Statistical Quantification of Discovery in Physics

Pontecorvo School 2017

Professor David A. van Dyk

Statistics Section, Department of Mathematics, Imperial College London
dvandyk@imperial.ac.uk

August 29, 2017

Contents

1 Background

I am a statistician, not a neutrino physicist...

- I collaborate with astro, solar, and particle physicists on statistical methodology.
- First contact with neutrino physics: PhyStat-ν *...about a year ago*

Today: Summarize statistical issues pertaining to discovery in (neutrino) physics.

1.1 Motivating Examples

Neutrino Oscillation

- Neutrino created as electron, muon or tau may later be measured with different flavor.
- Flavor probability varies periodically as neutrino travels; *depends on several parameters.*

Mass Hierarchy

....ordering of the mass eigenstates

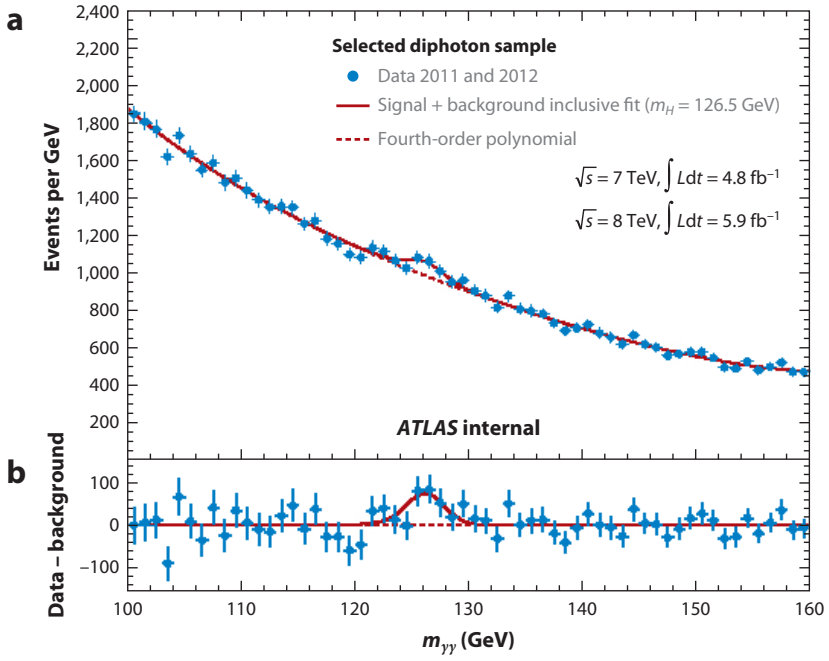
- We would like to compare the "normal hierarchy" (i.e, mass difference, $\Delta m^2_{32} > 0$) with the "inverted hierarchy" (i.e., $\Delta m^2_{32} < 0$).
- Which is more consistent with data? Which is correct?
- **Challenge:** While $|\Delta m^2_{32}|$ is well constrained, the sign is degenerate with θ_{23} or δ_{CP}.

CP-violation

- Is there evidence to counter the assumption that $\delta_{\text{CP}} \in \{0, \pi\}$?
- The current data is limited.

Bump Hunting (e.g., Higgs serach)

- **Question:** is there a bump above background or not?
- The location of the possible bump is unknown.
- What is the bump location if there is no bump? What does this mean?



We return to these examples during the lectures.

- Discuss both philosophical and technical differences between the three examples.

1.2 Outline of Statistical Methods Covered

There are two predominant statistical perspectives on how to formulate the question of Discovery.

Frequentist Model Selection: Suppose there were no new _____. What is the chance that we would see data as extreme as this?

Higgs Search:

- If there were no Higgs boson, what would be the chance that we would see a bump at least this large?
- What is the chance that a random fluctuation in background would result in a bump at least this large?
- This chance is larger if we consider a wider search region.

CP Violation:

- If there is no CP-violation, what is the chance of seeing data as extreme or more extreme than the data we have?
- "As extreme": We must define a statistic that is small [large] if there is no CP-violation but becomes larger [smaller] with increased CP-violation.

Mass Hierarchy:

- The situation is more complicated because there is not default model.

Bayesian Model Selection: Give the data that we have observed, what is the probability of the a new _____?

Higgs Search: Give the observed data, what is the probability that there is a Higgs boson?

CP Violation: Give the observed data, what is the probability of a CP-violation?

Mass Hierarchy: Give the observed data, what is the probability of the normal hierarchy?

NOTES:

1. Bayesian discovery is conceptually easier.
2. There is a reversal in conditioning between frequency-based and Bayesian discovery.
   - Frequency methods compute probabilities of data given a model (e.g., background only, no Higgs).
   - Bayesian methods compute the probability of the model given the data.
   - $\Pr(A \mid B)$ may be quite different than $\Pr(B \mid A)$.

     Example:

3. Main challenges
   - Do frequency method really answer the right question?
   - Bayesian answers depend on the choice of prior distribution.
   - The two perspectives may give seemingly contradictory results.
4. Because of this model selection remains controversial and challenging.
   - Model selection is harder that parameter estimation, both conceptually and technically.
   - This is particularly challenging because the science questions have a higher profile and are generally more central.

   Higgs Search: Compare the discovery of the Higgs boson to follow up studies that refine the estimates of its mass. Which gets more press?
5. To understand these subtleties, we must first compare frequency-based and Bayesian parameter estimation.

1.3 Likelihood Based Statistical Methods

Example 1. Consider a single bin detector where the data is an event count and we wish to estimate the source count rate per unit time, λ. (For simplicity we assume there is not background contamination.) Denote the observed event count by y and suppose that

$$y \sim \text{POISSON}(t\lambda),$$

where t is time.

Definition. *The Likelihood Function: The likelihood function is the distribution function (formally the probability density or probability mass function) of the data given the model parameters.*

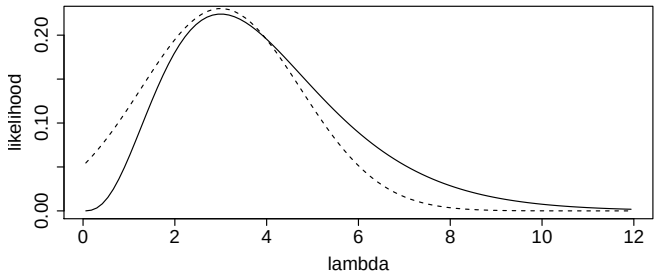
Generic notation: We denote the data by $\boldsymbol{y}$, the parameter by $\boldsymbol{\theta}$, and the likelihood function by $L(\boldsymbol{\theta} \mid y)$, or more succinctly $L(\boldsymbol{\theta})$ Both $\boldsymbol{y}$ and $\boldsymbol{\theta}$ may be multivariate (vectors); we use bold to represent vectors.

In our example,

$$\begin{aligned} \text{likelihood}(\lambda \mid y) &= e^{-t\lambda}(t\lambda)^y/y! \equiv L(\lambda \mid y) \\ \log \text{likelihood}(\lambda \mid y) &= -t\lambda + y\ln(t\lambda) - \ln(y!) \end{aligned}$$

Definition. *Maximum Likelihood Estimate (MLE): The value of the unknown parameter that maximizes the likelihood function (among the values in the range of scientifically feasible values). We denote the generic MLE by* $\hat{\theta}_{\text{MLE}}$.

