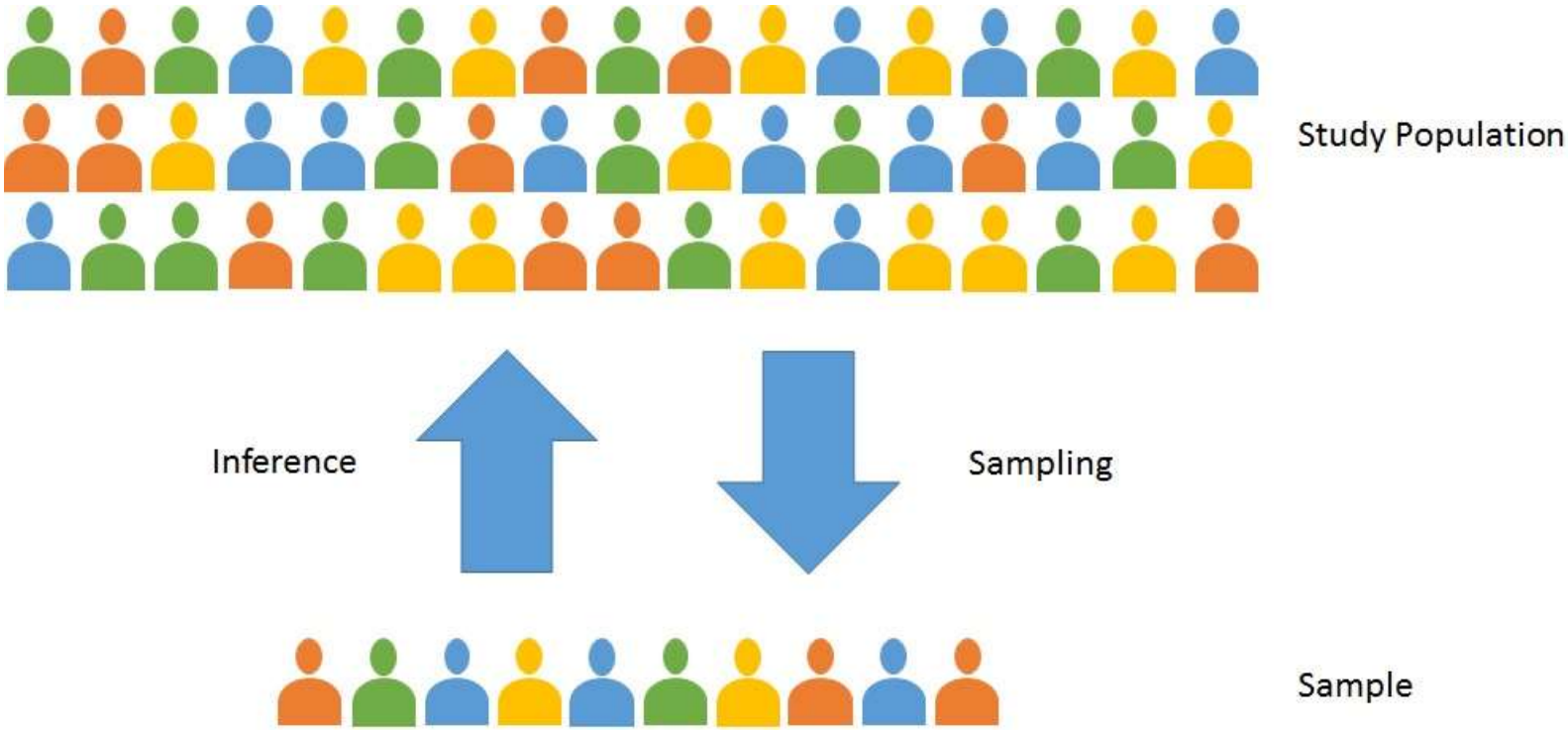
Clinical Trial Designs Frequentist and Bayesian approaches

February, 2023

Márcio Augusto Diniz, PhD
Madhu Mazumdar, PhD

Mount Sinai

Introduction

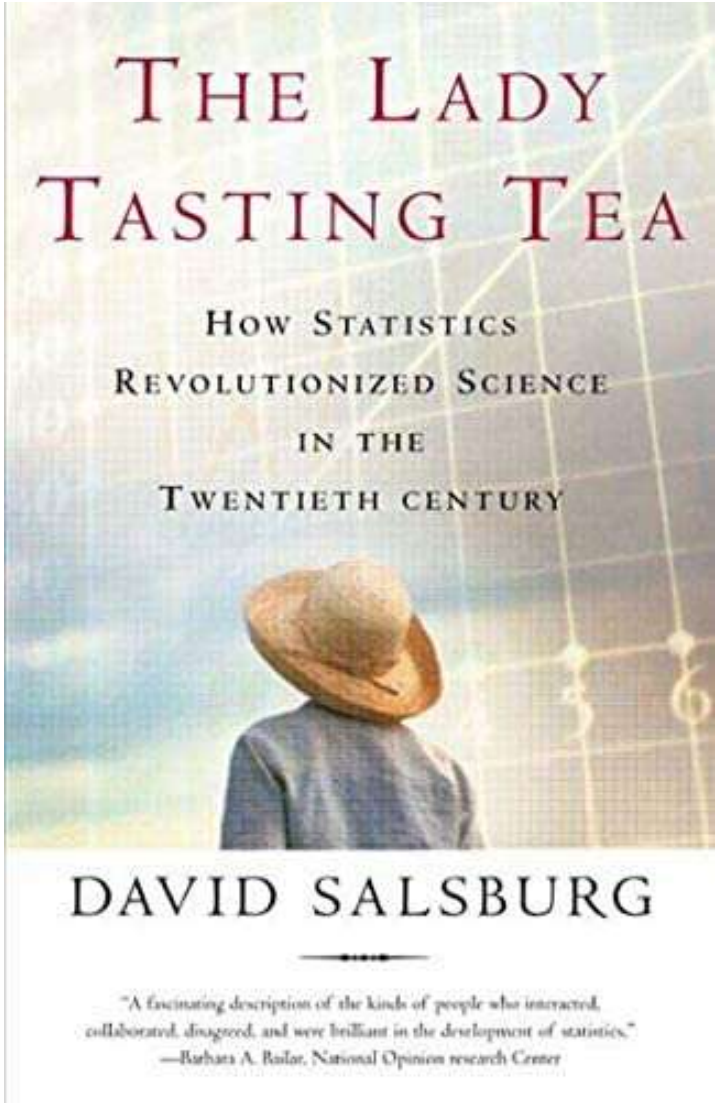


- Currently, inferences can be conducted follow two approaches: frequentism and Bayesianism.

