Module 12: An Olio of Topics in Applied Probabilistic Reasoning

To understand God's thoughts we must study statistics for these are the measure of His purpose.

– Florence Nightingale.

Abstract: The last module is a collection of topics in applied probabilistic reasoning that were all too small to command their own separate modules. Topics include: 1) the randomized response method as a way of asking sensitive questions and hopefully receiving truthful answers; 2) the use of surrogate end points (or proxies) in the study of some phenomenon where the connections to "real" outcomes of interest (for example, to mortality) are indirect and probabilistically linked (for example, to lowered cholesterol levels); 3) the comparison between a normative theory of choice and decision making derived from probability theory and actual human performance; 4) permutation tests and statistical inference derived directly from how a randomized controlled study was conducted. As an oddity that can occur for this type of statistical inference procedure, the famous 1954 Salk polio vaccine trials are discussed. Also, three brief subsections are given that summarize the jackknife, the bootstrap, and permutation tests involving correlational measures. This latter material is provided in an abbreviated form suitable for slide presentation in class, and where further explanatory detail would be given by an instructor.

Contents

1 The Randomized Response Method

As noted elsewhere, how questions are framed and the context in which they are asked are crucial for understanding the meaning of the given responses. This is true both in matters of opinion polling and for collecting data on, say, the health practices of subjects. In these situations, the questions asked are usually not sensitive, and when framed correctly, honest answers are expected. For more sensitive questions about illegal behavior, (reprehensible) personal habits, suspect health-related behaviors, questionable attitudes, and so on, asking a question outright may not garner a truthful answer.

The randomized response method is one mechanism for obtaining "accurate" data for a sensitive matter at a group level (but not

at the individual level). It was first proposed in 1965 by Stanley Warner in the *Journal of the American Statistical Association*, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias" (*60*, 63–69). A modified strategy was proposed by Bernard Greenberg and colleagues in 1969, again in *JASA*: "The Unrelated Question Randomized Response Model: Theoretical Framework" (*64*, 520–539). We first illustrate Warner's method and then Greenberg's with an example.

Let $\mathcal{Q}$ be the question: "Have you ever smoked pot (and inhaled)?"; and $\bar{\mathcal{Q}}$ the complement: "Have you never smoked pot (and inhaled)?" With some known probability, θ, the subject is asked $\mathcal{Q}$; and with probability $(1-\theta)$, is given $\bar{\mathcal{Q}}$ to answer. The respondent determines which question is posed by means of a probability mechanism under his or her control. For example, if the respondent rolls a single die and a 1 or 2 appears, question $\mathcal{Q}$ is given; if 3, 4, 5, or 6 occurs, $\bar{\mathcal{Q}}$ is given. So, in this case, $\theta = 1/3$.

As notation, let p be the proportion in the population for which the true response to $\mathcal{Q}$ is "yes"; $1-p$ is then the proportion giving a "yes" to $\bar{\mathcal{Q}}$. Letting P_{yes} denote the observed proportion of "yes" responses generally, its expected value is $\theta p + (1-\theta)(1-p)$; thus, p can be estimated as

$$\hat{p}_w = \frac{P_{yes} - (1-\theta)}{2\theta - 1} \ ,$$

where the subscript w is used to denote Warner's method of estimation. Obviously, θ cannot be $1/2$ because the denominator would then be zero; but all other values are legitimate. The extremes of θ

being 0 or 1, however, do not insure the "privacy" of a subject's response because the question actually answered would then be known.

The Greenberg method is referred to as the unrelated (or innocuous) question technique. The complement question $\bar{\mathcal{Q}}$ is replaced with an unrelated question, say, $\mathcal{Q}_U$, with a known probability of giving a "yes" response, say γ. For example, $\mathcal{Q}_U$ could be "Flip a coin. Did you get a head?" Here, $\gamma = 1/2$ for a "yes" response; the expected value of P_{yes} is $\theta p + (1-\theta)\gamma$, leading to

$$\hat{p}_g = \frac{P_{yes} - (1-\theta)\gamma}{\theta} \ ,$$

where the subscript g now refers to Greenberg's method of estimation.

