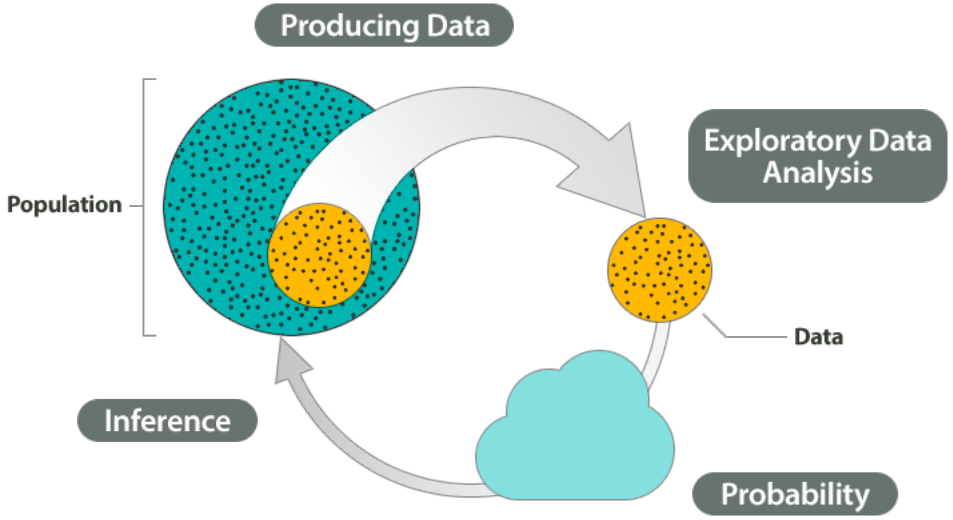
BMR 617

Samples and Populations
Confidence Intervals of Proportions

Marshall University Joan C. Edwards School of Medicine

The “Big Picture”

- In a typical experiment/study:
  - Collect data from a “population”
  - Only collect a small sample of the possible data
  - Explore the data to understand the sample
  - Use probability to make inferences about the population (not just the sample)



Example: Pfizer COVID-19 vaccine trial

- As an example, consider the Pfizer COVID-19 vaccine trial we discussed last time
- A total of 36,523 participants were recruited to the trial
  - Essentially a random selection from the population
  - We will assume they are representative of the population as a whole
- Randomly assigned to one of two groups
  - One group received the vaccine, the other the placebo
- We compared the infection rate between the two groups

Sample vs population

- Last time, we explored the data from the trial and computed various measures indicating the effectiveness of the vaccine
- The relative risk for the vaccine group compared to the placebo group was 4.97%
  - Those vaccinated had only 4.97% of the risk of infection of those receiving the placebo
- However, this measure was computed for the *sample*
  - For the people in the study
- We are really interested in the relative risk for the *population*

What we'd really like to know

- What we really want to know is how effective the vaccine is in the entire population
- The experiment we want to do, but obviously can't, is to see how many people in the entire population get infected without the vaccine, then roll back time and vaccinate everyone, and see how many people get infected with the vaccine. Then pick the best course of action.
- Using a test sample is our best option
  - The question is "How representative of the data for the entire population are the data we get from the sample?"

Sources of error

- There are two distinct sources of error:
- Systematic bias
  - Systematic bias occurs when elements of the experimental design lead to an unrepresentative sample
- Random effects
  - Even with a perfect experimental design (which is unobtainable), random effects may lead to data from the sample which are not representative of the population as a whole

Systematic bias

- Clinical trials always have some systematic bias associated with them
- In the study design, we try to mitigate these as far as possible
- Acknowledge sources of bias, and try to limit them
- In the COVID-19 vaccine trial, who might be willing to participate?
- How might the willingness to participate affect the outcome of the trial?

Random effects

- Let's assume that the sample are truly representative of the population
- Participants were randomly assigned to either the placebo group or the vaccine group
- Some people are naturally more susceptible to the virus than others
  - E.g. the virus invades the body by binding to the ACE2 receptor
  - People naturally expressing more ACE2 are more likely to be infected
- Is it possible that, for example, the placebo group has many more people who are more highly susceptible to infection than the vaccine group?
  - What would happen in this case?

General statistical approach

- We start by assuming
  - There is some fixed, but unknown, measure for the entire population
    - E.g. the vaccine reduces risk in the general population by some fixed, but unknown, factor
  - We have eliminated (or accounted for) systematic biases
- We then measure the data in the sample
- And then compute some measure of confidence we have that the data from the sample is representative of the data from the population
  - Or some measure of how representative it is
  - The key concept here is a *confidence interval*

Example: data expressed as a proportion

- In a study of cardiovascular disease in West Virginia, 89 overweight (25<BMI≤30) subjects who had no history of cardiovascular disease were genotyped for the SNP rs5880 in the CETP gene. Of these, 75 (84%) were homozygous (GG) while 14 (16%) were heterozygous (CG).
  - We would like to know the proportion of our population who are homozygous (GG) at this locus.
  - What is the population?
  - Given the data, what is our best estimate of the proportion who are homozygous-GG at this locus?
  - How confident are we this is representative of the true proportion?

