UNIVERSITY OF SYDNEY
STAT2011 Statistical Models
sydney.edu.au/science/maths/stat2011

Semester 1 2012 Lecturer: Michael Stewart

Tutorial Week 8

Remember the three steps for computing a standard error for an estimate $\hat{\theta}(\mathbf{x})$ *based on data* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ *modelled as values taken by random variables* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$*:*

- Write down $\mathrm{MSE}_\theta(\hat{\theta}(\mathbf{X}))$ (possibly depending on θ);
- (if necessary) plug in the value of the estimate $\hat{\theta}(\mathbf{x})$ into this expression;
- take the square root; the standard error then takes the form

$$\mathrm{se}(\hat{\theta}(\mathbf{x})) = \sqrt{\mathrm{MSE}_{\hat{\theta}(\mathbf{x})}(\hat{\theta}(\mathbf{X}))}\,.$$

(note that the real number (i.e. the estimate) $\hat{\theta}(\mathbf{x})$ is the value taken by the random variable (i.e. the estimator) $\hat{\theta}(\mathbf{X})$).

1. [*Election night tracking*] Suppose that at an election, a certain electorate has 98,614 registered voters. There are only two candidates. At 8.30pm, 65,750 votes had been counted (about $\frac{2}{3}$) and one of the two candidates (say for Party A) had received 33,200 votes.

   We are going to use an urn model to estimate (and provide a corresponding standard error for) the overall proportion $0 < \theta < 1$ of voters who voted for Party A. In particular we model that the 65,750 votes counted represents a *simple random sample* of all 98,614 votes; it is modelled as having been obtained in such a way that all possible samples of size $n = 65,750$ (from the population of $N = 98,614$ votes) were equally likely (this may or may not be reasonable – on election night usually certain "booths" are counted more quickly than others and there may be some difference between the corresponding sub-districts distributed throughout an electorate).

   (a) Estimate the overall proportion $0 < \theta < 1$ of voters in the electorate voting for Party A.

   (b) Using an appropriate urn model, compute a standard error for the estimate in the previous part (see the three-point check above).

   (c) A common "rule of thumb" is that the estimate should be within a "few standard errors" of the true value, if the model is "correct". Taking this to mean ±3 se's, should the candidate declare victory (or concede defeat!) now, or wait until more votes are counted?

   (d) How would the answer to the previous question change if we forgot to take into account the fact that we are sampling without replacement here? That is, recompute the estimate and standard errors (if necessary) under a different assumption, that the random sample was taken with replacement.

2. [*Capture-recapture population size estimation*] When studying certain (particularly endangered) species, it is often desired to estimate how many actual individuals are present in a particular habitat. The basic idea is that a certain number w of animals are captured, tagged and then released back into the habitat. After a certain period of time (long enough for all the animals to "randomly mix") a certain number n are captured (the "recapture stage"). We can perhaps use a sampling-without-replacement urn model, where we have w tagged and b untagged and we wish to get some idea of the population size $N = w + b$. The number X of tagged animals in the "recapture" sample can be used to get some information concerning N. In fact, suppose the parameter we are interested in is the *reciprocal* $\theta = 1/N$.

   (a) Write down the expected value of X.

   (b) Derive an unbiased estimator of $\theta = 1/N$.

   (c) Suppose that in an example, $w = 10$ animals are tagged initially and then a sample of size $n = 30$ is taken at the recapture stage. If $x = 5$ animals in the recapture sample are tagged, compute an estimate for $\theta = 1/N$ and provide a standard error for this estimate.

   (d) A "rough interval estimate" can be provided using (*estimate* ±2 *standard errors*). Construct such a rough interval estimate and convert this into a corresponding rough interval estimate for N.

3. Nylon bars were tested for brittleness. Each of 280 bars was moulded under similar conditions and was tested in five places. Assuming that each bar has uniform composition, the number of breaks on a given bar should behave like a *Bin*$(5, \theta)$ random variable. The frequencies of the numbers of breaks for the 280 bars appears below.

Breaks/bar	0	1	2	3	4	5
Frequency	157	69	35	17	1	1

(a) This frequency table is a summary of the *original data* $x_1, x_2, \ldots, x_{280}$. Compute the average number of breaks per bar $\bar{x} = \frac{1}{280}\sum_{i=1}^{280} x_i$.

