Introduction to Bayesian inference

Joachim Vandekerckhove

Prefacing notes: Reference materials

- For a theoretical introduction to Bayesian statistics, the standard work is "Bayesian Data Analysis" by Andrew Gelman et al.

Prefacing notes: Reference materials

- For a theoretical introduction to Bayesian statistics, the standard work is “Bayesian Data Analysis” by Andrew Gelman et al.
- For a first-principles journey into statistical reasoning and logic, I really like “Probability Theory: The Logic of Science” by Edwin T. Jaynes

Prefacing notes: Reference materials

- For a theoretical introduction to Bayesian statistics, the standard work is “Bayesian Data Analysis” by Andrew Gelman et al.
- For a first-principles journey into statistical reasoning and logic, I really like “Probability Theory: The Logic of Science” by Edwin T. Jaynes
- A practical course in Bayesian statistics is “A Course in Bayesian Graphical Modeling for Cognitive Science” by Michael D. Lee and Eric-Jan Wagenmakers
  http://www.bayesmodels.com/

Prefacing notes: Reference materials

- For a theoretical introduction to Bayesian statistics, the standard work is "Bayesian Data Analysis" by Andrew Gelman et al.
- For a first-principles journey into statistical reasoning and logic, I really like "Probability Theory: The Logic of Science" by Edwin T. Jaynes
- A practical course in Bayesian statistics is "A Course in Bayesian Graphical Modeling for Cognitive Science" by Michael D. Lee and Eric-Jan Wagenmakers http://www.bayesmodels.com/
- I have not yet read Richard McElreath's "Statistical Rethinking" but I hear very good things

Prefacing notes: Reference materials

- For a theoretical introduction to Bayesian statistics, the standard work is “Bayesian Data Analysis” by Andrew Gelman et al.
- For a first-principles journey into statistical reasoning and logic, I really like “Probability Theory: The Logic of Science” by Edwin T. Jaynes
- A practical course in Bayesian statistics is “A Course in Bayesian Graphical Modeling for Cognitive Science” by Michael D. Lee and Eric-Jan Wagenmakers http://www.bayesmodels.com/
- I have not yet read Richard McElreath’s “Statistical Rethinking” but I hear very good things
- Special issue of Psychonomic Bulletin & Review (volume 25, 2018)

Prefacing notes: This is not psychology

- Bayesian statistics is a set of formal methods for statistical inference, used by statisticians and scientists to make statements about unobserved parameters starting from the observed data

Prefacing notes: This is not psychology

- Bayesian statistics is a set of formal methods for statistical inference, used by statisticians and scientists to make statements about unobserved parameters starting from the observed data
- Not to be confused with *"Bayes-in-the-head"*, a set of psychological theories about how lay humans perform inference in daily life

Prefacing notes: What is probability?

- We will be dealing with the concept of *probability*, which is in some sense ambiguous

Prefacing notes: What is probability?

- We will be dealing with the concept of *probability*, which is in some sense ambiguous
- In one meaning—the Bayesian meaning—probability is a *degree of belief*

Prefacing notes: What is probability?

- We will be dealing with the concept of *probability*, which is in some sense ambiguous
- In one meaning—the Bayesian meaning—probability is a *degree of belief*
- In another meaning—the classical, or frequentist meaning—probability is a statement of *expected frequency over many repetitions*

Prefacing notes: What is probability?

- This distinction is directly relevant for empirical psychology

Prefacing notes: What is probability?

- This distinction is directly relevant for empirical psychology
- In the overwhelming majority of cases, psychologists are interested in making probabilistic statements about singular events: an hypothesis is either true or not; an effect is either zero or not; the effect size is likely to be between X and Y; either the one model or the other is more likely given the data. . .

Prefacing notes: What is probability?

- This distinction is directly relevant for empirical psychology
- In the overwhelming majority of cases, psychologists are interested in making probabilistic statements about singular events: an hypothesis is either true or not; an effect is either zero or not; the effect size is likely to be between X and Y; either the one model or the other is more likely given the data...
- We are not usually interested in the frequency with which a well-defined process will achieve a certain outcome

The Sum and Product Rules of probability

Some notational conventions:

- (A) is an *event*: a statement that can be true or false

Some notational conventions:

- (A) is an *event*: a statement that can be true or false
- $(A|B)$ is the *conditional* event: A is true *if B is true*

Some notational conventions:

- (A) is an *event*: a statement that can be true or false
- $(A|B)$ is the *conditional* event: A is true *if B is true*
    - In other words, *B implies A*: $B \rightarrow A$

Some notational conventions:

- (A) is an *event*: a statement that can be true or false
- $(A|B)$ is the *conditional* event: A is true *if B is true*
  - In other words, B *implies* A: $B \rightarrow A$
  - Independence means $P(A) = P(A|B)$

Some notational conventions:

