Birthdays: Probability, Statistics and Data Visualisation

Zak Varty

Question

Talk to your neighbour:

- What is the probability that two or more people in this room share a birthday?

- What assumptions are you making to arrive at your answer?

03:00

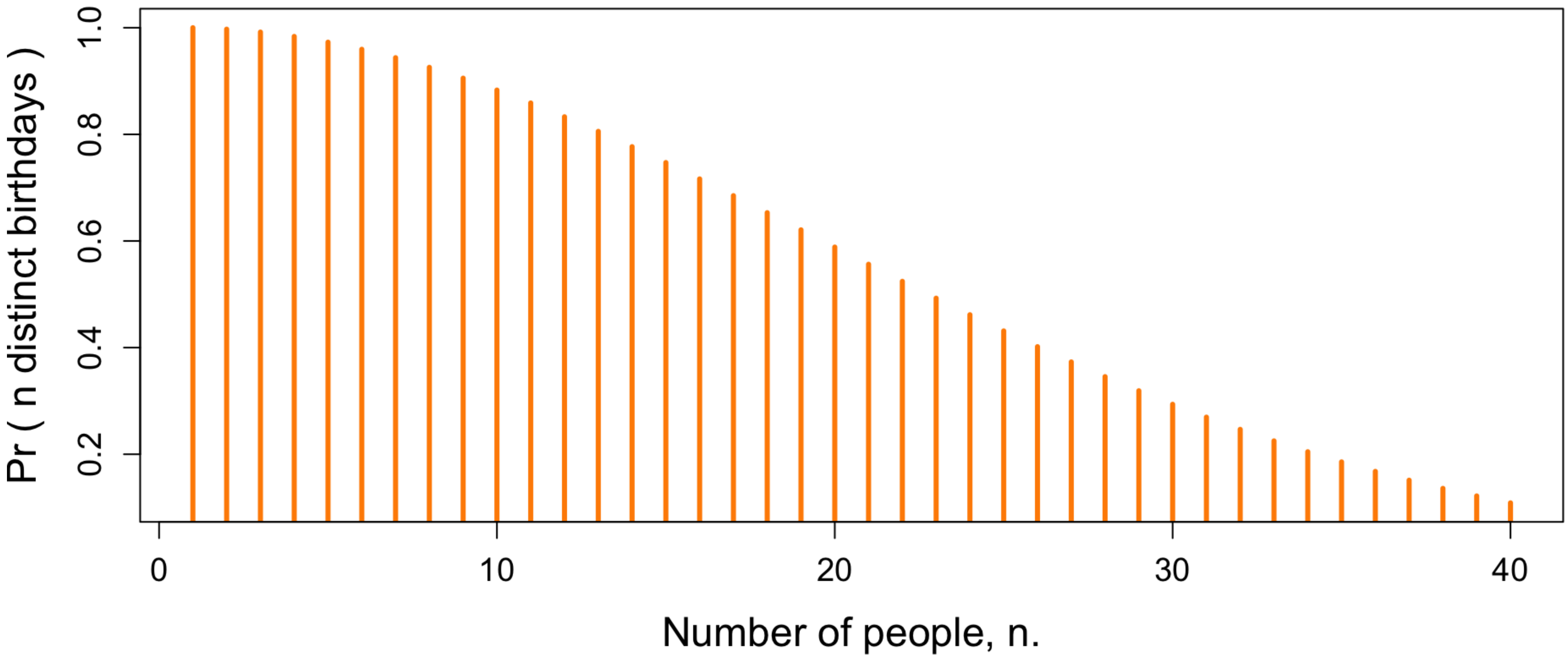
Typical Assumptions

- No twins, triplets, etc. (birth dates are independent)
- All days are equally probable.
- There are 365 days in the year.
- Pr(Shared birthday) = 1 - Pr(n distinct birthdays).

$$\Pr(\text{n distinct birthdays}) = \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365-(n-1)}{365}$$

