Bayesian Statistical Inference

An example of statistical inference is as follows.

Example. Consider a production process where the overall proportion of defectives, θ, is of interest. A random sample of size n of products from this process is checked for defectives. Let X denote the number of defectives found in the sample. Then $X \sim \text{Binomial}(n, \theta)$. i.e.,

$$P(X = x|\theta) = f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x = 0, 1, \ldots, n.$$

The unknown quantity θ indexes the model P_θ for X. What is the 'best fit' for θ if $X = x$ is observed? We may find the mle of θ: $l(\theta|x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, as a function of θ for given x is the likelihood function of θ. This is a measure of how likely that the model with proportion θ produced the data x. With this interpretation, it makes sense to maximize this likelihood function to estimate θ.
$\hat{\theta}_{\text{mle}} : \max_\theta l(\theta|x)$

In the example, $l(\theta|x) = c(x)\theta^x(1-\theta)^{n-x}$ has unique maximum at $\hat{\theta} = x/n =$ sample proportion of defectives. Good! Now we have an estimate (for θ). How good is this estimate? What is the estimation error? What is a confidence interval for θ?

These questions cannot be answered by the likelihood approach. In the *Frequentist* approach, they require the sampling distribution of the estimator – on repeated sampling how does $\hat{\theta}$ behave? Let us consider confidence statements.

Confidence Set. Any random set (related to X) which captures θ with a prescribed level of confidence.

If n is large, a 95% confidence interval for θ is $\hat{\theta} \pm 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}$. This is because, since $X \sim \text{Binomial}(n, \theta)$, for large n, X/n is approximately $N(\theta, \theta(1-\theta)/n)$, or

$$\frac{X/n - \theta}{\sqrt{\theta(1-\theta)/n}} \sim N(0, 1).$$

Then, approximately,

$$P\left(|\frac{X}{n} - \theta| \leq 1.96\sqrt{\theta(1-\theta)/n}\right) = 0.95.$$

For large n, $\hat{\theta}(X)$ is close to θ, so

$$P\left(|\frac{X}{n} - \theta| \leq 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}\right) = 0.95.$$

Therefore,

$$\theta \in \left(\hat{\theta}(X) \pm 1.96\sqrt{\hat{\theta}(X)(1-\hat{\theta}(X))/n}\right)$$

with probability 0.95 for all θ.

There are two issues with this approach. First, instead of binomial sampling, suppose we did inverse binomial sampling. i.e., check products until x (same count as what we got with binomial sampling) defectives are spotted. Then, we have:
binomial likelihood: $\theta^x(1-\theta)^{n-x}$
inverse binomial likelihood: $\theta^x(1-\theta)^y$

Suppose $n - x = y$; then the observed likelihood is the same for both the models, so $\hat{\theta} = x/n$ for both. However, the confidence intervals will be different since the variances of $\hat{\theta}$ will be different.

The second issue involves the interpretation of the confidence interval. If we sample again and again from the production process and construct 95% confidence intervals with each of the samples, in about 19 cases out of 20 the intervals will contain θ. However, for the given sample we get a fixed (not random) interval: $\hat{\theta}(x) \pm 1.96\sqrt{\hat{\theta}(x)(1-\hat{\theta}(x))/n}$. What is the interpretation for this interval? Surely, θ can only lie in that interval with probability either 0 or 1.

Classical statistics is frequentist – talks only in terms of optimality w.r.t. long-run average behaviour of statistical procedures. It cannot condition on data, and cannot interpret procedures with respect to fixed data. Why is conditioning needed if we have procedures which have good long-run behaviour?

The need for conditioning on data.

First of all, repetition of experiments (as in frequentist sense) may not be meaningful – what are the chances of another catastrophe like covid-19? Another point is illustrated below.

Example. Let X_1 and X_2 be i.i.d. with

$$X_i = \begin{cases} \theta - 1 & \text{with probability } 1/2; \\ \theta + 1 & \text{with probability } 1/2, \end{cases}$$

where $-\infty < \theta < \infty$. Define a confidence set for θ as follows.

$$C(X_1, X_2) = \begin{cases} \{\frac{X_1+X_2}{2}\} & \text{if } X_1 \neq X_2; \\ \{X_1 - 1\} & \text{if } X_1 = X_2. \end{cases}$$

Then, we get,

$$\begin{aligned} P_\theta(\theta \in C) &= P_\theta\left(\theta = \frac{X_1+X_2}{2} \middle| X_1 \neq X_2\right) P_\theta(X_1 \neq X_2) \\ &\quad + P(\theta = X_1 - 1 \mid X_1 = X_2) P_\theta(X_1 = X_2) \\ &= 1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{3}{4}, \end{aligned}$$

for all θ, so $C(X_1, X_2)$ is a 75% confidence set for θ. Thus, if we use this procedure repeatedly, we will be correct about θ three times out of four. But, if we observe $x_1 \neq x_2$, are we not 100% sure that $\theta = (x_1 + x_2)/2$? Why say that we are only 75% sure? This shows that ther are situations where pre-experimental optimality is not the appropriate approach for inference. However, the frequentist approach does not permit any (observed) data dependent confidence statements. There are many examples like this.

Example. To estimate μ in $N(\mu, \sigma^2)$, toss a fair coin. Have a sample of size $n = 2$ if it is a head and take $n = 1000$ if it is a tail. An unbiased estimate of μ is $\bar{X}_n = \sum_{i=1}^n X_i/n$ with variance $= \frac{1}{2}\{\frac{\sigma^2}{2} + \frac{\sigma^2}{1000}\} \sim \frac{\sigma^2}{4}$. Suppose it was a tail. Would you believe $\sigma^2/4$ is a measure of accuracy?

