Conditioning

Luc Demortier

CDF Statistics Committee
Oxford meeting, March 17-19, 2003

- Basic definitions
- Applications of conditioning
- The Borel paradox
- Berger and Wolpert's example
- Efron's solution
- Zech conditioning
- Roe-Woodroofe conditioning
- Conditioning in fits
- Conditional frequentist testing: simple vs. simple
- Conditional frequentist testing: simple vs. composite
- Final comments

Basic definitions

Introductory probability courses define conditional probabilities of events as ratios:

$$P(A \,|\, B) \;=\; \frac{P(A \,\&\, B)}{P(B)}, \qquad (1)$$

provided $P(B) \neq 0$. The division by $P(B)$ ensures that $P(\cdot \,|\, B)$ is also a probability measure, which assigns zero probability when event B is not realized: $P(B^c \,|\, B) = 0$.

In many applications one is conditioning on a continuous random variable X, for which $P(X = x) = 0$. One is then forced to hope that something like

$$P(A \,|\, X = x) = \lim_{\delta \to 0} P(A \,|\, x \leq X \leq x + \delta) \qquad (2)$$

exists and is a probability measure for each fixed x. A more rigorous approach requires notions from measure theory.

From the perspective of ensembles, conditioning is equivalent with partitioning. For example, suppose you have a sample space Ω and a statistic T that can take the values $0, 1, 2, \ldots$. Then, conditioning on $T = t$ selects a subspace in Ω, and subspaces with different t values are disjoint.

One can also condition on an inequality such as $T \leq t$; this is not fundamentally different from the case $T = t$.

Applications of conditioning

A common application is to derive probabilities and expectations from conditional distributions constructed by appeals to symmetry or modelling assumptions.

Example: Suppose a deck of 26 red and 26 black cards is well shuffled, that is, all 52! permutations of the cards are equally likely.

- Let A be the event $\{$top and bottom cards are red$\}$.
- Let T be the statistic that gives the color of the top card.

The distribution of T is then:

$$\mathbb{P}(T = \mathrm{red}) \;=\; 1/2 \quad \text{and} \quad \mathbb{P}(T = \mathrm{black}) \;=\; 1/2. \tag{3}$$

By symmetry, all permutations of the 51 remaining cards are equally likely. Therefore:

$$\mathbb{P}(A \,|\, T = \mathrm{red}) \;=\; 25/51 \tag{4}$$

$$\mathbb{P}(A \,|\, T = \mathrm{black}) \;=\; 0/51, \tag{5}$$

so that:

$$\mathbb{P}(A) \;=\; \mathbb{P}(T = \mathrm{red})25/51 + \mathbb{P}(T = \mathrm{black})0/51 \;=\; 25/102. \tag{6}$$

The Borel paradox

"Let x be a point chosen randomly from the uniform distribution on the surface of the Earth. Intuitively, if the point lies somewhere along the equator, then the longitude should be uniformly distributed over the range $[-180^\circ, +180^\circ]$. But what is so special about the equator? Given that the point lies on any particular great circle, its position should be uniformly distributed around that circle. In particular, for a great circle through the poles (i.e. conditional on the longitude), there should be conditional probability 1/4 that the point lies north of latitude 45°N. Average out over the longitude to deduce that the point has probability 1/4 of lying in the spherical cap extending from the North pole down to the 45° parallel of latitude. Unfortunately, that cap does not cover 1/4 of the Earth's surface area, as would be required for a point uniformly distributed over the whole surface."

As Kolmogorov put it: "The concept of a conditional probability with regard to an isolated hypothesis whose probability equals 0 is inadmissible. For we can obtain a probability distribution for [the latitude] on the meridian circle only if we regard this circle as an element of the decomposition of the entire spherical surface into meridian circles with the given poles."

References

[1] David Pollard, "A user's guide to measure theoretic probability", Cambridge Univ. Press, 2002, 251 pp.

