Confidence distributions in statistical inference

Sergei I. Bityukov
Institute for High Energy Physics, Protvino, Russia
Nikolai V. Krasnikov
Institute for Nuclear Research RAS, Moscow, Russia

Plan

- **Motivation**
- **A bit of history**
- **Confidence distributions (CD and aCD)**
- **Examples and inferential information contained in a CD**
- **Inference: a brief summary**
- **Combination of CDs**
- **Conclusion**
- **References**

Motivation (I) Common sense

If we have the procedure which states the one-to-one conformity between the observed value of random variable and the confidence interval of any level of significance then we can reconstruct the confidence density of the parameter and, correspondingly, the confidence distribution by single way.

The confidence distribution is interpreted here in the same way as confidence interval. From the duality between testing and confidence interval estimation, the cumulative confidence distribution function for parameter μ evaluated at μ_0, $F(\mu_0)$, is the $p-$value of testing $H_0 : \mu \geq \mu_0$ against its one-side alternative. It is thus a compact format of representing the information regarding μ contained in the data and given the model. Before the data have been observed, the confidence distribution is a stochastic element with quantile intervals $(F^{-1}(\alpha_1), F^{-1}(1-\alpha_2))$ which covers the unknown parameter with probability $1-\alpha_1-\alpha_2$. After having observed the data, the realized confidence distribution is not a distribution of probabilities in the frequentist sense, but of confidence attached to interval statements concerning μ. Note, the confidence distribution here is a frequentist concept and does not rely on a prior distribution.

Motivation (II) Construction by R.A. Fisher

Example of the construction by R.A. Fisher (B. Efron (1978))

Random variable x with parameter μ

$$x \sim \mathcal{N}(\mu, 1). \tag{1}$$

Probability density function here is

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

We can write

$$x = \mu + \epsilon, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and μ is a constant.

Let we have got $\hat{x}$ realization of x. It is an unbias estimator of parameter μ, then

$$\mu = \hat{x} - \epsilon. \tag{4}$$

As known $(-\epsilon) \sim \mathcal{N}(0, 1)$, because of the symmetry of the bell-shaped curve about its central point, i.e.

$$\mu|\hat{x} \sim \mathcal{N}(\hat{x}, 1). \tag{5}$$

It means that we construct the confidence density of the parameter

$$\tilde{\varphi}(\mu|\hat{x}) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(\hat{x}-\mu)^2}{2}} \tag{6}$$

by the single way for each value of $\hat{x}$.

Motivation (III) The presence of invariant

The construction above is direct consequence of the identity

$$\int_{-\infty}^{\hat{x}-\alpha_1} \varphi(x|\hat{x})dx + \int_{\hat{x}-\alpha_1}^{\hat{x}+\alpha_2} \tilde{\varphi}(\mu|\hat{x})d\mu + \int_{\hat{x}+\alpha_2}^{\infty} \varphi(x|\hat{x})dx = 1, \quad (7)$$

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

The presence of the identities of such type (Eq.7) is a property of statistically self-dual distributions (S. Bityukov, V. Taperechkina, V. Smirnova (2004); S. Bityukov, N. Krasnikov (2005)):

normal and normal

$$\varphi(x|\mu,\sigma) = \tilde{\varphi}(\mu|x,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \sigma = const$$

Cauchy and Cauchy

$$f(x|\mu,b) = \tilde{f}(\mu|x,b) = \frac{b}{\pi(b^2+(x-\mu)^2)}, \quad b = const$$

Laplace and Laplace

$$f(x|\mu,b) = \tilde{f}(\mu|x,b) = \frac{1}{2b}e^{-\frac{|x-\mu|}{b}}, \quad b = const$$

These identities allow to reconstruct the corresponding confidence densities by the single way.

