The Correlation Coefficient

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Outline

The Bivariate Normal Distribution

The Covariance Distribution

The Correlation Distribution

The Transformation Distributions

Discussion

Homework

The Bivariate Normal Distribution

If a random variable $\underset{2\times 1}{x}$ has a normal distribution with mean vector $\underset{2\times 1}{\mu}$ and variance-covariance matrix $\underset{2\times 2}{\Sigma}$, then

$$\underset{2\times 1}{x}=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \underset{2\times 1}{\mu}=\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \underset{2\times 2}{\Sigma}=\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

mean vector

mean vector

$$f(\underset{2\times 1}{x} \mid \underset{2\times 1}{\mu}, \underset{2\times 2}{\Sigma}) = (2\pi)^{-p/2} \, |\Sigma|^{-1/2} \, e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

covariance matrix

covariance matrix

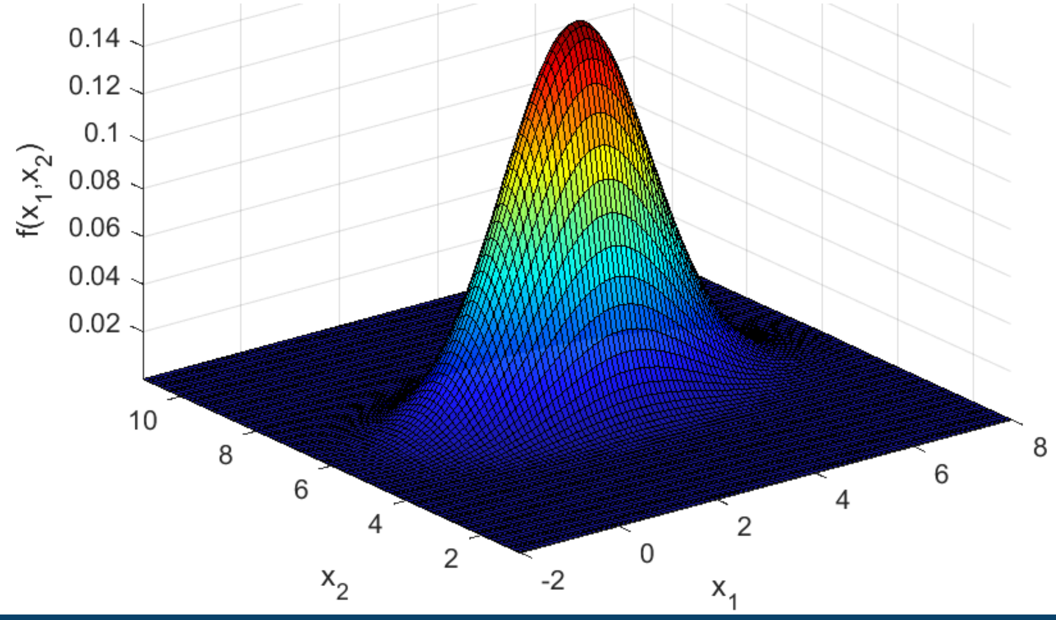
$$x, \mu \in \mathbb{R}^p \qquad p = 2 \qquad \Sigma > 0$$

set of pos def matrices

mean vector

covariance matrix

and we write $\underset{2\times 1}{x} \sim N(\underset{2\times 1}{\mu}, \underset{2\times 2}{\Sigma})$. The covariance matrix $\underset{2\times 2}{\Sigma}$, has to well-conditioned for an inverse.



The Bivariate Normal Distribution

This form may be more familiar

$$\underset{2\times 1}{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$f_X(x_1, x_2 \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}Q}$$

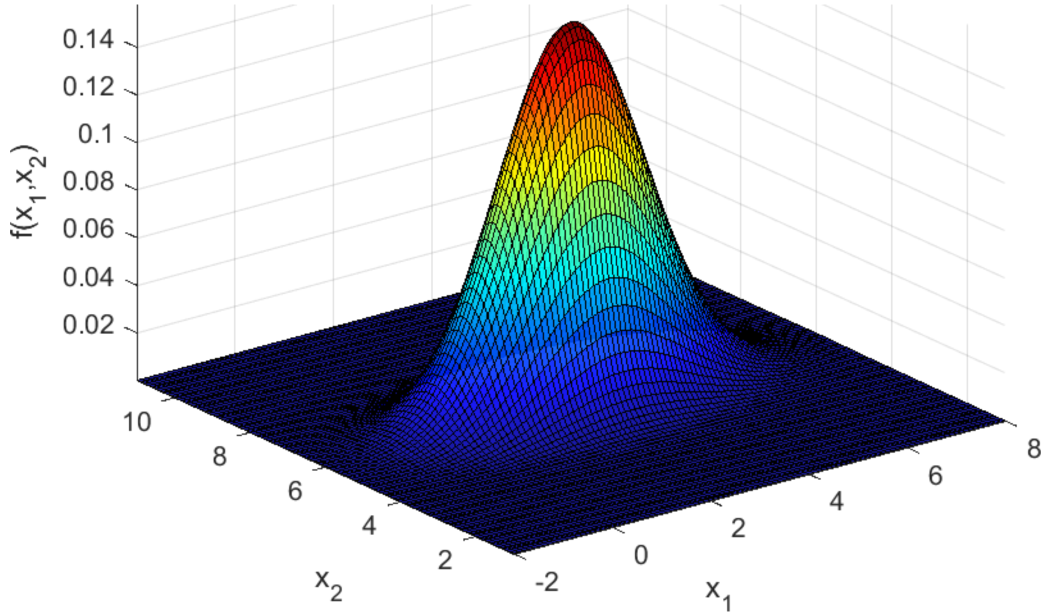
$$\underset{2\times 1}{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$Q = \frac{1}{(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]$$

$$\underset{2\times 2}{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

$$\sigma_1 > 0, \sigma_2 > 0, -1 < \rho < 1$$

$$\rho = \sigma_{12} / (\sigma_1\sigma_2) \qquad \sigma_{12} = \mathrm{cov}(x_1, x_2)$$



to avoid matrices.

