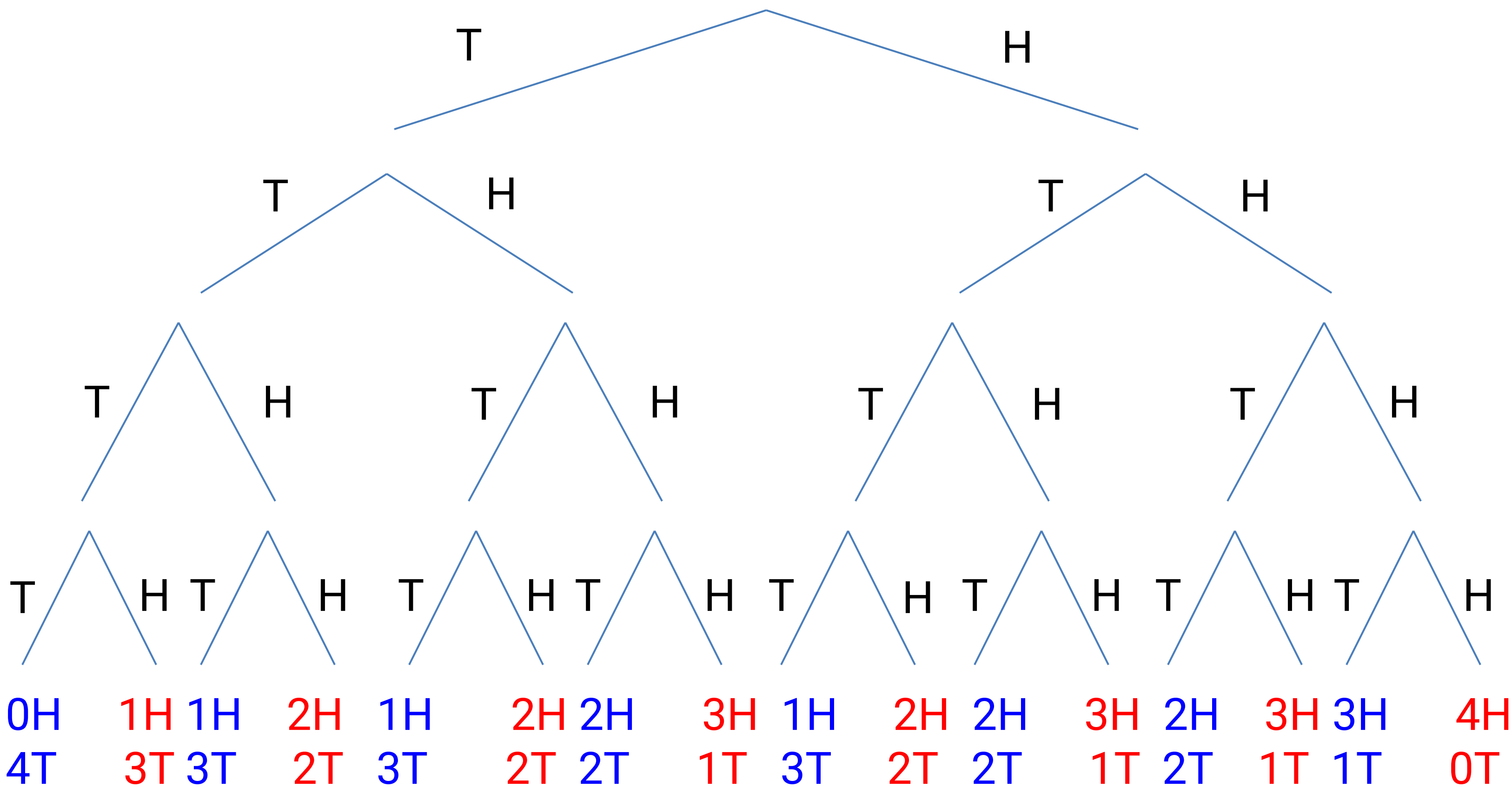
The binomial test

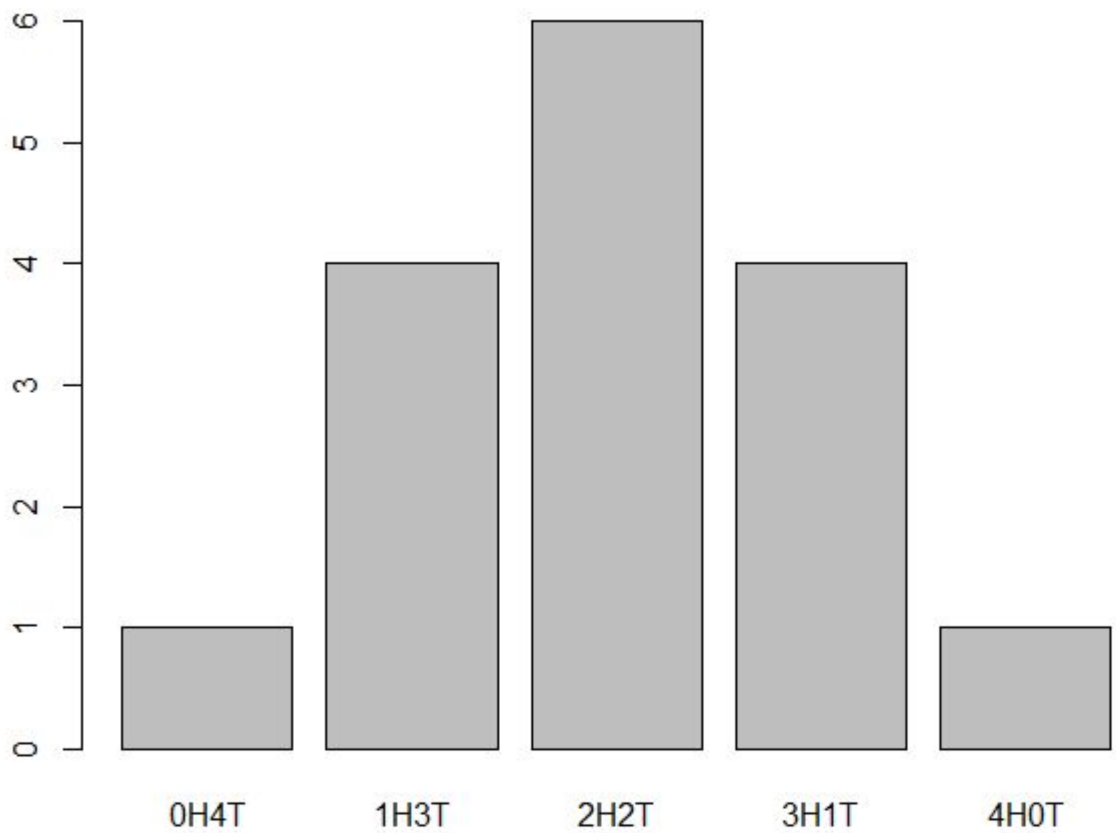
A preview of (almost) everything

Is the coin fair?



All the possible outcomes for 4 tosses of a coin

Distribution of frequencies:
Binomial distribution, with *p*(Head) = 0.5



The binomial distribution

Let:

- *n* (or *N*): number of dichotomous ("Bernoulli") trials
- *x*: number of successful trials
- *p*: probability of a successful trial

Then the probability of obtaining exactly *x* successful trials is given by:

```
> choose(n, x) *
  p^x (1 - p)^(n - x)
```

Introducing `dbinom`

```
> dbinom(0, size = 4, prob = .5)
[1] 0.0625
> dbinom(1, size = 4, prob = .5)
[1] 0.25
> dbinom(2, size = 4, prob = .5)
[1] 0.375
> dbinom(3, size = 4, prob = .5)
[1] 0.25
> dbinom(4, size = 4, prob = .5)
[1] 0.0625
```

Introducing `pbinom`

```
> pbinom(0, size = 4, prob = .5)
[1] 0.0625
> pbinom(1, size = 4, prob = .5)
[1] 0.3125
> pbinom(2, size = 4, prob = .5)
[1] 0.6875
> pbinom(3, size = 4, prob = .5)
[1] 0.9375
