Course:	Mathematical Statistics
Term:	Fall 2018
Instructor:	Gordan Žitković

Lecture 12
Bayesian Statistics

12.1 The frequentist and subjectivist interpretations of probability

There are two major schools of thought, **frequentist** and **subjectivist**, when it comes to the interpretation of the meaning of probability[1]. The frequentist (frequency) interpretation argues that the only way to interpret the probability of an event is to repeat the "experiment" a large number of time and compute the relative *frequency* of that event. That relative frequency, for a frequentist, *is* the probability of that event. A subjectivist, on the other hand, thinks of probability of an event as a *degree of belief*; for him or her, a probability is a subjective matter, which may differ from an individual to an individual. It is not an objective state of reality, but simply a numerical description of the state of an individual's knowledge and belief about a certain occurrence.

The usual laws of probability, most notably the *additivity* $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$, for $A \cap B = \emptyset$, hold in both interpretations, but for different reasons. In the frequentist's world, additivity of probability follows from the fact that the frequencies are additive (the "number of times A occurred" plus "the number of times B occurred" is exactly the same as "the number of times A or B occurred".)

The explanation is different for a subjectivist. She argues as follows: if an individual believes in A to a degree p_A, she will be indifferent between the cash amount of $\$p_A$ and a bet denoted by $\mathbf{1}_A$ which pays \$1 is A happens, and \$0 if it does not. Similarly, she will be indifferent between $\$p_B$ in cash, and the bet $\mathbf{1}_B$. Unless she is willing to part with all of her possessions, that same individual will then also need to be indifferent between $p_A + p_B$ in cash and the bet $\mathbf{1}_{A\cup B}$. Why? Suppose that she is not consistent in this way and that she prefers the cash to the bet. That means that she would be willing to accept a tiny bit less cash (say $p_A + p_B - \varepsilon$) in exchange for the bet $\mathbf{1}_{A\cup B}$ and still be happy. Anybody else could then come in and sell the separate bets $\mathbf{1}_A$ and $\mathbf{1}_B$ for p_A and p_B, respectively, to her, and then, come back and buy the

[1] this is a gross simplification - many sub-interpretations are lumped here under two umbrella notions. Third points of view, also exist.

aggregate bet $\mathbf{1}_{A\cup B} = \mathbf{1}_A + \mathbf{1}_B$ for $p_A + p_B - \varepsilon$. The net of these transactions is that our inconsistent subjectivist loses $\varepsilon > 0$ and the other party makes $\varepsilon > 0$ in profit, without assuming any risk. If more profit is needed, this game can be repeated n times, with the total profit of $n\varepsilon$. In other words, one cannot hold arbitrary subjective beliefs about likelihoods of different events without losing all their money to an opportunistic counter party.

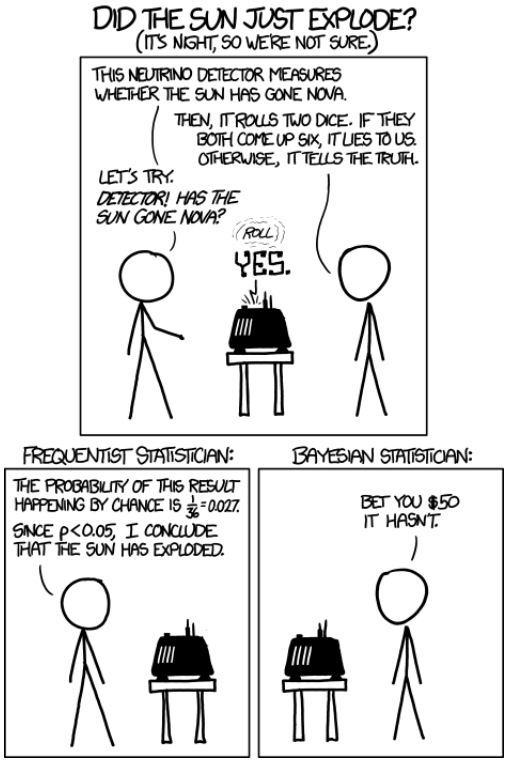

Figure 1. An XKCD comic on Frequentists and Bayesians

12.2 Frequentist vs. Bayesian statistics

The two schools of thought described above gave birth to two prevailing schools of thought on how to perform statistical inference. Frequentists think of the parameter θ as fixed, but unknown. Once an estimator is chosen, its performance is assessed by "repeating" the experiment many times, and thinking of the confidence levels (say) as the relative frequency of the number of experiments in which the "true parameter" θ ended up in the constructed confidence interval.

