Hypothesis Testing

Marc Meredith*

*Introduction to Statistical Methods

Week 5

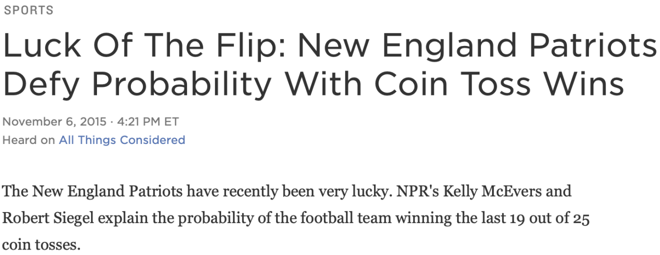
- Last week we developed the concept of estimators
  - Functions that produce estimates of population parameters based on the information contained in a sample of data
- This week will focus on how we test hypotheses about the values of population parameters using estimators
  - With a focus on how we test hypotheses about the values of population parameters after running least squares regressions
- Doing so allows us to establish whether an estimator produces a "statistically significant estimate"
  - One of the most misunderstood statistical concepts

Agenda for week:

- Introduce the concept of hypothesis testing
- Define what it means to reject a hypothesis and the two potential forms of error that are present in hypothesis testing
- Present the concept of a p-value and how it relates to confidence intervals
- Use R to conduct a power test
- Demonstrate how to test a hypothesis about a population mean or the difference between two population means using a regression
- Show how multiple hypotheses can be tested simultaneously

Key takeaways:

- There is a tradeoff in hypothesis testing between rejecting true hypotheses and failing to reject false hypotheses
- We shouldn't apply too rigid of a standard when deciding whether a result is statistically significant
- Power testing is essential when deciding how much data to collect
- Many hypotheses can be refined as statements about the population mean or the difference in two population means and tested using a least squares regression
- Seemingly improbable events often seem less improbable once we consider how many hypotheses are being tested



SPORTS

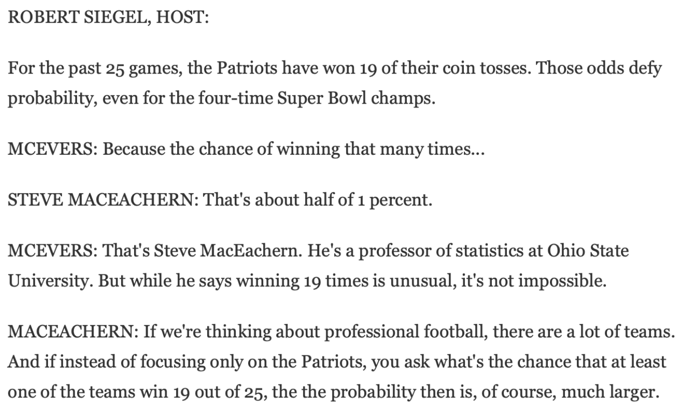
Luck Of The Flip: New England Patriots Defy Probability With Coin Toss Wins

November 6, 2015 · 4:21 PM ET
Heard on All Things Considered

The New England Patriots have recently been very lucky. NPR's Kelly McEvers and Robert Siegel explain the probability of the football team winning the last 19 out of 25 coin tosses.

Source: `https://n.pr/2SOTcsF`

Our goal for the week is to understanding this exchange:



ROBERT SIEGEL, HOST:

For the past 25 games, the Patriots have won 19 of their coin tosses. Those odds defy probability, even for the four-time Super Bowl champs.

MCEVERS: Because the chance of winning that many times...

STEVE MACEACHERN: That's about half of 1 percent.

MCEVERS: That's Steve MacEachern. He's a professor of statistics at Ohio State University. But while he says winning 19 times is unusual, it's not impossible.

MACEACHERN: If we're thinking about professional football, there are a lot of teams. And if instead of focusing only on the Patriots, you ask what's the chance that at least one of the teams win 19 out of 25, the the probability then is, of course, much larger.

Source: `https://n.pr/2SOTcsF`

- Steps of a hypothesis test:
  1. The null hypothesis is specified
  2. The alternative hypothesis is specified
  3. Rejection region is selected
  4. Test statistic is calculated
  5. Null hypothesis is rejected if the test statistic falls into the rejection region, and otherwise it is not rejected
- We'll use the lady tasting tea problem to illustrate each of these steps

- The lady tasting team problem refers to a famous exchange between Muriel Bristol, a noted early 20th Century British scientist, and statistician Ronald Fisher
- Bristol claimed she could tell whether milk was added first or last to tea
- Fisher was skeptical and devised the following experiment to test her claim:
    - Fisher poured eight cups of tea
        - Four of which he added milk first
        - Four of which he added milk last
    - Fisher had Bristol taste each and state whether she thought milk was added first or last
    - If Bristol could correctly identify all eight, he would accept her premise

- Fisher's method for evaluating Bristol's claim contains all of the key elements of a hypothesis test:
  1. The null hypothesis is specified:
     - Bristol cannot tell whether milk is added first or last
  2. The alternative hypothesis is specified:
     - Bristol can tell whether milk is added first or last
  3. Rejection region is selected:
     - Reject null if Bristol current identifies all 8
  4. Test statistic is calculated:
     - Sum the number of teas identified correctly
  5. Reject the null hypothesis if the test statistic falls into the rejection region, and otherwise do not reject

