1

Part 3

Lecture 3 Recap on Confounding

Medical Imaging
UNIVERSITY OF TORONTO

MiDATA

Who I am…

Pascal Tyrrell, PhD

Associate Professor

Department of Medical Imaging, Faculty of Medicine

Institute of Medical Science, Faculty of Medicine

Department of Statistical Sciences, Faculty of Arts and Science

BIAS AND CONFOUNDING

The factory workers of Art Smith are concerned about having severe Shortness of Breath (SOB). 100 employees exposed and a 100 not exposed were selected from the workforce. The following table summarizes the results for the SOB variable.

EXPOSURE AND PERCENTAGE OF SOB

		DISEASE		Total	%SOB	
		YES	NO			
TOXIN	YES	38	62	100	38%	
	NO	15	85	100	15%	Fisher’s p=0.0004

Odds Ratio = 3.48 95% CI 1.76 - 6.87

Relative Risk = 2.53 95% CI 1.49 - 4.30

REVIEW OF RELATIVE RISK AND ODDS RATIO

EXPOSURE AND PERCENTAGE OF SOB

		DISEASE		Total	%SOB	
		YES	NO			
TOXIN	YES	38	62	100	38%	
	NO	15	85	100	15%	Fisher's p=0.0004

RELATIVE RISK = P1/P0 = 0.38 / 0.15 = 2.53

ODDS RATIO = (P1/Q1) / (P0/Q0)
= (0.38/0.62) / (0.15/0.85)
= 0.613 / 0.176 = 3.48

His friend John Smith was surprised because his workers are exposed to the same chemical and have reported no increase in SOB. He then remembered that most of his employees are women. This observation led Art to look at the results for his female employees.

```
    EXPOSURE AND SOB PERCENT IN FEMALES

          DISEASE   Total %SOB
            YES  NO
        YES 2    18   20    10.00%
TOXIN
        NO  7    73   80    8.75%       Fisher's p=1.0

Odds Ratio   = 1.16   95% CI  0.22 - 6.06
Relative Risk= 1.14   95% CI  0.26 - 5.01
```

Art now assumed the excess risk must be among his male workers and was surprised that his males workers experienced no increased risk.(p = 0.86).

EXPOSURE AND % SOB DISEASE IN MALES

		DISEASE		Total	%SOB
		YES	NO		
TOXIN	YES	36	44	80	45
	NO	8	12	20	40

Fisher’s p=0.86

Relative Risk=1.13 95% CI 0.62 - 2.03
Odds Ratio =1.23 95% CI 0.45 - 3.33

PERCENTAGE OF SOB
BY TOXIN EXPOSURE AND SEX

	FEMALES	MALES	OVERALL
YES	2/20= 10.00%	36/80= 45.00%	38/100
NO	7/80= 8.75%	8/20= 40.00%	15/100
RR	1.14	1.13	2.53
Fisher's p	1.00	0.86	0.0004

LANGUAGE OF CONFOUNDING

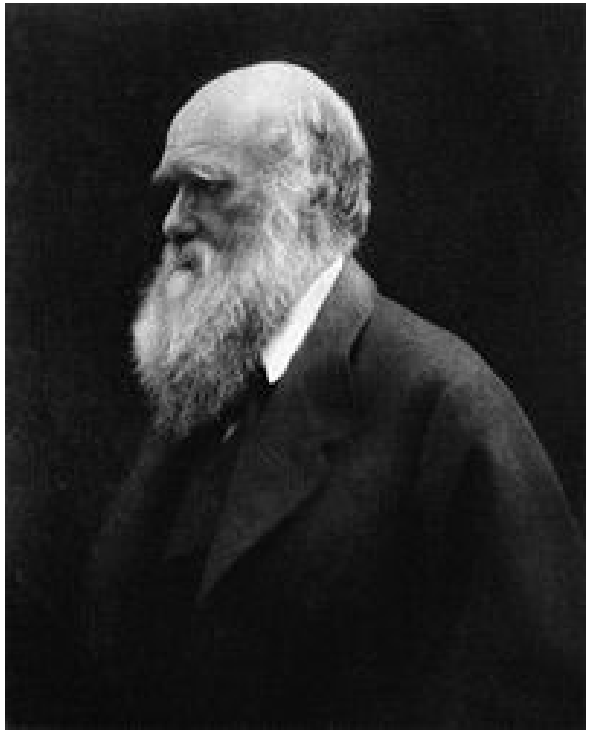
- Males are at a higher risk of disease
- Higher proportion of males exposed to toxin
- The increased risk of SOB in the group of exposed workers is not due to the toxin but because it has a greater percentage of males who have a higher percentage of SOB.