Berger and Wolpert's example

Two observations, X_1 and X_2, are to be taken, where

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2. \end{cases}$$

Consider the confidence set for the unknown θ

$$C(X_1, X_2) = \begin{cases} \text{the point } \{\frac{1}{2}(X_1 + X_2)\} & \text{if } X_1 \neq X_2 \\ \text{the point } \{X_1 - 1\} & \text{if } X_1 = X_2. \end{cases}$$

The unconditional frequentist coverage for this confidence set can be determined from a table:

X_1	X_2	Prob.	Cover?
$\theta - 1$	$\theta - 1$	0.25	No
$\theta - 1$	$\theta + 1$	0.25	Yes
$\theta + 1$	$\theta - 1$	0.25	Yes
$\theta + 1$	$\theta + 1$	0.25	Yes

So we have:

$$P_\theta(\theta \in C(X_1, X_2)) = 0.75. \quad (7)$$

However, this is not a sensible way to report the result of our measurement. For if we observe $X_1 \neq X_2$, then we know the value of θ with 100% confidence.

On the other hand, if we observe $X_1 = X_2$, then our calculation of θ has a 50% chance of being right.

It would be much smarter to introduce a conditioning statistic such as $S \equiv |X_1 - X_2|$, and to write:

$$P_\theta(\theta \in C(X_1, X_2) \,|\, S = 2) \quad = \quad 1; \tag{8}$$

$$P_\theta(\theta \in C(X_1, X_2) \,|\, S = 0) \quad = \quad 0.50. \tag{9}$$

The statistic S can be thought of as measuring the "strength of evidence" in the data: $S = 2$ indicates data with high strength of evidence, i.e. maximally informative, whereas $S = 0$ indicates data with minimal evidential content.

Conditional frequentist measures are fully frequentist, in the sense that they have the same unconditional properties as their unconditional counterparts. In the example above, one will report 100% confidence half the time and 50% confidence half the time, resulting in an "average" confidence of 75%, in agreement with the unconditional report. The conditional report however, is much more accurate.

References

[1] James O. Berger, "Could Fisher, Jeffreys, and Neyman have agreed on testing?", ISDS Discussion paper 02-01 (2002).

Efron's solution to the p-value problem in the CDF top evidence paper

CDF's first evidence for top quark production at the Tevatron was based on an observation of 12 events over an expected background of 5.7 ± 0.465. If we ignore correlations between the 12 events, what is the significance of this observation?

To answer this question we need to specify the model in greater detail. A judicious specification is as follows:

1. The observation of $n_1 = 12$ events is a Poisson variate with mean μ_1;

2. An independent estimate of the background is based on the observation of $n_2 = 150$ events, a Poisson variate with mean μ_2;

3. The null hypothesis is that $\mu_1 = c \cdot \mu_2$, with $c = 0.03793$ (the uncertainty on c is negligible).

The question is now, what is the probability of observing $n_1 \geq 12$ under the null hypothesis?

To calculate this probability directly we need both μ_1 and μ_2, but we only know their ratio c. The solution is to condition on the statistic $n_1 + n_2$, which arguably does not contain much information about the difference between μ_1 and μ_2.

Using abbreviated notation, one has:

$$
\begin{aligned}
&\mathrm{Prob}(n_1 \,|\, n_1{+}n_2) && (10)\\
&\quad = \frac{\mathrm{Prob}(n_1 \;\&\; n_1{+}n_2)}{\mathrm{Prob}(n_1{+}n_2)} && (11)\\
&\quad = \frac{\mathrm{Prob}(n_1 \;\&\; n_2)}{\mathrm{Prob}(n_1{+}n_2)} && (12)\\
&\quad = \frac{\mathrm{Prob}(n_1)\;\mathrm{Prob}(n_2)}{\mathrm{Prob}(n_1{+}n_2)} && (13)\\
&\quad = \frac{\frac{\mu_1^{n_1}}{n_1!}\,e^{-\mu_1}\;\frac{\mu_2^{n_2}}{n_2!}\,e^{-\mu_2}}{\frac{(\mu_1+\mu_2)^{n_1+n_2}}{(n_1+n_2)!}\,e^{-\mu_1-\mu_2}} && (14)\\
&\quad = \frac{(n_1+n_2)!}{n_1!\;n_2!}\;\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{n_1}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{n_2} && (15)\\
&\quad = \mathrm{Binomial}\left(n_1\;;\;n_1{+}n_2\,,\,\frac{c}{c+1}\right). && (16)
\end{aligned}
$$

For our example the numerical values are:

$$
\begin{aligned}
n_1 + n_2 &= 162 && (17)\\
\frac{c}{c+1} &= 0.03655 && (18)
\end{aligned}
$$

and we are now ready to compute the requested significance:

$$\begin{aligned}
&\mathrm{Prob}(n_1 \geq 12 \,|\, n_1{+}n_2{=}162 \;\&\; H_\circ) && (19)\\
&\quad = \;\; \mathrm{Prob}\,(\mathrm{Binomial}(n_1\;;\;162\,,\,0.03655) \geq 12) && (20)\\
&\quad \cong \;\; 1.6\% && (21)
\end{aligned}$$

This is actually the same result as the one reported by CDF using a Bayesian treatment of the nuisance parameter μ_1, assuming it to be Gaussian with mean 5.7 and width 0.465.

