Constructing Ensembles of Pseudo-Experiments

Luc Demortier
The Rockefeller University, New York, NY 10021, USA

The frequentist interpretation of measurement results requires the specification of an ensemble of independent replications of the same experiment. For complex calculations of bias, coverage, significance, etc., this ensemble is often simulated by running Monte Carlo pseudo-experiments. In order to be valid, the latter must obey the Frequentist Principle and the Anticipation Criterion. We formulate these two principles and describe some of their consequences in relation to stopping rules, conditioning, and nuisance parameters. The discussion is illustrated with examples taken from high-energy physics.

1. INTRODUCTION

Many statistical analyses in physics are based on a frequency interpretation of probability. For example, the result of measuring a physical constant θ can be reported in the form of a $1 - \alpha$ confidence interval $[X_1, X_2]$, with the understanding that if the measurement is replicated a large number of times, one will have $X_1 \leq \theta \leq X_2$ in a fraction $1 - \alpha$ of the replications. This type of interpretation therefore requires the definition of a *reference set* of similar measurements:

> *The reference set of a measurement is the ensemble of experiments in which the actually performed experiment is considered to be embedded for the purpose of interpreting its results in a frequentist framework.*

A major appeal of frequentism among physicists is its empirical definition of probability. By the strong law of large numbers, probabilities can be approximated in finite ensembles, and such approximations converge to the true value as the ensemble size increases. In other words, frequentist confidence statements are experimentally verifiable.

Physicists use Monte Carlo generated ensembles in various applications: to check a fitting algorithm for the presence of bias, non-Gaussian pulls, or other pathologies; to calculate the coverage of confidence intervals or upper limits; to average out statistical fluctuations in order to isolate systematic effects; to calculate goodness-of-fit measures and significances; to design experiments; etc. When constructing ensembles to address these questions, one needs to pay attention to a number of subtle issues that arise in a frequentist framework: what is the correct stopping rule?; is it appropriate to condition, and if so, on what statistic?; how should nuisance parameters be handled?

The aim of this paper is to draw attention to these issues and to propose some recommendations where possible. We start by discussing basic frequentist principles in section 2 and illustrate them with an example of conditioning in section 3. The importance of stopping rules is argued in section 4. Finally, some

purely frequentist methods to handle nuisance parameters are described in section 5.

2. FREQUENTIST PRINCIPLES

In order to deserve the label frequentist, a statistical procedure and its associated ensemble must satisfy two core principles, which we examine in the next two subsections.

2.1. The Frequentist Guarantee

The first principle states the aims of frequentism:

> **Frequentist Guarantee [1]:**
> *In repeated use of a statistical procedure, the long-run average actual accuracy should not be less than (and ideally should equal) the long-run average reported accuracy.*

To clarify this principle, we return to the $1 - \alpha$ confidence interval procedure mentioned in the Introduction. Let $\mathcal{E}$ be an ensemble of intervals obtained by applying this procedure many times on different, independent data. The *actual accuracy* of an interval in $\mathcal{E}$ is 1 or 0: either the interval covers the true value of the parameter of interest, or it does not. The average actual accuracy is therefore simply the fraction of intervals in $\mathcal{E}$ that cover. On the other hand, the average *reported accuracy* is $1 - \alpha$. The reported accuracy is often the same for all intervals in $\mathcal{E}$, but in some settings it is possible to report a different, data-dependent accuracy for each interval. Thus, averaging the reported accuracy is not necessarily a trivial operation. A procedure that satisfies the Frequentist Guarantee is said to have coverage.

In a sense, the Frequentist Guarantee is only weakly constraining, because it does not require a procedure to have coverage when applied to repeated measurements of the *same* quantity. To see how this is relevant, consider the construction of a 68% confidence interval for the mean μ of a Poisson distribution. One procedure is to take all μ values satisfying

$(n - \mu)^2/\mu \leq 1$, where n is the observed number of events. The resulting interval actually undercovers for many values of μ and overcovers for other values, so that the Frequentist Guarantee appears to be satisfied *on average*. To make this statement more precise we need a weighting function with which to carry out the average over μ. A simple proposal is to perform local smoothing of the coverage function, resulting in local average coverage [1].

