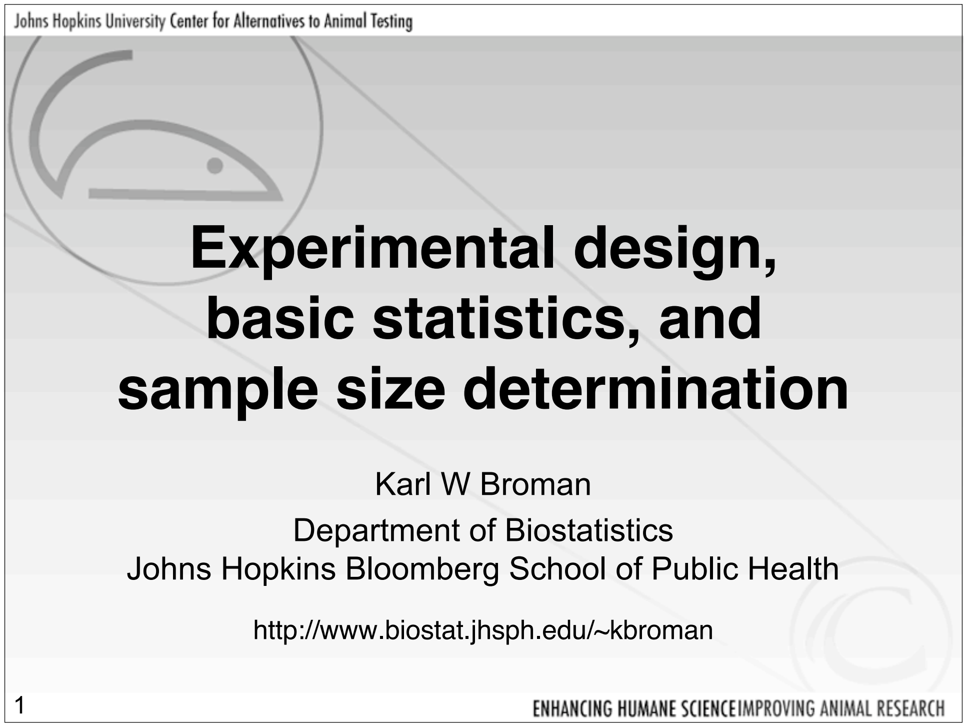


Johns Hopkins University Center for Alternatives to Animal Testing

Experimental design, basic statistics, and sample size determination

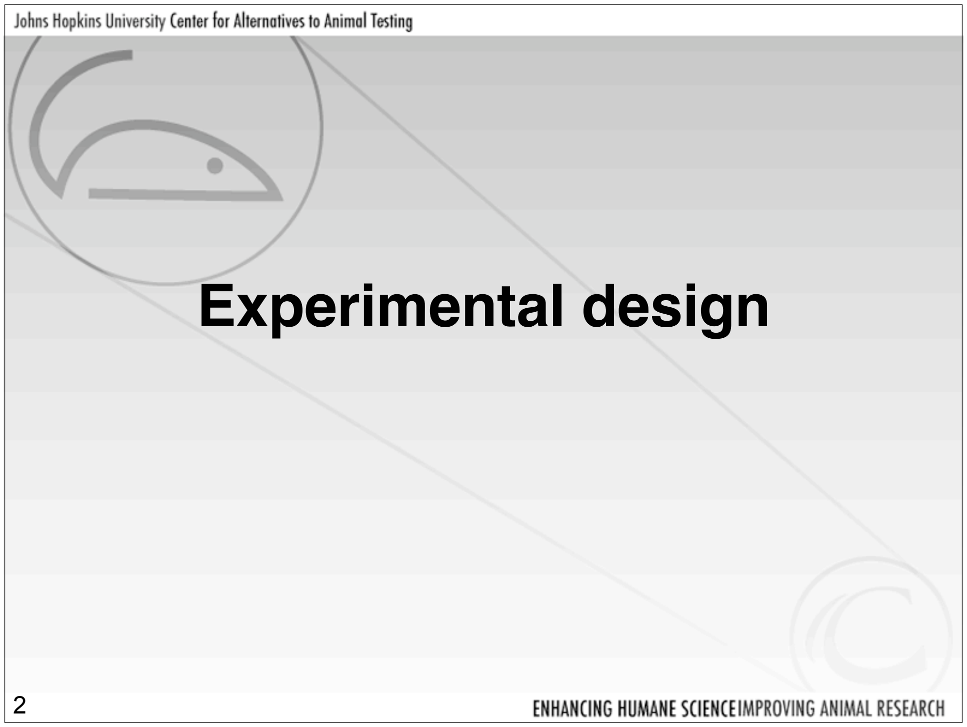
Karl W Broman

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

http://www.biostat.jhsph.edu/~kbroman

1

ENHANCING HUMANE SCIENCEIMPROVING ANIMAL RESEARCH

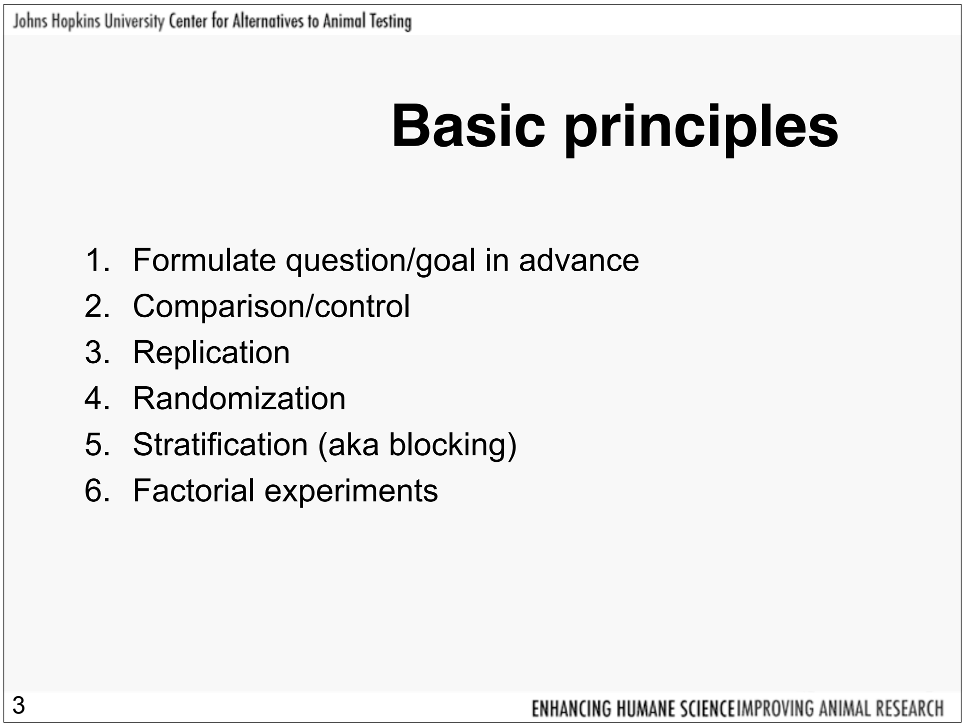


Johns Hopkins University Center for Alternatives to Animal Testing

Experimental design

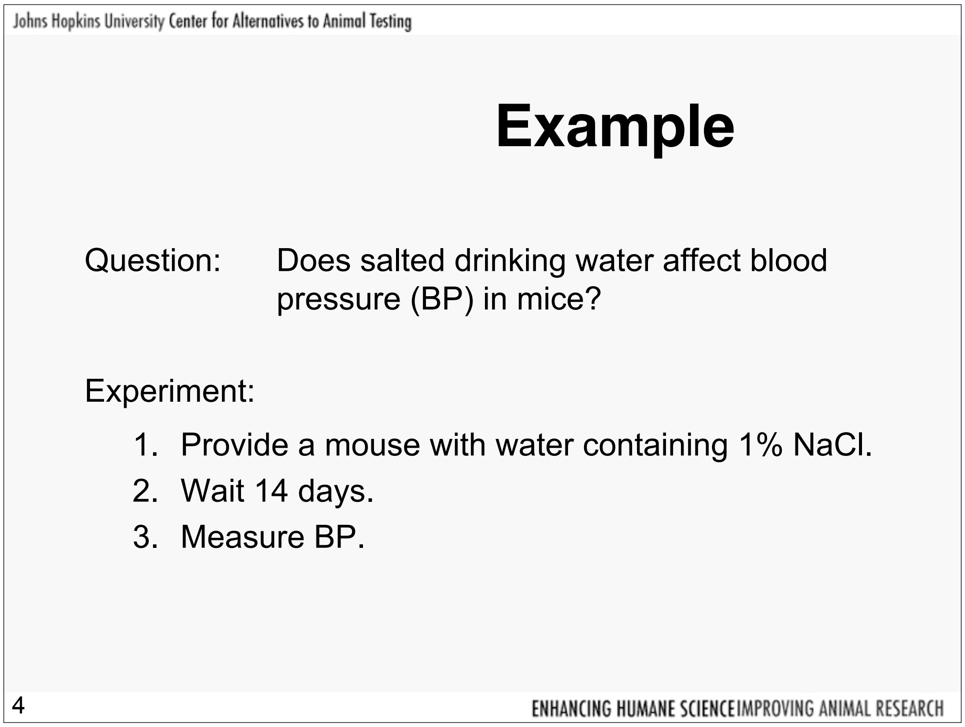
2

ENHANCING HUMANE SCIENCEIMPROVING ANIMAL RESEARCH



Basic principles

1. Formulate question/goal in advance
2. Comparison/control
3. Replication
4. Randomization
5. Stratification (aka blocking)
6. Factorial experiments

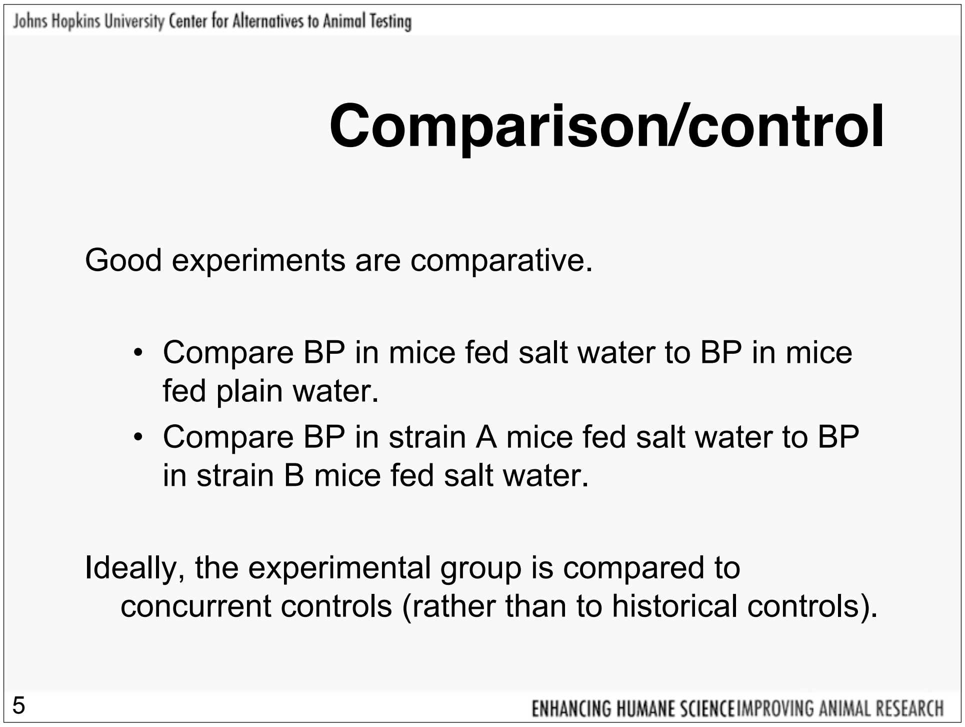


Example

Question: Does salted drinking water affect blood pressure (BP) in mice?

Experiment:

1. Provide a mouse with water containing 1% NaCl.
2. Wait 14 days.
3. Measure BP.

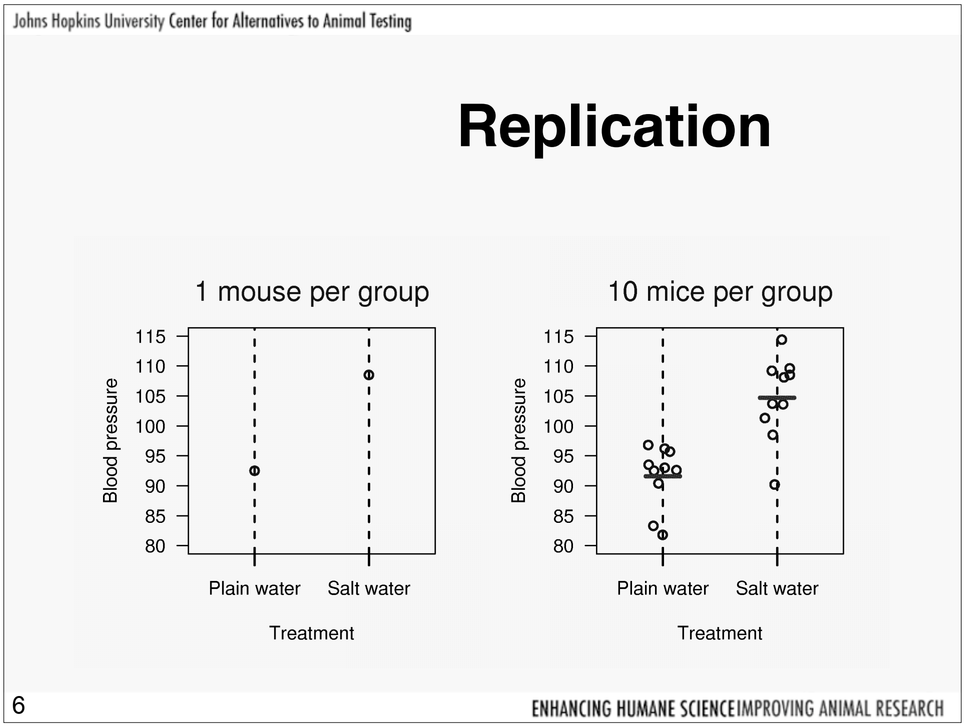


Comparison/control

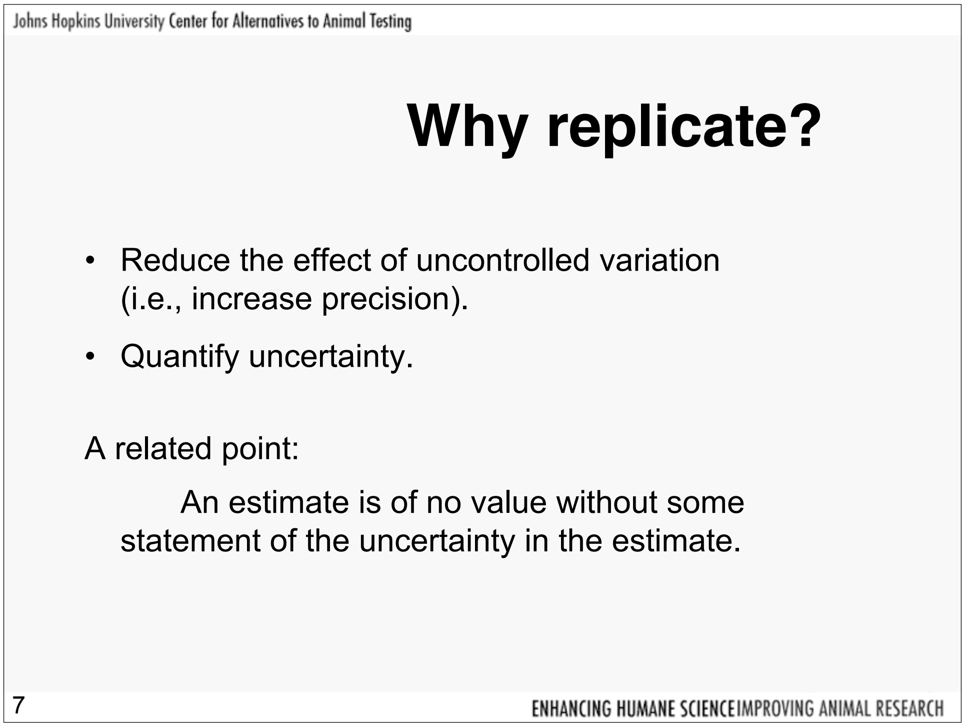
Good experiments are comparative.

