Arbetslogg 2016 vecka 8

Erik Bulow

22 februari 2016

Contents

1 Förberedelser

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

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

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

2 2016-02-22

Läser (R. A. Fisher 1921) som är teoretisk och innehåller härledningar för fördelning av hans z. Står att där finns en negativ bias för små stickprovsstorlekar(s. 213) och att orsaken till detta är lätt att förstå: medelvärden av termer i avvikelsen ej oberoende och summeras.

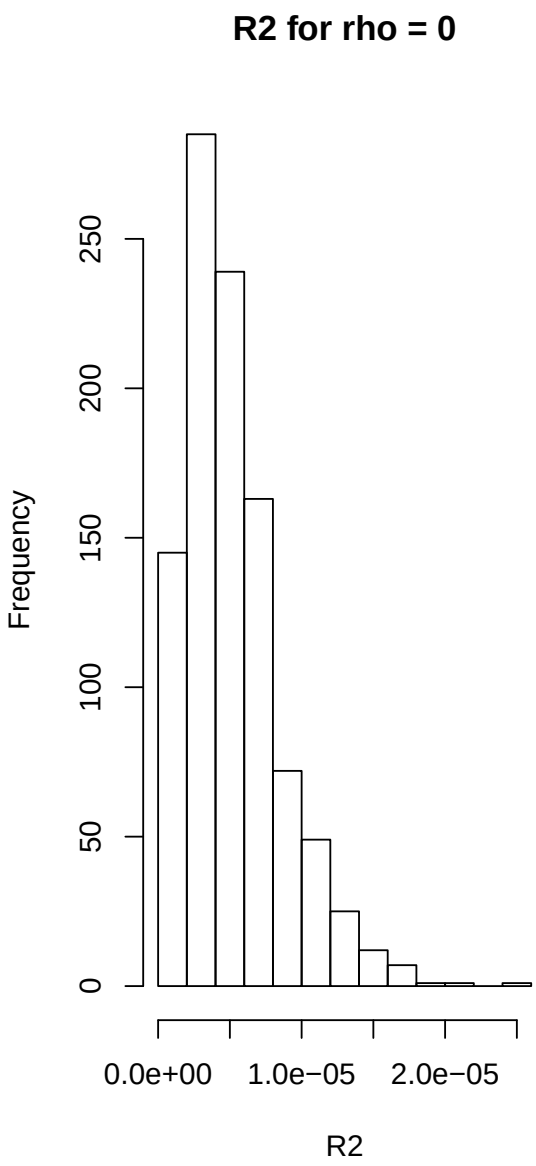
Sanna kurvan av z är ngt leptocurtic (smalare än vanlig normalfördelning), vilket leder till viss felskattning när normalapproximation används eftersom extremavvikelser blir vanligare(?).

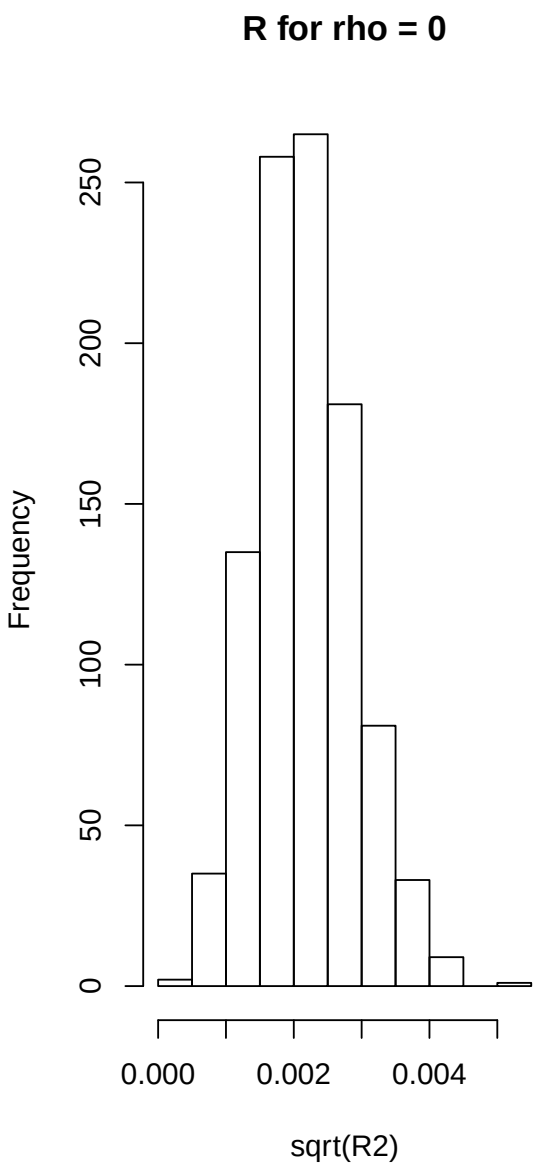
2.1 Skevare fördelning för ρ nära +-1

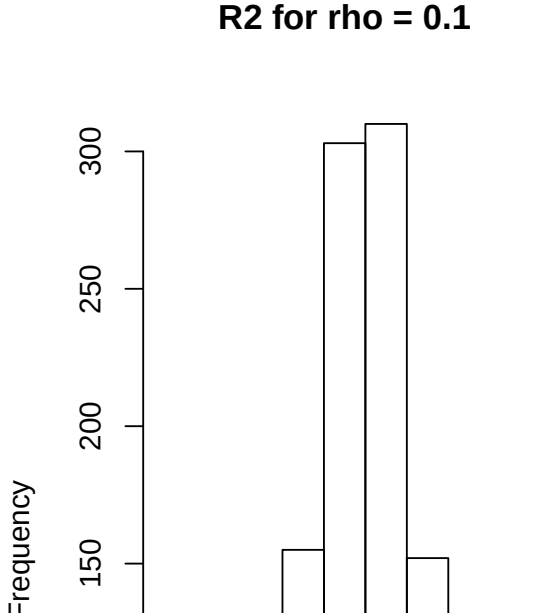
I 2016_w07 nämns att fördelningen av ρ blir extra skev nära 1. Vi kompletterar fördelningsgraf ovan (avs 2.2 2016.w07 förra veckan) (som avsåg $\rho = 0$) med graf för detta.

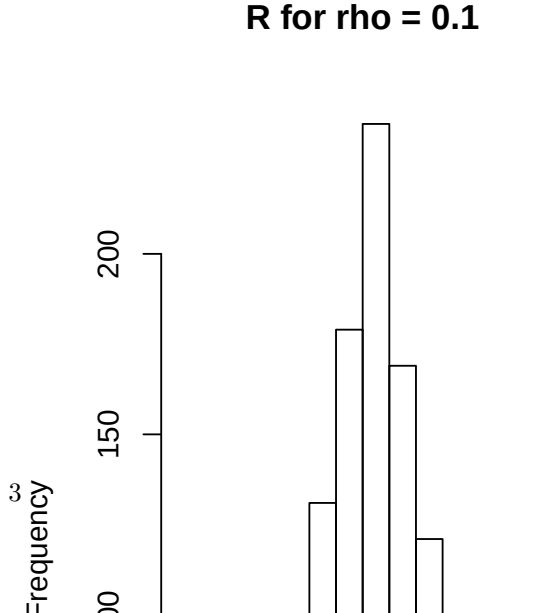
```
plot_emp_r2 <- function(r2) {
  R2 <- replicate(1000, summary(lm(Y ~., data = sim_data(r2)))$r.squared)
  par(mfrow = c(1, 2))
  hist(R2, main = paste("R2 for rho =", r2))
  hist(sqrt(R2), main = paste("R for rho =", r2))
}
```

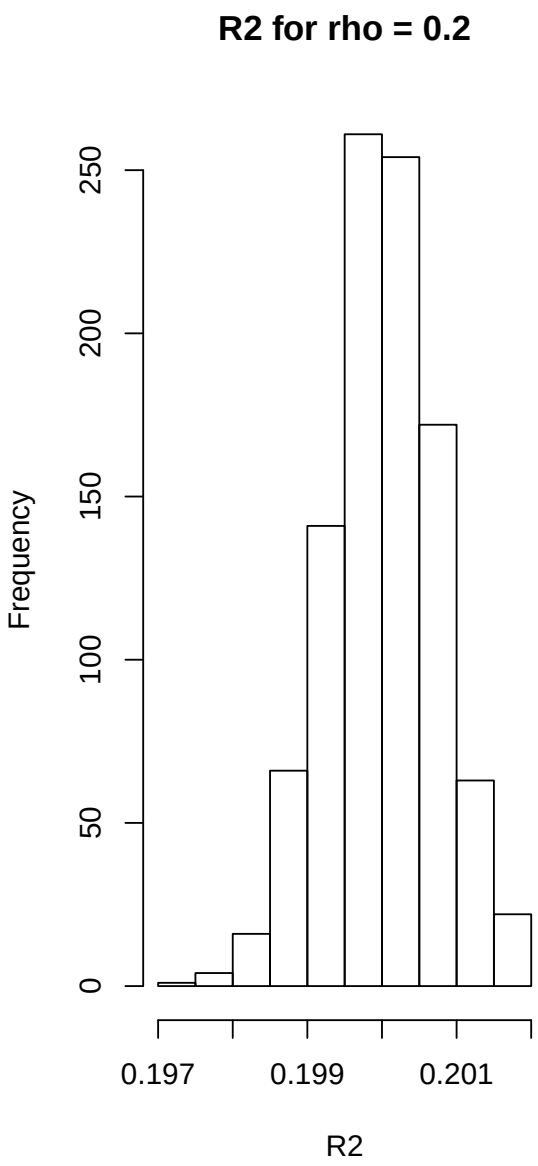
```
for (r2 in seq(0, 1, .1)) plot_emp_r2(r2)
```