- (A) is an *event*: a statement that can be true or false
- $(A|B)$ is the *conditional* event: A is true *if B is true*
  - In other words, B *implies* A: $B \to A$
  - Independence means $P(A) = P(A|B)$
- (A, B) is the *joint* event: A and B are both true

Some notational conventions:

- (A) is an *event*: a statement that can be true or false
- $(A|B)$ is the *conditional* event: A is true *if B is true*
  - In other words, B *implies* A: $B \to A$
  - Independence means $P(A) = P(A|B)$
- (A, B) is the *joint* event: A and B are both true
  - Of course $P(A, B) = P(B, A)$

Some notational conventions:

- (A) is an *event*: a statement that can be true or false
- $(A|B)$ is the *conditional* event: A is true *if B is true*
  - In other words, B *implies* A: $B \rightarrow A$
  - Independence means $P(A) = P(A|B)$
- (A, B) is the *joint* event: A and B are both true
  - Of course $P(A, B) = P(B, A)$
- $(\neg A)$ is the negation of A: A is false

Some notational conventions:

- (A) is an *event*: a statement that can be true or false
- $(A|B)$ is the *conditional* event: A is true *if B is true*
  - In other words, B *implies* A: $B \rightarrow A$
  - Independence means $P(A) = P(A|B)$
- (A, B) is the *joint* event: A and B are both true
  - Of course $P(A, B) = P(B, A)$
- $(\neg A)$ is the negation of A: A is false

Some notational conventions:

- (A) is an *event*: a statement that can be true or false
- $(A|B)$ is the *conditional* event: A is true *if B is true*
  - In other words, B *implies* A: $B \rightarrow A$
  - Independence means $P(A) = P(A|B)$
- (A, B) is the *joint* event: A and B are both true
  - Of course $P(A, B) = P(B, A)$
- $(\neg A)$ is the negation of A: A is false

These notations can be combined: $P(A, B|\neg C, \neg D)$ is the probability that A and B are both true assuming that C and D are both false.

The Product Rule of probability

With this notation in mind, we introduce the **Product Rule of probability**:

$$P(A, B) = P(B)P(A|B)$$

In words: the probability that A and B are both true is equal to the probability of B multiplied by the conditional probability of A *assuming B is true*.

The Product Rule of probability

With this notation in mind, we introduce the **Product Rule of probability**:

$$P(A, B) = P(B)P(A|B)$$

In words: the probability that A and B are both true is equal to the probability of B multiplied by the conditional probability of A *assuming B is true*.

Of course,

$$\begin{aligned} P(A, B) &= P(B)P(A|B) \\ = P(B, A) &= P(A)P(B|A). \end{aligned}$$

The Sum Rule of probability

The **Sum Rule of probability** requires one further concept: the *disjunctive set*:

- A disjunctive set is a collection of events, exactly one of which must be true. For example:

The Sum Rule of probability

The **Sum Rule of probability** requires one further concept: the *disjunctive set*:

- A disjunctive set is a collection of events, exactly one of which must be true. For example:
  - An event and its denial: $\{B, \neg B\}$

The Sum Rule of probability

The **Sum Rule of probability** requires one further concept: the *disjunctive set*:

- A disjunctive set is a collection of events, exactly one of which must be true. For example:
  - An event and its denial: $\{B, \neg B\}$
  - The possible outcomes of a coin flip: {heads, tails}

The Sum Rule of probability

The **Sum Rule of probability** requires one further concept: the *disjunctive set*:

- A disjunctive set is a collection of events, exactly one of which must be true. For example:
  - An event and its denial: $\{B, \neg B\}$
  - The possible outcomes of a coin flip: {heads, tails}
  - The possible outcomes of a roll of a six-sided die: {1, 2, 3, 4, 5, 6}

The Sum Rule of probability

The **Sum Rule of probability** requires one further concept: the *disjunctive set*:

- A disjunctive set is a collection of events, exactly one of which must be true. For example:
  - An event and its denial: $\{B, \neg B\}$
  - The possible outcomes of a coin flip: {heads, tails}
  - The possible outcomes of a roll of a six-sided die: {1, 2, 3, 4, 5, 6}
  - The truth of some hypothesis H, which must be either true or false: $\{H, \neg H\}$

The Sum Rule of Probability

- Suppose B represents the event "It will rain tomorrow"

The Sum Rule of Probability

- Suppose B represents the event "It will rain tomorrow"
  - Then $\neg B$ represents the event "It will not rain tomorrow"

The Sum Rule of Probability

- Suppose B represents the event "It will rain tomorrow"
  - Then $\neg B$ represents the event "It will not rain tomorrow"
  - One and only one of these events can occur, so $\{B, \neg B\}$ forms a disjunctive set.

