1. 5-1, 5-2, 5-3

Large n normal approximations (Central Limit Theorem).

$xBAR \sim N[mu, sigma^2 / n]$
(sketch a normal with mean mu and sd = sigma / root(n)).

$pHAT \sim N[p, pq / n]$
(sketch a normal with mean p and sd = root(pq) / root(n)).

Use FPC = root((N-n) / (N-1)) as appropriate.

ME

xBAR +/- 1.96 s / root(n)

pHAT +/- 1.96 root(pHAT qHAT) / root(n)

2. 6-3, 6-4, 6-6

Large n confidence intervals

xBAR +/- z s / root(n)

pHAT +/- z root(pHAT qHAT) / root(n)

Use FPC = root((N-n) / (N-1)) as appropriate.

Claim: P(mu in z-CI) ~ P(Z in [-z, z]) n large
P(p in z-CI) ~ P(Z in [-z, z]) n large

3. Any n > 1 confidence intervals if population is <u>normal</u>.

xBAR +/- t s / root(n)

Example problem. A population distribution is known to be "in control," that is "approximately normal." A sample of 12 finds xBAR = 16.3 with s = 7.1. Give a 95% confidence interval for population mean mu.

ans. We need the t-score (replaces z-score) for degrees of freedom n – 1 = 12 – 1 = 11). Consult the t-table, inside front cover to the right, for degrees of freedom = n – 1). Here is how it looks,

Critical Values of the t Distribution

Degrees of Freedom	
11	2.201
infinity	1.96
C.I.	95%

So the desired t-score for 11 degrees of freedom and confidence level 0.95 is t = 2.201. The 95% CI is therefore

xBAR +/- t s / root(n)
16.3 +/- 2.201 7.1 / root(12)

Claim: For sample of any size n > 1 from a **normal** population :
P(mu in xBAR +/- t s / root(n)) = 0.95 (not just ~ 0.95).

4. Supplement to the readings: Stratified sampling produces a better estimate for mu (general x scores).

Example question. A population of accounts is divided into two types, those having good credit and those not. We are interested in estimating the population average balance mu of all accounts. However, we will need to audit each account we include in our

sample and this is costly so we need to do things as efficiently as possible on a per-sample basis. Suppose it is known that

22% of accounts do not have good credit

We decide to invest in 100 samples. A statistically trained employee suggests that we draw a random sample of 22 accounts from those that do not have good credit and 78 from those that do. This is done, from which we find

xBAR from 22 = 147.21 xBAR from 78 = 194.49

s from 22 = 75.38 s from 78 = 110.33

What is the stratified estimator of the overall population mean mu based upon this stratified sample and what is a 95% confidence interval for mu based on this stratified estimator?

ans. The stratified estimate of mu is

$$xBAR_{STRAT} = 0.22 \;\; 147.21 + 0.78 \;\; 194.49$$

and the margin of error is

$$+/- \; 1.96 \; \text{root}(0.22^2 \; 75.38^2 \; / \; 22 + 0.78^2 \;\; 110.33^2 \; / \; 78)$$

The sample of 22 would be regarded as rather too small for application of this method based on large n z-approximations.

Defining the stratified estimator.

a. Divide the population into two or more disjoint sub-populations (strata).

b. From each of stratum$_i$ select a with-replacement sample of size n_i. The samples are to be independent between strata also.

c. Form the stratified estimate of the population mean mu

$$xBAR_{strat} = \text{sum, over all strata i, of } (W_i \;\; xBAR_i)$$

where

$W_i = N_i / N$ is the relative size of stratum i in the population

$xBAR_i$ is the mean of stratum i sample scores.

Using E and Var rules to evaluate the expectation and variance of the stratified estimator.

$$E\ xBAR_{STRAT} = \text{sum of } W_i\ \ E\ xBAR_i \quad \text{(linearity of E)}$$
$$= \text{sum of } W_i\ mu_i \quad \text{(each } xBAR_i \text{ is unbiased)}$$
$$= mu \quad \text{(mu is the wtd avg of subpopulation means)}$$

That is, the stratified estimator $xBAR_{strat}$ is **unbiased** as an estimator of the population mean mu.

$$Var\ xBAR_{STRAT} = \text{sum of } Var\ (W_i\ \ xBAR_i) \quad (xBAR_i \text{ are indep})$$
$$= \text{sum of } (W_i^2\ \ Var\ xBAR_i)$$
$$= \text{sum of } (W_i^{\ 2}\ \ sigma_i^2 / n_i)$$

ME

This leads to the following definition of ME for the stratified estimator of mu

ME for $xBAR_{STRAT}$ = +/- 1.96 root(sum of $(W_i^{\ 2}\ \ s_i^2 / n_i)$)

where s_i denotes the sample standard deviation obtained from the sample of stratum i.

