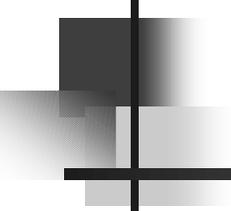
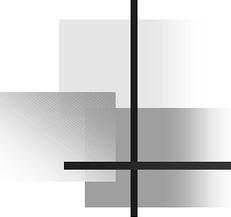


# **Statistical Hypothesis Testing**



---

**Helena Chmura Kraemer  
Stanford University**



# Recapitulation

---

- **Have a theory.**
- **Sampling, Design, measurement, treatment protocols.**
- **Need to set rule as to what evidence from the RCT would lead to recommendation for T over C.**
  - **Most common current method: Null Hypothesis Statistical Testing (NHST).**
  - **Greater Emphasis in Future (?): Effect sizes.**

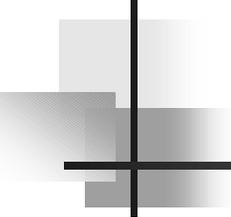
# The Overuse, Misuse, Abuse of NHST

---

- **Should NHST be “outlawed”?**
- **Signs of abuse:**
  - **Tables and text full of \*\*\*, NS, p-values**
  - **“Statistical significance” interpreted as big, important, useful (when it may be trivial)**
  - **“Non statistical significance” interpreted as “proof” of the equivalence (when it indicates poor**

# Analogy: Trial by Jury vs. NHST

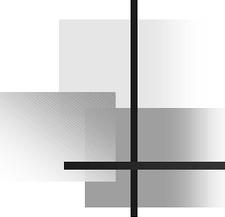
- **Trial by Jury**
  - You: The Prosecutor
  - Evaluation of Evidence: Judge/Jury
- **NHST**
  - You: The investigator
  - Evaluation of Evidence: Other scientists, reviewers, editors, readers, clinicians, policy makers, medical consumers or advocates.
- **Biostatisticians: The gadflies? The lawyers?**



# Exploratory Phase

---

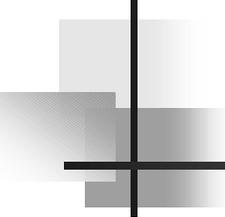
- **Gather evidence, testimony, etc. until have enough to bring charges, indict.**
- **Theory, Animal Studies, Clinical Observation, Pilot Studies, Phase I, II studies, until have rationale and justification for your theory than T is better than (different from) C.**



# Hypothesis Specification

---

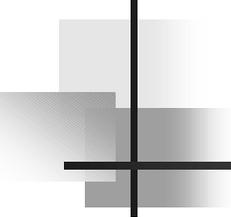
- **Charges are few and specific.**  
**Some may be dropped during the trial, but none added during the trial in response to evidence.**
- **Hypotheses are few and specific.**  
**Some may be dropped during the RCT, but none added “post hoc”.**



# Preparing for the Trial

---

- **Assemble the judge, jury to hear the evidence and render the verdict. Instructions to the jury to prevent mistrial.**
- **Design the RCT to generate evidence needed to adequately and fairly test theory. Set the rule that will support your theory “a priori”. Submit the proposal for review.**
  - **IRB**  
**Peer review group**



# Objectivity

---

- **The defendant is presumed innocent until proven beyond reasonable doubt to be guilty of stated charges.**
- **The “null hypothesis”, i.e. the denial of your theory, is presumed true until you prove beyond reasonable doubt that it is false.**
  - **“Beyond reasonable doubt” means that the probability of claiming that your theory is true when it is not (null hypothesis true) is less than an a priori set significance level (usually 5% or 1%).**

# Interpretation of the Verdict-1

---

- **“Guilty” means evidence was sufficient to prove guilt of stated charges beyond reasonable doubt.**
  - **May appeal the verdict.**
- **“Not guilty” means evidence was not sufficient to prove guilt of stated charges beyond reasonable doubt.**
  - **No double jeopardy**

