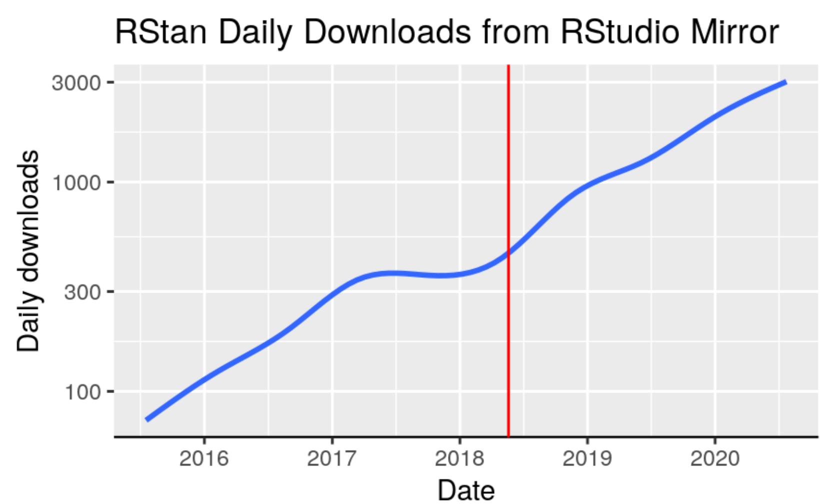
# Bayesian Inference without Probability Density Functions

Ben Goodrich ( benjamin.goodrich@columbia.edu ) YouTube Playlist
July 28, 2020

# Obligatory Disclosure

- Ben is an employee of Columbia University, which has received several research grants to develop Stan
- Ben is also a manager of GG Statistics LLC, which uses Stan for business
- According to Columbia University policy, any such employee who has any equity stake in, a title (such as officer or director) with, or is expected to earn at least $\$5,000.00$ per year from a private company is required to disclose these facts in presentations

0:00 / 0:19




*Season 3, Episode 9 of Billions*

# Main Points

- The majority of Stan users, the vast majority of potential Stan users, and nearly all Stan beginners should not be using Probability Density Functions
- Prior beliefs about unknowns are better articulated through quantile functions
- Just a handful of very flexible quantile functions can replace a multitude of well-known probability distributions that lack explicit quantile functions

# Bayes Rule Gets Unintuitive

- If $X$ and $Y$ are defined on discrete sample spaces, Bayes' Rule is intuitive:

$$\Pr(y \mid x) = \frac{\Pr(y) \times \Pr(x \mid y)}{\Pr(x)} = \frac{\Pr(y) \times \Pr(x \mid y)}{\sum_{y \in \Omega_Y} \Pr(y) \Pr(x \mid y)}$$

- If $X$ and $\theta$ are defined on continuous sample / parameter spaces, Bayes' Rule is less intuitive because it involves many Probability Density Functions (PDFs)

$$f(\theta \mid x) = \frac{f(\theta) \times f(x \mid \theta)}{f(x)} = \frac{f(\theta) \times f(x \mid \theta)}{\int_\Theta f(\theta) \times f(x \mid \theta)\, d\theta}$$

- But Bayes' Rule can be re-written under a change-of-variables from $\theta$ to $p$

$$f(p \mid x) = \left| \frac{\partial}{\partial p} \theta(p) \right| \frac{f(\theta(p)) \times f(x \mid \theta(p))}{f(x)} = \frac{f(p)\, f(x \mid \theta(p))}{f(x)}$$

# RNGs Are More Intuitive than PDFs

- Generative modeling is more fundamental to Bayesianism than Bayes' Rule is
- Prior predictive matching is fairly intuitive even on continuous parameter spaces since it operates at the RNG level (where $\sim$ reads as "is drawn from"):

$$\tilde{\theta} \sim \mathcal{B}eta\left(a, b\right); \tilde{x} \sim \mathcal{B}inomial\left(n, \tilde{\theta}\right)$$

  and then keep $\tilde{\theta}$ iff $\tilde{x} = x$. Acceptance proportion converges to $\Pr\left(x\right)$ and each kept $\tilde{\theta} \sim$ Beta$\left(\theta \mid a + x, b + n - x\right)$ (i.e. the posterior distribution)

- But in the Stan language, $\sim$ does NOT read as "is drawn from"

# Common Probability Distributions Are Not Useful

- There are too many probability distributions, leading to a paradox of choice
- None were originally intended to be used as priors
- Most common probability distributions were derived well before computers were invented to have elementary expressions for $\mu$ and $\sigma^2$
- People do not have prior expectations in their heads
- Historically, prior distribution families were chosen to do Gibbs sampling.
- Why has no one asked (until recently) “What probability distributions are most useful for expressing beliefs about unknowns?”

# The Beta Distribution Is Particularly Not Useful

- PDF is not elementary but $\mu = \frac{a}{a+b}$ and $\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$ are
- Can reparameterize as $a = \mu \left( \frac{\mu(1-\mu)}{\sigma^2} - 1 \right)$ and $b = (1-\mu) \left( \frac{\mu(1-\mu)}{\sigma^2} - 1 \right)$
- Beta distribution has the maximum differential entropy among all probability distributions over $\Theta = [0,1]$ that have a given $\mathbb{E} \ln \theta$ and $\mathbb{E} \ln(1-\theta)$

# Inverse Cumulative Distribution Functions (ICDFs)

- A Cumulative Distribution Function (CDF), $F\left(\theta \mid \ldots\right)$, is an increasing function from $\Theta$ to $[0,1]$ so its inverse is an increasing function from $[0,1]$ to $\Theta$
- $F^{-1}\left(0.5 \mid \ldots\right)$ is the median, while $F^{-1}\left(0.25 \mid \ldots\right)$ and $F^{-1}\left(0.75 \mid \ldots\right)$ are the lower and upper quartiles, so an ICDF is also called a quantile function
- If $\tilde{p} \sim Uniform\left(0,1\right)$ and $\tilde{\theta} = F^{-1}\left(\tilde{p} \mid \ldots\right)$, then $\tilde{\theta}$ is a realization from a probability distribution defined by that ICDF
- $\mathbb{E}\theta = \int_0^1 F^{-1}\left(p \mid \ldots\right) dp = \int_\Theta \theta f\left(\theta \mid \ldots\right) d\theta$ iff the integrals converge
- But CDFs and especially ICDFs rarely have explicit forms, whereas PDFs do

# Stan Skeleton with Inverse CDF Transformations

```
data {
  int<lower = 0> N;                    // number of observations
  vector[N] y;                         // observed outcomes
  ...                                  // known hyperparameters
}
parameters {
  real<lower = 0, upper = 1> p;        // cumulative probability
}
transformed parameters {
  real theta = some_icdf(p, ...);      // parameter of interest
}
model {
  y ~ likelihood(theta);               // function of p, not y
} // no explicit prior distribution for p because implicitly uniform
generated quantities {
  real prior_y = likelihood_rng(some_icdf(uniform_rng(0, 1), ...))
  real post_y  = likelihood_rng(theta);
}
```

