- Problem Set 2 will be posted late this afternoon and due on Monday
- Problem Set 1 is graded and an answer key is posted (written with optimism at 10:30am)

My name	Other names	R function	faux function
Treatment	Treatment, Dummy, Dummy, Simple	contr.treatment	contr_code_treatment
Deviation	Deviation, Simple, Contrast	contr.treatment - 1/k	contr_code_deviation
Difference	Forward/Backward, Contrast, Repeated	MASS::contr.sdif	contr_code_difference
Sum	Sum, Sum, Deviation, Deviation, Effects	contr.sum	contr_code_sum
Helmert	Reverse Helmert, Reverse Helmert, Difference	contr.helmert / column_i	contr_code_helmert
Polynomial	Polynomial, Orthogonal, Trend	contr.poly	contr_code_poly

1

categorical predictors (part 2: dichotomous predictors, ANOVA, and *t*-tests)

January 31, 2024

2

review 1

- in a design with <u>two</u> groups, we can create <u>one</u> numeric predictor, *X*, to indicate group membership
- for this single predictor, we estimate one parameter, a slope
- the numbers chosen for the groups matter for the interpretation of the slope parameter estimate

3

review 2

- generally, we will prefer contrast codes to indicate group membership
- contrast codes are those that sum to zero
- for two-group designs, $X = \pm½$ will have some nice benefits for interpretation of the associated slope

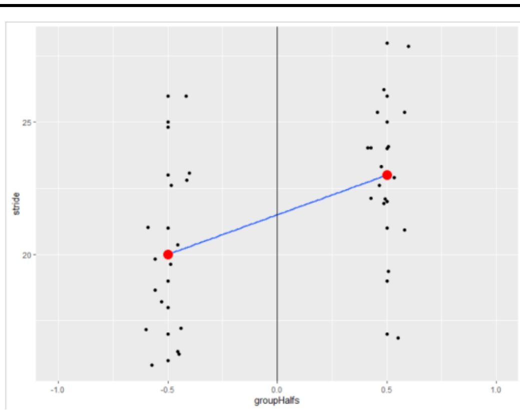
4

what does X = ±½ buy us?

- data from Monday ($M_{exercise} = 23$, $M_{control} = 20$)
  - (assign the positive value to higher mean to get a positive slope)

```
Coefficients:
            Estimate      SE     t Pr(>|t|)
(Intercept)  21.5000  0.5244 41.00  < 2e-16 ***
groupHalfs    3.0000  1.0488  2.86  0.00763 **
```

5



6

one other common coding scheme: *dummy codes*

- if one group is given a value of X = 0 and the other a value of $X = 1$, this is called *dummy coding*
- this is R's default if you don't create your own X
- typically 0 is assigned to a control/comparison/ reference group (but R will assign it alphabetically)
- what does dummy-coding buy us?

```
             Estimate      SE     t Pr(>|t|)
(Intercept)  20.0000  0.7416 26.97  < 2e-16 ***
groupDummy    3.0000  1.0488  2.86  0.00763 **
```

7

building an ANOVA summary table

$$SSE(C) = \sum (Y_i - \hat{Y}_{iC})^2$$

$$SSE(C) = \sum (Y_i - \bar{Y})^2$$

often called SS_{total}

8

building an ANOVA summary table

$$SSE(A) = \sum (Y_i - \hat{Y}_{iA})^2$$

$$SSE(A) = \sum (Y_i - \bar{Y}_{group})^2$$

many aliases: SS_{residual}, SS_{within}, SS_{error}

9

building an ANOVA summary table

$$SSR = \sum_i (\hat{Y}_{iA} - \hat{Y}_{iC})^2$$

$$SSR = \sum_i (\bar{Y}_{group} - \bar{Y})^2$$

$$SSR = \sum_k n_k (\bar{Y}_{group} - \bar{Y})^2$$

many aliases: $SS_{reduction}$, $SS_{between}$, $SS_{regression}$

10

the ANOVA table

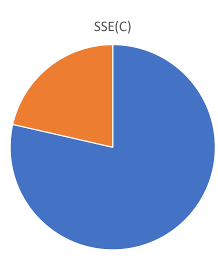
source	SS	df	MS	F	PRE
reduction	SSR	PA – PC	MSR	MSR/MSE(A)	SSR/SSE(C)
error	SSE(A)	n – PA	MSE(A)		
total	SSE(C)	n – 1	MSE(C)		
reduction	72	1	72	8.18	.214
error	264	30	8.8		
total	336	31			

```
> anova(modelBest)
Analysis of Variance Table

Response: stride
           Df Sum Sq Mean Sq F value   Pr(>F)
groupHalfs  1     72    72.0  8.1818 0.007633 **
Residuals  30    264     8.8
```

