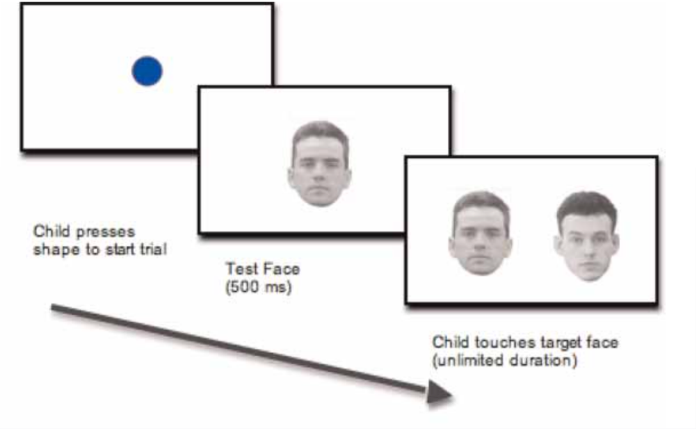
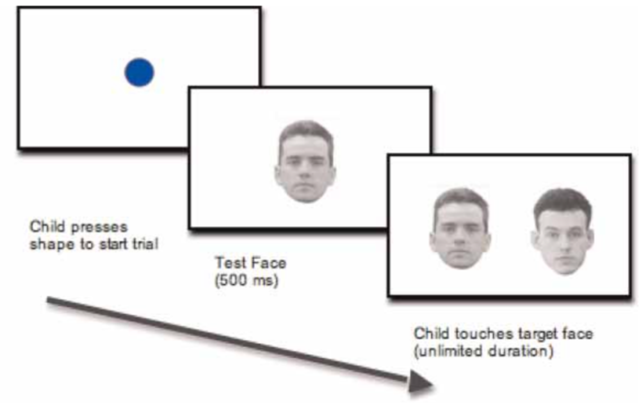
Logistic regression



Dr. Margriet A. Groen

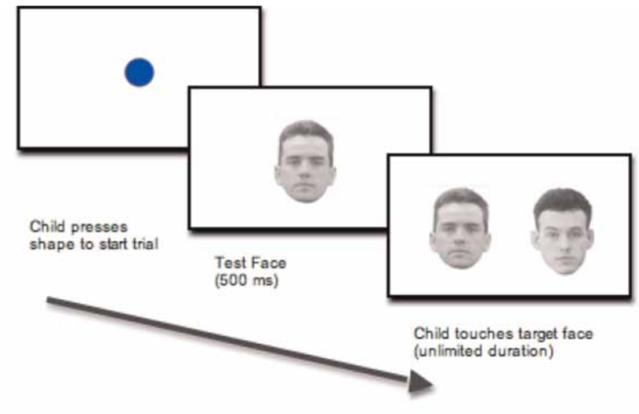
Discrete or categorical outcome variables

- Response accuracy: correct or incorrect?
- Group membership: good vs. poor reader?
- Eye movement: left vs right?
- Ordered categories: Likert rating scales at the point 1, 2, 3, 4, or 5?
- Group membership (out of multiple groups): participant in one of several groups like religious or ethnic or degree class group?
- Frequency of occurrence of an event: number of hallucinations occurring in different patient groups

