Discussion 4 - Solution

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?
   Solution: $E(\hat{\mu_1}) = \mu$

   b. Find the variance of $\hat{\mu_1}$.
   Solution: $\text{Var}(\hat{\mu_1}) = 3 \cdot \sigma^2$

   c. Consider the estimator $\hat{\mu_2} = \frac{X_1+X_2+X_3}{3}$. What is the mean of this estimator?
   Solution: $E(\hat{\mu_2}) = \mu$

   d. Find the variance of $\hat{\mu_2}$.
   Solution: $\text{Var}(\hat{\mu_2}) = \frac{\sigma^2}{3}$

   e. Now, consider the estimator $\hat{\mu_3} = \frac{X_1+2X_2+3X_3}{6}$. What is the mean of this estimator?
   Solution: $E(\hat{\mu_3}) = \mu$

   f. Find the variance of $\hat{\mu_3}$.
   Solution: $\text{Var}(\hat{\mu_3}) = \frac{7}{18}\sigma^2$

   g. Which of these three estimators is preferable? Why?
   Solution: all unbiased, so on average, all will find the truth. However, μ_2 has the smallest variance.

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 $P(X > 51)$. Draw a sketch that visually indicates what this probability is. (I.e. normal curve, annotated with mean, SD, cut-off, shaded area)
   Solution:

```
library(distributions3)
```

```
##
## Attaching package: 'distributions3'

## The following objects are masked from 'package:stats':
##
##     Gamma, quantile

## The following object is masked from 'package:grDevices':
##
##     pdf
```

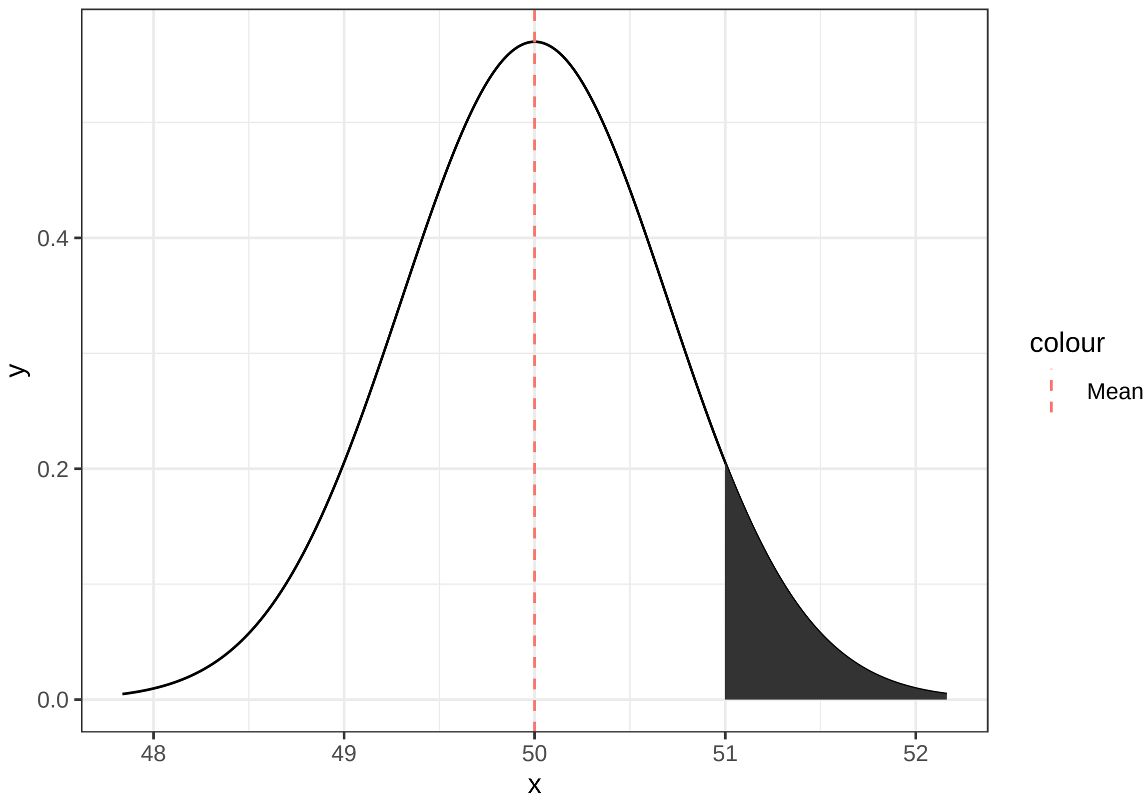
```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
X <- Normal(mu = 50, sigma = 0.7)

1 - cdf(X, 51)
```

```
## [1] 0.07656373
```