The Sum Rule of Probability

- Suppose B represents the event "It will rain tomorrow"
  - Then $\neg B$ represents the event "It will not rain tomorrow"
  - One and only one of these events can occur, so $\{B, \neg B\}$ forms a disjunctive set.
- Suppose A represents the event "It will rain today"

The Sum Rule of Probability

- Suppose B represents the event “It will rain tomorrow”
  - Then $\neg B$ represents the event “It will not rain tomorrow”
  - One and only one of these events can occur, so $\{B, \neg B\}$ forms a disjunctive set.
- Suppose A represents the event “It will rain today”
  - $\neg A$ represents “It will not rain today”

The Sum Rule of Probability

- Suppose B represents the event "It will rain tomorrow"
  - Then $\neg B$ represents the event "It will not rain tomorrow"
  - One and only one of these events can occur, so $\{B, \neg B\}$ forms a disjunctive set.
- Suppose A represents the event "It will rain today"
  - $\neg A$ represents "It will not rain today"
  - $\{A, \neg A\}$ is another disjunctive set

The Sum Rule of Probability

- Suppose B represents the event "It will rain tomorrow"
  - Then $\neg B$ represents the event "It will not rain tomorrow"
  - One and only one of these events can occur, so $\{B, \neg B\}$ forms a disjunctive set.
- Suppose A represents the event "It will rain today"
  - $\neg A$ represents "It will not rain today"
  - $\{A, \neg A\}$ is another disjunctive set
- Now there are four possible combinations of *joint events*: (A, B), $(A, \neg B)$, $(\neg A, B)$, and $(\neg A, \neg B)$

The Sum Rule of Probability

- Suppose B represents the event "It will rain tomorrow"
  - Then $\neg B$ represents the event "It will not rain tomorrow"
  - One and only one of these events can occur, so $\{B, \neg B\}$ forms a disjunctive set.
- Suppose A represents the event "It will rain today"
  - $\neg A$ represents "It will not rain today"
  - $\{A, \neg A\}$ is another disjunctive set
- Now there are four possible combinations of *joint events*: (A, B), $(A, \neg B)$, $(\neg A, B)$, and $(\neg A, \neg B)$
  - These again form a disjunctive set

The Sum Rule of Probability

- With that in mind, we can formulate a simplified Sum Rule

The Sum Rule of Probability

- With that in mind, we can formulate a simplified Sum Rule
- The probability of a single one of these events alone, say A, can be found by adding up the probabilities of all of the joint events that contain A as follows:

$$P(A) = P(A, B) + P(A, \neg B)$$

The Sum Rule of Probability

- With that in mind, we can formulate a simplified Sum Rule
- The probability of a single one of these events alone, say A, can be found by adding up the probabilities of all of the joint events that contain A as follows:

$$P(A) = P(A, B) + P(A, \neg B)$$

- In words, the probability it rains today is the sum of two joint probabilities: (1) the probability it rains today and tomorrow, and (2) the probability it rains today but not tomorrow.

The Sum Rule of Probability

$$P(A) = P(A, B) + P(A, \neg B)$$

is a simplification of the Sum Rule.

The Sum Rule of Probability

$$P(A) = P(A, B) + P(A, \neg B)$$

is a simplification of the Sum Rule.

In general, if $\{B_1, B_2, \ldots, B_K\}$ is a disjunctive set, the **Sum Rule of probability** states

$$P(A) = \sum_{k=1}^{K} P(A, B_k).$$

The Sum Rule of Probability

$$P(A) = P(A, B) + P(A, \neg B)$$

is a simplification of the Sum Rule.

In general, if $\{B_1, B_2, \ldots, B_K\}$ is a disjunctive set, the **Sum Rule of probability** states

$$P(A) = \sum_{k=1}^{K} P(A, B_k).$$

That is, the probability of event A alone is the sum of all the joint probabilities between A and the elements of a disjunctive set.

What is Bayesian inference?

What is Bayesian inference?

Bayesian inference is the application of the product and sum rules of probability to real problems of inference.

What is Bayesian inference?

Bayesian inference is the application of the product and sum rules of probability to real problems of inference.

"The sum and product rules. . . are a factory for building methods of inference" (-anon.)

What is Bayesian inference?

Bayesian inference is the application of the product and sum rules of probability to real problems of inference.

"The sum and product rules... are a factory for building methods of inference" (-anon.)

Together, these two rules allow us to calculate probabilities in an incredible variety of circumstances. One combination of the two rules in particular is useful for scientific inference is *hypothesis testing*.