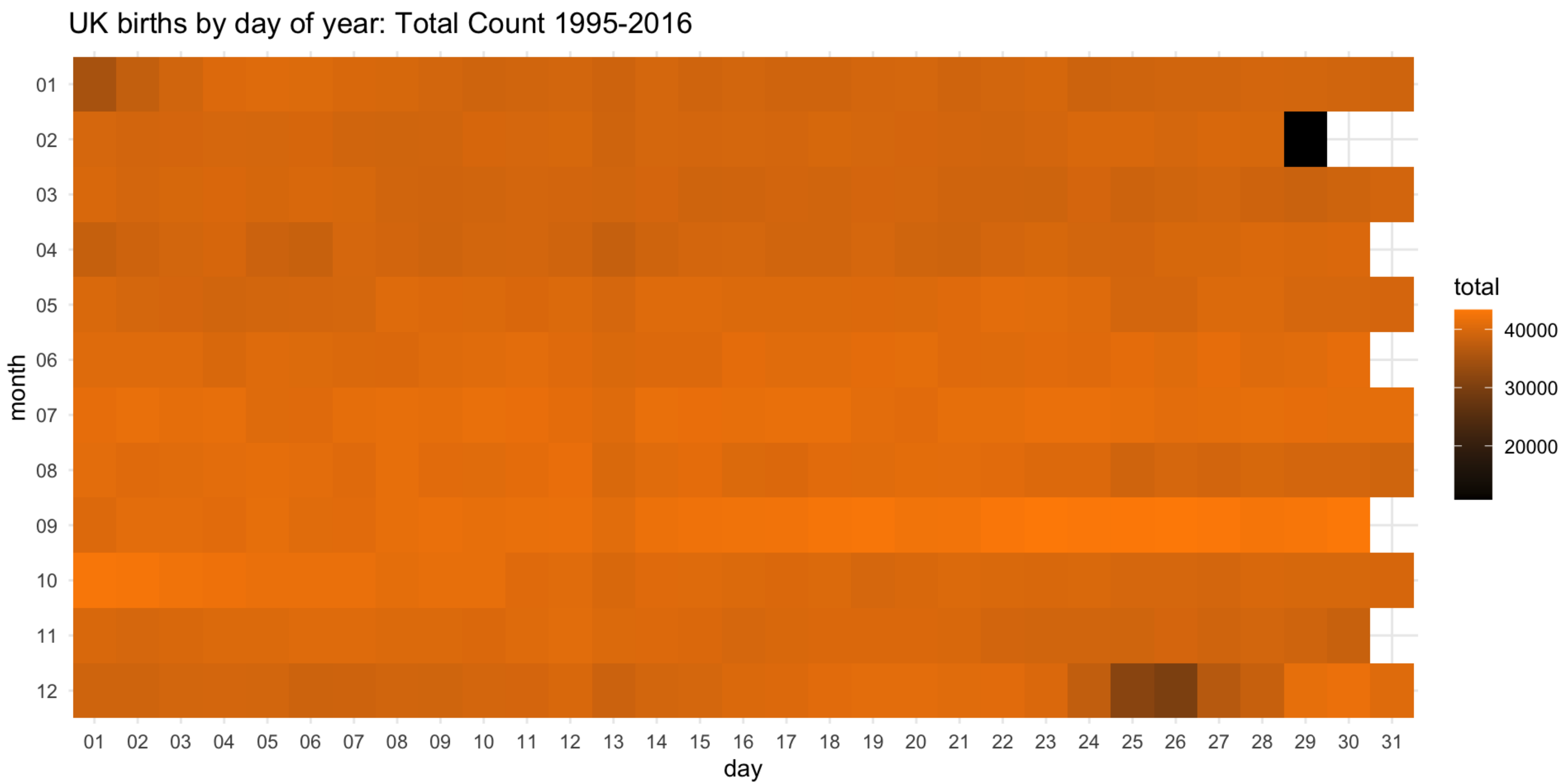
$$= \prod_{i=0}^{n-1} \left\{ \frac{365-i}{365} \right\}.$$