To decide which strategy might be the better, the variances of the two estimates can be compared though closed-form formulas:

$$\mathrm{Var}(\hat{p}_w) = \frac{p(1-p)}{n} + \frac{\theta(1-\theta)}{n(2\theta-1)^2} \ ;$$

$$\mathrm{Var}(\hat{p}_g) =$$

$$\frac{p(1-p)}{n} + \frac{(1-\theta)^2\gamma(1-\gamma) + \theta(1-\theta)(p(1-\gamma)+\gamma(1-p))}{n\theta^2} \ ,$$

where the number of respondents is denoted by n. As an example, suppose θ is .6; the coin flip defines $\mathcal{Q}_U$ so γ is .5; and let the true proportion p be .3. Using the variance formulas above: $\mathrm{Var}(\hat{p}_w) = 6.21/n$ and $\mathrm{Var}(\hat{p}_g) = .654/n$. Here, the Greenberg "innocuous question" variance is only about a tenth of that for the Warner estimate, making the Greenberg method much more efficient

in this instance (that is, the sampling variance for the Greenberg estimate is much less than that for the Warner estimate).

The use of innocuous questions is the most common implementation of a randomized response method. This is likely due to the generally smaller variance for the Greenberg estimator compared to that for Warner; also, the possible confusion caused by using "ever" and "never" and responding "yes" and "no" in Warner's method is avoided by the use of an innocuous question. As a practical example of the unrelated question implementation of randomized response, several excerpts are presented below from a *New York Times* article by Tom Rohan (August 22, 2013), entitled "Antidoping Agency Delays Publication of Research":

> Doping experts have long known that drug tests catch only a tiny fraction of the athletes who use banned substances because athletes are constantly finding new drugs and techniques to evade detection. So in 2011, the World Anti-Doping Agency convened a team of researchers to try to determine more accurately how many athletes use performance-enhancing drugs.
>
> More than 2,000 track and field athletes participated in the study, and according to the findings, which were reviewed by *The New York Times*, an estimated 29 percent of the athletes at the 2011 world championships and 45 percent of the athletes at the 2011 Pan-Arab Games said in anonymous surveys that they had doped in the past year.
>
> ...
>
> The project began in 2011 when the researchers created a randomized-response survey, a common research technique that is used to ask sensitive questions while ensuring a subject's confidentiality. The researchers conducted their interviews at two major track and field events: the world championships in Daegu, South Korea, and the Pan-Arab Games in Doha, Qatar.
>
> Athletes at the events answered questions on tablet computers and were asked initially to think of a birthday, either their own or that of someone close

to them. Then, depending on the date of the birthday, they were instructed to answer one of two questions that appeared on the same screen: one asked if the birthday fell sometime between January and June, and the other asked, "Have you knowingly violated anti-doping regulations by using a prohibited substance or method in the past 12 months?"

The study was designed this way, the researchers said, so only the athlete knew which of the two questions he or she was answering. Then, using statistical analysis, the researchers could estimate how many of the athletes admitted to doping.

The researchers noted that not every athlete participated, and those who did could have lied on the questionnaire, or chosen to answer the birthday question. They concluded that their results, which found that nearly a third of the athletes at the world championships and nearly half at the Pan-Arab Games had doped in the past year, probably underestimated the reality.

2 Surrogate End Points and Proxies

The presentation of data is an obvious area of concern when developing the basics of statistical literacy. Some aspects may be obvious, such as not making up data or suppressing analyses or information that don't conform to prior expectations. At times, however, it is possible to contextualize (or to "frame") the same information in different ways that might lead to differing probabilistic interpretations. An earlier module on the (mis)reporting of data was devoted more extensively to the review of Gigerenzer et al. (2007), where the distinctions are made between survival and mortality rates, absolute versus relative risks, natural frequencies versus probabilities, among others. Generally, the presentation of information should be as honest, clear, and transparent as possible. One such example given by Gigerenzer et al. (2007) suggests the use of frequency statements in-

stead of single-event probabilities, thereby removing the ambiguity of the reference class: instead of saying "there is a 30% to 50% probability of developing sexual problems with Prozac," use "out of every 10 patients who take Prozac, 3 to 5 experience a sexual problem." Thus, a male taking Prozac won't expect that 30% to 50% of his personal sexual encounters will result in a "failure."

