PS0700
Basic Statistical Methods: Statistical Inference

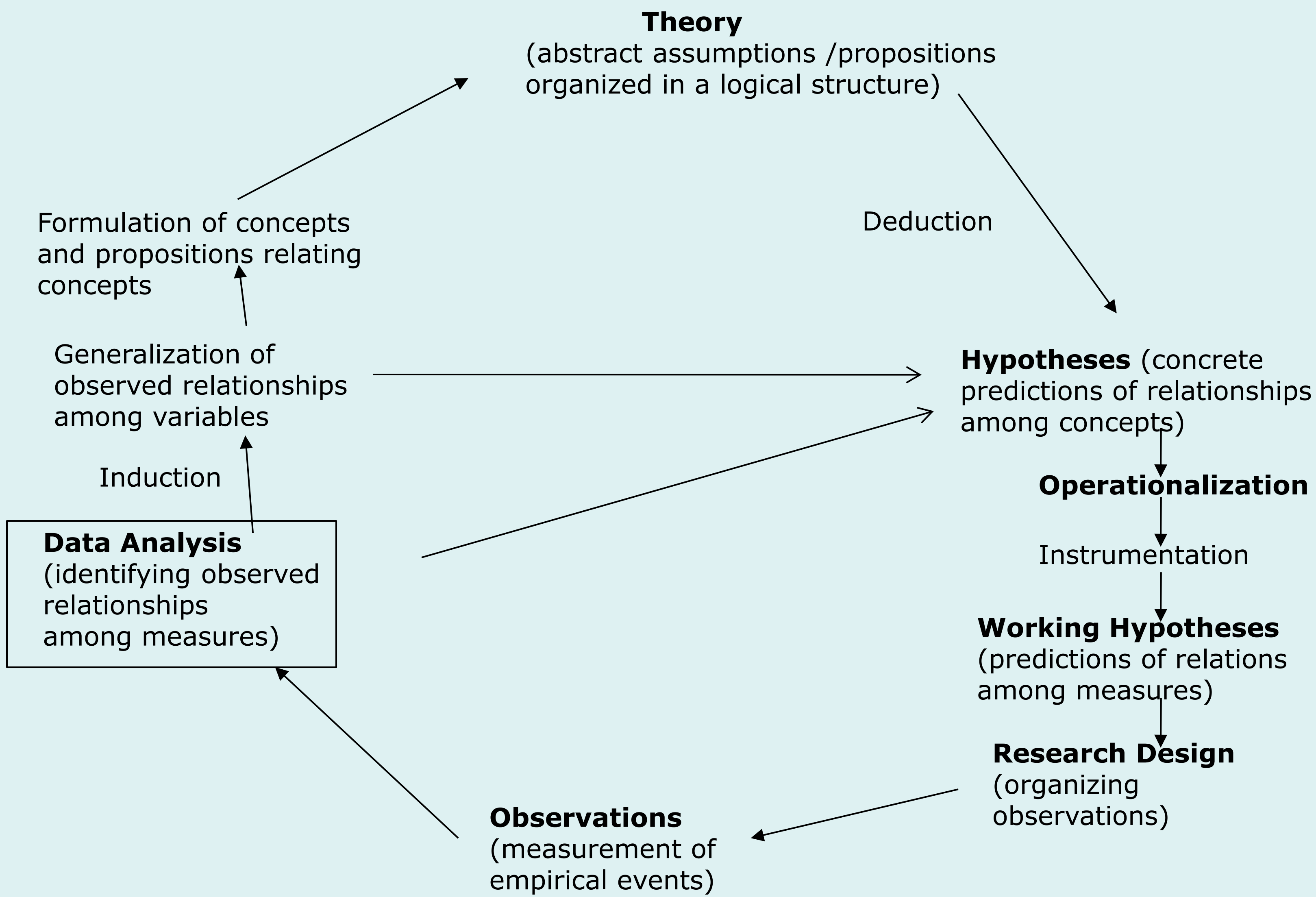
Political Science Research Methods

Professor Steven Finkel

Fall Semester 2022

Week 9b-10a

A Model of the Research Process



Inferential Statistics

- The process of *inferring* properties of the population of interest from properties of a sample of (randomly selected) observations from that population
  - Given sample mean $\bar{Y}$, what is the likely value of the population mean μ? And what is the likely range of values of μ, i.e. what is the *confidence interval* around our best guess of μ?

