Practice with functions and graphs to analyze distributions

Kaitlyn Watson-Group 9

M4 ICA3

Body temperature data

A recent paper, Decreasing human body temperature in the United States since the Industrial Revolution, presented evidence that human body temperatures in the United States have been decreasing over the past one hundred or so years. (Many scientists dispute the conclusions of the paper.) One of the data sets in the paper is taken from the NHANES (National Health and Nutrition Examination Study), and is available in the file `NHANES_processed.csv`. There are many variables in the data, but our focus will be on the variable `temp` that provides resting oral body temperatures.

(1) Draw a histogram of the body temperature variable

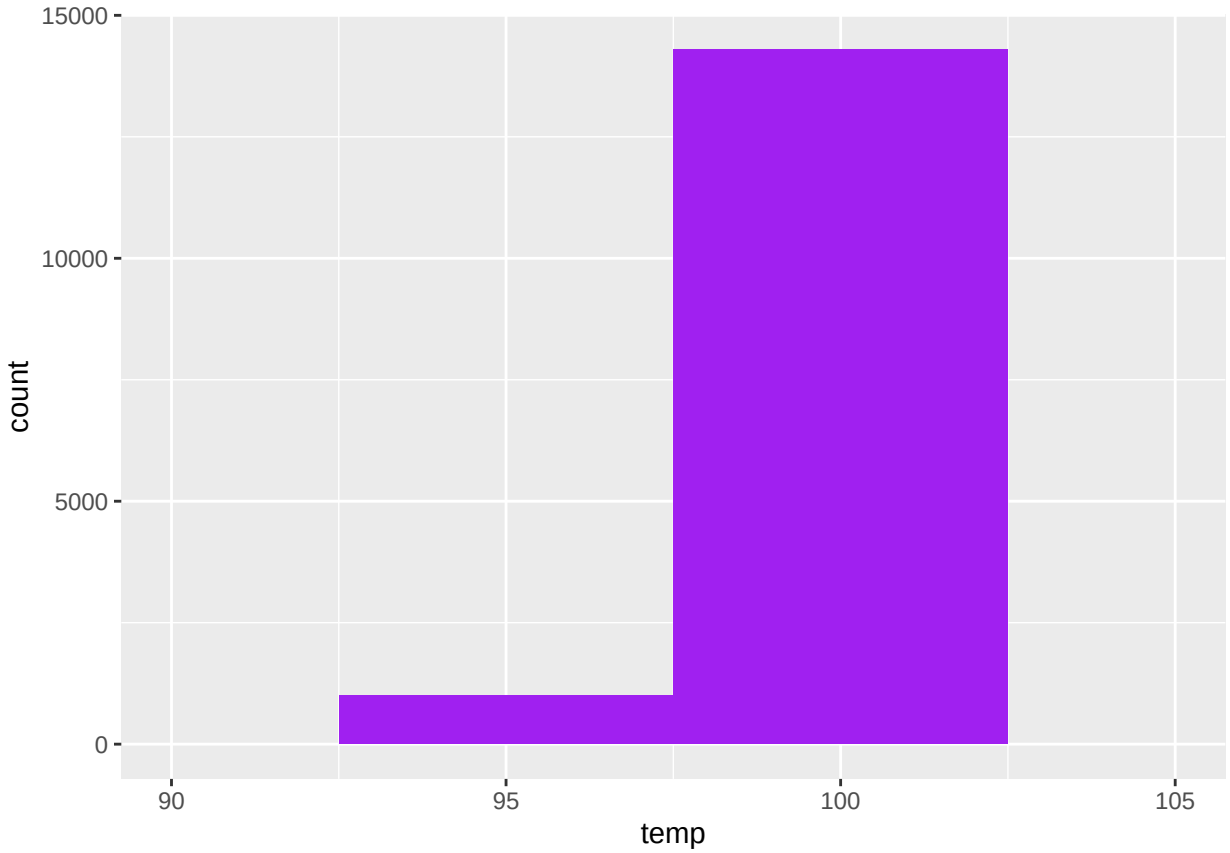
```
NHANES<-read.csv("NHANES_processed.csv")
head(NHANES)
```

```
  X  V1               study_ID sample_weights temp time_HR  race    sex age
1 1 Inf NH_805520010000142711 1           1645 98.4      NA white female  39
2 2 Inf NH_805520010000242711 1           3318 98.0      NA white   male  61
3 3 Inf NH_805520010000342711 1           2280 99.0      NA black female  25
4 4 Inf NH_805520010000442711 1           1574 98.4      NA black female  24
5 7 Inf NH_805520010000742711 1           1627 98.2      NA black female  35
6 8 Inf NH_805520010000842711 2          14604 97.6      NA white female  35
  year_of_birth  exam_date exam_year exammonth exam_findings exam_ICD exam_ICD2
1          1931 1971-05-20      1971         5             2     9999      9999
2          1909 1971-05-22      1971         5             2     9999      9999
3          1946 1971-04-28      1971         4             2      401      9999
4          1946 1971-05-12      1971         5             2     9999      9999
5          1935 1971-05-19      1971         5             2     9999      9999
6          1935 1971-04-28      1971         4             2      401      9999
     region birth_cohort head_eyes_ears_nose_findings thyroid_findings
1 NORTHEAST        1930s                        FALSE             TRUE
2 NORTHEAST        1900s                         TRUE             TRUE
