The Binomial Distribution

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Plan

Derive the Bernoulli and binomial distributions

Analyze the empirical-probability estimator

Bernoulli distribution

Coin flip such that probability of heads is θ

Bernoulli random variable with parameter θ

$$p_\theta(1) = \theta$$

$$p_\theta(0) = 1 - \theta$$

Coin flips

We flip a coin with bias θ independently n times, probability of a heads?

Strategy: Decompose into union of events

$$\begin{aligned}\{a \text{ heads}\} = &\{H_1, H_2, \ldots, H_a, T_{a+1}, T_{a+2}, \ldots, T_n\}\\ &\cup \{H_1, T_2, H_3, \ldots, H_{a+1}, T_{a+2}, \ldots, T_n\}\\ &\cup \ldots\end{aligned}$$

Coin flips

$$
\begin{aligned}
&\mathrm{P}\left(\{H_1, H_2, \ldots, H_a, T_{a+1}, T_{a+2}, \ldots, T_n\}\right) \\
&= \mathrm{P}\left(H_1\right) \cdots \mathrm{P}\left(H_a\right) \mathrm{P}\left(T_{a+1}\right) \cdots \mathrm{P}\left(T_n\right) \\
&= \theta^a \left(1 - \theta\right)^{n-a}
\end{aligned}
$$

What about $\{H_1, T_2, H_3, \ldots, H_{a+1}, T_{a+2}, \ldots, T_n\}$?

Example: Coin flips

How many possible orders are there?

$$\binom{n}{a} := \frac{n!}{a!\,(n-a)!}$$

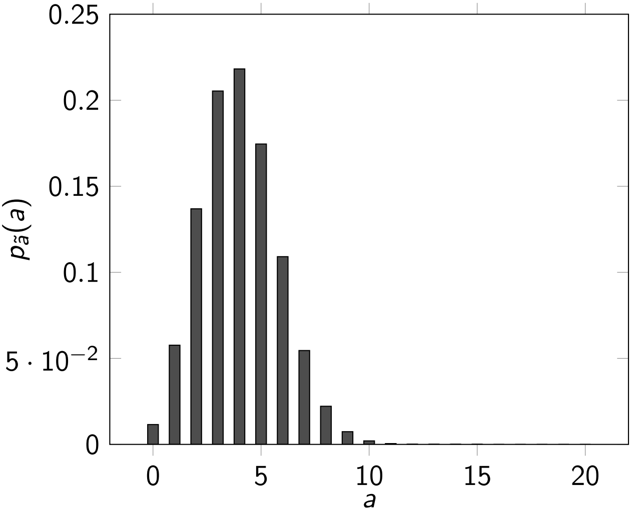
$$\mathrm{P}(\{a \text{ heads}\}) = \binom{n}{a}\,\theta^{a}\,(1-\theta)^{n-a}$$

Binomial distribution

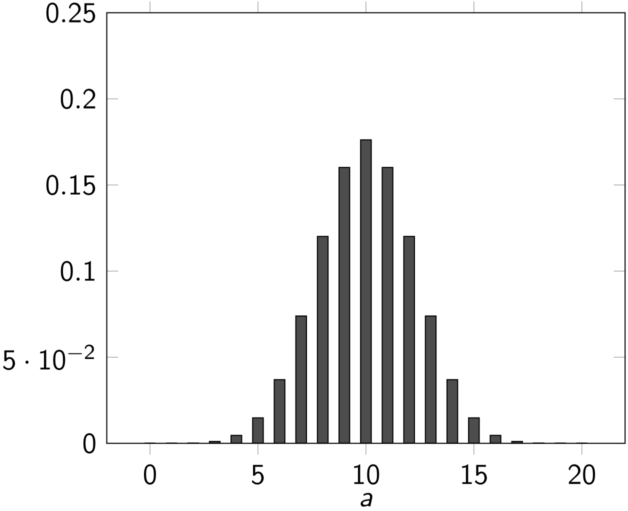
The pmf of a binomial random variable $\tilde{a}$ with parameters n and θ is

$$p_{\tilde{a}}(a) = \binom{n}{a} \theta^{a} (1-\theta)^{(n-a)} \qquad a = 0, 1, \ldots, n$$

