Use and Abuse of Inference

- **We have studied the two major forms of statistical inference: confidence intervals and significance tests.**
- **Modern technology (computers, calculators, etc.) has made it easy to perform these methods on data sets.**
- **We must be careful about how we interpret results, however.**
- **We need to understand our data and how they were collected, use the right method for our research question, and understand the implications of our results.**

Be Aware of the Sampling Design

- **What type of sample did the data come from?**
- **If it was a Simple Random Sample (SRS), the inference should be valid.**
- **If your data come from a more complicated sampling design (stratified, cluster, etc.) the methods we learned will not be exactly correct.**
- **There are other methods that are designed to work with these more complicated sampling designs.**

Be Aware of the Sampling Design (continued)

- **If there is nonresponse or dropout, this will affect the correctness of the inference unless specialized methods are used to account for this.**

- **If the data were collected in a haphazard or possibly biased way (convenience sample, volunteer sample), then no methods of inference will be exactly correct.**

Clicker Quiz 1

The confidence interval formulas from Chapter 21 are appropriate for data that come from which type of sample?

A. Convenience sample

B. Cluster sample

C. Stratified random sample

D. Simple random sample

Sampling Design (Continued)

- **Sometimes data that are not really from a SRS can be treated as if they were from a SRS.**
- **Example: Psychology experiment of visual perception.**
- **Can a class of psychology students be treated as a SRS of people with ordinary vision?**
- **Another example: Sociology study about attitudes toward poor people and antipoverty programs.**
- **Can a class of sociology students be treated as a SRS of Americans?**
- **Can a class of sociology students be treated as a SRS of college students?**

Cautions about Confidence Intervals

- **We know that a confidence interval we have obtained is not guaranteed to contain the parameter of interest.**
- **When we find a 95% interval, there's a 5% chance that the sample we obtain is "weird" enough that the interval doesn't contain the parameter of interest.**
- **Is this risk too high? Could use a 98% or 99% interval. But** . . .

Cautions about Confidence Intervals

- **The 98% interval is wider, and less precise/informative about the true value of the parameter.**
- **To get high confidence and a narrow interval, we need to take a large sample.**
- **For a confidence interval about p: To cut the interval width in half, we would need to take four times as many observations in our sample.**

Clicker Quiz 2

A 95% confidence interval about p based on n = 50 observations is (0.57, 0.83), that is, has width 0.26. To get a 95% confidence interval with width 0.13, about how many observations would we need?

A. 300

B. 50

C. 200

D. 100

Clicker Quiz 3

A 95% confidence interval about p based on n = 50 observations is (0.57, 0.83). What is a possible 98% confidence interval, based on the same data set?

A. (0.55, 0.85)

B. (0.59, 0.81)

C. (0.53, 0.83)

D. (0.57, 0.83)

Be Aware of the Requirements of Methods

- **Most of the types of inference we have seen require large samples.**
- **If you have small samples, don't use the "large-sample" formulas!**
- **There are other methods that are designed for small samples, skewed data, data with outliers, etc.**
- **When reading reports about data analyses, ask yourself: Have the authors used the right methods for their sample?**
- **What type of sample did their data come from? Are they reporting confidence intervals and/or P-values?**

Cautions about Significance Tests

- **What do we really want to know when doing a significance test?**
- **Some would say we want to know, "What is the probability that H_0 is true?"**
- **The P-value does NOT tell us this.**
- **It doesn't tell us the probability that H_0 is true, given the data.**
- **It tells us the probability of seeing data like we saw, given that H_0 is true.**

Cautions about Significance Tests (continued)

- **Some would argue that the classical P-value approach is backward, overly complicated reasoning.**
- **Some scientists discourage classical significance tests for this reason.**
- **One scientific journal (*Basic and Applied Social Psychology*) recently banned significance testing from the articles it publishes!**
- **Alternative approaches (Bayesian inference) try to find the probability that H_0 is true, given the data.**

Effect Sizes in Significance Tests

- **Significance tests are designed to detect some sort of “effect” or “difference”.**
- **Maybe it’s the effect of a new drug treatment on patients.**
- **Maybe it’s the difference between 0.5 and the probability a coin comes up “heads”.**
- **If the “effect” or “difference” is large, the significance test will usually detect it, even when the sample size is not huge.**
- **What if the “effect” or “difference” is present, but very small?**

Effect Sizes in Significance Tests (continued)

- **Suppose a new treatment increases mean survival time by 2 days, on average, compared to the standard treatment.**
- **That's a difference, but is it practically important?**
- **Suppose a coin has probability 0.502 of coming up "heads". Is that imbalance practically important?**
- **Issue: If the "effect" or "difference" is very small, the significance test will almost certainly detect it, *if the sample size is huge!***
- **If the "effect" is small but important, the significance test will quite possibly NOT detect it, *if the sample size is small*.**

Statistical Significance and Practical Importance

- **If a coin has probability 0.502 of coming up "heads", and we toss it 1,000,000 times, the resulting data and P-value will probably lead us to reject $H_0 : \boldsymbol{p}$ = 0.5 in favor of $H_0 : \boldsymbol{p} \neq$ 0.5.**
- **Is that what we want?**
- **Lesson: A result can be statistically significant without being practically important.**
- **Other Lesson: With enough data, you can reject just about any null hypothesis!**

Statistical Significance and Practical Importance (continued)

