STA303H1S/1002H Week 7

Distributions and generalized linear models (GLMs)

Prof. Liza Bolton

Mar 01, 2021

Topics

You can click the following links to navigate through the slides (in the HTML version).

- Follow-ups from the check-in survey
- Distributions
- Categorical variables: tables, odds ratios and relative risks
- Generalized linear models (GLMs)
- Case study: Challenger disaster
- Extra for the curious (NOT assessed)

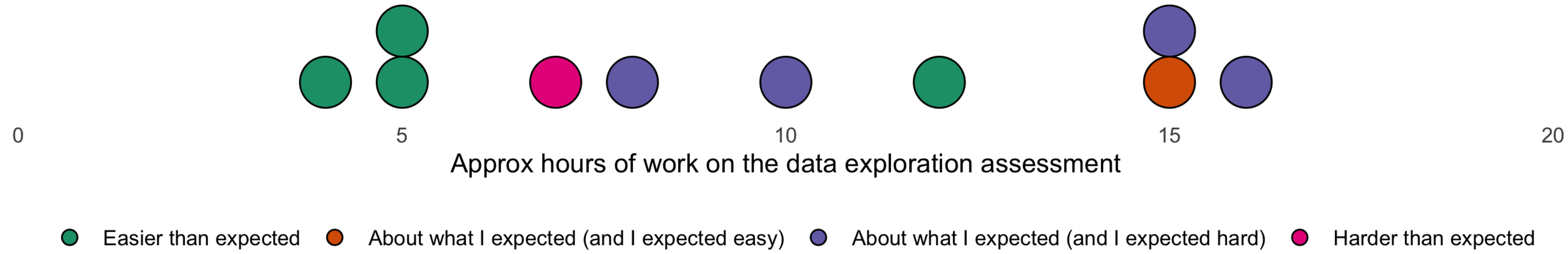
Go to topics list

Follow-ups from the check-in survey

Thank you for completing the end-of-week check-in!

- Happy belated birthday to the friends and family members you told me you were celebrating with over reading week. Hope you had a great time playing board games, watching K-dramas and anime, playing piano, going for walks by yourself or with your dog, sledding and going to hot springs (jealous!). But maybe don't tell me you're doing illegal things in a class survey...(Also, please don't do illegal things.)

- What did you think of the data exploration assessment and how long did it take you? Small sample, and almost certainly biased because this survey is opt-in...but this looks pretty reasonable to me.



Communication policy reminder

- All content and logistics questions must be asked on Piazza.
- Personal or private course matters should be emailed to sta303@utoronto.ca.
- Quercus mail or emails sent directly to teaching team members will not be answered.
- If you've missed an assessment due to illness or emergency, please fill out the appropriate form as soon as possible.

Upcoming assessments: weekly

- Week 7 quiz (due Wednesday, Mar 3 at 10:00 a.m. ET)
- Week 7 writing
  - Create phase due Mar 4 at 6:00 p.m. ET
  - Assess phase due Mar 5 at 6:00 p.m. ET
  - Reflect phase due Mar 8 at 6:00 p.m. ET

Upcoming assessments: non-weekly

Polished writing 2 (due Mar 12 at 6:00 p.m. ET)

Polished writing 2 must be a response to one of the prompts from Week 4 writing, Week 5 writing or Week 7 writing. You do not have to have completed the activity for that week to be able to submit your response as your polished writing, but the intention is that you are submitting a piece improved based on feedback from your peers.

Confirm project group/individual status (due Mar 19 at 6:00 p.m. ET)

More information about how to do this, and the project in general, **coming soon**. BUT you can already start thinking about if you'd like to work as a group (and if so, whom with) or an individual.

Professional development evidence and reflection (due Mar 26 at 6:00 p.m. ET)

Grading updates

Marks for the professional development proposal (median mark: B+) and polished writing 1 (median mark: A+) will be released today.

Grading is still underway for the data exploration assessment and the mixed assessment. I will share the marking rubrics as soon as all accommodated submissions are in.

Go to topics list

Distributions

Introduction

Much of this section should be recap of things you've learned in second year statistics courses.

Reading: Chapter 3 (§ 3.3.1, 3.3.2, 3.3.4, 3.3.6, 3.4.2, 3.5) of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). https://bookdown.org/roback/bookdown-BeyondMLR/.

Sections to read

- 3.3 Discrete random variables
    - 3.3.1 Binary Random Variable
    - 3.3.2 Binomial Random Variable
    - 3.3.4 Negative Binomial Random Variable
- 3.4 Continuous random variables
    - 3.4.2 Gamma Random Variable
    - 3.4.4 Beta Random Variable
- 3.5 Distributions Used in Testing
    - 3.5.1 χ^2 Distribution
    - 3.5.2 Student's t-Distribution
    - 3.5.3 F-Distribution

Reading guide

Try to answer the following for the selected distributions:

- What is the probability distribution function?
- What is/are the parameter(s)?
- How do changes to the the parameter(s) effect the response?
- What are the **mean** and **variance**?
- What values can your response variable take?
- When might you use this distribution? Come up with an example.
- What R code can you use to explore the density of this distribution?
- Can you simulate the distribution? Play with the parameters for yourself.

