

Using the normal distribution
for inference

Overview

Very quick review of hypothesis tests for more than 2 means and for correlation

Theories of hypothesis testing

Parametric probability distributions

The normal distribution

- Generating random data
- Visualizing the normal distribution
- Calculating probabilities from a normal distribution
- Calculating quantiles from a normal distribution

Very quick review: Hypothesis test for multiple means

Step 1:

$$H_0: \mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$$

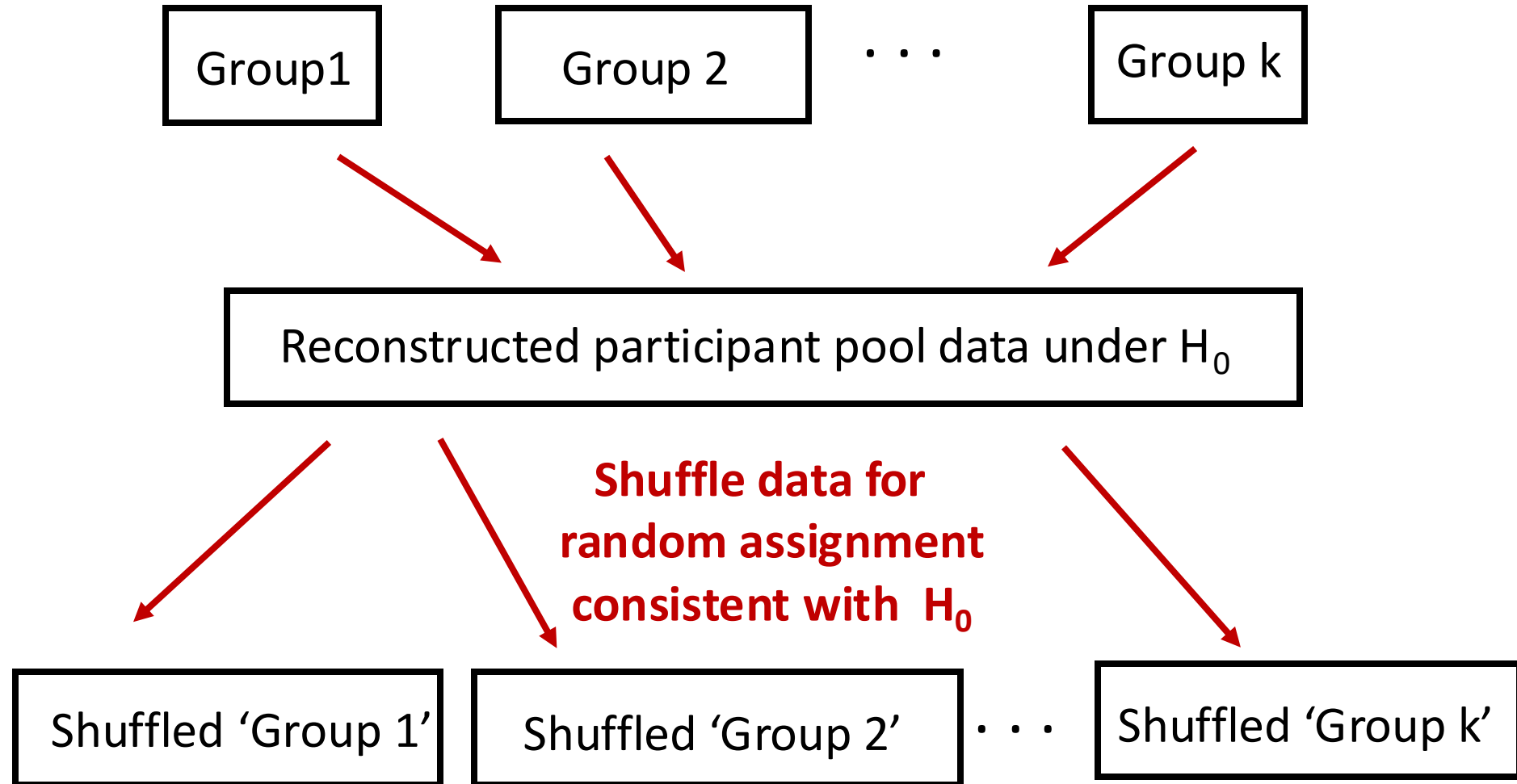
$$H_A: \mu_i \neq \mu_j \text{ for one pair of fields of study}$$

Step 2:

`boxplot(quant_data ~ cat_data)`

$$MAD = (|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$$

3. Create the null distribution!



4. p-value



Compute MAD statistic from shuffled groups

Very quick review: Hypothesis test for correlation

1. State the null and alternative hypothesis:

$$H_0: \rho = 0$$

Possible alternative hypotheses:

- a. $H_A: \rho < 0$
- b. $H_A: \rho > 0$
- c. $H_A: \rho \neq 0$

2. `cor(x, y)` and `plot(x, y)`

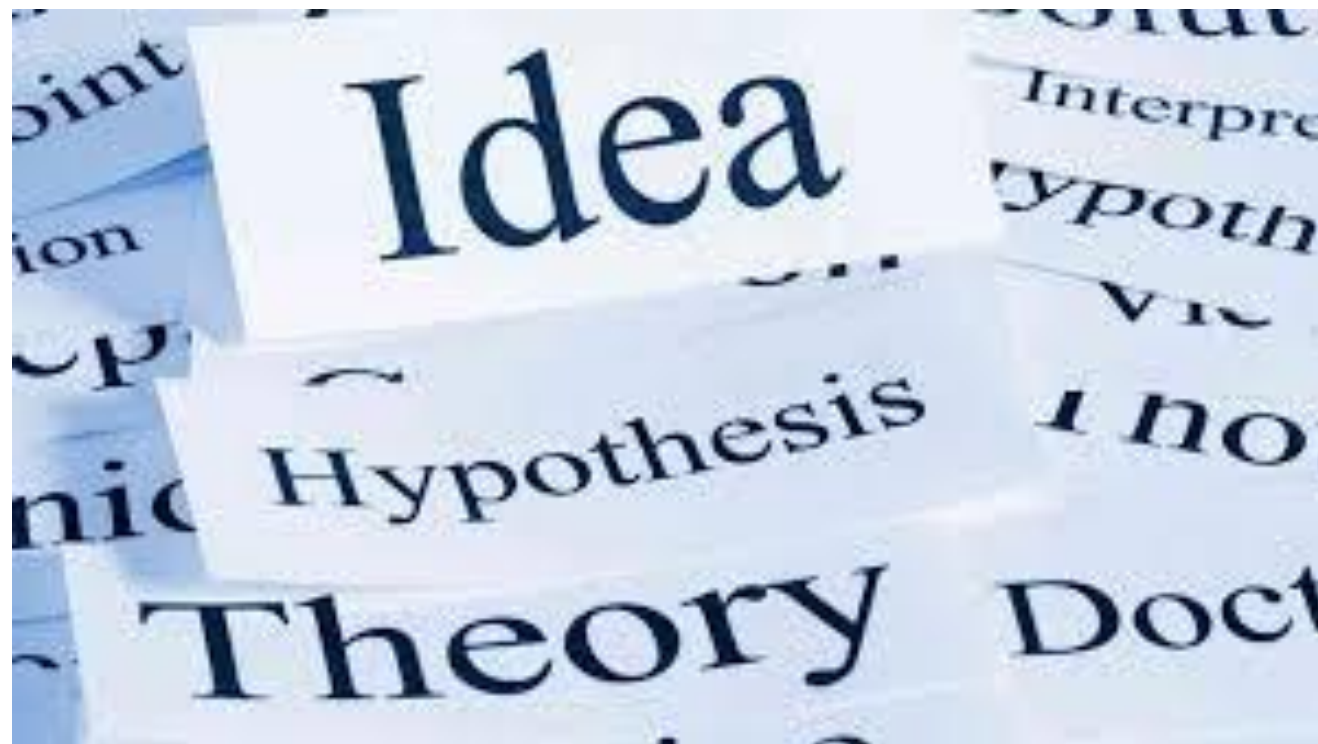
3. `cor(x, shuffle(y))`

4. p-value

5.



