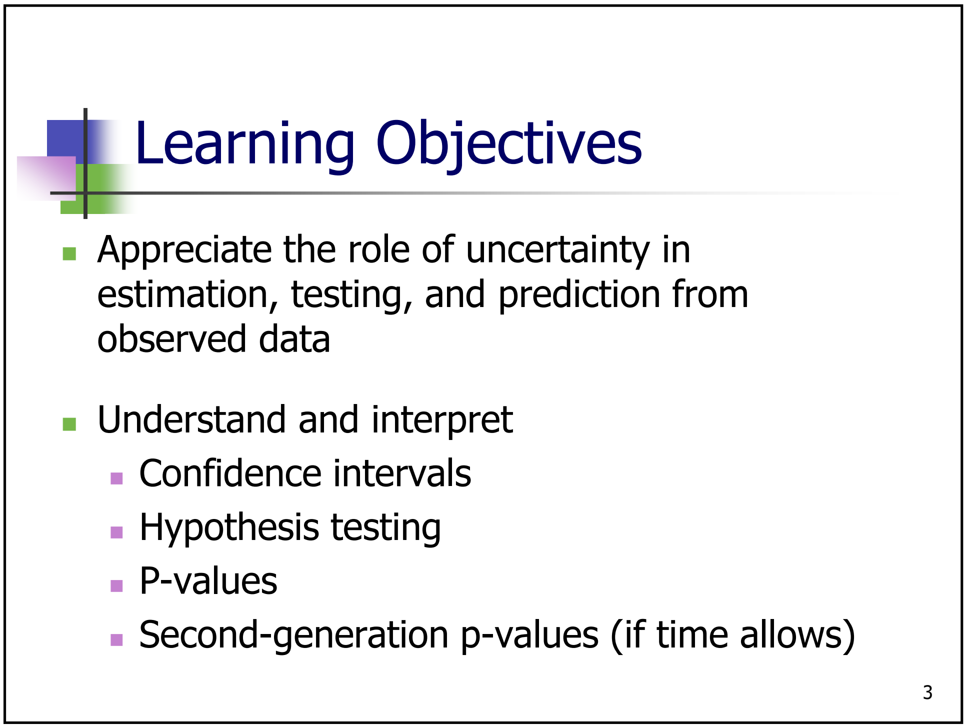
Review of Statistical Concepts for Imaging Sciences

Jeffrey D. Blume
School of Data Science
University of Virginia

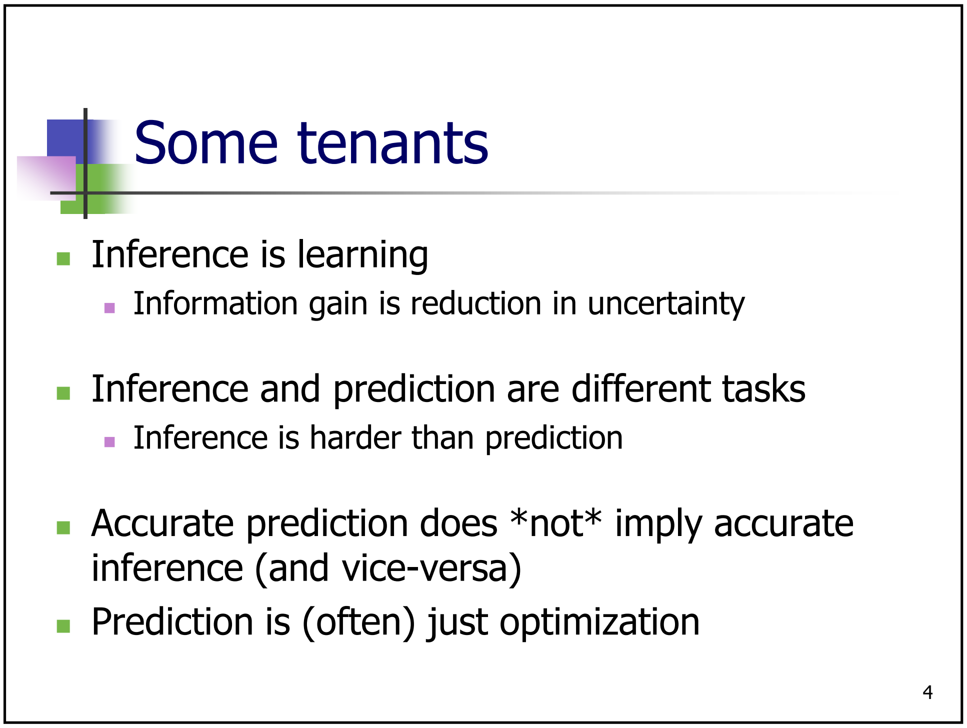
Acknowledgements

- Many iterations
- Thanks to
  - Diana Miglioretti for some imaging examples.
  - Todd Alonzo for improving these slides.
  - Nancy Obuchowski for designing the talk.



Learning Objectives

- Appreciate the role of uncertainty in estimation, testing, and prediction from observed data
- Understand and interpret
  - Confidence intervals
  - Hypothesis testing
  - P-values
  - Second-generation p-values (if time allows)



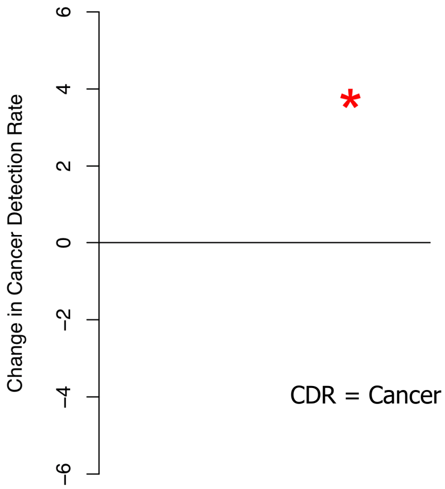
Some tenants

- Inference is learning
  - Information gain is reduction in uncertainty
- Inference and prediction are different tasks
  - Inference is harder than prediction
- Accurate prediction does *not* imply accurate inference (and vice-versa)
- Prediction is (often) just optimization

Uncertainty unlocks information

- Is digital mammography more ‘accurate’ than film-screen at detecting incident breast cancer in screening population?
- One radiologist’ s or facility’ s experience might suggest digital mammography detects more cancers and/or reduces false positives.
  - Is this true for the general population of eligible women, radiologists, and facilities?

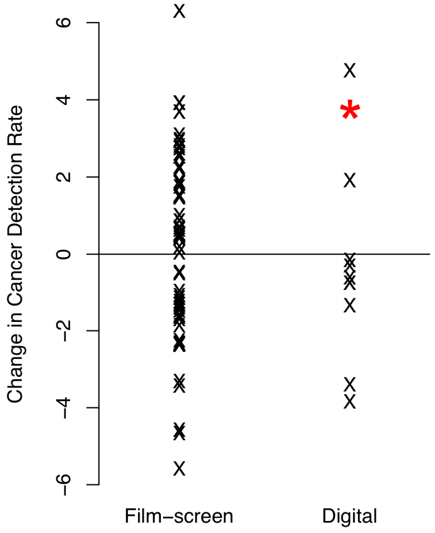
A published study argued digital mammography improves CDR based on a single facility’ s observed change after conversion.



CDR = Cancer Detection Rate

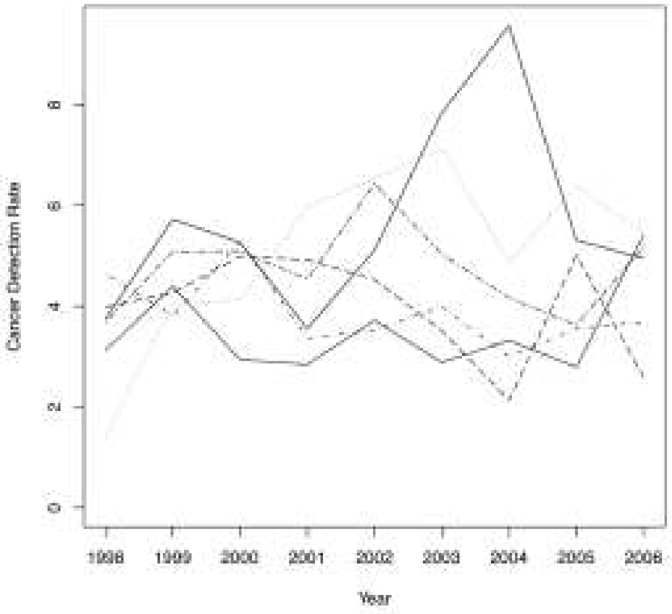
Vernacchia, et al. *AJR*, 2009

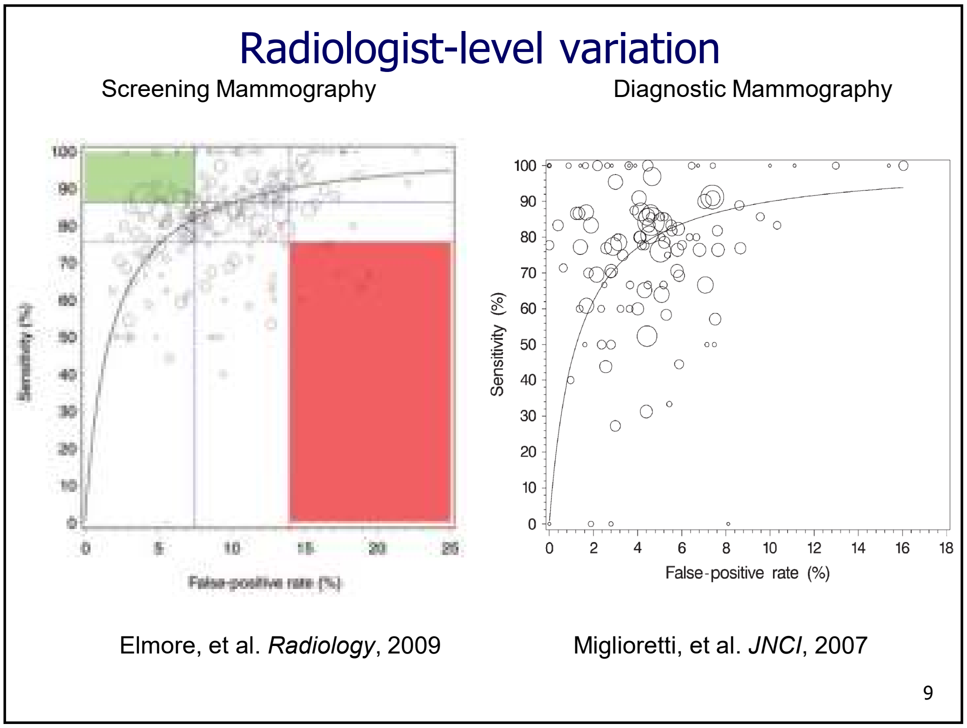
Important to consider *variation* within and between facilities.



Observed change in CDR from BCSC facilities that did and did not switch to digital.

