Here are some summary notes on sampling distributions of point estimators, and classical confidence intervals, all compared with reference Bayesian results in the key case of normal random samples. Reading in DeGroot (7.1, 7.4, 7.5, 7.6). Additional comments and examples from other models highlight some of the ideas.

**Example 1: Normal Model Review**

We are to observe data $X = \{x_1, \ldots, x_n\}$ in the normal random sample model $(x_i|\theta) \sim N(\theta, \sigma^2)$ with variance known. We now know that the reference posterior is $(\theta|X) \sim N(\bar{x}, \sigma^2/n)$. Some key facts we’ve used in inference about $\theta$ are:

- The posterior mean is $E(\theta|X) = \bar{x} - \bar{x}$, this is the key choice of a simple point estimate of $\theta$. It is also the posterior median, the posterior mode and the MLE of $\theta$.
- Central probability intervals are of the form $\bar{x} \pm z\sigma/\sqrt{n}$ where $z > 0$ is an upper-tail percentile of the standard normal distribution. For example, the 95% interval has $z = 1.96$. More generally, for a specified interval probability of $1 - p$ for some $p$, $z$ is the upper 100$(1 - p/2)$% percentile (=quantile) of the standard normal distribution (i.e., the point above which the probability is $p/2$).
- The intervals arise from the fact that the standardised random quantity $Z = (\theta - \bar{x})/\sqrt{\sigma^2/n}$ is $N(0, 1)$. So for any percentile $z > 0$,

$$Pr(-z < Z < z) = Pr(\bar{x} - z\sigma/\sqrt{n} < \theta < \bar{x} + z\sigma/\sqrt{n}|X).$$

This is all conditional on the data $X$, and hence the fixed value of $\bar{x}$ from the sample. The standardisation to $Z$ means simply that we can compute the interval probabilities from a known distribution, the standard normal.

- With non-reference priors, the results will be modified. Recall the reference prior is like a normal prior for $\theta$ with a very large variance.

We now compare this with elements of the traditional Sampling Theory approaches to point and interval estimation in this key example of normal mean estimation.

Prior to making the observations, the sample mean $\bar{x}$ is a random variable – we don’t know its value yet, but we do know that, from normal distribution theory, it has the sampling distribution

$$(\bar{x}|\theta) \sim N(\theta, \sigma^2/n)$$

where we are conditioning on the true value of $\theta$. When we see $X$, we will choose $\bar{x}$ as a point estimate of $\theta$ – classical statistics refers to $\bar{x}$ as an estimator of $\theta$ and is interested in making probability statements under the sampling distribution of the estimator $p(\bar{x}|\theta)$ to look at how “good” an estimator it is. Some examples:

**UNBIASED ESTIMATION.**

As $E(\bar{x}|\theta) = \theta$, we say that $\bar{x}$ is an unbiased estimator of $\theta$. Before seeing the data, we expect $\bar{x}$ to be at or close to $\theta$. The sampling distribution of the estimator is centered at the “correct” value.
MAXIMUM LIKELIHOOD ESTIMATORS.

Recall that, in the normal example here, $\bar{x}$ is also the MLE of $\theta$. In modern times, the standard sampling theory method of finding point estimators of parameters is to use the MLE (if it exists and can be found). Then the focus for inference is on features of the sampling distributions of MLEs themselves. In many other examples of simple random sampling models lead to MLEs of parameters that are functions of sample means, and are often unbiased estimates. For example, in a random sample of Poisson counts with mean $\mu$, $x_i \sim \text{Pois}(\mu)$, the sample mean $\bar{x}$ has a distribution that is not a standard form, but we know that it has mean and variance $E(\bar{x}|\mu) = \mu$ and $V(\bar{x}|\mu) = \mu/n$. So $\bar{x}$ is an unbiased estimator of $\mu$. Note also that the variance decreases towards zero as $n$ increases, meaning that $|\bar{x} - \mu|$ gets small for large $n$, as would be hoped; this relates to consistency of the estimator, and is a feature that parallels the convergence of posterior distributions for $\mu$ about $\mu$ for large $n$. (We are familiar with these ideas from the binomial sampling model.)

