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.

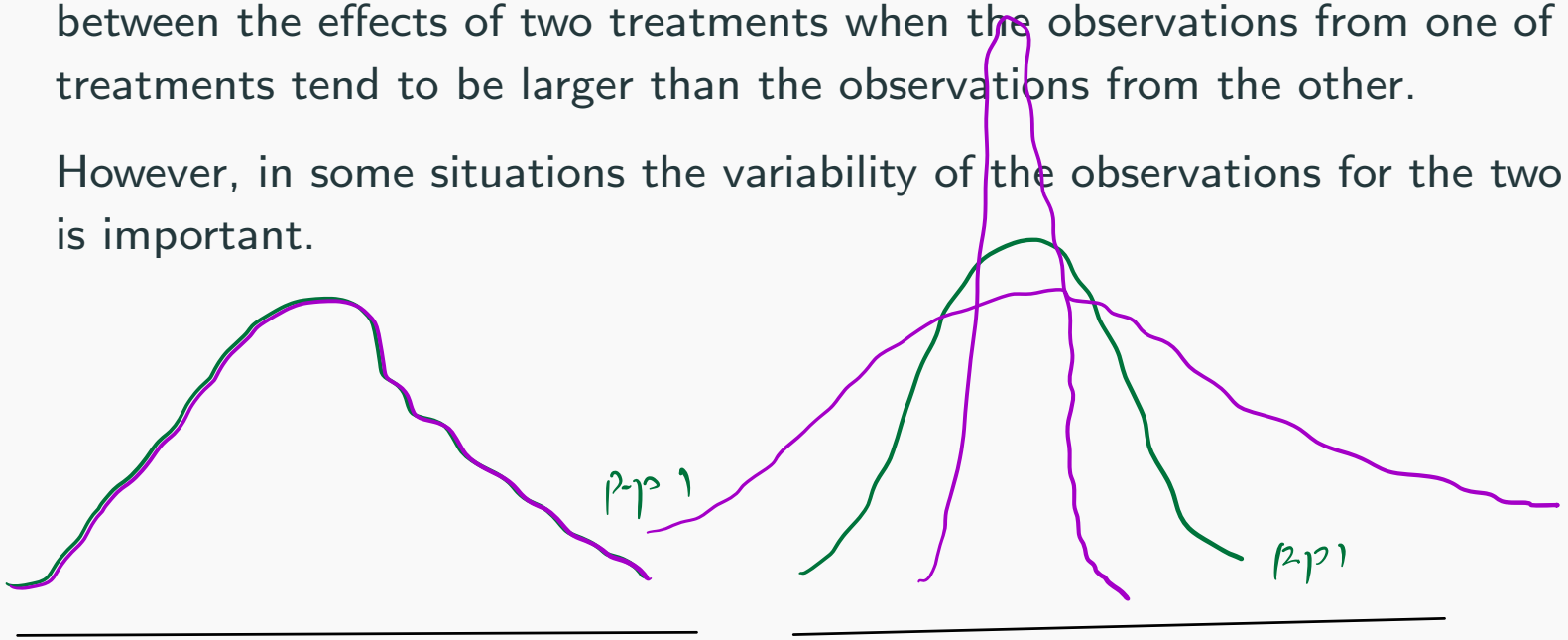
On the basis of these samples, we wish to investigate the presence of a treatment effect that results in a shift of location.

Table of contents

Tests for Equality of Scale Parameters

The tests we have discussed to this point are particularly designed to distinguish between the effects of two treatments when the observations from one of the treatments tend to be larger than the observations from the other.

However, in some situations the variability of the observations for the two treatments is important.



Suppose a machine for bottling a soft drink is designed to fill containers with 16 ounces of the beverage. Observations on the process may show that the data are centered around 16 as they should be, but there is excessive variability. This finding could lead an engineer to identify a problem that, if fixed, would reduce the variability of product quality.

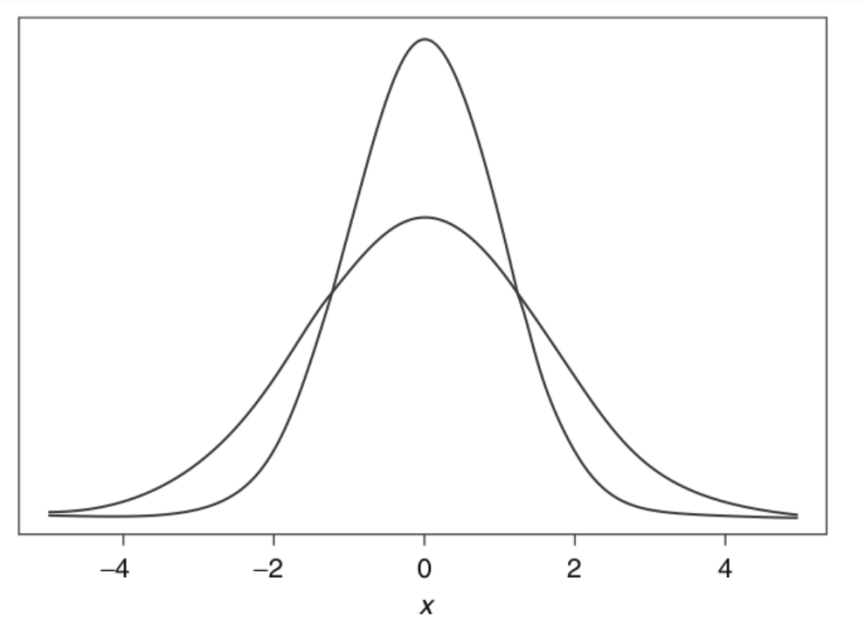


Figure 1: distributions of observations of the type that we would expect to see if the treatments affect the scale parameters of the population distributions but not the location parameters.

