CHAPTER 3

Simple Exposure Analysis

1. Facebook Insights

Facebook Insights. If you own a facebook page and you have 30 “likes” then you get access to facebook insights.

Facebook believes these Insights are useful - particularly to business.

The School of Computing, Engineering and Mathematics has a facebook page.

Likes, Reach and Talking About. Facebook Insights records data (over a fixed period) relating to the number of likes, the reach and the number “Talking About”

This can be at the *page* or *post* level

- Likes are just the number of unique people who click the *like* button on a page or post.
- Reach is the number of unique people who might have seen a page or post. It includes likes, but also includes people who have seen it because it was *shared* etc.
- “Talking about” means actively interacting with a page or post. That is likes, comments, tags or shares etc.

These measures can be obtained for various time periods, daily, weekly, 28-day, as well as a cumulative total.

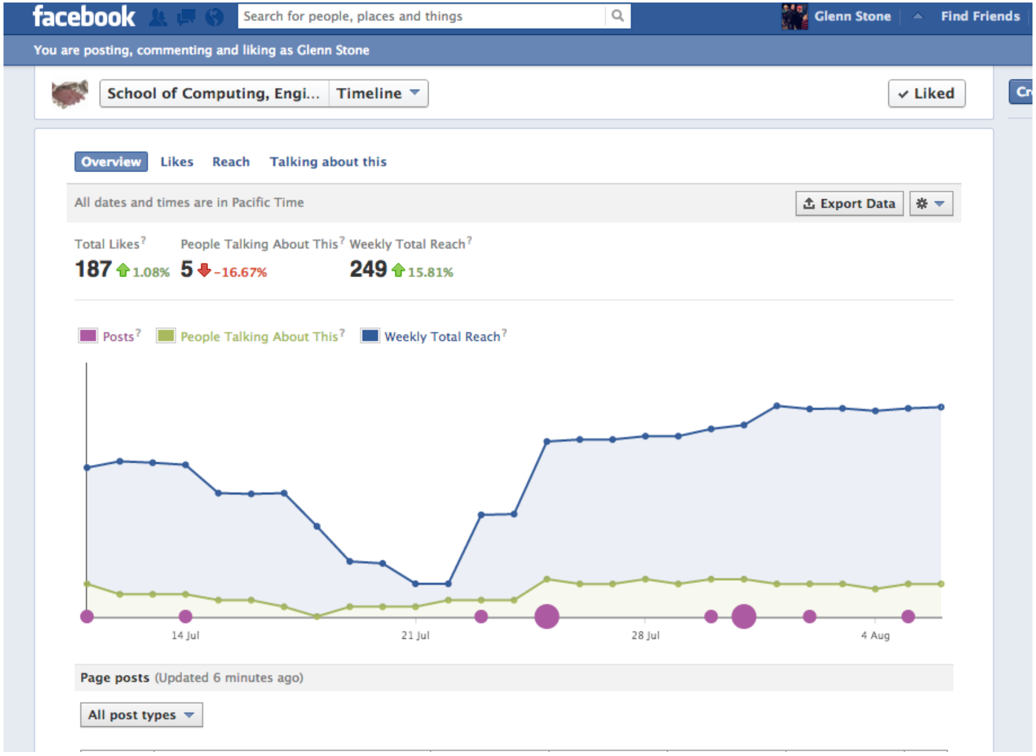
The measures are only available for pages that 30 or more of that measure.

Business Questions. The sort of things an organisation might be interested in include.

- What is the Reach of our facebook presence?
- What are the demographics of that Reach — are we getting to our target audience?
- What impacts have changes to our page/posts/presence made on Reach?

Facebook provide graphical visualisations of this data. We will look at replicating this, and doing some simple statistical analysis.

Facebook Insights.



You are posting, commenting and liking as Glenn Stone

All dates and times are in Pacific Time

Total Likes	People Talking About This	Weekly Total Reach
187 ⬆1.08%	5 ⬇-16.67%	249 ⬆15.81%

Page posts (Updated 6 minutes ago)

