# Bayesian Hypothesis Testing

Jim Berger

Duke University

*PHYSTATnu*
*CERN*
*January 22, 2019*

# Outline

- Bayesian testing and $p$-values (chi-square)
- Ockham's razor
- Precise and imprecise hypotheses
- Choice of prior distributions for testing
- Examples

# Two Types of Bayesian Problems

**I. Estimation (confidence limit) problems:** In principle, they are straightforward.

- There are optimal prior distributions for most problems (e.g., *reference priors* - although their derivation can be difficult).
- Implementation of Bayes is usually easy, through MCMC.

**II. Hypothesis testing or model uncertainty problems:** Not so easy.

- Sometimes one can use the optimal estimation priors, but often not.
- In the latter case, the answers can be quite sensitive to the choice of prior,
  - so that one often seeks a *robust* conclusion over the choice.
- Computations can be much more difficult.

# Bayesian Testing and $p$-values

Data: $N = \#$ events observed in time $T$ that are characteristic of Higgs boson production in LHC particle collisions.

Statistical Model: $N$ has density

$$\text{Poisson}(N \mid s + b) = \frac{(s+b)^N e^{-(s+b)}}{N!}\,,$$

where

- $s$ is the mean rate of production of Higgs events in time $T$;
- $b$ is the (known) mean rate of production of events with the same characteristics from background sources in time $T$.

To test: $H_0 : s = 0$ vs $H_1 : s > 0$. ($H_0$ corresponds to 'no Higgs.')

P-value: $P(N \geq N_{\text{observed}} \mid b, s = 0) = \sum_{j=N}^{\infty} \text{Poisson}(j \mid 0 + b)$

*Case 1:* $p = 0.00025$ if $N = 7$, $b = 1.2$

*Case 2:* $p = 0.025$ if $N = 6$, $b = 2.2$.

- Those who understand $p$-values know their use is difficult:
  *Luc Demortier:* In any search for new physics, a small p value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery.
- Bayesian analysis directly measures if the alternative hypothesis provides a better explanation.

Sequential testing: This is actually a sequential experiment, so $p$ should be adjusted to account for multiple looks at the data. Bayesian analysis does not need such a correction.

Bayes factor (*evidence* in physics) of $H_0$ to $H_1$: *ratio of likelihood under* $H_0$ *to average likelihood under* $H_1$ (or "odds" of $H_0$ to $H_1$)

$$B_{01}(N) = \frac{\text{Poisson}(N \mid 0+b)}{\int_0^\infty \text{Poisson}(N \mid s+b)\pi(s)\, ds} = \frac{b^N\ e^{-b}}{\int_0^\infty (s+b)^N\ e^{-(s+b)}\pi(s)\, ds}\,.$$

Subjective approach: Choose $\pi(s)$ subjectively (e.g., using the standard physics model predictions of the mass of the Higgs).

Objective approach: Choose $\pi(s)$ to be the 'intrinsic prior' (not discussed here) $\pi^I(s) = b(s+b)^{-2}$. (Note that this prior is proper and has median $b$.)

Bayes factor: is then given by

$$B_{01} = \frac{b^N\ e^{-b}}{\int_0^\infty (s+b)^N\ e^{-(s+b)} b(s+b)^{-2}\ ds} = \frac{b^{(N-1)}\ e^{-b}}{\Gamma(N-1, b)}\,,$$

where $\Gamma$ is the incomplete gamma function.

*Case 1:* $B_{01} = 0.0075$ (recall $p = 0.00025$)

*Case 2:* $B_{01} = 0.26$ (recall $p = 0.025$)

Posterior probability of the null hypothesis: The objective choice of prior probabilities of the hypotheses is $\Pr(H_0) = \Pr(H_1) = 0.5$ (after the Higgs discovery, we would use $Pr(H_0) = 0.000005$ for new data), in which case

$$\Pr(H_0 \mid N) = (1 + B_{01}^{-1})^{-1} .$$

*Case 1:* $\Pr(H_0 \mid N) = 0.0075$ (recall $p = 0.00025$)
*Case 2:* $\Pr(H_0 \mid N) = 0.21$ (recall $p = 0.025$)

Complete posterior distribution: is given by

- $\Pr(H_0 \mid N)$, the posterior probability of null hypothesis
- $\pi(s \mid N, H_1)$, the posterior distribution of $s$ under $H_1$

A useful summary of the complete posterior is $\Pr(H_0 \mid N)$ and $C$, a (say) 95% posterior credible set for $s$ under $H_1$.

