MARCH 19 - 23

PCRS 2025

PACIFIC WAVES - EXPLORING SCIENTIFIC FRONTIERS IN AN EVOLVING SOCIETY

INDIAN WELLS, CA

PACIFIC COAST
REPRODUCTIVE
SOCIETY

Methods and Statistics

- Micah J Hill, DO
- Fellowship Director, NIH

MARCH 19 - 23

PCRS 2025

PACIFIC WAVES - EXPLORING SCIENTIFIC FRONTIERS IN AN EVOLVING SOCIETY

INDIAN WELLS, CA

Disclosure Slide

- Neither I nor members of my immediate family have any actual or potential financial interests to disclose relating to the content of this presentation.

Needs Assessment Statement and Expected Learning Outcomes

- Describe strengths and limitations of common study designs
- Differentiate association from prediction
- Discuss how to identify and account for confounding
- Select appropriate statistical methods for various study designs

Example Oral Question

- You are asked to design a study to assess the obstetric safety of natural versus programed frozen embryo transfer
  - What is your hypothesis?
  - What is your primary outcome?
  - What are your secondary outcomes
  - What study design types could address this question?
  - What are the strengths and limitations of each design?
  - What are the key steps of conducting a clinical trial?
  - What is the basic analysis plan for this study?
  - How would your analysis plan change if you conduct a cohort study?

Hypothesis

- Good research is always hypothesis driven
- State the hypothesis (alternate hypothesis)
- Null hypothesis
- The study results should matter regardless of the direction of the findings
- Far better an approximate answer to the right question, than an exact answer to the wrong question, which can always be made with precision
  - John Tukey

Study Designs

- Experimental
  - Randomized clinical trial
- Observational
  - Cohort
    - Prospective
    - Retrospective
  - Case-control
    - Retrospective
  - Cross-sectional
    - Prevalent cases
- Descriptive
  - Case reports

Randomized Trial

1- Assemble the study population

inclusion/exclusion criteria

recruit adequate sample size (to avoid type-II error)

2- Evaluate baseline characteristics

3- Randomly assign subjects to study groups

subject blinded to intervention (single) :

diminishes error in subject evaluation /follow-up

investigator blinded to assignment and allocation sequence (double) :

diminishes selection bias

4- Apply intervention/placebo

5- Measure outcome variable

Randomized Trial

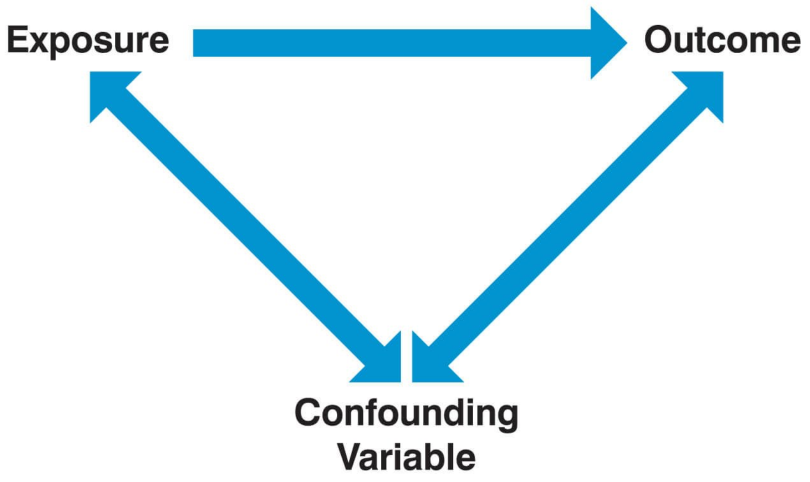
- Strengths
  - minimizes bias
  - minimizes confounding variables
  - Demonstrates causality
- Weaknesses
  - Expensive
  - Time consuming
  - Address a narrow question in a defined population

Bias and Confounding

- What is bias versus confounding?
- What are examples of biases in medical research?

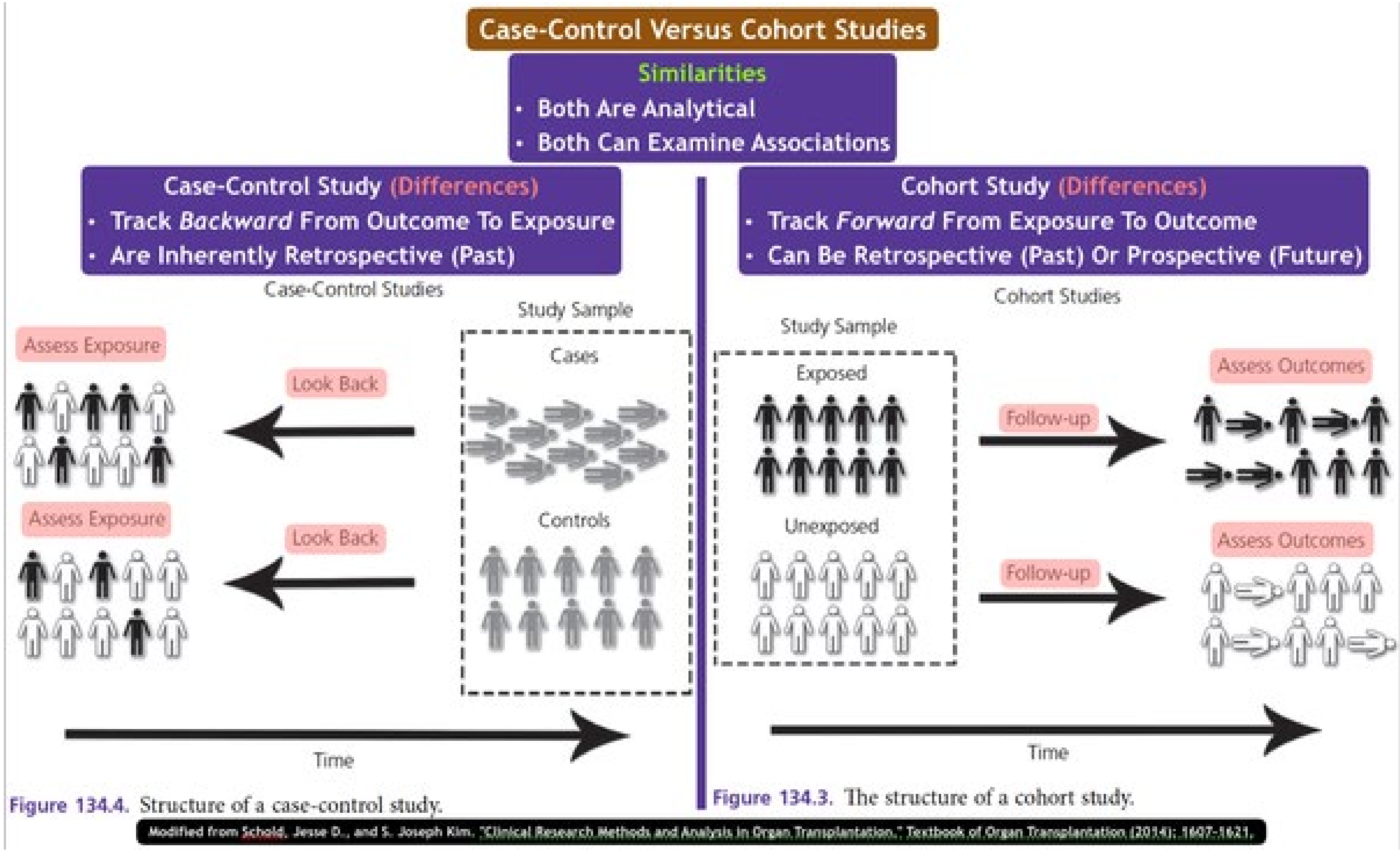
Bias and Confounding

- Bias
  - Systematic errors -> incorrect estimations of association
- Confounding
  - Inaccuracy in the estimated measure of association when exposures are mixed with other factors that are associated the outcome



Cohort versus Case-Controlled Study

- 100 patients who had a P4 over 2 on day of hCG
- 100 controls matched for age and antral follicle count with a P4 below 2
- Cases and controls are compared for live birth
- What type of study design is this?



Case-Control Versus Cohort Studies

Similarities

- Both Are Analytical
- Both Can Examine Associations

Case-Control Study (Differences)

- Track *Backward* From Outcome To Exposure
- Are Inherently Retrospective (Past)

Cohort Study (Differences)

- Track *Forward* From Exposure To Outcome
- Can Be Retrospective (Past) Or Prospective (Future)

Figure 134.4. Structure of a case-control study.

Figure 134.3. The structure of a cohort study.

Modified from Schold, Jesse D., and S. Joseph Kim. "Clinical Research Methods and Analysis in Organ Transplantation." Textbook of Organ Transplantation (2014): 1607-1621.

Cohort Studies

Observational, non-experimental, prospective or retrospective

Investigator does not manipulate intervention

Patients are assembled that have been “exposed” & compared to an unexposed control group (cohort)

These two groups are then followed longitudinally (maybe be prospective or retrospective) for outcome.

Designed to detect association, not causation

Prospective versus Retrospective Cohort

- Both level 2 evidence
- Prospective may help you collect confounding variables better
- Retrospective can be cheaper and just as good as prospective cohort studies

Case Control

Begins at the end

Good for studying diseases with low incidence

Here, a group of women with the disease (cases) are compared to a group without (controls) with respect to an *earlier exposure(s).*

Cohort versus Case-Controlled Study

Cohort

- ***Works forwards in time***
- ***Starts with exposure and looks for outcome***
- ***Eg natural versus programmed FET -> preeclampsia***

Case-control

- ***Works backwards in time***
- ***Starts with outcome and looks for exposure***
- ***Eg preeclampsia -> prevalence of natural versus programmed in those with and without pre-e***

Cohort Studies

- Strengths
  - Cheap
  - Easy to collect data
  - Data may already exist (retrospective)
- Weaknesses
  - Cannot prove causality, only association
  - Inherent bias
  - Confounding variables

Case Control Studies

- Strengths
  - Allows the study of rare diseases
  - Cheap
  - Easy to collect data
- Weaknesses
  - Cannot calculate prevalence or RR
  - Can only have a single outcome
  - Very susceptible to bias
    - Separate sampling of cases and controls
    - Retrospective measurements of predictors

