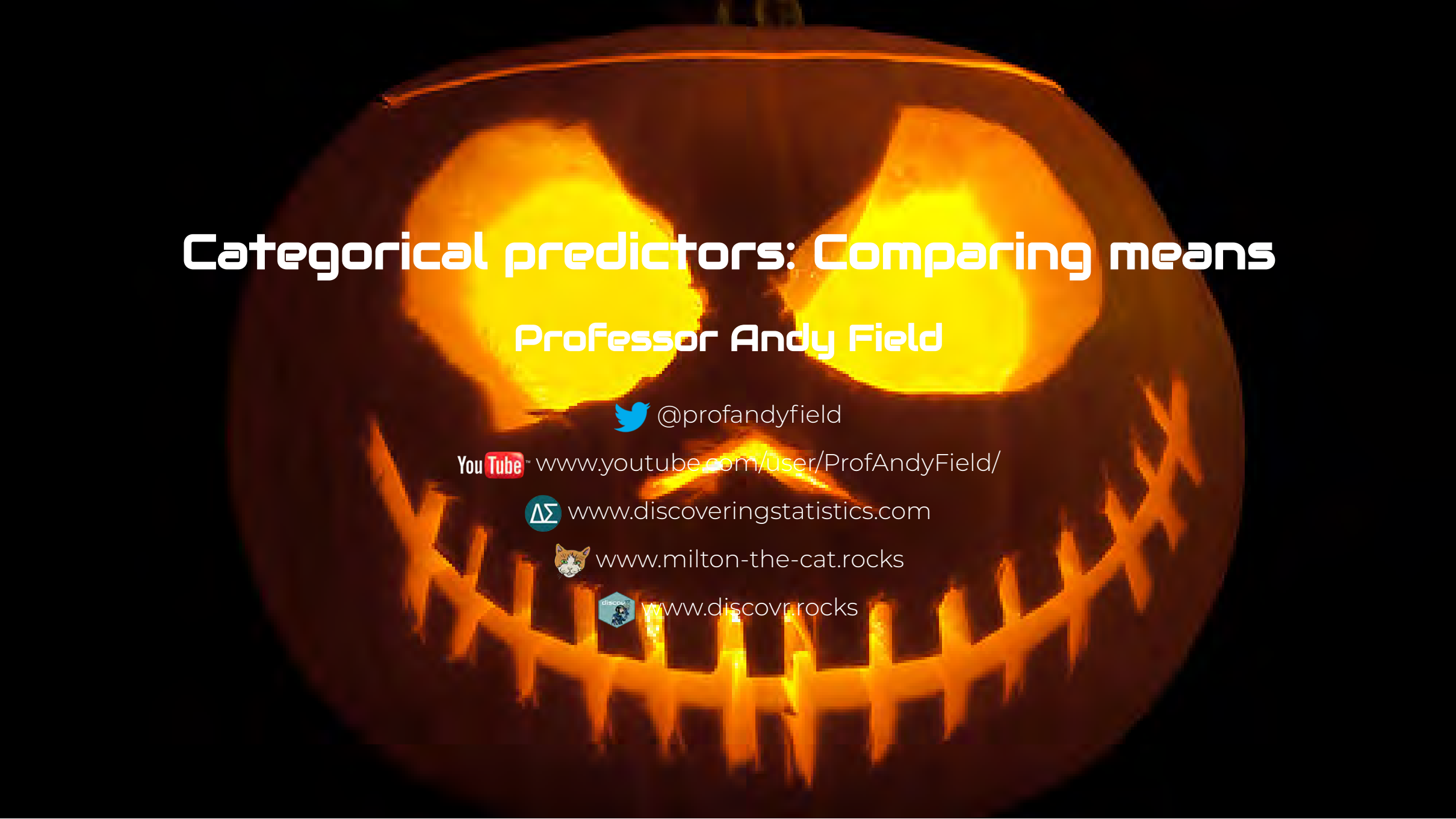


Categorical predictors: Comparing means

Professor Andy Field

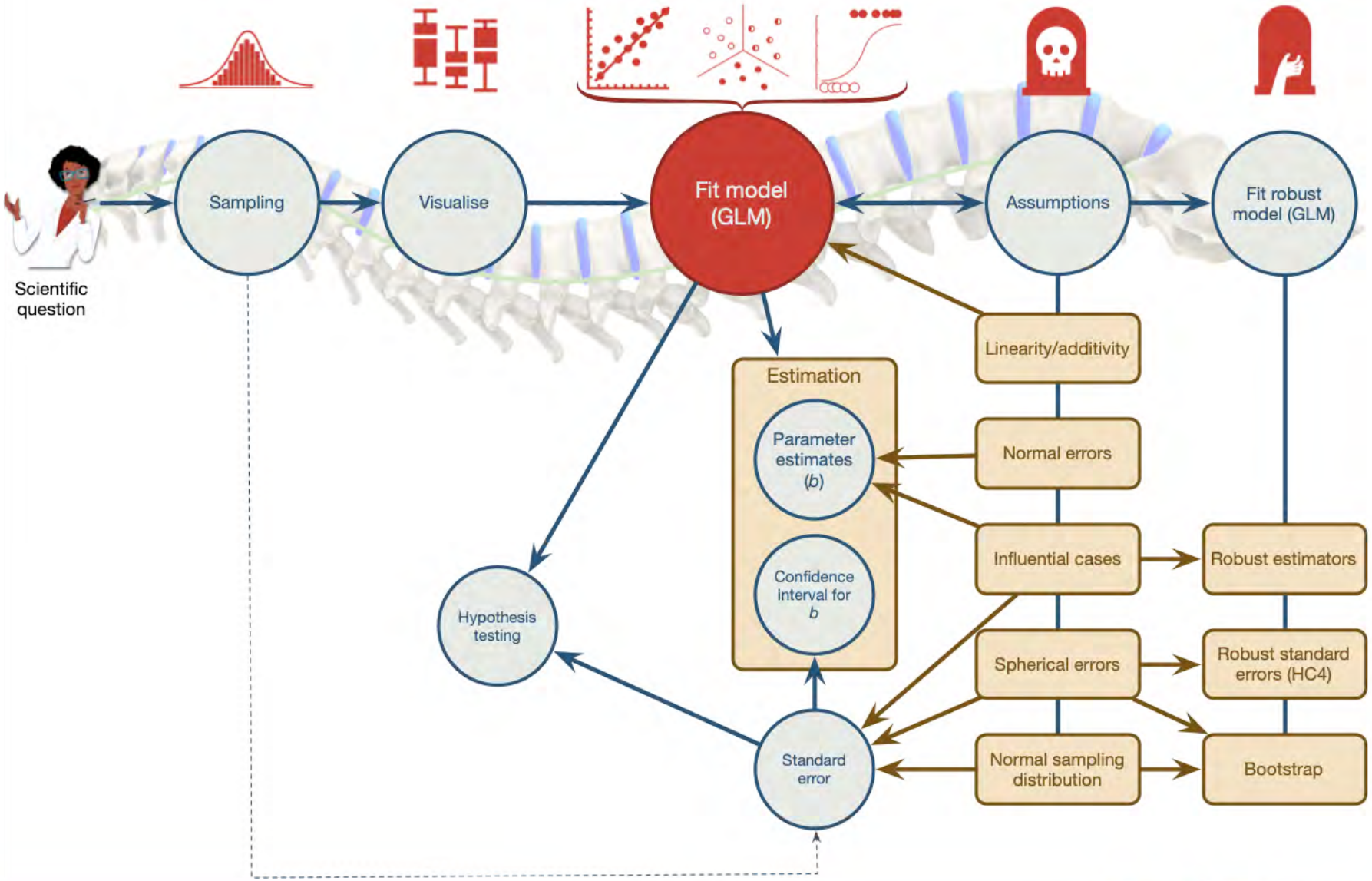
@profandyfield

www.youtube.com/user/ProfAndyField/

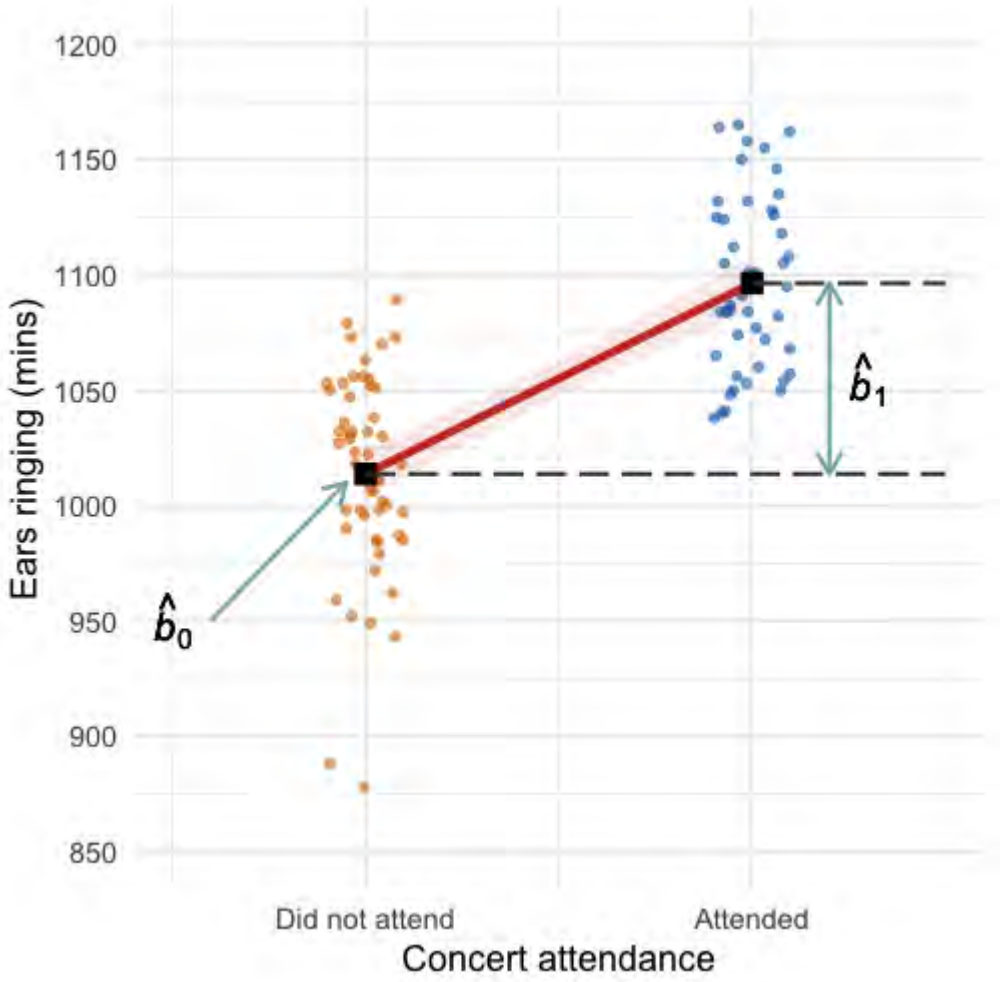
www.discoveringstatistics.com

www.milton-the-cat.rocks

www.discovr.rocks


Scientific question
Sampling
Visualise
Fit model (GLM)
Assumptions
