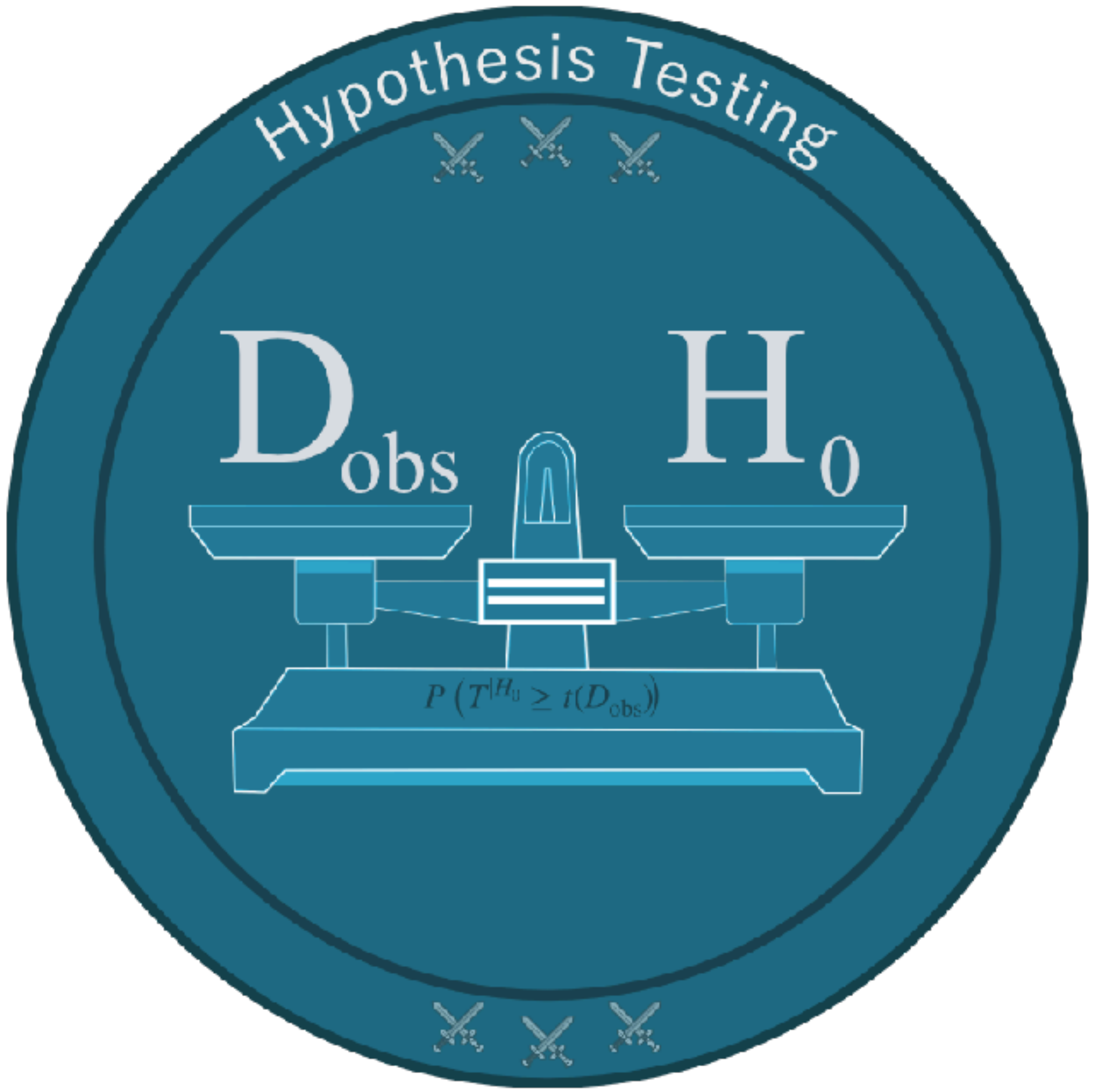


INTRODUCTION TO DATA ANALYSIS

HYPOTHESIS TESTING

PART I

RECAP & OUTLOOK

BAYESIAN PARAMETER ESTIMATION

- model M captures prior beliefs about data-generating process
  - prior over latent parameters
  - likelihood of data
- Bayesian posterior inference using observed data D_{obs}
- compare posterior beliefs to some parameter value of interest

FREQUENTIST HYPOTHESIS TESTING

- model M captures a hypothetically assumed data-generating process
  - fix parameter value of interest
  - likelihood of data
- single out some aspect of the data as most important (test statistic)
- look at distribution of test statistic given the assumed model (sampling distribution)
- check likelihood of test statistic applied to the observed data D_{obs}