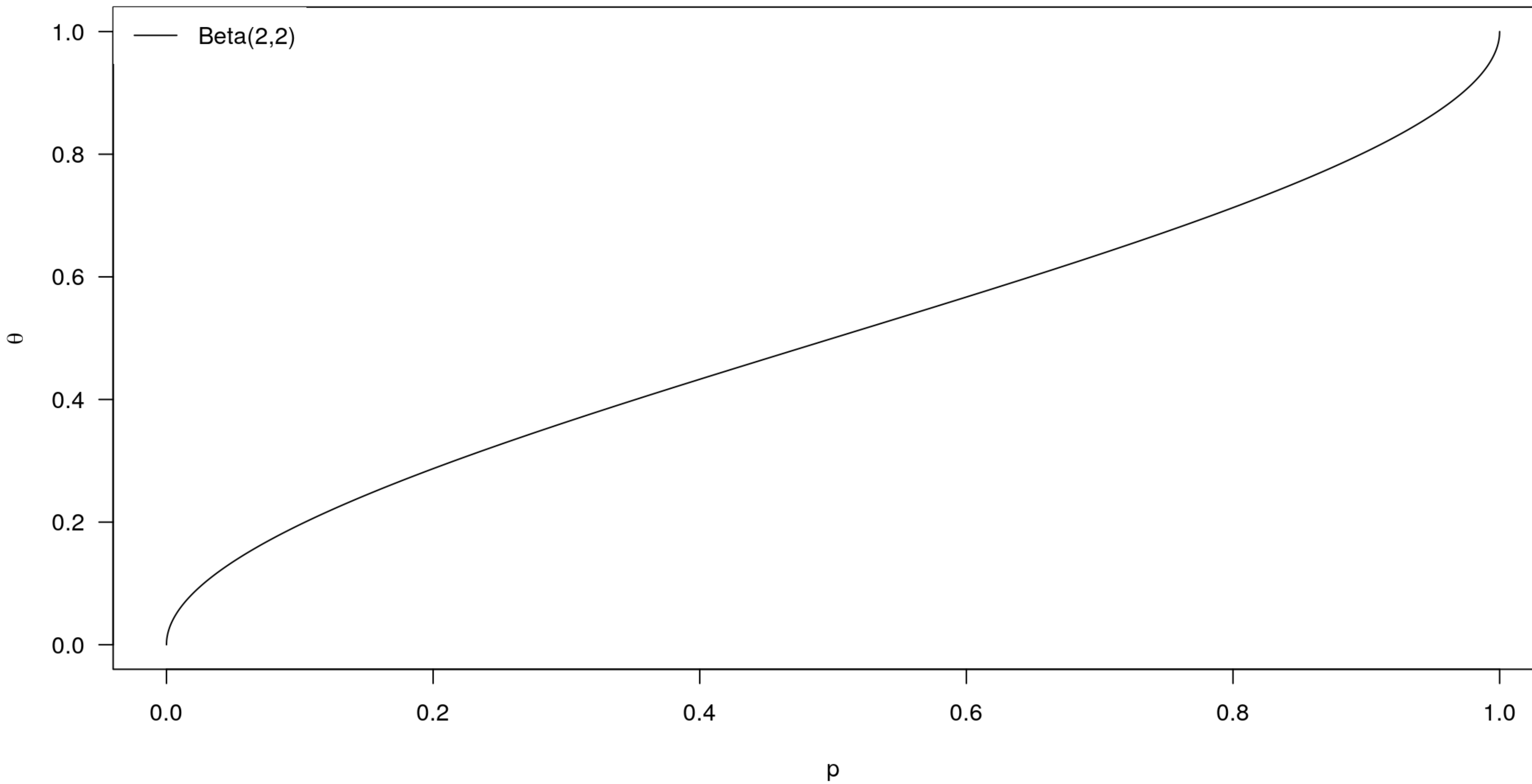
# Chebyshev Approximations of the 1st Kind $\left(T_k\right)$

- Suppose you wanted to approximate the ICDF of the Beta(2, 2) distribution
- Let $F^{-1}\left(p \mid a=2, b=2\right) = \sum_{k=0}^{\infty} c_k T_k\left(2p-1\right)$, where for all $k > 1$

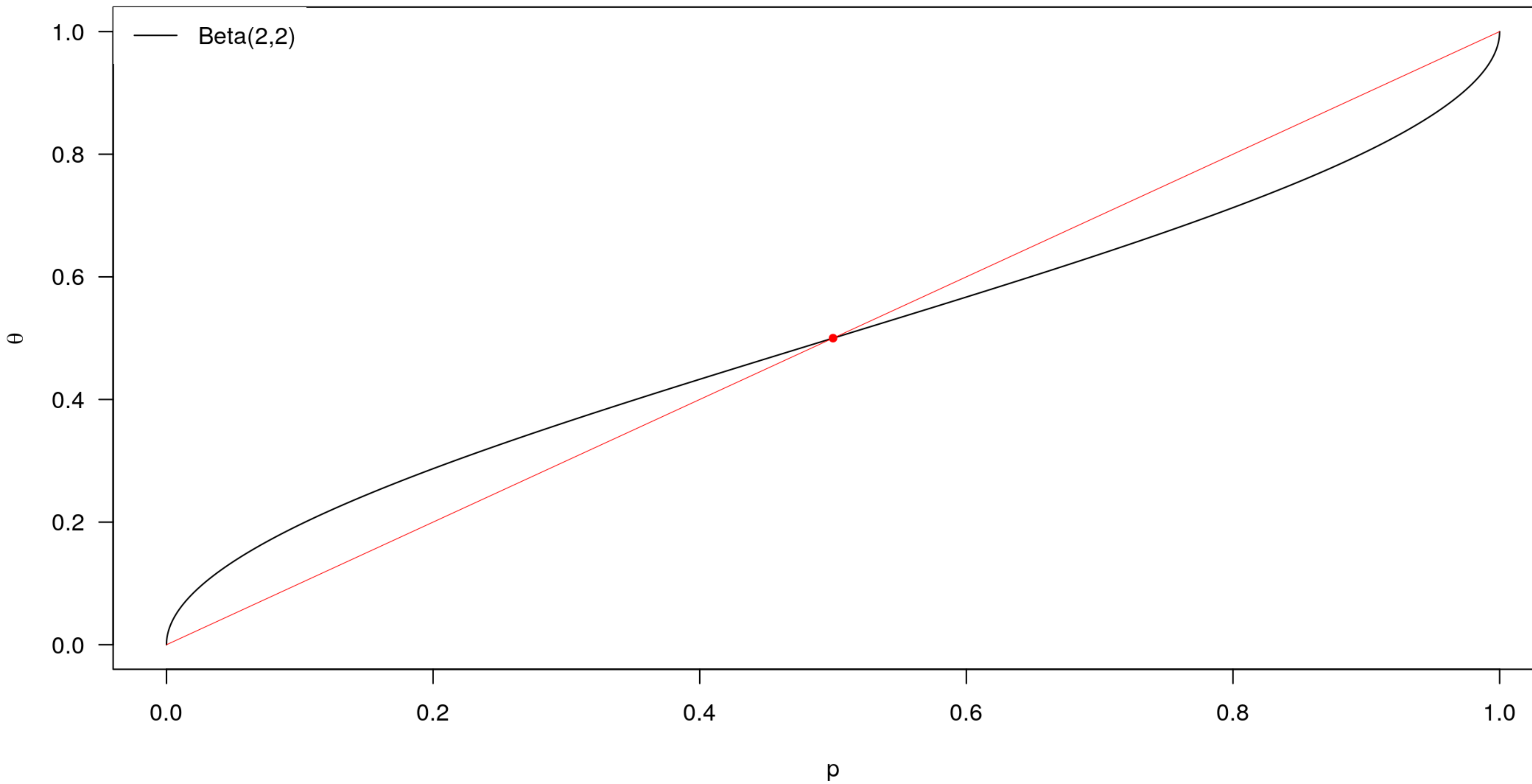
$$T_k\left(2p-1\right) \equiv 2\left(2p-1\right)T_{k-1}\left(2p-1\right) - T_{k-2}\left(2p-1\right)$$

  with base cases $T_0\left(2p-1\right) = 1$ and $T_1\left(2p-1\right) = 2p-1$
- $F^{-1}\left(p \mid a=2, b=2\right) \approx \sum_{k=0}^{K} c_k T_k\left(2p-1\right)$ for a given finite $K$
- Chebyshev approximation converges as $K \uparrow \infty$ for any Lipschitz-continuous ICDF in a nearly minimax way & the minimax way is rarely analytically feasible