> pbinom(4, size = 4, prob = .5)
[1] 1
```

Binomial Probabilities for P = .5

N = total number of events
X = number of events of one particular type (e.g., heads)
p(X) = prob that the number of events of this type is exactly X
p(<=X) = prob that the number of events of this type is less than or equal to X

N	X	p(X)	p(<=X)
1	0	0.500	0.500
1	1	0.500	1.000
2	0	0.250	0.250
2	1	0.500	0.750
2	2	0.250	1.000
3	0	0.125	0.125
3	1	0.375	0.500
3	2	0.375	0.875
3	3	0.125	1.000
4	0	0.063	0.063
4	1	0.250	0.313
4	2	0.375	0.688
4	3	0.250	0.938
4	4	0.063	1.000
5	0	0.031	0.031
5	1	0.156	0.188
5	2	0.313	0.500
5	3	0.313	0.813
5	4	0.156	0.969
5	5	0.031	1.000
6	0	0.016	0.016
6	1	0.094	0.109
6	2	0.234	0.344
6	3	0.313	0.656
6	4	0.234	0.891
6	5	0.094	0.984
6	6	0.016	1.000
7	0	0.008	0.008
7	1	0.055	0.063
7	2	0.164	0.227
7	3	0.273	0.500
7	4	0.273	0.773
7	5	0.164	0.938
7	6	0.055	0.992
7	7	0.008	1.000
8	0	0.004	0.004
8	1	0.031	0.035
8	2	0.109	0.145
8	3	0.219	0.363
8	4	0.273	0.637
8	5	0.219	0.855
8	6	0.109	0.965
8	7	0.031	0.996
8	8	0.004	1.000