- Compare BP in mice fed salt water to BP in mice fed plain water.
- Compare BP in strain A mice fed salt water to BP in strain B mice fed salt water.

Ideally, the experimental group is compared to concurrent controls (rather than to historical controls).



Replication

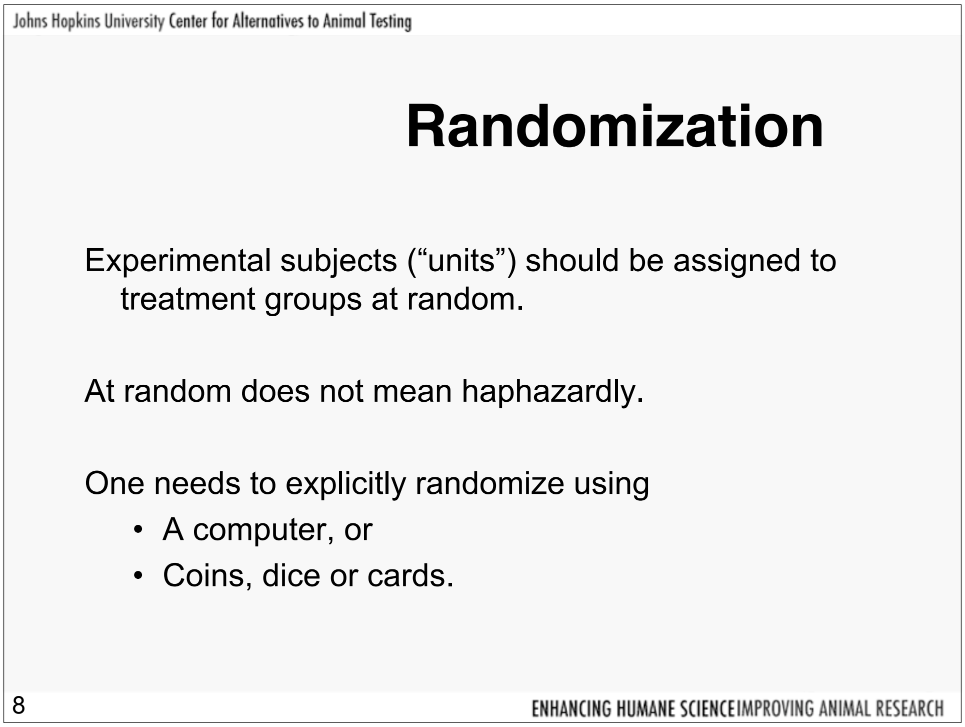


Why replicate?

- Reduce the effect of uncontrolled variation (i.e., increase precision).
- Quantify uncertainty.

A related point:

An estimate is of no value without some statement of the uncertainty in the estimate.



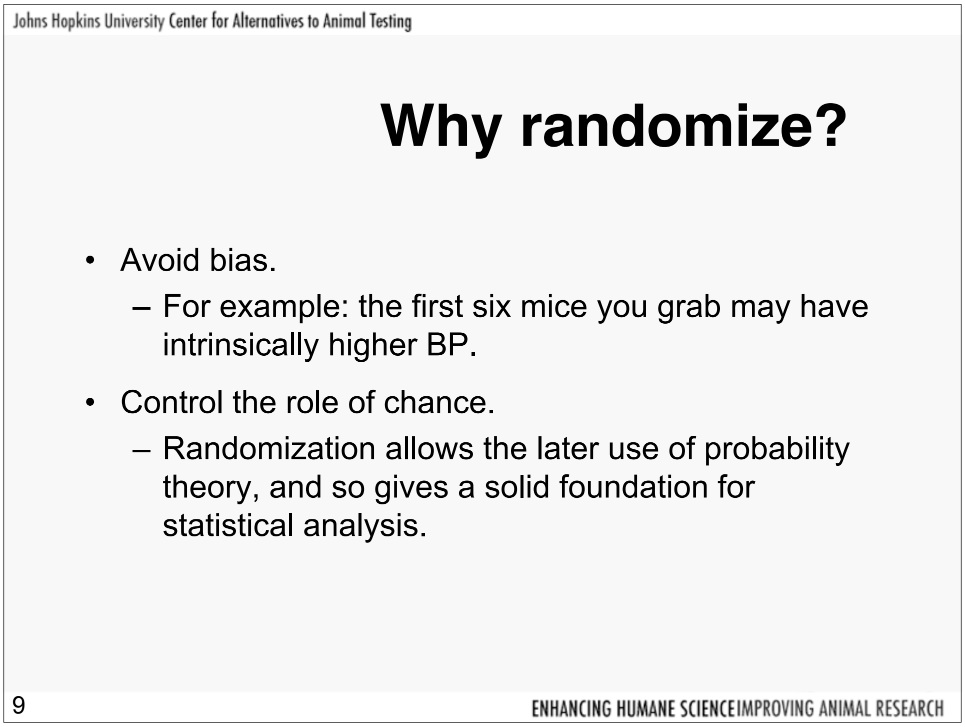
Randomization

Experimental subjects (“units”) should be assigned to treatment groups at random.

At random does not mean haphazardly.

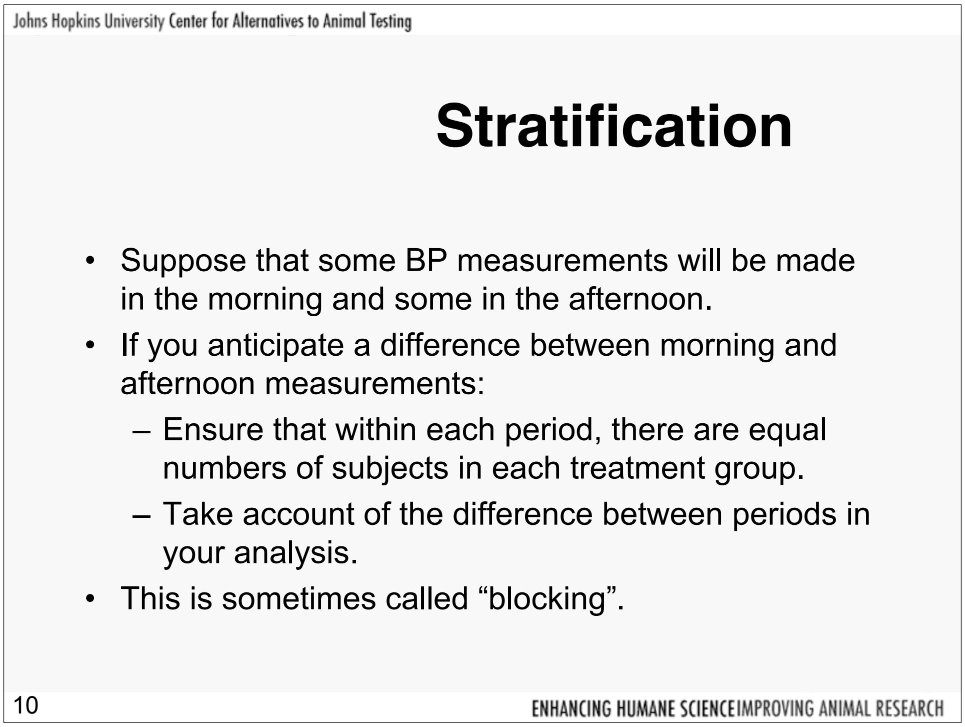
One needs to explicitly randomize using

- A computer, or
- Coins, dice or cards.



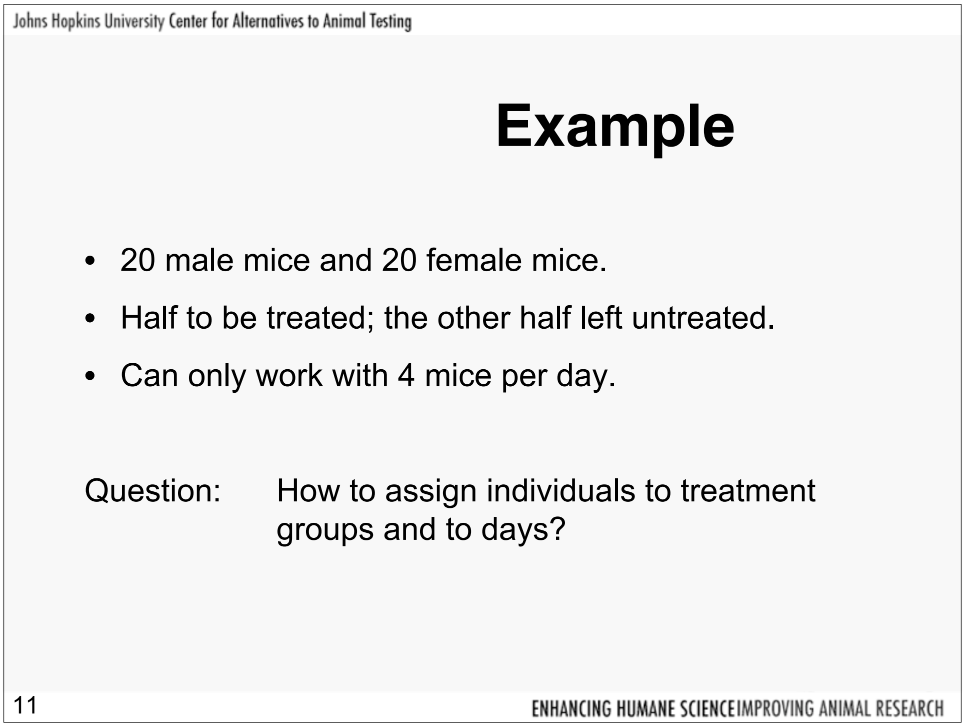
Why randomize?

- Avoid bias.
  - For example: the first six mice you grab may have intrinsically higher BP.
- Control the role of chance.
  - Randomization allows the later use of probability theory, and so gives a solid foundation for statistical analysis.



Stratification

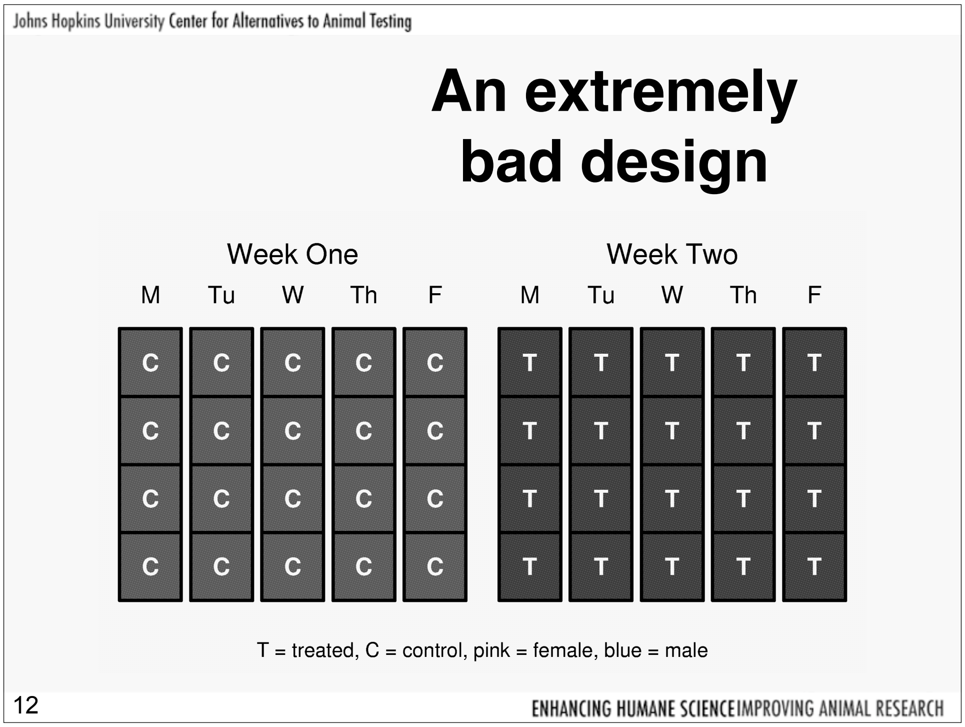
- Suppose that some BP measurements will be made in the morning and some in the afternoon.
- If you anticipate a difference between morning and afternoon measurements:
  - Ensure that within each period, there are equal numbers of subjects in each treatment group.
  - Take account of the difference between periods in your analysis.
- This is sometimes called “blocking”.



Example

- 20 male mice and 20 female mice.
- Half to be treated; the other half left untreated.
- Can only work with 4 mice per day.