- We need to develop some notation that we can use to represent the null and alternative hypotheses
    - The notation $H_o : \theta = c$ is used to communicate the null hypothesis that population parameter θ is equal to c
    - The notation $H_a : \theta \neq c$ is used to communicate the alternative hypothesis that population parameter θ is not equal to c
        - Used when we lack a strong theoretical sense of which direction our null hypothesis is most likely to be wrong
    - The notation $H_a : \theta < c$ or $H_a : \theta > c$ is used to communicate the alternative hypothesis that population parameter θ is less than c or greater than c, respectively
        - Used when we have a clear theoretical sense of which direction our null hypothesis is most likely to be wrong

- One way to restate Fisher's hypothesis is as a hypothesis about π, which represents the probability that Bristol will correctly identify whether the milk was added first or last
- Lets show how we can represent the null and alternative hypotheses using the notation that we developed on the previous slide
  - The notation $H_o : \theta = c$ is used to communicate the null hypothesis that population parameter θ is equal to c
    - E.g., $H_o : \pi = \frac{1}{2}$
  - The notation $H_a : \theta > c$ is used to communicate the alternative hypothesis that population parameter θ is greater than c
    - E.g., $H_a : \pi > \frac{1}{2}$

- When evaluating a null hypothesis, we make a choice about whether we reject or fail to reject the null hypothesis
  - Meaning that we never "accept" a null hypothesis
- We can make two types of error when evaluating a hypothesis
  1. Reject a null hypothesis even though it is correct (e.g., reject that $\pi = \frac{1}{2}$ even though Bristol has no special skills)
     - This is called type I error
     - Let α represent the probability of type I error
  2. Fail to reject a null hypothesis even though it is incorrect (e.g. fail to reject that $\pi = \frac{1}{2}$ even though Bristol has special skills)
     - This is called type II error
     - Let β represent the probability of type II error

- Once we establish a rejection region it often is straightforward to calculate α (e.g., the probability of making type I error)
- For example, $\alpha = \frac{1}{70}$ if Fisher rejects the null hypothesis when Bristol identifies all 8 teas correctly
    - When $\pi = \frac{1}{2}$, Bristol picks a sequence of eights teas, four of which have milk added first and four of which have milked added last
        - E.g., guessing FLLFLFLF, meaning that tea got added first to cup #1, added last to cup #2, added last to cup #3, ...
    - So $\alpha = \frac{1}{\#Sequences} = \frac{1}{70}$
        - A formula that we didn't have time to develop in this class ($\frac{8!}{4!4!} = 70$) shows that there are 70 potential ways to sequences eight teas, four of which have milk added first and four of which have milked added last
        - As each sequence is equally likely to occur, whatever sequence is guessed has a $\frac{1}{70}$ of being correct

- Calculating β (e.g., the probability of making type II error) is much less straightforward even once a rejection region is established
- Unlike with α, there is not a unique value of β, because it usually depends on the degree to which the null hypothesis is incorrect
- In the Lady Tasting Tea Problem:
  - $\beta = 0$ if Bristol gets it right all of the time
  - $\beta > 0$ if Bristol gets it right most of the time
  - $\beta \approx \frac{69}{70}$ if Bristol gets it right just over half of the time

- General principles of hypothesis testing highlighted by this example that underlie the math moving forward:
    - More likely to reject a hypothesis the more what you see in data diverges from the your expectations given the hypothesis
    - Less likely to reject a hypothesis the greater the sample variance given a specific amount of divergence between what you see in data and your expectations given the hypothesis
    - Hypothesis test can be conceptualized as
        1. Measuring the discrepancy between what you observe in data and your expectations given the hypothesis
        2. Calibrating how much sampling error could be affecting what you observe in data
        3. Calculating the likelihood of observing enough sampling error to cause the discrepancy between what you observe in data and your expectations given the hypothesis

- When selecting a rejection region, the standard practice is to pick the amount of type I error that we are willing to tolerate (e.g., fix α and accept whatever amount of β that this α generates)
  - Find intervals such that the probability that the test statistic is outside the interval if the null is correct is α
    - With $\alpha = .05$ being a semi-established standard
  - And choose the interval most consistent with the alternative hypothesis
- There is a tradeoff between type I and type II error
  - More certainty that we are not incorrectly rejecting a true null (e.g., lower α) means a greater likelihood of failing to reject a false null (e.g., higher β)

- Lets formalize the previous slide when testing a hypothesis about a sample mean $\bar{Y}_n$
  - E.g., we collect a sample of $Y_1, Y_2, ..., Y_n$ that we use to construct $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- $H_o : E[Y_i] = \mu_y$ and $H_a : E[Y_i] \neq \mu_y$
- We want to define a symmetric rejection region using lb, ub st $p(lb < \bar{Y}_n < ub \mid H_o) = 1 - \alpha$
  - Meaning that $p(\bar{Y}_n < lb \mid H_o) = p(\bar{Y}_n > ub \mid H_o) = \frac{\alpha}{2}$
- When $\alpha = .05$ this means
  - 2.5% chance that we could observe a value of $\bar{Y}_n$ that is less than lb when $E[Y_i] = \mu_y$
  - 2.5% chance that we could observe a value of $\bar{Y}_n$ that is greater than ub when $E[Y_i] = \mu_y$

- The values of *lb* and *ub* correspond to the bounds on a $100 * (1-\alpha)\%$ symmetric confidence interval assuming that the null hypothesis is true
- Given $H_o : E[Y_i] = \mu_y,\ var(Y_i) = {\sigma_y}^2$
  - $\frac{\bar{Y}_n - \mu_y}{\sqrt{\frac{{\sigma_y}^2}{n}}} \sim N(0,1) \implies$ $p(\Phi^{-1}(\frac{\alpha}{2}) < \frac{\bar{Y}_n - \mu_y}{\sqrt{\frac{{\sigma_y}^2}{n}}} < \Phi^{-1}(1-\frac{\alpha}{2})) = 1-\alpha \implies$ $p(\mu_y + \Phi^{-1}(\frac{\alpha}{2})\sqrt{\frac{{\sigma_y}^2}{n}} < \bar{Y}_n < \mu_y + \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{{\sigma_y}^2}{n}}) = 1-\alpha \implies$
  - $lb = \mu_y + \Phi^{-1}(\frac{\alpha}{2})\sqrt{\frac{{\sigma_y}^2}{n}}$

