Statistical Methods for Affective Computing

Jeffrey M. Girard, *Carnegie Mellon University*
Jeffrey F. Cohn, *University of Pittsburgh*

Tutorial Resources are available at: **http://jmgirard.com/fg2018**

Presentation Overview

1. **Measurement and Validity**
   - Contemporary Validity Theory*
   - Inter-Rater Reliability*
   - Criterion Validity*
2. **Estimation and Uncertainty**
   - Effect Sizes and Confidence Intervals*
   - Generalized Linear Modeling
   - Preview of Advanced Topics

Measurement and Validity

Part 1: Contemporary Validity Theory

Contemporary Validity Theory

- **What is measurement?**

  "Measurement is the assignment of a number to a characteristic of an object or event, which can be compared with other objects or events."

  - Measurement involves the assignment of **numbers**
  - Measurement facilitates **comparisons** and decision-making
  - Measurement attempts to capture/describe **characteristics**

Contemporary Validity Theory

- **What is being measured?**
  - Some characteristics are directly observable, such as length or mass
  - Other characteristics are not directly observable and are called *constructs*
  - Constructs are hypothetical units and must be inferred from their effects
  - The linkage between constructs and effects is usually derived from theory
  - *Example constructs: gravity, dark matter, intelligence, emotions, personality*

Contemporary Validity Theory

- **How is measurement done?**
  1. **Instruments** combine theory and methodology
  2. **Measurements** apply instruments in a particular context
  3. **Inferences** interpret measurements using a logic system
  4. **Actions** are decided based on inferences and a value system
- Errors in any of these steps can lead to problems
- Errors in one step are "passed on" to all later steps

Contemporary Validity Theory

Instrument (Theory + Method)	**Measurement** (Instrument + Context)	**Inference** (Measurement + Logic)	**Action** (Inference + Values)
Patients endorse or deny 9 symptoms of clinical depression	Nurse calls patient on phone and patient endorses 4 symptoms	Infer that patient is not suffering from clinical depression	Do not refer patient for treatment or further assessment
Percentage of objects for which algorithm predicted the label	In testing dataset, baseline = 90% and new algorithm = 92%	Infer that new algorithm is better than the baseline	Replace all instances of baseline algorithm with new algorithm

How might error creep into each step of these examples?

Contemporary Validity Theory

- **What is validity?**

  "Validity is the extent to which available evidence supports interpretations of scores to reflect standing with respect to the specified construct"

  - Validity reflects the connection between measurements and constructs
  - Validity is described along a continuum of support (not dichotomous)
  - Validity pertains to inferences based on measurements (not instruments)
  - Validation is an on-going process that evolves (not a one-time activity)

Contemporary Validity Theory

- **How are measurement-based inferences validated?**
  1. Evidence based on test content
     - The instrument covers all important aspects of the construct
     - The instrument respects the "boundaries" of the construct
  2. Evidence based on hypothesized relationships among variables
     - Measurements generalize/are consistent across changes in context
     - Measurements are (un)related to variables that they should (not) be
     - *We will talk in depth about reliability and criterion validity*

Contemporary Validity Theory

- **Where can I read more?**
  - Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.
  - Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212–225.
  - Girard, J. M., & Cohn, J. F. (2016). A primer on observational measurement. *Assessment*, *23*(4), 404–413.

Measurement and Validity

Part 2: Inter-Rater Reliability

Inter-Rater Reliability

- **What is reliability?**

  "Reliability quantifies how similar measurements are across contexts"

  - Do scores differ across times, versions, places, raters, groups, etc.?
  - Test-retest reliability is reliability across time/administrations
  - Inter-rater reliability is reliability across raters/judges/scorers
  - Generalizability is a broad term for reliability across groups, places, etc.

Inter-Rater Reliability

- **Why is reliability important?**
  - Reliability imposes a limit on the validity of measurements
  - If reliability is poor, then measurements are too inconsistent to be useful
  - Inferences and actions based on measurements require consistency
    - Imagine a scale that told you that your weight was changing ±5 lbs every minute
    - Imagine a clinical test that some doctors read as positive and others read as negative
    - Imagine a facial biometric system that recognized you indoors but not outdoors

Inter-Rater Reliability

- **What role does reliability play in affective computing?**
  - Measurements are often used to train supervised learning algorithms
  - These training labels are called "ground truth" and *assumed* to be correct
  - The reliability of training labels can impact algorithm performance
  - Any biases inherent to the labels will likely be inherited by the algorithm
  - If you have training labels, you must give an estimate of their reliability
  - If you generate new measurements, you should explore their reliability

Inter-Rater Reliability

- **How can reliability be estimated or explored?**
  - Similar (or identical) objects are measured in different contexts
    - Have multiple observers watch and label the same media files
    - Have multiple participants rate their experience of the same tasks
    - Have the same participants engage in the same tasks in different settings
  - These measurements are then compared using statistical methods
    - There are many approaches to estimating the different types of reliability
    - Each approach has its own set of advantages and disadvantages

Inter-Rater Reliability

- **Why focus on inter-rater reliability?**
  - The methods used for all types of reliability are similar (or identical)
  - The most common use of reliability in AC is between raters for labels
  - This allows you to provide evidence that your labels are reliable/valid
  - When there is no ground truth, we settle for consistency among raters
- **What specific approaches will we explore?**
  - For categorical measurements, we will discuss agreement indexes
  - For dimensional measurements, we will discuss correlation coefficients

Inter-Rater Reliability

- **How should measurement data be formatted?**
  - Create an object-by-rater ($n \times r$) matrix or data frame
  - All elements in the matrix or data frame should be numbers, so convert categories if necessary (e.g., {No, Yes} → {0, 1} or {A, B, C} → {1, 2, 3})
  - Include object-by-rater omissions but mark them as missing (NA or NaN)

	R1	R2	R3	R4	R5
1	2	2	3	2	2
2	2	2	2	2	2
3	2	NaN	2	2	1
4	1	2	2	2	2

‘Categorical-Data.csv’ dataset
Categories include {1, 2, 3}
Includes 4 objects and 5 raters
One omission added for illustration

Inter-Rater Reliability

- **What is an agreement index?**
  - Agreements are when two raters assign an object to the same category
  - With multiple raters, we can calculate agreement between pairs of raters

$$p_o = \frac{\text{Observed Agreement}}{\text{Possible Agreement}}$$

  - When categories are ordered, some disagreements are worse than others
  - In such cases, each pair of raters can be awarded varying degrees of credit
  - The amount of credit awarded is determined by a weighting scheme

Inter-Rater Reliability

```
% Import data from CSV file
>> CODES = csvread('Categorical-Data.csv');

%Compute agreement for unordered categories
>> mAGREE(CODES, 1:3, 'identity')
Percent observed agreement = 0.675

% Compute agreement for linear categories
>> mAGREE(CODES, 1:3, 'linear')
Percent observed agreement = 0.838

% Compute agreement for quadratic categories
>> mAGREE(CODES, 1:3, 'quadratic')
Percent observed agreement = 0.919
```

Example Dataset

	R1	R2	R3	R4	R5
1	2	2	3	2	2
2	2	2	2	2	2
3	2	NaN	2	2	1
4	1	2	2	2	2

Amount of Credit Awarded

	Same	1 Away	2 Away
Identity	1.00	0.00	0.00
Linear	1.00	0.50	0.00
Quadratic	1.00	0.75	0.00

Inter-Rater Reliability

- **Are there issues with agreement?**
  - What if raters guess and end up agreeing by chance?
  - What is the right "baseline" to compare agreement to?
  - What if some categories are more common than others?
  - What if agreement is higher for some categories than others?
- **Can we address these issues?**
  - Chance-adjusted agreement indexes try to address the first two issues
  - Category-specific agreement indexes try to address the last two issues

Inter-Rater Reliability

- **What is a chance-adjusted agreement index?**
  - How much agreement would occur "by chance" alone (p_c)?
  - If we know p_c, we can adjust observed agreement by this amount

$$r_i = \frac{p_o - p_c}{1 - p_c} = \frac{\text{Observed Nonchance Agreement}}{\text{Possible Nonchance Agreement}}$$

  - This yields the general form of a chance-adjusted agreement index (r_i)
  - It is still the ratio of observed to possible agreement, but p_c is removed
  - "When raters could agree *honestly*, how much did they do so?"

