

## Lecture 7

1. GLM: continuous and categorical predictors
2. Another ANOVA example: LSmeans, Estimate, Contrast
3. Missing values

1

### Multicenter Depression trial (HAMD) continued

Participants with major depression at 5 centers received an experimental drug (D) or placebo (P). Endpoint was change in Hamilton depression rating scale from baseline to the end of the 9-week treatment.

Were the treatment groups different at baseline?

#### The TTEST Procedure

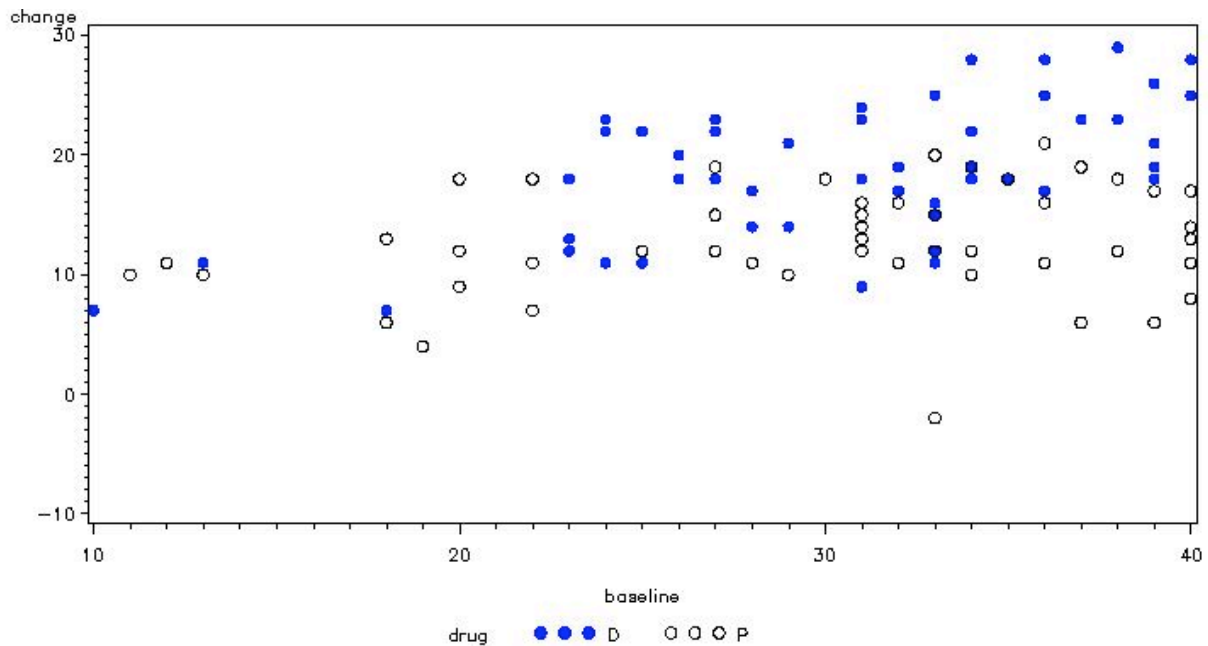
Variable	drug	N	Lower CL		Upper CL		Std Dev	Std Err
			Mean	Mean	Mean	Mean		
baseline	D	50	28.612	30.5	32.388	6.6432	0.9395	
baseline	P	50	27.497	29.82	32.143	8.1734	1.1559	
baseline	Diff (1-2)		-2.276	0.68	3.6359	7.4477	1.4895	

Variable	Method	Variances	DF	t Value	Pr >  t
baseline	Pooled	Equal	98	0.46	0.6490
baseline	Satterthwaite	Unequal	94.1	0.46	0.6491

2

Was there an effect of baseline on the size of the change?



3

```
Goptions reset=all vsize=4in ftext=simplex device=gif lfactor=2
noborder gsfname=graphout gsfmode= replace;
filename graphout "C: ... HAMD03.gif";
```

```
symbol1 value=dot color=blue h=0.85; value (v) = plotting character
```

```
symbol2 value=circle color=black h=1 ; height (h) gives character height
```

```
Proc Gplot data=ph6470.hamd2;
plot change * baseline = drug;
```

Drug values are D and P.

Alphabetically, D comes first and so gets symbol1.