Physicists may object to this notion of local average coverage on the grounds that they sometimes repeatedly measure a given constant of nature and are then interested in the coverage obtained for that particular constant, not in an average coverage over "nearby" constants. A possible answer is that one rarely measures the quantity of interest directly. Rather, one measures a combination of the quantity of interest with calibration constants, efficiencies, sample sizes, etc., all of which vary from one measurement to the next, so that an effective averaging does take place.

Finally, it could be argued that even Bayesians should subscribe to some form of the Frequentist Guarantee. If, over repeated use, a 95% credible Bayesian interval fails to cover the true value more than 30% of the time (say), then there must be something seriously wrong with that interval.

2.2. The Anticipation Criterion

Although the Frequentist Guarantee specifies how a statistical procedure should behave under many repetitions of a measurement, it does not indicate what constitutes a valid repetition, and hence a valid ensemble. To the extent that this question involves the notion of randomness, it is well beyond the scope of this paper. From a practical standpoint however, one would like to stipulate that all effects susceptible to interfere with that randomness must be recognized as such and included in the construction of the ensemble, i.e. "anticipated"[2]. Hence the second principle:

> **Anticipation Criterion:**
> *Ensembles must anticipate all elements of chance and all elements of choice of the actual experiments they serve to interpret.*

To clarify, "elements of chance" refers to statistical fluctuations of course, but also to systematic uncertainties when the latter come from nuisance parameters that are determined by auxiliary measurements. On the other hand, "elements of choice" refers to actions by experimenters, in particular how they decide to stop the experiment, and what decisions they make after stopping.

One can identify several levels of anticipation. At the highest level, the data collection and analysis methods, as well as the reference ensemble used to interpret results, are fully specified at the outset. They do not change once the data is observed. The reference ensemble is called "unconditional".

At the second highest level, the data collection and analysis methods are fully specified at the outset, but the reference ensemble is not. The latter will be fully determined once the data is observed, and is therefore "conditional". Although a conditional ensemble is not known before observing the data, it is a subset in a known partition of a known unconditional ensemble.

The lowest level of anticipation is occupied by Bayesian methods, which fully condition on the observed data. The reference ensemble collapses to a point and can therefore no longer be used as a reference.

As the level of anticipation decreases, the reference ensemble becomes smaller. A remarkable result is that within the second level of anticipation one can refine the conditioning partition to the point where it is possible to give a Bayesian interpretation to frequentist conclusions, and vice-versa [3].

3. CONDITIONING

To illustrate the interplay between anticipation and conditioning, we present here a famous example originally due to Cox [4]. Suppose we make one observation of a rare particle and wish to estimate its mass μ from the momenta of its decay products. For the sake of simplicity, assume that the estimator X of μ is normal with mean μ and variance σ^2. There is a 50% chance that the particle decays hadronically, in which case $\sigma = 10$; otherwise the particle decays leptonically and $\sigma = 1$. Consider the following 68% confidence interval procedures:

1. Unconditional
   If the particle decayed hadronically, report $X \pm \delta_h$, otherwise report $X \pm \delta_\ell$, where δ_h and δ_ℓ are chosen so as to minimize the expected length $\langle \delta \rangle = \delta_h + \delta_\ell$ subject to the constraint of 68% coverage. This yields $\delta_\ell = 2.20$ and $\delta_h = 5.06$. The expected length is 7.26.

2. Conditional
   If we condition on the decay mode, then the best interval is $X \pm 10$ if the particle decayed hadronically, and $X \pm 1$ otherwise. So the expected length is 11.0 in this case.

Note that in both cases we used all the information available: the measurement X as well as the decay mode. Both procedures are valid; the only difference between them is the reference frame. The unconditional ensemble includes both decay modes, whereas the conditional one only includes the observed decay mode.

