Unit 2C - Comparing Probabilities Between Groups

Adam Petrie
Department of Business Analytics
University of Tennessee

February 8, 2021

1. Motivation for Comparing Probabilities
    - Examples
    - Recap: making inference on p
2. Comparing Two Probabilities
    - Intuition
    - Master Formulas
    - Examples and `prop.test`
3. Comparing Multiple Probabilities
    - Motivation (and using `aggregate` and `mosaic`)
    - Tukey's Honest Significant Difference
    - Connecting Letters Report
4. Summary

Motivation for Comparing Probabilities

Motivation - comparing groups

In business analytics, it is more common to compare probabilities of two or more events than it is to make an inference about the probability that one particular event occurs.

- Three inventory control policies are being considered that have roughly equal costs. Which (if any) has the lowest probability of a stockout?
- There are two competing designs for a banner ad. Which has the higher click-thru rate (probability that a randomly picked surfer clicks the ad)?
- There are 83 distinct undergraduate degrees at UT, concentrated into 75 majors within 24 broad fields of study. Which major(s) have graduates with the highest probability of donating at least $1000 to UT within five years of graduation?

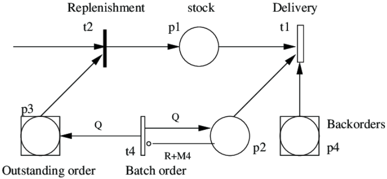
Motivation - comparing groups

Comparing the probability of two events (e.g., clicking on ad A, clicking on ad B), is one type of **A/B Testing**. This task is part of a business analytics practitioner's daily life!

Based on a Monte Carlo simulation we have ran, or on the data that we have collected, can we discern any difference in the probabilities of two events (is the click-thru rate of ad A higher than of B)?

Can we definitively rank the probabilities of multiple events from highest to lowest?

Motivation - stockout probabilities



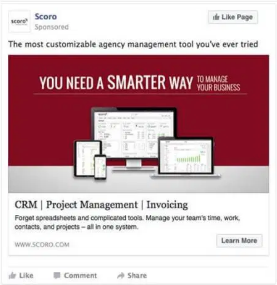
You possess a realistic model for daily demand probabilities. A Monte Carlo simulation is run for each of three inventory control policies to find the probability of a stockout for each. Only limited exploration has taken place so far.

Policy	1	2	3
Trials	80	200	50
Stockouts	12	20	7
$\hat{p}$	0.15	0.10	0.14

Does this suggest Policy 1 is worst and Policy 2 is best? Or are these estimates too close to each other to determine if there's any real difference?

Motivation - click-thru rates

4000 people saw the ad on the left; 4000 other people saw the ad on the right.



Scoro
Sponsored

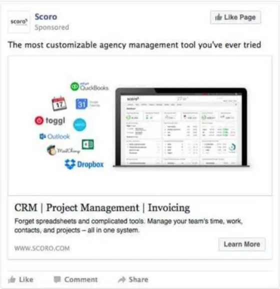
The most customizable agency management tool you've ever tried

YOU NEED A SMARTER WAY TO MANAGE YOUR BUSINESS

CRM | Project Management | Invoicing

Forget spreadsheets and complicated tools. Manage your team's time, work, contacts, and projects – all in one system.

WWW.SCORO.COM



Scoro
Sponsored

The most customizable agency management tool you've ever tried

CRM | Project Management | Invoicing

Forget spreadsheets and complicated tools. Manage your team's time, work, contacts, and projects – all in one system.

WWW.SCORO.COM

78 people clicked the left ad (1.95%). Less than half of that number (35) clicked on the right ad (0.875%). What's a range of plausible values for the difference in "true" click-thru rates, and does this suggest one ad is better than the other?

Motivation - UT donations

Among graduates of the College of Business who earned their degree in 2009 or before, many have donated something back to UT within five years of graduating. What majors have different donation probabilities from each other?

```
                                No  Yes  p.hat
Accounting                    4723 1588  .336
Business Administration       2245  527  .235
Finance                       3998  935  .234
Management                    1625  346  .213
General Business              1503  313  .208
Logistics & Transportation 1532  310  .202
Marketing                     4718  950  .201
Logistics                      429   80  .186
Economics                      934  141  .151
Transportation/Logistics      1970  288  .146
```

Recap: making inference on p

In Unit 3, we became experts at estimating the probability of a single event (like winning a game of GEMS).

- p - the actual probability of the event of interest occurring
- $\hat{p}$ - the estimated probability
- n - the number of trials in the Monte Carlo simulation, or the sample size of the collected data
- SE - the **standard error** of $\hat{p}$; its value will be our "best guess" of how far $\hat{p}$ is from p

Standard error of $\hat{p}$

- After we've conducted a Monte Carlo simulation, we do not expect the proportion of trials where our event of interest occurred ($\hat{p}$) to equal p, but it should be "close".
- The typical difference between p and $\hat{p}$ is referred to as the "standard error" of $\hat{p}$, and we have figured out a formula for it:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Usually (meaning about 2/3rds of the time), the actual value of p will be at most one standard error away from $\hat{p}$.

