# HYPOTHESIS TESTING

- Goals: → use data to verify or disprove a theory or hypothesis.
  → choose between alternative hypotheses

Simple hypothesis = hypothesis which is completely specified
E.g.: Theoretical model and ~~model~~ parameter values

Composite hypothesis = ensemble of more than one simple hypotheses
E.g.: model with free parameters (equivalent to infinite list of hypotheses for all possible values of the parameter).

Goals (more specific wording):

→ Take $H_0$ as the null hypothesis (background)
$H_1$ as the alternative hypothesis (signal + background)
$H_0$ and $H_1$ are a complete set: $P(H_0) + P(H_1) = 1$ (Bayesian)

Test of hypothesis = use data to verify/disprove $H_0$ ~~vs~~ $H_1$

→ Take $H_0$ as a given hypothesis
$\overline{H_0}$ as all other (unspecified) ~~hy~~ possible hypotheses

Goodness of fit = use data to verify/disprove $H_0$ vs $\overline{H_0}$

- Test statistic

Let $\vec{x}$ be some measured data distributed as:

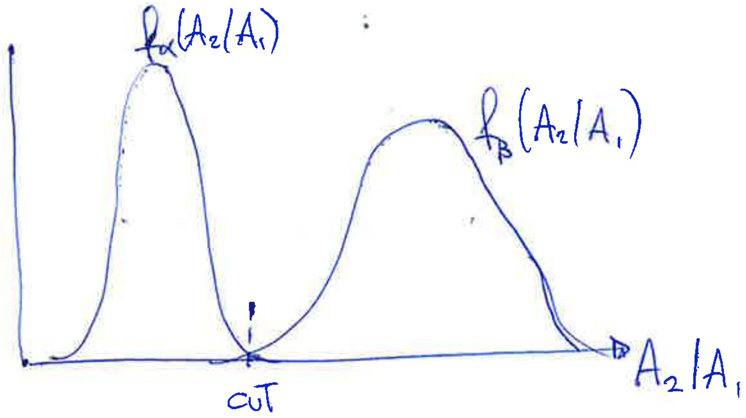
$f_0(\vec{x}|H_0)$ if $H_0$ is true

$f_1(\vec{x}|H_1)$ if $H_1$ is true

Let $H_0$ and $H_1$ be a complete set of ~~the~~ alternative hypotheses.

We want to develop a method to determine whether the observed data agree better with $H_0$ or $H_1$.

3) Decide some cut on $A_2/A_1$




4) Measure the "physics data" (whatever they are) and use the previous method to distinguish $\alpha$ from $\beta$

- Selection, ~~and~~ misidentification and significance

→ Selection efficiency = ~~expected~~ fraction of signal events that are expected to be correctly identified
$\varepsilon_s = 1-\beta$

→ Misidentification probability = fraction of background events that are expected to be erroneously identified as signal
$\varepsilon_b = \alpha$ = significance

→ Critical region = region where we expect the signal
$w$

→ Acceptance region = region where we expect the background
$W-w$ $\stackrel{\text{!}}{=}$ region where we accept $H_0$ as true

In general, the misidentification probability is also called "significance level".
When we design a hypothesis test, we need to specify the desired level of significance $\alpha$, ~~i.e. the amount of fraction of backgr probability~~
i.e. to which extend we are willing to accept the misidentification of data induced by $H_0$ with data induced by $H_1$:

$$P(t(\vec{x}) \in w \mid H_0) = \alpha$$

Hyp 3

Given a predefined value of $\alpha$, we want to find the region $w$ which maximizes $(1-\beta)$.

We can rewrite:

$$1-\beta = \int_w \frac{f_1(\vec{x}|H_1)}{f_0(\vec{x}|H_0)} f_0(\vec{x}|H_0)\, d\vec{x} = E_w\left[\frac{f_1(\vec{x}|H_1)}{f_0(\vec{x}|H_0)}\right]$$

The best critical region $w$ is the one that satisfies:

$$\boxed{\lambda(\vec{x}) = \frac{f_1(\vec{x}|H_1)}{f_0(\vec{x}|H_0)} \geq k_\alpha}$$

with $k_\alpha$ chosen so that the desired significance is achieved.

This is the Neyman-Pearson lemma.

Notice that: → the NP lemma is valid only if the PDFs are known (including the values of their parameters).

