Outline

• Lecture 1
  • Introduction
  • The Frequentist Principle
  • Confidence Intervals
  • The Profile Likelihood
• Lecture 2
  • Hypothesis Tests
  • Bayesian Inference
• Lecture 3
  • Introduction to Machine Learning
Hypothesis Tests – 1

The Basic Idea:

1. Decide which hypothesis is to be rejected and call it the null hypothesis. At the LHC, this is usually the background-only hypothesis.

2. Construct a function of the data called a test statistic, such that large values of it would cast doubt on the veracity of the null hypothesis.

3. Choose a test statistic threshold above which you are inclined to reject the null. Do the experiment, compute the statistic, and reject the null if the threshold is breached.
Hypothesis Tests – 2

There are two closely related approaches:

1. **Fisher**: reject the null if the test statistic is large enough.

2. **Neyman**: compare the null to an alternative hypothesis using a statistic that depends on both hypotheses. Reject the null if the alternative is preferred.
Fisher’s Approach: \textit{Null} hypothesis ($H_0$), e.g., background-only

The null hypothesis is \textit{rejected} if the p-value is judged to be small enough.

\[ p(x|H_0) \]

\[ p\text{–value} = P(x \geq x_0|H_0) \]

$x_0$ is the observed value of the test statistic.
Example: **Hypothesis Test (No Unc.)**

Background, $B = 9.4$ events (ignoring uncertainty)

\[ p(N|H_0) = \text{Poisson}(N, B = 9.4) \]

$N = 25$ observed count

\[
p\text{-value} = \sum_{k=N}^{\infty} \text{Poisson}(k, 9.4) = 1.76 \times 10^{-5}
\]
Example: Hypothesis Test (No Unc.)

Background, $B = 9.4$ events (ignoring uncertainty)

$p(N|H_0) = \text{Poisson}(N, B = 9.4)$

$p\text{-value} = 1.76 \times 10^{-5}$

$N = 25$ observed count

Since a small p-value is a bit non-intuitive, we usually map it to a Z-value, that is, the number of standard deviations away from the null if the distribution were a Gaussian.

This yields $Z = 4.14$ (see notebook 11_16_hzz4l.ipynb).
**Hypothesis Tests – 4**

**Neyman’s Approach**: *Null* hypothesis ($H_0$) + alternative ($H_1$)

Neyman argued that it is *necessary* to consider alternative hypotheses.

$$\alpha = p-value(x_\alpha)$$  Choose a *fixed* value of $\alpha$ *before* data are analyzed, which depends on *both* hypotheses.

Reject the null in favor of the alternative if the p-value $< \alpha$.  

$\alpha$ is called the *significance* (or size) of the test.
In Neyman’s approach, hypothesis tests are a contest between significance and power, i.e., the probability to accept a true alternative.

\[
\alpha = \int_{x_\alpha}^{\infty} p(x \mid H_0) \, dx
\]

significance of test

\[
p = \int_{x_\alpha}^{\infty} p(x \mid H_1) \, dx
\]

power of test
The Neyman-Pearson Test

The optimal test for fully specified hypotheses, the so-called, simple hypotheses, is to reject the null if the ratio $p(x|H_1)/p(x|H_0) > \lambda$ for some threshold $\lambda$.

$p(x|H_0)$

$p(x|H_1)$

$x$ $x_\alpha$

$\alpha = \int_{x_\alpha}^{\infty} p(x|H_0) \, dx$

significance of test

$p = \int_{x_\alpha}^{\infty} p(x|H_1) \, dx$

power of test
The Neyman-Pearson Test

Power curve
power vs. significance.
Note: in general, *no analysis* is uniformly the most powerful.

Blue is the more powerful below the cross-over point and green is the more powerful above.

\[ \alpha = \int_{x_\alpha}^{\infty} p(x|H_0) \, dx \]

significance of test

\[ p = \int_{x_\alpha}^{\infty} p(x|H_1) \, dx \]

power of test
Hypothesis Tests – 5

Any non-trivial analysis contains nuisance parameters. We need to get rid of them in order to perform an hypothesis test.

There two primary ways:

**Method 1:** Use the profile likelihood.

**Method 2:** Use a likelihood integrated over the nuisance parameters.
Example: Hypothesis Test (1)

Method 1:
In our example, the likelihood is a 2-parameter function
\( L(s, b) \equiv p(D \mid s, b) \) in which we replace \( b \) by \( \hat{b}(s) \) to get the profile likelihood \( L_p(s) = L(s, \hat{b}(s)) \).

Recall that the quantity
\[
t(s) = -2 \ln[L_p(s)/L_p(\hat{s})]
\]
can be used to compute approximate confidence intervals.

