10.4 Bayesian inference

Interesting results and many new statistical methods can be obtained when we take a rather different look at statistical problems.

The difference is in our treatment of *uncertainty*.

So far, random samples were the only source of uncertainty in all the discussed statistical problems. The only distributions, expectations, and variances considered so far were distributions, expectations, and variances of data and various statistics computed from data. Population parameters were considered fixed. Statistical procedures were based on the distribution of data given these parameters,

$$f(\boldsymbol{x} \mid \theta) = f(X_1, \ldots, X_n \mid \theta).$$

This is the **frequentist approach**. According to it, all probabilities refer to random samples of data and possible long-run frequencies, and so do such concepts as unbiasedness, consistency, confidence level, and significance level:

- an estimator $\widehat{\theta}$ is *unbiased* if in a long run of random samples, it averages to the parameter θ;
- a test has significance level α if in a long run of random samples, $(100\alpha)\%$ of times the true hypothesis is rejected;
- an interval has confidence level $(1-\alpha)$ if in a long run of random samples, $(1-\alpha)100\%$ of obtained confidence intervals contain the parameter, as shown in Figure 9.2, p. 255;
- and so on.

However, there is another approach: the **Bayesian approach**. According to it, uncertainty is also attributed to the unknown parameter θ. Some values of θ are more likely than others. Then, as long as we talk about the likelihood, we can define a whole distribution of values of θ. Let us call it a *prior distribution*. It reflects our ideas, beliefs, and past experiences about the parameter before we collect and use the data.

Example 10.23 (Salaries). What do you think is the average starting annual salary of a Computer Science graduate? Is it \$20,000 per year? Unlikely, that's too low. Perhaps, \$200,000 per year? No, that's too high for a fresh graduate. Between \$40,000 and \$70,000 sounds like a reasonable range. We can certainly collect data on 100 recent graduates, compute their average salary, and use it as an estimate, but before that, we already have our beliefs on what the mean salary may be. We can express it as some distribution with the most likely range between \$40,000 and \$70,000 (Figure 10.9). ◊

Collected data may force us to change our initial idea about the unknown parameter. Probabilities of different values of θ may change. Then we'll have a *posterior distribution* of θ.

One benefit of this approach is that we no longer have to explain our results in terms of a "long run." Often we collect just one sample for our analysis and don't experience any long runs of samples. Instead, with the Bayesian approach, we can state the result in terms of the posterior distribution of θ. For example, we can clearly state the *posterior probability* for a parameter to belong to the obtained confidence interval, or the *posterior probability* that the hypothesis is true.

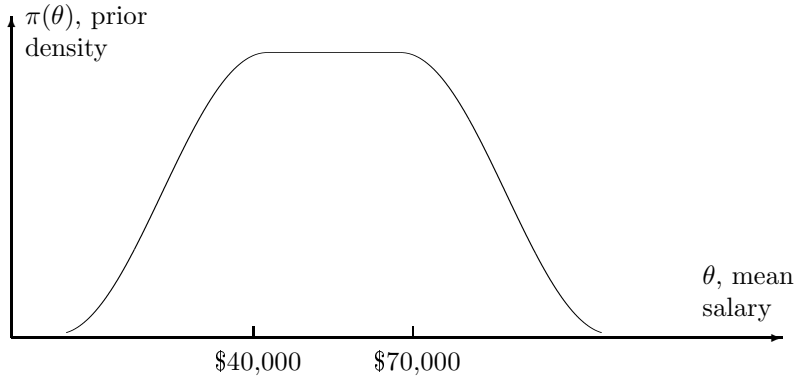

FIGURE 10.9: *Our prior distribution for the average starting salary.*

10.4.1 Prior and posterior

Now we have two sources of information to use in our Bayesian inference:

1. collected and observed data;
2. prior distribution of the parameter.

Here is how these two pieces are combined via the **Bayes formula** (see p. 29 and Figure 10.10).

Prior to the experiment, our knowledge about the parameter θ is expressed in terms of the **prior distribution** (prior pmf or pdf)

$$\pi(\theta).$$

The observed sample of data $\boldsymbol{X} = (X_1, \ldots, X_n)$ has distribution (pmf or pdf)

$$f(\boldsymbol{x}|\theta) = f(x_1, \ldots, x_n|\theta).$$

This distribution is conditional on θ. That is, different values of the parameter θ generate different distributions of data, and thus, conditional probabilities about $\boldsymbol{X}$ generally depend on the condition, θ.

Observed data add information about the parameter. The updated knowledge about θ can be expressed as the **posterior distribution**.

Posterior distribution

$$\boxed{\pi(\theta|\boldsymbol{x}) = \pi(\theta|\boldsymbol{X} = \boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})}.} \qquad (10.12)$$

Posterior distribution of the parameter θ is now conditioned on data $\boldsymbol{X} = \boldsymbol{x}$. Naturally, conditional distributions $f(\boldsymbol{x}|\theta)$ and $\pi(\theta|\boldsymbol{x})$ are related via the Bayes Rule (2.9).