    $ub = \mu_y + \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{{\sigma_y}^2}{n}}$
- We reject the null hypothesis at the α level when $\bar{Y}_n$ is not contained in this confidence interval

- We won't always know the value of $var(Y)$ under the null hypothesis
- We still can approximate ${\sigma_y}^2$ with S^2, replacing $\Phi^{-1}()$ with ${T_{n-1}}^{-1}()$
- In such a case:
  - $lb = \mu_y + {T_{n-1}}^{-1}(\frac{\alpha}{2})\sqrt{\frac{S^2}{n}}$
    $ub = \mu_y + {T_{n-1}}^{-1}(1 - \frac{\alpha}{2})\sqrt{\frac{S^2}{n}}$
- We again reject the null hypothesis at the α level when $\bar{Y}_n$ is not contained in this confidence interval

- When the alternative hypothesis is directional, we might want to select different values of *lb* and *ub*
    - In the Lady Tasting Tea problem, observing that she did not correctly identify any of the cups in which milk got poured in first does not support the alternative hypothesis
- When $H_a : E[Y_i] > \mu_y$ we set $lb = -\infty$ and select *ub* st $p(\bar{Y}_n > ub \mid H_o) = \alpha$
    - So that $p(-\infty < \bar{Y}_n < ub \mid H_o) = 1 - \alpha$
- Similarly, when $H_a : E[Y_i] < \mu_y$ we set $ub = \infty$ and select *lb* such that $p(\bar{Y}_n < lb \mid H_o) = \alpha$
    - So that $p(lb < \bar{Y}_n < \infty \mid H_o) = 1 - \alpha$

Example of rejection regions when testing a non-directional (aka two-tailed or two-sided) and directional (aka one-tailed or one-sided) alternative hypothesis with $\alpha = .05$

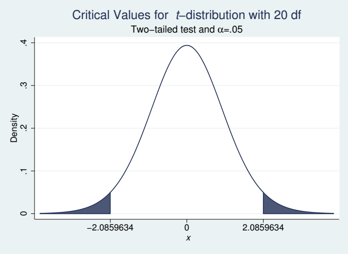


Figure: Two-Tailed Test

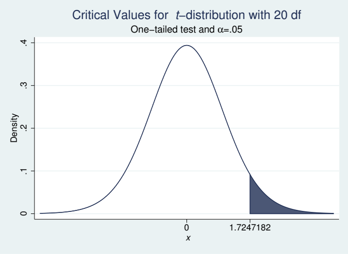


Figure: One-Tailed Test

- To illustrate the concept of a rejection region, consider the following example:
    - $H_o : \pi = \frac{9}{20}$ of the population approves of the president
      $H_a : \pi \neq \frac{9}{20}$ of the population approves of the president
    - My test statistic is the number of people who report supporting the president when I randomly survey 100 members of the population
    - And I (arbitrarily) reject the null hypothesis when the test statistic is between $[0, 39]$ or $[51, 100]$
- On the next slide I show that the probability (α) of making a type I error given this rejection region is .269

- Let X be a r.v. that represents the number of people who approve of the president in a sample of 100
- Given H_o :
  - X is a binomial r.v. with $\pi = \frac{9}{20}$ and $n = 100 \implies p(x \mid H_o) = \binom{100}{x} \frac{9}{20}^{x} \frac{11}{20}^{100-x}$
  - To calculate α, we can use $p(x \mid H_o)$ to calculate the probability that X has a realization in the rejection region when the null hypothesis is correct
  - $\alpha =$ $p(X \leqslant 39 \mid H_o) + p(X \geqslant 51 \mid H_o) =$ $1 - p(40 \leqslant X \leqslant 50 \mid H_o) =$ $1 - \sum_{i=40}^{50} \binom{100}{i} \frac{9}{20}^{i} \frac{11}{20}^{100-i} = .269$
    - Solved with “1 - (pbinom(50, 100, .45) - pbinom(39, 100, .45))” in R

- My probability (β) of making a type II error depends on the true value of π
  - When $\pi = \frac{11}{20}$, then $p(x) = \binom{100}{x} \frac{11}{20}^{x} \frac{9}{20}^{100-x} \implies \beta = \sum_{i=40}^{50} \binom{100}{i} \frac{11}{20}^{i} \frac{9}{20}^{100-i} \approx .182$
    - Solved with "pbinom(50, 100, .55) - pbinom(39, 100, .55)" in R
  - When $\pi = \frac{13}{20}$, then $p(x) = \binom{100}{x} \frac{13}{20}^{x} \frac{7}{20}^{100-x} \implies \beta = \sum_{i=40}^{50} \binom{100}{i} \frac{13}{20}^{i} \frac{7}{20}^{100-i} \approx .007$
    - Solved with "pbinom(50, 100, .65) - pbinom(39, 100, .65)" in R
- The closer that true value of π is to $\frac{9}{20}$, the greater the chance at type II error