Example 1 (con't). Returning to Example 1, it is easy to show that $\hat{\lambda}_{\text{MLE}} = y/t$. Below is a plot of the likelihood function with $t = 1$ if we observe $y = 3$.



The likelihood (solid line) and its normal approximation (dashed line).

NOTES:

- The normal approximation is the normal curve with the same mode and curvature at the mode.
- The standard likelihood-based error bar is computed as the standard deviation of this normal approximation. For a univariate parameter, the error bar is computed as

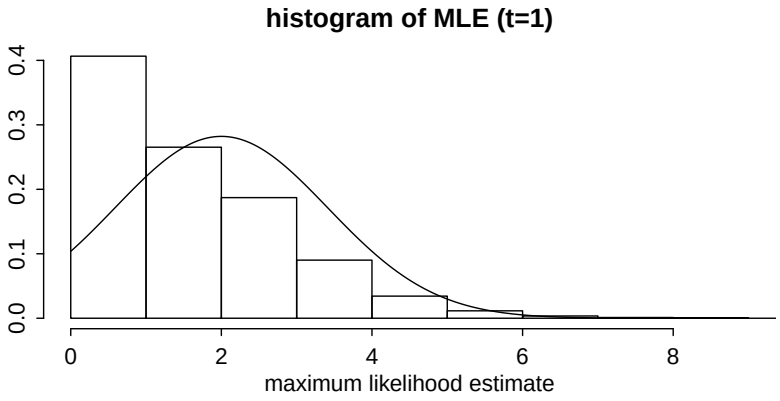
$$\text{se}(\hat{\theta}) = -\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln L(\theta \mid y)\big|_{\theta=\hat{\theta}}\right)^{-1}.$$

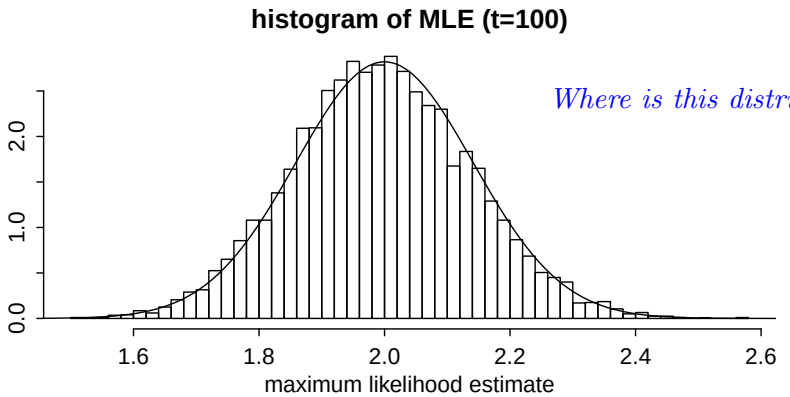
- Under the the "central limit theorem" the likelihood and its normal approximation become evermore similar for larger data sets.
- Given that likelihood based method condition "backwards", what is the justification for their use?

SIMULATION: (Example 1, con't)

1. Suppose $\lambda = 2$ and $t = 1$. Repeat the experiment 1000 times and consider the distribution of $\hat{\lambda}$.
2. Repeat but with $t = 100$.

Results:





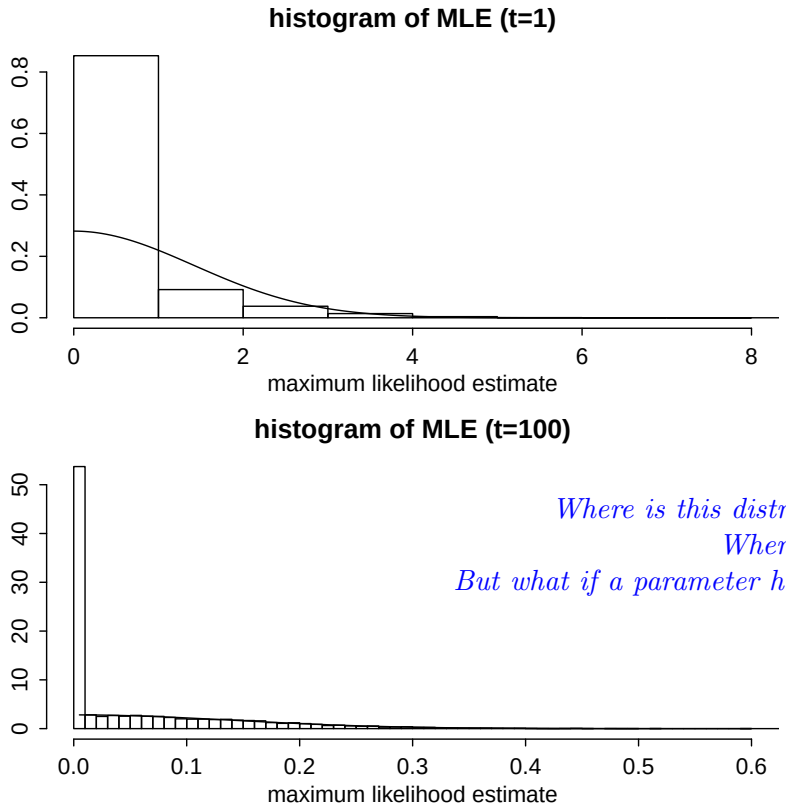
Where is this distribution centered?

Now suppose

$$y \sim \text{POISSON}\Big(t(2+\lambda)\Big),$$

i.e., we expect 2 counts due to background per unit time. It can be shown that under this model the MLE is $\hat{\lambda}_{\text{MLE}} = \max\Big(0, (y-2)/t\Big)$. We repeat the simulation under this model, with $\lambda = 0$ (i.e., no source).

Results:



Where is this distribution centered?
Where does it pile up?
But what if a parameter has no true value?

Asymptotic Frequency Properties of the MLE: For large data sets (and univariate θ)
If θ is not on the boundary of its set of possible values, (among other "regularity conditions")

1.

2. The interval

$$\hat{\theta} \pm \qquad \text{includes } \theta \quad \_\_\_\_\_ \ \% \text{ of the time.}$$

(That is, $\hat{\theta} \pm 1.96 \text{ se}(\hat{\theta})$ is an (asymptotic) 95% confidence interval for θ.)

Definition. *Asymptotic Properties: Properties of an estimator or procedure that hold more and more exactly as the size of the data grows.*

1.4 Bayesian Statistical Methods

Bayesian methods are based completely on the concepts and tools of probability. That is, they use probability density/mass functions, means, variances, and probability intervals. There is no need for new structures such as the likelihood function, maximum likelihood estimates, or confidence intervals. Bayesian statisticians accomplish this by assigning subjective probability distributions to the unknown parameters that are the subject of statistical inference.

Example 2. Consider again a simple Poisson model: $Y|\lambda \sim \text{Poisson}(t\lambda)$ and let $L(\lambda \mid y)$ be the likelkhood function. We aim to infer what values of λ are most probable *after* having observed y. We begin by using a probability distribution to describe our knowledge about likely values of λ *before* observing y.

Definition. *The prior distribution summarizes our state of knowledge before we observe the data and is denoted* $p(\boldsymbol{\theta})$.

Definition. *The posterior distribution summarizes our state of knowledge after we observe the data and is denoted* $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

We use the posterior distribution to infer what value of the parameter, $\boldsymbol{\theta}$ are probable, after having observed the data. $\boldsymbol{y}$. For simplicity here we assume θ is univariate.

