Figure 2.12. UNSTRATIFIED POPULAIONS One-Stage EPSWOR of Individual Elements Estimating the Difference of Two Proportions

Two previous figures in these STAT 332 Course Materials have presented the theory of survey sampling involving equiprobable selecting (EPS) from a respondent population of $\mathbf{N}$ elements (or units).

- Figure 2.3 developed the theory for estimating $\overline{\mathbf{Y}}$ (the respondent population *average*) and (the closely related theory for estimating) $_T\mathbf{Y}$ (a *total*) under EPS for a quantitative response variate, $\mathbf{Y}$, of the elements (or units) which make up the population;
- Figure 2.10 adapted the theory of Figure 2.3 to estimate $\mathbf{P}$ (the respondent population *proportion*) and (the closely related theory for estimating) $\mathbf{NP}$ (a *number* or *frequency*) for a *qualitative* response variate in *two* categories [which we denoted C and ^cC (the complement of C)] under EPS of the elements (or units) which make up the population.

In this Figure 2.12, we extend the ideas of Figure 2.10 to a qualitative response in *more than two* categories, C_1, C_2, C_3, ..., say; our particular concern is with estimating the *difference of two proportions*, denoted $\mathbf{P}_1 - \mathbf{P}_2$, from among two or more such proportions. We distinguish two cases:

* the two proportions of interest can reasonably be modelled as being *probabilistically independent*;
* the two proportions of interest can*not* reasonably be modelled as being probabilistically independent.

Using equation (2.12.1) (*e.g.*, from STAT 230) at the right for the standard deviation of the difference of two random variables, P_1 and P_2

$$s.d.(P_1 - P_2) = \sqrt{[s.d.(P_1)]^2 + [s.d.(P_2)]^2 - 2cov(P_1, P_2)} \qquad \text{-----(2.12.1)}$$

representing the two (sample) proportions, allows us to estimate the difference of interest; if P_1 and P_2 are probabilistically independent, the covariance term on the RHS of equation (2.12.1) is zero.

1. Case 1: Estimating $\mathbf{P}_1 - \mathbf{P}_2$ when $\mathbf{P}_1$ and $\mathbf{P}_2$ Involve *Different* Populations

An instance of Case 1 is where $\mathbf{P}_1$ is the proportion of adult females who are regular cigarette smokers in Canada, and $\mathbf{P}_2$ is the corresponding proportion of adult males; a sample would be selected from *each* group. Due to factors like possible similarities in behaviour within couples, the values of $\mathbf{P}_1$ and $\mathbf{P}_2$ are likely *not* independent but, for the two samples selected *separately* (by EPS), it is reasonable to model as probabilistically independent the random variables P_1 and P_2 representing the sample proportions p_1 and p_2. In this Figure 2.12, we think of two such (respondent) populations as Population 1 (of size $\mathbf{N}_1$ elements) and Population 2 (of size $\mathbf{N}_2$ elements); the corresponding samples are Sample 1 (of size n_1) and Sample 2 (of size n_2). To obtain an expression for a confidence interval for $\mathbf{P}_1 - \mathbf{P}_2$, we reason as follows from the theory presented in Figure 2.10:

- The random variable P_1 is an unbiased estimator of $\mathbf{P}_1$, the proportion of interest in Population 1 – the corresponding *estimate* is Sample 1 proportion p_1; we denote $1-p_1$ by q_1.
- The random variable P_2 is an unbiased estimator of $\mathbf{P}_2$, the proportion of interest in Population 2 – the corresponding *estimate* is Sample 2 proportion p_2; we denote $1-p_2$ by q_2.
- The random variable $P_1 - P_2$ is an unbiased estimator of $\mathbf{P}_1 - \mathbf{P}_2$ (Why?) – the corresponding *estimate* is $p_1 - p_2$.
- The *estimated* standard deviations of P_1 and P_2 are as shown in equations (2.12.2) and (2.12.3) at the right.
- Using the result (2.12.1), the *estimated* standard deviation of $P_1 - P_2$ is as shown in equation (2.12.4); if $\mathbf{N}_1 \gg n_1$ and $\mathbf{N}_2 \gg n_2$, equation (2.12.5) can be used as an approximation for equation (2.12.4).
- Finally, a $100(1-\alpha)\%$ confidence interval for $\mathbf{P}_1 - \mathbf{P}_2$ is as shown in equation (2.12.6).