The result of the above calculation however, has a different interpretation: *it is the probability, under the null-hypothesis that the background scale factor c is 0.03793, of observing 12 or more top candidate events in an ensemble of experiments in which the sum of the number of top candidates and the unrescaled background estimate is 162.*

Note that in the above ensemble μ_1 and μ_2 are unspecified; only their ratio is prescribed. One could for example generate pseudo-experiments with $\mu_2 = 900,000,000$ and $\mu_1 = c \cdot \mu_2 = 34,137,000$. After enormous amounts of "generation time", one would be able to assemble a subensemble of experiments with $n_1 + n_2 = 162$. Within this subensemble, 1.6% of pseudo-experiments have $n_1 \geq 12$.

Zech conditioning

Suppose we are searching for a rare process and observe a small number of events N that can be explained by background. If the expected background b is known with high precision, an α confidence level upper limit u on the true signal can be derived by solving the following equation:

$$1 - \alpha = P(n \leq N\,;\, u + b), \tag{22}$$

where n is a Poisson variate with mean $u + b$:

$$P(n \leq N\,;\, u + b) \equiv \sum_{n=0}^{N} \frac{(u+b)^n}{n!}\, e^{-(u+b)}. \tag{23}$$

This method yields a very unappealing solution in the case that zero events are observed:

$$u(N = 0) = -\ln(1 - \alpha) - b. \tag{24}$$

As the expected background increases, the upper limit decreases and could even become negative!

To solve this problem one needs to incorporate more information from the data into the calculation of the upper limit. This requires *conditioning*. There are two approaches, Bayesian and frequentist.

1. Bayesian approach

In a sense, this approach amounts to "maximal conditioning" on the observed data, but it requires prior probabilities for signal and background. For the problem at hand, Bayes' theorem takes the following form:

$$p(s, b \,|\, N) \;=\; \frac{p(N \,|\, s+b)\,\pi(s)\,\pi(b)}{\int_0^\infty ds \int_0^\infty db\; p(N \,|\, s+b)\,\pi(s)\,\pi(b)} \tag{25}$$

where:

$$p(N \,|\, s+b) \;=\; \frac{(s+b)^N}{N!}\,e^{-(s+b)}, \tag{26}$$

$$\pi(s) \;=\; 1, \tag{27}$$

$$\pi(b) \;=\; \delta(b-\bar{b}). \tag{28}$$

After performing a couple of simple integrals, one finds:

$$p(s \,|\, N) \;=\; \int_0^\infty db\; p(s, b \,|\, N) \;=\; \frac{\frac{(s+\bar{b})^N}{N!}\;e^{-(s+\bar{b})}}{\sum_{i=0}^{N} \frac{\bar{b}^i}{i!}\;e^{-\bar{b}}}, \tag{29}$$

and the α confidence level upper limit u is solution of:

$$\alpha \;=\; \int_0^u ds\; p(s \,|\, N). \tag{30}$$

This can be rewritten as:

$$1 \;-\; \alpha \;=\; \frac{P(n \leq N \,|\, u+\bar{b})}{P(n_b \leq N \,|\, \bar{b})}. \tag{31}$$

In the case of zero events observed, this solution is perfectly reasonable:

$$u(N = 0) \;=\; -\ln(1 - \alpha). \tag{32}$$

2. Conditional frequentist approach
   Zech's idea was to condition on the fact that the actual background must be smaller than the observed number of events. This is very easy to implement:

$$\begin{aligned}
1 \;-\; \alpha \;&=\; P(n \leq N \,|\, n_b \leq N\,;\, u, b) \qquad (33)\\
&=\; \frac{P(n \leq N \;\&\; n_b \leq N\;;\; u, b)}{P(n_b \leq N\;;\; b)} \qquad (34)\\
&=\; \frac{P(n \leq N\;;\; u, b)}{P(n_b \leq N\;;\; b)}. \qquad (35)
\end{aligned}$$

   This is exactly the same equation as that obtained in the Bayesian approach, but now we have a frequentist interpretation!

