## Announcements

- Make sure to sign in on the google form (I send a list of what section questions are useful for what pset questions afterwards)
- Pset 4 due Friday 2/24

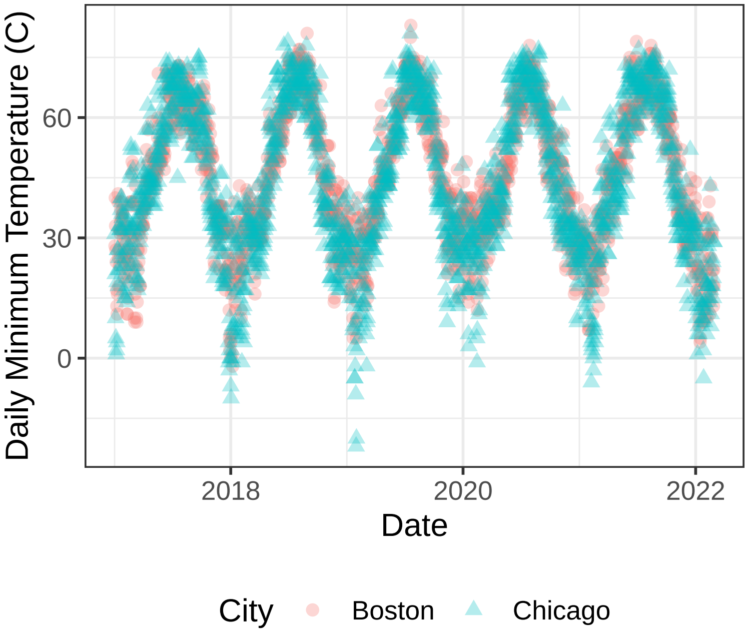
## Kahneman (Warm-up)

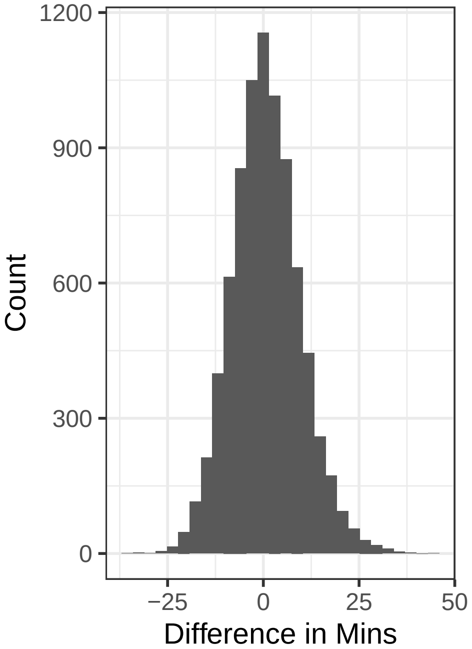
Without looking anything up, give 90% confidence intervals for the following quantities: (These are reproduced from Russo and Schoemaker 1989 but the answers are updated.)

| Statement | Value |
|---|---|
| Martin Luther King Jr.'s age at death | 39 years |
| Length of the Nile River | 4132 mi. |
| Number of countries that are members of OPEC | 13 |
| Number of books in the Catholic Old Testament | 46 |
| Diameter of the moon | 2159 mi. |
| Weight of an empty Boeing 747 | 412,300 lb. |
| Year in which Wolfgang Amadeus Mozart was born | 1756 |
| Gestation period (in days) of an Asian elephant | 680 days |
| Air distance from London to Tokyo | 5975 mi. |
| Deepest recorded point in the oceans | 35,814 ft. |
| Proportion of people taking this quiz in the room who got at least 8/10 | |

## Brr

The following questions deal with the 2000-2022 temperatures of Boston and Chicago available here.





In this problem, we'd like to determine whether daily minimum temperatures are significantly different between Boston and Chicago. To do this, we'll explore the student-$t$ distribution and confidence intervals. Our strategy will be to find the null distribution of some statistic assuming the true difference is 0 and then see how likely we are to have observed the crystallized version of that statistic under the null.