CDR for six randomly selected BCSC facilities exclusively performing film-screen mammography from 1998 – 2006.





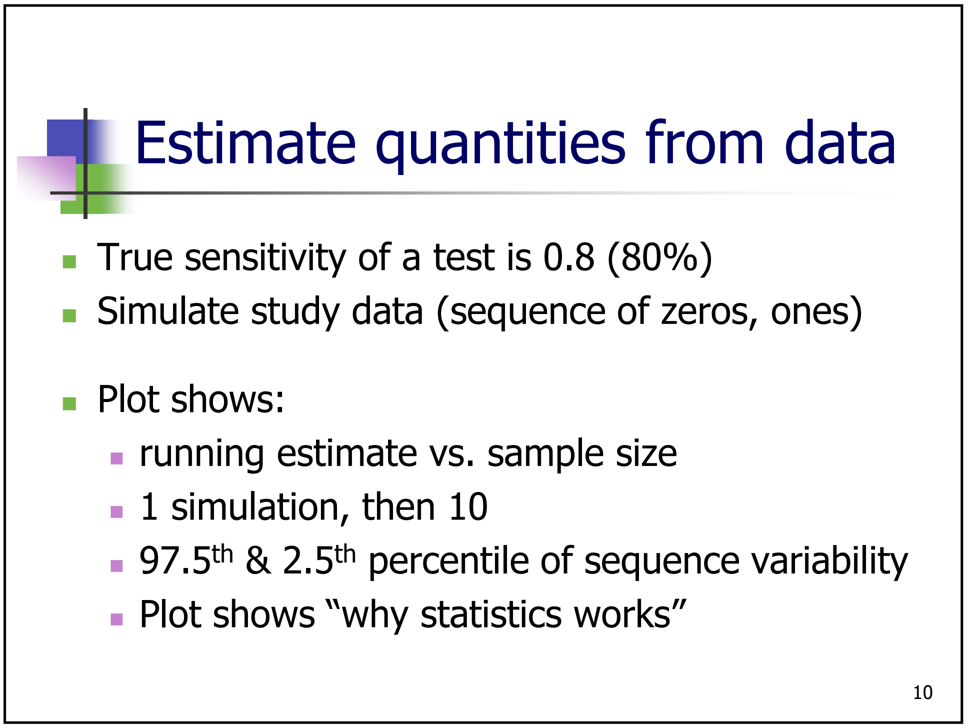
Radiologist-level variation

Screening Mammography

Diagnostic Mammography

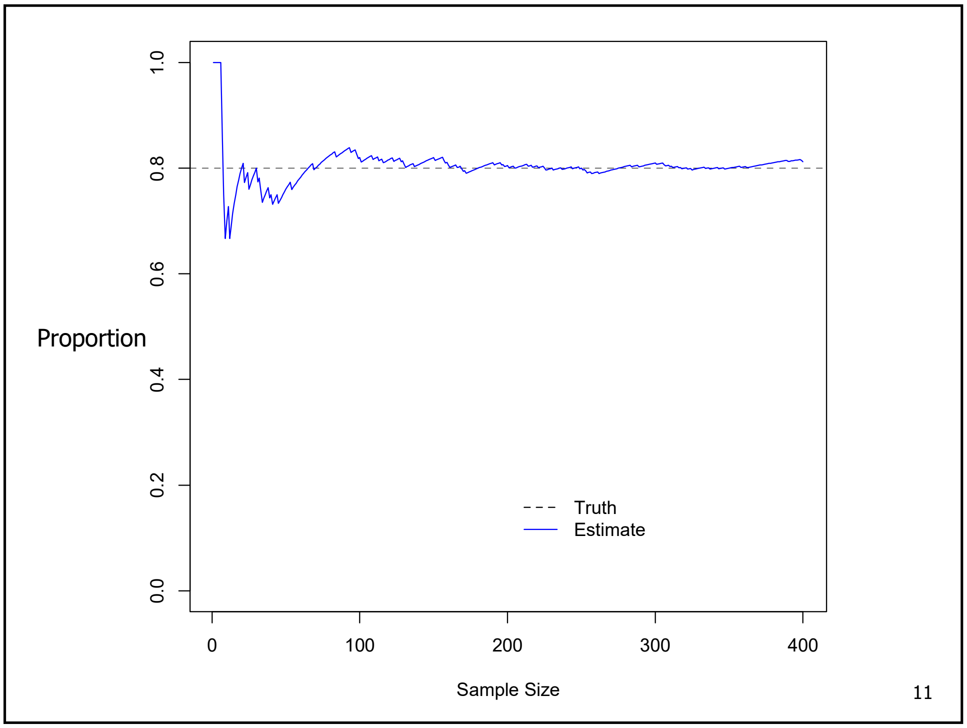
Elmore, et al. *Radiology*, 2009

Miglioretti, et al. *JNCI*, 2007

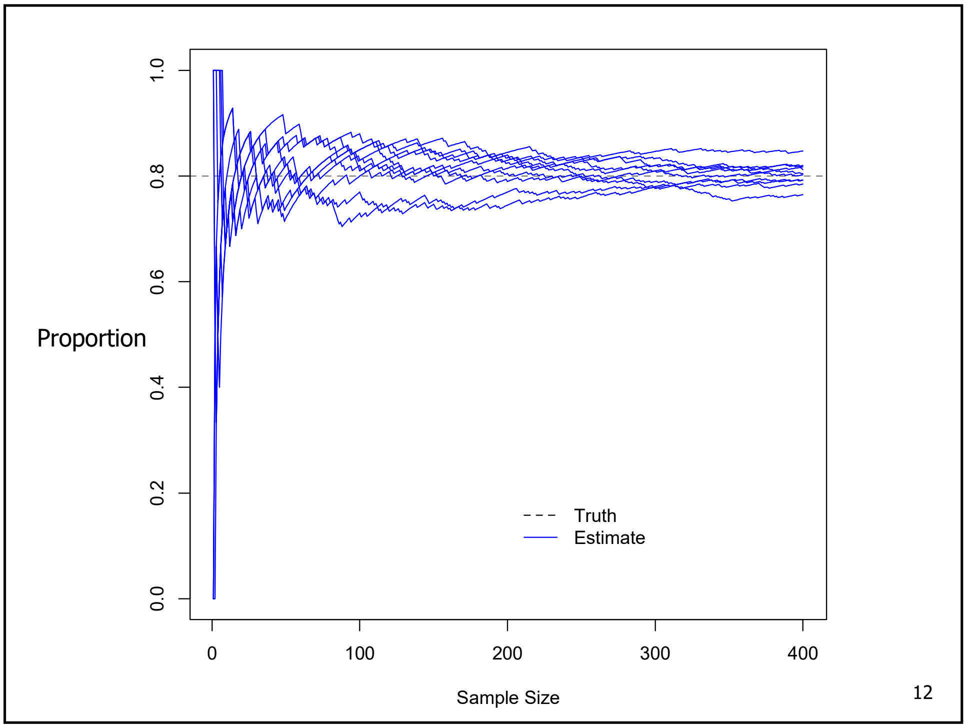


Estimate quantities from data

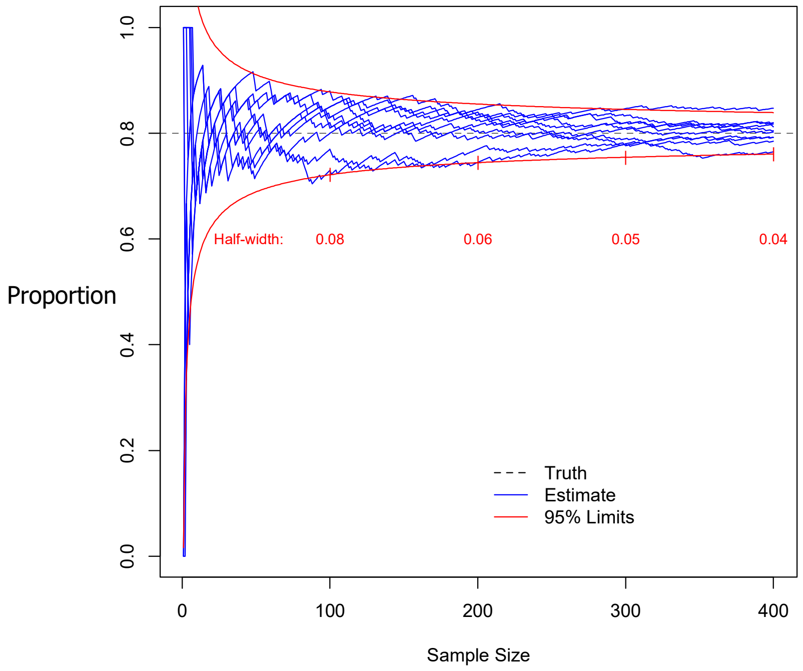
- True sensitivity of a test is 0.8 (80%)
- Simulate study data (sequence of zeros, ones)
- Plot shows:
  - running estimate vs. sample size
  - 1 simulation, then 10
  - 97.5th & 2.5th percentile of sequence variability
  - Plot shows “why statistics works”


Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
Sample Size
Truth
Estimate
11

11


Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
Sample Size
Truth
Estimate
12

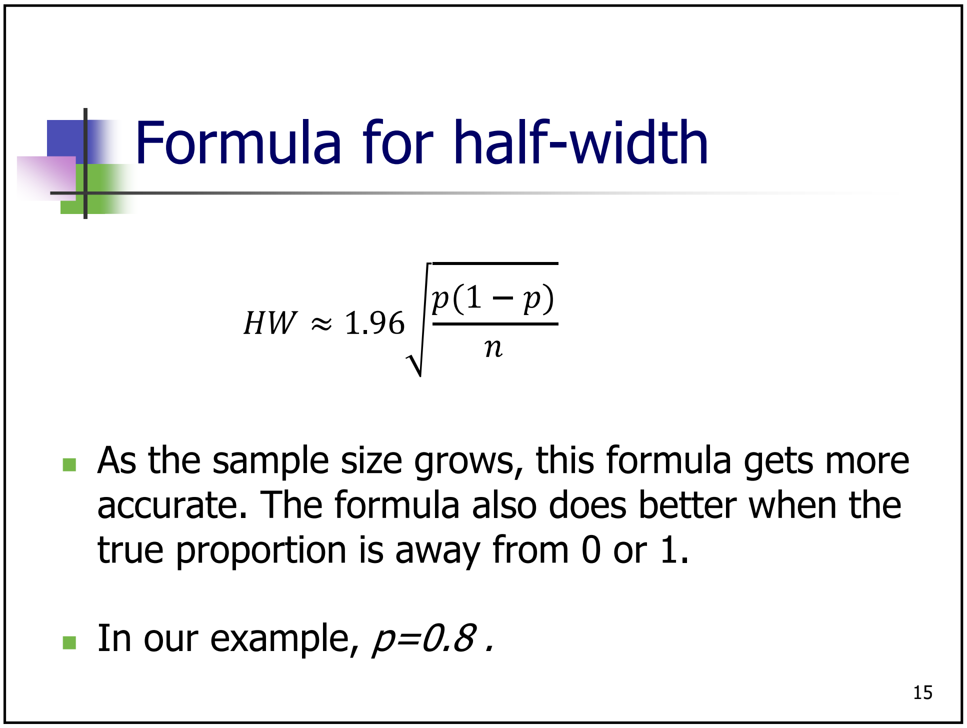
12



Relation to Sample Size

- As the sample size grows…
  - the sequences become more concentrated near the true proportion (in this case, 0.8)