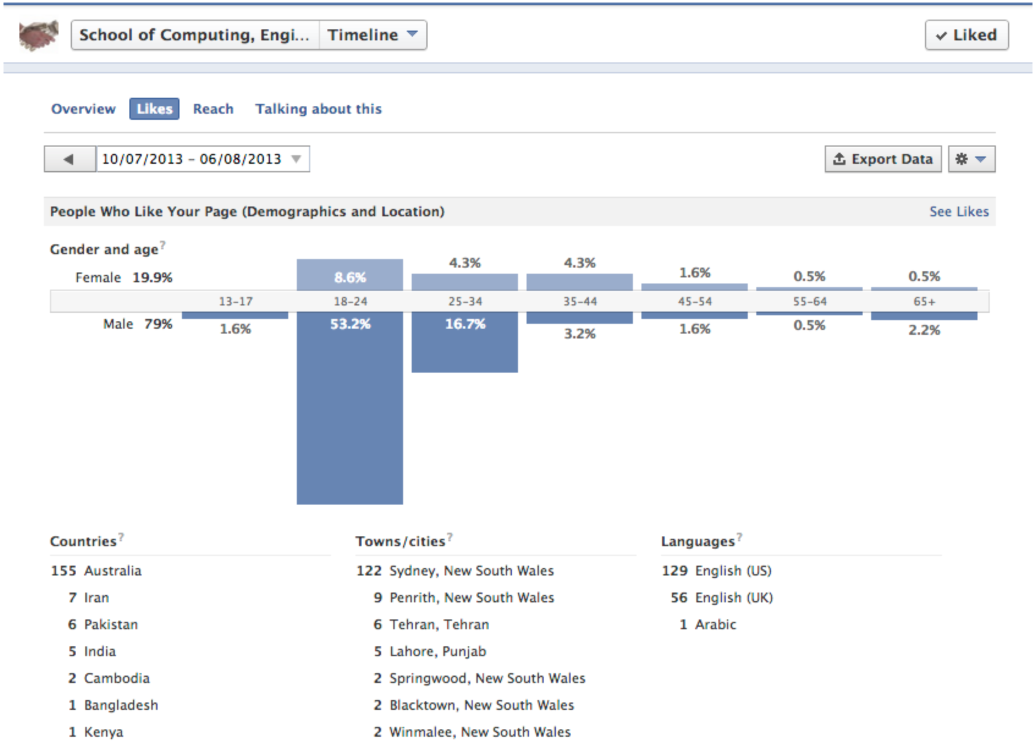
Facebook Insights.

- The overview screen shows the Reach (and “Talking About”) of the Page
- The figures at the top are current Likes, and cumulative TA and Reach for the last 7 days.
- The graph shows the TA and Reach for 7 day period ending on a particular day, plotted against that day.
- The purple circles represent the number of posts each day.

This is all for the past month.

(There is a table of per post info also)

Facebook Insights.



Countries	
155	Australia
7	Iran
6	Pakistan
5	India
2	Cambodia
1	Bangladesh
1	Kenya

Towns/cities	
122	Sydney, New South Wales
9	Penrith, New South Wales
6	Tehran, Tehran
5	Lahore, Punjab
2	Springwood, New South Wales
2	Blacktown, New South Wales
2	Winmalee, New South Wales

Languages	
129	English (US)
56	English (UK)
1	Arabic

Facebook Insights. On the second screen (its the same for Likes, Reach or Talking About), the Likes are broken down by several demographic factors

- Gender, Age, Country, Town, and Language.

The graph is a *bar plot* of proportions by gender and age.

You can choose the time period that this covers, up to a 92 day period.

You can also Export the data for 180 days.

Export Data. The small button to the right of this second screen allows us to export data.

- Data can be exported at a page or post level.
- Dates can be chosen up to 180 days long.
- XLS or CSV format can be used. XLS contains more information.

We will look at Page level, XLS data.

As exported by Facebook there are sometimes problems with this file, so open it in Excel and re-save as an XLSX file.

Export Data. R has some facilities to read XLSX files. The readxl library is one. (This will need to be installed)

```
library(readxl, quietly=TRUE)
```

```
sheet=Excel_sheets(”Facebook Insights Data Export 2013-08-07.xlsx”)

length(sheets)

## [1] 38

sheets

##  [1] ”Key metrics”                     ”Daily Like sources”
##  [3] ”Daily Viral Reach by story type” ”Weekly Viral Reach by story...”
##  [5] ”28 days Viral Reach by story...” ”Daily Viral Impressions by s...”
##  [7] ”Weekly Viral Impressions by...”  ”28 days Viral Impressions by...”
##  [9] ”Daily Total frequency distri...” ”Weekly Total frequency distr...”
## [11] ”28 days Total frequency dist...” ”Daily Page posts frequency d...”
## [13] ”Weekly Page posts frequency...”  ”28 days Page posts frequency...”
## [15] ”Daily Viral frequency distri...” ”Weekly Viral frequency distr...”
## [17] ”28 days Viral frequency dist...” ”Daily Talking About This by...”
## [19] ”Weekly Talking About This by...” ”28 days Talking About This b...”
## [21] ”Daily Page stories by story...”  ”Weekly Page stories by story...”
## [23] ”28 days Page stories by stor...” ”Daily People who interacted...”
## [25] ”Weekly People who interacted...” ”28 days People who interacte...”
## [27] ”Daily Page consumptions by type” ”Weekly Page consumptions by...”
## [29] ”28 days Page consumptions by...” ”Lifetime Likes by gender and...”
## [31] ”Lifetime Likes by country”      ”Lifetime Likes by City”
## [33] ”Lifetime Likes by Language”     ”Weekly Reach Demographics”
## [35] ”Weekly Reach by Country”        ”Weekly Reach by City”
## [37] ”Weekly Reach by Language”       ”Daily Demographics People Ta...”
```

