4 Classical goodness-of-fit tests

4.1 The binned data regime

Binned data may originate from the fact that the resolution of the detector is not sufficiently high to produce a continuous stream of data. Alternatively, researchers may decide to artificially bin the data. This can be done, for instance, for the sake of simplifying the estimation process or introducing Poisson uncertainties.

Suppose that our search region $\mathcal{X}$ is the energy interval $1 - 50$GeV and it has been divided into $i = 1, \ldots, n$ bins. Denote with x_i the value of the energy at the center of bin i. Moreover, let Y_i be the number of events in bin i. We typically assume that

$$Y_i \sim \text{Multinomial}[N, p(x_i, \boldsymbol{\theta})] \quad \text{or} \quad Y_i \sim \text{Poisson}[m(x_i, \boldsymbol{\theta})]$$

$N = \sum_{i=1}^{n} Y_i$

In the first case, the total number of events observed N, is treated as fixed, and $p(x_i, \boldsymbol{\theta})$ is the probability to observe an event in bin i, whereas, the expected number of events in bin i is $m(x_i, \boldsymbol{\theta}) = Np(x_i, \boldsymbol{\theta})$. In the second case, the total number of events observed is itself random and $m(x_i, \boldsymbol{\theta})$ is tells us how many event we expect to observe in bin i. Finally, $\boldsymbol{\theta}$ is a set of p (potentially unknown) parameters characterizing the shape of our mean function $m(x_i, \boldsymbol{\theta})$. Our goal is to assess if the our model $m(x_i, \boldsymbol{\theta})$ for the number of expected events in each of the bins is valid. Notice that, if $\boldsymbol{\theta}$ is unknown, we may proceed by estimating it via maximum likelihood (see Section 3.1). In this setting, however, we must consider the binned data likelihood, i.e.,

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{N!}{\prod_{i=1}^{n} y_i!} \prod_{i=1}^{n} \Big[p(x_i, \boldsymbol{\theta})\Big]^{y_i} \quad \text{for } Y_i \sim \text{Multinomial.}$$

there is a "-" here

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} \frac{[m(x_i, \boldsymbol{\theta})]^{y_i}}{y_i!} \exp\Big\{m(x_i, \boldsymbol{\theta})\Big\} \quad \text{for } Y_i \sim \text{Poisson,}$$

4.1.1 Pearson X^2 test

- Specification of the null hypothesis.

$$H_0 : E[Y_i] = m(x_i, \boldsymbol{\theta}) \qquad (\text{versus } H_1 : E[Y_i] \neq m(x_i, \boldsymbol{\theta}))$$

for all $i = 1, \ldots, n$.

- Specification of the test statistic.

$$X^2 = \sum_{i=1}^{n} \frac{[Y_i - m(x_i, \boldsymbol{\theta})]^2}{m(x_i, \boldsymbol{\theta})}$$

- Derivation of the distribution of the test statistic under H_0.

  - If , θ known:
    If $m(x_i, \theta) > 7$ for all $i = 1, \ldots, n$, as $N \to \infty$, under H_0,

    $$X^2 \approx \chi^2_{n-1}$$ Pearson 1900

  - If , θ unknown:
    If $m(x_i, \theta) > 7$ for all $i = 1, \ldots, n$, as $N \to \infty$, under H_0,

    $$X^2 \approx \chi^2_{n-p-1}$$ Fisher 1924

- Computation of the p-value.
  Let X^2_{obs} be the value of the test statistic X^2 evaluated on the data,

  - If , θ known: p-value$=P(\chi^2_{n-1} \geq X^2_{obs})$

  - If , θ unknown: p-value$=P(\chi^2_{n-p-1} \geq X^2_{obs})$

- What do we do after we have computed our p-value?
  We check if it is larger or smaller than α.

Warning: If the number of events in the bins is small, X^2 may lead to biased (non-admissible) tests. That is true even if the χ^2 approximation hold!

4.1.2 G^2 test (also known as deviance test or likelihood ratio)

- Specification of the null hypothesis: same as Pearson X^2.
- Specification of the test statistic.

if $Y_i \sim$ Poisson $\quad G^2 = 2\sum_{i=1}^{n}\left[Y_i \log \frac{Y_i}{m(x_i,\theta)} - (Y_i - m(x_i,\theta))\right]$

if $Y_i \sim$ Multinomial $G^2 = 2\sum_{i=1}^{n}\left[Y_i \log \frac{Y_i}{m(x_i;\theta)}\right]$

- Derivation of the distribution of the test statistic under H_0: same as Pearson X^2.
- Computation of the p-value: same as Pearson X^2.

