# TRACKING AND STATISTICAL KNOWLEDGE

KONSTANTIN GENIN

Abstract. Is there a tracking account on which diligent hypothesis testing generates knowledge? What about belief in frequentist confidence intervals? If there is something right about tracking, it ought to be able to explain how these workhorse methods of scientific inquiry generate knowledge. If it cannot, we must either abandon tracking or embrace scientific skepticism. This paper examines what kind of tracking conditions could make sense of our statistical practice.

## Sensitivity and Adherence

Nozick (1981) analyzes the knowledge relation as follows:

$S$ knows that $p$ if

(1) $S$ believes that $p$;
(2) $p$ is true;
(3) If $p$ were not the case, $S$ would not believe that $p$ (Sensitivity);
(4) If $p$ were the case, $S$ would believe that $p$ (Adherence).

If we are interested in scientific knowledge, it is natural to draw the connection with Neyman-Pearson theory. In classical frequentist hypothesis testing, we have a parametric model

$$\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$$

where $\Theta \subset \mathbb{R}$ and $p$ is a probability density function from some parametric family. We can think of parameters propositionally: a single $\theta \in \Theta$ individuates a possible world and $\Theta_0 \subset \Theta$ picks out a set of possible worlds. Given two disjoint propositions $\Theta_0$ and $\Theta_1$ we can ask whether the true world $\theta$ is a member of $\Theta_0$ or $\Theta_1$. In the former case we say that the null hypothesis is true, in the latter that the alternative hypothesis is true. A statistical test is an epistemic decision procedure with two possible outcomes: either you retain the null hypothesis or you reject it in favor of the alternative. There are two kinds of errors that a statistician faced with such a decision problem can make. Either she can falsely reject the null (Type I) or she can falsely retain the null (Type II).

| | $\theta \in \Theta_0$ | $\theta \in \Theta_1$ |
|---|---|---|
| Retain $\Theta_0$ (Believe $\Theta_0$) | No error | Type II error (false negative) |
| Reject $\Theta_0$ (Believe $\Theta_1$) | Type I error (false positive) | No error |

Belief in the null hypothesis is not usually meant to be taken very seriously. Belief in the alternative is meant to be a momentous epistemic decision. This is a reflection of an asymmetry in the utilities usually associated with the respective outcomes. Belief in the null hypothesis usually recommends no action, whereas rejection counsels some potentially dangerous decision. A common motivating example comes from the pharmaceutical setting. If we are researchers at BigPharma Co. our null might be that our new medication is no more effective than placebo. If we retain, we have wasted R&D money. If we reject, we prescribe a potentially dangerous new medication to our patients. Clearly, it is to be hoped that rejection of the null tends to generate knowledge of the alternative. The errors associated with our epistemic decision immediately suggest a statistical gloss of Nozick's tracking conditionals. Say that $S$ knows that $\Theta_1$ if

(1) $S$ believes $\Theta_1$;
(2) $\theta \in \Theta_1$;
(3) If $\Theta_1$ were false, $S$ would not believe $\Theta_1$ ($S$ avoids Type I errors);
(4) If $\Theta_1$ were true, $S$ would believe $\Theta_1$ ($S$ avoids Type II errors).

A well-designed test of a statistical hypothesis is expressly intended to minimize errors of Type I and II, so it seems as if Nozick's tracking conditionals – suitably paraphrased – promise an account of how tests of statistical hypotheses generate knowledge. The goal of this paper is to examine to what extent this promise is fulfilled.

## Tracking with Probabilities

The goal of Neyman-Pearson theory is to design tests with reasonable probability of avoiding errors of Type I and II. Of course, Nozick's tracking conditionals are articulated in terms of counterfactuals, so we must somehow translate the one kind of talk into the other. Roush (2005) suggests a straightforward importation of the tracking conditionals into probabilistic language. Say that $S$ knows that $p$ if

(1) $S$ believes that $p$;
(2) $p$ is true;
(3) P($S$ does not believe that $p \mid p$ is not the case) $\geq$ 1-$\alpha$ (Sensitivity);
(4) P($S$ believes that $p \mid p$ is the case) $\geq$ 1-$\alpha$ (Adherence).