A 95% confidence interval for p

If we want a range of plausible values for p based on the data we have collected, the "industry standard" is a 95% confidence interval.

$$\hat{p} \pm 2SE$$

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\left(\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Master formula for confidence interval for p

If we want a different level of confidence, we write the desired confidence as a number between 0-1 and find

Arbitrary confidence interval for p

$$\hat{p} \pm \texttt{qnorm(1-(1-confidence)/2)} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

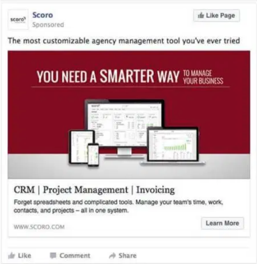
Extending analysis to multiple events

So how does the analysis extend to comparing probabilities of multiple events, e.g. stockout using policy 1 vs. stockout using policy 2?

- How do we "best guess" the probability for each event?
- How far off is the estimated difference in probabilities of two events from the "true" difference?
- Can we come up with a range of plausible values for the true difference?
- Is it possible to definitively rank multiple events from highest probability to lowest? If so, when?

Comparing Two Probabilities

Intuition regarding comparing two probabilities

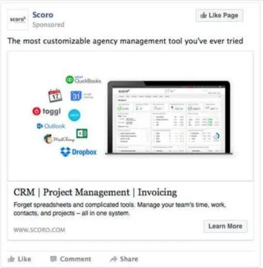


The most customizable agency management tool you've ever tried

CRM | Project Management | Invoicing

Forget spreadsheets and complicated tools. Manage your team's time, work, contacts, and projects – all in one system.

WWW.SCORO.COM



The most customizable agency management tool you've ever tried

CRM | Project Management | Invoicing

Forget spreadsheets and complicated tools. Manage your team's time, work, contacts, and projects – all in one system.

WWW.SCORO.COM

4000 people saw ad A, a different 4000 saw ad B. 78 people clicked on ad A (1.95%), and less than half of that number (35) clicked on ad B (0.875%); a difference of $\hat{p}_A - \hat{p}_B = 1.075\%$. What's a range of plausible values for the difference in true click-thru rates, and does this suggest one ad is better than the other?

Intuition regarding comparing two probabilities

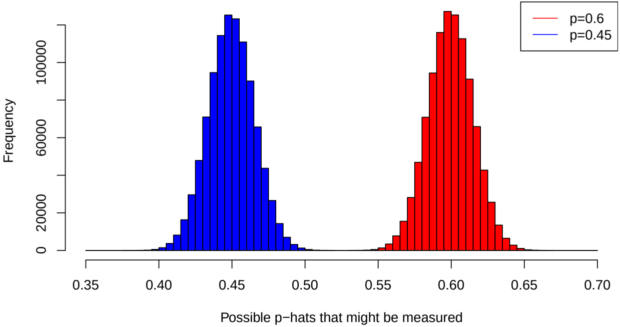
Often when we are compare the probabilities of two events, we don't actually care about the value of each individual event's probability. Instead, we care about whether there's a difference between the two, and if there is, which is larger/smaller.

- $\hat{p}_A = 0.0195$ is "close" to the true click-thru rate of ad A, but p_A is not directly relevant to our question
- $\hat{p}_B = 0.00875$ is "close" to the true click-thru rate of ad B, but p_B is not directly relevant to our question either
- $\hat{p}_A - \hat{p}_B = 0.01075$ is "close" to the true difference in click-thru rates, and *this* is what interests us. Is this a big enough difference to suggest ad A is better than B?

We need to ask ourselves: "how far is the estimated difference in probabilities $(\hat{p}_A - \hat{p}_B)$ from the true difference $(p_A - p_B)$?"

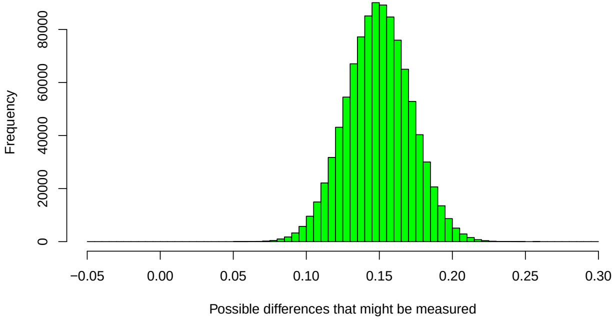
Intuition regarding comparing two probabilities

Let event A have $p_A = 0.60$ and event B have $p_B = 0.45$ so that $p_A - p_B = 0.15$. Let's have 1 million people each run a Monte Carlo simulation (with 1000 trials) and calculate $\hat{p}_A$, then have each run another and calculate $\hat{p}_B$. What possible $\hat{p}_A$'s and $\hat{p}_B$'s might we measure when we run Monte Carlo simulations (say each has 1000 trials)?



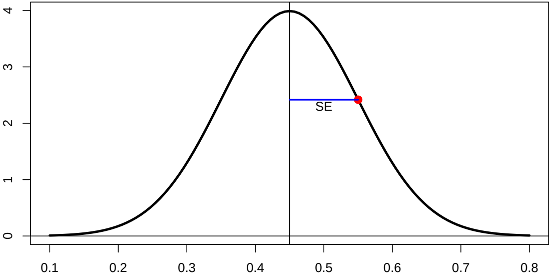
Intuition regarding comparing two probabilities

What about the possible *differences* $\hat{p}_A - \hat{p}_B$ each of these 1 million people might measure?



