	-.815	.303	-.026	-.181	-.180
F5	.472	.771	.102	.328	.177
F6	.431	-.597	-.093	.286	.147
F7	-.740	-.068	-.326	.410	-.004
F8	-.487	-.762	-.113	-.318	-.186
F9	-.527	-.484	.459	-.270	-.197
F10	.459	.683	.106	.328	.169
F11	-.245	-.072	-.280	.580	.475
F12	-.566	-.579	-.084	.241	.275
F13	-.623	-.450	.381	.144	.104
Variable Principal Normalization.					

Table-8

Object Scores

Respondent	Dimension				
	1	2	3	4	5
12	-.992	.344	-.075	-.098	-.047
13	-.549	.668	.287	.128	-.333
14	.822	-1.002	.176	.198	-.060
15	1.151	.091	.248	.489	-.054
16	.249	-.865	-.051	1.158	.292
17	.249	-.865	-.051	1.158	.292
18	.874	-.593	.762	-1.405	.352
19	1.181	-.055	.553	-2.516	-1.224
20	.726	-.378	.641	-.541	.494
21	.389	-1.111	.962	-.226	2.533
22	.633	-.850	.526	-.664	1.347
23	.608	-.933	.999	-.705	1.212
24	1.287	-.093	.257	-1.948	-1.242
25	1.256	-.099	.784	-1.505	-1.051
26	1.407	-.418	.038	-1.987	-1.590
27	1.383	-.153	-.173	-2.629	-1.875
28	.567	-.844	.790	-.979	.838
29	1.863	1.037	-.238	-2.459	-3.375
30	.342	-1.436	.859	.678	2.689
36	.380	-1.184	.665	1.017	1.376

37	.829	-.079	.327	-.098	.242
38	.524	-.509	.111	-.026	-.441
40	.511	-.506	.252	.593	.136
42	.766	-.556	.854	.803	-1.437
43	.786	-.288	-.297	-.086	-1.271
44	.577	-.906	-.163	1.020	-.963
45	.468	-1.391	1.016	.370	2.344
46	.465	-1.564	.470	.704	1.667
47	1.026	-.660	.309	-1.145	-.586
48	.555	-1.164	.676	-.153	1.997
49	.545	-1.483	.686	-.271	2.696
50	.589	-.731	.540	-.791	.787
51	-.992	.344	-.075	-.098	-.047
52	-.992	.344	-.075	-.098	-.047
53	-.992	.344	-.075	-.098	-.047
54	-.992	.344	-.075	-.098	-.047
55	-.992	.344	-.075	-.098	-.047
56	-.992	.344	-.075	-.098	-.047
57	-.992	.344	-.075	-.098	-.047
58	-.992	.344	-.075	-.098	-.047
59	-.992	.344	-.075	-.098	-.047
60	-.992	.344	-.075	-.098	-.047
61	-.992	.344	-.075	-.098	-.047

62	-.992	.344	-.075	-.098	-.047
63	1.180	-.473	-5.153	-.204	.994
64	.447	-1.384	.057	2.436	-1.825
65	.810	-.461	.680	-.971	.100
66	1.982	3.235	.427	1.374	.743
67	-.992	.344	-.075	-.098	-.047
68	-.992	.344	-.075	-.098	-.047
69	-.992	.344	-.075	-.098	-.047
70	-.992	.344	-.075	-.098	-.047
74	.447	-1.384	.057	2.436	-1.825
75	.810	-.461	.680	-.971	.100
76	1.982	3.235	.427	1.374	.743
77	-.992	.344	-.075	-.098	-.047
78	-.992	.344	-.075	-.098	-.047
79	-.992	.344	-.075	-.098	-.047
80	-.992	.344	-.075	-.098	-.047
81	-.992	.344	-.075	-.098	-.047
82	-.992	.344	-.075	-.098	-.047
83	-.992	.344	-.075	-.098	-.047
84	-.992	.344	-.075	-.098	-.047
85	-.992	.344	-.075	-.098	-.047
86	1.180	-.473	-5.153	-.204	.994
87	.447	-1.384	.057	2.436	-1.825

88	.810	-.461	.680	-.971	.100
89	1.982	3.235	.427	1.374	.743
90	-.992	.344	-.075	-.098	-.047
91	-.992	.344	-.075	-.098	-.047
92	-.992	.344	-.075	-.098	-.047
93	-.992	.344	-.075	-.098	-.047
97	.447	-1.384	.057	2.436	-1.825
98	.810	-.461	.680	-.971	.100
99	1.982	3.235	.427	1.374	.743
100	-.992	.344	-.075	-.098	-.047
101	-.992	.344	-.075	-.098	-.047
102	-.992	.344	-.075	-.098	-.047
103	-.992	.344	-.075	-.098	-.047
104	-.992	.344	-.075	-.098	-.047
105	-.992	.344	-.075	-.098	-.047
106	-.992	.344	-.075	-.098	-.047
107	-.992	.344	-.075	-.098	-.047
108	-.992	.344	-.075	-.098	-.047
109	1.180	-.473	-5.153	-.204	.994
110	.447	-1.384	.057	2.436	-1.825
111	.810	-.461	.680	-.971	.100
112	1.982	3.235	.427	1.374	.743
113	-.992	.344	-.075	-.098	-.047

