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 ~~X~~ + 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

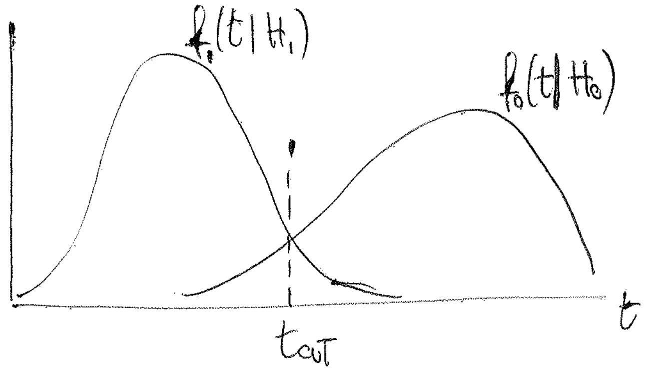
$f_1(\vec{x}|H_1)$ if H_1 is true

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.

Instead of using all data $\vec{x}$, we can use some statistic $t = t(\vec{x})$ and use its PDF to test the hypotheses H_0 and H_1.
Such statistic is called "Test statistic".

In general, we will have that $f_0(t|H_0)$ differs from $f_1(t|H_1)$.
Therefore we can set a cut on t, ~~and~~ (before doing the measurement), and "choose" H_0 or H_1 depending on the measured value of t, $\hat{t}$:

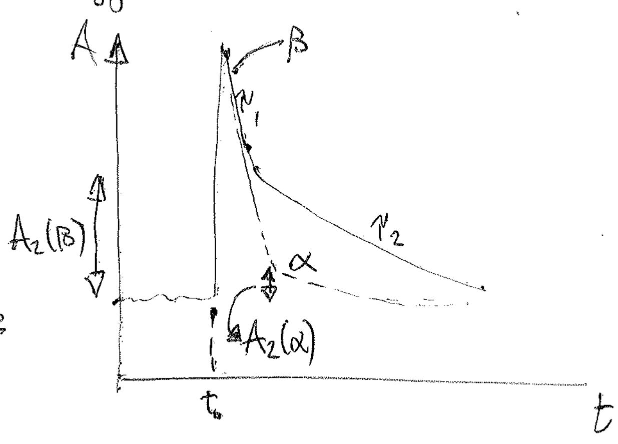


→ Example: Particle identification with scintillating crystal

Suppose we have a scintillating crystal producing light with 2 time constants:
$\tau_1 = 50$ ns
$\tau_2 = 200$ ns

Suppose the amplitude ~~of the~~ connected to the second time constant depends on the particle depositing the energy, so that

$$A_2(\tau_2, \alpha) << A_2(\tau_2, \beta)$$



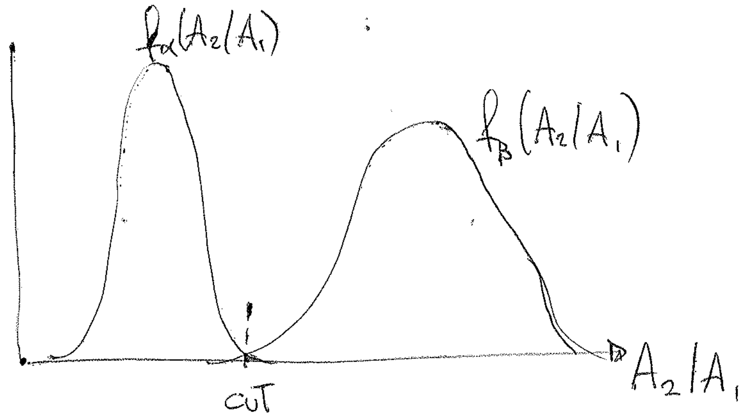
Instead of using the full pulse shape, we can fit the pulses with something like: $A(t) = A_1 e^{-\frac{t-t_0}{\tau_1}} + A_2 e^{-\frac{t-t_0}{\tau_2}}$

$t = A_2/A_1$
And use A_2/A_1 as a test statistic.