Inter-Rater Reliability

- **How can chance agreement be estimated?**
  - We need to build a "baseline" model to compare raters to
  - In practice, p_c is only an estimate of chance agreement
  - Different chance-adjusted indexes are based on different assumptions
  - They usually use the same general form (r_i) but estimate p_c differently
  - There are two primary types of assumptions about chance agreement

Inter-Rater Reliability

- **What are the category-based assumptions?**
  - Each category has an equal probability of being randomly selected
  - So chance is modeled as "flipping coins" or "rolling dice"
  - Bennett et al.'s (1954) *S* score was the first version of this approach
- **What are the distribution-based assumptions?**
  - Each category's probability of being randomly selected is equal to its prevalence
  - So chance is modeled as "meeting a quota" for each category
  - Quotas may be rater-specific (Cohen's κ) or shared (Scott's π, Krippendorff's α)

Inter-Rater Reliability

% Compute S for unordered categories
>> mSSCORE(CODES, 1:3, 'identity')
Percent observed agreement = 0.675
Percent chance agreement = 0.333
Bennett et al.'s S score = 0.513

% Compute kappa for unordered categories
>> mKAPPA(CODES, 1:3, 'identity')
Percent observed agreement = 0.675
Percent chance agreement = 0.725
Cohen's kappa coefficient = −0.182

Example Dataset

	R1	R2	R3	R4	R5
1	2	2	3	2	2
2	2	2	2	2	2
3	2	NaN	2	2	1
4	1	2	2	2	2

Inter-Rater Reliability

- **What is the category-specific agreement index?**
  - What if some categories are more difficult/ambiguous than others?
  - To explore this, we can calculate agreement for specific categories

$$SA_k = \frac{\text{Observed Agreement on Category } k}{\text{Possible Agreement on Category } k}$$

  - SA_k is the conditional probability of a random rater assigning a random object to category k given that another random rater already did so
  - SA_k is based on the work of Dice (1945) and Sørensen (1948)

Inter-Rater Reliability

% Compute specific agreement
>> mSPECIFIC(CODES, 1:3, 'identity')
Specific agreement for category 1 = 0.000
Specific agreement for category 2 = 0.820
Specific agreement for category 3 = 0.000

Example Dataset

	R1	R2	R3	R4	R5
1	2	2	3	2	2
2	2	2	2	2	2
3	2	NaN	2	2	1
4	1	2	2	2	2

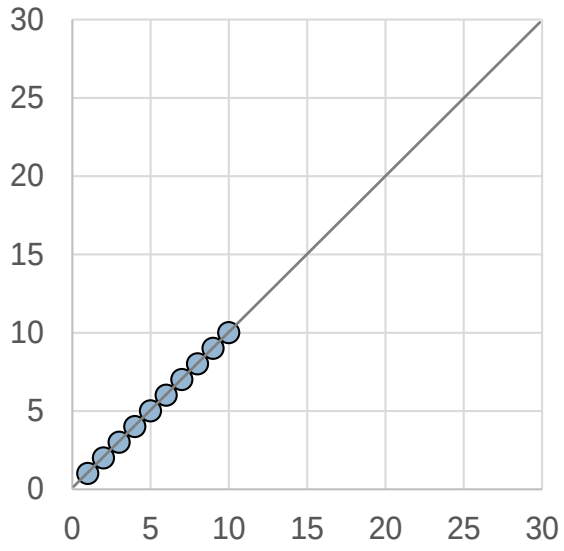
Compare these SA_k results to $S = 0.513$ and $\kappa = -0.182$
Each approach tells a very different story about reliability.
Which assumptions are we most comfortable with?
What are the pros and cons of each approach?

Inter-Rater Reliability

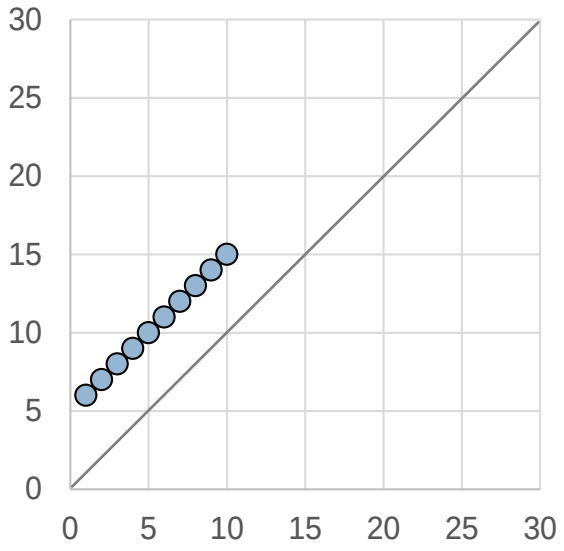
- **What is a correlation coefficient?**
  - Variance is a measure of the amount of spread or dispersion in a variable
  - Variance comes from difference sources and can be partitioned by source
  - Correlation coefficients are normalized measures of co-variance (−1 to 1)
  - Various correlation coefficients can be used to measure reliability
  - *We will discuss several intra-class correlation coefficients (ICCs)*

Inter-Rater Reliability

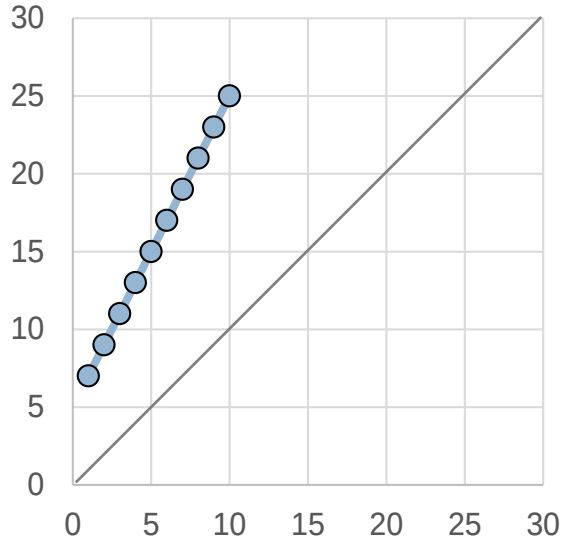
Agreement ICC	Consistency ICC
Requires $Y = X$	Allows $Y = X + b$



$A = 1.0$ $C = 1.0$



$A = 0.4$ $C = 1.0$



$A = 0.2$ $C = 0.8$

Inter-Rater Reliability

Agreement ICC (Intra-class correlation)	**Consistency ICC** (Intra-class correlation)
$A = \frac{\sigma_{row}^2}{\sigma_{row}^2 + \sigma_{col}^2 + \sigma_{err}^2}$	$C = \frac{\sigma_{row}^2}{\sigma_{row}^2 + \sigma_{err}^2}$
$\frac{\text{Object}}{\text{Object} + \text{Rater} + \text{Error}}$	$\frac{\text{Object}}{\text{Object} + \text{Error}}$
High A comes from high object, low rater, and low error variance	High C comes from high object variance and low error variance

Note. Because it is the numerator in the ICC formulas, low object variance makes a high ICC almost impossible.

Inter-Rater Reliability

```
% Import data from CSV file
>> RATINGS = csvread('Dimensional-Data.csv');

% Compute ICCs for single measures
>> ICC_C_1(RATINGS)
Single measures consistency ICC = 0.622
>> ICC_A_1(RATINGS)
Single measures agreement ICC = 0.558

% Compute ICCs for average measures
>> ICC_C_k(RATINGS)
Average measures consistency ICC = 0.767
>> ICC_A_k(RATINGS)
Average measures agreement ICC = 0.716
```

Example dimensional data

	R1	R2
1	7.800	7.800
2	7.800	-34.000
3	42.170	-120.556
4	101.950	-123.600
5	184.033	-151.630
…	…	…

Inter-Rater Reliability

- **What is some practical advice about reliability?**
  - Select the appropriate measure(s) of reliability
    - For categorical data, report: agreement, S score, kappa, and specific agreement
    - Use identity weights for unordered categories and linear weights for ordered ones
    - For dimensional data, report: consistency ICC and agreement ICC
    - Use average scores ICCs only if the mean of all raters will be analyzed
  - Select an appropriate sample for reliability analysis
    - Include as many and as varied objects as possible (minimum = 30)
    - Include all raters of interest and have them all rate the same objects

Inter-Rater Reliability

- **What's a good score on a measure of reliability?**

Chance-adjusted Agreement		Intra-class Correlation	
.80 – 1.00	Very Strong	.75 – 1.00	Excellent
.60 – .79	Strong	.60 – .74	Good
.40 – .59	Moderate	.40 – .59	Fair
.20 – .49	Weak	Less than .40	Poor
Less than .20	Very Weak		

A reasonable goal is 0.80 or higher and a reasonable minimum is 0.60

Inter-Rater Reliability

- **Where can I read more?**
  - Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Gaithersburg, MD: Advanced Analytics.
  - McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.
  - Zhao, X., Liu, J. S., & Deng, K. (2012). Assumptions behind inter-coder reliability indices. In C. T. Salmon (Ed.), *Communication Yearbook* (pp. 418–480). Routledge.
- **Where can I find those functions?**
  - http://mreliability.jmgirard.com (MATLAB)
  - http://www.agreestat.com/r_functions.html (R)