$$\hat{s.d.}(P_1) = \sqrt{\frac{n_1 p_1 q_1}{n_1 - 1}\left(\frac{1}{n_1} - \frac{1}{\mathbf{N}_1}\right)} \qquad \text{-----(2.12.2)}$$

$$\hat{s.d.}(P_2) = \sqrt{\frac{n_2 p_2 q_2}{n_2 - 1}\left(\frac{1}{n_2} - \frac{1}{\mathbf{N}_2}\right)} \qquad \text{-----(2.12.3)}$$

$$\hat{s.d.}(P_1 - P_2) = \sqrt{\frac{n_1 p_1 q_1}{n_1 - 1}\left(\frac{1}{n_1} - \frac{1}{\mathbf{N}_1}\right) + \frac{n_2 p_2 q_2}{n_2 - 1}\left(\frac{1}{n_2} - \frac{1}{\mathbf{N}_2}\right)} \qquad \text{-----(2.12.4)}$$

$$\hat{s.d.}(P_1 - P_2) \simeq \sqrt{\frac{p_1 q_1}{n_1 - 1} + \frac{p_2 q_2}{n_2 - 1}} \qquad \text{-----(2.12.5)}$$

$$p_1 - p_2 \pm z_\alpha^* \times \hat{s.d.}(P_1 - P_2) \qquad \text{-----(2.12.6)}$$

2. Case 2: Estimating $\mathbf{P}_1 - \mathbf{P}_2$ when $\mathbf{P}_1$ and $\mathbf{P}_2$ Involve the *Same* Population

A situation in which it is *not* reasonable to model as probabilistically independent the random variables P_1 and P_2 representing the sample proportions p_1 and p_2 is when the respondent population attributes $\mathbf{P}_1$ and $\mathbf{P}_2$ are defined over the elements of the *same* population; for instance, $\mathbf{P}_1$ might be the proportion of Canadian voters who are Liberal supporters in a federal election, and $\mathbf{P}_2$ the corresponding proportion of PC supporters; *other* categories of voters include NDP supporters, Reform supporters, BQ supporters and undecided voters. In a situation like this, the proportions across all categories of voters are constrained to add to 1 (or 100%), so a change in one proportion results in a change in at least one of the others. In the sample, we model this constraint as probabilistic *de*pendence of P_1 and P_2 We think of the (respondent) population in this Case 2 as having $\mathbf{N}$ elements and the sample obtained by EPS from this population as having n units (or elements).

To obtain an expression for a confidence interval for $\mathbf{P}_1 - \mathbf{P}_2$, we extend the reasoning given above for the case of *in*dependent proportions. We recall from Note 12 on page 2.84 of Figure 2.10 that, to estimate $\mathbf{P}$ (the respondent proportion of elements in one of only *two* categories), we can use a *bi*nomial probability model; when there are *more than two* categories, the corresponding model is *multi*nomial. Because the marginal distributions of the multinomial distribution are *bi*nomial, we deal

(*continued overleaf*)

with the two terms on the right-hand side of (2.12.1) involving the standard deviation of P_1 and of P_2 in the *same* way as in the *in*dependent case above; what *differs* is that we must now also include the *covariance* term.