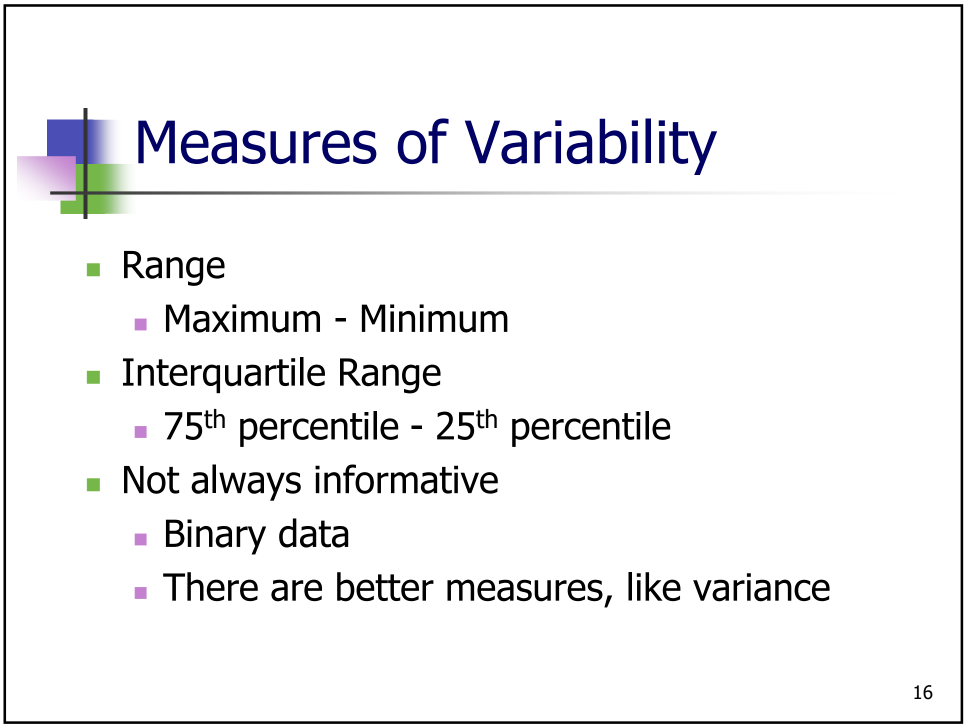
- Red lines comes from theory and …
  - captures 95% of the sequences at each sample size
  - shows the "half-width" ; the distance from the true proportion to the red line
  - illustrates a sample size projection



Formula for half-width

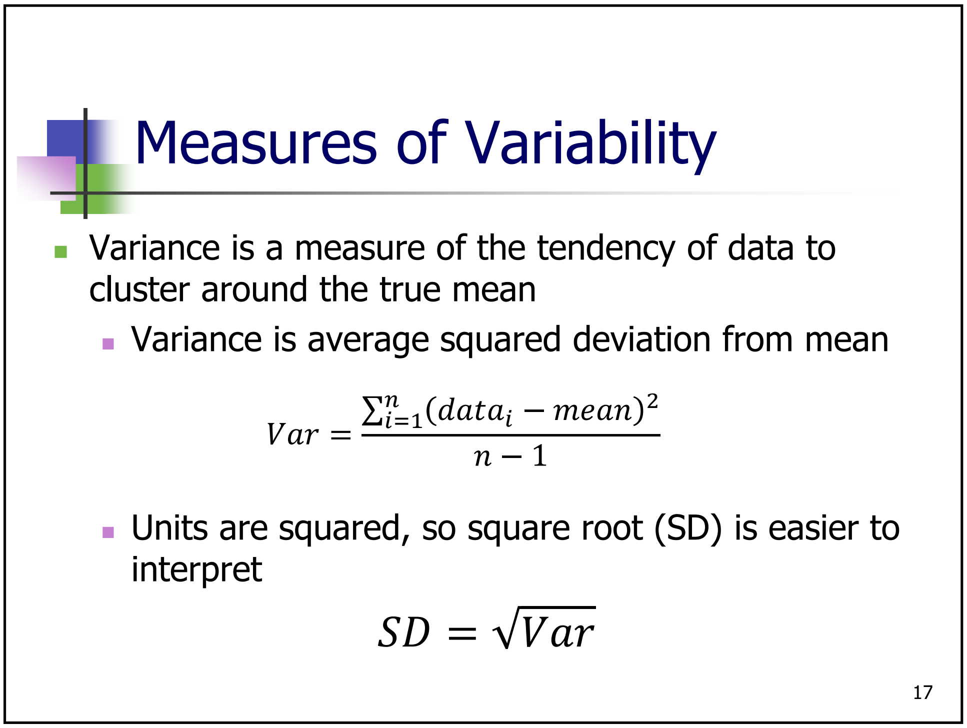
$$HW \approx 1.96\sqrt{\frac{p(1-p)}{n}}$$

- As the sample size grows, this formula gets more accurate. The formula also does better when the true proportion is away from 0 or 1.
- In our example, *p=0.8 .*



Measures of Variability

- Range
  - Maximum - Minimum
- Interquartile Range
  - 75th percentile - 25th percentile
- Not always informative
  - Binary data
  - There are better measures, like variance



Measures of Variability

- Variance is a measure of the tendency of data to cluster around the true mean
  - Variance is average squared deviation from mean

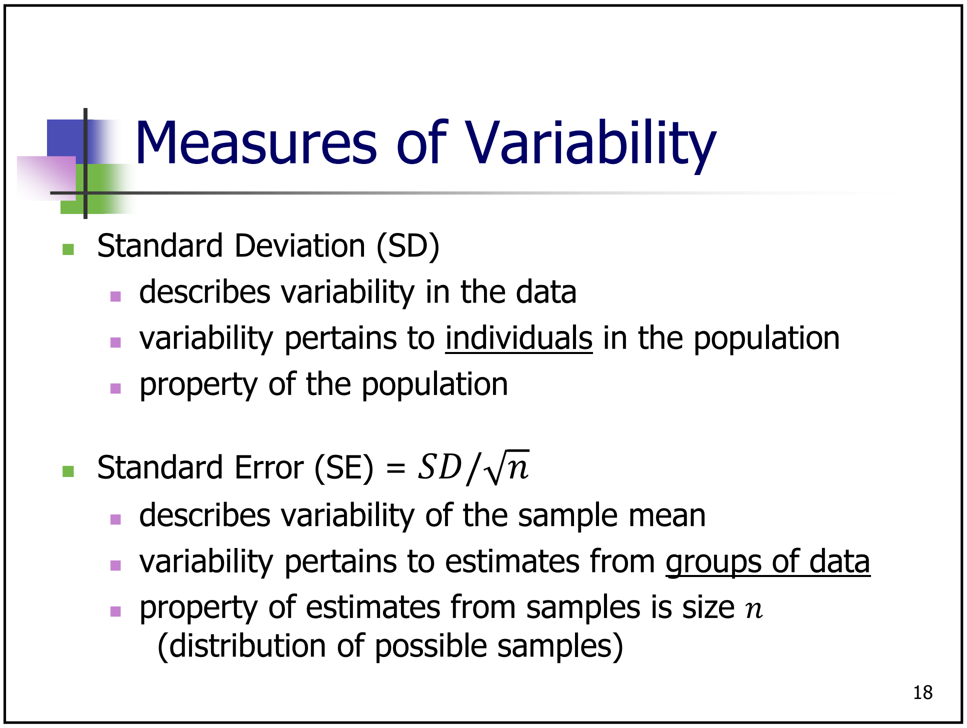
$$Var = \frac{\sum_{i=1}^{n}(data_i - mean)^2}{n-1}$$

  - Units are squared, so square root (SD) is easier to interpret

$$SD = \sqrt{Var}$$

17

17



Measures of Variability

- Standard Deviation (SD)
  - describes variability in the data
  - variability pertains to <u>individuals</u> in the population
  - property of the population

- Standard Error (SE) = $SD/\sqrt{n}$
  - describes variability of the sample mean
  - variability pertains to estimates from <u>groups of data</u>
  - property of estimates from samples is size n (distribution of possible samples)

18

18

Why focus on the mean?

- Good example for illustrating general principles
- Proportions and rates are means
- Estimates of complicated quantities often behave like means
- Means are not perfect; sensitive to outliers and population skewness

- Alternatives: median (middle value when ordered) and mode (most frequent value)

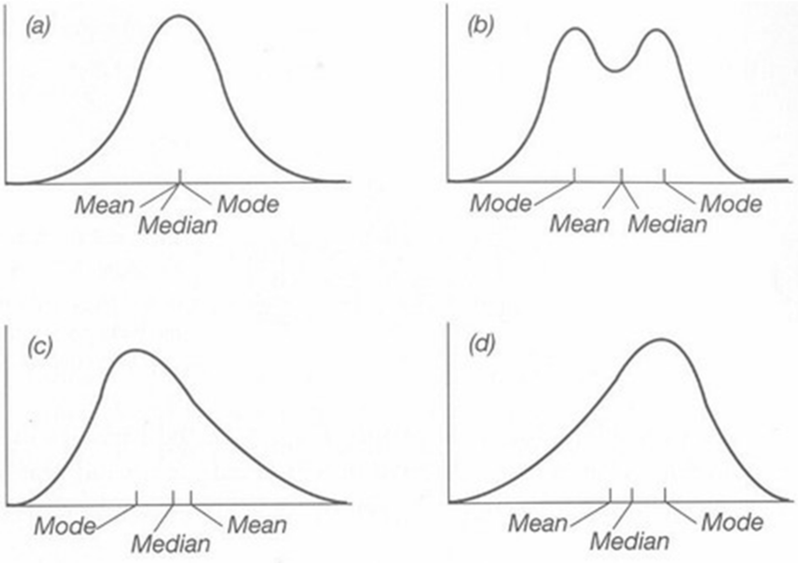


Figure 3.2 Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is positively skewed, and (d) is negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution

Combining Estimates & Variability

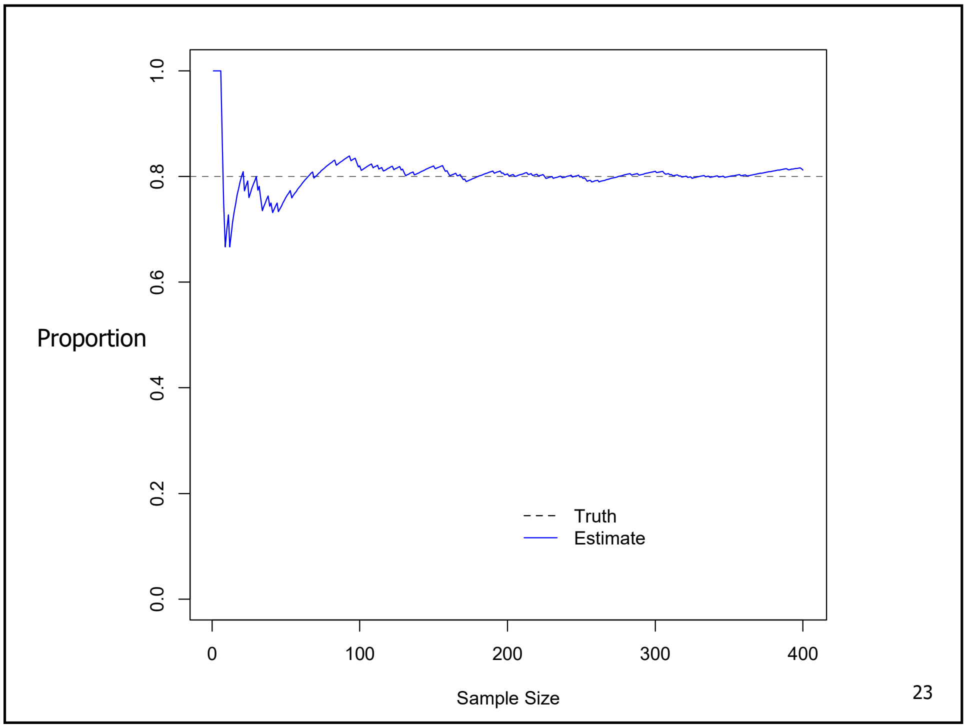
- An estimate alone is not informative
- Variability is the key
  - Low variability translates to high precision
  - High variability translates to low precision
- Confidence intervals (CI) express location and magnitude of variability
- They provide a range of estimates that are well supported by the data
  - Values in the CI are equally well supported by the data (even the pesky ones at the interval edges)

Confidence Intervals

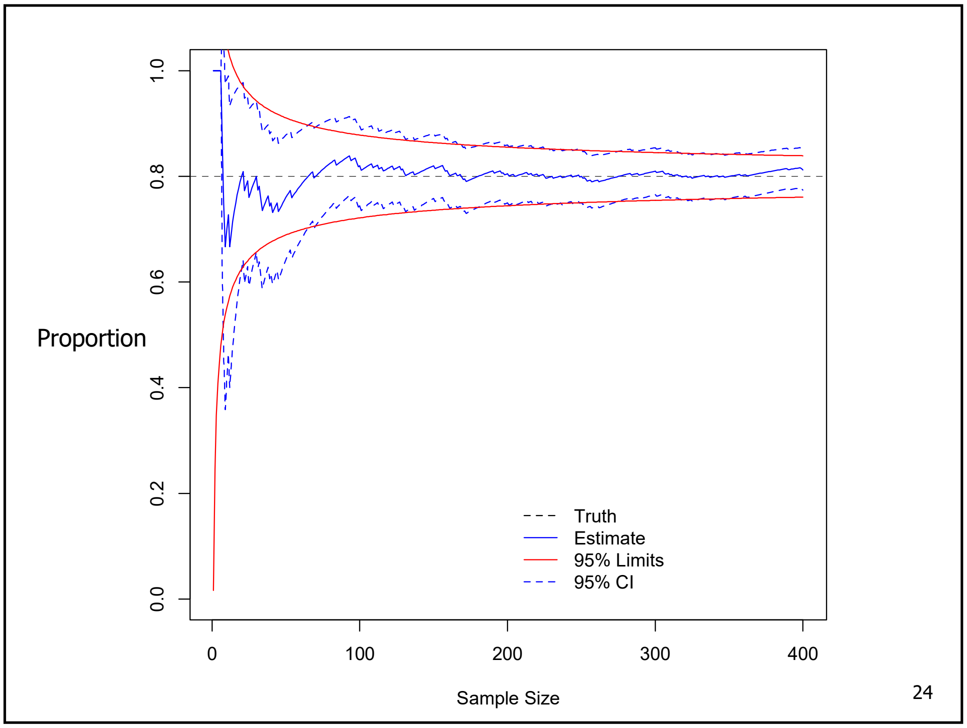
- Most 95% confidence intervals look like

Estimate ± 1.96*SE

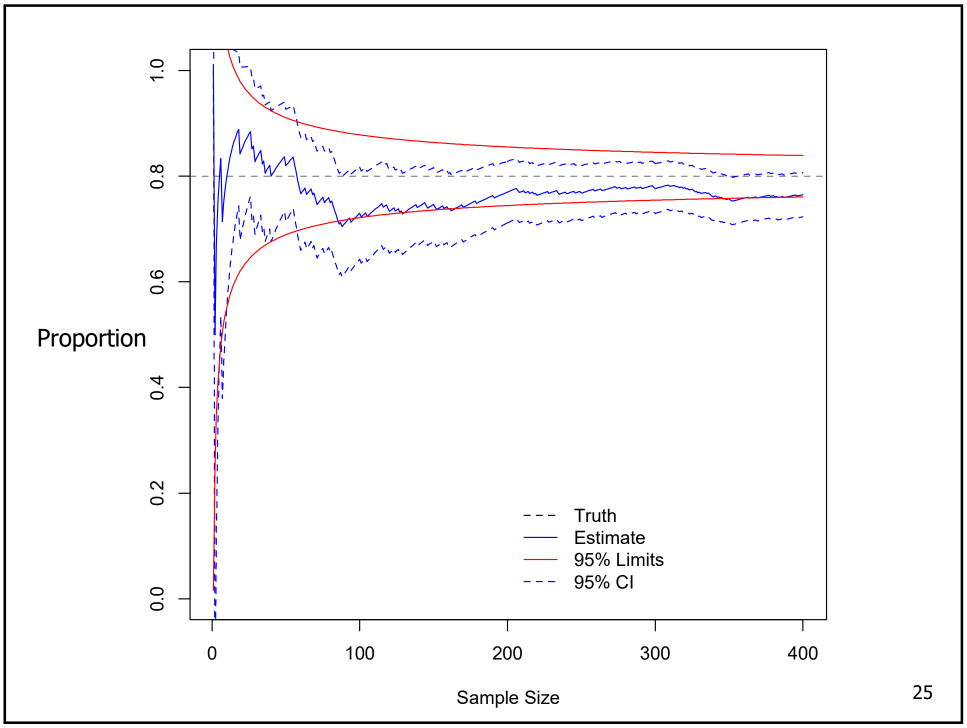
- when…
  - the sample size is ‘large enough’
  - the statistician is in a good mood
- 1.96*SE is the “margin of error” or “half-width”


Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Truth
Estimate
0
100
200
300
400
Sample Size
23

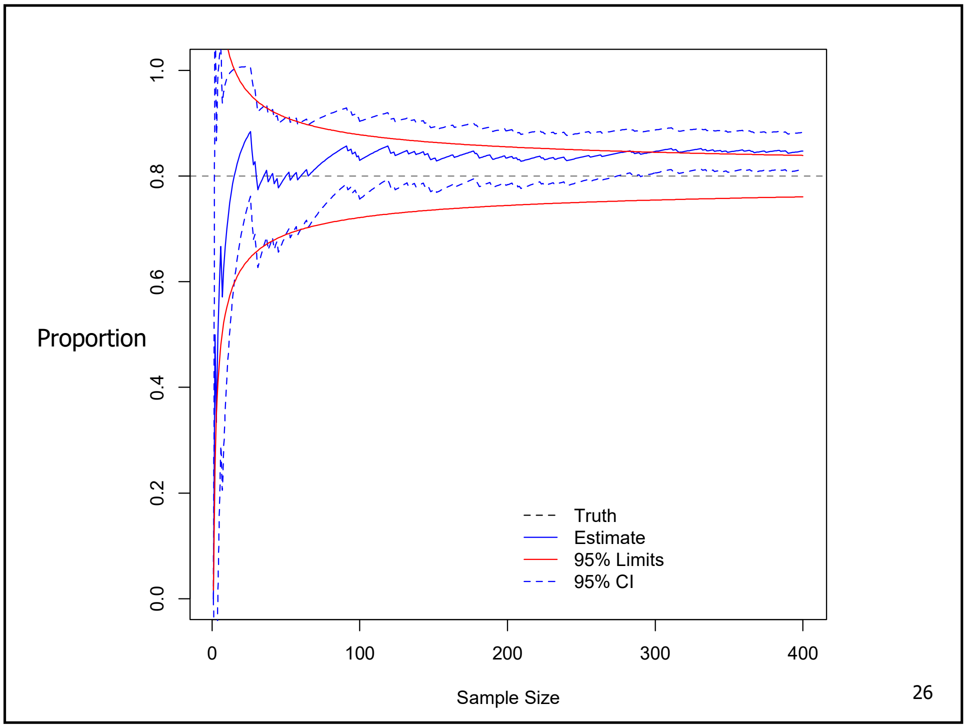
23


Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Truth
Estimate
95% Limits
95% CI
0
100
200
300
400
Sample Size
24

24

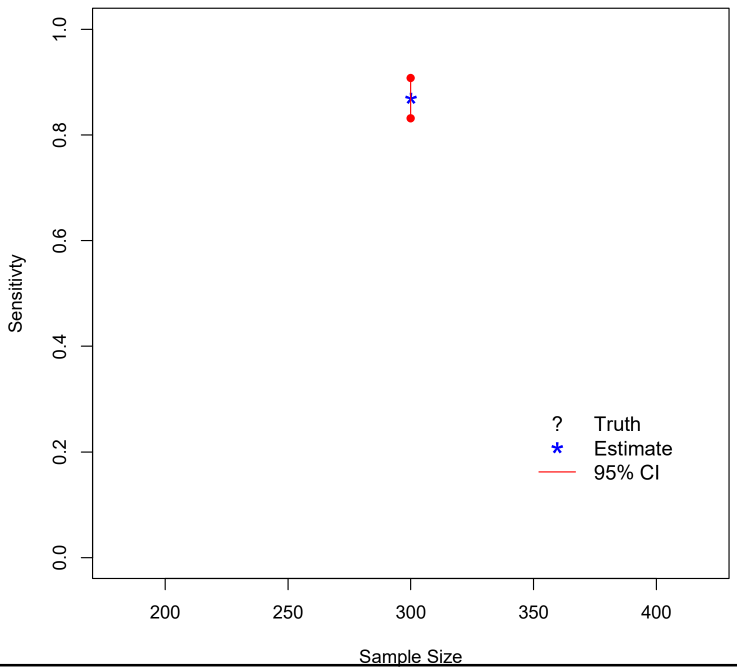

Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
Sample Size
Truth
Estimate
95% Limits
95% CI
25

25

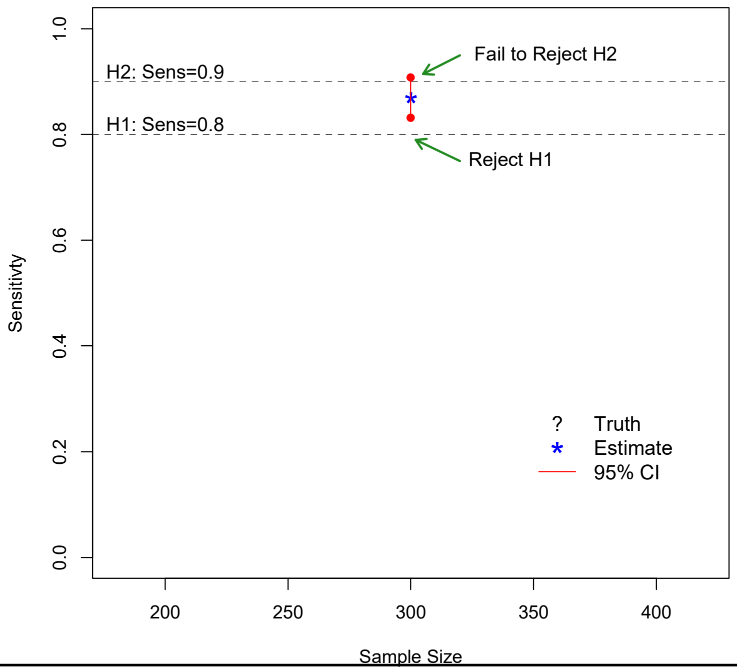

Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
Sample Size
Truth
Estimate
95% Limits
95% CI
26

26

What real results look like (n=300)...



