Hypothesis Testing and Estimation

Branko Miladinovic, Ph.D.
Luis Maldonado, MD

March 31, 2011

USF HEALTH CENTER FOR EVIDENCE BASED MEDICINE

UNIVERSITY OF SOUTH FLORIDA

Which is not true about the variance?

A. It is the square of the standard deviation.

B. It is a measure of the spread of data.

C. The units of the variance are different from the units of the original data set.

D. It is not affected by outliers.

Which is not true about the variance?

A. It is the square of the standard deviation.

B. It is a measure of the spread of data.

C. The units of the variance are different from the units of the original data set.

D. It is not affected by outliers.

Review Example: Ramipril is an angiotensin-converting enzyme (ACE) inhibitor which has been tested for use in patients at high risk of cardiovascular events. In one study published in the *New England Journal of Medicine, a total of 9,297 patients were recruited into a randomized, double-blind,* controlled trial.

	Cardio Event (including death) Present	Cardio Event Absent	Total
Ramparil Group	651 (14.0%)	3,994 (84%)	4,645
Placebo Group	826 (17.8%)	3,826 (82.2%)	4,652
Total	1,477	7,820	9,297

- These data indicate that fewer people treated with ramipril suffered a cardiovascular event: (14.0%) compared with those in the placebo group (17.8%).
- This gives a relative risk of 0.14/0.178 = 0.78, or a reduction in (relative) risk of 22%.

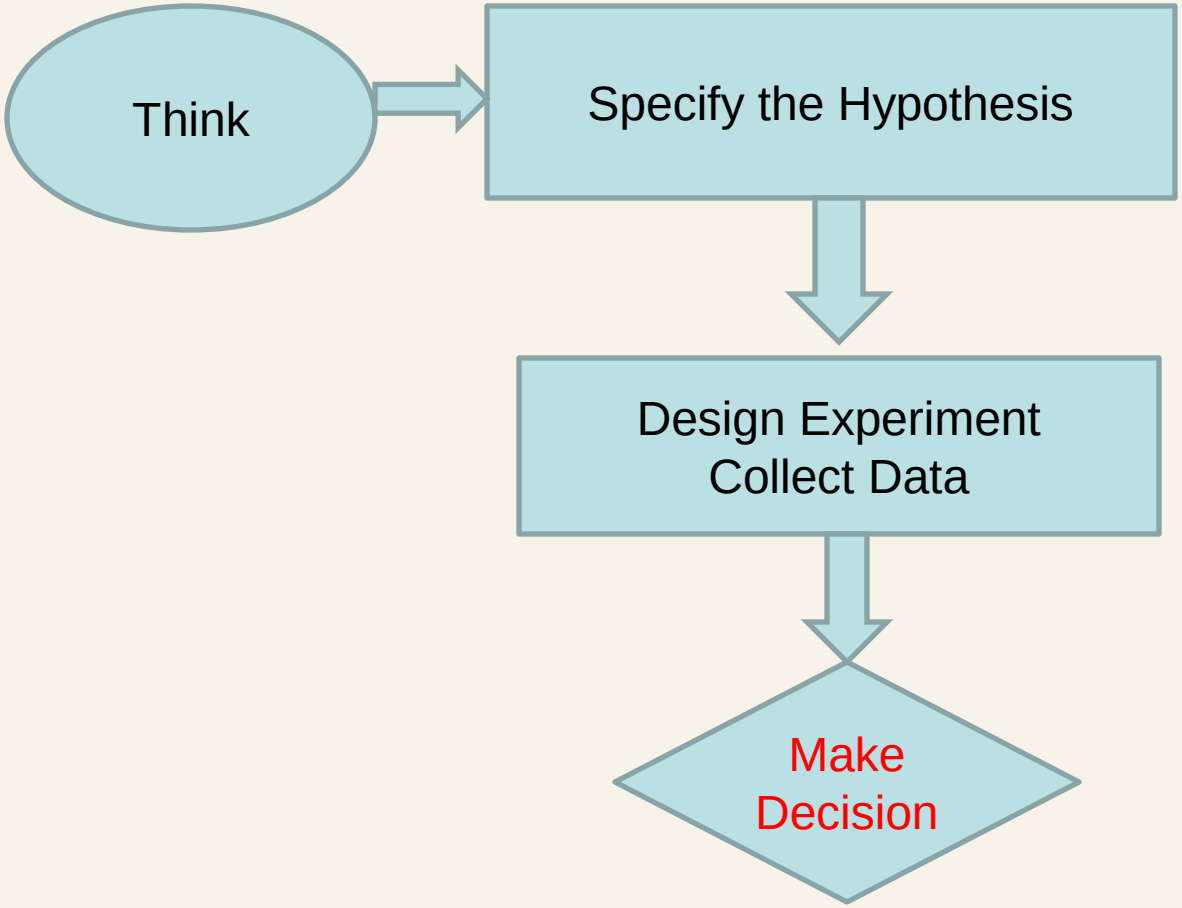
Today's lecture will help us understand the following statements:

1. The 95% confidence interval for this estimate of the relative risk runs from 0.70 to 0.86. Two observations can then be made from this confidence interval.
2. The observed difference is statistically significant at the 5% level, because the interval does not embrace a relative risk of one (point of no difference).
3. The observed data are consistent with as much as a 30% reduction in relative risk or as little as 14% reduction in risk

Outline

- Hypothesis Testing
- Estimation and Power
- Confidence Intervals

Hypothesis Testing



Prevalent Paradigm

Hypothesis testing is decision making based on the available evidence, i.e. the experiment is a procedure for choosing between two competing hypothesis, mindful that we do not commit errors and choose one when the other is wrong.

Hypothesis Testing

- Fundamental assumption of any quantitative research: There exists a true underlying treatment effect that any one experiment can only estimate.

- In terms of hypothesis:

  There is no difference between outcomes (Null).

  There is significant difference between outcomes (Alternative)

- Statistics probabilistically quantifies the differences so that decisions can be made.

Example: Lung cancer and chemo therapy.

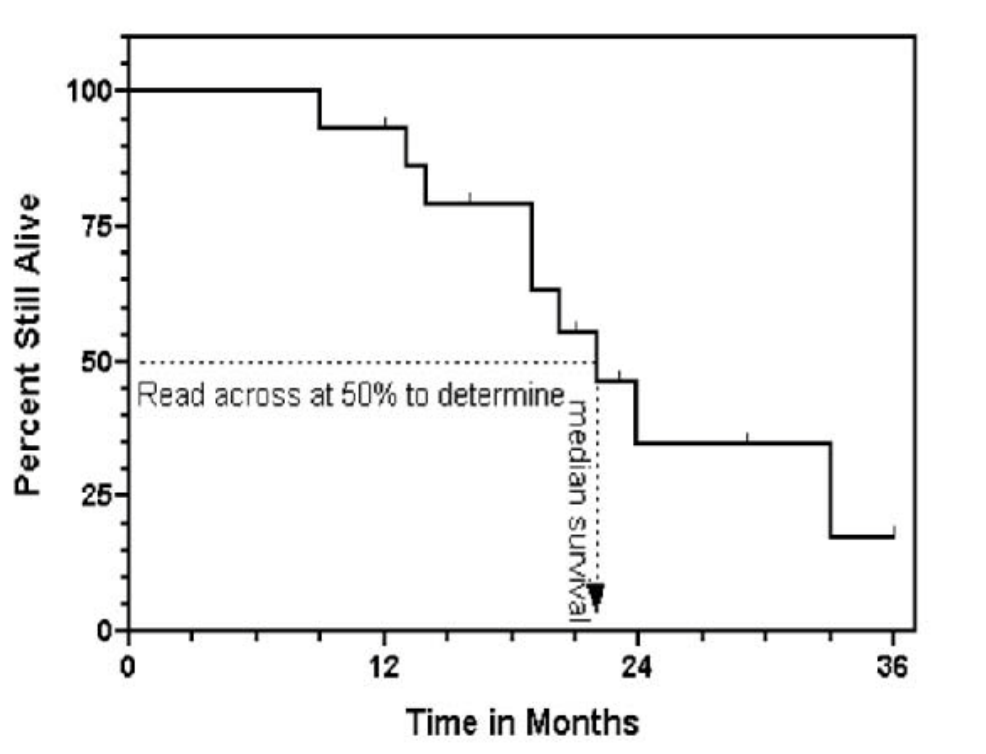
Patient says: I have lung cancer. Will I be able to attend my grand daughter's graduation a year from now if I start with a new chemo therapy today?

Doctor says: Yes…the median survival time is 22 months with a new treatment.

Patient replies: What is does that mean?

Survival curve: median survival time

- Follow-up time 36 months.
- Median survival time: the time at which half the subjects have died and half are still alive. It isn't estimable if the survival curve doesn't reach to 50% or below.



