STA 114: Statistics

Notes 13. Prediction

Consider data X modeled as $X \sim f(x|\theta)$, $\theta \in \Theta$. Suppose we want to predict an unobserved quantity X^*, which depends on the same parameter θ, based on an observation $X = x$.

Example (Hurricane counts)**.** Based on count data $X = (X_1, \cdots, X_n)$ from n consecutive years, we might be interested in forecasting the number of TCs X_{n+1} in the coming year. Here $X^* = X_{n+1}$ and a reasonable model is $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Poisson}(\mu)$, $X^* = X_{n+1} \sim \mathsf{Poisson}(\mu)$ and X and X^* are independent, where $\mu \in (0, \infty)$ is an unknown model parameter.

Example (Hurricant counts (contd.))**.** In the same setting, we might be interested in whether the next year's count exceeds a certain cut-off mark, say 15. In this case the variable of interest is the binary variable X^*, with $X^* = 1$ when $X_{n+1} > 15$ and $X^* = 0$ when $X_{n+1} \leq 15$, where X_{n+1} is the count for the coming year. Borrowing from the description of X and X_{n+1} above, we can describe X and X^* as: $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Poisson}(\mu)$, $X_{n+1} \sim \mathsf{Bernoulli}(p(\mu))$ where $p(\mu) = \sum_{k>15} e^{-\mu}\mu^k/k!$, and X and X^* are independent.

Example (Food expenditure)**.** Suppose we collect data from n Duke undergraduates on their (average) weekly expenditure on food $X_1, \cdots, X_n$. We might be interested in predicting $X^* = X_{n+1}$, the (average) amount a (hypothetical) future student is likely to pay on food per week. We can model $X_1, \cdots, X_n, X_{n+1} \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, with $(\mu, \sigma^2) \in (-\infty, \infty) \times (0, \infty)$ as unknown model parameters.

Example (Food expenditure (contd.))**.** We might also be interested in predicting the difference $X^* = X_{n+1} - X_{n+2}$ in expenditures for two (hypothetical) future students. If we model $X_1, \cdots, X_n, X_{n+1}, X_{n+2} \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, then we have the following model on X and X^*: $X_1, \cdots, X_n \sim \mathsf{Normal}(\mu, \sigma^2)$, $X^* \sim \mathsf{Normal}(0, 2\sigma^2)$, X and X^* independent.

From the above examples it is clear that we are discussing prediction of a variable X^*, given observation on data X in the following context: $X \sim f(x|\theta)$, $X^* \sim f^*(x^*|\theta)$, $\theta \in \Theta^*$, for some collections of pdfs/pmfs $f(x|\theta)$ and $f^*(x^*|\theta)$ indexed by a common parameter $\theta \in \Theta$. Also note that in all of the above examples the model parameters are not real, physical quantities that we could measure if we had more resources (unlike the opinion poll example where the parameter is the actual proportion of supporters, a measurable quantity). For such examples, prediction might be a more useful data analysis task than inference on the model parameters. Below we discussion classical and Bayesian approaches to prediction.

Classical approach

The main vehicle of prediction in classical statistics is the so-called plug-in approach. Suppose we obtain an estimate $\hat{\theta}(x)$ of θ from observation $X = x$ (based on ML or other considerations). Then the predictive description of X^* given $X = x$ is the pdf/pmf $\hat{f}^*(x^*|x) = f^*(x^*|\hat{\theta}(x))$. Although this is a reasonable approach, there is one difficulty. We essentially took the point summary $\hat{\theta}(x)$ to capture all uncertainty about θ. This goes against our intuition of uncertainty associated with statistical modeling that encouraged us to consider interval summaries over point summaries.

This difficulty can be explored formally as follows. Consider the model $X_1, \cdots, X_n, X_{n+1} \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, σ^2 fixed and $X^* = X_{n+1}$. The ML plug-in predictive distribution X^* given $X = x$ is $\mathsf{Normal}(\bar{x}, \sigma^2)$. A 95% central interval for this distribution is $\bar{x} \mp 1.96\sigma/\sqrt{n}$. But does this interval guarantee a 95% coverage of capturing X^*? [Recall Practice Problem #5]. We expect the answer to be "no" because we are not accounting for the uncertainty about μ estimated by $\hat{\mu}_{\text{MLE}}(x) = \bar{x}$. By simple calculations:

$$P_{[X,X^*|\mu]}(X^* \in \bar{X} \mp 1.96\sigma) = P_{[X,X^*|\mu]}(X^* - \bar{X} \in \mp 1.96\sigma) = P_{[Z|\mu]}\left(Z \in \mp \frac{1.96}{\sqrt{1+1/n}}\right)$$

