Understanding and Applying Good Statistical Principles

Jessica Utts
Department of Statistics
University of California, Irvine
http://www.ics.uci.edu/~jutts
jutts@uci.edu

How Statistical Inference Works

- Create a *model* of a process or population
  - May include unknown “parameters”
  - “All models are wrong, but some are useful”
- Collect data
- Hypothesis tests
  - Compare the observed data to “chance”
- Confidence intervals
  - Estimate the unknown “parameters”

Example: Ganzfeld & Remote Viewing

- Assume targets are arranged in packs of 4 dissimilar choices.
- Target pack is randomly selected, then correct target within pack is selected
- Session takes place
- Judge shown the 4 choices from the pack
- Use "direct hit" only – judge either picks correct target or not.
- Data for experiment is number of direct hits

Model using *binomial experiment*

1. There are *n* "trials" where *n* is determined in advance. (I.e., no “optional stopping” allowed.)
2. There are *the same two possible outcomes* on each trial, called "success" and "failure" and denoted S and F.
3. The *outcomes are independent* from one trial to the next. Knowledge of one does not help predict the next one.
4. The probability of a "success" *remains the same* from one trial to the next, and this probability is denoted by *p*. The probability of "failure" is $(1 - p)$ for every trial. [Ganzfeld & r.v., $p = ¼$ by chance.]

Comment about this model

- Binomial model may be too simplistic
- Probability of a hit may depend on other factors, like creative or not, meditator or not, etc.
- Can use more complex models, but will not discuss today

Probabilities for Binomial

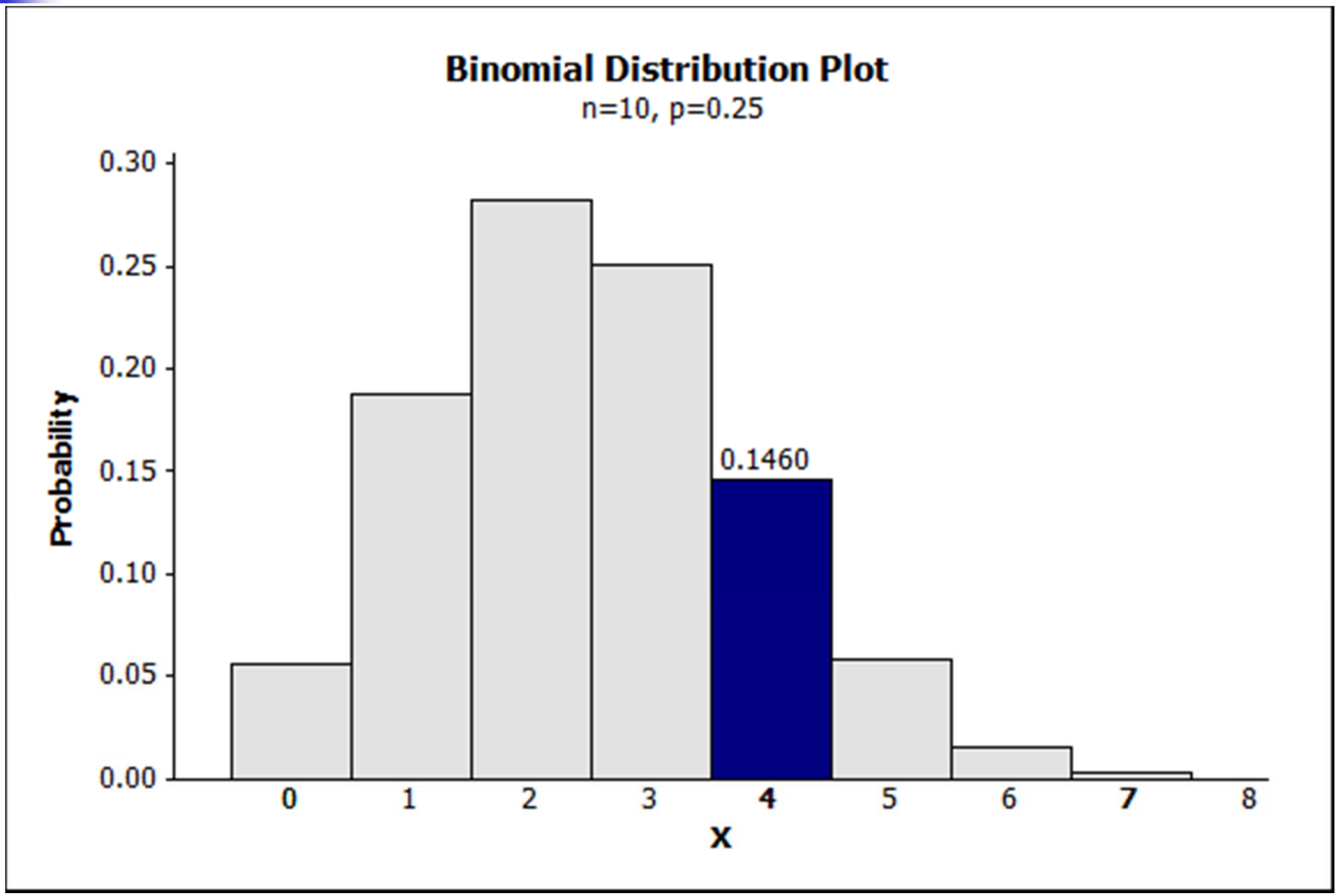
- For a binomial experiment with n trials, if X = number of successes, then for k = 0, 1, ..., n