Roush argues that her account has most of the advantages of Nozick's, without the arbitrary jury-rigging of similarity relations between possible worlds. Why worry about possible-world semantics if you can get by with conditional probabilities? The account is attractive, but several technical problems quickly arise. Probability zero events abound in statistics. Suppose I have observed a one-point sample $X_1 \sim F$ where $F$ is some continuous distribution and that $x_1 \neq 0$. Presumably, I know that $x_1$ is not equal to zero without doing much more than examining the sample. In order to evaluate the sensitivity condition, I need to investigate $P(\text{I do not believe that } x_1 \neq 0 | x_1 = 0)$. But since $F$ is a continuous distribution, $P(x_1 = 0) = 0$ so the conditional probability in the sensitivity

condition is undefined. Surely a tracking account ought to be able to handle such a trivial case.

We ought not make too much of this last criticism. Conditioning on probability zero events is a problem for everyone – not just Roush – and there are ways it can be remedied. More immediately problematic are events which have no probability. Suppose we have observed a sample $x_1, ..., x_N$ drawn from $F \sim \mathcal{N}(\mu, 1)$ where $\mu$ is unknown. We have tested the null hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu \neq 0$ and rejected. Suppose furthermore that we have designed our test so that the probability of false rejection is very low. We would like to say that that we know $H_1$, but in order to evaluate the sensitivity condition we need to compute $P(\text{retaining } H_0 \mid \mu = 0)$. But there is no frequentist probability associated with the event "$\mu = 0$". Perhaps Roush has in mind some sort of Bayesian quantity, but it does not seem that a tracking semantics could be easily represented in a Bayesian framework. After all, epistemic agents are meant to be tracking the *truth* of a hypothesis, not their own Bayesian posterior probabilities. Rigging up some way to condition on the event "$\mu = 0$" does violence to the realist intuition that statistical parameters are simple "out there" and are not the product of a stochastic process. Neyman-Pearson theory asks us to consider what the probability disposition of experiments would be like were the world to be governed by the parameter $\theta$. The theory is expressed explicitly in terms of possible worlds and a tracking semantics for frequentist statistics ought to respect the standard usage. Conditional probability is simply not enough for a frequentist semantics.

We turn now to defining sensitivity and adherence for beliefs governed by a statistical test. As before let

$$\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$$

be a family of probability measures having densitites $p(x, \theta)$. Suppose that we have independent and identically distributed observations $(X_1, ..., X_n)$ distributed according to $\mathcal{P}$. Let $\Theta_0, \Theta_1$ be disjoint, exhaustive subsets of $\Theta$. A test of $\Theta_0$ against $\Theta_1$ for the sample size $n$ is a mapping

$$\Psi_n : X^n \mapsto \{0, 1\}$$

where we use 1 to indicate rejection of $\Theta_0$. Define the *power function* by

$$\beta(\theta, \Psi_n) = P_\theta(\Psi_n(X^n) = 1)$$

for $\theta \in \Theta$. Finally, say that $\Psi_n$ is *level* $\alpha$ if

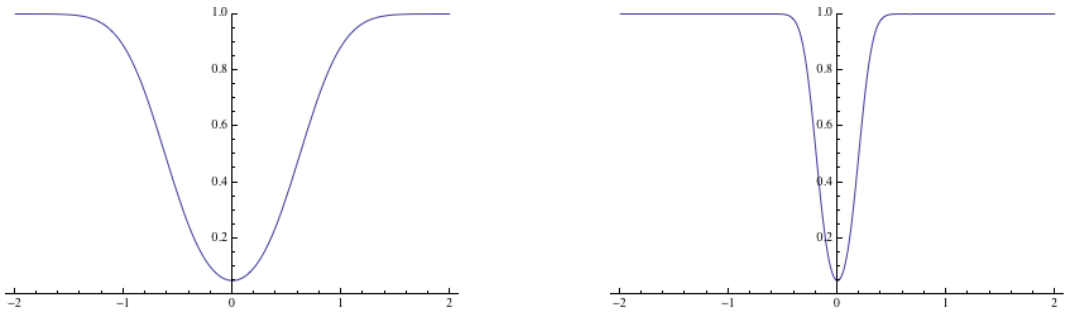
$$\sup_{\theta \in \Theta_0} \beta(\theta, \Psi_n) \leq \alpha$$

and that $\Psi_n$ has *power* $\beta$ if

$$\inf_{\theta \in \Theta_1} \beta(\theta, \Psi_n) \geq \beta.$$

The Neyman-Pearson testing procedure is to fix a level $\alpha \in [0, 1]$ and then maximize $\beta(\theta)$ for $\theta \in \Theta_1$ subject to $\beta(\theta) \leq \alpha$ for $\theta \in \Theta_0$. Now we can say that $S$ is $\alpha$-*sensitive* to $\Theta_1$ at sample size $n$ according to $\Psi_n$ if

(1) $S$ believes $\Theta_0$ if $\Psi_n(X^n) = 0$;
(2) $\Psi_n$ is level $\alpha$.