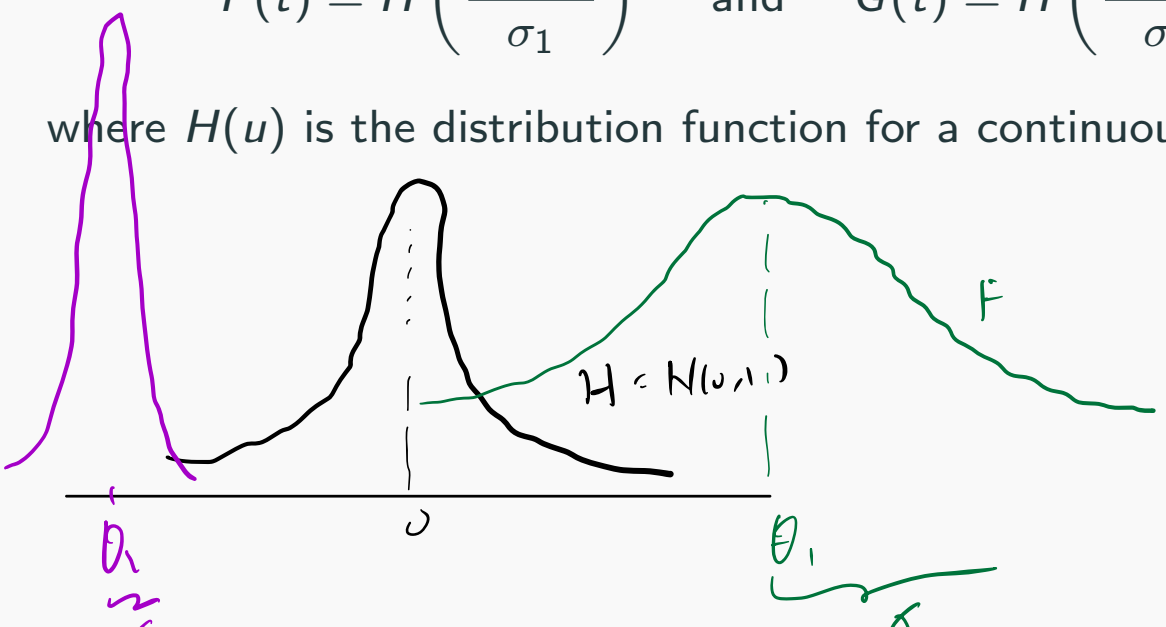
Setting

Let F and G be the distribution functions corresponding to populations 1 and 2, respectively.

The two-sample dispersion problem specifies that the Y population has greater (or less) variability associated with it than does the X population. We work under the location scale parameter model:

$$F(t) = H\left(\frac{t-\theta_1}{\sigma_1}\right) \quad \text{and} \quad G(t) = H\left(\frac{t-\theta_2}{\sigma_2}\right), \quad -\infty < t < \infty,$$

where $H(u)$ is the distribution function for a continuous distribution with median 0.



Setting

General assumptions of location scale model:

- θ_1 and θ_2 are the location parameter (~~medians~~ means) for the X and Y distributions, respectively.
- σ_1 and σ_2 are the scale parameters (variance) associated with the X and Y distributions, respectively.
- Y population has the same general form as the X population, but they could have different ~~medians~~ means and scale parameters.
- Another way to express this is to write

$$\frac{X-\theta_1}{\sigma_1} \stackrel{d}{=} \frac{Y-\theta_2}{\sigma_2},$$

We assume $\theta_1 = \theta_2 = \theta$, i.e. equal median.

$$\frac{X-\theta}{\sigma_1} \stackrel{d}{=} \frac{Y-\theta}{\sigma_2},$$

Hypothesis (Ansari-Bradley test)

The parameter of interest is the ratio of the scale parameters, $\gamma^2 = \frac{\sigma_1^2}{\sigma_2^2}$

Two-Sided Test:

$$H_0 : \gamma^2 = 1 \text{ versus } H_a : \gamma^2 \neq 1$$

One-Sided Upper-Tail Test:

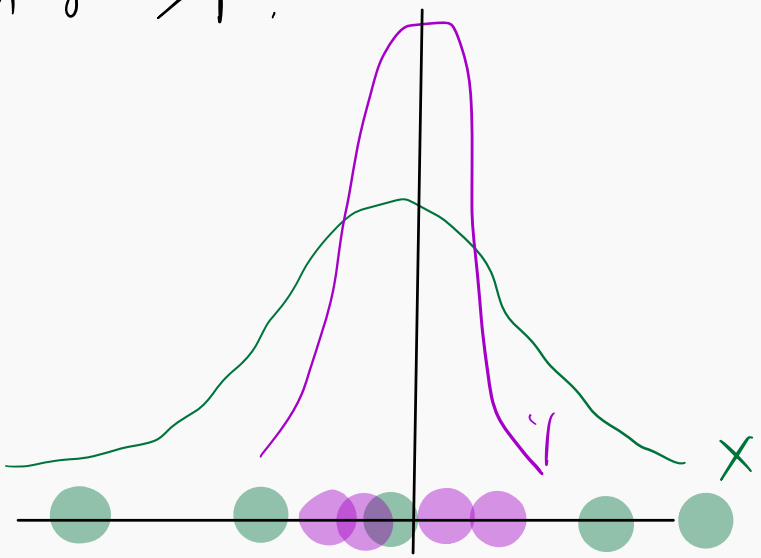
$$H_0 : \gamma^2 = 1 \text{ versus } H_a : \gamma^2 > 1$$