3 NORTHEAST        1940s                         TRUE             TRUE
4 NORTHEAST        1940s                         TRUE             TRUE
5 NORTHEAST        1930s                         TRUE             TRUE
6 NORTHEAST        1930s                         TRUE             TRUE
  chest_findings cardiovascular_findings abdominal_findings
1           TRUE                    TRUE              FALSE
2           TRUE                    TRUE               TRUE
3           TRUE                   FALSE               TRUE
4           TRUE                    TRUE               TRUE
5           TRUE                    TRUE               TRUE
```

```
6           TRUE                    TRUE              TRUE
  musculosceletal_findings neurological_findings skin_findings general_findings
1                     TRUE                  TRUE          TRUE             TRUE
2                     TRUE                  TRUE          TRUE             TRUE
3                     TRUE                  TRUE          TRUE             TRUE
4                     TRUE                  TRUE          TRUE             TRUE
5                     TRUE                  TRUE          TRUE             TRUE
6                     TRUE                  TRUE          TRUE             TRUE
  no_findings thyroid weight_KG height_CM      BMI   temp_C    bmi_adj
1       FALSE       1     61.80    164.36 22.87685 36.88889 -0.9879365
2        TRUE       1     77.11    168.76 27.07520 36.66667  2.8025235
3       FALSE       1    107.64    163.16 40.43396 37.22222 16.6804189
4        TRUE       3     62.60    162.36 23.74741 36.88889  0.0680290
5        TRUE       1     94.46    174.26 31.10660 36.77778  6.3240588
6        TRUE       1     57.72    163.96 21.47091 36.44444 -2.3567922
  height_norm weight_norm
1    9.257842    2.886066
2   13.657842   18.196066
3    8.057842   48.726066
4    7.257842    3.686066
5   19.157842   35.546066
6    8.857842   -1.193934
```

```
NHANES %>%
  select(temp) %>%
  ggplot(mapping=aes(temp)) +
    geom_histogram(binwidth=5, fill="purple", position="dodge") +
      xlim(c(90,105))
```

```
Warning: Removed 2 rows containing missing values (geom_bar).
```

