TREATMENT CHOICE WITH TRIAL DATA:

Statistical Decision Theory Should Supplant Hypothesis Testing

Charles F. Manski

Department of Economics and Institute for Policy Research

forthcoming in *The American Statistician*,

Special Issue on "Statistical Inference in the 21st Century: A World Beyond $P<0.05$"

Using Hypothesis Tests to Compare Treatments

A common procedure when comparing two treatments in a trial is to view one as the status quo and the other as an innovation.

The usual null hypothesis is that the innovation is no better than the status quo and the alternative is that the innovation is better.

If the null hypothesis is not rejected, it is recommended that the status quo treatment continue to be used in clinical practice.

If the null is rejected, it is recommended that the innovation become the treatment of choice.

The convention has been to fix the probability of a Type I error. Sample size determines the probability of a Type II error.

International Conference on Harmonisation (1999) provides guidance for the design and conduct of trials evaluating pharmaceuticals, stating (p. 1923):

> "Conventionally the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. The probability of type II error is conventionally set at 10% to 20%."

Manski and Tetenov (*PNAS*, 2016) give several reasons why hypothesis testing may yield unsatisfactory results for medical decisions and other treatment choices.

1. Use of Conventional Asymmetric Error Probabilities

It has been standard to fix the probabilities of Type I and II errors at 5% and 10-20%. Testing theory gives no rationale for selection of these error probabilities. It does not explain why a clinician concerned with patient welfare should find it reasonable to make treatment choices that have a substantially greater probability of Type II than Type I error.

2. *Inattention to Magnitudes of Losses to Welfare When Errors Occur*

A clinician should care about the magnitudes of the losses to patient welfare that arise when errors occur. A given error probability should be less acceptable when the welfare difference between treatments is larger. Testing theory does not take this into account.

3. Limitation to Settings with Two Treatments

A clinician often chooses among several treatments. Many trials compare more than two treatments. The standard theory of hypothesis testing only contemplates choice between two treatments. Statisticians have struggled to extend it to deal sensibly with comparisons of multiple treatments.

Example of Issues 1 and 2

A terminal form of cancer may be treated by a status quo treatment or an innovation.

Mean patient life span with the status quo treatment is known to be one year. Researchers see two possibilities for the innovation. It may be less effective than the status quo, yielding mean life span of 1/3 of a year, or it may be more effective, with mean life span of 5 years.

A trial is performed to learn the effectiveness of the innovation. The data are used to perform a test comparing the innovation and the status quo. The error probabilities are set at 0.05 and 0.20. The result is used to choose between the treatments.

A Type I error occurs with probability 0.05 and reduces mean patient life span by 2/3 of a year (1 year minus 1/3 year).

A Type II error occurs with probability 0.20 and reduces mean patient life span by 4 years (5 years minus 1 year).

Use of the test to choose between the status quo and the innovation implies that society is willing to tolerate a large (0.20) chance of a large welfare loss (4 years) when making a Type II error, but only a small (0.05) chance of a small welfare loss (2/3 of a year) when making a Type I error.

The theory of hypothesis testing does not motivate this asymmetry.

Principles of Statistical Decision Theory

Wald (1950) considered the uses of sample data to make decisions under uncertainty. He posed the task as choice of a statistical decision function, which maps potential data into a choice among the feasible actions.

He recommended evaluation of statistical decision functions as procedures, specifying how a decision maker would use whatever data are realized. Thus, the theory is frequentist.

He proposed evaluate of a statistical decision function by the distribution of loss that it yields across realizations of the sampling process. He focused attention on mean sampling performance.

He prescribed a three-step decision process.

1. Specify the state space (parameter space), which indexes the set of values of unknown quantities that the decision maker deems possible.

2. Eliminate inadmissible statistical decision functions.

    A decision function is inadmissible (weakly dominated) if there exists another one that yields at least as good sampling performance in every possible state of nature and strictly better performance in some state.

3. Use some criterion to choose an admissible statistical decision function. Leading criteria are maximization of subjective expected welfare (Bayes rule), maximin, and minimax regret.

Early applications of the Wald theory focused on prediction rather than treatment choice.

A literature on treatment choice has developed since the early 2000s. See Manski (2004, 2005, 2007), Manski and Tetenov (2007, 2014, 2016, 2018), Hirano and Porter (2009), Schlag (2006), Stoye (2009, 2012), Tetenov (2012), and Kitagawa and Tetenov (2018).

A statistical decision function uses the data to choose a treatment allocation, so such a function has been called a *statistical treatment rule* (STR). The mean sampling performance of an STR is its *expected welfare*.

The state space specifies the feasible distributions of treatment response.

The objective has been maximization of a social welfare function that sums treatment outcomes across the population. The literature mainly studies the minimax-regret criterion.