The expected length is shorter for unconditional intervals than for conditional ones. Does this mean we

should quote the former? If our aim is to report what we learned from the data we observed, then clearly we should report the conditional interval. Suppose indeed that we observed a hadronic decay. The unconditional interval width is then 10.12, compared to 20.0 for the conditional one. The reason the unconditional interval is shorter is that, if we could repeat the experiment, we might observe the particle decaying into the leptonic mode. However, this is irrelevant to the interpretation of the observation we actually made. This example illustrates a general feature of conditioning, that it usually increases expected length, and reduces power in test settings.

Another aspect of the previous example is that the conditioning statistic (the decay mode) is ancillary: its distribution does not depend on the parameter of interest (the particle mass). This is not always the case. Suppose for example that we are given a sample from a normal distribution with unit variance and unknown mean θ, and that we wish to test $H_0 : \theta = -1$ versus $H_1 : \theta = +1$. The standard symmetric Neyman-Pearson test based on the sample mean $\bar{X}$ as test statistic rejects H_0 if $\bar{X} > 0$. It makes no distinction between $\bar{X} = 0.5$ and $\bar{X} = 5$, even though in the latter case we certainly feel more confident in our rejection of H_0. Although $\bar{X}$ is not ancillary, it is possible to use it to calculate a conditional "measure of confidence" to help characterize one's decision regarding H_0 [5]. Unfortunately, a general theory for choosing such conditioning statistics does not exist.

4. STOPPING RULES

Stopping rules specify how an experiment is to be terminated. High-energy physics experiments are often sequential, so it is important to properly incorporate stopping rules in the construction of ensembles.

As a first example, consider the measurement of the branching fraction θ for the decay of a rare particle A into a particle B. Suppose we observe a total of $n = 12$ decays, $x = 9$ of which are $A \rightarrow B$ transitions, and the rest, $r = 3$, are $A \nrightarrow B$ transitions. We wish to test $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$.

A possible stopping rule is to stop the experiment after observing a total number of decays n. The probability mass function (pmf) is then binomial:

$$f(x\,;\,\theta) \;=\; \binom{n}{x}\, \theta^x\,(1-\theta)^{n-x}, \qquad (1)$$

and the p value for testing H_0 is:

$$p_b \;=\; \sum_{i=9}^{12} \binom{12}{i}\, \theta^i\,(1-\theta)^{12-i} \;=\; 0.075. \quad (2)$$

An equally valid stopping rule is to stop the experiment after observing a number r of $A \nrightarrow B$ decays.

Now the pmf is negative binomial:

$$f(x\,;\,\theta) \;=\; \binom{r+x-1}{x}\, \theta^x\,(1-\theta)^r, \qquad (3)$$

and the p value is:

$$p_{nb} \;=\; \sum_{i=9}^{\infty} \binom{2+i}{i}\, \theta^i\,(1-\theta)^3 \;=\; 0.0325. \quad (4)$$

If we adopt a 5% threshold for accepting or rejecting H_0, we see that the binomial model leads to acceptance, whereas the negative binomial model leads to rejection.

Here is a more intriguing example [6]. Imagine a physicist working at some famous particle accelerator and developping a procedure to select collision events that contain a Higgs boson. Assume that the expected rate of background events accepted by this procedure is known very accurately. Applying his technique to a given dataset, the physicist observes 68 events and expects a background of 50. The (Poisson) probability for 50 to fluctuate up to 68 or more is 0.89%, and the physicist concludes that there is significant evidence against H_0, the background-only hypothesis, at the 1% level.

Is this conclusion correct? Perhaps the physicist just decided to take a single sample. But what would he have done if this sample had not yielded a significant result? Perhaps he would have taken another sample! So the real procedure the physicist was considering is actually of the form:

- Take a data sample, count the number n_1 of Higgs candidates, and calculate the expected background b;

- If $\mathbb{P}(N \geq n_1 \,|\, b) \leq \alpha$ then stop and reject H_0;

- Otherwise, take a second sample with the same expected background, count the number n_2 of Higgs candidates and reject H_0 if $\mathbb{P}(N \geq n_1 + n_2 \,|\, 2b) \leq \alpha$.