The covariance of the random variables Y_i and Y_j of a multinomial distribution is $-n\pi_i\pi_j$, where π_i and π_j are the respective probabilities of outcomes *i* and *j* among the k possible outcomes of each repetition of the process being modelled by a multinomial distribution. Hence, we can reason intuitively that, in our *finite* population case, the *estimated* covariance of P_1 and P_2 is $-[np_1p_2/(n-1)](1/n - 1/\mathbf{N})$. Thus, the final expressions we need so we can deal with the *de*pendent case are:

– using equation (2.12.1) given overleaf on page 2.91, the *estimated* standard deviation of $P_1 - P_2$ is as shown in equation (2.12.7) at the right; if $\mathbf{N} \gg n$, equation (2.12.8) can be used as an approximation for equation (2.12.7).

$$\hat{s.d.}(P_1 - P_2) = \sqrt{\frac{n}{n-1}(p_1q_1 + p_2q_2 + 2p_1p_2)(\frac{1}{n} - \frac{1}{\mathbf{N}})} \qquad \text{-----(2.12.7)}$$

$$\hat{s.d.}(P_1 - P_2) \simeq \sqrt{\frac{1}{n-1}(p_1q_1 + p_2q_2 + 2p_1p_2)} \qquad \text{-----(2.12.8)}$$

– a $100(1-\alpha)\%$ confidence interval for $\mathbf{P}_1 - \mathbf{P}_2$ is as shown in equation (2.12.9).

$$p_1 - p_2 \pm z^*_\alpha \times \hat{s.d.}(P_1 - P_2) \qquad \text{-----(2.12.9)}$$

NOTES: 1. Comparing the expressionsxi (2.12.4) overleaf on page 2.91 and (2.12.7) above for $\hat{s.d.}(P_1 - P_2)$, we see that for *given* values of p_1 and p_2, the effect of *lack* of independence is to *broaden* the confidence interval; *i.e.*, *de*pendence tends to *increase* (sampling) imprecision, thus imposing *greater* limitation on answers from this category of error.

2. The results given above for the *de*pendent case can be derived by a method involving *indicator variables* – see Question **A3 – 6** of Assignment 3.

- The three-term sum in parentheses under the square root in equations (2.12.7) and (2.12.8) above for the estimated standard deviation of $P_1 - P_2$ can also be written as: $(p_1 + p_2) - (p_1 - p_2)^2$.

Example 2.12.1: In 1983, the *Kitchener-Waterloo Record* reported the results of two pre-federal election polls taken about a month apart. On September 8, *The Record* reported that a Gallup poll in August had found 50% of decided voters said they would vote for the Progressive Conservative party from among 1,055 respondents; 26% of the respondents said they were undecided. Another Gallup poll in September, reported by *The Record* on October 13, found 62% PC supporters among decided voters; there were 1,039 respondents to this poll, of whom 34% said they were undecided. [The relevant articles EM8302 and EM8303 are reprinted in Appendices 1 and 2 on page 2.96.] We *assume* that people were obtained for both polls by probability selecting from the population of eligible Canadian voters.

(a) Was there a *real* increase in PC support between the two polls, or could the reported difference reasonably be due to chance?

(b) What *assumptions* underlie the calculations needed to answer (a)?

(c) Explain briefly the meaning of the phrase in (a) ..... *could the reported difference reasonably be due to chance?*

Solution 1: (a) The two respondent populations attributes of interest in this question are the proportions of PC supporters, $\mathbf{P}_A$ and $\mathbf{P}_S$, in August and September, 1983. It is reasonable to assume that the two polls were csrried out in a way that allows us to model the two sample proportions by *probabilistically independent* random varisbles, P_A and P_S, as in Case 1 overleaf on page 2.91.

For *decided* voters, estimated to be 74% in August and 66% in September, the respective sample sizes are $n_A = 1{,}055 \times 0.74 = 781$ and $n_S = 1{,}039 \times 0.66 = 686$; also, the *data* are $p_A = 0.50$ and $p_S = 0.62$.

The newspaper articles do not give the population sizes but we know they are *much* larger than the sample sizes; we therefore approximate $(1/n - 1/\mathbf{N})$ by $1/n$ in the estimated s.d. expressions.

