INTRODUCTION TO BIOSTATISTICS FOR BIOMEDICAL RESEARCH

Frank E Harrell Jr
James C Slaughter
Department of Biostatistics
Vanderbilt University School of Medicine
f.harrell@vanderbilt.edu
james.c.slaughter@vanderbilt.edu

biostat.mc.vanderbilt.edu/ClinStat

Updated February 8, 2010

Chapter 1

Correlation

1.1 Overview

Outcome	Predictor	Normality?	Linearity?	Analysis Method
Interval	Binary	Yes		2-sample t-test **or linear regression**
Ordinal	Binary	No		Wilcoxon 2-sample test
Categorical	Categorical			Pearson χ^2 test
Interval	Interval	Yes	Yes	**Correlation or linear regression**
Ordinal	Ordinal	No	No	**Spearman's rank correlation**

- Examine association between continuous/interval outcome (y) and continous/interval predictor (x)

- Scatterplot of y versus x

1.2 Pearson's correlation coefficient

R11.1, .7-.8, K5.7.A

- $r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$

- Range: $-1 \leq r \leq 1$

- Correlation coefficient is a unitless index of strength of association between two variables (+ = positive association, - = negative, 0 = no association)

- Measures the linear relationship between X and Y

- Can test for significant association by testing whether the population correlation is zero

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

which is identical to the t-test used to test whether the population r is zero; d.f.=$n-2$.

- Use probability calculator for t distribution to get P-value (2-tailed if interested in association in either direction)

- 1-tailed test for a positive correlation between X and Y tests H_0 : when $X \uparrow$ does $Y \uparrow$ in the population?

- Confidence intervals for population r calculated using Fisher's Z transformation

R11.8
A89-91

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

  - For large n, Z follows a Normal distribution with standard error $\frac{1}{\sqrt{n-3}}$

  - To calculate a confidence interval for r, first find the confidence interval for Z then transform back to the r scale

$$\begin{aligned} Z &= \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \\ 2*Z &= \log_e\left(\frac{1+r}{1-r}\right) \end{aligned}$$

$$\begin{aligned}
\exp(2 * Z) &= \left(\frac{1+r}{1-r}\right) \\
\exp(2 * Z) * (1-r) &= 1+r \\
\exp(2 * Z) - r * \exp(2 * Z) &= 1+r \\
\exp(2 * Z) - 1 &= r * \exp(2 * Z) + r \\
\exp(2 * Z) - 1 &= r\left(\exp(2 * Z) + 1\right) \\
\frac{\exp(2 * Z) - 1}{\exp(2 * Z) + 1} &= r
\end{aligned}$$

- Example (Altman 89-90): Pearson's r for a study investigating the association of basal metabolic rate with total energy expenditure was calculated to be 0.7283 in a study of 13 women. Derive a 95% confidence interval for r.

  $$Z = \frac{1}{2}\log_{\mathrm{e}}\left(\frac{1+0.7283}{1-0.7283}\right) = 0.9251$$

  The lower limit of a 95% CI for Z is given by

  $$0.9251 - 1.96 * \frac{1}{13-3} = 0.3053$$

  and the upper limit is

  $$0.9251 + 1.96 * \frac{1}{13-3} = 1.545$$

  A 95% CI for the population correlation coefficient is given by transforming these limits from the Z scale back to the r scale

  $$\frac{\exp(2 * 0.3053) - 1}{\exp(2 * 0.3053) + 1} \quad \text{to} \quad \frac{\exp(2 * 1.545) - 1}{\exp(2 * 1.545) + 1}$$

  Which gives a 95% CI from 0.30 to 0.91 for the population correlation

1.3 Spearman's Rank Correlation

R11.12
K5.7.B

- Pearson's r assumes linear relationship between X and Y

- Spearman's ρ (sometimes labeled r_s) assumes monotonic relationship between X and Y
  - when $X \uparrow$, Y always $\uparrow$ or stays flat, or Y always $\downarrow$ or stays flat
  - does not assume linearity
- $\rho = r$ once replace column of Xs by their ranks and column of Ys by ranks
- To test $H_0 : \rho = 0$ without assuming linearity or normality, being damaged by outliers, or sacrificing much power (even if data are normal), use a t statistic:

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

  which is identical to the t-test used to test whether the population r is zero; d.f.=$n-2$.
- Use probability calculator for t distribution to get P-value (2-tailed if interested in association in either direction)
- 1-tailed test for a positive correlation between X and Y tests H_0 : when $X \uparrow$ does $Y \uparrow$ in the population?