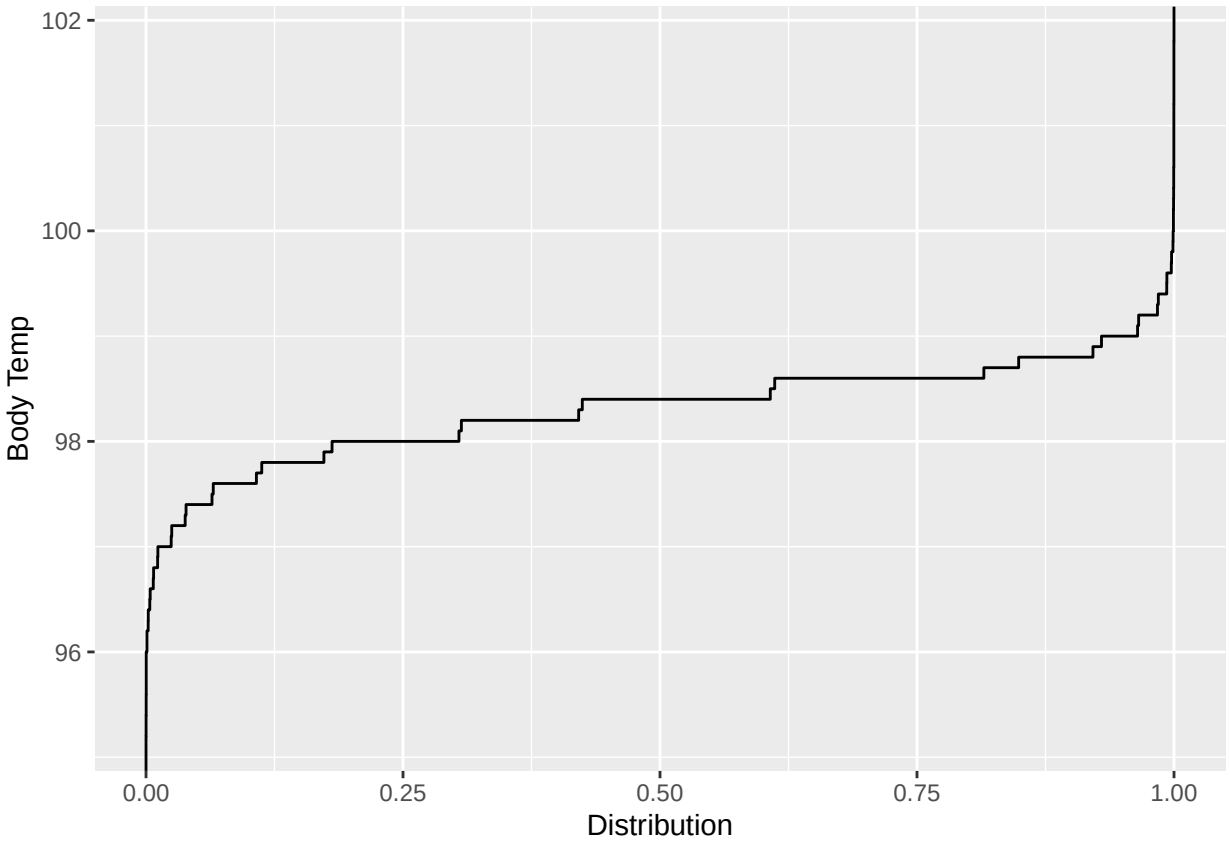


The *empirical cumulative distribution function* (ecdf) gives the proportion of data values at or below a particular value. In the case of the body temperature data, ecdf(98.1), for example, gives the proportion of temperatures in the data set below 98.1 degrees. A few minutes of thought reveals that

- ecdf(x) = 0 if x is less than the minimum temperature in the data set;
- ecdf(x) = 1 if x is greater than or equal to the maximum temperature in the data set;
- ecdf(x) is a non-decreasing function of x.

(2) Use the `stat_ecdf` function in `ggplot2` to draw an "empirical cumulative distribution function" for the body temperature variable.

```
NHANES %>%
  select(temp) %>%
  ggplot(mapping=aes(temp)) +
    stat_ecdf( mapping = NULL,
  data = NULL,
  geom = "step", pad = TRUE,
  position = "identity")+
   labs(x="Body Temp", y="Distribution")+
  coord_flip()
```



Look at the plot of the ecdf to gain more insight into the behavior of the ecdf.

(3) Estimate the 25th, 50th, and 75th percentiles of the temperature data from the ecdf plot. How do these compare with the percentiles computed via the `summary()` function?

```
NHANES %>%
  summarise(quantile(temp))
```

```
  quantile(temp)
1           95.2
2           98.0
3           98.4
4           98.6
5          101.8
```

The data found on the graph and the summary function was identical. However, we had to estimate more using the graph rather than having exact values found in the summarise() function

(4) How does the mean temperature compare to the median temperature?

```
NHANES %>%
  summarise(mean(temp), median(temp))
```

```
  mean(temp) median(temp)
1   98.30529         98.4
```

The median is slightly higher than the mean temperature.

