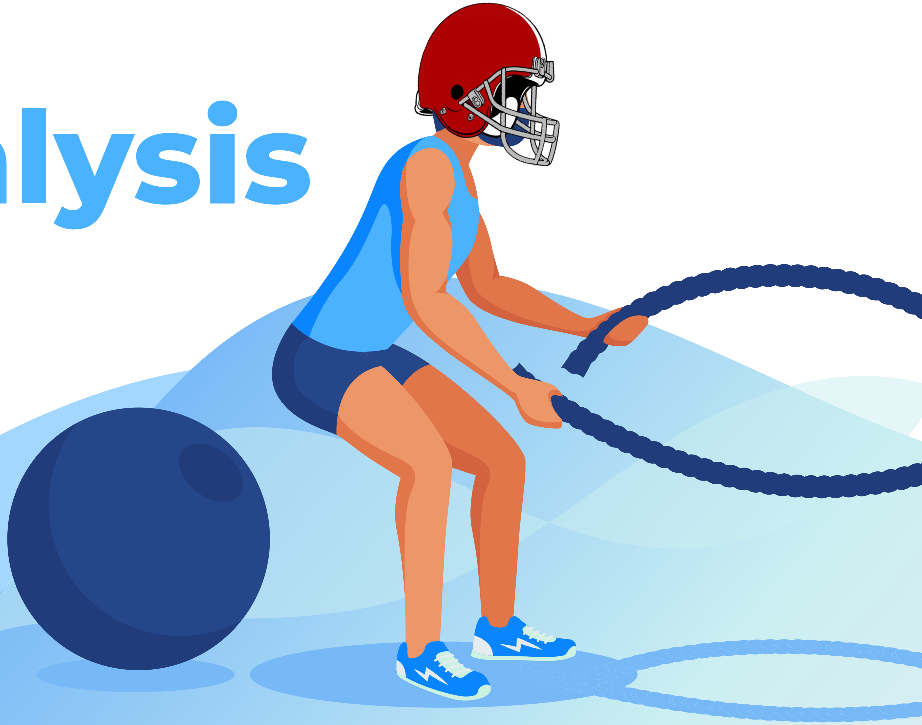
NFL Injuries Bayesian Analysis

Brandon Owens



Background

- **National Football League**
  - **$163B American Football organization**
  - **Schedule**
    - **Preseason: 3 games**
    - **Season: 17 games**
    - **Postseason: 0-4 games**

The Problem: Injuries

MOST GUARANTEED MONEY
Current Contracts

Deshaun Watson	$230M
Kyler Murray	$189.5M
Aaron Rodgers	$150.7M
Josh Allen	$150M
Patrick Mahomes	$141.5M
Matthew Stafford	$130M
Dak Prescott	$126M

Athlete Wellbeing

- Physical/mental health
- Ability to function/live

Organizational Cost

- “Guaranteed” contracts
- Rehabilitation process

Fan Interest

- Pay to see their favorite players
- Devalues the product

The NFL has implemented over 50 rule changes intended to reduce player danger since 2002 alone.

Rule Changes: Helping?

"I don't like [the new rules] at all. I think the hip drop is pointless."

Fans often claim controversy over these changes: **are they effective?**

Q1: Have Injury Rates Improved?