b. Three bags are selected randomly. Find the probability that at least two weigh more than 51kg.
Solution: Since the bags are independent and sampled from the same distribution, each has the same chance to weigh over the threshold. So this is asking about a binomial distribution. The probability of "success" (a bag weighing over 51kg) was found in part (a) as 0.077. So letting $M \sim \text{Bin}(3, 0.077)$, we want $P(M \geq 2) = 1 - P(M \leq 1)$. Use the binomial pmf or R.
In R, we can do either:

```r
p <- 1-cdf(X, 51)
Y <- Binomial(size = 3, p = p)

1- cdf(Y,1)
```

```
## [1] 0.01668838
```

or

```
pmf(Y, 2) + pmf(Y, 3)
```

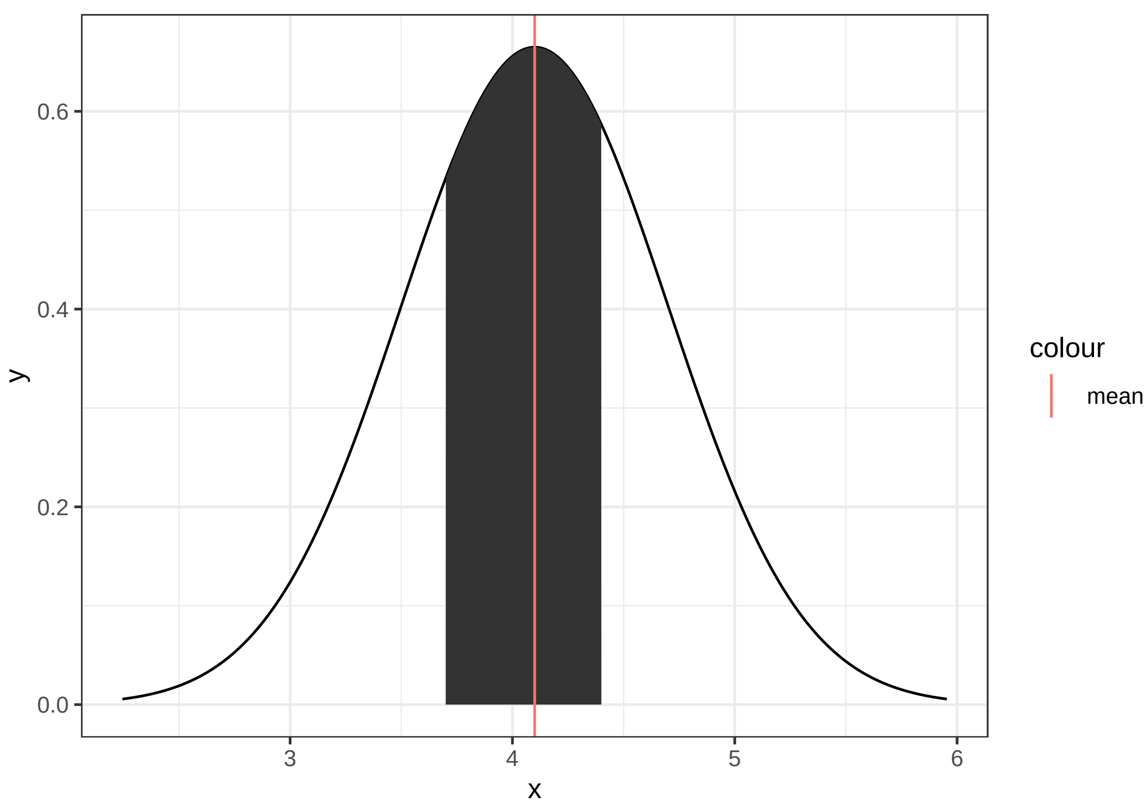
```
## [1] 0.01668838
```

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? Draw a sketch that visually indicates what this probability is. (I.e. normal curve, annotated with mean, SD, cut-off, shaded area)
    Solution:

```
X <- Normal(mu = 4.1, sigma = 0.6)
cdf(X, 4.4) - cdf(X, 3.7)
```

