UNIVERSITY OF SWAZILAND

FINAL EXAMINATION PAPER 2012

TITLE OF PAPER : SAMPLE SURVEY THEORY

COURSE CODE : ST306

TIME ALLOWED : TWO (2) HOURS

REQUIREMENTS : CALCULATOR AND STATISTICAL TABLES

INSTRUCTIONS : **ANSWER ANY THREE QUESTIONS**

Question 1

[20 marks, 8+4+4+4]

- (a) A sample of n=1000 cases (elements) yields the sample totals $\sum y_j=2000$, and $\sum y_j^2=40000$. How large is the standard error of the mean \bar{y} of the sample? What assumptions did you make? How might your answers change if the assumptions do not fit the actual situation?
- (b) A survey of the 9^{th} graders in Ndunayithini is intended to determine the proportion intending to go to a four-year college. A preliminary estimate of p=0.55 was obtained from a small informal survey. How large must the survey be to provide an estimator with error at most 0.05 with probability at least 99%?
- (c) A survey is to be made on the prevalence of the common diseases in a large population. For any disease that affects at least 1% of the individuals in the population, it is desired to estimate the total number of cases, with a coefficient of variation of not more than 20%. What size of a simple random sample is needed, assuming that the presence of the disease can be recognized without mistakes?
- (d) An opinion poll on Swaziland's health concern was conducted by the Swaziland National Aids Program between April 10-15, 2011. The survey reported that 89% of adults consider AIDS as the most urgent health problem of the US, with a margin of error of $\pm 3\%$. The result was based on telephone interviews of 872 adults.
 - (i) What was the target population?
 - (ii) What was the sample population?
 - (iii) How was the survey was conducted?
 - (iv) How was the sample selected?

Question 2

[20 marks, 2+6+6+6]

A simple random sample of 10 hospitals was selected from a population of 33 hospitals that had received state funding to upgrade their emergency medical services. Within each of the selected hospitals, the records of all patients hospitalised in the past 12 months for traumatic injuries (i.e. accidents, poisonings, violence, burns, etc.) were examined. The numbers of patients hospitalised for trauma conditions and the numbers who died for the selected hospitals are given below.

	Number of patients	
Hospital	hospitalised	Number with trauma
	for trauma conditions	conditions who died
1	560	4
2	190	4
3	260	2
4	370	4
5	190	4
6	. 130	0
7	170	9
8	170	2
9	60	0
10	110	1

- (a) Explain why this design may be considered as a cluster sample. What are the first-stage and second-stage units?
- (b) Obtain a point estimate and an approximate 95% confidence interval for the total number of persons hospitalised for trauma conditions for the 33 hospitals. State the properties of your estimator.
- (c) Obtain a point estimate of the proportion of persons who died among those hospitalised for trauma conditions for the 33 hospitals, using the cluster totals. Hence calculate an approximate 95% confidence interval for this proportion, and comment on the validity of the assumptions necessary for this calculation.
- (d) Give reasons why, for this survey, cluster sampling might be preferred to stratified random sampling. What might be the drawbacks of cluster sampling? Discuss, with reasons, any improvements you might make if another survey was being planned on the same topic.

Question 3 [20 marks]

The NIC (Pty) Ltd Corporation wishes to obtain information on the effectiveness of a business machine. A number of division heads will be interviewed by telephone and asked to rate the equipment on a numerical scale. The divisions are located in Swaziland, South Africa and Botswana. The costs are larger for interviewing division heads located outside Swaziland. The accompanying table gives the costs per interview, approximate variances of the ratings, and N_i 's that have been established. The corporation wants to estimate the average rating with $Var(\bar{y}_{st}) = 0.1$. Choose the sample size n that achieves this bound, and find the appropriate allocation.