Theorem. *(Bayes Theorem) Let* $\boldsymbol{y}$ *be a random vector and* θ *be random variable, with* $\boldsymbol{y}$ *representing the data and* θ *the model parameter. The posterior distribution of* θ *is given by*

$$p(\theta \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \theta)p(\theta)}{p(\boldsymbol{y})} \propto L(\theta \mid \boldsymbol{y})p(\theta),$$

where $p(\boldsymbol{y} \mid \theta) = L(\theta \mid \boldsymbol{y})$ *is the model for the data given the parameter,* $p(\theta)$ *is the prior distribution, and*

$$p(\boldsymbol{y}) = \int_{-\infty}^{\infty} p(\boldsymbol{y} \mid \theta)p(\theta)\mathrm{d}\theta$$

if θ *is continuous and*

$$p(\boldsymbol{y}) = \sum p(\boldsymbol{y} \mid \theta)p(\theta)$$

if θ *is discrete.*

Bayesian Parameter Estimation: We can estimate the unknown parameter, θ, by its *posterior mean*, $\mathrm{E}(\theta \mid \boldsymbol{y})$, and describe uncertainty in this estimate with its *posterior variance*, $\mathrm{Var}(\theta \mid \boldsymbol{y})$, or its *posterior standard deviation*, $\sqrt{\mathrm{Var}(\theta \mid \boldsymbol{y})}$.

Bayesian Interval Estimation: We can compute a *posterior probability interval* for Θ. (We assume Θ is continuous, but similar definitions can be made if Θ is discrete.)

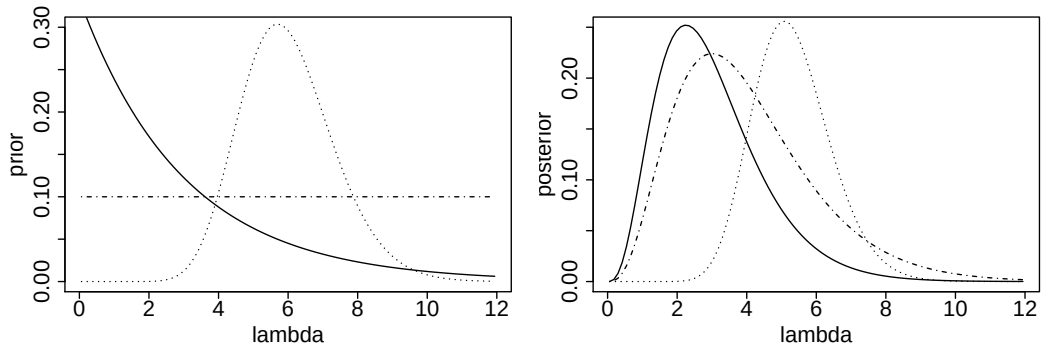
Definition. *For any* $\alpha \in [0, 1]$, *a* $100 \times (1 - \alpha)\%$ *posterior interval for the parameter,* θ *is any interval* $\mathcal{I}$, *such that*

$$\int_{\mathcal{I}} p(\theta \mid \boldsymbol{y})\mathrm{d}\theta = 1 - \alpha$$

Example 2 (con't). Returning to the Example 2,

$$\text{posterior}(\lambda) \propto \text{Likelihood}(\lambda) \ \times \ \text{prior}(\lambda)$$

The posterior distribution combines past information with that contained in the and current data. Again supposing that $y = 3$, we can compare three different prior distributions:



NOTES:

1. For small samples the prior distribution can be quite influential.
2. Without its asymptotic properties, the MLE also has little justification for small samples.
3. The effect of the prior dissipates as the the size of the data grows. The posterior mean and standard deviation enjoy essentially the same asymptotic frequency properties as the MLE and se($\hat{\theta}$).
4. For large data sets, Bayesian and Likelihood-based methods tend to give similar results.

1.5 A Statistical Framework for Discovery

Statistically, discovery is formulated in terms of model or hypothesis testing.

H_0: The null hypothesis (e.g., no CP-violoation, $\delta_{\text{CP}} = 0$) or no Higgs boson

H_A: The alternative hypothesis (e.g., CP-violation)

Without further evidence, H_0 is presumed true.

- "Deciding" on H_A means scientific discovery: new physics. (Higgs Boson or CP-violation)

Note. *Sometimes there is no presumed model. For example in the mass hierarchy problem, neither the normal nor the inverted hierarchy is a null-model. We may refer to Model Selection rather than testing in such cases.*

Errors.

TYPE-I: If we decide on H_A when actually H_0 is true, the result is a *False Detection* or *False Discovery*. Statisticians refer to this as a *Type-I Error*.

TYPE-II: There is another type of error. If H_A holds and we conclude there is insufficient evidence to reject H_0, the result is a missed detection. Statisticians refer to this as a *Type-II Error*.

As we shall see, the **choice of appropriate statistical approach** depends on a mix of philosophical and technical issues:

- Is there a *presumed* model? If not, what is meant by H_0?
- Are there more than 2 possible models? How can the framework be generalized?
- Are the models nested. That is, is H_0 a special case of H_A?
- Are there unknown parameters under H_0? We call such parameters, nuisance parameters.
- Are there parameters that have no value under H_0?

  Example:

- Are the values of the parameter under H_0 on the boundary of the set of possible values of the parameter under H_A?

  Example:

- Bayesian vs. Frequentist methods.

2 Frequentist Hypothesis Testing

2.1 Neyman Pearson

Example 3. Consider again a single bin detector where the data is an event count. Now suppose we wish to test for signal above background. Again denote the observed event count by y but now suppose that

$$y \sim \text{POISSON}(\lambda_B + \lambda),$$

where λ_B is the expected background count and λ is the expected source count. We wish to choose between the two hypotheses

$$H_0 : \lambda = 0 \quad \text{and} \quad H_A : \lambda > 0$$

Question: Why is this the appropriate choice of H_0 and H_A?

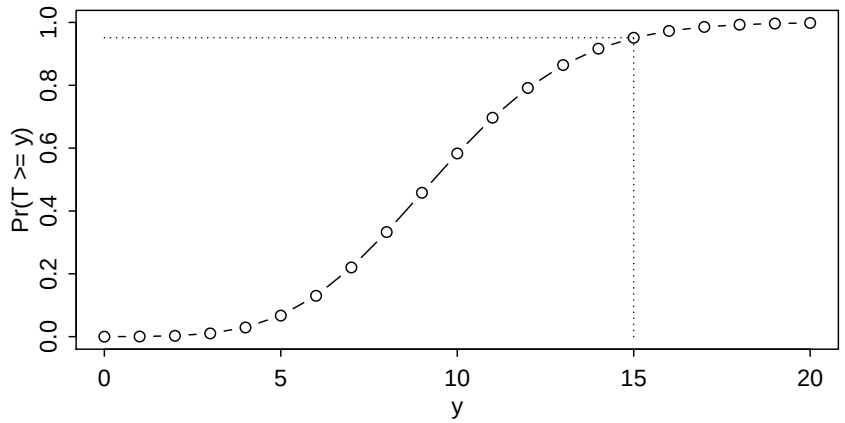
Neyman-Pearson Framework for Model Testing: Wish to choose between H_0 and H_A.

