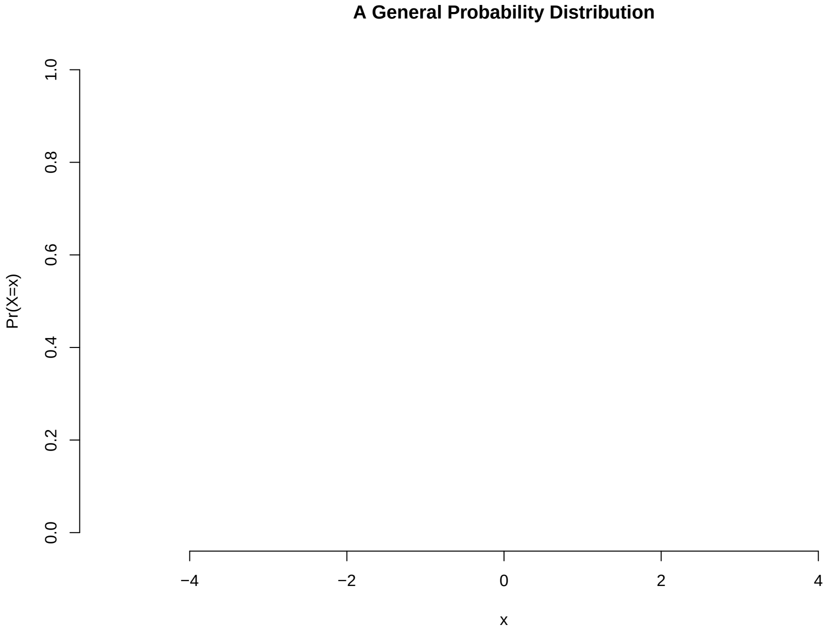
Probability Distributions

Robert W. Walker

2020-02-12

Probability: The Logic of Science

General Representation of Probability


A General Probability Distribution
Pr(X=x)
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
x

Probability Distributions of Two Forms

The Poster and Examples

Continuous vs. Discrete Distributions

Expectation

$$E(X) = \sum_{x \in X} x \cdot Pr(X = x)$$

$$E(X) = \int_{x \in X} x \cdot f(x) dx$$

Variance

$$E[(X - \mu)^2] = \sum_{x \in X} (x - \mu)^2 \cdot Pr(X = x)$$

$$E((X - \mu)^2) = \int_{x \in X} (x - \mu)^2 \cdot f(x) dx$$

The z-transform

In samples, the 0 and 1 are exact; these are features of the mean and *degrees of freedom* from last time.

$$z = \frac{x - \overline{x}}{s_x}$$

.

where $\overline{x}$ is the sample mean of x and s_x is the sample standard deviation of x. Take the example of earnings.

Suppose earnings in a community have mean 55,000 and standard deviation 10,000. This is in dollars. Suppose I earn 75,000 dollars. First, if we take the top part of the fraction in the z equation, we see that I earn 20,000 dollars more than the average (75000 - 55000). Finishing the calculation of z, I would divide that 20,000 dollars by 10,000 dollars per standard deviation. Let's show that.

$$z = \frac{75000 dollars - 55000 dollars}{\frac{10000 dollars}{SD}} = +2SD$$

.

I am 2 standard deviations above the average (the $+$) earnings. All z does is re-scale the original data to standard deviations with zero as the mean.

Suppose I earn 35,000. That makes me 20,000 below the average and gives me a z score of -2. I am 2 standard deviations below average (the -) earnings.

z is an easy way to assess symmetry. The mean of z is always zero but the distribution of z to the left and right of zero is informative. If they are roughly even, then symmetry is likely. If the signs are uneven, then symmetry is unlikely. In R, *z* is automated with the *scale()* command. The last line uses a table and the *sign* command to show me the positive and negative z.

```r
# Generate random normal income
Hypo.Income <- rnorm(1000, 55000, 10000)
# z-transform income [mean 55000ish, std. dev. 10000ish]
z.Income <- scale(Hypo.Income)
# Combine them into a data.frame
Income <- data.frame(Hypo.Income,z.Income)
# Show the data.frame
head(Income)
```

```
##   Hypo.Income   z.Income
## 1    71700.80  1.6890125
## 2    60749.91  0.5791407
## 3    58941.91  0.3959002
## 4    34258.31 -2.1057800
## 5    48504.53 -0.6619266
## 6    49148.68 -0.5966424
```

```r
table(sign(z.Income))
```

```
##
##  -1   1
```

Probability Distributions

A Grape Escape?