1.4 Correlation Examples

- Correlation difficult to judge by eye
- Example plots on following pages

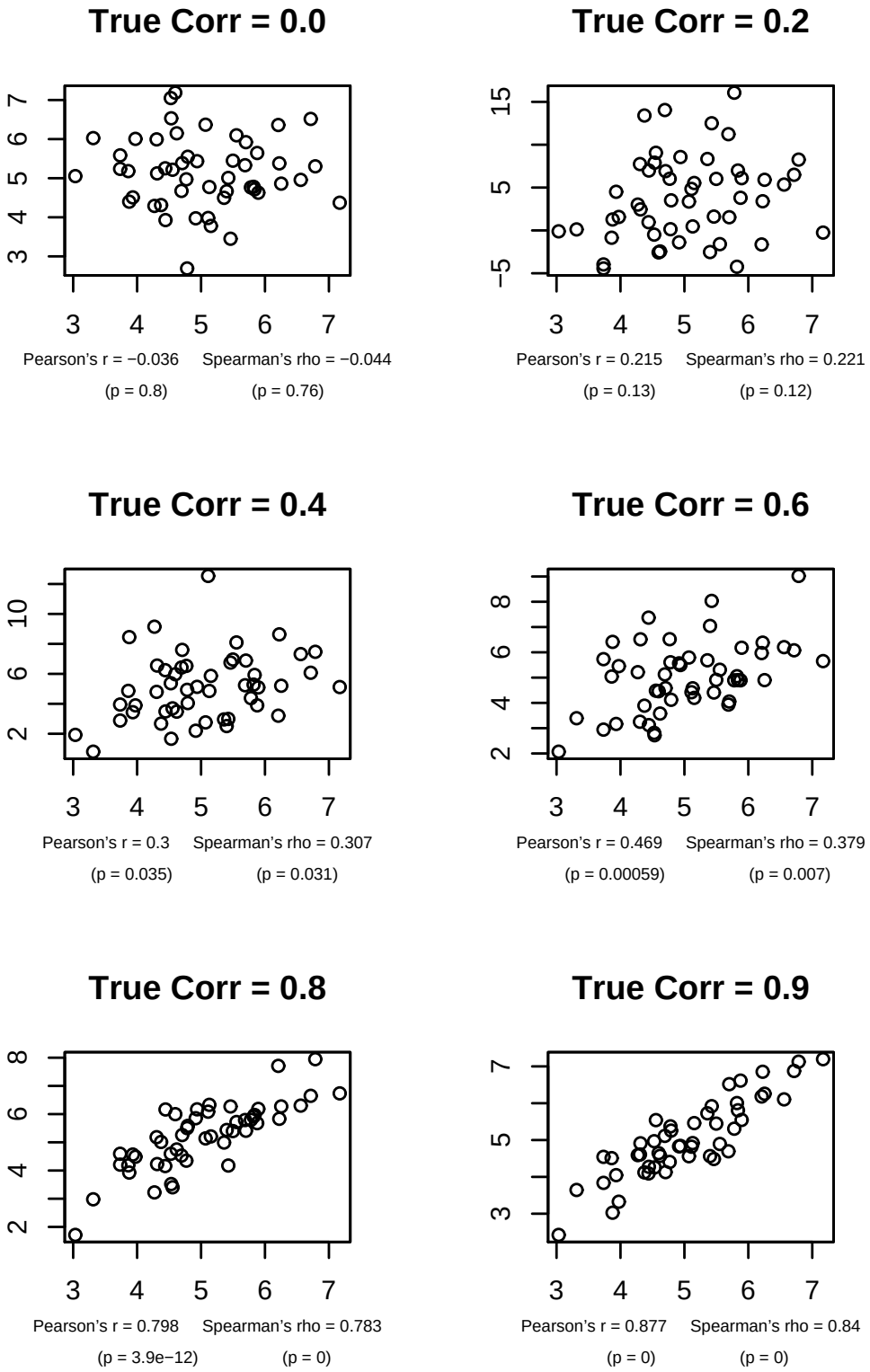

Figure 1.1: *X and Y are drawn from bivariate Normal populations with correlations ranging from 0.0 to 0.9. Pearson and Spearman sample correlations are shown for samples of size 50.*

