THE STRENGTH OF EVIDENCE VERSUS THE POWER OF BELIEF: ARE WE ALL BAYESIANS?

Jessica Utts[1], Michelle Norris[2], Eric Suess[3] and Wesley Johnson[1]
[1]Department of Statistics, University of California, Irvine, CA, USA
[2]Department of Mathematics and Statistics, California State University Sacramento, USA
[3]Department of Statistics and Biostatistics, California State University East Bay, USA
jutts@uci.edu

Although statisticians have the job of making conclusions based on data, for many questions in science and society prior beliefs are strong and may take precedence over data when people make decisions. For other questions, there are experts who could shed light on the situation that may not be captured with available data. One of the appealing aspects of Bayesian statistics is that the methods allow prior beliefs and expert knowledge to be incorporated into the analysis along with the data. One domain where beliefs are almost sure to have a role is in the evaluation of scientific data for extrasensory perception (ESP). Experiments to test ESP often are binomial, and they have a clear null hypothesis, so they are an excellent way to illustrate hypothesis testing. Incorporating beliefs makes them an excellent example for the use of Bayesian analysis as well. In this paper, data from one type of ESP study are analyzed using both frequentist and Bayesian methods.

INTRODUCTION

Teachers of Statistics often look for interesting yet simple examples. This paper focuses on a question that has fascinated humankind for centuries, for which there is still no definitive answer, and for which there is now enough data to make an intriguing statistical study. Further, it is a topic that naturally lends itself to frequentist testing and estimation, as well as to Bayesian inference. Many different aspects can be used as classroom examples.

In 1986 I was offered an opportunity to become the statistical consultant to a classified program in which the US government was investigating the use of psychics to gain intelligence information. I continued work on that program for a decade, including a year's leave of absence from my university position to work on the project full-time. That led to other opportunities to work with parapsychologists, scientists who use laboratory studies to investigate whether psychic functioning is possible, and if so, how it might work.

The parapsychology research community has developed a growing body of evidence that it might be possible to obtain information in ways that can't be explained through our normal senses. But I have noticed that many people have a hard time being objective when examining the results of these studies. People with a prior belief that psychic abilities are real tend to overlook possible alternative explanations, while people with a prior belief that psychic abilities can't possibly be real tend to propose and believe very unlikely alternative explanations for the data. Bayesian methods allow these prior beliefs to be quantified and combined with the data to make it more transparent that one's assessment of the evidence is influenced by more than just a frequentist analysis of the data. Thus, these studies provide an excellent yet simple example of useful Bayesian analyses.

TWO REASONS TO BE A BAYESIAN

In the same year that I began working with the scientific investigation of psychic abilities, Brad Efron published a paper in *The American Statistician* with the provocative title "Why isn't everyone a Bayesian?" (Efron, 1986). The question could equally well be asked today. Although Bayesian methods are becoming more prominent across many subject-matter disciplines, it is still true today that the vast majority of research studies that use statistical analyses report frequentist results. So another question might be, "Why should anyone be a Bayesian?"

I see two reasons why Bayesian methods are appealing–one is philosophical, and the other one is practical. The philosophical reason has to do with how we interpret probability. One possible interpretation is that the probability of an event is an objective, physical quantity based on relative frequency. Suppose a woman is about to become pregnant. Using relative frequency with data from the past, we can say that the probability that her baby will be a boy is about 0.512. But now suppose she is pregnant already. The baby's sex is determined. Now what do we mean when we say the probability that the baby will be a boy is .512? Now we are expressing our belief about

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

something that has already happened. Let's suppose that the woman's doctor has conducted a test and knows the sex of the baby, but the woman does not. Now there are two different probabilities associated with the outcome that the baby will be a boy. The doctor's probability is either 0 or 1, but the woman's probability is still .512. Each of them is now expressing the degree to which they believe in the "boy" outcome. Bayesian methods utilize this "degree of belief" interpretation of probability to model all uncertainty, whereas the relative frequency interpretation of probability is limited to repeatable events.