Perhaps males are given dirtier jobs or choose not to wear protective clothing. The SEX variable is called a CONFOUNDER.

Famous discussion between Charles Darwin and his cousin Francis Galton

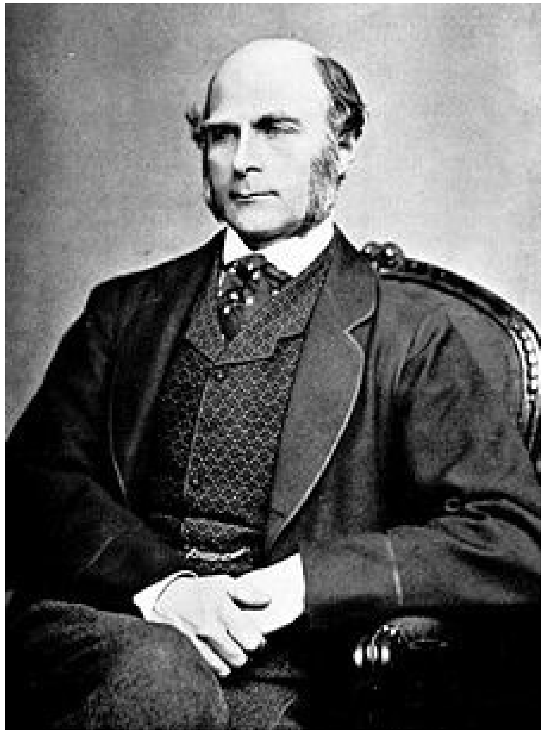
Charles Darwin

1809 – 1882



Francis Galton

1822 – 1911



Darwin, concerned about the small sample size of his experiment, wrote:

"As only a moderate number of crossed and self-fertilized were measured, it was of great importance to me to learn how far the averages were trustworthy. I therefore asked Mr. Galton, who has much experience in statistical researches to examine some of my tables of measurement ….."

The typical survey in the social sciences in the 19th century was very large. Darwin, a biologist, understood that a difference between two means of only 15 subjects selected from two different countries would unlikely result in any sound conclusions because of the wide variability among the observations.

However, Darwin believed the small sample size of his experiment was compensated by the care that he took in the design and execution of his study:

"But the case is somewhat different with my crossed and self-fertilized plants, as they were of exactly the same age, were subjected from the first to last to the same conditions and were descended from the same parents."

CREATING WIDE DATASET FOR PAIRED DATA

```
DATA INBRED ; INPUT INBRED @@ ;
GROUP = "INBRED " ; DATALINES;
139 163 160 160 147 149 149 122
132 144 130 144 102 124 144
RUN;

DATA CROSSED ; INPUT CROSSED @@ ;
GROUP = "CROSSED" ; DATALINES;
188  96 168 176 153 172 177 163
146 173 186 168 177 184  96
RUN;
```

Create a “paired dataset”