N	X	p(X)	p(<=X)
9	0	0.002	0.002
9	1	0.018	0.020
9	2	0.070	0.090
9	3	0.164	0.254
9	4	0.246	0.500
9	5	0.246	0.746
9	6	0.164	0.910
9	7	0.070	0.980
9	8	0.018	0.998
9	9	0.002	1.000
10	0	0.001	0.001
10	1	0.010	0.011
10	2	0.044	0.055
10	3	0.117	0.172
10	4	0.205	0.377
10	5	0.246	0.623
10	6	0.205	0.828
10	7	0.117	0.945
10	8	0.044	0.989
10	9	0.010	0.999
10	10	0.001	1.000
11	0	0.000	0.000
11	1	0.005	0.006
11	2	0.027	0.033
11	3	0.081	0.113
11	4	0.161	0.274
11	5	0.226	0.500
11	6	0.226	0.726
11	7	0.161	0.887
11	8	0.081	0.967
11	9	0.027	0.994
11	10	0.005	1.000
11	11	0.000	1.000
12	0	0.000	0.000
12	1	0.003	0.003
12	2	0.016	0.019
12	3	0.054	0.073
12	4	0.121	0.194
12	5	0.193	0.387
12	6	0.226	0.613
12	7	0.193	0.806
12	8	0.121	0.927
12	9	0.054	0.981
12	10	0.016	0.997
12	11	0.003	1.000
12	12	0.000	1.000

N	X	p(X)	p(<=X)
13	0	0.000	0.000
13	1	0.002	0.002
13	2	0.010	0.011
13	3	0.035	0.046
13	4	0.087	0.133
13	5	0.157	0.291
13	6	0.209	0.500
13	7	0.209	0.709
13	8	0.157	0.867
13	9	0.087	0.954
13	10	0.035	0.989
13	11	0.010	0.998
13	12	0.002	1.000
13	13	0.000	1.000
14	0	0.000	0.000
14	1	0.001	0.001
14	2	0.006	0.006
14	3	0.022	0.029
14	4	0.061	0.090
14	5	0.122	0.212
14	6	0.183	0.395
14	7	0.209	0.605
14	8	0.183	0.788
14	9	0.122	0.910
14	10	0.061	0.971
14	11	0.022	0.994
14	12	0.006	0.999
14	13	0.001	1.000
14	14	0.000	1.000
15	0	0.000	0.000
15	1	0.000	0.000
15	2	0.003	0.004
15	3	0.014	0.018
15	4	0.042	0.059
15	5	0.092	0.151
15	6	0.153	0.304
15	7	0.196	0.500
15	8	0.196	0.696
15	9	0.153	0.849
15	10	0.092	0.941
15	11	0.042	0.982
15	12	0.014	0.996
15	13	0.003	1.000
15	14	0.000	1.000
15	15	0.000	1.000

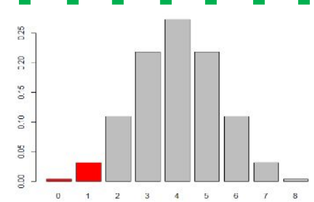
The sign test

Do infants distinguish between their mother's voice and that of another adult female?
Nine infants were tested to see how much time they would spend looking at a loud-
speaker playing their mother's voice or that of another woman. Here are the hypothetical data in seconds of looking time:

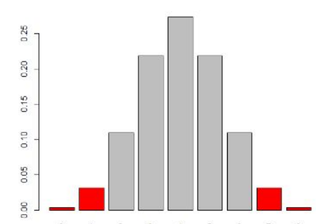
Infant:	a	b	c	d	e	f	g	h	I
Mother's voice	5	7	4	2	5	5	7	6	4
Other's voice	3	6	5	2	4	4	6	5	3
Sign	+	+	-	tie	+	+	+	+	+
(Heads & Tails:	**[H]**	**[H]**	**[T]**		**[H]**	**[H]**	**[H]**	**[H]**	**[H])**

