Discrete distributions

Probabilities and statistics for biology (CMB STAT1 - STAT2)

Jacques van Helden

2020-02-20

Negative binomial for over-dispersed counts

Discrete distributions of probabilities

The expression ***discrete distribution*** denotes probability distribution of variables that only take discrete values (by opposition to continuous distributions).

Notes:

- In probabilities, the observed variable (x) usually represents the number of successes of a series of tests, or the counts of some observation. In such cases, its values are natural numbers ($x \in \mathbb{N}$).
- The probability $\mathrm{P}(x)$ takes real values comprised between 0 and 1, but its distribution is said *discrete¨since it is only defined fora set of discrete values of X. It is generally represented by a step function.

Geometric distribution

Application: waiting time until the first appeearance of an event in a Bernoulli schema.

Examples:

- In a series of dices rollings, count the number rolls (x) before the first occurrence of a 6 (this occurrence itself is not taken into account).
- Length of a DNA sequence before the first occurrence of a cytosine.

Mass function of the geometric distribution

The ***Probability Mass Function*** (***PMF***) indicates the probability to observe a particular result.

For the geometric distribution, it indicates the probability to observe exactly x failures before the first success, in a series of independent trials with a probability of success p.

$$\mathrm{P}(X = x) = (1 - p)^x \cdot p$$

Justification:

- The probability of failure for the first trial is $q = 1 - p$ (complementary events).
- Bernoulli schema $\rightarrow$ the trials are independent $\rightarrow$ the probability of the series is the product of probabilities of its successive outcomes.
- One thus computes the product of probabilities of the x initial

Geometric PMF

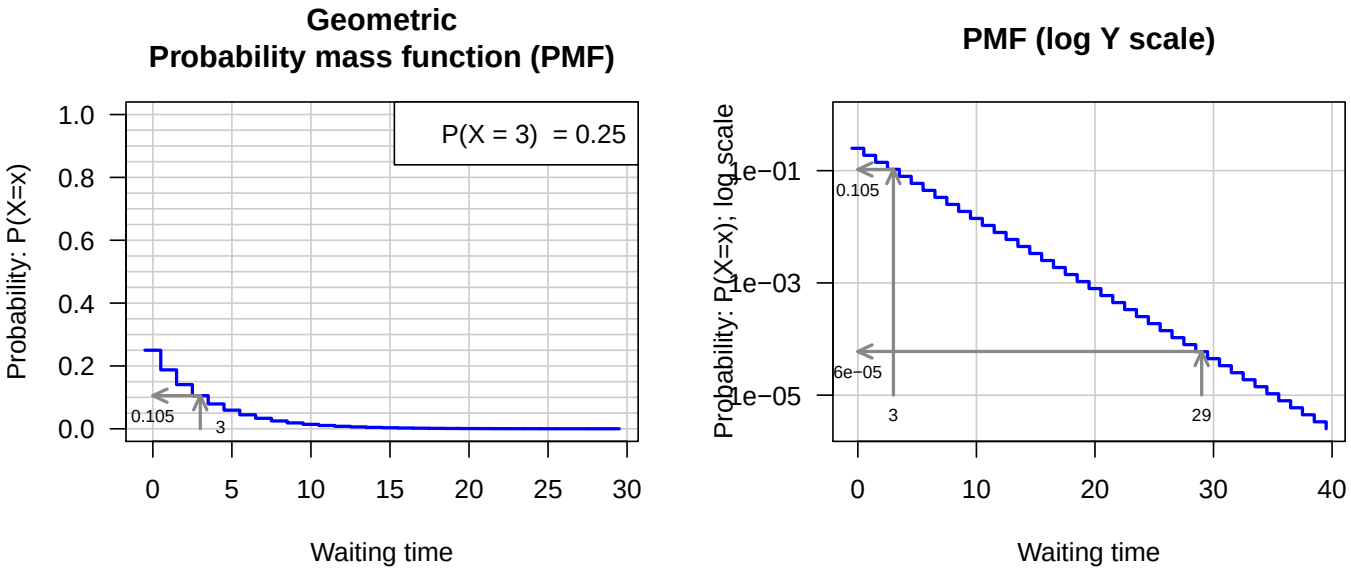

Figure 1: **Fonction de masse de la loi géométrique**. Gauche: ordonnée en échelle logarithmique.

