Ancillary Statistics

In a parametric model $f(\mathbf{y};\boldsymbol{\theta})$ for a **random variable** or vector $\mathbf{Y}$, a statistic $\mathbf{A} = a(\mathbf{Y})$ is ancillary for $\boldsymbol{\theta}$ if the distribution of $\mathbf{A}$ does not depend on $\boldsymbol{\theta}$. As a very simple example, if $\mathbf{Y}$ is a vector of independent, identically distributed random variables each with mean $\boldsymbol{\theta}$, and the sample size is determined randomly, rather than being fixed in advance, then $\mathbf{A}$ = number of observations in $\mathbf{Y}$ is an ancillary statistic. This example could be generalized to more complex structure for the observations $\mathbf{Y}$, and to examples in which the sample size depends on some further parameters that are unrelated to $\boldsymbol{\theta}$. Such models might well be appropriate for certain types of sequentially collected data arising, for example, in clinical trials.

Fisher [5] introduced the concept of an ancillary statistic, with particular emphasis on the usefulness of an ancillary statistic in recovering **information** that is lost by reduction of the sample to the **maximum likelihood** estimate $\hat{\boldsymbol{\theta}}$, when the maximum likelihood estimate is not minimal **sufficient**.

An illustrative, if somewhat artificial, example is a sample $(Y_1, \ldots, Y_n)$, where now n is fixed, from the uniform distribution on $(\theta, \theta + 1)$. The largest and smallest observations, $(Y_{(1)}, Y_{(n)})$, say, form a minimal sufficient statistic for θ. The maximum likelihood estimator of θ is any value in the interval $(Y_{(n)} - 1, Y_{(1)})$, and the **range** $Y_{(n)} - Y_{(1)}$ is an ancillary statistic. In this example, while the range does not provide any information about the value of θ that generated the data, it does provide information on the precision of $\hat{\theta}$. In a sample for which the range is 1, $\hat{\theta}$ is exactly equal to θ, whereas a sample with a range of 0 is the least informative about θ.

A theoretically important example discussed in Fisher [5] is the location model (*see* **Location–Scale Family**), in which $\mathbf{Y} = (Y_1, \ldots, Y_n)$, and each Y_i follows the model $f(y - \theta)$, with $f(\cdot)$ known but θ unknown. The vector of **residuals** $\mathbf{A} = (Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$, where $\overline{Y} = n^{-1}\sum Y_i$, has a distribution free of θ, as is intuitively obvious, since both Y_i and $\overline{Y}$ are centered at θ. The uniform example discussed above is a special case of the location model, and the range $Y_{(n)} - Y_{(1)}$ is also an ancillary statistic for the present example. In fact, the vector $\mathbf{B} = (Y_{(2)} - Y_{(1)}, \ldots, Y_{(n)} - Y_{(1)})$ is also ancillary, as is $\mathbf{C} = (Y_1 - \hat{\theta}, \ldots, Y_n - \hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimate of θ. A *maximal* ancillary provides the largest possible conditioning set, or the largest possible reduction in dimension, and is analogous to a minimal sufficient statistic. **A**, **B**, and **C** are maximal ancillary statistics for the location model, but the range is only a maximal ancillary in the location uniform.

An important property of the location model is that the exact conditional distribution of the maximum likelihood estimator $\hat{\theta}$, given the maximal ancillary **C**, can be easily obtained simply by renormalizing the **likelihood** function:

$$p(\hat{\theta}|\mathbf{c};\theta) = \frac{L(\theta;\mathbf{y})}{\int L(\theta;\mathbf{y})\,\mathrm{d}\theta}, \qquad (1)$$

where $L(\theta;\mathbf{y}) = \prod f(y_i;\theta)$ is the likelihood function for the sample $\mathbf{y} = (y_1, \ldots, y_n)$, and the right-hand side is to be interpreted as depending on $\hat{\theta}$ and $\mathbf{c}$, using the equations $\sum \partial\{\log f(y_i;\theta)\}/\partial\theta|_{\hat{\theta}} = 0$ and $\mathbf{c} = \mathbf{y} - \hat{\theta}$.