Classical Statistical Methods

Classical Statistical Methods

- It is based largely on the work of Karl Pearson (1857 - 1936) and Ronald Fisher (1890 – 1962).
- Karl Pearson introduced the standard deviation, Chi-squared test, p-value and regression methods;
- Ronald Fisher introduced the Fisher's Exact test, analysis of variance and popularized the p-value.
- A mathematical framework was proposed by Jerzy Neyman (1894 – 1981) and Egon Pearson (1895 – 1980).



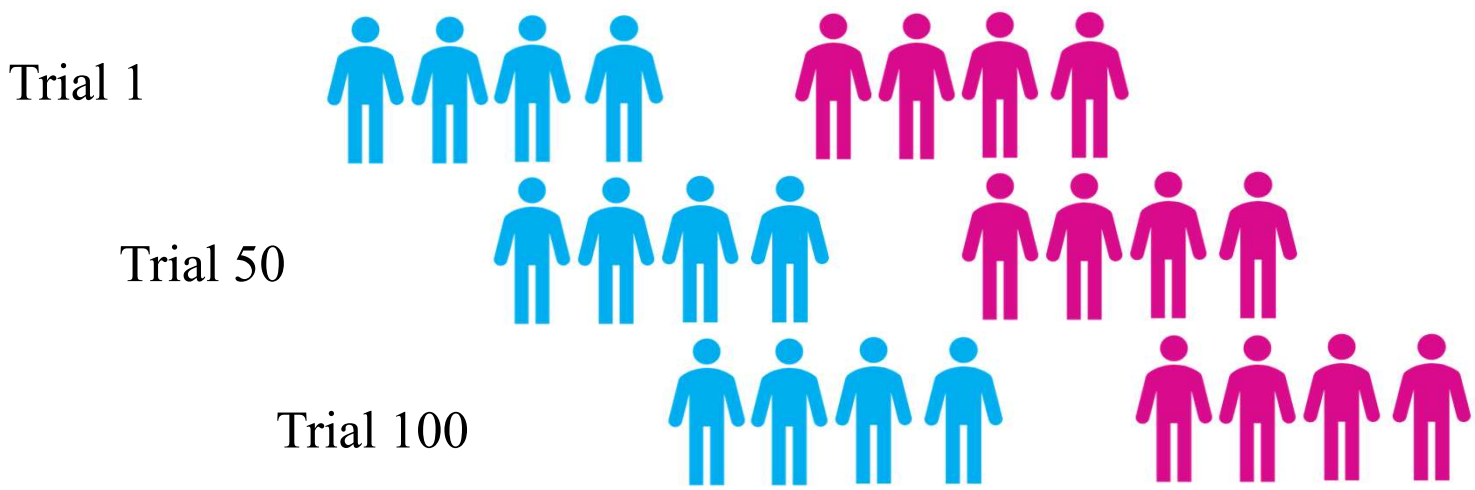
Classical Statistical Methods

- We are interested in populational quantities, known as parameters, that are unknown and fixed: prevalence of disease, treatment effect of a new drug, association between exposure and outcome;

- Experiments are conduct to collect data to estimate the parameters;
- However, any procedure in statistics carries uncertainty as we have limited information provided by our sample;
- How can we measure this uncertainty?
    - In the classical approach, we can measure this uncertainty when we repeat a statistical procedure in all possible samples that could have been sampled from the study population.

Classical Statistical Methods

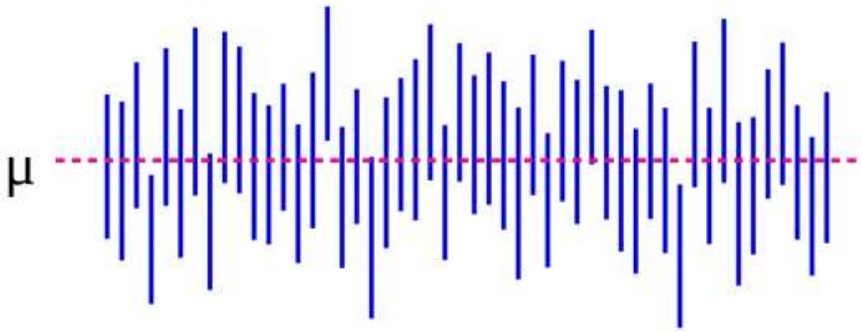
- Example: Our interest is to compare a new drug with a control group to evaluate whether there are differences in a biomarker. Data will be collected from 100 patients that will be equally randomized between drug and control arms.
- The classical approach assumes that this trial could be repeated several times under the same conditions, in other words, there is uncertainty about the sample.



Classical Statistical Methods

- 95% Confidence Interval (CI) for treatment effect: 2 [0.04 to 3.95];
- How do you interpret a 95% CI?
- We cannot state that there is 95% of probability that the treatment effect is between 0.04 and 3.95 because the parameter treatment effect is fixed, in other words, we cannot associate probability to parameters.

- The correct interpretation is based on the sampling uncertainty: If we calculate a 95% confidence interval for the treatment effect in each trial, then 95 out of 100 confidence intervals will contain the true and unknown treatment effect.



Classical Statistical Methods

- Testing the hypotheses:

Null: Treatment effect is negative or zero
Alternative: Treatment effect is positive

- Uncertainty: 5% type I error and 20% type 2 error (80% power);
    - How do we interpret those measures of uncertainty?