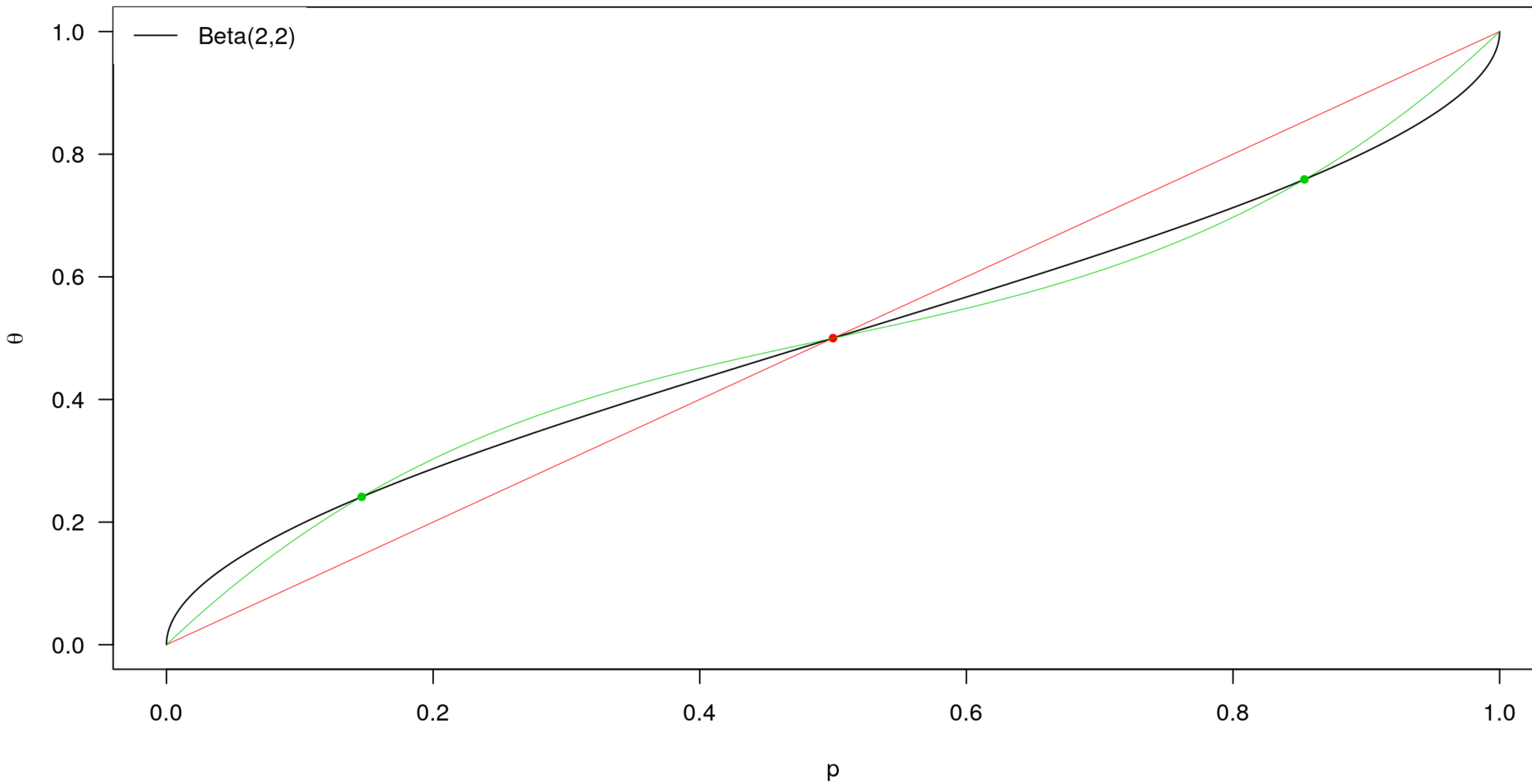
# Chebyshev Approximation of the Beta(2, 2) ICDF



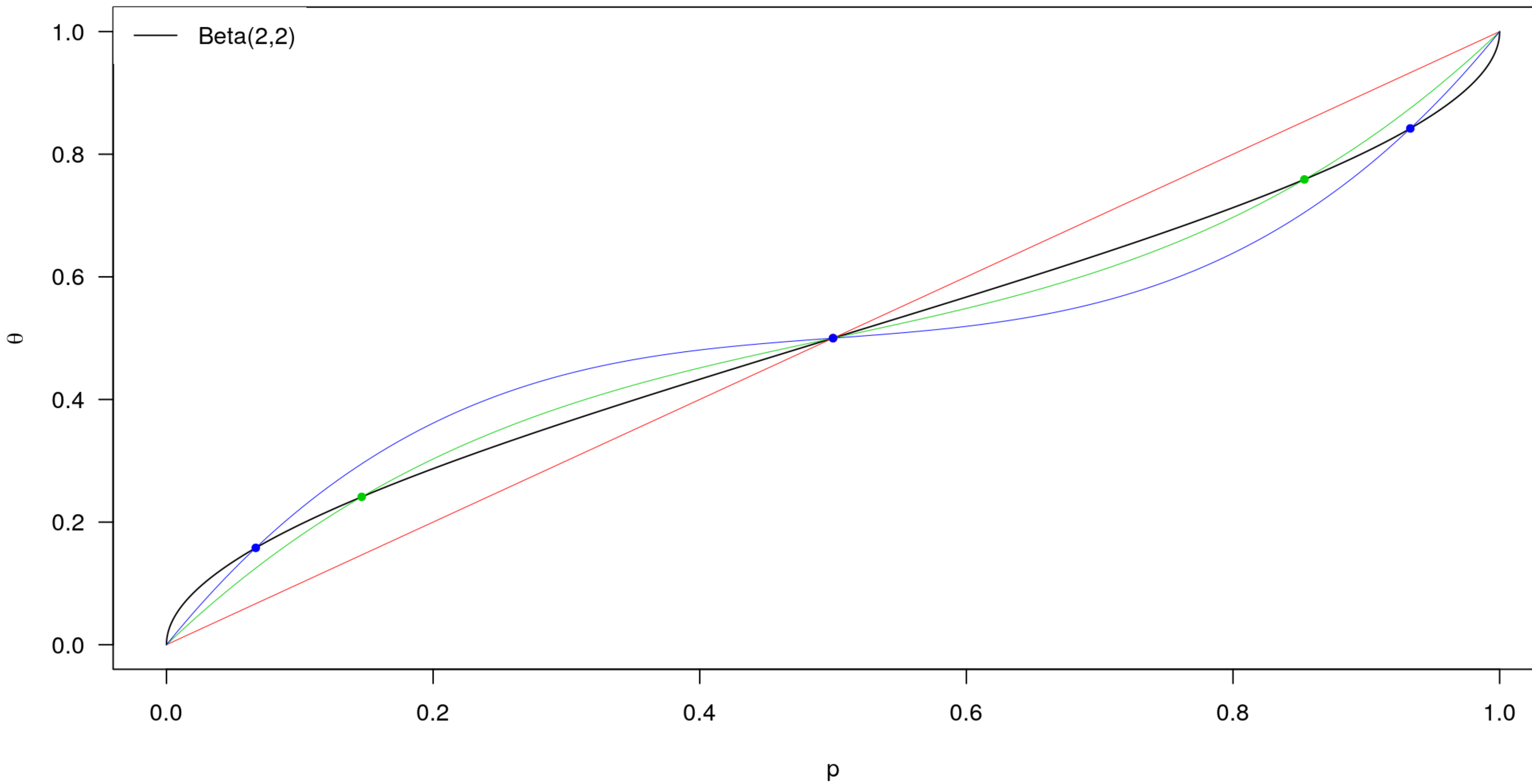
# Approximation with 1 interior and 2 end points