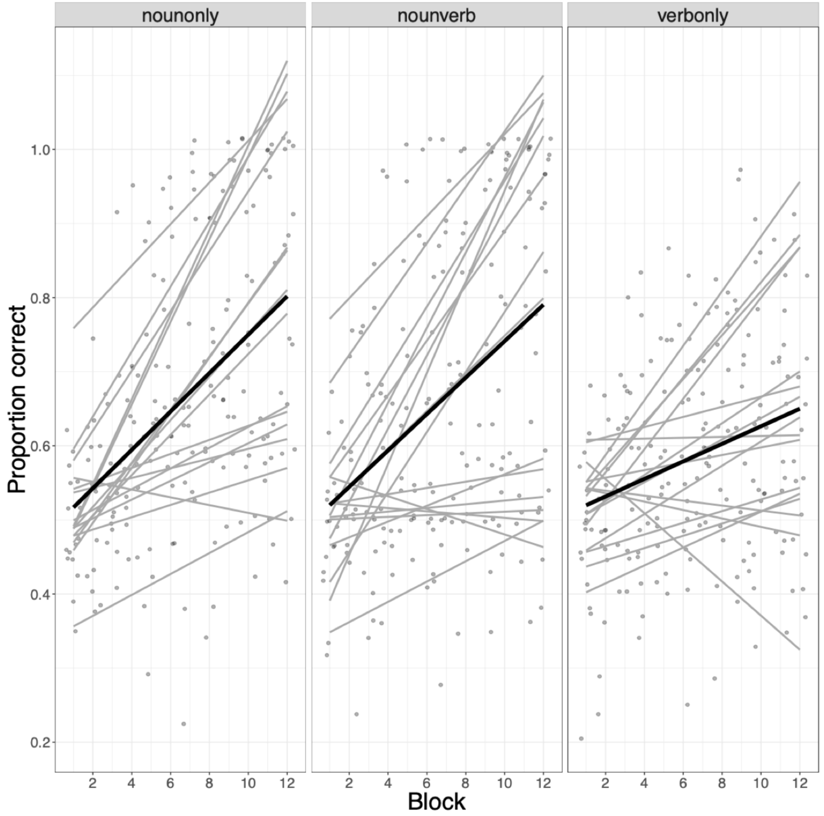


Why not model discrete events as continuous?

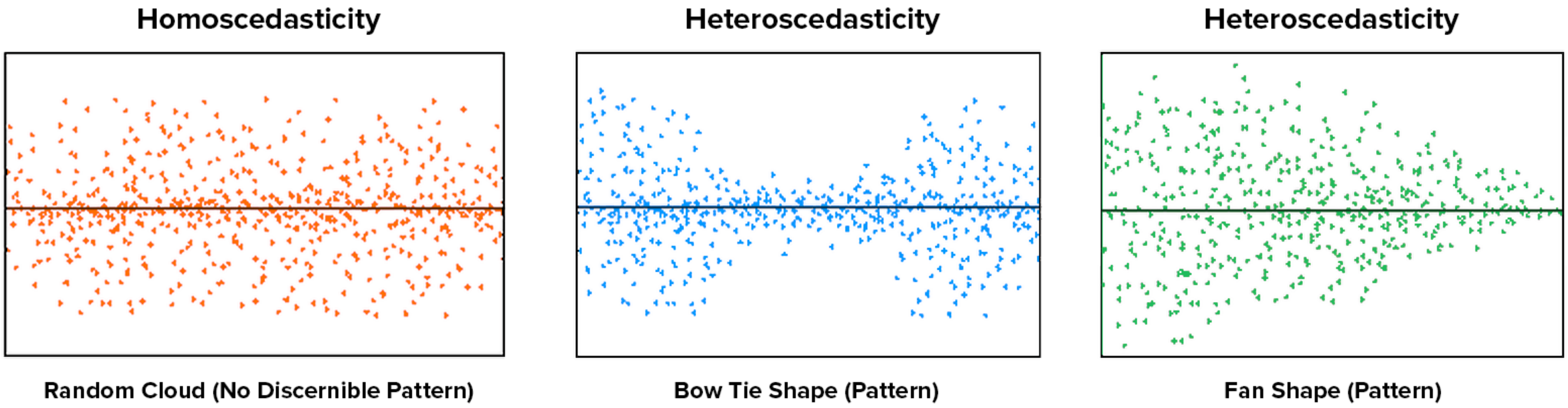
- Forced-choice task -> inaccurate = 0, accurate = 1.
- Could calculate the proportion of accurate responses for each participant (percent correct), and many people do.
- This is a **bad idea** because:
  1. Bounded scale
     - Spurious interaction effects
  2. Variance is proportional to the mean