CAVEAT

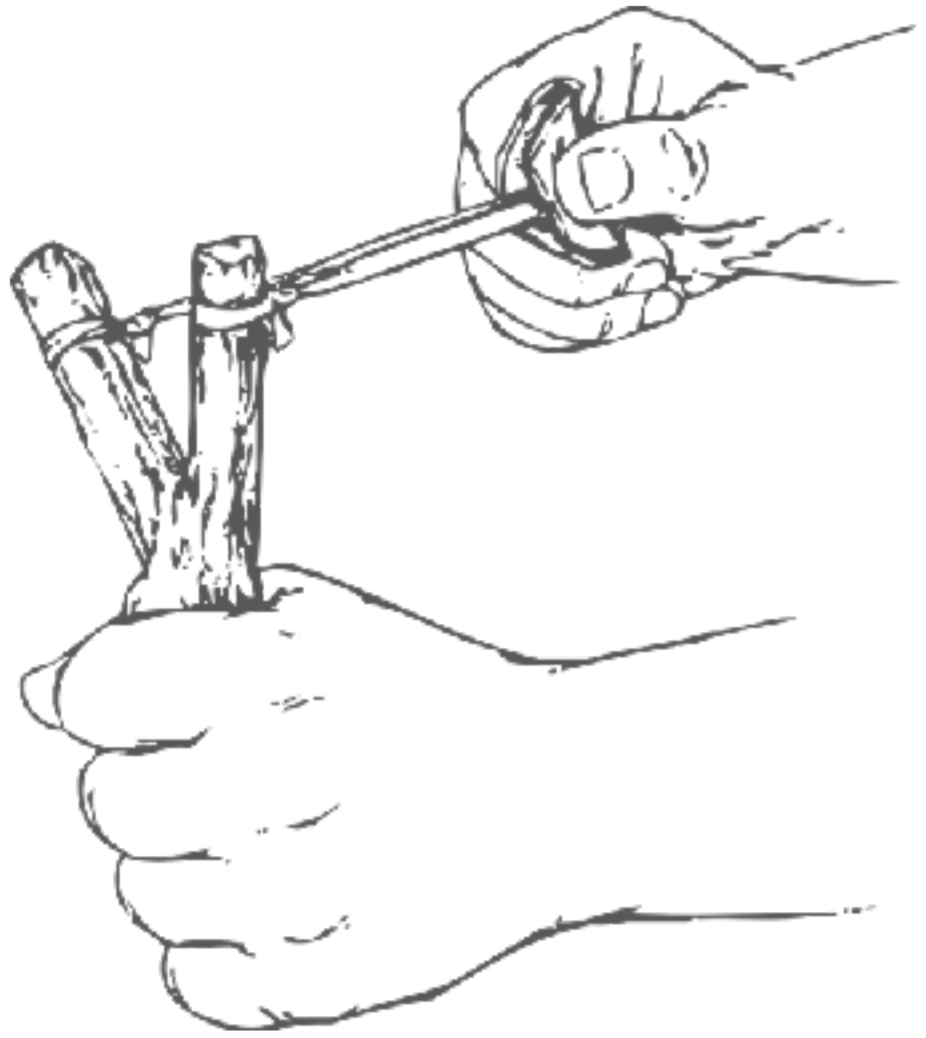
FREQUENTIST HYPOTHESIS TESTING

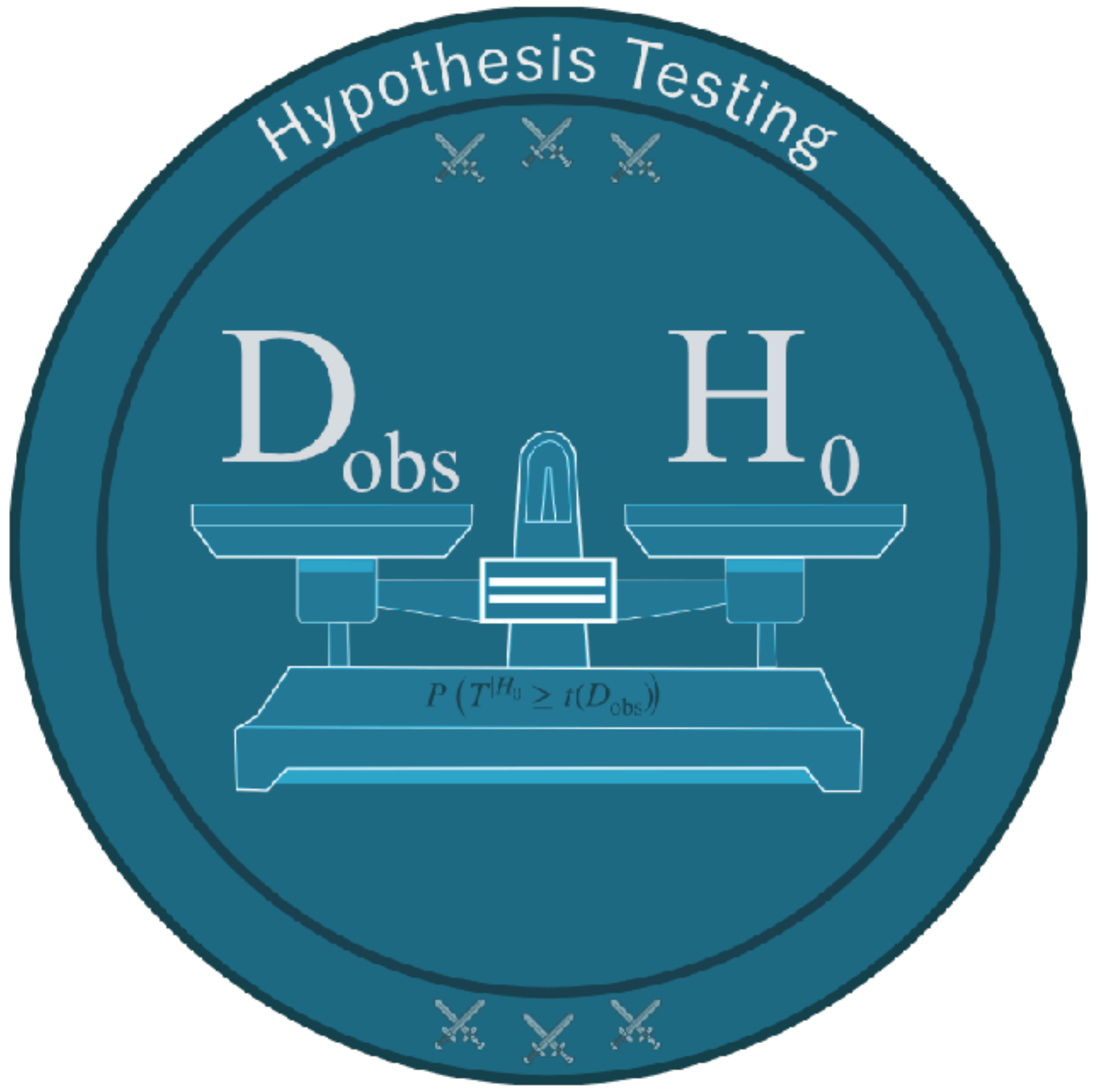
- there are at least three flavors of frequentist hypothesis testing
  - Fisher
  - Neyman-Pearson
  - modern hybrid NHST
    [null-hypothesis significance testing]
- not every text book is clear on these differences and/or which flavor it endorses
- there is also no unanimity of practice between or within research fields

!

LEARNING GOALS

- understand basic idea of frequentist hypothesis testing
- understand what a p-value is
    - definition, one- vs two-sided
    - test statistic & sampling distribution
    - relation to confidence intervals
    - significance levels & α-error





p-value

PRELIMINARIES

- research hypothesis: theoretically implied answer to a main question of interest for research
  - e.g., truth-judgements of sentences with presupposition failure at chance level? (King of France)
  - e.g., faster reactions in *reaction time* trials than in *go/No-go* trials? (Mental Chronometry)
- null hypothesis: specific assumption made for purposes of analysis
  - fix parameter value in a data-generating model for technical reasons
  - analogy: useful assumption in mathematical proof (e.g., in reductio ad absurdum)
- alternative hypothesis: the antagonist of the null hypothesis, specified to relate the null hypothesis to the research hypothesis

P-VALUE

Definition p-value. The p-value associated with observed data D_{obs} gives the probability, derived from the assumption that H_0 is true, of observing an outcome for the chosen test statistic that is at least as extreme evidence against H_0 as the observed outcome.

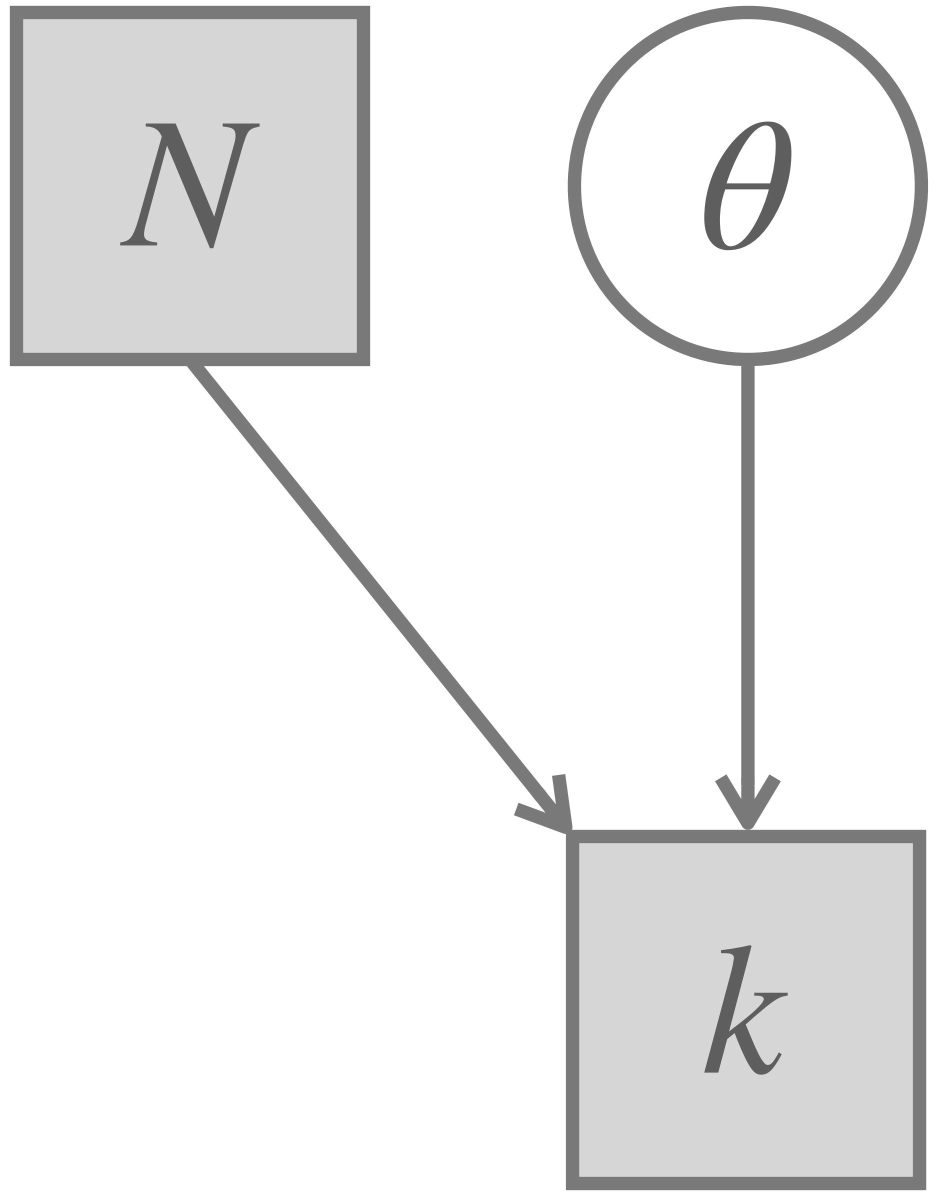
Formally, the p-value of observed data D_{obs} is:

$$p(D_{\text{obs}}) = P(T^{|H_0} \succeq^{H_{0,a}} t(D_{\text{obs}}))$$

where $t \colon \mathcal{D} \rightarrow \mathbb{R}$ is a **test statistic** which picks out a relevant summary statistic of each potential data observation, $T^{|H_0}$ is the **sampling distribution**, namely the random variable derived from test statistic t and the assumption that H_0 is true, and $\succeq^{H_{0,a}}$ is a linear order on the image of t such that $t(D_1) \succeq^{H_{0,a}} t(D_2)$ expresses that test value $t(D_1)$ is at least as extreme evidence *against* H_0 as test value $t(D_2)$.[1]

Binomial Model

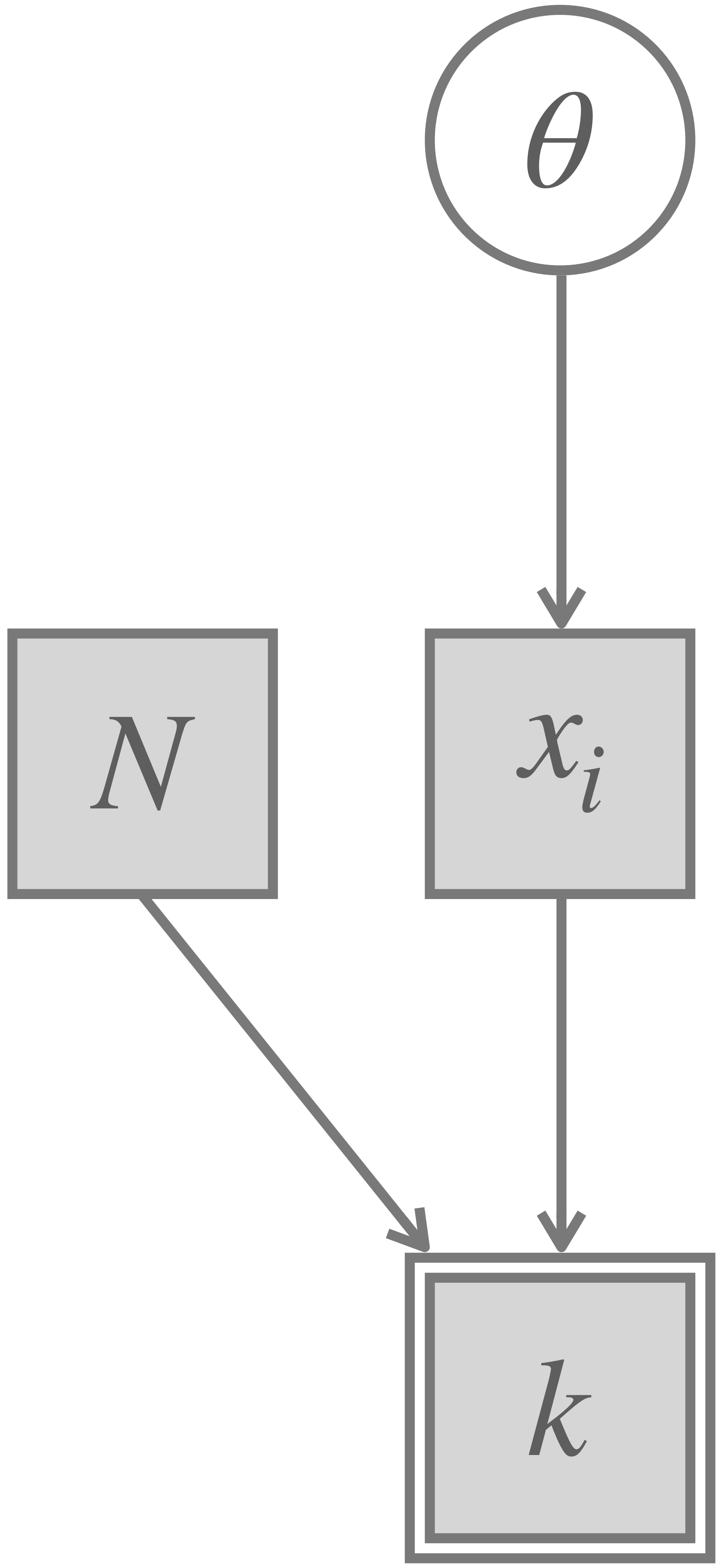
BAYESIAN BINOMIAL MODEL (AS ORIGINALLY INTRODUCED)



$$\theta \sim \text{Beta}(\ldots)$$

$$k \sim \text{Binomial}(\theta, N)$$

BAYESIAN BINOMIAL MODEL (EXTENDED)