*Case 1:* $\Pr(H_0 \mid N) = 0.0075$; $C = (1.0, 10.5)$

*Case 2:* $\Pr(H_0 \mid N) = 0.21$; $C = (0.2, 8.2)$

**Note: For testing precise hypotheses, confidence intervals alone are *not* a satisfactory inferential summary.**

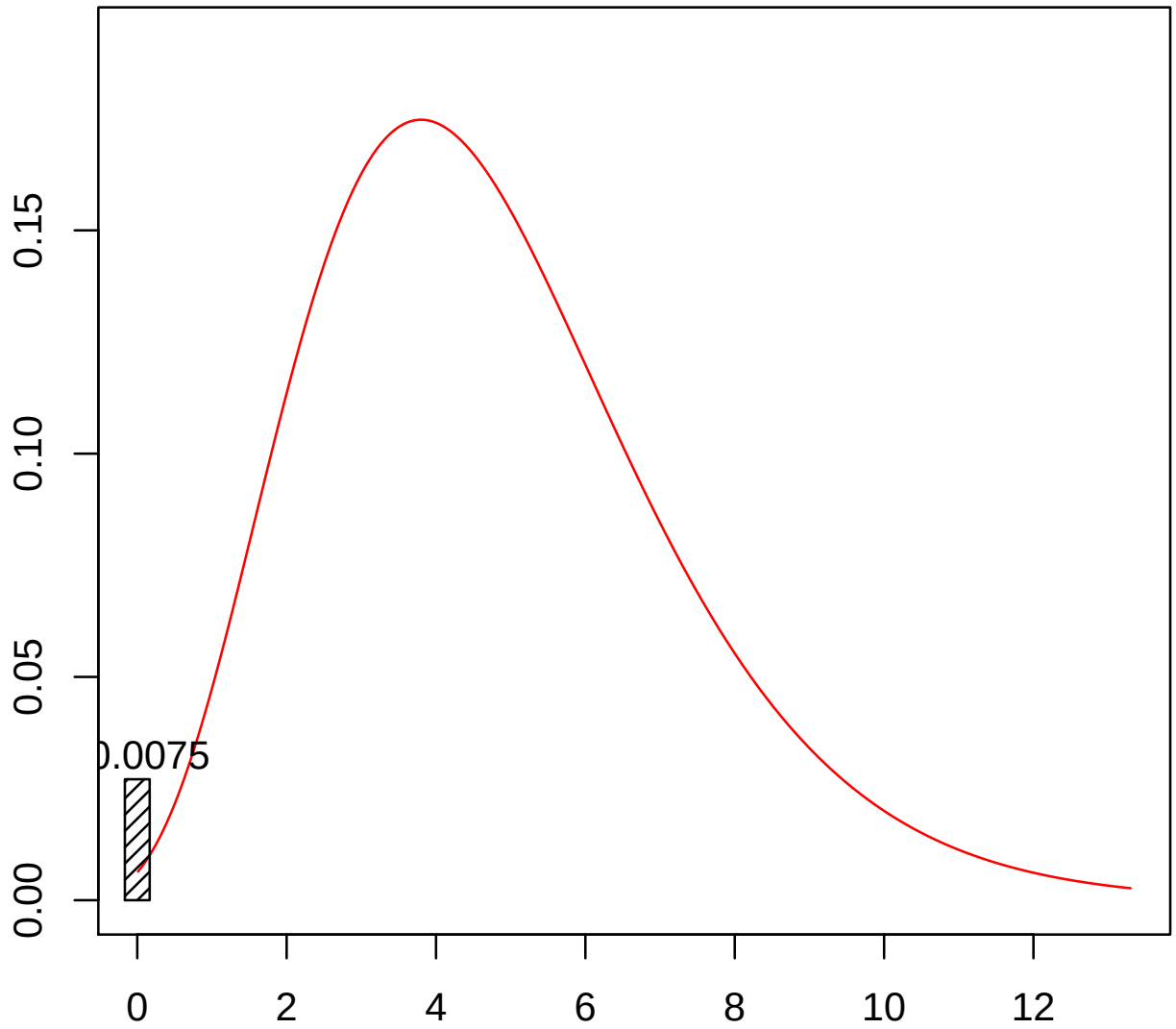



Figure 1: $\Pr(H_0 \mid N)$ (the vertical bar), and the posterior density for $s$ given $N$ and $H_1$.

## Is the discrepancy between $p$-values and Bayes factors due to choice of the prior?

A lower bound on the likelihood ratio (or Bayes factor): choose $\pi(s)$ to be a point mass at $\hat{s}$, yielding

$$B_{01}(N) = \frac{\text{Poisson}(N \mid 0+b)}{\int_0^\infty \text{Poisson}(N \mid s+b)\pi(s)\, ds} \geq \frac{\text{Poisson}(N \mid 0+b)}{\text{Poisson}(N \mid \hat{s}+b)}$$

$$= \min\{1, \left(\frac{b}{N}\right)^N e^{N-b}\}\,.$$

*Case 1:* $B_{01} \geq 0.0014$ (recall $p = 0.00025$)

*Case 2:* $B_{01} \geq 0.11$ (recall $p = 0.025$)

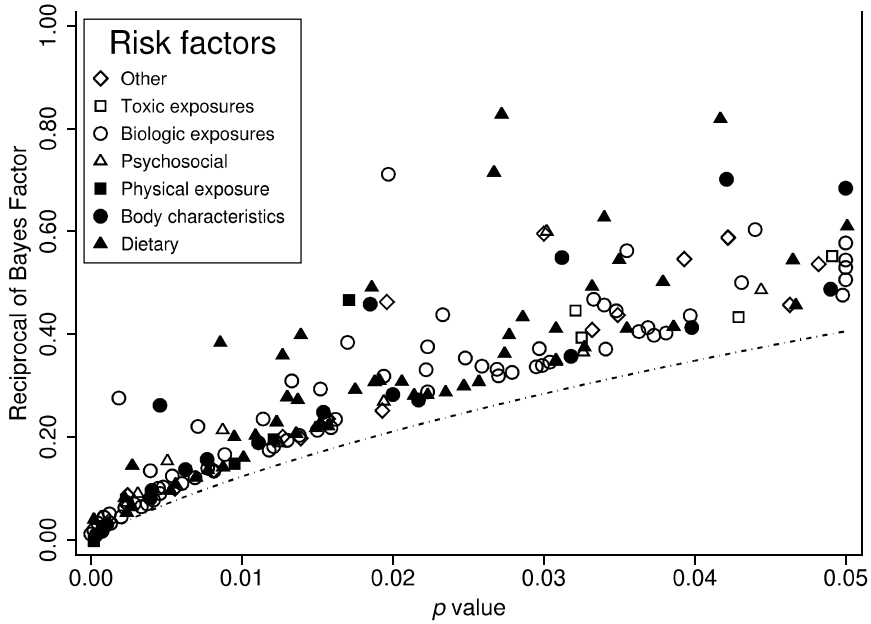
**Note:** This use of *robust Bayesian analysis* was done in Edwards, Lindman and Savage (1963) and Berger and Sellke (1987); many generalizations followed; indeed, there is now the Society of Imprecise Probabilities.