Classical Statistical Methods

		Truth	
		H_0	H_1
Decision	Not Reject H_0	No error	Type II
	Reject H_0	Type I	No error

- Identify a false treatment effect - false positive: obtaining a statistically significant p-value in 5 out 100 trials <u>under the scenario that there is no treatment effect</u>.
- Miss a true treatment effect - false negative: obtaining a non-statistically significant p-value in 20 out 100 trials <u>under the scenario that there is a treatment effect</u>.

Classical Statistical Methods

- Classical statistical methods are also known as frequentist statistical methods because they allow us to make conclusions regarding all possible samples without repeating a trial several times;

- In order to make such frequentist conclusions, calculations are heavily based on mathematical assumptions (large samples, average, normality), therefore, procedures are pre-specified such that calculations can be performed.

Frequentist Trial Designs

Example – Simon's two stage design

- Study Design: A single-arm two-stage study for a categorical outcome
- Hypotheses:
  - Null: Complete response after new drug is at most 10%
  - Alternative: Complete response after new drug is at least 30%
- Uncertainty:
  - Type I error: 5%
  - Type II error: 20% (power = 80%)
  - It minimizes the sample size when the drug is not effective.

Example – Simon's two stage design

- Stage 1:
  - Sample size: 11 patients

  Decision rules:
  - If one or fewer complete responses are observed, then the study will stop and drug will be declared futile;
  - If two or more complete responses are observed, the study will continue to stage 2;

- Stage 2:
  - Sample size: 16 patients

  Decision rules:
  - If at least 5 complete responses are observed among the total of 27 patients (11 + 16), then the drug is declared effective.

- When the drug is not effective:
  - Probability of early termination is 0.69;
  - The expected sample size is 15.

Example - Two stages using O'Brien-Fleming boundaries

- Study Design: A two-arm two-stage study for a continuous outcome

- Endpoint:
  - High values indicate good prognosis;
  - Minimum Clinically Important Difference = 5 units.

- Hypotheses:
  - Null: Biomarker average is the same in intervention and control groups;
  - Alternative: Biomarker average in different between the intervention and control groups.

- Uncertainty:
  - Type I error: 5%
  - Type II error: 10% (power = 90%)

Example - Two stages using O'Brien-Fleming boundaries

- O'Brien-Fleming boundaries indicates how type I and II errors should be spent in each stage.

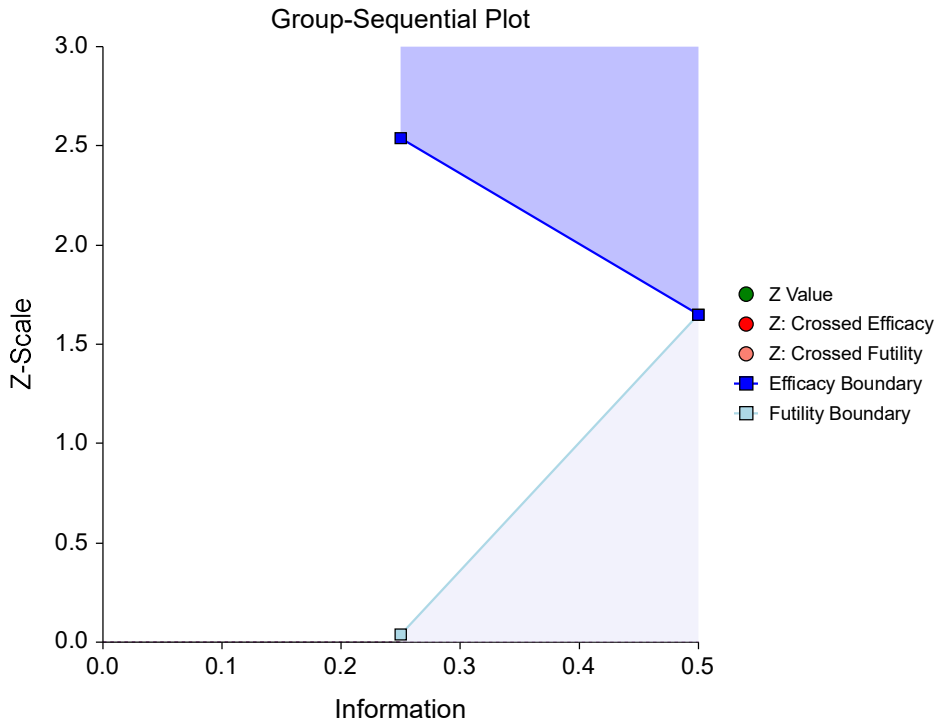
- Stage 1:
  - Sample size: 50 for each group
  - Type I error: 0.56%
  - Type II error: 2.0%

- Stage 2:
  - Sample size: 50 for each group
  - Type I error: 4.44%
  - Type II error: 8.0%