   Unfortunately, things are not as rosy as they may appear. In real life, we will never be able to check the coverage of this upper limit. This is because it is defined with respect to an ensemble of experiments, each of which has N or less background events, where N is the total number of events observed in "this" experiment.

Given an arbitrary experiment somewhere in the real world, it is impossible to tell whether that experiment belongs to our ensemble, because we don't know if $n_b \leq N$ for that experiment. We only have this kind of information for Monte Carlo experiments.

References

[1] O.Helene, "Upper limit of peak area", Nucl. Instrum. and Meth. **212**, 319 (1983).

[2] G.Zech, "Upper limits in experiments with background or measurement errors", Nucl. Instrum. and Meth. **A277**, 608 (1989).

[3] V.L.Highland, "Comment on 'Upper limits in experiments with background or measurement errors' [Nucl. Instr. and Meth. A 277 (1989) 608-610]", Nucl. Instrum. and Meth. **A398**, 429 (1997).

[4] G. Zech, "Reply to 'Comment on "Upper limits in experiments with background or measurement errors" [Nucl. Instr. and Meth. A 277 (1989) 608-610]' ", Nucl. Instrum. and Meth. **A398**, 431 (1997).

Roe-Woodroofe conditioning

The Feldman-Cousins method for calculating intervals also has the defect that in the absence of any observed event, the upper limit decreases as the expected background increases.

In the words of Roe & Woodroofe, "The background radiation is called an *ancillary variable*, because its distribution does not depend on unknown parameters, and conventional statistical wisdom calls for conditioning on ancillary variables when possible".

In the standard Feldman-Cousins (FC) method, the ordering principle is based on the likelihood ratio:

$$R_b(s, N) \;=\; \frac{L_b(s \,|\, N)}{L_b(\hat{s} \,|\, N)}, \tag{36}$$

where $L_b(s \,|\, N) = P(n = N \,;\, b + s)$, N is the observed event count, and $\hat{s}$ is the maximum likelihood estimator of the expected signal s. Roe and Woodroofe propose replacing $P(n = N \,;\, b + s)$ by the conditional probability of observing exactly N events total given at most N_b background events. The latter is:

$$q(N \,|\, N_b \,;\, b, s) = \begin{cases} \frac{P(n=N \,;\, b+s)}{P(n_b \le N_b \,;\, b)} & \text{if } N \le N_b \\ \sum_{i=0}^{N_b} \frac{P(i \,;\, b) P(N-i \,;\, s)}{P(n_b \le N_b \,;\, b)} & \text{if } N > N_b \end{cases}$$

If we sum $q(N \,|\, N_b\,;\, b, s)$ for $N = 0, \ldots, N_b$, we obtain the conditional probability of at most N_b events total, given at most N_b background events:

$$\sum_{N=0}^{N_b} q(N \,|\, N_b\,;\, b, s) \;=\; \frac{P(n \leq N_b\,;\, b+s)}{P(n_b \leq N_b\,;\, b)}, \tag{37}$$

which is the conditional probability Zech uses to calculate upper limits. Next, let $\tilde{R}_{b,N_b}(s, N)$ denote the likelihood ratio obtained using $q(N \,|\, N_b\,;\, b, s)$, i.e.:

$$\tilde{R}_{b,N_b}(s, N) \;=\; \frac{q(N \,|\, N_b\,;\, b, s)}{\max_{s'} q(N \,|\, N_b\,;\, b, s')} \tag{38}$$

and let $\tilde{c}(s, N_b)$ be the largest value of c for which

$$\sum_{N:\; \tilde{R}_{b,N_b}(s,N) < c} q(N \,|\, N_b\,;\, b, s) \;\leq\; \alpha. \tag{39}$$

Then the Roe-Woodroofe modified confidence interval consists of those s values for which

$$\tilde{R}_{b,N_b}(s, N_b) \;\geq\; \tilde{c}(s, N_b). \tag{40}$$

I suspect that Roe-Woodroofe conditioning is identical to Zech conditioning, i.e. they both use the same ensemble as "reference frame".