Discard ties

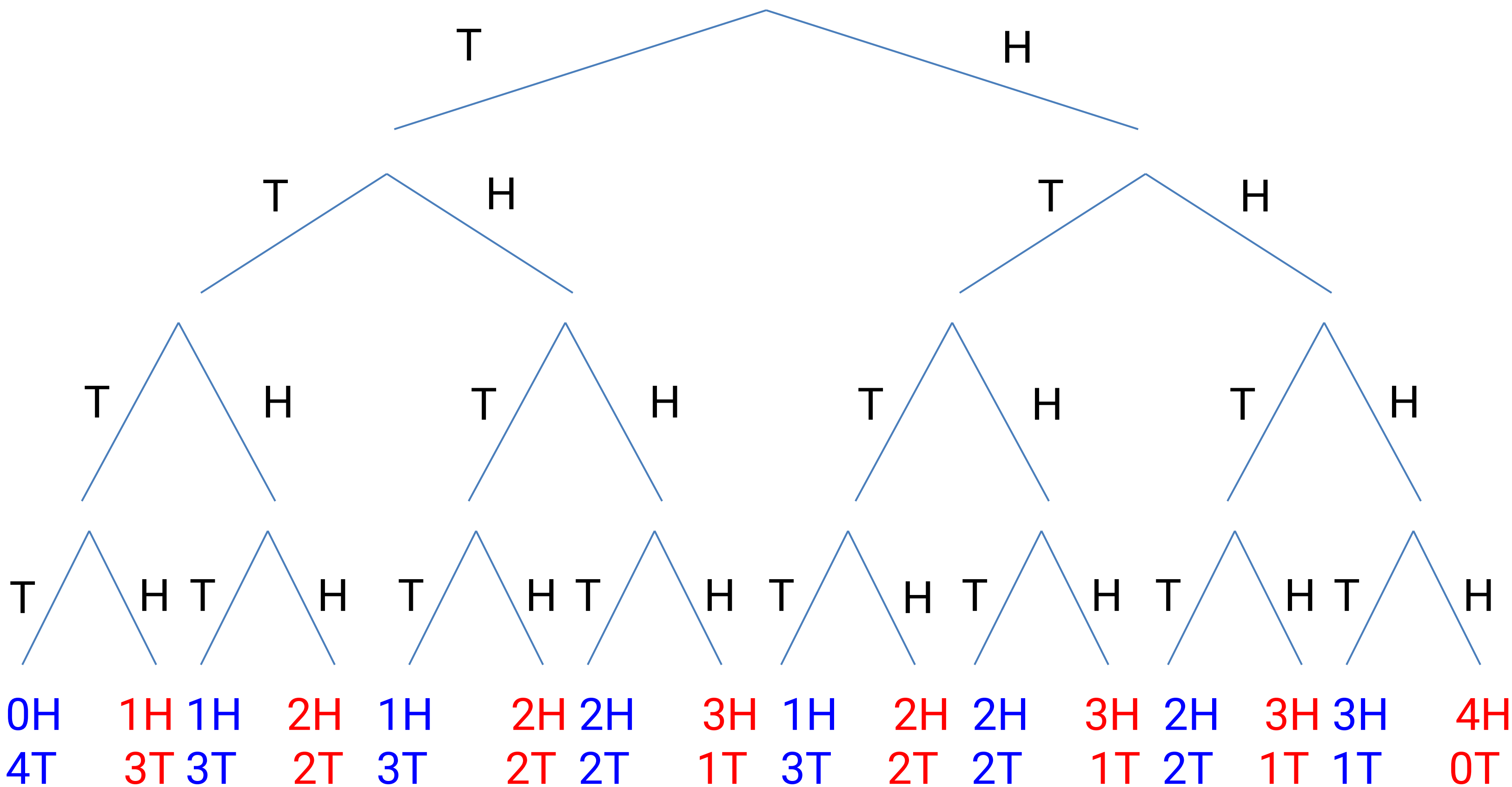
H_0: equally likely to have + or –.
> pbinom(1, size = 8, prob = .5)
[1] 0.03515625



2-tailed = p(number of – 's ≤ 1) + p(number of + 's ≤ 1)
= .035 + .035
= .070



Is the coin fair?



The probability is the product of probabilities of the individual tosses; e.g., .4 x .4 x .4 x .4.