Example - Two stages using O'Brien-Fleming boundaries

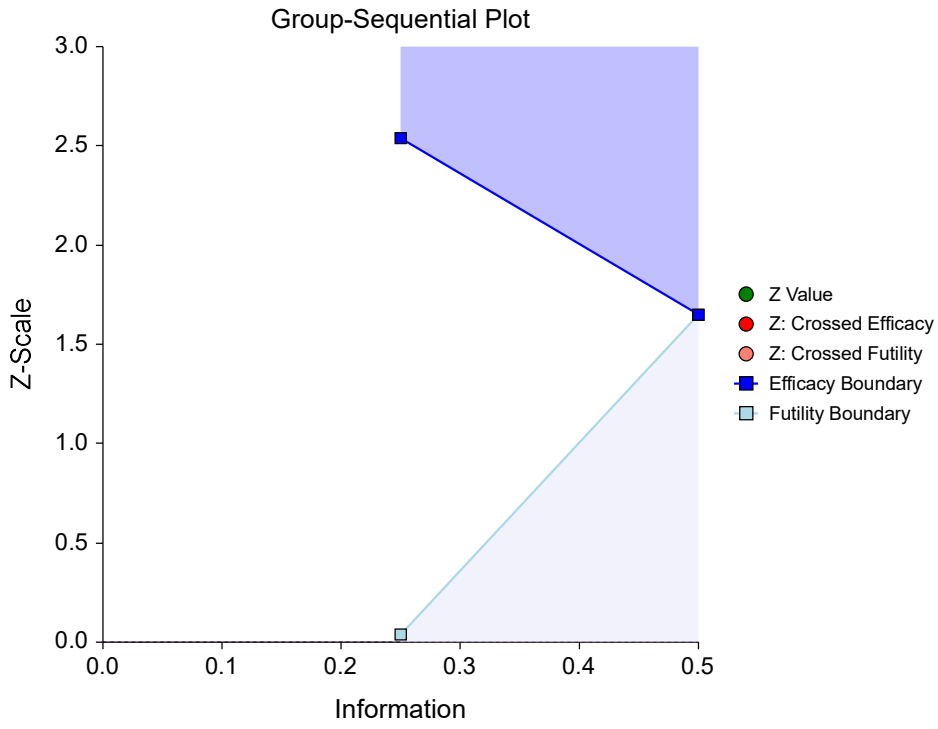
Decision rules:

- At the end of stage 1, a standardized difference between groups (Z-Score or Z-Scale) is calculated. Then,
  - If the Z-score is below the Futility boundary, then intervention is declared futile;
  - If Z-score is above the Efficacy boundary, then intervention is declared efficacious;
  - If Z-score is between Efficacy and Futility boundaries, the intervention is declared promising and the trial proceeds to stage 2.



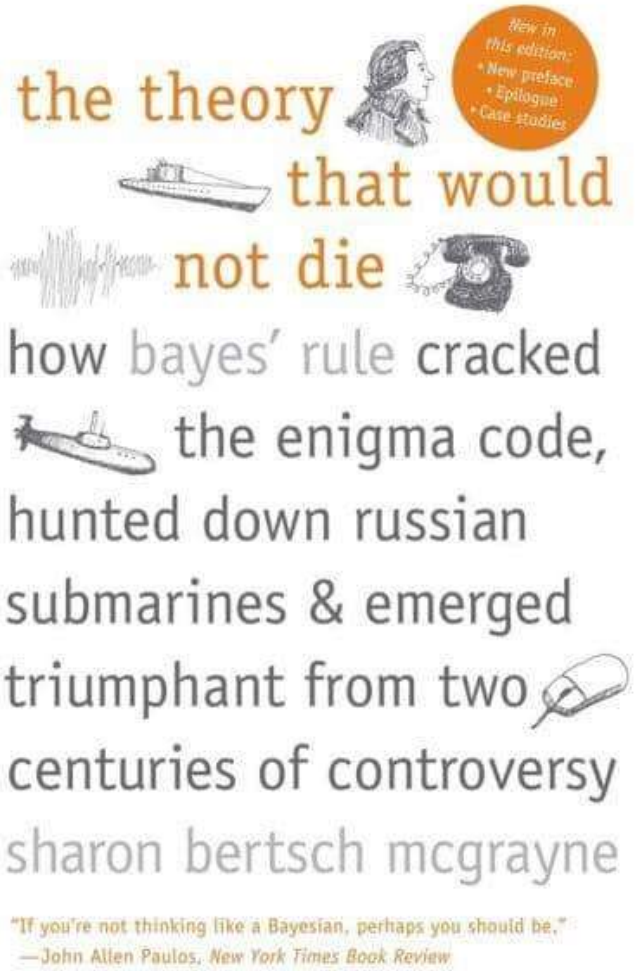
Example - Two stages using O'Brien-Fleming boundaries

- At the end of stage 2:
  - If the Z-score is below the Futility boundary, then intervention is declared futile;
  - If Z-score is above the Efficacy boundary, then intervention is declared efficacious;



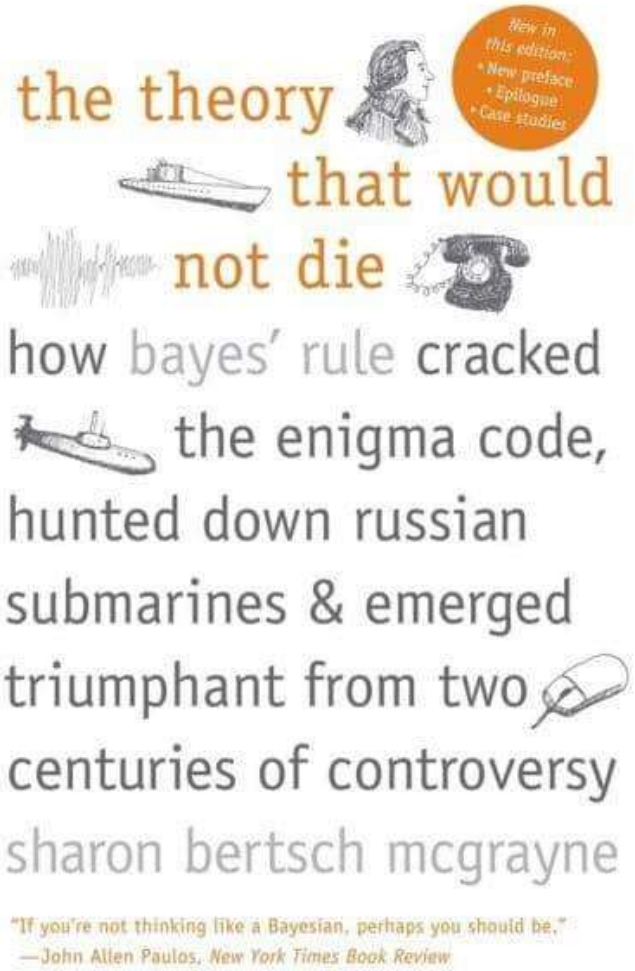
Bayesian Statistical Methods

Bayesian Statistical Methods



- It is based on the seminal work of Thomas Bayes (1701 – 1761);
- Bayes' system was: Initial Belief + New Data → Improved Belief;
- It was rediscovered and popularized by Pierre-Simon Laplace (1749 – 1827) who applied this approach to astronomy;
- Late in his life, Laplace discovered a mathematical result – the Central Limit Theorem - that led him to support the frequentist approach.