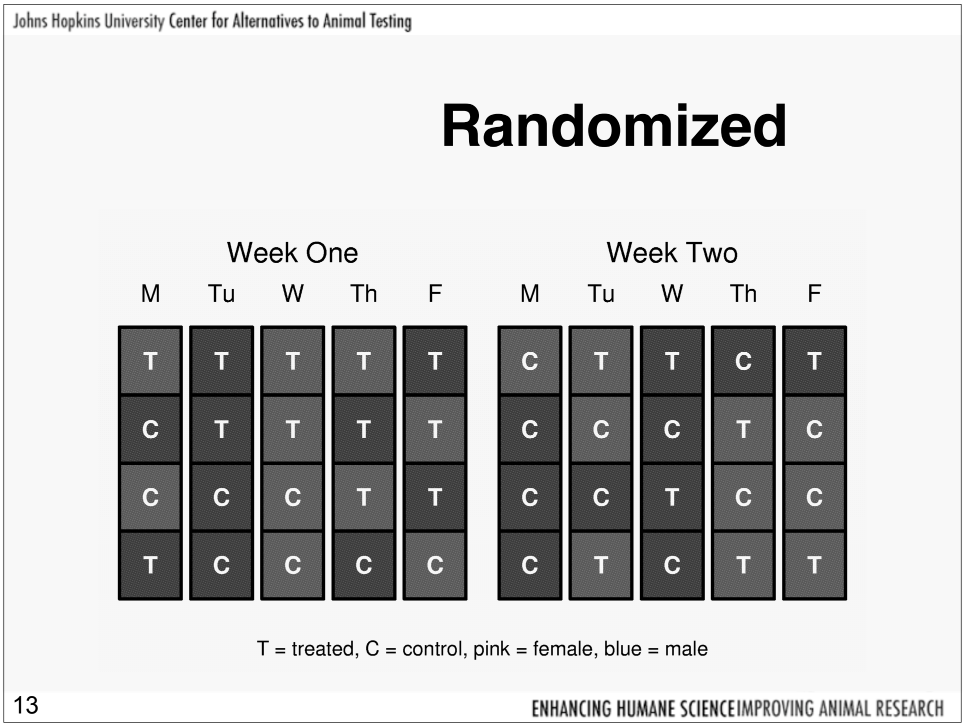
Question: How to assign individuals to treatment groups and to days?



An extremely bad design

Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T

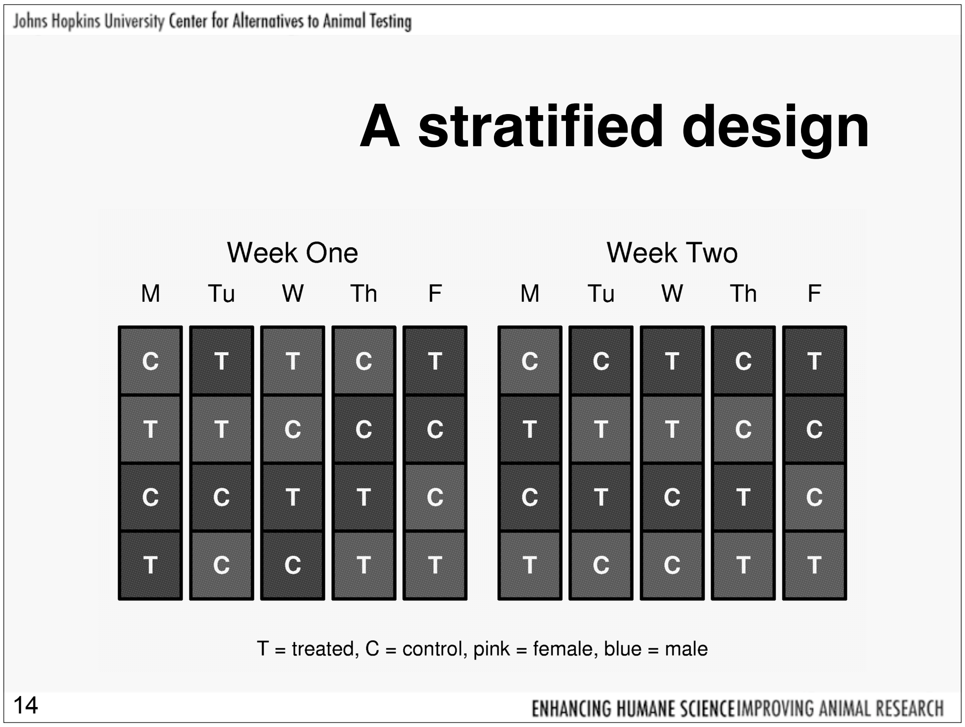
T = treated, C = control, pink = female, blue = male



Randomized

Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
T	T	T	T	T	C	T	T	C	T
C	T	T	T	T	C	C	C	T	C
C	C	C	T	T	C	C	T	C	C
T	C	C	C	C	C	T	C	T	T

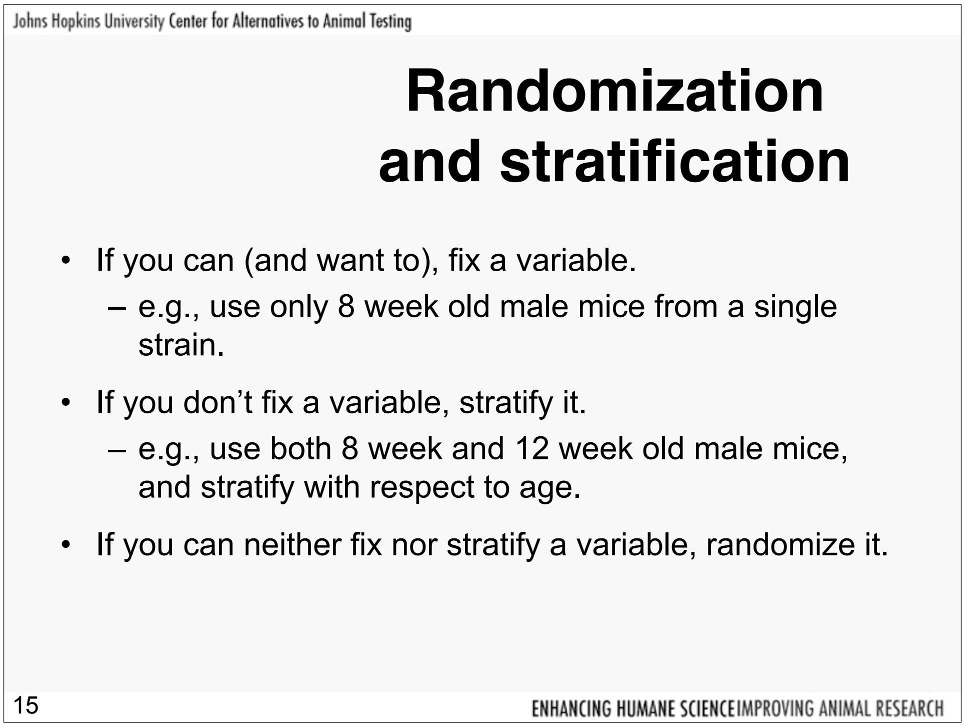
T = treated, C = control, pink = female, blue = male



A stratified design

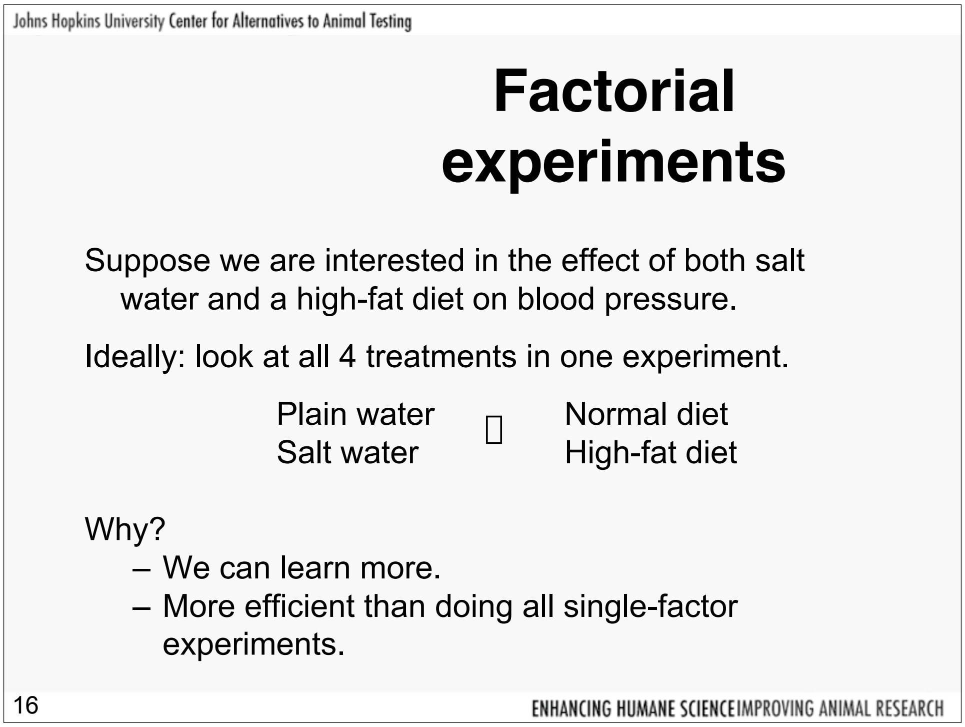
Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	T	T	C	T	C	C	T	C	T
T	T	C	C	C	T	T	T	C	C
C	C	T	T	C	C	T	C	T	C
T	C	C	T	T	T	C	C	T	T

T = treated, C = control, pink = female, blue = male



Randomization and stratification

- If you can (and want to), fix a variable.
  - e.g., use only 8 week old male mice from a single strain.
- If you don't fix a variable, stratify it.
  - e.g., use both 8 week and 12 week old male mice, and stratify with respect to age.
- If you can neither fix nor stratify a variable, randomize it.



Factorial experiments

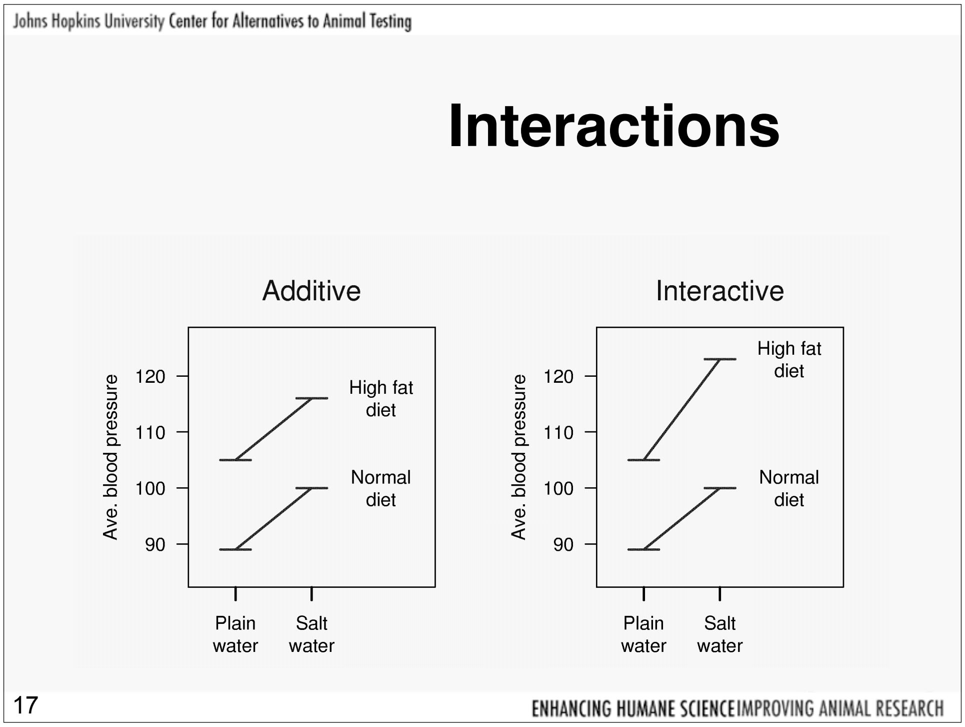
Suppose we are interested in the effect of both salt water and a high-fat diet on blood pressure.

Ideally: look at all 4 treatments in one experiment.

Plain water	×	Normal diet
Salt water		High-fat diet

Why?

- We can learn more.
- More efficient than doing all single-factor experiments.

