Sampling from Probability Distributions

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Many applications in data science require random samples from specific probability distributions.
- Sometimes it may be difficult to solve problems mathematically.
- *Forms the basis of simulation studies.*
- At its base, this involves generating random numbers.

Pseudo Random Numbers

- Used to generate random samples from probability distributions.

- **Random number** is uniformly distributed between 0 and 1.
    - *Technically not truly random.*

- In R: `runif(n, min = 0, max = 1)`

Sampling from Discrete Probability Distributions

- *Recall* that the values of the CDF divide the interval from 0 to 1.
- By generating random numbers between 0 and 1 we can match the corresponding outcomes.
- Then any random number falls within one of the intervals.

Example 1

- Assume the following discrete probability distribution:

Outcome:	1	2	3	4
$p(x)$ PMF:	0.2	0.3	0.1	0.4
$F(x)$ CDF:	0.2	0.5	0.6	1

- Use R to generate $n = 10$ random numbers from this distribution.

	Interval		Outcome
0	to	0.2	1
0.2	to	0.5	2
0.5	to	0.6	3
0.6	to	1.0	4

Sampling from Continuous Probability Distributions

- Again, the values of the CDF divide the interval from 0 to 1.
- By generating random numbers between 0 and 1 we can match the corresponding outcomes.
- Input the values into the CDF can be converted back to outcomes (inverse transform sampling).
  1. Generate $U \sim \text{Unif}(0,1)$
  2. Let $X = F_X^{-1}(U)$

Example 2

- Using R, generate $n = 1000$ random observations for the standard uniform distribution `runif(1000, min = 0, max = 1)`
- Next, use those observations and the inverse of the exponential CDF to generate observations from an exponential distribution with $\lambda = 2$.
- Hint:

$$F(X) = 1 - e^{-2x}$$

and

$$F^{-1}(U) = -\frac{ln(1-u)}{2}$$

Random Numbers in R

- There are existing functions in R for generating random numbers from common distributions.
- Generally: `rdistribution.name(n, ...)`

R Functions

- Uniform Distribution: `runif(n, min = 0, max = 1)`
- Normal Distribution: `rnorm(n, mean = 0, sd = 1)`
- Exponential Distribution: `rexp(n, rate = 0.5)`
- Poisson Distribution: `rpois(n, lambda = 3)`
- Binomial Distribution: `rbinom(n, size = 10, prob = 0.3)`
- Geometric Distribution: `rgeom(n, prob = 0.2)`

Thinking About Sample Sizes

- Increasing the number of simulations (n) will cause the sample parameters converge to the distribution.
- This occurs at the expense of computing time.

Example 3

- Set your seed to be 1000.
- In R, generate 3 samples from a normal distribution where $\mu = 1$ and $\sigma = 2$.
    1. $n_1 = 2$
    2. $n_2 = 20$
    3. $n_3 = 200$
- Calculate the sample mean ($\bar{x}$) and sample variance (s).
- Comment on your findings.

Final Thoughts

- Simulations are often used to test models before using them on real data.
- We will speak more about simulations in another term.
- A *representative* sample is better than a *large* sample.

Exercise 1

- Assume the following discrete probability distribution:

Outcome:	1	2	3
$p(x)$ PMF:	0.25	0.45	0.30
$F(x)$ CDF:	0.25	0.70	1.0

- Use R to generate $n_1 = 10$ & $n_2 = 100$ random numbers from this distribution.

	Interval		Outcome
0	to	0.25	1
0.25	to	0.7	2
0.7	to	1.0	3

- Comment on your results.

Exercise 2

- Using R, generate $n = 1000$ random observations for the standard uniform distribution `runif(1000, min = 0, max = 1)`
- Use the uniform observations and the `qnorm(p, mean, sd)` function to generate observations from a normal distribution with mean $= 3$ and standard deviation (sd) $= 2$.

Exercise 3

- In R, generate 3 samples from a **binomial** distribution where size (parameter n) $= 10$ and $p = 0.2$.
- Sample sizes:
  1. $n_1 = 2$
  2. $n_2 = 20$
  3. $n_3 = 200$
- Calculate the sample mean $(\bar{x})$ and sample variance (s).
- Compare your findings to the theoretical *expected value* and *variance*.

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.

- Inverse Transform Sampling