Hypothesis testing

- Call event H (the truth of) an hypothesis that a researcher holds and call $\neg H$ a competing hypothesis

Hypothesis testing

- Call event H (the truth of) an hypothesis that a researcher holds and call $\neg H$ a competing hypothesis
- Before any data are collected, the researcher has some level of prior belief in these competing hypotheses, which manifest as *prior probabilities* and are denoted $P(H)$ and $P(\neg H)$

Hypothesis testing

- Call event H (the truth of) an hypothesis that a researcher holds and call $\neg H$ a competing hypothesis
- Before any data are collected, the researcher has some level of prior belief in these competing hypotheses, which manifest as *prior probabilities* and are denoted $P(H)$ and $P(\neg H)$
- The hypotheses are well-defined and make specific predictions about the outcome of an experiment, $P(X|H)$ and $P(X|\neg H)$, where event X denotes the data collected in an experiment

Hypothesis testing

- Call event H (the truth of) an hypothesis that a researcher holds and call $\neg H$ a competing hypothesis
- Before any data are collected, the researcher has some level of prior belief in these competing hypotheses, which manifest as *prior probabilities* and are denoted $P(H)$ and $P(\neg H)$
- The hypotheses are well-defined and make specific predictions about the outcome of an experiment, $P(X|H)$ and $P(X|\neg H)$, where event X denotes the data collected in an experiment
  - $P(X|H)$ is called the *likelihood function* and can be thought of as how strongly the data are implied by an hypothesis

Hypothesis testing

- Call event H (the truth of) an hypothesis that a researcher holds and call $\neg H$ a competing hypothesis
- Before any data are collected, the researcher has some level of prior belief in these competing hypotheses, which manifest as *prior probabilities* and are denoted $P(H)$ and $P(\neg H)$
- The hypotheses are well-defined and make specific predictions about the outcome of an experiment, $P(X|H)$ and $P(X|\neg H)$, where event X denotes the data collected in an experiment
  - $P(X|H)$ is called the *likelihood function* and can be thought of as how strongly the data are implied by an hypothesis
  - *Conditional* on the truth of an hypothesis, likelihood functions specify the probability of a given outcome and are usually only interpretable in relation to other hypotheses' likelihoods

Hypothesis testing

- Call event H (the truth of) an hypothesis that a researcher holds and call $\neg H$ a competing hypothesis
- Before any data are collected, the researcher has some level of prior belief in these competing hypotheses, which manifest as *prior probabilities* and are denoted $P(H)$ and $P(\neg H)$
- The hypotheses are well-defined and make specific predictions about the outcome of an experiment, $P(X|H)$ and $P(X|\neg H)$, where event X denotes the data collected in an experiment
  - $P(X|H)$ is called the *likelihood function* and can be thought of as how strongly the data are implied by an hypothesis
  - *Conditional* on the truth of an hypothesis, likelihood functions specify the probability of a given outcome and are usually only interpretable in relation to other hypotheses' likelihoods
- Of interest is the probability that H is true, given the data X, or $P(H|X)$.

Hypothesis testing

From $P(H, X) = P(X)P(H|X)$, we can derive that

$$P(H|X) = \frac{P(H, X)}{P(X)}.$$

Hypothesis testing

From $P(H, X) = P(X)P(H|X)$, we can derive that

$$P(H|X) = \frac{P(H, X)}{P(X)}.$$

And since it is also true that $P(H, X) = P(H)P(X|H)$, we see that this is equivalent to

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)}.$$

Hypothesis testing

From $P(H, X) = P(X)P(H|X)$, we can derive that

$$P(H|X) = \frac{P(H, X)}{P(X)}.$$

And since it is also true that $P(H, X) = P(H)P(X|H)$, we see that this is equivalent to

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)}.$$

This is one common formulation of **Bayes' Rule**, and analogous versions can be written for each of the other competing hypotheses; for example, Bayes' Rule for $\neg H$ is

$$P(\neg H|X) = \frac{P(\neg H)P(X|\neg H)}{P(X)}.$$

Hypothesis testing

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)}$$

In words, this reads: “The probability of an hypothesis given the data is equal to the probability of the hypothesis before seeing the data, multiplied by the probability that the data occur if that hypothesis is true, divided by the probability of the observed data.”

Hypothesis testing

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)}$$

In words, this reads: "The probability of an hypothesis given the data is equal to the probability of the hypothesis before seeing the data, multiplied by the probability that the data occur if that hypothesis is true, divided by the probability of the observed data."

Many of these quantities we know:

- We must have some prior probability (belief) that the hypothesis is true if we are considering the hypothesis

Hypothesis testing

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)}$$

In words, this reads: “The probability of an hypothesis given the data is equal to the probability of the hypothesis before seeing the data, multiplied by the probability that the data occur if that hypothesis is true, divided by the probability of the observed data.”

Many of these quantities we know:

- We must have some prior probability (belief) that the hypothesis is true if we are considering the hypothesis
- If the hypothesis is well-described it should attach a particular probability to the observed data

Hypothesis testing

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)}$$

In words, this reads: "The probability of an hypothesis given the data is equal to the probability of the hypothesis before seeing the data, multiplied by the probability that the data occur if that hypothesis is true, divided by the probability of the observed data."

Many of these quantities we know:

- We must have some prior probability (belief) that the hypothesis is true if we are considering the hypothesis
- If the hypothesis is well-described it should attach a particular probability to the observed data
- What remains is the denominator $P(X)$

$P(X)$

- $P(X)$ goes by many names, including *the normalizing constant*, *the marginal likelihood*, *the evidence*, or *the prior predictive probability of the data*.