Export Data. The XLSX file contains 38 sheets. We are primarily going to look at “Key Metrics” and “Weekly Reach by Demographics”

(These are saved as CSV for ease of use, but...)

```
keyMetrics <- read.xls(”Facebook Insights Data Export 2013-08-07.xlsx”, 1)
WeekReach <- read.xls(”Facebook Insights Data Export 2013-08-07.xlsx”, 34)

keyMetrics <- read.csv(”keyMetrics.csv”, as.is=TRUE)
dim(keyMetrics)

## [1] 161  90

WeekReach <- read.csv(”WeeklyReachDemog.csv”)
dim(WeekReach)

## [1] 160  21
```

The “as.is=TRUE” prevents the key metrics being treated as factors. (see later)

2. Reach and Demographics

Demographics. Demographics refers to the characteristics of individuals. In Facebook, the most useful demographics are Age group and Gender. City, Country and Language are also recorded.

Several sheets in the XLSX file have breakdown by demographics.

We will look at “Weekly Reach Demographics”

Reach and Demographics. Reading in the CSV file gives the following. At the moment, we are interested in a particular (recent) date.

```
xx <- WeekReach[158,]
print(xx)
```

```
##     Description   Date F.13.17 F.18.24 F.25.34 F.35.44 F.45.54 F.55.64
## 158             8/5/13       1      45      15       6       2       3
##     F.65. M.13.17 M.18.24 M.25.34 M.35.44 M.45.54 M.55.64 M.65. U.18.24
## 158     2       2     121      31       7       4       2     3       1
##     U.25.34 U.35.44 U.45.54 U.65.
## 158      NA       1       1    NA
```

Reach and Demographics. First we make this into a table (matrix) after discarding the Date, and “U” categories. And we set up meaningful row and columns names

```
tab <- matrix(as.numeric(xx[3:16]), nrow=2, byrow=TRUE)
colnames(tab) <- c(”13-17”, ”18-24”, ”25-34”, ”35-44”,
                   ”45-54”, ”55-64”, ”65+”)
rownames(tab) <- c(”Female”,”Male”)
print(tab)
```

```
##        13-17 18-24 25-34 35-44 45-54 55-64 65+
## Female     1    45    15     6     2     3   2
## Male       2   121    31     7     4     2   3
```

Reach and Demographics. So we can draw a graph similar to the Facebook one (bar plot)

```
barplot(tab, legend=TRUE, col=c(”pink”,”lightblue”))
```

Reach and Demographics. However, side by side bars are sometimes easier to compare.

```
barplot(tab, legend=TRUE, col=c(”pink”,”lightblue”),
        beside=TRUE)
```

Questions of interest. Business (or page owners) might be interested to know...

- What proportion of Reach is ... eg. Male or 18 to 24? If the page, represents a product or service you might be interested in whether you are reaching Males more than Females?
- Is this proportion changing? (see later lecture)
- Are the age *profiles* different for males and females? Is there something different about the age profiles that you are reaching?

These are statistical questions...provided we are prepared to assume that the reach is a *random sample* of all possible Reach. (Discuss?)