Taking $p_A = p_A = 0.50$ and $p_S = p_S = 0.62$ and using equation (2.12.5), we have:

$$p_S - p_A = 0.12, \qquad \hat{s.d.}(P_S - P_A) \simeq \sqrt{\frac{0.62 \times 0.38}{685} + \frac{0.50 \times 0.50}{780}} = 0.025\,777\,013.$$

With both sample sizes of around 700, we hope we are justified in assuming reasonable normality for $P_S - P_A$ as a consequence of the Central Limit Theorem; also, because we are dealing with *counted* data, we use $z^*_\alpha = 1.95996$ to find an approximate 95% confidence interval as:

$$0.12 \pm 1.95996 \times 0.025\,777\,013 \;=\; 0.12 \pm 0.0505 \;\Rightarrow\; (0.0695,\ 0.1705) \text{ or about } (7,\ 17)\%.$$

Because zero lies comfortably *out*side the approximate 95% CI, the answer is that these poll data provide eveidence of a real increase in PC support from August to September of 1983.

Solution 1: (b) The accuracy of the confidence interval given above depends on five categories of error being adequately managed in the Plan and execution of the two polls.

- Study error: we assume that the frame used to specify the study population had adequate overlap with the target population of elegible Canadian voters.

(*continued*)

Figure 2.12. UNSTRATIFIED POPULAIONS: One-Stage EPSWOR of Individual Elements Estimating the Difference of Two Proportions (continued 1)

Solution 1: (cont.) (b)

- Non-response error: we assume that clear quetions were asked by properly trained interviewers so as to provide adequate incentives for response and for truthful responses (and so hopefully to make adequately close the values of the relevant attributes of the respondent and study populations).
- Sample error: we assume that the sample was selected by a probability selecting method so as to provide the probabilistic basis of the theory which underlies the confidence interval calculations.
- Measurement error: we assume that clear questions and truthful responses, together with correct data processing, managed measurement error adequately.
- Model error: we assume that probabilistic independence is an adequate model for the 'independence' of voters in different polls taken a month apart.

(c) The phrase asks how likely is it that sample error could be as large as the observed difference (12 percentage points in this instance). [Zero lying so far *out*side the 95% confidence interval in this case shows that a sample error of 12 percentage points is *very* unlikely (although, of course, *not* impossible).]

Solution 2: (a) Alternatively, instead of using a confidence interval for the difference of the two relevant proportions, we can answers the question more formally using a *test of (statistical) significance.*

We see from the statement of the question that we are dealing with a situation involving *two* study populations – eligible Canadian voters in August and in September of 1983.

Because we are dealing with *counted* data, we will use a *z*-test for the difference of two *proportions.*

Model: The random variable P_S is an unbiased estimator of $\mathbf{P}_S$, the proportion of PC supporters in the respondent population; the corresponding *estimate* is p_S which we take as the data $p_S = 0.62$ from the 686 *decided* respondents in the September poll.

The random variable P_A is an unbiased estimator of $\mathbf{P}_A$, the proportion of PC supporters in the respondent population; the corresponding *estimate* is p_A which we take as the data $p_A = 0.50$ from the 781 *decided* respondents in the August poll.

A model for the random variable $P_S - P_A$, an unbiased estimator of $\mathbf{P}_S - \mathbf{P}_A$ which is the difference of interest here, is given in equation (2.12.10) above, where $\mathbf{Q} = 1 - \mathbf{P}$.

$$P_S - P_A \;\dot{\sim}\; N\left[\mathbf{P}_S - \mathbf{P}_A, \sqrt{\frac{\mathbf{N}_S}{\mathbf{N}_S - 1}\mathbf{P}_S\mathbf{Q}_S\left(\frac{1}{n_S} - \frac{1}{\mathbf{N}_S}\right) + \frac{\mathbf{N}_A}{\mathbf{N}_A - 1}\mathbf{P}_A\mathbf{Q}_A\left(\frac{1}{n_A} - \frac{1}{\mathbf{N}_A}\right)}\right] \qquad \text{-----(2.12.10)}$$

We assume that the method of selecting the samples containing the 781 and 686 decided respondents was such as to make the limitation imposed on the answer by sample error acceptable in the investigation context.

H_o: $\mathbf{P}_S = \mathbf{P}_A$; *i.e.*, the proportions of decided PC supporters is the *same* in the two respondent populations.