(A) $n = 10$ (B) $n = 100$

FIGURE 1. Plot of $\beta(\mu)$

And $S$ is $\beta$-*adherent* to $\Theta_1$ at sample size $n$ according to $\Psi_n$ if

(1) $S$ believes $\Theta_1$ if $\Psi_n(X^n) = 1$;
(2) $\Psi_n$ has power $\beta$.

As desired, if $\alpha$ is small, then it is probable that $S$ does not believe $\Theta_1$ if it is false; and if $\beta$ is large, it is probable that $S$ believes $\Theta_1$ if it is true. We can now straightforwardly translate Nozick's tracking conditions. Say that $S$ knows that $\Theta_1$ according to test $\Psi_n$ and the sample $x^n$ if

(1) $\Psi_n(x^n) = 1$;
(2) $\theta \in \Theta_1$;
(3) $S$ is .05-sensitive to $\Theta_1$ at sample size $n$ according to $\Psi_n$;
(4) $S$ is .95-adherent to $\Theta_1$ at sample size $n$ according to $\Psi_n$.

Does the preceding definition allow $S$ statistical knowledge in typical cases? Let us examine a canonical example. Suppose we have independent and identically distributed observations $(X_1, .., X_n)$ and that $X_i \sim \mathcal{N}(\mu, 1)$ where $\mu$ is unknown. We are interested in testing the null hypothesis $\mu = 0$ against the alternative $\mu \neq 0$. The standard $\alpha$-level test $\Psi_n$ rejects the null when $|\bar{x}| > \frac{1}{\sqrt{n}} z_{\alpha/2}$ where $z_\alpha$ is the $\alpha$-quantile of the standard normal distribution. Figure 1 plots the power function of this test at sample size 10 and 100 for $\alpha = .05$.

Now suppose $S$ believes the alternative hypothesis because her test rejected at this significance level. Do the tracking conditions grant $S$ knowledge of the alternative hypothesis? In fact, they do not. $S$ satisfies sensitivity, since she has a 5% chance of falsely believing that $\mu \neq 0$. But she spectacularly fails the adherence condition, since her test has almost no power at parameters near 0. The situation is even worse, because $S$ would fail the current adherence condition *in the limit* – no matter how much data $S$ saw, there would be a parameter sufficiently close to zero against which she had no power. By a simple continuity argument:

$$\lim_{n \to \infty} \inf_{\mu \in \Theta_1} \beta(\mu, \Psi_n) = \alpha$$

The best we can say for $S$ is that the power of her test converges pointwise to 1:

$$\inf_{\mu \in \Theta_1} \lim_{n \to \infty} \beta(\mu, \Psi_n) = 1$$

But this is not sufficient to make her adherent to $\mu \neq 0$. If $S$ fails to reject, the situation is dual with respect to knowledge of the null. Since $S$ has a 95% chance of believing $\mu = 0$ if it is true, she satisfies adherence. But since there are parameters arbitrarily close to 0 at which she has very little power, she fails sensitivity.

Perhaps we have stated the tracking conditions too strongly. Nozick does not insist that we satisfy the tracking conditions at all possible worlds, only at the worlds nearby to ours. Recourse to a similarity relation over worlds might yield the required conditions. In this case, we are faced with only unpalatable options. It is precisely those parameters that are *near* the 0-world that cause trouble; the standard test has plenty of power against far-off parameters. To say that the nearby worlds are the worlds in which our test is already adherent (or sensitive) seems ad-hoc: it is to say that if the mean weren't zero we would have noticed it already, even at small sample sizes. This would make much of the practice in frequentist statistics look inexplicable: why should we be wary of inferences from small sample sizes, if even at small $n$ our tests are already effective in the counterfactual situations that we think are plausible? The argument from back-tracking conditionals given by Nozick (1981) and Roush (2005) looks reasonable in the case of massive violations of natural laws, but looks wildly ad hoc in the statistical setting. The tracking conditions are simply too strong to give an account of how statistical testing yields knowledge.