Measurement and Validity

Part 3: Criterion Validity

Criterion Validity

- **What is criterion validity?**

  "Criterion validity quantifies how well measurements match a gold standard"

  - Criterion validity is an important source of validity evidence
  - It is based on hypothesized relationships among variables
  - Do scores converge with those of established, trusted instruments?
    - Similar to reliability in that it quantifies consistency among variables
    - Different from reliability in that there is a "correct" score for each object

Criterion Validity

- **Why is criterion validity important?**
  - If the criterion measure is correct, then our measure should match it
  - Such a match provides evidence that we are measuring the same thing
  - It suggests that the new instrument may be used in place of the criterion
    - So if the new instrument is more time or cost-effective, this can be a big gain
- **What are the dangers and limitations of criterion validity?**
  - Criterion measurements may be noisy, incorrect, or biased themselves
  - Associations between new and criterion measures may be confounded

Criterion Validity

- **How can criterion validity be estimated or explored?**
  - Collect measurements of the same (or similar) objects using the new instrument and the trusted criterion instrument(s)
  - Explore the similarity between the new and criterion measurements
    - This can be done using a number of statistical and graphical approaches
    - The reliability indexes can be used for criterion validity (with only two columns)
    - However, there are other approaches that leverage our trust in the criterion scores

Criterion Validity

- **How is criterion validity estimated for categorical measurements?**
  - In the case of multiple categories, you can use an agreement index
  - However, there are special approaches for dichotomous measurements
  - These approaches are often based on the 2x2 contingency table

	Criterion = Yes Patient w/Condition	Criterion = No Patient w/o Condition
Test = Yes Positive Test Result	True Positives (TP)	False Positives (FP)
Test = No Negative Test Result	False Negatives (FN)	True Negatives (TN)

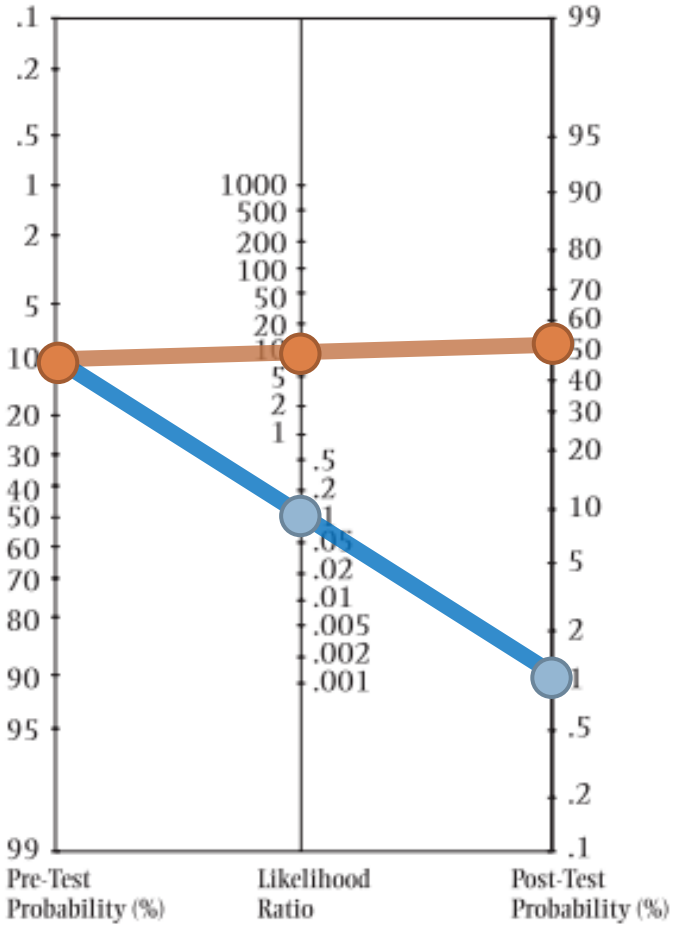
Criterion Validity

- **What are the most popular indexes derived from the 2x2 table?**
  - Sensitivity = TP / (TP + FN)
    - Probability of patient with condition receiving a positive test result
  - Specificity = TN / (TN + FP)
    - Probability of patient without condition receiving a negative test result
  - Positive Likelihood Ratio = Sensitivity / (1 – Specificity)
    - Increase in likelihood of diagnosis given a positive test result
  - Negative Likelihood Ratio = (1 – Sensitivity) / Specificity
    - Reduction in likelihood of diagnosis given a negative test result

Criterion Validity

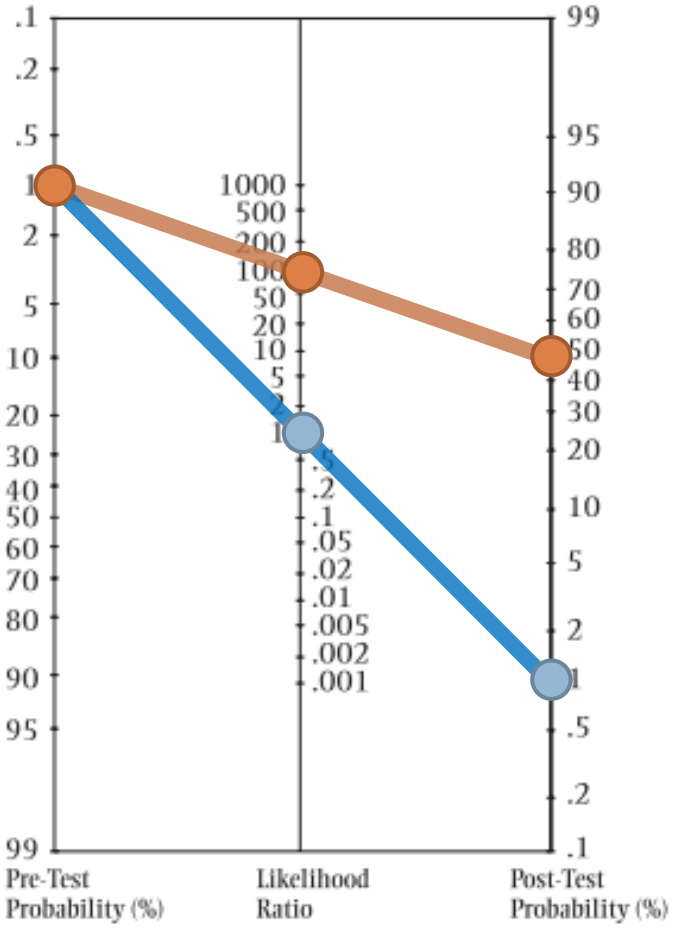
- **Why are likelihood ratios useful?**
  - Before testing, we can guess at the pre-test probability of a diagnosis
  - This probability is derived from the condition's estimated prevalence rate
  - A positive/negative test result should increase/decrease this probability
  - How much this post-test probability changes is based on likelihood ratios
  - We can visualize these probabilities and ratios using Fagan's nomogram

Criterion Validity



Pre = 10%
LR+ = 10.0
Post = 50%

Pre = 10%
LR- = 0.1
Post = 1%



Pre = 1%
LR+ = 100.0
Post = 50%

Pre = 1%
LR- = 1.0
Post = 1%

Criterion Validity

```
% Load data from CSV file
>> dat = csvread('Binary-Data.csv');

% Calculate performance measures
>> results = mBINARY(dat);

% Extract likelihood ratios from results
>> results.PLR
Positive Likelihood Ratio = 2.1429
>> results.NLR
Negative Likelihood Ratio = 0.4286
```

Example binary data

	Test Result	Criterion
1	1	1
2	1	0
3	0	0
4	1	1
5	0	1
6	0	1
7	0	0
8	1	1
9	1	1
10	1	1

Criterion Validity

- **Final thoughts on categorical measurements**
  - Pre-test probabilities may differ across contexts (e.g., locations or groups)
  - This can have a big impact on post-test probabilities and thus usefulness

LR+ < 1.0	1.0 < LR+ < 5.0	5.0 < LR+ < 10.0	LR+ > 10.0
Not Meaningful	Small	Moderate	Large

  - There are many other alternatives to explore (e.g., ROC analysis)
  - It is helpful to explore, and to report, several criterion validity measures

Criterion Validity

- **How is criterion validity estimated for dimensional measurements?**
  - The ICC approaches can be used here if interpreted properly
  - The inter-class correlation (PCC) is also sometimes used here
  - There are also a few additional approaches for criterion validity
  - Rather than partitioning variance, these approaches quantify error

Criterion Validity

- **How can we quantify error for criterion validity?**
  - We can simply take the average of the absolute differences

