

## Ch. 3: Correlation & Linear Regression

414

Psychology 402 - Fall 2022 - Dr. Michael Diehr

Copyright © 2022 Michael Diehr  
All Rights Reserved  
For use only by students enrolled  
in my sections of Psyc 402  
through the end of the semester.  
May not be posted, shared or uploaded  
online without permission.

415

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Ch. 3: Correlation & Linear Regression

- Relationships between 2 variables
- Scatterplots
- Linear Regression
- Exercise 2
- Correlation
- Race / DNA

423

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Number of variables

- One variable, one dimension
- Number Line
- Frequency Distribution / Histogram
  - 2 dimensional graph of 1D data
- Difference Score
  - 1 dimension
  - 2 dimensions

424

Psychology 402 - Fall 2022 - Dr. Michael Diehr

425

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Bivariate relationships

- “is factor A related to factor B”?
- Methods of analysis...
  - Anecdotal / Clinical
  - Numerical : simple 2x2 analysis
  - Visually -- scatterplots
    - see relationships and problems w/data
    - can’t test hypothesis
  - Statistically -- correlation & regression
    - hard to detect problems w/data
    - easy to test hypothesis

426

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Anecdotal / Clinical

- Many interesting findings began from non-scientific approaches
- “Intuition” that something is related through experiencing multiple situations
- Pattern recognition - Good and Bad
- Problems -- faulty memory, confirmation biases, prejudice, etc...
- Next step after a “gut” feeling : design experiment and collect data.

427

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Simple numerical analysis

- Simplify:
  - use categorical variables
  - or convert continuous variables to categorical
- Use extreme cases to maximize effect
- Compute percentages in a 2x2 matrix
- Do the results suggest an effect?

- Compute Chi-square statistic to judge significance

428

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Dichotomous Variables

- The simplest form of categorical
- Aka “binary”
- Examples:
  - 1/0
  - yes/no
  - pass/fail
  - true/false
  - healthy/sick
  - normal/impaired
  - etc.

429

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Example

- “I think there is brain dysfunction in HIV disease” as measured by neuropsychological testing
- Medical status: control vs. HIV+ symptomatic
- NP test results: normal vs. impaired

		Medical Status	
		Control	HIV+
NP Status	Normal	85%	52%
	Impaired	15%	48%

430

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## 2x2 Analysis

- Pro: easy to understand
- Con: using binary categories reduces *statistical power*
- Conclusion: other Graphical and Statistical methods should be used as well.

432

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Scatterplots

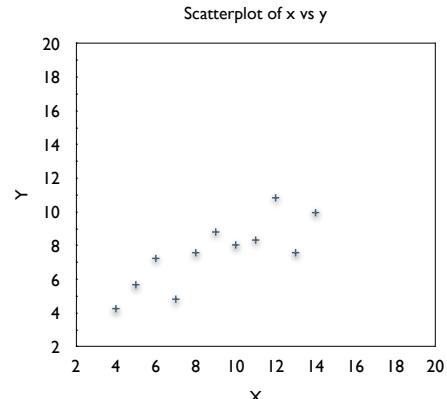
- Graph two variables in relation to each other on two-dimensional X, Y axis
- Easy to see
  - relations
  - problems
- Can’t prove relationship is “significant”
- Difficult to interpret clinically or in “common sense” terms

433

Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Scatterplots

x	y
10	8.04
8	7.58
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68



434

Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Linear Regression

- Assume X and Y are related
- Assume relationship is linear
- Model with single straight line
- Pick the line that best “fits” our data
- Other names: fitting a line, finding the trend, creating a trendline, best fit line...
- Residuals = difference between prediction and actual value
- Linear Regression minimizes the square of the residuals, often called “Ordinary Least Squares”

435

Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Why “Regression”

- Frances Galton
- Height of children vs parents.
- Tall parents have tall children (and vice versa)
- But children are closer to the mean than their parents (by a factor of ~2/3)
- Galton called this “Regression to the Mean”
- His paper fit\*\* straight lines to data points.
- The technique has been called “regression” ever since
- \*\* He never calculated the lines, he just eyeballed them

