Parametric inference on proportions

Overview

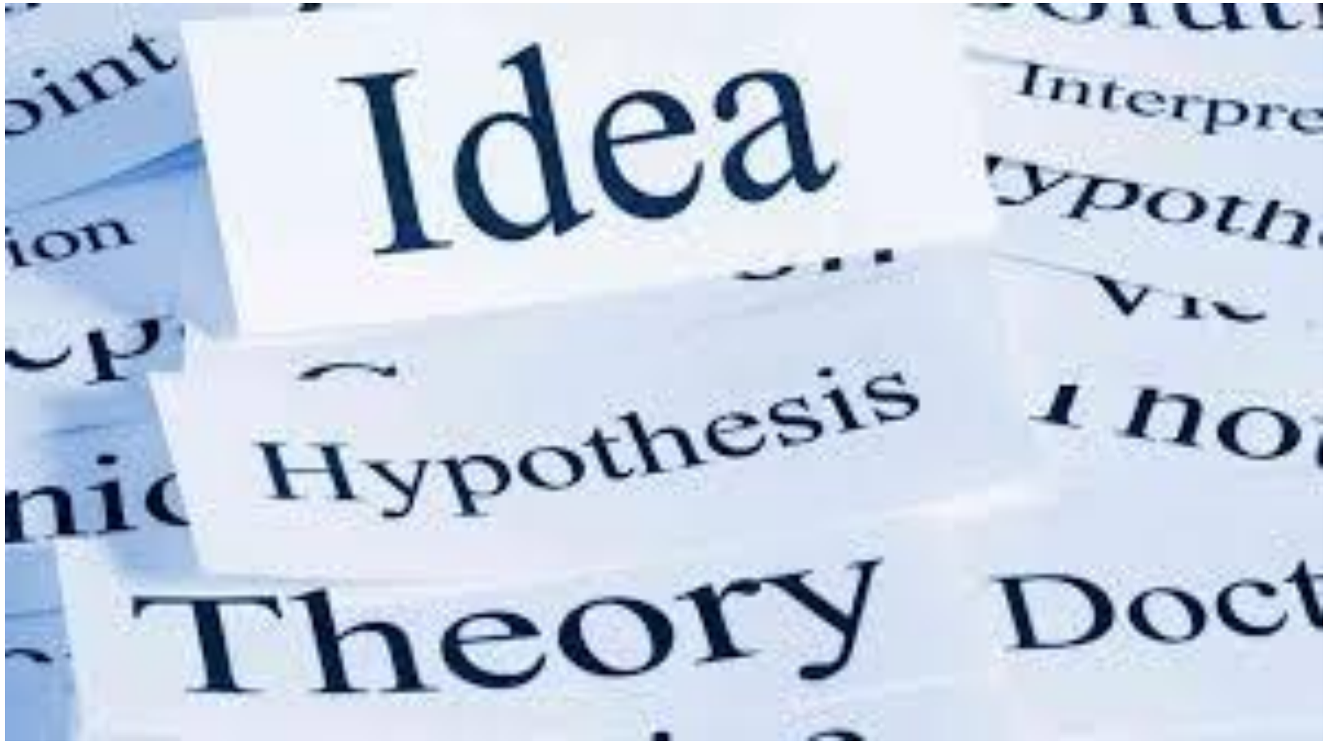
Review of theories of hypothesis testing

Review and continuation of probability distributions and normal distributions

Hypothesis tests and CI using normal distributions

If there is time: parametric inference on proportions

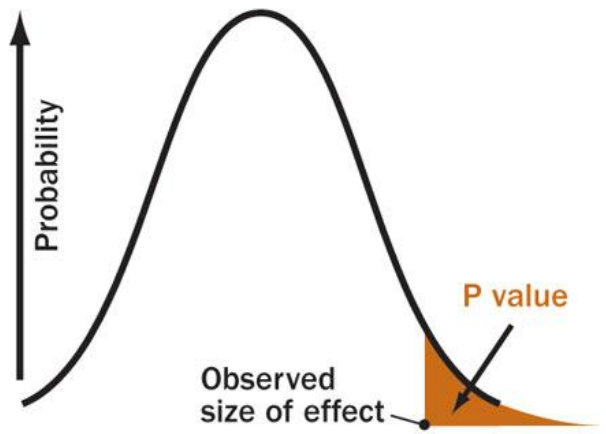
Quick review of theories of hypothesis tests



Two theories of hypothesis testing

1. **Significance testing** of Ronald Fisher
    - p-value as strength of evidence against the null hypothesis

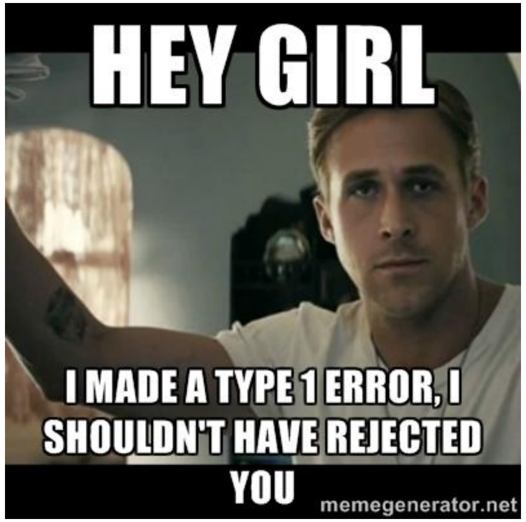
2. **Hypothesis testing** of Jezy Neyman and Egon Pearson
    - Make a formal decision of whether to reject H_0 (if p-value < predefined α value)

