Probative Foundations for Bayesian Statistics

David Mwakima
dmwakima@uci.edu
University of California, Irvine
Visiting Fellow Strathmore Institute of Mathematical Sciences

August 30th 2023

Background

- I believe that a naturalistic epistemology for science should follow something along the lines of what is called "A belief revision model of knowledge" (See Barrett and Huttegger draft notes for details)

Background

- I believe that a naturalistic epistemology for science should follow something along the lines of what is called “A belief revision model of knowledge” (See Barrett and Huttegger draft notes for details)
- The Bayesian paradigm provides a good place to implement this model.

Background

- I believe that a naturalistic epistemology for science should follow something along the lines of what is called "A belief revision model of knowledge" (See Barrett and Huttegger draft notes for details)
- The Bayesian paradigm provides a good place to implement this model.
- In science, we need some measures of statistical evidence given the data to guide this process.

Background

- I believe that a naturalistic epistemology for science should follow something along the lines of what is called "A belief revision model of knowledge" (See Barrett and Huttegger draft notes for details)
- The Bayesian paradigm provides a good place to implement this model.
- In science, we need some measures of statistical evidence given the data to guide this process.
- Classical statistics has for some time been the main way to do this. See Fletcher, Samuel C. and Mayo-Wilson, Conor (forthcoming). "Evidence in Classical Statistics". In *Routledge Handbook of the Philosophy of Evidence* (edited by Maria Lasonen-Aarnio and Clayton Littlejohn). Routledge.

Background

- Bayesian statistics, using Bayes Factors, has recently been proposed as an alternative, mostly in the field of psychology due to work by Wagenmakers, Rouder, Morey and their collaborators in response to the replication crisis and spurious scientific findings.

Background

- Bayesian statistics, using Bayes Factors, has recently been proposed as an alternative, mostly in the field of psychology due to work by Wagenmakers, Rouder, Morey and their collaborators in response to the replication crisis and spurious scientific findings.
- Recently some philosophers of physics/science have been looking at the possibility of using Bayes Factors elsewhere other than psychology. See Massimi, Michela (2021). "A Philosopher's Look at the Dark Energy Survey: Reflections on The Use of The Bayes Factor in Cosmology". *In The Dark Energy Survey: The Story of a Cosmological Experiment*, pages 357–372. World Scientific.

Background

- Bayesian statistics, using Bayes Factors, has recently been proposed as an alternative, mostly in the field of psychology due to work by Wagenmakers, Rouder, Morey and their collaborators in response to the replication crisis and spurious scientific findings.
- Recently some philosophers of physics/science have been looking at the possibility of using Bayes Factors elsewhere other than psychology. See Massimi, Michela (2021). "A Philosopher's Look at the Dark Energy Survey: Reflections on The Use of The Bayes Factor in Cosmology". *In The Dark Energy Survey: The Story of a Cosmological Experiment*, pages 357–372. World Scientific.
- Compare with Roberto Trotta (2008) "Bayes in the sky: Bayesian Inference and Model Selection in Cosmology", Contemporary Physics, 49:2, 71-104

Background

- In Bayesian statistics, we use background information in our prior π to arrive at a revised belief p given a sampling model/likelihood f.

Background

- In Bayesian statistics, we use background information in our prior π to arrive at a revised belief p given a sampling model/likelihood f.
- From Bayes' Theorem, we get the following relation

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Suppose $\frac{\pi(M_0)}{\pi(M_1)}$ measures our **prior odds** for two models M_0 and M_1. Then we expect the **posterior odds** $\frac{p(M_0|X)}{p(M_1|X)}$ to be given by:

$$\frac{p(M_0|X)}{p(M_1|X)} = U\frac{\pi(M_0)}{\pi(M_1)}$$

Background

- In Bayesian statistics, we use background information in our prior π to arrive at a revised belief p given a sampling model/likelihood f.

- From Bayes' Theorem, we get the following relation

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Suppose $\frac{\pi(M_0)}{\pi(M_1)}$ measures our **prior odds** for two models M_0 and M_1. Then we expect the **posterior odds** $\frac{p(M_0|X)}{p(M_1|X)}$ to be given by:

$$\frac{p(M_0|X)}{p(M_1|X)} = U\frac{\pi(M_0)}{\pi(M_1)}$$

- Bayes Factor B_{01} is the updating factor U that quantifies the **relative predictive accuracy** of our models (Rouder and Morey 2019).

$$B_{01} = \frac{f(X|M_0)}{f(X|M_1)}$$

Background

- Challenges to this proposal

  Either your methodology picks up on influences on error probing capacities of methods or it does not. If it does, then you are in sync with the minimal severity requirement. We may compare our different ways of satisfying it. If it does not, then we've hit a crucial nerve. If you care, but your method fails to reflect that concern, then a supplement is in order. Opposition in methodology of statistics is fighting over trifles if it papers over this crucial point. If there is to be a meaningful "reconciliation," it will have to be here. Mayo (2018, 270)

Background

- The **Minimal Requirement for Evidence**

Background

- The **Minimal Requirement for Evidence**

  One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false. If data x agree with a claim C but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with C even if they exist, then we have bad evidence, no test (BENT). Mayo (2018, 5)

Background

- The **Minimal Requirement for Evidence**

  One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false. If data x agree with a claim C but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with C even if they exist, then we have bad evidence, no test (BENT). Mayo (2018, 5)

- What does "false" or "flaws" mean?

Background

- The **Minimal Requirement for Evidence**

  One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false. If data x agree with a claim C but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with C even if they exist, then we have bad evidence, no test (BENT). Mayo (2018, 5)

- What does "false" or "flaws" mean?
- What does "practically guaranteed" mean?

Background

- The **Minimal Requirement for Evidence**

  One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false. If data x agree with a claim C but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with C even if they exist, then we have bad evidence, no test (BENT). Mayo (2018, 5)

- What does “false” or “flaws” mean?
- What does “practically guaranteed” mean?
- What is a “capability of finding flaws”?

Background

- Herbert Hoijtink, Pascal van Kooten & Koenraad Hulsker (2016) "Why Bayesian Psychologists Should Change the Way They Use the Bayes Factor", *Multivariate Behavioral Research*, 51:1, 2-10

Background

- Herbert Hoijtink, Pascal van Kooten & Koenraad Hulsker (2016) "Why Bayesian Psychologists Should Change the Way They Use the Bayes Factor", *Multivariate Behavioral Research*, 51:1, 2-10
- Richard D. Morey, Eric-Jan Wagenmakers & Jeffrey N. Rouder (2016) "Calibrated Bayes Factors Should Not Be Used: A Reply to Hoijtink, van Kooten, and Hulsker", *Multivariate Behavioral Research*, 51:1, 11-19

Background

- Herbert Hoijtink, Pascal van Kooten & Koenraad Hulsker (2016) "Why Bayesian Psychologists Should Change the Way They Use the Bayes Factor", *Multivariate Behavioral Research*, 51:1, 2-10
- Richard D. Morey, Eric-Jan Wagenmakers & Jeffrey N. Rouder (2016) "Calibrated Bayes Factors Should Not Be Used: A Reply to Hoijtink, van Kooten, and Hulsker", *Multivariate Behavioral Research*, 51:1, 11-19
- Wagenmakers, E. J., Gronau, Q. F., & Vandekerckhove, J. (2019). "Five Bayesian intuitions for the stopping rule principle." Unpublished manuscript. *PsyArXiv*, 7th March 2019.

