Mathematical Concepts (G6012)

Lecture 20

Thomas Nowotny
Chichester I, Room CI-105
Office hours: Tuesdays 15:00 - 16:45
T.Nowotny@sussex.ac.uk

Binomial distribution

- Is the probability distribution for the number of successes (1s) in a so-called Bernoulli process
- Bernoulli process:

$$\Omega = \{0,1\}^n = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \{0,1\}\}$$

$$P(\{\omega_i = 1\}) = p \qquad P(\{\omega_i = 0\}) = 1 - p$$

And all such events are independent, such that

$$P(\{(\omega_1, \ldots, \omega_n)\}) = p^k (1-p)^{n-k}$$

k = number of 1s, n-k= number of 0s.

Binomial distribution

$$X : \Omega \to \mathbb{R}$$

$$X\big((\omega_1, \ldots, \omega_n)\big) = \sum_{i=1}^{n} \omega_i$$

X is the number of 1s (number of successes)

What is $P(\{X = k\})$?

$$P(\{(\omega_1, \ldots, \omega_n) \in \Omega : \sum_{i=1}^{n} \omega_i = k\})$$

In other words the probability of all elementary events with k 1s and n-k 0s.

Binomial distribution

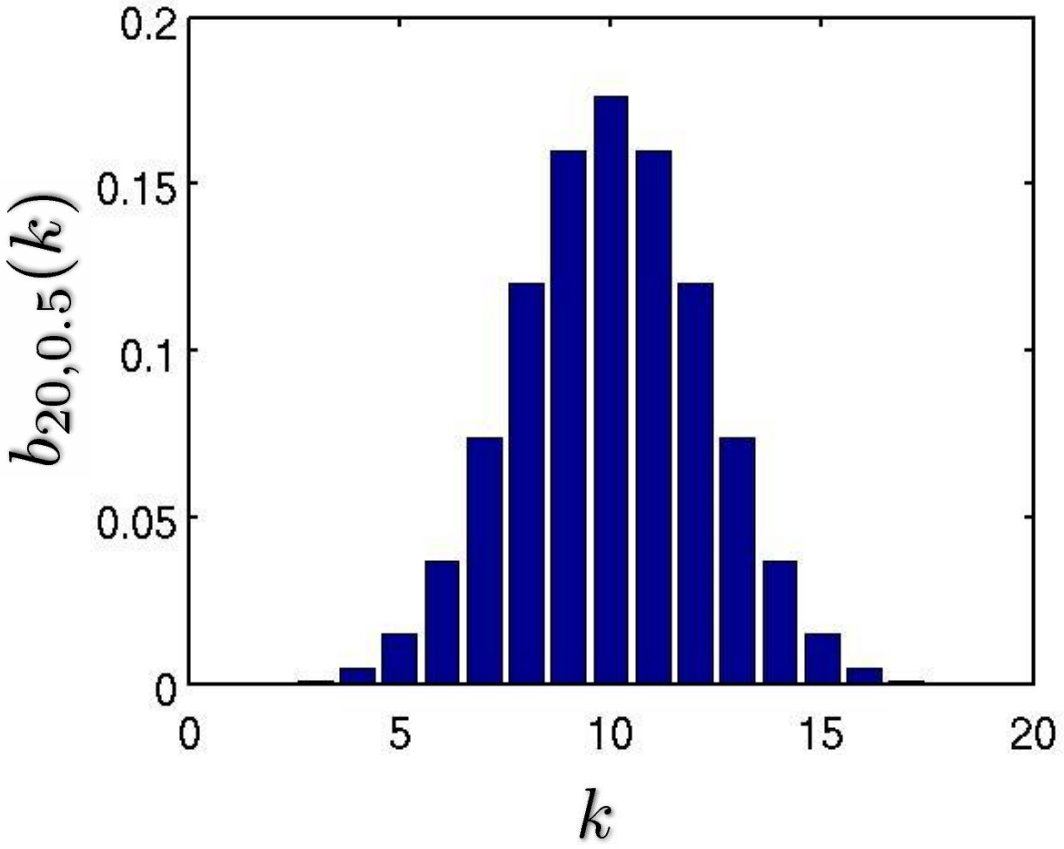
$$P(\{X = k\}) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

$$P(\{X = k\}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial coefficient

