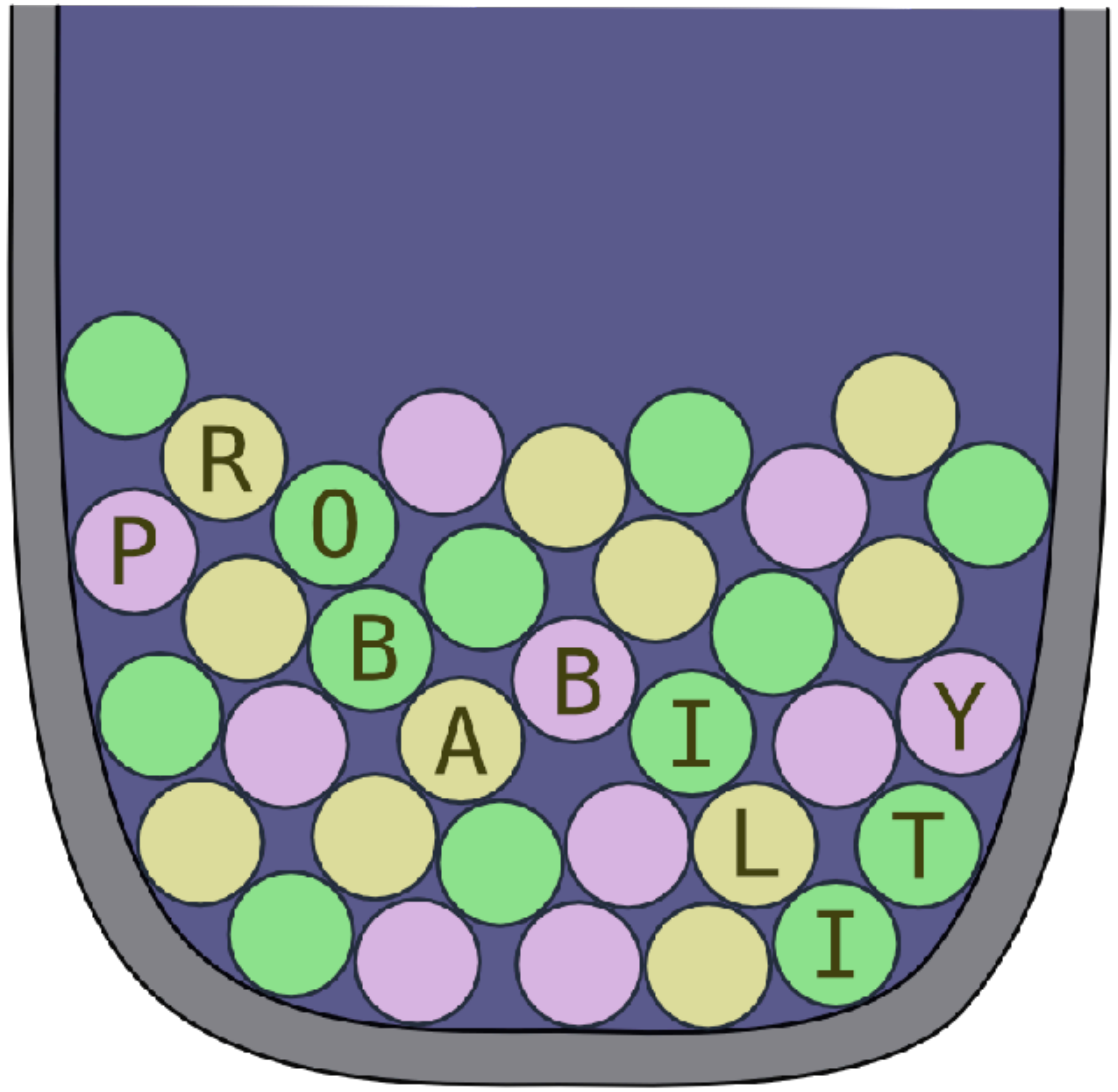
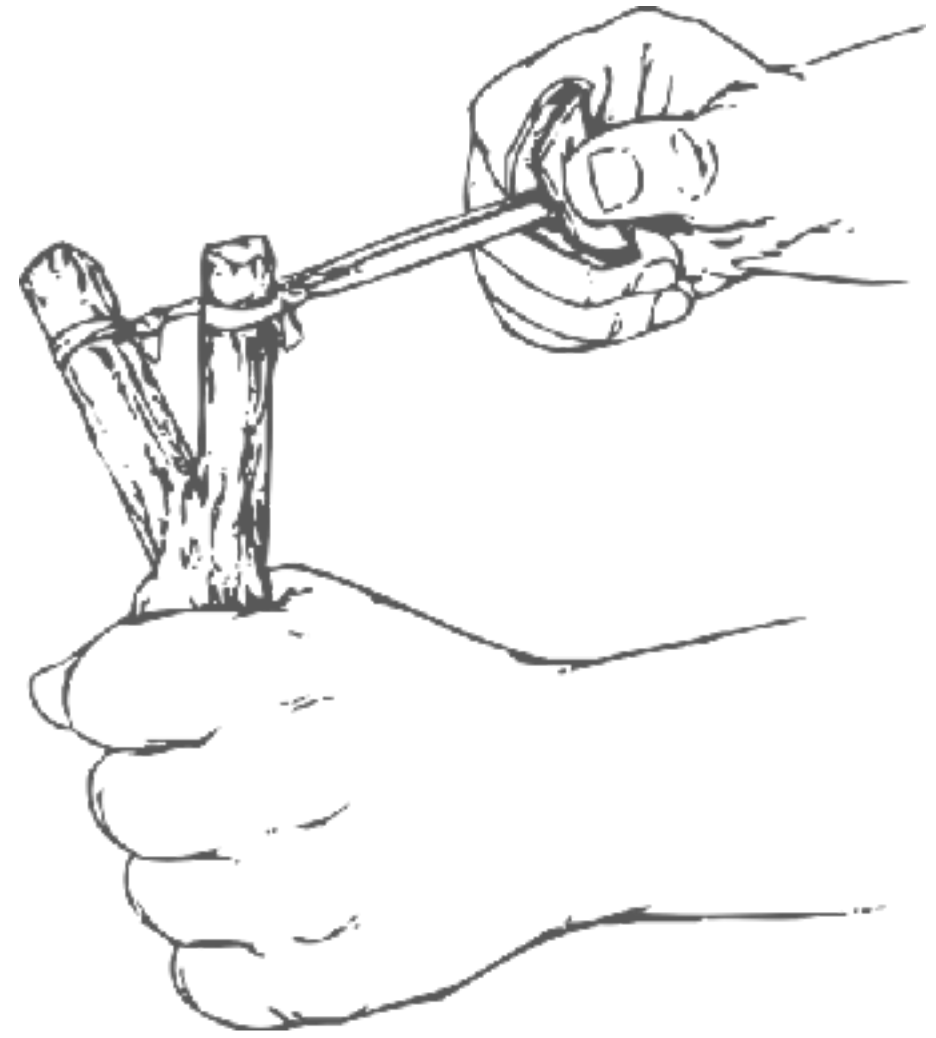


