DATA 755
September 8, 2020

- **Join Zoom Meeting**
  https://us02web.zoom.us/j/83671716150?pwd=NTY4V3Y0VFRHN204dWhFMi9qM1FNUT09

  Meeting ID: 836 7171 6150
  Passcode: 6162816

1st Lessons from Last Week's Homework

- ALL email Soc755.CFTurner@gmail.com
- Make YOUR NAME part of the file name, e.g., Turner_Homework_1.docx
- PUT name at top of file. In STATA make your first command, *Student X

2nd Lessons from Last Week's Homework

- NEVER (or almost never) "Save" your active data when you exit STATA
- PUT name at top of file. In STATA make your first command, * Student X
- BETTER –> carry forward all recodes into DO file for next analyses
- STATA log files should be generated as text files not as SMCL. STATA command —> *set logtype text, perm*
- Generate more compact LOGs by seting linesize to 100 or more
- BRAVO to those students who figured out how to use DO files

Exploring large datasets

- **Good programming hygiene:**
  comments everywhere, save commands
- **Verify Integrity of Data**
- **USE the documentation**
  Sample design, Codebook, Questionnaire

REJECTING NULL HYPOTHESIS

- Statement

The obtained chi-square is in the critical region.

134.478 > 21.026 (chi2 critical when df = (5-1)(4-1)=12)

The probability that sexual orientation and region of residence are independent of each other is 0%. We reject the null hypothesis of independence. There is a statistically significant relationship between sexual orientation and region of residence. Therefore, there is association between sexual orientation and region of residence.

BODY MASS INDEX (BMI)

The formula for BMI was devised in the 1830s by Belgian mathematician Adolphe Quetelet. BMI is universally expressed in kg/m2.

SOURCE:
www.thecalculatorsite.com/articles/health/bmi-formula-for-bmi-calculations.php

Weight lbs	100	105	115	110	120	125	130	135	140	145	150	155	160	165	170	175	180	185	190	195	200	205	210	215
Weight kgs	455	477	523	505	545	568	591	614	636	659	682	705	727	750	773	795	818	841	864	886	909	932	955	977
5'0" / 152.4	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
5'1" / 154.9	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
5'2" / 157.4	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	33	34	35	36	37	38	39
5'3" / 160.0	17	18	19	20	21	22	23	24	24	25	26	27	28	29	30	31	32	32	33	34	35	36	37	38
5'4" / 162.5	17	18	18	19	20	21	22	23	24	24	25	26	27	28	29	30	31	31	32	33	34	35	36	37
5'5" / 165.1	16	17	18	19	20	20	21	22	23	24	25	25	26	27	28	29	30	31	31	32	33	34	35	35
5'6" / 167.6	16	17	17	18	19	20	21	21	22	23	24	25	25	26	27	28	29	29	30	31	32	33	34	34
5'7" / 170.1	15	16	17	18	18	19	20	21	22	22	23	24	25	25	26	27	28	29	29	30	31	32	33	33
5'8" / 172.7	15	16	16	17	18	19	19	20	21	22	22	23	24	25	25	26	27	28	28	29	30	31	32	32
5'9" / 175.2	14	15	16	17	17	18	19	20	20	21	22	22	23	24	25	25	26	27	28	28	29	30	31	31
5'10" / 177.8	14	15	15	16	17	18	18	19	20	20	21	22	23	23	24	25	25	26	27	28	28	29	30	30
5'11" / 180.3	14	14	15	16	16	17	18	18	19	20	20	21	21	22	23	23	24	25	25	26	27	28	28	29	30
6'0" / 182.8	13	14	14	15	16	16	17	17	18	19	19	20	21	21	22	23	23	24	25	25	26	27	27	28	29
6'1" / 185.4	13	13	14	15	15	16	16	17	17	18	19	19	20	21	21	22	23	23	24	25	25	26	27	27	28
6'2" / 187.6	12	13	14	14	15	15	16	16	17	18	18	19	19	20	21	21	22	23	23	24	25	25	26	27	27
6'3" / 190.5	12	13	13	14	15	15	16	16	17	18	18	19	20	20	21	21	22	23	23	24	25	25	26	26
6'4" / 193.0	12	12	13	14	14	15	15	16	17	17	18	18	19	20	20	21	22	22	23	23	24	25	25	26
Height in/cm	Underweight			Healthy			Overweight			Obese			Extremely obese											

IN-CLASS EXERCISE

- **Examine BMI in adult 1997-2017 probability segment of NHIS sample (N = 646 thousand)**
- NOTE we are using 1997-2017 because of anomaly in 1997-2018 dataset

SEE List of World's Heaviest People

From Wikipedia

Coping with Skew and Kurtosis

- Read article posted on website
- A distribution that is fully normal will have a skew of zero and a kurtosis of 3
- Variable transformations can produce more nearly normal distributions of observations

Homework due 11pm on 9-13

- Identify 5 key variables associated with BMI
- Run basic descriptive analyses
- Identify problems that require remedy
- Remedy the problems
- Re-run descriptive analyses to verify fixes
- Save COMMAND files as DO file
- **Do NOT re-save the data file**