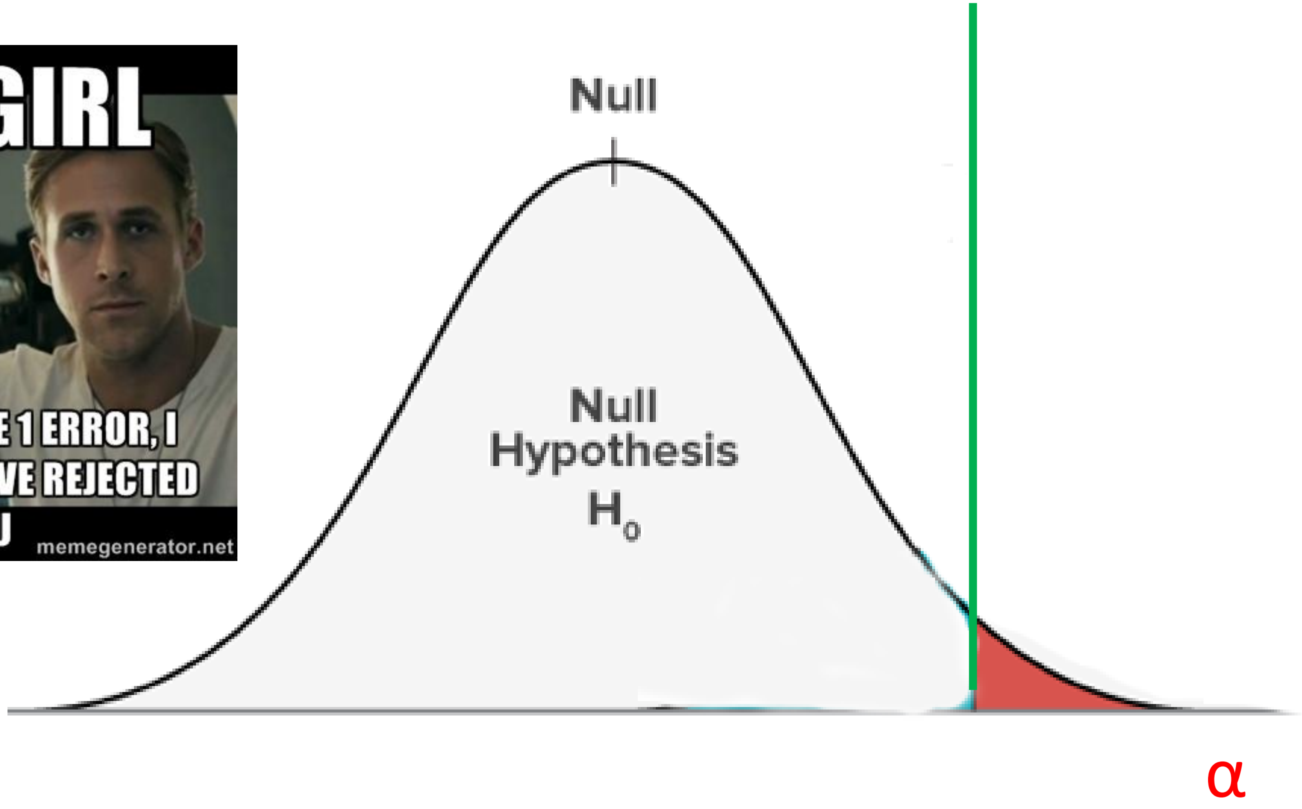


Significance testing

Hypothesis testing

Neyman-Pearson Frequentist logic

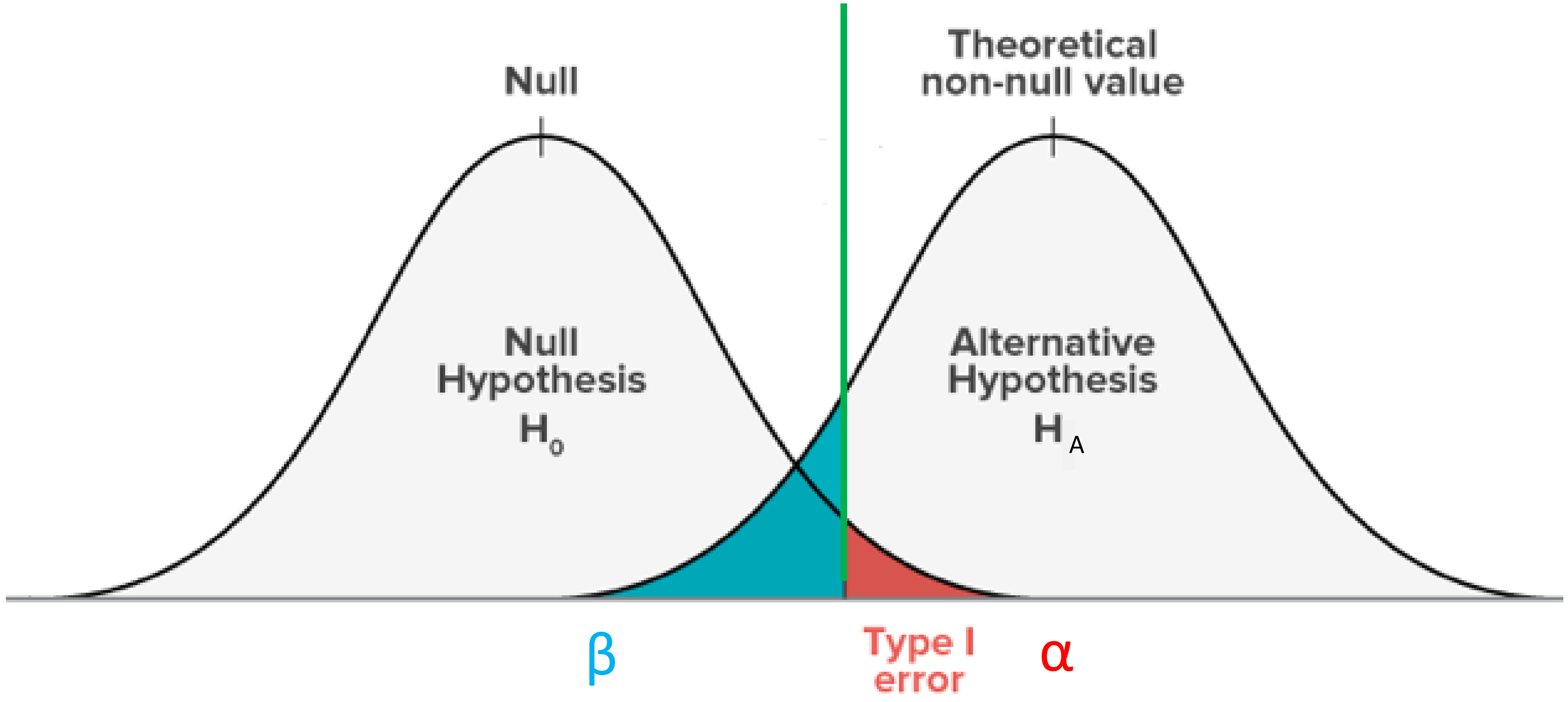




If Neyman-Pearson null hypothesis testing paradigm was followed perfectly, then only ~5% of all published research findings would be type I errors (for α = 0.05)

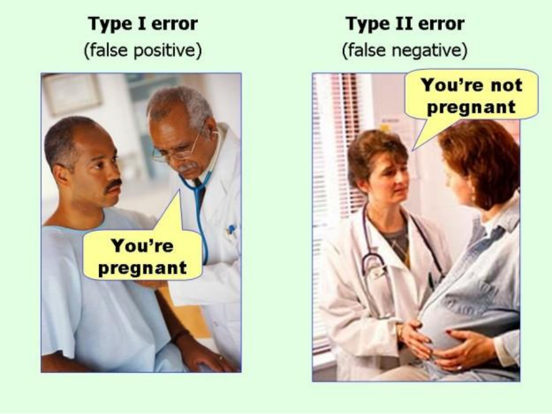
Type I error: incorrectly rejecting the null hypothesis when it is true