Background

- Herbert Hoijtink, Pascal van Kooten & Koenraad Hulsker (2016) "Why Bayesian Psychologists Should Change the Way They Use the Bayes Factor", *Multivariate Behavioral Research*, 51:1, 2-10
- Richard D. Morey, Eric-Jan Wagenmakers & Jeffrey N. Rouder (2016) "Calibrated Bayes Factors Should Not Be Used: A Reply to Hoijtink, van Kooten, and Hulsker", *Multivariate Behavioral Research*, 51:1, 11-19
- Wagenmakers, E. J., Gronau, Q. F., & Vandekerckhove, J. (2019). "Five Bayesian intuitions for the stopping rule principle." Unpublished manuscript. *PsyArXiv*, 7th March 2019.
- Morey, R. (2022) "Bayes factors, p-values, and the replication crisis" Workshop on 22nd of September 2022 of *Statistics Wars and Their Casualities* organized by Deborah Mayo. https://cardiffunipsychstats.co.uk/statswars2022/

My proposal:

- Let us try to get clear on what Mayo's **minimal requirement for evidence** is. Do this as charitably and as historically and mathematically accurate as we can be in order to meet her challenge in her own terms. This is meaningful reconciliation.

My proposal:

- Let us try to get clear on what Mayo's **minimal requirement for evidence** is. Do this as charitably and as historically and mathematically accurate as we can be in order to meet her challenge in her own terms. This is meaningful reconciliation.
- Distinguish between **model validation** and **model comparison**.

My proposal:

- Let us try to get clear on what Mayo's **minimal requirement for evidence** is. Do this as charitably and as historically and mathematically accurate as we can be in order to meet her challenge in her own terms. This is meaningful reconciliation.
- Distinguish between **model validation** and **model comparison**.
- Distinguish between **controlling** error and **considering** error.

My proposal:

- Let us try to get clear on what Mayo's **minimal requirement for evidence** is. Do this as charitably and as historically and mathematically accurate as we can be in order to meet her challenge in her own terms. This is meaningful reconciliation.
- Distinguish between **model validation** and **model comparison**.
- Distinguish between **controlling** error and **considering** error.
- Grant Mayo the point that Bayes Factors can't help with model validation but argue that Bayes Factors can meet the minimum requirement of evidence for model comparison.

My proposal:

- Let us try to get clear on what Mayo's **minimal requirement for evidence** is. Do this as charitably and as historically and mathematically accurate as we can be in order to meet her challenge in her own terms. This is meaningful reconciliation.
- Distinguish between **model validation** and **model comparison**.
- Distinguish between **controlling** error and **considering** error.
- Grant Mayo the point that Bayes Factors can't help with model validation but argue that Bayes Factors can meet the minimum requirement of evidence for model comparison.
- The key to my proposal is to argue that Mayo is wrong to criticize Bayes Factors because her criticism assumes that Bayes Factors are used in isolation. A subjective Bayesian will **consider** error by modeling the variability due to chance in their priors. Call this the **Full Package View of Bayes Factors** (Full Package View, in short).

Thesis

I will argue that the Full Package View is a Small Sample Optimal Test of Hypotheses.

What is involved in statistical "testing" of models?

- Is my model correct?

What is involved in statistical "testing" of models?

- Is my model correct?
- How do I compare models?

What is involved in statistical “testing” of models?

- Is my model correct?
- How do I compare models?
- These are two different questions.

What is involved in statistical "testing" of models?

- Is my model correct?
- How do I compare models?
- These are two different questions.
- Call the problem raised by question 1 the **Model Validation** problem.

What is involved in statistical "testing" of models?

- Is my model correct?
- How do I compare models?
- These are two different questions.
- Call the problem raised by question 1 the **Model Validation** problem.
- Call the problem raised by question 2 the **Model Comparison** problem.

Model Validation

- In the Frequentist approach to statistics, model validation is also called **significance testing**, following Fisher. The key measure here is the p-**value**.

Model Validation

- In the Frequentist approach to statistics, model validation is also called **significance testing**, following Fisher. The key measure here is the p-**value**.
- In the Bayesian approach to statistics the key measures here (as of today are):

Model Validation

- In the Frequentist approach to statistics, model validation is also called **significance testing**, following Fisher. The key measure here is the *p*-**value**.
- In the Bayesian approach to statistics the key measures here (as of today are):
  - **Marginal p-values** (Box, 1980)

Model Validation

- In the Frequentist approach to statistics, model validation is also called **significance testing**, following Fisher. The key measure here is the *p*-**value**.
- In the Bayesian approach to statistics the key measures here (as of today are):
  - **Marginal p-values** (Box, 1980)
  - **Predictive p-values** (Rubin, 1984; Gelman et al. 2004)

Model Comparison: Frequentist Approach

- In the Frequentist approach to statistics, model comparison follows the Gosset-Neyman-Pearson-Wald theory of hypothesis testing. According to this theory, model comparison is a constrained optimization decision problem.

Model Comparison: Frequentist Approach

- In the Frequentist approach to statistics, model comparison follows the Gosset-Neyman-Pearson-Wald theory of hypothesis testing. According to this theory, model comparison is a constrained optimization decision problem.

- Let us call this theory Small Sample Optimal Test Theory (for model comparison). Small Sample Optimal Test Theory is popularly known as Neyman-Pearson Hypothesis Testing.

Model Comparison: Frequentist Approach

- In the Frequentist approach to statistics, model comparison follows the Gosset-Neyman-Pearson-Wald theory of hypothesis testing. According to this theory, model comparison is a constrained optimization decision problem.
- Let us call this theory Small Sample Optimal Test Theory (for model comparison). Small Sample Optimal Test Theory is popularly known as Neyman-Pearson Hypothesis Testing.
- Why introduce a new label?

Model Comparison: Frequentist Approach

- In the Frequentist approach to statistics, model comparison follows the Gosset-Neyman-Pearson-Wald theory of hypothesis testing. According to this theory, model comparison is a constrained optimization decision problem.
- Let us call this theory Small Sample Optimal Test Theory (for model comparison). Small Sample Optimal Test Theory is popularly known as Neyman-Pearson Hypothesis Testing.
- Why introduce a new label?
  - It is more historically (and mathematically) accurate.