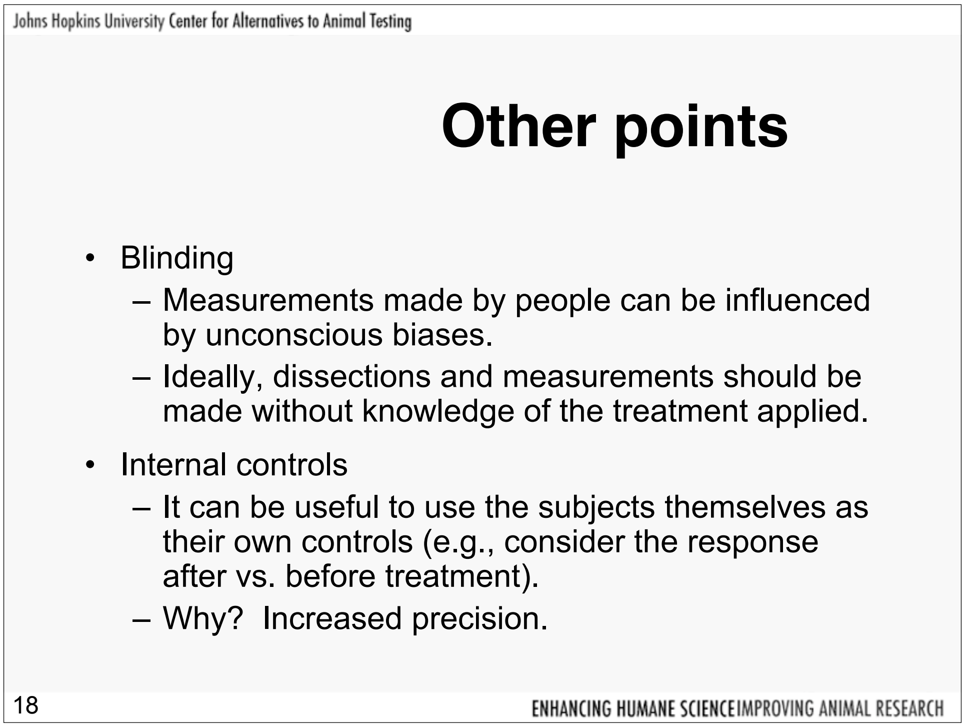


Johns Hopkins University Center for Alternatives to Animal Testing

Interactions

17

ENHANCING HUMANE SCIENCEIMPROVING ANIMAL RESEARCH



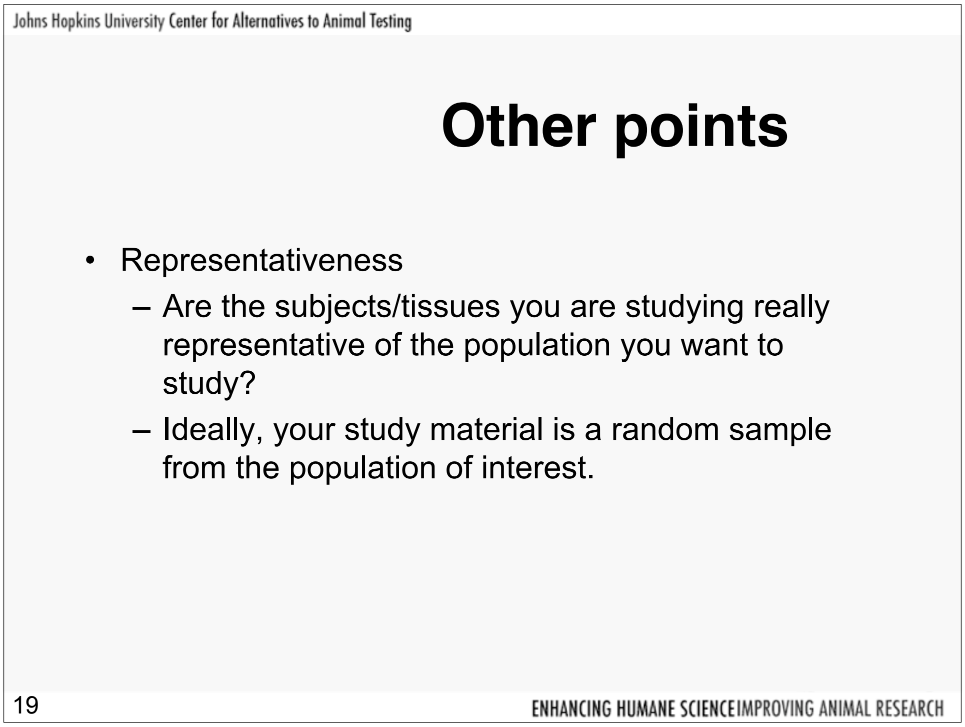
Johns Hopkins University Center for Alternatives to Animal Testing

Other points

- Blinding
  - Measurements made by people can be influenced by unconscious biases.
  - Ideally, dissections and measurements should be made without knowledge of the treatment applied.
- Internal controls
  - It can be useful to use the subjects themselves as their own controls (e.g., consider the response after vs. before treatment).
  - Why? Increased precision.

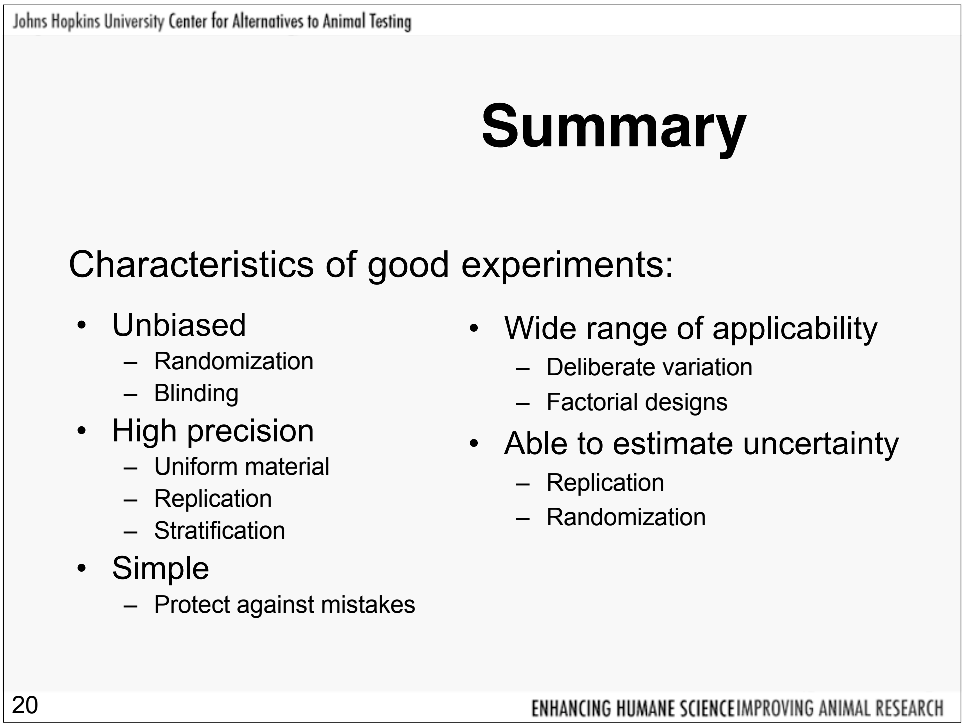
18

ENHANCING HUMANE SCIENCEIMPROVING ANIMAL RESEARCH



Other points

- Representativeness
  - Are the subjects/tissues you are studying really representative of the population you want to study?
  - Ideally, your study material is a random sample from the population of interest.

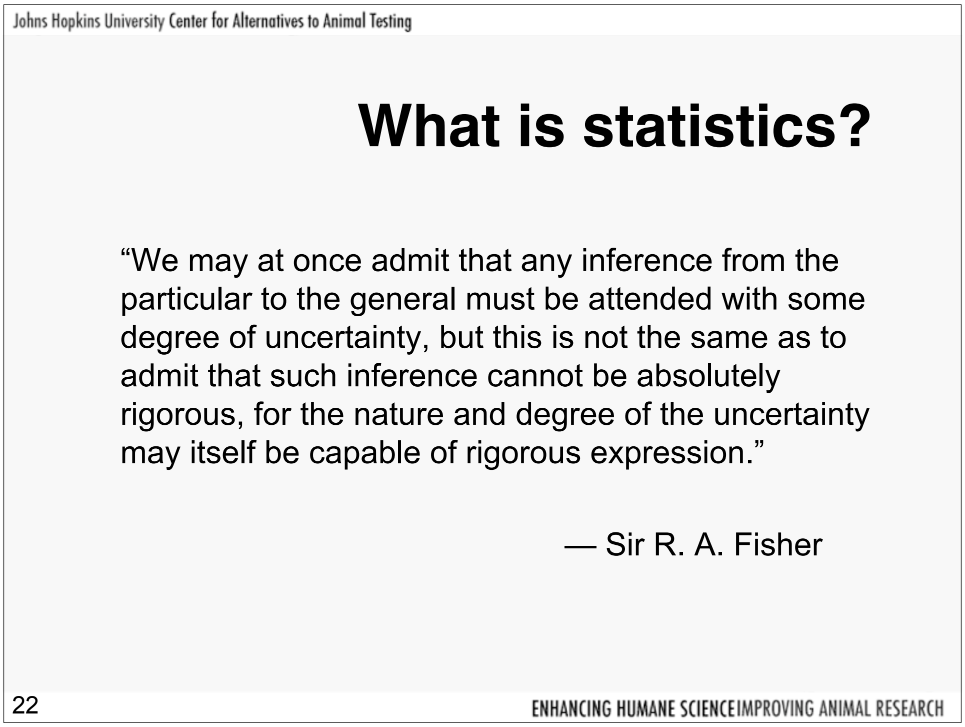


Summary

Characteristics of good experiments:

- Unbiased
  - Randomization
  - Blinding
- High precision
  - Uniform material
  - Replication
  - Stratification
- Simple
  - Protect against mistakes
- Wide range of applicability
  - Deliberate variation
  - Factorial designs
- Able to estimate uncertainty
  - Replication
  - Randomization

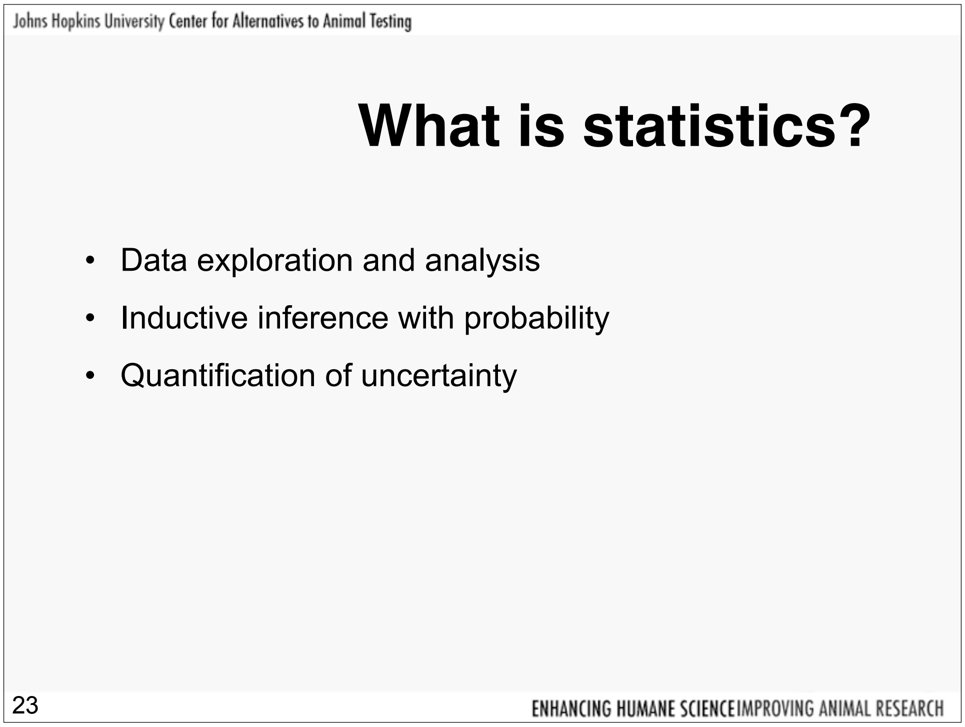
Basic statistics



What is statistics?

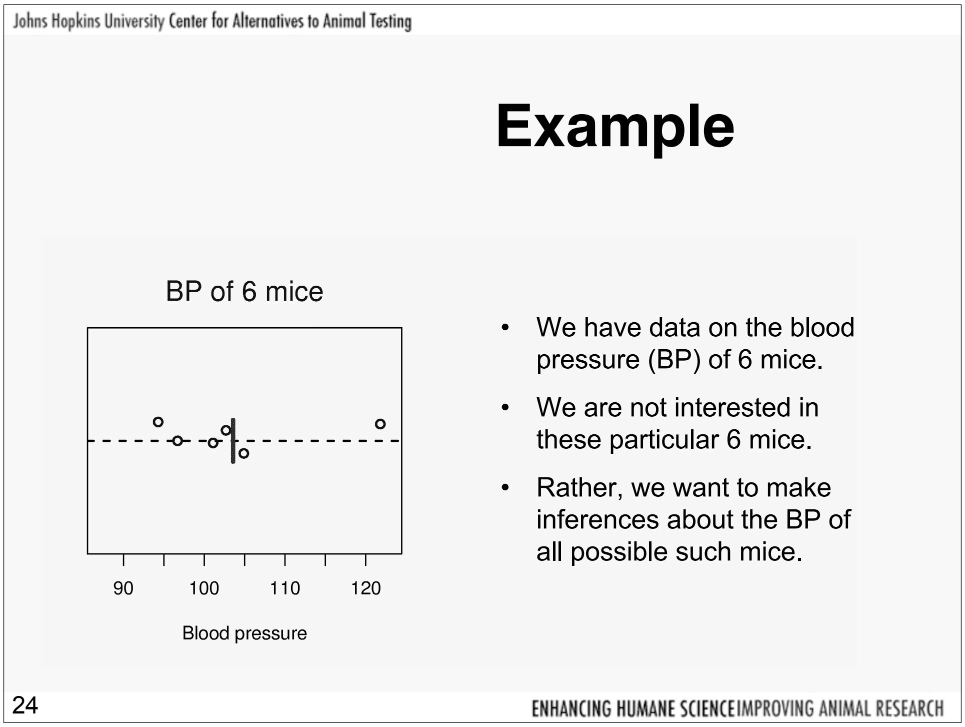
"We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression."

— Sir R. A. Fisher



What is statistics?