For the distributions used in tests:

- When might you use this distribution in a test? Come up with an example.
- What R code can you use to explore the density of this distribution? Can you plot the distribution? Play with the parameters for yourself.

Cheat sheet template

I've made a template for a 'cheat sheet' on which you can take these notes/play with code. You can access it with the `sta303_w7` package.

Just open RStudio, either locally or on the JupyterHub (wherever you can Knit) and run the following code in the **console.**

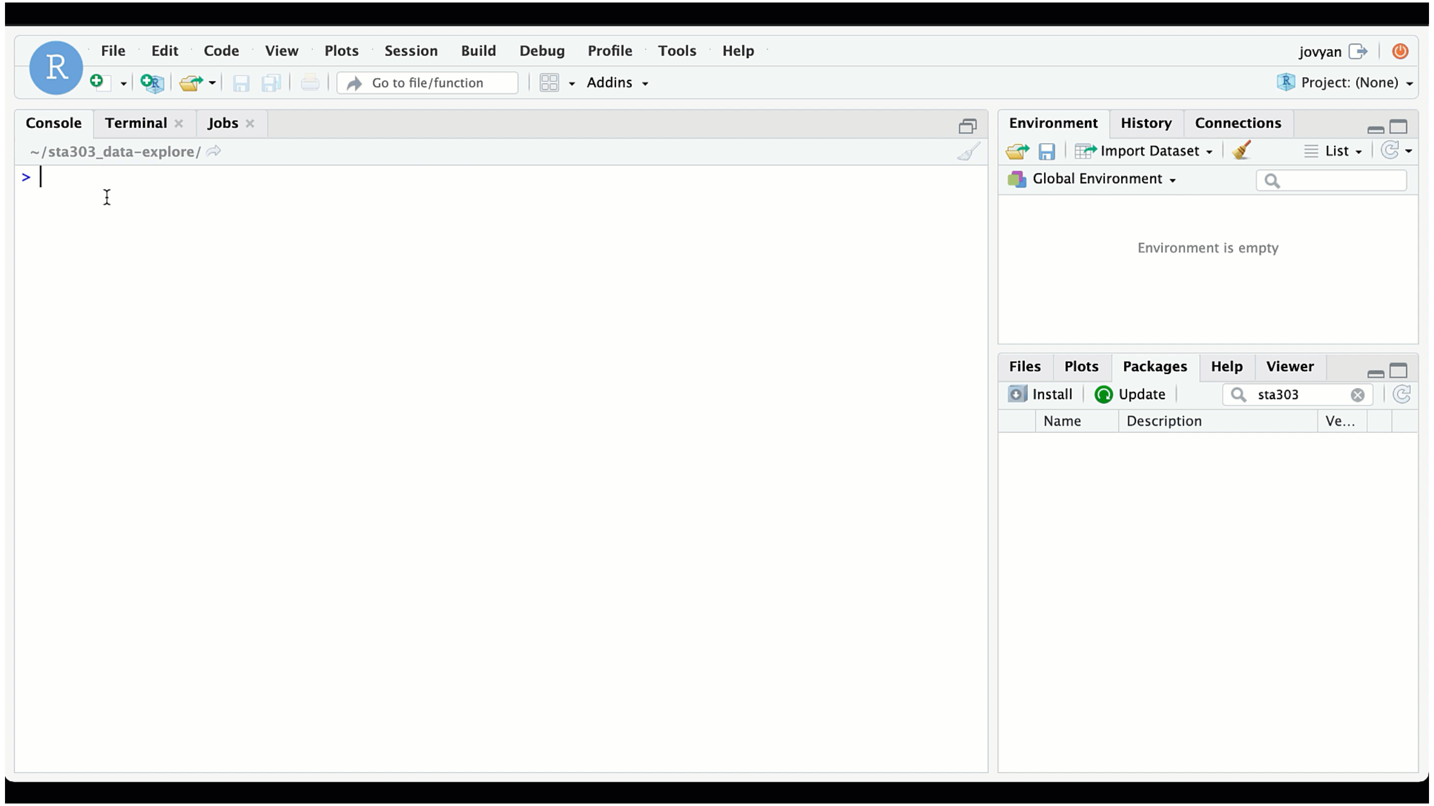
```
devtools::install_github("sta303-bolton/sta303w7")
# If it prompts you to update packages,
# just click 'Enter' or 'Return' or input 3
```

Then, go to File -> New File -> R Markdown and choose 'From template' and select 'Distribution cheat sheet {sta303w7}'.