Cross Sectional Studies

Observational, snap-shot in time

Measures prevalence of cases

Prevalence is the proportion of individuals w/ the disease *at a specific time*

Incidence refers to new cases that have developed over a period of time

Thus, temporal relationships cannot be established w/ cross-sectional studies

Case Report / Case Series

Observational, descriptive

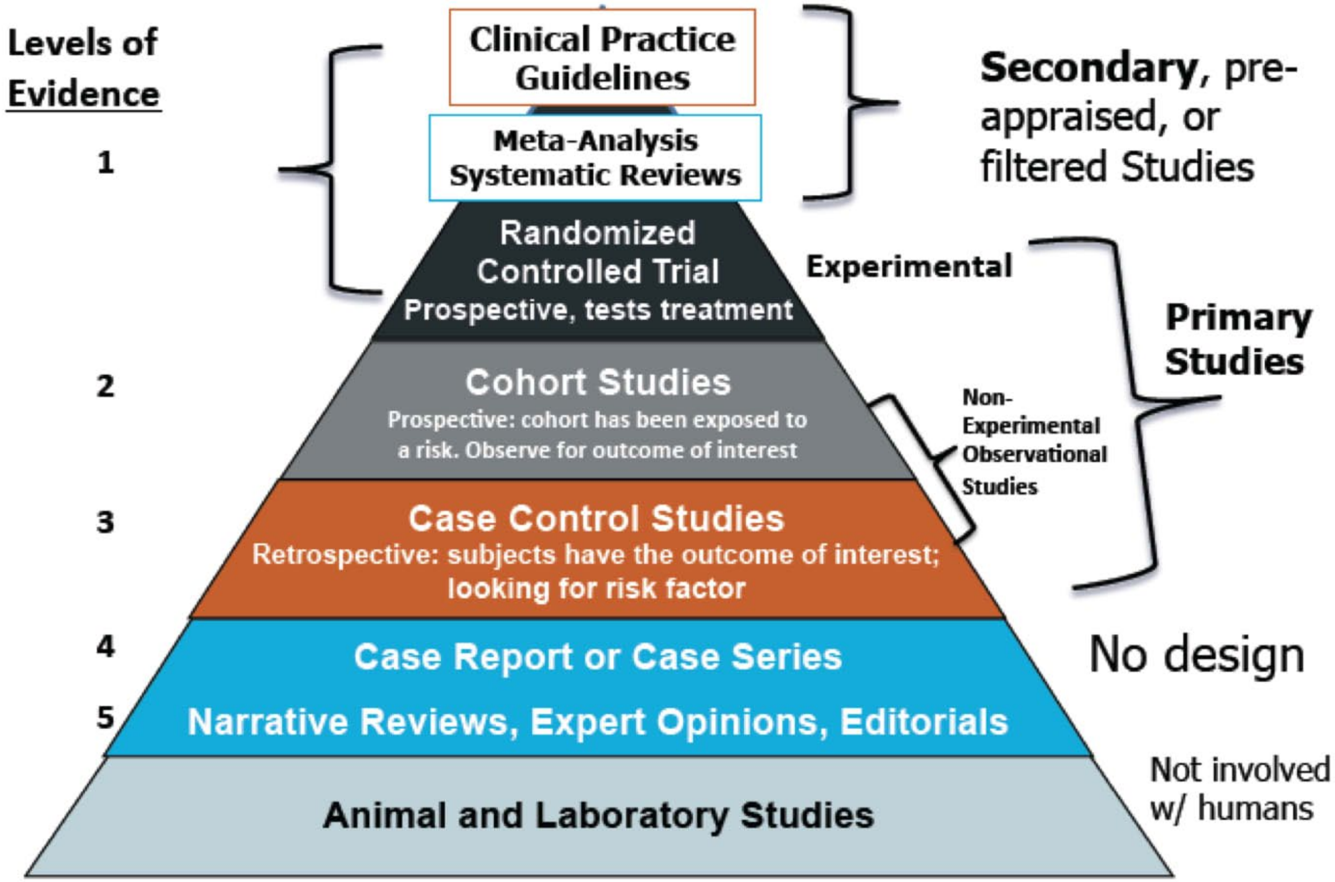
Assesses and describes a finding

Lacks a comparison group

Establishing cause and effect is not possible

Hypothesis generating

Evidence Pyramid



Levels of Evidence

1

Clinical Practice Guidelines

Meta-Analysis Systematic Reviews

Secondary, pre-appraised, or filtered Studies

Randomized Controlled Trial
Prospective, tests treatment

Experimental

Primary Studies

2

Cohort Studies
Prospective: cohort has been exposed to a risk. Observe for outcome of interest

Non-Experimental Observational Studies

3

Case Control Studies
Retrospective: subjects have the outcome of interest; looking for risk factor

4

Case Report or Case Series

No design

5

Narrative Reviews, Expert Opinions, Editorials

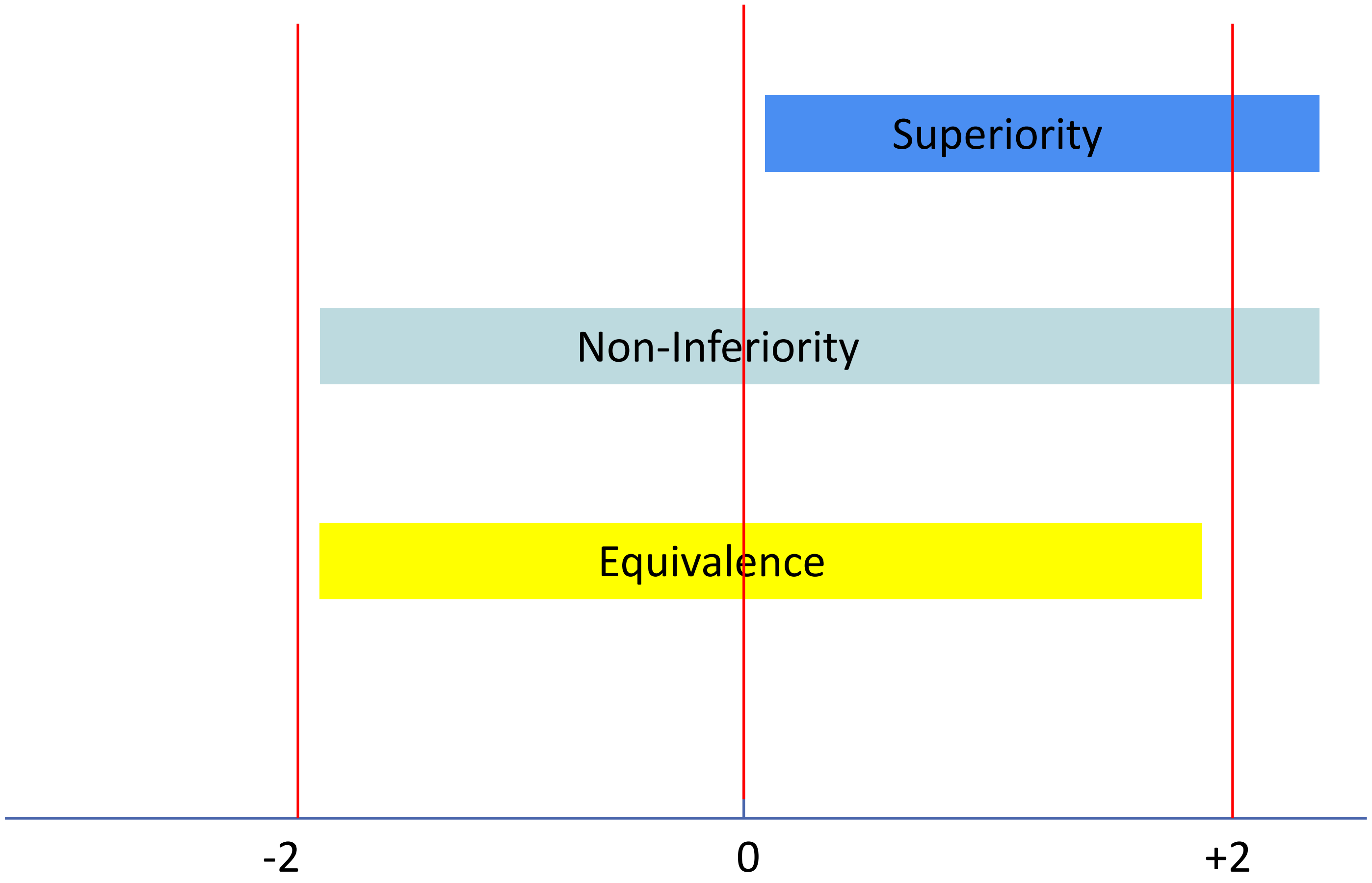
Animal and Laboratory Studies

Not involved w/ humans

Non-inferiority Trials

- Define superiority, non-inferiority, and equivalence

Trial Types



Trial Types

- Superiority
  - Study is designed to ask if a treatment **is better**
  - Superiority is found if we reject the null hypothesis that the treatments are similar
  - Superiority is found if the difference does not
    - Cross 0 (for a continuous variable)
    - Cross 1 (for a dichotomous variable)

- Non-Inferiority
  - Study is designed to ask if a treatment **is not unacceptably worse**
  - Unacceptably worse should be defined by meta-analysis or minimally acceptable clinical difference
  - Superiority is found if we reject the null hypothesis that the treatments are different
  - Non-inferiority is found if the lower 95% CI does not cross the predetermined threshold
  - Threshold should be the minimal difference that would be clinically important

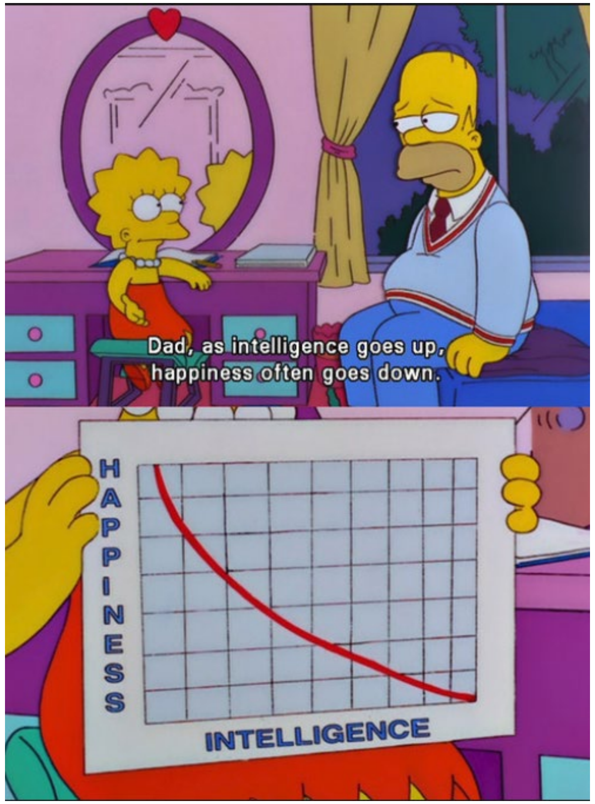
- Equivalence
  - Study is designed to ask if a treatment **is neither unacceptably worse or better**
  - Equivalence is found if both the upper and the lower 95% CI do not cross the predetermined threshold
  - Threshold should be the minimal difference that would be clinically important

