Are We All Bayesians? Data versus Belief in Evaluating Studies of Extra Sensory Perception

Jessica Utts

Department of Statistics

University of California, Irvine

http://www.ics.uci.edu/~jutts

jutts@uci.edu

CSU Fullerton, October 6, 2011

Collaborators for Bayesian Part

- Michelle Norris
  - Dept of Math and Stat, California State University, Sacramento
- Eric Suess
  - Dept of Stat and Biostat, California State University, East Bay
- Wesley Johnson
  - Dept of Statistics, UCI

Overview of why to use Bayes

- Statisticians have the job of making conclusions based on data, but for many questions prior beliefs are strong and may take precedence over data when people make decisions.
- Bayesian methods allow prior beliefs and expert knowledge to be incorporated into analysis along with data.
- One domain where beliefs are almost sure to play a role is in the evaluation of scientific data for extrasensory perception.
- Experiments to test ESP often are binomial, and they have a clear null hypothesis (psychic abilities are not real), so they are an excellent way to illustrate hypothesis testing.

Outline of Talk

- Introduction and Background
  - On my involvement with research in extrasensory perception (ESP)
  - On reasons to be a Bayesian in general
- How research in ESP (“Parapsychology”) is done
- Frequentist analysis of ESP data
- Simple Bayesian analysis
- More complicated Bayesian analysis
- Conclusions

Why This Topic? Some Background

- My involvement started in 1986 as consultant to classified US government program testing psychic abilities for spying
- Continued to consult with parapsychology researchers through the years
- Noticed that many people (on both sides) ignore data and base conclusions on belief
- Makes this topic a natural for Bayesian statistics
- Also an excellent example for hypothesis testing because there is a clear null hypothesis

Why Be a Bayesian? Reason 1: Philosophical

- Interpretation of probability as degree of belief fits all situations; rel. freq. does not
  - Before conception, P(birth is boy) = .512
  - Pregnant woman doesn't know sex of baby, but her doctor does. What is P(boy)? Is it 0/1, or is it .512? Different for woman and her doctor?
  - What about non-repeatable situations, such as probability of major earthquake in California?
- Bayesian probability is "degree of belief" in outcome, can be assessed for all situations.

Why Be a Bayesian?
Reason 1: Philosophical, continued

- *p*-values don't really answer what we want to know. Bayesian results do.
- *p*-values are highly dependent on sample size; Bayesian results get updated with more data in a logical way.
- Bayesian results assess likely values of parameter before looking at data (prior), and update them after looking at data (posterior).

Why Be a Bayesian? Reason 2: Practical

- It's rare that we have *no* prior information. Bayesian methods build that into analysis.
  - Estimate proportion of community infected with HIV. Could it really be anything from 0 to 1?
  - Estimate mean change in blood pressure after program in meditation. Do we really think it could be anything from $-\infty$ to ∞?
- Most statistical analyses are now done as a collaboration between statisticians and experts who have prior knowledge. Why not use that knowledge?

Psi/Psychic/ESP/Anomalous Cognition

Having information that could not have been gained through the known senses.

- Telepathy: Info from another person
- Clairvoyance: Info from another place
- Precognition: Info from the future
- Correlation: Simultaneous access to info

Note: For proof, source doesn't matter. For explanation, it does!

Controlled experiments to Test ESP

Crucial elements:

1. Safeguards to rule out cheating or ordinary communication
2. Knowledge of probabilities of outcomes by chance alone

Examples… are these okay?

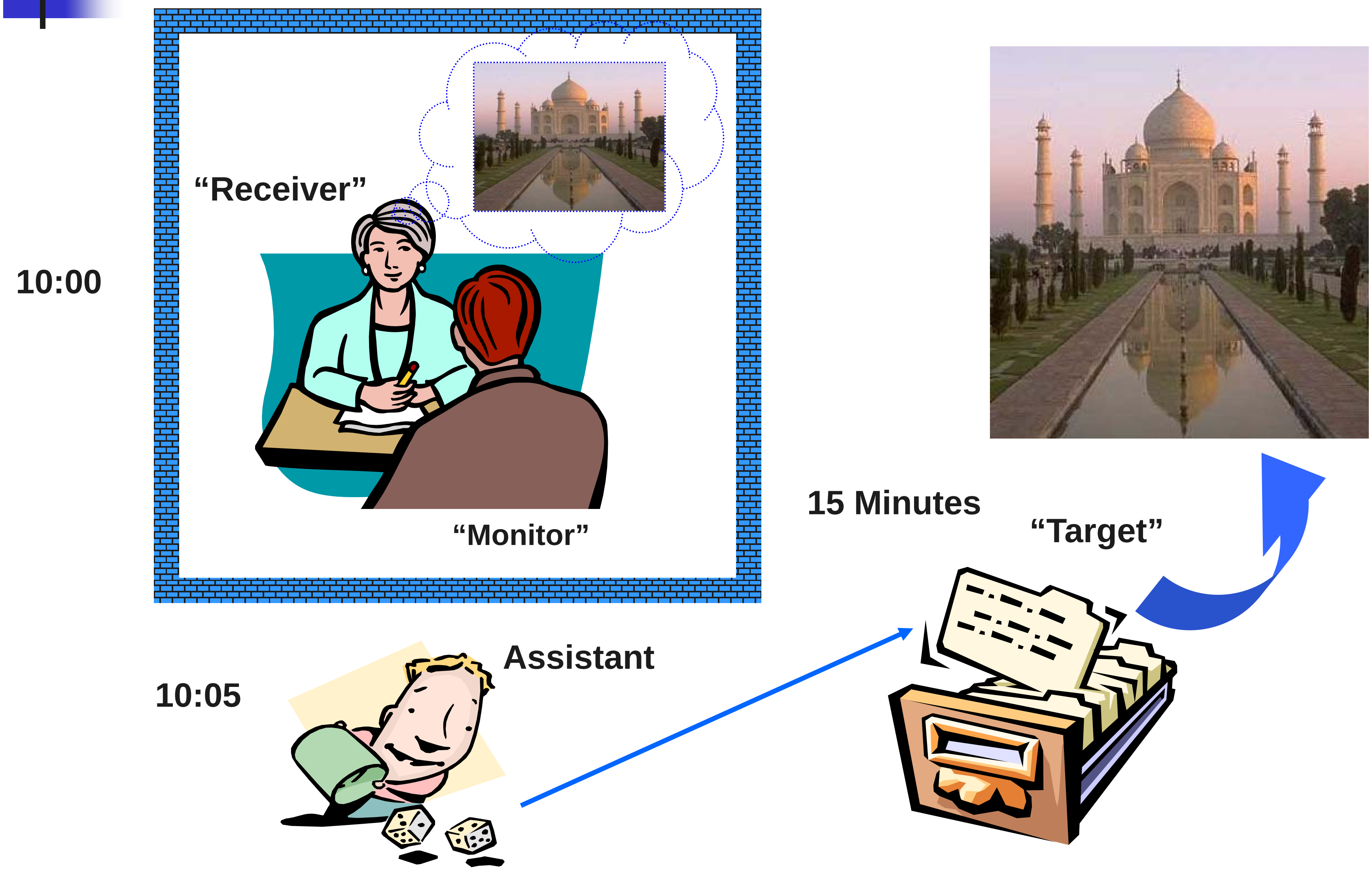
1. I am thinking of a number from 1 to 5. Guess it.
2. My assistant on 2nd floor has shuffled a deck of cards (well!) and picked one at random. What suit is it? (Example of *forced choice* experiment)