Another philosophical reason to be a Bayesian is that the results of statistical analyses have a more intuitively appealing interpretation. A *p*-value incorporates probabilities of events that didn't happen. Why does that make sense? Further, *p*-values are highly dependent on sample size, yet most practitioners of statistics don't understand that. And to make matters even more complicated, very few people understand the convoluted interpretation of a *p*-value. Many students think it is the probability that the null hypothesis is true. Even seasoned journalists often express it in as "the probability that the results occurred by chance." Bayesian results, on the other hand, answer questions that make more sense to ask: What probabilities did we assign to various possible parameter values before we had data, and what probabilities do we assign to them after looking at data?

The other reason to be Bayesian is a practical one. It is rare that we have no information about a situation before we collect data. If we want to estimate the proportion of a community that is infected with HIV, do we really believe it is equally likely to be anything from 0 to 1? If we want to estimate the mean change in blood pressure after 10 weeks of meditation, do we really believe it could be anything from $-\infty$ to ∞? Even the choice of what hypotheses to test, and whether to make them one-sided or two-sided is an illustration of using prior knowledge. Bayesian methods allow us to quantify that knowledge, and make it explicit. And if there are experts to consult, the situation is even better because we can incorporate their knowledge into our analysis. Their knowledge was probably obtained from some form of past data or theoretical understanding, so it is not generally as "subjective" as critics of Bayesian analysis might have us believe.

So why isn't everyone a Bayesian? Two immediate reasons come to mind. First, most of us weren't taught to do things that way (or to think that way), and second, before modern computing tools became available it was difficult to do the computations necessary for Bayesian analyses. Perhaps another reason is one that Efron noted in his paper, "Bayesian theory requires a great deal of thought about the situation to apply sensibly" (Efron, 1986, p. 1). But none of these are good reasons now to avoid using Bayesian methods. Computational tools are readily available, as are examples of how to apply Bayesian methods sensibly to a variety of situations. See Christensen et al (2010) for many examples.

STUDIES IN PARAPSYCHOLOGY

There are several interesting kinds of studies being done in parapsychology, only one of which will be analyzed in this paper. A few categorizations will help clarify things. First, there are multiple alleged psychic abilities, characterized by the purported outcome and/or source of information. *Psychokinesis* (PK) refers to the ability to change the physical world without apparent physical force. It is further divided into *micro* and *macro* PK, with the former referring to the manipulation of quantum systems and the latter referring to manipulation of non-quantum (macro) systems. Macro PK is very difficult to study under controlled conditions, and is susceptible to all sorts of magicians' tricks. Micro PK can be investigated under well-controlled conditions, and a meta-analysis of studies of micro-PK found a weak effect (Bösch et al., 2006), but there are several complex issues related to those studies that would require an extensive discussion to sort out, and they will not be discussed further here.

The term *extrasensory perception* (ESP) refers to the ability to acquire information without using the five known senses. ESP is further divided by the supposed source of the information. With *telepathy* the information is thought to come from another individual, with *clairvoyance* it is thought to come from another place but not from a person, and with *precognition* the information is thought to come from the future. In *proof-oriented* research the alleged source of information isn't so important, because the emphasis is on testing whether or not the null hypothesis of "no ESP" can be rejected. In more recent years, emphasis has moved in the direction of *process-oriented* research, in which the goal is to identify the underlying mechanism, and for which the alleged source is important.

A general term used to describe both PK and ESP abilities is *psi*, from the letter that begins the Greek word ψυχή, loosely translated as psyche, soul, or spirit.

Early work investigating ESP was done using *forced-choice* experiments, similar to a multiple-choice test. In the last several decades these have largely been replaced by *free-response* experiments, in which the "target" is a location, photograph, short segment from a movie, or some other material with a very large number of possibilities. The two most common types of free-response experiments are *remote viewing* studies, and *ganzfeld* studies. It is the latter type that is the focus on the analyses in this paper, because they constitute a body of work that has been scrutinized by skeptics and found to be methodologically sound.