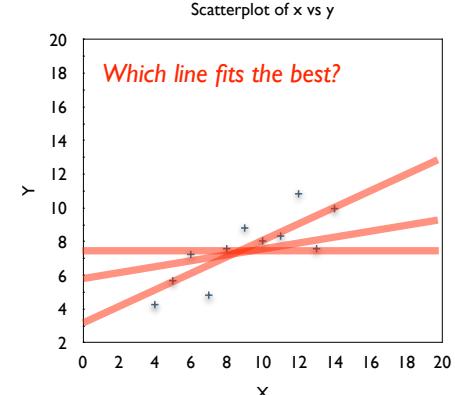
436

Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Linear Regression

**Equation:**  
 $y = 3.0 + 0.5x$

**Correlation**  
 $r_{xy} = 0.816$

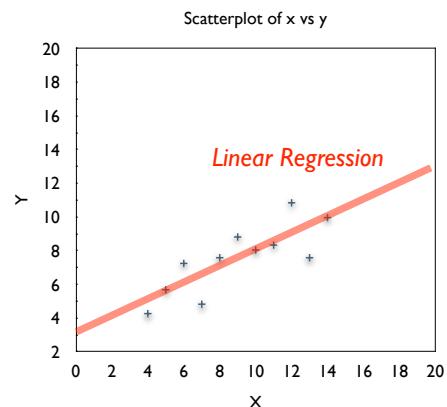


438

Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Anscombe's Quartet I

x	y
10	8.04
8	7.58
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

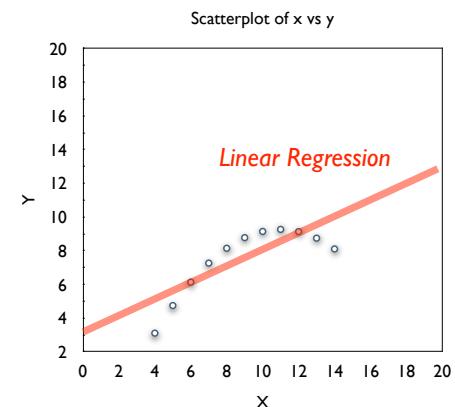


439

Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Anscombe's Quartet II

x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74

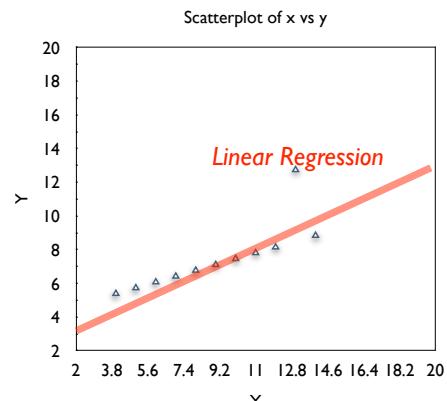


440

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Anscombe's Quartet III

x	y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

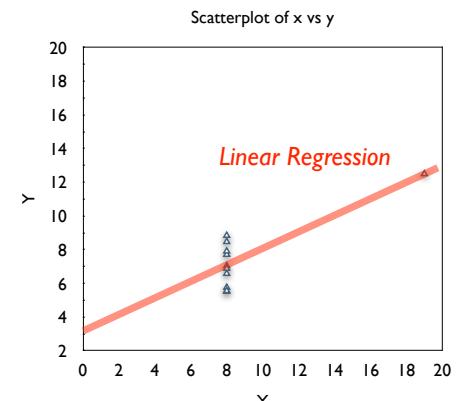


441

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Anscombe's Quartet IV

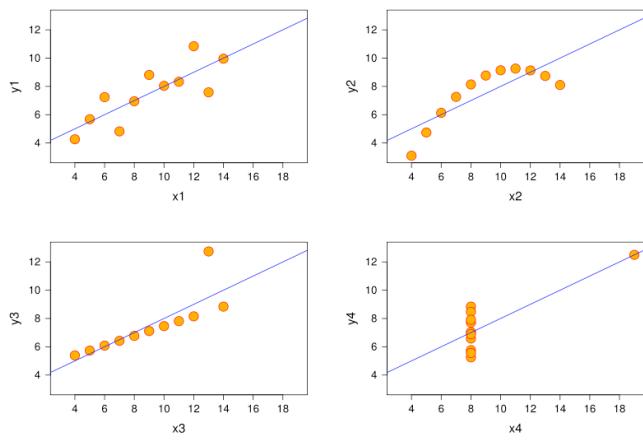
x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.52
19	12.5
8	5.56
8	7.91
8	6.89



442

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Anscombe's Quartet



443

## Anscombe's Quartet Summary

- Each series has the same Quantitative stats:
  - linear regression equations
  - correlations
- Each one is Qualitatively different
- Each series needs special handling
- Lesson? Graph Your Data!