USF HEALTH CENTER FOR EVIDENCE BASED MEDICINE

- Patient says, before I decide, I would like to randomly sample 10 patients who took the treatment and see if it worked for them at median time of 22 months (remember follow-up = 36 mo).

- Take a random sample of 10 patients:

E, NE, E, E, NE, E, NE, E, E, E

7 Effective, 3 Not Effective

Question: Is the treatment effective?

Characteristics of this Problem

1. Population: All treated lung cancer patients
2. Unknown Population Param: Proportion effective
3. Null Hypothesis : Proportion = 0.5
4. Alternative Hypothesis : Proportion > 0.5
5. Sample: 10 randomly selected patients
6. Sample Statistic : 7 effective

- Question: Is 7 or more effective treatments in 10 patients unusual?

What is the probability of 7 or more effective

Treatments, if the treatment is not effective i.e.

Pr(effective) = Pr(not effective) = .5

Note: Chance is <u>somewhat</u> responsible for the variation in the results.

Ten Randomly Selected Patients

x	Prob of x or more effective
6	0.377
7	0.172
8	0.054
9	0.011
10	0.001

Unlikely due to chance alone

Interpretation: If we select 10 patients at random, the probability of observing 7 or more successes (effective) is 0.172, assuming no difference (that treatment is not effective). IOW, we are17.2 % likely to observe this by chance alone and we cannot pin it on the treatment itself!

- Statistical inquiry is designed to determine whether the unexpected is attributable to chance alone or to another cause (treatment effects).
- P-value: If the null hypothesis is true (if the treatment was not effective) and we were to repeat the experiment (select 10 patients many times over), the probability of obtaining 10 consecutive effective treatments is 1 in 1000 by chance alone. So it should not be chance.
- Misconception: If p = 0.05, the null has only 5% of being true. Remember, we are assuming the null is true all along!

Important Point

- A statistician can not prove that a treatment is effective or not. Since we are always inferring about a population from a sample, we can never be certain. However, we can quantify that uncertainty and assess the strength of the evidence.
- P-values measure the strength of evidence.
- Our Motto: Being a statistician means never having to say you're certain.

Rational Strategy

- Reject the null hypothesis in favor of the alternative (and say treatment is effective) if the probability of getting a result obtained or one more extreme when the null hypothesis is true is less than α for some pre-assigned small value of α (i.e. p-value < α).
- α is called “significance level, often $\alpha = 0.05$.

To sum up….

- We always test the **null hypothesis (Ho), which assumes no** effect (e.g. the difference in means equals zero) in the population.

Proportion (effective) = 0.5

- We then define the **alternative hypothesis (H_A), which** holds if the null hypothesis is not true.

Proportion (effective) ≠ 0.5

Specifies direction

Example

A randomized, double-blind controlled trial was carried out to the study the prophylactic effect of inhaled corticosteroids (beclomethasone) on wheezing episodes associated with viral infection in school age children, over a 6 month period. Outcome of interest is forced expiratory volume (FEV).

Treatment	Placebo
Size = 50	Size = 48
Mean = 1.64	Mean = 1.54
Std Dev = 0.29	Std Dev = 0.25

Null: Population means are equal,

Alternative: Not equal (two sided)

Question: Why not choose one sided? (Principle of Equipoise)

- Question: Why not choose one sided test?

- Principle of Equipoise (clinical uncertainty): A subject may be enrolled in a trial only if there is a true uncertainty about the trial outcome.

- One sided test would imply that we know something about the treatment effectiveness.

In This Case

1. Population: All infected school age children
2. Unknown Population Parameters: Mean forced expiratory volume
3. Null Hypothesis: Mean changes are equal
4. Alternative Hypothesis: Mean changes are unequal
6. Decision: Reject the Null Hypothesis if p-value < α.
7. Two sided P-value = 0.0711. Fail to reject null hypothesis and conclude A and B do not differ..

Going back to the “chemo treatment effectiveness”

x	Prob of x or more effective	Decision*
6	0.387	don’t participate
7	0.172	don’t participate
8	0.054	don’t participate
9	0.011	participate
10	0.001	participate

*Based on significance level of 0.05

Decision Rule

- Reject the null hypothesis (that the treatment is ineffective) if you get 9 or 10 effective out of 10 randomly selected patients.

- This gives a significance level of 0.05; i.e. a probability of < 0.05 of rejecting the null hypothesis if it is true.

What to do?

There are 2 types of error that I can make and we'd like the probabilities of both to be low.