According to the Bayes Rule, the denominator of (10.12), $m(\boldsymbol{x})$, represents the unconditional distribution of data $\boldsymbol{X}$. This is the **marginal distribution** (pmf or pdf) of the sample $\boldsymbol{X}$. Being unconditional means that it is constant for different values of the parameter θ. It can be computed by the *Law of Total Probability* (p. 31) or its continuous-case version below.

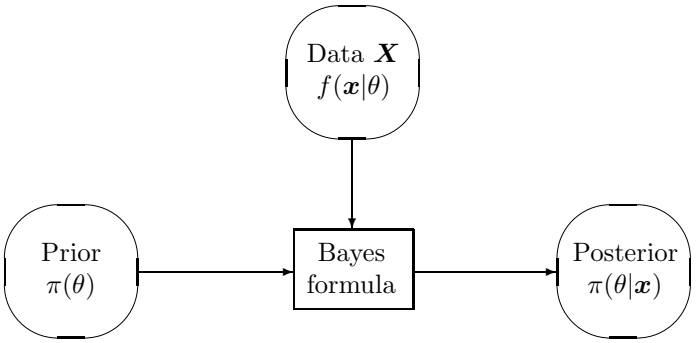

FIGURE 10.10: *Two sources of information about the parameter θ.*

Marginal distribution of data

$$m(\boldsymbol{x}) = \sum_{\theta} f(x|\theta)\pi(\theta)$$
for discrete prior distributions π

$$m(\boldsymbol{x}) = \int_{\theta} f(x|\theta)\pi(\theta)d\theta$$
for continuous prior distributions π (10.13)

Example 10.24 (Quality inspection). A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 10%. We have to decide whether to accept or to reject the shipment based on θ, the proportion of defective parts.

Before we see the real data, let's assign a 50-50 chance to both suggested values of θ, i.e.,

$$\pi(0.05) = \pi(0.10) = 0.5.$$

A random sample of 20 parts has 3 defective ones. Calculate the posterior distribution of θ.

Solution. Apply the Bayes formula (10.12). Given θ, the distribution of the number of defective parts X is Binomial$(n = 20, \theta)$. For $x = 3$, Table A2 gives

$$f(x \mid \theta = 0.05) = F(3 \mid \theta = 0.05) - F(2 \mid \theta = 0.05) = 0.9841 - 0.9245 = 0.0596$$

and

$$f(x \mid \theta = 0.10) = F(3 \mid \theta = 0.10) - F(2 \mid \theta = 0.10) = 0.8670 - 0.6769 = 0.1901.$$

The marginal distribution of X (for $x = 3$) is

$$\begin{aligned} m(3) &= f(x \mid 0.05)\pi(0.05) + f(x \mid 0.10)\pi(0.10) \\ &= (0.0596)(0.5) + (0.1901)(0.5) = 0.12485. \end{aligned}$$

Posterior probabilities of $\theta = 0.05$ and $\theta = 0.10$ are now computed as

$$\pi(0.05 \mid X = 3) = \frac{f(x \mid 0.05)\pi(0.05)}{m(3)} = \frac{(0.0596)(0.5)}{0.1248} = 0.2387;$$

$$\pi(0.10 \mid X = 3) = \frac{f(x \mid 0.10)\pi(0.10)}{m(3)} = \frac{(0.1901)(0.5)}{0.1248} = 0.7613.$$

Conclusion. In the beginning, we had no preference between the two suggested values of θ. Then we observed a rather high proportion of defective parts, 3/20=15%. Taking this into account, $\theta = 0.10$ is now about three times as likely than $\theta = 0.05$. $\Diamond$

NOTATION			
	$\pi(\theta)$	=	prior distribution
	$\pi(\theta \mid \boldsymbol{x})$	=	posterior distribution
	$f(x\|\theta)$	=	distribution of data (model)
	$m(\boldsymbol{x})$	=	marginal distribution of data
	$\boldsymbol{X}$	=	$(X_1, \ldots, X_n)$, sample of data
	$\boldsymbol{x}$	=	$(x_1, \ldots, x_n)$, observed values of $X_1, \ldots, X_n$.

Conjugate distribution families

A suitably chosen prior distribution of θ may lead to a very tractable form of the posterior.

DEFINITION 10.4

A family of prior distributions π is **conjugate** to the model $f(\boldsymbol{x}|\theta)$ if the posterior distribution belongs to the same family.

Three classical examples of conjugate families are given below.

Gamma family is conjugate to the Poisson model

Let $(X_1, \ldots, X_n)$ be a sample from Poisson(θ) distribution with a Gamma(α, λ) prior distribution of θ.

Then

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \sim e^{-n\theta}\theta^{\sum x_i}. \tag{10.14}$$

Remark about dropping constant coefficients. In the end of (10.14), we dropped $(x_i!)$ and wrote that the result is "proportional" ($\sim$) to $e^{-n\theta}\theta^{\sum x_i}$. Dropping terms that don't contain θ often simplifies the computation. The form of the posterior distribution can be obtained without the constant term, and if needed, we can eventually evaluate the normalizing constant in the end, making $\pi(\theta|\boldsymbol{x})$ a fine distribution with the total probability 1, for example, as we did in Example 4.1 on p. 77. In particular, the marginal distribution $m(\boldsymbol{x})$ can be

dropped because it is θ-free. But keep in mind that in this case we obtain the posterior distribution "up to a constant coefficient."