Fit robust model (GLM)
Linearity/additivity
Estimation
Parameter estimates (b)
Confidence interval for b
Normal errors
Influential cases
Robust estimators
Hypothesis testing
Spherical errors
Robust standard errors (HC4)
Standard error
Normal sampling distribution
Bootstrap

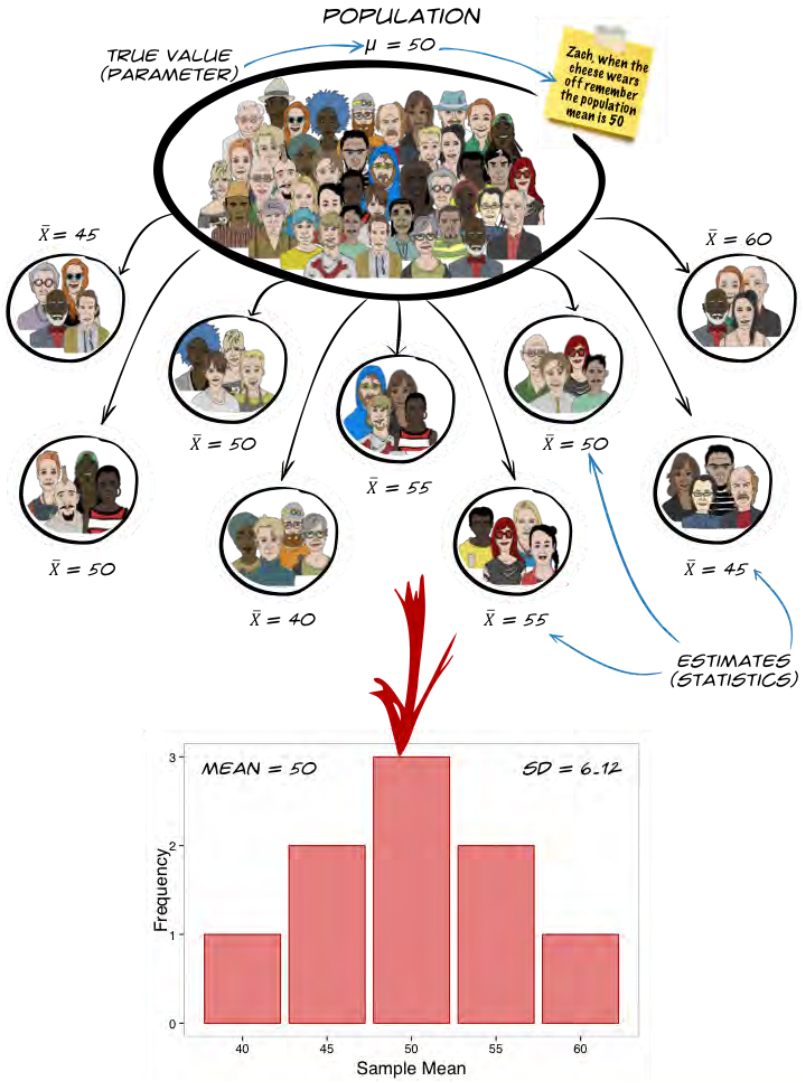
$$\hat{\text{ringing}}_i = \hat{b}_0 + \hat{b}_1 \text{attendance}_i + e_i$$