## Tracking in the Limit

It is hard to believe that statistical tests do not yield knowledge. They are, after all, a workhorse methodology for working scientists. If we think there is something right about the tracking analysis of knowledge, we have to liberalize the tracking conditions to make sense of statistical and scientific practice. Kelly (2013) writes that inductive learning is a matter of believing the truth now and eliminating error in other possible worlds *eventually*. We might reverse his lexicography and say that statistical knowledge is a matter of minimizing Type I errors *now* and eliminating Type II errors *eventually*. Say that $S$ is *weakly asymptotically $\beta$-adherent* to $\Theta_1$ according to the sequence of tests $\{\Psi_n\}$ if

(1) For all $n$, $S$ believes $\Theta_1$ if $\Psi_n(X^n) = 1$;
(2) $\inf_{\theta \in \Theta_1} \lim_{n \to \infty} \beta(\theta, \Psi_n) = \beta$.

Say that $S$ is *weakly asymptotically adherent* to $\Theta_1$ if S is weakly asymptotically $\beta$-adherent for $\beta = 1$. The natural thing to say is that $S$ knows that $\Theta_1$ according to test $\Psi_n$ and the sample $x^n$ if

(1) $\Psi_n(x^n) = 1$;
(2) $\theta \in \Theta_1$;
(3) $S$ is .05-sensitive to $\Theta_1$ at sample size $n$ according to $\Psi_n$;
(4) $S$ is weakly asymptotically adherent to $\Theta_1$ according to $\{\Psi_n\}$.

$S$ believes $\Theta_1$; if $\Theta_1$ were false, $S$ would probably not believe it; and in all $\Theta_1$ worlds, $S$ is committed to a testing regime that guarantees she will believe $\Theta_1$ eventually. The fact that $S$ has rejected the null *already* is a bit of epistemic luck, but her disposition is such that she would have eventually rejected the null no matter how $\Theta_1$ were true.

This analysis looks plausible. It demands sensitivity at finite samples, but only pointwise asymptotic adherence. This lexicographic preference is consistent both with Nozick's emphasis on the sensitivity condition and the standard Neyman-Pearson practice of privileging Type I over Type II errors. Unfortunately, it cannot be correct. Suppose $S$ were determined to reject the null hypothesis $\mu = 0$ in the canonical problem we examined previously. At each $n$ she computes the sample mean, and rejects only if $|\bar{x}| > \frac{1}{\sqrt{n}} z_{\alpha/2}$ where $\alpha$ is fixed at .05. If she fails to reject, she continues to sample, making sure to appropriately adjust her test to achieve .05 significance on the larger sample. When she sees a sufficiently extremal statistic, she halts the experimental procedure and publishes a significant result. This is an example of "sampling to a foregone conclusion" a classical bugbear in statistical inference. The problem with such a procedure is that with probability one, $S$ will reject the null at some sample size, even if it is true. At all sample sizes $S$ is .05 sensitive to $\Theta_1$. Furthermore, she is asymptotically adherent, since her test is powering up on all $\mu \in \Theta_1$ as $n$ increases. But $S$ cannot know $\Theta_1$. The reason that $S$ does not know $\Theta_1$ – even if she correctly rejects – is that her testing regime is not actually sensitive to $\Theta_1$. Even though at all sample sizes she uses a test that is sensitive at the 5% level, she is guaranteed to believe $\Theta_1$ *eventually*, even if it is false. $S$ is synchronically sensitive, but diachronically insensitive. It is not enough to be .05-sensitive at all sample sizes. To have knowledge one needs to be increasingly sensitive as the sample sizes grow larger. Say that $S$ is *strongly asymptotically $\alpha$-sensitive* to $\Theta_1$ according to the sequence of tests $\{\Psi_n\}$ if

(1) For all $n$, $S$ believes $\Theta_0$ if $\Psi_n(X^n) = 0$;
(2) $\lim_{n\to\infty} \sup_{\theta\in\Theta_0} \beta(\theta, \Psi_n) = \alpha$

