Check-in 6

Jack Rellamas

Remember, **follow the instructions below and use R Markdown to create a pdf document with your code and answers to the following questions on Gradescope.** You may find a template file by clicking "Code" in the top right corner of this page.

1. The following code downloads two sequences of 256 digits. One of them was generated by asking ChatGPT to generate to "sample 1000 random digits from 0 to 9 with replacement" (I only got 256). The other was generated using the R command `sample(0:9, 256, replace = T)`. Which is from R and which is from ChatGPT? Explain and justify your answer.

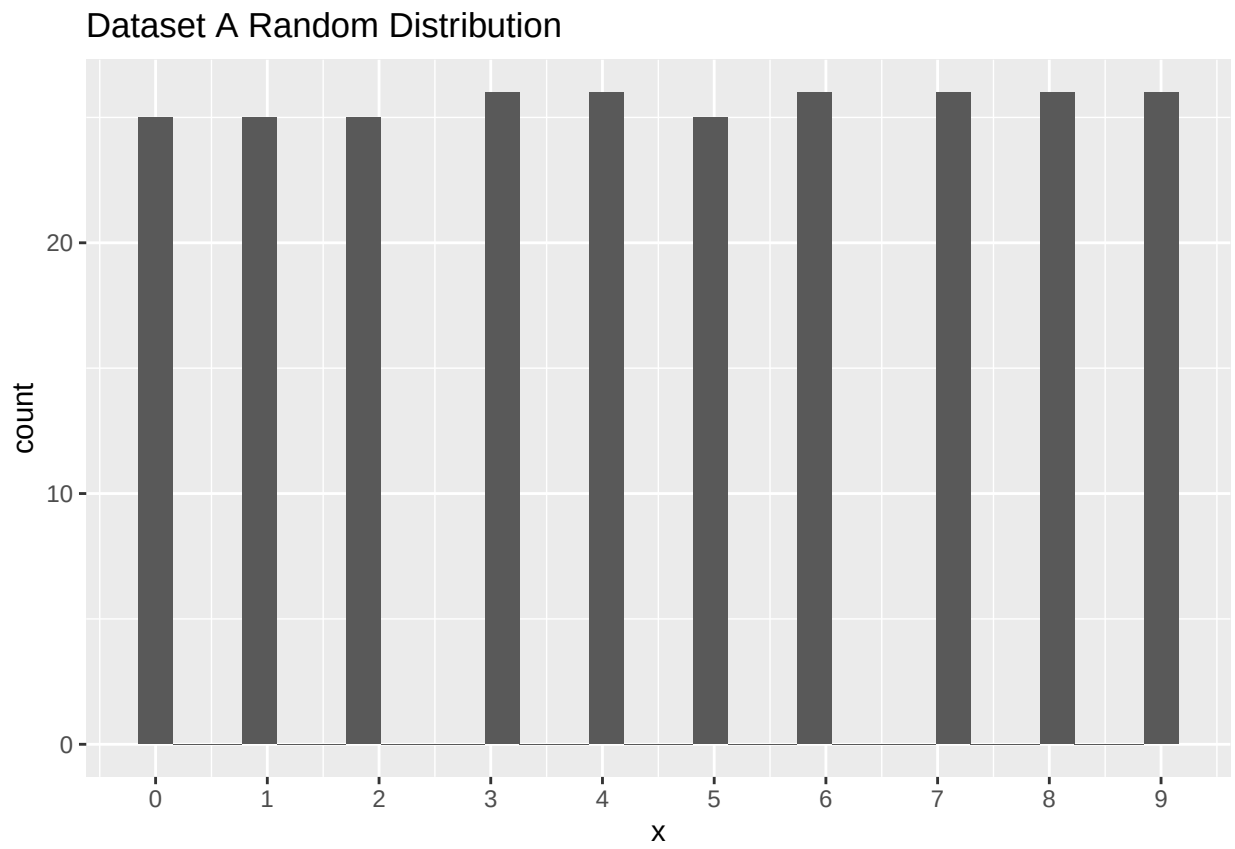
From looking at histograms of Datasets A and B, it is clear to see that Dataset A has a distribution that is much more uniform than Dataset B. However, both have the same number of random samples. So, when I used the same command as provided, with the same sample size of 256, I got a graph (Graph C) that was more randomly distributed. The random distribution looks closer to that of Dataset B. I made another graph, Graph D, with n = 1,000 and the distribution was still not as perfect as Dataset A. Finally I made Graph E with n = 10,000, and finally the distribution looked like Dataset A's distribution. We can conclude that Dataset A was made with ChatGPT and Dataset B was made with the given command. Dataset A's distribution is much too uniform for a relatively small sample size.