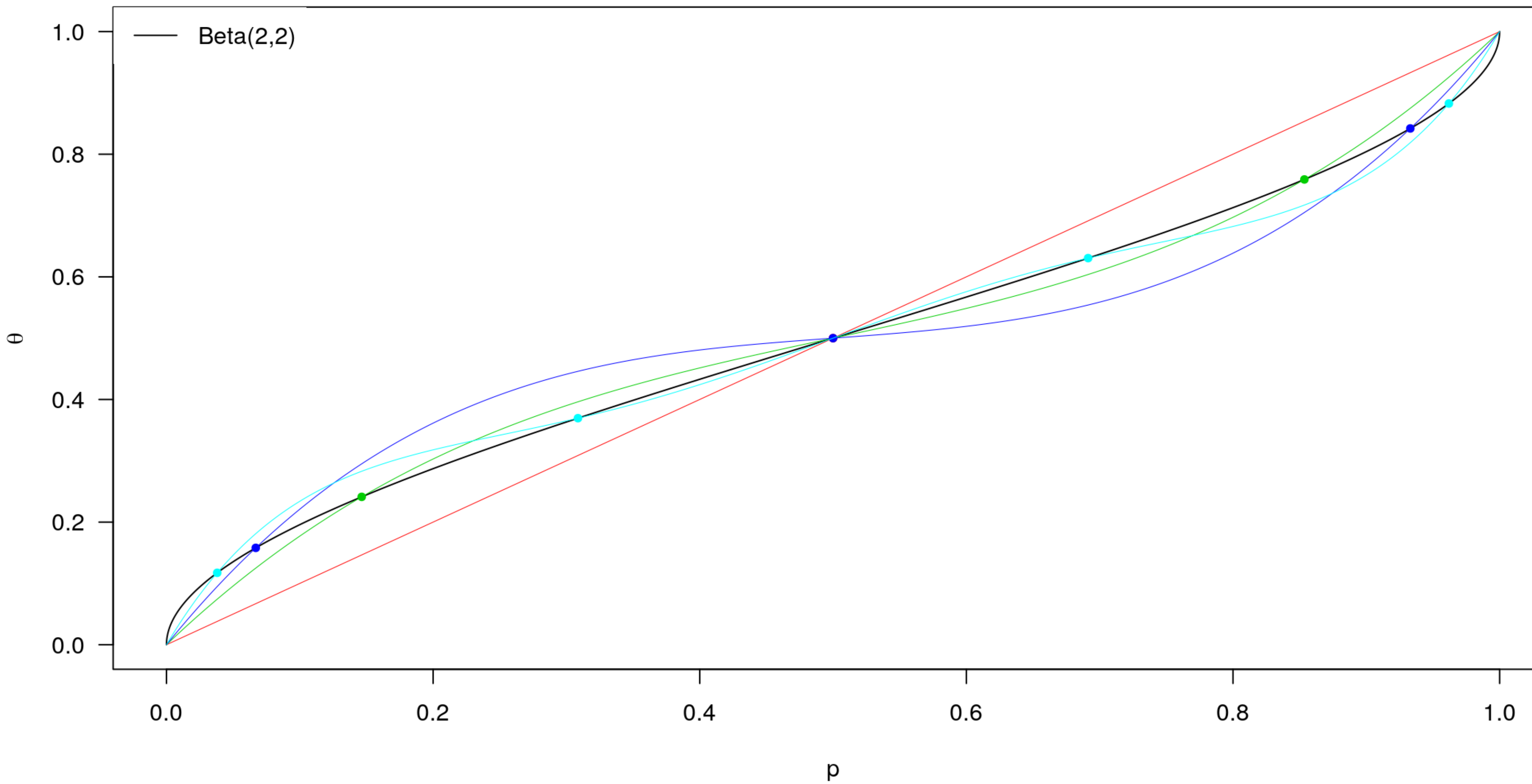
# Approximation with 2 interior and 2 end points



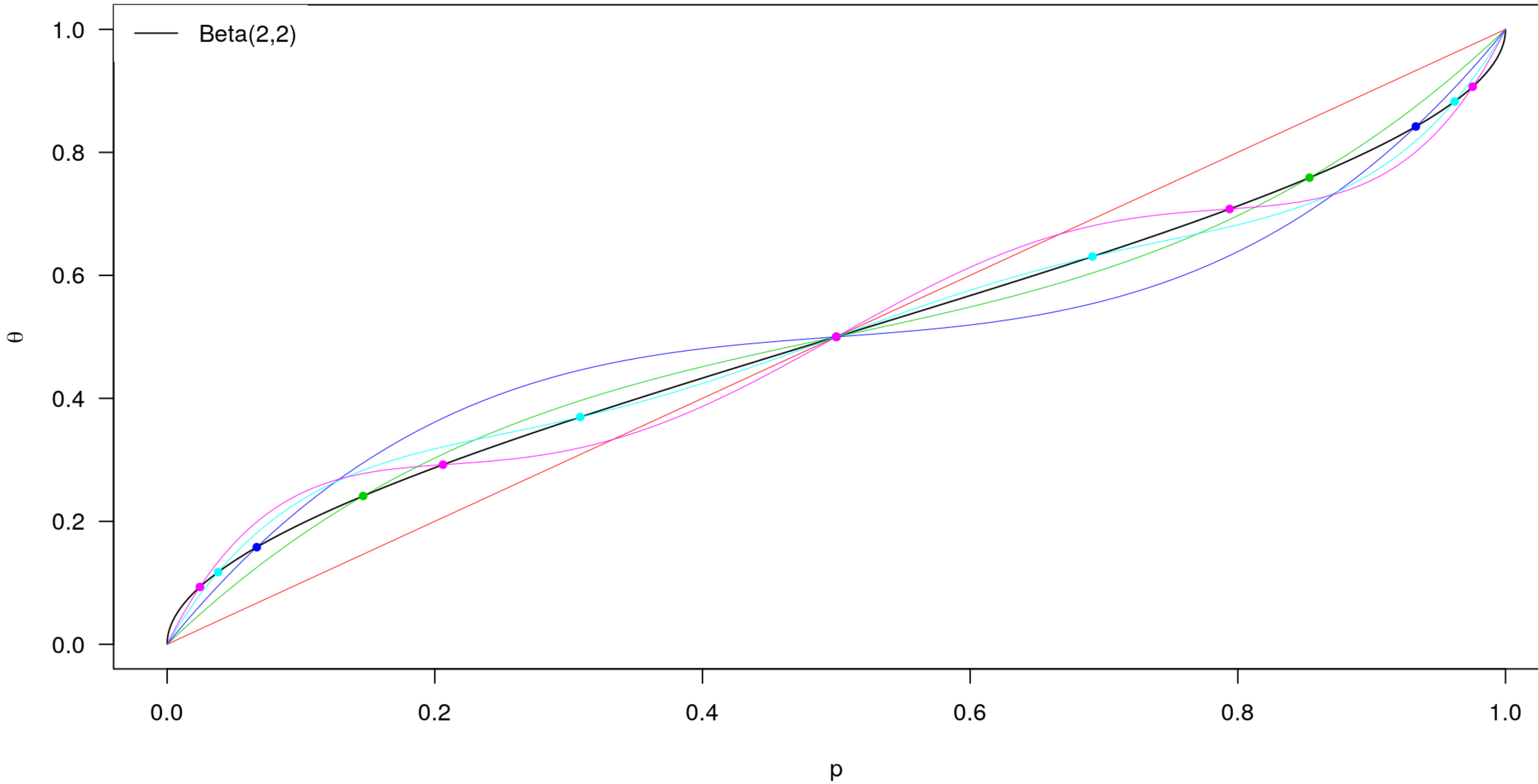
# Approximation with 3 interior and 2 end points



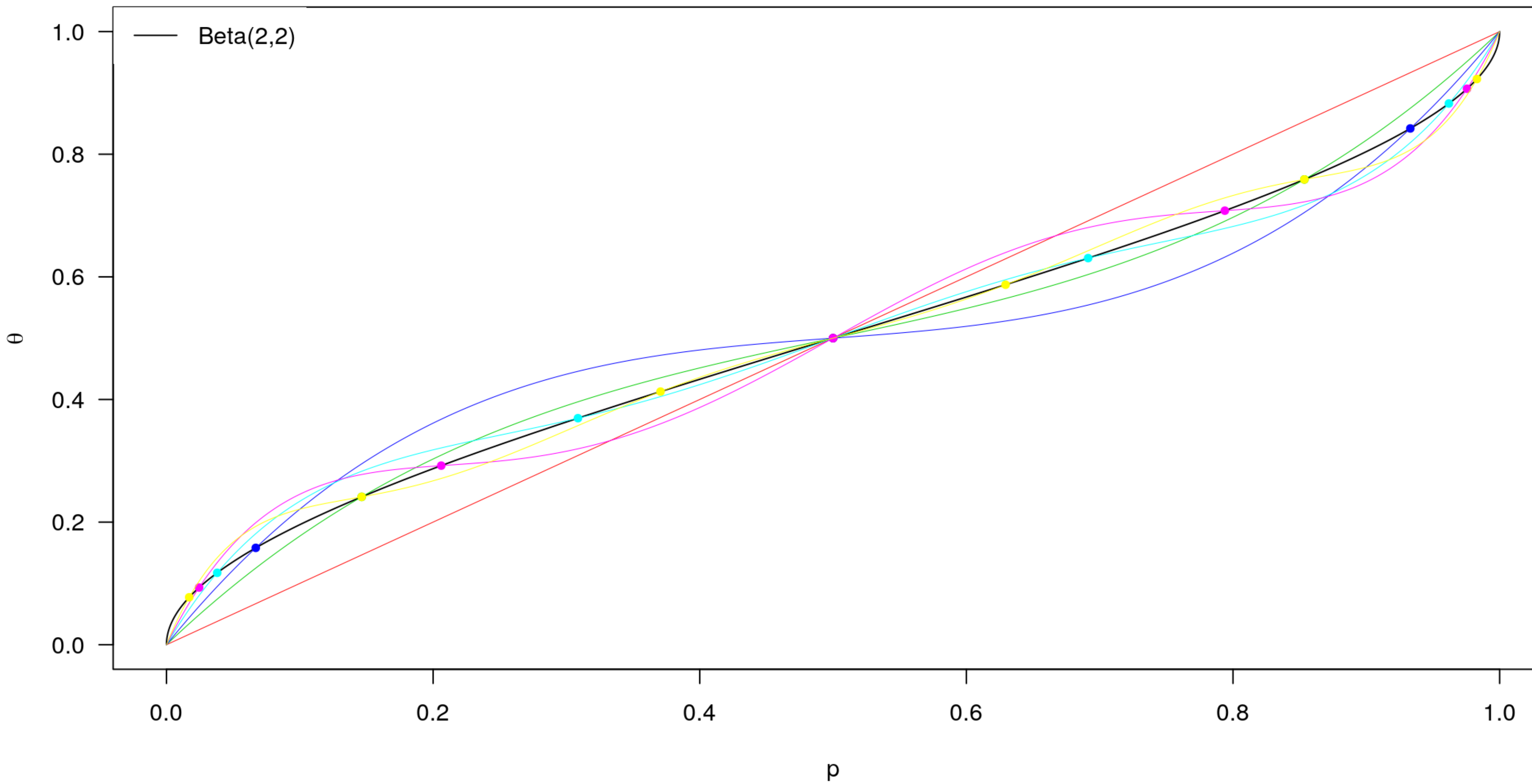
# Approximation with 4 interior and 2 end points