Bayesian Statistical Methods



- After Laplace's death, the use of Bayesian methods declined in science as the modern science could not be based on anything that was considered subjective.
- Nonetheless, Bayesian methods had continued to be used to solve practical problems with a wide applications during WWII, election polls from the 60s to 80s, etc.
- Only in the early 90s with the availability of computational power, the Bayesian methods have become popular again.

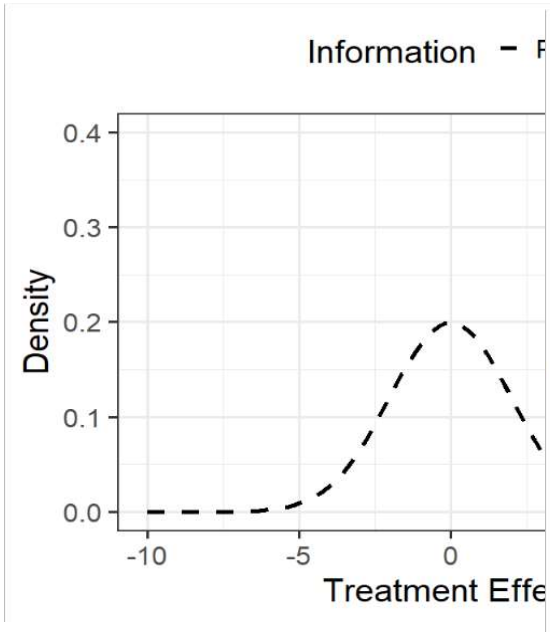
Bayesian Statistical Methods

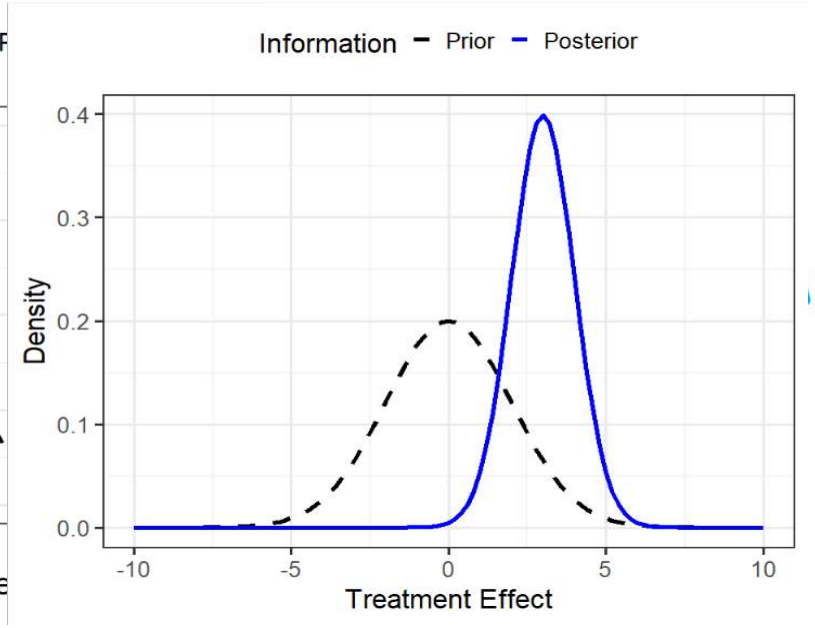
- We are in interested in populational quantities, known as parameters, that are unknown and random: prevalence of disease, treatment effect of a new drug, association between exposure and outcome.

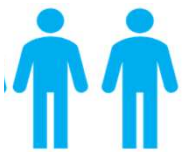
- Experiments are conduct to collect data to estimate the parameters;
- As previously, any procedure in statistics carries uncertainty as we have limited information provided by our sample.
- How can we measure this uncertainty?
    - We can measure this uncertainty when we considered all possible values for the parameters of interest with their associated probabilities.

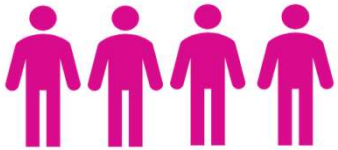
Bayesian Statistical Methods

- Example: Our interest is to compare a new drug with a control group to evaluate whether there are differences in a biomarker. Data will be collected from 100 patients that will be equally randomized between drug and control arms.
- The initial belief about the parameter is known as prior distribution;
- The Bayesian approach assumes that the only sample is the observed sample;
- Once data is observed, the updated belief is known as posterior distribution.