H_a: $\mathbf{P}_S > \mathbf{P}_A$; *i.e.*, the proportion of decided PC supporters is *higher* in the September respondent population. [We use a *one*-sided test because the statement of the question is concerned with an *in*crease in PC support in September.]

Dis. meas.: By standardizing in the model for $P_S - P_A$ in equation (2.12.10), we obtain the discrepancy measure (2.12.11) at the right, where the expression in the denominator is now the approximate *estimated* standard deviation of $P_S - P_A$, based on equation (2.12.5) when we are *not* given the sizes, $\mathbf{N}_S$ and $\mathbf{N}_A$, of the respective respondent populations but we know that $\mathbf{N}_S \gg n_S$ and $\mathbf{N}_A \gg n_A$.

$$\frac{(P_S - P_A) - (\mathbf{P}_S - \mathbf{P}_A)}{\sqrt{\dfrac{p_S q_S}{n_S - 1} + \dfrac{p_A q_A}{n_A - 1}}} \;\dot{\sim}\; N(0, 1) \qquad \text{-----(2.12.11)}$$

Observed: $p_S - p_A = 0.62 - 0.50 = 0.12$ $(= p_S - p_A)$;
expected: $\mathbf{P}_S - \mathbf{P}_A = 0$ (under the null hypothesis).

Value of discrepancy measure: $\frac{0.12 - 0}{0.025\,777} = 4.655\,310.$

P-value: $\Pr[N(0, 1) \geq 4.655\,310] = 1.617\,465 \times 10^{-6} \simeq 0.000\,002.$

Answer: Using a one-sided *z*-test for the difference of two proportions, the null hypothesis is rejected at the 1% (and also at the 0.001%) level (P $\simeq$ 0.000 002); thus, the data provide very highly statistically significant evidence of a real increase in PC support among decided voters in September compared with August of 1983.

(*continued overleaf*)

NOTES: 3. When testing for the *equality* of two proportions [involving either equation (2.12.4) or (2.12.5)], it is customary, in the denominator of the discrepancy measure, to replace the *individual* sample proportions, p_1 and p_2, by the *combined* proportion (p_c) calculated as shown in equation (2.12.12) at the right above. However, this refinement has no practically important effect on the answer from the test in Solution 2 above of Example 2.12.1.

$$p_c = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad \text{-----(2.12.12)}$$

4. The statement of the question in Example 2.12.1 gives no indication that there was reason to suspect an increase in PC support *prior to* obtaining the September poll results. The *one*-sided test we have performed may therefore be based on *data snooping* and so not be justified; instead, a *two-sided* test would avoid this possible *mis*-use of significance testing.
   - In the situation of Example 2.12.1, a *two*-sided test would similarly provide very highly statistically significant evidence (P $\simeq$ 0.000 003) of a *difference* in the level of PC support among decided voters in the August and September respondent populations.

Example 2.12.2: In the first of the two polls referred to in Example 2.12.1, the standings of the three main political parties in that election, among the 74% of decided respondents polled, were reported as follows: 50% PC supporters, 28% Liberal supporters and 20% NDP supporters. The number of respondents was 1,055 in the poll – see the article EM8302 reprinted on the upper half of page 2.96.

(a) Do these data provide evidence of a *real* lead in Liberal support over that of the NDP, or could the reported difference reasonably be due to chance?

(b) What *assumptions* underlie the calculations needed to answer (a)

Solution 1: (a) The situation in this question differs from that in Example 2.12.1 in that we are now dealing with the difference of two *de*pendent proportions – the proportions of Liberal and NDP supporters in the *same* population of voters. We therefore use the theory for Case 2 developed in this Figure 2.12 on page 2.91.

We work with *decided* voters – here, 74% of the sample of size 1,055 – so we have:
$n = 1{,}055 \times 0.74 \simeq 781$; also, the *data* are $p_L = 0.28$ and $p_N = 0.20$.