Four possible outcomes under the hypothesis testing paradigm

		The State of Nature	
		The Null is True	The Null is False
Decision	Reject the Null Hypothesis	Type I Error or alpha (Rejecting a true null)	Correct Decision
	Fail to reject the Null Hypothesis	Correct Decision	Type II Error or beta (failing to reject a false null)

USF HEALTH CENTER FOR EVIDENCE BASED MEDICINE

Confidence Intervals

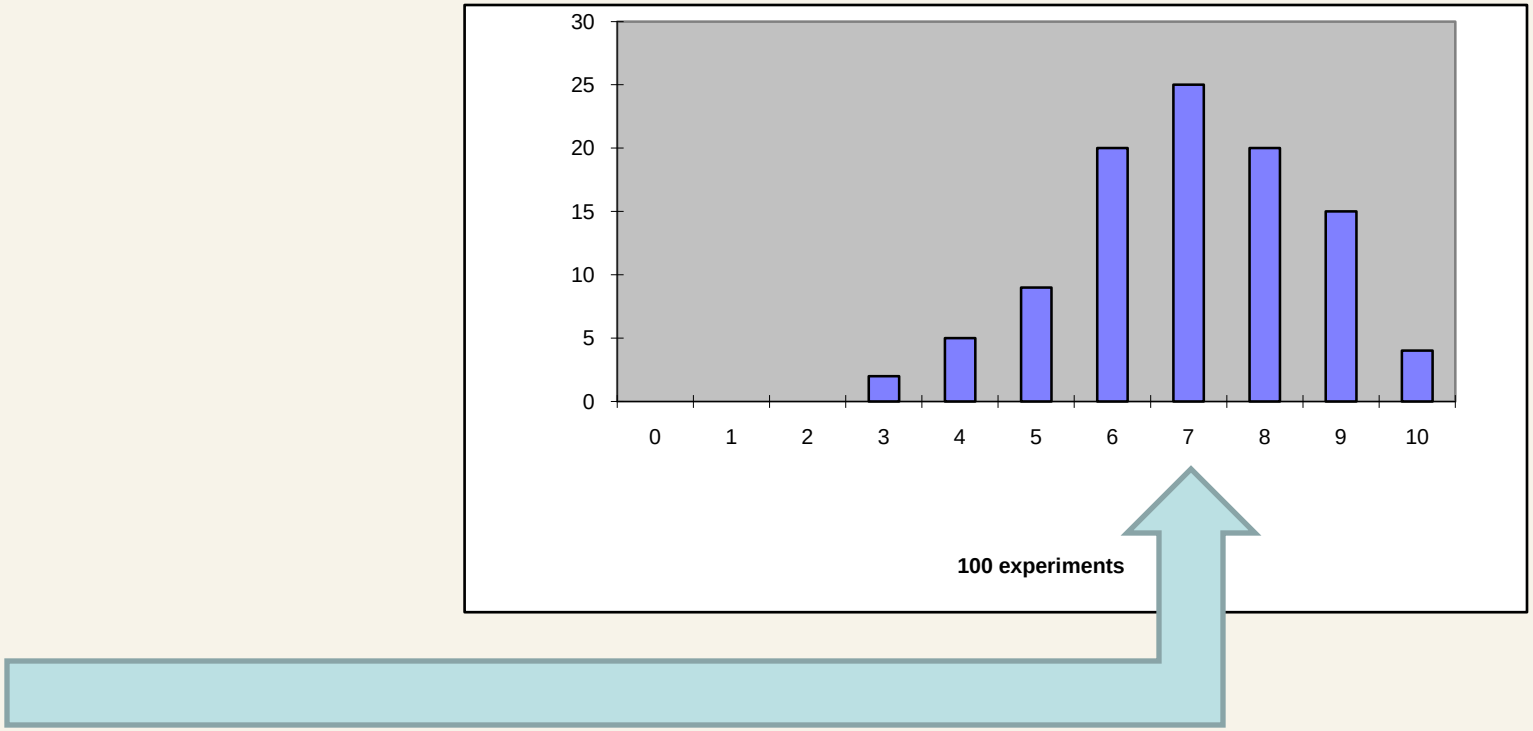
(A better) alternative to Hypothesis Testing Using P-values

- Lets assume the treatment is ineffective (that null is true) and start sampling.