4

## General Linear Model: both continuous and categorical predictors

An advantage of Proc GLM is that it handles both continuous and categorical predictors, and interactions, with a simple notation.

```
Proc GLM data=ph6470.hamd2;
  class drug center;
  model change =
    baseline center drug center*drug drug*baseline / solution;
```

Proc GLM creates the necessary indicator variables for the categorical variables and for the interactions.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
baseline	1	258.1363512	258.1363512	13.28	0.0005
center	4	17.9698453	4.4924613	0.23	0.9203
drug	1	8.3260953	8.3260953	0.43	0.5146
drug*center	4	259.9092872	64.9773218	3.34	0.0135
baseline*drug	1	80.4379038	80.4379038	4.14	0.0450

5

Parameter		Estimate	Standard Error	t Value	Pr >  t
Intercept		7.66883781 B	2.70462990	2.84	0.0057
baseline		0.10729512 B	0.08128172	1.32	0.1902
center	1	0.96968565 B	2.19621794	0.44	0.6599
center	2	5.14296429 B	2.48512246	2.07	0.0414
center	3	2.48723240 B	2.20038999	1.13	0.2614
center	4	1.91531659 B	2.38841382	0.80	0.4248
center	5	0.00000000 B	.	.	.
drug	D	0.66487374 B	4.69496468	0.14	0.8877
drug	P	0.00000000 B	.	.	.
drug*center	D 1	-0.53646056 B	3.06416965	-0.18	0.8614
drug*center	D 2	-10.87981194 B	3.55467935	-3.06	0.0029
drug*center	D 3	-3.36353166 B	2.98152048	-1.13	0.2623
drug*center	D 4	-1.95995181 B	3.26245873	-0.60	0.5495
drug*center	D 5	0.00000000 B	.	.	.
drug*center	P 1	0.00000000 B	.	.	.
drug*center	P 2	0.00000000 B	.	.	.
drug*center	P 3	0.00000000 B	.	.	.
drug*center	P 4	0.00000000 B	.	.	.
drug*center	P 5	0.00000000 B	.	.	.
baseline*drug	D	0.27115038 B	0.13331367	2.03	0.0450
baseline*drug	P	0.00000000 B	.	.	.

What is the slope for baseline in each group? Significantly different?

6

lsmeans drug / pdiff stderr E ; asks for values used to compute prediction

	change	Standard	H0:LSMEAN=0	H0:LSMean1=
drug	LSMEAN	Error	Pr >  t	LSMean2
D	18.5027164	0.6542649	<.0001	Pr >  t
P	13.0078985	0.6612902	<.0001	<.0001

This is slightly different from LSmeans fitted without baseline:

	change	Standard	H0:LSMEAN=0	H0:LSMean1=
drug	LSMEAN	Error	Pr >  t	LSMean2
D	18.5399711	0.6983708	<.0001	Pr >  t
P	12.9308425	0.7032031	<.0001	<.0001

7

#### Coefficients for drug Least Square Means

Effect	drug Level	
	D	P
Intercept	1	1
baseline	30.16	30.16
center	1	0.2
center	2	0.2
center	3	0.2
center	4	0.2
center	5	0.2
drug	D	1
drug	P	0
drug*center	D 1	0
drug*center	D 2	0
drug*center	D 3	0
drug*center	D 4	0
drug*center	D 5	0
drug*center	P 1	0.2
drug*center	P 2	0.2
drug*center	P 3	0.2
drug*center	P 4	0.2
drug*center	P 5	0.2
baseline*drug	D	30.16
baseline*drug	P	0

8

Where does 30.16 come from?

```
The MEANS Procedure  
  
Analysis Variable : baseline  
  
          Mean  
-----  
        30.160000  
-----
```

Just as SAS computes a fitted value with equal weights for each center, it gives both groups the mean value of the baseline.

Rules for constructing LSmeans: SAS Help > SAS Products > SAS/STAT > SAS/STAT User's Guide > Proc GLM > LSMeans Statement

then click on the reference to: Construction of Least-Squares Means

9

### **Another two-factor ANOVA example**

Researchers at Brigham and Women's Hospital in Boston introduced the Physical Capacity Evaluation (PCE), a new method for rating physical ability and impairment in the elderly (*Am J Public Health*, 1995; 85:558-560).

