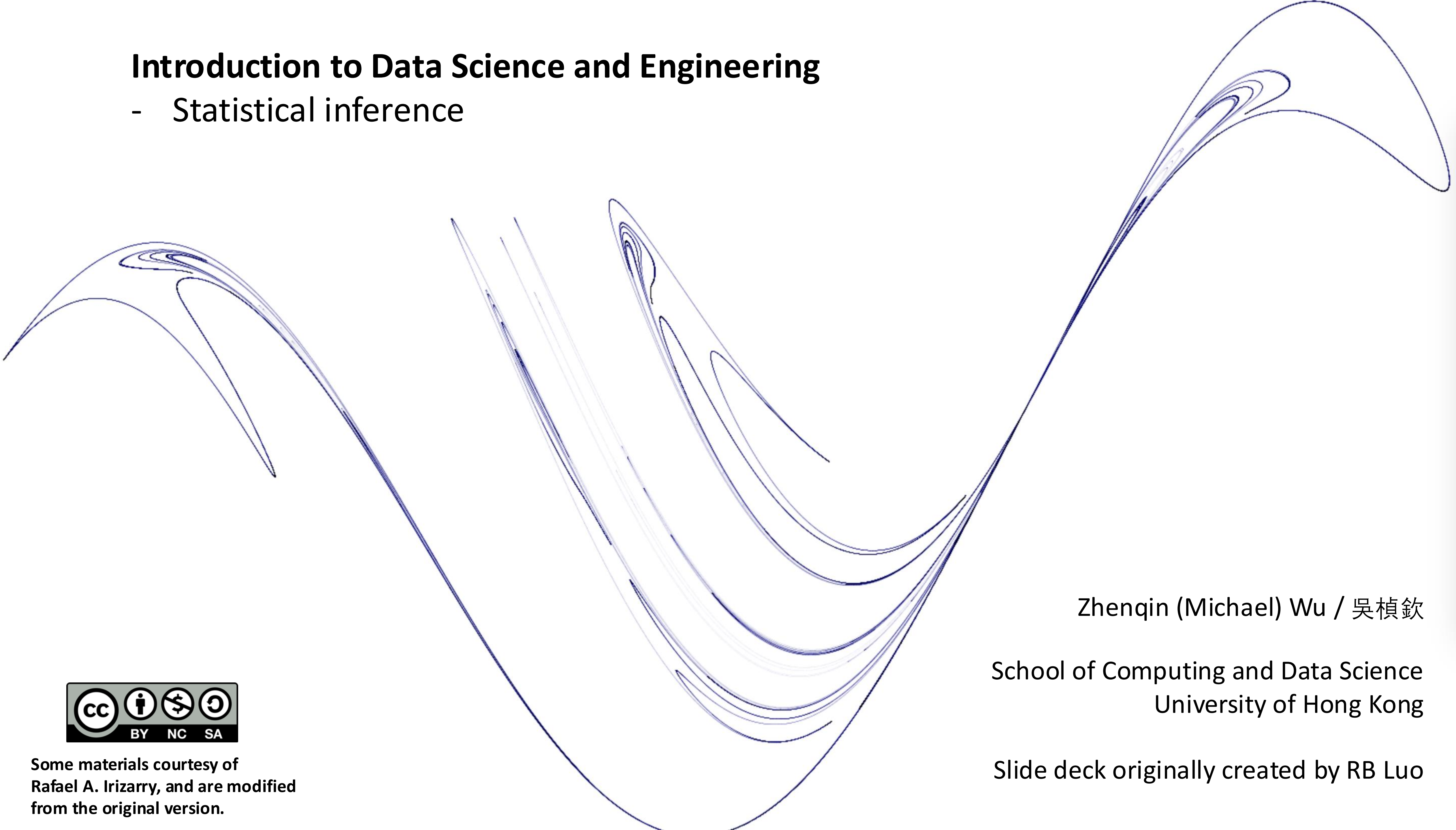
Introduction to Data Science and Engineering

- Statistical inference



Zhenqin (Michael) Wu / 吳楨欽

School of Computing and Data Science
University of Hong Kong

Slide deck originally created by RB Luo

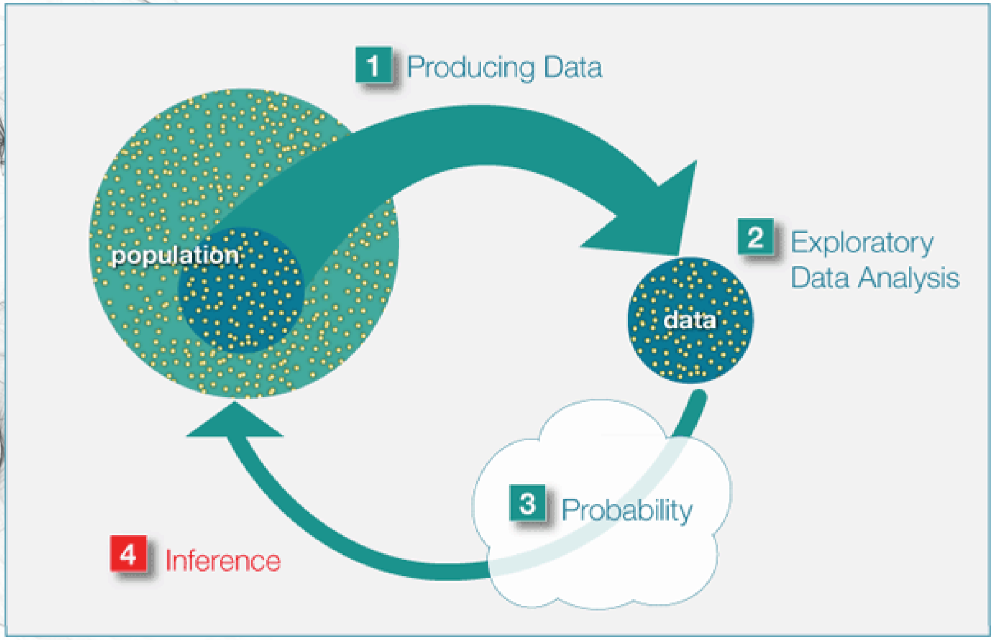
In this lecture

- What is statistical inference?
- Standard deviation, standard error, confidence interval
- Power
- p-value

In this lecture

- **What is statistical inference?**
- Standard deviation, standard error, confidence interval
- Power
- p-value

Statistical Inference



- A population have its intrinsic characteristics (i.e., parameter):
  - Average wealth
  - Opinion towards a matter
- In most cases, we can only observe a subset of the population, a.k.a. the samples.
- How can we derive conclusion on population-level characteristics based on observations from samples?

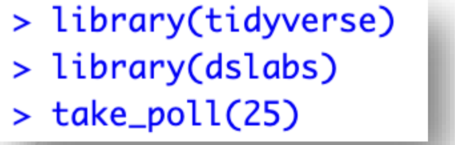
Motivating example: poll

- On a specific matter (e.g., presidential election), assume every citizen in the population has a ground truth preference.

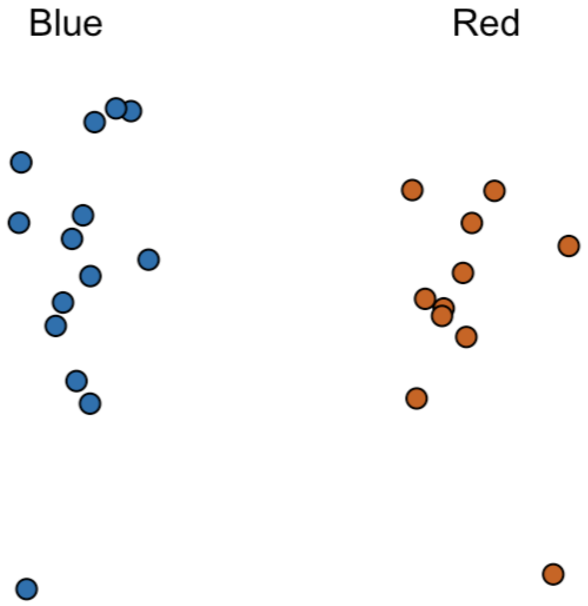
- In a binary case, denote p as the proportion of population that prefers one side (e.g., candidate A) over the other (candidate B). Statistical inference tries to find and estimate p.
  - If we can ask every single citizen, then p is automatically revealed. This is what happens in the actual election.
  - In a limited resource setting, we only have access to a subset of the population, what can we say about p?

Motivating example: poll

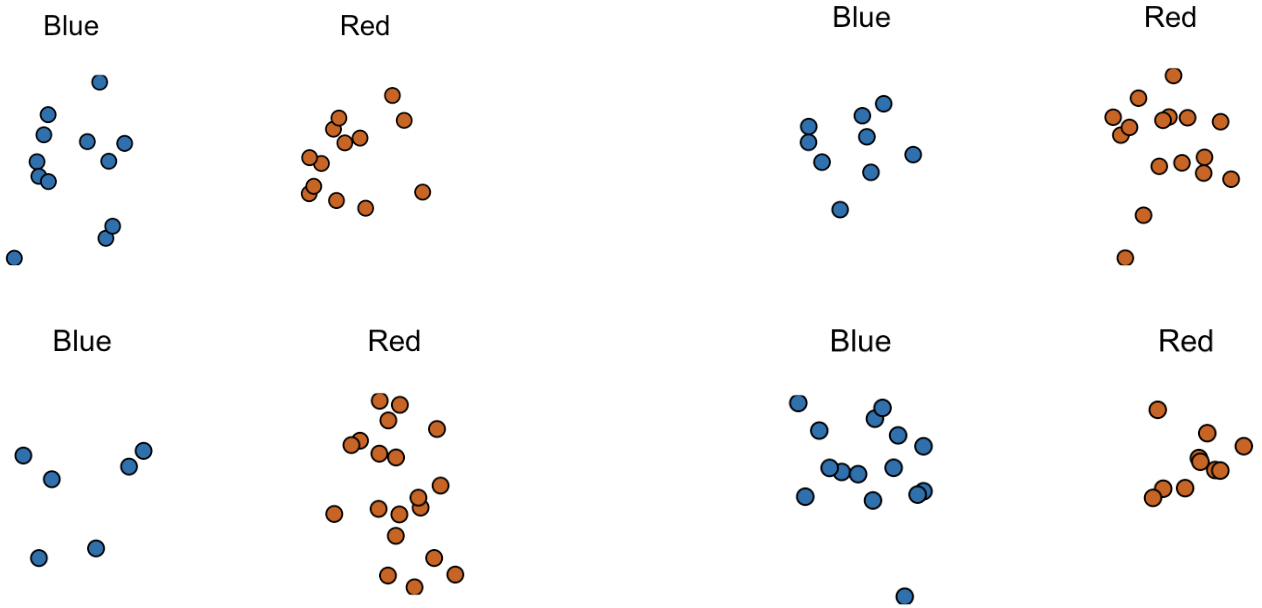
- Imagine this jar as the entire population, each bead represents an individual, and its color representing his/her preference.
- We can take some samples:



```
> library(tidyverse)
> library(dslabs)
> take_poll(25)
```



