

10. Probability distributions

These notes consider the Chapter 10 of the [handbook](#) on various probability distributions.

Discrete distributions

For random variables with countable number of possible values.

Binomial distribution

See page 200 for proper definition.

Useful when a random variable X has exactly two exclusive possible outcomes (e.g. success/fail) with known probabilities $p \in [0, 1]$ (success) and $q = 1 - p$ (fail). For $n \in \mathbb{N}$ trials with $k = 0, 1, 2, \dots, n$ successes

$$X \sim \text{Bin}(n, p)$$

Probability Mass Function (PMF)

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

Cumulative Distribution Function (CDF)

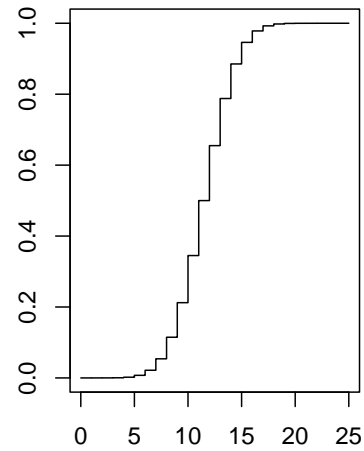
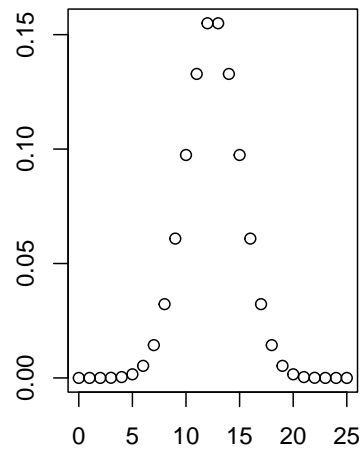
$$F(k) = P(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i q^{n-i}$$

In the following code examples, equal probabilities are assumed, $n = \text{size}$, $p = \text{prob} = 0.5$.

```
# Sequence for visualization
binomial_seq <- seq(0, 25, by = 1)

# Functions
binomial_pmf <- dbinom(x = binomial_seq, size = 25, prob = 0.5)
binomial_cdf <- pbinom(q = binomial_seq, size = 25, prob = 0.5)

# Plot
par(mfrow = c(1, 2))
plot(binomial_seq, binomial_pmf, ann = FALSE)
plot(binomial_seq, binomial_cdf, type = "S", ann = FALSE)
```



```
# Probability for exactly 3 successes out of 10 trials
dbinom(x = 3, size = 10, prob = 0.5)
```

```
## [1] 0.1171875
```

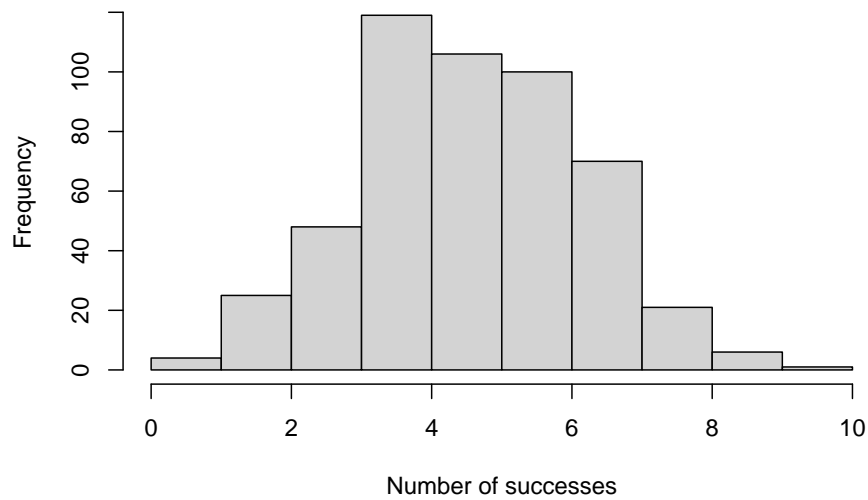
```
# Probability for up to 3 successes out of 10 trials
pbinom(q = 3, size = 10, prob = 0.5)
```

```
## [1] 0.171875
```

```
# Simulate 10 times how many successes there is using random numbers
rbinom(n = 10, size = 10, prob = 0.5)
```

```
## [1] 6 6 6 6 7 4 7 4 5 5
```

```
# With large enough n, expected value (np = 5) should become visible
rbinom500 <- rbinom(n = 500, size = 10, prob = 0.5)
hist(rbinom500, xlab = "Number of successes", ylab = "Frequency", main = NULL)
```



```
summary(rbinom(500))
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   4.00   5.00   5.05   6.00   10.00
```

