Description of the experimental statistics

The experimental statistics on recreational fishing introduces probability calculus-based estimates of the number of households that engaged in recreational fishing in 2022. The statistics publish numerical data on households that fished, based on the stratum division used in the sampling framework. The statistics based on a new calculation method rely on Bayesian inference and are published one-off as experimental.

Statistical unit

The target of data collection and the primary statistical unit is the household. A household consists of persons permanently residing in the same dwelling.

Statistical concepts and definitions

A household is considered to have fished if any member of the household has used any type of fishing gear at least once during the year. A person is considered to have fished even if they only rowed or steered a boat while someone else fished. Recreational fishing refers to all fishing activities (including crayfishing) conducted by Finnish households, excluding fishing carried out by professional fishermen and their households.

Population

The population of the statistics consists of Finnish households ($N = 2,816,206$). Stratified sampling is used as the sampling design. In the sample design for the recreational fishing statistics, the Finnish household population is logically divided into seven strata such that $N = N_1 + N_2 + \cdots + N_7$. An invitation to participate in the survey was sent to the samples of size n_k drawn from each stratum k ($k = 1, 2, \ldots, 7$).

When forming stratum's 1-6 (see Figure 1), information on the target person's municipality of residence (capital region, rest of Southern Finland, Western Finland, Eastern Finland, Oulu area, Lapland, and Åland), municipality type (urban, densely populated, and rural), and maritime classification (archipelago municipality, coastal municipality, and inland municipality) is utilized. Households that, according to Metsähallitus records, had paid the fisheries management fee in the statistical year form their own stratum (strata 7). For sampling and estimation purposes, households in stratum 7 are removed from the geographically based population data (strata 1-6). The total sample in 2022 comprised 11,000 households.

Data collection method

The data collection is dual-phased. In the first phase, a questionnaire is sent to the selected sample and can be returned by mail or online. The households who do not respond to the survey, and for whom a phone number is available, are targeted for a telephone survey with a narrower content (sc. loss interview).

Methods

Bayesian experimental statistics combine prior knowledge and observed data to provide updated probability-based assessments. Bayesian inference relies on Bayes' theorem, which defines how probability distributions are updated with new information. This method allows to perform dynamic statistical analysis, where prior knowledge or hypotheses (priors) are updated based on observed data, resulting to the target distribution of interest (posterior). The experimental statistics were conducted using the R software [3], version 4.2.2. Markov Chain Monte Carlo sampling was performed with the JAGS software [4], which was called via the `runjags` [2] R package. The R package `geofi` [5] was utilized for drawing the maps.