The Gamma prior distribution of θ has density

$$\pi(\theta) \sim \theta^{\alpha-1} e^{-\lambda\theta}.$$

As a function of θ, this prior density has the same form as the model $f(\boldsymbol{x}|\theta)$ – a power of θ multiplied by an exponential function. This is the general idea behind conjugate families.

Then, the posterior distribution of θ given $\boldsymbol{X} = \boldsymbol{x}$ is

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &\sim f(\boldsymbol{x}|\theta)\pi(\theta) \\ &\sim \left(e^{-n\theta}\theta^{\sum x_i}\right)\left(\theta^{\alpha-1}e^{-\lambda\theta}\right) \\ &\sim \theta^{\alpha+\sum x_i-1} e^{-(\lambda+n)\theta}. \end{aligned}$$

Comparing with the general form of a Gamma density (say, (4.7) on p. 85), we see that $\pi(\theta|\boldsymbol{x})$ is the Gamma distribution with new parameters,

$$\alpha_x = \alpha + \sum_{i=1}^{n} x_i \text{ and } \lambda_x = \lambda + n.$$

We can conclude that:

1. Gamma family of prior distributions is conjugate to Poisson models.
2. Having observed a Poisson sample $\boldsymbol{X} = \boldsymbol{x}$, we update the Gamma$(\alpha, \lambda)$ prior distribution of θ to the Gamma$(\alpha + \sum x_i, \lambda + n)$ posterior.

Gamma distribution family is rather rich; it has two parameters. There is often a good chance to find a member of this large family that suitably reflects our knowledge about θ.

Example 10.25 (Network blackouts). The number of network blackouts each week has Poisson(θ) distribution. The weekly rate of blackouts θ is not known exactly, but according to the past experience with similar networks, it averages 4 blackouts with a standard deviation of 2.

There exists a Gamma distribution with the given mean $\mu = \alpha/\lambda = 4$ and standard deviation $\sigma = \sqrt{\alpha}/\lambda = 2$. Its parameters α and λ can be obtained by solving the system,

$$\left\{ \begin{array}{rcl} \alpha/\lambda & = & 4 \\ \sqrt{\alpha}/\lambda & = & 2 \end{array} \right. \Rightarrow \left\{ \begin{array}{rcccl} \alpha & = & (4/2)^2 & = & 4 \\ \lambda & = & 2^2/4 & = & 1 \end{array} \right.$$

Hence, we can assume the Gamma$(\alpha = 4,\ \lambda = 1)$ prior distribution θ. It is convenient to have a conjugate prior because the posterior will then belong to the Gamma family too.

Suppose there were $X_1 = 2$ blackouts this week. Given that, the posterior distribution of θ is Gamma with parameters

$$\alpha_x = \alpha + 2 = 6, \quad \lambda_x = \lambda + 1 = 2.$$

If no blackouts occur during the next week, the updated posterior parameters become

$$\alpha_x = \alpha + 2 + 0 = 6, \quad \lambda_x = \lambda + 2 = 3.$$

This posterior distribution has the average weekly rate of $6/3 = 2$ blackouts per week. Two weeks with very few blackouts reduced our estimate of the average rate from 4 to 2. $\diamondsuit$

Beta family is conjugate to the Binomial model

A sample from Binomial(k, θ) distribution (assume k is known) has the probability mass function

$$f(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} \binom{k}{x_i} \theta^{x_i}(1-\theta)^{k-x_i} \sim \theta^{\sum x_i}(1-\theta)^{nk-\sum x_i}.$$

Density of Beta(α, β) prior distribution has the same form, as a function of θ,

$$\pi(\theta) \sim \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } 0 < \theta < 1$$

(see Section 12.2.2 in the Appendix). Then, the posterior density of θ is

$$\pi(\theta \mid \boldsymbol{x}) \sim f(\boldsymbol{x} \mid \theta)\pi(\theta) \sim \theta^{\alpha+\sum_{i=1}^{n} x_i - 1}(1-\theta)^{\beta+nk-\sum_{i=1}^{n} x_i - 1},$$

and we recognize the Beta density with new parameters

$$\alpha_x = \alpha + \sum_{i=1}^{n} x_i \text{ and } \beta_x = \beta + nk - \sum_{i=1}^{n} x_i.$$

Hence,

1. Beta family of prior distributions is conjugate to the Binomial model.
2. Posterior parameters are $\alpha_x = \alpha + \sum x_i$ and $\beta_x = \beta + nk - \sum x_i$.

