Correlation
Edpsy 580

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBRANA-CHAMPAIGN

Overview: Correlation & Regression

- Pearson correlation coefficient
- Simple Linear Regression.
  - What and why?
  - How (interpretation, estimation & diagnostics).
  - Statistical Inference.
  - Comments regarding interpretation.
- Bi-variate regression
- Multiple regression
- General Linear Model

Outline: Pearson Correlation Coefficient

- Definition & Properties.
- Statistical Inference
  - t-test that correlation equals 0.
  - Fisher's Z-Transformation.
  - Confidence intervals for ρ.
  - Test of $H_o : \rho = K$.
  - Test of $H_o : \rho_1 = \rho_2$ (2 independent populations).

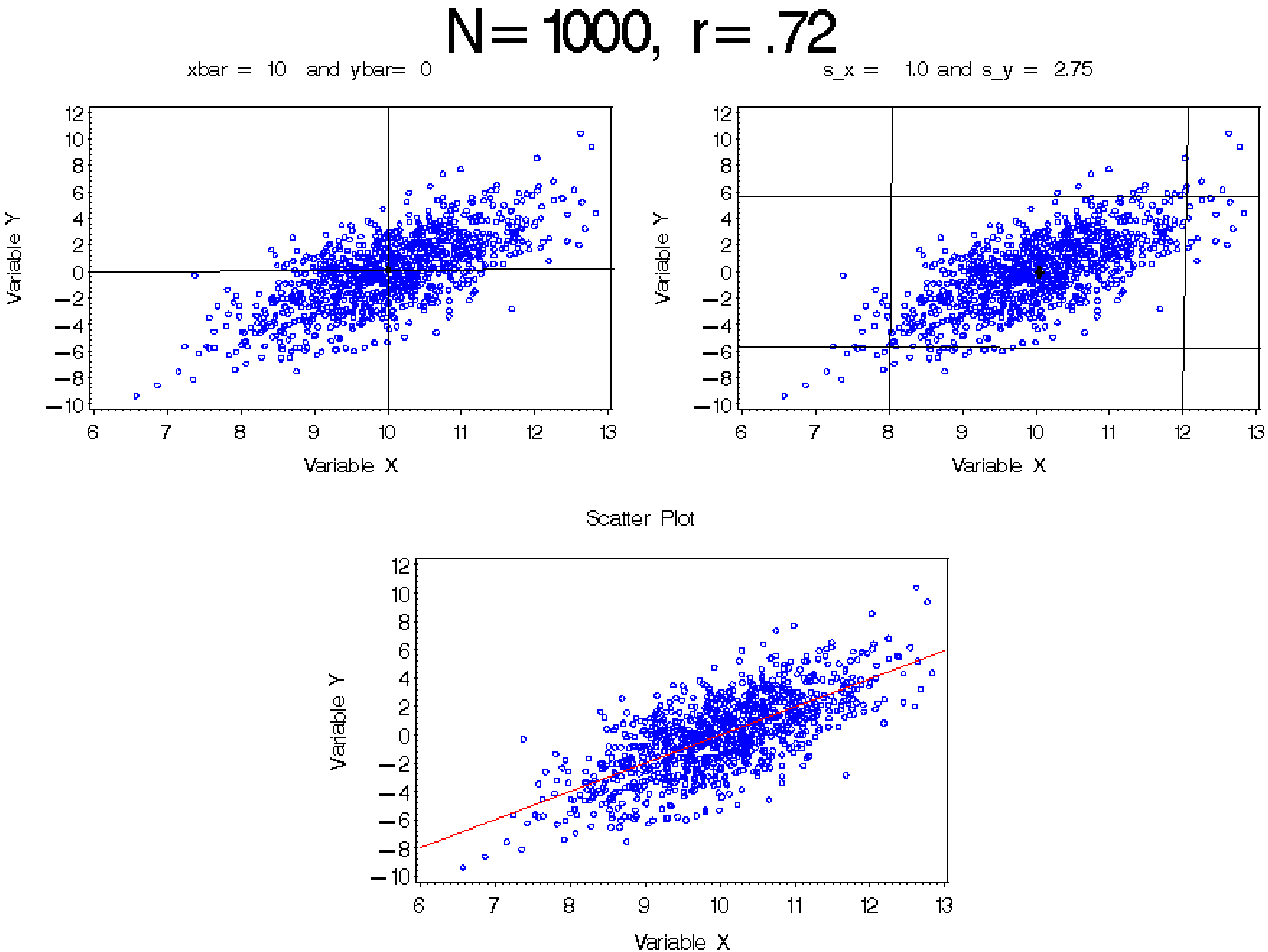
Correlation: Definition & Properties

- "Pearson Product Moment Correlation"
- Two numerical variables measured on same individual,

  (X_i, Y_i) for $i = 1, \ldots, n$. e.g.,

  - Height and weight.
  - Math and science scores.
  - Salary and merit.
  - High school GPA and college GPA.
  - Cost of wine and annual rainfall.
  - Conservative Party donors and people who buy garden bulbs by mail.