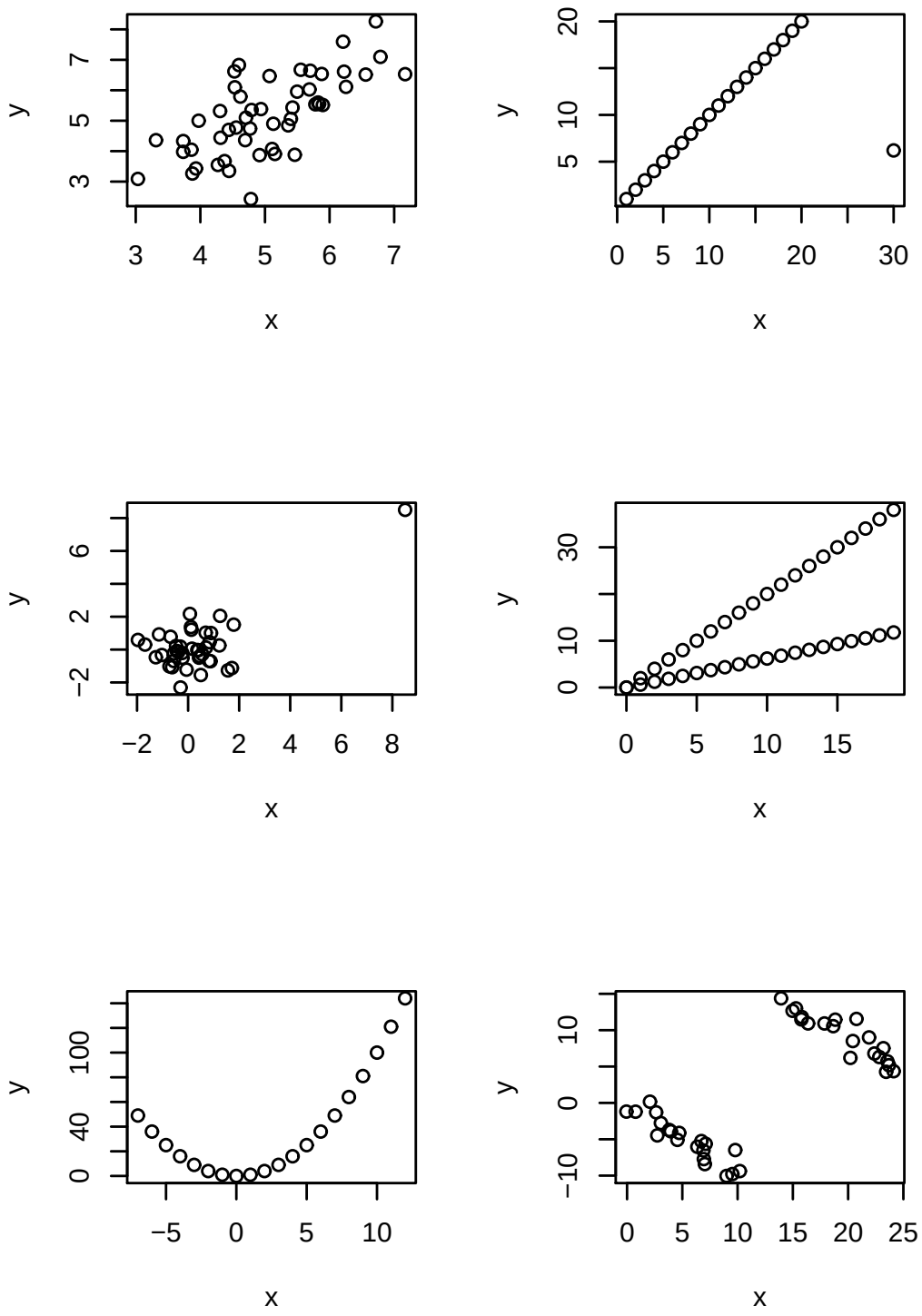

Figure 1.2: *Different observed datasets that have the same correlation. All six plots have a sample Pearson's correlation of* 0.7.

