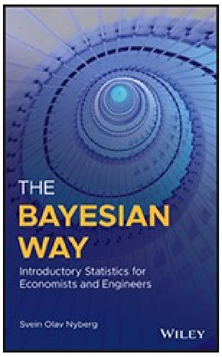


The Bayesian Way: Introductory Statistics for Economists and Engineers

by Svein Olav Nyberg

Chapter 11: Introduction

11.1 Mindful of the Observations

In the old days, Bayesian statistics was often called *inverse* statistics. That is a rather accurate way of putting it, if by "ordinary" (non-inverse) statistics we mean working *from* a model *to* predictions of observations, as shown in Figure 11.1. The model is then summed up in probability distributions for the observations, as for instance a Normal distribution φ, or distributions $bern_p$ for Bernoulli processes, and $erl_{(k,\,\lambda)}$ and $pois_\lambda$ for Poisson processes.

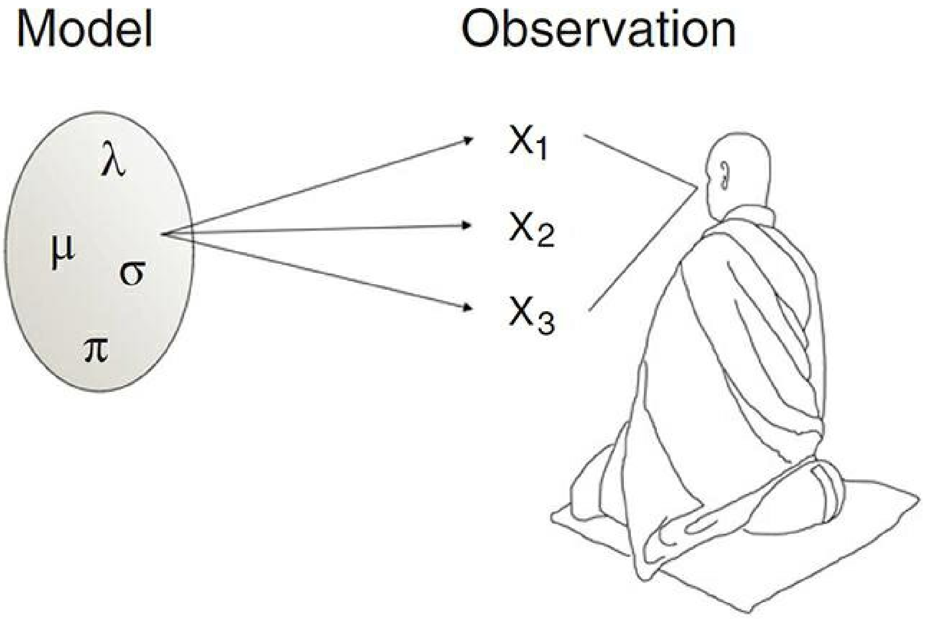


Figure 11.1: The zen monk Sōzen sees both the model and the observations

The "forward" direction in this paradigm is conclusion from model to observation, whereas *inverse* statistics concludes from observation to model, and then after that to the *next* observation. We have borrowed abbot Sōzen from the Norwegian zen temple Bugaku in order to illustrate how we perform Bayesian statistical inference.

We had our first glimpse of statistical inference in Chapter 6. Our core example was the Gamesmaster's dice, Example 6.3.5, which we expanded in Example 6.7. Gamesmaster picked a random die from a bag containing one each of the dice D_4, D_6, D_8, D_{10}, D_{12}, D_{20} that are all painted red on four of the faces and white on the remainder. He then tossed the die behind a screen, and the players were told only if the die had landed red or white, but not which die he had. From this, they were to guess which die Gamesmaster had picked, and to give probabilities for the reds and whites of the next observations.

We then looked at probability distributions, and got a taste of the extended concept of "population": when we have a finite population, an observation is equivalent to a random *sampling* from that population. But often, our observations arise from an ongoing process; there is no fixed set from which to sample. It is often still noticeable, however, that the observations follow a statistical distribution $f(x)$, as if they were sampled from an infinitely large population whose measurement distributions are described by $f(x)$. So, in the extended concept of "population", we will say that the essential feature of the model and the population is not the precise elements, but the *distribution* $f(x)$ of the values. We say that we *sample from a probability distribution* $f(x)$.

11.1.1 Models and Walls

We let abbot Sōzen illustrate, while Bard and Frederick and their friends make comments.

