Bayesian and Traditional Hypothesis Testing

School of Information Studies
Syracuse University

Learning Topics for This Week

Confidence intervals in context

Bayesian thinking about data

Opening up Bayes' theorem

Bayesian thinking applied to mtcars

Searching for posterior probabilities

- Markov chain / Monte Carlo method

Posterior distribution of mean differences

- Highest density interval

The null hypothesis significance test

Confidence Intervals in Context

Confidence Intervals in Context

Last week we considered a comparison between two sample means

Samples never provide a clear or perfect picture of the underlying population

Confidence intervals give a basic picture of uncertainty in the form of the width of the interval around the difference in means

This week we will develop a more sophisticated view of uncertainty with the help of Bayesian thinking

Bayesian Thinking

A Return to Toast

	Down	Up
Jelly	2	1
Butter	3	4

	Down	Up	Marginal
Jelly	0.2	0.1	0.3
Butter	0.3	0.4	0.7
Marginal	0.5	0.5	

Contingency Tables Are Bayesian Thinking

Bayesian thinking describes how we update our probabilities given new information.

Here's new information: There's a piece of toast lying on the ground with the topping facing down, what's the probability that the topping is jelly?

Observing toast-down gives new information.

New information updates our beliefs about jelly vs. butter.

	Down	Up	Marginal
Jelly	0.2	0.1	0.3
Butter	0.3	0.4	0.7
Marginal	0.5	0.5	

p(jelly|down) vs. p(down|jelly)

p(jelly|down): The probability of observing jelly, given that the toast is facing down

p(jelly|down) = 0.2 / 0.5 = 0.4
(using the marginal calculation method from chapter 2)

Notice that p(down|jelly)=0.2/0.3, where the denominator is the marginal probability of jelly

	Down	Up	Marginal
Jelly	0.2	0.1	0.3
Butter	0.3	0.4	0.7
Marginal	0.5	0.5	

Opening Up Bayes' Theorem

Bayes Theorem: Same Toast, New Formula

Our Toast Example: What's the probability that the topping is jelly, given that the toast is down?

$$p(\text{jelly} \mid \text{down}) = \frac{p(\text{down} \mid \text{jelly}) * p(\text{jelly})}{p(\text{down})} = 0.04$$

Same result as the contingency table method!

The probability of observing toast down, given it is jelly:
p(down | jelly) = 0.2/0.3 = 0.667 (where the 0.3 is the marginal total for jelly)

Then, p(jelly) is the marginal total for jelly (0.3)

And p(down) is the marginal for toast down (0.5)

	Down	Up	Marginal
Jelly	0.2	0.1	0.3
Butter	0.3	0.4	0.7
Marginal	0.5	0.5	

Bayes Theorem: Same Toast, New Formula

The "official" Bayes theorem:

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}$$

We're seeking p(H|D), the probability of the hypothesis being "true" given data we have observed; we need three ingredients in order to figure this out:

- The probability of observing these data when the hypothesis is "true," p(D|H), which Bayesian folks call the **likelihood**
- The **prior probability**, p(H), in other words our baseline belief about the "truth" of the hypothesis
- The probability of observing these data under any and all conditions, p(D), or in Bayes-speak, the **evidence**

Why so Darned Complicated?

We can restate our mathematical equation with the Bayesian words filled in:

$$posterior = \frac{likelihood \cdot prior}{evidence} = prior\left(\frac{likelihood}{evidence}\right)$$

Bayes wanted to be able to address the general case and, in particular, situations where the prior belief was just a qualitative statement about how likely something was

In our toast example, we know the precise marginal likelihood of p(jelly)=0.3 with the result that it cancels itself out—the 0.3 value is both p(jelly) and the denominator of p(down | jelly)

That leaves 0.2/0.5 = 0.4 which is precisely the same as the marginal method from Chapter 2

If our prior belief is more along the lines of "kinda likely" or "not too likely," we're less concerned with Bayes theorem as a math formula and more interested in the general idea of transforming a prior belief into a posterior probability

Bayesian Thinking Applied to mtcars

For Example, Back to the mtcars Data

In the rows, we have different possible prior beliefs about whether the type of a car's transmission makes any difference to fuel economy. We could ask some mechanics about this.