Theories of hypothesis tests



Two theories of hypothesis testing

Null-hypothesis significance testing (NHST) is a hybrid of two theories:

1. Significance testing of Ronald Fisher
2. Hypothesis testing of Jezy Neyman and Egon Pearson



Fisher (1890-1962)



Neyman (1894-1981)

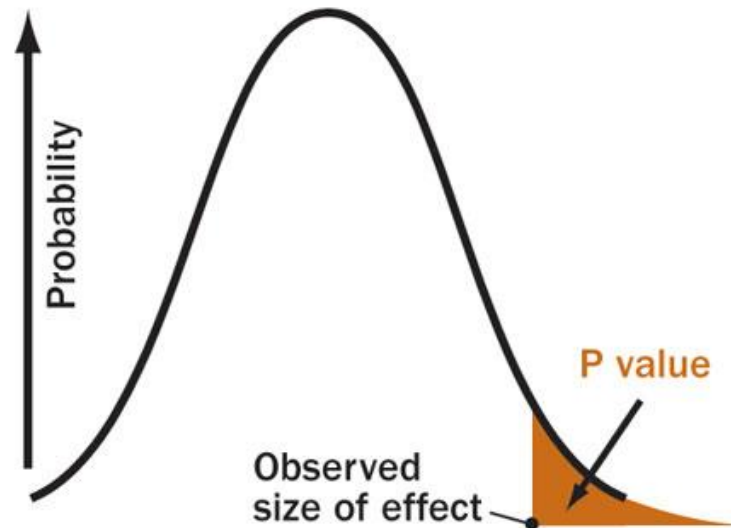


Pearson (1895-1980)

Ronald Fisher's significance testing

Views the p-value as strength of evidence against the null hypothesis

- P-values part of an on-going scientific process: tells the experimenter “what results to ignore”



Neyman-Pearson null hypothesis testing

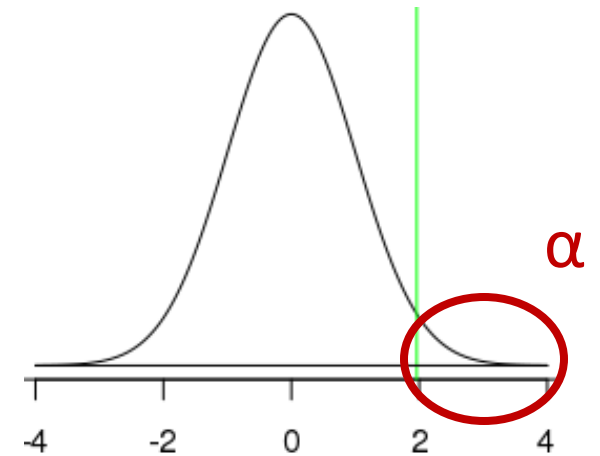
Makes ***a formal decision*** in statistical tests

Reject H_0 : if the observed sample statistic is beyond a **fixed value**

- i.e., reject H_0 if the p-value is less than some predetermined **significance level α**

Do not reject H_0 : if the observed sample statistic is not beyond a **fixed value**. This means the test is inconclusive

