Confidence distributions in statistical inference

Sergey Bityukov*, Nikolai Krasnikov†, Saralees Nadarajah** and Vera Smirnova*

*Institute for high energy physics, 142281 Protvino, Russia
†Institute for nuclear research RAS, Prospect 60-letiya Octyabrya, 7a, 117312 Moscow, Russia
**University of Manchester, Manchester M13 9PL, United Kingdom

Abstract. This paper reviews the new methodology for statistical inferences. Point estimators, confidence intervals and $p-$values are fundamental tools for frequentist statisticians. Confidence distributions, which can be viewed as "distribution estimators", are often convenient for constructing all of the above statistical procedures and more.

Keywords: Uncertainty, Confidence distributions, Estimation
PACS: 02.50.Tt, 06.20.Dk, 07.05.Kf

Introduction

If a procedure states one-to-one correspondence between the observed value of a random variable and the confidence interval of any level of significance then we can reconstruct a unique confidence density of the parameter and, correspondingly, a unique confidence distribution. The confidence distribution is a very useful tool in statistical reporting, and should be a competitive frequentist analogue of the Bayesian posterior distribution.

The following example shows the construction by R.A. Fisher [1, 2].

A random variable x with parameter θ $x \sim \mathbf{N}(\theta, 1)$, where the symbol $\sim$ means "distributed as". The probability density function (pdf) is

$$\varphi(x|\theta) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(x-\theta)^2}{2}}. \qquad (1)$$

We can write $x = \theta + \varepsilon$, where $\varepsilon \sim \mathbf{N}(0,1)$ and θ is a constant.

Let $\hat{x}$ be a single realization of x. For a normal distribution it is an unbiased estimator of parameter θ, i.e. $\hat{\theta} = \hat{x}$, therefore $\theta|\hat{x} = \hat{x} - \varepsilon$.

As is known $(-\varepsilon) \sim \mathbf{N}(0,1)$ due to the symmetry of the bell-shaped curve about its central point, i.e. $\theta|\hat{x} \sim \mathbf{N}(\hat{x}, 1)$.

Thus we construct the confidence density of the parameter

$$\tilde{\varphi}(\theta|\hat{x}) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(\hat{x}-\theta)^2}{2}} \qquad (2)$$

uniquely for each value of $\hat{x}$.

As pointed in paper [2] "Fisher [3, 4] gave correct interpretation of this "tempting" result. But starting in 1935 [5], he really believed he had changed the status of θ from that of a fixed unknown constant to that of a random variable on the parameter space with known distribution" [1]. In principle, the parameter θ can be a random variable in the case of the random origin of parameter. We will not discuss here this possibility.

The construction above is a direct consequence of the following identity

$$\int_{-\infty}^{\hat{x}-\alpha_1} \varphi(x|\hat{x})dx + \int_{\hat{x}-\alpha_1}^{\hat{x}+\alpha_2} \tilde{\varphi}(\theta|\hat{x})d\theta + \int_{\hat{x}+\alpha_2}^{\infty} \varphi(x|\hat{x})dx = 1, \qquad (3)$$

where $\hat{x}$ is the observed value of random variable x, and $\hat{x}-\alpha_1$ and $\hat{x}+\alpha_2$ are confidence interval bounds for location parameter θ.

In the case of Poisson and Gamma-distributions we can also exchange the random variable and the parameter, preserving the same formula for the probability distribution:

$$f(i|\theta) = \tilde{f}(\theta|i) = \frac{\theta^i e^{-\theta}}{i!}$$

In this case we can use another identity to relate the pdf of random variable and confidence density of the parameter for the unique reconstruction of confidence density [9, 10] (any other reconstruction is inconsistent with the identity and, correspondingly, breaks the probability conservation):

$$\sum_{i=\hat{x}+1}^{\infty} f(i|\theta_1) + \int_{\theta_1}^{\theta_2} \tilde{f}(\theta|\hat{x})d\theta + \sum_{i=0}^{\hat{x}} f(i|\theta_2) = 1, \qquad (4)$$

for any real $\theta_1 \geq 0$ and $\theta_2 \geq 0$ and non-negative integer $\hat{x}$. Confidence density $\tilde{f}(\theta|i)$ is the pdf of Gamma-distribution $\Gamma_{1,i+1}$ and $\hat{x}$ is the number of observed events.