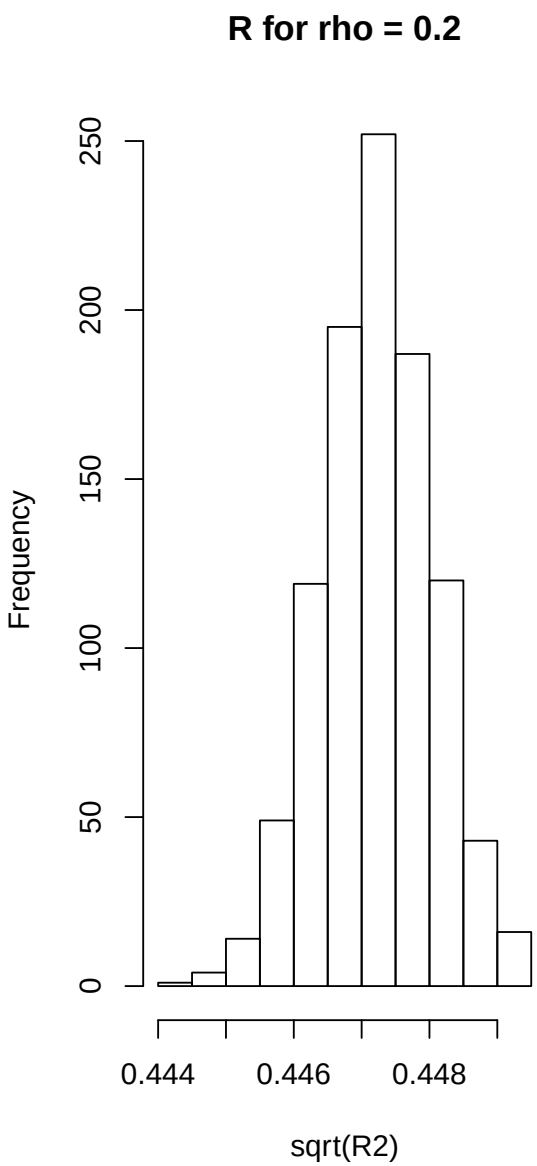

R2 for rho = 0
Frequency
0
50
100
150
200
250
0.0e+00
1.0e−05
2.0e−05
R2

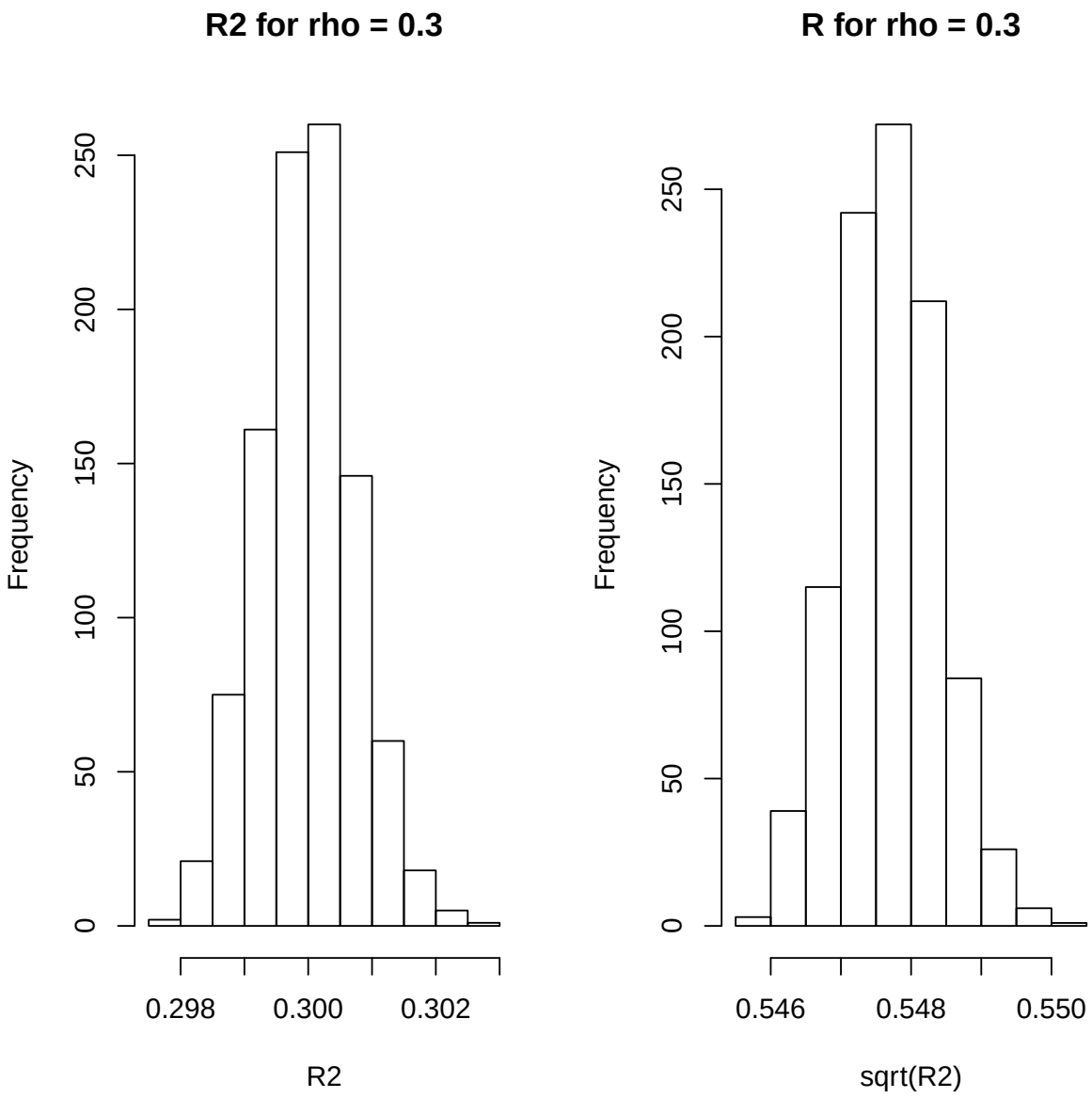

R for rho = 0
Frequency
0
50
100
150
200
250
0.000
0.002
0.004
sqrt(R2)

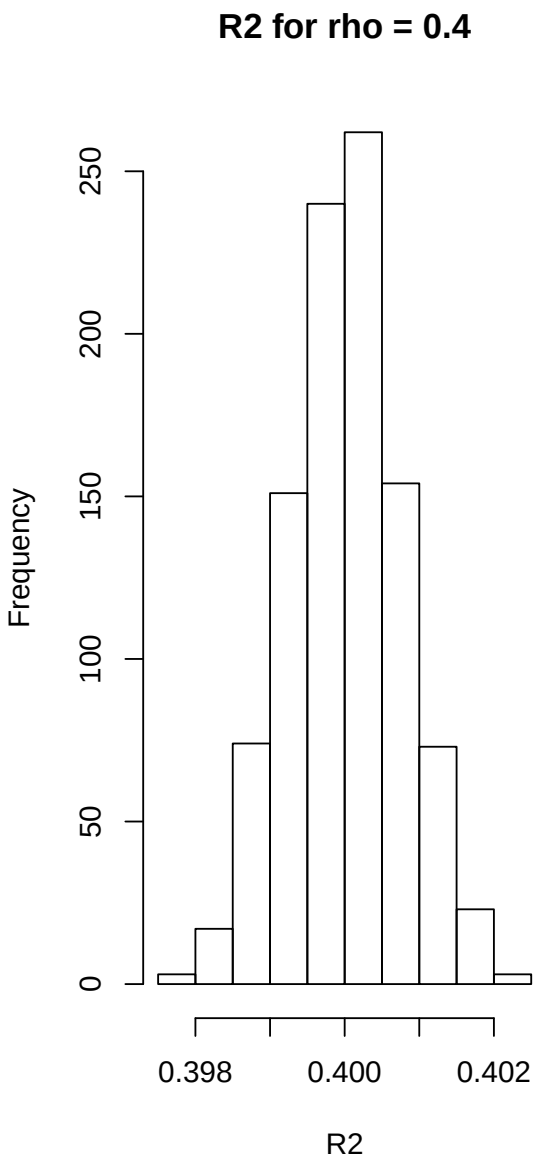

R2 for rho = 0.1
Frequency
150
200
250
300

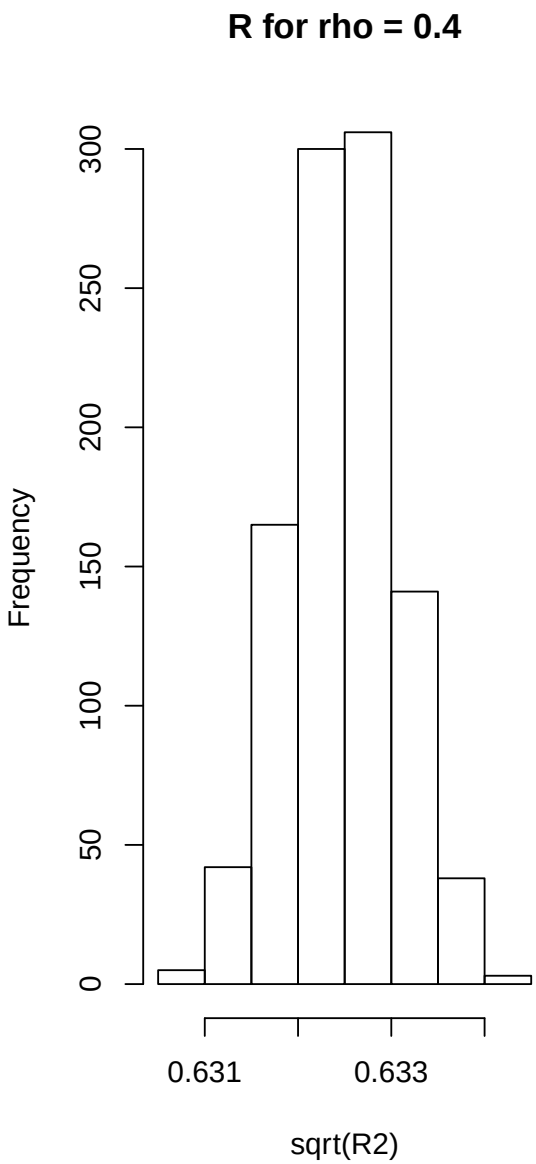

R for rho = 0.1
3
Frequency
150
200

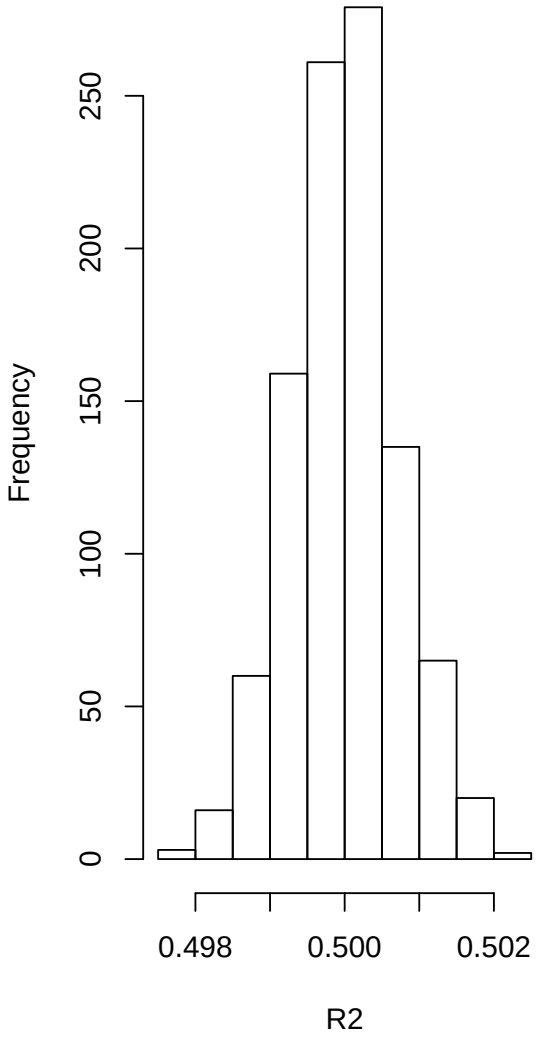

R2 for rho = 0.2
Frequency
0
50
100
150
200
250
0.197
0.199
0.201
R2

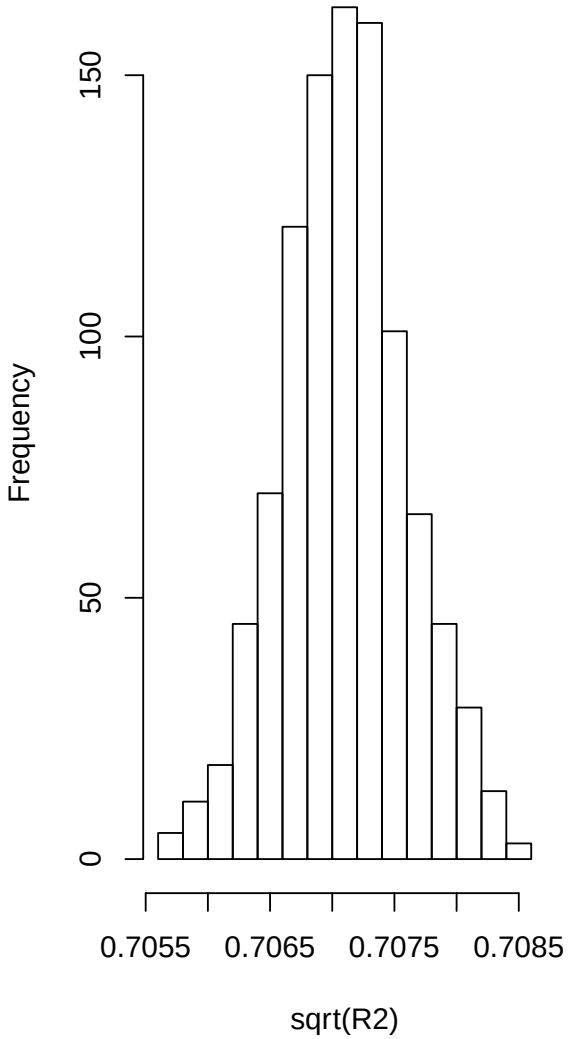

R for rho = 0.2
Frequency
0
50
100
150
200
250
0.444
0.446
0.448
sqrt(R2)

