An Introduction to Statistical Data Analysis

February 12, 2015

Daohai Yu, Ph.D.

Associate Professor of Biostatistics

Department of Clinical Sciences

Temple Clinical Research Institute

Temple University School of Medicine

Outline

1. Introduction
2. Overview of Statistical Tests and Inference Methods
3. Two-sample T-test (Parametric) vs. Wilcoxon Test (Non-parametric)
4. Categorical Data Analysis: 2-Way Contingency Table
5. Correlation Coefficient & Simple Linear Regression
6. Logistic Regression Model (binary endpoint)
7. Multivariable/Multiple Regression Models (continuous endpoint)

Introduction: A Step-by-step Approach to Scientific Discovery

1. Specify the scientific question.
2. Describe the question in the form of a null hypothesis and an alternative hypothesis.
3. Determine the variables (response, predictors, data type).
4. Choose a test procedure or statistical model.
5. Conduct power analysis/sample size calculations.
6. Run the experiment (e.g., a clinical trial).
7. Collect and analyze the data.
8. Interpret and report the results.

Source: Handbook of Biological Statistics.

Statistical Data Analysis Approaches

1. Parametric approach – assumes a particular distribution (e.g., binomial, normal) for the endpoint being measured with only a few unknown parameters. Analysis centers around the estimation/inference of those parameters.
2. Non-parametric approach – “distribution-free”. Most often the data are ranked, e.g., from low to high, and analysis of the ranks is done, often using parametric distribution theory.
3. Bayesian approach – allows for different prior opinions which then lead to different posterior distributions and inferences (less commonly used in practice).

Estimation vs. Hypothesis Testing

- Estimation problems: for example,
  - What will the 5-year survival rate be for this new therapy?
  - How many of these products will we sell next year?

  -- This involves data analysis about a single distribution (one at a time).

- Hypothesis testing for comparing two or more groups: for example, for comparing the complete remission rate (or income) of two or more groups (e.g., between the new and current treatment or male and female, respectively).

  -- This concerns data analysis about multiple distributions.

Real Example in Lung Cancer Study

- Step 1: Specify the scientific question.

  To determine anticancer effect of entinostat in combination with pemetrexed in advanced and previously treated patients with NSCLC.

- Step 2: Describe the question in the form of a null hypothesis and an alternative hypothesis.

  We expect that 6-month PFS (call this rate "p") is at least 30% for the new regimen (currently 6-month PFS at most 12%). H_0: $p \le 0.12$ vs. H_a: $p \ge 0.30$.

- Step 3: Determine the variables.

  Primary endpoint: percent of patients who are alive and progression-free at 6 months after initiation of study agents, binary type of data.

- Step 4: Choose a test procedure or statistical model.

  This is a phase II trial, and the Simon's two-stage design is in general used based on the binomial distribution.

Real Example in Lung Cancer Study-cont'd

- Step 5: Conduct power analysis/sample size calculations.
  1. Type I error rate = 10%, power (=1-type II error) = 90%.
  2. The Simon's two-stage optimal design: 19 patients in 1st stage & 15 in 2nd stage for a total of up to 34 patients (sample size).
  3. Stop the trial if ≤2 successes (alive and progression free at 6 months) in 19 patients in the 1st stage. Otherwise move on to the 2nd stage.
  4. Will not reject H_0: $p \leq 0.12$ (i.e., combination therapy is not effective) if ≤6/34 were alive and progression free at 6 months. Otherwise, if ≥ 7/34 (20.6%) were alive and progression free at 6 months, the null hypothesis H_0 will be rejected (i.e., 6-month PFS is ≥30% for the new regimen).
- Step 6: Run the experiment.
- Step 7: Collect and Analyze data.
- Step 8: Interpret and report the results.

Overview: Commonly Used Statistical Data Analysis Methods

- **Parametric Methods (on two variables, say, X and Y) on independent samples:**

X: Predictor variable	Y: Response (outcome) variable	
	Categorical	Continuous
Categorical	Chi-square (≥2 groups)	ANOVA (≥2 groups)
	Fisher's Exact Test (2x2) McNemar's Test (paired)	T-test (2 groups) Paired T-test (correlated)
Continuous	Logistic Regression/GLM	Linear Regression/GLM
		Pearson Correlation

Overview: Commonly Used Statistical Data Analysis Methods-cont'd

- **Non-parametric Methods (on two variables, say, X and Y) on independent samples:**

X: Predictor variable	Y: Response (outcome) variable	
	Categorical	Continuous
Categorical	Chi-square (≥2 groups)	Kruskal-Wallis (3 groups)
	Fisher's Exact Test (2x2) McNemar's Test (paired)	Wilcoxon Rank-sum (2 groups) Wilcoxon Signed Rank (paired)
Continuous	Logistic Rank Regression/GLM	Non-parametric Rank Regression/GLM
		Spearman Correlation

Paired Testing

- Sometimes individuals are tested before and after some events (e.g., Intervention). Or, each sample may be read by two raters. The independence assumption is thereby violated for individual data points, but not for paired data samples! This feature requires special data analysis methods:
- For a continuous dependent variable, either a paired t-test (parametric) or Wilcoxon signed rank test (non-parametric) on the differences can be employed.
- For a 2x2 table, McNemar's test is appropriate.

Case I

Outcome: Continuous

Predictor: Categorical (2 groups)

Unpaired Samples

T-test and Non-parametric Alternatives

- The T-test was developed by William Sealy Gosset (Student) at the Guinness Brewery in Dublin in 1908.
- Outcome: continuous; Predictor: categorical (2 groups).
- ***Three key assumptions of T-test:***

1. The raw data are **normally distributed** (it is actually enough that the mean is normally distributed).
2. **The variances of the two groups are equal.**
3. The data points are statistically independent (no correlated data!).

