Repeat Consistency of a Result

Frank Porter
Caltech
April 18, 2002

Introduction

We address the following issue: Suppose that we have taken a dataset consisting of a set of events. We do an analysis on this dataset, and obtain a result of interest. We then take the same dataset, and repeat the analysis, possibly with differences, to again obtain a result on the same matter of interest. How do we determine whether our two results are "consistent"?

Let us first attempt to ensure that we understand the question. Consider the limit in which we repeat the identical analysis. In this case, the result should be identical. If it isn't, there is an "inconsistency". Clearly, "inconsistency" here means "mistake" – the assertion that the analyses are identical must be incorrect. Now consider the limit in which the first analysis includes a random selection of one-half of the events, and the second analysis includes a selection of precisely the half not used in the first analysis. The two analyses are otherwise identical. In this case the difference between the two results should be of purely statistical origin. "Consistency" in this case is somewhat less trivial to determine, because the results can differ due to statistical fluctuations. However, a conclusion of "inconsistency" is still an assertion that a "mistake" was made, up to whatever small probability is permitted for a large statistical fluctuation.

The key point is that we are trying to evaluate whether any observed difference in the two analyses is an indication that there is something wrong with one or both analyses, or whether the difference is consistent with being due to statistical fluctuations.

How to Evaluate Consistency

Let θ stand for a physical parameter of interest. Let $\hat{\theta}_1$ be the estimated value of θ in the first analysis, and $\hat{\theta}_2$ be the estimated value in the second analysis. The most obvious statistic for evaluating consistency between the two analyses is $\Delta\theta \equiv \hat{\theta}_2 - \hat{\theta}_1$. We need the pdf, $p(\Delta\theta)$, for this difference in order to make a test for consistency. Given this, we simply compute:

$$\alpha = 1 - \int_{-\Delta\theta}^{\Delta\theta} p(x)dx. \tag{1}$$

α is the probability that we would observe a difference greater than the observed difference, due to statistical fluctuations. If α is bigger than some number, perhaps 0.05, then we conclude all is well, if not, we worry that something is wrong.

Actually, we have made a simplifying assumption here: If $p(x)$ depends on other parameters, including θ itself, then the analysis is complicated by the need to know these parameters. In the case that θ is a location parameter for $\hat{\theta}_1$ and $\hat{\theta}_2$, then there is no dependence of $p(x)$ on θ. As this is often at least approximately the case, our simplifying assumption is of interest. We'll stick to this assumption here, while recognizing that in practice additional complications may need to be treated.

The question we wish to address for now is: What is $p(x)$? In the first limiting case of the introduction, $p(x) = \delta(x)$. In the second limiting case, denote $q_i(\hat{\theta}_i;\theta),\ i = 1,2$ as the pdf's for the results of the two analyses. By statistical independence, we have $p(\Delta\theta) = \int_{-\infty}^{\infty} q_1(x)q_2(x+\Delta\theta)\,dx$. More generally, there may be correlation between the samplings, with a joint pdf $q(\hat{\theta}_1, \hat{\theta}_2;\theta)$. In this case,

$$p(\Delta\theta) = \int_{-\infty}^{\infty} q(x, x+\Delta\theta)\,dx. \tag{2}$$

Example: Normal Distribution

Suppose that our samplings are from a bivariate normal distribution with common mean:

$$q(x_1, x_2; \theta) = A \exp\left\{-\frac{1}{2}\left[(x_1-\theta)^2 W_{11} + (x_2-\theta)^2 W_{22} + 2(x_1-\theta)(x_2-\theta)W_{12}\right]\right\}, \tag{3}$$

where W is the inverse of the covariance matrix. If the covariance matrix is known, then the consistency may be evaluated by a simple χ^2 goodness-of-fit test:

$$\chi^2(1 \text{ dof}) = (x_1-\hat\theta)^2 W_{11} + (x_2-\hat\theta)^2 W_{22} + 2(x_1-\hat\theta)(x_2-\hat\theta)W_{12}, \tag{4}$$

where $\hat\theta$ is the value of θ which minimizes the χ^2.