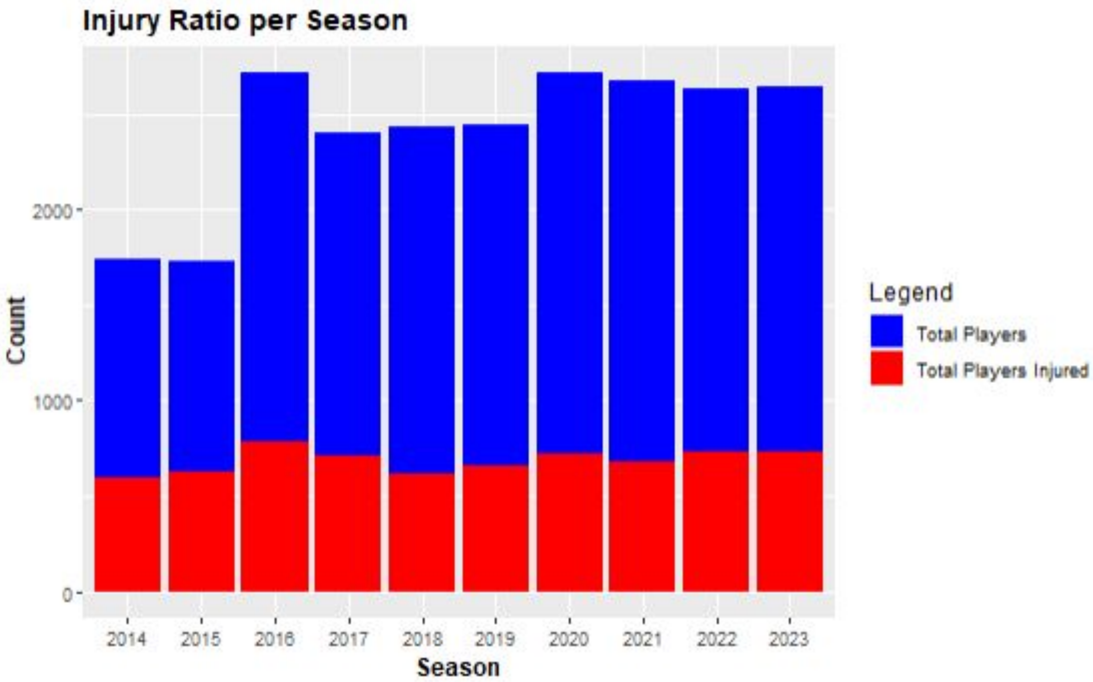
Q1: The Data

cleaned*

	year	n_players	n_inj	pct_inj
1	2009	1715	581	0.3387755
2	2010	1716	597	0.3479021
3	2011	1715	617	0.3597668
4	2012	1715	565	0.3294461
5	2013	1734	556	0.3206459
6	2014	1736	596	0.3433180
7	2015	1730	624	0.3606936
8	2016	2714	782	0.2881356
9	2017	2401	714	0.2973761
10	2018	2430	613	0.2522634
11	2019	2440	661	0.2709016
12	2020	2716	719	0.2647275
13	2021	2677	682	0.2547628
14	2022	2631	734	0.2789814
15	2023	2645	735	0.2778828

*status: ACT, RSN, DEV, PUP, INA
*report_status: Out, DBT, QST

Q1: EDA



Injury Ratio per Season
Count
2000
1000
0
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Season
Legend
Total Players
Total Players Injured

Q1: The Model (Beta Conjugate)

- Prior

$$\text{Beta}(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

$$\begin{aligned} p(\theta|y,n,M) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \end{aligned}$$

after normalization

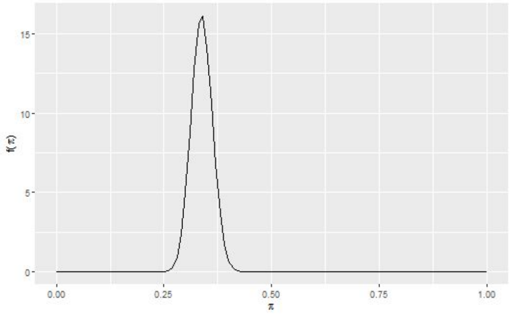
$$p(\theta|y,n,M) = \text{Beta}(\theta|\alpha+y,\beta+n-y)$$

Q1: The Model (Hyperparameters)

- Empirical Hyperparameters

3.1.2 Tuning the Beta prior

With a sense for how the Beta(α, β) model works, let's *tune* the shape hyperparameters α and β to reflect our prior information about Michelle's election support π. We saw in Figure 3.1 (left) that across 30 previous polls, Michelle's average support was around 45 percentage points, though she roughly polled as low as 25 and as high as 65 percentage points. Our Beta(α, β) prior should have similar patterns. For example, we want to pick α and β for which π tends to be around 0.45, $E(\pi) = \alpha/(\alpha+\beta) \approx 0.45$. Or, after some rearranging, [Bayes Rules Book]



Prior based on 2009-2013 years (~30 rule changes for players safety since):
E(pi) = alpha/(alpha+beta) ~ 0.30 - 0.36

Thus: alpha = 125; beta = 245

Q1: Sampling (Gibbs)

```
```{r hier}
iter <- 10000
n = df$number_of_players
x <- array(NA, c(iter, 10))
theta <- array(NA, c(iter, 10))

alpha <- 125
beta <-245

x[1, ]=df$distinct_injuries
theta[1, ]=df$prob_inj

for(i in 2:iter)
{
  x[i, ] <- rbinom(10, size = n, prob = theta[i - 1])
  theta[i, ] = rbeta(10, alpha+(x[i]), beta+n-(x[i]))
}

```
```

Q1: Gibbs Diagnostics

Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 12 14030 3746 3.75	Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 10 10700 3746 2.86	Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 8 10956 3746 2.92	Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 10 11344 3746 3.03	Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 12 17655 3746 4.71
Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 15 18006 3746 4.81	Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 10 11022 3746 2.94	Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 15 16932 3746 4.52	Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 10 10584 3746 2.83	Quantile (q) = 0.025 Accuracy (r) = +/- 0.005 Probability (s) = 0.95 Burn-in Total Lower bound Dependence (M) (N) (Nmin) factor (I) 15 14979 3746 4