- Take a random sample of 10 patients 100 and 500 times respectively

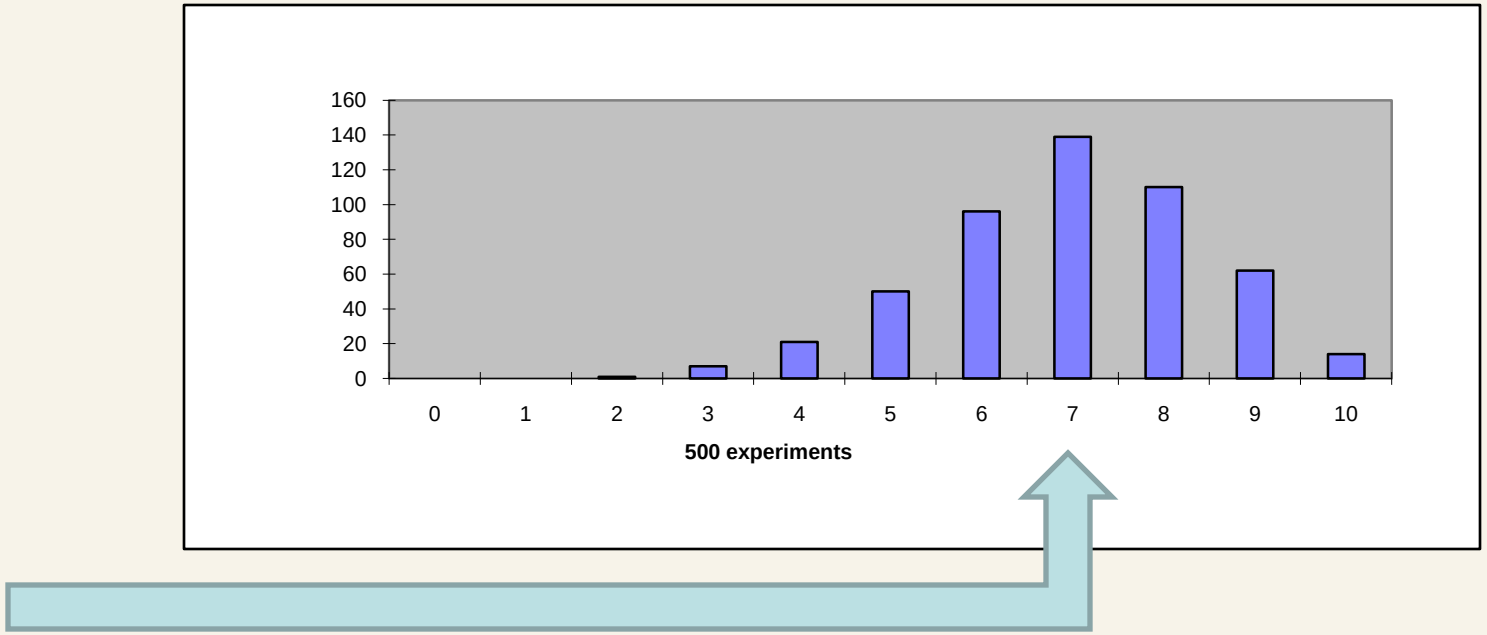
Success vs Frequency for 100 random patients

Observe approximately normal distribution



Mean number of success = 7/10 = 0.7

Success vs Frequency for 500 random patients

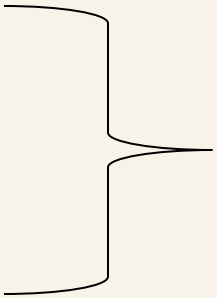


Mean number of success = 7/10 = 0.7

- Since we have a normal distribution, we can use Empirical Rule and construct a confidence interval around the observed proportion of 0.70, which is

$$0.70 \pm 1.96 * \text{Standard Error}.$$

- After 100 : (0.61, 0.79)
- After 500 : (0.76, 0.74)



Why different?

- Note 0.5 is not in the interval, so the treatment is effective at 95% level of confidence.

- What does “the treatment is effective at 95% level of confidence” mean?

- If we repeated a study a large number of times, 95% of the estimates (means, proportions…) would be within 1.96 standard errors of the true mean. This is called the confidence interval.

Note 1: Point estimate ± two standard errors has approximately 95% coverage for a wide variety of distributions and for sample sizes n > 20. This (oddly enough) works only for 95% coverage.

