18.650 – Fundamentals of Statistics

5. Nonparametric hypothesis testing

Goodness of fit tests

Let X be a r.v. Given i.i.d copies of X we want to answer the following types of questions:

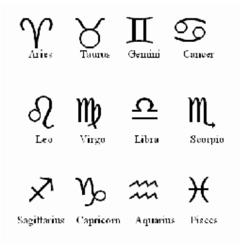
- Does X have distribution $\mathcal{N}(0,1)$? (Cf. Student's T distribution)
- Does X have distribution $\mathcal{U}([0,1])$?
- Does X have PMF $p_1 = 0.3,\ p_2 = 0.5,\ p_3 = 0.2$

These are all *goodness of fit* (GoF) tests: we want to know if the hypothesized distribution is a good fit for the data.

Key characteristic of GoF tests: no parametric modeling.

The zodiac sign of the most powerful people is....

Can your zodiac sign predict how successful you will be later in life? Fortune magazine collected the signs of 256 heads of the Fortune 500.



Fyi:
256/12
=21.33

Sign	Count
Aries	23
Taurus	20
Gemini	18
Cancer	23
Leo	20
Virgo	19
Libra	18
Scorpio	21
Sagittarius	19
Capricorn	22
Aquarius	24
Pisces	29

The zodiac sign of the most successful people is....

Sign	Count
Aries	23
Taurus	20
Gemini	18
Cancer	23
Leo	20
Virgo	19
Libra	18
Scorpio	21
Sagittarius	19
Capricorn	22
Aquarius	24
Pisces	29

In view of this data, is there statistical evidence that successful people are more likely to be born under some sign than others?

275 jurors with identified racial group.
We want to know if the jury is representative of the population of this county.

Race	White	Black	Hispanic	Other	Total
# jurors	205	26	25	19	275
proportion in county	0.72	0.07	0.12	0.09	1

Discrete distribution

Let $E = \{a_1, \ldots, a_K\}$ be a finite space and $(\mathbb{P}_{\mathbf{p}})_{\mathbf{p} \in \Delta_K}$ be the family of all probability distributions on E:

- $$\Delta_K = \left\{ \mathbf{p} = (p_1, \ldots, p_K) \in (0,1)^K : \sum_{j=1}^{K} p_j = 1 \right\}.$$
- For $\mathbf{p} \in \Delta_K$ and $X \sim \mathbb{P}_{\mathbf{p}}$,

$$\mathbb{P}_{\mathbf{p}}[X = a_j] = p_j, \quad j = 1, \ldots, K.$$

Goodness of fit test

- Let $X_1, \ldots, X_n \stackrel{iid}{\sim} \mathbb{P}_{\mathbf{p}}$, for some unknown $\mathbf{p} \in \Delta_K$, and let $\mathbf{p}^0 \in \Delta_K$ be fixed.

- We want to test:
$$H_0\colon\ \mathbf{p} = \mathbf{p}^0 \text{ vs. } H_1\colon\ \mathbf{p} \neq \mathbf{p}^0$$
with asymptotic level $\alpha \in (0, 1)$.

- Example: If $\mathbf{p}^0 = (1/K, 1/K, \ldots, 1/K)$, we are testing whether $\mathbb{P}_{\mathbf{p}}$ is the uniform distribution on E.

PMF, likelihood and maximum likelihood estimator

- Let $X \in \{a_1, \dots, a_K\}$ have pmf
$$p(a_j) = \mathbb{P}[X = a_j] = p_j, \quad j = 1, \dots, K$$
We can write
$$p(x) = \prod_{j=1}^{K} p_j^{\mathbb{1}(x=a_j)}$$
- Likelihood of the model:
$$L_n(X_1, \dots, X_n, \mathbf{p}) = p_1^{N_1} p_2^{N_2} \dots p_K^{N_K},$$
where $N_j = \#\{i = 1, \dots, n : X_i = a_j\}$.
- Let $\hat{\mathbf{p}}$ be the MLE: $\hat{\mathbf{p}}_j = \dfrac{N_j}{n}, \quad j = 1, \dots, K.$