≥3 groups: ANOVA. Need similar assumptions to be valid!!!

T-test and Non-parametric Alternatives-cont'd

	Equal Variances	Unequal Variances
Normal	Two-sample T-test	Unequal variance T-test Satterthwaite (Welch)
Not Normal	Wilcoxon rank-sum Normal scores, or T-test (for large n)	Wilcoxon rank-sum Normal scores, or Welch's T test (for large n)

Problem: How do you know if the data are normally distributed and the variances are equal?

T-test and Non-parametric Alternatives: A small sample example

- $n_1, n_2 = 12$
- X_1 = 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5
- X_2 = 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6

- **Parametric Results:**
  1. Preliminary F-test for homogeneity of variance, p = 1.00
  2. T-test, p = .0410
  3. Satterthwaite T-test, p = .0410
- **Non-parametric Test Results: Wilcoxon,**
  1. Kruskal-Wallis, p = .0735
  2. Normal approximation, p = .0783
  3. T-approximation, p = .0915
  4. Exact Wilcoxon, p = .0780
  5. Normal scores, p = .0830
  6. Exact normal scores, p = .0830

T-test and Non-parametric Alternatives: A small sample example-cont'd

Q: What if a new value, 15, is added to Group 2?

T-test and Non-parametric Alternatives: A small sample example-cont'd

- $n_1 = 12$, $n_2 = 13$
- Preliminary F-test for homogeneity of variance, p = .0014
- **Parametric Results:**
  1. T-test, p = .0741
  2. Satterthwaite T-test, p = .0722
- **Non-parametric Test Results: Wilcoxon,**
  1. Kruskal-Wallis, p = .0442
  2. Normal approximation, p = .0471
  3. T-approximation, p = .0586
  4. Exact Wilcoxon, p = .0457
  5. Normal scores, p = .0477
  6. Exact normal scores, p = .0458

T-test p-value depending upon the new value

	New Value						
	5	6	7	8	9	10	11
T-test	.025	.022	.022	.025	.030	.035	.042
Satterthwaite	.025	.022	.022	.024	.028	.034	.040
Eq. variances?	.963	.814	.538	.304	.159	.071	.032
Normal?	.441	.442	.646	.258	.054	.012	.003
	12	13	14	15	30	100	1000
T-test	.050	.058	.066	.074	.173	.288	.341
Satterthwaite	.048	.056	.064	.072	.167	.279	.331
Eq. variances?	.014	.006	.003	.001	<.0001	<.0001	<.0001
Normal?	.001	.0004	.0002	.0001	<.0001	<.0001	<.0001

A non-parametric test should be used when, say New Value ≥ 9.

Conclusion from power analysis

- ***Power*** – the probability of rejecting the null hypothesis for a given alternative hypothesis. A test with higher power is more likely to reject the null when the alternative hypothesis is true.

- "The power analysis results suggest that on the basis of power, at least for large samples, both the Wilcoxon and normal scores test are preferable to the T-test for general use".

  JL Hodges and EL Lehmann, 1961

  Fourth Berkeley Symposium

- ***Myth: The T-test should be used unless the test assumptions are violated.***

Alternative approach: Transform first to use a parametric method

- The transformations (e.g., log, square root,…) are one way to **improve the normality of the data**.
- That is, instead of using a non-parametric test, one can *sometimes* use a parametric test on the transformed data (if the test assumptions are met following the transformation).
- Example: Gene expression data analysis, cell growth model analysis…

Case II

Outcome: Categorical (binary)

Predictor: Categorical (binary)

Unpaired Samples

Categorical Data Analysis: 2 X 2 Table

	No Cure	Cure	Total
Treatment A	1	5	6
Treatment B	5	1	6
Total	6	6	12

- Chi-Square Test Value = 5.33 p-value = 0.0209
- Fisher Exact Test (2-sided) p-value = 0.0801
- Fisher mid p-value (when n's are small) p-value = 0.0411

Categorical Data Analysis: 2 X 2 Table-cont'd

	No Cure	Cure	Total
Treatment A	1 3	5 3	6
Treatment B	5 3	1 3	6
Total	6	6	12

- Chi-square = $\Sigma(O - E)^2/E = 4 \times 2^2/3 = 5.33$.
- The assumptional problem has to do with the denominator of the ratios. When small, the statistic fails to yield the proper p-values.
- Rule of thumb: Expected values under 5 are problematic.

Real Example: Data from a clinical trial

	No Infection	Infection	Total
Treatment A	40 42	5 3	45
Treatment B	31 29	0 2	31
Total	71	5	76

- Antifungal prophylaxis to prevent breakthrough aspergillus infections in BMT patients; Initial dose of glucocorticoids = 2 mg/kg.
- Test Results
  1. Chi-Square Test Value = 3.69 p-value = 0.0548
  2. Fisher Exact Test (2-sided) p-value = 0.0753
  3. Fisher mid p-value (when n's are small) p-value = 0.0422

WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Case III

Outcome: continuous

Predictor: continuous

-- Correlation Coefficient

--Simple Linear Regression

Assessing correlation for two continuous variables

- Pearson's correlation coefficient (*r*)

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y}$$

Subject ID (Female)	X=Height (inches)	Y=Shoe Size	Individual Height - Mean Height (Xi - X)	Individual Size - Mean Size (Yi - Y)
1	63	6.5	-1.5	-2
2	63	8	-1.5	-0.5
3	57	6	-7.5	-2.5
4	63	8.5	-1.5	0
5	66	8.5	1.5	0
6	69	9	4.5	0.5
7	69	11	4.5	2.5
8	67	8	2.5	-0.5
9	68	10	3.5	1.5
10	62	11	-2.5	2.5
11	64	7.5	-0.5	-1
12	61	6.5	-3.5	-2
13	58	7	-6.5	-1.5
14	62	11	-2.5	2.5
15	64	6.5	-0.5	-2
16	65	7.5	0.5	-1
17	63	7.5	-1.5	-1
18	66	8.5	1.5	0
19	68	9	3.5	0.5
20	72	12	7.5	3.5
Mean	64.5	8.5		

