

# AMS 572 Data Analysis I

## Inference on two population means and two population variances

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

# Two population setting

- ▶ Single cross-sectional sample, comparing two sub-samples
- ▶ Comparing samples from two different populations (two cross-sectional samples, case control study)
- ▶ Single sample; subjects randomly allocated to different interventions (experiment, clinical trials)
- ▶ Paired data; examples:
  - ▶ Study of how a characteristic changes from before to after a treatment
  - ▶ Study of twins
  - ▶ Individually-matched case-control study

# Overview of tests for two population means

- ▶ The samples are paired  $\Rightarrow$  paired samples t-test
- ▶ The samples are independent  $\Rightarrow$  independent samples t-test
  - ▶  $\sigma_1^2 = \sigma_2^2 \Rightarrow$  pooled-variance t-test (also known as equal variance t-test)
  - ▶  $\sigma_1^2 \neq \sigma_2^2 \Rightarrow$  unpooled-variance t-test (also known as unequal variance t-test)

## Inference on two population means: Paired samples

Example: Suppose that we are interested in comparing the IQ between fraternal twins consisting of a male and female

Pair	IQ Male	IQ Female	Difference $d_i$
1	200	199	1
2	150	120	30
...	...	...	...
n	109	112	-3

- ▶ Because of the inherent pairing or matching structure, a natural way is to consider the difference  $d_i$  between each pair.
- ▶ By considering the paired differences, the two population/sample problem reduces to a one population/one sample problem.
- ▶ One can apply the tests from Chapter 7 by treating  $d_i$ 's as the observations.

# Inference on two population means: Paired samples

The corresponding one or two-tailed hypothesis tests for paired samples reduce to:

$$\begin{cases} H_0 : \mu_m = \mu_f \\ H_a : \mu_m > \mu_f \end{cases} \Leftrightarrow \begin{cases} H_0 : \mu_d = 0 \\ H_a : \mu_d > 0 \end{cases}$$

$$\begin{cases} H_0 : \mu_m = \mu_f \\ H_a : \mu_m < \mu_f \end{cases} \Leftrightarrow \begin{cases} H_0 : \mu_d = 0 \\ H_a : \mu_d < 0 \end{cases}$$

$$\begin{cases} H_0 : \mu_m = \mu_f \\ H_a : \mu_m \neq \mu_f \end{cases} \Leftrightarrow \begin{cases} H_0 : \mu_d = 0 \\ H_a : \mu_d \neq 0 \end{cases}$$

where  $\mu_f$  = population mean IQ of female twin,  $\mu_m$  = population mean IQ of male twin,  $\mu_d$  = population mean difference in IQ

# SAS Code

```
data twins;
input IQmale IQfemale;
diff = IQmale-IQfemale;
datalines;
200 199
150 120
...
109 112
;
run;

proc univariate data=twins normal;
var diff;
run;
```

Example: To study the effectiveness of wall insulation in saving energy for home heating, the energy consumption (in MWh) for 5 houses in Bristol, England, was recorded for two winters; the first winter was before insulation and the second winter was after insulation:

House	1	2	3	4	5
Before	12.1	10.6	13.4	13.8	15.5
After	12.0	11.0	14.1	11.2	15.3

- (a) Provide a 95% confidence interval for the difference between the mean energy consumption before and after the wall insulation is installed. What assumptions are necessary for your inference?
- (b) Can you conclude that there is a difference in mean energy consumption before and after the wall insulation is installed at the significance level 0.05? Perform the hypothesis test and compute the p-value of your test. What assumptions are necessary for your inference?
- (c) Write a SAS and R program to perform the test and examine the necessary assumptions given in (b).



