Using Statistical Tests

EES 4891/5891

Probability & Statistics for Geosciences

Jonathan Gilligan

Class #16: Thursday, March 06 2025

Setting Up

Setting Up

- Accept the GitHub Classroom exercise at https://classroom.github.com/a/rpZxd6KB

- Create a new RStudio project from the repository you create from the exercise assignment.

Learning Goals

Learning Goals

- Understand how to apply Chi-Squared (χ^2) and Komogorov-Smirnoff tests and when to use one or the other.
- Understand how Bayesian hypothesis testing can work for binomial data
- Understand how to use R to perform many of these tests.

Comments on Using χ^2 Tests

Comments on Using χ^2 Tests

- χ^2 test:
  - H_0: your data O_i are described by the theoretical distribution.

1. Make a histogram of your data with N_b bins, $O_i = \#$ **observations in bin**$_i$
2. Make a histogram of the theoretically expected counts E_i
3. Calculate test statistic ξ^2

$$\Xi^2 = \sum_{i=1}^{N_b} \frac{(E_i - O_i)^2}{E_i}$$

  - The closer Ξ^2 is to zero, the closer your observed data are to the theoretical expectation.

- If H_0 is true, then Ξ^2 behaves like a random variable drawn from a $\chi^2_{\nu-1}$ distribution.
  - The probability that you would see $\Xi^2 \geq$ what you observed, if H_0 is true is

$$1 - \mathrm{CDF}_{\chi^2_{\nu-1}}(\Xi^2) = \int_{-\infty}^{\Xi^2} \chi^2_{\nu-1}(x)\mathrm{d}x$$

4. For a test level α, reject H_0 if $1 - \mathrm{CDF}_{\chi^2_{\nu-1}} < \alpha$
   - α is the probability of a Type-I error (false positive)

- How many bins (N_b) to use?
  - Rule of thumb:
    - Roughly 80% of the bins should have $E_i \geq 5$

Kolmogorov-Smirnov Test

- Kolmogorov-Smirnov Test
  - Data: x_i, $i = 1, 2, \ldots N$

1. Sort your data from smallest to largest. The empirical cumulative distribuion $F_N(x)$ is given by the pairs $(x_i, i/N)$ with y range from 0 to 1.
2. Calculate $F(x)$: CDF for the theoretical distribution
3. The test statistic is D:

$$D = \max_x |F_N(x) - F(x)|$$

4. Reject H_0 at level α if $D > C_N(\alpha)$
   - $C_N(\alpha)$ is universal. It doesn't depends on the theoretical distribution

- Choosing between χ^2 and K-S tests:
  - KS only works for data with continuous values. χ^2 also works with discrete values.
  - KS doesn't account for diminishing degrees of freedom if you estimate parameters from data.
    - You need a lot of data
  - $F(x)$ can be very hard to calculate for some distributions
  - KS can be preferable if the theoretical distribution can't satisfy the χ^2 condition of 80% of bins having $E_i > 5$.
  - χ^2 can be inconvenient to set up with continuous values.
    - KS can be preferable for date with continuous values.

Bayesian Hypothesis Testing

Sunny Resort Example

- A resort advertises that $6/7 = 85.7\%$ of its days are cloud-free (H_0).
- Observations of 25 days of weather report 15 cloud-free days (60%).
- Can we reject H_0 and conclude that the resort's advertising is misleading?
- Frequentist approach:
  1. Test statistic $f = k/N = 15/25 = 0.6$
  2. H_0: $f = f_c = 6/7 = 0.857$
  3. H_a: $f < f_c$, the resort is exaggerating.
  4. Null distribution: Probability of X cloud-free days out of 25, if the true fraction of cloud-free days is f_c

$$\mathbb{P}(X = k | f = f_c) = \binom{25}{k} f_c^k (1 - f_c)^{25-k}$$

- Test whether to reject H_0:
  5. p-value:

$$\mathbb{P}(X \leq 15) = \sum_{k=0}^{15} \binom{25}{k} f_c^k (1 - f_c)^{25-k}$$

```
f_c <- 6 / 7
seq <- map_dbl(1:15, \(k) choose(25, k) * f_c^k * (1
        - f_c)^{25 - k})
p <- sum(seq)
p
```

```
## [1] 0.001466964
```

  - $p = 0.0015$, So it is very unlikely that we could see this data if H_0 were true.