$$\Pr(X = k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

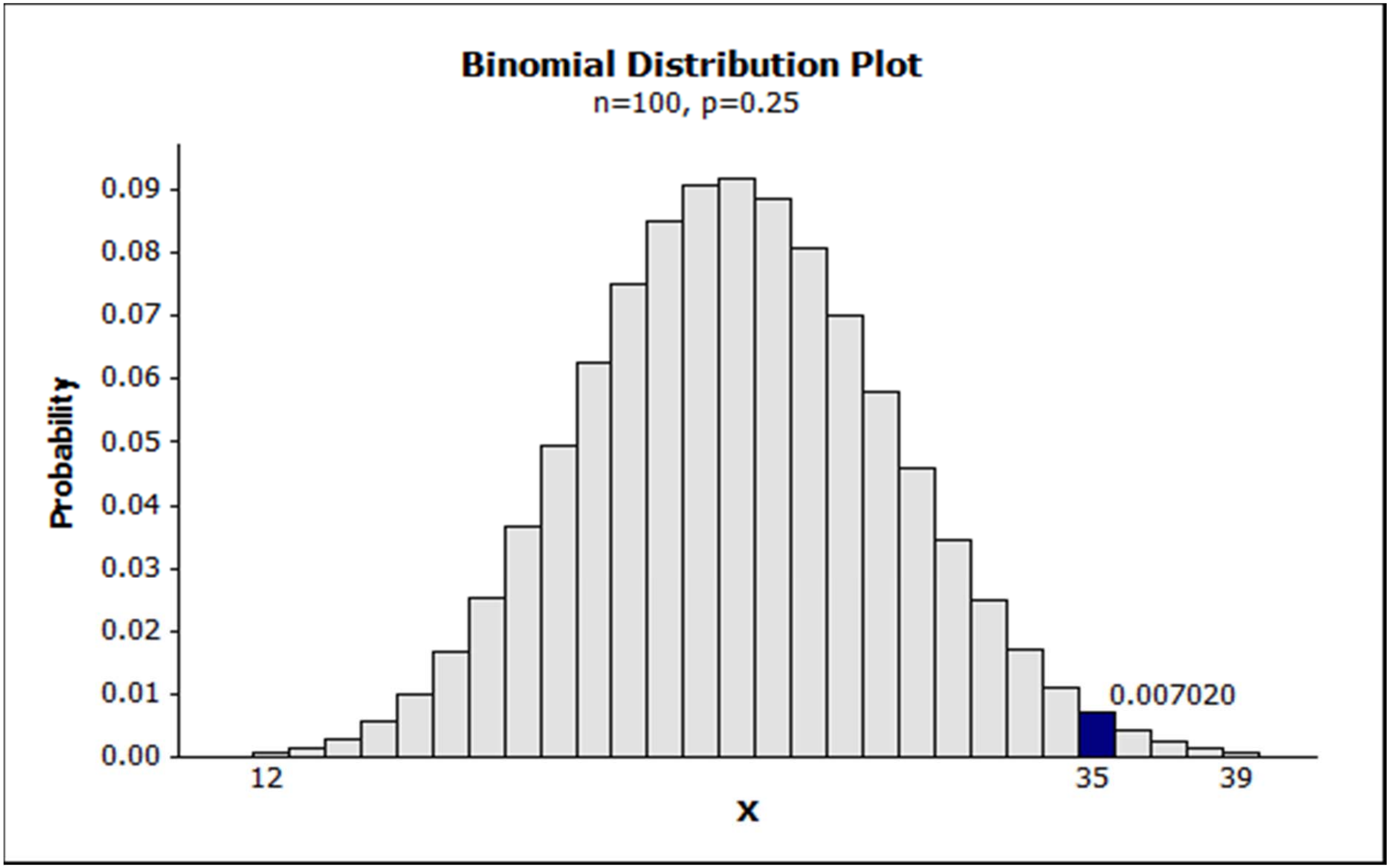
- Ex: Suppose n = 10, p = .25, $X = 4$

$$\Pr(X = 4) = \frac{10!}{4!(6)!} .25^4 \, (.75)^6 = .146$$

Probability distribution, n = 10, p = .25
Probability of 4 hits = .146



Suppose there are 35 hits in 100 trials
Probability of 35 hits = .007



Probability Question

- When we observe *k* hits in *n* trials, we could ask:
  - "What is the probability of exactly *k* hits by chance alone?" For example:
  - Probability of 4 hits in 10 trials = .146
  - Probability of 35 hits in 100 trials = .007
- More appropriate question:
  - What is the probability of *at least* k hits by chance alone?
  - This is the rationale behind the *p*-value of a test.

General Steps for Testing Hypotheses

1. Determine the **null** hypothesis and the **alternative** hypothesis.
2. Collect data and summarize with a single number called a **test statistic**.
3. Determine how **unlikely** test statistic would be ***if the null hypothesis were true***. This is the *p*-value.
4. Make a statistical **decision**.
5. Make a conclusion in **context**.

Step 1: The Hypotheses

- General:
  - Null hypothesis is there is no effect, no relationship, no difference, etc.
  - Alternative hypothesis is that there is an effect
- Ganzfeld and remote viewing, 4 choices
  - Use binomial experiment as the model
  - Define p = probability of a direct hit
  - Null hypothesis: $p = 1/4$ (or .25)
  - Alternative hypothesis: $p > 1/4$

Step 2: Data and test statistic

- General:
  - For a binomial experiment, test statistic = number of successes.
  - For many other situations the test statistic is a z-score or t-score, measuring how far data value is from the null hypothesis value.
- Ganzfeld and remote viewing:
  - Test statistic = number of direct hits
  - Sometimes use *z*-score instead (too detailed to explain here), but number of direct hits is better

