Handout 11 - Goodness-of-Fit

1 Tests of hypotheses vs Goodness-of-Fit

Statistical inference is the area of statistics which aims to develop and study reliable tools to make conclusion on the phenomenon/population under study based on what has been observed on a data sample. Among the main inferential tools we have

- **Tests of hypotheses:** Given a postulated model for the data, we test it against an alternative model.
- **Goodness-of-fit tests:** Given a postulated model for the data we test it against all possible alternatives.

For instance

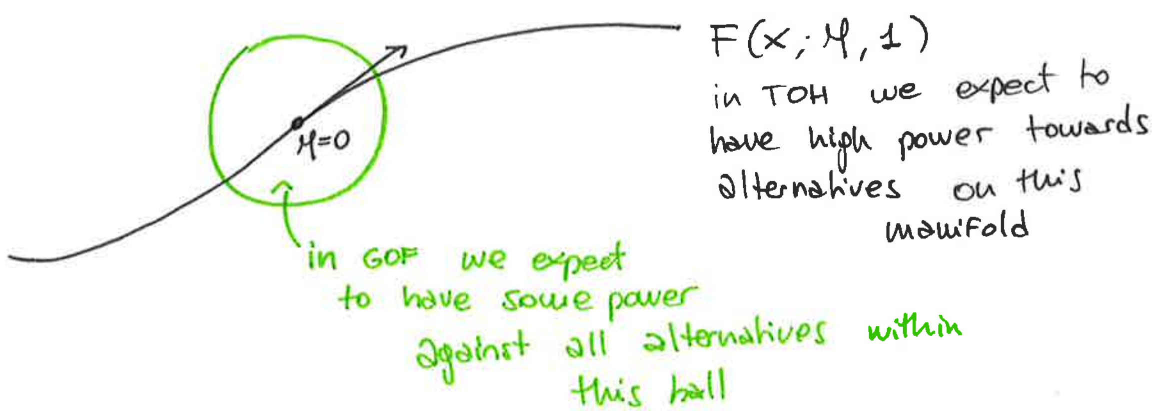
- Tests of hypotheses: e.g., we expect that $X \sim N(0,1)$, we test

  (TOH)
  $$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$

- Goodness-of-fit: e.g., we expect that $X \sim N(0,1)$, we test

  (GOF)
  $$H_0 : X \sim N(0,1) \quad (\text{versus} \quad H_1 : X \not\sim N(0,1)).$$

From a geometrical perspective:



2 The univariate case

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

$$H_0 : F = G \quad \text{versus} \quad H_1 : F \neq G \tag{1}$$

To perform this test, given n i.i.d. random variables $X_1, \ldots, X_n \sim F$, it is sensible to work with the process

$$\boxed{v_n(x; G, F) = \sqrt{n}[F_n(x) - G(x)]} \tag{2}$$

where $F_n(x)$ is the so-called *empirical distribution function* and it is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$$

we can look at $F_n(x)$ as an estimator of $F(x) = P(X \leq x)$

- **Suppose x is fixed to a specific value. What does $F_n(x)$ converge to?**
  - $\mathbb{1}_{\{X_i \leq x\}} \sim$ Bernoulli $(F(x))$
  - $F_n(x) \xrightarrow[n \to +\infty]{a.s.} F(x)$ SLLN

- **Based on the considerations above, if H_0 in (1) is true, what can we say about the asymptotic distribution of the components of $v_n(x; G, F)$ in (4)?**

- $nF_n(x) \sim \text{Binomial}(n, F(x))$

- $$\underbrace{\sqrt{n}\left[F_n(x) - F(x)\right]}_{V_n(x)} \xrightarrow[n\to+\infty]{d} N(0, F(x)(1-F(x)))$$ by De Moivre - Laplace

Suppose we consider x and x' fixed

$$\begin{aligned} \text{cov}(V_n(x), V_n(x')) &= E[V_n(x) V_n(x')] - \underbrace{E[V_n(x)]}_{=0} E[V_n(x')] \\ &= E[V_n(x) V_n(x')] \\ &= F(\underbrace{x \wedge x'}_{\text{minimum between } x \text{ and } x'}) - F(x)F(x') \end{aligned}$$

The stochastic process $v_n(x; F) = v_n(x; F, F)$, i.e.,

$$v_n(x; F) = \sqrt{n}[F_n(x) - F(x)] \tag{3}$$

is called *empirical process* and one can show that, as $n \to \infty$ is converges to a process called *Brownian Bridge* or *F-Brownian Bridge*, a Gaussian process with mean zero and covariance function:

$$F(x \wedge x') - F(x)F(x')$$

it follows from Donsker's theorem → extension of CLT for empirical processes

$\sup_t |F_n(t) - F(t)| \longrightarrow 0$ by Glivenko - Cantelli lemma → extension of LLN for empirical processes

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

- **What is the distribution of $T = F(X)$?**

$$P(T \leq t) = P(F(X) \leq F(x)) = P(X \leq F^{-1}(F(x))) = P(X \leq x) = F(x) = t$$

When applying such transformation, the empirical process $v_n(x; F)$ is transformed in the so called *uniform empirical process*, $u_n(t)$, i.e.,

$$u_n(t) = \sqrt{n}\,[F_n(t) - t]$$

$$\hookrightarrow F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{F(X_i) \leq t\}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{T_i \leq t\}}$$

where $F_n(t) = \frac{1}{n} I_{\{F(x_i) \leq t\}}$.