- Suppose instead that I want to solve for a symmetric rejection region such that $\alpha = .05$ because $H_a : \pi \neq \frac{9}{20}$
- On the next slide, I solve that I would therefore reject the null if 34 or fewer, or 56 or more people supported the president in a sample of 100 given H_o
- Solving this problem requires knowledge of the inverse CDF my test statistic under the null
- Thus, I use the normal approximation of the binomial distribution rather than the exact binomial distribution to model the distribution of the test statistic under the null
    - Because the binomial is the sum of Bernoulli r.v.s, the Central Limit Theorem says it will become approximately normal as n gets large
    - It is easier to work with the inverse CDF of the normal distribution than the inverse CDF of the binomial distribution

- Let $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$
  - Y_i equals 1 if someone approves of the president and 0 if someone doesn't approve of the president
- Given $H_o : \pi = .45$, $\frac{\bar{Y}_{100} - .45}{\sqrt{\frac{.45 \ast .55}{100}}} \sim N(0,1)$
  - When $H_o : \pi = \pi_o$, $\frac{\bar{Y}_n - \pi_o}{\sqrt{\frac{\pi_o(1-\pi_o)}{n}}} \sim N(0,1)$
    - Keeping in mind that the variance of a Bernoulli r.v. is $\pi(1-\pi)$
- $\Phi^{-1}(.975) = 1.96 \implies$
  - Solved using “qnorm(.975)”

$$p\left(-1.96 < \frac{\bar{Y}_{100} - .45}{\sqrt{\frac{.45 \ast .55}{100}}} < 1.96\right) = .95 \implies$$

$$p\left(.45 - 1.96\sqrt{\frac{.45.55}{100}} < \bar{Y}_{100} < .45 + 1.96\sqrt{\frac{.45.55}{100}}\right) = .95 \implies$$

$$p(.352 < \bar{Y}_{100} < .548) = .95$$

- Suppose that $H_a : \pi > \frac{9}{20}$, so I want to use a one-tailed test
- Below shows why we would choose to reject the null if 54 or more people supported the president in a sample of 100 given H_o :
  - $\Phi^{-1}(.95) = 1.645 \implies$
    - Solved using "qnorm(.95)"

$$p(-\infty < \frac{\bar{Y}_{100} - .45}{\sqrt{\frac{.45.55}{100}}} < 1.645) = .95 \implies$$

$$p(-\infty < \bar{Y}_{100} < .45 + 1.645\sqrt{\frac{.45.55}{100}}) = .95 \implies$$

$$p(-\infty < \bar{Y}_{100} < .532) = .95$$

- Comparing the last two slides highlights the difference between directional and non-directional hypothesis testing
  - Both tests reject the null when 56 or more people support the president
  - Both tests fail to reject the null when 35 to 53 people support the president
  - The symmetric (but not the one-tailed) test rejects the null when 0 to 34 people support the president
  - The one-tailed (but not the symmetric) test rejects the null when 54 or 55 people support the president
- Requires us to think ahead of time about what constitutes evidence against the null hypothesis

- An alternative way of thinking about hypothesis testing is the concept of a p-value
- A p-value (p) is the smallest level of α for which the observed data indicate that the null hypothesis should be rejected
  - E.g., the most amount of type I error that we could tolerate and still reject the null hypothesis
  - So when $p = .03$
    - We would reject the null if we are willing to tolerate a three percent chance that we are rejecting a true null hypothesis , which is true if we would tolerate a three, five, ten, . . . percent chance that we are rejecting a true null hypothesis
    - We would fail to reject the null if we aren't willing to tolerate a three percent chance that we are rejecting a true null hypothesis, which is true if we would only tolerate a two, one, one-tenth, . . . percent chance that we are rejecting a true null hypothesis

- We solve for the p-value of a non-directional hypothesis test by calculating the likelihood of observing sufficient sampling error to cause a deviation that is at least large as the observed deviation between sample mean and its expected value given the null hypothesis
- Consider the case of the sample mean with $H_o : E[Y_i] = \mu_y$ and $H_a : E[Y_i] \neq \mu_y$
- Define $q = \frac{|\bar{Y}_n - \mu_y|}{\sqrt{\frac{{\sigma_y}^2}{n}}}$
    - $\bar{Y}_n$ is q standard deviations away from its expected value under the null hypothesis
- $p = 2(1 - \Phi(q))$
    - Because $Z = \frac{\bar{Y}_n - \mu_y}{\sqrt{\frac{{\sigma_y}^2}{n}}} \sim N(0,1)$ given $H_o \implies$ $p(\mid Z \mid > q) = 1 - p(\mid Z \mid < q) =$ $1 - \Phi(q) - \Phi(-q)) = 1 - (2\Phi(q) - 1) = 2(1 - \Phi(q))$

- We solve for the p-value on a directional hypothesis test about a population mean in a similar manner
- Consider the case of the sample mean with $H_o : E[Y_i] = \mu_y$ and $H_a : E[Y_i] > \mu_y$
    - Thus, observing values of $\bar{Y}_n < \mu_y$ is no longer evidence supporting the alternative hypothesis
- Define $q = \frac{\bar{Y}_n - \mu_y}{\sqrt{\frac{{\sigma_y}^2}{n}}}$
    - $\bar{Y}_n$ is q standard deviations above its expected value under the null hypothesis
- $p = 1 - \Phi(q)$
    - Because $Z = \frac{\bar{Y}_n - \mu_y}{\sqrt{\frac{{\sigma_y}^2}{n}}} \sim N(0,1)$ given $H_o \implies p(Z > q) = 1 - \Phi(q)$