Step 3: The *p*-value

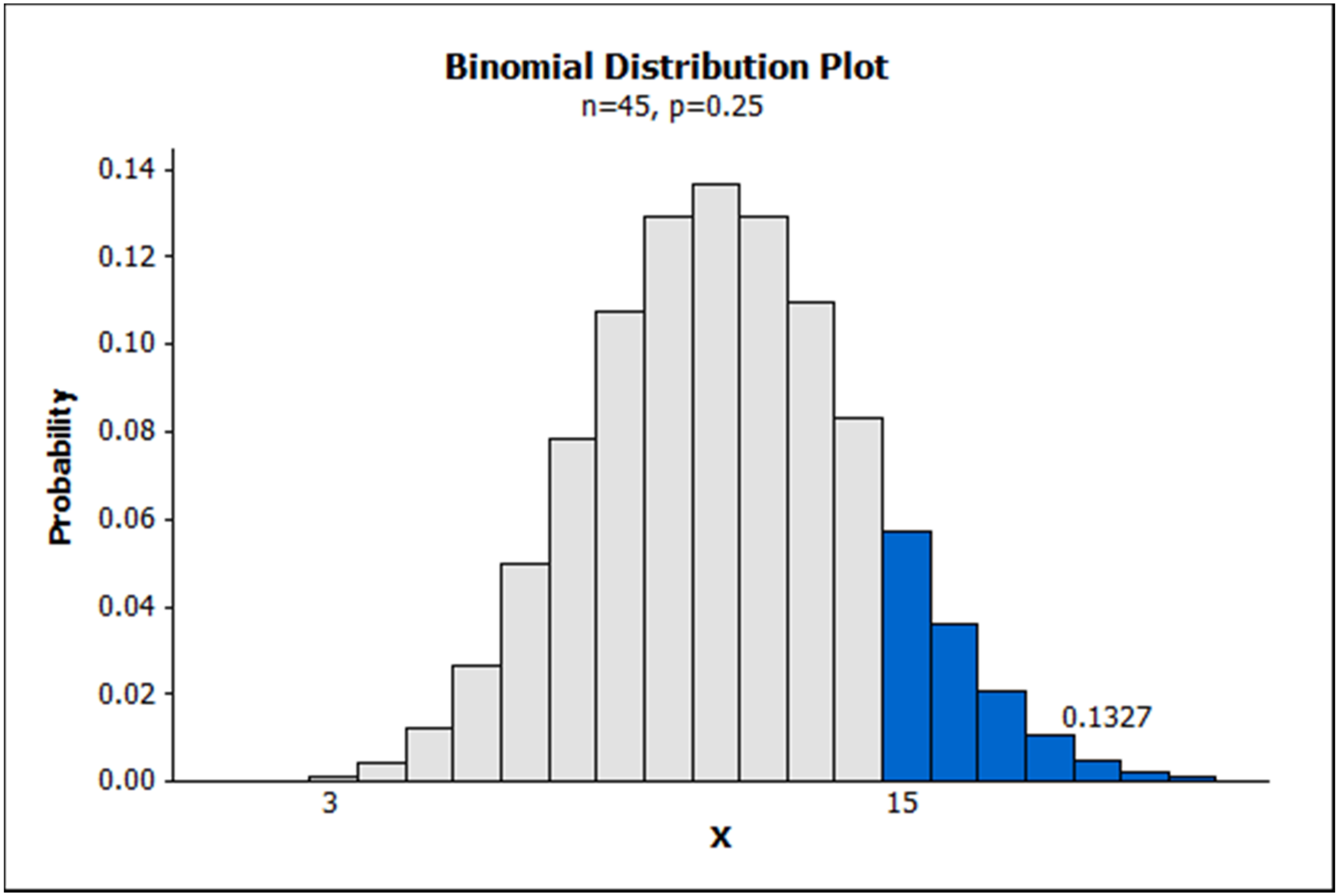
- This is the trickiest part!
- It is a *conditional* probability
- The *p*-value is the answer to this question:
  - What is the probability of observing a test statistic as large as the one observed or larger,
  - in the direction that supports the alternative hypothesis,
  - *if* the null hypothesis is true.

The *p*-value for ganzfeld & r.v.

- X = number of direct hits in n trials
- Null hypothesis is that probability of a hit on each trial is ¼ or .25
- Alternative hypothesis includes only values above ¼
- Therefore, if there are k hits, *p-value* is

Probability of *k or more hits* for a binomial distribution with n trials and success $p = 1/4$.

Example: Suppose $n = 45$, $k = 15$ Probability of at least 15 hits is .1327



Steps 4 and 5: Make a decision

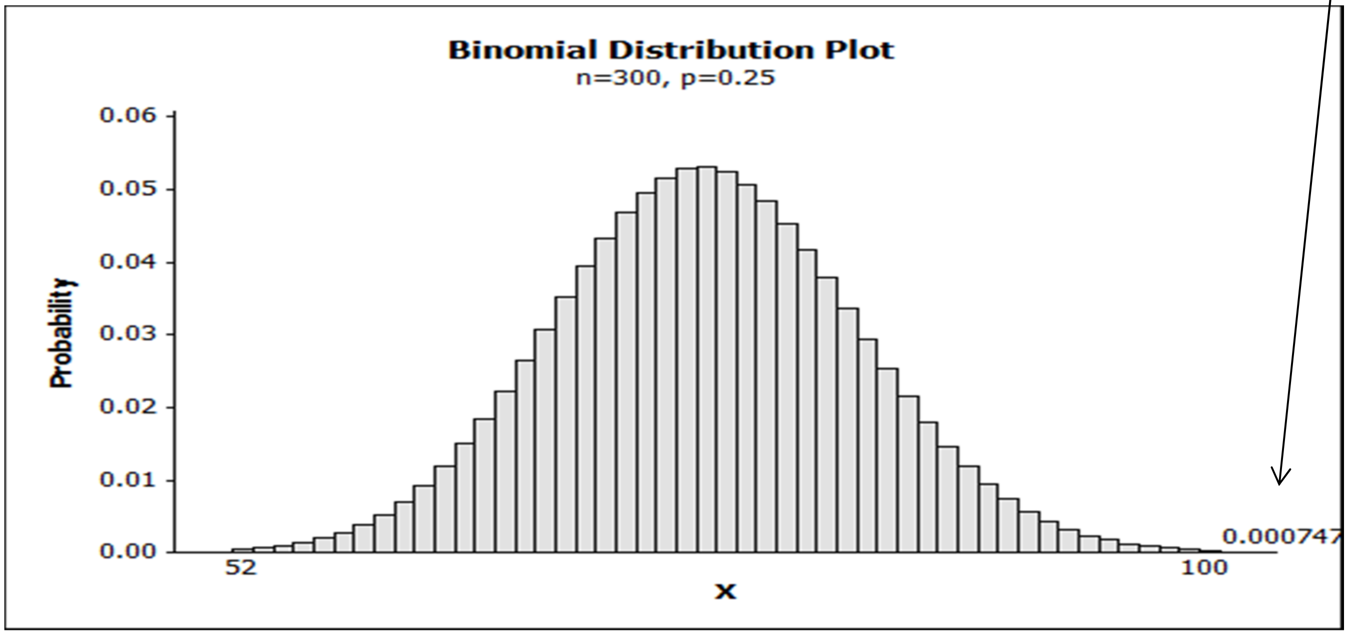
- Standard is to use .05 "level of significance"
- If *p*-value > .05
  - Cannot reject the null hypothesis
  - Result is not "statistically significant"
- If *p*-value ≤ .05
  - Reject the null hypothesis
  - Accept the alternative hypothesis
  - Result is "statistically significant"

Some issues with *p*-values

- A *p*-value is *not* the probability that the null hypothesis is true, as some think.
- A *p*-value > .05 *does not* mean the null hypothesis is true and can be *accepted.*
- A *p*-value < .05 *does not* mean the effect is large, even if the *p*-value is much smaller than .05.