Bayesians, on the other hand, think of data-gathering as a procedure which gradually adjusts their subjective opinions of the value of the "true parameter" θ. Being consistent in their beliefs (as explained in the previous section), Bayesians must adhere to the standard rules of probability. In particular, their beliefs need to satisfy the **Bayes rule**; for simplicity, let us assume for now that the unknown parameter θ can take one of n discrete values $\theta_1, \ldots, \theta_n$:

$$\mathbb{P}[\theta = \theta_k | Y = y] = \frac{\mathbb{P}[\theta = \theta_k]\,\mathbb{P}[Y = y | \theta = \theta_k]}{\sum_i \mathbb{P}[\theta = \theta_i]\,\mathbb{P}[Y = y | \theta = \theta_i]}$$

It relates the **posterior probability** $\mathbb{P}[\theta = \theta_k | Y = y]$ to the **prior probabilities** $\mathbb{P}[\theta = \theta_i]$, $i = 1, \ldots, n$. In other words, it tells us how a Bayesian's beliefs about various values of the parameter θ change, once $Y = y$ is observed. This procedure is called **Bayesian updating** and is the foundation of the entire subject of Bayesian statistics.

It is important to remark that the Bayes rule prescribes how a belief needs to be updated once new information ($Y = y$) arrives, but it says nothing about the prior belief. Often, it, itself, is a result of a similar updating procedure performed in the past, but that only kicks the can down the road. In fact, the prior belief is an input to any Bayesian model. In practice, one often holds a strong belief about θ from a experience or knowledge, and then quantifies it as a probability distribution. Alternatively, if no prior information whatsoever is available, one tries to assign a so-called **uninformative prior**. For example, in the simple case described above, an uninformative prior would assign equal probabilities to each possible value of the parameter θ, i.e., we would have $\mathbb{P}[\theta = \theta_i] = \frac{1}{n}$ for each $i = 1, \ldots, n$.

Example 12.2.1. We continue the setting of Example 11.2.1. The possible values of the parameter θ are $\theta_1 = 1$, $\theta_2 = 2$ and $\theta_3 = 3$. Without any other prior information of the probabilities $\mathbb{P}[\theta = \theta_i]$ for $i = 1, \ldots, 3$, we assign an uninformative prior

$$\mathbb{P}[\theta = 1] = \mathbb{P}[\theta = 2] = \mathbb{P}[\theta = 3] = \tfrac{1}{3}.$$

This allows us to compute the posterior probabilities, given the observation $Y = G$:

$$\begin{aligned}\mathbb{P}[\theta = 1 | Y = G] &= \frac{\mathbb{P}[Y = G | \theta = 1]\,\mathbb{P}[\theta = 1]}{\sum_{i=1}^{3} \mathbb{P}[Y = G | \theta = i]\,\mathbb{P}[\theta = i]} \\ &= \frac{0.9 \times \frac{1}{3}}{0.9 \times \frac{1}{3} + 0.15 \times \frac{1}{3} + 0.5 \times \frac{1}{3}} \approx 0.58\end{aligned}$$

Similarly,

$$\mathbb{P}[\theta = 2 | Y = G] = \frac{0.15 \times \frac{1}{3}}{0.9 \times \frac{1}{3} + 0.15 \times \frac{1}{3} + 0.5 \times \frac{1}{3}} \approx 0.10$$

and

$$\mathbb{P}[\theta = 3|Y = G] = \frac{0.5 \times \frac{1}{3}}{0.9 \times \frac{1}{3} + 0.15 \times \frac{1}{3} + 0.5 \times \frac{1}{3}} \approx 0.32.$$

In other words, the uniform prior distribution

	1	2	3
	0.33	0.33	0.33

turned into the posterior distribution

	1	2	3
	0.58	0.10	0.32

after we learned that $Y = G$.