Motivation (IV) The invariant in the case of asymmetric distributions

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

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

In this case (here estimator of parameter with bias) we can use another identity as invariant for reconstruction of confidence density (S. Bityukov, N. Krasnikov, V. Taperechkina (2000); S. Bityukov, N Krasnikov (2002)) by the single way (any another reconstruction is inconsistent with the identity and, correspondingly, breaks the conserving of probability):

$$\sum_{i=\hat{x}+1}^{\infty} \frac{\mu_1^i e^{-\mu_1}}{i!} + \int_{\mu_1}^{\mu_2} \frac{\mu^{\hat{x}} e^{-\mu}}{\hat{x}!} d\mu + \sum_{i=0}^{\hat{x}} \frac{\mu_2^i e^{-\mu_2}}{i!} = 1 \qquad (8)$$

for any real $\mu_1 \geq 0$ and $\mu_2 \geq 0$ and non-negative integer $\hat{x}$, i.e.

$$\sum_{i=\hat{x}+1}^{\infty} f(i|\mu_1) + \int_{\mu_1}^{\mu_2} \tilde{f}(\mu|\hat{x}) d\mu + \sum_{i=0}^{\hat{x}} f(i|\mu_2) = 1, \qquad (9)$$

where $f(i|\mu) = \tilde{f}(\mu|i) = \frac{\mu^i e^{-\mu}}{i!}$. $\tilde{f}(\mu|i)$ is a pdf of Gamma-distribution $\Gamma_{1,i+1}$ and $\hat{x}$ usually is a number of observed events.

A bit of history

Point estimators, confidence intervals and p-values have long been fundamental tools for frequentist statisticians. Confidence distributions (CDs), which can be viewed as "distribution estimators", are often convenient devices for constructing all the above statistical procedures plus more. The basic notion of CDs traces back to the fiducial distribution of Fisher (1930); however, it can be viewed as a pure frequentist concept. Indeed, as pointed out in Schweder, Hjort (2002) the CD concept is "Neymannian interpretation of Fisher's fiducial distribution" [Neyman (1941)]. Its development has proceeded from Fisher (1930) though recent contributions, just to name a few, of Efron (1993; 1998), Fraser (1991; 1996), Lehmann (1993), Singh, Xie, Strawderman (2001; 2005; 2007), Schweder, Hjort (2002) and others. In the works (Bityukov, Krasnikov (2002); Bityukov, Taperechkina, Smirnova (2004)) the approach for reconstruction of the confidence distribution densities by the using of the corresponding identities is developed. Recently (Bickel (2006)), the method for incorporating expert knowledge into frequentist approach by combining generalized confidence distributions is proposed.

So, this methodology is in active progressing.

Confidence distributions (I)

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

Definition 1 (Singh, Xie, Strawderman (2005)): 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 $\mathcal{X}_n \in \chi$, $H_n(\cdot)$ is a continuous cumulative distribution function;**
***(ii)* at the true parameter value $\theta = \theta_0, H_n(\theta_0) = H_n(\mathcal{X}_n, \theta_0)$, as a function of the sample $\mathcal{X}_n$, 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 $\theta = \theta_0, H_n(\mathcal{X}_n, \theta_0) \xrightarrow{W} U(0, 1)$ as $n \to +\infty$, and the continuity requirment on $H_n(\cdot)$ is dropped.

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

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

Item *(ii)* basically states that the function $H_n(\cdot)$ contains the right amount of information about the true θ_0.

Confidence distributions (II)

It follows from the definition of CD that if $\theta < \theta_0,\ H_n(\theta) \overset{sto}{\leq} 1 - H_n(\theta)$, and if $\theta > \theta_0,\ 1 - H_n(\theta) \overset{sto}{\leq} H_n(\theta)$. Here $\overset{sto}{\leq}$ is a stochastic comparison between two random variables; i.e. for two random variable Y_1 and Y_2, $Y_1 \overset{sto}{\leq} Y_2$, if $P(Y_1 \leq t) \geq P(Y_2 \leq t)$ for all t. Thus a CD works, in a sense, like a compass needle. It points towards θ_0, when placed at $\theta \neq \theta_0$, by assigning more mass stochastically to that side (left or right) of θ that contains θ_0. When placed at θ_0 itself, $H_n(\theta) = H_n(\theta_0)$ has the uniform $U[0,1]$ distribution and thus it is noninformative in direction.