Important: If you're working in the JupyterHub, make sure you save the .Rmd in a subfolder of your Home directory so it doesn't cause class interactives to fail in future. (Can't have any .Rmd in your Home directory.)

GIF of how to access the template

(This is just a visual of the instructions on the previous page.)



Categorical variables: tables, odds ratios and relative risks

Ontario COVID hospitalizations

Suppose you're interested in hospitalizations by age group in Ontario due to COVID. This table also shows the estimated population in each age group.

This data was retrieved from https://www.publichealthontario.ca/en/data-and-analysis/infectious-disease/covid-19-data-surveillance/covid-19-data-tool?tab=ageSex on 2021-02-28.

Table: COVID-19 hospitalizations in Ontario, by age group

Age	Hospitalizations	Population
0 to 09	98	1518527
10 to 19	92	1617937
20 to 29	402	2100175
30 to 39	668	2056056
40 to 49	1010	1876585
50 to 59	1932	2060937
60 to 69	2714	1795047
70 to 79	3151	1159898
80 to 89	3354	539715
90+	1645	139551

Creating tables in R

Suppose our raw data has an observation per row.

```r
ontario <- readRDS("data/w7/ontario_full.RDS")
head(ontario)
```

```
Rows: 14,864,428
Columns: 2
$ Age    <fct> 0 to 09, 0 to 09, 0 to 09, 0 to 09, 0 to 09, 0 to 09, 0 to 09, 0 …
$ Status <chr> "Hospitalized", "Hospitalized", "Hospitalized", "Hospitalized", "…
```

Creating tables

You can create a table with the `table()` function in R. (The `xtabs` function is also useful for creating cross (x) tables.)

```
hosp_table <- table(ontario$Age, ontario$Status)
hosp_table
```

```
          Hospitalized Not hospitalized
  0 to 09           98          1518429
  10 to 19          92          1617845
  20 to 29         402          2099773
  30 to 39         668          2055388
  40 to 49        1010          1875575
  50 to 59        1932          2059005
  60 to 69        2714          1792333
  70 to 79        3151          1156747
  80 to 89        3354           536361
  90+             1645           137906
```

Calculations with tables

From this kind of table there are three types of proportions that we can calculate.

- **Joint**
  - Joint proportions reflect the proportion total observation for which given levels of your categorical variables co-occur. I.e., what proportion of people were over 90 and hospitalized?
  - General calculation: Cell value over the grand total.
- **Marginal**
  - Marginal proportions sum across rows or columns. I.e., what is the proportion the Ontario population that has been hospitalized? We'd need to add up all the hospitalized and then divide that by the sum of all the cells.
  - General calculation: Row or columns sums over the grand total
- **Conditional**
  - Conditional proportions hold one variable level as given, it is a bit like zooming in to only one row or one column.
  - General calculation: Cell value of a row or column sum.

The `prop.table()` function will be very helpful to us!

Joint probabilities

Calculate joint proportions:

$$\frac{n_{ij}}{\Sigma n_{ij}}$$

```
hosp_table/sum(hosp_table)
```

```
##
##            Hospitalized Not hospitalized
##   0 to 09  6.592921e-06     1.021519e-01
##   10 to 19 6.189273e-06     1.088400e-01
##   20 to 29 2.704443e-05     1.412616e-01
##   30 to 39 4.493950e-05     1.382756e-01
##   40 to 49 6.794745e-05     1.261788e-01
##   50 to 59 1.299747e-04     1.385190e-01
##   60 to 69 1.825835e-04     1.205787e-01
##   70 to 79 2.119826e-04     7.781981e-02
##   80 to 89 2.256394e-04     3.608353e-02
##   90+      1.106669e-04     9.277585e-03
```

Marginal probabilities

Proportion of people in each age group:

```
round(margin.table(hosp_table, margin = 1)/sum(margin.table(hosp_table, margin = 1)), 3)
```

```
## 
##  0 to 09 10 to 19 20 to 29 30 to 39 40 to 49 50 to 59 60 to 69 70 to 79 
##    0.102    0.109    0.141    0.138    0.126    0.139    0.121    0.078 
## 80 to 89      90+ 
##    0.036    0.009 
```

Proportion of people in each hospitalization status group:

```
round(margin.table(hosp_table, margin = 2)/sum(margin.table(hosp_table, margin = 2)), 3)
```

```
## 
##     Hospitalized Not hospitalized 
##            0.001            0.999 
```