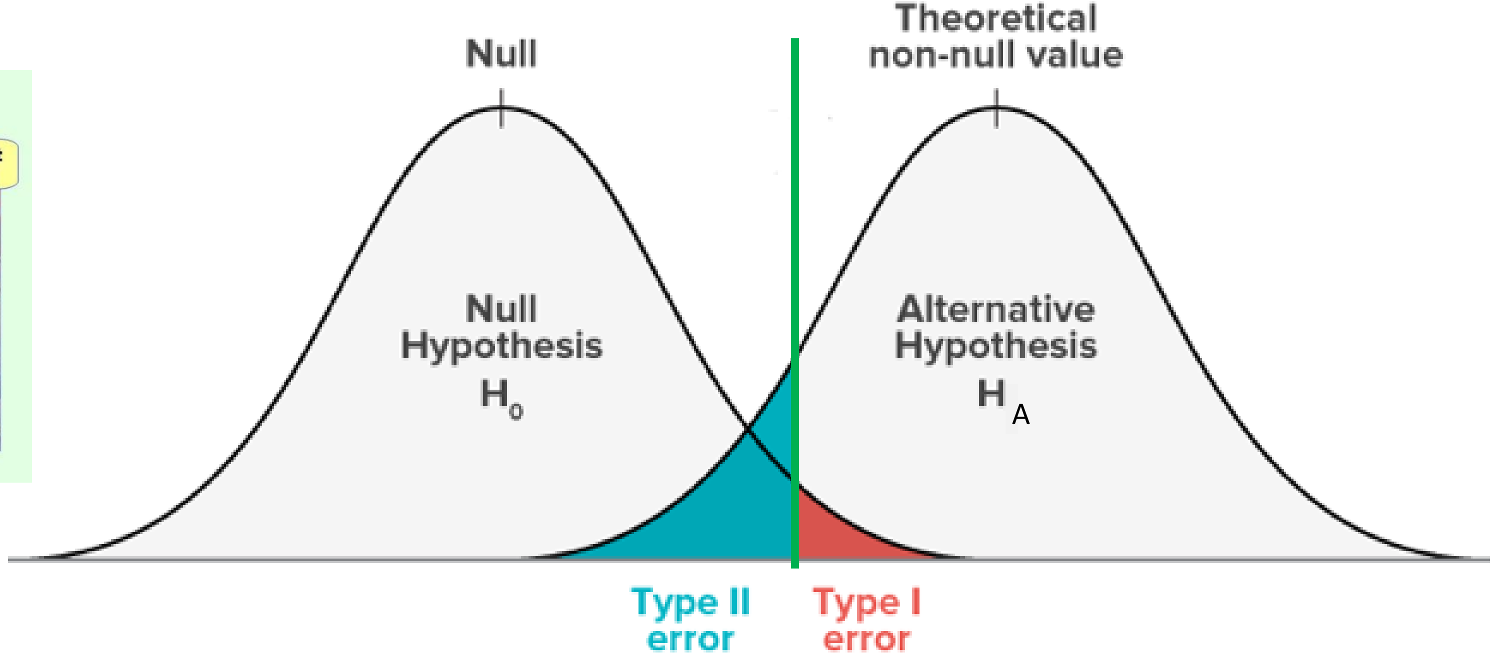
Neyman-Pearson Frequentist logic



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

Type I and Type II Errors





	H_0 is true	H_A is true (H_o 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)

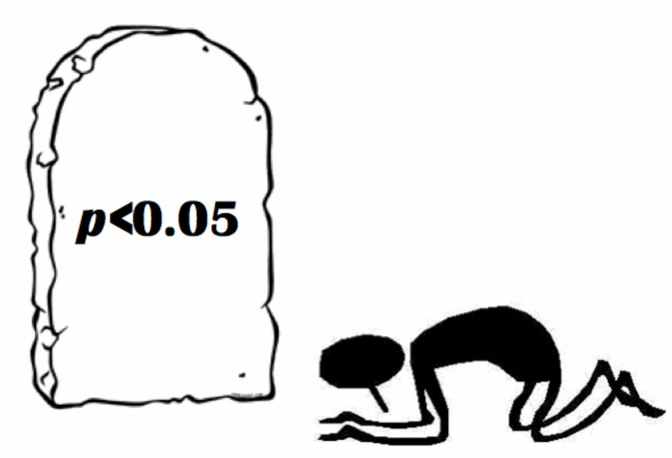
Problems with the NP hypothesis tests

<u>Problem 1</u>: 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, …

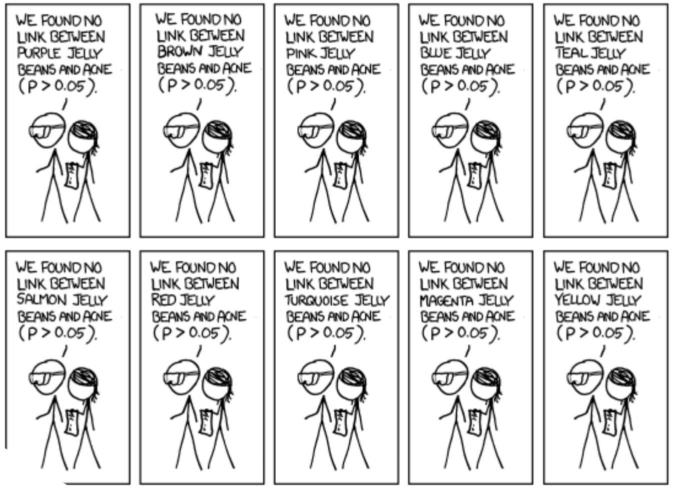
<u>Problem 2</u>: Arbitrary thresholds for alpha levels

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

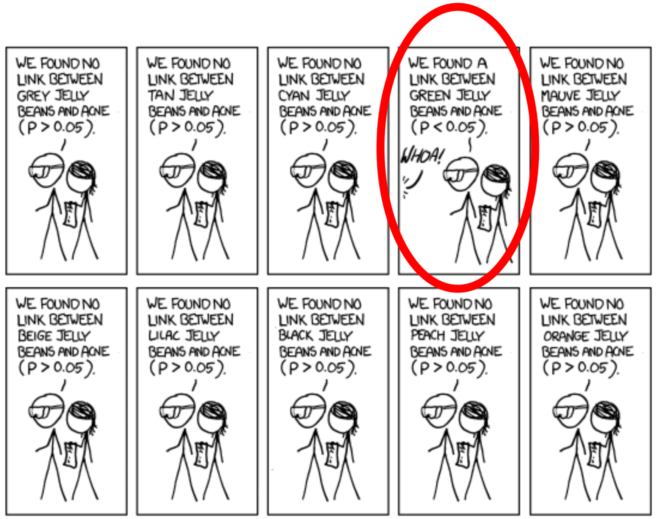


<u>Problem 3</u>: running many tests can give rise to a high number of type 1 errors

Multiple hypothesis tests



Don't ever do this!



Replication crisis

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

The file drawer effect

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

Some thoughts…

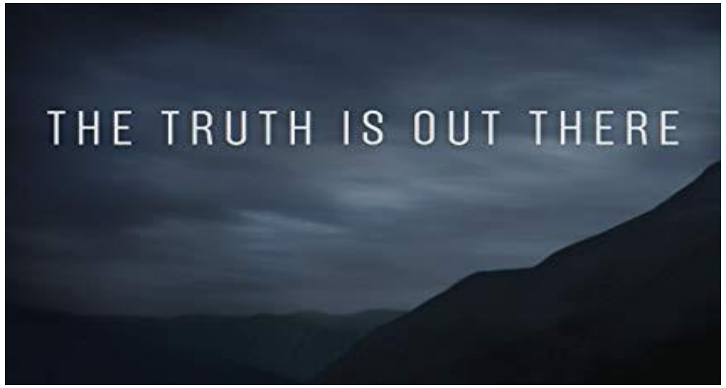
Better to have hypothesis tests than none at all. Just need to think carefully and use your judgment.

