

STAT 238 - Bayesian Statistics

Lecture Four

Spring 2026, UC Berkeley

Aditya Guntuboyina

28 Jan 2026

In the last lecture, we discussed the following problem.

1 Example 3: Inference from measurements

Problem 1.1. *Suppose a scientist makes 6 numerical measurements 26.6, 38.5, 34.4, 34, 31, 23.6 on an unknown real-valued physical quantity θ . On the basis of these measurements, what can be inferred about θ ?*

1.1 Frequentist Solution

Here is the standard frequentist solution to this problem. Use the model:

$$X_1, \dots, X_n \stackrel{\text{i.i.d}}{\sim} N(\theta, \sigma^2). \quad (1)$$

It then follows that $\bar{X}_n := (X_1 + \dots + X_n)/n \sim N(\theta, \sigma^2/n)$ which implies:

$$\frac{\sqrt{n}(\bar{X}_n - \theta)}{\sigma} \sim N(0, 1). \quad (2)$$

If σ is known, this gives the confidence interval $\bar{X}_n \pm \frac{\sigma}{\sqrt{n}}z_{\alpha/2}$ for θ (here $z_{\alpha/2}$ satisfies $\mathbb{P}\{N(0, 1) > z_{\alpha/2}\} = \alpha/2$). But this confidence interval cannot be computed as σ is unknown. It is natural to replace σ by the natural estimator:

$$\hat{\sigma} := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}.$$

But then the normal distribution in (2) needs to be changed to the Student t distribution with $n - 1$ degrees of freedom:

$$\frac{\sqrt{n}(\bar{X}_n - \theta)}{\hat{\sigma}} \sim t_{n-1}. \quad (3)$$

This leads to the confidence interval:

$$\left[\bar{X}_n - \frac{t_{n-1, \alpha/2}}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}, \bar{X}_n + \frac{t_{n-1, \alpha/2}}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2} \right]. \quad (4)$$

When $\alpha = 0.05$ and one plugs in the observed data $X_1 = 26.6, X_2 = 38.5, X_3 = 34.4, X_4 = 34, X_5 = 31, X_6 = 23.6$ with $n = 6$ in the above interval, we obtain the interval [25.598, 37.102].

1.2 Bayesian Solution

The Bayesian solution also leads to the same interval but with a different reasoning. We went over the calculations in the last lecture. Here are the main facts. We use the likelihood:

$$\text{Likelihood} = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right). \quad (5)$$

where $n = 6$ and $x_1 = 26.6, x_2 = 38.5, x_3 = 34.4, x_4 = 34, x_5 = 31, x_6 = 23.6$ denote the observed data points.

The unknown parameters are θ and σ . The prior is given by:

$$\theta, \log \sigma \stackrel{\text{i.i.d}}{\sim} \text{uniform}(-C, C) \quad (6)$$

for a very large positive constant C . In terms of the densities, (6) is the same as

$$\text{prior} = f_{\theta, \sigma}(\theta, \sigma) \propto \frac{\mathbf{1}\{-C < \theta < C, -C < \log \sigma < C\}}{\sigma}.$$

For this model, the posterior becomes:

$$\text{posterior} = f_{\theta, \sigma | \text{data}}(\theta, \sigma) \propto \sigma^{-n-1} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right) \mathbf{1}\{-C < \theta < C, -C < \log \sigma < C\}.$$

This is the joint posterior density of θ and σ . The posterior of θ alone is obtained by integrating out σ :

$$f_{\theta | \text{data}}(\theta) \propto \mathbf{1}\{-C < \theta < C\} \int_{e^{-C}}^{e^C} \sigma^{-n-1} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right) d\sigma$$

Because C is large, the limits of the integral can be taken to be 0 and ∞ . The integral then can be calculated precisely to obtain

$$f_{\theta | \text{data}}(\theta) \propto \mathbf{1}\{-C < \theta < C\} \left(\frac{1}{S(\theta)}\right)^{n/2}$$

where $S(\theta)$ is the sum of squares term:

$$S(\theta) = \sum_{i=1}^n (x_i - \theta)^2.$$

If C is large, then the indicator can be dropped (because it will essentially be always 1) so the posterior becomes:

$$f_{\theta | \text{data}}(\theta) \propto \left(\frac{1}{S(\theta)}\right)^{n/2} \propto \left(\frac{S(\hat{\theta})}{S(\theta)}\right)^{n/2}.$$

where $\hat{\theta} = \bar{x} = (x_1 + \dots + x_n)/n$ is the least squares estimator (which minimizes $S(\theta)$ over all θ). Thus the posterior mode is the mean.