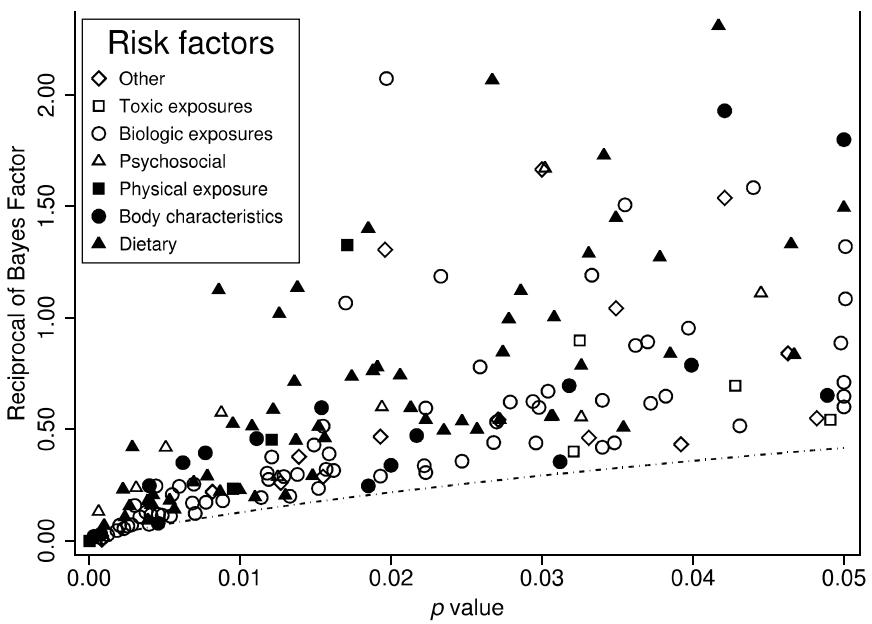
The following studies

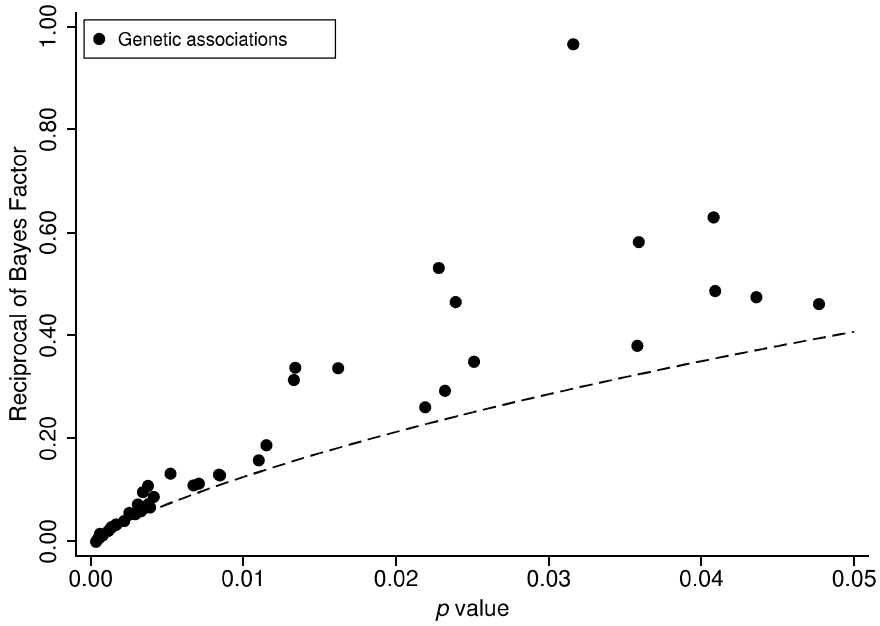
- look at large collections of published studies where $0 < p < 0.05$;
- compute a Bayes factor, $B_{01}$ for each study;
- graph the Bayes factors versus the corresponding $p$-values;
- the dashed lines are the general Bayes factor lower bound, for a given $p$-value $p$ (from Sellke, Bayarri and Berger, 2001),

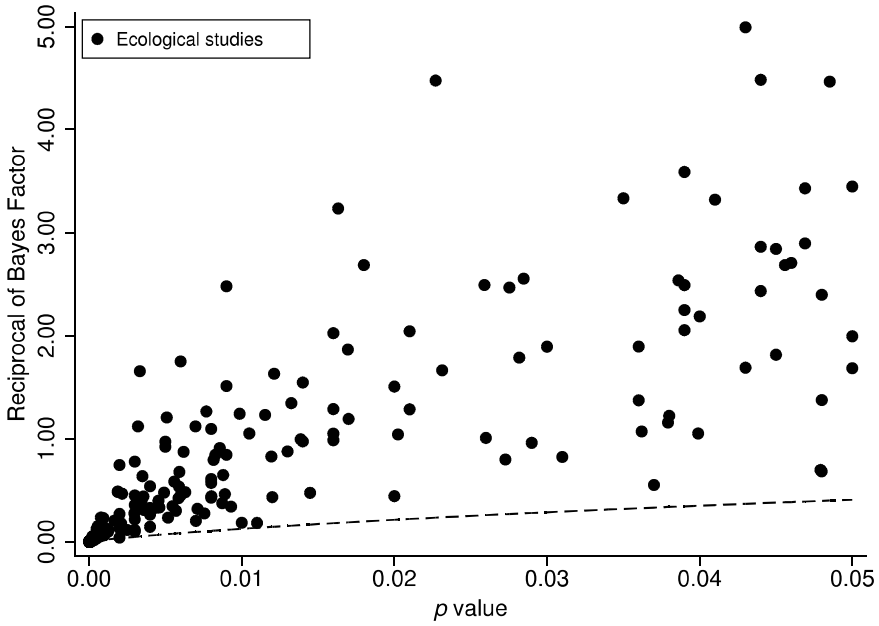
$$B_{01} \geq -e\ p\ \log(p)\,.$$

The first two graphs are for 272 'significant' epidemiological studies with two different choices of the prior; the third for 50 'significant' meta-analyses (these three from J.P. Ioannides, Am J Epidemiology, 2008); and the last is for 314 ecological studies (reported in Elgersma and Green, 2011).


Risk factors
Other
Toxic exposures
Biologic exposures
Psychosocial
Physical exposure
Body characteristics
Dietary
Reciprocal of Bayes Factor
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Risk factors
Other
Toxic exposures
Biologic exposures
Psychosocial
Physical exposure
Body characteristics
Dietary
Reciprocal of Bayes Factor
0.00
0.50
1.00
1.50
2.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Genetic associations
Reciprocal of Bayes Factor
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Ecological studies
Reciprocal of Bayes Factor
0.00
1.00
2.00
3.00
4.00
5.00
0.00
0.01
0.02
0.03
0.04
0.05
p value

## Posterior probabilities can equal $p$-values in one-sided testing:

- $X \mid \mu \sim N(x \mid \mu, \sigma^2)$
- One-sided testing

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

- Choose the usual estimation objective prior $\pi(\mu) = 1$, for which the posterior distribution, $\pi(\mu \mid x)$, can be shown to be $N(\mu \mid x, \sigma^2)$.
- Posterior probability of $H_0$:

$$\begin{aligned} \Pr(H_0 \mid x) = \Pr(\mu \leq \mu_0 \mid x) &= \Phi\left(\frac{\mu_0 - x}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{x - \mu_0}{\sigma}\right) = \Pr(X > x \mid \mu_0) = p\text{-value}\,. \end{aligned}$$

- Note that, here, we did not start out with prior probabilities of the hypotheses; these are *implied* by the prior distribution on the full parameter space $\mu \in (-\infty, \infty)$,
  - at least if we used $\mu \sim Uniform(-10^9, 10^9)$ instead of $\pi(\mu) = 1$.