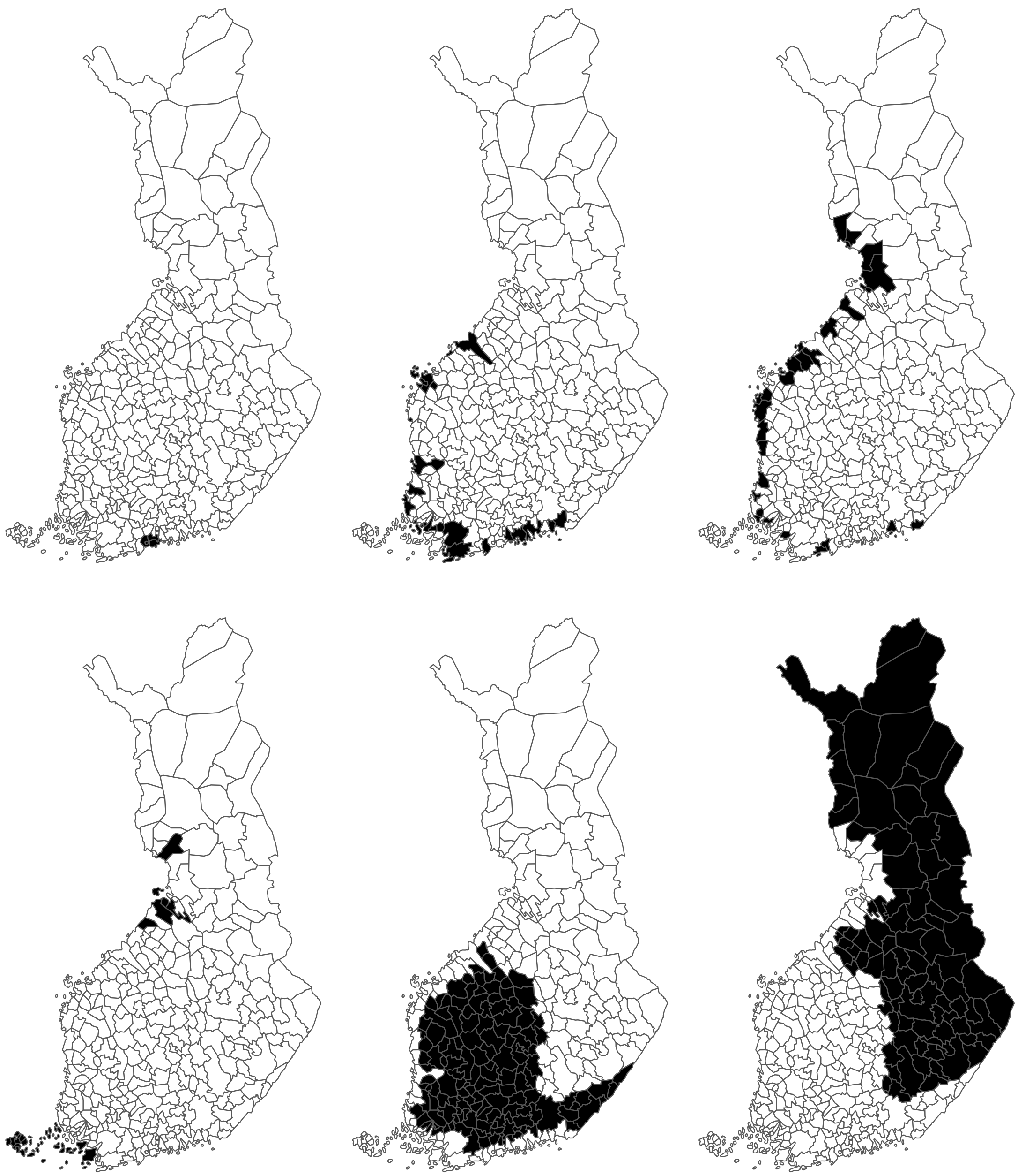

Figure 1: The strata according to the sampling design, numbered 1-6, respectively.

Statistical model

A statistical model can be formulated based on the sampling design and data collection method of the recreational fishing survey. The sample, where the total sample size $n = n_1 + \cdots + n_7$ is the sum of the stratum-specific sample sizes, is modeled in the first phase. Let $K \in \{1, 2\}$ denote household's fishing such that 1 indicates fishing and 2 indicates non-fishing. Define q_k as the probability that a household ik in stratum n_k fishes at least once during the year. In other words, $P(K_{ik} = 1) = q_k$ and $P(K_{ik} = 2) = 1 - q_k$. Thus, the quantity K can be thought to be generated through a Bernoulli distribution, which can be expressed using categorical distribution. Therefore, the fishing component of the model can be written as

$$
\begin{aligned}
K_{ik} &\overset{\text{ind}}{\sim} \text{Cat}(Q_k) \quad k = 1, \ldots, 7, \; i = 1, \ldots, n_k \\
q_k &\sim p(),
\end{aligned}
$$

where $Q_k = [q_k, 1 - q_k]$ and $p()$ is some arbitrary distribution. Furthermore, if a household fishes, it is thought that the household will respond to the mail survey with some probability p_{11}, and if it does not fish, it will respond to the mail survey with some probability p_{12}. Moreover, if a household does not respond to the mail survey, an attempt is made to find a phone number for the head of the household selected in the sample. It is thought that the phone number is found with some probability r. If the phone number is found, it is considered that a fishing household will respond to the phone interview with some probability p_{21}, and if the household does not fish, it will respond to the phone survey with some probability p_{22}. The outcome of the dual-phase data collection is denoted by the quantity $V \in \{1, 2, 3, 4\}$, whose realization is known for each household selected in the sample.