# SAS Code

```
Data energy;  
Input before after @@;  
Diff=before - after;  
Datalines;  
12.1 12.0 10.6 11.0 13.4 14.1 13.8 11.2 15.5 15.3  
;  
run;  
Proc univariate data = energy normal;  
Var Diff;  
run;
```

# Main SAS Output

The UNIVARIATE Procedure

Variable: Diff

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value-----
Student's t	t 0.616851	Pr >  t  0.5707
Sign	M 0.5	Pr >=  M  1.0000
Signed Rank	S 0.5	Pr >=  S  1.0000

Tests for Normality

Test	--Statistic--	-----p Value-----
Shapiro-Wilk	W 0.80253	Pr < W 0.0850
Kolmogorov-Smirnov	D 0.348791	Pr > D 0.0442
Cramer-von Mises	W-Sq 0.10017	Pr > W-Sq 0.0859
Anderson-Darling	A-Sq 0.551915	Pr > A-Sq 0.0779

## R Code and Output

```
> before <- c(12.1,10.6,13.4,13.8,15.5)
> after <- c(12,11,14.1,11.2,15.3)
> t.test(before,after,paired=TRUE)
```

Paired t-test

```
data:  before and after
t = 0.6169, df = 4, p-value = 0.5707
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.26036  1.98036
sample estimates:
mean of the differences
                0.36
```

```
> shapiro.test(before-after)
```

Shapiro-Wilk normality test

```
data:  before - after
W = 0.8025, p-value = 0.08496
```

# Inference on two population means: pooled variance t-test

# Setting

- ▶ Notation:

$$X_1, \dots, X_{n_1} \stackrel{iid}{\sim} N(\mu_1, \sigma_1^2)$$

$$Y_1, \dots, Y_{n_2} \stackrel{iid}{\sim} N(\mu_2, \sigma_2^2)$$

- ▶ Two independent samples of sample sizes  $n_1$  and  $n_2$ , the population variances are unknown but equal ( $\sigma_1^2 = \sigma_2^2 = \sigma^2$ )
- ▶ Both samples are normally distributed.
- ▶ Goal: Compare  $\mu_1$  and  $\mu_2$

# Review

- ▶ Two random variables  $X$  and  $Y$  are *independent* if for all  $x$  and  $y$

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

- ▶ Result A. If  $X$  and  $Y$  are independent random variables, then for any two constants  $a$  and  $b$  the random variable

$$W = aX + bY$$

has mean and variance

$$E(W) = aE(X) + bE(Y)$$

$$\text{Var}(W) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

- ▶ Result B. If  $X$  and  $Y$  are independent random variables that are **normally** distributed, then

$$W = aX + bY$$

is normally distributed with mean and variance given by Result A

- ▶ Corollary: If  $\bar{X}$  and  $\bar{Y}$  are based on two independent random samples of size  $n_1$  and  $n_2$  from two normal distributions with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$ , then

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

## Point estimator for $\mu_1 - \mu_2$

A point estimator for  $\mu_1 - \mu_2$

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{X} - \bar{Y} \sim N \left( \mu_1 - \mu_2, \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \sigma^2 \right)$$

under equal variance assumption  $\sigma_1^2 = \sigma_2^2 = \sigma^2$



# Review

- $W = Z_1^2 + Z_2^2 + \cdots + Z_k^2$ , where  $Z_i \stackrel{i.i.d.}{\sim} N(0, 1)$ , then

$$W \sim \chi_k^2$$

- $Z \sim N(0, 1)$ ,  $W \sim \chi_k^2$ , and  $Z$  and  $W$  are independent, then

$$T = \frac{Z}{\sqrt{\frac{W}{k}}} \sim t_k$$

## Pivotal quantity for inference on $\mu_1 - \mu_2$

- ▶ Result C. If  $\bar{X}$  and  $\bar{Y}$  are based on two independent random samples of size  $n_1$  and  $n_2$  from two normal distributions with means  $\mu_1$  and  $\mu_2$  and the same variances  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , then

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- ▶ Note: If  $n_1 = n_2$  then

$$s_p^2 = (s_1^2 + s_2^2)/2$$

## Confidence interval for $\mu_1 - \mu_2$

$$P(-t_{n_1+n_2-2, \frac{\alpha}{2}} \leq T \leq t_{n_1+n_2-2, \frac{\alpha}{2}}) = 1 - \alpha$$

$$P(-t_{n_1+n_2-2, \frac{\alpha}{2}} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{n_1+n_2-2, \frac{\alpha}{2}}) = 1 - \alpha$$

$$\begin{aligned} P(\bar{X} - \bar{Y} - t_{n_1+n_2-2, \frac{\alpha}{2}} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \\ \leq \bar{X} - \bar{Y} + t_{n_1+n_2-2, \frac{\alpha}{2}} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}) = 1 - \alpha \end{aligned}$$