Model Comparison: Frequentist Approach

- In the Frequentist approach to statistics, model comparison follows the Gosset-Neyman-Pearson-Wald theory of hypothesis testing. According to this theory, model comparison is a constrained optimization decision problem.
- Let us call this theory Small Sample Optimal Test Theory (for model comparison). Small Sample Optimal Test Theory is popularly known as Neyman-Pearson Hypothesis Testing.
- Why introduce a new label?
  - It is more historically (and mathematically) accurate.
  - The words "testing" and talk of "error rates" are distracting from the issues. There's severe experimental testing (Karl Popper), significance testing, Neyman-Pearson hypothesis testing, severe (statistical) testing developed by Deborah Mayo and Aris Spanos.

Model Comparison: Frequentist Approach

- In the Frequentist approach to statistics, model comparison follows the Gosset-Neyman-Pearson-Wald theory of hypothesis testing. According to this theory, model comparison is a constrained optimization decision problem.
- Let us call this theory Small Sample Optimal Test Theory (for model comparison). Small Sample Optimal Test Theory is popularly known as Neyman-Pearson Hypothesis Testing.
- Why introduce a new label?
  - It is more historically (and mathematically) accurate.
  - The words "testing" and talk of "error rates" are distracting from the issues. There's severe experimental testing (Karl Popper), significance testing, Neyman-Pearson hypothesis testing, severe (statistical) testing developed by Deborah Mayo and Aris Spanos.
  - Clarifies the issues. The challenge (raised by Mayo) is for Bayesian statistics is to develop an analogous small sample optimal test theory using Bayes Factors.

More Historically (and Mathematically) Accurate

- A statistical test is a random experiment for model validation or model comparison.

More Historically (and Mathematically) Accurate

- A statistical test is a random experiment for model validation or model comparison.
- To fix terms and to illustrate the main ideas let $\theta \in \Theta_0$ and $\theta \in \Theta_1$ be a disjoint and exhaustive set of statistical hypothesis, i.e., $\Theta = \Theta_0 \cup \Theta_1$ is the parameter space for θ for a sampling model $Y \sim f(\theta)$.

More Historically (and Mathematically) Accurate

- A statistical test is a random experiment for model validation or model comparison.
- To fix terms and to illustrate the main ideas let $\theta \in \Theta_0$ and $\theta \in \Theta_1$ be a disjoint and exhaustive set of statistical hypothesis, i.e., $\Theta = \Theta_0 \cup \Theta_1$ is the parameter space for θ for a sampling model $Y \sim f(\theta)$.
- Suppose that we only observe one value $\mathrm{Y} = \mathrm{y}$, where $y \in \{1, 2, 3, 4, 5, 6\}$ and $f(Y; \theta)$ is the probability mass function for $Y = y$. Now consider the following sampling models of our data Y and the likelihood ratio (LR) in each case.

y	1	2	3	4	5	6
$f(Y = y\,;\,\theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\,\theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\mathrm{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

More Historically (and Mathematically) Accurate

- Let y_0 denote the observed value of y. Given a sampling model, define the *p*-**value** as:

  $$P(Y \geq y_0) \quad \text{or} \quad P(Y \leq y_0)$$

  We want the $\geq$ or $\leq$ in order to get non-trivial probabilities in the support of Y in the continuous case.

More Historically (and Mathematically) Accurate

- Let y_0 denote the observed value of y. Given a sampling model, define the p-**value** as:

$$P(Y \geq y_0) \quad \text{or} \quad P(Y \leq y_0)$$

We want the $\geq$ or $\leq$ in order to get non-trivial probabilities in the support of Y in the continuous case.

- Consider our sampling models. What are the associated p-values of these sampling models for $y_0 = 1$, $y_0 = 3$, $y_0 = 6$?

y	1	2	3	4	5	6
$f(Y = y\,;\,\theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\,\theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\text{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

More Historically (and Mathematically) Accurate

- Let y_0 denote the observed value of y. Given a sampling model, define the p-**value** as:

$$P(Y \geq y_0) \quad \text{or} \quad P(Y \leq y_0)$$

We want the $\geq$ or $\leq$ in order to get non-trivial probabilities in the support of Y in the continuous case.

- Consider our sampling models. What are the associated p-values of these sampling models for $y_0 = 1$, $y_0 = 3$, $y_0 = 6$?

y	1	2	3	4	5	6
$f(Y = y\,;\,\theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\,\theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\text{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

- It has nothing to do with "extremeness" or "weirdness" of our data. We can ask for the p-value of $Y = y_0 = 1$, after all in the $f(Y = y\,;\,\theta \in \Theta_0)$ case. Although a **low** p-value may invalidate our model.

More Historically (and Mathematically) Accurate

- There are no modal claims of "possible data" or "what we could have observed but didn't." This casts serious doubts on readings of the *p*-value in modal terms. Fletcher and Mayo-Wilson (forthcoming)

More Historically (and Mathematically) Accurate

- There are no modal claims of "possible data" or "what we could have observed but didn't." This casts serious doubts on readings of the *p*-value in modal terms. Fletcher and Mayo-Wilson (forthcoming)
- There is no mention of alternatives or the null hypothesis in the definition of the *p*-value **in general**. In the **special case**, where we're looking at the sampling model where $\theta \in \Theta_0$ asserts "there's no effect" we get what is usually called the null model.

More Historically (and Mathematically) Accurate

- There are no modal claims of "possible data" or "what we could have observed but didn't." This casts serious doubts on readings of the *p*-value in modal terms. Fletcher and Mayo-Wilson (forthcoming)
- There is no mention of alternatives or the null hypothesis in the definition of the *p*-value **in general**. In the **special case**, where we're looking at the sampling model where $\theta \in \Theta_0$ asserts "there's no effect" we get what is usually called the null model.
- *p*-values are not conditional probabilities.

More Historically (and Mathematically) Accurate

- There are no modal claims of "possible data" or "what we could have observed but didn't." This casts serious doubts on readings of the *p*-value in modal terms. Fletcher and Mayo-Wilson (forthcoming)
- There is no mention of alternatives or the null hypothesis in the definition of the *p*-value **in general**. In the **special case**, where we're looking at the sampling model where $\theta \in \Theta_0$ asserts "there's no effect" we get what is usually called the null model.
- *p*-values are not conditional probabilities.
- *p*-values are the right tools for **Model Validation**. More of this later.

More Historically (and Mathematically) Accurate

- Typically we don't know which of our sampling models we should use; after all, statistical testing is a random experiment. For certain realizations of Y it looks like one sampling model is "more plausible" than another. Rouder and Morey (2019) prefer the expression "predictively accurate" to "more plausible". I agree.

