Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Pset 0 due Friday 9/15

Introductions

- Name
- Year
- Previous stats courses
- One question or thought related to lecture last week

Goals each week

- Hand out and explain R code for the week. New relative to last year, we'll plan to not do any in-section coding questions. LLMs are good enough now to do most of your coding for you (and they're allowed in this class!).
- See similar examples to the homework (both in code and analysis).
- Learn something about the world.

Effective sample size

The following problems are intended as a review of Stat 110.

1. Suppose there is a gambler who goes to the casino for n days and makes $Z_1, Z_2, \ldots, Z_n \sim \mathcal{N}(0,1)$ each day where the winnings are independent of each other. (You can assume these are in thousands if the stakes aren't high enough.) What is the distribution of $\bar{Z}$?

2. Now, suppose the gambler tends to win and lose in streaks. In particular, let $X_1, X_2, \ldots, X_n \sim \mathcal{N}(0,1)$ marginally be the winnings, but assume neighboring days have correlation ρ. That is,

$$\vec{X} \sim \text{MVN}(\vec{0}, \mathbf{\Sigma}), \mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & 0 & 0 & \ldots \\ \rho & 1 & \rho & 0 & \ldots \\ 0 & \rho & 1 & \rho & \ldots \\ 0 & 0 & \rho & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

   Intuitively, should the variance of $\bar{X}$ be higher or lower than the variance of $\bar{Z}$?

3. What is the distribution of $\bar{X}$?

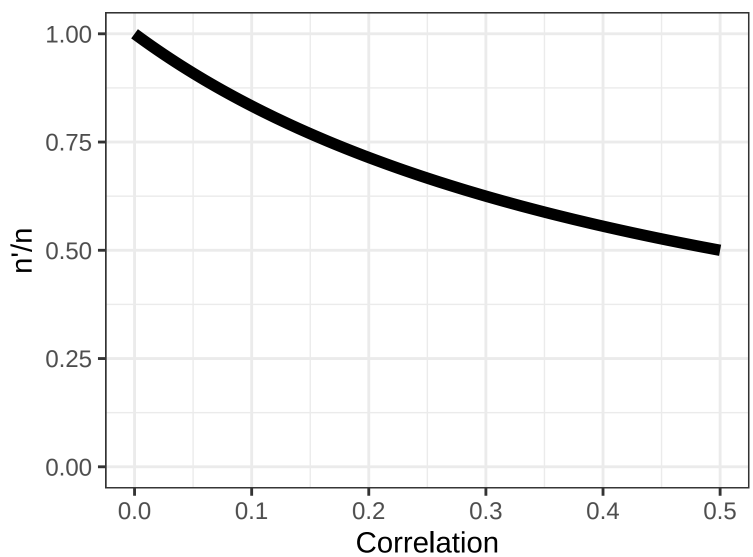
4. What would the distribution be if the X_i had variance σ^2 instead of 1 but everything else remained the same?

5. Show that the variance can be written as $\vec{c}^T \boldsymbol{\Sigma} \vec{c}$ where $\vec{c}$ is a vector of $1/n$.

6. What is the approximate distribution for large n?

7. By comparing the distributions in (1) and (6), determine the effective sample size n' when there are n random variables with the correlation structure of (2). That is, if you had n' independent Normals rather than n dependent Normals, what would n' have to be so that the variances of the sample means are the same?

8. Here is a plot of how the effective sample size changes with ρ.



We can test that our calculations are right by using a simulation. Explain what the following code does and whether the results agree with our expectations.

```r
library(MASS) # For Multivariate Normal
set.seed(139)

nsim <- 10^5
n <- 70
p <- 0.2
n_eff <- as.integer(n / (1 + 2 * p))

Sigma = matrix(0, nrow = n, ncol = n)
diag(Sigma) <- 1
for (i in 2:n) {
  Sigma[i, i-1] <- p
  Sigma[i-1, i] <- p
}

outputs <- matrix(nrow = nsim, ncol = 3)
for (i in 1:nsim) {
  x <- rnorm(n, 0, 1)
  outputs[i,1] <- mean(x)

  x <- rnorm(n_eff, 0, 1)
  outputs[i,2] <- mean(x)

  x <- mvrnorm(n = 1, rep(0, n), Sigma)
  outputs[i,3] <- mean(x)
}

variances_out <- apply(outputs, 2, var) # Apply over columns
names(variances_out) <- c("Independent n", "Independent n'", "Dependent n")
variances_out
```

```
##  Independent n Independent n'    Dependent n
##     0.01432340     0.01990641     0.01996290
```

9. You might have noticed that the plot of effective sample size versus correlation stops at a correlation of 0.5. Correlation ranges from -1 to 1, but our set-up actually doesn't work if $\rho > 0.5$ and n is large enough. To have a valid $\mathbf{\Sigma}$ matrix, it must satisfy the property that $\vec{x}^T\mathbf{\Sigma}\vec{x} \geq 0$ for all $\vec{x} \in \mathbb{R}^n$ (that is, it must be positive, semi-definite). Show that for $\rho > 0.5$, choosing the vector $\vec{x} = (-1, 1, -1, ..., -1)^T$ implies $\vec{x}^T\mathbf{\Sigma}\vec{x} < 0$ if n is large enough, violating the requirements for $\mathbf{\Sigma}$. (For simplicity, let n be odd.)