Treatment Choice with Existing Trial Data

Consider a trial with two treatments and a population of observationally identical patients.

Suppose that a health planner must assign treatment A or B to each member of patient population J.

Each patient $j \in J$ has response function $y_j(\cdot)$: $T \rightarrow Y$ mapping treatments $t \in T$ into individual outcomes $y_j(t) \in R$. Let P denote the distribution of treatment response in the population.

The members of the population may respond heterogeneously to treatment, but they are observationally identical to the planner.

For any $\delta \in [0, 1]$, the planner can allocate a fraction δ of patients to treatment B and $1 - \delta$ to A. The planner wants to choose δ to maximize an additive welfare function

$$U(\delta, P) = E[y(A)]\cdot(1 - \delta) + E[y(B)]\cdot\delta = \alpha\cdot(1 - \delta) + \beta\cdot\delta = \alpha + (\beta - \alpha)\cdot\delta,$$

where $\alpha \equiv E[y(A)]$ and $\beta \equiv E[y(B)]$.

$\beta - \alpha$ is the average treatment effect. It is optimal to set $\delta = 1$ if $\beta - \alpha > 0$ $\delta = 0$ if $\beta - \alpha < 0$.

The problem of interest is treatment choice when incomplete knowledge of P makes it impossible to determine the sign of $\beta - \alpha$.

Sample data are available, with sample space Ψ and sampling distribution Q.

An STR $\delta(\cdot): \Psi \to [0, 1]$ maps the data into a treatment allocation. The welfare realized with data ψ is the random variable

$$U(\delta, P, \psi) = \alpha + (\beta - \alpha)\cdot\delta(\psi).$$

The state space $[(P_s, Q_s), s \in S]$ is the set of (P, Q) pairs that the planner deems possible.

Expected welfare in state s is

$$W(\delta, P_s, Q_s) = \alpha_s + (\beta_s - \alpha_s)\cdot E_s[\delta(\psi)],$$

where $E_s[\delta(\psi)] \equiv \int_\Psi \delta(\psi) dQ_s(\psi)$.

The Bayes, maximin, and MR rules are

Bayes rule: $\max_{\delta \in [0,1]} \int_S W(\delta, P_s, Q_s)d\pi(s)$,

where π is a subjective distribution on the state space.

Maximin rule: $\max_{\delta \in [0,1]} \min_{s \in S} W(\delta, P_s, Q_s)$.

Minimax-regret rule: $\min_{\delta \in [0,1]} \max_{s \in S} [\max(\alpha_s, \beta_s) - W(\delta, P_s, Q_s)]$.

Measuring Performance by Maximum Regret

Practical Appeal

MR decisions behave more reasonably than maximin ones in the context of treatment choice.

When a trial has a balanced design and outcomes take a bounded range of values, it has been found that the MR rule is well approximated by the *empirical success* (ES) rule, which chooses the treatment with the highest observed average outcome in the trial.

In contrast, the maximin rule commonly ignores the trial data, whatever they may be.

Conceptual Appeal

Maximum regret quantifies how lack of knowledge of the true state diminishes the quality of decisions. An STR with small maximum regret is uniformly near-optimal across all states.

The concept is especially transparent when there are two treatments, say A and B.

In a state where A is better, the regret of an STR is the product of its probability of a Type I error (choosing B) and the magnitude of the loss in expected welfare that occurs when choosing B.

In a state where B is better, regret is the probability of a Type II error (choosing A) times the magnitude of the loss in expected welfare when choosing A.

A particularly simple case occurs when there are two states of nature and when Type I and Type II errors yield equal losses in expected welfare, say L. Then the maximum regret of an STR is L times the maximum of its probabilities of Type I and Type II errors.

Suppose that sample size has been chosen to give 0.05 probability of Type I error and 0.20 probability of Type II error, using a conventional test. Consider the STR that uses this test to choose a treatment. The maximum regret of this "test rule" is L times 0.20.

One can obtain a test rule with smaller maximum regret by enlarging the critical region for the test. Enlarging the critical region increases the probability of Type I error and reduces that of Type II error.

Maximum regret decreases until one enlarges the critical region to the degree that it equalizes the probabilities of Type I and Type II errors.

Designing Trials to Enable Near-Optimal Treatment Choice

(Manski and Tetenov, *PNAS*, 2016)

An ideal objective is to collect data that enable implementation of an *optimal* rule−−one whose expected welfare equals the welfare of the best treatment in every state of nature.

Optimality is not achievable in general, but *ε-optimal* rules do exist when trials have large enough sample size.

An ε-optimal rule has expected welfare within ε of the welfare of the best treatment in every state. Equivalently, it has maximum regret no larger than ε.

Implementation of the idea requires specification of a value for ε.

