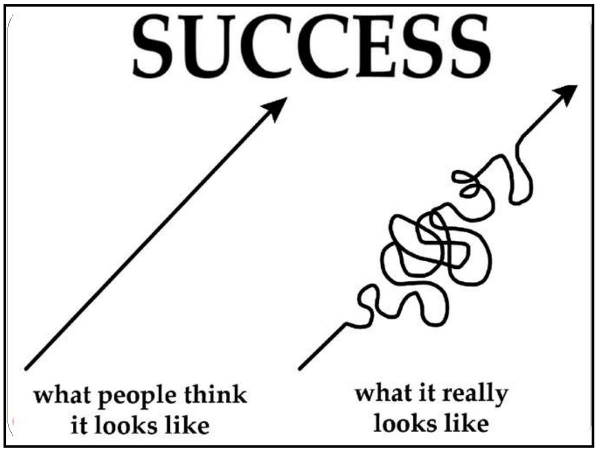


SOC424 – Statistics w/ Dr. Ellis Godard

Sampling Effects

My girl just texted me this she double not pregnant that was close🙆

Please read *all* instructions ;)

Outline for Today…

- **Review –** esp. sample vs population
  - Please stop using "sample population"; meaningless & a red flag
- **Bias –** What makes a good sample?
- **Types** of samples
- **Distributions** (esp. *Sampling*)
  - **Example –** using student evals of EG
- **Standard Error**
- **Lab**

Review to Date

- **Goal:**
  - Want **inferences from sample statistics** (data about what we observe)
  - **to population parameters** (numbers that describe a larger group of cases, beyond those we've been able to observe)
- **Constraints:**
  - How that happens **depends heavily on** how measured (especially **whether nominal, ordinal, or interval**)
  - Other main area of constraint: how cases were selected

Cases, Samples, & Populations

- A **case** is a **particular instance** studied
  - Individuals, groups, schools, states, etc.
- A **sample** is a **known**, practical idea
  - the set of just the cases actually observed
- A **population** is (usually) a **hypothetical** idea
  - total set of cases of interest in a study
  - May not be identifiable, or even known
- **Sample** = **subset** of the population of interest
  - e.g. class, CSUN, California residents
  - envision as concentric circles

Why Samples Matter: Ideas

- **Need to sample** - can't study all cases
  - Too expensive?
  - Too time consuming?
  - Or may not even know who the cases are
- **If follow rules, can use *sampling theory***
  - Allows inferences to unobserved cases
  - Under certain conditions / If certain conditions met

Homogeneity vs. Heterogeneity

- **Sometimes ok to assume all cases alike**
  - Blood samples
  - Social psychology
    - Studies based on sophomores in a psych class
- **Not typically a good assumption**
  - Sociologists, esp., *focus* on differences
  - Failure of early polls
    - Not adequately representing full range of voters
- **Data *reports* should also recognize this**
  - Lengthy example in notes

What makes a good sample?

- Ideal is to be **representative**
  - Aggregate characteristics clearly resemble same aggregate characteristics of the pop
  - Variation not *lacking* or *unbalanced*
  - In practice, may not *know* parameters
- Needn't be representative on all respects
  - Only on characteristics relevant to the study
  - But might not know what those are in advance
  - And don't want criticism that lack of representativeness hides some "control" variable

Good Samples, cont'd

- **Size matters**
  - But larger isn't always "better"
  - More important to be **unbiased**
- **Best practice is to *reduce bias***
  - risks making sampling unrepresentative
  - However, any method introduces some bias
    - assumptions about availability of subjects
    - Assumes results approximately representative

Decisions that Help/Hurt

- **Sampling Frame**
  - List of each & every case in the population from which the sample is to be drawn
  - Ideally, this is all possible cases
  - In practice, this is all the cases known or listed
    - e.g. mental patients, victims, pastors
  - Sampling with or without Replacement
- **Sampling Method**
  - Probability sampling improves chances by avoiding bias and allows statistical inferences based on assumptions of random samples
- **Weighting Data, when needed**

Weighting & Sample Weights

- Used if sample is disproportionate
  - E.g. population is 12.8% African-American, but sample is 6.4% African-American
  - Use a data "trick": pretend there were twice as many African-Americans
  - "weight by race" in SPSS
- Something to remember & consider
- But you won't weight data this semester
- And **don't pick "sample weight" as variable** ☹
  - *Not* interval in conventional sense
  - *Can't* be used to describe a sample