Two examples, both with 1/3 hits

- If $n = 45$, hits = 15, p-value = .1327.
  - Do not reject the null hypothesis.
- If $n = 300$, hits = 100, p-value = .000747
  - Clearly reject the null hypothesis

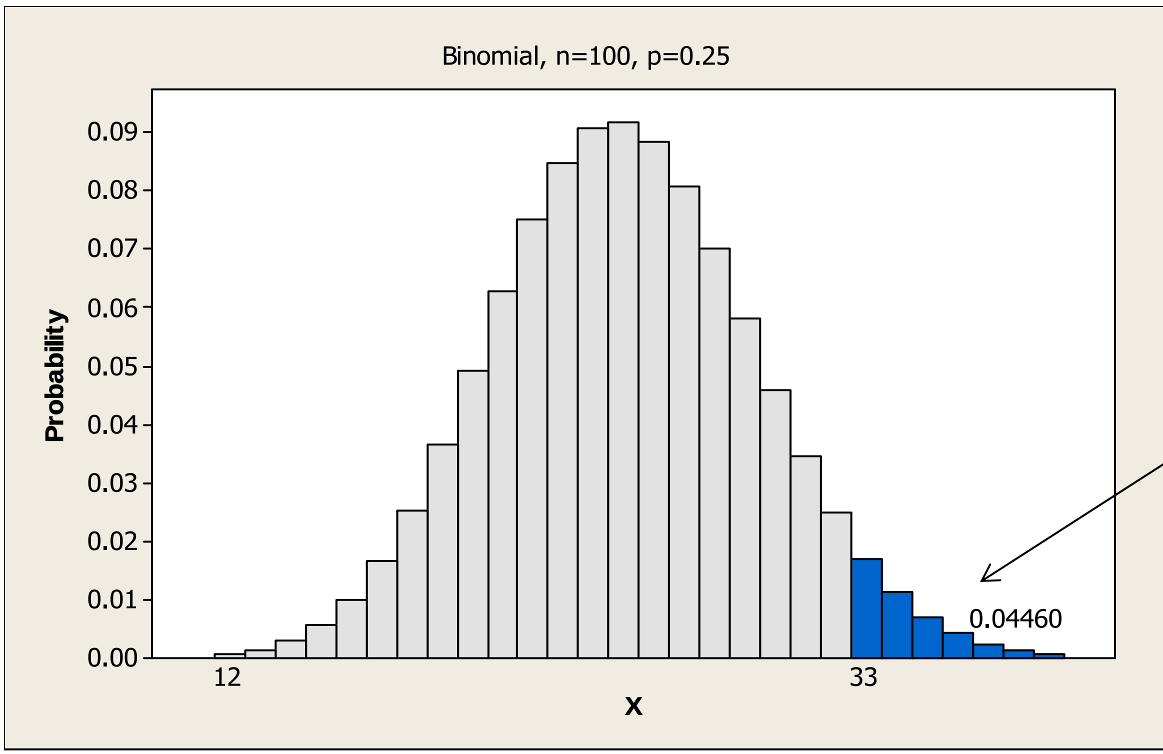


Two Types of Error: Type 1

- Only happens when the null hypothesis is true
- The error is that the null hypothesis is rejected
- Similar to a “false positive”
- Probability of a Type 1 error is whatever is used as the level of significance, usually .05.
- The claim about “extraordinary claims requiring extraordinary evidence” is saying that the level of significance should be set very low, to avoid a Type 1 error.

Example: For n = 100, when is null rejected?

Would need at least 33 hits because when null is true, probability that $X \geq 33$ is .0446



If null is true, a Type 1 error occurs if $X \geq 33$. Probability of that is .0446.

Two Types of Error: Type 2

- Only happens when the alternative hypothesis is true
- The error is that the null hypothesis is *not* rejected
- Similar to a “false negative”
- Unlike the null hypothesis, the alternative hypothesis includes a whole range of values
- Probability of a Type 2 error *depends* on *what value* in the alternative hypothesis is true.
- Power = 1 – Probability of Type 2 error

How is Power Calculated?

- Specify a value in the alternative hypothesis (let's call it p_a) for which you want power
- Specify the number of trials you will do
- Specify the level of significance (.05?)
- Find the number of successes that would lead to rejecting the null hypothesis
- Power = the probability of that many or more successes, *if* the value p_a is true

Example of finding power

- Experiment has 100 sessions, use .05 level of significance; find power if true $p = .33$
- How many successes are required to reject the null hypothesis?
  - With 33 successes, p-value is .0446
  - With only 32 successes, p-value is .069
  - So need 33 or more successes to reject null.
- Power = Prob. of at least 33 successes when the true hit rate is .33 = .5375

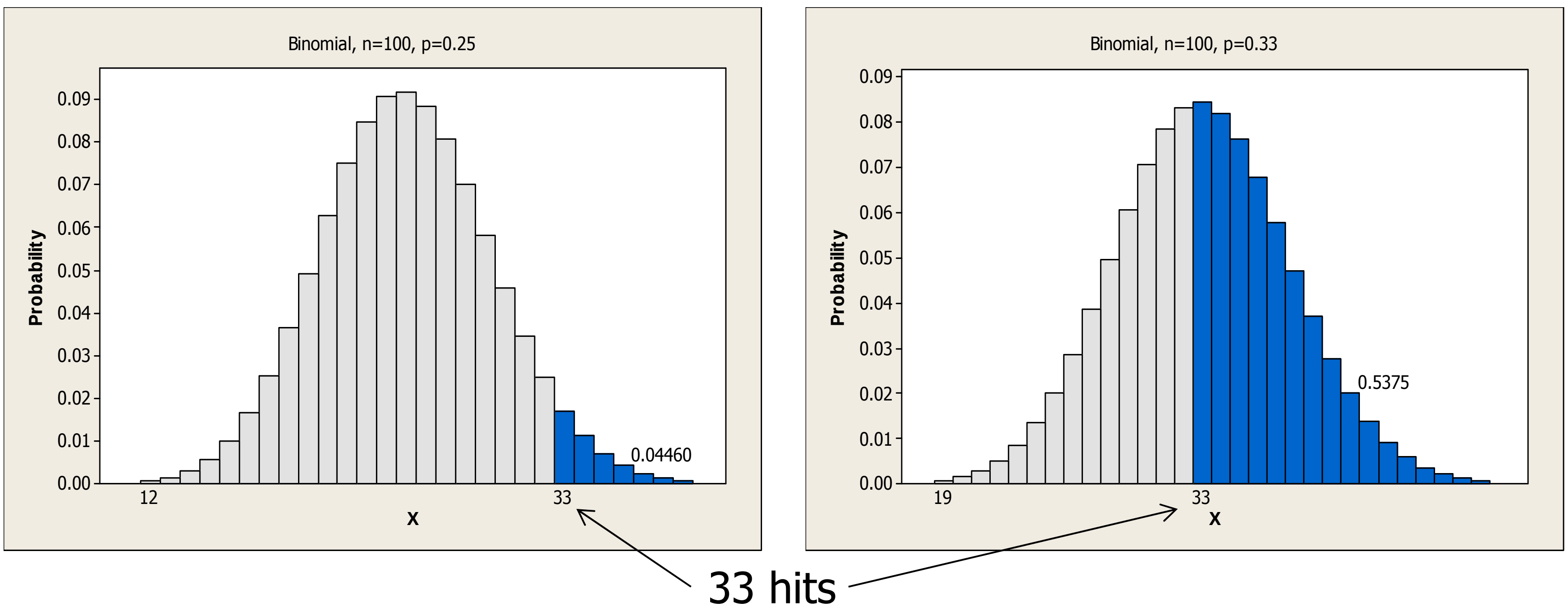
Type 1 error (left) and Power (right)