Null distribution

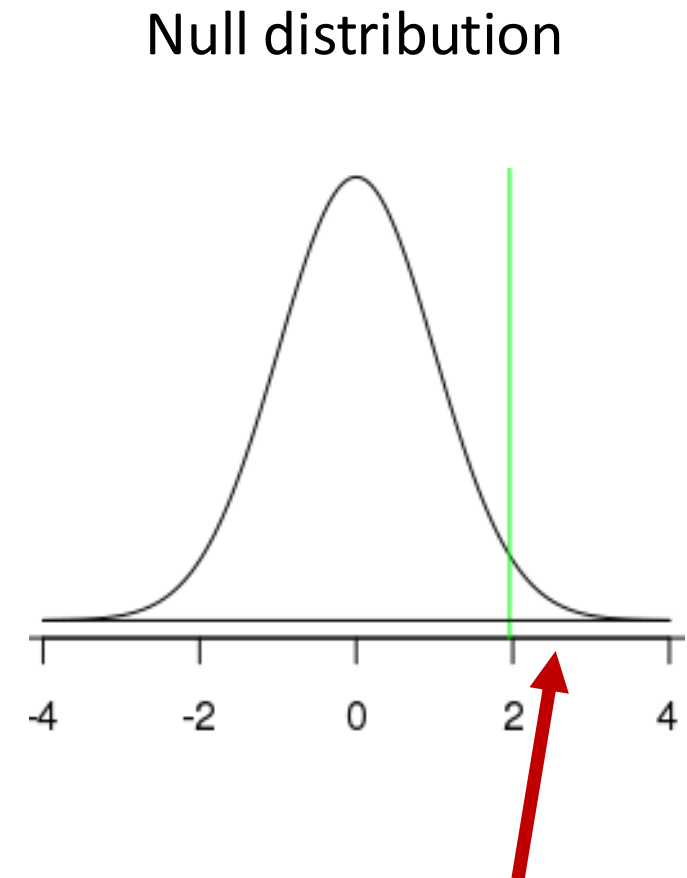


Neyman-Pearson frequentist logic

Type I error: incorrectly rejecting the null hypothesis when it is true

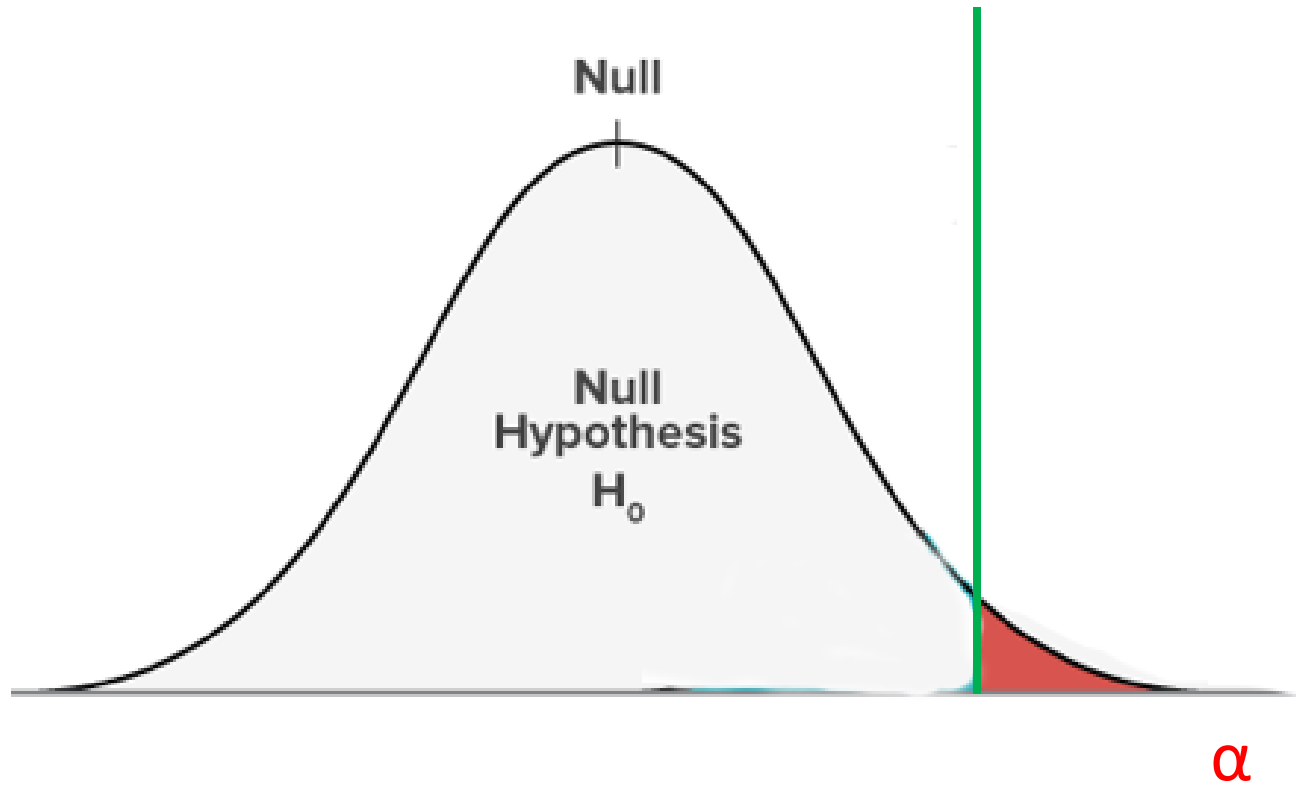
When the null hypothesis is true, only ~5% of the time will we falsely reject it (for $\alpha = 0.05$)