- 20,000 iterations
- Burn-in: 20
- Thin: 10

Raftery-Lewis
Heidelberg-Welch: Passed

Q1: Sampling (Gibbs… again)

```{r sampled}
iter <- 20000
n = df$number_of_players
x <- array(NA, c(iter, 10))
theta <- array(NA, c(iter, 10))

alpha <- 125
beta <- 245

x[1, ]=df$distinct_injuries
theta[1, ]=df$prob_inj

for(i in 2:iter)
{
  x[i, ] <- rbinom(10, size = n, prob = theta[i - 1])
  theta[i, ] = rbeta(10, alpha+(x[i]), beta+n-(x[i]))
}

x <- as.data.frame(x)
theta <- as.data.frame(theta)

binded <- cbind(x, theta)

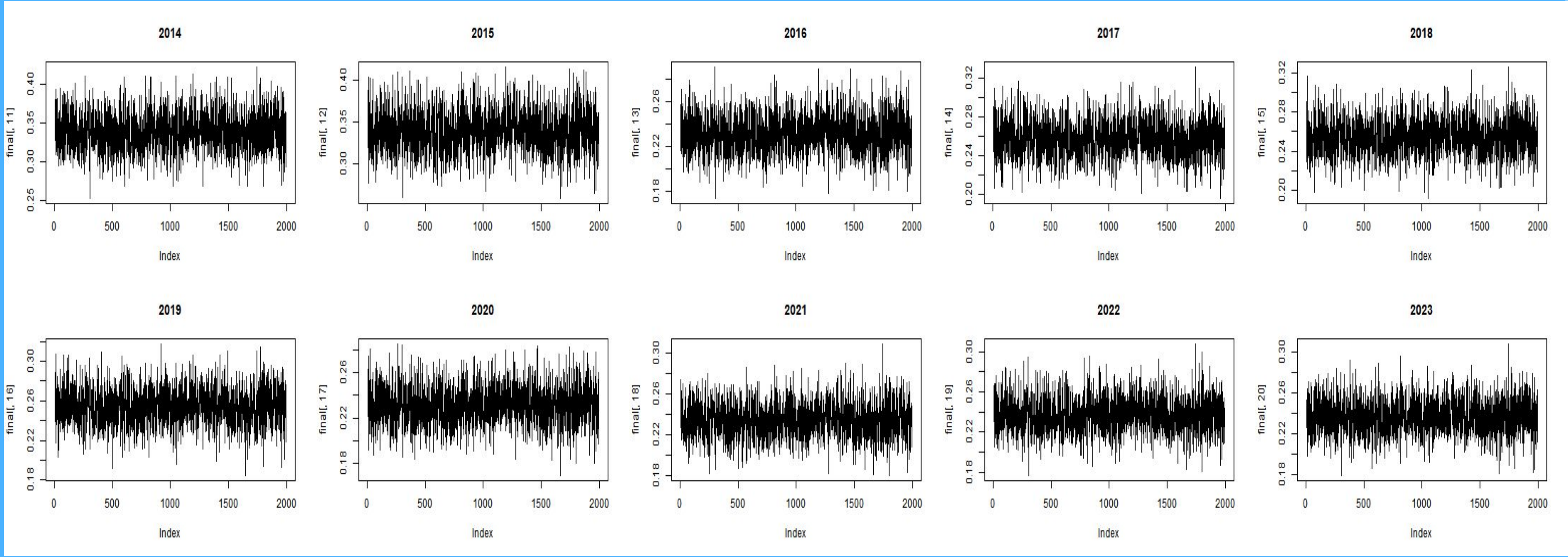
burn <- 20
b <- burn + 1
k=10
N = round((iter-b)/k)

burned_df <- binded[b:iter, ]
final <- burned_df[k*(1:N), ]

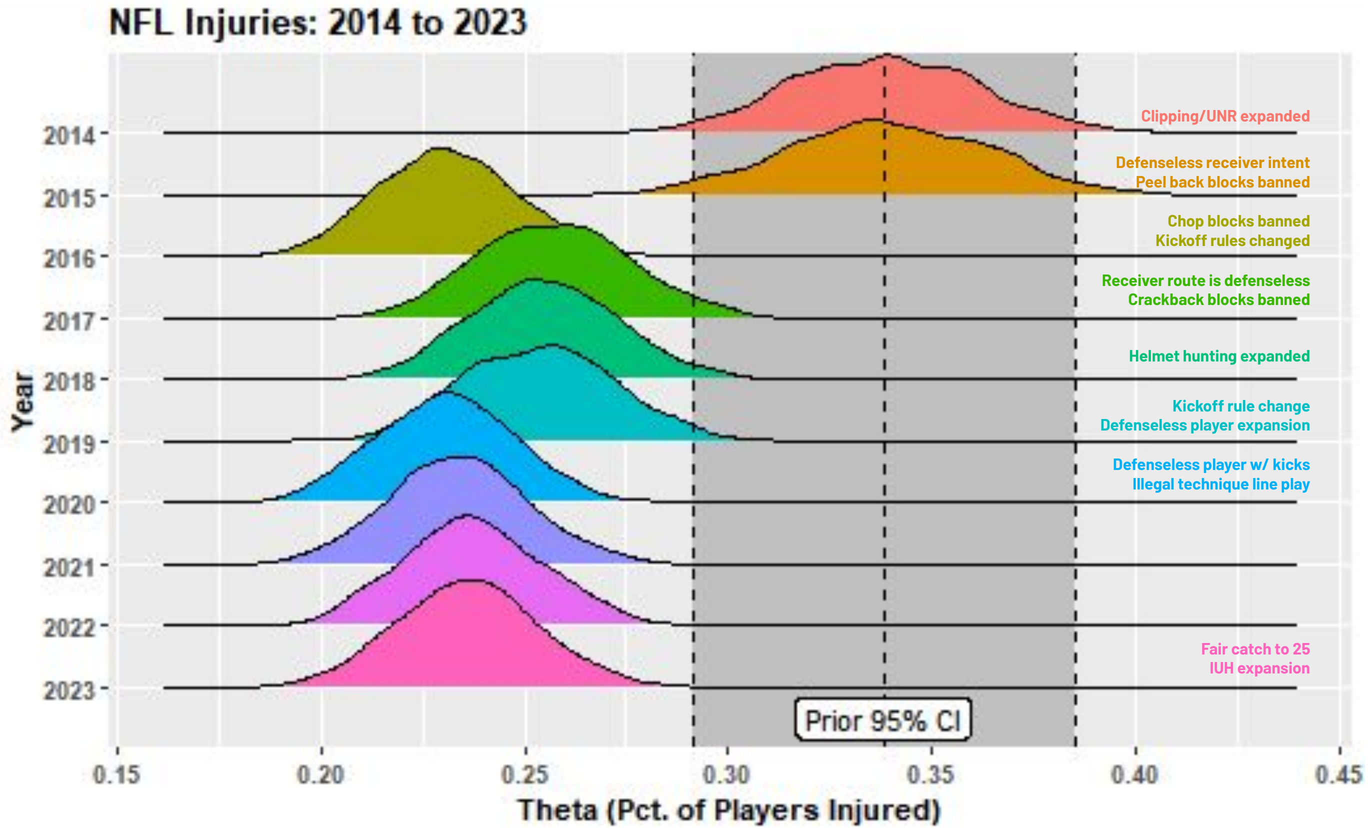
```