Normal family is conjugate to the Normal model

Consider now a sample from Normal distribution with an unknown mean θ and a known variance σ^2:

$$\begin{aligned} f(\boldsymbol{x} \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\} \sim \exp\left\{-\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2\sigma^2}\right\} \\ &\sim \exp\left\{\theta\frac{\sum x_i}{\sigma^2} - \theta^2\frac{n}{2\sigma^2}\right\} = \exp\left\{\left(\theta\overline{X} - \frac{\theta^2}{2}\right)\frac{n}{\sigma^2}\right\}. \end{aligned}$$

If the prior distribution of θ is also Normal, with prior mean μ and prior variance τ^2, then

$$\pi(\theta) \sim \exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\} \sim \exp\left\{\left(\theta\mu - \frac{\theta^2}{2}\right)\frac{1}{\tau^2}\right\},$$

and again, it has a similar form as $f(\boldsymbol{x}|\theta)$.

The posterior density of θ equals

$$\begin{aligned} \pi(\theta \mid \boldsymbol{x}) &\sim f(\boldsymbol{x} \mid \theta)\pi(\theta) \sim \exp\left\{\theta\left(\frac{n\overline{X}}{\sigma^2} + \frac{\mu}{\tau^2}\right) - \frac{\theta^2}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\right\} \\ &\sim \exp\left\{-\frac{(\theta-\mu_x)^2}{2\tau_x^2}\right\}, \end{aligned}$$

where

$$\mu_x = \frac{n\overline{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} \text{ and } \tau_x^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}. \tag{10.15}$$

This posterior distribution is certainly Normal with parameters μ_x and τ_x.

We can conclude that:

1. Normal family of prior distributions is conjugate to the Normal model with unknown mean;
2. Posterior parameters are given by (10.15).

We see that the posterior mean μ_x is a weighted average of the prior mean μ and the sample mean $\overline{X}$. This is how the prior information and the observed data are combined in case of Normal distributions.

How will the posterior mean behave when it is computed from a large sample? As the sample size n increases, we get more information from the data, and as a result, the frequentist estimator will dominate. According to (10.15), the posterior mean converges to the sample mean $\overline{X}$ as $n \to \infty$.

Posterior mean will also converge to $\overline{X}$ when $\tau \to \infty$. Large τ means a lot of uncertainty in the prior distribution of θ; thus, naturally, we should rather use observed data as a more reliable source of information in this case.

On the other hand, large σ indicates a lot of uncertainty in the observed sample. If that is the case, the prior distribution is more reliable, and as we see in (10.15), $\mu_x \approx \mu$ for large σ.

Results of this section are summarized in Table 10.2. You will find more examples of conjugate prior distributions among the exercises.

Model $f(\boldsymbol{x}\|\theta)$	Prior $\pi(\theta)$	Posterior $\pi(\theta\|\boldsymbol{x})$
Poisson(θ)	Gamma(α, λ)	Gamma($\alpha + n\overline{X}$, $\lambda + n$)
Binomial(k, θ)	Beta(α, β)	Beta($\alpha + n\overline{X}$, $\beta + n(k - \overline{X})$)
Normal(θ, σ)	Normal(μ, τ)	Normal$\left(\dfrac{n\overline{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \dfrac{1}{\sqrt{n/\sigma^2 + 1/\tau^2}}\right)$

TABLE 10.2: *Three classical conjugate families.*

10.4.2 Bayesian estimation

We have already completed the most important step in Bayesian inference. We obtained the posterior distribution. All the knowledge about the unknown parameter is now included in the posterior, and that is what we'll use for further statistical analysis (Figure 10.11).

To estimate θ, we simply compute the **posterior mean**,

$$\widehat{\theta}_{\mathrm{B}} = \mathbf{E}\{\theta|\boldsymbol{X} = \boldsymbol{x}\} = \begin{cases} \displaystyle\sum_{\theta} \theta\pi(\theta|\boldsymbol{x}) = \frac{\sum \theta f(\boldsymbol{x}|\theta)\pi(\theta)}{\sum f(\boldsymbol{x}|\theta)\pi(\theta)} & \text{if } \theta \text{ is discrete} \\ \displaystyle\int_{\theta} \theta\pi(\theta|\boldsymbol{x})d\theta = \frac{\int \theta f(\boldsymbol{x}|\theta)\pi(\theta)d\theta}{\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta} & \text{if } \theta \text{ is continuous} \end{cases}$$

The result is a conditional expectation of θ given data $\boldsymbol{X}$. In abstract terms, the **Bayes estimator** $\widehat{\theta}_{\mathrm{B}}$ is what we "expect" θ to be, after we observed a sample.