1.5 Correlation and Agreement

- Compare two methods of measuring the same underlying value
  - Lung function measured using a spirometer (expensive, accurate) or peak flow meter (cheap, less accurate)
  - Two devices (Restech and Sandhill) used to mesured acidity (pH) in the esophagus
- Typical (incorrect) approach begins with scatterplot of Restech versus Sandhill with a 1:1 line indicating perfect agreement

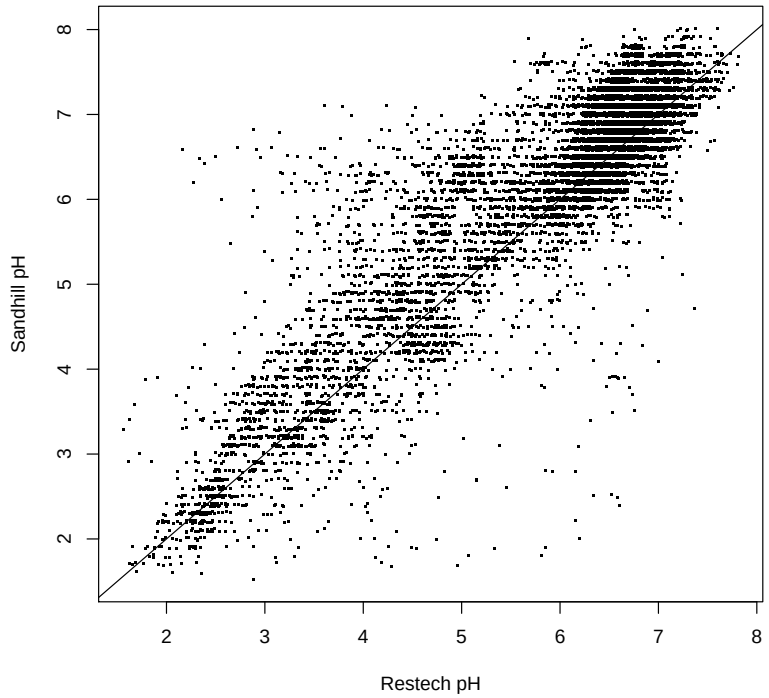

Figure 1.3: *Scatter plot of Restech and Sandhill pH readings. A 1:1 line is included to indicate "perfect" agreement between the two devices.*

- Incorrect approach would report a high correlation ($r = 0.90$) and conclude good agreement

- Problems with the correlation approach

  1. r measures the degree of linear association between two variables, not the agreement. If,for example, the Sandhill consistently gave pH values that were 0.5 unit higher than the Restech, we could still have high correlation, but poor agreement between the two devices. We can have high correlation if the two devices lie closely to any line, not just a 1:1 line that indicates perfect agreement.
  2. A change in scale does not affect correlation, but does influence agreement. For example, if the Sandhill always registered 2 times larger than the Restech, we would have perfect correlation but the agreement would get progressively worse for larger values of pH.
  3. Correlation depends on the range of the data so that larger ranges lead to larger correlations. This can lead to vary strange interpretations

	r	ρ
all data	0.90	0.73
avg pH ≤ 4	0.51	0.58
avg pH > 4	0.74	0.65

Table 1.1: *Pearson (r) and Spearman (ρ) correlations for Restech and Sandhill pH data. The correlation calculated using all of the data is larger than the correlation calculated using a retricted range of the data. However, it would be difficult to claim that the overall agreement is better than both the agreement when pH is less than 4 and when pH is greater than 4.*

  4. Tests of significance (testing if $r = 0$) are irrelevant to the question at hand, but often reported to demonstrate a significant association. The two devices are measuring the same quantity, so it would be shocking if we did not observe a highly significant p-value. A $p < .0001$ is not impressive. A regression analysis with a highly significant slope would be similarly unimpressive.
  5. Data can have high correlation, but poor agreement. There are many examples in the literature, but even in our analysis with $r = 0.90$, the correlation is high, but we will show that the agreement is not as good as the high correlation implies.