Distribution tails and cumulative distribution function

The ***tails*** of a distribution are the areas comprised under the density curve up to a given value (***left tail***) or staring from a given value (***right tail***).

- The ***right tail*** indicates the probability to observe a result (X) **smaller than or equal to** a given value (x): $P(X \leq x)$.
  - **Definition:** the ***Cumulative Density Function*** (***CDF***) $P(X \leq x)$ indicates the probability for a random variable X to take a value smaller than or equal to a given value (x). It corresponds to the left tail of the distribution (including the x value).
- The **left tail** of a distribution indicates the probability to observe a result **higher than or equal to** a given value: $P(X \geq x)$.

Distribution tails and cumulative distribution function

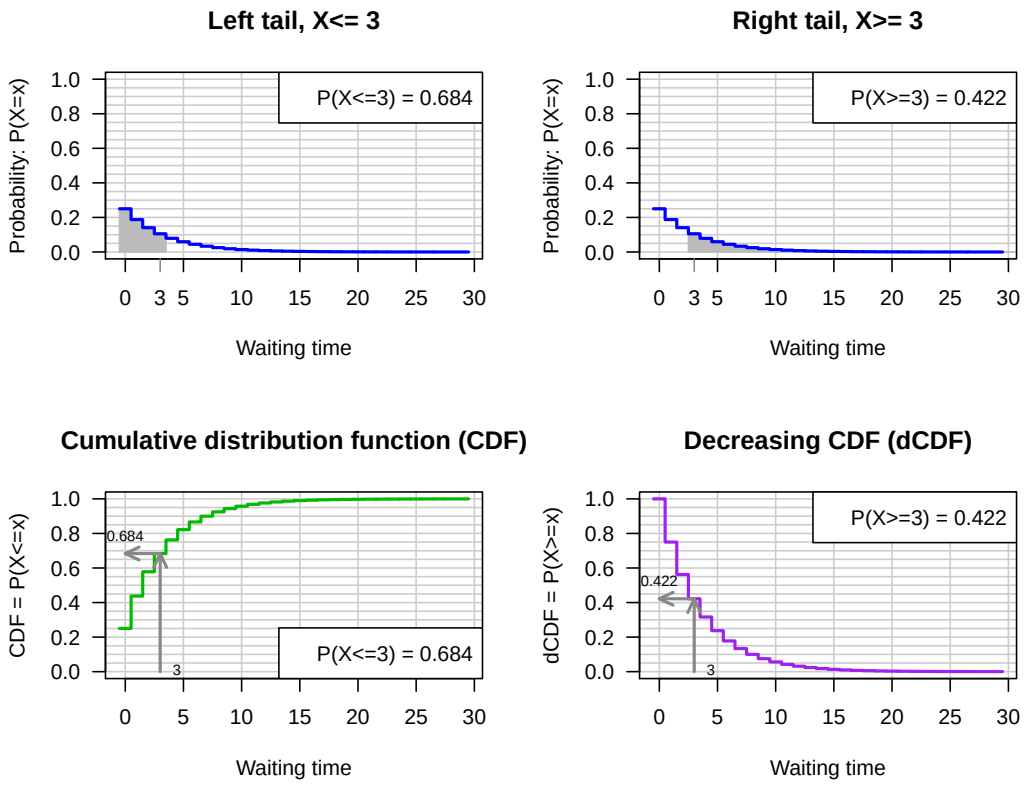

Figure 2: **Tails and Cumulative Density Function of the geometric distribution**.