For this test procedure to have a level of 1%, α must be set at 0.67%. Since the *actual* data had a p value of 0.89%, the physicist should not have rejected H_0.

So now the physicist finds himself forced to take another sample. There are two interesting cases:

1. The second sample yields 57 candidate events, for a total of 125. The probability for the expected background (100 events now) to fluctuate up to 125 or more is 0.88% > 0.67%, so the result is not significant. However, the result would have been significant if the physicist had not stopped halfway through data taking to calculate the p value!

2. The second sample yields 59 candidate events, for a total of 127. The p value is now 0.52% and significance has been obtained, unless of course the physicist was planning to take a third sample in the event of no significance.

Bayesian methods are generally independent of the stopping rule. It is therefore somewhat ironic that frequentists, who start from an objective definition of probability, should end up with results that depend on the thought processes of the experimenter.

5. NUISANCE PARAMETERS

Most problems of inference involve nuisance parameters, i.e. uninteresting parameters that are incompletely known and therefore add to the overall uncertainty on the parameters of interest. To fix ideas, assume that we have a sample $\{x_1, \ldots, x_n\}$ whose probability density function (pdf) $f(\vec{x}; \mu, \nu)$ depends on a parameter of interest μ and a nuisance parameter ν, and that the latter can be determined from a separate sample $\{y_1, \ldots, y_m\}$ with pdf $g(\vec{y}; \nu)$. Correct inference about μ must then be derived from the joint pdf

$$h(\vec{x}, \vec{y}; \mu, \nu) \equiv f(\vec{x}; \mu, \nu) \, g(\vec{y}; \nu). \qquad (5)$$

What is often done in practive however, is to first obtain a distribution $\pi(\nu)$ for ν, usually by combining measurement results with a sensible guess for the form of $\pi(\nu)$. Inference about μ is then based on:

$$h'(\vec{x}; \mu) \equiv \int f(\vec{x}; \mu, \nu) \, \pi(\nu) \, d\nu. \qquad (6)$$

Although this technique borrows elements from both Bayesian and frequentist methodologies, it really belongs to neither and is more properly referred to as a hybrid non-frequentist/non-Bayesian approach.

We illustrate the handling of nuisance parameters with a simple p value calculation. Suppose that a search for a new particle ends with a sample of $n_0 = 12$ candidates over a separately measured background of $\nu_0 = 5.7 \pm 0.47$, where we ignore the uncertainty on the standard error 0.47. Let μ be the unknown expected number of new particles among the 12 candidates. We wish to test $H_0 : \mu = 0$ versus $H_1 : \mu > 0$.

A typical model for this problem consists of a Poisson density for the number of observed candidates and a Gaussian for the background measurement. Using equation (6) with a simple Monte Carlo integration routine, one obtains a p value of $\sim 1.6\%$. For reference, when there is no uncertainty on ν_0 the p value is $\sim 1.4\%$.

While there are many purely frequentist approaches to the elimination of nuisance parameters, few of these have general applicability. Concentrating on the latter, we discuss the likelihood ratio and confidence interval methods in the next two subsections.

5.1. Likelihood Ratio Method

The likelihood ratio statistic λ is defined by:

$$\lambda = \frac{\sup\limits_{\substack{\mu=0 \\ \nu \geq 0}} \mathcal{L}(\mu, \nu \,|\, n_0, \nu_0)}{\sup\limits_{\substack{\mu \geq 0 \\ \nu \geq 0}} \mathcal{L}(\mu, \nu \,|\, n_0, \nu_0)}, \qquad (7)$$

where, for $\nu_0 \gg \Delta\nu$:

$$\mathcal{L}(\mu, \nu \,|\, n_0, \nu_0) \propto \frac{(\mu+\nu)^{n_0}}{n_0!} \, e^{-\mu-\nu} \, e^{-\frac{1}{2}\left(\frac{\nu-\nu_0}{\Delta\nu}\right)^2}.$$