The necessity to choose an effect size of interest when designing trials already arises in conventional practice, where the trial planner must specify the alternative hypothesis to be compared with the null.

A possibility is to base ε on the *minimum clinically meaningful difference* (MCMD) in the average treatment effect comparing alternative treatments.

Many medical writers call an average treatment effect clinically significant if its magnitude is greater than ε for a specified value of ε deemed minimally consequential in clinical practice.

We consider trials that draw predetermined numbers of subjects at random within groups stratified by covariates and treatments.

The analytical findings are simple sufficient conditions on sample sizes that ensure existence of ε-optimal treatment rules when outcomes are bounded.

These conditions are obtained by application of Hoeffding (*JASA*, 1963) large deviations inequalities to evaluate the performance of empirical success rules.

We provide exact computations of minimal sample sizes enabling ε-optimality that hold when there are two treatments and outcomes are binary.

Findings with Binary Outcomes, Two Treatments, and Balanced Designs

Determination of sample sizes that enable near-optimal treatment is simple in settings with binary outcomes, two treatments, and a balanced design which assigns the same number of subjects to each treatment group.

We compute the minimum sample size that enables ε-optimality when a clinician uses one of three different treatment rules, for various values of ε.

ε	ES Rule	One-Sided 5% z-Test	One-Sided 1% z-Test
0.01	145	3488	7963
0.03	17	382	879
0.05	6	138	310
0.10	2	33	79
0.15	1	16	35

Based on our exact calculations and analytical findings with large-deviations inequalities, we conclude that sample sizes determined by clinically relevant near-optimality criteria tend to be much smaller than ones set by conventional statistical power criteria.

Reduction of total sample size can lower the cost of executing trials, the time needed to recruit subjects, and the complexity of managing trials across centers.

Reduction of sample size per treatment arm can make it feasible to perform trials that increase the number of treatment arms and, hence, yield information about a wider variety of treatment options.

Trial Size for Near-Optimal Choice Between Surveillance and Aggressive Treatment: Reconsidering MSLT-II (Manski and Tetenov, 2018)

We develop and apply a refined version of the analysis in Manski and Tetenov (P*NAS*, 2016) to trials that compare aggressive treatment of patients with surveillance.

Internists choose between prescription of pharmaceuticals and surveillance when treating patients at risk of heart disease or diabetes. Oncologists choose between surveillance and aggressive treatments such as surgery or chemotherapy when treating cancer patients at risk of metastasis.

Aggressive treatment may be appealing to the extent that it better prevents onset or reduces the severity of illness. Surveillance may be attractive to the extent that it avoids side effects that may occur with aggressive treatment.

The need for a refined version of the analysis arises because our earlier work studied settings in which there is only a primary health outcome of interest, without secondary outcomes.

An important aspect of choice between surveillance and aggressive treatment is that the latter may have side effects.

The prevailing approach to choice of sample size in trials has been to focus entirely on the primary outcome of a treatment, without considering secondary outcomes.

When aggressive treatment may have serious side effects, it is more reasonable to consider how the primary outcome and side effects jointly determine patient welfare.

This requires new analysis.

As a case study, we reconsider a recent trial comparing *nodal observation* and *lymph node dissection* when treating patients with early-stage cutaneous melanoma at risk of metastasis.

Nodal observation is surveillance of lymph nodes by ultrasound scan, a procedure that has negligible side effects. Lymph node dissection is a surgical procedure in which the lymph nodes in the relevant regional basin are removed.

Dissection is commonly viewed as an aggressive treatment. A particularly concerning side effect is lymphedema, which may reduce patient quality of life substantially.

Choice between nodal observation and lymph node dissection is a common decision faced in early treatment of melanoma, breast cancer, and other forms of localized cancer.

The Multicenter Selective Lymphadenectomy Trial II (MSLT-II) compared dissection and observation for melanoma patients who had recently undergone sentinel lymph-node biopsy and who had obtained a positive finding of malignancy.

The primary outcome was defined to be melanoma-specific survival for three years following the date of randomization.

Findings were reported in Faries *et al*. (*NEJM*, 2017).

Our concern is choice of sample size in the trial.

Using a conventional statistical power calculation, the investigators assigned 971 patients to dissection and 968 to observation.

Choosing the MSLT-II Sample Size to Enable Near-Optimal Treatment

We assume a simple patient welfare function.

Welfare with nodal observation equal 1 if a patient survives for three years and 0 otherwise.

Welfare with dissection depends on whether a patient experiences lymphedema.

When a patient does not experience lymphedema, welfare with dissection equals 1 if the patient survives for three years and 0 otherwise.

When a patient experiences lymphedema, welfare is lowered by a specified fraction h, whose value expresses the harm associated with lymphedema. A patient who experiences lymphedema has welfare $1 - h$ if he survives and $-h$ if he does not survive.

