7 Inference

Statistical inference is the process of learning from data. The aim is to use information from a sample to draw conclusions about the population from which the sample was taken. Typically the sample is relatively small compared to the much larger (effectively infinite) population. We talk about making *inferences* from the sample, and often want to quantify the accuracy or reliability of the inferences.

There are two main schools of thought regarding the appropriate framework for inference: frequentist and Bayesian. One of the fundamental differences between the two schools is their interpretation of probability, with frequentists taking the "limiting relative frequency" viewpoint and Bayesians taking the subjective viewpoint, as discussed in Section 4.3. In this introductory course we adopt the more traditional frequentist approach and consider:

- Estimation:
  - Point estimation;
  - Maximum likehood;
  - Confidence intervals;
- Hypothesis testing.

Example:
Suppose we are interested in the time it takes bacteria to divide, and we model this with an exponential random variable with probability density function (pdf)

$$f(x) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

Given a random sample of observations we need to estimate the value of θ which allows our model to best describe the data. We might then be able to compare different strains of bacteria, or bacteria grown under different laboratory conditions. Suppose we estimate θ_1 and θ_2 for two different strains. *Statistical inference* would be used to address the following typical questions:

1. How accurate are the estimates θ_1, θ_2? They won't be perfect since both are based on limited noisy samples.
2. How plausible is it that the two strains grow at the same rate $\theta_1 = \theta_2$?
3. Are the data really consistent with our exponential model?

7.1 Sampling

In order to learn about a population from a sample it is very important that the sample is obtained in such a way as to be *representative* of the population. To illustrate, consider a trial in which we are investigating the effect of a new treatment on reducing blood pressure

in patients with hypertension. If our results are to be generalizable to the population of all hypertension patients, it is important that the individuals recruited to the study are representative of this entire population. For example, if we estimate the treatment effect using only women (or some other sub-population) we might introduce *bias* into our estimate. The question of how to construct samples in the best way possible forms an entire branch of statistics (called experimental design). Very often, it can be achieved by taking a *random sample*, generated by a mechanism by which all possible samples are *equally likely* to be chosen.

7.2 Parameter estimation

Many problems in statistical inference have the following form.

- Suppose we model measurements of an experimental quantity using a random variable X and specify that X comes from one of the standard distributions we have seen on this course, e.g. Normal, Poisson etc.
- Usually the distribution of X will depend on one or more *parameters*, for example the rate λ of a Poisson distribution.
- We might then consider repeated measurements as being independent identically distributed realisations of X.
- Given a set of observations $x_1, x_2, \ldots, x_n$ we want to obtain an *estimate* of the model parameters that *best fit* the data.

This represents the simplest kind of 'model' for a set of measurements. More complicated models might have dependence between observations, quantities drawn from several different (possibly dependent) distributions, and many underlying model parameters. However, the fundamental question of how best to estimate the model parameters remains the same.

7.2.1 Estimators

Suppose we have a model for measurements $X_1, \ldots, X_n$ that is parameterized by some parameter θ. An *estimator* for θ is a random variable T that is defined as some function $g(\cdot)$ of the measurements

$$T = g(X_1, \ldots, X_n)$$

such that

$$\theta \approx g(x_1, \ldots, x_n)$$

for observations $x_1, \ldots, x_n$. The distribution of T is called its *sampling distribution*. We have been deliberately flexible about the nature of the approximation in the equation above – much of this section will deal with choosing 'good' approximations.

Example:
Suppose $X_1, X_2, \ldots, X_n$ are IID with common mean μ and variance σ^2.

1. In the previous section we looked at the sample mean:

$$T = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

   We saw that the sample mean was a good estimator for the population mean μ, in that it gives the correct answer *on average*, i.e. over repeated random samples. Moreover, the Central Limit Theorem described the distribution of the sample mean around the population mean, enabling us to assign a degree of uncertainty to our estimates.

2. Instead of the sample mean, we might consider

$$T_1 = X_1.$$

   In other words our estimate of μ is just the first observation we make.

3. Similarly, we might consider

$$T_2 = \frac{1}{n}X_n$$

   the nth observation divided by n.

These last two estimators, T_1 and T_2, look strange. In what way are they any better or worse than the sample mean? To answer that question we need to consider the expectation and variance of our estimators.

We know that

$$E[T] = E[\bar{X}] = \mu, \qquad \text{and} \qquad \mathrm{Var}(T) = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Now consider $T_1 = X_1$. We have that

$$E[T_1] = E[X_1] = \mu$$

so just like $\bar{X}$ we get the right value on average. But what about the variability? We have that

$$\mathrm{Var}(T_1) = \mathrm{Var}(X_1) = \sigma^2,$$

and so clearly $\mathrm{Var}(T_1) \geq \mathrm{Var}(T)$ (for $n = 1, 2, \ldots$). In other words, using T_1 to estimate the mean will always lead to a more variable (less precise) estimator. So we would always choose $\bar{X}$ to estimate the mean rather than just taking the first observation.