```
A <- read.csv("https://math167r-s24.github.io/static/digits-a.csv")
B <- read.csv("https://math167r-s24.github.io/static/digits-b.csv")
```

```
library(tidyverse)
```

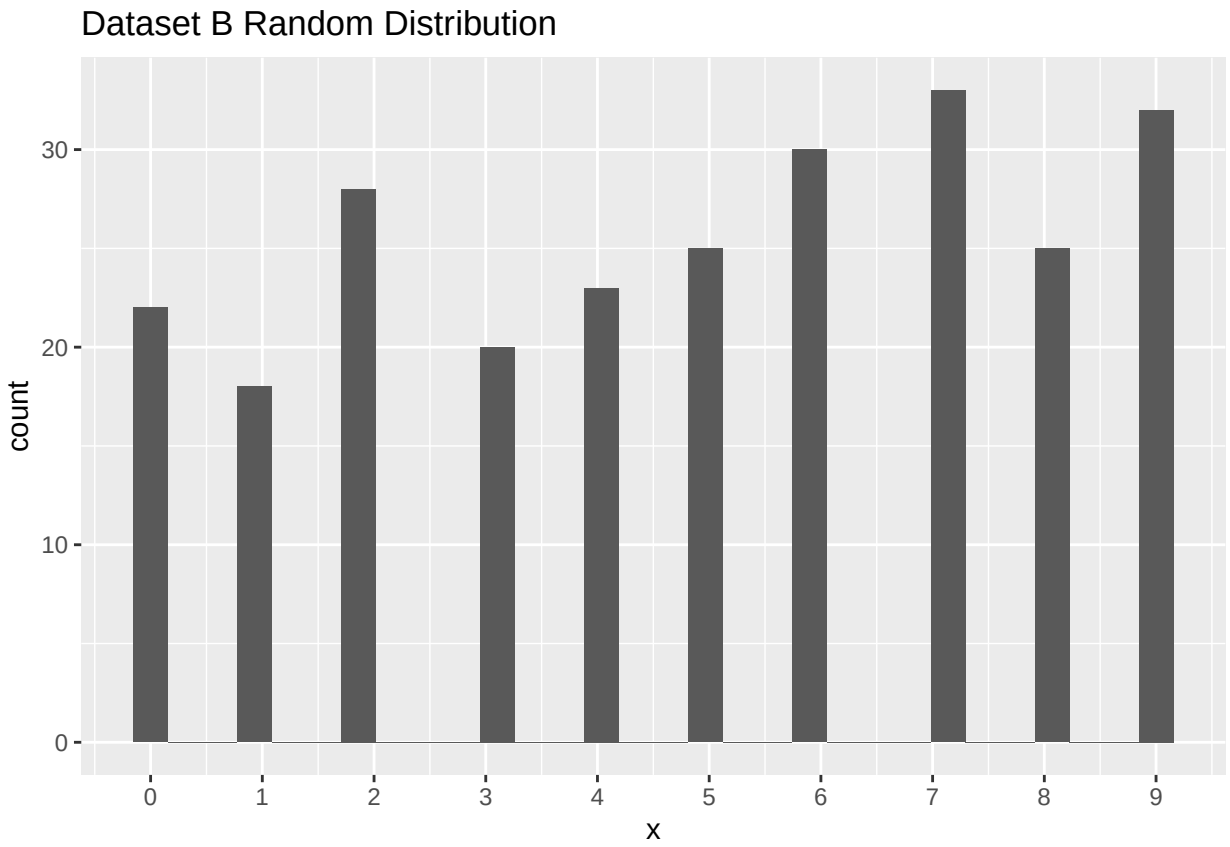
```