In presenting data to persuade, and because of the "lead-time bias" medical screening produces, it is ethically questionable to promote any kind of screening based on improved five-year survival rates, or to compare such survival rates across countries where screening practices vary. As a somewhat jaded view of our current health situation, we have physicians practicing defensive medicine because there are no legal consequences for overdiagnosis and overtreatment, but only for underdiagnosis. Or, as the editor of the *Lancet* commented (as quoted by R. Horton, *New York Review of Books*, March 11, 2004), "journals have devolved into information laundering operations for the pharmaceutical industry." The issues involved in medical screening and its associated consequences are psychologically important; for example, months after false positives for HIV, mammograms, or prostate cancer, considerable and possibly dysfunctional anxiety may still exist.

When data are presented to make a health-related point, it is common practice to give the argument in terms of a "surrogate endpoint." Instead of providing direct evidence based on a clinically desired outcome (for example, if you engage in this recommended behavior, the chance of dying from, say, a heart attack is reduced by such and such amount), the case is stated in terms of a proxy (for example,

if you engage in this recommended behavior, your cholesterol levels will be reduced). In general, a surrogate end point or biomarker is a measure of a certain treatment that may correlate with a real clinical endpoint, but the relationship is probabilistically determined and not guaranteed. This caution can be rephrased as "a correlate does not a surrogate make."

It is a common misconception that something correlated with the true clinical outcome must automatically then be usable as a valid surrogate end point and can act as a proxy replacement for the clinical outcome of primary interest. As is true for all correlational phenomena, causal extrapolation requires further argument. In this case, it is that the effect of the intervention on the surrogate directly predicts the clinical outcome. Obviously, this is a more demanding requirement.

Outside of the medical arena, proxies play prominently in the current climate-change debate. When actual surface temperatures are unavailable, surrogates for these are typically used (for example, tree-ring growth, coral accumulation, evidence in ice). Whether these are satisfactory stand-ins for the actual surface temperatures is questionable. Before automatically accepting a causal statement (for example, that greenhouse gases are wholly responsible for the apparent recent increase in earth temperature), pointed (statistical) questions should be raised, such as:

(a) why don't the tree-ring proxies show the effects of certain climate periods in our history—the Medieval Warm Period (circa 1200) and the Little Ice Age (circa 1600)?;

(b) over the last century or so, why has the tree-ring and surface temperature relationship been corrupted so that various graphical "tricks" need to be used to obtain the "hockey stick" graphic demonstrating the apparent catastrophic increase in earth temperature over the last century?;

(c) what effect do the various solar cycles that the sun goes through have on our climate; could these be an alternative mechanism for what we are seeing in climate change?;

(d) or, is it some random process and we are on the up-turn of something comparable to the Medieval Warm Period, with some later downturn expected into another Little Ice Age?

3 The Normative Theory of Probability and Human Decision Making

One important area of interest in developing statistical literacy skills and learning to reason probabilistically is the large body of work produced by psychologists. This work compares the normative theory of choice and decisions derivable from probability theory, and how this may not be the best guide to the actual reasoning processes individuals use. The contributions of Tversky and Kahneman (for example, 1971, 1974, 1981) are particularly germane to our understanding of reasoning. People rely on various simplifying heuristic principles to assess probabilities and engage in judgments under uncertainty. We give a classic Tversky and Kahneman (1983) illustration to show how various reasoning heuristics might operate:

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination

and social justice, and also participated in anti-nuclear demonstrations. Which ... [is] more probable?

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

Eighty-five percent of one group of subjects chose option 2, even though the conjunction of two events must be less likely than either of the constituent events. Tversky and Kahneman argue that this "conjunction fallacy" occurs because the "representativeness heuristic" is being used to make the judgment; the second option seems more representative of Linda based on the description given for her.