What about T_2? Its variance is

$$\mathrm{Var}(T_2) = \mathrm{Var}\left(\frac{X_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X_n) = \frac{\sigma^2}{n^2},$$

which will never be greater than the variance of $\bar{X}$, since

$$\frac{\sigma^2}{n^2} \leq \frac{\sigma^2}{n}, \quad n = 1, 2, \ldots$$

So, does that mean that T_2 is a better estimator of μ than $\bar{X}$? Well, not necessarily, because if we consider the expectation of T_2 we get

$$E[T_2] = E\left[\frac{X_n}{n}\right] = \frac{1}{n}E[X_n] = \frac{\mu}{n},$$

which is only equal to μ when $n = 1$. And so whilst it is a more precise estimator than the sample mean $\bar{X}$ it is also less accurate since $E[T_2] \neq \mu$ in general.

7.2.2 Biased and unbiased estimators

An estimator T of an unknown parameter θ is *unbiased* if

$$\mathrm{E}[T] = \theta$$

i.e. if on average it gives the correct answer. In the example above, $T = \bar{X}$ and $T_1 = X_1$ are unbiased estimators of μ but $T_2 = X_n/n$ is biased (when $n > 1$).

Given two or more unbiased estimators for θ it makes sense to choose the one with the smallest variance. We define the most *efficient* estimator to be the estimator with the smallest variance. For example, both $T = \bar{X}$ and $T_1 = X_1$ are unbiased estimators of μ but $\mathrm{Var}(\bar{X}) \leq \mathrm{Var}(X_1)$ and so $\bar{X}$ is preferred. (It turns out that out of all unbiased estimators for μ, $\bar{X}$ has the smallest variance.)

While in general it is often desirable to have unbiased estimators, these are by no means always the *best*.

7.2.3 Consistency

Suppose we have an estimator T_n for a parameter θ defined in terms of measurements $X_1, \ldots, X_n$ as $n \to \infty$. The estimator T_n is *consistent* if

$$\mathrm{E}[T_n] \to \theta \quad \text{and} \quad \mathrm{Var}\,(T_n) \to 0$$

as $n \to \infty$. In other words, the estimator is consistent if larger samples can be expected to give more precise estimates.

For example, $\bar{X}$ is a consistent estimator of μ, since $E[\bar{X}] = \mu$ for all n, and $\mathrm{Var}\left(\bar{X}\right) = \sigma^2/n \to 0$ as $n \to \infty$. Neither T_1 nor T_2 are consistent estimators of μ. Why?

7.2.4 Examples

Example 7.1 (Estimating the population variance)**:**
We have already seen that if $X_1, \ldots, X_n$ are IID with mean μ and variance σ^2, then the sample mean $\bar{X}$ is a consistent estimator for μ. Therefore, given observations $x_1, x_2, \ldots, x_n$ we estimate μ by the sample mean $\bar{x}$. It should therefore be no surprise that the sample variance s^2 provides a good estimate for σ^2.

Recall that the sample variance is defined as

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

so the corresponding *estimator* is

$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2.$$

It can be shown that S^2 is an unbiased (and moreover a consistent) estimator of σ^2.

Example 7.2 (Estimating the upper limit of a uniform random variable)**:**
Suppose $X_1, X_2, \ldots, X_n$ are IID Unif$(0, b)$ random variables, for some unknown value of b. Consider the following estimators for b:

$$T = \max(X_1, \ldots, X_n) \quad \text{and} \quad U = 2\bar{X}.$$

Which do you think is the better estimator?

It turns out that the estimator T is biased with $E[T] = nb/(n+1)$ while U is unbiased. Both T and U are consistent estimators of b. However, it can be shown that T has the smaller variance:

$$\text{Var}\,(T) = \frac{nb^2}{(n+2)(n+1)^2}, \quad \text{Var}\,(U) = \frac{b^2}{3n}.$$

Example 7.3 (Estimating both limits of a uniform random variable)**:**
Suppose $X_1, X_2, \ldots, X_n$ are IID Unif(a, b) random variables where both a and b are unknown. The previous example can be adapted to this new case. Two estimators are:

$$T_a = \min(X_1, \ldots, X_n) \quad \text{and } T_b = \max(X_1, \ldots, X_n).$$

Use the expectation and variance of the uniform distribution to find estimators for a and b analogous to U above.

This is an example of the *method of moments* – a simple but commonly used way to construct estimators. It works in the following way:

1. Suppose there are two parameters to estimate. Express the expectation and variance of the measurements X in terms of the parameters.
2. Obtain two equations by equating the expectation and variance of X to the sample mean and sample variance.
3. Solve the equations for the model parameters.

This method can easily be generalised to allow an arbitrary number of parameters.