

STA 104 Applied Nonparametric Statistics

Chapter 3: Two-Sample Methods

Xiner Zhou

Department of Statistics, University of California, Davis

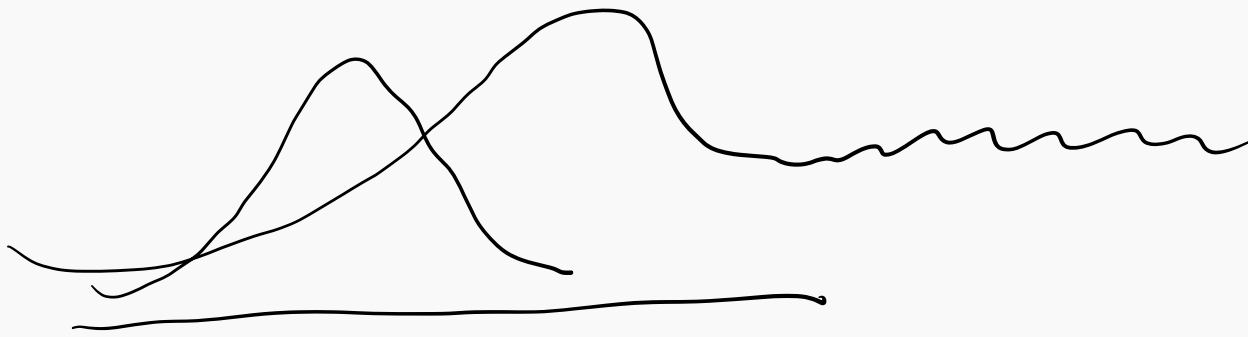
In this chapter the data consist of two random samples, a sample from the control population and an independent sample from the treatment population.

On the basis of these samples, we wish to investigate the presence of a treatment effect that results in a shift of location.

Table of contents

1. Two-Sample Permutation Test
2. Wilcoxon Rank-Sum Test
3. Why Ranks? Scoring Systems
4. Tests for Equality of Scale Parameters
5. An Omnibus Test for general differences in two populations (Kolmogorov-Smirnov test)

**An Omnibus Test for general
differences in two populations
(Kolmogorov-Smirnov test)**



Suppose it is not known how a difference between two ~~treatments~~ ^{populations} might manifest itself in the data.

It might cause observations in one treatment to be larger than observations in the other, or it might affect the variability of the observations, or it might affect the shapes of the distributions in some other way.

What we would like is an omnibus test -that is, a test designed to pick up differences among ~~treatments~~ ^{populations} regardless of the nature of the differences.

The Kolmogorov-Smirnov test is appropriate for this situation.

We obtain $N = m + n$ observations X_1, \dots, X_m and Y_1, \dots, Y_n .

- The observations X_1, \dots, X_m are a random sample from population 1; that is, the X 's are independent and identically distributed. The observations Y_1, \dots, Y_n are a random sample from population 2; that is, the Y 's are independent and identically distributed.
- The X 's and Y 's are mutually independent. Thus, in addition to assumptions of independence within each sample, we also assume independence between the two samples.
- Populations 1 and 2 have continuous distributions with distribution functions F and G , respectively.

We are interested in assessing whether there are any differences whatsoever between the X and Y probability distributions.

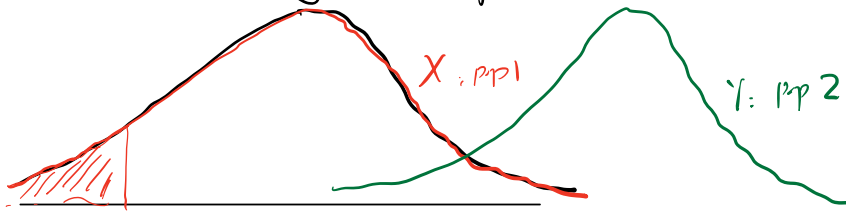
$$H_0 : F(t) = G(t) \text{ for all } t$$