Correlation Coefficient: A measure of **linear** relationship between two continuous variables

- The correlation coefficient *r* is between -1 and +1.
- Strength of linear relationship --| *r* |: the closer to 1 [or *r* to +/-1], the stronger the linear relationship.
- Direction of linear relationship:
  - $r > 0$ – variables move in the same direction,
  - $r < 0$ – variables move in opposite directions.
- $r = 0$ (or close to 0) indicates no (very weak) linear relationship.

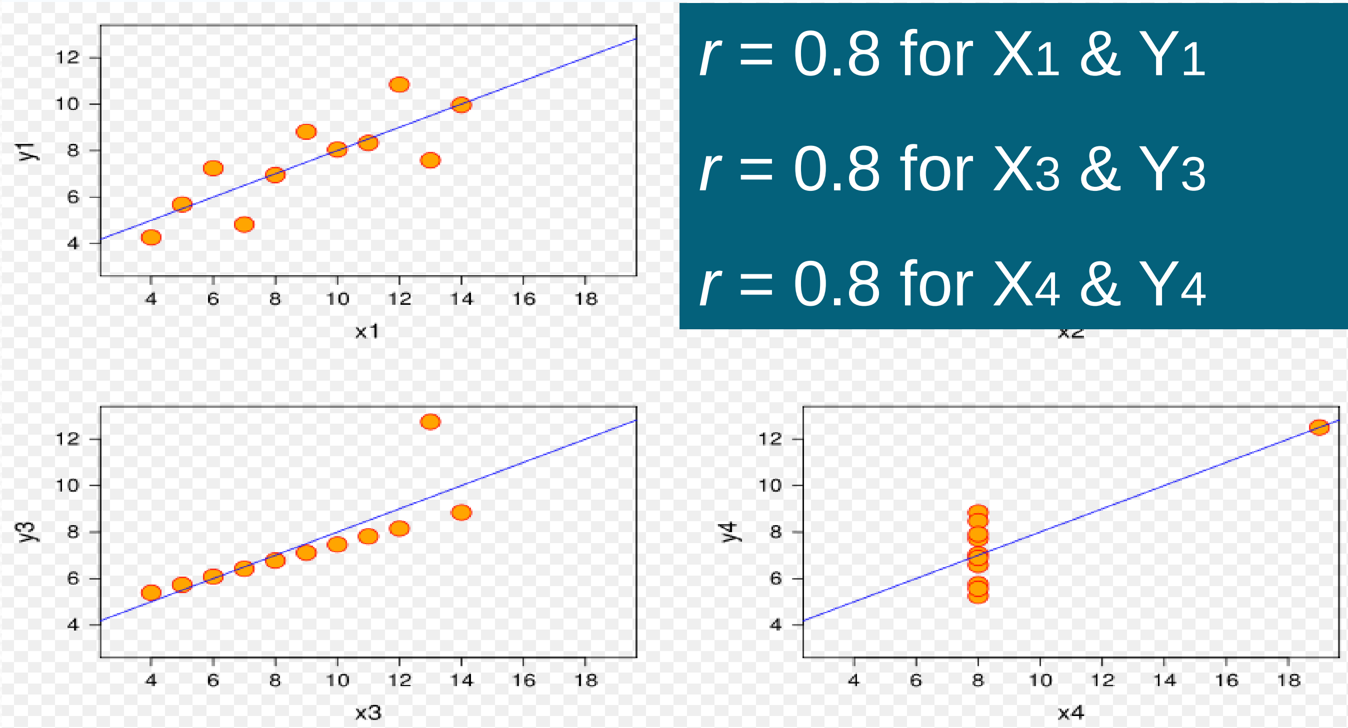
Pearson's Correlation Coefficient

- Computed as: $$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y}$$
- Sensitive to outliers/extreme values:

r = 0.8 for X_1 & Y_1

r = 0.8 for X_3 & Y_3

r = 0.8 for X_4 & Y_4



Source: wikipedia.org

Spearman's Correlation Coefficient

- Spearman's correlation coefficient (ρ) is mathematically equivalent to the Pearson's correlation coefficient after converting the data to ranks.