# Ockham's Razor

- Attributed to thirteen-century Franciscan monk William of Ockham (Occam in latin)

  *"Pluralitas non est ponenda sine necessitate."*

  (Plurality must never be posited without necessity.)

  *"Frustra fit per plura quod potest fieri per pauciora."*

  (It is vain to do with more what can be done with fewer.)

- Preferring the simpler of two hypothesis to the more complex when both agree with data is an old principle in science.
- Regard $H_0$ as *simpler* than $H_1$ if it makes *sharper predictions* about what data will be observed.
- Hypotheses are more complex if they have extra adjustable parameters that allow them to be tweaked to accommodate a wider variety of data.
  - "coin is fair" is a simpler model than "coin has unknown bias $\mu$"
  - $s = a + ut + \frac{1}{2}\,gt^2$ is simpler than $s = a + ut + \frac{1}{2}\,gt^2 + ct^3$

**Example:** *Perihelion of Mercury* (with Bill Jefferys)

In the 19th century it was known that there was an unexplained residual motion of Mercury's perihelion (the point in its orbit where the planet was closest to the Sun) in the amount of approximately 43 seconds of arc per century.

Various hypotheses:

- A planet 'Vulcan' close to the sun.
- A ring of matter around the sun.
- Oblateness of the sun.
- $H_G$: Law of gravity is not inverse square but inverse $(2 + \epsilon)$.

All these hypotheses had a parameter that could be adjusted to deal with whatever data on the motion of Mercury existed.

**To test:** $H_G$ versus $H_E$: General Relativity.

**Data in 1920:** $X = 41.6$ where $X \sim N(\mu, 2^2)$, $\mu$ being the perihelion advance of Mercury, with measurement standard deviation of 2.

**Prior probabilities of hypotheses:** $Pr(H_G) = Pr(H_E) = 1/2$.

**Prior (before data) for gravity hypothesis $H_G$:** $\pi_G(\mu) = N(0, 50^2)$.

- Symmetric about 0 (corresponding to inverse square law).
- Decreasing away from zero; normality is convenient.
- Initially, $\tau = 50$, because a gravity effect which would yield $\mu > 100$ would have had other observed effects.
- We will also consider utilization of classes of priors:
  - The class of all $N(0, \tau^2)$ priors, $0 < \tau < 60$.
  - The class of all symmetric priors that are nonincreasing in $|\mu|$, having probability one on $|\mu| < 110$.

**General Relativity hypothesis $H_E$:** Predicted $\mu_E = 42.9$, so no prior is needed. (Thus General Relativity is the 'simpler' hypothesis.)

**Bayes factor of $H_E$ to $H_G$:**

$$
\begin{aligned}
B_{EG} &= \frac{f_E(41.6)}{\int f_G(41.6 \mid \mu)\pi_G(\mu)} \, d\mu \\
&= \frac{\frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{8}(41.6-\mu_E)^2\right)}{\int \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{8}(41.6-\mu)^2\right) \frac{1}{50\sqrt{2\pi}} \exp\left(-\frac{1}{2\cdot 50^2}\mu^2\right) \, d\mu} \\
&= \frac{\frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{8}(41.6-42.9)^2\right)}{\frac{1}{\sqrt{2\cdot 2504\pi}} \exp\left(-\frac{1}{2\cdot 2504}(41.6-0)^2\right)} = 28.6
\end{aligned}
$$

This strongly favors General Relativity, even though the gravity hypothesis could fit the data better than General Relativity.

If one considers the class of possible normal priors, $B_{EG} \in [27.76, 10^{93}]$.

If one considers the class of possible symmetric nonincreasing priors, $B_{EG} \in [15.04, 10^{93}]$.

# Plausibility of precise hypotheses

A *precise hypothesis* is an hypothesis of lower dimension than the alternative (e.g. $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$).

This is in contrast to, say, testing $H_0 : \mu < 0$ versus $H_1 : \mu > 0$,

**A Key Issue: Is the precise hypothesis being tested plausible?**
This is so if it has a reasonable prior probability of being true.

*Example:* Let $\mu$ denote the difference in mean treatment effects for cancer treatments A and B, and test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$.

Scenario 1: Treatment A = standard chemotherapy
Treatment B = standard chemotherapy + steroids

Scenario 2: Treatment A = standard chemotherapy
Treatment B = a new radiation therapy

$H_0 : \mu = 0$ is plausible in Scenario 1, but not in Scenario 2; in the latter case, instead test $H_0 : \mu < 0$ versus $H_1 : \mu > 0$.

**Plausible precise null hypotheses:**

- $H_0$ : Gene A is not associated with Disease B.
- $H_0$: There is no psychokinetic effect.
- $H_0$: Vitamin C has no effect on the common cold.
- $H_0$: A new HIV vaccine has no effect.
- $H_0$: Cosmic microwave background radiation is isotropic.
- $H_0$ : Males and females have the same distribution of eye color.
- $H_0$ : Pollutant A does not cause disease B.

**Implausible precise null hypotheses:**

- $H_0$ : Small mammals are as abundant on livestock grazing land as on non-grazing land
- $H_0$ : Bird abundance does not depend on the type of forest habitat they occupy
- $H_0$ : Children of different ages react the same to a given stimulus.

## Approximating a believable precise hypothesis by an exact precise null hypothesis

A precise null, like $H_0 : \mu = \mu_0$, is typically never true *exactly*; rather, it is used as a surrogate for a 'real null'

$$H_0^\epsilon : |\mu - \mu_0| < \epsilon, \quad \epsilon \text{ small}\,.$$

(Even if $\mu = \mu_0$ in nature, the experiment studying $\mu$ will typically have a small unknown bias, introducing an $\epsilon$.)