- Will extend these procedures to *test statistical hypotheses*. Using the ideas we develop today, we will begin to make inferences about the relationship between variables *in the population* based on the relationship between the variables that we observe *in our sample*

Sampling Distributions

- Statistical Inference depends on understanding the concept of a *sampling distribution* of a statistic

- Definition: A theoretical frequency or probability distribution of a statistic obtained through repeated (infinite) random samples from a population of a given size

- Imagine taking an infinite number of random samples of a certain size from a population and plotting the histogram of the sample means (or the sample proportions on "1" of a 0/1 variable)

Example

- Assume a population has an average salary of \$5000/month, with a S.D. of 2000. Take a random sample of 100 individuals, plot the mean salary in this sample (say \$4000). Then take another random sample of 100 individuals, plot the mean salary in this sample (say \$6500). Take another, then another, then another, then infinite number of random samples of size 100. What would the histogram of all these sample means look like?

The Central Limit Theorem

- If repeated random samples of size N are drawn from a distribution with mean μ and standard deviation σ, then as N → ∞, the sampling distribution of sample means ($\bar{Y}$) will be normal, with mean μ and standard deviation $\frac{\sigma}{\sqrt{N}}$

- This is a remarkable theorem!!! It says that, no matter what the shape of the population, plotting an infinite number of random samples will produce a *normal* sampling distribution of a given statistic, provided that the sample size is large enough.
- See any number of web sites such as:
  - https://onlinestatbook.com/stat_sim/sampling_dist/index.html

- This means we can make use of the properties of the normal curve to conduct all kinds of statistical inferences about means (and proportions)!!!

Our Example: μ = 5000, σ=2000

- N=100
  - Sampling distribution of means will be normal, centered around 5000, with standard deviation of $2000/\sqrt{100}$, or 200.
  - We call the "standard deviation of a sampling distribution" the **STANDARD ERROR**, in this case the **STANDARD ERROR OF** $\bar{Y}$
  - We represent this with the symbol: $\sigma_{\bar{Y}}$

Following the Properties of Normal Distributions:

- 68% of all sample means $\bar{Y}$ will be between 5000 ± 200, or between 4,800 and 5,200
- 95% of all sample means will be between 5000 ± 1.96*200, or 5000 ± 392, or between 4608 and 5392
- 99% of all sample means will be between 5000 ± 2.58*200, or 5000 ± 516, or between 4484 and 5516
- **Therefore we know the exact probability of obtaining sample means with *any and all* particular values from that population with sample size 100**
- We can also calculate the Z score for any value on this *sampling distribution* as $z = \dfrac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$
- Exercise: What proportion of samples of size 100 from this population will have means between 4,900 and 5,300?

$$Z(4900) = \frac{4900 - 5000}{200} = \frac{-100}{200} = -.5$$

$$Z(5300) = \frac{5300 - 5000}{200} = \frac{300}{200} = 1.5$$

- 31% of all cases in a normal distribution are less then z=-.5, so 19% of sample means will be between z=-.5 and μ of 5000
- 6.6% of all cases in a normal distribution are greater than z=1.5, so 43.4% of sample means will be between z=1.5 and μ of 5000
- **So: 62.4% of all samples of size 100 from this population will have means between 4900 and 5300**

- Increasing the sample size will result in smaller standard errors, so even more of the samples from any population will lie closer to the population mean

N=100 $\sigma_{\bar{Y}} = \frac{2000}{\sqrt{100}} = 200$ 95% of samples $\mu \pm 392$

N=625 $\sigma_{\bar{Y}} = \frac{2000}{\sqrt{625}} = 80$ 95% of samples $\mu \pm 156.8$

$\mathrm{N} = 900$ $\sigma_{\bar{Y}} = \frac{2000}{\sqrt{900}} = 66.7$ 95% of samples $\mu \pm 130.7$

$\mathrm{N} = 1600$ $\sigma_{\bar{Y}} = \frac{2000}{\sqrt{1600}} = 50$ 95% of samples $\mu \pm 98$