Scatter Diagram & Summary Statistics



Definition: Correlation coefficient

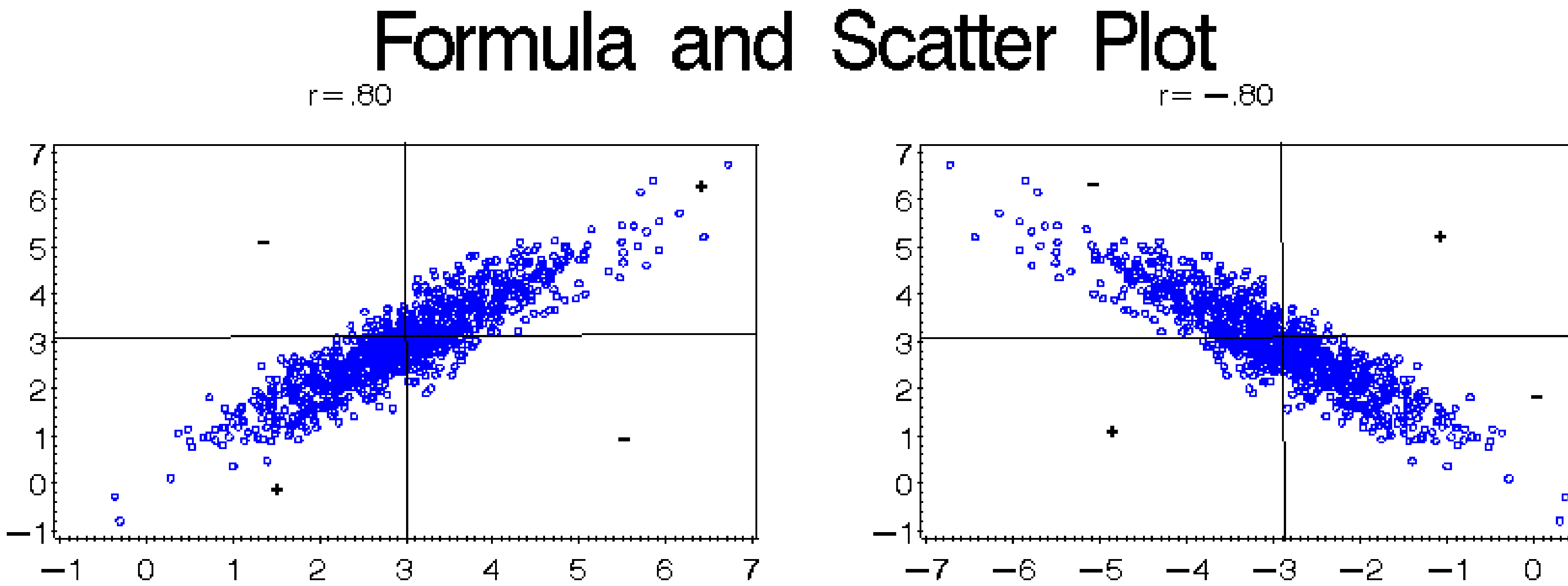
- ρ(Greek "rho") $=$ population correlation.
- $r =$ sample correlation.
- Formal definition

$$
\begin{aligned}
r &= \frac{\text{cov}(X,Y)}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \\
&= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}
\end{aligned}
$$

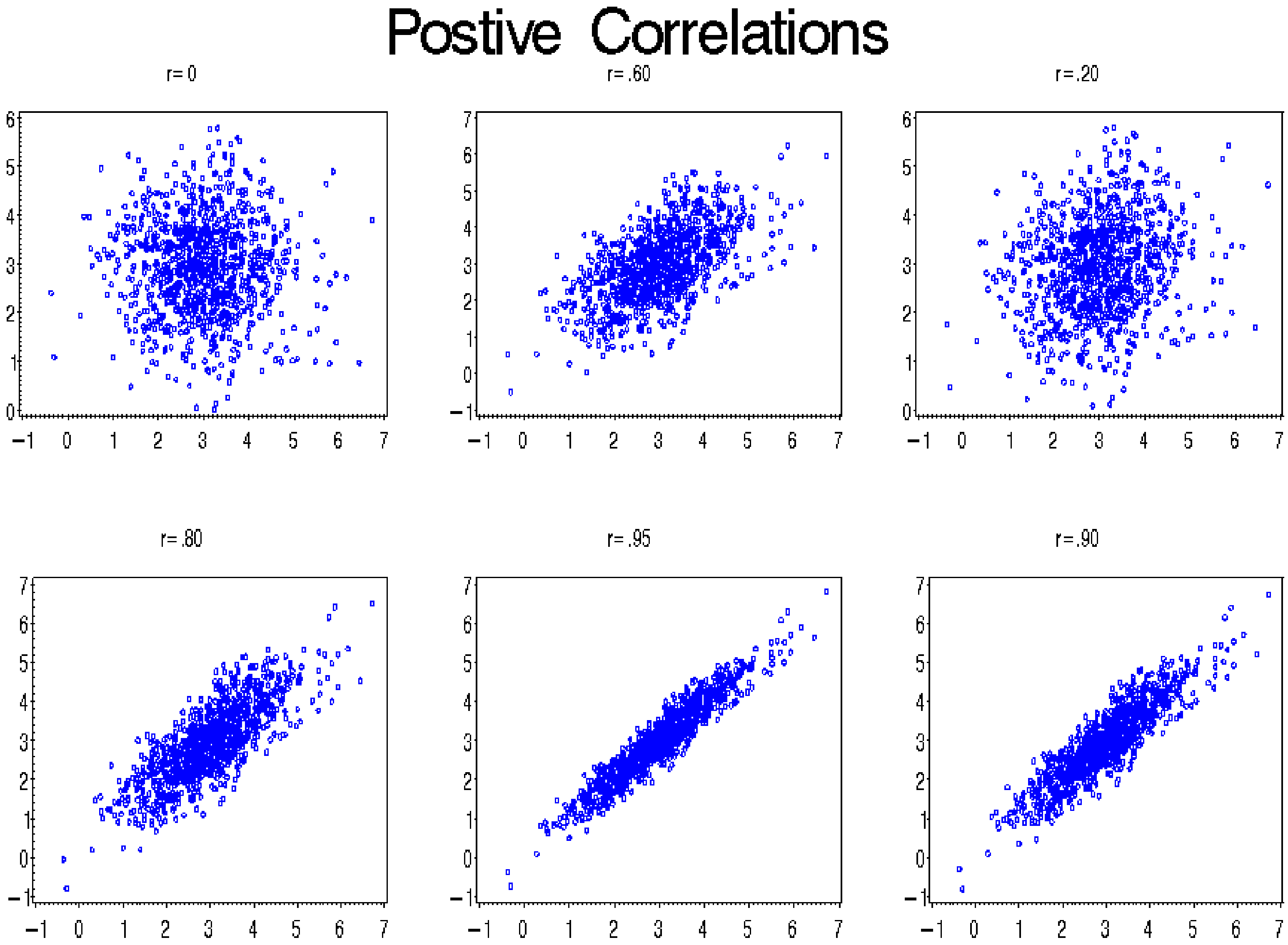
- It measures the extent to which two random variables are linearly related.