A  $100(1 - \alpha)\%$  C.I for  $(\mu_1 - \mu_2)$

$$\bar{X} - \bar{Y} \pm t_{n_1+n_2-2, \frac{\alpha}{2}} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

# Hypothesis test for $\mu_1 - \mu_2$

Suppose  $H_0 : \mu_1 - \mu_2 = c_0$ ,

Test statistic:

$$T_0 = \frac{(\bar{X} - \bar{Y}) - c_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ under } H_0$$

# Hypothesis test for $\mu_1 - \mu_2$

1. One tailed:  $H_0 : \mu_1 - \mu_2 = c_0$  vs  $H_a : \mu_1 - \mu_2 > c_0$ 
  - ▶ If  $c_0 = 0$ , this reduces to  $H_0 : \mu_1 = \mu_2$
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $T_0 \geq t_{n_1+n_2-2,\alpha}$
  - ▶ Alternatively, if p-value =  $P(T_0 \geq t_0 | H_0) \leq \alpha$ , reject  $H_0$
2. One tailed:  $H_0 : \mu_1 - \mu_2 = c_0$  vs  $H_a : \mu_1 - \mu_2 < c_0$ 
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $T_0 \leq -t_{n_1+n_2-2,\alpha}$
  - ▶ Alternatively, if p-value =  $P(T_0 \leq t_0 | H_0) \leq \alpha$ , reject  $H_0$
3. Two tailed:  $H_0 : \mu_1 - \mu_2 = c_0$  vs  $H_a : \mu_1 - \mu_2 \neq c_0$ 
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $|T_0| \geq t_{n_1+n_2-2,\alpha/2}$
  - ▶ Alternatively, if p-value =  $2 \cdot P(T_0 \geq |t_0| | H_0) \leq \alpha$ , reject  $H_0$

Example: An experiment was conducted to compare the mean number of tapeworms in the stomachs of sheep that had been treated for worms against the mean number in those that were untreated. A sample of 14 worm-infected lambs was randomly divided into 2 groups. Seven were injected with the drug and the remainders were left untreated. After a 6-month period, the lambs were slaughtered and the following worm counts were recorded:

Drug-treated sheep	18	43	28	50	16	32	13
Untreated sheep	40	54	26	63	21	37	39

- (a) Test at  $\alpha = 0.05$  whether the treatment is effective or not (For now, you may assume equal variance, we will learn formal statistical procedure to check this assumption later in Notes 8).
- (b) What assumptions do you need for the inference in part (a)?
- (c) Write a SAS and R program necessary to answer questions raised in (a) and (b).

# SAS Code

```
data sheep;
input group worms @@;
datalines;
1 18 1 43 1 28 1 50 1 16 1 32 1 13
2 40 2 54 2 26 2 63 2 21 2 37 2 39
;
run;
proc univariate data=sheep normal;
class group;
var worms;
title 'Check for normality';
run;
proc ttest data=sheep sides=L;
class group;
var worms;
title 'Independent samples t-test';
run;
```



# SAS Output

```
proc npar1way data=sheep wilcoxon;  
class group;  
var worms;  
title 'Nonparametric test for two-mean comparisons';  
run;
```

# SAS Output

Check for normality

The UNIVARIATE Procedure

Variable: worms

group = 1

Moments

N	7	Sum Weights	7
Mean	28.5714286	Sum Observations	200
Std Deviation	14.093227	Variance	198.619048
Skewness	0.495124	Kurtosis	-1.2527576
Uncorrected SS	6906	Corrected SS	1191.71429
Coeff Variation	49.3262945	Std Error Mean	5.32673912

Basic Statistical Measures

Location		Variability	
Mean	28.57143	Std Deviation	14.09323
Median	28.00000	Variance	198.61905
Mode	.	Range	37.00000
		Interquartile Range	27.00000