- 20,000 iterations
- Burn-in: 20
- Thin: 10

Q1: Gibbs Convergence

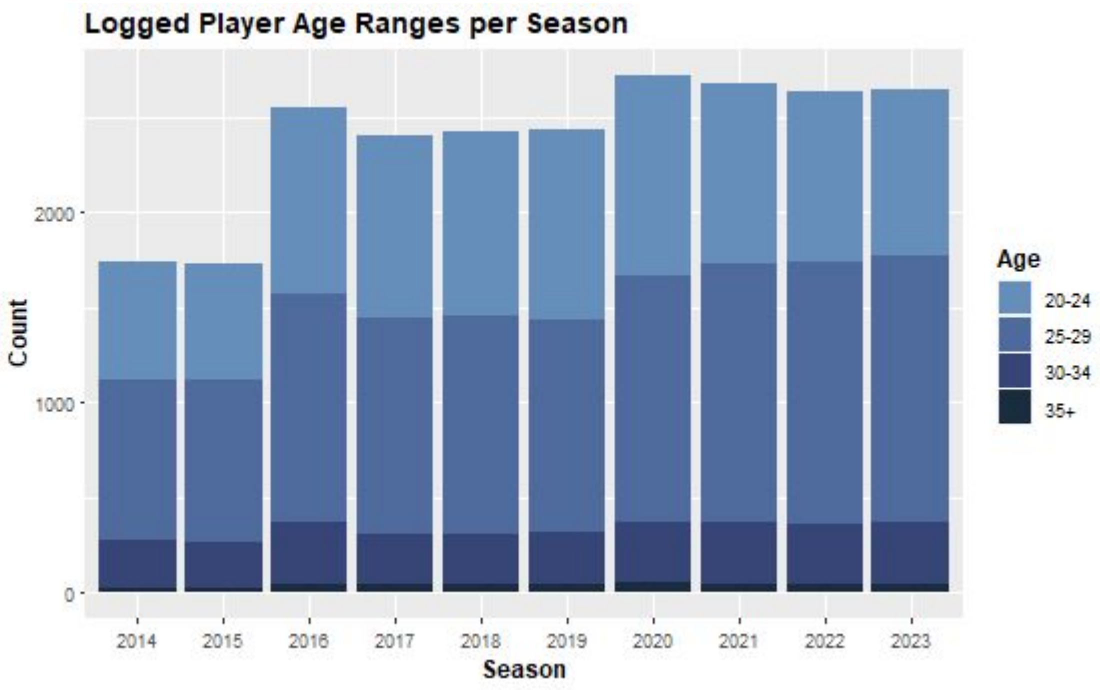


Q1: Posteriors / Discussion



Question 2: Has Age Composition Changed?

Q2: EDA


Logged Player Age Ranges per Season
Count
2000
1000
0
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Season
Age
20-24
25-29
30-34
35+

Q2: The Data

prob_inj	age_20_24	age_20_24_hurt	age_25_29	age_25_29_hurt	age_30_34	age_30_34_hurt	age_35_up	age_35_up_hurt
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0.3433180	614	453	848	131	246	8	28	0
0.3606936	613	504	845	105	243	6	29	0
0.2881356	974	680	1201	97	325	4	49	0
0.2973761	960	642	1133	66	265	5	43	0
0.2522634	975	565	1142	43	263	4	47	1
0.2709016	1008	605	1112	53	267	2	51	1
0.2647275	1045	666	1295	50	321	3	55	0
0.2547628	948	644	1361	37	323	1	45	0
0.2789814	894	689	1376	42	316	2	45	1
0.2778828	877	688	1396	41	326	3	46	0

Q2: The Model (Dirichlet Conjugate)

Dirichlet as a conjugate prior

$$P(\mathbf{X}|\boldsymbol{\pi}) = \prod_n \mathrm{Cat}(\mathbf{x}_n|\boldsymbol{\pi}) = \prod_n \prod_c \pi_c^{x_{cn}} = \prod_c \pi_c^{m_c}$$

$$\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_c \alpha_c)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_C)} \prod_{c=1} \pi_c^{\alpha_c - 1}$$

$$p(\boldsymbol{\pi}|\mathbf{X}) = \frac{P(\mathbf{X}|\boldsymbol{\pi})p(\boldsymbol{\pi})}{P(\mathbf{X})} \propto \prod_c \pi_c^{m_c} \prod_c \pi_c^{\alpha_c - 1}$$

$$= \prod_{c=1} \pi_c^{m_c + \alpha_c - 1} \propto \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha} + \boxed{\mathbf{m}})$$

Sufficient statistics

- Using **Dirichlet as a prior for Categorical parameter** π results in **Dirichlet posterior** distribution ➔ **Dirichlet is conjugate prior to Categorical dist.**
- $\alpha_c - 1$ can be seen as a prior count for the individual events.