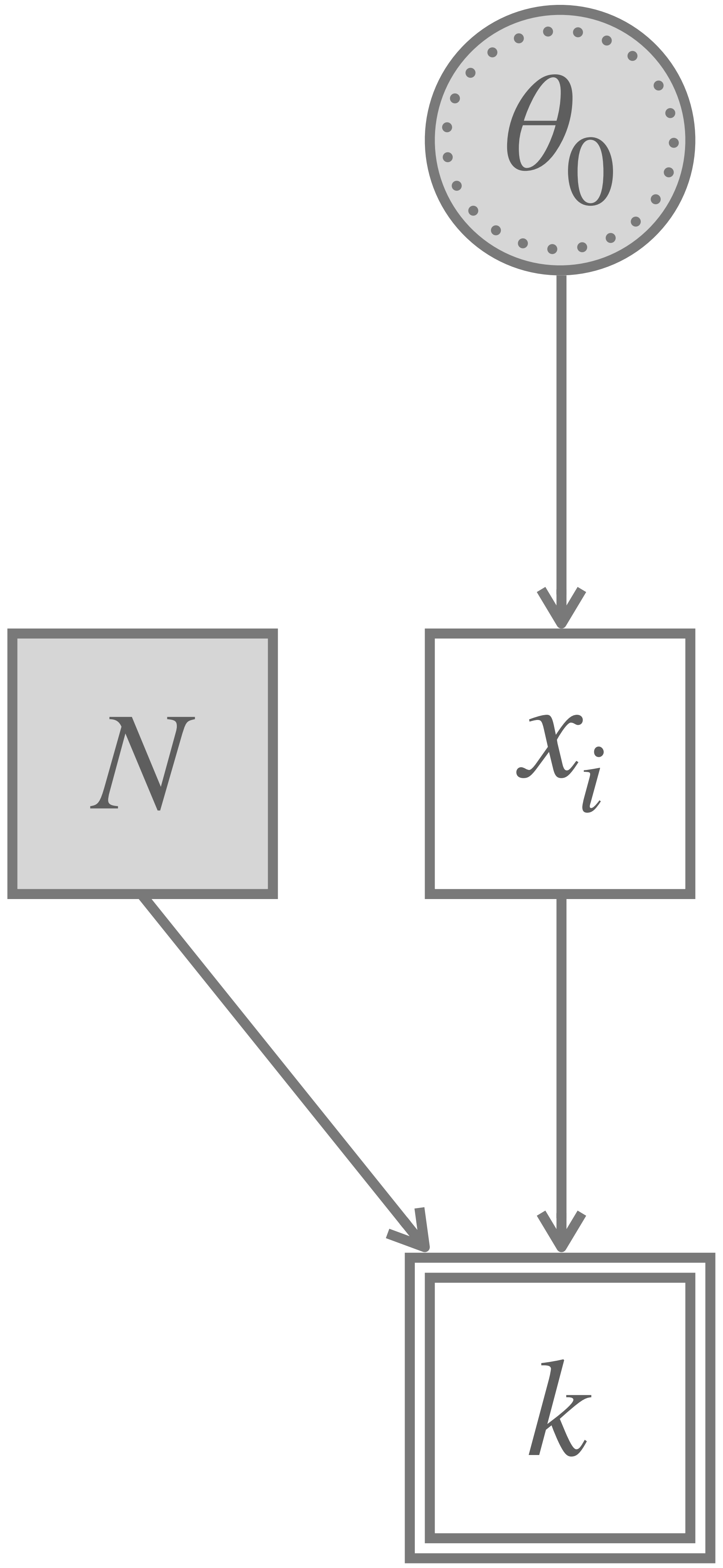
$$\theta \sim \text{Beta}(\ldots)$$

$$x_i \sim \text{Bernoulli}(\theta_0)$$

$$k = \sum_{i=1}^{N} x_i$$

FREQUENTIST BINOMIAL MODEL

[doted line = "working assumption"]



$$x_i \sim \text{Bernoulli}(\theta_0)$$ [likelihood of "raw" data]

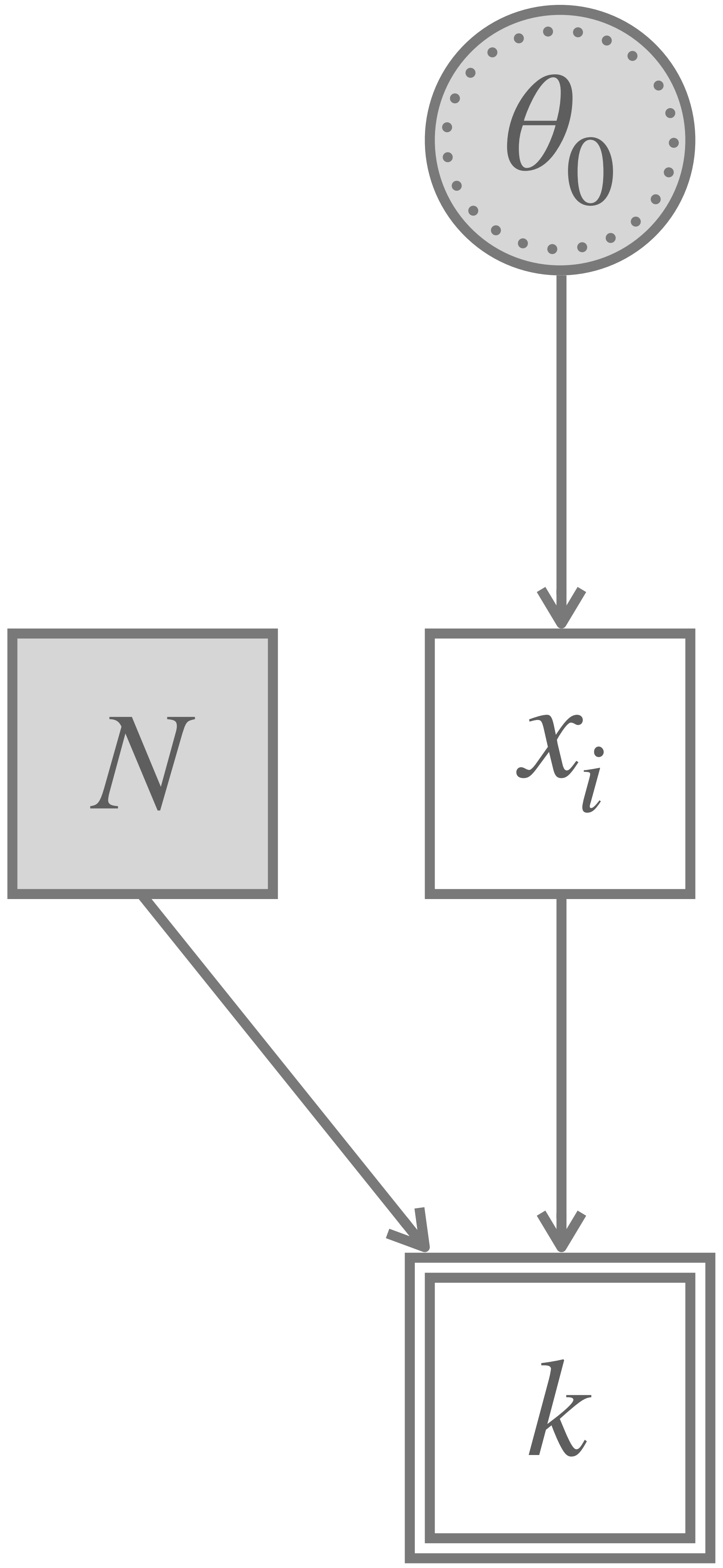
$$k = \sum_{i=1}^{N} x_i$$ [test statistic (derived from "raw" data)]

FACT:

The sampling distribution of k is:

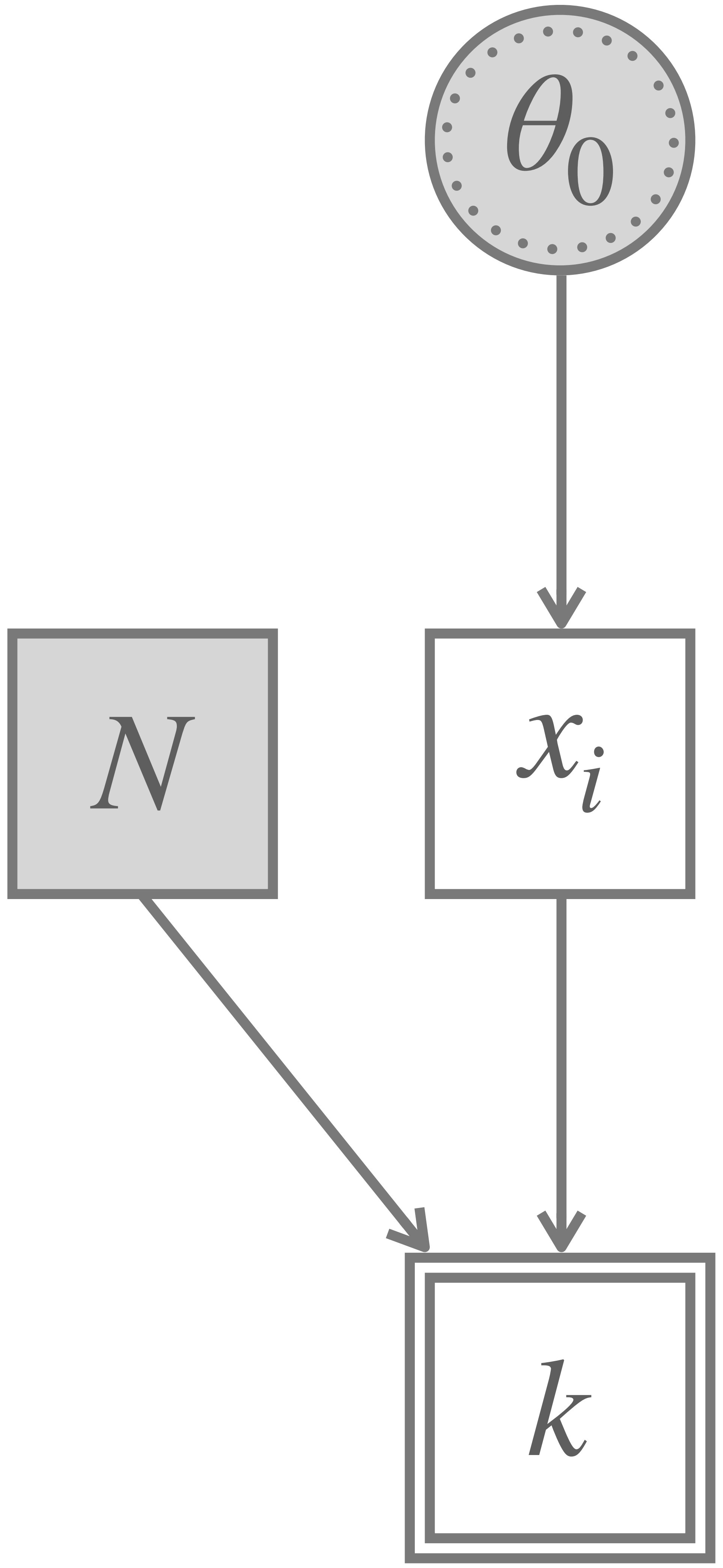
$$k \sim \text{Binomial}(\theta_0, N)$$

FREQUENTIST BINOMIAL MODEL



- null-hypothesis: $\theta = \theta_0$
- test statistic: k derived from "raw" data $\overrightarrow{x}$
  - the most important (numerical) aspect of the data for the current testing purposes
- sampling distribution: likelihood of observing a particular value of k in this model
- notice: the observed data D_{obs} has not yet made any appearance
  - remark: sometimes summary statistics of D_{obs} other than the test statistic might be used in the model

FREQUENTIST BINOMIAL MODEL



- likelihood of data: random variable $\mathcal{D}^{|H_0}$

$$P(\mathcal{D}^{|H_0} = \langle x_1, \ldots, x_N \rangle) = \prod_{i=1}^{N} \text{Bernoulli}(x_i, \theta_0)$$

- sampling distribution: random variable $T^{|H_0}$

$$P(T^{|H_0} = k) = \text{Binomial}(k, \theta_0, N)$$

Binomial p-values

BINOMIAL TEST

- 24/7 example: $N = 24$ and $k = 7$
  - $t(D_{\text{obs}}) = 7$
  - $P(T^{|H_0} = k) = \text{Binomial}(k, \theta_0, N)$
- p-value definition:

$$p(D_{\text{obs}}) = P(\boxed{T^{|H_0}}\ \boxed{\succeq^{H_{0,a}}}\ \boxed{t(D_{\text{obs}})})$$

we know this ??? we know this

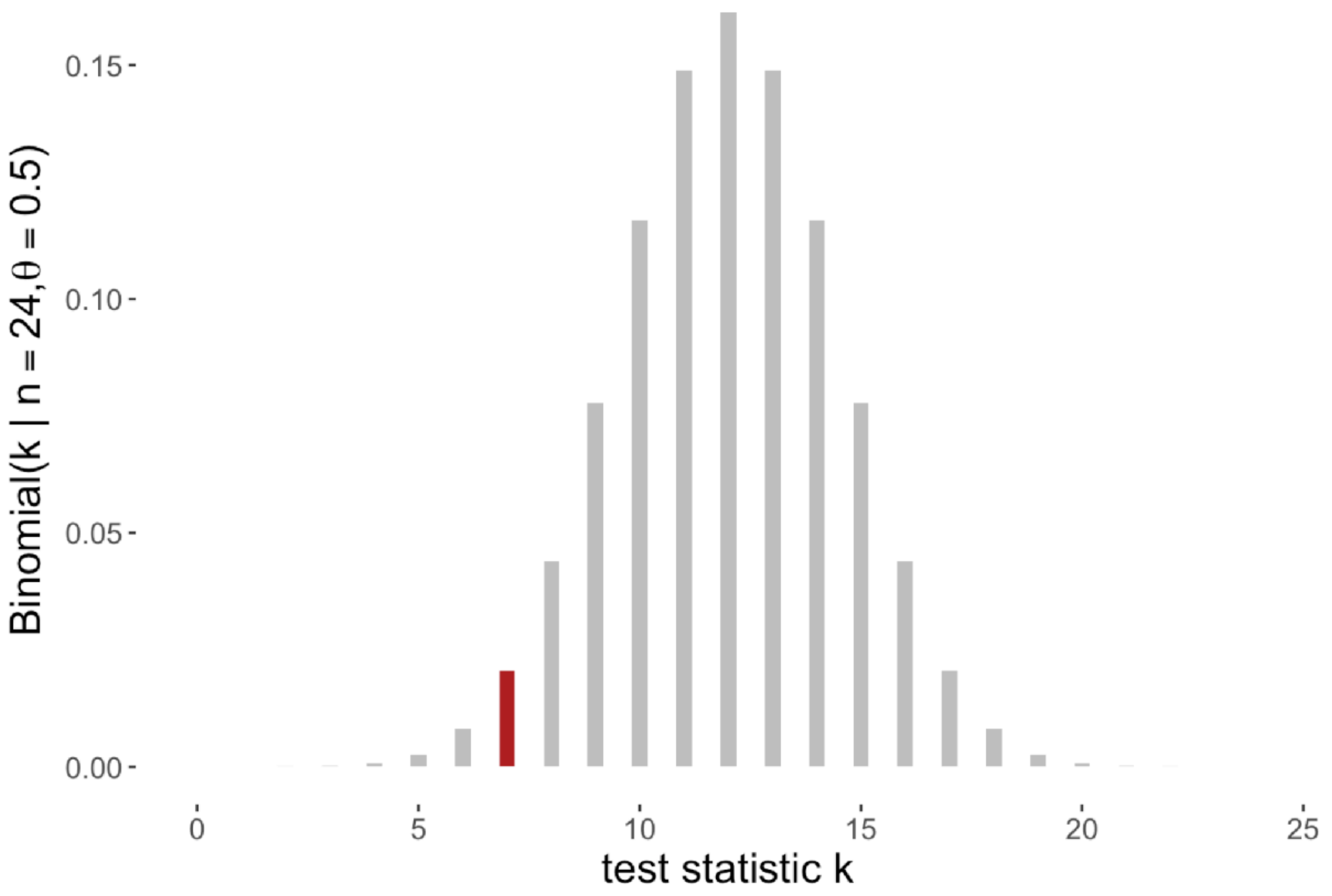
What counts as "more extreme evidence against the null hypothesis" is a context-sensitive notion that depends on the null-hypothesis *and* the alternative hypothesis because only when put together do null- and alternative hypothesis address the research question in the background.

