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 $\bar{Z}$?

$$\bar{Z} \sim \mathcal{N}(0, 1/n)$$

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 $\bar{Z}$?

It should be higher if ρ is positive because each observation shares some information with neighboring observations and therefore contributes less new information about the mean.

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

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, which is a linear combination of the multivariate Normal vector, so $\bar{X}$ is still Normal.

$E(\bar{X}) = 0$ by linearity of expectation.

$$\begin{aligned}
\text{Var}(\bar{X}) &= \sum_{i=1}^{n} \text{Var}(X_i/n) + \sum_{i,j \text{ s.t. } i \neq j} \text{Cov}(X_i/n, X_j/n) \\
&= \sum_{i=1}^{n} \text{Var}(X_i/n) + \sum_{i,j \text{ are neighbors}} \rho \sqrt{\text{Var}(X_i/n)\text{Var}(X_j/n)} \\
&= \text{Var}(X_1)/n + \sum_{i,j \text{ are neighbors}} \frac{\rho}{n^2} \sqrt{\text{Var}(X_i)\text{Var}(X_j)} \\
&= \text{Var}(X_1)/n + \frac{\rho(2n-2)}{n^2} \text{Var}(X_1)
\end{aligned}$$

where step 2 used the formula for correlation and step 4 used the fact that we have $2n-2$ covariance terms since all X_i have two neighbors except the first and last. Thus, using $\text{Var}(X_1) = 1$,

$$\bar{X} \sim \mathcal{N}\left(0, \frac{n + \rho(2n-2)}{n^2}\right)$$

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

We can see that the only place we used $\text{Var}(X_i)$ was in the last step, so we can just plug in σ^2 instead to get

$$\bar{X} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{n + \rho(2n-2)}{n^2}\right)\right)$$

5. What is the approximate distribution for large n?

As $n \to \infty$,

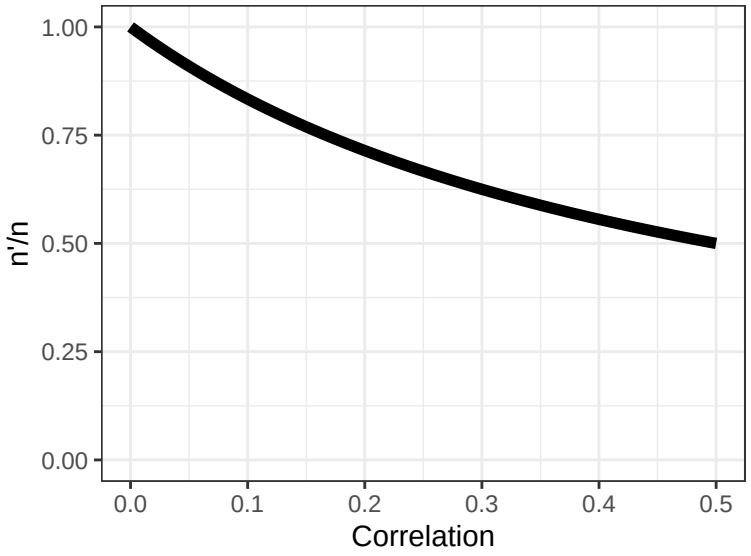
$$\frac{n + \rho(2n-2)}{n^2} \to \frac{n + 2\rho n}{n^2} \implies \bar{X} \dot{\sim} \mathcal{N}\left(0, \frac{1+2\rho}{n}\right)$$

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?

We need n' such that:

$$\frac{1}{n'} = \frac{1+2\rho}{n} \implies n' = \frac{n}{1+2\rho}$$

7. 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.

```
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
```

This simulation is generating many draws from three scenarios: (1) n independent observations, (2) n' independent observations, and (3) n dependent observations. For each draw, it computes the sample mean, stores it, and then finds the variance of the sample mean. As expected, the dependent n sample mean variance comes out very close to the independent n' sample mean variance but quite a bit higher than the independent n sample mean variance.

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.)

Let $\vec{x} = (-1, 1, -1, ..., -1)^T$. Then,

$$\mathbf{\Sigma}\vec{x} = \begin{bmatrix} -1+\rho \\ 1-2\rho \\ -(1-2\rho) \\ \vdots \\ -(1-2\rho) \\ 1-2\rho \\ -1+\rho \end{bmatrix} \implies \vec{x}^T\mathbf{\Sigma}\vec{x} = -2(-1+\rho) + \sum_{i=2}^{n-1}(1-2\rho)$$

Since $0.5 < \rho < 1$, $0 < -2(-1+\rho) < 1$ which means the first term is at most 1, but each $1-2\rho$ term is negative if $\rho > 0.5$, and we can have arbitrarily many of them. Therefore, we can choose an n large enough that the whole expression is negative, meaning the $\mathbf{\Sigma}$ matrix is invalid.

