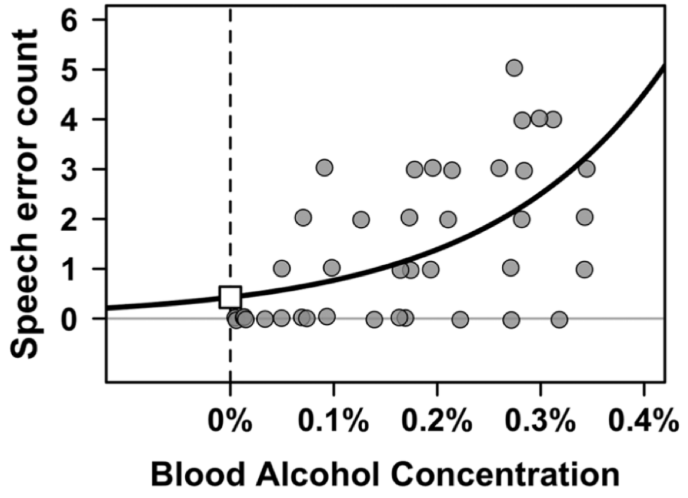
Poisson regression



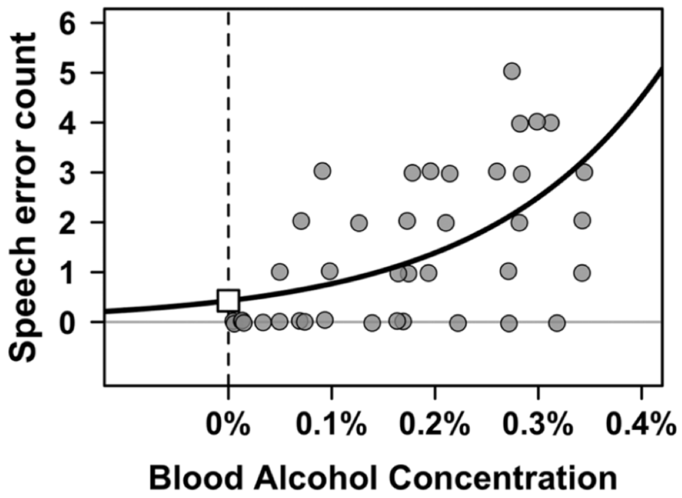
Dr. Margriet A. Groen

Count data as outcome variable

A count variable …

… is a variable that takes on discrete values (0, 1, 2, …) reflecting the number of occurrences of an event in a fixed period of time.

… can only take on positive integer values or zero.