How r Works

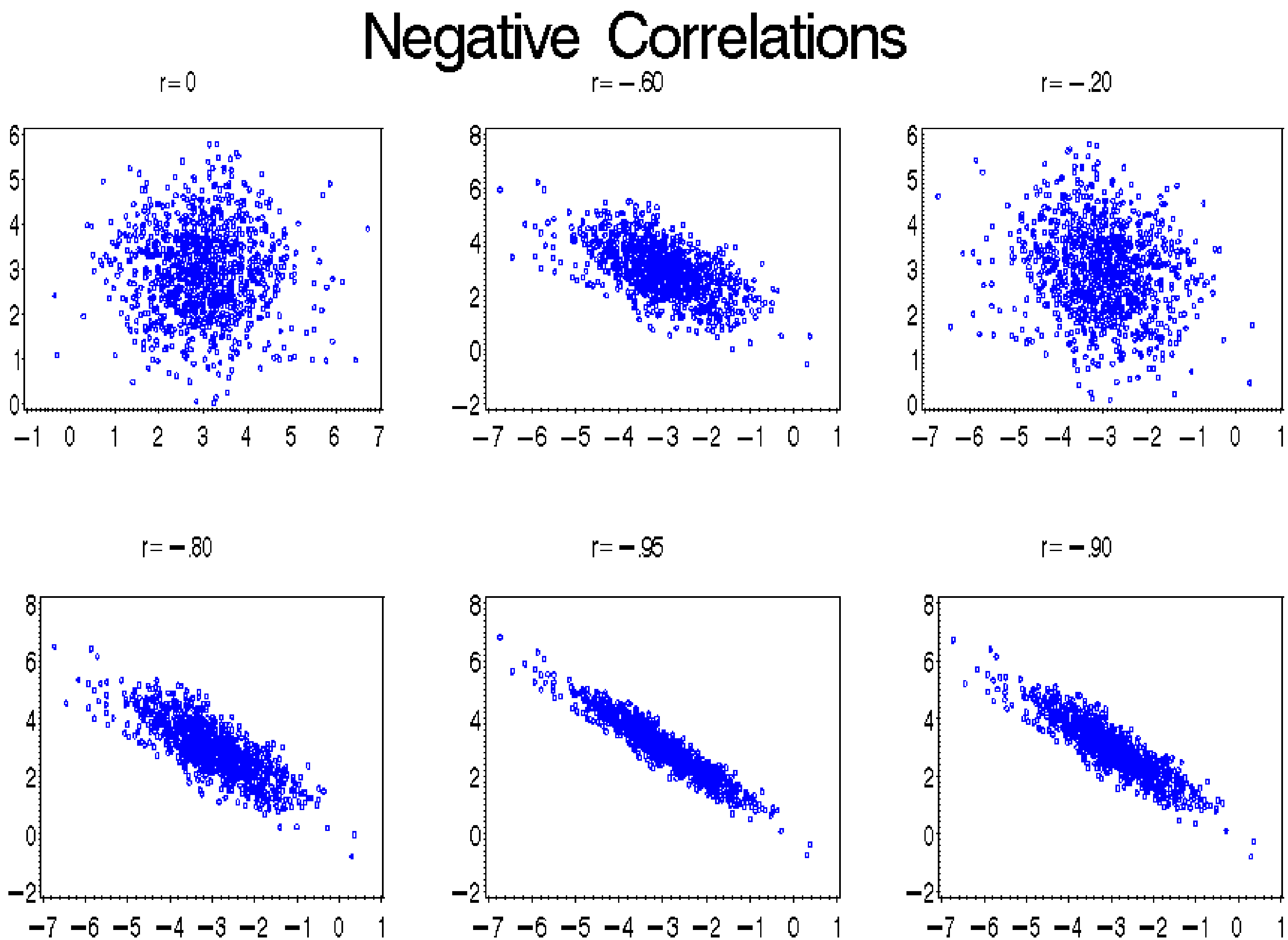
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \sum_{i=1}^{n} z_{x_i} z_{y_i}$$



