18.650 – Fundamentals of Statistics

4. Hypothesis testing

Goals

We have seen the basic notions of hypothesis testing:

- Hypotheses H_0/H_1,
- Type 1/Type 2 error, level and power
- Test statistics and rejection region
- p-value

Our tests were based on CLT (and sometimes Slutsky)...

- What if data is Gaussian, σ^2 is unknown and Slutsky does not apply?
- Can we use asymptotic normality of MLE?
- Tests about multivariate parameters $\theta = (\theta_1, \ldots, \theta_d)$ (e.g.: $\theta_1 = \theta_2$)?
- More complex tests: "Does my data follow a Gaussian distribution"?

Parametric hypothesis testing

Clinical trials

Let us go through an example to remind the main notions of hypothesis testing.

- Pharmaceutical companies use hypothesis testing to test if a new drug is efficient.
- To do so, they administer a drug to a group of patients (test group) and a placebo to another group (control group).
- We consider testing a drug that is supposed to lower LDL (low-density lipoprotein), a.k.a "bad cholesterol" among patients with a high level of LDL (above 200 mg/dL)

Notation and modelling

- Let $\Delta_{\mathrm{d}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the drug.
- Let $\Delta_{\mathrm{c}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the placebo.
- We want to know if
- We observe two independent samples:
  - $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\quad, \sigma_{\mathrm{d}}^2)$ from the test group and
  - $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\quad, \sigma_{\mathrm{c}}^2)$ from the group.

Hypothesis testing

- Hypotheses:

$$H_0 : \qquad \text{vs.} \quad H_1 :$$

- Since the data is Gaussian by assumption we don't need the

- We have

$$\bar{X}_n \sim \qquad \text{and} \qquad \bar{Y}_m \sim$$

- Therefore

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_{\mathrm{d}} - \Delta_{\mathrm{c}})}{} \sim \mathcal{N}(0,1)$$

Asymptotic test

- Assume that $m = cn$ and $n \to \infty$
- Using lemma, we also have

$$\xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

where

$$\hat{\sigma}_{\mathrm{d}}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{and} \quad \hat{\sigma}_{\mathrm{c}}^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2$$

- We get the the following test at asymptotic level α:

$$R_\psi = \left\{ \qquad\qquad \right\}$$

- This is -sided, -sample test.

Asymptotic test

- Example $n = 70, m = 50$, $\bar{X}_n = 156.4$, $\bar{Y}_m = 132.7$, $\hat{\sigma}_{\mathrm{d}}^2 = 5198.4$, $\hat{\sigma}_{\mathrm{c}}^2 = 3867.0$,

$$\frac{156.4 - 132.7}{\sqrt{\frac{5198.4}{70} + \frac{3867.0}{50}}} = 1.57$$

  Since $q_{5\%} = 1.645$, we

- We can also compute the p-value:

$$\text{p-value} = \qquad\qquad\qquad\qquad = 0.0582$$


PHARMACOLOGICAL
DRUG TRIAL RESULTS
OUR TRIALS SHOW THAT
THE NEW DRUG PERFORMS
NO BETTER THAN PLACEBO
MAYBE WE SHOULD
INVEST IN PLACEBOS
CHRIS
MADDEN

Small sample size

- What if $n = 20, m = 12$?
- We cannot realistically apply Slutsky's lemma
- We needed it to find the (asymptotic) distribution of quantities of the form
$$\frac{\bar{X}_n - \mu}{\hat{\sigma}^2}$$
when $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.
- It turns out that this distribution *does not depend on* μ *or* σ so we can compute its

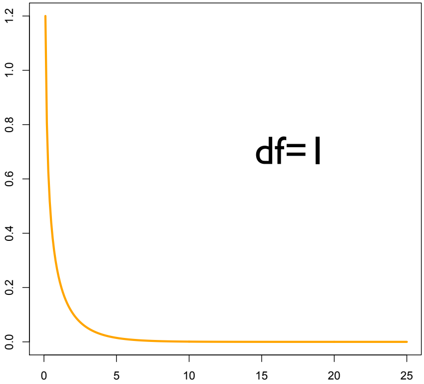
The χ^2 distribution

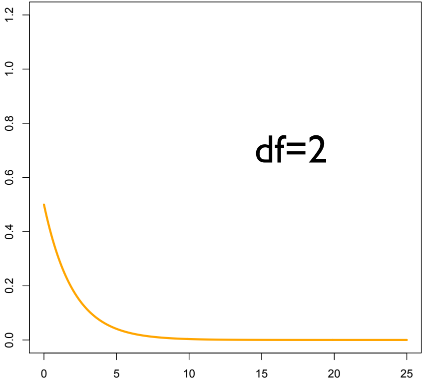
Definition

For a positive integer d, the χ^2 *(pronounced "Kai-squared") distribution with d degrees of freedom* is the law of the random variable $Z_1^2 + Z_2^2 + \ldots + Z_d^2$, where $Z_1, \ldots, Z_d \stackrel{iid}{\sim} \mathcal{N}(0, 1)$.