We will proceed as follows:
1) Calibrate with an α source to evaluate $f_\alpha(A_2/A_1)$
2) Calibrate with a β source to evaluate $f_\beta(A_2/A_1)$

Hyp. 2

3) Decide some cut on A_2/A_1



4) Measure the "physics data" (whatever they are) and use the previous method to distinguish α from β

- 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 w = region where we expect the signal

→ Acceptance region $W-w$ = region where we expect the background $\stackrel{!}{=}$ 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 α, ~~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:

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

Similarly, we can define the "power" of a test, as the probability of data produced by H_1 to be correctly identified:

$$P(t(\vec{x}) \in w \mid H_1) = 1 - \beta = \varepsilon_s$$

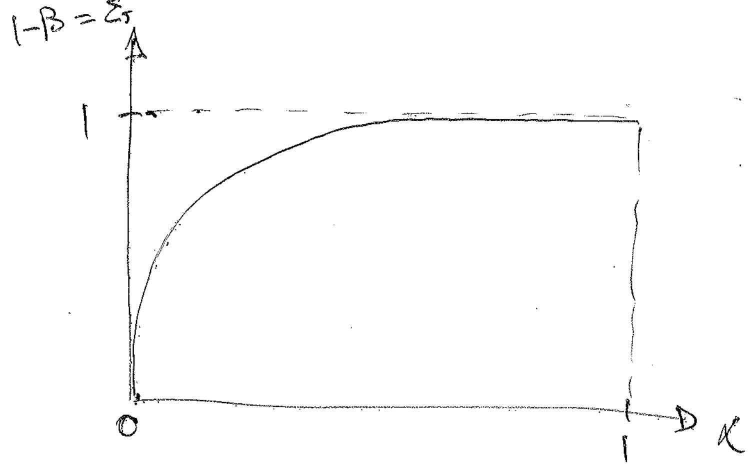
Conversely, β is the probability to misidentify data produced by H_1 as if H_0 were true.

We can define two type of errors:

a) Error of the first kind: rejecting H_0 when it is true $\Rightarrow$ prob. $= \alpha$

b) Error of the second kind: accepting H_0 when it is false $\Rightarrow$ prob. $= \beta$

We can plot the so-called "receiver operating characteristic" (ROC) curve:



Clearly, we would like to find the "most powerful" test of hypothesis!

- Neyman - Pearson lemma

Finding the most powerful test is equivalent to finding the best critical region in x-space.

Suppose we measure the variable $\vec{x}$ with the usual PDFs $f_0(\vec{x}|H_0)$ and $f_1(\vec{x}|H_1)$

Using the measurement $\vec{x}$ itself as a statistic, we have:

$$\int_w \mathcal{L}_0(\vec{x} \mid H_0)\, d\vec{x} = \alpha$$

$$\int_w \mathcal{L}_1(\vec{x} \mid H_1)\, d\vec{x} = 1 - \beta$$

where $\mathcal{L}_i$ is f_i evaluated at the measured data $\vec{x}$

$\boxed{\text{Hyp. 4}}$

Given a predefined value of α, 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_α chosen so that the ~~signific~~ 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

• Projective likelihood ratio Test : Wilks' Theorem

Suppose we have Two "nested" hypotheses H_0 and H_1, so That ~~The~~ The parameter values of H_0 are a special case of H_1 (e.g. signal = 0).

~~We can define The~~

We can divide The total parameter space Ω in Ω_0 and Ω_1

$$H_0 : \vec{\vartheta} \in \Omega_0 \subset \Omega$$

$$H_1 : \vec{\vartheta} \in \Omega_1 = \Omega - \Omega_0$$

We can define The maximum $\mathcal{L}$ ratio :

$$\boxed{\lambda = \frac{\max\limits_{\vec{\vartheta} \in \Omega_0} \mathcal{L}_0(\vec{x} \mid H_0)}{\max\limits_{\vec{\vartheta} \in \Omega_1} \mathcal{L}_1(\vec{x} \mid H_1)}} \Rightarrow -2 \ln \lambda = \chi^2$$

Wilks Theorem : If H_0 is True and for $n \to \infty$,
$-2 \ln \lambda$ has a χ^2 distribution with NDOF = $\dim(\Omega_1) - \dim(\Omega_0)$

What does it mean?

