Notes 4: Correlation and Geometry

Nathaniel E. Helwig

Department of Statistics
University of Illinois at Urbana-Champaign

Stat 420: Methods of Applied Statistics
Section N1U/N1G – Spring 2014

Outline of Notes

1) Pearson's Correlation:

- Population vs. sample
- Important properties
- Sampling distribution

2) Correlation and Regression:

- Simple linear regression
- Reinterpreting correlation
- Connecting the two

3) Inferences with Correlations:

- Hypothesis testing ($\rho = 0$)
- Hypothesis testing ($\rho \neq 0$)
- GPA Example

4) Geometrical Interpretations:

- Some sum-of-squares
- Sum-of-squares total
- Correlation coefficient

Pearson's Correlation Coefficient: Population

Pearson's product-moment correlation coefficient is defined as

$$\begin{aligned}\rho_{XY} &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \\ &= \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\mathrm{E}[(X-\mu_X)^2]}\sqrt{\mathrm{E}[(Y-\mu_Y)^2]}}\end{aligned}$$

where

- σ_{XY} is the population covariance between X and Y
- σ_X^2 is the population variance of X
- σ_Y^2 is the population variance of Y

Pearson's Correlation Coefficient: Sample

Given a sample of observations (x_i, y_i) for $i \in \{1, \ldots, n\}$, Pearson's product-moment correlation coefficient is defined as

$$
\begin{aligned}
r_{xy} &= \frac{s_{xy}}{s_x s_y} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}
\end{aligned}
$$

where

- $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ is sample covariance between x_i and y_i
- $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ is sample variance of x_i
- $s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ is sample variance of y_i

Properties of Pearson's Correlation

ρ_{XY} measures *linear* dependence between X and Y

- Not about prediction…just a measure of linear dependence
- CORRELATION $\neq$ CAUSATION

ρ_{XY} is bounded: $\quad -1 \leq \rho_{XY} \leq 1$

- $\rho_{XY} = -1$ implies perfect negative linear relationship
- $\rho_{XY} = 1$ implies perfect positive linear relationship
- $\rho_{XY} = 0$ implies no linear relationship

ρ_{XY} is independent of the units of measurement of X and Y

Properties of Pearson's Correlation (continued)

Magnitude of ρ_{XY} is unaffected by linear transformations

- Suppose that $W = aX + b$ and $Z = cY + d$
- If $\mathrm{sign}(a) = \mathrm{sign}(c)$, then $\rho_{WZ} = \rho_{XY}$
- If $\mathrm{sign}(a) \neq \mathrm{sign}(c)$, then $\rho_{WZ} = -\rho_{XY}$

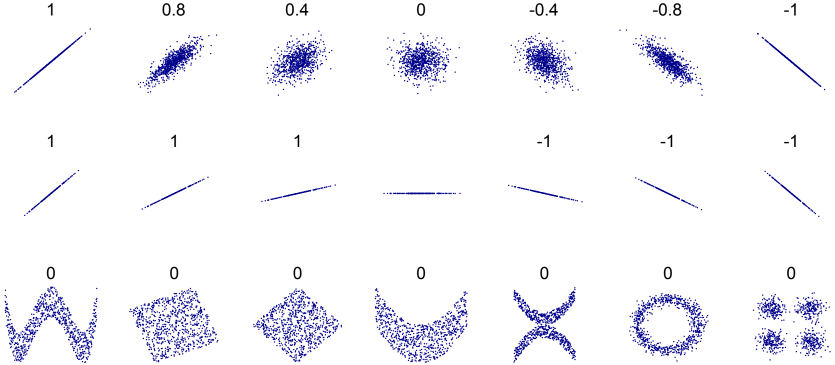
Sample correlation r_{xy} is sensitive to outliers

ρ_{XY} can be affected by moderator variables

- Need to think about possible moderator variables
- Can have different patterns of correlation in different subgroups