The presence of the identities of such type (Eqs. 3, 4) is a property of statistically dual distributions [11, 12] [2].

Confidence distributions (CDs), which can be viewed as "distribution estimators", are often convenient for constructing all the above statistical procedures and more. The basic notion of CDs traces back to the fiducial distribution of Fisher [3]; however, it can be viewed as a pure frequentist concept. Indeed, as pointed out in [13] the CD concept is a "Neymannian interpretation of Fisher's fiducial distribution" [14]. Its development has proceeded from Fisher [3] through various contributions, just to name a few, of Kolmogorov [15], Pitman [16], Efron [17, 18], Fraser [19, 20], Lehmann [21], Singh, Xie and Strawderman [22, 23], Schweder and Hjort [13, 24] and others. Bityukov [9] and Bityukov and Krasnikov [10, 11] developed the approach for reconstruction of the confidence distribution densities by using the corresponding identities. The results were further tested by Monte Carlo simulation [25].

[1] The history and generalization of the last approach can be found in paper [6]. The fiducial argument is very attractive notion and sometimes it reopen (see, as an example, [7] and corresponding critique [8]).

[2] Let $\phi(x,\theta)$ be a function of two variables. If the same function can be considered both as a family of the pdfs $\varphi(x|\theta)$ of the random variable x with parameter θ and as another family of pdfs $\tilde{\varphi}(\theta|x)$ of the random variable θ with parameter x (i.e. $\phi(x,\theta) = \varphi(x|\theta) = \tilde{\varphi}(\theta|x)$), then this pair of families of distributions can be named as *statistically dual distributions*. If x and θ play the symmetric role then these distributions can be named as *statistically self-dual distributions*.

Another useful application of CD is for meta-analysis [3]. The consecutive theory of combining information from independent sources through CD is proposed in [27]. Recently Bickel [28] suggested a method for incorporating expert knowledge into frequentist approach by combining generalized confidence distributions.

Confidence distributions

Basic definitions [27]

Suppose $X_1, X_2, \ldots, X_n$ are n independent random draws from a population $\mathbf{F}$ and χ the sample space corresponding to the data set $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)^T$. Let θ be a parameter of interest associated with $\mathbf{F}$ ($\mathbf{F}$ may contain other nuisance parameters), and let Θ be the parameter space.

Definition 1: *A function $H_n(\cdot) = H_n(X_n, (\cdot))$ on $\chi \times \Theta \to [0,1]$ is called a confidence distribution (CD) for a parameter θ if*
(i) for each given $\mathbf{X}_n \in \chi$, $H_n(\cdot)$ is a continuous cumulative distribution function;
(ii) at the true parameter value $\theta = \theta_0, H_n(\theta_0) = H_n(\mathbf{X}_n, \theta_0)$, as a function of the sample $\mathbf{X}_n$, has the uniform distribution $U(0,1)$.

The function $H_n(\cdot)$ is called an asymptotic confidence distribution (aCD) if requirement (ii) above is replaced by (ii)': at $\theta = \theta_0, H_n(\mathbf{X}_n, \theta_0) \xrightarrow{W} U(0,1)$ as $n \to +\infty$, and the continuity requirement on $H_n(\cdot)$ is dropped.

We call, when it exists, $h_n(\theta) = H_n^{'}(\theta)$ a confidence or CD density.

Item *(i)* requires the function $H_n(\cdot)$ to be a distribution function for each given sample.

Item *(ii)* states that the function $H_n(\cdot)$ brings the information onto the probability scale and thus provides confidence intervals and $p-$values.

A CD contains a wealth of information, somewhat comparable to, but different than, a Bayesian posterior distribution. A CD (or aCD) derived from a likelihood function can also be interpreted as an objective Bayesian posterior.