Overview: Two General Types

- Not necessarily survey (e.g. fingerprints)
  1. **Probability** – probability of selection of each case is known (though not necessarily equal)
  2. **Nonprobability** – probability of selection is *not* known (e.g. passersby on the street) – some bias or limitation
- With *random probability sampling* techniques:
  - *More likely* to be representative
    - Can't guarantee; never "perfect" (LOL)
  - Can estimate *degree* of representativeness
- Few samples truly RPS, but try best possible

Non-probability Samples (1st 5)

1. **Available Subjects:** prone to bias; ok for pretest
2. **Volunteer/Convenience:** e.g. self-admin insert
   - sampling biased to extremes (those w/ interest) and to negatives (those w/ complaints)
   - Oft used in marketing but not much scientific use
3. **Purposive/Judgmental:** id typical group
   - Good for initial design w/ accessible subset
   - Deviant/Trouble cases (what not fit gen. pattern)
4. **Snowball:** start w/ known informants, ask them for others, and accumulate more as you roll on
5. **Informants:** usually marginal so may not know; may be powerful, public, or flamboyant
6. **Quota Sampling:** select based on intersections of demographics, to ensure get a proportionate (?) sample

Probability Samples

1. **Simple Random Sample**
   - Random number table picks the cases
2. **Systematic Random Sample**
   - Use a "skip number" to take every Kth case
3. **Stratified Random Sampling**
   - Dictated by theory & data parameters
   - Internally homogenous
4. **Cluster Sampling**
   - Chosen for convenience; practicality/real life guide
   - Internally heterogenous

An(other) example…

MTV's *Real World*:

- **The cast of each season is a cluster**
  - It's a practical group of possible cases
  - Any one cast is pretty much like the rest
  - Heterogenous within each
- **Strata would be groups across casts**
  - E.g. age, race, gender, sexual orientation
  - Would need list of all the members of all the casts, or at least a list of percentages of each subgroup (male and female; black white and other, etc.)

Possible # of Samples HUGE!

For a population size (N) is 26, how many different samples of size n=10 are there?

Number of Permutations:

$$\frac{26!}{(26-10)!} = 26 \times 25 \times 24 \times 23 \times 22 \times 21 \times 20 \times 19 \times 18 \times 17 = 1.927 \times 10^{13}$$

Number of Combinations (Samples):

$$\frac{26!}{10!(26-10)!} = 312{,}455$$

They Make 4th Type Distribution

- ***Population Distribution:*** "Real" arrangement of a variable's data, in the population being studied
- ***Sample Distribution:*** Arrangement (as illustrated in a histogram) of the data actually collected or observed
- ***Normal Distribution:*** An arrangement w/ particular shape characteristics (similar to a "bell curve", but much more specific)
- ***Sampling Distribution:*** The distribution of sample means (that is, the collection of all of the means of all possible samples, for a given variable and population)

Really Two Different Types

- **Data Distributions** - what we want to relate
  - Sample (empirical)
  - Population (usually hypothetical)
- **Probability Distributions** - what we use to do it
  - Refers to likelihood of specific values
  - Two forms discussed:
    - Normal (symmetric, bell-shaped)
    - Standard Normal (last 2 lectures)
  - One key example: Sampling Distribution
    - Refers to the probabilities associated with some statistic
    - Probability distribution for some sample statistic.
      - Mean, variance, standard deviation, etc.

Distribution of Sample Means ($\bar{Y}$)

Consider this (tedious) procedure:

a. draw a sample of size n from a population;

b. compute the mean for this sample;

c. repeat (a) and (b) for *every possible sample of size n* you can draw from this population;

d. draw a histogram of the sample means obtained in (c) and compute the mean and standard deviation corresponding to this histogram.

Those Sample Means as Data

- **Their collection is a distribution itself**
  - It has a shape
  - It has (a?) central tendency
  - It has dispersion
  - Each of these can be measured
- **Histogram of all possible sample means**
  - Call this a *sampling distribution*, distinct from a *sample distribution*
  - It's standard deviation is the *standard error*

Illustrate w/ a Small Population

- **3 students as a population**
- **Variable is satisfaction w/ class**
- **N=3, with values of 1, 2, and 3**
  - Mu = 2
- **3 possible samples of 2:** 1,2 1,3 2,3
  - Means of those samples: 1.5 2 2.5
  - Mean of those means: 2

Illustrate w/ Larger Population

- **N=4, with values of 1, 2, 3, and 4**
  - Mu = 2.5
- **6 possible samples of n=2:** 12 13 23 14 24 34
  - Means of those samples: 1.5 2 2.5 2.5 3 3.5
  - Mean of those means: 2.5
- **4 possible samples of n=3:** 123 124 134 234
  - Means of those samples: 2 2.33 2.66 3
  - Mean of those means: 2.5
  - *Notice that the sample means are closer together!*