Student-t vs Normal

The following problems are intended as a review of Stat 111. We'll prove that the student-t distribution converges to the Normal distribution as its degrees of freedom increase and then analyze this convergence. This fact is useful for large n approximations.

1. Let $T_n \sim t_n$, so T_n can be represented as

$$T_n = \frac{Z}{\sqrt{V_n/n}}, Z \sim \mathcal{N}(0,1), V_n \sim \chi_n^2$$

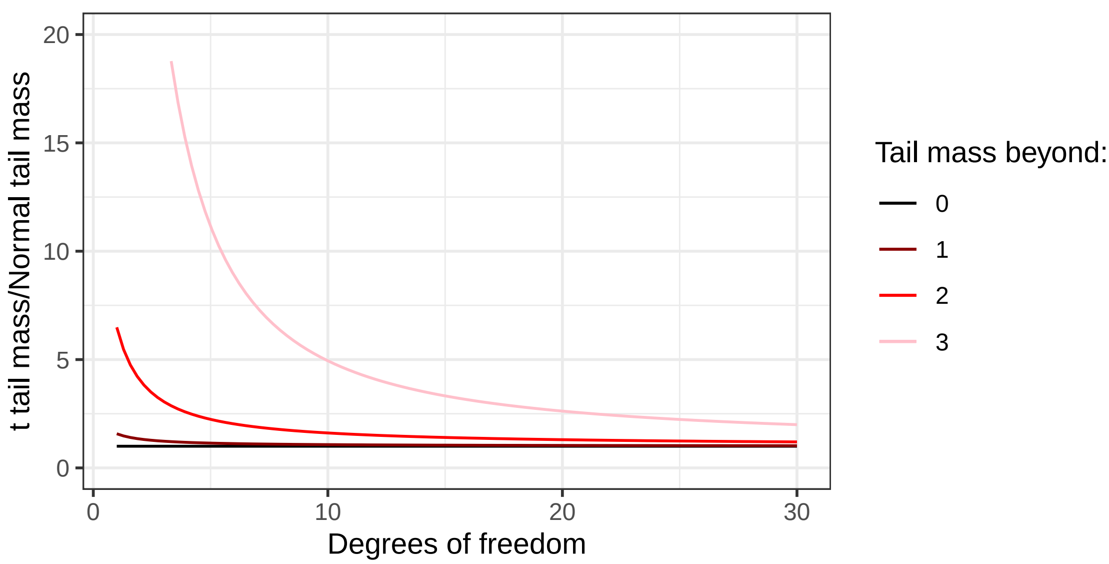
   which also means V_n can be represented as $V_n = \sum_{i=1}^n Z_i^2$ for $Z_i \sim \mathcal{N}(0,1)$. Show that $V_n/n \xrightarrow{p} 1$.

2. What tells us that if $V_n/n \xrightarrow{p} 1$, $\frac{1}{\sqrt{V_n/n}} \xrightarrow{p} 1$?

3. What tells us that if $Z \sim \mathcal{N}(0,1)$ and $\frac{1}{\sqrt{V_n/n}} \xrightarrow{p} 1$, $\frac{Z}{\sqrt{V_n/n}} \xrightarrow{d} \mathcal{N}(0,1)$

4. What does this mean about the distribution of T_n as $n \to \infty$?

5. Do the centers or the tails converge faster?



6. What does this imply about generating p-values from a Normal approximation to the student-t distribution?

Country demographics

These problems will deal with a data set of country-level statistics from UNdata and Varieties of Democracy.