Binomial distribution

The **binomial distribution** indicates the probability to observe a given number of successes (x) in a series of n independent trials with constant success probability p (Bernoulli schema).

Binomial PMF

$$\mathrm{P}(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} = C_n^x p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^n$$

Binomial CDF

$$\mathrm{P}(X \geq x) = \sum_{i=x}^{n} P(X = i) = \sum_{i=x}^{n} C_n^i p^i (1-p)^{n-i}$$

Properties

i-shaped binomial distribution

The binomial distribution can take various shapes depending on the values of its parameters (success probability p, and number of trials n).

When the expectation $(p \cdot n)$ is very small, the binomial distribution is monotonously decreasing and is qualified of *i*-***shaped***.

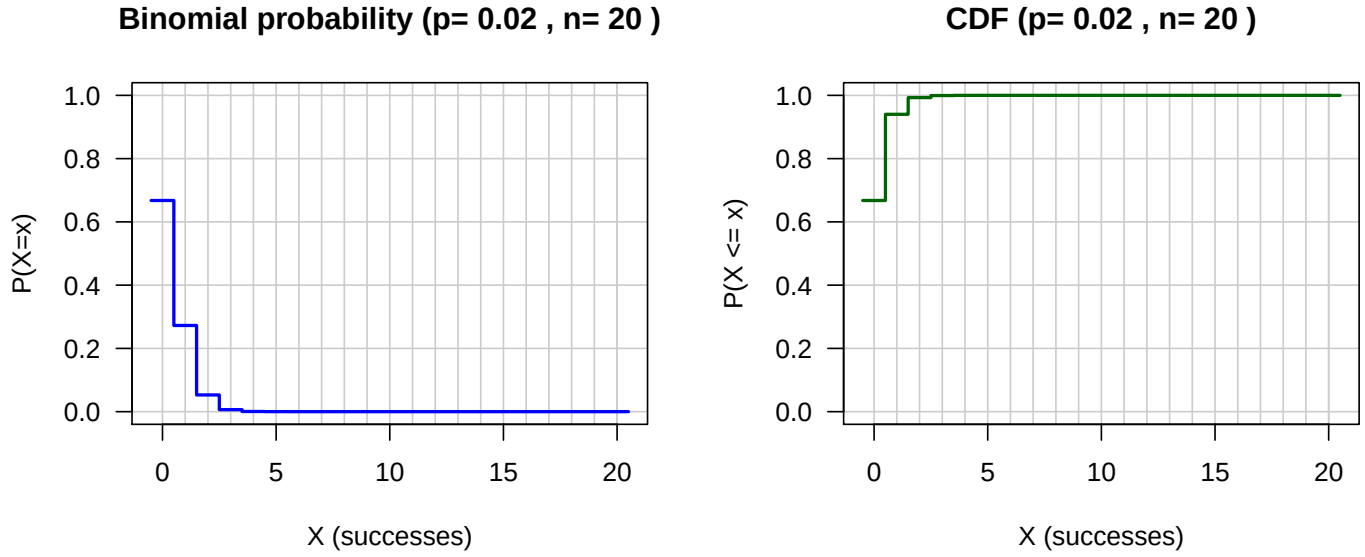

Figure 3: Distribution binomiale en forme de i.

Asymmetric bell-shaped binomial distribution

When the probability is relatively high but still lower than 0.5, the distribution takes the shape of an asymmetric bell.