Say that $S$ is *strongly asymptotically sensitive* to $\Theta_1$ if S is asymptotically $\alpha$-sensitive for $\alpha = 0$. Note that this definition preserves the lexicographic order, since we demand uniform convergence of sensitivity (significance) and only pointwise convergence of the adherence (power). Now we say that $S$ knows that $\Theta_1$ according to test $\Psi_n$ and the sample $x^n$ if

(1) $\Psi_n(x^n) = 1$;
(2) $\theta \in \Theta_1$;
(3) $S$ is at least .05-sensitive to $\Theta_1$ at sample size $n$ according to $\Psi_n$;
(4) $S$ is weakly asymptotically adherent to $\Theta_1$ according to $\{\Psi_n\}$;
(5) $S$ is strongly asymptotically sensitive to $\Theta_1$ according to $\{\Psi_n\}$.

$S$ believes $\Theta_1$ right now; if $\Theta_1$ were false, $S$ would probably not now believe it; in all $\Theta_1$ worlds, $S$ is committed to a testing regime that guarantees she will believe $\Theta_1$ eventually; and in all $\Theta_0$ worlds, $S$ is committed to a testing regime that guarantees she will believe $\Theta_0$ eventually. There is now no way to cheat the knowledge definition by sampling to a

foregone conclusion – the way to salvage the tracking analysis is to go fully asymptotic. Knowledge is a matter of controlling the probability of Type I errors now, and being disposed to eliminate Type I and II errors in the limit. A slogan for this view might be "sensitivity *now*, adherence in the limit." Such a testing regime exists for the toy inference problem we have been considering. Whether such a regime exists for a given inference problem is an interesting question in asymptotic testing theory.

## Tracking the Null

The preceding knowledge analysis would never allow knowledge of $\Theta_0$, because the third and fourth condition could never be satisfied. No matter how much data $S$ sees, she can never be more than $(1-\alpha)$-sensitive to $\Theta_0$. Obviously, if $\alpha \to 0$ then $S$ gets only less sensitive as she continues her regime of testing. Furthermore, $S$ cannot be asymptotically sensitive, since we have required uniform convergence of sensitivity and $S$ can at best be pointwise sensitive to $\Theta_0$ in the limit. Nevertheless, we might want to say that knowledge of the null is possible. Mayo (1996) argues that hypotheses that pass increasingly severe tests are more confirmed. It is to be hoped that confirmation has something to do with knowledge. How would we have to alter the tracking conditions to get knowledge of the null? Say that $S$ is *weakly asymptotically sensitive* to $\Theta_0$ according to the sequence of tests $\{\Psi_n\}$ if

(1) For all $n$, $S$ believes $\Theta_1$ if $\Psi_n(X^n) = 1$;
(2) $\inf_{\theta \in \Theta_1} \lim_{n\to\infty} \beta(\mu, \Psi_n) = 1$

and that $S$ is *strongly asymptotically adherent* to $\Theta_0$ according to the sequence of tests $\{\Psi_n\}$ if

(1) For all $n$, $S$ believes $\Theta_0$ if $\Psi_n(X^n) = 0$;
(2) $\lim_{n\to\infty} \inf_{\theta \in \Theta_0} \beta(\theta, \Psi_n) = 1$.

Now we might say that $S$ knows that $\Theta_0$ according to test $\Psi_n$ and the sample $x^n$ if

(1) $\Psi_n(x^n) = 0$;
(2) $\theta \in \Theta_0$;
(3) $S$ is at least .95-adherent to $\Theta_0$ at sample size $n$ according to $\Psi_n$;
(4) $S$ is strongly asymptotically adherent to $\Theta_0$ according to $\{\Psi_n\}$;
(5) $S$ is weakly asymptotically sensitive to $\Theta_0$ according to $\{\Psi_n\}$.

As is to be expected, the epistemic situation with respect to the null is dual to that that of the alternative. Such an analysis would allow knowledge of the null, but only by dropping the synchronic sensitivity requirement entirely and significantly weakening asymptotic sensitivity. A slogan for this view is "adherence *now*, sensitivity in the limit." Doing so yields a lot of lucky knowledge. It is easy to be .95-adherent; it is simply a matter of picking the right test. If we allowed the previous analysis to stand, knowledge would be a matter of believing a true null hypothesis and resolving to test it appropriately. This doesn't seem right, but it may be the price of a lot of scientific knowledge.