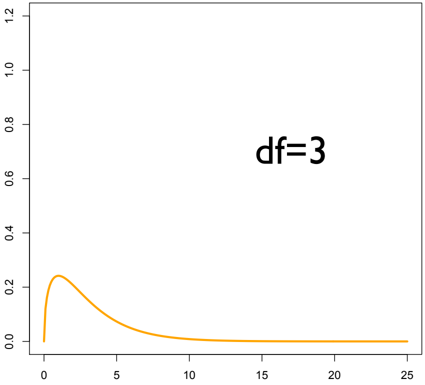
Examples:

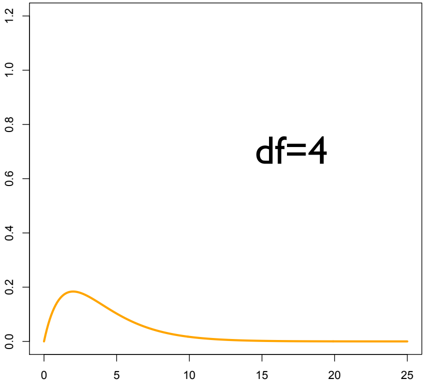
- If $Z \sim \mathcal{N}_d(\mathbf{0}, I_d)$, then $\|Z\|_2^2 \sim$
- $\chi_2^2 = \mathsf{Exp}(1/2)$.

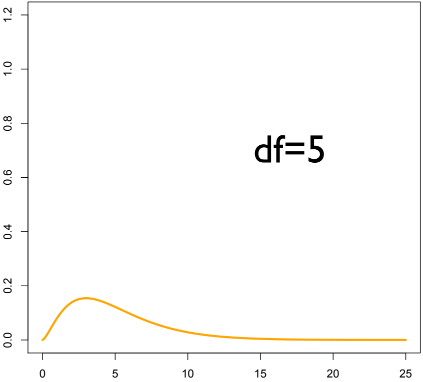

df=1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25



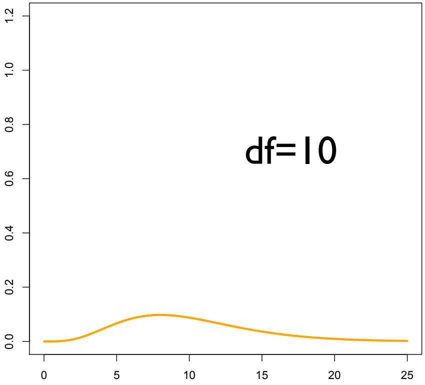
df=2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25

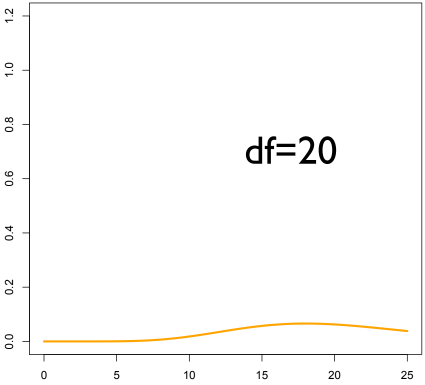

df=3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25


df=4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25



df=5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25


df=10
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25


df=20
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25

Properties χ^2 distribution (2)

Definition
For a positive integer d, the χ^2 *(pronounced "Kai-squared") distribution with d degrees of freedom* is the law of the random variable $Z_1^2 + Z_2^2 + \ldots + Z_d^2$, where $Z_1, \ldots, Z_d \overset{iid}{\sim} \mathcal{N}(0,1)$.

Properties: If $V \sim \chi_k^2$, then

- $\mathbb{E}[V] =$
- $\mathrm{var}[V] =$

Important example: the sample variance

- Recall that the sample variance is given by
  $$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2$$
- Cochran's theorem states that for $X_1, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, if S_n is the sample variance, then
  - $\bar{X}_n \perp\!\!\!\perp S_n$;
  - $\frac{nS_n}{\sigma^2} \sim \chi^2_{n-1}$.
  - We often prefer the unbiased estimator of σ^2:

Student's T distribution

Definition

For a positive integer d, the *Student's T distribution with d degrees of freedom* (denoted by t_d) is the law of the random variable $\dfrac{Z}{\sqrt{V/d}}$, where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_d$ and $Z \perp\!\!\!\perp V$ (Z is independent of V).