→ Example : Deviation of parameter μ from predicted value

H_0 : $\mu = \mu_0$ = value predicted by Theory

H_1 : $\mu \neq \mu_0$

$$\chi^2(\mu_0) = -2 \ln \frac{\max\limits_{\vec{\vartheta}} \prod_{i=1}^{n} \mathcal{L}(\vec{x}_i \mid \mu_0, \vec{\vartheta})}{\max\limits_{\mu, \vec{\vartheta}} \prod_{i=1}^{n} \mathcal{L}(\vec{x}_i \mid \mu, \vec{\vartheta})}$$

→ $\mathcal{L}$ maximized over $\vec{\vartheta}$ for a fixed value of $\mu = \mu_0$

→ max $\mathcal{L}$ over entire parameter space

- Discoveries and upper limits

- Suppose we are searching for a new physics process. We make a measurement and we need to quote a result. How do we decide wether the data tells 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

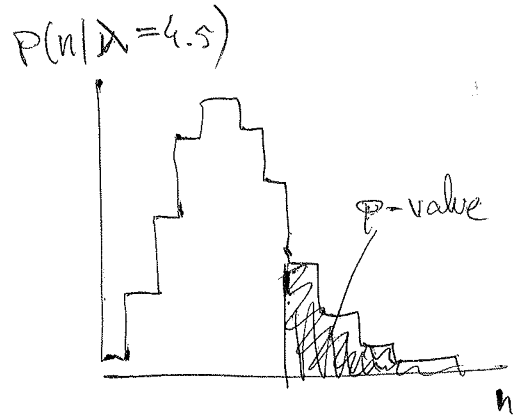
↳ The p-value tends to have small values if H_1 is true

→ 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.

→ Example: p-value for Poissonian Counting



H_0: Poisson counting with $\lambda_b = 4.5$, $\lambda_s = 0$

H_1: ~~λ_s~~ $\lambda_b = 4.5$, $\lambda_s > 0$

Measurement → $n = 8$

$$p\text{-value} = P(n \geq 8 \mid H_0) = \sum_{n=8}^{\infty} \text{Poisson}(n, 4.5) = \dots = 0.087$$

- Significance level

Instead of quoting a p-value, we normally quote the number of STDs that correspond to an area equal to the p-value under the right tail of a standard normal distribution:

$$p = \int_z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1 - G(z) = G(-z) = \frac{1}{2}\left[1 - \text{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$$

z is called "significance level".

By convention, we claim: → evidence if $z \geq 3$

→ observation/discovery if $z \geq 5$

- Significance for Poissonian counting

H_0: expected b events (b known)

H_1: expected $b + s$ events ($s \geq 0$ unknown)

The likelihood is: $\mathcal{L}(n|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$

To compute the significance, we need to compare the measured number of events n with the expected background b under the H_0 hypothesis ($s=0$).

→ 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 ~~base~~ is large and has some large uncertainty σ_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 $H_0 = H_1(\cancel{s} = 0)$
↳ signal strength

We can define the test statistic:

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

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

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

According to Wilks Theorem, $2\ln\lambda$ follows a χ^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~~ the other parameters $\vec{\theta}$.

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

- Significance with Toy-MC

A more general approach can be ~~used~~ obtained with Toy-MC.

1) Generate many Toy-MC datasets with no signal ($s=0$), obtaining an approximate distribution of $-2\ln\lambda$

2) The p-value is the probability that λ is $\leq$ the observed value $\hat{\lambda}$:

$$p = P(\lambda(\theta) \leq \hat{\lambda})$$

This is equal to the fraction of Toy-MC for which $\lambda(\theta) \leq \hat{\lambda}$.