- Assume H_0 to be true unless we are overwhelmed by evidence to the contrary.
- Require a *test statistic*, T, with known distribution under H_0 and power under H_A. Common choices for the the test statistic include $\Delta\chi^2$ or the likelihood ratio statistic.
- Compute a threshold $T^\star$ that is (say) the smallest value such that

  $$\text{tail probability:} \quad \Pr(T > T^\star \mid H_0) \leq \alpha,$$

  where α is the bound on the probability of falsely discovery, i.e., of falsely rejecting H_0.
- If $T > T^\star$ we conclude that there is sufficient evidence to reject H_0.

Example 3 (con't). In Example 3, a simple choice of Test statistic is $T = y$. Suppose $\lambda_B = 10$ and we wish to limit the probability of false detection to be less than 5%. Under H_0, $T \sim \text{POISSON}(10)$ and $\Pr(T > 14 \mid H_0) = 0.0487$, so we reject H_0 in favor of H_A is $y \geq 15$.



NOTES:

1. Well-defined frequency properties: Bounded Pr(false detection).
2. There is no characterization of the strength of the evidence. You reject H_0 if $T > T^\star$ with no distinction made if T exceeds $T^\star$ by a mile or millimeter.
3. The test statistic can be difficult to find, especially when there are nuisance parameters. What if λ_B where unknown in Example 3?
4. What happens if neither hypothesis holds? What do Type I/II *errors* mean then?

2.2 Fisher's P-value

The Neyman-Pearson framework allows a decision between H_0 and H_A, but provides no measure of the degree of support in the data for one hypothesis or the other.

Key Question: Is the current data plausible under the H_0?

To quantify the degree of evidence, a *p-value* is often reported:

$$\text{p-value} = \Pr(T > T^{\text{obs}} \mid H_0).$$

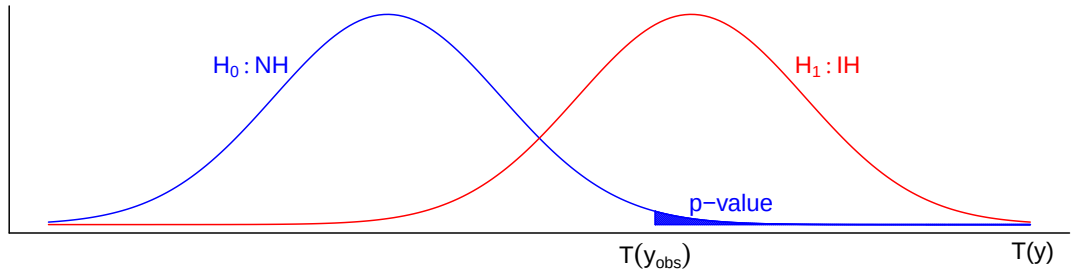
The p-value is the most commonly used criterion for hypothesis testing in general and discovery in physics.

Example 3 (con't). Suppose $T^{\text{obs}} = 9$, what is the p-value? Or $T^{\text{obs}} = 19$?

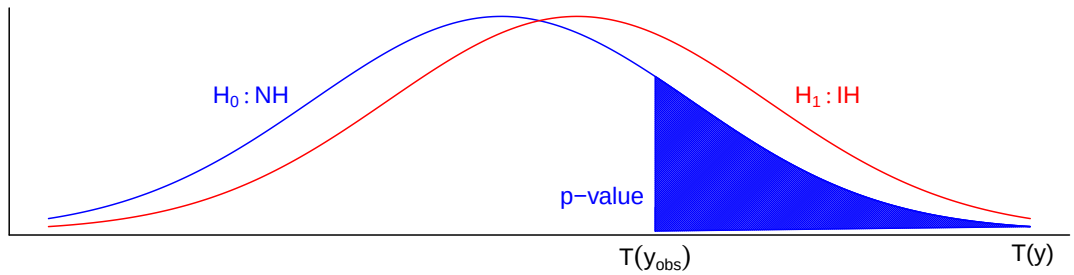
Example 4. Consider the mass-hierarchy example and suppose we wish to compare

$$H_0 : \text{Normal hierarchy} \quad \text{versus} \quad H_A : \text{Inverted hierarchy}$$

Case 1



Case 2



The distribution of $T(y)$ under H_0 (blue) and under H_A (red). The test statistic is more powerful in Case 1 than in Case 2.

NOTES:

1. This formulation assumes a test statistic, T, can be found whose distribution is completely specified under H_0. (There are no nuisance parameters.)

2. No alternative hypothesis is required for the p-value. Fisher originally intended that the p-value be used for model checking rather than model selection.

2.3 Likelihood Ratio Test and Standard Asymptotics

A common and generally efficient choice of test statistic is the *likelihood ratio test statistic.* Suppose we wish to compare

$H_0 :\ L_0(\theta \mid y)$ for $\theta \in \Theta$ and

$H_A :\ L_A(\phi \mid y)$ for $\phi \in \Phi$.

Here Θ and Φ correspond to the sets of possible parameter values under the two models.

Definition. *The likelihood ratio test statistic is defined as*

$$T_{\text{LRT}}(y) = -2 \ln \left(\frac{\max_{\theta \in \Theta} L_0(\theta \mid y)}{\max_{\phi \in \Phi} L_A(\phi \mid y)} \right).$$

Example 5. In the Mass Hierarchy problem, the two models correspond to normal and inverted hierarchy. In this case the parameters of the two models overlap, but neither model is a special case of the other.

Suppose the model under H_0 is a special case of the one under H_A. In this case we can consider a common likelihood function $L(\theta \mid y)$ where Θ is the the set of possible parameters and compare

$H_0 :\ \theta \in \Theta_0$ and

$H_A :\ \theta \in \Theta_0^c$.

Here Θ_0 is a subset of Θ.

Definition. *We say two models are nested if one is a special case of the other.*

.
If the models under comparison are nested, the likelihood ratio test statistic is written

$$T_{\text{LRT}}(y) = -2 \ln \left(\frac{\max_{\theta \in \Theta_0} L(\theta \mid y)}{\max_{\theta \in \Theta} L(\theta \mid y)} \right).$$

Example 6. In the Higgs Search, if the size of the bump is treated as a parameter, are the two models nested?

Example 7. In the CP-violation problem, if δ_{CP} is treated as a parameter to be fit, are the two models nested?

NOTES:

1. The likelihood ratio test is generally the go-to test statistic for statisticians.
2. Neyman-Pearson Lemma: If there are no unknown parameters under either H_0 or H_A, the likelihood ratio test is the most powerful test for any bound on the probability of false discovery / detection. *This means that for any bound we choose on the probability of a false discovery / detection, the likelihood ratio test is most likely to detect a true source.*

Example 8. Suppose $y_i \sim \text{POISSON}(\lambda)$ are an independent and identically distributed sample of size n, and we wish to test

$$H_0 : \lambda = 2 \quad \text{versus} \quad H_A : \lambda \neq 2.$$

It is easy to derive that $\hat{\lambda}_{\text{MLE}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and the likelihood ratio test statistic is

$$T_{\text{LRT}}(y) = 2\left(-\sum_{i=1}^{n} y_i + \ln(\bar{y})\sum_{i=1}^{n} y_i + 2n - \ln(2)\sum_{i=1}^{n} y_i\right).$$

To compute $T^\star$ such that $\Pr(T_{\text{LRT}}(y) > T^\star \mid H_0) = \alpha$, we need to know (an approximate?) distribution of T_{LRT} when H_0 is true.