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

Report the p-values rather than accept/reject H_0

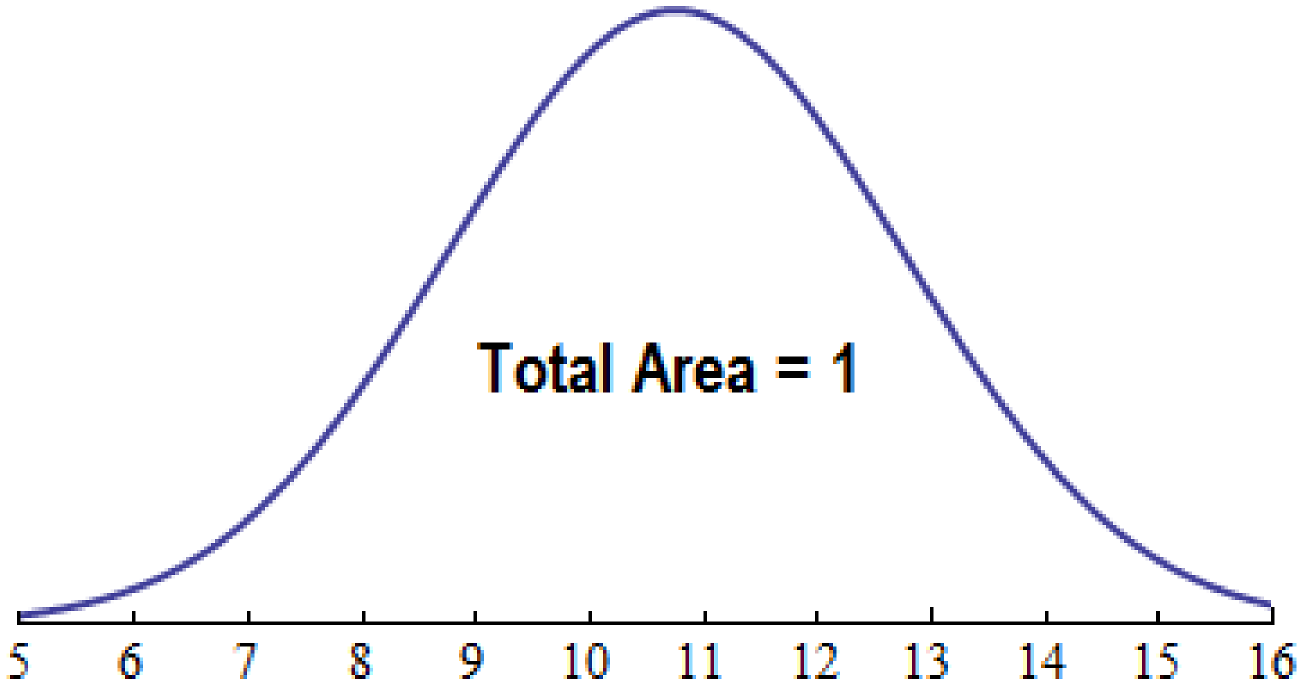
- i.e., report $p = 0.023$ not $p < 0.05$

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!

Review and continuation of inference using parametric probability distributions



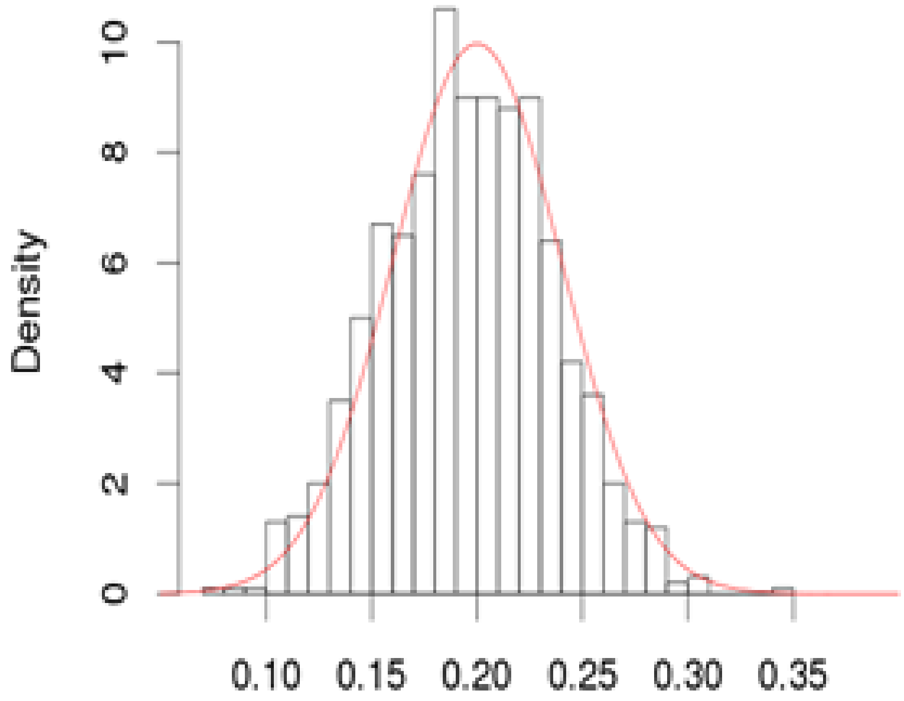
Inference using parametric probability distributions

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

- E.g., we can use these mathematical functions for null distributions rather than using computer simulations

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”



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**)

Main use case: Normal probability distributions

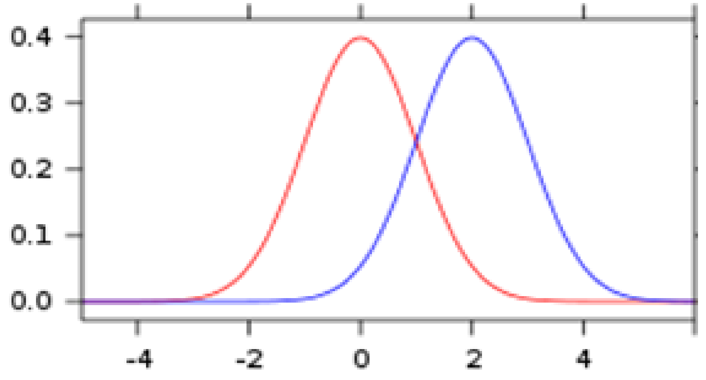
Normal distributions are a family of bell-shaped curves with two parameters

- The mean: μ
- The standard deviation: σ

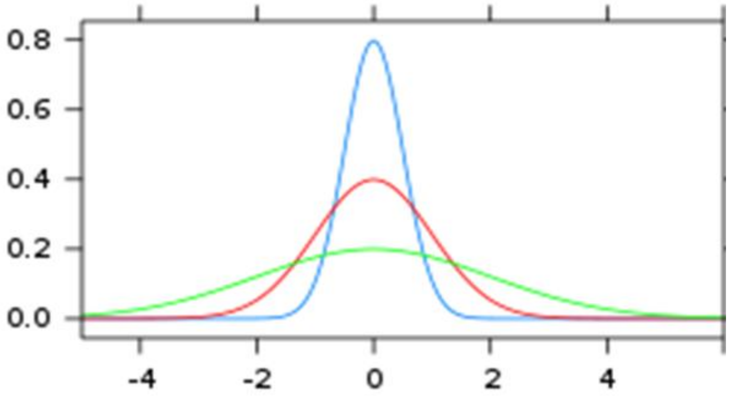
Once μ and σ are set, we have a specific probability distribution that we use to get probabilities of events, generate random data, etc.