Ganzfeld Studies

Based on clues from previous research, parapsychologists speculated that psi performance might be enhanced if participants were in a relaxed state and deprived of other explicit sensory input. Thus, they began studies using a technique that had been previously developed in psychology, called the *ganzfeld procedure*. Ganzfeld is German for "whole field;" the reason for this name will become apparent.

In most ganzfeld studies there is a "sender" and a "receiver." They are placed in separate acoustically-shielded rooms and the session begins by having them each listen to a relaxation tape. For the next 15 to 30 minutes, the sender looks at a "target" image on a television screen (which may be a static photograph or a short movie segment playing repeatedly) and attempts to transmit information about the target to the receiver. The receiver is seated in a reclining chair, with headphones playing white noise and a "whole field" of red light shining in his/her eyes. The light is diffused with halved ping-pong balls placed over the receiver's eyes. The theory is that the ears and eyes are receiving input, but with no sense or pattern, so the mind, looking for input, hopefully will receive the information being sent by the sender. During this time period, the receiver provides an ongoing verbal report, picked up by a microphone, of his or her thoughts, images and feelings. This is called the "mentation" period.

At the end of the mentation period the receiver is shown four possible choices of targets, all of the same type (static photograph or movie segment). One is the correct target, and the other three are "decoys." The receiver must choose the one s/he thinks best matches the description given during the mentation period. An experimenter, who does not know the correct answer, but who has listened during the mentation period, will provide information about what was said if requested by the receiver.

If the correct target is chosen by the receiver, the session is a "hit." Otherwise, it is a miss. In some studies the receiver provides a rank-order from 1 to 4 (most to least likely) and/or a rating on a scale from 1 to 100 for each possible choice, but the analysis for any of these methods can be reduced to a simple hit or miss, and that is what will be used in this paper.

Before a study begins, a "target pool" is assembled, usually consisting of hundreds of potential targets (photographs and/or video segments). The pool is then organized into sets of four targets that are of the same type (photo or video) but that otherwise are as dissimilar as possible. For each session, one of the sets of four is randomly selected, and then one of those is randomly selected to be the actual target. The remaining three are the decoys used for judging. Therefore, the target is randomly selected from a larger pool in such a way that the probability that it will best match what the receiver said is .25 by chance alone, as it is for each decoy.

Studies Included in Our Analyses

There have been a few meta-analyses of ganzfeld studies published over the years, and we included all of the studies from those that met criteria for methodological rigor and adherence to standard ganzfeld procedures. These included 16 of the studies in Table 2 of Dawson (1991), with eight eliminated because of unresolved allegations of methodological flaws, all 11 studies in Table 1 of Bem and Honorton (1994) and the 29 studies with a "standardness score" of more than 4 in Table 1 of Bem, Palmer and Broughton (2001). The standardness score is a measure of how closely the procedures used in the study conform to the standard ganzfeld protocol. For instance, a study that used music as targets instead of photos or videos would receive a low standardness score. An updated meta-analysis of ganzfeld studies is in press (Storm et al, 2010) but details of the most recent studies included in that investigation were not available at the time of our analysis. Studies included in our analysis are listed in Table 1, with number of sessions, number of hits and hit rate.