444

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Linear Regression Equation

$$Y' = a + bX$$

$Y'$  = predicted  $Y$

$X$  = actual  $X$

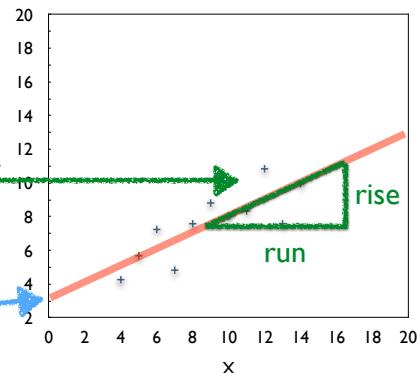
$b$  = slope

$dY/dX$   
(rise over run)

$a$  = intercept

$Y$  value when  $X = 0$

Scatterplot of x vs y



445

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Residuals in Linear Regression

- $X$  : independent variable
- $Y$  : dependent variable
- Model: predict  $Y$  from  $X$
- $Y'$  : “ $Y$  prime” : predicted  $Y$
- $Y' = a + bX$
- Prediction is imperfect.
- Difference between predicted ( $Y'$ ) and actual ( $Y$ ) is called a “Residual” =  $(Y - Y')$
- Calculation of best fit line minimizes the sum of the squared residuals  $\Sigma(Y - Y')^2$

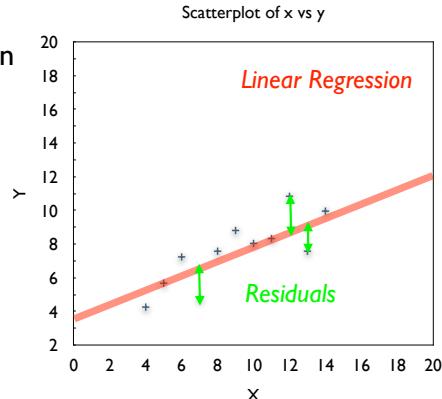
446

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Residuals in Linear Regression

Residuals are difference between actual Y and predicted Y' ( $Y - Y'$ )

Graphically it is equal to how far away (vertically) a point is from the linear regression line

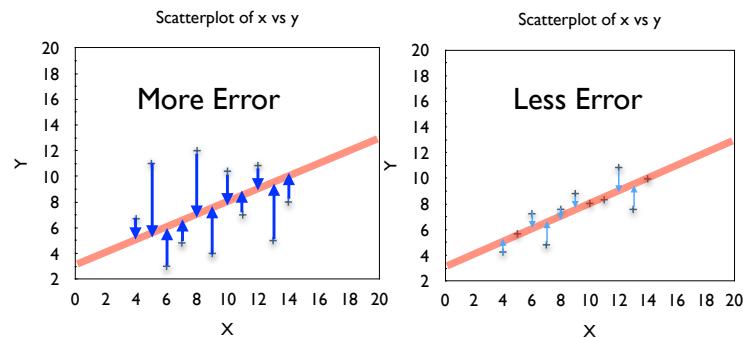


447

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Residuals and Error

Residuals (error) are greater when Y values are further from prediction.



448

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Residuals

$$d_i = y_i - y_i'$$

- In linear regression, the difference between the predicted y and actual y

449

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Measuring “fit”

- Can we use residuals to measure how well the predicted values measure the actual values?
- E.g. how big are the residuals
- Similar to how we calculate Standard Deviation with a single X variable*

450

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Sum of Squared Residuals

$$SSR = \sum_{i=1}^N d_i^2$$

$$SSR = \sum_{i=1}^N (y_i - y_i')^2$$

451

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Sum of Squared Residuals

- Residual =  $(Y_i - Y_i')$
- Squared residual =  $(Y - Y')^2$
- SSR: Sum of squared residuals
  - Linear regression minimizes this value
- SSR is hard to interpret
- Can we standardize SSR?
- Need to compare SSR to something else

452

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Sum of Squares Total

- What can we compare SSR to?
- SST
  - similar to the null hypothesis:
  - “what would SSR be if X and Y aren’t related at all?”
  - uses the mean of Y as the model

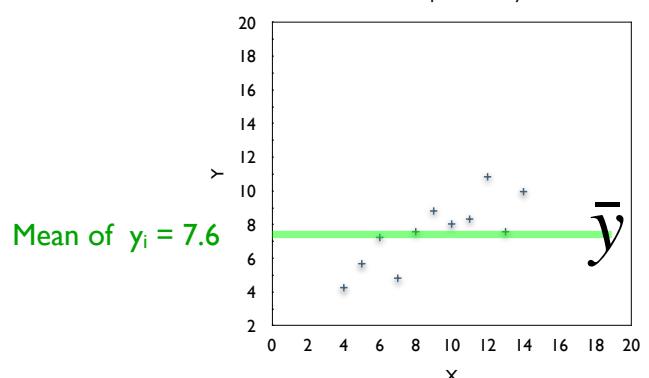
$$SST = \sum_{i=1}^N (y_i - \bar{y})^2$$

453

Psychology 402 - Fall 2022 - Dr. Michael Diehr

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

Scatterplot of x vs y



454

Psychology 402 - Fall 2022 - Dr. Michael Diehr

$R^2$

$$R^2 = 1 - \frac{SSR}{SST}$$

- $R^2 = 1 - (SSR/SST)$
- Ranges from 0 to 1 (0% to 100%)