- We solve for the p-value on a directional hypothesis test about a population mean in a similar manner
- Consider the case of the sample mean with $H_o : E[Y_i] = \mu_y$ and $H_a : E[Y_i] < \mu_y$
  - Thus, observing values of $\bar{Y}_n > \mu_y$ is no longer evidence supporting the alternative hypothesis
- Define $q = \frac{\bar{Y}_n - \mu_y}{\sqrt{\frac{{\sigma_y}^2}{n}}}$
  - $\bar{Y}_n$ is q standard deviations below its expected value under the null hypothesis
- $p = \Phi(q)$
  - Because $\mathrm{Z} = \frac{\bar{Y}_n - \mu_y}{\sqrt{\frac{{\sigma_y}^2}{n}}} \sim N(0,1)$ given $H_o \implies$ $p(Z < q) = \Phi(q)$

- Suppose we observe 58 people supported the president in our polling example
- Let $H_o : \pi = .45$ and $H_a : \pi \neq .45$
- $q = \frac{.58 - .45}{\sqrt{\frac{.58 .42}{100}}} \implies q = 2.634$
- $p = 2(1 - \Phi(2.634)) = .0084$
  - Solved using "2*(1 - pnorm(2.634))" in R
- Interpretation: There is a 0.84 percent chance that sampling error would cause a deviation of at least 2.634 standard deviation units from the hypothesized value when the null is correct

- How does this change when we have directional alternative hypotheses?
- When $H_o : \pi = .45$ and $H_a : \pi > .45$
  - $p = (1 - \Phi(2.634)) = .0042$
  - Interpretation: There is a 0.42 percent chance that sampling error would cause a deviation of more than 2.634 standard deviation units when the null is correct and the null specifies the minimum possible value
- When $H_o : \pi = .45$ and $H_a : \pi < .45$
  - $p = \Phi(-2.634) = .9958$
  - Interpretation: There is a 99.58 percent chance that sampling error would cause a deviation of more than 2.634 standard deviation units when the null is correct and the null specifies the maximum possible value

Notes about p-values:

- The phrase “statistically significant” is often used to refer to a hypothesis with a small p-value
    - Most commonly a p-value less than .05
- There is a distinction between statistical and substantive significance
    - Statistical significance is about the certainty with which a null hypothesis can be rejected
    - Substantive significance refers to whether the difference between your estimated and hypothesized value is meaningful

Notes about p-values (continued):

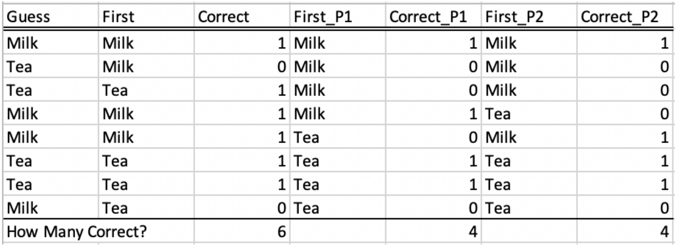
- In large samples, a substantively insignificantly deviation from a null hypothesis can be statistically significant because the sampling variance is small
- In small samples, a substantively significant deviation from a null hypothesis can be statistically insignificant because the sampling variance is large
- A high p-value generally should not be taken as evidence that the null hypothesis is correct, especially in small samples

- Having faster computers have caused techniques that simulate, rather than calculate, p-values to become increasingly popular
    - Chapter 3 of Bruce & Bruce highlights some of these approaches
- While we don't have time to go into too much detail in this course, I will introduce two of these approaches in case you want to read further on your own
    1. Exhaustive permutation test
    2. Bootstrap permutation test

Exhaustive permutation test:

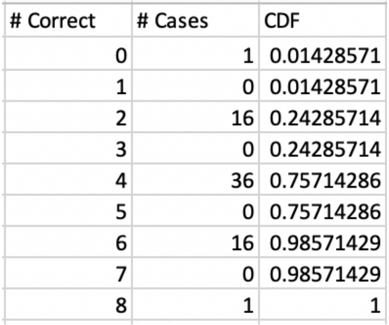
1. Construct a statistic of interest in some data
2. Count how many observations take-on each value of an independent variable in that data
3. Permute the values of the independent variable, keeping the total number of observations assigned to a value of the independent variable equal to the total number with that value in the original data
4. Construct the statistic of interest using this permuted dataset
5. Repeat steps 3 and 4 for every possible combination of independent variables that keeps the total number of observations assigned to a value of the independent variable equal to the total number with that value in the original data
6. Calculate how often you observe a test statistic of greater magnitude in the permuted data than in the original data

We could use a permutation test to solve the lady tasting tea problem:



Guess	First	Correct	First_P1	Correct_P1	First_P2	Correct_P2
Milk	Milk	1	Milk	1	Milk	1
Tea	Milk	0	Milk	0	Milk	0
Tea	Tea	1	Milk	0	Milk	0
Milk	Milk	1	Milk	1	Tea	0
Milk	Milk	1	Tea	0	Milk	1
Tea	Tea	1	Tea	1	Tea	1
Tea	Tea	1	Tea	1	Tea	1
Milk	Tea	0	Tea	0	Tea	0
How Many Correct?		6		4		4

When we extend to all 70 possible permutation we solve for the following distribution:



# Correct	# Cases	CDF
0	1	0.01428571
1	0	0.01428571
2	16	0.24285714
3	0	0.24285714
4	36	0.75714286
5	0	0.75714286
6	16	0.98571429
7	0	0.98571429
8	1	1

- We can use the fisher.test() function to apply this test in R:

```
> print(Fisher)
# A tibble: 8 x 2
  Guess First
  <chr> <chr>
1 Milk  Milk
2 Tea   Milk
3 Tea   Tea
4 Milk  Milk
5 Milk  Milk
6 Tea   Tea
7 Tea   Tea
8 Milk  Tea
```