Visualization of Pearson's Correlation

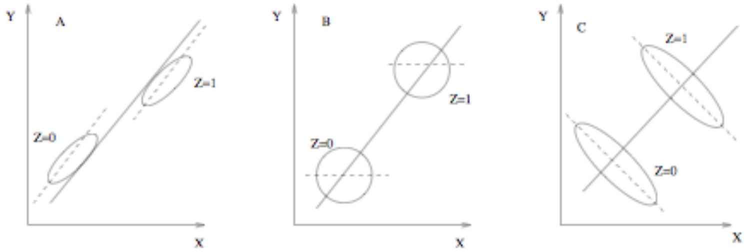
Plots of x_i versus y_i for different r_{xy} values:



From http://en.wikipedia.org/wiki/File:Correlation_examples2.svg

Visualization of Pearson's Correlation (continued)

Plots of x_i versus y_i for two groups of observations:



From *Practical Regression and Anova using R*, Faraway (2002)

Example #1: Pizza Data

The owner of Momma Leona's Pizza restaurant chain believes that if a restaurant is located near a college campus, then there is a linear relationship between sales and the size of the student population. Suppose data were collected from a sample of 10 Momma Leona's Pizza restaurants located near college campuses.

Population (1000s): x	2	6	8	8	12	16	20	20	22	26
Sales ($1000s): y	58	105	88	118	117	137	157	169	149	202

We want to find the correlation between student population (x) and quarterly pizza sales (y).

Example #1: Correlation Calculation

Remember from the definition: $\hat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

Remember from the SLR notes:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Example #1: Correlation Calculation (continued)

	x	y	x^2	y^2	xy
	2	58	4	3364	116
	6	105	36	11025	630
	8	88	64	7744	704
	8	118	64	13924	944
	12	117	144	13689	1404
	16	137	256	18769	2192
	20	157	400	24649	3140
	20	169	400	28561	3380
	22	149	484	22201	3278
	26	202	676	40804	5252
$\sum$	140	1300	2528	184730	21040

$$\bar{x} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{1300}{10} = 130$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2528 - 10(14^2) = 568$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 184730 - 10(130^2) = 15730$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 21040 - 10(14)(130) = 2840$$

Example #1: Correlation Calculation (continued)

Using the results from the previous slides:

$$\begin{aligned}\hat{r} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{2840}{\sqrt{568}\sqrt{15730}} \\ &= 0.950123\end{aligned}$$

Strong positive correlation: as the number of students (x) increases, the quarterly pizza sales (y) increases linearly.

Sampling Distribution of r

Fisher (1929) derived the sampling distribution of Pearson's r:

$$f(r) = \frac{n-2}{\pi}\left(1-\rho^2\right)^{\frac{1}{2}(n-1)}\left(1-r^2\right)^{\frac{1}{2}(n-4)}\int_0^\infty [\cosh(z)-\rho r]^{-(n-1)}\,\mathrm{d}z$$

where

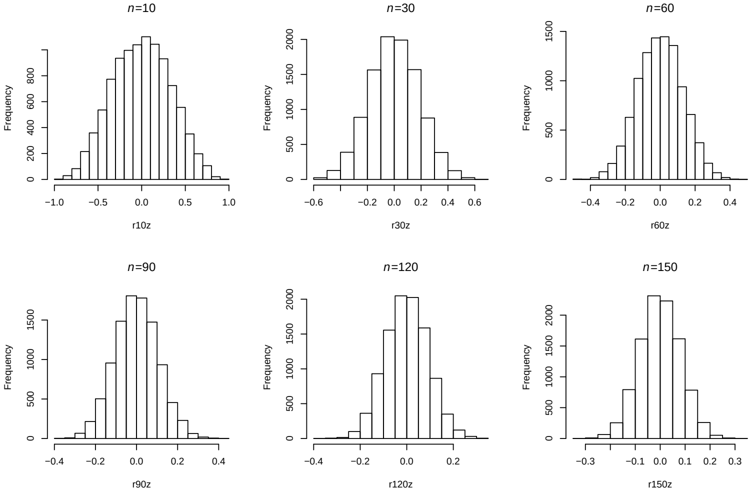
- ρ is the true population correlation coefficient
- n is the observed sample size

For a given ρ and n, we can simulate r using:

```
rsim<-function(rho,n){
  x=rnorm(n)
  y=rho*x+rnorm(n,sd=sqrt(1-rho^2))
  cor(x,y)
}
```

Empirical Distribution with $\rho = 0$

If $\rho = 0$ then $f(r)$ is symmetric and approximately normal for large n:



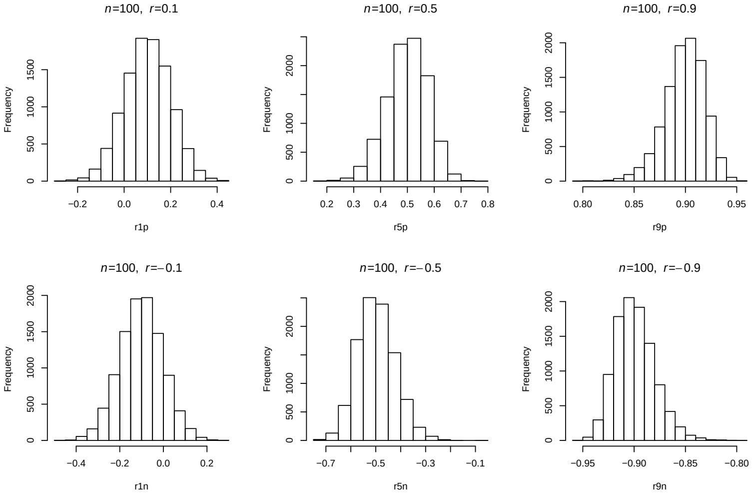
Empirical Distribution with $\rho = 0$ (continued)

R code:

```
> set.seed(1234)
> r10z=replicate(10000,rsim(rho=0,n=10))
> r30z=replicate(10000,rsim(rho=0,n=30))
> r60z=replicate(10000,rsim(rho=0,n=60))
> r90z=replicate(10000,rsim(rho=0,n=90))
> r120z=replicate(10000,rsim(rho=0,n=120))
> r150z=replicate(10000,rsim(rho=0,n=150))
> par(mfrow=c(2,3))
> hist(r10z,main=expression(italic(n)*"="*10))
> hist(r30z,main=expression(italic(n)*"="*30))
> hist(r60z,main=expression(italic(n)*"="*60))
> hist(r90z,main=expression(italic(n)*"="*90))
> hist(r120z,main=expression(italic(n)*"="*120))
> hist(r150z,main=expression(italic(n)*"="*150))
```

Empirical Distribution with $\rho \neq 0$

If $\rho \neq 0$ then $f(r)$ is skewed in opposite direction of correlation sign:



Empirical Distribution with $\rho \neq 0$ (continued)

R code:

```
> set.seed(1234)
> r1p=replicate(10000,rsim(rho=0.1,n=100))
> r5p=replicate(10000,rsim(rho=0.5,n=100))
> r9p=replicate(10000,rsim(rho=0.9,n=100))
> r1n=replicate(10000,rsim(rho=-0.1,n=100))
> r5n=replicate(10000,rsim(rho=-0.5,n=100))
> r9n=replicate(10000,rsim(rho=-0.9,n=100))
> par(mfrow=c(2,3))
> hist(r1p,main=expression(italic(n)*"="*100*",  "*italic(r)*"="*0.1))
> hist(r5p,main=expression(italic(n)*"="*100*",  "*italic(r)*"="*0.5))
> hist(r9p,main=expression(italic(n)*"="*100*",  "*italic(r)*"="*0.9))
> hist(r1n,main=expression(italic(n)*"="*100*",  "*italic(r)*"="*-0.1))
> hist(r5n,main=expression(italic(n)*"="*100*",  "*italic(r)*"="*-0.5))
> hist(r9n,main=expression(italic(n)*"="*100*",  "*italic(r)*"="*-0.9))
```

Simple Linear Regression Model

The simple linear regression model has the form

$$y_i = b_0 + b_1 x_i + e_i$$

for $i \in \{1, \ldots, n\}$ where

- $y_i \in \mathbb{R}$ is the real-valued response for the i-th observation
- $b_0 \in \mathbb{R}$ is the regression intercept
- $b_1 \in \mathbb{R}$ is the regression slope
- $x_i \in \mathbb{R}$ is the predictor for the i-th observation
- $e_i \stackrel{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is Gaussian measurement error

Ordinary Least Squares Solution

The ordinary least squares (OLS) problem is

$$\min_{b_0, b_1 \in \mathbb{R}} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

and the OLS solution has the form

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$ and $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$

Revisiting Sample Correlation r_{xy}

Note that we can rewrite the sample correlation coefficient as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$= \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

where

- $z_{x_i} = \frac{x_i - \bar{x}}{s_x}$ with $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- $z_{y_i} = \frac{y_i - \bar{y}}{s_y}$ with $s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$

Revisiting Sample Correlation r_{xy} (continued)

Note that z_{x_i} and z_{y_i} are standardized to have mean 0 and variance 1.