Picture when p = .25

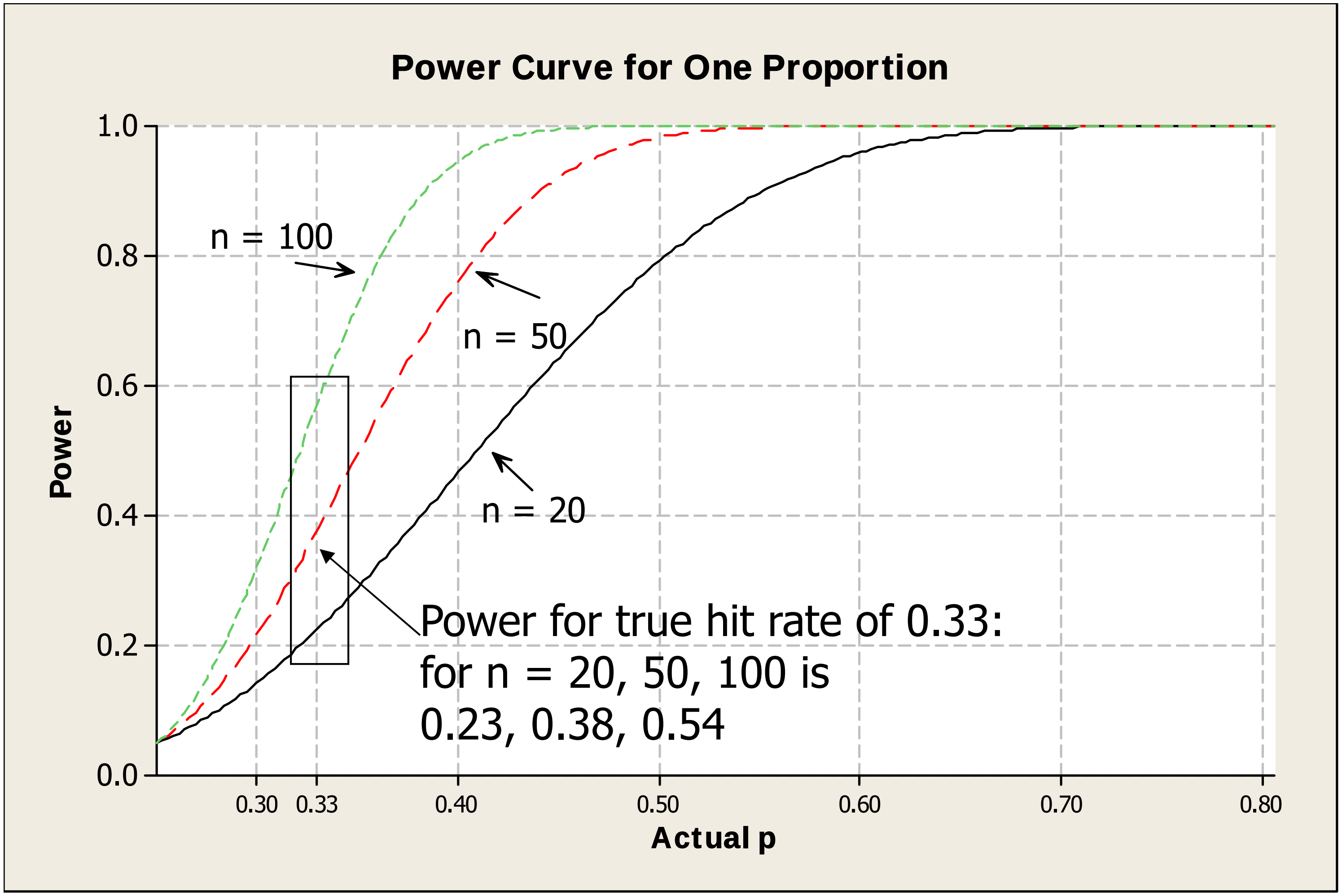
Shaded area = prob of 33 or more hits = .0446

Picture when p = .33

Shaded area = prob of 33 or more hits = .5375



Power curves:
One-sided binomial test of p = .25



Useful website for finding power

- http://www.statpages.org
- Click on “power, sample size and experimental design”
- Click on the type of test you want, e.g. Power/Sample size to compare a proportion to a specific value
- Put in your values
- Can also specify power and find required number of trials to achieve it.

Confidence Intervals

- A **parameter** is a population characteristic – value is usually unknown. Ex: True probability of a success.
- A **statistic,** or **estimate,** is a characteristic of a sample. A statistic estimates a parameter. Ex: Hit rate in a study.
- A **confidence interval** is an interval of values computed from sample data that is likely to include the true population value.
- The **confidence level** (often .95) for an interval describes our confidence in the procedure we used. *We are confident* that most of the confidence intervals we compute using our procedure will contain the true population value.

The Confidence Level Concept

- Applet to demonstrate confidence interval concept

http://www.rossmanchance.com/applets/NewConfsim/Confsim.html

- Note that on average, about 19 out of 20 or 95 out of 100 of all 95% confidence intervals should cover the true population value.

Confidence Interval Width

The width of a confidence interval is determined by:

- Sample size (*n* = number of trials)
  - Larger *n* provides greater accuracy, so more narrow interval
- Confidence level
  - Higher confidence requires wider interval
  - Extreme would be 100% confident that true hit rate is between 0 and 1!