Table 1. Number of Sessions (n), Hits (x) and Hit Rates ($\hat{p}$) for the 56 Studies in Our Analyses

ID	n	x	$\hat{p}$	ID	n	x	$\hat{p}$	ID	n	x	$\hat{p}$	ID	n	x	$\hat{p}$
1	32	14	.44	15	60	27	.45	29	50	12	.24	43	30	14	.47
2	7	6	.86	16	48	10	.21	30	50	12	.24	44	30	11	.37
3	30	13	.43	17	22	8	.36	31	50	9	.18	45	97	32	.33
4	30	7	.23	18	9	3	.33	32	51	19	.37	46	22	2	.09
5	20	2	.10	19	35	10	.29	33	29	12	.41	47	50	13	.26
6	10	9	.90	20	50	12	.24	34	128	60	.47	48	32	8	.25
7	10	4	.40	21	50	18	.36	35	32	13	.41	49	58	11	.19
8	28	8	.29	22	50	15	.30	36	50	11	.22	50	46	12	.26
9	10	4	.40	23	36	12	.33	37	8	3	.38	51	20	6	.30
10	20	7	.35	24	20	10	.50	38	40	8	.20	52	30	6	.20
11	20	12	.60	25	7	3	.43	39	65	24	.37	53	42	5	.12
12	100	41	.41	26	50	15	.30	40	50	18	.36	54	32	14	.44
13	40	13	.33	27	25	16	.64	41	30	11	.37	55	40	16	.40
14	27	11	.41	28	50	13	.26	42	30	11	.37	56	36	13	.36

The studies we included were conducted by many different investigators and at a variety of laboratories in multiple countries. There is unlikely to be a large "file drawer" of unpublished studies because the ganzfeld procedure requires a special laboratory, parapsychology is a small field in which most researchers know each other, there are a limited number of journals in which such results would be published, and the journals in parapsychology have a policy of publishing studies even if they produce non-significant results.

FREQUENTIST ANALYSIS

Because of the random selection of target material and the way the judging is done, the results of the ganzfeld experiments can be modeled using a binomial distribution, with n = number of sessions and X = number of hits. Under the null hypothesis of "no psi" the probability of a hit for each session is $p = .25$. Because the studies were all done using the same procedure, we can combine the results to conduct a test and get a confidence interval.

Combining the data from the studies in Table 1 results in a total of $n = 2124$ sessions and $X = 709$ hits. This means that 33.4% of the sessions resulted in hits, with a 95% confidence interval ranging from 31.4% to 35.4%. A test of the null hypothesis that $p = .25$ results in $z = 8.92$, and an exact (one-tailed) binomial p-value of 2.26×10^{-18}.

We can also do separate tests and confidence intervals, but in most cases the sample size is so small that the test will not have adequate power. If in fact the true probability of a hit is .334, the sample proportion observed for these studies, then even for the largest study in the table ($n = 128$) the power for a one-sided test with $\alpha = .05$ is only .69. The power for a test with the median sample size for the studies ($n = 32$) is a meager .308. This may be one reason that critics of parapsychology think that there are very few "successful" studies and that there is a "replication" problem.

A SIMPLE BAYESIAN ANALYSIS

There are many questions addressed by science for which belief seems to play a major role in how many people view the evidence. The possible existence of psi phenomena is one of those questions. Thus, it makes sense to address the question through a Bayesian analysis that combines prior belief with experimental results.

The simplest Bayesian approach for binomial data is to use a Beta distribution to model one's prior beliefs about the probability of success, p. Combining the data with the prior produces a Beta posterior distribution for p. Fortunately there is free software that makes it very easy to model one's prior beliefs with the appropriate Beta distribution. It's called "BetaBuster" and is available for free to download at the website http://www.epi.ucdavis.edu/diagnostictests/betabuster.html.

To use BetaBuster, the prior distribution is elicited by asking two simple questions:

- In your opinion, what is the most likely value for *p*?
- In your opinion, you are 95% certain that *p* is below what value? (Or, if the answer to the first question is above .5, you are 95% certain that *p* is above what value?)

The answer to these two questions provides enough information to specify the two shape parameters for the Beta distribution, *a* and *b*. Once these are determined, the posterior distribution for *p* is also Beta, with shape parameters $X + a$ and $n – X + b$.