Conditional probabilities

Conditional on each row:

```
kable(prop.table(hosp_table, margin = 1))
```

	Hospitalized	Not hospitalized
0 to 09	0.0000645	0.9999355
10 to 19	0.0000569	0.9999431
20 to 29	0.0001914	0.9998086
30 to 39	0.0003249	0.9996751
40 to 49	0.0005382	0.9994618
50 to 59	0.0009374	0.9990626
60 to 69	0.0015119	0.9984881
70 to 79	0.0027166	0.9972834
80 to 89	0.0062144	0.9937856
90+	0.0117878	0.9882122

Conditional on each column:

```
kable(round(prop.table(hosp_table, margin = 2), 3))
```

	Hospitalized	Not hospitalized
0 to 09	0.007	0.102
10 to 19	0.006	0.109
20 to 29	0.027	0.141
30 to 39	0.044	0.138
40 to 49	0.067	0.126
50 to 59	0.128	0.139
60 to 69	0.180	0.121
70 to 79	0.209	0.078
80 to 89	0.223	0.036
90+	0.109	0.009

Risk and odds

"Risk" refers to the probability of occurrence of an event or outcome. Statistically, risk = chance of the outcome of interest/all possible outcomes. The term “odds” is often used instead of risk. "Odds" refers to the probability of occurrence of an event/probability of the event not occurring. At first glance, though these two concepts seem similar and interchangeable, there are important differences that dictate where the use of either of these is appropriate.

~ From Common pitfalls in statistical analysis: Odds versus risk

Hospitalization risk and odds

Let's folks on folks 80 to 89.

```
           Hospitalized Not hospitalized
  80 to 89         3354           536361
```

This **risk** of being hospitalized for this group is $\frac{3354}{3354+536361} = 0.0062$.

The **odds** of being hospitalized are $\frac{3354}{536361} = 0.0063$.

These values look fairly similar. Odds and risks **will** be similar when the outcome of interest is rare. This can be seen by the fact that the only difference between the two calculations is whether the count of the outcome is included in the denominator or not. As a rule of thumb, an outcome is 'rare' if it occurs less than 10% of the time.

Odds ratio and risk ratios

Risk ratios are also called 'relative' risks. Risk ratios and odds ratios are...ratios of risks and odds respectively.

They are used to make comparisons between groups. Let's for example, compare 80 to 89 year olds with 10 to 19 year olds.

```
           Hospitalized Not hospitalized
  10 to 19           92          1617845
  80 to 89         3354           536361
```

$$OR = \frac{3354/536361}{92/1617845} = 110$$

$$RR = \frac{3354/(3354 + 536361)}{92/(92 + 1617845)} = 109$$

Once again, these values are similar because being hospitalized is (thankfully!) rare.

When do we use RR vs OR?

Calculation of *risk* requires as to know how many people are 'at risk'. As we'll see next week, in case-control studies, where such totals are not available to us, we cannot calculate a relative risk. BUT, we can calculate odds ratios and make a comment on the strength of association between our exposure and the outcome.

In cohort studies, where we do have the number number exposed, we can calculate either/both.

Logistic regression, which we'll be seeing more of in the next few weeks, calculates adjusted ORs and not RRs and so being able to interpret them is going to be important to us.

Generalized linear models (GLMs)

Generalized linear models (GLMs)

Generalized linear models are a **flexible** class of models that let us *generalize* from the linear model to include more types of response variables, such as *count, binary, and proportion data*.

Let's get flexible, flexible...

Assumptions of the Generalized Linear Model

- The data $Y_1, Y_2, \ldots, Y_n$ are independently distributed, i.e., cases are independent.
    - Thus errors are independent... but NOT necessarily normally distributed.

Assumptions of the Generalized Linear Model

- The data $Y_1, Y_2, \ldots, Y_n$ are independently distributed, i.e., cases are independent.
  - Thus errors are independent... but NOT necessarily normally distributed.
- The dependent variable Y_i does NOT need to be normally distributed, but it assumes a distribution, typically from an exponential family (e.g. binomial, Poisson, gamma,...)

Assumptions of the Generalized Linear Model

- The data $Y_1, Y_2, \ldots, Y_n$ are independently distributed, i.e., cases are independent.
  - Thus errors are independent... but NOT necessarily normally distributed.
- The dependent variable Y_i does NOT need to be normally distributed, but it assumes a distribution, typically from an exponential family (e.g. binomial, Poisson, gamma,...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but **it does assume a linear relationship between the transformed response (in terms of the link function) and the explanatory variables**; e.g., for binary logistic regression $logit(p) = \beta_0 + \beta_1 X$.