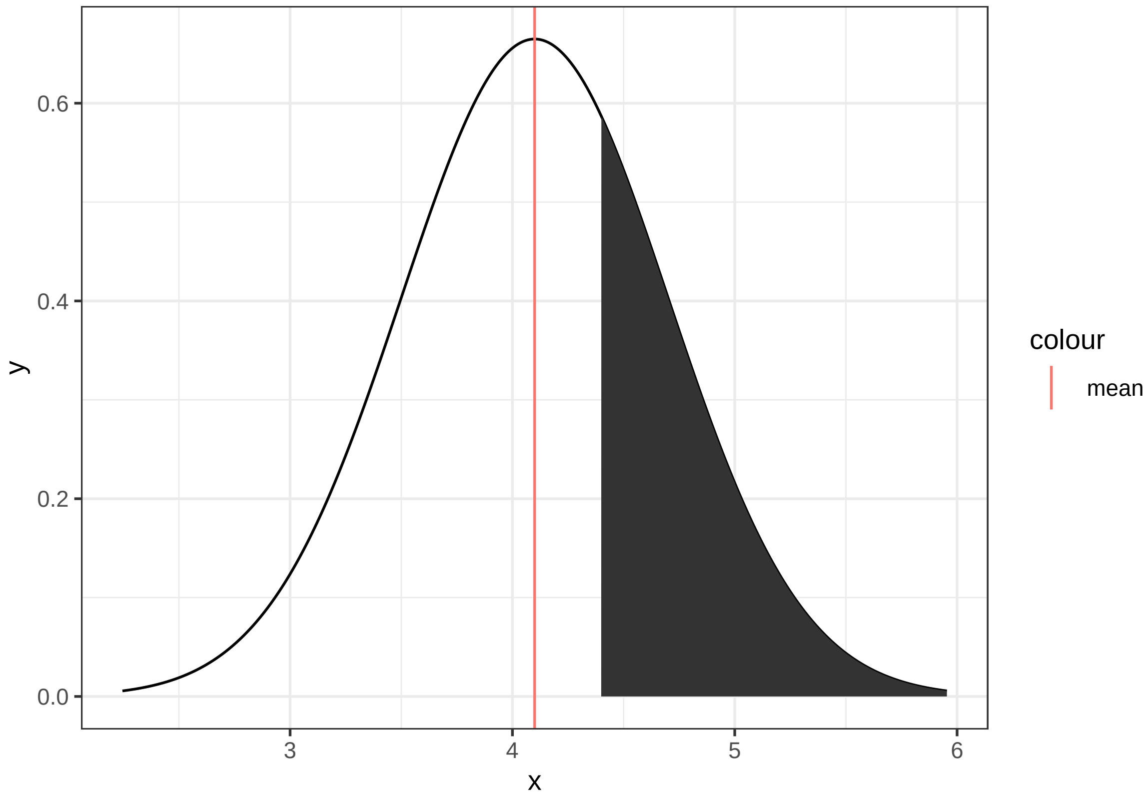
```
## [1] 0.4389699
```



    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? Draw a sketch that visually indicates what this probability is. (I.e. normal curve, annotated with mean, SD, cut-off, shaded area)
    Solution

```r
1 - cdf(X, 4.1 + 0.5*0.6)
```