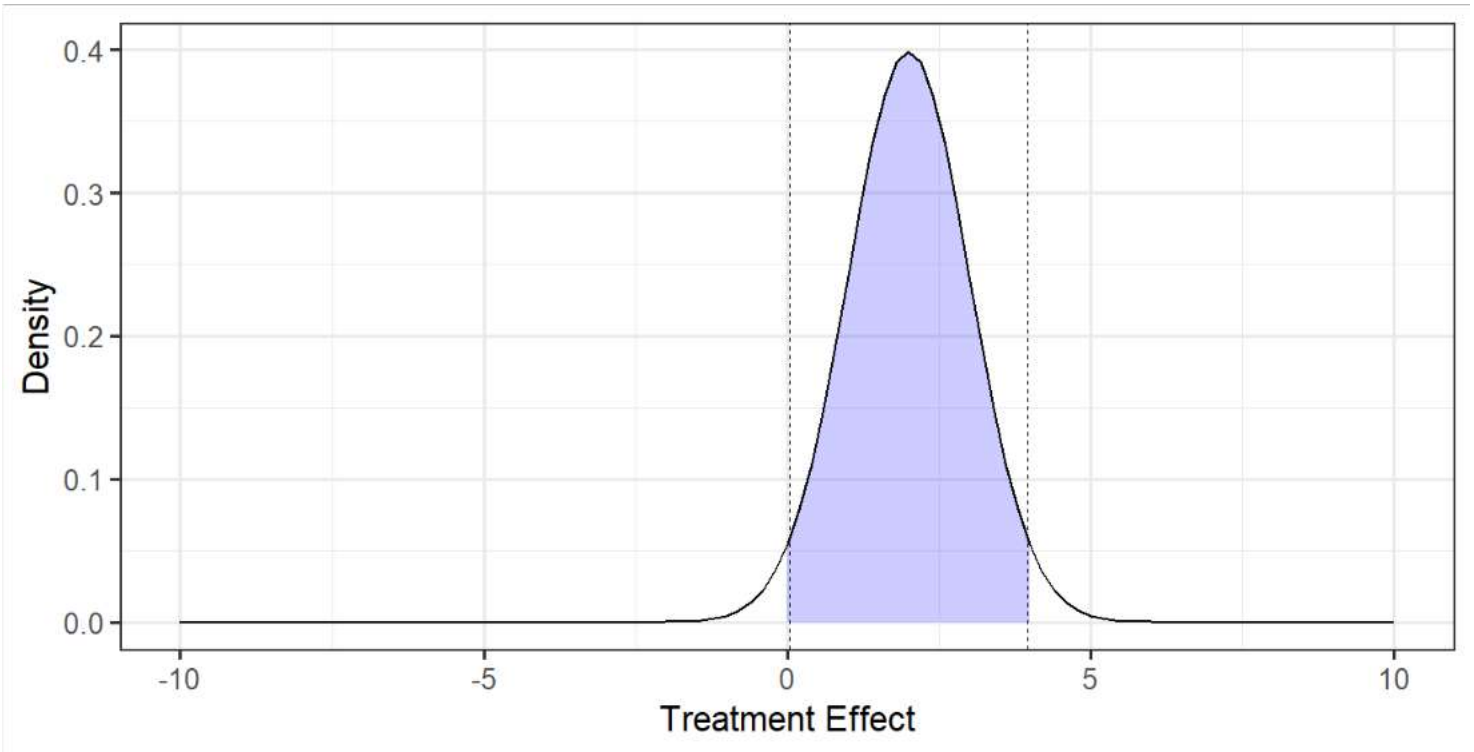






Bayesian Statistical Methods

- 95% Credibility Interval (CI) for treatment effect: 2 [0.04 to 3.95];
- How do you interpret a 95% CI?
- Now, we can state there is 95% of probability that the treatment effect is between 0.04 and 3.95.

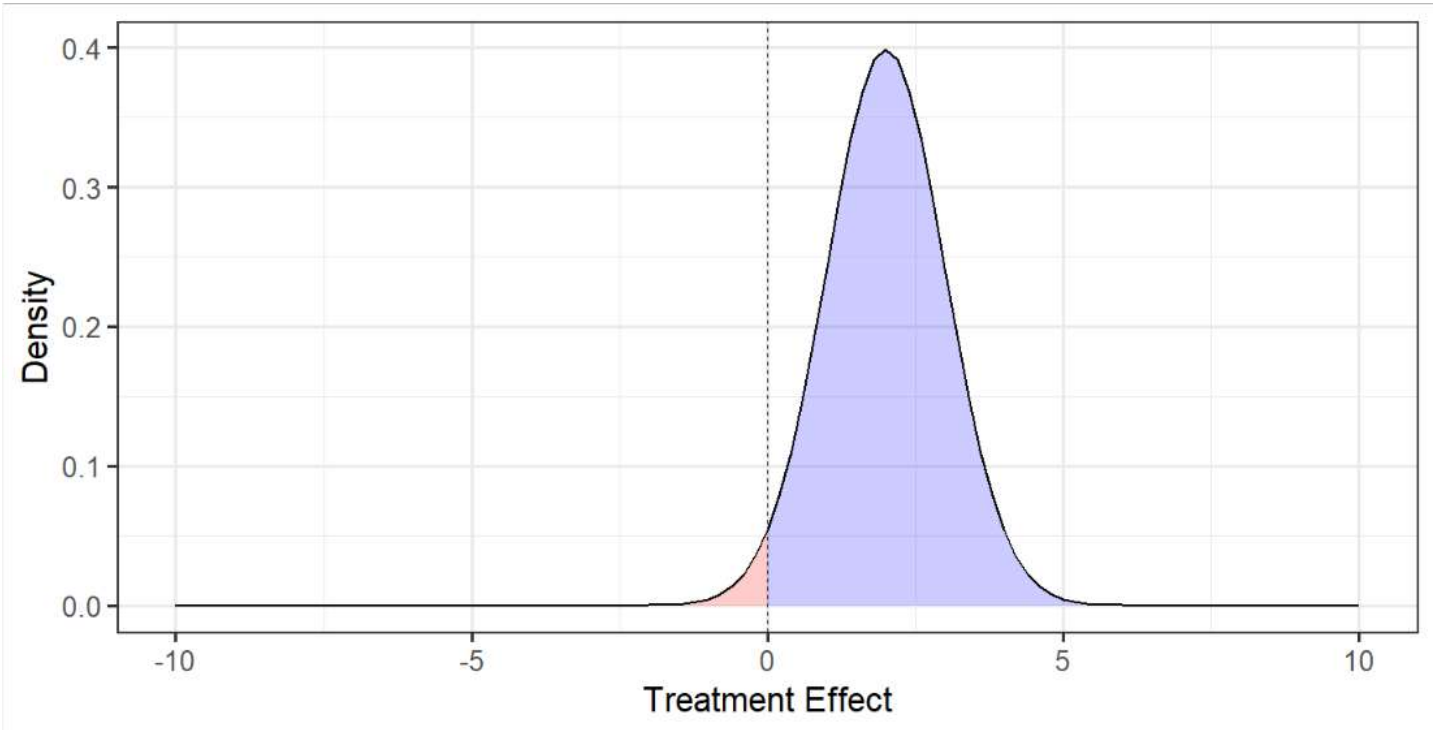


Bayesian Statistical Methods

We are interested in testing the hypotheses:

Null: Treatment effect is zero or less
Alternative: Treatment effect is greater than zero

- Probability(Null hypothesis) = 0.0228
- Probability(Alternative hypothesis) = 0.972



Bayesian Statistical Methods

- Based on the posterior distribution (uncertainty about the parameter after data is collected), decision rules or procedures can be created as a block building game.