# Interpretation of the Verdict-2

---

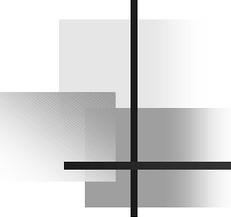
- **“Statistically significant” means evidence was sufficient to prove beyond reasonable doubt (5% or 1%) that the null hypothesis is not true, and hence provides support for your theory.**
  - **Replication and independent confirmation always required. Meta Analysis?**
  - **Does not mean “large” or “important”. It may not indicate clinical or policy significance.**
- **“Not statistically significant” means your evidence was not sufficient: inadequate power.**
  - **Learn from your mistakes!**

# **Caveats:**

## **The Burden of Proof is on You**

---

- **Don't initiate trial until preliminary evidence is strong enough.**
  - **Present the evidence competently.**
- **Don't initiate RCT without sufficient rationale and justification**
  - **Valid sampling, design, analytic procedure**
  - **Reliable and valid outcome, and few of them.**
  - **High enough power.**
  - **Stick to your own protocol!**
  - **Don't over generalize or exaggerate your results.**



# Example-1

---

**Theory:  $T > C$**

**Treatment, design and measurement protocols**

**Sample N patients**

**Randomly assign proportion P to T.  $P' = 1 - P$  to C.**

**Measure response to treatment with bias controlled.**

**Analytic plan:**

**Compare T versus C, and if response to T is sufficiently better than that to C, reject the null hypothesis (here that  $T < C$  one-tailed)**

# **Example-2: Specifically how?**

---

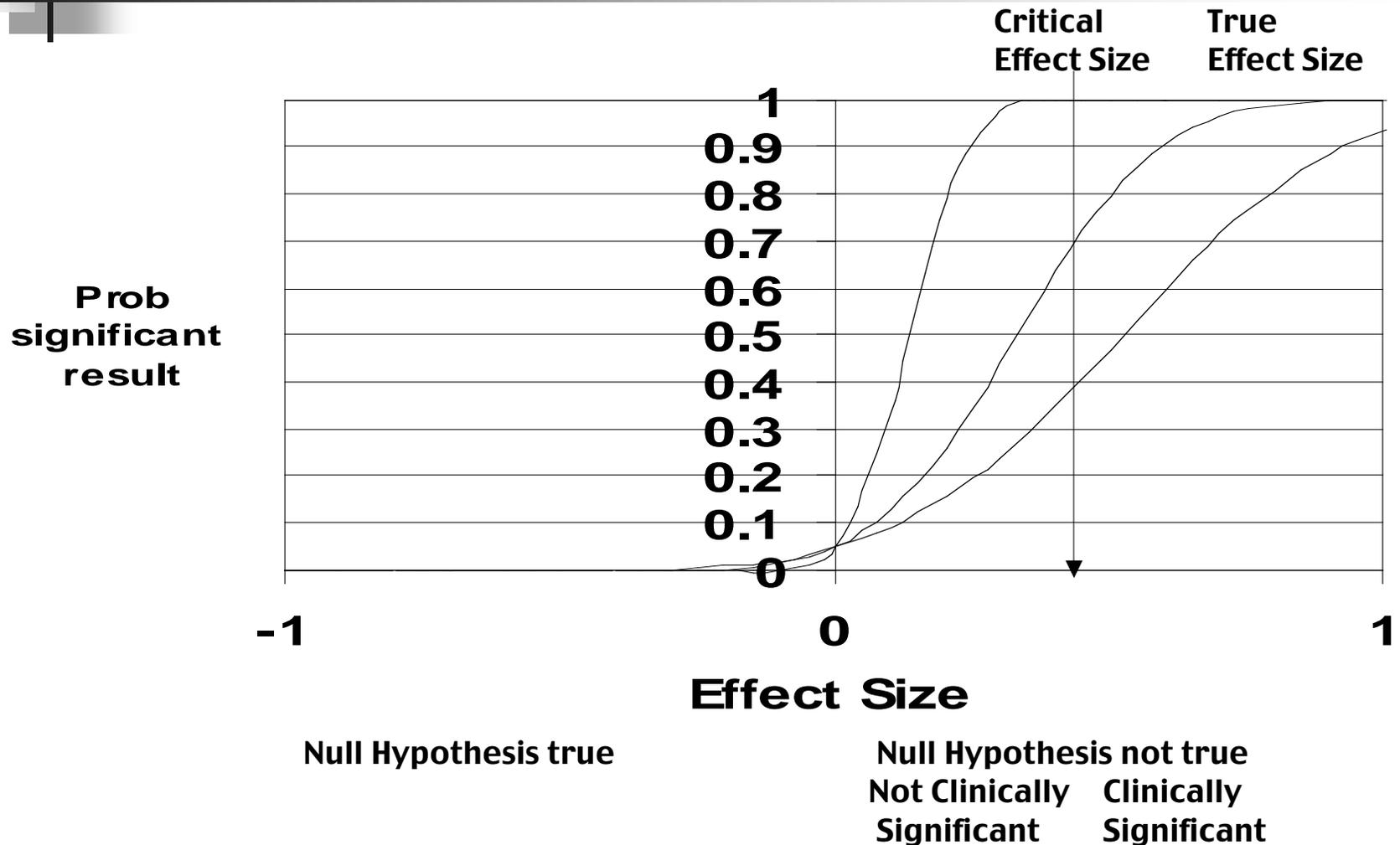
**Student's t-test: Compute t-statistic, and compute the p-value: a statistic estimating the probability of rejecting null hypothesis when the effect size is that observed. If p-value <5%, the reject null hypothesis (declare statistically significant at the 5% level).**