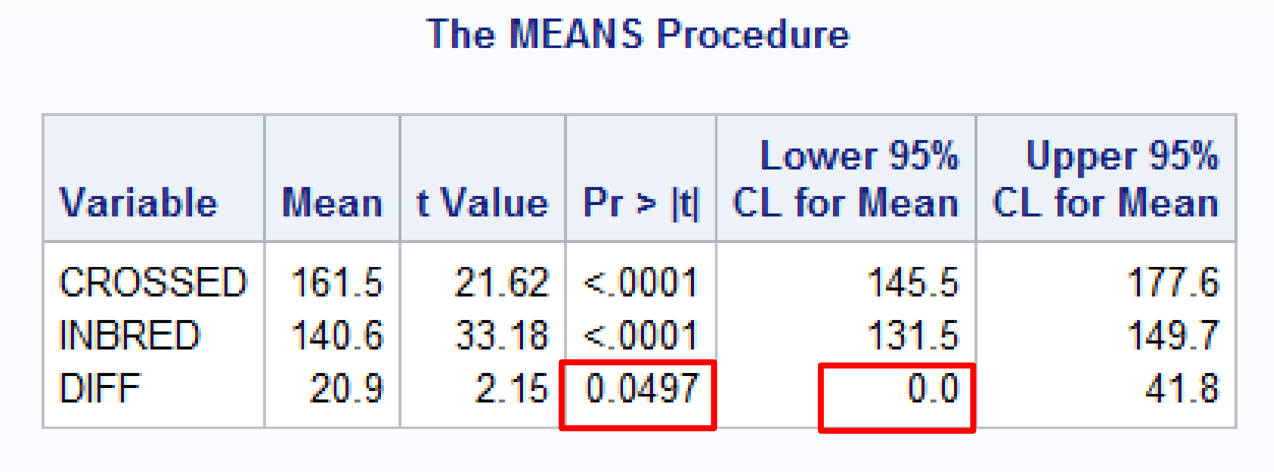
```
DATA WIDE; MERGE CROSSED INBRED;
DROP GROUP; PAIR = _N_ ;
DIFF = CROSSED - INBRED ; RUN;

TITLE1 "CROSSED AND INBRED PLANTS";
PROC PRINT DATA = WIDE NOOBS; RUN;
```

CROSSED AND INBRED PLANTS

PAIR	CROSSED	INBRED	DIFF
1	188	139	49
2	96	163	-67
3	168	160	8
4	176	160	16
5	153	147	6
6	172	149	23
7	177	149	28
8	163	122	41
9	146	132	14
10	173	144	29
11	186	130	56
12	168	144	24
13	177	102	75
14	184	124	60
15	96	144	-48
MEAN	161.53	140.60	20.93
VARIANCE	837.27	269.40	1424.64
SKEWNESS	-1.73	-0.80	-1.11

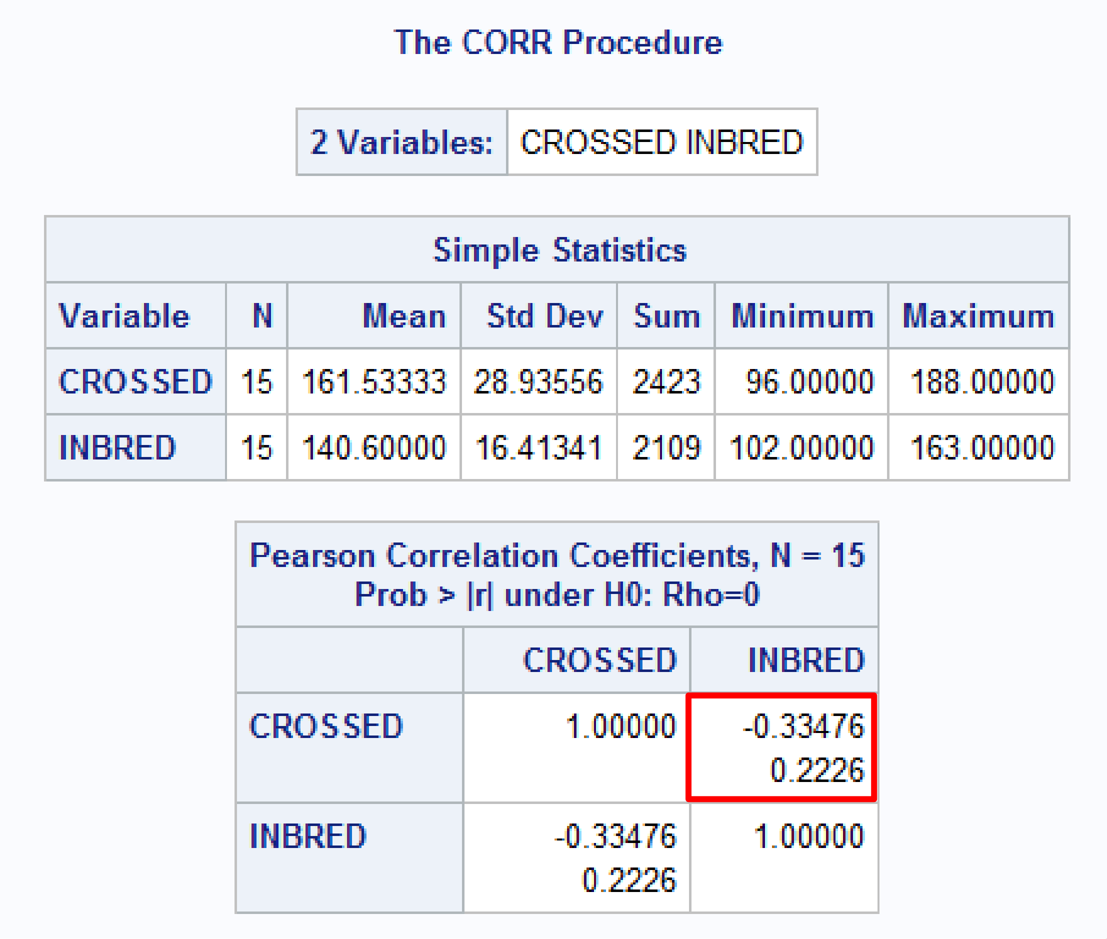
COMPARING MEANS USING THE MEANS OR TTEST PROCEDURES

```
TITLE1 "ASSUMING PLANTS WERE PAIRED";
PROC MEANS DATA=WIDE MEAN T PRT CLM MAXDEC=1;
VAR CROSSED INBRED DIFF ; RUN ;

PROC TTEST DATA=WIDE; PAIRED CROSSED*INBRED; RUN;

TITLE1 "PEARSON CORRELATION COEFFICIENT";
PROC CORR  DATA=WIDE; VAR CROSSED INBRED; RUN;
```