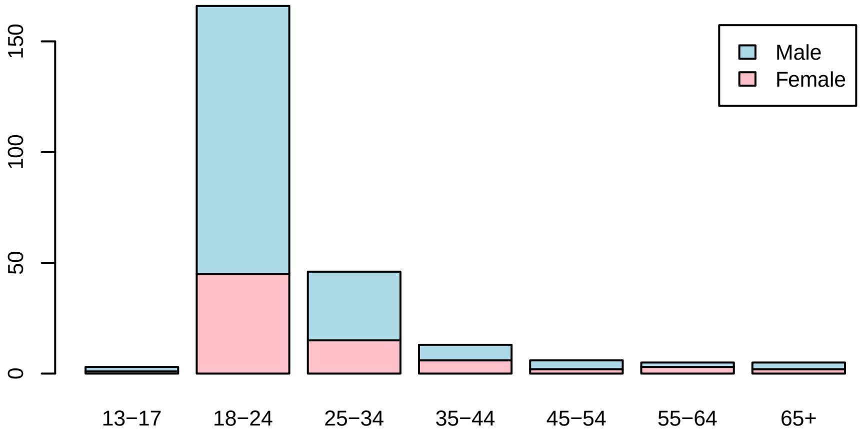


FIGURE 1. plot of chunk unnamed-chunk-7

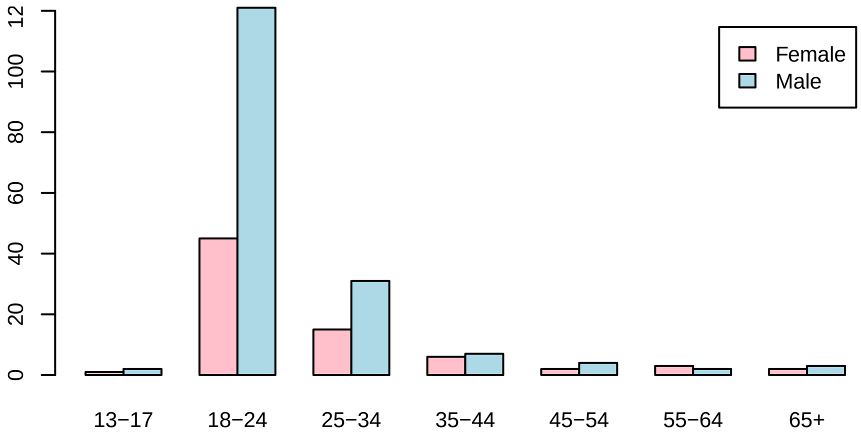


FIGURE 2. plot of chunk unnamed-chunk-8

Confidence interval for proportion.

Proportions. The simple (point) estimate of the proportion in category XYZ, is simply the number in that category divided by the total.

- Number of males $(2 + 121 + 31 + 7 + 4 + 2 + 3) = 170$
- Number of females $(1 + 45 + 15 + 6 + 2 + 3 + 2) = 74$,

- Total = 170+74 = 244

Therefore, our estimate of the proportion of males is $\hat{p} = 170/244 = 0.697$ or 69.7%

But note that this is an *estimate* of the true proportion p. Providing only the estimate $\hat{p}$ does not tell us how different $\hat{p}$ is to p.

Variance in sample proportions.

Example. We want to estimate the proportion of 1s rolled when using a six sided dice. In this experiment, we know that $p = 1/6$ (assuming the dice is fair). After rolling the dice $n = 10$ times, we obtain the sample:

6, 2, 4, 3, 1, 3, 4, 3, 3, 6

Giving $\hat{p} = 0.1$.

After rolling the dice another 10 times:

6, 6, 6, 1, 3, 4, 5, 1, 4, 5

Giving $\hat{p} = 0.2$.

Sample proportion distribution. If we repeat the experiment 1000 times, we get the following distribution. If $p = 1/6$, we can get values of $\hat{p}$ from 0 to 0.7, when $n = 10$.

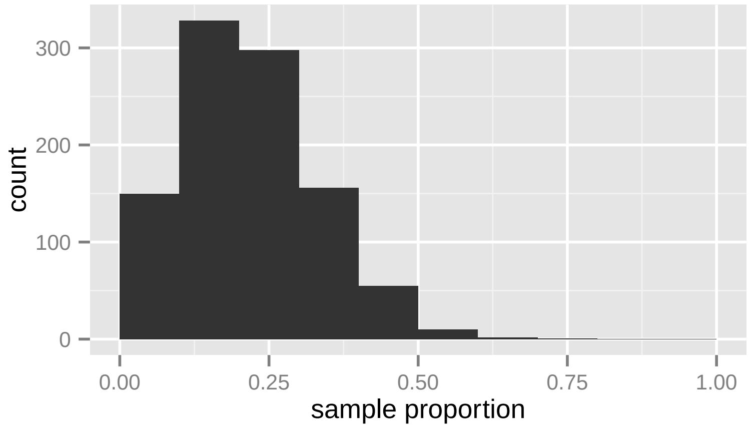


FIGURE 3. Distribution of sample proportion.

Confidence Interval for Proportion. We just showed that if we know the proportion p, we can estimate the variation of $\hat{p}$. But when we obtain a sample, we have $\hat{p}$ (not p), so we have the reverse question.

Confidence interval for p

Giving a sample proportion $\hat{p}$ and sample size n, what range of values could p take, with probability $1 - \alpha$?

For a 95% confidence interval:

- We must estimate the variation of $\hat{p}$
- Using bootstrapping, we sample *with replacement* from the original sample and compute the bootstrap statistic $\hat{p}_b$
- Repeat the process many times (at least 1000) to obtain a distribution of the bootstrap statistic.
- The confidence interval for p is the middle 95% of the bootstrap distribution.

Confidence interval for proportion of males. The sample proportion of males $\hat{p} = 0.697$ with $n = 244$.

```
bootDist = replicate(1000, mean(sample(c(”M”,”F”), size = 244,
     prob = c(0.697, 0.303), replace = TRUE) == ”M”))
```

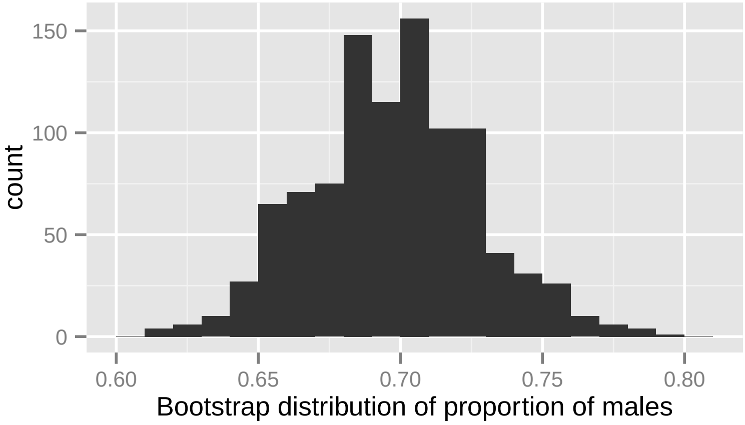


FIGURE 4. Bootstrap distribution of proportion of males.