The Covariance Distribution

In multivariate statistics if $\underset{2\times1}{x_1}, \underset{2\times1}{x_2},.., \underset{2\times1}{x_n}$ are IID $N(\underset{2\times1}{\mu}, \underset{2\times2}{\Sigma})$

and we calculate the covariance matrix

$$\underset{2\times2}{S} = \frac{1}{n-1}\sum_{i=1}^{n} \underbrace{(\underset{2\times1}{x_i} - \underset{2\times1}{\bar{x}})}_{2\times1}\underbrace{(\underset{2\times1}{x_i} - \underset{2\times1}{\bar{x}})'}_{1\times2} = \begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix}$$, then the joint PDF of

$$\underset{2\times1}{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \underset{2\times1}{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \underset{2\times2}{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

the covariance matrix $\underset{2\times2}{S}$ has a Wishart distribution

$$f(\underset{2\times2}{S} \mid \underset{2\times2}{\Sigma}, \nu) = k_W \left|\underset{2\times2}{\Sigma / \nu}\right|^{-\frac{\nu}{2}} \left|\underset{2\times2}{S}\right|^{\frac{\nu-\overset{p}{2}-1}{2}} e^{-\frac{1}{2}tr\left((\underset{2\times2}{\Sigma/\nu})^{-1}\underset{2\times2}{S}\right)} .$$

just a function of 3 variables

normalizing constant

$S,\Sigma > 0$

$\nu = n-1$

$tr()$=trace

If p=1

$$f(s^2 \mid \nu, \sigma^2) = k\left|\frac{\sigma^2}{\nu}\right|^{-\frac{\nu}{2}} \left|s^2\right|^{\frac{\nu-\overset{p}{1}-1}{2}} e^{-\frac{1}{2}\left(\frac{\sigma^2}{\nu}\right)^{-1}s^2}$$

The Covariance Distribution

The Wishart matrix probability density function

$$f(\underset{2\times 2}{S}|\Sigma,\nu)=k_W\,|\underset{2\times 2}{\Sigma}/\nu|^{-\frac{\nu}{2}}\,|\underset{2\times 2}{S}|^{\frac{\nu-p-1}{2}}\,e^{-\frac{1}{2}tr(\underset{2\times 2}{\Sigma}/\nu)^{-1}\underset{2\times 2}{S}}\;.$$

$$k_W^{-1}=2^{\frac{\nu p}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{j=1}^{p}\Gamma\left(\frac{\nu+1-j}{2}\right)$$

is the joint PDF of s_1^2, s_2^2, and s_{12}.

with mean, variance, and covariance of its elements

$$E(\underset{2\times 2}{S}|\Sigma,\nu)=\underset{2\times 2}{\Sigma}$$

$$\mathrm{var}(S_{ij}|\Sigma,\nu)=(\Sigma_{ij}^2+\Sigma_{ii}\Sigma_{jj})/\nu$$

$$i=1,2\quad j=1,2$$

$$\mathrm{cov}(S_{ij}S_{kl}|\Sigma,\nu)=(\Sigma_{ik}\Sigma_{jl}+\Sigma_{il}\Sigma_{jk})/\nu$$

$$i=1,2\quad j=1,2\quad k=1,2\quad l=1,2$$

If p=1

$$E(s|\sigma^2,\nu)=\sigma^2$$

$$var(s^2|\sigma^2,\nu)=\sigma^4/\nu$$

$$\underset{2\times 2}{S}=\begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix}$$

$$\underset{2\times 2}{\Sigma}=\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

The Covariance Distribution

We are often interested in the marginal PDF of the elements of S.

$$\underset{2\times 2}{S} = \begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix} \text{ which estimates } \underset{2\times 2}{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}.$$

Theorem:
$Q = A\ \ S\ \ A'$
$Q \sim W(\Delta = A\,\Sigma\,A' / \nu, \nu)$
$A=[1,0]$ *or* $A=[0,1]$

It can be shown that the variance s_1^2 has PDF

$$s_1^2 \sim \Gamma(\alpha = \frac{\nu}{2}, \beta_1 = \frac{2\sigma_1^2}{\nu}) \text{ , AKA } \frac{(n-1)s_1^2}{\sigma_1^2} \sim \chi^2(n-1) \ ,$$

$$E(s_i^2) = \sigma_i^2$$
$$\mathrm{var}(s_i^2) = 2\sigma_i^4 / \nu$$
$$i = 1,2$$

and the variance s_2^2 has PDF

$$s_2^2 \sim \Gamma(\alpha = \frac{\nu}{2}, \beta_2 = \frac{2\sigma_2^2}{\nu}) \text{ , AKA } \frac{(n-1)s_2^2}{\sigma_2^2} \sim \chi^2(n-1) \ ,$$

but the covariance has a more complicated marginal PDF.

https://en.wikipedia.org/wiki/Wishart_distribution#Marginal_distribution_of_matrix_elements

The Covariance Distribution

$$\underset{2\times 2}{S} = \begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix}$$

$$\underset{2\times 2}{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

The covariance s_{12} has the *Variance-Gamma* distribution

$$f(s_{12}) = \frac{\nu |\nu s_{12}|^{\frac{\nu-1}{2}}}{\Gamma(\nu/2)\sqrt{2^{\nu-1}\pi(1-\rho^2)(\sigma_1\sigma_2)^{\nu+1}}} K_{\frac{\nu-1}{2}}\left(\frac{|\nu s_{12}|}{\sigma_1\sigma_2(1-\rho^2)}\right)\exp\left(\frac{\rho\nu s_{12}}{\sigma_1\sigma_2(1-\rho^2)}\right).$$

K is the modified Bessel function of the second kind

The Variance-Gamma marginal PDF for s_{12} can also be written as

$$f(s_{12}) = \frac{\gamma^{2\lambda}|s_{12}-\mu_{s_{12}}|^{\lambda-\frac{1}{2}}}{\sqrt{\pi}\Gamma(\lambda)(2\alpha)^{\lambda-\frac{1}{2}}} K_{\lambda-\frac{1}{2}}\left(\alpha|s_{12}-\mu_{s_{12}}|\right)e^{\beta(s_{12}-\mu_{s_{12}})}$$

with mean and variance identified as

$$E(s_{12}) = \mu_{s_{12}} + \frac{2\beta\lambda}{\gamma^2} \quad \text{and} \quad \operatorname{var}(s_{12}) = \frac{2\lambda}{\gamma^2}\left(1+\frac{2\beta^2}{\gamma^2}\right).$$

The Covariance Distribution

$$\underset{2\times 1}{\mu} = \begin{pmatrix} 67 \\ 150 \end{pmatrix} \qquad \underset{2\times 2}{\Sigma} = \begin{pmatrix} 4 & 2 \\ 2 & 16 \end{pmatrix} \qquad \nu = 9$$

Example: Generated $\underset{2\times 1}{x_1}, \underset{2\times 1}{x_2}, \ldots, \underset{2\times 1}{x_{10}}$ from $N(\underset{2\times 1}{\mu}, \underset{2\times 2}{\Sigma})$ and calculated $\underset{2\times 1}{\bar{x}}$, subtracted mean $\underset{2\times 1}{\bar{x}}$ from each, transpose multiplied each deviation $\underset{2\times 2}{(x_i - \bar{x})'(x_i - \bar{x})}$, added the n=10 squared deviations and divided by ν=n-1=9 to form $\underset{2\times 2}{S} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})'(x_i - \bar{x})$. Repeated L=10^6 times to get $\underset{2\times 2}{S_{(1)}}, \ldots, \underset{2\times 2}{S_{(L)}}$. The $\underset{2\times 2}{S}$'s are now $W(\underset{2\times 2}{\Sigma/\nu}, \nu)$.