More Historically (and Mathematically) Accurate

- Typically we don't know which of our sampling models we should use; after all, statistical testing is a random experiment. For certain realizations of Y it looks like one sampling model is "more plausible" than another. Rouder and Morey (2019) prefer the expression "predictively accurate" to "more plausible". I agree.
- So we consider alternatives, which turns our inquiry into a model comparison problem.

More Historically (and Mathematically) Accurate

- Typically we don't know which of our sampling models we should use; after all, statistical testing is a random experiment. For certain realizations of Y it looks like one sampling model is "more plausible" than another. Rouder and Morey (2019) prefer the expression "predictively accurate" to "more plausible". I agree.
- So we consider alternatives, which turns our inquiry into a model comparison problem.
- p-values can't help us here because we have seen that depending on what we observed, we could have selected either sampling model.

More Historically (and Mathematically) Accurate

- Typically we don't know which of our sampling models we should use; after all, statistical testing is a random experiment. For certain realizations of Y it looks like one sampling model is "more plausible" than another. Rouder and Morey (2019) prefer the expression "predictively accurate" to "more plausible". I agree.
- So we consider alternatives, which turns our inquiry into a model comparison problem.
- p-values can't help us here because we have seen that depending on what we observed, we could have selected either sampling model.
- This was the gist of Jerzy Neyman and Egon Pearson's criticism of Ronald Fisher's **pure tests of significance** based on p-values.

More Historically (and Mathematically) Accurate

- Typically we don't know which of our sampling models we should use; after all, statistical testing is a random experiment. For certain realizations of Y it looks like one sampling model is "more plausible" than another. Rouder and Morey (2019) prefer the expression "predictively accurate" to "more plausible". I agree.
- So we consider alternatives, which turns our inquiry into a model comparison problem.
- p-values can't help us here because we have seen that depending on what we observed, we could have selected either sampling model.
- This was the gist of Jerzy Neyman and Egon Pearson's criticism of Ronald Fisher's **pure tests of significance** based on p-values.
- What did they propose instead?

More Historically (and Mathematically) Accurate

- If we index the partition of the parameter space and call Θ_0 the null (partition of the parameter) space and Θ_1 the alternative (partition of the parameter) space, we can label $\theta \in \Theta_0$ as the **null model** and $\theta \in \Theta_1$ as the **alternative model**.

More Historically (and Mathematically) Accurate

- If we index the partition of the parameter space and call Θ_0 the null (partition of the parameter) space and Θ_1 the alternative (partition of the parameter) space, we can label $\theta \in \Theta_0$ as the **null model** and $\theta \in \Theta_1$ as the **alternative model**.
- Define the **rejection region** R of a statistical test as the values of the random variable Y that if observed will lead us to reject $\theta \in \Theta_0$

More Historically (and Mathematically) Accurate

- If we index the partition of the parameter space and call Θ_0 the null (partition of the parameter) space and Θ_1 the alternative (partition of the parameter) space, we can label $\theta \in \Theta_0$ as the **null model** and $\theta \in \Theta_1$ as the **alternative model**.
- Define the **rejection region** R of a statistical test as the values of the random variable Y that if observed will lead us to reject $\theta \in \Theta_0$
- In fact, one may think of a statistical test as just this **choice** of a rejection region **before** we observe our data.

More Historically (and Mathematically) Accurate

- If we index the partition of the parameter space and call Θ_0 the null (partition of the parameter) space and Θ_1 the alternative (partition of the parameter) space, we can label $\theta \in \Theta_0$ as the **null model** and $\theta \in \Theta_1$ as the **alternative model**.
- Define the **rejection region** R of a statistical test as the values of the random variable Y that if observed will lead us to reject $\theta \in \Theta_0$
- In fact, one may think of a statistical test as just this **choice** of a rejection region **before** we observe our data.
- There is no talk of "truth" just assertions about the parameter space. (More of this later.)

More Historically (and Mathematically) Accurate

- If we index the partition of the parameter space and call Θ_0 the null (partition of the parameter) space and Θ_1 the alternative (partition of the parameter) space, we can label $\theta \in \Theta_0$ as the **null model** and $\theta \in \Theta_1$ as the **alternative model**.
- Define the **rejection region** R of a statistical test as the values of the random variable Y that if observed will lead us to reject $\theta \in \Theta_0$
- In fact, one may think of a statistical test as just this **choice** of a rejection region **before** we observe our data.
- There is no talk of "truth" just assertions about the parameter space. (More of this later.)
- Call this **the choice of rejection region** view of statistical tests. (**Choice View**, for short)

More Historically (and Mathematically) Accurate

- If we index the partition of the parameter space and call Θ_0 the null (partition of the parameter) space and Θ_1 the alternative (partition of the parameter) space, we can label $\theta \in \Theta_0$ as the **null model** and $\theta \in \Theta_1$ as the **alternative model**.
- Define the **rejection region** R of a statistical test as the values of the random variable Y that if observed will lead us to reject $\theta \in \Theta_0$
- In fact, one may think of a statistical test as just this **choice** of a rejection region **before** we observe our data.
- There is no talk of "truth" just assertions about the parameter space. (More of this later.)
- Call this **the choice of rejection region** view of statistical tests. (**Choice View**, for short)
- On the Choice View, a statistical test is a decision problem (something which Wald (1945) later realized).

More Historically (and Mathematically) Accurate

- Call a statistical hypothesis **simple** if it is specified as $\theta \in \{\theta_0\} \subset \Theta$. A statistical hypothesis is **composite** if it specified as $\theta \in \Theta$ where Θ includes **a range** of parameter values.

More Historically (and Mathematically) Accurate

- Call a statistical hypothesis **simple** if it is specified as $\theta \in \{\theta_0\} \subset \Theta$. A statistical hypothesis is **composite** if it specified as $\theta \in \Theta$ where Θ includes **a range** of parameter values.

- Consider our sampling models again.

y	1	2	3	4	5	6
$f(Y = y\,;\,\theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\,\theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\mathrm{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

More Historically (and Mathematically) Accurate

- Call a statistical hypothesis **simple** if it is specified as $\theta \in \{\theta_0\} \subset \Theta$. A statistical hypothesis is **composite** if it specified as $\theta \in \Theta$ where Θ includes **a range** of parameter values.

- Consider our sampling models again.

y	1	2	3	4	5	6
$f(Y = y\,;\,\theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\,\theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\text{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

- Here I assumed that $\theta \in \Theta_0$ and $\theta \in \Theta_1$ are simple, say $\theta = \theta_0$ and $\theta = \theta_1$ to simplify computation of LR and in order to appeal to the Neyman-Pearson Lemma.