N, the size of the respondent population, is not given in the statement of the question but we know that that **N** $\gg$ n and so we will use the *approximate* result of equation (2.12.8) near the top of page 2.92 when calculating $\hat{s.d.}(P_L - P_N)$; taking $p_L = p_L = 0.28$ and $p_N = p_N = 0.20$, we have:

$$p_L - p_N = 0.28 - 0.20 = 0.08,$$

$$\hat{s.d.}(P_L - P_N) \simeq \sqrt{\frac{1}{n-1}(p_L q_L + p_N q_N + 2p_L p_N)} \simeq \sqrt{\frac{1}{780}(0.28\times0.72+0.20\times0.80+2\times0.28\times0.20)} = 0.024\,641.$$

With a sample size as large as 781, we hope we are justified in assuming reasonable normality for $P_L - P_N$ as a consequence of the Central Limit Theorem; also, we estimated the respondent population proportions ($\mathbf{P}_L$ and $\mathbf{P}_N$) by the sample proportions (p_L and p_N) and, because we are dealing with *counted* data, we will use $z^*_\alpha = 1.95996$ to find approximate 95% confidence intervals for $\mathbf{P}_L - \mathbf{P}_N$ (the difference in the support for the Liberals and the NDP among decided voters) as:

$$0.08 \pm 1.95996 \times 0.024\,641 \;=\; 0.08 \pm 0.0483 \;\Rightarrow\; (0.0317,\ 0.1283) \text{ or about } (0.03_2,\ 0.12_8).$$

Because zero lies comfortably *out*side this approximate 95% confidence interval, the answer is that these data *do* provide evidence of a real lead in Liberal support over that of the NDP; *i.e.*, the reported difference can*not* reasonably be considered to be due merely to sample error.

Solution 1: (b) Error management to try to attain accuracy for the confidence interval in (a) is generally as discussed on pages 2.92 and 2.93 in Solution 1 of Example 2.12.1 but with two additions.

- An answer about a *difference* in population attributes (like that between two proportions of party support in a poll) may have less severe limitations than an answer involving estimating either attribute individually, because some components of some error categories – like an inaccurate measuring process – may tend to compensate for each other – see also Note 6 on the facing page 2.95.
  - This is similar to the idea underlying a clinical trial like the Physicians Health Study, where the effect of taking aspirin on the risk of a heart attack used treatment and control groups each of around 11,000 male doctors as participants. It is likely that doctors in particular have a *different* risk of heart attack from that of males in general but it was hoped that the *difference* in heart attack risk for male *doctors* under aspirin and placebo would be adequately close to this *difference* for males in general.
- Model error: we assume that the model for *lack* of probabilistic independence, which involves including covariance, adequately describes the constraints on propoortions of party support among voters in any poll.

(*continued*)

Figure 2.12. UNSTRATIFIED POPULAIONS: One-Stage EPSWOR of Individual Elements Estimating the Difference of Two Proportions (continued 2)

Solution 2: (a) Alternatively, instead of using a confidence interval for the difference of the two relevant proportions, we can address the question more formally using a *test of (statistical) significance.*

This question involves two *de*pendent proportions in *one* study population – eligible Canadian voters in August of 1983; the 'dependence' is the constraint that the set of proportions over all categories of voters must sum to one (or 100%) and so any change in value must involve at least two of the proportions.

The question involves *counted* data so we use a z-test for the difference of two dependent *proportions.*

Model: Let the random variable P_L represent p_L, the proportion of Liberal supporters in the sample of size $n = 781$ *decided* respondents obtained for the August poll; it is an unbiased estimator of $\mathbf{P}_L$, the proportion of decided Liberal supporters in the *respondent* population.

Similarly, let the random variable P_N represent p_N, the proportion of NDP supporters in the sample of size $n = 781$ *decided* respondents obtained for the same August poll; it is an unbiased estimator of $\mathbf{P}_N$, the proportion of decided NDP supporters in the *respondent* population.

A model for the random variable $P_L - P_N$, representing the *difference* $p_L - p_N$

$$P_L - P_N \div N[\mathbf{P}_L - \mathbf{P}_N, \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}(\mathbf{P}_L\mathbf{Q}_L + \mathbf{P}_N\mathbf{Q}_N + 2\mathbf{P}_L\mathbf{P}_N)(\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})}] \quad \text{-----(2.12.13)}$$

(and which is an unbiased estimator of $\mathbf{P}_L - \mathbf{P}_N$, the difference of interest here), is equation (2.12.13) above.

