## Getting Started

*Utilizarei o símbolo > para indicar o input do console e >>> para o output
-As funções básicas são o R base, o restante são adicionados pela comunidade com CRAN ou github.

## INTRODUCTION TO INFERENCE

-*Polls* são importantes quando é impossível perguntar a todos de uma população (logicamente impossível)
-A estratégia então é perguntar a um grupo menor escolhidos aleatoriamente e **inferir** que este grupo representa a opnião da população total.
-Eleições são casos de *pools* de opiniões.

### SAMPLING MODEL PARAMETERSAND ESTIMATES

-Em uma urna, tem-se a proporção de bolas azuis P, bolas vermelhas 1-p e o *spread*[propagação] P-(1-P) que é simplificado para 2P-1.
As bolas na urna são chamadas de *POPULAÇÃO* e as bolas azuis(P) são chamadas de *PARÂMETROS.*

### THE SAMPLE AVERAGE

-Precisamos estimar um parâmetro dado uma amostra; através de processos estimamos P. Por esta estimativa podemos calcular a propagação estimada, 2P-1.

-Em estatística, uma barra em cima de um símbolo denota a média, $\bar{X}$ .

### POLLING vs FORECASTING

-Se uma pesquisa de eleição é conduzida 4 meses antes das eleições, o P valerá para aquele momento, e não para o dia das eleições.

### PROPERTIES OF OUR ESTIMATE

- $\bar{X}$ é a soma de sorteios independentes de uma variável aleatória vezes uma constante 1/N:
$\mathrm{E}(\bar{X}) = p$

-Conforme o número de sorteios N aumenta, o desvio padrão da nossa estimativa diminui. O desvio padrão da média de X em N sorteios é: $\mathrm{SE}(\bar{X}) = \sqrt{p(1-p)/N}$
Porém, vale lembrar que na prática, aumentar N também implica em mais custo, trabalho e tempo.

## THE CENTRAL LIMIT THEOREM IN PRACTICE

-O teorema do limite central nos diz que a *distribution function* de uma soma de sorteios é aproximadamente normal.

### Key points

- Because $\bar{X}$ is the sum of random draws divided by a constant, the distribution of $\bar{X}$ is approximately normal.
- We can convert $\bar{X}$ to a standard normal random variable $Z$:

$$Z = \frac{\bar{X} - \mathrm{E}(\bar{X})}{\mathrm{SE}(\bar{X})}$$

- The probability that $\bar{X}$ is within .01 of the actual value of $p$ is:

$$\Pr\left(Z \leq .01/\sqrt{p\left(1-p\right)/N}\right) - \Pr\left(Z \leq -.01/\sqrt{p\left(1-p\right)/N}\right)$$

- The Central Limit Theorem (CLT) still works if $\bar{X}$ is used in place of $p$. This is called a *plug-in estimate*. Hats over values denote estimates. Therefore:

$$\hat{\mathrm{SE}}\left(\bar{X}\right) = \sqrt{\bar{X}\left(1-\bar{X}\right)/N}$$

- Using the CLT, the probability that $\bar{X}$ is within .01 of the actual value of $p$ is:

$$\Pr\left(Z \leq .01/\sqrt{\bar{X}\left(1-\bar{X}\right)/N}\right) - \Pr\left(Z \leq -.01/\sqrt{\bar{X}\left(1-\bar{X}\right)/N}\right)$$

### Code: Computing the probability of $\bar{X}$ being within .01 of $p$

```
X_hat <- 0.48
se <- sqrt(X_hat*(1-X_hat)/25)
pnorm(0.01/se) - pnorm(-0.01/se)
```

**MARGIN OF ERROR**

-É equivalente a 2*(standard error)

**THE SPREAD**

-A propagação entre 2 resultados com propabilidades p e 1-p é de **2p-1**

-O valor esperado da propação é de $2\bar{X} - 1.$

-O desvio padrão da propagação é de $2\hat{\mathrm{SE}}\left(\bar{X}\right).$

-A margem de erro da propagação é de 2 vezes a margem de erro de $\bar{X}.$

**POR QUE NÃO RODAR UMA PESQUISA GIGANTESCA?**

-Alto custo(R$)

-Pesquisas são mais complicadas que sortear números de uma urna, pessoas podem mentir nas pesquisas, ao contrário das bolas da urna que são azuis ou vermelhas. Além disso, há vários fatores extras quando se compara humanos a objetos.