Motivating example: poll

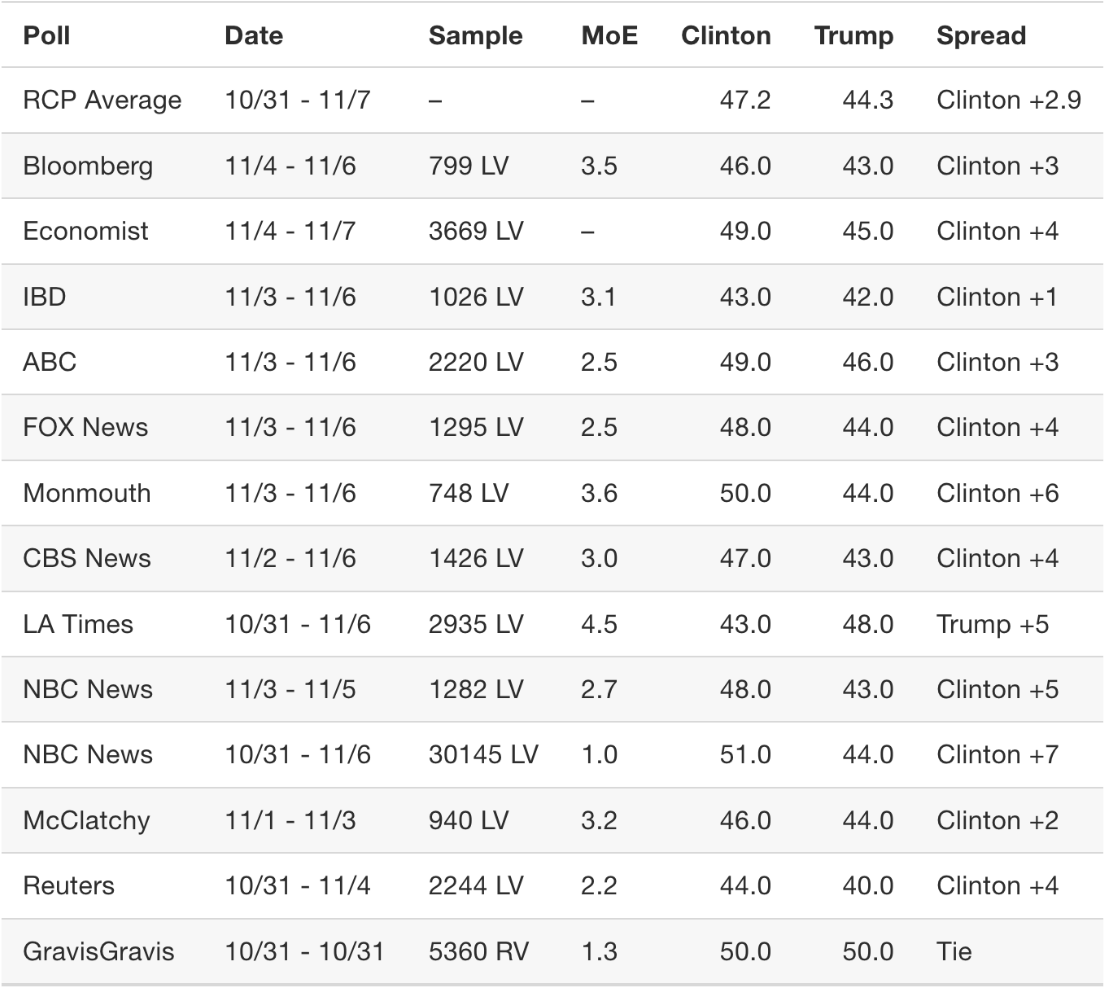


- Denote p as the proportion of blue beads in the jar
- `take_poll(25)` four times, p ranges from 0.24 to 0.60; What can we say about p?

Motivating example: poll

- Similar experiments are conducted by news organizations before presidential election.

- Real Clear Politics (RCP) organized and published poll results from different news agencies. (right: estimates of the popular vote for the 2016 U.S. presidential election)



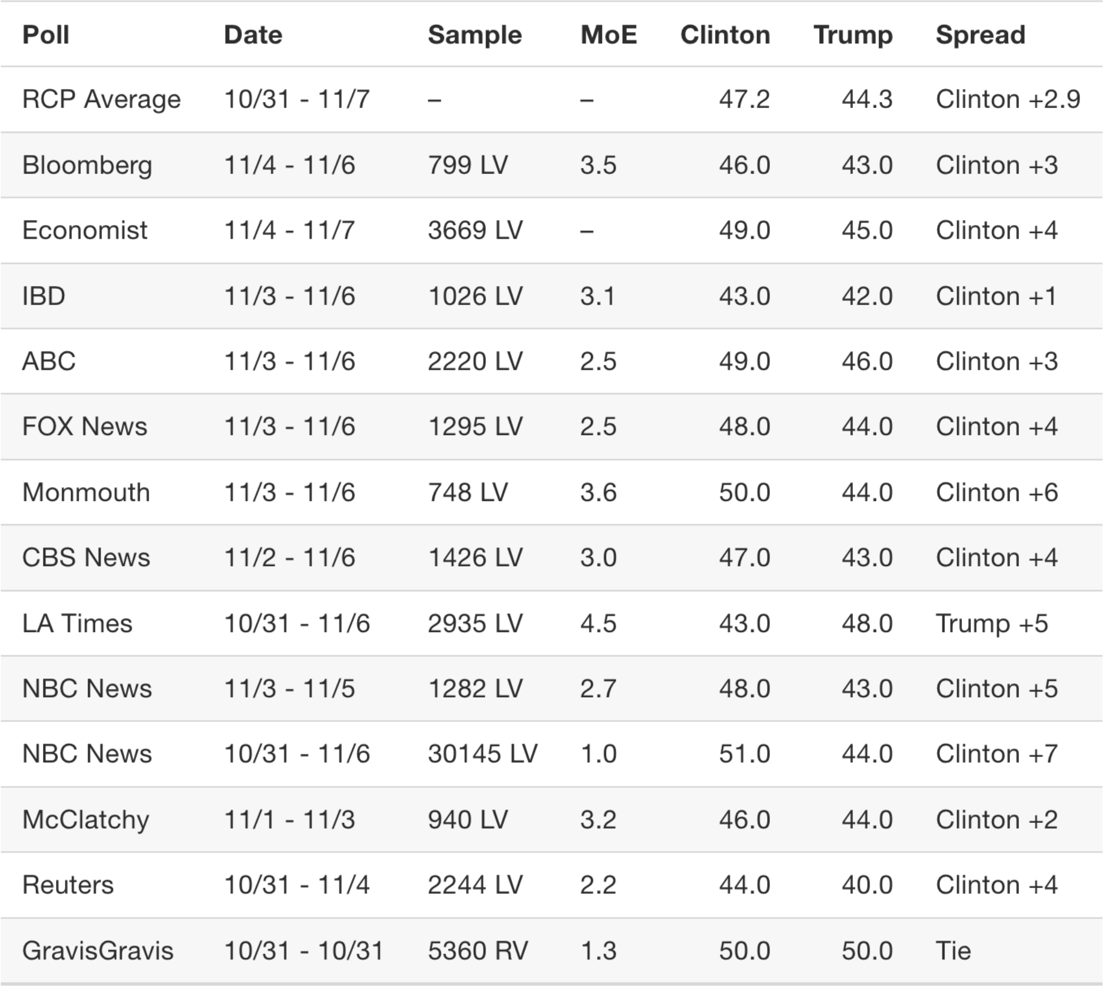
Poll	Date	Sample	MoE	Clinton	Trump	Spread
RCP Average	10/31 - 11/7	--	--	47.2	44.3	Clinton +2.9
Bloomberg	11/4 - 11/6	799 LV	3.5	46.0	43.0	Clinton +3
Economist	11/4 - 11/7	3669 LV	--	49.0	45.0	Clinton +4
IBD	11/3 - 11/6	1026 LV	3.1	43.0	42.0	Clinton +1
ABC	11/3 - 11/6	2220 LV	2.5	49.0	46.0	Clinton +3
FOX News	11/3 - 11/6	1295 LV	2.5	48.0	44.0	Clinton +4
Monmouth	11/3 - 11/6	748 LV	3.6	50.0	44.0	Clinton +6
CBS News	11/2 - 11/6	1426 LV	3.0	47.0	43.0	Clinton +4
LA Times	10/31 - 11/6	2935 LV	4.5	43.0	48.0	Trump +5
NBC News	11/3 - 11/5	1282 LV	2.7	48.0	43.0	Clinton +5
NBC News	10/31 - 11/6	30145 LV	1.0	51.0	44.0	Clinton +7
McClatchy	11/1 - 11/3	940 LV	3.2	46.0	44.0	Clinton +2
Reuters	10/31 - 11/4	2244 LV	2.2	44.0	40.0	Clinton +4
GravisGravis	10/31 - 10/31	5360 RV	1.3	50.0	50.0	Tie

RealClearPolitics - Live Opinion, News, Analysis, Video and Polls

Motivating example: poll

- Denote preference for candidate 1 (Clinton): p_1
- For candidate 2 (Trump): p_2

- Spread: $p_1 - p_2$
- MoE (Margin of Error): uncertainty about the estimation of spread, more on this later
  - In the second row, the estimation for spread is a range (-0.5%, 6.5%)



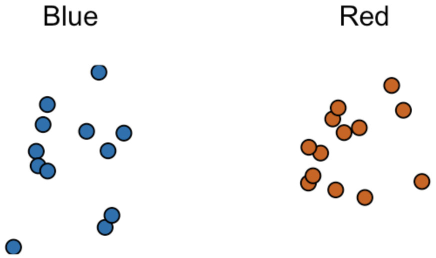
Poll	Date	Sample	MoE	Clinton	Trump	Spread
RCP Average	10/31 - 11/7	–	–	47.2	44.3	Clinton +2.9
Bloomberg	11/4 - 11/6	799 LV	3.5	46.0	43.0	Clinton +3
Economist	11/4 - 11/7	3669 LV	–	49.0	45.0	Clinton +4
IBD	11/3 - 11/6	1026 LV	3.1	43.0	42.0	Clinton +1
ABC	11/3 - 11/6	2220 LV	2.5	49.0	46.0	Clinton +3
FOX News	11/3 - 11/6	1295 LV	2.5	48.0	44.0	Clinton +4
Monmouth	11/3 - 11/6	748 LV	3.6	50.0	44.0	Clinton +6
CBS News	11/2 - 11/6	1426 LV	3.0	47.0	43.0	Clinton +4
LA Times	10/31 - 11/6	2935 LV	4.5	43.0	48.0	Trump +5
NBC News	11/3 - 11/5	1282 LV	2.7	48.0	43.0	Clinton +5
NBC News	10/31 - 11/6	30145 LV	1.0	51.0	44.0	Clinton +7
McClatchy	11/1 - 11/3	940 LV	3.2	46.0	44.0	Clinton +2
Reuters	10/31 - 11/4	2244 LV	2.2	44.0	40.0	Clinton +4
GravisGravis	10/31 - 10/31	5360 RV	1.3	50.0	50.0	Tie

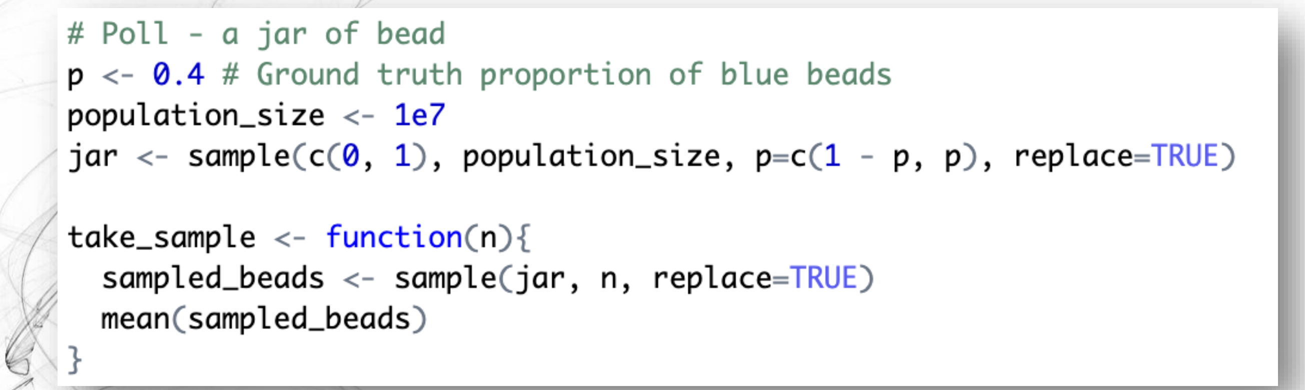
RealClearPolitics - Live Opinion, News, Analysis, Video and Polls

Motivating example: poll

- Back to the jar example:
  - Denote proportion of blue beads as p
  - Proportion of red beads is $1 - p$
  - Spread is $2p - 1$

- We did a poll on 25 beads. How can we estimate p? How certain are we?