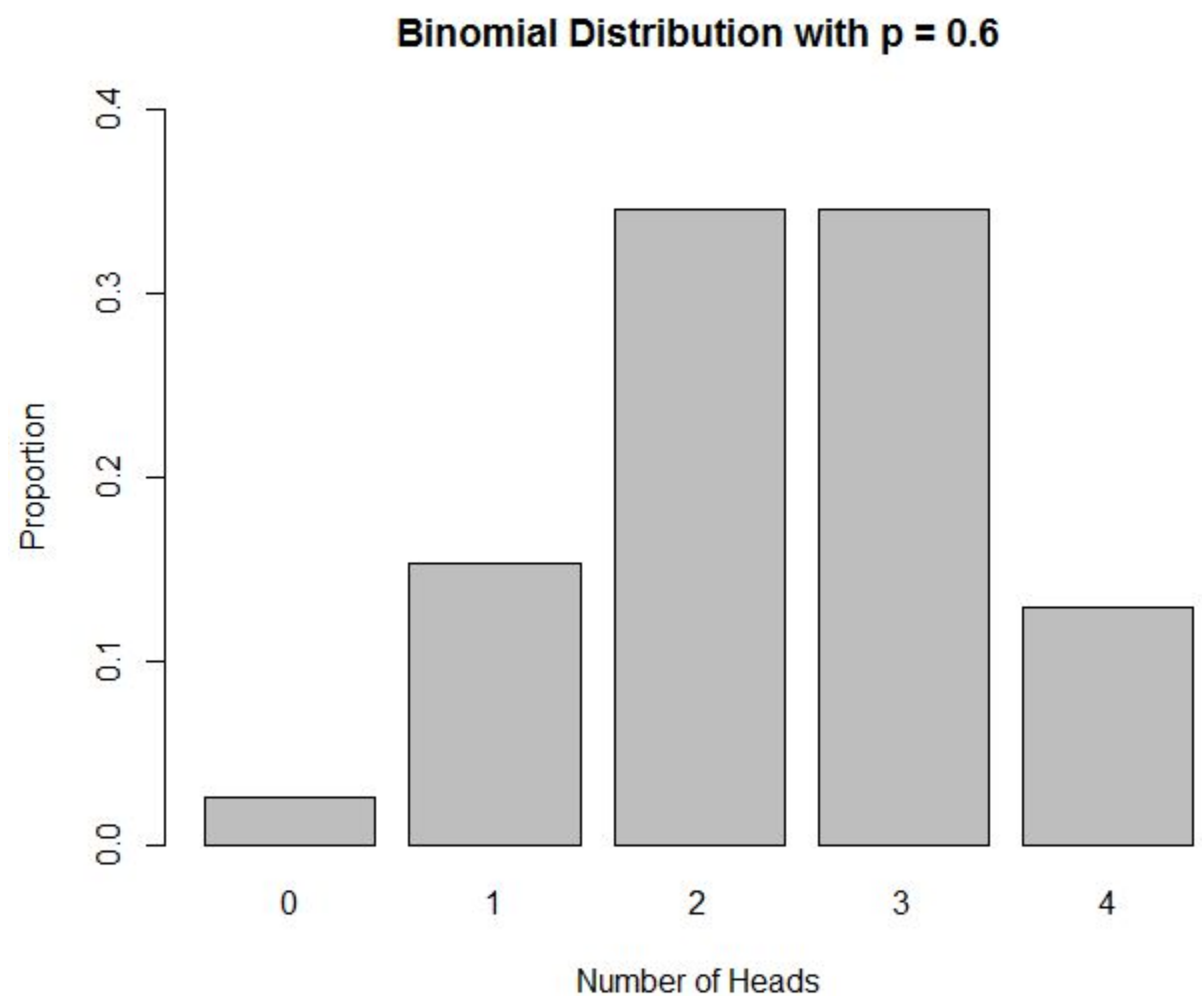
Comparing two binomials

With p = .5:

```
> dbinom(0, 4, .5)
[1] 0.0625
> dbinom(1, 4, .5)
[1] 0.25
> dbinom(2, 4, .5)
[1] 0.375
> dbinom(3, 4, .5)
[1] 0.25
> dbinom(4, 4, .5)
[1] 0.0625
```

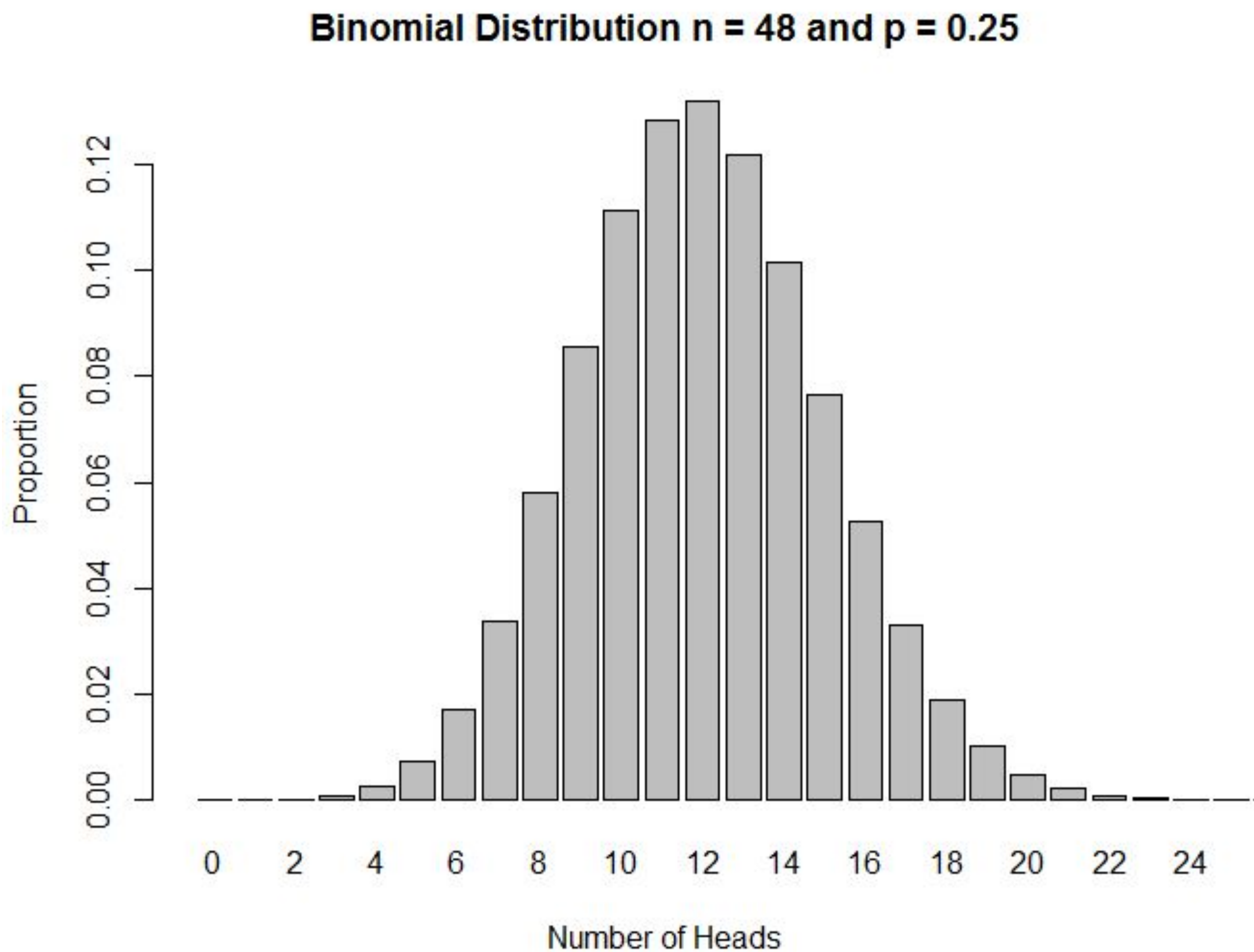
With p = .6:

```
> dbinom(0, 4, .6)
[1] 0.0256
> dbinom(1, 4, .6)
[1] 0.1536
> dbinom(2, 4, .6)
[1] 0.3456
> dbinom(3, 4, .6)
[1] 0.3456
> dbinom(4, 4, .6)
[1] 0.1296
```