Examples: Number of …

… depressive symptoms that a child exhibits

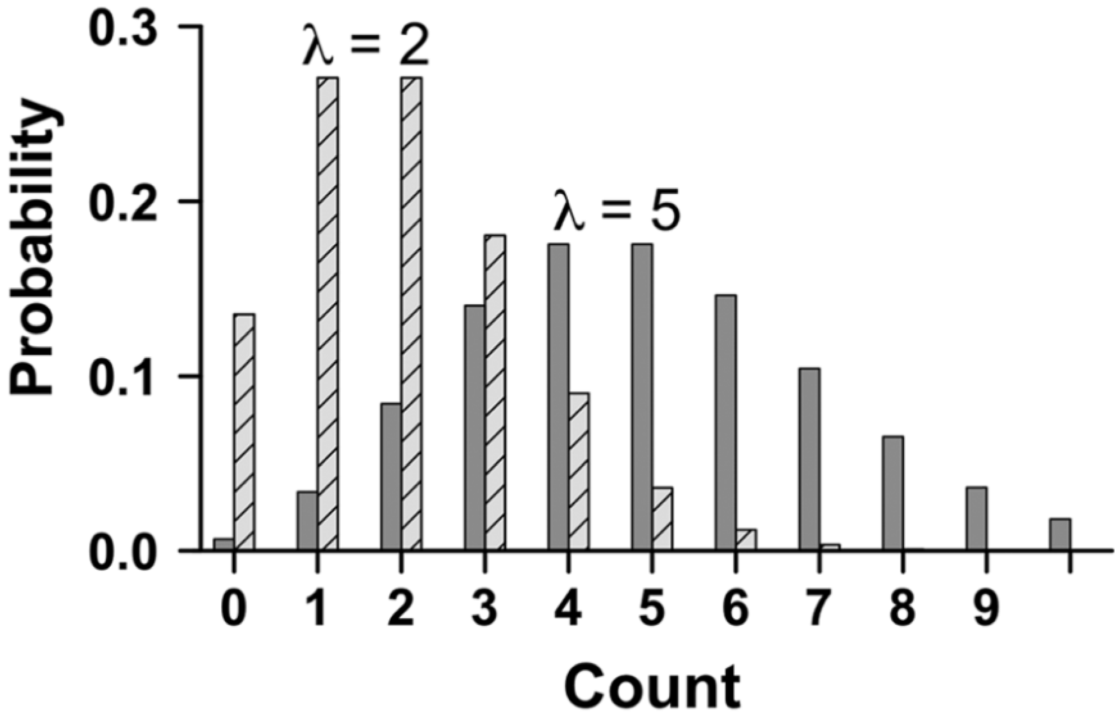
… alcoholic drinks consumed per day

… readmissions to alcohol detoxification programmes

… disciplinary incidents among a group of prison inmates

… fillers (such as uh and oh) as a function of politeness context

The Poisson distribution



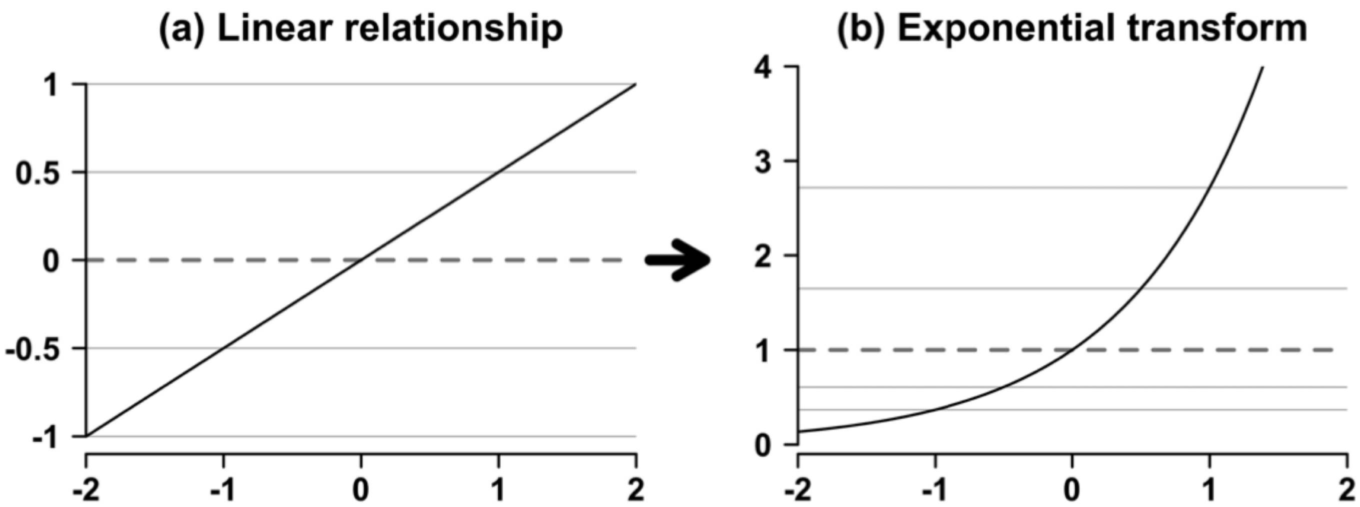
... has one parameter: λ 'lambda'

... cannot be negative

... contains only integers

... variance is associated with λ

Exponential transformation



$$\lambda_i = exp\left(\beta_0 + \beta_1 * x_i\right)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 * x_i$$