## Tracking Confidence Intervals

Confidence interval procedures are another methodology at the heart of frequentist statistics. It is natural to ask whether there is some tracking account on which these procedures generate knowledge. As before let

$$\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$$

be a family of probability measures having densitites $p(x, \theta)$. Suppose that we have independent and identically distributed observations $(X_1, ..., X_n)$ distributed according to $\mathcal{P}$. A confidence interval with coverage $1 - \alpha$ is a random interval

$$C_n(X^n) = [L(X^n), U(X^n)] \subset \Theta$$

such that

$$\inf_{\theta \in \Theta} P_\theta(\theta \in C_n(X^n)) \geq 1 - \alpha$$

Since this interval is just a set of worlds, we can think of $C_n(X^n)$ as a random proposition. A good confidence interval procedure generates random propositions that are probably true. That is, if we had a procedure with 95% coverage and repeated it many times on independent data sets of the same size, we would only fail to trap the true parameter 5% of the time.

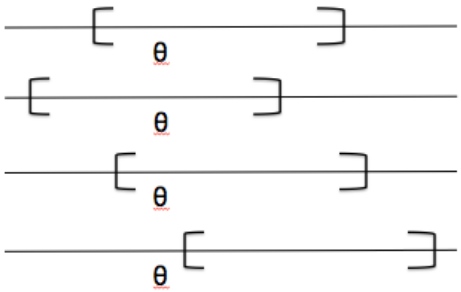



Figure 2. The confidence interval traps $\theta$ with probability $1 - \alpha$.

Suppose $S$ believes the proposition $C_n(x^n)$ that results from a confidence interval procedure on a particular sample $x^n$. Sensitivity asks us to evaluate the probability that $S$ would not believe $C_n(x^n)$ if it were false, i.e.

$$\inf_{\theta \in \ \Theta \setminus C_n(x^n)} P_\theta \left( C_n(X^n) \neq C_n(x^n) \right)$$

The trouble is that this quantity is always low, whether $C_n(x^n)$ is true or false. So long as $\mathcal{P}$ is a continuous family, the probability that $S$ believes any *particular* proposition is low, since a small change in the sample could yield a slightly different interval. So in this sense, $S$ is only trivially sensitive to $C_n(x^n)$. Clearly, $S$ cannot be adherent to $C_n(x^n)$ either. What we can say is that if $C_n(x^n)$ is true, then $S$ very probably won't believe its negation $\Theta \setminus C_n(x^n)$[1], since she will believe some proposition that overlaps $C_n(x^n)$ at the

[1]In what follows we write this proposition as $\neg C_n(x^n)$.

true parameter:

$$\sup_{\theta \in C_n(x^n)} P_\theta \left( C_n(X^n) \cap C_n(x^n) = \varnothing \right) \leq \alpha.$$

This is a straightforward consequence of the confidence property. Since $C_n(x^n)$ is true, $\theta \in C_n(x^n)$ and therefore it is probable that $\theta$ would also be in any other confidence interval $S$ would construct with her procedure. So if $C_n(x^n)$ were true, $S$ would not rule it out with high probability. This is a kind of adherence property. Furthermore, if $C_n(x^n)$ were false, there is no guarantee that $S$ would believe $\neg C_n(x^n)$, since the true parameter could be just outside of $C_n(x^n)$ and any confidence interval constructed there would probably overlap with $C_n(x^n)$. By a simple continuity argument:

$$\inf_{\theta \notin C_n(x^n)} P_\theta \left( C_n(X^n) \cap C_n(x^n) = \varnothing) \right) \leq \alpha.$$

So if $C_n(x^n)$ were false, we have no guarantee that $S$ would rule it out. The best we can say is that if $C_n(x^n)$ were false, $S$ would probably not rule out $\neg C_n(x^n)$:

$$\sup_{\theta \notin C_n(x^n)} P_\theta \left( C_n(X^n) \cap \Theta \setminus C_n(x^n) = \varnothing \right) \leq \alpha.$$

The trouble is that if $C_n(x^n)$ were true, there are parameters at which $S$ would probably not rule out $\neg C_n(x^n)$ either. What we can say is that if $C_n(x^n)$ is false, then $S$ would be able to rule it out *eventually* – the shrinking confidence intervals around the true parameter would eventually probably separate $S$ from $C_n(x^n)$:

$$\inf_{\theta \notin C_n(x^n)} \lim_{m \to \infty} P_\theta \left( C_m(X^m) \cap C_n(x^n) = \varnothing \right) = 1$$

So only the "adherence *now*, sensitivity in the limit" theory underwrites knowledge of confidence intervals. As with knowledge of the null, insisting on sensitivity at finite samples makes knowledge of confidence intervals impossible.