- i.e., we only make type I errors 5% of the time

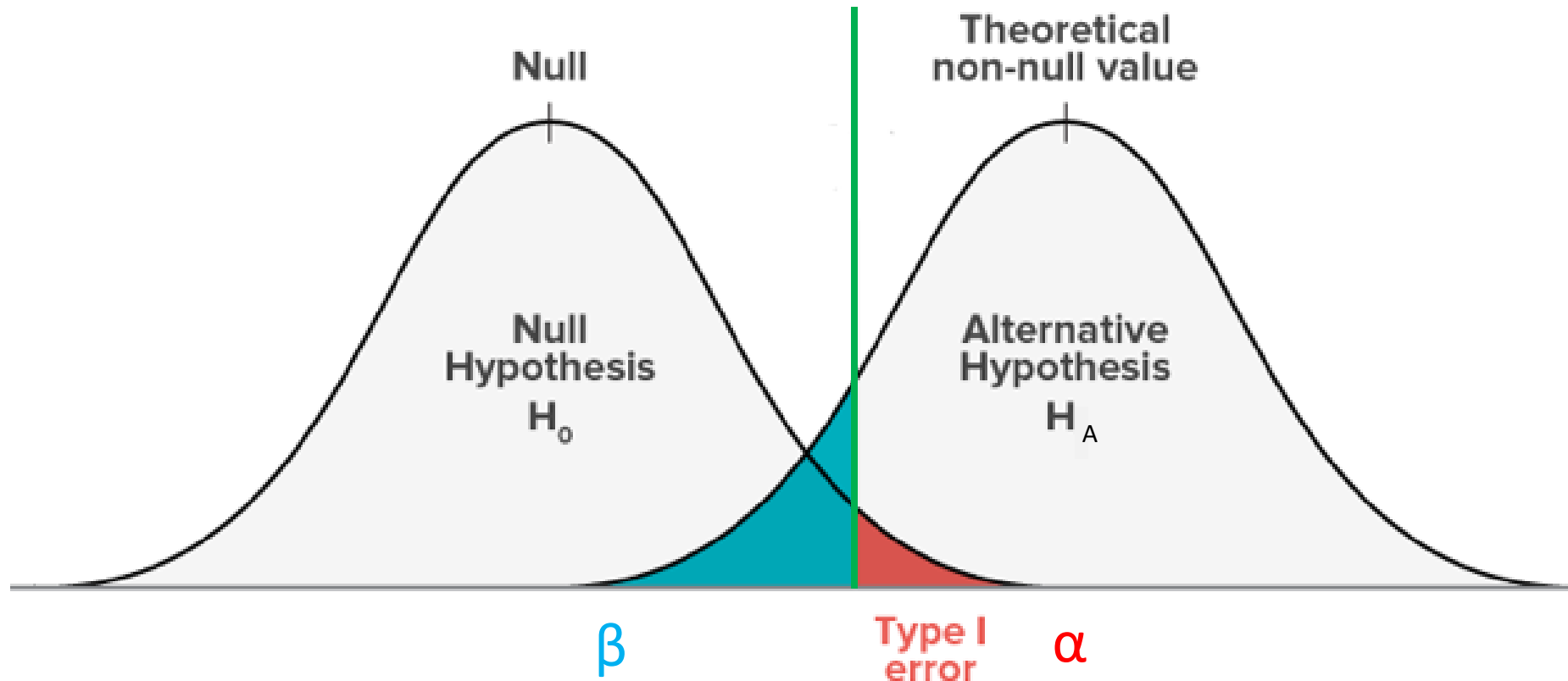


The null distribution is true but statistic landed here

Neyman-Pearson Frequentist logic



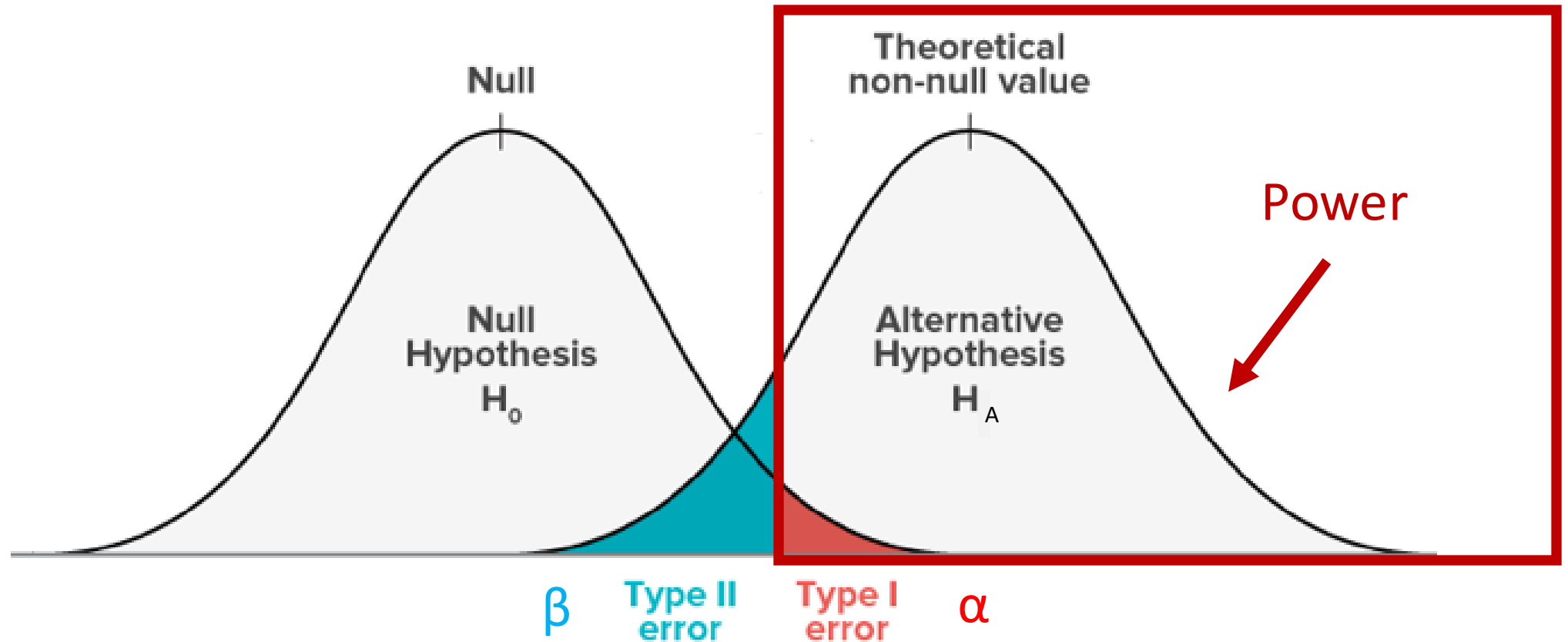
Neyman-Pearson Frequentist logic



Type II error: incorrectly failing to reject H_0 when it is false

- The rate at which we make type II errors is often denoted with the symbol β

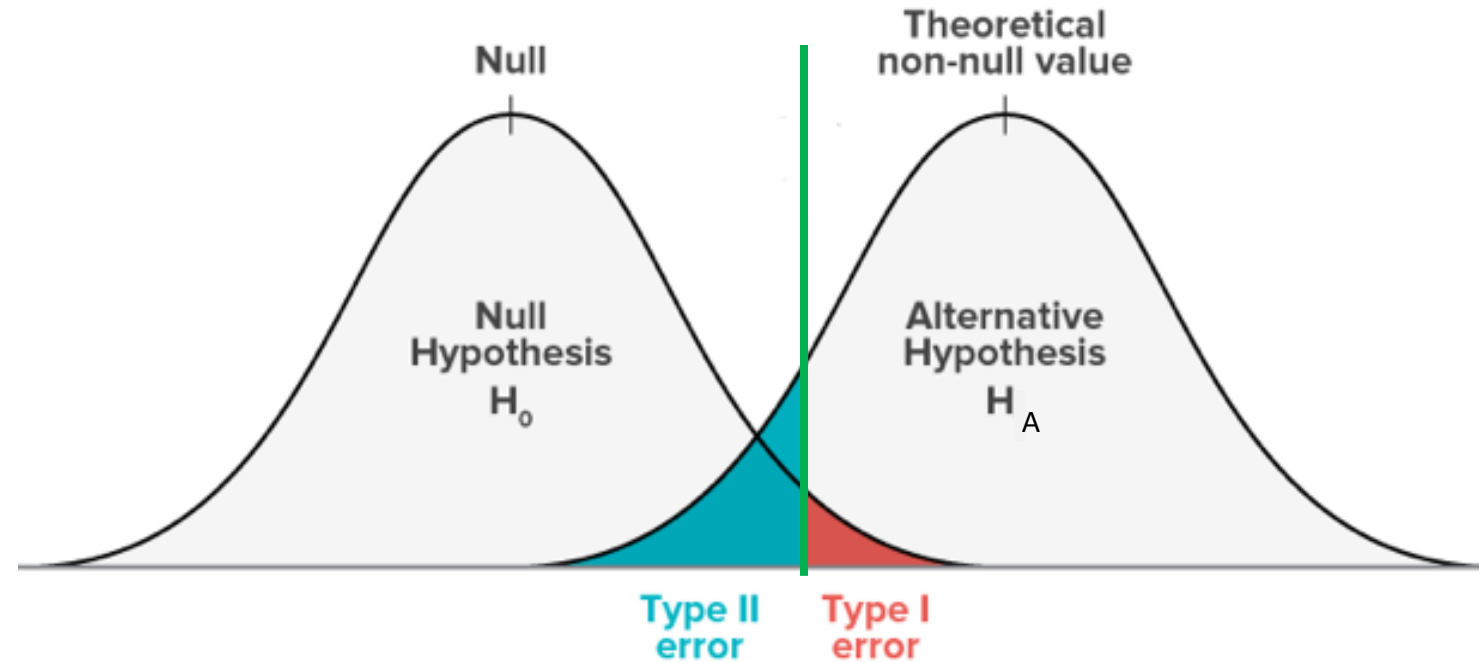
Neyman-Pearson Frequentist logic



The **power** of a test is the probability we reject the H_0 when it is **false**

- $1 - \beta$
- For a fixed α level, it would be best to use the most powerful test

Type I and Type II Errors



	H_0 is true	H_A is true (H_0 is false)
Reject H_0	Type I error (α) (false positive)	No error
Do not reject H_0	No error	Type II error (β) (false negative)

Problems with the NP hypothesis tests

Problem 1: we are interested in the results of a specific experiment, not whether we are right most of the time

- E.g., 95% of these statements are true:
 - Calcium is good for your heart, Paul is psychic, Buzz and Doris can communicate, ...

Problem 2: Arbitrary thresholds for alpha levels

- P-value = 0.051, we don't reject H_0

Problems with the NP hypothesis tests

Problem 1: we are interested in the results of a specific experiment, not whether we are right most of the time

- E.g., 95% of these statements are true:
 - Calcium is good for your heart, Paul is psychic, Buzz and Doris can communicate, ...

Problem 2: Arbitrary thresholds for alpha levels

- P-value = 0.051, we don't reject H_0 ?