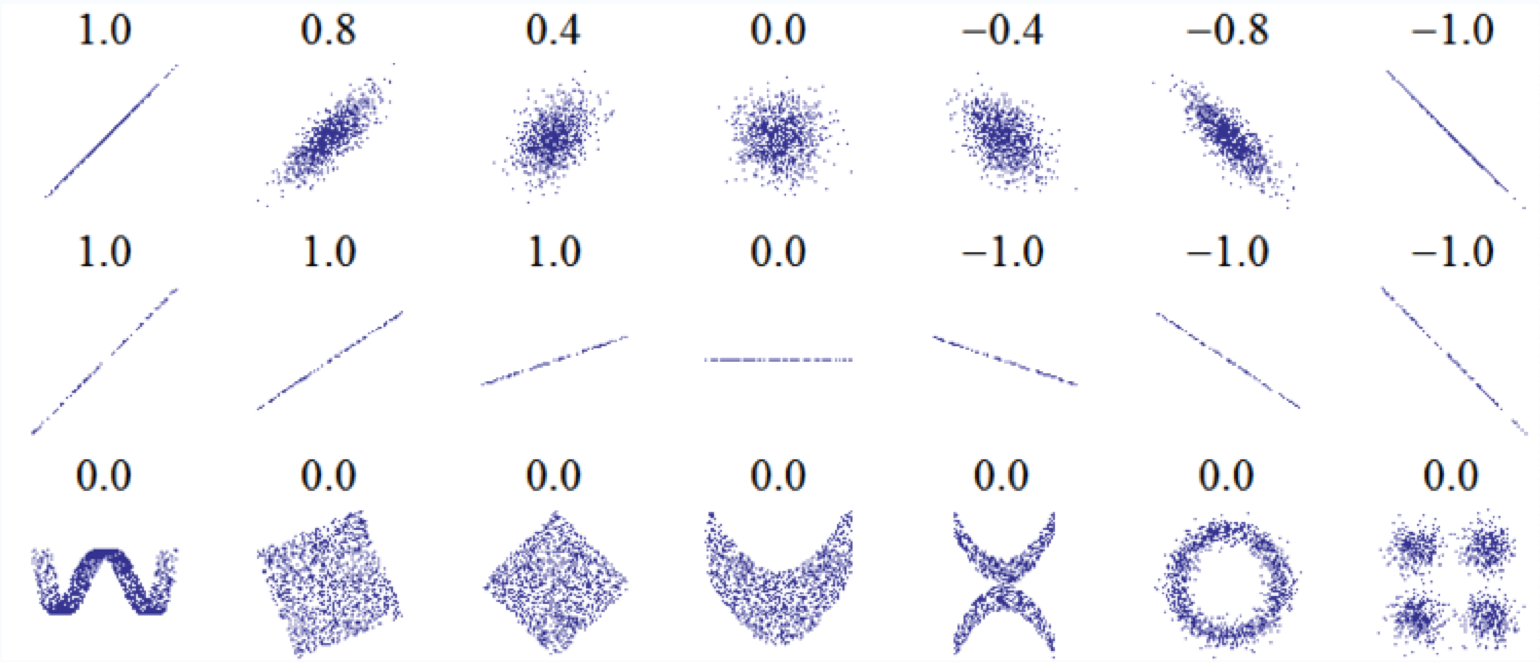
- Spearman's correlation coefficient is non-parametric -- the counterpart of Pearson's correlation without distribution assumptions.

Limitations of Correlation Coefficient

- Measures a linear relationship only
- Lacks predictive ability

Limitations of Correlation Coefficient-cont'd

1. Two variables having the same correlation coefficient can have different linear relationships.
2. When the correlation coefficient is zero, two variables may still have some (non-linear) relationships.



Source: wikipedia.org

Overview: Statistical Modeling

Benefits of Regression Modeling:

1. Evaluates and quantify association between a dependent endpoint and predictors; the association does not have to be linear.
2. Predicts unknown/future outcome; provides both a point estimate/prediction and an interval estimate/prediction (precision or accuracy).

Elements of A Statistical Model

- A distributional assumption for dependent variable Y, e.g., binomial (for response rate), normal (for tumor size).
- A formulated quantitative relationship model between Y and X (called predictors or covariates).
  - For example, $Y = \alpha + \beta^* X$, where parameters (α,β) are of interest and estimated.
- Estimation method:
  - The least squares estimates (LSE)—the fit with the smallest sum of the squares of the residuals. E.g., in ANOVA and linear regression.
  - The maximum likelihood estimate (MLE)—the fit maximizing the likelihood function of the data. E.g., in logistic model and GLM.
- Statistical software often used, e.g., SAS, R, STATA, and SPSS.

Case III

Outcome: Continuous

Predictor: Continuous (or Categorical)

-- Correlation Coefficient

-- Simple Linear Regression

(or ANOVA)