The representativeness heuristic operates where probabilities are evaluated by the degree to which A is representative of B; if highly representative, the probability that A originates from B is assessed to be higher. When representativeness heuristics are in operation, a number of related characteristics of the attendant reasoning processes become apparent: prior probabilities (base rates) are ignored; insensitivity develops to the operation of sample size on variability; an expectation that a sequence of events generated by some random process, even when the sequence is short, will still possess all the essential characteristics of the process itself. This leads to the "gambler's fallacy" (or, "the doctrine of the maturity of chances"), where certain events must be "due" to bring the string more in line with representativeness; as one should know, corrections are not made in a chance process but only diluted as the process unfolds. When a belief is present in the "law of small numbers," even small samples must be highly representative of the parent population; thus, researchers put too much faith in what is seen in small samples and overestimate replicability. Also, people may fail to recognize regression toward the

mean because predicted outcomes should be maximally representative of the input and therefore be exactly as extreme.

A second powerful reasoning heuristic is *availability*. We quote from Tversky and Kahneman (1974):

> Lifelong experience has taught us that, in general, instances of large classes are recalled better and faster than instances of less frequent classes; that likely occurrences are easier to imagine than unlikely ones; and that the associative connections between events are strengthened when the events frequently co-occur. As a result, man has at his disposal a procedure (the availability heuristic) for estimating the numerosity of a class, the likelihood of an event, or the frequency of co-occurrences, by the ease with which the relevant mental operations of retrieval, construction, or association can be performed. (p. 1128)

Because retrievability can be influenced by differential familiarity and saliences, the probability of an event may not be best estimated by the ease to which occurrences come to mind. A third reasoning heuristic is one of *anchoring and adjustment*, which may also be prone to various biasing effects. Here, estimates are made based on some initial value that is then adjusted (Tversky & Kahneman, 1974).

When required to reason about an individual's motives in some ethical context, it is prudent to remember the operation of the *fundamental attribution error*, where people presume that actions of others are indicative of the true ilk of a person, and not just that the situation compels the behavior. As one example from the courts, even when confessions are extracted that can be demonstrably shown false, there is still a greater likelihood of inferring guilt compared to the situation where a false confession was not heard. The classic

experiment on the fundamental attribution error is from Jones and Harris (1967); we quote a summary given in the Wikipedia article on the fundamental attribution error:

> Subjects read pro- and anti-Fidel Castro essays. Subjects were asked to rate the pro-Castro attitudes of the writers. When the subjects believed that the writers freely chose the positions they took (for or against Castro), they naturally rated the people who spoke in favor of Castro as having a more positive attitude toward Castro. However, contradicting Jones and Harris' initial hypothesis, when the subjects were told that the writer's positions were determined by a coin toss, they still rated writers who spoke in favor of Castro as having, on average, a more positive attitude towards Castro than those who spoke against him. In other words, the subjects were unable to see the influence of the situational constraints placed upon the writers; they could not refrain from attributing sincere belief to the writers.

A particulary egregious example of making the fundamental attribution error (and moreover, for nefarious political purposes), is Liz Cheney and her ad on the website "Keep America Safe" regarding those lawyers currently at the Justice Department who worked as advocates for "enemy combatants" at Guantanamo Bay, Cuba. We give an article that lays out the issues by Michael Stone of the *Portland Progressive Examiner* (March 5, 2010; "Toxic Politics: Liz Cheney's Keep America Safe 'Al Qaeda Seven' Ad"):

> Liz Cheney, daughter of former Vice President Dick Cheney and co-founder of the advocacy group "Keep America Safe," is taking heat for a controversial ad questioning the values of Justice Department lawyers who represented Guantanamo Bay detainees.
>
> Several top political appointees at the Justice Department previously worked as lawyers or advocates for 'enemy combatants' confined at Guantanamo Bay, Cuba. In their ad, Cheney's group derides the unidentified appointees as the 'Al Qaeda 7.' The ad implies the appointees share terrorist values.

Aside from questioning the values of these Justice Department lawyers, the ad is using fear and insinuations to smear both the Justice Department lawyers and the Obama administration.

Demonizing Department of Justice attorneys as terrorist sympathizers for their past legal work defending Gitmo detainees is wrong. The unfounded attacks are vicious, and reminiscent of McCarthyism.

Indeed, the ad itself puts into question Cheney's values, her patriotism, her loyalty. One thing is certain: her understanding of US history, the founding of our country, and the US Constitution, is left seriously wanting.