An example: Nettle's (1999) linguistic diversity data (1)

```
# A tibble: 74 x 5
       Country Population  Area   MGS Langs
         <chr>      <dbl> <dbl> <dbl> <int>
 1     Algeria       4.41  6.38  6.60    18
 2      Angola       4.01  6.10  6.22    42
 3   Australia       4.24  6.89  6.00   234
 4  Bangladesh       5.07  5.16  7.40    37
 5       Benin       3.69  5.05  7.14    52
 6     Bolivia       3.88  6.04  6.92    38
 7    Botswana       3.13  5.76  4.60    27
 8      Brazil       5.19  6.93  9.71   209
 9 Burkina Faso      3.97  5.44  5.17    75
10         CAR       3.50  5.79  8.08    94
# ... with 64 more rows
```

```
range(nettle$MGS)
```

```
[1] 0 12
```

An example: Nettle's (1999) linguistic diversity data (2)

```
filter(nettle, MGS %in% range(MGS))
```

```
# A tibble: 6 x 5
  Country         Population  Area   MGS Langs
  <chr>                <dbl> <dbl> <dbl> <int>
1 Guyana                2.90  5.33   12.    14
2 Oman                  3.19  5.33    0.     8
3 Solomon Islands       3.52  4.46   12.    66
4 Suriname              2.63  5.21   12.    17
5 Vanuatu               2.21  4.09   12.   111
6 Yemen                 4.09  5.72    0.     6
```

The model (1)

```
MGS_mdl <- glm(Langs ~ MGS, data = nettle,
               family = 'poisson')

tidy(MGS_mdl)
```

```
         term  estimate   std.error statistic       p.value
1 (Intercept) 3.4162953 0.039223267  87.09869  0.000000e+00
2         MGS 0.1411044 0.004526387  31.17375 2.417883e-213
```

$$\log(\lambda_i) = \beta_0 + \beta_1 * x_i$$

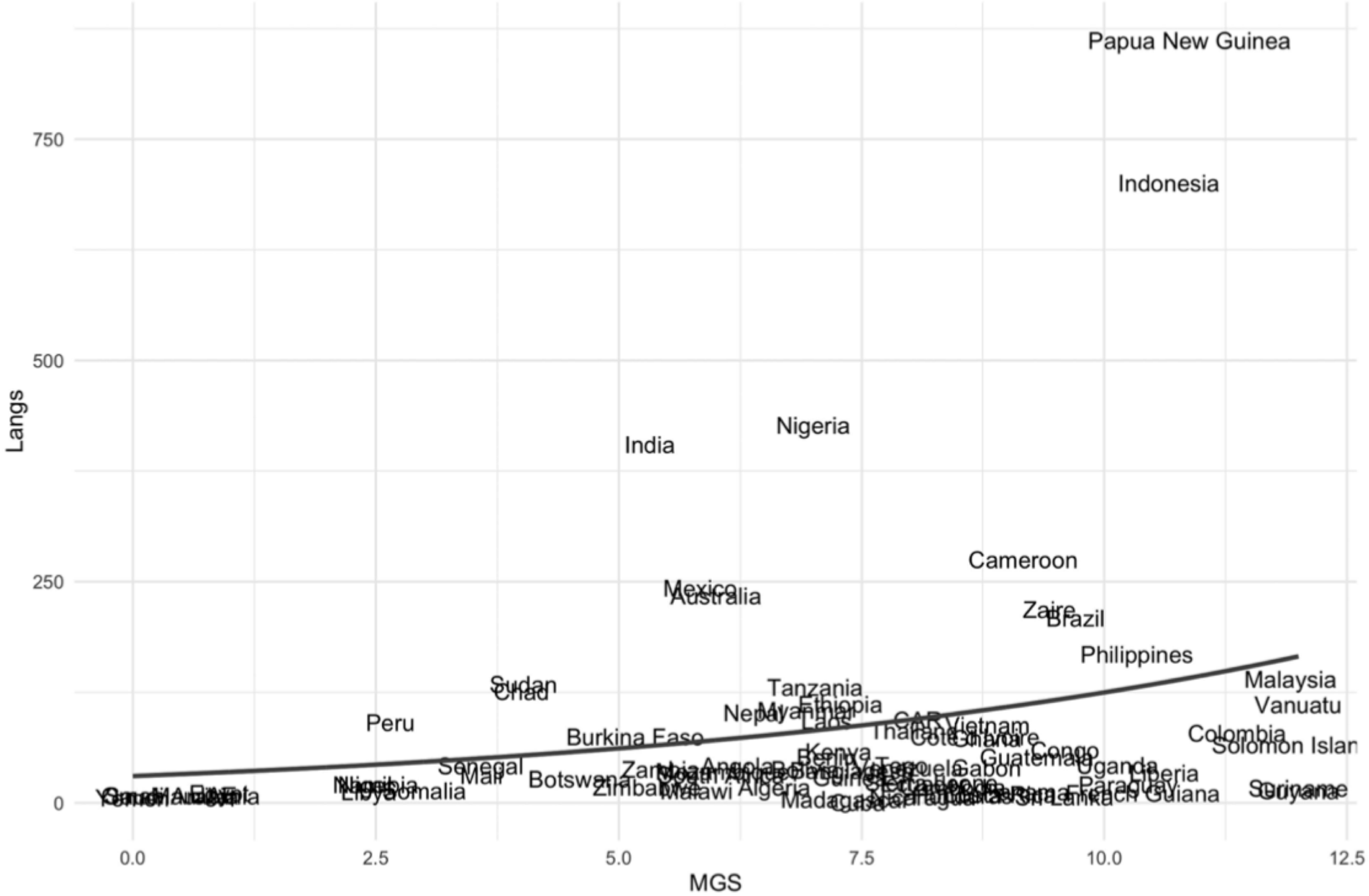
The model (2)

```
intercept + 0:12 * slope
```

```
 [1] 3.416295 3.557400 3.698504 3.839609 3.980713 4.121818
 [7] 4.262922 4.404026 4.545131 4.686235 4.827340 4.968444
[13] 5.109549
```

```
exp(intercept + 0:12 * slope)
```

```
 [1]  30.45637  35.07188  40.38685  46.50727  53.55521
 [6]  61.67123  71.01719  81.77948  94.17275 108.44415
[11] 124.87831 143.80298 165.59559
```