Plotting this Solution

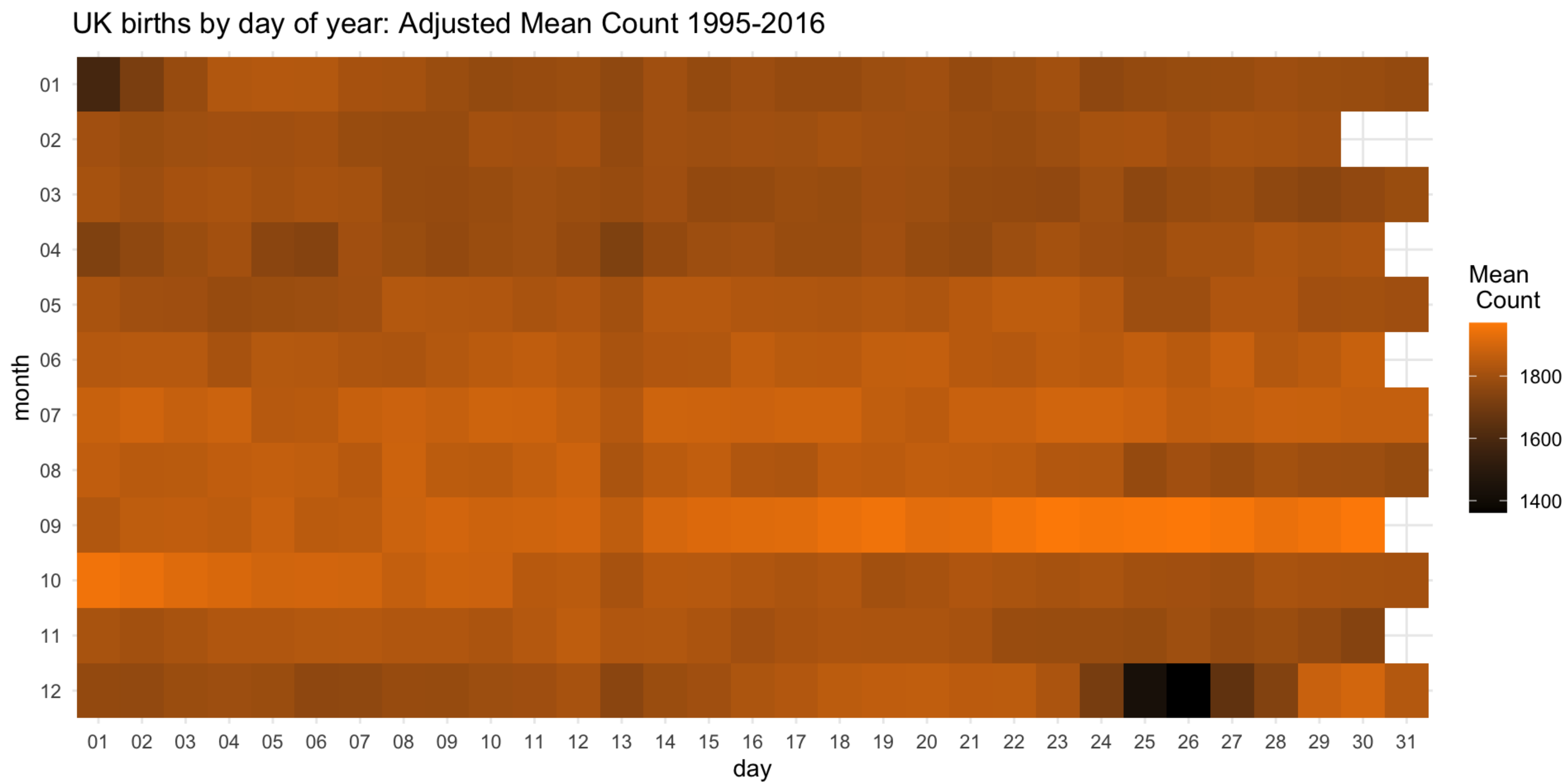


Testing Our Assumptions