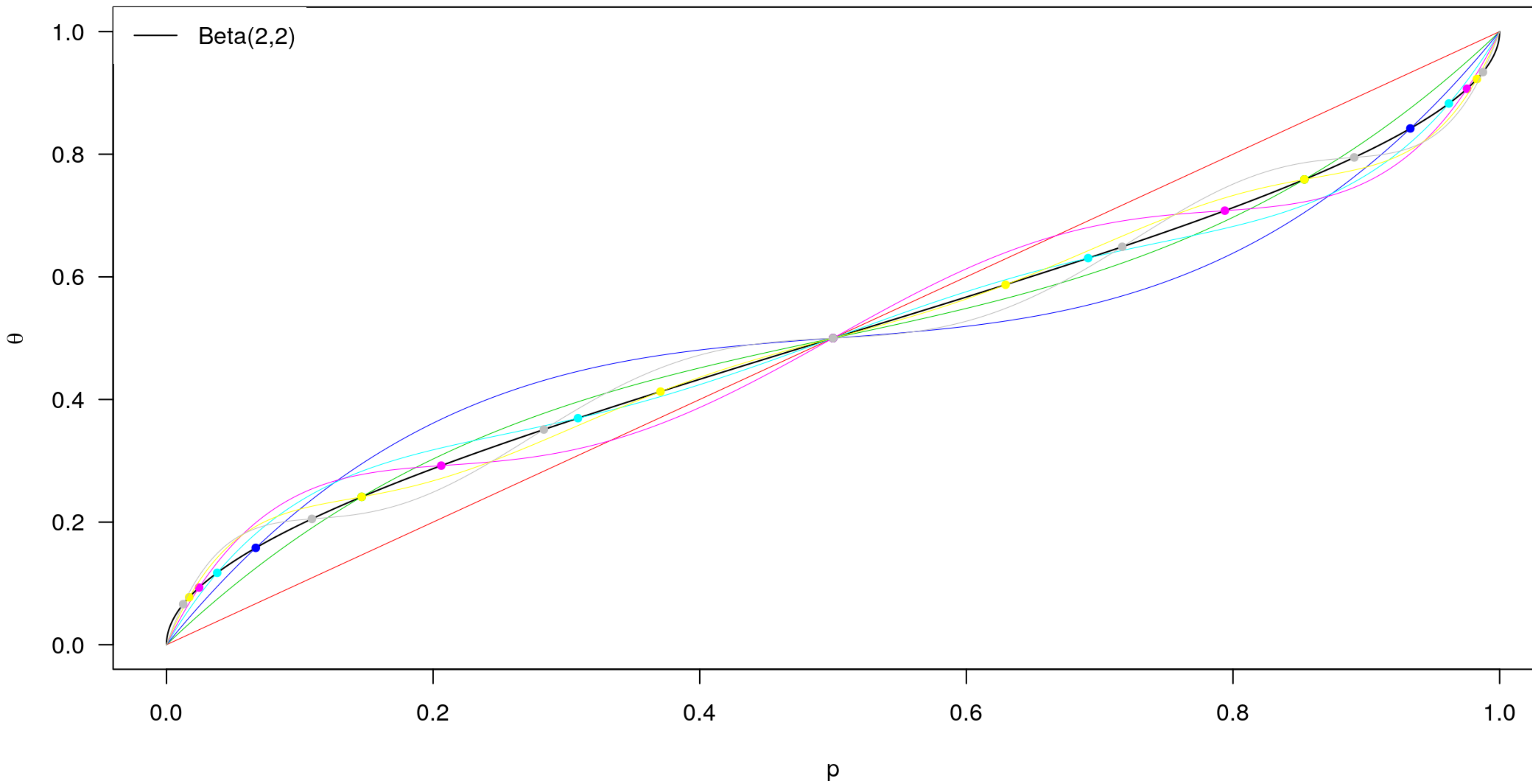
# Approximation with 5 interior and 2 end points



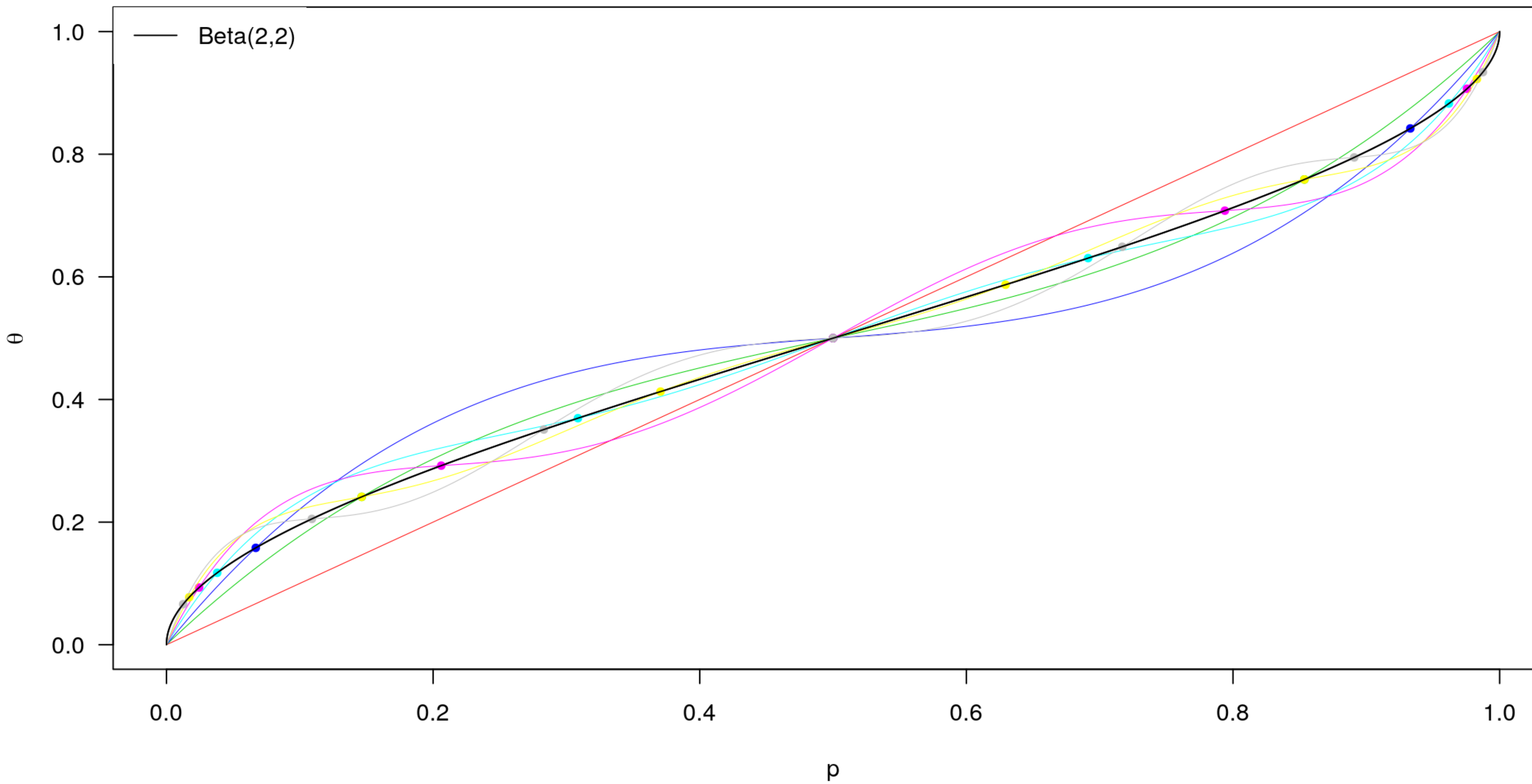
# Approximation with 6 interior and 2 end points