Examples of Confidence Intervals

- Using exact binomial, C.I. for true prob of hit
  - http://www.statpages.org/confint.html
- 100 sessions, 33 hits, 95% C.I. is .239 to .431
- 45 sessions, 15 hits (33% hits):
  - 90% confidence interval is .218 to .466
  - 95% confidence interval is .200 to .490
  - 99% confidence interval is .157 to .535
- 45 sessions, 18 hits (40% hits):
  - 95% C.I. is .257 to .557
  - Lower end just barely above .25, even with 40% hits!

Relationship between test and C.I.

- For a two-sided alternative hypothesis of the form "Population value ≠ null value"
  - If the null value *is* covered by a 95% C.I., then you cannot reject the null hypothesis at .05. The null value is a *plausible* value.
  - If the null value is *not* covered by 95% C.I., you *can* reject the null hypothesis (and accept the alternative) at .05.
- For a one-sided (>) alternative, use a 90% C.I. and reject null hypothesis at .05 if the entire interval is above null the value.

Confidence interval or hypothesis test? I recommend presenting both!

- Confidence interval gives the *magnitude* of the effect.
- Confidence interval illustrates how much uncertainty there is (width of the interval)
- Confidence intervals are easier to interpret
- But, hypothesis tests provide information on how unlikely results would be if the null hypothesis were true.

Effect Size

- An effect size measures how far the true parameter value is from the null value, usually in terms of standard deviations.
- Effect size for binomial is harder to interpret, so we'll switch to a more mundane example.

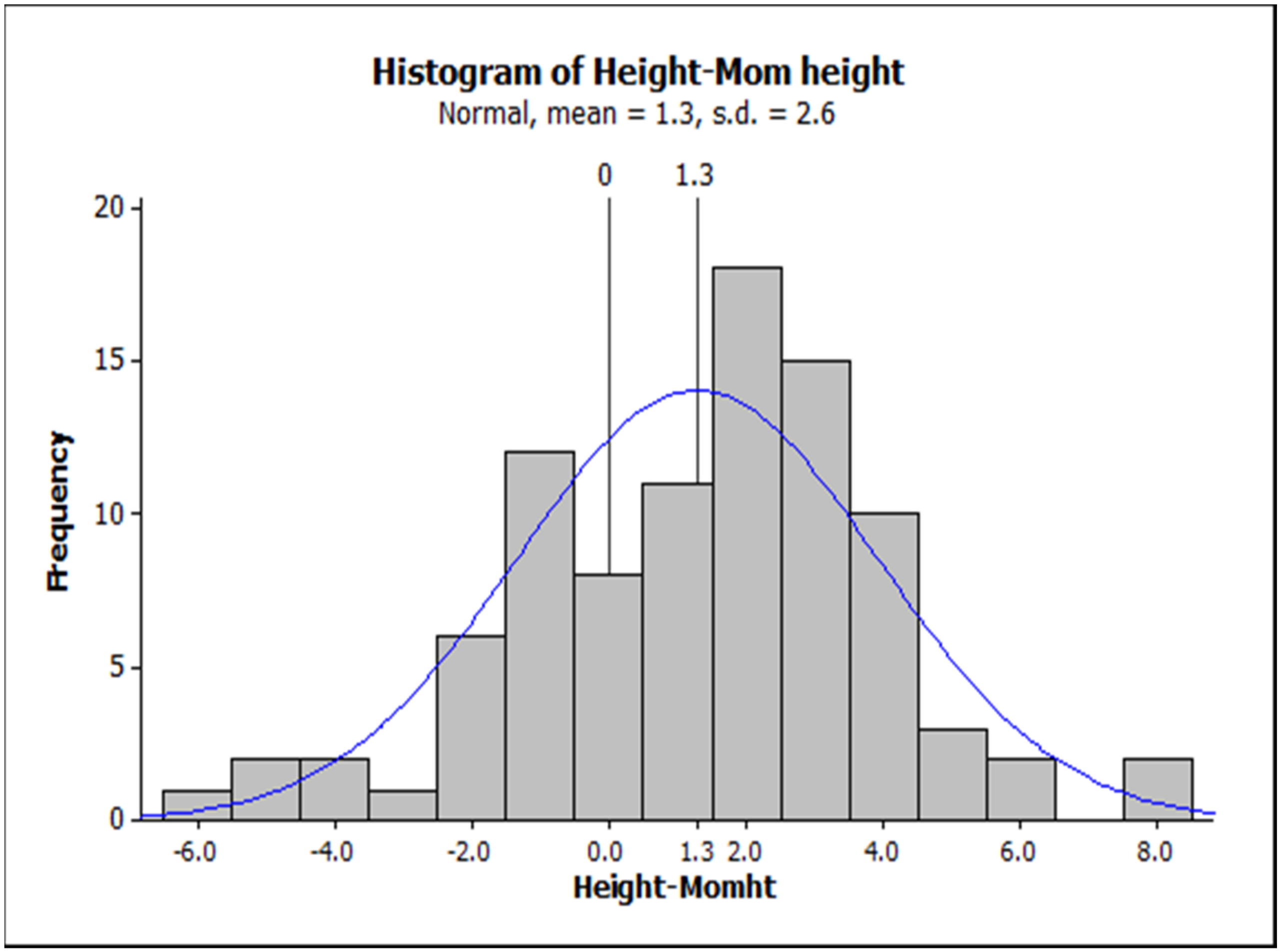
Effect size for comparing heights

- Suppose you want to compare the heights of college women and their mothers to see if the average heights are equal.
- Measure *n* pairs and find differences.
- Hypotheses:
  - Null: Mean of *population* of differences = 0
  - Alternative: Mean of population is > 0
- Effect size = True difference/(Std. dev.) = number of standard deviations true difference is from 0.

Effect size, continued

- Estimated effect size = $\frac{Sample\ mean\ difference}{Std.dev.of\ differences}$
- Test statistic is $t = \sqrt{n} \times$ Est. effect size
- Example: Data from my class
  - $n = 93$ pairs, mean diff = 1.30 in., s.d. = 2.6 in.
  - Estimated effect size = 1.3/2.6 = 0.5
  - Test statistic is $t = \sqrt{93} \times 0.5 = 4.8$, p-value ≈ 0
  - Conclude women students today are taller than their mothers, on average.

Illustration of effect size



Mean of 1.3 is 0.5 standard deviations above null value of 0.

