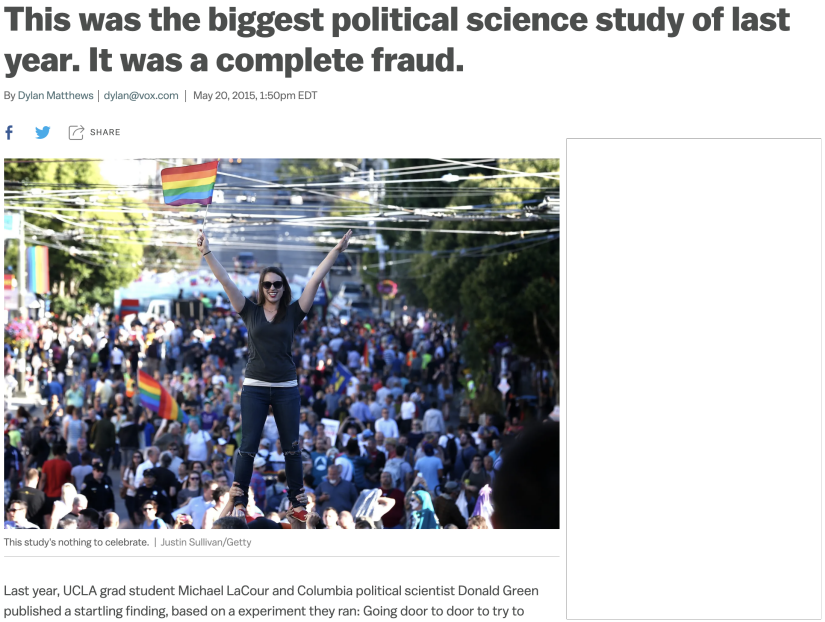
GOVT 707 Lab 2, Probability Distributions in R: AKA How to Fake Your Way into PNAS

Theodore Landsman

September 2, 2021

Introduction: Major Retractions in Social Science



This was the biggest political science study of last year. It was a complete fraud.

By Dylan Matthews | dylan@vox.com | May 20, 2015, 1:50pm EDT

SHARE

This study's nothing to celebrate. | Justin Sullivan/Getty

Last year, UCLA grad student Michael LaCour and Columbia political scientist Donald Green published a startling finding, based on a experiment they ran: Going door to door to try to

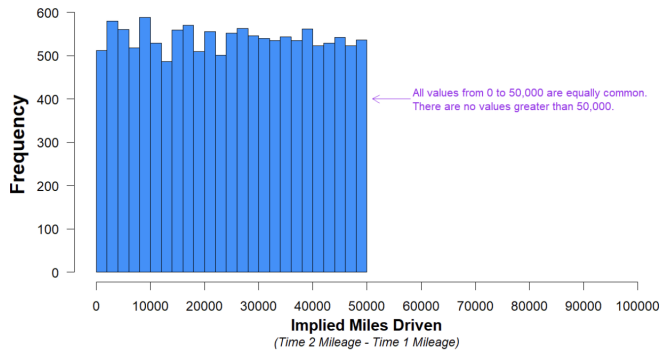
Figure 1: Press on LaCour and Green Retraction

- Big embarassing replication failures in political science have typically involved misprocessing data (eg replication of Gerber and Green 2000 by Imai 2005) or wholesale data fabrication (eg replication of LaCour and Green by Broockman and Kalla 2015)
- Alteration of data seems to be more rare in part because it is harder to detect.
- However alteration of data can still be detected if it is done in clumsy enough ways.

How Researchers on Dishonesty Clumsily Faked Their Data

like this:

Figure 1. Histogram of Miles Driven - Car #1 (N=13,488)



This histogram shows miles driven for the first car in the dataset. There are two important features of this distribution.

Figure 2: Analysis of Shu, Mazar, Gino, Ariely, and Bazerman by DataColada

-Understanding check, why is this obviously fake?