Simple calculus leads to:

$$\begin{aligned} -2\ln\lambda &= 2\left(n_0 \ln\frac{n_0}{\hat{\nu}} + \hat{\nu} - n_0\right) + \left(\frac{\hat{\nu}-\nu_0}{\Delta\nu}\right)^2 \quad \text{if } n_0 > \nu_0, \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{if } n_0 \leq \nu_0, \end{aligned}$$

with: $\hat{\nu} = \frac{\nu_0 - \Delta\nu^2}{2} + \sqrt{\left(\frac{\nu_0 - \Delta\nu^2}{2}\right)^2 + n_0 \, \Delta\nu^2}.$

Since λ depends on n_0 and ν_0, its distribution under H_0 depends on the true expected background ν_t. A natural simplification is to examine the limit $\nu_t \to \infty$. Application of theorems describing the asymptotic behavior of $-2\ln\lambda$ must take into account that for $n_0 < \nu_0$ the analytical maximum of the likelihood lies outside the physical region $\mu \geq 0$. The correct asymptotic result is that, under H_0, half a unit of probability is carried by the singleton $\{-2\ln\lambda = 0\}$, and the other half is distributed as a chisquared with one degree of freedom over $0 < -2\ln\lambda < +\infty$.

For our example the expected background is only 5.7 particles however, so one may wonder how close this is to the asymptotic limit. Here is an algorithm to check this. Choose a true number of background events ν_t and repeat the following three steps a large number of times:

1. Generate a Gaussian variate ν_0 with mean ν_t and width $\Delta\nu$;

2. Generate a Poisson variate n_0 with mean ν_t;

3. Calculate λ from the generated ν_0 and n_0.

The p value is then equal to the fraction of pseudo-experiments that yield a likelihood ratio λ smaller than the λ_0 obtained from the observed data.

Note that this algorithm does not "smear" the true value of any parameter, in contrast with equation (6). The price for this is that the result depends on the choice of ν_t. For ν_t varying from 0.5 to 50, the p value ranges from ~ 0.48 to $\sim 1.2\%$. A general prescription for dealing with a p value dependence on nuisance parameters is to use the so-called supremum p value:

$$p_{\text{sup}} = \sup_{\nu} \, \mathbb{P}(-2\ln\lambda \geq -2\ln\lambda_0 \,|\, \mu, \nu)\big|_{\mu=0}$$

From a frequentist point of view, the supremum p value is *valid*, in the sense that:

$$\mathbb{P}(p_{\text{sup}} \leq \alpha) \ \leq \ \alpha, \quad \text{for each } \alpha \in [0,1], \qquad (8)$$

regardless of the true value of the nuisance parameter. Although it is often difficult to calculate a supremum, in this case it turns out to equal the asymptotic limit to a good approximation. In our example $-2\ln\lambda_0 = 5.02$ and corresponds to $p_{\text{sup}} \approx p_\infty = 1.25\%$.

As the attentive reader will have noticed, the p value is smaller for $\Delta\nu = 0.47$ than for $\Delta\nu = 0$. This is a consequence of the discreteness of Poisson statistics; it does not violate inequality (8) because p_{sup} actually overcovers a little when $\Delta\nu = 0$. To avoid the bias resulting from this overcoverage, the use of mid-p values is sometimes advocated for the purpose of comparing or combining p values [7].

5.2. Confidence Interval Method

The supremum p value introduced in the previous section can be defined for any test statistic, although it will not always give useful results. If for example in our new particle search we take the total number n_0 of observed candidates as test statistic, the p value will be 100% since the background ν is unbounded from above. A more satisfactory method proceeds as follows [8, 9]. First, construct a $1-\beta$ confidence interval C_β for the nuisance parameter ν, then maximize the p value over that interval, and finally correct the result for the fact that $\beta \neq 0$:

$$p_\beta \ = \ \sup_{\nu \in C_\beta} \ \mathbb{P}(N \geq n_0 \,|\, \mu, \nu)|_{\mu=0} \ + \ \beta.$$

It can be shown that this is also a *valid* p value.