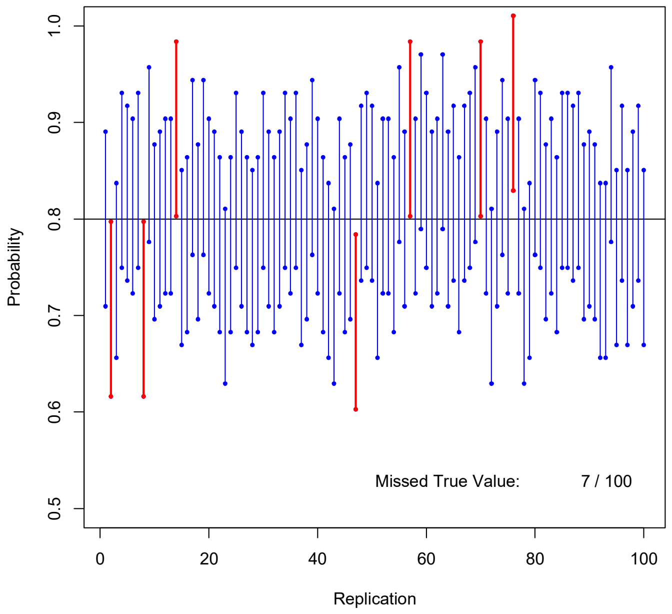
...with Hypothesis testing

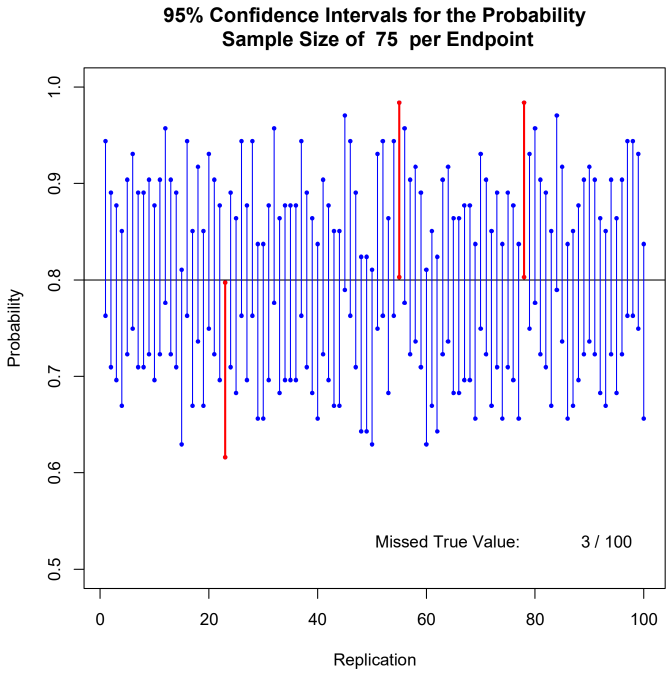


CIs can miss (bummer)

- 95% CI formulas will exclude the truth 5% of the time
- The problem is that you never know if any particular interval computed from data misses or not
- Increasing the sample size…
  - Does not change the miss rate (!)
  - Reduces the width of the CI
  - Reduces the amount by which the CIs misses the truth (on average) (!)

95% Confidence Intervals for the Probability
Sample Size of 75 per Endpoint

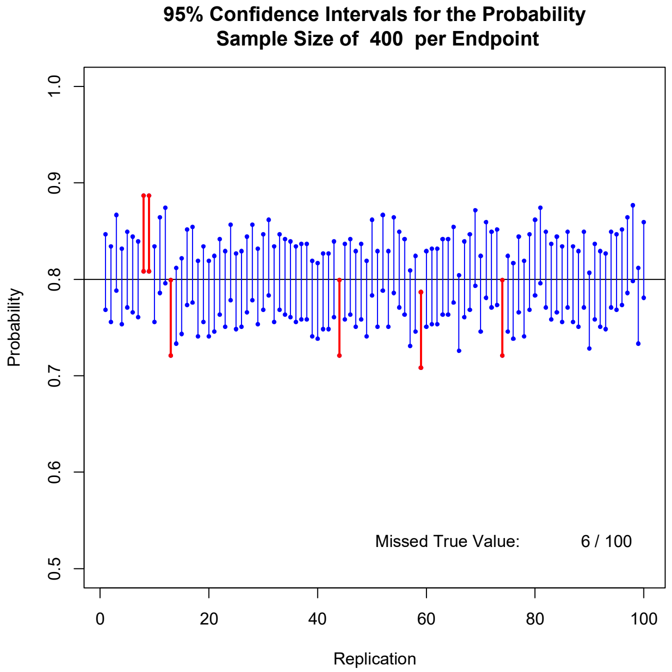




95% Confidence Intervals for the Probability
Sample Size of 75 per Endpoint

31

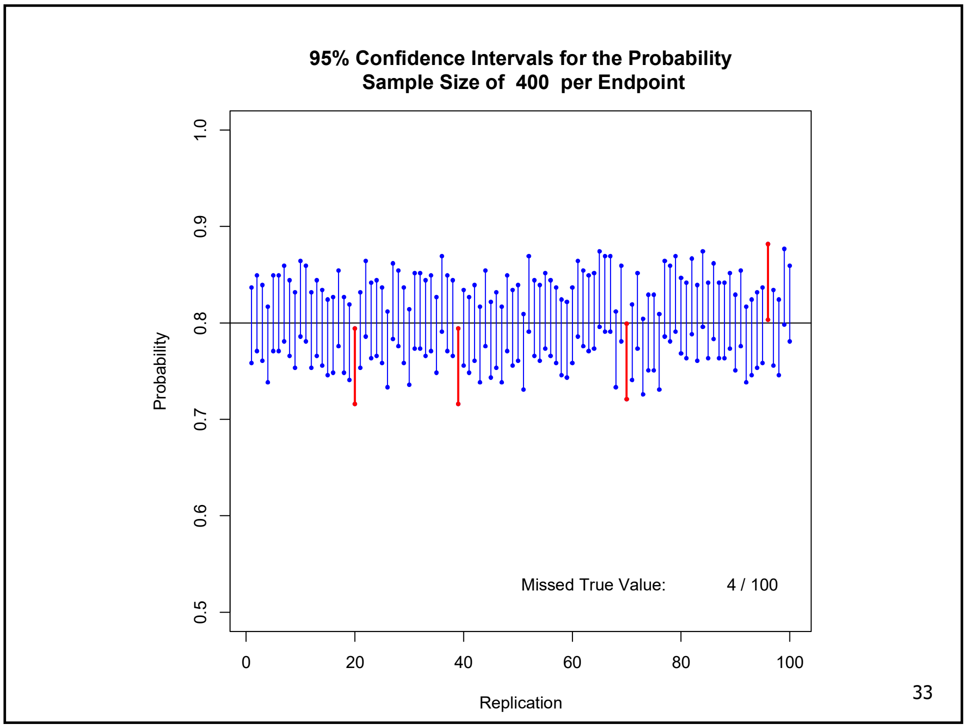
31



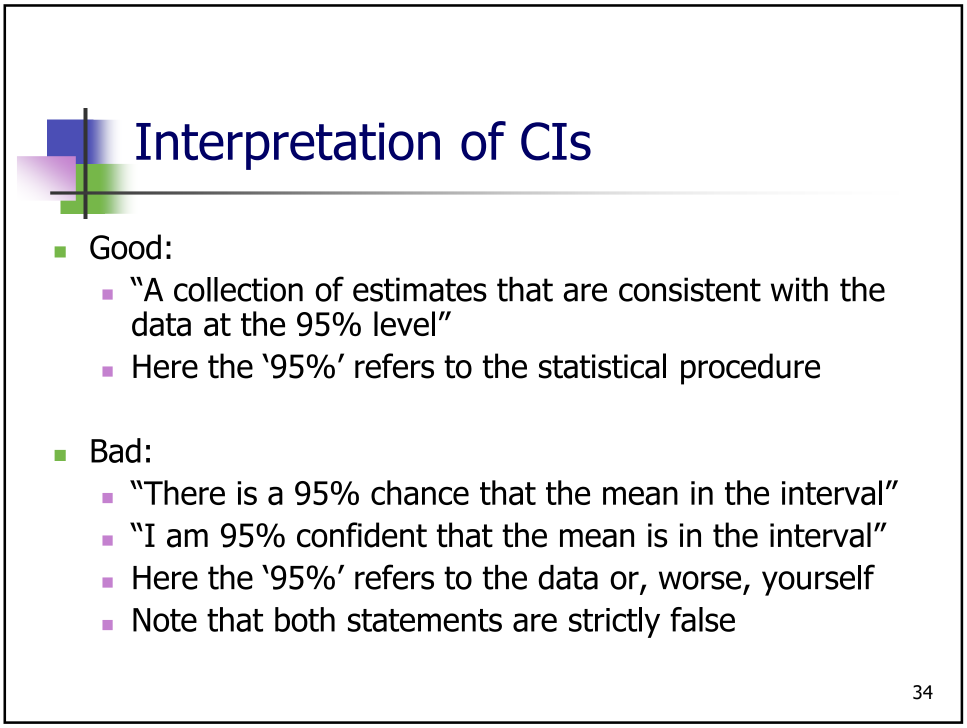
95% Confidence Intervals for the Probability
Sample Size of 400 per Endpoint

