Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.52201.03

Lecture 25: Mixed Models

Jason Mezey

April 25, 2024 (Th) 8:40-9:55

Summary of lecture 25: Mixed Models

- Last lecture, we completed our discussion of logistic regression
- Today, we will (briefly) introduce Mixed Models!
- We will also begin our (very brief) introduction to Bayesian Statistics!

(Brief) introduction to mixed models I

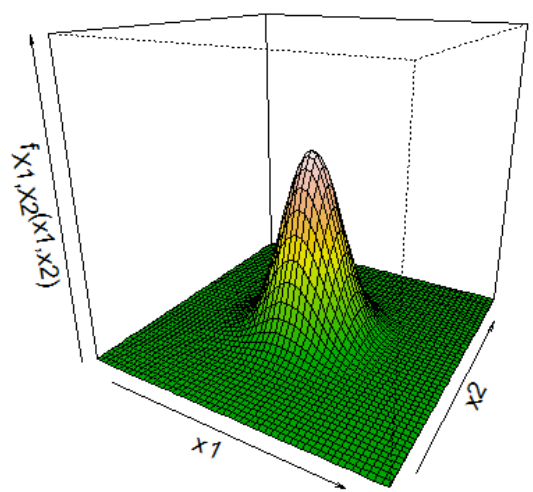
- A *mixed model* describes a class of models that have played an important role in early quantitative genetic (and other types) of statistical analysis before genomics (if you are interested, look up variance component estimation)
- These models are now used extensively in GWAS analysis as a tool for model covariates (often population structure!)
- These models considered effects as either "fixed" (they types of regression coefficients we have discussed in the class) and "random" (which just indicates a different model assumption) where the appropriateness of modeling covariates as fixed or random depends on the context (fuzzy rules!) - you will generally not have to deal with these issues in GWAS

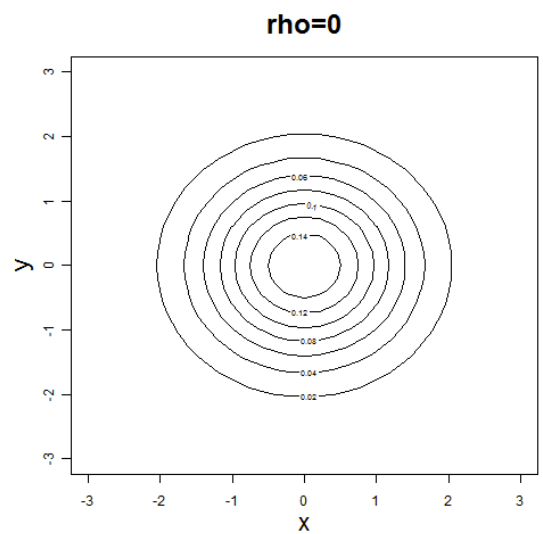
Introduction to mixed models II

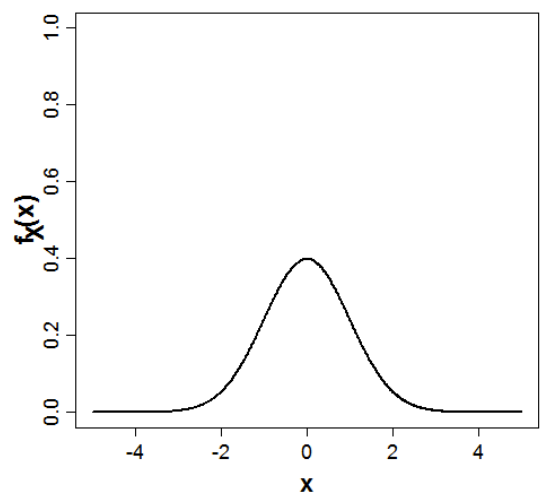
- Recall that for a linear regression of sample size n, we model the distributions of n total yi phenotypes using a linear regression model with normal error:

$$y_i = \beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

- A reminder about how to think about / visualize multivariate (bivariate) normal distributions and marginal normal distributions:







- We can therefore consider the entire sample of yi and their associated error in an equivalent multivariate setting:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon \qquad \epsilon \sim multiN(\mathbf{0}, \mathbf{I}\sigma_\epsilon^2)$$

Introduction to mixed models III

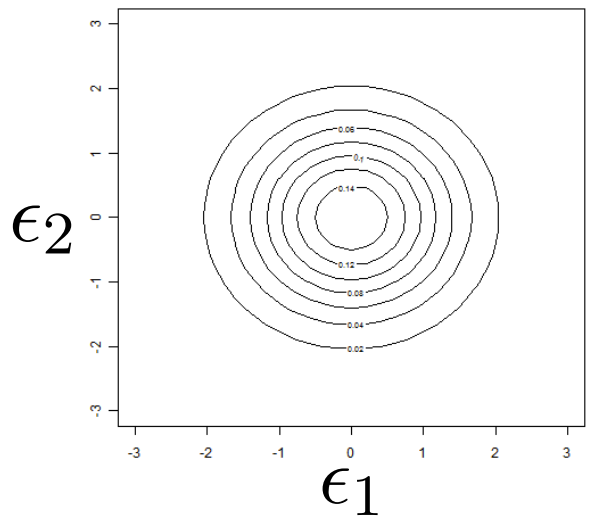
- Recall our linear regression model has the following structure:

$$y_i = \beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2_\epsilon)$$

- For example, for *n*=2:

$$y_1 = \beta_\mu + X_{1,a}\beta_a + X_{1,d}\beta_d + \epsilon_1$$

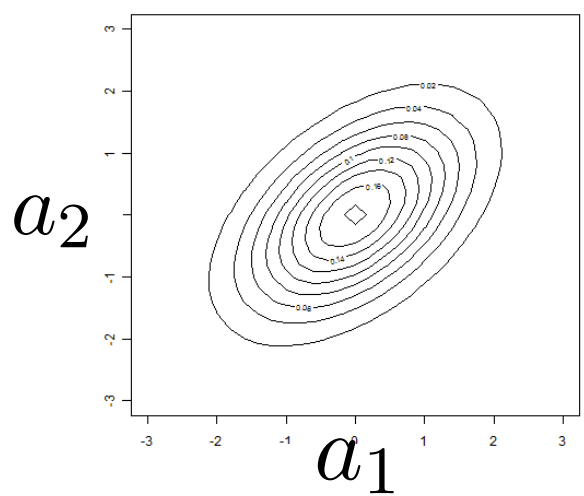
$$y_2 = \beta_\mu + X_{2,a}\beta_a + X_{2,d}\beta_d + \epsilon_2$$



- What if we introduced a correlation?

$$y_1 = \beta_\mu + X_{1,a}\beta_a + X_{1,d}\beta_d + a_1$$

$$y_2 = \beta_\mu + X_{2,a}\beta_a + X_{2,d}\beta_d + a_2$$



Introduction to mixed models IV

- The formal structure of a mixed model is as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{a} + \epsilon$$

$$\epsilon \sim multiN(\mathbf{0}, \mathbf{I}\sigma^2_\epsilon) \qquad \mathbf{a} \sim multiN(\mathbf{0}, \mathbf{A}\sigma^2_\mathbf{a})$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{i,a} & X_{i,d} \\ 1 & X_{i,a} & X_{i,d} \\ 1 & X_{i,a} & X_{i,d} \\ \vdots & \vdots & \vdots \\ 1 & X_{i,a} & X_{i,d} \end{bmatrix} \begin{bmatrix} \beta_\mu \\ \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \ldots & \ldots & \ldots & 1 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \mathbf{a}_3 \\ \vdots \\ \mathbf{a}_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Note that **X** is called the "design" matrix (as with a GLM), **Z** is called the "incidence" matrix, the **a** is the vector of random effects and note that the **A** matrix determines the correlation among the ai values where the structure of **A** is provided from external information (!!)