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value-----	
Student's t	t 5.363775	Pr >  t	0.0017
Sign	M 3.5	Pr >=  M	0.0156
Signed Rank	S 14	Pr >=  S	0.0156

# SAS Output

## Tests for Normality

Test	--Statistic---		-----p Value-----	
Shapiro-Wilk	W	0.925604	Pr < W	0.5142
Kolmogorov-Smirnov	D	0.201976	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.039988	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.271024	Pr > A-Sq	>0.2500

# SAS Output

Check for normality

The UNIVARIATE Procedure

Variable: worms

group = 2

Moments

N	7	Sum Weights	7
Mean	40	Sum Observations	280
Std Deviation	14.6742405	Variance	215.333333
Skewness	0.38989154	Kurtosis	-0.4979133
Uncorrected SS	12492	Corrected SS	1292
Coeff Variation	36.6856012	Std Error Mean	5.54634157

Basic Statistical Measures

Location		Variability	
Mean	40.00000	Std Deviation	14.67424
Median	39.00000	Variance	215.33333
Mode	.	Range	42.00000
		Interquartile Range	28.00000

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value-----	
Student's t	t 7.211961	Pr >  t	0.0004
Sign	M 3.5	Pr >=  M	0.0156
Signed Rank	S 14	Pr >=  S	0.0156

# SAS Output

```
Tests for Normality

Test                --Statistic---    -----p Value-----
Shapiro-Wilk        W      0.952397    Pr < W      0.7515
Kolmogorov-Smirnov  D      0.214286    Pr > D      >0.1500
Cramer-von Mises    W-Sq  0.041652    Pr > W-Sq   >0.2500
Anderson-Darling    A-Sq  0.240103    Pr > A-Sq   >0.2500
```

## Independent samples t-test

5

## The TTEST Procedure

Variable: worms

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	7	28.5714	14.0932	5.3267	13.0000	50.0000
2	7	40.0000	14.6742	5.5463	21.0000	63.0000
Diff (1-2)		-11.4286	14.3867	7.6900		

  

group	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
1		28.5714	15.5374 41.6055	14.0932	9.0816 31.0342
2		40.0000	26.4286 53.5714	14.6742	9.4560 32.3136
Diff (1-2)	Pooled	-11.4286	-Infy 2.2772	14.3867	10.3165 23.7486
Diff (1-2)	Satterthwaite	-11.4286	-Infy 2.2791		

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	12	-1.49	0.0815
Satterthwaite	Unequal	11.98	-1.49	0.0815

## Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	6	6	1.08	0.9244

# R Code and Output

```
> drug <- c(18,43,28,50,16,32,13)
> untrt <- c(40,54,26,63,21,37,39)
> shapiro.test(drug)
```

Shapiro-Wilk normality test

```
data:  drug
W = 0.9256, p-value = 0.5142
> shapiro.test(untrt)
```

Shapiro-Wilk normality test

```
data:  untrt
W = 0.9524, p-value = 0.7515
```

# R Code and Output

```
> var.test(drug,untrt)
```

F test to compare two variances

data: drug and untrt

F = 0.9224, num df = 6, denom df = 6, p-value = 0.9244

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.1584911 5.3680240

sample estimates:

ratio of variances

0.9223795



# R Code and Output

```
> t.test(drug,untrt,alternative='less',var.equal=TRUE)
```

Two Sample t-test

data: drug and untrt

t = -1.4862, df = 12, p-value = 0.08152

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 2.277215

sample estimates:

mean of x mean of y

28.57143 40.00000

# R Code and Output

```
> wilcox.test(drug,untrt,alternative='less')
```

```
Wilcoxon rank sum test
```

```
data:  drug and untrt
```

```
W = 14, p-value = 0.1043
```

```
alternative hypothesis: true location shift is less than 0
```

Inference on two population means: unequal  
variance  $\sigma_1^2 \neq \sigma_2^2$

## Effect on testing $\mu_1 = \mu_2$

What if  $\sigma_1^2 \neq \sigma_2^2$  and unknown?

