Topic: Generalized Linear Models

Models can handle more complicated situations and analyze the simultaneous effects of multiple variables, including mixtures of categorical and continuous variables. For example, the Breslow-Day statistics only works for $2 \times 2 \times K$ tables, while log-linear models will allow us to test of homogeneous associations in $I \times J \times K$ and higher-dimensional tables. We will focus on a special class of models known as the ***generalized linear models (GLIMs or GLMs*** in Agresti***)***.

The structural form of the model describes the patterns of interactions and associations. The model parameters provide measures of strength of associations. In models, ***the focus is on estimating the model parameters.*** The basic inference tools (e.g., point estimation, hypothesis testing, and confidence intervals) will be applied to these parameters. When discussing models, we will keep in mind:

- Objective
- Model structure (e.g. variables, formula, equation)
- Model assumptions
- Parameter estimates and interpretation
- Model fit (e.g. goodness-of-fit tests and statistics)
- Model selection

For example, recall a *simple linear regression model*

- *Objective:* model the expected value of a continuous variable, Y, as a linear function of the continuous predictor, X, $E(Y_i) = \beta_0 + \beta_1 x_i$
- *Model structure:* $Y_i = \beta_0 + \beta_1 x_i + e_i$
- *Model assumptions:* Y is is normally distributed, errors are normally distributed, $e_i \sim N(0, \sigma^2)$, and independent, and X is fixed, and constant variance σ^2.
- *Parameter estimates and interpretation:* ^β0β^0 is estimate of β_0 or the intercept, and ^β1β^1 is estimate of the slope, etc... Do you recall, what is the interpretation of the intercept and the slope?
- *Model fit:* R^2, residual analysis, F-statistic
- *Model selection:* From a plethora of possible predictors, which variables to include?

For a review, if you wish, see a handout labeled LinRegExample.doc [1] on modeling average water usage given the amount of bread production, e.g., estimated water production is positively related to the bread production:

Water = 2273 + 0.0799 Production

Generalized Linear Models (GLMs)

First, let’s clear up some potential misunderstandings about terminology. The term *general linear model* (GLM) usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors. It includes multiple linear regression, as well as ANOVA and ANCOVA (with fixed effects only). The form is yi∼N(xTiβ,σ2),yi∼N(xiTβ,σ2), where xixi contains known covariates and ββ contains the coefficients to be estimated. These models are fit by least squares and weighted least squares using, for example: SAS Proc GLM or R functions lsfit() (older, uses matrices) and lm() (newer, uses data frames).

The term ***generalized linear model*** (GLIM or GLM) refers to a larger class of models popularized by McCullagh and Nelder (1982, 2nd edition 1989). In these models, the response variable yiyi is assumed to follow an exponential family distribution with mean μiμi, which is assumed to be some (often nonlinear) function of xTiβxiTβ. Some would call these “nonlinear” because μiμi is often a nonlinear function of the covariates, but McCullagh and Nelder consider them to be linear, because the covariates affect the distribution of yiyi only through the linear combination xTiβxiTβ. The first widely used software package for fitting these models was called GLIM. Because of this program, “GLIM” became a well-accepted abbreviation for generalized linear models, as opposed to “GLM” which often is used for general linear models. Today, GLIM’s are fit by many packages, including **SAS Proc Genmod** and **R function glm()**. Notice, however, that Agresti uses GLM instead of GLIM short-hand, and we will use GLM.

Model	**Random**	**Link**	**Systematic**
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical

Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

The generalized linear models (GLMs) are a broad class of models that include linear regression, ANOVA, Poisson regression, log-linear models etc. The table below provides a good summary of GLMs following Agresti (ch. 4, 2013):

There are three components to any GLM:

- *Random Component* – refers to the probability distribution of the response variable (Y); e.g. normal distribution for Y in the linear regression, or binomial distribution for Y in the binary logistic regression. Also called a noise model or error model. How is random error added to the prediction that comes out of the link function?

- *Systematic Component* - specifies the explanatory variables (X_1, X_2, ... X_k) in the model, more specifically their linear combination in creating the so called *linear predictor*; e.g., $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ as we have seen in a linear regression, or as we will see in a logistic regression in this lesson.

- *Link Function, η or $g(\mu)$* - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables; e.g., $\eta = g(E(Y_i)) = E(Y_i)$ for linear regression, or $\eta = logit(\pi)$ for logistic regression.

Assumptions:

- The data Y_1, Y_2, ..., Y_n are independently distributed, i.e., cases are independent.
- The dependent variable Y_i does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship

between the transformed response in terms of the link function and the explanatory variables; e.g., for binary logistic regression *logit*(π) = $\beta_0 + \beta X$.

- Independent (explanatory) variables can be even the power terms or some other nonlinear transformations of the original independent variables.
- The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure, and *overdispersion* (when the observed variance is larger than what the model assumes) maybe present.
- Errors need to be independent but NOT normally distributed.
- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.
- Goodness-of-fit measures rely on sufficiently large samples, where a heuristic rule is that not more than 20% of the expected cells counts are less than 5.

For a more detailed discussion refer to Agresti(2007), Ch. 3, Agresti (2013), Ch.4, and/or McCullagh & Nelder (1989).

Following are examples of GLM components for models that we are already familiar, such as linear regression, and for some of the models that we will cover in this class, such as logistic regression and log-linear models.

Simple Linear Regression models how mean expected value of a continuous response variable depends on a set of explanatory variables, where index *i* stands for each data point:

Yi=β0+βxi+ϵiYi=β0+βxi+ϵi

or

E(Yi)=β0+βxiE(Yi)=β0+βxi

- *Random component: Y* is a response variable and has a normal distribution, and generally we assume errors, $e_i \sim N(0, \sigma^2)$.
- *Systematic component: X* is the explanatory variable (can be continuous or discrete) and is linear in the parameters $\beta_0 + \beta x_i$. Notice that with a multiple linear regression where we have more than one explanatory variable, e.g., (X_1, X_2, ... X_k), we would have a linear combination of these *Xs* in terms of regression parameters *β's*, but the explanatory variables themselves could be transformed, e.g., X^2, or *log(X).*

- *Link function: Identity Link,* $\eta = g(E(Y_i)) = E(Y_i)$ --- identity because we are modeling the mean directly; this is the simplest link function.

Binary Logistic Regression models how binary response variable *Y* depends on a set of *k* explanatory variables, $X=(X_1, X_2, ... X_k)$.

logit(π)=log(π1−π)=β0+βxi+…+β0+βxk′logit(π)=log(π1−π)=β0+βxi+…+β0+βxk′

which models the log odds of probability of "success" as a function of explanatory variables.

- *Random component:* The distribution of *Y* is assumed to be *Binomial(n,π), where π is a probability of "success".*
- *Systematic component: X*'s are explanatory variables (can be continuous, discrete, or both) and are linear in the parameters, e.g., $\beta_0 + \beta x_i + ... + \beta_0 + \beta x_k$. Again, transformation of the X's themselves are allowed like in linear regression; this holds for any GLM.
- *Link function: Logit link:*

  η=logit(π)=log(π1−π)η=logit(π)=log(π1−π)

  More generally, the logit link models the log odds of the mean, and the mean here is π. Binary logistic regression models are also known as logit models when the predictors are all categorical.

Log-linear Model models the expected cell counts as a function of levels of categorical variables, e.g., for a two-way table the saturated model

log(μij)=λ+λAi+λBj+λABijlog(μij)=λ+λiA+λjB+λijAB

where $\mu_{ij}=E(n_{ij})$ as before are expected cell counts (mean in each cell of the two-way table), *A* and *B* represent two categorical variables, and λ_{ij}*'s* are model parameters, and we are modeling the natural log of the expected counts.

- *Random component:* The distribution of counts, which are the responses, is *Poisson*
- *Systematic component: X's* are discrete variables used in cross-classification, and are linear in the parameters λ+λX1i+λX2j+…+λXkk+…λ+λiX1+λjX2+…+λkXk+…
- *Link Function: Log link,* $\eta = log(\mu)$ --- *log* because we are modeling the log of the cell means.

The log-linear models are more general than logit models, and some logit models are equivalent to certain log-linear models [2]. Log-linear model is also equivalent to Poisson regression model when all explanatory variables are discrete. For additional details see Agresti(2007), Sec. 3.3, Agresti (2013), Section 4.3 (for counts), Section 9.2 (for rates), and Section 13.2 (for random effects).

Topic: Hierarchical and Multilevel Models

Hierarchical linear models and multilevel models are variant terms for what are broadly called linear mixed models (LMM). These models handle data where observations are not independent, correctly modeling correlated error. Uncorrelated error is an important but often violated assumption of statistical procedures in the general linear model family, which includes analysis of variance, correlation, regression, and factor analysis. Violations occur when error terms are not independent but instead cluster by one or more grouping variables. For instance, predicted student test scores and errors in predicting them may cluster by classroom, school, and municipality. When clustering occurs due to a grouping factor (this is the rule, not the exception), then the standard errors computed for prediction parameters will be wrong (ex., wrong b coefficients in regression).

Linear mixed modeling, including hierarchical linear modeling, can lead to substantially different conclusions compared to conventional regression analysis.