1. Let $Y_1, ..., Y_n \sim \mathcal{N}(0, \sigma^2)$ i.i.d. with $n = 8091$. Find the distribution of $\bar{Y}$ assuming $\sigma^2$ is known, and use that to give a standardized distribution for $\bar{Y}$.

From properties of the Normal, we know that $\bar{Y} \sim \mathcal{N}(0, \sigma^2/n)$, so

$$\frac{\bar{Y}}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

2. Let $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ be the sample variance. Then, by 10.4.3 in the Stat 110 book, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Show that the sample variance is independent of the sample mean by using facts about Multivariate Normals and the vector $(\bar{Y}, Y_1 - \bar{Y}, ..., Y_n - \bar{Y})$. (Hint: In a MVN vector, zero covariance implies independence. Also, a function of a random vector independent of a random variable is also independent of the random variable.)

The vector is Multivariate Normal because each element is a linear combination of Normals. Next, $\text{Cov}(\bar{Y}, Y_i - \bar{Y}) = \text{Cov}(\bar{Y}, Y_i) - \text{Var}(\bar{Y})$. $\bar{Y} = \frac{1}{n}\sum_{j=1}^n Y_j$, and $Y_i$ is independent of all the $Y_j$ except for when $j = i$, so $\text{Cov}(\bar{Y}, Y_i) = \text{Cov}(Y_i/n, Y_i) = \text{Var}(Y_i)/n = \sigma^2/n$. Also, $\text{Var}(\bar{Y}) = \sigma^2/n$, so $\text{Cov}(\bar{Y}, Y_i - \bar{Y}) = \sigma^2/n - \sigma^2/n = 0$, which means $\bar{Y}$ is independent of every $Y_i - \bar{Y}$. Since the function of a random vector that is independent of a random variable is independent of that random variable, $S^2$ as a function of the $Y_i - \bar{Y}$ is independent of $\bar{Y}$.

3. Using the results above, write a function of $\bar{Y}$ and $S^2$ that has the $t_{n-1}$ distribution (this is our pivot). Recall that the $t_n$ distribution is defined as $\frac{Z}{\sqrt{V/n}}$ where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_n$ are independent.

$$\frac{\bar{Y}}{\sqrt{S^2/n}} = \frac{\frac{\bar{Y}}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2} \cdot \frac{1}{n-1}}} \sim t_{n-1}$$

since the standard normal in the numerator is a function of only $\bar{Y}$ and is therefore independent of the $\chi^2_{n-1}$ in the denominator, which is a function of only the sample variance. Note that the formula on the left has no mean because we specified the mean to be 0. However, normally there would be a $\bar{Y} - \mu$ in the numerator.

4. In terms of a CDF $F$, determine $P\left(\frac{\bar{Y}}{\sqrt{S^2/n}} > \tau\right)$ for a fixed $\tau$. Describe what this probability means in the context of the problem.

Letting $F_{t_{n-1}}$ be the $t_{n-1}$ CDF, $P\left(\frac{\bar{Y}}{\sqrt{S^2/n}} > \tau\right) = 1 - F_{t_{n-1}}(\tau)$. In the context of the problem, this says that if the $Y_i$ actually have mean 0, the probability our $t$-statistic will be greater than some threshold $\tau$ is $1 - F_{t_{n-1}}(\tau)$. For example, with $n = 8091$, the probability the $t$-statistic will be greater than 3 is about 0.00135.

```
dim(temps_diff)
```

```
## [1] 8091    6
```

```
1-pt(3, 8090)
```

```
## [1] 0.00135401
```

5. Using the pivot, find a 95% confidence interval for 0 and interpret what this means.

$$P\left(Q_{t_{n-1}}(0.025) < \frac{\bar{Y}}{\sqrt{S^2/n}} < Q_{t_{n-1}}(0.975)\right) = 0.95$$

$$\implies P\left(Q_{t_{n-1}}(0.025)\sqrt{S^2/n} - \bar{Y} < 0 < Q_{t_{n-1}}(0.975)\sqrt{S^2/n} - \bar{Y}\right) = 0.95$$

Therefore, our 95% confidence interval is $\left[Q_{t_{n-1}}(0.025)\sqrt{S^2/n} - \bar{Y}, Q_{t_{n-1}}(0.975)\sqrt{S^2/n} - \bar{Y}\right]$. This means that under the null distribution, if we generated $n$ observations many times, we would capture 0 with this interval 95% of the time. (Note that if you were to do this with a mean $\mu$ you would have to multiply everything by $-1$ and flip the inequalities.)