2. *The mean of the normal random process of filling is known to be 16.004 ounces with standard deviation 0.028 ounces.*

```
pnorm(15.95, 16.004, 0.028) + pnorm(16.05, 16.004, 0.028,
```

```
## [1] 0.07709829
```

```
1-pnorm(16.1, 16.004, 0.028)
```

```
## [1] 0.0003033834
```

```
pnorm(16.04, 16.004, 0.028)
```

```
## [1] 0.9007286
```

2. *The mean of the normal random process of filling is known to be 16.004 ounces with standard deviation 0.028 ounces.*

- What is the probability that a random jar is outside of requirements? NB: *norm* is the noun with mean (default 0) and sd (default 1).

```
pnorm(15.95, 16.004, 0.028) + pnorm(16.05, 16.004, 0.028, l
```

```
## [1] 0.07709829
```

```
1-pnorm(16.1, 16.004, 0.028)
```

```
## [1] 0.0003033834
```

```
pnorm(16.04, 16.004, 0.028)
```

```
## [1] 0.9007286
```

2. *The mean of the normal random process of filling is known to be 16.004 ounces with standard deviation 0.028 ounces.*

- What is the probability that a random jar is outside of requirements? NB: *norm* is the noun with mean (default 0) and sd (default 1).

```
pnorm(15.95, 16.004, 0.028) + pnorm(16.05, 16.004, 0.028,
```

```
## [1] 0.07709829
```

- What is the probability that a random jar contains more than 16.1 ounces?

```
1-pnorm(16.1, 16.004, 0.028)
```

```
## [1] 0.0003033834
```

```
pnorm(16.04, 16.004, 0.028)
```

```
## [1] 0.9007286
```

2. *The mean of the normal random process of filling is known to be 16.004 ounces with standard deviation 0.028 ounces.*

- What is the probability that a random jar is outside of requirements? NB: *norm* is the noun with mean (default 0) and sd (default 1).

```
pnorm(15.95, 16.004, 0.028) + pnorm(16.05, 16.004, 0.028,
```

```
## [1] 0.07709829
```

- What is the probability that a random jar contains more than 16.1 ounces?

```
1-pnorm(16.1, 16.004, 0.028)
```

```
## [1] 0.0003033834
```

- What is the probability that a random jar contains less than 16.04 ounces?

```
pnorm(16.04, 16.004, 0.028)
```

```
## [1] 0.9007286
```

2. *The mean of the normal random process of filling is known to be 16.004 ounces with standard deviation 0.028 ounces.*

- What is the probability that a random jar is outside of requirements? NB: *norm* is the noun with mean (default 0) and sd (default 1).

```
pnorm(15.95, 16.004, 0.028) + pnorm(16.05, 16.004, 0.028,
```

```
## [1] 0.07709829
```

- What is the probability that a random jar contains more than 16.1 ounces?

```
1-pnorm(16.1, 16.004, 0.028)
```

```
## [1] 0.0003033834
```

- What is the probability that a random jar contains less than 16.04 ounces?

```
pnorm(16.04, 16.004, 0.028)
```

```
## [1] 0.9007286
```

2. *The mean of the normal random process of filling is known to be 16.004 ounces with standard deviation 0.028 ounces.*

- What is the probability that a random jar is outside of requirements? NB: *norm* is the noun with mean (default 0) and sd (default 1).

```
pnorm(15.95, 16.004, 0.028) + pnorm(16.05, 16.004, 0.028, l
```

```
## [1] 0.07709829
```

- What is the probability that a random jar contains more than 16.1 ounces?

```
1-pnorm(16.1, 16.004, 0.028)
```

```
## [1] 0.0003033834
```

- What is the probability that a random jar contains less than 16.04 ounces?

```
pnorm(16.04, 16.004, 0.028)
```

```
## [1] 0.9007286
```

2. *The mean of the normal random process of filling is known to be 16.004 ounces with standard deviation 0.028 ounces.*

- What is the probability that a random jar is outside of requirements? NB: *norm* is the noun with mean (default 0) and sd (default 1).

```
pnorm(15.95, 16.004, 0.028) + pnorm(16.05, 16.004, 0.028,
```

```
## [1] 0.07709829
```

- What is the probability that a random jar contains more than 16.1 ounces?

```
1-pnorm(16.1, 16.004, 0.028)
```

```
## [1] 0.0003033834
```

- What is the probability that a random jar contains less than 16.04 ounces?

```
pnorm(16.04, 16.004, 0.028)
```

```
## [1] 0.9007286
```

Scottish Pounds

The Median is a Binomial with p=0.5

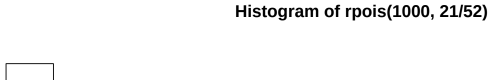
Air Traffic Controllers

FAA Decision: Expend or do not expend scarce resources investigating claimed staffing shortages at the Cleveland Air Route Traffic Control Center.

