

Arbetslogg 2016 vecka 9

Erik Bulow

29 februari 2016

Contents

1	Förberedelser	1
2	2016-20-29	2
2.1	Läsning av (Pearson 1895)	2
2.2	Läsning av (Pearson 1896)	2
2.3	Läsning av (Nemes et al. 2009)	3
2.4	Läsning av (Cowden 1952)	3
2.5	Läsning av (Kymn 1968)	4
2.6	Undersöker icke-central betafördelning	4
3	2016-03-01	6
3.1	Diskussion med SN	8
3.2	Läsning av (???)	14
4	2016-03-02	27
4.1	Testar att modifera beta-funktionerna för att också skatta ncp (misslyckas)	28
4.2	Jämför resultat för olika fördelningar	28
4.3	För ökande p	42
4.4	Läsning av (???)	62
	Referenser	66

1 Förrberedelser

```
# Try it out!
memory.limit(50000)
```

```
## [1] 50000
```

```
options(samplemetric.log = TRUE)
set.seed(123)
```

2 2016-20-29

Har udner helgen roat mig med att läsa Sigma och bl a den först kända publicerade artikeln om statistik (från 1600-talet). Intressant kuriositet även om det kanske inte har ngn direkt nytt för jobbet just nu.

2.1 Läsnings av (Pearson 1895)

(Läste egentligen denna förra veckan.)

Är lite okklar över referenserna. Tror jag hittade referens som pekade på denna artikel som upphov till korrelationskoefficienten (men hittade också annan referens som ist pekade på (Pearson 1895)). Men den nämns i denna artikel också och då denna publicerades före den senare så bör det kanske vara sant. Men här hänvisas också till Galtons formel så egentligen var det inte helt nytt.

Välldigt kort note som egentligen är del av längre paper som inte han färdigställas pga hälsoproblem.

2.2 Läsnings av (Pearson 1896)

Sägs vara källan till korrelationskoefficienten. Framkommer dock att konceptet de facto var känt sedan tidigare. Texten utgår från ganska praktiska exempel med heriditet och sexuell reproduktion etc. Ger äran till Bravais och delvis också till Edgeworth, Galton och Weldon. Innehåller rtedan här en del teori om fördelning baserat på χ^2 etc. Refereras också till r som Galtons funktion (Galton myntade f.ö. även uttrycket regression). Redan då användes normalapproximation. Texten utgår från ganska teoretiska beräkningar men konstateras att den praktiska formeln för att beräkna r är den Bravais föreslagit men utan att ha visat att det verkligen var den bästa formeln. Se s 265 för formeln.

Konstaterar redan här att:

Thus we may say that with sufficient accuracy for most cases the standard deviation of a coefficient of correlation is:

$$\frac{1 - r^2}{\sqrt{n(1 + r^2)}}$$

or its probable error =

$$.674506 \frac{1 - r^2}{\sqrt{n(1 + r^2)}}$$

[...] It will be sufficient therefore, for most practical purposes to assume that the probable error of a coefficient of correlation

$$= .674506 \frac{1 - r^2}{\sqrt{n(1 + r^2)}}$$

Här talas dock också om ganska stora stickprovsstorlekar såsom $n = 1000$. Skriver om ett dataset med 200 samlpes att:

The number is not sufficiently great to make the probable error of quite small enough dimensions in several cases, and so allow of definite conclusions.

(F.ö. ett sample baserat på övre medelklass så kudne därav inte heller nyttjas för generella slutsatser om populationen. Återkänner även på s 273 till att vi inte kan anta normalfördelning här. Refererar till att normalfördelning kunde antas vid studie av 900 kraniemätningar utförda vid tidigare studie.)

f.ö. undersöks i artikeln relationen mellan föräldrars längd och kön på avkomma. Konstateras (med viss reservation) att t ex längre fäder tenderar få döttrar i ngt högre utsträckning än söner. Dock svårare att se mönster för mödrar. Ser även att korrelation för längder tycks ärvas starkare på fädernet än mödernet även sett över flera generationsled.

F.ö. intressant att artikeln blandar både teori men också ganska utförliga praktiska beskrivningar. Känns både konkret och väl underbyggt på samma gång.

Gör inget fördelningsantagande för data vi samplar ifrån.

Noterar f.ö. att han nämner korrelation och standradavvikelse etc men gör inga referenser till kovarians.

Behandlar också fallet med tre grupper att jämföra och därmed tre parvisa korrelationer.

Görs också studier av korrelation av ansiktsbehåring, dvs ärftheten av detta. Även referenser till att färgblindhet ärvs från morfar till dotterson.

Behandlar även fall med 4 korrelationer. Denna teknik tycks användas där man idag istället skulle använda regression i modern mening.

Efter ett par generationer kommer familjära särdrag suddas ut varpå släkten alltmer liknar populationen. Detta gäller även vid selektive breeding. Skulle behövas experiment för att empiriskt utröna effekten av selektive breeding etc! :-)

2.3 Läsning av (Nemes et al. 2009)

Tipsad av denna av SN. Inte för att ämnet i sig är direkt relevant men då upplägget på själva artikeln kan antas liknande nu föreliggande förutsättningar. Konstaterar bias för mindre stickprovsstorlekar. Även här ses approximativt normalfördelning av oddskvoten för stora n. Finns även här en skev bakomliggande fördelning. Även här är problemen kända teoretiskt men inte bland praktiker. Finns också förslag på bias-korrigeringar.

Ger ingen rekommendation om sample size men konstaterar att andra föreslagit minst 100 och helst mer än 500. Diskret data kräver större smaple, liksom starkt korrelerade data.

Biasen påverkar på så sätt att små stickprovsstorlekar påvisar större effekt än för större samples.

Beskriver risken med detta att man publicerar material som inte stämmer med verkligheten. Även problem vid metastudier då man inte tänker på detta då flera studier jämförs.

På det hela taget mkt intressant och viktigt!

2.4 Läsning av (Cowden 1952)

Handlar om multipel-partial correlation coefficient. Förklarar att "multiple correlation coefficient" är vår koefficient där observerad vs predicted values korreleras och där pred beror på en eller flera variabler. Artikeln inför också "multiple-partial correlation coefficient", en justerad correlation mellan utfall samt två eller fler oberoende variabler.

Innehåller mkt härledningar och teori. Känns dock inte helt relevant i sammanhangert så lämnar den ej färdigläst.

2.5 Läsning av (Kymn 1968)

Det är känt sedan tidigare att:

$$F = r^2 \frac{n-2}{1-r^2} \sim F_{1,n-2}$$

samt

$$t = r \frac{\sqrt{n-2}}{1-r^2} \sim t_{n-2}$$

Denna artikel visar nu att

$$S = \frac{1+r}{1-r} \sim F_{n-2,n-2}$$

Fördelen med denna är att fördelningen är symmetrisk samt ev att S inte beror på n (men det gör ju å andra sidan F så jag vet inte riktigt varför det skulle vara så stor skillnad).

OBS! Bygger på att x, y är bivariat normalfördelade och oberoende $\rho = 0$ så nyttan av detta kanske är begränsad?

Noterar här att enligt (Hotelling 1953) krävs dock inte bivariat normalfördelning just då $\rho = 0$

2.6 Undersöker icke-central betafördelning

Tar en avstickare och försöker skapa funktion för icke-central betafördelning. Noterar att:

- x ska antas fix och har därmed ingen känd fördelning
-

```
#' Parameters for the noncentral beta distribution of R2
#'
#' @param ncp1 first part of the con centrality parameter
#' as given by \code{\link{ncp1}}
#' @param x object of class \code{\link{sim_data}}
#' @return List with "shape1", "shape2" and "ncp" parameters
#' as used for corresponding arguments in the \code{\link{Beta}}
#' functions.
#' @export
r2_beta_param <- function(ncp1, x) {
  stopifnot(ncol(x) == 2)
  list(
    shape1 = .5,
    shape2 = (nrow(x) - 2) / 2,
    ncp = ncp1 * sum((x$X1 - mean(x$X1)) ^ 2)
  )
}

#' Calculate the first half of the non centrally parameter of R2
#'
```

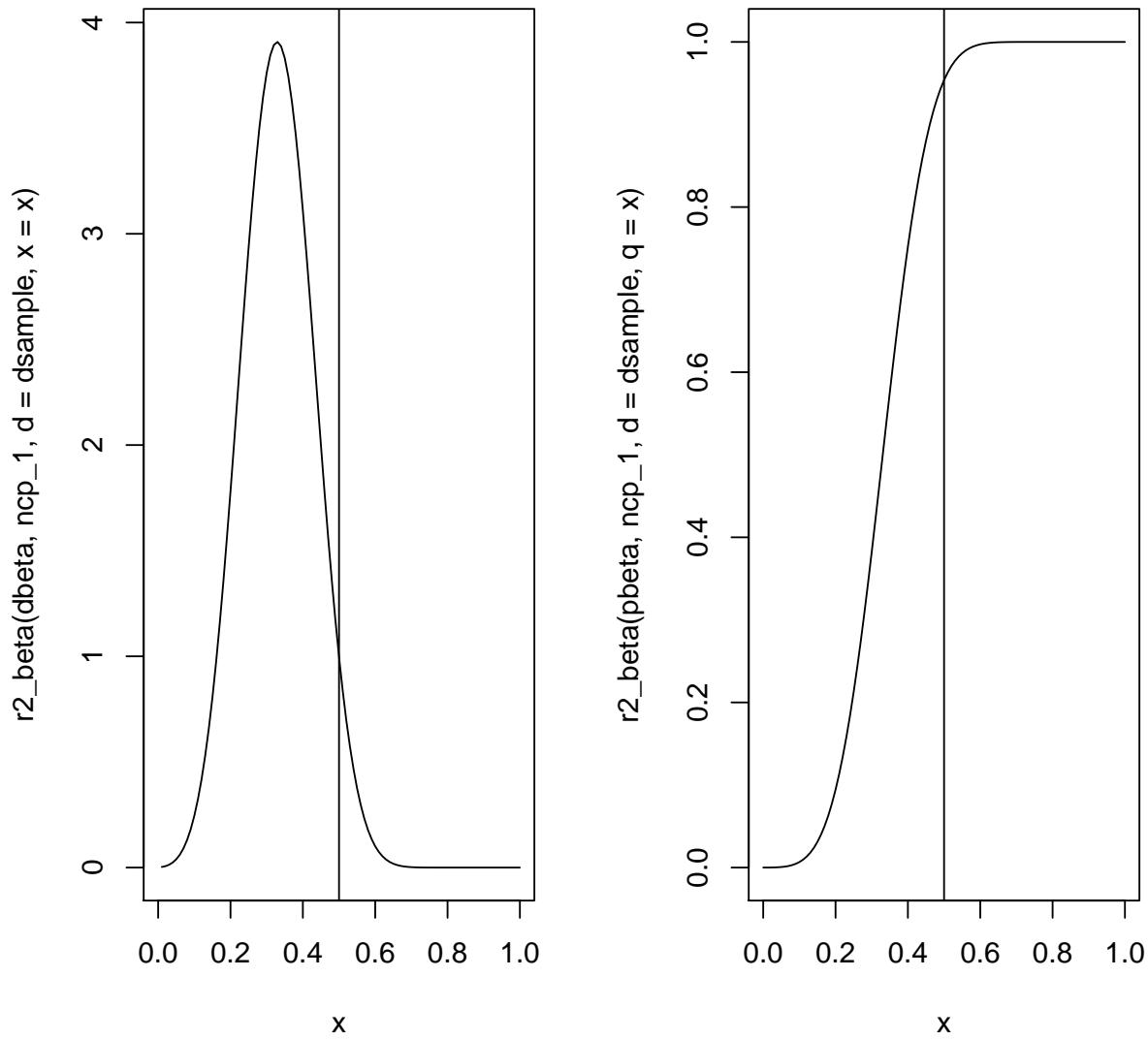
```

#' Calculate the non obsevral dependent part of the
#' centrality parameter used as argument
#' "ncp" in the \code{\link{Beta}} family of functions
#'
#' @param x object of class \code{\link{sim_data}}
#' @return numeric vector of length one
#' @export
#' @examples
#' ncp1 <- ncp1(sim_data())
ncp1 <- function(x) {
  stopifnot(ncol(x) == 2)
  fit    <- lm(Y ~ ., data = x)
  beta   <- fit$coefficients[2]
  sigma2 <- var(fit$residuals)
  (beta ^ 2) / (2 * sigma2)
}

#' The R2 disrtibution based on the Beta distribution
#'
#' @param fun one of the functions listed at \code{\link{Beta}}
#' @param ncp1 value given by \code{\link{ncp1}}
#' @param d object of class \code{\link{sim_data}} with columns
#' \code{Y} and \code{X1}.
#' @param ... arguments passed to \code{fun}
#' @return Value returned by call to \code{fun}
r2_beta <- function(fun, ncp1, d, ...) {
  do.call(fun, c(r2_beta_param(ncp1, d), list(...)))
}

d <- sim_data(r2 = .5, p = 1)
dsample <- dplyr::sample_n(d, 50)
ncp_1 <- ncp1(d)
# r2_beta(dbeta, ncp_1, d = dsample, x = seq(0.01, 1, .01))
par(mfrow = c(1, 2))
curve(r2_beta(dbeta, ncp_1, d = dsample, x = x))
abline(v = .5)
curve(r2_beta(pbeta, ncp_1, d = dsample, q = x))
abline(v = .5)

```



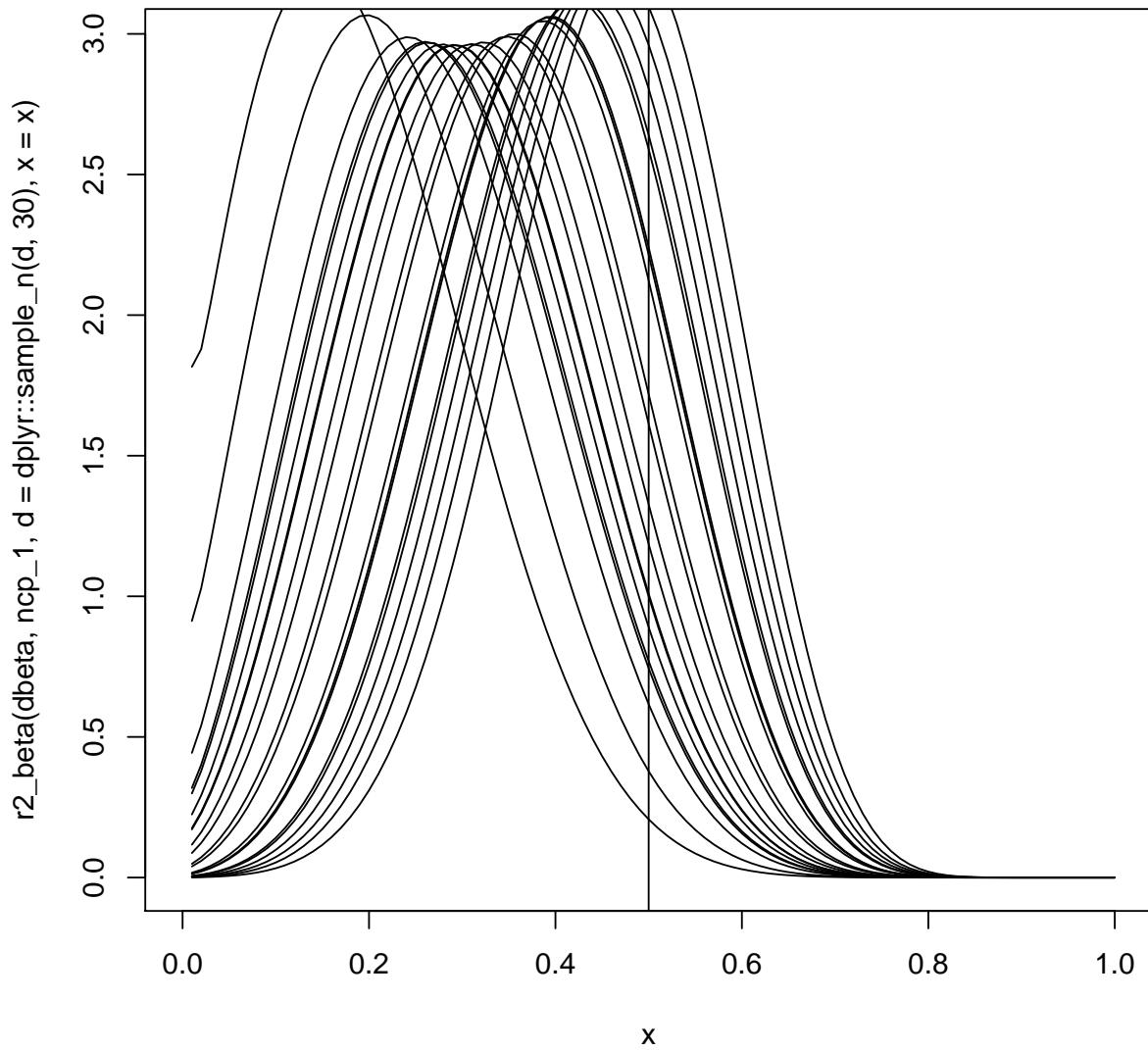
Är detta enligt förväntan? Ser ut som vi underskattar r^2 väldigt grovt ...?

3 2016-03-01

Fortsätter titta på simuleringarna ovan. Gör om några ggr och finner att det nog bara var slump att det blve så biased. Beöhver simulerar flera ggr men lite osäker på hur. Det bli rju olika fördelningar varje gång. Ska jag beräkna medelvärden för ncp eller på ngt sätt för hela fördelningen?

```
par(mfrow = c(1,1))
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x)); abline(v = .5)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
```

```
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
curve(r2_beta(dbeta, ncp_1, d = dplyr::sample_n(d, 30), x = x), add = TRUE)
```



3.1 Diskussion med SN

Är ovanligt och lite konstigt att fördelningen i detta fall beror på observerade data. För t ex t- och F-fördelning finns ju ett beroende av frihetsgrad (stickprovsstorlek) men inte av själva datapunkterna. Att ha ett sådant beroende känns lite märkligt då man på ngt sätt rättfärdigar ett observerat resultat genom teorier byggda på samma observerade resultat, vilket känns som ett cirkelresonameng. Å andra sidan är väl just detta anledningen till att x enl teorin heller inte är att betrakta som en slumpvariabel utan som fix.

Slutsatser:

1. Det går inte att finna en enda teoretisk fördelning (facit) då den alltid kommer att bero på slumpen.
2. Vad vi kan göra är att koncentrera oss på t ex olika moment av fördelningen. Vi kan t ex ta ett stickprov och för detta beräkna både den semiteoretiska fördelning detta ger upphov till, samt skatta R2 direkt.

Vi jämför sedan skattningen mot värdet givet av väntevärdet givet av fördelningen. Vi upprepar många ggr och plottar dessa värden med qqplot för att undersöka ev bias.

3. Kan också vara av värde att undersöka ifall det finns metod att skatta betafördelningens parametrar utifrån data på ngt mer generellt sätt.

Vi försöker göra enl (2) ovan. Dock behöver vi för detta jämföra skattningen inte mot mean utan mot mode för att få det korrekt. Dock svårt att finna ngn formel för mode av icke cenrtal betafördelning. Finns funktion `modeest::betaMode` men den funktionen hanterar ändå inte detta.

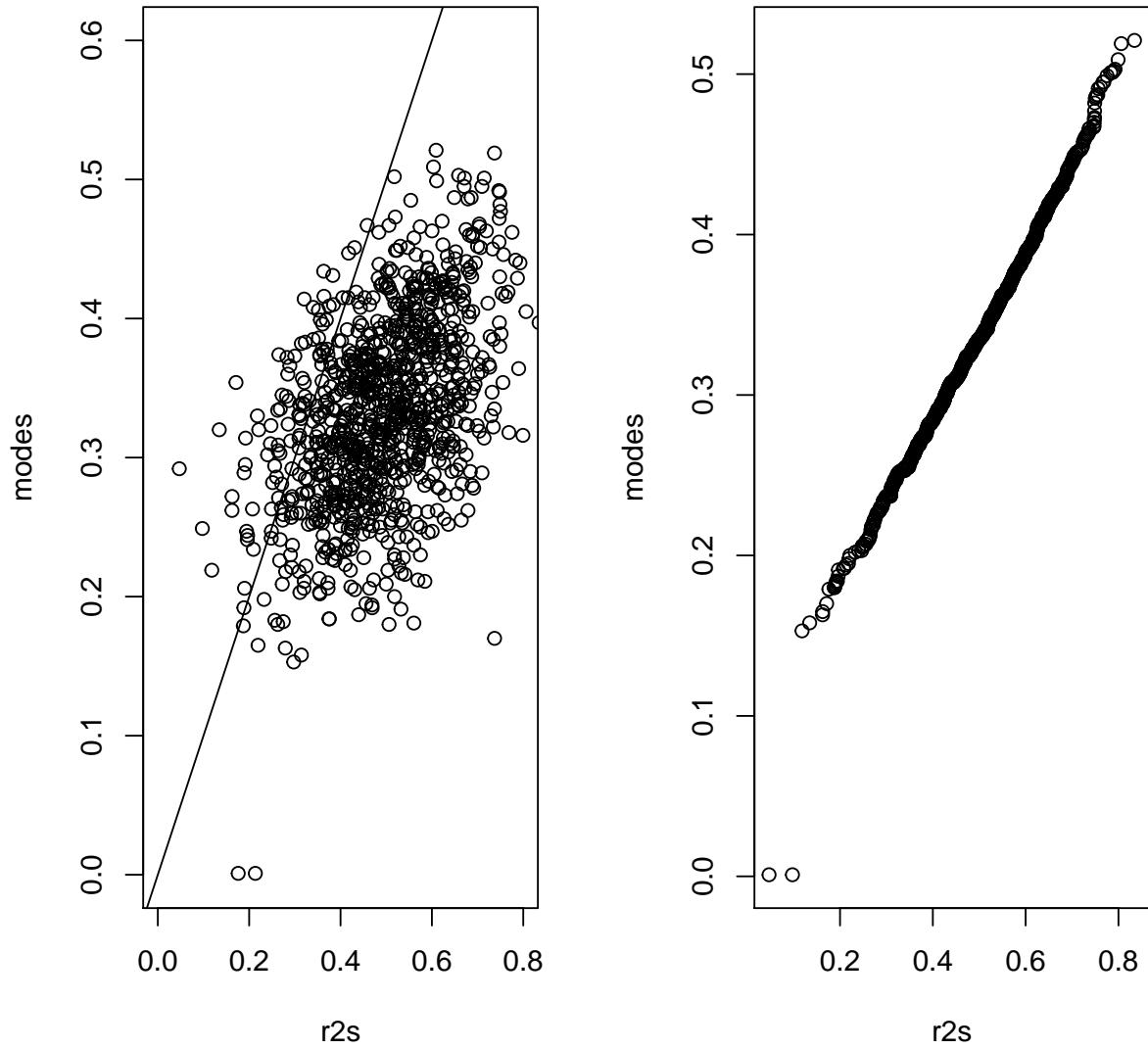
Enl (Park 1964) krävs numerisk approximation för skattning av mode för icke-central beta. Formel presenteras i (3.2) men bygger på antaganden (såsom att $r^2 \rightarrow 1$), vilket gör det hela ointeressant. Ett alternativ blir då att skatta ett numeriskt värde (såsom vi också gjort vid tidigare simulerings), vilket vi lätt kan göra om vi antar att fördelningen är unimodal, vilket vi här kan. Observera dock att vi här inte ska basera mode-skattningen på vårt slumpmässiga urval utan på fördelningens värde för $\forall x : x \in [0, 1]$ för relevant fördelning. Kanske skulle man också kunna undersöka metoder för att finna mode via paketet `modehunt`. Jag har ännu inte fördjupat mig i det och vet således inte ifall det skulle ge annat resultat än min egen mode-funktion.

```
r2_beta_mode <- function(ncp1, d, ...) {
  x <- seq(0.001, 1, .001)
  y <- r2_beta(dbeta, ncp1 = ncp1, d = d, x = x, ...)
  x[y == max(y)]
}

# Prepare data sets
d <- sim_data(r2 = .5, p = 1)
ss <- subsamples(d, n.max = 30, N = 1000)
ncp_1 <- ncp1(d)

# Calculate "theoretical modes" and "observed r2"
modes <- vapply(ss, function(d) r2_beta_mode(ncp_1, d), numeric(1))
r2s <- metrics(ss, n.sample = 30)$Rsquared

# Plot and compare
par(mfrow = c(1, 2))
plot(r2s, modes, xlim = c(0, .8), ylim = c(0, .6))
abline(0, 1)
qqplot(r2s, modes)
```



Vi ser här att vår teoretiska mode underskattar vårt observerade värde. Är det trots allt så att vi inte bör jämföra mot mode utan mot mean? Det finns visserligen en teoretisk formel för att beräkna mean av icke central betafordelning men den behöver i sin tur en confluent hypergeometric function. Det finns dock fler versioner av detta och de som finns implementerade i R tycks inte motsvara den här aktuella. Vi får därför skatta mean pss som vi tidigare skattade mode. Å andra sidan har vi också en formel för väntevärdet av r^2 för selektive samplning given av (Warren 1971) avs 2.2. Även här refereras till en confluent hypergeometric function men då denna har endast tre variabler är sannolikheten större att denna är samma som t ex `hypergeo::genhypergeo`. Dock krävs fler parametrar som jag inble blir riktigt klok på (tycks sma att man ska slumpa n värden från varje punkt tagen med selective samplning men jag får inte riktigt ihop det).

