University of Waterloo — STAT 332 – W. H. Cherry

Tutorial 4 Outline Solution

T4 – 1. (a) For the hardness measurements on the sample of n = 10 castings, we find that, in B.I.D. units:

the sample average is: $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = 4.4 \;(= \bar{y})$,

the square of the sample standard deviation is: $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 = 0.01\dot{3} \simeq 0.115\,470\,053^2 \;(= s^2)$;

the estimated standard deviation of the sample average is: $\hat{s.d.}(\bar{Y}) = s\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} = 0.036\,423\,435.$

An approximate 95% confidence interval for $\bar{\mathbf{Y}}$, the average hardness of the lot of 2,000 castings, is then:

$$\bar{y} \pm t^*_{n-1} \times \hat{s.d.}(\bar{Y}) = 4.4 \pm 2.26216 \times 0.036423 = (4.318,\ 4.482) \text{ or about } (4.3_2,\ 4.4_8) \text{ B.I.D. units.}$$

where t^*_{n-1} is the *97.5*th percentile of the t_9 distribution.

(b) A confidence interval is a way of estimating the value of $\bar{\mathbf{Y}}$ (the *respondent* population average) which takes selecting variability into account – it quantifies the likely size of sample error, due to uncertainty arising from the incompleteness that is a consequence of investigating only a *subset* of the population elements (or units). A confidence interval constructed from the values of the attributes of a particular sample [taken as model values as in (a)] will cover the (measured) value of $\bar{\mathbf{Y}}$ in about 95% of the time under repeated selecting. [In this context, we must not forget the *other* (three) error categories which also impose limitations on sample survey answers.]

(c) Four quantities affect the width of a confidence interval for $\bar{\mathbf{Y}}$:

- $\mathbf{N}$, the size of the respondent population – unless the sampling fraction is relatively *high*, the width of a confidence interval is *in*sensitive to the value of $\mathbf{N}$;
- $\mathbf{S}$, a measure of the variation in the value of $\mathbf{Y}$ from element to element in the respondent population (in the case of this *inanimate* population of castings, the respondent and study populations are likely to be essentially identical) – the width of a confidence interval goes *up* directly with the magnitude of $\mathbf{S}$;
- n, the sample size (the number of respondents among the units selected for participation in the sample survey) – the width of a confidence interval goes *down* with approximately the *square root* of the magnitude of n;
- t^*_{n-1}, the confidence coefficient, whose value for given n is determined by the confidence *level* required for the confidence interval – the *higher* the level, the *wider* the interval (for *particular* values of sample attributes).

The ability to *control* the values of these quantities is as follows:

- $\mathbf{N}$ is an *attribute* of the respondent population whose value is *not* under the control of the person doing the selecting;
- $\mathbf{S}$ is another attribute of the respondent population whose value is *not* under the control of the person doing the selecting;
  - when *calculating* the end-points of a *specific* confidence interval, $\mathbf{S}$ usually has to be *estimated* by s [the *sample* (data) standard deviation], so that selecting variability *in*creases, in a sense, the limitation imposed on answers by our inability to alter the magnitude of $\mathbf{S}$; BUT:
  - as we will learn later in the course (when we discuss selecting from a *stratified* population in Part 4), limitations imposed on answers arising from variation in the value of $\mathbf{Y}$ among the respondent population elements can be *managed* by a Plan involving a more complicated sampling protocol than equiprobable selecting (EPS) from an unstratified population and by its (subsequent) data *analysis.*
- n *is* under the control of the person doing the selecting but is constrained by available resources in practice.
- t^*_{n-1} is also under the control of the person doing the selecting but, in practice, what are acceptable limitations on investigation Answer(s) may restrict the ability to control t^*_{n-1}.

T4 – 2. (a) This question involves calculating a sample size when estimating $\mathbf{P}$ under EPS when there is an *absolute* specification of precision (here, h = 5 percentage points or 0.05) at a 95% level of confidence.