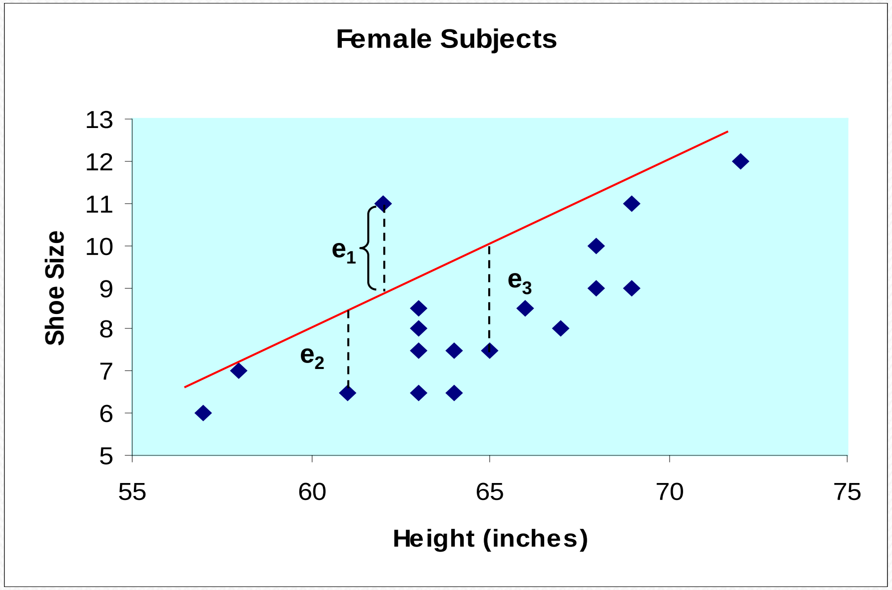
Simple Linear Regression

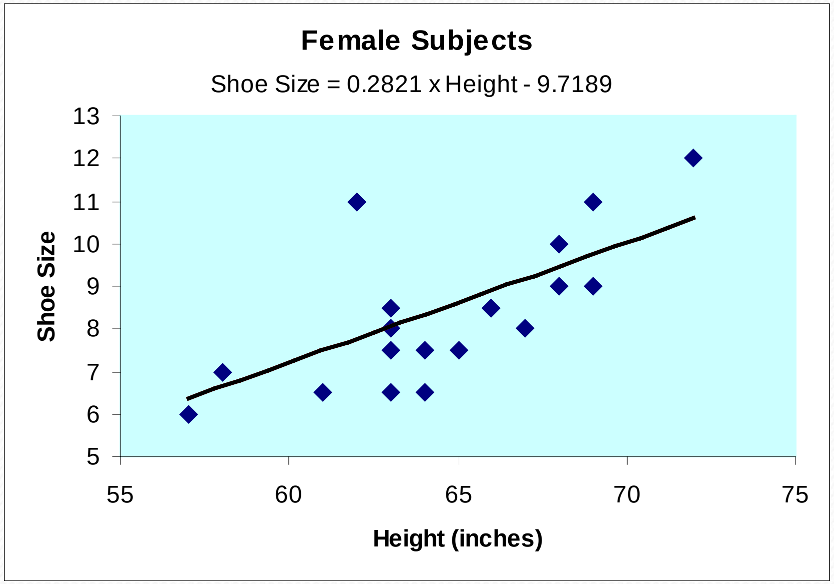
- Y: continuous, X: continuous,
  - e.g., Y: shoe size, X: height.
- Y: ***normally distributed*** ; Y at X's: ***independent***; and Y at X's : the ***same variance*** (homogeneity).
- Relationship between Y and X is formulated as follows:

  (Shoe Size) = α + β* (Height) + measurement error (random)
- Goal: estimate (α,β) and predict unknown shoe size for any given height.
- ***Remarks:*** when X is discrete, it's called (one-way) ***ANOVA***. ***Similar model assumptions needed for correct results!!!***

Simple Linear Regression: Model Fitting

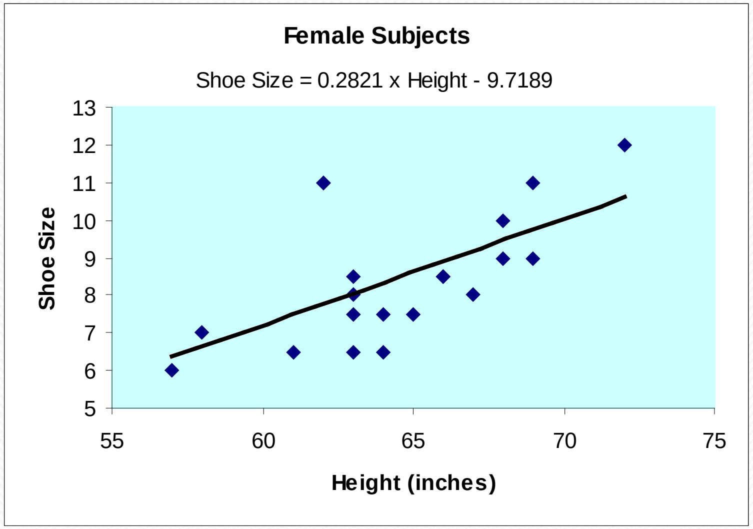
- The optimal line (minimizing the sum of the squared errors-LSE) is Shoe Size = – 9.72 + 0.28 * Height, which can be used for prediction.
- Estimated (α,β) = (– 9.72, 0.28).





Simple Linear Regression: Interpretation

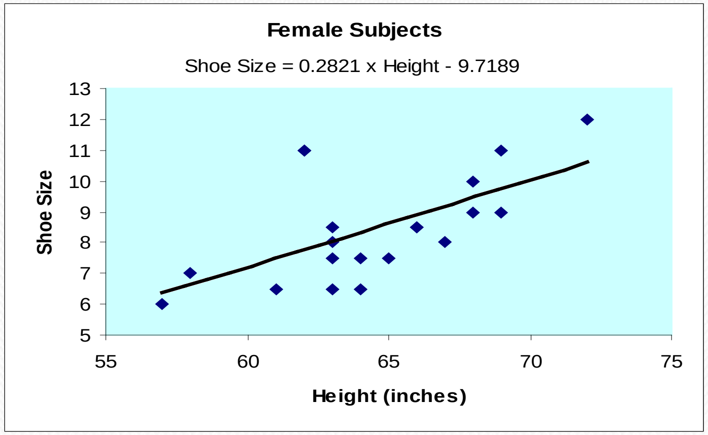
- Shoe Size = – 9.72 + 0.28 * Height



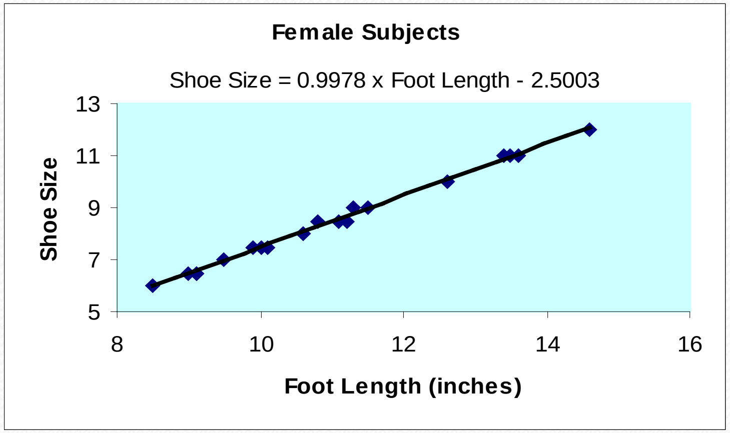
- Interpretation and prediction are available only for Height 57-72.
- When the Height is between 57 and 72, the increment of 1 inch in height results in the increment of 0.28 in the shoe size on average.
- Prediction. For example: if a female friend's height is 67 inches, the predicted shoe size for her is – 9.72 + 0.28*67 = about size 9.