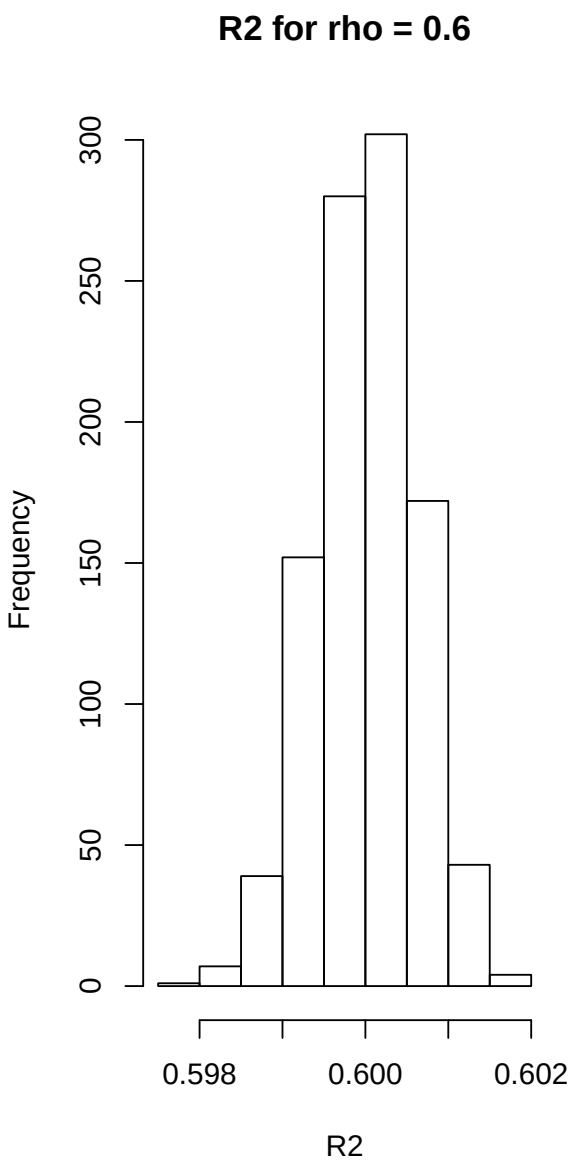

R2 for rho = 0.3
Frequency
0
50
100
150
200
250
0.298
0.300
0.302
R2
R for rho = 0.3
Frequency
0
50
100
150
200
250
0.546
0.548
0.550
sqrt(R2)

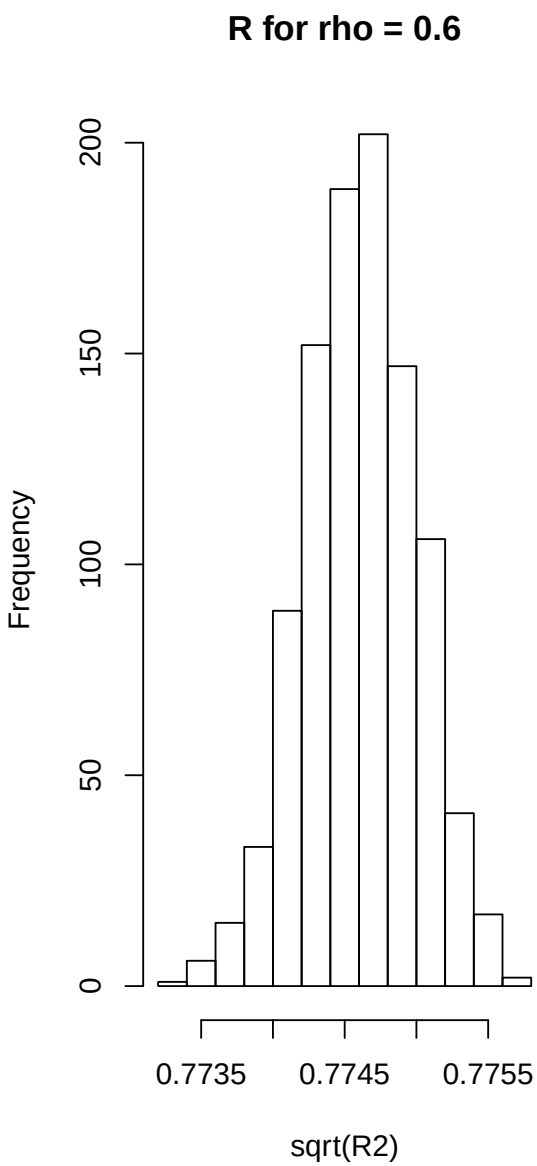

R2 for rho = 0.4
Frequency
0
50
100
150
200
250
0.398
0.400
0.402
R2

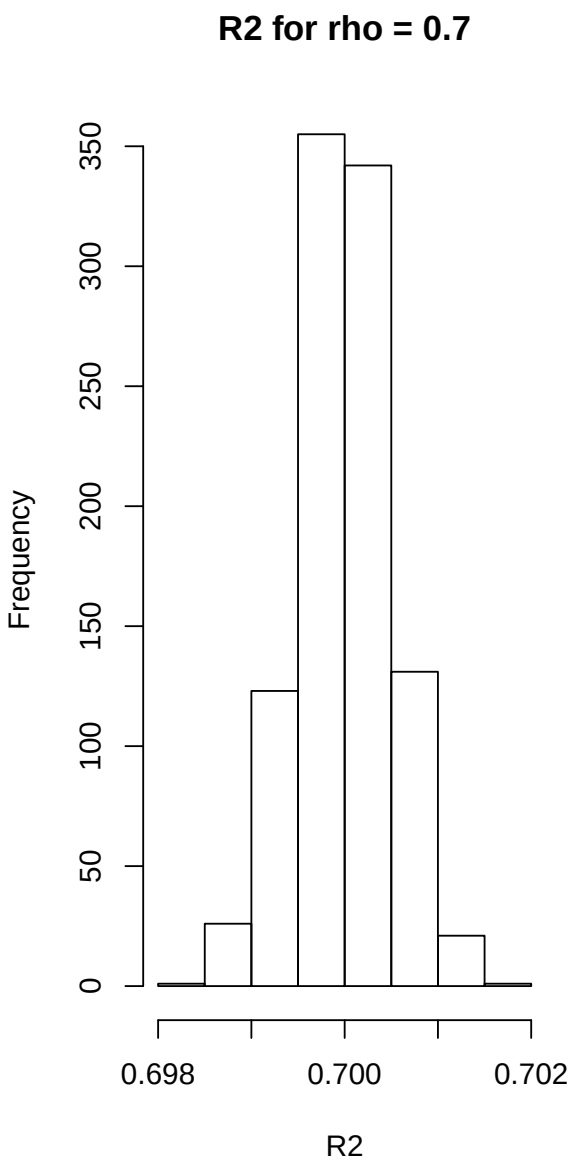

R for rho = 0.4
Frequency
0
50
100
150
200
250
300
0.631
0.633
sqrt(R2)

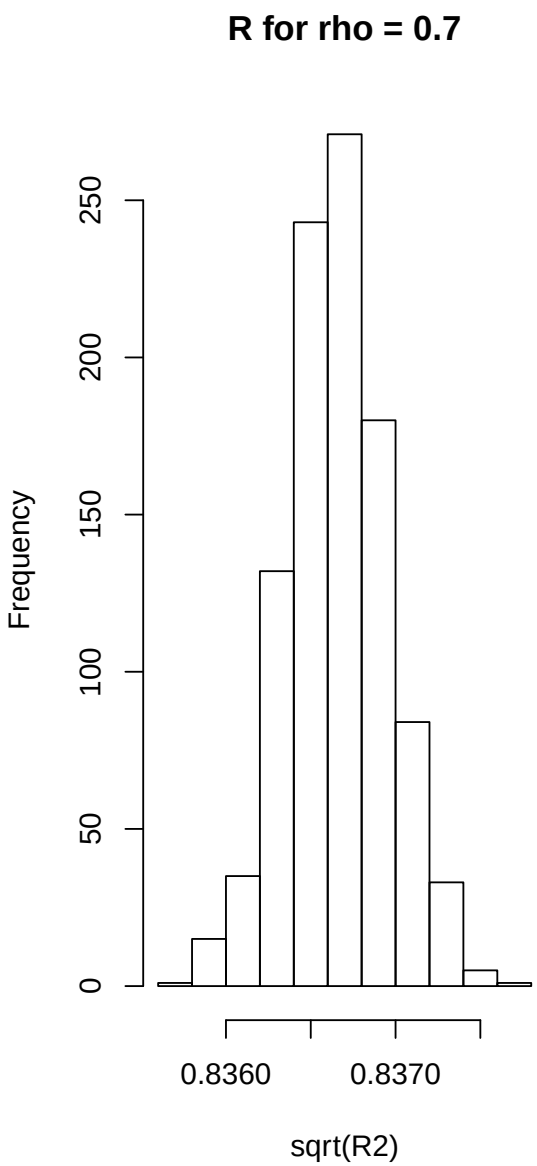

R2 for rho = 0.5
Frequency
0
50
100
150
200
250
0.498
0.500
0.502
R2

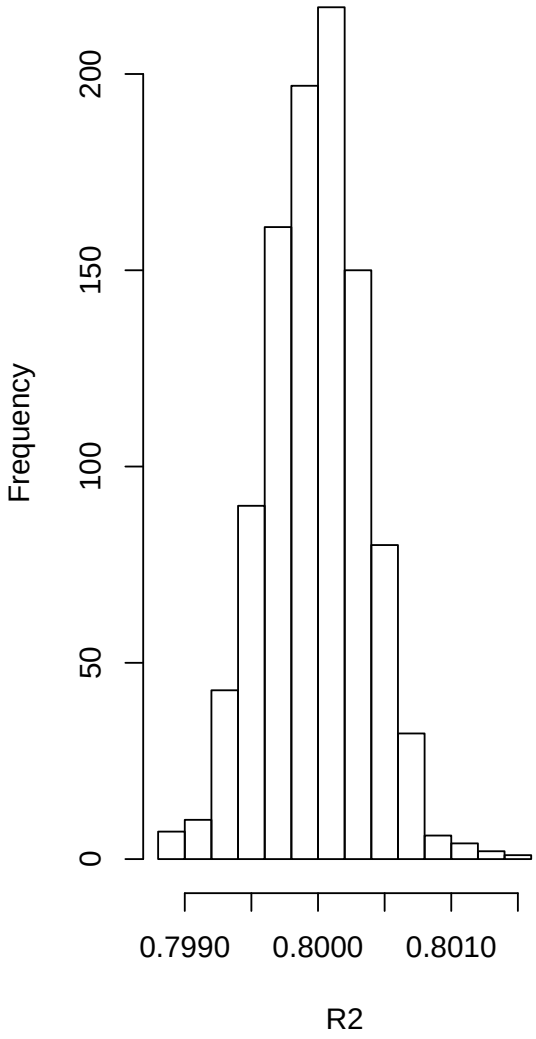

R for rho = 0.5
Frequency
0
50
100
150
0.7055
0.7065
0.7075
0.7085
sqrt(R2)