Definition. *The distribution of a test statistic when H_0 is true is called its null distribuiton.*

Theorem. *Wilk's Theorem: Suppose $y = (y_1, \ldots, y_n)$ is an independent and identically distributed sample with likelihood function $L(\theta \mid y)$, then under certain regularity conditions $T_{\text{LRT}}(y)$ converges to a chi-squared distribution as the sample size n goes to infinity. The degrees of freedom of this chi-square distribution is the difference between the number of free parameters under H_0 and under H_A.*

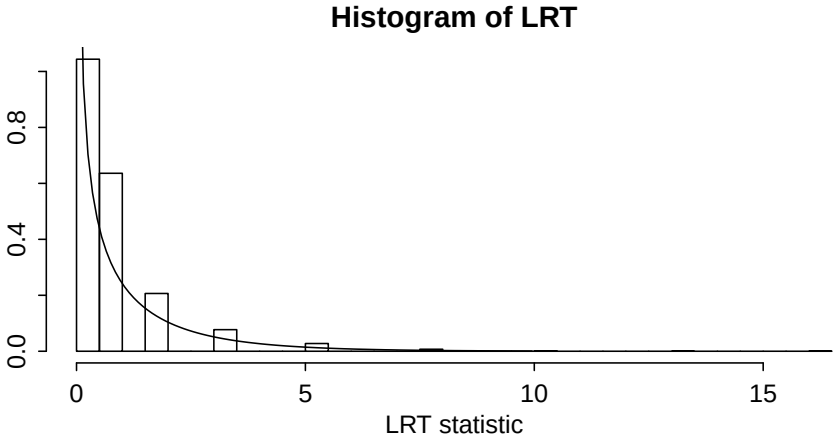
SIMULATION: Returning to Example 8,

1. Obtain a sample from the null distribution of the LRT statistics for $n = 1$.
   For $\ell = 1, \ldots, L$, here with $L = 10,000$.

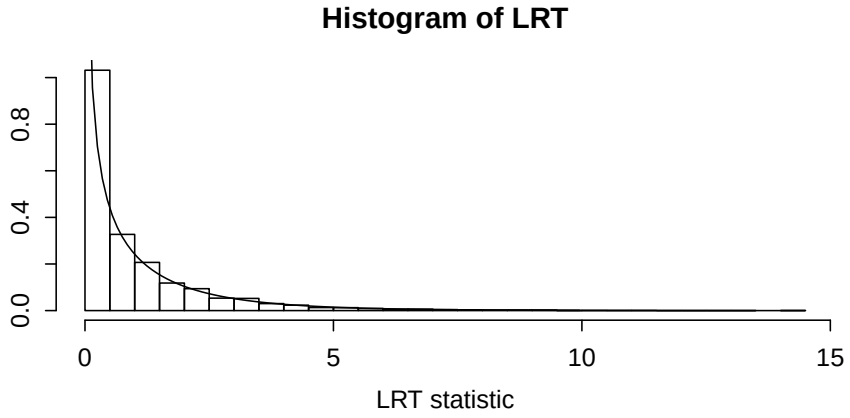
   (a) Sample a replicate data set,
   $$\tilde{y}^{(\ell)} \sim p(\tilde{y} \mid \theta, H_0).$$

   (b) Compute the LRT statistic for data set $\tilde{y}^{(\ell)}$.

   What is the distribution of the LRT statistics sampled from the null distribution?



2. Obtain a sample from the null distribution of the LRT statistics for $n = 1000$.



Regularity Conditions

Wilk's theorem requieres a number of regularity conditions that do not always hold in examples in physics.

1. The models must be nested.
2. None of the parameter values specified by H_0 may be on the boundary of their possible values under H_A. (Mathematically we sea Θ_0 must be in the interior of Θ.)
3. All of the parameters must have values specified under H_0.
4. The asymptotic null distribution of the MLE must be Gaussian.

Example 9. The Higgs search. Consider comparing the null hypothesis that $y_i \sim \text{POISSON}(\beta_1)$ for $i = 1, \ldots, n$, with the alternative hypothesis that

$$H_A : y_i \sim \text{POISSON}(\beta_i + \mu \mathcal{I}\{i = 20\}), \quad \text{for} \quad i = 1, \ldots, n,$$

where β_i is the expected background count in bin i and μ is the size of the bump. (See plot.) For simplicity we assume that

1. each β_i is known,
2. the bump is contained completely in one bin,
3. and we know which bin it is that contains the bump.

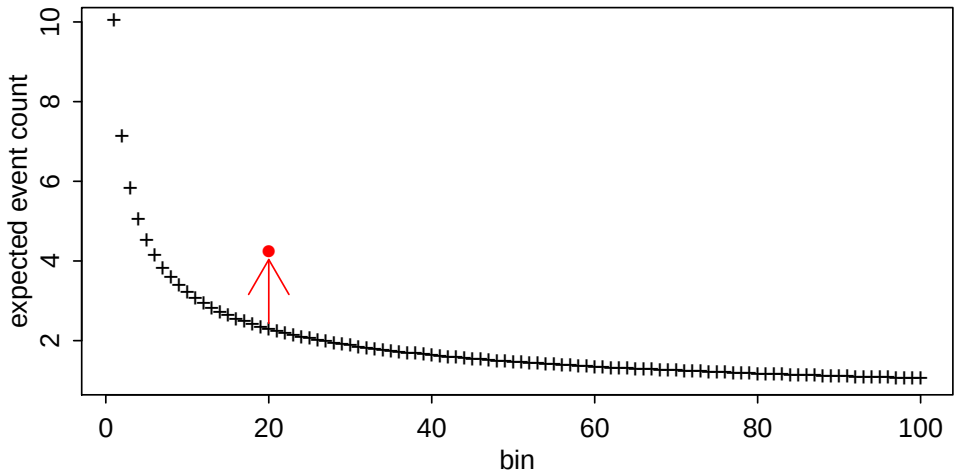


Illustration of the Higgs model in Example 9.

The range of values for μ under H_A is $\mu > 0$, whereas under the null hypothesis $\mu = 0$. Because the null value of μ in on the boundary, the standard asymptotic for the LRT do not hold.

The MLE $\hat{\mu}_{\text{MLE}}$ cannot be asymptotically normal with mean $\mu = 0$ under H_0. This can easily be seen because

$$\Pr(\hat{\mu}_{\text{MLE}} \geq 0 \mid H_0) = 1.$$

Simply stated, $\hat{\mu}_{\text{MLE}}$ cannot be negative. The standard asymptotic distribution of the LRT statistics depends on the normality of the MLE under H_0. So the standard asymptotic distribution of the LRT does not apply. In this case, the asymptotic distribution of the LRT statistic is a mixture of χ^2 random variables (see Section 2.4).

Now consider comparing the same null hypothesis with the more general alternative hypothesis that

$$H_A : y_i \sim \text{Poisson}(\beta_i + \mu \mathcal{I}\{i = m\}), \quad \text{for} \quad i = 1, \ldots, n,$$

Here we do not assume we know the location of the bump and fit m the index of the bin that contains the bump. The situation here is more complicated. Under H_0, $\mu = 0$, but m is undefined; *m has no value under H_0*. In this case the LRT statistic has no known distribution.

Do the regularity conditions hold?