It can also be used to test hypotheses.
Example: Hypothesis Test (1)

We’ll use $t(s)$ to test the null hypothesis, $s = s_0 = 0$.

Wilks’ theorem, applied to the Higgs boson example, states that for large samples the density of the signal estimate $\hat{s}$ will be approximately Gaussian.

Moreover, if $s_0$ is also the true value, then the distribution of $t(s_0)$ will be approximately a $\chi^2$ density of one degree of freedom.

This implies that the density of $t(s_0)$ is independent of all the parameters of the problem!
Since we know the form of the probability density of $t(s_0)$, we can calculate the p-value:

$$p\text{-value} = P[t(s_0) \geq t_{obs}(s_0)]$$

given the observed value $t_{obs}(s_0)$, of $t(s_0)$. Then, if the p-value < $\alpha$, the agreed upon significance of our test, we reject the $s_0$ hypothesis. In addition, we should report the p-value.

But, since $Z = \sqrt{t_{obs}(s_0)}$, we can sidestep the calculation of the p-value and just report $Z$. 

**Example: Hypothesis Test (1)**
Example: Hypothesis Test (1)

Background, $B = 9.4 \pm 0.5$ events. For this example, $t_{\text{obs}}(0) = 17.05$

therefore, $Z = \sqrt{t_{\text{obs}}(0)} = 4.13$

$L_p(s) = L(s, \hat{b}(s))$

$t(s) = -2 \ln[L_p(s)/L_p(\hat{s})]$

$\hat{b}(s) = \frac{g + \sqrt{g^2 + 4(1+k)Ms}}{2(1+k)}$

$g = N + M - (1 + k)s$

Exercise 12: Verify this calculation
Example: Hypothesis Test (2)

Method 2: We eliminate \( b \) from the problem through integration*:

\[
p(D \mid s) = \int_{0}^{\infty} \text{Poisson}(N \mid s + b) \, \text{Poisson}(M \mid kb) \, d(kb)
\]

\[
= \frac{(1 - x)^2}{M} \sum_{r=0}^{N} \text{beta}(x, r + 1, M) \, \text{Poisson}(N - r, s)
\]

where, \( x = \frac{1}{1+k} \), and \( \text{beta}(x, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, x^{a-1} (1 - x)^{b-1} \)

Exercise 13: Show this

\( p(D \mid s) \) is called a **marginal (or integrated) likelihood**.

**Example: Hypothesis Test (2)**

Note that $p(D \mid s) = p(N, M, k \mid s)$ satisfies the sum rule

$$\sum_{n=0}^{\infty} p(n, M, k \mid s) = 1$$

Notice also that the likelihood for the null hypothesis, $s = 0$, is

$$p(N \mid H_0) \equiv p(D \mid s = 0) = \frac{(1 - x)^2}{M} \sum_{r=0}^{N} \text{beta}(x, r + 1)$$

while the likelihood for the alternative hypothesis, $s \neq 0$, is

$$p(N \mid H_1) \equiv p(D \mid s \neq 0)$$
Example: Hypothesis Test (2)

Background, $B = 9.4 \pm 0.5$ events

$$p(N|H_0) \equiv p(D | s = 0)$$

$$p\text{-value} = \sum_{k=N}^{\infty} p(k|H_0) = 1 - \sum_{k=0}^{N-1} p(k|H_0) = 2.53 \times 10^{-5}$$

This is equivalent to $4.05 \sigma$ effect, to be compared with the $4.14 \sigma$ obtained earlier.

Exercise 14: Verify this calculation
An Aside on $s / \sqrt{b}$

The quantity $s / \sqrt{b}$ is often used as a rough measure of significance on the “$n$-$\sigma$” scale. But, it should be used with caution.

In our example, $s \sim 25 - 9.4 = 15.6$ events.

So according to this measure, the CMS result is a $15.6/\sqrt{9.4} \sim 5.1\sigma$ effect, which is to be compared with $4.1\sigma$!

So, beware of $s / \sqrt{b}$!
Outline

• Lecture 1
  • Introduction
  • The Frequentist Principle
  • Confidence Intervals
  • The Profile Likelihood

• Lecture 2
  • Hypothesis Tests
  • Bayesian Inference

• Lecture 3
  • Introduction to Machine Learning
Bayesian Inference – 1

**Definition:**

A method is Bayesian if

1. it is based on the *degree of belief* interpretation of probability and if
2. it uses Bayes’ theorem

\[
p(\theta, \nu | D) = \frac{p(D | \theta, \nu) \pi(\theta, \nu)}{p(D)}
\]

for *all* inferences.