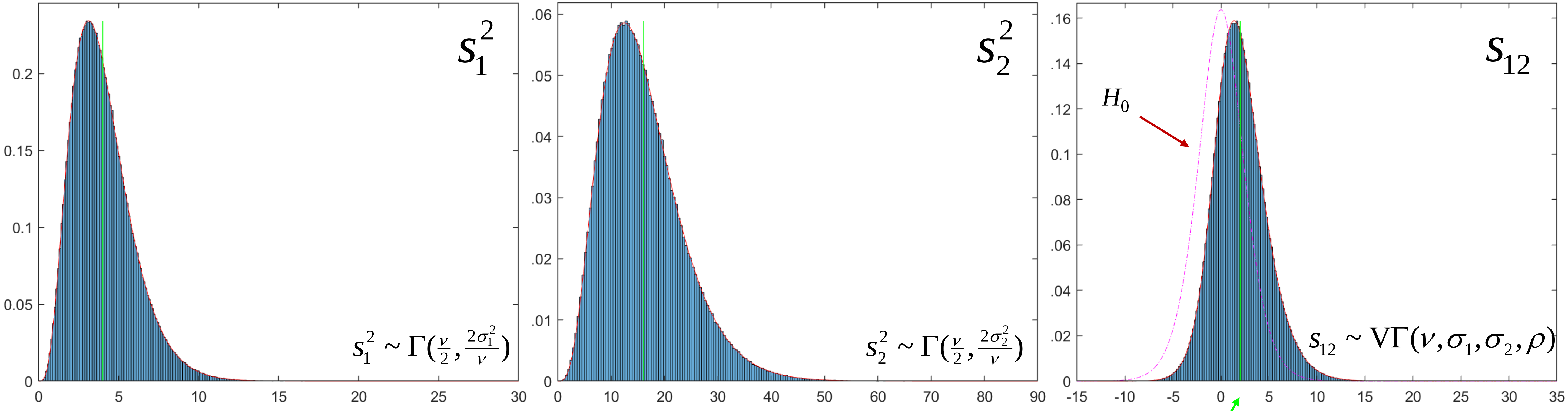
$$f(\underset{2\times 2}{S} \mid \Sigma, \nu) = k_W \left|\underset{2\times 2}{\Sigma / \nu}\right|^{-\frac{\nu}{2}} \left|\underset{2\times 2}{S}\right|^{\frac{\nu-p-1}{2}} e^{-\frac{1}{2} tr(\underset{2\times 2}{\Sigma/\nu})^{-1} \underset{2\times 2}{S}}$$

The Covariance Distribution

$$\underset{2\times1}{\mu} = \begin{pmatrix} 67 \\ 150 \end{pmatrix} \quad \underset{2\times2}{\Sigma} = \begin{pmatrix} 4 & 2 \\ 2 & 16 \end{pmatrix}$$

$$\nu = 9$$

The S's, $S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})'(x_i - \bar{x})$ are now $W(\Sigma/\nu,\nu)$.



$E(s_{12})$

$$E(S\,|\,\Sigma,\nu) = \Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 16 \end{pmatrix} \qquad \mathrm{var}(S_{ij}\,|\,\Sigma,\nu) = (\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj})/\nu = \begin{pmatrix} 3.56 & 7.56 \\ 7.56 & 56.89 \end{pmatrix}$$

$$\Sigma = AA'$$

Note histograms normalized with exact PDF superimposed.

$$A = \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{15} \end{pmatrix}$$

The Covariance Distribution

```
clear all
close all
rng('default')
warning off

% set parameters
nbins=200;
n=10; m=10^6;
mu=[67;150]; % True mean
Sigma=[4,2;2,16] % Alternative Hypothesis Cov
%Sigma=[4,0;0,16] % Null Hypothesis Cov
rho=Sigma(1,2)/sqrt(Sigma(1,1)*Sigma(2,2))
A=chol(Sigma)';
nu=n-1; a=nu/2;
b11=2*Sigma(1,1)/nu; b22=2*Sigma(2,2)/nu;
b12=2*Sigma(1,2)/nu;
% generate data
zz=A*randn(2,n*m)+mu;
xx=reshape(zz(1,:),[n,m]);
yy=reshape(zz(2,:),[n,m]);
clear zz
```

```
% calculate statistics
xbar=mean(xx); ybar=mean(yy);
simVarX= sum((xx-repmat(xbar,n,1)).*(xx-repmat(xbar,n,1)))'/nu;
simVarY= sum((yy-repmat(ybar,n,1)).*(yy-repmat(ybar,n,1)))'/nu;
simCovXY=sum((xx-repmat(xbar,n,1)).*(yy-repmat(ybar,n,1)))'/nu;
simCorXY=simCovXY./sqrt(simVarX.*simVarY);

% mean x
figure;
histogram(xbar,nbins,'normalization','pdf')
xlim([mu(1,1)-5*sqrt(Sigma(1,1)/n),mu(1,1)+5*sqrt(Sigma(1,1)/n)])

% mean y
figure;
histogram(ybar,nbins,'normalization','pdf')
xlim([mu(2,1)-5*sqrt(Sigma(2,2)/n),mu(2,1)+5*sqrt(Sigma(2,2)/n)])
```

The Covariance Distribution

```
% var x
[mean(simVarX),var(simVarX)]
Es11=Sigma(1,1);, vars11=2*Sigma(1,1)^2/nu;
figure;
H=histogram(simVarX,nbins,'normalization','pdf');
sorted=(sortrows(H.Values')); maxval=sorted(nbins,1);
xlim([0,30]), ylim([0,1.05*maxval])
hold on
fs11 = @(s11) s11^(a-1)*exp(-s11/b11)/(gamma(a)*b11^a);
fplot(fs11,[0,35],'r')
line([Sigma(1,1) Sigma(1,1)], [0 maxval],'Color','green')
xlim([0,30]), ylim([0,1.05*maxval])
```

```
% var y
[mean(simVarY),var(simVarY)]
Es22=Sigma(2,2), vars22=2*Sigma(2,2)^2/nu
figure;
H=histogram(simVarY,nbins,'normalization','pdf');
sorted=(sortrows(H.Values')); maxval=sorted(nbins,1);
xlim([0,90]), ylim([0,1.05*maxval])
hold on
fs22 = @(s22) s22^(a-1)*exp(-s22/b22)/(gamma(a)*b22^a);
fplot(fs22,[0,90],'r')
line([Sigma(2,2) Sigma(2,2)], [0 maxval],'Color','green')
xlim([0,90]), , ylim([0,1.05*maxval])
```