	Swaziland	South Africa	Botswana
Cost (in SZL)	9	35	36
Variance	2.25	3.24	3.24
N_i	112	68	39

Question 4

[20 marks, 4+4+6+6]

- (a) Consider a population of size N=5 divided into two strata where the response (y) values for the first stratum are 2, 5, and 8 and for the second stratum are 10 and 13. A stratified random sample consisting of one observation from each stratum will be taken. Let y_1 denote the sample observation from the first stratum and y_2 the sample observation from the second stratum.
 - (i) Let $\bar{y} = \frac{1}{2}(y_1 + y_2)$. Derive the sampling distribution of \bar{y} and show that it is a biased estimator of the population mean μ .
 - (ii) Let $\bar{y}_s = \frac{3}{5}y_1 + \frac{2}{5}y_2$. Derive the sampling distribution of \bar{y}_s and show that it is an unbiased estimator of μ .
- (b) Consider a population of farms on a 25×25 grid of varying sizes and shapes. If we randomly select a single square on this grid, then letting x_i = the area of farm i and A = 625 total units, the probability that farm i is selected is: $p_i = \frac{x_i}{A} = \frac{x_i}{625}$.

$y_i = Workers$	$p_i = rac{x_i}{A} = rac{Size \; of \; Farm}{Total \; Area}$
2	5/625
8	28/625
4	12/625
8	14/625
3	13/625

The table above shows a replacement sample of 5 farms selected with probability-proportional-to-size (PPS). Compute:

- (i) The estimated number of workers (and associated standard errors).
- (ii) The estimated number of farms (and associated standard errors).

using the Hansen-Hurwitz estimator.

Question 5

[20 marks, 6+8+6]

A simple random sample of 1 in 20 households in a small town provided the following data about the availability of cars and the number of adults in households.

Number of cars (y_i) in	Adults in household (x_i)					
the household	1	2	3	4	5	Total
0	58	127	9	6	0	200
1	68	140	27	4	1	240
2	4	30	5	8	3	50
3	0	3	4	2	1	10
Total	130	300	45	20	5	500

Obtain point estimates, and approximate 95% confidence intervals for the following (NB: Summing over all 500 households, $\sum x_i y_i = 795$):

- (a) the total number of cars in the town's households,
- (b) the ratio of cars per adult in the town's households,
- (c) the proportion of households with 1 or more cars per adult.

Useful formulas

$$\begin{split} s^2 &= \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1} \\ \hat{\mu}_{srs} &= \bar{y} \\ \hat{\rho}_{srs} &= N \hat{\mu}_{srs} \\ \hat{\rho}_{srs} &= N \hat{\mu}_{srs} \\ \hat{\rho}_{srs} &= \sum_{i=1}^{n} \frac{y_i}{n} \\ \hat{\tau}_{srs} &= N \hat{\mu}_{srs} \\ \hat{\rho}_{srs} &= \sum_{i=1}^{n} \frac{y_i}{n} \\ \hat{\tau}_{hh} &= \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{n} \\ \hat{\tau}_{hh} &= \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{n} \\ \hat{\tau}_{hh} &= \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{n} \\ \hat{\tau}_{hh} &= \frac{\hat{\tau}_{hh}}{N} \\ \hat{\tau}_{hl} &= \sum_{i=1}^{\nu} \frac{y_i}{\pi_i} \\ \hat{\tau}_{hl$$

$$\hat{\tau}_{cl} = \frac{M}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} y_{ij} = \frac{N}{n} \sum_{i=1}^{n} \sum_{j=1}^{L} y_{ij} = \frac{N}{n} \sum_{i=1}^{n} y_{i} = N\bar{y}$$

$$\hat{\mu}_{cl} = \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} y_{ij} = \frac{1}{nL} \sum_{i=1}^{n} y_{i} = \frac{\bar{y}}{L} = \frac{\hat{\tau}_{cl}}{M}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{\hat{\tau}_{cl}}{N}$

$$\hat{\mathsf{V}}(\hat{\tau}_{cl}) = N(N-n)\frac{s_u^2}{n} \qquad \qquad \hat{\mathsf{V}}(\hat{\mu}_{cl}) = \frac{N(N-n)}{M^2}\frac{s_u^2}{n}$$

where $s_u^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$.

$$\hat{\mu}_1 = \bar{y} = \frac{\hat{\tau}_{cl}}{N} \qquad \qquad \hat{V}(\hat{\mu}_1 = \frac{N-n}{N} \frac{s_u^2}{n})$$

The formulas for systematic sampling are the same as those used for one-stage cluster sampling. Change the subscript *cl* to *sys* to denote the fact that data were collected under systematic sampling.