# Plotting Histogram for Dataset A
ggplot(data = A, aes(x = x)) +
  geom_histogram() +
  scale_x_continuous(breaks = 0:10) +
  ggtitle("Dataset A Random Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



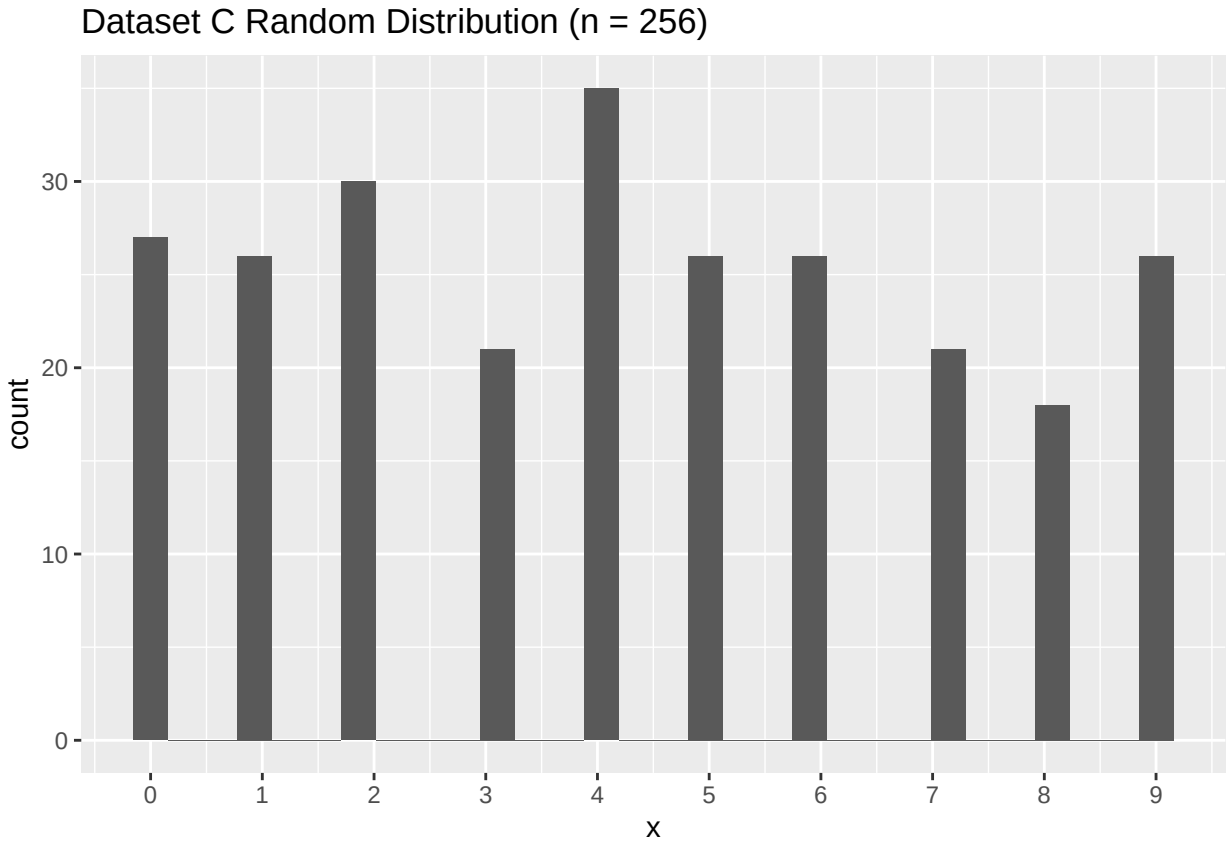

```