6. Using the data, compute this interval.

```
n <- dim(temps_diff)[1]
lb = qt(0.025, n-1) * sqrt(var(temps_diff$diff)/n) - mean(temps_diff$diff)
ub = qt(0.975, n-1) * sqrt(var(temps_diff$diff)/n) - mean(temps_diff$diff)
c(lb, ub)
```

```
## [1] -0.9713260 -0.5763196
```

7. Based on your interval, comment on whether it seems likely that the data follows the stated distribution (i.e. that the true mean difference is 0).

The interval we calculated did not contain 0, but if the difference in temperatures actually had mean 0 there would be a 95% chance of generating data that captured 0. Therefore, we suspect that the data might not actually follow a distribution with mean 0.

8. Show that $t_n \xrightarrow{d} Z$ with $Z \sim \mathcal{N}(0,1)$ as $n \to \infty$. (Hint: write the denominator as a sum of squared random variables and apply asymptotic tools.)

We can write $t_n$ as

$$t_n = \frac{Z}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Z_i^2}}$$

where $Z, Z_1, ..., Z_n$ are i.i.d. Standard Normals and we use the same first $n$ Standard Normals for each $t_n$. $E(Z_i^2) = \text{Var}(Z_i) + E(Z_i)^2 = 1$, so by the law of large numbers $\frac{1}{n}\sum_{i=1}^{n} Z_i^2 \xrightarrow{p} 1$. The continuous mapping theorem then shows $\frac{1}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Z_i^2}} \to 1$. Finally, since $Z \to \mathcal{N}(0,1)$ trivially, Slutsky's gives $t_n \xrightarrow{d} \mathcal{N}(0,1)$.

9. How close is the $t_{n-1}$ interval above to an interval using the Standard Normal?

```
n <- dim(temps_diff)[1]
lb = qnorm(0.025) * sqrt(var(temps_diff$diff)/n) - mean(temps_diff$diff)
ub = qnorm(0.975) * sqrt(var(temps_diff$diff)/n) - mean(temps_diff$diff)
c(lb, ub)
```

```
## [1] -0.9712964 -0.5763491
```

Almost identical; $n$ is very large.

10. Now, assume we have $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2$ known. We have seen before that $\bar{Y}$ is the unbiased MLE for $\mu$. Does $\bar{Y}$ achieve the Cramér-Rao lower bound? (The Cramér-Rao lower bound is the reciprocal of the Fisher information for the dataset for $\mu$.)

$$\ell(\mu; \vec{y}) = \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2$$

$$\implies s(\mu; \vec{y}) = \sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma^2}\right)$$

$$\mathcal{I}_n(\mu) = \text{Var}(s(\mu; \vec{Y})) = \frac{1}{\sigma^4}\sum_{i=1}^{n} \text{Var}(Y_i) = \frac{n}{\sigma^2}$$

Therefore, the CRLB is $\mathcal{I}_n^{-1}(\mu) = \frac{\sigma^2}{n}$. This is the minimum variance of any unbiased estimator of $\mu$. Since $\text{Var}(\bar{Y}) = \sigma^2/n$, it achieves the CRLB.