More Historically (and Mathematically) Accurate

- What is the optimal rejection regions/tests on the Choice View? Suppose I observe $y_0 = 3$? What probability, under $\theta \in \Theta_0$, would I consider so low that it invalidates my model?

y	1	2	3	4	5	6
$f(Y = y\,;\,\theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\,\theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\text{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

More Historically (and Mathematically) Accurate

- What is the optimal rejection regions/tests on the Choice View? Suppose I observe $y_0 = 3$? What probability, under $\theta \in \Theta_0$, would I consider so low that it invalidates my model?
- The point is that the optimal choice is underdetermined unless I specify an additional constraint. This was Gosset's (aka "Student") idea that led Neyman and Pearson to their theory of hypothesis testing. The view I am calling Small Sample Optimal Test Theory

y	1	2	3	4	5	6
$f(Y = y\,;\,\theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\,\theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\mathrm{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

More Historically (and Mathematically) Accurate

- In a letter dated May 11, 1926., Gosset had written to Pearson to say:

More Historically (and Mathematically) Accurate

- In a letter dated May 11, 1926., Gosset had written to Pearson to say:
  ... even if the chance is very small, say .00001, that doesn't in itself necessarily prove that the sample is not drawn randomly from the population [specified by the hypothesis]: what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say .05 (such as that it belongs to a different population or that the sample wasn't random or whatever will do the trick) you will be very much more inclined to consider that the original hypothesis is not true.

More Historically (and Mathematically) Accurate

- In a letter dated May 11, 1926., Gosset had written to Pearson to say:
  ... even if the chance is very small, say .00001, that doesn't in itself necessarily prove that the sample is not drawn randomly from the population [specified by the hypothesis]: what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say .05 (such as that it belongs to a different population or that the sample wasn't random or whatever will do the trick) you will be very much more inclined to consider that the original hypothesis is not true.
- "That the original hypothesis is not true" is a distraction. We might just substitute "the alternative" here without loss of meaning.

More Historically (and Mathematically) Accurate

- Pearson passed the suggestion on to Neyman who was spending the year in Paris and who acknowledged it in a letter of December 6, 1926, agreeing that "to have the possibility of testing, it is necessary to adopt such a principle as Student's" (see Lehmann (1999))

More Historically (and Mathematically) Accurate

- Pearson passed the suggestion on to Neyman who was spending the year in Paris and who acknowledged it in a letter of December 6, 1926, agreeing that "to have the possibility of testing, it is necessary to adopt such a principle as Student's" (see Lehmann (1999))
- This is the More Historically (and Mathematically) Accurate appellation I have been suggesting, namely, the Gosset-Neyman-Pearson theory of hypothesis testing. Understood as a decision problem this is the Gosset-Neyman-Pearson-Wald theory of hypothesis testing.

More Historically (and Mathematically) Accurate

- Pearson passed the suggestion on to Neyman who was spending the year in Paris and who acknowledged it in a letter of December 6, 1926, agreeing that “to have the possibility of testing, it is necessary to adopt such a principle as Student’s” (see Lehmann (1999))
- This is the More Historically (and Mathematically) Accurate appellation I have been suggesting, namely, the Gosset-Neyman-Pearson theory of hypothesis testing. Understood as a decision problem this is the Gosset-Neyman-Pearson-Wald theory of hypothesis testing.
- In Neyman and Pearson (1928) and Neyman and Pearson (1933) we get what are called today **Neyman-Pearson Likelihood Ratio Tests**. The general (decision theoretic) theory was developed by Wald (1945).

Uniformly Most Powerful Tests

- The additional constraint is defined in terms of what is called the power function. Let $\eta : \Omega \times \Theta \to \mathbb{R}$. Where $R \subset \Omega$, which is the sample space for Y, define the **power function** as:

$$\eta(R, \theta) = P_{\theta \in \Theta}(Y \in R)$$

Uniformly Most Powerful Tests

- The additional constraint is defined in terms of what is called the power function. Let $\eta : \Omega \times \Theta \to \mathbb{R}$. Where $R \subset \Omega$, which is the sample space for Y, define the **power function** as:

$$\eta(R, \theta) = P_{\theta \in \Theta}(Y \in R)$$

- Define the **power** of the test $\beta = \eta \upharpoonright \Omega \times \Theta_1$

$$\beta = P_{\theta \in \Theta_1}(Y \in R)$$

  Compare Cohen (1988, 4).

Uniformly Most Powerful Tests

- The additional constraint is defined in terms of what is called the power function. Let $\eta : \Omega \times \Theta \to \mathbb{R}$. Where $R \subset \Omega$, which is the sample space for Y, define the **power function** as:

$$\eta(R, \theta) = P_{\theta \in \Theta}(Y \in R)$$

- Define the **power** of the test $\beta = \eta \restriction \Omega \times \Theta_1$

$$\beta = P_{\theta \in \Theta_1}(Y \in R)$$

  Compare Cohen (1988, 4).

- Define the **size** of the test $\alpha = \sup \eta \restriction \Omega \times \Theta_0$

$$\alpha = \sup_{\theta \in \Theta_0} P(Y \in R)$$

  Why "supremum"? We want a unique upper bound.

Uniformly Most Powerful Tests

- What is α and β for $R = \{3, 4\}$, $R = \{3, 5\}$, $R = \{4, 5\}$, $R = \{5, 6\}$, $R = \{4, 5, 6\}$?

y	1	2	3	4	5	6
$f(Y = y\,;\,\theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\,\theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\text{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

Uniformly Most Powerful Tests

- The test or choice of critical region $R = \{5, 6\}$ maximized the power of the test.

Uniformly Most Powerful Tests

- The test or choice of critical region $R = \{5, 6\}$ maximized the power of the test.
- This is the idea behind Uniformly Most Powerful Tests (UMP Test).

Uniformly Most Powerful Tests

- The test or choice of critical region $R = \{5, 6\}$ maximized the power of the test.
- This is the idea behind Uniformly Most Powerful Tests (UMP Test).
- As alluded to earlier, this is a constrained optimization decision problem, which we may summarize as follows on the Choice View: choose R such that the power function is maximized subject to the constraint α.

Uniformly Most Powerful Tests

- The test or choice of critical region $R = \{5, 6\}$ maximized the power of the test.
- This is the idea behind Uniformly Most Powerful Tests (UMP Test).
- As alluded to earlier, this is a constrained optimization decision problem, which we may summarize as follows on the Choice View: choose R such that the power function is maximized subject to the constraint α.
- It turns out that for a simple null model and exponential families of distributions, the UMP Test can be found by appealing to either the Neyman-Pearson lemma or the Karlin-Rubin theorem. (I shall not go into this)

Uniformly Most Powerful Tests

- The test or choice of critical region $R = \{5, 6\}$ maximized the power of the test.
- This is the idea behind Uniformly Most Powerful Tests (UMP Test).
- As alluded to earlier, this is a constrained optimization decision problem, which we may summarize as follows on the Choice View: choose R such that the power function is maximized subject to the constraint α.
- It turns out that for a simple null model and exponential families of distributions, the UMP Test can be found by appealing to either the Neyman-Pearson lemma or the Karlin-Rubin theorem. (I shall not go into this)
- The UMP Test is also where LR is highest. See next slide.