We assume that the method of selecting the sample containing the 781 decided respondents was such as to make the limitation imposed on the answer by sample error acceptable in the investigation context.

H_o: $\mathbf{P}_L = \mathbf{P}_N$; *i.e.*, the proportion of decided Liberal supporters is the *same* in that of decided NDP supporters.

H_a: $\mathbf{P}_L > \mathbf{P}_N$; *i.e.*, the proportion of decided Liberal supporters is *higher* than that of decided NDP supporters.

[We use a *one*-sided test because the statement of the question is concerned with *higher* support for the Liberals than for the NDP.]

Dis. meas.: By standardizing in the model (2.12.13) above for $P_L - P_N$, we obtain the discrepancy measure (2.12.14) at the right, where the expression in the denominator is now

$$\frac{(P_L - P_N) - (\mathbf{P}_L - \mathbf{P}_N)}{\sqrt{\frac{1}{n-1}(p_L q_L + p_N q_N + 2p_L p_N)}} \div N(0, 1) \quad \text{-----(2.12.14)}$$

the *estimated* standard deviation $P_L - P_N$, based on equation (2.12.8) on page 2.92 when we are *not* given the sizes, $\mathbf{N}_L$ and $\mathbf{N}_N$, of the respective respondent populations but we know that $\mathbf{N}_L \gg n_L$ and $\mathbf{N}_N \gg n_N$.

Observed: $p_L - p_N = 0.28 - 0.20 = 0.08$ ($= p_L - p_N$);

expected: $\mathbf{P}_L - \mathbf{P}_N = 0$ (under the null hypothesis).

Value of discrepancy measure: $\frac{0.08 - 0}{0.024\,641} = 3.246\,620.$

P-value: $\Pr[N(0, 1) \geq 3.246\,620] = 5.839\,212 \times 10^{-4} \simeq 0.0006.$

Answer: Using a one-sided z-test for the difference of two dependent proportions, the null hypothesis is rejected at the 1% (and also at the 0.1%) level ($P \simeq 0.0006$); thus, the data provide highly statistically significant evidence of a real lead in Liberal support over that of the NDP in August of 1983 – the reported difference (of 8 percentage points) can*not* reasonably be considered to be due merely to sample error.

NOTES: 5. As in Example 2.12.1, the lack of evidence in the statement of Example 2.12.2 that the Liberals were thought to have higher support than the NDP *prior to* the availability of the August poll results raises concern about *data snooping*; performing a *two*-sided test, rather than the *one*-sided test give above, would alleviate this concern but in this situation, as in Example 2.12.1, make no practically important difference to the outcome from the test.

6. When estimating the *difference* of two population attribute values, like $\mathbf{P}_1 - \mathbf{P}_2$ in this Figure 2.12, the consequences for the Answer of inadequate error management may fortuituously (occasionally?) be somewhat mitigated. For example:

- it is possible that the *difference* in the attribute values of two similarly inadequate study populations is closer to the *difference* in the corresponding target population attribute values than are these corresponding attribute values to each other;
- it is possible that the *difference* in two measured values from the same inadequate measuring system is closer to the true *difference* than are the two measured values to their true values.

However, data-based investigating is an area not noted for bestowing unmerited benefits on the careless or the ignorant.

 (*continued overleaf*)

3. Appendix 1 – Newspaper Article for Examples 2.12.1 and 2.12.2

The article EM8302 reprinted below is a source of the information used in Example 2.12.1 on pages 2.92 to 2.94 and the source for Example 2.12.2 on pages 2.94 and 2.95.

EM8302: The Kitchener-Waterloo Record, September 8, 1983

NDP gains while Tories slip in poll

TORONTO (CP) – Support for the federal Progressive Conservatives is slipping among committed voters while the New Democratic Party is gaining favour, a Gallup poll released today suggests.