32

32

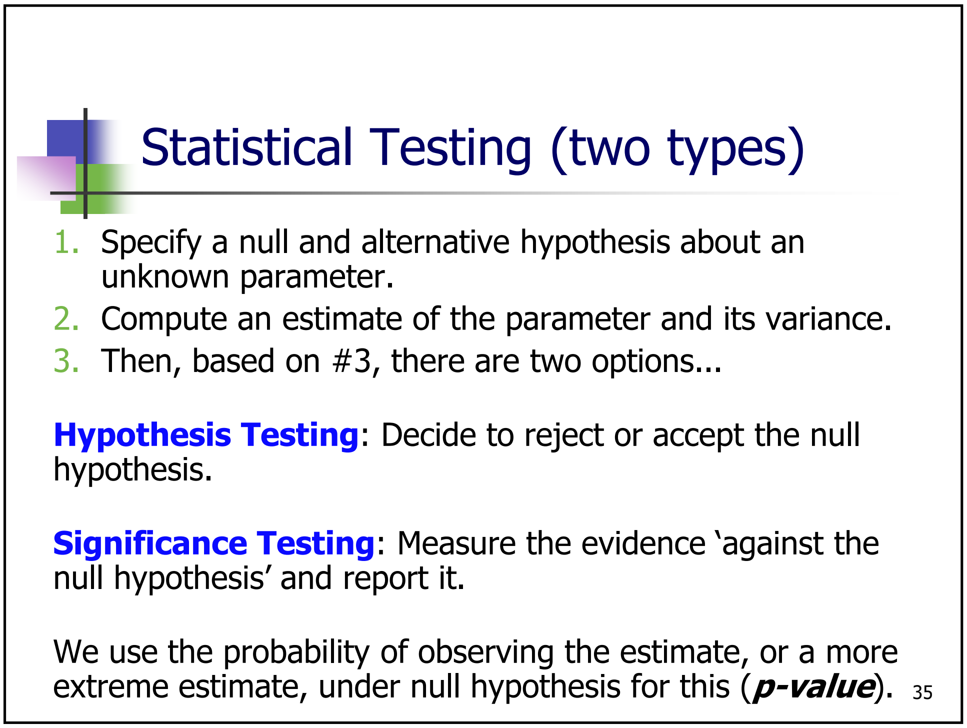


95% Confidence Intervals for the Probability
Sample Size of 400 per Endpoint



Interpretation of CIs

- Good:
  - "A collection of estimates that are consistent with the data at the 95% level"
  - Here the '95%' refers to the statistical procedure
- Bad:
  - "There is a 95% chance that the mean in the interval"
  - "I am 95% confident that the mean is in the interval"
  - Here the '95%' refers to the data or, worse, yourself
  - Note that both statements are strictly false



Statistical Testing (two types)

1. Specify a null and alternative hypothesis about an unknown parameter.
2. Compute an estimate of the parameter and its variance.
3. Then, based on #3, there are two options...

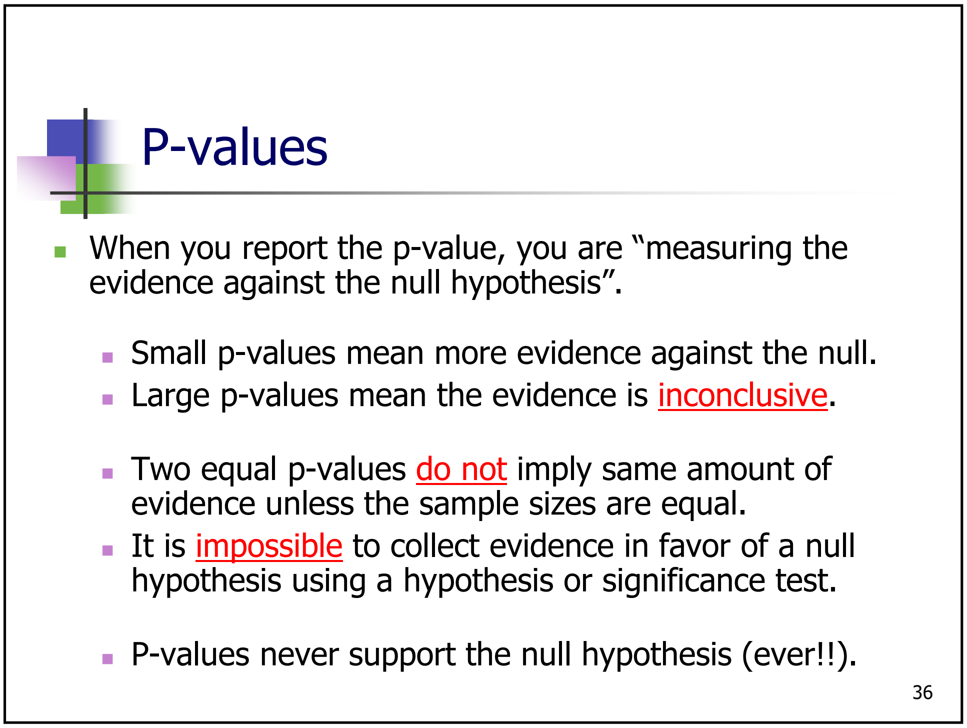
Hypothesis Testing: Decide to reject or accept the null hypothesis.

Significance Testing: Measure the evidence 'against the null hypothesis' and report it.

We use the probability of observing the estimate, or a more extreme estimate, under null hypothesis for this (***p-value***).

35

35

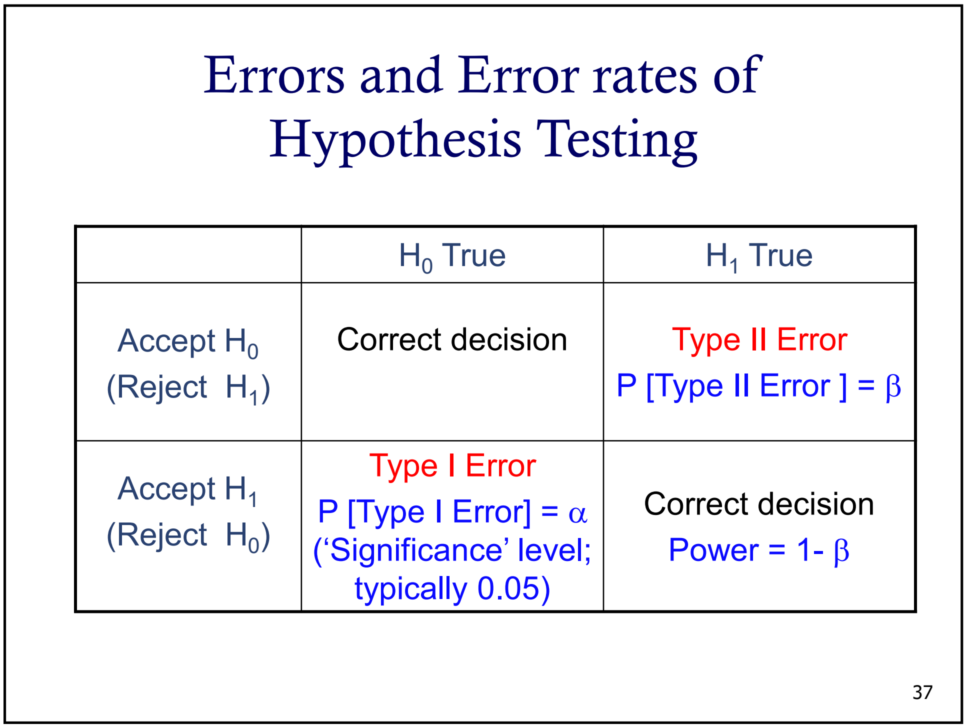


P-values

- When you report the p-value, you are "measuring the evidence against the null hypothesis".
  - Small p-values mean more evidence against the null.
  - Large p-values mean the evidence is inconclusive.
  - Two equal p-values do not imply same amount of evidence unless the sample sizes are equal.
  - It is impossible to collect evidence in favor of a null hypothesis using a hypothesis or significance test.
  - P-values never support the null hypothesis (ever!!).

