The best distributions of our lives

Outline

1. Choosing the distribution
2. Simulating from a distribution
   - An example: `VanKilled` as a count variable
   - An example: `PetrolPrice` as a continuous variable
3. Summary

Learning Objectives

1. Appreciate the need for building a vocabulary of distributions.
2. Appreciate the need to fit distributions.
3. Utilise a set of guidelines to help you fit your data to a distribution.

Choosing the distribution

Why do we need a "vocabulary" of distributions?

- Real data will not be perfectly described by known distributions.
- There will be several good models to fit our data to.
- Importantly, a specific distribution provides an *approximate* description of our data.
  - Once a distribution is selected, one can then exploit its properties.
  - E.g., suppose we fitted our data to an exponential distribution.
    - ⋆ We expect any data points that falls outside of say, the 99th percentile, to be outliers.

Some tips

1. What kind of variable are we interested in?
    - Discrete quantity: e.g., number of people in a queue.
    - Continuous quantity: e.g., amount of waiting time.
2. How was your data obtained?
    - Based on how the data was collected, we may get some clues.
    - E.g., counts of traffic accidents typically follow a Poisson distribution.
3. What is the support of your data?
    - Different distributions typically have different supports.
    - E.g., if there are negative numbers in your data, then eliminate distributions with non-negative supports.
4. What is the shape of your histogram?
    - The histogram of your data should be the pdf/pmf of the distribution that it came from.

Simulating from a distribution

Simulating from a distribution

- Suppose we have some sample data.
- After fitting the data to a distribution, one can then make predictions.
  - This means that each (observed) data point can be treated as a random variable.
  - $X \sim p(x|\theta)$ or $X \sim f(x|\theta)$.
  - We are then further assuming that the population data follows the chosen distribution.
- **Important**: We are *not* generating more observations.
  - Instead, we are making additional assumptions about the data.
  - This can be used to make inferences about the parameters.
  - This can also be used to **mimic** real-world scenarios.
- How do we determine the values of the parameter(s) θ?
  - So far, it seems as though we already know the values of the parameters.
  - This is not always possible, like in the case of the example considered for the gamma distribution.
  - The `fitdist()` function from the `fitdistrplus` package can estimate these parameters.
  - The syntax is

```
fitdist(data = <data>, distr = <distribution>)
```

The `Seatbelts` toy dataset

- Commonly used toy dataset.

```
Seatbelts
```

- Let us convert this dataset to a data frame in "R".

```
seatbelts_df <- as.data.frame(Seatbelts)
```

- Monthly totals of drivers in the Great Britain who were killed or serious injured from 1969 to 1984.
- 192 rows (observations) and 8 columns (variables).
- Let us focus on the last 4 years.

```
seatbelts_4 <- tail(seatbelts_df , n = 4*12)
```

The seatbelts_4 toy dataset

```
str(seatbelts_4)
```

```
## 'data.frame':    48 obs. of  8 variables:
##  $ DriversKilled: num  111 106 98 84 94 105 123 109 130 153 ...
##  $ drivers      : num  1474 1458 1542 1404 1522 ...
##  $ front        : num  704 691 688 714 814 736 876 829 818 942 ...
##  $ rear         : num  284 316 321 358 378 382 433 506 428 479 ...
##  $ kms          : num  15226 14932 16846 16854 18146 ...
##  $ PetrolPrice  : num  0.105 0.104 0.117 0.115 0.113 ...
##  $ VanKilled    : num  8 6 7 6 5 4 5 10 7 10 ...
##  $ law          : num  0 0 0 0 0 0 0 0 0 0 ...
```

We shall focus on the `VanKilled` and `PetrolPrice` variables for our examples.

An example: `VanKilled` as a count variable

An example: `VanKilled` as a count variable

- Let us focus on the `VanKilled` variable from the `seatbelts_4` dataset.
- This amounts to using the $ notation:

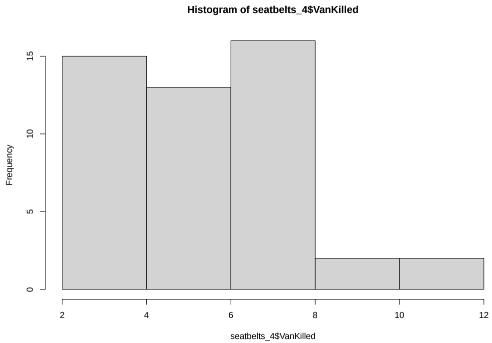
```
seatbelts_4$VanKilled
```

```
 [1]  8  6  7  6  5  4  5 10  7 10 12
[12]  7  4  5  6  4  4  8  8  3  ...
```

- Count variable: Consider a discrete distribution.
- Seemingly no upper limit: Support of variable can be taken to be $x = 0, 1, 2, \cdots, \infty$.
- Let us examine the shape of the histogram by using the `hist()` function:

```
hist(seatbelts_4$VanKilled)
```

- Based on these considerations, a *Poisson distribution* seems like a good choice.



