

Reliability & Statistical Power: Evolving trends & practices

Bryan Riemann, PhD, ATC, FNATA
Armstrong State University



National Strength and Conditioning Association
Master of Science in Sports Medicine Program



Research & Statistics

How to Make a Monster (1958)



Research

Statistics

Disclaimers

- I am **NOT** a statistician!!!



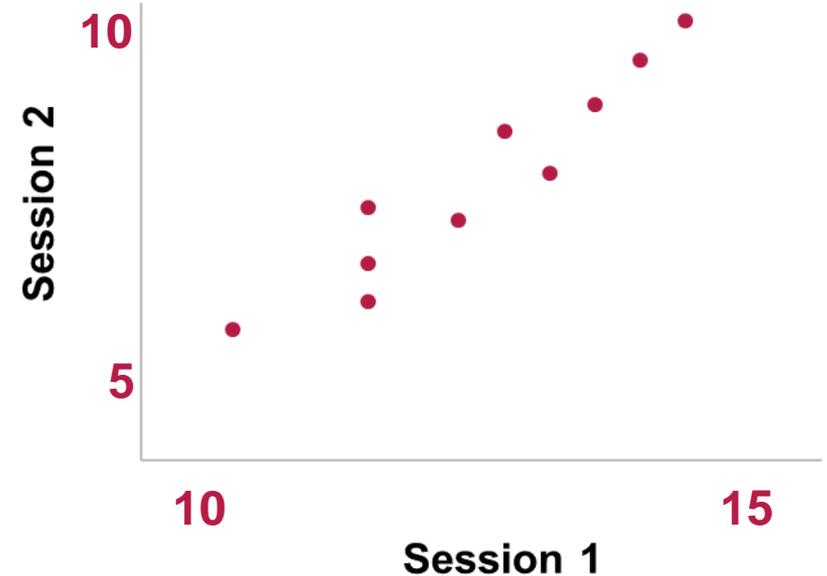
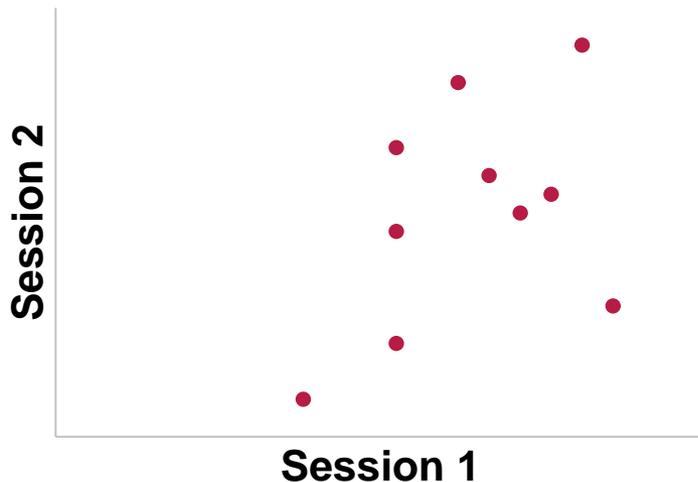
Session Overview

- **Reliability**
 - Absolute reliability
 - MDD
 - Heteroscedasticity
- **MID**
 - Establishment, interpretation & limitations
- **Power planning**
 - What can go wrong



Reliability

- **Three perspectives:**
 - Relative
 - Systematic bias
 - Absolute



Absolute Reliability

- **Most pertinent form for clinicians**
 - **Estimate of expected error from true or test-retest**
 - **Expressed in original units or proportion percentage of measurement values**
- **Common forms:**
 - **Standard error of measurement**
 - **Minimal detectable difference**
 - **Coefficient of variation**

Standard Error of Measurement

- **Often reported with ICC**

- Most common form:

$$SEM = SD\sqrt{1 - ICC}$$

- Thus, subject to same ICC factors

- (model, variability of scores)