1. Large sample approximation
2. Normality: Welch-Satterthwaite approximation (Behrens-Fisher problem)

# Large sample approximation

- ▶ If  $n_1$  and  $n_2$  are large, equal variance assumption is not important
- ▶ Recall CLT plus Slutsky implies

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0, 1)$$

- ▶ Thus

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$$

- ▶  $Z$  is the pivotal quantity for inference on  $\mu_1 - \mu_2$  for large sample
- ▶ Generally, require  $n_j \geq 30$  for  $j = 1, 2$
- ▶ Note assumption that  $X_i$ 's and  $Y_i$ 's normally distributed no longer needed either (CLT)

# Confidence interval for $\mu_1 - \mu_2$ for large sample

A  $100(1 - \alpha)\%$  C.I for  $(\mu_1 - \mu_2)$

# Hypothesis test for $\mu_1 - \mu_2$ for large sample

Suppose  $H_0 : \mu_1 - \mu_2 = c_0$ ,

Test statistic:

$$Z_0 = \frac{(\bar{X} - \bar{Y}) - c_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1) \text{ under } H_0$$

# Hypothesis test for $\mu_1 - \mu_2$ for large sample

1. One tailed:  $H_0 : \mu_1 - \mu_2 = c_0$  vs  $H_a : \mu_1 - \mu_2 > c_0$ 
  - ▶ If  $c_0 = 0$ , this reduces to  $H_0 : \mu_1 = \mu_2$
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $Z_0 \geq z_\alpha$
  - ▶ Alternatively, if p-value =  $P(Z_0 \geq z_0 | H_0) \leq \alpha$ , reject  $H_0$
2. One tailed:  $H_0 : \mu_1 - \mu_2 = c_0$  vs  $H_a : \mu_1 - \mu_2 < c_0$ 
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $Z_0 \leq -z_\alpha$
  - ▶ Alternatively, if p-value =  $P(Z_0 \leq z_0 | H_0) \leq \alpha$ , reject  $H_0$
3. Two tailed:  $H_0 : \mu_1 - \mu_2 = c_0$  vs  $H_a : \mu_1 - \mu_2 \neq c_0$ 
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $|Z_0| \geq z_{\alpha/2}$
  - ▶ Alternatively, if p-value =  $2 \cdot P(Z_0 \geq |z_0| | H_0) \leq \alpha$ , reject  $H_0$



# Welch-Satterthwaite approximation for normal distribution and small samples

- ▶ Assume normality;  $n_1, n_2$  small
- ▶ Pivotal quantity for inference on  $\mu_1 - \mu_2$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df}$$

- ▶ Welch (Biometrika 1938),

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

## Confidence interval for $\mu_1 - \mu_2$ for normal distribution and small samples

A  $100(1 - \alpha)\%$  C.I for  $(\mu_1 - \mu_2)$

$$\bar{X} - \bar{Y} \pm t_{df, \frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

# Hypothesis test for $\mu_1 - \mu_2$ for normal distribution and small samples

Suppose  $H_0 : \mu_1 - \mu_2 = c_0$ ,

Test statistic:

$$T_0 = \frac{(\bar{X} - \bar{Y}) - c_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df} \text{ under } H_0$$

# Hypothesis test for $\mu_1 - \mu_2$ for normal distribution and small samples

1. One tailed:  $H_0 : \mu_1 - \mu_2 = c_0$  vs  $H_a : \mu_1 - \mu_2 > c_0$ 
  - ▶ If  $c_0 = 0$ , this reduces to  $H_0 : \mu_1 = \mu_2$
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $T_0 \geq t_{df,\alpha}$
  - ▶ Alternatively, if p-value =  $P(T_0 \geq t_0 | H_0) \leq \alpha$ , reject  $H_0$
2. One tailed:  $H_0 : \mu_1 - \mu_2 = c_0$  vs  $H_a : \mu_1 - \mu_2 < c_0$ 
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $T_0 \leq -t_{df,\alpha}$
  - ▶ Alternatively, if p-value =  $P(T_0 \leq t_0 | H_0) \leq \alpha$ , reject  $H_0$
3. Two tailed:  $H_0 : \mu_1 - \mu_2 = c_0$  vs  $H_a : \mu_1 - \mu_2 \neq c_0$ 
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $|T_0| \geq t_{df,\alpha/2}$
  - ▶ Alternatively, if p-value =  $2 \cdot P(T_0 \geq |t_0| | H_0) \leq \alpha$ , reject  $H_0$