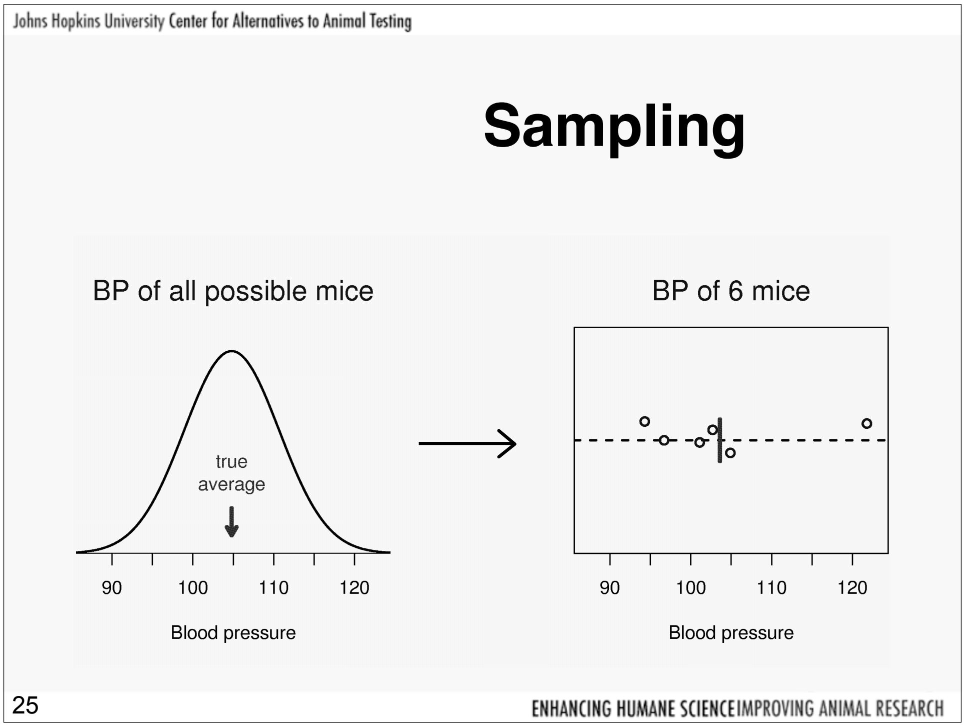
- Data exploration and analysis
- Inductive inference with probability
- Quantification of uncertainty



Example

BP of 6 mice

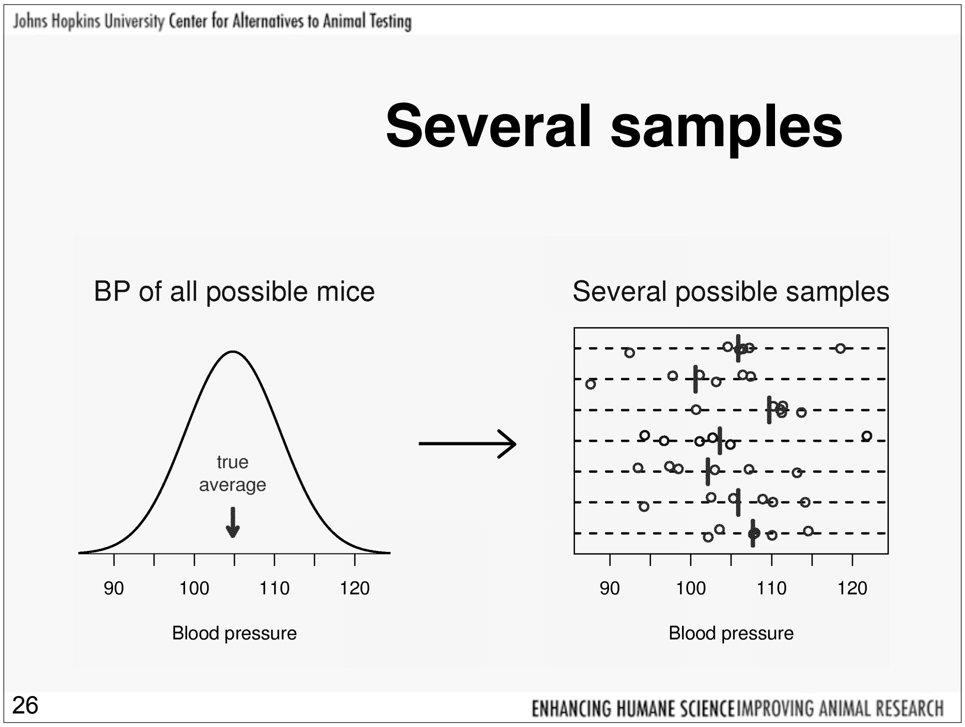
- We have data on the blood pressure (BP) of 6 mice.
- We are not interested in these particular 6 mice.
- Rather, we want to make inferences about the BP of all possible such mice.



Sampling

BP of all possible mice

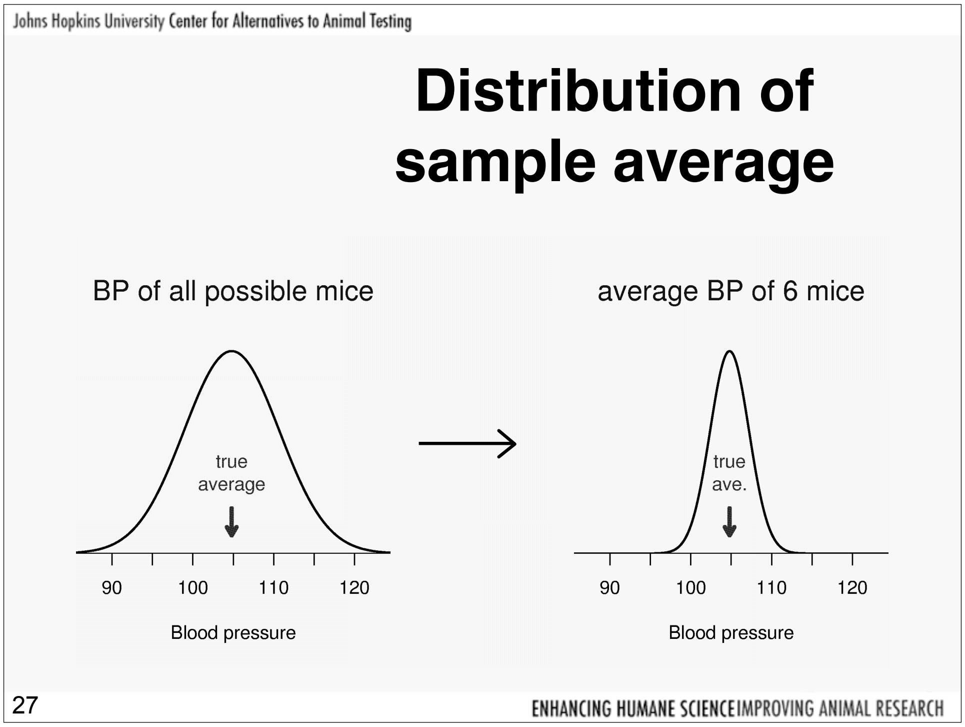
BP of 6 mice



Several samples

BP of all possible mice

Several possible samples

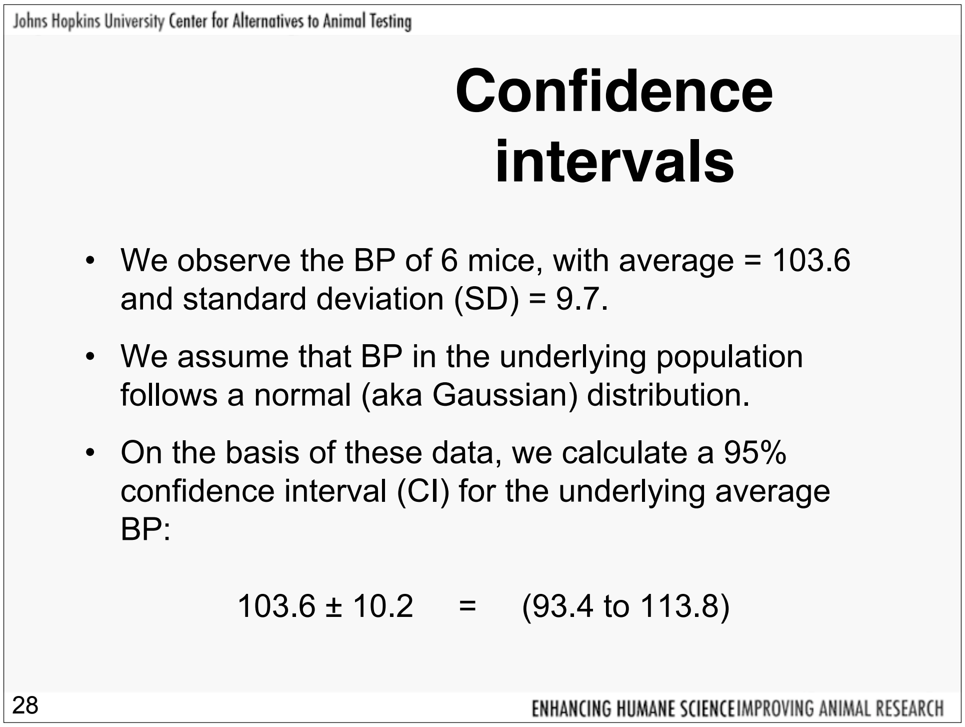


Johns Hopkins University Center for Alternatives to Animal Testing

Distribution of sample average

27

ENHANCING HUMANE SCIENCEIMPROVING ANIMAL RESEARCH



Johns Hopkins University Center for Alternatives to Animal Testing

Confidence intervals

- We observe the BP of 6 mice, with average = 103.6 and standard deviation (SD) = 9.7.
- We assume that BP in the underlying population follows a normal (aka Gaussian) distribution.
- On the basis of these data, we calculate a 95% confidence interval (CI) for the underlying average BP:

103.6 ± 10.2 = (93.4 to 113.8)

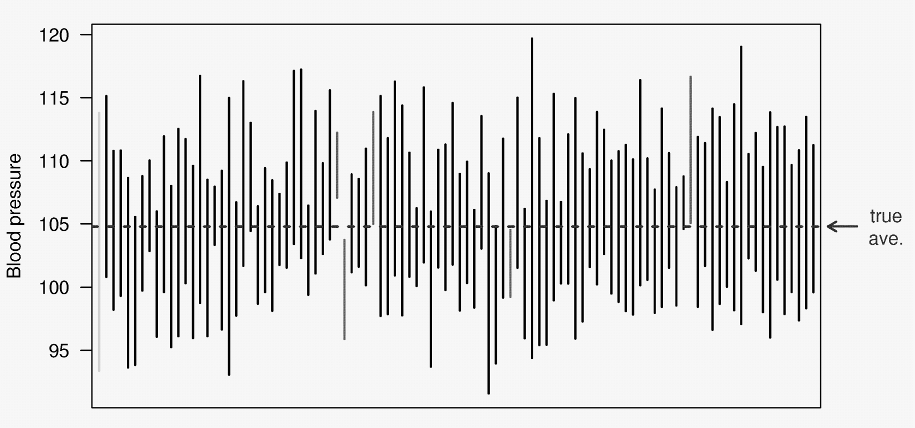
28

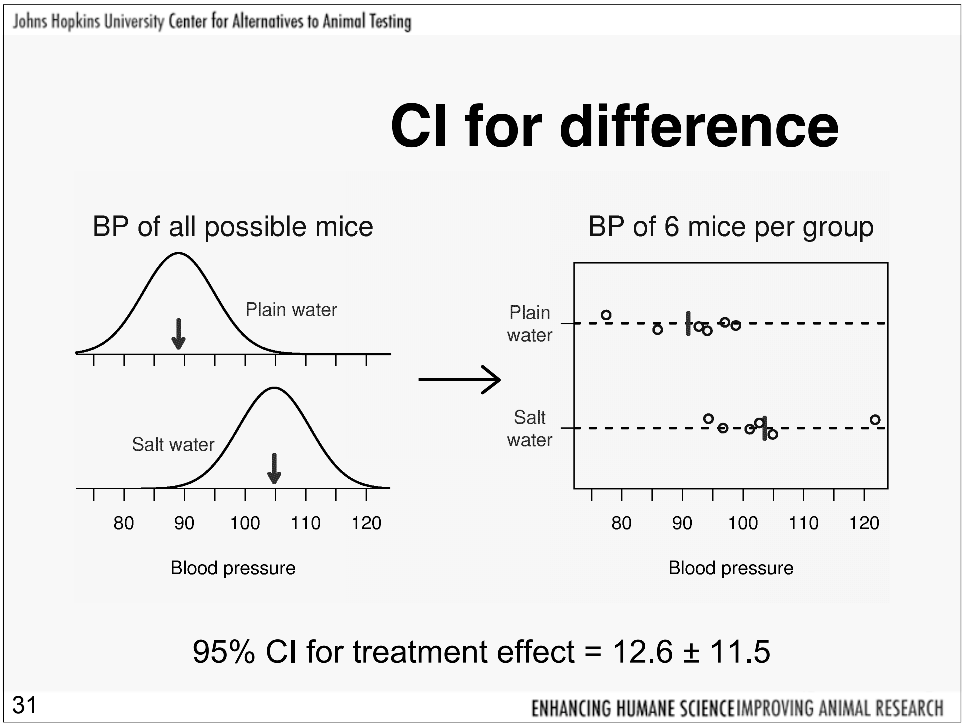
ENHANCING HUMANE SCIENCEIMPROVING ANIMAL RESEARCH

What is a CI?

- The plausible values for the underlying population average BP, given the data on the six mice.
- In advance, there is a 95% chance of obtaining an interval that contains the population average.

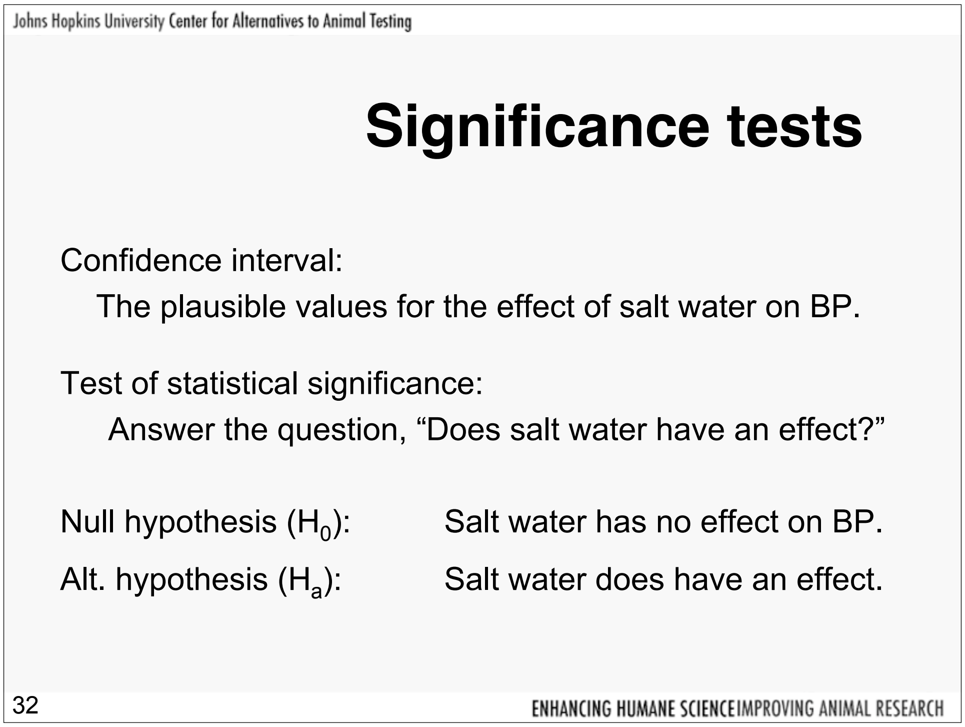
100 CIs





CI for difference

95% CI for treatment effect = 12.6 ± 11.5



Significance tests

Confidence interval:

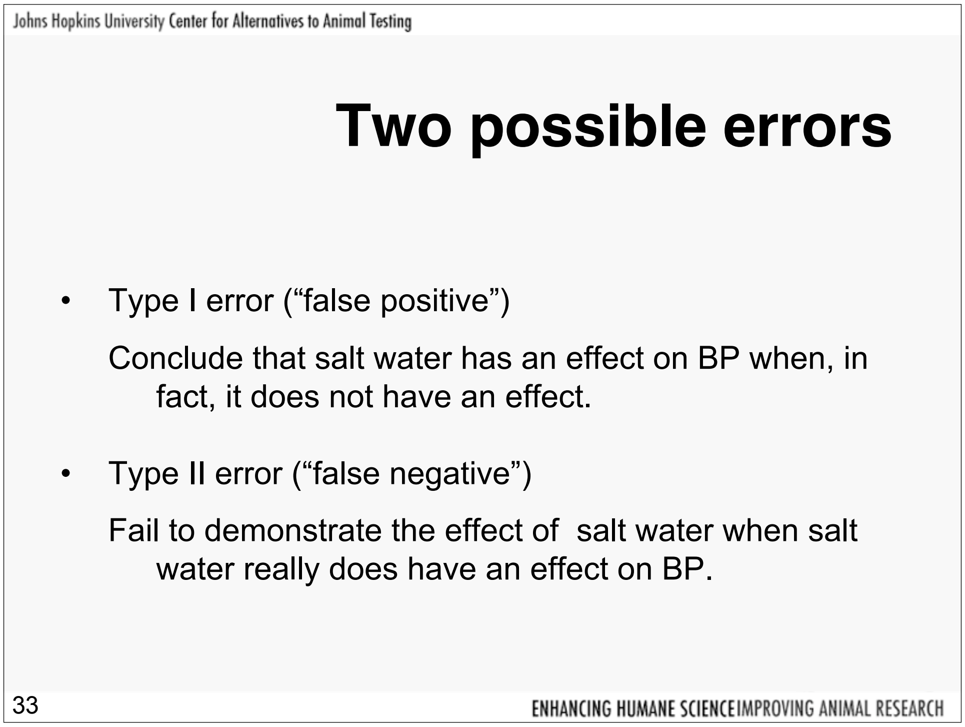
The plausible values for the effect of salt water on BP.