## Uniformity

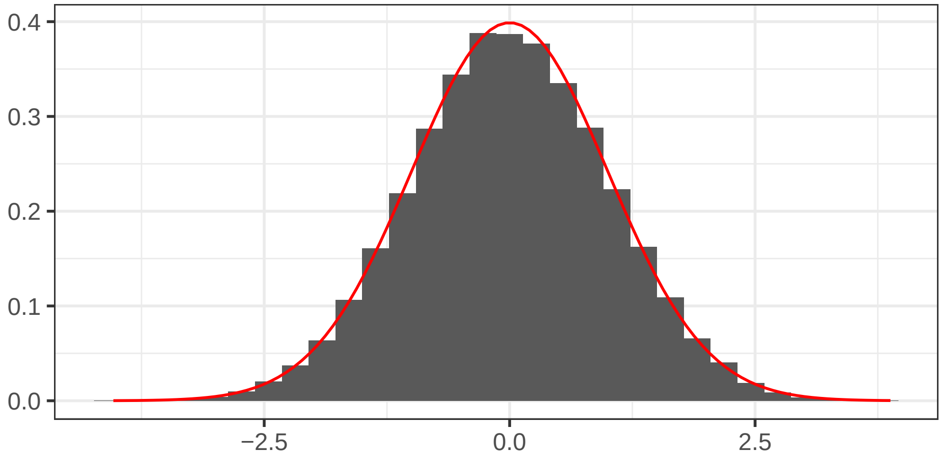
1. One fast way of computing Normal-like random variables is to take the sum of 12 $\text{Unif}(0,1)$ random variables and subtract 6. Find the expectation and variance of the resulting distribution and plot draws from it.

Let $U_i = \text{Unif}(0,1)$ for $i \in \{1, ..., 12\}$ and let $Z = (\sum_{i=1}^{12} U_i) - 6$.

$$E(Z) = \left(\sum_{i=1}^{12} 0.5\right) - 6 = 0$$

$$\text{Var}(Z) = \sum_{i=1}^{12} \text{Var}(U_i) = \sum_{i=1}^{12} \frac{1}{12} = 1$$

```
df = data.frame(x = replicate(100000, sum(runif(12)) - 6))
ggplot(df, aes(x = x)) +
  geom_histogram(bins=30, aes(y = after_stat(density))) +
  stat_function(fun = dnorm, args = c(0, 1), n = 100, col = "red") +
  theme_bw() +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank())
```



2. Now, suppose you have $n$ $\text{Unif}(\alpha - \beta/2, \alpha + \beta/2)$ i.i.d. random variables. Write the likelihood function for $\alpha$ and $\beta$.

$$L(\alpha, \beta; \vec{y}) = \prod_{i=1}^{n} \frac{I(\alpha - \beta/2 \leq y_i \leq \alpha + \beta/2)}{\beta} = \frac{I(\alpha - \beta/2 \leq y_{(1)})I(y_{(n)} \leq \alpha + \beta/2)}{\beta^n}$$

3. For the case of $\alpha = 10$ and $\beta = 5$, simulate $n = 100$ uniform random variables and use the `optim` function in R to estimate $\alpha$ and $\beta$ from starting guesses of $(20, 100)$ by maximizing the likelihood function. How do these compare to the true values?

```
set.seed(111)
alpha = 10
beta = 5
n <- 100
x <- runif(n, alpha - beta/2, alpha + beta/2)

loglikelihood <- function(params) {
  # Penalize lower and upper bounds close to the min and max
  return (- ((10^12)^((params[1] - params[2]/2) - min(x))) -
    ((10^12)^(max(x) - (params[1] + params[2]/2))) - n * params[2])
}

# Initial guess
params <- c(20, 100)

# Run the optimization
optimization <- optim(params, loglikelihood, method="L-BFGS-B",
                      control=list(fnscale=-1), lower=c(-Inf,0), upper=c(Inf,Inf))
optimization$par
```

```
## [1] 9.994105 4.885858
```

This turns out to be pretty hard to optimize. Our log likelihood only has a non-zero gradient with respect to $\beta$ and jumps to $-\infty$ once the indicators fail. Probably the best thing to do here is to use differentiable functions to very severely penalize $\alpha$ and $\beta$ that get close to the min and max. (On the pset, you can use the `optim` command in a similar way, but you don't need to worry about the lower and upper bounds or indicators.)