Linear mixed models are a generalization of general linear models to better support analysis of a continuous dependent variable for the following:

1. *Random effects:* For when the set of values of a categorical predictor variable are seen not as the complete set but rather as a random sample of all values (ex., when the variable “product” has values representing only 30 of a possible 142 brands). Random effects modeling allows the researcher to make inferences over a wider population than is possible with regression or other general linear model (GLM) methods.

2. *Hierarchical effects:* For when predictor variables are measured at more than one level (ex., reading achievement scores at the student level and teacher–student ratios at the school level; or sentencing lengths at the offender level, gender of judges at the court level, and budgets of judicial districts at the district level). The researcher can assess the effects of higher levels on the intercepts and coefficients at the lowest level (ex., assess judge-level effects on predictions of sentencing length at the offender level).

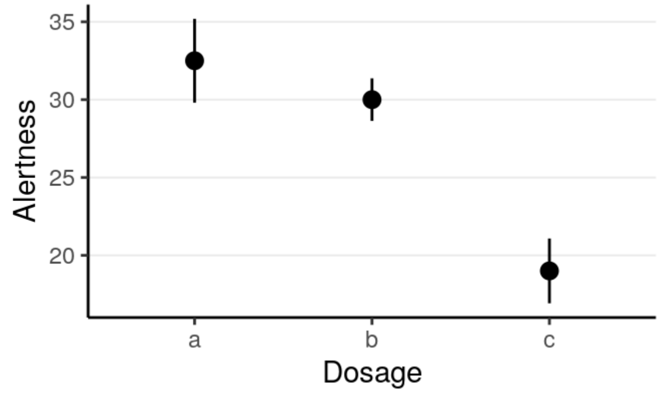
3. *Repeated measures:* For when observations are correlated rather than independent (ex., before–after studies, time series data, matched-pairs designs). In repeated measures, the lowest level is the observation level (ex., student test scores on multiple occasions), grouped by observation unit (ex., students) such that each unit (student) has multiple data rows, one for each observation occasion.

Following sections elaborates the Models.

1 ANOVA

We'll use some demonstration data that usually corresponds to a typical psychological experiment that uses one-way ANOVA:

```
>#  No summary function supplied, defaulting to `mean_se()
```



As can be seen, Dosage `c` has lower mean than others.

1.1 "Frequentist" ANOVA

In frequentist analyses, we generally first perform an omnibus test:

```
>#             Df Sum Sq Mean Sq F value  Pr(>F)
>#  Dosage       2    619   309.5    11.5 0.00094 ***
>#  Residuals   15    404    26.9
>#  ---
>#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And then there will be post hoc comparisons with adjustment on \(p\) values

```
>#
>#  Pairwise comparisons using t tests with pooled SD
>#
>#  data:  alert$Alertness and alert$Dosage
>#
>#    a     b
>#  b 0.417 -
>#  c 0.001 0.005
>#
>#  P value adjustment method: holm
```

which shows that `c` was lower than both `a` and `b`.

1.2 Bayesian ANOVA

In Bayesian, it is more common to treat grouping variables, especially with more than three or four categories, as clusters in hierarchical modeling. Specifically, we start with the normal model: \[\texttt{Alertness}_{ij} \sim \mathcal{N}(\mu_j, \sigma)\] but in the priors, we assume that the \(\mu_j\)s are exchangeable and have a common prior distribution such that \[\mu_j \sim \mathcal{N}(\gamma, \tau)\] This means that we believe the group means themselves are from a normal distribution with mean \(\gamma\) and *SD* \(\tau\). \(\gamma\) is the grand mean of `Alertness` averaged across the conditions, and \(\tau\) is the between-condition *SD*. They are called hyperparameters, and they also need priors (i.e., hyperpriors). Because the prior for \(\mu_j\) consists of hyperparameters that themselves have prior (hyperprior) distributions, this is also called *hierarchical priors*. We'll use: \[\begin{align*} \gamma & \sim \mathcal{N}(0, 50) \\ \tau & \sim \textrm{Gamma}(2, 1 / 8) \end{align*}\] Note that the Gamma prior was recommended in previous papers for hierarchical models, with the 8 in 1/8 being the prior belief of what the maximum value of \(\tau\) can be.

```
m1 <- brm(Alertness ~ 1 + (1 | Dosage), data = alert,
          prior = c(# for gamma
            prior(normal(0, 50), class = "Intercept"),
            # for sigma
            prior(student_t(4, 0, 10), class = "sigma"),
            # for tau
```

```
      prior(gamma(2, 0.125), class = "sd", coef = "Intercept",
         group = "Dosage")
     ),
     # Hierarchical models generally require smaller stepsize
     control = list(adapt_delta = .99))
```

Term	estimate	std.error	lower	upper
b_Intercept	26.97	7.46	15.34	38.45
sd_Dosage__Intercept	11.01	6.17	4.10	22.70
sigma	5.63	1.16	4.04	7.75
r_Dosage[a,Intercept]	5.11	7.61	-6.05	17.31
r_Dosage[b,Intercept]	2.80	7.56	-8.67	14.97
r_Dosage[c,Intercept]	-7.29	7.64	-19.25	3.92
lp__	-67.11	2.16	-71.12	-64.34

From the results, the posterior mean for \(\gamma\) is 26.973 (*SD* = 7.461), which was the grand mean Alertness level. The between-group *SD* was estimated to be \(\tau\) = 11.006, whereas the within-group *SD* was estimated to be \(\sigma\) = 5.628.

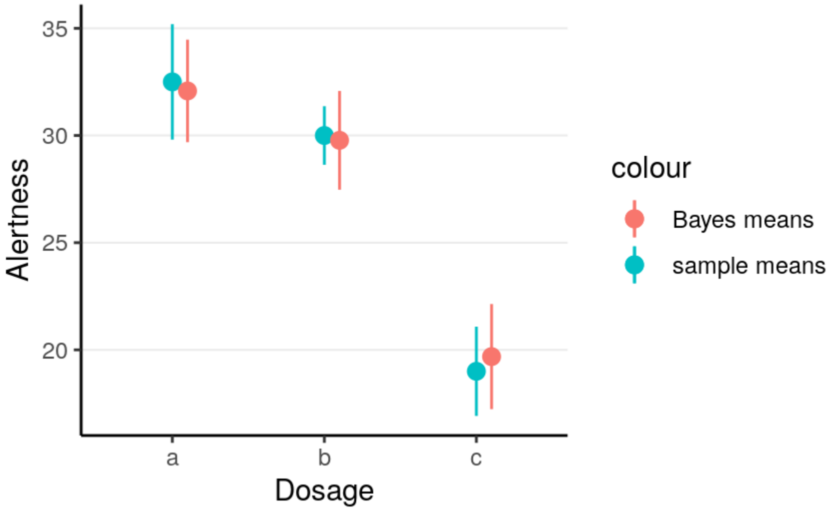
You can get the posterior mean for the mean of each group (i.e., \(\mu_j\)) using

```
>#   Estimate Est.Error Q2.5 Q97.5
># a     32.1      2.39 27.2  36.9
># b     29.8      2.30 25.1  34.4
># c     19.7      2.45 14.9  24.8
```

1.2.1 Shrinkage

Note that in the above model, the Bayes estimates of the group means are different from the sample group means, as shown in the following graph:

```
># No summary function supplied, defaulting to `mean_se()
```

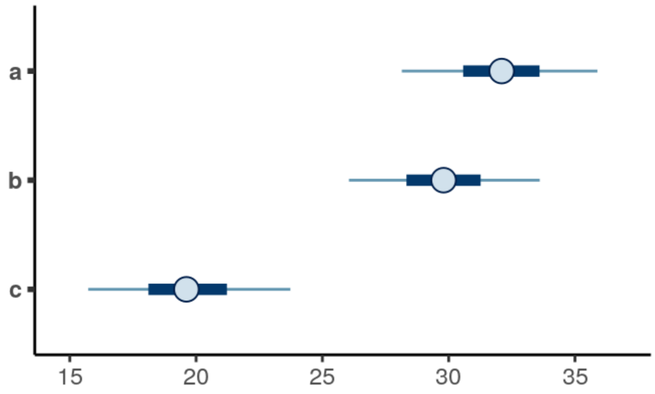


If you look more carefully, you can see that the Bayes estimates are closer to the middle. This *shrinkage* effect may seem odd at first, but it has a good reason. The hierarchical assumes that there are something in common for observations in different groups, so it performs *partial pooling* by borrowing information from other groups.

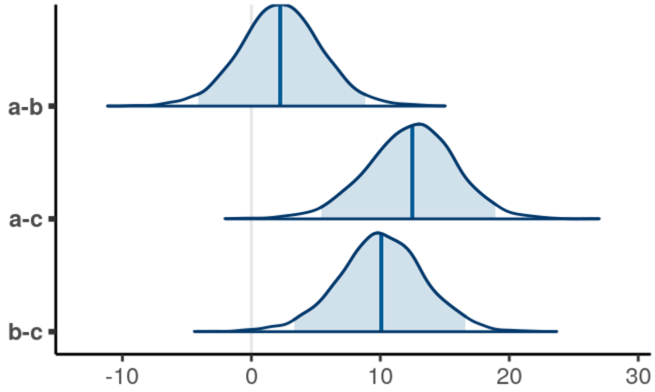
To illustrate the strength of partial pooling, I went through a thought experiment with my students in my multilevel modeling class. Imagine it’s your first time visiting Macau, my hometown, and you are about to go to a McDonald’s there. You’ve never been to any restaurants in Macau. So what do you expect? You probably will use your experience of eating at McDonald’s in the US as a reference. The Bayesian hierarchical model here is the same: it assumes that even though participants received different Dosage, there are something similar among them, so information from one group should provide some information for another group. And for many of our problems in research, hierarchical models have been shown to make better predictions and inferences, compared to traditional ANOVA. See Kruschke and Liddell (2018) for some more discussion.

1.2.2 Notes on multiple comparisons

With hierarchical models, the common recommendation is that no further control for multiple comparison is needed (see Gelman, Hill, and Yajima 2012). For one, we don’t use \(p\) values in Bayesian. For the other, by shrinking the group means closer to the grand mean in a hierarchical model, the comparisons in some sense have already been adjusted. You can plot the estimated group means by:



And below it shows the posterior of the differences:



And the results in this example are similar to the post hoc comparisons.

2 Multilevel Modeling (MLM)

Multilevel modeling is the set of techniques that built on the previous hierarchical model. It is proposed kind of separately in multiple disciplines, including education and other social sciences, and so historically it has been referred to by many different names, such as:

- Mixed/Mixed-effect models
- Hierarchical linear models
- Variance component models

It allows us to build models on different groups/clusters, and allows the parameters to be different across clusters. However, it does *partial pooling* by borrowing information from one cluster to another, which is especially beneficial when some groups have only a few people, where borrowing information from other clusters would help stabilize the parameter estimates.

2.1 Examples of clustering

There are many different forms of clustering in data across different disciplines. We’ve seen the example of people clustered in experimental conditions. Other examples include:

- Students in schools
- Clients nested within therapists within clinics
- Employees nested within organizations
- Citizens nested within employees
- Repeated measures nested within persons

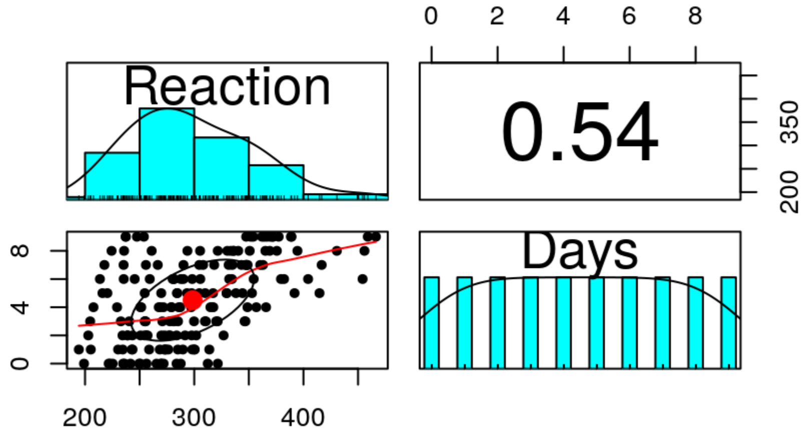
They can be represented in network graphs like the following (students within schools):

Sometimes there are more than one level of clustering, like students clustered by both middle schools and high schools. This is called a *crossed* structure as shown in the following, where we say that students are cross-classified by both middle and high schools. Another example commonly happened in psychological experiments is when participants see multiple stimuli, each as an item, so the observations are cross-classified by both persons and items.

The repeated measures nested within persons one is particularly relevant as that means essentially all longitudinal data are multilevel data and should be modelled accordingly. It allows one to build individualized model to look at within-person changes, as well as between-person differences of those changes. Techniques such as dependent-sample \(t\)-test, repeated-measures ANOVA, growth curve modeling, and time-series analyses, can all be represented in the multilevel modeling framework. Therefore, some authors, such as McElreath (2016), would suggest that MLM should be the default model that we use for analyses, rather than regression.

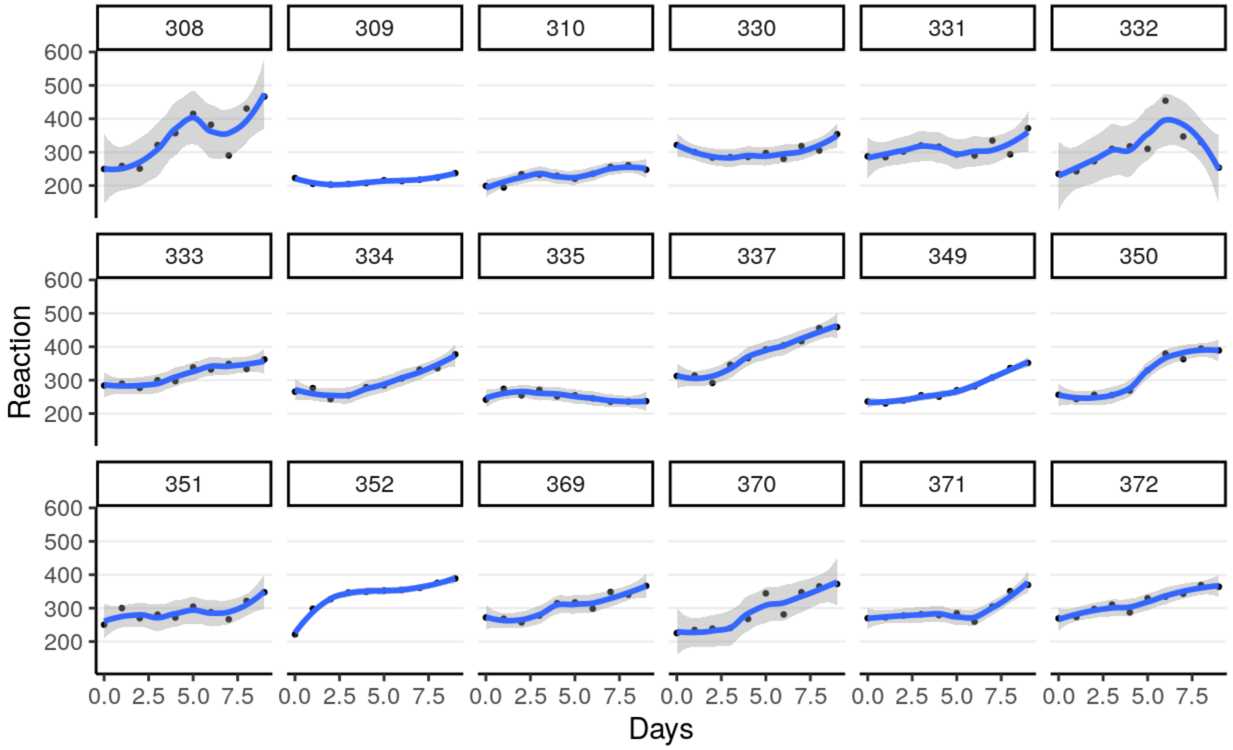
2.2 Data

We will use the data set sleepstudy from the lme4 package, which is the package for frequentist multilevel modeling. The data set contains 18 participants, each with 10 observations. It examines the change in average reaction time per day with increasing sleep deprivation. See `?lme4::sleepstudy` for more of the description. Here is a plot of the data:



This data set has clustering because it is repeated measures nested within persons. It is more useful to plot the change in the outcome:

```
># `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



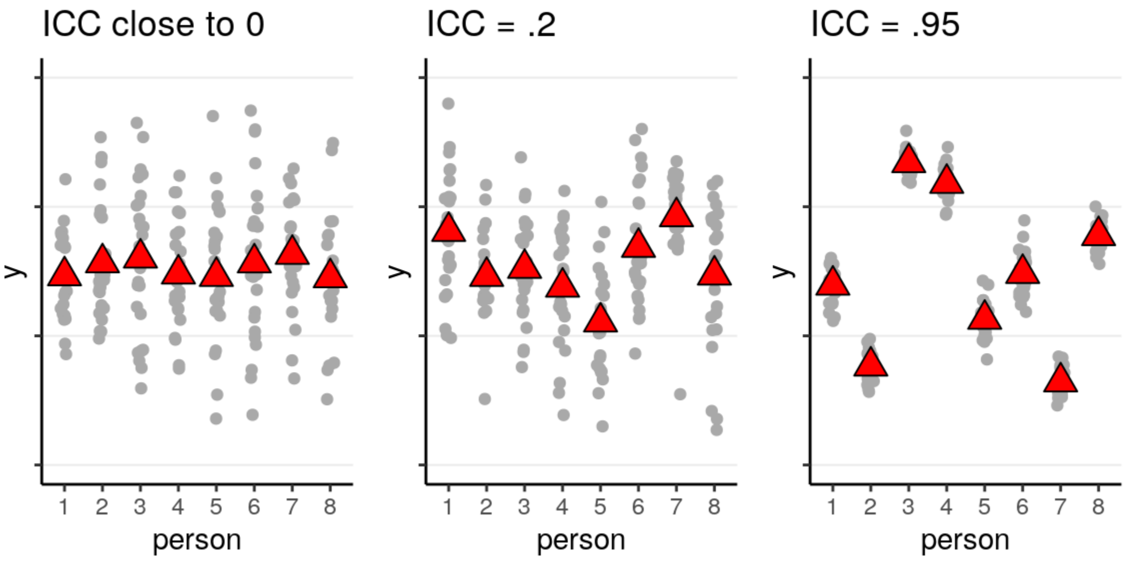
As you can see, most people experience increases in reaction time, although there are certainly differences across individuals.