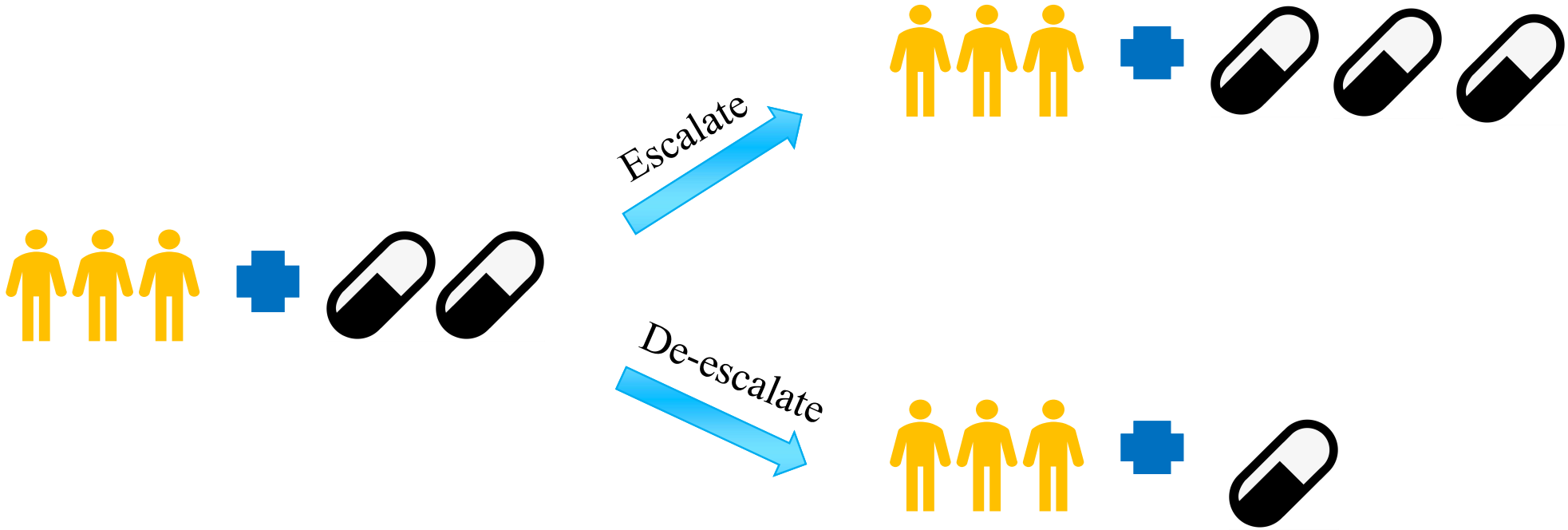
- After computational power has become widely available, frequentist properties of Bayesian procedures can be calculated based on computational simulations.

Bayesian Trial Designs

Example: Phase I clinical trials

- In phase I clinical trials, investigators want to identify the maximum tolerable dose (MTD) for a cytotoxic agent, in other words, a dose that has an acceptable toxicity rate.
- As it is the first study of a new drug in humans, sample sizes are limited, and the study is conducted by stages.

Example: Phase I clinical trials



- Based on the results, investigators decide how to escalate/de-escalate the dose for the next cohort of patients:

Initial Belief + New Data → Improved Belief;

Example: Phase I clinical trials

- There are two classes of methods that establish decision rules to escalate/de-escalate doses based on different criteria:
    - Model-based designs:

    Continual Reassessment Method (CRM),

    Escalation with Overdose Control (EWOC);

    - Model-assisted designs:

    Bayesian Optimal Interval (BOIN),

    Modified Toxicity Probability Interval (mTPI);