4. Find method of moments estimators for $\alpha$ and $\beta$ and see how well these perform on the simulated data.

$E(Y_i) = \alpha$, so $\bar{Y}$ is a method of moments estimator for $\alpha$. For a uniform random variable,

$$\text{Var}(Y_i) = E(Y_i^2) - E(Y_i)^2 = \frac{\beta^2}{12} \implies \hat{\beta} = \sqrt{12(\overline{Y^2} - \bar{Y}^2)}$$

```
c(mean(x), sqrt(12 * (mean(x^2) - mean(x)^2)))
```

```
## [1] 9.947620 4.751651
```

These perform pretty well on the single sample.

5. Another set of estimators is $\hat{\alpha} = Y_{(n/2+1/2)}$ and $\hat{\beta} = 2(Y_{(n/2+1/2)} - Y_{(1)})$. Describe the logic of these estimators and then find their biases and variances. Assume $n$ is odd. (Note that $U_{(k)} - U_{(j)} \sim \text{Beta}(k-j, n-(k-j)+1)$ for Standard Uniforms.)

The parameter $\alpha$ defines the center of the uniform distribution, so taking the median to estimate it makes sense. The parameter $\beta$ is the width of the uniform distribution, so taking the distance between the median and the smallest value and multiplying it by 2 is a way to find the approximate width.

By order statistics for uniform random variables and rescaling the uniform,

$$\frac{\hat{\alpha} - \alpha}{\beta} + \frac{1}{2} \sim \text{Beta}(n/2 + 1/2, n + 1 - (n/2 + 1/2)) = \text{Beta}(n/2 + 1/2, n/2 + 1/2)$$

which has mean $\frac{1}{2}$ and variance $\frac{1}{4(n+2)}$. Therefore, $E(\hat{\alpha}) = \beta\left(\frac{1}{2} - \frac{1}{2}\right) + \alpha = \alpha$, so it is unbiased. Likewise, $\text{Var}(\hat{\alpha}) = \frac{\beta^2}{4(n+2)}$. Using the note,

$$\frac{\hat{\beta}}{2\beta} \sim \text{Beta}(n/2 - 1/2, n - (n/2 - 1/2) + 1) = \text{Beta}(n/2 - 1/2, n/2 + 3/2)$$

with mean $\frac{n-1}{2(n+1)}$ and variance

$$\frac{\frac{n-1}{2(n+1)}\frac{n+3}{2(n+1)}}{n+2} = \frac{(n-1)(n+3)}{4(n+1)^2(n+2)}$$

Therefore,

$$E(\hat{\beta}) = 2\beta\left(\frac{n-1}{2(n+1)}\right)$$

which gives bias

$$\beta\left(\frac{n-1}{n+1}-1\right)$$

which means we are underestimating the spread as expected. Also,

$$\text{Var}(\hat{\beta})=(2\beta)^2\frac{(n-1)(n+3)}{4(n+1)^2(n+2)}$$

6. Make the $\hat{\beta}$ estimator unbiased.

Multiplying by the inverse of its biasing coefficient, we can make $\hat{\beta}$ unbiased by multiplying by

$$\frac{n+1}{n-1}$$

We can inspect the results above with a simulation:

```r
nsims <- 100000
n <- 11
alpha = 5
beta = 10
alpha_hat <- vector(length = nsims)
beta_hat <- vector(length = nsims)
for (i in 1:nsims) {
    x <- sort(runif(n, alpha - beta/2, alpha + beta/2))
    alpha_hat[i] <- median(x)
    beta_hat[i] <- 2*(median(x) - x[1])
}

c("True" = alpha, "Sample" = mean(alpha_hat))
```

```
##     True   Sample
## 5.000000 5.002801
```

```r
c("True" = beta, "Sample" = mean(beta_hat * (n+1)/(n-1))) # Unbiasing
```

```
##     True   Sample
## 10.00000 10.00618
```

```r
c("True" = beta^2/(4*(n+2)), "Sample" = var(alpha_hat))
```

```
##     True   Sample
## 1.923077 1.917702
```

```r
c("True" = (2*beta)^2 * (n-1) * (n+3) * 1/4 * 1/(n+1)^2 * 1/(n+2),
  "Sample" = var(beta_hat))
```

```
##     True   Sample
## 7.478632 7.425970
```