- **If a coin has probability 0.55 of coming up "heads", and we toss it 15 times, the resulting data and P-value may not lead us to reject $H_0 : \boldsymbol{p}$ = 0.5 in favor of $H_0 : \boldsymbol{p} \neq$ 0.5.**
- **Is that what we want?**
- **Lesson: Lack of statistical significance does not mean there is no effect; we just haven't found strong enough evidence in our sample.**
- **Other Lesson: Small samples often miss important effects that are really present in the population.**

Statistical Significance and Practical Importance (continued)

- **With small data sets, you have to pay careful attention to using the right method.**
- **A confidence interval may be a better method of inference about the probability.**
- **It will explicitly reveal the “uncertainty” about the parameter.**

Clicker Quiz 4

We test the balance of a coin ($H_0 : \boldsymbol{p} = 0.5$ against $H_a : \boldsymbol{p} \neq 0.5$) based on 16 heads out of 23 flips, using a significance level of $\boldsymbol{\alpha}$ = 0.05. Our p-value turns out to be 0.061. What is a reasonable conclusion?

A. The coin is certainly balanced.

B. The coin is certainly unbalanced.

C. The coin may be unbalanced, but we don't have strong enough evidence to conclude that for sure.

D. The coin is marginally balanced.

P-value alone is not Enough

- **In the previous clicker example, $\hat{p}$ = 16/23 = 0.696.**
- **What if we conducted the same test, and got the same $\hat{p}$ = 0.696, but we had 230 tosses?**
- **The P-value then would be 0.000000003.**
- **A $\hat{p}$ that was not convincing evidence of imbalance in 23 tosses is VERY convincing evidence of imbalance in 230 tosses.**
- **The P-value depends on BOTH the true parameter value AND the sample size.**
- **When reporting a P-value, we should also report the sample size and the value of the statistic that estimates the parameter.**

P-value alone is not Enough (continued)

- **In Count Buffon's coin-tossing experiment he got 2048 heads in 4040 tosses, so $\hat{p}$ = 2048/4040 = 0.507.**
- **The P-value would be 0.37: not enough evidence to say it is imbalanced.**
- **What if he conducted the same test, and got the same $\hat{p}$ = 0.507, but he had 40,400 tosses?**
- **The P-value then would be 0.0053.**
- **A $\hat{p}$ that was not convincing evidence of imbalance in 4040 tosses is VERY convincing evidence of imbalance in 40,400 tosses.**
- **So is the coin imbalanced or not?**

Confidence Intervals can be Better

- **It's usually better to report confidence intervals rather than P-values.**
- **The 95% confidence interval for p in Buffon's experiment (4040 tosses) is (0.492, 0.522).**
- **The 95% confidence interval for p in Buffon's experiment (40400 tosses) is (0.502, 0.512).**
- **These intervals make it clear what we suspect about the probability of heads, given the sample data we got.**
- **They are easier to interpret than just the P-values in each case (0.37 and 0.0053).**

Why a 5% significance level?

- **Many scientists hold on to "P-value $<$ 0.05" as a gold standard for whether an effect is "true" or not.**
- **There's no magic border at 0.05; it's just an arbitrary number.**
- **It was picked in an article long ago ("one chance out of twenty") for no really good reason, and people have held onto it ever since.**
- **It was convenient in the days before fast computers, when people had to rely on tables to do calculations.**
- **A P-value of 0.049 and another of 0.051 present essentially equally strong evidence that H_0 is false.**

Searching for Significance

- **Suppose you are looking for a “significant association” between success and a company and some background variable.**
- **You gather a sample of past employees and observe dozens of variables on them.**
- **For each variable, you test whether that variable is significantly associated with success.**
- **For most of the background variables, the results aren’t significant at α = 0.05, but a couple are.**
- **Should you conclude there is a true association between success and those two variables?**

Searching for Significance

- **Remember, even if EVERY null hypothesis (of no association) were true, 5% of these tests would produce significant results just by chance.**
- **Beware of doing many simultaneous significance tests! (See jelly bean comic strip)**

Searching for Significance: Publication Bias

- **Imagine: 20 scientific teams are gathering data and performing significance tests to answer the same research question: Does "substance X" reduce cancer rates in lab mice?**
- **For 19 of the teams, their results are not significant at α = 0.05, so they don't bother to write up and publish the work.**
- **For the other team, their results ARE significant at α = 0.05, so they write up and publish the work in an important scientific journal.**
- **A news item then says:** ***Scientists discover that "substance X" reduces cancer rates!***
- **Do you believe it?**
- **What could be done about this?**

Searching for Significance: Unrevealed Replicate Studies

- **Imagine: A scientific team gathers data and performs a significance test to answer the research question: Does "substance X" reduce cancer rates in lab mice?**

- **They try 19 samples of mice, but don't get significant results based on any of these samples.**

- **For the 20th sample of mice, their results ARE significant at α = 0.05, so they do write up the work based on the 20th sample and publish the work in an important scientific journal.**

- **The article doesn't mention that this was their 20th try at the experiment.**

- **A news item then says:** ***Scientists discover that "substance X" reduces cancer rates!***
- **Do you believe it?**
- **What could be done about this?**

Clicker Quiz 5

What is a downside to the problems of publication bias and multiple tests?

A. Most data are not from simple random samples.

B. Many published research findings of significance are in fact false.

C. Many published research findings do not include significance tests.

D. Research findings take too long to publish.