versus

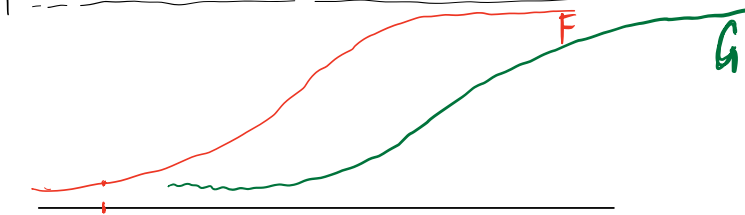
$$H_a : F(t) \neq G(t) \text{ for at least some } t$$

continuous data (random variables):

p.d.f (probability density function) $p(x)$



distribution function $F(x) \triangleq P(X \leq x)$



$H_0: F(t) = G(t)$ for $\forall t$

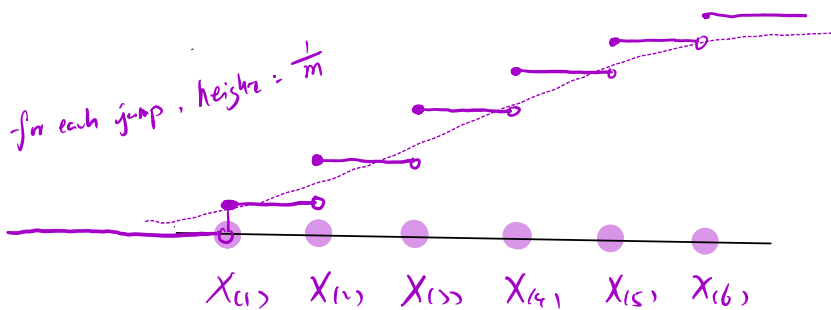
$H_A: F(t) \neq G(t)$

for at least some t

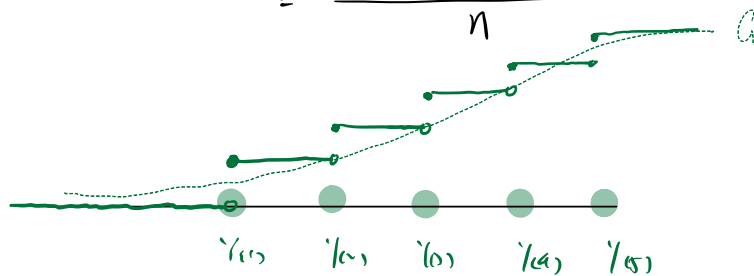
idea: use sample to estimate F , G . and then test

Defn: Empirical distribution function

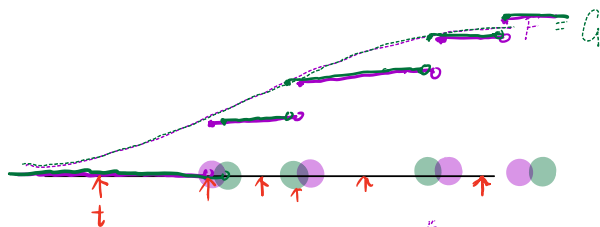
$$\begin{aligned} \text{for } X \text{ pop: } F_m(t) &= \text{proportion of } X \text{ samples } \leq t \\ &= \frac{\# X_s \leq t}{m} \end{aligned}$$



for 1 grp: $G_n(t) = \text{proportion of } Y \text{ samples} \leq t$
 $= \frac{\# Y \text{ samples} \leq t}{n}$