Let's see what happens with three sets of prior beliefs. For the *Skeptics prior*, we specify the most likely value for *p* as 0.25, and the skeptic is 95% certain that *p* is below 0.255. For the *Open-minded prior*, the most likely value of *p* is still 0.25, but the open-minded person is 95% certain that *p* is below 0.30. For the *Believer's prior*, the most likely value for *p* is 0.33, and the believer is 95% certain that *p* is below 0.36.

Figure 1 shows the prior and posterior distributions for *p* for the three priors applied to all of the data in Table 1, as well as for the open-minded prior applied to a single study with $n = 50$ and $X = 18$, for a sample proportion of 0.36 (as in Studies 21 and 40). Notice that the skeptic's opinion is not swayed much by the data, even with a sample as large as 2124 sessions. The observed data are consistent with the believer's initial beliefs, so the believer's posterior is not shifted much from the prior, but does have less variability. The open-minded person benefits the most from observing the data. With only one study with $n = .50$, the posterior distribution is slightly higher than the prior distribution. With $n = 2124$, the posterior distribution is much different from the prior distribution.

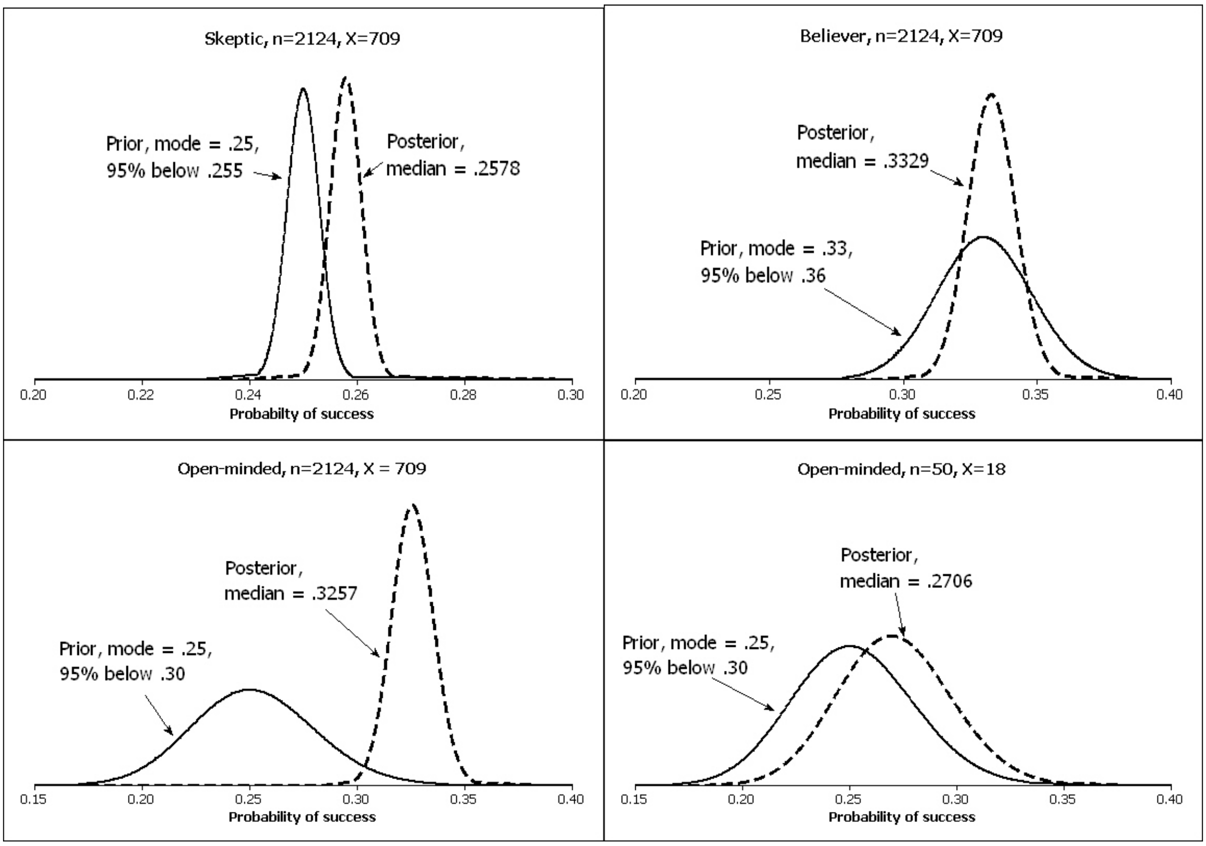

Figure 1. Skeptic, Believer and Open-Minded Prior and Posterior Beta Distributions for p

