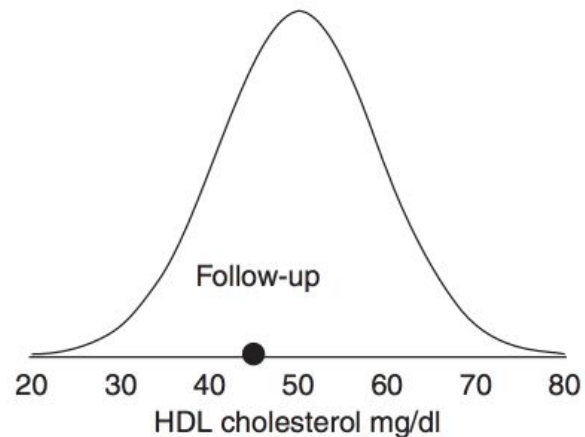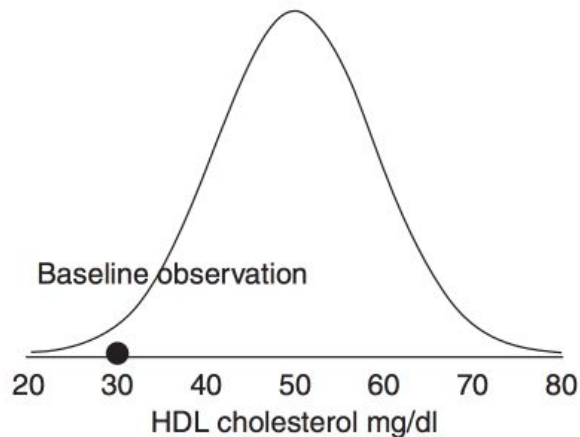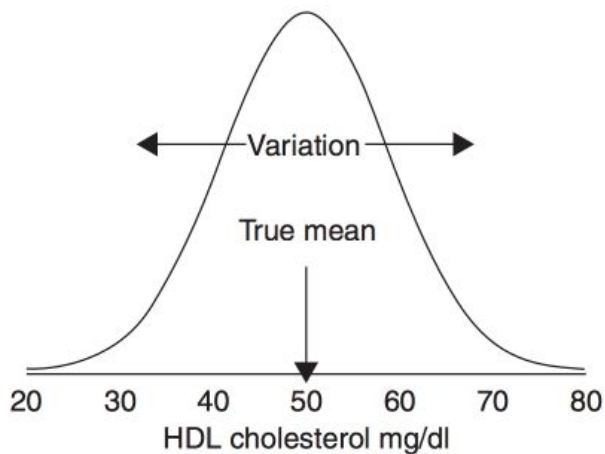# Topic 3: Circularity & Regression