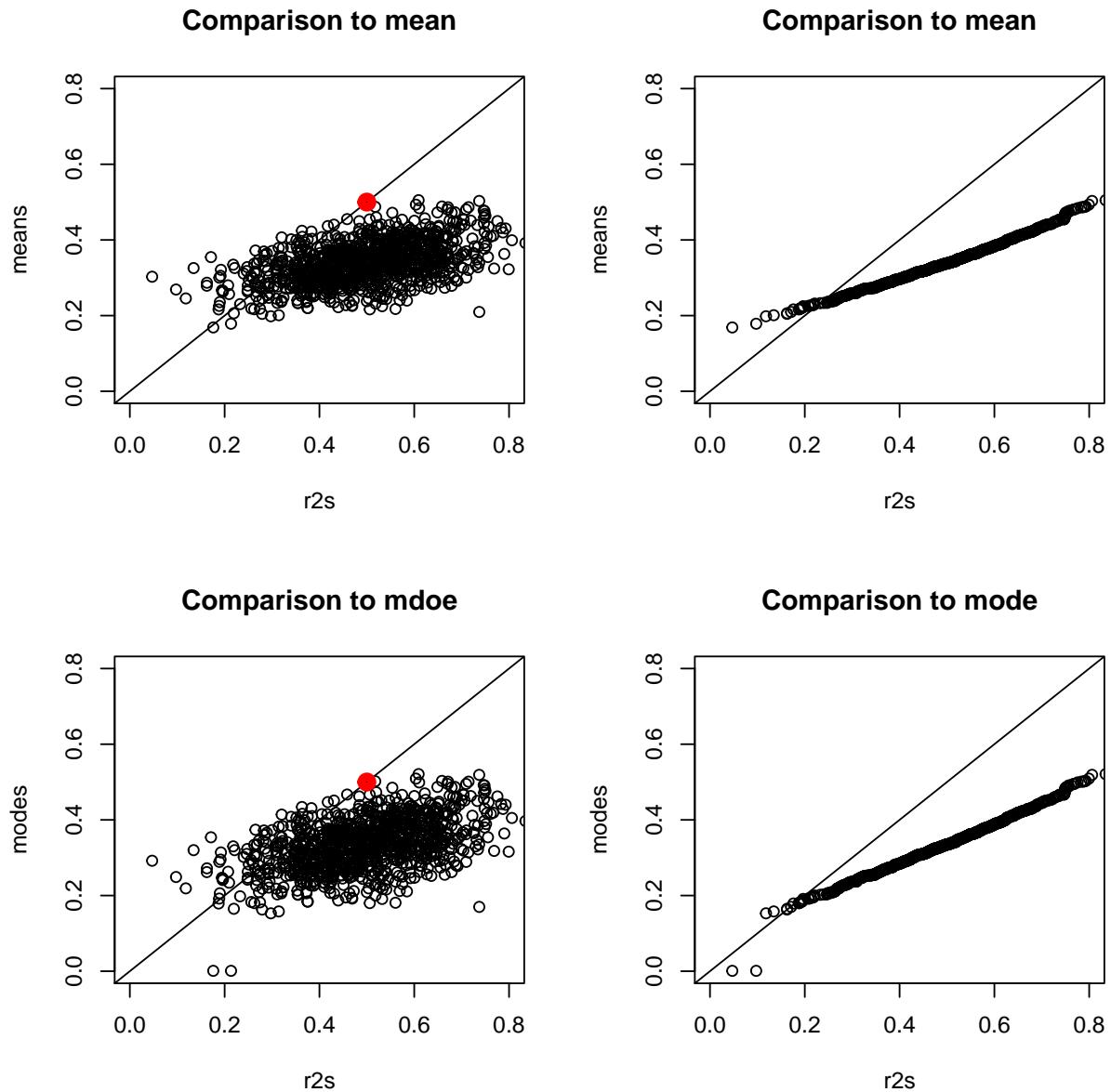
```
r2_beta_mean <- function(ncp1, d, ...) {
  x <- seq(0.001, 1, .001)
  y <- r2_beta(dbeta, ncp1 = ncp1, d = d, x = x, ...)
  sum((y / sum(y)) * x)
}
```

```

# Calculate "theoretical modes" and "observed r2"
means <- vapply(ss, function(d) r2_beta_mean(ncp_1, d), numeric(1))

# Plot and compare
par(mfrow = c(2, 2))
plot(r2s, means, xlim = c(0, .8), ylim = c(0, .8), main = "Comparison to mean")
abline(0, 1); points(.5, .5, lwd = 5, col = "red")
qqplot(r2s, means, xlim = c(0, .8), ylim = c(0, .8), main = "Comparison to mean")
abline(0, 1)
plot(r2s, modes, xlim = c(0, .8), ylim = c(0, .8), main = "Comparison to mode")
abline(0, 1); points(.5, .5, lwd = 5, col = "red")
qqplot(r2s, modes, xlim = c(0, .8), ylim = c(0, .8), main = "Comparison to mode")
abline(0, 1)

```



Röda prickar markerar ρ^2 men observera här att värden på y-axeln (mean resp mode) inte syftar till att approximera det teoretiska värdet utan värdet för r^2 , vilket vi vet underskattar ρ^2 för varje enskild observation. Om r^2 skulle följa den ickecentala betafördelningen skulle vi dock se observationer centrerade kring linjen i graferna. Det gör vi inte. Vad vi ser är istället att den teoretiska fördelningen tycks underskatta observerade r^2 systematiskt. Vi ser endast marginell skillnad mellan mode och mean (tyder väl på att den teoretiska fördelningen är mindre skev än den verkliga?) men möjligt att mode är lite bättre (vilket stämmer med teorin).

Slutsats: Den icke centrala betafördelningen enligt (Hogben 1968) underskattar r^2 .

Men för att sammanfatta så avviker jag också från teorin enl:

1. Mitt x slumpas (ej fixt). Vet dock inte riktigt hur detta bör påverka resultatet.
2. Mode och mean från fördelningen är skattade på kanske inte allra bästa sätt? Ett alternativ är kanske att nyttja fördelningen till att slumpa fram en massa värden och sedan beräkna mode och mean av det. Tycker dock inte det borde bli ngn skillnad ... men kan ju förstås testa ...

```
mode <- function(d) {
  z <- density(d)
  z$x[z$y == max(z$y)]
}

# Mode baserat på simulerings
r2_beta_mode_r <- function(ncp1, d, ...) {
  y <- r2_beta(rbeta, ncp1 = ncp1, d = d, n = 1e4, ...)
  mode(y)
}
modesr <- vapply(ss, function(d) r2_beta_mode_r(ncp_1, d), numeric(1))

# Kollar om det finns ngn skillnad mellan de två sätten
t.test(modes, modesr)
```

```
##
##  Welch Two Sample t-test
##
## data:  modes and modesr
## t = 0.010579, df = 1997.9, p-value = 0.9916
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.005944838  0.006009319
## sample estimates:
## mean of x mean of y
## 0.3340960 0.3340638
```

```
# Mean baserat på simulerings
r2_beta_mean_r <- function(ncp1, d, ...) {
  y <- r2_beta(rbeta, ncp1 = ncp1, d = d, n = 1e4, ...)
  mean(y)
}
meansr <- vapply(ss, function(d) r2_beta_mean_r(ncp_1, d), numeric(1))

# Kollar om det finns ngn skillnad mellan de två sätten
t.test(means, meansr)
```

```

## 
## Welch Two Sample t-test
## 
## data: means and meansr
## t = 1.3111, df = 1923.7, p-value = 0.19
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.001819645 0.009159424
## sample estimates:
## mean of x mean of y
## 0.3393171 0.3356472

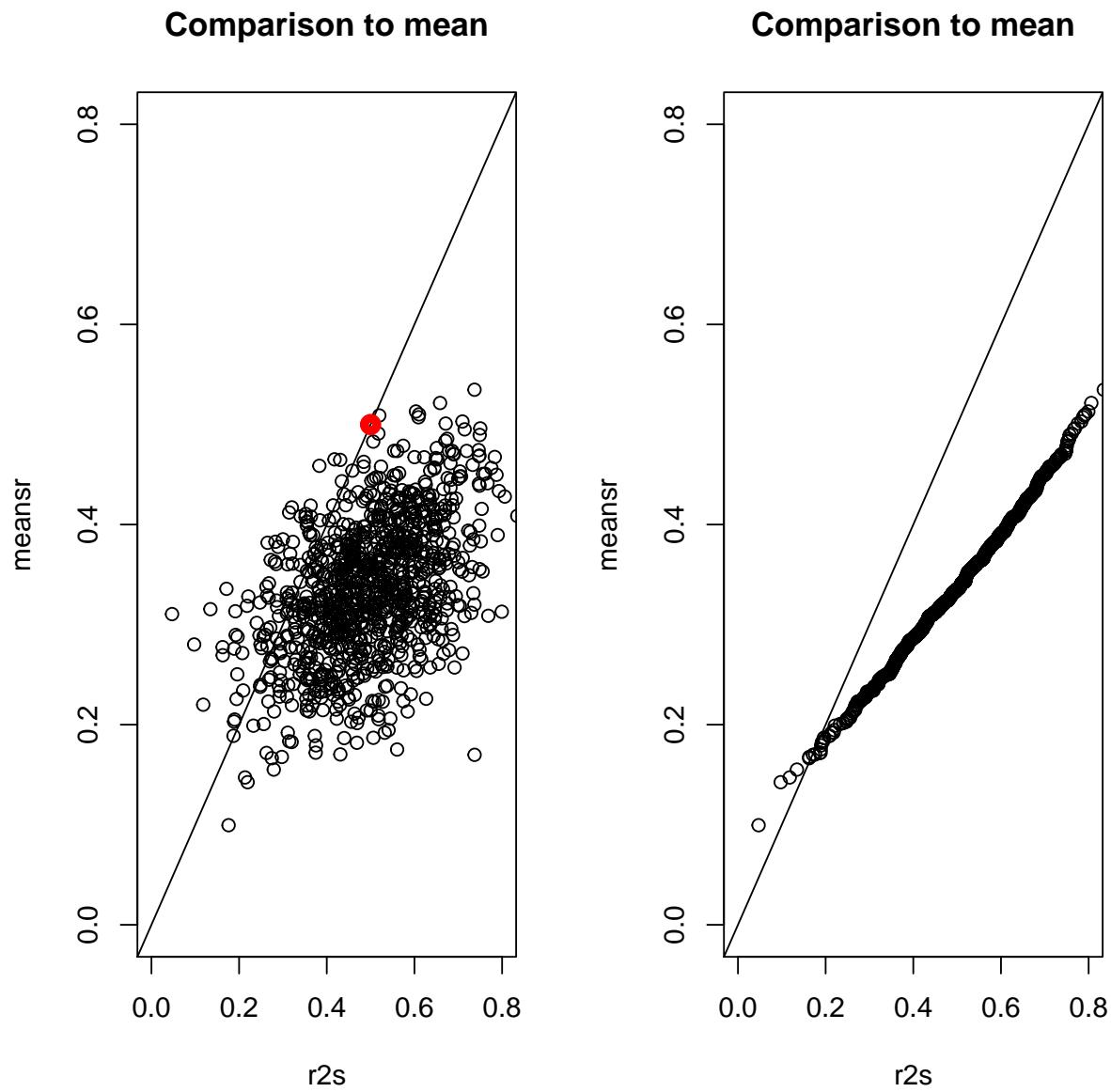
```

Alltså ingen skillnad för mode. Skillanden för mean är större men fortfarande inte signifikant.

```

par(mfrow = c(1, 2))
plot(r2s, meansr, xlim = c(0, .8), ylim = c(0, .8), main = "Comparison to mean")
abline(0, 1); points(.5, .5, lwd = 5, col = "red")
qqplot(r2s, meansr, xlim = c(0, .8), ylim = c(0, .8), main = "Comparison to mean")
abline(0, 1)

```



Noterar f.ö. att endast (Warren 1971) (samt en artikel som tycks irrelevant i sammanhanget) refererar till (Hogben 1968). Gissar därför att man inte har nyttjat dessa resultat i ngn större utsträckning. Det finns fler referenser till (Warren 1971). Jag har gått igenom dem och lagt till i läslistan.

Hade kanske varit intressant att t ex också jämföra med normalfördelning för att se vilken fördelning av dessa som är bäst och i så fall hur stor skillnaden kan vara.

3.2 Läsning av (???)

Utgår från ML-skattningar. Kan skatta både fördelning och dess parametrar. Kan också baseras på “maximum goodness-of-fit”.

`descdist`-funktionen medger också att beräkningar kan ta hänsyn till bias eller inte. Tyvärr tycke det inte möjligt att inkludera ickecentraliseringssparametern för skattning utan bara shape-paramterarna. Har försökt

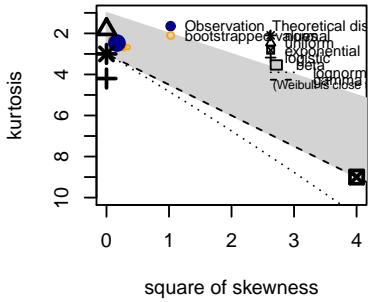
studera koden i paketet men finnser där ingen klar förklaring till varför. Kan det vara ngn skillnad som finns inbyggd i själva betafunktionerna? Samtliga Beta-funktioner nyttjar intern C-kod men jag kan se att man gör tydlig skillnad på just ncp-parametern (dock baserat på om den är missing och inte 0, vilket faktiskt är default ... förstår inte riktigt!?)

Går app skapa enskilda plotter av fitdist-objekt mha denscomp, cdfcomp, qqcomp och ppcomp.

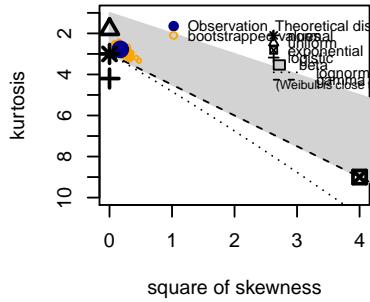
```
library("fitdistrplus")
ss <- sim_data(r2 = .5, p = 1) %>%
  subsamples(n.max = 500, N = 1000)
m <- metrics(ss, n.sample = c(10, 20, 30, 50, 100, 200, 300, 400, 500))
R2 <- as.data.frame(m$Rsquared)

# Tycks här som att vi har en beta fördelning för n upp till ca 30
# Dock för n = 200 tycks vi kunna använda gammal fördelning för approximation
# Från kanske n = 200 tycks normalapproximation kunna fungera bra.
par(mfrow = c(3, 3)); lapply(R2, descdist, boot = 1000)
```

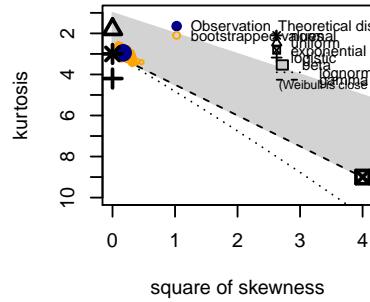
Cullen and Frey graph



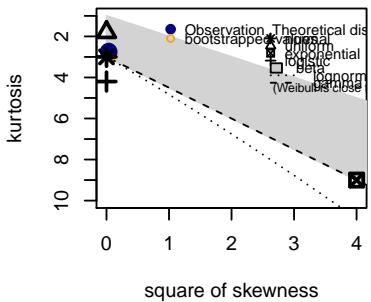
Cullen and Frey graph



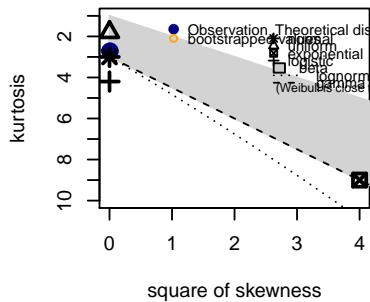
Cullen and Frey graph



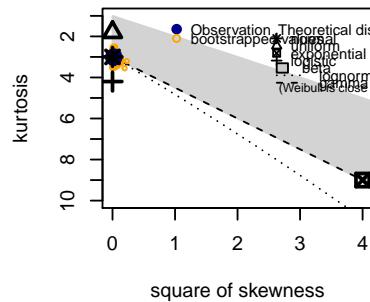
Cullen and Frey graph



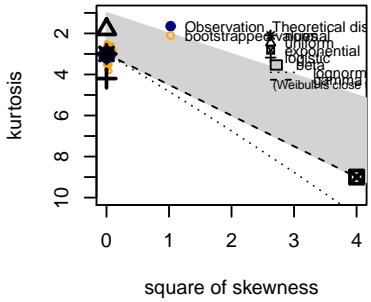
Cullen and Frey graph



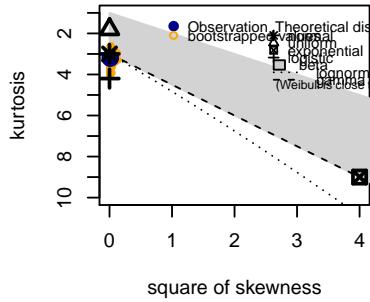
Cullen and Frey graph



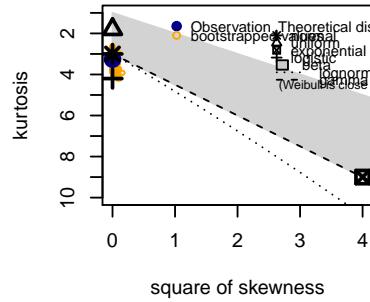
Cullen and Frey graph



Cullen and Frey graph

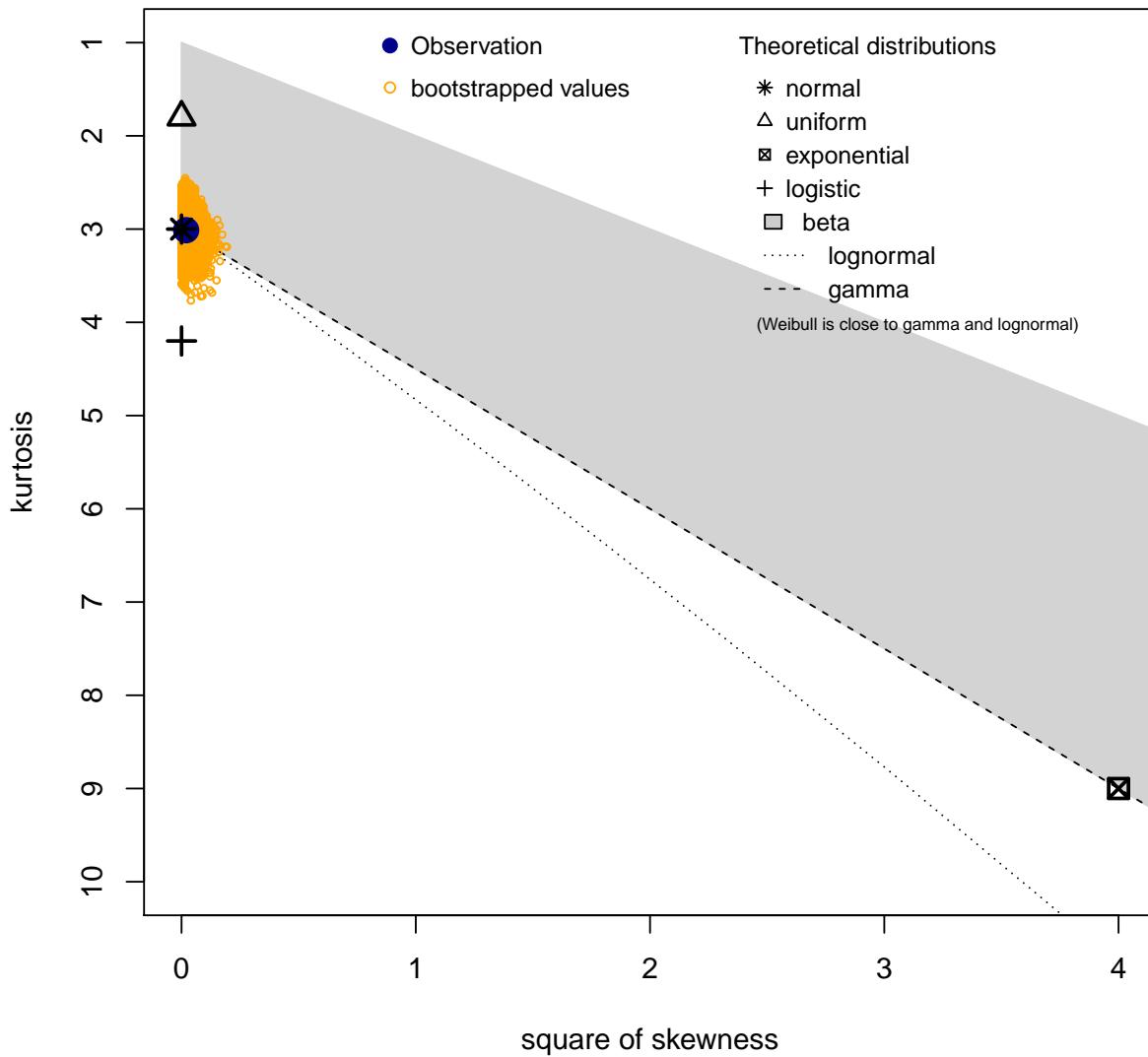


Cullen and Frey graph



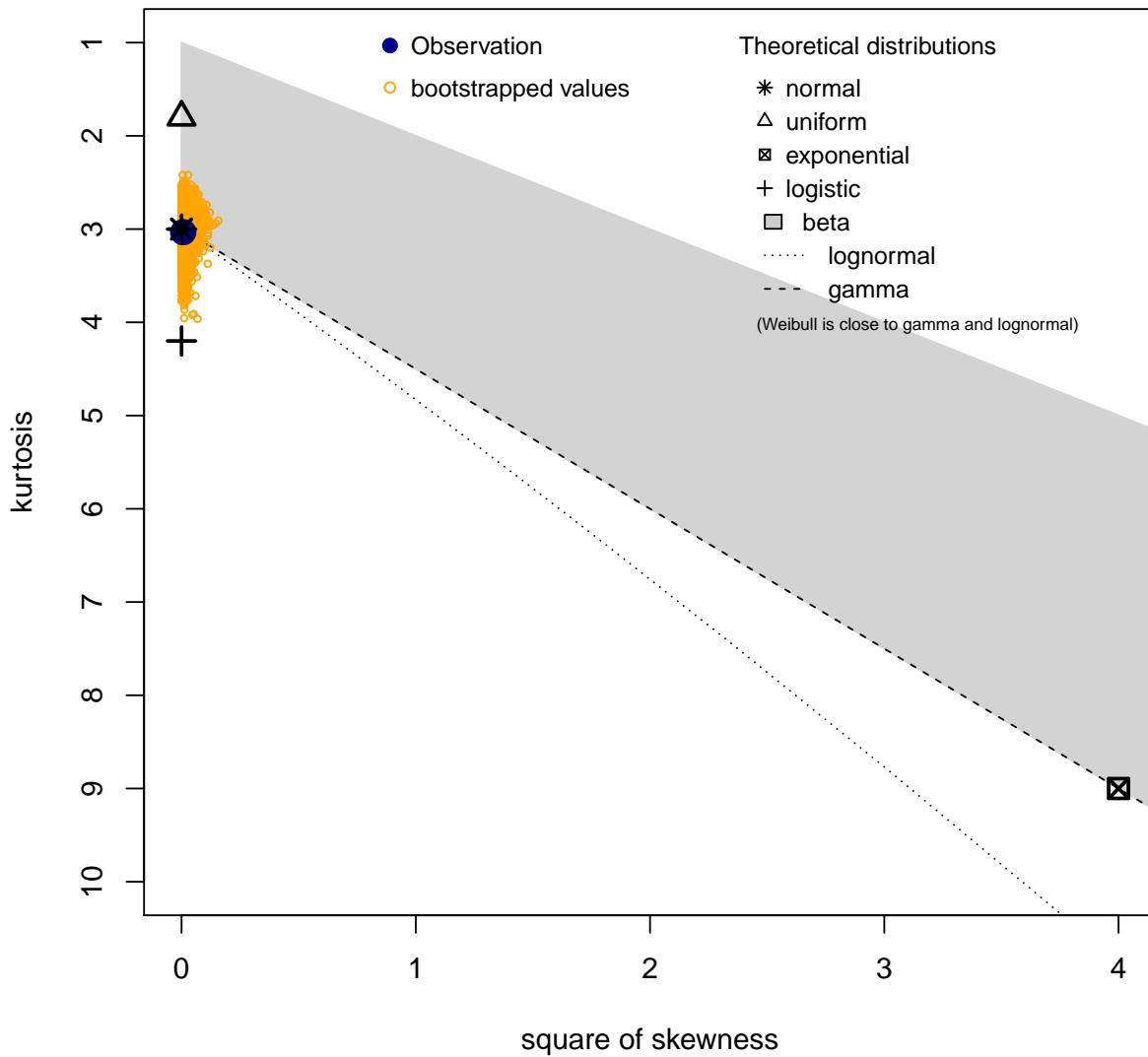
```
# Jag gör en större plot för just n = 200 för att kolla lite närmare på just detta  
# Ser här att en gammafördelning verkar kunna passa rätt bra.  
par(mfrow = c(1, 1)); descdist(R2[, 6], boot = 10000)
```

Cullen and Frey graph



```
# Ser här att en normalfördelning verkar kunna passa rätt bra vid n = 300
par(mfrow = c(1, 1)); descdist(R2[, 7], boot = 10000)
```

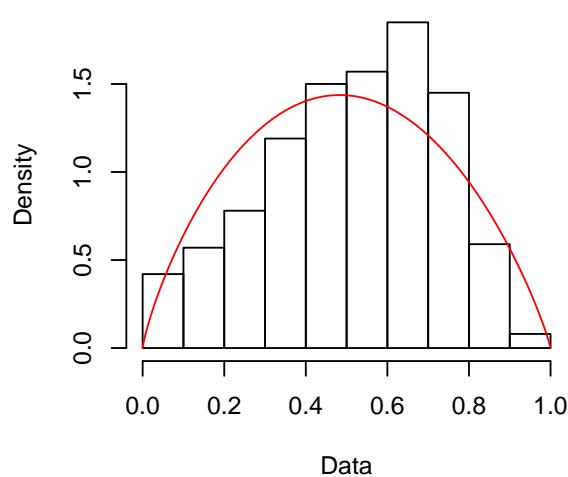
Cullen and Frey graph



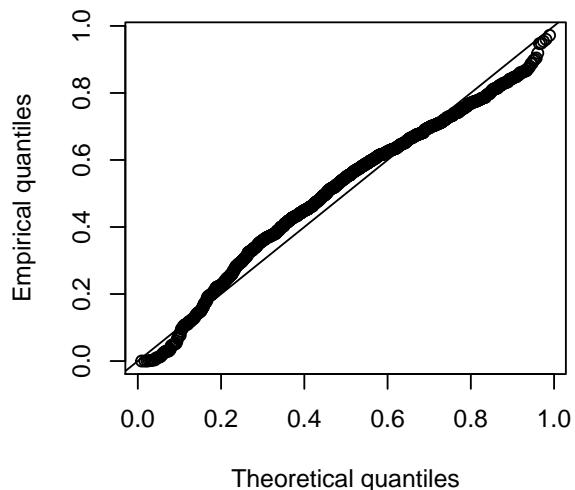
```
# Vill testa att anpassa en betafördelning
# Tydligen estimeras bara shape1 och shape2, inte ncp (vilket gör att resultatet inte blir jättebra)

lapply(R2, function(x) {
  fit <- fitdist(x, "beta")
  plotdist(x, "beta", para = list(shape1 = fit$estimate[1], shape2 = fit$estimate[2]))
})
```

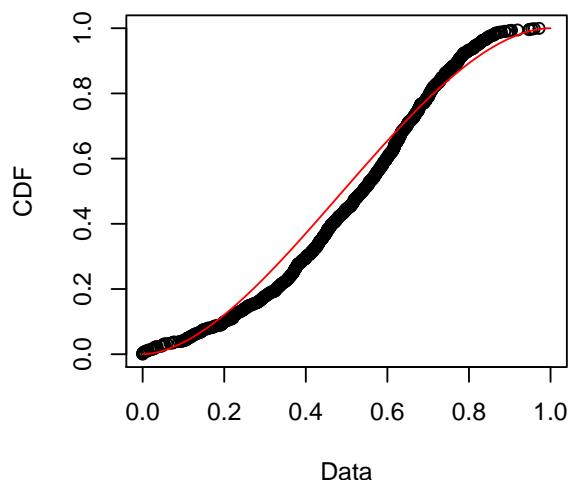
Empirical and theoretical dens.