John Aloysius Farrell, writing in the Thomas Jefferson Street blog, for *US News and World Report*, explains:

There are reasons why the founding fathers ... in the Bill of Rights, strove to protect the rights of citizens arrested and put on trial by the government in amendments number 4, 5, 6, 7, and 8.

The founders had just fought a long and bloody revolution against King George, and knew well how tyrants like the British sovereign perpetuated power with arbitrary arrests, imprisonments, and executions. And so, along with guarantees like the right to due process, and protection from unreasonable searches and cruel and unusual punishment, the first patriots also included, in the Sixth Amendment, the right of an American to a speedy trial, by an impartial jury, with "the Assistance of Counsel for his defense."

John Adams regarded his defense of the British soldiers responsible for the Boston Massacre as one of the noblest acts of his life for good reason. Our adversarial system of justice depends upon suspects receiving a vigorous defense. That means all suspects must receive adequate legal counsel, including those accused of the most heinous crimes: murder, rape, child abuse and yes, even terrorism.

Defending a terrorist in court does not mean that one is a terrorist or shares terrorist values. Implying otherwise is despicable. Cheney's attacks are a dangerous politicization and polarization of the terrorism issue. Those who would honor our system of law and justice by defending suspected terrorists deserve our respect. Instead Cheney and her group smear these patriots in an attempt to score points against political enemies.

4 Permutation Tests and Statistical Inference

The aim of any well-designed experimental study is to make a causal claim, such as "the difference observed between two groups is *caused* by the different treatments administered." To make such a claim we need to know the counterfactual: what would have happened if this group had not received the treatment? This counterfactual is answered most credibly when subjects are assigned to the treatment and control groups at random. In this instance, there is no reason to believe that the group receiving the treatment condition would have reacted any differently (than the control condition) had it received the control condition. If there is no differential experimental mortality to obscure this initial randomness, one can even justify the analyses used by how the groups were formed (for example, by randomization tests, or their approximations defined by the usual analysis methods based on normal theory assumptions). As noted by R. A. Fisher (1971, p. 34), "the actual and physical conduct of an experiment must govern the statistical procedure of its interpretation." When the gold standard of inferring causality is not met, however, we are in the realm of quasi-experimentation, where causality must be approached differently.

An important benefit from designing an experiment with random assignment of subjects to conditions, possibly with blocking in various ways, is that the method of analysis through randomization tests is automatically provided. As might be expected, the original philosophy behind this approach is due to R. A. Fisher, but it also has been developed and generalized extensively by others (see Edgington & Onghena, 2007). In Fisher's time, and although randomization

methods may have been the preferred strategy, approximations were developed based on the usual normal theory assumptions to serve as computationally feasible alternatives. But with this view, our standard methods are just approximations to what the preferred analyses should be. A short quotation from Fisher's *The Design of Experiments* (1971) makes this point well (and one that expands on the short phrase given in the previous paragraph):

> In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. The arithmetical procedure of such an examination is tedious, and we shall only give the results of its application ... to show the possibility of an independent check on the more expeditious methods in common use. (p. 45)

A randomization (or permutation) test uses the given data to generate an exact null distribution for a chosen test statistic. The observed test statistic for the way the data actually arose is compared to this null distribution to obtain a p-value, defined as the probability (if the null distribution were true) of an observed test statistic being as or more extreme than what it actually was. Three situations lead to the most common randomization tests: K-dependent samples, K-independent samples, and correlation. When ranks are used instead of the original data, all of the common nonparametric tests arise. In practice, null randomization distributions are obtained either by complete enumeration, sampling (a Monte Carlo strategy), or through various kinds of large sample approximations (for example, normal or chi-squared distributions).

Permutation tests can be generalized beyond the usual correlational framework or that of K-dependent or K-independent samples. Much of this work falls under a rubric of combinatorial data analysis (CDA), where the concerns are generally with comparing various kinds of complete matrices (such as proximity or data matrices) using a variety of test statistics. The most comprehensive source for this material is Hubert (1987), but the basic matrix comparison strategies are available in a number of places, for example, see discussions of the "Mantel Test" in many packages in R (as one example, see the "Mantel–Hubert general spatial cross-product statistic" in the package, `spdep`). Even more generally, one can at times tailor a test statistic in nonstandard situations and then implement a permutation strategy for its evaluation through the principles developed in CDA.