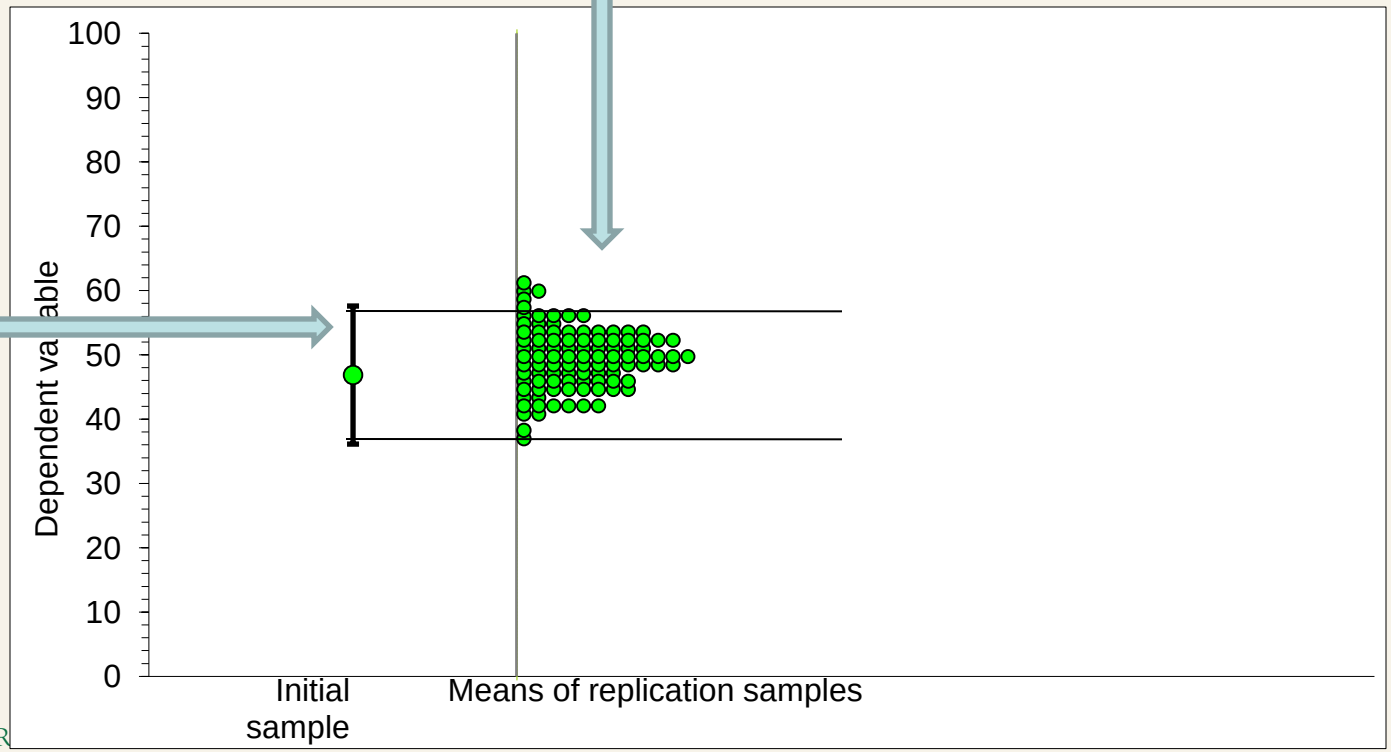
Note 2:

- 95% Confidence interval: If we repeat a study a large number of times, 95% of the estimates would be within 1.96 standard errors of the true mean.
- Below is a 95% CI and sample means 100 independent samples of size 30 simulated from normal distribution with mean $\mu = 50$ and std. dev $\sigma = 30$. Not that approximately three are outside.

100 subsequent sample means

Initial 95% CI:

(35.3, 58.5)



- On average, if you repeatedly calculate 95% CI, 95% of them will contain the population parameter of interest.