Test of statistical significance:

Answer the question, “Does salt water have an effect?”

Null hypothesis (H_0): Salt water has no effect on BP.

Alt. hypothesis (H_a): Salt water does have an effect.



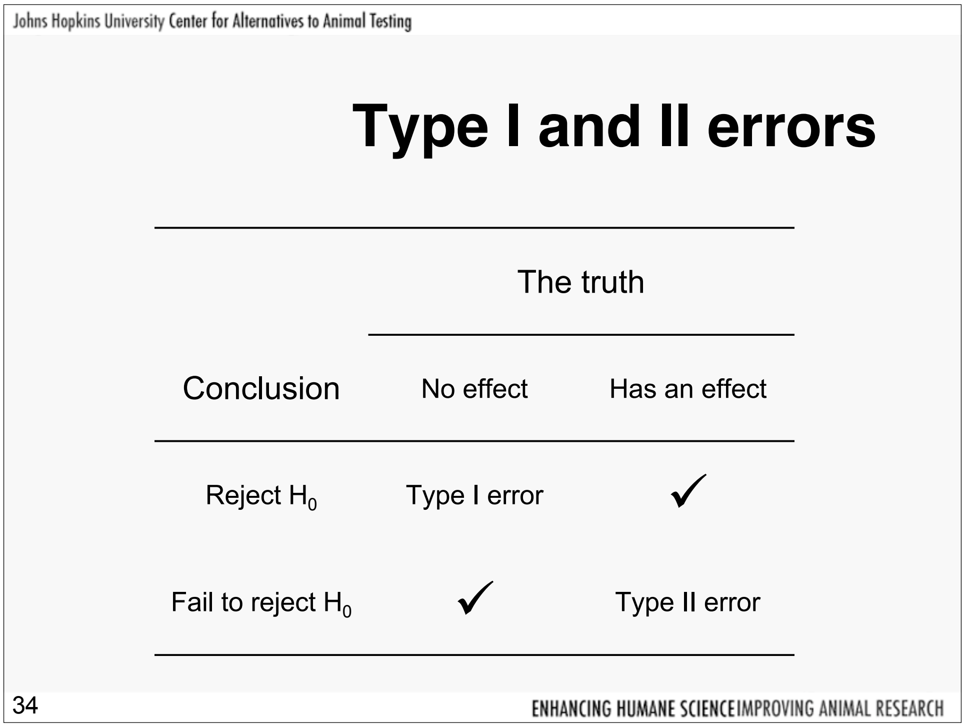
Two possible errors

- Type I error (“false positive”)

  Conclude that salt water has an effect on BP when, in fact, it does not have an effect.

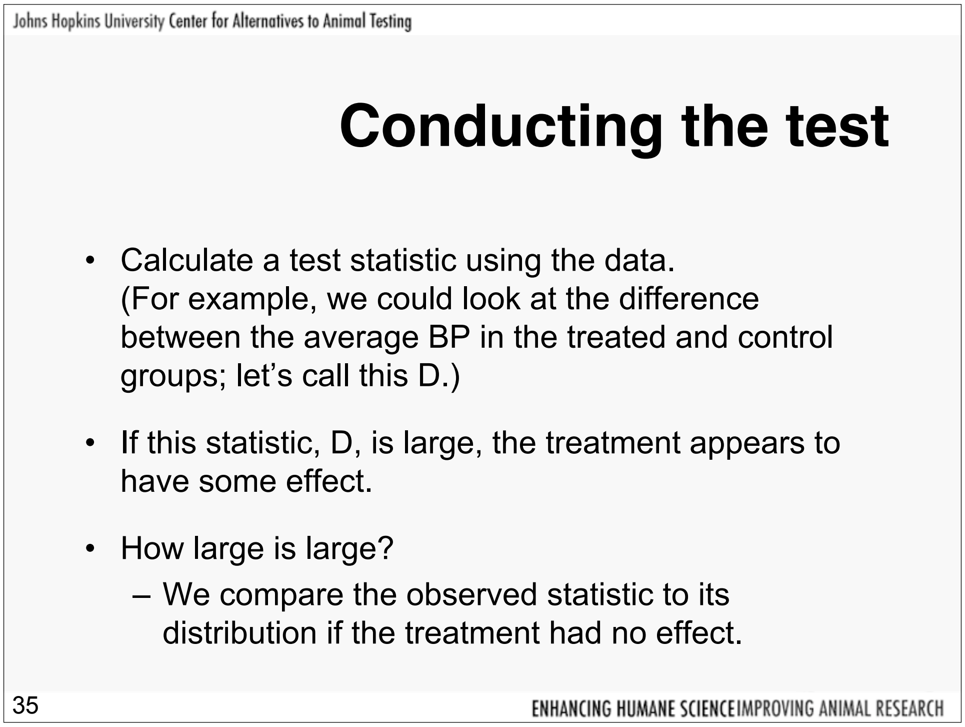
- Type II error (“false negative”)

  Fail to demonstrate the effect of salt water when salt water really does have an effect on BP.



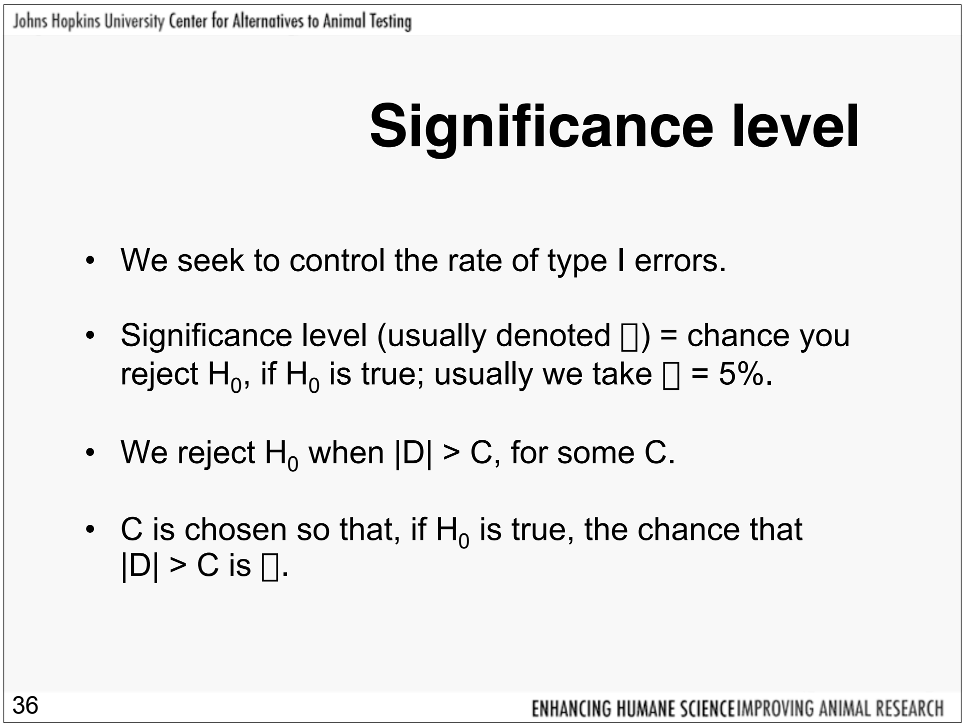
Type I and II errors

	The truth	
Conclusion	No effect	Has an effect
Reject H_0	Type I error	✓
Fail to reject H_0	✓	Type II error



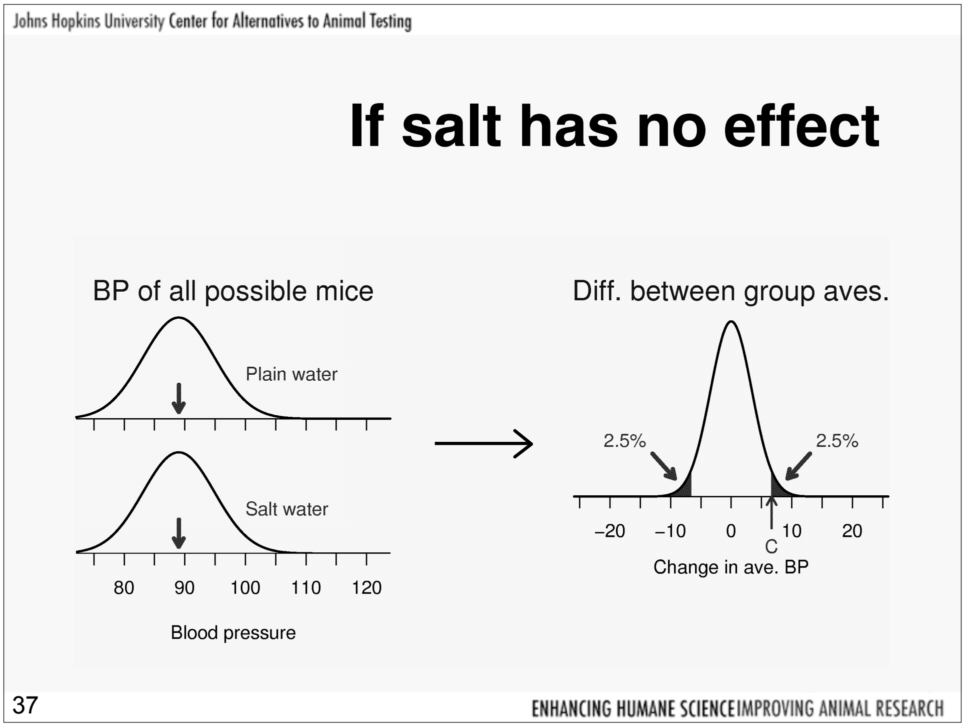
Conducting the test

- Calculate a test statistic using the data. (For example, we could look at the difference between the average BP in the treated and control groups; let's call this D.)
- If this statistic, D, is large, the treatment appears to have some effect.
- How large is large?
  - We compare the observed statistic to its distribution if the treatment had no effect.



Significance level

- We seek to control the rate of type I errors.
- Significance level (usually denoted □) = chance you reject H_0, if H_0 is true; usually we take □ = 5%.
- We reject H_0 when $|D| > C$, for some C.
- C is chosen so that, if H_0 is true, the chance that $|D| > C$ is □.

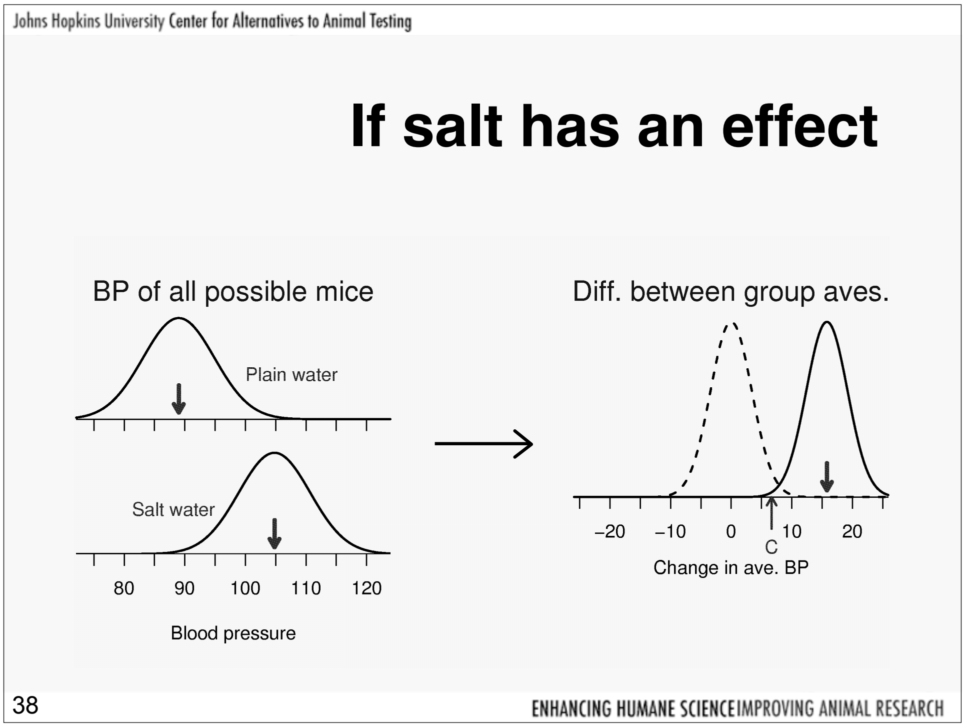


Johns Hopkins University Center for Alternatives to Animal Testing

If salt has no effect

37

ENHANCING HUMANE SCIENCE IMPROVING ANIMAL RESEARCH



Johns Hopkins University Center for Alternatives to Animal Testing

If salt has an effect

38

ENHANCING HUMANE SCIENCE IMPROVING ANIMAL RESEARCH

P-values

- A P-value is the probability of obtaining data as extreme as was observed, if the null hypothesis were true (i.e., if the treatment has no effect).
- If your P-value is smaller than your chosen significance level (□), you reject the null hypothesis.
- We seek to reject the null hypothesis (we seek to show that there is a treatment effect), and so small P-values are good.