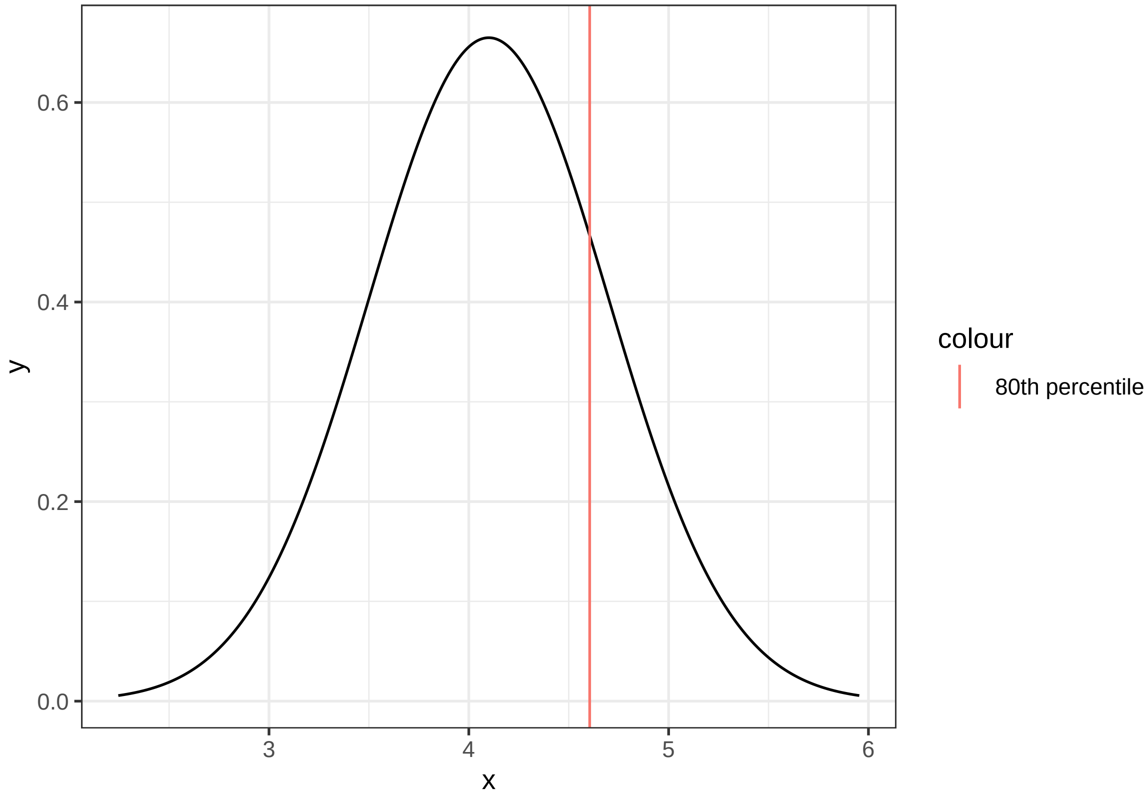
```
## [1] 0.3085375
```



c. How heavy is a female cat whose weight is on the 80th percentile? Draw a sketch that visually indicates what value we are looking for. (I.e. normal curve, annotated with mean, SD, cut-off, shaded area)
   Solution

```r
quantile(X, 0.8)
```

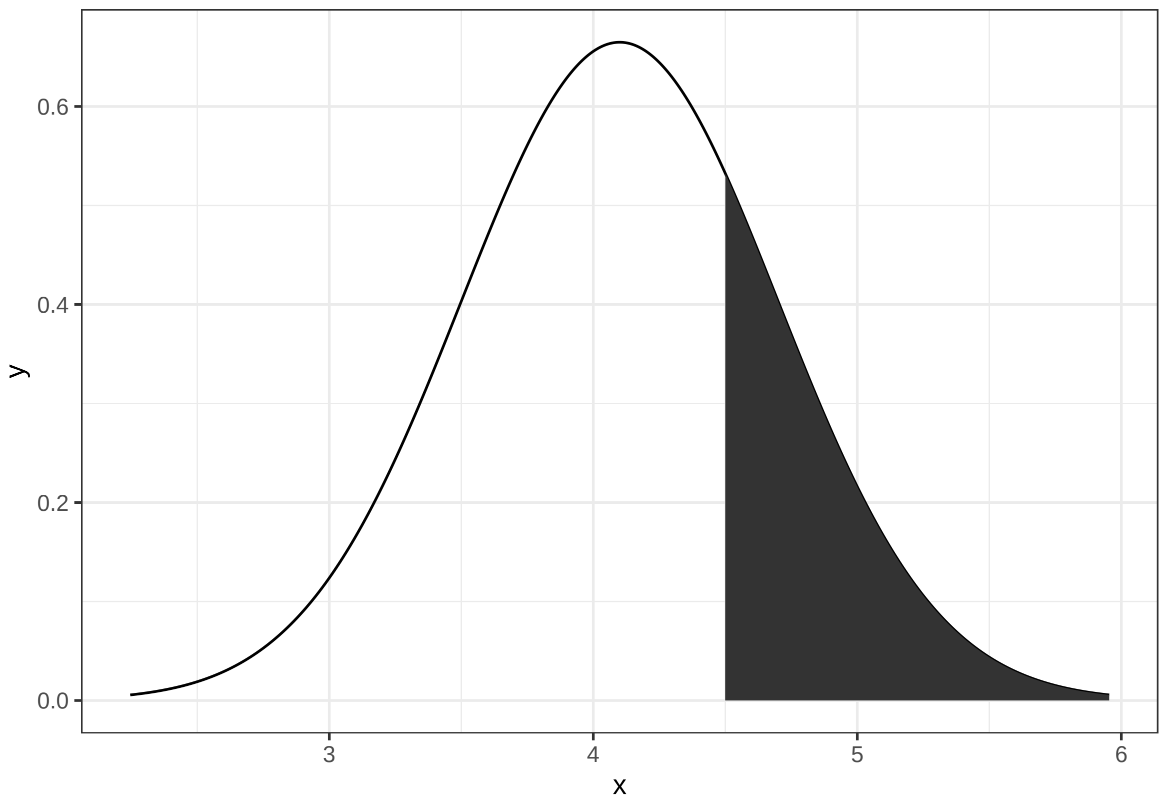
```
## [1] 4.604973
```



d. A female cat is chosen at random. What is the probability that she weighs more than 4.5 kg? Draw a sketch that visually indicates what this probability is. (I.e. normal curve, annotated with mean, SD, cut-off, shaded area)
Solution

```
1-cdf(X, 4.5)
```

```
## [1] 0.2524925
```