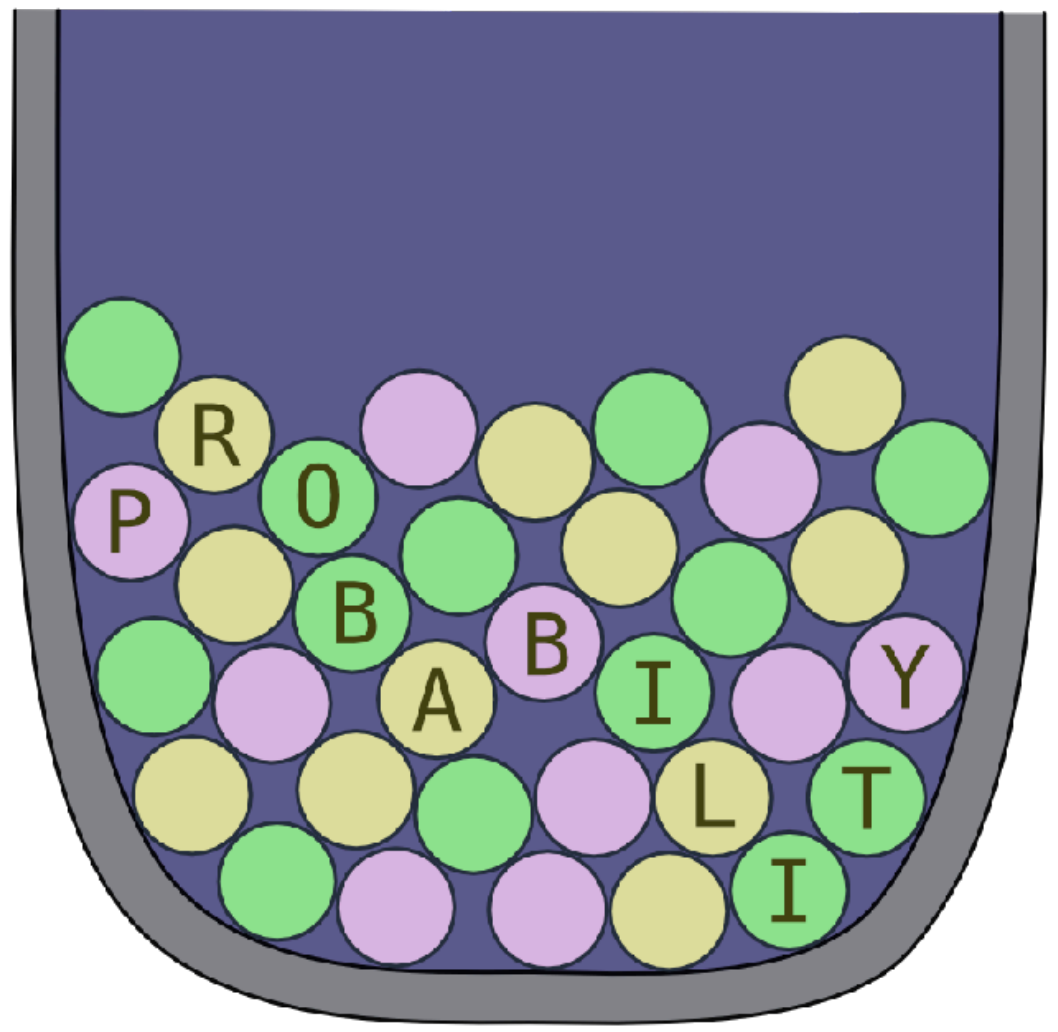
INTRODUCTION TO DATA ANALYSIS

PROBABILITY BASICS

LEARNING GOALS



- become familiar with the notion of probability
  - axiomatic definition & interpretation
  - joint, marginal & conditional probability
- Bayes rule
- random variables
- probability distributions in R
- probability distributions as approximated by samples



Probability

ELEMENTARY OUTCOMES AND EVENTS

- a random process has **elementary outcomes** $\Omega = \{\omega_1, \omega_2, \dots\}$
  - elementary outcomes are mutually exclusive
  - Ω exhausts the space of possibilities

> **Example.** The set of elementary outcomes of a single coin flip is $\Omega_{\text{coin flip}} = \{\text{heads}, \text{tails}\}$. The elementary outcomes of tossing a six-sided die is $\Omega_{\text{standard die}} = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$.[27]

- any $A \subseteq \Omega$ is an event
  - standard set-theoretic notation for negation, conjunction, disjunction etc.
  - example "rolling an odd number" $A = \{$⚀, ⚂, ⚄$\}$

PROBABILITY DISTRIBUTION

A **probability distribution** P over Ω is a function $P : \mathfrak{P}(\Omega) \rightarrow \mathbb{R}$ that assigns to all events $A \subseteq \Omega$ a real number (from the unit interval, see A1 below), such that the following (so-called Kolmogorov axioms) are satisfied:

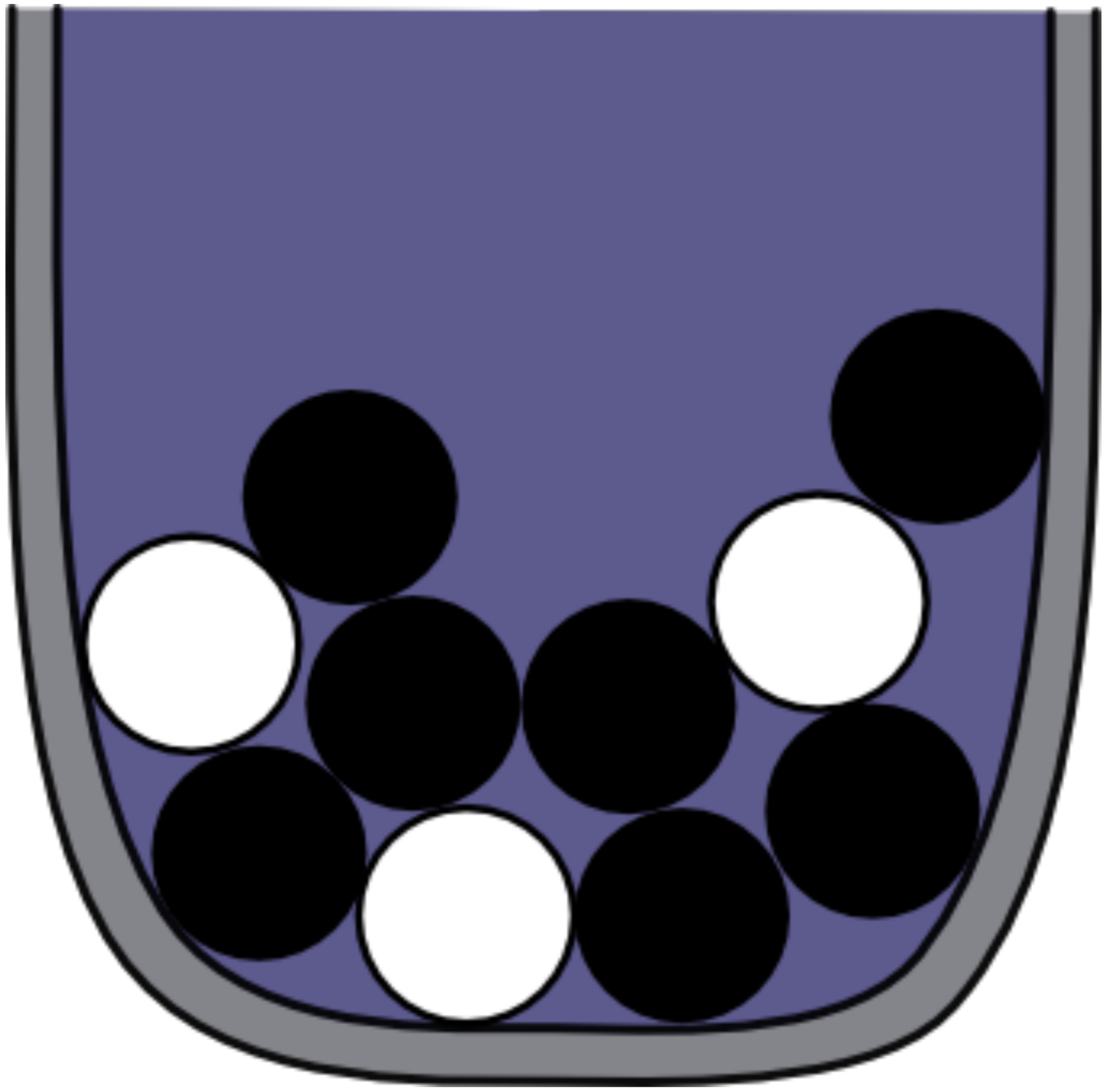
A1. $0 \leq P(A) \leq 1$

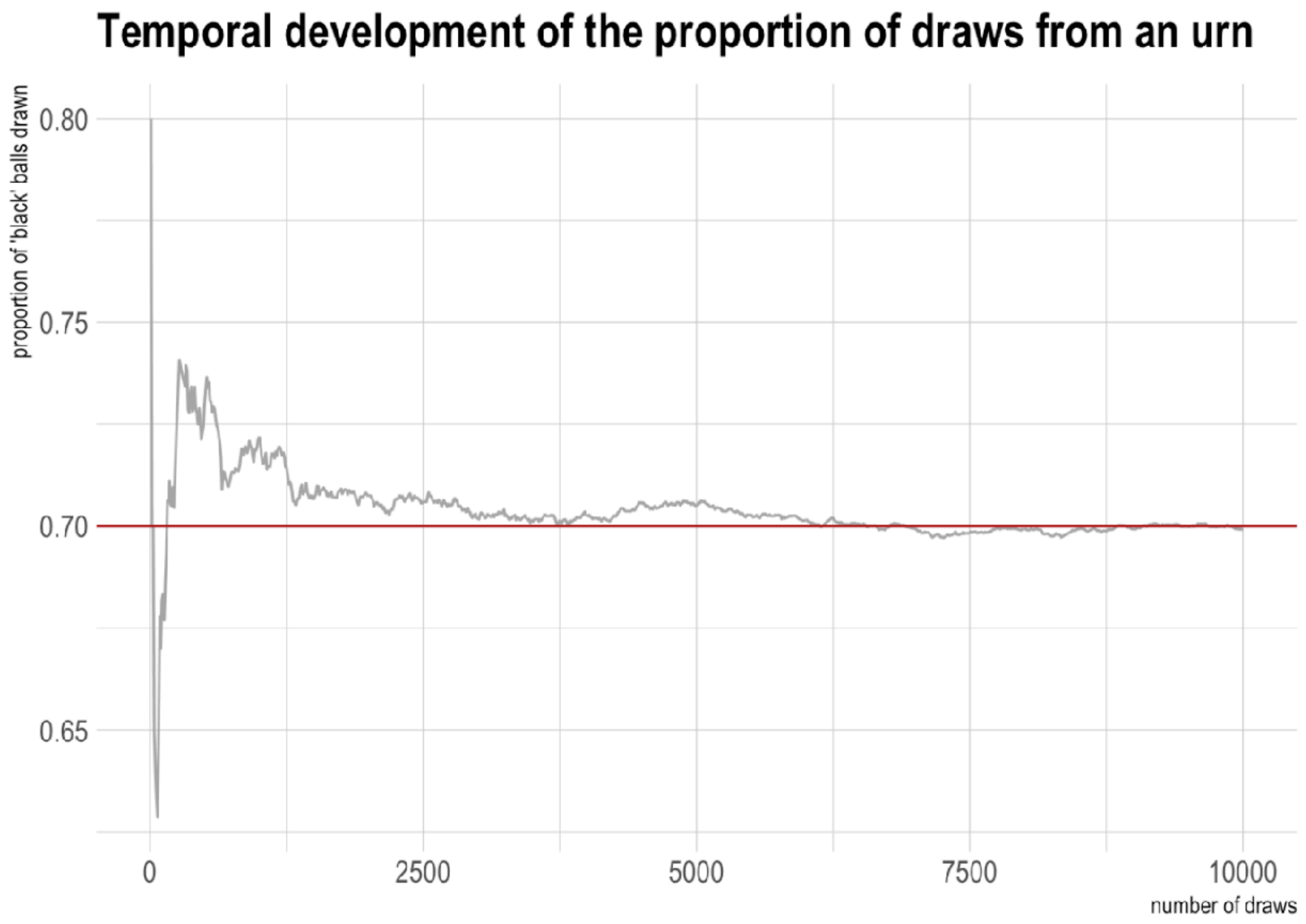
A2. $P(\Omega) = 1$

A3. $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$ whenever $A_1, A_2, A_3, \ldots$ are mutually exclusive

INTERPRETATIONS OF PROBABILITY

- Frequentist: probabilities are generalizations of intuitions/facts about frequencies of events in repeated executions of a random event.
- Subjectivist: probabilities are subjective beliefs by a rational agent who is uncertain about the outcome of a random event.
- Realist: probabilities are a property of an intrinsically random world.