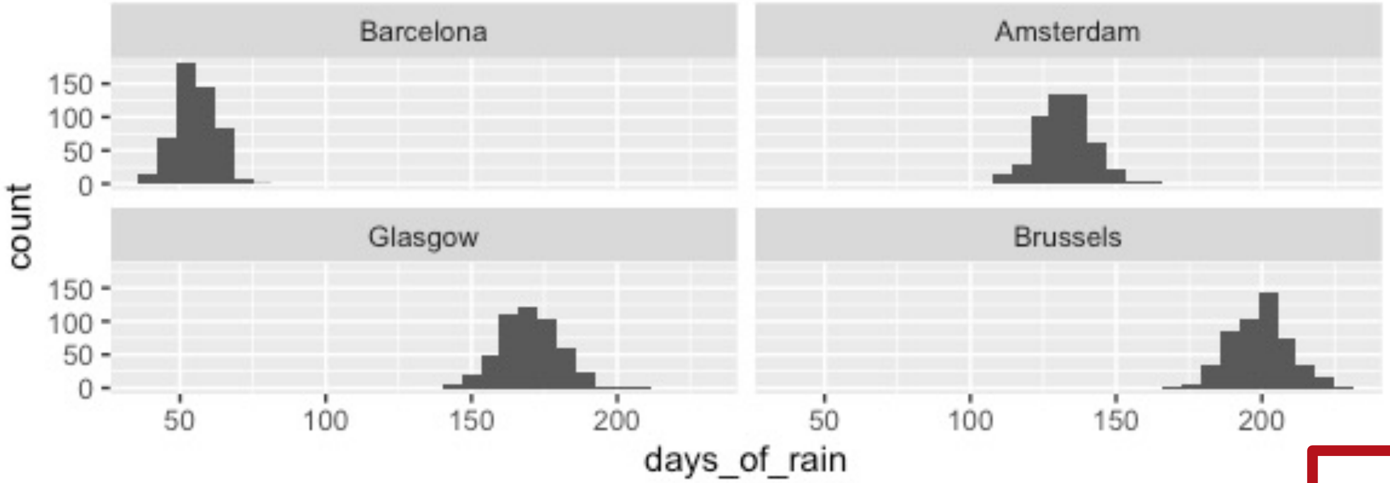
1. Bounded scale



Homoscedasticity



2. Variance is proportional to the mean

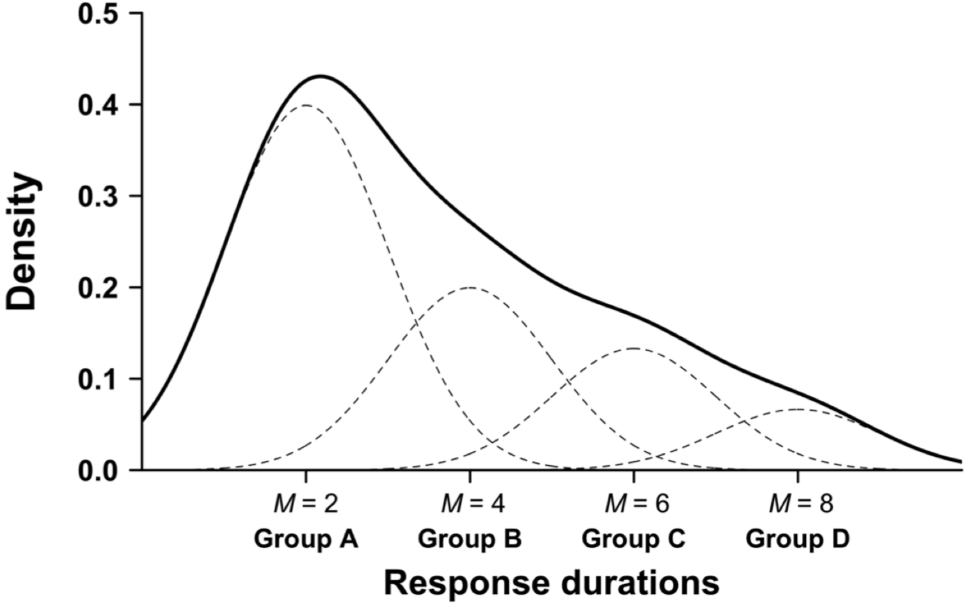


Summary 1

- Linear models assume outcomes are **unbounded** so allow predictions that are impossible when outcomes are, in fact, bounded as is the case for accuracy or other categorical variables
- Linear models assume **homogeneity of variance** but that is unlikely and anyway cannot be predicted in advance when outcomes are categorical variables