- Bayesian method for hypothesis testing

~~In case~~

Suppose we have two hypotheses H_0 and H_1 representing a complete set:

$$P(H_0|\vec{x}) + P(H_1|\vec{x}) = 1$$

We can apply the Bayes Theorem to the two hypotheses:

$$P(H_i|\vec{x}) = \frac{P(\vec{x}|H_i)\,\pi(H_i)}{P(\vec{x})}$$

Where: $P(\vec{x}) = P(\vec{x}|H_0)\pi(H_0) + P(\vec{x}|H_1)\pi(H_1)$

and $$P(\vec{x}|H_0) = \int_{\Omega_\theta} \mathcal{L}_0(\vec{x}|\vec{\theta})\,\pi(\vec{\theta})\,d\vec{\theta} \quad = \text{"evidence" of } H_0$$

$$P(\vec{x}|H_1) = \int_{\Omega_{\theta,s}} \mathcal{L}_1(\vec{x}|s,\vec{\theta})\,\pi(\vec{\theta})\,\pi(s)\,ds\,d\vec{\theta} \quad s = \text{signal parameter}$$

Again, we need to ~~de~~ assign some prior both to the parameters $s, \vec{\theta}$, and to the models H_0 and H_1.

~~for~~ A general choice is: $\pi(H_0) = \pi(H_1) = 0.5$

At this point we set a threshold on $P(H_0)$ in order to claim the "evidence" or discovery.

- 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: <u>Evidence</u> (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. 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!

• ~~Goodness of fit~~

• Sensitivity

Before doing a measurement, or when planning an experiment, we can ask ourself what is the expected probability to make a discovery, or what's the expected 90% limit we will place.

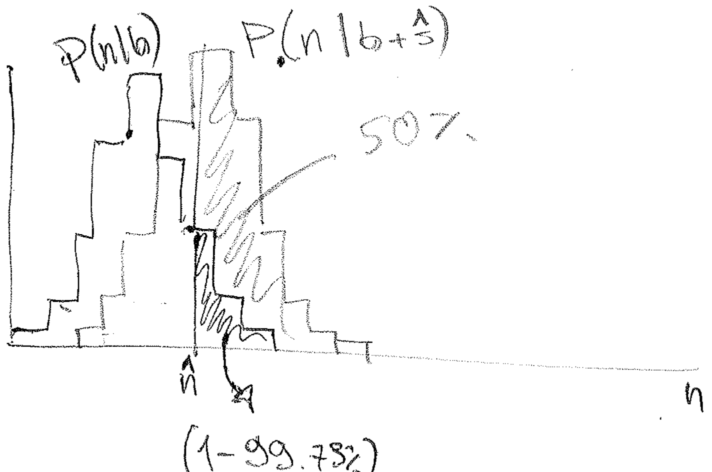
→ Discovery ~~probability~~ sensitivity = expected number of signal counts (or signal strength) for which an experiment has a 50% chance to observe an excess over the background at 3 σ or 5 σ ~~the~~ significance.

→ Exclusion sensitivity = expected number of signal events (or signal strength) that an experiment has 50% chance to exclude at ~~..~~ 90% CL

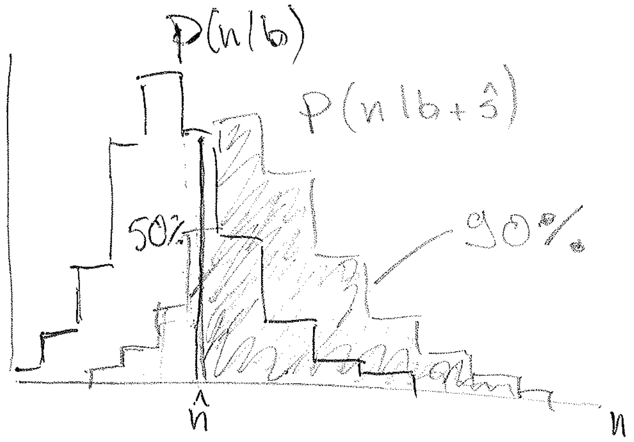
→ For a counting experiment with known expected background b, the two sensitivities can be computed by solving these equations:

$$\text{Discovery:}\quad \begin{cases} P(n \leq \hat{n} \mid b) \geq 99.73\% \quad (\text{for } 3\sigma) \\ P(n \geq \hat{n} \mid b+\hat{s}) \geq 50\% \end{cases}$$