Name data

The file `CensusNames.csv` contains data on surnames in the United States, including the name, the rank of the name among all names, and the number of people in the United States with that name. (Some uncommon names are not included in the data.)

```
CensusNames<-read.csv("CensusNames.csv")
head(CensusNames)
```

```
      name rank   count
1    SMITH    1 2442977
2  JOHNSON    2 1932812
3 WILLIAMS    3 1625252
4    BROWN    4 1437026
5    JONES    5 1425470
6   GARCIA    6 1166120
```

(5) Is your surname in the list? If so, what is its rank, and how many people in the United States have that name?

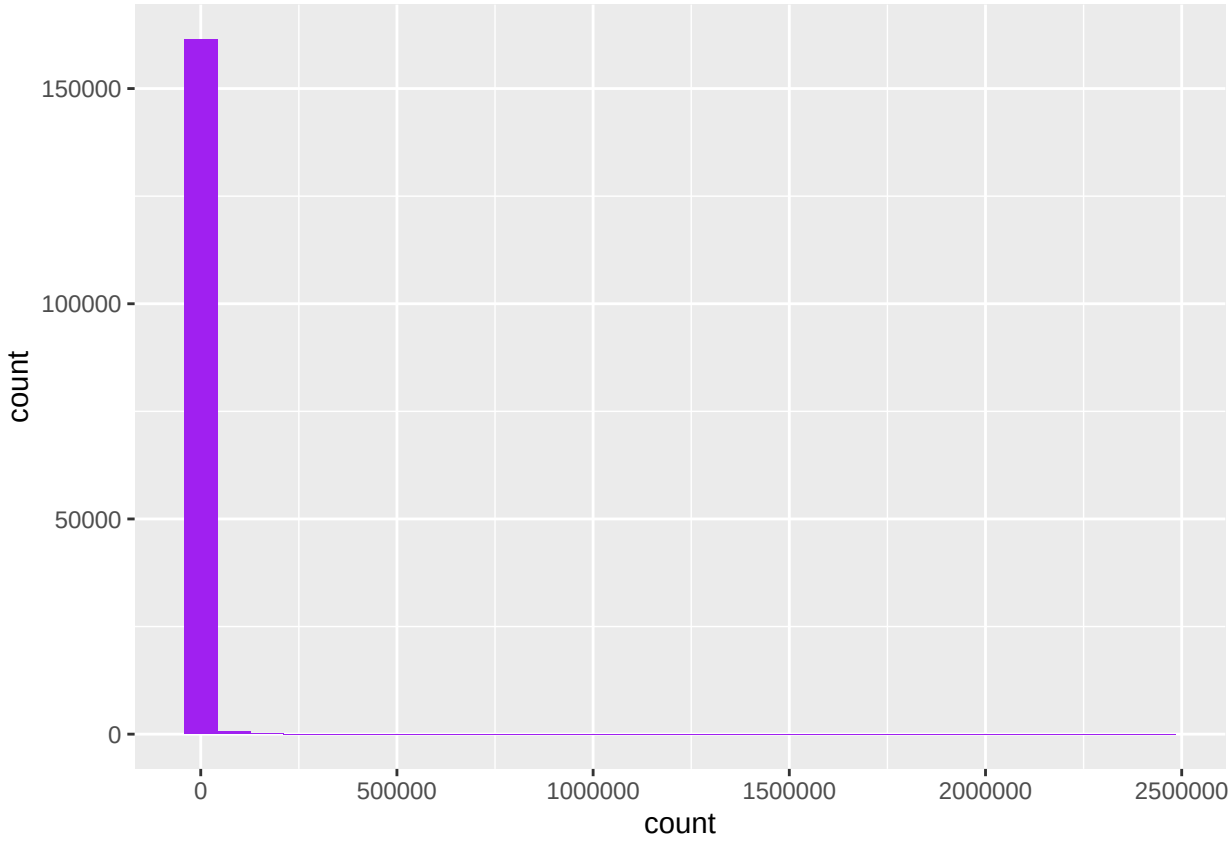
```
CensusNames %>%
  subset(CensusNames$name == "WATSON")
```

```
     name rank  count
81 WATSON   81 252579
```

