1 Review of Chapter 7 in Wild and Seber

The main theme of Chapter 7 is:

The sample is not the population.

To keep this in mind, we distinguish

sample characteristics also called *estimates*.

and

population characteristics also called *parameters*.

Parameters are **not random**. There is only one population (in one particular application). Hence only one population mean, only one population standard deviation, and so forth. In real life, parameters are unknown quantities. You don't have population data. That's why you are using statistics on sample data.

Estimates are **random variables**. If the sample is random, then so are quantities (estimates) calculated from it. So, like any other random variable, *an estimate has a probability distribution and a mean and standard deviation.*

1.1 Three Conventions

1.1.1 Roman and Greek Letters

To clearly distinguish estimates and parameters we (sometimes) use

Roman letters for *estimates*

like

$\bar{x}$	for the sample mean
s_X	for the sample standard deviation

and

Greek letters for *parameters*

like

μ_X	for the population mean
σ_X	for the population standard deviation

Note: When there is only one variable X under discussion, we often drop the subscripts, writing s, μ, and σ rather than s_X, μ_X, σ_X.

1.1.2 Hats

There is also an entirely different convention for the same thing. To clearly distinguish estimates and parameters we (at other times) use

> *letters decorated with "hats"* for *estimates*

like

$\hat{p}$	for the sample proportion
$\hat{\theta}$	for a generic estimate (sample characteristic)

and

> *undecorated letters* for *parameters*

like

p	for the population proportion
θ	for a generic parameter (population characteristic)

1.1.3 Capital and Small Letters

A fairly subtle convention, not nearly as important as the two preceeding (*so if you have to skip something in this review, skip this*), distinguishes

> *capital letters*, like X, $\overline{X}$, and S_X, for *random variables*

and

> *small letters*, like x, $\bar{x}$, and s_X, for *observed values* of those random variables.

The subtile point is that after a random variable is observed, it is no longer random. When I calculate that the sample mean of my data is 2.716, then I write

$$\bar{x} = 2.716$$

because 2.716 is not random. It's just a plain ordinary number. In contrast, the equation

$$\overline{X} = 2.716$$

may be either true or false depending on what the value of $\overline{X}$ turns out to be when observed. In fact, for a continuous probability model

$$\operatorname{pr}(\overline{X} = 2.716)$$

is zero because continuous probability models give positive probability only to *intervals* not single numbers (p. 236 in Wild and Seber).

It's often not clear whether you should use $\overline{X}$ or $\bar{x}$. Sometimes it depends on context which has not been made clear enough to decide.

There are, however, two places where *capital letters are required*:

> in *probabilities and expectations* like $\operatorname{pr}(\overline{X} > 10)$ and $E(\overline{X})$, and
> in *subscripts indicating random variables* like s_X and σ_X.

1.2 Means and Standard Deviations of Estimators

If $\overline{X}$ is the mean and S_X the standard deviation of a random sample of size n from a population with mean μ and standard deviation σ, then

$$\begin{aligned}
E(\overline{X}) &= \mu \\
\mathrm{sd}(\overline{X}) &= \frac{\sigma}{\sqrt{n}} \\
\mathrm{se}(\overline{X}) &= \frac{s_X}{\sqrt{n}}
\end{aligned}$$

If $\widehat{P}$ is the sample proportion for a random sample of size n from a population with population proportion p, then

$$\begin{aligned}
E(\widehat{P}) &= p \\
\mathrm{sd}(\widehat{P}) &= \sqrt{\frac{p(1-p)}{n}} \\
\mathrm{se}(\widehat{P}) &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\end{aligned}$$

If $\overline{X}_1$ and $\overline{X}_2$ are the means and S_1 and S_2 the standard deviations of two **independent** random samples of sizes n_1 and n_2 from populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, then

$$\begin{aligned}
E(\overline{X}_1 - \overline{X}_2) &= \mu_1 - \mu_2 \\
\mathrm{sd}(\overline{X}_1 - \overline{X}_2) &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\
\mathrm{se}(\overline{X}_1 - \overline{X}_2) &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\end{aligned}$$

If $\widehat{P}_1$ and $\widehat{P}_2$ are the sample proportions for **independent** random samples of size n_1 and n_2 from populations with population proportions p_1 and p_2, then

$$\begin{aligned}
E(\widehat{P}_1 - \widehat{P}_2) &= p_1 - p_2 \\
\mathrm{sd}(\widehat{P}_1 - \widehat{P}_2) &= \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}} \\
\mathrm{se}(\widehat{P}_1 - \widehat{P}_2) &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}
\end{aligned}$$

1.3 Large Sample Theory

All of the estimates mentioned in the preceeding section ($\bar{X}$, $\widehat{P}$, $\overline{X}_1 - \overline{X}_2$, and $\widehat{P}_1 - \widehat{P}_2$) have approximately normal distributions for large enough sample sizes (n for one sample, n_1 and n_2 for two samples).

1.4 Confidence Intervals

In Chapter 7 Wild and Seber call these "two standard error intervals," but in Chapter 8 we find out they are really called *confidence intervals*.

1.4.1 Interval Estimates and Point Estimates

The thingies discussed in this section are called *interval estimates*. For contrast, the estimates previously discussed, like $\bar{x}$ and $\hat{p}$ are called *point estimates*.

1.4.2 Large Sample (Approximate) Intervals

For any point estimate having an approximately normal sampling distribution

$$\text{point estimate} \pm 2\,\text{se}(\text{point estimate})$$

is an approximate 95% confidence interval for the parameter that the point estimate estimates.