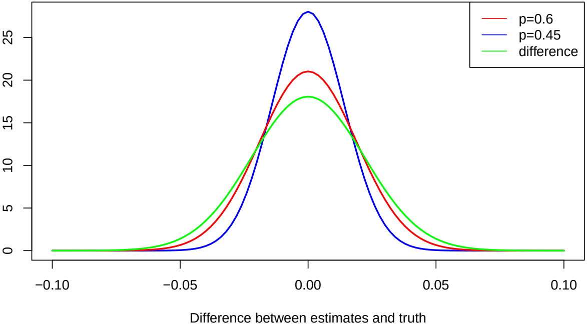
Intuition regarding comparing two probabilities

The bell-curve strikes again! Let's measure the "standard error" as the distance from the peak to the "inflection point" where the curve goes from sloping downward to sloping outward.



Intuition regarding comparing two probabilities

Let's change scale a little bit and look at how far off $\hat{p}_A$ might be from p_A, how far off $\hat{p}_B$ might be from p_B, and how far off $\hat{p}_A - \hat{p}_B$ might be from $p_A - p_B$.



Intuition regarding comparing two probabilities

Observations:

- The distribution of possible $\hat{p}_A - \hat{p}_B$ we might observe closely resembles a bell curve (much like the distribution of possible $\hat{p}_A$'s or possible $\hat{p}_B$'s that we might measure individually).
- The bell curve describing the possible differences is a little wider than the bell curve describing the $\hat{p}_A$'s or $\hat{p}_B$'s, but not quite the sum of those widths.
- Looking at inflection points on the bell-curves, it's clear that the standard error of the difference is related to the standard errors of $\hat{p}_A$ and $\hat{p}_B$, but it's not a simple sum.

SE for difference in two probabilities

Mathematically, or by running just the right set of investigations, it's possible to show that there is a formula that gives the *SE* of the measured difference in estimated probabilities $\hat{p}_A - \hat{p}_B$.

$$SE_{difference} = SE_{\hat{p}_A - \hat{p}_B} \approx \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$

- $\hat{p}_A$, $\hat{p}_B$ are the estimated probabilities from each sample.
- n_A, n_B are sample sizes of each sample (e.g., number of Monte Carlo simulations)

Note: the formula for the *SE* is a "good guess"; the guideline for "good" is to have at least 10 of both kinds of event (e.g. clicks and not-clicks; heads and not heads; etc.) in each sample.

A 95% confidence interval for $p_A - p_B$

Since the bell-curve described the variety in possible differences we might observe, we can use the same logic we used to find a 95% confidence interval for p to get a 95% confidence interval for $p_A - p_B$.

95% confidence interval for $p_A - p_B$

$$\hat{p}_A - \hat{p}_B \pm 2SE$$

$$\hat{p}_A - \hat{p}_B \pm 2\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$

Since 95% of values are within 2 SE of the peak of a bell curve, the difference in $\hat{p}$'s that we end up measuring has about a 95% chance of being within 2 SE of the true difference in p's. Thus, the procedure of taking the observed difference and going $\pm 2SE$ away from it has a 95% chance of yielding an interval that covers the true difference in p.

Master formula for confidence interval for $p_A - p_B$

For other levels of confidence, we'll use the theoretical properties of the bell curve, which can be accessed using the `qnorm` function in R.

Arbitrary confidence interval for p

$$\hat{p}_A - \hat{p}_B \pm \texttt{qnorm(1-(1-confidence)/2)} \cdot \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$

Comparing two probabilities summary ("Classic" procedure)

1. Calculate $\hat{p}_A$ and $\hat{p}_B$, the estimated probabilities from each sample.
2. Note n_A and n_B, the sizes of each sample
3. Calculate
$$SE_{difference} = \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$
4. For a 95% confidence interval:
$$Difference \pm 2SE_{difference}$$
5. For a different level of confidence, the 2 will become `qnorm(1-(1-confidence)/2)` where the confidence is written a decimal between 0-1.

Comparing two probabilities summary

How do we interpret the confidence interval for $p_A - p_B$? In other words, how do we determine whether the data can discern a difference in the two probabilities?

- Check if 0 is inside the confidence interval.
- If 0 is inside, then 0 is a plausible value for the true difference in probabilities. Although it's doubtful that the two events have *equal* probabilities, the data cannot discern a difference between them.
- If 0 is not inside, then 0 is not a plausible for the true difference. If all values are positive, then the data suggests that event A has the higher probability. If all values are negative, then the data suggests that event B has the higher probability.
- Remember that when 0 isn't a plausible value for the difference, we haven't outright *eliminated* the possibility that the true difference is 0. It's possible that the probabilities are equal and the data was just a bit quirky, but the procedure making the confidence interval guards against that from happening very often.

Click-thru rate example

Let's wrap up the click-thru rate comparison example.