Problem 3: running many tests can give rise to a high number of type I errors

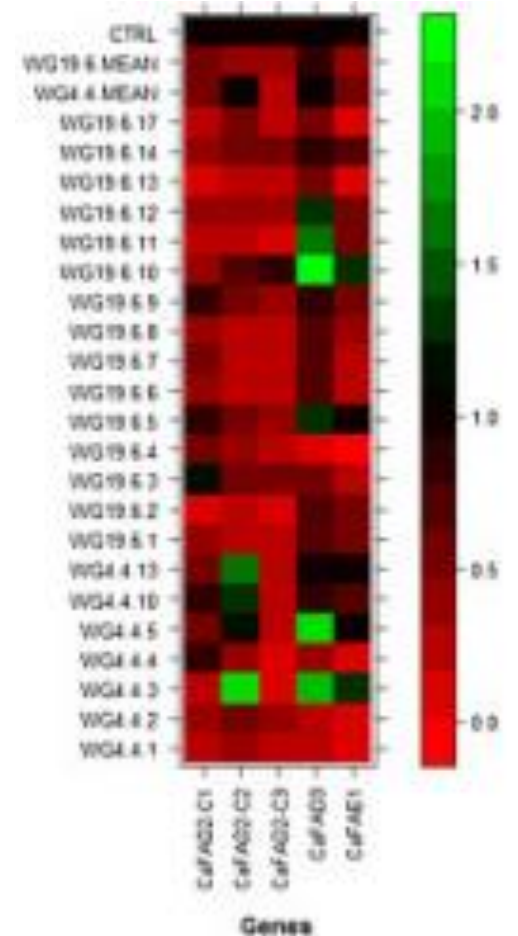
Genes and leukemia example

Scientists collected 7129 gene expression levels from 38 patients to find genetic differences between two types leukemia (L1 and L2)

Suppose there was no genetic differences between the types of leukemia

- $H_0: \mu_{L1} = \mu_{L2}$ is true for all genes

Q: If each gene was tested separately using a significance level of $\alpha = 0.05$, approximately how many type I errors would be expected?



Side note: correcting for multiple comparisons

There are methods that try to correct for running multiple hypothesis tests

The ***Bonferroni correction*** is one way that controls the probability of ***any*** hypothesis test giving a type I error

- i.e., controls the familywise error rate (no type I errors for any of the tests run)

It works by dividing the initial α level by the number of tests run

- E.g., $\alpha = 0.05/7129 = 0.000007$
- All p-values need to be below this level to be considered statistically significant
- This can lead to many type II errors (Type II error: failure to reject H_0 when it is false)

The problem of multiple testing

For $\alpha = 0.05$, ~5% of all published research findings should incorrectly reject the null hypothesis

Publication bias (file drawer effect): Generally positive results are more likely to be published, so if you read the literature, the proportion of incorrect results could be greater than 5%

Why Most Published Research Findings Are False

John P. A. Ioannidis

The Earth Is Round ($p < .05$)

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including

sure how to test H_0 , chi-square with Yates's (1951) correction or the Fisher exact test, and wonders whether he has enough power. Would you believe it? And would you believe that if he tried to publish this result without a

[American Statistical Association's Statement on p-values](#)

Some thoughts...

Better to have hypothesis tests than none at all. Just need to think carefully and use your judgment.

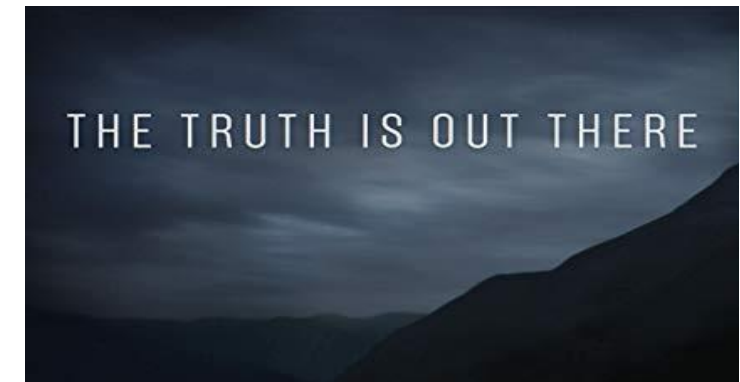
Report effect size in most cases – i.e., confidence intervals

Report the p-values rather than accept/reject H_0

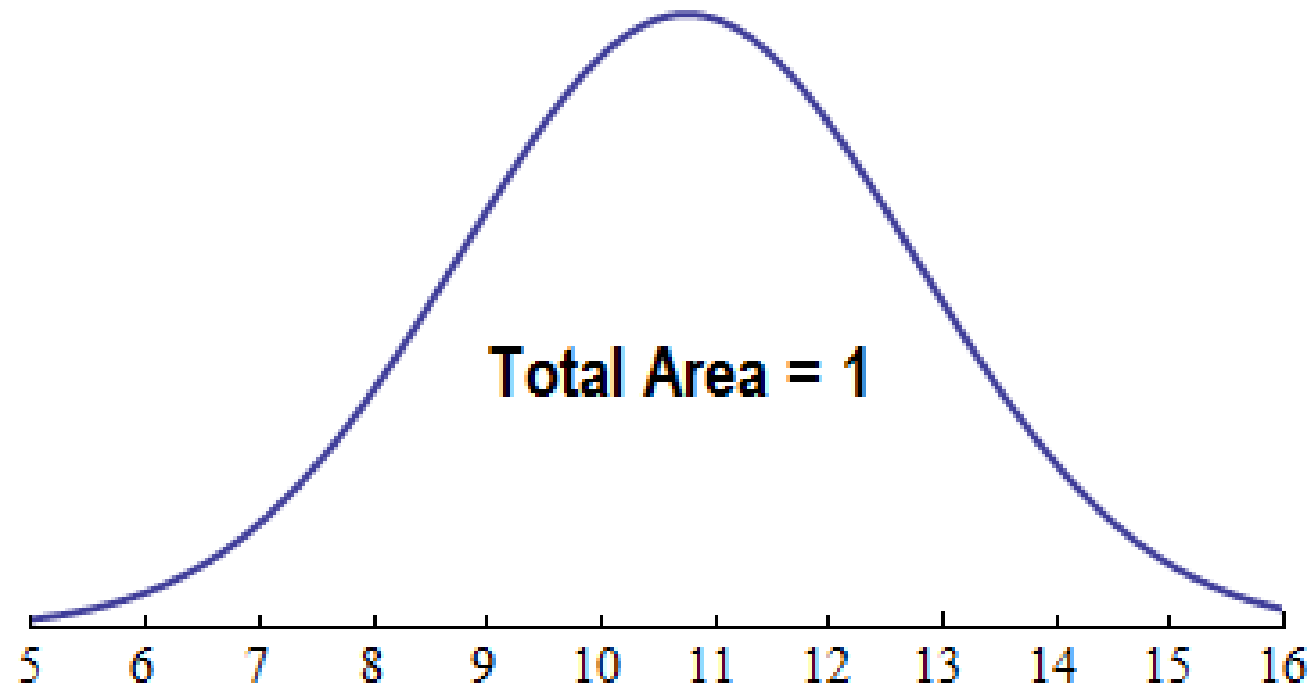
- i.e., report $p = 0.023$ not $p < 0.05$

Replicate findings (perhaps in different contexts) to make sure you get the same results

Be a good/honest scientists and try to get at the Truth!