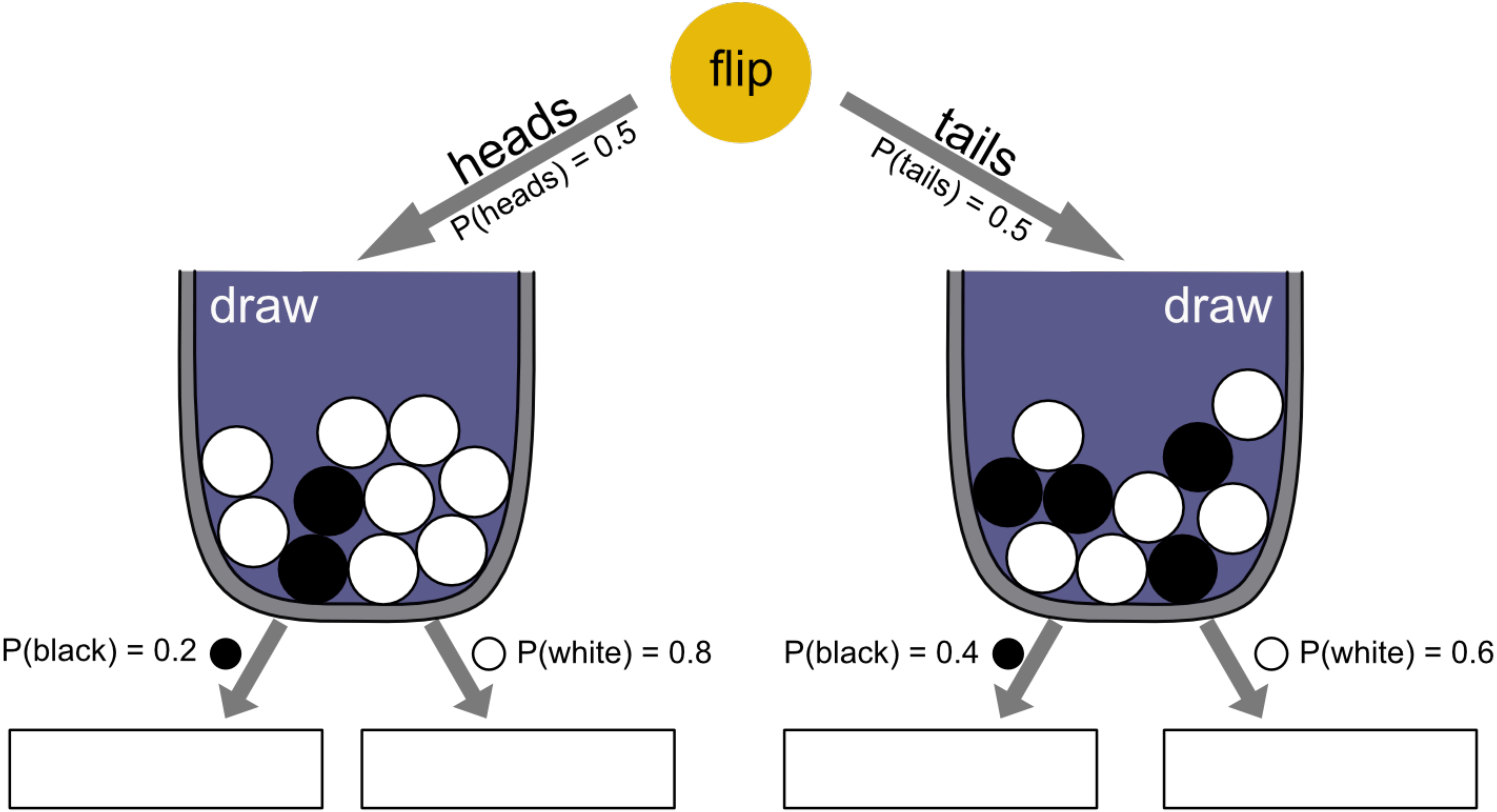



Temporal development of the proportion of draws from an urn
proportion of 'black' balls drawn
0.80
0.75
0.70
0.65
0
2500
5000
7500
10000
number of draws

PROBABILITY DISTRIBUTIONS AS SAMPLES

- No matter our preferred metaphysical interpretation, we can approximate a probability distribution by either:
  - a large set of representative samples; or
  - an oracle that returns a sample if needed.

Structured events


flip
heads
P(heads) = 0.5
tails
P(tails) = 0.5
draw
draw
P(black) = 0.2
P(white) = 0.8
P(black) = 0.4
P(white) = 0.6

JOINT PROBABILITY DISTRIBUTIONS

- Structured elementary outcomes: $\Omega_{flip-\&-draw} = \Omega_{flip} \times \Omega_{draw}$
  - shorthand notation $P(\text{heads, black})$ instead of $P(\langle \text{heads, black} \rangle)$

	heads	**tails**
black	$0.5 \times 0.2 = 0.1$	$0.5 \times 0.4 = 0.2$
white	$0.5 \times 0.8 = 0.4$	$0.5 \times 0.6 = 0.3$

MARGINAL DISTRIBUTIONS

- if $\Omega = \Omega_1 \times \dots \Omega_n$ and $A_i \subseteq \Omega_i$, the **marginal probability** of A_i is:

$$P(A_i) = \sum_{A_1 \subseteq \Omega_1, \dots, A_{i-1} \subseteq \Omega_{i-1}, A_{i+1} \subseteq \Omega_{i+1}, \dots, A_n \subseteq \Omega_n} P(A_1, \dots, A_{i-1}, A_i, A_{i+1}, \dots A_n)$$

	heads	**tails**	$\sum$
black	$0.5 \times 0.2 = 0.1$	$0.5 \times 0.4 = 0.2$	$P(\text{black}) = 0.3$
white	$0.5 \times 0.8 = 0.4$	$0.5 \times 0.6 = 0.3$	$P(\text{white}) = 0.7$
$\sum$	$P(\text{heads}) = 0.5$	$P(\text{tails}) = 0.5$	

Conditional probability & Bayes rule

CONDITIONAL PROBABILITY

- the conditional probability of A given B is: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

$$P(\text{black} \mid \text{heads}) = \frac{P(\text{black, heads})}{P(\text{heads})} = \frac{0.1}{0.5} = 0.2$$

	heads	tails	$\sum$
black	$0.5 \times 0.2 = 0.1$	$0.5 \times 0.4 = 0.2$	$P(\text{black}) = 0.3$
white	$0.5 \times 0.8 = 0.4$	$0.5 \times 0.6 = 0.3$	$P(\text{white}) = 0.7$
$\sum$	$P(\text{heads}) = 0.5$	$P(\text{tails}) = 0.5$	

