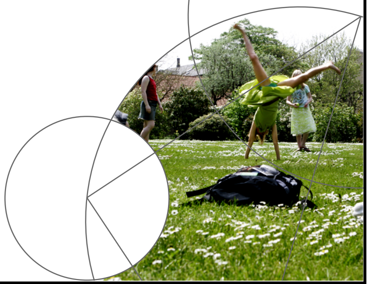
Faculty of Life Sciences

The binomial distribution

Claus Ekstrøm & Ib Skovgaard
E-mail: ims@life.ku.dk



Program

- Independent trials
- The binomial distribution
  - Estimation, mean and standard deviation
  - Normal approximation
  - Inference for the binomial distribution
- Comparison of to binomial distributions

Independent trials

- n trials
- Two possible results: success, failure
- Same probability of success, p.
- Trials are *independent*

Example: germination of seeds

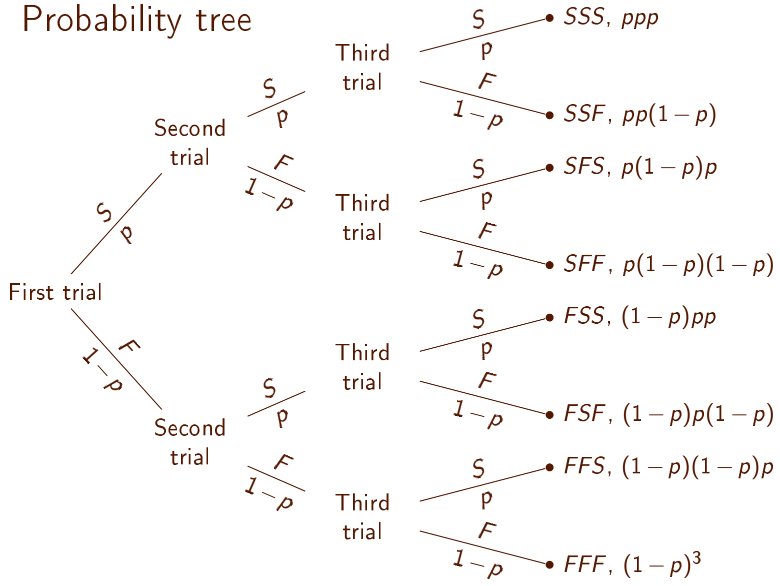
Assume that each seed has germination probability 0.6

Consider $n = 3$ seeds.

- What is the probability for precisely one of the seeds to germinate?
- What is the probability for at least one seed to germinate?

What are the probabilities for larger n?

Probability tree



The binomial distribution

Let Y denote the number of successes from n independent trials.
Then the binomial distribution is given by

$$P(j \text{ "successes"}) = P(Y = j) = \binom{n}{j} \cdot p^j \cdot (1-p)^{n-j},$$

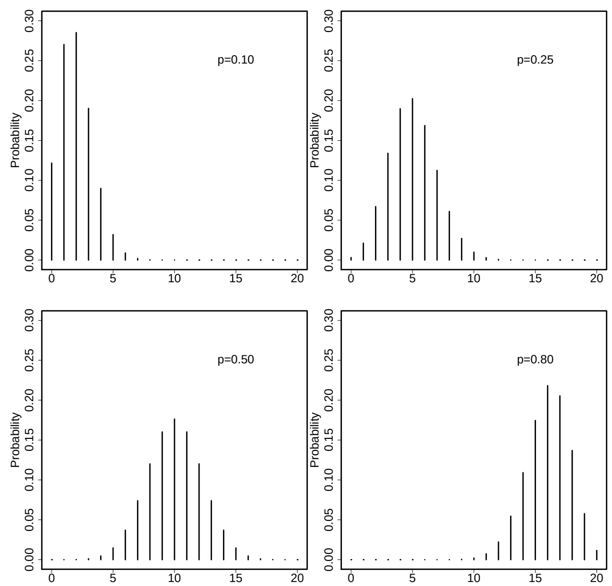
where the binomial coefficients are defined as

$$\binom{n}{j} = \frac{n!}{j!(n-j)!}$$

We write

$$Y \sim \text{bin}(n, p)$$

Binomial distributions



Mean, variance and standard deviation

For a binomially distributed variable $Y \sim \text{bin}(n, p)$

the mean is

$$EY = n \cdot p,$$

the variance is

$$\text{var}(Y) = n \cdot p \cdot (1-p),$$

and hence the standard deviation is

$$\text{SD}(Y) = \sqrt{n \cdot p \cdot (1-p)}.$$

Note that Y is a sum of n binomials with $n = 1$ (n zero-one trials), and this agrees with the rule of adding means and variances.