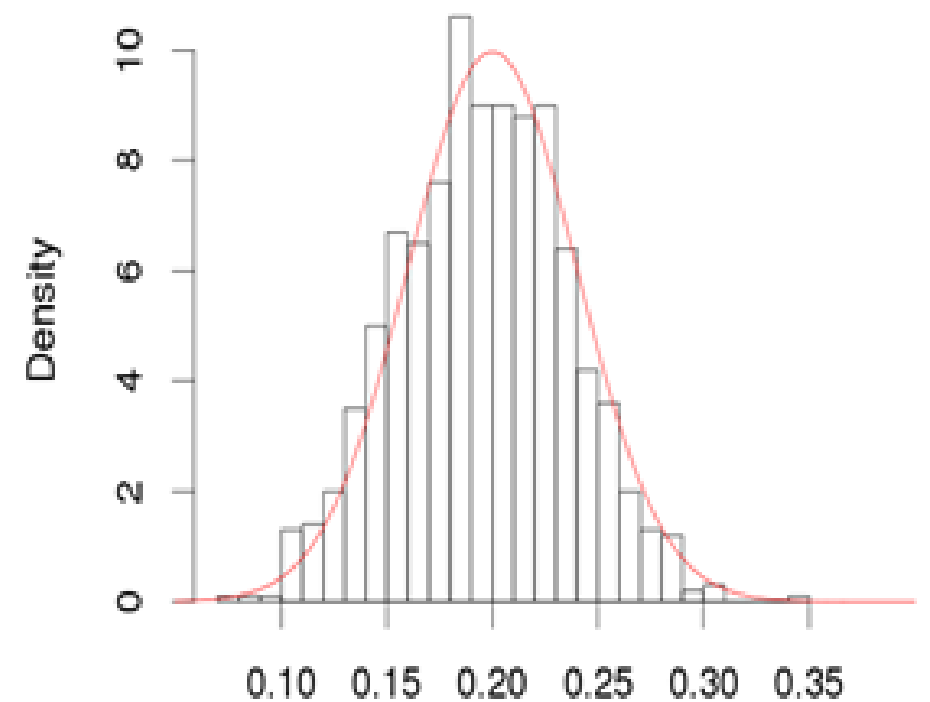
Inference using parametric probability distributions



Inference using parametric probability distributions

We can use mathematical functions called **probability distributions** to do statistical inference

For example, instead of running computer simulations to create null distributions, we can just mathematical probability distributions



Inference using parametric probability distributions

There are several functions that allow use to interact with probability distributions:

1. **Random number generation** functions (**r**)
2. **Density functions** to visual probability distributions (**d**)
3. **Cumulative probability distribution** functions to get the probability of events (**p**)
4. **Quantile functions** to get quantiles of a probability distributions (**q**)

Note: Many probability distributions have small number of ***parameters*** that control the shape of the probability distribution

- Using these probability models for statistical inference is called “parametric statistics”

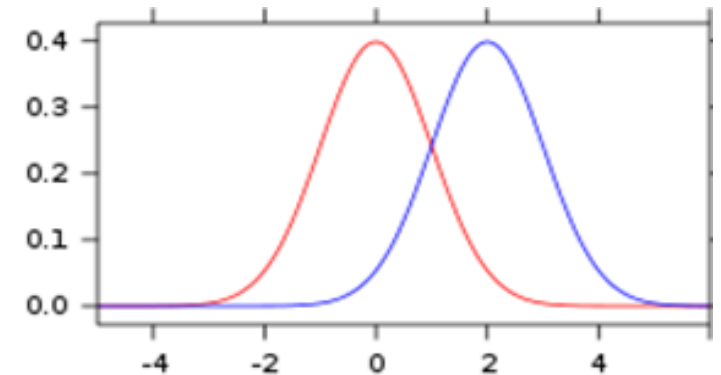
Main use case: Normal probability distributions

Normal distributions are a family of bell-shaped curves with two parameters

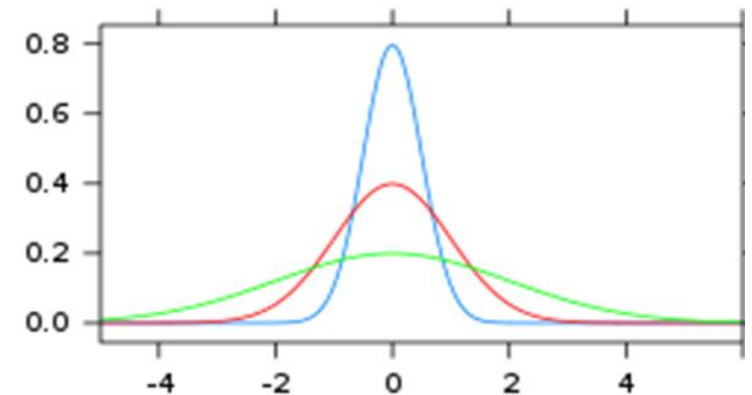
- The mean: μ
- The standard deviation: σ

Once μ and σ are set, we have a specific probability distribution that we use to get probabilities of events, generate random data, etc.

Changing μ



Changing σ



1. Generating random numbers (**r**norm)

Random numbers (i.e., random variables) that we have not yet observed are denoted with capital letters

- E.g., X, Z, etc.

We denote random numbers that come from a **normal distribution** using $X \sim N(\mu, \sigma)$

We can generate ***n*** random from a normal distribution using the **rnorm(n, mean, sd)**



Example: IQ scores

Intelligence quotient (IQ) is a score derived from a set of standardized tests to assess human intelligence

IQ scores are normally distributed and have a mean of 100 and a standard deviation of 15

Let's use the `rnorm()` to simulate some random IQ scores

- i.e., to generate values: $X \sim N(100, 15)$

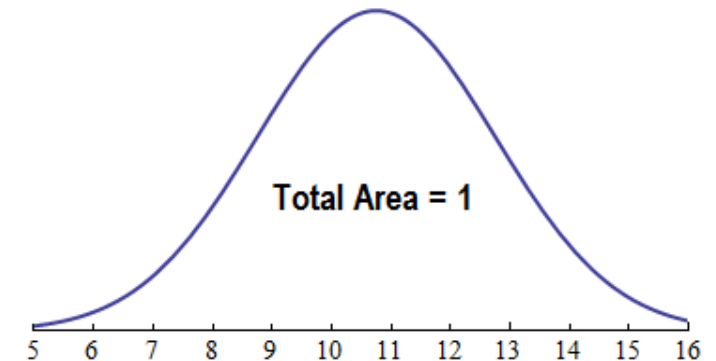
Let's try it in R!

2. Density Curves (dnorm)

A **density curve** is a mathematical function $f(x)$ that can be used to calculate probabilities

Density curves have two defining properties:

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

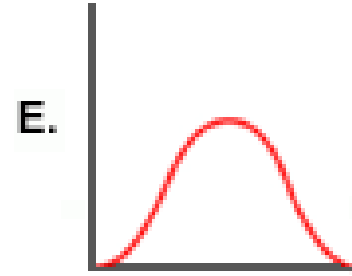
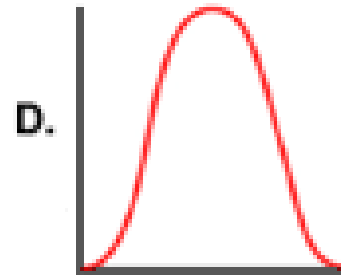
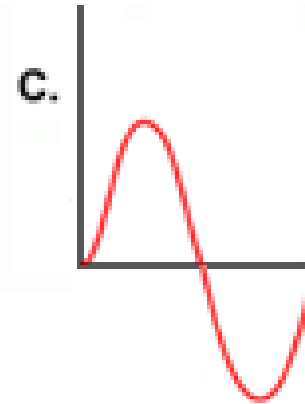
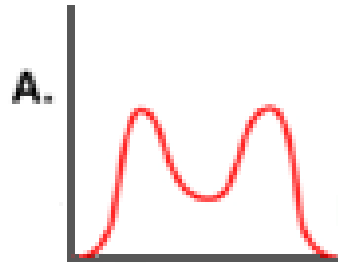