The MEANS Procedure

Variable	Mean	t Value	Pr > |t|	Lower 95% CL for Mean	Upper 95% CL for Mean
CROSSED	161.5	21.62	<.0001	145.5	177.6
INBRED	140.6	33.18	<.0001	131.5	149.7
DIFF	20.9	2.15	0.0497	0.0	41.8



The CORR Procedure

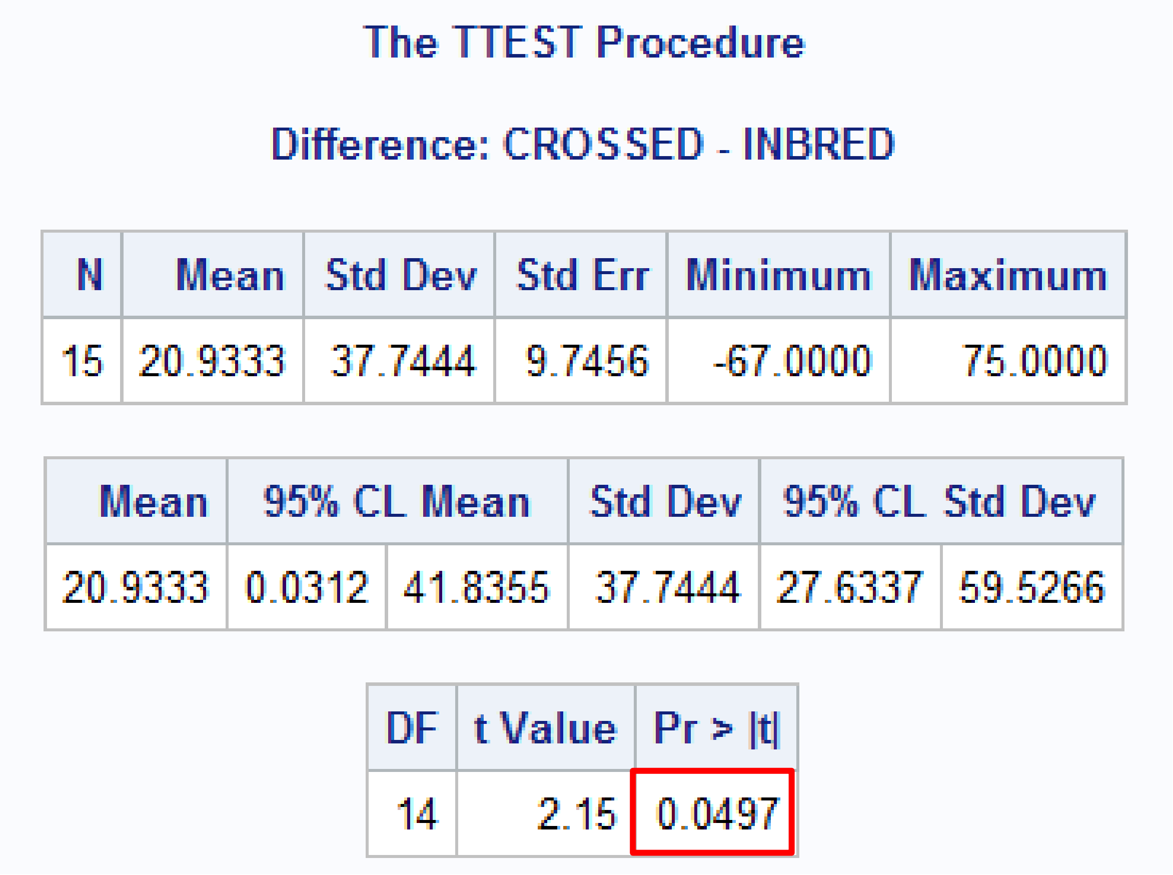
2 Variables:	CROSSED INBRED

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CROSSED	15	161.53333	28.93556	2423	96.00000	188.00000