These results shed light on why skeptics of parapsychology are not convinced by the substantial accumulation of data showing positive results. By putting a prior distribution on the probability of success, one's prior beliefs can be made explicit, leading skeptics and believers to a rational reason for "agreeing to disagree" on whether or not psi abilities are real.

MODELING STUDY DIFFERENCES WITH A HIERARCHICAL BAYESIAN MODEL

Using the binomial model for the combined data requires us to assume that the probability of a hit remains constant across studies. If psi abilities are real, that may not be a valid assumption. For instance, it is well-known in psychology that there can be an "experimenter effect," in which some experimenters are able to put participants more at ease and thus obtain more favorable results than others. Another reason for differences may be that the participant pool differs. For instance for two of the studies in Table 1 (#24 and #34) the participants were creative artists, dancers and musicians, and those studies obtained very successful results. In our hierarchical Bayesian analysis we will allow for the possibility of different true probabilities of success for each study.

Hierarchical Model

We use a Bayesian hierarchical model that assumes a constant probability of a hit *within* a study, but the possibility of different probabilities *across* studies. Let p_i be the probability of a hit for Study i, $i = 1, 2, \ldots, 56$. As before, n_i and X_i are the number of sessions and number of hits, respectively, for Study i. We assume X_i has a binomial (n_i, p_i) distribution.

Now that we are allowing studies to have different probabilities of success, our goal is no longer one of simple estimation and testing of a single parameter. Instead, we will estimate the median and variability of the distribution of possible p_i's for all potential ganzfeld studies, not just the ones for which we have data. We use the variance-stabilizing arcsin transformation for proportions, then transform the results back to the original units when finished.

Define the observed hit rate in Study i to be $\hat{p}_i$. Then the arcsin transformed population and sample proportions are:

$$\theta_i = \arcsin\sqrt{p_i}, \quad y_i = \arcsin\sqrt{\hat{p}_i}, \quad i = 1,\ldots,56.$$

For the next layer of the hierarchical model, we assume that the θ_i are independent observations from a normal distribution, $N(\mu, \sigma^2)$. Note that μ is the median (as well as the mean) of the distribution of θ_i*'s*, and therefore we can do a simple transformation to get the median of the p's; median(p) = $\sin^2(\mu)$. We would like to estimate the median(p), which is the value such that half of the possible studies would have the probability of success at or below it, and half would have success probability at or above it.

The variance of the distribution of θ_i*'s* is of interest as well. A small value of σ^2 indicates that the true probabilities of success are similar across all possible studies, whereas a large value indicates that there is substantial variability.

The final layer of the hierarchical model is to put a prior distribution on the median of the possible values of p_i, which is equivalent to putting a prior distribution on μ, the mean (median) of the θ_i's. This is the step that allows one's prior beliefs to be incorporated into the analysis.

Prior Beliefs and Hierarchical Model Results

We allowed for different prior beliefs by using four different possible prior distributions for the median of the distribution of p_i's, the possible hit probabilities across all studies. We considered individuals who have no prior idea at all, those who aren't sure, but who would put the middle of their guesses at the chance hit rate of 0.25 with a somewhat wide range of options surrounding that (similar to our open-minded prior in the simpler analysis), those who believe psi is possible, and those who are fairly certain that psi is not possible. Of course if someone is completely convinced that psi is not possible, the entire prior would be a point mass at 0.25, and no amount of data would change that person's belief. Details of the priors we used for this analysis are as follows:

- Non-informative prior for μ puts equal probability on all real numbers (improper). This is probably not realistic; the true probability of a hit is unlikely to be as low as 0 or as high as 1.
- Open-minded prior, which uses median(p) = 0.25 as the best guess, and for which we are 90% sure the median(p) is between 0.12 and 0.41.