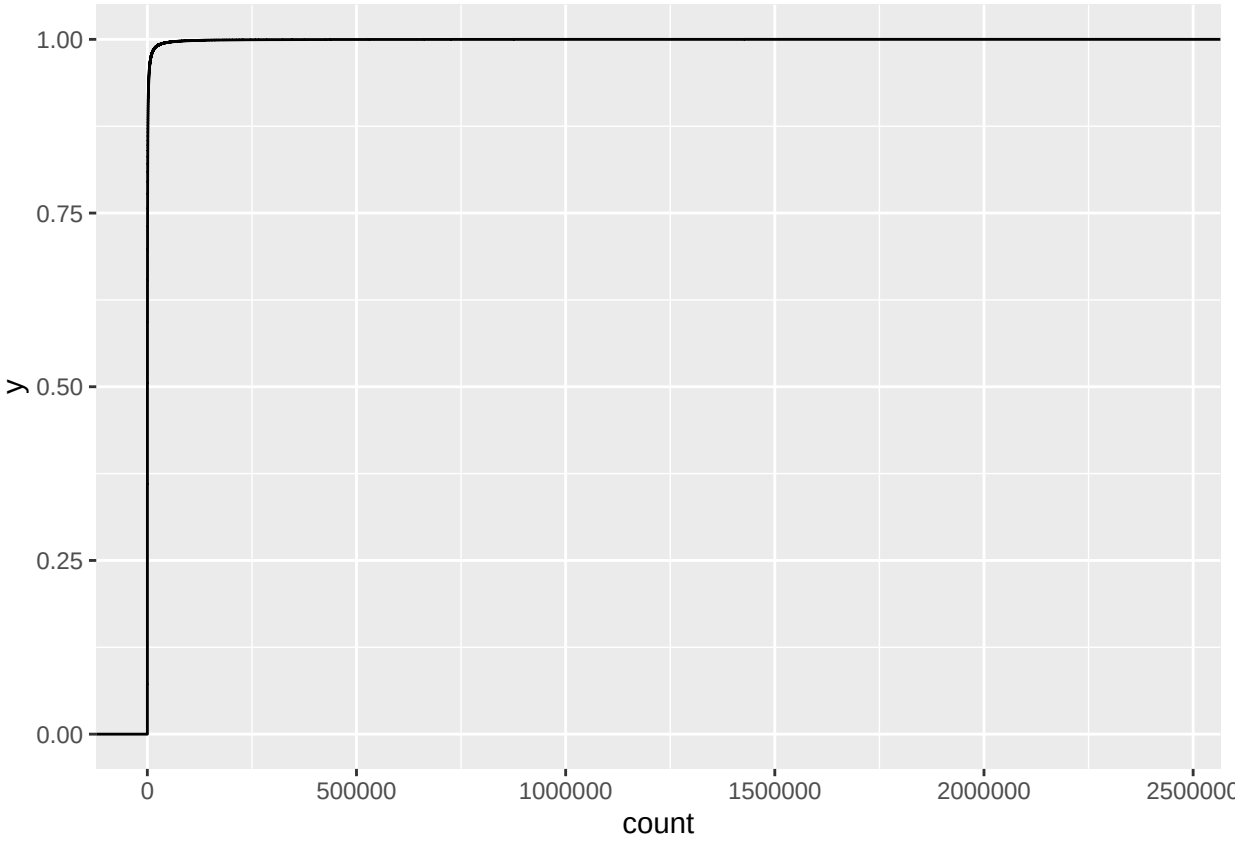
(6) Draw a histogram and an ecdf plot of the variable which counts the number of people with a name. What do you notice from these plots?

```
CensusNames %>%
  ggplot(mapping=aes(count)) +
    geom_histogram(fill="purple", position="dodge")
```