The Covariance Distribution

```
% cov x,y
[mean(simCovXY),var(simCovXY)]
Es22=Sigma(1,2)
figure;
H=histogram(simCovXY,nbins,'normalization','pdf');
sorted=(sortrows(H.Values')); maxval=sorted(nbins,1);
xlim([-15,35]), ylim([0,1.05*maxval])
sorted=(sortrows(H.Values')); maxval=sorted(nbins,1);
line([Sigma(1,2) Sigma(1,2)], [0 maxval],'Color','green')
hold on
fs12 = @(s12) nu*abs(s12*nu)^((nu-1)/2)/( ...
gamma(nu/2)*sqrt( 2^(nu-1)*pi*(1-rho^2)*...
sqrt( Sigma(1,1)*Sigma(2,2))^(nu+1) ) )...
*besselk( (nu-1)/2,abs(s12)*nu/((1-rho^2)*...
sqrt(Sigma(1,1)*Sigma(2,2))) )...
*exp( rho*s12*nu/((1-rho^2)*sqrt(Sigma(1,1)*Sigma(2,2))) );
muK=0;
alphaK=nu/((1-rho^2)*sqrt(Sigma(1,1)*Sigma(2,2)));
betaK=rho*alphaK;
lambdaK=nu/2;
gammaK=(1-rho^2)^2;
Es12=muK+2*betaK*lambdaK/gammaK^2;
fplot(fs12,[-15,35],'r')
xlim([-15,35]), ylim([0,1.05*maxval])
rho0=0; %null hypothesis distribution
fs12 = @(s12) nu*abs(s12*nu)^((nu-1)/2)/( gamma(nu/2)...
*sqrt( 2^(nu-1)*pi*(1-rho0^2)*sqrt( Sigma(1,1)*Sigma(2,2))^(nu+1) )
)...
*besselk( (nu-1)/2,abs(s12*nu)/((1-
rho0^2)*sqrt(Sigma(1,1)*Sigma(2,2))) )...
*exp( rho0*s12*nu/((1-rho0^2)*sqrt(Sigma(1,1)*Sigma(2,2))) );
fplot(fs12,[-15,35],'m-.')
line([Sigma(1,2) Sigma(1,2)], [0 maxval],'Color','green')
xlim([-15,35]), ylim([0,1.05*maxval])
```

The Correlation Distribution

From the variances s_1^2 , s_2^2 , and covariance s_{12} we can perform a transformation of variable to obtain the correlation coefficient $r = \dfrac{s_{12}}{s_1 s_2}$.

It has been shown that the alternative hypothesis ($\rho \neq 0$) PDF is

$$f(r) = \frac{n-2}{\sqrt{2\pi}} \frac{\Gamma(n-1)}{\Gamma\left(n-\frac{1}{2}\right)} \frac{(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}}}{(1-\rho r)^{n-\frac{3}{2}}} \quad {}_2F_1\left(\tfrac{1}{2},\tfrac{1}{2},n-\tfrac{1}{2},\tfrac{1}{2}(1+\rho r)\right)$$

F is the hypergeometric function

which under the null hypothesis ($\rho=0$) becomes

$$f(r \mid H_0) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}} .$$

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

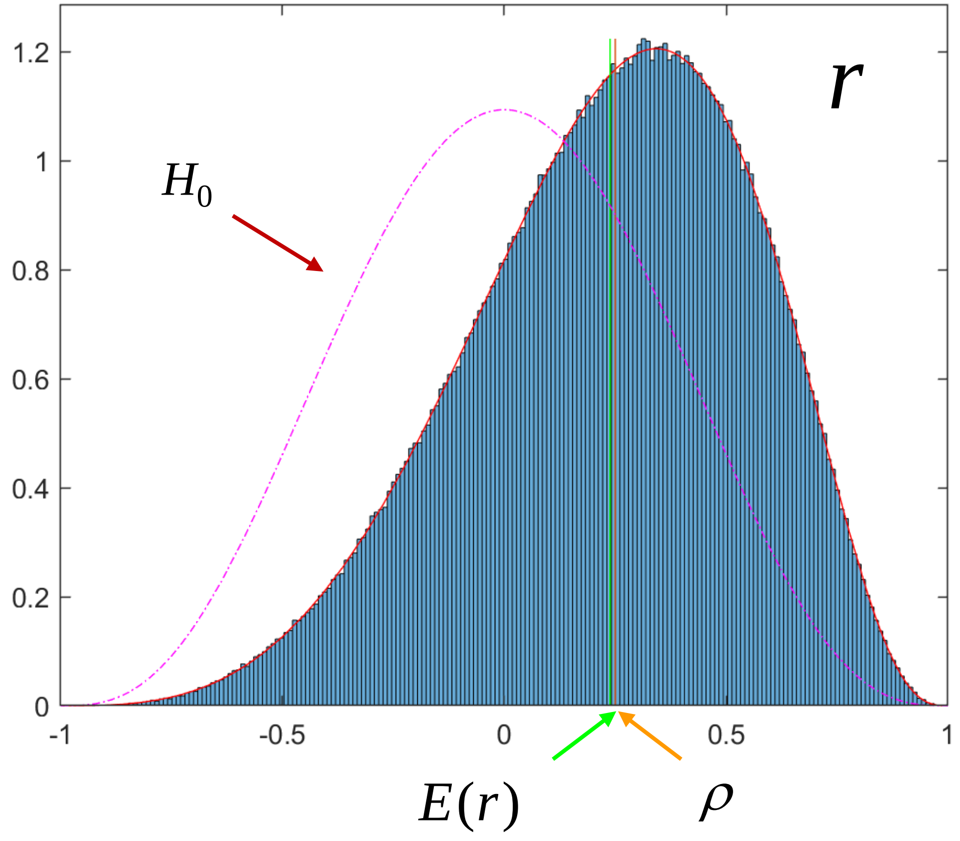
Hotelling: New Light on the Correlation Coefficient and its Transforms. JRSS-B, 15(2), 193-232, 1953.