Papua New Guinea
Indonesia
Nigeria
India
Cameroon
Mexico
Australia
Zaire
Brazil
Philippines
Malaysia
Vanuatu
Sudan
Chad
Tanzania
Ethiopia
Peru
Burkina Faso
Vietnam
Colombia
Solomon Islan
Kenya
Congo
Gabon
Uganda
Liberia
Senegal
Mali
Paraguay
Zimbabwe
Malawi
Algeria
Libya
Somalia
Guyana
750
500
250
0
Langs
0.0
2.5
5.0
7.5
10.0
12.5
MGS

Exposure variables (1)

An exposure variable …

… is a variable that potentially allows for more opportunities to observe a higher count

Examples:

space (e.g., size of a country)

time (e.g., trial duration/number of hours observed)

You can adjust a rate (count) by an exposure variable.

$$\lambda = \frac{\mu}{\tau};$$

Exposure variables (2)

```
MGS_mdl <- glm(Langs ~ MGS, data = nettle,
               family = 'poisson')

tidy(MGS_mdl)
```

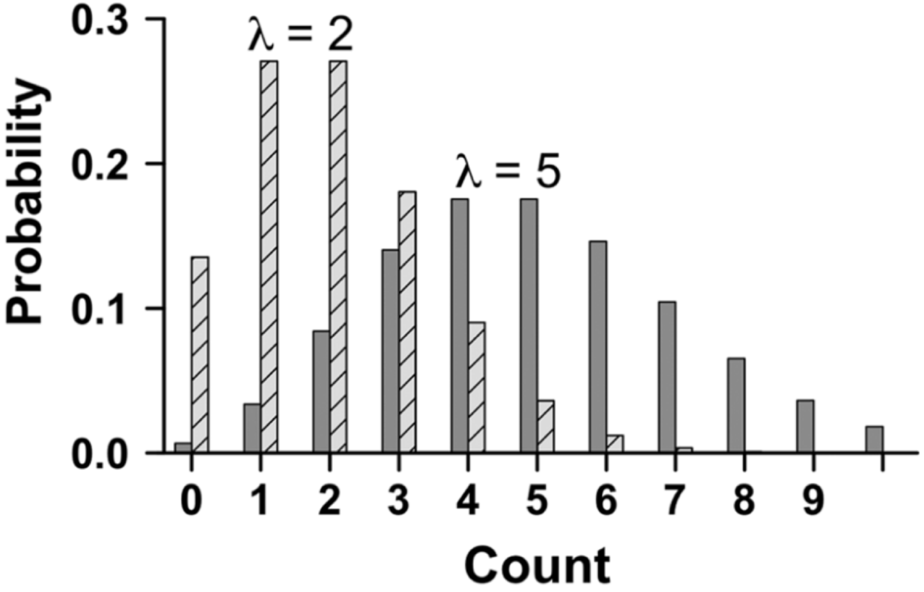
```
         term   estimate   std.error statistic       p.value
1 (Intercept) 3.4162953 0.039223267  87.09869  0.000000e+00
2         MGS 0.1411044 0.004526387  31.17375 2.417883e-213
```

```
MGS_mdl_exposure <- glm(Langs ~ MGS + offset(Area),
                        data = nettle, family = 'poisson')

tidy(MGS_mdl_exposure)
```

```
         term   estimate   std.error statistic p.value
1 (Intercept) -2.8230092 0.040738134 -69.29648       0
2         MGS  0.2092749 0.004719774  44.34003       0
```

Overdispersion



Variance of the Poisson distribution scales with the mean: the higher the mean rate, the more variable the counts.

If the variance is larger than theoretically expected for a given lambda, you are dealing with what's called '**overdispersion**' or 'excess variance'.

Negative binom

```
MGS_mdl_exposure <- glm(Langs ~ MGS + offset(Area),
                        data = nettle, family = 'poisson')

tidy(MGS_mdl_exposure)
```

```
         term   estimate   std.error statistic p.value
1 (Intercept) -2.8230092 0.040738134 -69.29648       0
2         MGS  0.2092749 0.004719774  44.34003       0
```

```
library(MASS)

# Fit negative binomial regression:

MGS_mdl_nb <- glm.nb(Langs ~ MGS + offset(Area),
                     data = nettle)

tidy(MGS_mdl_nb)
```

```
         term   estimate  std.error  statistic      p.value
1 (Intercept) -3.0527417 0.26388398 -11.568500 5.951432e-31
2         MGS  0.2296025 0.03418441   6.716585 1.860333e-11
```

Negative binomial regression

```
library(pscl)

# Perform overdispersion test:

odTest(MGS_mdl_nb)
```

```
Likelihood ratio test of H0: Poisson, as restricted NB
model:
n.b., the distribution of the test-statistic under H0 is
non-standard
e.g., see help(odTest) for details/references

Critical value of test statistic at the alpha= 0.05 level:
2.7055
Chi-Square Test Statistic = 5533.0321 p-value = < 2.2e-16
```

Generalized Linear Model Framework

$$I(\beta_0 + \beta_1 * x_i)$$
$$\downarrow$$
$$y_i \sim Normal(\mu_i, \sigma)$$

Linear regression

$$logistic(\beta_0 + \beta_1 * x_i)$$
$$\downarrow$$
$$y_i \sim Bernoulli(p_i)$$

Logistic regression

$$exp(\beta_0 + \beta_1 * x_i)$$
$$\downarrow$$
$$y_i \sim Poisson(\lambda_i)$$

Poisson regression

Summary

- You have learned how to model count data with **Poisson regression**, and its extension, **negative binomial regression**.
- The coefficients of a Poisson model are shown as **log coefficients**, which means that, after calculating the **log predictions**, you need to use **exponentiation** to interpret your model in terms of average rates.
- To control for differential exposure, **exposure variables** can be added.
- Negative binomial regression was used to account for **overdispersion**.
- Each GLM has three components: a **distribution** for the data-generating process, **a linear predictor** and **a link function**.