The concept of unbiasedness of point estimation is an old idea that has been used as a standard “criterion” for selecting point estimates. Note that it cannot be interpreted as any kind of Bayesian criterion as it is based on long-run, repeat use of the specific estimator in repeat experimentation. In Bayesian approaches posterior point estimates are simply summaries of the posterior distribution conditional on the observed data, and represent prior combined with likelihood, and the issue of properties in future experiments is not relevant. Also, with parameters in more than one dimension, unbiased estimators tend to be very poor in the frequency sense as well. Here we’ll note a simple example that indicates some potential shortcomings of the general concept from a practical viewpoint.

Example

- Recall the example of student drug use assessment in an early homework exercise. There the population proportion of drug users is $\theta$ and, through a sampling experiment designed to preserve confidentiality of responses, the observed number of students replying “Yes” to a question is binomial with probability $\mu = (\theta + 1)/2$. That is, $(y|\theta) \sim \text{Bin}(n, \mu)$ so that $E(y|\theta) = \mu = n(\theta + 1)/2$. Define the statistic $\hat{\theta} = 2(y/n) - 1$. This has some sampling distribution, with expectation $E(\hat{\theta}|\theta) = \theta$, so that $\hat{\theta}$ is an unbiased estimator of $\theta$. Now it is clear that this particular unbiased estimator may not be so good in any given experiment: suppose that $n = 100$ and we observed $y = 40$ “Yes” responses; this leads to $\hat{\theta} = 0.2$, and no-one wants a negative estimate of probability! So the concept is flawed. Note that, from that homework exercise, the MLE of $\theta$ corrects this flaw, and is $\hat{\theta} = \max\{0, \theta\}$. The MLE is not unbiased as an estimator of $\theta$, although the difference between $E(\hat{\theta}|\theta)$ and $\theta$ (this difference is the “bias”) becomes very small as $n$ gets large; this is related to the fact that MLEs are generally known to be asymptotically unbiased and consistent estimators.
CONFIDENCE INTERVALS.

In the normal model example, we might ask how close the (unbiased) estimate \( \bar{x} \) is likely to be to the true value \( \theta \). Since \( (\bar{x} | \theta) \sim N(\theta, \sigma^2/n) \) we can say that

\[
Pr(\theta - 1.96\sigma/\sqrt{n} < \bar{x} < \theta + 1.96\sigma/\sqrt{n}) = 0.95.
\] (2)

Similarly, if \( z > 0 \) is the upper 100(1 - p/2)% quantile of the standard normal distribution, the interval \( \theta \pm z\sigma/\sqrt{n} \) has probability \( 1 - p \) under the sampling distribution of \( \bar{x} \).

Now, the interval in (2) can be rewritten as

\[
Pr(\bar{x} - 1.96\sigma/\sqrt{n} < \theta < \bar{x} + 1.96\sigma/\sqrt{n}) = 0.95,
\] (3)

still remembering that this is a probability statement from \( p(\bar{x} | \theta) \). This gives us a 95% confidence interval for \( \theta \), namely the interval

\[
\bar{x} - 1.96\sigma/\sqrt{n} < \theta < \bar{x} + 1.96\sigma/\sqrt{n}
\] (4)

or \( \bar{x} \pm 1.96\sigma/\sqrt{n} \). Similarly, a 100(1 - p)% confidence interval is given by \( \bar{x} \pm z\sigma/\sqrt{n} \) if \( z > 0 \) is the upper 100(1 - p/2)% quantile of the standard normal distribution,

FREQUENCY INTERPRETATIONS.

In a future experiment with some true value of \( \theta \), the \( \bar{x} \) value observed is expected to be close to \( \theta \). In a long series of such future experiments, suppose you compute the interval (4) in each: about 95% of these random intervals will actually cover the true value of \( \theta \). This is the meaning and nature of the standard use of the term confidence: it refers to use of the interval in many future experiments – as a “procedure.” The theory says absolutely nothing more about the current experiment – sampling theory looks at the long-run, and does not provide statements conditional on the actually observed data. Similarly, unbiasedness of an estimator must be interpreted in the long-run frequency sense.

By contrast, Bayesian inference is conditional on the observed data and makes no reference to any hypothetical future experiments in inference on \( \theta \); it is concerned only with what the current data have to say about \( \theta \) in the current problem under study.

NUMERICAL EQUIVALENCES.