- \(D\) observed data
- \(\theta\) parameter of interest
- \(\nu\) nuisance parameters
- \(\pi\) *prior density*
Nuisance parameters are removed by marginalization:

\[
p(\theta | D) = \int p(\theta, \nu | D) d\nu = \frac{\int p(D | \theta, \nu) \pi(\theta, \nu) d\nu}{p(D)}
\]

in contrast to profiling, which, however, can be regarded as marginalization with respect to a specific prior, namely, \( \pi(\theta, \nu) = \delta(\nu - \hat{\nu}(\theta)) \).

Many frequentist procedures can be cast as Bayesian procedures with suitable priors.
Bayesian Inference – 3

Bayes’ theorem can be used to compute the probability of a model or hypothesis $H$.
To do so, first compute the full posterior density:

$$
p(\theta_H, \nu, H | D) = \frac{p(D | \theta_H, \nu, H) \pi(\theta, \nu, H)}{p(D)}
$$

- $D$ observed data
- $H$ model or hypothesis
- $\theta_H$ parameters of model $H$
- $\nu$ nuisance parameters
- $\pi$ prior density
Bayesian Inference – 4

Then,

1. factorize the priors: \(\pi(\theta_H, \nu, H) = \pi(\theta_H, \nu | H) \pi(H)\)

2. and, for each model, \(H\), compute the function

\[
p(D | H) = \int \int p(D | \theta_H, \nu, H)\pi(\theta_H, \nu | H) d\theta_H d\nu
\]

3. finally, compute the probability of each model, \(H\)

\[
p(H | D) = \frac{p(D | H)\pi(H)}{\sum_H p(D | H)\pi(H)}
\]
In order to compute $p(H | D)$, however, two things are needed:

1. Priors that integrate to one over the parameter spaces

$$\int \int \pi(\theta_H, \nu | H) d\theta_H d\nu = 1$$

2. The priors $\pi(H)$.

In practice, we compute the Bayes factor:

$$\frac{p(H_1 | D)}{p(H_0 | D)} = \frac{\frac{p(D | H_1)}{\pi(H_1)}}{\frac{p(D | H_0)}{\pi(H_0)}}$$

which is the ratio in the first bracket, denoted by $B_{10}$. 
Example: Bayesian Analysis \( H \to 4l \)

**Step 1:** Construct a probability model for the observations

\[
p(D|s, b) = \frac{(s+b)^N e^{-(s+b)}}{N!} \cdot \frac{(kb)^M e^{-kb}}{\Gamma(M+1)}
\]

**Knowns:**
- \( N = 25 \) observed event count
- \( M = 353.4 \) effective background event count
- \( k = 37.6 \) effective background scale factor

**Unknowns:**
- \( b \) expected background count
- \( s \) expected signal count
- \( d = s + b \) expected event count
Example: Bayesian Analysis $H \rightarrow 4l$

Step 2: Write down Bayes’ theorem:

$$p(s, b|D) = \frac{p(D | s, b) \pi(s, b)}{p(D)}$$

and specify the prior:

$$\pi(s, b) = \pi(b|s) \pi(s)$$

It is often convenient first to compute the marginal likelihood (as we did earlier) by integrating over the nuisance parameter, $b$.

$$p(D|s) = \int_0^\infty p(D | s, b) \pi(b|s) db$$
Example: Bayesian Analysis $H \rightarrow 4l$

The Prior:

What does

$$\pi(s, b) = \pi(b|s) \pi(s)$$

represent?

The prior encodes what we know, or assume, about the mean background and signal in the absence of new observations. We shall assume that $s$ and $b$ are non-negative.

Unfortunately, there is no unique way to encode such vague information.
Example: Bayesian Analysis $H \rightarrow 4l$

For simplicity, we shall take $\pi(b \mid s) = 1$, though one can do better*.

We have already calculated the integral and found

$$p(D \mid s) = \frac{(1 - x)^2}{M} \sum_{r=0}^{N} \text{beta}(x, r + 1, M) \text{ Poisson}(N - r, s)$$

where, $x = \frac{1}{1+k}$.

Example: Bayesian Analysis $H \rightarrow 4l$

$L(s) = P(25 \mid s)$ is the marginal likelihood for the expected signal $s$.

Here, we compare the marginal and profile likelihoods. For this problem they are found to be almost identical.

But, this does not always happen!
Example: Bayesian Analysis $H \rightarrow 4l$

Given the marginal likelihood $p(D \mid s)$ we can compute the posterior density

$$p(s \mid D) = \frac{p(D \mid s)p(s)}{p(D)}$$

Again, for simplicity, assume $\pi(s) = 1$, then

$$p(s \mid D) = \frac{\sum_{r=0}^{N} \beta(x, r + 1, M) \text{Poisson}(N - r, s)}{\sum_{r=0}^{N} \beta(x, r + 1, M)}$$

**Exercise 15**: Derive an expression for $p(s \mid D)$ assuming a gamma prior $\Gamma(qs, U + 1)$ for $\pi(s)$.
Example: **Bayesian Analysis** $H \rightarrow 4l$

Computing Credible (or Bayesian Confidence) Intervals

By solving,

$$
\int_{0}^{l(N)} p(s \mid D) \, ds = (1 - CL)/2
$$

$$
\int_{0}^{u(N)} p(s \mid D) \, ds = (1 + CL)/2
$$

with $CL = 0.683$, we obtain $s \in [11.5, 21.7]$ at 68% CL.