- Psi believer's prior, which uses median(p) = 0.33 and for which we are 90% sure the median(p) is between 0.30 and 0.36.
- Psi skeptics prior, which uses median(p) = 0.25 and for which we are 90% sure the median(p) is between 0.245 and 0.255.

We used the software R to carry out the analysis. Table 2 gives a summary of some of the results. Columns 3 to 5 of the table apply to the posterior distribution of median(*p*), the median of the distribution of possible success probabilities across all hypothetical ganzfeld studies that might ever be conducted. For each prior, the table shows the median of that posterior distribution and a 95% Bayesian interval, which represents the lower and upper 2.5% cutoffs for the posterior distribution of the median(*p*). In addition to the Bayesian results, we used maximum likelihood estimation for the more complex model allowing for different probabilities of a hit in each study. The MLE for the median, and the lower and upper end points of a 95% confidence interval are listed in columns 3 to 5 in the "frequentist" row of the table.

The last column of the table provides a way to assess the range of possibilities for *p*, rather than just the median. It gives the medians of the posterior distributions for the 2.5% and 97.5% cutoffs for the distribution of possible values for *p*. Notice that the range of possibilities for the skeptic has shifted slightly to above 0.25, but remains a very narrow range. For the non-informative and open-minded priors and the frequentist analysis, a wide range of possible values of *p* are likely. This result is in contrast to the frequentist confidence interval when we assumed a single value of *p*, which ranged from only .314 to .354. Therefore, the assumption that all studies have the same probability of success is probably not valid. Many more results are given in Utts et al, 2010.

Table 2. Median and 95% Intervals for the Posterior Distribution of Median(*p*); 95% Range for *p*

Type of Prior:	Prior median(p); 90% sure	2.5%	50%	97.5%	95% Range
Non-informative	N/A	0.30	0.33	0.36	0.19, 0.49
Open-minded	0.25; 0.12 and 0.41	0.29	0.33	0.36	0.17, 0.51
Psi believer	0.33; 0.30 and 0.36	0.308	0.326	0.345	0.281, 0.374
Psi skeptic	0.25; 0.245 and 0.255	0.251	0.257	0.262	0.254, 0.260
Frequentist	N/A	0.31	MLE = 0.33	0.36	MLE: 18, .50

STUDENT ACTIVITIES TO TEST PSYCHIC ABILITIES

Methodologically sound tests of psychic abilities must be designed so that ordinary means of communication are ruled out, and so that the probabilities of various results by chance alone are known. It is not easy to design in-class student activities that meet these criteria, but it is still possible to use tests of psi in class to illustrate the various concepts covered in this paper. Further, there are online tests that students can use to obtain their own data; see http://www.gotpsi.org for some examples. There are also applications students can download for their iPods and similar devices; for example see http://www.espresearch.com/iphone.

Of course it is not possible to do a ganzfeld experiment in class, but it is possible to conduct a free-response ESP test commonly called "remote viewing." Students are told that they will be shown a photograph or object, and are asked to describe what they think it will be using words and drawings.

One problem with conducting a group free-response test in class is what parapsychologists call the "stacking effect." This is simply the recognition that responses are not independent. For instance, if the target photograph is more visually appealing than the decoys, or is something people are likely to describe when they have no information about the target, then they are more likely to get a hit. Similarly, if one of the decoys is more appealing, it is more likely to be selected. The randomness in these experiments is in the choice of the target, not in the response. Let's suppose, as an extreme, that when someone is asked to describe a hidden target, they always describe an ocean and beach. In each packet of four targets, one of them will be the one that best matches an ocean and beach (even if none match it well). Using proper random selection, the one that does match best will be the correct target with probability 0.25. So, by chance alone, the response and correct target will match $1/4^{th}$ of the time.