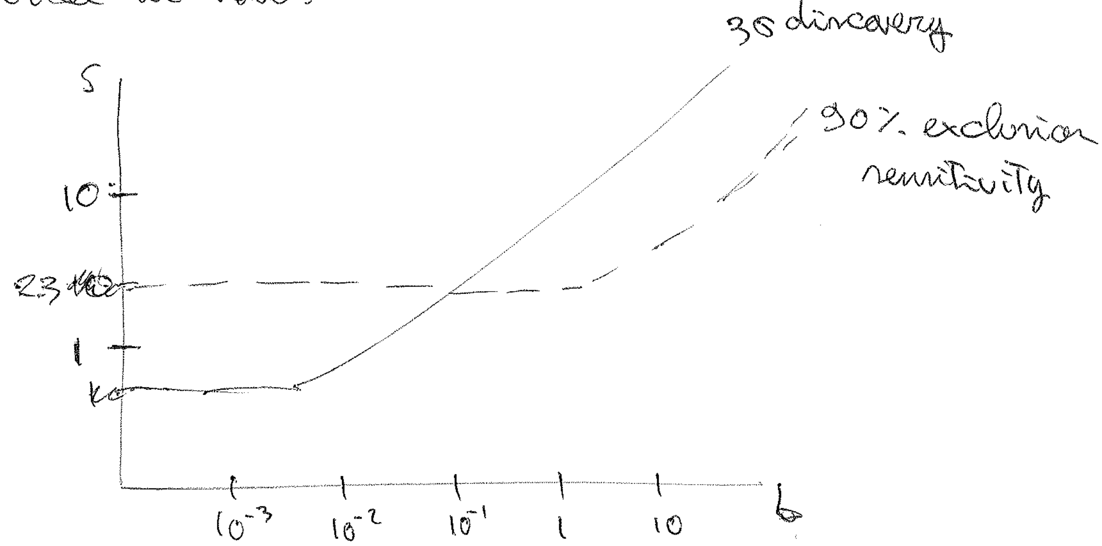
↳ Find $\min(\hat{n})$ that satisfies eq. 1
↳ Find $\hat{s} = \min(s)$ that satisfies eq. 2
↳ $\hat{s}$ is the discovery sensitivity



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



Overall we have:



GOODNESS OF FIT

- Address the question: How well do the data agree with the functional form predicted by a hypotesis H_0?

→ Here we're testing only one null hypothesis H_0, or one hypothesis H_0 against the infinite and unspecified set of alternatives to H_0.
↳ The theoretical basis is less robust than in hypothesis testing. Nevertheless we can successfully perform the test, using a frequentist approach.

- Procedure (general for all methods):

1) Choose/construct a test statistic $t(\vec{x})$ which is sensitive to the level of agreement between the data and the hypothesis H_0.
Let's assume that we choose t so that larger values of t indicate a worse agreement.

~~2) Compute the probability p that, assuming H_0 to be true, the~~

2) Assume H_0 true and compute the probability p that, repeating the measurement many times, we get a value of t greater or equal than the actually measured one. This is the p-value.
↳ Small p = bad agreement between data and H_0 = bad fit
↳ p-value calculation could depend or not on the PDF of t.

- EX.: Test statistic for Poisson distro

Assume we measure a discrete variable n with Poisson PDF with $\lambda = 17.3$, and obtain $n = 12$.
What is the p-value, or the level of compatibility with the hypotesis of n belonging to a Poisson with $\lambda = 17.3$?