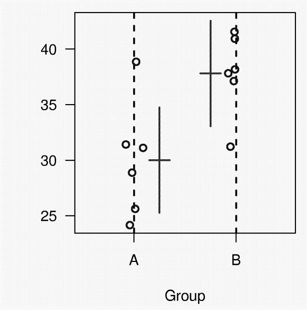
Summary

- Confidence interval
  - Plausible values for the true population average or treatment effect, given the observed data.
- Test of statistical significance
  - Use the observed data to answer a yes/no question, such as “Does the treatment have an effect?”
- P-value
  - Summarizes the result of the significance test.
  - Small P-value □ conclude that there is an effect.

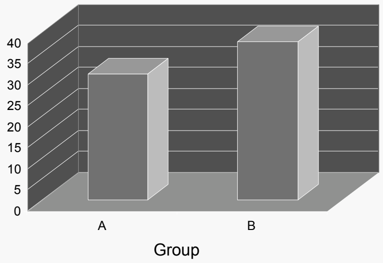
Never cite a P-value without a confidence interval.

Data presentation

Good plot



Bad plot



Data presentation

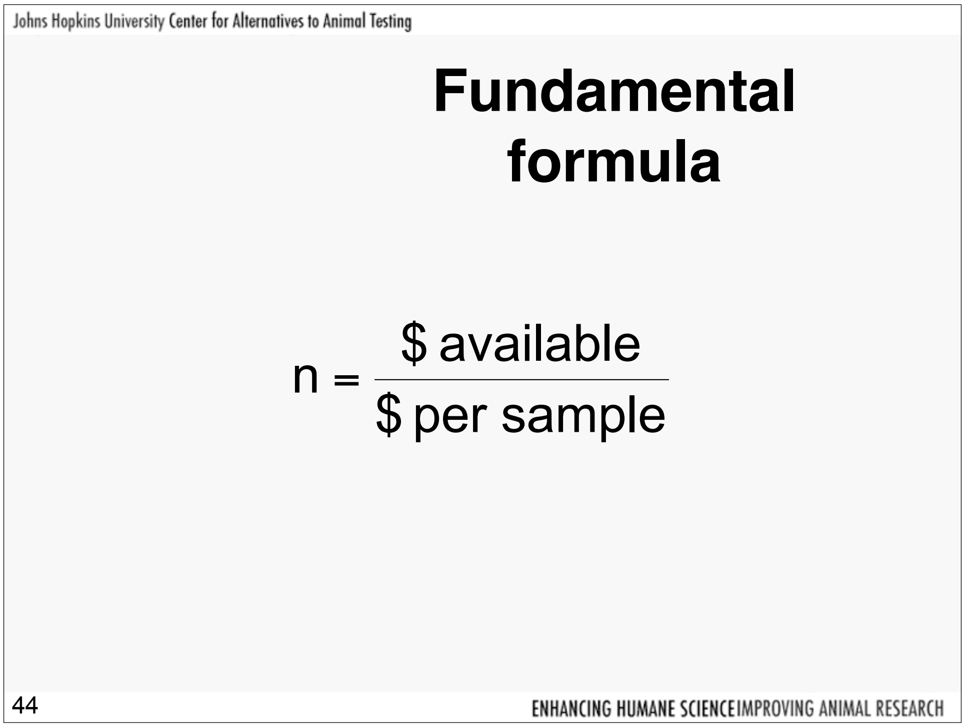
Good table

Treatment	Mean	(SEM)
A	11.3	(0.6)
B	13.5	(0.8)
C	14.7	(0.6)

Bad table

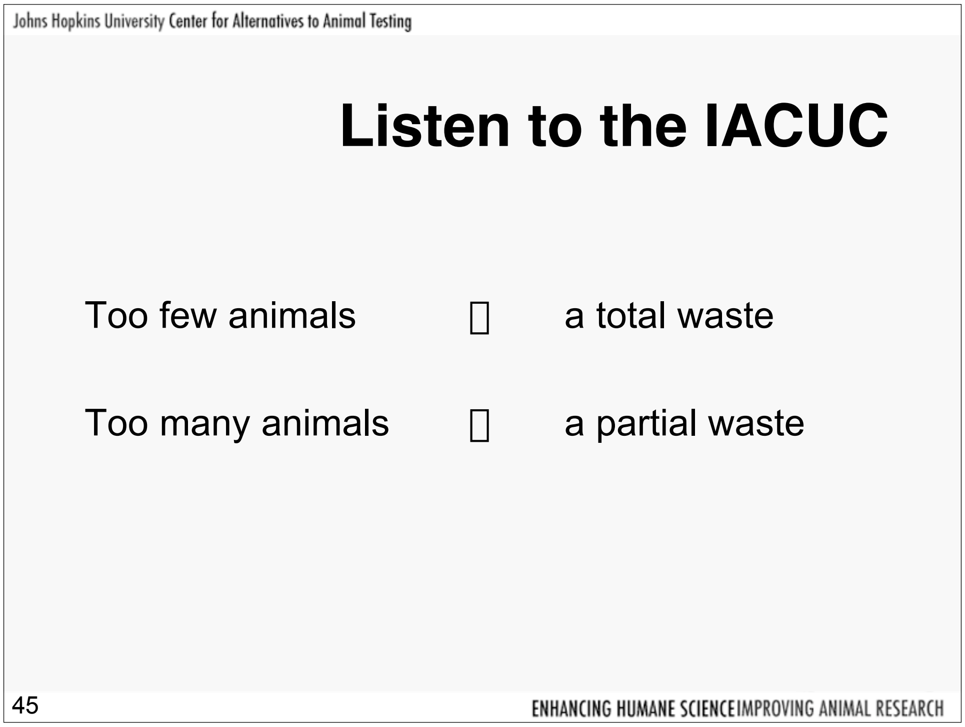
Treatment	Mean	(SEM)
A	11.2965	(0.63)
B	13.49	(0.7913)
C	14.727	(0.6108)

Sample size determination



Fundamental formula

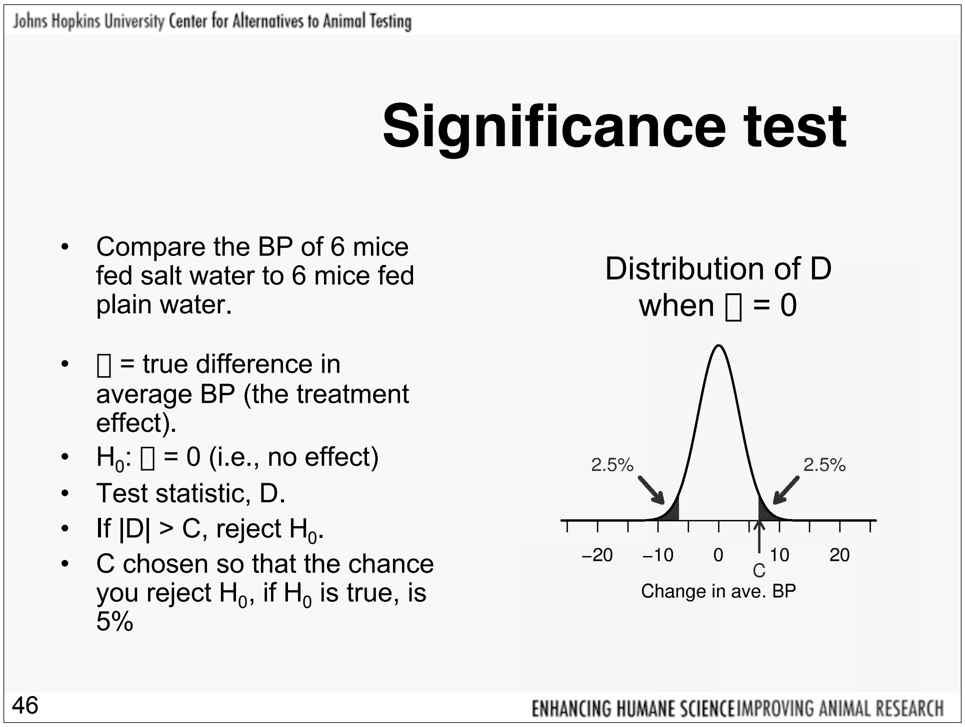
$$n = \frac{\$\text{ available}}{\$\text{ per sample}}$$



Listen to the IACUC

Too few animals a total waste

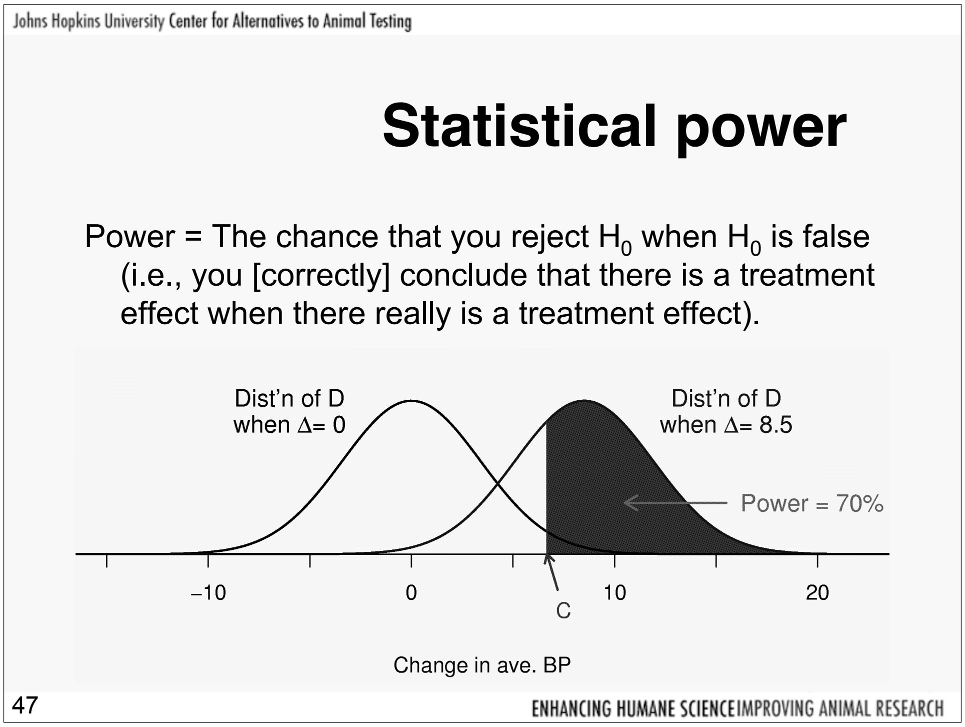
Too many animals a partial waste



Significance test

- Compare the BP of 6 mice fed salt water to 6 mice fed plain water.
- = true difference in average BP (the treatment effect).
- H_0: = 0 (i.e., no effect)
- Test statistic, D.
- If |D| > C, reject H_0.
- C chosen so that the chance you reject H_0, if H_0 is true, is 5%

Distribution of D when = 0



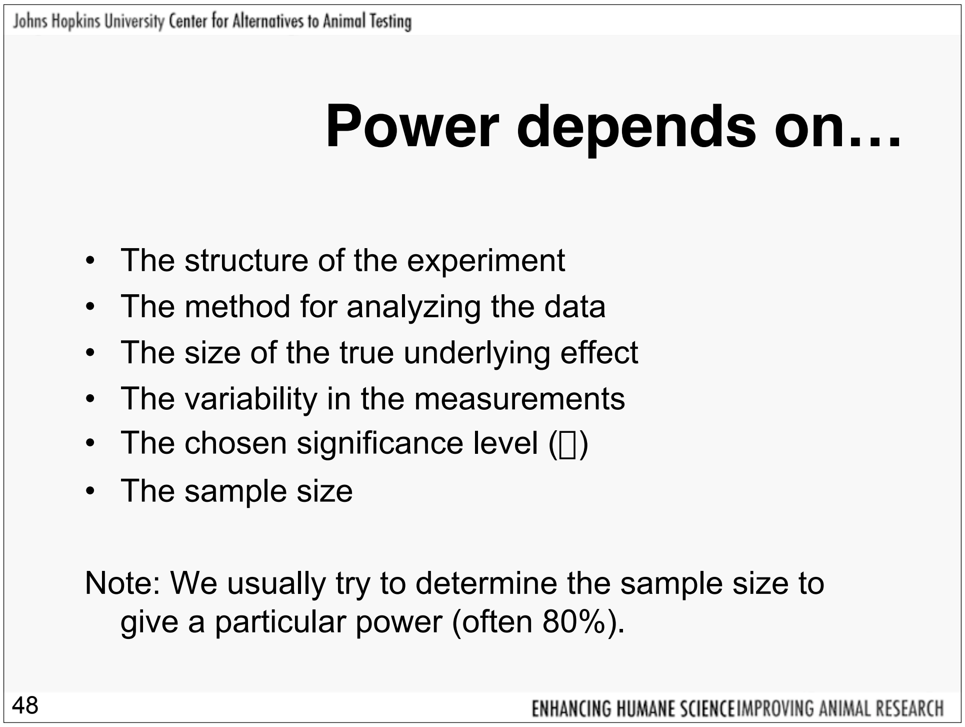
Johns Hopkins University Center for Alternatives to Animal Testing

Statistical power

Power = The chance that you reject H_0 when H_0 is false (i.e., you [correctly] conclude that there is a treatment effect when there really is a treatment effect).