See the following handout for simple approaches to assessing agreement and analyzing observer variability studies: `http://biostat.mc.vanderbilt.edu/twiki/pub/Main/ClinStat/obsVar.pdf`

1.5.1 Bland-Altman Plots

EMS36.4

- See Bland and Altman (1986, Lancet)
- Create plots of the difference in measurements on the y-axis versus the average value of the two devices on the x-axis
- If the two devices agree, the difference should be about zero
- The average of the two devices is our best estimate of the true, unknown (pH) value that is we are trying to measure
- Measurements will often vary in a systematic way over the range of measurement. By plotting the difference versus the average, we can visually determine if the difference changes over our estimate of the truth.
- Solid line indicated the mean, dashed lines are approximate 95% confidence intervals (assuming Normality)

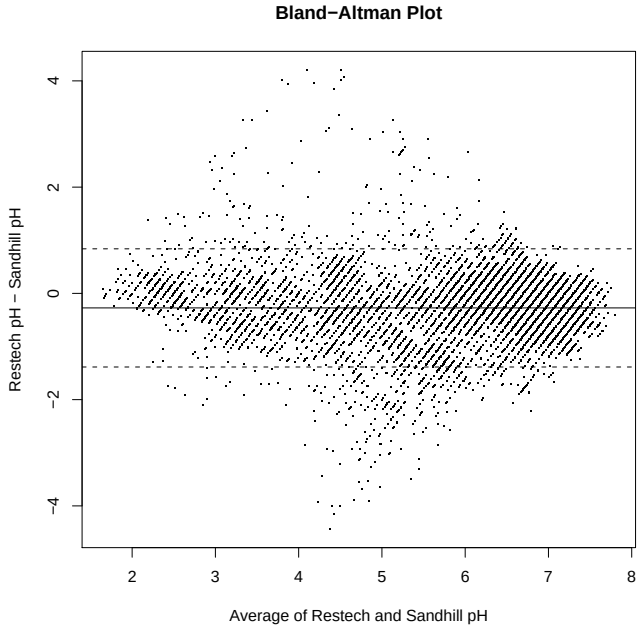

Figure 1.4: *Bland-Altman plot for the Restech and Sandhill pH data. The difference in pH mesaurements (Restech - Sandhill) is presented on the y-axis and the average of the two devices on the x-axis. We see poor agreement around pH values of 4-5*

- In our example, we will also consider differences in the two measurements over the time of day

- The added smooth curve is called a locally weighted scatterplot smooth (lowess)