INBRED	15	140.60000	16.41341	2109	102.00000	163.00000

Pearson Correlation Coefficients, N = 15
Prob > |r| under H0: Rho=0

	CROSSED	INBRED
CROSSED	1.00000	-0.33476 0.2226
INBRED	-0.33476 0.2226	1.00000



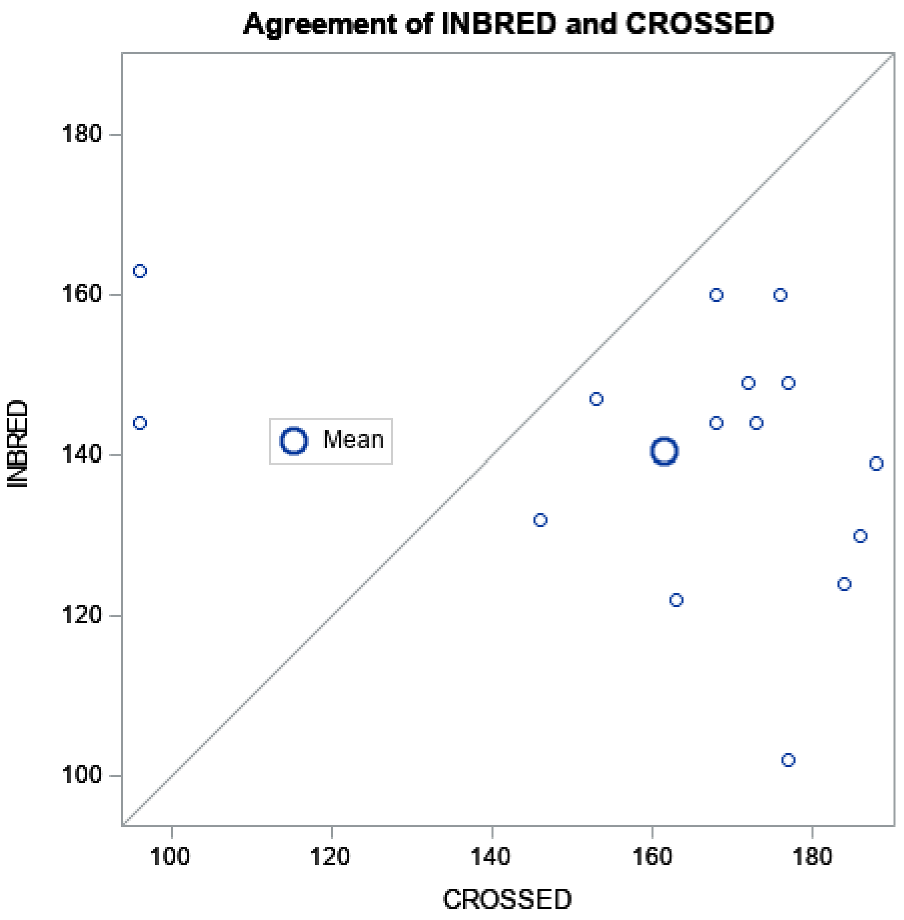
The TTEST Procedure

Difference: CROSSED - INBRED

N	Mean	Std Dev	Std Err	Minimum	Maximum
15	20.9333	37.7444	9.7456	-67.0000	75.0000