BINOMIAL TEST

- compare two research questions
  1. Is the coin fair?
     - $H_0: \theta = 0.5$
     - $H_a: \theta \neq 0.5$
  2. Is the coin biased towards heads?
     - $H_0: \theta = 0.5$
     - $H_a: \theta < 0.5$
- we still use a point-valued null-hypothesis for technical reasons
- the alternative hypothesis is important to fix the meaning of $\succeq^{H_{0,a}}$

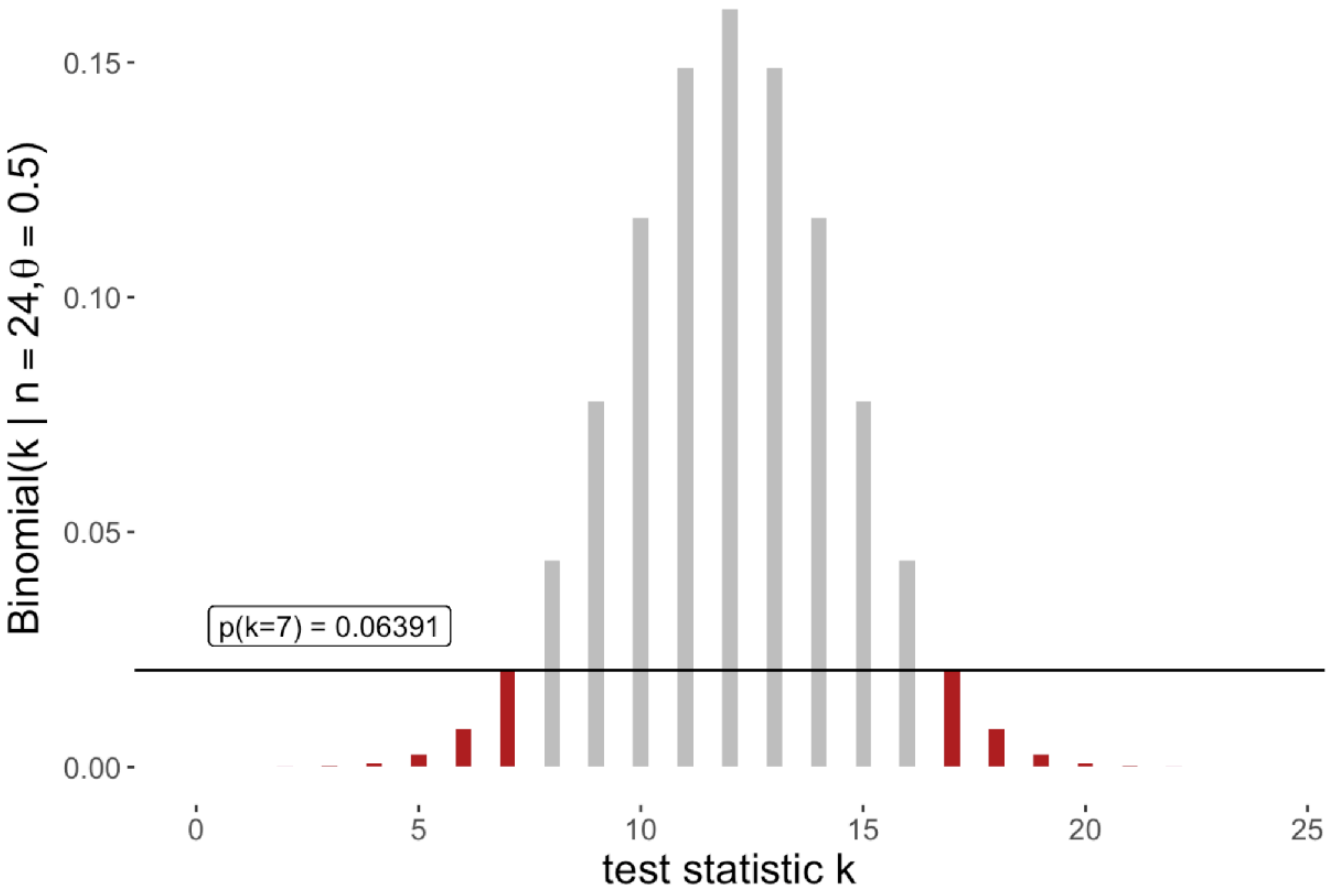
BINOMIAL TEST

- Case 1: Is the coin fair?
  - $H_0: \theta = 0.5$
  - $H_a: \theta \neq 0.5$
- which values of k are more extreme evidence against H_0?

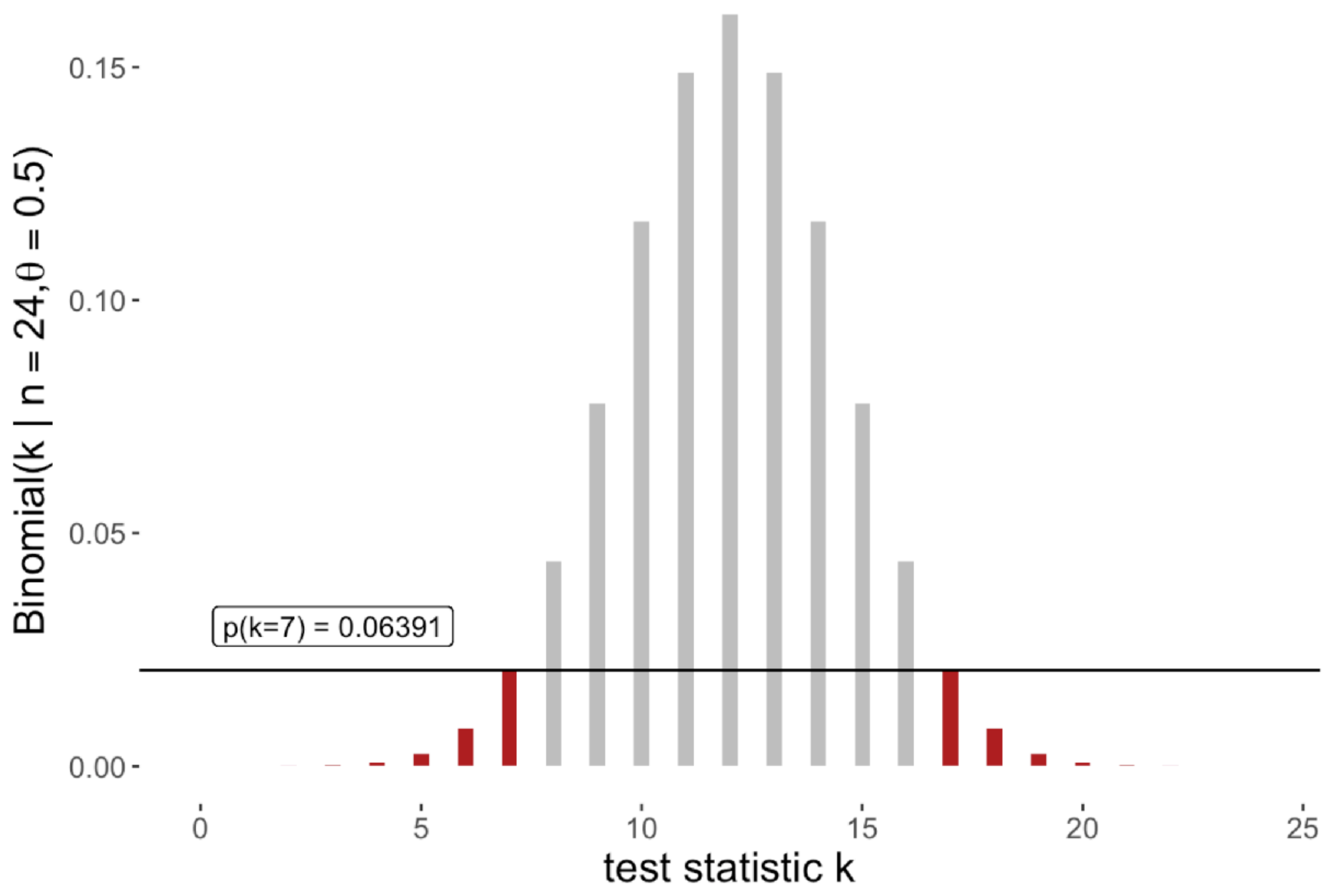


BINOMIAL TEST

- Case 1: Is the coin fair?
  - $H_0: \theta = 0.5$
  - $H_a: \theta \neq 0.5$
- which values of k are more extreme evidence against H_0?
  - anything that's even less likely to occur