References

[1] B.P. Roe and M.B. Woodroofe, "Improved probability method for estimating signal in the presence of background", Phys. Rev. **D60**, 053009 (1999).

[2] R.D. Cousins, "Additional comments on methods for setting confidence limits", FNAL Confidence Limits workshop, 28 March 2000.

Conditioning in fits

We make a likelihood fit of a histogram with Poisson statistics in the bins:

$$\mathcal{L}(\vec{x}) \;=\; \prod_{i=1}^{b} \frac{\mu_i^{n_i}}{n_i!}\, e^{-\mu_i}, \tag{41}$$

where the Poisson means μ_i are functions of a vector of parameters $\vec{x}$.

We can then study the result of the fit in terms of bias, coverage, etc. What is the proper ensemble for such a study? It would seem to be the ensemble obtained by letting b Poisson processes with means μ_i run freely over all possible outcomes n_i.

However, the all-zero outcome, "$n_i = 0$ for all i", cannot be fitted, since the corresponding likelihood function is constant, leading to infinite likelihood intervals. The all-zero outcome must therefore be removed from the ensemble, which is equivalent to conditioning on $\sum_{i=0}^{b} n_i > 0$.

Further conditioning could be useful if it allowed one to define subensembles of experiments with good properties in terms of bias, coverage, etc.

Conditional frequentist testing: Simple hypothesis versus simple hypothesis

Suppose we make an observation x in some sample space $\mathcal{X}$ and wish to test a null hypothesis

$$H_\circ : \quad x \sim f_\circ(x) \tag{42}$$

versus an alternative hypothesis

$$H_1 : \quad x \sim f_1(x). \tag{43}$$

To construct a conditional frequentist test we first calculate the p-values:

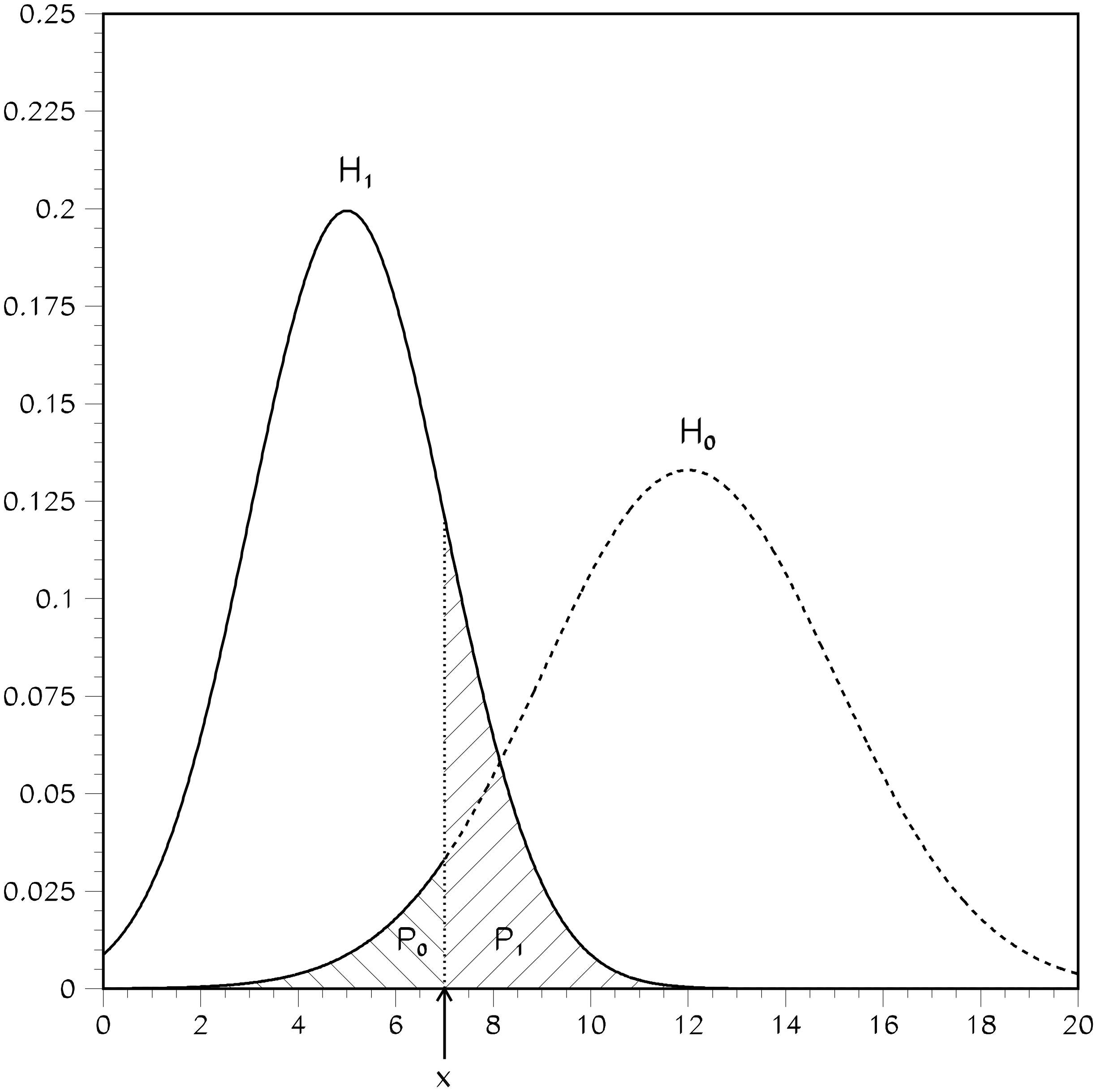
$$p_\circ \;=\; \mathcal{P}_\circ(X \leq x) \;=\; F_\circ(x) \tag{44}$$