The Correlation Distribution

Example: Using the same $S_{(1)},\ldots,S_{(L)}$ calculated $r_{(1)},\ldots,r_{(L)}$.

$$\underset{2\times 2}{\Sigma}=\begin{pmatrix}4 & 2\\ 2 & 16\end{pmatrix}$$

$$\rho=0.25$$



$$E(r)\approx 0.2383$$

$$\bar{r}_{adj}=0.2453$$

Of note is that $E(r)$ is biased.

$$E(r)=\int_{-1}^{1} rf(r)\,dr$$

$$f(r)=\frac{n-2}{\sqrt{2\pi}}\frac{\Gamma(n-1)}{\Gamma\left(n-\frac{1}{2}\right)}\frac{(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}}}{(1-\rho r)^{n-\frac{3}{2}}}\;{}_2F_1\left(\tfrac{1}{2},\tfrac{1}{2},n-\tfrac{1}{2},\tfrac{1}{2}(1+\rho r)\right)$$

$$E(r)=\rho+(1-\rho^2)\left(-\frac{\rho}{2n}-\frac{\rho-9\rho^3}{8n^2}+\frac{\rho+42\rho^3-75\rho^5}{16n^3}+\cdots\right)$$

$$E(r)\approx\rho-\frac{\rho(1-\rho^2)}{2n}\quad\longrightarrow\quad r_{adj}\approx r\left[1+\frac{1-r^2}{2n}\right]$$

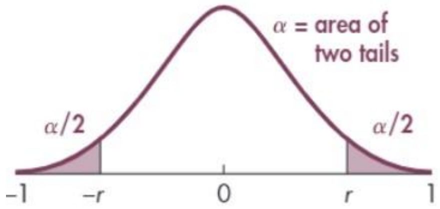
The Correlation Distribution

```
% cor x,y
figure;
H=histogram(simCorXY,nbins,'normalization','pdf');
sorted=(sortrows(H.Values')); maxval=sorted(nbins,1);
xlim([-1,1]), ylim([0,1.05*maxval])
sorted=(sortrows(H.Values')); maxval=sorted(nbins,1);
%print(gcf,'-dtiffn','-r200',['frhist'])
hold on
fr = @(r) (n-2)*gamma(n-1)*(1-rho^2)^((n-1)/2)*(1-r^2)^((n-4)/2)/...
( sqrt(2*pi)*gamma(n-1/2)*(1-rho*r)^(n-3/2) )...
*hypergeom([1/2,1/2],(2*n-1)/2,(rho*r+1)/2);
fplot(fr,[-1,1],'r')
Er=rho-rho*(1-rho^2)/2/n %biased
radj=mean( simCorXY.*(1+(1-simCorXY.^2)/2/n) )
line([Er, Er], [0 maxval],'Color','green')
line([rho,rho], [0 maxval],'Color',[0.8500 0.3250 0.0980])
fr0 = @(r) (gamma((n-1)/2)/gamma((n-2)/2)/sqrt(pi))*(1-r.^2).^((n-4)/2);
fplot(fr0,[-1,1],'m-.')
xlim([-1,1]), ylim([0,1.05*maxval])
```

TABLE 11

Critical Values of r When $\rho = 0$

The entries in this table are the critical values of r for a two-tailed test at α. For simple correlation, df $= n - 2$, where n is the number of pairs of data in the sample. For a one-tailed test, the value of α shown at the top of the table is double the value of α being used in the hypothesis test.



df \ α	0.10	0.05	0.02	0.01
1	0.988	0.997	1.000	1.000
2	0.900	0.950	0.980	0.990
3	0.805	0.878	0.934	0.959
4	0.729	0.811	0.882	0.917
5	0.669	0.754	0.833	0.875
6	0.621	0.707	0.789	0.834
7	0.582	0.666	0.750	0.798
8	0.549	0.632	0.715	0.765
9	0.521	0.602	0.685	0.735
10	0.497	0.576	0.658	0.708
11	0.476	0.553	0.634	0.684
12	0.458	0.532	0.612	0.661
13	0.441	0.514	0.592	0.641
14	0.426	0.497	0.574	0.623
15	0.412	0.482	0.558	0.606
16	0.400	0.468	0.543	0.590
17	0.389	0.456	0.529	0.575
18	0.378	0.444	0.516	0.561
19	0.369	0.433	0.503	0.549
20	0.360	0.423	0.492	0.537
25	0.323	0.381	0.445	0.487
30	0.296	0.349	0.409	0.449
35	0.275	0.325	0.381	0.418
40	0.257	0.304	0.358	0.393
45	0.243	0.288	0.338	0.372
50	0.231	0.273	0.322	0.354
60	0.211	0.250	0.295	0.325
70	0.195	0.232	0.274	0.302
80	0.183	0.217	0.256	0.283
90	0.173	0.205	0.242	0.267
100	0.164	0.195	0.230	0.254

For specific details about using this table to find p-values and critical values, see pages 621–623.

Johnson & Kuby

The Transformation Distributions

The exact PDF for r is generally difficult for non-Statisticians to understand, let alone get percentiles from it for hypothesis testing and/or confidence intervals.

The true ($\rho \neq 0$) PDF is also not needed for hypothesis testing.

So generally transformations of r that have “friendly” PDFs are used.

The Transformation Distributions

It has been shown that under the null hypothesis (ρ=0)