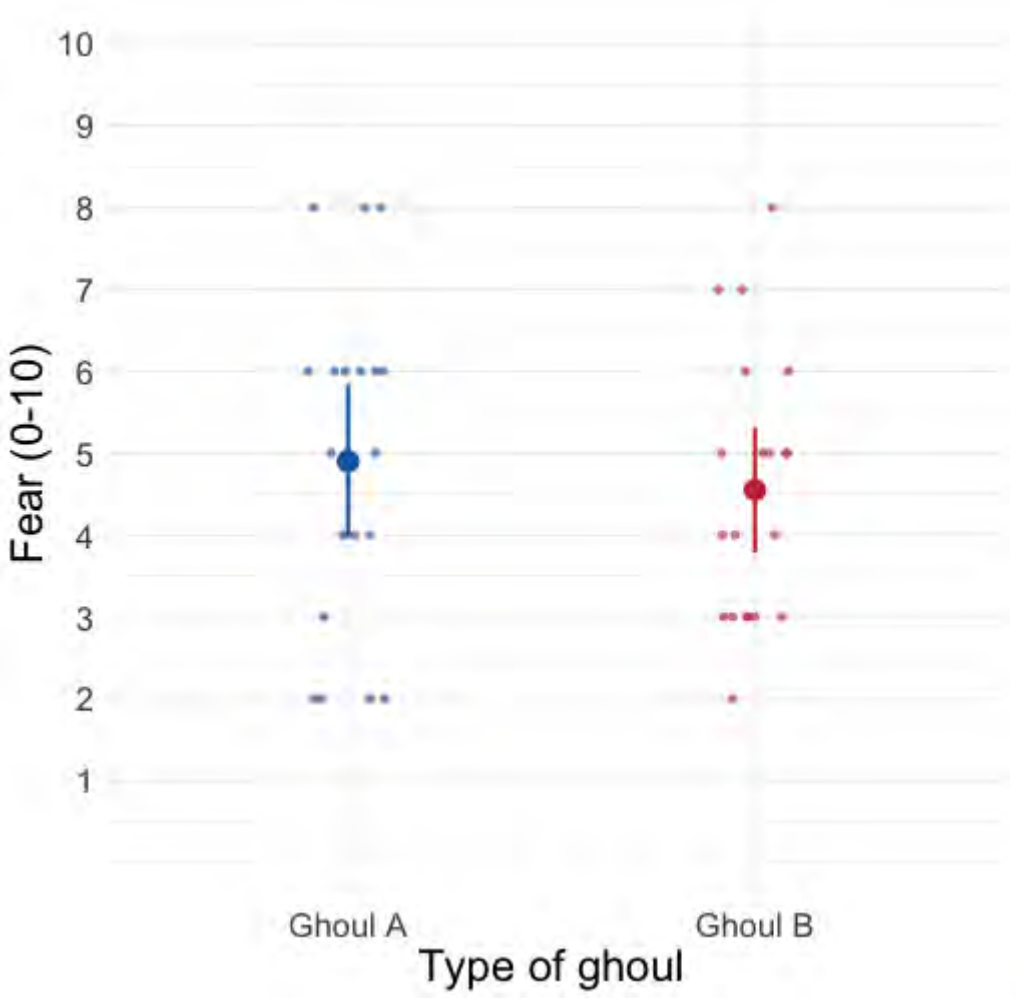
The GLM and experiments

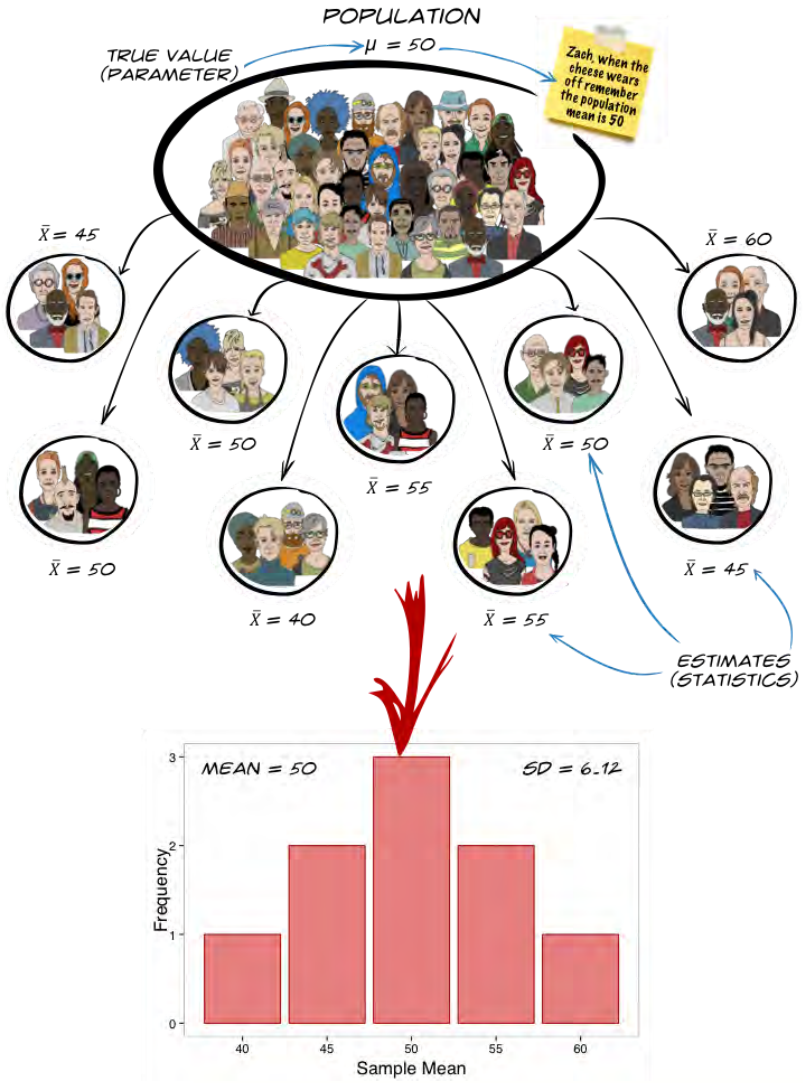
- In experimental research, predictors in the linear model are defined by a manipulation
  - By manipulating a predictor variable can we cause (and therefore predict) a change in behaviour?
- The *F*-statistic
  - Still quantifies the fit of the model to the data
  - Still has an associated significance test
  - The 'fit' represents the experimental manipulation (which defines the predictor)
  - 'Significant' fit equates to a 'significant' effect of the experimental manipulation

A ghoulish example


POPULATION
TRUE VALUE (PARAMETER)
μ = 50
Zach, when the cheese wears off remember the population mean is 50
X̄ = 45
X̄ = 60
X̄ = 50
X̄ = 50
X̄ = 55
X̄ = 50
X̄ = 50
X̄ = 40
X̄ = 55
X̄ = 45
ESTIMATES (STATISTICS)
MEAN = 50
SD = 6.12
Frequency
Sample Mean