Free response ESP experiments meeting crucial elements:

- Remote Viewing, originally done by US Government
- Similar type of experiment called "ganzfeld" (will describe)

Remote Viewing Protocol

Special thanks to Dr. Edwin May for this and other SRI slides



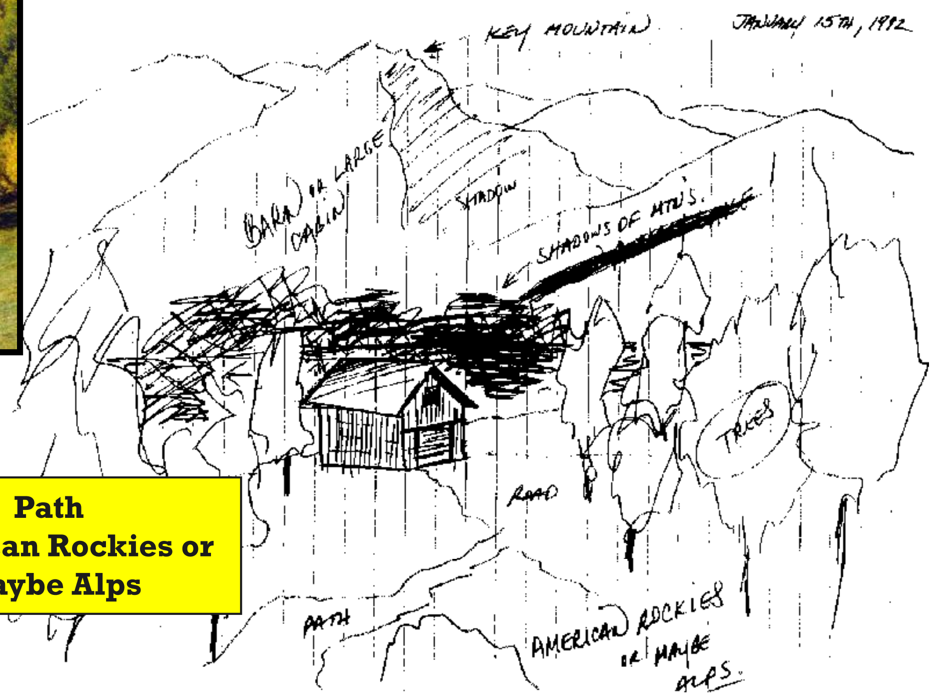
Some Additional Details

- After the session, drawings & descriptions are copied and secured so they can't be altered.
- Feedback to the remote viewer is given by showing him/her the copy of what (s)he drew, along with the target photo or video.
- Results are judged. In some labs, viewer is judge and feedback is given after judging. In others there is an independent judge.
- Meets condition #1: Safeguards to rule out cheating or ordinary means of communication

Example of an Excellent Match

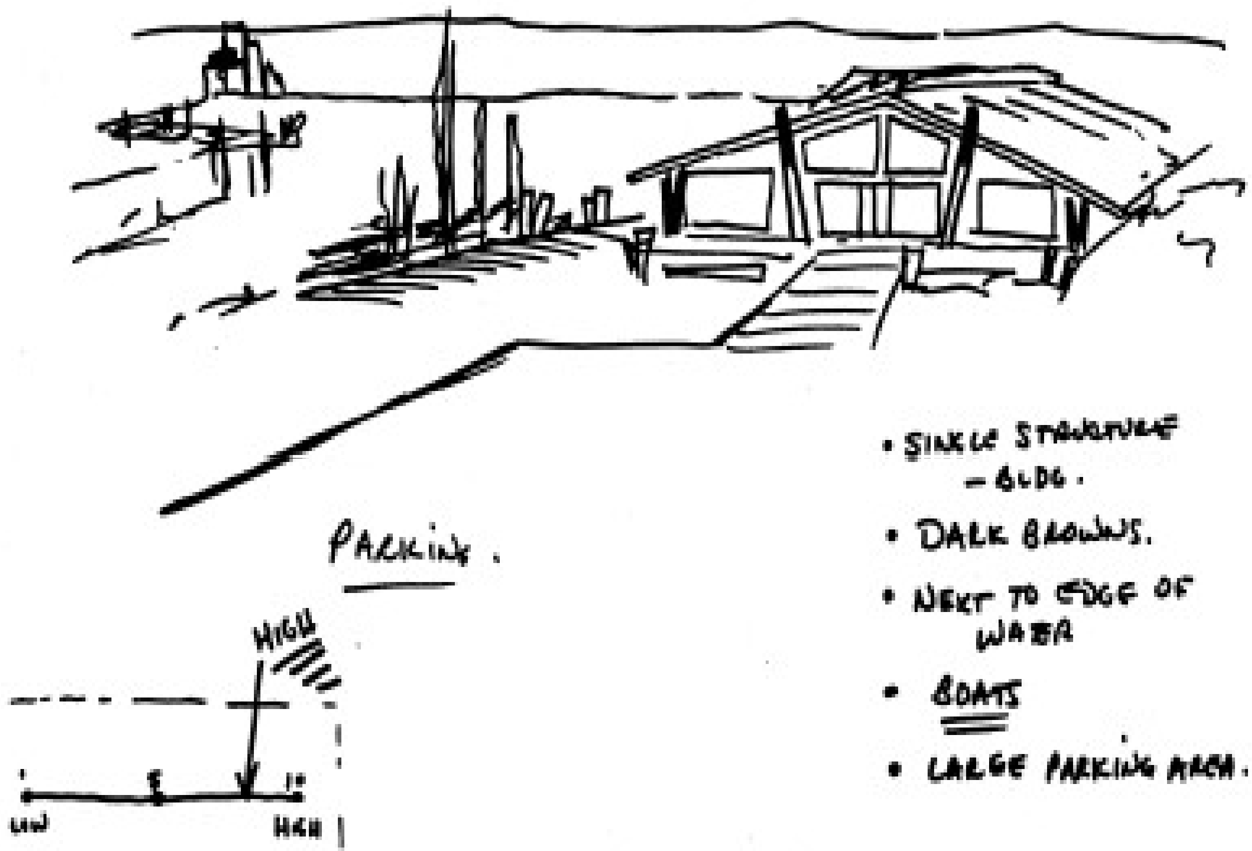
(Experiment at SAIC/Stanford)