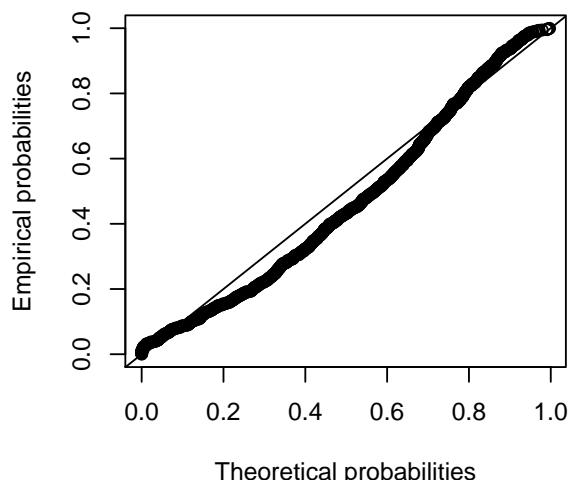
Q–Q plot



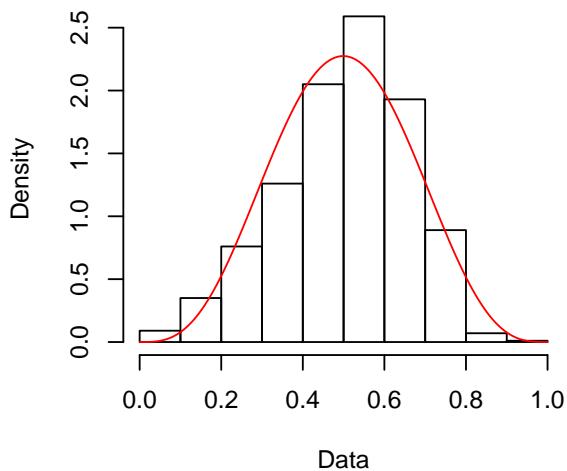
Empirical and theoretical CDFs



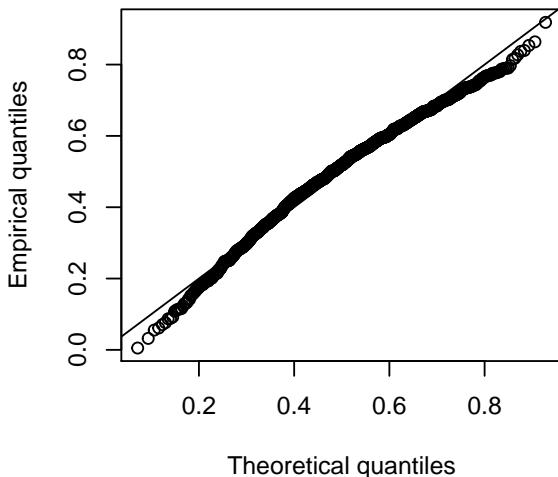
P–P plot



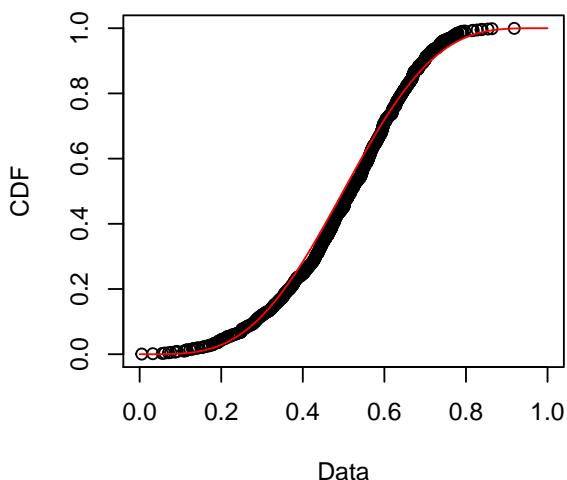
Empirical and theoretical dens.



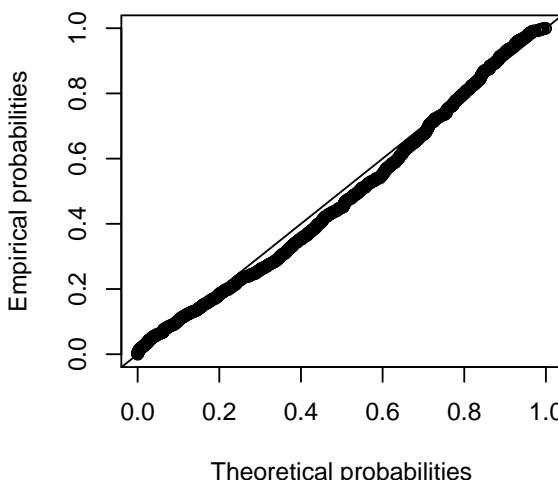
Q–Q plot



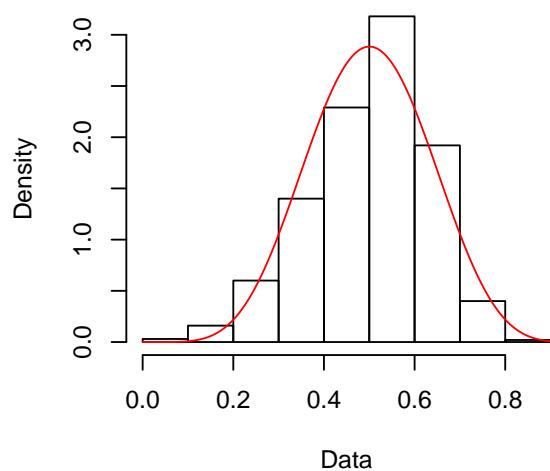
Empirical and theoretical CDFs



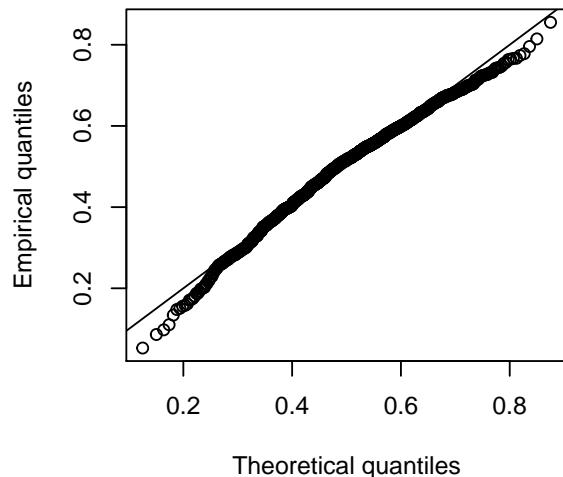
P–P plot



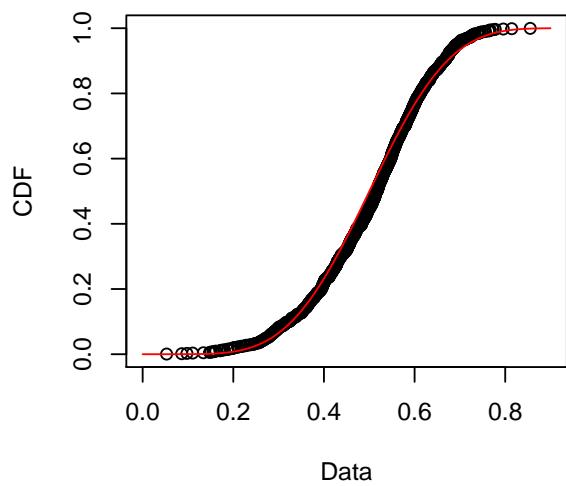
Empirical and theoretical dens.



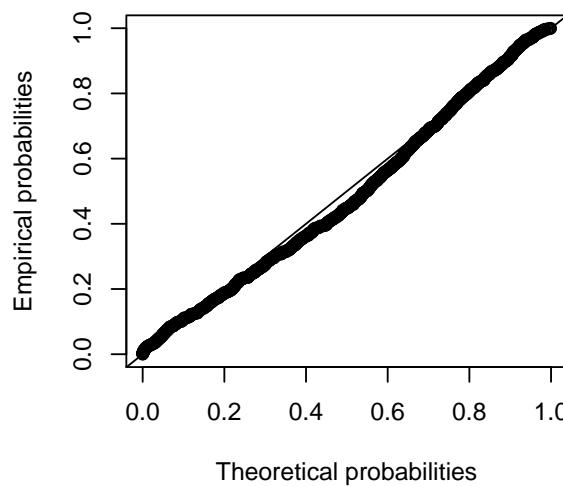
Q–Q plot



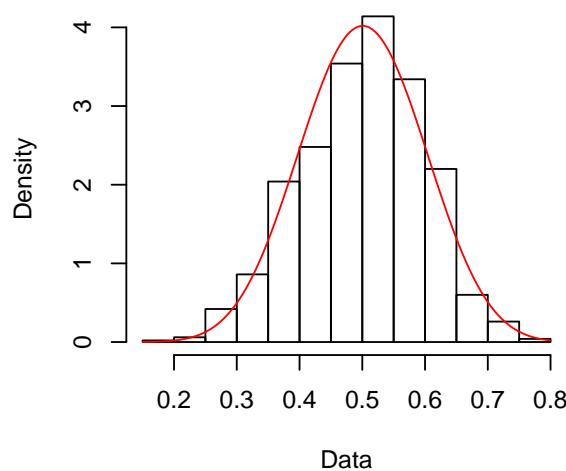
Empirical and theoretical CDFs



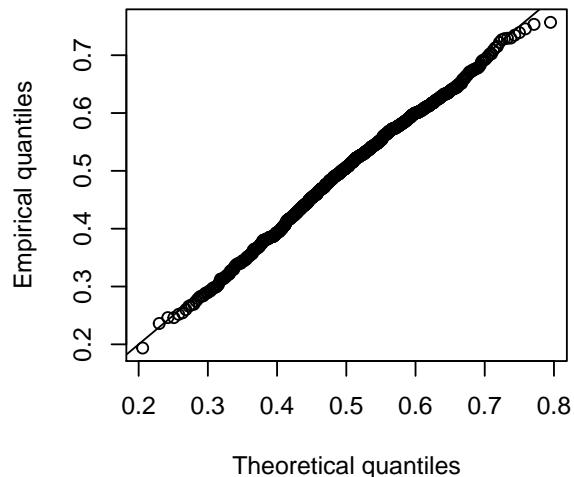
P–P plot



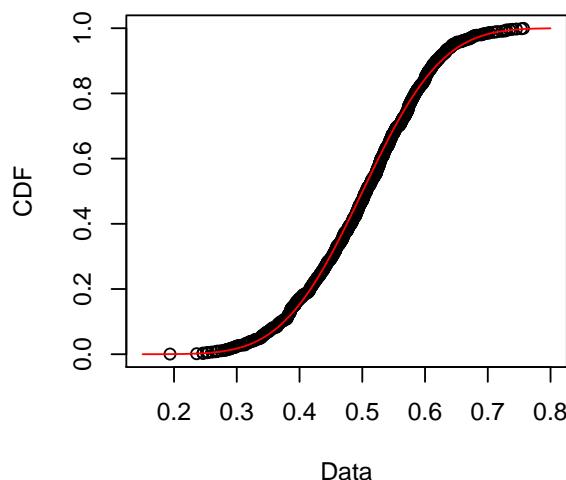
Empirical and theoretical dens.



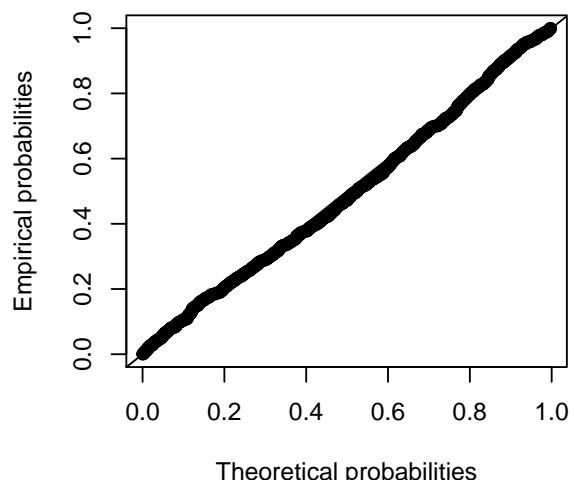
Q–Q plot



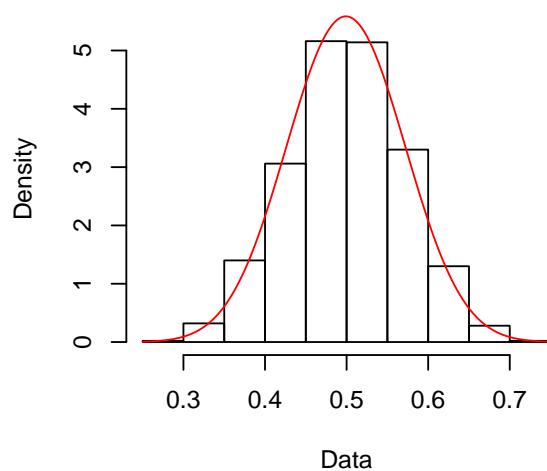
Empirical and theoretical CDFs



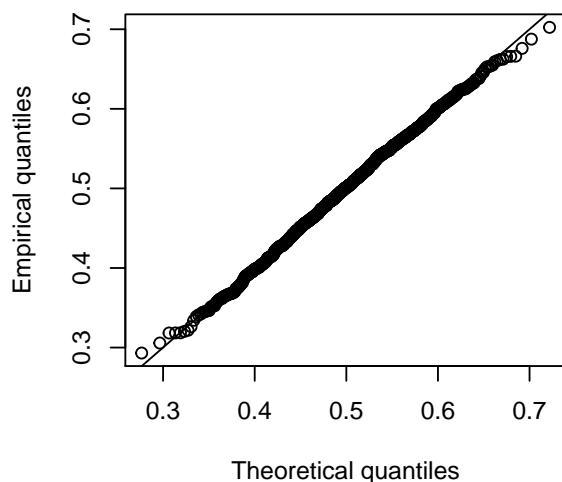
P–P plot



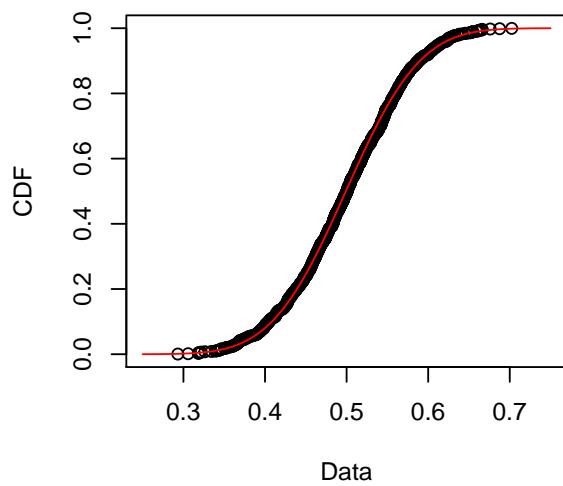
Empirical and theoretical dens.



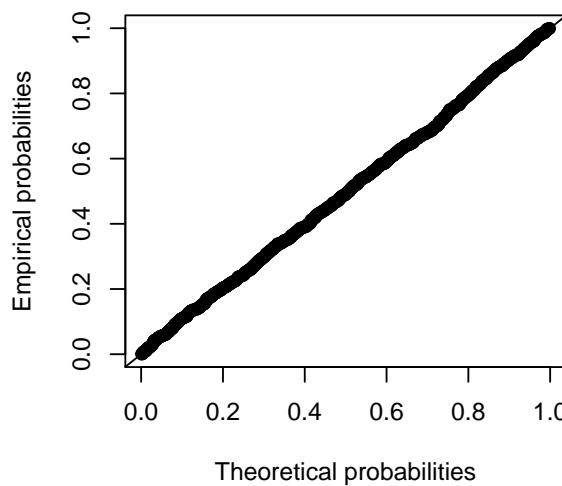
Q–Q plot



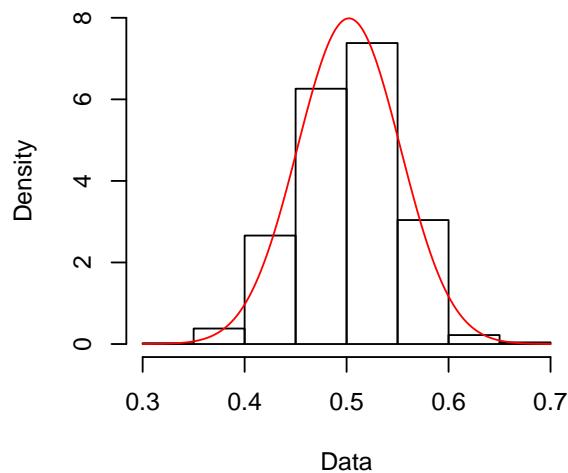
Empirical and theoretical CDFs



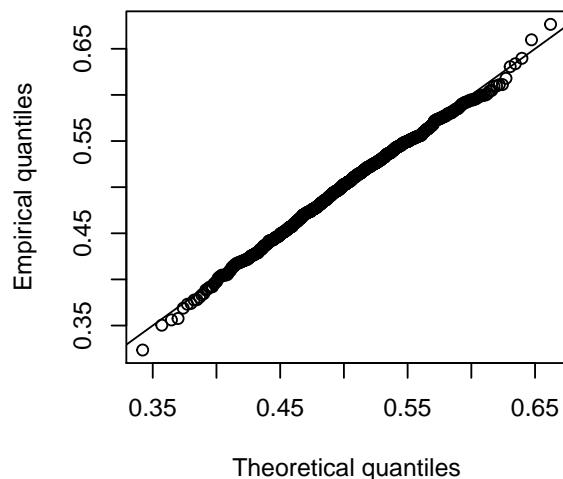
P–P plot



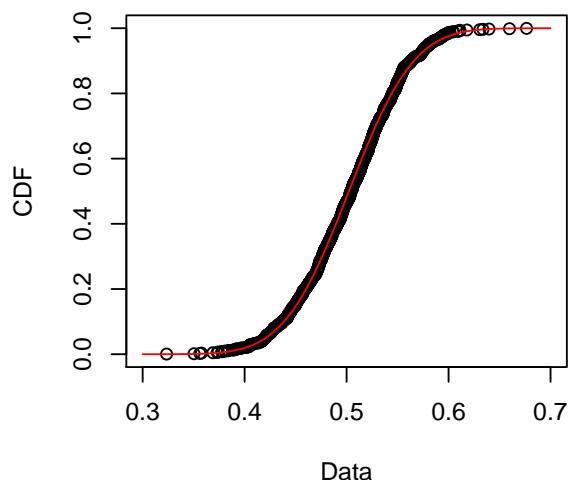
Empirical and theoretical dens.



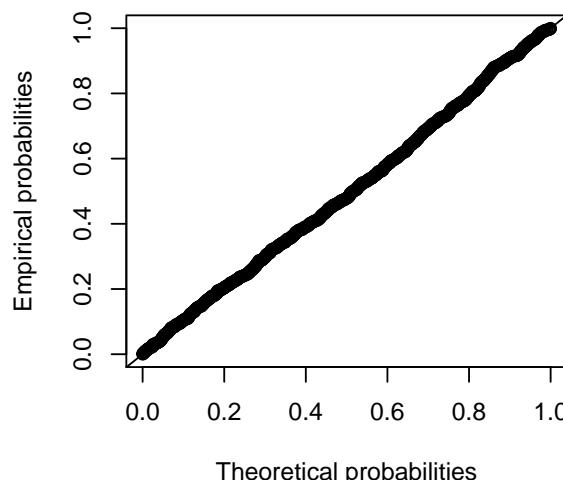
Q–Q plot



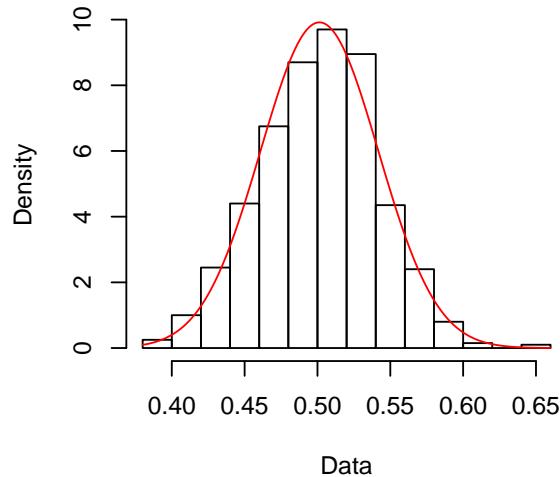
Empirical and theoretical CDFs



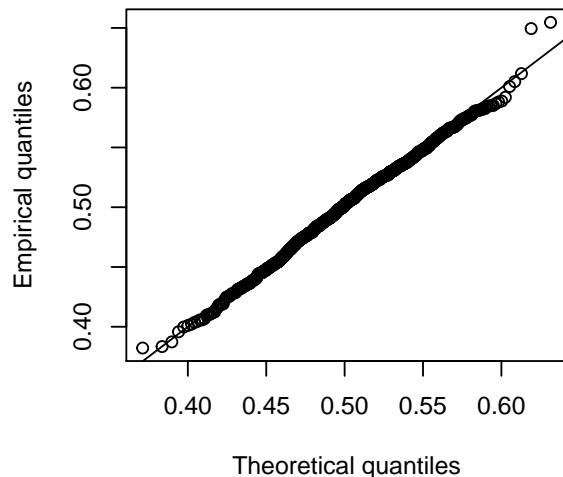
P–P plot



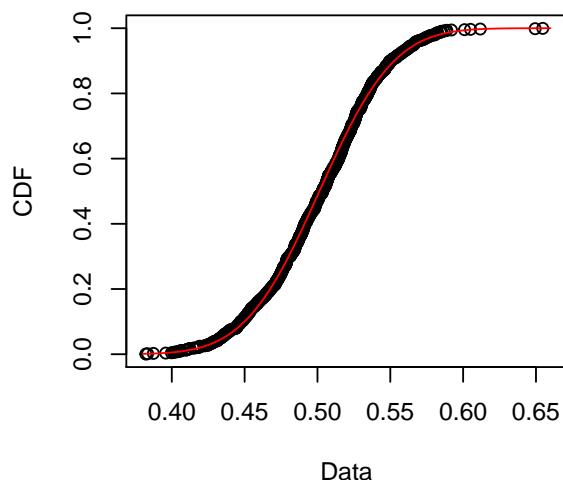
Empirical and theoretical dens.



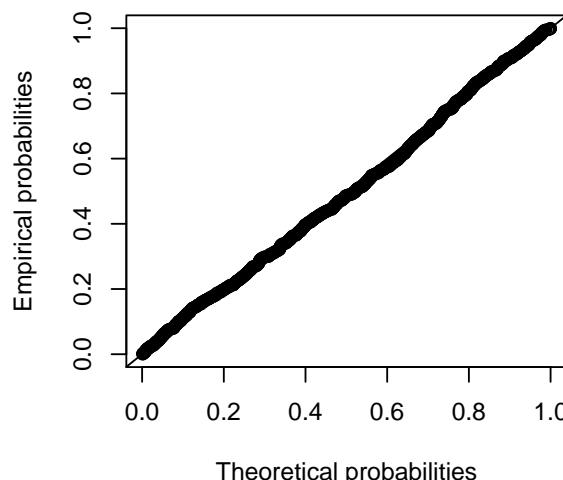
Q–Q plot



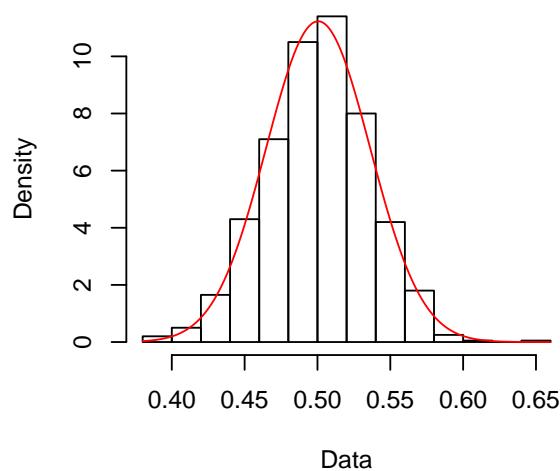
Empirical and theoretical CDFs



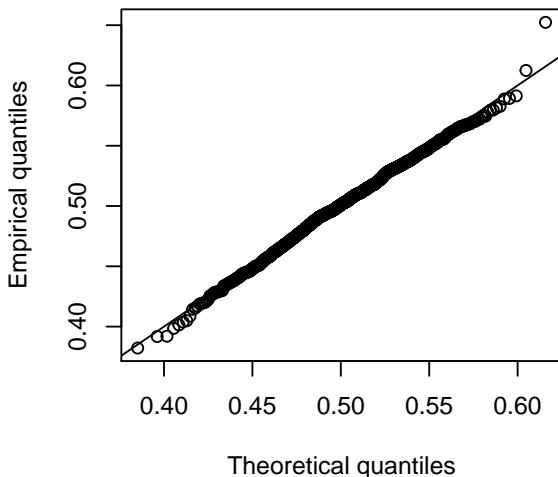
P–P plot



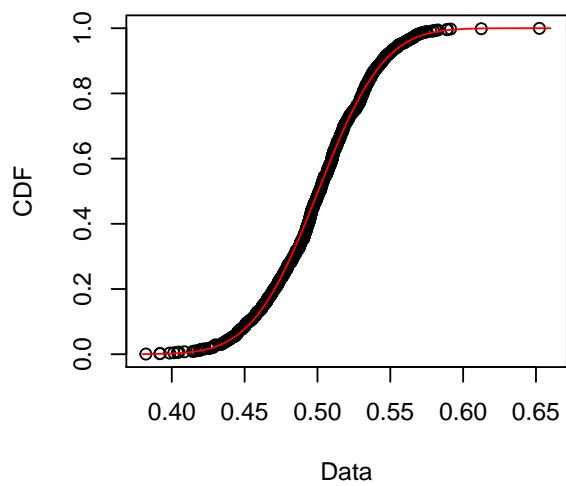
Empirical and theoretical dens.