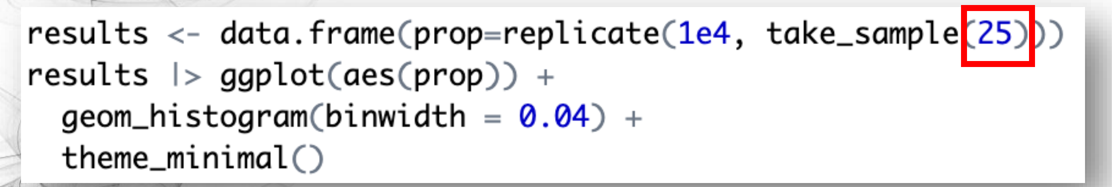




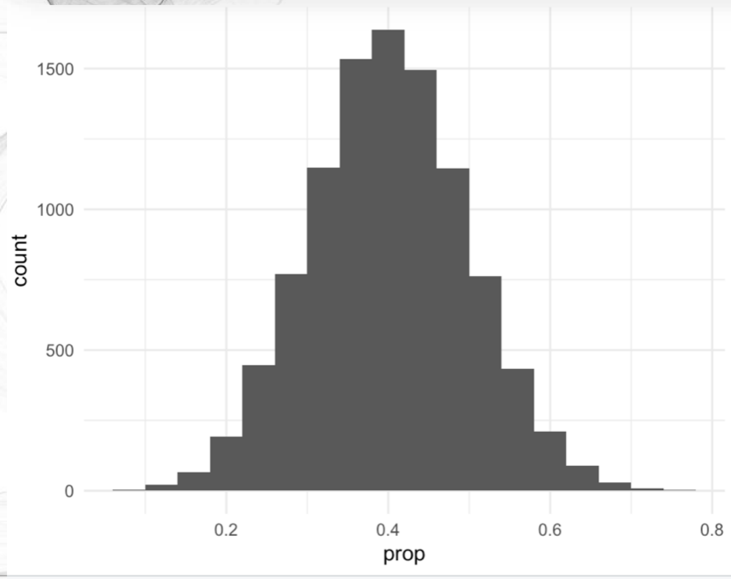
```r
# Poll - a jar of bead
p <- 0.4 # Ground truth proportion of blue beads
population_size <- 1e7
jar <- sample(c(0, 1), population_size, p=c(1 - p, p), replace=TRUE)

take_sample <- function(n){
  sampled_beads <- sample(jar, n, replace=TRUE)
  mean(sampled_beads)
}
```

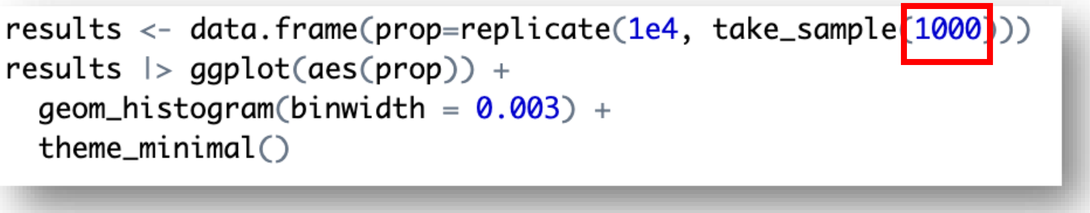
- The simplest idea is to use the proportion of blue beads in our sample as estimate for p
- When the number of picked beads is large enough, it should converge to p. (why?)



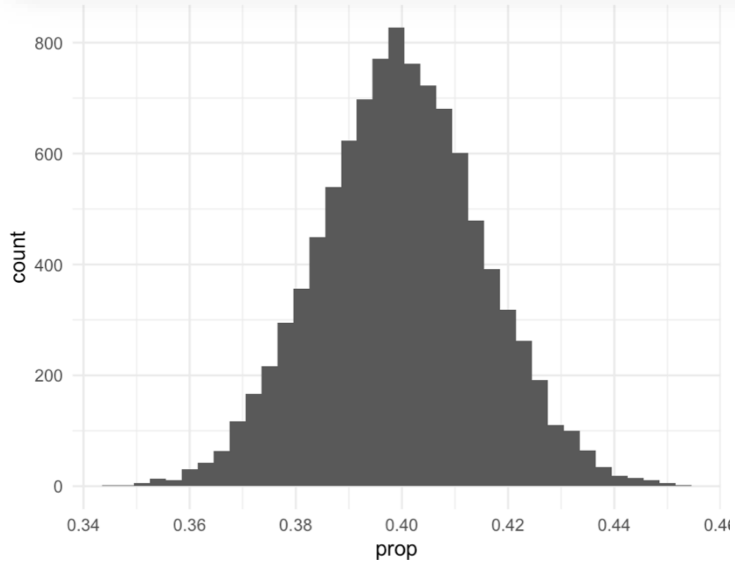
```r
results <- data.frame(prop=replicate(1e4, take_sample(25)))
results |> ggplot(aes(prop)) +
  geom_histogram(binwidth = 0.04) +
  theme_minimal()
```



Range is [0.2, 0.6]



```r
results <- data.frame(prop=replicate(1e4, take_sample(1000)))
results |> ggplot(aes(prop)) +
  geom_histogram(binwidth = 0.003) +
  theme_minimal()
```



Range is [0.37, 0.43]

Applying CLT

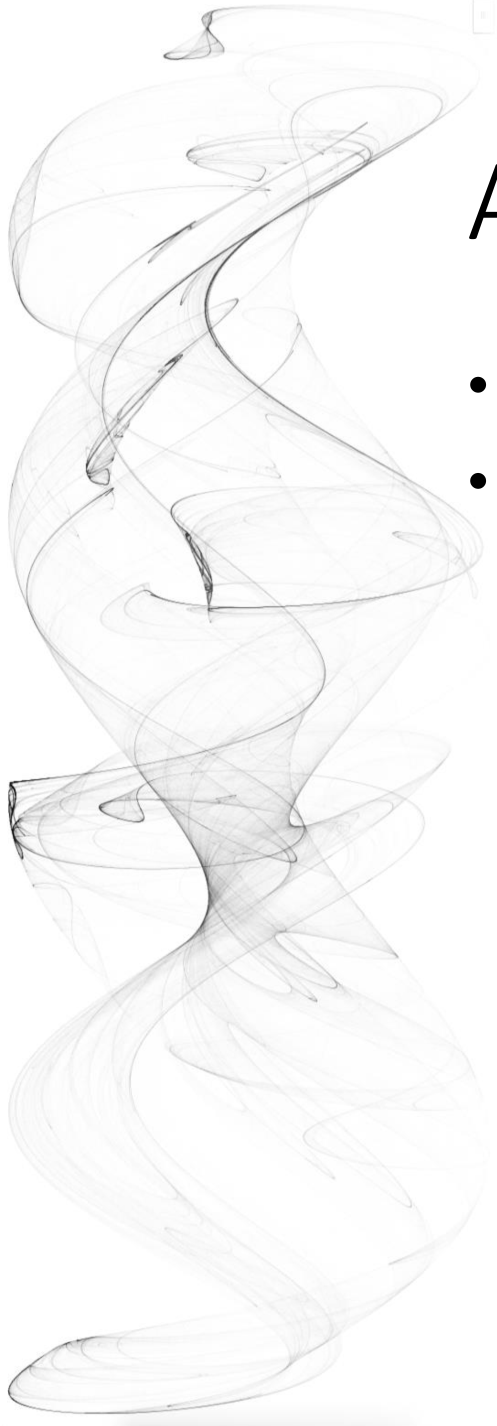
- Draw n beads $(X_1, X_2, \dots X_n)$ and calculate the proportion of blue beads is equivalent to calculating sample mean for a sample of 25 observations (0 for red, or 1 for blue). Denote this sample mean as $\bar{X}$

$$\mathrm{E}(\bar{X}) = \mathrm{E}(X_1) = p$$

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_1)}{n} = \frac{p(1-p)}{n}$$

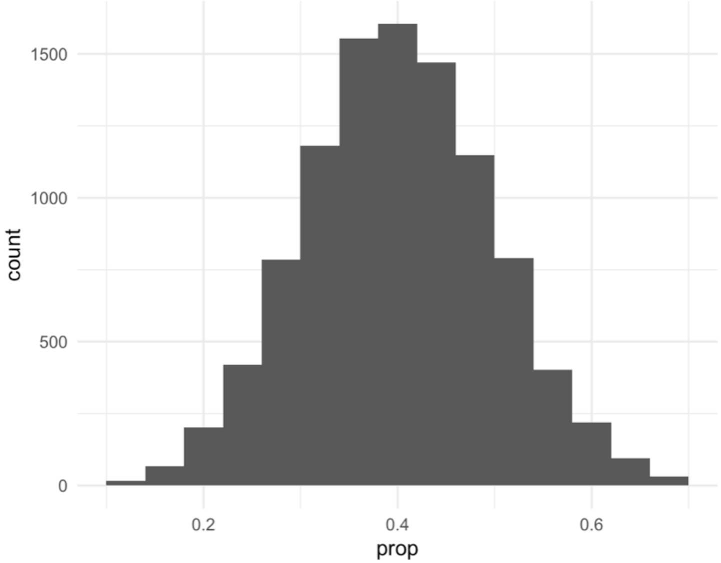
$$\mathrm{SD}(\bar{X}) = \sqrt{\frac{p(1-p)}{n}}$$

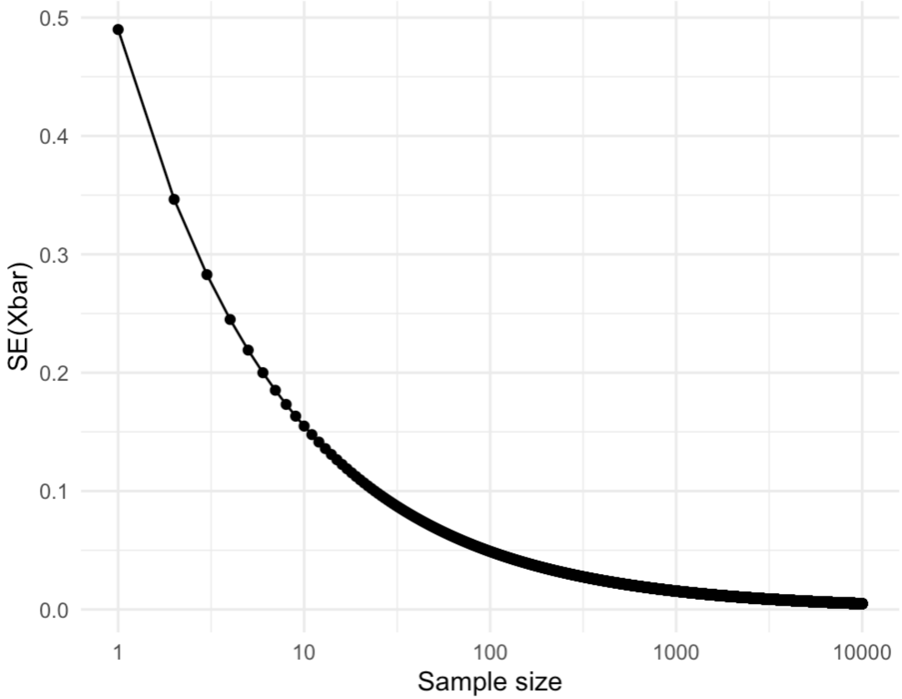
(also denoted as the standard error of our estimate)

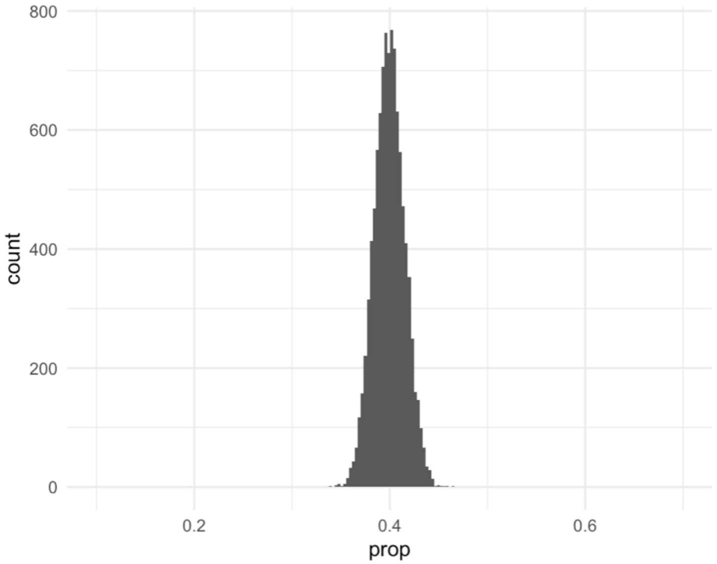


Applying CLT

- (right top) 25 samples: $\mathrm{SE}(\bar{X}) = 0.097$
- (right bottom) 1000 samples: $\mathrm{SE}(\bar{X}) = 0.015$