Introduction to mixed models V

- The matrix **A** is an *n*x*n* covariance matrix (what is the form of a covariance matrix?)
- Where does **A** come from? This depends on the modeling application...
- In GWAS, the random effect is usually used to account for population structure OR relatedness among individuals
  - For population structure, a matrix is constructed from the covariance (or similarity) among individuals based on their genotypes
  - For relatedness, we use estimates of identity by descent, which can be estimated from a pedigree or genotype data

Introduction to mixed models VI

- We perform inference (estimation and hypothesis testing) for the mixed model just as we would for a linear regression (!!)
- Note that in some applications, people might be interested in estimating the variance components $\sigma^2_\epsilon, \sigma^2_a$ but for GWAS, we are generally interested in regression parameters for our genotype (as before!): β_a, β_d
- For a GWAS, we will therefore determine the MLE of the genotype association parameters and use a LRT for the hypothesis test, where we will compare a null and alternative model (what is the difference between these models?)

Mixed models: inference I

- To estimate parameters, we will use the MLE, so we are concerned with the form of the likelihood equation

$$L(\beta, \sigma_a^2, \sigma_\epsilon^2|\mathbf{y}) = \int_{-\infty}^{\infty} Pr(\mathbf{y}|\beta, \mathbf{a}, \sigma_\epsilon^2) Pr(\mathbf{a}|\mathbf{A}\sigma_a^2) d\mathbf{a}$$

$$L(\beta, \sigma_a^2, \sigma_\epsilon^2|\mathbf{y}) = |\mathbf{I}\sigma_\epsilon^2|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma_\epsilon^2}[\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{a}]^{\mathrm{T}}[\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{a}]} |\mathbf{A}\sigma_a^2|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma_a^2}\mathbf{a}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{a}}$$

$$l(\beta, \sigma_a^2, \sigma_\epsilon^2|\mathbf{y}) \propto -\frac{n}{2} ln\sigma_\epsilon^2 - -\frac{n}{2} ln\sigma_a^2 - \frac{1}{2\sigma_\epsilon^2} [\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{a}]^{\mathrm{T}} [\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{a}] - \frac{1}{2\sigma_a^2}\mathbf{a}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{a}$$

- Unfortunately, there is no closed form for the MLE since they have the following form:

$$MLE(\hat{\beta}) = (\mathbf{X}\hat{\mathbf{V}}^{-1}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\mathbf{Y}$$

$$MLE(\hat{\mathbf{V}}) = f(\mathbf{X}, \hat{\mathbf{V}}, \mathbf{Y}, \mathbf{A})$$

$$\mathbf{V} = \sigma_a^2\mathbf{A} + \sigma_\epsilon^2\mathbf{I}$$

Mixed models: inference II

- We therefore need an algorithm to find the MLE for the mixed model
- We will discuss the use of an EM (Expectation-Maximization) algorithm for this purpose, which is an algorithm with good theoretical and practical properties, e.g. hill-climbing algorithm, guaranteed to converge to a (local) maximum, it is a stable algorithm, etc.
- We do not have time to introduce these properties in detail so we will just show the steps / equations you need to implement this algorithm (such that you can implement it yourself = see computer lab this week!)