More precisely, participation in the original survey ($V = 1$) is thought to be associated with some probability believed to be different for fishing and non-fishing households. Therefore, let $P(V_{ik} = 1|K_{ik} = 1) = p_{11}$ and $P(V_{ik} = 1|K_{ik} = 2) = p_{12}$, as mentioned earlier. Additionally, if r denotes the selection probability for the second sample (phone survey), then $P(V_{ik} = 2|K_{ik} = 1) = (1 - r)(1 - p_{11})$ and $P(V_{ik} = 2|K_{ik} = 2) = (1 - r)(1 - p_{12})$, which corresponds to the situation where the household did not respond to the original survey and no phone number was found. If a phone number is available, attempts are made to reach the household for a phone interview. Since the data collection method changes, the response probability is expected to be different. The probabilities of responding to the phone survey, assuming that the household did not initially respond and a phone number was available are $P(V_{ik} = 3|K_{ik} = 1) = p_{21}(1 - p_{11})r$ and $P(V_{ik} = 3|K_{ik} = 2) = p_{22}(1 - p_{12})r$. The fourth or last category consists of households that did not respond to the original survey, for which a phone number was available, but did not respond to the phone survey either. The probabilities for the fourth category are $P(V_{ik} = 4|K_{ik} = 1) = (1 - p_{21})(1 - p_{11})r$ and $P(V_{ik} = 4|K_{ik} = 2) = (1 - p_{22})(1 - p_{12})r$. This means that V_{ik} can be taken as a random variable following a categorical distribution conditioned on fishing K_{ik}. Define a matrix P, which contains the previously defined class probabilities.

$$
P = \begin{bmatrix} p_{11} & (1 - r)(1 - p_{11}) & p_{21}(1 - p_{11})r & (1 - p_{21})(1 - p_{11})r \\ p_{12} & (1 - r)(1 - p_{12}) & p_{22}(1 - p_{12})r & (1 - p_{22})(1 - p_{12})r \end{bmatrix}
$$

Based on the above, a statistical model for the random variable V can be defined as

$$
\begin{aligned}
V_{ik} &\overset{\text{ind}}{\sim} \text{Cat}(P[K_{ik},]) \quad k = 1, \ldots, 7, \; i = 1, \ldots, n_k \\
p_{jl} &\sim p(), \; j, l = 1, 2,
\end{aligned}
$$

where $p()$ denotes arbitrary distribution and $P[K_{ik} \in \{1, 2\},]$ is the row of matrix P determined by the fishing status (estimated or reported) of the household. The total number of fishing households $K1^k_{\text{SAMP}}$ in sample n_k is obtained by summing individual fishing households.

$$
K1^k_{\text{SAMP}} = \sum_{i=1}^{n_k} -(K_{ik} - 2)
$$

Most households in stratum k do not end up in the sample. The purpose of the sampling model is to learn something about each stratum and then exploit this information for the population of households $N^k_{\text{OUT}} = N_k - n_k$ that are outside the sample. Because the households are randomly selected into the sample, they are exchangeable within the strata. Additionally, as the Bernoulli distribution is a special case of the binomial distribution, the probability q_k learned from the sample can be used to assess the number of fishing households $K1^k_{\text{OUT}}$ outside the sample. Hence, based on the aforementioned,

$$K1^k_{\text{OUT}} \sim \text{Bin}(N^k_{\text{OUT}}, q_k).$$

The chosen strategy is to first model the sample with size n_k, and then use this information to model the larger population N^k_{OUT} that has not been studied. Consequently, generalizing the results to the finite superpopulation of size $N = N_1 + \cdots + N_7 = n_1 + N^1_{\text{OUT}} + \cdots + n_7 + N^7_{\text{OUT}}$ is just a few summation operations away. This final step highlights the uniqueness of official statistics compared to conventional statistical inference. In the sample model, the population parameters were already inferred. However, in official statistics, the results are usually published as whole natural numbers. The final operation is simple. The estimate of the number of fishing households among Finnish households is obtained by calculating