BINOMIAL TEST



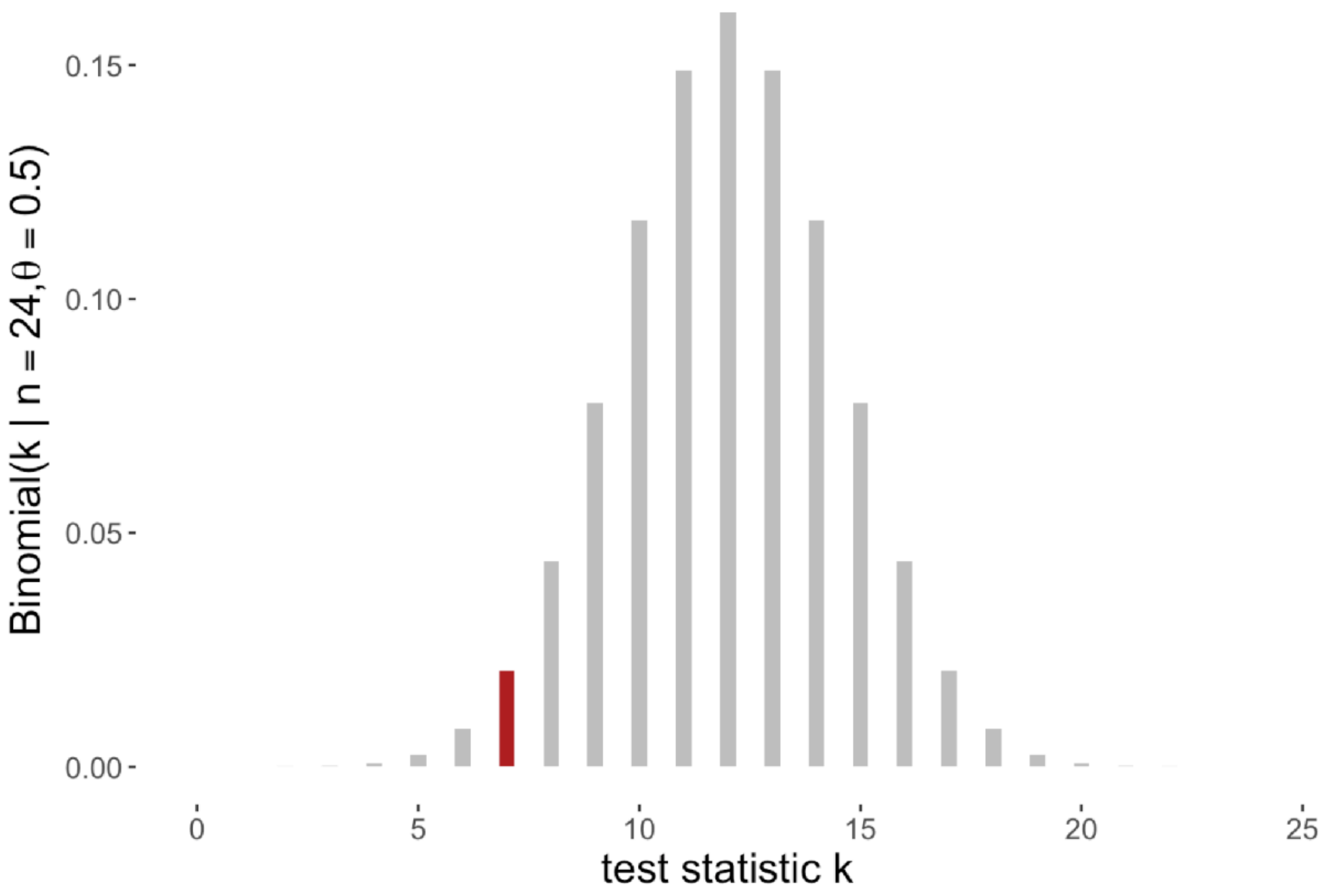

```r
# exact p-value for k=7 with N=24 and null-hypothesis theta = 0.5
k_obs <- 7
N <-  24
theta_0 <-  0.5
tibble( lh = dbinom(0:N, N, theta_0) ) %>%
  filter( lh <=  dbinom(k_obs, N, theta_0) ) %>%
  pull(lh) %>% sum %>% round(5)
```

```
## [1] 0.06391
```

$$p(k) = \sum_{k'=0}^{N} [\text{Binomial}(k', N, \theta_0) <= \text{Binomial}(k, N, \theta_0)] \text{ Binomial}(k', N, \theta_0)$$