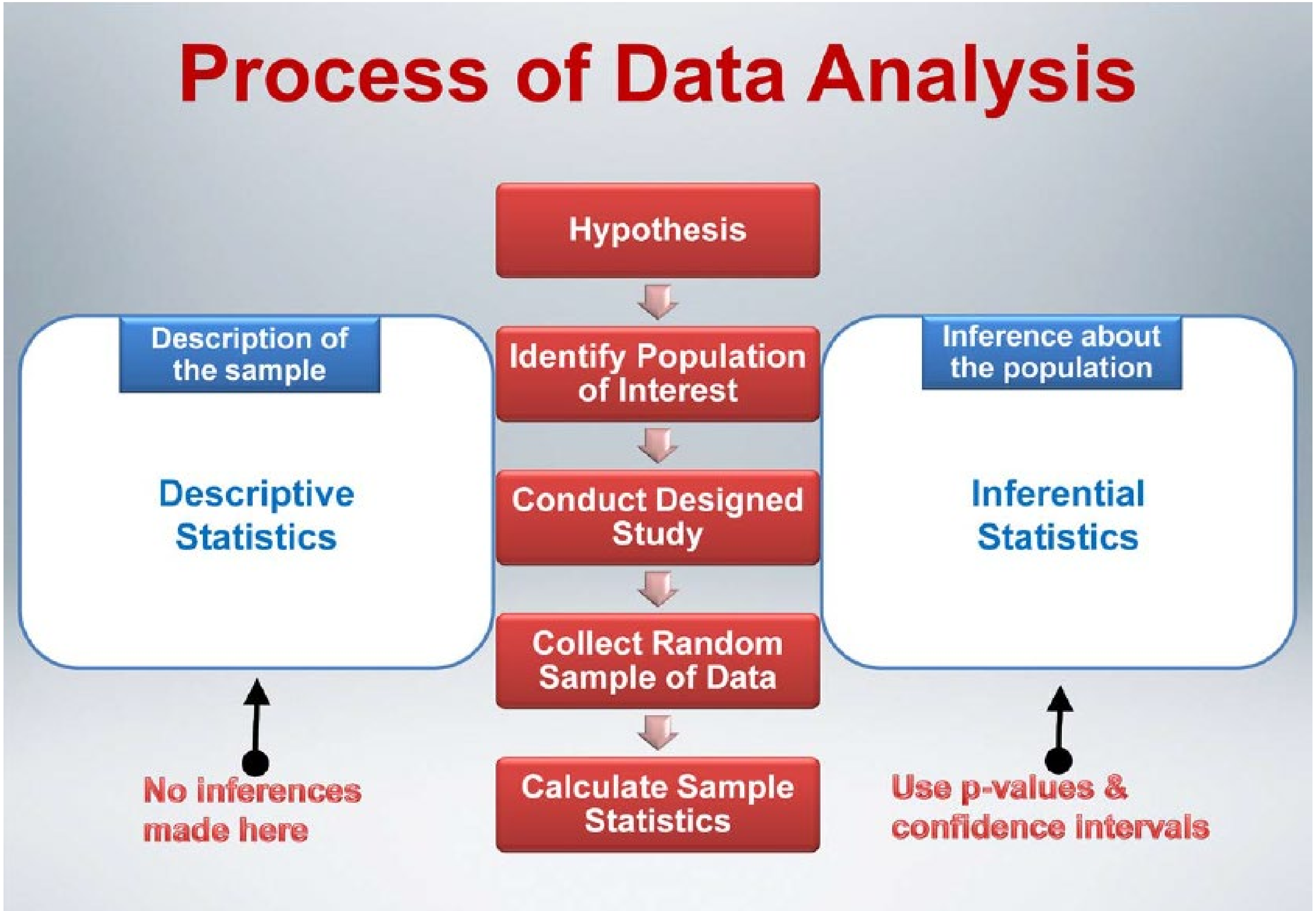
Sample Size Estimates

- *α-level*
- *Power*
- Baseline rate of events in control group
- Desired detectable difference in experimental group
- Ratio of controls : experimental subjects
- Paired or unpaired data

Descriptive Statistics

- The greatest value of a picture is when it forces us to notice what we never expected to see. — John W. Tukey



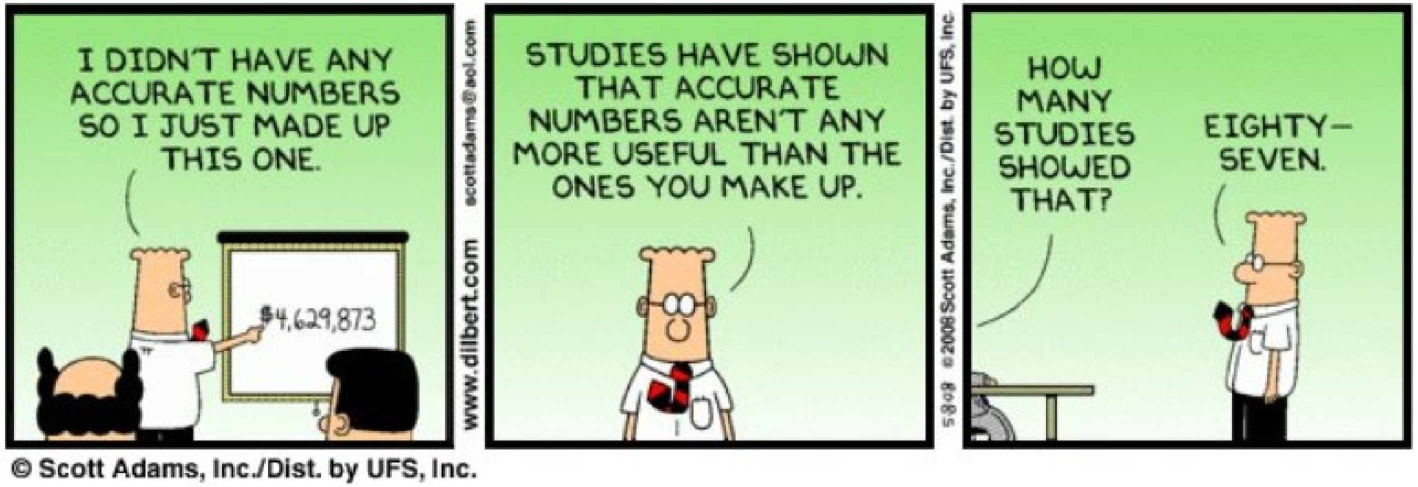


Process of Data Analysis

Hypothesis
Identify Population of Interest
Conduct Designed Study
Collect Random Sample of Data
Calculate Sample Statistics
Description of the sample
Descriptive Statistics
No inferences made here
Inference about the population
Inferential Statistics
Use p-values & confidence intervals

Descriptive Statistics

- Look at raw data before anything else!
  - Does it make sense?
  - Are there obvious errors?
  - Do the groups visually look different without the use of statistics?
  - Do the descriptive statistics inform your analysis further?


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.
www.dilbert.com
scottadams@aol.com
© 2008 Scott Adams, Inc./Dist. by UFS, Inc.
5-8-08

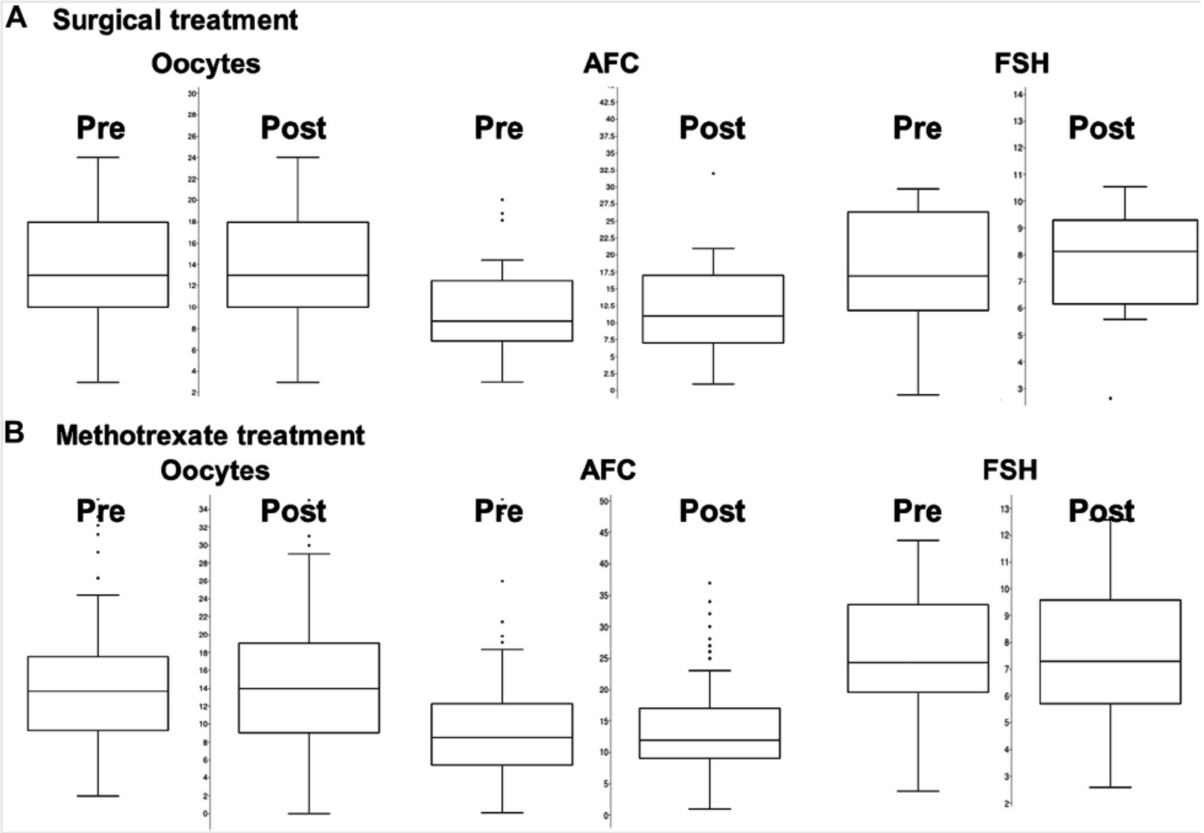
© Scott Adams, Inc./Dist. by UFS, Inc.

Descriptive Statistics

- Mean and median
- Range and IQR
- STDEV and SEM
- Line graphs
- Frequency histograms
- Box and whisker plots
- Scatter Plots

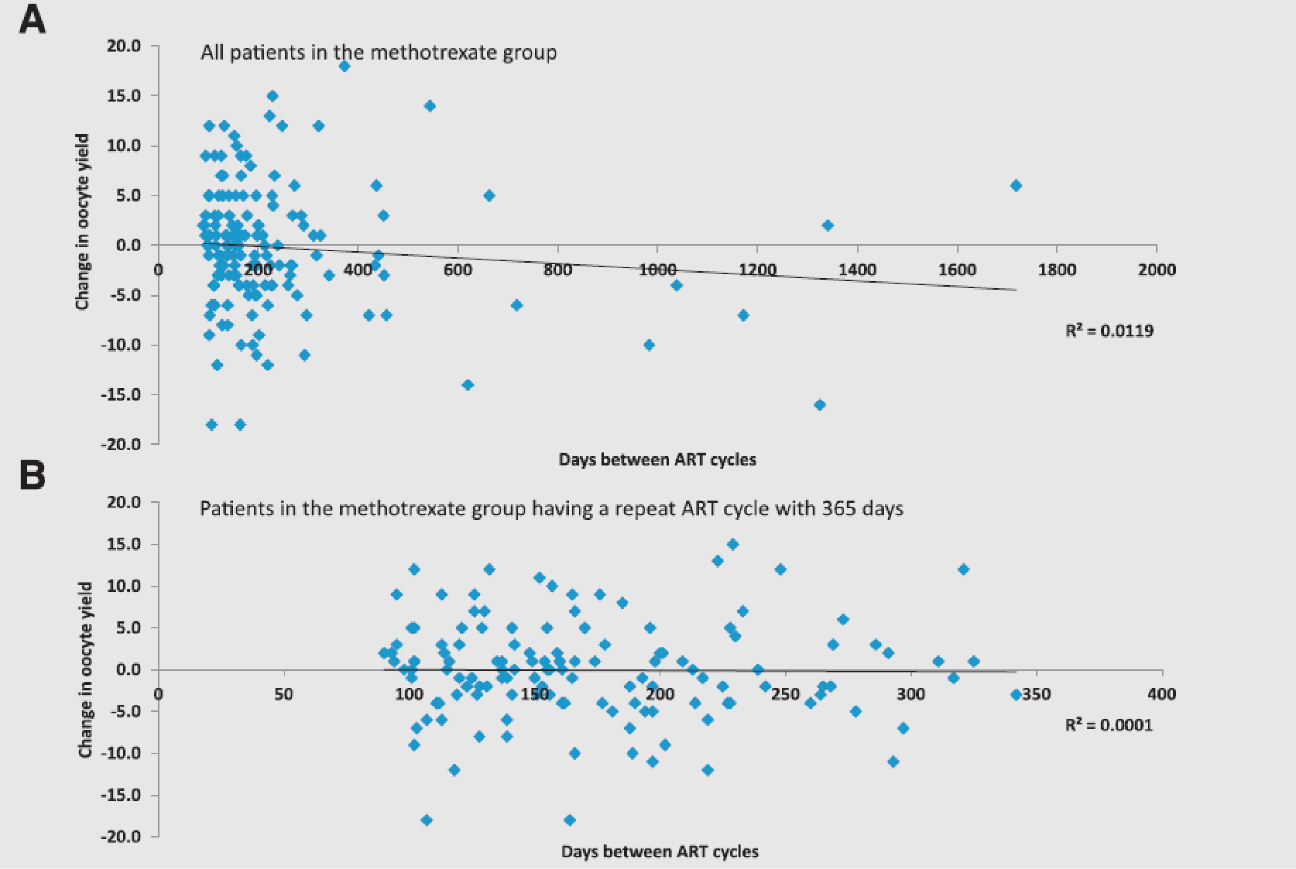
Example: MTX versus Surgery for IVF Ectopic