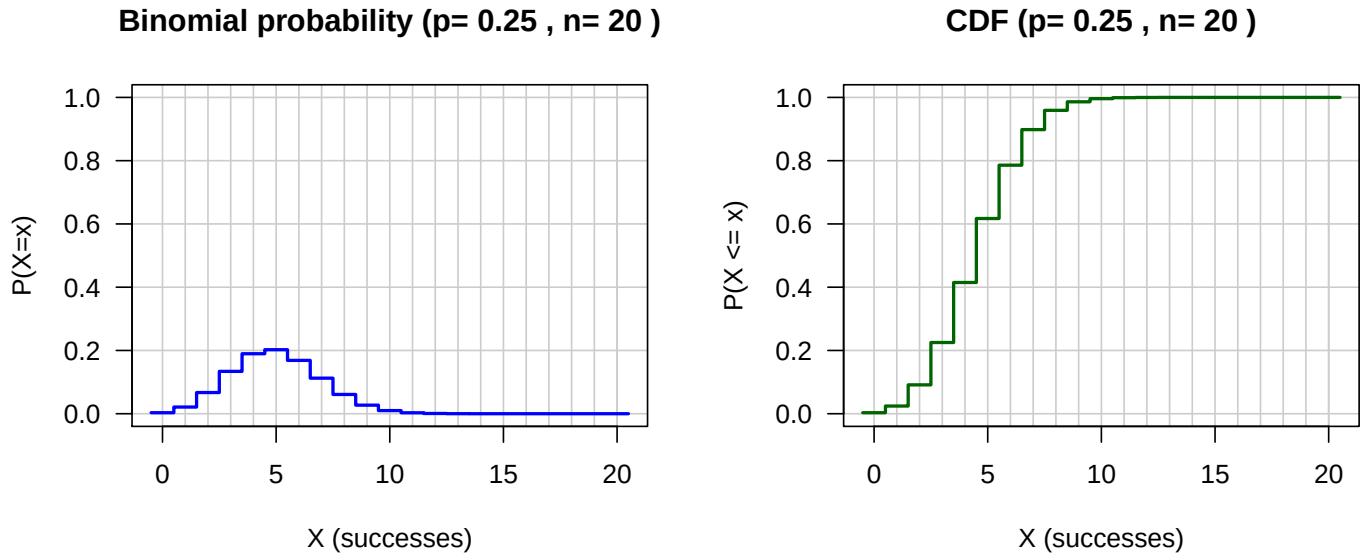

Figure 4: Distribution binomiale en forme de cloche asymétrique.

Symmetric bell-shaped binomial

When the success probability p is exactly 0.5, the binomial distribution takes the shape of a symmetrical bell.