⚠ $\hat{\mathbf{p}}$ maximizes $\log L_n(X_1, \dots, X_n, \mathbf{p})$ **under the constraint**
$$\sum_{j=1}^{K} p_j = 1.$$

χ^2 test

Theorem

$$\underbrace{n\sum_{j=1}^{K}\frac{\left(\hat{\mathbf{p}}_j-\mathbf{p}_j^0\right)^2}{\mathbf{p}_j^0}}_{T_n}\xrightarrow[n\to\infty]{(d)}\chi^2_{K-1}.$$

- χ^2 test with asymptotic level α: $\quad \psi=\mathbb{1}\{T_n>q_\alpha^{\chi^2_{K-1}}\}$, where $q_\alpha^{\chi^2_{K-1}}$ is the $(1-\alpha)$-quantile of χ^2_{K-1}.
- (Asymptotic) p-value of this test: $p-\text{value}=\mathbb{P}\left[Z>T_n^{\text{obs}}\right]$, where $Z\sim\chi^2_{K-1}$

CDF and empirical CDF

Let $X_1, \ldots, X_n$ be i.i.d. real random variables. Recall the cdf of X_1 is defined as:

$$F(t) = \mathbb{P}[X_1 \le t], \quad \forall t \in \mathbb{R}.$$

It completely characterizes the distribution of X_1.

Definition

The *empirical cdf* of the sample $X_1, \ldots, X_n$ is defined as:

$$\begin{aligned} F_n(t) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le t\} \\ &= \frac{\#\{i = 1, \ldots, n : X_i \le t\}}{n}, \quad \forall t \in \mathbb{R}. \end{aligned}$$

Consistency

By the LLN, for all $t \in \mathbb{R}$,

$$F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t).$$

Glivenko-Cantelli Theorem (*Fundamental theorem of statistics*)

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0.$$

Asymptotic normality

By the CLT, for all $t \in \mathbb{R}$,

$$\sqrt{n}\ \left(F_n(t) - F(t)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\big(0, F(t)\left(1 - F(t)\right)\big).$$

Donsker's Theorem

If F is continuous, then

$$\sqrt{n}\ \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{(d)} \sup_{0\le t\le 1} |\mathbb{B}(t)|,$$

where $\mathbb{B}$ is a Brownian bridge on $[0, 1]$.

Goodness of fit for continuous distributions

- Let $X_1, \ldots, X_n$ be i.i.d. real random variables with unknown cdf F and let F^0 be a **continuous** cdf.
- Consider the two hypotheses:

$$H_0 : F = F^0 \quad \text{v.s.} \quad H_1 : F \neq F^0.$$

- Let F_n be the empirical cdf of the sample $X_1, \ldots, X_n$.
- If $F = F^0$, then $F_n(t) \approx F^0(t)$, for all $t \in [0, 1]$.

Kolmogorov-Smirnov test

- Let $T_n = \sup_{t \in \mathbb{R}} \left| F_n(t) - F^0(t) \right|$.
- By Donsker's theorem, if H_0 is true, then $\sqrt{n} T_n \xrightarrow[n \to \infty]{(d)} Z$, where Z has a known distribution (supremum of a Brownian bridge).
- **KS test with asymptotic level α:**

$$\delta_\alpha^{KS} = \mathbb{1}\{T_n > q_\alpha / \sqrt{n}\},$$

where q_α is the $(1-\alpha)$-quantile of Z (obtained in tables).

- p-value of KS test: $\mathbb{P}[Z > T_n | T_n]$.

Computational issues

- In practice, how to compute T_n ?
- F^0 is non decreasing, F_n is piecewise constant, with jumps at $t_i = X_i, i = 1, \ldots, n$.
- Let $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ be the reordered sample.
- The expression for T_n reduces to the following practical formula:

$$T_n = \max_{i=1,\ldots,n} \left\{ \max \left(\left| \frac{i-1}{n} - F^0(X_{(i)}) \right|, \left| \frac{i}{n} - F^0(X_{(i)}) \right| \right) \right\}.$$

Pivotal distribution

- T_n is called a *pivotal statistic*: If H_0 is true, the distribution of T_n does not depend on the distribution of the X_i's and it is easy to reproduce it in simulations.