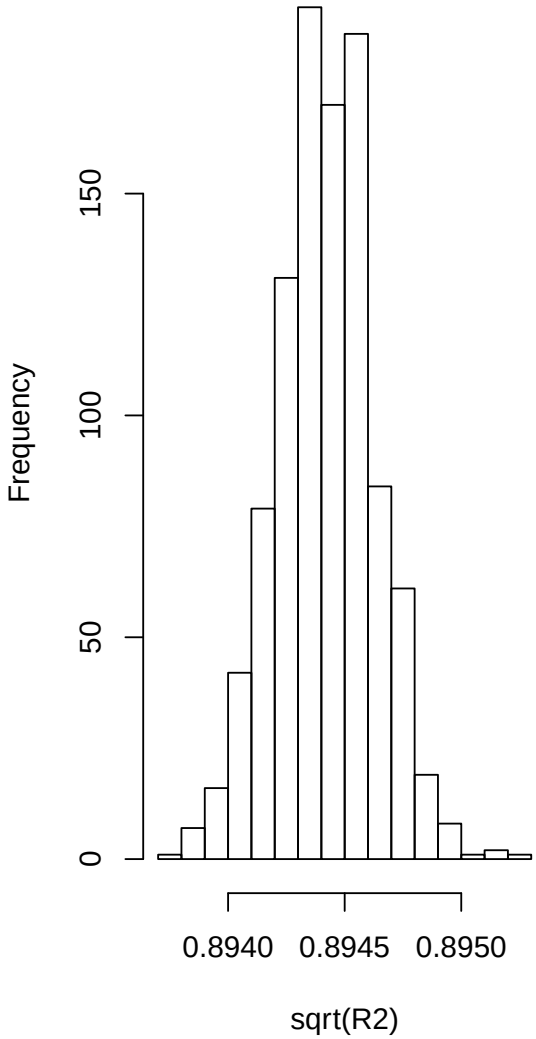

R2 for rho = 0.6
Frequency
0
50
100
150
200
250
300
0.598
0.600
0.602
R2

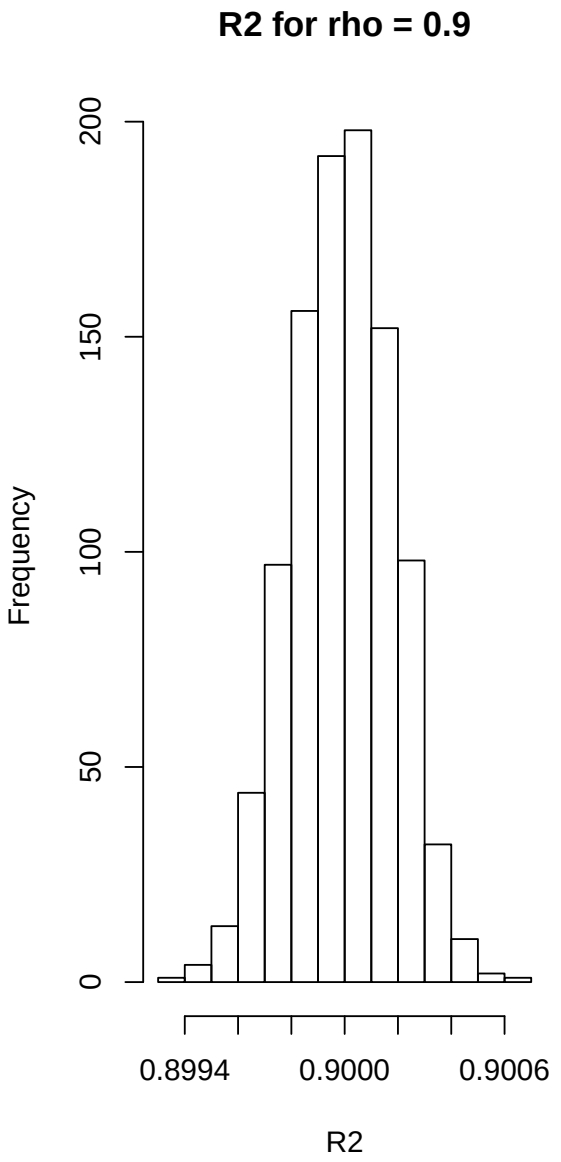

R for rho = 0.6
Frequency
0
50
100
150
200
0.7735
0.7745
0.7755
sqrt(R2)

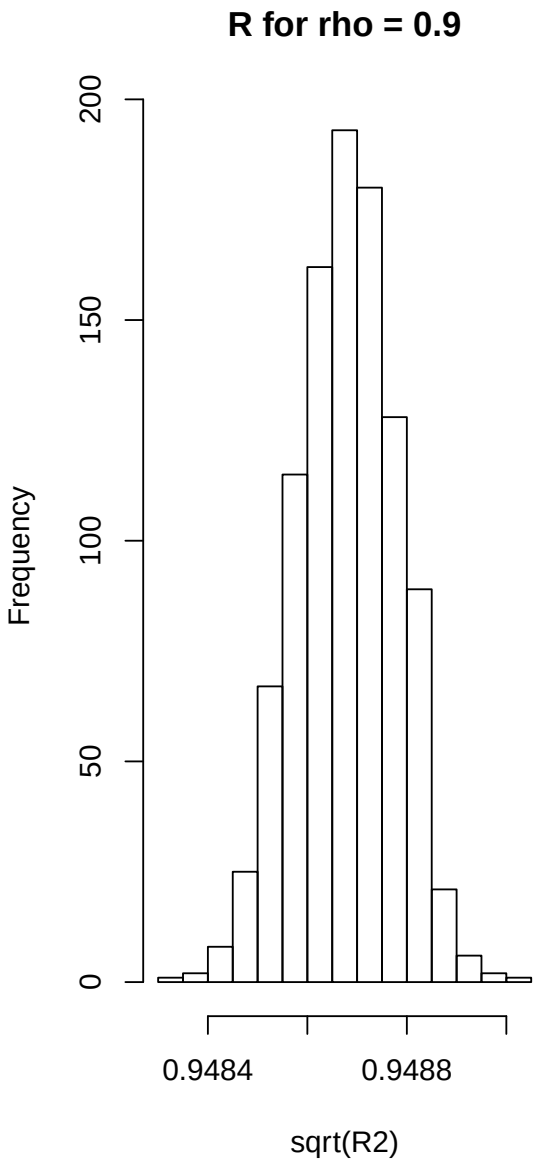

R2 for rho = 0.7
Frequency
0
50
100
150
200
250
300
350
0.698
0.700
0.702
R2

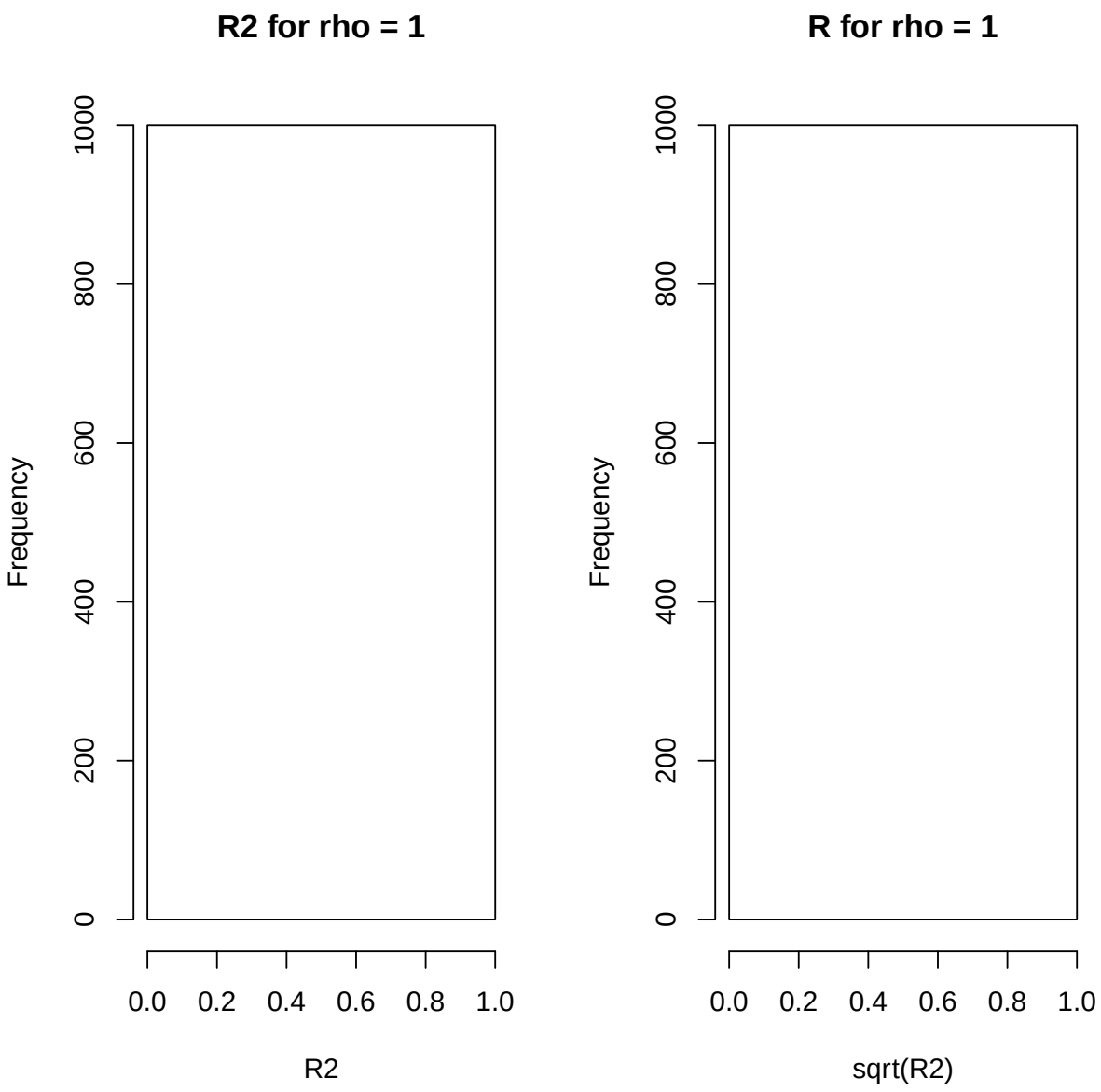

R for rho = 0.7
Frequency
0
50
100
150
200
250
0.8360
0.8370
sqrt(R2)


R2 for rho = 0.8
Frequency
0
50
100
150
200
0.7990
0.8000
0.8010
R2


R for rho = 0.8
Frequency
0
50
100
150
0.8940
0.8945
0.8950
sqrt(R2)


R2 for rho = 0.9
Frequency
0
50
100
150
200
0.8994
0.9000
0.9006
R2

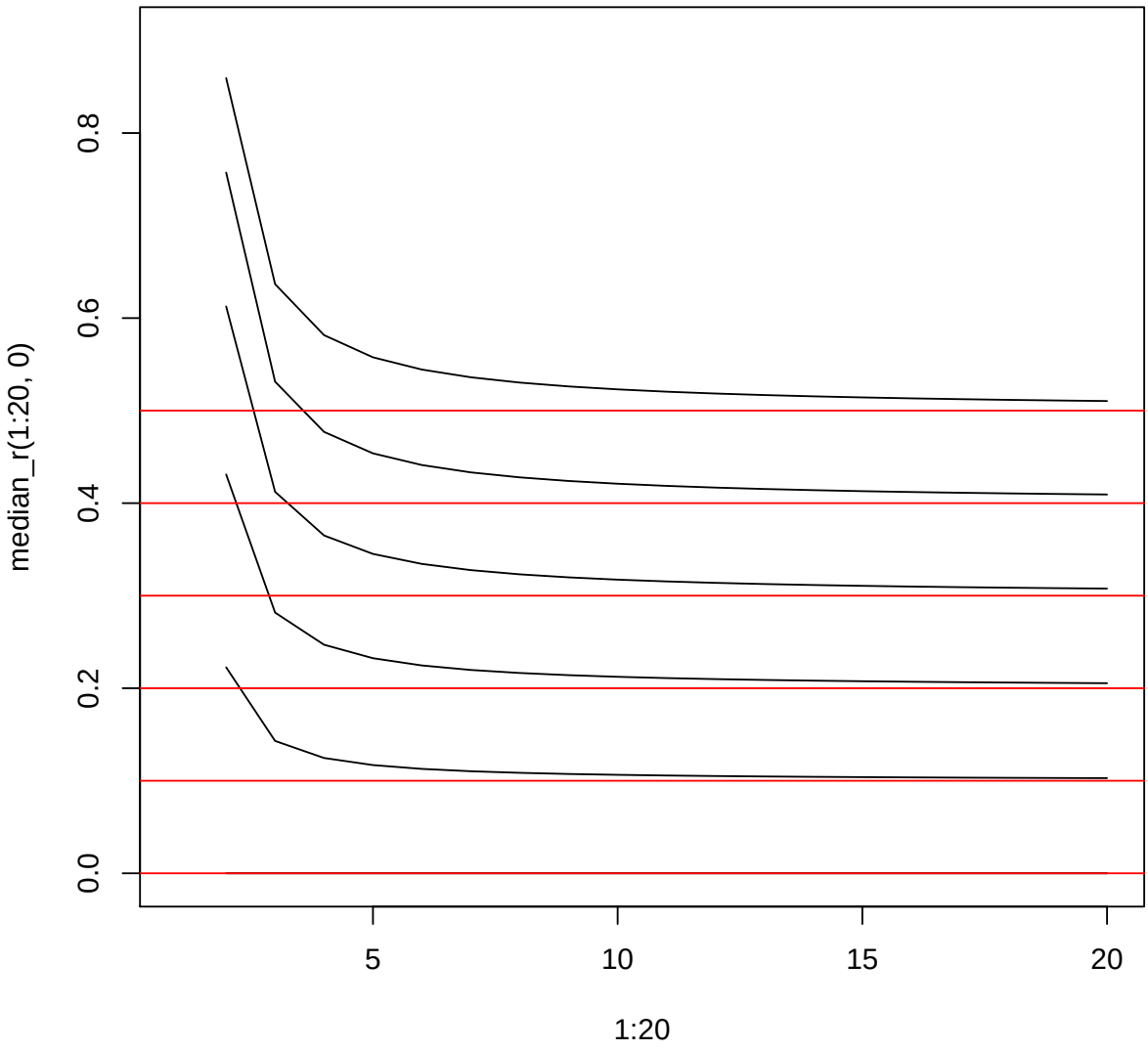

R for rho = 0.9
Frequency
0
50
100
150
200
0.9484
0.9488
sqrt(R2)



Refereras på s. 216 till median av r (obs! Ej R2, vilket skiljer vid jmför mot E[r2] ovan).

```
median_r <- function(n, rho) {
  rho +
    (rho * (1 - rho ^ 2)) / (2 * (n - 1)) *
    (1 + (9 - 14 * rho ^ 2) / (6 * (n - 1)))
}

plot(1:20, median_r(1:20, 0), type = "l", ylim = c(0, .9))
lines(1:20, median_r(1:20, .1), type = "l")
lines(1:20, median_r(1:20, .2), type = "l")
lines(1:20, median_r(1:20, .3), type = "l")
lines(1:20, median_r(1:20, .4), type = "l")
lines(1:20, median_r(1:20, .5), type = "l")
```

```r
abline(h = seq(0, .5, .1), col = "red")
```



Vi ser här att vi har väldigt stor bias för väldigt små n men att vi sedan har en väldigt snabb konvergens, som dock blir ngt sämre med större ρ. OBS! Fisher använder horisontella streck i sina formler som jag inte riktigt vet vad de betyder. Jag gissar dock här att de är istf parenteser men känner mig inte alls säker då de även används på ställen där det känns överflödigt.