Confidence interval for proportion of males. The 95% confidence interval is middle 95% of the bootstrap distribution:

```
lower = quantile(bootDist, 0.025)
upper = quantile(bootDist, 0.975)
print (c(lower, upper))
```

```
##      2.5%     97.5%
## 0.6434426 0.7581967
```

Therefore, we are 95%confident that the proportion of males is between 0.6434426 and 0.7581967. The 90% confidence interval is middle 90% of the bootstrap distribution:

```
lower = quantile(bootDist, 0.05)
upper = quantile(bootDist, 0.95)
print (c(lower, upper))
```

```
##        5%       95%
## 0.6516393 0.7459016
```

Problem: confidence intervals.

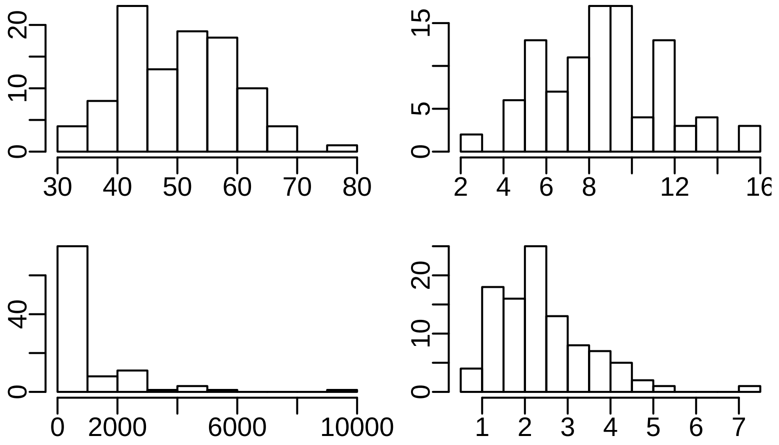


FIGURE 5. Bootstrap distribution of proportion of males.

Problem. Estimate the 95% confidence interval from the following bootstrap distributions.

Chi-squared test for independence.

Age profiles. **Question**

Are the age profiles of the males and females that we are reaching different? That is, neglecting the the difference in overall number is the spread across ages different?

```
print(tab)
```

```
##        13-17 18-24 25-34 35-44 45-54 55-64 65+
## Female     1    45    15     6     2     3   2
## Male       2   121    31     7     4     2   3
```

Is there evidence, that the age profiles of genders in our audience differs?

This is a *hypothesis test*. We want to test if the two variables are independent.

Independent events.

Example. The outcomes of tossing a coin and rolling a dice are independent. Here are the results from tossing a coin and rolling a dice 100 times.

	Dice					
	1	2	3	4	5	6
Head	12	8	6	9	6	4
Tail	8	9	8	10	10	10

If we ran the experiment again, we would get different numbers, but the two variables are still independent.

We know that the dice roll and coin toss are independent. For another experiment where we are unsure of the dependence, how different do the numbers have to be for us to have confidence that the variables are dependent?

Chi-squared statistic. The χ^2 (Chi-squared) statistic can be used to summarise the similarity of a table to the expected "independent" table.

$$\chi^2 = \sum_i \sum_j \frac{(X_{ij} - np_i q_j)^2}{np_i q_j}$$

- X_{ij} as the count of the ith row, jth column,
- n as the sample size (sum of all counts),
- p_i as the expected proportion of ith row and
- q_j as the expected proportion of the jth column.
- $np_i q_j$ is the expected count in cell ij, assuming independence between rows and columns.

If all X_{ij} are equal to the expected $np_i q_j$, $\chi^2 = 0$. The more different X_{ij} is to $np_i q_j$, the greater χ^2.

Chi-squared statistic of Dice and Coin sample.

Example.

	Dice					
	1	2	3	4	5	6
Head	12	8	6	9	6	4
Tail	8	9	8	10	10	10

- q_j for each j is 0.20, 0.17, 0.14, 0.19, 0.16, 0.14
- p_i for each i is 0.45, 0.55
- n = 100
- Giving $\chi^2 = 3.81$

We can compute the χ^2 value of our sample and obtain a number. If the number is large, then the sample table is not similar to the expected independent table, so the sample is probably not independent.

But how "large" is large? Is 3.81 large enough?

Examining the χ^2 distribution for independent tables. To determine what "large enough" means for the χ^2 statistic, we must examine the χ^2 statistic for tables where the random variables are *independent*. If our χ^2 statistic is larger than those, then our table variables are *probably not independent*.

How do we observe other samples with independent rows and columns?

We use randomisation of our table.

Randomisation to break dependence.

