Exam 1 is Wed. Feb. 12. **You are allowed 5 sheets of notes and a calculator.** The exam covers survey sampling.

Numbers refer to types of problems on exam.

A *population* is the entire set of (potential) measurements of interest while the *sample* is the subset actually studied.

A *voluntary response sample* consists of people who choose themselves from a general appeal (often from the media).

A *convenience sample* chooses the people easiest to reach (often bring surveys to a location).

Know that the above two sample designs are bad while the design below is good.

A *probability sample* gives each member of the population a known probability of being selected.

1) Given a story problem, be able to recognize whether the sample is a probability sample, a sample of convenience or a voluntary response sample. See HW1 1.

For a *simple random sample* (SRS) of size n from a population of size N. The probability that any member in the population gets in the sample is n/N, and all $\binom{N}{n}$ samples of size n are equally likely.

2) Given a list of n numbers, find the sample mean $\overline{y} = \sum_{i=1}^{n} y_i/n$, and sample variance $S^2 = \dfrac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$. Recall that $\overline{y}$ estimates the population mean μ and S^2 measures the population variance σ^2.

3) Know how to get a SRS using table B or with the *sample function*. See HW1 3,4. When using table B, label objects

```
 a1, ...,   a9, a10, a11, ..., a99, a100, a101, ..., a999, a1000 as
  1,  ...,   9,   0   for 10 or fewer objects
 01, ...,   09,  10,  11, ...,  99,   00  for between 11 and 100 objects
001 ,...,  009, 010, 011, ..., 099,  100,  101, ...,  999,   000
                                          for between 101 and 1000 objects
```

Want to estimate the population mean μ, the population total τ, or the population proportion p (or π).

4) SRS estimator for μ: Let $Y_1, ..., Y_N$ be the population. Let a SRS $y_1, ..., y_n$ of size n be taken. Then $\hat{\mu} = \overline{y}$ and

$$SE(\hat{\mu}) = \sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}.$$

and the 95% CI for μ is $\hat{\mu} \pm 1.96 SE(\hat{\mu})$. See HW1 5.

The quantity $\dfrac{N-n}{N}$ is called the *finite population correction* (fpc) and can be ignored if the *sampling fraction* $f = \dfrac{n}{N} \leq \dfrac{1}{20} = 0.05$.

5) Let $\theta = \mu, \tau$, or p. For an estimator of $\hat{\theta}$ of θ, the *bound on the error of estimation* or the *margin of error* is $1.96SE(\hat{\theta})$.

6) The accuracy of two samples for the mean is measured by $SE(\hat{\mu}_i)$ and the sample with the smaller SE is more accurate.

7) Suppose that you can not compute the two SE's. Suppose that N_1 and N_2 are the two sample sizes. Assume that $\sigma_1^2 \approx \sigma_2^2$, $n_1 < N_1/20$ and $n_1 < N_1/20$. Then use the following rule of thumb: The accuracy of the sample depends on the size of the sample (not the size of the population). Two SRS's of the same size have about the same accuracy while bigger samples have greater accuracy. Some times you will be given the sample size, sometimes a percentage of the population sampled (then you need to figure out which sample is larger). See HW1 2.

The above problem is especially useful for proportions since $\sigma_1^2 \approx \sigma_2^2$ if $0.3 < p_i < 0.7$ for $i = 1, 2$.

If Y has a *hypergeometric distribution*, $Y \sim \mathrm{HG(C, N - C, n)}$, then the data set contains N objects of two types. There are C objects of the first type (that you wish to count) and $N - C$ objects of the second type. Suppose that n objects are selected at random without replacement from the N objects. Then Y counts the number of the n selected objects that were of the first type. The pmf of Y is

$$f(y) = P(Y = y) = \frac{\binom{C}{y}\binom{N-C}{n-y}}{\binom{N}{n}}$$

where the integer y satisfies $\max(0, n - N + C) \leq y \leq \min(n, C)$. The right inequality is true since if n objects are selected, then the number of objects of the first type must be less than or equal to both n and C. The first inequality holds since $n - Y$ counts the number of objects of second type. Hence $n - Y \leq N - C$.