One-Sided Lower-Tail Test:

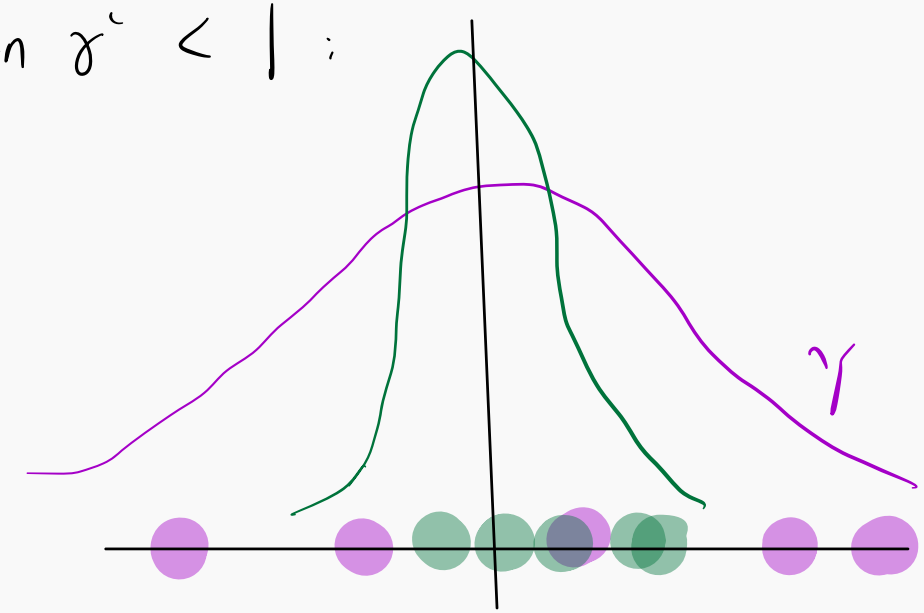
$$H_0 : \gamma^2 = 1 \text{ versus } H_a : \gamma^2 < 1$$

A rank based test for dispersion when median equal (Ansari-Bradley test)

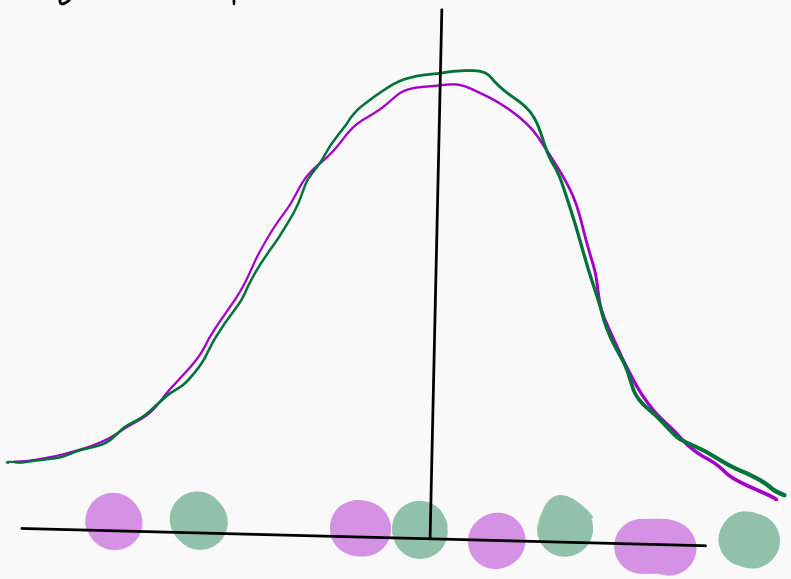
when $\gamma > 1$:



when $\gamma < 1$:



when $\gamma^2 \simeq 1$:



How to capture relative deviation from the center without using the exact values of X and Y?

- Order the combined sample of $N = (m + n)$ X-values and Y-values from least to greatest.
- Assign the score 1 to both the smallest and largest observations in this combined sample, assign the score 2 to the second smallest and second largest, and continue in the manner.
  - If N is an even integer, the array of assigned scores is $1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1$.
  - If N is an odd integer, the array of assigned scores is $1, 2, 3, \ldots, (N-1)/2, (N+1)/2, (N-1)/2, \ldots, 3, 2, 1$.