Example Normal mean and variance: Suppose $X_1, X_2, \ldots, X_n$ is a sample from $\mathbf{N}(\mu, \sigma^2)$, with both μ and σ^2 unknown. A CD for μ is $H_n(y) = F_{t_{n-1}}(\frac{y - X}{s_n/\sqrt{n}})$, where $\bar{X}$ and s^2 are, respectively, the sample mean and the sample variance, and $F_{t_{n-1}}(\cdot)$ is the cumulative distribution function of the Student t_{n-1}-distribution. A CD for σ^2 is $H_n(y) = 1 - F_{\chi^2_{n-1}}(\frac{(n-1)s_n^2}{y})$ for $y \geq 0$, where $F_{\chi^2_{n-1}}(\cdot)$ is the cumulative function of the χ^2_{n-1}-distribution.

[3] Meta-analysis is the modern term for combining results from different experiments or trials (see, for example, [26]).

Confidence distributions and pivots [29]

Consider the statistical model for the data X. The model consists of a family of probability distributions for X, indexed by the vector parameter (ψ,χ), where ψ is a scalar parameter of primary interest, and χ is a nuisance parameter (vector).

Definition 2 : *A univariate data-dependent distribution for ψ, with cumulative distribution function $C(\psi;X)$ and with quantile function $C^{-1}(\alpha;X)$ is an exact confidence distribution if $P_{\psi\chi}(\psi \leq C^{-1}(\alpha;X)) = P_{\psi\chi}(C(\psi;X) \leq \alpha) = \alpha$ for all $\alpha \in (0,1)$ and for all probability distributions in the statistical model.*

By definition, the stochastic interval $(\infty, C^{-1}(\alpha;X))$ covers ψ with probability α, and is a one-sided confidence interval method with coverage probability α. The interval $(C^{-1}(\alpha;X), C^{-1}(\beta;X))$ will for the same reason cover ψ with probability $\beta-\alpha$, and is a confidence interval method with this coverage probability. When data have been observed as $X=x$, the realized numerical interval $(C^{-1}(\alpha;x), C^{-1}(\beta;x))$ will either cover or not cover the unknown true value of ψ. The degree of confidence $\beta-\alpha$ that is attached to the realized interval is inherited from the coverage probability of the stochastic interval. The confidence distribution has the same dual property. *Ex ante* data, the confidence distribution is a stochastic entity with probabilistic properties. *Ex post* data, however, the confidence distribution is a distribution of confidence that can be attached to interval statement.

The realized confidence (degree of confidence) $C(\psi;x)$ is a p-value of the one-sided hypothesis $H_0 : \psi \leq \psi_0$ versus $\psi > \psi_0$ when data have been observed to be x. The *ex ante* confidence, $C(\psi;X)$ is by definition uniformly distributed. The p-value is just a transformation of the test statistic to the common scale of the uniform distributions (*ex ante*). The realized p-value when testing the two-sided hypothesis $H_0 : \psi = \psi_0$ versus $\psi \neq \psi_0$ is 2 $min\{C(\psi_0), 1-C(\psi_0)\}$.

Confidence distributions are easily found when pivots [30] can be identified [4].

A function of the data and the interest parameter, $p(X,\psi)$, is a pivot if the probability distribution of $p(X,\psi)$ is the same for all (ψ,χ), and the function $p(X,\psi)$ is increasing in ψ for almost all x.

If based on a pivot with cumulative distribution function F, the cumulative confidence distribution is $C(X,\psi) = F(p(X,\psi))$.

From the definition, a confidence distribution is exact if and only if $C(X,\psi) \sim U$ is a uniformly distributed pivot.

Applications of the CD notion

Confidence intervals for signal with expected background [12]

The confidence density is a more informative notion than the confidence interval. For example, the Gamma-distribution $\Gamma_{1,\hat{n}+1}$ is the confidence density of the parameter of

[4] The self-duality in Eq. 3 is equivalent to the existence of a linear and symmetrically distributed pivot.