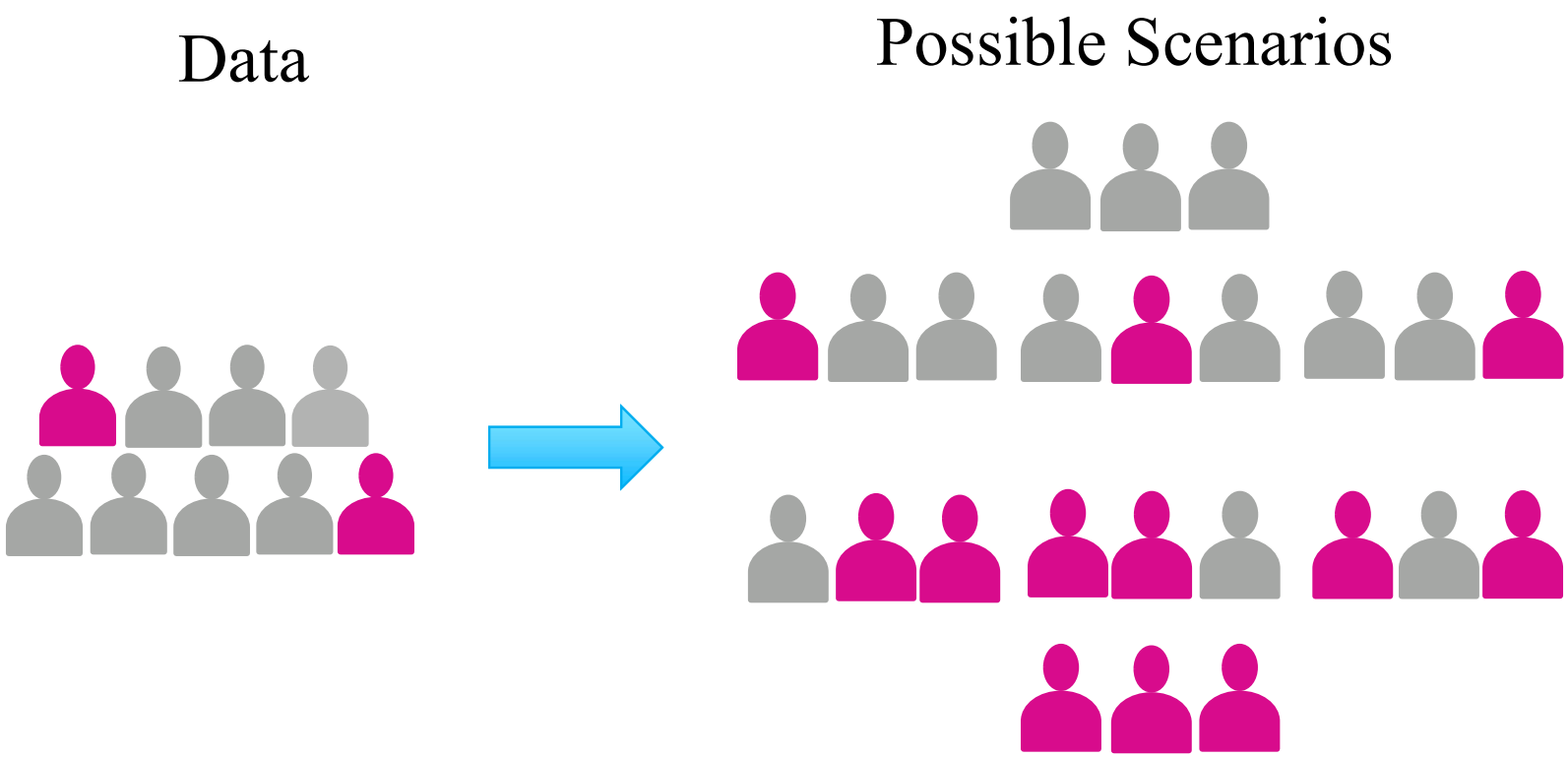
Example: Predictive Probability for Futility

- A single-arm study with an interim analysis to test the hypotheses:
  - Null: Complete response rate with the new drug is at most 10%
  - Alternative: Complete response rate with the new drug is at least 30%

- Ah-hoc Decision Rule: If we observe at least **4** CR out **12** patients, then the drug is declared efficacious;
- Interim Analysis will be conducted after **9** patients;

Example: Predictive Probability for Futility

- After 9 patients, we have observed 2 complete responses.
- What are the chances that the drug will be declared effective?



Example: Predictive Probability for Futility

- The trial will be successful only if 2 or 3 CR are observed.

Scenario	True CR Rate				Trial Conclusion
Out of 3 patients	0.1	2/9	0.3	0.9	Drug is effective
0 CR	0.3	0.47	0.34	0.01	No
1 CR	0.24	0.40	0.44	0.02	No
2 CR	0.03	0.12	0.19	0.24	Yes
3 CR	0.001	0.01	0.03	0.73	Yes

- Which scenario should be considered?
- Under the Bayesian approach, the predictive probability is a weighted average over all possible values of CR rate such that scenarios closer to the data have large weights while scenarios far away from the data have low weights.
  - Predictive probability that trial will be successful = 0.178

Example: Predictive Probability for Futility

- What are the frequentist properties of this trial design?

- Simulating 1000 trials under different scenarios:
  - Probability of false positives (type I error) under a scenario where the percentage of CR is 10%;
  - Probability of early termination under a scenario where the percentage of CR is 10%;
  - Expected sample size under a scenario where the percentage of CR is 10%.
  - Probability of false negatives (type II error) under a scenario where the percentage of CR is 30%;

Example: Phase I clinical trials

- What are the frequentist properties of this trial design?

- In a simulated trial, we can:
  - Check whether the selected MTD is the true MTD;
  - Calculate the percentage of patients receiving overly toxic doses;
  - Calculate the Toxicity rate.

- Simulating 1000 trials, the frequentist properties are
  - Probability of the selected MTD to be the true MTD;
  - Average percentage of patients receiving overly toxic doses;
  - Average toxicity rate.

Example: Response adaptive randomization (RAR)

1. Patients are initially equally randomized to control, drug A and drug B groups.

2. After every **20** patients are enrolled in each group, we can calculate the probability that treatment effect is greater than zero for drug A compared to control and drug B compared to control:

- p_A = probability that drug A is better than control arm = 0.45;
- p_B = probability that drug B is better than control arm = 0.9;

Example: Response adaptive randomization

3. Update probabilities of randomization for each group based on the probability that each dose performs better than the control group.

4. At the end of trial, declare that a dose is effective if pA or pB are very large, in other words,

$$p_A > \mathbf{0.9} \text{ or } p_B > \mathbf{0.9}.$$

Example: Response adaptive randomization

- What are the frequentist properties that can be calculated for this trial design?

- Simulating 1000 trials under different scenarios:
  - the probability to declare a drug is better than control in a scenario where drug A and drug B are equal to the control group (false positive);
  - the probability to declare a drug as ineffective in a scenario where drug A or/and drug B are better than control (false negative).

Concluding Remarks

Frequentist trial designs

- Advantages:
  - Strong control of false positives results which is often required by regulatory agencies;
  - Software is easily available;
  - It does not require a lot of input from investigators;

- Disadvantages:
  - There is not much flexibility with pre-specified procedures;
  - It assumes that the average response is approximately normally distributed which is true only with large sample sizes or normally distributed samples;
  - It is not possible to incorporate information from historical data or subjective information from the investigator.

Bayesian trial designs

- Advantages:
  - Incorporate sequential learning;
  - Use of predictive probabilities of future results;
  - Suitable for small sample sizes.
- Disadvantages:
  - It does not strongly control type I error;
  - It requires a lot of input from investigators;
  - It requires more time to design a trial and involvement of statistician during the trial;

Questions?