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



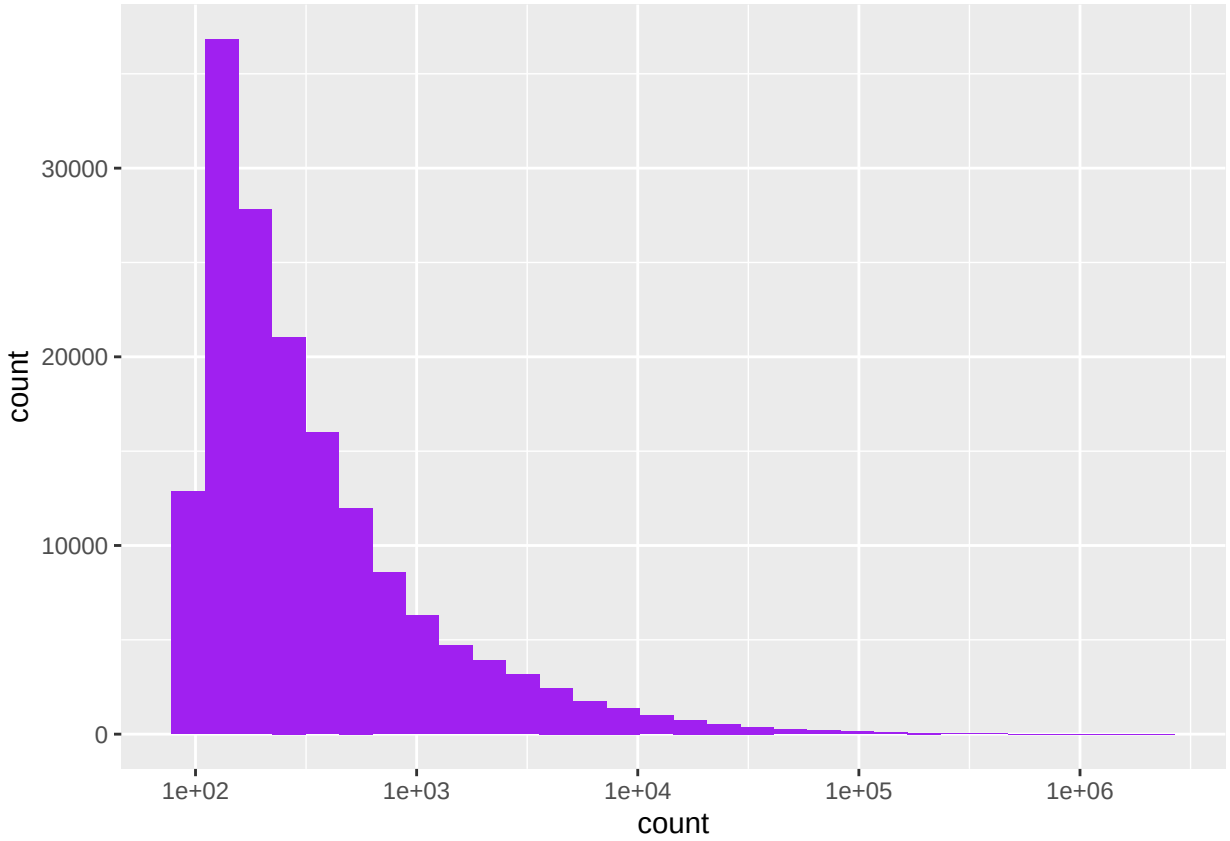

```
CensusNames %>%
  ggplot(mapping=aes(count)) +
  stat_ecdf( mapping = NULL,
  data = NULL,
  geom = "step", pad = TRUE,
  position = "identity")
```