Poisson distribution

See page 204.

Useful when estimating amounts in random processes where the expected value (λ) is known

$$X \sim Poi(\lambda)$$

PMF

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots, n$$

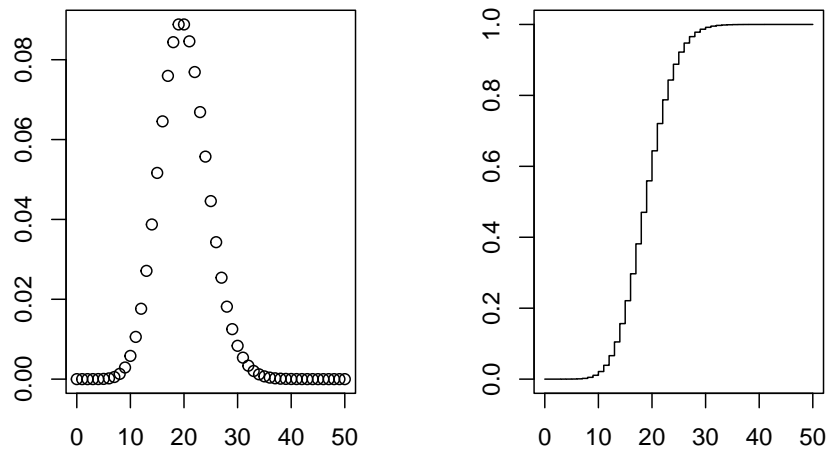
Note that the Poisson distribution can be used to approximate the binomial distribution when n is large and p is small. In this case, $\lambda = np$ and

$$P(X = k) = \frac{(np)^k e^{-np}}{k!}$$

```
# Sequence for visualization
poisson_seq <- seq(0, 50, by = 1)
```

```
# Functions
poisson_pmf <- dpois(x = poisson_seq, lambda = 20)
poisson_cdf <- ppois(q = poisson_seq, lambda = 20)
```

```
# Plot
par(mfrow = c(1, 2))
plot(poisson_seq, poisson_pmf, ann = FALSE)
plot(poisson_seq, poisson_cdf, type = "S", ann = FALSE)
```



```
# A bridge is crossed by 12 people per minute, on average.
# What is the probability that exactly 16 people crosses it in a minute?
dpois(x = 16, lambda = 12)
```

```
## [1] 0.05429334
```