Cohen’s suggested guidelines for a Small, medium, large effect size

- 0.2 is a small effect size and can only be detected using statistics
- 0.5 is a moderate effect size and can be detected by someone used to working with that type of data (Ex: difference in heights)
- 0.8 is a large effect size and should be detectable without statistics
- Note: Ganzfeld hit rate of .33 is effect size of about 0.18, so it’s a small effect size.

Hypothesis testing paradox: Effect size versus *p*-value

- Researcher conducts test with $n = 100$ and finds $t = 2.50$, *p*-value = 0.014, reject null
- Just to be sure, repeats with $n = 25$
- Uh-oh, finds $t = 1.25$, *p*-value = 0.22, cannot reject null! The effect has disappeared!
- To salvage, decides to combine data, so now $n = 125$. Finds $t = 2.795$, *p*-value = 0.006!
- Paradox: The 2nd study alone did not replicate finding, but when combined with 1st study, the effect seems even stronger than 1st study!

What's going on?

- The test statistic and *p*-value depend on the sample size.
- Both studies have the same effect size
- Combined data also has that effect size
  - effect size is test statistic/$\sqrt{n}$

Study	***n***	**Test statistic**	***P*-value**	**Effect size**
1	100	2.50	0.014	0.25
2	25	1.25	0.22	0.25
Combined	125	2.795	0.006	0.25

Why Effect Sizes are Important

- Unlike *p*-values, they don't depend on sample size (but accuracy of estimating them does).
- They are a measure of the true effect or difference in the population.
- They can be compared even when different units or different tests are used.
- Replication should be defined as getting approximately the same effect size, *not* as getting approximately the same *p*-value!

Bayesian Analysis

- Completely different statistical “model”
- Frequentist method: Parameters, such as binomial probability of success, are considered fixed but unknown.
- Bayesian method: Uncertainty about parameters is modeled by putting a distribution of possibilities on them.
- Prior belief in null vs alternative hypothesis is stated explicitly.

How to Incorporate Prior Beliefs

- Two ways, *both* required in a Bayesian analysis:
  - What do you think is the probability that the alternative hypothesis (psi) is true?
  - *If* the psi hypothesis is true, how large do you think the effect size is? (Or, what do you think is the probability of a hit?)
- This 2nd question is often ignored in doing Bayesian analysis. Can be very misleading if not done right! And, can be *hidden* in the analysis.

More Details

Simple Bayesian analysis of Ganzfeld:

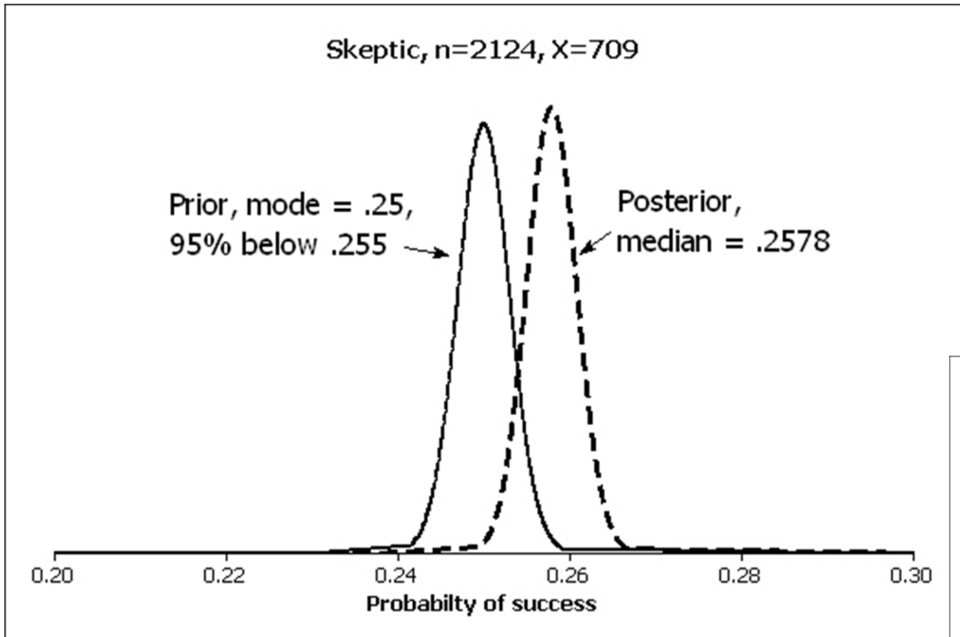
- "Prior" distribution on the hit rate provides the range of values one believes it *could* be, along with how likely they are.
- Combine prior distribution with data to get a "posterior" distribution for the hit rate.

Utts, Norris, Suess, Johnson (ICOTS 8) 56 studies, $n = 2124$, $X = 709$ (33.4%)

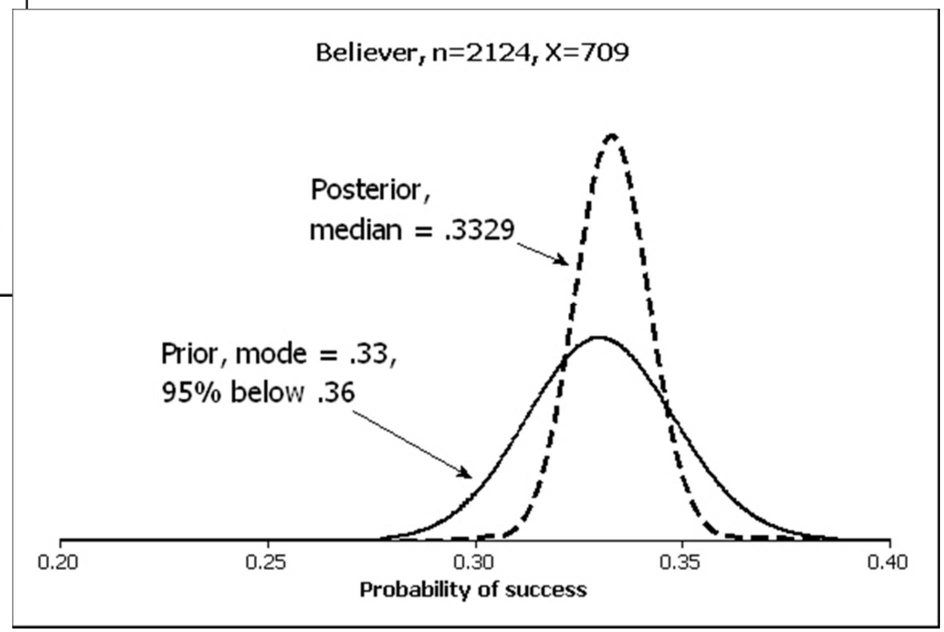
Simple analysis: 3 Prior Sets of Belief about p