Uniformly Most Powerful Tests

y	1	2	3	4	5	6
$f(Y = y\,;\, \theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\, \theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\text{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

Severe Testing

- Mayo goes further than Neyman-Pearson hypothesis testing using what she calls **severity** and **severe testing**.

Severe Testing

- Mayo goes further than Neyman-Pearson hypothesis testing using what she calls **severity** and **severe testing**.
- Her account is **not** about long-run error control, which is what many people have assumed her account is about.

Severe Testing

- Mayo goes further than Neyman-Pearson hypothesis testing using what she calls **severity** and **severe testing**.
- Her account is **not** about long-run error control, which is what many people have assumed her account is about.
- Mayo insists that it is about what can be **learned** from an error-statistical approach to experimental inquiry. The error-probabilities on her account and the severity function measure of evidence allow us to **learn** that certain errors (e.g., mistaking an artifact from a real effect) are absent with certain probabilities. See Mayo (1996, 94 - 95).

Severe Testing: So, what can be learned?

- **The Argument from Error**

Severe Testing: So, what can be learned?

- **The Argument from Error**

  It is learned that an error is absent to the extent that a procedure that was highly capable of detecting the error nevertheless fails to do so.

Severe Testing: So, what can be learned?

- **The Argument from Error**

  It is learned that an error is absent to the extent that a procedure that was highly capable of detecting the error nevertheless fails to do so.

- Let $\theta \in \Theta_0$ denote the null model, $\theta \in \Theta_1$ denote the alternative model and Ω denote the sample space for Y. For a given test or rejection region $R \subset \Omega$, Mayo's **Severity Function** is the function $\text{SEV} : \Omega \times \Theta \to \mathbb{R}$ is given by:

$$SEV(\theta, D) = \begin{cases} P_{\theta \in \Theta_0}(Y \leq D) & \text{if } D \in R \\ P_{\theta \in \Theta_1}(Y \geq D) & \text{if } D \notin R \end{cases}$$

  where D is a function of the *observed* y_0.

Severe Testing: So, what can be learned?

- **The Argument from Error**

  It is learned that an error is absent to the extent that a procedure that was highly capable of detecting the error nevertheless fails to do so.

- Let $\theta \in \Theta_0$ denote the null model, $\theta \in \Theta_1$ denote the alternative model and Ω denote the sample space for Y. For a given test or rejection region $R \subset \Omega$, Mayo's **Severity Function** is the function $\text{SEV} : \Omega \times \Theta \to \mathbb{R}$ is given by:

$$SEV(\theta, D) = \begin{cases} P_{\theta \in \Theta_0}(Y \leq D) & \text{if } D \in R \\ P_{\theta \in \Theta_1}(Y \geq D) & \text{if } D \notin R \end{cases}$$

  where D is a function of the *observed* y_0.

- Compare with the definition of the power function. Where $R \subset \Omega$, we defined the **power of a test** as the function $\beta = \eta \upharpoonright \Omega \times \Theta_1$ given by:

$$P_{\theta \in \Theta_1}(Y \in R)$$

Severe Testing: So, what do we learn with SEV?

$$SEV(\theta, D) = \begin{cases} P_{\theta\in\Theta_0}(Y \leq D) & \text{if } D \in R \\ P_{\theta\in\Theta_1}(Y \geq D) & \text{if } D \notin R \end{cases}$$

- What is the severity of the Test $R = \{5, 6\}$ and Test $R = \{3, 4\}$ with $D = 1$. Now compare these tests with $D = 3$, $D = 5$ and $D = 6$. See next slide.

y	1	2	3	4	5	6
$f(Y = y\,;\,\theta \in \Theta_0)$	0.89	0.07	0.01	0.01	0.01	0.01
$f(Y = y\,;\,\theta \in \Theta_1)$	0.01	0.01	0.01	0.05	0.28	0.64
$\text{LR} = \frac{f(Y=y\,;\,\theta\in\Theta_1)}{f(Y=y\,;\,\theta\in\Theta_0)}$	0.0112	0.143	1	5	28	64

Severe Testing: So, what do we learn with SEV?

Interpretation: If $D \in R$ and $SEV(\theta, D)$ is **high**, then we learn that $\theta \in \Theta_1$ passed a highly severe test. If $D \notin R$ and $SEV(\theta, D)$ is **high**, then we learn that $\theta \in \Theta_0$ passed a highly severe test.

	$R = \{5, 6\}$	
D = 1	SEV(θ, 1) = 1	$\theta \in \Theta_0$ passed
D = 3	SEV(θ, 3) = 0.98	$\theta \in \Theta_0$ passed
D = 5	SEV(θ, 5) = 0.99	$\theta \in \Theta_1$ passed
D = 6	SEV(θ, 6) = 1	$\theta \in \Theta_1$ passed

	$R = \{3, 4\}$	
D = 1	SEV(θ, 1) = 1	$\theta \in \Theta_0$ passed
D = 2	SEV(θ, 2) = 0.99	$\theta \in \Theta_0$ passed
D = 3	SEV(θ, 3) = 0.97	$\theta \in \Theta_1$ passed
D = 4	SEV(θ, 4) = 0.98	$\theta \in \Theta_1$ passed

Severe Testing: So, what do we learn with SEV?

- For some D, the Test $R = \{5, 6\}$ is not necessarily the most severe test of its size for $\theta \in \Theta_0$.

Severe Testing: So, what do we learn with SEV?

- For some D, the Test $R = \{5, 6\}$ is not necessarily the most severe test of its size for $\theta \in \Theta_0$.
- For some D and Test R, we can either pass $\theta \in \Theta_0$ or $\theta \in \Theta_1$.

Severe Testing: So, what do we learn with SEV?

- For some D, the Test $R = \{5, 6\}$ is not necessarily the most severe test of its size for $\theta \in \Theta_0$.
- For some D and Test R, we can either pass $\theta \in \Theta_0$ or $\theta \in \Theta_1$.
- A likelihood ratio for $D = 3$ is neutral between the two models. But with SEV and a given choice of R, we can say whether or not either $\theta \in \Theta_0$ or $\theta \in \Theta_1$ passed a highly severe test.

Severe Testing: So, what do we learn with SEV?

- For some D, the Test $R = \{5, 6\}$ is not necessarily the most severe test of its size for $\theta \in \Theta_0$.
- For some D and Test R, we can either pass $\theta \in \Theta_0$ or $\theta \in \Theta_1$.
- A likelihood ratio for $D = 3$ is neutral between the two models. But with SEV and a given choice of R, we can say whether or not either $\theta \in \Theta_0$ or $\theta \in \Theta_1$ passed a highly severe test.