$P(X)$

- $P(X)$ goes by many names, including *the normalizing constant*, *the marginal likelihood*, *the evidence*, or *the prior predictive probability of the data*.
- The name varies by how one uses Bayes' Rule

$P(X)$

- $P(X)$ goes by many names, including *the normalizing constant*, *the marginal likelihood*, *the evidence*, or *the prior predictive probability of the data*.
- The name varies by how one uses Bayes' Rule
- When one uses it to explain Bayes' Rule, *the prior predictive probability of the data* $P(X)$ is the probability of observing a given outcome in the experiment, taking into account all the possible hypotheses we are considering

$P(X)$

$P(X)$ can be thought of as the average likelihood under each hypothesis, weighted by the prior probability of each hypothesis.

$P(X)$

$P(X)$ can be thought of as the average likelihood under each hypothesis, weighted by the prior probability of each hypothesis.

Indeed, by the Sum Rule:

$$P(X) = P(X, H) + P(X, \neg H)$$

$P(X)$

$P(X)$ can be thought of as the average likelihood under each hypothesis, weighted by the prior probability of each hypothesis.

Indeed, by the Sum Rule:

$$P(X) = P(X, H) + P(X, \neg H)$$

And each of these terms can be obtained through application of the product rule, so we obtain the expression:

$$P(X) = P(H)P(X|H) + P(\neg H)P(X|\neg H).$$

$P(X)$

$P(X)$ can be thought of as the average likelihood under each hypothesis, weighted by the prior probability of each hypothesis.

Indeed, by the Sum Rule:

$$P(X) = P(X, H) + P(X, \neg H)$$

And each of these terms can be obtained through application of the product rule, so we obtain the expression:

$$P(X) = P(H)P(X|H) + P(\neg H)P(X|\neg H).$$

Which gives a weighted-average probability of observing the outcome.

A more complete formulation of Bayes' Rule

$$P(H|X) = \frac{P(H)P(X|H)}{P(H)P(X|H) + P(\neg H)P(X|\neg H)}.$$

Bayes' Rule is obtained as a necessary consequence of the Product Rule and the Sum Rule of probability.

Why is Bayes better?

The Lady Tasting Tea

The Lady Tasting Tea

This story is set in the early 1900s, in a dining hall at Rothamsted Agricultural Experiment Station, and involves Ronald Aylmer Fisher, a leading eugenicist, and his colleague, algologist Dr. Muriel Bristol.

The Lady Tasting Tea

This story is set in the early 1900s, in a dining hall at Rothamsted Agricultural Experiment Station, and involves Ronald Aylmer Fisher, a leading eugenicist, and his colleague, algologist Dr. Muriel Bristol.

Fisher put tea infusion into a cup and added milk.

The Lady Tasting Tea

This story is set in the early 1900s, in a dining hall at Rothamsted Agricultural Experiment Station, and involves Ronald Aylmer Fisher, a leading eugenicist, and his colleague, algologist Dr. Muriel Bristol.

Fisher put tea infusion into a cup and added milk.

But Lady Muriel protested, preferring her tea prepared the other way (first milk, then infusion).

The Lady Tasting Tea

This story is set in the early 1900s, in a dining hall at Rothamsted Agricultural Experiment Station, and involves Ronald Aylmer Fisher, a leading eugenicist, and his colleague, algologist Dr. Muriel Bristol.

Fisher put tea infusion into a cup and added milk.

But Lady Muriel protested, preferring her tea prepared the other way (first milk, then infusion).

Fisher, doubtful anyone could tell the difference, set up a blind discrimination experiment in which he prepared six pairs of cups.

The Lady Tasting Tea

This story is set in the early 1900s, in a dining hall at Rothamsted Agricultural Experiment Station, and involves Ronald Aylmer Fisher, a leading eugenicist, and his colleague, algologist Dr. Muriel Bristol.

Fisher put tea infusion into a cup and added milk.

But Lady Muriel protested, preferring her tea prepared the other way (first milk, then infusion).

Fisher, doubtful anyone could tell the difference, set up a blind discrimination experiment in which he prepared six pairs of cups.

In the experiment, the Lady Muriel was able to make the correct discrimination five times, but erred on the sixth: RRRRRW

The Lady Tasting Tea

This story is set in the early 1900s, in a dining hall at Rothamsted Agricultural Experiment Station, and involves Ronald Aylmer Fisher, a leading eugenicist, and his colleague, algologist Dr. Muriel Bristol.

Fisher put tea infusion into a cup and added milk.

But Lady Muriel protested, preferring her tea prepared the other way (first milk, then infusion).

Fisher, doubtful anyone could tell the difference, set up a blind discrimination experiment in which he prepared six pairs of cups.