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

Changing μ



Changing σ



1. Generating random numbers (rnorm)

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

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

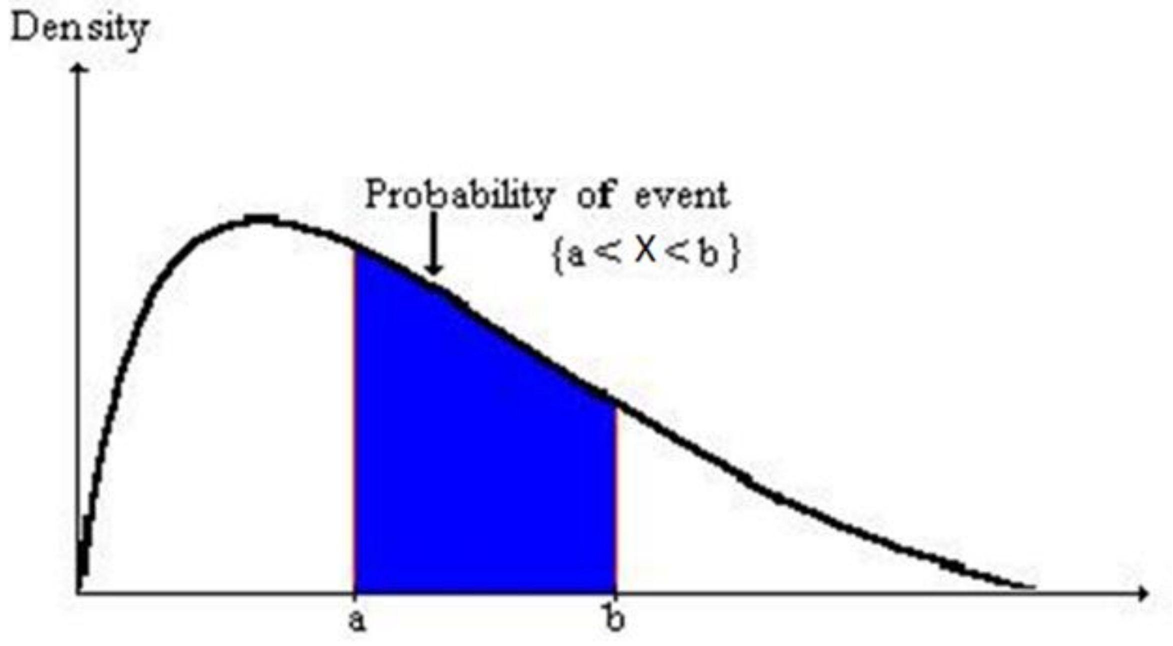
We denote random numbers that come from a **normal distribution** using X ~ N(μ, σ)

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

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



Density curve are functions f(x) that have two key properties:

1. The total area under the curve f(x) is equal to 1

2. The curve is always ≥ 0

2. Density Curves

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

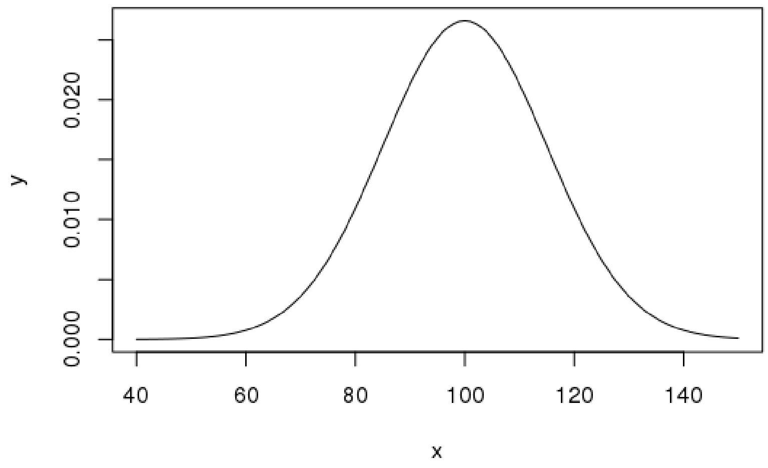
- We can think of them a normalized histogram that has infinite data

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

```
x <- seq(40, 150, length.out = 1000)

y <- dnorm(x, 100, 15)

plot(x, y, type = "l")
```



3. Cumulative distribution functions (pnorm)

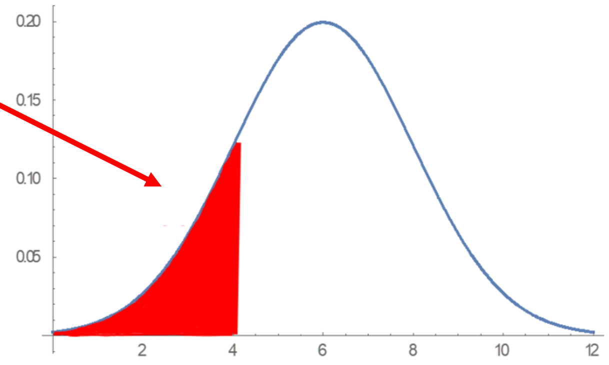
Cumulative distribution functions give the probability of getting a random value X less than or equal to an input value x: $P(X \leq x)$

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

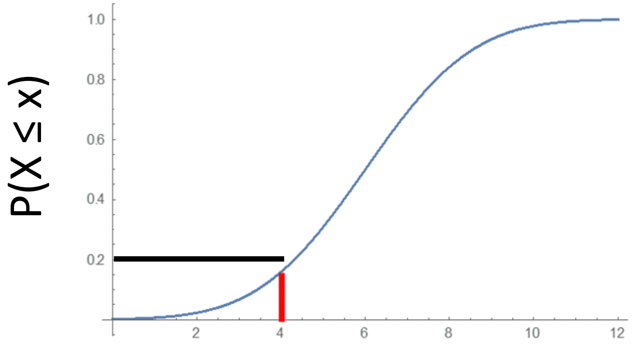
Cumulative distribution functions are obtained by calculating the area under a probability density function

Density: f(x)

P(X ≤ 4)



Cumulative distribution F(x)



$$P(X \leq x)$$

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

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

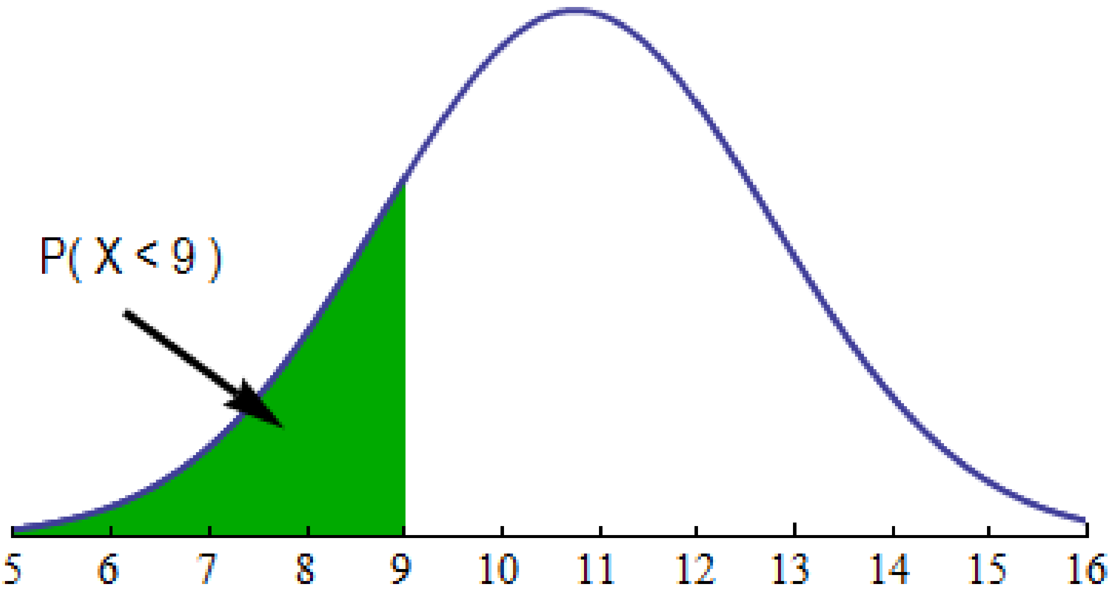
Finding normal probabilities of a normal curve