**Result** (Berger and Delampady, 1987 Statistical Science):
Robust Bayesian theory can be used to show that, under reasonable conditions, if $\epsilon < \frac{1}{4}\,\sigma_{\hat{\mu}}$, where $\sigma_{\hat{\mu}}$ is the standard error of the estimate of $\mu$, then

$$Pr(H_0^\epsilon \mid \mathbf{x}) \approx Pr(H_0 \mid \mathbf{x})\,.$$

*Note:* Typically, $\sigma_{\hat{\mu}} \approx \frac{c}{\sqrt{n}}$, where $n$ is the sample size, so for large $n$ the above condition can be violated, and using a precise null may not be appropriate, even if the real null is believable.

# Choice of Prior Distributions in Testing

## A. Choosing priors for "common parameters" in testing:

*Common parameters* in densities under two hypotheses are parameters that are present in both and have the *same role*.

Example: If the data $x_i$ are i.i.d. $N(x_i \mid \mu, \sigma^2)$, and it is desired to test

$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0\,,$$

the density under $H_0$ is $N(x_i \mid 0, \sigma^2)$ and that under $H_1$ is $N(x_i \mid \mu, \sigma^2)$, so $\sigma^2$ is a common parameter to both densities and has the same role in each.

**Priors for common parameters:** Use the standard 'objective prior' for the common parameters.

Example: For the normal testing problem, the standard objective prior for the variance is $\pi(\sigma^2) = 1/\sigma^2$.

## B. Choosing priors for non-common parameters

If subjective choice is not possible, be aware that

- Vague proper priors are often horrible: for instance, if $X \sim N(x \mid \mu, 1)$ and we test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ with a Uniform$(-c, c)$ prior for $\mu$, the Bayes factor is

$$B_{01}(c) = \frac{f(x \mid 0)}{\int_{-c}^{c} f(x \mid \mu)(2c)^{-1} d\mu} \approx \frac{2c\, f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu) d\mu} = 2c\ f(x \mid 0)$$

  for large $c$, which depends dramatically on the choice of $c$.

- Improper priors are problematical, because they are unnormalized; should we use $\pi(\mu) = 1$ or $\pi(\mu) = 2$, yielding

$$B_{01} = \frac{f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu)(1) d\mu} = f(x \mid 0) \quad \text{or} \quad B_{01} = \frac{f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu)(2) d\mu} = \frac{1}{2} f(x \mid 0)\,?$$

- It is curious here that use of vague proper priors is much worse than use of objective improper priors (though neither can be justified).

## Various proposed default priors for non-common parameters

- Conventional priors
  - Jeffreys choice and generalizations
- Priors induced from a single prior
- Intrinsic priors (derived from data or imaginary data)
- Fractional priors
- Expected posterior priors

# Conventional priors

**Jeffreys choices for the normal testing problem:**

- *Data:* $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$
- We are testing
$$H_0 : X_i \sim N(x_i \mid 0, \sigma_0^2) \quad \text{versus} \quad H_1 : X_i \sim N(x_i \mid \mu, \sigma_1^2).$$
- We thus seek $\pi_0(\sigma_0^2)$ and $\pi_1(\mu, \sigma_1^2) = \pi_1(\mu \mid \sigma_1^2)\pi_1(\sigma_1^2)$.
- Since $\sigma_0^2$ and $\sigma_1^2$ are common parameters with the same role, Jeffreys used the same objective prior $\pi_0(\sigma_0^2) = 1/\sigma_0^2$ and $\pi_1(\sigma_1^2) = 1/\sigma_1^2$.
- $\pi_1(\mu \mid \sigma_1^2)$ must be proper (and not vague), since $\mu$ only occurs in $H_1$. Jeffreys argued that it
  - should be centered at zero ($H_0$);
  - should have scale $\sigma_1$ (the 'natural' scale of the problem);
  - should be symmetric around zero;
  - should have no moments (more on this later).

  The 'simplest prior' satisfying these is the $Cauchy(\mu \mid 0, \sigma_1)$ prior,

resulting in

**Jeffreys proposal:**

$$\pi_0(\sigma_0^2) = \frac{1}{\sigma_0^2}\,, \qquad \pi_1(\mu, \sigma_1^2) = \frac{1}{\pi\sigma_1(1+(\mu/\sigma_1)^2)} \cdot \frac{1}{\sigma_1^2}\,.$$

**Predictive matching argument for these priors:**

For any location scale density $\frac{1}{\sigma}g(\frac{y-\mu}{\sigma})$ and one observation $y$

under $H_1 : \mu = 0$,

$$m_0(y) = \int \frac{1}{\sigma}\, g\left(\frac{y-\mu}{\sigma}\right)\, \frac{1}{\sigma^2}\, d\sigma^2 = \frac{2}{|y|}\,;$$

under $H_1 : \mu \neq 0$ and for any proper prior of the form $\frac{1}{\sigma}h\left(\frac{\mu}{\sigma}\right)$,

$$m_1(y) = \int \frac{1}{\sigma}\, g\left(\frac{y-\mu}{\sigma}\right)\, \frac{1}{\sigma}h\left(\frac{\mu}{\sigma}\right)\, \frac{1}{\sigma^2}\, d\mu\, d\sigma^2 = \frac{2}{|y|}\,,$$

so that $B_{01} = 1$ for one observation, as should be the case. (Of course, this doesn't say that the prior for $\mu$ should be Cauchy.)

**Sensitivity:** The choice of $\sigma_1$ for the prior under $H_1$ matters a lot.

## An example of inducing prior probabilities of hypotheses through an overall prior on the parameter space:

$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{C})$, where $\boldsymbol{C}$ is a correlation matrix with all correlations equal to $\rho$. It is desired to test

$$H_0 : \rho < 0 \quad \text{versus} \quad H_1 : \rho > 0\,.$$

(If $H_2 : \rho = 0$ is a plausible hypothesis, it should be added.)

- Choose $\pi(\boldsymbol{\mu}, \sigma^2) = 1/\sigma^2$ for the common parameters in $H_0$ and $H_1$.
- It might be tempting to choose $\pi(\rho) \propto 1$ for each hypotheses (or, better, the *reference prior* $\pi(\rho) \propto 1/\sqrt{1-\rho^2}$).
  - This is fine if $p = 2$, since the parameter space is then $\rho \in (-1, 1)$, so the induced prior probabilities of hypotheses are $Pr(H_0) = Pr(H_1) = 1/2$.
  - But $-(p-1)^{-1} < \rho < 1$ so, for $p \geq 3$, this would be assigning prior probabilities to each hypothesis of

$$Pr(H_0) = \int_{-(p-1)^{-1}}^{0} \frac{1}{(1+(p-1)^{-1})}\, d\rho = \frac{1}{p}, \quad Pr(H_1) = 1 - Pr(H_0) = 1 - \frac{1}{p}.$$

- Alternatively, the standard hypothesis testing approach woould be to choose $Pr(H_1) = 1 - Pr(H_0) = 1/2$, and then choose objective proper priors for the parameter under each hypotheses, e.g. $\pi_0(\rho) = (p-1)$ and $\pi_1(\rho) = 1$ (or the reference prior versions).
- These two approaches give very different answers and either could be correct.