~~for Take $t(n) = (n-\lambda)$~~ Take $t = |n - \lambda|$

$$p = \sum_{\substack{n \\ t \geq 5.3}} \frac{e^{-\lambda}\lambda^n}{n!} = \sum_{n=1}^{12} \frac{e^{-\lambda}\lambda^n}{n!} + \sum_{n=23}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} = 0.229$$

- 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 ~~[illegible]~~ 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 χ^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

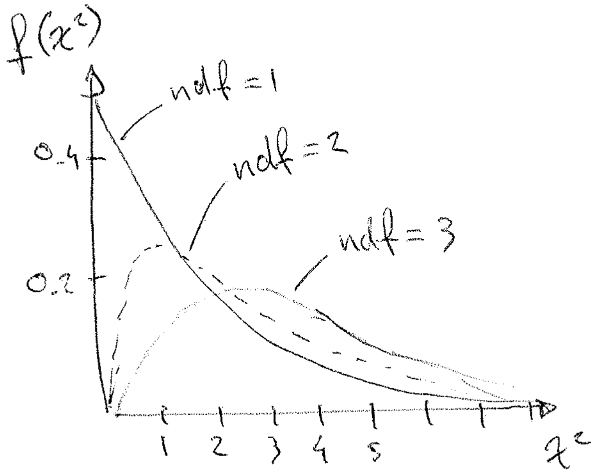
$\lambda_i =$ expectation value for bin i (depends on $f(x^2)$)

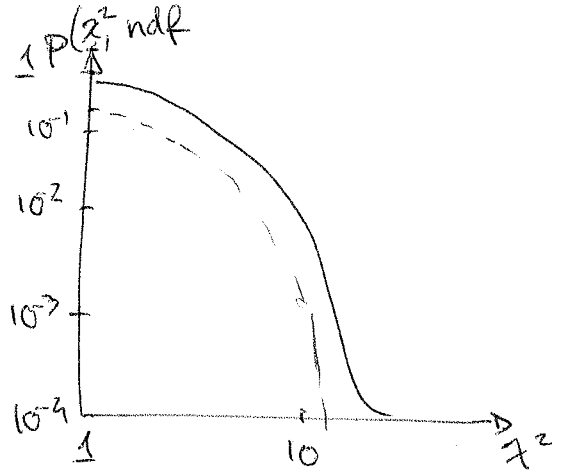
$V[\lambda_i] =$ variance for λ_i

The PDF of χ^2 is:

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

The p-value is: $p = P(\chi^2, n) = \int_{\chi^2}^{+\infty} f(z, n)\, dz$





If we fit $f(n|\vec{\theta})$ on the data, and $\dim(\vec{\theta}) = d$, the minimize value of χ^2 will follow a $\chi^2(n-d)$ distribution.

~~If the total~~

Possible situations:

- χ^2 too small (too good fit) → errors ~~ha~~ have been overestimated, or the data have been selected
  ↳ usually if $p < 0.05$
- χ^2 too large → Hypothesis H_0 is wrong, or there are unaccounted correlations in the bins

Reduced $\chi^2 = \dfrac{\chi^2}{ndf}$ → gives also some info on agreement between data and H_0
↳ not as informative as p-value

→ What if n_i is too small to use the χ^2 probability?
↳ Use a MC approach!

- Wald - Wolfowitz run-Test

Notice: The Pearson's χ^2 Test does not Take into account the sign of the deviations.

~~the~~ The following two cases would give exactly the same χ^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: $$E[r] = 1 + \frac{2 n_+ n_-}{n}$$

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

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

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

- Combining Tests

Suppose that we run the χ^2 and run-test on same data, obtaining the p-values $\hat{p}_1$ and $\hat{p}_2$. How do we combine them?

Let's assume that:
1) p_1 and p_2 are uniformly distributed in $[0, 1]$

~~2) p_1 and p_2 are independent~~

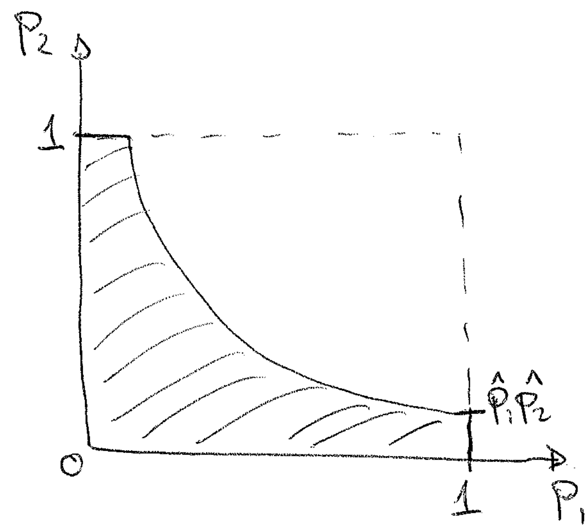
2) t_1 and t_2 are independent: $P(t_1, t_2 | H_0) = P(t_1|H_0) P(t_2|H_0)$

This might be hard to prove, but holds for the χ^2 and run-test, because the first does not use the information of the sign, and the second uses only the sign.