## CONFIDENCE INTERVALS AND P-VALUES

-O intervalo de propagação de 10% não é considerado um bom intervalo por ser muito amplo.
-Um intervalo pequeno de propagação mas que erra na maioria das vezes também não é considerado um bom intervalo.
-Intervalos de confiança de 95% são intervalos construídos com 95% de chance de incluir P. A marge de erro é aproximadamente de 95% *confidence intervalo*

For a confidence interval of size $q$, we solve for $z = 1 - \frac{1-q}{2}$.

To determine a 95% confidence interval, use `z <- qnorm(0.975)`. This value is slightly smaller than 2 times the standard error.

-Lembre-se que o intervalo de 95% de confiança é aleatório, mas P não é aleatório.

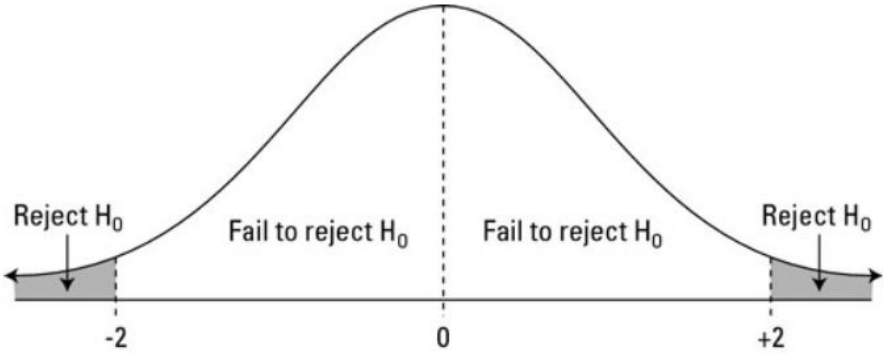
### P VALUES

The p-value is the probability of observing a value as extreme or more extreme than the result given that the null hypothesis is true.

In the context of the normal distribution, this refers to the probability of observing a Z-score whose absolute value is as high or higher than the Z-score of interest.

Suppose we want to find the p-value of an observation 2 standard deviations larger than the mean. This means we are looking for anything with $|z| \geq 2$.

Graphically, the p-value gives the probability of an observation that's at least as far away from the mean or further. This plot shows a standard normal distribution (centered at $z = 0$ with a standard deviation of 1). The shaded tails are the region of the graph that are 2 standard deviations or more away from the mean.




The right tail can be found with `1-pnorm(2)`. We want to have both tails, though, because we want to find the probability of any observation as far away from the mean or farther, in either direction. (This is what's meant by a two-tailed p-value.) Because the distribution is symmetrical, the right and left tails are the same size and we know that our desired value is just `2*(1-pnorm(2))`.

Recall that, by default, `pnorm()` gives the CDF for a normal distribution with a mean of $\mu = 0$ and standard deviation of $\sigma = 1$. To find p-values for a given z-score `z` in a normal distribution with mean mu and standard deviation `sigma`, use `2*(1-pnorm(z, mu, sigma))` instead.

## STATISTICAL MODELS

### POLL AGGREGATORS

- *Poll aggregators* combinam o resultado de muitas *polls* para simular *polls* com uma larga *sample size* e então gerar mais precisamente estimativas do que *polls* individuais.
- *Polls* podem ser simuladas com simulações de Monte Carlo e usadas para construir e estimar a propagação e confiança de intervalos.

**CODE:**

**Simulating polls**

```
//Note that to compute the exact 95% confidence interval, we would use qnorm(.975)*SE_hat instead of 2*SE_hat.

d <- 0.039
Ns <- c(1298, 533, 1342, 897, 774, 254, 812, 324, 1291, 1056, 2172, 516)
p <- (d+1)/2

# calculate confidence intervals of the spread
confidence_intervals <- sapply(Ns, function(N){
    X <- sample(c(0,1), size=N, replace=TRUE, prob = c(1-p, p))
    X_hat <- mean(X)
    SE_hat <- sqrt(X_hat*(1-X_hat)/N)
    2*c(X_hat, X_hat - 2*SE_hat, X_hat + 2*SE_hat) - 1
})

# generate a data frame storing results
polls <- data.frame(poll = 1:ncol(confidence_intervals),
                    t(confidence_intervals), sample_size = Ns)
names(polls) <- c("poll", "estimate", "low", "high", "sample_size")
polls
```