- **Alternate form (Weir, 2005):**

$$SEM = \sqrt{MS_{Error}}$$

- **May be too small**

- Covers only ~52% of test-retest differences not 68% (1 SD) of true score error

Atkinson and Nevill, 2000

Minimal Detectable Difference

- **Extension of SEM**
- **AKA: MDC or SDC**

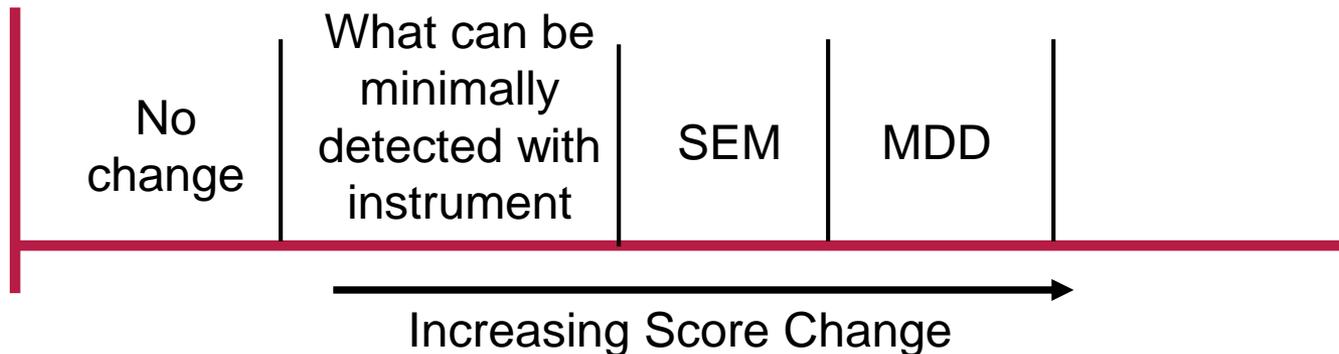
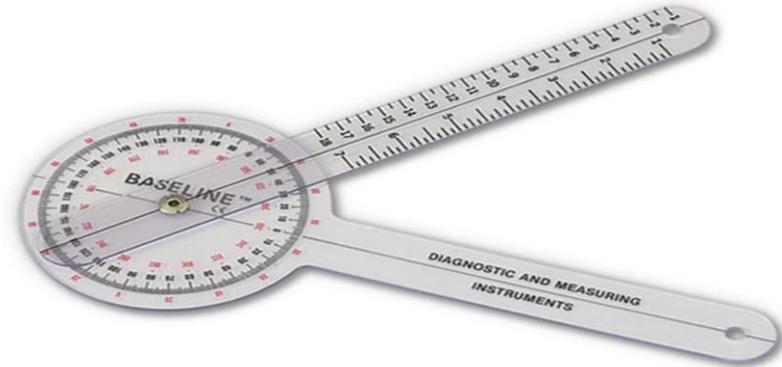
$$MDD_{90\%} = SEM * 2.33$$

↑
($\sqrt{2} * 1.65$)

$$MDD_{95\%} = SEM * 2.77$$

↑
($\sqrt{2} * 1.96$)

Minimal Detectable Difference



- **Changes outside boundaries considered real**
 - Large MDD can mislead conclusion that no “real” change despite patient appreciating change