Lectures 6 & 7

- Regression to the mean
- Stopping rules
- Circular analysis

Regression to the mean

Regression to the mean (RTM) is a statistical phenomenon that is characterized by the fact that unusually large or small measurements tend to be followed by measurements that are closer to the mean.
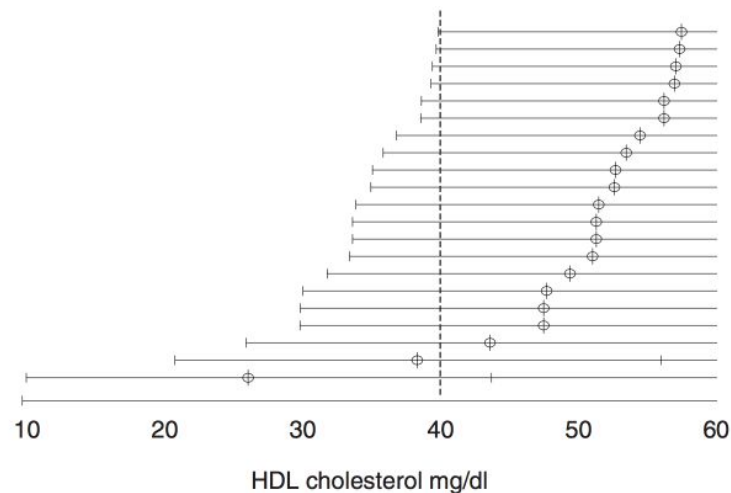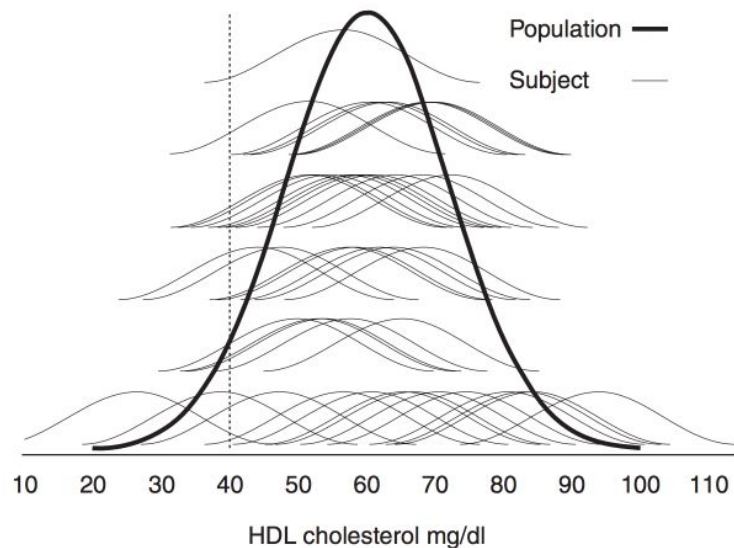
Regression to the mean

Regression to the mean (RTM) is a statistical phenomenon that is characterized by the fact that unusually large or small measurements tend to be followed by measurements that are closer to the mean.

- Occurs when repeated measurements are made on the same subject or unit of observation.
- Can make natural variation in repeated data look like real change.
- Happens because values are observed with random error (non-systematic variation like random measurement error or random fluctuations in a subject).
- There is almost no data without random error → makes RTM a common phenomenon.

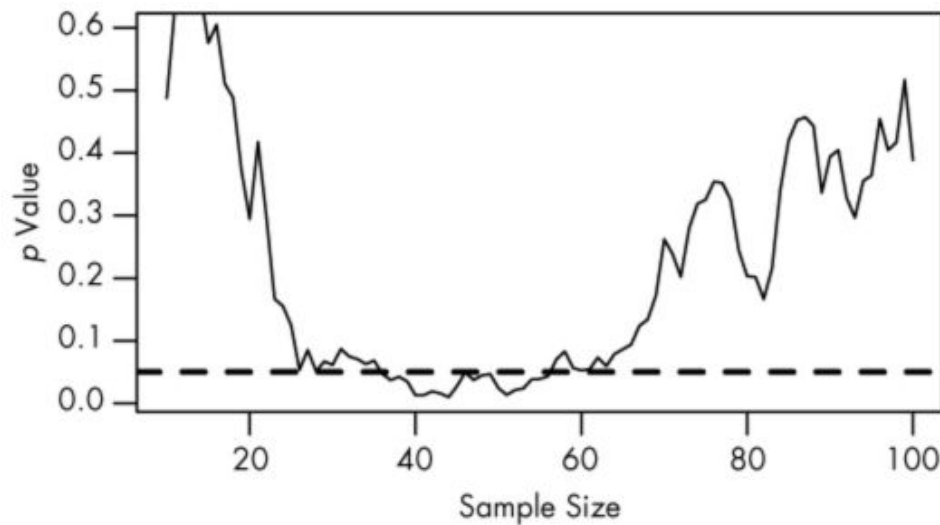# Regression to the mean

- Effect of RTM is compounded by categorizing subjects into groups based on baseline measurements.
- Variability in individual measurements > variability in the true means → Attenuation of association (regression dilution bias).