- $\hat{p}_A = 78/4000$, $\hat{p}_B = 35/4000$; $n_A = n_B = 4000$; $SE = 0.002635962$

$$SE_{\hat{A}-\hat{B}} \approx \sqrt{\frac{\frac{78}{4000}(1-\frac{78}{4000})}{4000} + \frac{\frac{35}{4000}(1-\frac{35}{4000})}{4000}} = 0.002635962$$

- 95% confidence interval: $Difference \pm 2SE_{difference} \rightarrow (0.005, 0.016)$
- Since 0 is not in the interval, the data is able to discern a difference in true click-thru rates. With all values inside being positive, the data suggests that the click-thru rate of ad A is higher than that of ad B.

```
p.A <- 78/4000; n.A <- 4000; p.B <- 35/4000; n.B <- 4000
SE <- sqrt( p.A*(1-p.A)/n.A + p.B*(1-p.B)/n.B ); SE

## [1] 0.002635962

(p.A-p.B) + c(-1,1)*2*SE

## [1] 0.005478076 0.016021924
```

In R: prop.test

It's easy to make a typo in these computations. The command prop.test in R will do all of them for automatically. **However**, it makes some extra adjustments to the interval (beyond the scope of this class), so it will give slightly different numbers than if you plug-and-chug by hand. I recommend always using prop.test!

```
#1st argument = vector giving # times event of interest occurred for each sample
#2nd argument = vector of sample sizes
prop.test( c(78,35), c(4000,4000), conf.level = 0.95 )

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(78, 35) out of c(4000, 4000)
## X-squared = 15.834, df = 1, p-value = 6.914e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.005333609 0.016166391
## sample estimates:
##  prop 1  prop 2
## 0.01950 0.00875
```

Example: stockout probabilities

One of the motivating examples looked at stockout probabilities of three different policies. Can the data discern any difference in stockout probabilities between Policy 1 and 2?

```
#Policy 1:  12 stockouts in 80 trials; Policy 2:  20 stockouts in 120 trials
prop.test( c(12,20), c(80,200), conf.level = 0.95 )

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(12, 20) out of c(80, 200)
## X-squared = 0.96056, df = 1, p-value = 0.327
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04735576  0.14735576
## sample estimates:
## prop 1 prop 2
##   0.15   0.10
```

Example: stockout probabilities

```
#Policy 1:  12 stockouts in 80 trials; Policy 2:  20 stockouts in 120 trials
prop.test( c(12,20), c(80,200), conf.level = 0.95 )$conf.int

## [1] -0.04735576  0.14735576
## attr(,"conf.level")
## [1] 0.95
```

Since 0 is inside the interval, 0 is a plausible value for the true difference in stockout probabilities based on the Monte Carlo simulations that have been run. Although it's doubtful that these two probabilities are *exactly* equal, the data cannot discern a difference between them.

Aside - In R: prop.test also works for a confidence interval for p

prop.test also works for getting a confidence interval for the probability of single event (again slightly different numbers come out from our formula due to some advanced adjustments it makes). I do not recommend using it and believe binom.test is a better option.

```r
#1st argument = # times event of interest occurred
#2nd argument = sample size
prop.test( 78, 4000, conf.level = 0.95 )

##
##  1-sample proportions test with continuity correction
##
## data:  78 out of 4000, null probability 0.5
## X-squared = 3692.2, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.01554175 0.02440700
## sample estimates:
##      p
## 0.0195
```

Comparing Multiple Probabilities

Comparing Multiple Probabilities

What majors have the "best" and "worst" probabilities of donating within 5 years after they graduate? Different numbers of alumni with each major make comparing the probabilities difficult when looking at a table.

```
ALUMNI <- read.csv("within5years.csv")
table(ALUMNI$Major,ALUMNI$Within5)

##
##                               No  Yes
##   Accounting                4723 1588
##   Business Administration   2245  527
##   Economics                  934  141
##   Finance                   3998  935
##   General Business          1503  313
##   Logistics                  429   80
##   Logistics & Transportation 1532  310
##   Management                1625  346
##   Marketing                 4718  950
##   Transportation/Logistics  1970  288
```

Comparing Multiple Probabilities

Use `aggregate` to quickly determine the fraction of entries that equal "Yes" (in the `Within5` column). This gives the $\hat{p}$'s for each group.

```
aggregate(Within5=="Yes" ~ Major,data=ALUMNI,FUN=mean)

##                        Major Within5 == "Yes"
## 1                 Accounting        0.2516241
## 2    Business Administration        0.1901154
## 3                  Economics        0.1311628
## 4                    Finance        0.1895398
## 5           General Business        0.1723568
## 6                  Logistics        0.1571709
## 7   Logistics & Transportation        0.1682953
## 8                 Management        0.1755454
## 9                  Marketing        0.1676076
## 10   Transportation/Logistics        0.1275465
```