- Box and Whisker



Hill et al, F&S 2014, PMID: 24269042

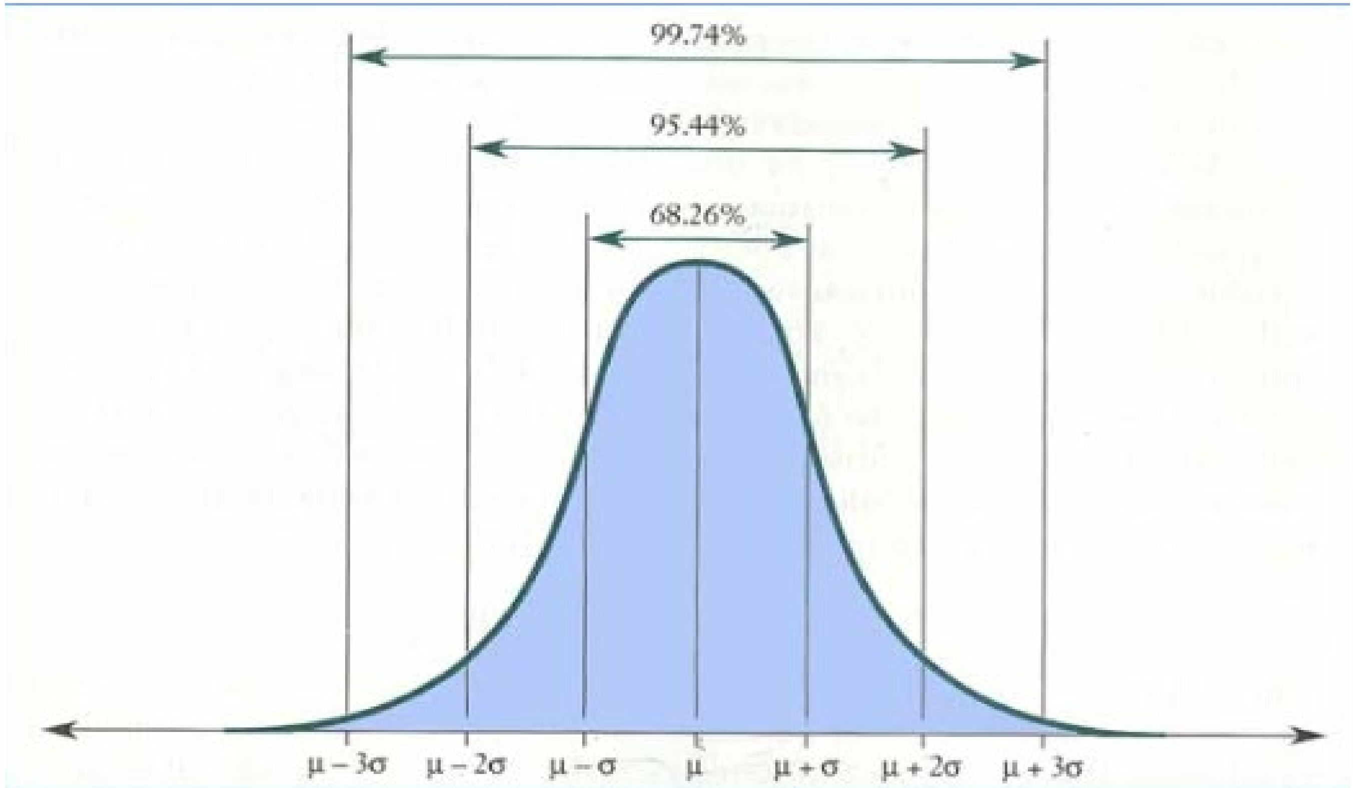
Scatter Plot



Normality

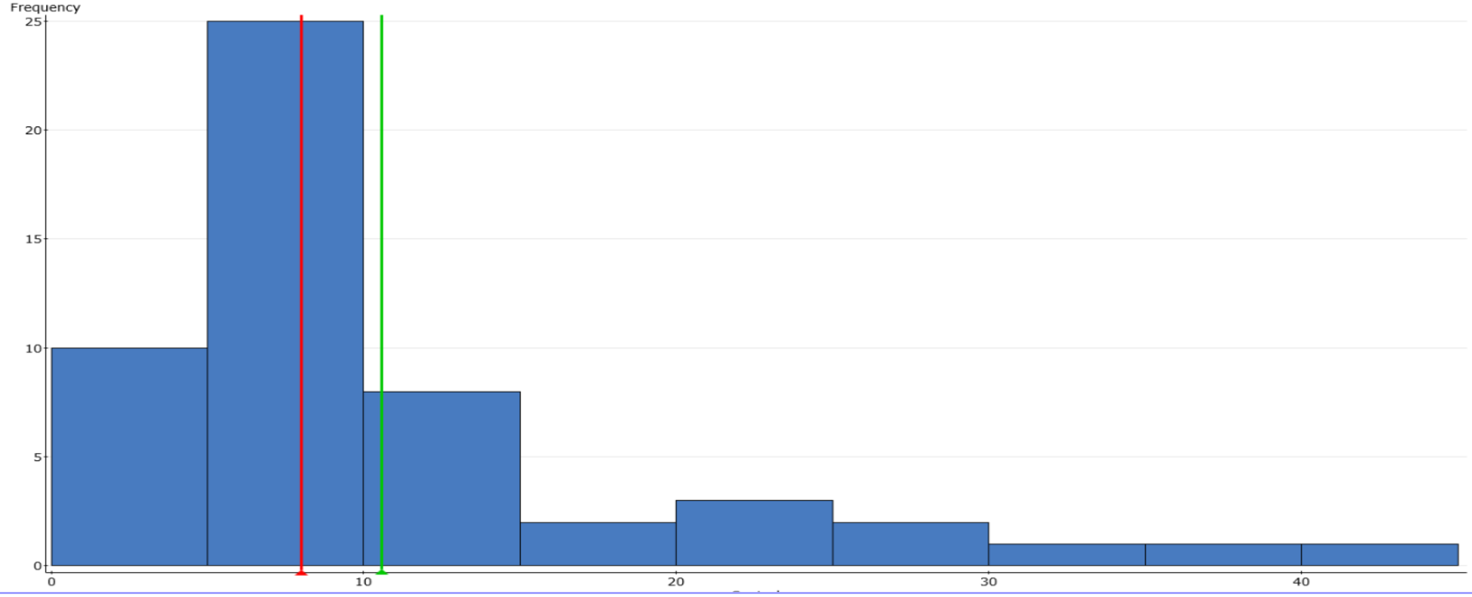
- Look at the data!
- Shapiro-Wilk test
- Komogorov-Smirnov test

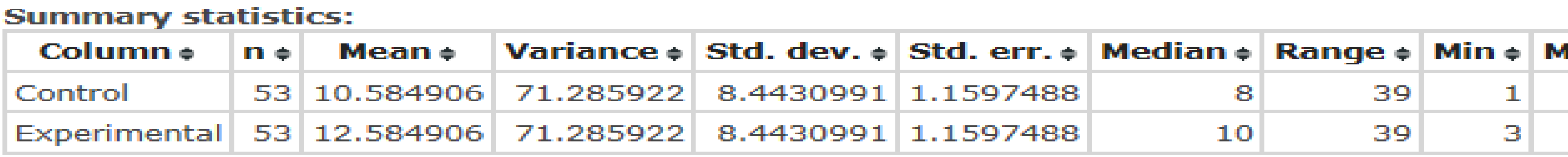
μ = mean
Ơ = standard deviation
1 STDEV 68% of data
2 STDEV 95% of data
3 STDEV 99% of data



MARCH 19 - 23
PCRS 2025
PACIFIC WAVES - EXPLORING SCIENTIFIC FRONTIERS IN AN EVOLVING SOCIETY
INDIAN WELLS, CA

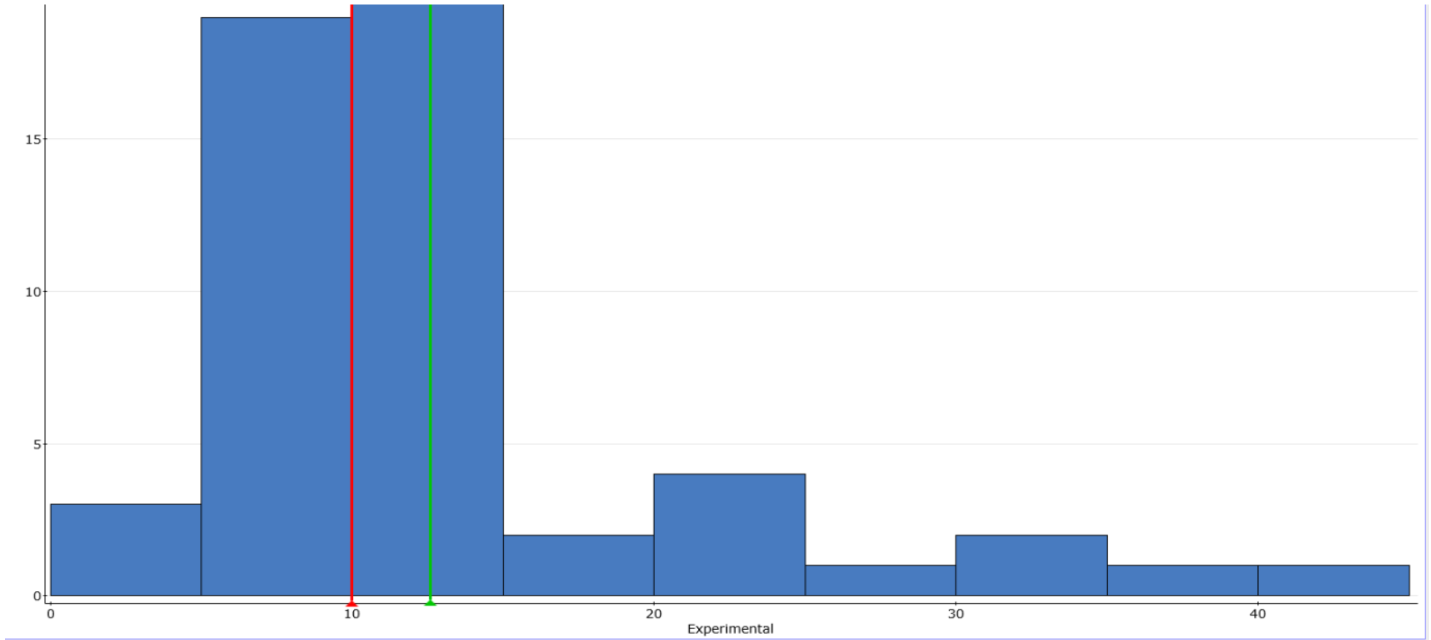
Frequency Histograms





Summary statistics:

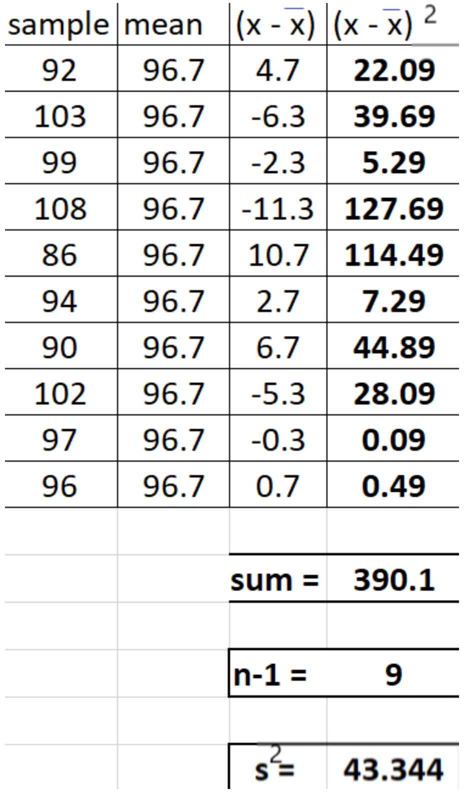
Column ⬍	n ⬍	Mean ⬍	Variance ⬍	Std. dev. ⬍	Std. err. ⬍	Median ⬍	Range ⬍	Min ⬍	M
Control	53	10.584906	71.285922	8.4430991	1.1597488	8	39	1	
Experimental	53	12.584906	71.285922	8.4430991	1.1597488	10	39	3	



Variance

- Sum of the differences of each value from the mean squared / sample size

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$



sample	mean	$(x - \bar{x})$	$(x - \bar{x})^2$
92	96.7	4.7	**22.09**
103	96.7	-6.3	**39.69**
99	96.7	-2.3	**5.29**
108	96.7	-11.3	**127.69**
86	96.7	10.7	**114.49**
94	96.7	2.7	**7.29**
90	96.7	6.7	**44.89**
102	96.7	-5.3	**28.09**
97	96.7	-0.3	**0.09**
96	96.7	0.7	**0.49**
		sum =	**390.1**
		n-1 =	**9**
		s^2=	**43.344**

- Measures the spread of the data

Standard Deviation

- Square root of variance

$$\text{variance} = \sigma^2 = \frac{\Sigma (x_r - \mu)^2}{n}$$

$$\text{standard deviation} \quad \sigma = \sqrt{\frac{\Sigma (x_r - \mu)^2}{n}}$$

$$\mu = \text{mean}$$

- A measure of the dispersion of a set of data from its mean

Standard Error of the Mean

- STDEV ÷ square root of the sample size

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

- Measures how precisely you know the population mean

STDEV and SEM

- Use STDEV error bars when you want to show the variability of your data

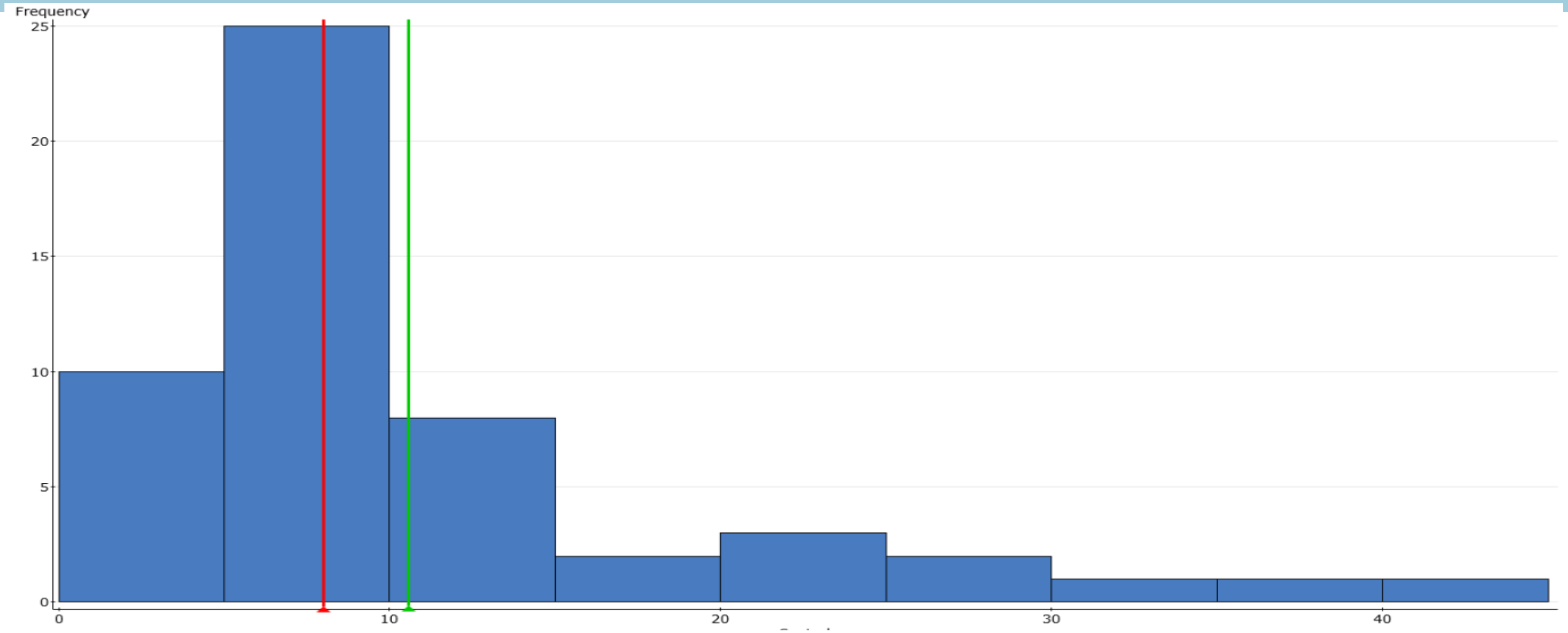
- Use SEM error bas when you want to the precision of the estimation of the population mean

- SEM will always have smaller error bars

STDEV and SEM

- You can run statistics on parametric data with just means and STDEV
- You cannot run statistics on non-parametric data without the raw data
- Overlapping STDEV **does not** tell you if the two groups are similar
- Overlapping SEM **does not** tell you if the two groups are similar

Is this normally distributed????



- Doesn’t look like a bell
- Mean and median are not similar
- Data with large right tail (positive or right skew)
- 2 standard deviations should encompass 95% of data
  - Mean 10.5 ± 8.4
  - 10.5 – 2SD = -6.5
  - You cant have negative oocytes

Inferential Statistics

"Data don't make any sense,
we will have to resort to statistics."

Outcomes Analysis

Comparison	Parametric	Non-Parametric
2 means	Student’s T test	Mann-Whitney U
2 paired means	Paired T test	Wilxocon signed rank
3 or more means	ANOVA	Kruskal-Wallis
3 or more repeated means	Repeated measures ANOVA	Friedman
Correlation	Pearson’s Coefficient	Spearman’s Coefficient

Comparison	<5 outcomes in any comparison	≥ outcomes in any comparison
Dichotomous 2 groups	Fisher’s exact test	Chi square
Dichotomous multiple groups	Fisher’s exact test	Chi square

Communicating Statistics

- Absolute risk
- NNT/NNH
- RR
- OR
- P value

EMPLOYEE SATISFACTION HAS DOUBLED SINCE LAST YEAR!
EMPLOYEE SATISFACTION
1% → 2%
DilbertCartoonist@gmail.com
Dilbert.com

Definitions

- Risk difference and absolute risk
    - Difference in risk between the exposure groups
- NNT
    - the number of patients treated to have 1 different outcome
- Odds Ratio
    - the **odds** that an outcome will occur given a particular exposure, compared to the **odds** of the outcome occurring in the absence of that exposure
- Relative Risk
    - the **risk** that an outcome will occur given a particular exposure, compared to the **risk** of the outcome occurring in the absence of that exposure

Risk versus Odds

- 80/100 patients get pregnant with a new drug
- Risk of pregnancy
    - # of positives ÷ total # of patients
    - 80/100
    - 80%
    - 0.8
- Odds of pregnancy
    - # of positives ÷ # of negatives
    - 80/20
    - 4:1
    - 4

80/100 patients get pregnant versus 40/100 patients get pregnant

- Risk difference is .80 – .40 = .40
- Absolute risk is 80% – 40% = 40%
- NNT is 100/40 = 2.5 -> 3
- RR = 80/100 ÷ 40/100 = 80/40 = 2
- OR = 80/20 ÷ 40/60 = 4/.666 = 6

NNT

- 100 ÷ Absolute risk
- If Absolute risk is 50%, NNT = 100/50 = 2
- If Absolute risk is 10%, NNT = 100/10 = 10
- If Absolute risk is 1%, NNT = 100/1 = 100

Estimating treatment effects

Group	Outcome	
	Positive	Negative
Treatment	a	b
Control	c	d

- Risk difference (RD)

$$\frac{a}{a+b}-\frac{c}{c+d}$$

- Relative risk (RR)

$$\frac{a/(a+b)}{c/(c+d)}$$

- Odds ratio (OR)

$$\frac{a/b}{c/d}$$

Estimating treatment effects

- Difference between how often something occurred in the two groups

- How often an event occurred/ number of patients between the two groups

- How often an event occurred/ how often an event did not occur between the two groups

