Inference notes

The big picture

Inference is all about *inferring* things from data. The general set-up in statistics is you observe a *sample* from some *population*. The population has a distribution which depends on some population parameters, inference is the process of making statements about the population based on the data observed in your sample. Any statistic you compute on a sample is a random variable because there is randomness in the sample - if you had a different sample, you would compute a different statistic. The distribution of the statistic is called the *sampling distribution*. If you can figure out the sampling distribution, then you can make probabilistic statements about the parameter of interest, By the central limit theorem, the sample distribution will (in most cases) be related to the population parameters in some way, so information about the sampling distribution will give you information about the underling population parameter.

That's all a bit abstract, so here's an example. Suppose you're trying to figure out the mean height of people in England. A sensible distribution for height would be would be $N(\mu, \sigma^2)$ - the problem is that you don't know μ so you want to estimate it. You go out and get a random sample of people, measure their heights and take the mean. Each sample you take has a different sample mean, so the sample mean is a random variable. The left plot shows some example samples, and the right plot shows the (empirical) sampling distribution for the mean. Notice that the sampling distribution - the distribution of the sample mean - is centered on the population mean. The sampling distribution contains information about the population parameter:

Some samples (left) & sampling distribution (right) of heights

Assuming heights are normal with mean = 175, sd = 5

Sample 1 Avg height = 173.5

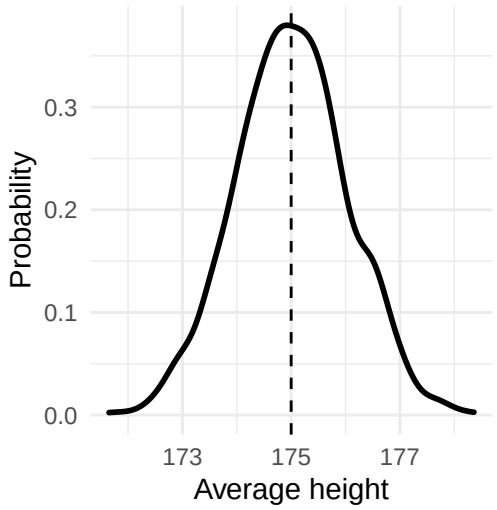
170	179	171	180	178	188	177	173	173	174	167	170
173	174	170	167	171	171	166	174	166	180	179	170

Sample 2 Avg height = 175.1

175	175	177	175	173	170	181	171	182	175	173	168
184	173	174	180	164	170	179	178	176	172	177	178

Sample 3 Avg height = 175.4

177	170	176	179	189	171	179	175	170	167	179	177
177	170	174	175	174	180	173	176	167	180	175	180



Properties of estimators

In the example from the last section, we talked about *estimating* the population mean. We did this using the sample mean. This is a very natural estimate to use, and in this specific example it all worked out. Let's formalise the process a bit.

Inference starts with you picking an *estimand* - the quantity you want to *estimate*. The process of computing the estimand is known as an *estimator*. Once you have some data, you can follow the instructions of the estimator to form an *estimate*. In the earlier example, the estimand was the population mean. The estimator was the sample mean. Once you have a sample you can compute the sample mean, that number is the estimate. The only random variable in this framework is the estimator. The estimand is usually a population parameter (which is fixed so isn't random), and the estimate is a specific realisation of the estimator after you put the sample data into it.

There are some *properties of estimators* which we might want, for example we might want an estimator $\hat{\theta}$ to be:

- *Consistent* - As the sample size increases, the estimator converges to the true parameter values ($lim_{n\to\infty}\hat{\theta} = \theta$)
- *Unbiased* - An estimator $\hat{\theta}$ has bias b if $E\hat{\theta} = \theta + b$. If $b = 0$ then the estimator is unbiased. This is essentially a weaker form of consistency, it says that on average the estimator takes on the parameter value
- *Efficient* - the estimator has small variance. A good estimator will only take on a small number of values (ideally including θ), so you'd usually choose the estimator with the smallest variance

Every consistent estimator is asymptotically efficient, since as $n \to \infty$ the estimator converges on the parameter, so only takes 1 value.