# Approximation with 7 interior and 2 end points



# Approximation with 7 interior and 2 end points



# The No Name Distribution of the 1st Kind

The no name distribution of the first kind has ICDF (provided it is increasing)

$$\theta\left(p; \mathbf{c}\right) \equiv \sum_{k=0}^{K} c_k T_k\left(2p - 1\right)$$

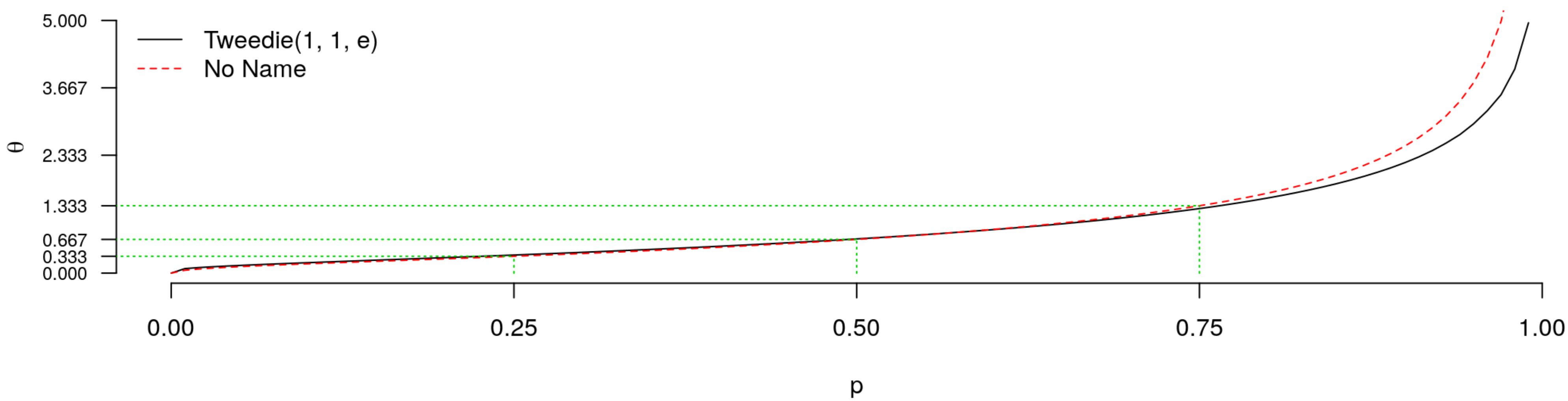
with $\mathbf{c}$ such that $\theta\left(p; \mathbf{c}\right)$ runs through the $K + 1$ quantiles the user provides

# Tweedie($\phi = 1, \mu = 1, \xi = e$) Example

- Tweedie distribution is defined over $\Theta \in [0, \infty)$ but does not have an explicit PDF, CDF, ICDF, or anything else. Nevertheless, it satisfies $\text{Var}\left(\theta\right) = \phi\mu^{\xi}$.

$$\theta\left(p; \mathbf{c}\right) \equiv e^{\tanh^{-1}\sum_{k=0}^{K} c_k T_k(2p-1)} \iff \tanh\log\theta\left(p; \mathbf{c}\right) \equiv \sum_{k=0}^{K} c_k T_k\left(2p-1\right)$$

```
q <- qno_name1(quantiles = c(0, 1 / 3, 2 / 3, 4 / 3, Inf), u = c(0, 0.25, 0.5, 0.75, 1))
```



# Standard Stable($\alpha = 1.9, \beta = 0.5$) Example

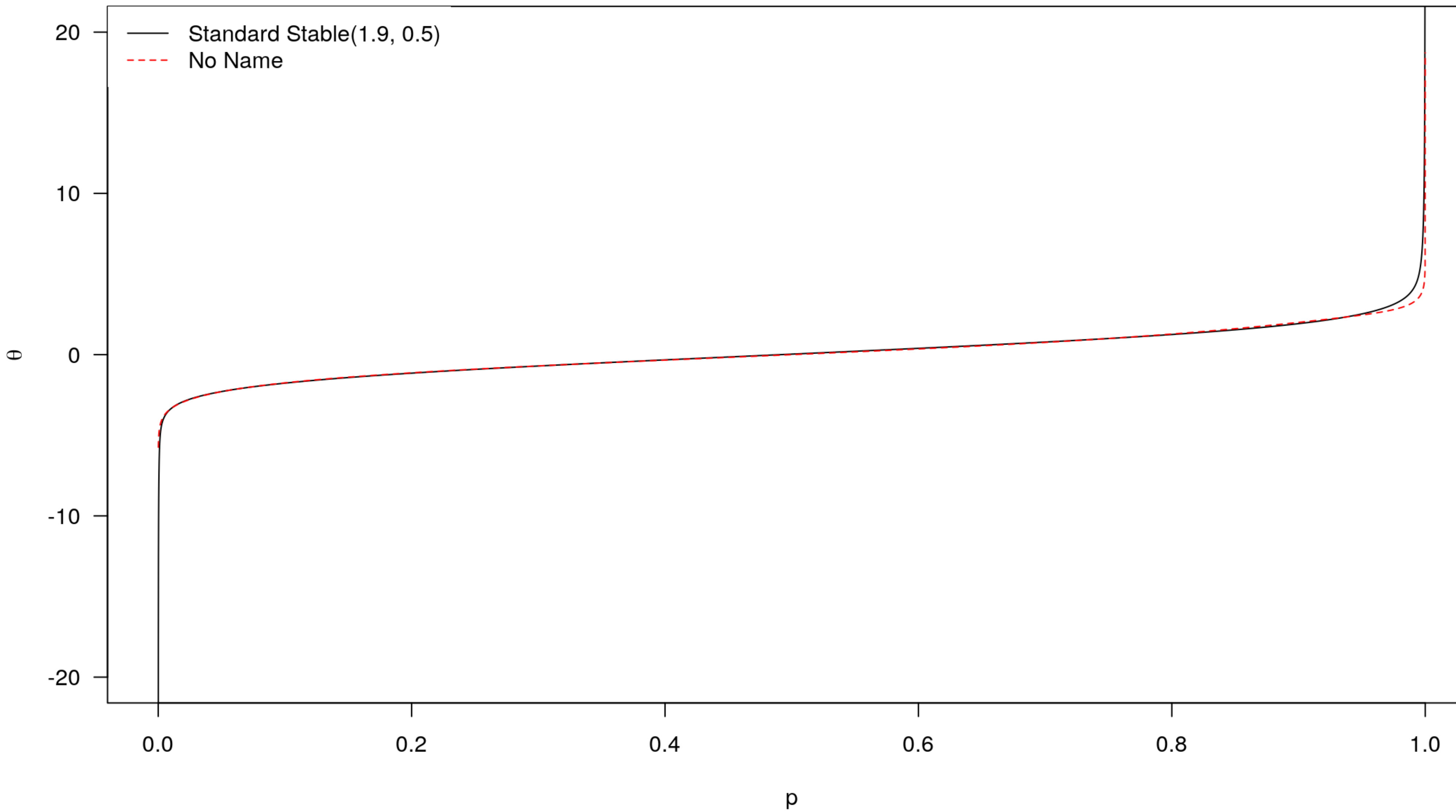
- Stable distribution is generically defined over $\Theta = \mathbb{R}$ but does not have an explicit PDF, CDF, or ICDF. It does have an elementary characteristic function.