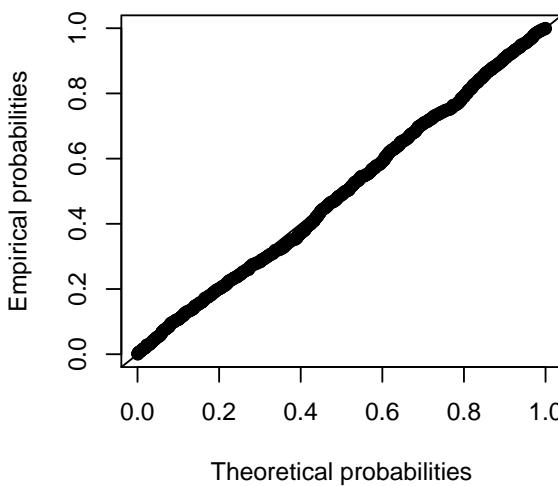
Q–Q plot



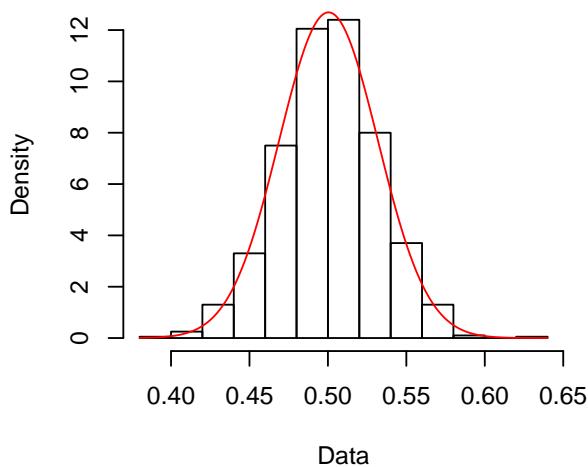
Empirical and theoretical CDFs



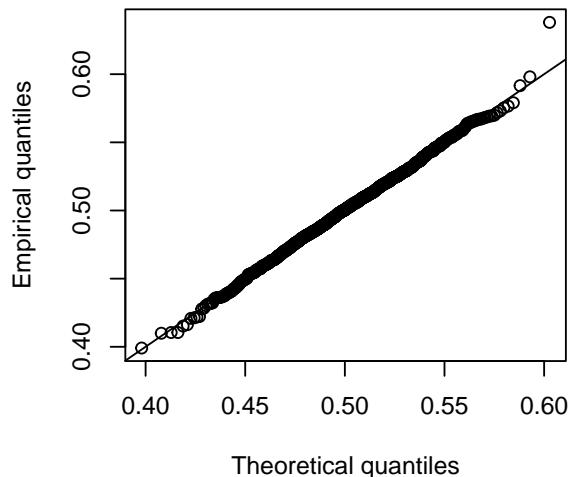
P–P plot



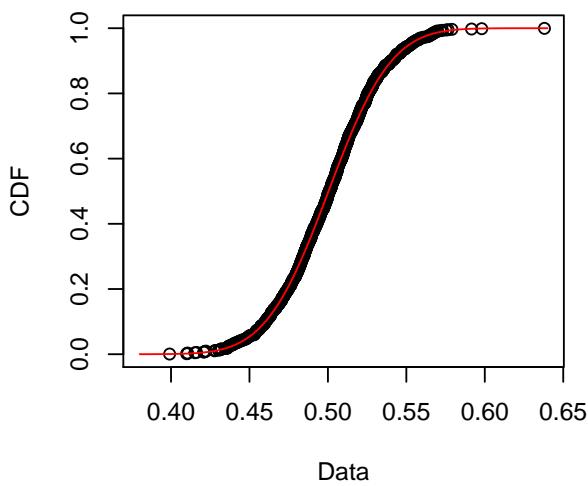
Empirical and theoretical dens.



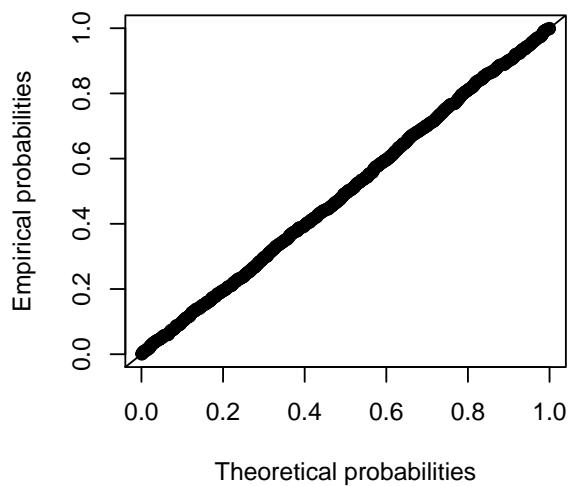
Q–Q plot



Empirical and theoretical CDFs



P–P plot



Vid första anblick tycks det här som att den vanliga betafördelningen trots allt kanske kan passa någorlunda? (Dock inte för alltför små n!)

- Finns det någon systematik i hur `shape1` och `shape2` skattas utifrån `n`? Hade varit jättenitressant i så fall!
- Kan vi använda dessa betafördelningar och jämföra mot den ickecentrala betafördelning pss som ovan?

4 2016-03-02

Fortsätter leka lite med paketet `fitdistrplus`

4.1 Testar att modifiera beta-funktionerna för att också skatta ncp (misslyckas)

```
pbeta <- function(q, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
  .Call(C_pnbeta, q, shape1, shape2, ncp, lower.tail, log.p)
dbeta <- function(x, shape1, shape2, ncp = 0, log = FALSE)
  .Call(C_dnbeta, x, shape1, shape2, ncp, log)
qbeta <- function(p, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
  .Call(C_qnbneta, p, shape1, shape2, ncp, lower.tail, log.p)
rbeta <- function(n, shape1, shape2, ncp = 0) {
  X <- rchisq(n, 2 * shape1, ncp = ncp)
  X/(X + rchisq(n, 2 * shape2))
}
fitdist(R2[, 1], "beta")

## Fitting of the distribution 'beta' by maximum likelihood
## Parameters:
##           estimate Std. Error
## shape1 1.823046 0.07720639
## shape2 1.881074 0.07998226
```

Detta upprepades också med alla tillgängliga val av method men utan att lyckas.

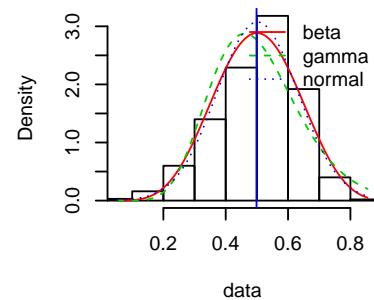
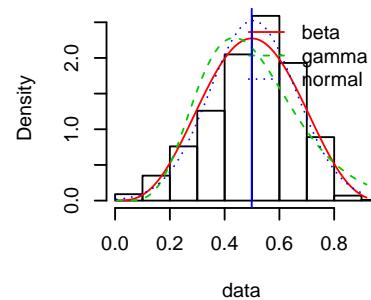
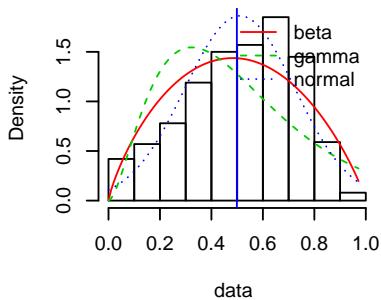
4.2 Jämför resultat för olika fördelningar

Vi har sett ovan att beta, gamma och normalfördelning kan funka för olika stickprovsstorlekar. Vi kan undersöka detta.

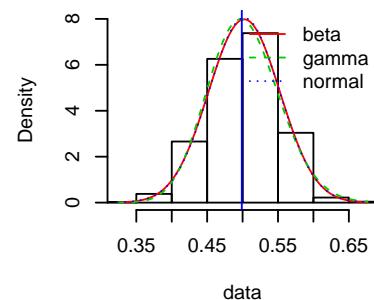
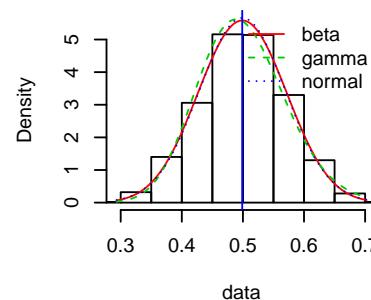
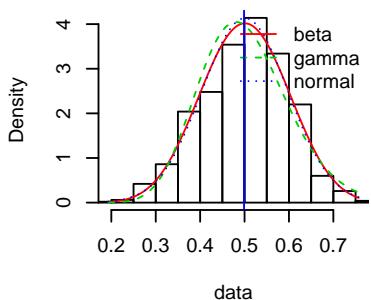
```
par(mfrow = c(3, 3))
plot.legend <- c("beta", "gamma", "normal")
fitdists <- function(x, distr = c("beta", "gamma", "norm"))
  lapply(distr, function(d) fitdist(x, d))

denscomps <- function(m, distr = c("beta", "gamma", "norm"), ...) {
  R2 <- as.data.frame(m$Rsquared)
  lapply(R2, function(r2) {
    denscomp(fitdists(r2, distr = distr), ...)
    abline(v = attr(m, "real_Rsquared"), col = "blue")
  })
}
denscomps(m, legendtext = plot.legend)
```

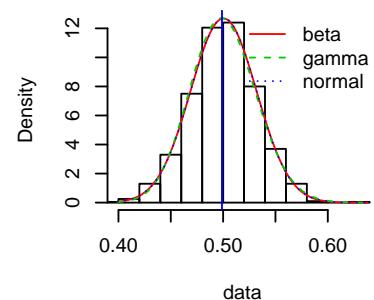
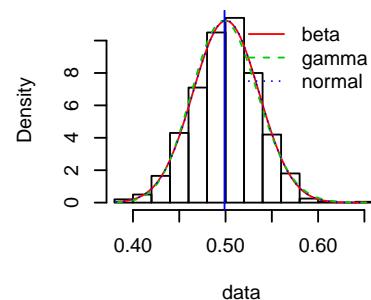
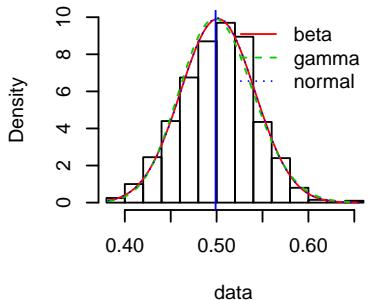
Histogram and theoretical densities



Histogram and theoretical densities



Histogram and theoretical densities



```
## $`10`  
## NULL  
##  
## $`20`  
## NULL  
##  
## $`30`  
## NULL  
##  
## $`50`  
## NULL  
##  
## $`100`  
## NULL
```

```

## 
## $`200` 
## NULL 
## 
## $`300` 
## NULL 
## 
## $`400` 
## NULL 
## 
## $`500` 
## NULL

```

Käns som att vi efter detta ganska klart kan förkasta gammafördelningen som lämplig kandidat men kanske att den ändå skulle passa bättre för mindre ρ (då den ju har en pos eskevhet).

OBS! Denna chunk sparas för att kunna härleda vad som gjort men ej visat sig fruktsamt. Koden ska inte behöva anropas etc på nytt!

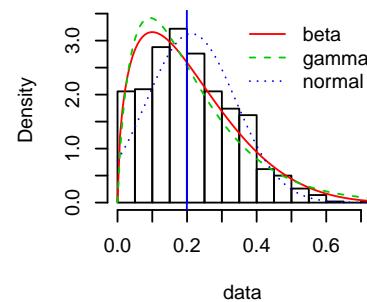
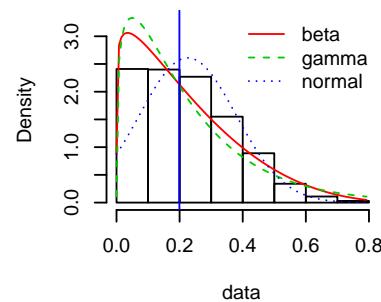
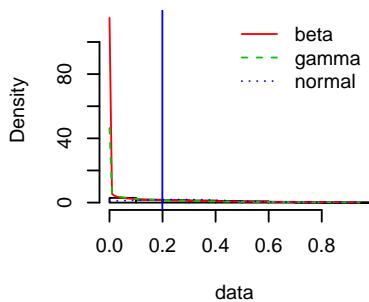
```

distplots <- function(r2 = .2, p = 1) {
  ss <- sim_data(r2 = r2, p = p) %>%
    subsamples(n.max = 500, N = 1000)
  m <- metrics(ss, n.sample = c(10, 20, 30, 50, 100, 200, 300, 400, 500))
  denscomps(m, legendtext = plot.legend)
}

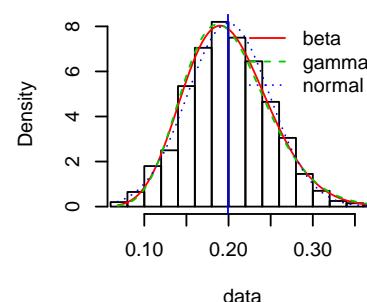
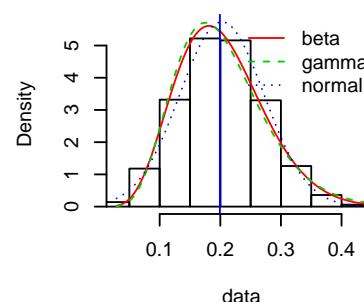
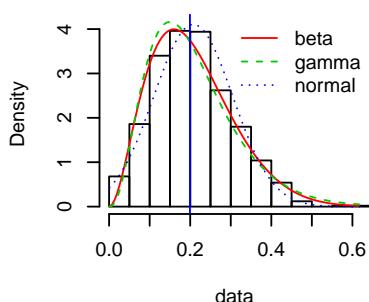
par(mfrow = c(3, 3))
lapply(c(.2, .5, .8), distplots)

```

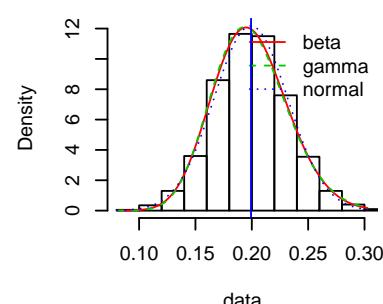
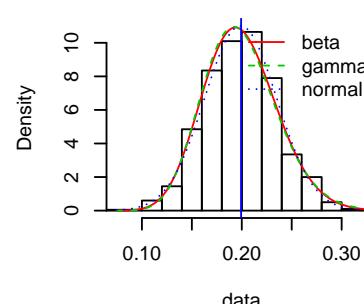
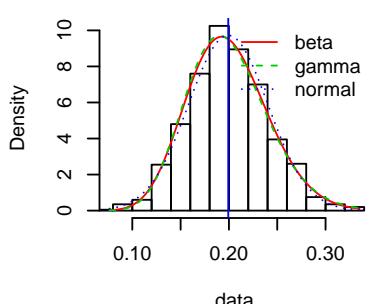
Histogram and theoretical densities



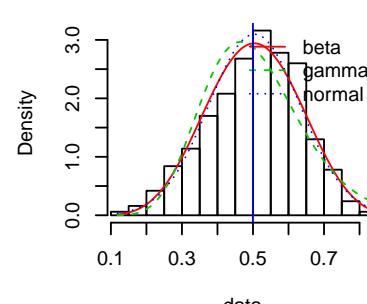
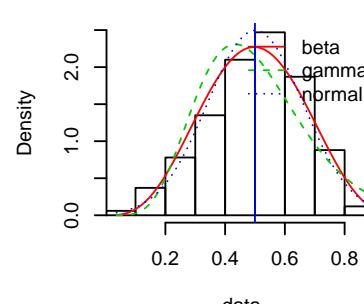
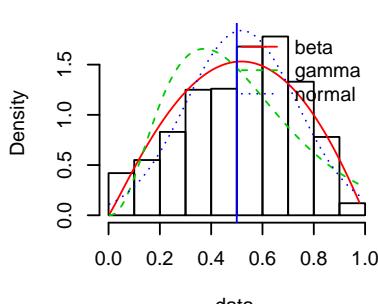
Histogram and theoretical densities



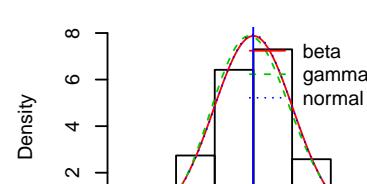
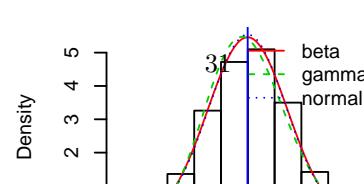
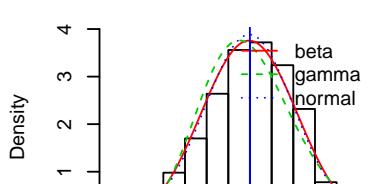
Histogram and theoretical densities