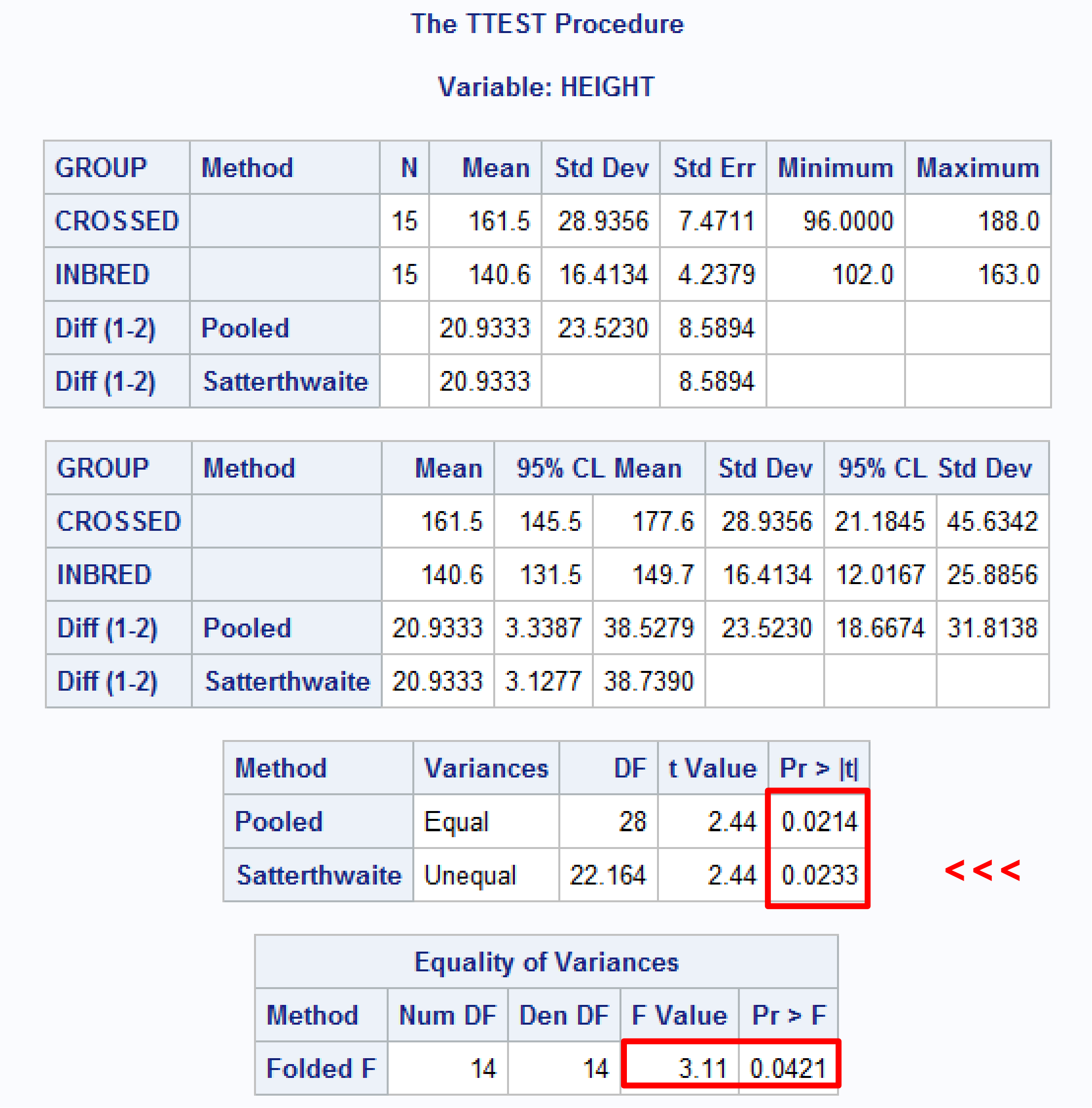
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
20.9333	0.0312	41.8355	37.7444	27.6337	59.5266