$$\hat{\mu}_{c(a)} = \frac{\sum_{i=1}^{n} y_{i}}{\sum_{i=1}^{n} M_{i}} = \frac{\sum_{i=1}^{n} y_{i}}{m} \qquad \hat{V}(\hat{\mu}_{c(a)}) = \frac{(N-n)N}{n(n-1)M^{2}} \sum_{i=1}^{n} M_{i}^{2} (\bar{y} - \hat{\mu}_{c(a)})^{2}$$

$$\hat{\mu}_{c(b)} = \frac{N}{M} \frac{\sum_{i=1}^{n} y_{i}}{n} = \frac{N}{nM} \sum_{i=1}^{n} y_{i} \qquad \hat{V}(\hat{\mu}_{c(b)}) = \frac{(N-n)N}{n(n-1)M^{2}} \sum_{i=1}^{n} (y_{i} - \bar{y})^{2} = \frac{(N-n)N}{nM^{2}} s_{u}^{2}$$

$$\hat{p}_{c} = \frac{\sum_{i=1}^{n} p_{i}}{n} \qquad \hat{V}(\hat{p}_{c}) = \left(\frac{N-Nn}{nN}\right) \sum_{i=1}^{n} \frac{(p_{i} - \hat{p}_{c})^{2}}{n-1} = \left(\frac{1-f}{n}\right) \sum_{i=1}^{n} \frac{(p_{i} - \hat{p}_{c})^{2}}{n-1}$$

$$\hat{p}_{c} = \frac{\sum_{i=1}^{n} y_{i}}{\sum_{i=1}^{n} M_{i}} \qquad \hat{V}(\hat{p}_{c}) = \left(\frac{1-f}{n\bar{m}^{2}}\right) \frac{\sum_{i=1}^{n} (y_{i} - \hat{p}_{c}M_{i})^{2}}{n-1}$$

To estimate τ , multiply $\hat{\mu}_{c(\cdot)}$ by M. To get the estimated variances, multiply $\hat{V}(\hat{\mu}_{c(\cdot)})$ by M^2 . If M is not known, substitute M with Nm/n. $\bar{m} = \sum_{i=1}^n M_i/n$.

$$n \text{ for } \mu \text{ SRS} \qquad n = \frac{N\sigma^2}{(N-1)(d^2/z^2) + \sigma^2}$$

$$n \text{ for } \tau \text{ SRS} \qquad n = \frac{N\sigma^2}{(N-1)(d^2/z^2N^2) + \sigma^2}$$

$$n \text{ for } p \text{ SRS} \qquad n = \frac{Np(1-p)}{(N-1)(d^2/z^2) + p(1-p)}$$

$$n \text{ for } \mu \text{ SYS} \qquad n = \frac{N\sigma^2}{(N-1)(d^2/z^2) + \sigma^2}$$

$$n \text{ for } \tau \text{ SYS} \qquad n = \frac{N\sigma^2}{(N-1)(d^2/z^2) + \sigma^2}$$

$$n \text{ for } \tau \text{ SYS} \qquad n = \frac{\sum_{h=1}^L N_h^2(\sigma_h^2/w_h)}{N^2(d^2/z^2) + \sum_{h=1}^L N_h\sigma_h^2}$$

$$n \text{ for } \tau \text{ STR} \qquad n = \frac{\sum_{h=1}^L N_h^2(\sigma_h^2/w_h)}{N^2(d^2/z^2N^2) + \sum_{h=1}^L N_h\sigma_h^2}$$

$$n \text{ for } \tau \text{ STR} \qquad n = \frac{\sum_{h=1}^L N_h^2(\sigma_h^2/w_h)}{N^2(d^2/z^2N^2) + \sum_{h=1}^L N_h\sigma_h^2}$$

where $w_h = \frac{n_h}{n}$. Allocations for STR μ :

$$n_{h} = (c - c_{0}) \left(\frac{N_{h} \sigma_{h} / \sqrt{c_{h}}}{\sum_{k=1}^{L} N_{k} \sigma_{k} \sqrt{c_{k}}} \right) \qquad (c - c_{0}) = \frac{\left(\sum_{k=1}^{L} N_{k} \sigma_{k} / \sqrt{c_{k}} \right) \left(\sum_{k=1}^{L} N_{k} \sigma_{k} \sqrt{c_{k}} \right)}{N^{2} (d^{2} / z^{2}) + \sum_{k=1}^{L} N_{k} \sigma_{k}^{2}}$$