Example. Let X_1, X_2 be i.i.d. $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Let $\bar{X} \pm C$ be a 95% confidence interval, $C > 0$ being suitably chosen. Suppose $X_1 = 2$ and $X_2 = 1$. Then we know for sure $\theta = (X_1 + X_2)/2$ and hence $\theta \in (\bar{X} - C, \bar{X} + C)$. Should we still claim we have only 95% confidence that the confidence interval covers θ?

How is then conditioning on data to be done? Consider the example below.

Example. A laboratory test (such as RAT for COVID-19) is needed to check whether a person has a particular disease. The result of the test is either positive ($x = 1$) or negative ($x = 0$). Let θ_1 denote 'disease is present', θ_2 be 'not present'. $P(X = x|\theta)$ is as follows.

	$x = 0$	$x = 1$
θ_1	0.2	0.8
θ_2	0.7	0.3

The test is not fool-proof. 30% false positives and 20% false negatives appear.

Now suppose a patient is sent to the laboratory for this test and the test result comes out positive. What is the doctor to conclude regarding the presence or absence of the disease? Note that the question of interest is not whether the test result is positive or negative. Instead, what are the chances of the disease being present? i.e., $P(\theta = \theta_1|X = 1) = ?$

What we have are $P(X = 1|\theta = \theta_1)$ and $P(X = 1|\theta = \theta_2)$. We have the 'wrong' conditional probabilities! They need to be reversed or inverted. But how?

Suppose, in the concerned community, the disease is present in 5% of the cases. i.e., $P(\theta = \theta_1) = 0.05$. This is, however, not part of the sample data. This is pre-experimental. The doctor has this information from experience in the field. Now,

$$P(\theta = \theta_1|X = x) = \frac{P(\theta = \theta_1 \text{ and } X = x)}{P(X = x)},$$

and

$$P(X = x) = P(X = x|\theta_1)P(\theta = \theta_1) + P(X = x|\theta_2)P(\theta = \theta_2),$$

so applying the *Bayes Theorem*,

$$P(\theta = \theta_1|X = x) = \frac{P(X = x|\theta_1)P(\theta = \theta_1)}{P(X = x|\theta_1)P(\theta = \theta_1) + P(X = x|\theta_2)P(\theta = \theta_2)}. \quad (1)$$

Therefore,

$$P(\theta = \theta_1|X = 1) = \frac{0.8 \times 0.05}{0.8 \times 0.05 + 0.3 \times 0.95} = \frac{0.04}{0.04 + 0.237} = 0.123,$$

and $P(\theta = \theta_2|X = 1) = 0.877$. Positive blood test indicates only a 12.3% chance of disease being present in a random member of the community, so further diagnostic measures may be needed. On the other hand, this is important since the risk has more than doubled, from 5% to 12.3%.

(1) shows how to 'invert' the given conditional probabilities, $P(X = x|\theta)$ to derive the desired conditional probabilities, $P(\theta = \theta_i|X = x)$, which is an application of the Bayes Theorem. Since this involves an inversion, the name, *Theory of inverse probability* is used for statistical inference based on this approach. This was the usage at the time of Bayes and Laplace – late 18th century, before Fisher and Pearson. However, these days it is known simply as Bayesian inference. Note that, to obtain $P(\theta = \theta_1|X = 1)$, it is

essential to have $P(\theta = \theta_1)$ (and hence $P(\theta = \theta_0) = 1 - P(\theta = \theta_1)$). Where does this come from, and what kind of a probability is this?

Ingredients of Bayesian inference

likelihood function, $l(\theta|x) \propto f(x|\theta)$

prior distribution, $\pi(\theta) = \begin{cases} \text{probability mass function, if } \theta \text{ is discrete;} \\ \text{probability density function, if } \theta \text{ is continuous} \end{cases}$

What are the implications of using a prior distribution (π) on the unknown quantity θ?

There is usually some information (prior to sample data collection) available about θ; sometimes this may be precise but not often. Thus, there is usually a lot of uncertainty about θ. What is a good way to quantify uncertainty? Probability is the only well-accepted mathematical approach. This does not necessarily mean that θ is random. Probability is a tool to incorporate uncertainty, that is all. There is no requirement that probabilities must have a relative frequency interpretation based on a repeatable experiment. However, past data is a common source for prior probabilities. Recall the example on laboratory test for diagnosis. In this case, the prior probability, $P(\theta = \theta_1) = 0.05$ in the concerned population is simply the prevalence of the disease, about which medical experts are expected to have information. In the quality control example, the manufacturer wants to monitor the quality of his products. Random samples are taken periodically to estimate the proportion p of defectives. But the manufacturer has a lot of other information about his production process including past data on the proportion of defectives.

An important aspect of Bayesian inference is this: Whether useful prior information is available or not, a prior distribution is needed for the implementation of conditioning on data using the Bayes theorem.

Technically, a Bayesian takes the view that all unknown quantities, namely the unknown parameter and the data before observation, have a probability distribution. For the data, the distribution, given θ, comes from a model that arises from past experience in handling similar data as well as subjective judgment. The distribution of θ arises as a quantification of the Bayesian's knowledge and belief. If her knowledge and belief are weak, she may fall back on a common objective distribution in such situations.