Let $p = C/N$. Then

$$E(Y) = \frac{nC}{N} = np$$

and

$$VAR(Y) = \frac{nC(N-C)}{N^2}\frac{N-n}{N-1} = np(1-p)\frac{N-n}{N-1}.$$

If n is small compared to both C and $N - C$ then $Y \approx \mathrm{BIN(n, p)}$. If n is large but n is small compared to both C and $N - C$ then $Y \approx N(np, np(1-p))$.

Here BIN stands for binomial and $N(\mu, \sigma^2)$ stands for a normal distribution with mean μ and variance σ^2.

8) The SRS size n needed to estimate μ to within an error of estimation B ($\approx$ half the CI width) with 95% confidence is

$$n = \frac{N\sigma^2}{(N-1)\left(\frac{B}{1.96}\right)^2 + \sigma^2}.$$

Round up to the nearest integer. In the above formula, use $\sigma^2 = S^2$ from a previous study or $\sigma^2 = (\text{range}/4)^2$ where the range = max value − min value obtained by theory, common sense or a previous study. See HW2 5.

9) SRS for a proportion p: $y_i = 0$ if the item is not counted and $y_i = 1$ if the item is counted. $\hat{p} = \overline{y} =$ (number of 1's)$/n$.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)}.$$

and the 95% CI for p is $\hat{p} \pm 1.96 SE(\hat{p})$. See HW2 4.

10) The SRS size n needed to estimate p to within an error of estimation B ($\approx$ half the CI width) with 95% confidence is

$$n = \frac{Np^*(1-p^*)}{(N-1)\left(\frac{B}{1.96}\right)^2 + p^*(1-p^*)}$$

where p^* is a given estimate of p. If no estimate is given, use $p^* = 0.5 = 1/2$. Round up to the nearest integer. See HW2 6.

A *stratified random sample* is obtained by separating the population into nonoverlapping groups called *strata*, and then taking a SRS from each *stratum*. Let $L =$ number of strata, $N_i =$ number of sampling units in stratum i, and $N = \sum_{i=1}^{L} N_i$. Let $n_i =$ SRS size in stratum i.

11) Stratified sample estimators for μ. Let $\overline{y}_i$ and S_i^2 be the sample mean and the sample variance from stratum i.

$$\hat{\mu}_{st} = \overline{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i\, \overline{y}_i.$$

$$SE(\hat{\mu}_{st}) = \sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{S_i^2}{n_i}} = \sqrt{\sum_{i=1}^{L}\left[\frac{N_i}{N} SE(\overline{y}_i)\right]^2}.$$

The 95% CI for μ is $\hat{\mu}_{st} \pm 1.96\ SE(\hat{\mu}_{st})$. See HW2 8

12) Stratified sample estimators for p. Let $\hat{p}_i$ be the sample proportion from stratum i.

$$\hat{p}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i\, \hat{p}_i.$$

$$SE(\hat{p}_{st}) = \sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1}} = \sqrt{\sum_{i=1}^{L}\left[\frac{N_i}{N} SE(\hat{p}_i)\right]^2}.$$

The 95% CI for μ is $\hat{p}_{st} \pm 1.96\ SE(\hat{p}_{st})$. See HW2 8.

The stratified sample estimates are obtained using a weighted combination of the formulas for a SRS. Given a table of summary statistics, you should be able to find the stratified sample estimators using 11) and 12) and SRS formulas for the jth stratum using 4) and 9).

A *systematic sample* is obtained by randomly selecting one element from the first k elements in the frame and every kth element thereafter.

If n is the sample size of the systematic sample and if $nk = N$, then every element has equal chance n/N of getting in the sample.

13) Assume that the numbers (possibly 0's or 1's) corresponding to the elements in the sampling frame are randomly ordered. Then the systematic sample estimators for μ are the same as the SRS estimators given by 4) and the systematic sample estimators for p are the same as the SRS estimators given by 9).

Result 13) also holds for systematic sample estimators of μ if the numbers corresponding to the elements in the sampling frame are ordered in magnitude.