"So what is the difference between Frederick's frequentism and Bard's Bayesianism here?" Sam asks.

"So far," Frederick replies, “nothing. Or almost nothing. The two of us have different interpretations of the meaning of the basic concept of probability, and in the Gamesmaster's dice example, we frequentists insist that, when the die has been picked, it has been picked. From there on, there is nothing random about it: it was either picked, or it wasn't picked. As you will see if you remember Example 5.1.6, we will say that after the die is picked, the value of the probability $P(D_8)$ is either 0 or 100%, regardless of our state of knowledge about it. But to the Bayesians, probability is something entirely different, so the dice example works nicely within their paradigm.

"But the dice example is still a useful one, for it illustrates both our similarities and our differences," Frederick continues, "and what we both do. Each die D_k corresponds to a model, a probability distribution for assigning probabilities to our red–white observations. But we do not know which model is the correct one. Is it D_8? Is it D_{20}? We do not know, for we do not see the probability distribution directly. As illustrated in Figure 11.2, we *only see the observations*. We see red and we see white, but we do not see the die. We do not see how many faces are white, and how many are red."

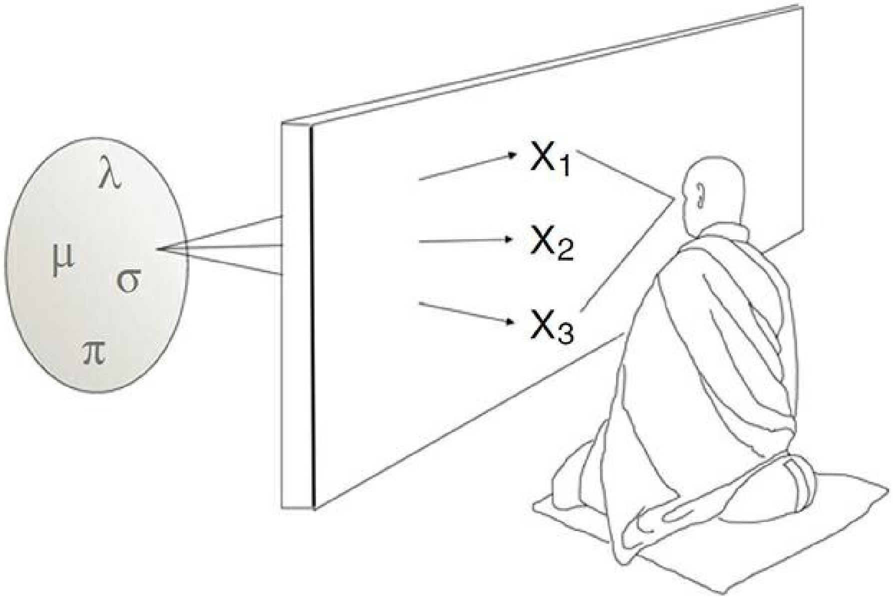


Figure 11.2: Zen monk Sōzen does not see the model, only the observations

"Statistical investigations are all essentially like that," Bard says, and takes over. "The underlying reality is not directly accessible to us, but is as if hidden behind a screen, just like the dice. We do not see the probability distribution from which the observations are sampled, but we see the observations. Likewise, we do not *see* the probability p that a coin will give *heads*, but we see the outcomes of individual coin flips. Neither do we see the waiting time parameter λ for a Poisson process, but we may measure individual actual waiting times."

"There are times when we *in principle* could have discerned the underlying model," Frederick interjects. "We consider the weights of salmon in the river Loppa to be Normally distributed $\varphi_{(\mu,\, \sigma)}$, and if we had managed the feat of emptying the river of every single fish, we *would* have the precise values of μ and σ. But note that this would be the weights of the salmon there and then, and not of one minute later, and it would work only if we captured and weighed *all* the salmon at once. So though theoretically thinkable, such exhaustive knowledge will *in practice* be impossible, and the population parameters will be as if hidden behind a screen."

"But what do you do if you can't know?" Sam asks, "do you just give up?"

"No," Frederick replies, "we estimate the parameters of the distribution, like Sōzen in his figure (Figure 11.3). But this is where Bard and I part ways. For Bard considers the parameters to be stochastic variables, and sets up probability distributions for them according to his observations. For my own part, I consider only the observations to be stochastic variables, whereas the parameters are fixed but unknown magnitudes whose probability of having such and such values, or of being within such and such intervals, is 0 or 100%. So Sōzen's illustration (Figure 11.3) illustrates Bard's viewpoint only, not mine."

