Foundations of Medical Statistics

Statistical Inference 2: Estimation and Precision

2.1 Estimators and their sampling distributions

Recall from Statistical Inference 1 that the basics of classical statistical methods involve making inferences from samples to population. In this session we are concerned with **estimation** of features of the population. Such population features will nearly always be represented by **parameters** in a **statistical model**. A statistical model is a probability distribution constructed to enable inferences to be made from the sample data about the population. (Non-parametric methods are much less useful when estimation is the main aim, mainly because an underlying probability distribution is not given in terms of the desired population parameter.)

The statistical model summarises relevant properties of the population **and the random process by which the sample is obtained**.

To estimate the unknown parameter, a statistic is calculated from the sample. Recall that a statistic is any function or combination of random variables (realised by the sample data) that does not depend on unknown parameters for its calculation. The statistic used to estimate the unknown population parameter is termed an **estimator** of the parameter; the term estimate is used for the value obtained by substituting sample data values into the formula for the estimator. The estimator, being a statistic which is a function of random variables, is itself a random variable; the estimate is a realisation of that random variable.

When sampling just one random variable, its sampling distribution is identical to the population distribution. However, in the context of estimation we are often concerned with sampling a number of random variables, n, from the population, and calculating a statistic from these to be used as estimator: here we need to obtain the sampling distribution of this statistic, and this will vary with n. In some contexts, however, one random variable suffices: care must be taken to note this, and in the following three examples the first is presented as requiring n random variables, while the other two are presented as requiring just one.

Example 2.1.1
The Forced Expiratory Volume in one second (FEV1) is a continuous measure of lung function. A group of n patients randomly sampled from the patients attending an asthma clinic have their FEV1 measured, with a view to estimating the mean FEV1 of patients generally attending that clinic.

Model assumed: Let $Y_1, ..., Y_n$ be random variables representing the random process whereby the FEV1 observations $y_1, ..., y_n$ are obtained in the sample. We will assume that each random variable is drawn from the same population distribution, and that the observations are independent of each other. We will use the abbreviation *iid* to indicate in this way that the random variables are **independent and identically distributed**. Further, we will assume that the population distribution is Normal with population mean μ and population variance σ^2.

This model can be compactly written as:

$$Y_i \overset{iid}{\sim} N(\mu, \sigma^2), \;\; i = 1,..., n.$$

Estimator for μ: we will propose the sample arithmetic mean, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The arithmetic mean is intuitively acceptable, but in Statistical Inference 3 we introduce a formal framework for justifying use of estimators. The observed sample arithmetic, $\bar{y}$, the realisation of $\bar{Y}$, is thus the estimate obtained for the unknown μ.

Sampling distribution of estimator: The sampling distribution of this statistic is known, given the model: it can be shown that

$$\bar{Y} \overset{iid}{\sim} N(\mu, \sigma^2/n).$$

EXERCISE 2.1.1

a) Show that, given the model for Y_i, the expectation and variance of $\bar{Y}$ are as given above. From Probability lectures recall that the sum of Normally distributed random variables is itself Normally distributed: the result above then follows directly.

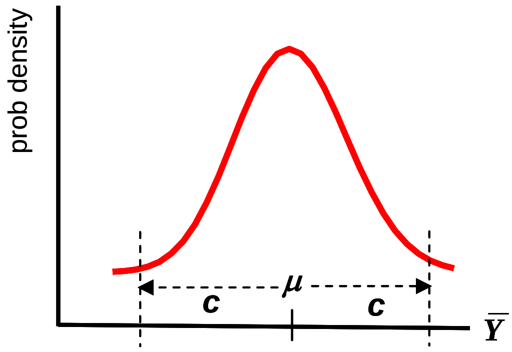
b) Consider $Z = \dfrac{\bar{Y} - \mu}{\sqrt{\mathrm{Var}(\bar{Y})}}$. Show that $Z \sim N(0,1)$