The process $u_n(t)$ converges to a *standard Brownian Bridge*, $u(t)$, a Gaussian process such that

- It is defined over the unit interval, i.e., $t \in [0, 1]$
- It has mean zero
- It has covariance function $\operatorname{cov}(u(s), u(t)) = t \wedge s - ts$

Moreover, $u(t)$ can be seen as a projection of the Brownian Motion on the unit interval, namely, $w(t)$, $t \in [0, 1]$, specifically:

$$u(t) = w(t) - tw(1).$$

Gaussian Process
- $w(0) = 0$
- $w(t) \sim N(0, t)$
- it has independent increment
- $\operatorname{cov}(w(t+h) - w(t), w(t)) = 0 \quad \forall h$

$\operatorname{cov}(w(t), w(s)) = t \wedge s$

- **What is the value of $u(t)$ at $t = 0$ and at $t = 1$?**

$u(0) = 0 = u(1)$

↑ this is why it is called "Bridge"

- **What is the advantage of referring to $u_n(t)$ instead of $v_n(x, F)$?**

it doesn't depend on F

$\Rightarrow$ it is said to be distribution - free !

Let's now go back to our original process in (4), i.e.,

$$v_n(x; G, F) = \sqrt{n}[F_n(x) - G(x)] \tag{4}$$

used to test

$H_0 : F = G$
vs
$H_1 : F \neq G$

- **What happens when $n \to \infty$?**

if H_0 is true $\longrightarrow 0$

if H_1 is true $\approx \sqrt{n}\,[F(x) - G(x)]$

the differences become more and more obvious as n increases!

So our test will be based on the process

$$u_n(t) = \sqrt{n}[G_n(t) - t] \tag{5}$$

with $t = G(x)$ and $G_n(t) = \frac{1}{n} I_{\{G(x_i) \leq t\}}$.
How can we use (5) to test $H_0 : F = G$ versus $H_1 : F \neq G$? We simply take functionals of it, e.g.,

- **Kolmogorov statistics:** $\sup_t |u_n(t)| \xrightarrow{d} \sup_t |u(t)|$
- **Cramer-von Mises statistics:** $\int |u_n(t)|^2 dt \xrightarrow{d} \int |u(t)|^2 dt$
- **Anderson-Darling statistics** $\int \left|\frac{u_n(t)}{\sqrt{t(1-t)}}\right|^2 dt \xrightarrow{d} \int \left|\frac{u(t)}{\sqrt{t(1-t)}}\right|^2 dt$

where the convergence is intended as $n \to \infty$, under H_0. These are just some possibilities among an entire family of test statistics and they are all distribution free!

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

each test statistic will be more sensitive than others towards different alternatives

e.g. Anderson-Darling will be more sensible to deviations in the tails of the distribution

3 The multivariate case

Now suppose that X takes value in $\mathbb{R}^p$. The (multivariate) empirical process

$$v_n(\boldsymbol{x};F) = \sqrt{n}[F_n(\boldsymbol{x}) - F(\boldsymbol{x})], \quad \boldsymbol{x} = (x_1, \ldots, x_p) \tag{6}$$

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. Nevertheless, we can still simulate the distribution of (6) under H_0 and which is that of an *F-Brownian Bridge* indexed by the sets of the form

$$(\infty, x_1] \times (\infty, x_2] \times \cdots \times (\infty, x_p]$$

and construct an entire family of test statistics which extend those we have seen for the univariate case, e.g.,

- **Kolmogorov's statistics:** $\sup_{\boldsymbol{x}} |v_n(\boldsymbol{x};F)| \xrightarrow{d} \sup_{\boldsymbol{x}} |v(\boldsymbol{x};F)|$
- **Cramer von Mises statistics:** $\int |v_n(\boldsymbol{x};F)|^2 dF(\boldsymbol{x}) \xrightarrow{d} \int |v(\boldsymbol{x};F)|^2 dF(\boldsymbol{x})$
- **Anderson-Darling statistics** $\int |\frac{v_n(\boldsymbol{x};F)}{\sqrt{F(\boldsymbol{x})(1-F(\boldsymbol{x}))}}|^2 dF(\boldsymbol{x}) \xrightarrow{d} \int |\frac{v(\boldsymbol{x};F)}{\sqrt{F(\boldsymbol{x})(1-F(\boldsymbol{x}))}}|^2 dF(\boldsymbol{x})$

where the convergence is intended as $n \to \infty$, under H_0.

There are ways to recover distribution-free in the multivariate setting and even when G depends on unknown parameters to be estimated. A recent advancement in this direction is the so called *Khamaladze-2* (K-2) transformation or *Khamaladze's rotation* and which, hopefully, we will be able to see by the end of the course. (If not, see Khmaladze, E. (2016). *Unitary transformations, empirical processes and distribution free testing. Bernoulli, 22(1), 563-588.*)