Assumptions of the Generalized Linear Model

- The data $Y_1, Y_2, \ldots, Y_n$ are independently distributed, i.e., cases are independent.
  - Thus errors are independent... but NOT necessarily normally distributed.
- The dependent variable Y_i does NOT need to be normally distributed, but it assumes a distribution, typically from an exponential family (e.g. binomial, Poisson, gamma,...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but **it does assume a linear relationship between the transformed response (in terms of the link function) and the explanatory variables**; e.g., for binary logistic regression $logit(p) = \beta_0 + \beta_1 X$.
- Explanatory variables can be even the power terms or some other non-linear transformations of the original independent variables.

Assumptions of the Generalized Linear Model

- The data $Y_1, Y_2, \ldots, Y_n$ are independently distributed, i.e., cases are independent.
  - Thus errors are independent... but NOT necessarily normally distributed.
- The dependent variable Y_i does NOT need to be normally distributed, but it assumes a distribution, typically from an exponential family (e.g. binomial, Poisson, gamma,...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but **it does assume a linear relationship between the transformed response (in terms of the link function) and the explanatory variables**; e.g., for binary logistic regression $logit(p) = \beta_0 + \beta_1 X$.
- Explanatory variables can be even the power terms or some other non-linear transformations of the original independent variables.
- The homogeneity of variance does NOT need to be satisfied.

Assumptions of the Generalized Linear Model

- The data $Y_1, Y_2, \ldots, Y_n$ are independently distributed, i.e., cases are independent.
  - Thus errors are independent... but NOT necessarily normally distributed.
- The dependent variable Y_i does NOT need to be normally distributed, but it assumes a distribution, typically from an exponential family (e.g. binomial, Poisson, gamma,...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but **it does assume a linear relationship between the transformed response (in terms of the link function) and the explanatory variables**; e.g., for binary logistic regression $logit(p) = \beta_0 + \beta_1 X$.
- Explanatory variables can be even the power terms or some other non-linear transformations of the original independent variables.
- The homogeneity of variance does NOT need to be satisfied.
- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.

Components of a Generalized Linear Model

Generalized linear models have three parts:

1. **random** component: the response and an associated probability distribution
2. **systematic** component: explanatory variables and relationships among them (e.g., interaction terms)
3. **link function**, which tell us about the relationship between the systematic component (or linear predictor) and the mean of the response

It is the **link function** that allows us to generalize the linear models for count, binomial and percent data. It ensures the linearity and constrains the predictions to be within a range of possible values.

Generalized Linear Models

$$Y_i \sim G(\mu_i, \theta)$$
$$h(\mu_i) = X_i^T \beta$$

- G is the distribution of the response variable
- μ_i is a location parameter for observation i
- θ are additional parameters for the density of G
- h is a link function
- X_i are covariates for observation i
- β is a vector of regression coefficients

Ordinary Least Squares again

GLM

$$\begin{aligned} Y_i &\sim G(\mu_i, \theta) \\ h(\mu_i) &= X_i^T \beta \end{aligned}$$

OLS

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= X_i^T \beta \end{aligned}$$

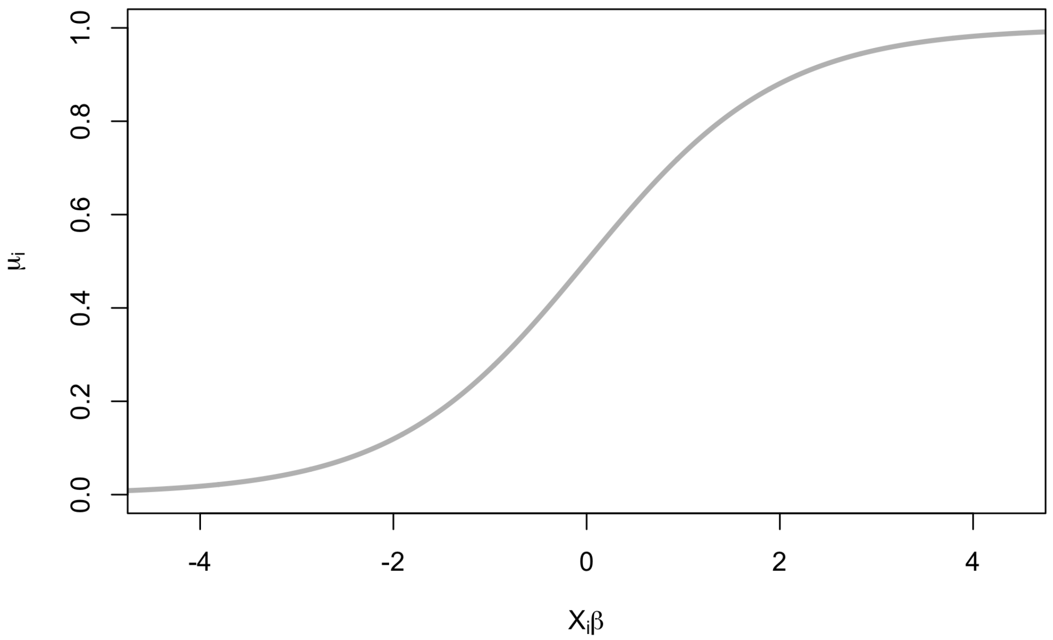
OLS is just a flavour of GLM when:

- G is a Normal distribution
- θ is the variance parameter, denoted σ^2
- h is the identity function

Binomial (or logistic) regression

$$\begin{aligned} Y_i &\sim \text{Binomial}(N_i, \mu_i) \\ \log\left(\frac{\mu_i}{1-\mu_i}\right) &= X_i\beta \end{aligned}$$

- G is a Binomial distribution
- ... or a Bernoulli if $N_i = 1$
- h is the logit link



- $X_i^T\beta$ can be negative
- μ_i is between 0 and 1.

Let's look at an example...

Case study: Challenger disaster

Shuttle data

On January 28, 1986, the Space Shuttle Challenger broke apart 73 seconds into its flight, killing all seven crew members. The spacecraft **disintegrated** over the Atlantic Ocean. The disintegration of the vehicle began after a joint in its right rocket booster failed at liftoff. The failure was caused by the **failure of O-ring seals** used in the joint that were not designed to handle the unusually cold conditions that existed at this launch.

We will look at a data set about the number of rubber O-rings showing thermal distress for 23 flights of the space shuttle, with the ambient temperature and pressure at which tests on the putty next to the rings were performed.

Follow along with the case study

You can follow along with this case study using the template available in the `sta303-bolton/sta303w7` package.

```r
# Only need to download if your didn't for the cheat sheet
# or if your JupyterHub session is new
devtools::install_github("sta303-bolton/sta303w7")

# If it prompts you to update packages,
# just click 'Enter' or 'Return' or input 3
```

Then, go to File -> New File -> R Markdown and choose 'From template' and select 'Challenger case study {sta303w7}'.

Important: If you're working in the JupyterHub, make sure you save the .Rmd in a subfolder of your Home directory so it doesn't cause class interactives to fail in future. (Can't have any .Rmd in your Home directory.)

```r
# install.packages("SMPracticals")
data('shuttle', package='SMPracticals')
rownames(shuttle) = as.character(rownames(shuttle))
shuttle[1:4,]
```

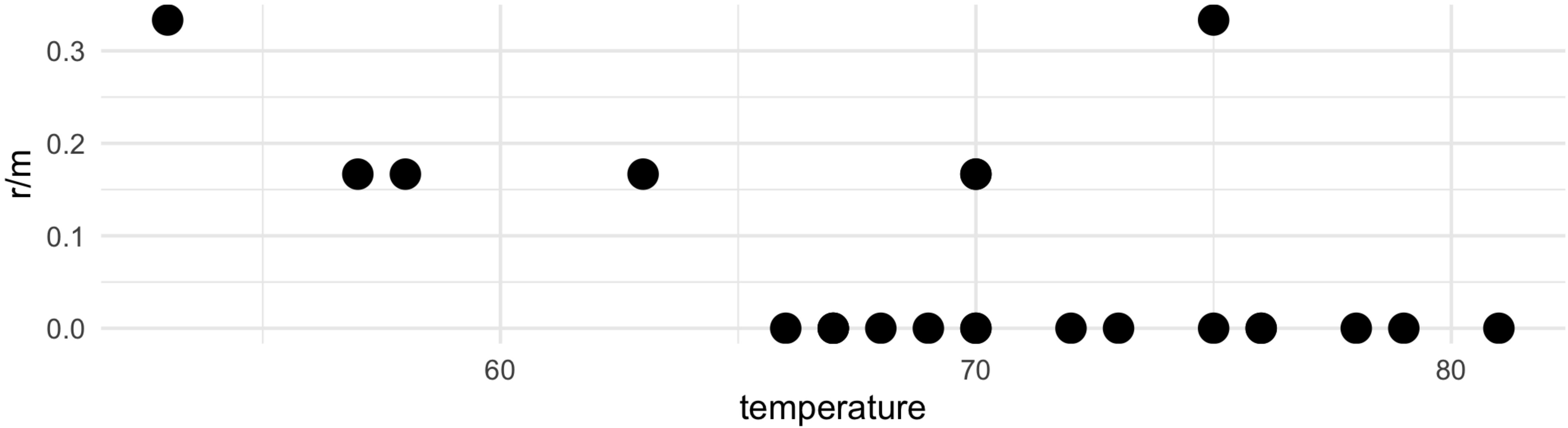
```
##   m r temperature pressure
## 1 6 0          66       50
## 2 6 1          70       50
## 3 6 0          69       50
## 4 6 0          68       50
```

- `m`: number of rings
- `r`: number of damaged rings

Thus we have a situation where we are interested in the number of successes out of a fixed number of trials. Hopefully your memories of the Binomial distribution are being triggered by that language.

```r
# Base R plot
# plot(shuttle$temperature, shuttle$r/shuttle$m)

# ggplot
shuttle %>%
  ggplot(aes(x = temperature, y = r/m)) +
  geom_point(size = 4) +
  theme_minimal()
```



Are shuttle rings more likely to get damaged in cold weather?