1. Models must be nested:

2. Parameter values specified under H_0 may not be on the boundary of their parameter space.

3. All parameters must be defined under H_0.

4. The asymptotic null distrubution of the MLE must be Gaussian.

Example 10. Mass hierarchy. Do the regularity conditions hold?

Example 11. CP-violation. Do the regularity conditions hold?

We return to methods that can be used when the regularity conditions fail in Section 2.4.

2.4 When Standard Asymptotic Methods Fail

When Wilk's Theorem fails, we require other methods to compute / approximate

$$\text{p-value} = \Pr\Big(T(y) \geq T(y_{\text{obs}}) \mid H_0\Big).$$

Specifically this requires other methods to approximate the null distribuiton of $T(y)$.

2.4.1 Numerical Methods

The easiest method is to use numerically simulate from the null distribution of $T(y)$. (Physicists sometimes call this generating "toys".)

Procedure:

1. Acquire a large sample of replicate data sets, of size L, under H_0:

$$\tilde{y}^{(\ell)} \sim p(\tilde{y} \mid \theta, H_0) \quad \text{for} \quad \ell = 1, \ldots, L.$$

2. Estimate the p-value by Monte Carlo:

$$\text{p-value} \approx \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{I}\left\{T(\tilde{y}^{(\ell)}) > T(y_{\text{obs}})\right\}.$$

   Here $T(y_{\text{obs}})$ is the value of the test statistic for the actual observed data.

NOTES:

1. This procedure does not specify how nuisance parameters should be handled. How should replicated data sets be sampled in Step 1 if the null model involves unknown parameters.
2. The standard strategy is to either fit the nuisance parameters or resample them accounting for uncertainty in their fit.
3. We discuss a Bayesian strategy for handling nuisance parameters.

Posterior Predictive P-values:
In principle, if there are unknown parameters under H_0, i.e., nuisance parameters, the p-value depends on these parameters:

$$\text{p-value}(\theta) = \Pr(T(\tilde{y}) > T(y_{\text{obs}}) \mid \theta, H_0).$$

Here $T(\tilde{y})$ is random; its distribution is generated by (1) generating replicate data sets under the model

$$\tilde{y} \sim p(\tilde{y} \mid \theta, H_0),$$

and (2) computing the test statistic for each replicate data set. Again, $T(y_{\text{obs}})$ is the test statistic evaluated with the observed data.

QUESTION: How do we compute p-value(θ) if θ is unknown?

Definition. *A posterior predictive p-value (or ppp-value) is expected value of the p-value under the (Bayesian) posterior distribution:*

$$\text{ppp-value} \;=\; \int \Pr(T(\tilde{y}) > T(y_{\text{obs}}) \mid \theta, H_0)\; p(\theta \mid y, H_0)\; d\theta \;=\; \Pr(T(\tilde{y}) > T(y_{\text{obs}}) \mid y, H_0).$$

Procedure for Compute ppp-values:

1. Acquire a large sample of size L from the posterior distribution:

$$\theta^{(\ell)} \sim p(\theta \mid y) \quad \text{for} \quad \ell = 1, \ldots, L.$$

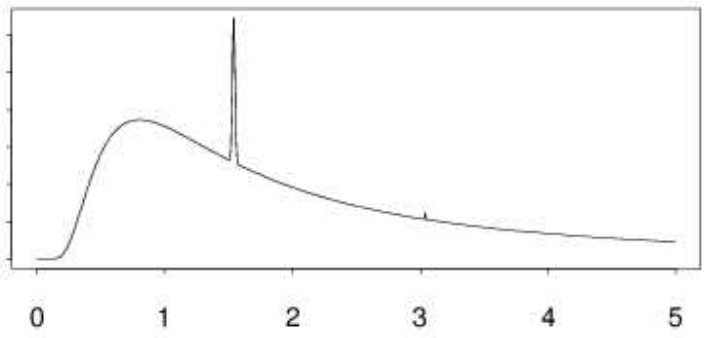
2. Use the posterior sample to generate a sample of replicated data sets from

$$\tilde{y}^{(\ell)} \sim p(\tilde{y} \mid \theta^{(\ell)}) \quad \text{for} \quad \ell = 1, \ldots, L.$$

3. Estimate the ppp-value by Monte Carlo:

$$\text{ppp-value} \approx \frac{1}{L} \sum_{\ell=1}^{L} I\left\{T(\tilde{y}^{(\ell)}) > T(y_{\text{obs}})\right\}.$$

Example 12. The distribution of energy originating from an astronomical source is called its *spectrum* and can be informative as to the physical processes at the source. A typical X-ray source spectrum might consist of an extended smooth component known as a continuum and one or more narrow features knows as spectral lines:



E = energy (keV)

Space-based instruments study X-ray spectra of astronomical sources via the count of photons recorded in each of a large number or narrow energy bins. Suppose we wish to compare several models for a stellar spectrum, each of which includes a simple power law continuum, $\alpha E^{-\beta}$. The models may or may not include one spectral line and if there is a line its location may or may not be known.

Model 0: $Y_i \overset{\text{iid}}{\sim} \text{POISSON}\left(\alpha E_i^{-\beta}\right)$.

Model 1: $Y_i \overset{\text{iid}}{\sim} \text{POISSON}\left(\alpha E_i^{-\beta} + \mu I_{\{i=m\}}\right)$, with μ known.

Model 2: $Y_i \overset{\text{iid}}{\sim} \text{POISSON}\left(\alpha E_i^{-\beta} + \mu I_{\{i=m\}}\right)$.

In Model 0 there is no spectral line. In Models 1 and 2 there is one spectral line, its location, μ, is fixed and known in Model 1 and fitted in Model 2. We treat Model 0 as the null hypothesis.[1]

NOTE:

1. If the spectral line's location is known, all parameters have values under Model 0 but the parameter space of Model 0 is on the boundary of the parameter space of Model 1.

2. If the spectral line's location is unknown, *not all parameters have values under Model 0.* The parameter space of Model 0 is on the boundary of the parameter space of Model 1.

3. The standard asymptotic distribution of the LRT does not apply for testing Model 0 against Model 1 or for testing Model 0 against Model 2. (See Example 9.)

[1] This is a somewhat simplified version of Models 2 and 3 compared to what was used in the simulation study. See Protossov, et al., 2002, ApJ for details.

We also consider one more model that includes an *absorption feature*. This means that there is a narrow range of energy with *fewer* expected photons then under Model 0. (This is the opposite of an emission line, where there are excess photons in a narrow range of energy.)