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$$

  - Or we can square the differences first to penalize larger errors

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

  - The MAE approach is easier to interpret and less sensitive to outliers

Criterion Validity

- **Can we normalize these values to a common metric?**
  - MAE and MRSE can be normalized by dividing by the range, mean, or SD
  - Either criterion-observed values or known/established values can be used

$NMAE_R = \frac{MAE}{y_{max} - y_{min}}$	$NRMSE_R = \frac{RMSE}{y_{max} - y_{min}}$
$NMAE_M = \frac{MAE}{\bar{y}}$	$NRMSE_M = \frac{RMSE}{\bar{y}}$
$NMAE_S = \frac{MAE}{s_y}$	$NRMSE_S = \frac{RMSE}{s_y}$

Criterion Validity

```
% Load data from CSV file
>> dat = csvread('Dimensional-Data.csv');

% Calculate MAE
>> result = mABSERROR(dat)
Mean Absolute Error = 336.0038

% Calculate MAE normalized by known range
>> result = mABSERROR(dat, 2000)
Normalized MAE = 0.1680

% Calculate MAE normalized by observed SD
>> result = mABSERROR(dat, 'std')
Normalized MAE = 0.7090
```

Example dimensional data

	R1	R2
1	7.800	7.800
2	7.800	-34.000
3	42.170	-120.556
4	101.950	-123.600
5	184.033	-151.630
…	…	…

Criterion Validity

- **Where can I read more?**
  - Bowden, S. C. (Ed.). (2017). *Neuropsychological assessment in the age of evidence-based practice*. New York, NY: Oxford University Press.
  - Luo, W., Phung, D., Tran, T., Gupta, S., Rana, … (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research*, *18*(12), e323.
  - Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82.
- **Where can I find those functions?**
  - http://mreliability.jmgirard.com (MATLAB)
  - https://achekroud.github.io/nomogrammer_vignette.html (R)

Estimation and Uncertainty

Part 1: Effect Sizes and Confidence Intervals

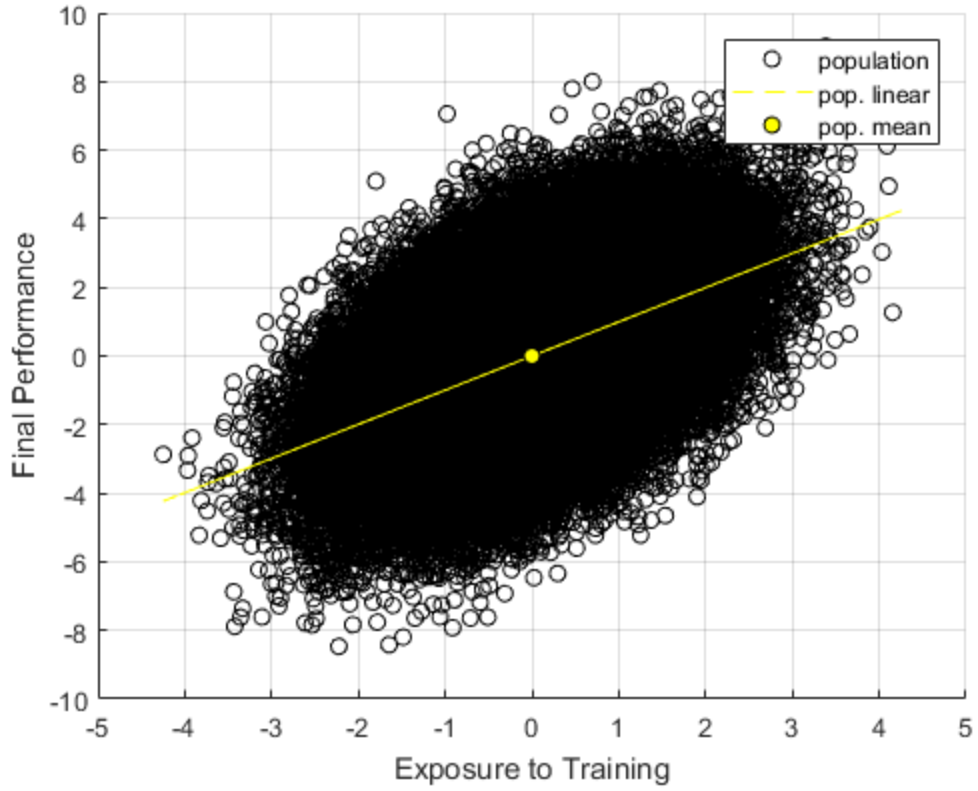
Effect Sizes and Confidence Intervals

- **What is the point of a research study?**
  - To understand variables and relationships in a population of interest
  - Population parameters are true means, differences, correlations, etc.
  - Sample statistics estimate parameters using a subset of the population
  - Sampling error is the difference between a parameter and its estimate
- **Goal #1:** Estimate the magnitude of population parameters
- **Goal #2:** Quantify the precision of estimates given sampling error

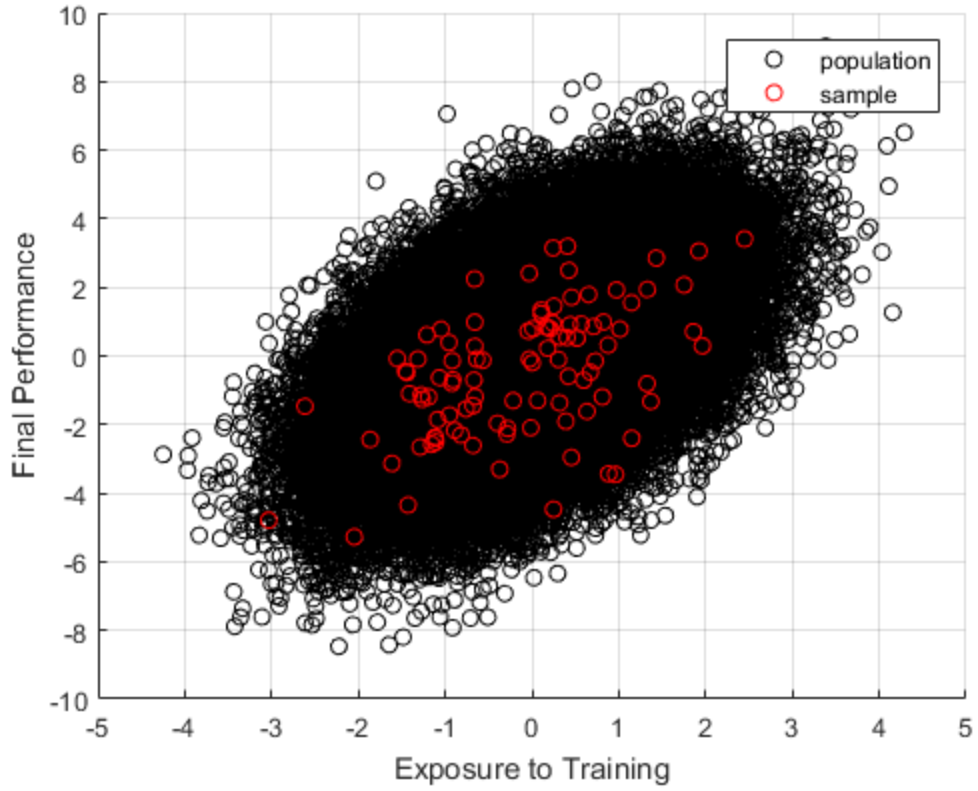
Effect Sizes and Confidence Intervals

- **When can effect sizes be useful in affective computing?**
  - When you want to estimate a single mean value
    - The F1 score for a classifier or MAE for a regression model
  - When you want to compare the means of two groups
    - The mean difference between pre (time 1) and post (time 2) scores
    - The mean difference between depressed and non-depressed groups
  - When you want to know the level of association between variables
    - The correlation between your measures of exposure to training and performance
    - The amount of variance in accuracy explained by subject gender, age, and race