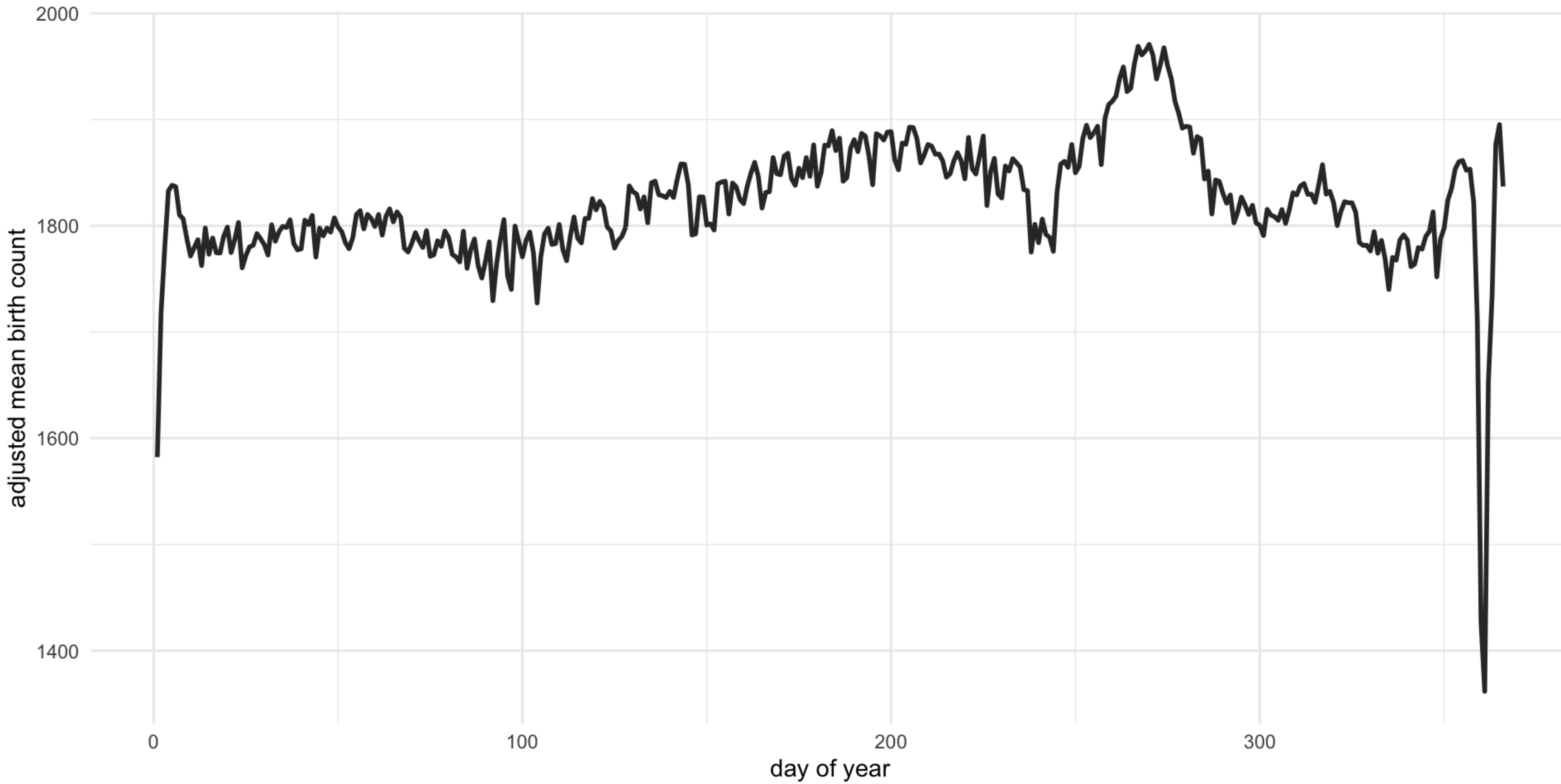
Are all days really equally likely?