In the experiment, the Lady Muriel was able to make the correct discrimination five times, but erred on the sixth: RRRRRW

So, can she truly tell the difference or not? How do we decide?

The Lady Tasting Tea (RRRRRW)

- Fisher proceeded to develop what is now known as Fisher's exact test in several steps

The Lady Tasting Tea (RRRRRW)

- Fisher proceeded to develop what is now known as Fisher's exact test in several steps
- This was his first attempt:

The Lady Tasting Tea (RRRRRW)

- Fisher proceeded to develop what is now known as Fisher's exact test in several steps
- This was his first attempt:
  - Suppose the lady is completely unable to do what she claims – she is effectively guessing. Let's call this the null hypothesis, H_0

The Lady Tasting Tea (RRRRRW)

- Fisher proceeded to develop what is now known as Fisher's exact test in several steps
- This was his first attempt:
  - Suppose the lady is completely unable to do what she claims – she is effectively guessing. Let's call this the null hypothesis, H_0
  - Then try to discredit (*nullify*) H_0 by demonstrating that *if* H_0 ***were*** *true, then the data would be very unlikely*:

The Lady Tasting Tea (RRRRRW)

- Fisher proceeded to develop what is now known as Fisher's exact test in several steps
- This was his first attempt:
  - Suppose the lady is completely unable to do what she claims – she is effectively guessing. Let's call this the null hypothesis, H_0
  - Then try to discredit (*nullify*) H_0 by demonstrating that *if* H_0 ***were*** *true, then the data would be very unlikely*:
  - In other words, set out to show that $p = P(X|H_0)$ is small

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt...

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt. . .
  - If the lady is guessing, then each trial will be R with $P_R = 0.5 = P_W$

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt...
  - If the lady is guessing, then each trial will be R with $P_R = 0.5 = P_W$
  - The data have probability $P_R^5 P_W = 2^{-6} = \frac{1}{64} \approx .016$

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt...
  - If the lady is guessing, then each trial will be R with $P_R = 0.5 = P_W$
  - The data have probability $P_R^5 P_W = 2^{-6} = \frac{1}{64} \approx .016$
- Fisher could now argue that either

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt...
  - If the lady is guessing, then each trial will be R with $P_R = 0.5 = P_W$
  - The data have probability $P_R^5 P_W = 2^{-6} = \frac{1}{64} \approx .016$
- Fisher could now argue that either
  1. H_0 is false and the lady can tell the difference, or

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt...
  - If the lady is guessing, then each trial will be R with $P_R = 0.5 = P_W$
  - The data have probability $P_R^5 P_W = 2^{-6} = \frac{1}{64} \approx .016$
- Fisher could now argue that either
  (1) H_0 is false and the lady can tell the difference, or
  (2) H_0 is true and an improbable event occurred

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt...
  - If the lady is guessing, then each trial will be R with $P_R = 0.5 = P_W$
  - The data have probability $P_R^5 P_W = 2^{-6} = \frac{1}{64} \approx .016$
- Fisher could now argue that either
  (1) H_0 is false and the lady can tell the difference, or
  (2) H_0 is true and an improbable event occurred
    - ... and improbable events are unlikely, so we prefer (1)

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt...
  - If the lady is guessing, then each trial will be R with $P_R = 0.5 = P_W$
  - The data have probability $P_R^5 P_W = 2^{-6} = \frac{1}{64} \approx .016$
- Fisher could now argue that either
  (1) H_0 is false and the lady can tell the difference, or
  (2) H_0 is true and an improbable event occurred
    - ... and improbable events are unlikely, so we prefer (1)
    - The result is said to be "significant" with p $=$ .016

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt is clearly nonsense

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt is clearly nonsense
  - *Every* possible outcome of the experiment (each of the 64 RWWWW, WRWRW, ...) has the same probability of .016

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt is clearly nonsense
  - *Every* possible outcome of the experiment (each of the 64 RWWWW, WRWRW, ...) has the same probability of .016
  - So using this reasoning, we would reject the H_0 *no matter what the data were*

The Lady Tasting Tea (RRRRRW)

- Fisher's first attempt is clearly nonsense
  - *Every* possible outcome of the experiment (each of the 64 RWWWW, WRWRW, ...) has the same probability of .016
  - So using this reasoning, we would reject the H_0 *no matter what the data were*
  - Fisher realized this absurdity, and made a second attempt

The Lady Tasting Tea (RRRRRW)

- Fisher's second attempt. . .