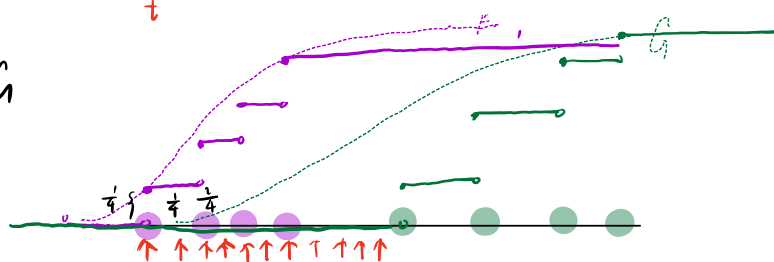


$H_0: F \equiv G$



$D \approx 0$
small in favor H_0

$H_a: F \neq G$



$D \nearrow$
large in favor H_a

$D = \max_{-\infty < t < \infty} |F_n(t) - G_n(t)|$
 difference

Motivation

Obtain the empirical distribution functions for the X and Y samples. For every t , let

$$F_m(t) = \frac{\text{number of sample } X \text{ 's } \leq t}{m}$$

and

$$G_n(t) = \frac{\text{number of sample } Y \text{ 's } \leq t}{n}.$$

(The functions $F_m(t)$ and $G_n(t)$ are called the **empirical distribution functions** for the X and Y samples, respectively.)

- if $F(t) = G(t)$
 $\max_{(-\infty < t < \infty)} \{|F_m(t) - G_n(t)|\}$ tend to \downarrow
- if $F(t) \neq G(t)$
 $\max_{(-\infty < t < \infty)} \{|F_m(t) - G_n(t)|\}$ tend to \uparrow

Kolmogorov-Smirnov statistic

$$D = \max_{(-\infty < t < \infty)} \{|F_m(t) - G_n(t)|\}$$

$F_m(t)$ and $G_n(t)$ are step functions changing functional values only at the observed X and Y sample observations, respectively. Thus, if we let $Z_{(1)} \leq \dots \leq Z_{(N)}$ denote the $N = (m + n)$ ordered values for the combined sample of X_1, \dots, X_m and Y_1, \dots, Y_n , then