Model 3: $Y_i \overset{\text{iid}}{\sim} \text{POISSON}\Big(\alpha E_i^{-\beta} - \mu I_{\{i=m\}}\Big)$

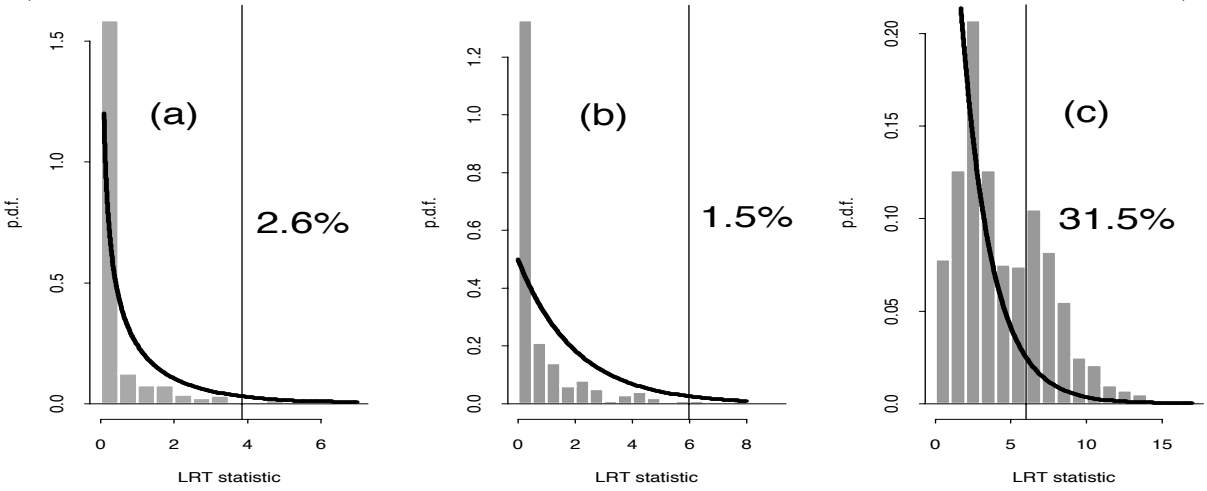
We use ppp-values with the LRT statistics to

(a) compare Model 0 with Model 1
(b) compare Model 0 with Model 2
(c) compare Model 0 with Model 3

The null distributions of each of the three LRT statistic appear below. The black curves are the standard asymptotic χ^2 distributions under Wilk's Theorem; the vertical lines represent nominal 5% cut offs; and the given percentages are the actual false positive rates corresponding to the 5% cut off points. The results show that the null distribution derived under Wilk's Theorem may be conservative or anti-conservative.

Asymptotic theory may be conservative or anti-conservative.

(Vertical lines mark nominal 5% cutoffs; percents are actual false positive rates when using nominal cutoffs.)



2.4.2 Non-standard Asymptotics

Simulating toys may be infeasible when a 4σ or 5σ detection criterion is required. Here we list some of methods that can be used.

Non-nested Models. When there are *no unknown parameters*, the central limit theorem (CLT) can be used to approximate the null distribution of the LRT statistic. Suppose we wish to compare

H_0**:** $y = (y_1, \ldots, y_n)$ are independently distributed according to the probability density $f_0(y)$.

H_A**:** $y = (y_1, \ldots, y_n)$ are independently distributed according to the probability density $f_1(y)$.

Because there are no unknown parametes, we can write the LRT statistic as

$$T_{\text{LRT}}(y) = -2 \ln \left(\frac{\prod_{i=1}^n f_0(y_i)}{\prod_{i=1}^n f_1(y_i)} \right) = -2 \sum_{i=1}^n \ln \big(f_0(y_i) - f_1(y_i) \big).$$

Because this is the sum of an independent random sample, the CLT implies that for large n,

$$T_{\text{LRT}}(y) \overset{\text{approx}}{\sim} \text{N}\Big(\text{mean} = -2n\text{E}\big(f_0(y_i) - f_1(y_i) \mid H_0\big), \text{variance} = 4n\text{Var}\big(f_0(y_i) - f_1(y_i) \mid H_0\big)\Big)$$

These two quantities, $\text{E}\big(f_0(y_i) - f_1(y_i) \mid H_0\big)$, and $\text{Var}\big(f_0(y_i) - f_1(y_i) \mid H_0\big)$, can easily be approximated with a moderate sample under H_0. See Cousins et al. (2005, J. High Energy Phys., 11, 046) for details

When H_0 is on the boundary of the set of possible parameters. A generalization of Wilk's theorem applies when H_0 is on the boundary of the set of possible parameters. This generalization is known as Chernoff's Theorem.

Example 13. Suppose we model $y \sim N(\mu, 1)$, where $\mu \geq 0$ and we wish to test

H_0**:** $\mu = 0$

H_A**:** $\mu > 0$.

Because the null distribution of the MLE of μ cannot be centered on its true value, 0 under H_0, Wilk's theorem fails. Instead, half the time the MLE of μ is zero (when $y \leq 0$) , and the LRT statistic is equal to zero. The other half the time the standard theory holds. Thus the LRT statistic is equal to 0 half the time and is a χ^2_1 variable the other half of the time. This has the effect of halving the p-values relative to thoes computed under Wilk's Theorem.

When there is a parameter that has no value under H_0. This case is summarized in a separate set of slides called "A Case Study: Bump Hunting".

3 Bayesian Model Selection

3.1 Bayes Factors and Posterior Probabilities

For Bayesian model selection, in principle, we can compute the posterior distribution of each model under consideration. Let π_0 be the prior probability of the null model. Using Bayes Theorem,

$$\Pr(H_0 \mid y) = \frac{p(y \mid H_0)\pi_0}{p(y \mid H_0)\pi_0 + p(y \mid H_A)(1-\pi_0)}.$$

The marginal distribution of the data under each model is given by the respective prior predictive distributions,

$$p(y \mid H_0) = \int p(y \mid \theta, H_0)p(\theta \mid H_0)d\theta \quad \text{and} \quad p(y \mid H_A) = \int p(y \mid \theta, H_A)p(\theta \mid H_A)d\theta$$

The posterior odds of H_0 is

$$\frac{p(H_0 \mid y)}{p(H_A \mid y)} = \frac{p(y \mid H_0)\pi_0}{p(y \mid H_A)\pi_A} = \frac{\pi_0}{1-\pi_0} \times \frac{p(y \mid H_0)}{p(y \mid H_A)},$$

where $\pi_0/(1-\pi_0)$ is the prior odds of H_0 and the

$$\text{Bayes Factor} = \frac{p(y \mid H_0)}{p(y \mid H_A)}.$$

NOTES:

1. Bayes Factors are often used in place of posterior odds to avoid specifying the prior odds.
2. Like the LRT statistic, the Bayes Factor is a *relative probability* and avoids the oddities of tail probabilities.
3. Unlike with the LRT the models under comparison need not be nested.
4. Unlike hypothesis testing, there is no inherent asymmetry between the models under consideration. We may decide stronger evidence is required for us to chose the "alternative model" than for us to stick with the "null model", but no such distinction between the two models is required. For this reason we typically do not us the "null" and "alternative" terminology.
5. I typically use the log Bayes Factor so that the two models are treated more symmetrically.
6. Jeffreys (1961, "The Theory of Probability") proposed a scale to qualify the degree of evidence for one of the models that is supported by a given value of a Bayes Factor:

Bayes Factor	$\ln_{10}$(Bayes Factor)	Degree of Evidence
> 100	> 2	Overwhelming evidence for H_0
30 to 100	1.5 to 2	Very Strong evidence for H_0
10 to 30	1 to 1.5	Strong evidence for H_0
3 to 10	0.5 to 1	Substantial evidence for H_0
$\frac{1}{3}$ to 3	-0.5 to 0.5	Barely worth mentioning
$\frac{1}{10}$ to $\frac{1}{3}$	-0.5 to -1	Substantial evidence for H_A
$\frac{1}{30}$ to $\frac{1}{10}$	-1 to -1.5	Strong evidence for H_A
$\frac{1}{100}$ to $\frac{1}{30}$	-1.5 to -2	Very Strong evidence for H_A
$< \frac{1}{100}$	< -2	Overwhelming evidence for H_A

7. How do we interpret the posterior odds or the Bayes Factor if <u>neither</u> model holds??

3.2 The Problem with P-values

Although the use of p-values in model selection is endemic, they cam be quite misleading.