Example: CG islands in DNA

In parts of the DNA consisting occurrence of the duo CG among the sequences of bases A, C, G and T, is more frequent than in most parts. These parts of the DNA are called CG-islands.

Suppose that the normal frequency of the duo CG is 0.06, and that a certain sample contains 125754 duos. If 7800 CG duos are found in the sample, does that indicate that the CG-frequency is elevated in the sample?

What is $P(Y \geq 7800)$?

UNIVERSITY OF COPENHAGEN FACULTY OF LIFE SCIENCES

Approximation with the normal distribution

This approximation may be useful when n is large. $\text{bin}(n,p)$ is approximated by $N(np, np(1-p))$:

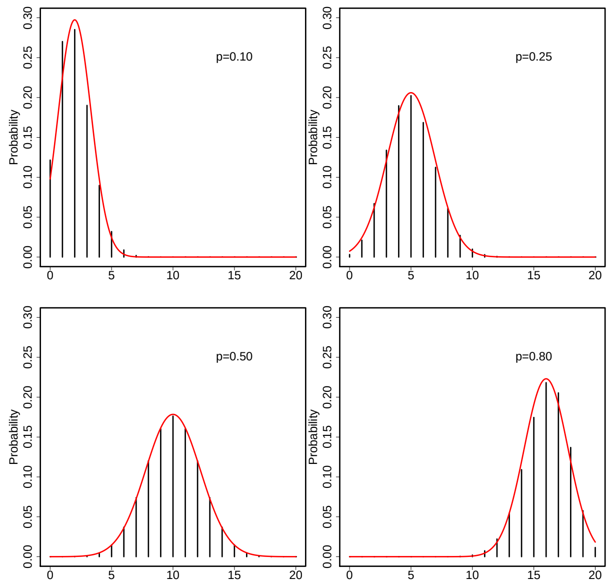
$$P(Y \leq y) \approx \Phi\left(\frac{(y+0.5)-np}{\sqrt{np(1-p)}}\right)$$

If $Y \sim \text{bin}(125754, 0.06)$ this gives

$$\begin{aligned} P(Y \geq 7800) &\approx 1-\Phi\left(\frac{(7800-0.5)-125754\cdot 0.06}{\sqrt{125754\cdot 0.06(1-0.06)}}\right) \\ &= 1-0.9987 = 0.0013. \end{aligned}$$

UNIVERSITY OF COPENHAGEN FACULTY OF LIFE SCIENCES

Binomial approximations



UNIVERSITY OF COPENHAGEN FACULTY OF LIFE SCIENCES

Inference

The estimate of the parameter p is

$$\hat{p} = \frac{\text{number of successes}}{\text{number of trials}}.$$

with corresponding standard error

$$\text{SE}(\hat{p}) = \frac{s}{\sqrt{n}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Hence, a confidence interval for the probability of "success", p, is

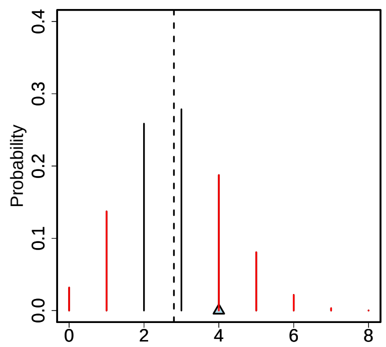
$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Inference — hypothesis testing

Alternative to confidence interval: test the hypothesis $H_0: p = p_0$.
Results, that are more extreme than the observed gives the p-value:

$$p\text{-value} = \sum_{\text{Extreme } y\text{'s}} P(Y = y)$$

where the sum is over y's shown in red in the graph, (namely those with a smaller H_0-probability than the observed).



Gender of offspring from speckled bear

Over a decade 63 speckled bears were born in European zoo's. Of these 40 were males and 23 were females.

Problem: Could this be due to chance if the two genders were equally likely?

For Brown bear 6 males and 12 females were born.

Comparison of two proportions

Lad $Y_1 \sim \text{bin}(n_1, p_1)$ and $Y_2 \sim \text{bin}(n_2, p_2)$. Interested in the difference

$$p_1 - p_2$$

We know that

$$\hat{p}_1 = \frac{y_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{y_2}{n_2}.$$

Hence,

$$\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2}.$$

Da Y_1 and Y_2 is independent is

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$$

Comparison of two proportions — II

Using

$$\text{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

we get the 95% confidence interval for the difference between the two proportions, $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Animal offspring III

To investigate the stability of the many male-births for Kirk's dik-dik, a comparison was made with a previous study.

	Males	Females
Last 10 years	154	96
Previous study	169	134

Do the two studies seem to agree?

Lecture summary: main points

- The binomial distribution
  - When should it be used?
  - Estimation, confidence intervals, tests of hypotheses
  - Approximation by the normal distribution
- Comparison of success probabilities for two binomials.