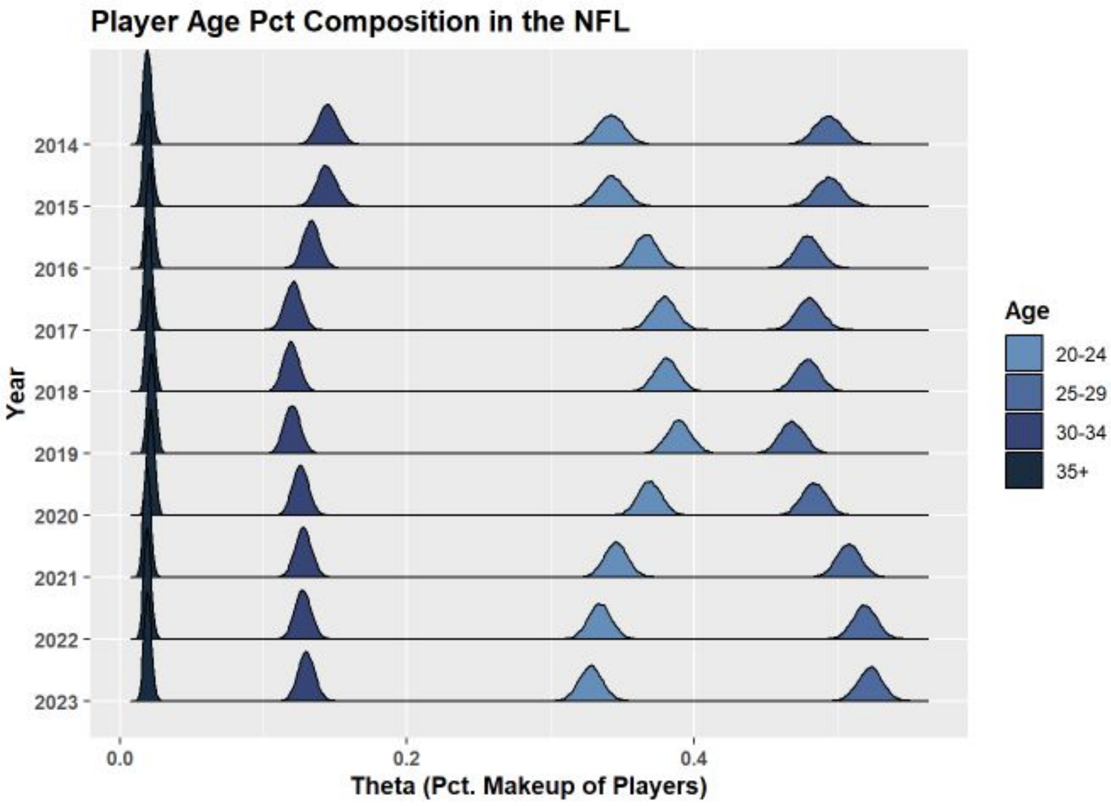
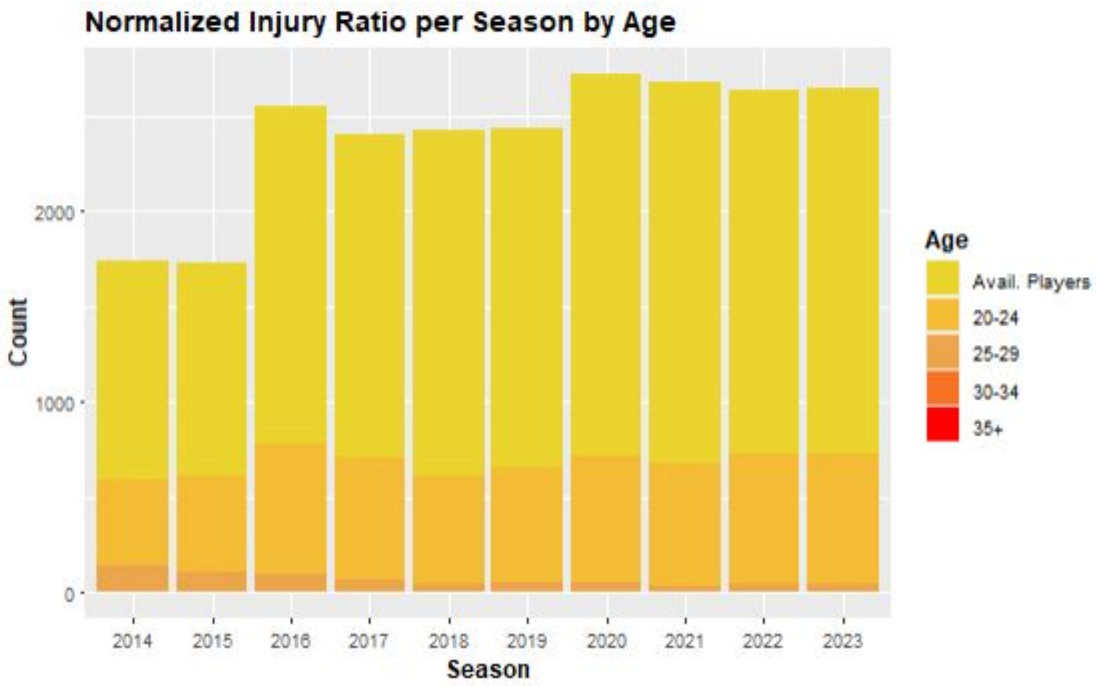
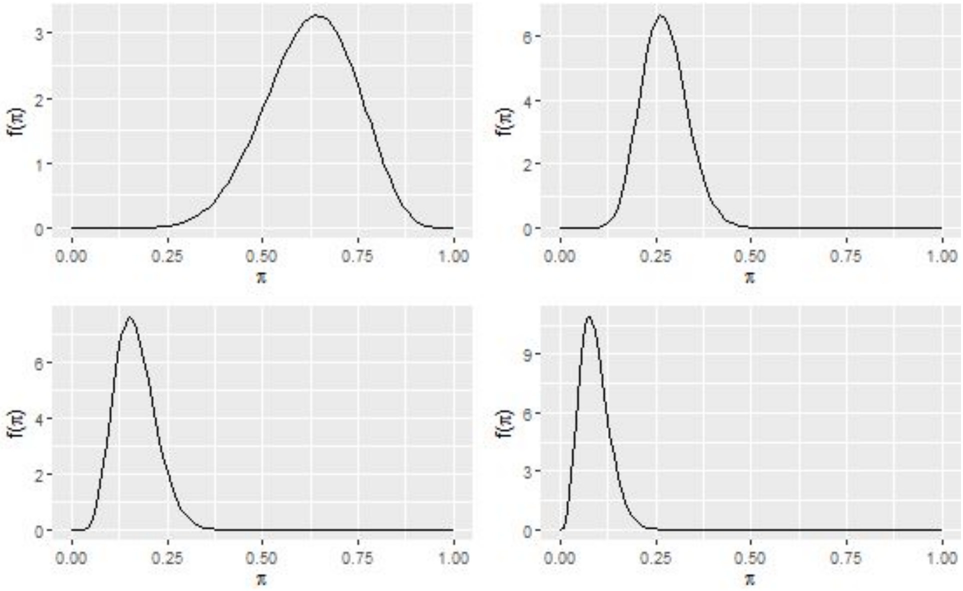
Q2: Model (Hyperparameters)