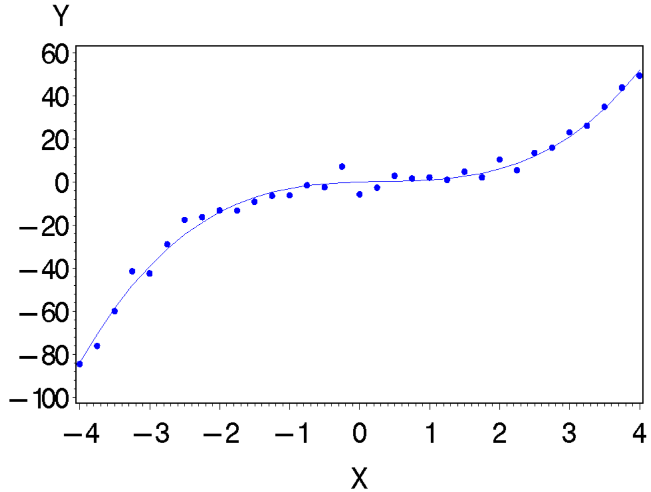
Examples of Different r's



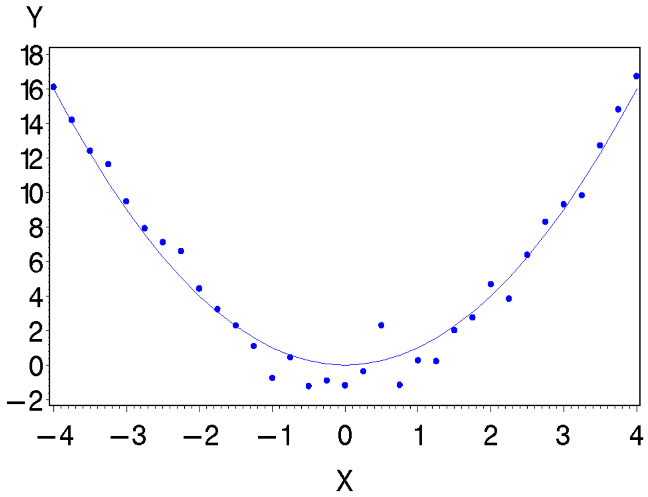
Examples of Different r's

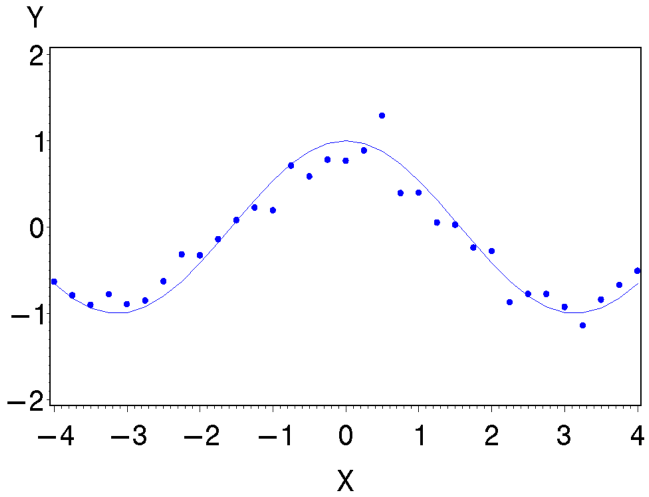


Non-Linear Relationships



$r = 0$





Properties: Correlation Coefficient

- $-1 \leq r \leq +1$
  - $-1 \leq r < 0 \longrightarrow$ small values of X go with large values of Y and large values of X go with small values of Y.
  - $0 < r \leq +1 \longrightarrow$ large values of X go with large values of Y and small values of X go with small values of Y.
  - $r = 0 \longrightarrow$ No linear relationship.
- r measures the strength of the relationship (magnitude) between two variables and the direction of the relationship (sign).

Properties: Correlation Coefficient

- r measures <u>linear</u> relationship.
- Linear transformations of X and/or Y do <u>not</u> change the size (magnitude) of r. Linear transformations do <u>not</u> change the direction (sign) as long as

$$X^* = aX + b$$

  where $a > 0$ (e.g., z scores).
- In a scatter plot, a linear transformation(s) (where $a > 0$) simply corresponds to relabelling axis (axes).