114	-.992	.344	-.075	-.098	-.047
115	-.992	.344	-.075	-.098	-.047
116	-.992	.344	-.075	-.098	-.047
120	.447	-1.384	.057	2.436	-1.825

The object scores represent the score of each respondent with respect to his responses to the question asked. One can note that based on an individual's response, we can measure the degree of work-life balance.

Measuring the Work-Life Balance

The object scores along with the variance explained by each component are considered to compute the work-life balance score. Here, the variance explained by each component is considered as a weight in calculation of the score.

Public Sector

There are 93 respondents under this sector and the object scores are presented in table-8. The percentage of variance explained is given in table-4. Using both the work-life balance scores for each respondent are calculated. For example, the score for the 12th respondent is calculated as in the following. The object scores for the 12th respondent are given for each component in the following table

-.992	.344	-.075	-.098	-.047

Now these scores are multiplied with the corresponding variance percentage explained respectively and the total is taken as the score of the 12th respondent. That is

41%*-0.992+ 16.7%*0.344+11.5%*-0.075+7.8%*-0.098+6.2%*-0.047=-0.368306729

Similarly, the scores for other respondent are calculated and given in the following table

Table 9

Measure of Work - Live - Balance

-0.368306729	-0.368306729	-0.368306729	-0.368306729	0.256508337
-0.091373176	-0.368306729	-0.368306729	-0.368306729	0.198949078
0.201607521	-0.368306729	-0.368306729	-0.368306729	0.157341181
0.550139366	-0.368306729	-0.368306729	-0.368306729	0.506156221
0.061152164	0.036486045	-0.368306729	0.036486045	0.219627492
0.061152164	0.263245628	-0.140074695	-0.368306729	0.199965237
0.258366426	1.554039569	0.036486045	-0.368306729	0.371532253
0.264905016	-0.368306729	0.263245628	-0.368306729	0.11320769
0.296202105	-0.368306729	-0.368306729	-0.368306729	-0.368306729
0.224376374	-0.368306729	-0.368306729	-0.140074695	-0.368306729
0.209817123	-0.368306729	-0.368306729	0.036486045	-0.140074695
0.228165879	-0.368306729	-0.368306729	0.263245628	0.036486045
0.311288729	-0.368306729	-0.368306729	-0.368306729	0.263245628
0.404697805	-0.368306729	-0.368306729	-0.368306729	-0.368306729
0.142523326	0.219554159	-0.368306729	-0.368306729	-0.368306729
0.208813471	0.223052465	0.251171575	-0.368306729	-0.368306729
0.292818807	0.201054082	0.263245628	-0.368306729	0.036486045
0.154390911	0.168390962	-0.368306729	0.086516442	

Note that four values have been removed as outliers. Using these scores, the average score and standard deviation have been calculated. The following table give the estimated sample size for the public sector organizations

Table 10

Sample Size Determination

Confidence Level Desired	95%
Half-Width Desired	0.01
Population Stdev.	0.341358638
Minimum Sample Size	4477

Private Sector

There are 27 respondents under private sector and work-life balance scores are calculated as in the case of public sector organizations. The following table gives the work-life scores.

Table 11

Measure of Work - Life - Balance

-0.05311669	0.402527476
-0.02081363	0.245972696
-0.05311669	0.10173117
-0.05311669	-0.03164484
-0.05311669	0.22246973
-0.05311669	-0.30352597
-0.05311669	-0.30352597
-0.05311669	-0.30352597
-0.05311669	-0.30352597
-0.05311669	-0.30352597
0.275617723	-0.30352597

Table 12

Sample Size Determination

The following table gives the sample size estimated for private sector organizations

Confidence Level Desired	95%
Half-Width Desired	0.01
Population Stdev.	0.200295024
Minimum Sample Size	1542

Note

In both public as well as private sectors, the half width is chosen as 0.01 at 95% confidence level, so that the actual average work-life score will be estimated with a distance of 0.01.

Dr.SLN has concluded the estimation of the sample size and suggested the researcher to collect the respective number of respondents to estimate the parameters to conclude the study.

As continuation of the case, one can look at designing a sampling technique to collect the sample responses. (Hint: One can divide the design into three stages. First stage can be stratified random sampling, second and third stages can be cluster and convenient sampling techniques respectively).