We can think of **m** as the number of trials, and **r** as the number of "successes". (It feels weird to call damage a success, but it is our outcome of interest, so we treat it as such.)

$$\begin{aligned} Y_i &\sim \text{Binomial}(N_i, \mu_i) \\ \log\left(\frac{\mu_i}{1-\mu_i}\right) &= X_i \beta \end{aligned}$$

- m: number of rings, N_i
- r: number of damaged rings Y_i
- pressure, temperature: covariates X_i
- μ_i: probability of a ring becoming damaged given X_i
- $\beta_{\text{temperature}}$: parameter of interest

Inference: parameter estimation

$$
\begin{aligned}
Y_i &\sim G(\mu_i, \theta) \\
h(\mu_i) &= X_i\beta \\
\pi(Y_1 \ldots Y_N; \beta, \theta) &= \prod_{i=1}^{N} f_G(Y_i; \mu_i, \theta) \\
\log L(\beta, \theta; y_1 \ldots y_N) &= \sum_{i=1}^{N} \log f_G(y_i; \mu_i, \theta)
\end{aligned}
$$

- The Y_i are *independently distributed*
- **Joint density** π of random variables $(Y_1 \ldots Y_N)$ is the product of the marginal densities f_G.
- **Likelihood function** L given observed data $y_1 \ldots y_N$ is a function of the parameters.
- **Maximum Likelihood Estimation**:

$$\hat{\beta}, \hat{\theta} = \text{argmax}_{\beta,\theta} L(\beta, \theta; y_1 \ldots y_N)$$

- The best parameters are those which are most likely to produce the observed data

Shuttle example in R

- `glm` works like `lm` with a `family` argument.
- Binomial models can take two types of inputs:
  - If, as in this case, we have groups of trials, we need our response to be a matrix with two columns: `y` and `N-y`.
  - If our `y` is a single 0/1 (or otherwise binary categorical variable) then we can set it up as usual, just a single column.

```
shuttle$notDamaged <- shuttle$m - shuttle$r
shuttle$y <- as.matrix(shuttle[,c('r','notDamaged')])
shuttleFit <- glm(y ~ temperature + pressure,
  family=binomial(link='logit'), data=shuttle)
shuttleFit$coef
```

```
##  (Intercept)  temperature     pressure
##  2.520194641 -0.098296750  0.008484021
```

Summary of fit

```
summary(shuttleFit)
```

```
##
## Call:
## glm(formula = y ~ temperature + pressure, family =
binomial(link = "logit"),
##     data = shuttle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0361  -0.6434  -0.5308  -0.1625   2.3418
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   3.486784   0.723   0.4698
## temperature -0.098297   0.044890  -2.190   0.0285
## pressure     0.008484   0.007677   1.105   0.2691
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.546  on 20  degrees of freedom
## AIC: 36.106
##
## Number of Fisher Scoring iterations: 5
```

```
confint(shuttleFit)
```

```
##                    2.5 %      97.5 %
## (Intercept) -4.322926283  9.77264497
## temperature -0.194071699 -0.01356289
## pressure    -0.004346403  0.02885221
```

There is no evidence that pressure is significantly associated with failure of O-rings...but how do we interpret these values?

Interpreting logistic models

$$
\begin{aligned}
Y_i &\sim \text{Binomial}(N_i, \mu_i) \\
\log\left(\frac{\mu_i}{1-\mu_i}\right) &= \sum_{p=1}^{P} X_{ip}\beta_p \\
\left(\frac{\mu_i}{1-\mu_i}\right) &= \prod_{p=1}^{P} \exp(\beta_p)^{X_{ip}}
\end{aligned}
$$

- μ_i is a probability
- $\log[\mu_i/(1-\mu_i)]$ is a log-odds
- $\mu_i/(1-\mu_i)$ is an odds
- If $\mu_i \approx 0$, then $\mu_i \approx \mu_i/(1-\mu_i)$

$$
\begin{aligned}
\beta_q =& \log\left(\frac{\mu_2}{1-\mu_2}\right) - \log\left(\frac{\mu_1}{1-\mu_1}\right) \\
\exp(\beta_q) =& \left(\frac{\mu_2}{1-\mu_2}\right) \Big/ \left(\frac{\mu_1}{1-\mu_1}\right)
\end{aligned}
$$

- β_q is the log-odds ratio
- $\exp(\beta_q)$ is the odds ratio
- $\exp(\text{intercept})$ is the baseline odds, when $X_1 \ldots X_n = 0$.