- z_{x_i} and z_{y_i} are Z-scores

Thus, the sample correlation coefficient

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}$$

is the sample covariance of the standardized scores of X and Y.

Connecting r_{xy} to OLS Solution

First convert x_i and y_i into Z-scores:

$$z_{x_i} = \frac{x_i - \bar{x}}{s_x} \qquad\qquad z_{y_i} = \frac{y_i - \bar{y}}{s_y}$$

Next suppose we want to predict z_{y_i} from z_{x_i}.

Now plug in the Z-scores to the OLS solution:

$$\hat{b}_0^{(z)} = \bar{z}_y - \hat{b}_1^{(z)} \bar{z}_x = 0$$
$$\hat{b}_1^{(z)} = \frac{\sum_{i=1}^n (z_{x_i} - \bar{z}_x)(z_{y_i} - \bar{z}_y)}{\sum_{i=1}^n (z_{x_i} - \bar{z}_x)^2} = r_{xy}$$

because $\bar{z}_x = \bar{z}_y = 0$ and $\frac{1}{n-1}\sum_{i=1}^n (z_{x_i} - \bar{z}_x)^2 = \frac{1}{n-1}\sum_{i=1}^n z_{x_i}^2 = 1$

Connecting r_{xy} to OLS Solution (continued)

Thus, we have that $\hat{z}_{y_i} = r_{xy} z_{x_i}$ is the fitted value for i-th observation.

In general, the relationship between r_{xy} and b_1 is:

$$\begin{aligned} b_1 &= \frac{s_{xy}}{s_x^2} \\ &= \left(\frac{s_{xy}}{s_x s_y} \right) \frac{s_y}{s_x} \\ &= r_{xy} \frac{s_y}{s_x} \end{aligned}$$

which implies that $r_{xy} = b_1 \frac{s_x}{s_y}$.

Connecting r_{xy} to SSR

Remember from the SLR notes that $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i = \bar{y} + \hat{b}_1(x_i - \bar{x})$ because $\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}$

Plugging $\hat{y}_i = \bar{y} + \hat{b}_1(x_i - \bar{x})$ into the definition of *SSR* produces

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}\hat{b}_1^2(x_i - \bar{x})^2$$

Now plugging in $\hat{b}_1 = \hat{r}_{xy}\frac{s_y}{s_x}$ to R^2 defintion produces

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}\hat{b}_1^2(x_i - \bar{x})^2}{(n-1)s_y^2}$$

$$= \frac{\sum_{i=1}^{n}\left(\hat{r}_{xy}\frac{s_y}{s_x}\right)^2(x_i - \bar{x})^2}{(n-1)s_y^2} = \hat{r}_{xy}^2$$

Example #1: Connecting $\hat{r}_{xy}$ to $\hat{b}_1$

From the SLR notes, remember that:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 2840/568 = 5$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x} = 130 - 5(14) = 60$$

And from the previous correlation calculations, remember that:

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{2840}{\sqrt{568}\sqrt{15730}} = 0.950123$$

Note that $0.950123 = \hat{r} = \hat{b}_1 \frac{s_x}{s_y} = 5\frac{\sqrt{568}}{\sqrt{15730}} = 0.950123$

Testing for Non-Zero Correlation

In most cases we want to test if there is a linear relationship:

$$H_0 : \ \rho_{XY} = 0$$
$$H_1 : \ \rho_{XY} \neq 0$$

If (X, Y) follow a bivariate normal distribution with $\rho = 0$ and if $\{(x_i, y_i)\}_{i=1}^{n}$ are independent samples then

$$T^* = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1 - r_{xy}^2}} \sim t_{n-2}$$

so we reject H_0 if $|T^*| \geq t_{n-2}^{(\alpha/2)}$ where $t_{n-2}^{(\alpha/2)}$ is critical t_{n-2} value such that $\mathrm{P}(T \geq t_{n-2}^{(\alpha/2)}) = \alpha/2$

Fisher z-Transformation

If $\rho \neq 0$, then we can use Fisher's z-transformation:

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

If (X, Y) follow a bivariate normal distribution and if $\{(x_i, y_i)\}_{i=1}^{n}$ are independent samples then z is approximately normal with

$$\mathrm{E}(z) = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$$
$$\mathrm{V}(z) = \frac{1}{n-3}$$

where ρ is the true population correlation coefficient.