Two *Sampling* Distributions

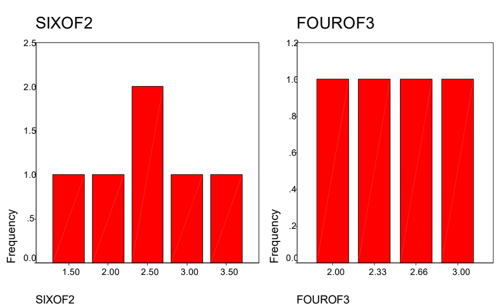
SIXOF2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.50	1	16.7	16.7	16.7
	2.00	1	16.7	16.7	33.3
	2.50	2	33.3	33.3	66.7
	3.00	1	16.7	16.7	83.3
	3.50	1	16.7	16.7	100.0
	Total	6	100.0	100.0	

FOUROF3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	1	16.7	25.0	25.0
	2.33	1	16.7	25.0	50.0
	2.66	1	16.7	25.0	75.0
	3.00	1	16.7	25.0	100.0
	Total	4	66.7	100.0	
Missing	System	2	33.3		
Total		6	100.0		

Sampling Distribution Histograms



Sampling Variances for $\bar{Y}$

Variance of the sampling distribution is equal to the variance of the population distribution averaged across the sample size:

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}$$

The square root of each side gives the "standard error", the sampling distribution's standard deviation:

$$\sqrt{\sigma_{\bar{Y}}^2} = \sqrt{\frac{\sigma_Y^2}{n}} \quad \cdots\rightarrow \quad \sigma_{\bar{Y}} = \frac{\sqrt{\sigma_Y^2}}{\sqrt{n}} = \frac{\sigma_Y}{\sqrt{n}}$$

Since we use the sample standard deviation as an estimate for the population standard deviation, we estimate the standard error:

$$\hat{\sigma}_{\bar{Y}} = \frac{s_Y}{\sqrt{n}}$$

Sampling Error (aka Sample Error)

- Every sample statistic differs somewhat from the population parameter for which it is a point estimate
- That difference is the *sampling error*
- If we knew the parameter, we could just subtract the statistic & calculate that error
- But since we rarely know the parameter, we need a means of *estimating* how big that difference (that "error") is likely to be

Standard Error: In a nutshell

- **Don't confuse with either the *standard deviation* or the *sampling error***
- **The *standard error* is the (estimated) standard deviation of a sampling distribution**
  - A standard deviation tells how much sample values are likely to differ from the mean of the *sample* distribution.
  - The standard error tells how much the means of random samples are likely to differ from the (grand) mean of the *sampling* distribution.
- **Key tool for inferences and estimations**

The Standard Error: Formula

- Simple but crucial:

Population Standard Deviation

$$\sigma_{\bar{Y}} = \sigma / \sqrt{n}$$

Standard Error (Standard Deviation of Sampling Distribution)

sample size (n)

Make sure that you distinguish the population standard deviation from the standard error.

They are not the same: subscript for standard error is "Y bar" – it's the *standard deviation of sample means*

Larger Samples, Smaller Errors

- As the sample size increases, the standard error decreases
  - A larger denominator makes the fraction smaller, since the standard deviation is divided by a larger number
- This is one of the implications of the *Central Limit Theorem*, the topic for the next lecture

Central Limit Theorem

If a sample is taken randomly and if it is sufficiently large, the *sampling* distribution is normal, even if the *data* distribution (that is, the sample or population data) is not normal

SPSS Example (Demo, as/if needed)

- Using gss88a.sav
- Look at histogram for "Number of Children"
  - Mean of all 1481 cases is 2.02
    - Don't round that – it's meaningful *as is*
- Randomly select a sample of ten respondents
  - Use DATA - "Select cases", just like before

1. Find mean number of children for that sample
2. What's the *sampling* error? $|\bar{Y} - \mu|$
3. What's the *standard* error? $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

For your next lab…

- **Lab form on Canvas ("Sampling")**
  - Using musicB.sav
  - Looking at the mean for 1606, and taking a subsample, then a smaller subsample (like HW3)
  - Will need to provide the sample means, and *calculate* sampling errors and standard errors *(show work!!)*
- **NOT a long lab (that's 3 shorts in a row!)**
  - Plenty of time – unless you're new to SPSS ☹
  - Use time wisely – extra time for HWs etc.