SLR: R^2--Predictive Power

- Coefficient of determination (R^2): the proportion of variability in the data that is accounted for by the statistical model (function).
- R^2 is the squared correlation coefficient between the observed outcome values and the outcome values predicted based on the statistical model (function).



$R^2 = 0.365$



$R^2 = 0.996$

Simple Linear Regression: Testing the Slope (β)

- $\mathbf{Y} = \alpha + \beta * \mathbf{X}$

- Null and alternative hypotheses:

  H_0: $\beta = 0$ → No linear association between X and Y

  H_a: $\beta \neq 0$ → Linear association between X and Y

- For example, (Shoe Size) = – 9.72 + 0.28 * Height

  H_0: Shoe size and Height have no association

  H_a: Shoe size and Height have a linear association

Case IV

Outcome: Categorical (binary)
Predictor: Continuous or Categorical

--Simple Logistic Regression

Simple Logistic Regression

- Y: binary; X: Continuous or Categorical.

  e.g., Y: whether or not having a disease (e.g., ovarian cancer);

  Y=1 if Yes, Y=0 if No.

  X: a measurement of biomarker CA125.

- Assume Y= 1 with probability ***Pr***; observed Y's at different X's independently.

- Relationship between Y and X:

  1. ***Pr*** is modeled as: log(***Pr***/(1-***Pr***)) = $\alpha + \beta * X$. (logit model)
  2. Example: log(ODDS of having ovarian cancer) = $\alpha + \beta *$ CA125.

Simple Logistic Regression: Interpretation

- When CA125 is continuous:

  Estimated β = 0.2 → *Increasing CA125 by one unit will on average increase the patient's odds of having ovarian cancer by* ***exp(0.2X1) = 1.22-fold****.*

- When CA125 is treated as categorical, such as low (CA125 = 0) and high (CA125 = 1) :

  Estimated β = 0.2 → ***Odds ratio (OR)*** *of having ovarian cancer between patients having high CA125 and low CA125 is exp(0.2) =* ***1.22****.*

Multivariable/Multiple Regression Models

- More than one predictor/covariate associated with outcome.

- Example 1: Multiple linear regression--
  (both female and male subjects in the population/sample)
  Shoe Size = $\alpha + \beta_1 *$ Height $+ \beta_2 *$ Gender

- Example 2: Multiple logistic regression--
  $\log(\text{odds}) = \alpha + \beta_1 *$ CA125 $+ \beta_2 *$ biomarker LPA2

Model Diagnosis & Variable Selection

- **Variable selection/model development**:
  - Forward selection;
  - **Backward elimination**;
  - Stepwise selection.
  - Bayesian variable selection.
- **Checking model assumptions**:
  - Check the normality assumption of Y;
  - Check the constant variance assumption of Y.
- Outliers and high leverage points.
- Model validation/goodness of fit.

Model Validation

- Validation of a fitted regression model is the confirmation that model is sound and effective for the purpose for which it was intended.
- Assessing the effectiveness of the fitted equation against an independent set of data.
- One criterion: Mean squared error of prediction.

Some Useful Entry-level Books

- Biostatistics: The Bare Essentials, 3rd Edition by GR Norman and DL Streiner.

- Fundamentals of Biostatistics, 6th Edition by Bernard Rosner.