Their evaluation combines measures of hand function, grip strength, coordination, flexibility, balance, dressing, and walking. The paper reports results from 289 elderly adults and compares the PCE with an accepted self-report scale. The researchers divided their participants by gender and by age:

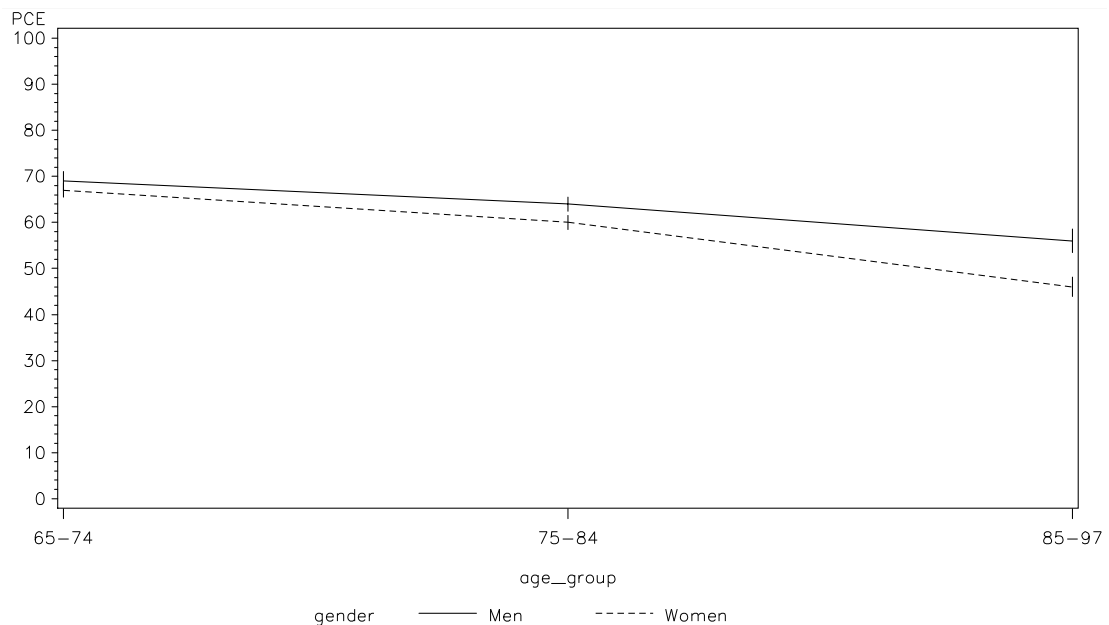
- $n = 89$ , age 65 to 74
- $n = 121$ , age 75 to 84
- $n = 79$ , age 85 to 97

In this example, we are interested in effects of both factors.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
gender	1	1959.91771	1959.91771	12.19	0.0006
age_group	2	12129.00755	6064.50377	37.73	<.0001
gender*age_group	2	710.38038	355.19019	2.21	0.1116

Level of gender	Level of age_group	N	Mean	Std Dev
Men	65-74	38	69.000000	13.000000
Men	75-84	61	64.000000	12.000000
Men	85-97	34	56.000000	15.000000
Women	65-74	51	67.000000	11.000000
Women	75-84	60	60.000000	12.000000
Women	85-97	45	46.000000	14.000000

11



“Men outperformed women in tasks involving strength (grip, walk, foot tap, tandem stand). Age effects were most pronounced among women in the two oldest groups.”

12

## What is the difference between men's and women's PCE scores?

```
Proc GLM data=a;
class gender age_group;
model pce = gender age_group gender*age_group;
means gender gender*age_group ;
LSmeans gender / pdiff stderr ;
```

Level of		-----PCE-----		
gender	N	Mean	Std Dev	
Men	133	63.3834586	13.8699161	
Women	156	58.2500000	14.8130827	

gender	PCE LSMEAN	Standard Error	H0:LSMEAN=0 Pr >  t	H0:LSMean1= LSMean2 Pr >  t
Men	63.0000000	1.1349188	<.0001	0.0006
Women	57.6666667	1.0221150	<.0001	

Slightly different means, so means and LSmeans are doing different calculations.

13

We can reproduce the means results with a two-sample *t*-test:

```
Proc Ttest ci=none data=a;
class gender;
var PCE;
```

### The TTEST Procedure

Variable	gender	N	Lower CL Mean	Mean	Upper CL Mean	Std Dev	Std Err
PCE	Men	133	61.004	63.383	65.762	13.87	1.2027
PCE	Women	156	55.907	58.25	60.593	14.813	1.186
PCE	Diff (1-2)		1.7914	5.1335	8.4755	14.387	1.698

Variable	Method	Variances	DF	t Value	Pr >  t
PCE	Pooled	Equal	287	3.02	0.0027
PCE	Satterthwaite	Unequal	284	3.04	0.0026

This gives us an estimated gender difference where every study participant has equal weight.