Because of the stacking effect, if a free response experiment is done in class, it must be done so that each student has a different target. If all students have access to computers and the internet, the online tests mentioned above can be used. Another way to do this is to have four choices of target, but have students randomize which one is the correct answer for them. You can do this, for instance, by giving each student a card from a poker deck, and having the suit of the card designate which target is the right answer for that student. Of course they cannot look at the card until after they have chosen which of the four targets they think they were trying to describe.

Student projects that are conducted outside of class are more amenable to tests of psi because students can work with participants individually. I taught an honors seminar several times at University of California, Davis, in which students worked in teams of four to carry out a psi experiment of their own design over several weeks. One of these was a study of possible psi in dreams, and led to a publication for the students (Dalton et al., 2000).

CONCLUSION

Tests of psychic abilities can be an interesting and entertaining way to teach many different concepts at varying levels of statistics courses. For frequentist analyses, there is a clear point null hypothesis, unlike in most simple situations of testing one parameter. Psi studies are an excellent way to illustrate the concept of power. For instance, although the combined data in Table 1 led to rejecting the null hypothesis with a *p*-value close to 0, the individual studies have sample sizes so small that many of them have results that are not statistically significantly different from chance. This fact can be used to emphasize why one should never *accept* the null hypothesis, especially without examining power first.

Psi studies can be used to illustrate the concept of independence by explaining the stacking effect discussed in the previous section. For more advanced courses, meta-analysis can be illustrated by using psi studies to compare hit rates, assess methodological quality of different studies, and determine what inclusion and exclusion criteria should be used. See Bem and Honorton (1994), Bem et al. (2001) and Storm et al. (2010), for meta-analyses of ganzfeld studies with varying degrees of complexity.

Psi studies provide a natural context for Bayesian analysis, in which prior beliefs can easily be incorporated. They also provide a simple example of eliciting a prior that makes sense and is easy to find, in a context for which it is then easy to find the appropriate posterior distribution. More complex Bayesian hierarchical models can be illustrated as well, in higher level courses, as shown in this paper.

REFERENCES

Bem, D. J., & Honorton, C. (1994). Does Psi exist? Replicable Evidence for an Anomalous Process of Information Transfer. *Psychological Bulletin, 115*(1), 4-18.

Bem, D. J., Palmer, J., & Broughton, R. S. (2001). Updating the Ganzfeld Database: A Victim of Its Own Success? *Journal of Parapsychology, 65,* 207-218.

Bösch, H., Steinkamp, F,. & Boller, E. (2006). Examining Psychokinesis: The Interaction of Human Intention with Random Number Generators--A Meta-Analysis. *Psychological Bulletin. 132*(4), 497-523.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T.E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians.* Boca Raton, FL: CRC Press.

Dalton, K., Utts, J., Novotny, G., Sickafoose, L., Burrone, J., & Phillips, C. (2000). Dream GESP and Consensus Vote: A Replication. In *Proceedings of the 43rd Annual Convention of the Parapsychological Association,* (pp. 74-85).

Dawson, R. (1991). Comment on “Replication and Meta-Analysis in Parapsychology” by J. Utts, *Statistical Science, 6*(4), 382-385.

Efron, B. (1986). Why Isn’t Everyone a Bayesian? *The American Statistician, 40*(1), 1-5.

Storm, L., Tressoldi, P. E., & DiRisio, L. (2010). Meta-Analysis of Free-Response Studies, 1992-2008: Assessing the Noise Reduction Model in Parapsychology. *Psychological Bulletin* (In press).

Utts, J., Norris, M., Suess, E., & Johnson, W. (2010). A Hierarchical Model for Ganzfeld Studies. Unpublished manuscript.