(West Virginia Medical Journal, Vol 108, January 2012)

Confidence Interval for Proportion Data

- The population is the set of all West Virginia residents who are overweight ($25<BMI\leq30$) and who have no history of cardiovascular disease (according to the criteria in the publication).
- From our sample data, the best estimate is that 84% of this population are homozygous (GG) at the rs5880 locus.
- To express our confidence in this estimate, we can state a confidence interval:

  We are 95% confident that the range 76.7% to 91.8% contains the true proportion of homozygous (GG) individuals in this population.

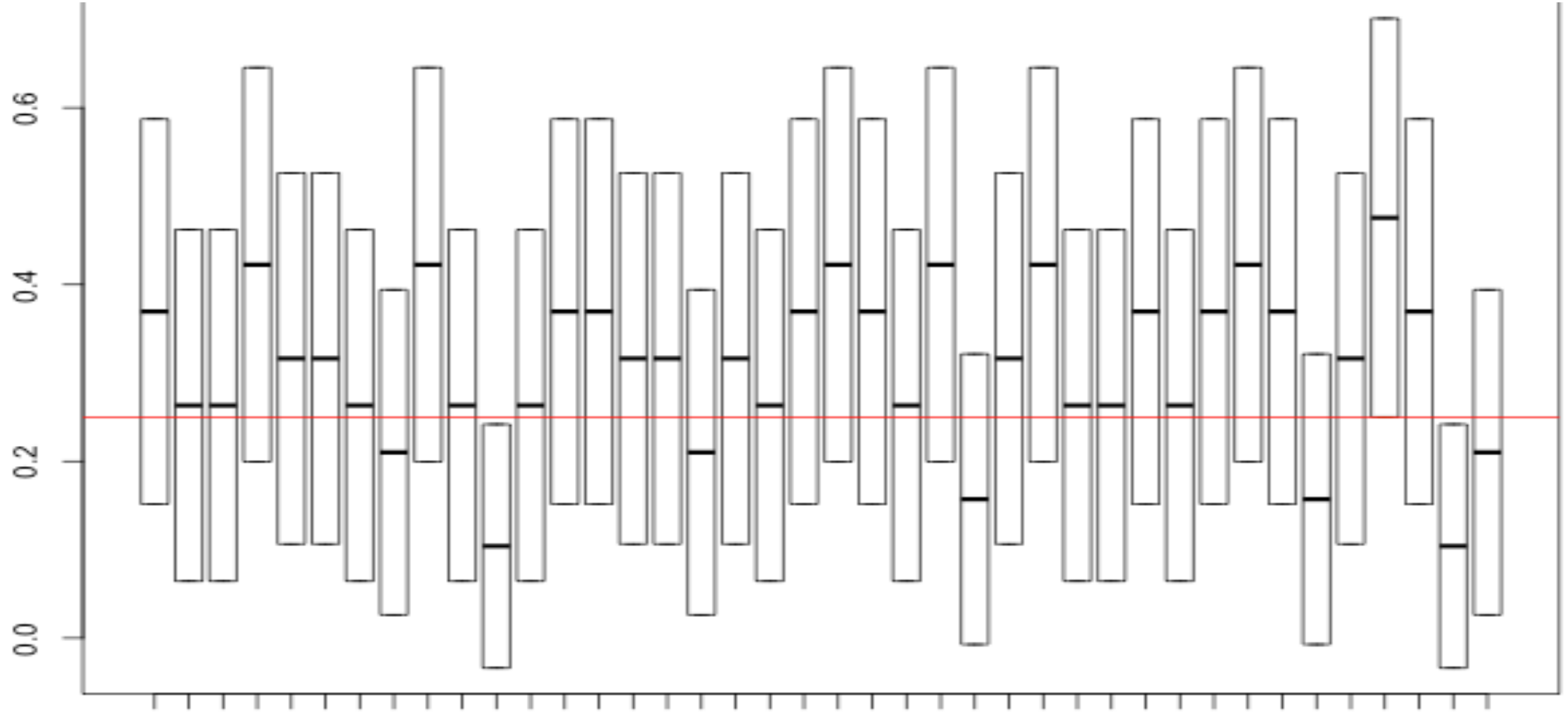
Understanding Confidence Intervals

- The logic of a confidence interval is often subtly misunderstood.
- It is a little logically misleading to say "There is a 95% chance the true population value lies between 76.7% and 91.8%”.
  - This implies the true population value is subject to random fluctuations.
  - It is the confidence interval that is subject to random fluctuations.
- It's better to say "There is a 95% chance that the interval [76.7%, 91.8%] includes the true population value."