Which one should be used?
In general, the χ^2 approximation is typically better for Pearson than it is for G^2, especially in the Poisson case. The χ^2 approximation may be better for G^2 if the counts are small. The literature on this is vast but a self-contained review can be found in

> Cressie and Read, 1989. *Pearson's X^2 and the Loglikelihood Ratio Statistic G^2: A Comparative Review.* International Statistical Review. `https://www.jstor.org/stable/1403582?seq=1#metadata_info_tab_contents`

Proof that G^2 is just the LRT between the model that assumes $E[Y_i] = m(x_i, \theta)$ and the model that assumes $E[Y_i] = y_i$ (Poisson case)

$$L(\theta; Y) = \prod_{i=1}^{n} \frac{[m(x_i, \theta)]^{y_i}}{y_i!} \exp\{-m(x_i, \theta)\}$$

$$\ell_0(\theta; y) = \sum_{i=1}^{n} [y_i \log m(x_i, \theta) - \log y_i! - m(x_i, \theta)]$$

$$\ell(y) = \sum_{i=1}^{n} [y_i \log y_i - \log y_i! - y_i]$$

$$LRT = -2[\ell_0(\theta; y) - \ell(y)] =$$

$$= 2 \sum_{i=1}^{n} [y_i \log y_i - \cancel{\log y_i!} - y_i - y_i \log m(x_i, \theta) + \cancel{\log y_i!} + m(x_i, \theta)]$$

$$= 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{m(x_i, \theta)} - (y_i - m(x_i, \theta))\right]$$

4.2 The unbinned data regime

4.2.1 The univariate case

Let X be continuous random variable taking values over the real line and distributed according to the distribution function $F(x) = P(X \leq x)$, i.e., $X \sim F$. When F is unknown, we are typically interested in testing if, for a given distribution function G,

$$H_0 : F = G \quad \text{versus} \quad H_1 : F \neq G \tag{3}$$

To perform this test, given a sample of N i.i.d. random variables from F, it is sensible to work with the process

$$v_G(x) = \sqrt{N}[F_N(x) - G(x)] \tag{4}$$

where $F_N(x)$ is the so-called *empirical distribution function*, it provides an estimate of $F(x)$ and it is defined as

$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{x_i \leq x\}}$$

The stochastic process $v_G(x)$ is called *empirical process*. Let's focus for a moment on understanding our $v_G(x)$...

- As $N \to \infty$, we have that $F_N(x) \to F(x)$.
  So, what do you think will happen to $v_G(x)$ when $N \to \infty$ (under H_0 and H_1, respectively)?
  - Under H_0: $v_G(x) \to 0 \quad \forall x$
  - Under H_1: the difference between F_N and G becomes more and more obvious and indeed it is magnified by a factor of $\sqrt{N}$

When dealing with univariate distributions, we typically consider the so-called *Probability Integral Transform* (PIT), that is, we set $T = G(X)$.

- *What is the distribution of $T = G(X)$ under H_0?*

$T \sim$ Uniform $[0, 1]$ for any G

cdf of T $\quad P(T \leq t) = t$

When applying such transformation, the empirical process $v_G(x)$ is transformed in the so called *uniform empirical process*, i.e.,

$$u(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} [\mathbb{1}_{\{t_i \leq t\}} - t] = \sqrt{N}\,[F_N(t) - t]$$

(empirical cdf of T)

How can we use $u(t)$ to test $H_0 : F = G$ versus $H_1 : F \neq G$? We simply take functionals of it, e.g.,

- **Kolmogorov statistic:** $\sup_t |u(t)|$
- **Cramer von Mises statistic:** $\int |u(t)|^2 dt$
- **Anderson-Darling statistic:** $\int |\frac{u(t)}{\sqrt{t(1-t)}}|^2 dt$

These are just some possibilities among an entire family of test statistics and they are all distribution free; that is, their distribution does not depend on the G that I am testing. Distribution-freeness is an important property of a test statistic as it allows us to avoid a case-by-case simulation/mathematical derivation.

- **Why is it useful to have an entire family of test statistics?**

5 The multivariate case

Now suppose that X takes value in $\mathbb{R}^D$ and suppose our G depends on a set of unknown parameters $\boldsymbol{\theta}$. The (multivariate parametric) empirical process

$$v_G(\boldsymbol{x}, \boldsymbol{\theta}) = \sqrt{N}[F_N(\boldsymbol{x}) - G(\boldsymbol{x}, \boldsymbol{\theta})], \quad \boldsymbol{x} = (x_1, \ldots, x_D) \tag{5}$$

can still be used to test $H_0 : F = G$ versus $H_1 : F \neq G$. In this case, however we can no longer exploit the PIT so we lose distribution-freeness. That is because, in this case,

$$T = G(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}) \not\sim \text{Unif}[0,1]$$

where $\widehat{\boldsymbol{\theta}}$ is an estimator (e.g., the MLE) of $\boldsymbol{\theta}$. Nonetheless, we can still construct an entire family of test statistics which extend those we have seen for the univariate case, e.g.,

- **Kolmogorov's statistic:** $\sup_x |v_G(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})|$
- **Cramer von Mises statistic:** $\int |v_G(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})|^2 dG(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})$
- **Anderson-Darling statistic:** $\int \left| \frac{v_G(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})}{\sqrt{G(\boldsymbol{x})\left[1 - G(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})\right]}} \right|^2 dG(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})$

and we can simulate their distribution via the parametric bootstrap.

There are ways to recover distribution-free and even make the simulation faster but these would required an entire new lecture. A suitable reference is

> Algeri, 2022. *K-2 rotated goodness-of-fit for multivariate data.* Physical Review D. `https://journals.aps.org/prd/abstract/10.1103/PhysRevD.105.035030`.

You may watch a seminar on this at `https://indico.cern.ch/event/1130770/` (starting from minute 46:00 of the recording).