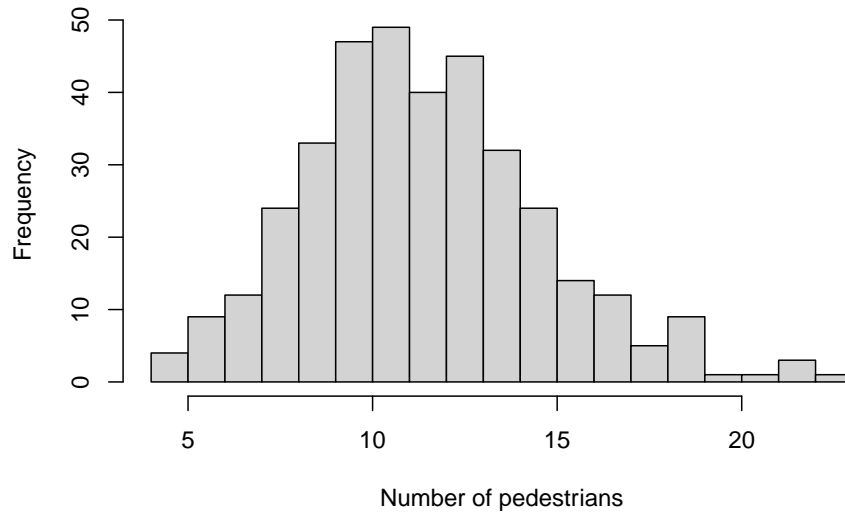
```
# A bridge is crossed by 12 people per minute, on average.
# What is the probability that up to 16 people crosses it in a minute?
ppois(q = 16, lambda = 12)
```

```
## [1] 0.898709
```

```
# Simulate amount of people per minute 10 times
rpois(n = 10, lambda = 12)
```

```
## [1] 11 7 10 9 15 13 15 9 16 9
```

```
# By repeating the observation once every day for a year,
# the expected value (lambda = 12) should become visible
poisson_pedestrians <- rpois(n = 365, lambda = 12)
hist(poisson_pedestrians, breaks = 20,
     xlab = "Number of pedestrians", ylab = "Frequency", main = NULL)
```



```
summary(poisson_pedestrians)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      4.00   10.00   12.00   11.91   14.00   23.00
```

Continuous distributions

For random variables with infinite number of possible values.

Generic definitions

See page 208.

The probability that an event happens between an interval $[a, b]$ can be calculated from the Probability Density Function (PDF) $f(x)$ as an integral

$$P(a \leq X \leq b) = \int_a^b f(x) \, dx$$

Note that the point probabilities for continuous random variables are intrinsically zero

$$\int_a^a f(x) \, dx = 0$$

CDF

$$P(X \leq t) = \int_{-\infty}^t f(x) \, dx$$

Normal distribution

See page 210.

Many phenomena in nature follow the normal distribution. Additionally, so called central limit theorem states that the averages of idenpendet random samples drawn from a population approximately follow the normal distribution - even if the population is better desccribed by some other distribution!

Normal distribution is defined by expected value (and median and mode due to symmetry) μ and (standard) deviation σ .

$$X \sim N(\mu, \sigma^2)$$

PDF

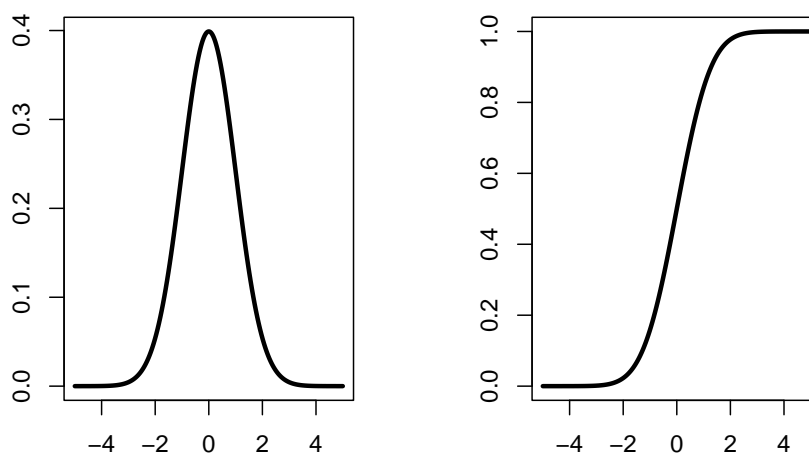
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

In the following code examples, $\mu = \text{mean}$, $\sigma = \text{sd}$.

```
# Sequence for visualization
normal_seq <- seq(-5, 5, by = 0.01)

# Functions
normal_pmf <- dnorm(x = normal_seq, mean = 0, sd = 1)
normal_cdf <- pnorm(q = normal_seq, mean = 0, sd = 1)

# Plot
par(mfrow = c(1, 2))
plot(normal_seq, normal_pmf, type = "l", lwd = 3, ann = FALSE)
plot(normal_seq, normal_cdf, type = "l", lwd = 3, ann = FALSE)
```



```
# The average length of a zebrafish is 22 mm with a standard deviation of 1 mm.
# How likely is a fish 23 mm long?
dnorm(x = 23, mean = 22, sd = 1)
```

```
## [1] 0.2419707
```