Algorithm Basics

- **algorithm** - a sequence of instructions for taking an input and producing an output
- We often use algorithms in estimation of parameters where the structure of the estimation equation (e.g., the log-likelihood) is so complicated that we cannot
  - Derive a simple (closed) form equation for the estimator
  - Cannot easily determine the value the estimator should take by other means (e.g., by graphical visualization)
- We will use algorithms to "search" for the parameter values that correspond to the estimator of interest
- In general: algorithms are not guaranteed to produce the correct value of the estimator (!!), because the algorithm may "converge" (=return) the wrong answer (e.g., converges to a "local" maximum or does not converge!) and because the compute time to converge to exactly the same answer is impractical for applications

Mixed models: EM algorithm

1. At step $[t]$ for $t = 0$, assign values to the parameters: $\beta^{[0]} = \left[\beta_{\mu}^{[0]}, \beta_{a}^{[0]}, \beta_{d}^{[0]}\right], \sigma_{a}^{2,[0]}, \sigma_{\epsilon}^{2,[0]}$. These need to be selected such that they are possible values of the parameters (e.g. no negative values for the variance parameters).

2. Calculate the expectation step for $[t]$:

$$\mathbf{a}^{[t]} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_{\epsilon}^{2,[t-1]}}{\sigma_{a}^{2,[t-1]}}\right)^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{y} - \mathbf{x}\beta^{[t-1]})$$

$$V_{\mathbf{a}}^{[t]} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_{\epsilon}^{2,[t-1]}}{\sigma_{a}^{2,[t-1]}}\right)^{-1}\sigma_{\epsilon}^{2,[t-1]}$$

3. Calculate the maximization step for $[t]$:

$$\beta^{[t]} = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}(\mathbf{y} - \mathbf{Z}\mathbf{a}^{[t]})$$

$$\sigma_{a}^{2,[t]} = \frac{1}{n}\left[\mathbf{a}^{[t]}\mathbf{A}^{-1}\mathbf{a}^{[t]} + tr(\mathbf{A}^{-1}V_{\mathbf{a}}^{[t]})\right]$$

$$\sigma_{\epsilon}^{2,[t]} = -\frac{1}{n}\left[\mathbf{y} - \mathbf{x}\beta^{[t]} - \mathbf{Z}\mathbf{a}^{[t]}\right]^{\mathrm{T}}\left[\mathbf{y} - \mathbf{x}\beta^{[t]} - \mathbf{Z}\mathbf{a}^{[t]}\right] + tr(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}V_{\mathbf{a}}^{[t]})$$

   where tr is a trace function, which is equal to the sum of the diagonal elements of a matrix.

4. Iterate steps 2, 3 until $(\beta^{[t]}, \sigma_{a}^{2,[t]}, \sigma_{\epsilon}^{2,[t]}) \approx (\beta^{[t+1]}, \sigma_{a}^{2,[t+1]}, \sigma_{\epsilon}^{2,[t+1]})$ (or alternatively $lnL^{[t]} \approx lnL^{[t+1]}$).

Mixed Model hypothesis testing I

- Recall that our null and alternative hypotheses are:

$$H_0 : \beta_a = 0 \cap \beta_d = 0$$

$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$

- We will use the LRT for the null (0) and alternative (1):

$$LRT = -2ln\Lambda = -2ln\frac{L(\hat{\theta}_0|\mathbf{y})}{L(\hat{\theta}_1|\mathbf{y})} \qquad LRT = -2ln\Lambda = 2l(\hat{\theta}_1|\mathbf{y}) - 2l(\hat{\theta}_0|\mathbf{y})$$

