# 9 Probability and Statistics

(AST230) R for Data Science

Md Rasel Biswas

# Generating random data

- The function `sample()` is used to generate random values from a vector, and it has the following arguments:
  - `x` → A vector of outcome you want to sample from
  - `size` → The number of samples (observations) you want to draw
  - `replace` → It can take either `TRUE` or `FALSE`
  - `prob` → Specifies probability of selection of different elements of `x`

```
sample(x = 1:10, size = 4, replace = F)
```

```
[1]  7  4 10  2
```

# Generating random data

## Select 10 numbers from 0 to 100

```
sample(x = 0:100, size = 10, replace = F) # replace=FALSE
```

```
[1] 76 23  9 93 19 31 15 38 65 11
```

```
sample(x = 0:100, size = 10, replace = T) # replace=TRUE
```

```
[1]  31  47  79  27   1  52  48  91  57 100
```

# Generating random data

- Select students' grades randomly

```
sample(replace = TRUE, x = LETTERS[1:4], size = 10)
```

```
[1] "B" "D" "A" "D" "C" "A" "A" "D" "D" "D"
```

- Tossing a fair coin 10 times

```
sample(replace = TRUE, x = c("H", "T"), size = 10)
```

```
[1] "H" "H" "H" "H" "H" "T" "H" "H" "T" "T"
```

- Tossing a biased coin 10 times

```
sample(replace = TRUE, x = c("H", "T"), size = 10, prob = c(.7, .25))
```

```
[1] "H" "H" "H" "H" "H" "T" "H" "H" "H" "H"
```

# Use of initial seed in generating random numbers

## Without seed:

```r
# No seed
sample(1:10, 3)
```

```
[1] 1 2 5
```

```r
# No seed
sample(1:10, 3)
```

```
[1] 4 5 7
```

```r
# No seed
sample(1:10, 3)
```

```
[1] 4 1 7
```

## With seed:

```r
set.seed(100)
sample(1:10, 3)
```

```
[1] 10  7  6
```

```r
set.seed(100)
sample(1:10, 3)
```

```
[1] 10  7  6
```

```r
set.seed(100)
sample(1:10, 3)
```

```
[1] 10  7  6
```

# Useful functions related to probability distributions:

- R has built in many functions for conveniently working with a large number of distributions.
- The quantities that are of main interest from any probability distribution are:
  - Probability density function (pdf for continuous variable) or probability mass function (pmf for discrete variable)
  - Cumulative distribution function (cdf)
  - Quantile function (inverse cdf)
  - Generating random sample from respective distributions.

# Useful functions related to probability distributions:

| Distribution | Density Function | Cumulative Distribution | Quantile | Random Variates |
|---|---|---|---|---|
| Normal | `dnorm()` | `pnorm()` | `qnorm()` | `rnorm()` |
| Poisson | `dpois()` | `ppois()` | `qpois()` | `rpois()` |
| Binomial | `dbinom()` | `pbinom()` | `qbinom()` | `rbinom()` |
| Uniform | `dunif()` | `punif()` | `qunif()` | `runif()` |

- Such functions are available for other probability distributions, such as exponential, logistic, Chi-squared etc.

# `rbinom()` and `rnorm`

- `rbinom()` is used to draw a sample from a **binomial distribution**
  - `size` $\rightarrow$ number of Bernoulli trials
  - `prob` $\rightarrow$ probability of success
  - `n` $\rightarrow$ number of observations
- Draw a sample of size 8 from $B(10, 0.75)$

```
rbinom(size = 10, prob = .75, n = 8)
```

```
[1] 9 8 8 6 8 7 9 7
```

# rbinom() and rnorm

- `rnorm()` is used to draw a sample from a **normal distribution**
  - `mean` $\rightarrow$ mean of the distribution $(\mu)$
  - `sd` $\rightarrow$ standard deviation of the distribution $(\sigma)$
  - `n` $\rightarrow$ number of observations
- Draw a sample of size 5 from $N(10, 16)$

```r
rnorm(mean = 10, sd = 4, n = 5)
```

```
[1] 14.743527  8.970948 11.748854  8.539669 11.986696
```

# `pnorm()`

- For $X \sim N(50, 3^2)$, find $P(45 < X < 55)$.
- $P(a < X \leq b) = F(b) - F(a)$

```r
pnorm(q = 55, mean = 50, sd = 3) -
  pnorm(q = 45, mean = 50, sd = 3)
```

```
[1] 0.9044193
```

# `dnorm()`

- For $X \sim Bin(10, 0.5)$, find $P(X = 5)$.

```r
dbinom(x = 5, size = 10, prob = 0.5)
```

```
[1] 0.2460938
```

# `qnorm()`

- Let $Z$ follows a standard normal distribution. Then the 0.975−quantile is $Z_{0.975} \approx 1.96$. It means the probability of sampling a value less than or equal to 1.96 is 0.975 or $97.5$

```
qnorm(p = 0.975, mean = 0, sd = 1)
```

```
[1] 1.959964
```