The Lady Tasting Tea (RRRRRW)

- Fisher's second attempt...
  - Instead, consider the probability of observing a *single* error, but at any point in the sequence: WRRRRR, RWRRRR, RRWRRR, RRRWRR, RRRRWR, RRRRRW

The Lady Tasting Tea (RRRRRW)

- Fisher's second attempt...
  - Instead, consider the probability of observing a *single* error, but at any point in the sequence: WRRRRR, RWRRRR, RRWRRR, RRRWRR, RRRRWR, RRRRRW
  - The probability of this happening at $P_R = 0.5$ is $6 \times P_R^5 P_W = \frac{6}{64} = 0.094$

The Lady Tasting Tea (RRRRRW)

- Fisher's second attempt...
  - Instead, consider the probability of observing a *single* error, but at any point in the sequence: WRRRRR, RWRRRR, RRWRRR, RRRWRR, RRRRWR, RRRRRW
  - The probability of this happening at $P_R = 0.5$ is $6 \times P_R^5 P_W = \frac{6}{64} = 0.094$
  - No longer "significant" at .05!

The Lady Tasting Tea (RRRRRW)

- Unfortunately, Fisher's second attempt has equally absurd implications

The Lady Tasting Tea (RRRRRW)

- Unfortunately, Fisher's second attempt has equally absurd implications
  - Consider a bigger tea-tasting study with 256 trials

The Lady Tasting Tea (RRRRRW)

- Unfortunately, Fisher's second attempt has equally absurd implications
  - Consider a bigger tea-tasting study with 256 trials
  - The lady performs exactly at chance level, with 128R and 128W (the most likely pattern)

The Lady Tasting Tea (RRRRRW)

- Unfortunately, Fisher's second attempt has equally absurd implications
  - Consider a bigger tea-tasting study with 256 trials
  - The lady performs exactly at chance level, with 128R and 128W (the most likely pattern)
  - Each pattern has a probability of 2^{-256}

The Lady Tasting Tea (RRRRRW)

- Unfortunately, Fisher's second attempt has equally absurd implications
  - Consider a bigger tea-tasting study with 256 trials
  - The lady performs exactly at chance level, with 128R and 128W (the most likely pattern)
  - Each pattern has a probability of 2^{-256}
  - ... but there are $5.7 \times 10^7 5$ of these "128R, 128W" patterns

The Lady Tasting Tea (RRRRRW)

- Unfortunately, Fisher's second attempt has equally absurd implications
  - Consider a bigger tea-tasting study with 256 trials
  - The lady performs exactly at chance level, with 128R and 128W (the most likely pattern)
  - Each pattern has a probability of 2^{-256}
  - ... but there are $5.7 \times 10^{7}5$ of these "128R, 128W" patterns
  - $p = 5.7 \times 10^{75} \times 2^{-256} \approx .049$, and we again reject H_0 for *every possible outcome*

The Lady Tasting Tea (RRRRRW)

- Fisher's third attempt...

The Lady Tasting Tea (RRRRRW)

- Fisher's third attempt. . .
  - If 1 error in 6 is significant, surely so is no error, or 6 R's

The Lady Tasting Tea (RRRRRW)

- Fisher's third attempt...
  - If 1 error in 6 is significant, surely so is no error, or 6 R's
  - We compute p as the probability of observing data that is *at least as extreme* as the real data, assuming that H_0 is true

The Lady Tasting Tea (RRRRRW)

- Fisher's third attempt...
  - If 1 error in 6 is significant, surely so is no error, or 6 R's
  - We compute p as the probability of observing data that is *at least as extreme* as the real data, assuming that H_0 is true
  - Absurdly, we now use as evidence an imaginary data pattern (RRRRRR) *that we did not observe and that no hypothesis we hold predicts* (more on this in a moment)

The Lady Tasting Tea (RRRRRW)

- Fisher's third attempt is now the dominant statistical paradigm in psychology

The Lady Tasting Tea (RRRRRW)

- Fisher's third attempt is now the dominant statistical paradigm in psychology
  - As it relies on frequency distributions of data in imagined universes, this approach is called "frequentist" as often as "classical"

The Lady Tasting Tea (RRRRRW)

- Fisher's third attempt is now the dominant statistical paradigm in psychology
  - As it relies on frequency distributions of data in imagined universes, this approach is called "frequentist" as often as "classical"
- For the lady tasting tea:

The Lady Tasting Tea (RRRRRW)

- Fisher's third attempt is now the dominant statistical paradigm in psychology
  - As it relies on frequency distributions of data in imagined universes, this approach is called "frequentist" as often as "classical"
- For the lady tasting tea:
  - For the outcome RRRRRW, there are 5 others as extreme and 1, with no errors, more extreme, giving 7 cases in all and a total probability of $7 \times 2^{-6} = .109$, not significant at 5%

The Lady Tasting Tea (RRRRRW)

But what is the consequence of this absurd reliance on hypothetical data sets such as RRRRRR?

The Lady Tasting Tea (RRRRRW)

But what is the consequence of this absurd reliance on hypothetical data sets such as RRRRRR?

- The frequentist is required to consider what results are as, or more, extreme

The Lady Tasting Tea (RRRRRW)

But what is the consequence of this absurd reliance on hypothetical data sets such as RRRRRR?