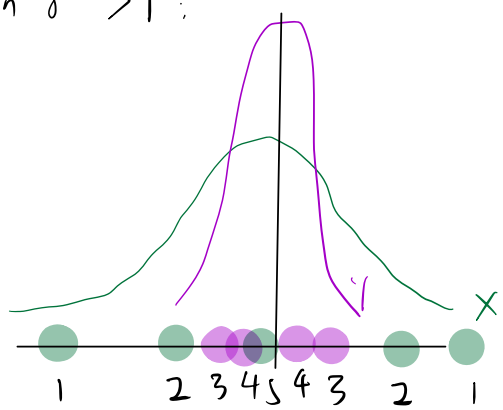
Let $R(Y_j)$ denote the score assigned in this manner to Y_j, for $j = 1, \ldots, n$, and

$$C = \sum_{j=1}^{n} R(Y_j)$$

Closer to the common center : rank ↑

Further away from the common center : rank ↓

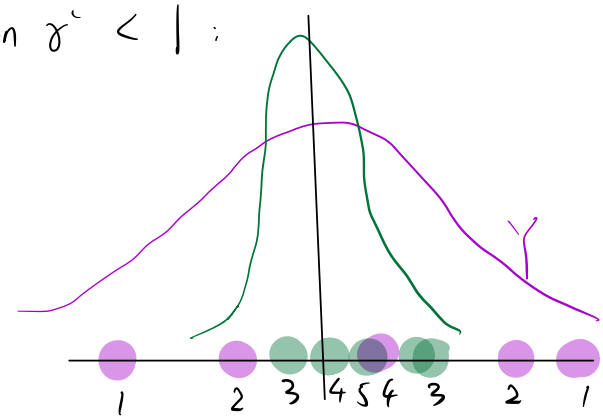
When $\gamma^2 > 1$:



C large

C : 3 + 4 + 4 + 3 = 14

When $\gamma^2 < 1$:

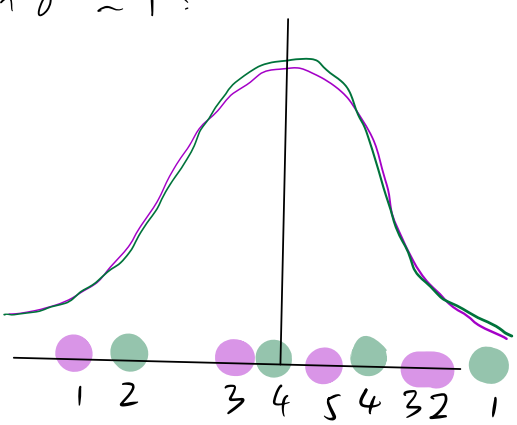


C small

C = 1 + 2 + 4 + 2 + 1

= 10

When $\gamma^2 \simeq 1$:



C not too large nor too small

C : 1 + 3 + 5 + 3 + 2 = 14

Derivation of null distribution using permutation

When H_0 is true: There are

$$\binom{n+m}{n} = \frac{(n+m)!}{n!\,m!}$$

possible assignments for $Y-$ranks (equal likely).

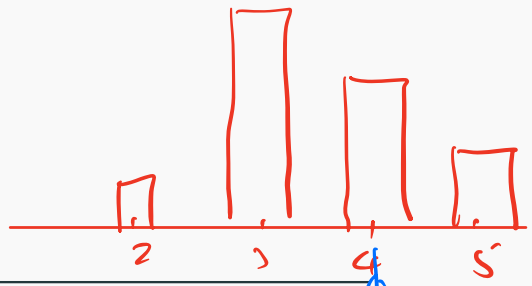
$m = 3, n = 2$ $\binom{5}{2} = \frac{5!}{2!\,3!} = \frac{5\times 4}{2} = 10$ permutations

~~Meshing~~	Probability	~~$(R^{(1)}, R^{(2)})$~~	$C = ~~R^{(1)} + R^{(2)}~~$
YYXXX	$\frac{1}{10}$	$(1,2)$	3
YXYXX	$\frac{1}{10}$	$(1,3)$	4
YXXYX	$\frac{1}{10}$	$(1,2)$	3
YXXXY	$\frac{1}{10}$	$(1,1)$	2
XYYXX	$\frac{1}{10}$	$(2,3)$	5
XYXYX	$\frac{1}{10}$	$(2,2)$	4
XYXXY	$\frac{1}{10}$	$(1,2)$	3
XXYYX	$\frac{1}{10}$	$(2,3)$	5
XXYXY	$\frac{1}{10}$	$(1,3)$	4
XXXYY	$\frac{1}{10}$	$(1,2)$	3

$R(1,) + R(1,)$

C: 2 3 4 5

=> null dist of C :



Possible value of C	Probability under H_0
2	$\frac{1}{10}$
3	$\frac{4}{10}$
4	$\frac{3}{10}$
5	$\frac{2}{10}$

$C_{obs} = 4$

$H_A: \gamma > 1$

Thus, for example, the probability, under H_0, that C is greater than or equal to 4 , for example, is therefore

$$P_0(C \geq 4) = P_0(C = 4) + P_0(C = 5) = .3 + .2 = .5,$$

Large sample approximation of null distribution

$$\frac{C}{n} = \frac{1}{n}\sum_{i=1}^{n} R(Y_i)$$

is the average of the scores assigned to the Y observations.

We want to know the behavior of: Under H_0, sample mean of a random sample of size n drawn without replacement from finite population of scores S_N, where $S_N = \{1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1\}$ if N is an even number and $S_N = \{1, 2, 3, \ldots, (N-1)/2, (N+1)/2, (N-1)/2, \ldots, 3, 2, 1\}$ if N is odd. [4]

[4] Facts from finite population theory:

- The mean is equal to the mean μ_{pop} of the finite population.
- The variance is equal to

$$\frac{\sigma^2_{\text{pop}}}{n} \times \frac{N-n}{N-1},$$

where σ^2_{pop} denotes the variance of the finite population and the factor $(N-n)/(N-1)$ is the finite population correction factor.