Data in the Real World

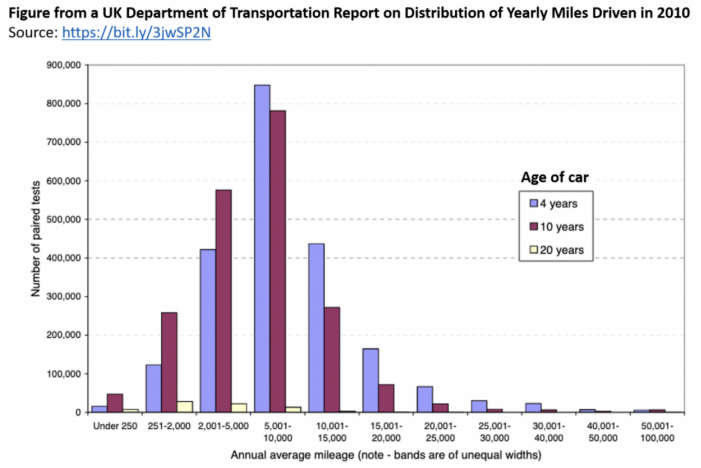


Figure 3: Analysis of Shu, Mazar, Gino, Ariely, and Bazerman by DataColada

-Real data is the product of random but probalistic events aggregated over time, e.g An individual may or may not drive their car on a given day (1,0) and may take quick trips (a few miles) or a huge trip (they have a mental break and decide they want to live on a commune in the Oregon woods).

Data in the Real World Part 2

- This means that real data has both outliers (guy who drives to Oregon), and bounds. You can't drive <0 miles there is also some upper limit on how far a car can go both physically (maximum speed x amount of time elapsed in your time series) and logistically. However, we should expect close to 0 cases near those upper bounds and few cases around the lower bounds.
- If a probabilistic process proceeds for long enough, the results will nearly always tend towards a normal distribution.

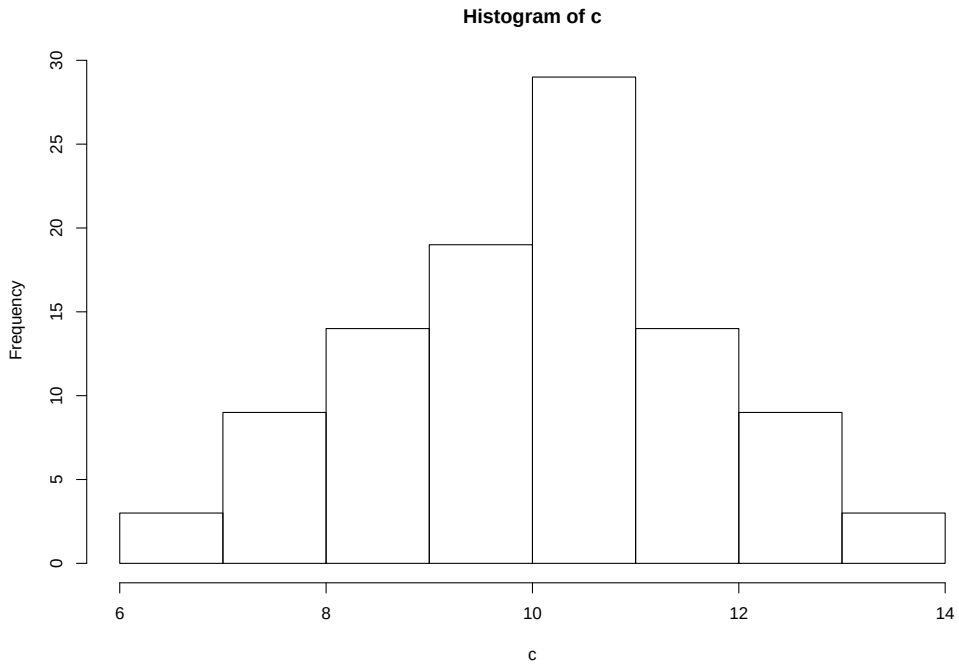
Combining Data

- Question, what should the distribution look like for a variable that is the linear combination (ie adding together) of two random normal variables?

Combining Data Part Two

- Answer: Still random normal!

```
a <- rnorm(100, mean = 5, sd = 1)
b <- rnorm(100, mean = 5, sd = 1)
c <- a + b
hist(c)
```



- Example, what are some causes of getting into Georgetown? How are these causes distributed within the population? Does the existence of causes imply that getting into Georgetown is non-random?

Distributions Typically Used in Statistics

- In data analysis we often deal with classes of probability distributions whose specific shape can be determined with a few parameters.
- Example: by specifying the value of μ and σ, we can determine the exact shape of the normal distribution;

$$\mathrm{N}(\mathrm{x}|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{1}{2\sigma^2}(\mathrm{x} - \mu)^2 \Big)$$

- R contains useful functions to deal with probability distributions.
  - For each class of distribution, R has 4 types of functions.
    - ★ PMF/PDF
    - ★ CDF
    - ★ Quantile
    - ★ Random number generator

Functions Related with Probability Distributions in R (1)

- **Probability Mass Functions (PMF)/Probability Density Function (PDF)**
  - PMF of a discrete probability distribution $f(X = x)$ describes the probability that the random variable X takes the value x.
  - PDF of a continuous probability distribution $f(x)$ describes the *density* where X equals to x (i.e., relative likelihood of occurence of value x)
  - In R, PMF/PDF of a distribution is represented with a function starting from the letter `d`.
    - ★ e.g., PDF of a normal distribution can be computed with `dnorm()` function
- This sounds really compliated, but what it means is that if we have a variable with a known distribution, we can query the probability of the variable taking on a given value with dnorm().