## Safety and Statistics

Tracking can only recover statistical knowledge by going asymptotic. Can the safety theorist do any better? The situation for hypothesis testing looks reasonably good. Say that $S$'s belief in $\Theta_1$ is $\alpha$-*safe* at sample size $n$ according to $\Psi_n$ if

$$\text{If } P_\theta(\Psi_n(X^n) = 1) \geq (1 - \alpha) \text{ then } \theta \in \Theta_1.$$

That is, if $S$ is likely to reject the null (believe the alternative), then the null is false (alternative is true). Any $\alpha$-level hypothesis test is safe e.g. if $S$ is likely to reject the null hypothesis $\mu = 0$, then the truth is far away enough from 0 that $S$'s test already has high power. Safety deals with this quite elegantly. But safety will have difficulty with confidence intervals. We have seen that the probability that $S$ believes any *particular* interval is always low. We might say that $S$'s belief in $C_n(x^n)$ is $\alpha$-*safe* if

$$\text{If } P_\theta(C_n(X^n) \cap C_n(x^n) \neq \varnothing) \geq (1 - \alpha) \text{ then } \theta \in C_n(x^n).$$

But then believing $C_n(x^n)$ is not safe! We have seen that if $\theta$ is outside of $C_n(x^n)$ but close to its boundary, the resulting confidence interval will probably intersect $C_n(x^n)$. What is true is that, if you are likely to rule out $C_n(x^n)$, then it is probably false:

$$\text{If } P_\theta(C_n(X^n) \cap C_n(x^n) = \varnothing) \geq (1 - \alpha) \text{ then } \theta \notin C_n(x^n).$$

This might be an interesting property, but it is not safety. Safety is not able to explain why we are interested in frequentist confidence intervals.

## Concluding Remarks

If we want to recover most of what we take to be statistical knowledge, it seems we have to go with an "adherence *now*, sensitivity in the limit" theory. Insisting otherwise would rule out confidence intervals as a kind of knowledge and make knowledge of the null hypothesis – no matter how severely tested – impossible. Can we live with such a view? Admittedly, it does some violence to our folk semantics. Suppose I am a pessimistic but persistent player of the lottery. I believe every ticket I buy is a loser. Nevertheless, I check the winning numbers when they are announced. I am always adherent to the "my ticket is a loser" hypothesis and I am sensitive in the limit. If my ticket is a winner, I will find out eventually. If my ticket is a loser, do I know it as soon as I buy the ticket? Perhaps I do. But it is intuitive to say that I don't know, because my belief fails sensitivity: even if my ticket were a winner, I would still believe it was a loser. "Adherence now, sensitivity in the limit" does not support this intuition. Perhaps I can know my ticket is not a winner. But suppose now that I am a sunny optimist. I believe every ticket I buy is a winner. Of course, I check the winning numbers when they are announced and I am usually disappointed. I am adherent to the "my ticket is a winner" hypothesis, and sensitive in the limit. But surely when my lucky day arrives I won't know that my ticket is a winner! "Adherence now, sensitivity in the limit" does not discriminate between the pessimist and the optimist. I am not sure what to think about this. Perhaps the epistemic standards for statistical knowledge are just different than the standards for knowledge of chancy propositions. After all, the true parameter is not announced at the end of the week. A more complete theory would systematically shift the epistemic standards with the shifting context of inquiry.

## References

[1] Kelly, Kevin T. "A Hyper-intensional Learning Semantics for Inductive Empirical Knowledge" in *Logical/Informational Dynamics, a Festschrift for Johan van Benthem*, Alexandru Baltag and Sonja Smets, eds, Dordrecht: Springer, 2013.

[2] Mayo, Deborah G. *Error and the growth of experimental knowledge.* University of Chicago Press, 1996.

[3] Nozick, Robert. *Philosophical Explanations.* Harvard University Press, 1981.

[4] Roush, Sherrilyn. *Tracking truth: Knowledge, evidence, and science.* Oxford: Clarendon Press, 2005.