Testing for Arbitrary Correlation

In some cases we want to test if there is a particular correlation:

$$H_0: \; \rho_{XY} = \rho_0$$
$$H_1: \; \rho_{XY} \neq \rho_0$$

In this case, we use Fisher's z-transformation; first define the standardized variable $Z^* = \left[z - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right]\sqrt{n-3}$

We reject H_0 if $|Z^*| \geq Z_{\alpha/2}$ where $Z_{\alpha/2}$ is critical Z value such that $\mathrm{P}(Z \geq Z_{\alpha/2}) = \alpha/2$

Confidence Intervals for r_{xy}

To form a CI around r_{xy} we use Fisher's z-transformation to form a CI on the transformed scale:

$$z \pm Z_{\alpha/2}/\sqrt{n-3}$$

Then we need to transform z limits back to r:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Example #1: Correlation Inference Questions

Returning to Momma Leona's Pizza example, suppose we want to...

(a) Test if there is a significant linear relationship between student population (x) and quarterly pizza sales (y), i.e., test $H_0 : \rho_{XY} = 0$ versus $H_1 : \rho_{XY} \neq 0$. Use $\alpha = .05$ significance level.

(b) Test $H_0 : \rho_{XY} = 0.8$ versus $H_1 : \rho_{XY} \neq 0.8$. Use $\alpha = .05$ level.

(b) Test $H_0 : \rho_{XY} = 0.8$ versus $H_1 : \rho_{XY} > 0.8$. Use $\alpha = .05$ level.

(d) Make a 90% CI for ρ_{XY}.

Example #1: Answer 1a

Question: Test $H_0 : \rho_{XY} = 0$ versus $H_1 : \rho_{XY} \neq 0$. Use $\alpha = .05$.

The needed t test statistic is

$$T^* = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0.950123\sqrt{8}}{\sqrt{1-0.950123^2}} = 8.616749$$

which follows a t_8 distribution.

The critical t_8 values are $t_8^{(.025)} = -2.306004$ and $t_8^{(.975)} = 2.306004$, so our decision is

$$t_8^{(.975)} = 2.306004 < 8.616749 = T^* \Longrightarrow \text{ Reject } H_0$$

Example #1: Answer 1b

Question: Test $H_0 : \rho_{XY} = 0.8$ versus $H_1 : \rho_{XY} \neq 0.8$. Use $\alpha = .05$.

First form the z-transformed variable

$$z = 0.5 \ln \left(\frac{1 + \hat{r}}{1 - \hat{r}} \right) = 0.5 \ln \left(\frac{1.950123}{0.04987704} \right) = 1.833043$$

which is approximately normal with mean and variance

$$\mathrm{E}(z) = 0.5 \ln \left(\frac{1 + \rho_0}{1 - \rho_0} \right) = 0.5 \ln \left(\frac{1.8}{0.2} \right) = 1.098612$$

$$\mathrm{V}(z) = \frac{1}{n - 3} = 1/7$$

under the null hypothesis $H_0 : \rho_{XY} = 0.8$.

Example #1: Answer 1b (continued)

Question: Test $H_0 : \rho_{XY} = 0.8$ versus $H_1 : \rho_{XY} \neq 0.8$. Use $\alpha = .05$.

Now form the standardized variable

$$Z^* = \frac{z - z_0}{\sqrt{\mathrm{V}(z)}} = \frac{1.833043 - 1.098612}{1/\sqrt{7}} = 1.943122$$

which is approximately $\mathrm{N}(0, 1)$ under $H_0 : \rho_{XY} = 0.8$.