It can be shown that this distribution is closely related to the t -distribution. Specifically,

$$\frac{\sqrt{n}(\theta - \hat{\theta})}{\sqrt{S(\hat{\theta})/(n-1)}} \mid \text{data} \sim t_{n-1} \quad (7)$$

where t_{n-1} denotes the t -density with $n - 1$ degrees of freedom. Note that $\hat{\theta} = \bar{x}$ and $S(\hat{\theta}) = \sum_{i=1}^n (x_i - \bar{x})^2$.

So the Bayesian point estimate is simply $\hat{\theta} = \bar{x}$ (this is the posterior mean, median and mode!). A $100(1 - \alpha)\%$ uncertainty interval for θ is given by:

$$\left[\hat{\theta} - \frac{1}{\sqrt{n}} t_{n-1, \alpha/2} \sqrt{\frac{S(\hat{\theta})}{n-1}}, \hat{\theta} + \frac{1}{\sqrt{n}} t_{n-1, \alpha/2} \sqrt{\frac{S(\hat{\theta})}{n-1}} \right] \quad (8)$$

where $t_{n-1, \alpha/2}$ is the $(1 - \alpha/2)$ quantile of the Student t -distribution with $n - 1$ degrees of freedom. This uncertainty interval is referred to as the Bayesian Credible Interval.

Thus, in problem 1.1, the standard frequentist and Bayesian solutions coincide.

1.3 Frequentist vs Bayes

However, it is very easy to break this coincidence. For example, consider the following problem:

Problem 1.2. *A scientist gets repeated measurements on an unknown physical quantity θ until a measurement smaller than 25 is obtained. Suppose the resulting data is 26.6, 38.5, 34.4, 34, 31, 23.6. On the basis of these measurements, what can be inferred about θ ?*

Note that the observed data is exactly the same as before.

The frequentist confidence interval (4) is no longer valid, because the frequentist probability statement (3) is no longer valid. This is because the number of datapoints n cannot be taken to be deterministically equal to 6. So the frequentist probability that we need to calculate is (below we denote the number of data points by N and treat it as a random variable):

$$\mathbb{P} \left\{ \bar{X}_N - \frac{t_{N-1, \alpha/2}}{\sqrt{n}} \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X}_N)^2} \leq \theta \leq \bar{X}_N + \frac{t_{N-1, \alpha/2}}{\sqrt{N}} \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X}_N)^2} \right\}$$

where X_1, X_2, \dots are i.i.d $N(\theta, \sigma^2)$ as before and

$$N := \inf \{n \geq 1 : X_n \leq 25\}.$$

The probability above is complicated and there is no reason for it to be exactly equal to $1 - \alpha$. Constructing valid frequentist confidence intervals in the presence of *stopping rules* (such as the rule of stopping as soon as we observe a data point smaller than 25) is, in fact, a problem of current research (see e.g., the paper <https://arxiv.org/abs/2210.01948>).

In contrast to frequentist inference, the Bayesian inference procedure will not change. This is because the likelihood function in Problem 1.2 is the same as the likelihood function in Problem 1.1. To verify this, consider the following likelihood in Problem 1.2 (below δ denotes the rounding error in the observations, which is extremely small)

$$\begin{aligned} & \text{Likelihood in Problem 1.2} \\ &= \mathbb{P} \{ \text{observed data} \mid \theta, \sigma \} \\ &= \mathbb{P} \{ X_1 \in [x_1 - \delta, x_1 + \delta], \dots, X_6 \in [x_6 - \delta, x_6 + \delta], X_1 \geq 25, \dots, X_5 \geq 25, X_6 < 25 \mid \theta, \sigma \} \end{aligned}$$

In other words, we are writing the probability that the first five observations are all larger than 25 while the sixth observation is smaller than 25, in addition to the exact values of the observations, in the likelihood. But it is clear that these additional constraints do not change the probability (as an example, just note that $\mathbb{P}(Z = 5) = \mathbb{P}(Z = 5, Z < 10)$). The additional restriction $Z < 10$ does not affect the probability because it is already covered in $Z = 5$). Thus

Likelihood in Problem 1.2

$$\begin{aligned} &= \mathbb{P}\{X_1 \in [x_1 - \delta, x_1 + \delta], \dots, X_6 \in [x_6 - \delta, x_6 + \delta], X_1 \geq 25, \dots, X_5 \geq 25, X_6 < 25 \mid \theta, \sigma\} \\ &= \mathbb{P}\{X_1 \in [x_1 - \delta, x_1 + \delta], \dots, X_6 \in [x_6 - \delta, x_6 + \delta] \mid \theta, \sigma\} \\ &= \text{Likelihood in Problem 1.1.} \end{aligned}$$

Since the likelihood is the same in both problems, Bayesian inference for both problems will be the same (note the priors will be the same as there is no reason to use different priors). Therefore, from the Bayesian perspective, stopping rules can be ignored for inferring θ , because they do not affect the likelihood.