**Mann-Whitney Wilcoxon: Compute the test statistic and compute the p-value etc.**

**2X2 Chi Square test: Compute the success rate in T and compare with that in C. Compute the test statistic and compute the p-value, etc.**

**Correlation Test: Compute the correlation coefficient between outcome and exposure to T Compute the p value etc**

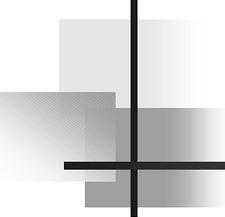
# Example-3: What is N? P? Will I have enough power?



# Example-4: Where do mistakes happen with power?

---

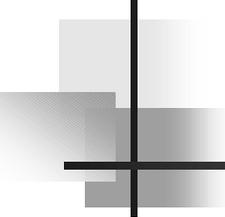
- **Critical effect size set at heart's desire rather than threshold of clinical significance.**
- **Simple miscalculation.**
  - **Proposing to do Chi-Square test, but computing power using t-test.**
  - **Making assumptions unlikely to be true.**
    - **Assuming normal distributions, equal variance when that is not true.**
    - **Assuming absence of site differences in a multi-site study.**



# The Problem of Effect Size

---

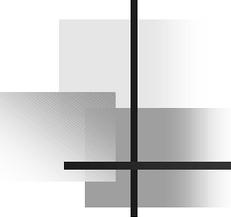
- **Common choices (Rules of Thumb):**
  - **T-test: Cohen's d, the standardized mean difference between the treatment means. (Null:  $d=0$ ; Small: .2; Medium: .5; Large: .8.)**
  - **Mann-Whitney-Wilcoxon:  $AUC = \text{Prob}(T > C) + .5\text{Prob}(T = C)$ . (Null: 50%; Small: 56%; Medium: 64%; Large: 71 %.)**
  - **2X2 Chi-Square: Odds Ratio, Risk Ratio, Risk Difference.**
    - $AUC = .5(1 + RD)$
  - **Correlation Coefficient: (Null: 0; Small: .1; Medium: .3; Large: .5.)**
- **To date, largely based on statistical, not clinical or policy considerations.**



# And when the RCT is done?

---

- **Write up the results, and celebrate!**
- **Learn from your mistakes.**
- **Formulate new hypotheses for future testing**
  - **Moderators of treatment: Factors measured at baseline that identify on whom or under what conditions the treatment works better or worse.**
    - **Why important? Selection for treatment; Inclusion/exclusion criteria, stratification for future studies.**
  - **Mediators of treatment: Events or changes during treatment that may help explain how or why the treatment works.**
    - **Why important? Suggestions for improvement of treatment efficacy or effectiveness.**



# Conclusion

---

- **If use NHST, always present effect sizes for any statistically significant result, and some measure of the accuracy of estimation.**
- **If don't use NHST, consider using effect sizes and some measure of the accuracy of estimation. Possibly Bayes' estimation?**
- **Statistical significance is necessary, but not sufficient! Ultimately the crucial issue is the benefit to the patients, i.e. clinical or policy significance.**