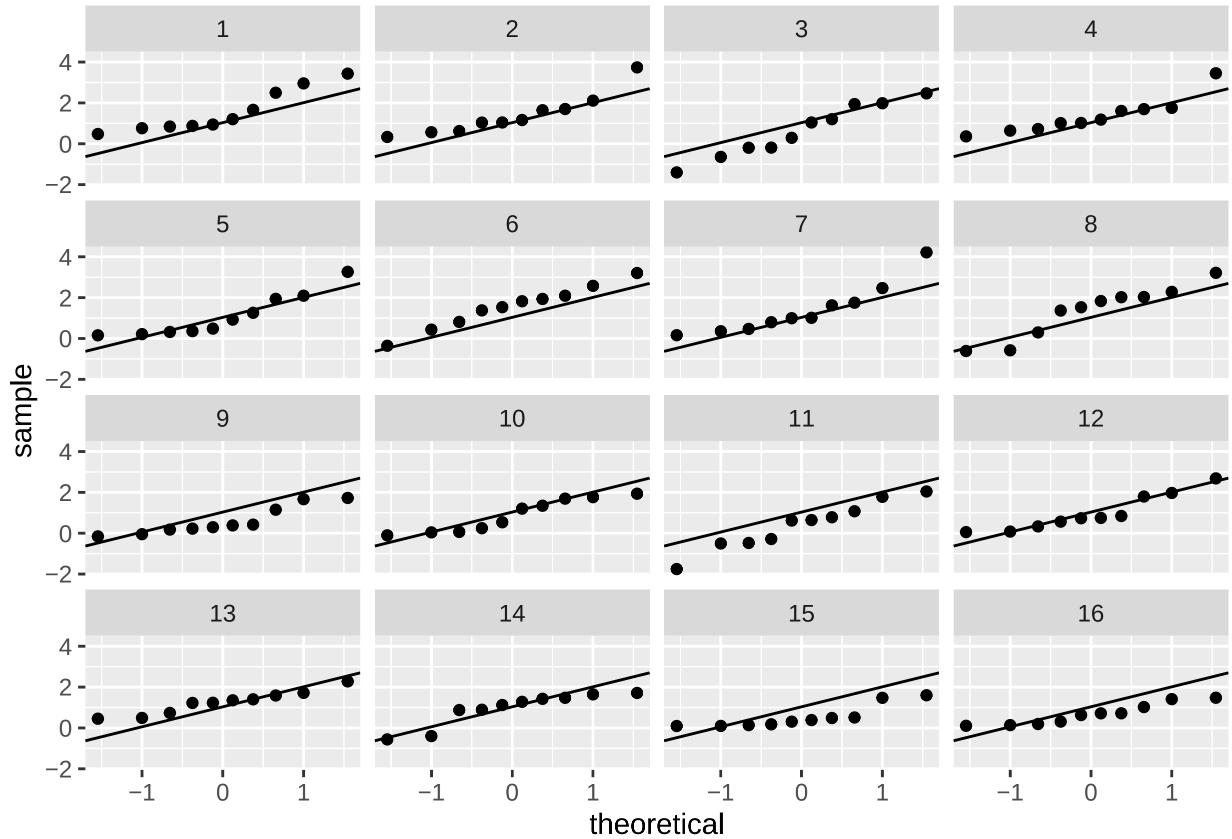


4. Take a look at the R-code below. [**Note**: everything after a pound symbol (#) is a comment, and NOT part of the R code.]

   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?
   Solution: I can maybe point out a few of the samples that come from the exponential distribution, but definitely not all.
   c. Repeat with `sample_size <- 100`. Does this change anything?
   Solution: Now I can point out all the exponentials.
   d. What does this tell you about the use of QQ-plots to determine normality of data?
   Solution: It sucks for small sample sizes.

```
library(tidyverse)
library(distributions3)

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

sample_size <- 10

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

```r
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)))
```




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

```
## # A tibble: 16 x 2
##        i distribution
##    <int> <chr>
##  1     1 Exponential
##  2     2 Exponential
##  3     4 Exponential
##  4     5 Exponential
##  5     7 Exponential
##  6    12 Exponential
##  7    15 Exponential
##  8    16 Exponential
##  9     3 Normal
## 10     6 Normal
## 11     8 Normal
## 12     9 Normal
## 13    10 Normal
## 14    11 Normal
## 15    13 Normal
## 16    14 Normal
```