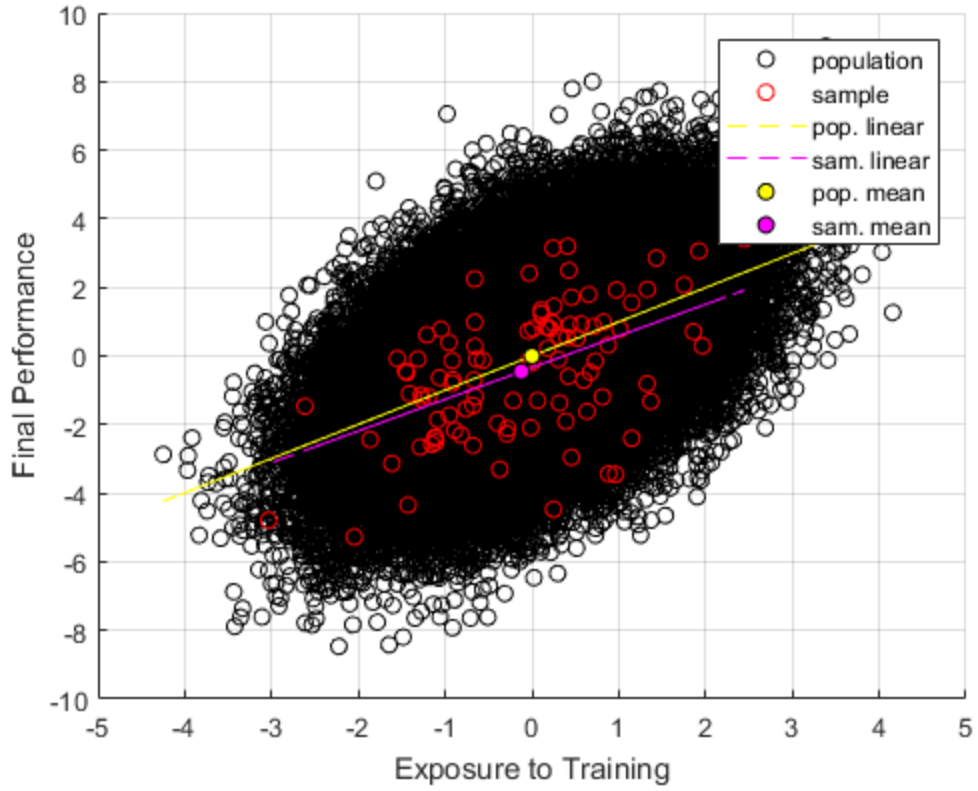
Effect Sizes and Confidence Intervals



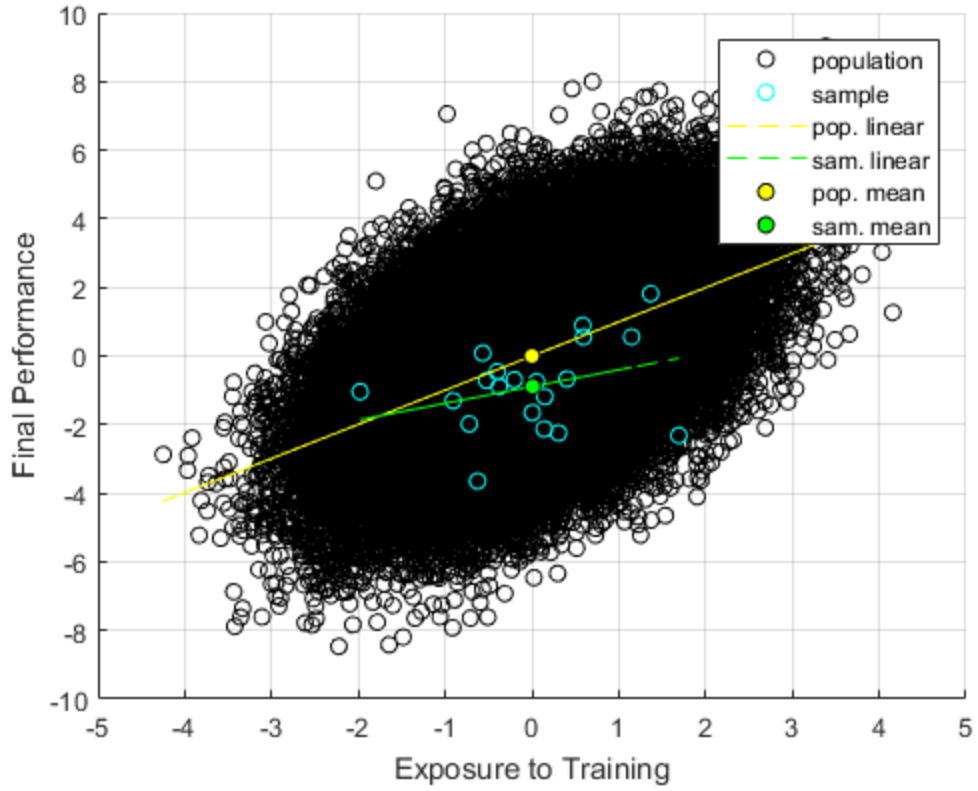
Effect Sizes and Confidence Intervals



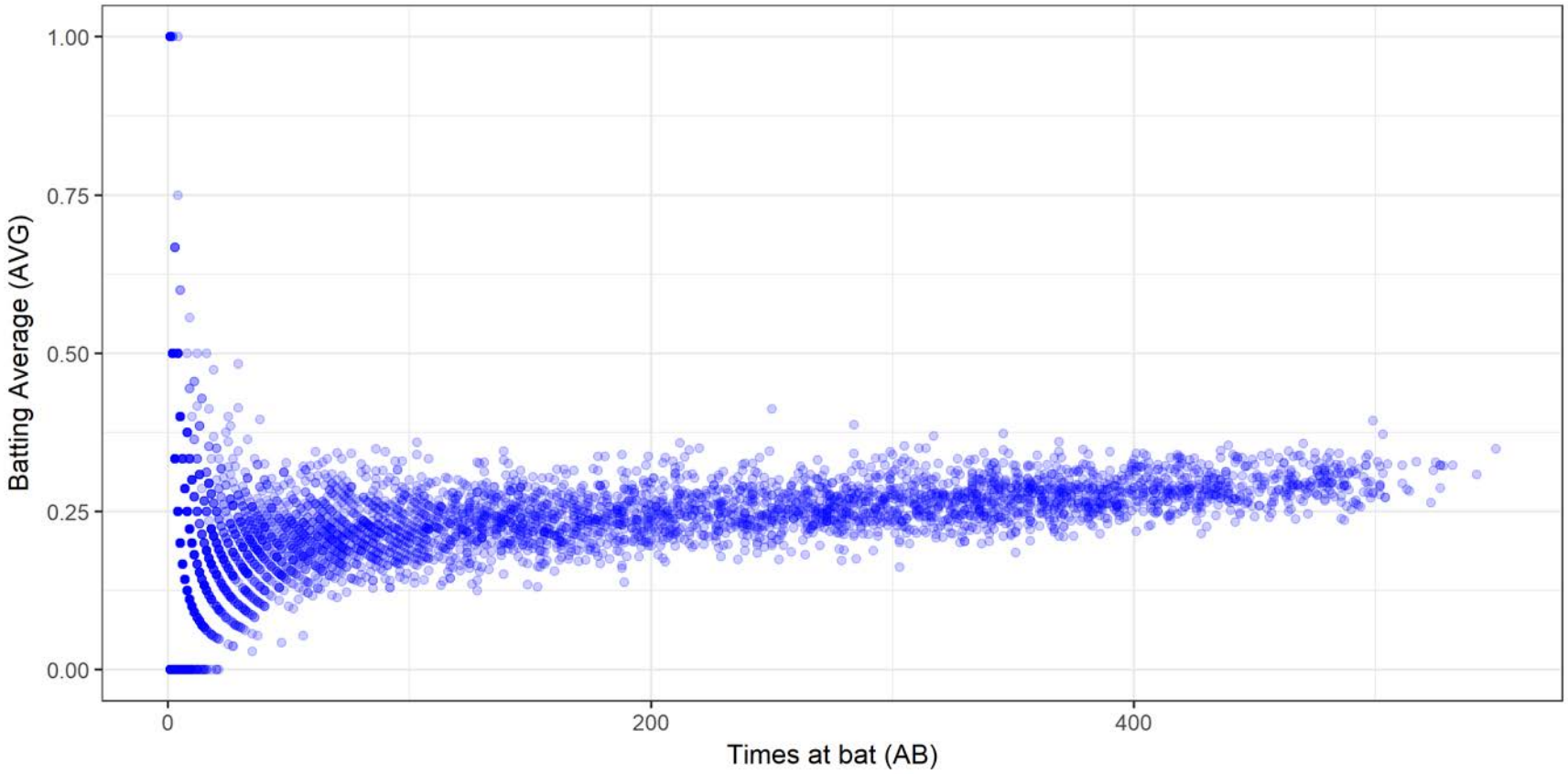
Effect Sizes and Confidence Intervals



Effect Sizes and Confidence Intervals



Effect Sizes and Confidence Intervals



Real data from 5236 players in the Korean Baseball Organization from 1982-2007

Effect Sizes and Confidence Intervals

- **Effect sizes quantify the magnitude of an effect**
  - Effects are means, differences between means, associations, etc.
  - Effect sizes are *point estimates* of population parameters
  - Provide an answer to "How large do we think this effect is?"
- **Confidence intervals quantify the precision of an estimate**
  - Account for the amount of variability in estimate and sampling error
  - Confidence intervals are *interval estimates* of population parameters
  - Provide an answer to "What other effect size values are plausible?"