1. Compare the following summary statistics for the 2010 populations (in millions of people) of Western African and Eastern African countries:

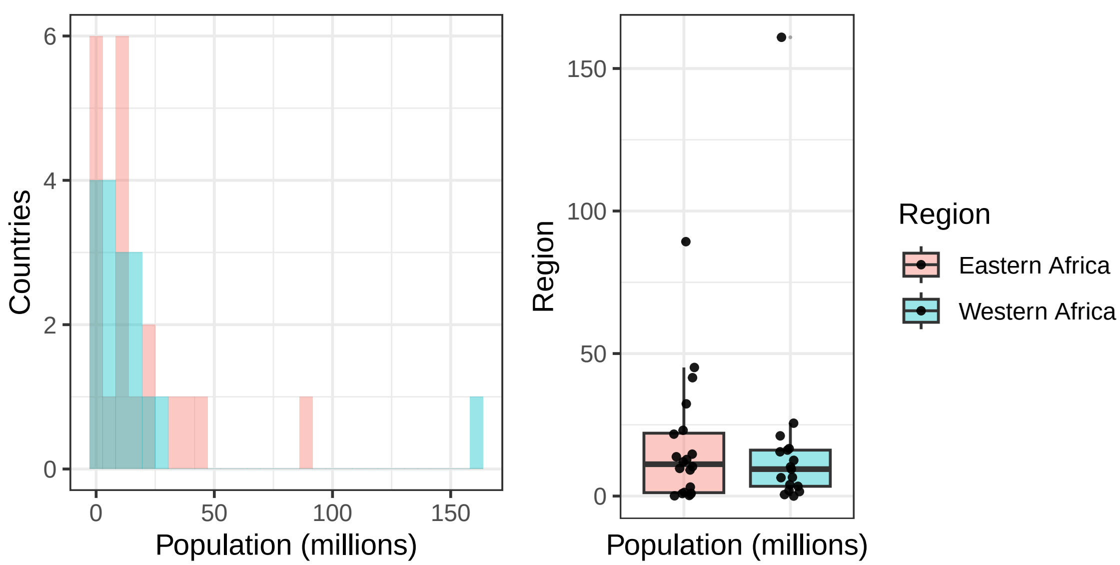
```
# Western Africa
pop1 <- countries[countries$Year == 2010 &
                    countries$Region == "Western Africa",
                  ]$`Population mid-year estimates (millions)`
round(c(summary(pop1), "SD" = sd(pop1)), 2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD
##    0.01    3.42    9.45   18.39   16.12  160.95   37.50
```

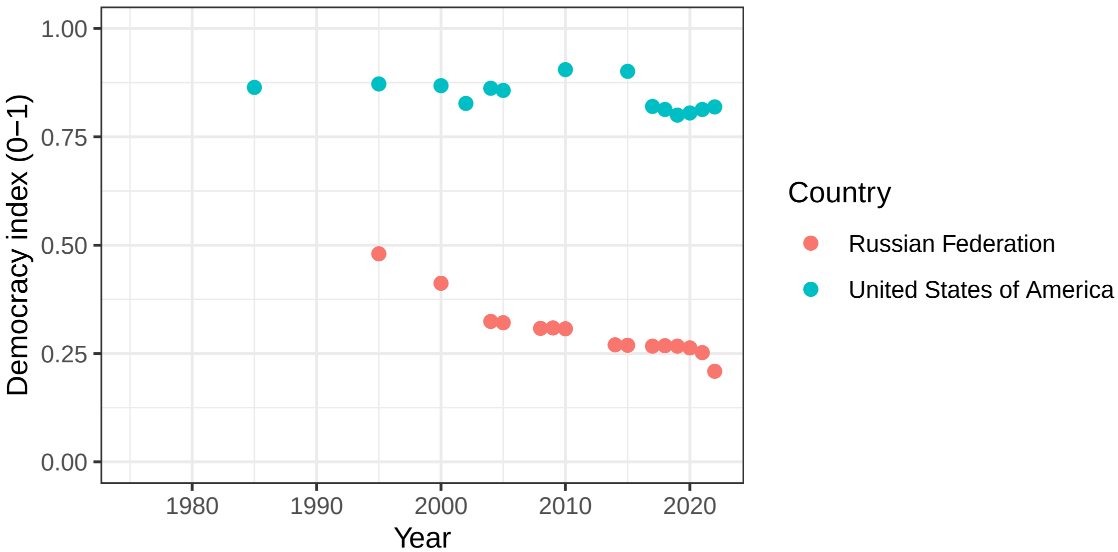
```
# Eastern Africa
pop2 <- countries[countries$Year == 2010 &
                    countries$Region == "Eastern Africa",
                  ]$`Population mid-year estimates (millions)`
round(c(summary(pop2), "SD" = sd(pop2)), 2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD
##    0.09    1.19   11.17   17.14   22.06   89.24   21.66
```

2. Compare the distributions. Would you expect to see a significant difference in a t-test?



3. Varieties of Democracy is a group of researchers that estimates a democracy score for each country each year based on a large compilation of data. Note any trends in the democracy index.


Democracy index (0–1)
1.00
0.75
0.50
0.25
0.00
1980
1990
2000
2010
2020
Year
Country
Russian Federation
United States of America