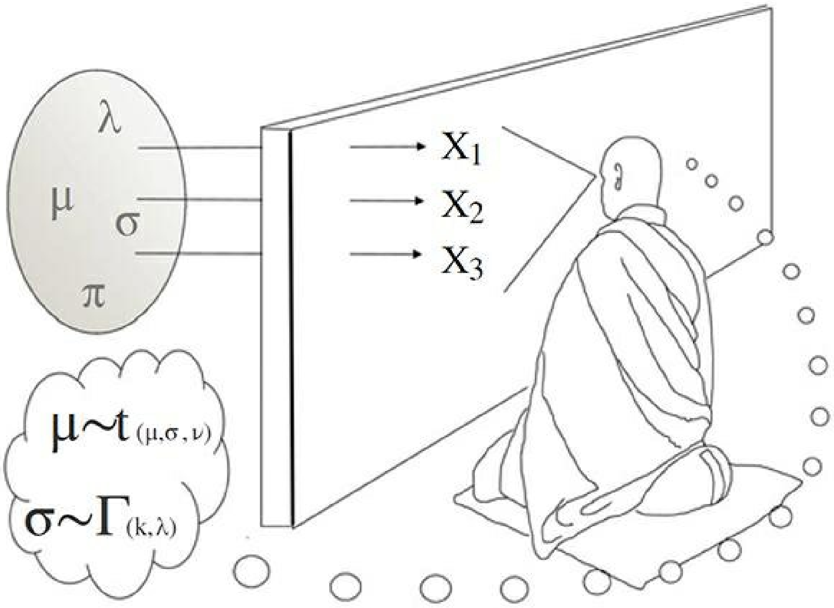


Figure 11.3: Zen monk Sōzen estimates the model parameters based on the observations

11.1.2 Randomness

"To get good estimates of the parameters," Bard says, "we have to make our observations in the right way. We need representative observations. If we wish to know what the British think about the question of Scottish independence, we will not be doing an overly good job of it if all we do is poll the first 20 Scotsmen exiting a football match between England and Scotland. But what *does* work well, is *random sampling* from the population, since randomness has no preference. Both Frederick's techniques and mine are based on such non-preferential sampling."

"But we still need to stay alert," Frederick takes over, "for what is *random* is not always easy to discern. Picking random persons outside the football match is random enough, but it is the wrong ‘random’. We need to make sure our randomness is *unbiased*, not preferring one part of the population to another."

"In all fairness, we should mention that *randomness* is a hard enough problem that it belongs to philosophy as well as to statistics," Bard interjects, "for what does it really mean for anything to be random? As we saw in the discussion on randomness in Section 5.14, a die may be ‘random’ in an everyday sense. But upon closer examination, when we looked at the physics of the situation, it was anything but random. Maybe randomness is simply another way of saying *I don't know* – an expression of our ignorance or partial ignorance, about the details of the system. Or maybe the proponents of the theory of *propensity* are right: that randomness is a fundamental property of some systems."

"And we tend not to know these underlying mechanisms," Frederick replies, "but what we know, is how we choose our observations. We should therefore examine what sources of bias there may be in this particular system, and then we should strive to eradicate these sources of bias from our samples, to the best of our ability. For instance: if we are looking at salmon weights in Loppa, we should examine whether certain parts of the river have smaller or bigger fish than the rest, or if different sizes of fish are more easily caught at different times of the day. We then remove the possibilities of such bias by not fishing at only one spot, or at only one time of day."

"And *then* we use the weight of the fish caught without such bias to estimate the weight of all the salmon of Loppa?" Sam asks.

"Precisely!" Bard and Frederick intone in unison.

11.1.3 Next Observation

"After that," Bard continues, “we estimate the next observation – the weight of the next salmon. This is what Sōzen does in Figure 11.4. Or maybe we don't. *That* depends what the goals of our investigations are.