Effect Sizes and Confidence Intervals

- **What are some popular and useful effect size measures?**
  - The *d* family measures the difference between two groups' mean scores
    - What is the difference between group 1's mean score and group 2's mean score?
    - What is the difference between group 1's mean score at time 1 and at time 2?
  - The *r* family measures the association between two (or more) variables
    - What is the association between variable 1 and variable 2?
    - How much variance in variable 1 is explained by variables 2, 3, and 4?
  - Direct estimates of population values of interest are also effect sizes
    - What is the mean score for this group? What is the standard deviation?

Effect Sizes and Confidence Intervals

- **What is the d-family of effect size measures?**
  - The d family includes mean contrasts in standard deviation units
    - $d = 0.75$ says the mean of group 1 is 3/4 of a SD higher than the mean of group 2
  - The form of the d family effect sizes is always the same

$$\text{Parameter } \delta = \frac{\mu_1 - \mu_2}{\sigma^*} \text{ and Statistic } d = \frac{M_1 - M_2}{s^*}$$

  - Different versions of d use different SDs for normalization

d_{pool}	d_{s1}	d_{s2}	d_{total}	d_{diff}	d_{Win}
Pooled SD	Sample 1 SD	Sample 2 SD	Total SD	Contrast SD	"Robust"

Effect Sizes and Confidence Intervals

```
% Load data from CSV file
>> dat = csvread('Dimensional-Data.csv');

% Calculate d using pooled SD
>> mDES(dat(:, 1), dat(:, 2), 'pool')
d_pool = 0.6311

% Calculate d using SD from group 2
>> mDES(dat(:, 1), dat(:, 2), 's2')
d_s2 = 0.4112

% Calculate d using paired differences
>> mDES(dat(:, 1), dat(:, 2), 'diff')
d_diff = 0.5534
```

Example dimensional data

	R1	R2
1	7.800	7.800
2	7.800	-34.000
3	42.170	-120.556
4	101.950	-123.600
5	184.033	-151.630
…	…	…

Effect Sizes and Confidence Intervals

- **What is the r-family of effect size measures?**
  - The r family includes associations based on correlation/regression
  - Correlation measures are signed and standardized from -1.0 to $+1.0$
    - r is a correlation between two continuous observed variables
    - There are other correlations (e.g., polychoric) for non-normal variables
  - Variance measures are unsigned and standardized from 0.0 to 1.0
    - r^2 is the amount of variance explained by a correlation coefficient
    - R^2 is the amount of variance explained by a multiple regression model

Effect Sizes and Confidence Intervals

```
% Load data from CSV file
>> dat = csvread('Dimensional-Data.csv');

% Calculate r using correlation function
>> r = corr(dat(:, 1), dat(:, 2))
Correlation Coefficient = 0.6693

% Calculate r^2 the old-fashioned way
>> rsq = r ^ 2
Coefficient of Determination = 0.4480
```

Example dimensional data

	R1	R2
1	7.800	7.800
2	7.800	-34.000
3	42.170	-120.556
4	101.950	-123.600
5	184.033	-151.630
…	…	…

Effect Sizes and Confidence Intervals

- **How can the magnitude of effect sizes be interpreted?**
  - Heuristic rules for interpreting effect size values exist
    - $|d| > .20$ = Small, $|d| > .50$ = Medium, $|d| > .80$ = Large
    - $|r| > .10$ = Small, $|r| > .30$ = Medium, $|r| > .50$ = Large
  - However, such rules can be misleading and context is crucial
    - Any given value may be small or large in different research domains
    - The best way to judge relative magnitude is to consult similar studies
  - An effect size is important if it has theoretical/practical implications
  - Papers' discussion sections should address "substantive significance"

Effect Sizes and Confidence Intervals

- **What is a confidence interval?**
  - A confidence interval extends above and below an effect size estimate
  - It shows a range of values for the estimate that are also highly plausible
  - Parametric CI's assume that the data are normally distributed
    - Take the estimate and add/subtract an estimate of sampling error ($SE \times t_{crit}$)
  - Non-parametric (e.g., bootstrapped) CI's do not assume normality
    - Take many samples with replacement and calculate the statistic in each
    - Look at the observed distribution of statistics to determine the CI bounds

Effect Sizes and Confidence Intervals

- **What is some practical advice about confidence intervals?**
  - Include a 95% confidence interval for every effect size you calculate
  - Use bias-corrected and accelerated bootstrap CI's with 2000+ resamples
  - Report results as "The sample mean was 100 units, 95% CI: [93.7, 106.3]."
  - Be careful to interpret confidence intervals correctly
    - **WRONG**: "This interval has a 95% chance of containing the parameter value."
    - **CORRECT**: "The interval is a set of values that are plausible for the parameter value."
    - **CORRECT**: "We are 95% confident that the interval contains the parameter value."

Effect Sizes and Confidence Intervals

```
% Load data from CSV file
>> dat = csvread('Dimensional-Data.csv');

% Calculate d using pooled SD
>> es = mDES(dat(:, 1), dat(:, 2), 'pool')
d_pool = 0.6311

% Calculate 95% bootstrap CI
>> ci = bootci(2000, @(x, y) mDES(x, y, 'pool'), dat(:, 1), dat(:, 2))
Lower Bound = 0.4272
Upper Bound = 1.2891
```

Effect Sizes and Confidence Intervals

```
% Calculate CI for reliability estimate
>> dat = csvread('Categorical-Data.csv');
>> ci = bootci(2000, @mAGREE, dat)
Lower Bound = 0.5500
Upper Bound = 0.9000

% Calculate CI for criterion validity estimate
>> dat = csvread('Binary-Data.csv');
>> ci = bootci(2000, @(x) mBINARY(x, 'F1S'), dat)
Lower Bound = 0.3636
Upper Bound = 0.9412
```

Effect Sizes and Confidence Intervals

- **Where can I read more?**
  - Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574.
  - Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *The American Psychologist*, *60*(2), 170–180.
  - Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.
- **Where can I find those functions?**
  - http://mreliability.jmgirard.com (MATLAB)
  - https://cran.r-project.org/package=bootES (R)

Estimation and Uncertainty

Part 2: Generalized Linear Modeling

Generalized Linear Modeling

- **When would we need to go beyond effect size measures?**
  - More complex models are required to analyze complex relationships
  - There are several complex modeling techniques with broad applicability
    - General linear modeling (LM) quantifies relationships for normal outcomes
    - Generalized linear modeling (GLM) is an extension of LM for non-normal outcomes
    - Popular techniques (ANOVA and regression) are special cases of LM and GLM
    - There are also other advanced techniques for specific needs and uses (MLM, SEM)
  - Linear modeling is highly interpretable and many relationships are linear

Generalized Linear Modeling

- **How does linear modeling work?**
  - The basis of LM is the simple regression formula

$$y = \alpha + \beta X + \varepsilon$$

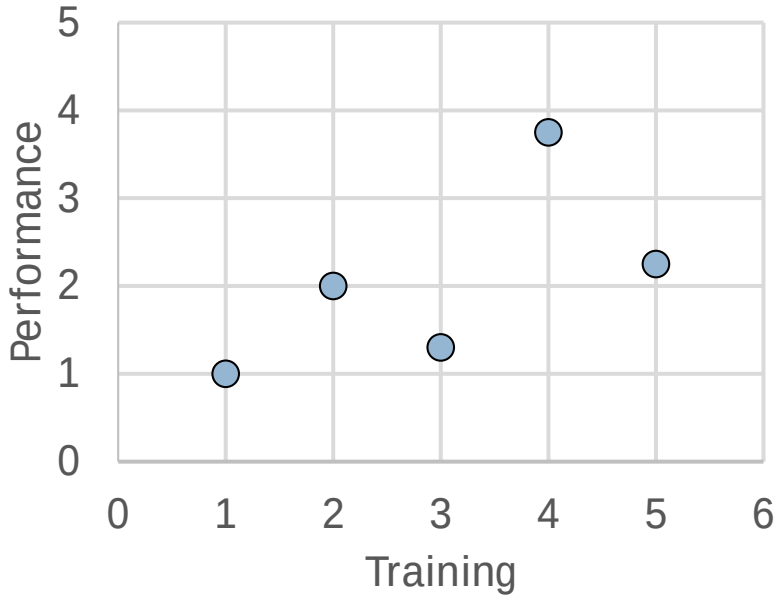
    - y is the outcome or predicted variable
    - α is the intercept or value of y when all X variables equal 0
    - β is the regression coefficient or weight
    - X is the predictor variable
    - ε is the residual (i.e., unexplained variance in y)

