Discussion 4

1. Consider a large population which has true mean μ and true standard deviation σ. We take a sample of size 3 from this population, thinking of the sample as the RVs X_1, X_2, X_3 where X_i can be considered iid (independent identically distributed). We are interested in estimating μ.

   a. Consider the estimator $\hat{\mu_1} = X_1 + X_2 - X_3$. What is the mean of this estimator?
   b. Find the variance of $\hat{\mu_1}$.
   c. Consider the estimator $\hat{\mu_2} = \frac{X_1+X_2+X_3}{3}$. Is this estimator biased?
   d. Find the variance of $\hat{\mu_2}$.
   e. Now, consider the estimator $\hat{\mu_3} = \frac{X_1+2X_2+3X_3}{6}$. What is the mean of this estimator?
   f. Find the variance of $\hat{\mu_3}$.
   g. Which of these three estimators is preferable? Why?

2. A packing plant fills bags with cement. The weight X kg of a bag of cement can be modeled by a normal distribution with mean 50kg and standard deviation 0.7kg.

   a. Find $P(X > 51)$
   b. Three bags are selected randomly. Find the probability that at least two weigh more than 51kg.

3. Weights of female cats are well approximated by a normal distribution with mean 4.1kg and standard deviation of 0.6kg $X \sim N(4.1, 0.6^2)$.

   a. What proportion of female cats have weights between 3.7 and 4.4kg?
   b. A certain female cat has a weight that is 0.5 standard deviations above the mean. What proportion of female cats are heavier than this one?
   c. How heavy is a female cat whose weight is on the 80th percentile?
   d. A female cat is chosen at random. What is the probability that she weighs more than 4.5 kg?

4. Take a look at the `R`-code below. [**Note**: You can also find this as a separate .Rmd file on Canvas. Everything after a pound symbol (#) is a comment, and NOT part of the `R` code.]

   a. Walk through it, one line at a time, and try to make sense of it.
   b. Run all the code, **EXCEPT FOR THE LAST LINE**. This should generate a bunch of QQ-plots. Half of them are from a normal distribution, the other half an exponential distribution. Can you tell which correspond to samples from the normal distribution, and which correspond to samples from the exponential distribution?
   c. Repeat with `sample_size <- 100`. Does this change anything?

```
library(tidyverse)
library(distributions3)

## Create two random variables -- one Normal and one Exponential
X <- Normal(mu = 1, sigma = 1)
Y <- Exponential(rate = 1)
```

```r
sample_size <- 10

## Create a vector of i's to use for scrambling
is <- sample(1:16)

normal_samples <- data.frame(i = is[1:8]) %>%
  mutate(distribution = "Normal",                    # Create column that just says "Normal".
                                                     # For book keeping.
         sample = map(i, random,                     # map runs the function "random" with
                      d = X, n = sample_size)) %>%   # arguments d = X, and n = sample_size
                                                     # for each i. I.e. it creates 8 random
                                                     # samples of size sample_size from the
                                                     # distribution X.
                                                     # The result is this weird list thing...
  unnest_longer(col = sample) # This get's rid of the list and simply gives us a
                              # data.frame that we can work with...

exponential_samples <- data.frame(i = is[9:16]) %>%
  mutate(distribution = "Exponential",
         sample = map(i, random, d = Y, n = sample_size)) %>%
  unnest_longer(col = sample)

# Here we just stack the two data.frame on top of each other (i.e. bind the rows together)
all_samples <- bind_rows(
  normal_samples,
  exponential_samples
)

# Create QQ-plots
ggplot(all_samples,
       aes(sample = sample)) +
  geom_qq() +
  facet_wrap(~i) +
  geom_abline(aes(slope = sd(sample), intercept = mean(sample)))

## To reveal the truth behind the i's, run this line:
all_samples %>% select(i, distribution) %>% unique()
```