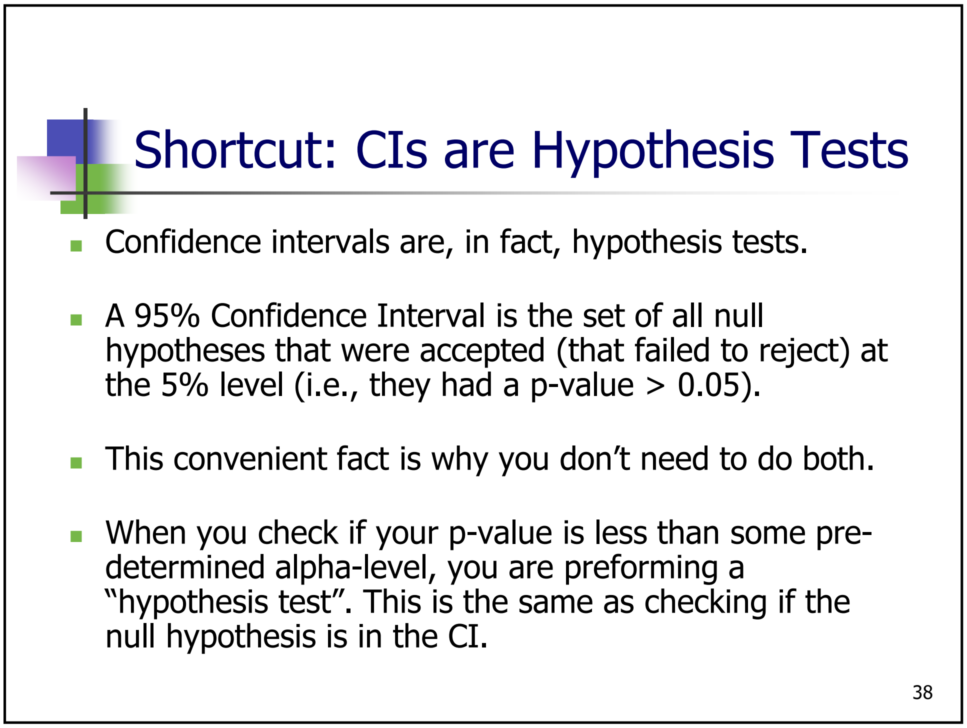
36

36



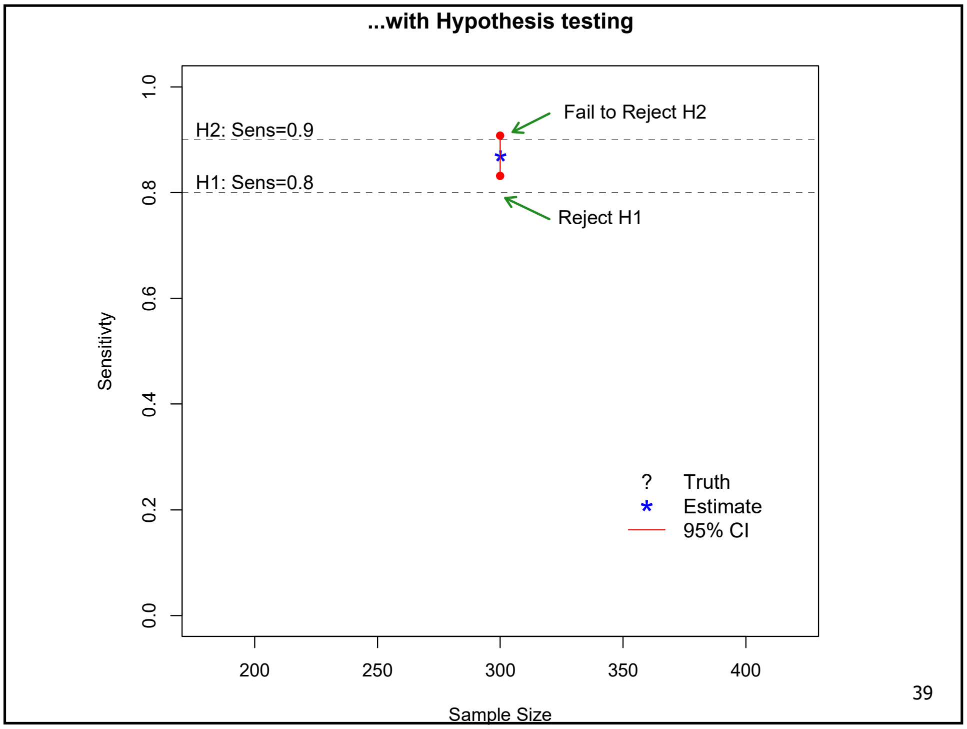
Errors and Error rates of Hypothesis Testing

	H_0 True	H_1 True
Accept H_0 (Reject H_1)	Correct decision	Type II Error P [Type II Error] = β
Accept H_1 (Reject H_0)	Type I Error P [Type I Error] = α ('Significance' level; typically 0.05)	Correct decision Power = 1- β

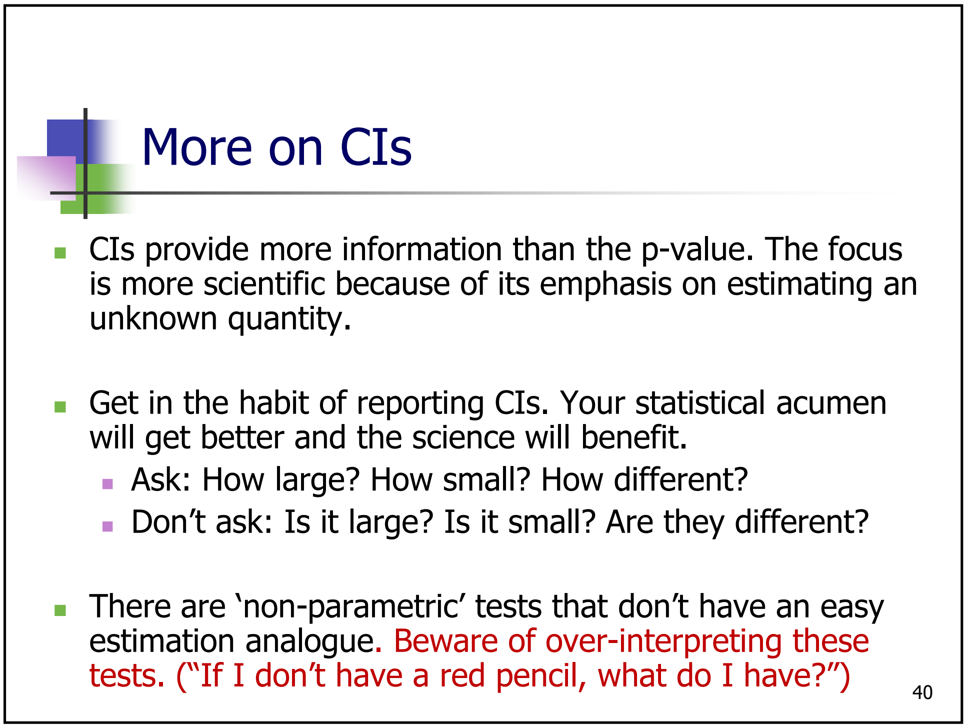


Shortcut: CIs are Hypothesis Tests

- Confidence intervals are, in fact, hypothesis tests.
- A 95% Confidence Interval is the set of all null hypotheses that were accepted (that failed to reject) at the 5% level (i.e., they had a p-value > 0.05).
- This convenient fact is why you don't need to do both.
- When you check if your p-value is less than some pre-determined alpha-level, you are preforming a "hypothesis test". This is the same as checking if the null hypothesis is in the CI.



39

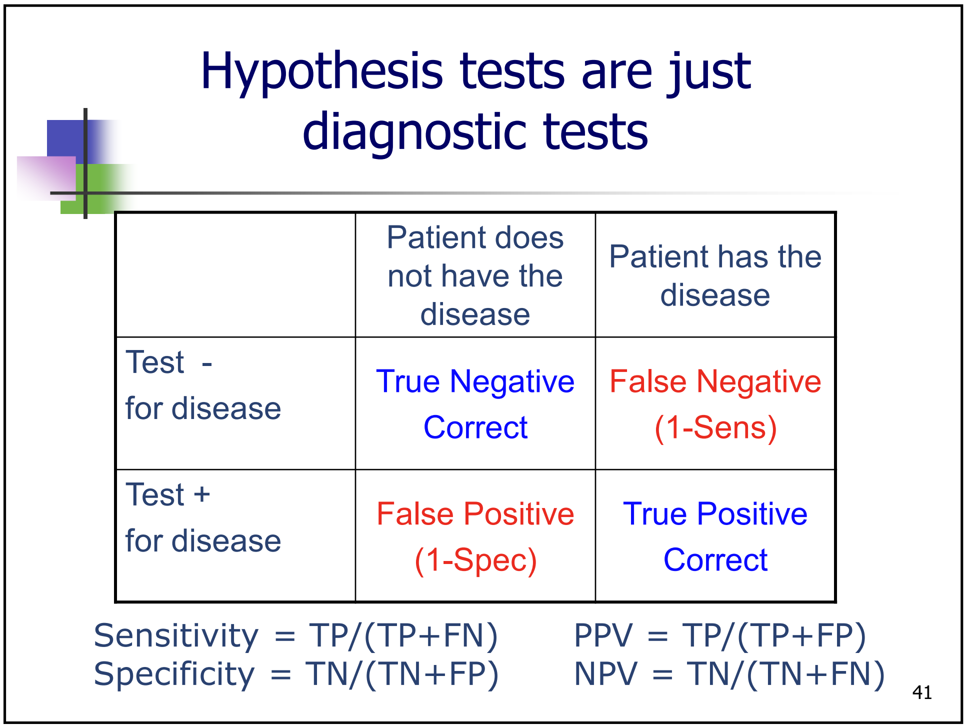


More on CIs

- CIs provide more information than the p-value. The focus is more scientific because of its emphasis on estimating an unknown quantity.
- Get in the habit of reporting CIs. Your statistical acumen will get better and the science will benefit.
  - Ask: How large? How small? How different?
  - Don't ask: Is it large? Is it small? Are they different?
- There are 'non-parametric' tests that don't have an easy estimation analogue. Beware of over-interpreting these tests. ("If I don't have a red pencil, what do I have?")

40

40



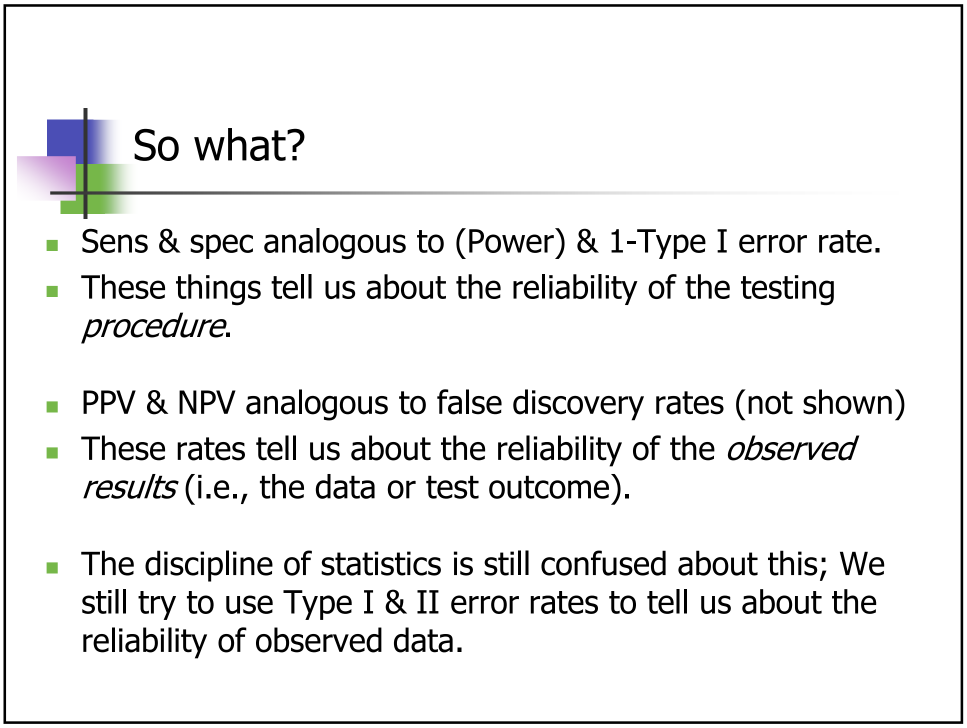
Hypothesis tests are just diagnostic tests

	Patient does not have the disease	Patient has the disease
Test - for disease	True Negative Correct	False Negative (1-Sens)
Test + for disease	False Positive (1-Spec)	True Positive Correct

Sensitivity = TP/(TP+FN) PPV = TP/(TP+FP)
Specificity = TN/(TN+FP) NPV = TN/(TN+FN)

41

41



So what?

- Sens & spec analogous to (Power) & 1-Type I error rate.
- These things tell us about the reliability of the testing *procedure*.

- PPV & NPV analogous to false discovery rates (not shown)
- These rates tell us about the reliability of the *observed results* (i.e., the data or test outcome).