where $Z = \frac{X^* - \bar{X}}{\sigma\sqrt{1+1/n}}$. Now, by our model on X and X^*, for any μ, $X^* \sim \mathsf{Normal}(\mu, \sigma^2)$, $\bar{X} \sim \mathsf{Normal}(\mu, \sigma^2/n)$ and they are independent. So for any μ, $X^* - \bar{X} \sim \mathsf{Normal}(0, \sigma^2 + \sigma^2/n)$ and hence $Z \sim \mathsf{Normal}(0, 1)$. So the above coverage probability is $2\Phi(1.96/\sqrt{1+1/n}) - 1$. For $n = 5$, this equals 93% rather than the touted coverage of 95%.

For the normal models (both known and unknown σ^2), this loss of coverage is easy to fix. In this above example, to get 95% coverage probability, we should use $\bar{x} \mp 1.96\sigma\sqrt{1+1/n}$. More generally, the predictive interval $\bar{x} \mp z(\alpha)\sigma\sqrt{1+1/n}$ yields a $100(1-\alpha)\%$ coverage of X^*. For the unknown σ^2 model, the same holds for the interval $\bar{x} \mp z_{n-1}(\alpha)s_x\sqrt{1+1/n}$.

However, such fixes are not generally available for non-normal models. Calculating the coverage can be a challenging task. Even normal approximations to the MLE may not salvage the situation, because we also need to account for X^*. However, simulations techniques (as we saw in labs) can be used to approximate coverage probabilities of a given predictive interval procedure.

Bayesian approach

Prediction under a Bayesian formulation is conceptually very straightforward. Suppose we have $X \sim f(x|\theta)$, $X^* \sim f^*(x^*|\theta)$, X and X^* independent, and $\theta \in \Theta$ is assigned a prior $\xi(\theta)$. Then we can talk about the joint plausibility scores of the triplet (X, X^*, θ) via the pdf/pmf/...

$$g(x, x^*, \theta) = f(x|\theta)f^*(x^*|\theta)\xi(\theta), \quad x \in S, x^* \in S^*, \theta \in \Theta.$$

To see why this is the case, assume for the moment that the spaces S, S^* and Θ are all discrete so that $f(x|\theta)$, $f^*(x^*|\theta)$ and $\xi(\theta)$ are all pmfs. Then,

$$g(x, x^*, \theta) = P(X = x, X^* = x^*|\theta)\xi(\theta) = P(X = x|\theta)P(X^* = x^*|\theta)\xi(\theta) = f(x|\theta)f^*(x^*|\theta)\xi(\theta)$$

where the middle inequality follows because X, X^* are independent given θ.

From the joint plausibility scores on (X, X^*, θ) we can extract conditional joint scores of (X^*, θ) given $X = x$ by the pdf/pmf/...

$$h^*(x^*, \theta|x) = f(x^*|\theta)\xi(\theta|x), \quad x^* \in S^*, \theta \in \Theta.$$

To see this, again working with pmfs, argue that

$$\begin{aligned} h^*(x^*, \theta|x) &= \frac{g(x, x^*, \theta)}{\sum_{\tilde{\theta}\in\Theta}\sum_{\tilde{x}^*\in S^*} g(x, \tilde{x}^*, \tilde{\theta})} \\ &= \frac{f(x|\theta)f^*(x^*|\theta)\xi(\theta)}{\sum_{\tilde{\theta}\in\Theta}\{\sum_{\tilde{x}^*\in S^*} f^*(x^*|\tilde{\theta})\}f(x|\tilde{\theta})\xi(\tilde{\theta})} \\ &= \frac{f^*(x^*|\theta)f(x|\theta)\xi(\theta)}{\sum_{\tilde{\theta}\in\Theta} f(x|\tilde{\theta})\xi(\tilde{\theta})} \\ &= f^*(x^*|\theta)\xi(\theta|x) \end{aligned}$$

From this we get the conditional plausibility scores of X^* given $X = x$ by the pdf/pmf/...

$$f^*(x^*|x) = \begin{cases} \sum_{\theta\in\Theta} f^*(x^*|\theta)\xi(\theta|x) & \text{if } \xi(\theta|x) \text{ is a pmf} \\ \int_\Theta f^*(x^*|\theta)\xi(\theta|x)d\theta & \text{if } \xi(\theta|x) \text{ is a pdf} \end{cases}$$

Intuitively, the predictive distribution $f^*(x^*|x)$ stands for the following. If we knew θ, we would use $f^*(x^*|\theta)$ to describe X^*. But we do not know θ and our understanding of it is represented by the posterior pdf $\xi(\theta|x)$ given $X = x$. So we must combine our representation of X^* given θ with our representation of θ to get $f^*(x^*|x) = \int_\Theta f^*(x^*|\theta)\xi(\theta|x)d\theta$.