~~Otherwise, this~~

→ The NP lemma provides ~~an upper~~ the most-powerful test. If we don't know the parameter values, the power of any test will be $\leq$ than that of NP.

Practical instructions (assuming parameter values are known):

1) Evaluate $f_0(\vec{x}|H_0)$ and $f_1(\vec{x}|H_1)$
2) Evaluate $\lambda(\vec{x})$ and find region $w$
3) Do your measurement, obtaining data $\vec{x}$.
4) If $\lambda(\vec{x}) > k_\alpha \Rightarrow H_1$ is considered true

   If $\lambda(\vec{x}) \leq k_\alpha \Rightarrow H_0$ is considered true

- Discoveries and upper limits

- Suppose we are searching for a new physics process. We make a measurement and we need to quote ~~the~~ a result. How do we decide wether the data tell us that there is new physics?

→ Frequentist approach: measure the "significance", i.e. the probability that a background statistical fluctuation produces a fake signal at least as intense as the measured one.

→ Bayesian approach: quantify the ~~degree~~ posterior degree of belief on the hypotheses $H_0$ and $H_1$.

- P-values

To claim a discovery, we need to determine that the data are sufficiently inconsistent with the bkg-only hypothesis $H_0$.

⇒ We can use a test statistics $t$ to measure such inconsistency!

p-value = probability $p$ that the test statistic $t$ ~~measures~~ assumes a value greater or equal to the measured value $\hat{t}$ due to an overfluctuation of the background.

↳ The p-value has a uniform distribution in $[0, 1[$ if $H_0$ is true

↳ The p-value tends to have small values if $H_1$ is true

→ Example: Event counting experiment

Take the number of observed events $n$ as a test statistic.

p-value = probability to measure $\geq n$ events under the $H_0$ hypothesis.

→ If $b$ is large, we can approximate the $\mathcal{L}$ with a Gaussian with $\mu = b$ and $\sigma = \sqrt{b}$.

An excess $n - b = s$ must be compared with $\sqrt{b}$.

The significance will be: $Z = \frac{n-b}{\sqrt{b}} = \frac{s}{\sqrt{b}}$

→ If $b$ ~~~~ is large and has some large uncertainty $\sigma_b$,

The significance will be: $Z = \frac{n-b}{\sqrt{b + \sigma_b^2}}$

→ If $b$ is small, one can prove that the significance is:

$$Z = \sqrt{2\left[(s+b)\ln\left(1 + \frac{s}{b}\right) - s\right]}$$

• Significance with likelihood ratio

Take again two nested hypotheses $H_0$ and $H_1$, with $H_0 = H_1(s = 0)$
↳ signal strength

We can define the test statistic:

$$\lambda(s, \vec{\theta}) = \frac{\mathcal{L}_{s+b}(\vec{x} \mid s, \vec{\theta})}{\mathcal{L}_b(\vec{x} \mid \vec{\theta})}$$

→ Notice that we inverted numerator and denominator w.r.t. Wilks Theorem

A minimum of $-2\ln\lambda$ at $s = \hat{s}$ indicates the possible presence of a signal with strength $\hat{s}$.

According to Wilks Theorem, $2\ln\lambda$ follows a $\chi^2$ distro with 1 DOF.

An approximate estimate of the significance is: $Z = \sqrt{2\ln\lambda(\hat{s})}$

→ This is a local significance that can be used if we have a "perfect" prior knowledge of the other parameters $\vec{\theta}$.

→ If we estimate $\vec{\theta}$ from the data, we need to consider the "look elsewhere effect".

- Bayes Factor / Ratio

If $H_0$ and $H_1$ are not a complete set of hypotheses, we can't compute ~~$P(H_i|\vec{x})$ becau~~ $P(\vec{x})$, and therefore $P(H_i|\vec{x})$.

However, we can compute the ratio:

$$\underbrace{\frac{P(H_1|\vec{x})}{P(H_0|\vec{x})}}_{\text{Posterior odds}} = \underbrace{\frac{P(\vec{x}|H_1)}{P(\vec{x}|H_0)}}_{\text{Bayes factor}} \underbrace{\frac{\pi(H_1)}{\pi(H_0)}}_{\text{prior odds}}$$

If $\pi(H_0) = \pi(H_1)$, the posterior odds are identical to the Bayes factor.

One can then set some thresholds on the Bayes factor (or on the posterior odds) to claim "evidence" and "discovery".