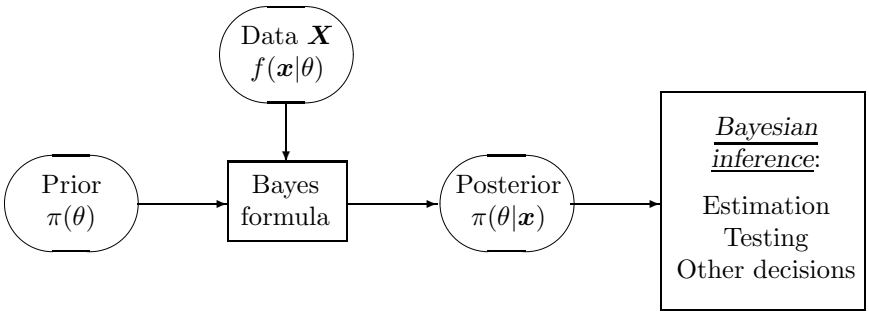

FIGURE 10.11: *Posterior distribution is the basis for Bayesian inference.*

How accurate is this estimator? Among all estimators, $\widehat{\theta}_{\mathrm{B}} = \mathbf{E}\{\theta|\boldsymbol{x}\}$ has the lowest squared-error **posterior risk**

$$\rho(\widehat{\theta}) = \mathbf{E}\left\{(\widehat{\theta}-\theta)^2 \mid \boldsymbol{X}=\boldsymbol{x}\right\}.$$

For the Bayes estimator $\widehat{\theta}_{\mathrm{B}} = \mathbf{E}\{\theta \mid \boldsymbol{x}\}$, posterior risk equals **posterior variance**,

$$\rho(\widehat{\theta}) = \mathbf{E}\left\{(\mathbf{E}\{\theta|\boldsymbol{x}\}-\theta)^2 \mid \boldsymbol{x}\right\} = \mathbf{E}\left\{(\theta - \mathbf{E}\{\theta|\boldsymbol{x}\})^2 \mid \boldsymbol{x}\right\} = \mathrm{Var}\{\theta|\boldsymbol{x}\},$$

which measures variability of θ around $\widehat{\theta}_{\mathrm{B}}$, according to the posterior distribution of θ.

Example 10.26 (Normal case). The Bayes estimator of the mean θ of Normal(θ,σ) distribution with a Normal(μ,τ) prior is

$$\widehat{\theta}_{\mathrm{B}} = \mu_x = \frac{n\overline{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2},$$

and its posterior risk is

$$\rho(\widehat{\theta}_{\mathrm{B}}) = \tau_x^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}$$

(Table 10.2). As we expect, this risk decreases to 0 as the sample size grows to infinity. ◊

Example 10.27 (Network blackouts, continued). After two weeks of data, the weekly rate of network blackouts, according to Example 10.25 on p. 356, has Gamma posterior distribution with parameters $\alpha_x = 6$ and $\lambda_x = 3$.

The Bayes estimator of the weekly rate θ is

$$\widehat{\theta}_{\mathrm{B}} = \mathbf{E}\{\theta|\boldsymbol{x}\} = \frac{\alpha_x}{\lambda_x} = 2 \text{ (blackouts per week)}$$

with a posterior risk

$$\rho(\widehat{\theta}_{\mathrm{B}}) = \mathrm{Var}\{\theta|\boldsymbol{x}\} = \frac{\alpha_x}{\lambda_x^2} = \frac{2}{3}.$$

◊

Although conjugate priors simplify our statistics, Bayesian inference can certainly be done for other priors too.

Example 10.28 (Quality inspection, continued). In Example 10.24 on p. 354, we computed posterior distribution of the proportion of defective parts θ. This was a discrete distribution,

$$\pi(0.05 \mid \boldsymbol{x}) = 0.2387; \quad \pi(0.10 \mid \boldsymbol{x}) = 0.7613.$$

Now, the Bayes estimator of θ is

$$\widehat{\theta}_{\mathrm{B}} = \sum_{\theta} \theta \pi(\theta \mid \boldsymbol{x}) = (0.05)(0.2387) + (0.10)(0.7613) = 0.0881.$$

It does not agree with the manufacturer (who claims $\theta = 0.05$) or with the quality inspector (who feels that $\theta = 0.10$) but its value is much closer to the inspector's estimate.

The posterior risk of $\widehat{\theta}_{\mathrm{B}}$ is

$$\begin{aligned} \mathrm{Var}\,\{\theta|\boldsymbol{x}\} &= \mathbf{E}\left\{\theta^2|\boldsymbol{x}\right\} - \mathbf{E}^2\,\{\theta|\boldsymbol{x}\} \\ &= (0.05)^2(0.2387) + (0.10)^2(0.7613) - (0.0881)^2 = 0.0004, \end{aligned}$$

which means a rather low posterior standard deviation of 0.02. ◊

10.4.3 Bayesian credible sets

Confidence intervals have a totally different meaning in Bayesian analysis. Having a posterior distribution of θ, we no longer have to explain the confidence level $(1-\alpha)$ in terms of a long run of samples. Instead, we can give an interval $[a, b]$ or a set C that has a posterior probability $(1-\alpha)$ and state that *the parameter θ belongs to this set with probability* $(1-\alpha)$. Such a statement was impossible before we considered prior and posterior distributions. This set is called a $(1-\alpha)100\%$ *credible set*.