## Another example - Equivalence Testing:

- $p_1$ is the cure rate of existing Drug 1 and $p_2$ is the cure rate of proposed generic Drug 2 (to provide a cheap alternative to Drug 1).
- The generic drug will be approved if $|p_1 - p_2| < 0.03$.
- Thus we are testing $H_1 : |p_1 - p_2| < 0.03$ versus $H_2 : |p_1 - p_2| > 0.03$.
- Regulatory agencies want to test with $Pr(H_1) = Pr(H_2) = 1/2$.
  - Past data on $p_1$ is summarized by the posterior $\pi(p_1)$ is $N(0.6, 0.02^2)$. But the new clinical trial will have a bias $b \sim N(0, 0.03^2)$, so use $\pi^*(p_1)$ equal to $N(0.6, 0.0013)$ (for both hypotheses).
  - Under $H_1$, choose the prior $\pi_1(p_2 \mid p_1) = Uniform(p_1 - .03, p_1 + 0.03)$, so that $\eta = p_1 - p_2$ is uniform on $(-0.03, 0.03)$.
  - Under $H_2$, choose $\pi_2(p_2 \mid p_1)$ to be uniform on $(p_1 - .03 - s, p_1 - .03) \bigcup (p_1 + .03, p_1 + .03 + s)$.
    - The generic company will suggest a choice of $s$ based on their studies.
    - The regulatory agency might choose $s$ based on past generic drug trials.
    - Probably best is to report the answer as a function of $s$, seeking a robust answer.

## The Neutrino problem

**To test** (before cosmological and neutrino oscillation data):

$$H_N : 0 < \nu_1 < \nu_2 < \nu_3 \quad \text{versus} \quad H_I : 0 < \nu_3 < \nu_1 < \nu_2, \, .$$

- Cosmological data implies things like $\sum \nu_i < 0.12$.
- Neutrino oscillation data gives constraints on some of the $(\nu_i - \nu_j)^2$.

**Priors that have been utilized** (before looking at the data):

- All properly assign essentially equal prior mass to $H_N$ and $H_I$.
- Choices of priors include the uniform, Jeffreys, reference, and a hierarchical prior.
- Most priors are developed for the $\nu_i$, but at least one is developed for the $\log(\nu_i)$.
- Answers vary from Bayes factors (evidence) of 5 to 1 in favor of $H_N$ to 42 to 1 in favor.

## Intrinsic priors

Discussion of these can be found in Berger and Pericchi (2001). One popular such prior, that applies to our testing problem, is the *intrinsic prior* defined as follows:

- Let $\pi^O(\mu)$ be a good estimation objective prior (using a constant prior will almost always work fine), with resulting posterior distribution and marginal distribution for data $\boldsymbol{x}$ given, respectively, by

$$\pi^O(\mu \mid \boldsymbol{x}) = f(\boldsymbol{x} \mid \mu)\pi^O(\mu)/m^O(\boldsymbol{x}), \quad m^O(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \mu)\pi^O(\mu)\, d\mu\,.$$

- Then the intrinsic prior (which will be proper) is

$$\pi^I(\mu) = \int \pi^O(\mu \mid \boldsymbol{x}^*) f(\boldsymbol{x}^* \mid \mu_0)\, d\boldsymbol{x}^*\,,$$

with $\boldsymbol{x}^* = (x_1^*, \ldots, x_q^*)$ being imaginary data of the smallest sample size $q$ such that $m^O(\boldsymbol{x}^*) < \infty$ (this is an imaginary bootstrap construction).

$\pi^I(\mu)$ is often available in closed form, but even if not, computation of the resulting Bayes factor is often a straightforward numerical exercise.

- The resulting Bayes factor is

$$B_{01}(\boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \mu_0)}{\int f(\boldsymbol{x} \mid \mu)\pi^I(\mu)d\mu} = \frac{f(\boldsymbol{x} \mid \mu_0)}{\int m^O(\boldsymbol{x} \mid \boldsymbol{x}^*)f(\boldsymbol{x}^* \mid \mu_0)d\boldsymbol{x}^*} .$$

*Example* (Higgs Boson Example): Test $H_0 : \mu = 0$ versus $H_0 : \mu > 0$, based on $X_i \sim f(x_i \mid \mu) = (\mu + b)\exp\{-(\mu + b)x_i\}$, where $b$ is known;

- Suppose we choose $\pi^O(\mu) = 1/(\mu + b)$ (the more natural square root is harder to work with).
- A minimal sample size for the resulting posterior to be proper is $q = 1$.
- Computation then yields
  $\pi^I(\mu) = \int \pi^O(\mu \mid x_1^*)f(x_1^* \mid 0)dx_1^* = b/(\mu + b)^2$.

# Dealing with interval probabilities

*The problem:* We know $p_i \in (a_i, b_i)$, $i = 1, \ldots, m$. Of interest: $P = \prod_{i=1}^{m} p_i$.

Example: A device has $m$ components, each of which will function with probability $p_i$. $P$ is the probability the device will function.

*Standard IP answer:* State that $P \in (\prod_i a_i, \prod_i b_i)$.

Example: $P \in (0.4, 0.98)$, a probably useless answer.

*Bad alternative:* use *midpoints*, so $P = \prod_i [(a_i + b_i)/2]$.

Example: $P = 0.92$, ignoring the IP issue.

*Laplace alternative (inverse probability):* $p_i \sim \text{Uniform}(a_i, b_i)$, find the equal-tailed 95% Bayesian confidence interval for $P$.

Example: $P \in (0.91, 0.94)$, sensible, but not a fully IP solution.

*Higher level IP:* Model actual beliefs about $p_i \in (a_i, b_i)$, but stay within the IP framework.