Calculating bias, efficiency, and consistency for estimators requires working with expectations and variances. Here's a couple of useful results which will come in handy when calculating properties of estimators. First is the *law of iterated expectations* (proof in probability notes):

$$E[EY|X] = EY$$

This says that the average of a variable is a weighted average of within-group averages. For example, suppose you knew the average height of left handed people ($X = 1$) and the average height of right handed people ($X = 0$). The law of iterated expectations says

$$\begin{aligned} EY &= E[Y|X=1]P(X=1) + E[Y|X=0]P(X=0) \\ \text{Avg height} &= \text{Avg height}_{Left} \times \%_{Left} + \text{Avg height}_{Right} \times \%_{Right} \end{aligned}$$

Which makes sense. Applying the law of iterated expectations to variance gives the *partition of variance* formula:

$$\begin{aligned} VY &= EY^2 - E[Y]^2 \\ &= E[EY^2|X] - E[EY|X]^2 \\ &= E[V[Y|X] + E[Y|X]^2] - E[EY|X]^2 \\ &= E[V[Y|X]] + (E[E[Y|X]^2] - E[EY|X]^2) \\ &= E[V[Y|X]] + V[E[Y|X]] \end{aligned}$$

Or, in words, if you have some variable Y which is split into groups X (think of left/right handed again):

$$\text{Variance in } Y = [\text{weighted average of within-group variances}] + [\text{variation between group means}]$$

Which also makes sense.

As an example, let's say $X \sim N(\mu, \sigma^2)$, you get a sample of size n, and you want to estimate the population mean μ and variance σ^2. Two intuitive estimators are $\hat{\mu} = \frac{1}{n}\sum_i x_i$ and $\widehat{\sigma^2} = \frac{1}{n}\sum_i (x_i - \hat{\mu})^2$. $\hat{\mu}$ is unbiased (and therefore consistent):

$$E\hat{\mu} = E\left[\frac{1}{n}\sum_i x_i\right] = \frac{1}{n}\sum_i Ex_i = \frac{1}{n}n\mu = \mu$$

But $\widehat{\sigma^2}$ is biased. To see this use the partition of variance formula to split the variance in X into 2 terms:

$$\begin{aligned} VX &= V[X|\hat{\mu}] + (\hat{\mu} - \mu)^2 \\ VX &= \widehat{\sigma^2} + (\hat{\mu} - \mu)^2 \end{aligned}$$

Rearranging & taking expectations:

$$\begin{aligned} E\left[\widehat{\sigma^2}\right] &= E[VX] - E(\hat{\mu} - \mu)^2 \\ &= \sigma^2 - V\hat{\mu} \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

The variance estimate is biased (although for large n the bias is negligible). An unbiased estimate for the variance would be

$$\frac{n}{n-1}\widehat{\sigma^2} = \frac{n}{n-1}\frac{1}{n}\sum_i (x_i - \hat{\mu})^2 = \frac{1}{n-1}\sum_i (x_i - \hat{\mu})^2$$

You need to divide by $n-1$ because you used 1 *degree of freedom* in estimating the sample mean $\hat{\mu}$. Roughly speaking, you use up 1 degree of freedom for each parameter you estimate from your data. You started off with n degrees of freedom (because you have n data points in your sample), then you lose one when you estimate the sample mean, leaving you with $n-1$ remaining.