Important fact: In the special case $n_i = W_i\ n$ (called proportionally stratified sampling and used in the example problem above) the theoretical sd of $xBAR_{STRAT}$ (obtained just above) is never larger than the sd of xBAR.

In practice it is usually the case that the estimator xBAR (without stratification) and its competitor $xBAR_{STRAT}$ (with stratification) are each approximately normally distributed. This makes it easy to compare their performances. Since they have the same expectation mu (each is unbiased for the population mean mu), and the proportionally stratified estimator $xBAR_{STRAT}$ has the smaller sd, it is the better estimator. To see it just think of comparing two normal curves, both having the same mean mu, but one having a smaller sd.

5. Supplement to the readings: Outline of a ME applicable to stratified sampling for p (0-1 scores).

Example question. A population of accounts is divided into two types, those having good credit and those not. We are interested in estimating the population fraction p of all accounts that will respond to an offer we plan to make them. However, it is costly to make this offer properly so we need to do things as efficiently as possible on a per-sample basis. Suppose it is known that

30% of accounts are held by women

We decide to invest in 50 samples. A statistically trained employee suggests that we draw a random sample of 50 0.3 = 15 accounts from those held by women and 35 from those held by men. This is done, from which we find

pHAT from 15 = 6/15 pHAT from 35 = 20/35

What is the stratified estimator of the overall population proportion p based upon this stratified sample and what is a 95% confidence interval for p based on this stratified estimator?

ans. The stratified estimate of p is

$$pHAT_{STRAT} = 0.3\ 6/15 + 0.7\ 20/35$$

and the margin of error is

+/- 1.96 root(0.3^2 (6/15 9/15) / 15 + 0.7^2 (20/35 15/35) / 35)

The sample size of 15 would be regarded as rather too small and insufficient to support the use of this large n method.

Defining the stratified estimator.

a. Divide the population into two or more disjoint sub-populations (strata).

b. From each of stratum$_i$ select a with-replacement sample of size n_i. The samples are to be independent between strata also.

c. Form the stratified estimate of the population proportion p having some particular characteristic of interest

$pHAT_{strat}$ = sum, over all strata i, of (W_i $pHAT_i$)

where

$W_i = N_i / N$ is the relative size of stratum i in the population

$pHAT_i$ is the stratum$_i$ proportion for sample scores.

Then

E $pHAT_{strat}$ = sum of W_i E $pHAT_i$ (linearity of E)

= sum of W_i p_i (each $pHAT_i$ is unbiased for p_i)

= p (p is the wtd avg of subpopulation proportions)

That is, the stratified estimator $pHAT_{strat}$ is also **unbiased** as an estimator of the population proportion p.

Var $pHAT_{strat}$ = sum of Var (W_i $pHAT_i$) ($pHAT_i$ are indep)

= sum of (W_i^2 Var $pHAT_i$)

= sum of (W_i^2 p_i q_i / n_i)

ME

This leads to the following definition of ME for the stratified estimator of mu

ME for $pHAT_{STRAT}$

= +/- 1.96 root(sum of (W_i^2 $pHAT_i$ $qHAT_i$ / n_i))

where $pHAT_i$ $qHAT_i$ is the estimated standard deviation of i-th stratum (0-1 scores), obtained from the sample of stratum i.

Important fact: In the special case $n_i = W_i$ n (called proportionally stratified sampling and used in the example problem above) the theoretical sd of $pHAT_{STRAT}$ (obtained just above) is never larger than the sd of pHAT.

In practice it is usually the case that the estimator pHAT (without stratification) and its competitor $pHAT_{STRAT}$ (with stratification)

are each approximately normally distributed. This makes it easy to compare their performances. Since they have the same expectation p, and the proportionally stratified estimator $pHAT_{strat}$ has the smaller sd, we prefer to use the stratified estimator. To see it just think of comparing two normal curves, both having the same mean mu, but one having a smaller sd.