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 two 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 \( n \)?
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

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

Mean, variance and standard deviation

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

the mean is

$$EY = np.$$

the variance is

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

and hence the standard deviation is

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

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 $P(Y \geq 7800)$?

### Inference

The estimate of the parameter $p$ is

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

with corresponding standard error

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

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:

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

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

Let \( Y_1 \sim \text{bin}(n_1, p_1) \) 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,

\[
\hat{p}_1 - \hat{p}_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2}.
\]

Da \( Y_1 \) and \( Y_2 \) is independent is

\[
\text{Var}(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}
\]
To investigate the stability of the many male-births for Kirk’s dik-dik, a comparison was made with a previous study.

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last 10 years</td>
<td>154</td>
<td>96</td>
</tr>
<tr>
<td>Previous study</td>
<td>169</td>
<td>134</td>
</tr>
</tbody>
</table>

Do the two studies seem to agree?