Inference & the Correlation Coefficient

- Preliminaries: bivariate normal distribution.
- This is a generalization of the normal distribution for two random variables
(say X and Y).
- The parameters of the bivariate normal distribution are:

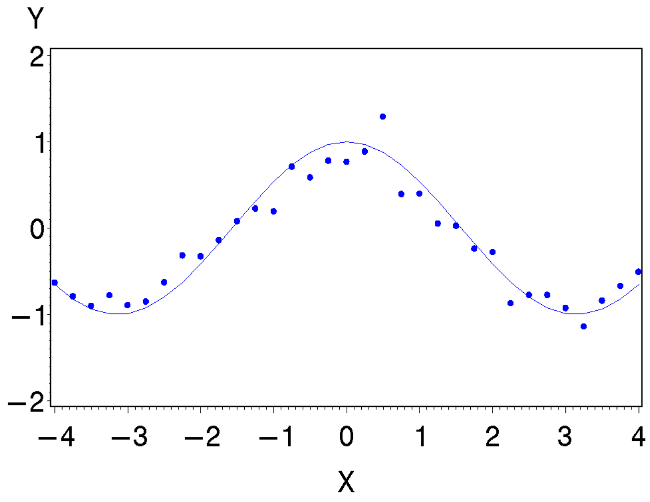
$$\mu_x, \quad \sigma_x^2, \quad \mu_y, \quad \sigma_y^2, \quad \text{and} \quad \rho_{xy}$$

- It looks like a bell or a little hill.
- MatLab program.

The Bivariate Normal Distribution

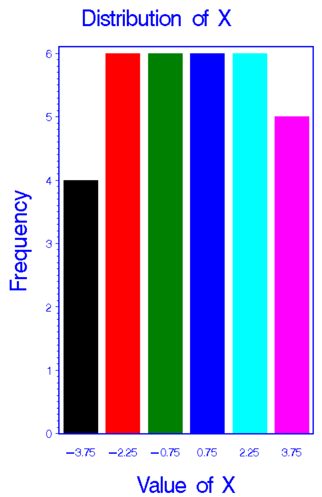
- If X and Y have a bivariate normal distribution, then
  - $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$
  - $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$
  - ρ_{xy} measures how related X and Y are.
- If X and Y are bivariate normal and $\rho_{xy} = 0$, then X and Y are statistically independent.
- If X and Y are statistically independent, then $\rho_{xy} = 0$.
- The case where $\rho_{xy} = 0$ and the (joint) distribution of X and Y is not bivariate normal does not imply that X and Y are statistically independent.

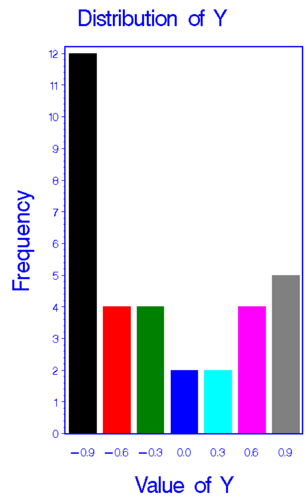
Example: $\rho = 0$ & dependent



$r = 0$

Marginal distributions of X and Y are *not* normal:





Hypothesis Testing

- Statistical Hypotheses: The most common case,

$$H_o : \rho = 0 \quad \text{versus} \qquad H_a : \rho \neq 0$$

- Assumptions:
  - X and Y are random variables whose joint distribution is bivariate normal.*** qualification.
  - Observations are independent.

- Test Statistic: Given the assumptions above and $H_o : \rho = 0$,

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

- Sampling Distribution of the test statistic is Student's t with $\nu = n - 2$.
- Note: the test statistic depends on both r and the sample size n. So for a given α-level, you do not have to compute the test statistic. . . just find the “critical” value for r.

Example Hypothesis Testing for ρ

- High School & Beyond: Reading scores and Motivation
- $H_o : \rho_{read,mot} = 0$ vs $H_a : \rho_{read,mot} \neq 0$.
- Test statistic

$$t = \frac{.21061}{\sqrt{\frac{(1-.21061^2)}{600-2}}} = \frac{.21061}{\sqrt{.9556/598}} = 5.269$$

- For $\nu = 600 = 2 = 598$, p value $= P(|t| \geq 5.269) < .001$; therefore, Reject H_o.
- Conclusion: The data provide evidence that there is a linear relationship between reading and motivation.

Alternative Method

- Find the critical r and compare to the observed r.
- Will reject $H_o : \rho = 0$ vs $H_a : \rho \neq 0$ whenever

$$observed \quad t_{n-2} \leq_{.025} t_{n-2} \quad \text{or} \quad observed \quad t_{n-2} \geq_{.975} t_{n-2}$$

- Take

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

and r as a function of t.

Alternative Method

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

- Square both sides and solve for r:

$$\begin{aligned}
t^2 &= r^2\frac{(n-2)}{1-r^2} \\
t^2\frac{(1-r^2)}{(n-2)} &= r^2 \\
\frac{t^2}{(n-2)} &= r^2(1+t^2/(n-2)) \\
r^2 &= \frac{t^2}{(n-2)(1+t^2/(n-2))}
\end{aligned}$$

Alternative Method

- So
$$r_{crit} = \frac{t_{crit}}{\sqrt{(n-2)(1+t^2_{crit}/(n-2))}}$$

- For our HSB example:
$$r_{crit} = \frac{1.9639}{\sqrt{598\left(1+\frac{(1.9639)^2}{598}\right)}} = \frac{1.9639}{\sqrt{601.85}} = .08$$

- Any correlation $> .08$ (or $< -.08$) would be "significant" for $n = 600$.
- Note: "Statistical significance" does not imply "importance".