*Disadvantage:* affected by sample sizes. If we had enrolled a few more of the oldest adults, the estimate would be bigger.

14

To find out what LSmeans is doing, we can ask:

```
LSmeans gender / pdiff stderr E ;
```

Coefficients for gender Least Square Means

Effect	gender Level	
	Men	Women
Intercept	1	1
gender Men	1	0
gender Women	0	1
age_group 65-74	0.33333333	0.33333333
age_group 75-84	0.33333333	0.33333333
age_group 85-97	0.33333333	0.33333333
gender*age_group Men 65-74	0.33333333	0
gender*age_group Men 75-84	0.33333333	0
gender*age_group Men 85-97	0.33333333	0
gender*age_group Women 65-74	0	0.33333333
gender*age_group Women 75-84	0	0.33333333
gender*age_group Women 85-97	0	0.33333333

These “coefficients” multiply the *regression coefficients*, not the means.

15

```
model pce = gender age_group gender*age_group / solution;
```

Parameter		Estimate	Standard Error	t Value	Pr >  t
Intercept		46.00000000 B	1.88994405	24.34	<.0001
gender	Men	10.00000000 B	2.88086584	3.47	0.0006
gender	Women	0.00000000 B	.	.	.
age_group	65-74	21.00000000 B	2.59298184	8.10	<.0001
age_group	75-84	14.00000000 B	2.50016097	5.60	<.0001
age_group	85-97	0.00000000 B	.	.	.
gender*age_group	Men 65-74	-8.00000000 B	3.95991442	-2.02	0.0443
gender*age_group	Men 75-84	-6.00000000 B	3.68962113	-1.63	0.1050
gender*age_group	Men 85-97	0.00000000 B	.	.	.
gender*age_group	Women 65-74	0.00000000 B	.	.	.
gender*age_group	Women 75-84	0.00000000 B	.	.	.
gender*age_group	Women 85-97	0.00000000 B	.	.	.

Men:  $46 + 10 + (21 + 14 + 0 - 8 - 6 + 0)/3 = 63$

Women:  $46 + 0 + (21 + 14 + 0 + 0)/3 = 57.66667$

16

The difference using `means` output is the average of the men less the average of the women.

What is the difference using `LSmeans` ?

Coefficients for gender Least Square Means

Effect		Men	Women
Intercept		1	1
gender	Men	1	0
gender	Women	0	1
age_group	65-74	0.33333333	0.33333333
age_group	75-84	0.33333333	0.33333333
age_group	85-97	0.33333333	0.33333333
gender*age_group	Men 65-74	0.33333333	0
gender*age_group	Men 75-84	0.33333333	0
gender*age_group	Men 85-97	0.33333333	0
gender*age_group	Women 65-74	0	0.33333333
gender*age_group	Women 75-84	0	0.33333333
gender*age_group	Women 85-97	0	0.33333333

Average of the 3 age-group means for men less the average of 3 age-group means for women. Age-groups get equal weight in adjustment for age.

17

### What is the difference between men's and women's PCE scores?

To answer this question, we must decide what to do about the other factor, age.

Ignore age:

`means` : average of males – average of females (participants get equal weights)

$$63.383 - 58.25 = 5.1335 \text{ units PCE}$$

Adjust for age:

`LSmeans` : age-groups get equal weights

$$63.0 - 57.6666667 = 5.33 \text{ units PCE}$$

`LSmeans` gives a fitted value. How about other fitted values?

18

Proc GLM has two statements that compute a linear combination of the regression coefficients  $\mathbf{l}^\top \hat{\boldsymbol{\beta}}$  and test the null hypothesis,  $\mathbf{l}^\top \boldsymbol{\beta} = 0$ .

`estimate` gives  $\mathbf{l}^\top \hat{\boldsymbol{\beta}}$ , its standard error, and  $t$ -test of  $H_0$

```
ESTIMATE "label" term <coefficients> ;
```

“Coefficients” is the vector  $\mathbf{l}$ , ordered as SAS orders the values of the variable(s)

`contrast` gives the  $F$ -test of  $\mathbf{l}^\top \boldsymbol{\beta} = 0$

```
CONTRAST "label" term <coefficients> ;
```

Format of the command is identical.