455

Psychology 402 - Fall 2022 - Dr. Michael Diehr

$R^2$

- Terminology
  - Coefficient of Determination
  - Explained Variance
  - Shared Variance
- Meaning
  - what % of variation in Y values can we predict from the variation in X values
- Careful: *Correlation* is not causation

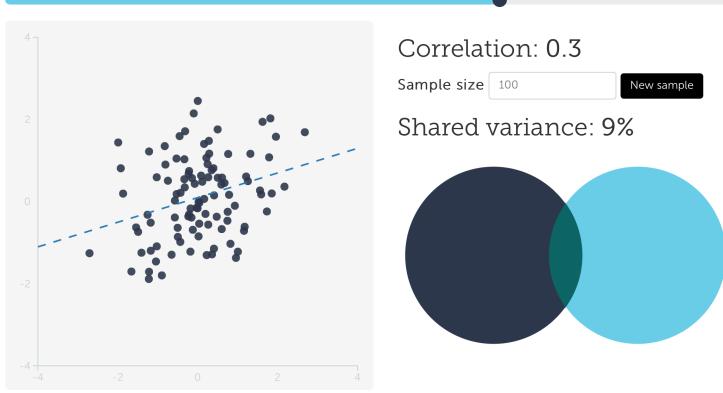
456

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Interactive Correlation Demo

- <http://rpsychologist.com/d3/correlation/>

Slide me



Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Ch. 3 - Part 2

462

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Review

- # of variables / dimensions
  - 1 Mean (SD)
  - 2 Linear Regression

466

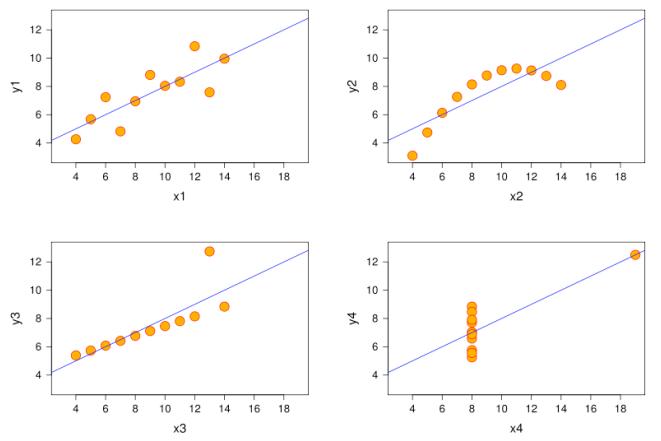
Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Review

467

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Anscombe's Quartet



472

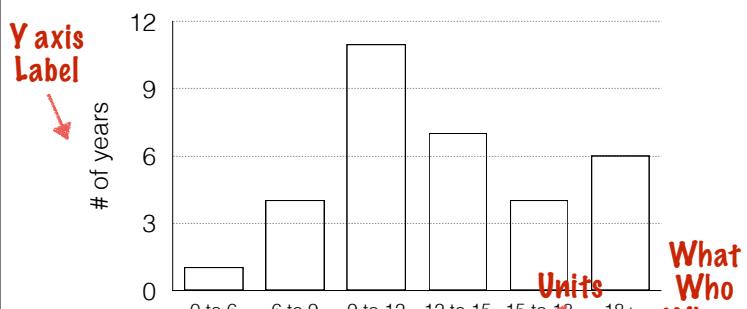
## Exercise 2 - GraphPad Prism

474

Psychology 402 - Fall 2022 - Dr. Michael Diehr

### Example of APA-7 style Histogram

**Figure 1**  
Frequency Distribution of Annual Rainfall in San Diego



**Note** Note. Annual rainfall for the years 1970-2002, measured at Lindberg Airport, by water year (October-September). Note the somewhat bi-modal distribution, with both 9 to 12 and 18 or more inches being most common.

Psychology 402 - Fall 2022 - Dr. Michael Diehr

### APA-7 Figure Example

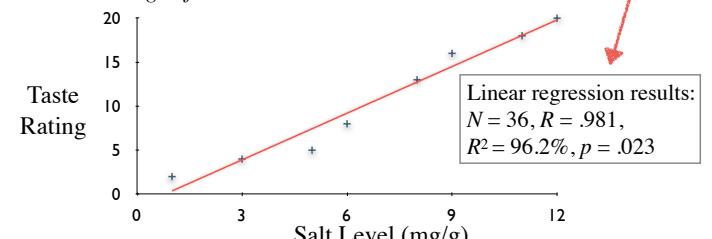
478

Title is above the figure

#### Figure 1

Taste Ratings of a Cracker in Relation to Salt Amount

Legend is within the figure



*Note.* Subjects ( $N=36$ ) ate a single dry cracker which varied in the amount of salt (milligrams per gram) and rated the taste on a 20 point scale.