Example (continued). If two random events are independent, we can independently shuffle the order of their outcomes without effecting the probability of their joint outcome.

Dice	5	2	2	1	5	1	3	6	4	6
Coin	T	H	H	H	H	H	H	H	H	H

↓ Random shuffle ↓

Dice	1	6	2	5	2	6	3	1	4	5
Coin	H	H	H	H	H	T	H	H	H	H

Note that the above two outcomes are not equivalent if dependence exists between the coin toss and dice roll outcomes.

Observing independent samples. By repeatedly using random shuffling and computing the χ^2 statistic of the new table, we can observe what the dice vs. coin χ^2 statistic will look like when the two are independent.

We now ask "what is the probability that we could obtain our χ^2 statistic or greater, given that the coin and dice are independent?"

- If the probability is low, then the dice and coin are probably not independent (since the shuffled tables are independent).
- If the probability is high, then we can't say anything. We can assume that they are independent.

Computing the p-value.

Example (continued).

```r
expectedIndependent = function(X) {
    n = sum(X)
    p = rowSums(X)/sum(X)
    q = colSums(X)/sum(X)
    return(p %o% q * n) # outer product creates table
}

chiSquaredStatistic = function(X, E) {
    return(sum((X - E)^2/E))
}

E = expectedIndependent(X) # compute expected counts if independent

x2 = replicate(1000, { # compute 1000 randomised chi-squared statistics
  diceShuffle = sample(dice)
  coinShuffle = sample(coin)
  Xindep = table(coinShuffle, diceShuffle)
  chiSquaredStatistic(Xindep, E)
})
```

Chi-squared distribution.

Problem: χ^2 test.

Problem. Our sample produces a χ^2 statistic of 15. Determine if our sample does not belong to the following χ^2 distributions.

Hypothesis test. We just performed a *Hypothesis test* on the dice vs. coin table.

- H_0: The dice and coin outcomes are independent.
- H_A: The dice and coin outcomes are not independent.

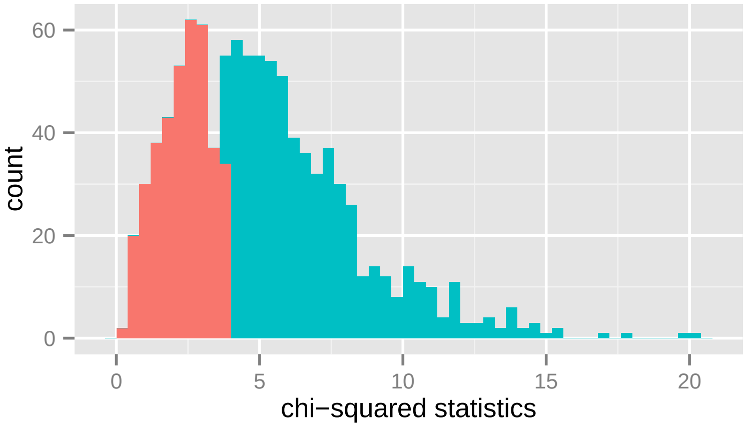


FIGURE 6. Randomisation distribution of the dice vs. coin chi-squared statistic. Blue region is the set of χ^2 values that are greater than our sample χ^2 statistic (3.81).

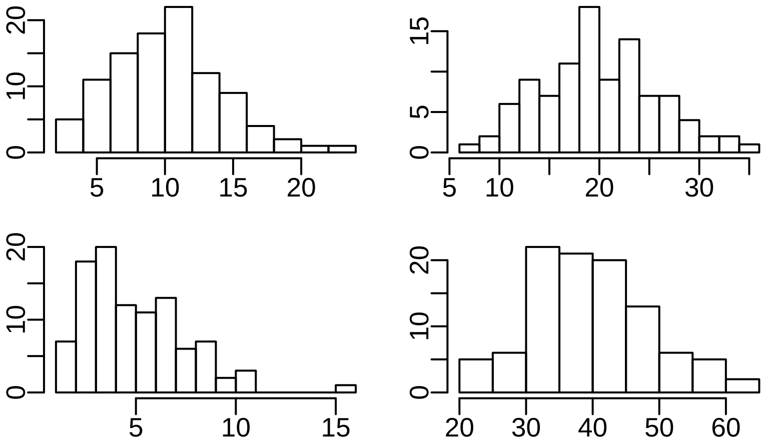


FIGURE 7. Bootstrap distribution of proportion of males.

The p-value for the test is the proportion of blue in the previous histogram (0.568)

Since the p-value is high, we cannot reject H_0.

We can also compute the p-value using the R function:

```
chisq.test(X, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
```

```
##
## data:  X
## X-squared = 3.8067, df = NA, p-value = 0.5902
```

Age profiles: Hypothesis test for independence. We now continue our analysis of age vs. gender.

```
##        13-17 18-24 25-34 35-44 45-54 55-64 65+
## Female     1    45    15     6     2     3   2
## Male       2   121    31     7     4     2   3
```

If we assume (H_0) that age and gender are independent (we get the same distribution of ages no matter what gender we observe), the expected frequencies are (np_iq_j)

```
##            13-17     18-24    25-34    35-44    45-54    55-64      65+
## Female 0.9098361  50.34426 13.95082 3.942623 1.819672 1.516393 1.516393
## Male   2.0901639 115.65574 32.04918 9.057377 4.180328 3.483607 3.483607
```

Giving us a χ^2 value of 4.8116605.