BAYES RULE

- Bayes rule follows straightforwardly from the definition of conditional probability:

$$P(B \mid A) = \frac{P(A \mid B)\ P(B)}{P(A)}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(B \mid A)\ P(A)$$

$$P(B \cap A) = P(A \mid B) \cdot P(B)$$

PREVIEW ::: BAYES RULE FOR DATA ANALYSIS

$$P(B \mid A) = \frac{P(A \mid B)\ P(B)}{P(A)}$$

likelihood of data prior over parameters

$$P(\theta \mid D) = \frac{P(D \mid \theta)\ P(\theta)}{P(D)}$$

posterior over parameters

marginal likelihood of data

Random variables

RANDOM VARIABLES

- a random variable is a function: $X : \quad \Omega \rightarrow \mathbb{R}$
  - if range of X is countable, we speak of a discrete random variable
  - otherwise, we speak of a continuous random variable
- think: distribution of a summary statistic
- notation:
  - shorthand notation $P(X = x)$ instead of $P(\{\omega \in \Omega \mid X(\omega) = 2\})$
  - similarly write stuff like $P(X \leq x)$ or $P(1 \leq X \leq 2)$

RANDOM VARIABLE ::: EXAMPLES

Example. For a single flip of a coin we have $\Omega_{\text{coin flip}} = \{\text{heads}, \text{tails}\}$. A usual way of mapping this onto numerical outcomes is to define $X_{\text{coin flip}} : \text{heads} \mapsto 1; \text{tails} \mapsto 0$. Less trivially, consider flipping a coin two times. Elementary outcomes should be individuated by the outcome of the first flip and the outcome of the second flip, so that we get:

$$\Omega_{\text{two flips}} = \{\langle\text{heads}, \text{heads}\rangle, \langle\text{heads}, \text{tails}\rangle, \langle\text{tails}, \text{heads}\rangle, \langle\text{tails}, \text{tails}\rangle\}$$

Consider the random variable $X_{\text{two flips}}$ that counts the total number of heads. Crucially, $X_{\text{two flips}}(\langle\text{heads}, \text{tails}\rangle) = 1 = X_{\text{two flips}}(\langle\text{tails}, \text{heads}\rangle)$. We assign the same numerical value to different elementary outcomes.

CUMULATIVE DISTRIBUTION & PROBABILITY MASS ::: DISCRETE RVs

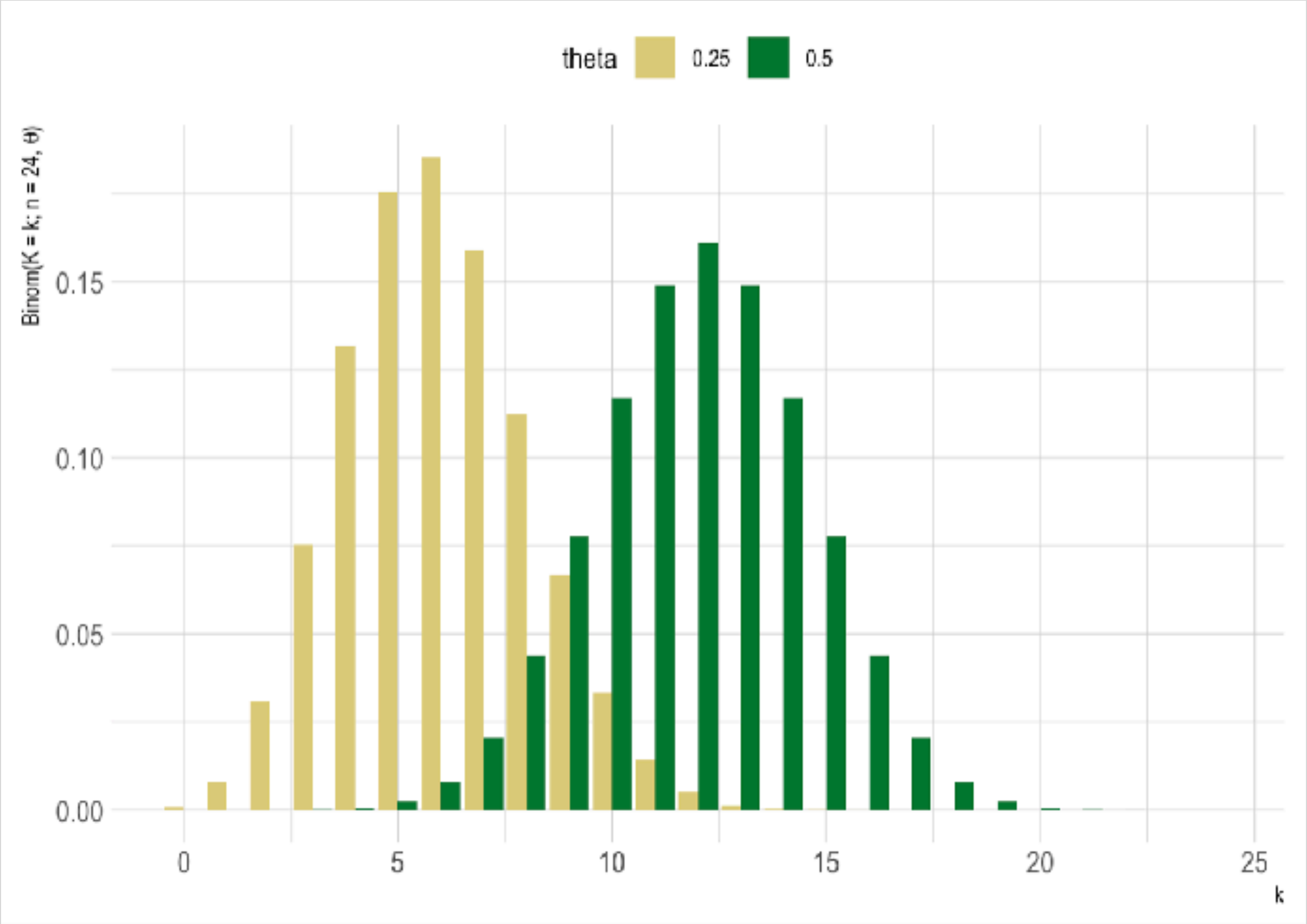
For a discrete random variable X, the **cumulative distribution function** F_X associated with X is defined as:

$$F_X(x) = P(X \leq x) = \sum_{x' \in \{\text{Rng}(X) \mid x' \leq x\}} P(X = x)$$