$$p_1 \;=\; \mathcal{P}_1(X \geq x) \;=\; 1 - F_1(x) \tag{45}$$

and define the statistic:

$$S(x) \;=\; \max\{p_\circ\,,\, p_1\} \tag{46}$$

Conditioning on $S(x)$ means that data with the same p-value are accorded the same "strength of evidence".


0.25
0.225
0.2
0.175
0.15
0.125
0.1
0.075
0.05
0.025
0
0
2
4
6
8
10
12
14
16
18
20
H1
H0
P0
P1
x

The corresponding conditional test is

1. to reject $H_\circ$ when $p_\circ \leq p_1$, and to accept $H_\circ$ otherwise;

2. to compute the Type I and Type II conditional error probabilities (CEP's).

If we define a critical value $c \in \mathcal{X}$ by

$$F_\circ(c) \;=\; 1 - F_1(c)\,, \tag{47}$$

then part 1. of the test can be restated as:

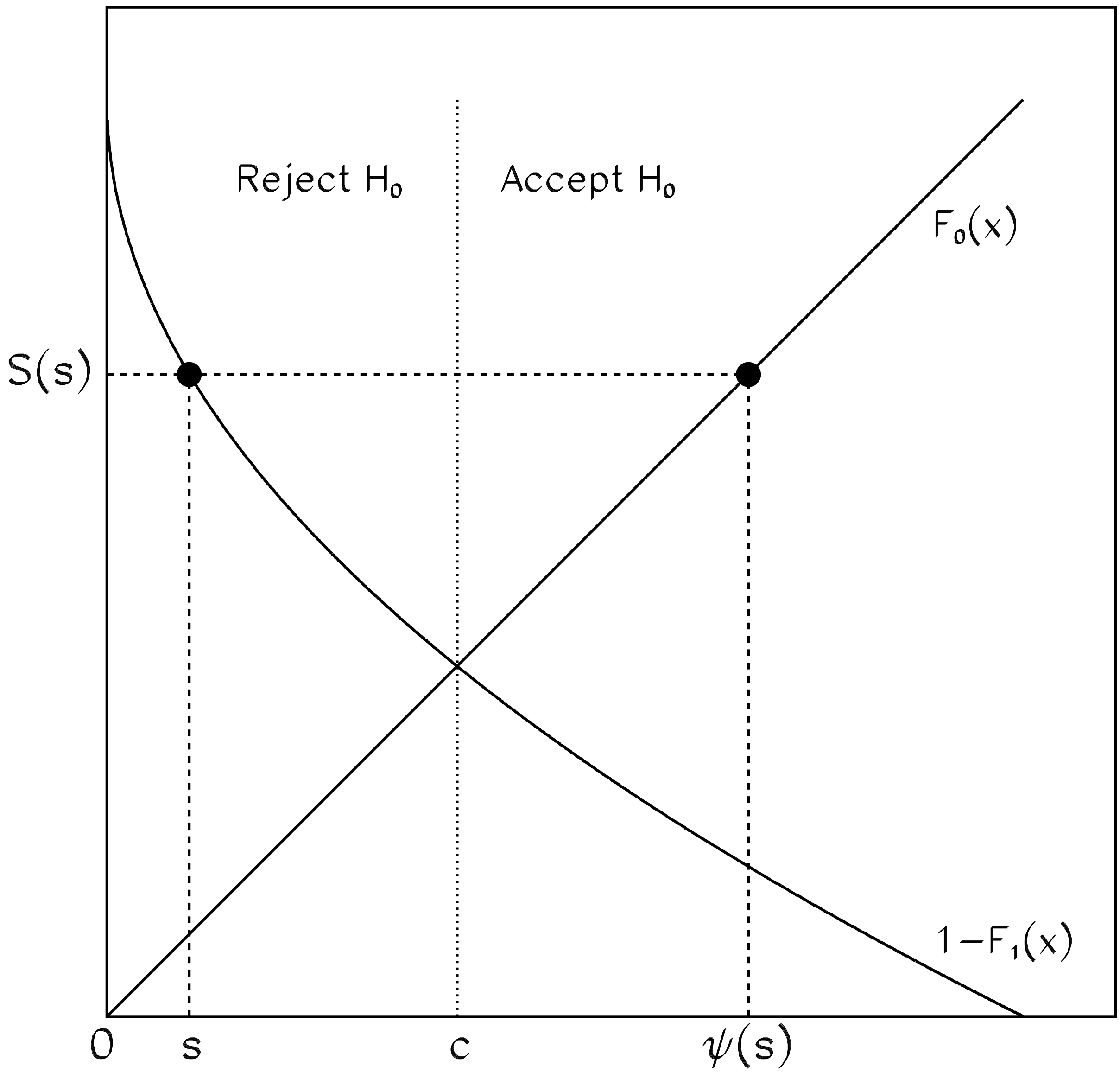
1. to reject $H_\circ$ when $x \leq c$, and to accept $H_\circ$ otherwise.

To calculate the CEP's, define, for $0 \leq s \leq c$:

$$\begin{aligned} \mathcal{X}_s &= \{x \in \mathcal{X} :\; S(x) = 1 - F_1(s)\} && (48) \\ &= \{x \in \mathcal{X} :\; x = s \;\text{ or }\; x = \psi(s)\}, && (49) \end{aligned}$$

where $\psi(s) = F_\circ^{-1}[1 - F_1(s)]$.

Then $\{\mathcal{X}_s :\; s \in [0, c]\}$ is a partition of the sample space $\mathcal{X}$.