DF	t Value	Pr > |t|
14	2.15	0.0497



```
TITLE1 "  CREATING TWO DATASETS  ";
TITLE2 "IGNORING POSSIBLE PAIRING";
DATA LONGCROSS; SET CROSSED;
GROUP = "CROSSED"; HEIGHT=CROSSED; RUN;

DATA LONGINBRED ; SET INBRED ;
GROUP = "INBRED "; HEIGHT=INBRED; RUN;

DATA LONG; SET LONGCROSS LONGINBRED; RUN;

PROC TTEST DATA = LONG ;
CLASS GROUP ;
VAR HEIGHT ; RUN ;
```

GLM ANALYSIS ASSUMING A COMPLETELY RANDOMIZED DESIGN



The TTEST Procedure

Variable: HEIGHT

GROUP	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
CROSSED		15	161.5	28.9356	7.4711	96.0000	188.0
INBRED		15	140.6	16.4134	4.2379	102.0	163.0
Diff (1-2)	Pooled		20.9333	23.5230	8.5894		
Diff (1-2)	Satterthwaite		20.9333		8.5894		

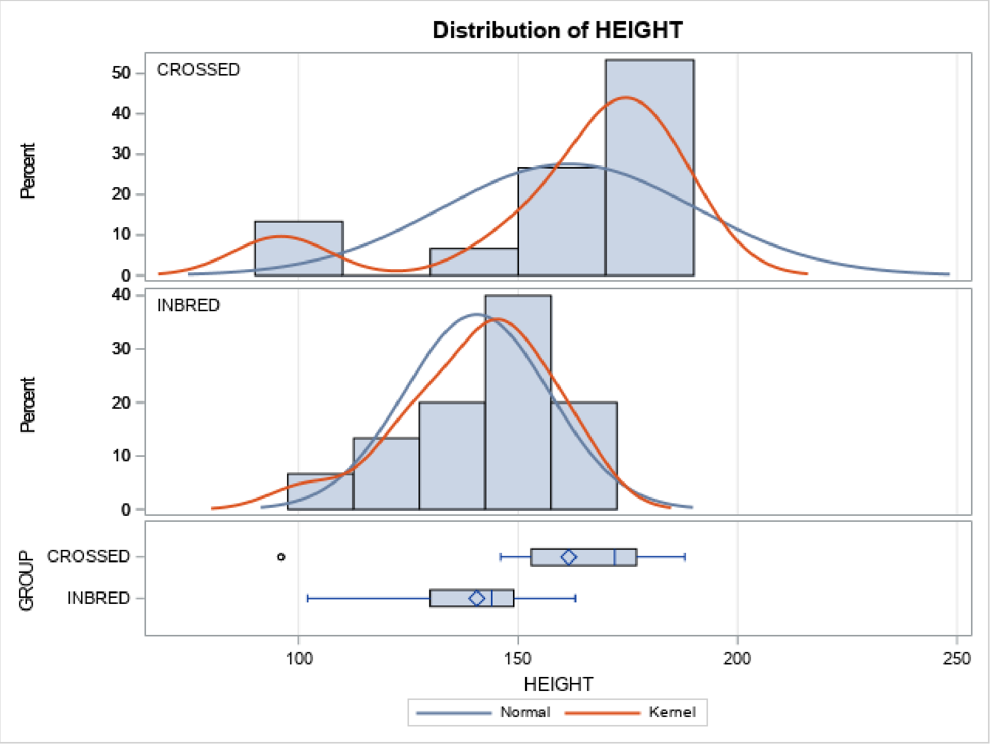
GROUP	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
CROSSED		161.5	145.5	177.6	28.9356	21.1845	45.6342
INBRED		140.6	131.5	149.7	16.4134	12.0167	25.8856
Diff (1-2)	Pooled	20.9333	3.3387	38.5279	23.5230	18.6674	31.8138
Diff (1-2)	Satterthwaite	20.9333	3.1277	38.7390			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	2.44	0.0214
Satterthwaite	Unequal	22.164	2.44	0.0233

<<<

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	14	3.11	0.0421



MY NOTE: RATIO OF SAMPLE VARIANCES $= (28.94/16.41)^2 = 3.11$

Medical Imaging
UNIVERSITY OF TORONTO

MiDATA

End of Lecture 3

Next up in Part 4 Lecture 1: Categorical Data