NHIS Sample Design

Cross-sectional household interview survey

- **Sampling and interviewing continuous thru year.**
- **Multistage area probability design**
- **Representative sample of households & group qtrs**
- **Sampled clusters of addresses in primary sampling units (PSU's).** PSU consists of — county, a small group of contiguous counties, or a metropolitan statistical area.
- **NO oversample race/ethnicity groups at household level.**
- **BUT persons aged 65 or older, blacks, Hispanics, Asians have higher chance to be selected in household/group quarters**
- **ALSO representative sample of children**

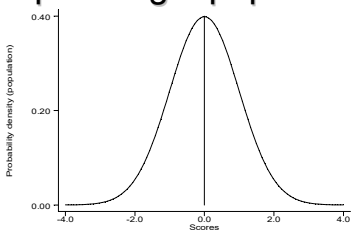
Basic Sampling Theory for Simple and Cluster Samples

Malcolm Rosier
Survey Design and Analysis Services Pty Ltd
http://survey-design.com.au

Sample design

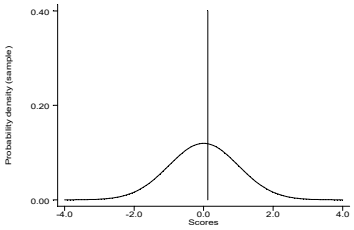
- The focus of the design for a sample must be on the magnitude of the standard errors of sampling not than on an arbitrary percentage of the target population.
- The standard errors are used to calculate confidence intervals around the sample data.

Graph: Target population



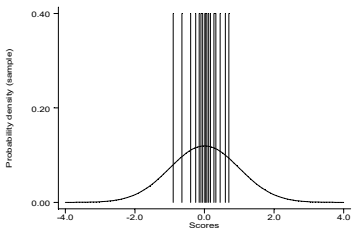
- Population: mean = μ, standard deviation = σ

Graph: Sample



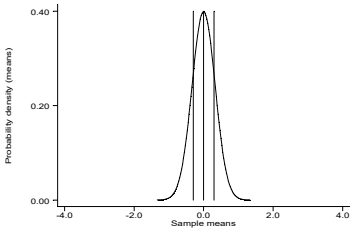
- Sample: mean = x, standard deviation = s

Graph: Means from many samples



- However we could get many different samples with different sample means from the population.

Graph: Distribution of sample means



- This gives us a sampling distribution of sample means:

Sampling distribution of sample means

- normal distribution
- mean = μ = mean of underlying population distribution
- standard deviation = $\sigma / \sqrt{n}$

Standard error of a population mean

The standard deviation of the sampling distribution of sample means is termed the standard error of a mean.

standard error of population mean = $\sigma / \sqrt{n}$

Standard error of a proportion (srs)

The standard error of a percentage (proportion) is:

$se(prop) = \sqrt{[p(1-p)/n]}$

Confidence intervals

- Confidence intervals are usually expressed at the 95 per cent level (1.96 standard errors of sampling for a proportion)

Table: Effect of sample size on standard error

Size	se(p50)	lower 95 ci	upper 95 ci
100	0.050	0.402	0.598
200	0.035	0.431	0.569
500	0.022	0.456	0.544
1000	0.016	0.469	0.531
2000	0.011	0.478	0.522

Two stage samples

- The most efficient method is usually sampling at the first stage with probability proportional to size (pps).
- This produces a self-weighting sample.
- Easier logistics for administration.

Two stage samples

Stage 1

Primary sampling units (psu) are selected with a probability proportional to the size of the target population in the psu.

Example of psu: schools

Two stage samples

Stage 2

A random cluster of secondary sampling units (ssu) is selected at random from each of the psu.

Example of ssu: students in schools

Deff

- Two-stage sampling is less efficient than a simple random sample (srs) of the same size.

deff = (standard error of sampling for complex sample)^2 / (standard error of sampling for srs)^2

Deft

- The square root of deff is deft, which gives the ratio of the standard errors of sampling.

deft = (standard error of sampling for complex sample) / (standard error of sampling for srs)

Simple equivalent sample

- The simple equivalent sample (ses) is the size of a simple random sample which has the same standard error as the complex sample.
- We sometime use the term effective sample (n_{eff})

Table: Values for deff and simple equivalent sample

	psu	ssu	total	rho	deff	ses
1	50	20	1000	0.05	1.95	513
2	50	20	1000	0.10	2.90	345
3	50	20	1000	0.20	4.80	208
4	20	50	1000	0.05	3.45	290
5	20	50	1000	0.10	5.90	169
6	20	50	1000	0.20	10.80	93

ESTIMATE	DEFT	Clustered Sample Size Needed to equal SRS Sample of 1,000
% Strong Republicans	1.24	1,538
% Libras	1.03	1,061
Mean Years of Education	1.92	3,686
% Black	2.57	6,605

WEIGHTING

- First-stage weights are typically inverse of selection probability
- Second-stage weights provide additional precision, e.g., ratio-adjustment to known Census marginals
- Weights may apply at several levels, e.g., household weights, person weights, etc.