$$n_{h} = n \left(\frac{N_{h}}{N} \right) \qquad n = \frac{\sum_{k=1}^{L} N_{k} \sigma_{k}}{N^{2} (d^{2} / z^{2}) + \frac{1}{N} \sum_{k=1}^{L} N_{k} \sigma_{k}^{2}}$$

$$n_{h} = n \left(\frac{N_{h} \sigma_{h}}{\sum_{k=1}^{L} N_{k} \sigma_{k}} \right) \qquad n = \frac{\left(\sum_{k=1}^{L} N_{k} \sigma_{k} \right)^{2}}{N^{2} (d^{2} / z^{2}) + \sum_{k=1}^{L} N_{k} \sigma_{k}^{2}}$$

Allocations for STR τ :

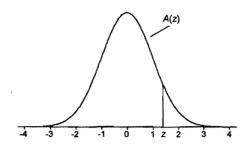
change
$$N^2(d^2/z^2)$$
 to $N^2(d^2/z^2N^2)$

Allocations for STR p:

$$n_h = n \left(\frac{N_i \sqrt{p_h(1 - p_h)/c_h}}{\sum_{k=1}^L N_k \sqrt{p_k(1 - p_k)/c_k}} \right) \qquad n = \frac{\sum_{k=1}^L N_k p_k(1 - p_k)/w_k}{N^2(d^2/z^2) + \sum_{k=1}^L N_k p_k(1 - p_k)}$$

Table A.1

Cumulative Standardized Normal Distribution



A(z) is the integral of the standardized normal distribution from $-\infty$ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of z of particular importance:

=	A(z)	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail

-	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0,6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	. 0.8238	0.8264	0.8289	0.8315	0,8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.862
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.901
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0,917
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.931
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.944
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0,9525	0.9535	0.954
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0,963
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.970
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.976
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0,9808	0.9812	0.981
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.985
2.2	0.9861	0.9864	0.9868	0,9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.991
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.993
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.995
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.996
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.997
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.998
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.998
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.999
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.999
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.999
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.999
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.999
3.5	0.9998	0.9998	0,9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.999
3.6	0.9998	0.9998	0.9999	-,						

TABLE A.2 t Distribution: Critical Values of t

		Significance level					
Degrees of freedom	Two-tailed test: One-tailed test:	10% 5%	5% 2.5%	2% 1%	1% 0.5%	0.2% 0.1%	0.1% 0.05%
1		6.314	12.706	31.821	63.657	318,309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8 9		1,860 1.833	2.306 2.262	2.896	3.355	4.501	5.041
10		1.812	2.228	2.821 2.764	3.250 3.169	4.297 4.144	4.781 4.587
11		1.796			3.106		4.437
12		1,782	2.201 2.179	2. 718 2.681	3.106	4.025 3.930	4.437
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2,120	2.583	2.921	3.686	4.015
17		1,740	2,110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2,093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3,485	3.768
24 25		1.711 1.708	2,064 2,060	2.492	2.797	3.467 3.450	3.745
				2.485	2.787		3.725
26 27		1.706	2.056	2.479	2.779	3.435	3,707
28		1,703 1,701	2.052 2.048	2.473 2.467	2.771 2,763	3.421	3.690
29		1.699	2.045	2.462	2.756	3,408 3,3 9 6	3.674 3.659
30		1,697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38 40		1.686	2.024	2.429	2.712	3.319	3.566
		1.684	2.021	2.423	2.704	3.307	3.551
42 44		1.682	2.018	2.418	2.698	3.296	3.538
46		1.680 1,679	2.015 2.013	2.414 2.410	2.692	3.286	3.526
48		1.677	2.013	2.410	2.687 2.682	3.277 3.269	3.515 3.505
50		1,676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2,390	2.660	3.232	3.460
70		1.667	1,994	2.381	2.648	3.211	3,435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3,160	3,373
150		1.655	1.976	2.351	2.609	3.145	3.357
200 300		1.653 1.650	1.972 1.968	2.345 2.339	2.601	3,131	3.340
400		1.649	1.966	2.339	2.592 2.588	3.118 3.111	3.323 3.315
500 600		1.648 1.647	1.965 1.964	2.334 2.333	2.586 2.584	3.107 3.104	3.310 3.307
∞		1.645	1,960			3.104	
w		1.045	1,900	2.326	2.576	3.090	3.291