Correcting for Feb 29



A different view of the same data



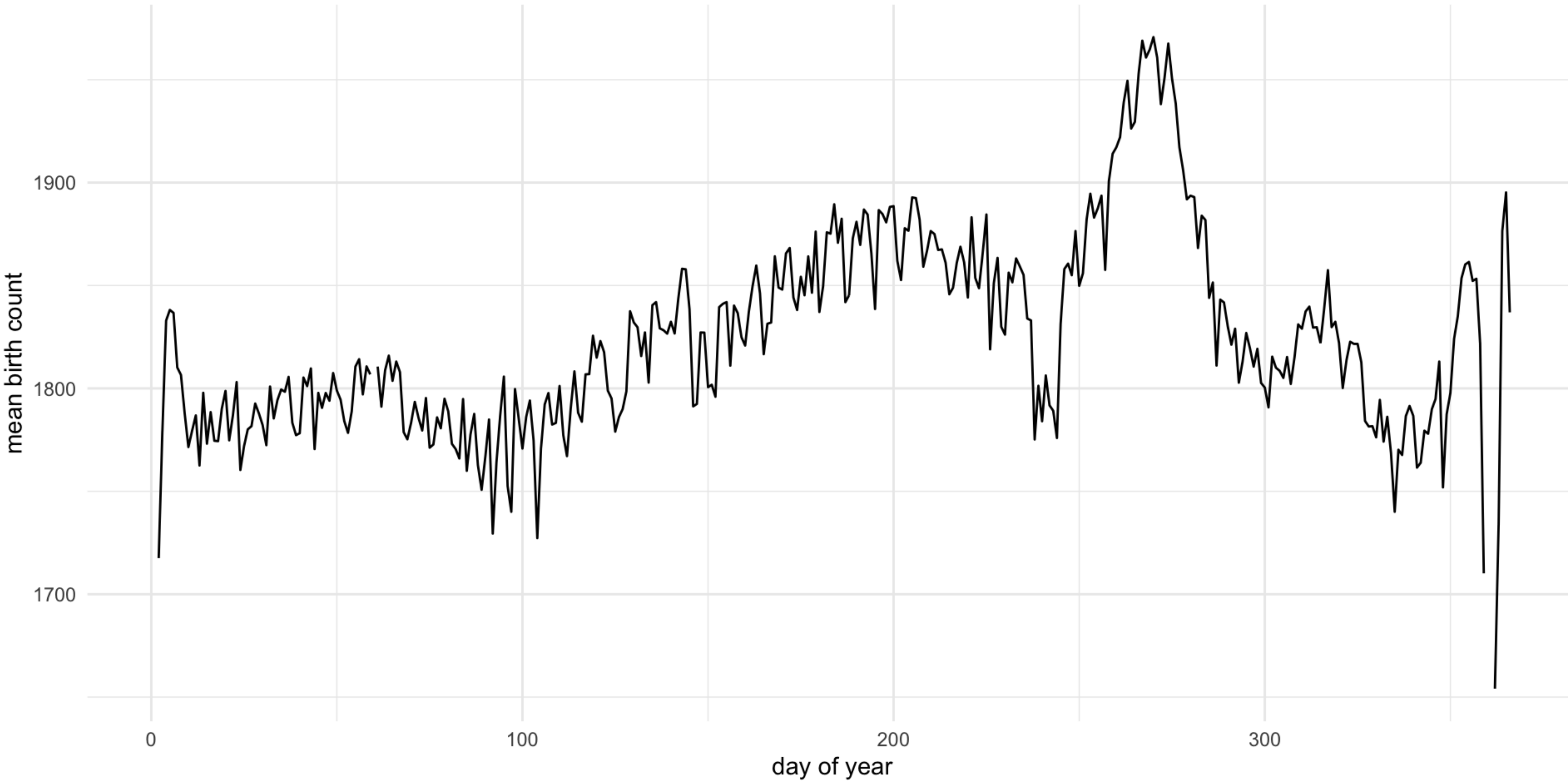
Handling holidays

- Only emergency treatments on Christmas, Boxing day and New Years Day.
- Treat these days separately. (Skip over hypothesis tests here)