For every α in $(0,1)$, let $(-\infty, \xi_n(\alpha)]$ be a $100\alpha\%$ lower-side confidence interval, where $\xi_n(\alpha) = \xi_n(\mathcal{X}_n, \alpha)$ is continuous and increasing in α for each sample $\mathcal{X}_n$. Then $H_n(\cdot) = \xi_n^{-1}(\cdot)$ is a CD in the usual Fisherian sense. In this case,

$$\{\mathcal{X}_n : H_n(\theta) \leq \alpha\} = \{\mathcal{X}_n : \theta \leq \xi_n(\alpha)\} \tag{10}$$

for any α in $(0,1)$ and θ in $\Theta \subseteq \mathrm{R}$. Thus, at $\theta = \theta_0, Pr\{H_n(\theta_0) \leq \alpha\} = \alpha$ and $H_n(\theta_0)$ is $U(0,1)$ distributed.

Definition 1 is very convenient for the purpose of verifying if a particular function is a CD or an aCD.

Confidence distributions (III)

There are another definition of confidence distribution.

Consider data X randomly drawn from a distribution parametrized in part by θ, the scalar parameter with true value θ_0. Define a distribution generator as a function T that maps X to a distribution of θ.

Definition 2 (Bickel (2006)): **If $F_{T(X)}(\theta)$, the corresponding cumulative distribution function, is uniform $U(0,1)$ when $\theta = \theta_0$, then $T(X)$ is called a confidence distribution for θ.**

It follows that, for continuous data, the $(1-\alpha_1-\alpha_2)$100% confidence interval $(F^{-1}_{T(X)}(\alpha_1), F^{-1}_{T(X)}(1-\alpha_2))$ has exact coverage in the sense that it includes the true value of the parameter with relative frequency $(1-\alpha_1-\alpha_2)$ over an infinite number of realizations of X and, if random, T.
Here $\alpha_1+\alpha_2 \leq 1, \alpha_1 \geq 0, \alpha_2 \geq 0$.

Examples and inferential information contained in a CD (I)

Example 1 Normal mean and variance (Singh, Xie, Strawderman (2005)): Suppose $X_1, X_2, \ldots, X_n$ is a sample from $\mathcal{N}(\mu, \sigma^2)$, with both μ and σ^2 unknown. A CD for μ is $H_n(y) = F_{t_{n-1}}(\frac{y - \bar{X}}{s_n/\sqrt{n}})$, where $\bar{X}$ and s^2 are, respectively, the sample mean and variance, and $F_{t_{n-1}}(\cdot)$ is a 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.

Example 2 p-value function (Singh, Xie, Strawderman (2005)): For any given $\tilde{\theta}$, let $p_n(\tilde{\theta}) = p_n(\mathcal{X}_n, \tilde{\theta})$ be a p-value for a one-sided test $K_0 : \theta \leq \tilde{\theta}$ versus $K_0 : \theta > \tilde{\theta}$. Assume that the p-value is available for all $\tilde{\theta}$. The function $p_n(\cdot)$ is called a p-value function. Typically, at the true value $\theta = \theta_0$, $p_n(\theta_0)$ as a function of $\mathcal{X}_n$ is exactly (or asymptotically) $U(0,1)$-distributed. Also, $H_n(\cdot) = p_n(\cdot)$ for every fixed sample is almost always a cumulative distribution function. Thus, usually $p_n(\cdot)$ satisfies the requirements for a CD(or aCD).

Examples and inferential information contained in a CD (II)

Example 3 Likelihood functions (Singh, Xie, Strawderman (2005)): There is a connection between the concepts of aCD and various types of likelihood functions. In an exponential family, both the profile likelihood and the implied likelihood (Efron(1993)) are aCD densities after a normalization. The work (Singh, Xie, Strawderman (2001)) provided a formal proof, with some specific conditions, which shows that $e^{l_n^*(\theta)}$ is proportional to an aCD density for the parameter θ, where $l_n^*(\theta) = l_n(\theta) - l_n(\hat{\theta})$, $l_n(\theta)$ is the log-profile likelihood function, and $\hat{\theta}$ is the maximum likelihood estimator of θ.

