S&DS 101
Intro Statistics: Life Sciences

Overview

Home questions

Quick review of the 5 steps for hypothesis testing

Using parametric null distributions for hypothesis testing

- The binomial distribution
- The normal distribution

Any questions about homework 1?

Notes:

Knit the document and then answer questions from your knitted code

- This will prevent answers that involve randomization from changing

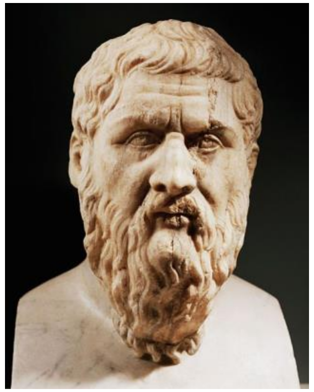
Do not print lots of output

Show your work: do all your calculations in R (not by hand outside R)

== vs =

Review of 5 steps for hypothesis testing

Five steps of hypothesis testing

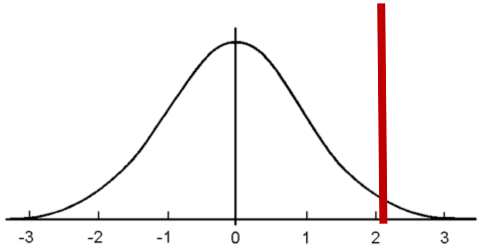


1. State H_0 and H_A
   - Assume Gorgias (H_0) was right

2. Calculate the actual observed statistic

3. Create a distribution of what statistics would look like if Gorgias is right
   - Create the **null distribution** (that is consistent with H_0)

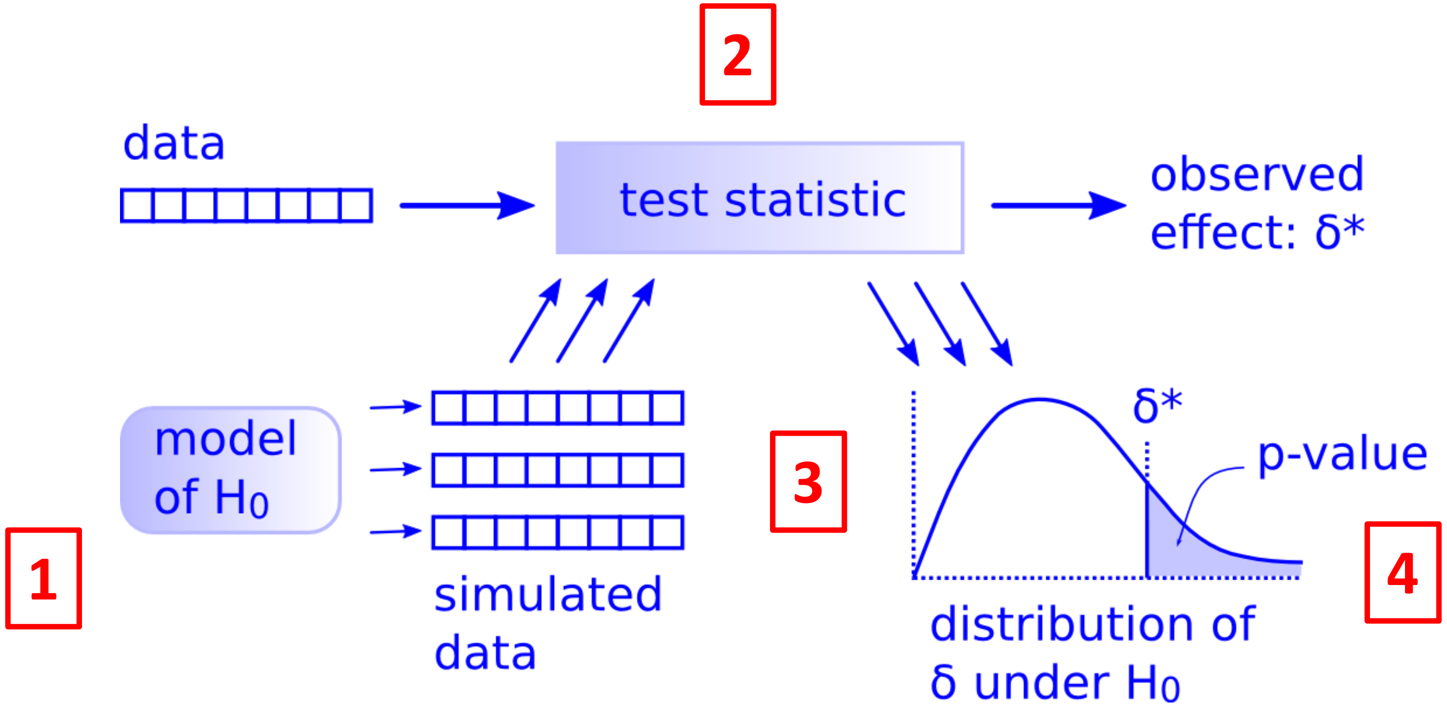
4. Get the probability we would get a statistic more than the observed statistic from the null distribution
   - p-value