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information

Who was Student?

This distribution was introduced by **William Sealy Gosset** (1876–1937) in 1908 while he worked for the Guinness brewery in Dublin, Ireland.

Student's T test (one sample, two-sided)

- Let $X_1, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown
- We want to test:

$$H_0 : \mu = 0, \quad \text{vs} \quad H_1 : \mu \neq 0$$

- Test statistic:

$$T_n = \frac{\sqrt{n}\bar{X}_n}{\sqrt{\tilde{S}_n}} =$$

- Since $\sqrt{n}\bar{X}_n/\sigma \sim$ (under) and $\tilde{S}_n/\sigma^2 \sim$ are independent by theorem, we have:

$$T_n \sim$$

- Student's test with (non asymptotic) level $\alpha \in (0, 1)$:

$$\psi_\alpha = \mathbb{1}\{|T_n| > q_{\alpha/2}\},$$

where $q_{\alpha/2}$ is the $(1 - \alpha/2)$-quantile of t_{n-1}.

Student's T test (one sample, one-sided)

- We want to test:

$$H_0 : \mu \le \mu_0, \quad \text{vs} \quad H_1 : \mu > \mu_0$$

- Test statistic:

$$T_n = \frac{\sqrt{n}(\bar{X}_n \qquad)}{\sqrt{\tilde{S}_n}} \sim$$

under H_0.

- Student's test with (non asymptotic) level $\alpha \in (0, 1)$:

$$\psi_\alpha = \mathbb{1}\Big\{ \qquad\qquad \Big\},$$

Two-sample T-test

- Back to our cholesterol example. What happens for small sample sizes?
- We want to know the distribution of

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_{\mathrm{d}} - \Delta_{\mathrm{c}})}{\sqrt{\frac{\hat{\sigma}_{\mathrm{d}}^2}{n} + \frac{\hat{\sigma}_{\mathrm{c}}^2}{m}}}$$

- We have approximately

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_{\mathrm{d}} - \Delta_{\mathrm{c}})}{\sqrt{\frac{\hat{\sigma}_{\mathrm{d}}^2}{n} + \frac{\hat{\sigma}_{\mathrm{c}}^2}{m}}} \sim t_N$$

where

$$N = \frac{\left(\hat{\sigma}_{\mathrm{d}}^2/n + \hat{\sigma}_{\mathrm{c}}^2/m\right)^2}{\frac{\hat{\sigma}_{\mathrm{d}}^4}{n^2(n-1)} + \frac{\hat{\sigma}_{\mathrm{c}}^4}{m^2(m-1)}} \geq \min(n, m)$$

(Welch-Satterthwaite formula)

Non-asymptotic test

- Example $n = 70, m = 50$, $\bar{X}_n = 156.4$, $\bar{Y}_m = 132.7$, $\hat{\sigma}_{\mathrm{d}}^2 = 5198.4$, $\hat{\sigma}_{\mathrm{c}}^2 = 3867.0$,

$$\frac{156.4 - 132.7}{\sqrt{\frac{5198.4}{70} + \frac{3867.0}{50}}} = 1.57$$

- Using the shorthand formula $N = \min(n, m) =$, we get $q_{5\%} = 1.68$ and

$$\text{p-value} = \qquad\qquad = 0.0614$$

- Using the W-S formula

$$N = \frac{\left(\frac{5198.4}{70} + \frac{3867.0}{50}\right)^2}{\frac{5198.4^2}{70^2(70-1)} + \frac{3867.0^2}{50^2(50-1)}} = 113.78$$

we round to .

- We get

$$\text{p-value} = \qquad\qquad = 0.0596$$

Discussion

Advantage of Student's test: Non asymptotic / Can be run on small samples

Drawback of Student's test: It relies on the assumption that the sample is Gaussian (soon we will see how to test this assumption)

A test based on the MLE

- Consider an i.i.d. sample $X_1, \ldots, X_n$ with statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$, where $\Theta \subseteq \mathbb{R}^d$ $(d \geq 1)$ and let $\theta_0 \in \Theta$ be fixed and given.
- Consider the following hypotheses:

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_0. \end{cases}$$

- Let $\hat{\theta}^{MLE}$ be the MLE. Assume the MLE technical conditions are satisfied.