Maximum likelihood estimators

All the estimators so far have been intuitive - you guessed that the sample mean is probably a good estimator for a population mean, and the sample variance is probably a good estimator for the population variance. Then you calculated expectations to see what properties these estimators have. *Maximum likelihood* is a method for constructing estimators & estimates based on observed data. The idea behind maximum likelihood is straightforward:

Suppose you have a sample D of some variable X, and the distribution of X has some parameters - which we'll write as $f(x;\theta)$. Once you observe your data D, the (joint) probability of the data is $f(D;\theta)$. The only variable here is θ, so you can think of the joint as a function of θ:

$$L(\theta) = f(D;\theta)$$

L is the *likelihood*. For each value of θ, $L(\theta)$ is the probability of observing data D for that specific parameter value θ. The *maximum likelihood estimator* is the value of θ which maximises the likelihood - i.e. it is the parameter value which is most consistent with the observed data. Finding the maximum is just a case of differentiation. In a lot of cases your sample will be IID, so the joint factors. Because of this it's usually easier to maximise the log likelihood instead of the likelihood - either way you'll still get the same estimator, because log is a monotonically increasing function. The maximum likelihood estimator is such that

$$\frac{\partial}{\partial\theta}L\bigg|_{\hat{\theta}_{ML}} = 0 \qquad \text{or equivalently} \qquad \frac{\partial}{\partial\theta}\ln L\bigg|_{\hat{\theta}_{ML}} = 0$$

For example, the log likelihood for n IID observations of a normal $N(\mu,\sigma^2)$ is

$$\ln L = -n\ln(\sqrt{2\pi}\,\sigma) - \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

Differentiating with respect to μ and setting equal to zero gives

$$0 = \sum_{i=1}^{n}(x_i - \hat{\mu}_{ML}) = \sum_{i=1}^{n}x_i - n\hat{\mu}_{ML} \quad \Longrightarrow \quad \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

The second derivative is $-2n$ which is always negative, so $\hat{\mu}_{ML}$ maximises the likelihood. The maximum likelihood estimator is exactly the one we thought of intuitively! This is nice because maximum likelihood is a general principle, if you have to construct an estimator in a new setting where you have no intuition about what a good estimator may be, you can *always* maximise the likelihood to get an estimator.

Under some *regularity conditions*, the following nice (asymptotic) properties of maximum likelihood estimators hold. as $n \to \infty$:

- $L(\theta)$ becomes normal, centered at the maximum likelihood estimate. This means that the log likelihood becomes a negative parabola
- the MLE is consistent, unbiased, and efficient. It has a normal distribution, centered at the population parameter
- The variance of the MLE approaches the *minimum variance bound* - more on this in the next section…
- MLEs are invariant to transforms - if $\hat{\theta}_{ML}$ is a MLE for θ, then $\ln\hat{\theta}_{ML}$ is a MLE for $\ln\theta$. The variance doesn't transform nicely however, so you'll need to calculate it

The regularity conditions are very reasonable, so these properties hold in a range of cases. The regularity conditions are:

- $L(\theta)$ must be differentiable and have a global maximum (same for $\ln L(\theta)$)
- The boundaries of data space must not depend on θ, and the maximum can't occur at the boundaries of the data space

Basically, as long as the likelihood isn't something crazy, maximum likelihood will work.

Just a quick note on variance of transformed MLEs before moving on. You can either calculate the variance using the MVB form the next section, or you can use the *delta method*. Taking a Taylor expansion of $g(\theta)$ about $\hat{\theta}_{ML}$ to first order and rearranging gives

$$g(\theta) - g\left(\hat{\theta}_{ML}\right) \approx g'\left(\hat{\theta}_{ML}\right)\left(\theta - \hat{\theta}_{ML}\right) \implies g(\theta) \sim N\left(g\left(\hat{\theta}_{ML}\right), \left[g'\left(\hat{\theta}_{ML}\right)\right]^2 \sigma^2_{ML}\right)$$

Which follows because $\hat{\theta}_{ML} \sim N(\theta, \sigma^2_{ML})$. So the transformed variance is a scaled version of the MLE variance.