For the sake of illustration with our example, we consider three choices of β and construct the corresponding $1-\beta$ confidence intervals for ν_t:

$$
\begin{aligned}
1-\beta = \ 99.5\% : &\quad C_{0.005} \ = [4.38\,,\,7.02]\\
1-\beta = \ 99.9\% : &\quad C_{0.001} \ = [4.15\,,\,7.25]\\
1-\beta = \ 99.99\% : &\quad C_{0.0001} = [3.87\,,\,7.53]
\end{aligned}
$$

To calculate the p value, a good choice of statistic is the maximum likelihood estimator of the signal, i.e. $\hat{s} \equiv n_0 - \nu_0$. Under H_0, the survivor function of $\hat{s}$ is given by:

$$\mathbb{P}(S \geq \hat{s}) \ = \ \sum_{k=0}^{\infty} \frac{1 + \text{erf}\left(\frac{k - \nu_t - \hat{s}}{\sqrt{2}\,\Delta\nu}\right)}{1 + \text{erf}\left(\frac{\nu_t}{\sqrt{2}\,\Delta\nu}\right)} \, \frac{\nu_t^k}{k!} \, e^{-\nu_t}$$

We then find:

$$
\begin{aligned}
1-\beta = \ 99.5\% : &\quad p_\beta \ = \ 1.6\% \ +0.5\% \ = \ 2.1\%\\
1-\beta = \ 99.9\% : &\quad p_\beta \ = \ 1.7\% \ +0.1\% \ = \ 1.8\%\\
1-\beta = \ 99.99\% : &\quad p_\beta \ = \ 1.88\% +0.01\% \ = \ 1.89\%
\end{aligned}
$$

An important point about the confidence interval method is that, in order to satisfy the Anticipation Criterion, the value of β and the confidence set C_β must be specified before looking at the data. Since p_β is never smaller than β, the latter should be small. In particular, if p_β is used in a level-α test, then β must be smaller than α for the test to be useful.

6. SUMMARY

From the practical point of view of someone analyzing data, the most critical property of frequentist ensembles is their "anticipatoriness." This requires that all the structural elements of an analysis (i.e. test sizes, interval procedures, bin boundaries, stopping rules, etc.) be in place before looking at the data. The only exception to this requirement occurs in situations where conditioning is both possible and appropriate. Even in that case, the conditioning partition itself must be specified beforehand.

References

[1] M. J. Bayarri and J. O. Berger, "The interplay of Bayesian and frequentist analysis," http://www.stat.duke.edu/~berger/papers/interplay.html, April 2003.

[2] L. D. Brown, "Comment on "Conditional confidence statements and confidence estimators"," *J. Amer. Statist. Assoc.* **72**, 810–812 (1977).

[3] J. Berger, B. Boukai, and Y. Wang, "Unified frequentist and Bayesian testing of a precise hypothesis," *Statist. Sci.* **12**, 133–160 (1997).

[4] D. R. Cox, "Some problems connected with statistical inference," *Ann. Math. Statist.* **29**, 357–372 (1958).

[5] J. Kiefer, "Conditional confidence statements and confidence estimators," *J. Amer. Statist. Assoc.* **72**, 789–808 (1977).

[6] J. O. Berger and D. A. Berry, "The relevance of stopping rules in statistical inference," in *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.), vol. 1, pp. 29–72, Springer-Verlag, 1988.

[7] G. Berry and P. Armitage, "Mid-P confidence intervals: a brief review," *Statistician* **44**, 417–423 (1995).

[8] R. L. Berger and D. D. Boos, "P values maximized over a confidence set for the nuisance parameter," *J. Amer. Statist. Assoc.* **89**, 1012–1016 (1994).

[9] M. J. Silvapulle, "A test in the presence of nuisance parameters," *J. Amer. Statist. Assoc.* **91**, 1690–1693, (1996); Correction: *ibid.* **92**, 801 (1997).