5. Make a judgement
   - Assess whether the results are statistically significant

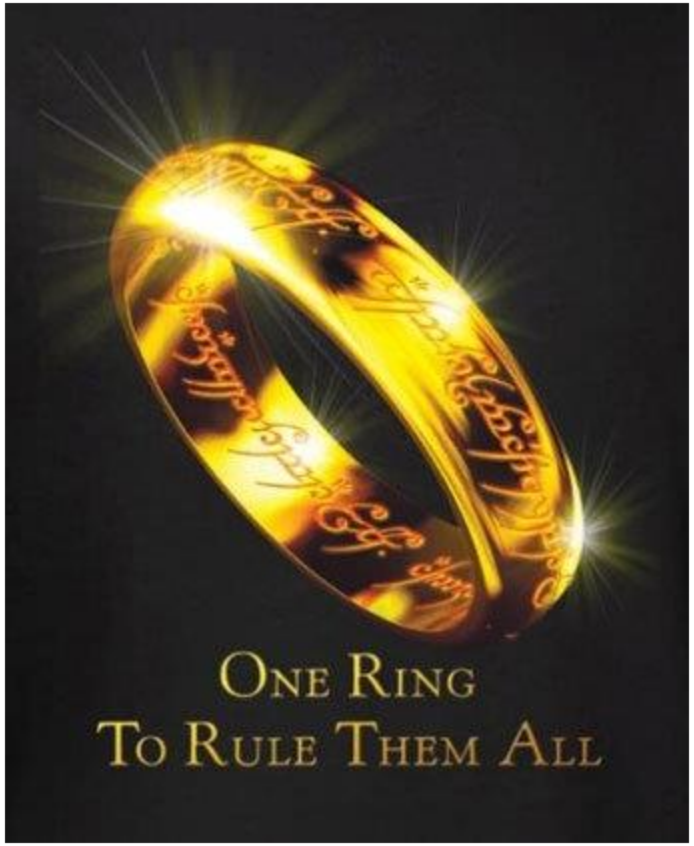
One test to rule them all

There is only one hypothesis test!



Just follow the 5 hypothesis tests steps!

5!



Hypothesis testing in 5 easy steps!

1. State the null hypothesis… and the alternative hypothesis
   - Buzz is just guessing so the results are due to chance: H_0: $\pi = 0.5$
   - Buzz is getting more correct results than expected by chance: H_A: $\pi > 0.5$

2. Calculate the observed statistic
   - Buzz got 15 out of 16 guesses correct, or $\hat{p}$ = .973

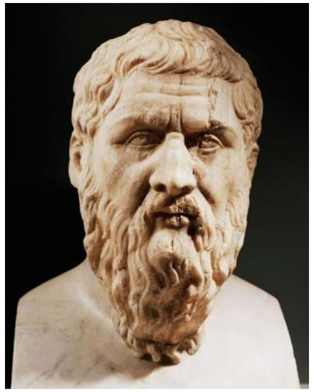
3. Create a null distribution that is consistent with the null hypothesis
   - i.e., what statistics would we expect if Buzz was just guessing

4. Examine how likely the observed statistic is to come from the null distribution
   - What is the probability that the dolphins would guess 15 or more correct?
   - i.e., what is the p-value

5. Make a judgement
   - If we have a small p-value, this means that $\pi = .5$ is unlikely and so $\pi > .5$
   - i.e., we say our results are 'statistically significant'

Five steps of hypothesis testing

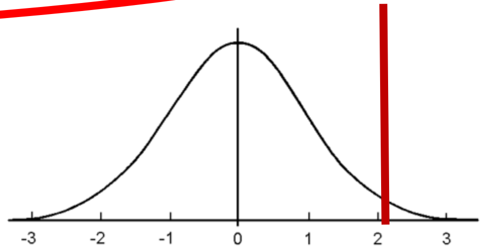
1. State H_0 and H_A
   - Assume Gorgias (H_0) was right



2. Calculate the actual observed statistic

3. Create a distribution of what statistics would look like if Gorgias is right
   - Create the **null distribution** (that is consistent with H_0)