Words: Key Mountain
Barn or Large Cabin
Shadow
Shadows of Mtns.
Trees
Road



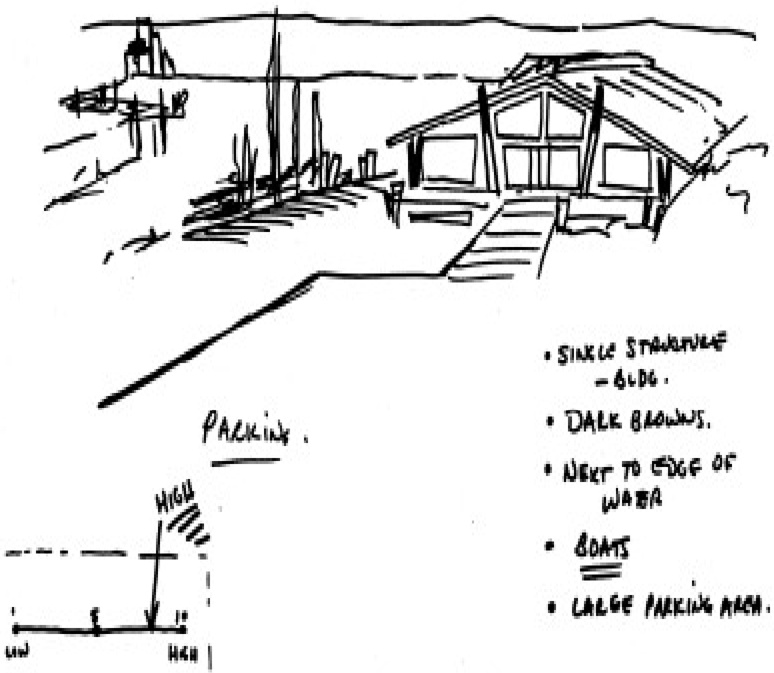
Path
American Rockies or
Maybe Alps

Early Remote Viewing Example (SRI)

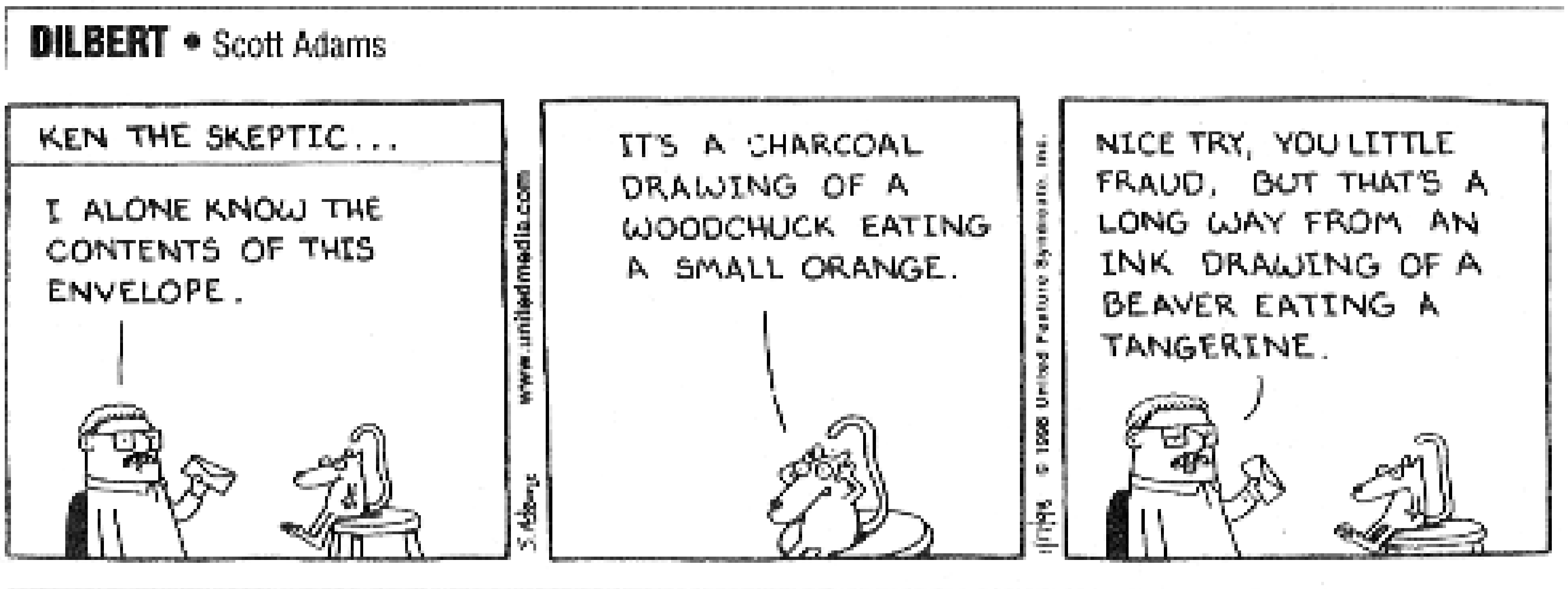


Target: Pete's Harbor Restaurant

How to Judge?


PARKING.
HIGH
LOW
HIGH
• SINGLE STRUCTURE
- BLDG.
• DARK BROWNS.
• NEXT TO EDGE OF
WATER
• BOATS
• LARGE PARKING AREA.

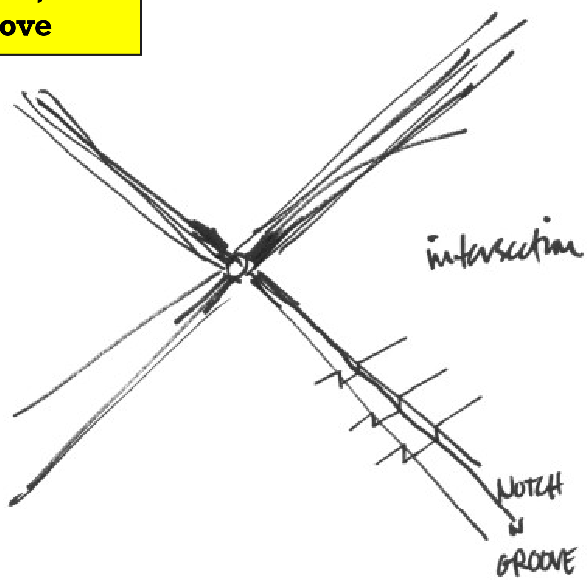
How NOT to Judge the Response

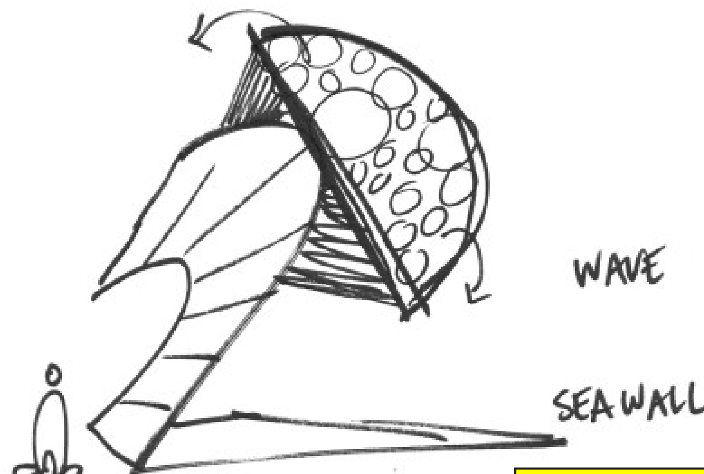


Can't use subjective probability of match – too much room for personal bias.