Centring parameters

```r
quantile(shuttle$temperature)
```

```
##    0%   25%   50%   75%  100% 
##    53    67    70    75    81 
```

```r
quantile(shuttle$pressure)
```

```
##    0%   25%   50%   75%  100% 
##    50    75   200   200   200 
```

- Currently the intercept is log-odds when temperature = 0 and pressure = 0
- centre the covariates so the intercept refers to:
  - temperature = 70 (degrees Farenheit)
  - pressure = 200 (pounds per square inch)

```r
shuttle$temperatureC <- shuttle$temperature  - 70
shuttle$pressureC <-  shuttle$pressure - 200
shuttleFit2 <-  glm(y ~ temperatureC + pressureC, family='binomial', data=shuttle)
```

Shuttle odds parameters

```r
par_table = cbind(est = summary(
  shuttleFit2)$coef[,1],
  confint(shuttleFit2))
rownames(par_table)[1]= "Baseline"
```

```r
round(exp(par_table),3)
```

```
##                est 2.5 % 97.5 %
## Baseline     0.070 0.023  0.155
## temperatureC 0.906 0.824  0.987
## pressureC    1.009 0.996  1.029
```

Table 1: MLEs of baseline odds and odds ratios, with 95% confidence intervals.

Interpreting shuttle parameters

- The odds of a ring being damaged when temperature = 70 and pressure = 200 is 0.0697, which corresponds to a probability of

```
round(exp(par_table[1,'est']) / (1+exp(par_table[1,'est'])), 3)
```

```
## [1] 0.065
```

- Each degree increase in temperature (in Fahrenheit) decreases the odds of damage by (in percent)

```
round(100*(1-exp(par_table[2,'est']) ), 3)
```

```
## [1] 9.362
```

Week 7 learning checklist

By the end of week 7, you should be able to:

- Recognize a form of the probability density function for Bernoulli, binomial, negative binomial, Poisson, gamma and
- Identify how changing values for a parameter affects the characteristics of the probability distribution.
- Identify the mean and variance for each distribution.
- Match the response for a study to a plausible random variable and provide reasons for ruling out other random variables.
- Match a histogram of sample data to plausible distributions.
- Create tables and calculate joint, marginal and conditional probabilities with them.
- Calculate odds, risks, odds ratios (OR) and risk ratios (RR).
- Understand why ORs and RRs are similar for rare outcomes.
- State the assumptions of GLMs.
- Interpret logistic regression output (more next week.)

Extra for the curious (NOT assessed)

The information on the following slides is **not assessed in this course**.

Consider taking a course like STA442 to go deeper!

Efficient maximization (for your reference only)

- Iteratively Reweighted Least Squares is the 'classic' algorithm when G is in the exponential family
- ... but GLMs are easy for any density which is differentiable
- The derivatives with respect to β are easy to compute with the chain rule

$$\frac{\partial}{\partial \beta_p} \log L(\beta, \theta; y_1 \ldots t_N) = \sum_{i=1}^{N} \left[\frac{d}{d\mu} \log f_G(Y_i; \mu, \theta) \right]_{\mu = h^{-1}(X_i^T \beta)} \left[\frac{d}{d\eta} h^{-1}(\eta) \right]_{\eta = X_i^T \beta} \cdot X_{ip}$$

- Analytical expressions exist for the derivatives of $\log f_G$ and h^{-1}
- Second derivatives are also tractable
- Numerical maximization to find $\hat{\beta}$ is fast when derivatives are available

Numerical maximizers (for your reference only)

- There are hundreds of them!
- `optim` is the standard `R` optimizer, which has 6 methods available.
    - some methods will use gradients if you provide them.
- `TrustOptim` uses derivatives and 'trust regions', the method used in INLA.
- `ipopt` is probably the cutting edge.
- Statisticians don't make enough use of of-the-shelf optimizers.

Automatic differentiation (for your reference only)

$$\sum_{i=1}^{N} \left[\frac{d}{d\mu} \log f_G(Y_i; \mu, \theta) \right]_{\mu = h^{-1}(X_i^T \beta)} \left[\frac{d}{d\eta} h^{-1}(\eta) \right]_{\eta = X_i^T \beta} \cdot X_{ip}$$

- Overkill for most GLMs, but infinitely extensible.
- Computers evaluate logs, sines, and other functions through some Taylor-series-like polynomial thing.
- ... which are easy to differentiate!
- AD programs can take computer code and figure out how to differentiate it.
- Used in Neural Nets, Hamiltonian MCMC, optimization, and many more.

Go to topics list

See you Wednesday for class!