The critical Z values are $Z_{.025} = -1.959964$ and $Z_{.975} = 1.959964$, so our decision is

$$Z_{.975} = 1.959964 > 1.943122 = Z^* \implies \text{Retain } H_0$$

Example #1: Answer 1c

Question: Test $H_0 : \rho_{XY} = 0.8$ versus $H_1 : \rho_{XY} > 0.8$. Use $\alpha = .05$.

We have the same transformed variable $z = 1.833043$ with $\mathrm{E}(z) = 1.098612$ and $\mathrm{V}(z) = 1/7$; results in the same

$$Z^* = \frac{z - z_0}{\sqrt{\mathrm{V}(z)}} = \frac{1.833043 - 1.098612}{1/\sqrt{7}} = 1.943122$$

which is approximately $\mathrm{N}(0, 1)$ under $H_0 : \rho_{XY} = 0.8$

The critical Z value is $Z_{.95} = 1.644854$, so our decision is

$$Z_{.95} = 1.644854 < 1.943122 = Z^* \implies \text{Reject } H_0$$

Example #1: Answer 1d

Question: Make a 90% CI for ρ_{XY}.

First form the z-transformed variable

$$z = 0.5 \ln\left(\frac{1+\hat{r}}{1-\hat{r}}\right) = 0.5 \ln\left(\frac{1.950123}{0.04987704}\right) = 1.833043$$

which is approximately normal with variance $\mathrm{V}(z) = 1/7$.

The critical Z value is $Z_{.95} = 1.644854$, so the 90% CI is given by

$$z \pm Z_{.95}\sqrt{\mathrm{V}(z)} = 1.833043 \pm 1.644854\sqrt{1/7} = [1.211347;\ 2.45474]$$

and converting the z limits back to the correlation scale produces

$$\left[\frac{e^{2(1.211347)}-1}{e^{2(1.211347)}+1};\ \frac{e^{2(2.45474)}-1}{e^{2(2.45474)}+1}\right] = [0.8370831;\ 0.9853554]$$

Data Overview

This example uses the *GPA* data set that we examined before.

- From http://onlinestatbook.com/2/regression/intro.html

Y: student's university grade point average.

X: student's high school grade point average.

Have data from $n = 105$ different students.

Correlation Calculation

Calculate Pearson's correlation with `cor` function:

```
> X=gpa$high_GPA
> Y=gpa$univ_GPA
> cor(X,Y)
[1] 0.7795631
```

Calculate Pearson's correlation with `cov` and `sd` functions:

```
> cov(X,Y)/(sd(X)*sd(Y))
[1] 0.7795631
```

Correlation Calculation (continued)

Calculate Pearson's correlation manually:

```
> mux=mean(X)
> muy=mean(Y)
> cxy=sum((X-mux)*(Y-muy))
> sx=sqrt(sum((X-mux)^2))
> sy=sqrt(sum((Y-muy)^2))
> cxy/(sx*sy)
[1] 0.7795631
```

Testing for Non-Zero Correlation

To test $H_0 : \rho_{XY} = 0$ versus $H_1 : \rho_{XY} \neq 0$ use the `cor.test` function:

```
> cor.test(X,Y)

 Pearson's product-moment correlation

data:  X and Y
t = 12.632, df = 103, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6911690 0.8449761
sample estimates:
      cor
0.7795631
```

Testing for Non-Zero Correlation (continued)

Note that we can get the same results manually using

```
> gpacr=cor(X,Y)
> tstar=gpacr*sqrt(length(X)-2)/sqrt(1-gpacr^2)
> tstar
[1] 12.63197
> 2*(1-pt(tstar,103))
[1] 0
> z=log((1+gpacr)/(1-gpacr))/2
> z
[1] 1.044256
> zlo=z-qnorm(.975)/sqrt(102)
> zhi=z+qnorm(.975)/sqrt(102)
> c(zlo,zhi)
[1] 0.8501905 1.2383212
> rlo=(exp(2*zlo)-1)/(exp(2*zlo)+1)
> rhi=(exp(2*zhi)-1)/(exp(2*zhi)+1)
> c(rlo,rhi)
[1] 0.6911690 0.8449761
```