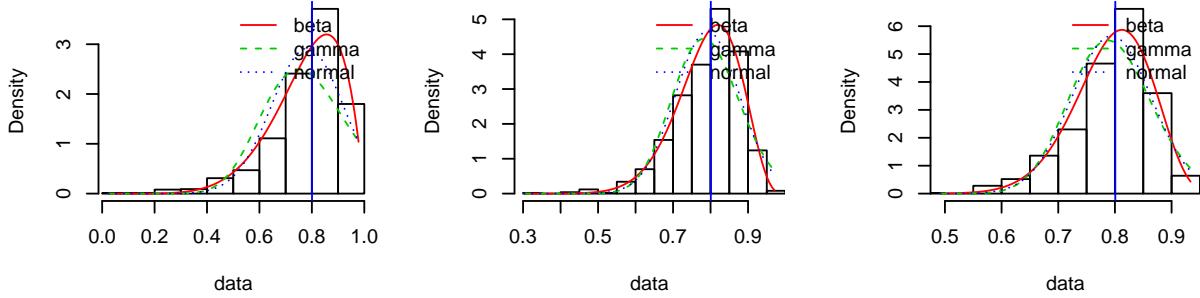
Histogram and theoretical densities



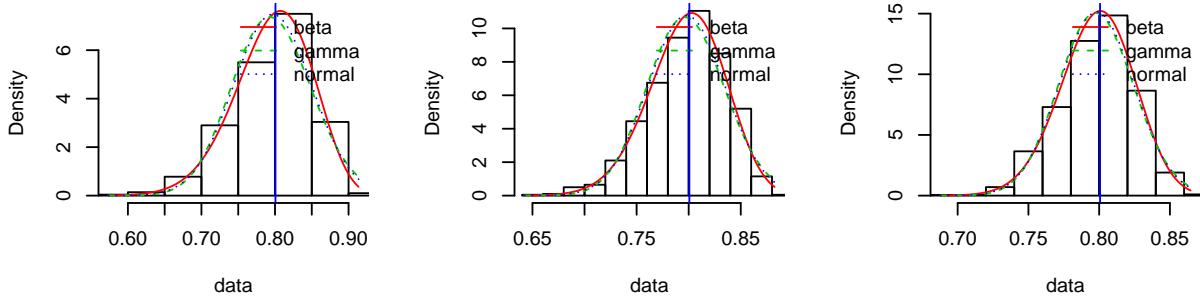
Histogram and theoretical densities



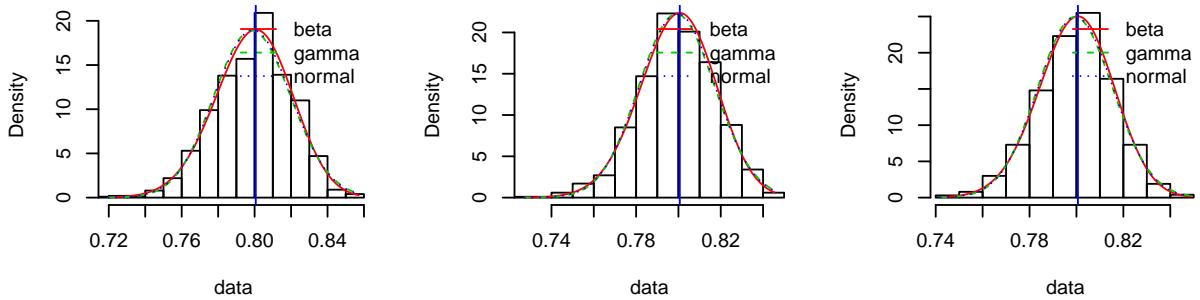
Histogram and theoretical densities



Histogram and theoretical densities



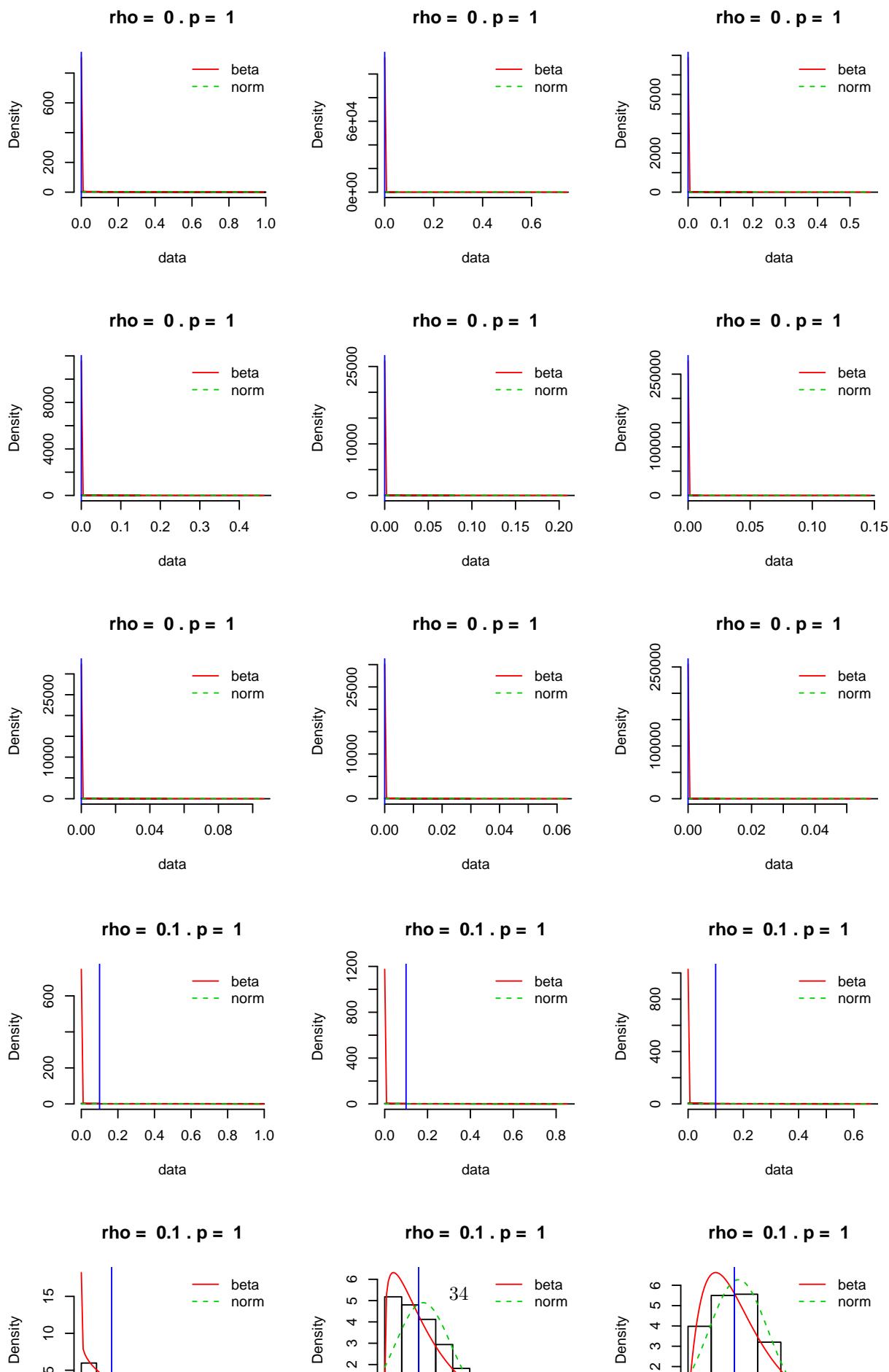
Histogram and theoretical densities

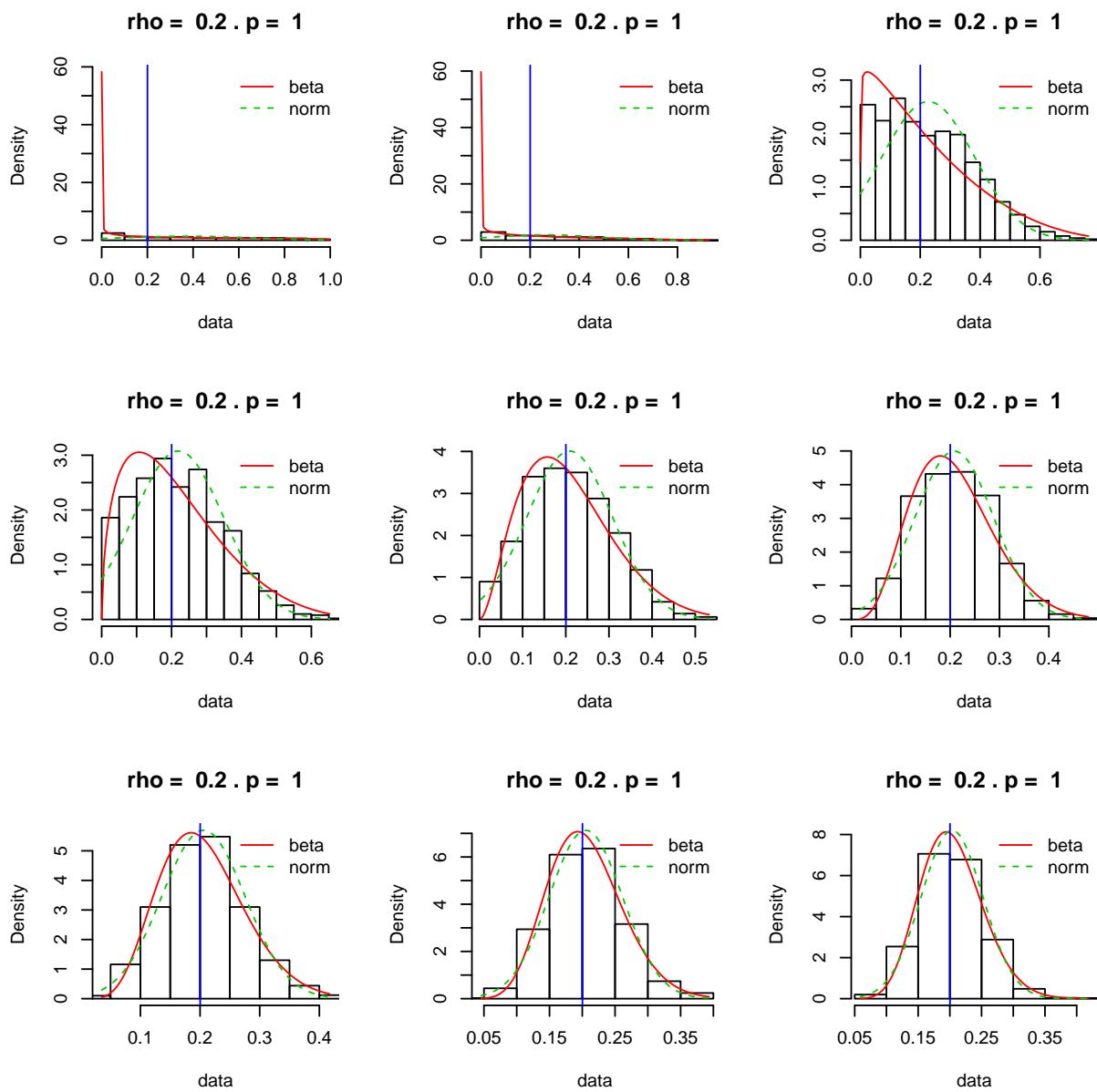


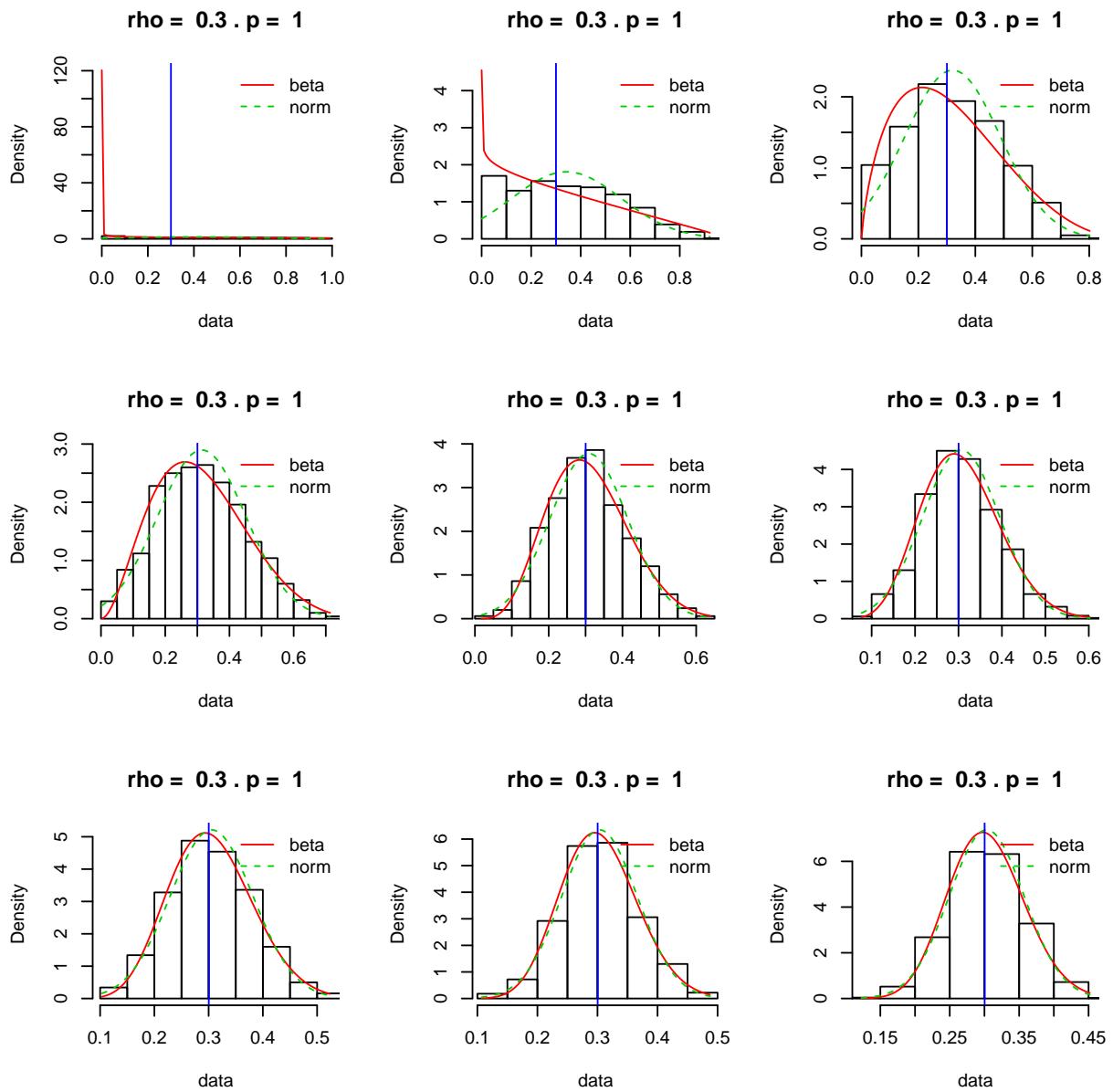
Finner att gamma de facto passar bättre för just små ρ men att beta ändå är ett bättre alternativ. Ser därför ingen anledning att fortsätta studera gamma-fördelningen i detta sammanhang. Ser också att resultaten är ganska samstämmiga för $n > 200$ så begränsar mig dit men tar istället in lite fler mellanliggande värden som kanske kan vara intressanta. Värjer mig nu heller inte för att ta ännu fler ρ .

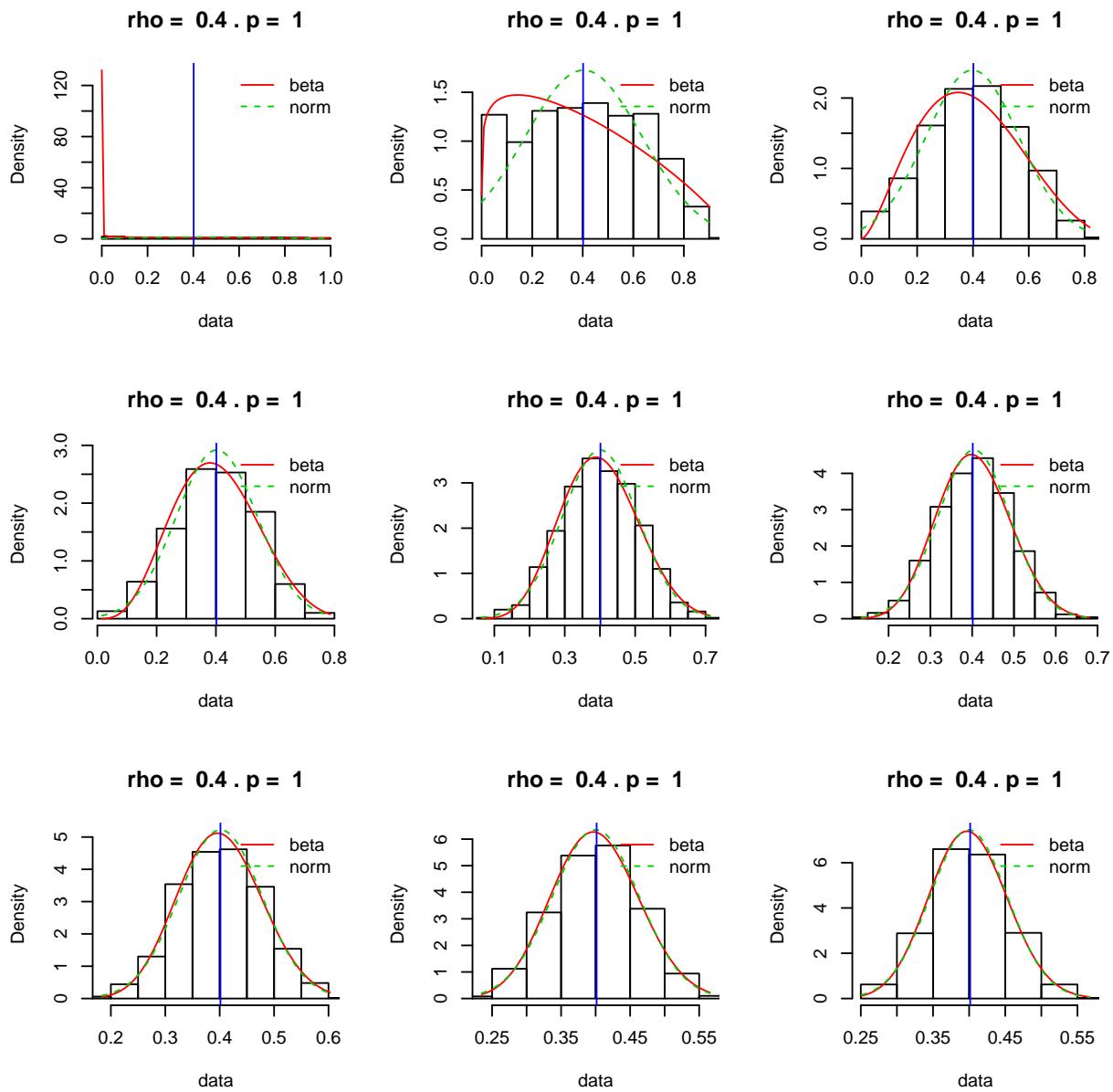
```
distplots <- function(r2 = .2, p = 1, distr = c("beta", "norm"),
                      n.sample = c(5, 10, 20, 30, 50, 75, 100, 150, 200), ...){
  ss <- sim_data(r2 = r2, p = p) %>%
    subsamples(n.max = max(n.sample), N = 1000)
  m <- metrics(ss, n.sample = n.sample)
  par(mfrow = c(floor(sqrt(length(n.sample))), ceiling(sqrt(length(n.sample))))))
  denscomps(m, distr = distr, legendtext = distr, main = paste("rho = ", r2, ". p = ", p), ...)
}

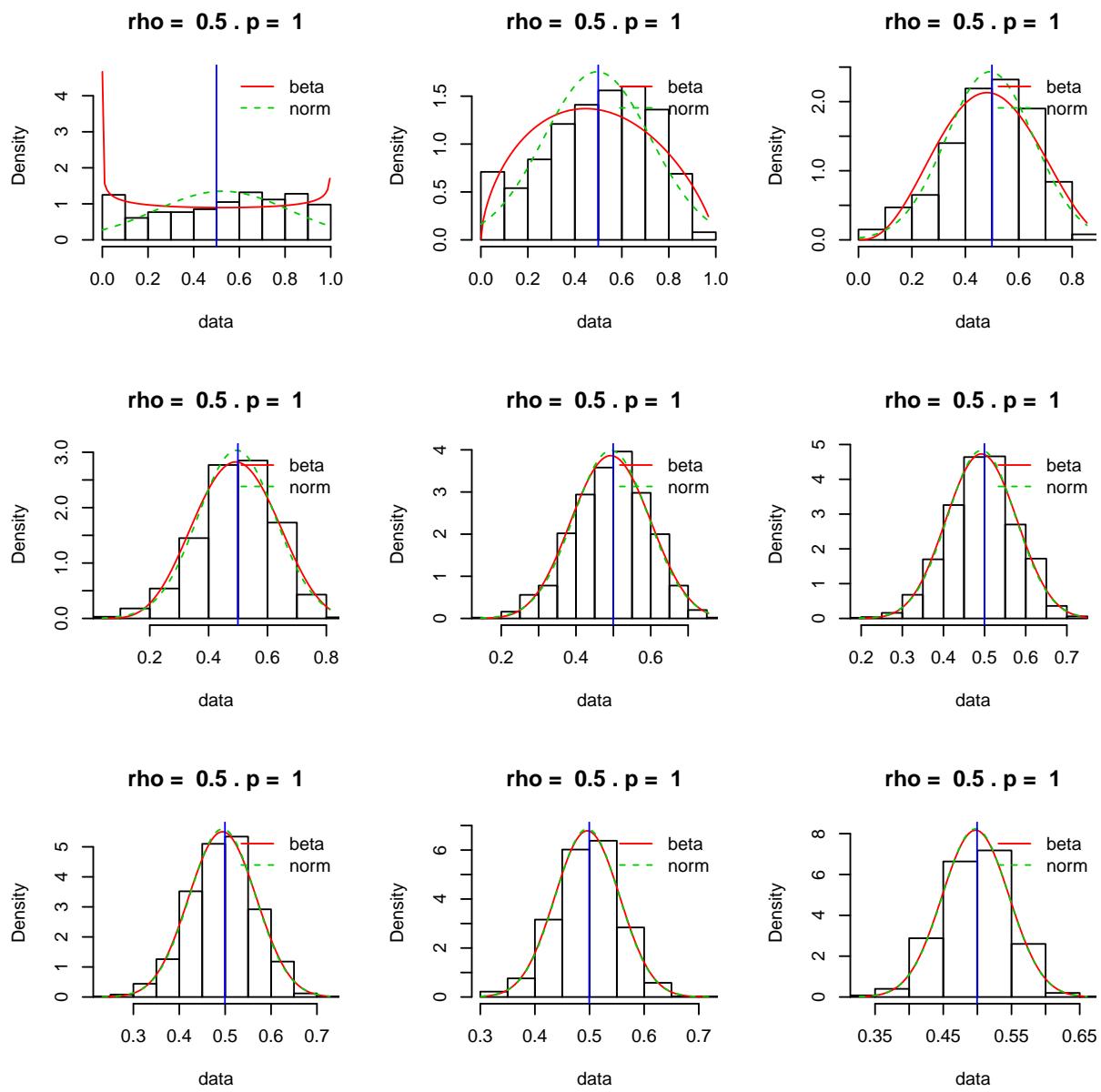
lapply(seq(0, .9, .1), distplots)
```

