Distribution-Free Tests ($\simeq$ Chapter 14)

All the methods we have seen so far rely on assumptions on the true distribution of data. Namely, we always assumed that $X_1, \dots, X_n \sim f_\theta$, for some $\theta \in \mathbb{R}^d$ (unknown) and a known form of f_θ

Ex: T-tests for means of continuous data assume $X_i \sim N(\mu, \sigma^2)$ (here $\theta = (\mu, \sigma^2)$)

Such an assumption is especially important to derive the distribution of the test statistic under the null, and hence design the testing procedure.

Parametric methods: Assume the data arise from a distribution described by a few parameters (think of $\theta = (\mu, \sigma^2)$ for Gaussians)

Nonparametric methods: Do not make parametric assumptions (most often based on ranks as opposed to raw values)

$\longrightarrow$ In this chapter, we discuss non-parametric alternatives to the one- and two-sample T-tests.

Examples (of when the parametric T-test goes wrong)

Extreme Outliers

Test 1: T-test comparing the two datasets

$$X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$
$$Y = \{7, 8, \ldots, 20\}$$

$\bar{X} = 5.5$, $\bar{Y} = 13.5$, p-value $= 1.9 \cdot 10^{-5}$

Test 2

$$X = \{1, 2, \ldots, 10\}$$
$$Y' = \{7, 8, \ldots, 20, \underline{200}\}$$

Outlier

$\bar{X} = 5.5$, $\bar{Y} = 25.9$, p-value $= 0.12$

Why is this happening? Technically speaking, because $S_{Y'}^2 = 2335$ (while $S_X^2 = 9.2$, $S_Y^2 = 17.5$)

Skew: Similarly, one can make up examples showing the distributional skew has T-tests fail.

Can you make your own?

Rk: (When to use nonparametric methods)

- With correct assumptions (e.g. normal distribution), parametric methods will be more efficient (= powerful) than nonparametric ones, but often not as much as you might think.
- If the normality assumption is grossly violated, nonparametric tests can be much more efficient and powerful than the parametric one.
- Nonparametric methods provide a well-founded way to deal with circumstances in which parametric methods perform poorly.

14.2 The sign test

Idea: (1) T-tests are about means, and fail with outliers or skewed data

(2) The median is a much better measure of centrality, when describing data which is skewed, or with outliers

(3) let's focus on tests about medians in such cases!

Recall: the median of a real random variable is the value $\tilde{\mu}_X$ such that $\mathbb{P}(X \leq \tilde{\mu}_X) = \mathbb{P}(X > \tilde{\mu}_X) = \frac{1}{2}$

Thm: (Sign Test)

Let $X_1, \dots, X_n \overset{iid}{\sim} f_X$ be a sample from a continuous distribution with density f_X. If $n \geq 10$ and $\bar{\mu}_X = \bar{\mu}_0$, then

$$T = \frac{K - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \overset{\text{Approx.}}{\sim} N(0,1),$$

where $K = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \geq \bar{\mu}_0\}}$ is the number of sample points that are bigger or equal to $\bar{\mu}_0$.

Proof: Let $Y_i = \mathbb{1}_{X_i \geq \bar{\mu}_0}$. If $\bar{\mu}_X = \bar{\mu}_0$, then $\sim \text{Binomial}(n, p)$, where $p = \mathbb{P}(X_i \geq \bar{\mu}_0) = \frac{1}{2}$. Conclude using the CLT.

Rk: • As a corollary, we can test $H_0 : \bar{\mu}_X = \bar{\mu}_0$ against

$H_1 : \bar{\mu}_X > \bar{\mu}_0$, with rejection region $\{T \geq z_\alpha\}$

$H_1 : \bar{\mu}_X < \bar{\mu}_0$ 〃 $\{T \leq -z_\alpha\}$

$H_1 : \bar{\mu}_X \neq \bar{\mu}_0$ 〃 $\{|T| \geq z_{\alpha/2}\}$

Question: when are we able to use this test for $H_0 : \mu_X = \mu_0$?