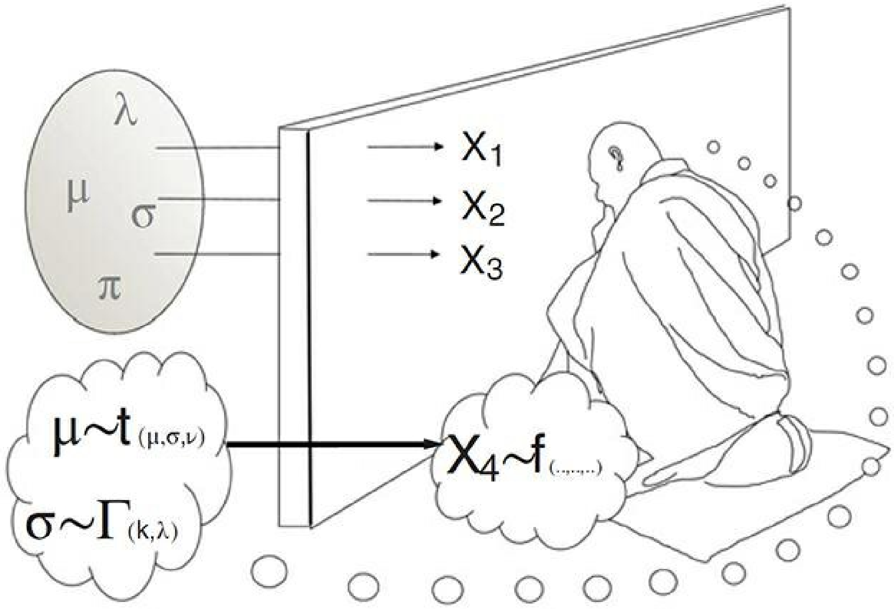

λ
μ
σ
π
X_1
X_2
X_3
$\mu \sim t_{(\mu,\sigma,\nu)}$
$\sigma \sim \Gamma_{(k,\lambda)}$
$X_4 \sim f_{(..,..,..)}$

Figure 11.4: Zen monk Sōzen uses his estimate of the model to estimate the next observation

“Sometimes, the question is whether a person has cancer or not, and we use diagnostic observations to determine that. In such instances, we are solely interested in the question of whether or not they have cancer, not in predicting the result of diagnostic observation.

"The example of Gamesmaster's dice is at the opposite end: here, our real interest is only in whether the next toss will yield a red or a white, whereas estimating the number of faces is interesting only as a means to that end: when I have a probability distribution for that parameter – the number of faces – I may also find a probability distribution for the next observation."

"I notice that you share the same goal, but evaluate the *methods* differently," Sam comments, "but could I ask you both to clarify one thing for me. I don't quite understand what Frederick means by *estimate* if it's not a probability distribution, and I don't understand how Bard is able to transmogrify observations into probability distributions. Would you care to explain?"

"That's why we're here," Bard replies. "The key to answering both questions lies in understanding our views on probability. *My* philosophy allows me a *prior* probability expressing my degree of knowledge and ignorance prior to the investigations. If I know nothing, I choose to work with a prior expressing my ignorance. But regardless of how much or how little I know, I have a prior probability. Having a prior probability allows me to apply Bayes’ theorem to the observations to find a *posterior* probability, just as we did in the Gamesmaster's dice example. That is the key to *my* side. Without a *prior* probability, it is not possible to obtain a *posterior* probability to estimate the model."

"That's me!" Frederick says and waves his hands. "I'm the prior-less guy, so that you can see the difference between us. For us frequentists, degree of knowledge or conviction, ignorance, or whatever, is irrelevant. For us, probability is not about degrees of knowledge, hence the kind of *prior probability* Bard talks about makes no sense within our framework. Frequentist estimates are properties of the data alone."

11.1.4 The Data Alone

"But aren't the estimates properties of the model as well?" Sam queries Frederick.

"Actually, no," Frederick replies. "I know it might sound strange, and though I consider my view to be far more correct than Bard's, I have in time realized that quite a few, even professors in other disciplines, never get a good grasp of what my methods are actually all about. We have had our meetings of concern in the statistics association, …"

"… and the solution is to teach a method that people understand: *Bayes!*" Bard interjects.

"… and the solution is *not agreed upon*," Frederick retorts, "but my father taught me something that might be of use to understand our two worlds. He was in control of one of the major casinos on the island I come from, and he told me that the trick to running a successful casino is not to win every game, but to win *in the long run*. This applies even though you as the controller of the casino often hold the same kind of privileged position that the Gamesmaster has in the dice example."

Frederick stops to see if Sam is paying attention, and continues: "The gambler may win or lose, because gambling is a hobby to him. Or at least ought to be. But if you are running the casino, this is your livelihood, so by the end of the day, in summary, you must be a net winner. You may lose individual games along the way, and this is even in your interest, for that is what attracts gamblers to your casino. But your net total, when the day is done, should be one where your winnings outrank your losses. Are you with me?"