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)
```

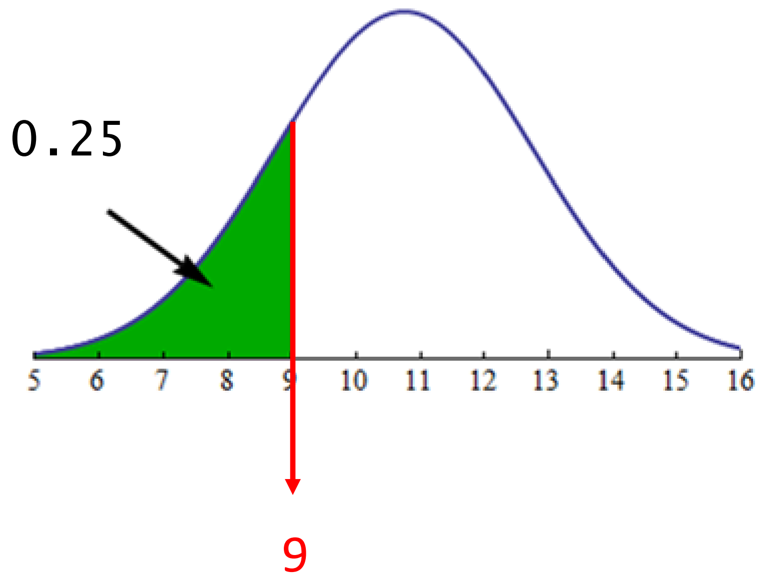


Let's quickly review these in R!

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



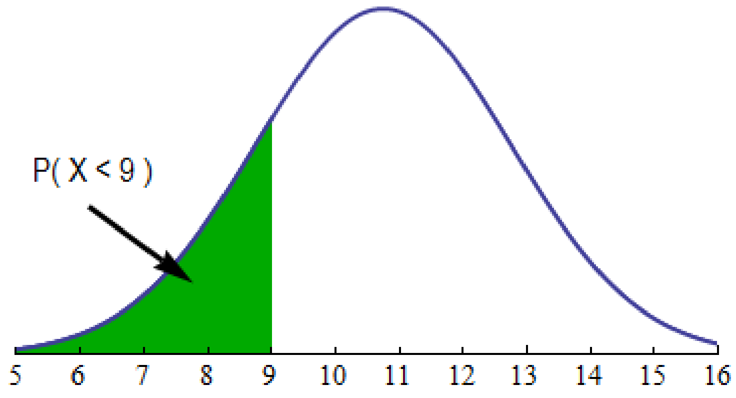
We can do this using quantile function:

```
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, mean, sd)

Plot the density curve

- dnorm(x_vec, mean, sd)

Get the probability that we would get a random value less than x: P(X < x)

- pnorm(x_vec, mean, sd)

Get the quantile value for a given proportion of the distribution

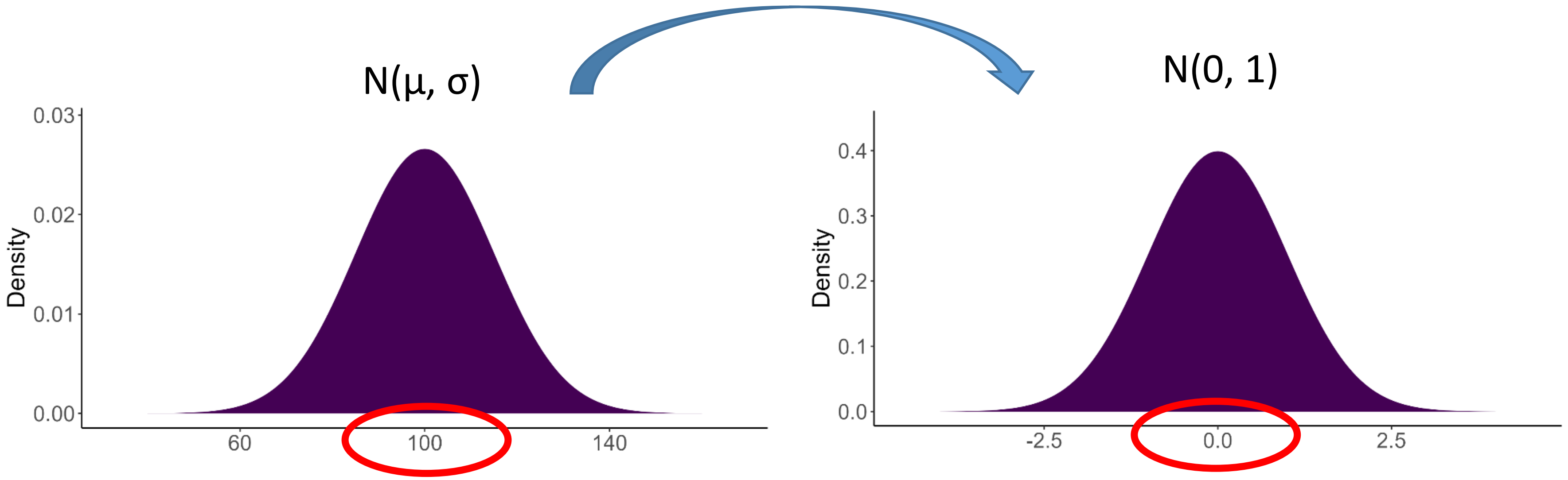
- qnorm(area, mean, sd)

The Standard Normal distribution

Standard Normal N(0, 1)

Since all normal distributions have the same shape, it is convenient to convert them to a standard scale with: **$\mu = 0$, $\sigma = 1$**

This is called the **standard normal** distribution: $Z \sim N(0, 1)$



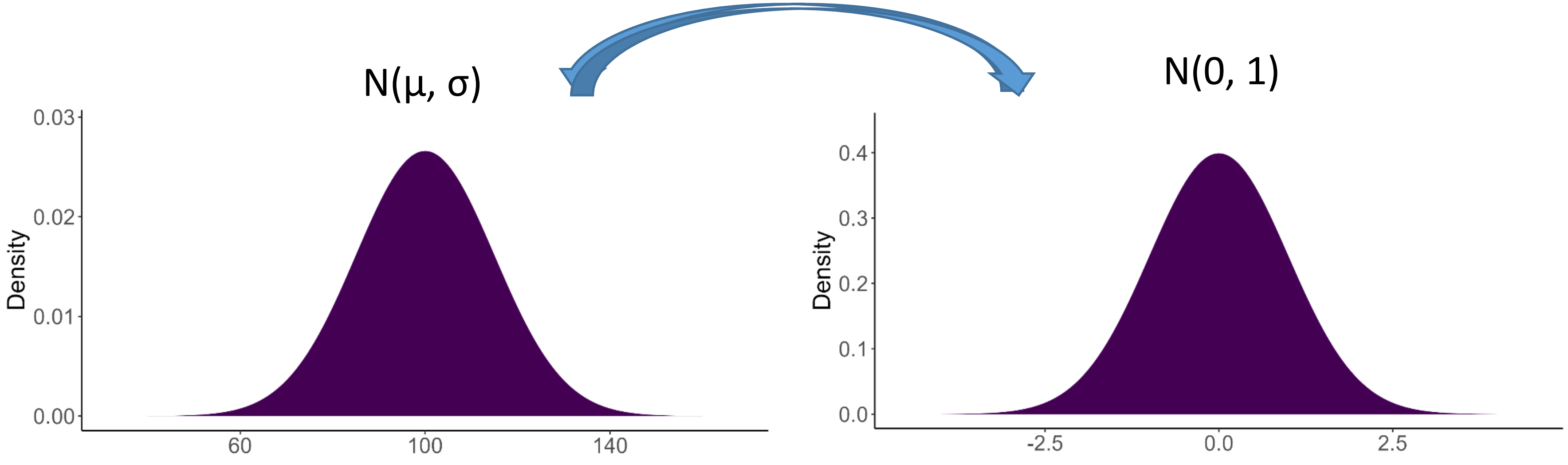
Converting to the standard normal distribution