  It's the idea of viewing statistical inference as severe testing that invites a non-trivial difference with probabilism. Mayo (2018, 346)

Truth and Error Rates

- Throughout this discussion I have not talked about truth or error rates, yet we have understood what is going on.

Truth and Error Rates

- Throughout this discussion I have not talked about truth or error rates, yet we have understood what is going on.
- This means that the interpretation of α as the probability of a Type I error rate is a distraction from the real issues. The error is not something we do, which we somehow have to control. (This is what Mayo calls the view of statistical testing in terms of "performance")

Truth and Error Rates

- Throughout this discussion I have not talked about truth or error rates, yet we have understood what is going on.
- This means that the interpretation of α as the probability of a Type I error rate is a distraction from the real issues. The error is not something we do, which we somehow have to control. (This is what Mayo calls the view of statistical testing in terms of "performance")
- The error is due to **sampling variability**. A random variable has intrinsic variability quantified by its distribution function. (Call this the variability view of error.)

Truth and Error Rates

- Throughout this discussion I have not talked about truth or error rates, yet we have understood what is going on.
- This means that the interpretation of α as the probability of a Type I error rate is a distraction from the real issues. The error is not something we do, which we somehow have to control. (This is what Mayo calls the view of statistical testing in terms of "performance")
- The error is due to **sampling variability**. A random variable has intrinsic variability quantified by its distribution function. (Call this the variability view of error.)
- On the variability view of error, we are not **controlling** errors, we are **considering** errors.

Truth and Error Rates

Consider the famous passage from Pearson and Neyman (1930, 106) "On the Problem of Two Samples."

> *But if we accept the criterion suggested by the method of likelihood it is still necessary to determine its sampling distribution in order to control the error involved in rejecting a true hypothesis, because a knowledge of* λ *alone is not adequate to insure control of this error.*

- Most people stop here and say, "Aha! See! It's about controlling error."

Truth and Error Rates

Consider the famous passage from Pearson and Neyman (1930, 106) "On the Problem of Two Samples."

> *But if we accept the criterion suggested by the method of likelihood it is still necessary to determine its sampling distribution in order to control the error involved in rejecting a true hypothesis, because a knowledge of λ alone is not adequate to insure control of this error.*

- Most people stop here and say, "Aha! See! It's about controlling error."
- Not so fast!

Truth and Error Rates

In the same paragraph they go on to say:

> *We cannot for example say in general that if $\lambda \leq \lambda_0 = 0.01$, we should be justified in rejecting the hypothesis. In order to fix a limit between "small" and "large" values of λ we must know how often such values appear when we deal with a true hypothesis. That is to say we must have knowledge of P_{λ_0}, the chance of obtaining $\lambda \leq \lambda_0$ in the case where the hypothesis tested is true. The frequency distribution of λ differs according to the size of samples and the nature of the hypothesis tested, and it may well happen that the modal value of λ is in the neighbourhood of zero.*

Here they are clearly talking about **considering error** due to **sampling variability**.

Truth and Error Rates

Concerning the notion of "truth" and "Type I and Type II errors", they had said this in the same paper.

> *We have discussed elsewhere (Neyman and Pearson (1928a, b); Neyman (1929a)) certain principles regarding the testing of hypotheses, which we believe to be intuitively sound. They are not, strictly speaking, mathematical results and may be rejected by those who do not believe in them.*

- So we may drop talk of "truth concerning statistical hypotheses", even the talk of "Type I and Type II errors".

Truth and Error Rates

Concerning the notion of "truth" and "Type I and Type II errors", they had said this in the same paper.

> *We have discussed elsewhere (Neyman and Pearson (1928a, b); Neyman (1929a)) certain principles regarding the testing of hypotheses, which we believe to be intuitively sound. They are not, strictly speaking, mathematical results and may be rejected by those who do not believe in them.*

- So we may drop talk of "truth concerning statistical hypotheses", even the talk of "Type I and Type II errors".
- But we can preserve the "intuitively sound" choice of α. While this choice is seemingly "arbitrary", it can be justified from a decision-theoretic point of view as the risk given a 0-1 loss function.

Truth and Error Rates

- So the real question is this: **is using α to constrain our optimization decision problem well-motivated?** I think the answer is yes. Because of considerations of variability in our data, we may need **calibration**.

Truth and Error Rates

- So the real question is this: **is using α to constrain our optimization decision problem well-motivated?** I think the answer is yes. Because of considerations of variability in our data, we may need **calibration**.
- "Calibration" is a loaded word. See Morey, Wagenmakers, Rouder (MWR) (2016) cited earlier and compare with Hoijtink, Kooten, and Hulsker (HKH) (2016) earlier.

Truth and Error Rates

- So the real question is this: **is using α to constrain our optimization decision problem well-motivated?** I think the answer is yes. Because of considerations of variability in our data, we may need **calibration**.
- "Calibration" is a loaded word. See Morey, Wagenmakers, Rouder (MWR) (2016) cited earlier and compare with Hoijtink, Kooten, and Hulsker (HKH) (2016) earlier.
- HKH are right in the problems they raise: specify your subjective priors ("break the Bayesian egg"), calibrate by providing subjective priors on effect sizes.

Truth and Error Rates

- So the real question is this: **is using α to constrain our optimization decision problem well-motivated?** I think the answer is yes. Because of considerations of variability in our data, we may need **calibration**.
- "Calibration" is a loaded word. See Morey, Wagenmakers, Rouder (MWR) (2016) cited earlier and compare with Hoijtink, Kooten, and Hulsker (HKH) (2016) earlier.
- HKH are right in the problems they raise: specify your subjective priors ("break the Bayesian egg"), calibrate by providing subjective priors on effect sizes.
- HKH are wrong to "elaborate how frequency calculations can be used to provide an interpretation of the size of the Bayes factor" because they focus on **controlling error**.

Truth and Error Rates

- Subjective priors based on effect sizes is a good idea if you're **considering errors** because by definition, you are considering the variability. Compare Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iver- son, G. J., & (2011). "Statistical evidence in experimental psychology: An empirical comparison using 855 t tests" Perspectives on Psychological Science, 6, 291–298

Truth and Error Rates

- Subjective priors based on effect sizes is a good idea if you're **considering errors** because by definition, you are considering the variability. Compare Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iver- son, G. J., & (2011). "Statistical evidence in experimental psychology: An empirical comparison using 855 t tests" Perspectives on Psychological Science, 6, 291–298
- "The default test provides an excellent starting point of analysis, one that may later be supplemented with a detailed problem-specific analysis." (same paper p. 294)

Truth and Error Rates

- Subjective priors based on effect sizes is a good idea if you're **considering errors** because by definition, you are considering the variability. Compare Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iver- son, G. J., & (2011). "Statistical evidence in experimental psychology: An empirical comparison using 855 t tests" Perspectives on Psychological Science, 6, 291–298
- "The default test provides an excellent starting point of analysis, one that may later be supplemented with a detailed problem-specific analysis." (same paper p. 294)
- The test is default because it applies regardless of the phenomenon under study: For every experiment, one uses the same prior distribution on effect size for the alternative hypothesis, the Cauchy (0,1) distribution.