DEFINITION 10.5

Set C is a $(1-\alpha)100\%$ **credible set** for the parameter θ if the posterior probability for θ to belong to C equals $(1-\alpha)$. That is,

$$\boldsymbol{P}\{\theta \in C \mid \boldsymbol{X} = \boldsymbol{x}\} = \int_C \pi(\theta|\boldsymbol{x}) d\theta = 1 - \alpha.$$

Such a set is not unique. Recall that for two-sided, left-tail, and right-tail hypothesis testing, we took different portions of the area under the Normal curve, all equal $(1-\alpha)$.

Minimizing the length of set C among all $(1-\alpha)100\%$ credible sets, we just have to include all the points θ with a high posterior density $\pi(\theta|\boldsymbol{x})$,

$$C = \{\theta : \ \pi(\theta|\boldsymbol{x}) \geq c\}$$

(see Figure 10.12). Such a set is called the **highest posterior density credible set**, or just the **HPD set**.

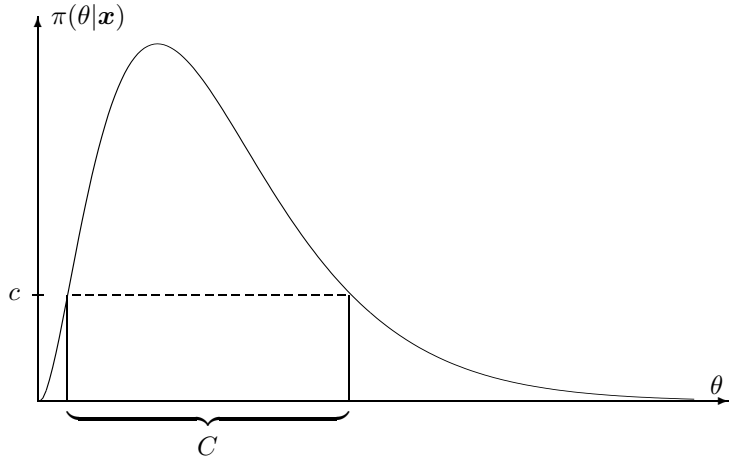

FIGURE 10.12: *The $(1-\alpha)100\%$ highest posterior density credible set.*

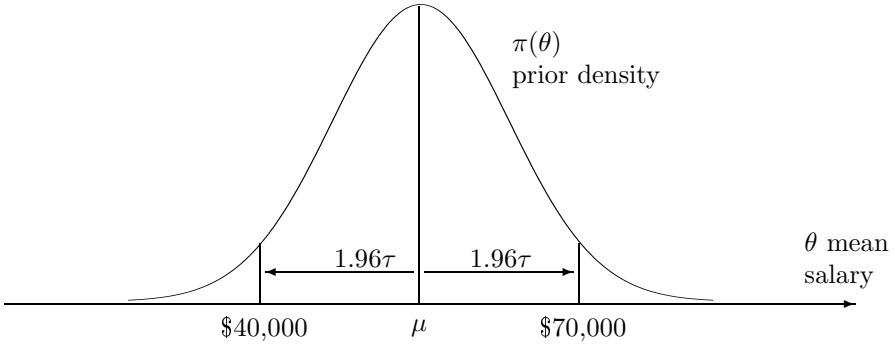

FIGURE 10.13: *Normal prior distribution and the 95% HPD credible set for the mean starting salary of Computer Science graduates (Example 10.29).*

For the Normal(μ_x, τ_x) posterior distribution of θ, the $(1-\alpha)100\%$ HPD set is

$$\mu_x \pm z_{\alpha/2}\tau_x = [\mu_x - z_{\alpha/2}\tau_x, \mu_x + z_{\alpha/2}\tau_x].$$

Example 10.29 (SALARIES, CONTINUED). In Example 10.23 on p. 352, we "decided" that the most likely range for the mean starting salary θ of Computer Science graduates is between \$40,000 and \$70,000. Expressing this in a form of a prior distribution, we let the prior mean be $\mu = (40,000 + 70,000)/2 = 55,000$. Further, if we feel that the range $[40,000; 70,000]$ is 95% likely, and we accept a Normal prior distribution for θ, then this range should be equal

$$[40,000; 70,000] = \mu \pm z_{0.025/2}\tau = \mu \pm 1.96\tau,$$

where τ is the prior standard deviation (Figure 10.13). We can now evaluate the prior standard deviation parameter τ from this information,

$$\tau = \frac{70,000 - 40,000}{2(1.96)} = 7,653.$$

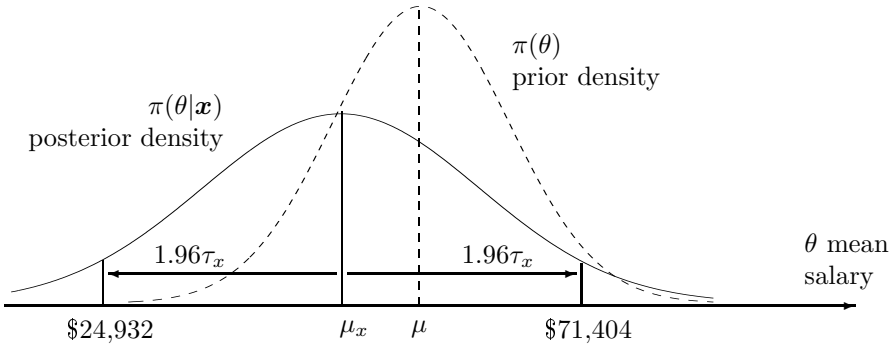

FIGURE 10.14: *Normal prior and posterior distributions for the mean starting salary (Example 10.29).*

This is the advantage of using a rich (two-parameter) family of prior distributions: we are likely to find a member of this family that reflects our prior beliefs adequately.