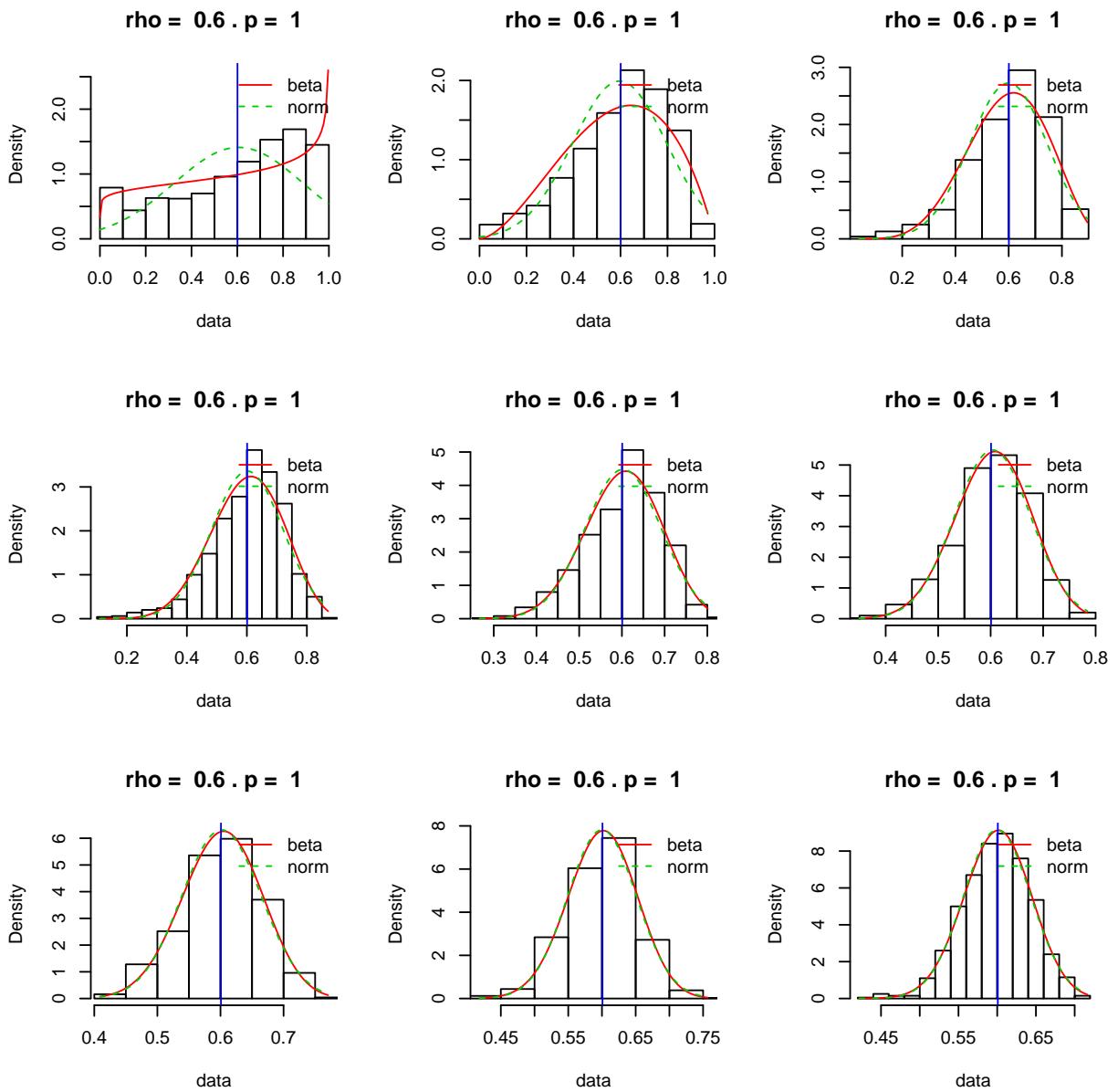



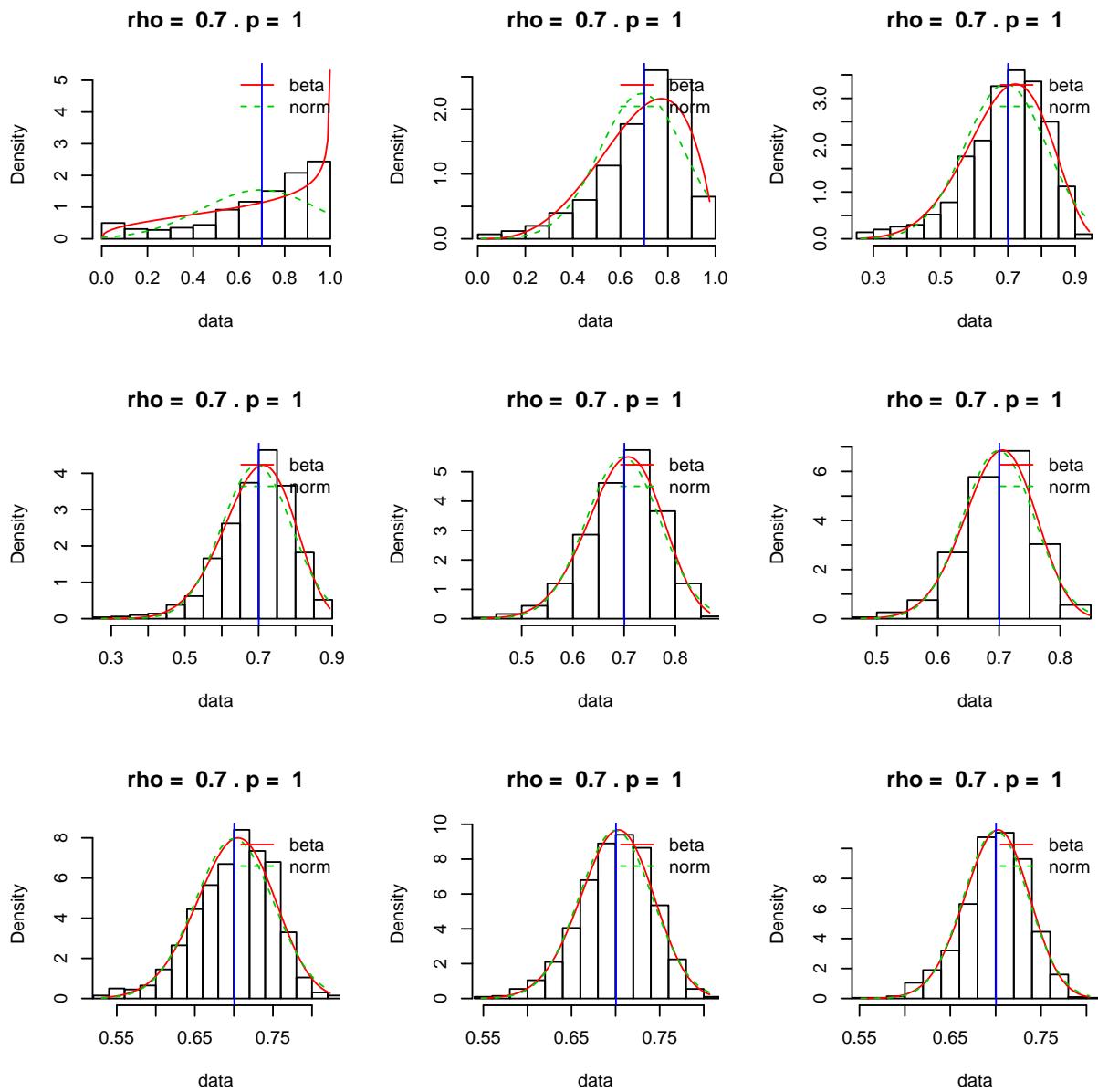


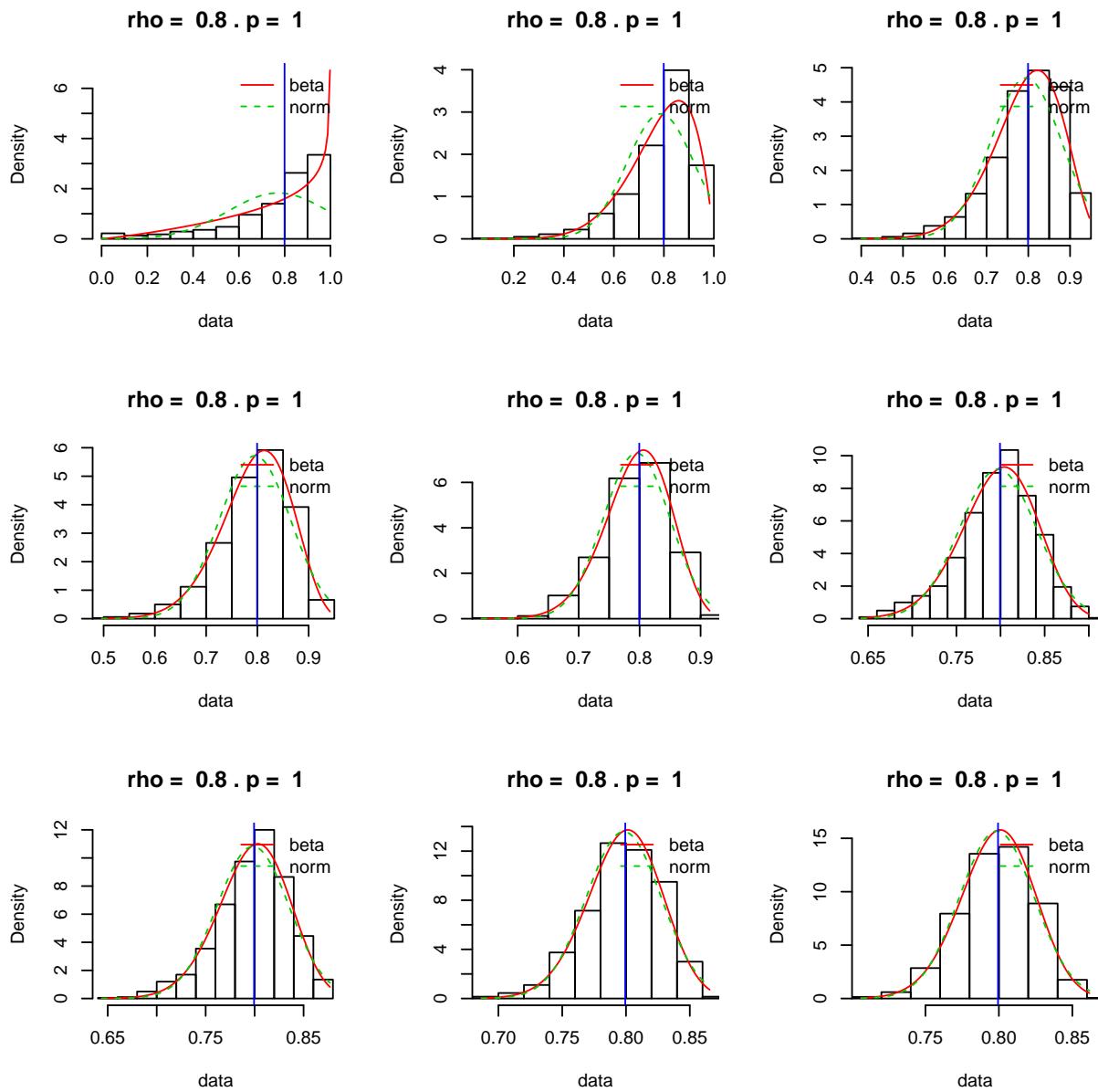


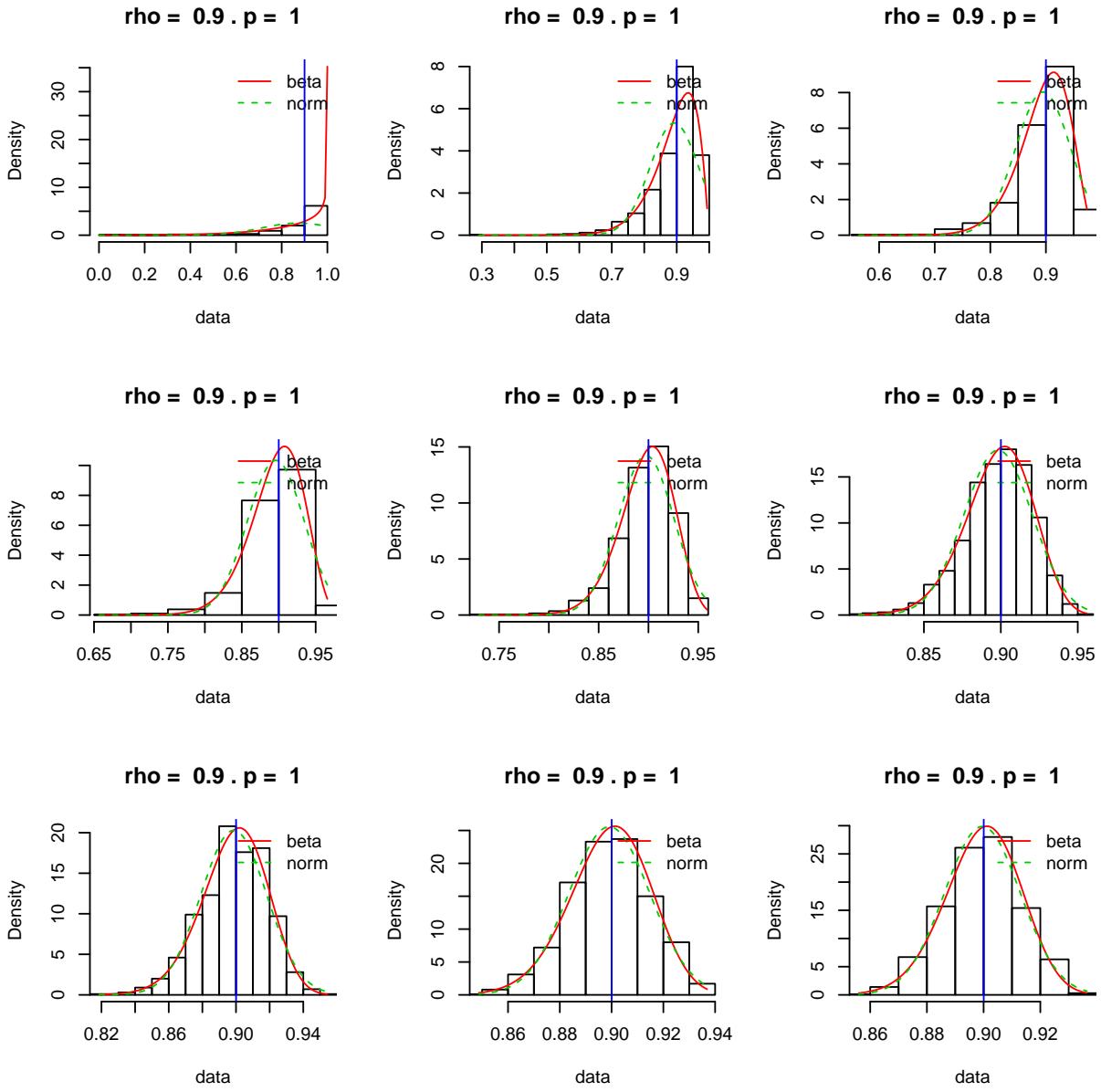












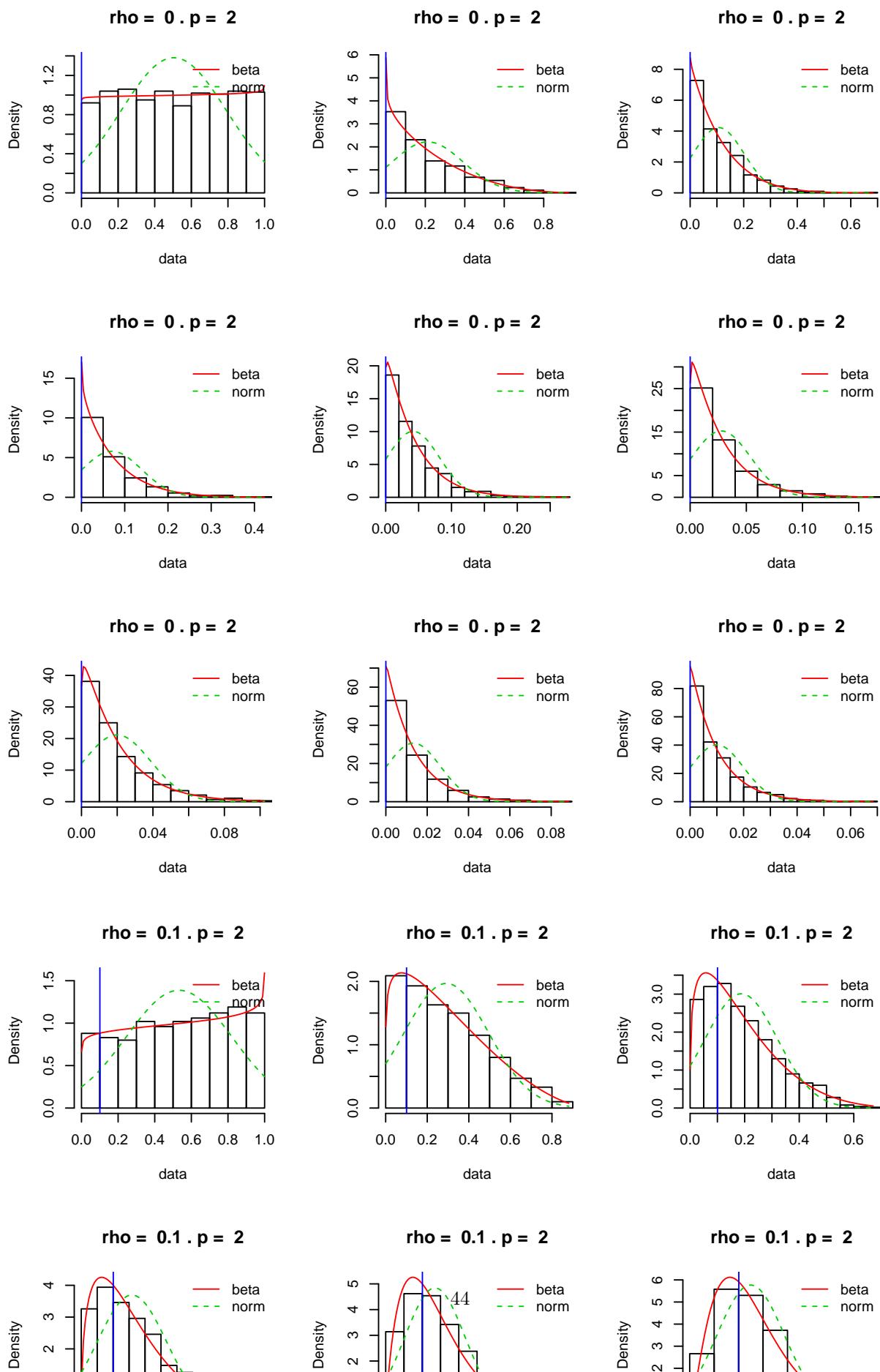
Slutsatser: Vi kan se att beta-fördelningen ger alldeles för höga skattningar för väldigt små ρ . Problemet är alltså inte att det inte blir några data men att y-axeln dras ut så mkt att dessa värden inte syns. När ρ blir större blir det lättare att anpassa fördelning även om n mindre. I några fall kan normalfördelningen tyckas t o m bättre än beta men gissar att detta trots allt beror på slumpen. När ρ dock ökar till att ligga närmare .5 börjar fördelningen krumbukta sig ganska ordentligt för ökande $n < 20$. För $\rho > .6$ tycks det som att betafördelningen inte riktigt når upp till högsta värdena i histogrammet. I detta avseende är t o m normalfördelningen bättre. För $n > 20$ tycks det dessutom som att beta och normal är approximativt likvärdiga. Att högsta stapeln ligger över betafördelningens mode tycks f.o. hålla i sig för växande ρ men effekten avtar med mindre n för ökade ρ .

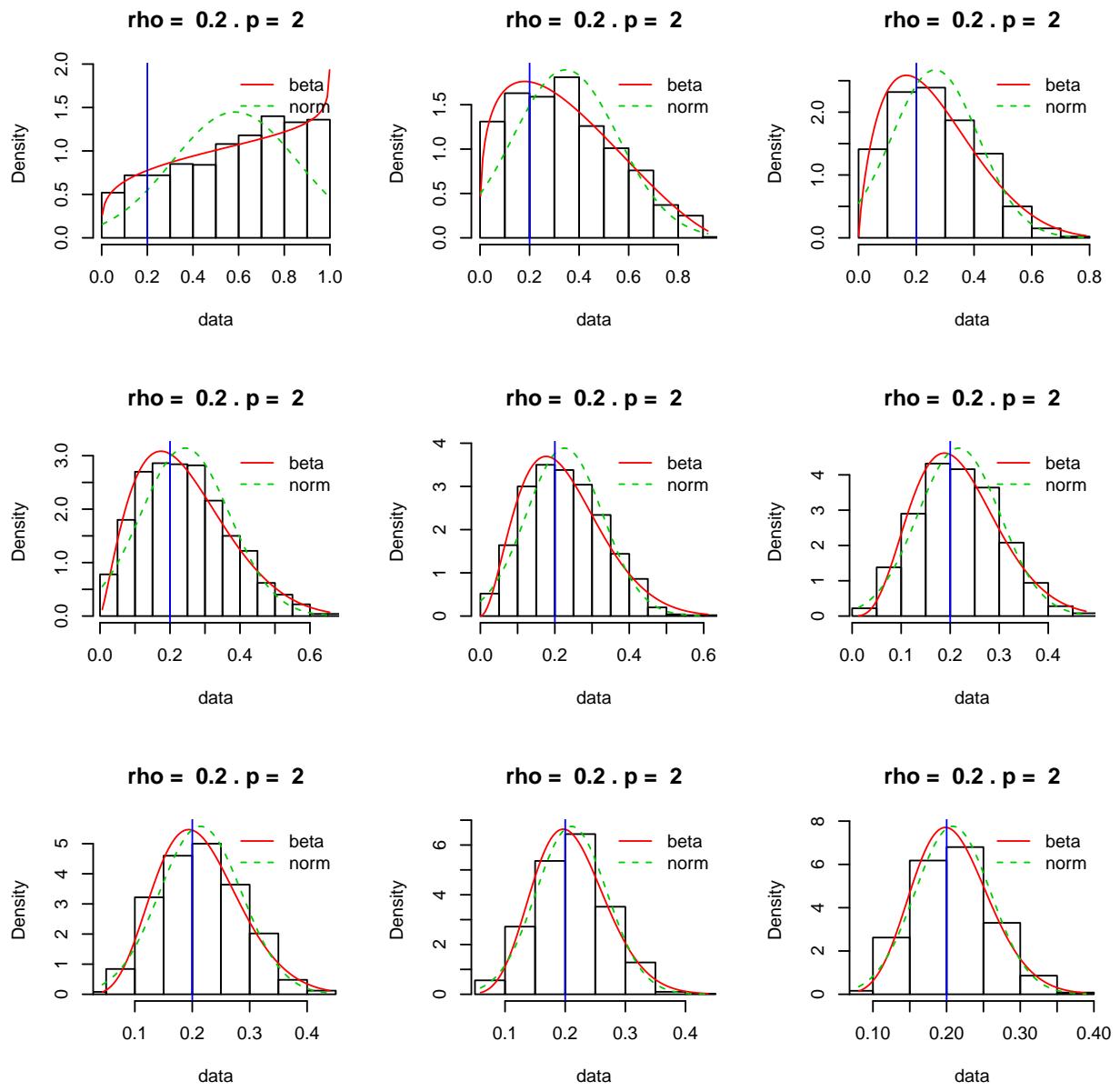
4.3 För ökande p

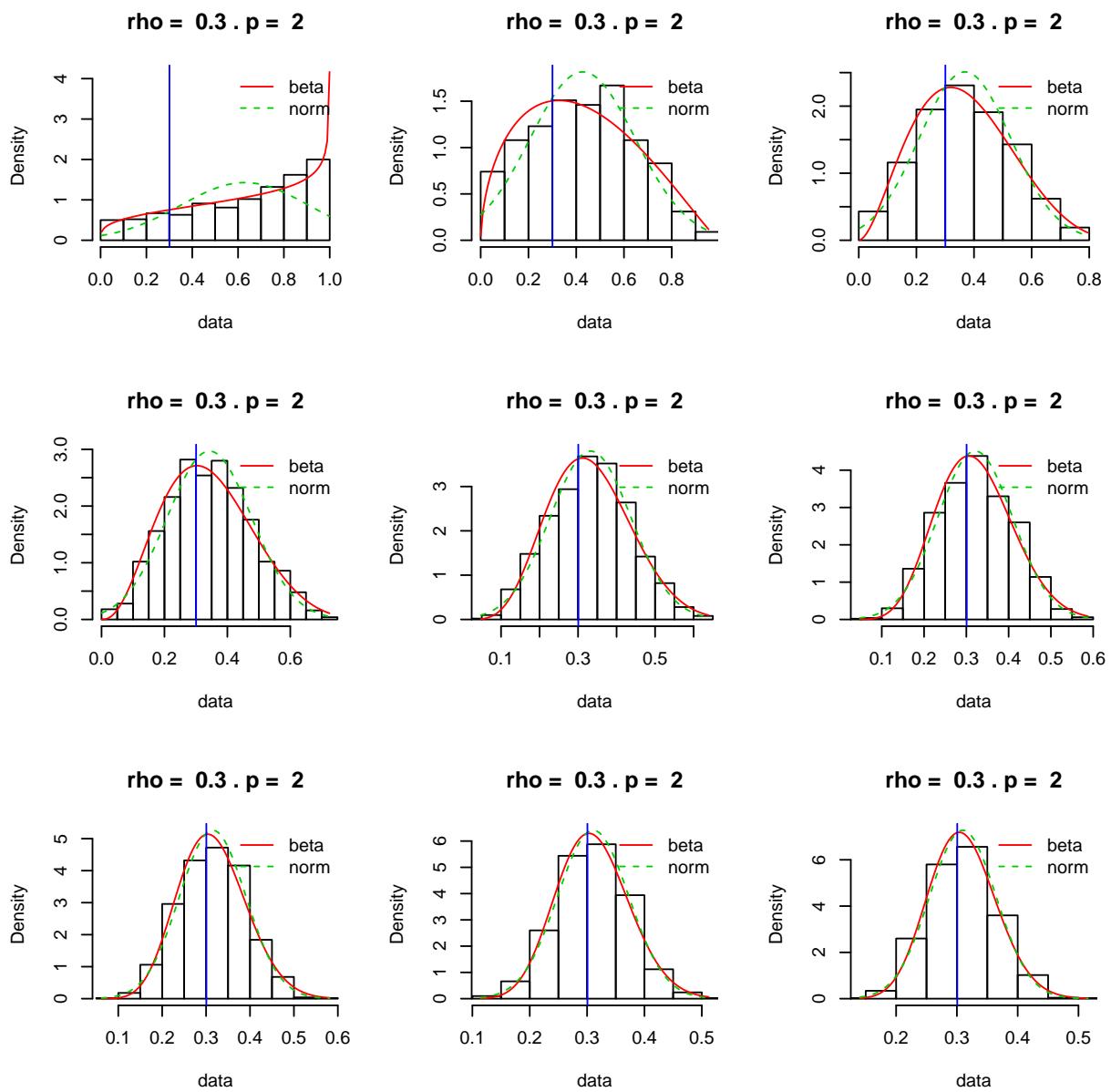
Vi vill också undersöka effekten om vi ökar antalet oberoende variabler (p)

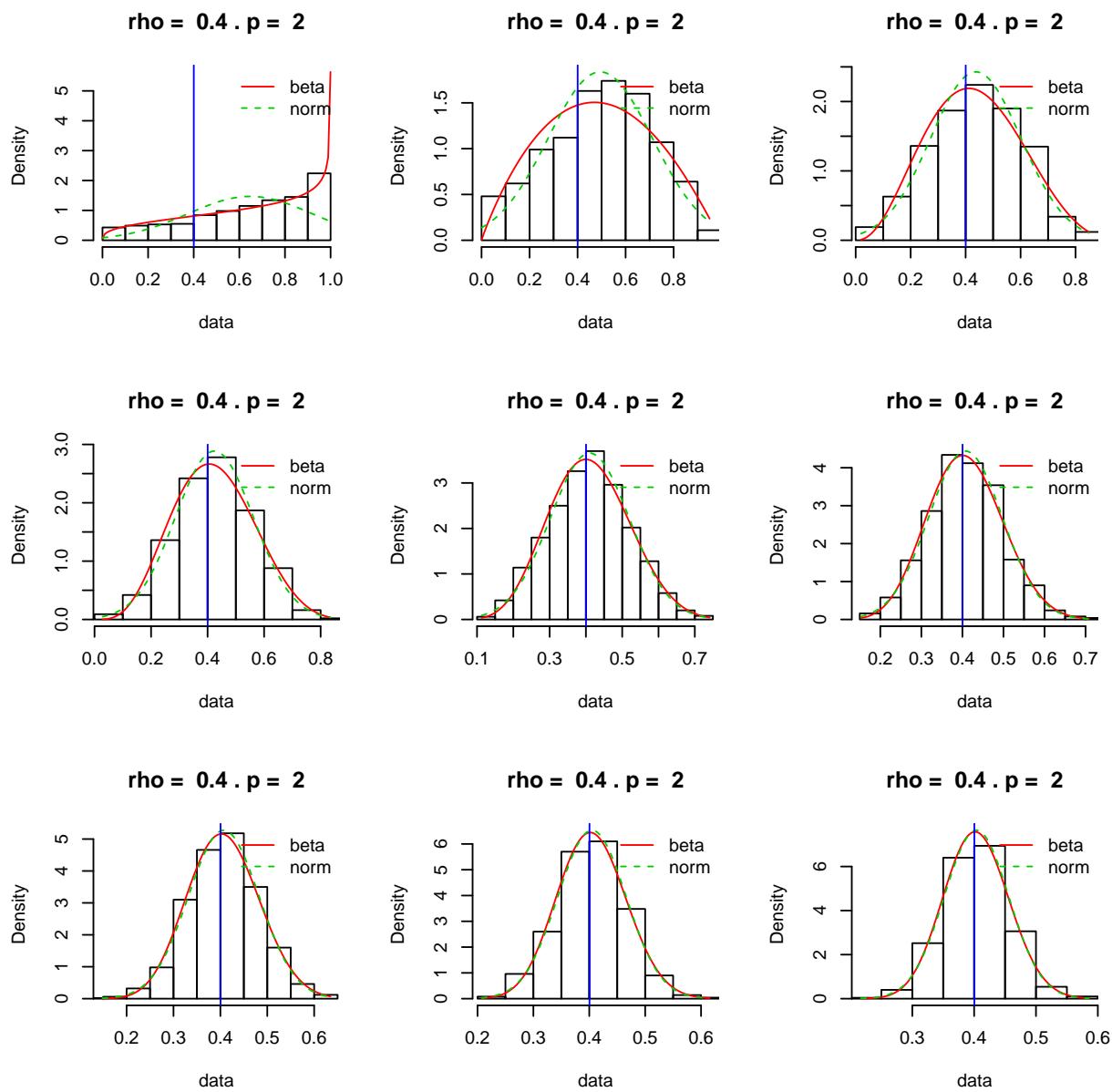
4.3.1 $p = 2$

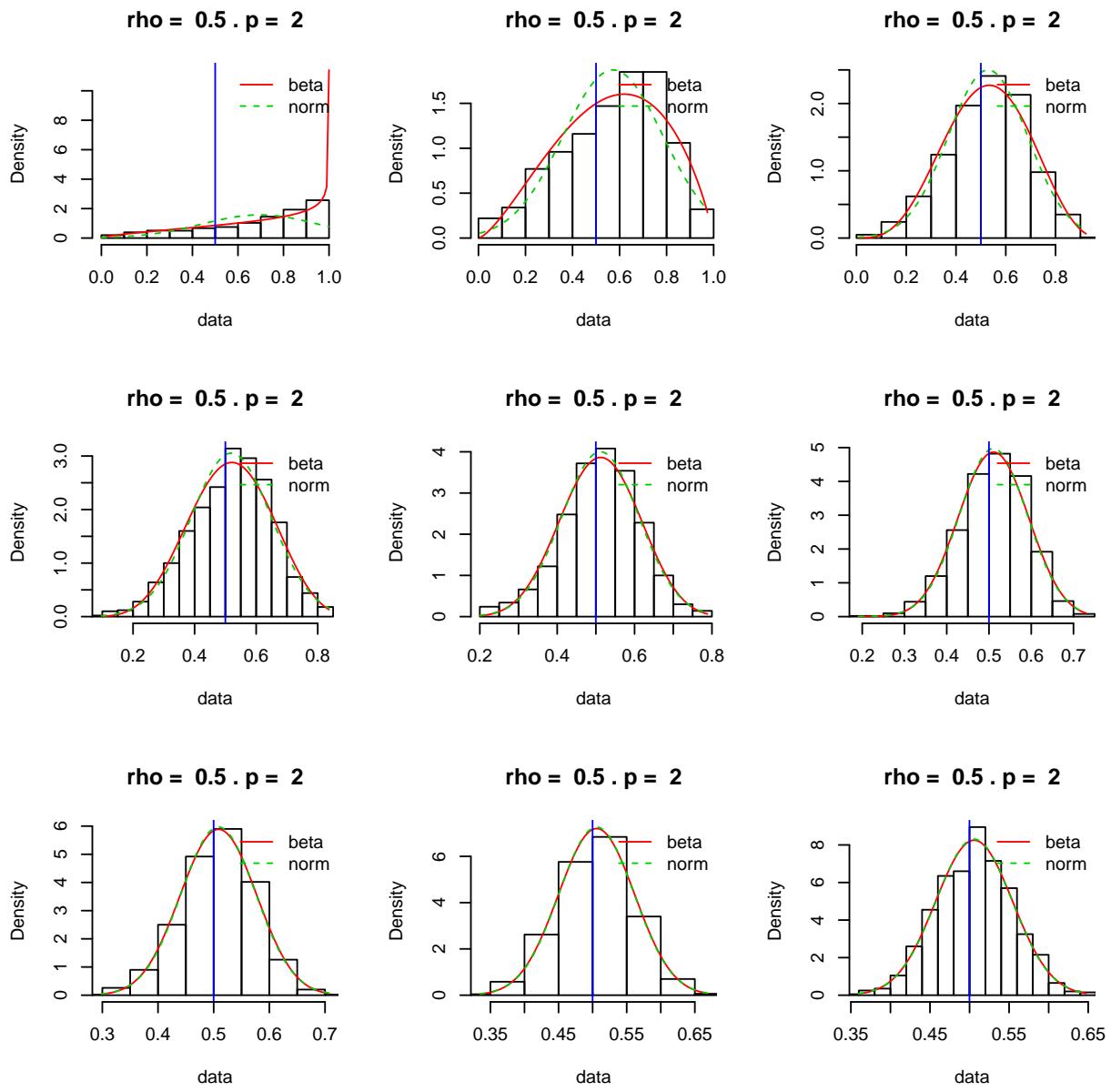
```
lapply(seq(0, .9, .1), distplots, p = 2)
```