Kan kanske vara värt att notera att Fisher skriver $r = tanh(z)$ när han egentligen tycks mena $arctanh$.

Det fanns en kontrovers mellan Fishers artikel 1915 och cooperative study 1916 som missförstod den föregående och vars kritik nu bemöts 1921. Kan kanske nämnas nu i samband med 100-årsjubileumet. 1916 antog att Fisher 1915 använt Byes sats på ett sätt han inte gjort. F.ö. noteras det lite roliga citatet (s 220):

> ... its value depends almost wholly upon the preconcieved opinions of thye computer and scarely at all upon the data supplied to him.

I det stora hela handlar artikeln mer om z och dess egenskaper än om r.

2.2 Läsning av (Student 1908)

Delvis empirisk. Behandlar fallet för bivariat fördelning med två variabler. Utgår fram tabell och drar slumpmässigt 750 stickprov av storlek 4 därifrån. Räknar även ut moment etc och redovisar fördelningarna i tabellform. Även försök med 100 stickprov á 30. Av tabellerna att döma tycks det som att även Student överskattar korrelationen. Dock skriver han om underskattning i löptexten och ger även sådan indikation i annan mindre tabell (s. 306). Det avser dock värden för den teoretiska fördelning han föreslår och inte för de impiriska observationer han gör (såvitt jag kan förstå). Han utvecklar alltså här teori för en fördelning men jag tror vi kan bortse från den då det kommit bättre förslag senare (inkl (Hotelling 1953)).

Gör en allmän reflektion att stickprovsstorlek om minst 30 tycks räcka för att undvika problematik med små stickprov.

2.3 Läsning av (Soper 1913)

Denna artikel svar på föregående (som uppmanade till ytterligare forskning). Konstaterar att då n stor och ρ ej nära +-1 har man approximativ normalfördelning men i övrigt får man observerade värden som motsvarar fördelningens mode, ej mean, vilket skiljer sig kraftigt ty fördelningen väldigt skev.

Skriver att det fnins ett arbete med beskrivning av ett par skeva fördelningar som brukar passa väldigt bra i väldigt många sammanhang. Skriver också att det i detta fall krävs en empirisk validering för att finna en relevant fördelning. Sådan föreslås och beräkningar sker.

Inledningsvis är artikeln väldigt teoretisk (massiva härledningar) men fr s. 107 lite mer deskriptiv i ord.

3 2016-02-23

3.1 Fortsatt läsning av (Soper 1913)

Fördelningen han föreslår (s. 107) har (för små n) en approximativ fördelnig för medelvärdet som:

$$\rho\big(1 - \frac{1-\rho^2}{2n}\big)$$

och standardavvikelse:

$$\frac{1-\rho^2}{\sqrt{n-1}}\big(1 + \frac{11\rho^2}{4n}$$

.

Notera att medelvärdet här blir mindre än ρ men detta rör förstås medelvärdet och inte typvärdet vilket är det vi observerar vid simulering, dvs vi beräknar medelvärdet av typvärdet för den sanna fördelningen. Även i denna artikel framkommer att samplat medelvärde kommer ge en överskattning av det sanna värdet. Typvärdet ges å andra sidan approximativt av (värt att jämföra med värde som ges ovan enl betafördelning etc):

$$\rho\left\{1 + \frac{3(1-\rho^2)}{2(n-1)} + \frac{(41+23\rho^2)(1-\rho^2)}{8(n-1)^2}\right\}$$

Konstaterar att dessa formler leder till totslt haveri för stickprovsstorlekar så små som 3, 4 el 5, dvs att product moment approximation inte funkar i dessa fall. Detta eftersom som vi då får typvärde = 0 etc. För

så små stickprovsstorlekar i kombination med små ρ får r en uniform fördelning $U(0,1)$. Min slutsats blir väl då att aktuellt fördelningsantagande bör ersättas med fördelning från (Hotelling 1953) etc. Jämförelse med dessa teoretiska modeller mot experiment i (Student 1908) visar inte helt samstämighet och skillnaderna är delvis större än vad slumpen medger. Dock tycker jag ändå fördelningarna i kurvorna ser ganska övertygande ut (så dåligt är det alltså inte).

F.ö tycks K Pearsson ha agerat handledare i detta arbete (nämns väldigt kort precis i slutet).

Tycker på det hela taget detta var en bra artikel om än väldigt teoretisk inledningsvis.

3.2 Läsning av (Soper, HE and Young, AW and Cave, BM and Lee, Alice and Pearson 1916)

Inleder med en väldigt bra sammanfattning av läget på den tiden. Här finns också formler för $\bar{r}, \sigma_r^2$ och typvärde skrivna i en lite annan form. Erkänner att beräkningarna varit väldigt ansträngande och tagit flera månader med hjälp av flera namnvigna computers.

Här införs bruk av hypergeometrisk fördelning (s. 333). Moment härleds olika för udda och jämna n.Här finner man snabb konvergens för formlerna från $n > 25$. Artikeln innehåller väldigt många tabeller med uppgifter om moment etc för olika n, ρ etc. Ett gediget arbete. På s 25 fnins också graf som illustrerar fall för $n = 25$ och där det framgår att fördelningen blivit tämligen symmetrisk men med mean ngt lägre än mode (enl texten är den långt från normal men följer bra en Pearsson-kurva). Här framkommer också att skattad mean ligger lite över teoretisk mean men under mode (dvs utgår från föreslagen teoretisk fördelning).

Artikeln (även fler) refererar till fördelning som "Pearson curve typ I och II" etc. Detta var tidigare benämningar på betafördelning samt ett specialfall av den föregående som inte längre brukar namnges explicit.

Artikeln ger f.ö. ett teoretiskt mått på diff mellan mean och mpde (xxvii).

De formler som utarbetats fram till s 337 sägs inte vara till nytta för färre fall än 25. Man gör därmed separata härledningar för $n = 5, ..., 24$ ($n < 5$ visas inget intresse).

Framkommer på s 350 att tidigare formel för typvärde enl (Soper 1913) tycks felaktig men resonerar och kommer fram til att det finns en naturlig förklaring. I vilket fall som helst presenteras här en alternativ formell (bet mer komplex). Tror dock vi kan bortse från detta och förkasta till förmån för senare forskning.

Man använder f.ö. flera olika metodwer för att beräkna liknande eller samma resultat men fördjupar mig inte i detta.

Poängterar s 351 att konvergens mpt normalfördelning sker väldigt långsamt och att fördelningen fortsatt är skev.

Artkeln inkluderar bilder på 6 st fysiska modeller som visualiserar den skeva fördelningen för olika val av n och ρ.

Som lite rolig kommentar noteras referens till att man undersökt just egenskaper för "human femur" och därmed noterar 400 korrelationer i anslutning till detta.

På s. 356 ges formel för $\hat{\rho}$, vilket ska vara "th emost likely value of the correlation in the sampled population" (s. 352). Formuleras också som "Suppose we have found the value of the correlation in a small sample to be r, what is the most reasonable value $\hat{\rho}$ to give to the correlation ρ of the sampled population?" själva uttrycket återges ej här men man skriver också själva att approximation tidigare given i (R. Fisher and Fisher 1915) enl

$$\rho = r\big(1 - \frac{1 - r^2}{2n}$$

är tillräckligt bra och användning av den formeln rekommenderas i praktiken.

Fig 2 (s. 368) illustrerar sambandet mellan mean och mode/antimode för olika ρ för n = 3.

Normalfördelningsapproximation uppnås inte för ens $n = 25$ eller 50 (s. 371) men för låga värden på ρ räcker $n = 100$ ganska bra men för $\rho > .5$ räcker inte ens detta.

Artikeln avslutas med att den blivit mycket längre än förväntat och att vissa delar därför utelämnats til förmån för en uppföljande del (vet in teom det kom ngn sådan? Hittar ingen med liknande namn).

Det läggs väldigt mkt möda på att skapa tabelllverk med väldigt mpnga decimaler.

3.3 Läsning av (Nair 1941)

Behandlar icke-normalfördelad data! Han utgår från tabellverk med 10400 slumptal och adderar dem alla parvis!!!

Undersöker samples med korrelation mellan mean och sd.

Konstaterar på s 391 att population har positiv skevhet men sample ger neg skevhet, vilket också kopnstateras av andra.

Gör empirisk undersökning av exponentialfördelad och pearson curve III-fördelad (gammafördelad) data, dels mot antagande om att data ändå skulle vara normalfördelad, dels genom Fishers två approximationer för z transformationen. Hans slutsats är att normalapproximationen ändå funkar förvånandsvärt bra och att denna t o m är bättre än att gå via z transformationen. Bygger på ett par hundra stickprov av storlek 6. Medger dock att det skulle behövas mer data för korrekt skattning av extremvärden etc.

På det hela taget en hyfsat intressant case study men inte med så mkt ny teori. Kanske av ungefär den kaliber som vi själva kan åstadkomma?

3.4 Läsning av (Gayen 1951)

Behandlar också icke normalfördelad data (för alla stickprovsstorlekar). Teoretisk. Många formler men också en del grafer.

Refererar till att teorin för bivariat normalfördelad data (R. Fisher and Fisher 1915) funkar även om fördelningsantagandet kan ifrågasättas ganska mkt. Kan dock bli problem med väldigt små stickprovsstorlekar.

Bygger på teorier som bör göra resultaten applicerbara nästan oberoende av fördelning etc och behandlar hela skalan av möjliga (men kända) ρ. Utgår från bivariat Edgeworth-"fördelning".

Tar fram en extremt krånglig fördelning (32) s. 224 samt mean och var för r enl (38) och (39) men även dessa väldigt långa. (32) kan delas upp i tre delar: (i) ordinär normalfördelning, (ii) + (iii) corrective functions due to poulation excess and skewness. Menar att normalapproximationen funkar bra för $\rho = 0, N \geq 11$, däremot behövs korrigeringen för t ex $\rho = 0.8$.

Handlar också en del om Fishers z transformation.

Denna artikel känns lite överteoretisk. Undrar om den fått så stort genomslag i praktiken? Ser att den citerats 177 ggr men även de som citerar tycks vara teoretiska.

4 2016-02-24

4.1 Läsning av (Ruben 1966)

Utgår från normalfördelad data. Utvecklar alternativ normalapproximation som funkar ngt bättre än tidigare förslag inkl (Hotelling 1953) samt ytterligare metod som funkar för stora n, ρ.

Bygger på transformation av r till kvot mellan två χ^2-variables. Visar att ytterligare transformation har standardiserad normalfördelning. Denna approx funkar lite bättre än Fishers z och Hotellings formler samt mkt bättre då n stort.