Suppose now that we know that the three busses arrived from three different neighboring towns of sizes $5,000$, $35,000$ and $10,000$, respectively. If we picked a random football fan and tried to guess his or her bus of origin (without asking about the team they support), we would probably guess bus 2 and assign the following prior probability distribution to the parameter θ:

$$\begin{array}{|c|c|c|c|}\hline & 1 & 2 & 3 \\ \hline & \frac{5000}{50000} & \frac{35000}{50000} & \frac{10000}{50000} \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline & 1 & 2 & 3 \\ \hline & 1/10 & 7/10 & 1/5 \\ \hline \end{array}.$$

If we subsequently learn that out fan supports the team G, i.e., that $Y = G$, Bayes formula would produce a new (posterior) distribution for θ:

$$\mathbb{P}[\theta = 1|Y = G] = \frac{0.9 \times \frac{1}{10}}{0.9 \times \frac{1}{10} + 0.15 \times \frac{7}{10} + 0.5 \times \frac{1}{5}} \approx 0.31$$

and similarly,

$$\mathbb{P}[\theta = 2|Y = G] \approx 0.35 \text{ and } \mathbb{P}[\theta = 3|Y = G] \approx 0.34.$$

The value of θ with the largest posterior is still $\theta = 2$, even though the proportion of the fans of the Orange team in town 2 is only 0.15. Note however, that this probability went down from 0.7. In other words, the new information that $Y = G$ prompted the Bayesian statistician to adjust his or her beliefs about θ quite significantly. The prior information, however, was so strong that even after that adjustment, the value $\theta = 2$ wins, albeit with a much smaller margin[1]

[1] our Bayesian statistician would have been willing to pay more for the bet $\mathbf{1}_{\{\theta=2\}}$ before seeing the data $Y = G$.

12.3 Priors and posteriors when distributions are continuous

We move on and apply the Bayesian ideas to the continuous case which is quite common in practice. The Bayes formula in this case looks pretty much the same as above, with the usual discrete→continuous changes. In particular, we assume that the prior distribution for the unknown parameter θ admits a pdf, which is usually denoted by $p(\theta)$ in Bayesian statistics. The posterior density is traditionally denoted by $p(\theta|y_1,\dots,y_n)$, the likelihood function $L(\theta; y_1,\dots,y_n)$ plays the role of the conditional probability, and the sum in the denominator is replaced by an integral:

$$p(\theta|y_1,\dots,y_n) = \frac{p(\theta)L(\theta|y_1,\dots,y_n)}{\int p(\tilde{\theta})L(\tilde{\theta}|y_1,\dots,y_n)\,d\tilde{\theta}}.$$

We use the notation $\tilde{\theta}$ for the variable of integration, and the integral is always taken from $-\infty$ to ∞; this can be effectively reduced to a smaller domain if $p(\theta)$ takes the value 0 on some subset of $\mathbb{R}$. It is important to note right away that the integral in the denominator does not depend on θ; it is effectively a constant.

Example 12.3.1. Suppose that $Y_1,\dots,Y_n$ is a random sample from the Normal distribution with the unknown mean μ and a known standard deviation $\sigma = 1$. The prior distribution for the parameter μ is assumed to be normal, too, with parameters 0 and 1 (a standard normal). The pdf of the prior distribution is

$$p(\mu) = (2\pi)^{-1/2}e^{-\frac{1}{2}\mu^2},$$

and the likelihood function is given by

$$L(\mu; y_1,\dots,y_n) = (2\pi)^{-n/2}e^{-\frac{1}{2}\sum(y_i-\mu)^2}.$$

Therefore, by the Bayes formula,

$$p(\mu|y_1,\dots,y_n) = \frac{(2\pi)^{-n/2-1/2}e^{-\frac{1}{2}(\mu^2+\sum_{i=1}^n(y_i-\mu)^2)}}{\int(2\pi)^{-n/2-1/2}e^{-\frac{1}{2}(\tilde{\mu}^2+\sum_{i=1}^n(y_i-\tilde{\mu})^2)}\,d\tilde{\mu}}$$