The binomial distribution is sometimes written as $b_{n,p}(k)$

Binomial distribution properties



Expectation value

$$\mathbb{E}\sum_{i=1}^{N} x_i = N \cdot p$$

Standard deviation

$$\sigma_{\sum_{i=1}^{N} x_i} = \sqrt{Np(1-p)}$$

Expectation value and maximum are *not* the same

BB Proof of Expectation value

$$\mathbb{E}X = \mathbb{E}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \mathbb{E}x_i$$

$$= \sum_{i=1}^{N} \left(1 \cdot p + 0 \cdot (1-p)\right)$$

$$= \sum_{i=1}^{N} p = N \cdot p$$

BG Proof of Standard Deviation

$$\sigma_X^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$$

$$\mathbb{E}X^2 = \mathbb{E}\sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}x_i x_j$$

$$= \sum_{i=1}^{N} \sum_{\substack{j \in \{1 \ldots N\} \\ j \neq i}} \mathbb{E}x_i x_j + \sum_{i=1}^{N} \mathbb{E}x_i^2$$

These are Independent, i.e. $= p^2$

$= p$

BG Proof of Standard Deviation

$$\mathbb{E}X^2 = (N^2 - N) \cdot p^2 + N \cdot p$$
$$= N^2p^2 + Np - Np^2$$

$$\left(\mathbb{E}X\right)^2 = (Np)^2 = N^2p^2$$

$$\sigma_X^2 = \mathbb{E}X^2 - \left(\mathbb{E}X\right)^2 = Np - Np^2 = Np(1-p)$$

Law of large numbers

- There are several different laws of large numbers.
- Here I would like to show you one example to give a feel for what these laws are about.

Law of large numbers for Bernoulli processes

$$\Omega_n = \{0,1\}^n$$

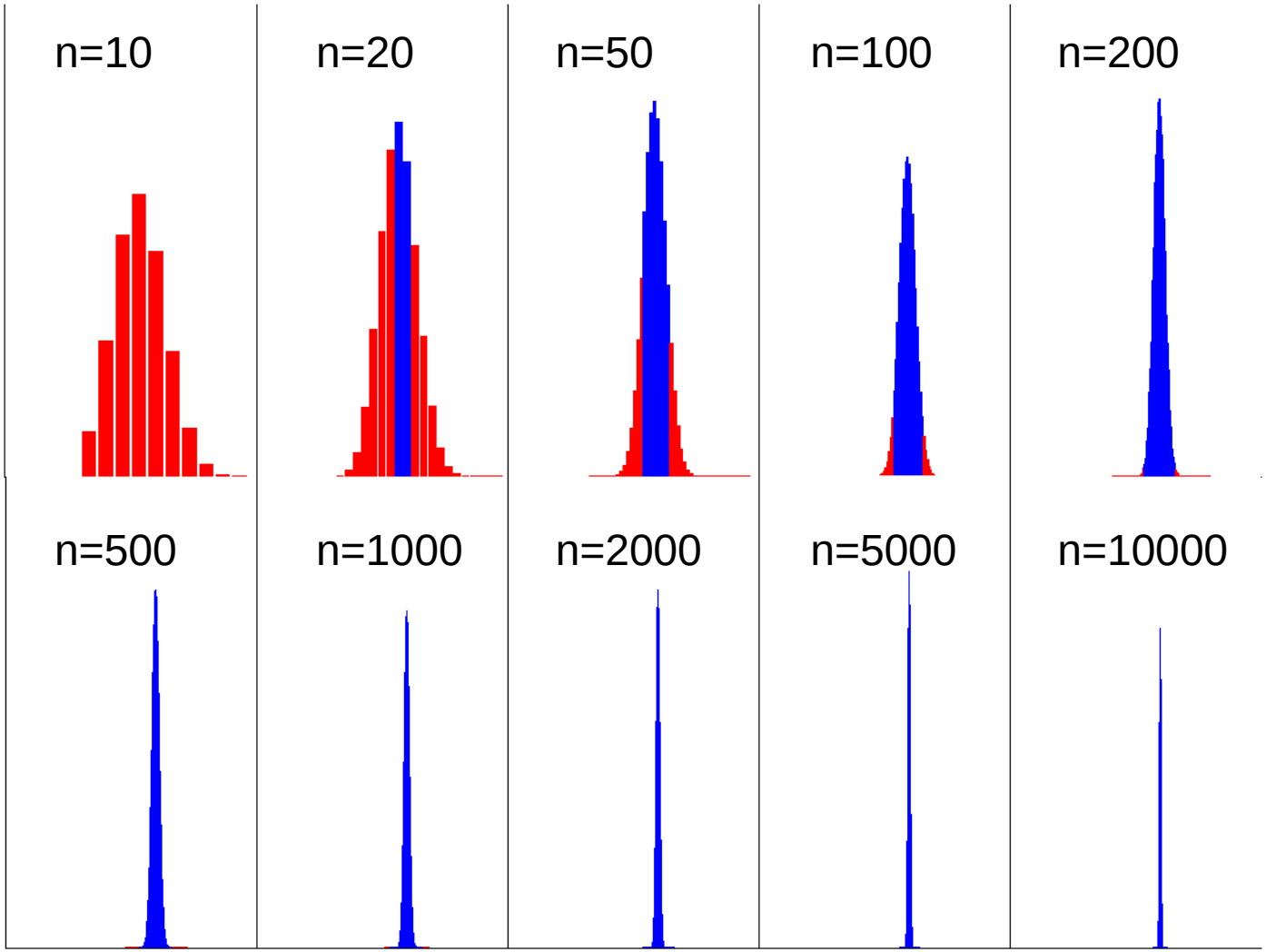
Random variable $X_n = \sum_{i=1}^{n} \omega_i$