Essential facts: The Cleveland ARTCC is the US's busiest in routing cross-country air traffic. In mid-August of 1998, it was reported that the first week of August experienced 3 errors in a one week period; an error occurs when flights come within five miles of one another by horizontal distance or 2000 feet by vertical distance. The Controller's union claims a staffing shortage though other factors could be responsible. 21 errors per year (21/52 errors per week) has been the norm in Cleveland for over a decade.

1. Plot a histogram of 1000 random weeks. NB: *pois* is the noun with no default for λ – the arrival rate.

```
hist(rpois(1000, 21/52))
```



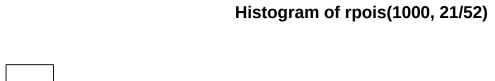
Air Traffic Controllers

FAA Decision: Expend or do not expend scarce resources investigating claimed staffing shortages at the Cleveland Air Route Traffic Control Center.

Essential facts: The Cleveland ARTCC is the US's busiest in routing cross-country air traffic. In mid-August of 1998, it was reported that the first week of August experienced 3 errors in a one week period; an error occurs when flights come within five miles of one another by horizontal distance or 2000 feet by vertical distance. The Controller's union claims a staffing shortage though other factors could be responsible. 21 errors per year (21/52 errors per week) has been the norm in Cleveland for over a decade.

1. Plot a histogram of 1000 random weeks. NB: *pois* is the noun with no default for λ – the arrival rate.

```
hist(rpois(1000, 21/52))
```



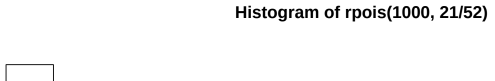
Air Traffic Controllers

FAA Decision: Expend or do not expend scarce resources investigating claimed staffing shortages at the Cleveland Air Route Traffic Control Center.

Essential facts: The Cleveland ARTCC is the US's busiest in routing cross-country air traffic. In mid-August of 1998, it was reported that the first week of August experienced 3 errors in a one week period; an error occurs when flights come within five miles of one another by horizontal distance or 2000 feet by vertical distance. The Controller's union claims a staffing shortage though other factors could be responsible. 21 errors per year (21/52 errors per week) has been the norm in Cleveland for over a decade.

1. Plot a histogram of 1000 random weeks. NB: *pois* is the noun with no default for λ – the arrival rate.

```
hist(rpois(1000, 21/52))
```



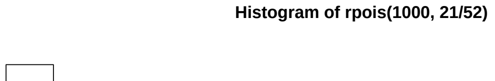
Air Traffic Controllers

FAA Decision: Expend or do not expend scarce resources investigating claimed staffing shortages at the Cleveland Air Route Traffic Control Center.

Essential facts: The Cleveland ARTCC is the US's busiest in routing cross-country air traffic. In mid-August of 1998, it was reported that the first week of August experienced 3 errors in a one week period; an error occurs when flights come within five miles of one another by horizontal distance or 2000 feet by vertical distance. The Controller's union claims a staffing shortage though other factors could be responsible. 21 errors per year (21/52 errors per week) has been the norm in Cleveland for over a decade.

1. Plot a histogram of 1000 random weeks. NB: *pois* is the noun with no default for λ – the arrival rate.

```
hist(rpois(1000, 21/52))
```

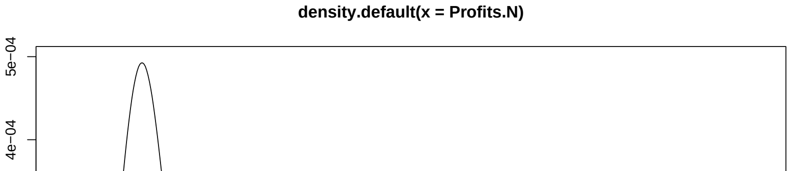


Deaths by Horse Kick in the Prussian cavalry?