At this point we could go ahead and try to evaluate the integral in the denominator. Alternatively, we could remember that the denominator is not a function of μ (it got "integrated away"), and that it can be taught of as "the constant the numerator must be divided by to obtain a true pdf". In other words, we have

$$p(\mu|y_1,\dots,y_n) = c\,e^{-\frac{1}{2}(\mu^2+\sum(y_i-\mu)^2)}$$

where c does not depend on μ and ensures that $p(\mu|y_1,\ldots,y_n)$ is a true pdf (i.e., integrates to 1). If we rearrange the terms inside the exponential function above, we get

$$\begin{aligned} p(\mu|y_1,\ldots,y_n) &= c\, e^{-\frac{1}{2}\left((n+1)\mu^2 - 2\mu\sum_i y_1 + \sum y_i^2\right)} \\ &= c\, e^{-\frac{1}{2}\sum y_i^2} e^{-\frac{1}{2}(n+1)\left(\mu^2 - 2\mu\frac{\sum_i y_1}{n+1}\right)} \\ &= c\, e^{-\frac{1}{2}\sum y_i^2} e^{-\frac{1}{2}(n+1)\left(\mu^2 - 2\mu\frac{\sum_i y_1}{n+1} + \left(\frac{\sum_i y_i}{n+1}\right)^2\right)} \\ &= \hat{c}\, e^{-\frac{1}{\frac{2}{n+1}}\left(\mu - \frac{1}{n+1}\sum_i y_i\right)^2} \end{aligned}$$

where $\hat{c} = c\, e^{-\frac{1}{2}\sum y_i^2 - \frac{1}{2(n+1)}\sum y_i^2}$ is another constant as far as μ is concerned. We recognize the expression above as the pdf of a normal distribution with mean $\frac{1}{n+1}\sum y_i$ and variance $(n+1)$. This automatically dictates the value of the constant $\hat{c}$. Indeed, it needs to be equal to $(2\pi(n+1))^{-1/2}$; otherwise, $p(\mu; y_1,\ldots,y_n)$ would not be a pdf. Therefore, we conclude

$$p(\mu|y_1,\ldots,y_n) = \frac{1}{\sqrt{2\pi(n+1)}} e^{-\frac{1}{\frac{2}{n+1}}\left(\mu - \frac{1}{n+1}\sum_i y_i\right)^2},$$

i.e., the posterior distribution of μ is normal, with mean $\frac{1}{n+1}\sum y_i$ and variance $n+1$.

It turns out that the general situation is a little harder to analyze, but easier to interpret. A bit more algebra than above shows that if the prior distribution of μ is normal with known parameters μ_{prior} and Σ_{prior}, and if $Y_1,\ldots,Y_n \sim N(\mu,\sigma_0)$, where σ_0 is known, then the posterior distribution of μ is also normal with parameters

$$\mu_{\text{posterior}} = \lambda\bar{y} + (1-\lambda)\mu_{\text{prior}} \quad \text{and} \quad \Sigma^2_{\text{posterior}} = \left(\frac{1}{\Sigma^2_{\text{prior}}} + \frac{1}{\sigma_0^2/n}\right)^{-1},$$

where $\lambda = \frac{\Sigma^2_{\text{prior}}}{\Sigma^2_{\text{prior}} + \sigma_0^2/n}$. The posterior mean is a weighted average of the prior mean and the sample mean, with weights that are inversely proportional to the prior and sample variances. The posterior variance is smaller than both the prior variance and the sample variance. Loosely speaking, the posterior belief about μ centers it somewhere between the prior mean and the sample mean, with a higher degree of certainty than before.

12.4 Statistical inference in the Bayesian framework

In theory, once we combine the data and the prior into a posterior distribution, our job is done - we have a complete description of our state of knowledge (belief) about the unknown parameter θ. In practice, the whole distribution (i.e., its pdf) is often hard to interpret directly and may contain an overwhelming amount of information. For that reason we introduce the Bayesian versions of point and interval estimators. We define them in the continuous case, with the discrete case being completely analogous:

Definition 12.4.1. Let $p(\theta|y_1,\dots,y_n)$ denote the posterior distribution for the unknown parameter θ.

1. The estimator
$$\hat{\theta} = \mathbb{E}^{\text{posterior}}[\theta] = \int \theta p(\theta|y_1,\dots,y_n)\,d\theta$$
is called the **Bayes estimator** for θ.
2. An interval estimator $(\hat{\theta}_L, \hat{\theta}_R)$ is called a (Bayesian) **credible interval of size** $1-\alpha$ if the posterior probability that $\theta \in (\hat{\theta}_L, \hat{\theta}_R)$ is (at least) $1-\alpha$, i.e., if
$$\mathbb{P}^{\text{posterior}}[\hat{\theta}_L \le \theta \le \hat{\theta}_R] = \int_{\hat{\theta}_L}^{\hat{\theta}_R} p(\theta|y_1,\dots,y_n)\,d\theta = 1-\alpha.$$

We illustrate the use of these concepts on a commonly used example. It features a new class of distributions (do not confuse the parameter α below with the significance level α above. They have nothing to do with each other.)