Alternative Method

- More correlations from the HSB data where $N = 600$, $\alpha = .05$, and $r_{crit} = .0877$:

	Locus of control	Self concept	Motivation
Reading	.38	.06	.21
(p-value)	($< .01$)	(.14)	($< .01$)
Science	.32	.07	.12
(p-value)	($< .01$)	(.09)	($< .01$)

- Note p-value$=$ Prob$(|r| \geq r$ given $\rho = 0)$, i.e., $H_o : \rho = 0$.

Computing Correlations: SAS

- SAS program command:

```
PROC CORR;
   VAR rdg sci;
   With locus concpt mot;
```

- or

```
PROC CORR;
   VAR rdg sci locus concpt mot;
```

- ASSIST
- ANALYST
- Interactive data analysis

Fisher's Z-Transformation

- Why? When $\rho \neq 0$, the sampling distribution for r is skewed.
- Fisher's Z-Transformation is a function of r whose sampling distribution of the transformed value is close to normal.
- Can compute confidence intervals and a variety of tests using Fisher's Z.
- Requirement: For the distribution of Z to be approximately normal,
  - Variables from a bivariate normal distribution.
  - Sample size should be $n \geq 10$ (and larger if question the bivariate normal assumption).

Fisher's Z-Transformation

Fisher's Z-Transformation:

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

where

- r is the sample correlation
- Z is the transformed value of r
- $\ln$ is the natural logarithm.
- Note: the natural logarithm has base equal to $\exp = e = 2.718281828$; that is,

$$\text{if} \quad \exp^a = x \quad \text{then} \quad \ln(x) = a$$

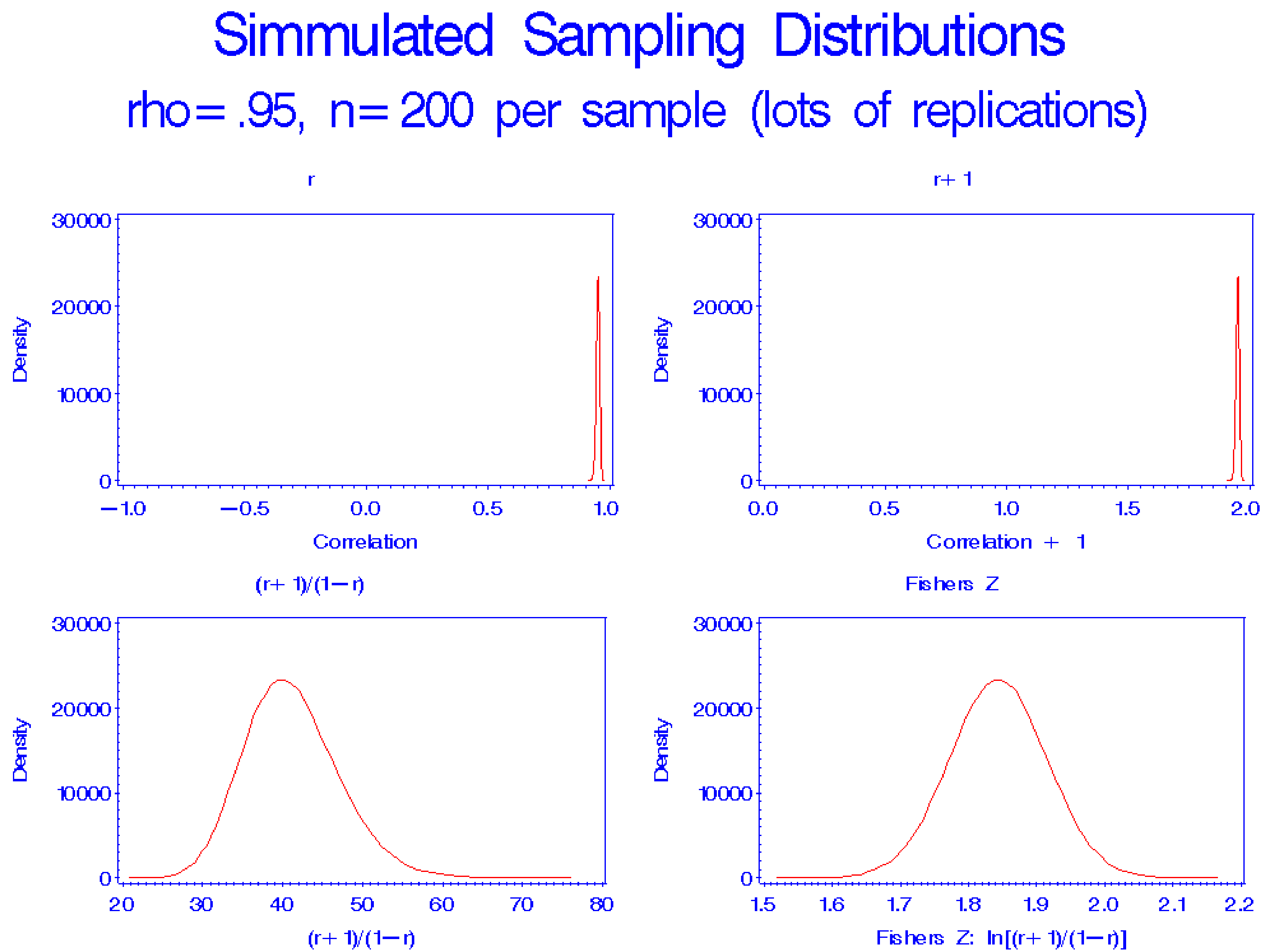
Fisher's Z-Transformation

- Taking the logarithm of numbers has the effect of "compressing" the differences or space between the larger values and "stretching" the space between smaller values.