We can apply a z-score transformation to any normally distributed random variable $X \sim N(\mu, \sigma)$ to convert it to the standard normal distribution $Z \sim N(0, 1)$:

$$Z = (X - \mu)/\sigma$$

To convert from $Z \sim N(0, 1)$ to any $X \sim N(\mu, \sigma)$, we reverse the standardization with:

$$X = \mu + Z \cdot \sigma$$



Converting to the standard normal distribution

1. What is the Z-score of someone who has an IQ score of 112?

$$Z = (X - \mu)/\ \sigma$$

2. What if someone has a Z-score of 2.2, what is their IQ score?

$$X = \mu + Z \cdot \sigma$$

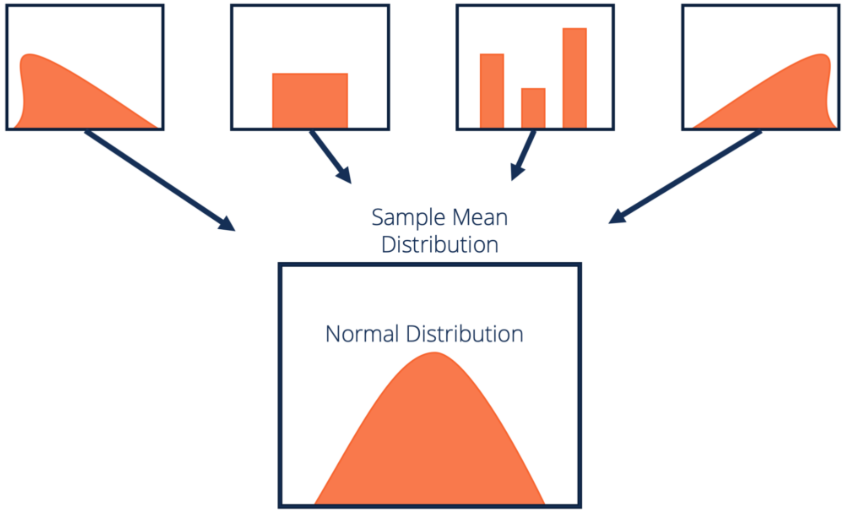
Let's quickly try it in R!

The Central Limit Theorem

Central limit theorem

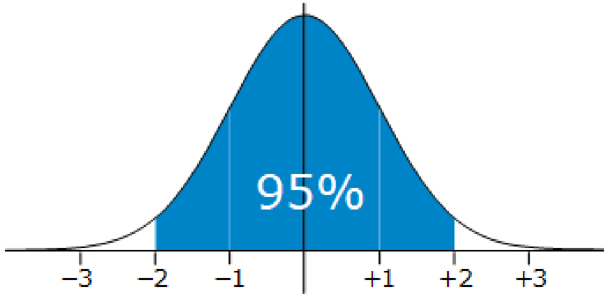
For random samples with a sufficiently large sample size (n), the distribution of sample statistics for a mean ($\bar{x}$) or a proportion ($\hat{p}$) is:

- Normally distributed
- Centered at the value of the population parameter



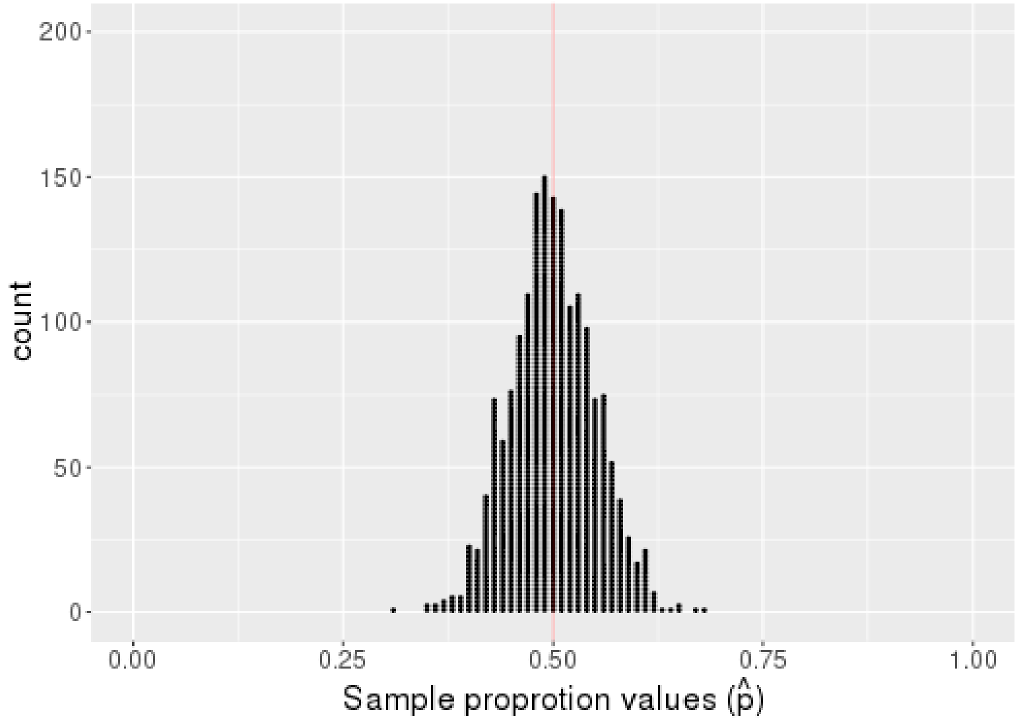
Upshot: We can create confidence intervals and run hypothesis tests using the normal distribution

- Rather than using computational methods like the bootstrap or a randomization methods to create a null distribution!



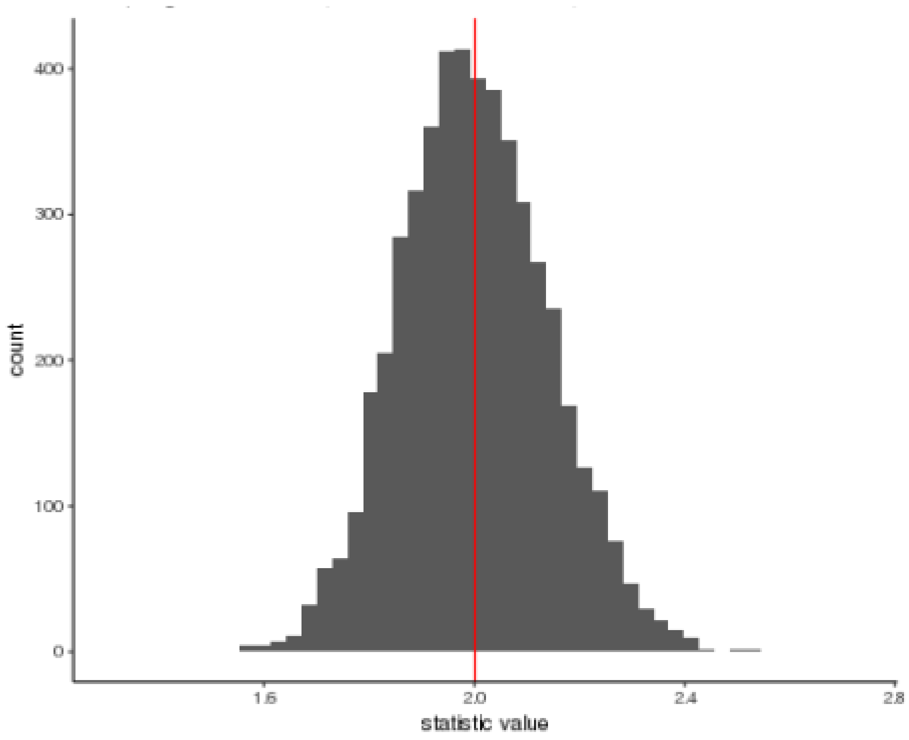
Central limit theorem

proportion ($\hat{p}$)