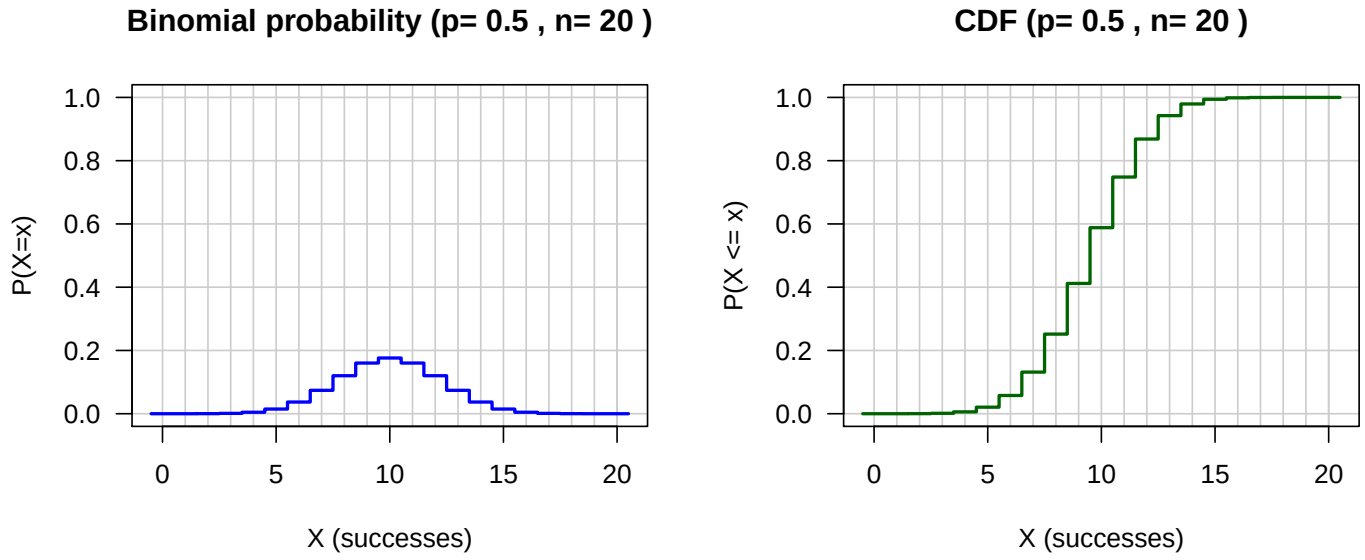

Figure 5: Distribution binomiale en forme de cloche symétrique (p=0.5).

j-shaped binomial distribution

Then the success probability is close to 1, the distirbution is monotonously increasing and is qualified of ***j-shaped distribution.

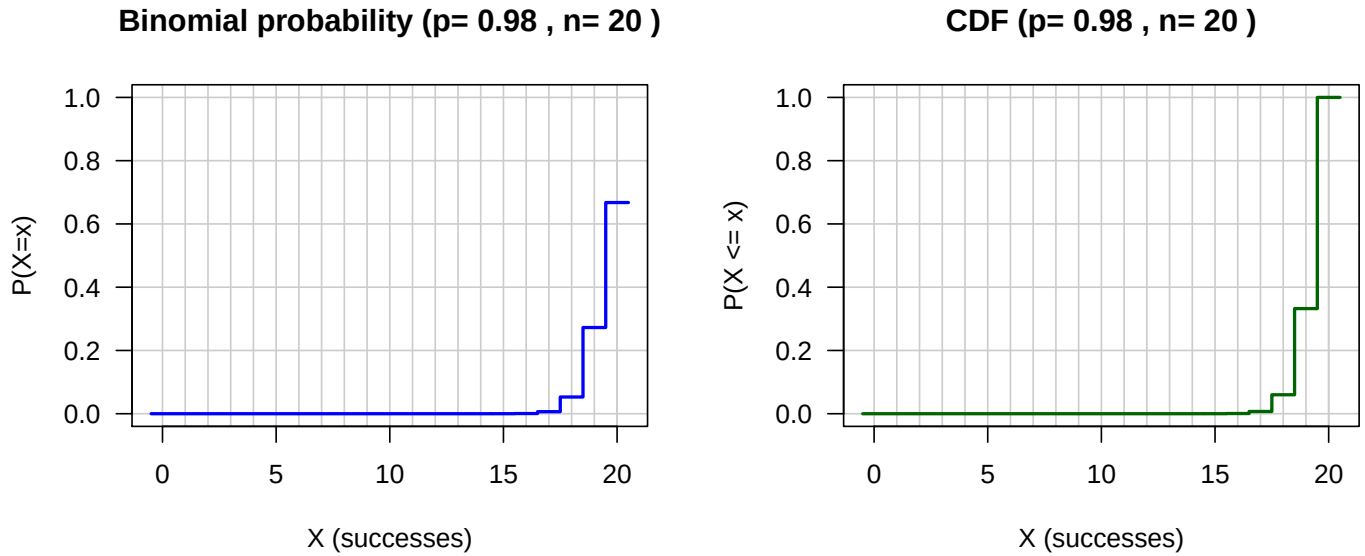

Figure 6: Distribution binomiale en forme de j.

Examples of applications of the binomial

1. **Dices**: number of 6 observed during a series of 10 dice rolls
2. **Sequence alignment**: number of identities between two sequences alignmed without gap and with an arhbitrary offset.
3. **Motif analysis**: number of occurrences of a given motif in a genome.

Note: the binomial assumes a Bernoulli schema. Forexamples 2 and 3 this amounts to consider that nucleotides are concatenated in an independent way, which is quite unrealistic.