Note the very strong correlation, suggesting higher salt levels are strongly related to taste ratings.

Note is below the figure

Psychology 402 - Fall 2022 - Dr. Michael Diehr

477

## Ch. 3 - Part 3

481

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Review

483

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Standard Error of Estimate

- Residual =  $(Y - Y')$
- Standard Deviation of residuals
  - measure of “average” error
  - aka “Standard Error of Estimate”
  - In Prism:  $S_{y,x}$

486

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Correlation : Pearson’s r

- Pearson’s Product-Moment Correlation
- Measures the strength of the linear relationship between two variables
- Ranges between -1.0 and +1.0
- Is a special case of linear regression, when both X and Y have been turned into Z scores.
- r is **transitive commutative** (correlation between X and Y is same as correlation between Y and X)
- $R^2$  = “explained variance” is the proportion of variation in the data explained by the model.
- $R^2$  ranges from 0 to 1.0 (0% to 100%)

488

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Regression vs. Correlation

	Linear Regression	Correlation
<b>Scores</b>	Raw	Z
<b>Mean, Std Dev</b>	sample means sample Std Dev	0 1
<b>Equation</b>	$Y' = a + bX$	$Y' = r X$
<b>Slope</b>	b = change in Y per change in X	r = correlation coefficient
<b>Slope<sup>2</sup></b>	meaningless	$R^2 = \%$ variance explained
<b>Commutative ?</b>	no	yes, $R_{xy} = R_{yx}$

489

Psychology 402 - Fall 2022 - Dr. Michael Diehr

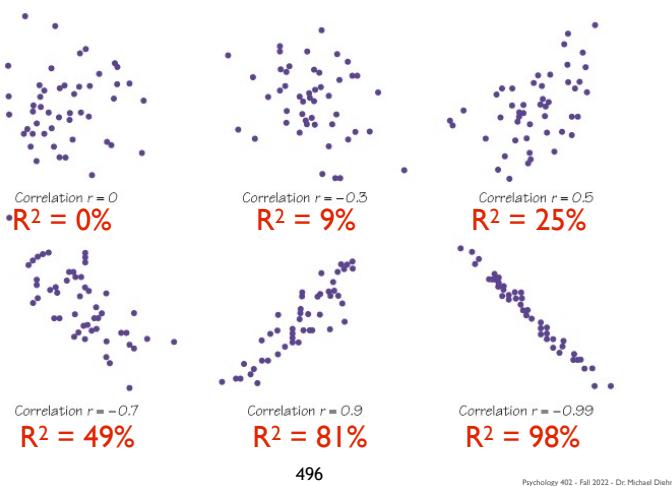
## R vs $R^2$

	R	$R^2$
<b>Minimum</b>	-1.0	0.0 (0%)
<b>Maximum</b>	1.0	1.0 (100%)
<b>Meaning</b>	correlation between X and Y	% of variance in Y explained by X
<b>AKA</b>	“correlation”, “correlation coefficient”	shared variance, explained variance, coefficient of determination
<b>Notes</b>	can be positive or negative	always positive (since it's squared)

490

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Correlations



## Interactive Correlation Example

- <http://rpsychologist.com/d3/correlation/>

- $R^2$  or “Explained Variance” is sometimes called “Shared Variance”

497

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Other Correlation Coefficients

- Continuous (interval & ratio): Pearson's  $r$
- Ordinal (Ranked): A B C D... 1st, 2nd, 3rd...
  - Spearman's Rho: correlation between two ordinal / ranked variables.
- Dichotomous (yes/no, one/zero, T/F, Male/Female, Pass/Fail...)
  - True vs. Artificial?

498

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Continuous vs. Dichotomous

Type of X / Type of Y	Continuous	Artificial Dichotomous	True Dichotomous
Continuous	Pearson $r$	Biserial $r$	Point biserial $r$
Artificial Dichotomous	Biserial $r$	Tetrachoric $r$	Phi
True Dichotomous	Point biserial $r$	Phi	Phi

499

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Correlation : Issues

- Technical / Calculation :
  - Non-normal distribution
  - Non-linear data and relationships
  - Outliers, data errors
  - Restricted Range
- Interpretation:
  - Correlation  $\Rightarrow$  Causation
  - Third variable explanations