We can think of density curves as histograms that have:

- Infinitely large data sample
- With infinitely small bins sizes
- Normalized to have an area of 1

2. Density Curves

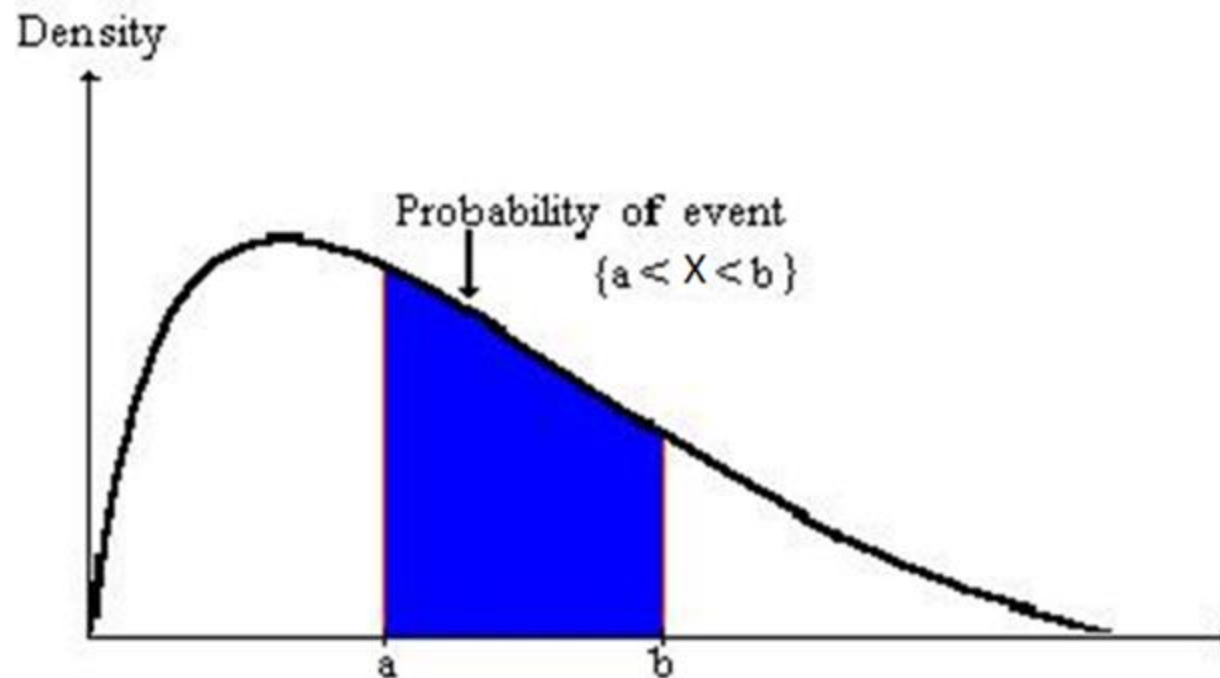
Which of these could **not** be a density curve?



2. Density Curves

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

$P(a < X < b)$ is the area under the curve from a to b



$$P(a < X < b) = \int_a^b f(x)dx$$

The Normal density function

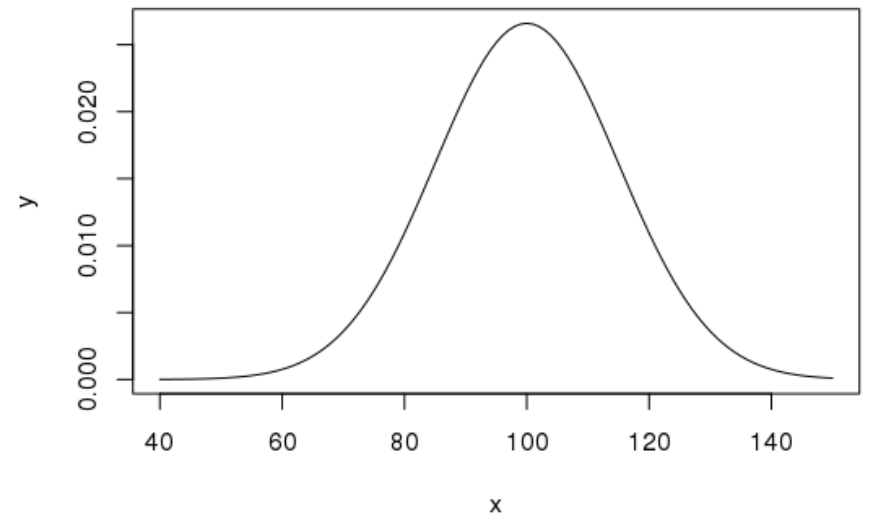
For normal distributions, the equation for the density function is:

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

R: `dnorm(x, mean, sd)`

In this class we will not be concerned with the details of this function

We just need to know that once μ and σ are set, it is just a regular function that takes in an x value and returns a y value



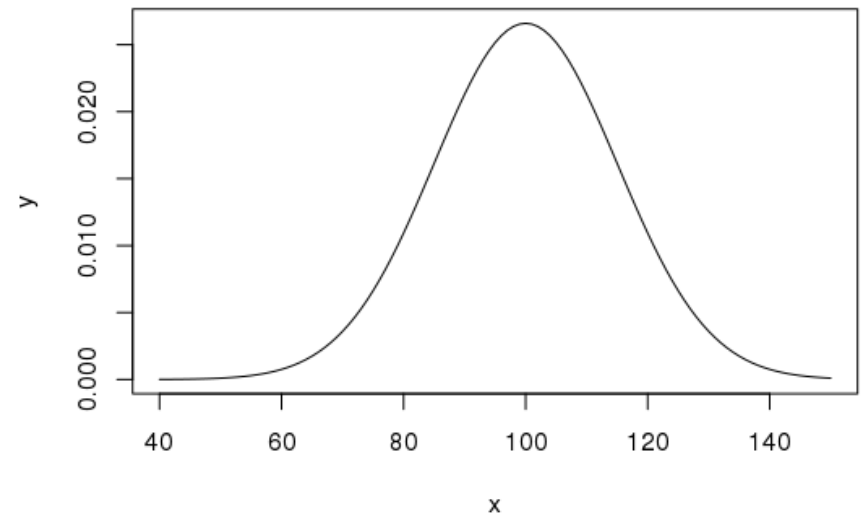
Graphing Normal Curves

We will use density functions to visualize which values are most likely

In R we will use the `dnorm()` function to do this

```
x <- 40:150  
y <- dnorm(x, 100, 15)  
plot(x, y, type = "l")
```