- To use the function fisher.test(), we first need to summarize these data with a contingency table
- Then send the resulting stored contingency table as an argument into fisher.test()

```
> exact <- table(Fisher$Guess, Fisher$First)
> print(exact)

       Milk Tea
  Milk    3   1
  Tea     1   3
> fisher.test(exact, simulate.p.value = FALSE)

        Fisher's Exact Test for Count Data

data:  exact
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

- Because it can be computationally challenging to compute the exact distribution, we may use a bootstrapped distribution to approximate it
- The bootstrap is a method for approximating the distribution of an estimator or test statistic by treating a sample of data as if it were the population and resampling from it
- Example:
  - Let $X_1 = 12$, $X_2 = 6$, $X_3 = 3$
  - If I resample 3 items from the distribution with replacement, I would get each of these bootstrap samples with probability $1/27$
    - $(3, 3, 3)$, $(3, 3, 6)$, $(3, 3, 12)$, $(3, 6, 3)$, $(3, 6, 6)$, $(3, 6, 12)$, $(3, 12, 3)$, $(3, 12, 6)$, $(3, 12, 12)$, $(6, 3, 3)$, $(6, 3, 6)$, $(6, 3, 12)$, $(6, 6, 3)$, $(6, 6, 6)$, $(6, 6, 12)$, $(6, 12, 3)$, $(6, 12, 6)$, $(6, 12, 12)$, $(12, 3, 3)$, $(12, 3, 6)$, $(12, 3, 12)$, $(12, 6, 3)$, $(12, 6, 6)$, $(12, 6, 12)$, $(12, 12, 3)$, $(12, 12, 6)$, $(12, 12, 12)$

Bootstrap permutation test:

1. Construct a statistic of interest in some data
2. Count how many observations take-on each value of an independent variable in that data
3. Permute the values of the independent variable, keeping the total number of observations assigned to a value of the independent variable equal to the total number with that value in the original data
4. Construct the statistic of interest using this permuted dataset
5. Repeat steps 3 and 4 n times, where $n <$ every possible combination of independent variables that keeps the total number of observations assigned to a value of the independent variable equal to the total number with that value in the original data
    - Need to decide about how to select the n cases (with or without replacement)
6. Calculate how often you observe a test statistic of greater magnitude in the permuted data than in the original data

- We can implement a bootstrap. permutation test by using the syntax "Fisher.test(XXX, simulate.p.value = TRUE, B = YYY)"
  - Although is not available when the contingency table is 2 X 2 because computing the exact distribution is not computationally challenging in this case
- XXX is the name of a stored contingency table
- YYY represents a number, that is specifying how many times you want to resample
  - Often observe people bootsrapping 500 to 1000 cases

- One of the most important applications of hypothesis testing is using your expectation about how a hypothesis test is likely to play out to guide decisions about how much data to collect
- Power testing is used to inform yourself of one of the following four variables given what you specify about the other three:
  - The sample size (n)
  - The effect size (f) that you want to detect
  - The significance level (α) at which the hypothesis test will be conducted
  - The probability ($1-\beta$) that you reject the null hypothesis given f and α
- Most often used to calculate how big n must be to reject the null hypothesis of no difference with probability p when you only want an α chance of making type I error and effect size is actually f

- To illustrate how to conduct a power test, we'll use the pwr.2p.test() function that is part of the "pwr" library in R
- This function allows us to do a power calculation for comparing proportions in two populations
  - E.g., Compare turnout in a mobilized group and a non-mobilized control group
- The syntax is "pwr.2p.test(h = , n = , sig.level = , power = , alternative =)":
  - h is the effect size
  - n is the sample size (in each group)
  - sig.level is the value of α
  - power is $1 - \beta$
  - alternative = "two.sided", "less", or "greater"

- Suppose we want to figure out how many people to include a voter mobilization experiment
- Before running our experiment we believe that:
  - Our mobilization increases turnout by 3 percentage points
  - We want to reject the null hypothesis that mobilization has no effect on turnout 60 percent of the time at the $\alpha = .05$ level
  - Our alternative hypothesis is that mobilization increases turnout
- The output below shows that we need to mobilize about 8,000 people (and have at least 8,000 people who we haven't mobilized)

```
> pwr.2p.test(h = .03, n = NULL, sig.level = 0.05,
                               power = 0.6, alternative = c("greater"))

              h = 0.03
              n = 8007.036
      sig.level = 0.05
          power = 0.6
    alternative = greater
```

- We'll now turn to thinking about how we test hypotheses about regression coefficients
- Let $\hat{\beta}_1$ be our estimate of β_1
- Let $var(\hat{\beta}_1)$ be the sampling variance of $\hat{\beta}_1$
- Under certain conditions, $\frac{\beta - \hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}} \sim T_{n-k-1}$
    - Where k is the number of explanatory variables included in the regression model (e.g., $k = 1$ when running a bivariate regression)
    - Requires either a large enough sample to apply the CLT or an assumption that $\epsilon_i \sim N(0, \sigma^2)$

- When $\frac{\beta-\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}} \sim T_{n-k-1}$ testing a hypothesis about a single regression coefficient is very straightforward
- Let $H_o : \beta_1 = c$ and $H_a : \beta_1 \neq c$
- Calculate the q such that $\mid \hat{\beta}_1 - c \mid = q\sqrt{var(\hat{\beta}_1)} \implies$ $q = \frac{|\hat{\beta}_1 - c|}{\sqrt{var(\hat{\beta}_1)}} \implies$ $p = 2 * (1 - T_{n-k}(q))$
- Example:
  - Suppose $\hat{\beta}_1 = 2$, $c = 1$, $var(\hat{\beta}_1) = \frac{1}{4}$, and $n - k = 100$
  - $q = \frac{|2-1|}{\sqrt{\frac{1}{4}}} = 2 \implies$ $p = 2 * (1 - T_{100}(2)) = .048$

- R includes the p-value on the hypothesis test that $H_o : \beta_j = 0$ and $H_a : \beta_j \neq 0$ automatically in its output
- When $c = 0$, then $q = \frac{|\hat{\beta}_j|}{\sqrt{var(\hat{\beta}_j)}} \implies$

  $p = 2 * (1 - T_{n-k}(\frac{|\hat{\beta}_j|}{\sqrt{var(\hat{\beta}_j)}}))$