Aside: prop.table is useful

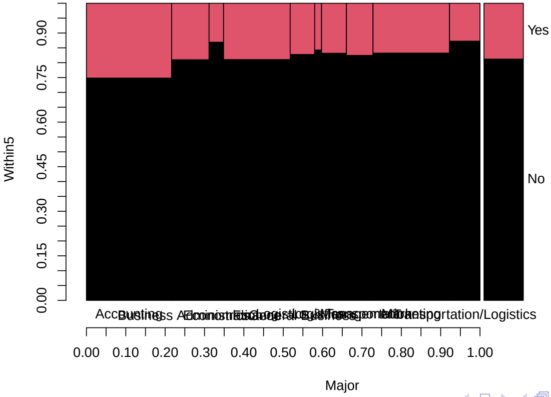
The prop.table command can also be used to get estimated probabilities for each major too.

```
prop.table( table(ALUMNI$Major,ALUMNI$Within5), margin=1 )

## 
##                                       No       Yes
##   Accounting                   0.7483759 0.2516241
##   Business Administration      0.8098846 0.1901154
##   Economics                    0.8688372 0.1311628
##   Finance                      0.8104602 0.1895398
##   General Business             0.8276432 0.1723568
##   Logistics                    0.8428291 0.1571709
##   Logistics & Transportation   0.8317047 0.1682953
##   Management                   0.8244546 0.1755454
##   Marketing                    0.8323924 0.1676076
##   Transportation/Logistics     0.8724535 0.1275465
```

Visualize the differences in proportions using mosaic

A **mosaic plot** is particularly useful for visually comparing groups to each other. In the `regclass` library, `mosaic` can be used.

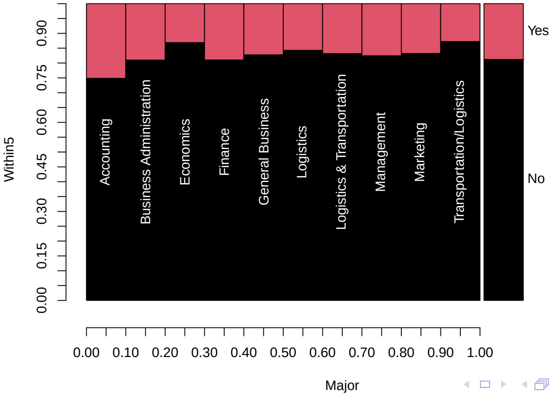
```
library(regclass)
mosaic(Within5 ~ Major,data=ALUMNI)
```



Options to mosaic

There are additional arguments you can give to make the plot more readable. By default, widths of bars are proportional to the number of observations in that group. Having equal=TRUE increases readability but hides that information.

```
mosaic(Within5 ~ Major,data=ALUMNI,inside=TRUE,equal=TRUE)
```



Comparing Multiple Probabilities

If we want to compare the 10 majors' donation probabilities by getting confidence intervals for each pair, we'll quickly discover a fundamental issue - an unacceptable number of those intervals will "miss" the true difference!

- With 10 majors, there are a total of 45 possible comparisons.
- The procedure constructing the confidence interval produces a range of values that covers the true difference 95% of the time (implying that the interval *won't* cover the difference 5% of the time).
- So 5% of the 45 confidence intervals we'd produce will "miss" the true difference. This works out to be $0.05 \times 45 = 2.25$.
- Comparing 10 groups, we expect about 2 confidence intervals to "miss" the true difference by design. Analogously, if we compared all 75 majors at UT, there would be 2775 comparisons, and about 139 would be expected to "miss" the true difference.

Fundamental Issue When Comparing Multiple Probabilities Illustration

Without adjusting the confidence intervals, too many will “miss” the true difference. Let's illustrate with an example.

Imagine 100 different kinds of coins (pennies, nickels, from all around the world). They each have a 50% chance of coming up heads. Let's flip each 500 times. Then, we'll construct confidence intervals for the difference in probabilities for each pair (4950 confidence intervals in all).

How many confidence interval will be “wrong” (0 is not inside the interval)?

Fundamental Issue When Comparing Multiple Probabilities Illustration

```r
#Make matrix where each row is the sequence of 500 flips
FLIPS <- matrix( sample(c(0,1),size=100*500,replace=TRUE), nrow=100, ncol=500)
#Loop through all pairs of coins, increase counter if CI was wrong
counter <- 0
for( coin1 in 1:99 ) {
  for (coin2 in (coin1+1):100 ) {
    CI <- prop.test( c(sum(FLIPS[coin1,]), sum(FLIPS[coin2,])),
                     c(500,500), conf.level = 0.95 )
    if( CI$conf.int[1] < 0 & 0 < CI$conf.int[2] ) { next } else { counter <- counter + 1 }
  }}
counter

## [1] 157

counter/4950

## [1] 0.03171717
```

157 confidence intervals got it wrong, since 0 was not inside the interval. That's about 0.032 of the 4950 confidence intervals (we expected about 0.05).

Adjusting each confidence interval comparing pairs of groups

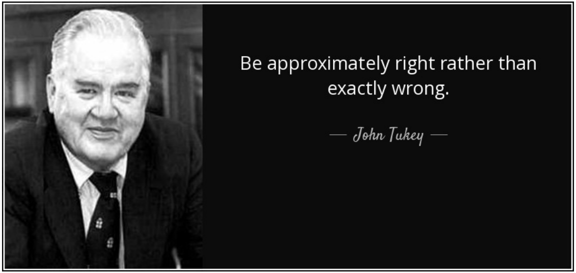
Whenever we we want to compare probabilities among n_{groups} groups, there are a total of $0.5n_{groups}(n_{groups} - 1)$ comparisons that can be made.