- Empirical Hyperparameters

Thus:

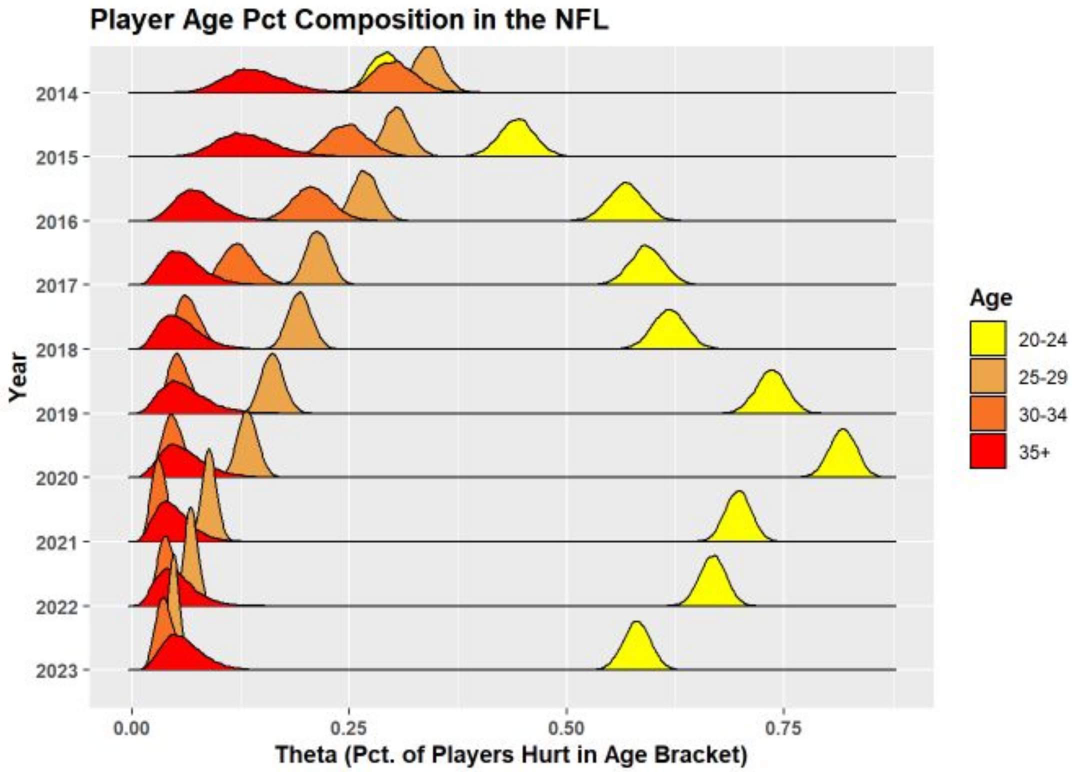
- alpha1 = 250
- alpha2 = 400
- alpha3 = 120
- alpha4 =20

Q2: Sampling (Analytical)

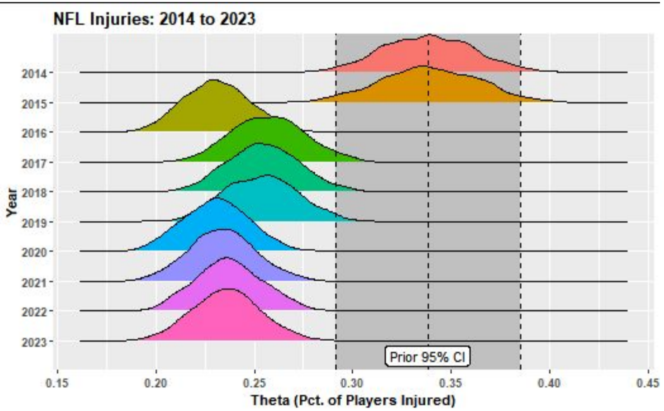
```
```{r analytical}
iter <- 10000

alpha <- c(250,400,120,20)
n_matrix <- subset(players_age[6:15, 2:5])
sims <- array(NA, c(iter, length(alpha), nrow(n_matrix)))

for (i in 1:10){
  sims[, ,i] = rdirichlet(iter, (alpha + n_matrix[i,]))
}

```
```

Q2: Posteriors / Discussion



Question 3: Which Ages are Getting Hurt?