The August poll, conducted eight weeks after Brian Mulroney was chosen the new Conservative leader, said 50 per cent of decided voters would support the Conservatives in an election, a drop of five percentage points from a July poll.

At the same time, the poll suggests, the NDP increased its share to 20 per cent from 16 per cent a month earlier, while the Liberals garnered 28 per cent, up one percentage point.

Twenty-six percent said they were undecided compared with 23 per cent in July.

"Nothing much has changed", Conservative party president Peter Elzinga said of the results. "We're still well ahead, the Liberals are still in serious trouble and the NDP has just climbed up from rock bottom."

NDP leader Ed Broadbent said the results "are a very good sign for the New Democratic Party right across Canada and it will mean that our own members will be going into Parliamant more determined than ever."

Parliament reopens Monday.

The poll is based on interviews with 1,055 people and is accurate to within four percentage points 19 times out of 20, says Gallup.

This means Conservative support could range fron 46 to 54 per cent of decided voters, the NDP's between 16 and 24 per cent and the Liberals' between 24 and 32 per cent.

The poll provides no precise regional breakdown because the smaller sample size for each region increases the margin of error. But it suggests the NDP's gains came at Conservative expense in the Atlantic provinces, Ontario and the West.

Reached Wednesday night in Liverpool, N.S., where he is attending a federal caucus meeting, Broadbent said Canadians are listening to the party's message on medicare and unemployment "and that's encouraging."

Federal Labour Miniser Andre Ouellet, remarking on the apparent dip in Conservative support, said "The situation seems to be coming back to normal and the gap between the Conservatives and the NDP should shrink in the coming months.

Ouellet, attending a Quebec Liberal caucus meeting at a Quebec resort, said he believes the Liberal popularity will increase as anti-inflation measures and job creation measures take effect.

"We're still relativey low but I'm not surprised by the situation."

4. Appendix 2 – Newspaper Article for Example 2.12.1

The article EM8303 reprinted below is a second source of information used in Example 2.12.1 on pages 2.92 to 2.94.

EM8303: The Kitchener-Waterloo Record, October 13, 1983

Poll gives PCs 62%; Liberals only 23%

TORONTO (CP) – The federal Progressive Conservative party has jumped to a record 62 per cent in popular support while the Liberals have plunged to an all-time low, a Gallup poll released today indicates.

The poll said the Conservatives, under new leader Brian Mulroney, had risen 12 percentage points from the 50-per-cent support they received in a similar poll last month. The Liberals were the choice of 23 per cent of decided voters polled, while the NDP had support from 14 per cent.

With the latest result, the Conservatives topped the previous record showing of 60 per cent in the 10-year-old history of the monthly Gallup poll. That mark was set in July, 1958, in the aftermath of John Diefenbaker's sweeping election victory that March.

The latest poll, taken only days after Mulroney won an Aug. 29 byelection in the Nova Scotia riding of Central Nova and before he entered the Commons, was based on in-home interviews with 1,039 eligible voters.

Such polls are generally considered accurate plus or minus four percentage points 19 out of 20 times, meaning Conservative support could be as high as 66 per cent or as low as 58 per cent and Liberal support as high as 27 per cent or as low as 19 per cent.

The results would be affected further with a shift either way in the number of undecided voters polled.

Gallup says the number of undecided respondents increased by eight percentage points to 34 per cent.

"The Gallup shows a massive repudiation of the Liberal party," Mulroney said in an interview.

"It is a historic situation. It shows growing confidence in the Conservatives in all areas of the country and in all linguistic groups and it is an urgent appeal for a new government."

So massive was the surge, the Conservatives tied the Liberals among those polled in Quebec, where they won only 12 per cent of the popular vote in the 1980 federal election.

The Conservatives held a 2-to-1 margin over the Liberals in Ontario and were supported by two of every three of those polled in the West.

But Senator Keith Davey, a senior Trudeau advisor, said he does not believe the Gallup results.

"It is not what our poll is showing," he said, adding the Conservatives "have nowhere to go but down."