Regression to the mean

Longitudinally tracking the effect of drug.

- Terminate the study early if there if it is clear that the drug has an effect.
- In fact, it is *unethical* to withhold the drug from the control group.

Issues:

- Null hypothesis should take varying group size into account.
- Truth inflation (lucky patients, not brilliant drugs): stopped trials exaggerate their effect by 29% more than trials that run their full course.

Regression to the mean

Longitudinally tracking the effect of drug.

- Terminate the study early if there if it is clear that the drug has an effect.
- In fact, it is *unethical* to withhold the drug from the control group.
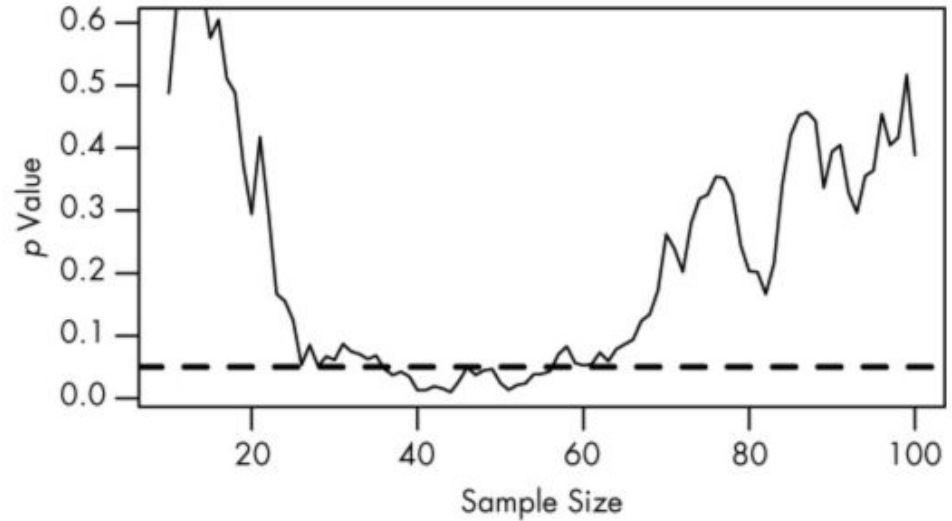
- Many published studies do not publish their original intended sample size or the stopping rule
- Preregistration!

Regression to the mean

Assigning a narrative or causal reasoning from observed data is often very hard:

- Galton's observation: Tall parents had (on average) children who were shorter than them, and short parents had (on average) children who were taller than them.

- "Norway had a great first jump; he will be tense, hoping to protect his lead and will probably do worse"; "Sweden had a bad first jump and now he knows he has nothing to lose and will be relaxed, which should help him do better.

- Why it is that highly intelligent individuals tend to marry those who are less intelligent than they are?

- Depressed children treated with an energy drink improve significantly over a three-month period.

Circular analysis & Double-dipping

Statistical analysis is often exploratory: no hypothesis is advance.

- Collect data → Poke around to see if there's something interesting → New hypotheses → Perform new experiments / Collect new data → Test the hypotheses.

- Collect data → Poke around to see if there's something interesting → New hypotheses → Take the subset of the original data that appears to show signal →Test the hypotheses.

  - Double-dipping → truth inflation.

  - Happens all the time in neuroimaging (apparently 40% of the literature), genetics, epidemiology.

Circular analysis & Double-dipping

- Collect data → Poke around to see if there's something interesting → New hypotheses → Take the subset of the original data that appears to show signal → Test the hypotheses.

  - Null hypothesis based on random chance is wrong at the final stage.

  - Only signals with the strongest random noise make it into further analysis.

- Mitigating this problem:

  - Split data in half; Reduces power.

  - Choose hypotheses based on prior knowledge.

Topic 4: Descriptive statistics & visualization

Lectures 8 & 9

- Descriptive statistics

- Spurious correlations

- Visualization challenges