- Values near the midpoints are often more likely than the endpoints.
- Beliefs are typically symmetric and unimodal in the intervals.
- Reflect IP concerns by forming, for all $i$, the credal set $\mathcal{P}_i$ of all distributions with the above two properties.
  - Often the $p_i$ are dependent, which would have to be incorporated into the overall class $\mathcal{P}$ of possible priors, challenging to do in an IP way.
- Find the extremal 95% confidence interval – here, the union of all 95% equal-tailed Bayesian confidence intervals from priors in $\mathcal{P}$.

Example: If the $p_i$ are independent, $P \in (0.91, 0.94)$. (The extremal 95% CI happens to be that from the uniform priors, in the independent case.)

## Comment on Unfolding

Shyamalkumar (unpublished) had the following interesting result about finding $\pi(\mu)$ such that

$$m_\pi(x) = \int f(x \mid \mu)\pi(\mu)\, d\mu$$

is as close as possible to an estimated $\hat{m}(x)$:

- choose any initial $\pi_0(\mu)$ that has support everywhere;
- iteratively compute

$$\pi_l(\mu) = \int \pi_{l-1}(\mu \mid x)\hat{m}(x)dx\,, \quad \text{where } \pi_{l-1}(\mu \mid x) = \frac{\pi_{l-1}(\mu)f(x \mid \mu)}{\int \pi_{l-1}(\mu)f(x \mid \mu)d\mu}\,.$$

*Fact:* $\pi^*(\mu) = \lim_{l\to\infty} \pi_l(\mu)$ is the density for which $m_\pi(x)$ is as close as possible to $\hat{m}(x)$ in Kullback-Leibler divergence.

## A Possibly Interesting Implementation via Particle Filtering:

- Represent $\pi_l(\mu)$ by a collection of *particles* $\{\mu_i\}$ with weights $\{w_i^{(l)}\}$. (Initialize with a random sample $\{\mu_i\}$ from $\pi_0(\mu)$, so the initial weights are equal.)
- Then $\pi_l(\mu \mid x)$ would be the same collection of particles but with modified weights

$$w_i^{(l)}(x) = \frac{w_i^{(l-1)} f(x \mid \mu_i)}{\sum_j w_j^{(l-1)} f(x \mid \mu_j)} ,$$

  and $\pi_l(\mu)$ would be the same collection of particles but with weights

$$w_i^{(l)} = \int w_i^{(l)}(x) \hat{m}(x) dx .$$

- As one progresses one will need to add new particles adapting to the evolving density, but there are likely techniques in particle filtering for doing this.

## Gott and the "doomsday argument:"

What is the posterior distribution of $N$, the total number of humans that will ever exist?

An individual - Joe - is just born and is number $n$ in the ordered list of all possible humans. ($n$ is 6 billion or so.)

- Since Joe "could have equally likely been born anywhere in the list," $p(n \mid N) = Uniform\{1, 2, \ldots, N\}$.
- $p(N) = 1/N$ (the 'right' objective Bayes prior distribution for $N$).
- Bayes theorem yields a posterior distribution for $N$,

$$p(N \mid n) \approx \frac{n}{N^2} \quad \text{on} \quad \{n+1, n+2, \ldots\}\,.$$

Mules: what if 3rd on made or one made today (with mules on the way out)

# The Jeffreys-Lindley 'Paradox'and Experimental Bias

Suppose that $H_0$ is truly precise (e.g. 0 psychic effect or 'no Higgs boson'), but that the experiment has some bias $b \sim N(b \mid 0, \delta^2)$. Then

$$\begin{aligned} \Pr(H_0 \mid \bar{x}) &= \left[ 1 + \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\exp\{\frac{1}{2}\, z_b^2 \, [1 + \frac{(\delta^2 + \sigma^2/n)}{v_0^2}]^{-1}\}}{\{1 + \frac{v_0^2}{\delta^2 + \sigma^2/n}\}^{1/2}} \right]^{-1} \\ &\approx 1 - \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\sqrt{\delta^2 + \sigma^2/n}}{v_0} \, \exp\{\frac{1}{2}\, z_b^2\}, \end{aligned}$$

when $\sqrt{\delta^2 + \sigma^2/n}$ is small, and where $z_b = |\bar{x}|/\sqrt{\delta^2 + \sigma^2/n}$ can be thought of as standard normal under $H_0$ in the presence of the bias. Then

$$\lim_{n \to \infty} \Pr(H_0 \mid \bar{x}) = 1 - \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\delta}{v_0} \, \exp\{\frac{(\bar{x})^2}{2\delta^2}\}$$

which does not go to 1. Also

$$\Pr(H_0 \mid \bar{x}) \approx 1 - \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\delta}{v_0} \, \exp\{\frac{z^2\sigma^2}{2n\delta^2}\},$$

for the interesting range of $z^2\sigma^2/[n\delta^2]$.

