Statistical Data Analysis

Dr. Jana de Wiljes

3. November 2021

Universität Potsdam

Random variables

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[\bar{X}_n] = \mu \text{ and } Var[\bar{X}_n] = \frac{\sigma^2}{n} \tag{1}$$

Proof

Law of large numbers

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$. Then

$$\bar{X}_n \to \mu \text{ for } n \to \infty \text{ (almost sure)} \tag{2}$$

Empirical variance

Definition: The empirical variance is defined by

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \tag{3}$$

Note: we will also use an analog notation for the random variables:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \tag{4}$$

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables. Then

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i^2 - n\bar{X}_n^2) \tag{5}$$

Proof

Proof

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[S_n^2] = \sigma^2 \tag{6}$$

Proof

Proof

Empirical standard deviation

Def: The empirical standard deviation is defined by

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \tag{7}$$

Order statistic

Def: Let $(x_1, \ldots, x_n) \in \mathbb{R}^n$ be a sample set. One can order the elements in an increasing manner:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \tag{8}$$

Then $x_{(i)}$ is referred to as the i-th order statistic of the sample set.

Sample median

Def: The sample median of a set of samples if given by

$$\mathrm{Med}_n = \mathrm{Med}_n(x_1, \ldots, x_n) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \textit{in case n is uneven} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{in case } n \text{ is even} \end{cases}$$

Then $x_{(i)}$ is referred to as the i-th order statistic of the sample set.

Example

Truncated mean

Def: The truncated mean samples $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined by

$$\frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

Empirische α-Quantile

Def: Let $(x_1, \ldots, x_n) \in \mathbb{R}^n$ be a set of samples and $\alpha \in (0, 1)$. The empirical α Quantil is defined by

$$q_\alpha = \begin{cases} x_{\lfloor n\alpha \rfloor + 1} & \text{falls } n\alpha \notin \mathbb{N} \\ \frac{1}{2}\left(x_{\lfloor n\alpha \rfloor} + x_{\lfloor n\alpha \rfloor + 1}\right) & \text{falls } n\alpha \in \mathbb{N} \end{cases}$$

Distibution of the order statistic

Proposition: Let $X_1, X_2, \ldots, X_n$ be independent and identical distributed random variables, that are absolute continuous with a density f and cumulative distibution function F. Let

$$X_{(1)} \leq X_{(2)} \leq, \ldots, \leq X_{(n)} \tag{9}$$

be the order statistics. Then the density of the random variable $X_{(i)}$ is

$$f_{X_{(i)}}(t) = \frac{n!}{(i-1)!(n-1)!} f(t) F(t)^{(i-1)} (1 - F(t))^{n-i} \tag{10}$$

Beta distribution

Def: For a and b larger than zero and

$$f(x) = \frac{1}{\mathrm{B}(a, b)} x^{a-1}(1-x)^{b-1}.$$

where the normalization is given by

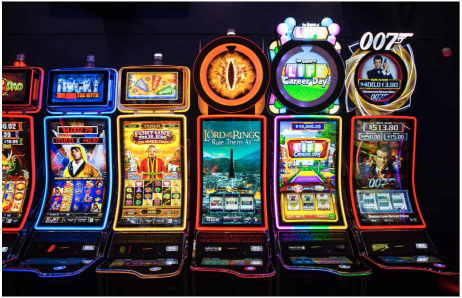
$$\mathrm{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$$

with $\Gamma(n) = (n-1)!$ being the gamma function.

Excurse to Bandits

Multi-armed bandits

Choose from K options
to receive a
high reward and
to educe loss after T rounds



Examples:

- Which advertising campaign generates the largest revenue
- Which restaurant to pick ?
- Which netflix series to streamen?
- Which vaccination should be further developed ?