Let nodal observation be treatment A. Let y(A) = 1 if a patient survives and y(A) = 0 otherwise. Mean patient welfare with observation is the survival probability P[y(A) = 1].

Let lymph node dissection be treatment B. Let y(B) = 1 if a patient survives and y(B) = 0 otherwise. Let s(B) = 1 if a patient experiences lymphedema and s(B) = 0 otherwise. Mean patient welfare with dissection is $P[y(B) = 1] - h \cdot P[s(B) = 1]$.

Observation is optimal if $P[y(A) = 1] > P[y(B) = 1] - h \cdot P[s(B) = 1]$.

The MSLT-II trial yields information about the probabilities of survival and lymphedema.

The sample size determines the extent of the information. For any positive ε, a sample of size N per treatment arm enables ε-optimal treatment if N is sufficiently large.

Findings

We compute sample sizes that enable ε-optimal treatment for any values of h and ε. We assume that treatment choice will be made with the empirical success rule.

We preform computations using two methods to determine maximum regret. One applies simulated annealing and the other uses a normal approximation to the finite-sample distribution of empirical success.

Table 1: Near-optimality (maximum regret) of ES rules
(N is the number of subjects per treatment arm)

		h = 0		*h* = 0.1		*h* = 0.2		*h* = 0.3		*h* = 0.4		*h* = 0.5	
N =		Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.
10		0.038209	0.037490	0.044905	0.039672	0.045017	0.041857	0.045794	0.044046	0.046704	0.046237	0.049236	0.048431
20		0.026947	0.026689	0.030401	0.028180	0.030479	0.029672	0.031212	0.031166	0.032803	0.032661	0.034487	0.034157
30		0.021983	0.021841	0.024046	0.023039	0.024516	0.024237	0.025874	0.025435	0.026805	0.026634	0.028039	0.027834
40		0.019029	0.018937	0.020710	0.019963	0.021105	0.020989	0.021930	0.022016	0.023172	0.023044	0.024218	0.024071
50		0.017016	0.016949	0.018182	0.017860	0.018829	0.018772	0.019865	0.019683	0.020688	0.020595	0.021621	0.021507
60		0.015530	0.015480	0.016640	0.016307	0.017170	0.017134	0.018019	0.017962	0.018859	0.018789	0.019709	0.019617
70		0.014376	0.014336	0.015228	0.015099	0.015890	0.015861	0.016708	0.016624	0.017444	0.017387	0.018227	0.018150
80		0.013447	0.013414	0.014291	0.014124	0.014861	0.014835	0.015612	0.015546	0.016306	0.016257	0.017034	0.016968
90		0.012677	0.012649	0.013369	0.013317	0.014009	0.013985	0.014690	0.014653	0.015365	0.015321	0.016048	0.015990
100		0.012025	0.012002	0.012724	0.012634	0.013287	0.013266	0.013952	0.013898	0.014570	0.014276	0.015215	0.014845
150		0.009817	0.009804	0.010330	0.010316	0.010841	0.010827	0.011359	0.011339	0.011876	0.011680	0.012395	0.012150
200		0.008501	0.008493	0.008941	0.008933	0.009384	0.009374	0.009826	0.009814	0.010274	0.010128	0.010720	0.010537
250		0.007603	0.007597	0.007995	0.007990	0.008390	0.008382	0.008786	0.008775	0.009183	0.009066	0.009580	0.009433

We focus on the sample size enabling near-optimal treatment when $h = 0.2$ and $\varepsilon = 0.0085$.

Setting $h = 0.2$ supposes that suffering from lymphedema reduces welfare by 0.2. This value is suggested by Cheville *et al.* (*Cancer*, 2010), who elicited from a group of patients their perspectives on the matter.

Setting $\varepsilon = 0.085$ follows from the way that the MSLT-II investigators performed their power calculation. They judged a difference of 5 percentage points in survival to be a clinically meaningful loss in welfare and they judged 0.17 to be an acceptable probability of Type II error. Regret equals the magnitude of welfare loss times the probability that the loss will occur. Thus, the investigators judged $0.17 \times 0.05 = 0.0085$ to be an acceptable level of regret.

ε-optimal treatment is achievable with 244 patients assigned to each of observation and dissection. The total sample size of 488 is much smaller than the 1939 in MSLT-II.

Conclusion

Science does not always progress monotonically. There are times when important, even fundamental, ideas are discovered and receive attention but then are neglected.

This occurred with statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten thereafter.

A revival focusing on treatment choice began in the early 2000s and has been growing.

The growing dissatisfaction of statisticians with hypothesis testing is exemplified by the ASA Statement in Wasserstein and Lazar (2016). I hope this will encourage statisticians and econometricians to relearn statistical decision theory and use it when studying not only treatment choice but decision making with sample data more generally.