An example: `VanKilled` as a count variable

cont'd

- Let us now load the library:

```
library(fitdistrplus)
```

- In this case, the `distr` argument in the `fitdist()` function must be `"pois"`.

```
(pois_killed <- fitdist(data = seatbelts_4$VanKilled,# Use VanKilled
                        distr = "pois")) # Use the Poisson model
```

```
## Fitting of the distribution ' pois ' by maximum likelihood
## Parameters:
##        estimate Std. Error
## lambda 5.916667  0.3510896
```

- Let us store this parameter:

```
lambda <- pois_killed$estimate[1]
```

An example: `VanKilled` as a count variable

cont'd

- What is the probability of 3 van drivers being killed in a given month?

```
dpois(x = 3, # x=3 number of VanKilled
      lambda = lambda) # Set lambda = lambda
```

```
## [1] 0.09300459
```

- We can also generate random numbers from this Poisson model.

```
rpois(n = 10, # Generate 10 random numbers from the Poisson model
      lambda = lambda) # Set lambda = lambda
```

```
##  [1] 7 4 4 7 5 6 2 8 6 2
```

An example: `PetrolPrice` as a continuous variable

An example: `PetrolPrice` as a continuous variable

- Let us focus on the `PetrolPrice` variable from the `seatbelts_4` dataset.
- This amounts to using the `$` notation:

```
seatbelts_4$PetrolPrice
```

```
##  [1] 0.1047603 0.1040025 0.1166555 0.1151615 0.1129895 0.1138606 0.1191181
##  [8] 0.1244900 0.1232229 0.1206779 0.1210490 0.1169686 0.1127503 0.1080793
## [15] 0.1088385 0.1112918 0.1113040 0.1154544 0.1147683 0.1172074 ...
```

- Continuous variable: Consider a continuous distribution.
- Suppose that based on the shape of the histogram, the support, etc., we determine the *normal distribution* to be a good candidate.

An example: `PetrolPrice` as a continuous variable

cont'd

- In this case, the `distr` argument in the `fitdist()` function must be `"norm"`.

```
(norm_petrol <- fitdist(data=seatbelts_4$PetrolPrice,
        distr = "norm")) # Select the normal dist.
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters:
##         estimate   Std. Error
## mean 0.115747909 0.0005724072
## sd   0.003965753 0.0002942506
```

- Let us store these parameters:

```
mean <- norm_petrol$estimate[1]
sd <- norm_petrol$estimate[2]
```

An example: `PetrolPrice` as a continuous variable

cont'd

- What is the probability of petrol prices being £0.10 or less in a given month?

```
pnorm(q = 0.1, # q=0.10 Pounds
      mean = mean, # Set mean = mean
      sd = sd) # Set sd = sd
```

```
## [1] 3.578943e-05
```

- We can also generate random numbers from this normal distribution.

```
rnorm(n = 8, # Generate 8 random numbers from normal dist.
      mean = mean, # Set mean = mean
      sd = sd) # Set sd = sd
```

```
[1] 0.1179422 0.1278694 0.1152226 0.1139953
[5] 0.1171129 0.1196494 0.1244046 0.1129438
```

Summary

Summary

- Having a vocabulary of distributions gives one an idea of what model to use, based on the data.
- One can use the `fitdist()` function to estimate the parameters of a chosen distribution.
- The `r<distribution>()` functions generate random variables from the chosen distribution.
  - **Important:** We are **not** generating more observations.

References

- R-data — seatbelts dataset.
- Delignette-Muller, M. L. and Dutang, C. (2015). fitdistrplus: An r package for fitting distributions. *Journal of statistical software*, 64:1–34.
- Henderson, H. V. and Velleman, P. F. (1981). Building multiple regression models interactively. *Biometrics*, pages 391–411.
- Wasserman, L. (2004). All of statistics springer new york.