μ σ



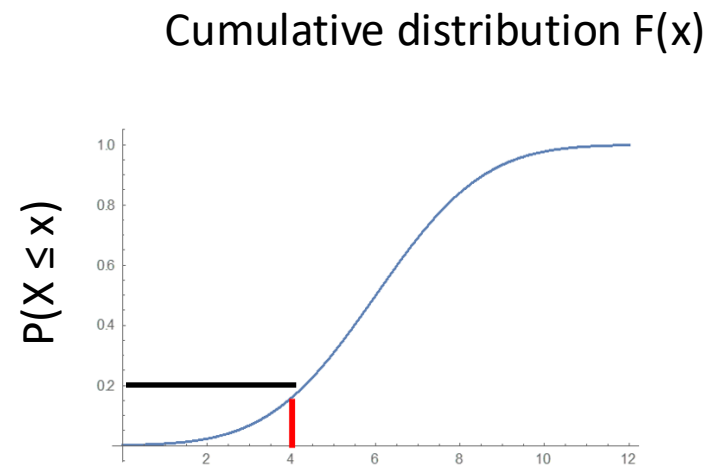
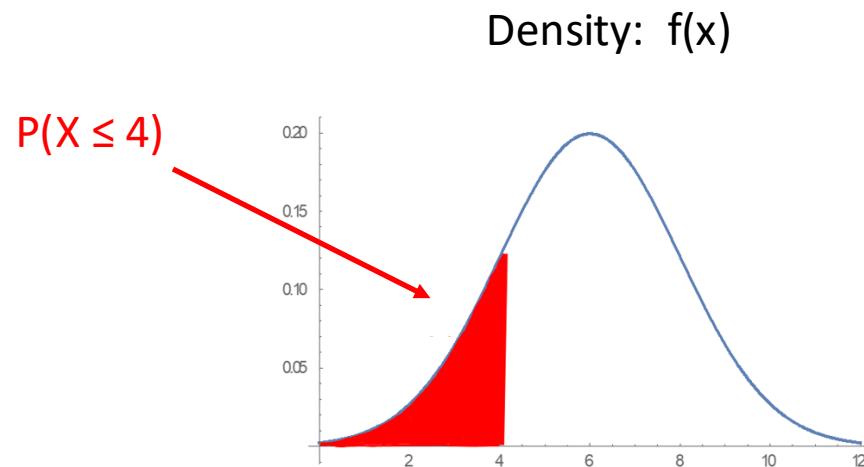
Let's try it in R!

3. Cumulative distribution functions (pnorm)

Cumulative distribution functions give the probability of getting a random value X less than or equal to an input value x : $P(X \leq x)$

- For example, we would write the probability of getting a random number X less than 2 as: $P(X \leq 2)$

Cumulative distribution functions are obtained by calculating the area under a probability density function



$$P(X \leq x)$$

$$= \text{pnorm}(x)$$

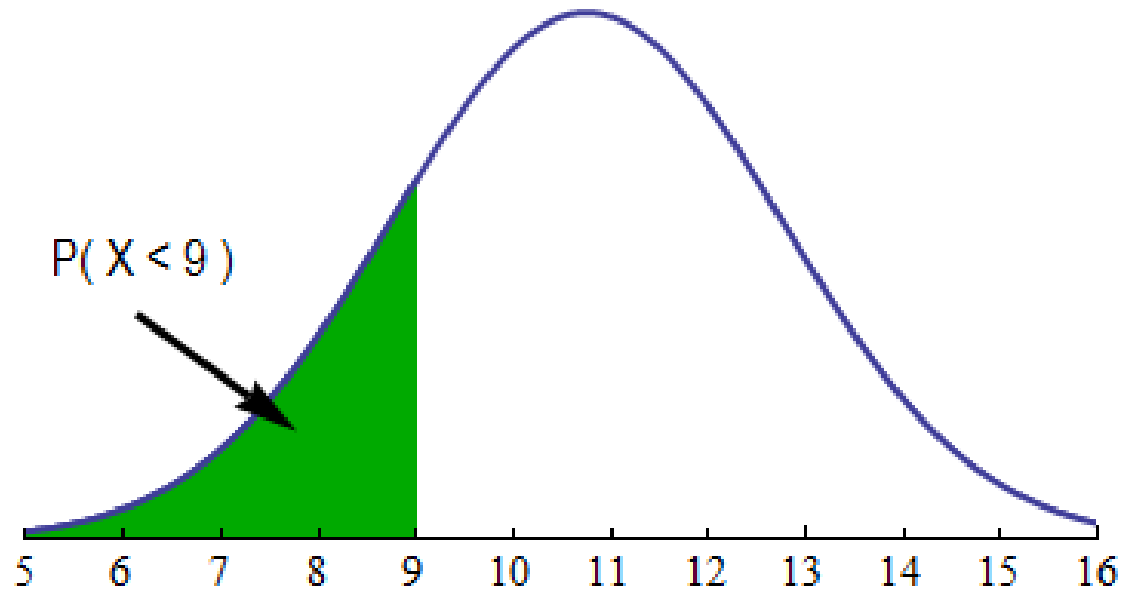
$$= \int_{-\infty}^x f(x) dx$$

Finding normal probabilities of a normal curve

To get the probability (area) from a normal distribution we can use the `pnorm()` function

```
pnorm(x, mean, sd)
```

$P(X < 9; 11, 3)$ μ σ
`pnorm(9, 11, 3)`



Let's try it in R!

4. Calculating quantiles (qnorm)

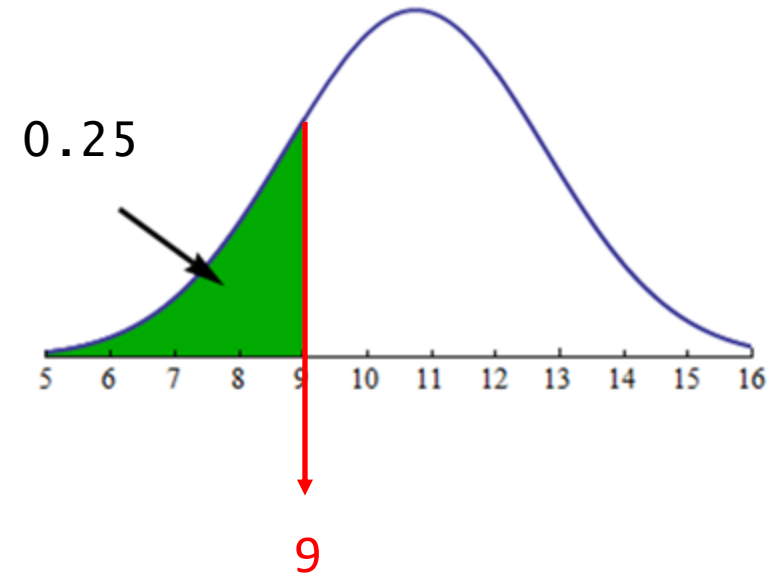
Recall that the p^{th} percentile is the value x such that $p\%$ of your data is less than x

Quantile function gives us the value x such that p proportion of the area of a distribution is less than x

We can do this using quantile function:

```
qnorm(quantile, mean, sd)
```

Value between 0 and 1



Let's try it in R!

Summary of R normal probability functions

Generate random data

- `rnorm(m, mu, sigma)`

Plot the density curve

- `dnorm(x_vec, mu, sigma)`

Get the probability that we would get a random value less than x

- `pnorm(x_vec, mu, sigma)`

Get the quantile value for a given proportion of the distribution

- `qnorm(area, mu, sigma)`