$$f(r \mid H_0) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}}$$

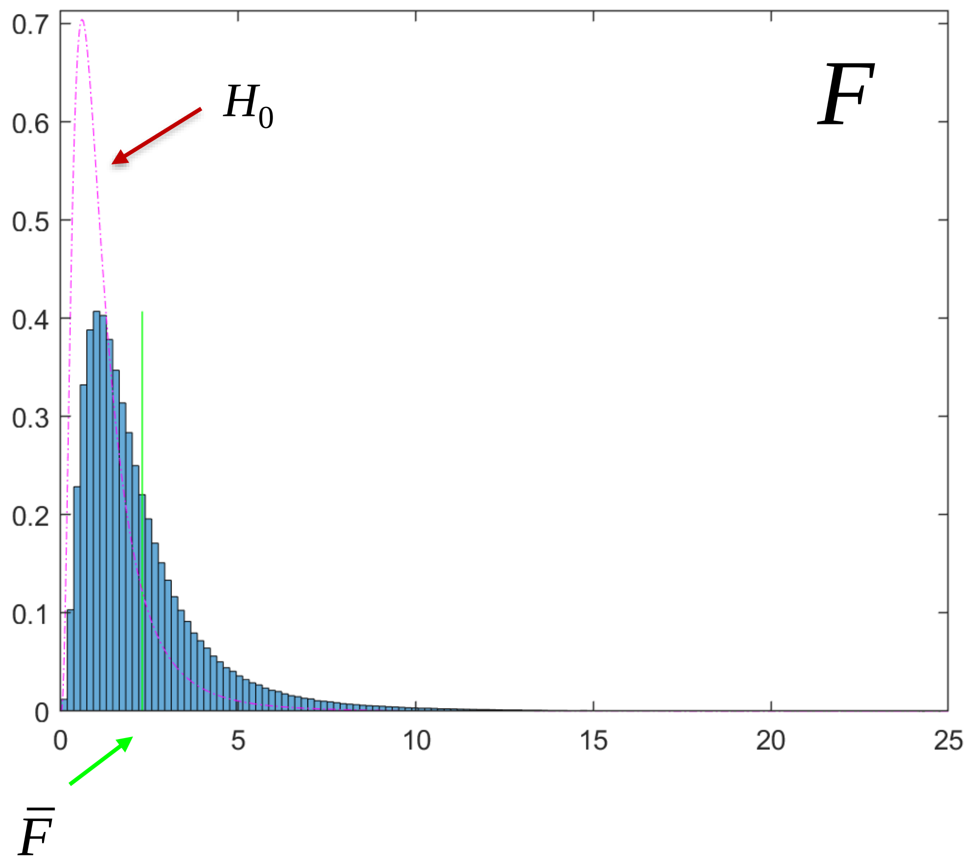
the transformation $F = \dfrac{1+r}{1-r}$ can be made resulting in F having an F distribution with n-2 numerator and n-2 denominator degrees of freedom, $F \sim F(n\text{-}2, n\text{-}2)$.

The Transformation Distributions

Example: Using the same $S_{(1)},\ldots,S_{(L)}$ calculated $F_{(1)},\ldots,F_{(L)}$.

$$\underset{2\times 2}{\Sigma}=\begin{pmatrix}4 & 2\\ 2 & 16\end{pmatrix}$$

$$\rho=0.25$$



There is not an expression for F under the alternative hypothesis. (No red curve on histogram.) Simulation can be used to build the alternative distribution.

$$F=\frac{1+r}{1-r}$$

The Transformation Distributions

It has been shown that under the null hypothesis (ρ=0)

$$f(r \mid H_0) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}}$$

the transformation $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ can be made resulting in z having a normal distribution, $z \mid H_0 \sim N\left(0, \frac{1}{n-3}\right)$.

So now we can form confidence intervals and perform hypothesis testing.

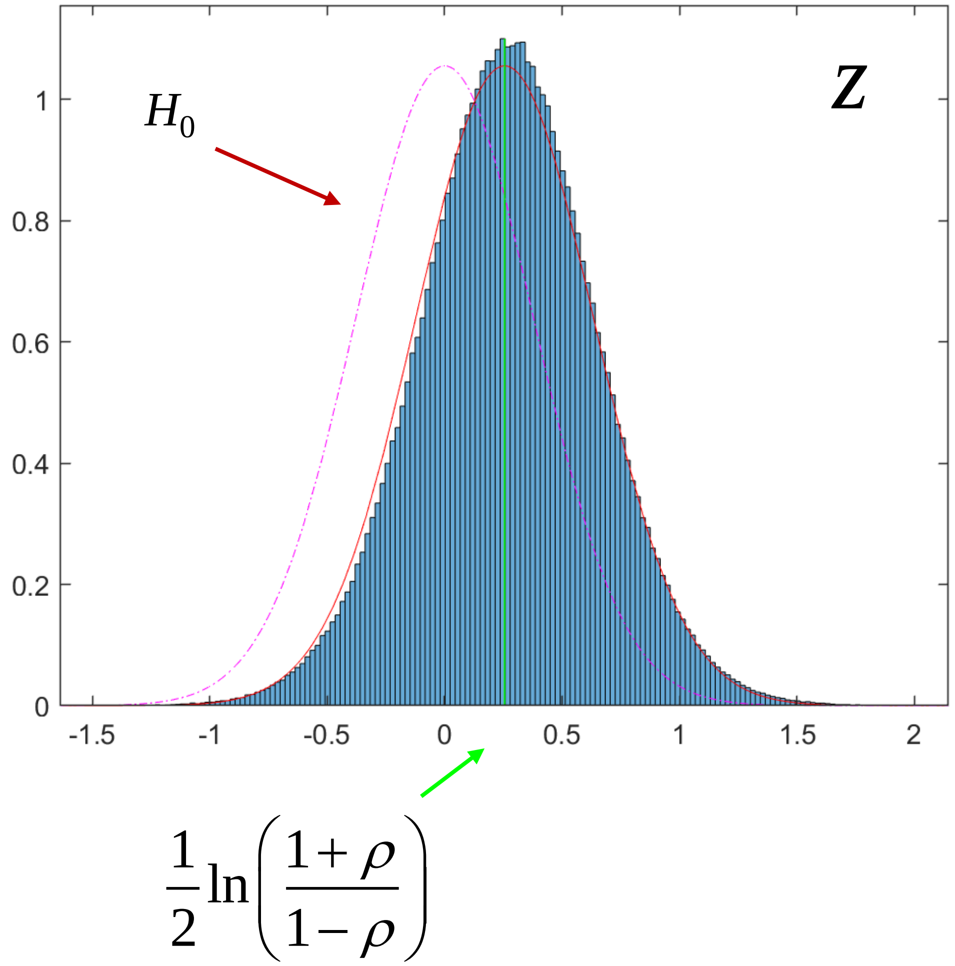
https://en.wikipedia.org/wiki/Fisher_transformation

The Transformation Distributions

Example: Using the same $S_{(1)},\ldots,S_{(L)}$ calculated $z_{(1)},\ldots,z_{(L)}$.

$$\underset{2\times 2}{\Sigma}=\begin{pmatrix}4 & 2\\ 2 & 16\end{pmatrix}$$

$$\rho=0.25$$



There is not an expression for z under the alternative hypothesis.

$$z=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

But an approximation exists.

$$z\overset{\circ}{\sim}N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right),\frac{1}{n-3}\right)$$

It is good in the tails for significance.