- 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

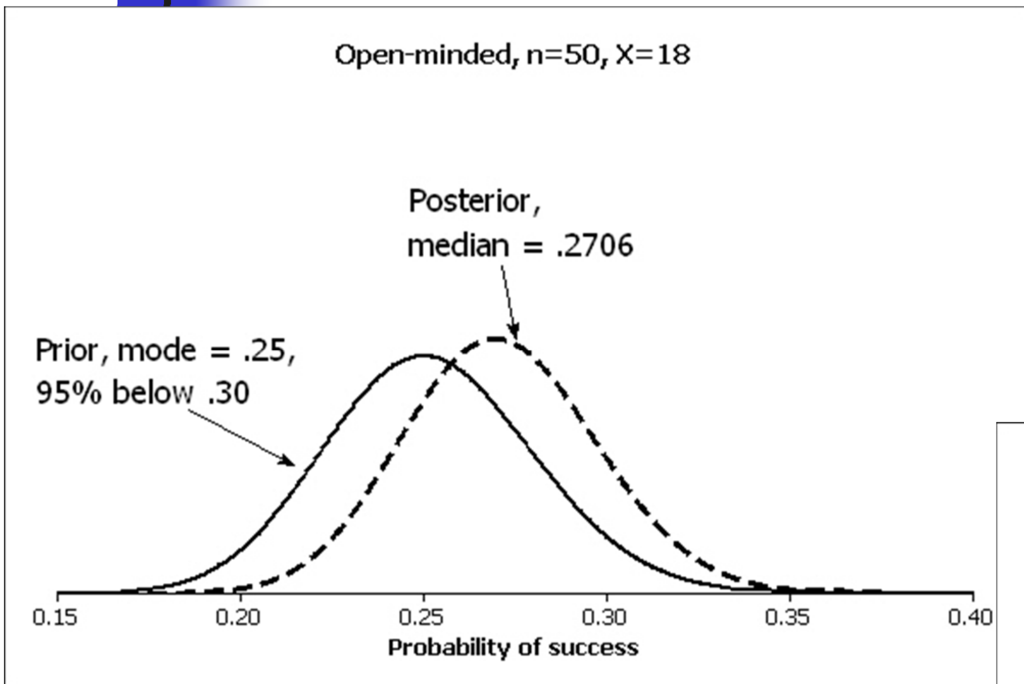


Data reduced the range of the believer's likely values for *p*



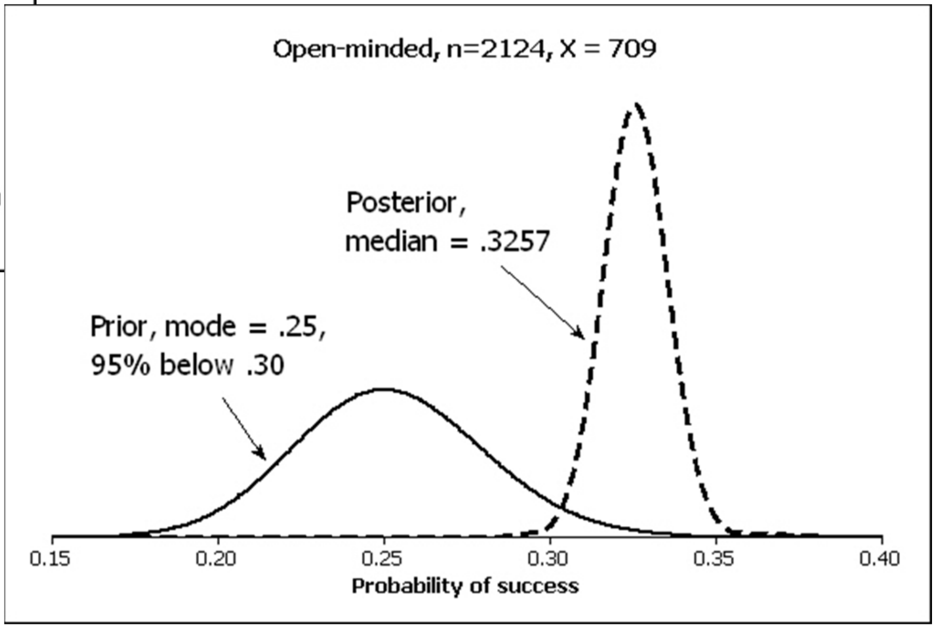
Data shifted the skeptic's belief very slightly. Posterior median = .2578

Open-minded: One study and all data



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

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



Summary of Simple Bayesian Analysis (ICOTS paper for more complex 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 extreme 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!

Bayesian Analyses of Bem’s experiments Wagemakers et al; Bem, Utts, Johnson

- Wagenmakers et al put prior probability on the psi hypothesis = $10^{-20} \approx 0$!
- Then, they used a prior distribution on values in the alternative with too much weight on large effects:
  - 57% chance that the true effect exceeds Cohen’s “large” effect size of 0.8 (hit rate about 63%)
  - 6% chance that it exceeds effect size of 10 (hit rate greater than 1)!

Bayesian Analyses of Bem's experiments Continued…

- So of course for that prior, data came closer to null than to this unrealistic alternative.
- We used more reasonable prior, putting 90% chance of effect size being less than .5 (hit rate of about 48%).

Bayesian Results

- Bayes Factor = Odds of alternative versus null, assuming equal prior belief:
  - Wagenmakers et al too-wide prior: 0.632 to 1
  - Our (more realistic) prior: 13,669 to 1
  - Multiply by *your* prior odds to get posterior odds
- Posterior probability of true null in all 9 studies:
  - Wagenmakers et al's too-wide prior: 0.61
  - Bem et al's realistic prior: 7.3×10^{-5}
  - Using p-values: 2.68×10^{-11} (two-tailed)

Summary

- Hypothesis tests, confidence intervals and Bayesian analysis are all methods for assessing the evidence.
- Unless the null hypothesis is exactly true, hypothesis test *p*-values depend on *n*.
- Effect sizes are a better way to measure the magnitude of an effect than testing.
- Bayesian methods require explicit statement of one's beliefs – that's why I like them!

QUESTIONS?

Contact info:

jutts@uci.edu

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

UCIrvine

UNIVERSITY OF CALIFORNIA, IRVINE