$$\theta\left(p;\mathbf{c}\right) \equiv \tanh^{-1} \sum_{k=0}^{K} c_k T_k\left(2p-1\right) \iff \tanh\theta\left(p;\mathbf{c}\right) \equiv \sum_{k=0}^{K} c_k T_k\left(2p-1\right)$$

```
q <- qno_name1(quantiles = c(-Inf, -0.9, 0, 1, Inf), u = c(0, 0.25, 0.5, 0.75, 1))
```

```
## Error in boyd(c):
## Implied quantile function is decreasing near 0.0063, 0.9447.
## Try increasing the number of quantiles and / or changing their values.
```

```
q <- qno_name1(quantiles = c(-Inf, -1.75, -0.9, 0, 1, 2, Inf),
               u = c(0, 0.1, 0.25, 0.5, 0.75, 0.9, 1))
```



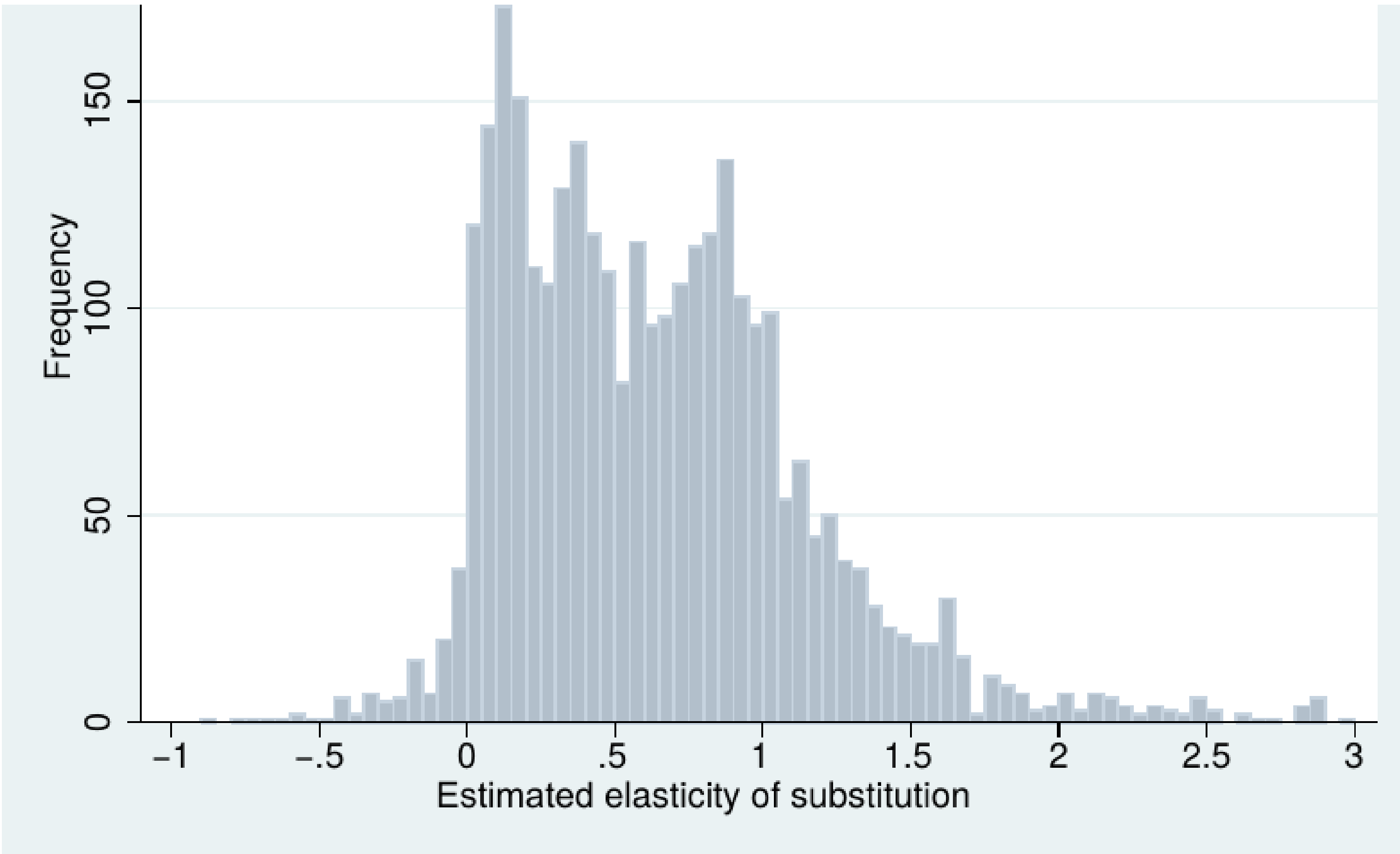
Standard Stable(1.9, 0.5)
No Name
20
10
0
-10
-20
θ
0.0
0.2
0.4
0.6
0.8
1.0
p

# Constant Elasticity of Substitution (CES) Models

$$Y_t \approx \gamma e^{\lambda(t-1)} \left( \delta \left( \delta_1 K_t^{-\rho_1} + (1-\delta_1) E_t^{-\rho_1} \right)^{\frac{\rho}{\rho_1}} + (1-\delta) L_t^{-\rho} \right)^{-\frac{\nu}{\rho}}$$

- $Y_t$ is value added, $K_t$ is capital, $E_t$ is energy, and $L_t$ is labor
- $\rho = \frac{1}{\sigma} - 1$ and $\rho_1 = \frac{1}{\sigma_1} - 1$ where $\sigma > 0$ is the elasticity of substitution between labor and both capital and energy (the quantity of interest), while $\sigma_1 > 0$ is the elasticity of substitution between capital and energy
- $\gamma, \lambda > 0, \delta \in (0,1), \delta_1 \in (0,1)$ and $\nu > 0$ are not that important today
- Take logarithms and assume Gaussian error with standard deviation $\omega > 0$. Informative priors on the parameters are essential to avoid divergences.

# Point Estimates of $\sigma_1$ from Gechert et al. (2019)

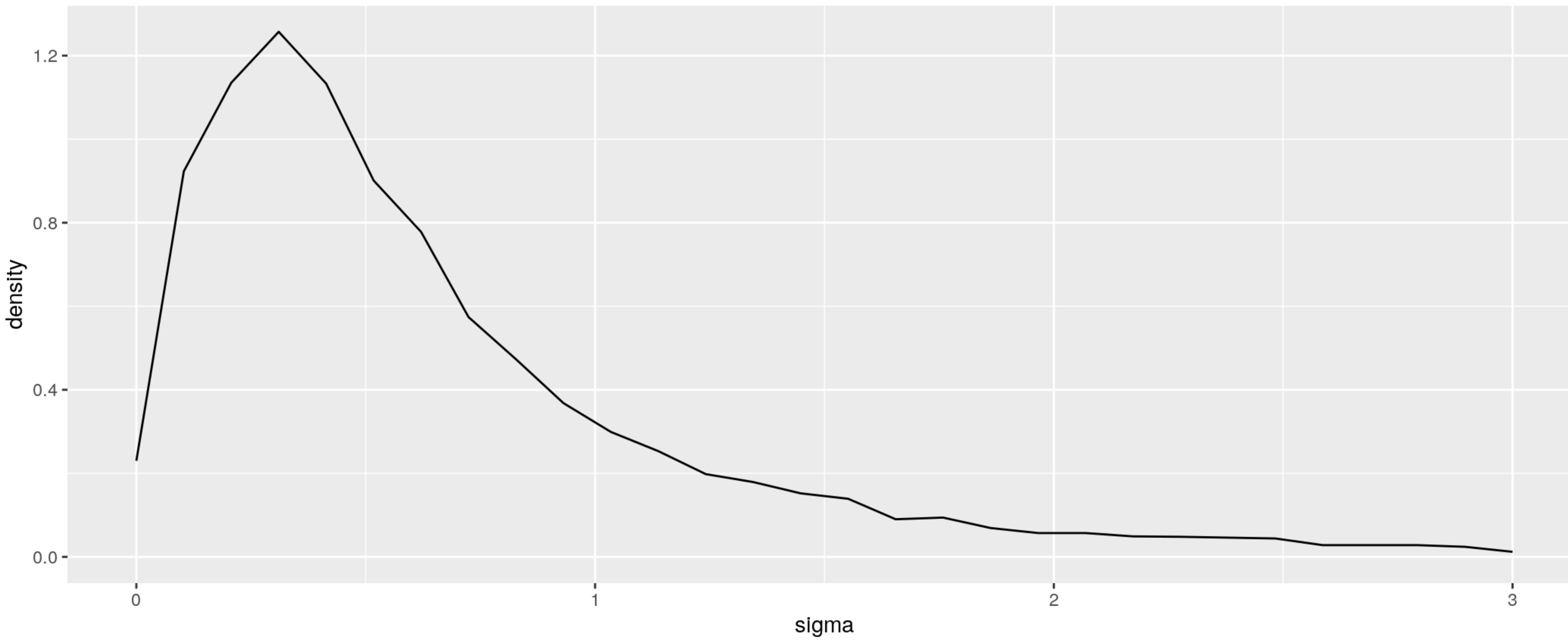



*Notes:* Estimates smaller than −1 and larger than 3 are excluded from the figure for ease of exposition but included in all statistical tests.