On the other hand, the covariance matrix may not be fully known, or may be a lot of trouble to estimate. In particular, we may not know the correlation coefficient (ρ). We can still make a test for consistency, though at a cost of power (probability of correctly deciding that the results are not the same) and/or at the cost of significance level (probability of incorrectly deciding that the results are different) due to the uncertainty in ρ, the correlation coefficient. As shown in the Appendix, the difference $x_2 - x_1$ is distributed according to a normal distribution with mean zero and variance $\sigma_{x_1}^2 + \sigma_{x_2}^2 - 2\rho\sigma_{x_1}\sigma_{x_2}$. Since ρ is bounded by $\rho \in (-1, 1)$, the variance of $x_2 - x_1$ is in the region $(\sigma_{x_1} - \sigma_{x_2})^2$ to $(\sigma_{x_1} + \sigma_{x_2})^2$.

If a test of the observed distance using the smaller variance gives consistency, then we may be confident that the results are consistent. If a test of the observed distance using the larger variance gives inconsistency, then we conclude that the results are inconsistent. This is as much as we can do in the absence of knowledge concerning ρ. However, we might at least know that the correlation is not negative, and thence be able to tighten the test. In this case, the maximum variance of the difference is $\sigma_{x_1}^2 + \sigma_{x_2}^2$.

Example: Branching Fraction

Assume that we have a dataset corresponding to N B decays. We wish to determine the branching fraction θ to a particular final state. We do two analyses, yielding samples of N_1 and N_2 events, with efficiencies ϵ_1 and ϵ_2, respectively. For simplicity, we assume that the uncertainties in N and ϵ_i are negligible, and that there is no background contribution. The mean number of events expected in each of these analyses is $\langle N_i \rangle = \theta\epsilon_i N$. The two estimates of the branching fraction are given by:

$$\hat\theta_i = \frac{N_i}{N\epsilon_i}, \quad i = 1, 2. \tag{5}$$

If we treat N as fixed, we may model our samplings according to the multinomial distribution. Let N_{12} be the number of events which are common to both samples N_1 and N_2, with corresponding efficiency ϵ_{12}. Thus, $\langle N_{12} \rangle = \theta\epsilon_{12} N$. Note that $\max(0, \epsilon_1 + \epsilon_2 - 1) \le \epsilon_{12} \le \min(\epsilon_1, \epsilon_2)$, and that statistically independent sampling corresponds to $\epsilon_{12} = \epsilon_1\epsilon_2$. Define $\bar N_i \equiv N_i - N_{12}$, *i. e.*, the number of events that are selected in analysis "i", but not in the other analysis. Also let $\bar\epsilon_i$ be the efficiency for an event to be selected in analysis i but not in the other analysis. We thus divide our entire sample of N events into four disjoint sets, with a partitioning given by:

$$\begin{aligned} &P(\bar N_1, \bar N_2, N_{12}; \theta) = \\ &\frac{N}{\bar N_1! \bar N_2! N_{12}! (N - \bar N_1 - \bar N_2 - N_{12})!} (\bar\epsilon_1\theta)^{\bar N_1} (\bar\epsilon_2\theta)^{\bar N_2} (\bar\epsilon_{12}\theta)^{\bar N_{12}} [1 - (\epsilon_1 + \epsilon_2 + \epsilon_{12})\theta]^{(N - \bar N_1 - \bar N_2 - N_{12})}. \end{aligned} \tag{6}$$

We typically may use the Poisson limit:

$$P(\bar N_1, \bar N_2, N_{12}; \theta) = \frac{\mu_1^{\bar N_1}}{\bar N_1!} \frac{\mu_2^{\bar N_2}}{\bar N_2!} \frac{\mu_{12}^{N_{12}}}{N_{12}!} e^{-\mu_1 - \mu_2 - \mu_{12}}, \tag{7}$$

where

$$\mu_i \equiv \bar\epsilon_i \theta \langle N \rangle, \qquad \mu_{12} \equiv \epsilon_{12}\theta\langle N \rangle. \tag{8}$$

