First things first

• NA, NaN and NULL
  - NA: Not Available → A missing value
    ```r
    > a = c(1.7, 4.9, 6.2, 2.5, NA, 7.9)
    > a
    [1] 1.7 4.9 6.2 2.5 NA  7.9
    > is.na(a)
    [1] FALSE FALSE FALSE FALSE TRUE FALSE
    ```
  - NaN: Not a Number → an undefined or unrepresentable value
    ```r
    > b <- log(a)
    Warning message:
    In log(a): NaNs produced
    > b
    [1]  NA 1.5892352 1.824593 0.9162907 5.0 2.0668628
    > is.nan(b)
    [1] TRUE FALSE FALSE FALSE FALSE FALSE
    ```
  - NULL: The null object
    ```r
    > c <- NULL
    > is.null(c)
    [1] TRUE
    > is.null(b)
    [1] FALSE
    ```
Probability distributions
• What is a probability distribution?
  – A function that describes the probability of a random variable taking certain values
  – There are many distributions (en.wikipedia.org/wiki/List_of_probability_distributions)
• Discrete vs. continuous

Discrete vs. Continuous
• Discrete distributions
  – E.g. Bernoulli, Binomial and Poisson distributions
  – E.g. The probability of finding six or more eggs in a bird's nest

Discrete vs. continuous
• Continuous distributions
  – E.g. Exponential, Normal, Student's t, Chi-squared and F-distributions
  – E.g. The probability of measuring a wing length of 85 cm

Photo: Per Harald Olsen
Snow Bunting nest

Snow Bunting
Four fundamental items

For a statistical distribution we can calculate:
1. Density or point probability (e.g. \( \text{dnorm}(), \text{dbinom}() \))
2. Cumulative probability (e.g. \( \text{pnorm}(), \text{pbinom}() \))
3. Quantiles (e.g. \( \text{qnorm}(), \text{qbinom}() \))
4. Random numbers (pseudo) (e.g. \( \text{rnorm}(), \text{rbinom}() \))

Density or point probability

- **Continuous** distribution
  - Characterized by a probability density function, \( f(x) \)
  - The probability of getting a value close to \( x \):
    \[
    P(a \leq X \leq b) = \int_a^b f(x) \, dx
    \]
  - Need to integrate \( f(x) \) on the interval from \( a \) to \( b \) to get the probability for continuous distributions

Density or point probability

- **Continuous** - Normal distribution
  \[
  P(a \leq X \leq b) = \int_a^b f(x) \, dx
  \]
  - Expectation = \( \mu = 100 \)
  - Variance = \( \sigma^2 = 100 \)
Density or point probability

**Continuous - Normal distribution**

\[
P(a \leq X \leq b) = \int_a^b f(x)dx = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx
\]

\(\mu = 100, \sigma = 10, a = 89.9, b = 90.1\)

```r
> dnorm(90, mean = 100, sd = 10)
[1] 0.02419707
> integrate(dnorm, lower = 89.9, upper = 90.1, + mean = 100, sd = 10)
0.04839444 with absolute error < 5.4e-17
```

**Discrete distribution**

- Characterized by a probability mass function, \(f(x)\)
- The probability of getting exactly the value \(x\)

\[P(X = x) = f(x)\]

You get the probability from \(f(x)\) directly with discrete distributions

**Discrete – Binomial distribution**

\[P(X = x) = f(x)\]

\[\text{Expectation } = np = 20 \cdot 0.35 = 7\]

\[\text{Variance } = np(1-p) = 4.55\]
Density or point probability

- Discrete – Binomial distribution
  
  \[ P(X = x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x} \]

  \[ x = 10, \ n = 20, \ p = 0.35 \]

  ```
  dBinom(10, 20, 0.35)
  [1] 0.06861497
  ```

Cumulative probability

Describes the probability of getting \( x \) or less in a given distribution

- Continuous distributions:
  
  \[ P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t) \, dt \]

- Discrete distributions:
  
  \[ P(X \leq x) = F(x) = \sum_{i \leq x} f(x_i) \]

Cumulative probability

- Continuous – Normal distribution
  
  \[ P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t) \, dt \]

  - Expectation = \( \mu = 100 \)
  - Variance = \( \sigma^2 = 100 \)
Cumulative probability

- **Continuous** – Normal distribution
  
  \[ P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t) \, dt \]

  mean = 100, sd = 10, x = 90

  > pnorm(90, mean = 100, sd = 10)
  [1] 0.1586553

- **Continuous** – Normal distribution
  
  \[ P(X > x) = 1 - F(x) \]

  mean = 100, sd = 10, x = 90

  > 1 - pnorm(90, mean = 100, sd = 10)
  [1] 0.8413447
  > pnorm(90, mean = 100, sd = 10, lower.tail = FALSE)
  [1] 0.1586543

