`gretl` Notes

Week 3

Module 1 Week 3

- Dataset used: `Health Behavior in School Aged Children 2002.gdt`
  - Open with `File` -> `Open data` -> `User file...`
- Open up a new script editor: `File` -> `Script files` -> `New script` -> `gretl script`

Estimate the Mean BMI:

- Capital letters are annoying; rename:

```
rename BMI bmi
```

- Output:

```
? rename BMI bmi
Listing 81 variables:
  0) const        1) TYPE         2) Q1           3) Q4
  4) AGE          5) RACE         6) Q7           7) Q9
  8) Q10A         9) Q10B        10) Q12         11) Q14
 12) Q15A1       13) Q15A2       14) Q15A_BRO    15) Q15A_SIS
 16) Q17         17) Q18         18) Q19A        19) Q19B
 20) Q20A        21) Q20B        22) Q21         23) Q22
 24) Q23A        25) Q23B        26) Q24A        27) Q24B
 28) Q25A        29) Q25B        30) Q26A        31) Q26B
 32) Q27A        33) Q27B        34) Q28A        35) Q28B
 36) Q28C        37) Q28D        38) Q28E        39) Q28F
 40) Q28G        41) Q28H        42) Q28I        43) Q28J
 44) Q28K        45) Q28L        46) Q28M        47) Q28N
 48) Q28O        49) Q29         50) Q30         51) Q31
 52) bmi         53) BMI_COMP    54) Q32         55) Q33
 56) Q34         57) Q60         58) Q61         59) Q62C
 60) Q62D        61) Q62G        62) Q64         63) Q65
 64) Q71         65) Q74         66) Q76         67) Q77
 68) Q80         69) Q81B        70) Q81C        71) Q82
 72) Q85A        73) Q85B        74) Q85C        75) Q86
 76) Q87         77) Q88A_COMP   78) Q88B_COMP   79) F_JOB4
 80) M_JOB4
```

- Calculate the *Estimate* of the Mean of BMI:

```
scalar meanbmi = mean(bmi)
```

- Output:

```
? scalar meanbmi = mean(bmi)
Generated scalar meanbmi = 20.9922
```

- Calculate a *95% Confidence Interval* for the Mean of BMI:
  - *Population Standard Deviation* is unknown => need to estimate it via the Student's t-distribution:

```
scalar sdbmi = sd(bmi)
```

- Recall: $CI = \bar{X} \pm t_{df} * SE$ and $df = n - 1$
- Note that 2060 out of 14817 observations are missing:

```
summary bmi
```

- Output:

```
? summary bmi

Summary statistics, using the observations 1 - 14817
for the variable 'bmi' (12757 valid observations)

  Mean                      20.992
  Median                    20.200
  Minimum                   10.980
  Maximum                   48.650
  Standard deviation        4.3384
  C.V.                     0.20667
  Skewness                  1.1085
  Ex. kurtosis              2.0534
  5% percentile             15.350
  95% percentile            29.521
  Interquartile range       4.9600
  Missing obs.                2060
```

- Calculate *Degrees of Freedom*:

```
scalar dfbmi = $nobs - 2060 - 1

print dfbmi
```

- Output:

```
? scalar dfbmi = $nobs - 2060 - 1
Generated scalar dfbmi = 12756
? print dfbmi

          dfbmi =  12756.000
```

- So we have $df = 12756$
- Calculate the *Standard Error* ($SE = \frac{SD}{\sqrt{n}}$); $n = df + 1$:

```
scalar se_meanbmi = sdbmi/sqrt(dfbmi+1)
```

- Output:

```
? scalar se_meanbmi = sdbmi/sqrt(dfbmi+1)
Generated scalar se_meanbmi = 0.0384112
```

- Thus, $SE \approx 0.0384112$. Small value, which is unsurprising because n is pretty large; larger n => better precision with estimates.
- Calculate the 95% CI:
  - `critical(t, dfbmi, .025)` to get a *T-Score* critical value with `dfbmi` Degrees of Freedom and 0.25 tail probability
  - the syntax when using `critical` --> `critical(distribution, parameters, tail prob)`:

```
scalar ubound = meanbmi + (critical(t, dfbmi, .025)*se_meanbmi)
scalar lbound = meanbmi - (critical(t, dfbmi, .025)*se_meanbmi)
print lbound ubound
```

- Output:

```
? scalar ubound = meanbmi + (critical(t, dfbmi, .025)*se_meanbmi)
Generated scalar ubound = 21.0675
? scalar lbound = meanbmi - (critical(t, dfbmi, .025)*se_meanbmi)
Generated scalar lbound = 20.9169
? print lbound ubound

         lbound =  20.916927

         ubound =  21.067511
```

- Thus, CI: $(20.916927, 21.067511)$

Estimate Single Proportion:

- Let's estimate the proportion of children whose father lives at home with them
  - We can find the name of the variable in the description; turns out to be `Q15A2`, where `1` means "father lives at home". But we still need to replace the `2` which means "father doesn't live at home" with `0`

```
freq Q15A2
```

- Output:

```
? freq Q15A2

Frequency distribution for Q15A2, obs 1-14817

          frequency    rel.     cum.

  1        9220     62.23%   62.23% **********************
  2        5597     37.77%  100.00% *************
```

- Make the 2's 0's:

```
# Better name than `Q15A2`:
series dadhome = (Q15A2==1)

# Tell `gretl` this is a discrete variable:
discrete dadhome

# Make sure all is good
freq dadhome
```