500

Psychology 402 - Fall 2022 - Dr. Michael Diehr

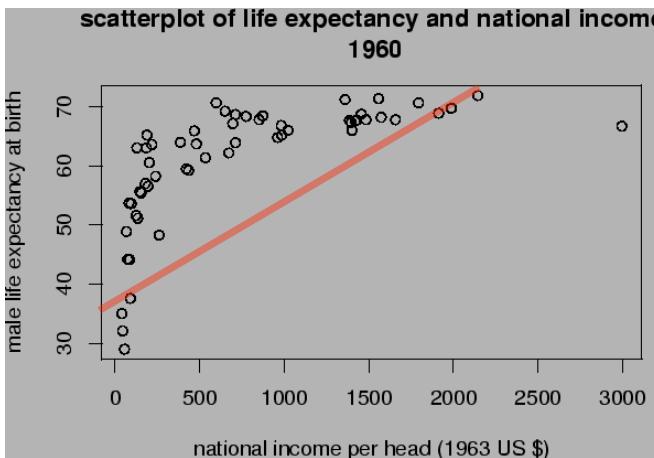
## Non-linearity

- Linear Regression & Correlation assume a linear relationship between X and Y
- When it's not linear:
  - Restrict the range of X
  - Transform (log, square root, etc.)
  - other statistical analyses (Spearman's Rho...)

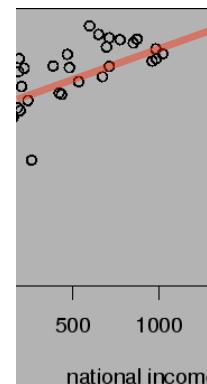
501

Psychology 402 - Fall 2022 - Dr. Michael Diehr

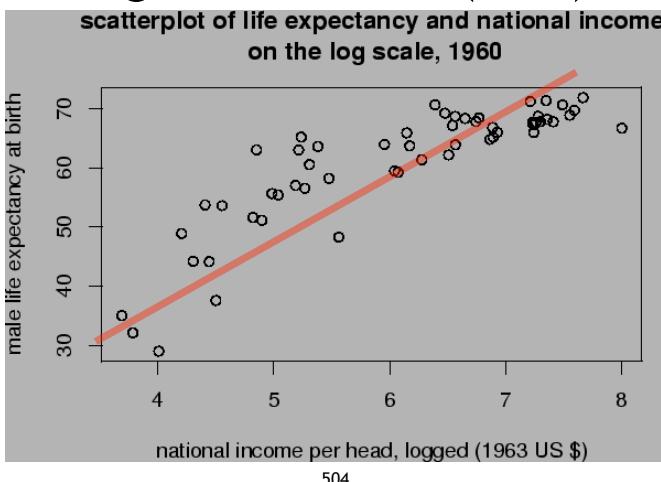
## Life expectancy / national income



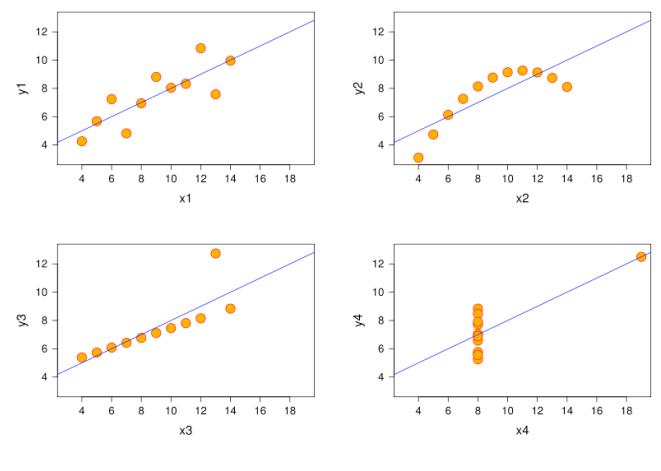
## Restrict range of X



## log transform X (or Y)



## Outliers & Data Errors?



## Correlation = Causation?

- A relationship (linear or otherwise) between X and Y tells us nothing about whether X causes Y
- Lack of correlation between X and Y does not mean that X doesn't cause Y
- Ice cream sales are positively related to increases in drowning deaths

## Hypothesis Testing

- Parameters estimated from sample data have error
- How do we know if a given estimate is correct?
- How big is the error likely to be (confidence intervals)?
- Inferential Statistics - covered later
  - Formulas to calculate probability, confidence intervals.
  - Higher N is better
  - “statistical significance” not the same as “clinical significance”

## Statistical vs Clinical Significance<sup>509</sup>

- Regarding the change in the Dependent Variable (DV)
- Statistical Significance:
  - Could the change be due to chance?
  - P value ( $p < .05$  : less than 5% probability)
- Clinical Significance
  - Was the change big enough to matter?
  - Effect Size ( $R^2$ )
  - Depends on context

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Significance vs. Effect Size

- Two coin flips : both heads (100%)
  - big effect size (50%)
  - not statistically significant ( $p=0.25$ )
- 1000 coin flips, 490 heads (49.0%)
  - small effect (1%)
  - statistically significant ( $p=0.02$ )
- 1000 coin flips, 350 heads (35%)
  - big effect (15%)
  - statistically significant ( $p<.00000001$ )

510

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Lies, damned lies, and statistics

- Statistical significance (P) is a function of...
- Errors of measurement (E)
- Effect Size (R)
- Sample Size (N)

- $P \sim E / (R \times N)$

511

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Reporting Results

- Headline: “Men had higher IQ than women. Results were significant  $p < .001$ ”
  - → “that’s very significant”
  - → “men are much smarter than women”
  
- P-value : statistically significant: Yes
- Effect Size : clinically significant: ? Unknown

512

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Review : Is race “real”?