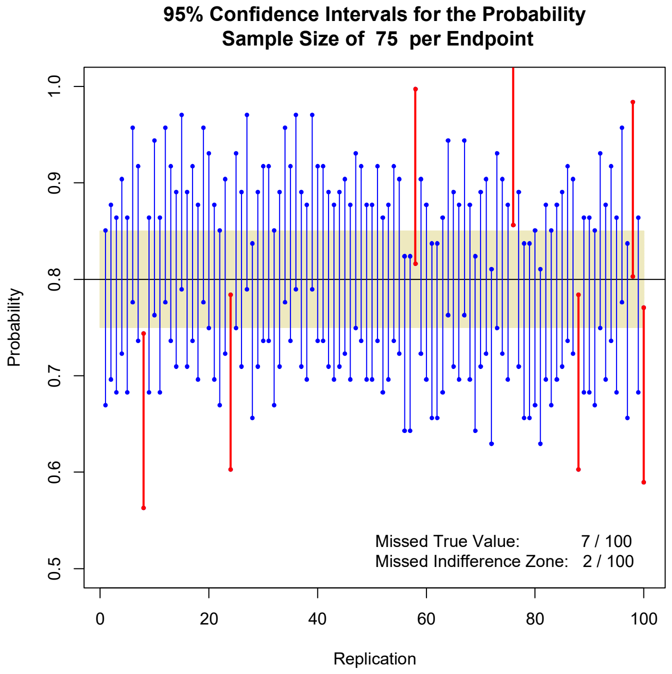
- The discipline of statistics is still confused about this; We still try to use Type I & II error rates to tell us about the reliability of observed data.

42

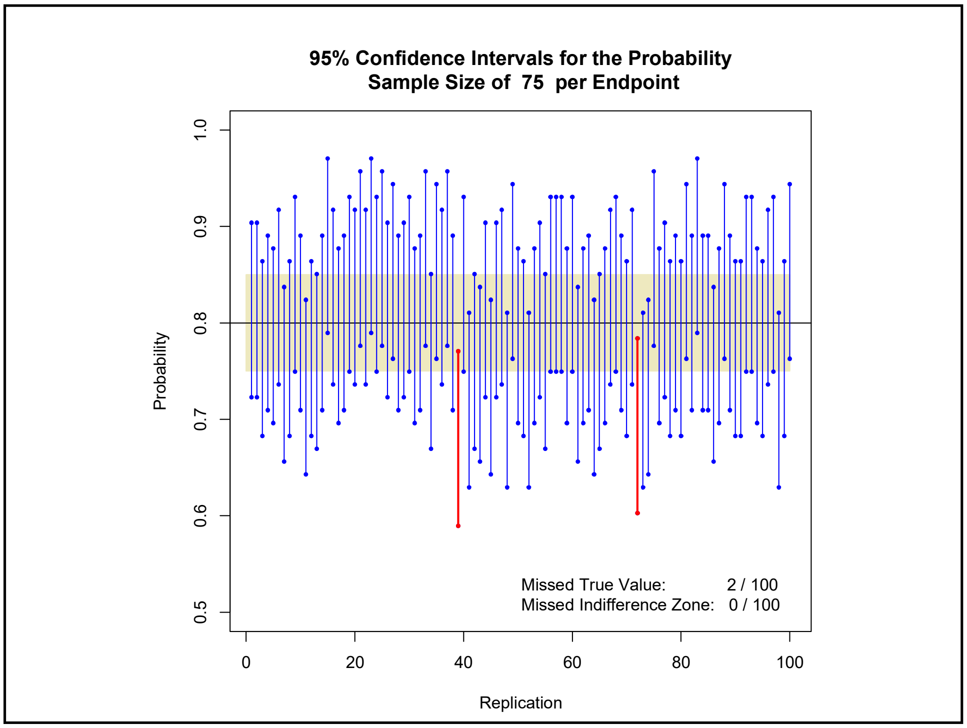
Usefulness of Indifference zones

- Use an indifference zone to represent null effect, practically null effects, & trivial effects.
- Indifference zones often represent clinical or practical equivalence.
- Indifference zones lower Type I Error rates, lower false discovery rates, and have improved statistical properties (but sometimes lower power).
- Indifference zones are the key tool that make equivalence studies and non-inferiority studies work.

43

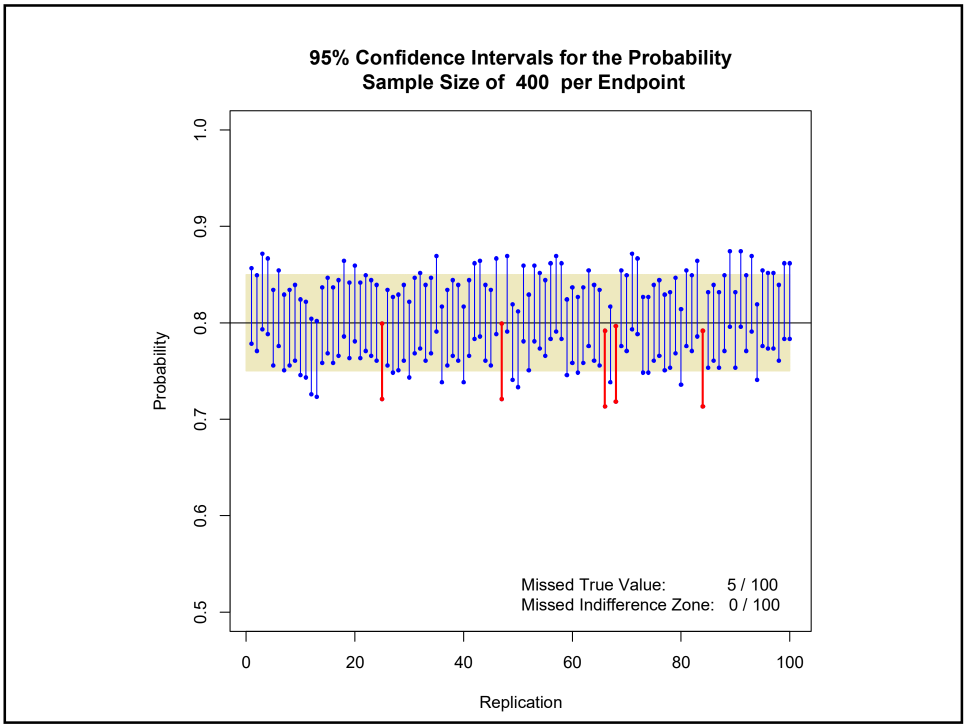


44



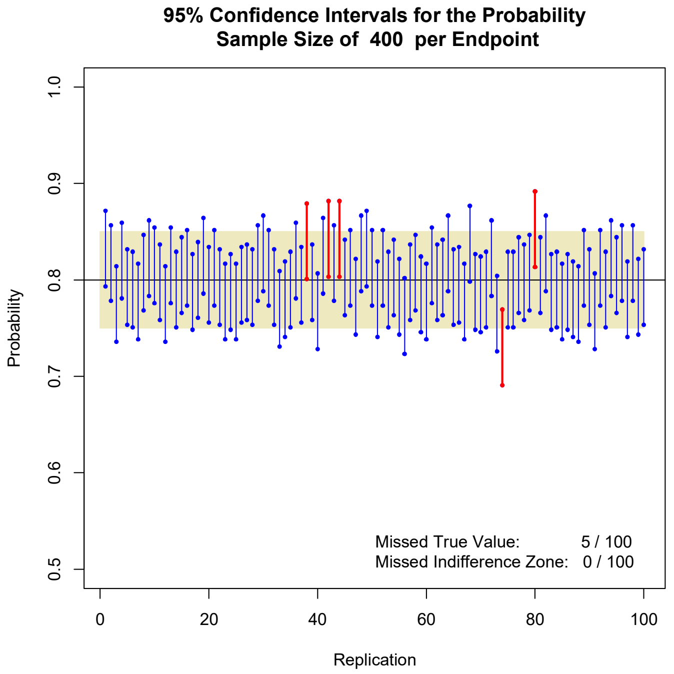
95% Confidence Intervals for the Probability
Sample Size of 75 per Endpoint

45



95% Confidence Intervals for the Probability
Sample Size of 400 per Endpoint

46

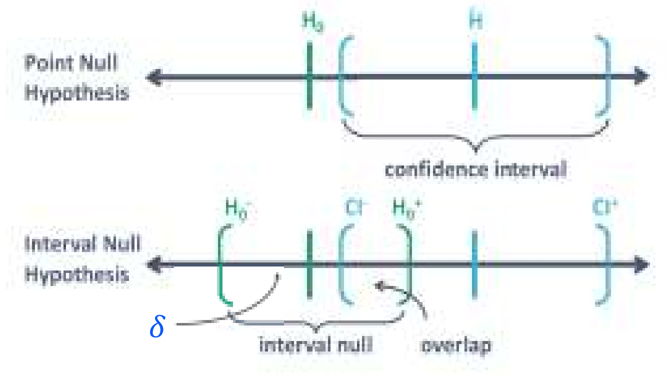

95% Confidence Intervals for the Probability
Sample Size of 400 per Endpoint

Missed True Value: 5 / 100
Missed Indifference Zone: 0 / 100

P-values for indifference zones

- A **second-generation p-value (SGPV)** uses a ‘interval null’ or null zone for inference purposes.
- The SGPV measures the overlap between the confidence interval and the indifference/null zone.
- SGPVs indicate when the data favor the alternative, favor the null, or are inconclusive.
- SGPVs can be used to improve reporting, study planning, equivalence testing, feature selection and more.

Second-generation p-values



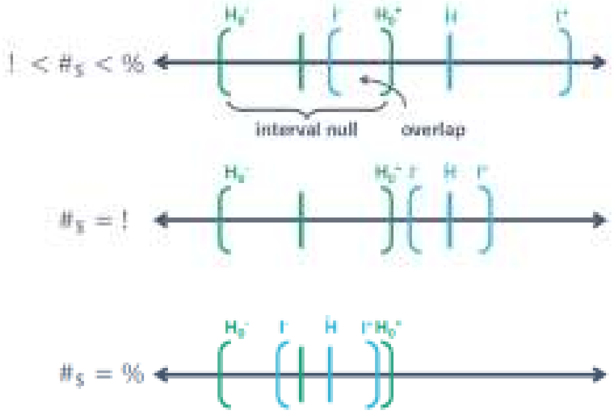
Point null hypothesis H_0 and interval null hypothesis $[\mathrm{H}_0^-, \mathrm{H}_0^+]$
Data-supported hypothesis $\widehat{\mathrm{H}}$ and confidence interval $[\mathrm{CI}^-, \mathrm{CI}^+]$

From Blume et al. PLOS One 2018

P-values with indifference zones

- When the CI does **not** overlap with the indifference zone we have **SPGV=0**. This implies clinically meaningful departures from the null.
- When the CI is completely contained in the indifference zone, we have **SPGV=1**. This implies clinical equivalence.
- When the CI partially overlaps with the indifference zone, we have **0<SGPV<1.** This implies the results are inconclusive.

SGPV Illustration



Works with confidence, credible, and support intervals

Blume et. al. PLOS One 2018

Take Home Messages

- Confidence intervals are versatile and they avoid some of the common pitfalls of statistical testing.
- The ‘art’ in statistics is in translating a scientific question into quantifiable statement that can be tested empirically.
- More on statistical testing: Blume and Peipert. *Journal of the American Association of Gynecologic Laparoscopists* 2003; 10(4): 439-444.
- Second-Generation p-values are a potential solution. See www.statisticalevidence.com (Blume et al *PLOS One* 2018)