The mean is given by Np = 4 x .6 = 2.4H1.6T. When $p \neq$ 0.5, the distribution is skewed, but skew decreases as N increases.

Binomial distributions for larger *N*



Rule of thumb: If $p \neq 0.5$ and $Np(1 - p) \geq 9$, then the normal distribution approximates the binomial.

From sample to population

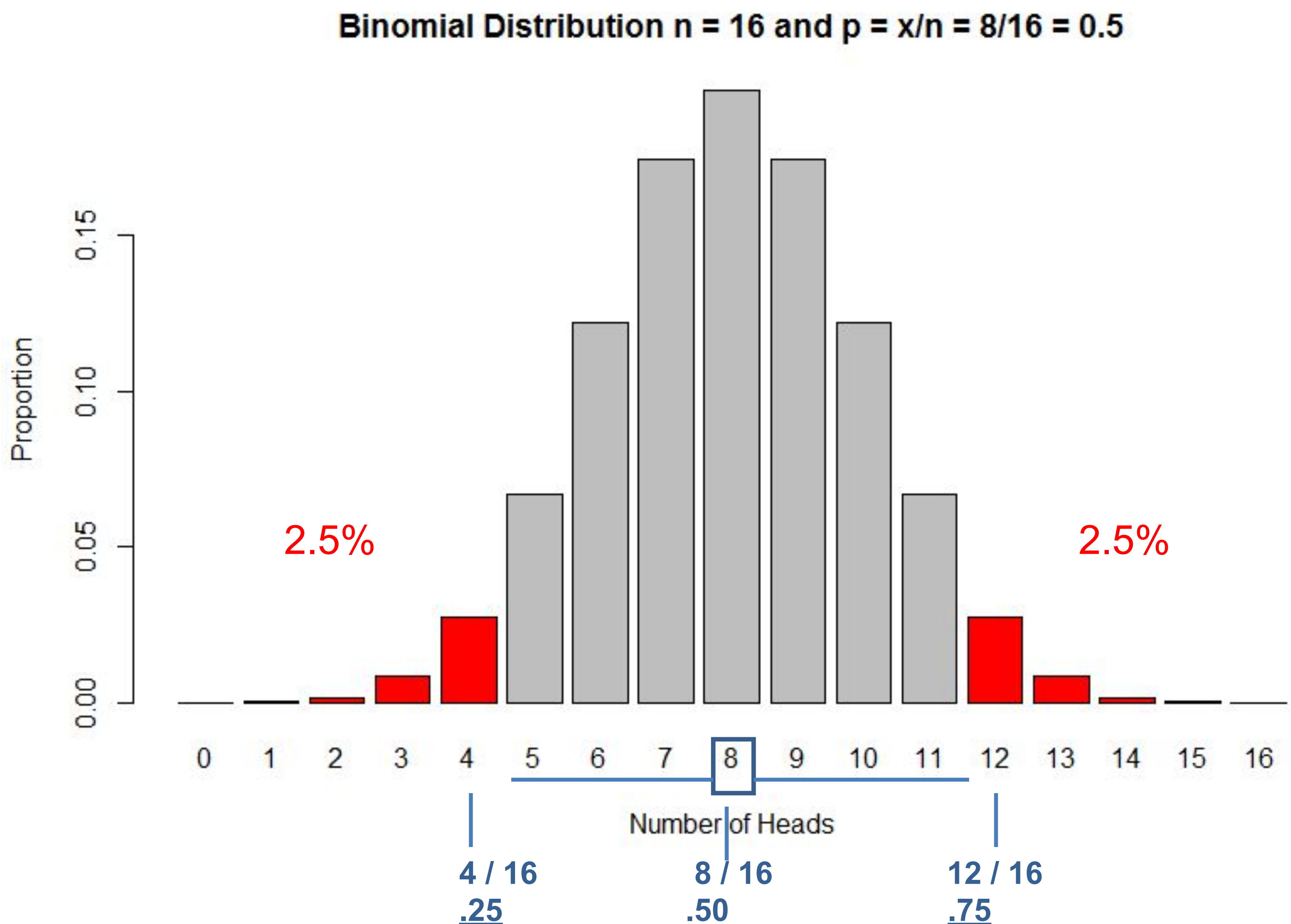
We:

- observe a **sample** of n Bernoulli trials with x outcomes of one type, and
- calculate a **statistic** p, the proportion of outcomes of that one type in the sample such that $p = x / n$, then
- given this sample, compute a range of likely values for the **population parameter** p, i.e., a confidence interval.

From population to sample

We:

- hypothesized the **parameter** p, the probability of an outcome of a particular type in a **binomial** population,
- observed a **sample** of n Bernoulli trials with x outcomes of that type, and
- calculated the probability that the sample came from that binomial population.



Based on the *sample* proportion (8 / 16 = .5), we would expect, with 95% confidence, that the value of the *population proportion* lies between 4 / 16 = .25 and 12 / 16 = .75.

Introducing `binom.test`

```
> binom.test(8, 16, .5)

    Exact binomial test

data:  8 and 16
number of successes = 8, number of trials = 16, p-value =
1
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
 0.2465101 0.7534899
sample estimates:
probability of success
                   0.5
```

Introducing `binom.test`

```
> binom.test(3, 16, .5)
```

    Exact binomial test

data: 3 and 16

number of successes = 3, number of trials = 16, **p-value = 0.02127**

alternative hypothesis: true probability of success is not equal to 0.5

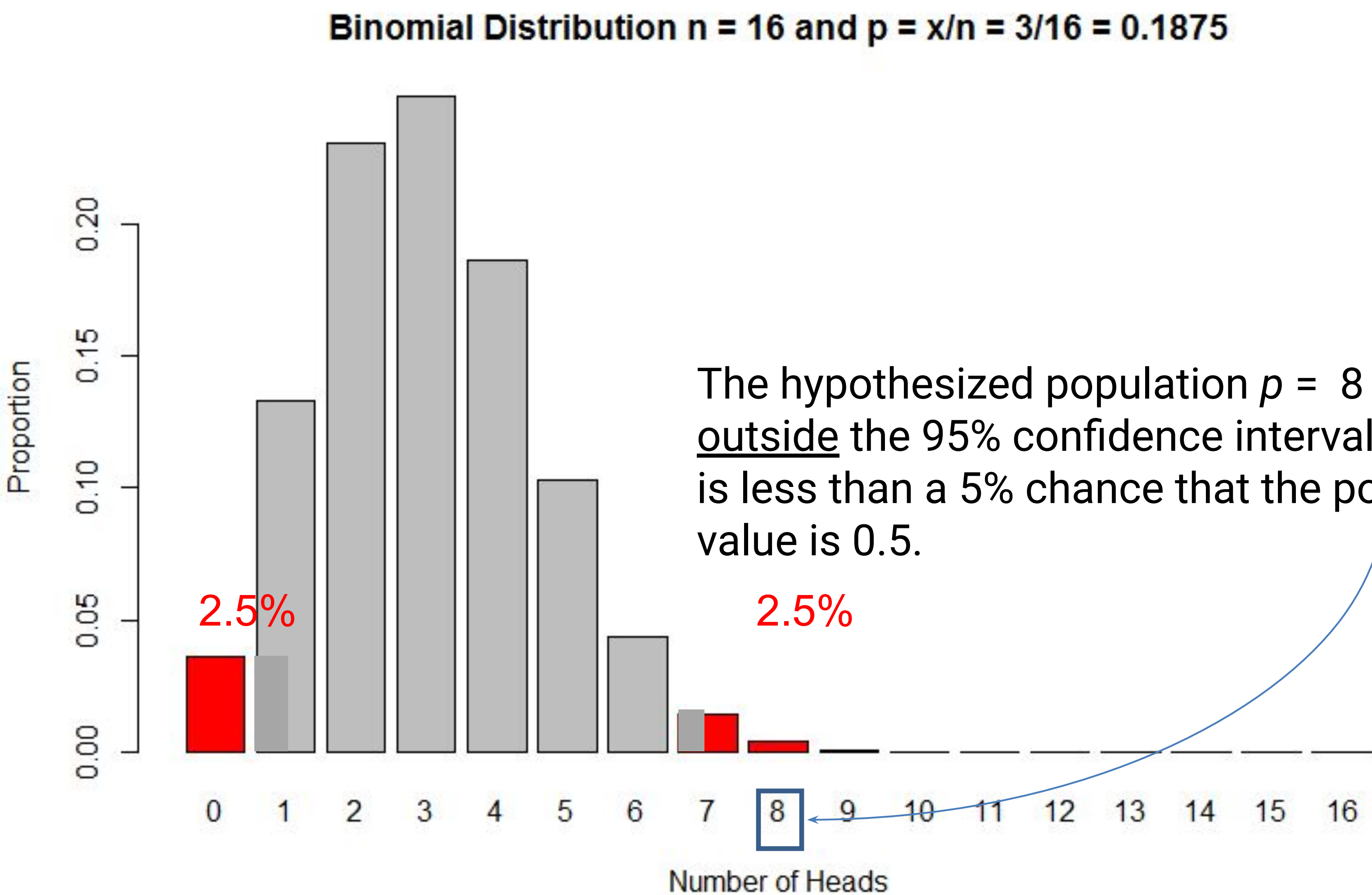
95 percent confidence interval:

0.04047373 0.45645655

sample estimates:

probability of success

0.1875



The hypothesized population *p* = 8 / 16) is <u>outside</u> the 95% confidence interval, so there is less than a 5% chance that the population value is 0.5.

Testing a null hypothesis (H_0)

- Often H_0 is the hypothesis of “no difference”, but in principle, it can be any hypothesis.
- We reject H_0 if the model it specifies does not fit our data well, i.e., if the mismatch between the data and the model is unlikely to be due simply to chance (sampling error).

Testing a null hypothesis (H_0)

- Alternatively, we can construct a confidence interval based on the sample, and if the H_0 value does not fall within that interval, we reject H_0.
- If H_0 is not rejected on the basis of the available evidence, we **have not** proven that it's true; it *may* be that the difference between the model and sample are washed out by sampling error.

Statistical significance

- Statistical significance means that the probability that the H_0 model assigns to the data is less than a small, arbitrarily-chosen α (alpha), often .05 or .01 by convention.
- Statistical significance is *not* a "gradable" quality.

Some applications of the binomial test (1/)

- 1 group, 1 dichotomous measure per participant; e.g., do Columbus residents have a preference for [str] vs. [ʃtr]?
- 1 group, 1 interval measure per participant (sign test): Take the difference between each measure and a hypothesized population value and use the signs of the differences; e.g., do male speakers of a dialect have f_0 that differs from that of the standard dialect?

Some applications of the binomial test (2/)

- 1 group, 1 interval measure per participant (sign test): Take the difference between each measure and a hypothesized population value and use the signs of the differences; e.g., do male speakers of a dialect have f_0 that differs from that of the standard dialect?

Some applications of the binomial test (3/)

- 1 group, 2 interval measures per participant (sign test); take the difference between each pair of measures and use the signs of the differences; e.g., do infants' looking times differ for mother's vs. other's voice?

Null hypothesis testing errors

- Type I: Rejecting the null hypothesis when it should not be rejected (i.e., there really is no difference)
  - probability of Type I error (α):
    - the critical value of the test statistic
    - smaller α -> wider CI -> fewer rejections of H_0
- Type II: Not rejecting the null hypothesis when it should be rejected (i.e., there really is a difference)
  - probability of Type II error (β):
    - Reduced by increasing *n*
    - Reduced by increasing *effect size*

Effect size

Effect size is simply a measure of the magnitude of the observed difference.

- For the binomial, one way to measure effect size is by means of the odds of one outcome
  - If we toss a coin 4 times and the results are 3H1T, then the odds of heads are 3 : 1 = 3 / 1.
- NB: effect size does not reflect sample size.
  - If we toss a coin 40 times and the results are 30H10T, then the odds of heads are the same.

Even small effect sizes can be significant with large samples

```
> binom.test(51,
  100, .5)$p.value
[1] 0.9204108
```

```
> binom.test(5100,
  10000,.5)$p.value
[1] 0.04658553
```

The problem with null hypothesis significance testing

Every non-directional null hypothesis is false.

- Zero effects are practically impossible when scores are continuously valued and measured with sufficient precision.
- Given enough data, even a miniscule effect size could be shown to be statistically significance.

The solution

Therefore, we report:

- effect sizes
- confidence intervals

in addition to significance and the associated *p*-value.

Alternatively, we can turn to Bayesian approaches...

Binomial test assumptions

- Samples can be sorted into two exhaustive, mutually exclusive categories.
- Every sample is independent of every other event.
- The population parameter *p* is fixed/cannot change during sampling.

In summary: the data consists of *dichotomous random variables which are independently and identically distributed (i.i.d)*.