- Risk difference (RD)

$$\frac{a}{a+b} - \frac{c}{c+d}$$

- Relative risk (RR)

$$\frac{a/(a+b)}{c/(c+d)}$$

- Odds ratio (OR)

$$\frac{a/b}{c/d}$$

A large RR ≠ A Large Absolute Risk

Group	Outcome	
	Positive	Negative
Treatment	*5*	*995*
Control	*1*	*999*

- Risk difference (RD)

$$\frac{5}{5+995}-\frac{1}{1+999}=0.004$$

- Relative risk (RR)

$$\frac{5/(5+995)}{1/(1+999)}=5.00$$

- Odds ratio (OR)

$$\frac{5/995}{1/999}=5.02$$

Absolute risk = 5/1000 – 1/1000 = 4/1000 = 0.4%
NNT = 100/0.4 = 250

RR and OR are similar when events are rare

Group	Outcome	
	Positive	Negative
Treatment	*5*	*995*
Control	*1*	*999*

- Relative risk (RR)

$$\frac{5/(5+995)}{1/(1+999)} = 5.00$$

- Odds ratio (OR)

$$\frac{5/995}{1/999} = 5.02$$

OR overstates the effect as eventsare more common

Group	Outcome	
	Positive	Negative
Treatment	*60*	*40*
Control	*20*	*80*

- Risk difference (RD)

$$\frac{60}{60+40} - \frac{20}{20+80} = 0.40$$

- Relative risk (RR)

$$\frac{60/(60+40)}{20/(20+80)} = 3.00$$

- Odds ratio (OR)

$$\frac{60/40}{20/80} = 6.00$$

RR and OR Relationship by Disease Prevalence

Control	Experimental	RR	OR	NNT
1 : 1000	2: 1000	2	2.001	1000
1 : 500	2: 500	2	2.004	500
1: 100	2 : 100	2	2.02	100
10: 100	20 : 100	2	2.25	10
40 : 100	80 : 100	2	6	3
45:100	90:100	2	11.25	3
49.5 : 100	99 : 100	2	101	2

RR versus OR

- OR and RR can always be calculated for binary outcomes
- RR cannot be calculated for case-control study designs (unknown denominator)
- RR is intuitively easier to understand than OR
- RR and OR are commonly (but mistakenly) interpreted as equivalent
  - OR interpreted as RR will always overstate effect size
  - RR and OR are similar when event rates are rare, but are increasingly different (OR more extreme) as event frequency increases
  - Differences between RR and OR increase with greater treatment effect sizes

Interpreting OR, RR, and Correlations

- What does a RR of 0.95 for live birth and age mean?

- What does a RR of 2.5 for live birth a embryo quality mean?
  - Poor embryo 10%
  - Fair embryo 25%
  - Good Embryo 62.5%
  - 2.5x increased risk for each increment 10% -> 25% -> 62.5%

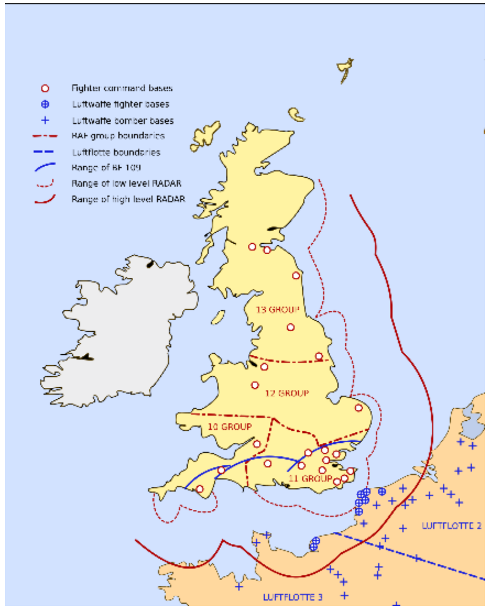
- R square = amount of change in one variable based on another

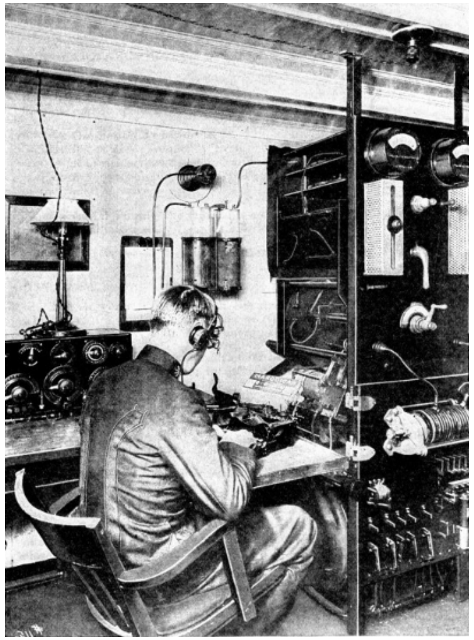
Prediction Statistics

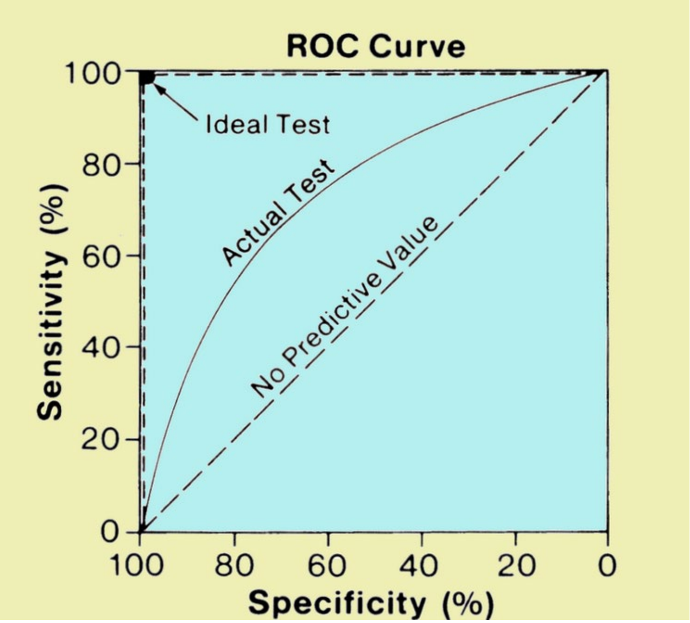
- Sensitivity- I have disease, what is the chance of positive test
- Specificity- I don’t have disease, what is the chance of a negative test
- PPV- I have a positive test, what is the chance of disease
- NPV- I have a negative test, what is the chance of no disease

- $\text{Sens} = TP \; / \; TP + FN$
- $\text{Spec} = TN / TN + FP$
- $\text{PPV} = TP / TP + FP$
- $\text{NPV} = TN / TN + FN$

ROC







Radar detector setting	Percent of German planes detected (sensitivity)	Percent of geese flocks correctly identified (specificity)	Percent of geese flocks incorrectly identified (1- specificity)
Off	0	100	0
Setting 1	35	93	7
Setting 2	60	85	15
Setting 3	85	70	30
Setting 4	92	30	70
Full	100	0	100

ROC Curve

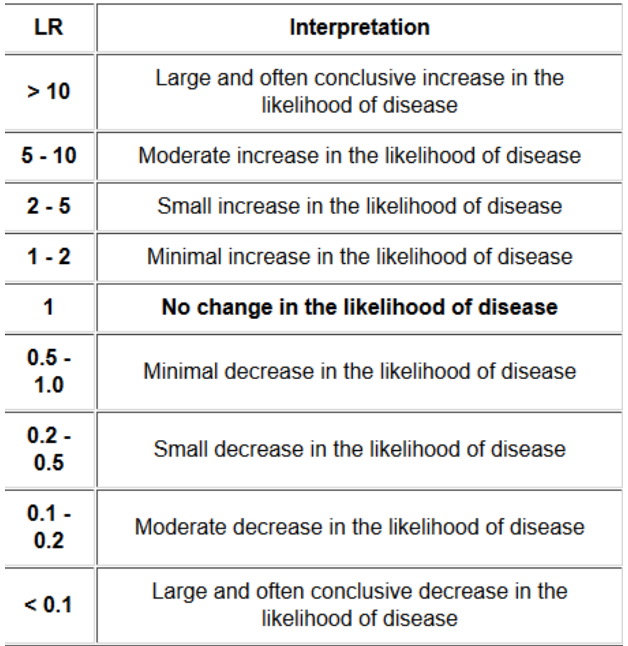
- Plot sensitivity versus 1-specificity
- Calculate the area under the curve (AUC)

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

- AUC should not be below .5   If it is, flip the question and AUC will flip in direction.

Likelihood Ratio

- How much do we shift our opinion based on a result?
- Probability of obtaining a + test in a diseased patient ÷ probability of a + test in a healthy patient
- Sensitivity ÷ (1 – Specificity)



LR	Interpretation
> 10	Large and often conclusive increase in the likelihood of disease
5 - 10	Moderate increase in the likelihood of disease
2 - 5	Small increase in the likelihood of disease
1 - 2	Minimal increase in the likelihood of disease
1	**No change in the likelihood of disease**
0.5 - 1.0	Minimal decrease in the likelihood of disease
0.2 - 0.5	Small decrease in the likelihood of disease
0.1 - 0.2	Moderate decrease in the likelihood of disease
< 0.1	Large and often conclusive decrease in the likelihood of disease

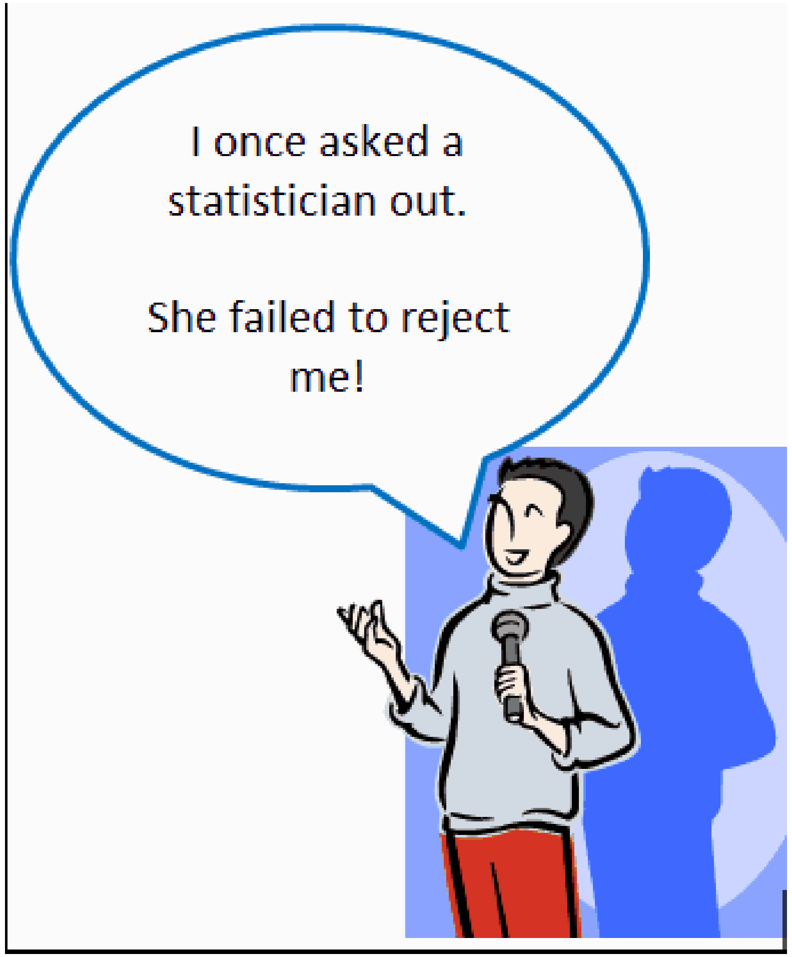
Post Test Probability

- Now that I have the result, how probable is the outcome?

