Math 1000
Mathematical Literacy in Today’s World

Lecture 13
Winter 2010
Midterm Exam

- Covers only Chapters 5-7
- Similar to Quizzes
- Given in this room at this time Wednesday March 10
Start of Chapter 8

- Lecture next Monday will start Chapter 8
- Review for the Midterm in your quiz sessions Tuesday
- The first worksheet for Chapter 8 will be on Thursday, March 11
- There will be a quiz on Chapter 8 after we’re done
Observational Studies

- The kind of statistical study we have looked at so far
- Measures variables of interest but does not attempt to influence possible responses
- The purpose is to describe relationships
Experiments

- Here some change (a treatment) is applied to the individuals
- Then changes in response are measured
- The purpose is to determine if there will be a measurable response to the treatment
Randomized Comparative Experiments

- A group of individuals is chosen for an experiment
- They are divided into two (or more) groups at random
- Treatment is applied to one (or more) group(s), but not to the control group
- Results are analyzed to see if there is a statistically significant difference
“Double-blind, placebo-controlled”

- It is often valuable to keep both the subjects of the experiment and the experimenters themselves from knowing which group is the control group and which is the treatment group.
- By making the two groups look like they are doing the same sort of activity.
Experiment versus real life

- An experiment is a test for cause-and-effect in miniature, just among the subjects selected to be in the test groups.
- There remains the question: If the test works in the experiment, will it work in the general population?
Example

- During drug trials, patients may be given better medical care than they would normally receive.
- The results may be that, while the drug works when getting the better medical care, it could be much less effective among general patients.
Example

- Center brake lights (in the rear window) were added to cars after an experimental study indicated a large reduction in rear-end collisions

- Now that drivers are used to seeing these lights, their effectiveness has severely decreased
Experiment versus real life

- In either case the experimental results had less application to the general population
- This is called “lack of realism”
- Lack of realism is a non-statistical concern for the experimenter
Observations still have value

- For the effects of smoking, we would not want to do an experiment
- Because that would involve forcing our treatment group to smoke
- And our experience suggests that the effect is long-term, so the experiment would have to go on a long time
Observations over time

- An observational study which follows individuals over time to find relationships between variables (like smoking versus health) is called a prospective study.

- Prospective studies are comparative, but not experimental.
Prospective studies

- To have statistical value, prospective studies need to involve a large number of individuals
- But they still might have confounding variables
Imaginary example

- Suppose that some people have a genetic background which has two effects: one causes enjoyment from smoking; the other causes a tendency to have lung cancer.
- Then prospective studies would show a strong correlation between smoking and lung cancer.
- BUT not a cause!
Statistical Inference

- From either an observational or an experimental study, we can infer some information about the general population from the results of the study.
- For example, if a poll indicates that 20% of the sample favors a certain candidate, we can infer that 20% of the general population favors that candidate.
Parameter of the population

- A parameter is a number that describes the population (like percentage favoring a certain candidate)
- Usually the value of the parameter is unknown and we can only infer its value from statistical samples
Statistics from a sample

- A **statistic** is a number that describes a sample.
- Once we take a sample, we compute the value of some statistics.
- We often use such statistics to estimate unknown parameters.
Why the distinction?

- We need to know how useful are our estimates of unknown parameters.
- For that we need to know how likely it is that statistics will suggest the **wrong** estimates for the parameters.
- We hope that it is **not** likely; in fact that it is rare that statistics give wrong estimates.
Our estimates for population parameters are statistically significant if the chances that we chose an unusual sample giving a quite different estimate is less than 5%.

But how do we tell what those chances are?
The idea of repeated samples

To determine how useful are our estimates, we imagine repeating the sample more than once.

Then we see how much the statistics vary.

Lots of variation implies not good estimates.
A national random sample of 2500 adults was asked whether they agreed that shopping for clothes was frustrating.

1650 of them agreed, giving a proportion of \( \frac{1650}{2500} = 0.66 = 66\% \) agreeing.

This proportion \( \hat{p} = 0.66 \) is a statistic from the sample; we can use it to estimate the parameter \( p \) for the entire population.
Repeating the sample

- If the poll was done again, probably a different number of adults would agree, resulting in a different statistic \( \hat{p} \).
- But how different might it be?
- Imagine repeating the sample many times and see how much it varies
- (Make a histogram of the various results)
Too expensive

- Generally it is beyond reasonable resources to be able to make that many repeated samples
- Could we simulate this?
- *(I.e., could a computer draw random samples from a population with unknown parameter to see how much the statistics vary?)*
Sampling distribution

- This is the distribution of the computed statistics from a large number of samples (ideally all possible ones)
- For purposes of estimating the parameter, we hope the sampling distribution shows more values near the actual value of the parameter; and we hope it doesn’t have much spread
Example from text

- For the example about frustration in buying clothes, we could repeat samples of size $n = 100$ and get statistics $\hat{p}$ from each sample.