Coefficient of Variation

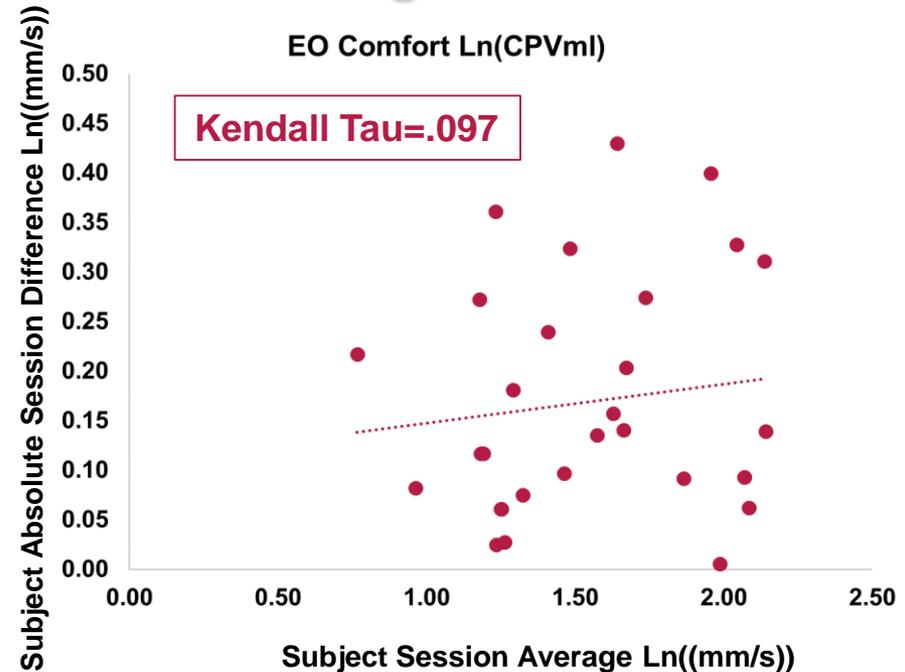
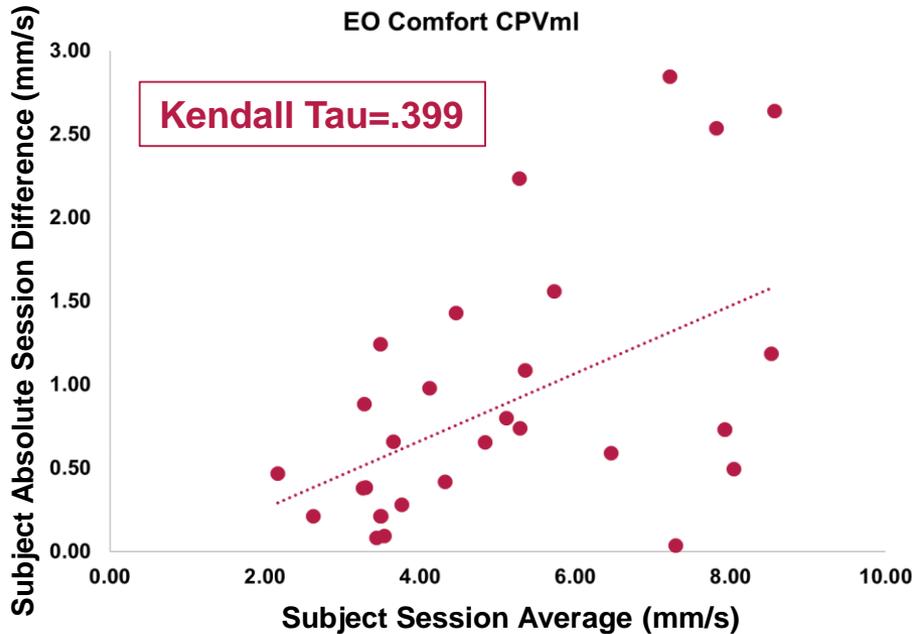
$$CV = \frac{s}{\bar{X}} * 100$$

- **Error as % of individuals mean score**
 - **Becomes unitless; enhances comparisons to other studies/measures**
 - **Can not be used when:**
 - **Scale includes negative values**
 - **Mean is/close to zero**
 - **Most useful with heteroscedasticity**

Heteroscedasticity

- **When variance meets magnitude**
 - **Error differs systematically between participants**
 - **Occurs with ratio scale measurements** (Nevill & Atkinson, 1997)
 - **Assumption with raw values is consistent error across participants**
 - **Violation: those with greater error will influence statistic**
- **Visualized with average vs. difference plot**
 - **Confirmed with Kendall's tau**

Heteroscedasticity



- **Common remedy: Natural log transformation**
 - Stabilizes variance, normalizes distribution
 - Can multiply by 100 to maintain consistent precision
- **Conduct reliability analysis on transformed**

Heteroscedasticity

- Recommendations when reviewing:

No mention reliability statistics were conducted on Ln transformed data in caption

Table 1 Descriptive statistics and reliability results between the two testing sessions ($1.9 \pm .7$ day separation) in thirty older adults

| | Session 1 | Session 2 | ICC | Systematic bias | | SEM (%) |
|---------------------------|-------------------------|-------------------------|-----|----------------------|---------|---------|
| | $\bar{X} \pm SD$ (mm/s) | $\bar{X} \pm SD$ (mm/s) | | \bar{X} change (%) | P value | |
| Self-selected eyes open | 4.9 ± 1.8 | 5.1 ± 2.1 | .86 | 3.1 | .427 | 15.9 |
| Self-selected eyes closed | 6.9 ± 3.7 | 6.5 ± 3.4 | .82 | -4.4 | .414 | 23.6 |
| Narrow-eyes open | 6.2 ± 2.7 | 6.7 ± 2.6 | .74 | 7.6 | .192 | 23.3 |
| Narrow-eyes closed | 9.1 ± 3.9 | 10.0 ± 5.1 | .81 | 6.8 | .200 | 21.2 |