47

ENHANCING HUMANE SCIENCEIMPROVING ANIMAL RESEARCH



Johns Hopkins University Center for Alternatives to Animal Testing

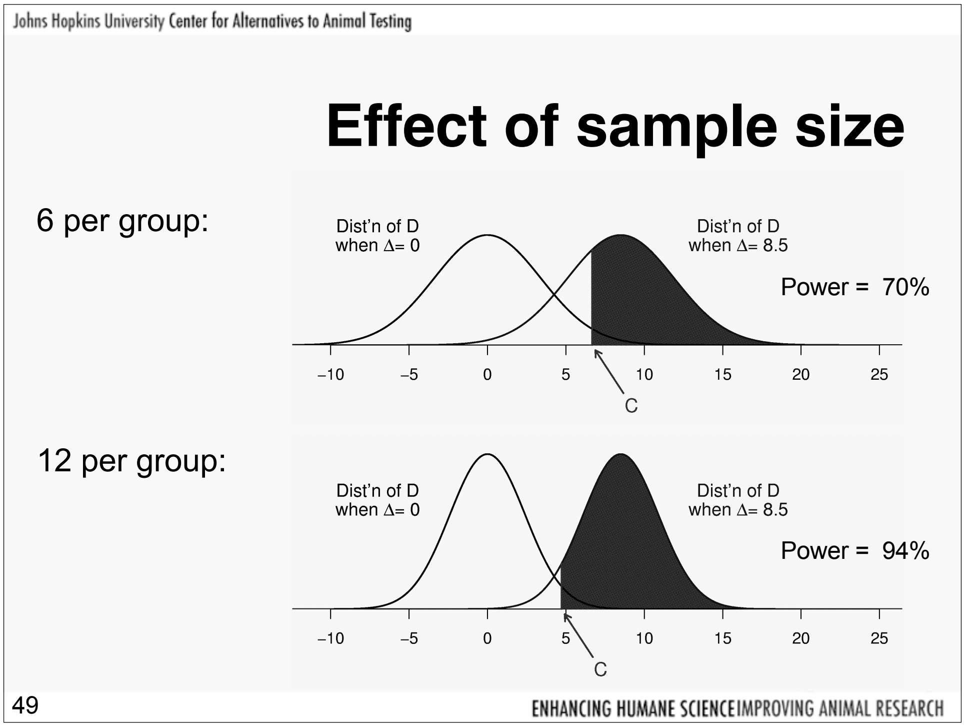
Power depends on…

- The structure of the experiment
- The method for analyzing the data
- The size of the true underlying effect
- The variability in the measurements
- The chosen significance level (□)
- The sample size

Note: We usually try to determine the sample size to give a particular power (often 80%).

48

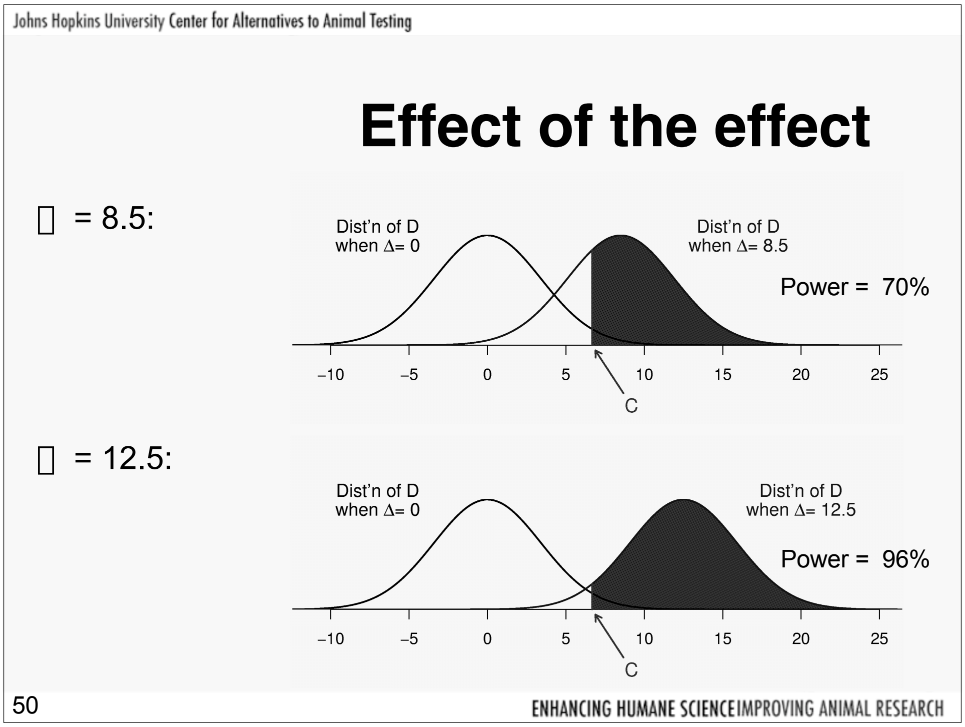
ENHANCING HUMANE SCIENCEIMPROVING ANIMAL RESEARCH



Effect of sample size

6 per group:

12 per group:



Effect of the effect

□ = 8.5:

□ = 12.5:

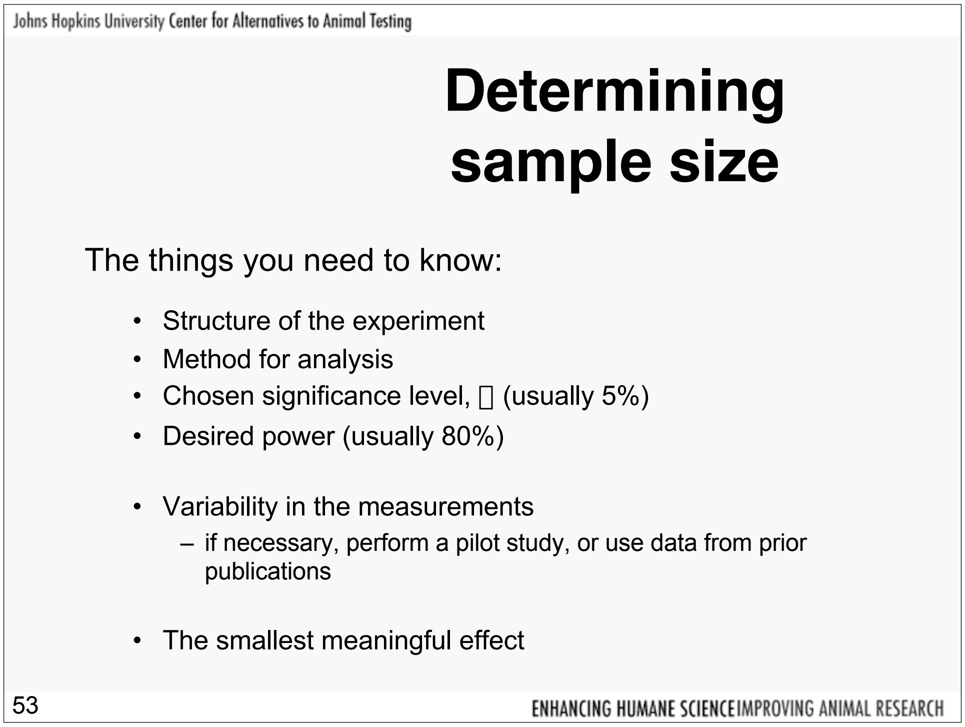
A formula

$$n = \frac{\square/\square^{2}}{\square-\square} \left[\quad \square + \square/\square \right]^{2} \square 2$$

Censored

Various effects

- Desired power ↑ □ sample size ↑
- Stringency of statistical test ↑ □ sample size ↑
- Measurement variability □ □ sample size □
- Treatment effect ↑ □ sample size □



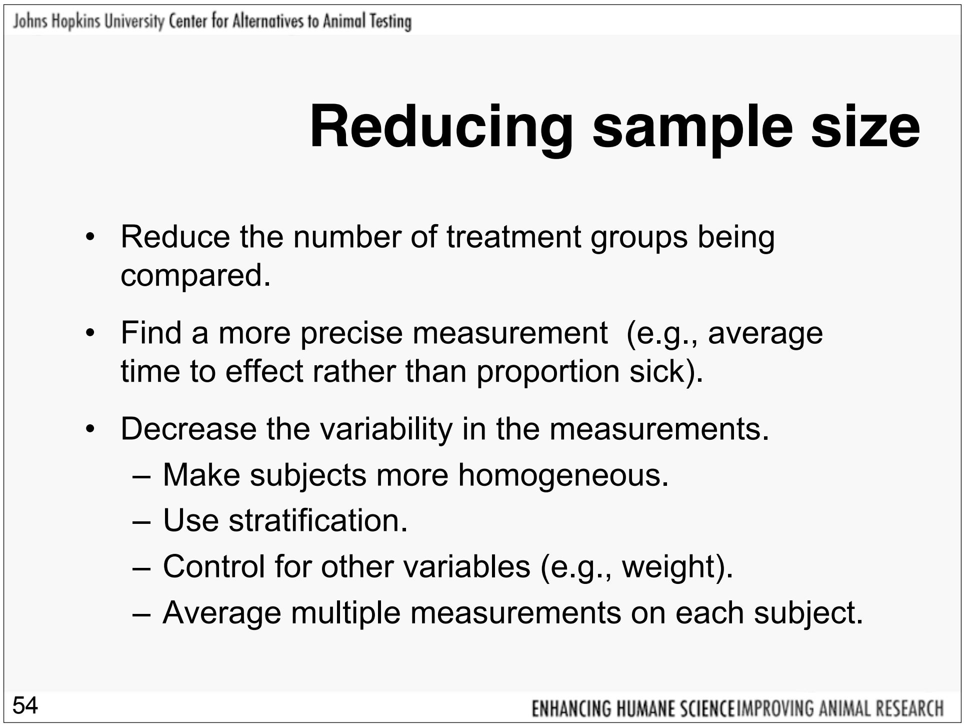
Determining sample size

The things you need to know:

- Structure of the experiment
- Method for analysis
- Chosen significance level, □ (usually 5%)
- Desired power (usually 80%)

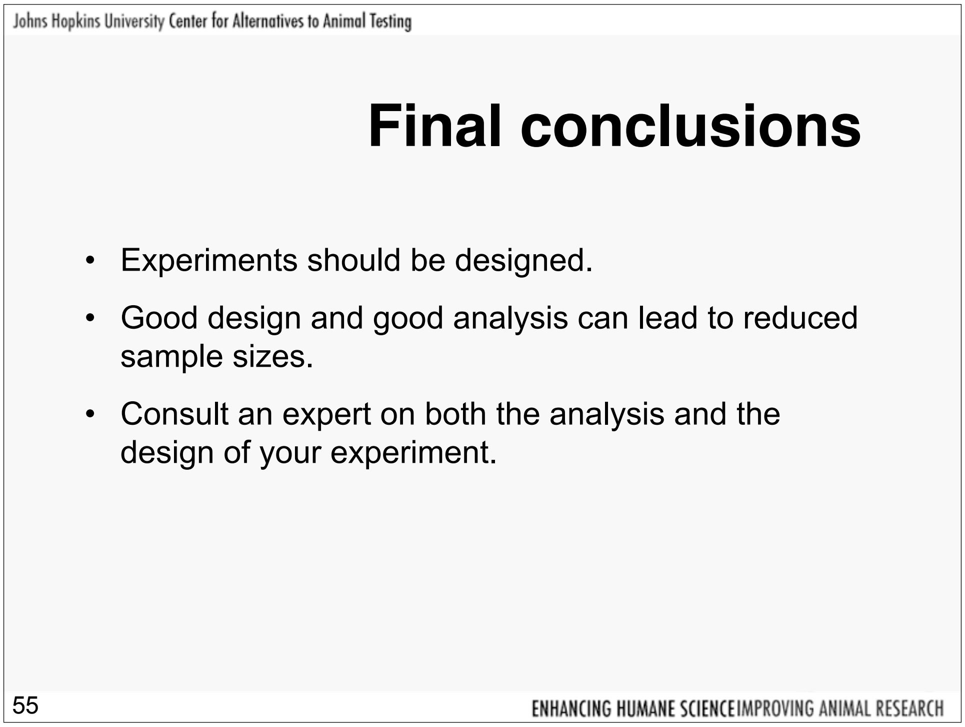
- Variability in the measurements
  - if necessary, perform a pilot study, or use data from prior publications

- The smallest meaningful effect



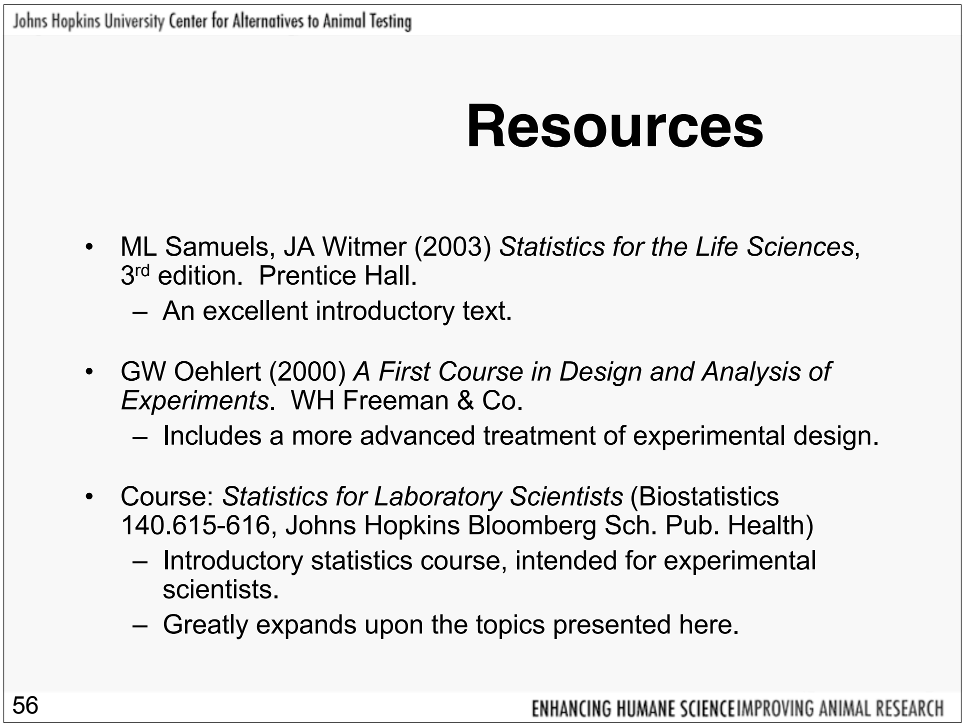
Reducing sample size

- Reduce the number of treatment groups being compared.
- Find a more precise measurement (e.g., average time to effect rather than proportion sick).
- Decrease the variability in the measurements.
  - Make subjects more homogeneous.
  - Use stratification.
  - Control for other variables (e.g., weight).
  - Average multiple measurements on each subject.



Final conclusions

- Experiments should be designed.
- Good design and good analysis can lead to reduced sample sizes.
- Consult an expert on both the analysis and the design of your experiment.



Resources

- ML Samuels, JA Witmer (2003) *Statistics for the Life Sciences*, 3rd edition. Prentice Hall.
  - An excellent introductory text.
- GW Oehlert (2000) *A First Course in Design and Analysis of Experiments*. WH Freeman & Co.
  - Includes a more advanced treatment of experimental design.
- Course: *Statistics for Laboratory Scientists* (Biostatistics 140.615-616, Johns Hopkins Bloomberg Sch. Pub. Health)
  - Introductory statistics course, intended for experimental scientists.
  - Greatly expands upon the topics presented here.