Reject H_0
Accept H_0
$F_0(x)$
$1-F_1(x)$
$S(s)$
0
s
c
$\psi(s)$

Simple calculus shows that

$$\frac{d}{ds}\psi(s) \;=\; \frac{d}{ds}F_\circ^{-1}[1 - F_1(s)] \;=\; \frac{-f_1(s)}{f_\circ[\psi(s)]}. \tag{50}$$

For the Type I CEP we find:

$$\begin{aligned}
\alpha(s) \;&=\; \mathcal{P}_\circ(\text{rejecting } H_\circ \,|\, \mathcal{X}_s) && (51)\\
&=\; \frac{\mathcal{P}_\circ(X = s)}{\mathcal{P}_\circ(X = s) \;+\; \mathcal{P}_1(X = \psi(s))} && (52)\\
&=\; \frac{f_\circ(s)}{f_\circ(s) + f_\circ(\psi(s))|\frac{d\psi}{ds}|} && (53)\\
&=\; \frac{f_\circ(s)}{f_\circ(s) + f_1(s)} \quad [\text{using } (50)] && (54)\\
&=\; \frac{B(s)}{1 + B(s)}, && (55)
\end{aligned}$$

where $B(s) = f_\circ(s)/f_1(s)$ is the Bayes factor in favor of $H_\circ$. From a Bayesian viewpoint, $\alpha(s)$ is the posterior probability for $H_\circ$, assuming equal prior probabilities for $H_\circ$ and H_1!! For the Type II CEP one finds similarly:

$$\beta(s) \;=\; \mathcal{P}_1(\text{accepting } H_\circ \,|\, \mathcal{X}_s) \;=\; \frac{1}{1 + B(s)}, \tag{56}$$

which can be interpreted as the posterior prob. for H_1.