2.3 Intraclass correlation

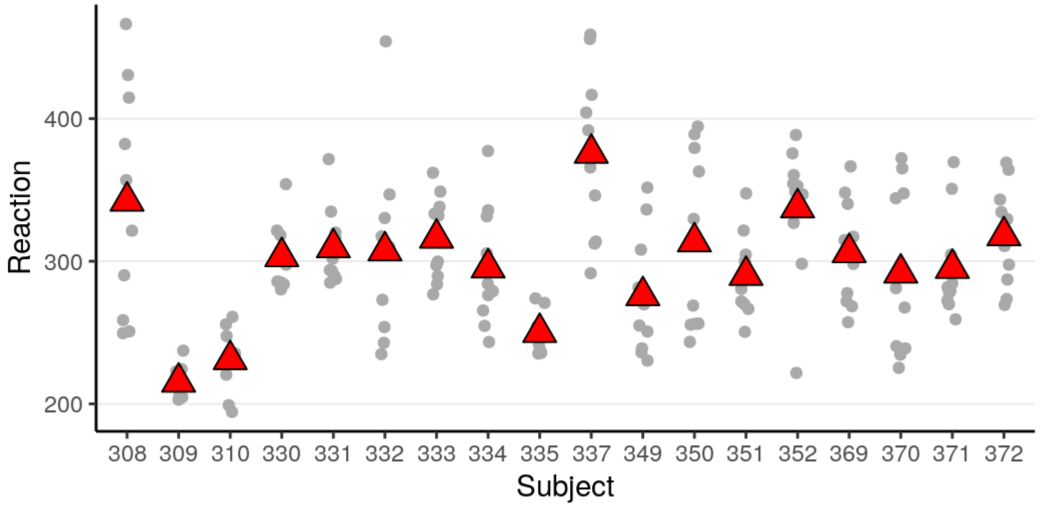
With multilevel data, the first question to ask is how much variation in the outcome is there at each level. This is quantified by the *intraclass correlation*, which, for a two-level model, is defined by $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$ where τ is the between-level *SD*, which is the *SD* of the cluster means (i.e., the variability of mean response time across persons in this example), and σ is the within-level *SD* (i.e., variability within a person, which is assumed constant across persons).

The ICC represents the proportion of variance of the outcome that are due to between-level (e.g., between-group, between-person) differences

Here is a graph from my MLM class showing how the data would be like with different ICC levels:



As you can see, the higher the ICC, the higher the variations in the cluster means, relative to the within-cluster variations. Below is the graph for the sleepstudy data:



Which has substantial between-person variations.

2.3.1 Computing ICC

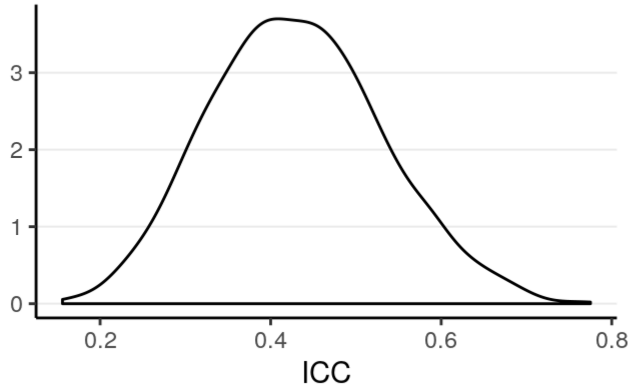
To compute the ICC, we need to first fit a multilevel model, which in this case is the *varying intercept* model: \[\begin{align*} \texttt{Reaction}_{ij} & \sim \mathcal{N}(\mu_j, \sigma) \\ \mu_j & \sim \mathcal{N}(\gamma, \tau) \end{align*}\] where \(\mu_j\) is the mean reaction for the \(j\)th person, and \(i\) indexes measurement occasions.

We'll rescale `Reaction` by 10:
To use weakly informative priors, we will set \[\begin{align*} \gamma & \sim \mathcal{N}(0, 50) \\ \sigma & \sim t^+(4, 0, 5) \\ \tau & \sim \textrm{Gamma}(2, 1 / 5) \end{align*}\]

```
m2 <- brm(Reaction10 ~ (1 | Subject), data = sleepstudy,
          prior = c(# for intercept
            prior(normal(0, 50), class = "Intercept"),
            # for tau
            prior(gamma(2, 0.2), class = "sd"),
            # for sigma
            prior(student_t(4, 0, 5), class = "sigma")),
          control = list(adapt_delta = .95),
          cores = 2L,
          seed = 2107)
```

Now use the posterior draws of \(\tau\) and \(\sigma\) to compute the posterior for the ICC:

term	estimate	std.error	lower	upper
b_Intercept	29.90	1.016	28.20	31.58
sd_Subject__Intercept	3.93	0.838	2.75	5.45
sigma	4.46	0.248	4.06	4.88



```
>#    vars    n mean  sd median trimmed mad  min  max range skew kurtosis se
>#  X1    1 4000 0.43 0.1   0.43    0.43 0.1 0.16 0.78  0.62  0.2    -0.21  0
```

2.3.2 Interpretations

The model suggested that the average reaction time across individuals and measurement occasions was 298.988 ms, 95% CI [278.958, 319.692]. It was estimated that 43.007%, 95% CI [24.463%, 63.909%] of the variations in reaction time was attributed to between-person differences.

2.4 Is MLM needed?

This is a commonly asked question. Based on Lai and Kwok (2015), you can compute the design effect index, which shows the inflation in variability of the estimates due to clustering. It is recommended to account for clustering if the design effect is larger than 1.1. It is defined as: \[\mathit{Deff}= 1 + (n - 1) \rho\] where \(n\) is the (average) number of observations in each cluster, and in our case it is 10. Therefore, the design effect in sleepstudy for Reaction is \[\mathit{Deff}= 1 + (10 - 1) (0.43)\] which is 4.871, so we do need to account for the clustering.

3 Varying Coefficients

The strength of a multilevel model is that it can allow researchers to build models that allow for cluster-specific coefficients. In our example data this is analogous to fitting separate models for each person, but instead of only using 10 data points for each model, MLM pools information from other people as it believes that we can learn something about one person by looking at data from other people.

For example, for each person, we'll fit a regression model using Days to predict Reaction10. Using our previous notations, \[\begin{align} \texttt{Reaction10}_i & \sim \mathcal{N}(\mu_i, \sigma) \\ \mu_i & = \beta_0 + \beta_1 \texttt{Days}_i \end{align}\] However, because we have more than one person, we'll use the subscript \(j\) to denote the person, so that the model becomes \[\begin{align} \texttt{Reaction10}_{ij} & \sim \mathcal{N}(\mu_{ij}, \sigma_j) \\ \mu_{ij} & = \beta_{0j} + \beta_{1j} \texttt{Days}_{ij} \end{align}\] which suggests that all three of \(\beta_0\), \(\beta_1\), and \(\sigma\) can be different across persons. We'll first start with varying \(\beta_0\), or *varying intercepts*.

3.1 Varying Intercepts

With varying intercepts model, we assumed that only \(\beta_0\) is different across persons, but \(\beta_1\) and \(\sigma\) are common parameters that do not change across persons. This is also referred to as a *random intercept model* in (frequentist) MLM literature. Specifically, the model and priors are: \[\begin{align} \text{Repeated-measure level:} \\ \texttt{Reaction10}_{ij} & \sim \mathcal{N}(\mu_{ij}, \sigma) \\ \mu_{ij} & = \beta_{0j} + \beta_{1} \texttt{Days}_{ij} \\ \text{Person level:} \\ \beta_{0j} & \sim \mathcal{N}(\mu^{[\beta_0]}, \tau^{[\beta_0]}) \\ \text{Priors:} \\ \mu^{[\beta_0]} & \sim \mathcal{N}(0, 50) \\ \tau^{[\beta_0]} & \sim \mathrm{Gamma}(2, 0.2) \\ \beta_1 & \sim \mathcal{N}(0, 10) \\ \sigma & \sim t^+(4, 0, 5) \end{align}\] where the \(\beta_{0j}\)s follow a common normal distribution with hyperparameters \(\mu^{[\beta_0]}\) and \(\tau^{[\beta_0]}\). Thus, \(\mu^{[\beta_0]}\) is the *grand intercept*, or the average intercept across persons, and \(\tau^{[\beta_0]}\) is the *SD* of those intercepts.