Less useful than estimate, unless contrast has > 1 df.

19

### Estimate 1: equal weights for each participant (within gender)

```
Proc GLM data=a;
  class gender age_group;
  model pce = gender ; one-factor model
  estimate "Men minus Women 1" gender 1 -1 ;
  contrast "Men minus Women 1" gender 1 -1 ;
```

Estimate	Estimate	Standard Error	t Value	Pr >  t
Parameter				
Men minus Women 1	5.13345865	1.69797086	3.02	0.0027

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Men minus Women 1	1	1891.901547	1891.901547	9.14	0.0027

These tests are equivalent:  $\sqrt{F} = \sqrt{9.14} = 3.02 = t$ , but ESTIMATE gives an estimate.

20

## 2. Difference of averages of age-group means = difference of LSmeans

```
proc glm data=a;
  class gender age_group;
  model pce = gender age_group gender*age_group; full 2-factor model
  ESTIMATE "Men minus Women 2 (equal)" gender 1 -1 / E ;
  same estimate statement, different model; E asks for coefficients
```

Parameter	Estimate	Standard Error	t Value	Pr >  t
Men minus Women 2 (equal)	5.3333333	1.52733749	3.49	0.0006

difference of LSmeans

$$63 - 57.6666667 = 5.33 \text{ units PCE}$$

21

Coefficients for Estimate Men minus Women 2 (equal)

Intercept		0
gender	Men	1
gender	Women	-1
age_group	65-74	0
age_group	75-84	0
age_group	85-97	0
gender*age_group	Men 65-74	0.333333333
gender*age_group	Men 75-84	0.333333333
gender*age_group	Men 85-97	0.333333333
gender*age_group	Women 65-74	-0.333333333
gender*age_group	Women 75-84	-0.333333333
gender*age_group	Women 85-97	-0.333333333

We only specified coefficients for gender—where did the 1/3s come from?

22

According to the SAS Documentation for Proc GLM ESTIMATE:

If you do not specify the [coefficients] for an effect that contains a specified effect, then the elements of the specified effect are equally distributed over the corresponding levels of the higher-order effect.

The interaction `gender*age_group` contains the gender effect.

SAS takes the +1 for male and divides it equally among the 3 `gender*age_group` terms for male: each gets weight 1/3.

Similarly, the -1 for female gives weights -1/3 for the 3 `gender*age_group` terms for female.

This gives equal weight to each `gender*age_group` mean in the overall gender average.

23

### 3. Difference of averages of age-group means, sample weighted

Here are the sample sizes for the age-group means:

Level of age_group	N	-----PCE----- Mean	Std Dev
65-74	89	67.8539326	11.8657745
75-84	121	62.0165289	12.1174691
85-97	79	50.3037975	15.185058

Weights for each term must sum to 1. To avoid rounding error, let SAS divide by the sum of the weights ( $89 + 121 + 79 = 289$ ).

```
Proc GLM data=a;
  class gender age_group;
  model pce = gender age_group gender*age_group;
  ESTIMATE "Men minus Women 3 (sample)"
    gender 289 -289 age_group 0 0 0
    gender*age_group 89 121 79 -89 -121 -79 / divisor=289 E ;
```

24

Use E option to check that the coefficients are correct:

Coefficients for Estimate Men minus Women 3 (sample)

Intercept		0
gender	Men	1
gender	Women	-1
age_group	65-74	0
age_group	75-84	0
age_group	85-97	0
gender*age_group	Men 65-74	0.3079584775
gender*age_group	Men 75-84	0.4186851211
gender*age_group	Men 85-97	0.2733564014
gender*age_group	Women 65-74	-0.307958478
gender*age_group	Women 75-84	-0.418685121
gender*age_group	Women 85-97	-0.273356401

25

#### 4. Difference of averages of age-group means, population weighted

Age-group totals in the US population: (Source: Table 11, *Statistical Abstract of the US: 2006*)