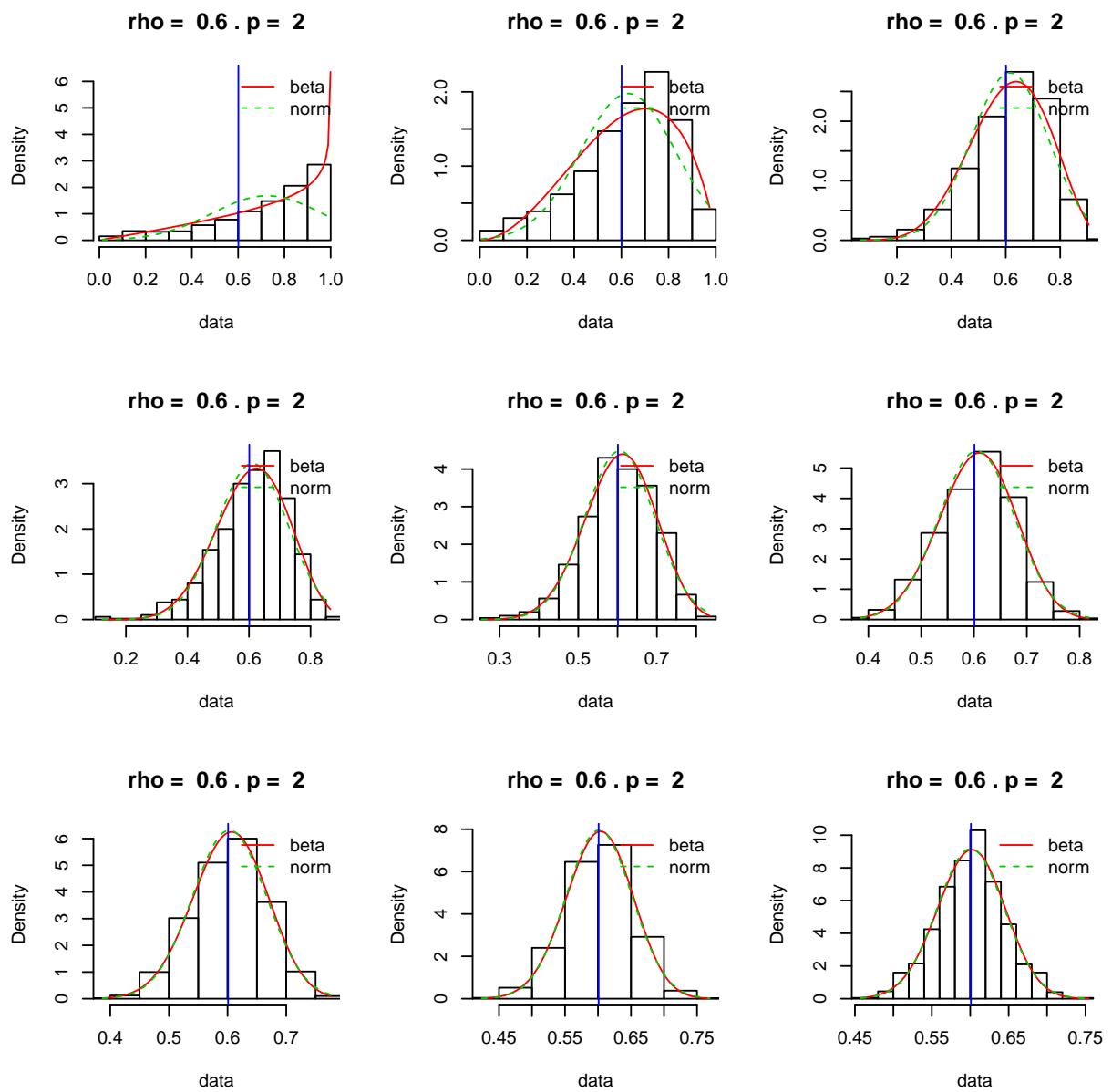


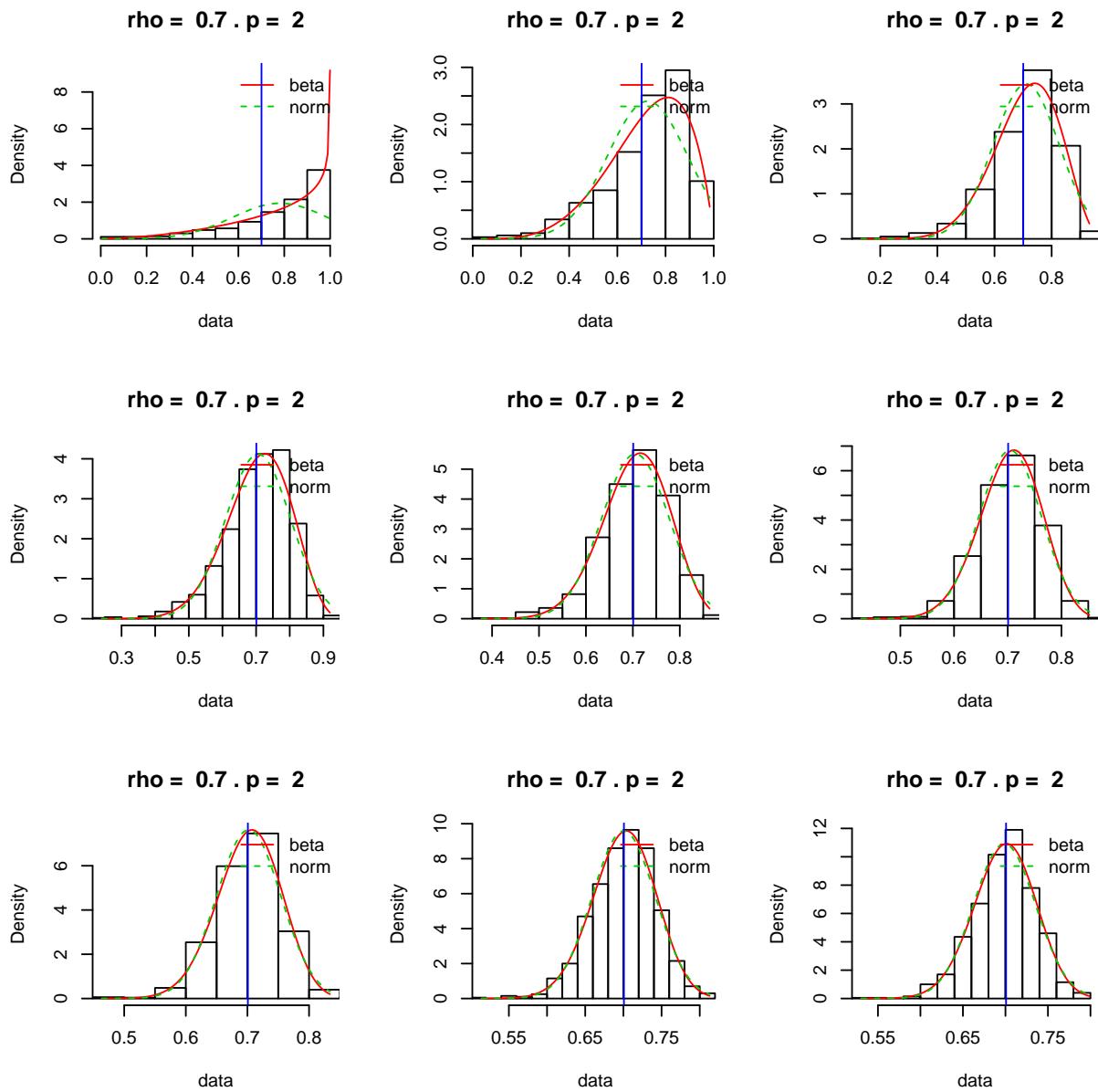


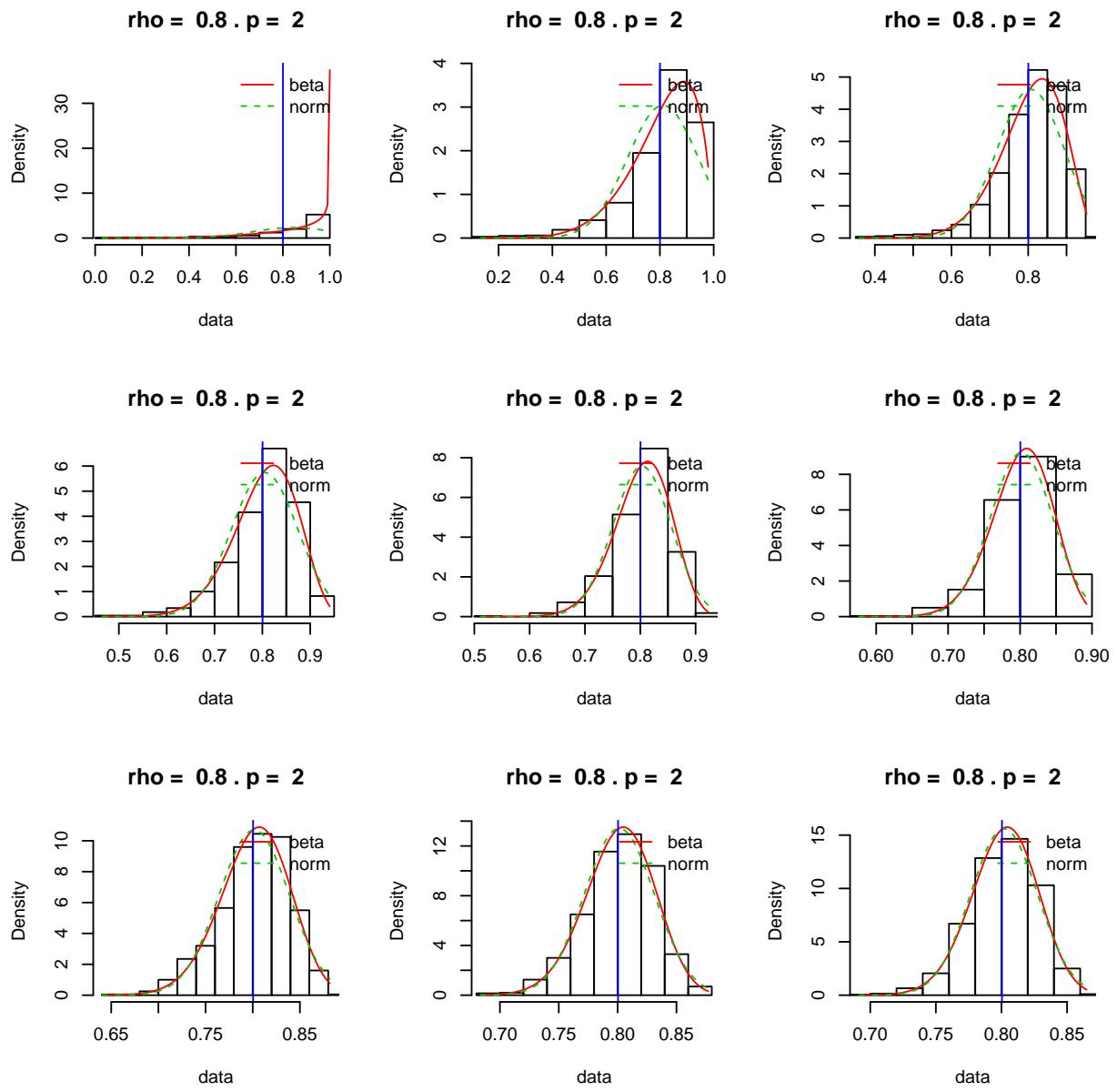


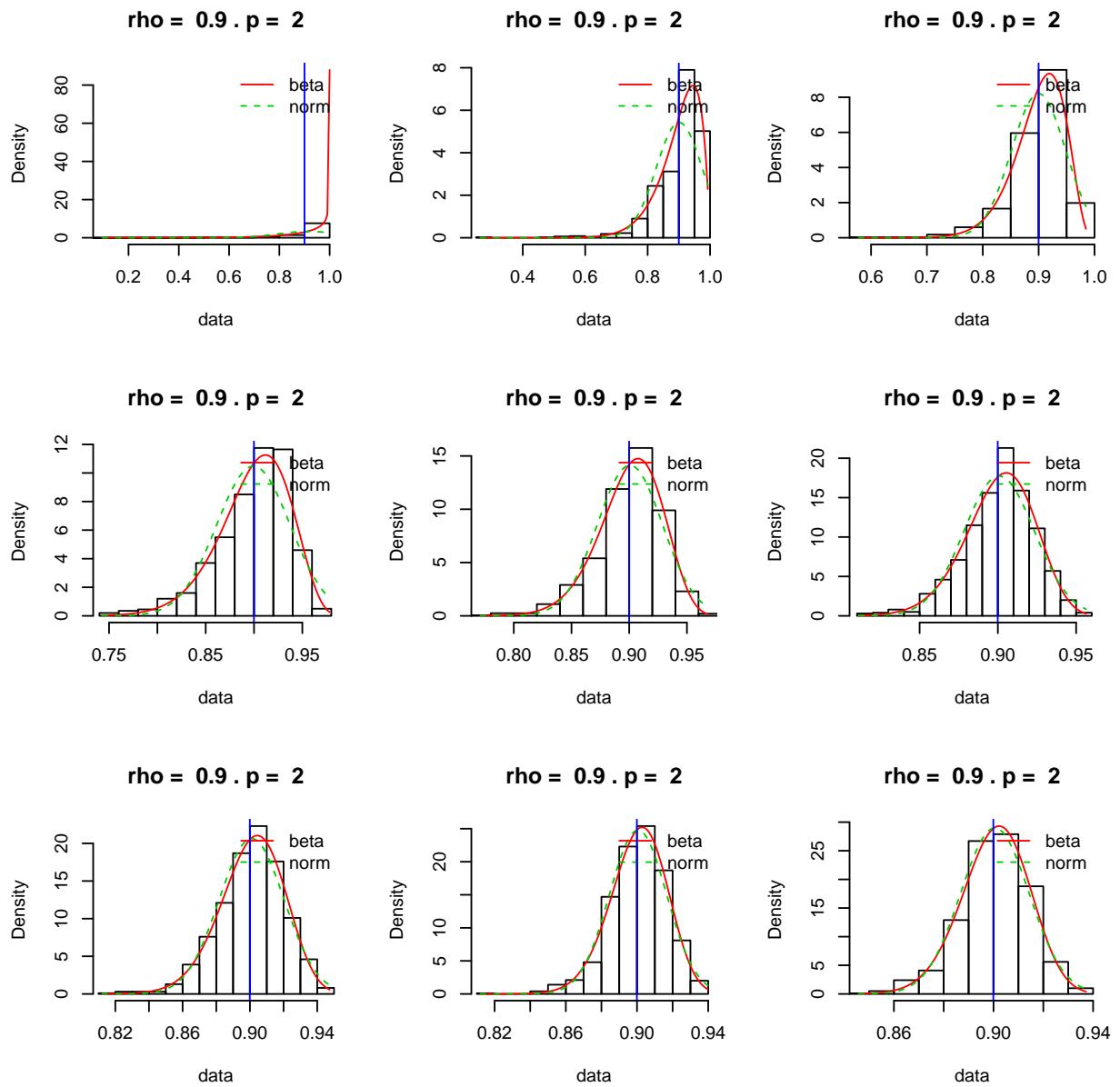






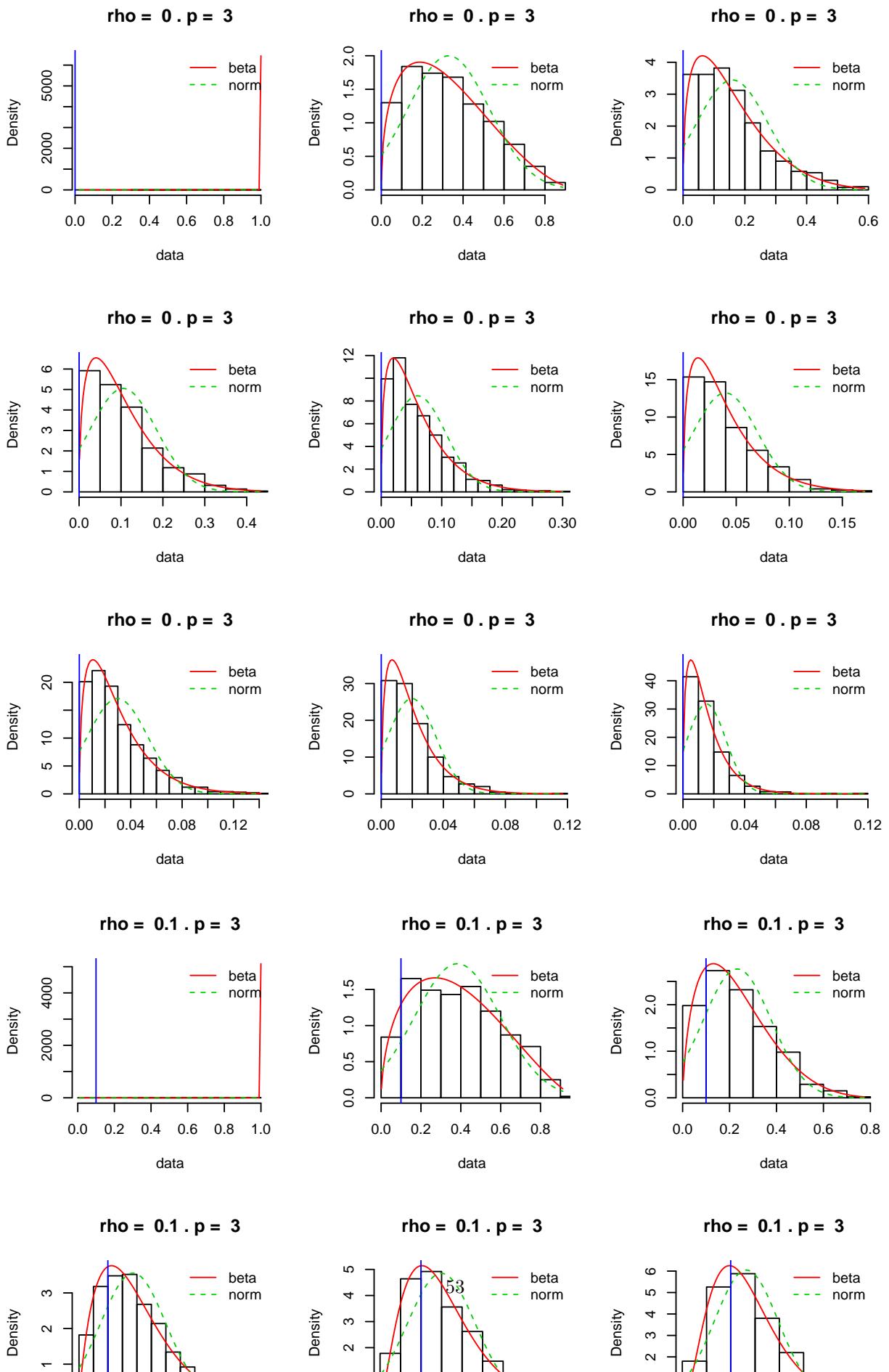


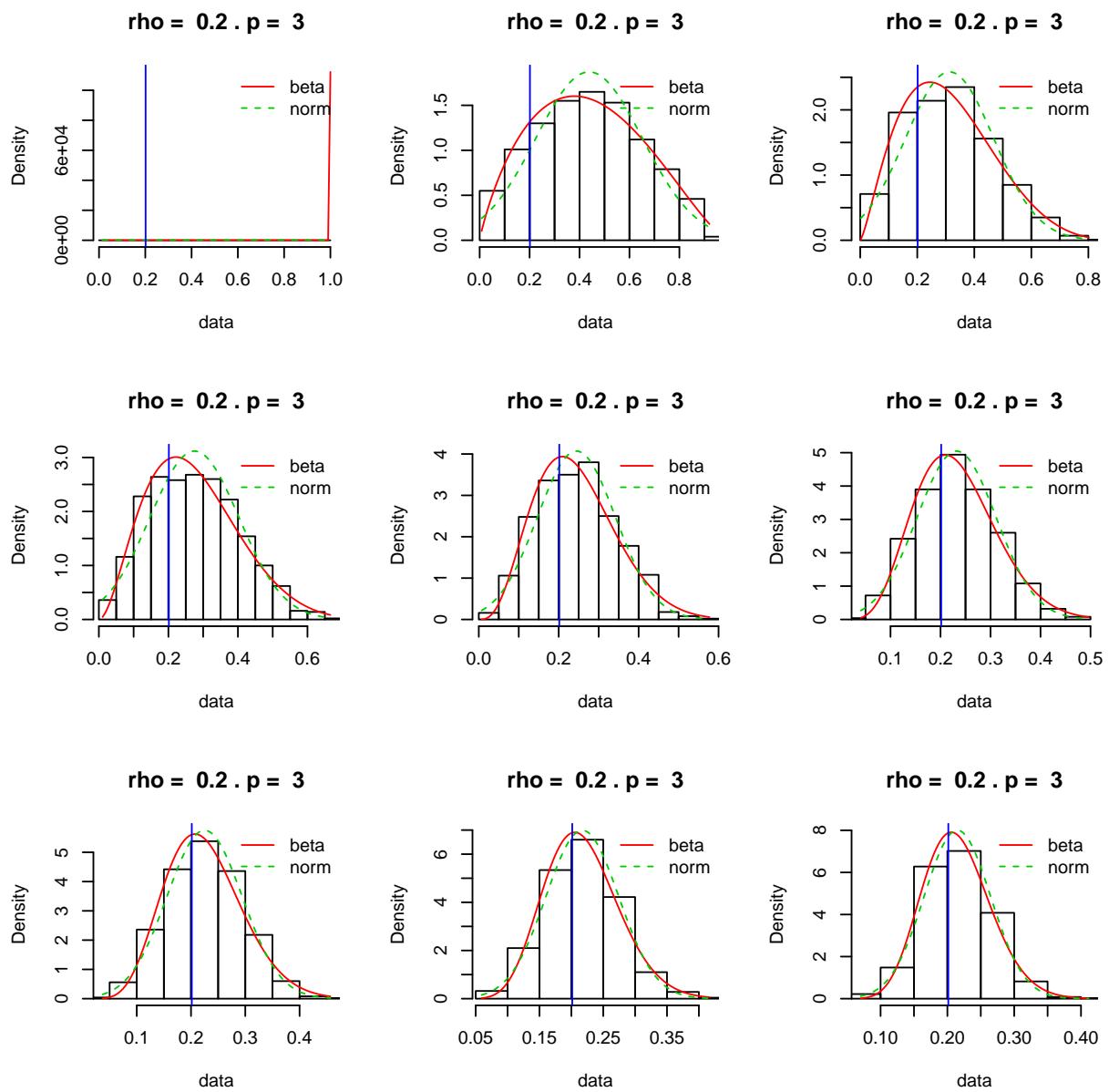


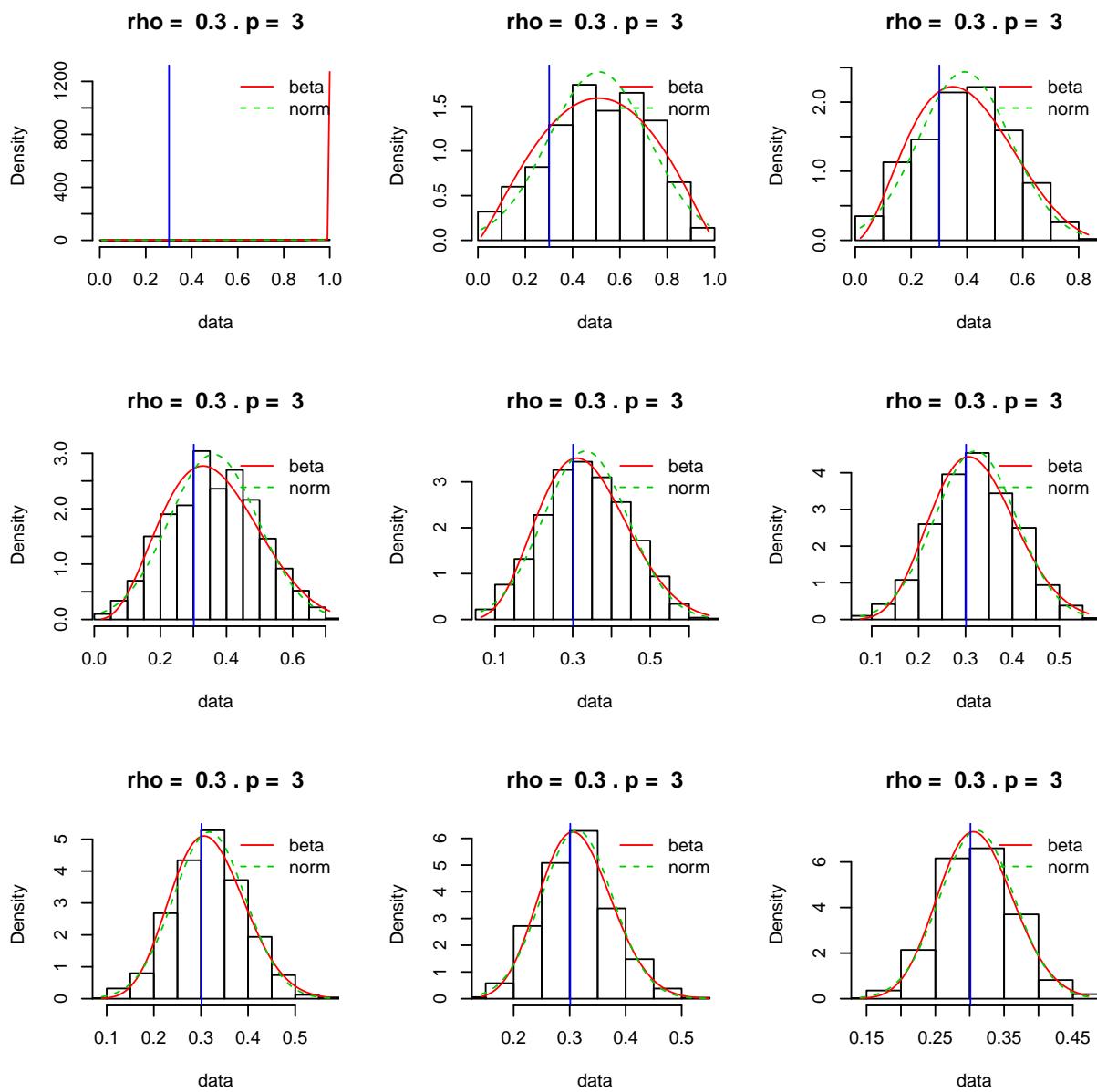


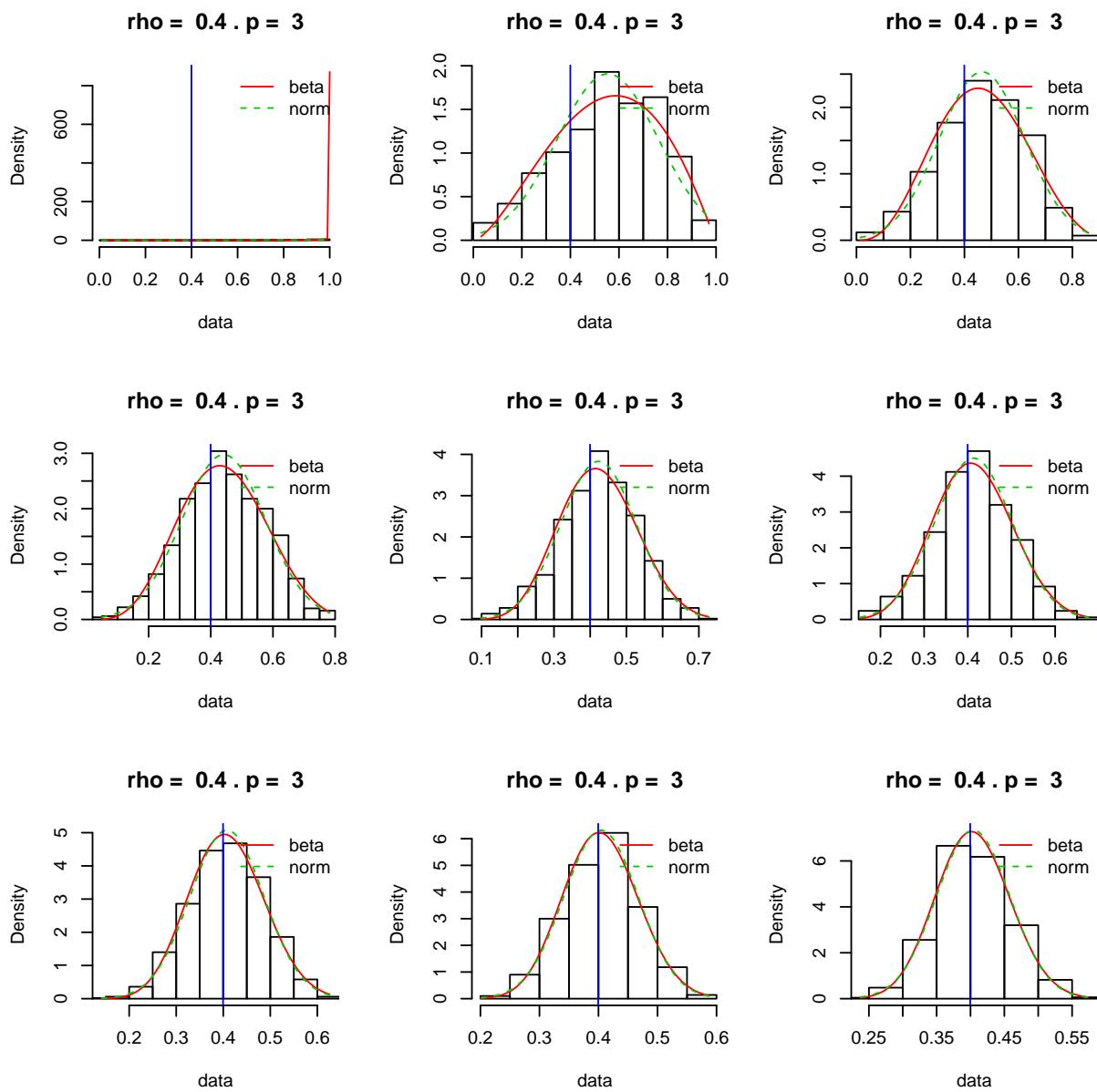
4.3.2 p = 3

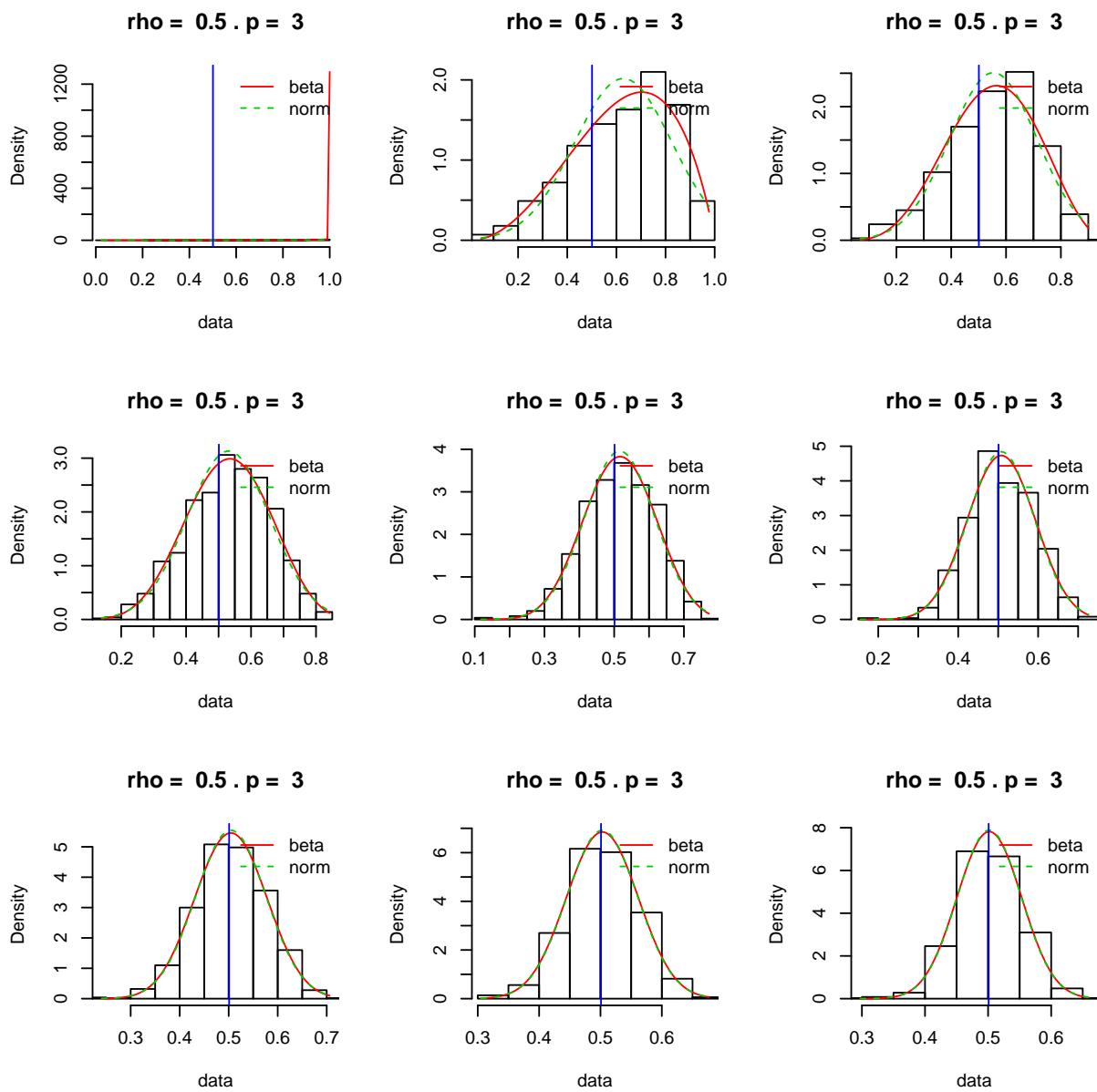
```
lapply(seq(0, .9, .1), distplots, p = 3)
```