The model can be fitted in brms:

```
m3 <- brm(Reaction10 ~ Days + (1 | Subject), data = sleepstudy,
          prior = c(# for intercept
            prior(normal(0, 50), class = "Intercept"),
            # for slope
            prior(normal(0, 10), class = "b"),
            # for tau
            prior(gamma(2, 0.2), class = "sd"),
            # for sigma
            prior(student_t(4, 0, 5), class = "sigma")),
          control = list(adapt_delta = .95),
          cores = 2L,
          seed = 2107)
```

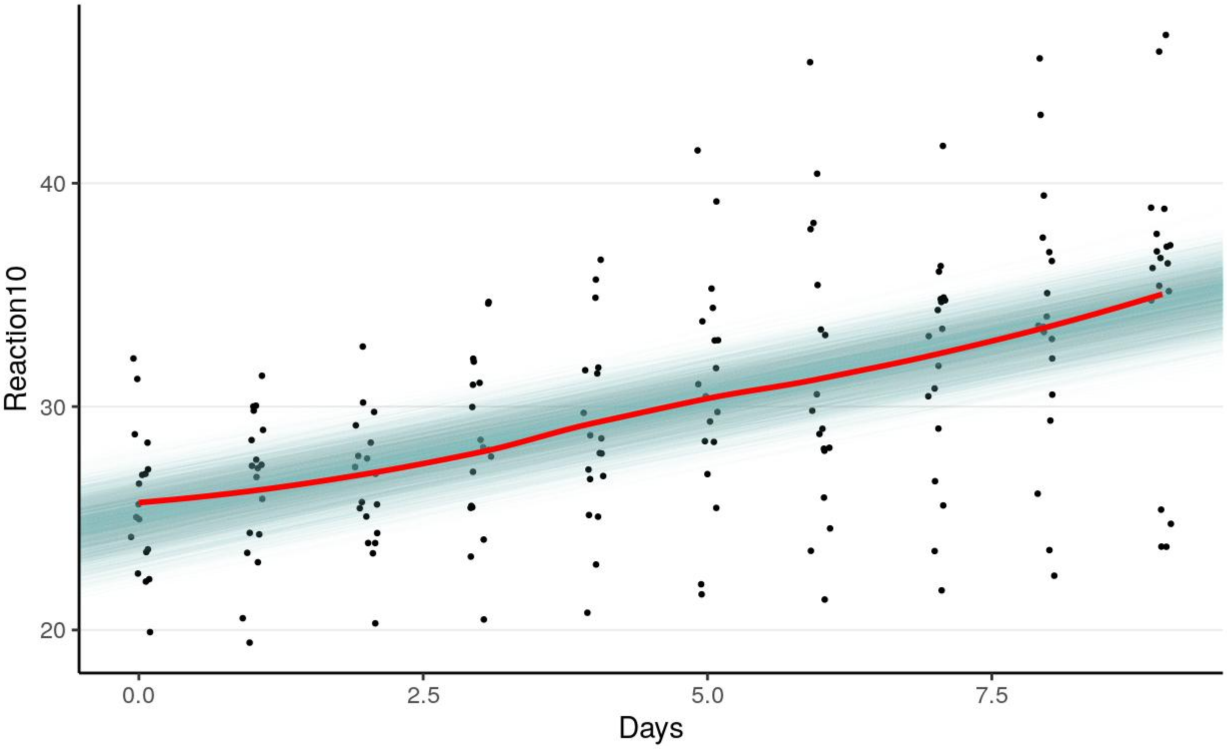
Below is a summary table of the results

Term	estimate	std.error	Lower	upper
b_Intercept	25.11	1.043	23.297	26.80
b_Days	1.04	0.084	0.909	1.19
sd_Subject__Intercept	4.08	0.831	2.953	5.65
Sigma	3.12	0.178	2.845	3.43

Let’s check the fit of the model to the data, first to the overall data and then to each individual specifically.

3.1.1 Fit of Overall data

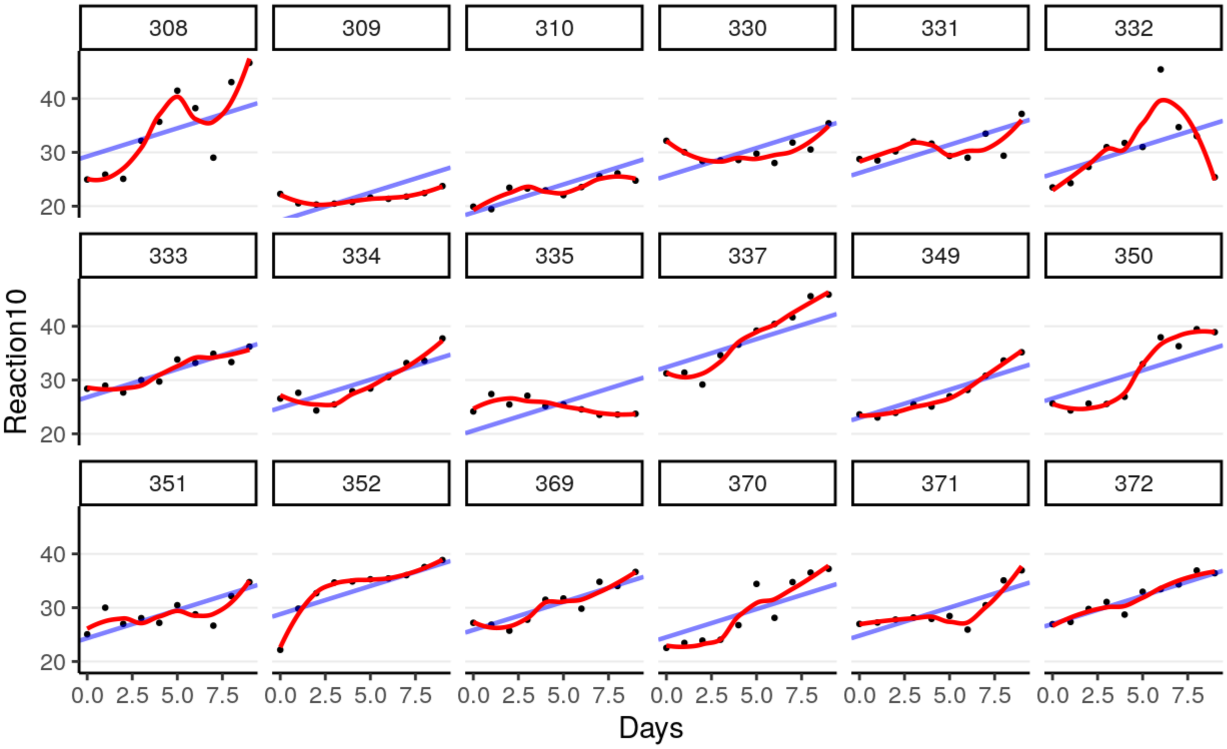
```
># `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



As can be seen, the estimated coefficient for `Days`, which was assumed constant for everyone, fit the overall data. However, does it fit each individual?

3.1.2 Fit of Individuals

```
># `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Obviously it only fit a few individuals, but not all. So let's also allow β_1 to vary.

3.2 Varying Slopes

We'll now also allow β_1 to vary across clusters, with the following model:

$$\begin{aligned} \text{Repeated-measure level:} \\ \texttt{Reaction10}_{ij} & \sim \mathcal{N}(\mu_{ij}, \sigma) \\ \mu_{ij} & = \beta_{0j} + \beta_{1j} \texttt{Days}_{ij} \\ \text{Person level:} \\ \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \end{bmatrix} & \sim \mathcal{N}_2\left(\begin{bmatrix} \mu^{[\beta_0]} \\ \mu^{[\beta_1]} \\ \end{bmatrix}, \boldsymbol{\mathbf{T}} \right) \end{aligned}$$

where

$$\boldsymbol{\mathbf{T}} = \begin{bmatrix} {\tau^{[\beta_0]}}^2 & \\ \tau^{\beta{10}} & {\tau^{[\beta_1]}}^2 \\ \end{bmatrix}$$

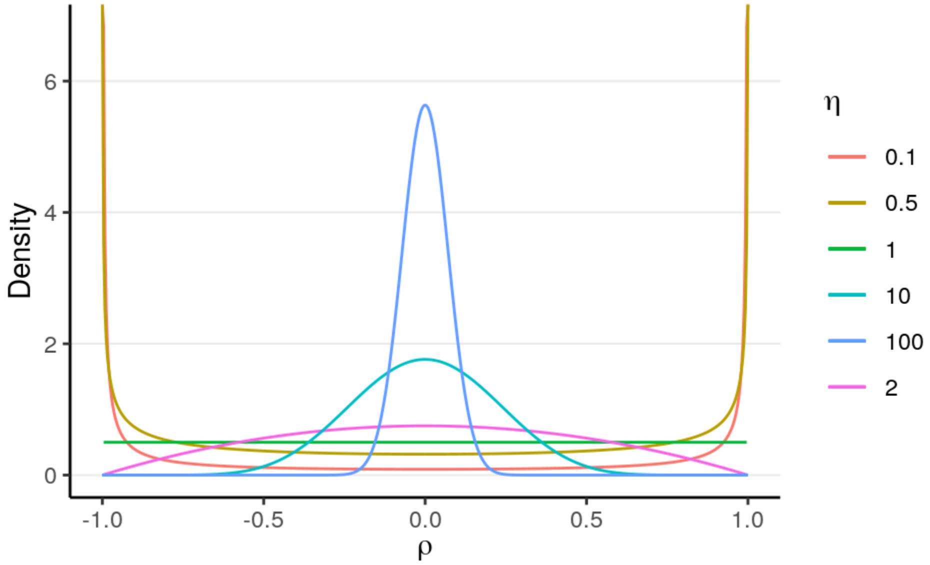
Note that $\mathcal{N}_2$ denotes a bivariate normal (i.e., 2-dimensional multivariate normal) distribution, because now we can talk about how β_0 and β_1 are associated at the person level. Generally I don't interpret the covariance between them because it largely depends on how the variables were centered, but nevertheless we should allow them to be correlated. The parameter \(\tau^{\beta{10}}\) thus denotes the covariance of them.