Confidence interval for μ: The more narrowly distributed an estimator is around its expectation (the smaller its variance), the greater the **precision** it is said to have. When the expectation of an estimator is exactly the target population parameter, the dispersion of the estimator around its expected value, as measured by the estimator variance, is an indication of how close to the population value the estimate is likely to be. A more useful and direct measure of this closeness is given by the confidence interval which the estimator can provide for the parameter. We can use our knowledge of the sampling distribution of the estimator, given the population parameters, to obtain a confidence interval. Recall from Statistical Inference 1 that the 95% confidence interval for μ is given by a region around the estimator such that the probability that this region contains μ is 0.95: in the long run, 95% of samples will contain μ within this region, and 5% of samples will fail to contain it. Since the sampling distribution of the estimator here is symmetrical, it suffices to find the symmetrical ‘tails’ each of which fail to contain μ with probability 0.025, and then exclude these tails from the interval. In other words, we need to find c such that:

$$\mathrm{Prob}(\bar{Y} > \mu + c) = 0.025$$
$$\mathrm{Prob}(\bar{Y} < \mu - c) = 0.025,$$

and thus define random variables for the 95% lower (L) and upper (U) confidence limits:

$$L = \bar{Y} - c \;\; \Rightarrow \;\; \mathrm{Prob}(L > \mu) = 0.025$$
$$U = \bar{Y} + c \;\; \Rightarrow \;\; \mathrm{Prob}(U < \mu) = 0.025$$



We can solve for c:

$$\text{Prob}(\bar{Y} > \mu + c) = \text{Prob}(\bar{Y} - \mu > c) = 0.025$$

$$\Rightarrow \quad \text{Prob}\left(\frac{\bar{Y} - \mu}{\sqrt{\text{Var}(\bar{Y})}} > \frac{c}{\sqrt{\text{Var}(\bar{Y})}} \right) = 0.025$$

$$\Rightarrow \quad \text{Prob}\left(Z > \frac{c}{\sqrt{\text{Var}(\bar{Y})}} \right) = 0.025$$

but we saw earlier that Z ~ N(0,1), and we know from the Probability lectures that Prob(Z>**1.96**) = 0.025, so

$$\frac{c}{\sqrt{\text{Var}(\bar{Y})}} = 1.96 \;\Rightarrow c = 1.96\sqrt{\text{Var}(\bar{Y})}$$

and so 95% CI for $\mu = \bar{Y} \pm 1.96\sqrt{\text{Var}(\bar{Y})} = \bar{Y} \pm 1.96\, \sigma/\sqrt{n}$

$\sqrt{\text{Var}(\bar{Y})}$ is known as the **standard error** of the estimator $\bar{Y}$.

Example 2.1.2

A routine surgical procedure is performed at a hospital to cure a certain medical condition. A group of n patients is randomly sampled from the patients undergoing this operation, with a view to estimating the proportion of patients that can generally be expected to be cured by this procedure.

Model assumed: Let the single random variable R represent the number of cured patients out of n sampled to undergo the procedure. We assume the model:

$$R \sim Bin(n, \pi)$$

in other words, that the number cured is binomially distributed with population proportion π being cured. Note that n is part of the sampling design: only π is an unknown parameter.

Estimator for π : we will propose the proportion $P = R/n$. Thus the sample based estimate will be $p = r/n$.

Sampling distribution of estimator: From Probability lectures, we know that

$$\mathrm{E}(P) = \pi \text{ and } \mathrm{Var}(P) = \frac{\pi(1-\pi)}{n}.$$

the sampling distribution of $P=R/n$ can be obtained directly from the binomial distribution of R above, since $\mathrm{Prob}(a < P < b) = \mathrm{Prob}(na < R < nb)$. However, to construct confidence intervals in this setting a Normal approximation is often used; both methods, using the binomial distribution and the Normal approximation, will be presented later in Analytical Techniques 2, which will also discuss issues arising from the discreteness of distributions such as the binomial.

Example 2.1.3
Following a serious surgical procedure, a number of subjects were followed up for a total of p person-years, with a view to estimating the rate of deaths following this surgery.

Model assumed: Let the single random variable D be the number of deaths observed over follow-up p person-years, then we assume D is drawn from a Poisson distribution:

$$D \sim Po(\mu),$$

where μ, the expected number of deaths, = λp, and λ is the rate parameter, the expected number of deaths per person-year.

Estimator for λ : we will propose the rate D/p. Thus the sample based estimate will be the observed rate d/p.

Sampling distribution of estimator: We know from Probability lectures that here $\mathrm{E}(D) = \mu$ and $\mathrm{Var}(D) = \mu$. We could use the Poisson distribution to obtain confidence intervals, but again this is often done by means of a Normal approximation; these methods for confidence intervals will be presented later in Analytical Techniques 2.