We need to make each confidence interval *wider* so that, as a *family*, there is a 95% chance that the *collection* only has at least one confidence interval that "misses" the true difference.

Without an adjustment, expect $0.025n_{groups}(n_{groups} - 1)$ confidence intervals to miss the true difference by design!

Tukey to the rescue!

John Tukey was a prolific statistician (and inventor of the word "bit" and arguably "software" in computer science) that invented many useful tools during the mid to late 1900s.



Be approximately right rather than exactly wrong.

— John Tukey —

Tukey's Honest Significant Difference

Tukey's **Honest Significant Difference** provides a basis to adjust the collection of confidence intervals comparing two groups so that the *collection* of intervals over all pairs will cover the set of true differences in p's about 95% of the time.

The mechanics of where the adjustment comes from is well-beyond the score of this course (PhD maths!), but we'll use his procedure whenever we need to compare all pairs of groups in some statistical way (so we'll see him again when we compare multiple groups' averages).

Setting up Tukey

The output is verbose with so many groups, but you can scan down the list and see that 0 is outside many intervals. Looks like there are noticeable differences in donation habits across majors!

```
AOV <- aov(Within5 == "Yes" ˜ Major, data=ALUMNI) #Compare proportions of Within5 that equal "Yes"
TUKEY <- TukeyHSD(AOV)
## TUKEY
## #                                                      diff    lwr    upr p adj
## #Business Administration-Accounting                  -0.062 -0.090 -0.033 0.000
## #Economics-Accounting                                -0.120 -0.161 -0.080 0.000
## #Finance-Accounting                                  -0.062 -0.085 -0.039 0.000
## #General Business-Accounting                         -0.079 -0.112 -0.047 0.000
## #Marketing-Logistics & Transportation                -0.001 -0.034  0.032 1.000
## #Transportation/Logistics-Logistics & Transportation -0.041 -0.079 -0.002 0.029
## #Marketing-Management                                -0.008 -0.040  0.024 0.999
## #Transportation/Logistics-Management                 -0.048 -0.086 -0.010 0.003
## #Transportation/Logistics-Marketing                  -0.040 -0.071 -0.009 0.001
```

Tukey's Analysis for Alumni data

Some insights:

- Business Administration-Accounting goes from -0.090 to -0.033, so 0 is not a plausible value for the true difference. The data can discern a difference in donation probabilities between these two majors (probability for Accounting majors is somewhere between 3.3% and 9.0% higher than for Business Administration).
- Marketing-Management goes from -0.040 to 0.024, or -4% to 2.4%. Since 0 is inside the interval, it is a plausible value for the true difference. Our data can't discern any difference in the donation probabilities of these two majors.

Tukey's Analysis for Alumni data

At this point it's helpful to once again look at a table of the proportions, ordered from smallest to biggest.

```
ALL <- aggregate( Within5 == "Yes" ~ Major, data=ALUMNI, FUN=mean )  #Get p for each group
ALL[order(ALL[,2],decreasing=TRUE),] #Order from largest to smallest

##                          Major Within5 == "Yes"
## 1                   Accounting        0.2516241
## 2      Business Administration        0.1901154
## 4                      Finance        0.1895398
## 8                   Management        0.1755454
## 5             General Business        0.1723568
## 7   Logistics & Transportation        0.1682953
## 9                    Marketing        0.1676076
## 6                    Logistics        0.1571709
## 3                    Economics        0.1311628
## 10    Transportation/Logistics        0.1275465
```

Connecting Letters Report for Alumni

Summarizing which groups having differences over 45 comparisons is a daunting task. A tool called the **connecting letters report** is helpful.

```
library(multcompView)
multcompLetters4(AOV,TUKEY)

## $Major
##                 Accounting    Business Administration
##                        "a"                        "b"
##                    Finance                 Management
##                        "b"                       "bc"
##           General Business Logistics & Transportation
##                       "bc"                       "bc"
##                  Marketing                  Logistics
##                       "bc"                      "bcd"
##                  Economics   Transportation/Logistics
##                       "cd"                        "d"
```

Connecting Letters Report for Alumni

To interpret:

- Groups are ordered from largest estimated probability (left, letters earlier in alphabet) to smallest (right, letters later in alphabet).
- If two groups share a letter beneath them (e.g., Finance and Logistics), then a confidence interval for the difference in probabilities will contain 0. The data can't discern a difference in probabilities between these two groups.
- If two groups do not share a letter beneath them (e.g., Accounting and Finance), then a confidence interval for the difference in probabilities will not contain 0. The data suggests that whichever group has the letter that is first alphabetically has the higher probability.

Connecting Letters Report for Alumni

```
multcompLetters4(AOV,TUKEY)

## $Major
##                 Accounting    Business Administration
##                        "a"                        "b"
##                    Finance                 Management
##                        "b"                       "bc"
##           General Business Logistics & Transportation
##                       "bc"                       "bc"
##                  Marketing                  Logistics
##                       "bc"                      "bcd"
##                  Economics   Transportation/Logistics
##                       "cd"                        "d"
```

Lessons here:

- Accounting has a larger probability than all the majors (only "a").
- Most other majors have similar probabilities to each other ("b"), except for Transportation/Logistics who has the lowest probability along with Economics (both "d").
- The summary here isn't perfect: it's rare that groups separate into clusters easily.