The minimum variance bound (MVB)

For any estimator $\hat{\theta}$ with bias b, the following inequality always holds:

$$V[\hat{\theta}] \geq \frac{-\left(1 + \dfrac{\partial b}{\partial \theta}\right)^2}{E\left[\dfrac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

This is the *minimum variance bound*. It's a very good inequality, and in most cases you can take it to be an approximate equality. Let's prove it:

The first thing to note is that the likelihood is a probability of data, so integrating over all possible observed data will be 1

$$\int dD\, L = 1$$

And - if the boundaries of data space don't depend on the parameter - we can differentiate this to get

$$\int dD\, \frac{\partial L}{\partial \theta} = 0$$

using the chain rule, we also have $\frac{\partial \ln L}{\partial \theta} = \frac{1}{L}\frac{\partial L}{\partial \theta}$. So we can rewrite the previous equation as

$$\int dD\, \frac{\partial \ln L}{\partial \theta} L = E\left[\frac{\partial \ln L}{\partial \theta}\right] = 0$$

Since the estimator has bias b we know

$$E\hat{\theta} = \int dD\, \hat{\theta} L = \theta + b$$

Differentiating and writing the derivative in terms of the log likelihood gives

$$\int dD\, \hat{\theta} \frac{\partial \ln L}{\partial \theta} L = 1 + \frac{\partial b}{\partial \theta}$$

There is no term involving $\frac{\partial \hat{\theta}}{\partial \theta}$, because an estimator only depends on the data (not the parameter in the model which generates the data). Now add zero in a clever way to the left hand side - add on $\theta E\left[\frac{\partial \ln L}{\partial \theta}\right]$ to get

$$\int dD\, \left(\hat{\theta} - \theta\right) \frac{\partial \ln L}{\partial \theta} L = 1 + \frac{\partial b}{\partial \theta}$$

Integrals are an example of an inner product, so the Cauchy Schwartz inequality applies - $\langle u, v\rangle^2 \leq \langle u, u\rangle\langle v, v\rangle$ (to prove this look at $\langle u + \lambda v, u + \lambda v\rangle$. It has to be ≥ 0, so the resulting polynomial in λ must have discriminant ≤ 0). Take $u = \left(\hat{\theta} - \theta\right)\sqrt{L}$, $v = \frac{\partial \ln L}{\partial \theta}\sqrt{L}$, then

$$\left[1 + \frac{\partial b}{\partial \theta}\right]^2 = \left[\int dD\, \left(\hat{\theta} - \theta\right) \frac{\partial \ln L}{\partial \theta} L\right]^2 \leq \left[\int dD\, \left(\hat{\theta} - \theta\right)^2 L\right] \left[\int dD\, \left(\frac{\partial \ln L}{\partial \theta}\right)^2 L\right]$$

Now we just have to simplify. Differentiating the expectation of the derivative of the log likelihood gives

$$0 = \frac{\partial}{\partial \theta} E\left[\frac{\partial \ln L}{\partial \theta}\right] = \int dD\, \frac{\partial^2 \ln L}{\partial \theta^2} L + \int dD\, \left(\frac{\partial \ln L}{\partial \theta}\right)^2 L$$

So, rewriting the integrals as expectations

$$E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = -E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]$$

Which gives the inequality we were trying to prove

$$\left[1 + \frac{\partial b}{\partial \theta}\right]^2 \leq V\hat{\theta}\left(-E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]\right) \qquad \blacksquare$$

The MVB is particularly useful for maximum likelihood estimates because, in the region around the MLE, the second derivative is roughly constant - so you can replace the expectation with the second derivative evaluated $\hat{\theta}_{ML}$. Also the bias of an MLE goes to zero as the sample size increases, further simplifying the formula to

$$V\hat{\theta}_{ML} \approx \frac{-1}{\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\hat{\theta}_{ML}}}$$