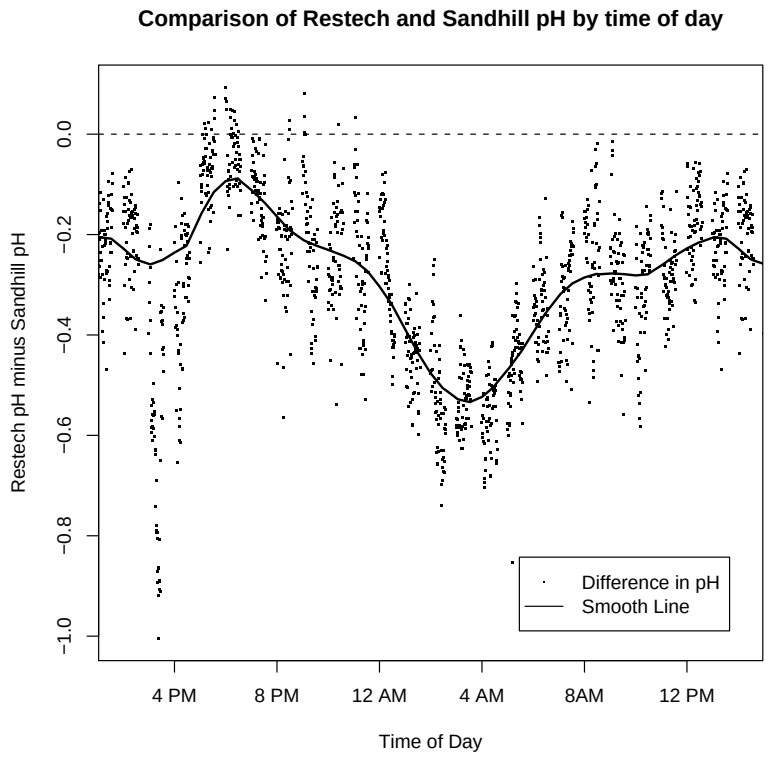

Figure 1.5: *Differene in pH measurements (Restech - Sandhill) by time of day. Is the difference modified by a subject being in a supine position rather than being upright?*

1.5.2 Using r to Compute Sample Size

- Without knowledge of population variances, etc., r can be useful for planning studies

- Choose n so that margin for error (half-width of C.L.) for r is acceptable

- Precision of r in estimating ρ is generally worst when $\rho = 0$

- This margin for error is shown in the figure below

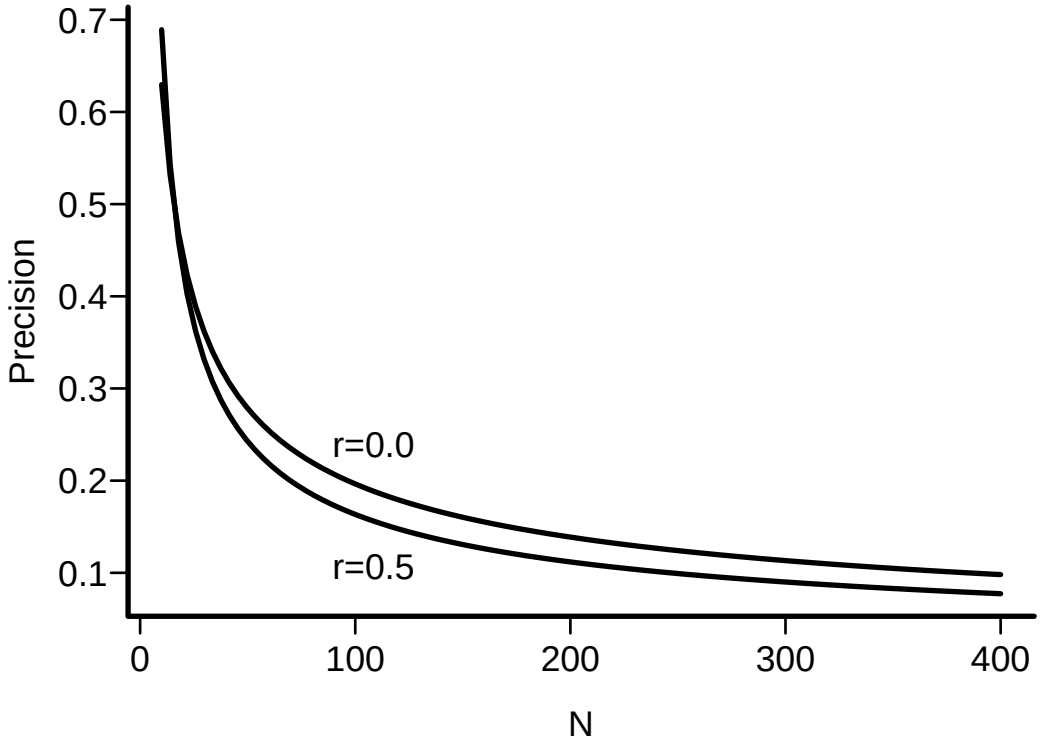

Figure 1.6: *Margin for error (length of longer side of asymmetric 0.95 confidence interval) for r in estimating ρ, when $\rho = 0$ and $\rho = 0.5$. Calculations are based on Fisher's z transformation of r.*

1.5.3 Comparing Two r's

- Rarely appropriate
- Two r's can be the same even though slopes may differ
- Usually better to compare effects on a real scale (slopes)