In the columns we have the observations we might obtain from measuring cars, in some cases a difference of greater than 3 mpg between automatics and manuals, and in other cases not. The choice of 3 mpg is arbitrary: What would be a large enough observed difference to matter to you?

	> 3 MPG	< 3 MPG	Row totals
Transmissions do matter			
Transmissions don't matter			
Column totals			

For Example, Back to the mtcars Data

Bayesian thinking gives us a way of reasoning about a hypothesis in the light of new evidence. Bayes' theorem transforms our prior beliefs into posterior probabilities using evidence from new data.

	> 3 MPG	< 3 MPG	Row totals
Transmissions do matter			
Transmissions don't matter			
Column totals			

Searching for Posterior Probabilities

Searching for Posterior Probabilities

When we calculated the confidence interval using t.test(), there was precisely one answer that was obtained from a simple mathematical formula

With Bayesian thinking, we generally do not have a precise value for the prior probability in mind—we must test a range of possible values; the computations are complex

So, rather than applying a simple mathematical formula to calculate the result, we "search" through a "space" of possible results, looking for the results that are most likely

This search process is called "Markov Chain Monte Carlo"

Applying Bayesian Thinking to Statistics With a Walking Robot

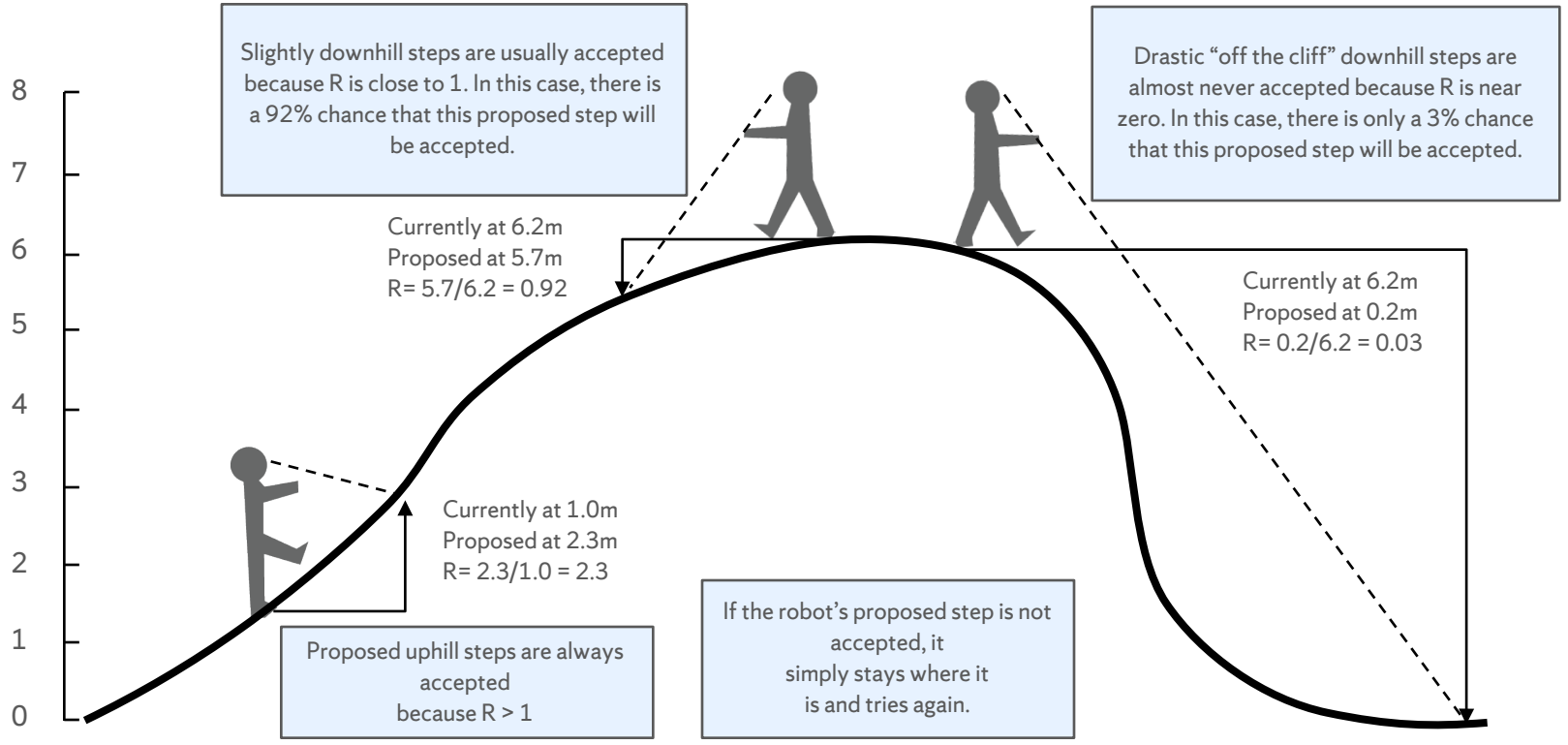
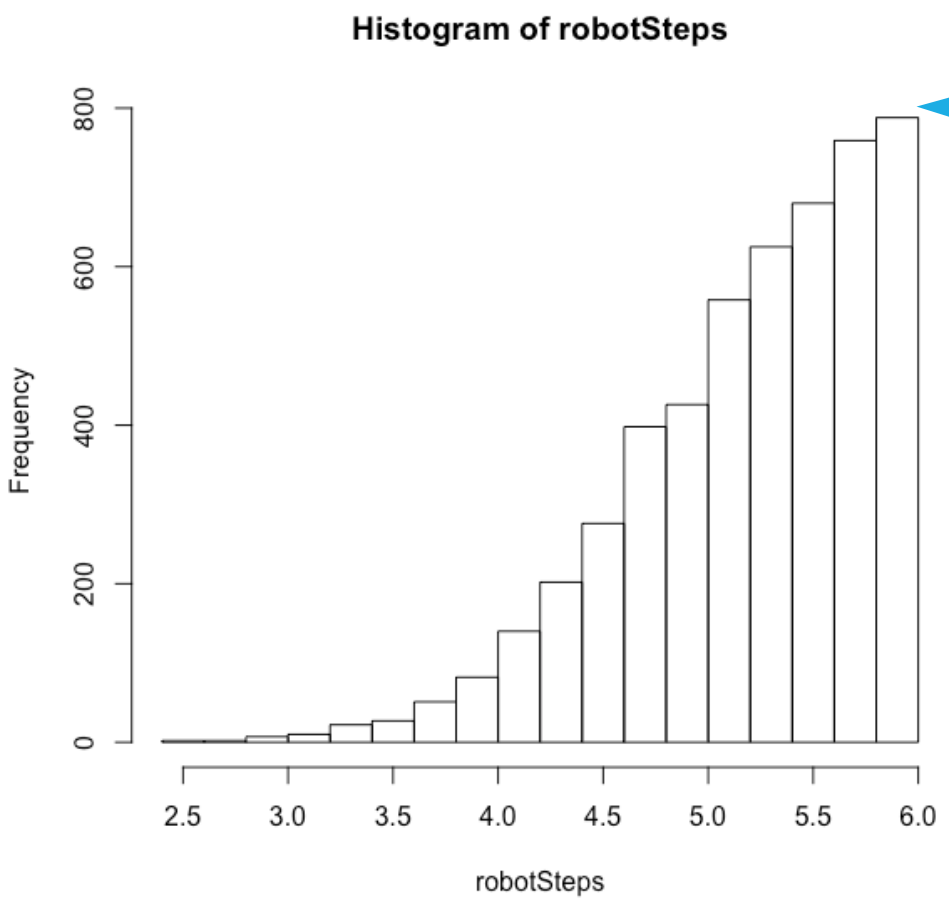


Image credit: Paul O. Lewis, http://marple.eeb.uconn.edu/mcmcrobot/?page_id=24

Applying Bayesian Thinking to Statistics With a Walking Robot



The robot spent the most amount of time walking around where the altitude was about 6.0