Distributions

(a) Positive skew arising from a Gaussian process



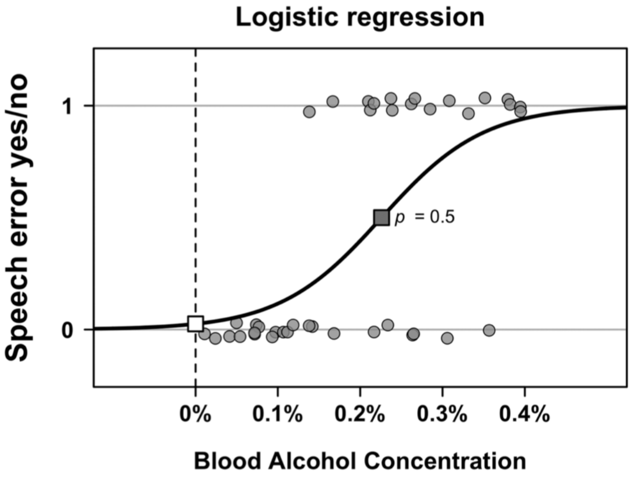
(b) Modeling the process that generates the response

$$\beta_0 + \beta_1 * x_i$$

$$\downarrow$$

$$y_i \sim Normal(\mu_i, \sigma)$$

Bernouille distribution

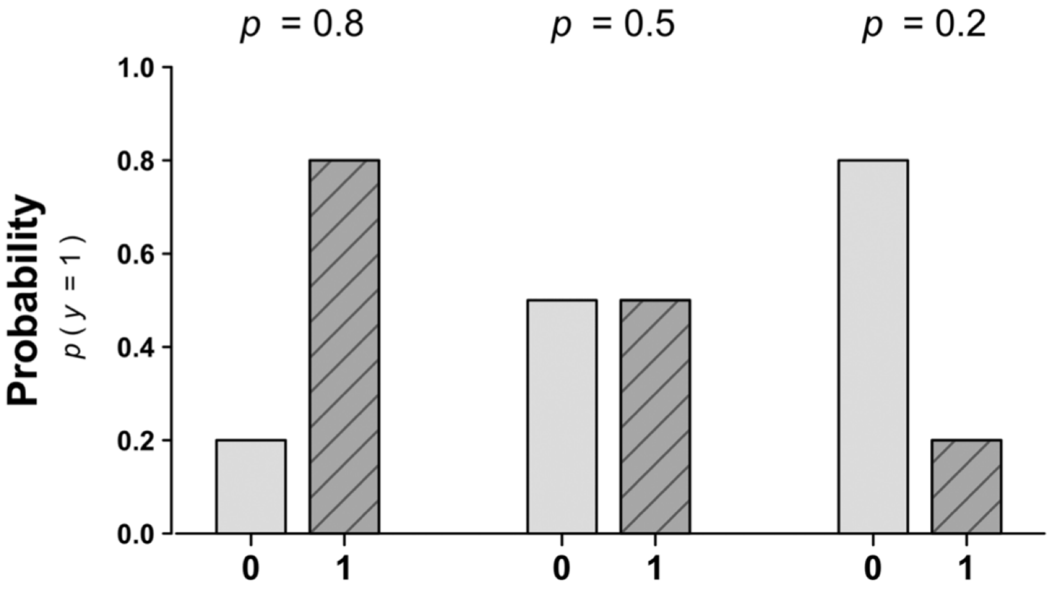


$$y \sim binomial(N = 1, p)$$

$$y \sim bernoulli(p)$$

Logistic regression

(a) The Bernoulli distribution



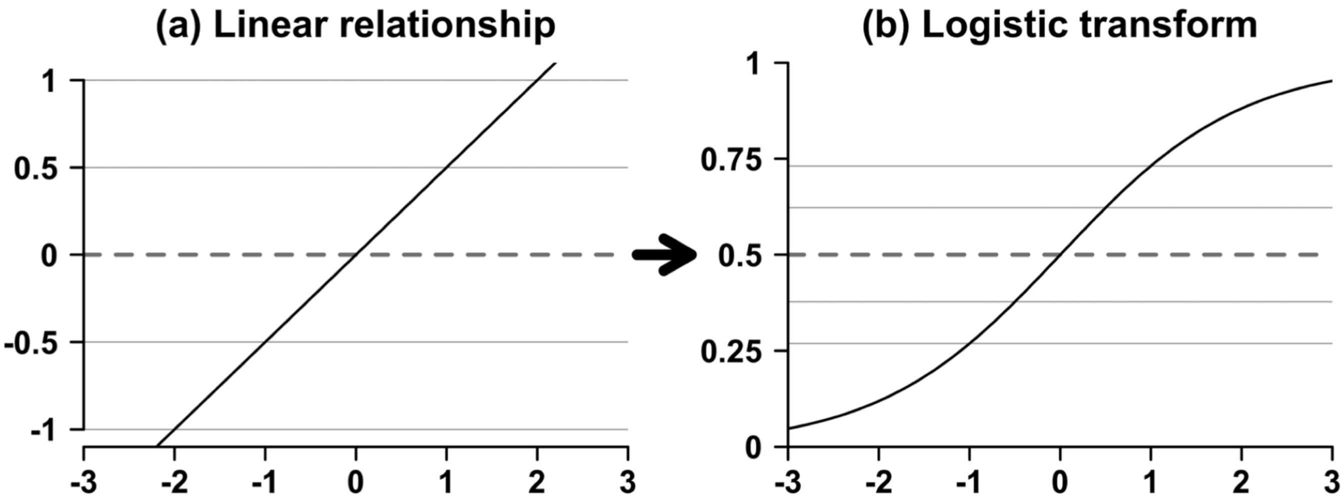
(b) Logistic regression

$$logistic(\beta_0 + \beta_1 * x_i)$$

$$\downarrow$$

$$y_i \sim Bernoulli(p_i)$$

Logistic transform




```
plogis()
```

How to estimate effects of a bounded outcome?

Transform a probability to odds