Multi-armed bandits

A stochastic K-Armed Bandit is defined via the tuple $\langle \mathcal{A}, \mathcal{Y}, P, r \rangle$ where

- $\mathcal{A}$ is the set of actions (arms) and $|\mathcal{A}| = K$
- $\mathcal{Y}$ is the set of possible outcomes
- $P(\cdot|a) \in \mathcal{P}(\mathcal{Y})$ is the outcome probability, conditioned on action $a \in \mathcal{A}$ being taken,
- $r(\mathcal{Y}) \in \mathcal{R}$ represents the reward obtained when outcome $Y \in \mathcal{Y}$ is observed

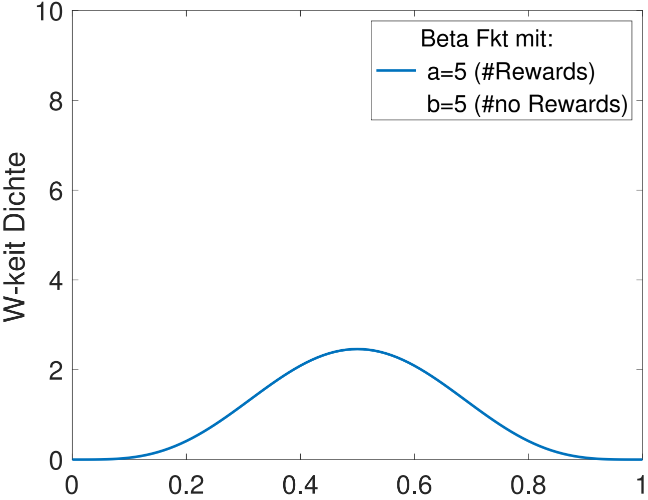
Regret

Def: Let $a^* \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_{y \sim P(\cdot|a)}[r(y)]$ denote the optimal arm. The T-period regret of the sequence of actions $a_1, \ldots, a_T$ is the random variable

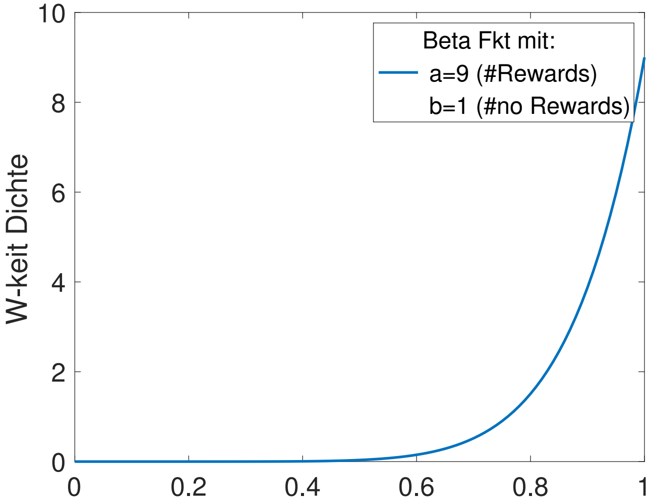
$$\mathbf{Regret}(T) = \sum_{t=1}^{T} \Big[r(Y_t(a^*)) - r(Y_t(a_t)) \Big] \tag{11}$$