2.2 General distribution

The importance of the Normal distribution in inference, and its application to models not themselves defined in terms of Normal random variables, is essentially due to the Central Limit Theorem, discussed in Probability lecture 4.

Provided that Y_i are drawn from a single distribution with expectation and variance, i.e. that

$$\mathrm{E}(Y_i) = \mu \quad \text{and} \quad \mathrm{Var}(Y_i) = \sigma^2$$

both exist, then as we saw earlier

$$\mathrm{E}(\bar{Y}) = \mu \quad \text{and} \quad \mathrm{Var}(\bar{Y}) = \sigma^2/n.$$

[Note that if the population is finite, draws must be with replacement, since without replacement dependence is induced. See Rice, Chapter on Simple Random Sampling.]

We saw in particular that $Y_i \sim N(\mu, \sigma^2) \Rightarrow \bar{Y} \sim N(\mu, \sigma^2/n)$. However, as a consequence of the Central Limit Theorem (Probability 4), we can go further. Even when we do not know the precise form of the distribution, $Y_i \sim (\mu, \sigma^2)$, but it has mean and variance, then except in very pathological settings, as n gets large the distribution of $\bar{Y}$ tends to Normal: **asymptotically**

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

Many common tests and procedures use this asymptotic result. It also implies that inferences about means can be made when very little is assumed about the original distribution, provided we can get a reasonable estimate of σ^2.

2.3 Properties of an estimator

Some important questions arise from the previous section:

1. What makes a sensible or good estimator?
2. How should we choose between alternative estimators?
3. How do we find estimators in more complicated contexts?

The last two questions depend on the answer to the first, which in turn depends on the sampling distribution of the estimator.

The principal properties of an estimator are based on the expectation and variance of its sampling distribution. The square root of the variance, that is, the standard deviation of the sampling distribution of the estimator, is called the **standard error** of the estimator.

2.3.1 Bias

Suppose T is an estimator for a parameter θ. In general we want the sampling distribution of the estimator to be centered in the “right place”, which of course is θ itself. In other words, we want

$$\mathrm{E}(T) = \theta.$$

If this true, the estimator is said to be **unbiased**. Otherwise the difference between the expectation and the parameter is called the bias:

$$\mathrm{bias}(T) = \mathrm{E}(T) - \theta.$$

It is not so important that the bias is exactly zero — many common estimators are biased. It is critical, however, that the bias tends to zero as the sample size gets larger, in which case it is asymptotically unbiased:

T is an **unbiased** estimator for θ if $\mathrm{E}(T) = \theta$.
T is an **asymptotically unbiased** estimator for θ if $\lim_{n\to\infty} \mathrm{E}(T) = \theta$.

A very closely related property of an estimator is **consistency**: essentially, an estimator is consistent if its distribution concentrates in an increasingly narrow region around the parameter as the sample size increases: in other words, if **both the bias and the variance of the estimator approach zero**. Estimators without this property are said to be **inconsistent**, which usually makes an estimator unusable. But if an estimator is asymptotically unbiased *and* its variance also tends to zero, then it will be consistent.

2.3.2 Efficiency

As we see above, it is not enough that an estimator should have small bias: the expectation of the distribution may be close to the target parameter θ, yet the dispersion of the distribution may be large: we want the variance of the estimator to be "small". Typically, of two estimators with the same bias, the one with smaller variance is the better; and the smaller the variance the more **precise** the estimator is said to be. The ratio of the variances $\text{Var}(S)/\text{Var}(T)$ is called the **relative efficiency** of estimator T relative to S. This comparison is most meaningful when both estimators are unbiased or have the same bias. Often the relative efficiency of two estimators varies with sample size, and then it is useful to consider **asymptotic relative efficiency**, the limit of the ratio as the sample size increases. However, note that the asymptotic relative efficiency may not be applicable to small samples.

The best estimator from several unbiased estimators, the one with the smallest variance (greatest precision), is called the **minimum variance unbiased estimator**, or an **efficient** estimator.

2.3.3 Relative efficiency of mean versus median

In the $N(\mu, \sigma^2)$ distribution, as in any symmetric distribution, the mean and median are both equal to the population parameter μ, and both the sample mean ($\bar{Y}$) and median ($\dot{Y}$) are unbiased for μ. Therefore it is sensible to choose between them on the basis of their precision.