- Post test probability = pretest probability * likelihood ratio

Hypothesis Testing

- You can reject or fail to reject the null hypothesis
- You cannot accept the null hypothesis
- You cannot statistically reject or accept the alternate hypothesis



What is a *P* value?

What is a *P* value?

- A measure of the probability that an effect size as large as the one observed (or larger) could have resulted from random chance

- Is calculated on the assumption that the null hypothesis is true
  - “if the null hypothesis is true, what is the chance that random sampling of a population would have led to the effect seen in the data?”

- $1 \geq P \geq 0$

- Only2 possible outcomes
  - Statistically significant
  - Not statistically significant

P value

- A P-value *does not*
  - indicate the strength of a relationship
  - indicate clinical significance
    - Statistically significant effects may not be clinically significant
    - Clinically significant effects may exist even if statistical significance is not found

Rejecting the Null Hypothesis

- *α-level*
  - significance level
  - the probability (*P* value) at or below which H_0 is rejected
  - the probability of rejecting an H_0 that is true
  - ***Type I error***
  - Typically $\boldsymbol{\alpha}$ = 0.05
  - False positive finding rate

- *β-level*
  - the probability of failing to reject an H_0 that is false
  - ***Type II error***
  - ***Typically β = 0.20***
  - (1 - β) = ***power***
  - the probability of rejecting an H_0 that is false
  - False negative finding rate

Type I & II Errors

- Type I error: falsely rejecting the null hypothesis
  - we find a difference that doesn’t exist
  - By convention we accept a 5% risk we are wrong in **rejecting** the null hypothesis

- Type II error: falsely accepting the null hypothesis
  - We don’t find a difference that truly exists
  - By convention we accept a 20% risk we are wrong in ***failing to reject*** the null hypothesis

- Law analogy
  - we would prefer to falsely find a murderer innocent (20% risk of letting the murdered go free)
  - over falsely convicting an innocent person (5% risk of wrongly imprisoning the prisoner)
  - You can be found guilty or not guilty
  - You cant be found innocent

Common Statistical Pitfalls

- Mistake association or correlation for causation
- Finding no difference does not prove the groups are equivalent (maybe type II error)
- Don’t say two groups were “different, but not statistically different”
- Don’t say there is a trend to significance for low P values
- Don’t say “very significant” or “highly significant” for low P values
- Express non-parametric data as mean ± STDEV

Unit of Analyses

- The unit of analysis should typically be the patient
- Must be the unit of randomization
- Using embryos as unit of analysis falsely increases power
- Comparisons of IVF-ET clinical pregnancy and implantation rates at SGF
- 4th quarter 2008 (n=649) versus 1st quarter 2009 (n=974)
- Using patients as the unit of analysis:
  - Pregnancy = 48.7% vs 51.1%, p = 0.34 (chi-square)
  - Implantation = 33.8% vs 36.6%
    - Chi square = 0.19
- Using embryos as the unit of analysis:
  - Implantation = 432/1465 (29.5%) vs 684/2093 (32.7%)
    - Chi-square test: p = 0.043

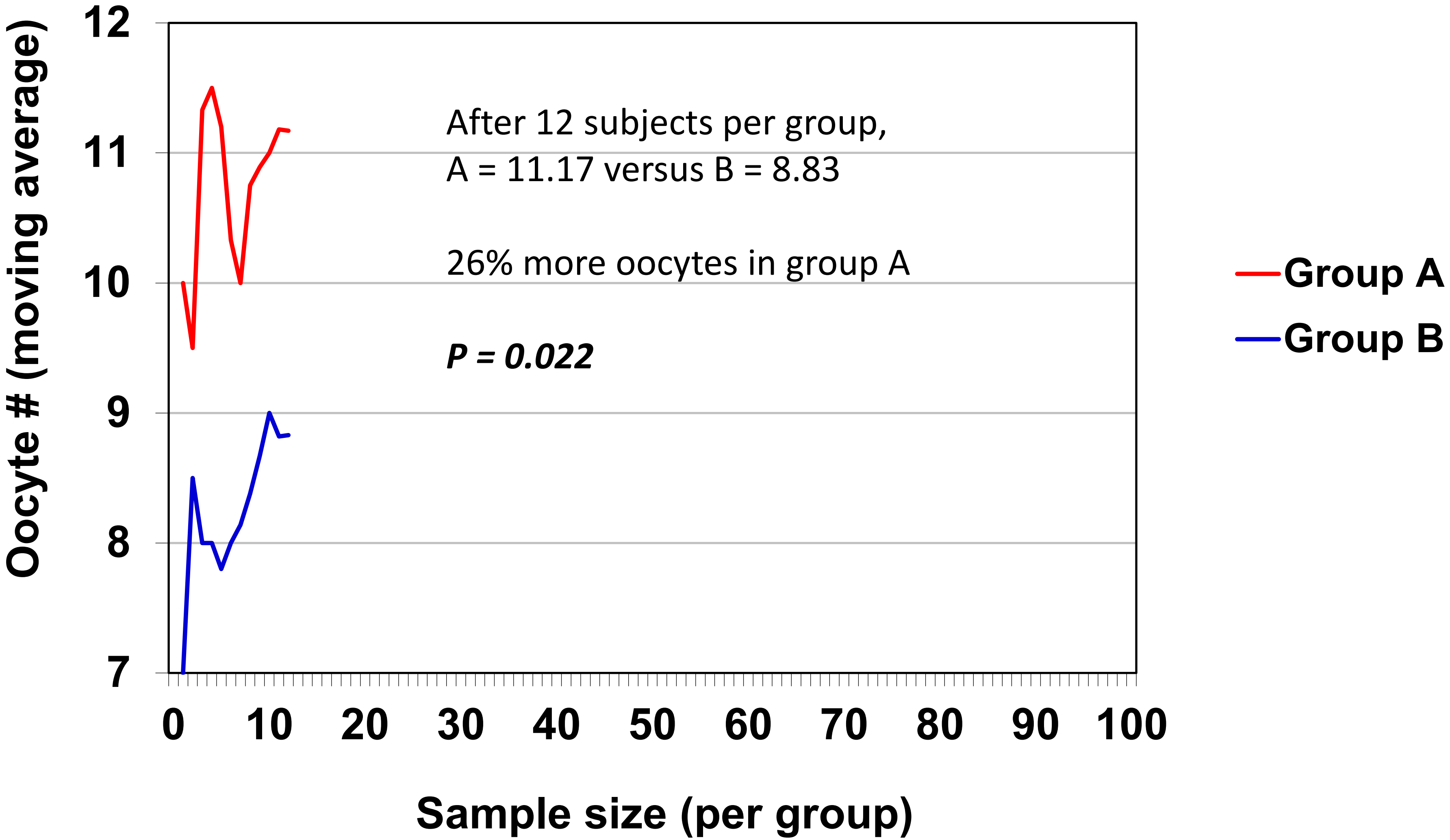
Ending a trial early

- Investigators will often end a prospective trial earlier than originally planned if an interim analysis indicates a statistically significant trend
- The problem: doing so will often give misleading results

Ending a trial early: example

- Simulation of randomized prospective trial of two stimulation protocols
- Study outcome: number of oocytes
- Two samples of 100 subjects each, simulated using a random number generator (excel)
- Both groups sampled from a population with mean = 10 (SD = 3) oocytes

Ending a trial early: example



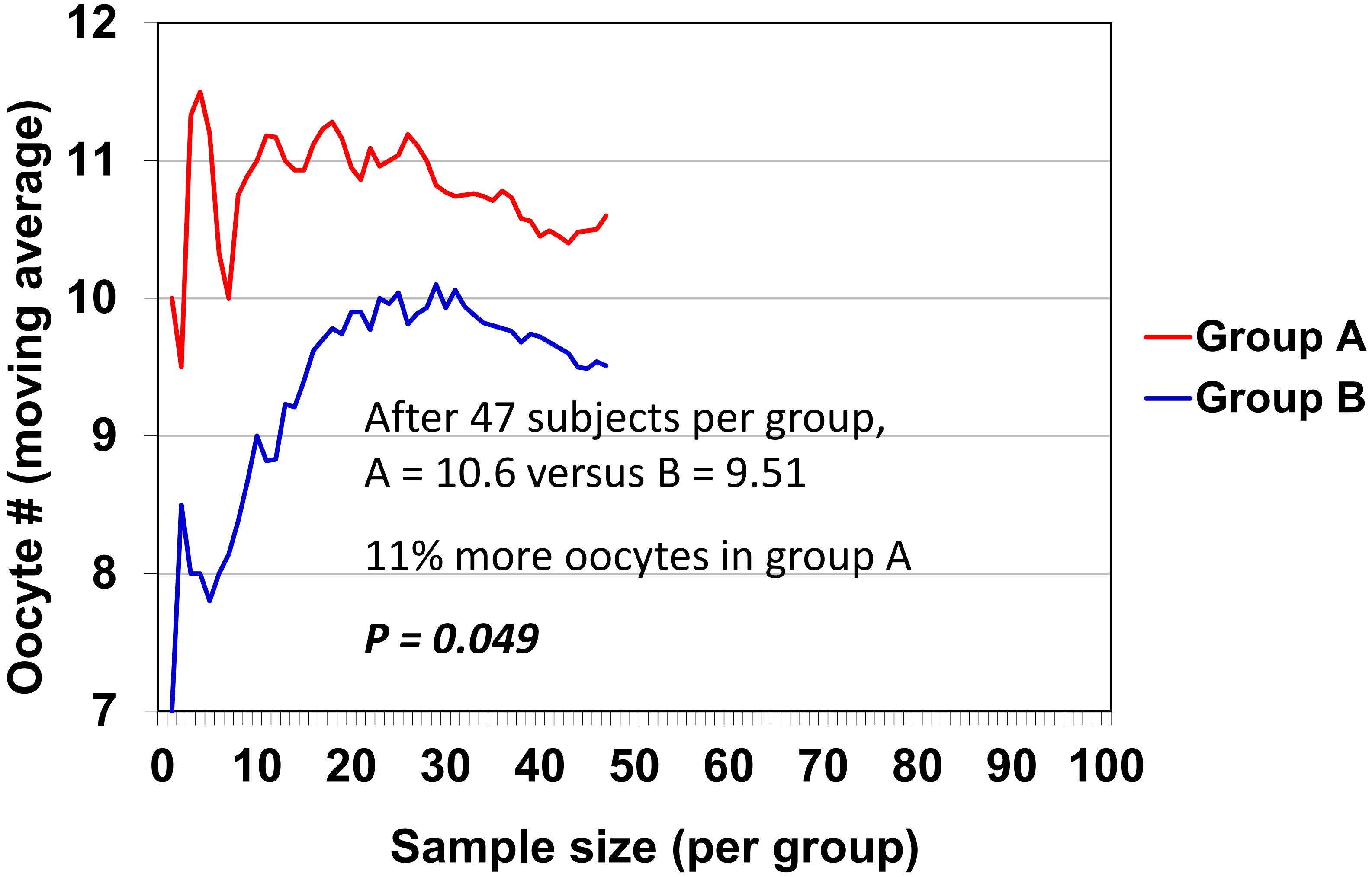
After 12 subjects per group,
A = 11.17 versus B = 8.83