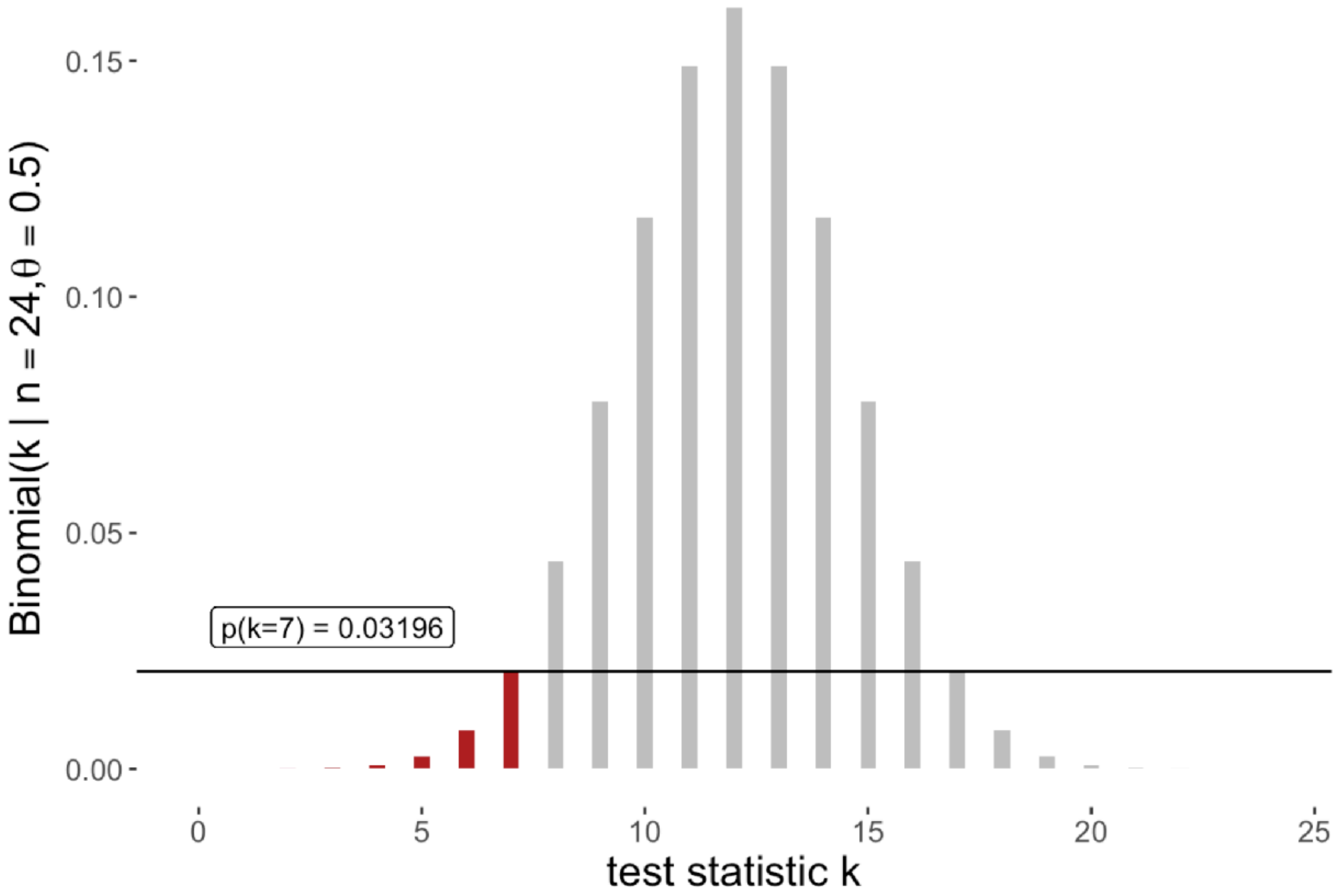
BINOMIAL TEST

- Case 2: Is the coin biased towards heads?
  - $H_0 : \theta = 0.5$
  - $H_a : \theta < 0.5$
- which values of k are more extreme evidence against H_0?

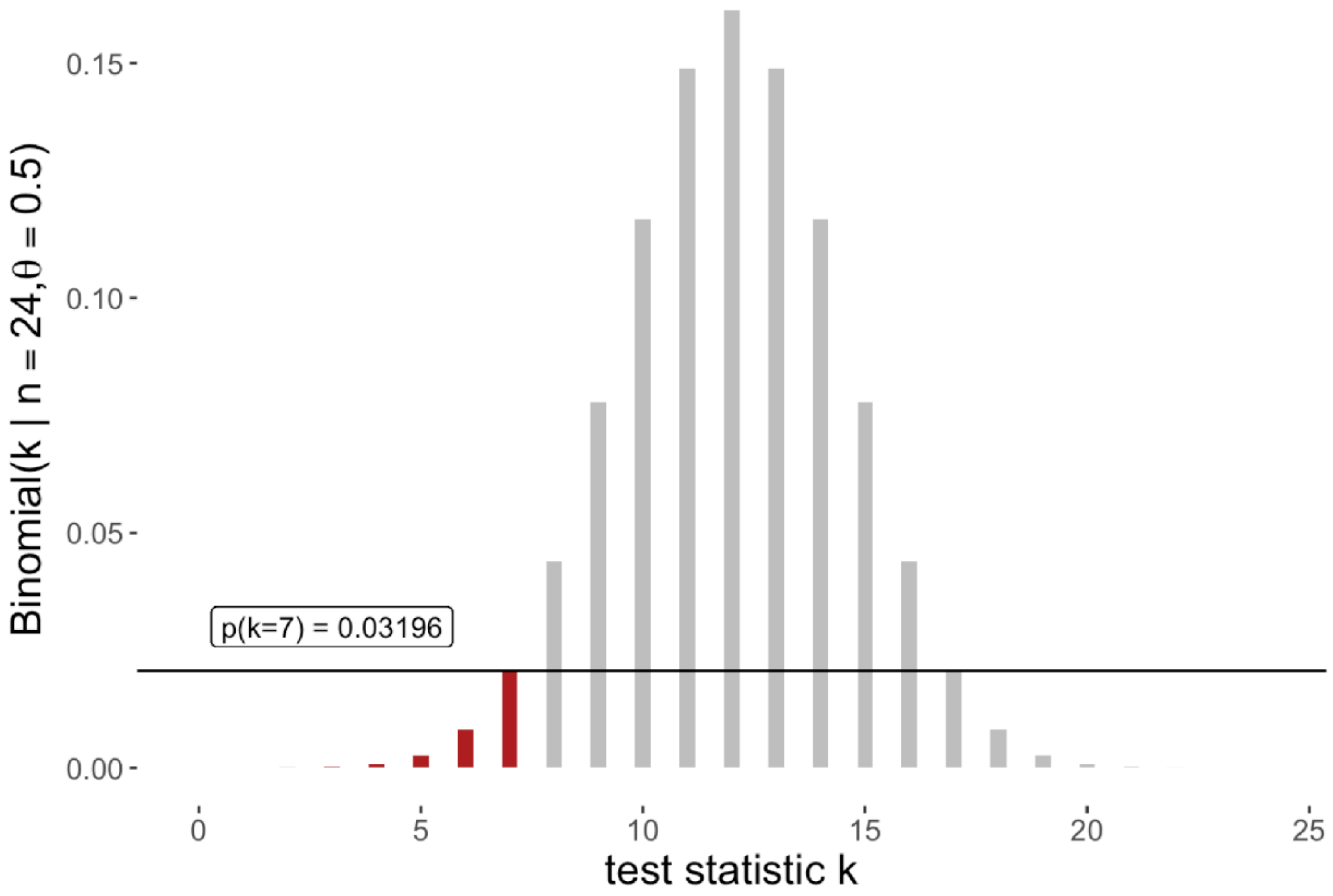


BINOMIAL TEST

- Case 2: Is the coin biased towards heads?
  - $H_0: \theta = 0.5$
  - $H_a: \theta < 0.5$
- which values of k are more extreme evidence against H_0?
  - anything even more in favor of H_a



BINOMIAL TEST




```r
binom.test(
  x = 7,       # observed successes
  n = 24,      # total nr. observations
  p = 0.5,      # null hypothesis
  alternative = "less" # the alternative to compare against is theta < 0.5
)
```

```
##
##  Exact binomial test
##
## data:  7 and 24
## number of successes = 7, number of trials = 24, p-value = 0.03196
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.4787279
## sample estimates:
## probability of success
##              0.2916667
```

p-value
revisit

P-VALUE

Definition p-value. The p-value associated with observed data D_{obs} gives the probability, derived from the assumption that H_0 is true, of observing an outcome for the chosen test statistic that is at least as extreme evidence against H_0 as the observed outcome.

Formally, the p-value of observed data D_{obs} is:

$$p(D_{\text{obs}}) = P(T^{|H_0} \succeq^{H_{0,a}} t(D_{\text{obs}}))$$

where $t \colon \mathcal{D} \to \mathbb{R}$ is a **test statistic** which picks out a relevant summary statistic of each potential data observation, $T^{|H_0}$ is the **sampling distribution**, namely the random variable derived from test statistic t and the assumption that H_0 is true, and $\succeq^{H_{0,a}}$ is a linear order on the image of t such that $t(D_1) \succeq^{H_{0,a}} t(D_2)$ expresses that test value $t(D_1)$ is at least as extreme evidence *against* H_0 as test value $t(D_2)$.[1]

significance and α-errors

SIGNIFICANCE LEVELS

- standardly we fix a significance level α before the test
- common values of α are:
  - $\alpha = 0.05$
  - $\alpha = 0.01$
  - $\alpha = 0.001$
- if the *p*-value for the observed data passes the pre-established threshold of significance, we say that the test result was significant
- a significant test result is conventionally regarded as "strong enough" evidence against the null-hypothesis, so that we can reject the null hypothesis as a viable explanation of the data
- non-significant results are interpreted differently in different approaches (more later)

α-ERROR

- an α-error (aka type-I error) occurs when we reject a true null hypothesis
- by definition this type of error occurs, in the long run, with a proportion of no more than α
- it is in this way that frequentist statistic is subscribed and cherishes a regime of long-term error control on research results
- Bayesian approaches (usually) are not concerned with long-term error control