4. Get the probability we would get a statistic more than the observed statistic from the null distribution
   - p-value

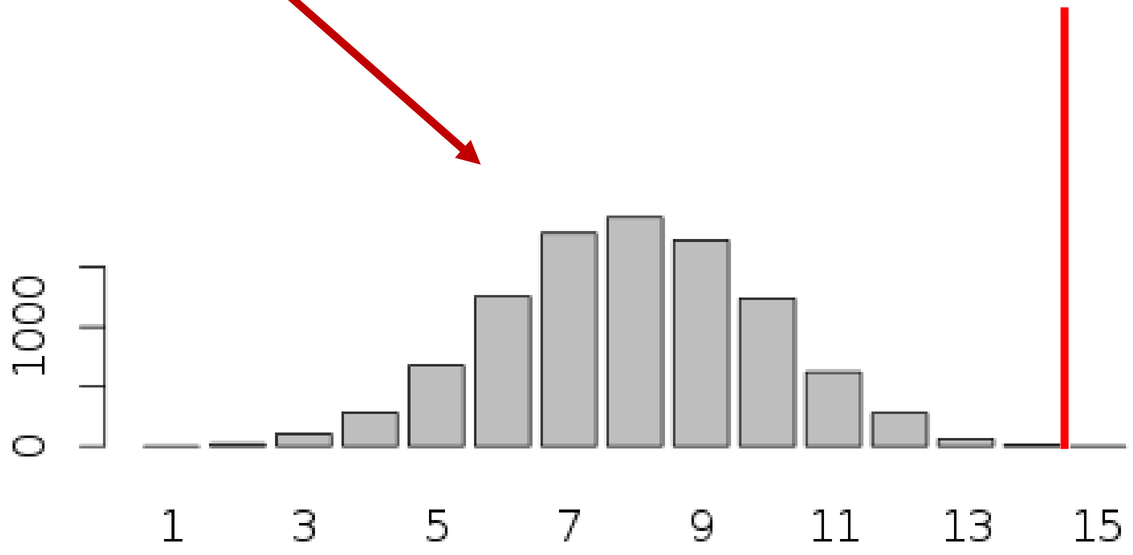


5. Make a judgement
   - Assess whether the results are statistically significant

Step 3: Buzz and Doris randomization null distribution

0	0
1	1
2	22
3	105
4	283
5	679
6	1257
7	1786
8	1920
9	1726
10	1238
11	623
12	279
13	63
14	15
15	3
16	0

Null distribution



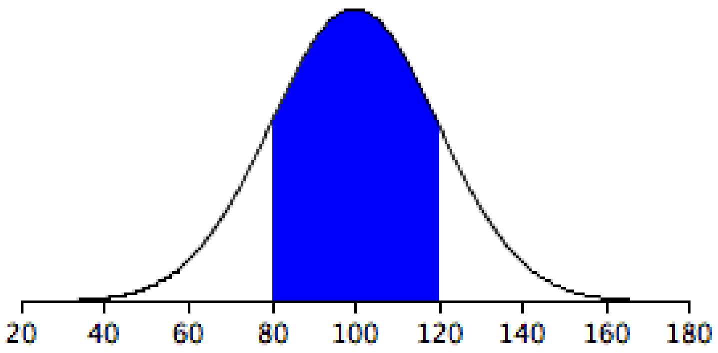
null_distribution <- rbinom(10000, 16, .5)

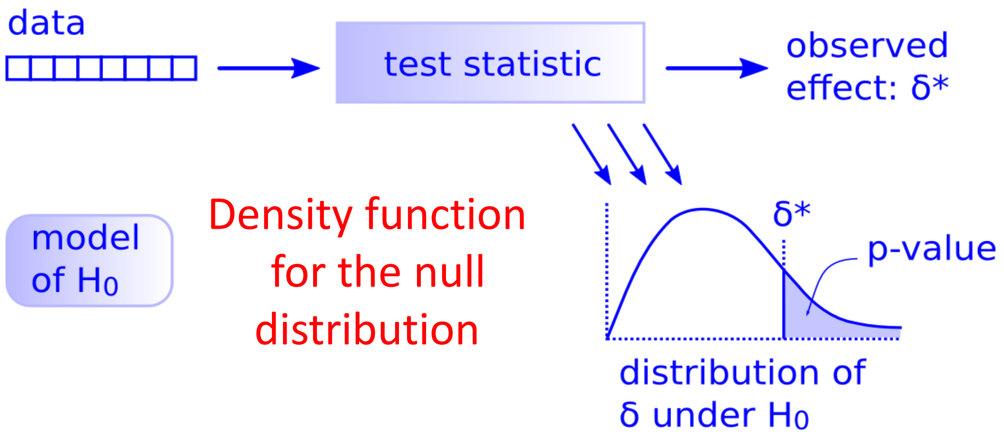
p_value = 3/10000 = .0003

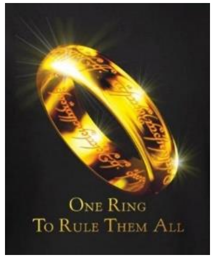
Parametric hypothesis tests

Parametric hypothesis tests are hypothesis test that use **density functions** for the **null distribution**