# Inference on two population variances

## F distribution

If  $X_1$  and  $X_2$  are independent rvs with  $X_1 \sim \chi_{v_1}^2$  and  $X_2 \sim \chi_{v_2}^2$ , then

$$\frac{X_1/v_1}{X_2/v_2} \sim F_{v_1, v_2}$$

Note:

- ▶ If  $F \sim F_{v_1, v_2}$ , then
- ▶ Thus, if the F-table only gives the upper bound  $F_{v_1, v_2, \alpha, U}$ , i.e.,  $P(F \geq F_{v_1, v_2, \alpha, U}) = \alpha$ , the lower bound can be obtained using the relationship above.

## Inference on $\sigma_1^2/\sigma_2^2$

- ▶ Want to test

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_a : \sigma_1^2 \neq \sigma_2^2$$

- ▶ We know that (assuming normality)

$$\frac{(n_k - 1)s_k^2}{\sigma_k^2} \sim \chi_{n_k-1}^2 \text{ for } k = 1, 2$$

## Inference on $\sigma_1^2/\sigma_2^2$

- ▶ Let

$$X = \frac{(n_1 - 1)s_1^2}{\sigma_1^2} \text{ and } Y = \frac{(n_2 - 1)s_2^2}{\sigma_2^2}$$

- ▶ It follows that

$$F = \frac{X/(n_1 - 1)}{Y/(n_2 - 1)} \sim F_{n_1 - 1, n_2 - 1}$$

- ▶ Thus

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1 - 1, n_2 - 1}$$

is a pivotal quantity for inference on  $\sigma_1^2/\sigma_2^2$



# Point estimator and confidence interval for $\sigma_1^2/\sigma_2^2$

- Point estimator

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{s_1^2}{s_2^2}$$



$$1 - \alpha = P(F_{n_1-1, n_2-1, \alpha/2, L} \leq F \leq F_{n_1-1, n_2-1, \alpha/2, U})$$

$$= P\left(\frac{\frac{s_1^2}{s_2^2}}{F_U} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{\frac{s_1^2}{s_2^2}}{F_L}\right)$$

A  $100(1 - \alpha)\%$  C.I for  $\sigma_1^2/\sigma_2^2$

$$\left[ \frac{1}{F_{n_1-1, n_2-1, \alpha/2, U}} \frac{s_1^2}{s_2^2}, \frac{1}{F_{n_1-1, n_2-1, \alpha/2, L}} \frac{s_1^2}{s_2^2} \right]$$

# Hypothesis test for $\sigma_1^2/\sigma_2^2$

Suppose  $H_0 : \sigma_1^2 = \sigma_2^2 \Leftrightarrow H_0 : \sigma_1^2/\sigma_2^2 = 1$

Test statistic:

$$F_0 = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1} \text{ under } H_0$$

# Hypothesis test for $\sigma_1^2/\sigma_2^2$

1. One tailed:  $H_0 : \sigma_1^2/\sigma_2^2 = 1$  vs  $H_a : \sigma_1^2/\sigma_2^2 > 1$ 
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $F_0 \geq F_{n_1-1, n_2-1, \alpha, U}$
2. One tailed:  $H_0 : \sigma_1^2/\sigma_2^2 = 1$  vs  $H_a : \sigma_1^2/\sigma_2^2 < 1$ 
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $F_0 \leq F_{n_1-1, n_2-1, \alpha, L}$
3. Two tailed:  $H_0 : \sigma_1^2/\sigma_2^2 = 1$  vs  $H_a : \sigma_1^2/\sigma_2^2 \neq 1$ 
  - ▶ At the significance level  $\alpha$ , reject  $H_0$  in favor of  $H_a$  if  $F_0 \geq F_{n_1-1, n_2-1, \alpha/2, U}$  or  $F_0 \leq F_{n_1-1, n_2-1, \alpha/2, L}$

Example: A new drug for reducing blood pressure (BP) is compared to an old drug. 20 patients with comparable high BP were recruited and randomized evenly to the 2 drugs.