- These answers would vary, but around some middle value (which would be our estimate for the parameter $p$).
Example from text

- Or we could repeat samples of size $n = 2500$ and get statistics $\hat{p}$ for each
- Answers would still vary, but presumably with much less spread around the middle value
- So that our estimate is more accurate in this case
Histogram for $p = 0.6$ with sample size $n = 100$
Histogram like normal distribution

- Looks like mean is 0.60, as we think
- And standard deviation about 0.05
Histogram for $p = 0.6$ with sample size $n = 2500$
Magnified

1000 samples each of 2500
Histogram like normal distribution

- Looks like mean is still 0.60, as we think
- And standard deviation now about 0.01
For a sampling proportion

- Choose a SRS of size \( n \) from a large population that has an unknown parameter \( p \) giving the proportion (fraction) of the population answering yes to some question (a so-called “success”)
- Find the number \( k \) in the sample answering yes and compute \( \hat{p} = \frac{k}{n} \) (the statistic from the sample)
Then the sampling distribution is approximately normal

- Then for large sampling sizes $n$, the sampling distribution is approximately a normal distribution
- With mean $p$ (center of distribution)
- And standard deviation $\sqrt{\frac{p(1-p)}{n}}$ (spread of distribution)
Computing standard deviation for samples

- For $p = 0.6$, $n = 100$, we get standard deviation $= \sqrt{0.6(1-0.6)/100} = \sqrt{0.6 \cdot 0.4/100} = \sqrt{0.24/100} = \sqrt{0.0024} = 0.049$

- For $p = 0.6$, $n = 2500$, standard deviation $= \sqrt{0.6(1-0.6)/2500} = \sqrt{0.6 \cdot 0.4/2500} = 0.0098$
For $n = 100$, standard deviation $= 0.049$
For \( n = 2500 \), standard deviation = 0.0098
Trusting a sample

- The normal distribution shows that 68% of sample results with $n = 2500$ will have results $p$-hat between $0.6 - 0.0098 = 0.5902$ and $0.6 + 0.0098 = 0.6098$

- So 68% of the sample results will be close to the actual value of the parameter $p$
95% confidence

- And, using two standard deviations, 95% of the samples will be within $2(0.0098) = 0.0196$ of the mean $0.6$.
- So 95% of the samples with $n = 2500$ will have $\hat{p}$-hat between $0.6 - 0.0196 = 0.5804$ and $0.6 + 0.0196 = 0.6196$.
- Almost all samples give useful results.
95% confidence interval

- The two numbers 0.5804 and 0.6196, between which 95% of the sample results will be, form an interval.
- It may be written [ 0.5804, 0.6196 ]
- Or as 0.60196
So in general

- To estimate a proportion parameter $p$ (indicating the proportion or fraction of the population agreeing with a given statement):

- Pick a SRS of $n$ individuals, and find the statistic $p$-hat (the proportion for the sample)
Continuing the procedure

- Use the p-hat to estimate the standard deviation \( s \) for the sample p-hats by computing \( s = \sqrt{\frac{(p\text{-hat})(1-(p\text{-hat}))}{n}} \)

- Then use \((p\text{-hat} - 2s)\) to \((p\text{-hat} + 2s)\) as a 95% confidence interval for the unknown parameter \( p \)
Then we can say:

Based on our random sample of n individuals, we can be 95% certain that the actual value of the proportion parameter p lies in the computed confidence interval.
But we cannot say:

- We cannot say that we are 95% certain that the parameter is the middle value in the interval (the p-hat)
- Nor can we say that we can be certain that the parameter lies in the interval
Example

- Suppose that a SRS of $n = 100$ people shows that $10$ favor Ron Paul’s candidacy.
- Then $p$-hat = $10/100 = 0.1$ (10%)
- Standard deviation $s$ = square root of $[0.1(0.9)/100] = $ square root of $0.0009 = 0.03$
- So interval is $0.1-2(0.03) = 0.04$ to $0.1+2(0.03) = 0.16$
- Thus we can be 95% certain that the actual percentage supporting Ron Paul is between 4% and 16% (plus or minus 6 points)
Example, continued

- Thus we can be 95% certain that the actual percentage supporting Ron Paul is 10%, plus or minus 6 percentage points (margin of error)

- But the other 5% of the time our sample was from unusual people whose proportion of support is much different from the rest of the population
If we want more accuracy

- To decrease the size of the 95% confidence interval, increase $n$ to reduce the size of the standard deviation (larger sample size)

- To increase your confidence level, use more than two standard deviations from the mean (but that increases the size of the confidence interval)
If we want more accuracy

- To decrease the size of the 95% confidence interval, increase \( n \) (larger sample size)
- For the Ron Paul example, increasing \( n \) to 1000 would give standard deviation \( s = \sqrt{\frac{0.1 \times 0.9}{1000}} = \sqrt{0.00009} = 0.0095 \)
- 95% confidence interval would be 0.1 ± 0.019 or 8.1% to 11.9%
If we want more confidence

- To increase your confidence level, use more than two standard deviations from the mean.
- For the original Ron Paul example, say use 3s.
- Then 99.7% confidence interval would be $0.13 \times 0.03 = 0.109$ or 1% to 19%.
- Error is bigger, but we are more confident that his actual support is in that interval.
Remember

- 68% of the data lies in an interval one standard deviation below the mean to one standard deviation above the mean
- 95% of the data lies in an interval two standard deviations below the mean to two standard deviations above the mean
- 99.7% of the data lies in an interval three standard deviations below the mean to three standard deviations above the mean
The End for Today