- i.e., by making some additional assumptions, we can know the mathematical form of the null distribution





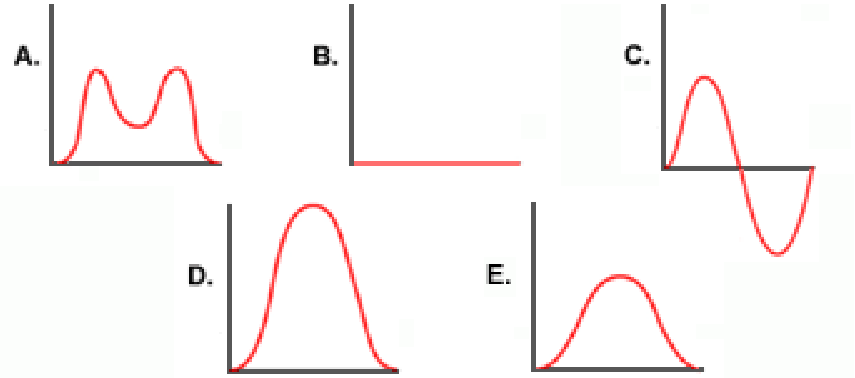


We need to make sure to check our assumptions to make sure these parametric tests are valid in practice

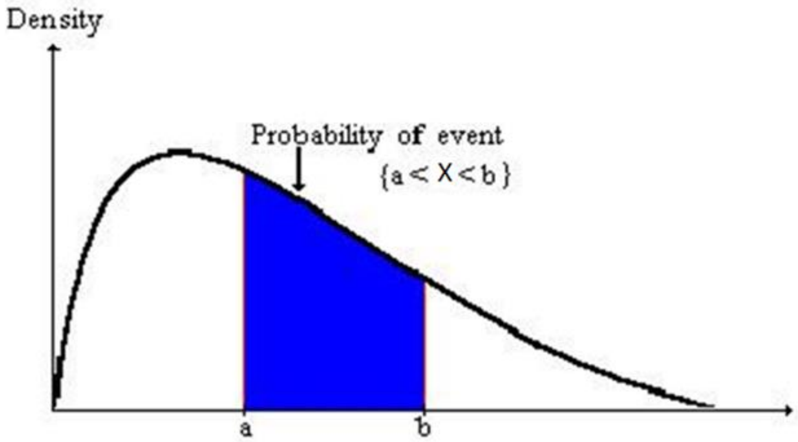
Review: density functions

A **density curve** is a mathematical function f(x) that has two properties:

1. The total area under the curve f(x) is equal to 1
2. The curve is always ≥ 0



Which of these could ***not*** be a density curve?



The area under the curve in an interval [a, b] models the probability that a random number X will be in the interval

- Pr(a < X < b) is the area under the curve from a to b

Review: Binomial distribution

Probability of getting ***k*** successes out of ***n*** trials is:

Parameters: π and n

$$Pr(X = k)$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

What values can X take on (i.e., what is the sample space)?