Note that the plug-in approach follows a similar logic, but instead of averaging unknown values of θ, it just plugs-in the estimate $\hat{\theta}(x)$ for θ to produce $\hat{f}^*(x^*|x) = f^*(x^*|\hat{\theta}(x))$.

Posterior predictive distribution of future observation for conjugate models

Consider data X and future observation X^* modeled as $X \sim \mathsf{Binomial}(n, p)$, $X^* \sim \mathsf{Binomial}(m, p)$, X and X^* are independent, $p \in [0, 1]$ assigned a $\mathsf{Beta}(a, b)$ prior pdf. Then,

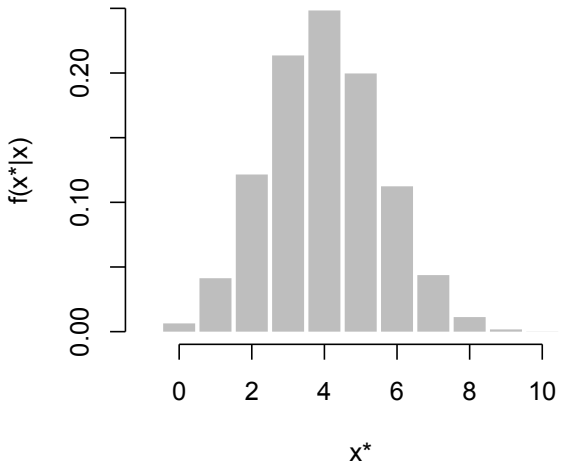
$$f^*(x^*|x) = \int_0^1 \binom{m}{x^*} p^{x^*}(1-p)^{m-x^*}\xi(p|x)dp, \quad x^* \in \{0, 1, \cdots, m\}.$$

But $\xi(p|x) = \mathsf{Beta}(a' = a + x, b' = b + n - x)$ and so, for any $x^* \in \{0, \cdots, m\}$,

$$\begin{aligned} f^*(x^*|x) &= \int_0^1 \binom{m}{x^*} p^{x^*}(1-p)^{m-x^*}\frac{p^{a'-1}(1-p)^{b'-1}}{B(a', b')}dp \\ &= \binom{m}{x^*}\frac{1}{B(a', b')}\int_0^1 p^{a'+x^*-1}(1-p)^{b'+m-x^*-1}dp \\ &= \binom{m}{x^*}\frac{B(a' + x^*, b' + m - x^*)}{B(a', b')}. \end{aligned}$$

Here we could evaluate the integral $\int_\Theta f^*(x^*|\theta)\xi(\theta|x)d\theta$ because it boils down to evaluating the normalizing constant of a function that is a constant multiple of a beta density. Similar calculations will be possible for any conjugate model (see homework).

Example (Opinion poll)**.** Suppose a researcher, having found $X = 200$ supporters of a policy among $n = 500$ students she surveyed, wants to predict the number of supporters X^* in another group of $m = 10$ students. Suppose p, the actual proportion of supporters in the college is assigned a $\mathsf{Uniform}(0, 1)$ prior. Then from the calculations above, $f^*(x^*|x = 200) = \binom{10}{x^*} B(201 + x^*, 301 + 10 - x^*)/B(201, 301)$ for $x^* = 0, \cdots, 10$ (and zero otherwise). A plot of this is shown below.



Special calculations for normal models

The same applies to a normal conjugate model, and we can carry out the integration $\int f^*(x^*|\theta)\xi(\theta|x)$ analytically (with $\theta = \mu$ of $\theta = (\mu, \sigma^2)$ and appropriate conjugate prior pdf $\xi(\theta)$) to derive the expression for $f^*(x^*|x)$ when $X^* = X_{n+1}$ is a future observation. Unsurprisingly, this predictive distribution has a recognizable form. Details are given below.

First, consider data $X = (X_1, \cdots, X_n)$ and future variable $X^* = X_{n+1}$ modeled as $X_1, \cdots, X_n, X_{n+1} \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, σ^2 fixed and suppose $\xi(\mu) = \mathsf{Normal}(a, b^2)$. Then $\xi(\mu|x) = \mathsf{Normal}(a', b'^2)$ where $a' = (nb^2\bar{x} + \sigma^2 a)/(nb^2 + \sigma^2)$ and $b'^2 = b^2\sigma^2/(nb^2 + \sigma^2)$. Then $f^*(x^*|x) = \mathsf{Normal}(a', b'^2 + \sigma^2)$. This follows from the result below. So a 95% central predictive credible interval for X^* is $a' \mp 1.96\sqrt{b'^2 + \sigma^2}$.