Bayesian Approach

- Bayesian approach:
  - Given the data, and a prior, calculate the posterior probability distribution for f.

1. Likelihood for N days of observations is a *binomial distribution*:

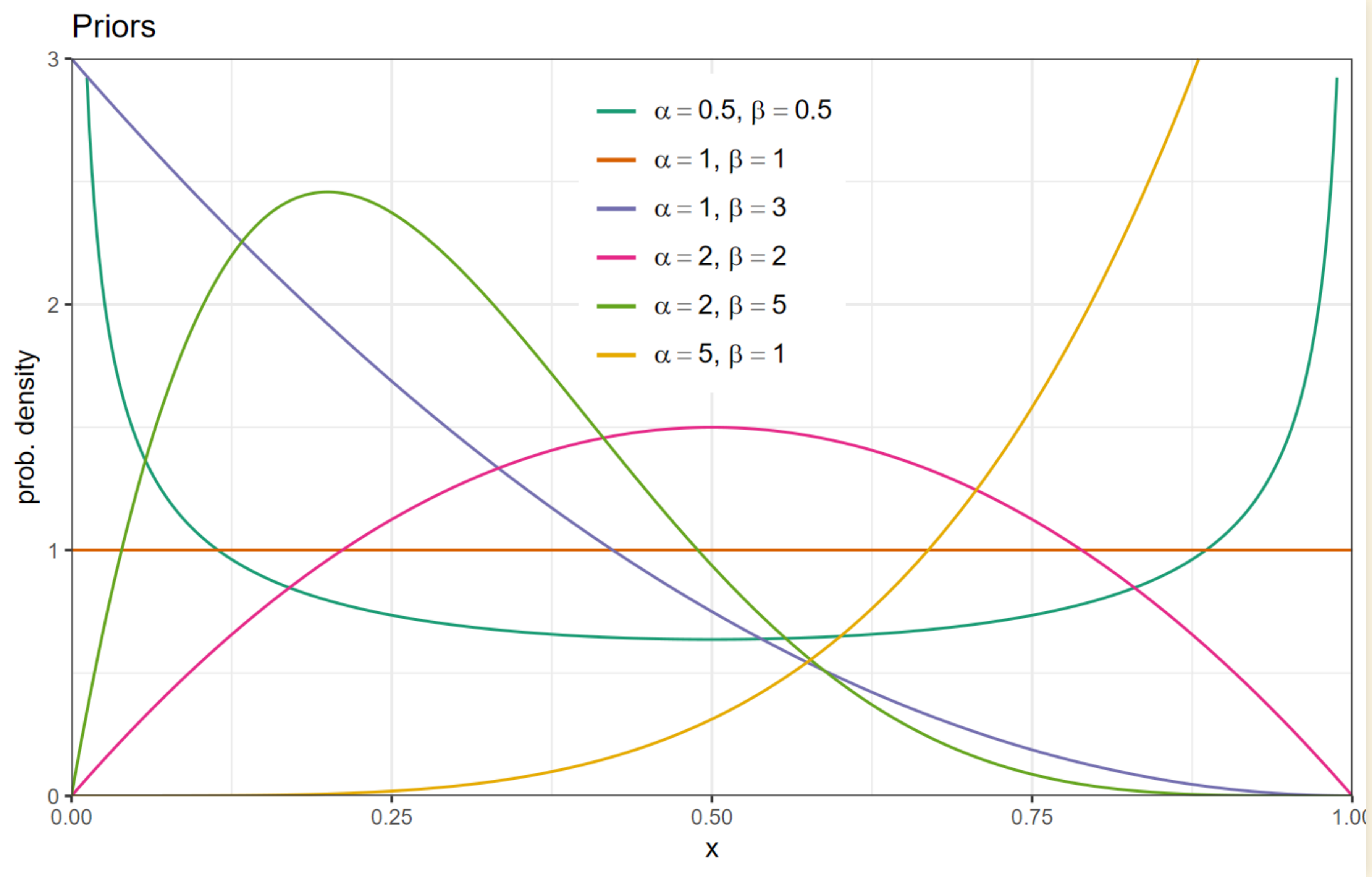
$$\mathbb{P}(X = k|f) = \binom{N}{k} f^k (1 - f)^{N-k}$$