Then, *prior to collecting any data*, the 95% HPD credible set of the mean starting salary θ is

$$\mu \pm z_{0.025}\tau = \underline{[40{,}000;\ 70{,}000]}.$$

Suppose a random sample of 100 graduates has the mean starting salary $\overline{X} = 48,000$ with a sample standard deviation $s = 12,000$. From Table 10.2, we determine the posterior mean and standard deviation,

$$\begin{aligned}
\mu_x &= \frac{n\overline{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} = \frac{(100)(48,000)/(12,000)^2 + (55,000)/(7,653)^2}{100/(12,000)^2 + 1/(7653)^2} \\
&= 48,168; \\
\tau_x &= \frac{1}{\sqrt{n/\sigma^2 + 1/\tau^2}} = \frac{1}{\sqrt{100/(12,000)^2 + 1/(7653)^2}} = 11,855.
\end{aligned}$$

We used the sample standard deviation s in place of the population standard deviation σ assuming that a sample of size 100 estimates the latter rather accurately. Alternatively, we could put a prior distribution on unknown σ too and estimate it by Bayesian methods. Since the observed sample mean is smaller than our prior mean, the resulting posterior distribution is shifted to the left of the prior (Figure 10.14).

Conclusion. After seeing the data, the Bayes estimator for the mean starting salary of CS graduates is

$$\widehat{\theta}_{\mathrm{B}} = \mu_x = 48,168 \text{ dollars},$$

and the 95% HPD credible set for this mean salary is

$$\mu_x \pm z_{0.025}\tau_x = 48,168 \pm (1.96)(11,855) = 48,168 \pm 23,236 = \underline{[24{,}932;\ 71{,}404]}$$

Lower observed salaries than the ones predicted a priori extended the lower end of our credible interval. $\Diamond$

Example 10.30 (TELEPHONE COMPANY). A new telephone company predicts to handle an average of 1000 calls per hour. During 10 randomly selected hours of operation, it handled a total of 7265 calls.

How should it update the initial estimate of the frequency of telephone calls? Construct a 95% HPD credible set. Telephone calls are placed according to a Poisson process. The hourly rate of calls has an Exponential prior distribution.

Solution. We need a Bayesian estimator of the frequency θ of telephone calls. The number of calls during 1 hour has Poisson(θ) distribution, where θ is unknown, with

$$\text{Exponential}(\lambda) = \text{Gamma}(1, \lambda)$$

prior distribution that has an expectation of

$$\mathbf{E}(\theta) = \frac{1}{\lambda} = 1000 \text{ calls.}$$

Hence, $\lambda = 0.001$. We observe a sample of size $n = 10$, totaling

$$\sum_{i=1}^{n} X_i = n\overline{X} = 7265 \text{ calls.}$$

As we know (see Table 10.2 on p. 358), the posterior distribution in this case is Gamma(α_x, λ_x) with

$$\begin{aligned} \alpha_x &= \alpha + n\overline{X} = 7266, \\ \lambda_x &= \lambda + n = 10.001. \end{aligned}$$

This distribution has mean

$$\mu_x = \alpha_x/\lambda_x = 726.53$$

and standard deviation

$$\tau_x = \sqrt{\alpha_x}/\lambda_x = 8.52.$$

The Bayes estimator of θ is

$$\mathbf{E}(\theta|\boldsymbol{X}) = \mu_x = \underline{726.53 \text{ calls per hour}}.$$

It almost coincides with the sample mean $\overline{X}$ showing that the sample was informative enough to dominate over the prior information.

For the credible set, we notice that α_x is sufficiently large to make the Gamma posterior distribution approximately equal the Normal distribution with parameters μ_x and τ_x. The 95% HPD credible set is then

$$\mu_x \pm z_{0.05/2}\tau_x = 726.53 \pm (1.96)(8.52) = 726.53 \pm 16.70 = \underline{[709.83,\ 743.23]}$$

$\Diamond$

10.4.4 Bayesian hypothesis testing

Bayesian hypothesis testing is very easy to interpret. We can compute prior and posterior probabilities for the hypothesis H_0 and alternative H_A to be true and decide from there which one to accept or to reject.

Computing such probabilities was not possible without prior and posterior distributions of the parameter θ. In non-Bayesian statistics, θ was not random, thus H_0 and H_A were either true (with probability 1) or false (with probability 1).