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 $V/n \xrightarrow{p} 1$.

$E(Z_i^2) = \text{Var}(Z_i) + (E(Z_i))^2 = 1$, so by the law of large numbers, $\frac{1}{n}\sum_{i=1}^n Z_i^2 \to E(Z_1^2) = 1$

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

Continuous mapping theorem

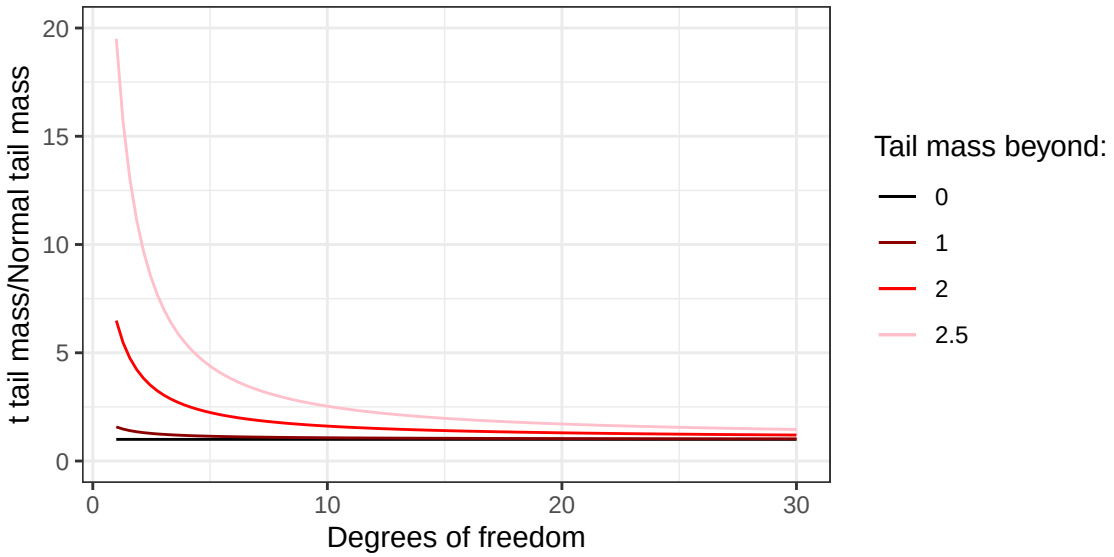
3. What tells us that if $Z \sim \mathcal{N}(0,1)$ and $\frac{1}{\sqrt{V/n}} \xrightarrow{p} 1$, $\frac{Z}{\sqrt{V/n}} \xrightarrow{d} \mathcal{N}(0,1)$

Slutsky's Theorem

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

$$T_n \xrightarrow{d} \mathcal{N}(0,1)$$

5. Do the centers or the tails converge faster?



The centers converge much faster. The plot is showing the ratio of the student-t and standard Normal distribution masses beyond a particular value. We can see that even at $n = 30$, usually considered a good sample size, the mass above 2.5 in a student-t distribution is about 30% larger than the mass above 2.5 for a standard Normal.

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

P-values generated from test statistics near 0 will be about the same for a student-t or a standard Normal distribution. However, p-values generated from large test statistics in a Normal approximation can significantly overstate the significance (be too low) relative 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)
```

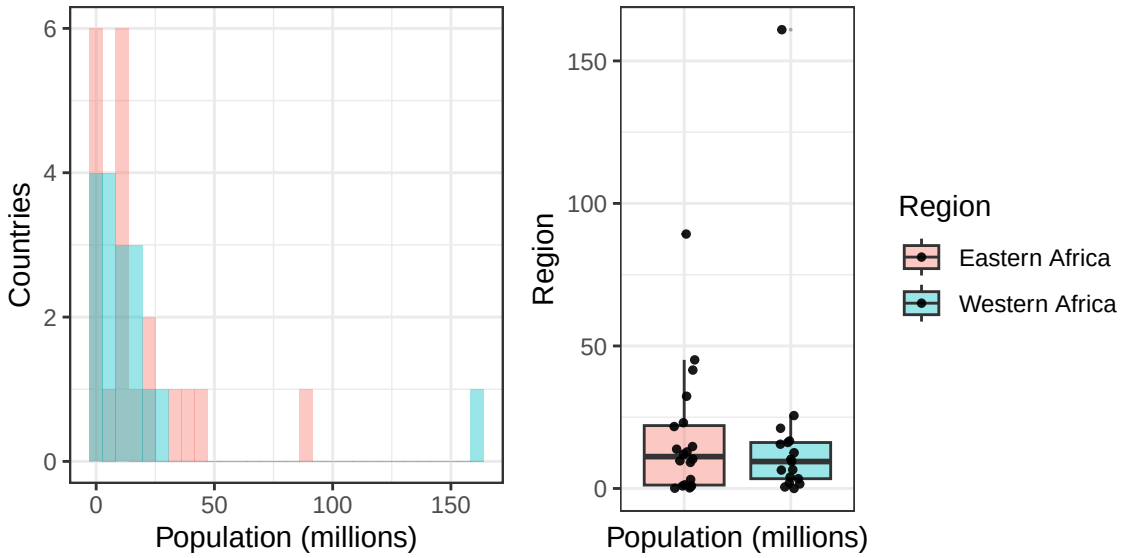
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD
##    0.01    3.42    9.45   18.39   16.12  160.95   37.50
```

```
# 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
```