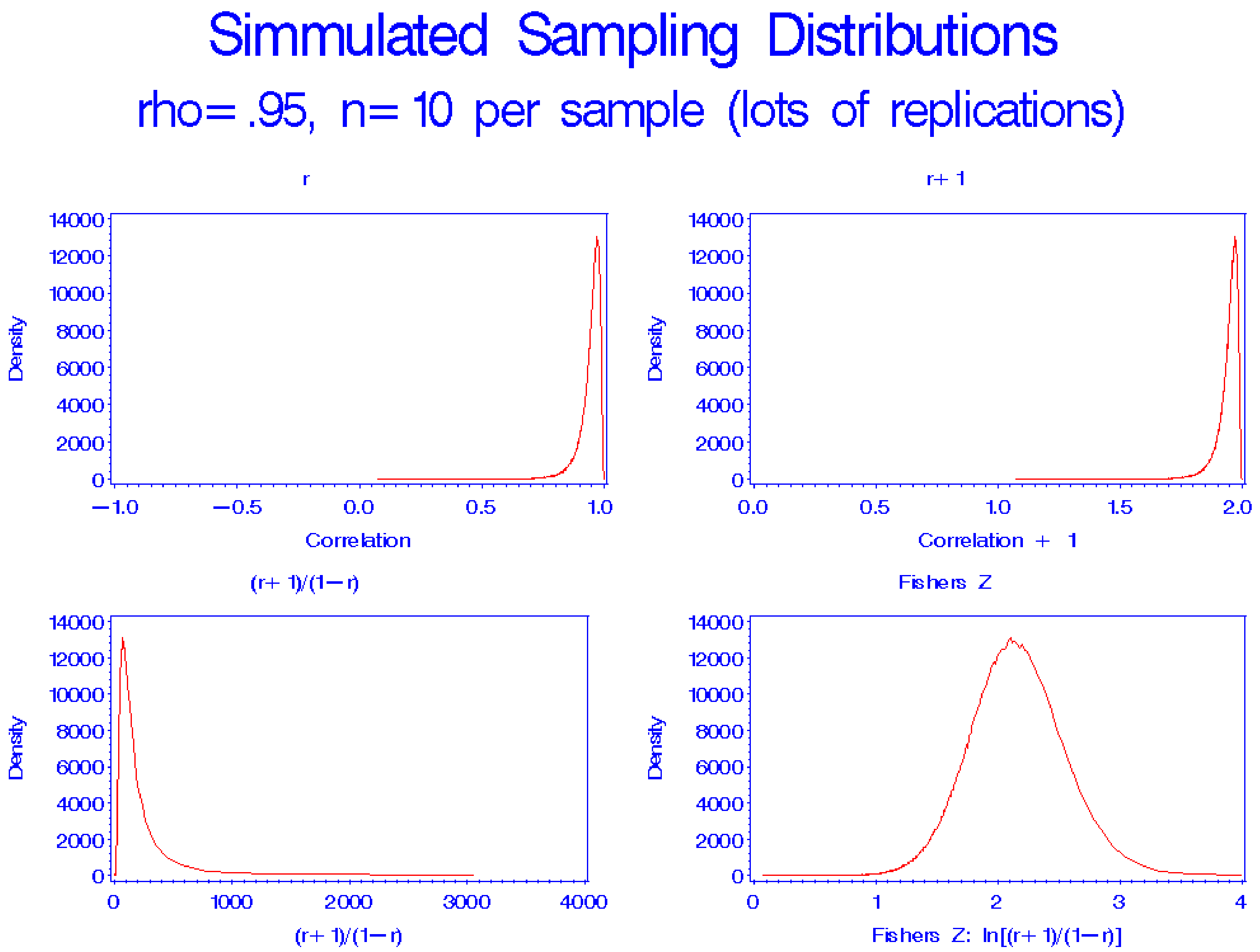
- If a distribution is positively skewed, the taking the logarithm has the effect of making the distribution more symmetric.

- How Fisher's Z-Transformation works. . .

How Fisher's Z-Transformation Works



. . . and for Smaller Samples. . .



Sampling Distribution of Fisher's Z

- IF Observations
  - Are from a bivariate normal distribution.
  - Are independent across individuals.
  - $n \geq 10$
- THEN the sampling distribution of Z is $\approx \mathcal{N}(\mu_Z, \sigma_Z^2)$ where

$$
\begin{aligned}
E(Z) = \mu_Z = Z_\rho &= \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) + \frac{\rho}{2(n-1)} \\
\sigma_Z^2 &= \frac{1}{n-3}
\end{aligned}
$$

- The value $\frac{\rho}{2(n-1)}$ is the bias factor, which in SAS you can request that a bias adjustment be used (in confidence intervals).
- μ_Z and σ_Z^2 are independent of each other.
- The transformation of r is known as the "inverse of the hyperbolic tangent of r".