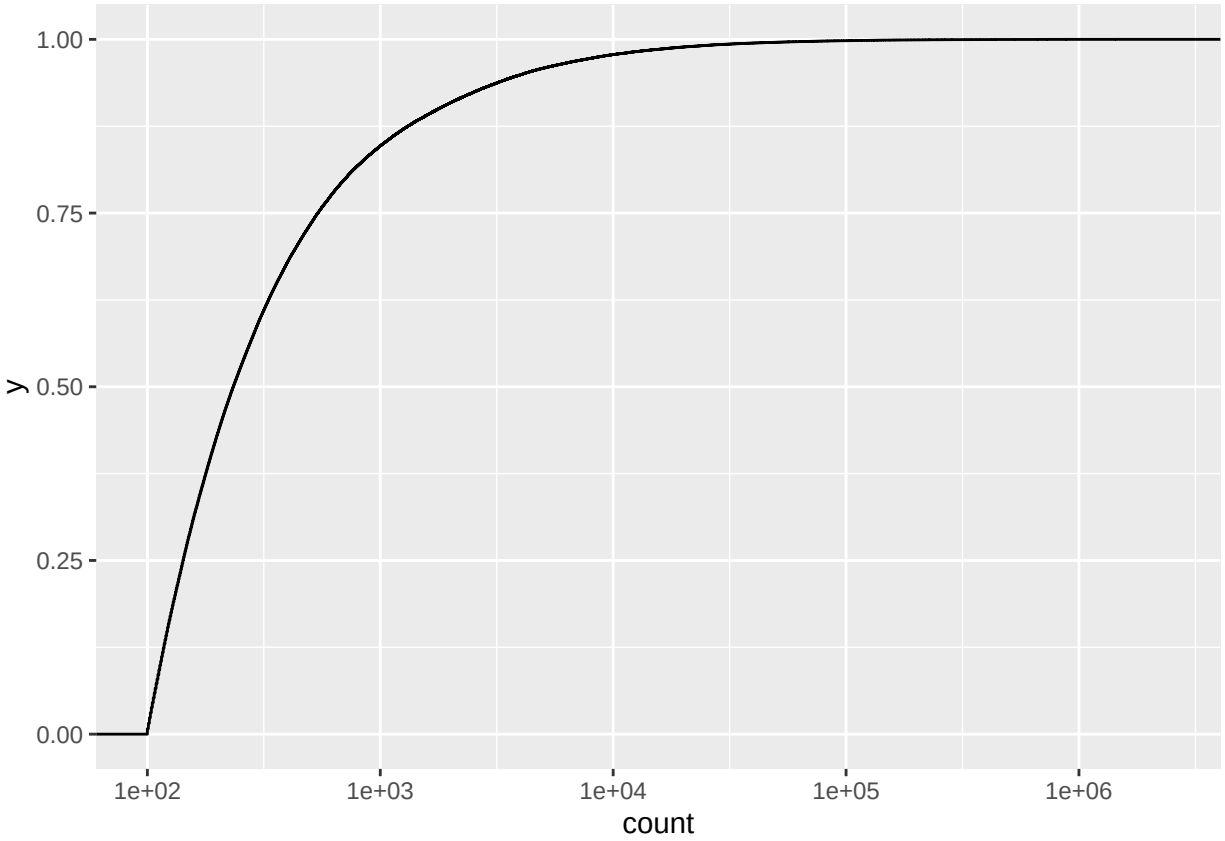
There are a lot of names that do not have many people associated with them, but there are a few names that have many people associated with them. (@) Draw a histogram and an ecdf plot of the base 10 logarithm of the counts. The R function `log10()` computes base 10 logarithms.

```
CensusNames %>%
  ggplot(mapping=aes(count)) +
  geom_histogram(fill="purple", position="dodge")+
  scale_x_log10()
```

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




```
CensusNames %>%
ggplot(mapping=aes(count)) +
stat_ecdf( mapping = NULL,
data = NULL,
geom = "step", pad = TRUE,
position = "identity")+
scale_x_log10()
```



Referenced from Statistics Globe, Accessed on October 27, 2021. https://statisticsglobe.com/draw-histogram-with-logarithmic-scale-in-r

(7) From the ecdf plot of the logarithm of the counts, estimate the 25th, 50th, and 75th percentiles of the counts. Compare these to the computed percentiles from the `summary()` function. Are they similar?

```
CensusNames %>%
  summarise(quantile(count))
```

```
  quantile(count)
1             100
2             142
3             234
4             539
5         2442977
```

The quantiles found using the summarise() function and the estimations made using the ECDF graph are very similar. However, the ECDF graph was estimated.

(8) How does the mean count compare to the median count? How does the mean count compare to the 75th percentile of the counts?

```
CensusNames %>%
  summarise(mean(count), median(count))
```

```
  mean(count) median(count)
1    1637.364           234
```

The median is significantly smaller than the mean. Additionally, the mean of the overall data is considerably larger than the 75% percentile.

US cities data

Extra practice: This section isn't very different from the previous section. Complete it later for your practice.(not graded)

The file `UnitedStatesCities.csv` contains data on cities in the United States, including the name of the city, the population rank of the city among all cities, and the number of people living in the city in 2010.

(9) What are the population and rank of East Lansing?

(10) Draw a histogram and an ecdf plot of the populations. Do these data seem more like the data on body temperatures or the data on the count of people with a particular name?

(11) Draw a histogram and an ecdf plot of the base-10 logarithm of the populations.

(12) From the ecdf plot of the logarithm of the populations, estimate the 25th, 50th, and 75th percentiles of the populations. Compare these to the computed percentiles from the `summary()` function. Are they similar?

(13) How does the mean population compare to the median population? How does the mean count compare to the 75th percentile of the populations?