- If H_0 is true, then

$$\times \left(\hat{\theta}_n^{MLE} - \theta_0\right) \xrightarrow[n \to \infty]{(d)} \mathcal{N}_d\left(0, I_d\right)$$

Wald's test

- Hence,

$$\underbrace{n\left(\hat{\theta}_n^{MLE}-\theta_0\right)^\top I(\hat{\theta}^{MLE})\left(\hat{\theta}_n^{MLE}-\theta_0\right)}_{T_n} \xrightarrow[n\to\infty]{(d)}$$

- Wald's test with asymptotic level $\alpha \in (0,1)$:

$$\psi = \mathbb{1}\{T_n > q_\alpha\},$$

where q_α is the $(1-\alpha)$-quantile of χ^2_d (see tables).

- Remark: Wald's test is also valid if H_1 has the form "$\theta > \theta_0$" or "$\theta < \theta_0$" or "$\theta = \theta_1$"...

A test based on the log-likelihood

- Consider an i.i.d. sample $X_1, \dots, X_n$ with statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$, where $\Theta \subseteq \mathbb{R}^d$ ($d \geq 1$).

- Suppose the null hypothesis has the form

$$H_0 : (\theta_{r+1}, \dots, \theta_d) = (\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)}),$$

  for some fixed and given numbers $\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)}$.

- Let

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell_n(\theta) \quad \text{(MLE)}$$

  and

$$\hat{\theta}_n^c = \underset{\theta \in \Theta_0}{\operatorname{argmax}}\ \ell_n(\theta) \quad \text{(“constrained MLE”)}$$

  where $\Theta_0 =$

Likelihood ratio test

Test statistic:

$$T_n = 2\left(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n^c)\right).$$

Wilks' Theorem

Assume H_0 is true and the MLE technical conditions are satisfied. Then,

$$T_n \xrightarrow[n\to\infty]{(d)}$$

Likelihood ratio test with asymptotic level $\alpha \in (0,1)$:

$$\psi = \mathbb{1}\{T_n > q_\alpha\},$$

where q_α is the $(1-\alpha)$-quantile of χ^2_{d-r} (see tables).

Implicit hypotheses

- Let $X_1, \ldots, X_n$ be i.i.d. random variables and let $\theta \in \mathbb{R}^d$ be a parameter associated with the distribution of X_1 (e.g. a moment, the parameter of a statistical model, etc...)

- Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be continuously differentiable (with $k < d$).

- Consider the following hypotheses:

$$\begin{cases} H_0 : & g(\theta) = 0 \\ H_1 : & g(\theta) \neq 0. \end{cases}$$

- E.g. $g(\theta) = (\theta_1, \theta_2)$ $(k = 2)$, or $g(\theta) =$ $(k = 1)$, or...

Delta method

- Suppose an asymptotically normal estimator $\hat{\theta}_n$ is available:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma(\theta)).$$

- Delta method:

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k\left(0, \Gamma(\theta)\right),$$

  where $\Gamma(\theta) = \nabla g(\theta)^\top \Sigma(\theta) \nabla g(\theta) \in \mathbb{R}^{k\times k}$.

- Assume $\Sigma(\theta)$ is invertible and $\nabla g(\theta)$ has rank k. So, $\Gamma(\theta)$ is invertible and

$$\sqrt{n}\,\Gamma(\theta)^{-1/2}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k\left(0, \quad\right).$$

Wald's test for implicit hypotheses

- Then, by Slutsky's theorem, if $\Gamma(\theta)$ is continuous in θ,

$$\sqrt{n}\,\Gamma(\quad)^{-1/2}\left(g(\hat{\theta}_n)-g(\theta)\right)\xrightarrow[n\to\infty]{(d)}\mathcal{N}_k\left(0,I_k\right).$$

- Hence, if H_0 is true, i.e., $g(\theta)=0$,

$$\underbrace{ng(\hat{\theta}_n)^\top\Gamma^{-1}(\hat{\theta}_n)g(\hat{\theta}_n)}_{T_n}\xrightarrow[n\to\infty]{(d)}\chi^2_k.$$

- Test with asymptotic level α:

$$\psi=\mathbb{1}\{\qquad\qquad\},$$

where q_α is the $(1-\alpha)$-quantile of χ^2_k (see tables).

Goodness of fit

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:

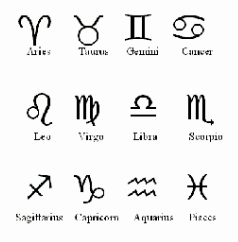
- Does X have distribution $\mathcal{N}(0,1)$? (Cf. Student's T distribution)
- Does X have distribution $\mathcal{U}([0,1])$?
- Does X have PMF $p_1 = 0.3,\ p_2 = 0.5,\ p_3 = 0.2$

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 $X \sim \mathbb{P}_{\mathbf{p}}$,

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

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 on E.