The **probability mass function** f_x associated with X is defined as:

$$f_X(x) = P(X = x)$$

$$\text{Binom}(K = k; n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$



probability mass function

CUMULATIVE DISTRIBUTION & PROBABILITY MASS ::: DISCRETE RVs

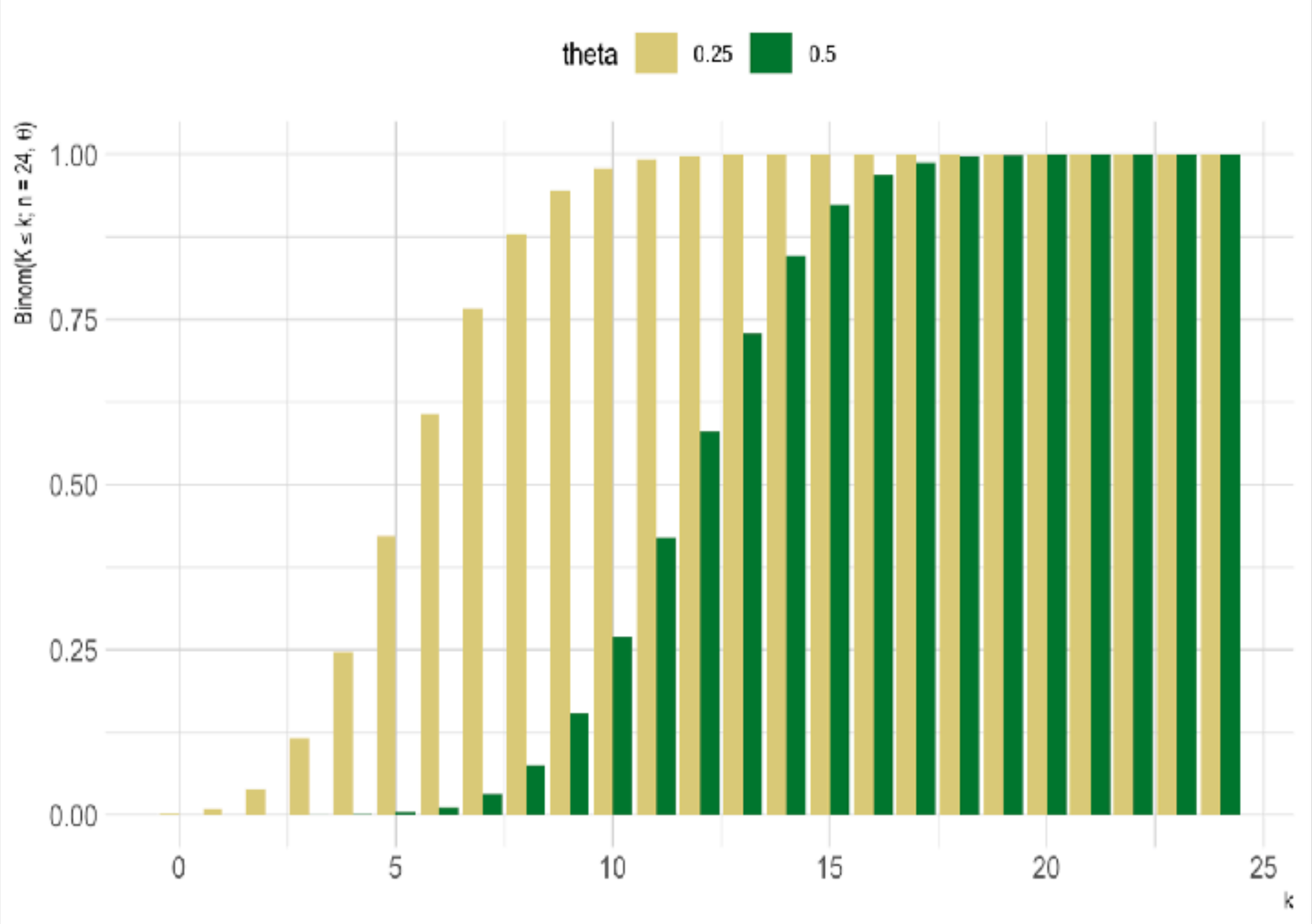
For a discrete random variable X, the **cumulative distribution function** F_X associated with X is defined as:

$$F_X(x) = P(X \leq x) = \sum_{x' \in \{\text{Rng}(X) \mid x' \leq x\}} P(X = x)$$

The **probability mass function** f_x associated with X is defined as:

$$f_X(x) = P(X = x)$$

$$\text{Binom}(K = k; n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$



cumulative probability function

CUMULATIVE DISTRIBUTION & PROBABILITY MASS ::: CONTINUOUS RVs

For a continuous random variable X, the probability $P(X = x)$ will usually be zero: it is virtually impossible that we will see precisely the value x realized in a random event that can realize uncountably many numerical values of X. However, $P(X \leq x)$ does take workable values and so we define the cumulative distribution function F_X associated with X as:

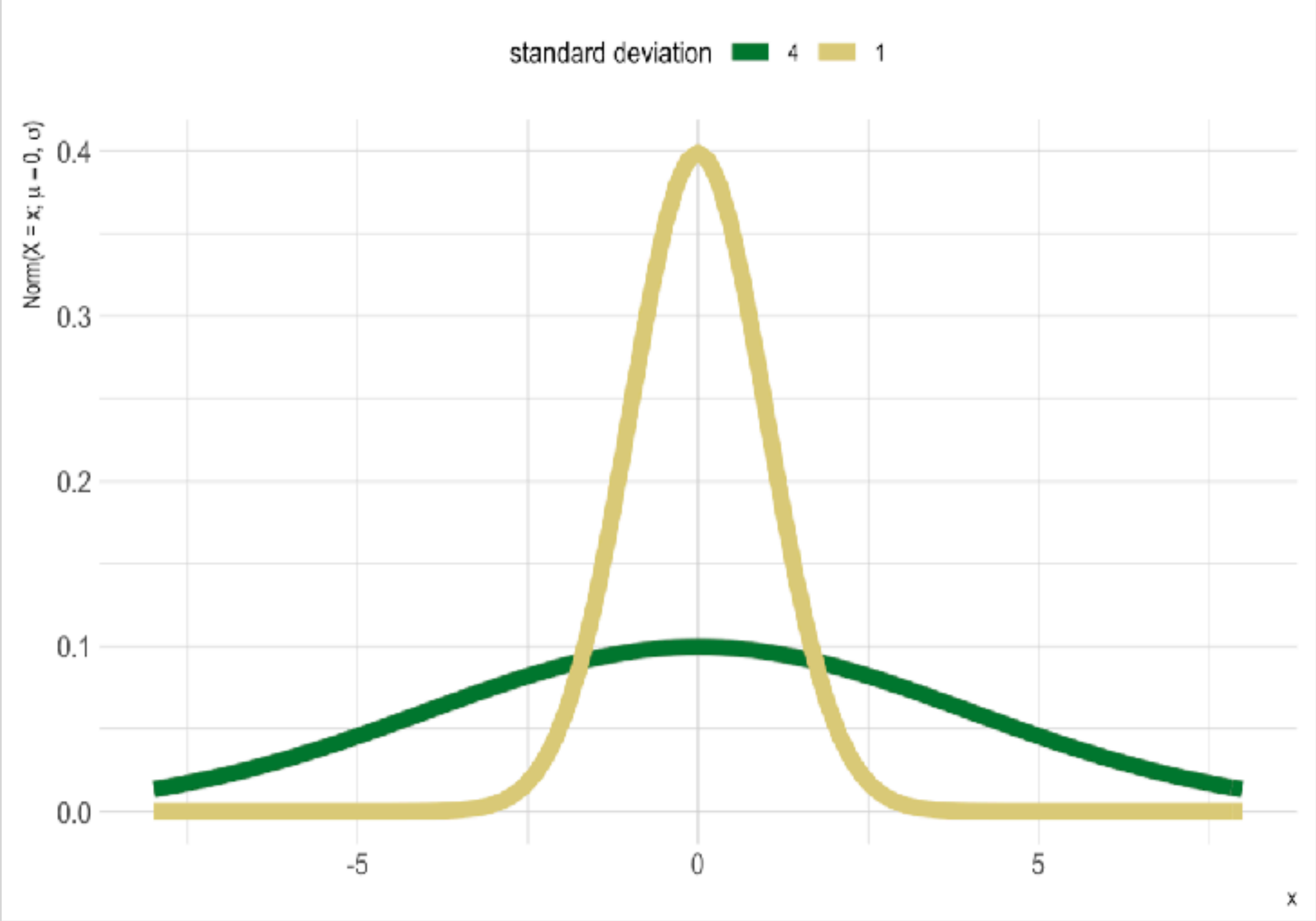
$$F_X(x) = P(X \leq x)$$

Instead of a probability **mass** function, we derive a **probability density function** from the cumulative function as:

$$f_X(x) = F'(x)$$

A probability density function can take values greater than one, unlike a probability mass function.

$$\mathcal{N}(X = x; \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



probability density function

CUMULATIVE DISTRIBUTION & PROBABILITY MASS ::: CONTINUOUS RVs

For a continuous random variable X, the probability $P(X = x)$ will usually be zero: it is virtually impossible that we will see precisely the value x realized in a random event that can realize uncountably many numerical values of X. However, $P(X \leq x)$ does take workable values and so we define the cumulative distribution function F_X associated with X as:

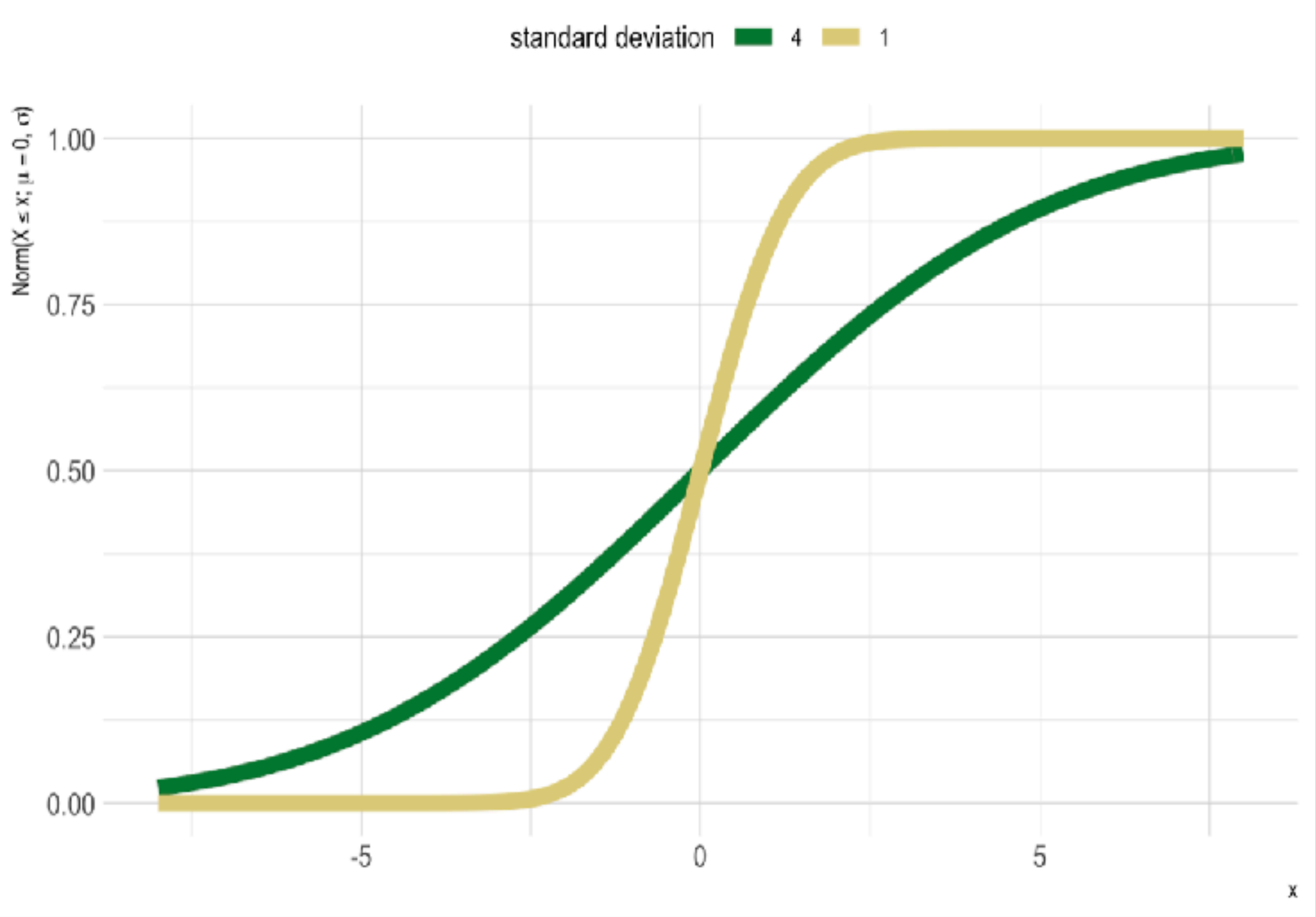
$$F_X(x) = P(X \leq x)$$

Instead of a probability **mass** function, we derive a **probability density function** from the cumulative function as:

$$f_X(x) = F'(x)$$

A probability density function can take values greater than one, unlike a probability mass function.