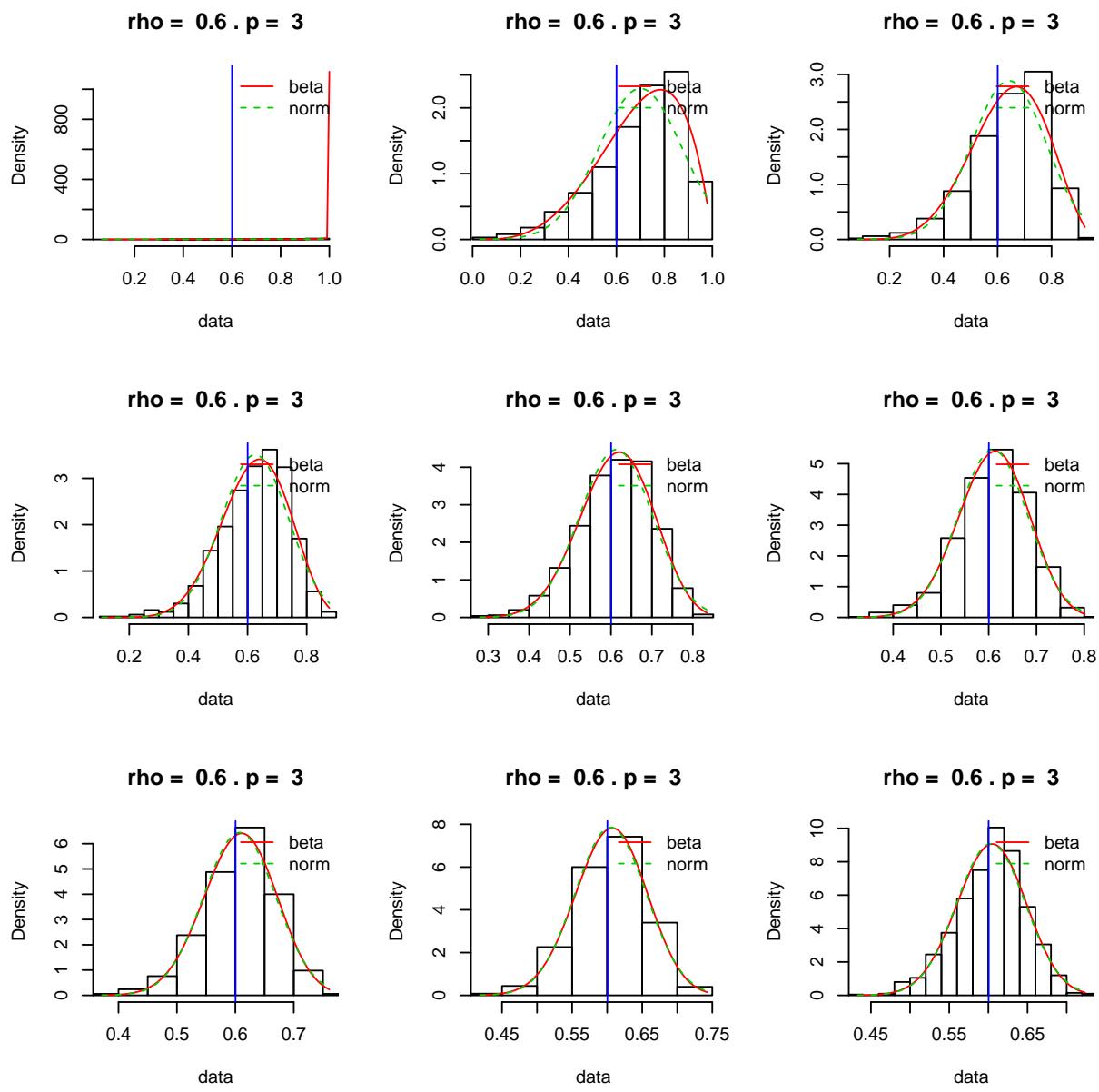


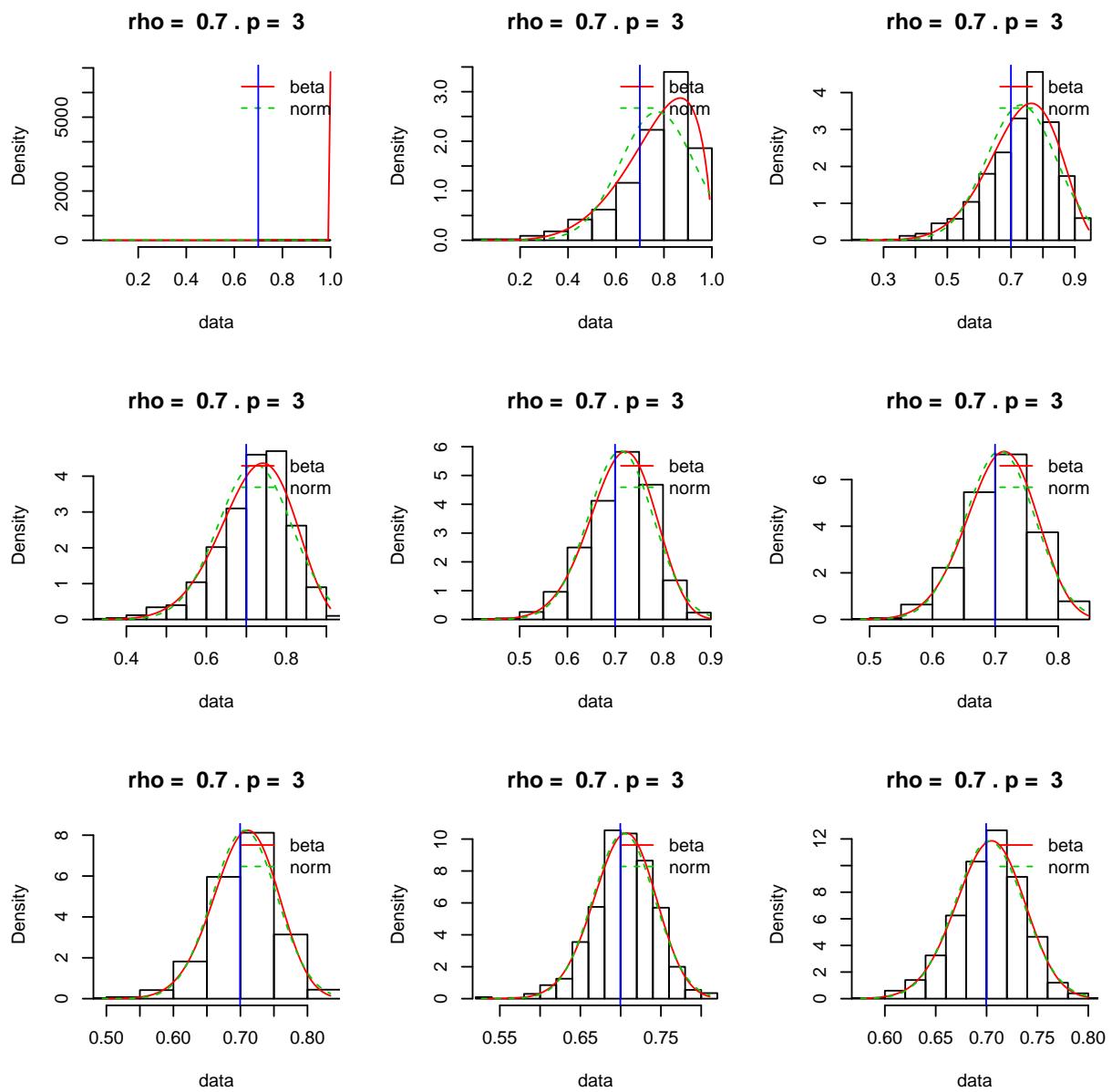


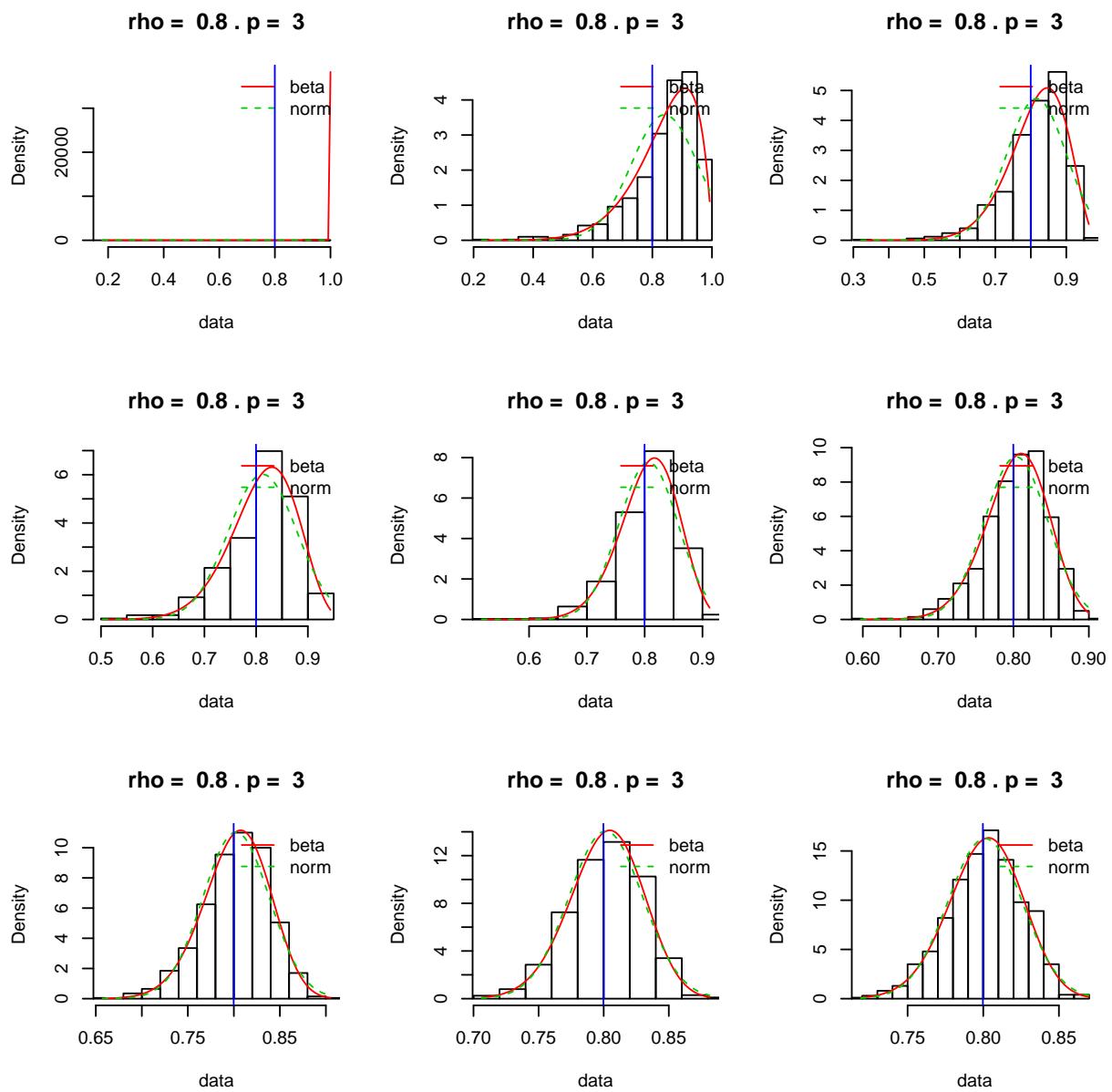


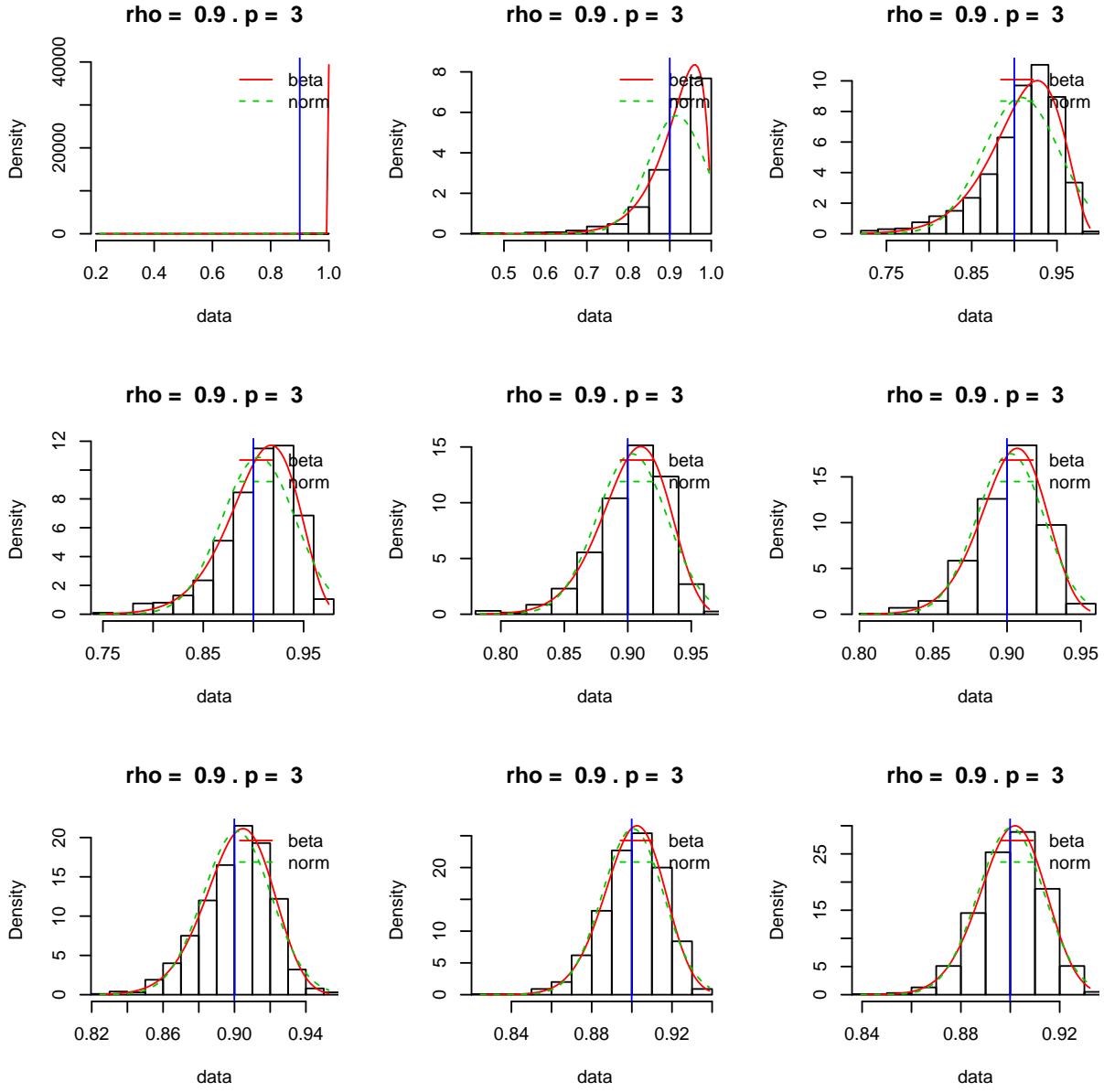












4.3.3 $p > 3$

När jag provar med större p får jag konvergensproblem av optim enl kod 100 på ?mle

Sedan är det förstås viktigt att också titta på CDF, QQ och PP-plottar där QQ illustrerar lack of fit i svansarna och PP i mitten (???). Funktionen `gofstat` kan också användas för att få fram en del teoretiska värden som beskriver goodness of fit. I så fall kan Anderson-Darling statistican vara bra att undersöka då den ger vikt till svansarna. Men finns också brister. AIC/BIC kan rekommenderas vid jämförelse mellan olika fördelningar.

Går också lätt att få konfidensintervall för parameterskattningarna gm bootstrap via funktionen `bootdist`.

4.4 Läsning av (???)

Väldigt lätt att t ex plotta en ickecentral betafördelning (här med defaultparametervärden).

```
library(distrMod)

## Loading required package: distr

## Loading required package: startupmsg

## Utilities for start-up messages (version 0.9)

## For more information see ?"startupmsg", NEWS("startupmsg")

## Loading required package: sfsmisc

##
## Attaching package: 'sfsmisc'

## The following object is masked from 'package:dplyr':
## 
##     last

## Loading required package: SWeaveListingUtils

## Utilities for Sweave together with TeX listings package (version 0.6.2)

## NOTE: Support for this package will stop soon.
## Package 'knitr' is providing the same functionality in a better way.
## Some functions from package 'base' are intentionally masked ---see SWeaveListingMASK().
## Note that global options are controlled by SWeaveListingoptions() ---c.f. ?"SWeaveListingoptions".

## For more information see ?"SWeaveListingUtils", NEWS("SWeaveListingUtils")
## There is a vignette to this package; try vignette("ExampleSWeaveListingUtils").

##
## Attaching package: 'SWeaveListingUtils'

## The following objects are masked from 'package:base':
## 
##     library, require

## Object oriented implementation of distributions (version 2.5.3)

## Attention: Arithmetics on distribution objects are understood as operations on corresponding random variables.
## Some functions from package 'stats' are intentionally masked ---see distrMASK().
## Note that global options are controlled by distroptions() ---c.f. ?"distroptions".

## For more information see ?"distr", NEWS("distr"), as well as
## http://distr.r-forge.r-project.org/
## Package "distrDoc" provides a vignette to this package as well as to several extension packages; try
```

```

## 
## Attaching package: 'distr'

## The following objects are masked from 'package:dplyr':
## 
##     location, n

## The following objects are masked from 'package:stats':
## 
##     df, qqplot, sd

## Loading required package: distrEx

## Extensions of package distr (version 2.5)

## Note: Packages "e1071", "moments", "fBasics" should be attached /before/ package "distrEx". See dist...
##       package "RobExtremes". See distrExMOVED().

## For more information see ?"distrEx", NEWS("distrEx"), as well as
##   http://distr.r-forge.r-project.org/
## Package "distrDoc" provides a vignette to this package as well as to several related packages; try v...

## 
## Attaching package: 'distrEx'

## The following objects are masked from 'package:stats':
## 
##     IQR, mad, median, var

## Loading required package: RandVar

## Implementation of random variables (version 0.9.2)

## For more information see ?"RandVar", NEWS("RandVar"), as well as
##   http://robast.r-forge.r-project.org/
## This package also includes a vignette; try vignette("RandVar").

## Loading required package: stats4

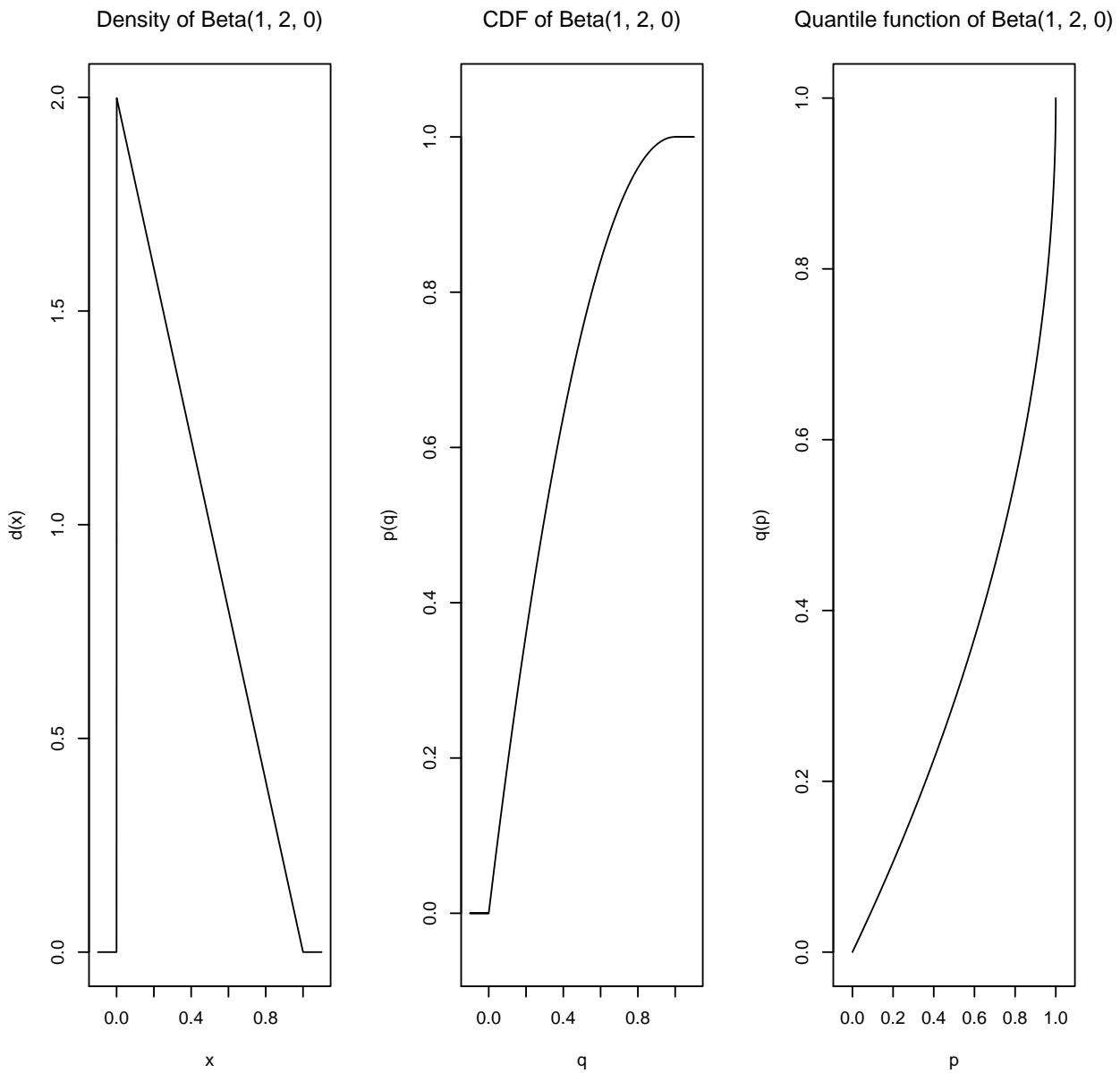
## Object Oriented Implementation of Probability Models (version 2.5.3)

## Some functions from pkg's 'base' and 'stats' are intentionally masked ---see distrModMASK().
## Note that global options are controlled by distrModoptions() ---c.f. ?"distrModoptions".

## For more information see ?"distrMod", NEWS("distrMod"), as well as
##   http://distr.r-forge.r-project.org/
## There is a vignette to this package; try vignette("distrMod").
## 
## Package "distrDoc" provides a vignette to the other distrXXX packages,
## as well as to several related packages; try vignette("distr").

```

```
##  
## Attaching package: 'distrMod'  
  
## The following object is masked from 'package:stats4':  
##  
##     confint  
  
## The following object is masked from 'package:stats':  
##  
##     confint  
  
## The following object is masked from 'package:base':  
##  
##     norm  
  
B <- Beta()  
plot(B)
```



Skattningar sker via **MCEstimator** (**MLEstimator**) men för detta krävs en familj att skatta emot. Det finns en **BetaFamily** men den omfattar inte ickecentral beta. Man kan definiera en egen familj via **L2ParamFamily** (se t ex källkoden för **BetaFamily**). Dock är jag osäker på argumentet **L2deriv.fct**. Det ska bestå av tre funktioner i en lista där varje funktion av x beskriver vänsterledet i $\frac{\partial \alpha}{\partial f} \ln \hat{f} = 0$ där f är fördelningsfunktionen för beta, dvs ett uttryck för att finna maximum likelihood-estimatet (digamma = derivatan av gammafördelningen). Dessa funktioner går att beräkna analytiskt för betafördelningen men går det för ickecentral beta? Fördelningsfunktionen finns här: https://en.wikipedia.org/wiki/Noncentral_beta_distribution#Probability_density_function

Om vi sedan deriverar Betafunktionen finns det uttrycket här: https://en.wikipedia.org/wiki/Beta_function#Derivatives

Referenser

- Cowden, Dudley J. 1952. "The Multiple-Partial Correlation Coefficient." *Journal of the American Statistical Association* 47 (259): 442–56. doi:[10.1080/01621459.1952.10501183](https://doi.org/10.1080/01621459.1952.10501183).
- Hogben, David. 1968. "The distribution of the sample correlation coefficient with one variable fixed." *Journal of Research of the National Bureau of Standards, Section B: Mathematical Sciences* 72B (1): 33. doi:[10.6028/jres.072B.007](https://doi.org/10.6028/jres.072B.007).
- Hotelling, Harold. 1953. "New Light on the Correlation Coefficient and its Transforms Author(s): Harold Hotelling." *Journal of the Royal Statistical Society. Series B (Methodological)*, 15 (2): 296–193–232.
- Kymn, Kern O . 1968. "The Distribution of the Sample Correlation Coefficient Under the Null Hypothesis." *Econometrica* 36 (1): 187–89.
- Nemes, Szilard, Junmei Miao Jonasson, Anna Genell, and Gunnar Steineck. 2009. "Bias in odds ratios by logistic regression modelling and sample size." *BMC Medical Research Methodology* 9: 56. doi:[10.1186/1471-2288-9-56](https://doi.org/10.1186/1471-2288-9-56).
- Park, John H Jr. 1964. "Variations of the Non-central t and Beta Distributions." *The Annals of Mathematical Statistics* 35 (4): 1583–93.
- Pearson, Karl. 1895. "Note on Regression and Inheritance in the Case of Two Parents." *Proceedings of the Royal Society of London (1854-1905)* 58: 240–42. doi:[10.1098/rspl.1895.0041](https://doi.org/10.1098/rspl.1895.0041).
- . 1896. "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia." *Philosophical Transactions of the Royal Society A* 187.
- Warren, W. G. 1971. "Correlation or Regression: Bias or Precision." *Applied Statistics* 20 (2): 148. doi:[10.2307/2346463](https://doi.org/10.2307/2346463).