Log likelihood ratios

The one slight negative about log likelihoods is that they're hard to interpret - what does a log likelihood of -4.5910 mean for example? To make it more interpretable, subtract off the maximum value

$$LLR(\theta) = \ln L(\theta) - \ln L\left(\hat{\theta}_{ML}\right) = \ln\left(\frac{L(\theta)}{L\left(\hat{\theta}_{ML}\right)}\right)$$

This is called the *log likelihood ratio*. It has a maximum value of zero, which makes the numbers slightly more interpretable. It also shows up naturally through the Taylor expansion of the log likelihood

$$\ln L(\theta) \approx \ln L\left(\hat{\theta}_{ML}\right) + \left.\frac{\partial \ln L}{\partial \theta}\right|_{\hat{\theta}_{ML}} \left(\theta - \hat{\theta}_{ML}\right) + \frac{1}{2}\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\hat{\theta}_{ML}} \left(\theta - \hat{\theta}_{ML}\right)^2$$

The first derivative is zero, and the second derivative we can get from the MVB (writing $\sigma_{ML} = \sqrt{V\left[\hat{\theta}\right]}$)

$$LLR(\theta) \approx -\frac{1}{2}\left(\frac{\theta - \hat{\theta}_{ML}}{\sigma_{ML}}\right)^2 \approx -\frac{1}{2}\left(\frac{\hat{\theta}_{ML} - \theta}{\sigma_{ML}}\right)^2$$

This is the *quadratic approximation* to the log likelihood ratio. Since the term in brackets is squared, the order of the numerator doesn't matter. Looking back at the asymptotics of the maximum likelihood estimator, $\hat{\theta}_{ML} \sim N(\theta, \sigma^2_{ML})$. so the term inside the brackets is distributed standard normal, and its square is distributed chi squared with 1 degree of freedom (proof in probability notes). So

$$-2LLR(\theta) \sim \chi^2_1$$

This relies on the fact that the distribution of $\hat{\theta}_{ML}$ is asymptotically normal, so it requires that the regularity conditions hold and that the sample size is infinite. Since you never have an infinite sample, you'll never get to the asymptotic case. As such, this is called a *normal approximation* to the log likelihood ratio. If there were n parameters to estimate then this would become $-2LLR(\theta) \sim \chi^2_n$. This can be used to calculate confidence intervals for the log likelihood ratio, by finding the values of θ such that $-2LLR(\theta) \leq \chi^2_{1,0.95} = 3.84$. This gives 95% confidence interval boundaries given by $LLR(\theta) = -1.92$, which can be read off the graph of the log likelihood ratio. These can be inverted to get intervals for $\hat{\theta}_{ML}$ - the boundaries of the confidence interval are solutions to the equation

$$\left(\frac{\theta - \hat{\theta}_{ML}}{\sigma_{ML}}\right)^2 = \chi^2_{1,0.95}$$