$$\mathcal{N}(X = x; \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



cumulative probability function

EXPECTED VALUE OF A RANDOM VARIABLE

- the expected value of random variable $X : \Omega \rightarrow \mathbb{R}$ is:

  if X is discreet: $\mathbb{E}_X = \sum_x x \, f_X(x)$

  if X is continuous: $\mathbb{E}_X = \int x \, f_X(x) \; \mathrm{d}x$

- think: mean of a representative sample of X

VARIANCE OF A RANDOM VARIABLE

- the variance of random variable $X : \Omega \rightarrow \mathbb{R}$ is:

  if X is discreet: $\text{Var}(X) = \sum_x (\mathbb{E}_X - x)^2 \ f_X(x)$

  if X is continuous: $\text{Var}(X) = \int (\mathbb{E}_X - x)^2 \ f_X(x) \ \mathrm{d}x$

- think: variance of a representative sample of X

COMPOSITE RANDOM VARIABLES

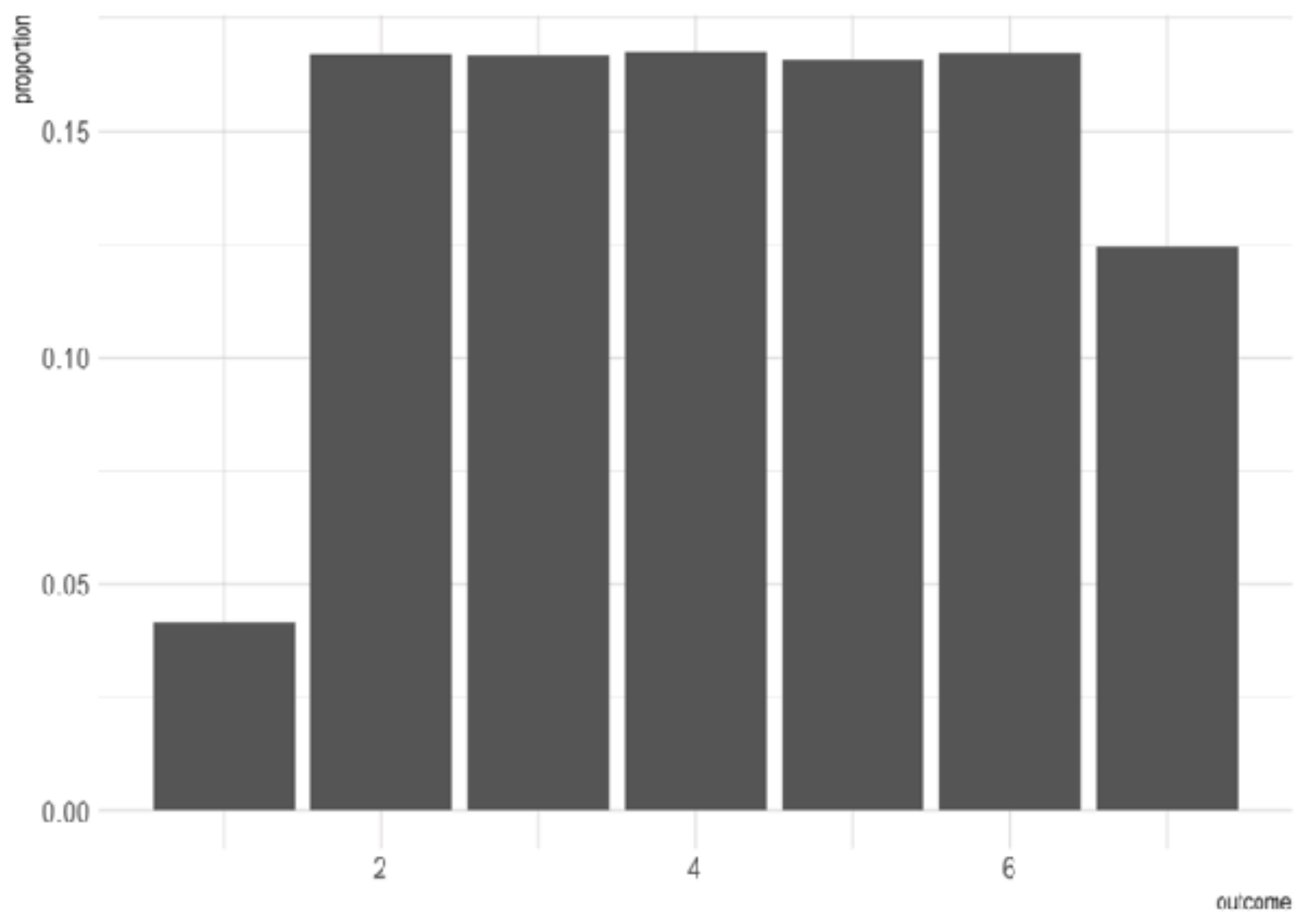
- we can compose random variables with standard mathematical operations
  e.g., $Z = X + Y$, where X and Y are random variables

- easy to conceive of this in terms of samples

```
n_samples <- 1e6
# `n_samples` rolls of a fair dice
samples_x <- sample(
  1:6,
  size = n_samples,
  replace = T
)
```

```
# `n_samples` flips of a biased coin
samples_y <- sample(
  c(0,1),
  prob = c(0.25, 0.75),
  size = n_samples,
  replace = T
)

samples_z <- samples_x + samples_y
```



Probability distributions in R

PROBABILITY DISTRIBUTIONS IN R

- for each distribution `mydist`, there are four types of functions
  - `dmydist(x, ...)` density function gives the (mass/density) $f(x)$ for x
  - `pmydist(x, ...)` cumulative probability function gives cumulative distribution $F(x)$ for x
  - `qmydist(p, ...)` quantile function gives value x with `p = pmydist(x, ...)`
  - `rmydist(n, ...)` random sample function returns `n` samples from the distribution

EXAMPLE ::: NORMAL DISTRIBUTION

```r
# density of standard normal at x = 1
dnorm(x = 1, mean = 0, sd = 1)
```

```
## [1] 0.2419707
```

```r
# cumulative density of standard normal at q = 0
pnorm(q = 0, mean = 0, sd = 1)
```

```
## [1] 0.5
```

```r
# point where the cumulative density of standard normal is p = 0
qnorm(p = 0.5, mean = 0, sd = 1)
```

```
## [1] 0
```

```r
# n = 3 random samples from a standard normal
rnorm(n = 3, mean = 0, sd = 1)
```

```
## [1]  0.5382749 -0.1837154 -0.3165524
```