Inferences to Unknown Population Parameters

- Usually we do not know μ or σ, in fact estimating μ from our sample mean $\bar{Y}$ is one of the main purposes for doing our research in the first place!
- So how do we use the information we've learned so far to make inferences from our sample mean $\bar{Y}$ to the unknown population parameter μ?
- First of all, we know that our best guess of μ is going to be what? That's right, $\bar{Y}$! Why? Because from the CLT we know that the sampling distribution of all $\bar{Y}$ is centered around μ – therefore the value of μ that had the highest likelihood of producing our sample mean $\bar{Y}$ is if μ= $\bar{Y}$

Confidence Intervals

- But we also know that there is a lot of uncertainty around that best guess. For example, our sample mean had a 95% chance of being in the interval,

$$\mu \pm 1.96 * \frac{\sigma}{\sqrt{N}}$$

or 1.96 STANDARD ERRORS away from μ

- If we knew σ, we could calculate this interval directly. But we only know $\hat{\sigma}$ (the standard deviation in our sample).
- (Remember that $\hat{\sigma} = \sqrt{\frac{\Sigma(Y_i - \bar{Y})^2}{N-1}}$)
- So we use our sample standard deviation to construct an ***estimated standard error*** $\frac{\hat{\sigma}}{\sqrt{N}}$

- THEREFORE: IF WE CONSTRUCT AN INTERVAL AROUND $\bar{Y}$ EQUAL TO 1.96* $\frac{\hat{\sigma}}{\sqrt{N}}$ IN EITHER DIRECTION, WE WILL BE 95% CONFIDENT THAT THIS INTERVAL WILL ENCOMPASS THE POPULATION MEAN μ, WHATEVER IT IS!!!!!

- Another way to look at it: If μ was ***not*** in that interval, it would have been ***very*** unlikely to have observed the value of $\bar{Y}$ that we did observe

- So: the 95% confidence interval for estimating a population mean from a sample mean:

$$\bar{Y} \pm 1.96 * \sqrt{\frac{\hat{\sigma}}{N}}$$

Example of Calculating a Confidence Interval

- We observe the hours per week spent studying of 900 randomly selected U.S. undergraduate students, and obtain an average of 40, with a standard deviation of 30. What is the 95% confidence interval for the population mean of hours studying per week for all U.S. undergraduates?
- Solution:

$$\bar{Y} = 40 \quad \hat{\sigma} = 30 \quad N = 900$$

Standard Error= $\dfrac{\hat{\sigma}}{\sqrt{N}}$ = 30/30=1

So 95% confidence interval for population μ is:

40 $\pm$ 1.96*1

38.04 --- 41.96

We are 95% confident that if we had observed all U.S. undergrad students, the average number of hours per week spent studying would be between 38.04 and 41.96

Comments on Confidence Intervals

- We will be wrong 5% of the time!!!! Sometimes you just get unlucky with an unusual sample – that is the nature of a probability sample!
- If you want higher degree of confidence, you can construct the 99% confidence interval as:

  $\bar{Y} \pm 2.58 * \frac{\hat{\sigma}}{\sqrt{N}}$ (This makes a wider interval!)

- And if you want a smaller interval, with *less* confidence that μ is really in that interval, you can construct the 90% confidence interval as:

  $\bar{Y} \pm 1.65 * \frac{\hat{\sigma}}{\sqrt{N}}$

- Remember: increasing sample size is the best way to have more confidence in your sample estimates!! If N=2500 in our example, the standard error would be .6, meaning that the 95% confidence interval would be 40 ± 1.96*.6, or between 38.8 and 41.2 (a smaller interval)

- And if N=3600 in our sample, the standard error would be .5 (because 30/60=.5), and the 95% confidence interval would be 40±1.96*.5 or between 39.02 and 40.98

- Also illustrates the **diminishing marginal benefit of larger sample sizes**

Confidence Intervals for Proportions

- This same procedure can be applied for estimating confidence intervals of ***proportions*** – e.g., proportion of voters favoring candidate X in a national survey with probability sampling
- 95% confidence interval for a sample proportion (p) is:

$p \pm 1.96 * \sqrt{\frac{P*(1-P)}{N}}$ with maximum value at P=.5

- With N=1000, 95% confidence interval is:
- $p = \pm 1.96 * \sqrt{\frac{.5(.5)}{1000}} = \pm .031$
- With N=1500, 95% confidence inverval is:
- $p = \pm 1.96 * \sqrt{\frac{.5(.5)}{1500}} = \pm .025$
- With N=2000, 95% confidence interval is $\pm .022 = \pm 2.2\%$