Generalized Linear Modeling

	Performance	Training
1	1.00	1.00
2	2.00	2.00
3	1.30	3.00
4	3.75	4.00
5	2.25	5.00

$\bar{Y} = 2.06$ $\bar{X} = 3.00$

$s_Y = 1.072$ $s_X = 1.581$



$r_{XY} = 0.627$

$$\beta = \frac{r_{XY} s_y}{s_x} = \frac{(0.627)(1.072)}{1.581} = 0.425$$

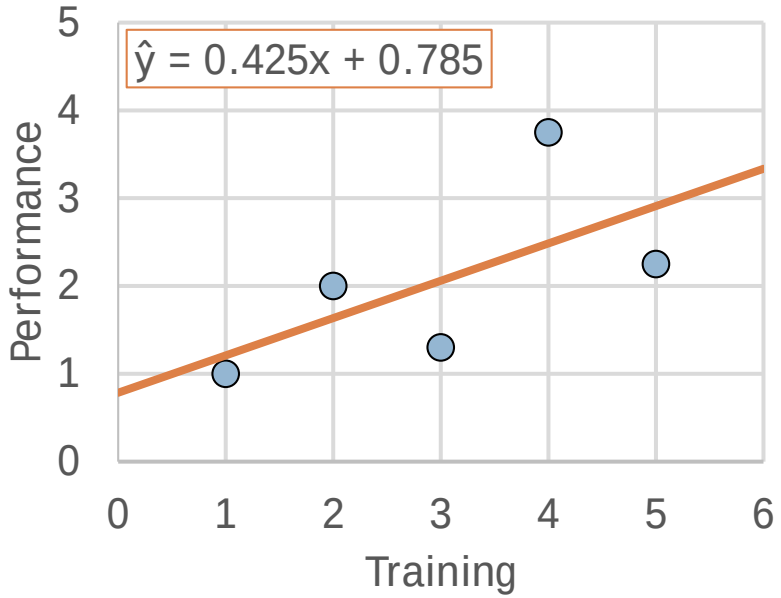
$$\alpha = \bar{Y} - \beta\bar{X} = 2.06 - (0.425)(3.00) = 0.785$$

Generalized Linear Modeling

	Performance	Training
1	1.00	1.00
2	2.00	2.00
3	1.30	3.00
4	3.75	4.00
5	2.25	5.00

$\bar{Y} = 2.06$ $\bar{X} = 3.00$

$s_Y = 1.072$ $s_X = 1.581$



$r_{XY} = 0.627$

$$\beta = \frac{r_{XY} s_y}{s_x} = \frac{(0.627)(1.072)}{1.581} = 0.425$$

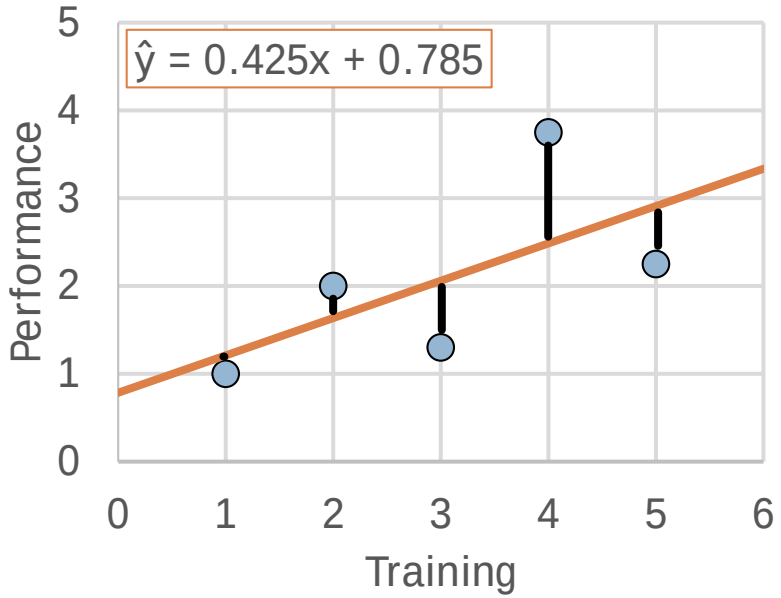
$$\alpha = \bar{Y} - \beta\bar{X} = 2.06 - (0.425)(3.00) = 0.785$$

Generalized Linear Modeling

	Performance	Training
1	1.00	1.00
2	2.00	2.00
3	1.30	3.00
4	3.75	4.00
5	2.25	5.00

$\bar{Y} = 2.06$ $\bar{X} = 3.00$

$s_Y = 1.072$ $s_X = 1.581$



$r_{XY} = 0.627$

$$\beta = \frac{r_{XY} s_y}{s_x} = \frac{(0.627)(1.072)}{1.581} = 0.425$$

$$\alpha = \bar{Y} - \beta\bar{X} = 2.06 - (0.425)(3.00) = 0.785$$

Generalized Linear Modeling

```
# Create vector variables for X and Y
> training <- c(1, 2, 3, 4, 5)
> performance <- c(1.00, 2.00, 1.30, 3.75, 2.25)

# Merge vectors into a data frame
> database <- data.frame(training, performance)

# Regress performance on training in data frame
> fit <- lm(performance ~ training, data=database)

# Display regression summary and confidence intervals
> summary(fit)
> confint(fit)
```

Generalized Linear Modeling

Residuals:

1	2	3	4	5
-0.210	0.365	-0.760	1.265	-0.660

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.785	1.012	0.776	0.494
training	0.425	0.305	1.393	0.258

Residual standard error: 0.965 on 3 degrees of freedom
Multiple R-squared: 0.393,Adjusted R-squared: 0.190
F-statistic: 1.942 on 1 and 3 DF, p-value: 0.258

	2.5 %	97.5 %
(Intercept)	-2.434	4.004
Training	-0.545	1.395

Generalized Linear Modeling

- **How can this be extended to include many predictor variables?**
  - The true power of LM is realized when many X variables are included
  - The intercepts and coefficients are labeled β_i starting at $i = 0$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j + \varepsilon$$

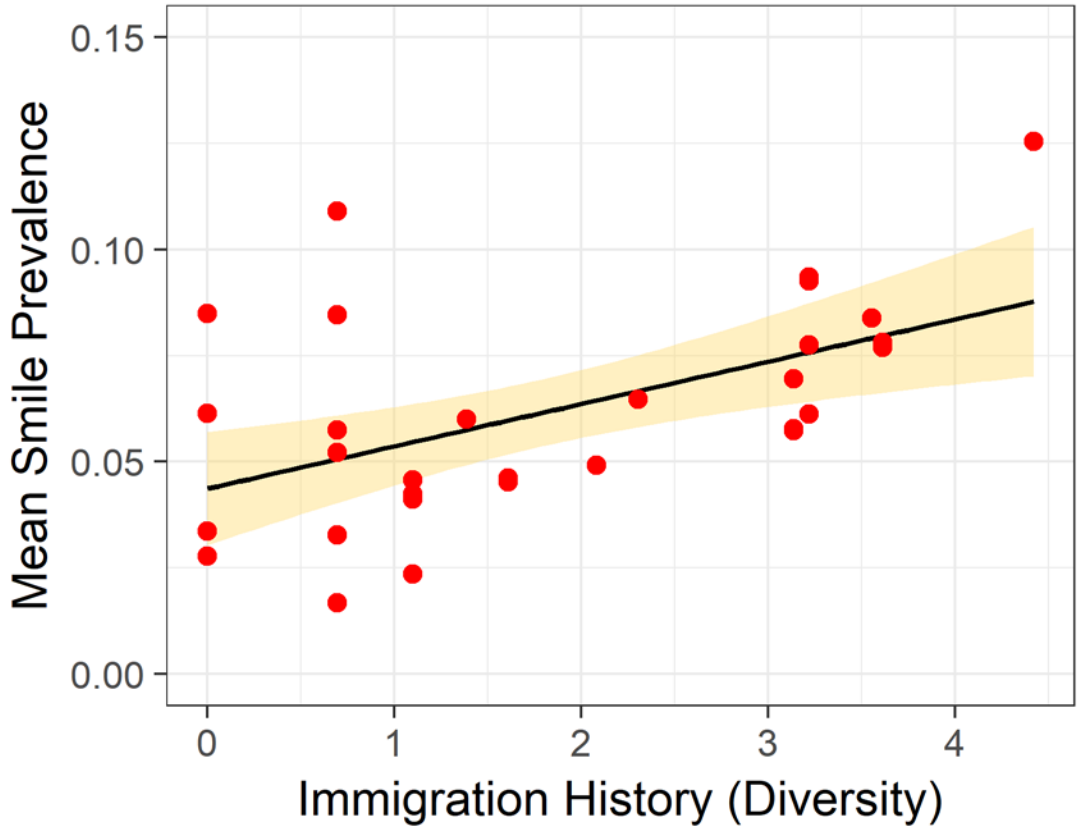
  - LM accounts for the fact that predictor variables may share variance
  - This changes the value and interpretation of regression coefficients
  - A regression coefficient (β_i) becomes the unique contribution of X_i
  - This is often called the effect of X_i "controlling for" all other X variables