\bar{X} mean, *SD* standard deviation, *ICC* intraclass correlation coefficient, *SEM* standard error of measurement

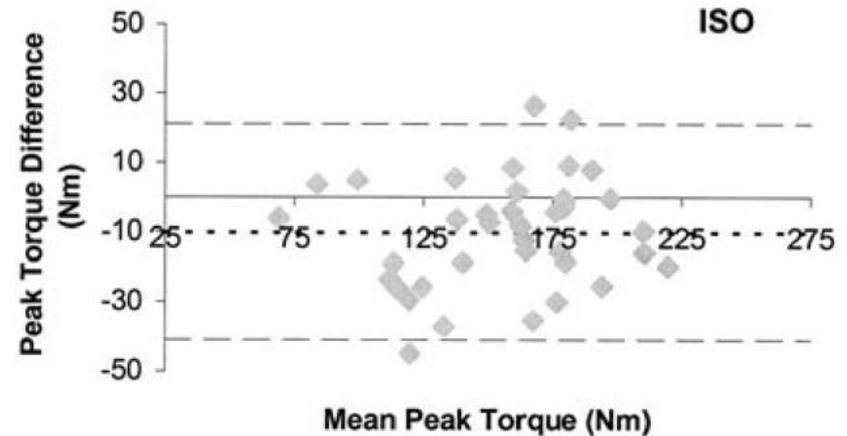
Having original units here OK but confusing since used Ln transformed data

Clue here: units are in percentages

Reliability of a Single-Session Isokinetic and Isometric Strength Measurement Protocol in Older Men

T. Brock Symons,¹ Anthony A. Vandervoort,^{1,2} Charles L. Rice,¹
 Tom J. Overend,² and Greg D. Marsh¹

Furthermore, we generated Bland-Altman plots (15) of the difference between test day 1 (TD1) and test day 2 (TD2) versus the mean of TD1 and TD2 for each participant using the raw data scores for all measures (Figures 1 and 2). We determined heteroscedasticity (nonuniform scatter) by visual inspection of the Bland-Altman plots and we deemed them to be present when the difference scores for participants at one end of the plot demonstrated a tendency for larger values (2).



| Measure | Average Torque | | |
|---|----------------|------|------|
| | CON | ISO | ECC |
| Intraclass correlation coefficient (ICC) | 0.90 | 0.92 | 0.93 |
| Lower confidence limit | 0.82 | 0.85 | 0.87 |
| Upper confidence limit | 0.95 | 0.96 | 0.96 |
| Typical error as a coefficient of variation (CV _{TE}) (%) | 10.92 | 7.71 | 7.65 |
| Lower confidence limit | 8.90 | 6.29 | 6.24 |
| Upper confidence limit | 14.13 | 9.97 | 9.90 |

Coefficient of Variation



Heteroscedasticity & Coefficient of Variation

- **Hopkins (2000): computing CV when data are log transformed**

$$S_{TE} = \frac{SD_{Diff\ between\ scores}}{\sqrt{2}} \qquad CV_{TE} = 100 * \left(e^{\frac{S_{TE}}{100}} - 1 \right)$$

- **×/÷ raw scores by CV expressed as factor to create error boundaries**
 - **Allows for measurement error to be scaled to magnitude**

Example: $CV_{TE} = 15\%$
Lower bound = $1/1.15 * \text{score}$
Upper bound = $1.15 * \text{score}$

Reviewing Recommendations

- **When reliability statistics reported:**
 - **Ensure all three perspectives provided**
 - **Make sure study population is clearly defined**
 - **Point estimates for population based on sample studied**
 - **Providing confidence intervals helps clinical utility**
 - **Potent factor: sample size**
- **SEM: make sure formula is defined**
 - **Understand the limitations of the ICC approach**

Reviewing Recommendations

- **Intervention studies with MDD reported**
 - Report specific MDD (90%, 95%) used
 - Include proportion of participants who exceed MDD