The distributions are similar with close means, medians, minimums, and 1st quartiles. They both have considerable right skews with a means nearly double the medians, and their tails differ with Eastern African countries having a higher third quartile but lower maximum (Nigeria vs. Ethiopia).

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



Probably not: The means look about the same, and the standard deviations are large enough that even dividing by the square root of the sample size still yields a standard error of the mean larger than the difference in means.

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
```

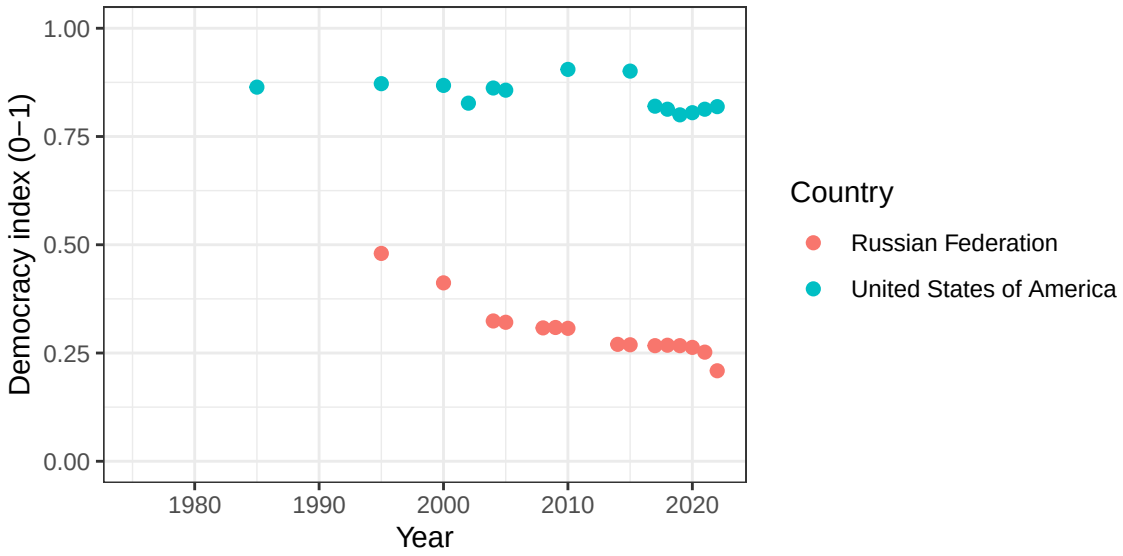
Let μ_0 be the mean population of countries in Western Africa and μ_1 be the mean population of countries in Eastern Africa. We are testing H_0: $\mu_0 = \mu_1$ vs H_a : the two means are different. We get a t-statistic of 0.12 with 24.7 degrees of freedom for a two-sided t-test, which corresponds to a p-value of $0.90 > 0.05$, so we fail to reject the null that there is a difference in mean populations. (The confidence interval for the differences in means is $(-20.0, 22.5)$, which includes 0, consistent with the t-test.)

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
```

Let p_{US} be the proportion of the population that is nurses or midwives in the US and p_{UK} be the equivalent proportion in the UK. We want to test $H_0 : p_{\text{US}} = p_{\text{UK}}$ vs. H_a : they are not equal. We get a z-statistic of $\sqrt{57941} = 240.7$ which gives a p-value less than 2.2×10^{-16}, so we reject the null and conclude that the proportion of the population that is nurses and midwives is significantly higher in the US.

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.



The United States' index has consistently been well above Russia's. The US index hovers around 0.85 while

Russia's has consistently declined. (The gaps are because I merged these scores with the UN data which were missing a few years.)