Generically,

$$\hat{\theta} \pm 2\,\text{se}(\hat{\theta})$$

is an approximate 95% confidence interval for θ.

And

$$\begin{gathered}
\bar{x} \pm 2\,\text{se}(\bar{x}) \\
\hat{p} \pm 2\,\text{se}(\hat{p}) \\
\bar{x}_1 - \bar{x}_2 \pm 2\,\text{se}(\bar{x}_1 - \bar{x}_2) \\
\hat{p}_1 - \hat{p}_2 \pm 2\,\text{se}(\hat{p}_1 - \hat{p}_2)
\end{gathered}$$

are approximate 95% confidence intervals for the parameters that the point estimates estimate, when the appropriate conditions are satisfied.

- The sample size is large (both sample sizes are large in the two-sample cases).
- In the two-sample cases, the samples are **independent**.

Plugging in the formulas for the standard errors,

$$\bar{x} \pm 2 \frac{s_X}{\sqrt{n}}$$

is an approximate 95% confidence interval for μ,

$$\hat{p} \pm 2 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

is an approximate 95% confidence interval for p,

$$\bar{x}_1 - \bar{x}_2 \pm 2 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

is an approximate 95% confidence interval for $\mu_1 - \mu_2$, and

$$\hat{p}_1 - \hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

is an approximate 95% confidence interval for $p_1 - p_2$.

Note: For confidence levels other than 95% see Section 1.4.4 below.

1.4.3 Small Sample (Exact) Intervals

Nothing in the preceeding section is useful for small samples. For *proportions* there is no small sample theory. But for *means* there is. In Chapter 7 we only do the one-sample case (the two-sample case will come later).

The confidence interval for μ in the preceeding section, was derived from the fact that

$$T = \frac{\overline{X} - \mu}{S_X/\sqrt{n}} \approx \text{Normal}(0, 1) \tag{1}$$

where the "double wiggle" sign means "approximately distributed as." An exact (small sample) confidence interval for μ can be derived from the analogous fact that

$$T = \frac{\overline{X} - \mu}{S_X/\sqrt{n}} \sim \text{Student}(n-1) \tag{2}$$

where the "single wiggle" sign means "exactly distributed as" and Student(d) means the Student's t-distribution with d degrees of freedom.

The difference between the two theories is that

- (1) holds (approximately) regardless of the population distribution for *sufficiently large sample size* n.
- (2) holds (exactly) for a *normal population distribution* regardless of the sample size n.

The relation between (1) and the confidence interval is that the two equations

$$\text{pr}\left(-2 < \frac{\overline{X} - \mu}{S_X/\sqrt{n}} < 2\right) \approx 0.95$$

$$\text{pr}\left(\overline{X} - 2\frac{S_X}{\sqrt{n}} < \mu < \overline{X} + 2\frac{S_X}{\sqrt{n}}\right) \approx 0.95$$

are equivalent, and the latter is the claim made for the confidence interval.

Hence in order to get *exact* (not approximate) confidence intervals *assuming a normal population distribution* we only need to substitute for 2 the t such that

$$\text{pr}(-t < T < t) = 0.95 \tag{3}$$

where $T \sim \text{Student}(n-1)$. This t is called the t *critical value* for 95% confidence and is *different for each sample size* n.

The t critical values for 95% confidence are given in the column headed **0.025** of Appendix 6 in Wild and Seber or by either of the R commands

```
qnorm(0.975, n - 1)
- qnorm(0.025, n - 1)
```

where `n` is the sample size (so `n - 1` is the degrees of freedom).

For example, if $n = 10$, then

$$\bar{x} \pm 2.262 \frac{s_X}{\sqrt{n}}$$

is an exact 95% confidence interval for μ, and if $n = 5$, then

$$\bar{x} \pm 2.571 \frac{s_X}{\sqrt{n}}$$

is an exact 95% confidence interval for μ.

Warning:

- These intervals are exact only if the *population distribution is exactly normal.*
- If the population distribution is close to but not exactly normal, then the these intervals are approximate (their actual coverage probability is near their nominal 95% coverage probability).
- If the population distribution is nowhere near normal, then these intervals are totally bogus.

Note: For confidence levels other than 95% see Section 1.4.4 below.

1.4.4 Different Confidence Levels

For confidence levels other than 95%, just change the 0.95 in (3) to some other number.

To get

$100(1 - \alpha)\%$ confidence

the t critical value is

the $1 - \alpha/2$ quantile of the Student$(n - 1)$ distribution

or

minus the $\alpha/2$ quantile of the Student$(n - 1)$ distribution.

Thus

confidence level	column of Appendix 6 headed
90%	0.05
95%	0.025
99%	0.005

and so forth.

Approximate Large-Sample Intervals The same trick works for large-sample intervals based on the approximate normality of the sampling distribution of a point estimate. Just use the Normal$(0, 1)$ distribution instead of the Student$(n-1)$ distribution. This the Student's t-distribution with "infinity degrees of freedom" in the bottom row of Appendix 6 in Wild and Seber. Hence

confidence level	z critical value
90%	1.645
95%	1.960
99%	2.576

(We call it a z critical value rather than a t critical value because a standard normal random variable is often denoted Z.)

Note also that a finicky person also uses 1.96 s. e. intervals rather than 2 s. e. intervals for 95% confidence (not that it really matters, it's only approximate anyway).