We know that if $Y_i \sim N(\mu, \sigma^2)$ then $\text{Var}(\bar{Y}) = \sigma^2/n$. However, it can be shown that for a Normal population like this, for large n approximately

$$\text{Var}(\dot{Y}) = (\pi/2)(\sigma^2/n) = 1.571\sigma^2/n$$

giving an asymptotic relative efficiency for the sample mean relative to the sample median of

$$\frac{\text{Var}(\dot{Y})}{\text{Var}(\bar{Y})} = 1.571.$$

Thus we would prefer to use the mean. Note, however, that this result depends on the assumption of Normality and a large sample size: if a population is symmetrically distributed but with **heavier tails** than Normal, this would not necessarily be true; in this case the distribution has outlying observations, and the sample mean may have larger variance than the median: recall that the mean takes into account all the observations, even distant outliers that the sample median ignores. On the other hand, if the population distribution were not symmetric the median would not be consistent for the population mean, and would not therefore be considered for use as an estimator for the population mean: the present comparison would not then be sensible.

2.3.4 Mean Square Error

To compare two estimators with different biases, we can measure the dispersions not around the expectations of each estimator (giving the two variances) but around the true population value. The expected variability of an estimator around the population parameter is called the **Mean Square Error (MSE)**.

Recall that

$$\mathrm{Var}(T) = \mathrm{E}\{[T - \mathrm{E}(T)]^2\},$$

thus

$$\mathrm{MSE}(T) = \mathrm{E}[(T - \theta)^2].$$

This can be partitioned into the sum of the variance and the square of the bias. First we partition the squared deviations, before taking expectations. This type of partitioning of squared deviations is frequently encountered in other contexts (see later in 2.4) and is of fundamental importance:

$$\begin{aligned}
\mathrm{MSE}(T) &= \mathrm{E}[(T - \theta)^2] = \mathrm{E}\{[[T - \mathrm{E}(T)] + [\mathrm{E}(T) - \theta]]^2\} \\
&= \mathrm{E}\{[T - \mathrm{E}(T)]^2 + [\mathrm{E}(T) - \theta]^2 + 2[T - \mathrm{E}(T)][\mathrm{E}(T) - \theta]\} \\
&= \mathrm{E}\{[T - \mathrm{E}(T)]^2\} + \mathrm{E}\{[\mathrm{E}(T) - \theta]^2\} + 0
\end{aligned}$$

(*last term is* 0 *since* $[\mathrm{E}(T) - \theta]$ *is a constant and* $\mathrm{E}[T - \mathrm{E}(T)]=0$)

$$\begin{aligned}
&= \mathrm{Var}(T) + [\mathrm{E}(T) - \theta]^2 \\
&= \mathrm{Var}(T) + [\mathrm{bias}(T)]^2
\end{aligned}$$

Among several alternative estimators, we are typically looking for the one with smallest MSE (though other issues such as simplicity and convenience play a part in practice).

2.4 Estimation of population variance; degrees of freedom

We have seen that the variance in a population is often required to make inferences about other statistics, particularly those based on the population mean. The definition of a population variance is, as we know, the average squared deviation about the population mean. So if $Y_i \sim (\mu, \sigma^2)$, where we have no distributional assumptions other than the existence of population mean μ and variance σ^2, then an intuitively appropriate estimator of the population variance σ^2, based on a sample of size n, might be:

$$V_\mu = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 \tag{1}$$

Note: In the above, $\mu = \mathrm{E}(Y)$ by definition; but Y is not necessarily unbiased: it depends whether μ is the target parameter to be estimated.

EXERCISE 2.4.1
Show that V_μ is unbiased for σ^2.

The population mean is usually unknown, however, so the best we can do is use an estimate of the population mean: the sample mean. Hence, for an estimator the population variance based on a sample of size *n* we might consider:

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \qquad (2)$$

Now, the squared deviation around the sample mean ignores the deviation between the sample mean and the population mean, so the formula above will tend slightly to underestimate the average squared deviation around the population mean, and thus to underestimate the population variance.