Is the difference between the observed frequencies and expected frequencies large enough for us to say that the age and gender are not independent?

Age profiles: randomisation distribution. Given the below distribution of χ^2 if gender and age are independent, and our sample having $\chi^2 = 4.8116605$, can we say that gender and age are not independent?

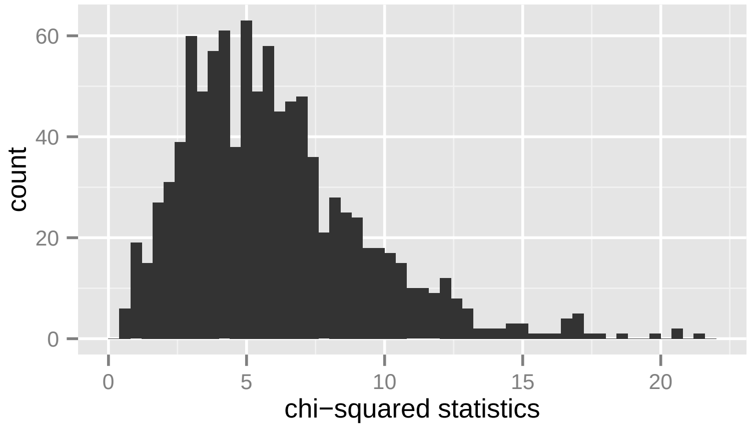


FIGURE 8. Randomisation distribution of the age vs. gender chi-squared statistic.

Age profiles: randomisation distribution. The blue region shows a p-value of 0.598, so we cannot say that gender and age are not independent (cannot reject H_0).

Other uses of the χ^2 test. We computed the χ^2 distribution where the expected frequencies were those generated using np_iq_j. This was to test if the two variables (age and gender) are independent.

Note that if yoU aRE not UsinG a Randomisation distRibUtion foR a χ2 tEst (E.G. UsinG bUilt in fUnction in R chisq.tEst(), thEn ExpEctEd coUnt foR Each cEll mUst bE GREatER than 5.

Note that we can test for any set of expected frequencies. If we want to test if the proportion of males is 60%and females is 40% and the sample size is n, the expected number of males is $0.6n$ and the expected number of females is $0.4n$.

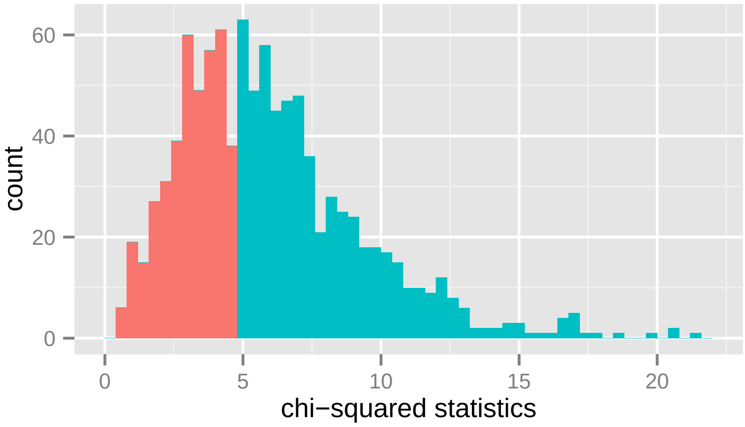


FIGURE 9. Randomisation distribution of the age vs. gender chi-squared statistic. Blue region is the set of χ^2 values that are greater than our sample χ^2 statistic (4.81).

We will go through a problem concerning this in the lab.

3. Key Metrics

Key Metrics. The Key Metrics data has a header row, and a second row that is a description. There are 90 columns: 1,15,18 and 24 are;

- **Date** The Date!
- **Daily Total Reach** Daily number of people who have seen any content associated with your Page. (Unique Users)
- **Daily Organic Reach** Daily number of people who visited your Page, or saw your Page or one of its posts in News Feed or ticker. These can be people who have liked your Page and people who haven't. (Unique Users)
- **Daily Viral Reach** Daily number of people who saw your Page or one of its posts from a story shared by a Friend. These stories include liking your Page, posting to your Page's Timeline, liking, commenting on or sharing one of your Page posts, answering a question you posted, responding to one of your events, mentioning your Page, tagging your Page in a photo or checking in at your location. (Unique Users)

Key Metrics.

```
dates <- keyMetrics[,1]
dates <- dates[-1]
dates <- strptime(dates, format="%m/%d/%y")
reach <- keyMetrics[,15]
reach[1]
```

```
## [1] "Daily The number of people who have seen any content associated with your Page. (Unique Users)"
```

```
reach <- as.numeric(reach[-1])
```