# Plotting Histogram for Dataset B
ggplot(data = B, aes(x = x)) +
  geom_histogram() +
  scale_x_continuous(breaks = 0:10) +
  ggtitle("Dataset B Random Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



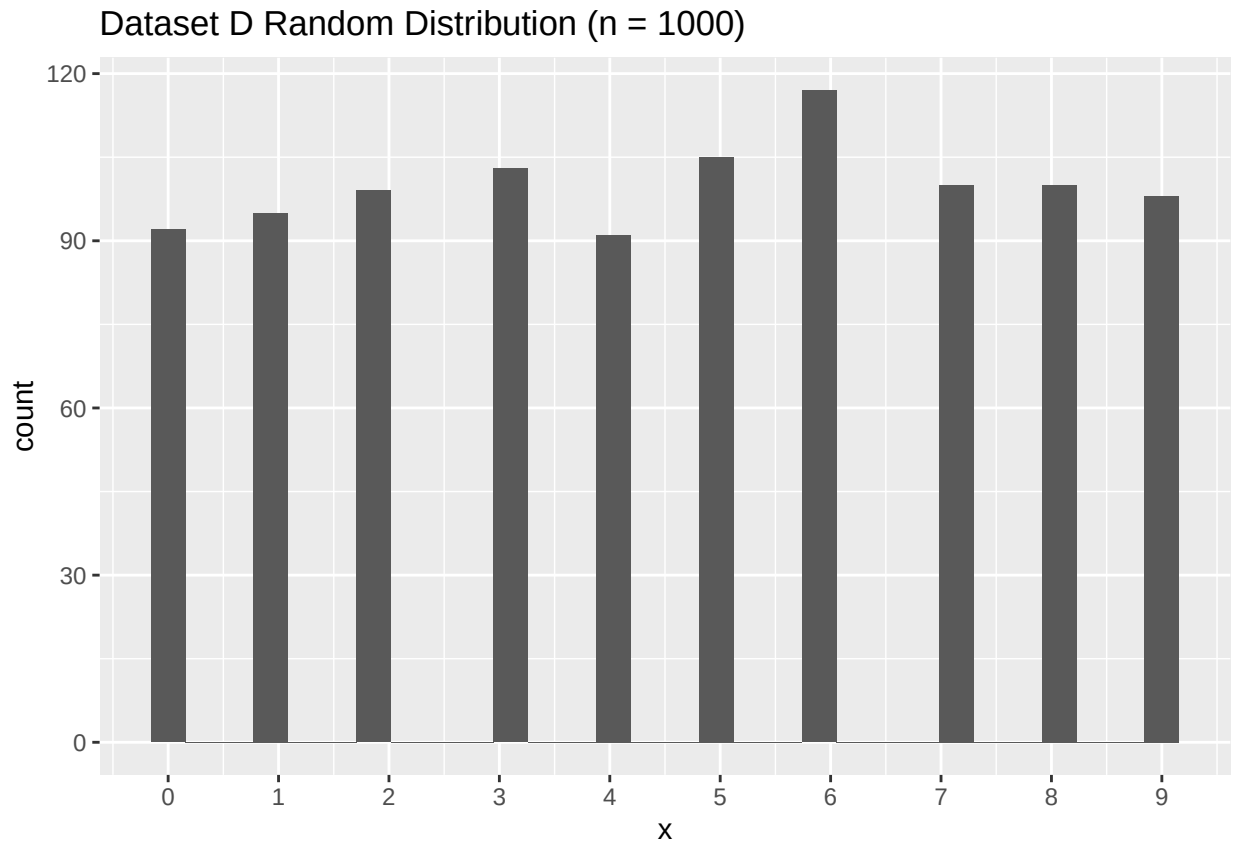