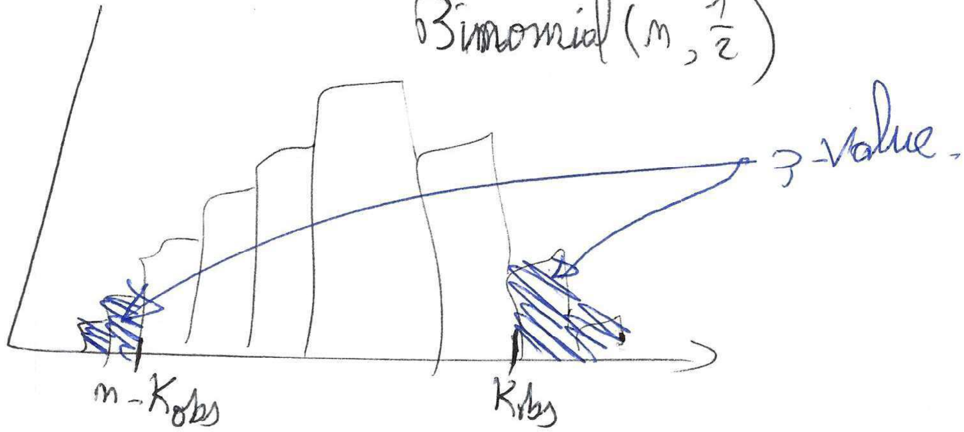
$\longrightarrow$ When f_X is symmetric about μ_X. Indeed, in such a situation, $\mu_X = \tilde{\mu}_X$. (Prove this!)

Rk: What to do if $n < 10$?

The best way to proceed is to derive p-values directly on K with an exact test. Indeed, under H_0, $K \sim \text{Binomial}(n, \frac{1}{2})$, so we can compute the exact p-value.

For instance, for $H_1 : \tilde{\mu}_X \neq \mu_0$, we get

$$p\text{-value} = \mathbb{P}_{H_0}\left(K \geq K_{observed} \text{ or } K \leq n - K_{observed} \right)$$



⑥ The sign test can also be used for paired data, by using the usual "difference" trick.

<u>Thm</u>: let $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ be two paired samples.

If $\tilde{\mu}_{X-Y} = 0$, then

$$K \begin{cases} \simeq N\left(\frac{n}{2}, \frac{n}{4}\right) & \text{if } n \to \infty \\ \sim \text{Binomial}\left(n, \frac{1}{2}\right) & \text{if } n \leqslant 10 \end{cases},$$

where $K = \sum \mathbb{1}_{X_i - Y_i \geqslant 0}$ is the number of sample differences that are bigger or equal to 0.

<u>Proof</u>: Apply the previous theorem to $\Delta_i = X_i - Y_i$.

A general remark about testing quantiles

Going from parametrics to nonparametrics, we basically have replaced the parameter of interest that describes a notion of centrality of the unknown distribution.

Namely, we have replaced $\mu_X = \mathbb{E}(X)$ by $\tilde{\mu}_X = \text{Median}(X)$, so that we test H_0: $\text{Median}(X) = \tilde{\mu}_0$, which translates to

$$H_0 : \mathbb{P}(X \leq \tilde{\mu}_0) = \frac{1}{2}.$$

This suggests to generalise this idea to any other quantile! Say you want to test

$$H_0 : \mathbb{P}(X \leq \tilde{\mu}_0) = p_0$$

= "The pth quantile of X is equal to p".

What would the test statistic be?

$\rightarrow$ Still taking $K = \sum_{i=1}^{n} \mathbb{1}_{\{X_i > \tilde{\mu}_0\}}$, we know that under the null,

$$K \sim \text{Binomial}(n, p_0) \approx N(np_0, np_0(1-p_0)).$$

when $np_0, n(1-p_0) \geq 5$

As a consequence, one should consider

$$T = \frac{K - np_0}{\sqrt{np_0(1-p_0)}},$$

and proceed as for the median (case $p = \frac{1}{2}$)

<u>Supplementary Question</u>: How do we do if $np_0 < 5$?

Two possibilities: 1) If n is small, use the exact binomial model (and p-value)

2) If n is large, use the Poisson approximation

$$\text{Binomial}(n, p_0) \simeq \text{Poisson}(np_0)$$

⑨

14.3 Wilcoxon Tests

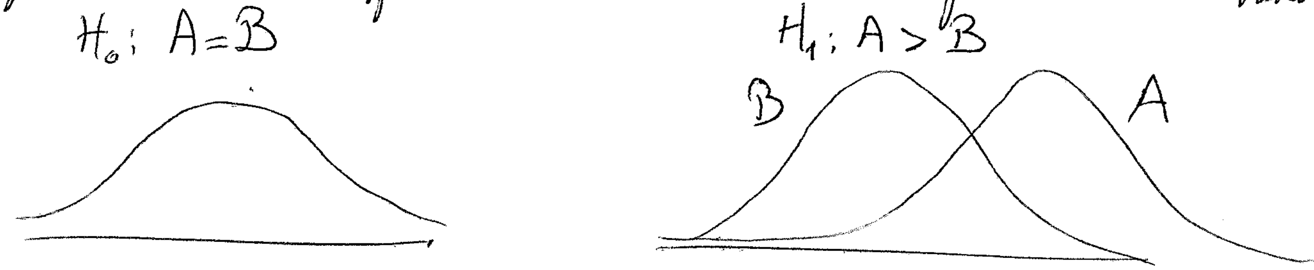
We now move to a more elaborate family of methods based on ranks, but that require a bit more of hypotheses.