Generalized Linear Modeling

```
% Read in data from CSV file
>> t = readtable('Country-Data.csv')

% Regress observed smiling on historical diversity
>> lm1 = fitlm(t, 'obsSmi ~ hisHet')
Estimated Coefficients:
                    Estimate       SE       tStat      pValue
    (Intercept)      4.972        0.455     10.927    8.509e-12
    hisHet           0.087        0.021     4.242     0.00021
Number of observations: 31, Error degrees of freedom: 29
Root Mean Squared Error: 2
R-squared: 0.383,  Adjusted R-Squared 0.362
F-statistic vs. constant model: 18, p-value = 0.00021
```

Generalized Linear Modeling



Generalized Linear Modeling

```
% Read in data from CSV file
>> t = readtable('Country-Data.csv')

% Regress observed smiling on individualism and historical diversity
>> lm1 = fitlm(t, 'obsSmi ~ indiv + hisHet')
```

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	4.565	0.648	7.037	1.1797e-07
hisHet	0.077	0.023	3.356	0.002
indiv	0.017	0.020	0.881	0.385

Number of observations: 31, Error degrees of freedom: 28
Root Mean Squared Error: 2
R-squared: 0.4, Adjusted R-Squared 0.357
F-statistic vs. constant model: 9.32, p-value = 0.001

Generalized Linear Modeling

- **What are the assumptions of LM?**
  - The relationship between y and the X variables is linear
  - The X variables are not redundant (i.e., too highly correlated)
  - The residuals are not correlated with the X variables
  - The residuals have equal variance across levels of X
  - The residuals are normally distributed with a mean of 0

Generalized Linear Modeling

- **How can the assumptions of LM be tested?**
  - Assumptions can be tested using the gvlma() function in R
  - There are statistical and graphical tests of the assumptions
- **What happens if the assumptions are violated?**
  - Estimates can become biased when some assumptions are violated
  - Some violations can be addressed through data transformations
  - Some violations can be addressed by moving from LM to GLM

Generalized Linear Modeling

- **How does GLM work?**
  - GLM works the same way as LM but allows for non-normal residuals
  - Link function transform outcomes and residuals to normal distributions

Distribution	Support	Link Name	Link Function
Normal	Real: $(-\infty, \infty)$	Identity	μ
Exponential	Real: $(0, \infty)$	Inverse	μ^{-1}
Poisson	Integer: $0,1,2, \ldots$	Log	$\ln(\mu)$
Logistic	Integer: $\{0, 1\}$ Integer: $[0, N]$	Logit	$\ln\left(\frac{\mu}{1-\mu}\right)$

Generalized Linear Modeling

- Predict frame-level agreement from properties of frame
- Are we more likely to misclassify an image with higher head pose?

$$Agree = \beta_0 + \beta_1(|Pitch|) + \beta_2(|Yaw|) + \beta_3(|Roll|) + \varepsilon$$

- However, agreement is dichotomous and so residuals are not normal
- For binomially distributed data, we can use the logit link function

$$logit(Agree) = \beta_0 + \beta_1(|Pitch|) + \beta_2(|Yaw|) + \beta_3(|Roll|) + \varepsilon$$

- Functions: in R use glm() and in MATLAB use fitglm()

Generalized Linear Modeling

```
% Import data from CSV
> library(readr)
> database <- read_csv('Pose-Data.csv')
% Regress agreement on head pose using LM
> fit <- lm(Agree ~ abs(Pitch) + abs(Yaw) + abs(Roll), data = database)
% Check regression assumptions using GVLMA
> gvlma(fit)
```

	Value	p-value	Decision
Global Stat	1.283e+06	0.0000	Assumptions NOT satisfied!
Skewness	4.885e+05	0.0000	Assumptions NOT satisfied!
Kurtosis	7.875e+05	0.0000	Assumptions NOT satisfied!
Link Function	5.693e-0 2	0.8114	Assumptions acceptable.
Heteroscedasticity	7.285e+03	0.0000	Assumptions NOT satisfied!

Generalized Linear Modeling

```
> gfit <- glm(agree ~ abs(pitch) + abs(yaw) + abs(roll), data = database, family = binomial())
> summary(gfit)
```

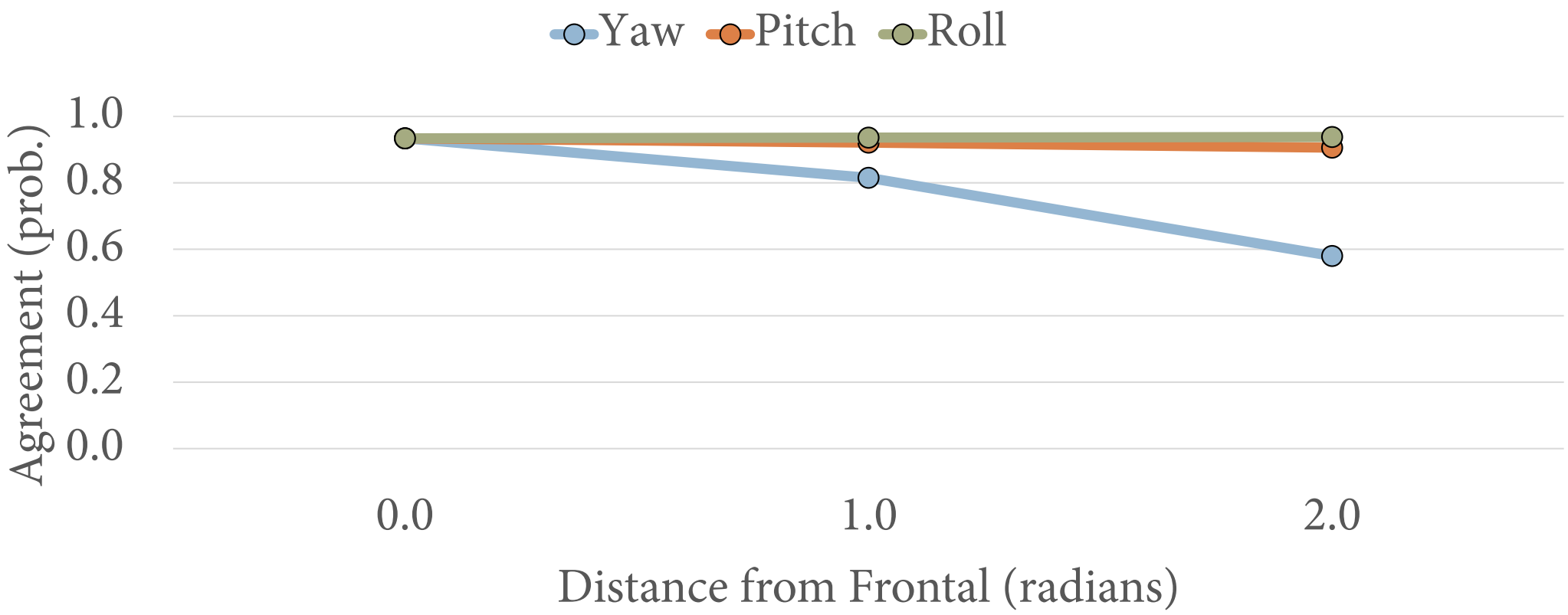
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.332	0.381	0.395	0.410	0.573

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.649	0.016	157.619	< 2e-16 ***
abs(Yaw)	-1.164	0.077	-14.961	< 2e-16 ***
abs(Pitch)	-0.192	0.064	-2.973	0.00295 **
abs(Roll)	0.031	0.079	0.398	0.69071

Generalized Linear Modeling



Generalized Linear Modeling

- **Where can I read more?**
  - Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York, NY: Routledge.
  - Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage Publications.
  - Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Estimation and Uncertainty

Part 3: Preview of Advanced Topics

Preview of Advanced Topics

- **How do I deal with nonlinear relationships within GLM?**
  - Polynomial regression
- **How do I deal with nested data and correlated residuals?**
  - Multilevel modeling
- **How do I deal with missing data?**
  - Missing data analysis
- **How do I deal with latent variables and multiple outcomes?**
  - Structural equation modeling

Questions and Consultation

Feel free to contact me in the future as well at j.girard@pitt.edu