You Judge this Typical Novice Response

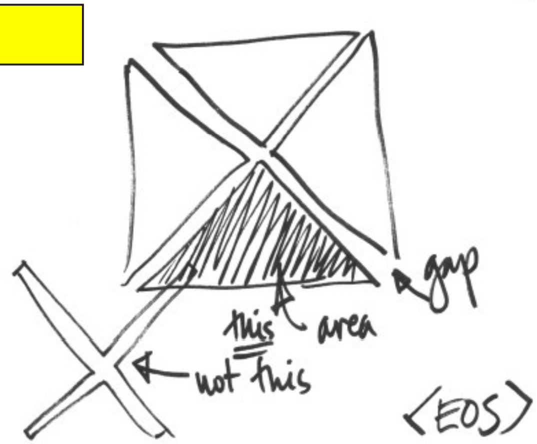
intersection, notch, groove



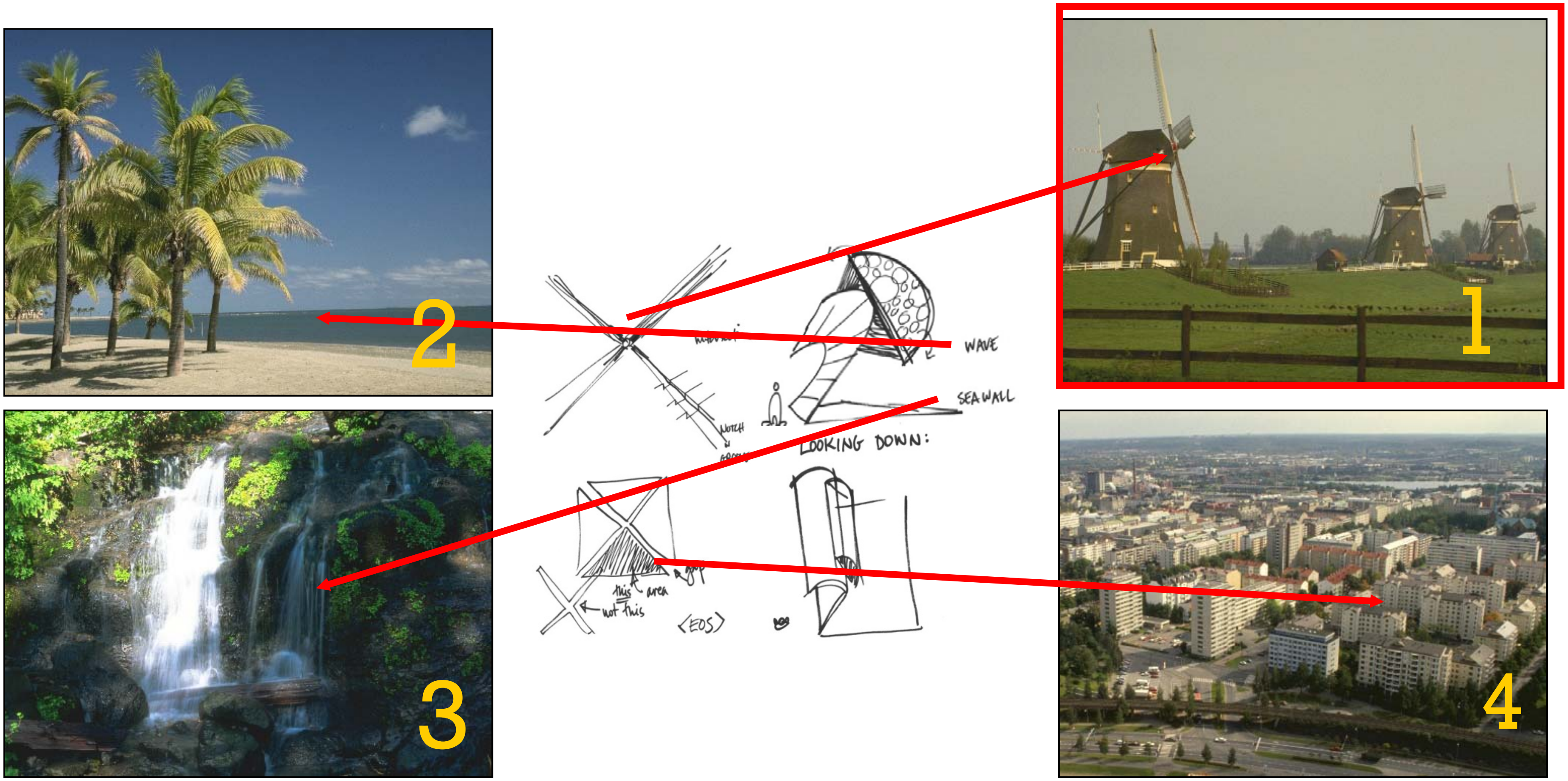


wave, sea wall

gap



Rank-Order Judging



Analysis Methods

- Before the *experiment*, targets put into packs of 4 dissimilar choices
- Before *session* begins a pack is randomly selected, then target within it (e.g. windmills). The *session* takes place, producing a response.
- After the *session*, a judge is given the response and the 4 choices and must assign ranks. Judge is blind to correct answer.
- For *session*, result = the *rank* assigned to correct target, *or* “direct hit” if it gets 1^{st} place rank. In some labs judge picks best match only.

Analysis Methods, Continued

- Summary statistic: Sum or ranks (some labs), or number of direct hits (others), for entire *experiment* (many *sessions*).
- Meets Condition #2: Knowledge of probabilities of various outcomes by chance alone.
- Note that randomness is in the selection of the *target*, not response. No matter what the response is, the randomly selected target is the best match by chance alone with probability ¼.

Automated Ganzfeld Experiments Similar to Remote Viewing

- Sender, receiver, experimenter. Target selected in same way as remote viewing (random, packs of 4)
- Sender in sound-shielded room, looking at target, which is a photograph or short video segment.
- Receiver in sound isolation room with red light in eyes, white noise in ears, comfy chair. Listens to relaxation tape. Then talks into microphone, attempting to describe the unknown target.
- Experimenter and sender listen. Then *receiver* judges response with 4 choices – actual target and 3 decoys. Direct hit analysis usually used.

Simplest Model for RV and Ganzfeld

- X = number of direct hits in experiment (proportion of successes in *n* sessions)
- Assume X ~ Binomial (n, p)
  - n = number of sessions
  - p = probability that the judge can identify the correct target, given the response.
- By chance alone, p = 1/4
- If psychic functioning occurs, expect response is a better match than chance, and $p > ¼$.