RESULT 1. If $W \sim \mathsf{Normal}(a, b^2)$ and $U|(W = w) \sim \mathsf{Normal}(w, c^2)$ then $U \sim \mathsf{Normal}(a, b^2 + c^2)$.

Proof. From the second property, it's OK to write $U = W + Z$ where $Z \sim \mathsf{Normal}(0, c^2)$ and is independent of W. But two independent normals add to a normal with means and variances added, therefore $U = W + Z \sim \mathsf{Normal}(a + 0, b^2 + c^2) = \mathsf{Normal}(a, b^2 + c^2)$. □

Now consider the case where σ^2 is unknown and $\xi(\mu, \sigma^2) = \mathsf{N}\chi^{-2}(m, k, r, s)$. Then $\xi(\mu, \sigma^2|x) = \mathsf{N}\chi^{-2}(m', k', r', s')$ with the usual update formulas. So $f^*(x^*|x)$ equals the distribution of a $t(r')$ random variable scaled by $\sqrt{s'(1+1/k')}$ and shifted by m'. More precisely,

$$\left.\frac{X^* - m'}{\sqrt{s'(1+1/k')}}\right|(X = x) \sim t(r').$$

This follows from the result below. Hence a $100(1-\alpha)\%$ central predictive credible interval for X^* is $m' \mp z_{r'}(\alpha)\sqrt{s'(1+1/k')}$.

RESULT 2. If $(W, V) \sim \mathsf{N}\chi^{-2}(m, k, r, s)$ and $U|(W = w, V = v) \sim \mathsf{Normal}(w, bv)$ then $T = \frac{U-m}{\sqrt{s(b+1/k)}} \sim t(r)$.

Proof. We know $rs/V \sim \chi^2(r)$. Think about the description of (U, W) given $V = v$. This is precisely, $U|(W = w) \sim \mathsf{Normal}(w, bv)$ and $W \sim \mathsf{Normal}(m, v/k)$, therefore, by the result above, still under the condition $V = v$, $U \sim \mathsf{Normal}(m, v(b + 1/k)) = \mathsf{Normal}(m, v/\tilde{k})$ where $\tilde{k} = 1/(b + 1/k)$. But this description of U given $V = v$, coupled with the description $rs/V \sim \chi^2(r)$ means that (U, V) must have $\mathsf{N}\chi^{-2}(m, \tilde{k}, r, s)$ distribution. From properties of this distribution we know, $T = \frac{U-m}{\sqrt{s/\tilde{k}}} \sim t(r)$, which yields the desired result. □

Simulating from $f^*(x^*|x)$

Even when $f^*(x^*|x)$ is not available in closed form, one might obtain summaries of this pdf/pmf by generating random samples from it. For example, if we have access to a sample of draws $\theta_1, \cdots, \theta_M$ from the posterior $\xi(\theta|x)$, then we can generate a sample of draws $x_1^*, \cdots, x_M^*$ from $f^*(x^*|x)$ simply by drawing $x_i^* \sim f^*(x^*|\theta = \theta_i)$. The justification behind this comes from the identity that the joint pdf/pmf/.. of (X^*, θ) given $X = x$ is $f^*(x^*|\theta)\xi(\theta|x)$ and so drawing a $\theta_i \sim \xi(\theta|x)$ and then drawing a $x_i^* \sim f^*(x^*|\theta = \theta_i)$ is precisely same as making a draw (x^*, θ) from the joint pdf/pmf/... Once we have the joint samples $(x_1^*, \theta_1), \cdots, (x_M^*, \theta_M)$ and ignore the θ_i's, the draws x_i^* must precisely be draws from the marginal $f^*(x^*|x)$.

Example (Opinion poll (contd.))**.** For the opinion poll example described above, we could get draws $x_1^*, \cdots, x_M^*$ from $f^*(x^*|x)$ as follows:

```
n <- 500; x <- 200;              ## data
a <- 1; b <- 1;                  ## prior Be(a = 1, b = 1)
a.x <- x + a; b.x <- n - x + b; ## posterior Be(a.x, b.x)
M <- 1000                        ## number of samples to draw
p.samp <- rbeta(M, a.x, b.x)     ## draw p from posterior
x.star <- rbinom(M, 10, p.samp) ## draw x.star[i] ~ Bin(10, p.samp[i])
hist(x.star, freq = FALSE, col = "gray", border = "white", breaks = 0:11 - 0.5)
```