Reviewing Recommendations

- **Look for mention of heteroscedasticity checking**
 - **How was it handled- transformation?**
 - **Not sure helpfulness of including Bland-Altman plots**
 - **Probably more space efficient to summarize coefficients**
 - **What data were reliability analyses conducted upon?**
 - **If transformed, do they report percentages or ratios?**

Session Overview

- **Reliability**
 - Absolute reliability
 - MDD
 - Heteroscedasticity
- **MID**
 - Establishment, interpretation & limitations
- **Power planning**
 - What can go wrong



Minimal Important Difference

- **Interpreting change/differences needs to consider statistical significance & clinical meaningfulness**
 - **Dependent upon perspective**
 - **What might be important for one patient group may not be the same for another**

Jaeschke et al (1989): Minimal clinically important difference

“the smallest difference in score in the domain of interest which patients perceive as beneficial”

Minimal Important Difference

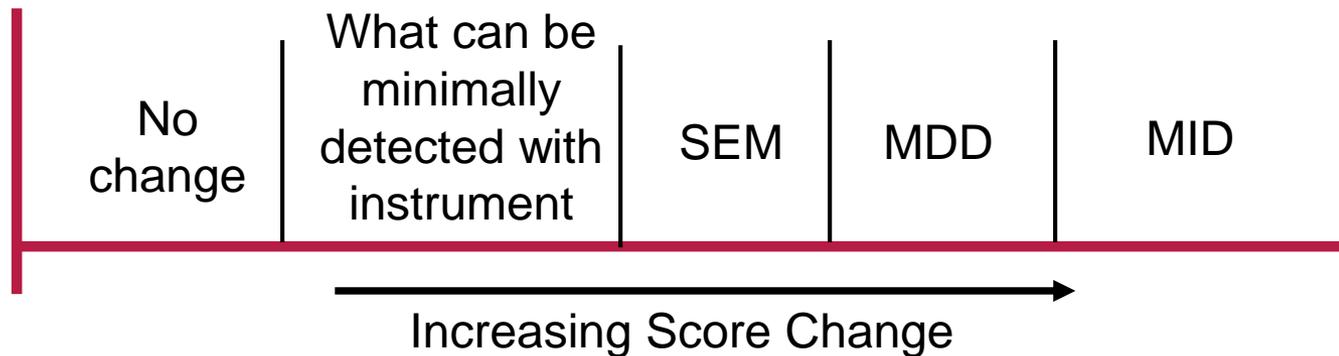
- **MCID transitioned to MID- more emphasis on patient's perspective**
 - Several other terms used, each with slightly different definition/derivation
- **Simply: threshold of change beyond MDD that patient perceives as meaningful & would elect to repeat intervention**
- **Majority literature considers MID for PRO but increasing trend for other commonly used measures**

Minimal Important Difference

- **Two approaches to establishing:**
 - **Anchor based (patient perception)**
 - Global Assessment Ratings
 - **Distribution based**
 - Statistical characteristics of data: effect sizes, absolute reliability, systematic bias testing
- **Strength/weaknesses of both, but little universal consensus**

Minimal Important Difference

- Typically expect $MID > MDD$



- **Some exceptions where $MID < MDD$:**
 - QuickDASH & ASES
 - Some question instrument utility, others suggest MID more impt
 - Highlights incongruence between distribution & anchor approaches

Minimal Important Difference

- **MID values have limited generalizability**
 - **Dependent upon:**
 - Patients used
 - MID methodology
- **Thus, need caution when using**
 - **Clinically- give priority to treatments that exceed MID**

MID Example

- **3 recreational tennis players with shoulder impingement without instability**
 - **Complete QuickDASH at initial & 3 week f/u**

QuickDASH

Please rate your ability to do the following activities in the last week by circling the number below the appropriate response.

| | NO DIFFICULTY | MILD DIFFICULTY | MODERATE DIFFICULTY | SEVERE DIFFICULTY | UNABLE |
|---|---------------|-----------------|---------------------|-------------------|--------|
| 1. Open a tight or new jar. | 1 | 2 | 3 | 4 | 5 |
| 2. Do heavy household chores (e.g., wash walls, floors). | 1 | 2 | 3 | 4 | 5 |
| 3. Carry a shopping bag or briefcase. | 1 | 2 | 3 | 4 | 5 |
| 4. Wash your back. | 1 | 2 | 3 | 4 | 5 |
| 5. Use a knife to cut food. | 1 | 2 | 3 | 4 | 5 |
| 6. Recreational activities in which you take some force or impact through your arm, shoulder or hand (e.g., golf, hammering, tennis, etc.). | 1 | 2 | 3 | 4 | 5 |