Poisson distribution in the case of the $\hat{n}$ observed events from the Poisson flow of events. It means that we can reconstruct any confidence interval (shortest, central, ...) by direct calculation of the pdf of a Gamma-distribution. The following example illustrates the advantages of the confidence density construction.

Let us consider the Poisson distribution with two components: a signal component with a parameter μ_s and a background component with a parameter μ_b, where μ_b is known. To construct confidence intervals for the parameter μ_s in the case of observed value $\hat{n}$, we must find the distribution $\tilde{f}(\mu_s|\hat{n})$.

First let us consider the simplest case $\hat{n} = \hat{s} + \hat{b} = 1$. Here $\hat{s}$ is the number of signal events and $\hat{b}$ is the number of background events among the observed number $\hat{n}$ of events.

$\hat{b}$ can be equal to 0 and 1. We know that $\hat{b}$ is equal to 0 with probability

$$p_0 = P(\hat{b} = 0) = \frac{\mu_b^0}{0!} e^{-\mu_b} = e^{-\mu_b}$$

and $\hat{b}$ is equal to 1 with probability

$$p_1 = P(\hat{b} = 1) = \frac{\mu_b^1}{1!} e^{-\mu_b} = \mu_b e^{-\mu_b}.$$

Correspondingly,

$$P(\hat{b} = 0|\hat{n} = 1) = P(\hat{s} = 1|\hat{n} = 1) = \frac{p_0}{p_0 + p_1} \text{ and}$$
$$P(\hat{b} = 1|\hat{n} = 1) = P(\hat{s} = 0|\hat{n} = 1) = \frac{p_1}{p_0 + p_1}.$$

It means that the distribution of the confidence density $\tilde{f}(\mu_s|\hat{n} = 1)$ is equal to the weighted sum of distributions

$$\hat{f}(\mu_s|\hat{n} = 1) = P(\hat{s} = 1|\hat{n} = 1)\tilde{f}(\mu_s|\hat{s} = 1) + P(\hat{s} = 0|\hat{n} = 1)\tilde{f}(\mu_s|\hat{s} = 0), \qquad (5)$$

where the confidence density $\tilde{f}(\mu_s|\hat{s} = 0)$ is the Gamma distribution $\Gamma_{1,1}$ with the pdf

$$\tilde{f}(\mu_s|\hat{s} = 0) = e^{-\mu_s}$$

and the confidence density $\tilde{f}(\mu_s|\hat{s} = 1)$ is the Gamma distribution $\Gamma_{1,2}$ with the pdf

$$\tilde{f}(\mu_s|\hat{s} = 1) = \mu_s e^{-\mu_s}.$$

As a result, we have the confidence density of the parameter μ_s

$$\tilde{f}(\mu_s|\hat{n} = 1) = \frac{\mu_s + \mu_b}{1 + \mu_b} e^{-\mu_s}.$$

Using this formula for $\tilde{f}(\mu_s|\hat{n} = 1)$, we can construct the shortest confidence interval of any confidence level trivially.

In this manner we can construct the confidence density $\tilde{f}(\mu_s|\hat{n})$ for any values of $\hat{n}$ and μ_b. From Eq. 4 we use the confidence densities $\tilde{f}(\mu_s|\hat{s} = i)$, $i = 0, \hat{n}$. Mixing together the confidence densities with corresponding conditional probability weights (in analogy with Eq. 5) yields the confidence density

$$\tilde{f}(\mu_s|\hat{n}) = \frac{(\mu_s+\mu_b)^{\hat{n}}}{\hat{n}! \sum\limits_{i=0}^{\hat{n}} \dfrac{\mu_b^i}{i!}} e^{-\mu_s}.$$

We have obtained the known formula [31, 32, 33]. The numerical results of the calculations of shortest confidence intervals using this confidence density coincide with Bayesian confidence intervals constructed using the uniform prior.