$$D = \max_{i=1, \dots, N} \{|F_m(Z_{(i)}) - G_n(Z_{(i)})|\}$$

$$J = \frac{mn}{d} \max_{i=1 \dots 1-1} \{|F_m(Z_{(i)}) - G_n(Z_{(i)})|\}$$

where d : greatest common divisor of m and n

Derivation of null distribution using permutation

$\binom{m+n}{n}$ permutations of (X, Y) samples

calculate D/J for each permutation
KS

Large sample approximation of null distribution

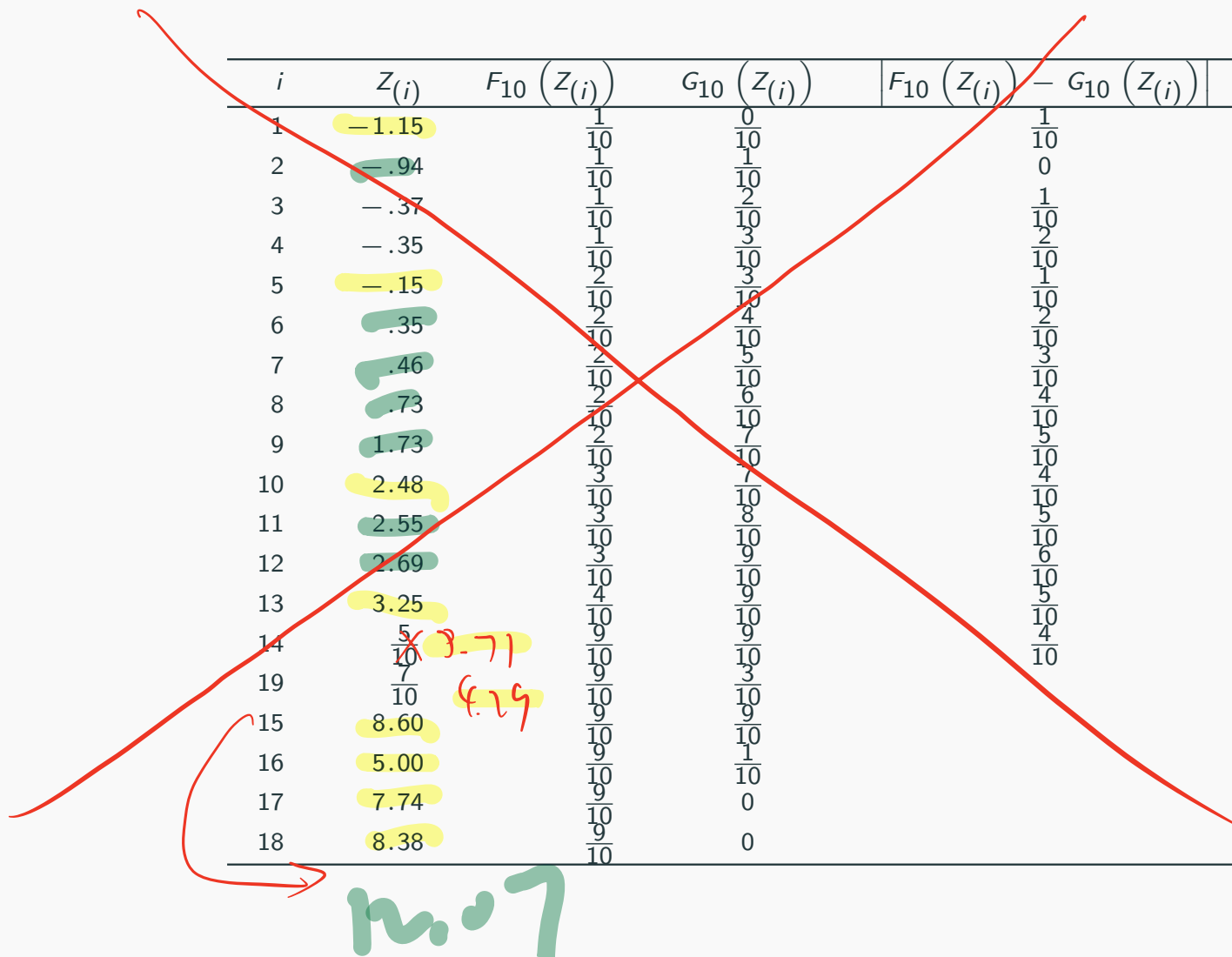
As $\min(m, n)$ tends to infinity,

$$J^* = \underbrace{\left(\frac{mn}{N}\right)^{1/2}}_{\text{rescaling factor}} \underbrace{\max_{i=1, \dots, N} \{ |F_m(Z_{(i)}) - G_n(Z_{(i)})| \}}_D \sim \text{Kolmogorov distribution}$$

Example: Effect of Feedback on Salivation Rate

The effect of enabling a subject to hear himself salivate while trying to increase or decrease his salivary rate has been studied by Delse and Feather (1968). Two groups of subjects were told to attempt to increase their salivary rates upon observing a light to the left and decrease their salivary rates upon observing a light to the right. Members of the feedback group received a 0.2-s, 1000-cps tone for each drop collected, whereas members of the no-feedback group did not receive any indication of their salivary rates.

Feedback group	No-Feedback group
— .15	2.55
8.60	12.07
5.00	.46
3.71	.35
4.29	2.69
7.74	— .94
2.48	1.73
3.25	.73
— 1.15	— .35
8.38	— .37