Definition 12.4.2. The continuous distribution with the pdf
$$f(y) = \tfrac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}\mathbf{1}_{[0,1]}(y),$$
where $\alpha,\beta > 0$ are parameters and $B(\alpha,\beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy$, is called the **beta distribution** (denoted by $\text{Beta}(\alpha,\beta)$).

The beta distribution generalizes the uniform distribution ($\alpha=\beta=1$), as well as some other important distributions supported on $(0,1)$. As we will see in the example below, it seems to be tailor-made for modeling proportions or unknown probabilities. The function $B(\alpha,\beta)$, defined so that $\int f(y)\,dy = 1$, is known as the **beta function** and it can be expressed in terms of the gamma function as follows:
$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha,\beta)}.$$

In particular, since $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$, we have $B(\alpha, \beta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} = (\alpha+\beta-1)\binom{\alpha+\beta-2}{\alpha-1}$, as soon as $\alpha, \beta \in \mathbb{N}$.

Both the expectation and the variance of the beta distribution have nice compact explicit forms:

$$\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta}, \ \mathrm{Var}[Y] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Example 12.4.3. Let $Y_1, \ldots, Y_n$ be a sample from the Bernoulli $B(\theta)$-distribution (we use θ instead of the usual p to avoid confusion with the pdfs which are also denoted by p in Bayesian statistics). The prior distribution for p is chosen to be $\mathrm{Beta}(\alpha, \beta)$ for some $\alpha, \beta > 0$. The posterior is then given by

$$p(\theta|y_1, \ldots, y_n) = \tfrac{1}{C} p(\theta) L(\theta; y_1, \ldots, y_n),$$

where $C = \int L(\tilde{\theta}|y_1, \ldots, y_n) p(\tilde{\theta})\, d\tilde{\theta}$ is a constant (as far as θ is concerned). We plug in the expressions for p and L (and remember that the function $B(\alpha, \beta)$ which appears in the pdf of the beta distribution does not contain θ, either):

$$\begin{aligned} p(\theta|y_1, \ldots, y_n) &= \tfrac{1}{CB(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \prod_i \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \tfrac{1}{CB(\alpha,\beta)} \theta^{\alpha-1+\sum_i y_i}(1-\theta)^{\beta-1+n-\sum_i y_i} \end{aligned}$$

As a function of θ, this looks exactly like the density of beta distribution, but with parameters $\alpha' = \alpha + \sum y_i$ and $\beta' = \beta + n - \sum y_i$. Therefore, the constant $CB(\alpha, \beta)$ must be equal to the normalizing constant for the $\mathrm{Beta}(\alpha', \beta')$ distribution, i.e., $B(\alpha', \beta')$.

Now that we have our posterior distribution, we need to compute its expectation to obtain an expression for the Bayes estimator:

$$\hat{\theta} = \tfrac{\alpha'}{\alpha'+\beta'} = \tfrac{\alpha+\sum_i y_i}{\alpha+\beta+n}.$$

The special case where $\alpha = \beta = 0$ corresponds to a so-called **uninformative prior**[1] we obtain $\hat{\theta} = \bar{y}$ - which is both the UMVUE and the MLE in this case. For integer values of α and β, the prior information amounts to adding α ones and β zeros to the sample. It is as if, we already collected a sample of size $\alpha + \beta$ before $y_1, \ldots, y_n$.

To find a credible interval, one needs to be able to compute quantiles of the posterior distribution. For example, when $\alpha = \beta = 0$, $n = 100$ and $\sum y_i = 60$, the posterior distribution is $\mathrm{Beta}(60, 40)$, whose 2.5% and 97.5% quantiles are 0.503 and 0.693, respectively. Therefore, in this case, the 95%-credible interval for θ is $(0.503, 0.693)$.

[1] strictly speaking, there is no Beta distribution with $\alpha = \beta = 0$ since its "density" $y^{-1}(1-y)^{-1}$ integrates to $+\infty$ on $[0, 1]$. What we have in mind is the limiting case $\alpha \to 0$ and $|beta \to 0$.