**CODE:**

**Calculating the spread of combined polls:**

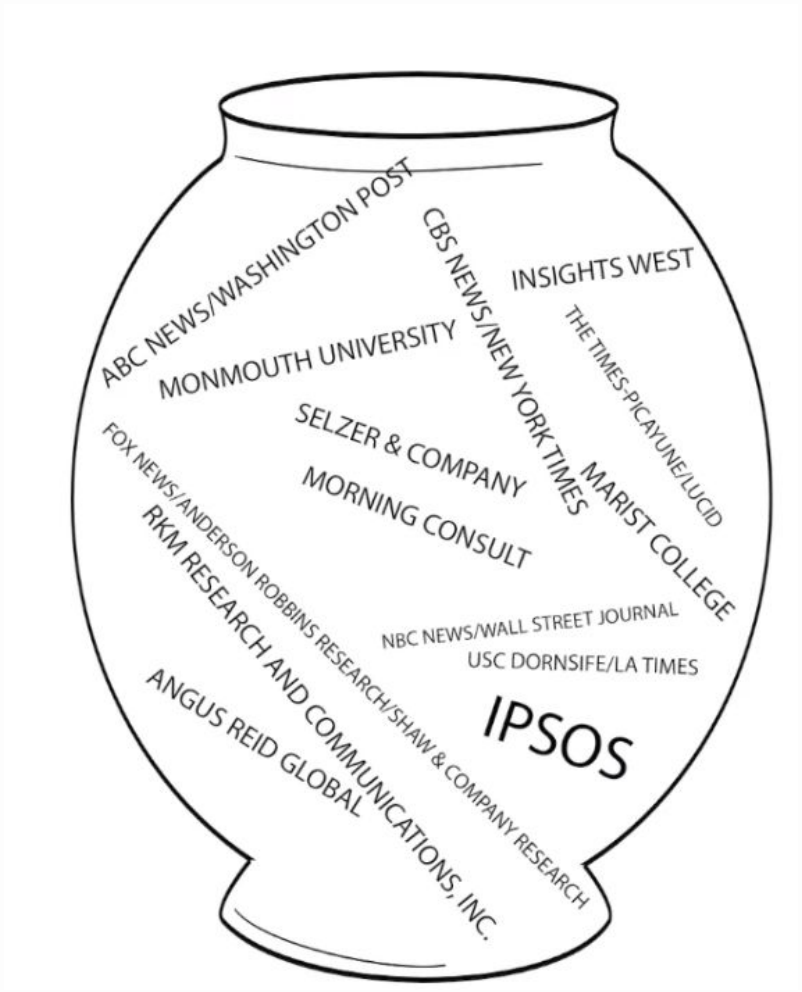
```
//Note that to compute the exact 95% confidence interval, we would use qnorm(.975) instead of 1.96.

d_hat <- polls %>%
    summarize(avg = sum(estimate*sample_size) / sum(sample_size)) %>%
    .$avg

p_hat <- (1+d_hat)/2
moe <- 2*1.96*sqrt(p_hat*(1-p_hat)/sum(polls$sample_size))
round(d_hat*100,1)
round(moe*100, 1)
```

### POLLSTER AND MULTILEVEL MODELS

-Diferentes *poll aggregators* geram diferentes predições de resultados de eleições mesmo vindo de uma mesma *poll* de dados. Isso acontece porque usam diferentes modelos estatísticos.

**DATA-DRIVEN MODELS**

-Usar o modelo de teoria de urna para combinar resultados pode não ser uma boa escolha algumas vezes por conta do *pollster effect.*
-Em vez de termos bolinhas com 0's e 1's dentro da urna, agora nossa urna conterá *poll results* de todas as possíveis *polls.*




-Assumimos que o valor esperado da urna é a propagação (*spread*), que é d = 2p - 1.
-Nosso desvio padrão agora inclui *pollster-to-pollster variability.*
-Nosso novo desvio padrão **σ** agora leva em consideração a varibilidade de *pollster-to-pollster.*E ele não pode mais ser calculado através de p e d, e agora é um parâmetro desconhecido.
-O teorema do limite central ainda funciona para estimar o tamanho da amostra porque a soma de muitas variáveis aleatórias é uma variável aleatória normalmente distribuída com valor esperado d e erro padrão σ/sqrt(N);
-Podemos estimar o σ não observado como o desvio padrão,que pode ser calculado com a função sd.