The probability that $p_1 p_2 \leq \hat{p}_1 \hat{p}_2$ is:

$$P(p_1 p_2 \leq \hat{p}_1 \hat{p}_2) = \int_0^1 dp_1 \int_0^{\hat{p}_1 \hat{p}_2 / p_1} p_1 p_2 \, dp_2$$

$$= \hat{p}_1 \hat{p}_2 \left[1 - \ln(\hat{p}_1 \hat{p}_2) \right]$$



$$\Rightarrow P(p_1 p_2 \leq \hat{p}_1 \hat{p}_2) > \hat{p}_1 \hat{p}_2$$

→ With n tests, we can ~~compute~~ prove that:

$$p = -2 \ln \prod_{i=1}^{n} \hat{p}_i$$

is distributed as a χ^2 with $2n$ dof.

- χ^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 χ^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 χ^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 their 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.

• Kolmogorov - Smirnov Test

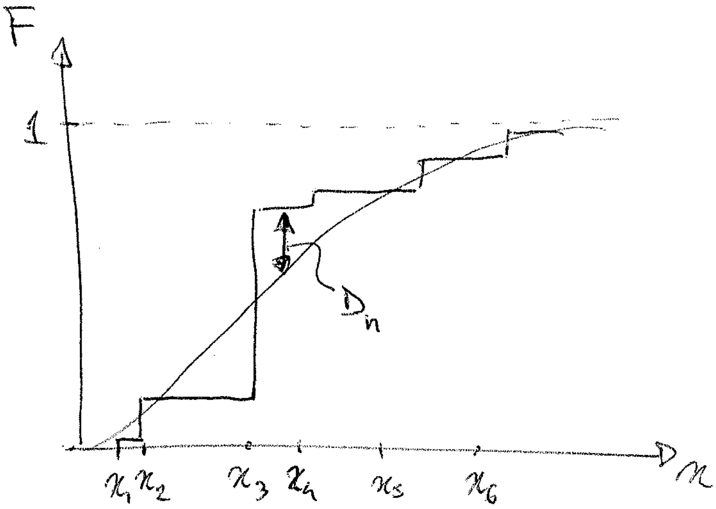
Take a set of n measurements $x_i = x_1, \dots, x_n$ ordered in increasing values of x.

The discrete cumulative distribution is:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \vartheta(x - x_i) \quad \text{with } \vartheta(x_i) = \begin{cases} 1 & \text{if } x \geq 0 \\ \emptyset & \text{if } x < 0 \end{cases}$$

This can be compared with the cumulative PDF of $f(x)$:

$$F(x) = \int_{-\infty}^{x} f(y)\, dy$$



We can take as a test-statistic:

$$D_n = \max \left| F_n(x) - F(x) \right|$$

For large n, D_n converges to $\emptyset$ in probability.

One can prove that: 1) The distribution of $K = \sqrt{n}\, D_n$ does not depend on $f(x)$

2) The probability that $K \leq k$ with k arbitrary is The Kolmogorov distribution:

$$P(K \leq k) = \frac{\sqrt{2\pi}}{k} \sum_{i=1}^{\infty} e^{-(2i-1)^2 \pi^2 / 8k^2}$$

If the parameters $\vec{\vartheta}$ of $f(x|\vec{\vartheta})$ are obtained from a fit, one cannot use the Kolmogorov distribution but has to empirically obtain the distribution of K from Toy-MC.

→ The K-S test can be used also to compare two measurements, and to test if they are produced from the same PDF!

- 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