Same ghoul


Fear (0-10)
Ghoul A
Ghoul B
Type of ghoul



Different ghouls

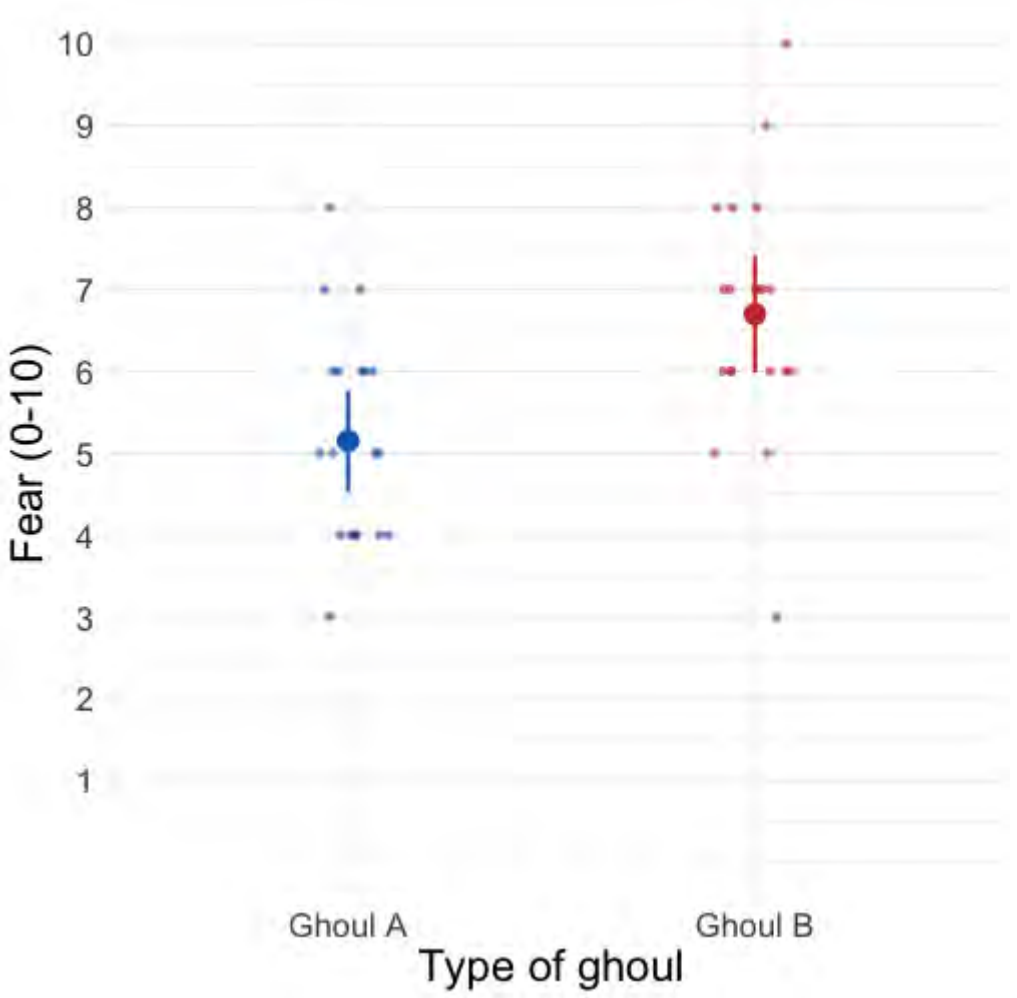


Table 1: Data for the zombie vs. werewolf costume experiment

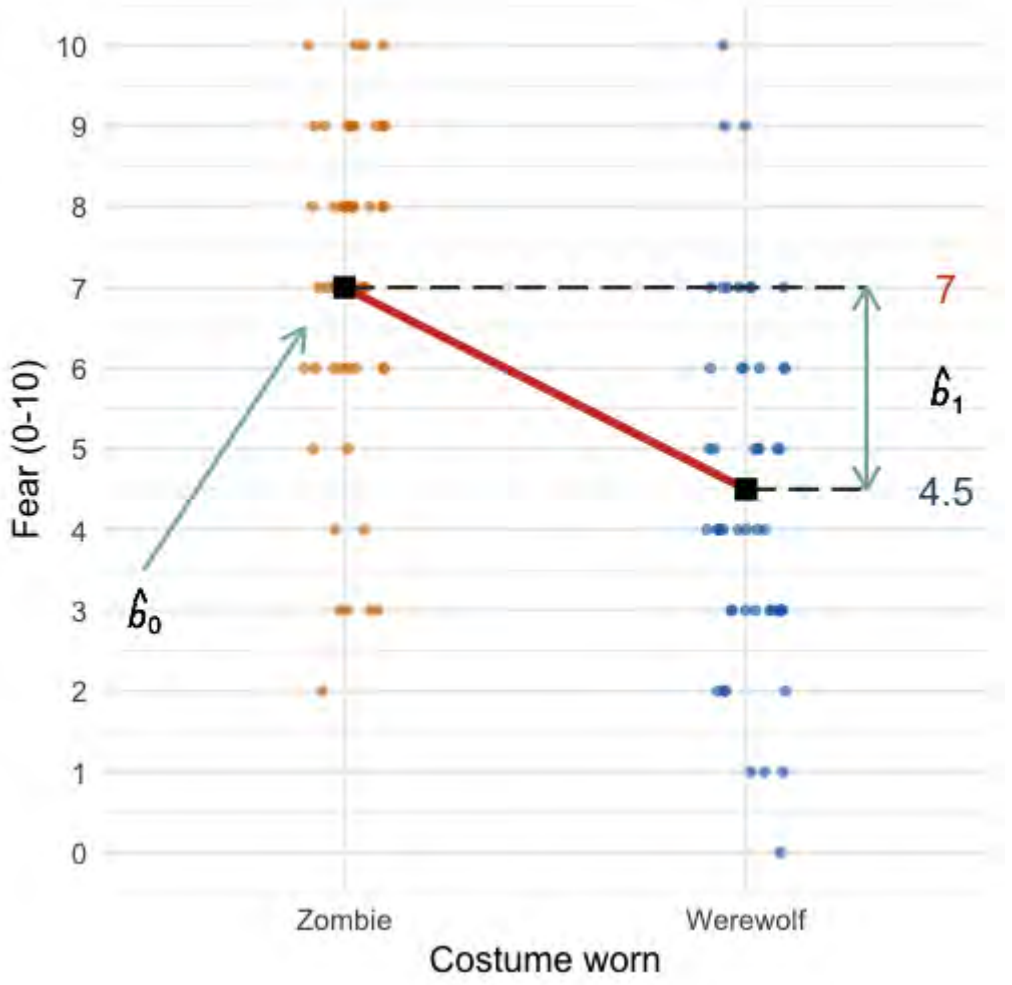
ID	Costume	Fear	Numeric code
1	Zombie	6	0
2	Zombie	8	0
3	Zombie	5	0
4	Zombie	6	0
5	Zombie	8	0
6	Zombie	8	0
7	Zombie	7	0
8	Zombie	10	0
9	Zombie	7	0
10	Zombie	3	0

Previous 1 2 3 4 5 ... 10 Next

The model

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i + e_i$$

$$\hat{\text{Fear}}_i = \hat{b}_0 + \hat{b}_1 \text{Costume}_i + e_i$$



Dummy coding: b_0

- Dummy Coding
  - Zombie = 0, Werewolf = 1
- When costume = zombie
  - Costume = 0
  - Predicted fear = mean of zombie group:

$$\hat{\text{fear}}_i = \hat{b}_0 + \hat{b}_1 \text{Costume}_i$$
$$\bar{X}_{\text{zombie}} = \hat{b}_0 + \hat{b}_1 \times 0$$
$$\bar{X}_{\text{zombie}} = \hat{b}_0$$

Dummy coding: b_1

- When costume = werewolf
  - Costume = 1
  - Predicted fear = mean of werewolf group:

$$\hat{\text{fear}}_i = \hat{b}_0 + \hat{b}_1 \text{Costume}_i$$
$$\bar{X}_{\text{werewolf}} = \hat{b}_0 + \hat{b}_1 \times 1$$
$$\bar{X}_{\text{werewolf}} = \hat{b}_0 + \hat{b}_1$$

$$\bar{X}_{\text{werewolf}} = \bar{X}_{\text{zombie}} + \hat{b}_1$$
$$\hat{b}_1 = \bar{X}_{\text{werewolf}} - \bar{X}_{\text{zombie}}$$