i	$Z_{(i)}$	$F_{10}(Z_{(i)})$	$G_{10}(Z_{(i)})$	$F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})$
1	-1.15	$\frac{1}{10}$	$\frac{0}{10}$	$\frac{1}{10}$
2	-.94	$\frac{1}{10}$	$\frac{1}{10}$	0
3	-.37	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$
4	-.35	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{2}{10}$
5	-.15	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$
6	.35	$\frac{2}{10}$	$\frac{4}{10}$	$\frac{2}{10}$
7	.46	$\frac{2}{10}$	$\frac{5}{10}$	$\frac{3}{10}$
8	.73	$\frac{2}{10}$	$\frac{6}{10}$	$\frac{4}{10}$
9	1.73	$\frac{2}{10}$	$\frac{7}{10}$	$\frac{5}{10}$
10	2.48	$\frac{3}{10}$	$\frac{7}{10}$	$\frac{4}{10}$
11	2.55	$\frac{3}{10}$	$\frac{8}{10}$	$\frac{5}{10}$
12	2.69	$\frac{3}{10}$	$\frac{9}{10}$	$\frac{6}{10}$
13	3.25	$\frac{4}{10}$	$\frac{9}{10}$	$\frac{5}{10}$
14	5 3.71	$\frac{5}{10}$	$\frac{9}{10}$	$\frac{4}{10}$
15	8.60	$\frac{7}{10}$	$\frac{3}{10}$	$\frac{4}{10}$
16	5.00	$\frac{7}{10}$	$\frac{9}{10}$	$\frac{1}{10}$
17	7.74	$\frac{7}{10}$	0	
18	8.38	$\frac{9}{10}$	0	

n=7

$$\max_{i=1, \dots, 20} \{ |F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})| \} = \frac{6}{10}$$

```

> x=c(-0.15,8.6,5,3.71,4.29,7.74,2.48,3.25,-1.15,8.38)
> y=c(2.55,12.07,0.46,0.35,2.69,-0.94,1.73,0.73,-0.35,-0.37)
>
> library(NSM3)
> pKolSmirn(x,y,method="Exact")
Number of X values: 10 Number of Y values: 10
Kolmogorov-Smirnov J Statistic: 6
Exact upper-tail probability: 0.0524
> pKolSmirn(x,y,method="Asymptotic")
Number of X values: 10 Number of Y values: 10
Kolmogorov-Smirnov J* Statistic: 1.3416
Asymptotic upper-tail probability: 0.0546

```

$$J^* = \sqrt{\frac{mn}{n}} D$$

Indicate some marginal evidence in the samples that feedback might have an effect on salivation rate.

Feedback group	No-Feedback group
-.15	2.55
8.60	12.07

$$\binom{4}{2} = \frac{4!}{2!2!} = 6$$

i	z_{1i2}	$F_1(z_{1i2})$	$G_2(z_{1i2})$	$ F - G $
1	-.15	.5	0	.5
2	2.55	1	0	.5
3	8.6	1	.5	.5
4	12.07	1	1	0

$$\Rightarrow D = .5$$

$$\Rightarrow J = \frac{2 \times 2}{2} D = 1$$

i	z_{1i2}	$F_1(z_{1i2})$	$G_2(z_{1i2})$	$ F - G $
1	-.15	.5	0	.5
2	2.55	.5	.5	0
3	8.6	1	.5	.5
4	12.07	1	1	0

$$\Rightarrow D = .5$$

$$J = 1$$

i	z_{1i2}	$F_1(z_{1i2})$	$G_2(z_{1i2})$	$ F - G $
1	-.15	.5	0	.5
2	2.55	.5	.5	0
3	8.6	.5	1	.5
4	12.07	1	1	0

$$\Rightarrow D = .5$$

$$J = 1$$

i	z_{1i2}	$F_1(z_{1i2})$	$G_2(z_{1i2})$	$ F - G $
1	-.15	0	.5	.5
2	2.55	.5	.5	0
3	8.6	1	.5	.5
4	12.07	1	1	0

$$\Rightarrow D = .5$$

$$J = 1$$

i	$z_{li,2}$	$F_1(z_{li,2})$	$G_2(z_{li,2})$	$ F - G $
1	- .15	0	.5	.5
2	2.55	.5	.5	0
3	8.6	.5	1	.5
4	12.07	1	1	0

$\Rightarrow D = .5$

$J = 1$

i	$z_{li,2}$	$F_1(z_{li,2})$	$G_2(z_{li,2})$	$ F - G $
1	- .15	0	.5	.5
2	2.55	0	1	1
3	8.6	.5	1	.5
4	12.07	1	1	0

$\Rightarrow D = 1$

$J = 2$