Functions Related with Probability Distributions in R (2)

- **Cumulative Density Function (CDF)**
  - CDF of a probability distribution $F(X = x)$ describes the probability that X is equal to or smaller than x (i.e., $F(X = x) = \Pr(X \leq x)$)
  - Mathematically,

$$F'(X) = f(X)$$

$$\int_{-\infty}^{x} f(X)dx = F(X = x)$$

  - In R, CDF of a distribution is represented with a function starting from the letter `p`
    - ⋆ e.g., CDF of a normal distribution can be computed with `pnorm()` function
- Again to make this simple, if we have a variable with a known distribution, we can query the probability of the variable taking range of values with qnorm().

Functions Related with Probability Distributions in R (3)

- **Inverse distribution function** or **Quantile function** is the inverse function of the CDF. Therefore, it takes the cumulative probability as the input and returns the value $X = x$ as the output.
  - In R, quantile function of a distribution is represented with a function starting from the letter `q`
    - ★ e.g., Quantile function of a normal distribution can be computed with `pnorm()` function
  - This is the inverse of dnorm() and qnorm() in other words, with pnorm() we ask say what is the 50th percentile of the probability distribution and we get a value of a random variable with the listed properties.

Functions Related with Probability Distributions in R (4)

- In R, we can simulate the value of X based on the PMF/PDF of a distribution
  - we call this practice as **random number generation**
- Functions to generate random numbers start from the letter `r`
  - e.g., we can generate random numbers from a normal distribution using the `rnorm()` command
- Strictly speaking, value of X simulated in R is not random; they are a deterministic sequence of numbers which look like random
  - To make sure that we can completely replicate the numbers, we determine the first number by specifying the seed through `set.seed()` command.
    - ★ e.g., First type `set.seed(123)` and then `rnorm(2)` into the R console. You'll get the two numbers -0.5604756 and -0.2301775 regardless of the computing environment.

Distributions Commonly Used in Statistics

Name	PMF/PDF	CDF	Quantile	Random Number
Uniform	`dunif`	`punif`	`qunif`	`runif`
Binomial	`dbinom`	`pbinom`	`qbinom`	`rbinom`
Poisson	`dpois`	`ppois`	`qpois`	`rpois`
Negative Binomial	`dnbinom`	`pnbinom`	`qnbinom`	`rnbinom`
Normal	`dnorm`	`pnorm`	`qnorm`	`rnorm`
Logistic	`dlogis`	`plogis`	`qlogis`	`rlogis`
t	`dt`	`pt`	`qt`	`rt`
F	`df`	`pf`	`qf`	`rf`
χ^2	`dchisq`	`pchisq`	`qchisq`	`rchisq`
Exponential	`dexp`	`pexp`	`qexp`	`rexp`
Weibull	`dweibull`	`pweibull`	`qweibull`	`rweibull`
Gamma	`dgamma`	`pgamma`	`qgamma`	`rgamma`

Examples

- The probability that a binomial random variable X equals to 3 when $n = 10$ and $p = 0.4$ is

```
dbinom(x = 3, size = 10, prob = 0.4)
```

```
## [1] 0.2149908
```

- The probability that a binomial random variable X is larger than 5 when $n = 13$ and $p = 0.3$ is

```
1 - pbinom(q = 5, size = 13, prob = 0.3)
```

```
## [1] 0.1653975
```

```
# or
pbinom(q = 5, size = 13, prob = 0.3, lower.tail = FALSE)
```

```
## [1] 0.1653975
```

Examples (cont.)

- The probability that a random variable X following the normal distribution with $\mu = 3$ and $\sigma^2 = 16$ falls between 5 and 7 is

```
pnorm(q = 7, mean = 3, sd = 4) - pnorm(q = 5, mean = 3, sd = 4)
```

```
## [1] 0.1498823
```

- If you want to find the value of $X \sim \mathrm{N}(-2, 4)$ where the cumulative probability just exceeds 0.4,

```
qnorm(p = 0.4, mean = -2, sd = 2)
```

```
## [1] -2.506694
```