Since this is a Bayesian calculation, this statement means:

*the probability that $s$ lies in $[11.5, 21.7]$ is 0.68.*
Example: **Bayesian Analysis** $H \rightarrow 4l$
Finally, we can test which of the two hypotheses, \( s = 0 \), or \( s \neq 0 \), is preferred. First calculate

\[
p(D \mid H_1) = \int_0^\infty p(D \mid s, H_1)\pi(s \mid H_1)\,ds
\]

But recall, to do so, we need to specify a *proper* prior for the signal, that is, a prior \( \pi(s \mid H_1) \) that integrates to one.

The simplest such prior is a \( \delta \)-function, e.g.:

\[
\pi(s \mid H_1) = \delta(s - 15.6),
\]

which yields

\[
p(D \mid H_1) \equiv p(D \mid s = 15.6) = 7.91 \times 10^{-2}.
\]

Note also that

\[
p(D \mid H_0) \equiv p(D \mid s = 0) = 1.59 \times 10^{-5}
\]
Example: Bayesian Analysis $H \rightarrow 4l$

From
\[
p(D \mid H_1) = 7.91 \times 10^{-2} \quad \text{and} \quad p(D \mid H_0) = 1.59 \times 10^{-5}
\]

we conclude that the odds in favor of the hypothesis $s = 15.6$ has \textit{increased} by $\sim 5000$ relative to the prior odds.

The increased odds can be converted to a $Z$-value (S. Sekmen, HBP) roughly equivalent to the frequentist measure using
\[
Z = \text{sign}(\ln B_{10}) \sqrt{2[\ln B_{10}]}
\]
This yields $Z = 4.13$.

Exercise 16: Verify this number
Generalization to Multiple Bins

The generalization to so-called “shape” analyses, that is, to multiple bins introduces no new concepts.

Here is a model for $M$ independent bins, each with $N$ sources:

1. Mean count in $i^{th}$ bin: $d_i = \sum_{j=1}^{N} p_j a_{ji}$, where each bin contains $N$ sources with mean counts $a_{ji}$. The $p_j$ are parameters such as the signal strength $\mu$.

2. Likelihood for $i^{th}$ bin: $p(D_i|d_i) = \text{Poisson}(D_i, d_i)$.

3. Likelihood for $i^{th}$ bin of $j^{th}$ source:
   $$p(A_{ji}|r_{ji} a_{ji}) = \text{Poisson}(A_{ji}, r_{ji} a_{ji})$$
   where $r_{ji}$ are known scale factors.
Generalization to Multiple Bins

The overall probability model is

\[
p(D|a) = \prod_{i=1}^{M} p(D_i|d_i) \prod_{j=1}^{N} p(A_{ji}|r_{ji}a_{ji})
\]

which can be marginalized with respect to \(a_{ji}\) exactly*:

\[
p(D|p_j, r_{ji}) = \prod_{i=1}^{M} \sum_{k_1,...,k_N=0}^{D_i} \prod_{j=1}^{N} \left(A_{ji}+k_j\right) p_j^{k_j} r_{ji}^{A_{ji}+k_j} (p_j + r_{ji})^{-k_j}
\]

with \(k_1 + \cdots + k_N = D_i\).

If the scale factors \(r_{ji}\) are not known precisely, the above can be extended to incorporate the appropriate uncertainties.

*FSU undergraduate Robert Orlando.
Foundation of Statistics: Probability

Two main interpretations:
1. Degree of belief
2. Relative frequency

Likelihood Function
Main ingredient in any full scale statistical analysis.

Frequentist Principle
Construct statements such that a fraction \( f \geq C.L. \) of them will be true over an infinite ensemble of statements.
Summary – 2

Frequentist Approach

1. Use likelihood function only.
2. Eliminate nuisance parameters by profiling.
3. Decide on a fixed threshold $\alpha$ for rejection and reject null if p-value < $\alpha$, but do so only if rejecting the null makes scientific sense, e.g.: the probability of the alternative is judged to be high enough.

Bayesian Approach

1. Model *all* uncertainty using probabilities and use Bayes’ theorem to make *all* inferences.
2. Eliminate nuisance parameters through marginalization.