Our difference test statistic is

$$\Delta\theta = \frac{1}{N}\left(\frac{N_2}{\epsilon_2} - \frac{N_1}{\epsilon_1}\right) = \frac{1}{N}\left[\frac{\bar{N}_2}{\epsilon_2} - \frac{\bar{N}_1}{\epsilon_1} + N_{12}\left(\frac{1}{\epsilon_1} - \frac{1}{\epsilon_2}\right)\right]. \tag{9}$$

Note that $\epsilon_i = \bar{\epsilon}_i + \epsilon_{12}$. The distribution of this test statistic may be evaluated according to the above probability distribution with Monte Carlo methods. The observed difference may then be compared with the predicted distribution in order to evaluate consistency. As in the Gaussian case, it is possible that the "correlation" parameter, ϵ_{12}, may not be known, and a similar treatment to that for the Gaussian example will be necessary. In many cases, it will probably be reasonable to assume that $\epsilon_1\epsilon_2 < \epsilon_{12} < \min(\epsilon_1, \epsilon_2)$, *i. e.*, that the selection is not anti-correlated.

It may be remarked that the test here checks for consistency between the two results. It doesn't check for other possible problems, such as whether we have consistency with the expected overlap. Other tests could be devised to address such questions.

It is possible that, in the case where we don't know the correlation parameter, we can estimate it from other data available to us. That is, we may have sets of signal-like events and background-like events from our two analyses which can be used to estimate the relative values of $\bar{\epsilon}_1$, $\bar{\epsilon}_2$, and ϵ_{12}. These can then be used to evaluate whether the observed signal numbers show consistent behavior.

Example: Two Analyses on the Same Events

A case that causes some confusion is when the event selection is identical, but two analyses are performed on the selected sample. Because different information may be used in the two analyses, some variation in the results may be expected. As the measured information is in the form of random variables, this is still a problem amenable to statistical analysis.[1] The question being asked is still whether the observed difference is consistent with statistical fluctuations, versus the possibility that there is an inconsistency (mistake).

A concrete illustration[2] is the following situation: Suppose that we have a set of events consisting of mass measurements, $\{m_1, \ldots, m_n\}$, of some resonance. Let the resonance mass be denoted θ. Assume for simplicity that the measurements are all made with Gaussian resolution functions, with possibly different widths, but all unbiased, and assume that the natural resonance width is negligible. We may form an estimate of the resonance mass by taking the sample mean of all the measurements:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} m_i. \tag{10}$$

This is an unbiased estimator for θ, since $\langle\hat{\theta}_1\rangle = \theta$.

Now suppose that we actually have, for each measurement, the resolution, σ_i, with which it is made. This additional information does not invalidate our estimator $\hat{\theta}_1$, but we can incorporate this information into another estimator:

$$\hat{\theta}_2 = \sum_{i=1}^{n} \frac{m_i}{\sigma_i^2} \Big/ \sum_{i=1}^{n} \frac{1}{\sigma_i^2}. \tag{11}$$

Again, we have an unbiased estimator for θ, since $\langle\hat{\theta}_2\rangle = \theta$.

Both $\hat{\theta}_1$ and $\hat{\theta}_2$ are normally distributed, with moment matrix:

$$M = \begin{pmatrix} \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 & 1/\sum_{i=1}^{n}\frac{1}{\sigma_i^2} \\ 1/\sum_{i=1}^{n}\frac{1}{\sigma_i^2} & 1/\sum_{i=1}^{n}\frac{1}{\sigma_i^2} \end{pmatrix}. \tag{12}$$

[1] Art Snyder has discussed the problem in the context of two maximum likelihood analyses: http://-www.slac.stanford.edu/∼snyder/shifts.ps.

[2] Suggested by Art Snyder.