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.

Inference: a brief summary (Singh, Xie, Strawderman (2005))

- *Confidence interval.* **From the definition, it is evident that the intervals** $(-\infty, H_n^{-1}(1-\alpha)], [H_n^{-1}(\alpha), +\infty)$ **and** $(H_n^{-1}(\alpha/2), H_n^{-1}(1-\alpha/2))$ **provide** $100(1-\alpha)\%$**-level confidence intervals of different kinds for** θ**, for any** $\alpha \in (0,1)$**. The same is true for an aCD, where the confidence level is achieved in limit.**

- *Point estimation.* **Natural choices of point estimators of the parameter** θ**, given** $H_n(\theta)$**, include the median** $M_n = H_n^{-1}(1/2)$**, the mean** $\bar{\theta} = \int_{-\infty}^{\infty} t dH_n(t)$ **and the maximum point of the CD density** $\hat{\theta} = arg\ max_\theta h_n(\theta),\ h_n(\theta) = H_n^{'}(\theta)$**.**

- *Hypothesis testing.* **From a CD, one can obtain p-values for various hypothesis testing problems. The work (Fraser (1991)) developed some results on such a topic through** p**-value functions. The natural line of thinking is to measure the support that** $H_n(\cdot)$ **lends to a null hypothesis** $K_0 : \theta \in C$**. There are possible two types of support:**

  1. **Strong-support** $p_s(C) = \int_C dH_n(\theta)$**.**
  2. **Weak-support** $p_w(C) = sup_{\theta \in C} 2min(H_n(\theta), 1-H_n(\theta))$**.**

  If K_0 **is of the type** $(-\infty, \theta_0]$ **or** $[\theta_0, +\infty)$ **or a union of finitely many intervals, the strong-support** $p_s(C)$ **leads to the classical p-values.**

  If K_0 **is a singleton, that is,** K_0 **is** $\theta = \theta_0$**, then the weak-support** $p_w(C)$ **leads to the classical p-values.**

Combination of CDs (Singh, Xie, Strawderman (2005))

The notion of a CD (or aCD) is attractive for the purpose of combining information. The main reasons are that there is a wealth of information on θ inside a CD, the concept of CD is quite broad, and the CDs are relatively easy to construct and interpret.

Let $H_1(y), \ldots, H_L(y)$ be L independent CDs, with the same true parameter value θ_0. Suppose $g_c(u_1, \ldots, u_L)$ is any continuous function from $[0,1]^L$ to R that is monotonic in each coordinate. A general way of combining, depending on $g_c(u_1, \ldots, u_L)$ can be described as follows: Define $H_c(u_1, \ldots, u_L) = G_c(g_c(u_1, \ldots, u_L))$, where $G_c(\cdot)$ is the continuous cumulative distribution function of $g_c(u_1, \ldots, u_L)$, and $U_1, \ldots, U_L$ are independent $U(0,1)$ distributed random variables.

Denote $H_c(y) = H_c(H_1(y), \ldots, H_L(y))$. It is easy to verify that $H_c(y)$ is a CD function for the parameter θ. $H_c(y)$ is a combined CD.

Let $F_0(\cdot)$ be any continuous cumulative distribution function and $F_0^{-1}(\cdot)$ be its inverse function. A convenient special case of the function g_c is

$$g_c(u_1, \ldots, u_L) = F_0^{-1}(u_1) + \ldots + F_0^{-1}(u_L).$$

In this case, $G_c(\cdot) = F_0 * \ldots * F_0(\cdot)$, where $*$ stands for convolution. Just like the p-value combination approach, this general CD combination recipe is simple and easy to implement. For example, let $F_0(t) = \Phi(t)$ is the cumulative distribution function of the standard normal. In this case

$$H_{NM}(y) = \Phi(\frac{1}{\sqrt{L}}[\Phi^{-1}(H_1(y)) + \ldots + \Phi^{-1}(H_L(y))])$$

Conclusion

The notion of confidence distribution, an enterely 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.