→ Example: Evidence (Bayes factor)

- Numerical ~~consi~~ and practical considerations

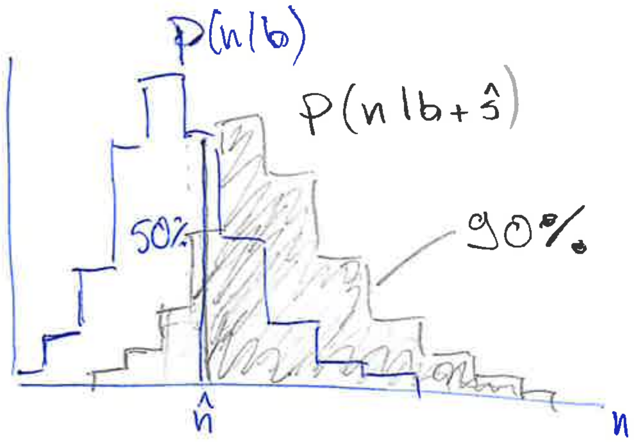
When running a Bayesian analysis, we might need to face three different problems involving 3 different algorithms:

1) Finding the global mode of posterior → Minimizer algorithm
2) ~~Finding~~ Interval estimation → MCMC
3) Computing "significance" (doing model testing or Bayes factor) → n-dimensional integration of full posterior PDF

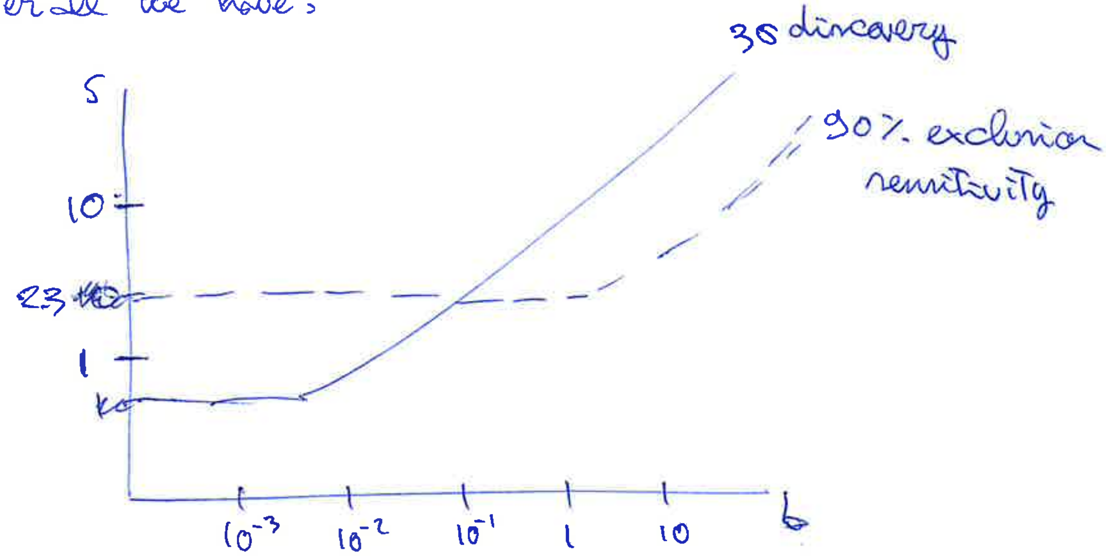
At the moment, there is <u>no algorithm</u> that does all 3 of them at the same time. ~~M~~ Moreover, MCMC and integrators are inefficient.

⇒ If you have an idea for an algorithm that can do all 3 things with a high efficiency (no discarded points) and that can work for dim > 50, please let me know, because I want to work with you!

Exclusion: $\begin{cases} P(n \le \hat{n} \mid b) \geq 50\% \\ P(n \geq \hat{n} \mid b+s) \geq 90\% \end{cases}$




Overall we have,



- Distribution-free Test

A goodness of fit test is distribution-free if the distribution of t is known independently of $H_0$.

↳ Also the p-value is independent of $H_0$

↳ We can compute p for any $H_0$, and compare it to tabulated data that were calculated once for all!

↳ Eventually p might depend on the number of events, number of bins in a histogram, or number of constrains in a fit.

- Distribution-free Tests for histograms

Suppose we measure n times a ~~random~~ variable $\vec{x}$ with PDF $f(\vec{x})$.

Then we have n values of the test statistic t, with PDF $f(t)$.

Assume n is a Poisson variable.