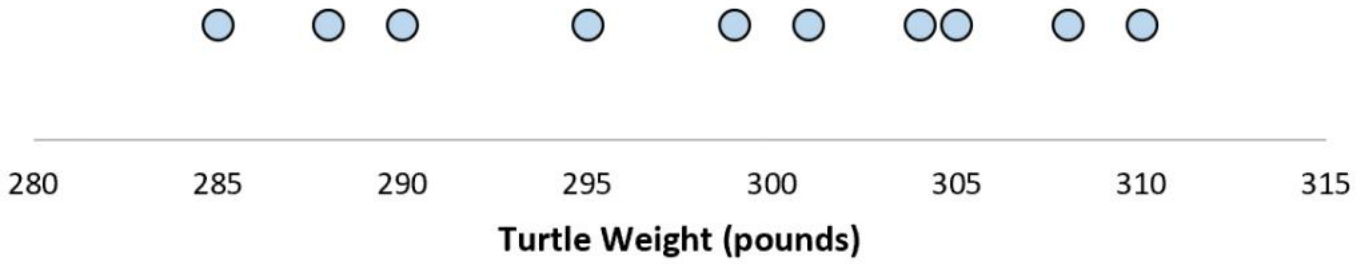


In this lecture

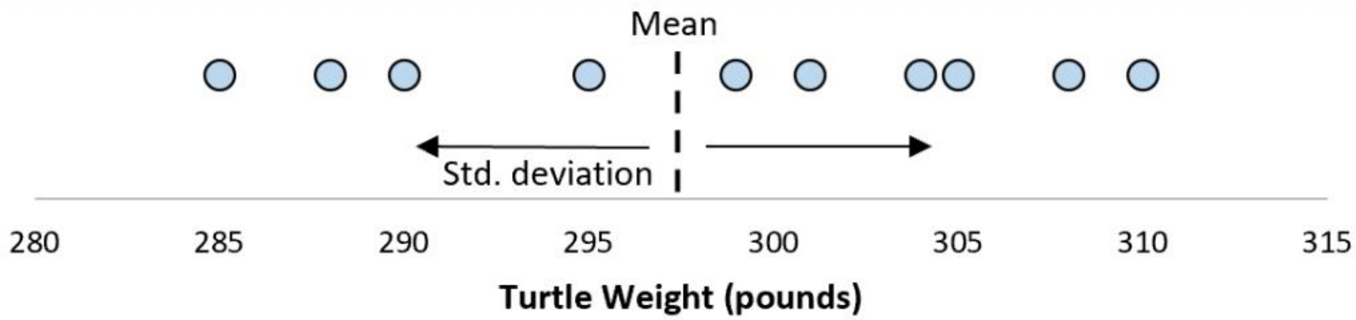
- What is statistical inference?
- **Standard deviation, standard error, confidence interval**
- Power
- p-value

Standard Deviation vs. Standard Error

Suppose we measure the weights of 10 different turtles

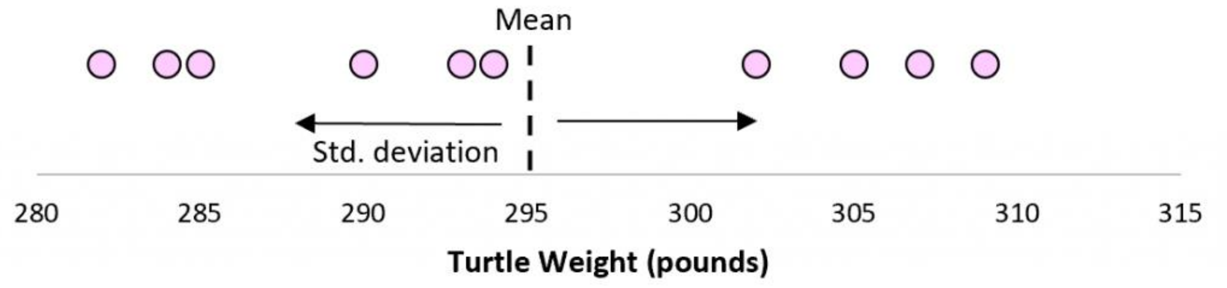


We can calculate the sample mean and the sample standard deviation

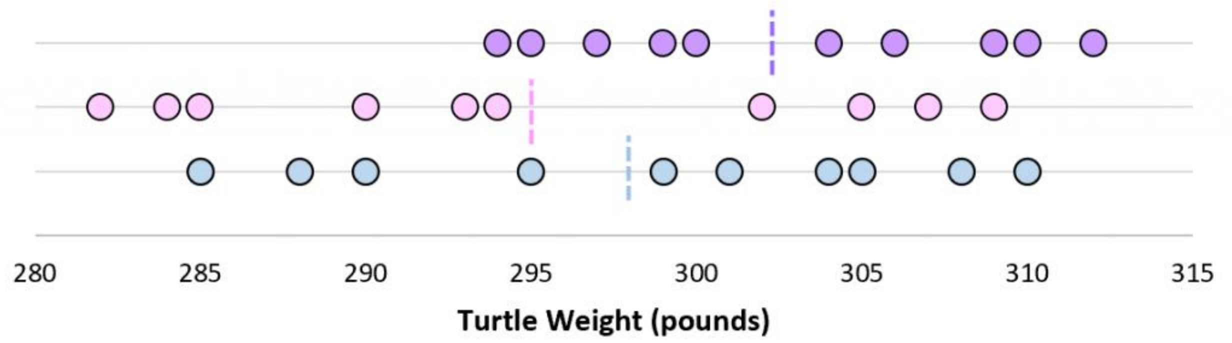


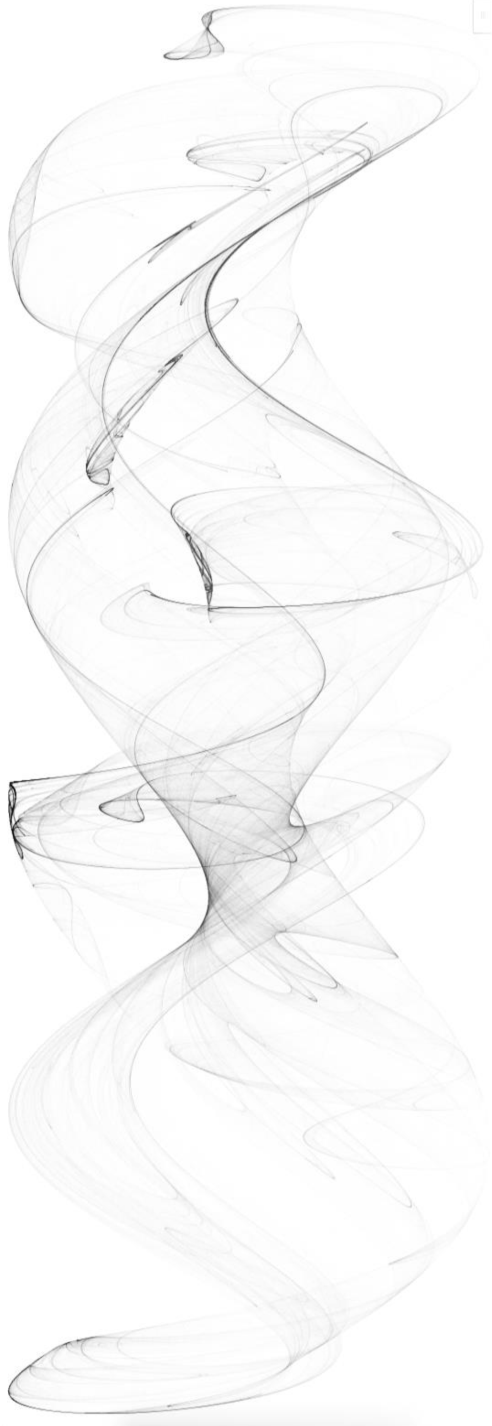
Standard Deviation vs. Standard Error

We can draw another sample of 10 turtles



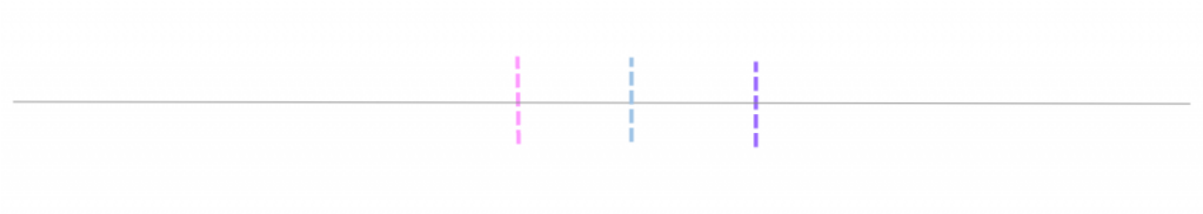
Or even more; each sample comes with a sample mean and a sample SD



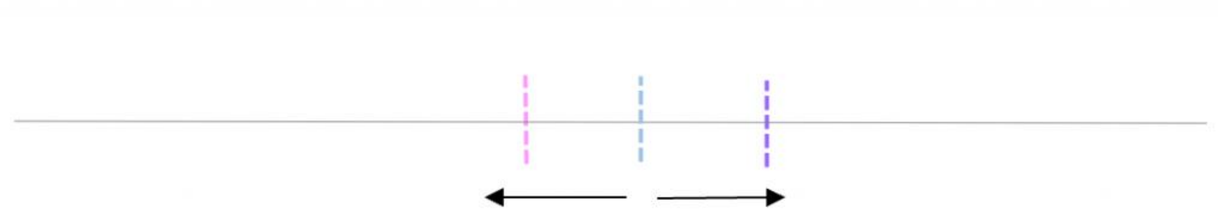


Standard Deviation vs. Standard Error

Now if we only look at the sample means:

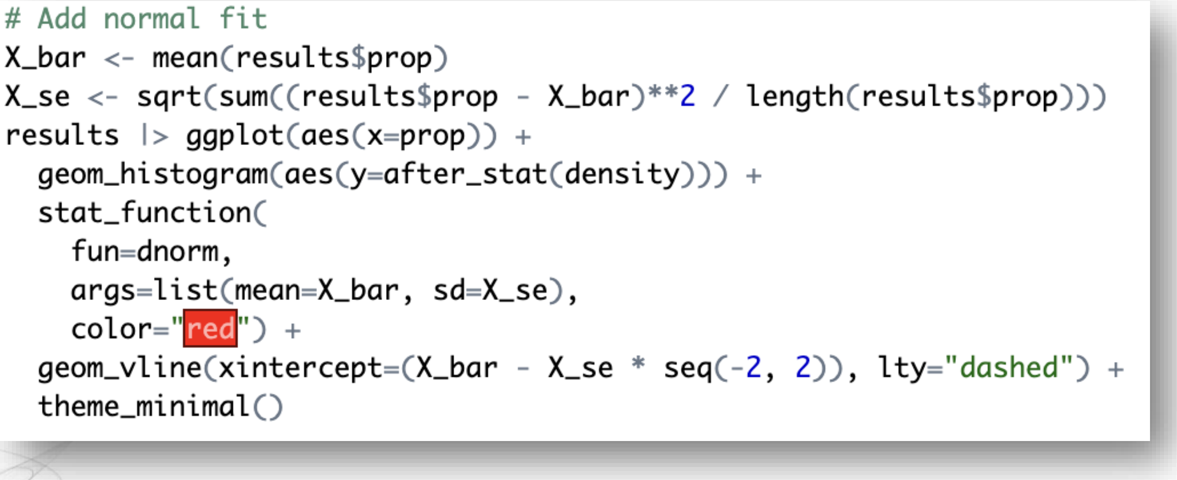


Their spread/uncertainty, calculated by the standard deviation of sample means, is known as the standard error:

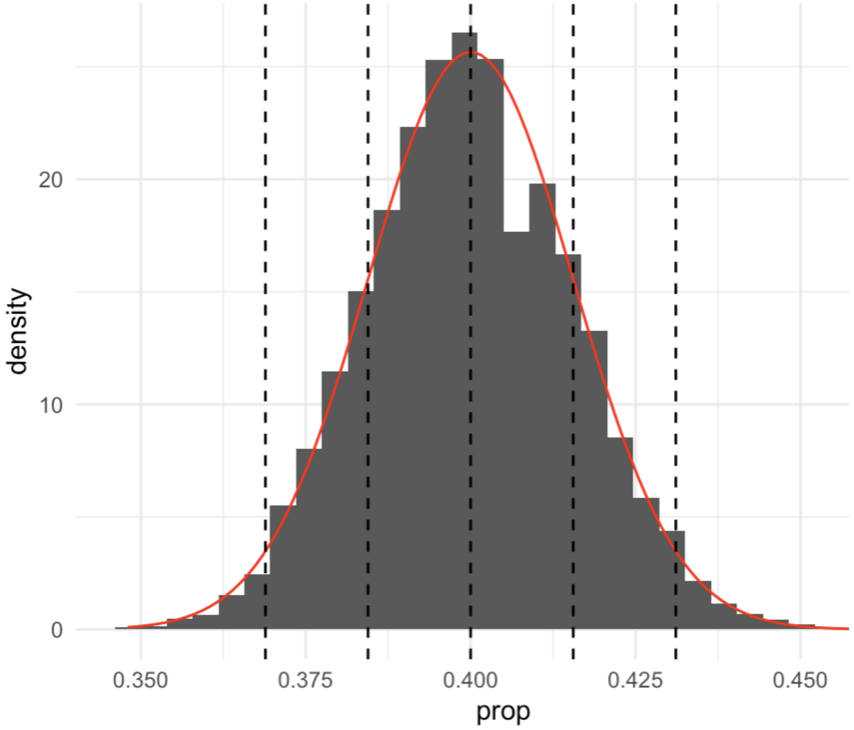


Sample mean’s standard error is directly relevant to the standard deviation of individual observations: $s/\sqrt{n}$

Standard Error => Confidence Interval



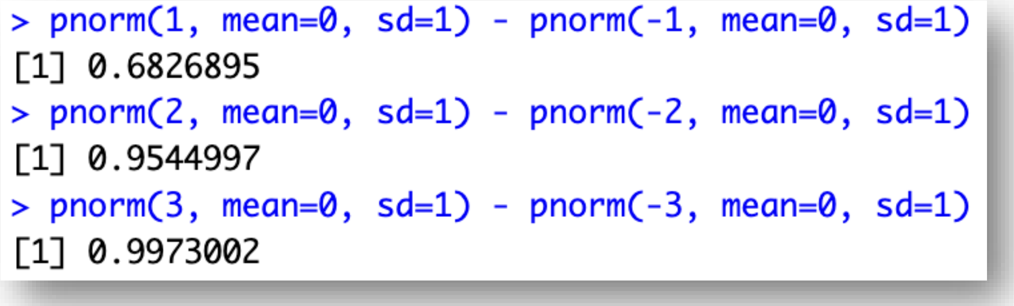
```
# Add normal fit
X_bar <- mean(results$prop)
X_se <- sqrt(sum((results$prop - X_bar)**2 / length(results$prop)))
results |> ggplot(aes(x=prop)) +
  geom_histogram(aes(y=after_stat(density))) +
  stat_function(
    fun=dnorm,
    args=list(mean=X_bar, sd=X_se),
    color="red") +
  geom_vline(xintercept=(X_bar - X_se * seq(-2, 2)), lty="dashed") +
  theme_minimal()
```



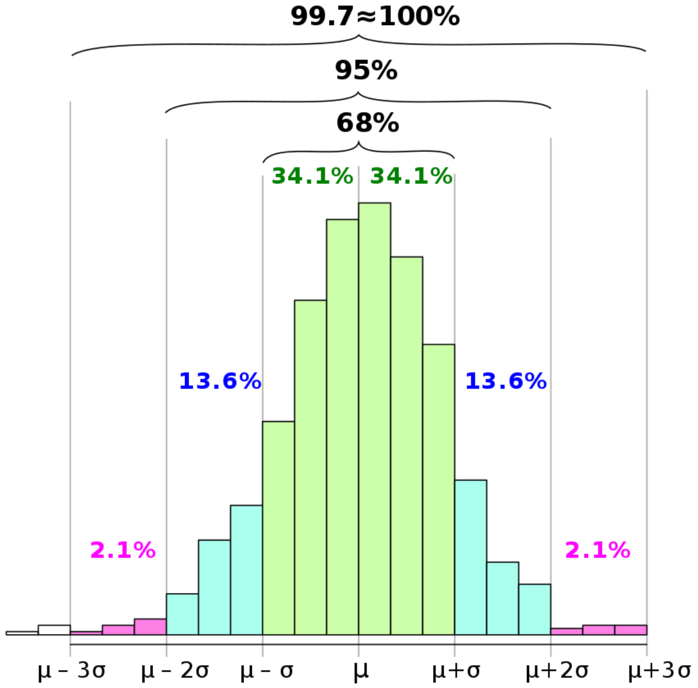
- Dashed lines are sample mean offset by (-2, -1, 0, 1, 2) times SE
- MoE (Margin of Error) in poll is defined as two times SE.

Standard Error => Confidence Interval

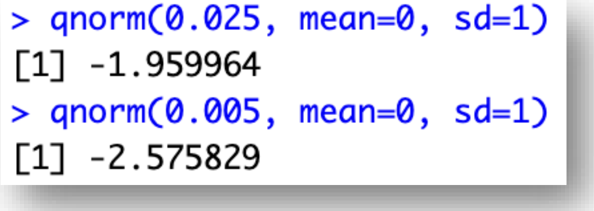
Under CLT, sample mean follows a normal distribution:



```
> pnorm(1, mean=0, sd=1) - pnorm(-1, mean=0, sd=1)
[1] 0.6826895
> pnorm(2, mean=0, sd=1) - pnorm(-2, mean=0, sd=1)
[1] 0.9544997
> pnorm(3, mean=0, sd=1) - pnorm(-3, mean=0, sd=1)
[1] 0.9973002
```



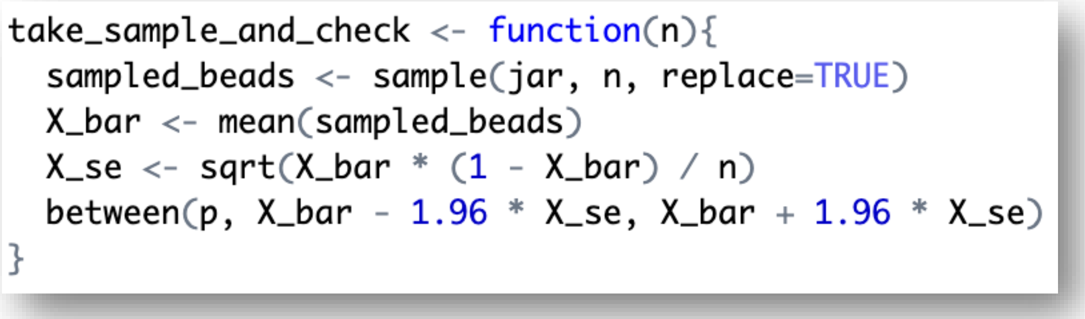
To make sure that your estimate is correct in 95% of the case, you will need a range of ±1.96 standard error (confidence interval)



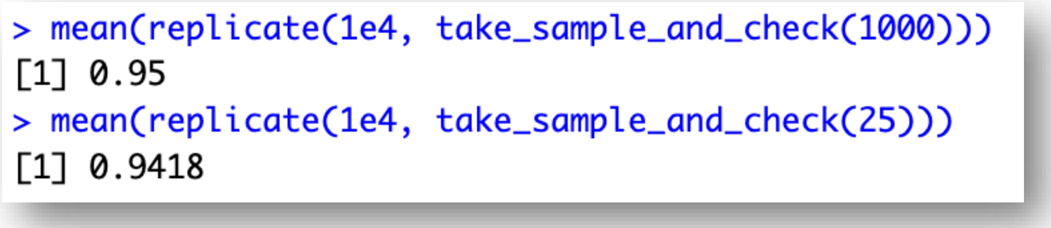
```
> qnorm(0.025, mean=0, sd=1)
[1] -1.959964
> qnorm(0.005, mean=0, sd=1)
[1] -2.575829
```

Standard Error => Confidence Interval

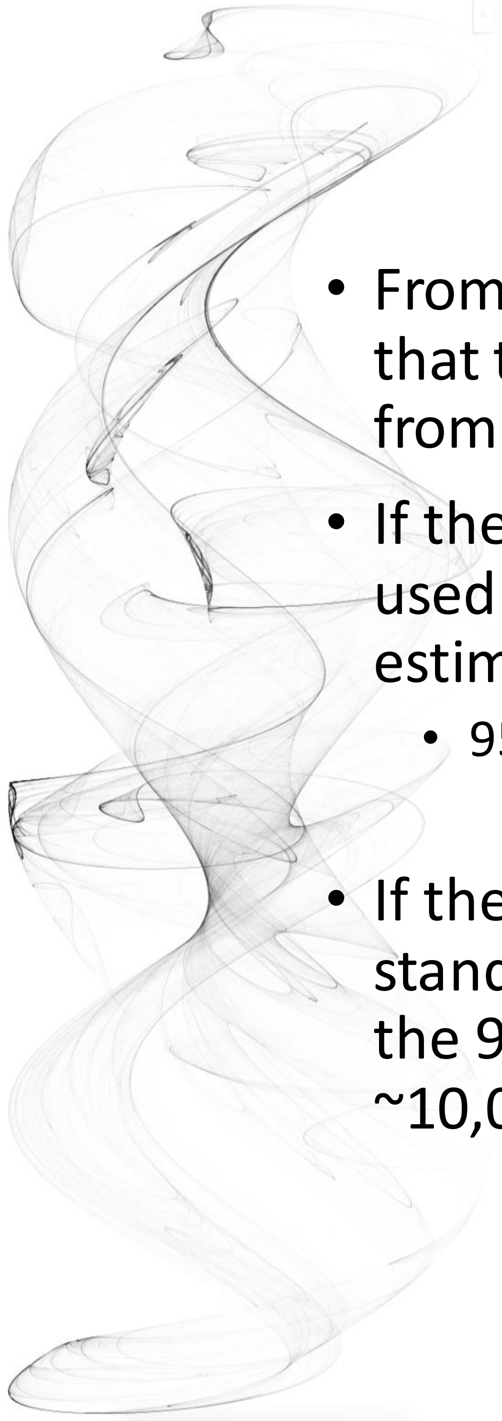
To make sure that your estimate is correct in 95% of the case, you will need a range of ±1.96 standard error