```
reg1 <- lm(Voted ~ Mobilized, data = LectureDifferenceMeans)
reg1summary <- summary(reg1)
print(reg1summary)

Call:
lm(formula = Voted ~ Mobilized, data = LectureDifferenceMeans)

Residuals:
   Min     1Q Median     3Q    Max
 -0.38  -0.38  -0.35   0.62   0.65

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.35000    0.02411  14.517   <2e-16 ***
Mobilized    0.03000    0.03235   0.927    0.354
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4822 on 898 degrees of freedom
Multiple R-squared:  0.0009569,        Adjusted R-squared:  -0.0001556
F-statistic: 0.8602 on 1 and 898 DF,  p-value: 0.3539
```

- The output on the previous slide shows the p-value on $\beta_{mobilized}$ equals 0.354
- How R calculates this
  - $\hat{\beta}_{mobilized} = 0.03000$, $se_{foreign} = 0.03235$, and 898 degrees of freedom (DF)
  - $H_o : \beta_{mobilized} = 0 \implies Q = \frac{\hat{\beta}_{mobilized} - 0}{se_{mobilized}} \sim T_{898}$
  - $p = p(\mid Q \mid > 0.927) = 2(1 - T_{898}(0.927)) = 2(1 - 0.823) = 0.354$
    - Solved in R using "2*(1 - pt(0.927, 898))"
- Thus, we conclude that we would have a 35.4 percent chance of making type I error if we rejected the null hypothesis that people who are mobilized vote at the same rate as people who are not mobilized based on these data

- We could also show that the p-value on $H_o : \beta_{mobilized} = 0$ is 0.354 by showing that 65.6% symmetric CI is the smallest symmetric CI for $\beta_{mobilized}$ that includes 0

```
> confint(reg1, level = 0.656)
                  17.2 %    82.8 %
(Intercept)   0.3271729 0.3728271
Mobilized    -0.0006258 0.0606258
```

- A t-test is a special case of a Wald test (or F-test), which is a more general framework used to test hypotheses about regression coefficient(s)
- Intuition of a Wald Test is as follows:
  - Create a discrepancy vector measuring the difference between the estimated regression coefficient(s) and the expected regression coefficient(s) given all of the null hypotheses
  - Standardize this discrepancy vector by a measure of the expected amount of sampling error
  - Generate a single number that summarizes the amount standardized discrepancy observed across all of the null hypotheses
  - Figure out the likelihood that sampling error would cause this aggregate amount of discrepancy from all of the null hypotheses
- We use the "linearHypothesis()" function in R, which is included in "library(car)", to implement Wald tests

- How to use the linearHypothesis function to replicate the t-test produced as part of R's baseline output:

```
> library(car)
Loading required package: carData
> linearHypothesis(reg1, "Mobilized = 0")
Linear hypothesis test

Hypothesis:
Mobilized = 0

Model 1: restricted model
Model 2: Voted ~ Mobilized

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    899 209.0
2    898 208.8  1       0.2 0.8602 0.3539
```

- How to use the linearHypothesis function to conduct a t-test beyond what is produced as part of R’s baseline output

```
> linearHypothesis(reg1, "Mobilized = 0.01")
Linear hypothesis test

Hypothesis:
Mobilized = 0.01

Model 1: restricted model
Model 2: Voted ˜ Mobilized

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    899 208.89
2    898 208.80  1  0.088889 0.3823 0.5365
```

- How to use the linearHypothesis function to conduct a F-test about a linear combination of two or more coefficients

```
> linearHypothesis(reg1,  "(Intercept) + Mobilized = 0.4")
Linear hypothesis test

Hypothesis:
(Intercept)  + Mobilized = 0.4

Model 1: restricted model
Model 2: Voted ˜ Mobilized

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    899 209.0
2    898 208.8  1       0.2 0.8602 0.3539
```

- How to use the linearHypothesis function to conduct a joint F-test about the value of two or more coefficients
    - Value will be clearer next week once we start including more than one explanatory variable in a regression

```
> linearHypothesis(reg1,  c("(Intercept) = 0.5", "Mobilized = 0.01"))
Linear hypothesis test

Hypothesis:
(Intercept) = 0.5
Mobilized = 0.01

Model 1: restricted model
Model 2: Voted ˜ Mobilized

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    900 226.25
2    898 208.80  2     17.45 37.524 2.232e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- We refer to the phenomenon of testing multiple hypotheses simultaneously as multiple testing (or multiple comparisons)
- The Wald test on the previous slide is an example of testing multiple hypothesis jointly
    - E.g., returned one p-value on the null hypothesis that the intercept was equal to .5 and coefficient on mobilized was 0.01
        - Rather than a p-value on the null hypothesis that the intercept was equal to .5 and the p-value on the null hypothesis that the coefficient on mobilized was 0.01
- Recently it has become better understood that it is important to test multiple hypothesis jointly
    - To avoid the so-called multiple testing problem