```
# Using the same command as provided.
set.seed(2000)
C <- sample(0:9, 256, replace = T)
C <- data.frame(C)
ggplot(data = C, aes(x = C)) +
  geom_histogram() +
  scale_x_continuous(breaks = 0:10) +
  xlab("x") +
  ggtitle("Dataset C Random Distribution (n = 256)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



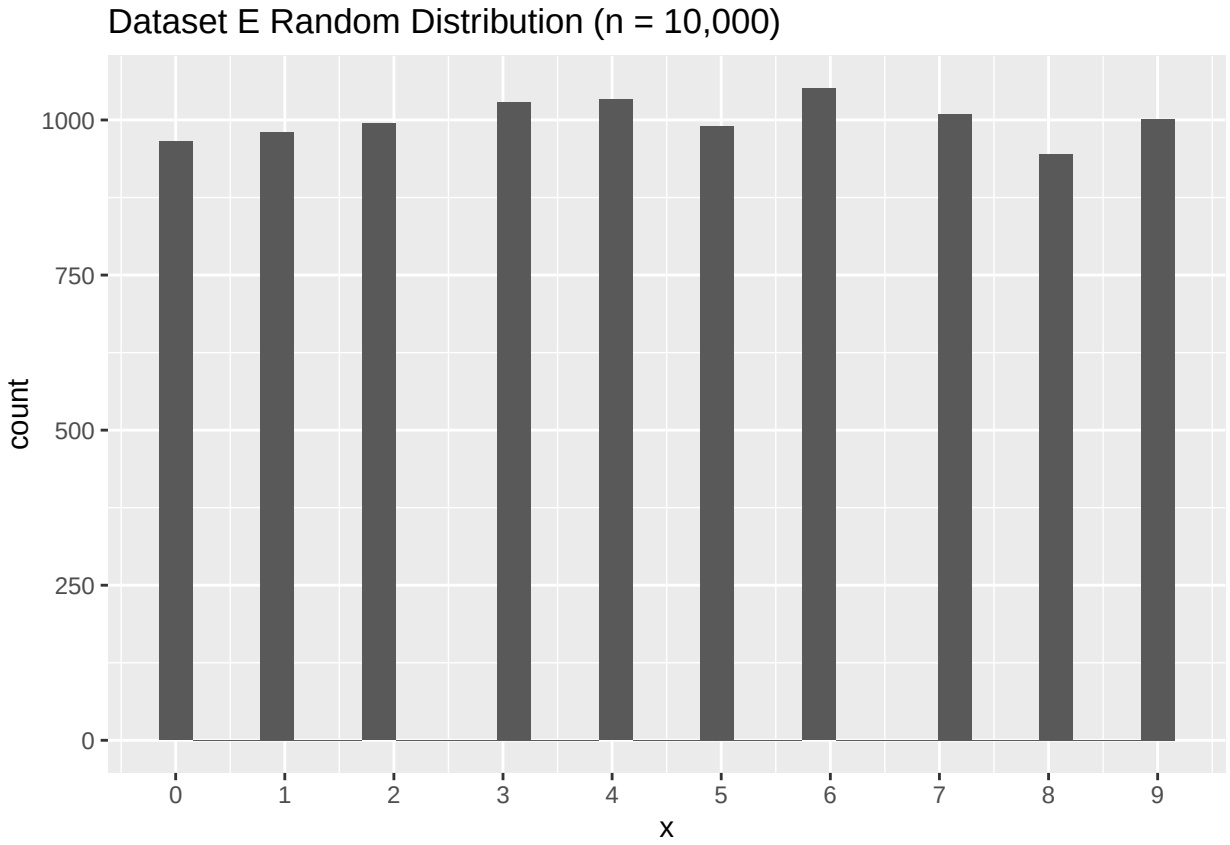

```
# Using the same command as provided.
set.seed(3000)
D <- sample(0:9, 1000, replace = T)
D <- data.frame(D)
ggplot(data = D, aes(x = D)) +
  geom_histogram() +
  scale_x_continuous(breaks = 0:10) +
  xlab("x") +
  ggtitle("Dataset D Random Distribution (n = 1000)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Using the same command as provided.
set.seed(4000)
E <- sample(0:9, 10000, replace = T)
E <- data.frame(E)
ggplot(data = E, aes(x = E)) +
  geom_histogram() +
  scale_x_continuous(breaks = 0:10) +
  xlab("x") +
  ggtitle("Dataset E Random Distribution (n = 10,000)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Dataset E Random Distribution (n = 10,000)
count
1000
750
500
250
0
0
1
2
3
4
5
6
7
8
9
x