- **Discrete** – Binomial distribution
  
  \[ P(X \leq x) = F(x) = \sum_{i=x}^{n} f(x_i) \]
Cumulative probability

• Discrete – Binomial distribution

\[ P(X \leq x) = F(x) = \sum_{i=x}^{n} f(x_i) \]

\( x = 10, \ n = 20, \ p = 0.35 \)

\[
\text{> pbinom(10, 20, 0.35)} \\
\{[1] 0.9468334\}
\]

Cumulative probability

• Discrete – Binomial distribution

\[ P(X > x) = 1 - F(x) \]

\( x = 10, \ n = 20, \ p = 0.35 \)

\[
\text{> 1-pbinom(10, 20, 0.35)} \\
\{[1] 0.0531661\}
\]

\[
\text{> pbinom(10, 20, 0.35, lower.tail = F)} \\
\{[1] 0.0531661\}
\]

Quantiles

\( x_p \), the \( p \)-quantile, is the value with the property that there is a probability \( p \) of getting a value less than or equal to it

• Continuous – Normal distribution

\[ P(X \leq x_p) = F(x_p) = p \]

\( p=0.95, \ \text{mean} = 100, \ sd = 10 \)

\[
\text{> qnorm(0.95, mean = 100, sd = 10)} \\
\{[1] 116.4485\}
\]
Quantiles

Quantiles, the $p$-quantile, is the value with the property that there is a probability $p$ of getting a value less than or equal to it.

- **Discrete - Binomial distribution**

$$P(X \leq x_p) = F(x_p) = p$$

$x = 10$, $n = 20$, $p = 0.35$

```
> qbinom(0.95, 20, 0.35)
[1] 11
```

Random numbers

```
> rnorm(10, 100, 10)
[5] 87.46305 106.48988
[7] 93.67608 95.85592
[9] 89.27771 100.35549
```

> rbinom(10, 20, 0.35)

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

A data set

- Saltmarsh sharp-tailed sparrows (*Ammodramus caudacutus*) (Spisshalespurv)

- Is distributed along the Atlantic coast of USA
- 246 ♀ measured:
  - Age
  - Weight
  - Wing length
  - Egg size
Linear regression

- Load data

```r
sparrow <- read.csv("C:/RData/sparrow.csv")
> head(sparrow, 10)
 wingcrd  wt  age  eggsize
 1   56.5 17.4 0 2555.731
 2   57.0 17.5 0 2464.543
 3  53.5 18.5 0 2869.429
 4  54.5 21.8 0 3308.374
 5  55.0 18.6 0 2969.068
 6  55.0 17.7 0 2555.187
 7  56.5 16.6 0 2570.419
 8  55.0 17.5 0 2582.778
 9  56.0 21.6 0 3309.669
10  55.5 17.2 0 2649.100
```

- Hypothesis:
  - As eggs are costly to produce, females with large fat reserves are able to produce larger eggs than females with less fat reserves

- Response variable
  - Egg size: continuous variable

- Explanatory variable
  - Female weight: continuous variable

- What model?
  - Linear regression: \( y = a + bx \)

Linear regression

\[
\text{Egg size} = \text{intercept} + \text{slope} \cdot \text{weight}
\]

```r
> lr <- lm(eggsize ~ wt, data = sparrow)
> lr

Call:
  lm(formula = eggsize ~ wt, data = sparrow)
Coefficients:
     (Intercept)          wt
      40.76         147.42
```
Linear regression – assumptions

1. Linearity of the relationship
2. Independence of errors
3. Homoscedasticity of the errors
4. Normality of the errors

Linear regression – parameters

> summary(lm)
Call:
  lm(formula = egg.size ~ wt, data = sparrow)
Residuals:
    Min      1Q  Median      3Q     Max
  -295.526  -72.435   -4.834   77.058  260.953
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.761     74.434    0.548  0.58446
wt           147.419     3.562  41.209  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 103.1 on 244 degrees of freedom
Multiple R-squared: 0.8992,  Adjusted R-squared: 0.8986
F-statistic: 1304 on 1 and 244 DF,  p-value: < 2.2e-16

Linear regression – graph

Egg size = 40.8 + 147.4 · weight
R resources
- R reference card by Tom Short
  - cran.r-project.org/doc/contrib/Short-refcard.pdf

Extras
- How to create the normal and binomial distribution density graph with red polygon or line with R
  - Normal density curve:
```r
x <- seq(10, 100, length = 200)
y <- dnorm(x, mean = 100, sd = 10)
plot(x, y, type = "l", ylab = "Density")
x2 <- seq(0.5, 99.5, length = 100)
y2 <- dnorm(x2, mean = 100, sd = 10)
polygon(x,y, x2,y2, col = "red", border = "red")
```
  - Binormal density (point probability) curve:
```r
s <- 1:20
plot(s, dbinom(s, 20, 0.35), ylab = "Point probability", type = "s")
hist(s, breaks=c(10,100), col=dbinom(s, 20, 0.35), col = "red")
```