The location model example is readily generalized to a **linear regression** model with nonnormal errors. Suppose that, for $i = 1, \ldots, n$, we have independent observations from the model $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i$, where the distribution of ε_i is known. The vector of standardized residuals $(Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/\hat{\sigma}$ is ancillary for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ and there is a formula similar to (1) for the distribution of $\hat{\boldsymbol{\theta}}$, given the residuals.

It is possible, and has been argued, that Fisher's meaning of ancillarity included more than the requirement of a distribution free of $\boldsymbol{\theta}$: that it included a notion of a physical mechanism for generating the data in which some elements of this mechanism are "clearly" not relevant for assessing the value of $\boldsymbol{\theta}$, but possibly relevant for assessing the accuracy of the inference about $\boldsymbol{\theta}$. Thus, Kalbfleisch [7] makes a distinction between an experimental and a mathematical ancillary statistic. Fraser [6] developed the notion of a structural model as a physically generated extension of the location model. Efron & Hinkley [4] gave particular attention to the role of an ancillary statistic in estimating the variance of the maximum likelihood estimator.

Two generalizations of the concept of ancillary statistic have become important in recent work in the theory of **inference**. The first is the notion of approximate ancillarity, in which the distribution of **A** is not required to be entirely free of $\boldsymbol{\theta}$, but free of $\boldsymbol{\theta}$ to some order of approximation. For example, we might require that the first few

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a15002

moments of $\mathbf{A}$ be constant (in $\boldsymbol{\theta}$), or that the distribution of $\mathbf{A}$ be free of $\boldsymbol{\theta}$ in a neighborhood of the true value $\boldsymbol{\theta}_0$, say. The definition used by Barndorff-Nielsen & Cox [1] is that $\mathbf{A}$ is qth order locally ancillary for $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$ if $f(\mathbf{a};\boldsymbol{\theta}_0+\delta/\sqrt{n}) = f(\mathbf{a};\boldsymbol{\theta}_0)+O(n^{-q/2})$. Approximate ancillary statistics are also discussed in McCullagh [9] and Reid [11]. The notion of an approximate ancillary statistic has turned out to be rather important for the asymptotic theory of statistical inference, because the location family model result given in (1) can be generalized, to give the result

$$p(\hat{\boldsymbol{\theta}}|\mathbf{a};\boldsymbol{\theta}) \doteq c(\boldsymbol{\theta},\mathbf{a})|j(\hat{\boldsymbol{\theta}})|^{1/2}\frac{L(\boldsymbol{\theta};\mathbf{y})}{L(\hat{\boldsymbol{\theta}};\mathbf{y})}, \qquad (2)$$

where c is a normalizing constant, $j(\boldsymbol{\theta}) = -\partial^2 \log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ is the observed Fisher information function, $\mathbf{a}$ is an approximately ancillary statistic, and in the right-hand side $\mathbf{y}$ is a function of $\hat{\boldsymbol{\theta}}, \mathbf{a}$. This approximation, which is typically much more accurate than the normal approximation to the distribution of $\hat{\boldsymbol{\theta}}$, is known as Barndorff-Nielsen's approximation, or the p^* approximation, and is reviewed in Reid [10] and considered in detail in Barndorff-Nielsen & Cox [1]. In (2) the likelihood function is normalized by a slightly more elaborate looking formula than the simple integral in (1), but the principle of renormalizing the likelihood function has still been applied. A distribution function approximation analogous to (2) is also available: see Barndorff-Nielsen & Cox [1] and Reid [11].

Suppose that the parameter $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, where $\boldsymbol{\psi}$ is the parameter of interest and $\boldsymbol{\lambda}$ is a **nuisance parameter**. For example, $\boldsymbol{\psi}$ might parameterize a regression model for survival time as a function of several covariates, and $\boldsymbol{\lambda}$ might parameterize the baseline **hazard** function. If we can partition the minimal sufficient statistic for $\boldsymbol{\theta}$ as $(\mathbf{S}, \mathbf{T})$, such that

$$f(\mathbf{s},\mathbf{t};\boldsymbol{\theta}) = f(\mathbf{s}|\mathbf{t};\boldsymbol{\psi})f(\mathbf{t};\boldsymbol{\lambda}), \qquad (3)$$

then $\mathbf{T}$ is an ancillary statistic for $\boldsymbol{\psi}$ in the sense of the above definition. Factorizations of the form given in (3) are the exception, though, and we more often have a factorization of the type