Thompson Sampling

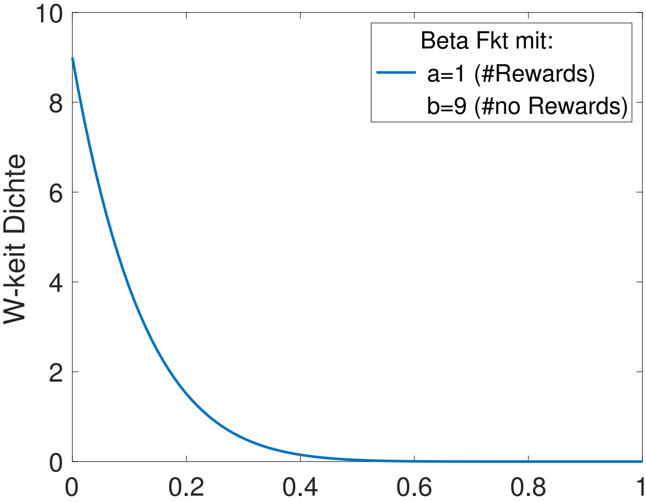
Beta distribution



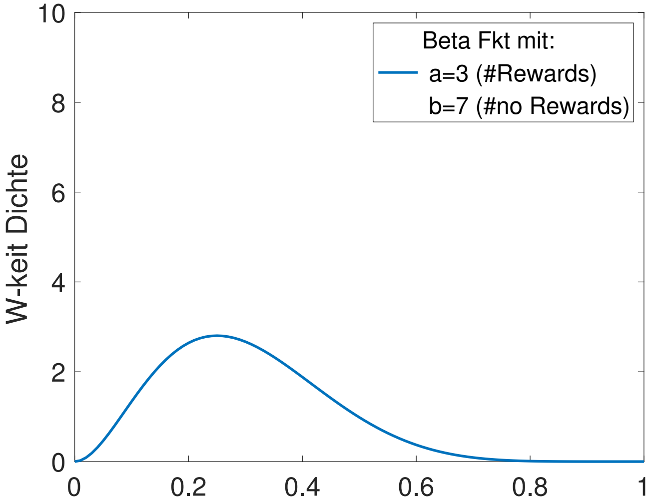
Beta distribution



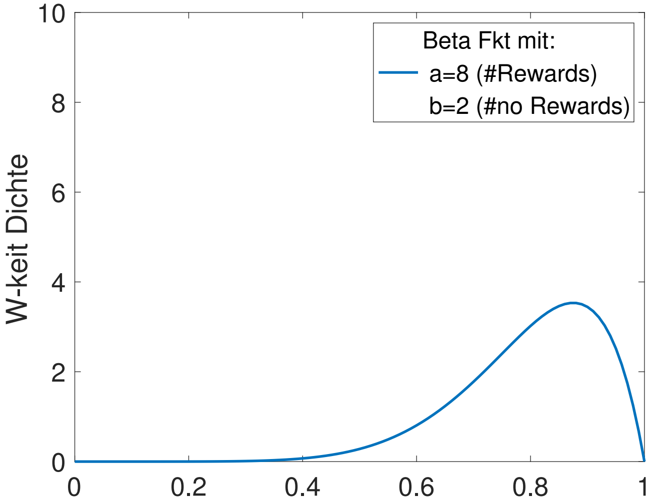
Beta distribution



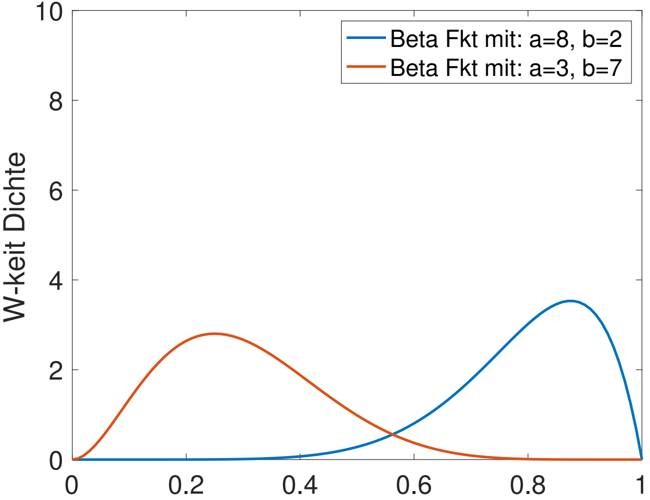
Beta distribution



Beta distribution



Beta distribution



Thompson Sampling

- **Problem setting:** Choose from K options to receive a high reward
- **Algorithm:** Iterated over the following steps:
  1. In each round save information on the choice of action and if a reward was received
  2. Draw from the beta distribution: defined via for each action by
     - a) how often performing action resulted in a reward
     - b) how often performing action did not resulted in a reward
  3. Choose the action that has the highest beta function value

Empirical cdf

The empirical cdf of a sample set $(x_1, \dots, x_n) \in \mathbb{R}^n$ is defined through

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_i \leq 1} = \frac{1}{n} \#\{i \in \{1, \dots, n\} : x_i \leq t\}, \quad t \in \mathbb{R} \tag{12}$$

Empirical cdf