Posterior Distribution of Mean Differences

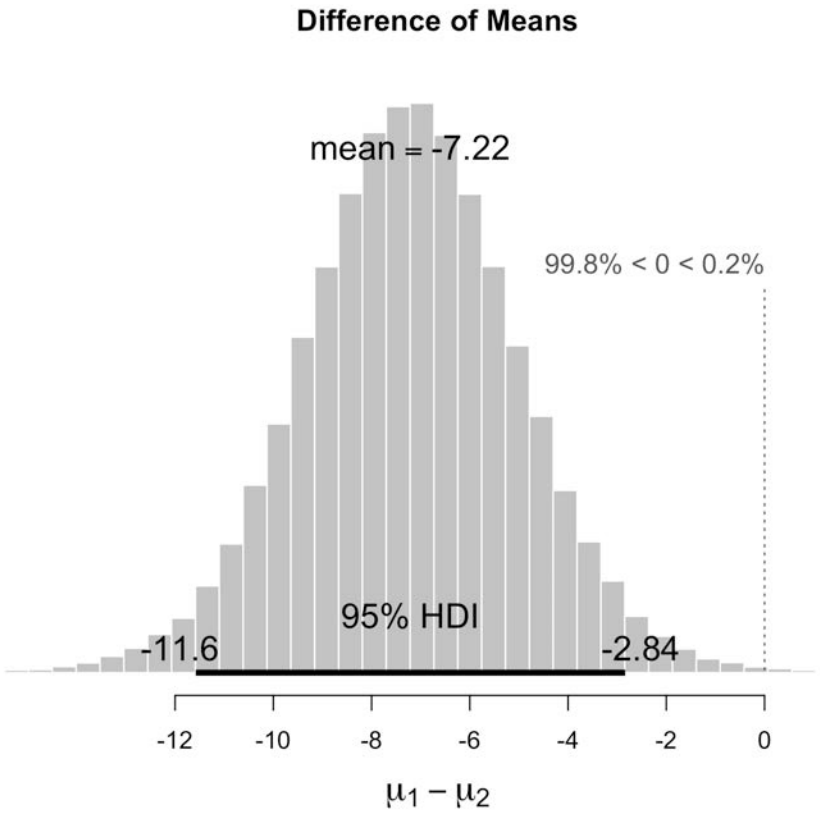
The Landscape of Mean Differences

The output on the right shows the results of a Markov Chain-Monte Carlo analysis of mean differences between automatic and manual transmissions.

The MCMC "robot" took 100,000 steps through a landscape of possible mean differences.

For each proposed step, the robot examined the mtcars data to see if it seemed to be headed in the right direction.

The result is a posterior distribution of possible values for the mean difference. What's the most likely posterior value?



Important Distinctions: HDIs vs. CIs

Highest Density Interval

The Bayesian analysis shows that the "Highest Density Interval" or HDI ranges from -11.6 up to -2.84 (miles per gallon).

This HDI has an intuitive interpretation: There is a 95% probability that the population mean difference between the two groups falls within this range. The histogram on the previous slide shows that the most likely value is -7.22.

These are statements about the posterior distribution of possible values of the mean difference.

Confidence Interval

From the analysis we conducted in Chapter 4, the 95 percent confidence interval ranged from -11.3 up to -3.2 (miles per gallon).

Statisticians say that if we could replicate our whole study 100 times, on average in 95 of those replications the calculated confidence interval would contain the actual population mean difference.

This is a statement about the long run possibilities, not about the accuracy of this particular confidence interval.