Ganzfeld Studies in This Analysis

- From meta-analyses of ganzfeld studies
- Included all ganzfeld studies from those meta-analyses that met criteria for safeguards and standard procedures
- Used 56 studies
  - Combined n = 2124 sessions
  - Combined X = 709 hits
  - X/n = .334, when .25 expected by chance

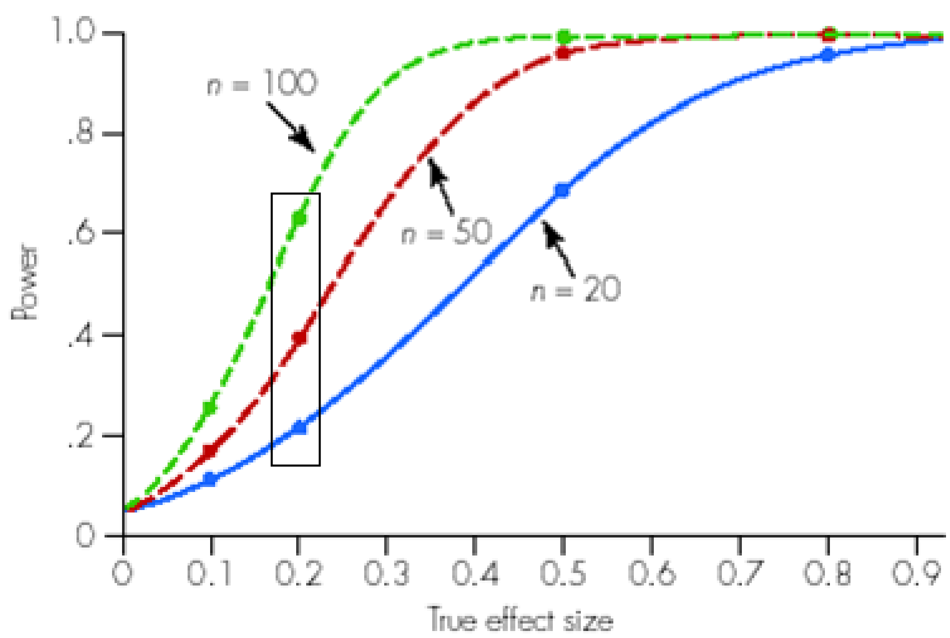
Binomial Analysis

- Define p = probability of a success in a session. Simple assumption (for now) is that p is fixed across sessions and studies.
- Hypothesis test:
  - Null: $p = .25$
  - Alternative: $p > .25$
  - *P-value* (exact binomial) = 2.26×10^{-18}
- Note that for individual studies, n ranged from 7 to 128, but mostly very small. Hard to get statistical significance for one study; power is too low. (For median n of 32, power is only .308 if true p is 1/3.)

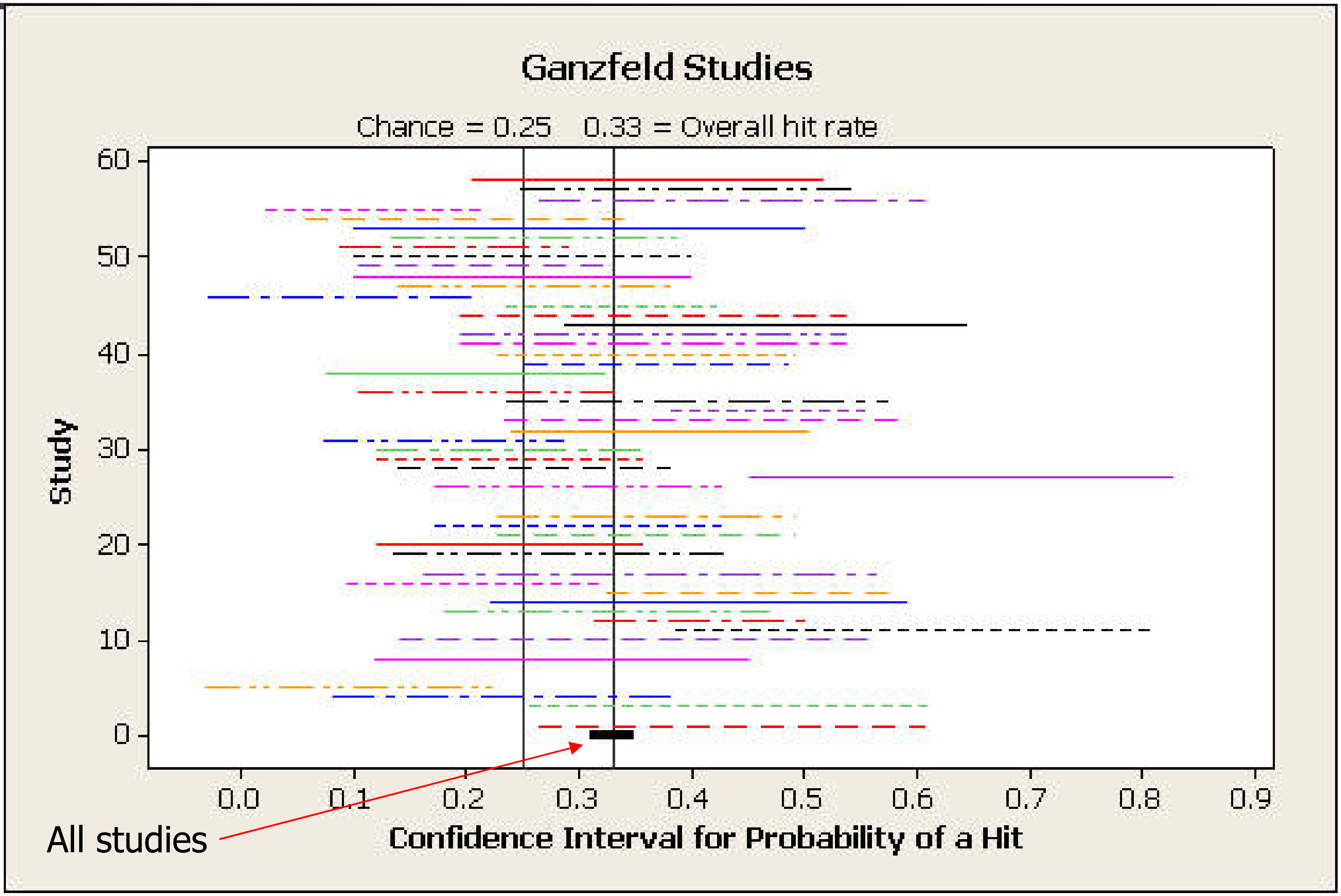
Power curves, one-sided *t-test* for one mean (from *Mind On Statistics,4th ed., Utts and Heckard*)

	True Effect Size			
Sample Size	**0.1**	**0.2**	**0.5**	**0.8**
n = 20	.11	.22	.695	.964
n = 50	.17	.40	.967	nearly 1
n = 100	.26	.63	.9996	nearly 1