Key Metrics.

```
plot(dates, reach, type="l")
```

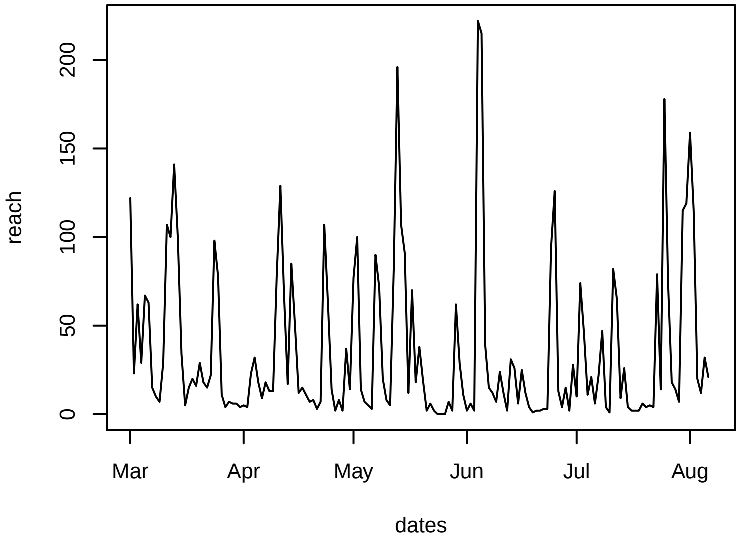


FIGURE 10. Daily Total Reach

Key Metrics.

```
keep <- (dates > as.POSIXlt("2013-07-01"))
plot(dates[keep], reach[keep], type="l")
```

Weekly Total Reach.

```
wreach <- as.numeric(keyMetrics[-1,16])
plot(dates[keep], wreach[keep], type="l")
```

Conclusion. So by exporting the data we can do a lot more than the usual insights provide

- Testing of Reach demographics
- Estimating features of Reach/Likes etc
- Graphing different parameters
- Graphing over different time periods

Next week. Text Mining 1: Indexing and Querying Text

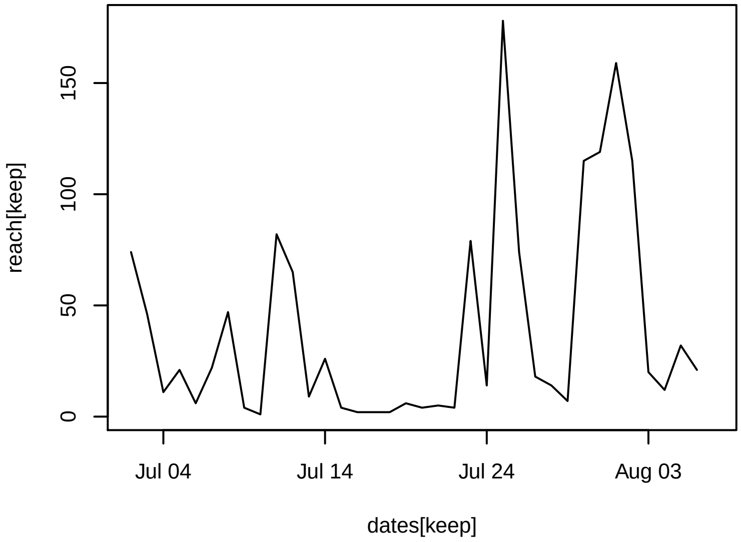


FIGURE 11. Daily Total Reach

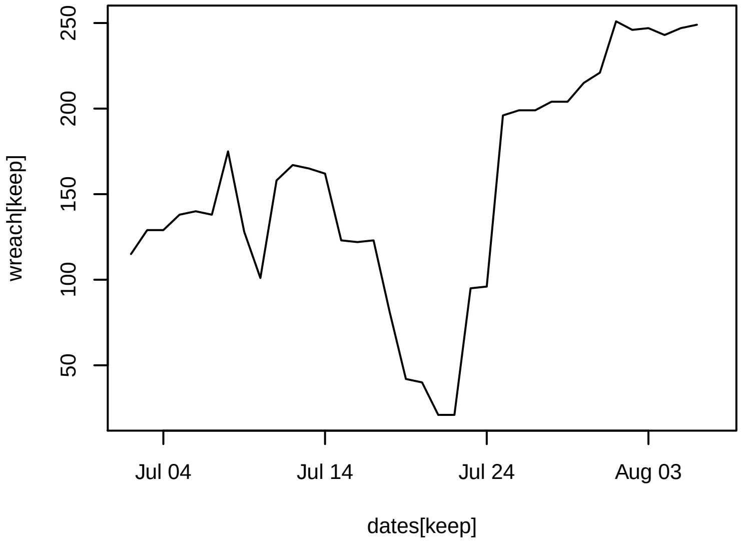


FIGURE 12. Weekly Total Reach