The linear model

- We can fit a linear model with fear as the outcome and the type of costume (zombie or werewolf) as the predictor, note:
  - Intercept (b_0) is the mean of 'zero coded' group
  - b for the dummy variable is the difference between the means of the two costume groups (4.5 − 7 = −2.5)

```
zombie_lm <- lm(Fear ~ Costume, zombie_tib)
broom::tidy(zombie_lm, conf.int = TRUE)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.0	0.299	23.400	0	6.406	7.594
CostumeWerewolf	-2.5	0.423	-5.909	0	-3.340	-1.660

Another ghoulish example

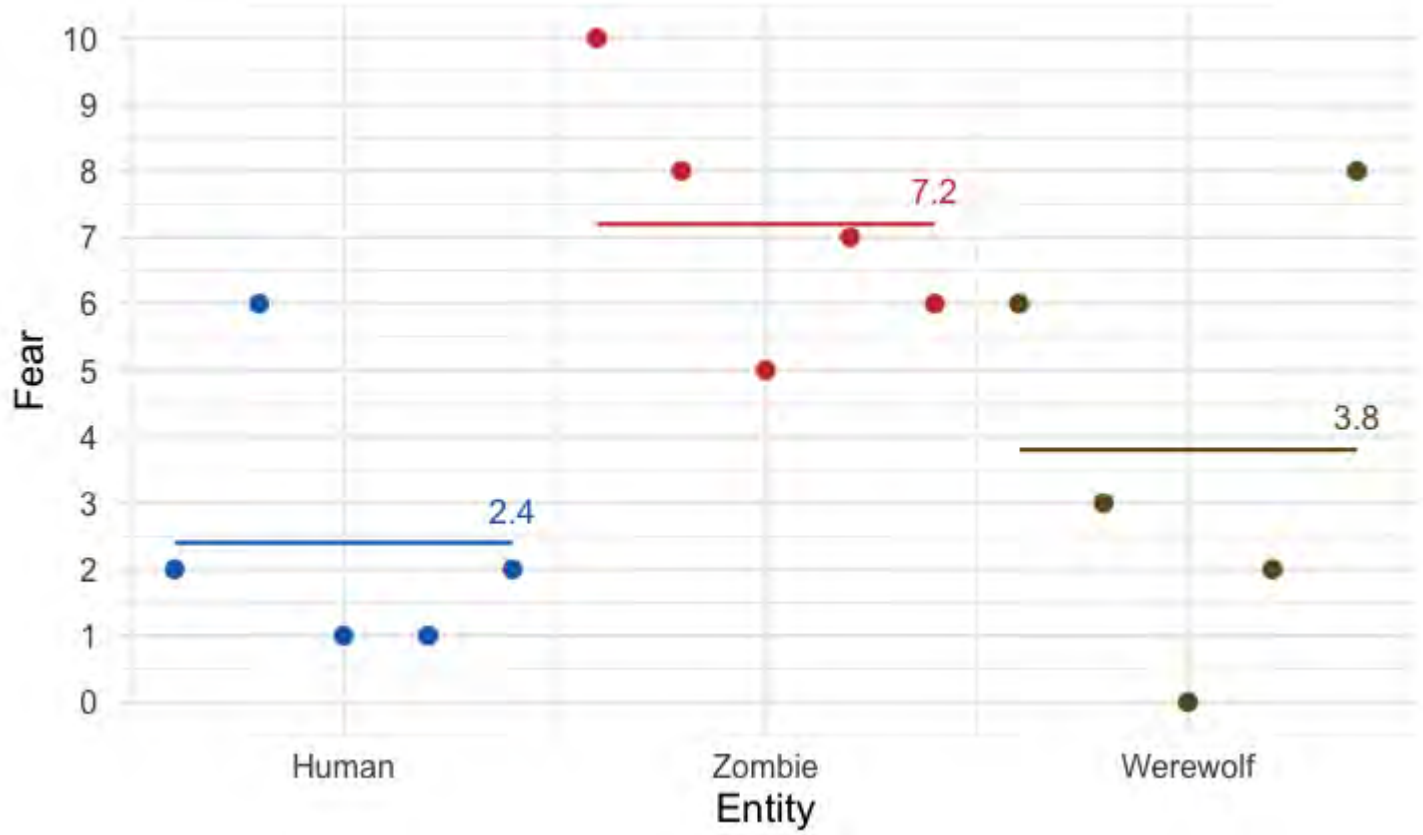
- Which is more scary?
  - Human
  - Zombie
  - Werewolf
- Design
  - Prank
- Outcome
  - Fear

The data

	Human	Zombie	Werewolf
	2	10	6
	6	8	3
	1	5	0
	1	7	2
	2	6	8
Mean	2.40	7.20	3.80
Variance (s^2)	4.30	3.70	10.20
Standard deviation (s)	2.07	1.92	3.19

$$\text{Overall mean } (\bar{X}_{\text{grand}}) = 4.47$$

The model



Dummy coding multiple categories

- You can code any categorical predictor into a series of dummy variables
  - Dummy variables must be entered in the same block
  - Choose a baseline category - it is always coded as 0
  - (By default R chooses the first level of the factor)
  - The *b* for each dummy variable will be the difference in means between each category and the baseline

Entity	Dummy 1 (Zombie vs. Human)	Dummy 1 (Werewolf vs. Human)
Human	0	0
Zombie	1	0
Werewolf	0	1