This underestimation can also be considered a symptom of the fact that we have not properly taken into account the **degrees of freedom** involved in the estimation. The degrees of freedom involved in an estimation is a fundamental but subtle concept in statistical inference: essentially the degrees of freedom are the *number of independent pieces of information* used in the estimation. We will meet and apply this concept throughout the course, and one of its subtleties is that pieces of information are different in different contexts: here, for example, the pieces of information correspond to the individual observations sampled; but in other contexts the correspondence is more subtle, and one piece of information may be the result of a count on a number of observations in a certain category. It is therefore hard to explain degrees of freedom in a manner which adequately prepares for all its applications. Here, however, we can use the intuitive sense of a piece of information supplied by each of the *n* sample observations, to illustrate the idea in a number of related ways:

- Note that in applying formula (1) above, all *n* observations are independent not just of each other, but of the population mean μ around which the dispersion is to be estimated: for example, μ could have a value entirely outside the range of the sample observations (this is mathematically possible, though not very probable in a large sample). On the other hand, in formula (2), the sample mean is not independent of all *n* sample observations: this is clear immediately in the fact that $\overline{Y}$ is constrained mathematically to fall within the range of sample observations (which contributes as mentioned to the underestimation of the population variance). However, given only *n*-1 of the *n* observations, the sample mean is still mathematically entirely unconstrained. Conversely, given the sample mean value, only *n*-1 observations are mathematically independent of it.

- If we took just one observation as the estimating sample, this would coincide with the sample mean for *n*=1, and could provide no information with which to estimate *dispersion* around the unknown population mean: only the subsequent *n*-1 observations could provide information on dispersion. The single observation can of course provide information on the location of the population mean, but it could only provide information on population variance if we also knew the population mean.

This suggests that the following estimator, with a slightly higher estimate, may be unbiased for the population variance:

$$V_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad (= \frac{n}{n-1}V_n) \tag{3}$$

We now prove that this is unbiased using the following strategy (see the examinable Appendix for an alternative proof):

i) We know that V_μ in (1) is unbiased for the population variance σ^2. So we start by partitioning V_μ, as suggested earlier, into an average squared deviation of observations around the sample mean, and the squared deviation of the sample mean around the population mean. Note the fundamental similarity between partitioning the sums of squared deviations below and the earlier partitioning of expected squared deviations for the MSE (2.3.4).

ii) The partition will now be recognizable as including V_n, and we will obtain its expectation. The expectation of V_{n-1} will then follow directly.

i) Partitioning V_μ

$$V_\mu = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}[(Y_i - \bar{Y}) + (\bar{Y} - \mu)]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}[(Y_i - \bar{Y})^2 + (\bar{Y} - \mu)^2 + 2(Y_i - \bar{Y})(\bar{Y} - \mu)]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \frac{1}{n}\sum_{i=1}^{n}(\bar{Y} - \mu)^2 + \frac{2}{n}(\bar{Y} - \mu)\sum_{i=1}^{n}(Y_i - \bar{Y})$$

$$= V_n + (\bar{Y} - \mu)^2 \quad \text{(note that } \sum_{i=1}^{n}(Y_i - \bar{Y}) = 0) \tag{4}$$

ii) Obtaining $E(V_n)$, given that $E(V_\mu) = Var(Y) = \sigma^2$:

$$(4) \Rightarrow \qquad E(V_n) = E(V_\mu) - E\left[(\bar{Y} - \mu)^2\right] = Var(Y) - Var(\bar{Y}) = \sigma^2 - \frac{\sigma^2}{n}$$

$$= \sigma^2\left(\frac{n-1}{n}\right)$$

So V_n is biased, though asymptotically unbiased. As anticipated, this estimator slightly underestimates σ^2, being biased downwards by the variance of the sample mean, which V_n ignores; but we can readily correct for this underestimation:

$$\mathrm{E}\left[\frac{n}{n-1}V_n\right] = \frac{n}{n-1}\mathrm{E}[V_n] = \sigma^2 \Rightarrow \text{(from (3)) } \mathrm{E}(V_{n\text{-}1}) = \sigma^2.$$

It is because of this that $V_{n\text{-}1}$, which is of course the standard expression for the **sample variance**, is preferred to V_n.

Typically, when estimation requires the use of sample estimators in place of their target population parameter, such as $\bar{Y}$ above for μ, one degree of freedom is lost for each estimator used.