$$K1_{\text{POP}} = \sum_k K1^k_{\text{SAMP}} + K1^k_{\text{OUT}}.$$

Selection of distributions $p()$

An essential part of Bayesian inference is the use of prior knowledge. Prior knowledge can come from expert judgment, results of previous studies, or logical reasoning. Priors are system inputs in the form of a probability distribution, and often simply referred as the prior distribution. Based on the results from previous years [1], the following prior distributions were chosen for the probabilities of fishing q_k.

$$q_1, q_2 \sim \text{Beta}(2, 7), \quad q_3, q_4, q_5 \sim \text{Beta}(3, 8), \quad q_6 \sim \text{Beta}(7, 8), \quad q_7 \sim \text{Beta}(10, 2).$$

Based on assessments of 2014-2020, it can be assumed that for households in strata 1 and 2 (parameters q_1, q_2), perhaps two out of ten households fish at least once a year, but certainly not more than half. The next three geographical strata correspond to parameters q_3, q_4, q_5, for which prior knowledge can be described verbally by saying that maybe about one-third of households fish at least once a year, but very likely not half or more. For stratum 6, it can be argued that perhaps about half of the households fish once a year. However, it is quite certain that more than two or three out of ten households fish, but very likely not more than 70%. For those who have paid the fisheries management fee (q_7), the idea is that if the fisheries management fee has been paid, the household very likely fishes at least once a year, although not all these households fish. The prior distributions for the parameters related to data collection were also defined based on estimates from previous years.

$$p_{11} \sim \text{Beta}(5, 10), \quad p_{21} \sim \text{Beta}(2, 10), \quad p_{12} \sim \text{Beta}(2, 5), \quad p_{22} \sim \text{Beta}(2, 5), \quad r \sim \text{Beta}(2, 6).$$

The chosen distributions reflect the idea that fishing households generally participate in the survey better than non-fishing households, the probability of participating in a telephone interview is higher than that of participating in a mail survey, and it is estimated that maybe less than half of the households in the sample who do not participate in the mail survey will have a phone number available.

It should be noted that the given distributions are not very peaked or tightly bounded. Thus, they can be characterized as *mildly* informative. Additionally, it is important to note that despite the low response rate, a large number of responses were received for the first question (whether the household fished or not). Therefore, the influence of the prior distributions on the final result is considerably less significant than the observed data.

Definition

Fishing is considered to occur when an individual uses any kind of fishing equipment or gear at least once within the duration of one year. A person is regarded to have fished even if only rowed or steered a boat while another person fished. A household is considered to have fished if at least one member of the household fished at least once during the year.

Timing

The experimental statistics on recreational fishing present assessments for the year 2022. The data was collected by the Natural Resources Institute Finland in the spring of 2023. The model was developed during 2023. The statistics were produced in February 2024.

References

[1] Antti Sykkö. The Bayesian Approach to Official Statistics: Problems & Benefits, 2023.

[2] Matthew J. Denwood. runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *Journal of Statistical Software*, 71(9):1–25, 2016.

[3] R Foundation for Statistical Computing - R Core Team. R: A Language and Environment for Statistical Computing, 2021. https://www.R-project.org/, [September 23, 2023].

[4] Sarah Depaoli and James P. Clifton and Patrice R. Cobb. Just Another Gibbs Sampler (JAGS): Flexible Software for MCMC Implementation. *Journal of Educational and Behavioral Statistics*, 41(6):628–649, 2016.

[5] M. Kainu, J. Lehtomaki, J. Parkkinen, J. Miettinen, P. Kantanen, S. Vesanen, and L. Lahti. *geofi: Access Finnish Geospatial Data*, 2023. R package version 1.0.8.