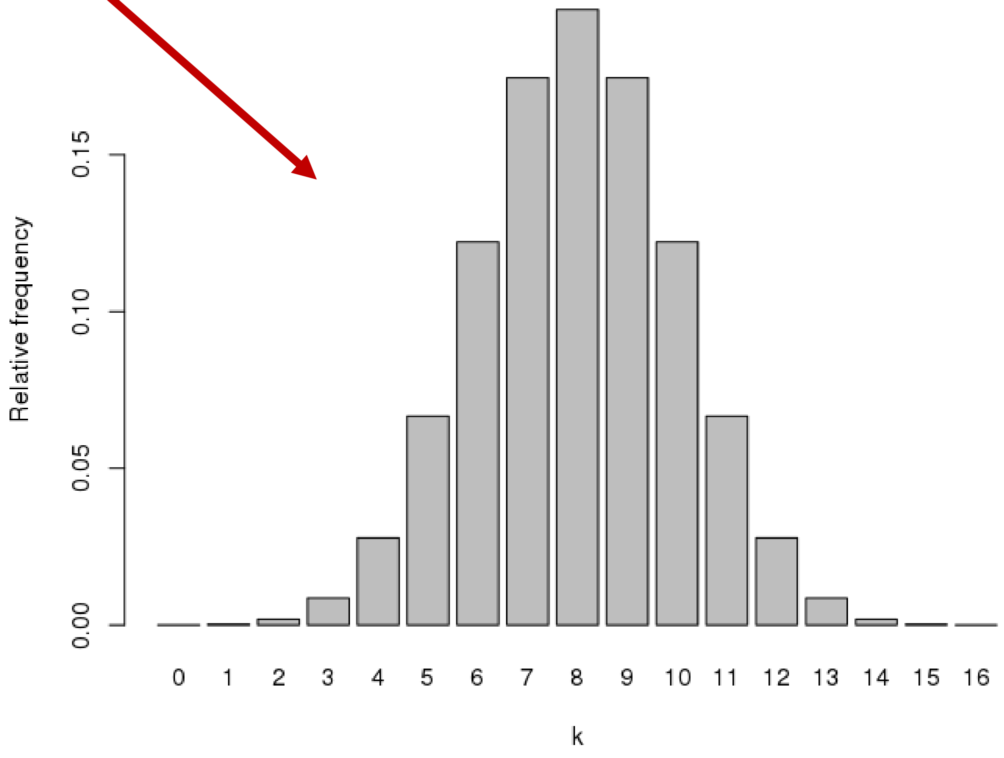
- k ranges from 0, 1, …, n

Step 3: Buzz and Doris binomial null distribution

Null distribution

$$Pr(X = k) =$$

$$\binom{n}{k}\pi^k(1-\pi)^{n-k}$$



For displaying the density function

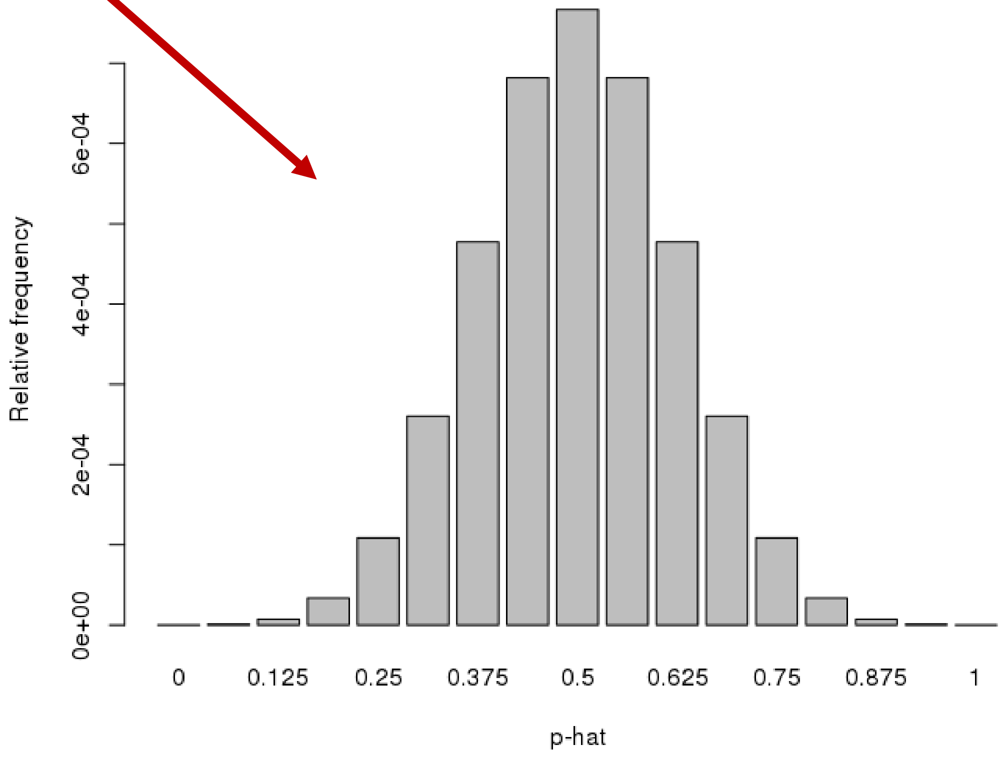
```
null_distribution  <-  dbinom(0:16, 16, .5)
```