2.5 Distribution of 'sample variance' estimator

The notation S^2 is commonly used for the sample variance estimator designated above as $V_{n\text{-}1}$:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

We have shown above that

$$\mathrm{E}(S^2) = \sigma^2.$$

However, a non-linear function (such as a square root) of an unbiased estimator of a parameter is not unbiased for the same function of that parameter. Thus S, the sample standard deviation, **is biased** for σ, the population standard deviation, though it is asymptotically unbiased (you will be asked to prove the bias in the practical).

It can also be shown (see Appendix) that

$$\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$$

The χ^2_m, 'chi-squared distribution on m degrees of freedom', is skewed to the right, and becomes more symmetric (asymptotically tending to Normal) as m increases. Since the expectation (m) and variance ($2m$) of a chi-squared distribution are entirely specified by the degrees of freedom m, the style of notation $\sim\chi^2(m, 2m)$ is not used and is instead generally condensed to $\sim\chi^2_m$.

Further details of the chi-squared distribution, which we met in Probability 4, and which will be very frequently encountered during the course, are given in the Appendix. It can also be shown (Appendix) that

$$\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}.$$

Appendix (*examinable*)

Chi-squared distribution

Some of these results are introduced in Probability 4:

$$X \sim N(0,1) \Rightarrow X^2 \sim \chi_1^2$$

$$X_1,..., X_n \overset{iid}{\sim} N(0,1) \Rightarrow \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

Below I will use X_n^2 to refer to a chi-squared df = n random variable.

Thus $X_m^2 + X_n^2 = X_{m+n}^2 \sim \chi_{m+n}^2$

E(X_1^2) = Var(X) + [E(X)]^2 = 1 + 0 = 1 (since $X \sim N(0,1)$)

Var(X_1^2) = 2 [As an exercise you could prove this; you will probably need to integrate by parts.]

E(X_n^2) = n

Var(X_n^2) = $2n$

$n \to \infty \Rightarrow X_n^2 \sim N(n,2n)$ [due to Central Limit Theorem]

Distribution of S^2 for Normal population

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Suppose $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1,..., n$. [Note the Normality assumption.]

$$\Rightarrow \quad \frac{X-\mu}{\sigma} \sim N(0, 1)$$

$$\Rightarrow \quad \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi_1^2$$

$$\Rightarrow \quad \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi_n^2 \qquad (5)$$

The strategy now is similar to that used in 2.4: note that in (5) the usually unknown parameter μ is required. If we substitute $\overline{X}$ for μ <u>we will lose one degree of freedom</u>:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi_{n-1}^2 \qquad (5b)$$

$$\Rightarrow \quad \frac{1}{\sigma^2}(n-1).S^2 \sim \chi_{n-1}^2$$

And we can achieve (5b) formally by partitioning as in 2.4:

Ignoring the divisor *n* throughout 2.4 (4) implies we can partition thus:

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{i=1}^{n}(\overline{X}-\mu)^2 \qquad (6)$$

(5),(6) $\Rightarrow$ $\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2+\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(\overline{X}-\mu)^2 \sim \chi_n^2$

$\Rightarrow$ $\dfrac{n-1}{\sigma^2}S^2+\dfrac{n}{\sigma^2}(\overline{X}-\mu)^2 \sim \chi_n^2$

$\Rightarrow$ $\dfrac{n-1}{\sigma^2}S^2+\left(\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi_n^2$ [note the second term is a *N*(0,1) squared.]

$\Rightarrow$ $\dfrac{n-1}{\sigma^2}S^2 + X_1^2 = X_n^2$

$\Rightarrow$ $\dfrac{n-1}{\sigma^2}S^2 = X_n^2 - X_1^2 = X_{n-1}^2$

$\Rightarrow$ $\dfrac{n-1}{\sigma^2}S^2 \sim \chi_{n-1}^2$, as required.

Another way of expressing this is that:

$$S^2=\frac{\sigma^2}{n-1}X_{n-1}^2 \qquad (7)$$

It follows from (7) that, since E(X_{n-1}^2) = n -1 (see E(X_n^2) = n earlier)

$$\mathrm{E}\left(S^2\right)=\frac{\sigma^2}{n-1}n-1$$

$\Rightarrow$ E(S^2) = σ^2 (as already proved in 2.4); and, also from (7)

Var(S^2) = $\left(\dfrac{\sigma^2}{n-1}\right)^2$ Var(X_{n-1}^2) = $\left(\dfrac{\sigma^2}{n-1}\right)^2$ 2(n –1) = $\dfrac{2\sigma^4}{n-1}$, as required.