The NHST

The Null Hypothesis Significance Test: Setup

An important reasoning tool used throughout the 20th century:

- Begin by asserting a null hypothesis that there is no mean difference between the means of two groups
- The "opposite" of the null hypothesis is the alternative hypothesis
- Choose an "alpha level" probability, beyond which the null hypothesis will be rejected—a common alpha level is 0.05; a more stringent level could be 0.01 or 0.005
- Collect data and conduct a statistical test, such as the t-test of two independent means, to calculate a significance value, designated by the letter "p"

The Null Hypothesis Significance Test: Evaluate

If the calculated value of p is *less than the alpha level* that was chosen above, for example if p = 0.049 when alpha was chosen as 0.05, then *reject* the null hypothesis

When the null hypothesis is rejected, this can be considered evidence in favor of an alternative hypothesis, though the results of that significance test do not say what that alternative hypothesis might be, or the probability that any particular alternative hypothesis may be correct

If p is *greater than the alpha level* that was chosen above, for example if p = 0.051 when alpha was chosen as 0.05, then *fail to reject* the null hypothesis; failing to reject the null hypothesis does not mean that we accept the null hypothesis, rather that we have no good evidence either way; likewise, the p-value does not inform the question of how likely the null hypothesis is

The NHST Graphically: The T-Test of mtcars

The graph at right shows 10,000 random values of the t-distribution, very similar to the normal distribution (slightly thicker tails).

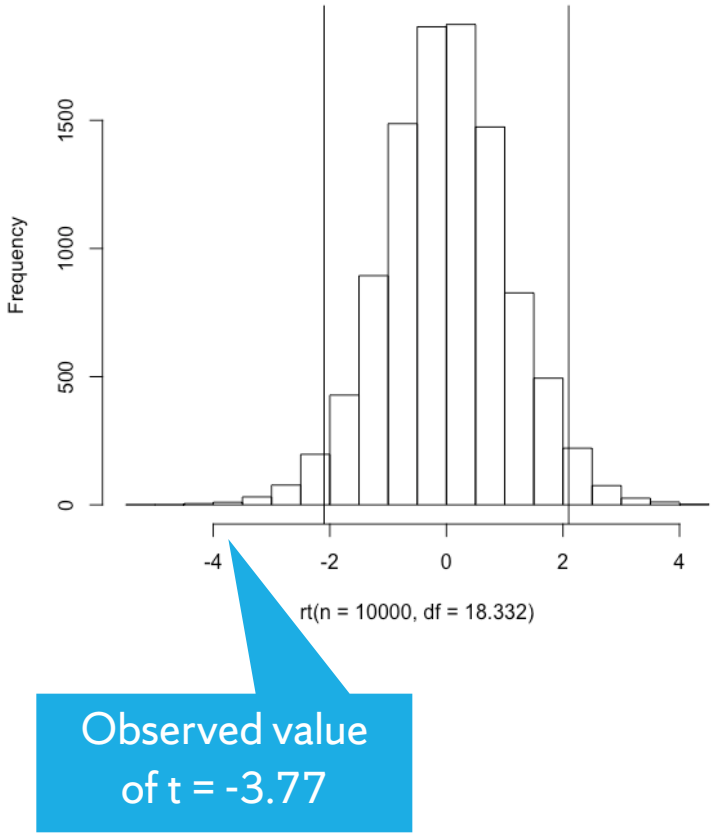
The distribution is centered on zero, reflecting our basic assumption of the null hypothesis.

The two vertical lines divide the graph into a central region with 95% of all values of t and two tails with 2.5% each.

The observed mean difference from the mtcars data of -7.2 mpg has been converted to a t-value of -3.77 using the formula for t.

The value t=-3.77 falls in the left hand tail of the distribution, beyond the left vertical. The associated p value is 0.001374.

We reject the null hypothesis.



Some Concerns About the NHST

Results published by Bohannon (2015) in Science suggested that perhaps fewer than 40% of articles published in the field of psychology could be replicated with similar statistical results and using the same research design.

Anyone who has spent time conducting statistical tests on a small data set knows that the addition or removal of just a few cases/observations can make the difference between p = 0.051 and p = 0.049. There's a term for this idea—p-hacking (Nuzzo, 2014).

Our system of publishing scientific studies has pushed researchers in the direction of seeking statistical significance, but having tests that simply tell us whether a difference is statistically significant does not show us whether a result is meaningful.

For researchers or engineers to start with a basic assumption that there would be no difference between the two types of transmissions gives us only a very low bar to cross in order to provide statistical support.

Fun With rjags