Truth and Error Rates

- Subjective priors based on effect sizes is a good idea if you're **considering errors** because by definition, you are considering the variability. Compare Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iver- son, G. J., & (2011). "Statistical evidence in experimental psychology: An empirical comparison using 855 t tests" Perspectives on Psychological Science, 6, 291–298
- "The default test provides an excellent starting point of analysis, one that may later be supplemented with a detailed problem-specific analysis." (same paper p. 294)
- The test is default because it applies regardless of the phenomenon under study: For every experiment, one uses the same prior distribution on effect size for the alternative hypothesis, the Cauchy (0,1) distribution.
- Is this enough calibration? This is what HKH were arguing is not enough.

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

- My proposal builds on the following idea.

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

- My proposal builds on the following idea.

  We note that, from a Bayesian perspective, the effect size can naturally be conceived as (a summary statistic of) the posterior distribution of a parameter representing the effect, under an uninformative prior distribution. In this sense, a standard Bayesian combination of parameter estimation and model selection could encompass all of the useful measures of evidence we observed. Wetzels et. al (2011, 296)

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

- My proposal builds on the following idea.

  We note that, from a Bayesian perspective, the effect size can naturally be conceived as (a summary statistic of) the posterior distribution of a parameter representing the effect, under an uninformative prior distribution. In this sense, a standard Bayesian combination of parameter estimation and model selection could encompass all of the useful measures of evidence we observed. Wetzels et. al (2011, 296)

- Why "uninformative?" Break the Bayesian egg.

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

- Let's think of calibration **not** from a Frequentist perspective but from a general point of view.

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

- Let's think of calibration **not** from a Frequentist perspective but from a general point of view.
- "Calibration" is related to **sensitivity** and **specificity** of instruments.

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

- Let's think of calibration **not** from a Frequentist perspective but from a general point of view.
- "Calibration" is related to **sensitivity** and **specificity** of instruments.
- So we can talk about the calibration of our instruments for measuring statistical evidence. This is what Mayo has recently called **probativeness**, which is different from **performance** and **probabilism**. See Mayo (2018, 396)

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

- Let's think of calibration **not** from a Frequentist perspective but from a general point of view.
- "Calibration" is related to **sensitivity** and **specificity** of instruments.
- So we can talk about the calibration of our instruments for measuring statistical evidence. This is what Mayo has recently called **probativeness**, which is different from **performance** and **probabilism**. See Mayo (2018, 396)
- Assuming sensitivity and specificity is a good thing (which I think it is), the challenge, again, is **to formulate a small sample optimal test theory for Bayesian statistics**.

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

Part 1

- But we can already do this. Think about medical diagnostics.

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

Part 1

- But we can already do this. Think about medical diagnostics.
- $P(+|\,\text{have disease})$ (Sensitivity) and $P(-|\,\text{don't have the disease})$ (Specificity)

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

Part 1

- But we can already do this. Think about medical diagnostics.
- $P(+|\,\text{have disease})$ (Sensitivity) and $P(-|\,\text{don't have the disease})$ (Specificity)
- But the Bayes Factor $\frac{f(X|M_0)}{f(X|M_1)}$ has this natural interpretation in terms of sensitivity and specificity.

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

Part 1

- But we can already do this. Think about medical diagnostics.
- $P(+|\,\text{have disease})$ (Sensitivity) and $P(-|\,\text{don't have the disease})$ (Specificity)
- But the Bayes Factor $\frac{f(X|M_0)}{f(X|M_1)}$ has this natural interpretation in terms of sensitivity and specificity.
- So my proposal is that we can adjust these within a unified framework of both Bayes Factors **and** priors by taking into account $P(\text{have disease})$ and $P(\text{don't have disease})$ to get the right posterior odds, $P(\text{have disease}\,|+)$ and $P(\text{don't have the disease}\,|-)$

Towards a Small Sample Optimal Test Theory for Bayesian Statistics

Part 1

- But we can already do this. Think about medical diagnostics.
- $P(+|\,\text{have disease})$ (Sensitivity) and $P(-|\,\text{don't have the disease})$ (Specificity)
- But the Bayes Factor $\frac{f(X|M_0)}{f(X|M_1)}$ has this natural interpretation in terms of sensitivity and specificity.
- So my proposal is that we can adjust these within a unified framework of both Bayes Factors **and** priors by taking into account $P(\text{have disease})$ and $P(\text{don't have disease})$ to get the right posterior odds, $P(\text{have disease}\,|+)$ and $P(\text{don't have the disease}\,|-)$
- Just as one uses the sampling distribution to consider variability one can *use the prior* to consider error.

Bayes Risk for using a Bayes Factor

Part 2

- Take the 0-1 Loss function and use a subjective prior on effect sizes to calculate Bayes Risk of using a given Bayes Factor. We report Bayes Factor and the Bayes Risk together.

Bayes Risk for using a Bayes Factor

Part 2

- Take the 0-1 Loss function and use a subjective prior on effect sizes to calculate Bayes Risk of using a given Bayes Factor. We report Bayes Factor and the Bayes Risk together.
- Compare Wetzels et. al (2011, 296)

Bayes Risk for using a Bayes Factor

Part 2

- Take the 0-1 Loss function and use a subjective prior on effect sizes to calculate Bayes Risk of using a given Bayes Factor. We report Bayes Factor and the Bayes Risk together.
- Compare Wetzels et. al (2011, 296)

  The Bayes factor is not a measure of the mere size of an effect. Hence, the measure of effect size confers additional information, particularly when small numbers of participants or trials are involved. So, especially for these sorts of studies, there is an argument for reporting both a Bayes factor and an effect size.

Bayes Risk for using a Bayes Factor

Part 2

- Take the 0-1 Loss function and use a subjective prior on effect sizes to calculate Bayes Risk of using a given Bayes Factor. We report Bayes Factor and the Bayes Risk together.
- Compare Wetzels et. al (2011, 296)

  The Bayes factor is not a measure of the mere size of an effect. Hence, the measure of effect size confers additional information, particularly when small numbers of participants or trials are involved. So, especially for these sorts of studies, there is an argument for reporting both a Bayes factor and an effect size.
- This is our sample small optimal test theory. The details will be added.