11

the SS values, graphically



12

a confidence interval for β_1

$$b \pm \sqrt{\frac{F_{crit} \times MSE}{(n-1)s_X^2 \times tolerance}}$$

$$3 \pm \sqrt{\frac{4.17 \times 8.8}{31 \times 0.26 \times 1}} = 3 \pm 2.14 = [0.86, 5.14]$$

- note that this is also the CI for $\mu_1 - \mu_2$
- if the CI does not contain 0, H_0 can be rejected

13

digression: the independent-samples *t*-test

- to test the null hypothesis

$$H_0: \mu_1 = \mu_2$$

- we can calculate t via the following steps
  1. find pooled variance
  2. find *SE* of the difference between two independent means
  3. find t

14

pooled variance

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

$$\frac{15 \times 10.667 + 15 \times 6.933}{30} = 8.8$$

15

SE

$$SE = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$SE = \sqrt{\frac{8.815}{16} + \frac{8.815}{16}} \approx 1.05$$

16

t

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{SE}$$

$$t = \frac{23 - 20}{1.05} = 2.86$$

17

an alternative (and popular) effect-size measure: Cohen's *d*

- R^2 (*PRE*) measures proportion of variance accounted for by Model A not accounted for by Model C
- Cohen's *d* is a measure of the difference between two means, standardized

$$d = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_p^2}}$$

$$d = \frac{23 - 20}{\sqrt{8.815}} = 1.01$$

18

what does Cohen's *d* indicate?

- how much difference there is between two group means, counting in *SD*s
  - (very much like a *z*-score)
- alternatively, it's an estimate of how different the average individual in one population is from an average individual in another population (in *SD*s)

$$d \hat{=} \delta$$

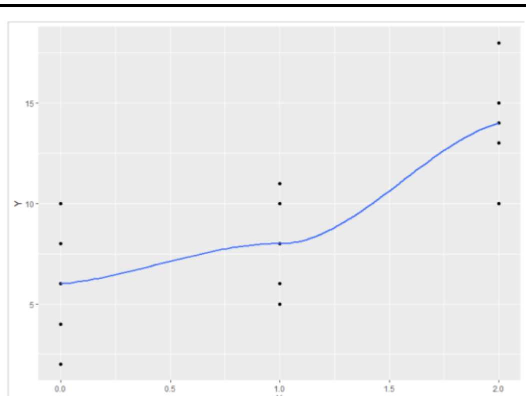
- under duress, Cohen suggested using .2, .5, and .8 as guidelines for small, medium, and large effects

19

getting ready for >2 groups

- the first lesson to learn: creating one *X* ***won't*** suffice
- let's try
- I have a data set with 3 groups, and I assigned values of X = 1, 2, and 3 to them, respectively

20



21

using one *X* for >2 groups will usually induce nonlinearity

- we will need $m - 1$ new variables to numerically code our m groups
- the numbers we choose to indicate group membership will depend on what we want our slopes to tell us (among other constraints)

22

we need more *X*s

- if we have m groups, we need $m - 1$ predictors (*X*s), ***no more, no less***
- the predictors should be contrast codes

$$\sum \lambda_k = 0$$

- in addition to using contrasts codes, the contrasts should be *orthogonal* (independent)

23

orthogonality

- defined mathematically

$$\sum \lambda_{1k} \lambda_{2k} = 0$$

- what?!
- let's look at some *X*s for a three-group design and check for orthogonality

24

three (m = 3) groups → two Xs, with the value of λ assigned to each

group	λ_1	λ_2	$\lambda_1\lambda_2$
A	1	0	
B	0	1	
C	0	0	

25

three (m = 3) groups → two Xs, with the value of λ assigned to each

group	λ_1	λ_2	$\lambda_1\lambda_2$
A	1	0	
B	0	1	
C	-1	-1	

26

three (m = 3) groups → two Xs, with the value of λ assigned to each

group	λ_1	λ_2	$\lambda_1\lambda_2$
A	2	0	
B	-1	1	
C	-1	-1	

27