Biostatistical Support at TUSM

Department of Clinical Sciences

Temple Clinical Research Institute

Kresge, 2nd Floor

Chair: Dr. Susan Fisher

Dr. Daohai Yu: DYu@Temple.Edu

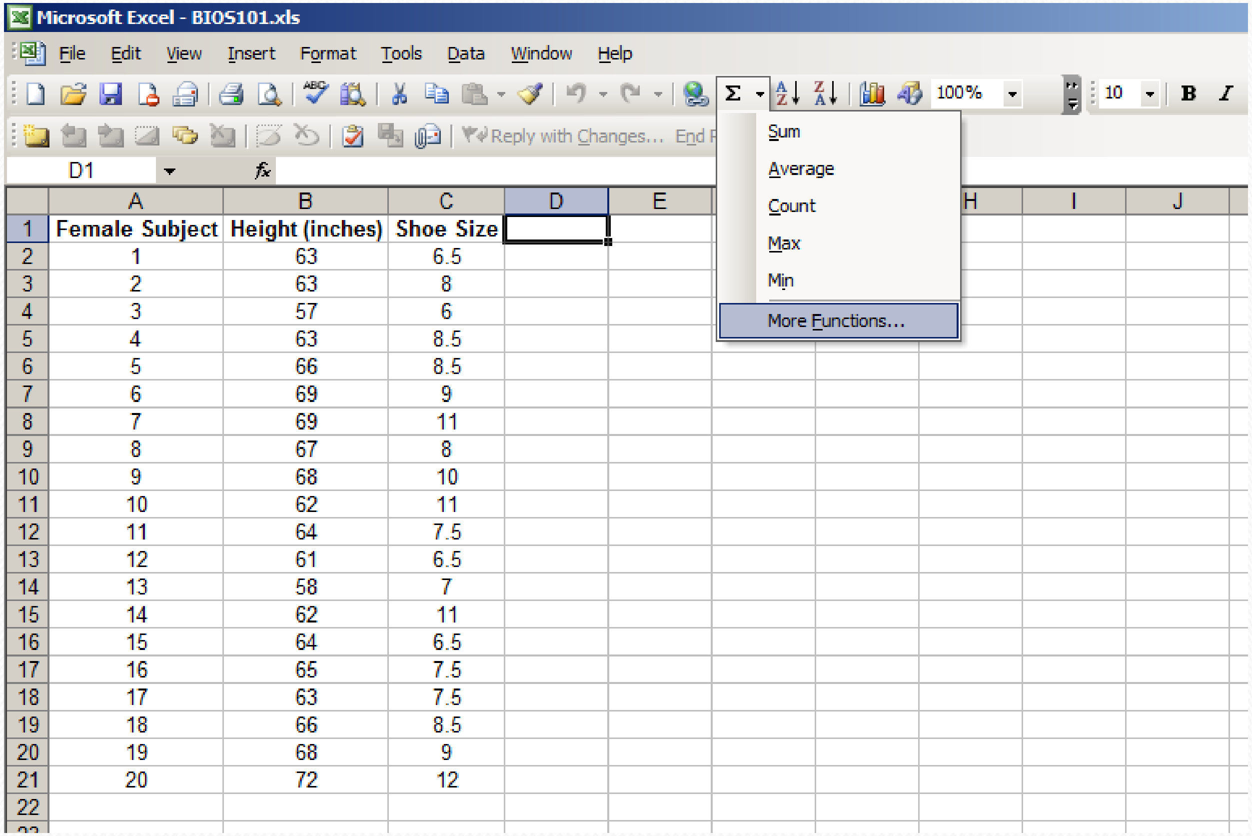
Dr. Huaqing Zhao: Zhao@Temple.Edu

We welcome collaborations/consultations!

Thank you!

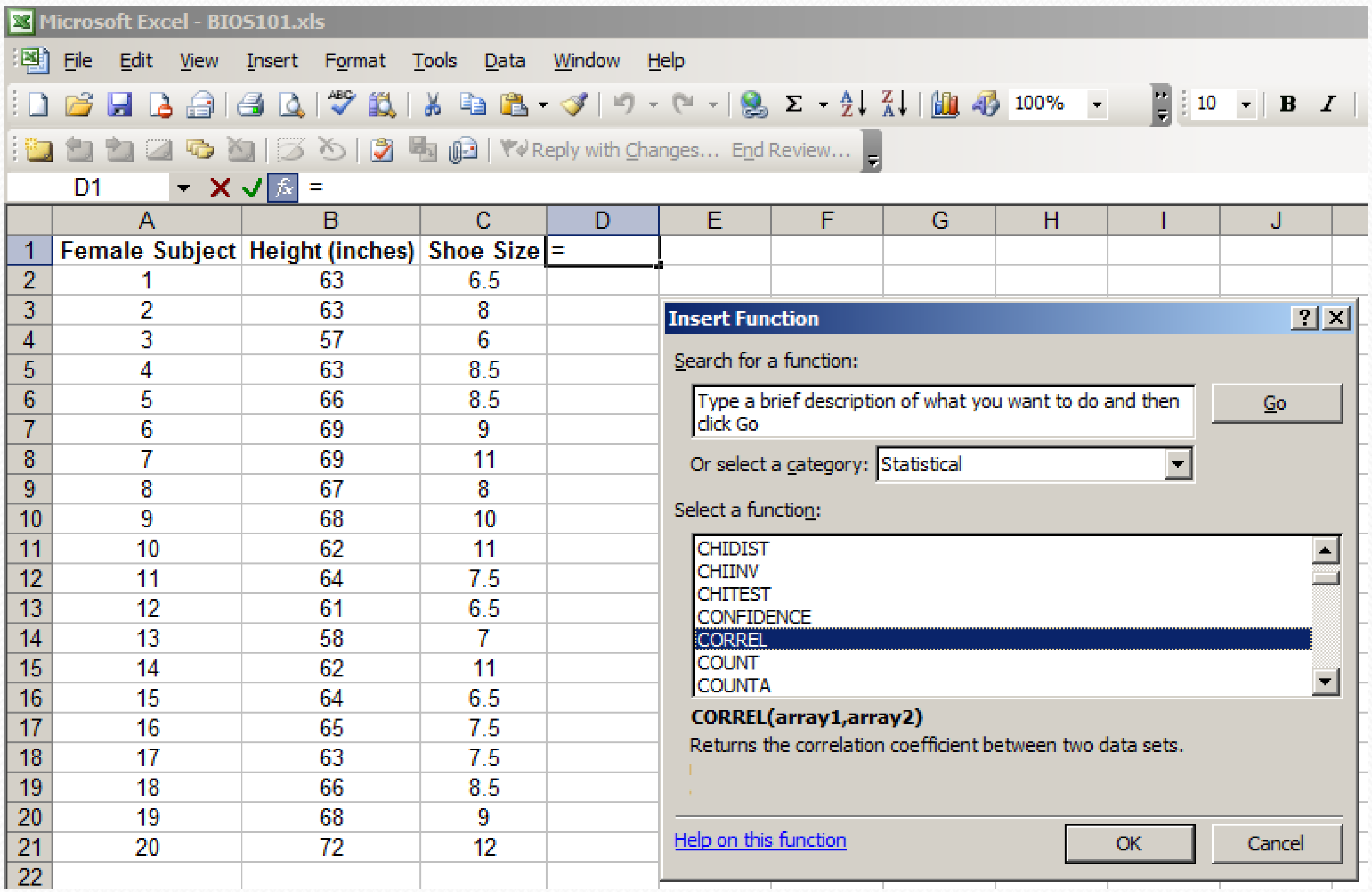
Questions?