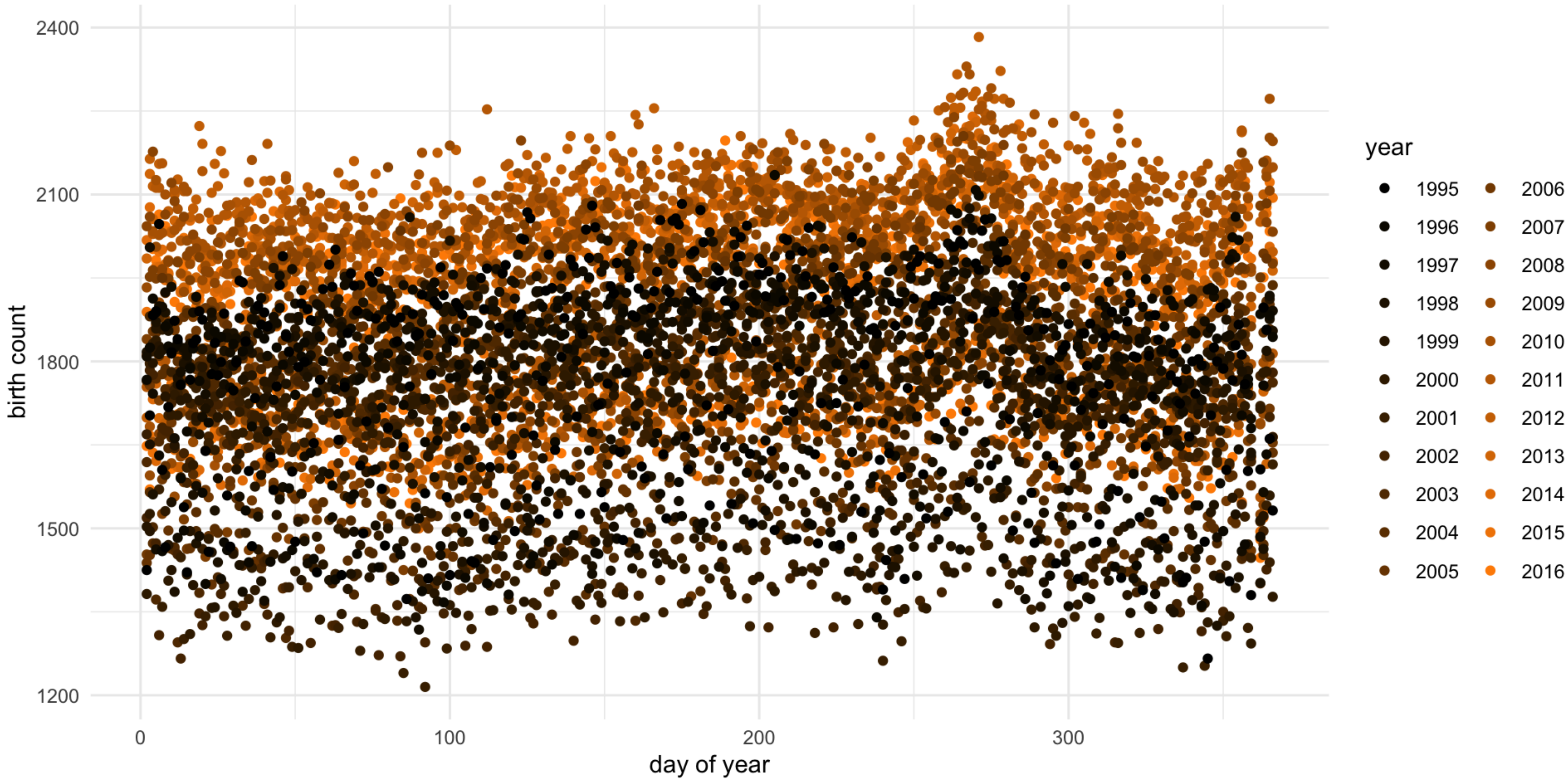
$$\hat{p} = \frac{\text{relative count}}{\text{total relative count}}$$

$$\hat{p}_{01/01} = 0.00237 \approx \frac{1}{421}, \quad \hat{p}_{25/12} = 0.00214 \approx \frac{1}{467}, \quad \hat{p}_{26/12} = 0.00204 \approx \frac{1}{490}.$$

What about the non-holidays?



But we have all of this information!



What are we testing?

$$\mathrm{H}_0 \ : \text{all non-holidays have the same mean birth count.}$$

$$\text{vs.}$$

$$\mathrm{H}_1 \ : \text{At least one day has a different mean.}$$

Bootstrap confidence interval - idea

- Under H_0 and the observations being i.i.d., the observed counts could have happened on any day
- We can build another day that we have not observed based on our data
- Build lots of these days and calculate the mean of each one
- Use the quantiles of these e.g. $(\bar{x}_{0.025}, \bar{x}_{0.975})$ to construct a confidence interval.