Connecting Letters Report for Stockout Policies

Three inventory control policies. In a Monte Carlo simulation, policy 1 had a stockout in 12 out of 80 trials, policy 2 had a stockout in 20 of 200 trials, and policy 3 had a stockout in 7 of 50 trials.

```
DATA <- data.frame(Policy = factor( c(rep(1,80), rep(2,200), rep(3,50) ) ),
     Stockout = factor( c(rep("Y",12),rep("N",68),rep("Y",20),rep("N",180),rep("Y",7),rep("N",43))) )
aggregate(Stockout == "Y" ~ Policy, data=DATA, FUN=mean)

##   Policy Stockout == "Y"
## 1      1            0.15
## 2      2            0.10
## 3      3            0.14
```

Connecting Letters Report for Stockout Policies

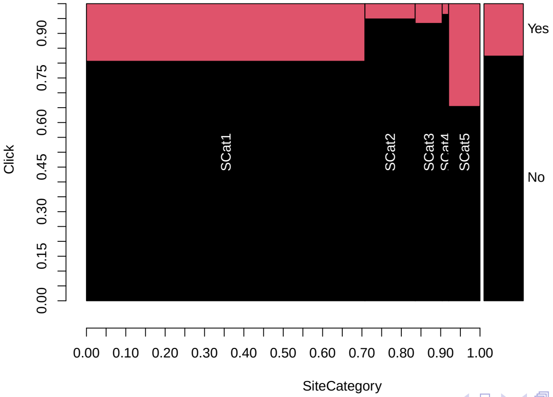
```
AOV <- aov(Stockout == "Y" ~ Policy, data=DATA)
TUKEY <- TukeyHSD(AOV)
TUKEY

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Stockout == "Y" ~ Policy, data = DATA)
##
## $Policy
##      diff         lwr        upr     p adj
## 2-1 -0.05 -0.15075625 0.05075625 0.4729461
## 3-1 -0.01 -0.14730763 0.12730763 0.9839220
## 3-2  0.04 -0.08042676 0.16042676 0.7143395
```

The confidence intervals for all pairs include 0. Thus, 0 is a plausible value for the difference in stockout probabilities for all pairs. If there is a difference in stockout probabilities, our simulations weren't extensive enough to detect them.

Connecting Letters Report for Click-Thru Rates

Click-thru rate from BAS 320 dataset. How do click-thru rates compare based on the site category (news, humor, etc.) that hosts the ad?

```
library(regclass); data("EX6.CLICK")
mosaic( Click ~ SiteCategory, data=EX6.CLICK, inside=TRUE )
```



Connecting Letters Report for Click-Thru Rates

Summarize each site category with a table of estimated probabilities.

```
aggregate(Click=="Yes" ~ SiteCategory, data=EX6.CLICK, FUN=mean)

##   SiteCategory Click == "Yes"
## 1        SCat1     0.19386694
## 2        SCat2     0.05132641
## 3        SCat3     0.06638116
## 4        SCat4     0.03603604
## 5        SCat5     0.34594096
```

Connecting Letters Report for Click-Thru Rates

```
AOV <- aov(Click=="Yes" ~ SiteCategory, data=EX6.CLICK)
TUKEY <- TukeyHSD(AOV)
TUKEY

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Click == "Yes" ~ SiteCategory, data = EX6.CLICK)
##
## $SiteCategory
##                     diff         lwr         upr     p adj
## SCat2-SCat1 -0.14254053 -0.16914582 -0.11593524 0.0000000
## SCat3-SCat1 -0.12748579 -0.16243627 -0.09253530 0.0000000
## SCat4-SCat1 -0.15783091 -0.22705914 -0.08860267 0.0000000
## SCat5-SCat1  0.15207402  0.11940196  0.18474607 0.0000000
## SCat3-SCat2  0.01505474 -0.02633574  0.05644523 0.8589396
## SCat4-SCat2 -0.01529038 -0.08798278  0.05740202 0.9788903
## SCat5-SCat2  0.29461455  0.25512911  0.33409998 0.0000000
## SCat4-SCat3 -0.03034512 -0.10648893  0.04579869 0.8132387
## SCat5-SCat3  0.27955980  0.23403185  0.32508775 0.0000000
## SCat5-SCat4  0.30990492  0.23477965  0.38503020 0.0000000
```

Connecting Letters Report for Click-Thru Rates

```
multcompLetters4(AOV,TUKEY)

## $SiteCategory
## SCat5 SCat1 SCat3 SCat2 SCat4
##   "a"   "b"   "c"   "c"   "c"
```

Looks like category 5 has the highest click-thru rate, followed by category 1, and category 2/3/4 are all tied for last. Rare example where there is a definitive ranking!

Summary

Summary regarding inference on p

The $\hat{p}$ we measure during a Monte Carlo simulation or by analyzing data isn't expected to equal p. It might be a little too high, or maybe too low, but it will be "close".

The standard error is our best guess for how far off $\hat{p}$ will be from p. Amazingly, there is a formula that gives its value:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The standard error decreases "slowly" with n: increasing n by a factor of 4 only halves the SE. Increasing n by a factor of 25 decreases the SE by a factor of 5.