- The frequentist is required to consider what results are as, or more, extreme
- In this example, Fisher takes other possibilities with 6 pairs of cups

The Lady Tasting Tea (RRRRRW)

But what is the consequence of this absurd reliance on hypothetical data sets such as RRRRRR?

- The frequentist is required to consider what results are as, or more, extreme
- In this example, Fisher takes other possibilities with 6 pairs of cups
- But why fix 6? Did they decide that in advance, or did Dr. Bristol have a meeting to go to after tea? Had the cups been prepared less efficiently, might she have done fewer?

The Lady Tasting Tea (RRRRRW)

- What if the experiment was to keep going until she made a mistake?

The Lady Tasting Tea (RRRRRW)

- What if the experiment was to keep going until she made a mistake?
  - The probability of the sequence RRRRRW is still 2^{-6}

The Lady Tasting Tea (RRRRRW)

- What if the experiment was to keep going until she made a mistake?
  - The probability of the sequence RRRRRW is still 2^{-6}
  - More extreme sequences are now those in which the first mistake occurs after the 6^{th} pair, so at the 7^{th} (probability 2^{-7}), or 8^{th} (probability 2^{-8}), and so on

The Lady Tasting Tea (RRRRRW)

- What if the experiment was to keep going until she made a mistake?
  - The probability of the sequence RRRRRW is still 2^{-6}
  - More extreme sequences are now those in which the first mistake occurs after the 6^{th} pair, so at the 7^{th} (probability 2^{-7}), or 8^{th} (probability 2^{-8}), and so on
  - The probability of the observed result *and more extreme ones* is $\sum_{i=6}^{+\infty} 2^{-i} = .031$

The Lady Tasting Tea (RRRRRW)

- What if the experiment was to keep going until she made a mistake?
  - The probability of the sequence RRRRRW is still 2^{-6}
  - More extreme sequences are now those in which the first mistake occurs after the 6^{th} pair, so at the 7^{th} (probability 2^{-7}), or 8^{th} (probability 2^{-8}), and so on
  - The probability of the observed result *and more extreme ones* is $\sum_{i=6}^{+\infty} 2^{-i} = .031$
  - In the previous calculation, we had $p = .109$, yet now we have significance with $p < 5\%$

The Lady Tasting Tea (RRRRRW)

- What if the experiment was to keep going until she made a mistake?
  - The probability of the sequence RRRRRW is still 2^{-6}
  - More extreme sequences are now those in which the first mistake occurs after the 6^{th} pair, so at the 7^{th} (probability 2^{-7}), or 8^{th} (probability 2^{-8}), and so on
  - The probability of the observed result *and more extreme ones* is $\sum_{i=6}^{+\infty} 2^{-i} = .031$
  - In the previous calculation, we had $p = .109$, yet now we have significance with $p < 5\%$
- This is absurd! What does it matter what might have happened, but didn't?

The Lady Tasting Tea (RRRRRW)

It's important to note that, throughout, we've been calculating the probability of data patterns if H_0 is true.

The Lady Tasting Tea (RRRRRW)

It's important to note that, throughout, we've been calculating the probability of data patterns if H_0 is true.

These calculations are all based on $p(x|H_0)$. For example, the last p-value was $\sum_{i=6}^{+\infty} p(i|H_0)$.

The Lady Tasting Tea (RRRRRW)

It's important to note that, throughout, we've been calculating the probability of data patterns if H_0 is true.

These calculations are all based on $p(x|H_0)$. For example, the last p-value was $\sum_{i=6}^{+\infty} p(i|H_0)$.

But $p(x|H_0)$ is not what we are after—we are interested in $p(H_0|x)$.

The Lady Tasting Tea (RRRRRW)

An alternative analysis

- So far, we have only considered what happens under H_0

The Lady Tasting Tea (RRRRRW)

An alternative analysis

- So far, we have only considered what happens under H_0
- But what if H_0 is false? What if she really can tell?

The Lady Tasting Tea (RRRRRW)

An alternative analysis

- So far, we have only considered what happens under H_0
- But what if H_0 is false? What if she really can tell?
- Under $H_0 : P_R = 0.5$, but under $H_A : P_R > 0.5$

The Lady Tasting Tea (RRRRRW)

An alternative analysis

- So far, we have only considered what happens under H_0
- But what if H_0 is false? What if she really can tell?
- Under $H_0 : P_R = 0.5$, but under $H_A : P_R > 0.5$
- What we need to do is to compare the probability under H_0 with probabilities under the alternative hypotheses (other values of P_R). But which values of P_R?

The Lady Tasting Tea (RRRRRW)

An alternative analysis

- So far, we have only considered what happens under H_0
- But what if H_0 is false? What if she really can tell?
- Under $H_0 : P_R = 0.5$, but under $H_A : P_R > 0.5$
- What we need to do is to compare the probability under H_0 with probabilities under the alternative hypotheses (other values of P_R). But which values of P_R?
- To answer this consider another lady...