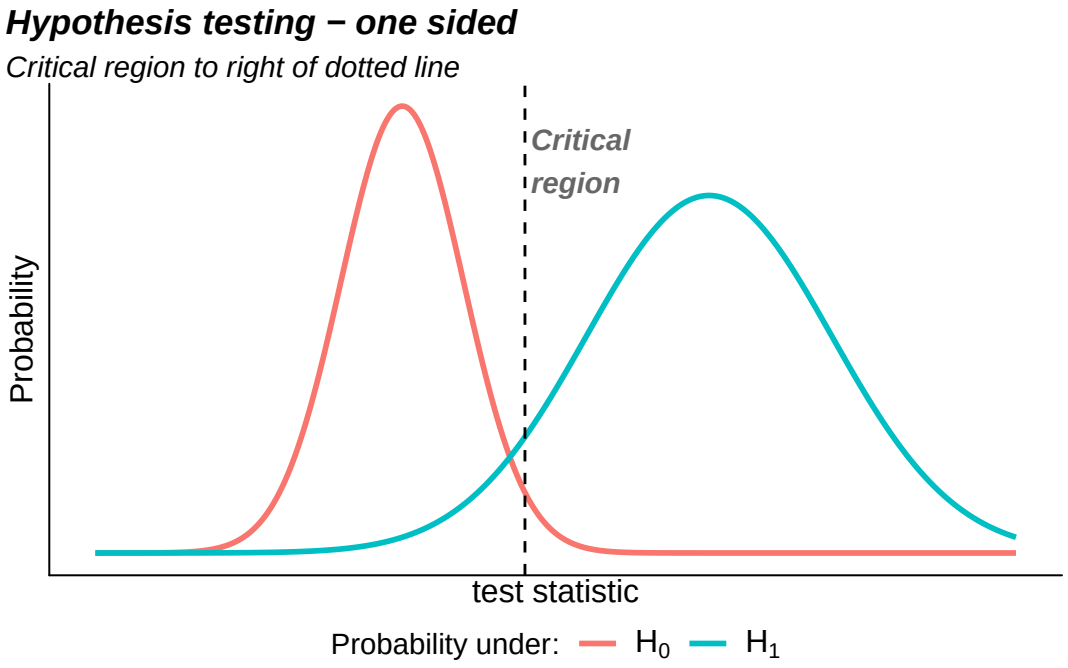
Solving for θ gives the 95% confidence interval boundaries as

$$\theta = \hat{\theta}_{ML} \pm \sqrt{\chi^2_{1,0.95}}\,\sigma_{ML} = \hat{\theta}_{ML} \pm 1.96\,\sigma_{ML}$$

You can either calculate σ_{ML} using the MVB, or you can estimate it from the graph of the log likelihood ratio - just find the values of θ such that $LLR(\theta) = -\frac{1}{2}$ (i.e. $\theta - \hat{\theta}_{ML} = \pm\sigma_{ML}$).

Hypothesis tests

A *hypothesis* is a True/False statement about parameters. They can be simple, or composite (where the hypothesis itself has parameters). For example $\theta = 2$ is a simple hypothesis, and $\theta > 2$ is a composite hypothesis. Hypotheses give the probability of data - if you have a hypothesis H_0 then you can talk about $P(D|H_0)$, the probability of observing data D under the hypothesis H_0. A *hypothesis test* is a procedure for rejecting a hypothesis in favour of another hypothesis. Typically, hypothesis testing involves 2 hypotheses - H_0, usually called the *null hypothesis*, and H_1, usually called the *alternative hypothesis*. You create a *test statistic* τ (more on how to do this later), and define a *critical region* W - a region of τ space where the probability of observing $\tau \in W$ under H_0 is small. You get some data and calculate the test statistic, if $\tau \in W$ then you reject H_0. A rough schematic would look something like this



The important part here is the critical region. Once you define the critical region W, the possible probabilities of observing a test statistic in W are

$$P(\tau \in W|H_0) = \alpha \qquad \text{The } \textit{size} \text{ of the test}$$
$$P(\tau \in W|H_1) = \beta \qquad \text{The } \textit{power} \text{ of the test}$$

The size of the test is is usually fixed ($\alpha = 0.05$ is pretty common), and then the critical region is chosen such that the power is maximised. In the picture above α is the area under the red curve to the right of the dotted line, and β is the area under the blue curve to the right of the dotted line.

Minimising α and maximising β reduce the chance of errors. There are two ways to be wrong when hypothesis testing:

- You could incorrectly reject the null. This is a false positive (aka *type 1 error*) - you incorrectly detected a signal which doesn't exist. It would happen if your test statistic falls in the critical region when the null is true, the probability of this happening is $P(\tau \in W|H_0) = \alpha$ by definition
- You could fail to reject the null. This is a false negative - you did not detect the signal. It would happen if your test statistic doesn't fall in the critical region when the alternative is true, the probability of this happening is $1 - P(\tau \in W|H_1) = 1 - \beta$