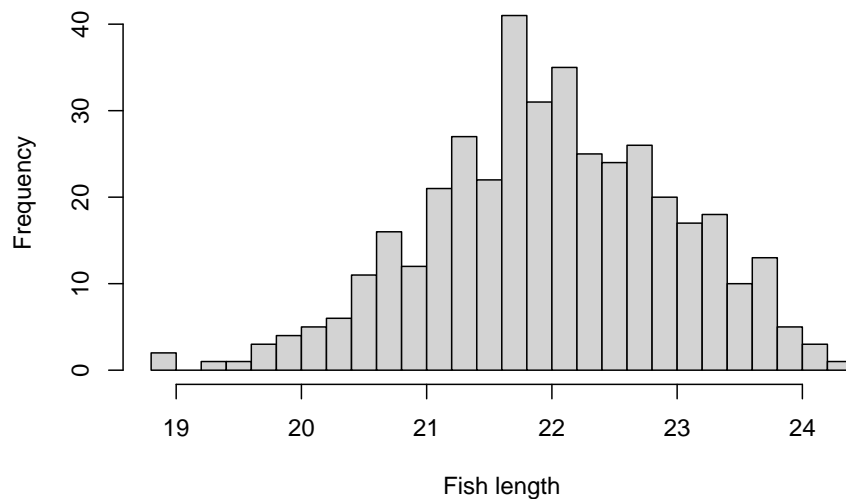
```
# What fraction of the fish are up to 23 mm long?
pnorm(q = 23, mean = 22, sd = 1)
```

```
## [1] 0.8413447
```

```
# Simulate length of five random fish
rnorm(n = 5, mean = 22, sd = 1)
```

```
## [1] 21.15717 22.14971 22.58035 23.01931 23.66488
```

```
# With large enough n, expected value should become visible
normal_zebrafish <- rnorm(n = 400, mean = 22, sd = 1)
hist(normal_zebrafish, breaks = 20,
      xlab = "Fish length", ylab = "Frequency", main = NULL)
```



```
summary(normal_zebrafish)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  18.91   21.35   21.98   21.99   22.70   24.37
```

Student's t-distribution

See page 215.

Similar to the normal distribution, Student's t distribution can be used to estimate the expected value of a population using large enough amount of sample averages.

The Student's t distribution is defined by just one parameter, namely the degrees of freedom (DoF) ν . Compared to the normal distribution, the advantage here is that prior knowledge of parameters μ and σ is not needed - which oftentimes is the case. Additionally, with a large ν , the Student's t distribution approaches the standard normal distribution $N(0,1)$.

PDF is defined using the Gamma function $\Gamma(n) = (n-1)!$.

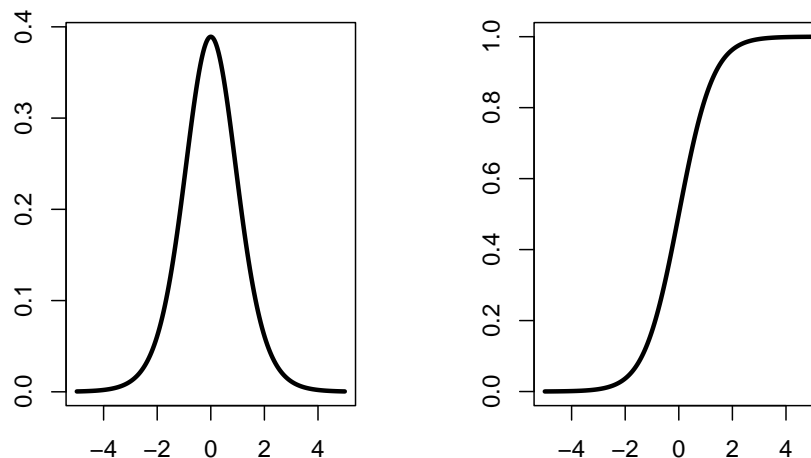
$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