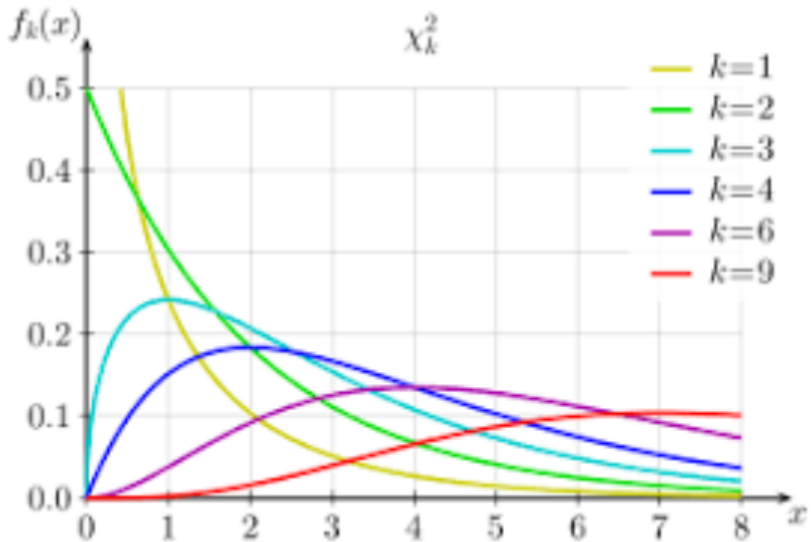
- To do this, run the EM algorithm twice, once for the null hypothesis (again what is this?) and once for the alternative (i.e. all parameters unrestricted) and then substitute the parameter values into the log-likelihood equations and calculate the LRT

Mixed Model p-value

- To calculate our p-value, we need to know the distribution of our LRT statistic under the null hypothesis
- There is no simple form for this distribution for any given n (contrast with F-statistics!!) but we know that as n goes to infinite, we know the distribution is i.e. ($n \rightarrow \infty$):

$$LRT = -2lnL = 2l(\hat{\theta}_1|\mathbf{y}) - 2l(\hat{\theta}_0|\mathbf{y})$$

$$LRT \rightarrow \chi^2_{df}$$



Mixed models: inference V

- In general, a mixed model is an advanced methodology for GWAS analysis but is proving to be an extremely useful technique for covariate modeling
- There is software for performing a mixed model analysis (e.g. R-packages, EMMAX, GEMMA, etc.)
- Mastering mixed models will take more time than we have to devote to the subject in this class, but what we have covered provides a foundation for understanding the topic

Construction of **A** matrix I

- The matrix **A** is an *n*x*n* covariance matrix (what is the form of a covariance matrix?)
- Where does **A** come from? This depends on the modeling application...
- In GWAS, the random effect is usually used to account for population structure OR relatedness among individuals
  - For relatedness, we use estimates of identity by descent, which can be estimated from a pedigree or genotype data
  - For population structure, a matrix is constructed from the covariance (or similarity) among individuals based on their genotypes

Construction of **A** matrix II

$$Data = \begin{bmatrix} z_{11} & \cdots & z_{1k} & y_{11} & \cdots & y_{1m} & x_{11} & \cdots & x_{1N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{n1} & \cdots & z_{nk} & y_{n1} & \cdots & y_{nm} & x_{11} & \cdots & x_{nN} \end{bmatrix}$$

- Calculate the *n*xn (*n*=sample size) covariance matrix for the individuals in your sample across all genotypes - this is a reasonable **A** matrix!
- There is software for calculating **A** and for performing a mixed model analysis (e.g. EMMAX, FAST-LMM, etc.)
- Mastering mixed models will take more time than we have to devote to the subject in this class, but what we have covered provides a foundation for understanding the topic

Introduction to Bayesian analysis 1

- Up to this point, we have considered statistical analysis (and inference) using a Frequentist formalism
- There is an alternative formalism called Bayesian that we will now introduce in a very brief manner
- Note that there is an important conceptual split between statisticians who consider themselves Frequentist of Bayesian but for GWAS analysis (and for most applications where we are concerned with analyzing data) we do not have a preference, i.e. we only care about getting the "right" biological answer so any (or both) frameworks that get us to this goal are useful
- In GWAS (and mapping) analysis, you will see both frequentist (i.e. the framework we have built up to this point!) and Bayesian approaches applied

Introduction to Bayesian analysis II

- In both frequentist and Bayesian analyses, we have the same probabilistic framework (sample spaces, random variables, probability models, etc.) and when assuming our probability model falls in a family of parameterized distributions, we assume that a single fixed parameter value(s) describes the true model that produced our sample
- However, in a Bayesian framework, we now allow the parameter to have it's own probability distribution (we DO NOT do this in a frequentist analysis), such that we treat it as a random variable
- This may seem strange - how can we consider a parameter to have a probability distribution if it is fixed?
- However, we can if we have some prior assumptions about what values the parameter value will take for our system compared to others and we can make this prior assumption rigorous by assuming there is a probability distribution associated with the parameter
- It turns out, this assumption produces major differences between the two analysis procedures (in how they consider probability, how they perform inference, etc.