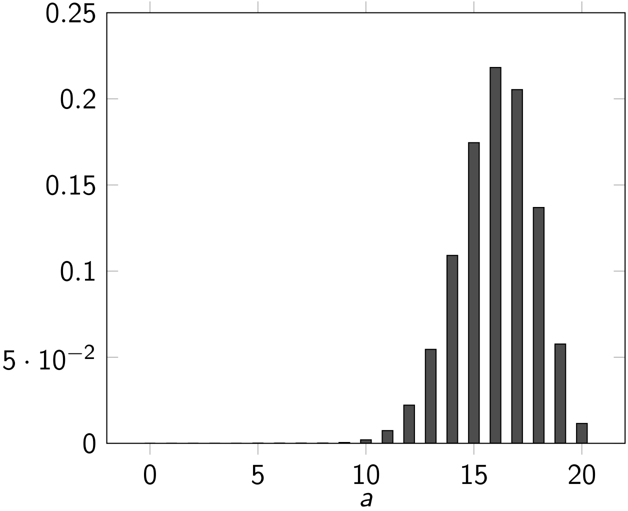
Binomial $n = 20$, $\theta = 0.2$



Binomial $n = 20$, $\theta = 0.5$



Binomial $n = 20$, $\theta = 0.8$



Empirical probability

Statistical estimator for the probability of an event

Let A be an event in a sample space Ω

Let $X := \{x_1, x_2, \ldots, x_n\}$ be a set of data with values in Ω

The empirical probability of A is

$$\mathrm{P}_X(A) := \frac{\sum_{i=1}^{n} 1_{x_i \in A}}{n}$$

where $1_{x_i \in A}$ is one if $x_i \in A$ and zero otherwise

We can use the binomial distribution to analyze this estimator

Coin flip

We simulate a fair coin flip twenty times

Heads (out of 20)	15	13	10	9	9	8	9	9	12	8
Empirical prob.	0.75	0.65	0.5	0.45	0.45	0.4	0.45	0.45	0.6	0.4

Analysis of empirical probability

Assume $\mathrm{P}(A) = \theta_{\text{true}}$

$B_i :=$ *data point i belongs to A*

Data are independent

$S_1, S_2, \ldots, S_n$ (where S_i is B_i or B_i^c) are all mutually independent

We model number of data in A by random variable $\tilde{c}$

Distribution of $\tilde{c}$?

Distribution of empirical probability

Binomial with parameters n and θ_{true}!

$$p_{\tilde{c}}(c) = \binom{n}{c} \theta_{\text{true}}^{c} \left(1 - \theta_{\text{true}}\right)^{(n-c)} \quad c = 0, 1, 2, \ldots, n$$

The empirical probability estimator is $\tilde{\theta} := \frac{\tilde{c}}{n}$ so

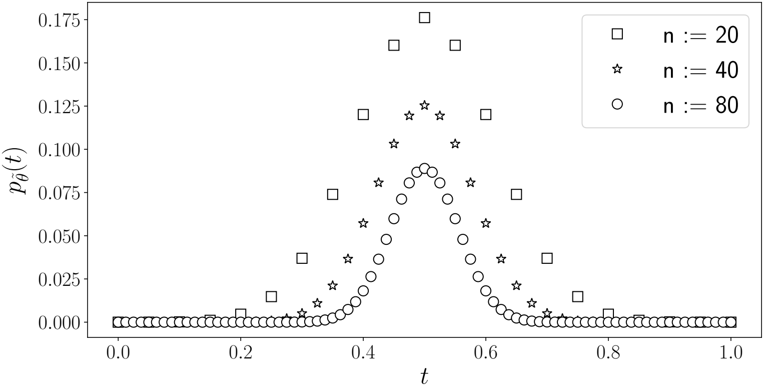
$$\begin{aligned} p_{\tilde{\theta}}(t) &= \mathrm{P}(\tilde{c} = nt) \\ &= \binom{n}{nt} \theta_{\text{true}}^{nt} \left(1 - \theta_{\text{true}}\right)^{n-nt} \quad t = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1 \end{aligned}$$

Coin flip

We simulate a fair coin flip twenty times

Heads (out of 20)	15	13	10	9	9	8	9	9	12	8
Empirical prob.	0.75	0.65	0.5	0.45	0.45	0.4	0.45	0.45	0.6	0.4

Distribution of empirical probability



$$\mathrm{P}(|\tilde{\theta} - \theta_{\text{true}}| \leq 0.1)?$$

0.737 ($n = 20$) 0.846 ($n = 40$) 0.943 ($n = 80$)

What have we learned?

Definition of the Bernoulli and binomial distributions

How to analyze the empirical-probability estimator