Step 3: Buzz and Doris binomial null distribution

$$Pr(X = k) =$$

$$\binom{n}{k}\pi^k(1-\pi)^{n-k}$$

Null distribution in terms of p̂



For displaying the density function

null_distribution <- dbinom(0:16, 16, .5)/16

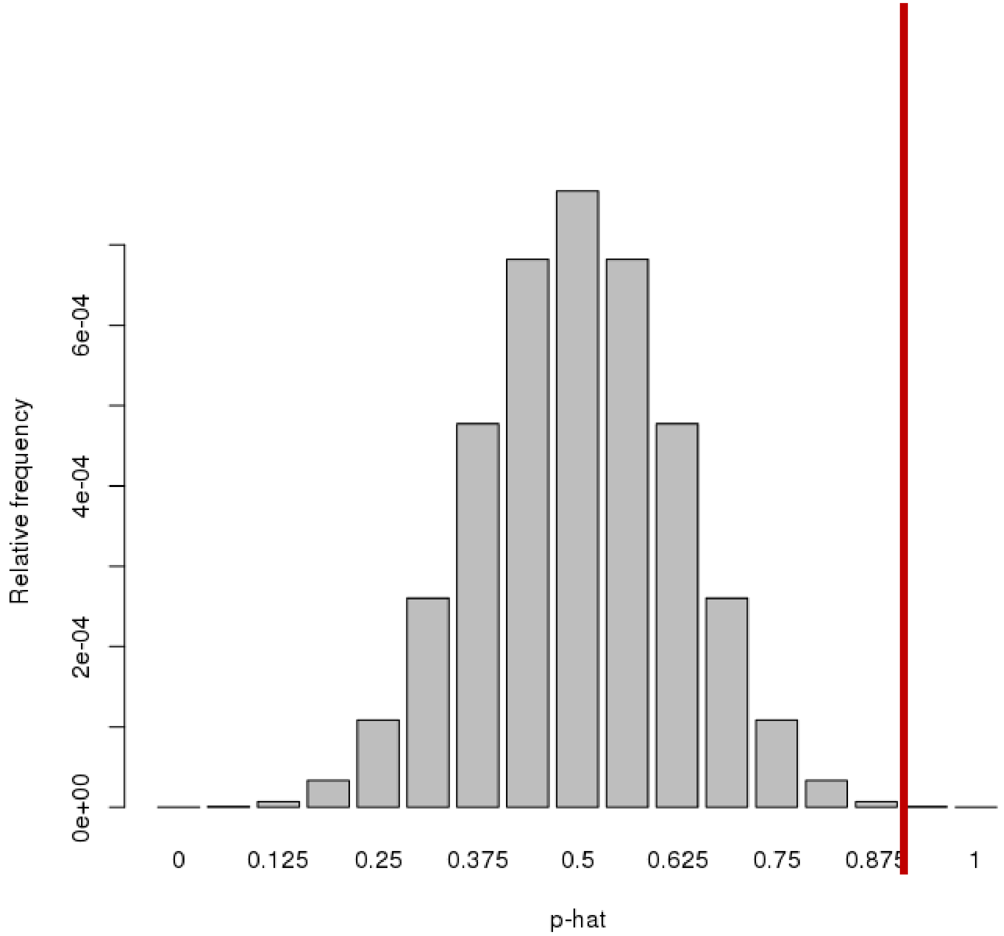
Step 3: Buzz and Doris binomial null distribution

The observed statistic was: $\hat{p} = 15/16$

Conceptually, how can we get a p-value?

$\Pr(X \geq \text{obs\_stat} \mid H_0)$

$\Pr(X \geq \text{obs\_stat} \mid n = 16, \pi = 0.5)$



p_value <- pbinom(14, 16, .5, lower.tail = FALSE)

Review: Binomial functions in R

1. rbinom(): generate random numbers from a binomial distribution
2. dbinom(): create the binomial density function $\Pr(X = k; n, \pi)$
3. pbinom(): create the cumulative distribution function $\Pr(X \le k; n, \pi)$

pbinom() is the cumulative sum of dbinom()

Let's try it in R…

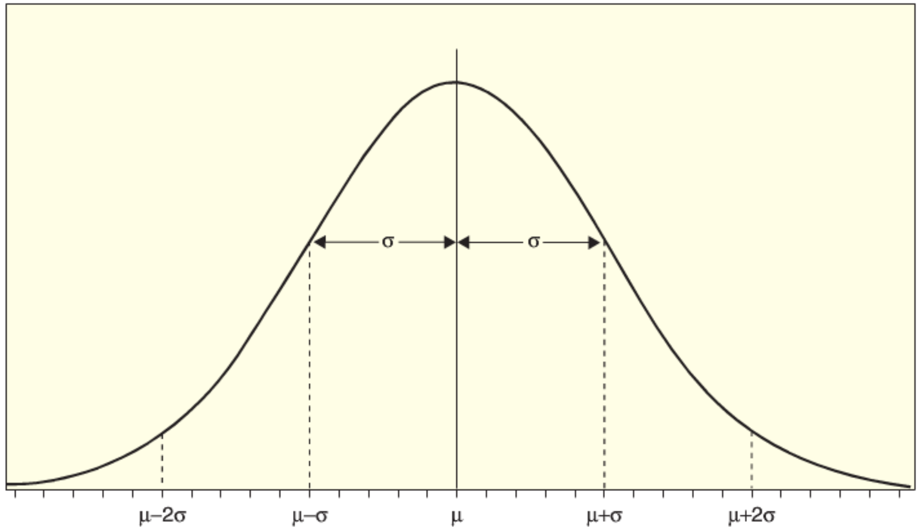
Normal density as an approximation to the binomial density function