3.2.1 LKJ Prior

The LKJ Prior is a probability distribution for correlation matrices. A correlation matrix has 1 on all the diagonal elements. For example, a 2 \(\times\) 2 correlation matrix is \[\begin{bmatrix} 1 & \\ 0.35 & 1 \end{bmatrix}\] where the correlation is 0.35. Therefore, with two variables, there is one correlation; with three or more variables, the number of correlations will be \(q (q - 1) / 2\), where \(q\) is the number of variables.

For a correlation matrix of a given size, the LKJ prior has one shape parameter, \(\eta\), where \(\eta = 1\) corresponds to a uniform distribution of the correlations such that any correlations are equally likely, \(\eta \geq 1\) favors a matrix closer to an identity matrix so that the correlations are closer to zero, and \(\eta \leq 1\) favors a matrix with larger correlations. For a 2 \(\times\) 2 matrix, the distribution of the correlation, \(\rho\), with different \(\eta\) values are shown in the graph below:

```
># Warning: Removed 2 rows containing missing values (geom_path).
```



As you can see, when \(\eta\) increases, the correlation is more concentrated to zero.

The default in brms is to use \(\eta\) = 1, which is non-informative. If you have a weak but informative belief that the correlations shouldn't be very large, using \(\eta\) = 2 is reasonable.

The resulting model and priors are:

\[\begin{align} \text{Repeated-measure level:} \\ \texttt{Reaction10}_{ij} & \sim \mathcal{N}(\mu_{ij}, \sigma) \\ \mu_{ij} & = \beta_{0j} + \beta_{1j} \texttt{Days}_{ij} \\ \text{Person level:} \\ \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \end{bmatrix} & \sim \mathcal{N}_2\left(\begin{bmatrix} \mu^{[\beta_0]} \\ \mu^{[\beta_1]} \\ \end{bmatrix}, \boldsymbol{\mathbf{T}} \right) \\ \boldsymbol{\mathbf{T}} & = \operatorname{diag}(\boldsymbol{\mathbf{\tau}}) \boldsymbol{\mathbf{\Omega }}\operatorname{diag}(\boldsymbol{\mathbf{\tau}}) \\ \text{Priors:} \\ \mu^{[\beta_0]} & \sim \mathcal{N}(0, 50) \\ \mu^{[\beta_1]} & \sim \mathcal{N}(0, 10) \\ \tau^{[\beta_m]} & \sim \mathrm{Gamma}(2, 0.2), \; m = 0, 1 \\ \boldsymbol{\mathbf{\Omega }}& \sim \mathrm{LKJ}(1) \\ \sigma & \sim t^+(4, 0, 5) \end{align}\]

```r
m4 <- brm(Reaction10 ~ Days + (Days | Subject),
          data = sleepstudy,
          prior = c(# for intercept
            prior(normal(0, 50), class = "Intercept"),
            # for slope
            prior(normal(0, 10), class = "b"),
            # for tau_beta0 and tau_beta1
            prior(gamma(2, 0.2), class = "sd", group = "Subject"),
            # for correlation
            prior(lkj(1), class = "cor"),
            # for sigma
            prior(student_t(4, 0, 5), class = "sigma")),
          control = list(adapt_delta = .95),
          cores = 2L,
          seed = 2107)
```

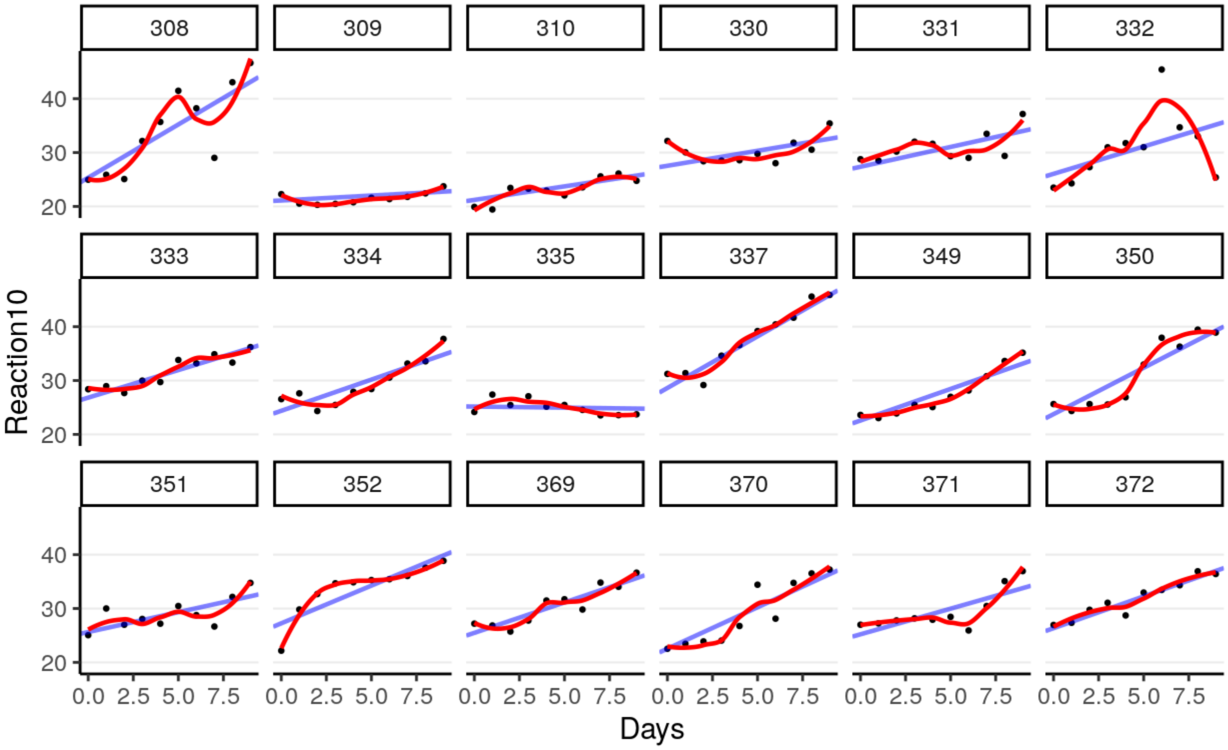
Below is a summary table of the results

term	estimate	std.error	lower	upper
b_Intercept	25.14	0.781	23.894	26.413
b_Days	1.04	0.183	0.751	1.339

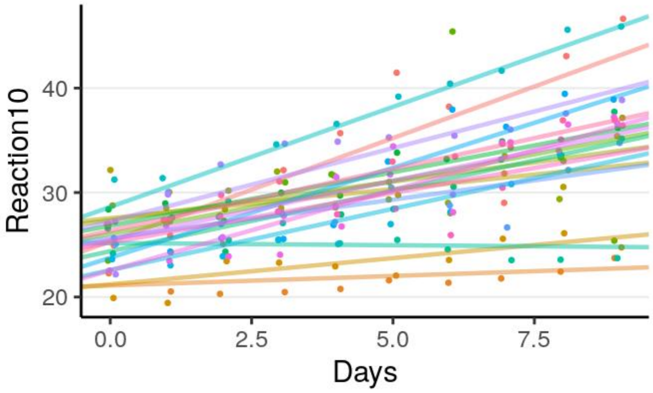
term	estimate	std.error	lower	upper
sd_Subject__Intercept	2.83	0.721	1.812	4.138
sd_Subject__Days	0.69	0.167	0.462	0.996
sigma	2.59	0.153	2.351	2.860

3.2.2 Fit of Individuals

```
># `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

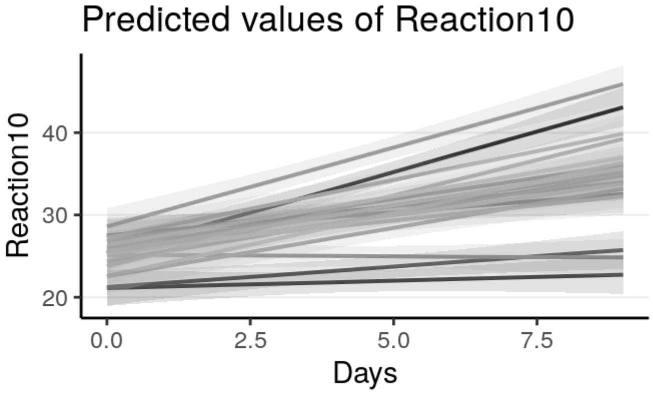


You can see that the fit is better. You can also visualize the varying regression lines:



Or using the sjPlot package:

```
># Note: uncertainty of error terms are not taken into account. You may want to use `rstantools::posterior_predict()`.
```



3.2.3 Fixed Effect Model

You can compare the previous model with one where have different slopes for different person, which can be modelled by including an interaction with the categorical Subject predictor. This is referred to as the *fixed-effect* model, as opposed to *random-effect* model used to describe hierarchical models with partial pooling. Below is an example:

```
m4_fixed <- brm(Reaction10 ~ Days * I(factor(Subject)),
        data = sleepstudy,
        prior = c(# for intercept
          prior(normal(0, 50), class = "Intercept"),
          # for slope
          prior(normal(0, 10), class = "b"),
          # for sigma
          prior(student_t(4, 0, 5), class = "sigma")),
        control = list(adapt_delta = .95),
      cores = 2L,
      seed = 2107)
```

3.2.4 Interpretations

Based on the model, at Day 0, the average reaction time across individuals was 251.417 ms, 95% CI [236.109, 266.954], and the *SD* at Day 0 was 28.278ms, 95% CI [16.561ms, 44.749ms].

The average growth rate per day in reaction time across individuals was 427 ms, 95% CI [6.789, 14.064], and the *SD* at Day 0 was 6.904ms, 95% CI [4.272ms, 10.769ms], as shown in the figure.

3.3 Varying σ

Finally, you can also allow σ to be different across individuals. This is typically used to relax the homogeneity of variance assumption, but recently there is also some interest in treating varying σ as an important outcome. Examples include fluctuations in mood, as two people with the same mean level of mood may fluctuate very differently, and mood swing can be an important outcome to assess. There has been some interesting applications in health research using ecological momentary assessment data. For an overview, see the paper by Hedeker, Mermelstein, and Demirtas (2008).

Without going into the details, here is the model and the priors:

$$\begin{align} \text{Repeated-measure level:} \\ \texttt{Reaction10}_{ij} & \sim \mathcal{N}(\mu_{ij}, \sigma_j) \\ \mu_{ij} & = \beta_{0j} + \beta_{1j} \texttt{Days}_{ij} \\ \text{Person level:} \\ \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \log(\sigma_j) \end{bmatrix} & \sim \mathcal{N}_2\left(\begin{bmatrix} \mu^{[\beta_0]} \\ \mu^{[\beta_1]} \\ \mu^{[s]} \end{bmatrix}, \boldsymbol{\mathbf{T}} \right) \\ \boldsymbol{\mathbf{T}} & = \operatorname{diag}(\boldsymbol{\mathbf{\tau}}) \boldsymbol{\mathbf{\Omega }}\operatorname{diag}(\boldsymbol{\mathbf{\tau}}) \\ \text{Priors:} \\ \mu^{[\beta_0]} & \sim \mathcal{N}(0, 50) \\ \mu^{[\beta_1]} & \sim \mathcal{N}(0, 10) \\ \mu^{[s]} & \sim t^+(4, 0, 1.6) \\ \tau^{[\beta_m]} & \sim \mathrm{Gamma}(2, 0.2), \; m = 0, 1 \\ \tau^{[s]} & \sim \mathrm{Gamma}(2, 0.625) \\ \boldsymbol{\mathbf{\Omega }}& \sim \mathrm{LKJ}(1) \end{align}$$

```
# Use |c| to estimate the covariance between the sigma and beta random effects
```

```r
m5 <- brm(bf(Reaction10 ~ Days + (Days |c| Subject),
             sigma ~ (1 |c| Subject)),
          data = sleepstudy,
          prior = c(# for intercept
            prior(normal(0, 50), class = "Intercept"),
            # for slope
            prior(normal(0, 10), class = "b"),
            # for tau_beta0
            prior(gamma(2, 0.2), class = "sd", coef = "Intercept",
                  group = "Subject"),
            # for tau_beta1
            prior(gamma(2, 0.2), class = "sd", coef = "Days",
                  group = "Subject"),
            # for correlation
            prior(lkj(1), class = "cor"),
            # for sigma
            prior(student_t(4, 0, 1.6), class = "Intercept", dpar = "sigma"),
            # for tau_sigma
            prior(gamma(2, 0.625), class = "sd", coef = "Intercept",
                  group = "Subject", dpar = "sigma")),
          control = list(adapt_delta = .95),
          cores = 2L,
          seed = 2107)
```

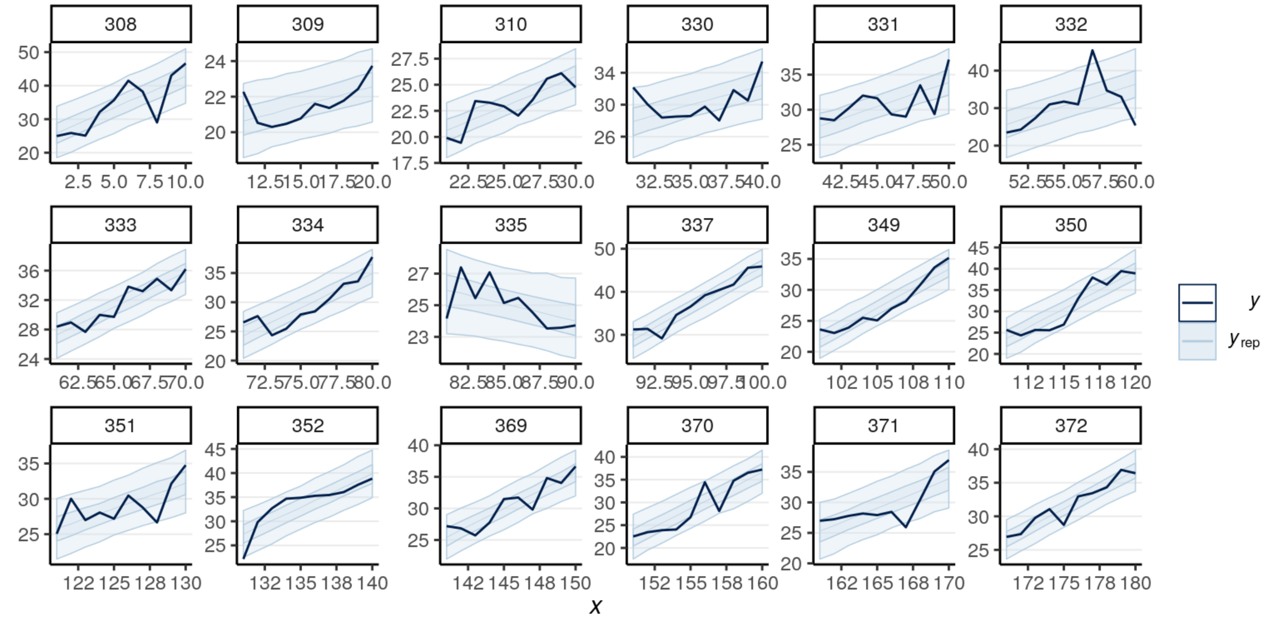
Below is a summary table of the results

Term	estimate	std.error	lower	upper

b_Intercept	25.156	0.839	23.773	26.553
b_sigma_Intercept	0.734	0.138	0.511	0.965
b_Days	1.042	0.179	0.747	1.332
sd_Subject__Intercept	3.184	0.722	2.187	4.499
sd_Subject__Days	0.708	0.159	0.493	1.019
sd_Subject__sigma_Intercept	0.512	0.124	0.340	0.739
cor_Subject__Intercept__Days	0.000	0.280	-0.465	0.463
cor_Subject__Intercept__sigma_Intercept	0.243	0.297	-0.278	0.706
cor_Subject__Days__sigma_Intercept	0.444	0.263	-0.044	0.810

And the posterior predictive check:

```
># Using all posterior samples for ppc type 'ribbon_grouped' by default.
```

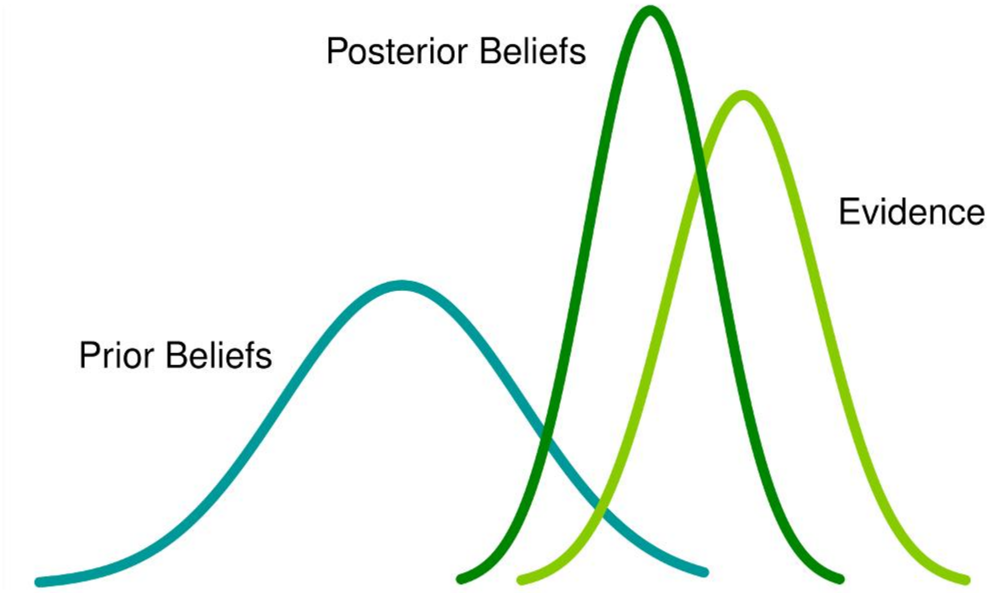


Topic: Bayesian Inference

Bayesian Statistics continues to remain incomprehensible in the ignited minds of many analysts. Being amazed by the incredible power of machine learning, a lot of us have become unfaithful to statistics. Our focus has narrowed down to exploring machine learning. Isn't it true?

We fail to understand that machine learning is not the only way to solve real world problems. In several situations, it does not help us solve business problems, even though there is data involved in these problems. To say the least, knowledge of statistics will allow you to work on complex analytical problems, irrespective of the size of data.

In 1770s, Thomas Bayes introduced 'Bayes Theorem'. Even after centuries later, the importance of 'Bayesian Statistics' hasn't faded away. In fact, today this topic is being taught in great depths in some of the world's leading universities.



Before we actually delve in Bayesian Statistics, let us spend a few minutes understanding *Frequentist Statistics*, the more popular version of statistics most of us come across and the inherent problems in that.

1. Frequentist Statistics

The debate between *frequentist* and *bayesian* have haunted beginners for centuries. Therefore, it is important to understand the difference between the two and how does there exists a thin line of demarcation!

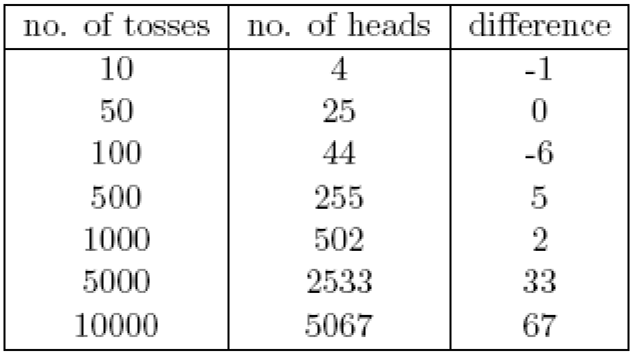
It is the most widely used inferential technique in the statistical world. Infact, generally it is the first school of thought that a person entering into the statistics world comes across.

Frequentist Statistics tests whether an event (hypothesis) occurs or not. It calculates the probability of an event in the long run of the experiment (i.e the experiment is repeated under the same conditions to obtain the outcome).

Here, the sampling distributions of **fixed size** are taken. Then, the experiment is theoretically repeated **infinite number of times** but practically done with a stopping intention. For example, I perform an experiment with a stopping intention in mind that I will stop the experiment when it is repeated 1000 times or I see minimum 300 heads in a coin toss.

Let's go deeper now.

Now, we'll understand *frequentist statistics* using an example of coin toss. The objective is to estimate the fairness of the coin. Below is a table representing the frequency of heads:



no. of tosses	no. of heads	difference
10	4	-1
50	25	0
100	44	-6
500	255	5
1000	502	2
5000	2533	33
10000	5067	67

We know that probability of getting a head on tossing a fair coin is 0.5. No. of heads represents the actual number of heads obtained. Difference is the difference between 0.5*(No. of tosses) - no. of heads.

An important thing is to note that, though the difference between the actual number of heads and expected number of heads(50% of number of tosses) increases as the number of tosses are increased, the proportion of number of heads to total number of tosses approaches 0.5 (for a fair coin).

This experiment presents us with a very common flaw found in frequentist approach i.e. *Dependence of the result of an experiment on the number of times the experiment is repeated*.

To know more about frequentist statistical methods, you can head to this excellent course on inferential statistics.

2. The Inherent Flaws in Frequentist Statistics

Till here, we've seen just one flaw in *frequentist statistics*. Well, it's just the beginning.

20th century saw a massive upsurge in the *frequentist statistics* being applied to numerical models to check whether one sample is different from the other, a parameter is important enough to be kept in the model and variousother manifestations of hypothesis testing. But *frequentist statistics* suffered some great flaws in its design and interpretation which posed a serious concern in all real life problems. For example:

1. p-values measured against a sample (fixed size) statistic with some stopping intention changes with change in intention and sample size. i.e If two persons work on the same data and have different stopping intention, they may get two different p- values for the same data, which is undesirable.

For example: Person A may choose to stop tossing a coin when the total count reaches 100 while B stops at 1000. For different sample sizes, we get different t-scores and different p-values. Similarly, intention to stop may change from fixed number of flips to total duration of flipping. In this case too, we are bound to get different *p-values*.

2- Confidence Interval (C.I) like p-value depends heavily on the sample size. This makes the stopping potential absolutely absurd since no matter how many persons perform the tests on the same data, the results should be consistent.

3- Confidence Intervals (C.I) are not probability distributions therefore they do not provide the most probable value for a parameter and the most probable values.

These three reasons are enough to get you going into thinking about the drawbacks of the *frequentist approach* and why is there a need for *bayesian approach*. Let's find it out.

From here, we'll first understand the basics of Bayesian Statistics.

3. Bayesian Statistics

"Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their beliefs in the evidence of new data."

You got that? Let me explain it with an example:

Suppose, out of all the 4 championship races (F1) between Niki Lauda and James hunt, Niki won 3 times while James managed only 1.

So, if you were to bet on the winner of next race, who would he be ?

I bet you would say Niki Lauda.

Here's the twist. What if you are told that it rained once when James won and once when Niki won and it is definite that it will rain on the next date. So, who would you bet your money on now ?

By intuition, it is easy to see that chances of winning for James have increased drastically. But the question is: how much ?

To understand the problem at hand, we need to become familiar with some concepts, first of which is conditional probability (explained below).

In addition, there are certain pre-requisites:

Pre-Requisites:

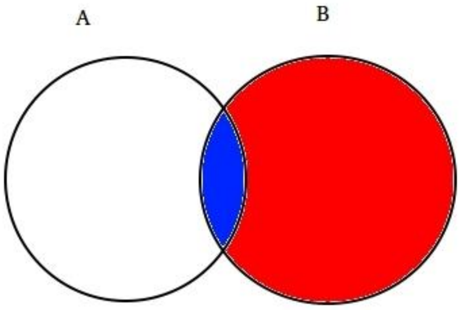
1. Linear Algebra : To refresh your basics, you can check out Khan's Academy Algebra.
2. Probability and Basic Statistics : To refresh your basics, you can check out another course by Khan Academy.

3.1 Conditional Probability

It is defined as the: Probability of an event A given B equals the probability of B and A happening together divided by the probability of B."

For example: Assume two partially intersecting sets A and B as shown below.

Set A represents one set of events and Set B represents another. We wish to calculate the probability of A given B has already happened. Lets represent the happening of event B by shading it with red.



Now since B has happened, the part which now matters for A is the part shaded in blue which is interestingly $A \cap B$. So, the probability of A given B turns out to be:

$$\frac{BlueArea}{RedArea + BlueArea}$$

Therefore, we can write the formula for event B given A has already occurred by:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

or

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Now, the second equation can be rewritten as :

$$P(A|B) = \frac{P(B|A) X P(A)}{P(B)}$$

This is known as **Conditional Probability**.

Let’s try to answer a betting problem with this technique.

Suppose, B be the *event of winning of James Hunt*. A be the *event of raining*. Therefore,

1. P(A) =1/2, since it rained twice out of four days.
2. P(B) is 1/4, since James won only one race out of four.
3. P(A|B)=1, since it rained every time when James won.

Substituting the values in the conditional probability formula, we get the probability to be around 50%, which is almost the double of 25% when rain was not taken into account (Solve it at your end).

This further strengthened our belief of James winning in the light of new *evidence* i.e rain. You must be wondering that this formula bears close resemblance to something you might have heard a lot about. Think!

Probably, you guessed it right. It looks like **Bayes Theorem**.

4. Bayesian Inference

There is no point in diving into the theoretical aspect of it. So, we’ll learn how it works! Let’s take an example of coin tossing to understand the idea behind *bayesian inference*.

An important part of *bayesian inference* is the establishment of *parameters* and *models*.

Models are the mathematical formulation of the observed events. Parameters are the factors in the models affecting the observed data. For example, in tossing a coin, **fairness of coin** may be defined as the parameter of coin denoted by θ. The outcome of the events may be denoted by D.

Answer this now. What is the probability of 4 heads out of 9 tosses(D) given the fairness of coin (θ). i.e $P(D|\theta)$

Wait, did I ask the right question? No.

We should be more interested in knowing : Given an outcome (D) what is the probbaility of coin being fair (θ=0.5)

Lets represent it using Bayes Theorem:

$$P(\theta|D)=(P(D|\theta)\ X\ P(\theta))/P(D)$$

Here, $P(\theta)$ is the ***prior*** i.e the strength of our belief in the fairness of coin before the toss. It is perfectly okay to believe that coin can have any degree of fairness between 0 and 1.

$P(D|\theta)$ is the likelihood of observing our result given our distribution for θ. If we knew that coin was fair, this gives the probability of observing the number of heads in a particular number of flips.

$P(D)$ is the evidence. This is the probability of data as determined by summing (or integrating) across all possible values of θ, weighted by how strongly we believe in those particular values of θ.

If we had multiple views of what the fairness of the coin is (but didn't know for sure), then this tells us the probability of seeing a certain sequence of flips for all possibilities of our belief in the coin's fairness.

$P(\theta|D)$ is the posterior belief of our parameters after observing the evidence i.e the number of heads .

From here, we’ll dive deeper into mathematical implications of this concept. Don’t worry. Once you understand them, getting to its *mathematics* is pretty easy.

To define our model correctly , we need two mathematical models before hand. One to represent the ***likelihood function P(D|θ)*** and the other for representing the distribution of ***prior beliefs .*** The product of these two gives the ***posterior belief P(θ|D)*** distribution.

Since prior and posterior are both beliefs about the distribution of fairness of coin, intuition tells us that both should have the same mathematical form. Keep this in mind. We will come back to it again.