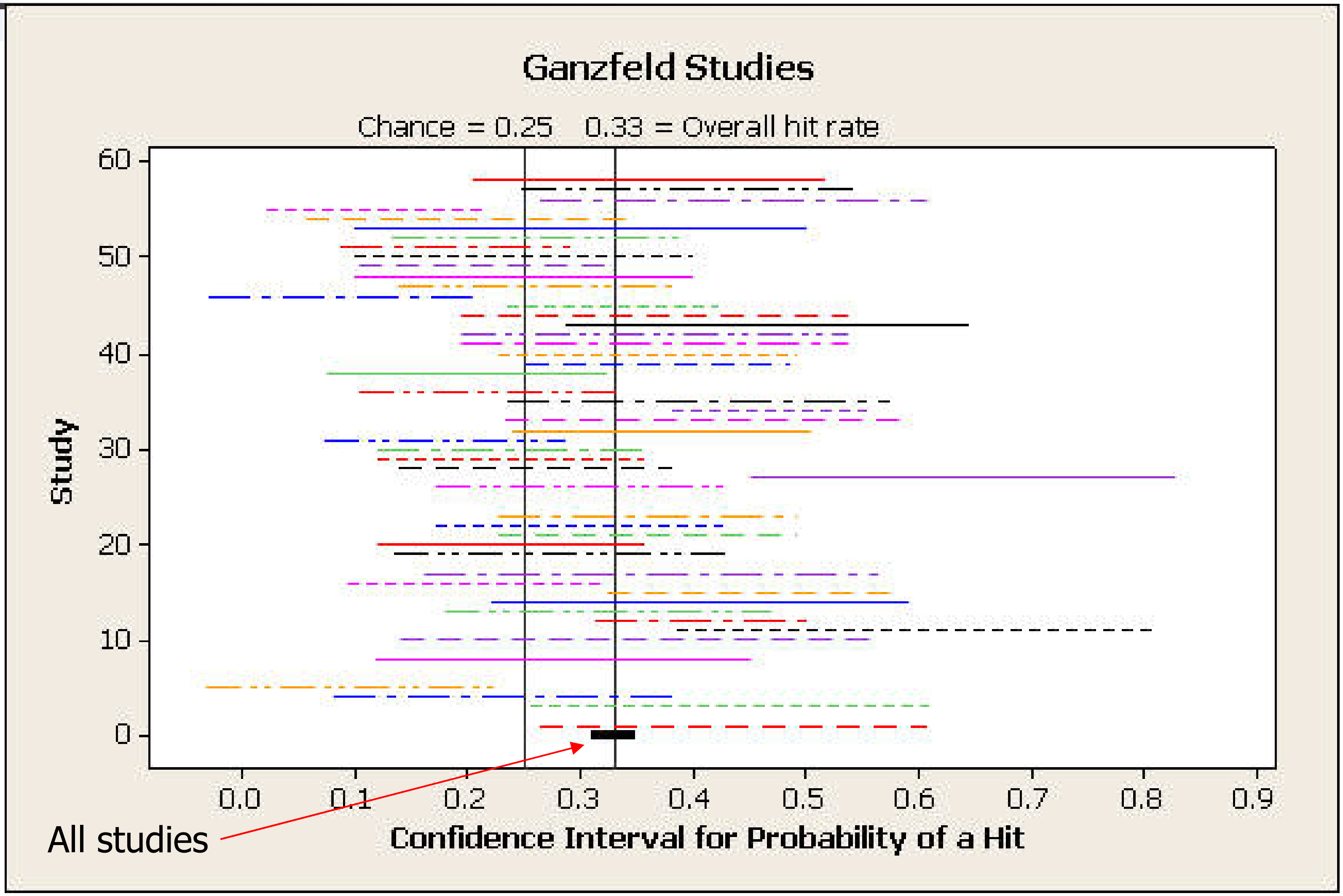
Typical psi effect size



Individual Confidence Intervals



Combined 95% CI is .314 to .354



Are You Convinced?

- Overall *p*-value is 2.26×10^{-18}
- Overall confidence interval is .314 to .354, when chance is .25.
- Yet, I have found that disbelievers don't change their minds when they see data.
- Why not? Perhaps we are all Bayesians!
- Note: Skeptics have tried unsuccessfully to find flaws with the experiments.
- In general, beliefs probably do play a role in how we interpret data!

Simple Bayesian Analysis

- Assume X = number of hits is binomial with fixed p = probability of a hit
  - $X \mid p \sim \text{Binomial}(2124, p)$
- Use Beta distribution to model prior belief about p ("conjugate prior")
  - $p \sim \text{Beta}(a, b)$
  - More about how to do this on next slide
- Posterior distribution for p is also Beta distribution
  - $\text{Beta}(X + a, n - X + b)$

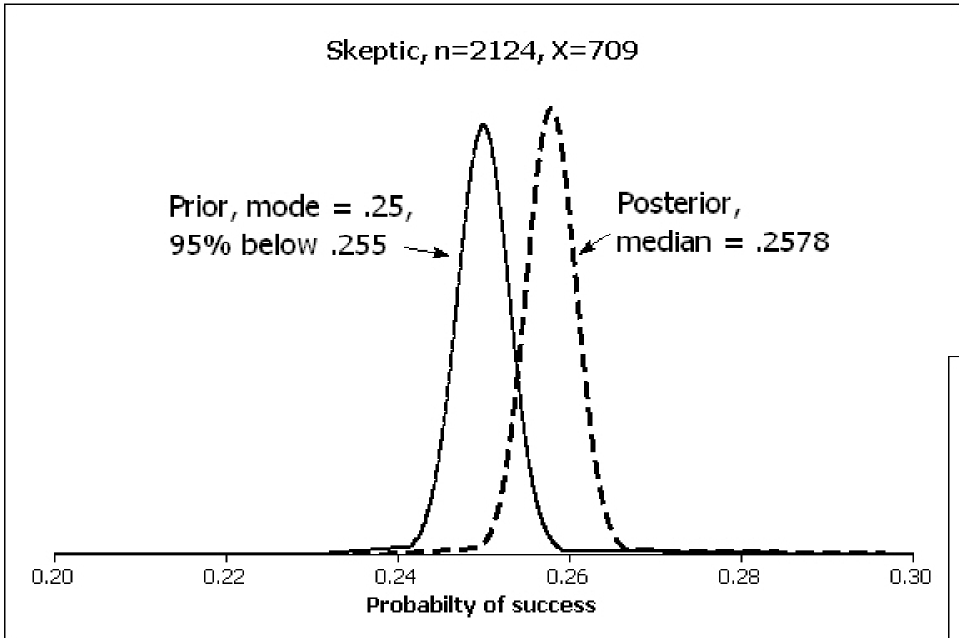
How to Determine Beta Prior

- Use free software called “BetaBuster”
  www.epi.ucdavis.edu/diagnostictests/betabuster.html
- Ask these questions to elicit the prior:
  - In your opinion, what is the most likely value for *p*? (This becomes the mode.)
  - Fill in the blank: I am 95% certain that *p* cannot exceed the value ______.
- The answers to these 2 questions determine the parameters for the Beta prior.