## Experimental biases:

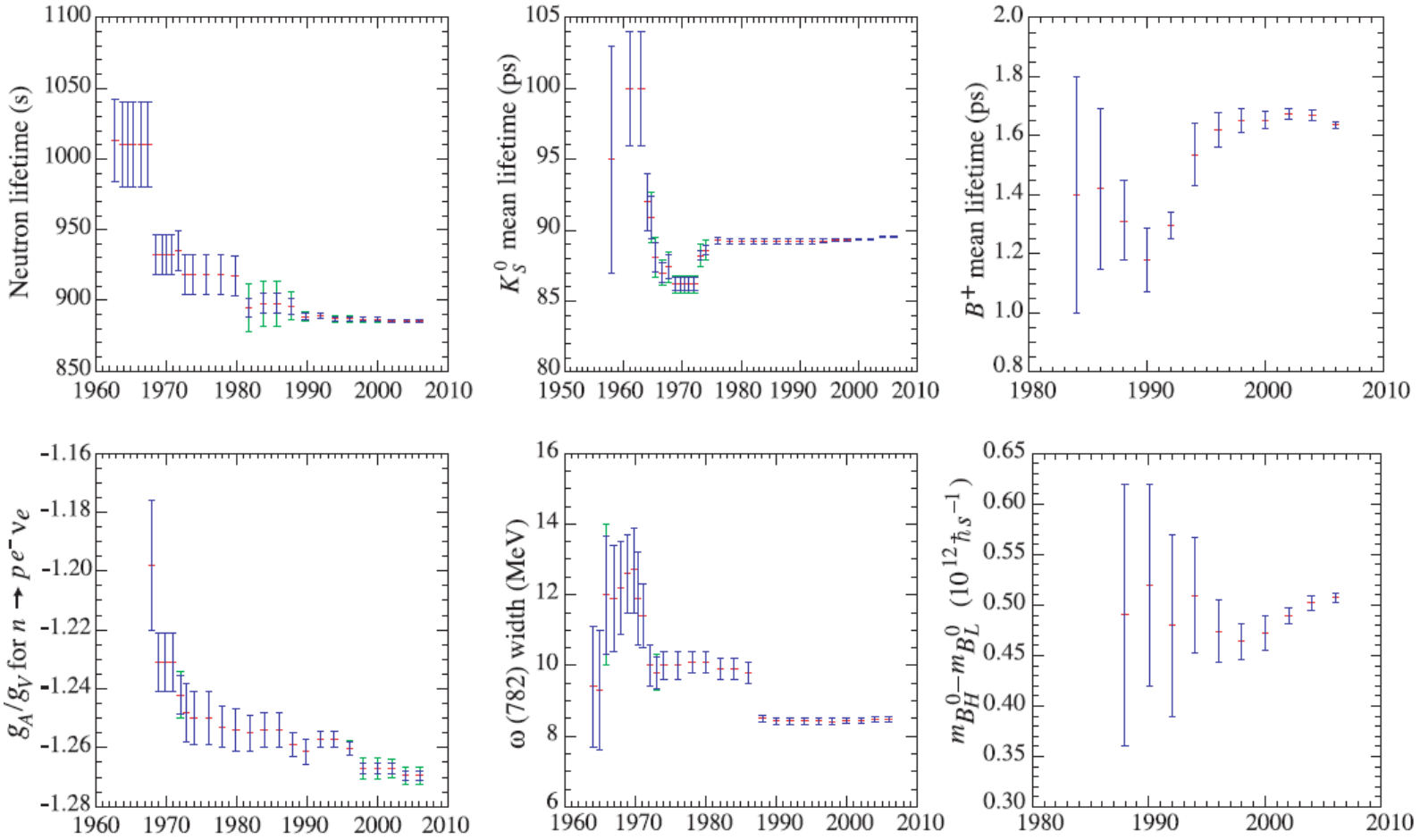



Figure 2: Historical record of values of some particle properties published over time, with quoted error bars (Particle Data Group).

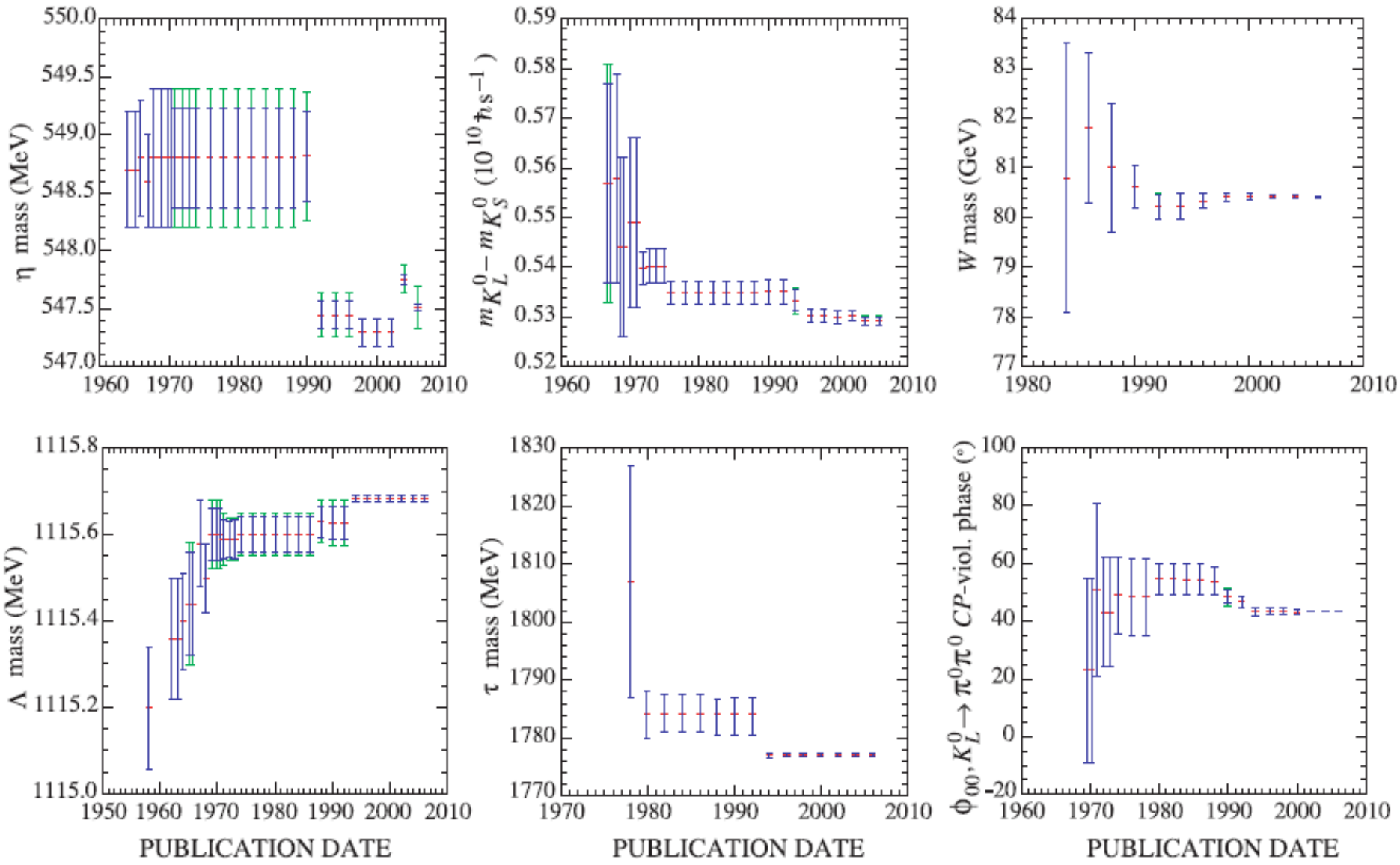



Figure 3: Historical record of values of some particle properties published over time, with quoted error bars (Particle Data Group).