Intressant stycke s. 518, historisk reflektion att då Fisher presenterade sin artikel 1915 fokuserade man på korrelation etc då regressionsanalys och ANOVA etc ännu inte introducerats. I alla fall diskuteras att en härledning liknar situationen för multilpe correlation för fler oberoende variabler etc. Denna var dock bara aktuell (åtm vid tidpunkten) för $\rho = 0$ då $p = 2$ (men funkade för $p > 2$).

Inser nu f.ö. frårn Wikipedia att det vi hittills kollat på ju är just "the coefficient of multiple correlation". Wikipedia (https://en.wikipedia.org/wiki/Multiple_correlation):

> The coefficient of multiple correlation, denoted R, is a scalar that is defined as the Pearson correlation coefficient between the predicted and the actual values of the dependent variable in a linear regression model that includes an intercept.

Assymptotiskt då $\lim_{n\to\infty} r \sim N(\rho, (1-\rho^2)^2/n)$ (s. 519)

Jag får känslan att man kanske vill ha en approximativ normalfördelning för att det är lättare rent beräkningsmässigt än att utgå från betafördelning etc.

4.2 Läsning av (Mukaka 2012)

Känner ett lpötsligt behov av ngt mer samtida som moväxling.

Kort artikel med få referenser. Grafer från Stata. Lättläst. För med publik. Bra ingång till ämnet. Förklarar vad korrelation är och hur det tolkas (t ex att det bara är linjärt och inte mäter andra typer av samband).

Står att pearsson korrelation bara används då båda variablerna normalfördelade. Spearmans rank correlation rekommenderas för skeva fördelningar eller ordinal data och är mer stabil för extremvärden. Tydliga exempel på när respektive mått är lämpat.

Inkluderar uppställning med olika ρ och hur de ska tolkas, vad som är lågt och högt etc.

4.3 Läsning av (Gorsuch and Lehmann 2010)

Artikeln från teologiskt seminarium. Påpekar problem då $\rho \neq 0$. Nyttjar Fishers z. Skattar för små stickprovsstorlekar 20 och 30 samt stora 50 och 100. Finner att Fisher ej behövs för stora n men att konfidensintervall rtots allt blir bättre.

Menar att det är allmänt känt att r underskattar ρ och hänvisar detta till (R. A. Fisher 1921).

Faktum är också att författarna här själva noterar överskattning men skriver:

> Note that the modal and median correlation encountered by investigators is considerably above the population ρ for the N = 20. So while the average correlation will be underestimated when averaging across studies, the average investigator will observe a correlation which overestimates the population value, a point that has not been widely discussed.

Beskriver pedagogiskt hur z används i praktiken. Menar att z (i motsats till r) överskattar ρ. Refererar också till (Olkin and Pratt 1958) men konstaterar att den formeln bara gör nytta för $n < 20$.

Resonerar kring CI antingen baserat på normalfördelning (och konstaterar problemen med assymetyrisk fördelning) eller via Fishers z. Lägger också ut ordne ifall man kan använda t-fördelning hellre än normalapproximation eller hur man kan approximera detta. Känns ganska långt från de tidigare teoretiska förslag som framlagts.

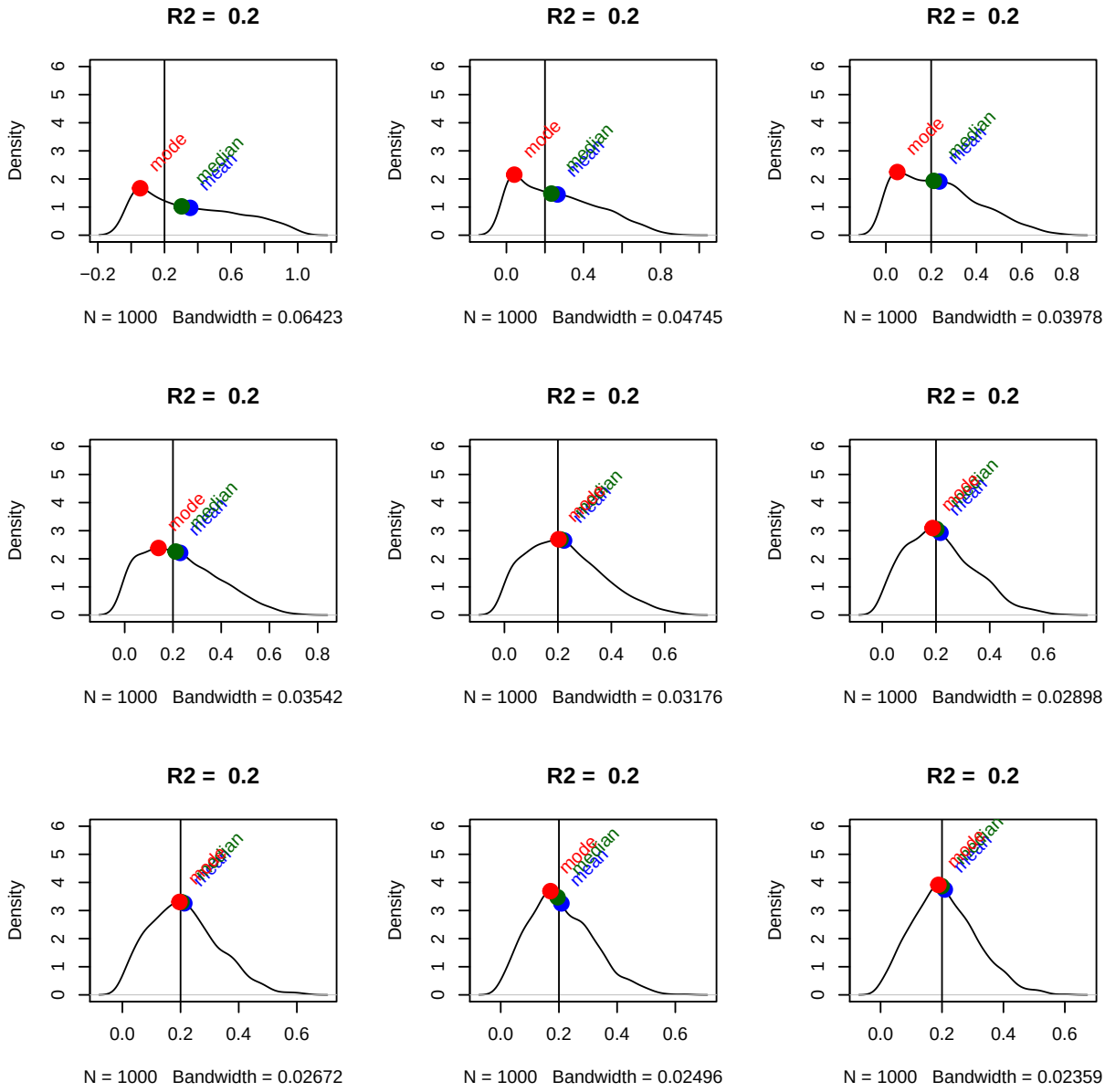
Utgår från simulering med 5000 samples i SPSS och korrelerar två variabler X och Y. Valde stickprovsstorlekar 20, 30, 50 och 100. Utgick bara från $\rho = 0.2, 0.5, 0.8$ (förutsåg mest bias för $\rho = 0.5$). För varje sample beräknades mean, median, sd, skew och kurtosis. Noterar också att skevheten i simulerade fördelningar ökar med ρ men minskar med n. Vi kan ju här f.ö. konstatera att våra simulöeringar ovan (avs 2.1) baseras på $n = 1000$ och att de därmed bör vara approximativt normalfördelade även om så inte är fallet för R^2 då $\rho = 0$.

Jag vill jämföra lite egna simuleringhar med R^2 istf r (vilket alltså innebär att mina uppgifter nite alls är jämförbara mot artikelns ... men ändå ...).

```
teologiska <- function(r2) {
  sim_data(r2, p = 1) %>%
  subsamples(n.max = 100, N = 1000) %>%
  metrics(n.sample = seq(5, 45, 5)) %>%
  .$Rsquared
}

# Function to add opint to density ponit
add_point <- function(d, x, text, col) {
  xs <- density(d)$x
  ys <- density(d)$y
  x2y <- function(x)  ys[which.min(abs(xs - x))]
  text(x, x2y(x), label = text, col = col, srt = 45, adj = c(-.5, 0))
  points(x, x2y(x), col = col, lwd = 5)
}

# Calculate mode frfom opssibly contnous distribution
mode <- function(d) {
  z <- density(d)
  z$x[z$y == max(z$y)]
}

# Illustrate the skewness dependent on rho and n
plot_skewness <- function(d, R2, ...) {
  plot(density(d), main = paste("R2 = ", R2), ...)
  abline(v = R2)
  add_point(d, mean(d),   text = "mean",   col = "blue")
  add_point(d, median(d), text = "median", col = "darkgreen")
  add_point(d, mode(d),   text = "mode",   col = "red")
}
```

För $\rho = .2$

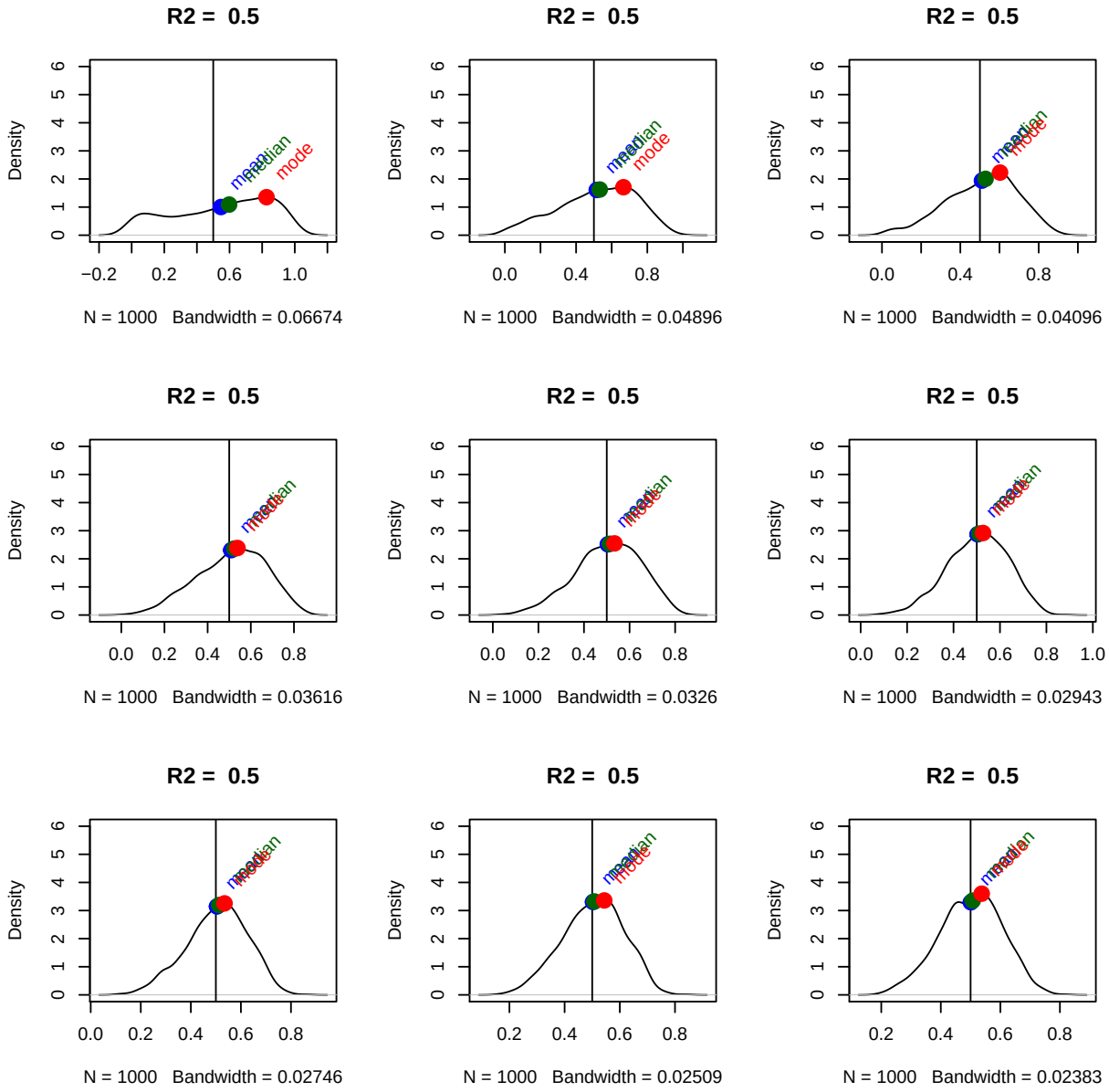
```
par(mfrow = c(3, 3))
apply(teologiska(.2), 2, plot_skewness, R2 = .2, ylim = c(0, 6))
```




```
## NULL
```