- if N even:

$$S_N = \{1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1\}$$
$$u_{\text{pop}} = \frac{2}{N} \sum_{i=1}^{N/2} i = \frac{(N/2)[(N/2)+1]}{2(N/2)} = \frac{N+2}{4}$$
$$E_0\left(\frac{C}{n}\right) = \frac{N+2}{4}$$
$$\text{var}_0\left(\frac{C}{n}\right) = \left[\frac{(N+2)(N-2)}{48n}\right]\left[\frac{N-n}{N-1}\right] = \frac{m(N+2)(N-2)}{48n(N-1)}$$
$$\Rightarrow E_0(C) = nE_0\left(\frac{C}{n}\right) = \frac{n(N+2)}{4}$$
$$\Rightarrow \text{var}_0(C) = n^2 \,\text{var}_0\left(\frac{C}{n}\right) = \frac{mn(N+2)(N-2)}{48(N-1)}$$

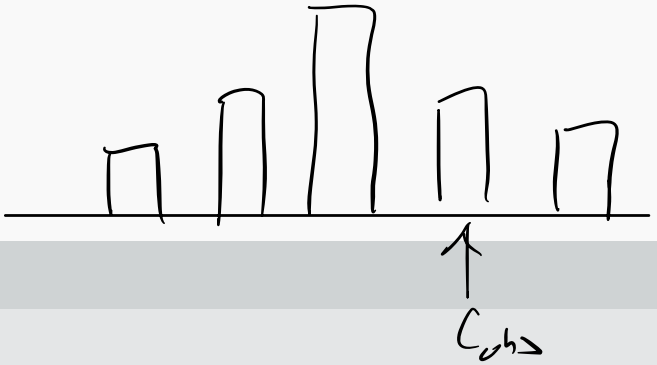
- if N odd:

$$S_N = \{1, 2, 3, \ldots, (N-1)/2, (N+1)/2, (N-1)/2, \ldots, 3, 2, 1\}$$
$$\Rightarrow E_0(C) = \frac{n(N+1)^2}{4N}$$
$$\Rightarrow \text{var}_0(C) = \frac{mn(N+1)\left(3+N^2\right)}{48N^2}$$

Asymptotic normality follows from standard theory for the mean of a sample.

$$C^* = \frac{C - E_0(C)}{\sqrt{\text{var}_0(C)}} \sim N(0, 1)$$

Procedure



a. One-Sided Upper-Tail Test

To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_1 : \gamma^2 > 1$$

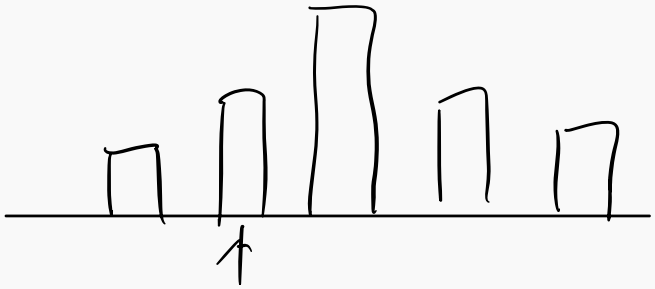
at the α level of significance,

Reject H_0 if $C \geq c_\alpha$; otherwise do not reject, where the constant w_α is chosen to make the type I error probability equal to α. (Or use p-value)

Large-sample approximation

Reject H_0 if $C^* \geq z_\alpha$; otherwise do not reject.

Procedure



b. One-Sided Lower-Tail Test

To test

$$H_0 : \gamma^2 = 1$$

versus

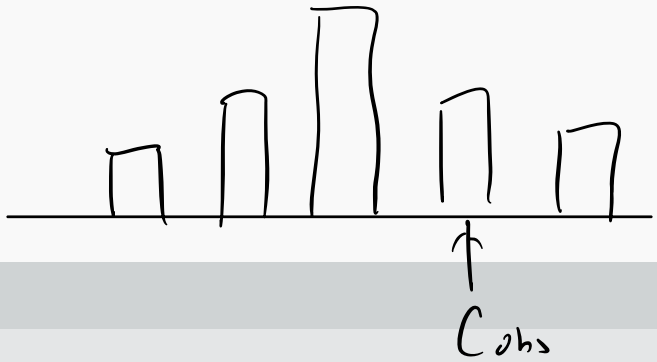
$$H_a : \gamma^2 < 1$$

at the α level of significance, Reject H_0 if $C \leq c_{1-\alpha}$; otherwise do not reject.

Large-sample approximation

Reject H_0 if $C^* \leq -z_\alpha$; otherwise do not reject.