Note that the form of this matrix is indicative of the fact that all of the information in the first analysis is used in the second analysis. According to our above analysis of the bivariate normal, the difference between the two estimators is distributed according to the normal distribution with standard deviation:

$$\sigma_{\Delta\theta} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 - \frac{1}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}}. \tag{13}$$

A simple χ^2 test can thus be used for checking consistency between the two results.

Dealing with Systematic Uncertainties

Typically our results have "systematic" uncertainties in addition to the statistical uncertainties we have so far been dealing with. In some cases, the systematic effect will be identical for both results. For example, both results might be based on the same estimate of the integrated luminosity. In this situation, there is no additional uncertainty in $\Delta\theta$ from this source.

On the other hand, it is possible that a systematic effect may be different in the two analyses. For example, the efficiency estimates in the two analyses may be made differently. This could lead to a "systematic" uncertainty in $\Delta\theta$ affecting our criteria for consistency. If possible, any common systematic should be separated out, leaving only the "independent" systematics, call them s_{x_1} and s_{x_2}. Then it is reasonable to assign a systematic uncertainty of $s_{\Delta\theta} = \sqrt{s_{x_1}^2 + s_{x_2}^2}$ to the difference, similarly to the result for uncorrelated statistical uncertainties. If it is too difficult to separate out the common systematics, then the best one can do is embark on a treatment similar to the discussion for the unknown correlation in the statistical errors.

Finally, a comment on what to do with the observed difference in the results of the two analyses. In general, the existence of two such results is a "fortuitous" circumstance – doing the two analyses is not a part of the experiment design. In particular, there is no plan that the purpose of doing the two analyses is in order to evaluate a "systematic uncertainty". The systematic uncertainties should be evaluated as appropriate in each analysis. The existence of more than one analysis may be used as a "check" for mistakes, but no new systematic uncertainty should be assigned to cover the difference between the two analyses.

APPENDIX: Bivariate Normal Distribution with Common Mean

Consider the distribution:

$$p(x,y;\theta) = A\exp\left\{-\frac{1}{2}\left[(x-\theta, y-\theta)W\begin{pmatrix} x-\theta \\ y-\theta \end{pmatrix}\right]\right\}, \tag{14}$$

where W is the inverse of the moment matrix, M, and $A = 1/2\pi\sqrt{\det(M)}$. The moment matrix is given by:

$$M = \begin{pmatrix} \langle(x-\theta)^2\rangle & \langle(x-\theta)(y-\theta)\rangle \\ \langle(x-\theta)(y-\theta)\rangle & \langle(y-\theta)^2\rangle \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \tag{15}$$

The inverse is:

$$W = M^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}\begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}. \tag{16}$$

We wish to change variables to $x = x, v = y - x$, with inverse mapping $x = x, y = v + x$:

$$q(x,v;\theta) = A\exp\left\{-\frac{1}{2}\left[(x-\theta, v+x-\theta)W\begin{pmatrix} x-\theta \\ v+x-\theta \end{pmatrix}\right]\right\}, \tag{17}$$

Further, we wish to integrate over x, in order to determine the distribution of the difference, v, independent of the unknown θ. Without the factor of $-1/2$, the expression in the exponential is, letting $w \equiv x - \theta$:

$$\chi^2 = w^2W_{11} + (v+w)^2W_{22} + 2w(v+w)W_{12} = w^2(W_{11}+W_{22}+2W_{12}) + 2wv(W_{22}+W_{12}) + v^2W_{22}. \tag{18}$$

We complete the square for w:

$$\chi^2 = (aw + b)^2 + c, \tag{19}$$

where

$$a = \sqrt{W_{11} + W_{22} + 2W_{12}}, \quad b = \frac{v(W_{22} + W_{12})}{\sqrt{W_{11} + W_{22} + 2W_{12}}}, \quad \text{and} \quad c = v^2 \frac{W_{11}W_{22} - W_{12}^2}{W_{11} + W_{22} + 2W_{12}}. \tag{20}$$

The coefficient of v^2 in c can be rewritten as $1/(\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y)$.

Thus, the difference between y and x is distributed as $N\left(0, \sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y}\right)$.