Q3: EDA



Normalized Injury Ratio per Season by Age
Count
2000
1000
0
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Season
Age
Avail. Players
20-24
25-29
30-34
35+

Q3: The Data

prob_inj <dbl>	age_20_24 <dbl>	age_20_24_hurt <dbl>	age_25_29 <dbl>	age_25_29_hurt <dbl>	age_30_34 <dbl>	age_30_34_hurt <dbl>	age_35_up <dbl>	age_35_up_hurt <dbl>
0.3433180	614	453	848	131	246	8	28	0
0.3606936	613	504	845	105	243	6	29	0
0.2881356	974	680	1201	97	325	4	49	0
0.2973761	960	642	1133	66	265	5	43	0
0.2522634	975	565	1142	43	263	4	47	1
0.2709016	1008	605	1112	53	267	2	51	1
0.2647275	1045	666	1295	50	321	3	55	0
0.2547628	948	644	1361	37	323	1	45	0
0.2789814	894	689	1376	42	316	2	45	1
0.2778828	877	688	1396	41	326	3	46	0

Q3: The Model (Beta Conjugate)

- Prior

$$\mathrm{Beta}(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

$$\begin{aligned} p(\theta|y,n,M) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \end{aligned}$$

after normalization

$$p(\theta|y,n,M) = \mathrm{Beta}(\theta|\alpha+y,\beta+n-y)$$

Q3: Model (Hyperparameters)

- Empirical Hyperparameters



Thus:

- alpha1 = 10, beta1 = 6
- alpha2 = 15, beta2 = 40
- alpha3 = 8, beta3 = 40
- alpha4 =5, beta4 = 50

Q3: Sampling (Analytical)

```
```{r beta_posterior}
alpha = c(10, 15, 8, 5)
beta = c(6, 40, 40, 50)

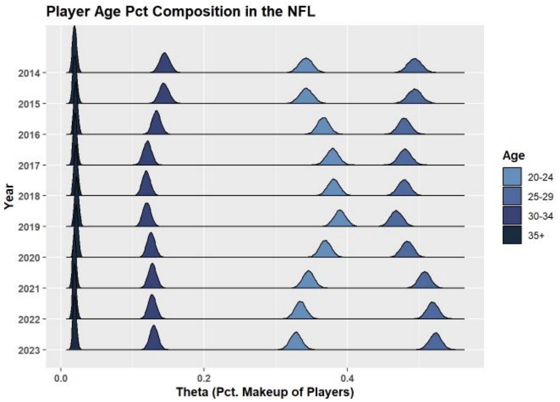
for (i in 1:10){
  for (j in 1:iter){
    sims[j, ,i] <- rbeta(4, alpha + (as.numeric(players_age2[i, 6:9])), beta +
(as.numeric(players_age2[i, 2:5]) - as.numeric(players_age2[i,6:9])))
}}
```
```

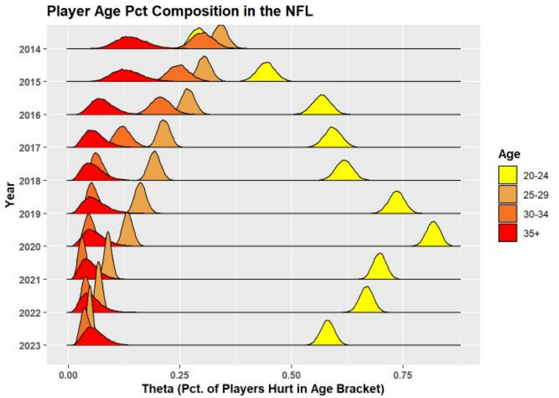
Q3: Posteriors / Discussion



Wrapping It All Up:







Conclusions

- 2016 and onwards has seen an impressive shift in the percentage of athletes getting injured compared to the past.
    - Which rulings are responsible?
    - Rulings or change in referee flexibility?
    - Both?... or another factor?
- Age has shown to be a significant determinant of the composition of players getting hurt (younger).
    - Sports Science / Inability to handle explosiveness
    - Inability to adapt to new league
    - Load Management
- NFL should continue implementing rule changes for player safety, but maybe with a focus towards protecting younger athletes.

Interested?: https://github.com/brandonowens24/NFL-Injuries-Bayesian-Analysis


That's all Folks!