Procedure



c. Two-Sided Test

To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_3 : \gamma^2 \neq 1$$

at the α level of significance, Reject H_0 if $C \geq c_{\alpha/2}$ or $C \leq c_{1-\alpha/2}$; otherwise do not reject

Large-sample approximation

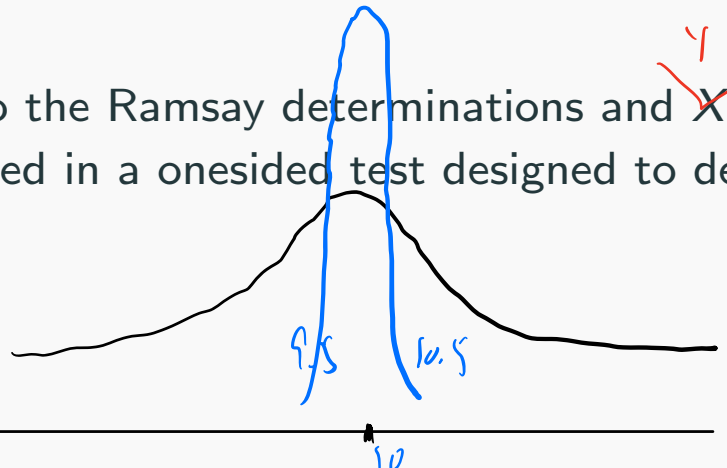
Reject H_0 if $|C^*| \geq z_{\alpha/2}$; otherwise do not reject.

Example:Serum Iron Determination

Jung and Parekh (1970) in a study concerned with techniques for direct determination of serum iron. In particular, they attempted to eliminate some of the problems associated with other commonly used methods, which often result in turbidity of the analyzed serum, as well as requiring large samples and slow, tedious analyses. To accomplish this, the authors proposed an improved method for serum iron determination based on a different detergent.

One of the purposes of their investigation was to study the accuracy of their method for serum iron determination in comparison to a method due to Ramsay (1957). From the point of view of procedural technique, the Jung-Parekh method competes favorably with the Ramsay method for serum iron determination. An additional concern, however, is whether there is a loss of accuracy when the Jung-Parekh procedure is used instead of the Ramsay procedure. As a result, the interest is greater dispersion or variation for the Jung-Parekh method of serum iron determination than for the method of Ramsay.

Hence, letting X correspond to the Ramsay determinations and Y to the Jung-Parekh determinations, we are interested in a onesided test designed to detect the alternative $H_1 : \gamma^2 > 1$.



true iron

old new

Ramsay method	Jung-Parekh method
111	107.5
107	108
99.5	105.5
98.5	98
102	105
106	103
109	110
108.5	106.5
103.5	104
99	100

Ramsay method	Jung-Parekh method
111	107.5
107	108
99.5	~~105.5~~

$m=3 \quad n=2$

order the combined sample : 99.5 107 107.5 108 111

rank : 1 2 3 2 1

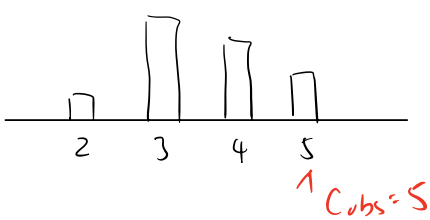
$\Rightarrow C = 5$

Permutation null distribution:

$m = 3, n = 2$

~~Meshing~~	Probability	~~$(R^{(1)}, R^{(2)})$~~	$C = ~~R^{(1)} + R^{(2)}~~$ $R(1,)+R(1,)$
YYXXX	$\frac{1}{10}$	$(1,2)$	3
YXYXX	$\frac{1}{10}$	$(1,3)$	4
YXXYX	$\frac{1}{10}$	$(1,2)$	3
YXXXY	$\frac{1}{10}$	$(1,1)$	2
XYYXX	$\frac{1}{10}$	$(2,3)$	5
XYXYX	$\frac{1}{10}$	$(2,2)$	4
XYXXY	$\frac{1}{10}$	$(1,2)$	3
XXYYX	$\frac{1}{10}$	$(2,3)$	5
XXYXY	$\frac{1}{10}$	$(1,3)$	4
XXXYY	$\frac{1}{10}$	$(1,2)$	3

Possible value of C	Probability under H_0
2	$\frac{1}{10}$
3	$\frac{4}{10}$
4	$\frac{3}{10}$
5	$\frac{2}{10}$



$H_1: \gamma^2 > 1$

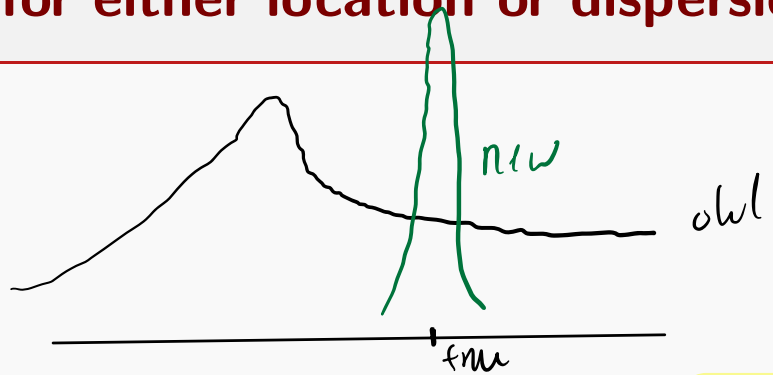
C larger

p-value $= P(C \geq 5) = P(C=5) = 0.2 > 0.05$

```
> ramsay <- c(111, 107, 99.5, 98.5, 102, 106, 109, 108.5, 103.5, 99)
> jung.parekh <- c(107.5, 108, 105.5, 98, 105, 103, 110, 106.5, 104,100)
>
> rank(c(ramsay,jung.parekh))
 [1] 20 14  4  2  6 12 16 17  8  3 15 18 11  1 10  7 19 13  9  5
>
> ansari.test(ramsay, jung.parekh, alternative="greater",exact=T)

        Ansari-Bradley test

data:  ramsay and jung.parekh
AB = 47, p-value = 0.1306
alternative hypothesis: true ratio of scales is greater than 1

> # large-sample normal approximation
> ansari.test(ramsay, jung.parekh, alternative="greater", exact=F)

        Ansari-Bradley test

data:  ramsay and jung.parekh
AB = 47, p-value = 0.1124
alternative hypothesis: true ratio of scales is greater than 1
```

Hence, there is not sufficient evidence to indicate loss of accuracy when the Jung-Parekh method is used instead of the Ramsay method.