Poisson law

The Poisson law describes the probability of the number of realisations of an event during a fixed time interval, assuming that the average number of events is constant, and that the events are independent (previous realisations do not affect the probabilities of future realisations).

Poisson Probability Mass Function

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

- x is the number of event realisations
- λ (Greek letter "lambda") represents the expectation, i.e. the average number of occurrences that would be obtained by running the same test an infinite number of times;
- e is the exponential base ($e = 2.718$).

Properties of the Poisson distribution

- **Expectation** (number of realisations expected by chnace): $< X >= \lambda$ (by construction)
- **variance**: $\sigma^2 = lambda$ (***the variance equals the mean!***)
- **Standard deviation**: $\sigma = \sqrt{\lambda}$

Application: mutagenesis

- A bacterial population is submitted to a mutagen (chemical agent, irradiations). Each cell is affected by a particular number of mutations.
- Taking into account the dosis of the mutagen (exposure time, intensity, concentration) one could take an empirical measure of the mean number of mutations by individual (expectation, λ).
- The Poisson law can be used to describe the probability for a given cell to have a given number of mutations ($x = 0, 1, 2, ...$).

Historical experiment by Luria-Delbruck (1943)

In 1943, Salvador Luria and Max Delbruck demonstrated that when cultured bacteria are treated by an antibiotic, the mutations that confer resistance are not induced by the antibiotic itself, but preexist. Their demonstration relies on the fact that the number of antibiotic-resistant cells follows a Poisson law (Luria & Delbruck, 1943. Genetics 28:491–511)

Convergence of the binomial towards the Poisson

Under some circumstances, the binmial law converges towards a Poisson.

- very small probability of success ($p \ll 1$)
- large number of trials (n)

TO DO

Netative binomial: number of successes before the r^{th} failure

The **negative binomial** distribution (also called **Pascal distribution**) indicates the probability of the number of successes (k) before the r^{th} failure, in a Bernoulli schema with success probability p.

$$\mathcal{NB}(k|r,p) = \binom{k+r-1}{k} p^k (1-p)^r$$

This formula is a simple adaptation of the binomial, with the difference that we know that the last trial must be a failure. The binomial coefficient is thus reduced to choose the k successes among the $n-1 = k+r-1$ trials preceding the r^{th} failure.

Negative binomial: alternative formulations

It can also be adapted to indicate related probabilities.

- Number of **failures** (r) before the k^{th} **success**.

$$\mathcal{NB}(r|k,p) = \binom{k+r-1}{r} p^k (1-p)^r$$

- Number of **trials** ($n = k + r - 1$) before the r^{th} **failure**.