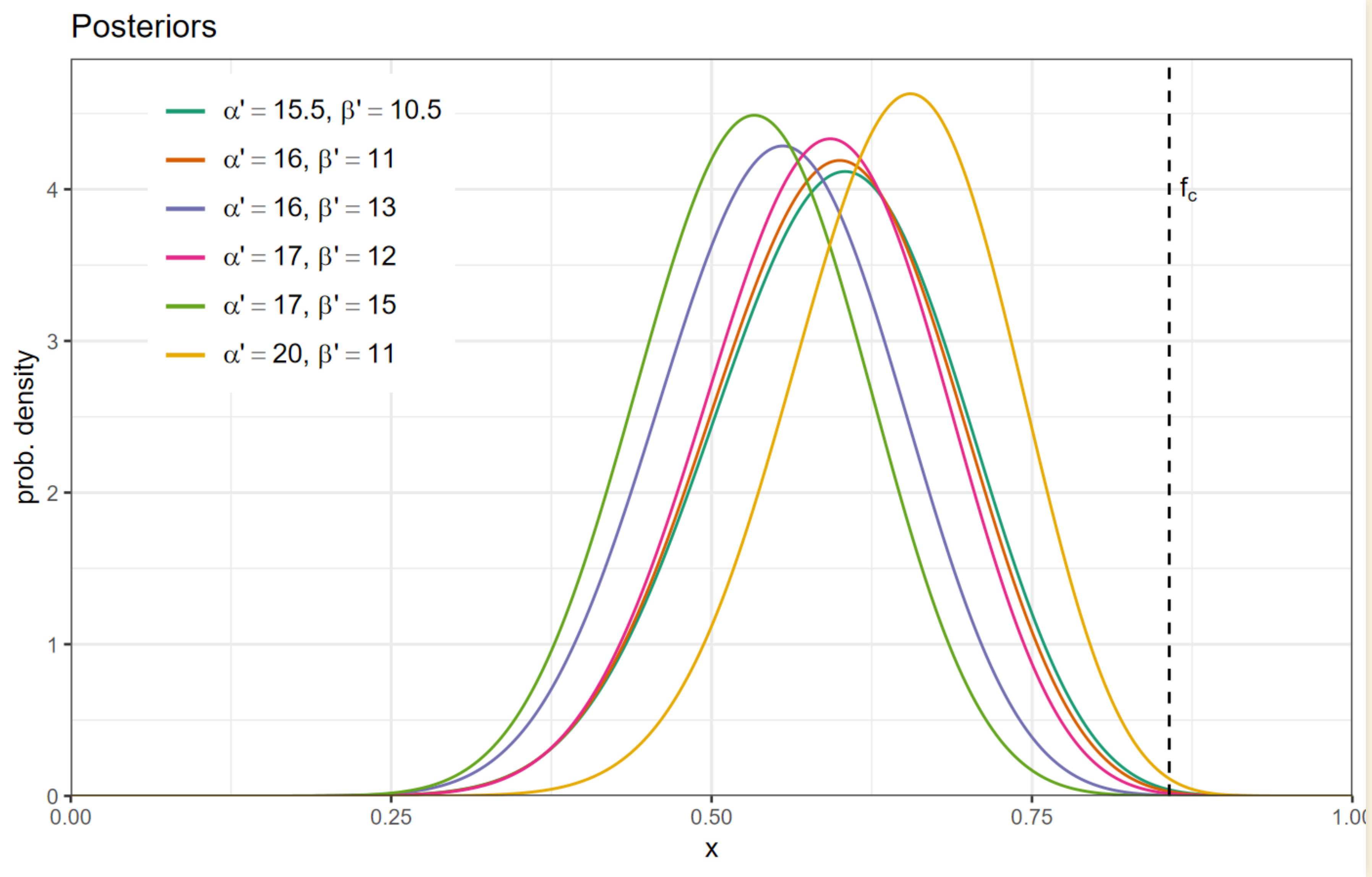
2. Prior for f:
   - We know that $0 \leq f \leq 1$.
   - A *beta* distribution $B(x|\alpha, \beta)$ is a good prior for probabilities.
   - $B(x|\alpha = 1, \beta = 1)$ is a uniform distribution with equal probabilities for all possible values of f.