| | NOT AT ALL | SLIGHTLY | MODERATELY | QUITE A BIT | EXTREMELY |
|--|------------|----------|------------|-------------|-----------|
| 7. During the past week, to what extent has your arm, shoulder or hand problem interfered with your normal social activities with family, friends, neighbours or groups? | 1 | 2 | 3 | 4 | 5 |

| | NOT LIMITED AT ALL | SLIGHTLY LIMITED | MODERATELY LIMITED | VERY LIMITED | UNABLE |
|---|--------------------|------------------|--------------------|--------------|--------|
| 8. During the past week, were you limited in your work or other regular daily activities as a result of your arm, shoulder or hand problem? | 1 | 2 | 3 | 4 | 5 |

Please rate the severity of the following symptoms in the last week. (circle number)

| | NONE | MILD | MODERATE | SEVERE | EXTREME |
|--|------|------|----------|--------|---------|
| 9. Arm, shoulder or hand pain. | 1 | 2 | 3 | 4 | 5 |
| 10. Tingling (pins and needles) in your arm, shoulder or hand. | 1 | 2 | 3 | 4 | 5 |

| | NO DIFFICULTY | MILD DIFFICULTY | MODERATE DIFFICULTY | SEVERE DIFFICULTY | SO MUCH DIFFICULTY THAT I CAN'T SLEEP |
|--|---------------|-----------------|---------------------|-------------------|---------------------------------------|
| 11. During the past week, how much difficulty have you had sleeping because of the pain in your arm, shoulder or hand? (circle number) | 1 | 2 | 3 | 4 | 5 |

MID Example

- Mintken et al (2009)
 - ICC (2,1) across 14d: .90
 - SEM: 4.8pts
 - $MDD_{90\%}$: 11.2pts (4.8 * 2.33)
 - MID: 8pts

| Patient # | Change between initial & follow up | Interpretation |
|-----------|------------------------------------|---|
| 1 | 4↓ | Neither beyond measurement error or clinical meaningfulness |
| 2 | 15↓ | Exceeds both measurement error and clinical meaningfulness |
| 3 | 9↓ | Suggests clinical meaningfulness in perception of change but change does not exceed measurement error |

Reviewing Recommendations

- **Intervention studies with MID reported**
 - **Report specific method used establish MID**
 - **Make sure patient population is clearly defined**
 - **Include proportion of participants who meet or exceed MID**
 - **Helps estimate likelihood that patients will respond favorably to similar treatment**



Session Overview

- **Reliability**
 - Absolute reliability
 - MDD
 - Heteroscedasticity
- **MID**
 - Establishment, interpretation & limitations
- **Power planning**
 - What can go wrong