A rank based test for either location or dispersion (Lepage test)



In many two-sample situations, we are interested in simultaneously detecting either location or scale differences between the X and Y populations.

We are interested in assessing whether there are differences in either the location parameters (i.e., medians) θ_1 and θ_2 or the scale parameters σ_1 and σ_2 for the X and Y populations.

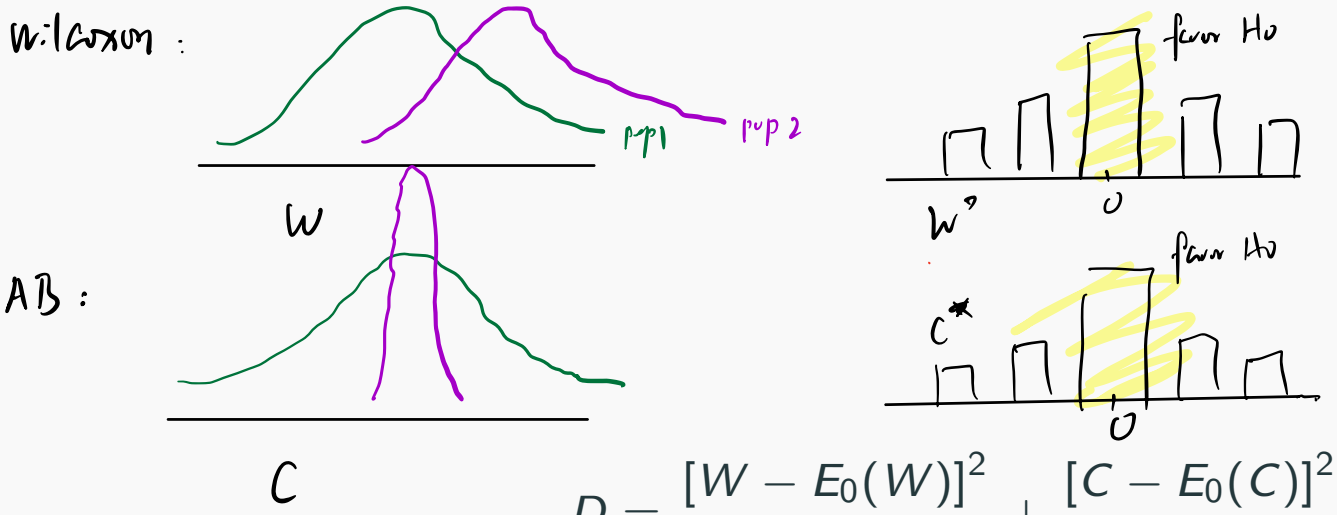
Thus, we are interested in testing

$$H_0 : \theta_1 = \theta_2, \sigma_1 = \sigma_2$$

vs

$$H_a : \theta_1 \neq \theta_2, \text{ or } \sigma_1 \neq \sigma_2$$

Motivation



$$D = \frac{[W - E_0(W)]^2}{\text{var}_0(W)} + \frac{[C - E_0(C)]^2}{\text{var}_0(C)}$$

$$= (W^*)^2 + (C^*)^2$$

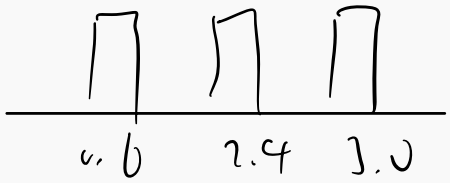
- A large value of $(W^*)^2$ is indicative of a possible difference in locations
- A large value of $(C^*)^2$ is indicative of a possible difference in dispersions
- D will be large if and only if $(W^*)^2$ is large or $(C^*)^2$ is large or both, then such a large value of D is indicative of alternative is true

Derivation of null distribution using permutation

$m = 2, n = 2$

	Probability	$D = (W^*)^2 + (C^*)^2$
$XXYY$	$\frac{1}{6}$	2.4
$XYXY$	$\frac{1}{6}$	.6
$YXXY$	$\frac{1}{6}$	3.0
$XYYX$	$\frac{1}{6}$	3.0
$YXYX$	$\frac{1}{6}$	.6
$YYXX$	$\frac{1}{6}$	2.4

Value of D	Probability under H_0
0.6	$\frac{1}{3}$
2.4	$\frac{1}{3}$
3.0	$\frac{1}{3}$



$n = m = 2 \quad N = n + m = 4$

$$W^* = \frac{W - \frac{n(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}} = \frac{W - 5}{\sqrt{\frac{5}{3}}}$$

$$C^* = \frac{C - \frac{n(N+2)}{4}}{\sqrt{\frac{mn(N+2)(N-2)}{48(N-1)}}} = \frac{C - 3}{\sqrt{\frac{1}{3}}}$$