$$odds = \frac{probability\ of\ something\ happening}{probability\ of\ that\ thing\ not\ happening}$$

$$\text{odds} = \frac{p}{1-p}$$

Odds are continuous ranging from zero to infinity

$$\text{log odds} = \log\left(\frac{p}{1-p}\right)$$

Use the natural logarithm of the odds, because it ranges from negative to positive infinity

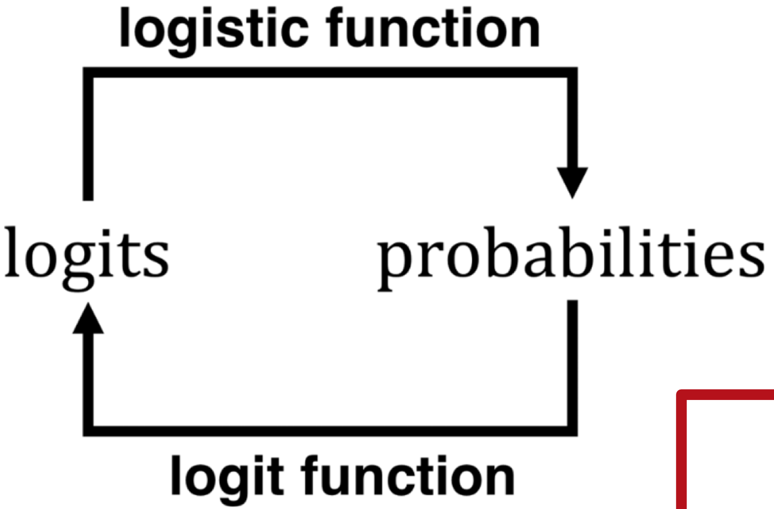
$$logit = ln\ \frac{probability\ of\ something\ happening}{probability\ of\ that\ thing\ not\ happening}$$

Odds and log odds

Probability	*Odds*	*Log odds ('logits')*
0.1	0.11 to 1	–2.20
0.2	0.25 to 1	–1.39
0.3	0.43 to 1	–0.85
0.4	0.67 to 1	–0.41
0.5	**1 to 1**	**0.00**
0.6	1.5 to 1	+0.41
0.7	2.33 to 1	+0.85
0.8	4 to 1	+1.39
0.9	9 to 1	+2.20