A statistical experiment

- This statistical experiment may clarify how confidence intervals work.
- Take a bag with 25 red balls and 75 black balls. We know that the true proportion of red balls in the bag is 25%, but we will try to estimate this by sampling (and computing a confidence interval).
  - Draw 15 balls from the bag.
  - Compute the proportion of those that are red.
  - Calculate the 95% confidence interval for that proportion.
  - Replace the balls, and repeat 40 times.
    - So we end up with 40 confidence intervals

Confidence Interval Experiment Results



Confidence Interval Experiment Results (continued)

- Each time we compute a 95% confidence interval, it has a 95% chance of containing the true population value.
- Since we computed 40, we'd expect, on average, 38 of these to contain the true population value.
- In this case, we actually know the true population value is 0.25, so we can check.
  - It turns out - in this example - that only 37 of the 40 (92.5%) contain the true value.
- The more repeats you do, the closer you'll get to 95% of the intervals containing the true value.

Different Levels of Confidence

- What's special about 95%?
  - Nothing at all!
  - Commonly used only by tradition.
  - Could just as well compute 90% confidence intervals, 99% confidence intervals, or any other value up to 100%
    - Would the 90% confidence interval for our genotype data be wider or narrower than the 95% confidence interval?
    - What would the 100% confidence interval be for these data?
      - Is this helpful?

Confidence Intervals: Assumptions

- Formulae for calculating confidence intervals (coming soon) only work if certain conditions are true. These are the assumptions in the formulae.
  - The sample is random (or representative) of the population.
  - The observations in the sample are independent of each other.
    - What could violate this in our genotype study?
  - The measurements are accurate.

Violating assumptions

- In practice, these assumptions rarely hold exactly.
  - Important to minimize deviation from the assumptions.
  - Acknowledge any violations.
  - Violating the assumptions will make the confidence intervals too optimistic.
    - Too narrow.

Computing confidence intervals for proportions

- Computing confidence intervals for proportions is technically complex.
- No general agreement on the best way to do this.
- In R, we can use the `binom` package to compute several estimates of 95% confidence intervals for proportions:

```
install.packages('binom')
library(binom)
```

95% Confidence Interval for CETP genotype data

- Recall our genotype data for the CETP gene in overweight West Virginians
  - Of a sample of 89 people, 75 had the GG genotype
- The proportion in the sample is 75/89=0.8427
- To compute the 95% confidence interval, we can use `binom.confint(75, 89)`
- Which gives a table of values

	method	x	n	mean	lower	upper
1	agresti-coull	75	89	0.8426966	0.7518448	0.9051893
2	asymptotic	75	89	0.8426966	0.7670555	0.9183377
3	bayes	75	89	0.8388889	0.7622935	0.9112045
4	cloglog	75	89	0.8426966	0.7488898	0.9036636
5	exact	75	89	0.8426966	0.7501772	0.9112457
6	logit	75	89	0.8426966	0.7517204	0.9045688
7	probit	75	89	0.8426966	0.7552917	0.9065772
8	profile	75	89	0.8426966	0.7578246	0.9081181
9	lrt	75	89	0.8426966	0.7578267	0.9081307
10	prop.test	75	89	0.8426966	0.7467288	0.9082794
11	wilson	75	89	0.8426966	0.7531121	0.9039219

Specifying a method

- To specify a particular method, use the `methods` parameter:
  ```
  binom.confint(75, 89, methods='asymptotic')
  ```
- The output here is
  ```
      method  x  n       mean     lower     upper
  1 asymptotic 75 89 0.8426966 0.7432874 0.9421059
  ```
- The interpretation is:
  "We are 95% confident that the range 0.7433 to 0.9421 contains the true proportion of overweight West Virginians who carry the GG genotype for the CETP gene."

Changing the confidence level

- To change the confidence level, use the conf.level parameter
- For example, to compute the 99% confidence level, use binom.confint(75, 89, methods="asymptotic", conf.level=0.99)
- Without running this, do you expect the confidence interval to be wider (contain more values) or narrower?
  - Run it and check
- What about the 90% confidence interval?
- What should a 100% confidence interval give? Why? Try it.

Choosing the method

- The "asymptotic" method assumes errors around the estimate are normally distributed
  - We'll discuss what this means soon in the course
- This assumption can fail for proportions very close to 0 or 1
- In these cases, the "exact" method can be better
  - What does the "exact" method give for the confidence level of 100%?

Pfizer vaccine trial

- Recall the Pfizer vaccine trial data we examined
- In the trial, what proportion of those receiving the placebo became infected? What proportion of those receiving the vaccine became infected?
- What are the 95% confidence intervals for each of these proportions?
  - What method is appropriate here?
  - What are the interpretations of these?

		Treatment		
		Placebo	Vaccine	Total
SARS-CoV-2 status	Infected	162	8	170
	Not infected	18163	18190	36353
	Total	18325	18198	36523

Summary

- When working with a sample, we can use the sample to make *estimates* of statistical measures for the population
- E.g. we used our clinical trial sample to estimate the proportion of people in the population who would become infected with COVID-19 with or without a vaccine
- The *confidence interval* is a range of values we compute from our sample data which we can say contains the population value with a specified degree of confidence
  - Note the level of confidence is chosen, not calculated