has distribution $b_{n,p}(k)$

Consider $x_n = X_n/n$ and $\epsilon > 0$

$P(|x_n - p| \geq \epsilon)$ is the area under the tails:

$P(|x_n - p| \geq \epsilon)$ for $\epsilon = 0.1$ and $p = 0.4$



Law of large numbers

For any $\epsilon > 0$

$$\lim_{n \to \infty} P(|x_n - \mathbb{E}x_n| \geq \epsilon) = 0$$

The probability for x_n to be more than ϵ away from its expectation value $\mathbb{E}x_n = p$ converges to 0 for $n \to \infty$.

STATISTICS

Statistics vs Probability Theory

- In statistics we do not know the underlying probability space (which is the starting point in probability theory)
- We have a number of samples rather than information on the system they originate from
- But: Both are related: We use probability theory to make sense of statistics.

Statistical measures: **mean**

- The **mean** of a set of observations

  $\{x_i\},\ i = 1, \ldots, N$

  is defined as

$$\overline{x} = \langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i$$

  it is also called the **average**.

Statistical measures: Standard deviation

- The **standard deviation** of a set of observations

$$\{x_i\},\ i = 1, \ldots, N$$

is defined as

$$std(x) = \sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

More on Standard Deviation

- This is the same name as the standard deviation in probability theory
- It is not exactly the same thing though
- The confusion about this is large (even among mathematicians)

- The truth is that "$\sigma_{\text{Statistics}}$" is an **estimator** of an underlying "$\sigma_{\text{Probability}}$".

Why the N-1 term?

- Most people put 1/(N-1) to achieve what is called an unbiased estimator:

$$\mathbb{E}(\sigma_x^{\mathrm{stat}})^2 = \mathbb{E}\left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2\right) = (\sigma_X^{\mathrm{prob}})^2$$

if x_i originate from observing the random variable X .

BG Proof of unbiased estimator

Assume N independent, identically distributed random variables x_i :

$$\mathbb{E}\,\hat{\sigma}_x^2 = \mathbb{E}\left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2\right)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\mathbb{E}\left(x_i-\frac{1}{N}\sum_{j=1}^{N}x_j\right)\left(x_i-\frac{1}{N}\sum_{k=1}^{N}x_k\right)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbb{E}x_i^2-\mathbb{E}\frac{1}{N}\sum_{j=1}^{N}x_ix_j-\mathbb{E}\frac{1}{N}\sum_{k=1}^{N}x_ix_k+\mathbb{E}\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}x_jx_k\right)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbb{E}x_i^2-\frac{2}{N}\sum_{j\neq i}^{N}\mathbb{E}x_i\mathbb{E}x_j-\frac{2}{N}\mathbb{E}x_i^2+\frac{1}{N^2}\sum_{j=1}^{N}\left(\sum_{k\neq j}^{N}\mathbb{E}x_j\mathbb{E}x_k+\mathbb{E}x_j^2\right)\right)$$