- Indeed, let $U_i = F^0(X_i), i = 1, \ldots, n$ and let G_n be the empirical cdf of $U_1, \ldots, U_n$.

- If H_0 is true, then $U_1, \ldots, U_n \overset{i.i.d.}{\sim} \mathcal{U}\left([0.1]\right)$

  and $T_n = \sup\limits_{0 \le x \le 1} |G_n(x) - x|$.

Quantiles and p-values

- For some large integer M:
  - Simulate M i.i.d. copies $T_n^1, \ldots, T_n^M$ of T_n;
  - Estimate the $(1-\alpha)$-quantile $q_\alpha^{(n)}$ of T_n by taking the sample $(1-\alpha)$-quantile $\hat{q}_\alpha^{(n,M)}$ of $T_n^1, \ldots, T_n^M$.
- Test with approximate level α:

$$\delta_\alpha = \mathbb{1}\{T_n > \hat{q}_\alpha^{(n,M)}/\sqrt{n}\}.$$

- Approximate p-value of this test:

$$\text{p-value} \approx \frac{\#\{j = 1, \ldots, M : T_n^j > T_n\}}{M}.$$

Kolmogorov–Smirnov Tables

Critical values, $d_{alpha};(n)^a$, of the maximum absolute difference between sample $F_n(x)$ and population $F(x)$ cumulative distribution.

Number of trials, n	Level of significance, α			
	0.10	0.05	0.02	0.01
1	0.95000	0.97500	0.99000	0.99500
2	0.77639	0.84189	0.90000	0.92929
3	0.63604	0.70760	0.78456	0.82900
4	0.56522	0.62394	0.68887	0.73424
5	0.50945	0.56328	0.62718	0.66853
6	0.46799	0.51926	0.57741	0.61661
7	0.43607	0.48342	0.53844	0.57581
8	0.40962	0.45427	0.50654	0.54179
9	0.38746	0.43001	0.47960	0.51332
10	0.36866	0.40925	0.45662	0.48893

Other goodness of fit tests

We want to measure the distance between two functions: $F_n(t)$ and $F(t)$. There are other ways, leading to other tests:

- Kolmogorov-Smirnov:

$$d(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

- Cramér-Von Mises:

$$d^2(F_n, F) = \int_{\mathbb{R}} [F_n(t) - F(t)]^2 \, dF(t)$$

- Anderson-Darling:

$$d^2(F_n, F) = \int_{\mathbb{R}} \frac{[F_n(t) - F(t)]^2}{F(t)(1 - F(t))} dF(t)$$

Composite goodness of fit tests

What if I want to test: "Does X have Gaussian distribution?" but I don't know the parameters?
Simple idea: plug-in

$$\sup_{t \in \mathbb{R}} \left| F_n(t) - \Phi_{\hat{\mu}, \hat{\sigma}^2}(t) \right|$$

where

$$\hat{\mu} = \bar{X}_n, \qquad \hat{\sigma}^2 = S_n^2$$

and $\Phi_{\hat{\mu}, \hat{\sigma}^2}(t)$ is the cdf of $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$.

In this case Donsker's theorem is *no longer valid*. This is a common and serious mistake!

Kolmogorov-Lilliefors test (1)

Instead, we compute the quantiles for the test statistic:

$$\sup_{t \in \mathbb{R}} \left| F_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t) \right|$$

They do not depend on unknown parameters!

This is the Kolmogorov-Lilliefors test.