Permutations	Prob	W	W^*	C	C^*	$D^2 = (W^*)^2 + (C^*)^2$
Y Y X X	$\frac{1}{6}$	3		3		
Y X Y X	$\frac{1}{6}$	4		3		
Y X X Y	$\frac{1}{6}$	5		2		
X Y Y X	$\frac{1}{6}$	5		4		
X Y X Y	$\frac{1}{6}$	6		3		
X X Y Y	$\frac{1}{6}$	7		3		

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

Large sample approximation of null distribution

$$D \sim \chi^2_{df=2}$$

$$D = (W')^2 + (C')^2$$

Example:Effect of Maternal Steroid Therapy on Platelet Counts of Newborn Infants.

Autoimmune thrombocytopenic purpura (ATP) is a disease in which the patient produces antibodies to his/her own platelets. Due to transplacental passage of antiplatelet antibodies during pregnancy, children of women with ATP are often born with low platelet counts. For this reason, there is medical concern that a vaginal delivery for a mother with ATP could result in intracranial hemorrhage for the infant. However, the proper obstetrical management of pregnant women with ATP is controversial. Most doctors have advocated cesarean section as the preferable method of delivery for mothers with ATP. Others suggest that cesarean section, with its obvious complications for both mother and infant, be avoided unless there is some additional obstetrical reason for it. Karpatkin, Porges, and Karpatkin (1981) studied the effect of administering the corticosteroid prednisone to pregnant women with ATP with the intent of raising the infants' platelet counts to safe levels during their deliveries.

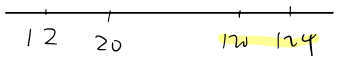
The data are a subset of the data obtained by Karpatkin et al. in their study of the effect that administration of prednisone to pregnant women with ATP had on their infants' platelet counts.

The primary interest in the study is in whether or not the predelivery administration of prednisone typically leads to an increased newborn platelet count. Thus, the principal statistical issue in the study is that of a possible difference in locations for the prednisone and nonprednisone populations. However, there is some concern that the administration of predelivery prednisone could also lead to a rather large increase in variability in the newborn platelet counts. (Such a finding would certainly affect our interpretation of any possible increase in typical platelet count resulting from the prednisone.)

We take the infant platelet count data for mothers given prednisone to be the Y sample ($n = 10$) and the corresponding control (nonprednisone) data to be the X sample ($m = 6$).

Mothers given prednisone	Mothers not given prednisone
120,000	12,000
124,000	20,000
215,000	112,000
90,000	32,000
67,000	60,000
95,000	40,000
190,000	
180,000	
135,000	
399,000	

Mothers given prednisone	Mothers not given prednisone
120,000	12,000
124,000	20,000



$n = m = 2 \qquad N = n + m = 4$

$W = 7$

$$W^* = \frac{W - \frac{n(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}} = \frac{W - 5}{\sqrt{\frac{5}{3}}} = \frac{2}{\sqrt{\frac{5}{3}}}$$

$C = 3$

$$C^* = \frac{C - \frac{n(N+2)}{4}}{\sqrt{\frac{mn(N+2)(N-2)}{48(N-1)}}} = \frac{C - 3}{\sqrt{\frac{1}{3}}} = 0$$

$$\Rightarrow D = \left(\frac{2}{\sqrt{\frac{5}{3}}}\right)^2 + 0^2 = 2.4$$

	Probability	$D = (W^*)^2 + (C^*)^2$
XXYY	$\frac{1}{6}$	2.4
XYXY	$\frac{1}{6}$	.6
YXXY	$\frac{1}{6}$	3.0
XYYX	$\frac{1}{6}$	3.0
YXYX	$\frac{1}{6}$	.6
YYXX	$\frac{1}{6}$	2.4

Value of D	Probability under H_0
0.6	$\frac{1}{3}$
2.4	$\frac{1}{3}$
3.0	$\frac{1}{3}$

$$\Rightarrow \text{p-value} = P(D \geq 2.4) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

$$> 0.05$$

$$W = 10 + 11 + 16 + 7 + 6 + 8 + 14 + 13 + 12 + 15 = 112$$

$$W^* = \frac{112 - \{10(6 + 10 + 1)/2\}}{\{6(10)(6 + 10 + 1)/12\}^{1/2}} = 2.929$$

$$C = 7 + 6 + 2 + 7 + 6 + 8 + 3 + 4 + 5 + 1 = 49$$

$$C^* = \frac{49 - \{10(16 + 2)/4\}}{\left\{\frac{10(6)(16+2)(16-2)}{48(16-1)}\right\}^{1/2}} = .873$$

$$D = (2.929)^2 + (.873)^2 = 9.34$$

[redacted]

```
> library(NSM3)
> # large-sample
> pLepage(x,y,method="Asymptotic")
Number of X values: 10 Number of Y values: 6
Lepage D Statistic: 9.3384
Asymptotic upper-tail probability: 0.0094
> # permutation
> pLepage(x,y,method="Exact")
Number of X values: 10 Number of Y values: 6
Lepage D Statistic: 9.3384
Exact upper-tail probability: 0.0025
```

We reject H_0 and there is significant difference in either locations or variabilities between the infant platelet counts for the prednisone and control populations.