Estimation of quality of planned experiment [10]

Let us consider the estimation of the quality of planned experiments as another example of the use of confidence density. The approach is based on the analysis of uncertainty, which will take place under the future hypotheses testing about the existence of a new phenomenon in Nature. We consider the Poisson distribution with parameter μ and we preserve the notation of the previous subsection. We test a simple statistical hypothesis H_0: *new physics is present in Nature* (i.e $\mu = \mu_s + \mu_b$) against a simple alternative hypothesis H_1: *new physics is absent* (i.e. $\mu = \mu_b$).

The value of uncertainty is determined by the values of the probability to reject the hypothesis H_0 when it is true (Type I error α) and the probability to accept the hypothesis H_0 when the hypothesis H_1 is true (Type II error β). This uncertainty characterizes the distinguishability of the hypotheses under the given choice of critical area.

Let both values μ_s and μ_b, which are defined in the previous Section, be exactly known. In this simplest case the errors of Type I and II, which will take place in testing of hypothesis H_0 versus hypothesis H_1, can be written as follows:

$$\begin{cases} \alpha = \sum\limits_{i=0}^{n_c} f(i|\mu_s+\mu_b), \\ \beta = 1 - \sum\limits_{i=0}^{n_c} f(i|\mu_b), \end{cases} \tag{6}$$

where f is a Poisson probability function and n_c is a critical value.

Let the values $\hat{\mu}_s = \hat{s}$ and $\hat{\mu}_b = \hat{b}$ be known, for example, from Monte Carlo experiment with integrated luminosity [5] which is exactly the same as the data luminosity later in the planned experiment. It means that we must include the uncertainties in values μ_s and μ_b to the system of the equations Eqs. 6. As is shown [9] (see, also, the generalized case in the same reference) we have the system

[5] In scattering theory and accelerator physics, luminosity is the number of particles per unit area per unit time times the opacity of the target. The integrated luminosity is the integral of the luminosity with respect to time.

$$\begin{cases} \alpha = \int_0^\infty \tilde{f}(\mu|\hat{s}+\hat{b}) \sum_{i=0}^{n_c} f(i|\mu) d\mu = \sum_{i=0}^{n_c} \frac{C^i_{\hat{s}+\hat{b}+i}}{2^{\hat{s}+\hat{b}+i+1}}, \\ \beta = 1 - \int_0^\infty \tilde{f}(\mu|\hat{b}) \sum_{i=0}^{n_c} f(i|\mu) d\mu = 1 - \sum_{i=0}^{n_c} \frac{C^i_{\hat{b}+i}}{2^{\hat{b}+i+1}}, \end{cases}$$

where n_c is a critical value of the hypotheses testing about the observability of signal and C^i_N is $\dfrac{N!}{i!(N-i)!}$.

Note, here the Poisson distribution is a prior distribution of the expected probabilities and the negative binomial (Pascal) distribution is a posterior distribution of the expected probabilities of the random variable. This is a transformation of the estimated confidence densities $\tilde{f}(\mu|\hat{s}+\hat{b})$ and $\tilde{f}(\mu|\hat{b})$ (pdfs of the corresponding $\Gamma-$distributions) to the space of the expected values of the random variable.

Conclusion

The notion of a confidence distribution, an entirely frequentist concept, is in essence a Neymanian interpretation of Fisher's fiducial distribution. It contains information related to every kind of frequentist inference. The confidence distribution is a direct generalization of the confidence interval, and is a useful format of presenting statistical inference.

ACKNOWLEDGMENTS

We are grateful to Vassili Kachanov, Victor Matveev and Nikolai Tyurin for interest and support of this work. We thank Luc Demortier, Bradley Efron, Kesar Singh and Minge Xie for helpful and constructive comments. Especially, we thanks Sir David Cox invited our attention to Kolmogorov's paper. We also thank Sergei Gleyzer, Yuri Gouz, Alexandre Nikitenko and Claudia Wulz for very useful discussions. This work was supported by the Grants RFBR 08-02-91007-CERN and 10-02-00468.