26% more oocytes in group A

P = 0.022

Simulation courtesy of Kevin Richter PhD

Ending a trial early: example



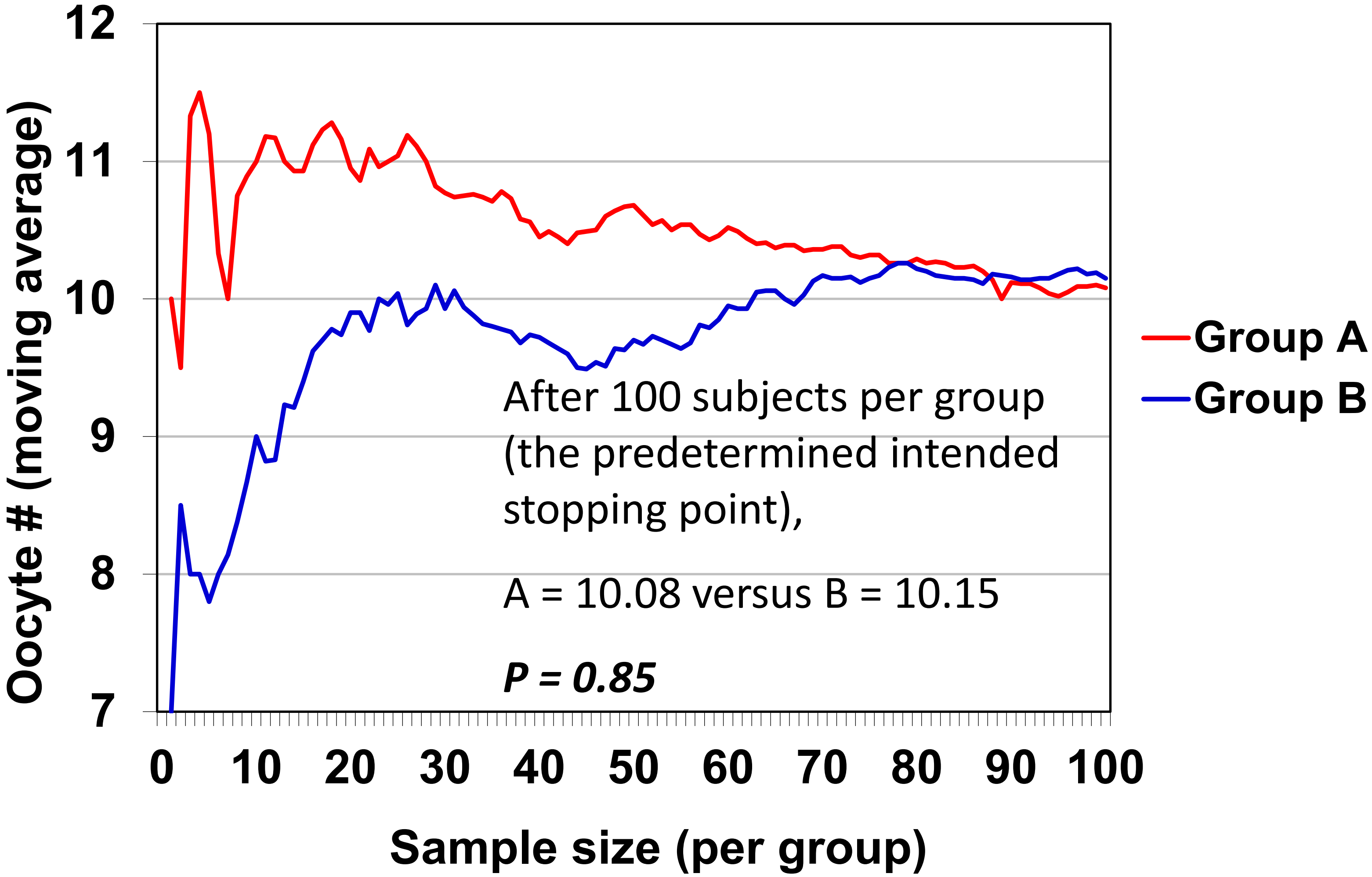
After 47 subjects per group,
A = 10.6 versus B = 9.51

11% more oocytes in group A

P = 0.049

Simulation courtesy of Kevin Richter PhD

Ending a trial early: example



After 100 subjects per group (the predetermined intended stopping point),

A = 10.08 versus B = 10.15

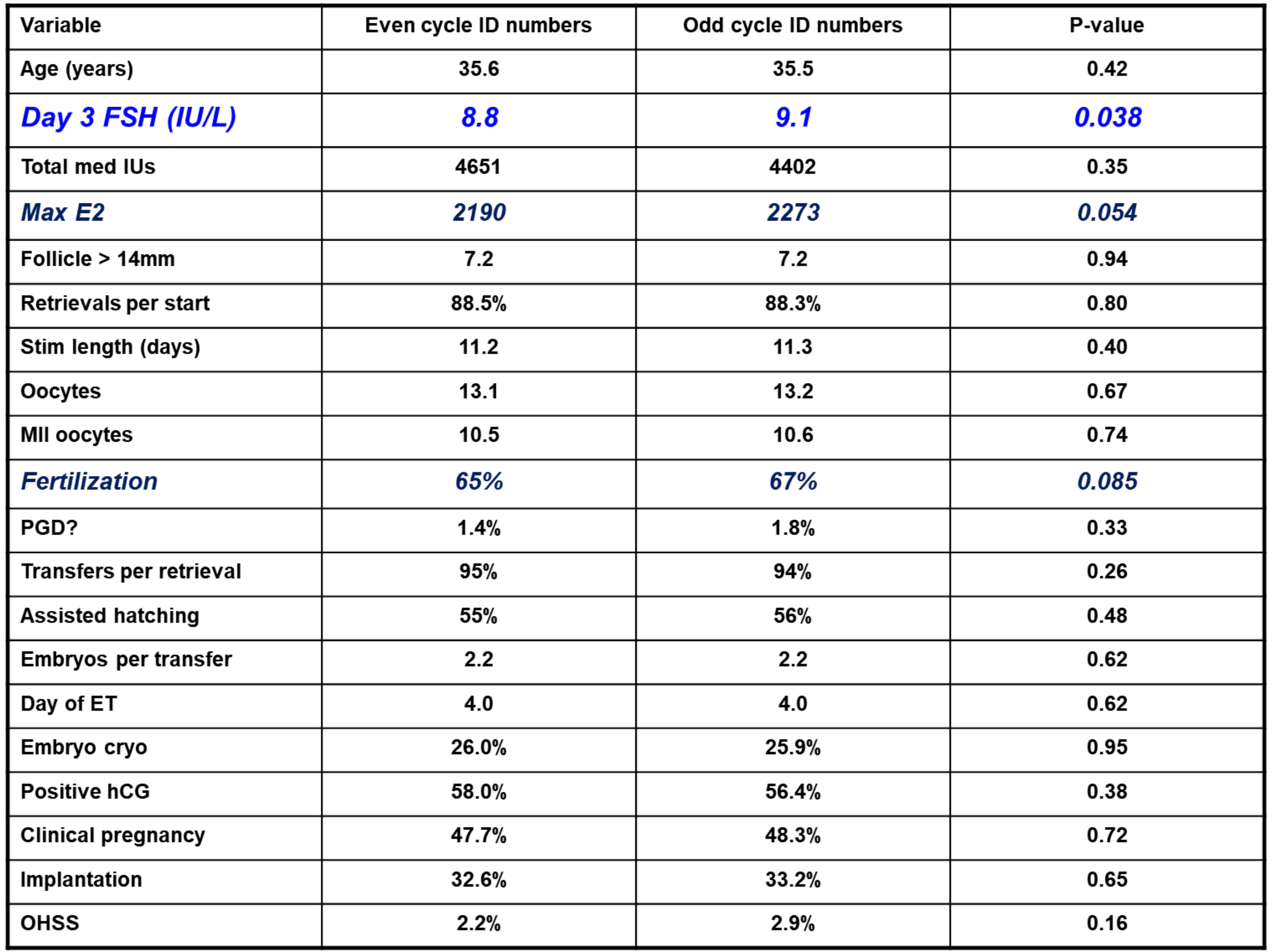
P = 0.85

Simulation courtesy of Kevin Richter PhD

Multiple Comparisons

- The problem:
  - A given p-value indicates the probability of Type-I error (*i.e.* mistakenly concluding that there is a difference when there really is not) for a ***single*** comparison
  - If more than one comparison is made, the chances of making a Type-I error for ***any*** of the comparisons is greater than indicated by the p-values for each comparison
  - The more comparisons that are made, the greater the chances of making one or more Type-I errors (unless the threshold for significance is adjusted appropriately)

Random Chance due to Multiple Comparison



Variable	Even cycle ID numbers	Odd cycle ID numbers	P-value
Age (years)	35.6	35.5	0.42
Day 3 FSH (IU/L)	*8.8*	*9.1*	*0.038*
Total med IUs	4651	4402	0.35
Max E2	*2190*	*2273*	*0.054*
Follicle > 14mm	7.2	7.2	0.94
Retrievals per start	88.5%	88.3%	0.80
Stim length (days)	11.2	11.3	0.40
Oocytes	13.1	13.2	0.67
MII oocytes	10.5	10.6	0.74
Fertilization	*65%*	*67%*	*0.085*
PGD?	1.4%	1.8%	0.33
Transfers per retrieval	95%	94%	0.26
Assisted hatching	55%	56%	0.48
Embryos per transfer	2.2	2.2	0.62
Day of ET	4.0	4.0	0.62
Embryo cryo	26.0%	25.9%	0.95
Positive hCG	58.0%	56.4%	0.38
Clinical pregnancy	47.7%	48.3%	0.72
Implantation	32.6%	33.2%	0.65
OHSS	2.2%	2.9%	0.16

Multiple Comparisons

Limit Comparisons

Define a single primary outcome

Adjust the threshold for defining statistical significance so that the chance of making any Type-I errors for any of the comparisons made is below the desired study-wide error rate (typically 0.05)

- **Bonferroni method**:
  - The desired study-wide error rate is divided by the number of comparisons made
  - For this example, 0.05 / 20 = 0.0025
- **Holm-Bonferroni method**:
  - The lowest p-value is compared to the adjusted threshold as above
  - If significant, the next lowest p-value is compared to the threshold adjusted for the number of remaining comparisons (*i.e.* 0.05 / 19 = 0.0026)
  - This process is continued until a comparison fails to meet the criterion for statistical significance

Multiple Comparisons

- Don’t correct in non-inferiority or equivalence studies

- Don’t necessarily correct if all the data consistently shows a difference
  - Eg implantation, clinical pregnancy, ongoing pregnancy and live birth all show similar difference

Questions?

© photography by asiya | 2013

References

- Modified from Schold, Jesse D., and S. Joseph Kim, "Clinical Research Methods and Analysis in Organ Transplantation", Textbook of Organ Transplantation (2014); 1607-1621
- McKibbon A, Eady A, Marks S. PDQ: Evidence-Based Principles and Practice. Hamilton, Ontario: B.C. Decker Inc., 1999.
- Micah J. Hill, Janelle C. Cooper, Gary Levy, Connie Alford, Kevin S. Richter, Alan H. DeCherney, Charles L. Katz, Eric D. Levens, Erin F. Wolff, "Ovarian reserve and subsequent assisted reproduction outcomes after methotrexate therapy for ectopic pregnancy or pregnancy of unknown location, Fertility and Sterility", Volume 101, Issue 2, 2014, Pages 413-419.e4, ISSN 0015-0282