3. Calculate the posterior.
   - The *beta* distribution has a special property:
   - It's a *conjugate prior* to a *binomial* distribution.
   - a *binomial* likelihood and a *beta* prior always gives a different *beta* for the posterior:

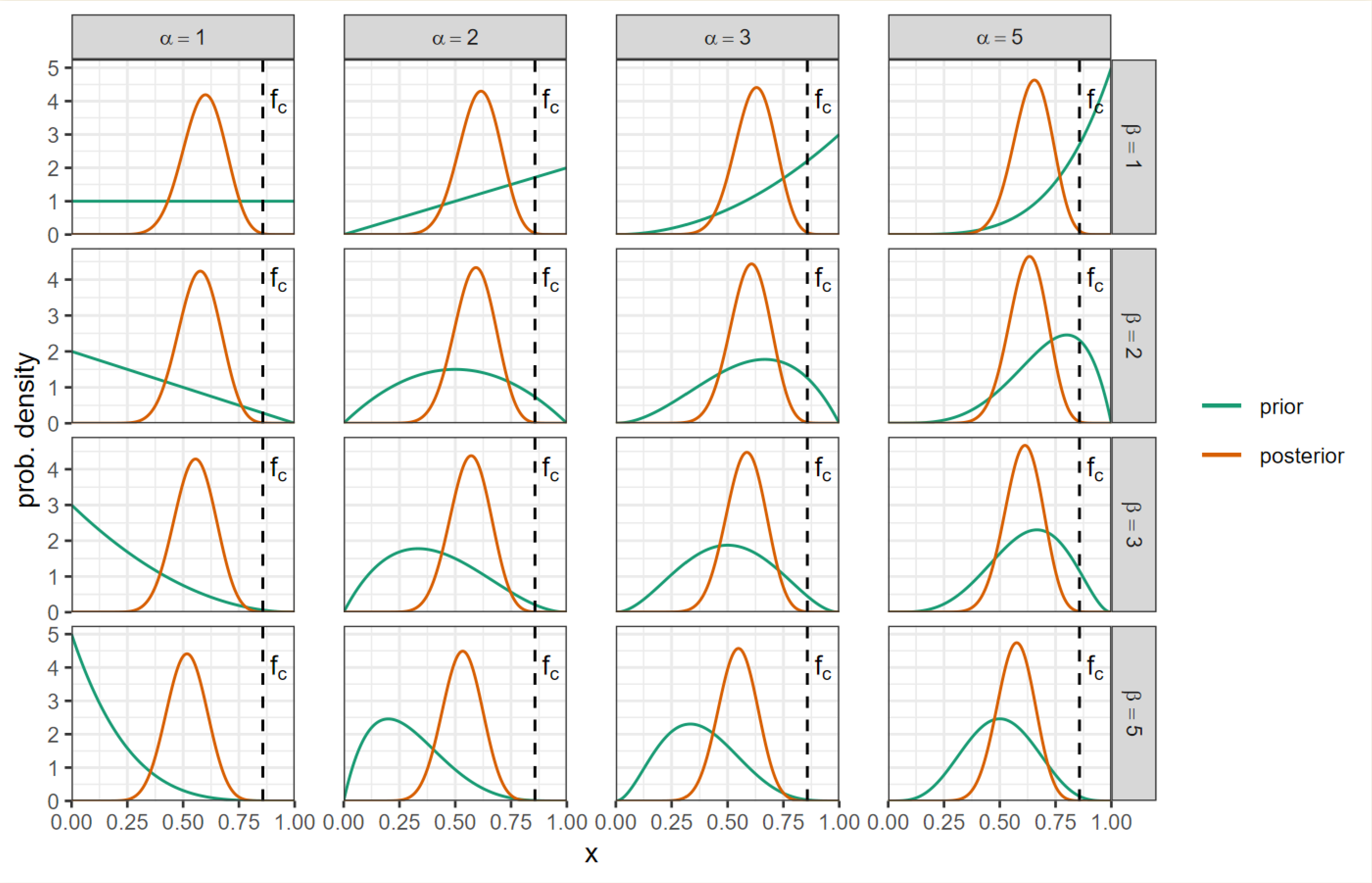
$$\mathbb{P}(f|X = k) = B(f|\alpha' = \alpha + k, \\ \beta' = N - k + \beta)$$

Beta Distributions





Priors and Posteriors



Exercises with R