If we bin the n values of t, we get a histogram where each bin follows a Poisson statistic.

↳ We have lost the dependence on $f(t)$

- Pearson's $\chi^2$ Test for histograms

Let's assume the number of entries in each bin $n_i$ is large enough that we can approximate the corresponding Poisson to a Gaussian.

Then we can use as a statistic:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - \lambda_i)^2}{V[\lambda_i]}$$

where $m$ = # of bins

$\lambda_i$ = expectation value for bin i (depends on $f(x^2)$)

$V[\lambda_i]$ = variance for $\lambda_i$

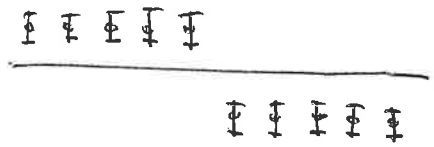
The PDF of $\chi^2$ is:

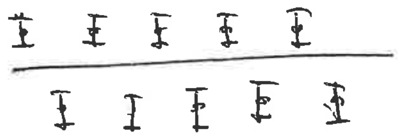
$$f(\chi^2, ndf) = \frac{1}{2^{\frac{ndf}{2}}\,\Gamma\left(\frac{ndf}{2}\right)} (\chi^2)^{\frac{ndf}{2}-1} e^{-\frac{\chi^2}{2}} \longrightarrow \begin{array}{l} \text{mean} = ndf \\ V[\chi^2] = 2\,ndf \end{array}$$

- Wald - Wolfowitz run-Test

Notice: The Pearson's $\chi^2$ Test does not take into account the sign of the deviations.

~~Ex~~ The following two cases would give exactly the same $\chi^2$:





Let's define as "run" each region ~~with~~ ~~of~~ measurements with residuals of the same sign.

↳ The number of runs $r$ is binomial

Denoting with $n_+$ = number of measurements with positive residuals

$n_-$ = " " " negative "

Number of possible combinations: $\dfrac{n!}{n_+!\, n_-!}$

Expected number of runs: $E[r] = 1 + \dfrac{2 n_+ n_-}{n}$

Variance: $V[r] = \dfrac{2 n_+ n_- \cdot (2 n_+ n_- - n)}{n^2 (n-1)}$

With $n \geq 20$, $r$ can be approximated by a Gaussian, so we have:

$$\varphi = \frac{r - E[r]}{\sqrt{V[r]}}$$

- $\chi^2$ Test for unbinned data

Suppose we measure $\vec{x}$ n times and fit it with $f(\vec{x}|\vec{\theta})$

We can still run a $\chi^2$ test by binning the data $x$ in $m$ bins:

$$\underbrace{\chi^2 = 2\sum_{\substack{i=1\\ n_i\neq 0}}^{m} n_i \ln\frac{n_i}{\hat{\lambda}_i}}_{\text{Multinomial case}}$$

where $n_i$ = number of events in bin $i$
$\hat{\lambda}_i$ = expectation value for $n_i$ obtained from fit

For Poisson distributed data:

$$\chi^2 = 2\sum_{\substack{i=1\\ n_i\neq 0}}^{m} n_i \ln\frac{n_i}{\hat{\lambda}_i} + \hat{\lambda}_i - n_i$$

$\rightarrow$ In large-sample limit, it follows a $\chi^2$ with $(m-d)$ dof

- Test using max-$\mathcal{L}$ estimate

Suppose we use $\mathcal{L}_{max}$ as a test statistic, and compare the measured value $\hat{\mathcal{L}}_{max}$ to the set of $\mathcal{L}_{max}$ from Toy-MC experiments, where we set the value of $\vec{\theta}$ to theirs expected true values.

We have that: 1) $\mathcal{L}_{max}$ distributions are not well separated under different hypotheses

$\hookrightarrow$ Do <u>not</u> use $\mathcal{L}_{max}$ as a test-statistics for GOF.

- Smirnov - Cramer - Von Mises Test

Use as test statistics: $$W^2 = \int_{-\infty}^{+\infty} \left[F_n(x) - F(x)\right]^2 \cancel{dF(x)}\ f(x)\,dx$$

↳ Instead of using the single point where the difference is largest, we use the integral of the squared difference

- Anderson - Darling Test:

$$A^2 = n \int_{-\infty}^{+\infty} \frac{\left(F_n(x) - F(x)\right)^2}{F(x)\left(1 - F(x)\right)}\, dF(x)$$

→ put more weight on the tails of the distribution