In some cases, reference Bayesian analyses can produce numerical results close or equal to those of standard sampling estimators and confidence intervals. This example is a case in point:

DISTRIBUTIONS FOR INFERENCE:
- reference posterior \( p(\theta | X) = N(\bar{x}, \sigma^2/n) \) so that \( Z = (\theta - \bar{x})\sqrt{n}/\sigma \) is \( N(0, 1) \)
- sampling distribution \( p(\bar{x} | \theta) = N(\theta, \sigma^2/n) \) so that \( Z' = (\bar{x} - \theta)\sqrt{n}/\sigma \) is \( N(0, 1) \)

POINT ESTIMATE: \( \bar{x} \) is
- unbiased posterior mode, mean, median (and MLE)
- unbiased for \( \theta \)

INTERVALS: For small \( p \) (such as 0.05) let \( z > 0 \) be the upper 100(1 - p/2)% quantile of the standard normal distribution. Then the 100(1 - p)% intervals are \( \bar{x} \pm z\sigma/\sqrt{n} \), either
- reference posterior central probability interval
- confidence interval

The Bayesian and frequentist concepts and interpretations are very different. The numerical results here are, however, the same – as is true in other models with reference Bayesian analyses. We saw this exemplified in binomial models. However, the numerical results can be very different in other cases and in non-reference Bayesian analyses.
Example 2: Normal Model with Unknown Variance

When the normal variance $\sigma^2$ is unknown as well as the mean $\theta$, the reference Bayesian posterior for $\theta$ is

$$ (\theta|X) \sim t_{n-1}(\bar{x}, s^2/n) \quad \text{or} \quad T = (\theta - \bar{x})\sqrt{n}/s \sim t_{n-1}(0, 1), $$

where $t_{n-1}$ is the Student t distribution on $n - 1$ degrees of freedom. We note also $\theta = \bar{x} + Ts/s\sqrt{n}$.

The use of these results parallel the normal model above but with
- the normal variance $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n - 1)$ estimating $\sigma^2$,
- the normal posterior replaced by the slightly fatter-tailed $t_{n-1}$, and
- the standard Student t random quantity $T$ replacing the standard normal $Z$.

The posterior mean, median, mode is still $\bar{x}$, and the $100(1 - p)$% symmetric posterior central interval for $\theta$ is

$$ \bar{x} \pm ts/\sqrt{n} \quad (5) $$

where $t > 0$ is the upper $100(1 - p/2)$% quantile of the standard $t_{n-1}$ distribution.

**SAMPLING THEORY RESULTS (DeGroot 7.5)**

Under the sampling model $(x_i, \theta, \sigma^2) \sim N(\theta, \sigma^2)$ define the statistic

$$ T = (\bar{x} - \theta)\sqrt{n}/s. $$

Here $(\theta, \sigma^2)$ are fixed at their true values and we are interested in the variation in $T$ induced by variation in $(\bar{x}, s^2)$ under repeated sampling - the frequency sampling distribution of $T$ under the model, or $p(T|\theta, \sigma^2)$. In general, such distributions depend on the values of the parameters conditioned on; it turns out that in this example the distribution of $T$ does not involve the parameters, and is $T \sim t_{n-1}(0, 1)$.

Hence we can construct confidence intervals as in the normal example above. If $t > 0$ is the upper $100(1 - p/2)$% quantile of the standard $t_{n-1}$ distribution,

$$ p = Pr(-t < T < t|\theta, \sigma^2) $$

and the formula for $T$ means we can rewrite this as

$$ p = Pr(\bar{x} - ts/\sqrt{n} < \theta < \bar{x} + ts/\sqrt{n}|\theta, \sigma^2), $$

so that $\bar{x} \pm ts/\sqrt{n}$ is a $100(1 - p)$% confidence interval for $\theta$; in a long run of future experiments of this kind, the intervals constructed this way will cover the true value of $\theta$ about 95% of the time. Again, notice that the intervals are numerically equivalent to the reference Bayesian intervals (5).

Example 3: Linear Regression Models

The same equivalences arise in linear models. We often use reference Bayesian posterior $t$ distributions for inference on linear model regression parameters. The posterior intervals constructed from these are exactly the same, numerically, as those arising from the frequentists sampling theory for the sampling distributions of least squares estimates of regression parameter vectors. Full details are in DeGroot (10.2, 10.3, 10.5). The point here is to note the equivalences - from an applied perspective, the distributional results agree in forming point estimates and interval estimates.