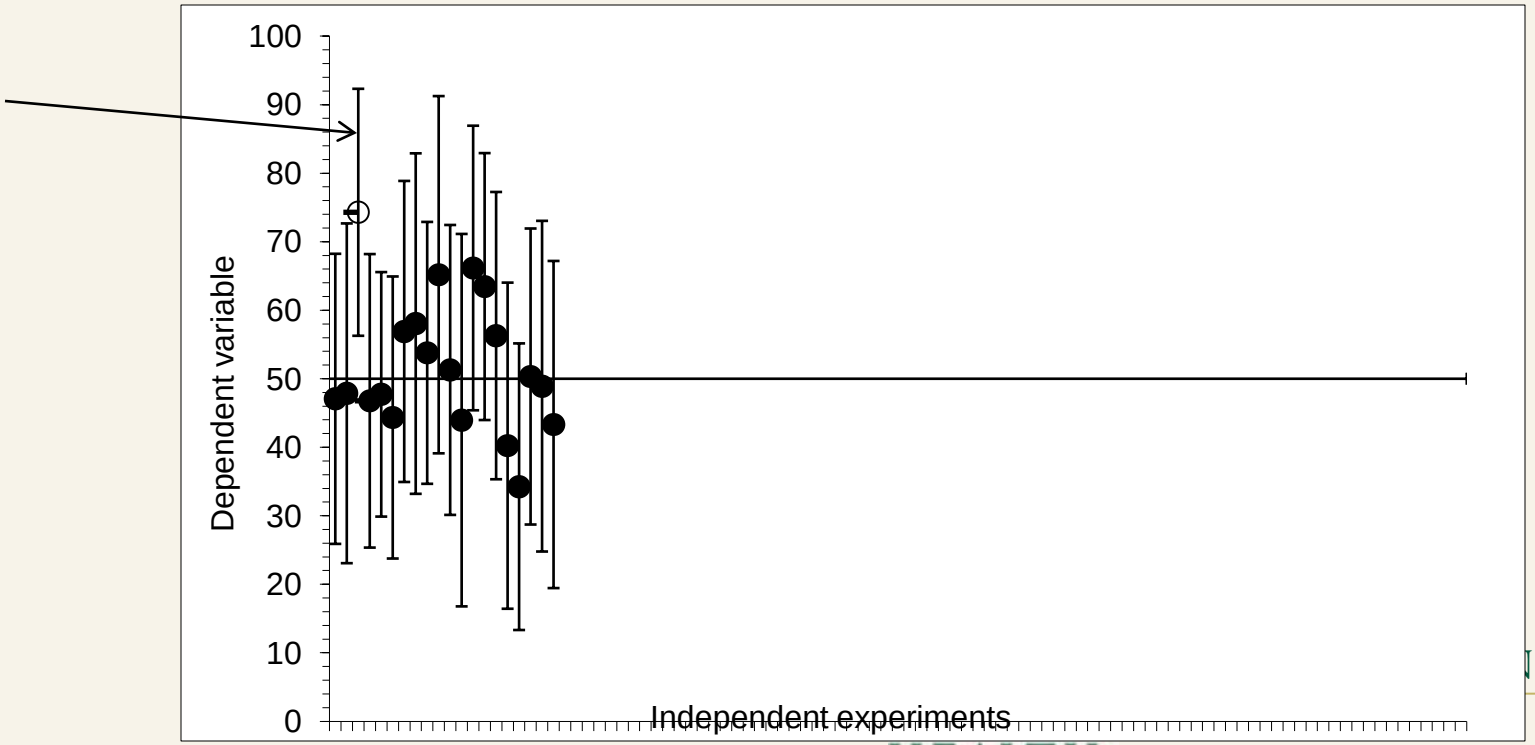
- It does not make sense to talk about the probability that a CI contains the parameter of interest (it either does or does not, i.e. probability is either 1 or 0).

Note 3:

- Below are 20 independent samples of size 20 simulated from normal distribution with mean μ = 50 and std. dev σ = 60. Note that 19 (or 95%) CI's contain the true mean = 50 and 1 does not.

Does not contain

μ



- Why are we so attached to 95% confidence?

An empirical study by Cowles & Davis (American Psychologist (1982), Vol 37, No. 5, pg. 553) suggested that for humans the “threshold of dismissal of the idea of chance is odds of 1 in 20.”