Conditional frequentist testing: Simple hypothesis versus composite hypothesis

Suppose we wish to test

$$H_\circ : \; \theta \; = \; \theta_\circ \tag{57}$$

versus

$$H_1 : \; \theta \; \in \; \Theta_1 \tag{58}$$

where $\theta_\circ \notin \Theta_1$.

Let's start with the Bayesian viewpoint. Assign a default prior probability $\pi_\circ = 1/2$ to $H_\circ$ and the prior density $g(\theta)/2$ to Θ_1, where $g(\theta)$ is a proper p.d.f. over Θ_1.

The Bayes factor in favor of $H_\circ$ is

$$B(x) \; = \; m_\circ(x)/m_1(x) \, , \tag{59}$$

with:

$$m_\circ(x) \;\; = \;\; f(x \,|\, \theta_\circ) \tag{60}$$

$$m_1(x) \;\; = \;\; \int_{\Theta_1} f(x \,|\, \theta) \, g(\theta) \, d\theta. \tag{61}$$

Hence the test:

For any b in the range of $B(x)$, let $\psi(b) = F_\circ^{-1}[1 - F_1(b)]$, where $F_\circ$ and F_1 are the cumulative distribution functions of B under $H_\circ$ and H_1 respectively. Define:

$$
\begin{array}{llllll}
r = 1 & \text{and} & a = \psi(1) & \text{if} & \psi(1) \geq 1 \\
r = \psi^{-1}(1) & \text{and} & a = 1 & \text{if} & \psi(1) < 1
\end{array}
$$

Then:

- if $B(x) \leq r$, reject $H_\circ$;
- if $r < B(x) < a$, make no decision;
- if $B(x) \geq a$, accept $H_\circ$.

In the first and third cases, report the posterior probabilities

$$
\begin{aligned}
\Pr(H_\circ \mid x) &= \alpha^\star(B(x)) = B(x)/(1 + B(x)) && (62) \\
\Pr(H_1 \mid x) &= \beta^\star(B(x)) = 1/(1 + B(x)). && (63)
\end{aligned}
$$

Similarly to the simple versus simple test, this test is also a conditional frequentist test, arising from the use of the conditioning statistic

$$S(X) \;=\; \min\{B(X)\,,\,\psi^{-1}(B(X))\} \tag{64}$$

over the domain $\mathcal{X}^\star = \{X \,:\; 0 \leq S(X) \leq r\}$ (the complement of $\mathcal{X}^\star$ is the no-decision region). The conditional error probabilities in this case are:

$$\alpha(s) \;\equiv\; \mathcal{P}_{\theta_\circ}(\text{rejecting } H_\circ \,|\, S(X) = s)\,, \tag{65}$$

$$\beta(\theta\,|\,s) \;\equiv\; \mathcal{P}_{\theta}(\text{accepting } H_\circ \,|\, S(X) = s)\,, \tag{66}$$

and one can prove that:

$$\alpha(s) \;=\; \alpha^\star(B) \tag{67}$$

$$E^{g(\theta\,|\,s)}[\beta(\theta\,|\,s)] \;=\; \beta^\star(B)\,, \tag{68}$$

where $g(\theta\,|\,s)$ is the posterior p.d.f. of θ conditional on H_1 being true and on the observed value of $S(X)$.

Some final comments

- The unconditional frequentist test (Neyman) effectively replaces the data by the indicator of its being in the acceptance or rejection region, thereby ignoring a maximum amount of information;

- The p-value test (Fisher) is even more unreasonable, when incorrectly viewed as an error probability;

- The unified Bayesian-frequentist conditioning tests have obvious scientific merits and pedagogical benefits. They yield actual error probabilities that can also be interpreted as posterior probabilities of hypotheses. Because they are Bayesianly justifiable, the results of these tests are not affected by the stopping rule.

- A somewhat sticky issue is the choice of conditioning statistic, corresponding to the arbitrariness in prior selection of the Bayesian formulation. Bayesians and frequentists have the same problem!

- Conditional frequentist methods can be developed for composite null hypotheses, basically by reducing the composite hypothesis to a simple one by Bayesian averaging. This is somewhat controversial because of the asymmetric treatment of null hypotheses and alternatives in classical statistics. Frequentists accept the notion of "average power", but may not like the idea of "average Type I error".