- Pre-DNA theory
- Post-DNA theory

513

Psychology 402 - Fall 2022 - Dr. Michael Diehr

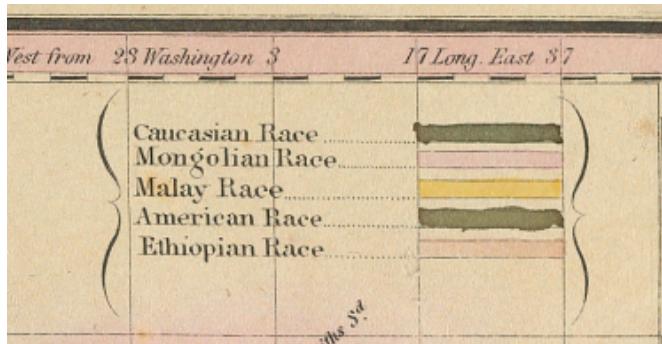
## Pre-DNA

- Gold, Silver, Brass, Iron -- Plato
- “There is a physical difference between the white and black races which I believe will forever forbid the two races living together on terms of social and political equality.” -- Abraham Lincoln

514

Psychology 402 - Fall 2022 - Dr. Michael Diehr

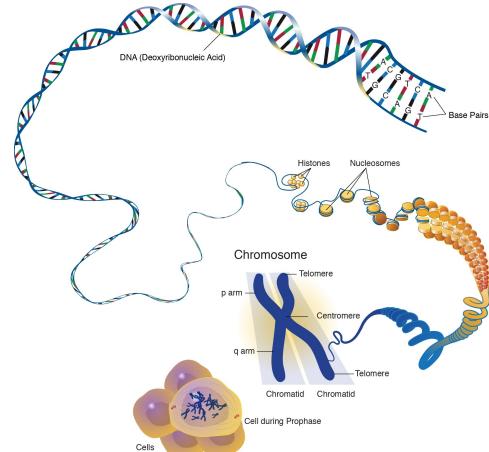
# Five Races?



515

Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Genetics : DNA



Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Genetics

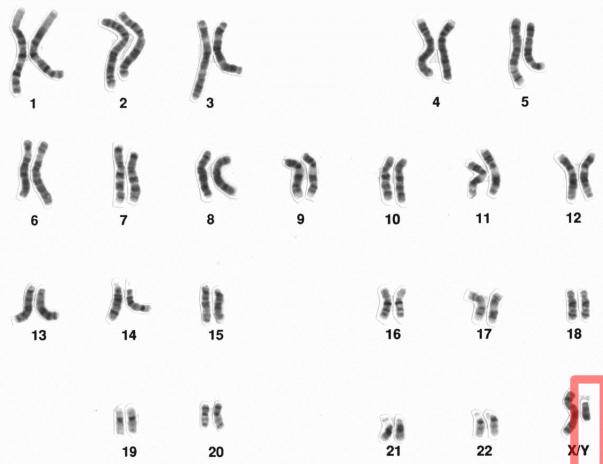
517

- Human genome contains about 3 billion pairs of deoxyribonucleic acid (DNA)
- DNA is Transcribed into RNA
- RNA is Translated into Proteins
- Proteins
  - serve as structural components
  - function as enzymes to catalyze biochemical reactions
- Human DNA is grouped into 46 chromosomes
  - 23 pairs, one of each pair comes from each parent
  - 22 pairs in both males and females (autosomes)
  - 1 pair determines sex: either "XX" (females) or "XY" (males)

Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Humans: 46 Chromosomes - 23 pairs

519



Michael Diehr

# Gene

520

- DNA is subdivided into Chromosomes
- Chromosomes are subdivided into Genes
- Gene is a functional unit of DNA
- makes one thing (single protein or RNA)

Psychology 402 - Fall 2022 - Dr. Michael Diehr

# Genetics : Species Differences

521

organism	estimated size (base pairs)	# genes	gene size	# chromosomes
Homo sapiens (human)	3.2 billion	~25,000	1 gene per 100,000 bases	46
Mus musculus (mouse)	2.6 billion	~25,000	1 gene per 100,000 bases	40
Drosophila melanogaster (fruit fly)	137 million	13,000	1 gene per 9,000 bases	8
Arabidopsis thaliana (plant)	100 million	25,000	1 gene per 4000 bases	10
Caenorhabditis elegans (roundworm)	97 million	19,000	1 gene per 5000 bases	12
Saccharomyces cerevisiae (yeast)	12.1 million	6000	1 gene per 2000 bases	32
Escherichia coli (bacteria)	4.6 million	3200	1 gene per 1400 bases	1
H. influenzae (bacteria)	1.8 million	1700	1 gene per 1000 bases	1

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Visible differences?