Using Fisher's Z

- HSB data: Are there relationships between psychological variables and achievement: motivation and reading?
- Observed correlation, $r = .21061$.
- If the true population correlation coefficient is $\rho > 0$, then the sampling distribution of r will be skewed.
- Use Fisher's Z transformation,

$$Z = \frac{1}{2}\ln\left(\frac{1+.21061}{1-.21061}\right) = \frac{1}{2}\ln(1.53360) = \frac{1}{2}(.4276) = .2138$$

- The standard deviation,

$$\sigma_Z = \frac{1}{\sqrt{600-3}} = .04093$$

Example Using Fisher's Z

- Suppose want to test $H_o : \rho = .25$ vs $H_a : \rho \neq .25$.
- Need the value of Z for $\rho = .25$,

$$Z_{.25} = \frac{1}{2} \ln\left(\frac{1 + .25}{1 - .25} = .2554 \right)$$

- Test statistic is

$$z = \frac{Z_{obs} - Z_{null}}{\sigma_Z} = \frac{.2138 - .2554}{.04093} = \frac{-.0416}{.04093} = -1.016$$

- Retain H_o. . . .
- Note: a lower case z is used for the test statistic and upper case Z is denotes Fisher's Z-transformed value of r.

Confidence Interval for ρ

- Another use for Fisher's Z-transformation.
- Suppose we want a 95% CI for correlation between motivation and reading scores.
- Steps:

1. Transform the sample correlation: $Z_{obs} = .2138$.
2. Compute the $(1-\alpha)\%$ CI for Z_ρ

$$\begin{array}{ccl} Z_{obs} & \pm & z_{\alpha/2}\sigma_Z \\ .2138 & \pm & 1.96(.04093) \Longrightarrow (.13, .29) \end{array}$$

3. Un-transform the end points of the CI above.

Confidence Interval for ρ

- Reversing the Fisher Z transformation. . . a little algebra gives

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

- Our example

$$r_{lower} = \frac{e^{2(.1336)} - 1}{e^{2(.1336)} + 1} = \frac{2.71828^{.2672} - 1}{2.71828^{.2672} + 1} = \frac{.3063}{2.3063} = .1328$$

$$r_{upper} = \frac{e^{2(.2940)} - 1}{e^{2(.2940)} + 1} = \frac{2.71828^{.5880} - 1}{2.71828^{.5880} + 1} = \frac{.8004}{2.8004} = .2858$$

- The 95% confidence interval for ρ between motivation and reading scores is $(.13, .29)$.

Fisher's Z in SAS


```
TITLE 'Testing Ho: rho=0 using Fisher-Z transformation';
proc corr data=hsb fisher;
    var mot rdg;
RUN;

TITLE 'Ho: rho= .25 , No bias adjustment';
proc corr data=hsb fisher(rho0=.25 biasadj=no alpha=.05);
var mot rdg;
RUN;

TITLE 'Ho: rho= .25 , With bias adjustment';
proc corr data=hsb fisher(rho0=.25 biasadj=yes alpha=.05);
var mot rdg;
RUN;
```

Two Independent Group Test

- Test whether the correlation from 2 independent groups are the same or different.
- The same procedure that we used for testing difference between mean for large samples.
- Statistical hypotheses:

$$H_o : \rho_1 = \rho_2 \quad \text{vs} \quad H_a : \rho_1 \neq \rho_2$$

- Assumptions:
  - Observations are independent within and between populations
  - The joint distribution of the two variables in each population is bivariate normal.

Two Independent Group Test

- Test Statistic:

$$z = \frac{Z_1 - Z_2}{\sigma_{Z_1 - Z_2}}$$

where

- Z_1 and Z_2 are Fisher Z-transformations of the sample correlations, r_1 and r_2, from the two groups.
- Standard deviation,

$$\sigma_{Z_1 - Z_2} = \sqrt{\sigma^2_{Z_1} + \sigma^2_{Z_2}} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

  Why?
- Sampling distribution of the test statistic is $\mathcal{N}(0, 1)$.

Example Two Independent Group Test

- Is the relationship between writing scores and locus of control the same or different for male and female high school students?
- The data: $n_{male} = 327$ and $r_{male} = .40196$
  $n_{female} = 273$ and $r_{female} = .28250$
- Statistical hypotheses:
  $H_o : \rho_{male} = \rho_{female}$ vs $H_a : \rho_{male} \neq \rho_{female}$
- Assumptions:
  - Scores come from bivariate normal populations.
  - Independence within and between groups.
- . . . so what's dependent?

Example Two Independent Group Test

Test Statistic:

$$
\begin{aligned}
Z_{female} &= \frac{1}{2}\ln\left(\frac{1+.28250}{1-.28250}\right) = \frac{1}{2}(.5808) = .29040 \\
Z_{male} &= \frac{1}{2}\ln\left(\frac{1+.40196}{1-.40196}\right) = \frac{1}{2}(.85197) = .425980 \\
\sigma_{(Z_m - Z_f)} &= \sqrt{\frac{1}{n_{male}-3} + \frac{1}{n_{female}-3}} = \sqrt{\frac{1}{324}+\frac{1}{270}} = .08240 \\
z &= \frac{.42598 - .29040}{.08240} = 1.645
\end{aligned}
$$

Conclusion: Retain H_o for $\alpha = .05$. The difference between the correlations more likely to be due to chance than reflect real a difference.