$$\hat{\text{Fear}}_i = \hat{b}_0 + \hat{b}_1 \text{Zombie vs. Human}_i + \hat{b}_2 \text{Werewolf vs. Human}_i + e_i$$

Dummy coding: b_0

- When entity = human
  - Zombie vs. Human = 0
  - Wolf vs. Human = 0
- Predicted fear = mean of human group:

Entity	Zombie vs. Human	Werewolf vs. Human
Human	0	0
Zombie	1	0
Werewolf	0	1

$$\hat{\text{Fear}}_i = \hat{b}_0 + \hat{b}_1 \text{Zombie vs. Human}_i + \hat{b}_2 \text{Werewolf vs. Human}_i + e_i$$

$$\bar{X}_{\text{human}} = \hat{b}_0 + \hat{b}_1 \times 0 + \hat{b}_2 \times 0$$

$$\hat{b}_0 = \bar{X}_{\text{human}}$$

Dummy coding: 4

- When entity = zombie
  - Zombie vs. Human = 1
  - Wolf vs. Human = 0
- Predicted fear = mean of zombie group:

Entity	Zombie vs. Human	Werewolf vs. Human
Human	0	0
Zombie	1	0
Werewolf	0	1

$$\hat{\text{Fear}}_i = \hat{b}_0 + \hat{b}_1 \text{Zombie vs. Human}_i + \hat{b}_2 \text{Werewolf vs. Human}_i + e_i$$

$$\bar{X}_{\text{zombie}} = \hat{b}_0 + \hat{b}_1 \times 1 + \hat{b}_2 \times 0$$

$$\bar{X}_{\text{zombie}} = \hat{b}_0 + \hat{b}_1$$

$$\hat{b}_1 = \bar{X}_{\text{zombie}} - \hat{b}_0$$

$$= \bar{X}_{\text{zombie}} - \bar{X}_{\text{human}}$$

Dummy coding: b_2

- When entity = werewolf
  - Zombie vs. Human = 0
  - Wolf vs. Human = 1
- Predicted fear = mean of werewolf group:

Entity	Zombie vs. Human	Werewolf vs. Human
Human	0	0
Zombie	1	0
Werewolf	0	1

$$\hat{\text{Fear}}_i = \hat{b}_0 + \hat{b}_1 \text{Zombie vs. Human}_i + \hat{b}_2 \text{Werewolf vs. Human}_i + e_i$$