Model with interaction term (centered)

$$logit(p_i) = \beta_0 + \beta_1 * x_i$$



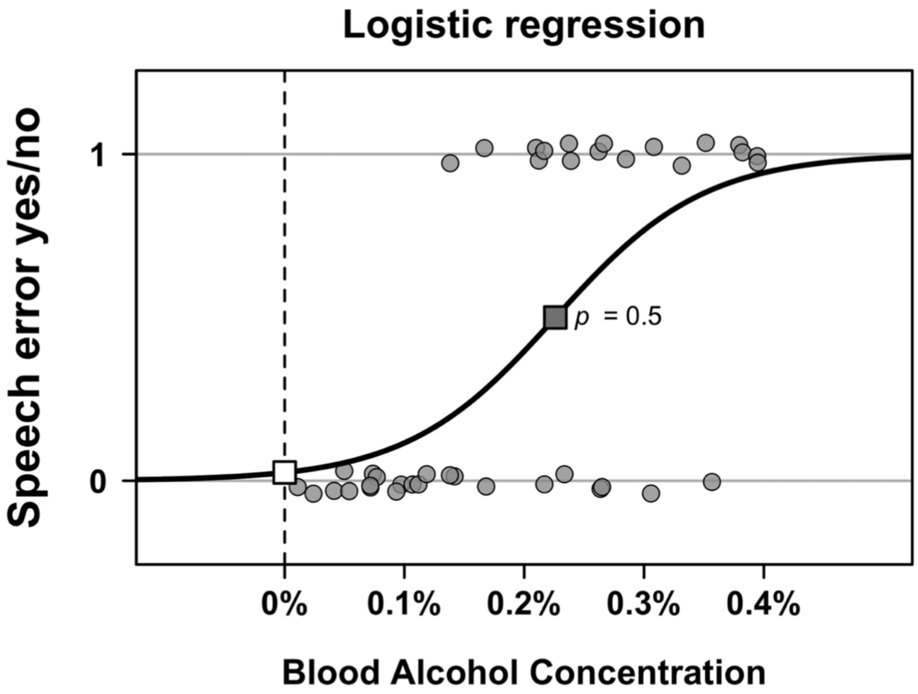
Summary 2

- Categorical outcome variable:
  - Bounded scale
  - Homogeneity of variance not met
- Bernouille distribution:

$$y \sim bernoulli(p)$$

- Transform probability to odds
- Transform odds to log odds

Example: Speech errors and blood alcohol concentration



Example: The data

```
library(tidyverse)
library(broom)

alcohol <- read_csv('speech_errors.csv')

alcohol
```

```
# A tibble: 40 x 2
      BAC speech_error
    <dbl>        <int>
 1 0.0737            0
 2 0.0973            0
 3 0.234             0
 4 0.138             1
 5 0.0933            0
 6 0.262             1
 7 0.357             0
 8 0.237             1
 9 0.352             1
10 0.379             1
# ... with 30 more rows
```

Example: Fitting the model

```
alcohol_mdl <- glm(speech_error ~ BAC,
                   data = alcohol, family = 'binomial')
```

Example: Interpreting the model (1)

```
tidy(alcohol_mdl)
```

```
           term  estimate std.error statistic     p.value
1 (Intercept) -3.643444  1.123176 -3.243878 0.0011791444
2         BAC 16.118147  4.856267  3.319041 0.000903273
```

”Blood alcohol concentration significantly predicted the occurrence of a speech error (logit coefficient: +16.11, SE = 4.86, z = 3.3, p = .0009).”

Example: Interpreting the model (2)

```
intercept <- tidy(alcohol_mdl)$estimate[1]
slope <- tidy(alcohol_mdl)$estimate[2]
intercept
```

```
[1] -3.643444
```

```
slope
```

```
[1] 16.11815
```

Example: Calculating predicted log odds

```
intercept + slope * 0 # BAC = 0
```

```
[1] -3.643444
```

```
intercept + slope * 0.3 # BAC = 0.3
```

```
[1] 1.192
```

Example: Calculating probabilities

```
plogis(intercept + slope * 0)
```

```
[1] 0.02549508
```

```
plogis(intercept + slope * 0.3)
```

```
[1] 0.7670986
```