$$f(\mathbf{s},\mathbf{t};\boldsymbol{\theta}) = f(\mathbf{s}|\mathbf{t};\boldsymbol{\psi})f(\mathbf{t};\boldsymbol{\psi},\boldsymbol{\lambda}) \qquad (4)$$

or

$$f(\mathbf{s},\mathbf{t};\boldsymbol{\theta}) = f(\mathbf{s}|\mathbf{t};\boldsymbol{\psi},\boldsymbol{\lambda})f(\mathbf{t};\boldsymbol{\lambda}). \qquad (5)$$

An example of (4) is the **two-by-two table**, with ψ the log **odds ratio**. The conditional distribution of a single cell entry, given the row and column totals, depends only on ψ: this is the basis for **Fisher's exact test** (*see* **Conditionality Principle**). Although it is sometimes claimed that the row total is an ancillary statistic for the parameter of interest, this is in fact not the case, at least according to the definition of ancillarity discussed here. Some more general notions of ancillarity have been proposed in the literature, but have not proved to be widely useful in theoretical developments. Further discussion of ancillarity and conditional inference in the presence of nuisance parameters can be found in Liang & Zeger [8] and Reid [10].

Ancillary statistics are defined for parametric models, so would not be defined, for example, in Cox's **proportional hazards** regression model (*see* **Cox Regression Model**). Cox [2] did originally argue, though, that the full likelihood function could be partitioned into a factor that provided information on the regression parameters $\boldsymbol{\beta}$ and a factor that provided no information about $\boldsymbol{\beta}$ in the absence of knowledge of the baseline hazard: the situation is analogous to (4) but, as was pointed out by several discussants of [2], the likelihood factor that is used in the analysis is not in fact the conditional likelihood for any observable random variables. Cox [3] developed the notion of **partial likelihood** to justify the now standard estimates of $\boldsymbol{\beta}$.

References

[1] Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London. (This is the only book currently available that gives a survey of many of the main ideas in statistical theory along with a detailed discussion of the asymptotic theory for likelihood inference that has been developed since 1980 (see **Large-sample Theory**). Chapter 2.5 discusses ancillary statistics, and Chapters 6 and 7 consider the p^* approximation and the related distribution function approximation.).

[2] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[3] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[4] Efron, B. & Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion), *Biometrika* **65**, 457–487.

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a15002

[5] Fisher, R.A. (1934). Two new properties of mathematical likelihood, *Proceedings of the Royal Society, Series A* **144**, 285–307.
[6] Fraser, D.A.S. (1968). *The Structure of Inference*. Wiley, New York.
[7] Kalbfleisch, J.D. (1975). Sufficiency and conditionality, *Biometrika* **62**, 251–259.
[8] Liang, K.-Y. & Zeger, S.L. (1995). Inference based on estimating functions in the presence of nuisance parameters, *Statistical Science* **10**, 158–172.
[9] McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman & Hall, London.
[10] Reid, N. (1988). Saddlepoint approximations in statistical inference, *Statistical Science* **3**, 213–238. (A review of the p^* approximation and related developments, up to 1987.).
[11] Reid, N. (1995). The roles of conditioning in inference, *Statistical Science* **10**, 138–157.

Further Reading

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator, *Biometrika* **70**, 343–365. (Among the early papers on the p^* approximation, this is one of the easier ones.)
Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio, *Biometrika* **73**, 307–322. (Introduces the distribution function approximation based on the p^* approximation.)
Fisher, R.A. (1973). *Statistical Methods and Scientific Inference*, 3rd Ed. Oliver & Boyd, Edinburgh. (Includes a discussion of ancillary statistics with many examples.)
Jorgensen, B. (1994). The rules of conditional inference: is there a universal definition of nonformation?, *Journal of the Italian Statistical Society* **3**, 355–384. (This considers several definitions of ancillarity in the presence of nuisance parameters.)
Kalbfleisch, J.G. (1985). *Probability and Statistical Inference*, Vol. II, 2nd Ed. Springer-Verlag, New York. (This is one of the few undergraduate textbooks that treats ancillarity in any depth, in Chapter 15.)

N. REID

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a15002