BG Proof of unbiased estimator

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbb{E}x_i^2 - \frac{2(N-1)}{N}(\mathbb{E}x_i)^2 - \frac{2}{N}\mathbb{E}x_i^2 + \frac{1}{N^2}\sum_{j=1}^{N}\left((N-1)(\mathbb{E}x_j)^2 + \mathbb{E}x_j^2\right)\right)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{N}{N}\mathbb{E}x_i^2 - \frac{2(N-1)}{N}(\mathbb{E}x_i)^2 - \frac{2}{N}\mathbb{E}x_i^2 + \frac{N-1}{N}(\mathbb{E}x_i)^2 + \frac{1}{N}\mathbb{E}x_i^2\right)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{N-1}{N}\mathbb{E}x_i^2 - \frac{N-1}{N}(\mathbb{E}x_i)^2\right)$$

$$= \mathbb{E}x_i^2 - (\mathbb{E}x_i)^2 = \sigma_x^2$$

Therefore, the estimator with 1/(N-1) is unbiased.

Common use of σ

- It is common to give "error bars" of $\pm\sigma$
  This means that we can expect further observations to be within the error bar with probability 68.3% *
- My climbing gear is tested such that it does not fail within 3σ around the specification,
  this implies the probability for it to fail is < 0.27%

* If we can assume a so-called Gaussian distribution

Statistical measures: **median**

- The median of a set of observations $\{x_i\},\ i = 1, \ldots, N$

  is the value x_j such that half of the x_i are larger and half of them are smaller than x_j (if the sample set has an even number of samples we take the middle (average) of the "middle samples")

Median: Examples

- $\{1, 2, 3, 4, 5, 6\}$ the median is: 3.5

  (the mean is: 3.5)

- $\{0, 0, 1, 1, 2, 2, 10\}$ the median is: 1

  (the mean is: $\frac{16}{7}$)

Mean and median are not the same! (not even close)

Median versus Mean

- Some people think the median is a better measure for a “typical value” than the mean – why?
- Because outliers can greatly change the mean but not the median:
  Consider a distribution of house prices, all are around 250k, but there is one castle for 10M
  …

STATISTICAL TESTS

Statistical Tests: Hypothesis testing

- The idea is that you have an observation of values $\{x_i\}$ nd a null-Hypothesis H
- E.g. you are playing chess against your boy/girlfriend and observe that you win 10 out of 15 games (observation).
  Your null-hypothesis is that you have a chance of 50% to win.
- Can you reject the hypothesis based on the observation?

In hypothesis testing you set a "significance niveau" α , e.g. $\alpha = 0.05 = 5\%$

Then you calculate the probability P of your observation or a more extreme one if the hypothesis were true.

If your observation lies in the region of extreme observations (with respect to α) you reject the hypothesis.

Example chess game:

Chess game statistical test

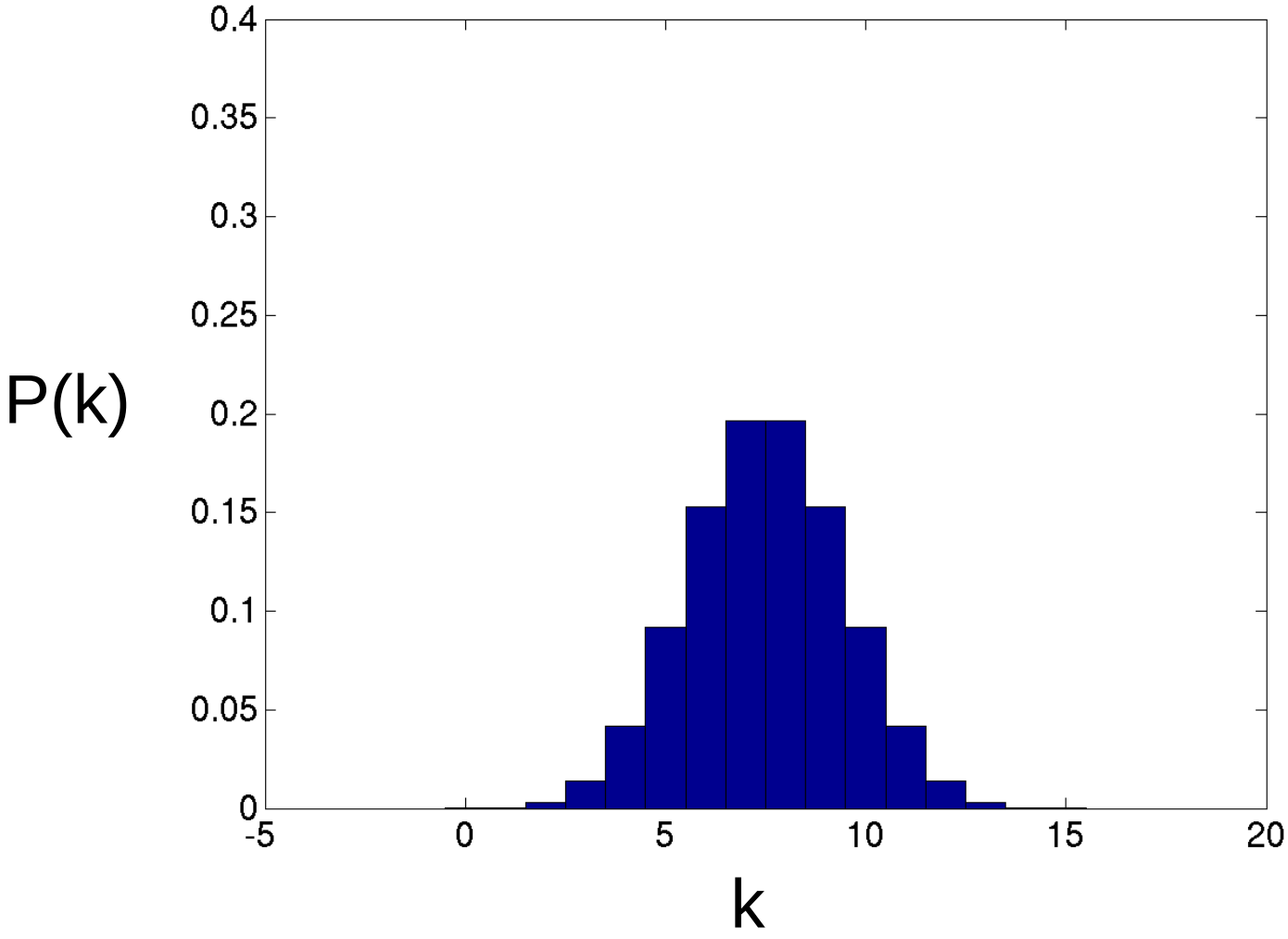
Probability space: $\Omega = \{0, 1\}$

$$P(\{0\}) = P(\{1\}) = \frac{1}{2}$$

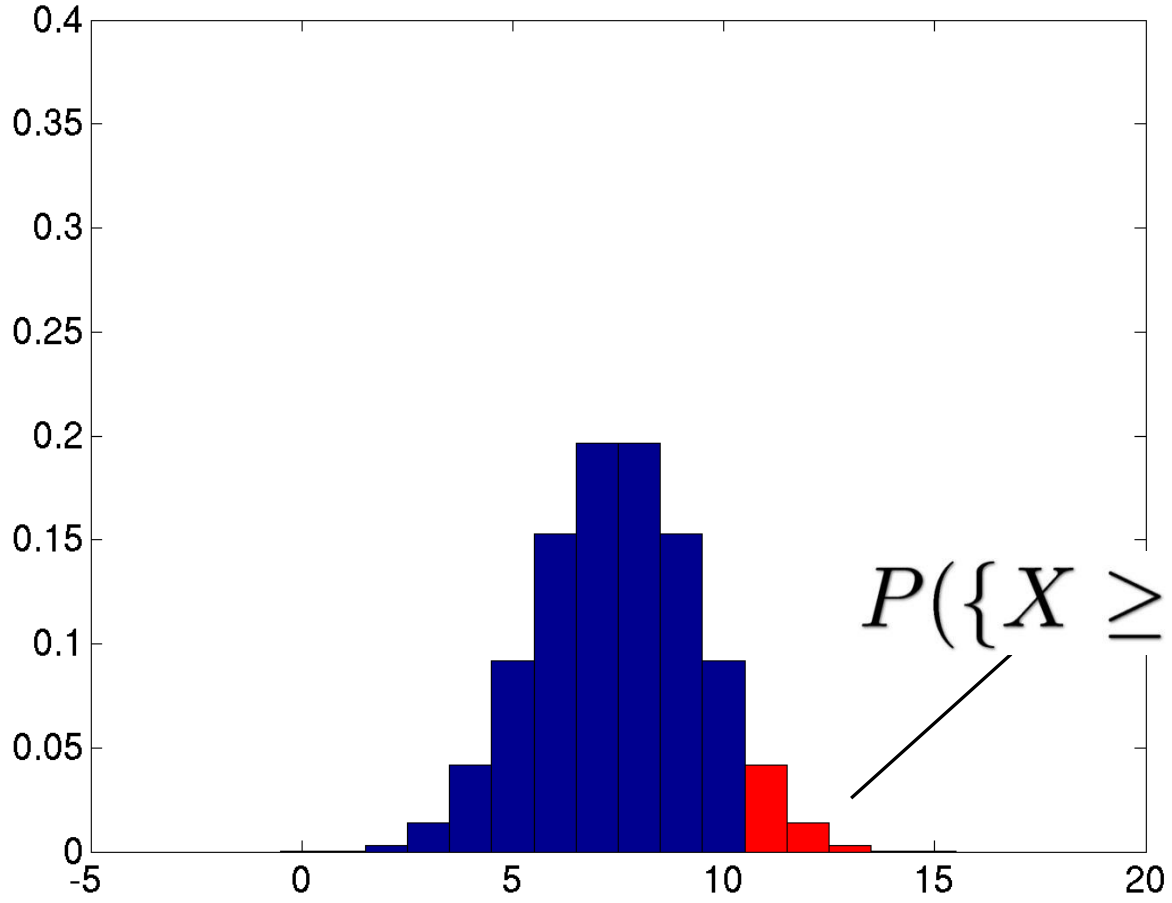
(This is the hypothesis H)

We can calculate the probability for n wins from this:

Probability for k wins



Tail probability



Chess hypothesis conclusion

If you had won 11 or more games out of 15 we would reject the hypothesis that the winning probability is 50%.
If you win 10 or less we cannot make a strong statement.

One-tailed versus two-tailed tests

- We only considered the alternative of being better in chess (right tail)
- In some problems it makes sense to consider deviations to both sides – the tests are called two-tailed tests.
- For two-tailed tests, one looks for a low limit and a high limit:

$$P(\{X \leq k_{\text{low}}\}) \leq \frac{\alpha}{2} \qquad P(\{X \geq k_{\text{high}}\}) \leq \frac{\alpha}{2}$$

Summary: Statistical tests so far

- Typically we test a **null-Hypothesis H_0**
- There is a **test statistic X** and a corresponding **probability distribution P(X=k)**
- We calculate the probability of an observed value **x** of the test statistic under the assumption that the null-Hypothesis is true
- Based on the this probability we reject the null-Hypothesis or "**do not reject**" it.
- There is no such thing as "accepting the null-Hypothesis"

Possible errors

- There are two possible errors we can make:

1. Errors of the first kind: The null-Hypothesis was true but we reject it. The probability for this type of error is limited by the significance niveau α

Possible errors

2. Errors of the second kind: The null-Hypothesis is false but we cannot reject it. The probability of this type of error is unknown and may be large – this is one of the reasons why not rejecting the null-Hypothesis should not be interpreted as accepting it.

Reporting significance (P-value)

- In the past the distributions of test statistics were taken from tables. To avoid imprecisions from this, scientists only reported $P < \alpha$ for a significance niveau α that was chosen up front.
- Nowadays, all distributions of test statistics can be calculated numerically to any precision – it is ok now to report observed P-values directly, e.g. P=0.015.