REFERENCES

1. B. Efron, *American Mathematical Monthly* **85**, 231–246 (1978).
2. F. Hampel, *Lecture Notes in Computer Science* **4123**, 512–526 (2006).
3. R. A. Fisher, *Proceedings of the Cambridge Philosophical Society* **26**, 528–535 (1930).
4. R. A. Fisher, *Proceedings of the Royal Society, London* **A139**, 343–348 (1933).
5. R. A. Fisher, *Annals of Eugenics* **6**, 391–398 (1935).
6. J. Hannig, H. Iyer, P. Patterson, *Journal of the American Statistical Association* **101**, 254–269 (2006).
7. A. Hassairi, A. Masmoudi, C. C. Kokonendji, *Communications in Statistics—Theory and Methods* **34**, 245–252 (2005).
8. N. Mukhopadhyay, *Communications in Statistics—Theory and Methods* **35**, 293–297 (2006).
9. S. I. Bityukov, *Journal of High Energy Physics* **09**, 060 (2002).

10. S. I. Bityukov, N. V. Krasnikov, *Nuclear Instruments and Methods in Physics Research* **A502**, 795–798 (2003).
11. S. I. Bityukov, N. V. Krasnikov, "Statistically dual distributions and conjugate families," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by K. H. Knuth et al., AIP Conference Proceedings 803, American Institute of Physics, New York, 2005, pp. 398–402.
12. S. I. Bityukov, N. V. Krasnikov, V. V. Smirnova, V. A. Taperechkina, *Proceedings of Science (ACAT)* **062**, 1–9 (2007).
13. T. Schweder, N.L. Hjort, *Scandinavian Journal of Statistics* **29**, 309–332 (2002).
14. J. Neyman, *Biometrika* **32**, 128–150 (1941).
15. A. N. Kolmogorov, *Bulletin of the Academy of Sciences*, USSR, *Mathematics Series* **6**, 3–32 (1942) (in russian).
16. P. J. G. Pitman, *Journal of the American Statistical Association* **52**, 322–330 (1957).
17. B. Efron, *Biometrika* **80**, 3–26 (1993).
18. B. Efron, *Statistical Science* **13**, 95–122 (1998).
19. D. A. S. Fraser, *Journal of the American Statistical Association* **86**, 258-265 (1991).
20. D. A. S. Fraser, *International Statistical Review* **64**, 231–235 (1996).
21. E. L. Lehmann, *Journal of the American Statistical Association* **88**, 1242–1249 (1993).
22. K. Singh, M. Xie, W. Strawderman, "Confidence distributions - concept, theory and applications," Technical report, Dept.Statistics, Rutgers Univ., 2001, Revised 2004.
23. K. Singh, M. Xie, W. Strawderman, *IMS Lecture Notes Monograph Series* **54**, 132–150 (2007).
24. T. Schweder, N. L. Hjort, "Frequentist analogies of priors and posteriors," in *Econometrics and the Philosophy of Economics*, Princeton University Press, 2003, pp.285–317.
25. S. I. Bityukov, V. A. Medvedev, V. V. Smirnova, Yu. V. Zernii, *Nuclear Instruments and Methods in Physics Research* **A534**, 228-231 (2004).
26. L. V. Hedges, I. Olkin, *Statistical Methods for Meta-analysis*, Academic Press, Orlando, 1985.
27. K. Singh, M. Xie, W. Strawderman, *Annals of Statistics* **33**, 159–183 (2005).
28. D. R. Bickel, *arXiv: math/0602377*[math.ST] 1–19 (2006).
29. T. Schweder, *Scientia Marina* **67**, 89-97 (2003).
30. O. E. Barndorff-Nielsen, D. R. Cox, *Inference and Asymptotics*, Chapman & Hall, London, 1994.
31. O. Helene, G.P. Yost et.al., Review of Particle Properties, *Physics Letters* **B204**, 81 (1988).
32. G. Zech, *Nuclear Instruments and Methods in Physics Research* **A277**, 608–610 (1989).
33. G. D'Agostini, *Bayesian Reasoning in Data Analysis, a Critical Introduction*. World Scientific, Hackensack, NJ, 2003.