Looks like needs a little negative skewness.

https://en.wikipedia.org/wiki/Fisher_transformation

The Transformation Distributions

It has been shown that under the null hypothesis (ρ=0)

$$f(r \mid H_0) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}}$$

the transformation $f = \dfrac{r^2(n-2)}{1-r^2}$ can be made resulting in f having an

F distribution with 1 numerator and n-2 denominator degrees of

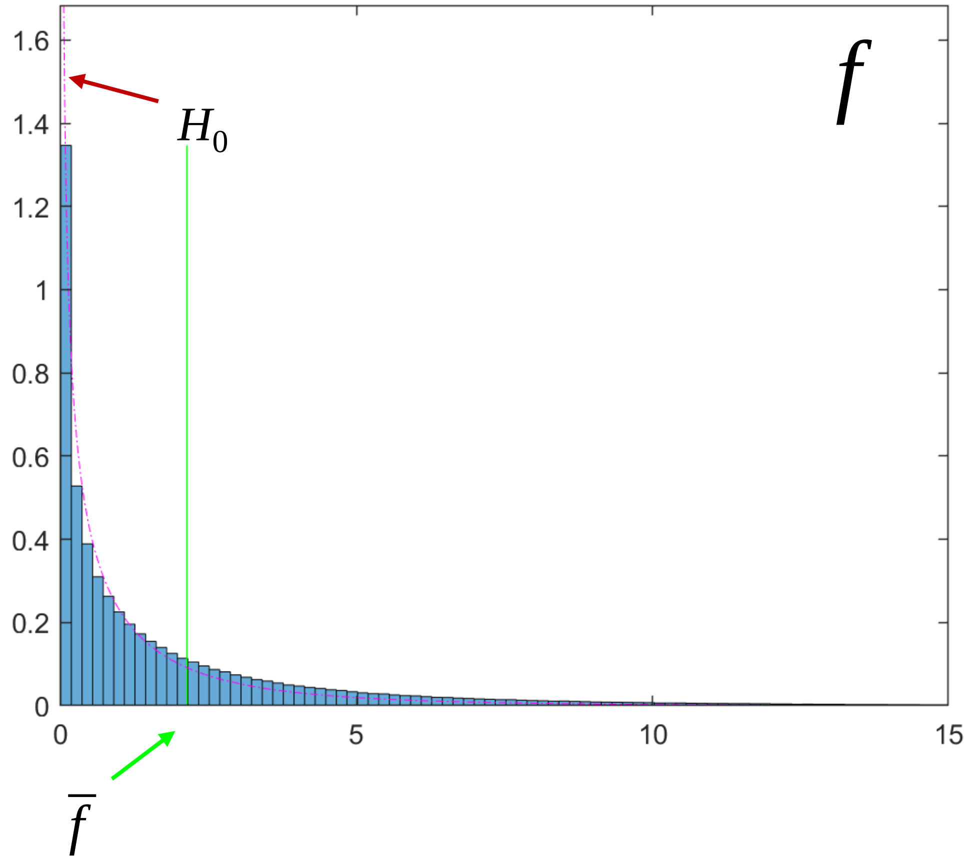
freedom, $F \sim F(1,n\text{-}2)$.

The Transformation Distributions

Example: Using the same $S_{(1)},\ldots,S_{(L)}$ calculated $f_{(1)},\ldots,f_{(L)}$.

$$\underset{2\times 2}{\Sigma}=\begin{pmatrix}4 & 2\\ 2 & 16\end{pmatrix}$$

$$\rho=0.25$$



There is not an expression for f under the alternative hypothesis. (No red curve on histogram.)

Simulation can be used to build the alternative distribution.

This statistic isn't as discriminative.

$$f=\frac{r^2(n-2)}{1-r^2}$$

The Transformation Distributions

It has been shown that under the null hypothesis (ρ=0)

$$f(r \mid H_0) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{n-2}{2}\right)}(1-r^2)^{\frac{\nu-3}{2}}$$

the transformation $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ can be made resulting in t having an t distribution with 1 numerator and n-2 denominator degrees of freedom, $t \sim t(n\text{-}2)$.

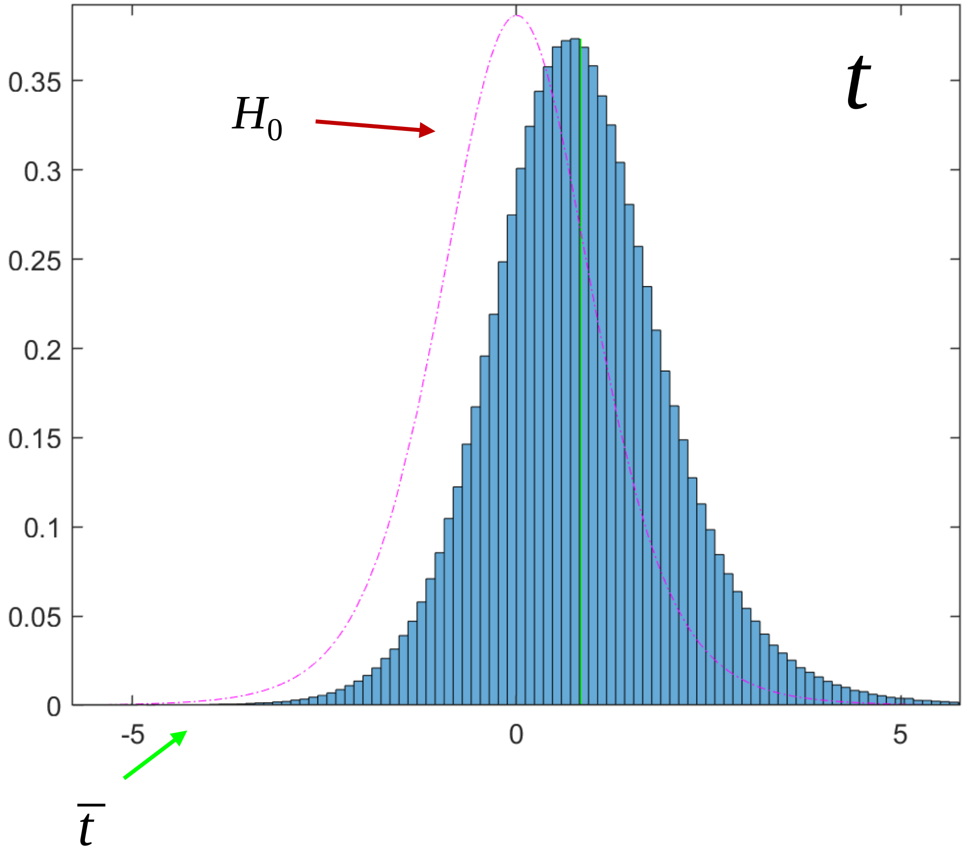
https://en.wikipedia.org/wiki/Fisher_transformation

The Transformation Distributions

Example: Using the same $S_{(1)},\ldots,S_{(L)}$ calculated $t_{(1)},\ldots,t_{(L)}$.