We can also use a normal density function as a null distribution which is an approximation to the binomial distribution

- All steps in the hypothesis test are the same except we us a normal distribution in step 3
- The normal approximation is reason able if $n \cdot \pi \geq 10$ and $n \cdot (1 - \pi) \geq 10$

Recall the normal density has the form

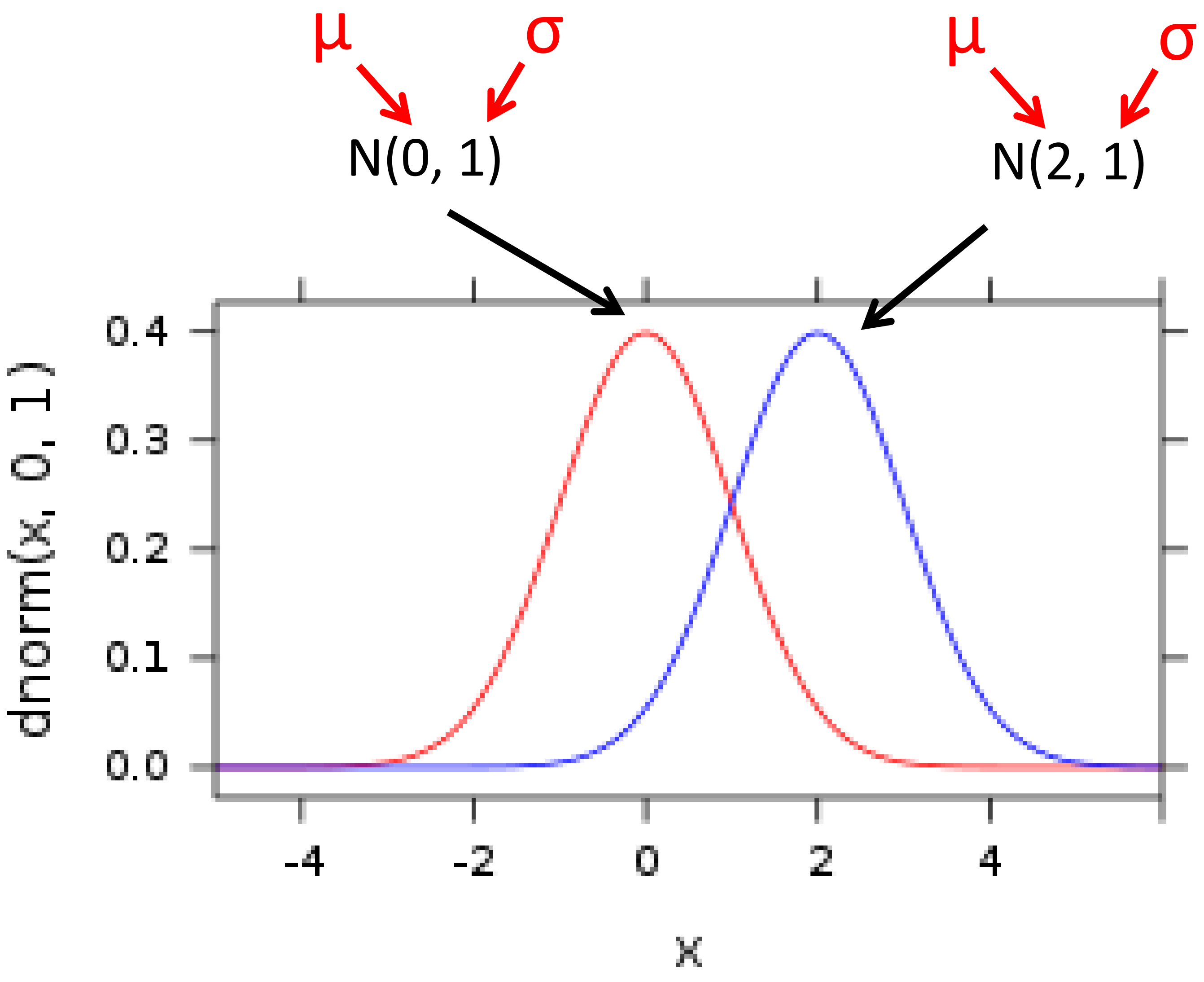
$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