**CODE:**

```
//Note that to compute the exact 95% confidence interval, we would use qnorm(.975) instead of 1.96.

# collect last result before the election for each pollster
one_poll_per_pollster <- polls %>% group_by(pollster) %>%
  filter(enddate == max(enddate)) %>%      # keep latest poll
  ungroup()
```

```
# histogram of spread estimates
one_poll_per_pollster %>%
   ggplot(aes(spread)) + geom_histogram(binwidth = 0.01)

# construct 95% confidence interval
results <- one_poll_per_pollster %>%
   summarize(avg = mean(spread), se = sd(spread)/sqrt(length(spread))) %>%
   mutate(start = avg - 1.96*se, end = avg + 1.96*se)
round(results*100, 1)
```

## BAYESIN STATISTICS

## BAYESIAN STATISTICS

Bayes' Theorem

A Probabilidade de um evento A acontecer dado que um evento B também acontece é igual a probabilidade de ambos acontecerem dividido pela probabilidade de B acontecer.

$$\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

-O teorema nos mostra que um teste para uma doença muito rara terá uma alta porcentagem de falsos positivos mesmo que a acuracia do teste seja alta.

Dado que:

+ representa um teste positivo
- representa um teste negativo

D = 0 indica sem doença

Probability of having the disease given a positive test: Pr(D=1 | +)

99% test accuracy when disease is present: Pr(+ | D=1)=0.99

99% test accuracy when disease is absent: Pr(− | D=0)=0.99

Rate of cystic fibrosis: Pr(D=1)=0.00025

$$\begin{aligned}
\Pr(D = 1 \mid +) &= \frac{\Pr(+ \mid D = 1) \cdot \Pr(D = 1)}{\Pr(+)} \\
&= \frac{\Pr(+ \mid D = 1) \cdot \Pr(D = 1)}{\Pr(+ \mid D = 1) \cdot \Pr(D = 1) + \Pr(+ \mid D = 0) \cdot \Pr(D = 0)} \\
&= \frac{0.99 \cdot \Pr(D = 1)}{0.99 \cdot \Pr(D = 1) + \Pr(+ \mid D = 0) \cdot \Pr(D = 0)} \\
&= \frac{0.99 \cdot 0.00025}{0.99 \cdot 0.00025 + \Pr(+ \mid D = 0) \cdot \Pr(D = 0)} \\
&= \frac{0.99 \cdot 0.00025}{0.99 \cdot 0.00025 + 0.01 \cdot \Pr(D = 0)} \\
&= \frac{0.99 \cdot 0.00025}{0.99 \cdot 0.00025 + 0.01 \cdot 0.99975} \\
&= 0.02
\end{aligned}$$

**CODE:**

Monte Carlo simulation

```
prev <- 0.00025    # disease prevalence
N <- 100000    # number of tests
outcome <- sample(c("Disease", "Healthy"), N, replace = TRUE, prob = c(prev, 1-prev))

N_D <- sum(outcome == "Disease")    # number with disease
N_H <- sum(outcome == "Healthy")    # number healthy

# for each person, randomly determine if test is + or -
accuracy <- 0.99
test <- vector("character", N)
test[outcome == "Disease"] <- sample(c("+", "-"), N_D, replace=TRUE, prob = c(accuracy, 1-accuracy))
test[outcome == "Healthy"] <- sample(c("-", "+"), N_H, replace=TRUE, prob = c(accuracy, 1-accuracy))

table(outcome, test)
```

**THE HIERARCHICAL MODEL**

- Hierarchical models use multiple levels of variability to model results. They are hierarchical because values in the lower levels of the model are computed using values from higher levels of the model.
- We model baseball player batting average using a hierarchical model with two levels of variability:
    - $p \sim N(\mu, \tau)$ describes player-to-player variability in natural ability to hit, which has a mean $\mu$ and standard deviation $\tau$.
    - $Y \mid p \sim N(p, \sigma)$ describes a player's observed batting average given their ability $p$, which has a mean $p$ and standard deviation $\sigma = \sqrt{p(1-p)/N}$. This represents variability due to luck.
    - In Bayesian hierarchical models, the first level is called the *prior distribution* and the second level is called the *sampling distribution*.
- The *posterior distribution* allows us to compute the probability distribution of $p$ given that we have observed data $Y$.
- By the continuous version of Bayes' rule, the *expected value of the posterior distribution* $p$ given $Y = y$ is a weighted average between the prior mean $\mu$ and the observed data $Y$:

$$\mathrm{E}(p \mid y) = B\mu + (1-B)Y \quad \text{where} \quad B = \frac{\sigma^2}{\sigma^2+\tau^2}$$

- The *standard error of the posterior distribution* $\mathrm{SE}(p \mid Y)^2$ is $\frac{1}{1/\sigma^2+1/\tau^2}$. Note that you will need to take the square root of both sides to solve for the standard error.
- This Bayesian approach is also known as *shrinking*. When $\sigma$ is large, $B$ is close to 1 and our prediction of $p$ shrinks towards the mean (\mu). When $\sigma$ is small, $B$ is close to 0 and our prediction of $p$ is more weighted towards the observed data $Y$.