The follow quotation from Efron(1998) on Fisher's contribution of the fiducial distribution seems quite relevant in the context of CDs:
"... but here is a safe prediction for the 21st century: statisticians will be asked to solve bigger and more complicated problems. I believe there is a good chance that objective Bayes methods will be developed for such problem, and that something like fiducial inference will play an important role in this development. Maybe Fisher's biggest blunder will become a big hit in the 21st centure!"

We are grateful to Bob Cousins, Vladimir Gavrilov, Vassili Kachanov, Louis Lyons, Victor Matveev and Nikolai Tyurin for interest and support of this work. We thank Prof. Kesar Singh and Prof. Minge Xie for helpful and constructive comments.
We also thank Yuri Gouz, Alexandre Nikitenko and Claudia Wulz for very useful discussions.

References

D.R. Bickel (2006), Incorporating expert knowledge into frequentist interface by combining generalized confidence distributions, e-Print: math/0602377, 2006.

S.I. Bityukov, N.V. Krasnikov, V.A. Taperechkina (2000), Confidence intervals for Poisson distribution parameter, Preprint IFVE 2000-61, Protvino, 2000; also, e-Print: hep-ex/0108020, 2001.

S.I.Bityukov (2002), Signal Significance in the Presence of Systematic and Statistical Uncertainties, JHEP 09 (2002) 060; e-Print: hep-ph/0207130; S.I. Bityukov, N.V. Krasnikov, NIM A502 (2003) 795-798.

S.I. Bityukov, V.A. Taperechkina, V.V. Smirnova (2004), Statistically dual distributions and estimation of the parameters, e-Print: math.ST/0411462, 2004.

S.I. Bityukov, N.V. Krasnikov (2005), Statistically dual distributions and conjugate families, in Proceedings of MaxEnt'05, August 2005, San Jose, CA, USA.

B. Efron (1978), Controversies in the Foundations of Statistics, The American Mathematical Monthly, 85(4) (1978) 231-246.

B. Efron (1993), Bayes and likelihood calculations from confidence intervals. Biometrika 80 (1993) 3-26. MR1225211.

B. Efron (1998), R.A. Fisher in the 21st Century. Stat.Sci. 13 (1998) 95-122.

R.A. Fisher (1930), Inverse probability. Proc. of the Cambridge Philosophical Society 26 (1930) 528-535.

D.A.S. Fraser (1991), Statistical inference: Likelihood to significance. J. Amer. Statist. Assoc. 86 (1991) 258-265. MR1137116.

D.A.S. Fraser (1996), Comments on "Pivotal inference and the fiducial argument", by G. A. Barnard. Internat. Statist. Rev. 64 (1996) 231-235.

E.L. Lehmann (1993). The Fisher, Neyman-Pearson theories of testing hypotheses:One theory or two? J. Amer. Statist. Assoc. 88 (1993) 1242-1249. MR1245356.

J. Neyman (1941), Fiducial argument and the theory of confidence intervals. Biometrika 32 (1941) 128-150. MR5582.

T. Schweder and N.L. Hjort (2005), Confidence and likelihood, Scand. J. Statist. 29 (2002) 309-332.

K. Singh, M. Xie, W. Strawderman. (2001), Confidence distributions - concept, theory and applications, Technical report, Dept.Statistics, Rutgers Univ., Revised 2004

K. Singh, M. Xie, W.E. Strawderman (2005), Combining information from independent sources through confidence distributions, The Annals of Statistics 33 (2005) 159-183.

K. Singh, M. Xie, W. Strawderman. (2007), Confidence distributions (CD) - Distribution Estimator of a Parameter. Yehuda Vardi Volume. IMS-LNMS. (to appear).