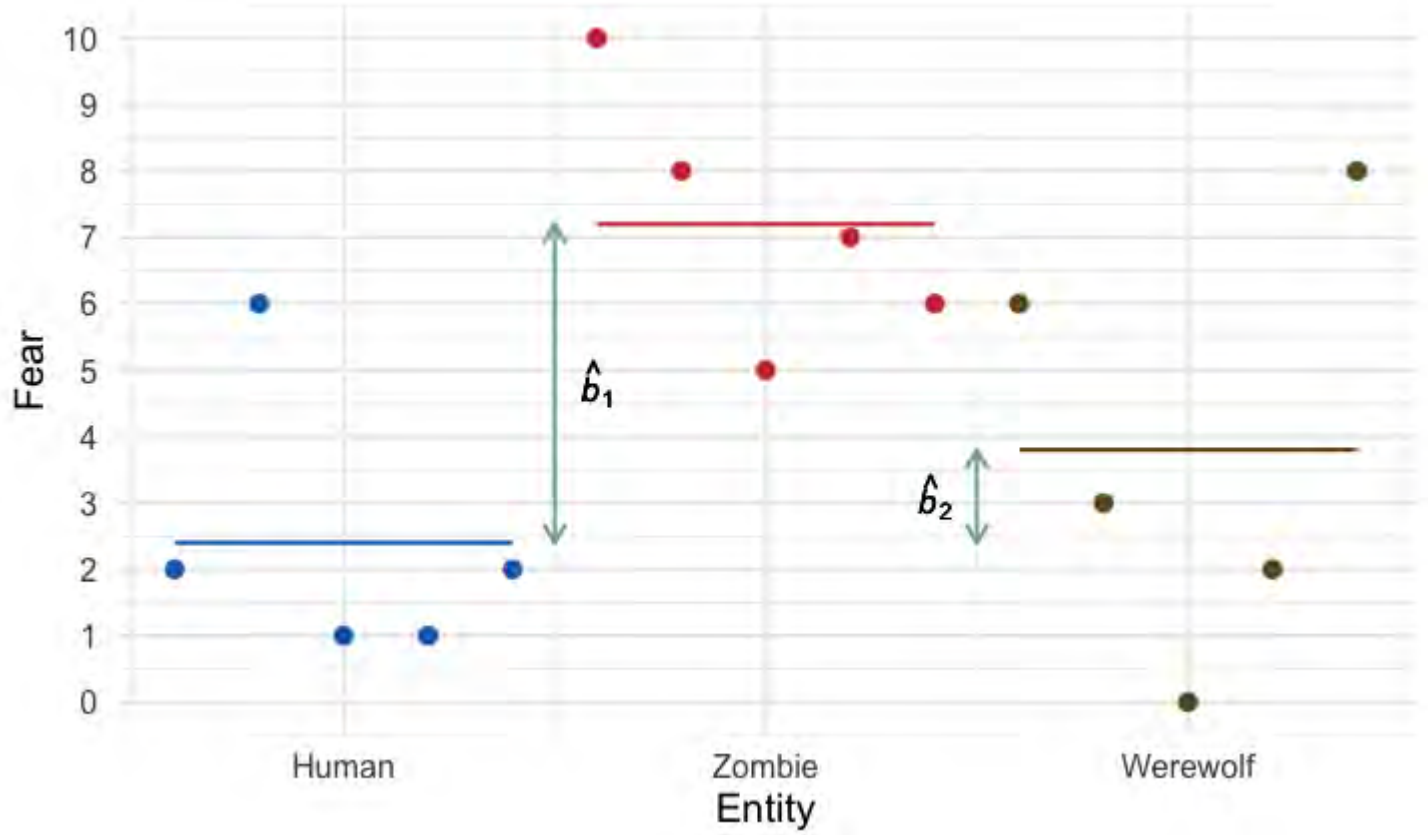
$$\bar{X}_{\text{werewolf}} = \hat{b}_0 + \hat{b}_1 \times 0 + \hat{b}_2 \times 1$$

$$\bar{X}_{\text{werewolf}} = \hat{b}_0 + \hat{b}_2$$

$$\hat{b}_2 = \bar{X}_{\text{werewolf}} - \hat{b}_0$$

$$= \bar{X}_{\text{werewolf}} - \bar{X}_{\text{human}}$$

The model


10
9
8
7
6
5
4
3
2
1
0
Fear
$\hat{b}_1$
$\hat{b}_2$
Human
Zombie
Werewolf
Entity

The parameter values

	Human	Zombie	Werewolf
	2	10	6
	6	8	3
	1	5	0
	1	7	2
	2	6	8
Mean	2.40	7.20	3.80

$$\hat{b}_0 = \bar{X}_{\text{human}} = 2.40$$

$$\hat{b}_1 = \bar{X}_{\text{zombie}} - \bar{X}_{\text{human}} = 7.20 - 2.40 = 4.80$$

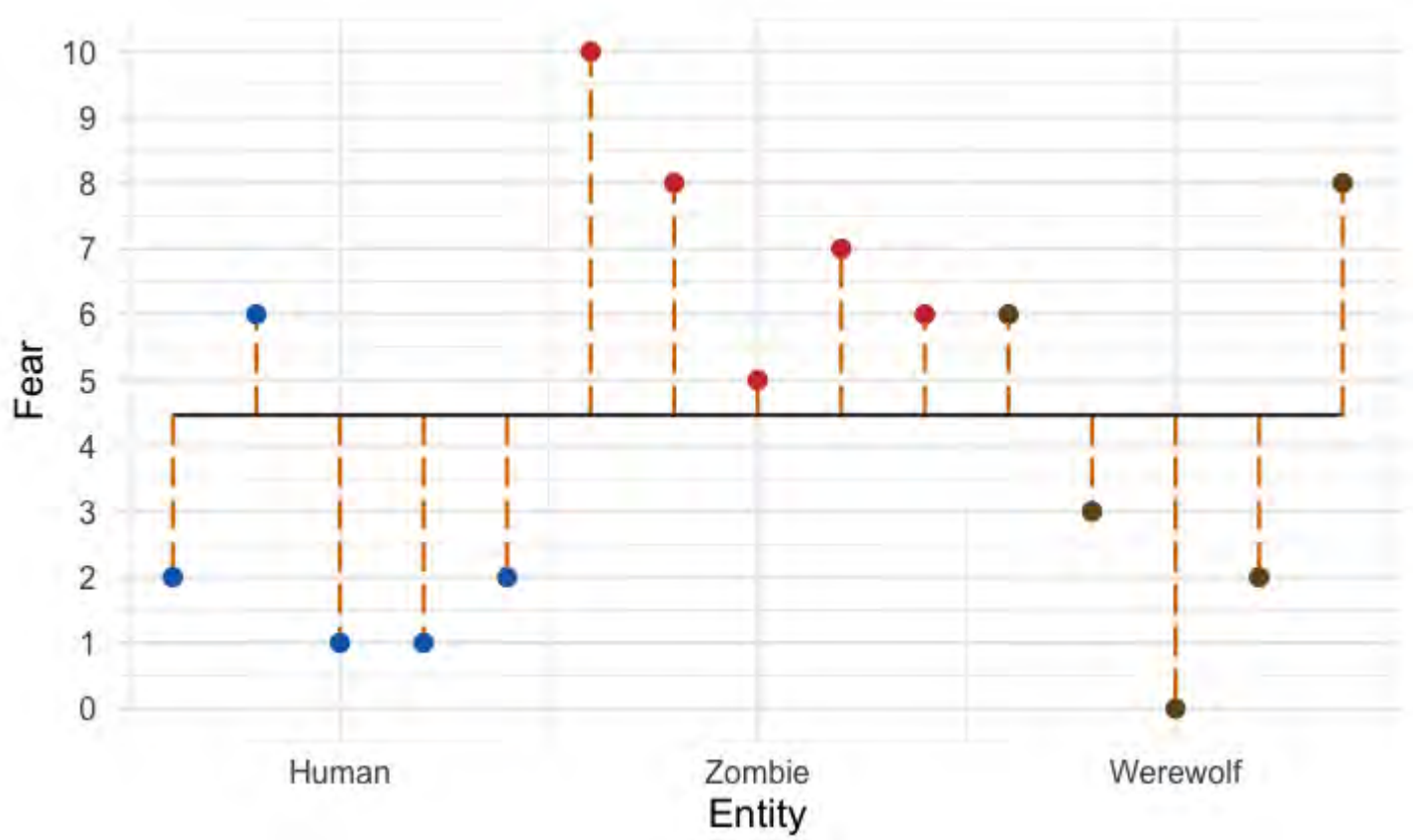
$$\hat{b}_2 = \bar{X}_{\text{werewolf}} - \bar{X}_{\text{human}} = 3.80 - 2.40 = 1.40$$

The model

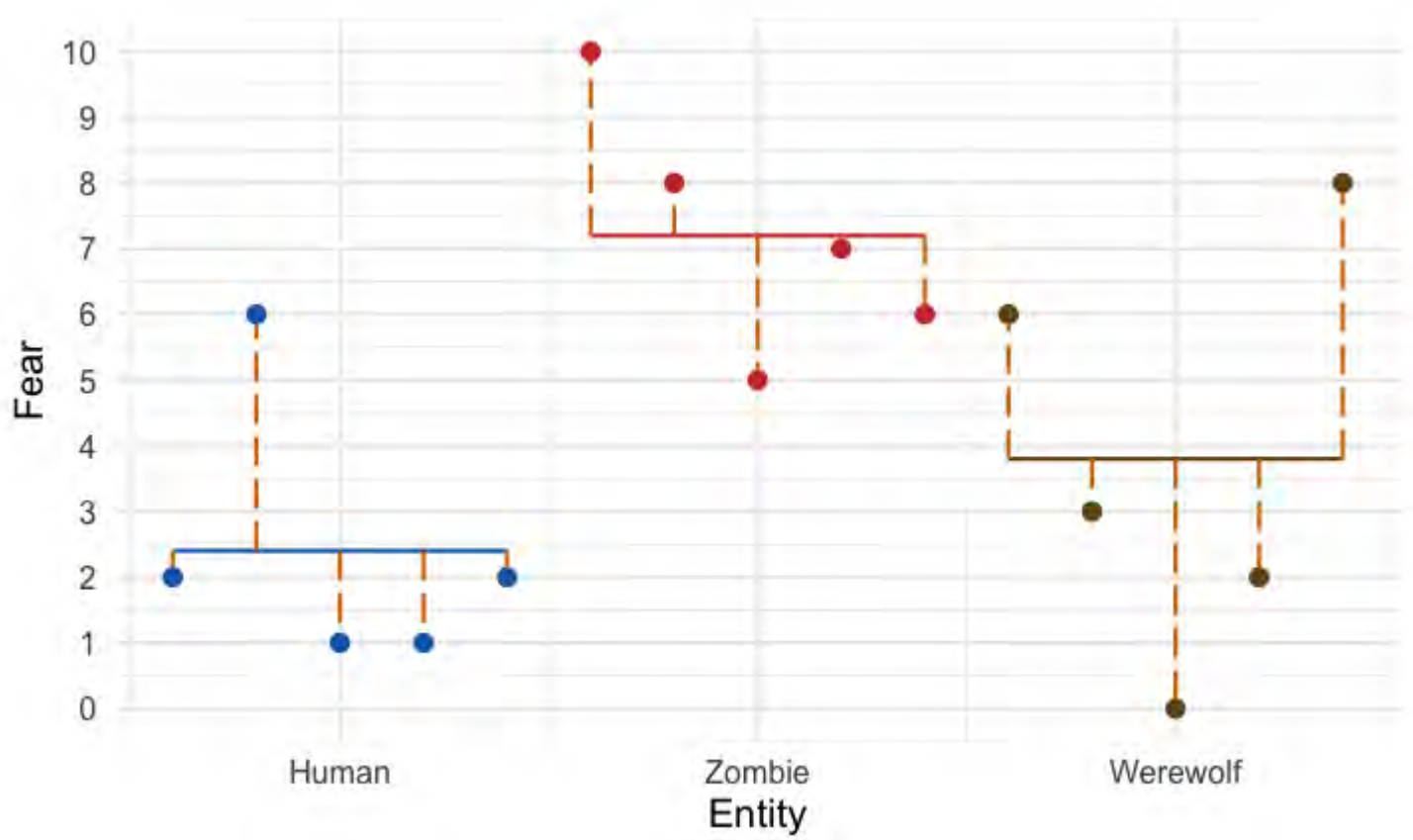
$$\hat{\text{Fear}}_i = \hat{b}_0 + \hat{b}_1 \text{Zombie vs. Human}_i + \hat{b}_2 \text{Werewolf vs. Human}_i + e_i$$

$$\hat{\text{Fear}}_i = 2.4 + 4.8 \text{Zombie vs. Human}_i + 1.4 \text{Werewolf vs. Human}_i + e_i$$

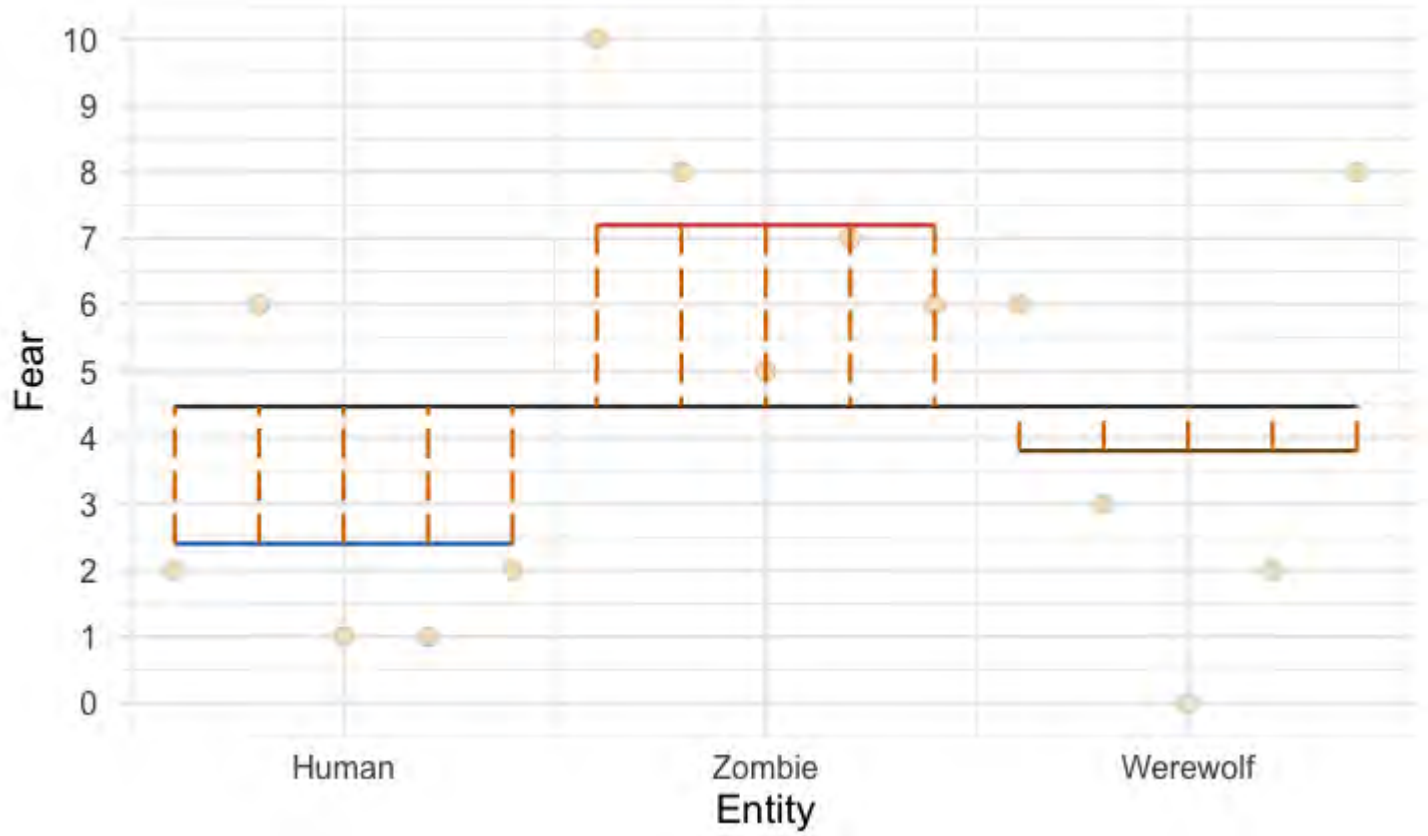
Total sum of squared error, SS_T



Residual sum of squared error, SS_R



Model sum of squared error, SS_M



Testing the model

Overall fit (*F*-statistic)

- The ratio of how well the model fits to how much error it has
- In the case of experiments:
  - The model = differences between means