$$\mathcal{NB}(n|r,p) = \binom{n-1}{r-1} p^{n-r} (1-p)^r$$

Negative binomial density

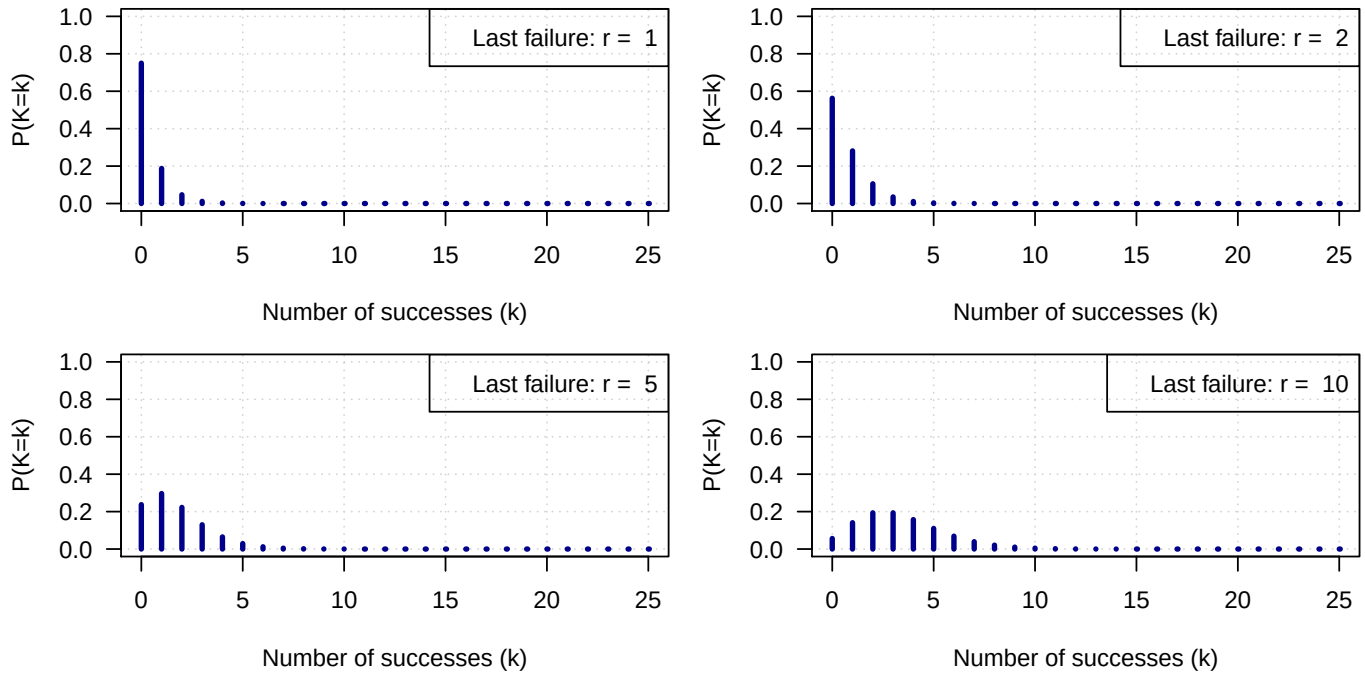

Figure 7: Negative binomial.

Properties of the negative binomial

The variance of the negative binomial is higher than its mean. It is therefore sometimes used to model distributions that are over-dispersed by comparisong with a Poisson.

$$\mathcal{NB}(r|k,p) = \binom{k+r-1}{r} p^k (1-p)^r$$

- Parameters:
  - p: probability of success at each trial
  - r: number of failures
  - k: number of successes before the r^{th} failure
- Mean: $\frac{pr}{1-p}$
- Variance: $\frac{p(1-p)}{p^2}$

Exercise – Negative binomial

Each student chooses a value for the maximal number of failures (r).

1. Read carefully the help of the negative binomial functions: `help(NegBinomial)`
2. **Random sampling**: draw of $rep = 100000$ random numbers from a negative binomial distribution (`rndbinom()`) to compute the distribution of the number of successes (k) before the r^{th} failure.
3. Compute the expected mean and variance of the negative binomial.
4. Compute the mean and variance from your sampling distribution.
5. Draw an histogram with the number of successes before the r^{th} failure.
6. Fill up the form on the collective result table

Solution to the exercise – negative binomial

```
r <- 6          # Number of failures
p <- 0.75       # Failure probability
rep <- 100000
k <- rnbinom(n = rep, size = r, prob = p)
max.k <- max(k)
exp.mean <- r*(1 - p)/p
rand.mean <- mean(k)
exp.var <- r*(1 - p)/p^2
rand.var <- var(k)
hist(k, breaks = -0.5:(max.k + 0.5), col = "grey", xlab = "
     las = 1, ylab = "", main = "Random sampling from negat
abline(v = rand.mean, col = "darkgreen", lwd = 2)
abline(v = exp.mean, col = "green", lty = "dashed")
arrows(rand.mean, rep/20, rand.mean + sqrt(rand.var), rep/2
       angle = 20, length = 0.1, col = "purple", lwd = 2)
text(x = rand.mean, y = rep/15, col = "purple",
```

Negative binomial for over-dispersed counts

Exercises

- html
- pdf
- Rmd