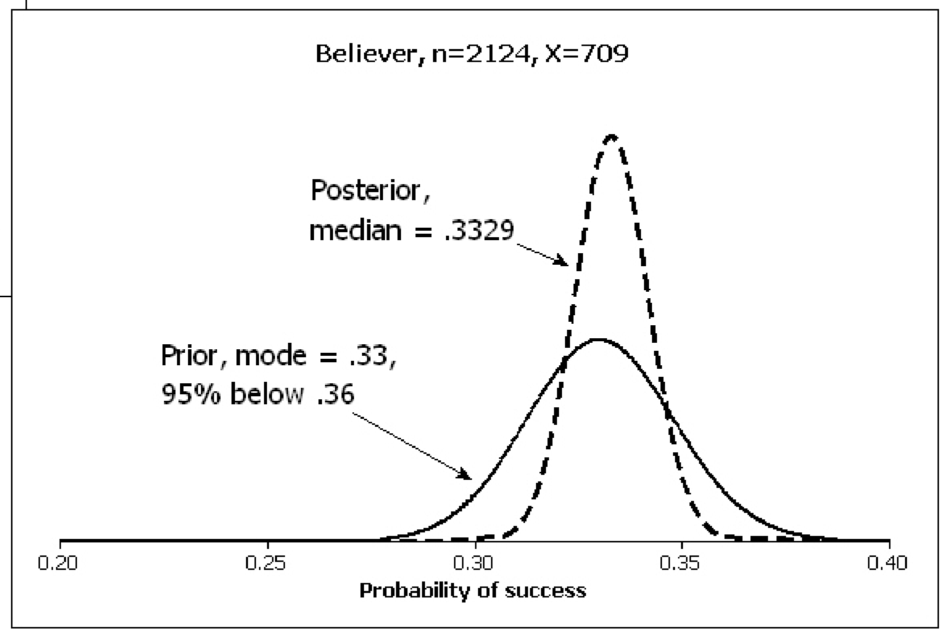
Consider 3 Prior Sets of Belief

- Skeptic:
  - Most likely value for *p* is .25 (chance)
  - 95% certain *p* is below .255
- Believer:
  - Most likely value for *p* is .33
  - 95% certain *p* is below .36
- Open-minded observer
  - Most likely value for *p* is .25 (chance)
  - 95% certain *p* is below .30

Posterior for *p*, Skeptic and Believer

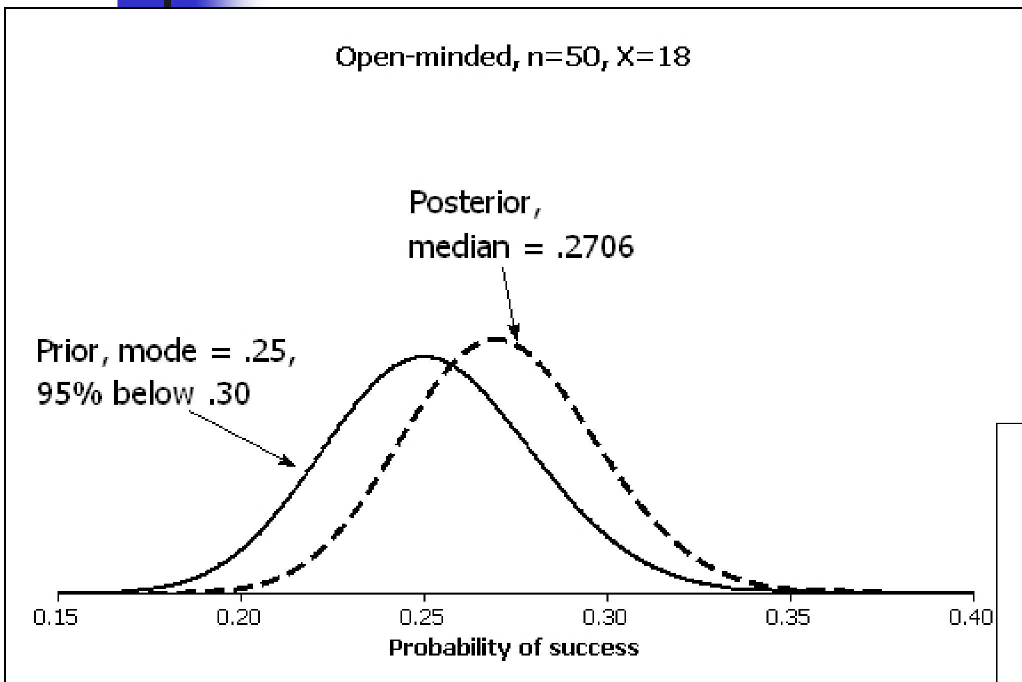


Data reduced the range of the believer's likely values for *p*

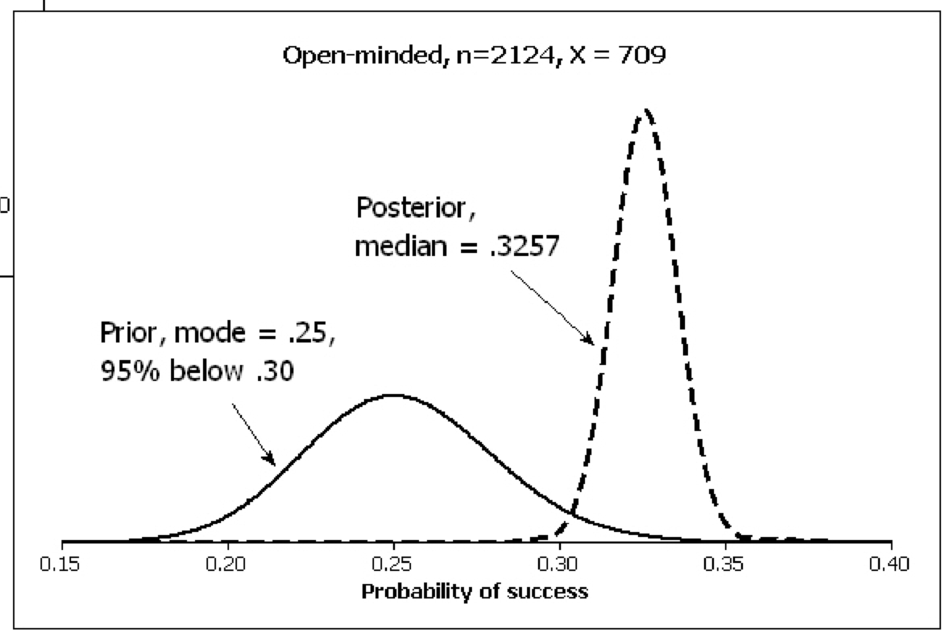


Data shifted the skeptic's belief very slightly.
Posterior median = .2578

Open-minded: One study and all data



Open-minded, all data, allows data to play major role



One study, n = 50, 36% hits, shifted the open-minded belief slightly.

Summary of Simple Analysis

- Skeptic's opinion was not changed much by the data, even with 2124 trials and 33% success rate.
- Open-minded prior allowed data to have a larger influence.
- Helps explain why skeptics still are not convinced by the evidence, even with a *p*-value of 2.26×10^{-18}
- Allows skeptics and believers to see why they disagree!