Appendix: Using MS Excel to Calculate Pearson Correlation *r*

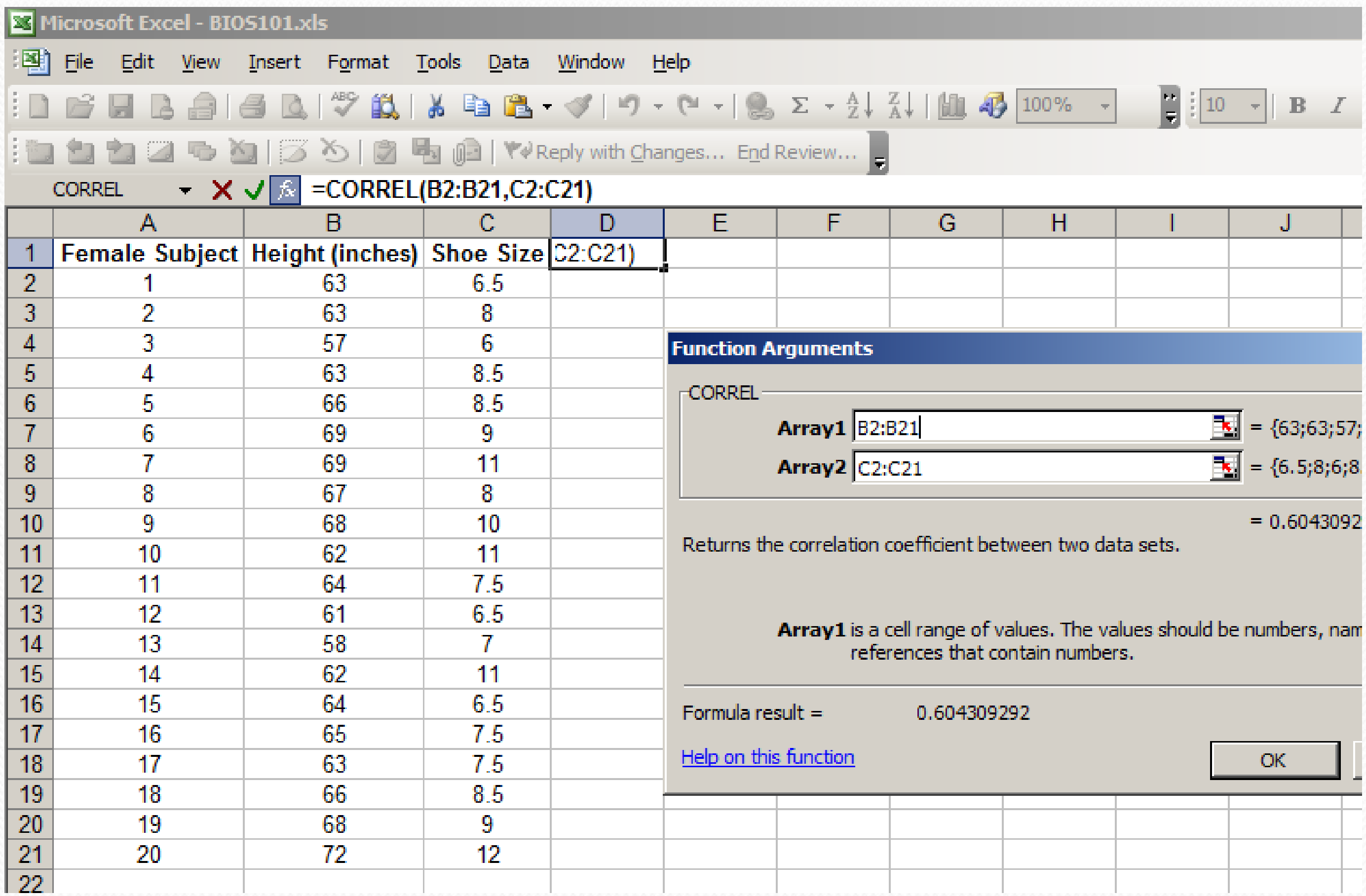


	A	B	C
1	Female Subject	Height (inches)	Shoe Size
2	1	63	6.5
3	2	63	8
4	3	57	6
5	4	63	8.5
6	5	66	8.5
7	6	69	9
8	7	69	11
9	8	67	8
10	9	68	10
11	10	62	11
12	11	64	7.5
13	12	61	6.5
14	13	58	7
15	14	62	11
16	15	64	6.5
17	16	65	7.5
18	17	63	7.5
19	18	66	8.5
20	19	68	9
21	20	72	12

Appendix: Using MS Excel to Calculate Pearson Correlation *r*-2



Appendix: Using MS Excel to Calculate Pearson Correlation *r*-3



	A	B	C
1	Female Subject	Height (inches)	Shoe Size
2	1	63	6.5
3	2	63	8
4	3	57	6
5	4	63	8.5
6	5	66	8.5
7	6	69	9
8	7	69	11
9	8	67	8
10	9	68	10
11	10	62	11
12	11	64	7.5
13	12	61	6.5
14	13	58	7
15	14	62	11
16	15	64	6.5
17	16	65	7.5
18	17	63	7.5
19	18	66	8.5
20	19	68	9
21	20	72	12
22			

Appendix: Using MS Excel to Calculate Pearson Correlation *r*-4

Microsoft Excel - BIOS101.xls

D1 =CORREL(B2:B21,C2:C21)

	A	B	C	D
1	Female Subject	Height (inches)	Shoe Size	0.604
2	1	63	6.5	
3	2	63	8	
4	3	57	6	
5	4	63	8.5	
6	5	66	8.5	
7	6	69	9	
8	7	69	11	
9	8	67	8	
10	9	68	10	
11	10	62	11	
12	11	64	7.5	
13	12	61	6.5	
14	13	58	7	
15	14	62	11	
16	15	64	6.5	
17	16	65	7.5	
18	17	63	7.5	
19	18	66	8.5	
20	19	68	9	
21	20	72	12	