För $\rho = .5$

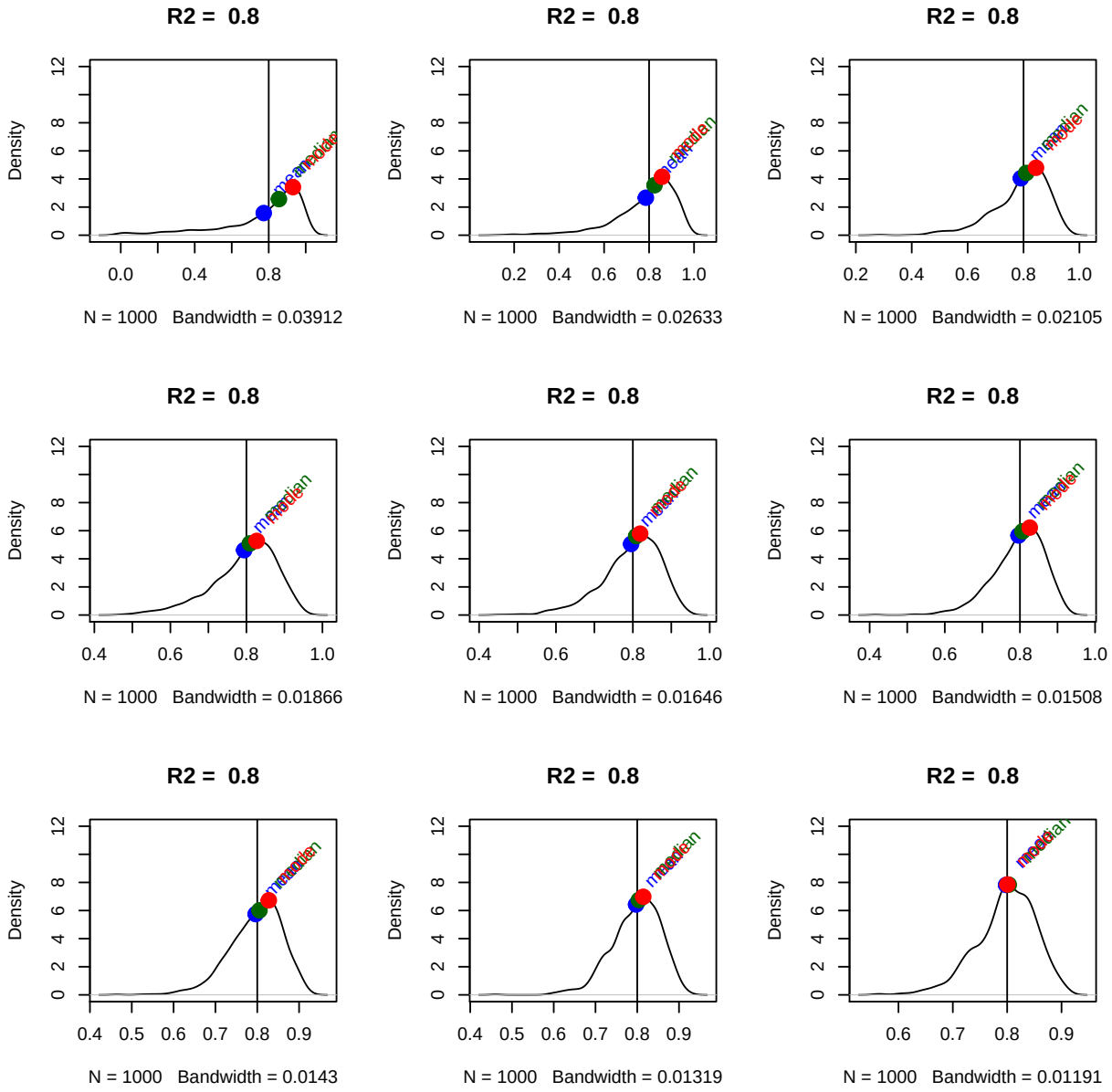
```
par(mfrow = c(3, 3))
apply(teologiska(.5), 2, plot_skewness, R2 = .5, ylim = c(0, 6))
```




```
## NULL
```

För $\rho = .8$

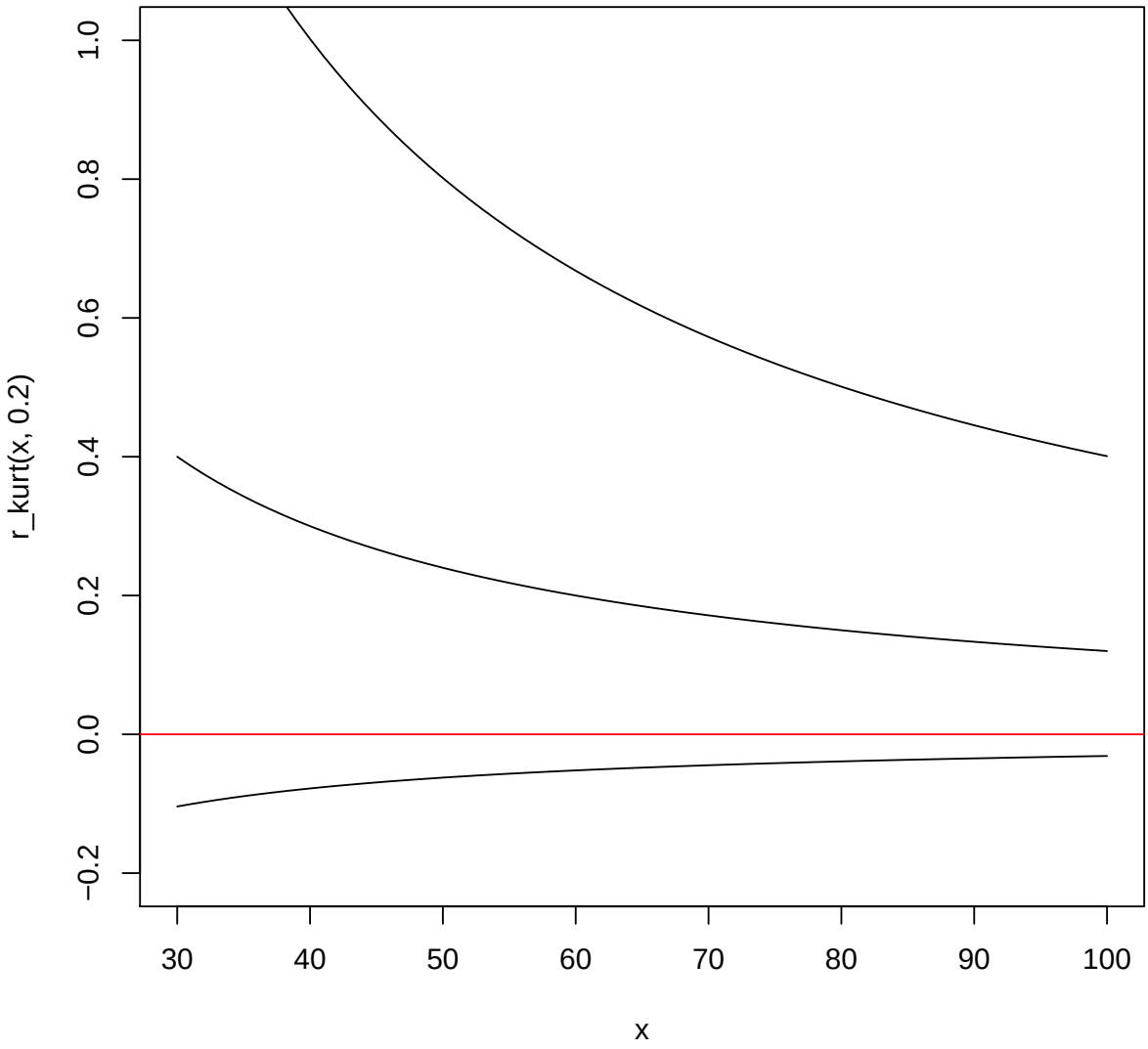
```
par(mfrow = c(3, 3))
apply(teologiska(.8), 2, plot_skewness, R2 = .8, ylim = c(0, 12))
```




```
## NULL
```

Står f.ö. att kurtosis skiftar från negativt till pos då ρ ökar men att detta tidigare inte nämnts i litteraturen. Stämmer inte! Detta överrensstämmer väl med (Hotelling 1953)? Enligt s 212 ges kurtosis approximativt av:

```
r_kurt <- function(n, rho) (6 / n) * (12 * rho ^ 2 - 1)

par(mfrow = c(1, 1))
x <- 30:100
plot(x, r_kurt(x, .2), type = "l", ylim = c(-.2, 1))
lines(x, r_kurt(x, .5), type = "l")
lines(x, r_kurt(x, .8), type = "l")
abline(h = 0, col = "red")
```




```
# beräknas teoretiska kurtosis för motsvarande n och rho som i artikeln
ns <-  c(20, 30, 50, 100)
rhos <- c(.2, .5, .8)
k <- outer(ns, rhos, r_kurt)
k <- t(k)
rownames(k) <- paste("rho = ", rhos)
knitr::kable(k, row.names = TRUE, col.names = paste("n = ", ns))
```

	n = 20	n = 30	n = 50	n = 100
rho = 0.2	-0.156	-0.104	-0.0624	-0.0312
rho = 0.5	0.600	0.400	0.2400	0.1200
rho = 0.8	2.004	1.336	0.8016	0.4008

Vi kan alltså konstatera att teorin visst medger att tecknet skiftar (dock får vi inte ngn exakt överrensstämmelse). Kanske kan vara intressant att göra egna simuleringar av motsvarande uppgifter för R2?

```
sim_data(.2, p = 1) %>%
  subsamples(n.max = 100, N = 5000) %>%
  metrics(n.sample = c(20, 30, 50, 100)) %>%
  .$Rsquared %>%
  apply(2, e1071::kurtosis, type = 2)
```

```
##           20           30           50          100
## -0.281068100 -0.228779116 -0.215273893  0.008309957
```

5 2016-02-25

Fortsätter läsning enl ovan.

De presenterar bias ngt deskreptivt för olika nivår men utan att jämföra med den teoretiska formeln (so minte alls nämns). De finner stört bias för större ρ (vilket väl också är vad teorin visar).

De jämför även bias för bias-korrigerad $G(r)$ enl (Olkin and Pratt 1958). CI beräknas i tab 4 men enl normalapproximation även för små N (konstateeras att detta inte blir bra). CI för z beräknas via transform och back-transform till N.

Slutsatser att bias negligerbar för $n > 30$ och att CI bör baseras på Fishers z. Den bia som finns är i huvud taget så pass liten att detta problem ofta är det minsta relaterat till alla möjliga påverkansfaktorer vid beräkning.

Hade vart intressant att undersöka en liknande metodik för R^2 men svårare då ju detta värde inte låter sig transformeras fram och tillbaka utan blir då absolutbelopp. Konstaterar f.ö. att standardavvikelse (och därmed varians) för r bara utgår från r^2 (och inte r). Detsamma gäller tyvärr inte för bias. Vi kan ju också påminna oss själva om att vi har $E[r^2]$ sedan tidigare.