Reductions in BP after 1 month of taking the drugs are as follows:

New drug: 0, 10, -3, 15, 2, 27, 19, 21, 18, 10

Old drug: 8, -4, 7, 5, 10, 11, 9, 12, 7, 8

- (a) Assume both populations are normal\*\*, test at  $\alpha = 0.05$  whether the new drug is better than the old one
- (b) Write a SAS and R program to do the above tests (including the normality test)

\*\*NOTE: In general, we have to test whether the samples are normally distributed. For this question, we make this assumption to illustrate how to apply the statistical procedure for normally distributed data.

# SAS Code

```
PROC UNIVARIATE DATA=BP NORMAL;  
CLASS DRUG;  
VAR BPR;  
RUN;
```

```
PROC TTEST DATA=BP SIDES=U;  
CLASS DRUG;  
VAR BPR;  
RUN;
```

# Selected SAS Output

## Tests for Normality

Test (class=1)	--Statistic--	-----p Value-----
Shapiro-Wilk	W      0.955526	Pr < W      0.7339
Kolmogorov-Smirnov	D      0.142059	Pr > D      >0.1500
Cramer-von Mises	W-Sq   0.037679	Pr > W-Sq   >0.2500
Anderson-Darling	A-Sq   0.238555	Pr > A-Sq   >0.2500

Test (class =2)	--Statistic--	-----p Value-----
Shapiro-Wilk	W      0.805147	Pr < W      0.0167
Kolmogorov-Smirnov	D      0.273266	Pr > D      0.0334
Cramer-von Mises	W-Sq   0.124461	Pr > W-Sq   0.0447
Anderson-Darling	A-Sq   0.776159	Pr > A-Sq   0.0294

## Selected SAS Output

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	18	1.34	0.0981
Satterthwaite	Unequal	12.547	1.34	0.1016

### Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	4.87	0.0274

# SAS Code and Output for Wilcoxon Rank Sum test

```
PROC NPAR1WAY DATA=BP;  
CLASS DRUG;  
VAR BPR;  
RUN;
```

## Wilcoxon Two-Sample Test

Statistic	123.0000
-----------	----------

### Normal Approximation

Z	1.3259
One-Sided Pr > Z	0.0924
Two-Sided Pr >  Z	0.1849

### t Approximation

One-Sided Pr > Z	0.1003
Two-Sided Pr >  Z	0.2006



# R Code and Output

```
> new <- c(0,10,-3,15,2,27,19,21,18,10)
> old <- c(8,-4,7,5,10,11,9,12,7,8)
> shapiro.test(new)
```

Shapiro-Wilk normality test

```
data:  new
W = 0.95553, p-value = 0.7339
> shapiro.test(old)
```

Shapiro-Wilk normality test

```
data:  old
W = 0.80515, p-value = 0.01674
```

# R Code and Output

```
> var.test(new,old)
```

F test to compare two variances

data: new and old

F = 4.869, num df = 9, denom df = 9, p-value = 0.02736

alternative hypothesis: true ratio of variances is not equal to  
95 percent confidence interval:

1.209381 19.602411

sample estimates:

ratio of variances

4.868962

# R Code and Output

```
> t.test(new,old,alternative='greater',var.equal=FALSE)
```

Welch Two Sample t-test

```
data:  new and old
```

```
t = 1.3423, df = 12.547, p-value = 0.1016
```

```
alternative hypothesis: true difference in means is greater than
```

```
95 percent confidence interval:
```

```
 -1.485809      Inf
```

```
sample estimates:
```

```
mean of x mean of y
```

```
    11.9      7.3
```

## R Code and Output

```
> wilcox.test(new,old,alternative='greater',exact=FALSE)
```

Wilcoxon rank sum test with continuity correction

data: new and old

W = 68, p-value = 0.09244

alternative hypothesis: true location shift is greater than 0