- Output:

```
? series dadhome = (Q15A2==1)
Generated series dadhome (ID 81)
? discrete dadhome
? freq dadhome

Frequency distribution for dadhome, obs 1-14817

          frequency    rel.     cum.

  0        5597     37.77%   37.77% *************
  1        9220     62.23%  100.00% **********************
```

- Get the proportion; note that having recoded to 0's and 1's from 1's and 2's allows us to use the `mean` command for proportion calculation:

```
scalar dadprop = mean(dadhome)
```

- Output:

```
? scalar dadprop = mean(dadhome)
Generated scalar dadprop = 0.622258
```

- Thus, $\hat{p} \approx 0.622258$
- Now let's calculate a 95% confidence interval for this estimated proportion:
  - Recall: We can approximate the sampling distribution for a proportion with a normal distribution with a mean $\hat{p}$ and standard deviation equal to $\sqrt{\hat{p}(1-\hat{p})/n}$
  - *Standard Error* for the *Proportion*:

```
scalar se_prop = sqrt((dadprop*(1-dadprop))/$nobs)
```

- Output:

```
? scalar se_prop = sqrt((dadprop*(1-dadprop))/$nobs)
Generated scalar se_prop = 0.00398293
```

- Thus, $SE \approx 0.00398293$
- Now, the 95% CI; no distribution params needed for `z`, because it's already a Standard Normal Distribution:

```
scalar uboundprop = dadprop + (critical(z, .025)*se_prop)
scalar lboundprop = dadprop - (critical(z, .025)*se_prop)
print lboundprop uboundprop
```

- Output:

```
? scalar uboundprop = dadprop + (critical(z, .025)*se_prop)
Generated scalar uboundprop = 0.630065
? scalar lboundprop = dadprop - (critical(z, .025)*se_prop)
Generated scalar lboundprop = 0.614452
? print lboundprop uboundprop

     lboundprop =  0.61445182

     uboundprop =  0.63006462
```

- Thus, CI: $(0.61445182, 0.63006462)$

One-Sample Hypothesis Test for the Mean

- Let's test if the mean BMI in the population is different from 22
  - H_o: $\mu_{BMI} = 22$
  - H_a: $\mu_{BMI} \neq 22$
  - Recall: *Test Statistic* $T = \frac{\text{Estimate} - \text{Mean}_{H_0}}{\text{SE}}$:

```
# `meanbmi` and `se_meanbmi` were calculated earlier:
scalar tstat_bmi = (meanbmi - 22)/se_meanbmi
```

- Output:

```
? scalar tstat_bmi = (meanbmi - 22)/se_meanbmi
Generated scalar tstat_bmi = -26.2366
```

- Thus, $T = -26.2366$
- What is the associated p-value?:

```
# `t` - means from T-Distribution
# `dfbmi` - degrees of freedom calculated earlier
# `tstat_bmi` - the Test Statistic calculated earlier
pvalue t dfbmi tstat_bmi
```

- Output:

```
? pvalue t dfbmi tstat_bmi
t(12756): area to the right of -26.2366 =~ 1
(to the left: 4.08235e-148)
(two-tailed value = 8.16469e-148; complement = 1)
```

- Thus, $p - value \approx 8.16469 * 10^{-148} \approx 0$ => reject the H_0
- What if we tested a BMI of 21 instead of 22?

```
scalar tstat_bmi2 = (meanbmi - 21)/se_meanbmi
pvalue t dfbmi tstat_bmi2
```

- Output:

```
? scalar tstat_bmi2 = (meanbmi - 21)/se_meanbmi
Generated scalar tstat_bmi2 = -0.202567
? pvalue t dfbmi tstat_bmi2
t(12756): area to the right of -0.202567 = 0.580262
(to the left: 0.419738)
(two-tailed value = 0.839477; complement = 0.160523)
```

- Thus, $p - value \approx 0.839477$ => We fail to reject the H_0
- Conclusion: there is a statistically significant difference between the Sample Mean of BMI and 21, but not 22.
    - Remind ourselves what the Sample Mean BMI was:

```
print meanbmi
```

- Output:

```
? print meanbmi

        meanbmi =  20.992219
```

- Interpretation: it is much harder to estimate a small difference or effect than a larger one

Hypothesis Test for a Single Proportion

- Let's test if the proportion of kids with dad at home is different from .63
    - H_0: $P = .63$
    - H_a: $P \neq .63$

```
scalar null_value = 0.63
scalar se_prop_null = sqrt((null_value*(1-null_value))/$nobs)
scalar zstatprop = (dadprop - null_value)/se_prop_null
pvalue z zstatprop
```

- Output:

```
? scalar null_value = 0.63
Replaced scalar null_value = 0.63
Warning: "= sqrt((null_value*(1-null_value))/$nobs)"
 obsolete use of "=" as Boolean test: please use "=="

? scalar se_prop_null = sqrt((null_value*(1-null_value))/$nobs)
Generated scalar se_prop_null = 0.00396635
? scalar zstatprop = (dadprop - null_value)/se_prop_null
Replaced scalar zstatprop = -1.95187
? pvalue z zstatprop
Standard normal: area to the right of -1.95187 = 0.974523
(to the left: 0.0254771)
(two-tailed value = 0.0509541; complement = 0.949046)
```

- Thus, we fail to reject H_0, because $0.051 > 0.05$
    - Very close, so we should be cautious. If possible, we should re-draw the sample and re-perform the test.