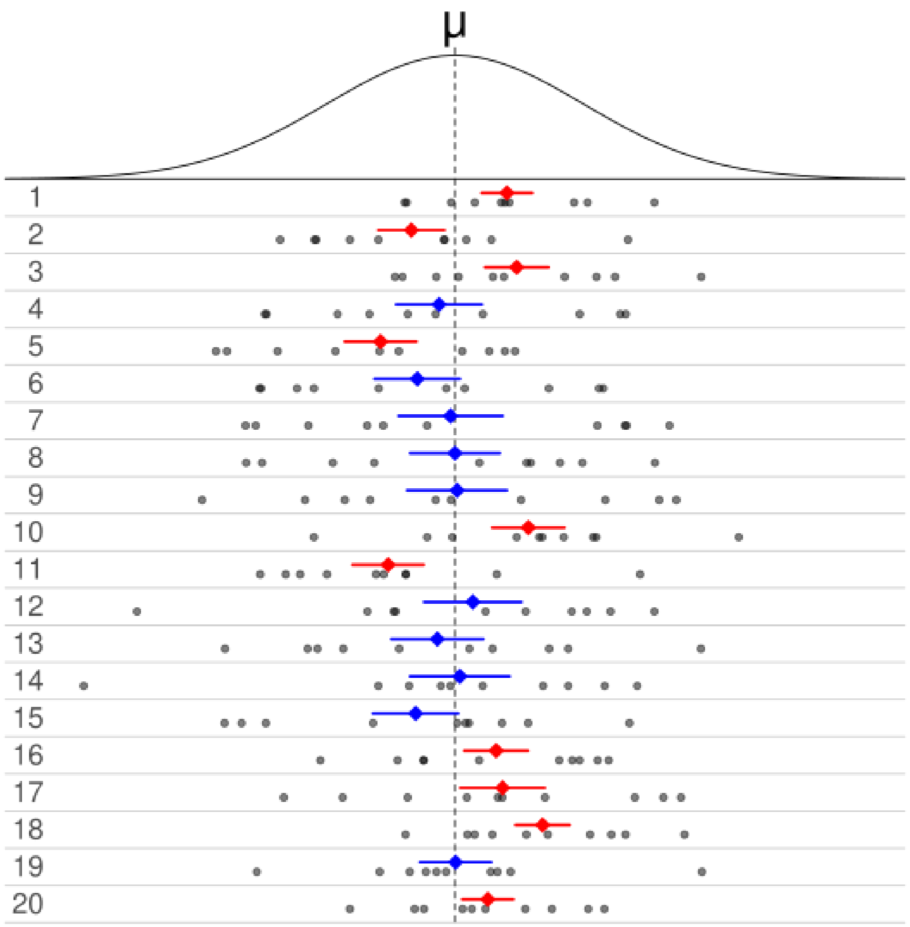
```r
take_sample_and_check <- function(n){
  sampled_beads <- sample(jar, n, replace=TRUE)
  X_bar <- mean(sampled_beads)
  X_se <- sqrt(X_bar * (1 - X_bar) / n)
  between(p, X_bar - 1.96 * X_se, X_bar + 1.96 * X_se)
}
```



```
> mean(replicate(1e4, take_sample_and_check(1000)))
[1] 0.95
> mean(replicate(1e4, take_sample_and_check(25)))
[1] 0.9418
```



- From the Real Clear Politics table, we learn that the sample sizes in opinion polls range from 500-3,500 people.
- If the observed sample mean is 0.51, and we used 1000 samples. Standard error of our estimate is ~0.0158:
    - 95% Confidence interval: 0.4784 - 0.5416
- If the underlying p = 0.51, and we need a standard error < 0.005 (so that 0.5 is not in the 95% CI). We need a sample size of ~10,000 people.



In this lecture

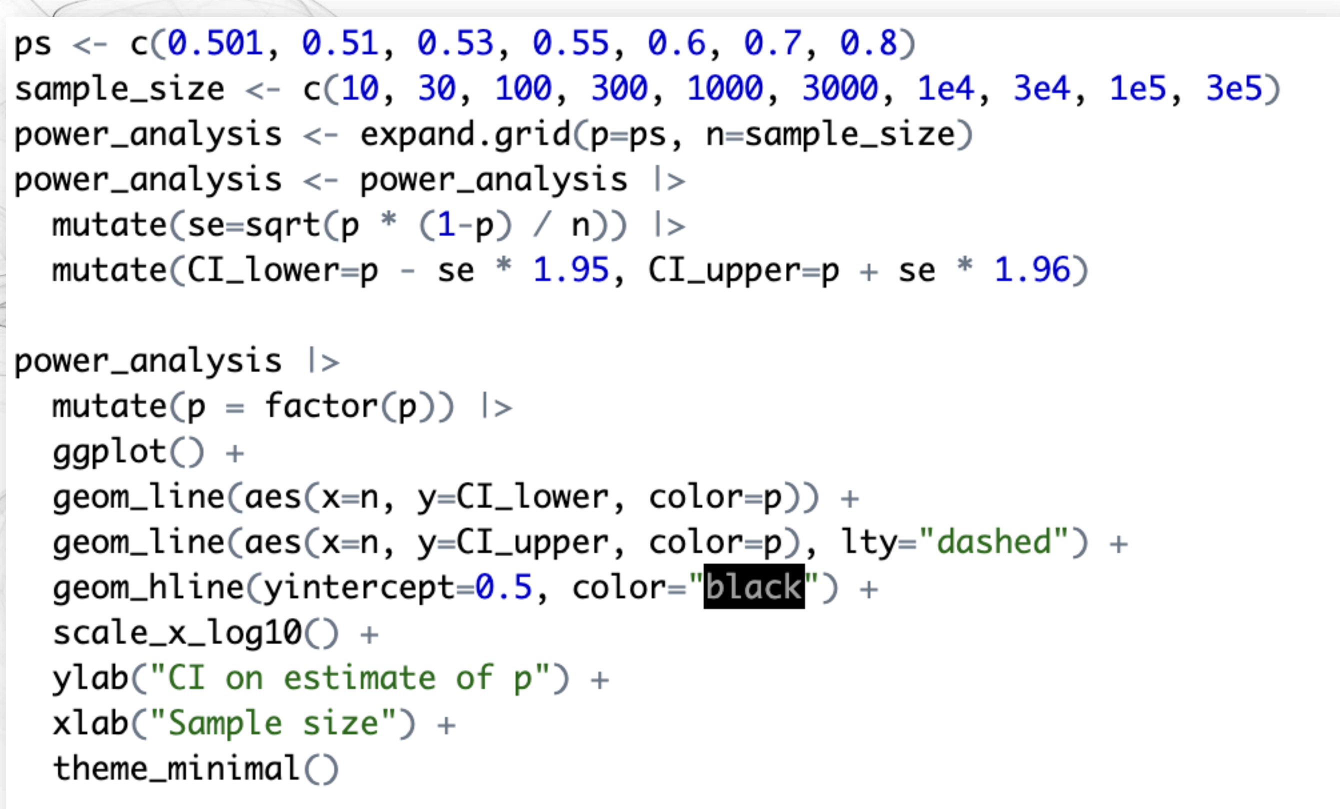
- What is statistical inference?
- Standard deviation, standard error, confidence interval
- **Power**
- p-value

Power

- In a poll with 1000 samples, the 95% Confidence interval is 0.4784 - 0.5416, which includes 0.5.
- In other words, the estimate of spread cannot rule out 0. We cannot be sure (with 95% confidence) who wins.
- But the election is not going to be a draw. It suggests that given the ground truth p (which could be close to 0.51), our sample size is too small to determine the winner. This is often called a lack of power.

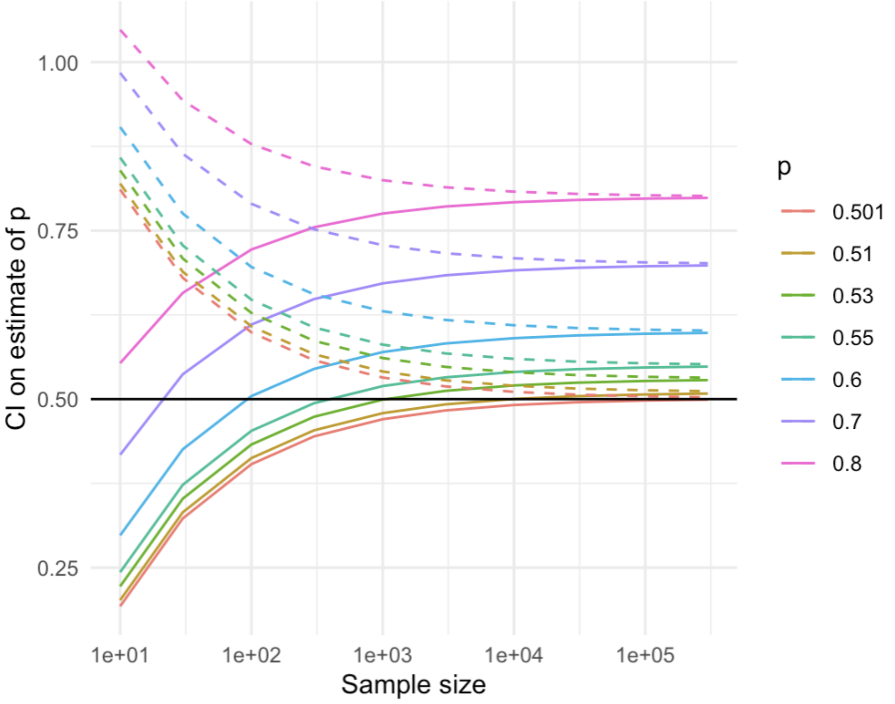
Power

- **Power analysis:** how many samples do we need to derive meaningful results (e.g., rule out 0.5 in the estimate for p)



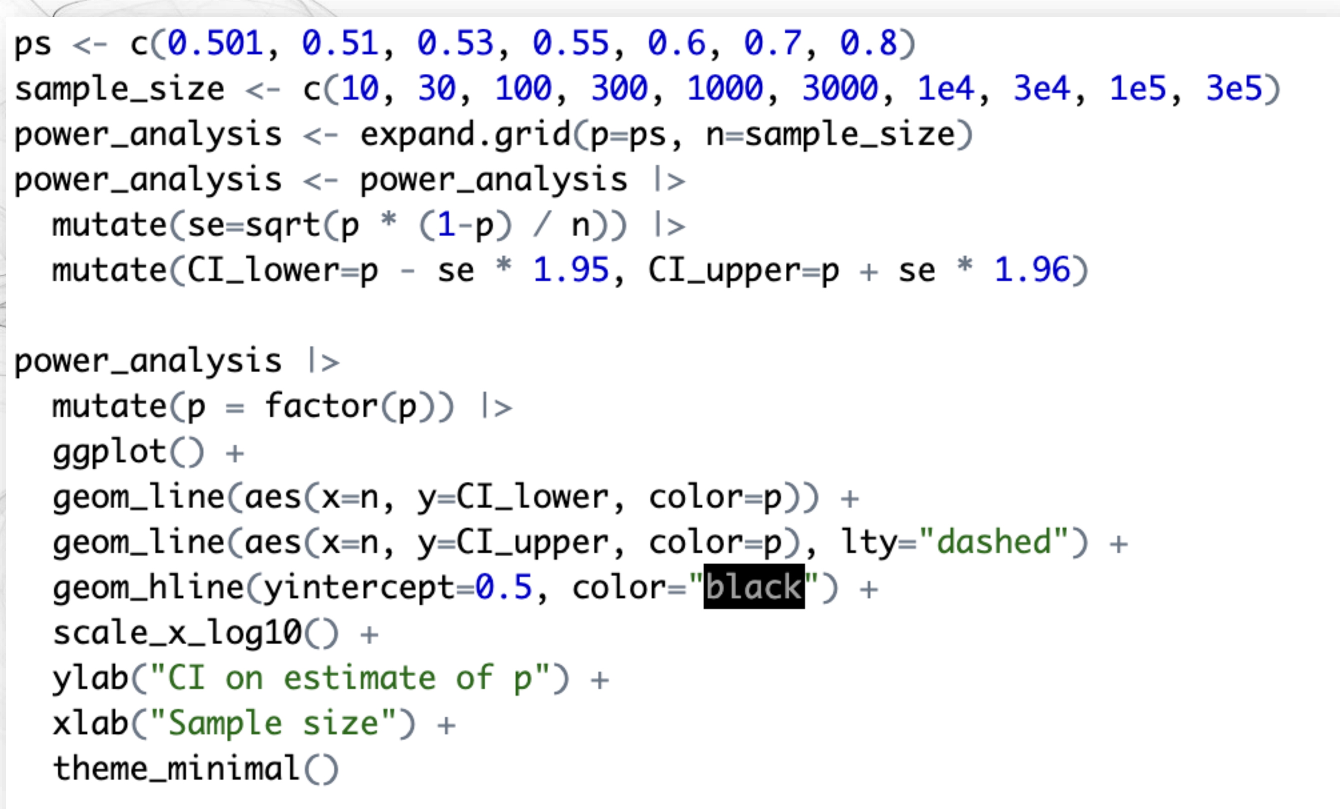
```r
ps <- c(0.501, 0.51, 0.53, 0.55, 0.6, 0.7, 0.8)
sample_size <- c(10, 30, 100, 300, 1000, 3000, 1e4, 3e4, 1e5, 3e5)
power_analysis <- expand.grid(p=ps, n=sample_size)
power_analysis <- power_analysis |>
  mutate(se=sqrt(p * (1-p) / n)) |>
  mutate(CI_lower=p - se * 1.95, CI_upper=p + se * 1.96)

power_analysis |>
  mutate(p = factor(p)) |>
  ggplot() +
  geom_line(aes(x=n, y=CI_lower, color=p)) +
  geom_line(aes(x=n, y=CI_upper, color=p), lty="dashed") +
  geom_hline(yintercept=0.5, color="black") +
  scale_x_log10() +
  ylab("CI on estimate of p") +
  xlab("Sample size") +
  theme_minimal()
```