[Given time] A Less Basic Monte Carlo Simulation:

1. Customers arriving at a car dealership at a rate of 6 per hour.

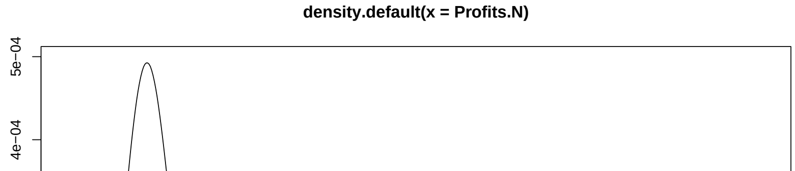
```
Customers <- rpois(1000, 6) # Customers ~ Poisson(6)
Purchasers <- rbinom(1000, size=Customers, prob=0.15) # P
# Next part needs a coding trick.  For each row [of 1000],
Profits.U <- sapply(c(1:1000), function(x) { sum(runif(Purc
Profits.N <- sapply(c(1:1000), function(x) { sum(rnorm(Purc
plot(density(Profits.N))
```



[Given time] A Less Basic Monte Carlo Simulation:

1. Customers arriving at a car dealership at a rate of 6 per hour.
2. Each customer has a 15% probability of making a purchase.

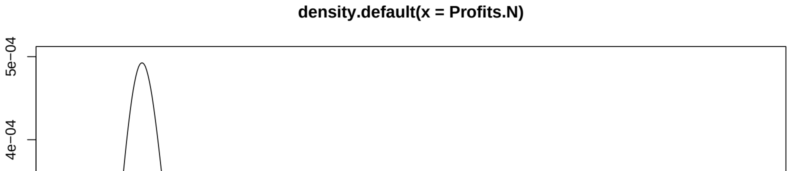
```
Customers <- rpois(1000, 6) # Customers ~ Poisson(6)
Purchasers <- rbinom(1000, size=Customers, prob=0.15) # P
# Next part needs a coding trick.  For each row [of 1000],
Profits.U <- sapply(c(1:1000), function(x) { sum(runif(Purc
Profits.N <- sapply(c(1:1000), function(x) { sum(rnorm(Purc
plot(density(Profits.N))
```



[Given time] A Less Basic Monte Carlo Simulation:

1. Customers arriving at a car dealership at a rate of 6 per hour.
2. Each customer has a 15% probability of making a purchase.
3. Purchasers yield [this part is harder]:

```
Customers <- rpois(1000, 6) # Customers ~ Poisson(6)
Purchasers <- rbinom(1000, size=Customers, prob=0.15) # P
# Next part needs a coding trick.  For each row [of 1000],
Profits.U <- sapply(c(1:1000), function(x) { sum(runif(Purc
Profits.N <- sapply(c(1:1000), function(x) { sum(rnorm(Purc
plot(density(Profits.N))
```

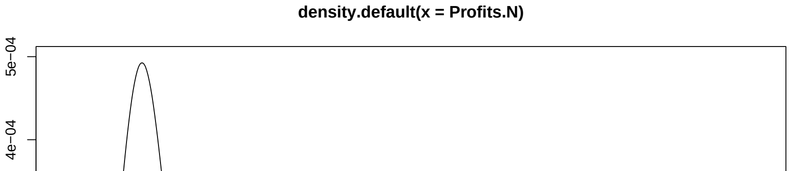


[Given time] A Less Basic Monte Carlo Simulation:

1. Customers arriving at a car dealership at a rate of 6 per hour.
2. Each customer has a 15% probability of making a purchase.
3. Purchasers yield [this part is harder]:

- Uniform profits over the interval \$1000-\$3000.

```
Customers <- rpois(1000, 6) # Customers ~ Poisson(6)
Purchasers <- rbinom(1000, size=Customers, prob=0.15) # P
# Next part needs a coding trick.  For each row [of 1000],
Profits.U <- sapply(c(1:1000), function(x) { sum(runif(Purc
Profits.N <- sapply(c(1:1000), function(x) { sum(rnorm(Purc
plot(density(Profits.N))
```



[Given time] A Less Basic Monte Carlo Simulation:

1. Customers arriving at a car dealership at a rate of 6 per hour.
2. Each customer has a 15% probability of making a purchase.
3. Purchasers yield [this part is harder]:

- Uniform profits over the interval \$1000-\$3000.
- Normal profits that average \$1500 with standard deviation \$500.

```
Customers <- rpois(1000, 6) # Customers ~ Poisson(6)
Purchasers <- rbinom(1000, size=Customers, prob=0.15) # P
# Next part needs a coding trick.  For each row [of 1000],
Profits.U <- sapply(c(1:1000), function(x) { sum(runif(Purc
Profits.N <- sapply(c(1:1000), function(x) { sum(rnorm(Purc
plot(density(Profits.N))
```