## 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 $\rho$. 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 $\bar{Z}$?

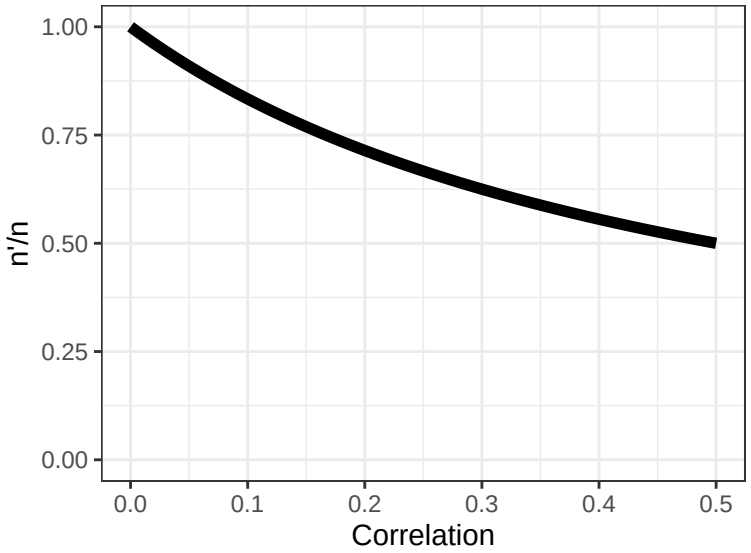
3. What is the distribution of $\bar{X}$?

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

5. What is the approximate distribution for large $n$?

6. By comparing the distributions in (1) and (5), determine the effective sample size $n'$ when there are $n$ random variables with the correlation structure of (2). That is, is 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?

7. Here is a plot of how the effective sample size changes with $\rho$.




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.

```
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)
```

```r
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
```

8. 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

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

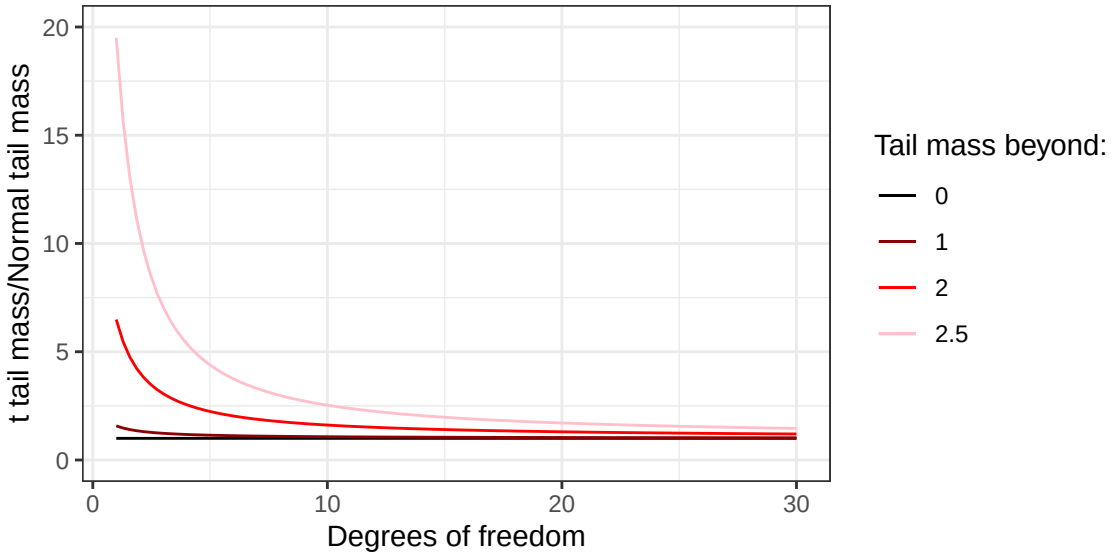
which also means $V$ can be represented as $V = \sum_{i=1}^{n} Z_i^2$ for $Z_i \sim \mathcal{N}(0,1)$. Show that $V/n \xrightarrow{p} 1$.

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

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

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:

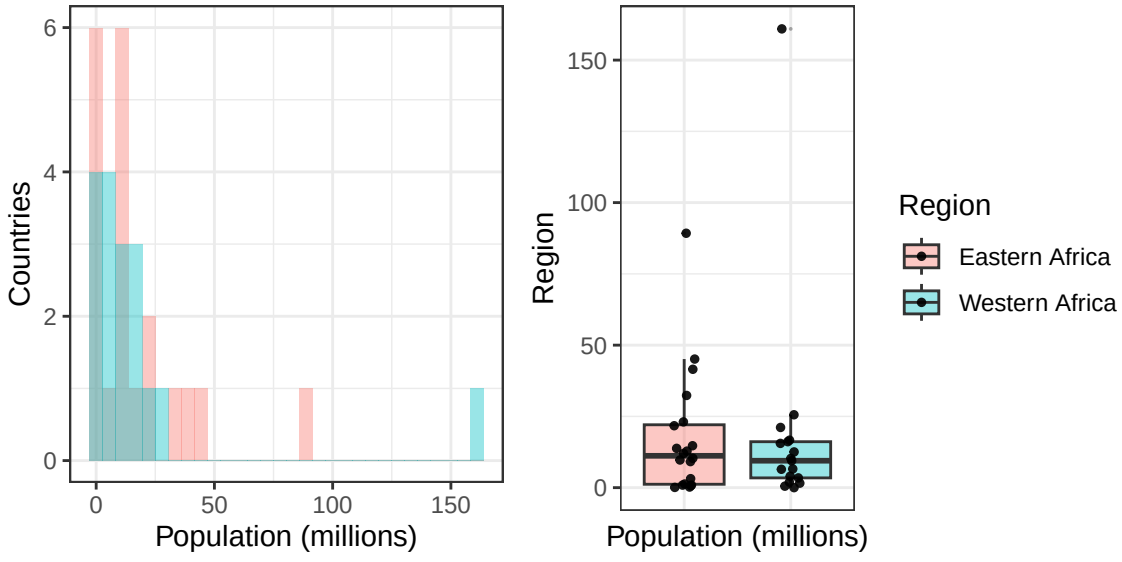
```
# Western Africa
pop1 <- countries[countries$Year == 2010 &
                    countries$Region == "Western Africa",
                  ]$`Population mid-year estimates (millions)`
round(c(summary(pop1), "SD" = sd(pop1)), 2)
```

```
##    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. Perform a formal $t$-test for the difference in population means between Western African and Eastern African countries.

```
##
##  Welch Two Sample t-test
##
## data:  west_african and east_african
## t = 0.12188, df = 24.688, p-value = 0.904
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.98061  22.49249
## sample estimates:
## mean of x mean of y
##  18.39294  17.13700
```

4. Perform a formal $z$-test for the difference in the proportions of the populations that are nurses or midwives in the US versus the UK in 2010.

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(us_nurses_midwives, uk_nurses_midwives) out of c(us_pop, uk_pop)
## X-squared = 57941, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.003585282 0.003637892
## sample estimates:
##      prop 1      prop 2
## 0.012504975 0.008893388
```

5. 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.