Indigenous  
Australian  
Melanesia  
African  
European

Australian and  
Africans are  
most genetically  
different



522

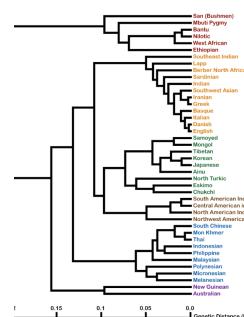
Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Genetic Differences

- Sub-Saharan African
- Indo-European
- East Asian
- Native American
- South Asian
- Aboriginal

$Fst$  = % of  
subpopulation  
variance

523



Psychology 402 - Fall 2022 - Dr. Michael Diehr

## DNA Variation

525

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## DNA Variation

- variation between individuals : 3mbp / person
- variation within groups : 85%
- variation between groups: 15%
  - 5% - within *population groups*
  - 10% - between *population groups*
- *Note: skin color is one of the few traits where the pattern is reversed*

526

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## DNA Differences

- Identical Twins
  - 0.0%
- Human vs. Human
  - 0.1%
- Humans vs Gorillas
  - 1.6%
- Humans vs Chimps:
  - 4.0%
- Humans vs. Cats
  - 10.0%



527

## Post-DNA theory

- Variance
  - variation between individuals
  - aka variation *within races population groups*
  - variation *between population groups*

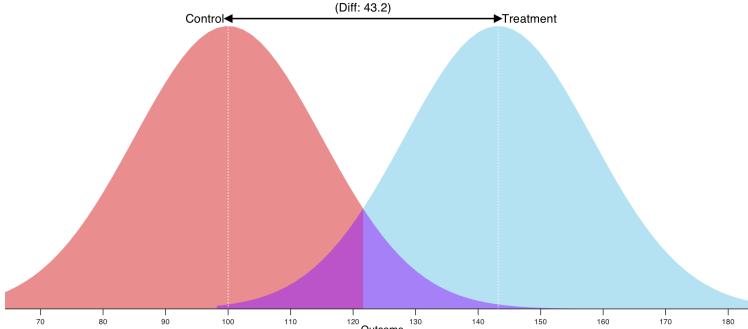
530

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Skin Color

Cohen's  $d$ : 2.9

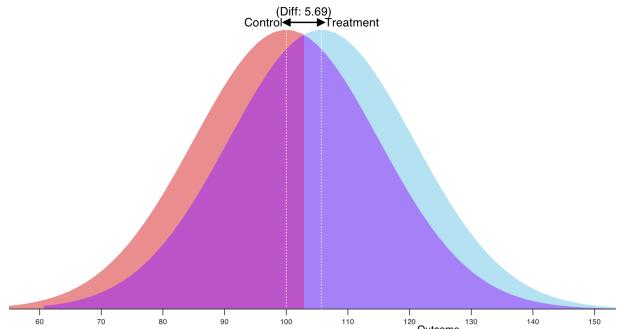
(Diff: 43.2)



## Many other traits

Cohen's  $d$ : 0.38

(Diff: 5.69)



- 85% between group, 15% within group
- 98% probability blue person higher than red

531

Psychology 402 - Fall 2022 - Dr. Michael Diehr

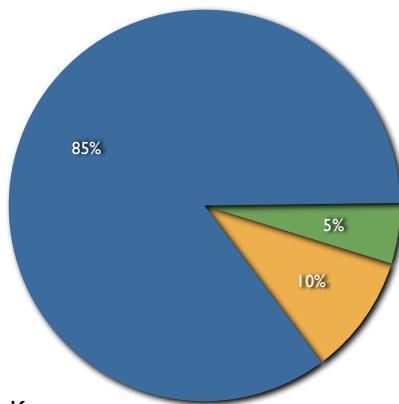
- 15% between group, 85% within group
- 61% chance blue person higher than red

532

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Variance: Genetic Variation

- Within local populations
- Within "race"
- Between "race"



For example:

- 85% within Japanese
- 5% between Japanese & Korean
- 10% between Asian and Caucasian

533

Psychology 402 - Fall 2022 - Dr. Michael Diehr

## Prehistorical Migration