Capture–recapture methods are used to estimate animal populations. First, t animals are captured and tagged. Assume that these t animals are a SRS of size t from the population of N animals where N is unknown. These tagged animals are released into the wild, and if p is the proportion of tagged animals now in the population, then $p = \dfrac{t}{N}$. To estimate N, a second sample of n animals are captured at a later date. Let x denote the number of tagged animals in the second sample. then $\hat{p} = \dfrac{x}{n}$,

14) Estimators of N for *direct sampling*. For *direct sampling*, the second sample is a SRS of n animals, and

$$\hat{N} = \frac{t}{\hat{p}} = \frac{nt}{x} \quad \text{and} \quad \mathrm{SE}(\hat{\mathrm{N}}) = \sqrt{\frac{\mathrm{t^2 n(n-x)}}{\mathrm{x^3}}}.$$

The 95% CI for N is $\hat{N} \pm 1.96 SE(\hat{N})$. See HW3 2a.

15) Estimators of N for *inverse sampling*. For inverse sampling, random sampling (capturing) of animals is done until exactly x animals are recaptured. Suppose that the sample size is n. (So x is fixed and n is random.) The formulas are the same as in 14) except

$$SE(\hat{N}) = \sqrt{\frac{t^2 n(n-x)}{x^2(x+1)}}.$$

See HW3 2b.

SRS's, stratified random samples, systematic samples, cluster samples, and ratio estimation with SRS's are too expensive to be used for interviewing from a large scattered population. *Multistage samples* are used instead.

A two stage sample takes a SRS of clusters of units. Then a SRS is taken from each selected cluster.

A three stage sample takes a SRS of big clusters. From each selected big cluster, a SRS of smaller clusters is taken. From each selected smaller cluster, a SRS of units is taken.

The k stage samples are similar. There is a biggest, 2nd biggest, 3rd biggest, ..., $(k-1)$th biggest clusters of units.

The *current population survey* is conducted by the US government every month. Economic indicators such as the unemployment rate are obtained by this survey. Interviewers are sent to about 50,000 households.

One of the most commonly used probability methods for opinion and voting polls is *random digit dialing*. This is a two stage sample. The clusters are area codes. From each selected area code, 7 digit random numbers are generated (possibly with constraints on the first 3 digits) that correspond to phone numbers. These households are contacted and interviewed over the phone.

16) Half sample estimators for τ. Divide the results from the multistage sample in half. Compute $\hat{\tau}_i$ from the 1st and 2nd half. Then

$$\hat{\tau} = \frac{\hat{\tau}_1 + \hat{\tau}_2}{2} \text{ and } \text{SE}(\hat{\tau}) = \frac{|\hat{\tau}_1 - \hat{\tau}_2|}{2}.$$

See HW3 1.

The SE actually used is often much more complex, but the basic idea is to take many half samples, compute the SE formula above, average the results and multiply the average by some constant. The formulas for two stage cluster samples can be computed with the aid of a calculator

Ratio estimators are useful if $y \approx bx$ where $b > 1$. To check this assumption, a scatterplot of x vs. y should be linear (but fan shaped) through the origin.

The population units comes in pairs $(Y_1, X_1), ..., (Y_N, X_N)$ and the data is a SRS of pairs: $(y_1, x_1), ..., (y_n, x_n)$.

17) Ratio estimators for $R = \mu_y/\mu_x = \tau_y/\tau_x$.

$$\hat{R} = r = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{\overline{y}}{\overline{x}}.$$

$$SE(\hat{R}) = \sqrt{\frac{N-n}{nN}\frac{1}{\mu_x^2}\frac{\sum_{i=1}^n (y_i - rx_i)^2}{n-1}}.$$

Use $(\overline{x})^2$ if μ_x^2 is unknown. See HW3 4.

18) Ratio estimators for μ_y.

$$\hat{\mu}_y = r\mu_x.$$

Use

$$\hat{\mu}_y = r\ \overline{x} = \overline{y}$$

if μ_x is unknown.

$$SE(\hat{\mu}_y) = \sqrt{\frac{N-n}{nN}\frac{\sum_{i=1}^n (y_i - rx_i)^2}{n-1}}.$$

See HW3 3.

19) Suppose that $\theta = \mu, p, \tau, N$, or R. If $\hat{\theta}$ and $SE(\hat{\theta})$ are given, then a 95% CI for θ is

$$\hat{\theta} \pm 1.96 SE(\hat{\theta}).$$