The idea of repeatedly using the sample itself to evaluate a hypothesis or to generate an estimate of the precision of a statistic, can be placed within the broader category of resampling statistics or sample reuse. Such methods include the bootstrap, jackknife, randomization and permutation tests, and exact tests (for example, Fisher's exact test for 2×2 contingency tables). Given the incorporation of these techniques into conveniently available software, such as R, there are now many options for gauging the stability of the results of one's data analysis.

4.1 The Jackknife

An idea similar to the "hold-out-some(one)-at-a-time" is Tukey's Jackknife.

This was devised by Tukey to obtain a confidence interval on a parameter (and indirectly to reduce the bias of an estimator that is not already unbiased).

In Psychology, there is an early discussion of the Jackknife in the *Handbook of Social Psychology (Volume II)* (Lindzey and Aronson; 1968) by Mosteller and Tukey: Data Analysis — Including Statistics.

General approach for the Jackknife:

suppose I have n observations $X_1, \ldots, X_n$ and let θ be an unknown parameter of the population.

We have a way of estimating θ (by, say, $\hat{\theta}$) –

Group the n observations into t groups of m; thus, $n = tm$:

$\{X_1, \ldots, X_m\}, \ldots, \{X_{(t-1)m+1}, \ldots, X_{tm}\}$

Let $\hat{\theta}_{-0}$ be the estimate based on all groups;

Let $\hat{\theta}_{-i}$ be the estimate based on all groups except the i^{th}

Define new estimates of θ, called "pseudo-values" as follows:

$\hat{\theta}_{*i} = t\hat{\theta}_{-0} - (t-1)\hat{\theta}_{-i}$, for $i = 1, \ldots, t$

The Jackknife estimate of θ is the mean of the pseudo-values:

$\hat{\theta}_{*\cdot} = \sum_{i=1}^{t} \frac{\hat{\theta}_{*i}}{t}$

An estimate of its standard error is

$s_{\hat{\theta}_{*\cdot}} = [\sum_{i=1}^{t} \frac{(\hat{\theta}_{*i} - \hat{\theta}_{*\cdot})^2}{t(t-1)}]^{1/2}$

Approximate confidence interval:

$\hat{\theta}_{*\cdot} \pm s_{\hat{\theta}_{*\cdot}} t_{\frac{\alpha}{2},t-1}$

We act as if the t pseudo-values $\hat{\theta}_{*1}, \ldots, \hat{\theta}_{*t}$ are independent and identically distributed observations.

We also reduce some bias in estimation if the original estimate was biased.

An example:

suppose I want to estimate μ based on $X_1, \ldots, X_n$

Choose $t = n$

$\hat{\theta}_{-0} = \frac{1}{n} \sum_{j=1}^{n} X_j$

$\hat{\theta}_{-i} = \frac{1}{n-1} \sum_{j=1, i \neq j}^{n} X_j$

$\hat{\theta}_{*i} = n(\frac{1}{n} \sum_{j=1}^{n} X_j) - (n-1)(\frac{1}{n-1} \sum_{j=1, i \neq j}^{n} X_j) = X_i$

Thus, $\hat{\theta}_{*\cdot} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$

$s_{\hat{\theta}_{*\cdot}} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2} =$

$\sqrt{s_X^2/n}$, where s_X^2 is an unbiased estimate of σ^2

Confidence interval:

$\bar{X} \pm (\sqrt{s_X^2/n})\ t_{\frac{\alpha}{2},t-1}$

4.2 The Bootstrap

Population ("Theory World"): the pair of random variables X and Y are, say, bivariate normal

Sample ("Data World"): n pairs of independent and identically distributed observations on (X, Y):

$(X_1, Y_1), \ldots, (X_n, Y_n)$; these could be used to give r_{XY} as an estimate of ρ_{XY}

Now, make Data World the Theory World Population:

$(X_1, Y_1), \ldots, (X_n, Y_n)$, and each occurs with probability $\frac{1}{n}$

Sample this Theory World Population (with replacement) to get one "bootstrap" sample (with possible repeats):