- To better understand the multiple-testing problem, lets return to New England Patriot coin-flipping story at the start of the lecture
- The story established that the New England Patriots won 19 out of 25 coin tosses
- To highlight the multiple-testing problem, we are going to
  - Test the null hypothesis $H_o : \pi_{ne} = .5$
    - Where π_{ne} is the probability that the Patriots win a coin toss
  - Test the null hypothesis $H_o : \pi_{ne} = \pi_{gb} = \pi_{phl} = \cdots = .5$
    - Where $\pi_{gb}, \pi_{phl}, \ldots,$ is the probability that the Packers, Eagles, and each of the remaining 29 NFL teams wins a coin toss

Testing the null hypothesis $H_o : \pi_{ne} = .5$

- Let X be a r.v. equal to the number coin flips NE wins out of 25
- If H_o is true, then $p(x) = \binom{25}{x} \frac{1}{2}^{x} \frac{1}{2}^{25-x}$
  - Where $p(19) = \binom{25}{19} \frac{1}{2}^{19} \frac{1}{2}^{6} = 0.0053$
- If the alternative hypothesis is $H_a : \pi_{ne} \neq .5$, then:

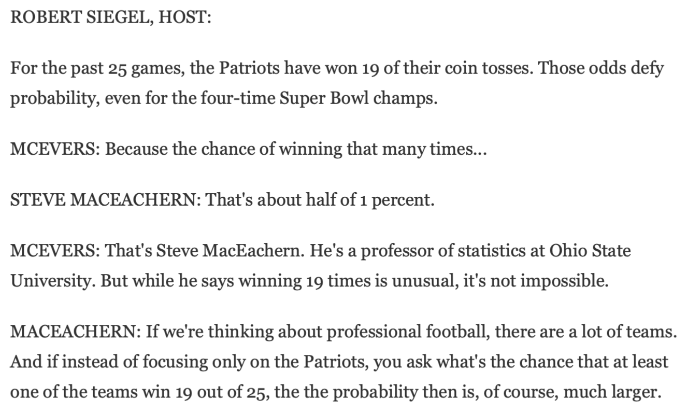
  $\alpha = \sum_{j=0}^{6} \frac{1}{2}^{j} \frac{1}{2}^{25-j} + \sum_{j=19}^{25} \frac{1}{2}^{j} \frac{1}{2}^{25-j} = 0.0146$
  - Solved in R using "1 - (pbinom(18, 25, .5) - pbinom(6, 25, .5))"
- If the alternative hypothesis instead is $H_a : \pi_{ne} > .5$, then:

  $\alpha = \sum_{j=19}^{25} \frac{1}{2}^{j} \frac{1}{2}^{25-j} = 0.0073$

- The previous slide shows there is a 0.73 percent chance that the Patriots would win 19 or more of 25 coin flips
- Let Y be a r.v. equal to the number of NFL teams that win 19 or more of 25 coin flips
- If $H_o : \pi_{ne} = \pi_{gb} = \pi_{phl} = \cdots = .5$, then $p(y) \approx \binom{32}{y} 0.0073^y 0.9927^{32-y}$
  - $p(y)$ is constructed using the binomial distribution
  - Y is only being approximated by the binomial distribution because the number of coin flips won by one team is not independent of the number of coin flips won by another
- $p(0) \approx 0.9927^{32} = 0.740$
  - There is about a 74 percent chance that each team flips 32 coins, none of the teams will win 19 or more of these coin tosses

- The New England Patriot coin-flipping story highlights the importance of thinking about your hypothesis and testing every element of it when constructing a p-value
  - We have less than 1 percent chance of making type I error if we reject the null hypothesis that the Patriots are 50% likely to win a coin toss based on these data
  - We have about a 24 percent chance of making type I error if we reject the null hypothesis that each NFL team is 50% likely to win a coin toss based on these data
- The broader lesson is to think clearly based on your theory and what it implies about your null hypothesis
- Important because evidence inconsistent with a null hypothesis is often more salient than evidence that is consistent with a null hypothesis

We should now be able to understand this exchange:



ROBERT SIEGEL, HOST:

For the past 25 games, the Patriots have won 19 of their coin tosses. Those odds defy probability, even for the four-time Super Bowl champs.

MCEVERS: Because the chance of winning that many times...

STEVE MACEACHERN: That's about half of 1 percent.

MCEVERS: That's Steve MacEachern. He's a professor of statistics at Ohio State University. But while he says winning 19 times is unusual, it's not impossible.

MACEACHERN: If we're thinking about professional football, there are a lot of teams. And if instead of focusing only on the Patriots, you ask what's the chance that at least one of the teams win 19 out of 25, the the probability then is, of course, much larger.

Source: `https://n.pr/2SOTcsF`

Key takeaways:

- There is a tradeoff in hypothesis testing between rejecting true hypotheses and failing to reject false hypotheses
  - Which we usually resolve by anchoring the probability of type I error and accepting whatever type II error this generates
- We shouldn't apply too rigid of a standard when deciding whether this likelihood is significant or not
  - The evidence supporting rejecting the null is pretty comparable when the p-value on a hypothesis test is 0.049 and 0.051
- Power testing is essential when deciding how much data to collect
  - Need to ensure we collect enough data to say something meaningful

Key takeaways:

- Many hypotheses can be refined as statements about the population mean or the difference in two population means
    - Keeping in mind the discussion from last week about what regression coefficients tell us about differences in conditional expectations
- Seemingly improbable events often seem less improbable once we consider how many hypotheses are being (implicitly) tested
    - E.g., No news stories are written about the team that wins the exact number of coin tosses that statistical theory would predict