Bootstrap confidence interval - procedure

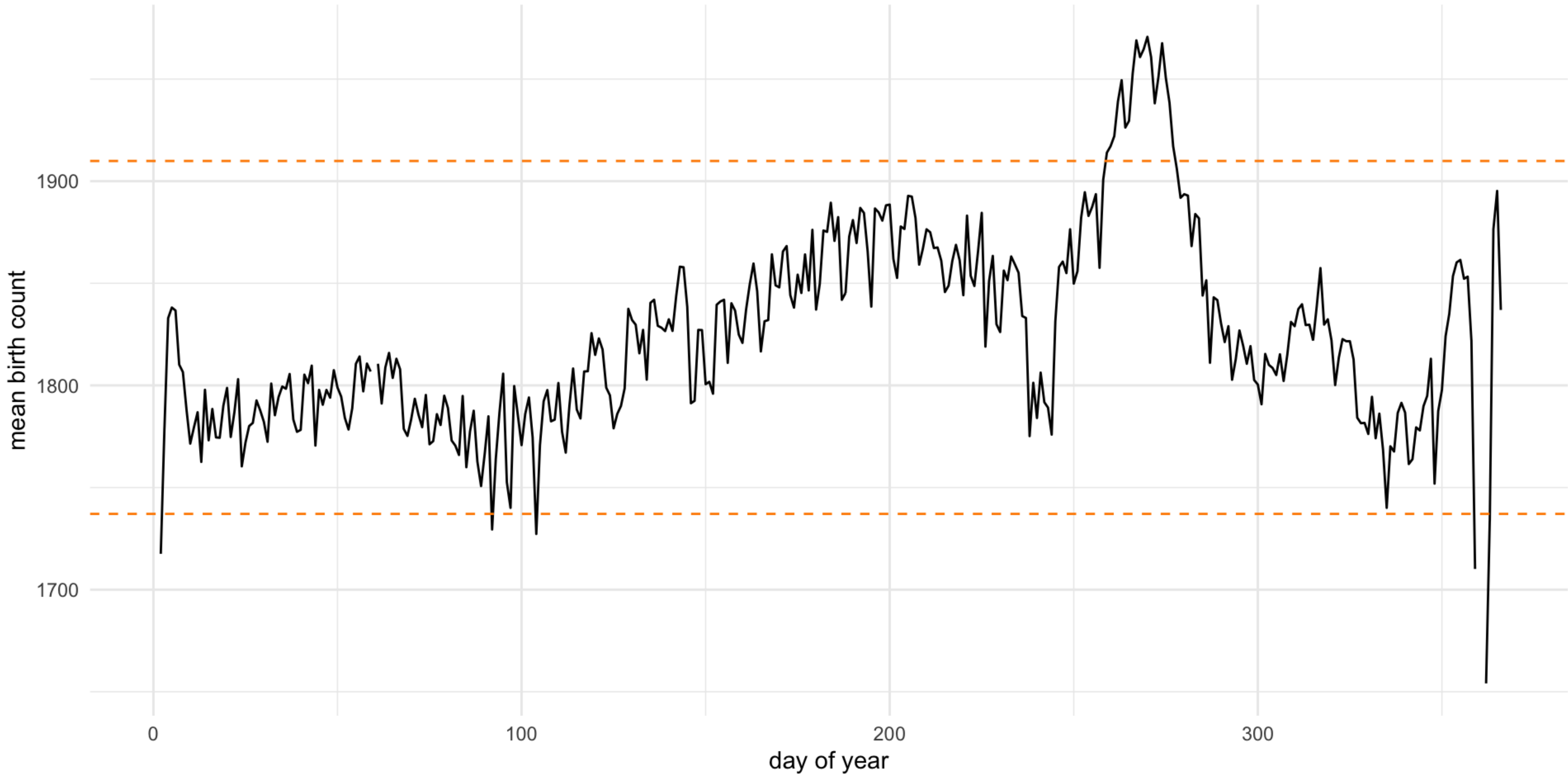
- Sample 22 non holiday birth counts at random, with replacement.
- Calculate their mean.
- Add this to a vector of means.
- Repeat until we have 10,000 means.
- Calculate the desired quantiles of the mean vector.
- Compare to observations.

Bootstrap confidence interval - code

```r
# 1: set up data & storage
boot_data <- as.vector(as.matrix(non_hols[,3:24]))
boot_sample <- rep(NA,22)

boot_size <- 1e5
boot_means  <- rep(NA,boot_size)

# 2: make fake data sets where day does not matter
for (i in 1:boot_size) {
  boot_sample <- sample(boot_data, size = 22, replace = TRUE)
  boot_means[i] <- mean(boot_sample,na.rm = TRUE)
}

# 3: find "typical" range if day does not influence count
boot_CI <- quantile(boot_means, probs = c(0.025,0.975))
boot_CI
```

```
    2.5%    97.5%
1737.091 1909.900
```

Bootstrap confidence interval - results



Other questions

We could take this analysis much further:

- Is there dependence on the previous count(s)?
- Is there an effect from the day of week?
- Is there an trend over time?

If you want to learn more:
https://pudding.cool/2018/04/birthday-paradox/

Takeaways

- Statistics is about understanding the assumptions.
- The more assumptions you make the more you can conclude.
- Display choices impact what is seen and what is hidden.
- A statistician's work is never done!

Any quick questions? I'll be around at coffee for longer ones!