Example 14. *(Jeffrey-Lindley Paradox)* Suppose we wish to repeatedly compare a null hypothesis with an alternative hypothesis. We happen to know that H_0 is actually true with probability π_0 and the alternative is true with probability $1-\pi_0$. To be precise, suppose we wish to compare $H_0 : Y \sim \mathrm{N}(0,1)$ with the $H_A : Y \sim \mathrm{N}(3.4,1)$ and compute a p-value based on the test statistic $T = Y$.

Suppose we observe Y and compute a p-value and find it to be significant with $0.05 >$ p-value > 0.04. We can reject H_0 in favor of the alternative.

Question: Given that $0.05 >$ p-value > 0.04, how often is H_0 true?

Because we know H_0 is true with probability π_0, we can compute

$$\Pr(H_0 \mid 0.05 > \text{ p-value } > 0.04) = \qquad (\star)$$

Compare this with the p-value of 0.05!!

Question: What does this mean? What have we learned from the p-value?

Note: By adjusting the mean under the alternative we can make $(\star)$ either much less than or much greater than π_0. *How can we interpret the p-value for model selection??*

3.3 The Problem with Prior Distributions

Example 15. Suppose $y \mid \mu \sim \mathrm{N}(\mu, 1)$ with prior distribution $\mu \sim \mathrm{N}(0, \tau^2)$. It is easy to derive the marginal distribution of y (often called the prior predictive distribution):

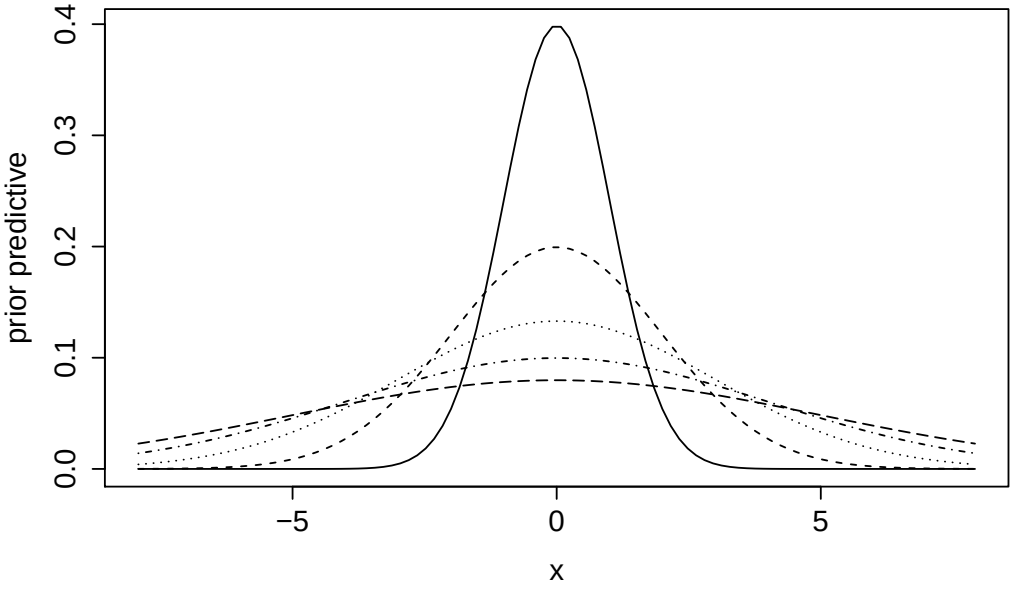
$$y \sim \mathrm{N}(0, 1+\tau^2).$$

To see this, note that the joint distribution of y and μ is

$$p(y,\mu) \propto \exp\left\{-\frac{1}{2}\Big((y-\mu)^2 + \mu^2/\tau^2\Big)\right\},$$

so (y,μ) are bivariate normal and y is marginally normal with $\mathrm{E}(y) = \mathrm{E}\big[\mathrm{E}(y \mid \mu)\big] = \mathrm{E}(\mu) = 0$ and $\mathrm{Var}(y) = \mathrm{E}\big[\mathrm{Var}(y \mid \mu)\big] + \mathrm{Var}\big[\mathrm{E}(y \mid \mu)\big] = \mathrm{E}(1) + \mathrm{Var}(\mu) = 1 + \tau^2$.

The marginal distribution of y is plotted for several choices of τ^2 below:



NOTE: The value of $p(y)$ depends on τ^2. Because Bayes Factors and posterior probabilities of H_A use prior predictive distributions, they must be used with great care. The choice of prior distribution must be carefully considered and *must always be reported.*

Now suppose we wish to compare

$$H_0 : y \sim \mathrm{N}(0, 1) \quad \text{with} \quad H_A : y \mid \mu \sim \mathrm{N}(\mu, 1),$$

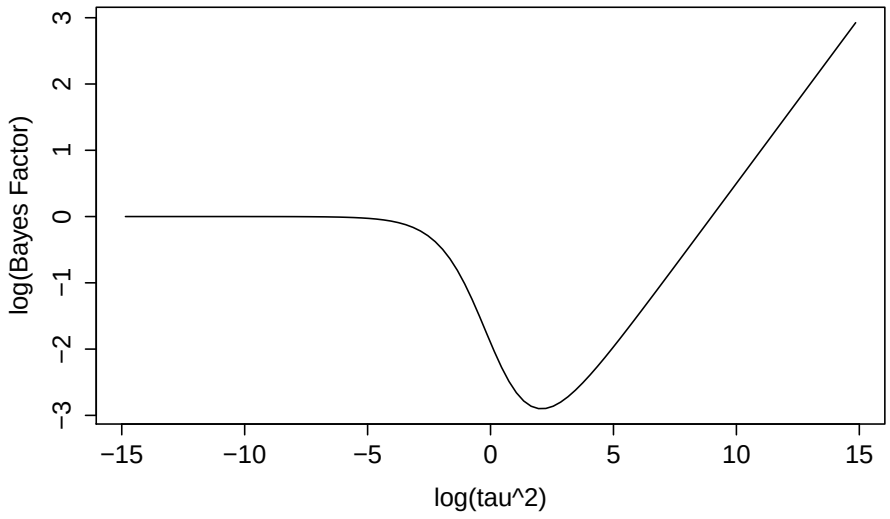
where $\mu \sim \mathrm{N}(0, \tau^2)$. The marginal distributions of y under the two models are

$$H_0 : y \sim \mathrm{N}(0, 1) \quad \text{and} \quad H_A : y \sim \mathrm{N}(\mu, 1 + \tau^2),$$

and

$$\text{Bayes Factor} = \frac{\dfrac{1}{\sqrt{2\pi}} \exp\left\{-\dfrac{y^2}{2}\right\}}{\dfrac{1}{\sqrt{2\pi(1+\tau^2)}} \exp\left\{-\dfrac{y^2}{2(1+\tau^2)}\right\}} = \sqrt{1+\tau^2} \exp\left\{-\frac{y^2\tau^2}{2(1+\tau^2)}\right\}$$

If we observe $y = 3$, the Bayes Factor can vary dramatically depending on our choice of prior,



Note. *Unfortunately, Bayes Factors depend heavily on the choice of prior distribution.*

3.4 A Case Study: Bump Hunting

The Case Study is presented in a separate set of slides.