K-L table

Sample Size N	Level of Significance for $D = \text{Max}\,\lvert F^*(X) - S_N(X)\rvert$				
	.20	.15	.10	.05	.01
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231

Quantile-Quantile (QQ) plots (1)

- Provide a visual way to perform GoF tests
- Not formal test but quick and easy check to see if a distribution is plausible.
- Main idea: we want to check visually if the plot of F_n is close to that of F or equivalently if the plot of F_n^{-1} is close to that of F^{-1}.
- More convenient to check if the points

$$\big(F^{-1}(\frac{1}{n}), F_n^{-1}(\frac{1}{n})\big), \big(F^{-1}(\frac{2}{n}), F_n^{-1}(\frac{2}{n})\big), \dots, \big(F^{-1}(\frac{n-1}{n}), F_n^{-1}(\frac{n-1}{n})\big)$$

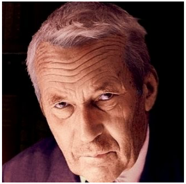
  are near the line $y = x$.
- F_n is not technically invertible but we define

$$F_n^{-1}(i/n) = X_{(i)},$$

  the ith largest observation.

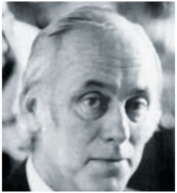
24 A

24 B



24 C

24 D



24 E

Quantile-Quantile (QQ) plots (2)

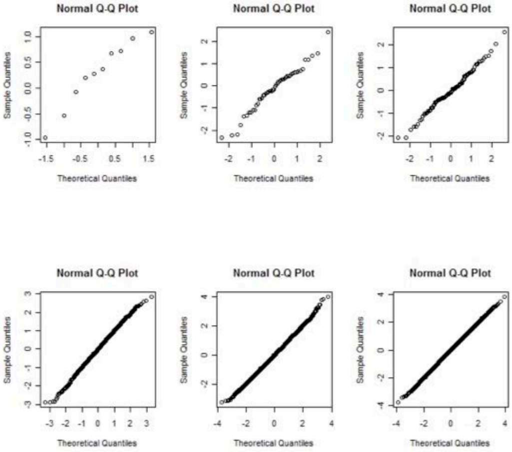


Figure 1: QQ-plots for samples of sizes 10, 50, 100, 1000, 5000, 10000 from a standard normal distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

Quantile-Quantile (QQ) plots (3)

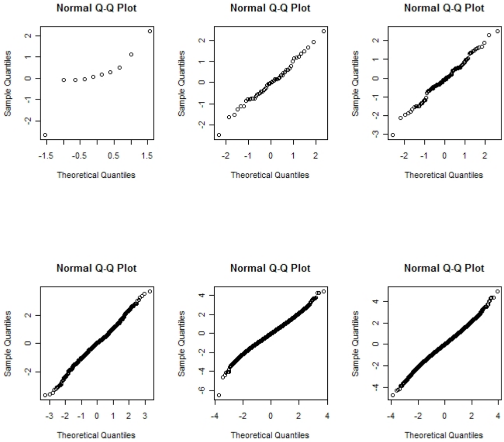


Figure 2: QQ-plots for samples of sizes $10, 50, 100, 1000, 5000, 10000$ from a t_{15} distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

ADDED ATTRIBUTION

Slide #24A

Image from the AIP Emilio Segrè Visual Archives, Von Mises Collection. © 2019 American Institute of Physics

Slide # 24B
https://commons.wikimedia.org/wiki/File:Abraham_Wald.jpg
Citation/Attribution -- Image on Wikimedia by Konrad Jacobs, Erlangen.
Copyright is MFO. (CC BY-SA) 2.0

Slide #24C
https://www.sapaviva.com/andrey-kolmogorov/
Image from Sapaviva.com
© 2017 Valentine O. Oduenyi

Slide # 24D
https://www.causeweb.org/cause/resources/fun/quotes/wellswilks-statistical-thinking
Citation/Attribution -- Image © 2019 Consortium for the Advancement of Undergraduate Statistics Education. (CC BY NC SA) 4.0

Slide # 524E
https://www.statmania.info/2017/09/Harald-cramer.html
Citation/Attribution -- Image (c) 2016 Stat Mania