elasticitiy estimates

# Prior Quantile Function for $\sigma$ and $\sigma_1$

```
q <- qno_name1(quantiles = c(0, 0.27, 0.5, 0.9, Inf), u = c(0, 0.25, 0.5, 0.75, 1))
ggplot(data.frame(sigma = q(runif(9999)))) + geom_freqpoly(aes(x = sigma, after_stat(density))
```



# Stan Program for a CES Model

```
// defines no_name1_icdf(p, u, theta)
#include no_name1.stan
data {
  int<lower = 0> T; // if T == 0, this draws from priors
  vector[T] log_Y;
  vector[T] log_K;
  vector[T] log_E;
  vector[T] log_L;
  positive_ordered[5] u;    ordered[5] theta[7];
  positive_ordered[6] u_lg; ordered[6] theta_lg;
}
parameters {
  vector<lower = 0, upper = 1>[8] p;
}  // cumulative probability primitives
transformed parameters {
  real sigma   = no_name1_icdf(p[1], u, theta[1]);
  real sigma_1 = no_name1_icdf(p[2], u, theta[2]);
  real delta   = no_name1_icdf(p[3], u, theta[3]);
  real delta_1 = no_name1_icdf(p[4], u, theta[4]);
  real nu      = no_name1_icdf(p[5], u, theta[5]);
  real omega   = no_name1_icdf(p[6], u, theta[6]);
  real lambda  = no_name1_icdf(p[7], u, theta[7]);
  real log_gamma = no_name1_icdf(p[8], u_lg, theta_lg);
}
```

```
model {
  real rho   = -1 + inv(sigma);
  real rho_1 = -1 + inv(sigma_1);
  real nu_rho = nu / rho;
  real log_delta = log(delta);
  real log_delta_1 = log(delta_1);
  real log1m_delta = log1m(delta);
  real log1m_delta_1 = log1m(delta_1);
  real rho_rho_1 = rho / rho_1;
  vector[T] mu;
  for (t in 1:T) // with numerical stability
    mu[t] = log_gamma
          + lambda * (t - 1)
          - nu_rho
          * log_sum_exp(log_delta + rho_rho_1
          * log_sum_exp(log_delta_1 -
                        rho_1 * log_K[t],
                        log1m_delta_1 -
                        rho_1 * log_E[t]),
                        log1m_delta -
                        rho * log_L[t]);
  log_Y ~ normal(mu, omega); // log-likelihood
} // MLEs invariant to the ICDF transformations
```

# Maximum Likelihood Estimates of a CES Model

```
data(GermanIndustry, package = "micEconCES")
GermanIndustry <- log(subset(GermanIndustry, year < 1973 | year > 1975)[ , 2:5])
colnames(GermanIndustry) <- paste0("log_", c('Y', 'K', 'L', 'E'))
dat <- c(list(T = nrow(GermanIndustry), u_lg = c(0, 0.25, 0.5, 0.75, 0.9, 1),
              theta_lg = c(-2, 1, 3, 5, 7, 10), u = c(0, 0.25, 0.5, 0.75, 1),
              theta = list(sigma   = c(0, 0.27, 0.5, 0.9, Inf),
                           sigma_1 = c(0, 0.27, 0.5, 0.9, Inf),
                           delta   = c(0, 1/3, 0.5, 2/3, 1),
                           delta_1 = c(0, 1/3, 0.5, 2/3, 1), nu = c(0, 0.6, 1.0, 1.4, Inf),
                           omega   = c(0, 0.016, 0.03, 0.05, 0.15),
                           lambda  = c(0, 0.01, 0.02, 0.03, 0.05))), GermanIndustry)

MLEs <- optimizing(CES, data = dat, as_vector = FALSE, refresh = 0, seed = 54321)
round(rbind(theta = unlist(MLEs$par[-1]), p = MLEs$par$p), digits = 3) # delta on boundary
```

```
##       sigma sigma_1 delta delta_1    nu omega lambda log_gamma
## theta 0.174   0.153 0.999   0.799 0.583 0.030  0.014     4.261
## p     0.143   0.122 0.999   0.872 0.240 0.489  0.347     0.659
```

# Posterior Estimates for a CES Model

```
post <- sampling(CES, data = dat, seed = 12345,
                 control = list(adapt_delta = 0.96, max_treedepth = 12), refresh = 0)
print(post, pars = "p", include = FALSE, probs = c(.025, .1, .25, .5, 0.75, .9, .975))
```

```
...
##             mean se_mean   sd  2.5%   10%   25%   50%   75%   90% 97.5% n_eff Rhat
## sigma       3.45    0.15 5.09  0.83  1.12  1.54  2.21  3.56  5.97 14.74  1205    1
## sigma_1     1.31    0.05 1.69  0.36  0.44  0.57  0.82  1.37  2.57  5.06  1281    1
## delta       0.72    0.01 0.20  0.32  0.44  0.56  0.74  0.89  0.97  0.99  1505    1
## delta_1     0.05    0.00 0.06  0.00  0.00  0.01  0.03  0.07  0.12  0.22  1785    1
## nu          0.90    0.00 0.08  0.75  0.80  0.85  0.89  0.94  1.00  1.06  1536    1
## omega       0.02    0.00 0.00  0.02  0.02  0.02  0.02  0.03  0.03  0.03  1840    1
## lambda      0.02    0.00 0.00  0.02  0.02  0.02  0.02  0.02  0.02  0.02  1173    1
## log_gamma   1.13    0.01 0.46  0.27  0.57  0.82  1.10  1.43  1.71  2.06  1373    1
## lp__       83.83    0.08 2.41 78.40 80.54 82.43 84.15 85.59 86.68 87.61   957    1
...
```

```
## Maximized likelihood is 93.64359
```

# Conclusions

- Your audience is unlikely to be equipped to understand prior PDFs
- Quantiles rather than expectations are an easier entry point
- Avoid prior PDFs by utilizing the logic of RNGs that apply an ICDF to a standard uniform random variate to obtain a random variate from the intended distribution
- We need to get ICDFs into Stan (many of them are in Boost)
- Construct a prior ICDF rather than choosing one from list

# References


- Chalabi, Y., 2012, *New Directions in Statistical Distributions, Parametric Modeling and Portfolio Selection*, Dissertation, ETH Zurich. Link
- Gechert, S. et al., 2019 “Death to the Cobb-Douglas Production Function? A Quantitative Survey of the Capital-Labor Substitution Elasticity”, ZBW – Leibniz Information Centre for Economics, Kiel, Hamburg Link
- Gil, A., Segura, J., and Temme, N., 2007, *Numerical Methods for Special Functions*, Society for Industrial and Applied Mathematics. Chapter 3
- Gilchrist, W., 2000, *Statistical Modelling with Quantile Functions*, CRC Press. Link
- Hadlock, C., 2017, *Quantile-Parameterized Methods for Quantifying Uncertainty in Decision Analysis*, Dissertation, University of Texas at Austin. Link
- Henningsen, A. and Henningsen G., 2014 “Econometric Estimation of the ‘Constant Elasticity of Substitution’ Function in R: Package micEconCES”, Vignette
- Keelin, T. and Powley B., 2011, “Quantile Parameterized Distributions”, *Decision Analysis*, 8(3) 206 – 2019. Link
- Trefethen, L., 2013, *Approximation Theory and Approximation Practice*, Society for Industrial and Applied Mathematics. Website