<u>Ex</u>: . Suppose we have m and n experimental units $\begin{cases} \cdot\ n \text{ control} \\ \cdot\ m \text{ treatment} \end{cases}$

. The assignments is made at random

. We're interested in testing if the treatment does has positive effect or not.

Very often, a positive effect would translate in <u>a shift</u> on the distribution,



without a change of shape, whatever the shape

$\longrightarrow$ let's use the shape shift phenomenon in a test.

<u>Idea</u>:
- Group all m and n observations together and rank them in order of increasing size
- Calculate the sum of ranks of the control group
- If this sum is too large, reject the null.

Thm (Wilcoxon Signed Ranks Test)

Let $(X_1, \cdots, X_n)$ be independent observations drawn from p.d.f's $f_{X_1}, \cdots, f_{X_n}$ all of which are continuous and symmetric with equal mean μ.

Then for testing

$$H_0 : \mu = \mu_0 \quad \text{VS} \quad H_1 : \mu \neq \mu_0 \,,$$

we use

$$T = \sum_{i=1}^{n} R_i Z_i \quad \begin{cases} \approx N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right) & \text{if } n \to \infty \\ \sim P(T=t) = \left(\frac{1}{2}\right)^n c(t) & \text{if } n \leq 12 \end{cases}$$

where the rank of $|X - \mu_0|$, and

$$Z_i = \begin{cases} 1 & \text{if } X_i - \mu_0 < 0 \\ 0 & \text{if } X_i - \mu_0 \geq 0 \end{cases}$$

and

$c(t)$ is the coefficient of e^{tx} in the expansion of $\prod_{i=1}^{n} (1 + e^{ix})$

11

<u>Rk</u>: - Again, we did not need to assume a specific family of distributions
- The X_i's don't need to be identically distributed
- We have both asymptotic <u>and</u> nonasymptotic

<u>Proof</u>: If H_0 is true, since $\bar{\mu}_{X_i} = \mu_{X_i}$ because of symmetry, T has the same distribution as

$$U = \sum_{i=1}^{n} U_i \, , \text{ where } \begin{cases} \mathbb{P}(U_i = 0) = \frac{1}{2} \\ \mathbb{P}(U_i = 1) = \frac{1}{2} \end{cases} , \; U_i\text{'s } \perp\!\!\!\perp$$

Hence, we can do the reasoning on U. To describe its distribution, we compute its moment generating function

$$M_U(t) = \mathbb{E}\left(e^{(\sum^{n} u_i) t}\right)$$

$$= \prod_{i=1}^{n} \mathbb{E}(e^{U_i t})$$

$$= \prod_{i=1}^{n} \left(\frac{1}{2} e^{0t} + \frac{1}{2} e^{1 \times t} \right)$$

$$= \left(\frac{1}{2}\right)^n \prod_{i=1}^{n} (1 + e^{it})$$

We conclude by noticing that since U is discrete,

$$M_U(t) = \mathbb{E}\left(e^{Ut}\right)$$
$$= \sum_{x \in \mathbb{N}} e^{tx} \mathbb{P}(U = x),$$

and by developping the previous product]

Rk: · The asymptotic normality is obtained using a version of the CLT (on the U_i's) that is out of the scope of this course. If interested, see "CLT for triangular arrays".

The following result is especially designed for two-sample problems, as opposed to the previous one

Thm (Wilcoxon Ranks Sum Test / Mann-Whitney)

Let $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_m)$ be two independent samples from p.d.f.'s f_X and f_Y which are continuous, symmetric, of the same shape and the same standard deviation. ($\equiv$ one might only be the shifted version of the other)

If $\mu_X = \mu_Y$,

$$T = \sum_{i=1}^{n+m} R_i Z_i \approx N\left(\frac{nm}{2}, \frac{nm(n+m+1)}{12}\right)$$

if $n, m \geq 10$,

where • R_i is the rank of the i^{th} observation in the join sample $(X_1, \dots, X_n, Y_1, \dots, Y_m)$

• $Z_i = \begin{cases} 1 & \text{if the } i^{th} \text{ observation comes from } f_X \\ 0 & \text{—— } f_Y \end{cases}$

Proof: Similar