For Bayesian tests, in order for H_0 to have a meaningful, non-zero probability, it often represents a set of parameter values instead of just one θ_0, and we are testing

$$H_0: \ \theta \in \Theta_0 \ \text{ vs } \ H_A: \ \theta \in \Theta_1.$$

This actually makes sense because exact equality $\theta = \theta_0$ is unlikely to hold anyway, and in practice it is understood as $\theta \approx \theta_0$.

Comparing posterior probabilities of H_0 and H_A,

$$\boldsymbol{P}\{\Theta_0 \mid \boldsymbol{X} = \boldsymbol{x}\} \ \text{ and } \ \boldsymbol{P}\{\Theta_1 \mid \boldsymbol{X} = \boldsymbol{x}\},$$

we decide whether $\boldsymbol{P}\{\Theta_1 \mid \boldsymbol{X} = \boldsymbol{x}\}$ is large enough to present significant evidence and to reject the null hypothesis. One can again compare it with $(1-\alpha)$ such as 0.90, 0.95, 0.99, or state the result in terms of likelihood, "the null hypothesis is this much likely to be true."

Example 10.31 (Telephone company, continued). Let us test whether the telephone company in Example 10.30 can actually face a call rate of 1000 calls *or more* per hour. We are testing

$$H_0: \ \theta \geq 1000 \ \text{ vs } \ H_A: \ \theta < 1000,$$

where θ is the hourly rate of telephone calls.

According to the Gamma(α_x, λ_x) posterior distribution of θ and its Normal $(\mu_x = 726.53, \tau_x = 72.65)$ approximation,

$$\boldsymbol{P}\{H_0 \mid \boldsymbol{X} = \boldsymbol{x}\} = \boldsymbol{P}\left\{\frac{\theta - \mu_x}{\tau_x} \geq \frac{1000 - \mu_x}{\tau_x}\right\} = 1 - \Phi(3.76) = 0.0001.$$

By the complement rule, $\boldsymbol{P}\{H_A \mid \boldsymbol{X} = \boldsymbol{x}\} = 0.9999$, and this presents sufficient evidence against H_0.

We conclude that it's extremely unlikely for this company to face a frequency of 1000+ calls per hour. ◊

Loss and risk

Often one can anticipate the consequences of Type I and Type II errors in hypothesis testing and assign a **loss** $L(\theta, a)$ associated with each possible error. Here θ is the parameter, and a is our action, the decision on whether we accept or reject the null hypothesis.

Each decision then has its **posterior risk** $\rho(a)$, defined as the expected loss computed under the posterior distribution. The action with the lower posterior risk is our **Bayes action**.

Suppose that the Type I error causes the loss

$$w_0 = Loss(\text{Type I error}) = L(\Theta_0, \text{reject } H_0),$$

and the Type II error causes the loss

$$w_1 = Loss(\text{Type II error}) = L(\Theta_1, \text{accept } H_0).$$

Posterior risks of each possible action are then computed as

$$\begin{aligned} \rho(\text{reject } H_0) &= w_0 \pi(\Theta_0 \mid \boldsymbol{x}), \\ \rho(\text{accept } H_0) &= w_1 \pi(\Theta_1 \mid \boldsymbol{x}). \end{aligned}$$

Now we can determine the Bayes action. If $w_0\pi(\Theta_0 \mid \boldsymbol{x}) \le w_1\pi(\Theta_1 \mid \boldsymbol{x})$, the Bayes action is to accept H_0. If $w_0\pi(\Theta_0 \mid \boldsymbol{x}) \ge w_1\pi(\Theta_1 \mid \boldsymbol{x})$, the Bayes action is to reject H_0.

Example 10.32 (Quality inspection, continued). In Example 10.24 on p. 354, we are testing

$$H_0: \ \theta = 0.05 \ \text{ vs } \ H_A: \ \theta = 0.10$$

for the proportion θ of defective parts. Suppose that the Type I error is three times as costly here as the Type II error. What is the Bayes action in this case?

Example 10.28 gives posterior probabilities

$$\pi(\Theta_0 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.2387 \text{ and } \pi(\Theta_1 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.7613.$$

Since $w_0 = 3w_1$, the posterior risks are

$$\begin{aligned} \rho(\text{reject } H_0) &= w_0 \pi(\Theta_0 \mid \boldsymbol{x}) = 3w_1(0.2387) = 0.7161 w_1, \\ \rho(\text{accept } H_0) &= w_1 \pi(\Theta_1 \mid \boldsymbol{x}) = 0.7613 w_1. \end{aligned}$$

Thus, rejecting H_0 has a lower posterior risk, and therefore, it is the Bayes action. Reject H_0. ◊

Summary and conclusions

A number of popular methods of Statistical Inference are presented in this chapter.

Chi-square tests represent a general technique based on counts. Comparing the observed counts with the counts expected under the null hypothesis through the chi-square statistic, one can test for the goodness of fit and for the independence of two factors. Contingency tables are widely used for the detection of significant relations between categorical variables.

Nonparametric statistical methods are not based on any particular distribution of data. So, they are often used when the distribution is unknown or complicated. They are also very handy when the sample may contain some outliers, and even when the data are not numerical.