More Complex:
Bayesian Hierarchical Model

- Binomial model relies on the assumption that *p* is constant from study to study and from session to session. (May be true only for null hypothesis!)
- To test this assumption, we need a more complicated model. We assume constant hit rate *within* a study, but different hit rates *across* studies.
- Let p_i, i=1,2,…,56 be the true hit rate for study *i*.
- n_i = number of trials in study *i*

Bayesian Hierarchical Model, continued

- Hierarchical model:
  - X_i = number of hits in study i,
  - $X_i \sim \text{Binomial}(n_i, p_i)$

- p_i are "study-specific" hit rates and are assumed to come from a probability distribution. Want to estimate the median and variation of the *distribution* of p_i's across all possible studies that could be done.

Some Technical Stuff…

- We transform to speed convergence to normality and stabilize variance:

$$\theta_i = \sin^{-1}\sqrt{p_i}, \quad y_i = \sin^{-1}\sqrt{\hat{p}_i} = \sin^{-1}\sqrt{\frac{x_i}{n_i}}, \quad i = 1,\ldots,56.$$

- For large samples:

$$\hat{p}_i \sim N(p_i, \frac{p_i(1-p_i)}{n_i}) \quad \text{and} \quad y_i \sim N(\theta_i, \frac{1}{4n_i})$$

(delta method)

More Technical Stuff…

- We need to specify a distribution for the p_i's. This is done by placing a distribution on

$$\theta_i = \sin^{-1}\sqrt{p_i}$$

- We assume $\theta_i \sim N(\mu, \sigma^2)$

- μ and σ^2 are parameters we wish to estimate
- μ is the median of the distribution of θ_i's, and since the transformation is one-to-one and increasing

$$\text{median}(p_i) = \sin^2 \mu$$

- A small σ^2 means the θ_i's are similar so that the p_i's are similar, whereas a large σ^2 means the p_i's vary a lot – so there are important differences in the study-to-study hit rates.

Prior Distributions

- Bayesian Analyses were run corresponding to 4 choices of priors:
  - *Non-informative prior*: The non-informative prior for μ puts equal probability on all real numbers (improper).
  - *Weakly informative prior* (similar to open-minded in simple case): Uses median(p) = 0.25 and 90% sure median(p) is between 0.12 and 0.41*
  - *Believer's prior*: Uses median(p) = 0.33 and 90% sure median(p) is between 0.30 and 0.36
  - *Skeptic's prior*: Uses median(p) = 0.25 and 90% sure median(p) is between 0.245 and 0.255

*Comes from prior on θ's being N($\sin^{-1}$(.25), .01)

Results

parameter	Bayesian noninformative prior 2.50%	50%	97.5%	sd	Frequentist MLE	95% CI low	95% CI upper	Bayesian weakly informative prior 2.50%	50%	97.50%	sd
Median(p_i)	0.30	0.33	0.36	0.02	0.33	0.31	0.36	0.29	0.33	0.36	0.02
95th percentile of p	0.42	0.49	0.57	0.04	0.50	0.45	0.56	0.44	0.51	0.59	0.04
5th percentile of p	0.13	0.19	0.24	0.03	0.18	0.14	0.21	0.12	0.17	0.22	0.03
μ	0.57	0.61	0.65	0.02	0.61	0.59	0.64	0.57	0.61	0.65	0.02
σ^2	0.0042	0.0100	0.0197	0.0040	0.0116	0.0061	0.0171	0.0059	0.0123	0.0237	0.0046

parameter	Bayesian: Skeptic's Prior 2.50%	50%	97.50%	sd	Believer's Prior 2.50%	50%	97.50%	sd
Median(p_i)	0.251	0.257	0.262	0.003	0.308	0.326	0.345	0.01
95th percentile of p	0.253	0.260	0.266	0.003	0.348	0.374	0.394	0.01
5th percentile of p	0.248	0.254	0.260	0.003	0.262	0.281	0.305	0.01
μ	0.525	0.531	0.537	0.003	0.59	0.61	0.63	0.01
σ^2	2.6E-8	4.4E-6	1.5E-5	4.6E-6	3.5E-4	9.5E-4	0.001	2.0E-4

Percentiles of Posterior Distribution of Median(p)

	2.5% of Median (p)	50% of Median (p)	97.5% of Median (p)
Non-inform	.30	.33	.36
Open-mind	.29	.33	.36
Frequentist	.31	MLE = .33	.36
Believer	.308	.326	.345
Skeptic	.251	.257	.262

95% Range for Individual *p*

- Non-informative: .19 to .49
- Open-minded: .17 to .51
- Frequentist (MLE) .18 to .50

All of the above are similar.

But these are narrower, especially skeptic:

- Believer: .281 to .374
- Skeptic: .254 to .260

Finding about Study-to-Study Variation

- Under the frequentist analysis, we obtain that 90% of the study-specific hit rates (p_i's) are in the interval (0.18, 0.50), weakly informative (open-minded) prior gives (0.17,0.51)
- The data DO indicate study-to-study differences in the hit rate. Thus, a binomial model may not be appropriate.

Comparing Bayesian and Frequentist Results

- Results under frequentist, Bayesian non-informative and weakly informative (open-minded) priors are very similar
  - 95% probability interval for median (p_i) is (0.30, 0.36)
- Bayesian analysis under informative priors is sensitive to priors
  - Skeptics prior gives 95% probability interval for median (p_i) as (0.251,0.262)
  - Believer's prior gives 95% probability interval for median (p_i) as (0.308, 0.345)

Some conclusions from the analyses

- "Average" hit rate (for population) seems to be slightly above 30%, whatever method is used (except skeptic's prior).
- Binomial model with fixed p is too simple; hit rates may change based on a number of factors.
- Statistical models need to incorporate additional information about participants, conditions of experiment, etc. Bayesian approach is most reasonable.

Summary

- ESP experiments are a good way to illustrate:
  - Testing a clear null hypothesis
  - Why "replication" should not be based on p-values (low power)
  - Simple Bayesian analysis
  - Why prior beliefs matter

Recent Controversy

- Bem, D. J. (2011). Feeling the Future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407-425.
- Response in same issue by Wagenmakers et al, Bayesian analysis that seemed to negate results.
- Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, 101, 716-719

Recent controversy, continued

- Wagenmakers et al used a Bayesian prior that was unrealistic – assumed if psi exists, the effect is huge. So the data supported the null (no psi), rather than that unrealistic alternative.
- We used a more reasonable prior – assumed if psi exists the effect is relatively small. Data overwhelming supported the psi hypothesis.

If you want to try it....

- On-line tests:

  www.gotpsi.org

- IPhone (iPod) application:

  www.espresearch.com/iphone

- Link to Bem, Utts, Johnson paper (written for psychologists, so easy to understand the Bayesian part):

  http://dl.dropbox.com/u/8290411/ResponseToWagenmakers.pdf

QUESTIONS?

Contact info:

jutts@uci.edu

http://www.ics.uci.edu/~jutts