Multinomial likelihood

- Likelihood of the model:

$$L_n(X_1, \ldots, X_n, \mathbf{p}) = p_1^{N_1} p_2^{N_2} \ldots p_K^{N_K},$$

where $N_j = \#\{i = 1, \ldots, n : X_i = a_j\}$.

- Let $\hat{\mathbf{p}}$ be the MLE:

$$\hat{\mathbf{p}}_j = \frac{N_j}{n}, \qquad j = 1, \ldots, K.$$

⚠ $\hat{\mathbf{p}}$ maximizes $\log L_n(X_1, \ldots, X_n, \mathbf{p})$ **under the constraint**

χ^2 test

- If H_0 is true, then $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}^0)$ is asymptotically normal, and the following holds.

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: $\psi_\alpha = \mathbb{1}\{T_n > q_\alpha\}$, where q_α is the $(1-\alpha)$-quantile of χ^2_{K-1}.
- Asymptotic p-value of this test: $p -$ value $= \mathbb{P}\left[Z > T_n | T_n\right]$, where $Z \sim \chi^2_{K-1}$ and $Z \perp\!\!\!\perp T_n$.

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:

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

It completely characterizes the distribution of X_1.

Definition

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

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

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 $t \in \mathbb{R}$,

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

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 $T_n = \sup_{t \in \mathbb{R}} \sqrt{n} \left| F_n(t) - F^0(t) \right|$.

- By Donsker's theorem, if H_0 is true, then $T_n \xrightarrow[n \to \infty]{(d)} Z$, where Z has a known distribution (supremum of a Brownian bridge).

- **KS test with asymptotic level α:**

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

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:

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

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 $U_1, \ldots, U_n \overset{i.i.d.}{\sim}$

  and $T_n = \sup_{0 \leq x \leq 1} \sqrt{n}\,|G_n(x) - x|$.

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:

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

- 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 $D = \text{Max} \left| F^*(X) - S_N(X) \right|$				
	.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

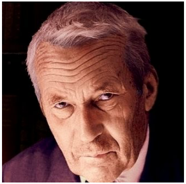
$$\big(F^{-1}(\frac{1}{n}), F_n^{-1}(\frac{1}{n})\big), \big(F^{-1}(\frac{2}{n}), F_n^{-1}(\frac{2}{n})\big), \dots, \big(F^{-1}(\frac{n-1}{n}), F_n^{-1}(\frac{n-1}{n})\big)$$

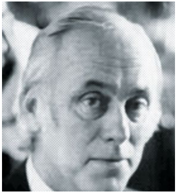
  are near the line $y = x$.
- F_n is not technically invertible but we define

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

  the ith largest observation.



Quantile-Quantile (QQ) plots (2)

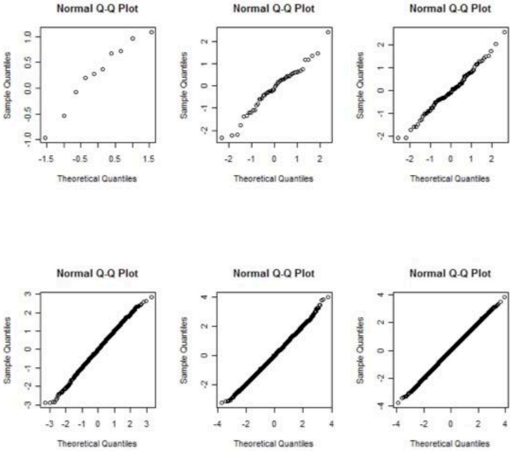


Figure 1: QQ-plots for samples of sizes $10, 50, 100, 1000, 5000, 10000$ from a standard normal distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

Quantile-Quantile (QQ) plots (3)

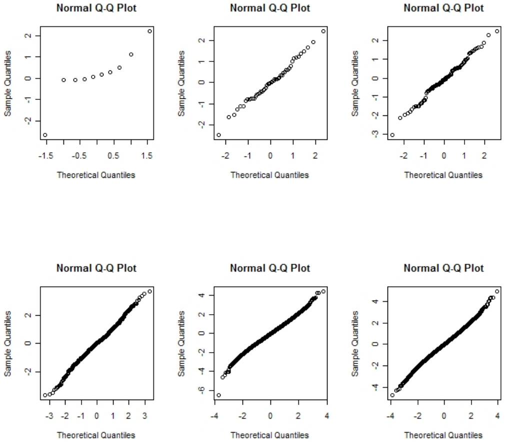


Figure 2: QQ-plots for samples of sizes $10, 50, 100, 1000, 5000, 10000$ from a t_{15} distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.