[Proportion sampling distribution app]

mean ($\bar{x}$)



[Sampling/Bootstrap distribution app]

Summary

For large n, the sampling distributions of $\bar{x}$ and $\hat{p}$ have normal distributions

We can convert any normal distribution $N(\mu, \sigma)$, into a standard normal distribution $N(0, 1)$

We are now (almost) ready to run hypothesis tests and compute confidence intervals for $\bar{x}$ and $\hat{p}$ using normal distributions

Let's quickly explore the central limit theorem in R!

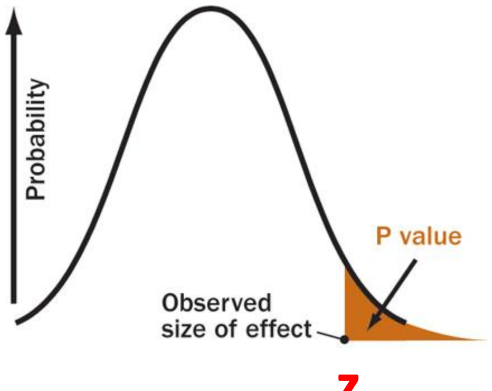
Hypothesis tests and confidence intervals using a normal distribution

Hypothesis tests based on a Normal Distribution

When the null distribution is normal, it is often convenient to use a standard normal test statistic using:

$$z = \frac{\textit{Sample Statistic} - \textit{Null Parameter}}{SE}$$

The p-value for the test is the probability a standard normal value is beyond this standardized test statistic



$P(\, Z \geq z_{obs} \; ; \quad \mu = 0, \; \sigma = 1)$

pnorm(z, 0, 1, lower.tail = FALSE)

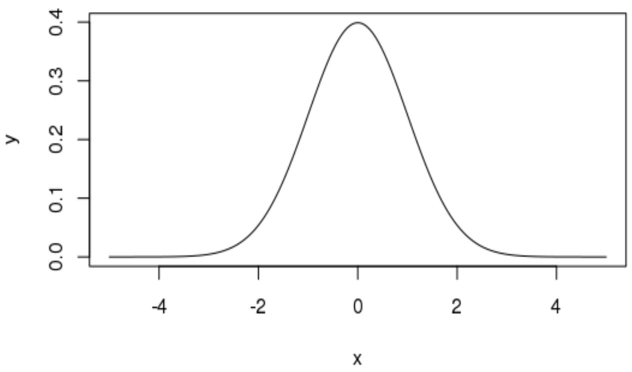
Hypothesis tests based on a Normal Distribution

To repeat what was on the last slide: we can transform our obs_stat to a z-statistic that comes from a standard normal distribution N(0, 1) using:

$$z = \frac{stat_{obs} - param_0}{SE}$$

The p-value is then the probability of obtaining a value from a standard normal distribution beyond this z statistic

`pnorm(z, 0, 1)`	if H_A: $\mu < param_0$
`1 - pnorm(z, 0, 1)`	if H_A: $\mu > param_0$
`2 * (1 - pnorm(abs(z), 0, 1))`	if H_A: $\mu \neq param_0$

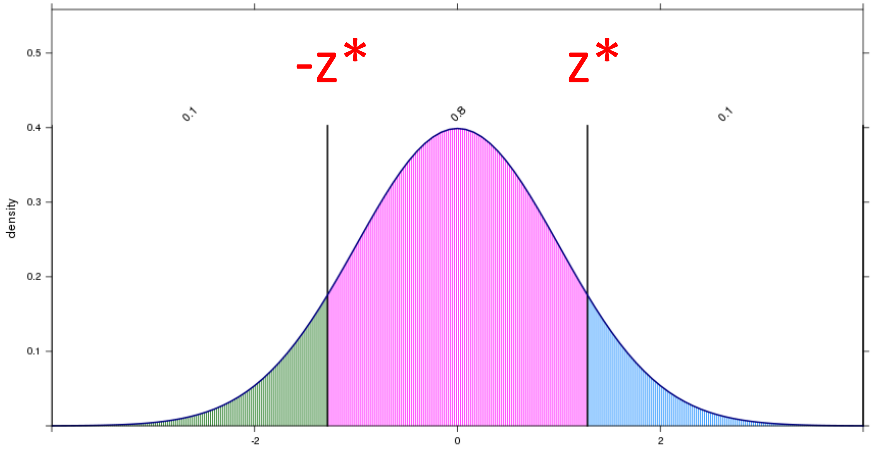


Confidence intervals based on a Normal Distribution

If the distribution for a statistic is normal with a standard error SE, we can find a confidence interval for the parameter using:

$$\text{sample statistic} \pm z^* \times \text{SE}$$

where z* is chosen so that the area between –z* and + z* in the standard normal distribution is the desired confidence level



Confidence level	80%	90%	95%	98%	99%
Z*	1.282	1.645	1.960	2.326	2.576

```
z_stars <- qnorm(c(.90, .95, .975, .99, .995), 0, 1)
```

Do goalies guess the direction of a penalty shot less than 50% of the time?

1. Start by stating H_0 and H_A

H_0: $\pi = .5$

H_A: $\pi < .5$

2. Calculate the observed statistic

- Goal keepers correctly guessed the direction 41% of the time out of 128 kicks
- With SE* = 0.043
  - (could you calculate this SE*?)

$$z = \frac{Sample\ \ Statistic\ \ -\ \ Null\ \ Parameter}{SE}$$

Can you compute a z-statistic?

Let's try the rest in R!

Do goalies guess the direction of a penalty shot less than 50% of the time?

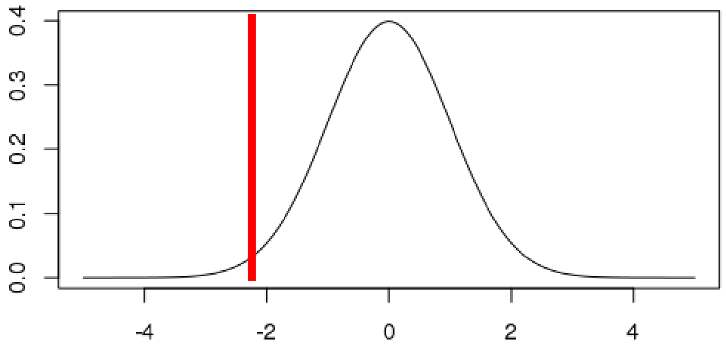
Steps: 3-4. What is the probability one would get a z-statistic as small or smaller than -2.093 from a standard normal distribution?

pnorm(-2.093, 0, 1)

[Normal area app P(X ≤ x)]

p-value = 0.018

Standard normal null distribution



Step 5?