This example shows clearly that frequentist inference violates the Likelihood Principle (the likelihood principle states that “all the evidence in a sample relevant to model parameters is contained in the likelihood function”). See https://en.wikipedia.org/wiki/Likelihood_principle for more information on the likelihood principle.

On the other hand, Bayesian inference always satisfies the likelihood principle (assuming that priors are the same), because data enters the Bayesian posterior calculation only through the likelihood.

Here is another example of violation of the likelihood principle in frequentist inference.

2 Example 4: Coin Fairness Testing

Problem 2.1. *Suppose a coin is tossed 12 times leading to the outcome: TTTHTHTTTTH (this has 3 heads and 9 tails). What is your assessment of the fairness of the coin?*

2.1 Frequentist Solution

For the usual frequentist answer to this question, we assume that the observed sequence of outcomes are the realization of random variables X_1, \dots, X_n (with $n = 12$) that are independently distributed according to the $\text{Bin}(n, p)$ distribution for some unknown p . We need to test the (null) hypothesis that $p = 0.5$ against, say, the alternative $p < 0.5$. This can be done by calculating the p -value which is the probability (under the assumption $p = 0.5$) of getting 3 or lower heads. The distribution of the number of heads under the null distribution is $\text{Bin}(n, 0.5)$ so the p -value is

$$\left(\binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0} \right) \frac{1}{2^{12}} = \frac{299}{4096} = 0.073 = 7.3\%$$

which does not lead to a rejection of the null hypothesis at the usual 5% level.

In this p -value calculation, we implicitly assumed that the experiment consisted of tossing the coin 12 times where 12 was *a priori* chosen by the coin tosser. Consider now the alternative scenario where the coin tosser wanted to toss the coin *until the point where 3*

heads are observed. Now for the same outcome, the p -value will change. Indeed now the random variable of interest will become $N =$ number of tosses and the p -value will equal the probability of needing to toss the coin 12 or more times to get the 3 heads (assuming fairness). This is calculated using the negative binomial distribution as:

$$1 - \sum_{n=3}^{11} \binom{n-1}{2} 2^{-n} = \frac{134}{4096} = 0.0327 = 3.27\%$$

and this leads to rejection of the null hypothesis at the 5% level.

Note that the “likelihood function” is the same function $p^3(1-p)^9$ whether the sample size was predetermined or whether the coin was tossed till 3 heads are observed. But the procedure obtained for testing $p = 0.5$ has changed from the binomial to the negative binomial case. This means that p -valued based frequentist inference violates the Likelihood Principle. Here is a story from the wikipedia article on the “Likelihood Principle” (see https://en.wikipedia.org/wiki/Likelihood_principle) which puts an interesting context to these numbers:

Suppose a number of scientists are assessing the probability of a certain outcome (which we shall call 'success') in experimental trials. Conventional wisdom suggests that if there is no bias towards success or failure then the success probability would be one half. Adam, a scientist, conducted 12 trials and obtains 3 successes and 9 failures. One of those successes was the 12th and last observation. Then Adam left the lab.

Bill, Adam's boss in the same lab, continued Adam's work and published Adam's results, along with a significance test. He tested the null hypothesis that θ , the success probability, is equal to a half, versus $\theta < 0.5$. The probability that out of 12 trials, 3 or fewer (i.e. more extreme) were successes, if H_0 is true, is 7.3%. Thus the null hypothesis is not rejected at the 5% significance level.

*Adam actually stopped immediately after 3 successes, because his boss Bill had instructed him to do so. After the publication of the statistical analysis by Bill, Adam realizes that he has missed a **later instruction** from Bill to instead conduct 12 trials, and that Bill's paper is based on this second instruction. Adam is very glad that he got his 3 successes after exactly 12 trials, and explains to his friend Charlotte that by coincidence he executed the second instruction. But Charlotte then explains to Adam that the p -value should now be changed to 3.27% and the result becomes significant at the 5% level. Adam is astonished to hear this.*

For more comments on the violation of the likelihood principle by p -values, read MacKay [1, Section 37.2].

We shall look at the Bayesian solution to this problem in the next lecture.

References

- [1] MacKay, D. J. C., (2003) *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.