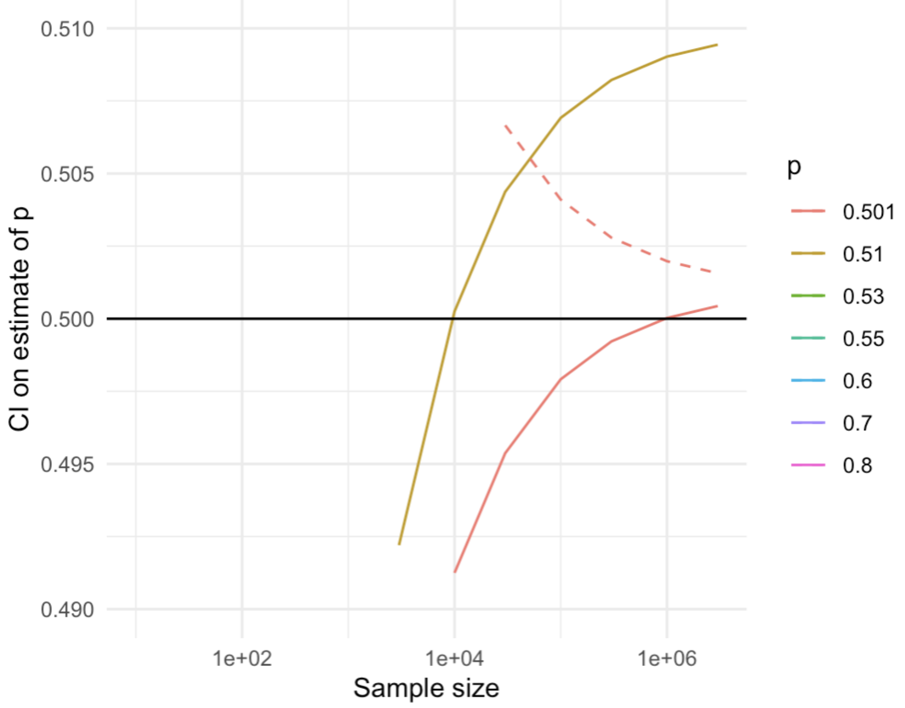
Power

- **Power analysis:** how many samples do we need to derive meaningful results (e.g., rule out 0.5 in the estimate for p)



```
ps <- c(0.501, 0.51, 0.53, 0.55, 0.6, 0.7, 0.8)
sample_size <- c(10, 30, 100, 300, 1000, 3000, 1e4, 3e4, 1e5, 3e5)
power_analysis <- expand.grid(p=ps, n=sample_size)
power_analysis <- power_analysis |>
  mutate(se=sqrt(p * (1-p) / n)) |>
  mutate(CI_lower=p - se * 1.95, CI_upper=p + se * 1.96)

power_analysis |>
  mutate(p = factor(p)) |>
  ggplot() +
  geom_line(aes(x=n, y=CI_lower, color=p)) +
  geom_line(aes(x=n, y=CI_upper, color=p), lty="dashed") +
  geom_hline(yintercept=0.5, color="black") +
  scale_x_log10() +
  ylab("CI on estimate of p") +
  xlab("Sample size") +
  theme_minimal()
```



In this lecture

- What is statistical inference?
- Standard deviation, standard error, confidence interval
- Power
- **p-value**

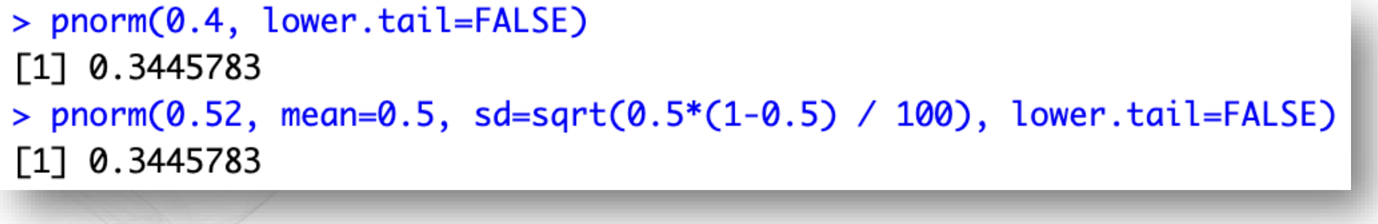
p-value

- p-values are another way of quantifying uncertainty
- In the jar of bead example: we don’t want an accurate estimate of the proportion; we just want to know: are there more blue beads or red beads?
  - Say we randomly took 100 beads and saw 52 blue beads. Since 0.52 > 0.5, are there more blue beads? How certain are we?

p-value

- H_0 (Null hypothesis): blue = red
  - Hypothesis testing is asking: would this null hypothesis be true? How likely would it be true?
  - Let’s assume that the null is true and calculate how likely that we acquire such an observation: mean of 100 samples is 0.52.

$$z = \frac{\bar{X} - 0.5}{SE(\bar{X})} = \frac{0.02}{\sqrt{\frac{0.5 \ * (1 - 0.5)}{100}}} = 0.4$$



```
> pnorm(0.4, lower.tail=FALSE)
[1] 0.3445783
> pnorm(0.52, mean=0.5, sd=sqrt(0.5*(1-0.5) / 100), lower.tail=FALSE)
[1] 0.3445783
```

This is a pretty high chance (34%)

- p ≤ 0.05, unlikely
- p ≤ 0.01, very unlikely

Know These Problems with P Values | by Andrew Plummer | Towards Data Science

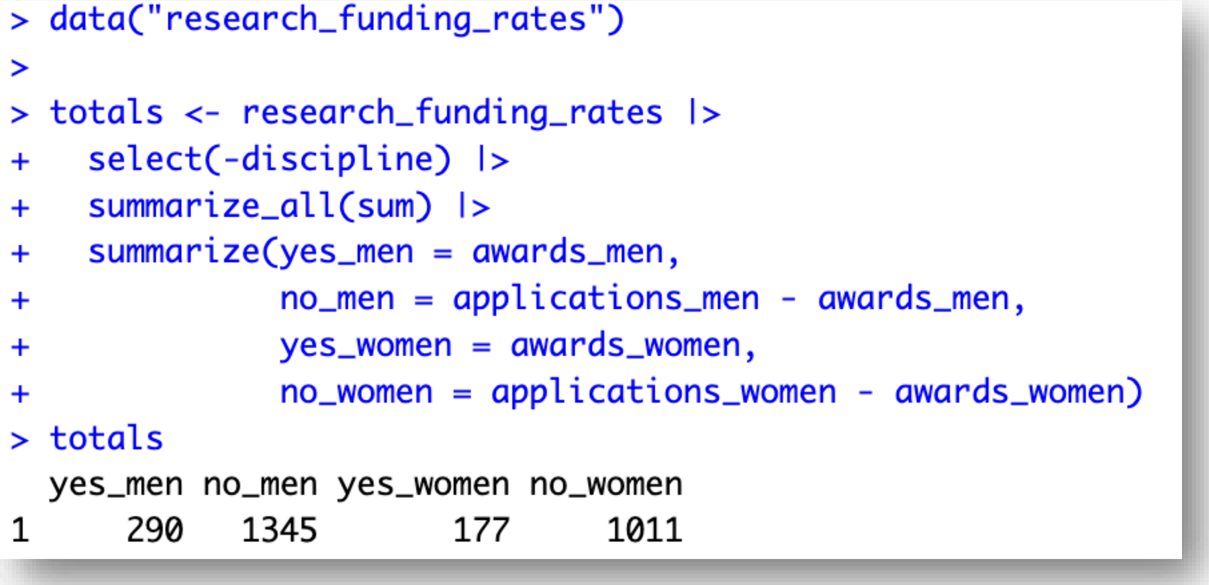
p-value

$$z = \frac{\bar{X} - 0.5}{SE(\bar{X})} = \frac{0.02}{\sqrt{\frac{0.5 \ * (1 - 0.5)}{100}}} = 0.4$$

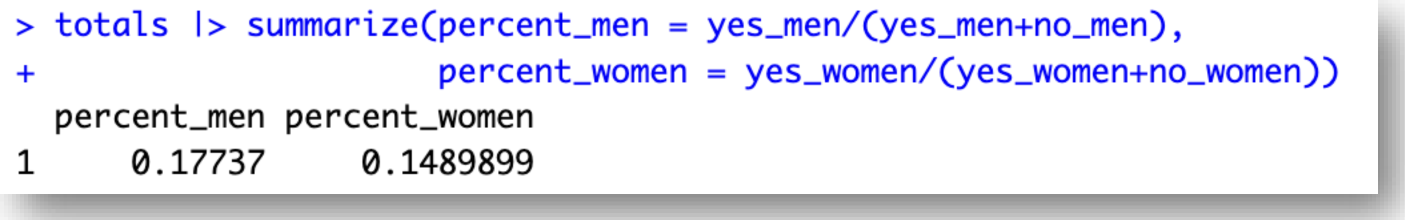
- Key assumption: in this calculation, we assumed that the observed sample mean should follow a normal distribution (CLT).
- Z-score here, therefore, should follow a standard normal.
- When calculating p-values, we always need to assume an underlying distribution (or approximate distribution) for the quantity that is being tested.

p-value

A 2014 PNAS paper analyzed success rates from funding agencies in the Netherlands and concluded that their results reveal gender bias favoring male applicants over female applicants:



```
> data("research_funding_rates")
>
> totals <- research_funding_rates |>
+   select(-discipline) |>
+   summarize_all(sum) |>
+   summarize(yes_men = awards_men,
+             no_men = applications_men - awards_men,
+             yes_women = awards_women,
+             no_women = applications_women - awards_women)
> totals
  yes_men no_men yes_women no_women
1     290   1345       177     1011
```



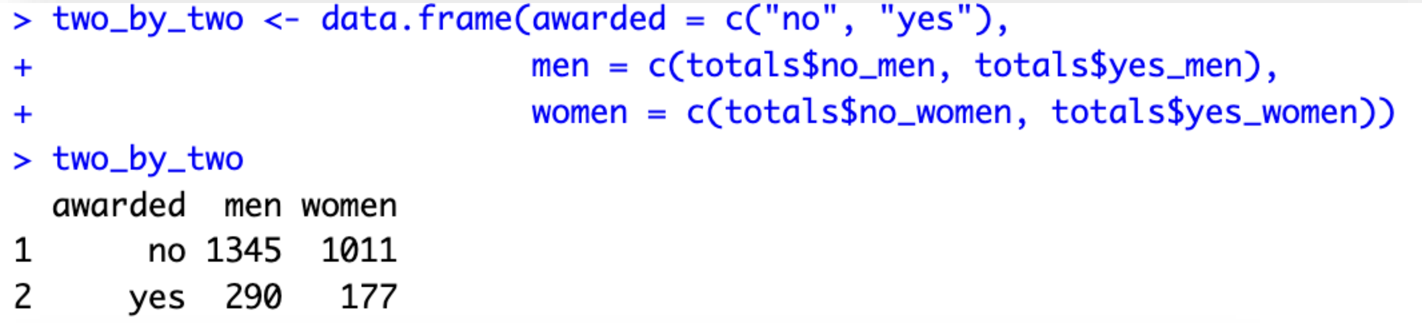
```
> totals |> summarize(percent_men = yes_men/(yes_men+no_men),
+                     percent_women = yes_women/(yes_women+no_women))
  percent_men percent_women
1     0.17737     0.1489899
```

0.177 > 0.149

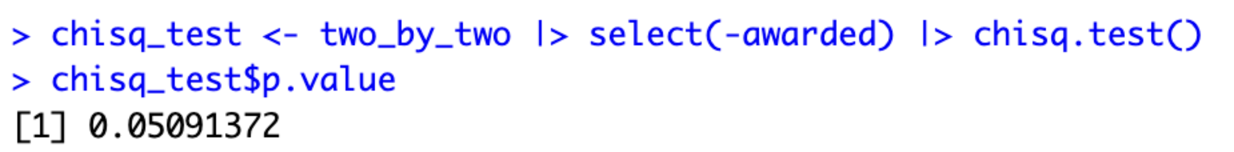
But could this appear by chance?

p-value

- Event 1: Male applicant
- Event 2: application approved
- H_0 (Null hypothesis): the two events are independent



```
> two_by_two <- data.frame(awarded = c("no", "yes"),
+                          men = c(totals$no_men, totals$yes_men),
+                          women = c(totals$no_women, totals$yes_women))
> two_by_two
  awarded  men women
1      no 1345  1011
2     yes  290   177
```



```
> chisq_test <- two_by_two |> select(-awarded) |> chisq.test()
> chisq_test$p.value
[1] 0.05091372
```

- $p \leq 0.05$, unlikely
- $p \leq 0.01$, very unlikely

- What is the quantity being tested? What is its underlying distribution?