Sam nods, "but what does this have to do with statistical estimates?"

"Everything!" Frederick replies, "at least in the frequentist world. For we use the data, and the data alone, to estimate the model. These estimates are, as I told you, properties of the data, and which data we get from our investigation is random. So whether our estimates capture or do not capture the parameter is …" says Frederick, and adds a dramatic pause, "random! You see: like the casino, we do not mind missing the target in individual cases. But our techniques are designed to capture the parameters as best as possible *in the long run*. Bard's techniques are more like those of a gambler, trying to win the individual gambles."

Sam's eyes show a glimmer of understanding, so Frederick adds: "But I will readily admit that our techniques do not bear the same semblance of unity that Bard's do."

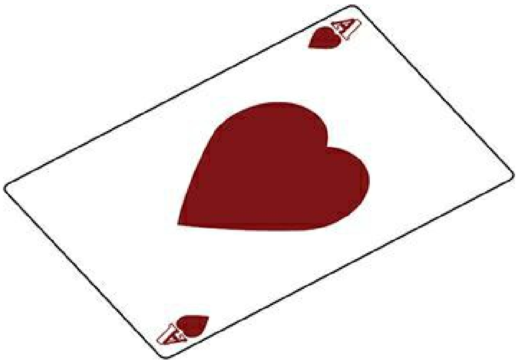
"That's right," Bard points out, "for whereas I find all that I need through the (*posterior*) probability distribution of the model, Frederick does not have any *core* engine driving all of his methods from a single framework. My colleagues have often called frequentist techniques *ad hoc inventions* …"

"And some of *mine* call your priors *superstition*," Frederick interjects.

"But despite the differences between our two schools, the two of us have agreed to present statistics together," Bard continues, "for the benefit of you students. We think that regardless of which of our two school you are attracted to, you will understand it better when it is contrasted to the other one. For the difference lies in different philosophies, and they again are best understood in each others' contrast. This is also the root source of our differences. That, and of course disagreements as to the practicality of our two approaches in handling the uncertainty that is after all the subject matter of statistics. And we know and respect each others' techniques for what they are."

"Amen!" Frederick agrees. "But as you surely understand, then, I would like to explain *my* techniques on a case basis, if that is all right with you, Sam. I've even got my own chapter in Bard's book, Chapter 16!"

Sam nods, but at the same time, Frederick and Bard remember that they are competitors in an area that is even closer to the hearts of both of them. For they are both in love with their common friend Mina.



Bard and Frederick are of course unable to separate statistics from their infatuations, and start discussing how large a proportion of her hugs Mina will be granting each of them. The statistical model in question is then her tendency of handing out hugs, parameter **p** for the probability, whereas the observations are the hugs themselves – or rather: to whom the hugs are given. The *future* observation is then who gets the next hug. Bard, who is the Bayesian, has a prior belief that the two of them have an equally large chance of getting the next hug, whereas the frequentist Frederick says he will wait and see until he has some observations to go by.

Then Mina gives the first hug to Frederick.

Frederick observes that he has received 100% of the hugs, and that thus the best estimate is that Mina will hug him only; he therefore predicts that Mina's next hug will be his as well. In Section 16.1, we will learn about the conceptual underpinnings of Frederick's estimate: *unbiased* point estimates. We will also be looking at a more nuanced and less infatuated form of Frederick's reasoning around proportions in Section 16.3.1.

Bard on his side concludes that Mina probably will be giving more hugs to Frederick than to himself, but not all of them. Bard weighs his initial model against the data gathered, and his estimate is that Mina will give $\frac{1}{3}$ of her future hugs to him, and $\frac{2}{3}$ to Frederick. This is a *Bayesian* point estimate. We will learn more about Bayesian estimates of proportions in Section 13.2.

After a week has passed, Bard has received 17 hugs, and Frederick 28. In *Frederick's* model, Bard's future share is $\frac{17}{17+28} \approx 37.8\%$ of the hugs, whereas Frederick himself gets the rest. I *Bard's* model, he will be receiving $\frac{17+1}{17+1+28+1} \approx 38.3\%$ of the hugs. So we see that even though the two friends develop their models in different ways, they converge to some kind of agreement as they gather more data. The end of the story? The end of this story is that Mina dates Sam. She is her own woman, and not the subject of possessive calculations. Or as Bard and Frederick head-shakingly agree, "in statistics and romance, nothing is certain!"