$(X_1', Y_1'), \ldots, (X_{n'}', Y_{n'}')$ (usually, n equals n')

Get B bootstrap samples and compute the correlation for each: $r_{XY}^{(1)}, \ldots, r_{XY}^{(B)}$

This last distribution could be used, for example, to obtain a confidence interval on ρ_{XY}

4.2.1 Permutation tests for correlation measures

We start at the same place as for the Bootstrap:

Population ("Theory World"): the pair of random variables X and Y are, say, bivariate normal

Sample ("Data World"): n pairs of independent and identically distributed observations on (X, Y):

$(X_1, Y_1), \ldots, (X_n, Y_n)$; these could be used to give r_{XY} as an estimate of ρ_{XY}

Now, to test $H_o : X$ and Y are statistically independent.

Under H_o, the X's and Y's are matched at random; so, assuming (without loss of generality) that we fix the X's, all $n!$ permutations of the Y's against the X's are equally likely to occur.

We can calculate a correlation for each of these $n!$ permutations and graph:

the distribution is symmetric and unimodal at zero; the range along the horizontal axis obviously goes from -1 to $+1$

p-value (one-tailed) = number of correlations as or larger than the observed correlation/$n!$

Also, as an approximation, $r_{XY} \sim N(0, \frac{1}{n-1})$;

Thus, the standard error is close to $\frac{1}{\sqrt{n}}$; this might be useful for quick "back-of-the-envelope" calculations

4.3 An Introductory Oddity: The 1954 Salk Polio Vaccine Trials

The 1954 Salk polio vaccine trials was the biggest public health experiment ever conducted. One field trial, labeled an observed control experiment, was carried out by the National Foundation for Infantile Paralysis. It involved the vaccination, with parental consent, of second graders at selected schools in selected parts of the country. A control group would be the first and third graders at these same

schools, and indirectly those second graders for whom parental consent was not obtained. The rates for polio contraction (per 100,000) are given below for the three groups (see Francis et al., 1955, for the definitive report on the Salk vaccine trials).[1]

Grade 2 (Vaccine): 25/100,000;
Grade 2 (No consent): 44/100,000;
Grades 1 and 3 (Controls): 54/100,000.

The interesting observation we will return to below is that the Grade 2 (No consent) group is between the other two in the probability of polio contraction. Counterintuitively, the refusal to give consent seems to be partially protective.

The second field trial was a (double-blind) randomized controlled experiment. A sample of children were chosen, all of whose parents consented to vaccination. The sample was randomly divided into two, with half receiving the Salk vaccine and the other half a placebo of inert salt water. There is a third group formed from those children with no parental consent and who therefore were not vaccinated. We give the rates of polio contraction (per 100,000) for the three groups:

Vaccinated: 28/100,000;
Control: 71/100,000;
No consent: 46/100,000.

Again, not giving consent appears to confer some type of immunity;

[1]The interpretation of results and the source of the information given in this section, *An Evaluation of the 1954 Poliomyelitis Vaccine Trials*, is by Thomas Francis, Robert Korns, and colleagues (1955) (in particular, see Table 2b: Summary of Study of Cases by Diagnostic Class and Vaccination Status; p. 35).

the probability for contracting polio for the "no consent" group is between the other two.

The seeming oddity in the ordering of probabilities, where "no consent" seems to confer some advantage, is commonly explained by two "facts": (a) children from higher-income families are more vulnerable to polio; children raised in less hygienic surroundings tend to contract mild polio and immunity early in childhood while still under protection from their mother's antibodies; (b) parental consent to vaccination appears to increase as a function of education and income, where the better-off parents are much more likely to give consent. The "no consent" groups appear to have more natural immunity to polio than children from the better-off families. This may be one of the only situations we know of where children growing up in more resource-constrained contexts are conferred some type of advantage.

References

[1] Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). New York: Chapman & Hall / CRC.

[2] Fisher, R. A. (1971). *The design of experiments* (9th ed.). New York: Hafner.

[3] Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

[4] Hubert, L. J. (1987). *Assignment methods in combinatorial data analysis*. New York: Marcel Dekker.

[5] Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*, 1–24.

[6] Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

[7] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

[8] Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

[9] Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.