In the following code examples, $\nu = \mathbf{df}$.

```
# Sequence for visualization
students_t_seq <- seq(-5, 5, by = 0.01)

# Functions
students_t_pmf <- dt(x = students_t_seq, df = 10)
students_t_cdf <- pt(q = students_t_seq, df = 10)

# Plot
par(mfrow = c(1, 2))
plot(normal_seq, students_t_pmf, type = "l", lwd = 3, ann = FALSE)
plot(normal_seq, students_t_cdf, type = "l", lwd = 3, ann = FALSE)
```




```
# Same as in the earlier examples  
dt(x = 2, df = 20)
```

```
## [1] 0.05808722
```

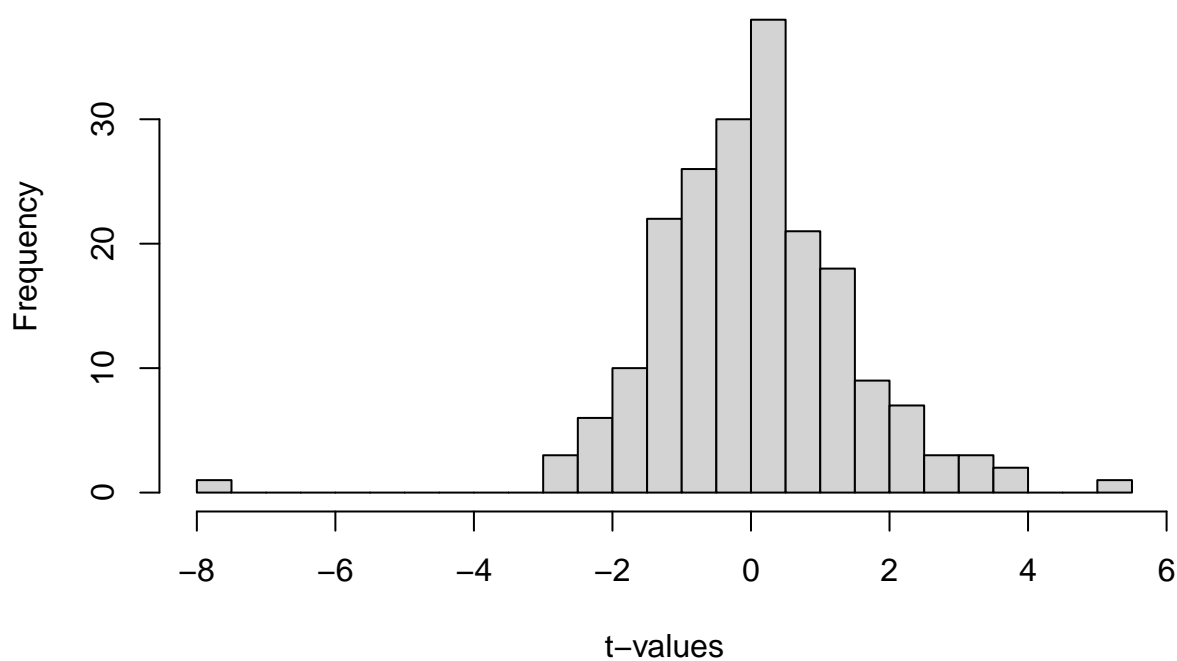
```
pt(q = 2, df = 20)
```

```
## [1] 0.9703672
```

```
rt(n = 5, df = 20)
```

```
## [1] -0.51208523 -0.08958681 0.26896335 -1.20457221 -0.25909449
```

```
# The mean should be around 0 regardless of DoF  
student_values <- rt(n = 200, df = 5)  
hist(student_values, breaks = 20,  
      xlab = "t-values", ylab = "Frequency", main = NULL)
```



```
summary(student_values)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
## -7.93293 -0.82388  0.06036  0.05325  0.84955  5.34079
```