Example: The CDC reported

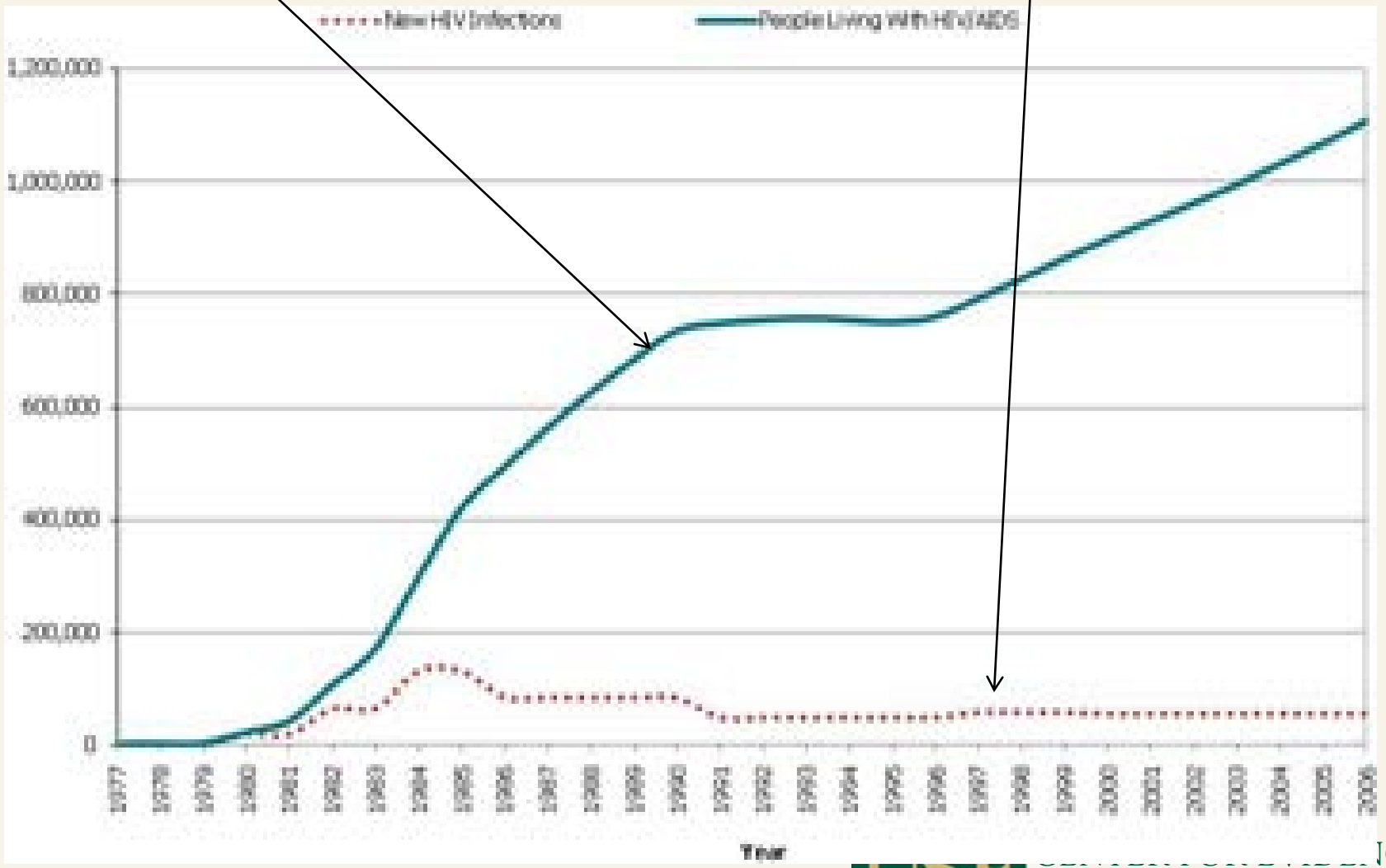
- "Based on the stratified extrapolation approach the incidence of HIV in the US for 2006 was 56,300 new infections (with a 95% confidence interval of 48,200 to 64,500) "

- Prevalence is the number of people living with HIV infection at the end of a given time period.

- Incidence is the number of new infections that occur during a given time period.

- Fatality Rate = Number of those who died within a given group.

Rates are often defined as person-year: If 100 people are followed for 6 years, we accumulate 600 person-years. If we observe 60 events, then the incidence rate is 60/600 = 0.1 events per person year.


Prevalence
Incidence
New HIV Infections
People Living With HIV/AIDS
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
Year

True or False

A clinical trial to compare a mouthwash against a control found a difference in plaque score after 1 year of 1.1 units, *P = 0.006 (two sided). Which of the following is true?*

(a) The probability that the null hypothesis is true is 0.006.

(b) If the null hypothesis were true, the probability of getting an observed result of 1.1 or greater is 0.006.

(c) The alternative hypothesis is a mean difference of 1.1.

(d) The probability of the alternative hypothesis being true is 0.994.

True or False

A clinical trial to compare a mouthwash against a control found a difference in plaque score after 1 year of 1.1 units, *P = 0.006 (two sided). Which of the following is true*

(a) The probability that the null hypothesis is true is 0.006.

(b) If the null hypothesis were true, the probability of getting an observed result of 1.1 or greater is 0.006.

(c) The alternative hypothesis is a mean difference of 1.1.

(d) The probability of the alternative hypothesis being true is 0.994.

Questions?