(b) Model the original data as values taken by IID RVs $X_1, X_2, \ldots, X_{280}$, each with a $Bin(5,\theta)$ distribution. Write down $E_\theta(X_1)$ and hence $E_\theta(\bar{X})$ where of course $\bar{X} = \frac{1}{280}\sum_{i=1}^{280} X_i$ is the sample average.

(c) Write down an unbaised estimator based on $\bar{X}$. Hence compute the corresponding estimate (recall that an estimate – a real number – is a value taken by an estimator – a random variable).

(d) Show that the estimator obtained above is also the maximum-likelihood estimator.

(e) Compute the sample variance of the original data (the $\frac{1}{n}$-version will do). Also write down $Var_\theta(X_1)$ as a function of θ, and plug-in the estimate to obtain an estimate of $Var_\theta(X_1)$. Does this suggest the IID binomial model is a good fit; why?

(f) Pool the last three classes (i.e. 3, 4 and 5 breaks) into a single class. Compute expected frequencies for the resultant 4 classes under the fitted binomial model, and hence compute standardised residuals. Do these suggest that the IID binomial model is a good fit, or not, and why?

(g) The sample variance suggests fitting the beta-binomial distribution (why?). Assuming a Pólya sampling model, what integers w and b give a population mean and variance closest to the sample versions? (**Hint:** you already have an estimate for $p = w/(w+b)$; use the sample variance to estimate $np(1-p)(N+n)/(N+1)$ where $N = w+b$, solve for N; finally round $w = pN$ and N to the nearest integer).

(h) Compute fitted beta-binomial probabilities for the values 0, 1 and 2. By subtraction determine the fitted probability for ≥ 3, the corresponding expected frequencies and resultant standardised residuals. Does this provide a reasonable fit; why?

4. [*Estimating a population variance*] Suppose we have IID RVs $X_1, X_2, \ldots, X_n$ and write

$$V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\,,$$

where as usual $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ denotes the average of the X_i's (note that both V and $\bar{X}$ depend on n!).

(a) By expanding the square show that $V = \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - \bar{X}^2$.

(b) Writing $\mu = E(X_1)$ and $\sigma^2 = Var(X_1)$, write down $Var(\bar{X})$ and hence $E\left[(\bar{X})^2\right]$.

(c) Show that V is biased as an estimator of σ^2.

(d) Write down $Var(\frac{1}{n}\sum_{i=1}^{n} X_i^2)$ in terms of μ, σ^2 and $\mu_4 = E(X_1^4)$.

(e) It can be shown that if

- two sequences of random variables $\{X_n\}$ and $\{Y_n\}$ are such that $X_n \xrightarrow{P} \mu$ and $Y_n \xrightarrow{P} \nu$;
- $g(x,y)$ as a function of two real variables x and y is continuous at the point (μ,ν),

then $g(X_n, Y_n) \xrightarrow{P} g(\mu,\nu)$. Use this to show that $V \xrightarrow{P} \sigma^2$.

5. [*Estimating a population mean*] Suppose that data $x_1, x_2, \ldots, x_n$ is modelled as values taken by IID RVs $X_1, X_2, \ldots, X_n$ with $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$ both unknown. It is desired to estimate μ; σ^2 here is known as a *nuisance parameter*: we don't really care what value it is, however not knowing it is not as convenient as knowing it.

(a) Write down $E(\bar{X})$ and $Var(\bar{X})$.

(b) Suppose that $n = 250$, the sample sum takes the value 2498 while the sample sum of squares takes the value 27008. Write down the value taken by the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the (unbiased) sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

(c) Write down the (theoretical) MSE of $\bar{X}$ viewed as an estimator of μ as a function of n and the parameters. In particular note that it depends on the nuisance parameter σ^2.

(d) Compute a standard error for the estimate $\bar{x}$ by plugging-in the value of s^2 as an estimate of σ^2 in the expression of the MSE and taking the square root.

6. Writing $Bias(\hat{\theta}(\mathbf{X})) = E_\theta(\hat{\theta}(\mathbf{X})) - \theta$, show that

$$\mathrm{MSE}_\theta(\hat{\theta}(\mathbf{X})) = E_\theta\{[\hat{\theta}(\mathbf{X})) - \theta]^2\} = Var(\hat{\theta}(\mathbf{X})) + [Bias(\hat{\theta}(X))]^2\,.$$