The size of the standard error is also tied in to what we estimated from the sample. The SE is at its largest when $\hat{p} = 0.5$.

Summary regarding inference on p

We can take $\hat{p}$ and the value of the SE to come up with a 95% confidence interval for p.

$$\hat{p} \pm 2SE \qquad\qquad \hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- The confidence interval gives a range of plausible values for p based on our data.
- p may or may not be in the interval you construct (and you'll never know if your intervals covers p)
- "In the long run", 95% of the time, your 95% confidence interval will cover p
- The "confidence" in an interval is the probability that the procedure that generated the interval ends up producing a range of values that covers p.
- NEVER say "the probability that p is in the interval is 95%". p is some fixed value; we just don't know it. We don't use probability to quantify our ignorance about the world.

Summary regarding inference on p

In R, there are 3 ways to get a confidence interval for p: math it yourself, use `prop.test`, or use `binom.test`. `binom.test` is *strongly* preferred! However, for larger values of n, the three intervals are very similar.

```
n.events <- 323;  ntrials <- 1034  #For example, event occured 323 out of 1034 trials
p.hat <- n.events/ntrials; p.hat

## [1] 0.3123791

SE <- sqrt( p.hat*(1-p.hat)/ntrials); SE

## [1] 0.01441303

p.hat + c(-1,1)*2*SE

## [1] 0.2835531 0.3412052

prop.test( n.events, ntrials )$conf.int[1:2]

## [1] 0.2843984 0.3417724

binom.test( n.events,ntrials )$conf.int[1:2]

## [1] 0.2842143 0.3416164
```

Summary regarding difference in probabilities $p_A - p_B$

Likewise, the difference in estimated probabilities we measure between two samples $\hat{p}_A - \hat{p}_B$ won't equal the true difference in probabilities, but it should be "close".

The formula for the standard error is more complex, but it still gives our best guess for how far off measured difference in probabilities will be from the true difference.

$$SE_{difference} = SE_{\hat{p}_A - \hat{p}_B} \approx \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$

A 95% confidence interval for the difference is still:

$$Difference \pm 2SE_{difference}$$

Summary regarding difference in probabilities $p_A - p_B$

Because of the relatively involved calculation for the *SE*, it's often "safer" to use `prop.test` instead of mathing it out manually.

```
n.A <- 1000;  n.B <- 2000  #number of trials of each group
events.A <- 563;  events.B <- 1233
phat.A <- events.A/n.A; phat.B <- events.B/n.B; phat.A; phat.B

## [1] 0.563
## [1] 0.6165

SE <- sqrt( phat.A*(1-phat.A)/n.A + phat.B*(1-phat.B)/n.B ); SE

## [1] 0.0190852

phat.A - phat.B + c(-1,1)*2*SE

## [1] -0.0916704 -0.0153296

prop.test( c(events.A,events.B), c(n.A,n.B) )$conf.int[1:2]

## [1] -0.09165631 -0.01534369
```

Summary comparing multiple probabilities

With k groups, there are a total of $0.5k(k-1)$ comparisons that can be made.

Because each confidence interval is designed to cover the true difference in probabilities 95% of the time, it is also designed to "fail" 5% of the time (not cover the true difference). After a large number of comparison, this means a lot of confidence intervals won't cover the true difference.

Each confidence interval needs to include more values (get wider) so that the *collection* of confidence intervals as a whole all cover the set of true differences 95% of the time.

Summary comparing multiple probabilities in R

In R, it is a multistep process to obtain the confidence intervals. For illustration, imagine a dataframe called DATA has columns Event (the value whose probability you want to compare across groups be called Yes) and Group (containing group identities).

```
#Table of p-hats; chance column and dataframe names, and change "Yes" to whatever
#encodes your event of interest, e.g. "Click", "Donate", etc.
aggregate(Event=="Yes"~Group,data=DATA,FUN=mean)
```

Summary comparing multiple probabilities in R

```
AOV <- aov( Event=="Yes" ~ Group, data=DATA )  #Setup
TUKEY <- TukeyHSD(AOV) #Get all confidence intervals for pairs
TUKEY # print to screen
multcompLetters4(AOV,TUKEY)  #Produce the connecting letters reports
```

Summary Connecting Letters Report

Interpreting connecting letters:

```
Group        TN    AL     GA    NC    KY     MS
Letters      "a" "abc"   "a"  "ab"  "bc"    "c"
```

- Letters earlier in the alphabet have higher estimated probabilities, and the list is ranked from left to right. Here, we know that *TN* has the largest estimated probability while *MS* has the lowest (check the `aggregate` command for numbers)
- If two groups share the same letter (e.g. TN and AL, KY and MS, AL and KY), the data cannot discern a difference in probabilities between them, i.e., the confidence interval for the difference in probabilities includes 0.
- If two groups don't share the same letter (e.g., TN and KY, NC and MS), the confidence interval for the difference in probabilities doesn't includes 0. The group with the earlier letter alphabetically has a higher probability (TN > KY, NC > MS)
- Sometimes an easy/coherent story emerges. Most of the time it's complicated!