Week 4

Wednesday, 8 September 2021 19:49

STATISTICS

GENOMIC DATA SCIENCE

BIOLOGY + COMPUTER SCIENCE + STATISTICS

IS IT POSSIBLE TO TAKE GENOMIC MEASUREMENTS AND PERSONALIZE THERAPIES

WE SHOULD CARE ABOUT STATISTICS
MOST PROBLEMS ARE DUE TO:

- LACK OF TRANSPARENCY

DATA SHOULD ALWAYS MADE AVAILABLE
REPRODUCIBILITY: CAN YOU REPERFORM THE ANALYSIS?

- LACK OF COOPERATION

SHARING BOTH DATA & CODE
DISCOVERING PROBLEMS IN DATA SCIENCE

- LACK OF EXPERTISE

PREDICTION RULES
STUDY DESIGN PROBLEMS
BATCH EFFECTS & CONFOUNDERS

EXPERIMENTAL DESIGN SHOULD BE IN PLACE TO HELP YOU IN THE ANALYSIS

CENTRAL DOGMA OF STATISTICS (INFERENCE)

MEASURING WHOLE POPULATION IS EXPENSIVE
WE WANT TO SAY SOMETHING TRUE ABOUT THE WHOLE POPULATION BY USING PROBABILITY

MAKING A GUESS ON WHAT THE POPULATION LOOKS LIKE

HOW CAN WE QUANTIFY VARIABILITY?

DATA SHARING PLAN
HOW TO SHARE DATA

1) RAW DATA

NO PROCESSING, COMPUTING SUMMERIZING OR DELETING

2) TIDY DATASED

1 VARIABLE × COLUMN
1 OBSERVATION × ROW
1 TABLE × DATATYPE
LINKING INDICATORS

DATASET PROCESSED, CLEAN & READY FOR ANALYSIS

3) CODEBOOK (DOCUMENTATION)

EXPLAINING YOUR DATASET,
HOW DATA IS GATHERED & MEASURED

4) METHODOLOGY (RECIPE)

EXPLICIT INSTRUCTIONS
SOFTWARE VERSION
RECIPE TO OBTAIN TIDY DATASET
SCRIPTS YOU USED

MAKING BIG DATA AS SMALL AS POSSIBLE AS QUICK AS POSSIBLE TO ENABLING SHARING

PLOTTING : UNDERSTANDING PROPERTIES & CHARACTERISTICS OF DATASET

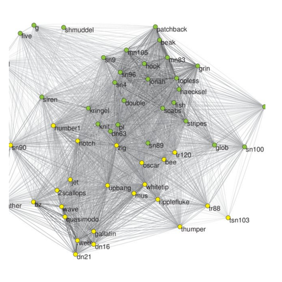
INTERACTIVITY ALLOWS DISCOVERY

SHOW AS MUCH OF THE DATA AS YOU CAN IN YOUR PLOTS

PLOT REPLICATES

MA PLOTS (BLAND-ALTMAN)

RIDICULOGRAMS



A NETWORK PLOT THAT LOOKS BEAUTIFUL BUT COMUNICATES VERY LITTLE INFORMATION AND APPEARS IN THE COVER OF MAJOR SCIENTIFIC PUBLICATIONS (NATURE, SCIENCE)

YOUR PLOTS SHOULD BE INTERPRETABLE

EXPERIMENTAL DESIGN

N = # OF MEASUREMENTS

$$N = \frac{\$ \text{ YOU HAVE}}{\$ / \text{MEASUREMENT}}$$ TERRIBLE IDEA

KEY TO SAMPLE SIZE:
UNDERSTANDING VARIABILITY

POWER: PROBABILITY THAT IF THERE IS A REAL DIFFERENCE IN YOUR DATA YOU'LL BE ABLE TO DETECT IT

N SAMPLE SIZE
Δ DIFFERENCE
δ STANDARD DEVIATION
$\mathbb{P}$ POWER

POWER.T.TEST $(N, \Delta, \delta) = \mathbb{P}$

HOW MANY SAMPLES DO WE NEED TO REACH A DESIDERED POWER $\mathbb{P}$

POWER.T.TEST $(\Delta, \delta, \mathbb{P}) = N$

EX: $\Delta = 5$, $\delta = 10$, $\mathbb{P} = 0.8$ $\rightarrow N = \lceil 63{,}8 \rceil = 64$

WE CAN EVEN DO ONESIDED ANALYSIS IF WE KNOW IN ADVANCE THAT THE EFFECT WILL ALWAYS BE EITHER HIGHER OR LOWER

POWER CALCULATIONS HYPOTESIS BASED ON WHAT WE THINK THE EFFECT SIZE MIGHT BE

GENOMIC VARIABILITY

- PHENOTYPIC VAR.
- MEASUREMENT ERRORS
- NATURAL BIOLOGICAL VAR.

BETTER TECH = LOWER VARIABILITY

NATURAL BIOLOGICAL VAR. CANNOT BE ELIMINATED WITH BETTER TECH

STATISTICAL SIGNIFICANCE

ARE OBSERVED DIFFERENCES REAL AND REPRODUCIBLE?

T-STATISTIC

$$T_0 = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_y^2}{N_y} + \frac{S_x^2}{N_x}}}$$

SCALING $\bar{y} - \bar{x}$ BY THEIR UNIT OF VARIABILITY

$T_0 \ll$ LESS LIKELY
$T_0 \gg$ MORE LIKELY

MULTIPLE TESTS

P-VALUES ARE NOT DESIGNED FOR MULTIPLE TESTS

P-VALUES ARE UNIFORMLY DISTRIBUTED

- FAMILY-WISE ERROR RATE

$\mathbb{P}(\# \text{ FALSE POSITIVES} \geq 1)$

- FALSE DISCOVERY RATE

$$\mathbb{E}\left[\frac{\# \text{ FALSE POSITIVE}}{\# \text{ DISCOVERIES}}\right]$$

PUBLICATION BIAS

IT'S IMPORTANT TO REPORT NEGATIVE RESULTS EVEN IF YOU CAN'T PUBLISH THEM IN THE BEST JOURNALS

AVOID P-HACKING!

CONFOUNDING

SHOE SIZE   LITERACY
AGE

PARAMETERS RELATED TO OTHER

BATCH EFFECT

DISTANT DATES
DIFFERENT METODOLOGY
DIFFERENT TECHNOLOGY
INCOMPATIBLE DATASETS

RANDOMIZATION HELPS TO ELIMINATE CONFOUNDERS

BREAKING DOWN RELATIONSHIP WITH CONFOUNDING VARIABLE