Introduction to Bayesian analysis III

- To introduce Bayesian statistics, we need to begin by introducing Bayes theorem
- Consider a set of events (remember events!?) $\mathcal{A} = \mathcal{A}_1 ... \mathcal{A}_k$ of a sample space Ω (where k may be infinite), which form a partition of the sample space, i.e. $\bigcup_i^k \mathcal{A}_i = \Omega \text{ and } \mathcal{A}_i \cap \mathcal{A}_j = \emptyset \text{ for all } i \neq j$
- For another event $\mathcal{B} \subset \Omega$ (which may be Ω itself) define the Law of total probability:

$$Pr(\mathcal{B}) = \sum_{i=1}^{k} Pr(\mathcal{B} \cap \mathcal{A}_i) = \sum_{i=1}^{k} Pr(\mathcal{B}|\mathcal{A}_i) Pr(\mathcal{A}_i)$$

- Now we can state Bayes theorem:

$$Pr(\mathcal{A}_i|\mathcal{B}) = \frac{Pr(\mathcal{A}_i \cap \mathcal{B})}{Pr(\mathcal{B})} = \frac{Pr(\mathcal{B}|\mathcal{A}_i) Pr(\mathcal{A}_i)}{Pr(\mathcal{B})} = \frac{Pr(\mathcal{B}|\mathcal{A}_i) Pr(\mathcal{A})}{\sum_{i=1}^{k} Pr(\mathcal{B}|\mathcal{A}_i) Pr(\mathcal{A}_i)}$$

Introduction to Bayesian analysis IV

- Remember that in a Bayesian (not frequentist!) framework, our parameter(s) have a probability distribution associated with them that reflects our belief in the values that might be the true value of the parameter

- Since we are treating the parameter as a random variable, we can consider the joint distribution of the parameter AND a sample **Y** produced under a probability model:

$$Pr(\theta \cap \mathbf{Y})$$

- Fo inference, we are interested in the probability the parameter takes a certain value given a sample:

$$Pr(\theta|\mathbf{y})$$

- Using Bayes theorem, we can write:

$$Pr(\theta|\mathbf{y}) = \frac{Pr(\mathbf{y}|\theta)Pr(\theta)}{Pr(\mathbf{y})}$$

- Also note that since the sample is fixed (i.e. we are considering a single sample) $Pr(\mathbf{y}) = c,$ we can rewrite this as follows:

$$Pr(\theta|\mathbf{y}) \propto Pr(\mathbf{y}|\theta)Pr(\theta)$$

Introduction to Bayesian analysis V

- Let's consider the structure of our main equation in Bayesian statistics:

$$Pr(\theta|\mathbf{y}) \propto Pr(\mathbf{y}|\theta)Pr(\theta)$$

- Note that the left hand side is called the posterior probability:

$$Pr(\theta|\mathbf{y})$$

- The first term of the right hand side is something we have seen before, i.e. the likelihood (!!):

$$Pr(\mathbf{y}|\theta) = L(\theta|\mathbf{y})$$

- The second term of the right hand side is new and is called the prior:

$$Pr(\theta)$$

- Note that the prior is how we incorporate our assumptions concerning the values the true parameter value may take

- In a Bayesian framework, we are making two assumptions (unlike a frequentist where we make one assumption): 1. the probability distribution that generated the sample, 2. the probability distribution of the parameter

That's it for today

- Next OPTIONAL lectures: more Bayesian Statistics (!!)