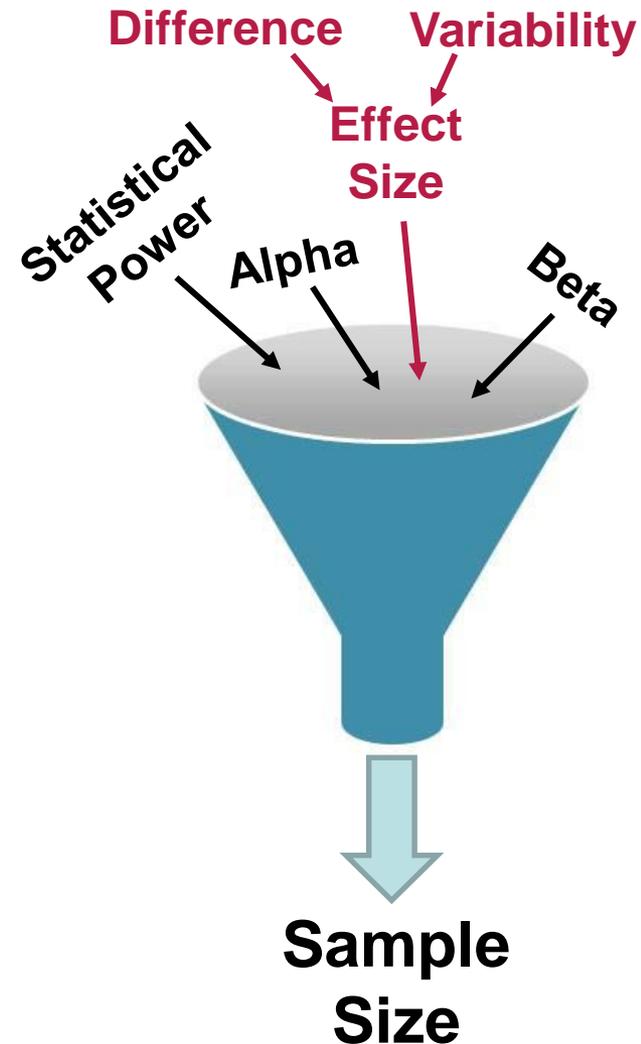
Power Planning

- **Probability of statistically detecting a treatment effect when one truly exists**
- **Influenced by:**
 - **Alpha**
 - **Beta**
 - **Effect size**
 - **Difference & variability**
 - **Sample size**



Power Planning

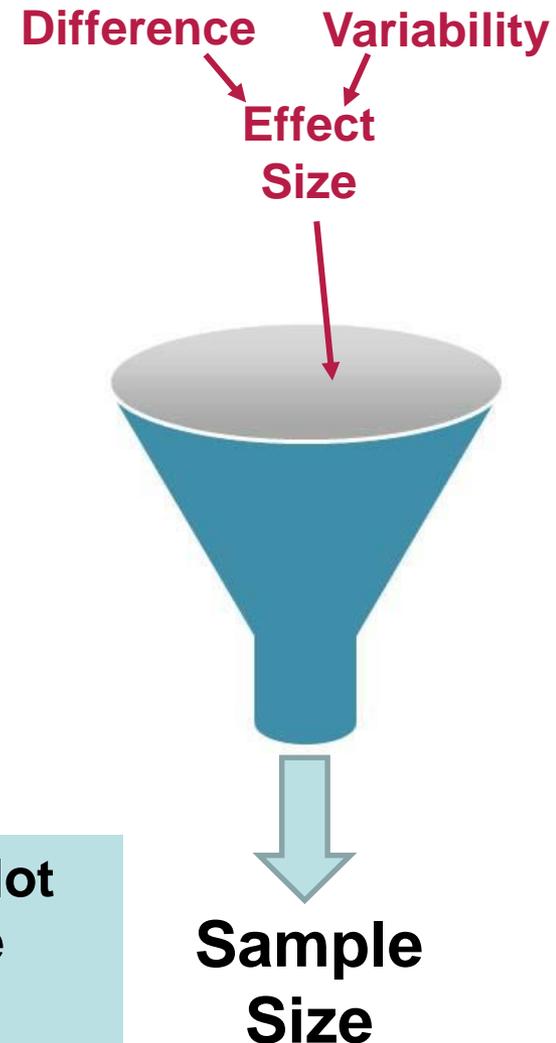
- Solving for sample size, provides basis for initial study design
- Challenge estimating:
 - Difference:
 - Balance between prompting big changes & clinical practicality
 - Variability:
 - Inclusion/exclusion criteria
 - Measurement reliability



Power Planning

- **Where do estimates come from?**
 - **Difference:**
 - What a clinically relevant change would be
 - Previous research/Pilot work
 - **Variability:**
 - Previous research/pilot work: must match target sample characteristics
 - **Occasionally:** effect size chosen without regard to clinical relevance or patient variability

Need to realize: when using previous research/pilot work for differences & variability, they are sample estimates & subject to sampling fluctuations



Reviewing Recommendations

- **Study with non-significant findings: is it truly no effect or type II error?**
 - **Personal bias: those that have non-significant findings but well designed/executed should still be published**
- **Studies with *a priori* power analysis**
 - **Scrutinize source of difference/variance estimates**
 - **Sufficient detail provided to evaluate?**
 - **Do they come from a similarly chosen population?**
 - **Do the difference/variance estimates have clinical relevance?**
 - **If no significance, could the estimates prompted too small size?**
 - **Did they add some buffer to account for sampling fluctuation?**

Session Overview

- **Reliability**
 - Absolute reliability
 - MDD
 - Heteroscedasticity
- **MID**
 - Establishment, interpretation & limitations
- **Power planning**
 - What can go wrong





*Thank
You*