$$\underset{2\times 2}{\Sigma}=\begin{pmatrix} 4 & 2 \\ 2 & 16 \end{pmatrix}$$

$$\rho=0.25$$



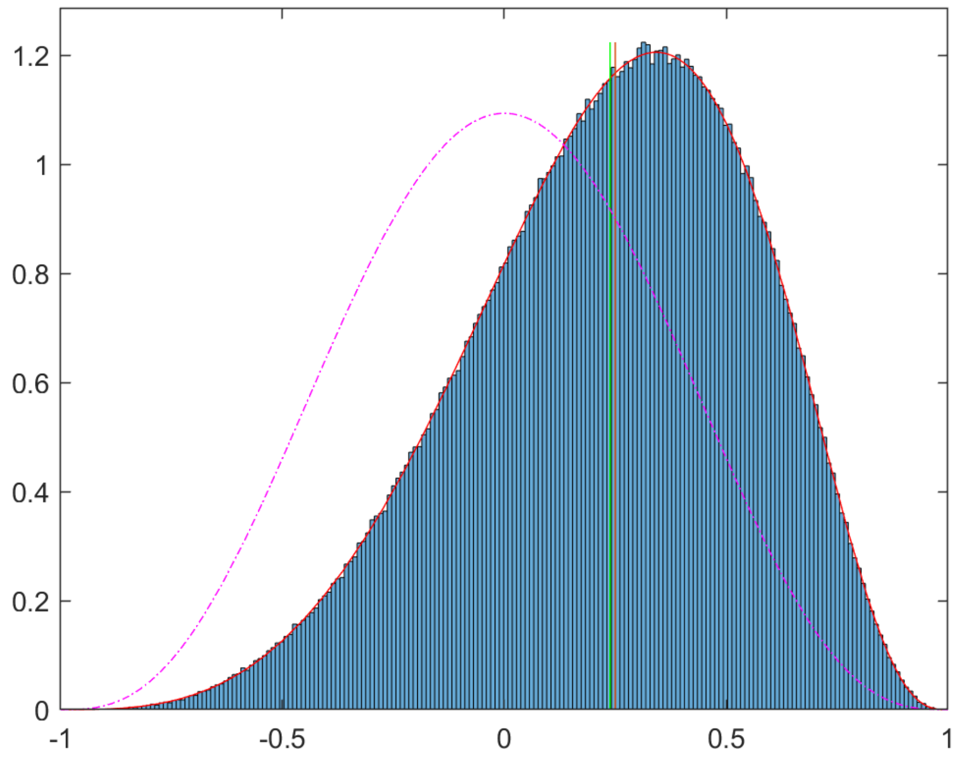
There is not an expression for t under the alternative hypothesis. (No red curve on histogram.) Simulation can be used to build the alternative distribution.

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

https://en.wikipedia.org/wiki/Fisher_transformation

Discussion

There are many complicated subtleties to learn about the correlation.
Since we are confident with our math and computation abilities,
I recommend that we work with the exact null distribution for



$$f(r|H_0)=\frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}}$$

calculate percentiles and estimate by

$$r_{adj}\approx r\left[1+\frac{1-r^2}{2n}\right]$$

Discussion

Questions?

Homework 5

1. Generate L=10^6 data sets of size n=15. Use $\underset{2\times 1}{\mu}=\begin{pmatrix} 67 \\ 150 \end{pmatrix}$ and $\underset{2\times 2}{\Sigma}=\begin{pmatrix} 4 & 4 \\ 4 & 16 \end{pmatrix}$.

   Calculate s_{12} and r from each set.
   Make a normalized histogram of the s_{12}'s and superimpose $f(s_{12})$.
   Calculate the sample mean and variance of the s_{12}'s and compare to the expected values. Comment.

2. Make a normalized histogram of the r's and superimpose $f(r)$.
   Calculate the sample mean and variance of the r's and compare to the approximate expected values. Comment.

Homework 5

**3. Generate one additional data set of size $n=15$ and compute r.

Perform a hypothesis test of H_0: $\rho=0$ vs. H_1: $\rho\neq 0$.

Compute the 2.5^{th} and 97.5^{th} percentile of $f(r|H_0)$.

Reject the null hypothesis if r less than 2.5^{th} percentile or larger than the 97.5^{th} percentile.

$$f(r\,|\,H_0)=\frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}}$$

**For students in MSSC 5770 that want to demonstrate how smart they are.

Homework 5

4. Convert each of your L=10^6 r's to

$$F=\frac{1+r}{1-r} \qquad z=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad f=\frac{r^2(n-2)}{1-r^2} \qquad t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

determine the 2.5^{th} and 97.5^{th} percentile of each statistic.

Compare your simulation percentiles to a theoretical percentile if possible.

Covert your one additional r from #3 to each statistic.

Perform a hypothesis test for H_0: ρ=0 vs. H_1: $\rho\neq0$ from each.

Do you get the same results from each hypothesis type?

Homework 5

**5. Make up your own interesting problem to solve about r.
Present any theoretical or simulation results and any data results.
Be imaginative and interesting.

**For students in MSSC 5770 that want to demonstrate how smart they are.

Homework 5

**6. For each of ρ=0, .2, .4, .6, and .8, generate L=10^6 data sets of size n=15.

Calculate r from each so you have 5 sets of L=10^6 r's.

On the same graph plot the 5 histograms.

When ρ=0, find the 95^{th} percentile $r_{.95}$. This is α=0.05.

For each of ρ=.2, .4, .6, and .8, find the fraction less than $r_{.95}$.

The fraction less than $r_{.95}$ is β. $P(\text{not reject } H_0|H_0 \text{ False})=\beta$.

Make a plot of ρ vs. β. i.e. $(\rho_{.0},\beta_{.0}),(\rho_{.2},\beta_{.2}), (\rho_{.4},\beta_{.4}), (\rho_{.6},\beta_{.6}), (\rho_{.8},\beta_{.8})$.

For $(\rho_{.0},\beta_{.0})$ use $(0,.95)$.

Comment.

Repeat for each of F, z, f, and t.

**For students in MSSC 5770 that want to demonstrate how smart they are.

$$\underset{2\times 1}{\mu}=\begin{pmatrix}67\\150\end{pmatrix}$$

$$\underset{2\times 2}{\Sigma}=\begin{pmatrix}4 & 8\rho\\ 8\rho & 16\end{pmatrix}$$

	H_0 True	H_0 False
Fail to Reject H_0	Type A Correct Decision $(1\text{-}\alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	Type B Correct Decision $(1\text{-}\beta)$