5.1 Allmän reflektion

Jag går tillbaka till (Hogben 1968) och noterar referensen till att r^2 har en icke central betafördelning (Seber 1963). Hogben transformerar r till $r = W/\sqrt{(w^2 + X^2)}$ där $W \sim N(..., 1), X \sim \chi^2_{n-1}$vilket leder till $r^2 = W^2/(W^2 + X^2)$ där $W^2 \sim$ ickecentral χ^2_1 (ickecentral ty mean $\neq 0$). Enligt (Seber 1963) gäller att sådan statistika har ickecentral betafördelning typ I med shape $1 = 1/2$, shape $2 = (n-2)/2$ och icke centralitetsparameter

$$\lambda = \frac{\theta^2}{2} = \frac{\beta^2}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

där $Y_i = \alpha + \beta x_i + \epsilon_i, i = 1, 2, ..., n, \epsilon \sim N(0, \sigma^2)$. Observera dock att detta endast gäller då x är fix! Dvs Y är stokastisk men inte x. Är lite osäker på vad detta egentligen betyder i praktiken!? Kan man undersöka empiriskt hur väl detta stämmer då även x är stokastisk!? Vilka artiklar finns som refererar till denna? Kanske finns där ngn som utvecklar det hela? R-funktionsfamiljen `Beta` kan därmed anropas med argumenten `shape1`, `shape2` och `ncp` för dessa parametrar.

Antagandet om x_i fixt är enl SN för att man då inte underskattar β. I praktiken bygger x_i ändå på realisernig av en underliggande stokastisk variabel. Genom att ignorera detta går man misste om möjligheten att skatta x från Y (man kan bara gå åt andra hållet). SN hänvisar också till (Faraway 2005, s 77) som beskriver precis samma sak men förklarar det i formler. När vi tillämpar detta på riktiga data kan vi då kanske anta att vi de facto har en realisation via observationerna. Kanske dock ngt svårare motivera vid simulerad data.

5.2 Läsning av (Skidmore and Thompson 2011)

Använder monte carlo-simuleringar med $6 \beta 6 = 108$ val av ρ, fördelningar med olika skevhet samt stickprovsstorlekar. Finner att förslag från (Olkin and Pratt 1958) funkar i fler situationer än förutspått.

Betonar att studier av effect size ofta lyfts fram som viktigt och att användningen är stor exempelvis inom metaanalys.

Nämns att även om det föreslagits korrigering typ Olkin, så sker det sällan.

Ger referenser för :

> Previous researchers have studied the efficacy of (a) multiple R2 correction formulas with multiple predictors (Cattin, 1980; Claudy, 1978; Huberty&Mourad, 1980; Raju, Bilgic, Edwards, & Fleer, 1999; Shieh, 2008; Yin & Fan, 2001), (b) multiple R2 correction formulas with one predictor (i.e., the Pearson r2) (Cattin, 1980; Wang & Thompson, 2007),

Känns som att dessa kan vara mkt viktiga att kolla upp!

Förklarar tre källor till bias:

> Three study design features affect bias (Thompson, 2006a, 2006b). First, greater sampling error, and thus greater tendency on the average to overestimate true population effect size, occurs in studies with smaller sample sizes, for obvious reasons. Second, greater sampling error tends to occur in studies involving more mea- sured variables, for less obvious reasons. [Förklaras mkt pedagogiskt men klipper ej in här]. Third, least obvious is the fact that greater sampling error (and thus more inflated effect estimates) tends to occur in samples drawn from populations with smaller population parameter effect sizes.

Observera här att vi alltså har det omvända förhållandet från texter som handlar om ρ, ty i och med att vi kvadrerar tal i intervallet [0,1] så blir ju de störe talen mindre och vice versa.

Artikeln är i det stora hela pedagogisk och bra samt stödjer sig på mkt teori etc. Använder själva SAS men erbjuder också ett Excel-blad för den som vill göra egna beräkningar.

Simulerar 5000 ggr för var och en av de 108 kombinationerna.

För låga ρ och n visade sig även de korrigerade värdena ha pos bias. Dock bättre för större ρ, n.

Konstaterar att fördelningsskillnader inte påverkade resultatet nämnvärt, dvs att data inte behöver vara bivariat normalfördelat (styrker intresset för (Hogben 1968)).

Rekommendationer har gjort att justera R^2-skattning för $N < 50$.

Nämner inget om den ickecentrala betafördelningen.

F.ö. noterar jag från http://stats.stackexchange.com/questions/48703/what-is-the-adjusted-r-squared-formula-in-lm-in-r-and-l att den version som används i `summary.lm` är den som går under olika namn: “Wherry formula”, “Ezekiel formlua”, “Wherry/McNemar formula”, and “Cohen/Cohen formula”. Ges av: Ezekiel (1929; 1930, p. 121)

Vi kan alltså kopnstatera att den version som R använder egentligen inte är så bra. Varför har man valt just den?

5.3 Läsning av (Ezekei 1929)

Blir lite nyfiken på just denna version av adjusted R2 så läser dfen artikeln. Vi denna tidpunkt uttrycktes att man fortfarande inte sett ngn praktisk användning av varken (Student 1908) eller (R. Fisher and Fisher 1915).

Utgår från multiple correlation.

OBS! formenl utrycks lite annorlunda än hur den sedan implementerats i `lm.summary` I `lm.summary`:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

men i artikeln:

$$\bar{R}^2 = 1 - \frac{1-R^2}{1-\frac{m}{n}} = [m = p+1] = 1 - \frac{1-R^2}{1-\frac{p+1}{n}} = 1 - \frac{1-R^2}{\frac{n}{n} - \frac{p+1}{n}} = 1 - \frac{1-R^2}{\frac{n-(p+1)}{n}} = 1 - n\frac{1-R^2}{n-(p+1)} = 1 - (1-R^2)\frac{n}{n-(p+1)}$$

där vi alltså i täljaren har n istf $n-1$. Numera används n - 1 då detta är frihetsgraden. Frihetsgrad infördes 1908 och populariserades 1922. Var det ännu inte helt main stream 1929?

5.4 Läsning av (R. a Fisher 1924)

Beskriver att r är invariant till ortogonala transformationer av x och y.

Utökar till "partial correlation coefficient", dvs hur korrelationen påverkas av ytterligare variabler.

Partiella correlationer kan beräknas mha paketet "ppcor":

```
d <- sim_data()
cor(d)
```

```
##            Y            X1            X2            X3            X4
## Y  1.0000000  0.2010178998  0.1988989637  0.1982098456  0.2007194135
## X1 0.2010179  1.0000000000  0.0003330857  0.0007151523 -0.0011523315
## X2 0.1988990  0.0003330857  1.0000000000  0.0002002419 -0.0018095128
## X3 0.1982098  0.0007151523  0.0002002419  1.0000000000 -0.0005790543
## X4 0.2007194 -0.0011523315 -0.0018095128 -0.0005790543  1.0000000000
## X5 0.2020532  0.0008563422 -0.0005585796 -0.0002395177  0.0016090724
##               X5
## Y   0.2020532257
## X1  0.0008563422
## X2 -0.0005585796
## X3 -0.0002395177
## X4  0.0016090724
## X5  1.0000000000
```

```
ppcor::pcor(d)$estimate
```

```
##            Y          X1          X2          X3          X4          X5
## Y  1.0000000  0.21917469  0.21751185  0.21639080  0.21941605  0.22005476
## X1 0.2191747  1.00000000 -0.04735755 -0.04674666 -0.04918771 -0.04741329
## X2 0.2175118 -0.04735755  1.00000000 -0.04687810 -0.04944741 -0.04839381
## X3 0.2163908 -0.04674666 -0.04687810  1.00000000 -0.04802968 -0.04784553
## X4 0.2194160 -0.04918771 -0.04944741 -0.04802968  1.00000000 -0.04675231
## X5 0.2200548 -0.04741329 -0.04839381 -0.04784553 -0.04675231  1.00000000
```

Vi ser här att våra simulerade data är partiellt korrelerade! Hur detta tolkas förklaras pedagogiskt här: http://www.psychwiki.com/wiki/What_is_a_partial_correlation%3F

Det innebär kanske att våra data trots allt nite är helt oberoende! Kan detta förklara varför vår bias inte är helt additiv då vi har fler oberoende variabler?

5.5 Läsning av (Wherry 1931)

Formlerna i denna artikel är handskrivna :-) Refererar till samma formel som ovan. Här nämns dock att denna formel visserligen presenterats av Ezekei men att den ursprungligen utvecklats av en B B Smith.

Beskrivs att metoden använts ganska mkt redan men att (Larson 1931) i tidigare artikel uppmärksammat via empiri att bias återstår. Detta är då ett alternativ som ska fungera bättre, vilket motiveras både via teori och empiri.

Observera att denna metod ej inkluderas vid jämförelsen i (Skidmore and Thompson 2011) vilket ju är lite synd kanske.

5.6 Läsning av (Larson 1931)

Denna artikel publicerades alltså tidigare än föregående. Konstaterar att redan sedan flera år konstaterats från flera håll att R2 växer vid införande av fler oberoende variabler. Situationen anses allvarlig då det kan leda till att folk helt överger användning av regression etc. Han jämför alltså också med (Ezekei 1929) men får då en felskattning åt andra hållet.

5.7 Notering

Noterar f.ö. i helt annat sammanhang att Karl Pearson introducerade korrelationskoefficienten 1896. Lyckades nu också hitta den och har lagt in den i Mendeley! Tjoho!

Referenser

Ezekei, Mordecai. 1929. "The Application of the Theory of Error to Multiple and Curvilinear Correlation." *Journal of the American Statistical Association* 24 (165): 99–104.

Faraway, Julian J. 2005. *Linear Models with R.* http://www.stat.lsa.umich.edu/{~}faraway/LMR/.

Fisher, R A. 1921. "On the probable error of a coefficient of correlation deduced from a small samlpe."

Fisher, R a. 1924. "The distribution of the partial correlation coefficient."

Fisher, R.a., and R.a. Fisher. 1915. "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population." *Biometrika* 10 (4): 507–21. doi:10.2307/2331838.

Gayen, A. K. 1951. "The Frequency Distribution of the Product-Moment Correlation Coefficient in Random Samples of Any Size Drawn from Non-Normal Universes." *Biometrika* 38 (1/2): 219–47.

Gorsuch, Rl, and Cs Lehmann. 2010. "Correlation Coefficients: Mean Bias and Confidence Interval Distortions." *Journal of Methods and Measurement in the Social Sciences* 1 (2): 52–65. https://journals.uair.arizona.edu/index.php/jmmss/article/download/114/118.

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.

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.

Larson, S C. 1931. "The shrinkage of the coefficient of multiple correlation." *Journal of Educational Psychology* 22 (1): 45–55. doi:10.1037/h0072400.

Mukaka, M. M. 2012. "Statistics corner: A guide to appropriate use of correlation coefficient in medical

research." *Malawi Medical Journal* 24 (3): 69–71.

Nair, A N Krishnan. 1941. "Distribution of Students 't' and the Correlation Coefficient in Samples from Non-Normal Populations." *The Indian Journal of Statistics* 5 (4): 383–400.

Olkin, Ingram, and J.W. Pratt. 1958. "Unbiased estimation of certain correlation coefficients." *The Annals of Mathematical Statistics* 29 (1): 201–11. doi:10.2307/2237306.

Ruben, Harold. 1966. "Some New Results on the Coefficient of the Sample Correlation Distribution." *Journal of the Royal Statistical Society. Series B (Methodological),* 28 (3): 513–25.

Seber, G A F. 1963. "The Non-Central Chi-Squared and Beta Distributions." *Biometrika* 50 (3): 542–44.

Skidmore, Susan Troncoso, and Bruce Thompson. 2011. "Choosing the Best Correction Formula for the Pearson r 2 Effect Size." *The Journal of Experimental Education* 79 (3): 257–78. doi:10.1080/00220973.2010.484437.

Soper, H. E. 1913. "On the peobable error of the correlation coefficient to a second approximation." *Biometrika* 9 (1-2): 91–115. doi:10.1093/biomet/9.1-2.91.

Soper, HE and Young, AW and Cave, BM and Lee, Alice and Pearson, Karl. 1916. "On the Distribution of the Correlation Coefficient in Small Samples. Appendix II to the Papers of 'Student' and R. A. Fisher." *Biometrika* 11 (4): 328–413.

Student. 1908. "Probable Error of a Correlation Coefficient." *Biometrika* 6 (2-3): 302–10. http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/6.2-3.302 papers2://publication/doi/10.1093/biomet/6.2-3.302.

Wherry, R. 1931. "A new formula for predicting the shrinkage of the coefficient of multiple correlation." *The Annals of Mathematical Statistics* 2 (4): 440–57. http://www.jstor.org/stable/2957681$/backslash$npapers2://publication/uuid/F3D4916B-BB98-4094-A459-DF4387AC9610.