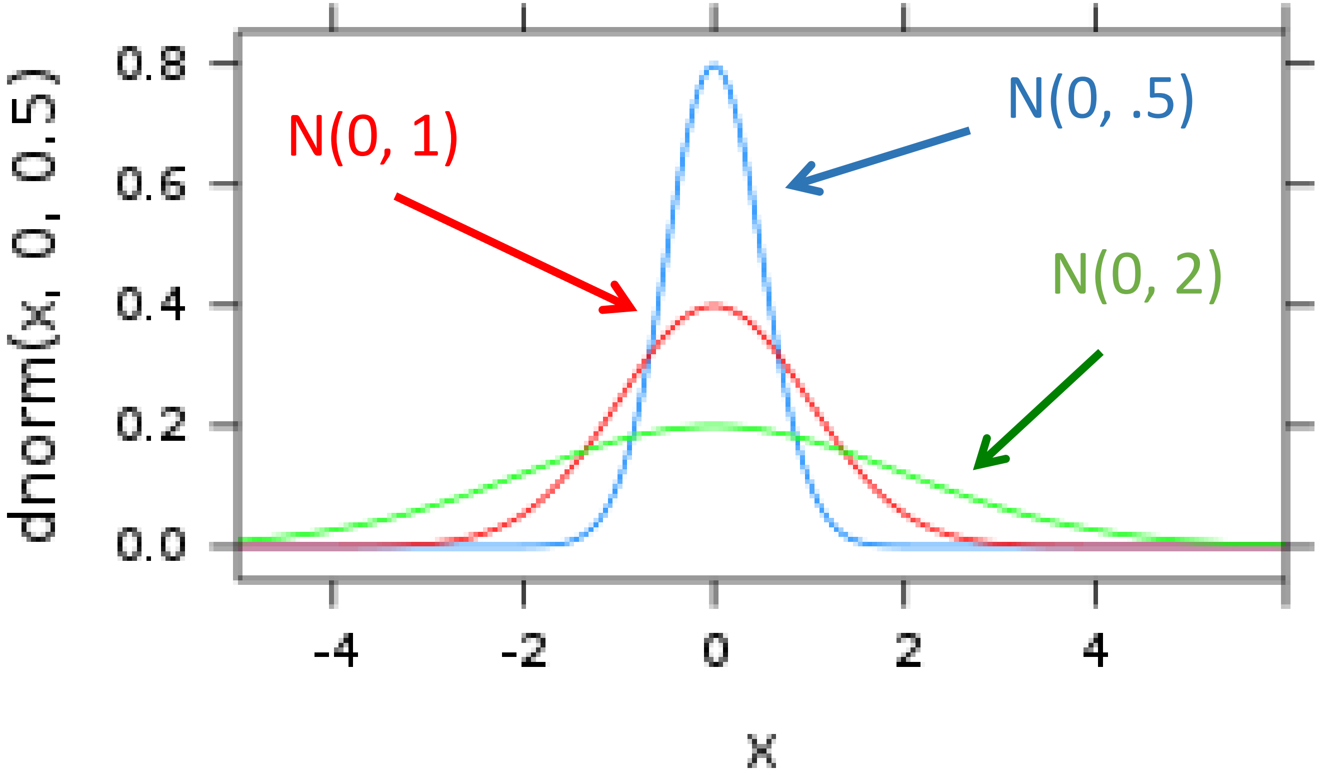
Important part: normal distributions have two parameters: μ and σ

$$N(X; \mu, \sigma)$$

Normal curves with different means



Normal curves with different variances



Normal density as a null distribution

Recall our hypothesis test for a proportion is:

- H_0: $\pi = \pi_0$ ($\pi_0 = 0.5$ for the Doris and Buzz example)
- H_A: $\pi \neq \pi_0$

When using the normal density function as a null distribution for a proportion, we have the following values for the parameters

The mean of the normal distribution is given by $\mu = \pi_0$

The standard error is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

Normal distribution functions in R

1. rnorm(): generate random numbers from a normal distribution

2. dnorm(): create the normal density function $f(x; \mu, \sigma)$

3. pnorm(): create the cumulative distribution function $Pr(X \leq x; \mu, \sigma)$

pnorm() is the integral of dnorm()

Let's try it in R…