  - *F* is the ratio of the experimental effect to the background 'error'
  - Tests whether group means differ overall

Parameter estimates (and *t*-tests)

- Break down the overall fit
- Tell us, specifically, which means differ

Overall fit

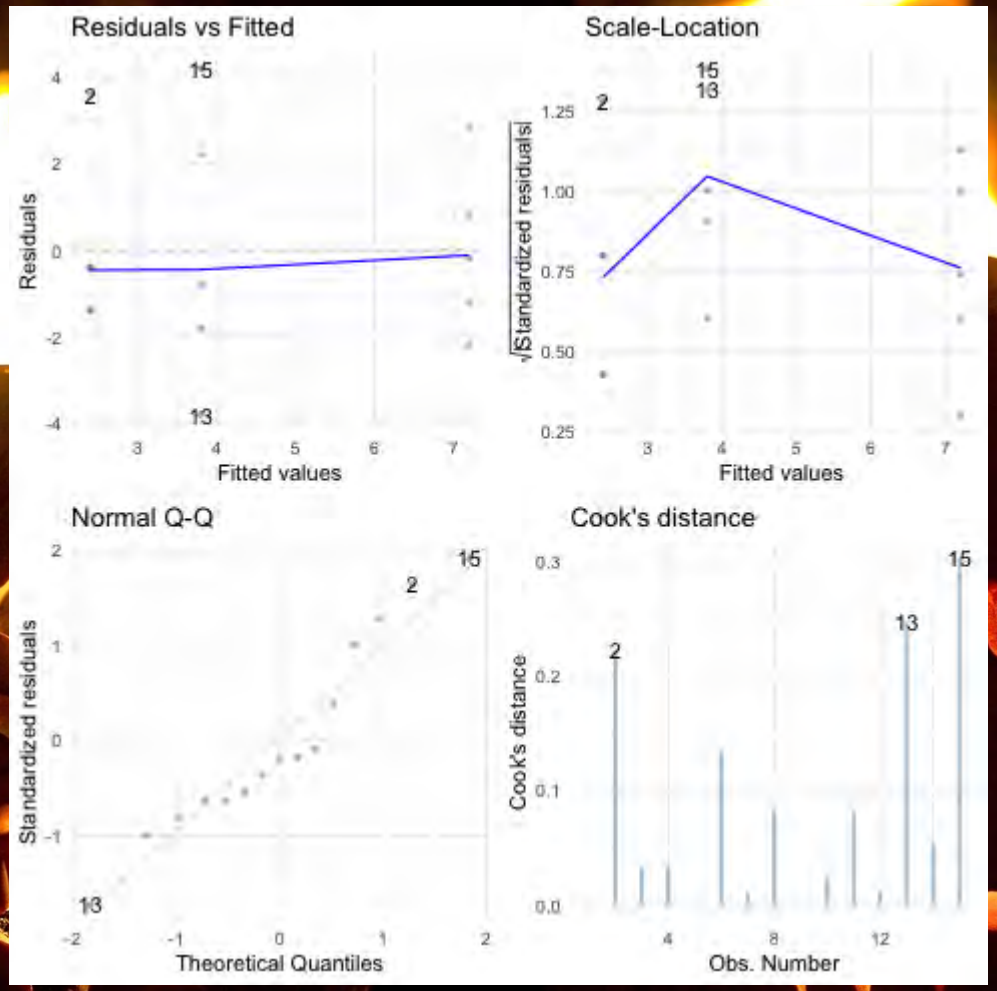
r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.46	0.36	2.46	5.02	0.03	2	-33.13	74.26	77.1	72.8	12	15

The type of monster in the prank had a significant effect on fear levels, $F(2, 12) = 5.02$, $p = 0.026$, $R^2 = 0.46$.

Parameter estimates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.4	1.10	2.18	0.05	0.00	4.80
entityZombie	4.8	1.56	3.08	0.01	1.41	8.19
entityWerewolf	1.4	1.56	0.90	0.39	-1.99	4.79

Testing assumptions



Robust *F*-statistic

```
stats::oneway.test(fear ~ entity, data = humans_tib)
```

```
##
##     One-way analysis of means (not assuming equal variances)
##
## data:  fear and entity
## F = 6.8917, num df = 2.0000, denom df = 7.7365, p-value = 0.01911
```

The type of monster in the prank had a significant effect on fear levels, $F(2, 7.74) = 6.89$, $p = 0.019$.

Robust model using HC4 Standard errors

```
parameters::parameters(human_lm, robust = TRUE, vcov.type = "HC4")
```

Parameter	Coefficient	SE	CI_low	CI_high	t	df_error	p
(Intercept)	2.4	1.037	0.141	4.659	2.315	12	0.039
entityZombie	4.8	1.414	1.719	7.881	3.394	12	0.005
entityWerewolf	1.4	1.904	-2.748	5.548	0.735	12	0.476