We are given that $\mathbf{N}$ = 4,205 students, but no information is provided about the range of likely values of $\mathbf{P}$; we will therefore make a 'worst case' calculation using $\mathbf{P} = \mathbf{Q} = 0.5 = ½$.

We have: $n_0 = \frac{\mathbf{PQ}}{D^2}$ and $n = n_0\left[1 + \frac{n_0 - 1}{\mathbf{N}}\right]^{-1}$, where: $D = \frac{h}{z^*_\alpha}$.

so that with h = 0.05 and $z^*_\alpha = 1.95996$ to meet the requirements of an estimate of $\mathbf{P}$ to within 5 percentage points 19 times out of 20: $D = \frac{0.05}{1.95996} = 0.025\,5107$;

hence: $n_0 = \frac{½ \times ½}{0.025\,5107^2} = 384.14$ and: $n = 384.14\left[1 + \frac{383.14}{4{,}205}\right]^{-1} = 352.08 \simeq 352$;

i.e., about 350 students should be selected *if* there will be no non-response.

[In practice, there *is* likely to be non-response, so that it would probably be necessary to select initially more than 400 students for the sample survey.]

(*continued overleaf*)

T4 – 2. (cont.) (b) The question now involves calculating a sample size when estimating $\mathbf{P}$ under EPS when there is a *relative* specification of precision (here, r = 0.05 or 5 percent) at a 95% level of confdence.

Again, we are given that $\mathbf{N}$ = 4,205 students, but no information is provided about the range of likely values of $\mathbf{P}$ and, in this situation, we can*not* make a 'worst case' calculation similar to that in (a). [Why?]

We have: $n_0 = \dfrac{\mathbf{Q}/\mathbf{P}}{(r/z^*_\alpha)^2}$ and $n = n_0\left[1 + \dfrac{n_0 - 1}{\mathbf{N}}\right]^{-1}$,

so that with r = 0.05 and $z^*_\alpha = 1.95996$ to meet the requirements of estimating $\mathbf{P}$ to within 5 percent 19 times out of 20: $n_0 = \dfrac{\mathbf{Q}/\mathbf{P}}{0.025\,5107^2}$;

this and the previous expression for n are those required for calculating the sample size for this situation.

NOTE: Although we cannot make a 'worst case' calculation in (b) like that in (a), if we again set $\mathbf{P} = \mathbf{Q} = 0.5$, we obtain $n_0 = 1{,}536.6 \simeq 1{,}537$; this result is *four times* that for n_0 in (a), reflecting the fact that, in this instance, the specified precision would be *twice* as high in (b) as in (a) – 5 percent of $\mathbf{P} = 0.5$ is 2.5 (rather than 5) percentage points.

More generally, (a) and (b) become equivalent questions *only* as $\mathbf{P} \Rightarrow 1$ so that $r\mathbf{P} \Rightarrow 0.05$, the value of h in (a). For $\mathbf{P} < 1$, the specification of precision in (b) is *more* stringent than in (a) and the required sample size is therefore larger in (b).

***T4 – 3.** A sample selecting procedure is called *equiprobable selecting* if it makes all possible *samples* of a given size equally likely to be selected.

(a) Under this procedure, not all samples of the size proposed for the sample survey are equally probable; for instance, a sample which includes *both* members of a two-adult household is not possible, *i.e.*, such a sample has *zero* probability of being selected.

(b) Under this procedure, not all samples of the size proposed for the sample survey are equally probable; for instance, a sample which includes only *one* member of a two-adult household is not possible, *i.e.*, such a sample has *zero* probability of being selected.

NOTE: Part (a) of this question can also be answered by showing that the selection probabilities for *individual* elements are not all equal; the danger of such an approach is that it may make us more likely to forget that the *definition* of EPS is in terms of sample, *not* individual unit (or element), selection probabilities. Arguing from *individual* inclusion probabilities in (b) leads to a *wrong* answer.

You might like to *quantify* (or *formalize*) the argument based on individual inclusion probabilities for (a).

*Bonus question