Testing for Arbitrary Correlation

To test $H_0 : \rho_{XY} = 0.7$ versus $H_1 : \rho_{XY} \neq 0.7$ define `fisherz` function

```
fisherz=function(r,n,rho0=0){
  z=log((1+r)/(1-r))/2
  z0=log((1+rho0)/(1-rho0))/2
  zstar=(z-z0)*sqrt(n-3)
  pval=2*(1-pnorm(zstar))
  list(z=z,pval=pval)
}
```

and then use

```
> fisherz(cor(X,Y),105,rho0=0.7)
$z
[1] 1.044256

$pval
[1] 0.07391138
```

Testing for Arbitrary Correlation (continued)

Note that we could also test $H_0 : \rho_{XY} = 0.7$ versus $H_1 : \rho_{XY} \neq 0.7$ using the output from the `cor.test` function.

Output 95% CI from `cor.test` function is $[0.6911690, \ 0.8449761]$, which contains the null hypothesis value of $\rho_{XY} = 0.7$.

So, we retain the null hypothesis at the $\alpha = .05$ level.

Geometry of Sum-of-Squares

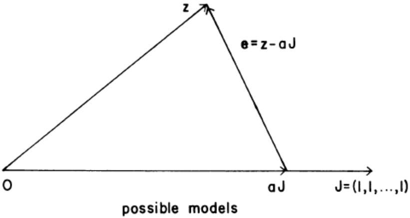


Figure 8. A simple statistical model.

Bryant, P. (1984). Geometry, statistics, probability: Variations on a common theme. *The American Statistician, 38*, 38–48.

$\mathbf{J}$ is an $n \times 1$ vector of ones

Let $a\mathbf{J}$ denote a constant vector

Let $\mathbf{z} = (z_1, \ldots, z_n)'$ denote any n-dimensional vector

$SS = \sum_{i=1}^{n}(z_i - a)^2 = \|\mathbf{e}\|^2$ with $\mathbf{e} = \mathbf{z} - a\mathbf{J}$

Geometry of Sum-of-Squares Total

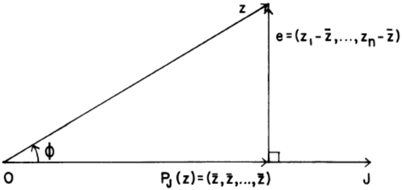


Figure 9. Derivation of the sample mean.

Bryant, P. (1984). Geometry, statistics, probability: Variations on a common theme. *The American Statistician, 38*, 38–48.

J is an $n \times 1$ vector of ones

- $n^{-1}\mathbf{J}\mathbf{J}'$ is projection matrix

Let $\mathbf{z} = (z_1, \ldots, z_n)'$ denote any n-dimensional vector

Let $P_J(\mathbf{z}) = n^{-1}\mathbf{J}\mathbf{J}'\mathbf{z} = \bar{z}\mathbf{J}$ denote the projection of **z** onto **J**

Note: $\mathbf{z} - P_J(\mathbf{z})$ is orthogonal to **J**

Geometry of Pearson's Correlation

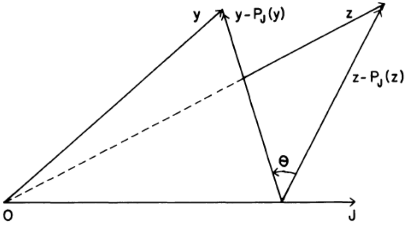


Figure 10. The simple correlation coefficient.

Bryant, P. (1984). Geometry, statistics, probability: Variations on a common theme. *The American Statistician, 38*, 38–48.

Let $\mathbf{y} = \{y_i\}_{i=1}^n$ and $P_J(\mathbf{y}) = \bar{y}\mathbf{J}$ denote the projection of $\mathbf{y}$ onto $\mathbf{J}$.

Correlation is cosine of angle between $\mathbf{y} - P_J(\mathbf{y})$ and $\mathbf{z} - P_J(\mathbf{z})$:

$$
\begin{aligned}
r &= \cos(\theta) \\
&= \frac{(\mathbf{y} - P_J(\mathbf{y}))'(\mathbf{z} - P_J(\mathbf{z}))}{\|\mathbf{y} - P_J(\mathbf{y})\|\|\mathbf{z} - P_J(\mathbf{z})\|}
\end{aligned}
$$