Age group	2004		
	Total	Male	Female
<b>Total</b> . . . . .	<b>293,655</b>	<b>144,537</b>	<b>149,118</b>
Under 5 years . . . . .	20,071	10,263	9,808
5 to 9 years . . . . .	19,606	10,029	9,576
10 to 14 years . . . . .	21,145	10,831	10,314
15 to 19 years . . . . .	20,730	10,635	10,094
20 to 24 years . . . . .	20,971	10,803	10,168
25 to 29 years . . . . .	19,561	9,995	9,566
30 to 34 years . . . . .	20,471	10,341	10,130
35 to 39 years . . . . .	21,052	10,571	10,482
40 to 44 years . . . . .	23,056	11,463	11,593
45 to 49 years . . . . .	22,123	10,918	11,205
50 to 54 years . . . . .	19,496	9,535	9,961
55 to 59 years . . . . .	16,490	8,001	8,488
60 to 64 years . . . . .	12,589	5,998	6,591
65 to 74 years . . . . .	18,463	8,428	10,036
75 to 84 years . . . . .	12,971	5,218	7,753
85 years and over . . . . .	4,860	1,508	3,352
5 to 13 years . . . . .	36,376	18,617	17,759
14 to 17 years . . . . .	16,831	8,625	8,206
18 to 24 years . . . . .	29,245	15,057	14,189
18 years and over . . . . .	220,377	107,032	113,345
55 years and over . . . . .	65,373	29,153	36,220
65 years and over . . . . .	36,294	15,154	21,140
75 years and over . . . . .	17,831	6,726	11,104
Median age (years) . . . . .	36.0	34.7	37.4

As in estimate 3, we want totals:  $18463 + 12971 + 4860 = 36294$

26

```

proc glm data=a;
  class gender age_group;
  model pce = gender age_group gender*age_group;
  ESTIMATE "Men minus Women 4 (US pop)"

      gender 36294 -36294 age_group 0 0 0

      gender*age_group 18463 12971 4860 -18463 -12971 -4860

/ divisor=36294 E ;

```

27

Coefficients for Estimate Men minus Women 4 (US pop)

Intercept		0
gender	Men	1
gender	Women	-1
age_group	65-74	0
age_group	75-84	0
age_group	85-97	0
gender*age_group	Men 65-74	0.5087066733
gender*age_group	Men 75-84	0.3573868959
gender*age_group	Men 85-97	0.1339064308
gender*age_group	Women 65-74	-0.508706673
gender*age_group	Women 75-84	-0.357386896
gender*age_group	Women 85-97	-0.133906431

28

from one-factor ANOVA:

Parameter	Estimate	Standard Error	t Value	Pr >  t
Men minus Women 1	5.13345865	1.69797086	3.02	0.0027

from two-factor ANOVA:

Parameter	Estimate	Standard Error	t Value	Pr >  t
Men minus Women 2 (equal)	5.3333333	1.52733749	3.49	0.0006
Men minus Women 3 (sample)	5.0242215	1.50057623	3.35	0.0009
Men minus Women 4 (US pop)	3.7860252	1.65461462	2.29	0.0229

ESTIMATE computes predicted means using the fitted regression coefficients. Standard error of prediction is based on the Root MSE and correlations among fitted regression coefficients.

Choice of weighting scheme depends on what we want to estimate.

29

## Missing Values

Returning to the multi-center depression study (HAMD), suppose that some of the baseline and final surveys were lost.

Proc GLM ignores any observation with a missing value for any term in the model.

The GLM Procedure

Class Level Information

Class	Levels	Values
drug	2	D P
center	5	1 2 3 4 5

Number of Observations Read	100
Number of Observations Used	89

30

Big problems occur when missing data is missing for reasons related to the value of the missing observation itself or other observations. Suppose patients with the worst side-effects dropped out of the study and didn't fill out the final survey.

This is non-random missing data, the missing process contains information about the response. The missing data cannot be ignored.

Ignoring the missing data and analyzing the complete cases (Proc GLM default) leaves you with a non-representative sample and biased estimates.

31

**Missing completely at random (MCAR):** data are missing independently of both observed and unobserved data.

*Example:* a participant flips a coin to decide whether to complete the depression survey.

**Missing at random (MAR):** given the observed data, data are missing independently of unobserved data.

*Example:* male participants are more likely to refuse to fill out the depression survey, but it does not depend on the level of their depression.

MCAR implies MAR, but not the other way round.

*Statistical Analysis with Missing Data, Second Edition* (2002) by Roderick Little and Donald Rubin

32

If we have MAR, then a complete case analysis will be unbiased but inefficient, because we have lost observations.

Another approach we can use with MAR is **multiple imputation**:

1. Fill in missing values  $M$  times using distribution of observed values, creating  $M$  complete data sets.
2. Analyze each of the  $M$  complete data sets.
3. Combine the results of the  $M$  analyses to draw conclusions.