11.2 Technically …

In statistical inference, there are a few technical phrases that are well worth noting, and we have compiled the following list of key terms.

Parameter: We implicitly assume that the underlying population may indeed be well described by means of a probability distribution, and that this distribution again belongs to a certain class and may be specified by means of a few parameters p_1, p_2, …, p_k. If, for instance, the population is the salmon weights in Loppa, and we say that the salmon weight follows the probability distribution $\varphi_{(3,\, 0.7)}$, then $\mu = 3$ and $\sigma = 0.7$ are the parameters in our investigation. In our Loppa examples, we knew the values of μ and σ, but in statistical inference, the values of these parameters are usually hidden, and known only through *estimates*.

Observation: Before our observations, our future observations are *stochastic variables* X_1, X_2, … following the probability distribution(s) of the population. When we have observed X_k, have a concrete value x_k, we call this concrete value x_k a *realized value* for X_k. The Loppa salmon weights are $X_k \sim \varphi_{(3,\, 0.7)}$ prior to weighing. After weighing, their weight are realized values. For instance, the unknown X_5 has materialized as the concrete $x_5 = 4.1$.

Statistic: We rarely need the individual observations when we perform inference; what we need are mathematical summaries of the data; such a summary number is called a *statistic*. Fundamentally, any number that is a function of the observations is a statistic, but the interesting and relevant ones tend to be the ones we used for summing up our data in Chapter 2: mean ($\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$), variance, standard deviation, median, and percentile. Here too, we differentiate between the statistic Ψ as a stochastic variable prior to observation, and its realized value ψ afterwards.

A collection of statistics is *sufficient* if they contain enough information for our inferences. For instance, for the Loppa salmon, the mean and the variance are together *sufficient statistics* for inference on the parameters μ and σ. The realized value of $\bar{X}$ is in this instance $\bar{x}$.

Estimator: A statistic $\hat{\Theta}$ is an *estimator* if it is an estimate (a guess) at the value of the parameter θ. For instance, for the Loppa salmon, $\bar{X}$ is an estimator for μ. Since using "the data alone" for estimation is primarily a frequentist notion, estimators are a topic in Section 16.1.

Posterior: Bayesians typically code and extract all information about a parameter through its *posterior* probability distribution. This is uniquely Bayesian. A Bayesian may for instance say, after 15 observations, that μ, the mean salmon weight in Loppa, follows a Student's *t* distribution, $\mu \sim t_{(2.9,\, 0.1,\, 14)}$, whereas he after 200 observations may come up with the more precise estimate $\mu \sim t_{(2.93,\, 0.04,\, 199)}$. This is the topic of Chapters 12 and 13.

Prior: Another key Bayesian concept. Whereas the *posterior* codes the total information available after the new observations, the *prior* codes the information prior to these observations. The observations themselves are coded into the *likelihood*.

11.3 Reflections

?

1. Bayesian/frequentist
   a. Who makes their estimates of the model parameters from the data alone?
   b. Who speaks of *P*(observation | model)?
   c. Who speaks of *P*(model | observation)?

d. Who presupposes *randomness* for their methods?

e. Who does all their inference through a probability distribution?

f. Who speaks of *unbiased* estimates?

2. What are the two main purposes of statistical inference mentioned by Bard and Frederick? What is the difference between these two purposes, and how are the purposes related?

3. Why is *randomness* important?

4. May you observe a population, a model, or the model's parameters directly?

5. Your company has acquired the Chuck Wood's lumber mill. Along with the mill itself, they also got the mill's inventory. Your job is to estimate the humidity of the lumber by measuring 100 units. Discuss in groups which factors may bias the selection and sampling of units.

6. Discuss strengths and weaknesses in Bard's and Frederick's estimates of the proportion of hugs Mina will give to each of them. May one of the ways of analysing fit better in one context, and the other better in another context? If so: which kind of analysis fits which kind of context best?

Answers

1. Bayesian/frequentist

a. The frequentists.

b. Both frequentists and Bayesians.

c. The Bayesians.

d. Both frequentists and Bayesians.

e. The Bayesians.

f. The frequentists.

2. These assignments are for reflection, to be discussed in groups. There is no fixed right answer.

3. These assignments are for reflection, to be discussed in groups. There is no fixed right answer.

4. These assignments are for reflection, to be discussed in groups. There is no fixed right answer.

5. These assignments are for reflection, to be discussed in groups. There is no fixed right answer.

6. These assignments are for reflection, to be discussed in groups. There is no fixed right answer.