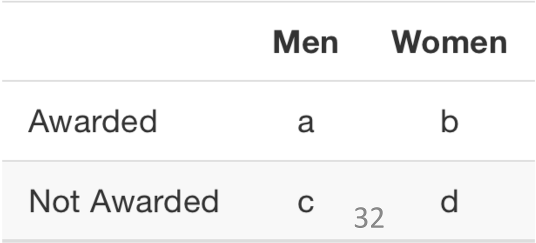
Association tests

Two major ways to verify if two categorical variables are independent:

- Testing the two-by-two table
  - Fisher’s exact test (hypergeometric distribution)
  - **Chi-square test:**
    - Defined a quantity that measures how much the observed 2-by-2 table deviates from a perfectly independent table, which follows χ2 distribution
- Testing the odds ratio:

$$\mathrm{OR} = \frac{ad}{bc} = 0.812$$

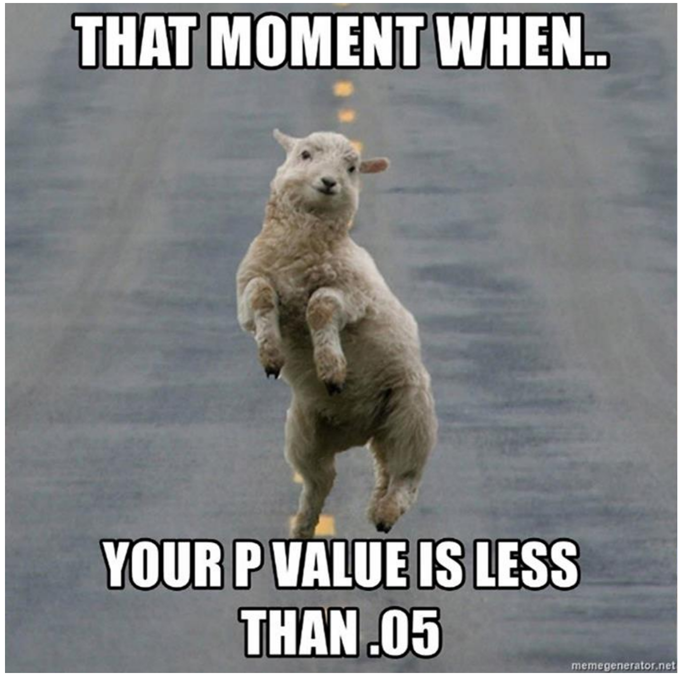
  - When a, b, c, d are large, $\log(OR)$ approximately follows normal distribution (this is not CLT) with standard error $\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
  - BTW, this test generates p-value of 0.0454



	Men	Women
Awarded	a	b
Not Awarded	c	d

p-value

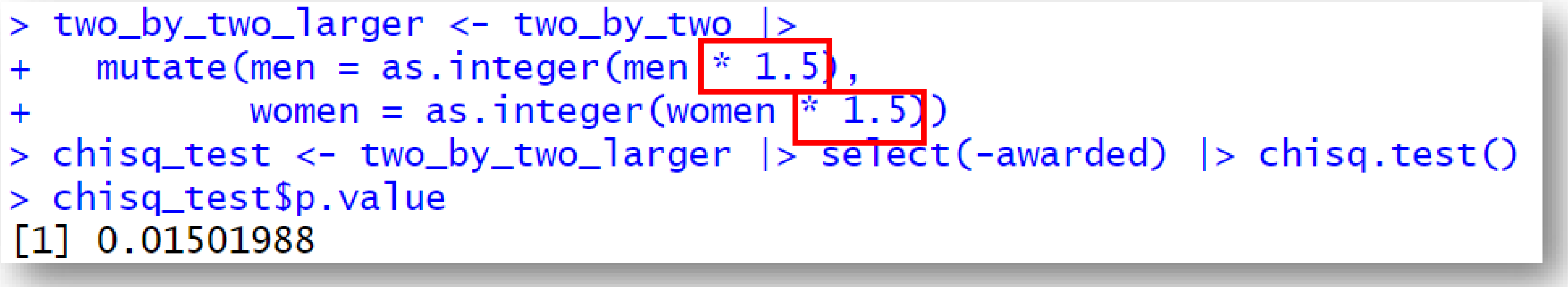
- Major critiques:
  - The threshold of 0.05 is arbitrary
  - P-hacking, multiple hypothesis testing
  - Sensitive to sample size



Larger samples, smaller p-values

- Some studies having large sample sizes tend to report impressively small p-values. Yet, the actual effect size might be modest.

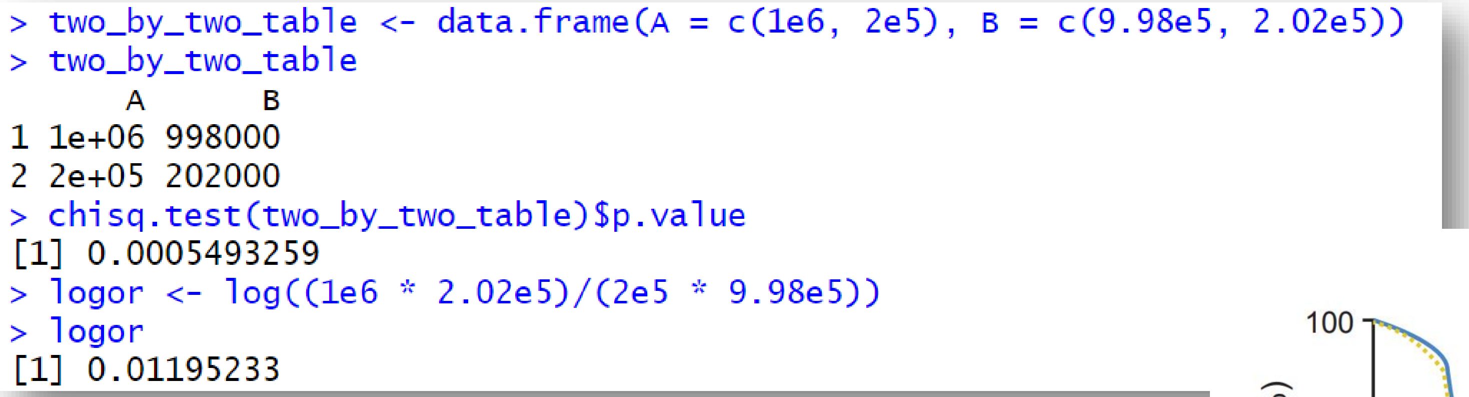
- Same effect size, slightly larger data:



```
> two_by_two_larger <- two_by_two |>
+   mutate(men = as.integer(men * 1.5),
+          women = as.integer(women * 1.5))
> chisq_test <- two_by_two_larger |> select(-awarded) |> chisq.test()
> chisq_test$p.value
[1] 0.01501988
```

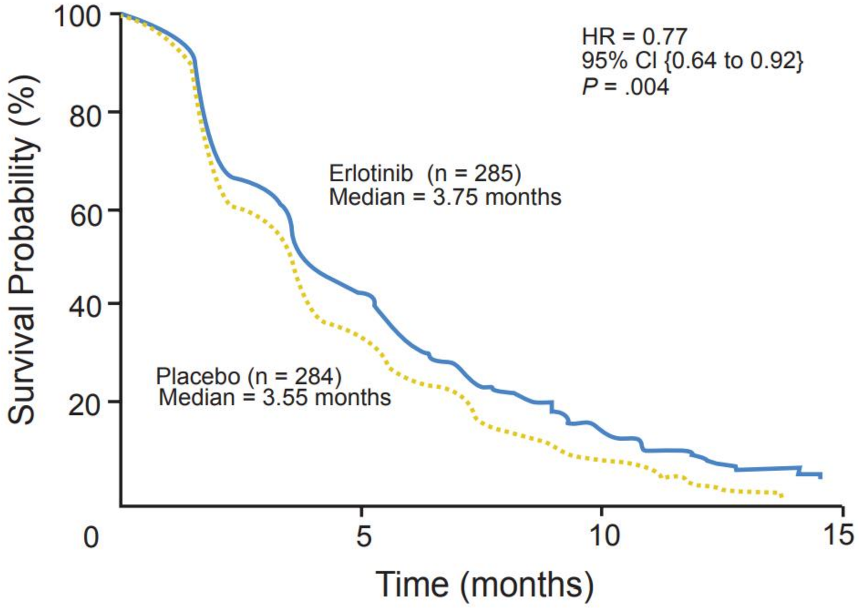
p-value versus effect size

- In the extreme case, even tiny effect could be statistically significant with enough amount of data points



```
> two_by_two_table <- data.frame(A = c(1e6, 2e5), B = c(9.98e5, 2.02e5))
> two_by_two_table
       A      B
1 1e+06 998000
2 2e+05 202000
> chisq.test(two_by_two_table)$p.value
[1] 0.0005493259
> logor <- log((1e6 * 2.02e5)/(2e5 * 9.98e5))
> logor
[1] 0.01195233
```

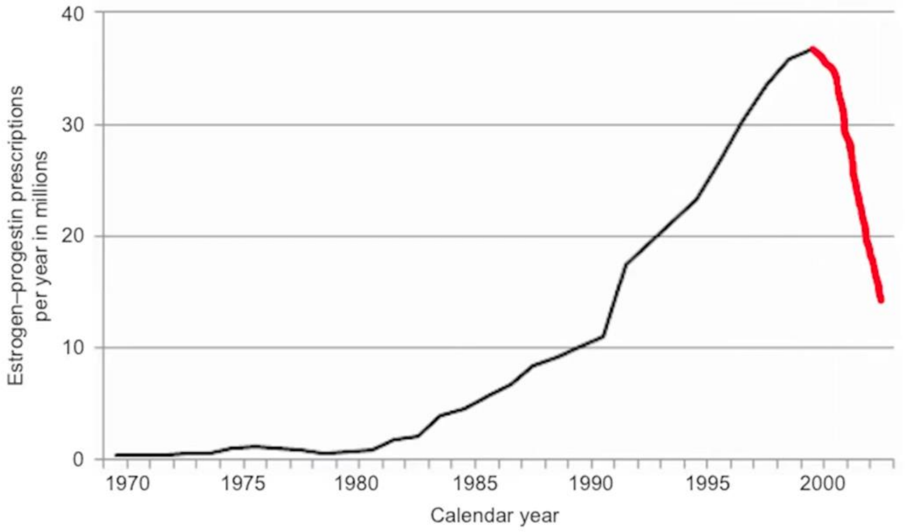
- GWAS: a variant might be associated with a trait with $p < 10^{-20}$, but only explained 1% of the variance/the trait itself.
- Clinical trial: an anti-cancer drug that significantly extend disease-free survival by 0.2 month



Kelley, Robin K., and Andrew H. Ko. "Erlotinib in the treatment of advanced pancreatic cancer." *Biologics: targets and therapy* 2.1 (2008): 83-95.

Large samples with a little bit of bias

- In 1980s, large observational studies show that postmenopausal women on hormone replacement therapies had lower CVD risk.
- However, this was later found to be confounded by other factors; women on therapies are wealthier and more health-conscious.
- Randomized controlled trials later in 1990s and 2000s show that such therapies might increase CVD risks.



To learn more: https://www.youtube.com/watch?v=MVYWqWu2Za4

In this lecture

- What is statistical inference?
- Standard deviation, standard error, confidence interval
- Power
- p-value