

001 - Introduction to Inferential Statistics

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar

Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

`sahir.bhatnagar@mcgill.ca`

slides compiled on September 1, 2020



Objectives for this course

1. Visualize/Analyze/Interpret data using statistical methods with a computer

Objectives for this course

1. Visualize/Analyze/Interpret data using statistical methods with a computer
2. Gather data into analysis ready format

Objectives for this course

1. Visualize/Analyze/Interpret data using statistical methods with a computer
2. Gather data into analysis ready format
3. Learn regression

Objectives for this course

1. Visualize/Analyze/Interpret data using statistical methods with a computer
2. Gather data into analysis ready format
3. Learn regression
4. Understand the statistical results in a scientific paper

Objectives for this course

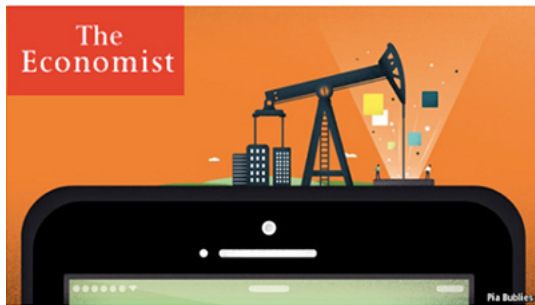
1. Visualize/Analyze/Interpret data using statistical methods with a computer
2. Gather data into analysis ready format
3. Learn regression
4. Understand the statistical results in a scientific paper
5. Learn the tools for creating reproducible analyses

Data is the new oil¹

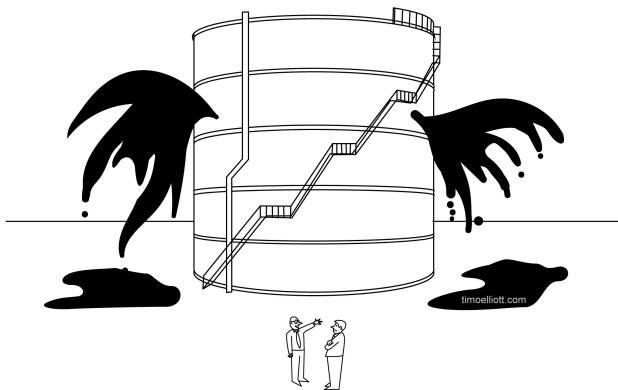
Fuel of the future

Data is giving rise to a new economy

How is it shaping up?



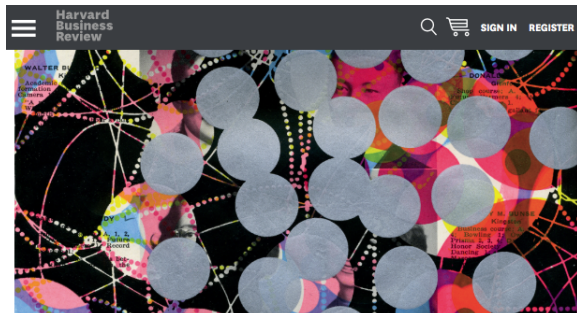
¹<https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy>



“Data is the new oil? Absolutely—toxic if mishandled!...”

²

<https://timoelliott.com/blog/2018/03/data-is-the-new-oil-yes-toxic-if-mishandled.html>



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

³

<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

Why R ?

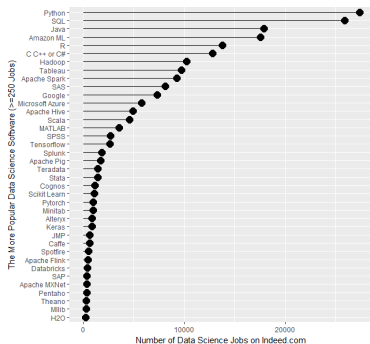


Figure: Data as of May 2019

<http://r4stats.com/articles/popularity/>

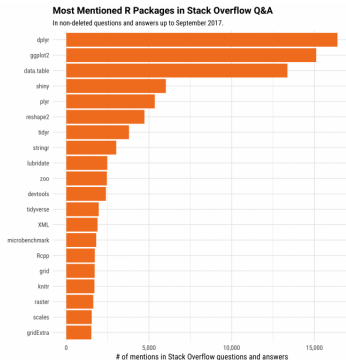


Figure: Popular R packages [https://](https://stackoverflow.blog/2017/10/10/impressive-growth-r/)

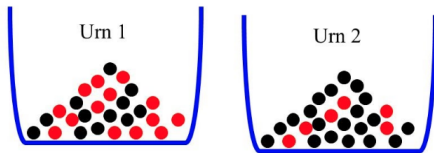
stackoverflow.blog/2017/10/10/impressive-growth-r/

First day in a statistics course

Example:

We have two urns. Urn 1 contains 14 red balls and 12 black balls. Urn 2 contains 6 red balls and 20 black balls.

An Urn is selected at random and a ball is selected from that urn.



If the ball turns out to be red what is the probability that it came from the first urn?

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa

Second day in a statistics course



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Tidy data

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational units forms a table
- Tidy data is ready for regression routines and plotting

country	year	cases	population
Afghanistan	1999	181	19987071
Afghanistan	2000	2666	200095360
Brazil	1999	30737	172006362
Brazil	2000	80488	174004898
China	1999	210258	1272015272
China	2000	210766	1280008583

variables

country	year	cases	population
Afghanistan	1999	181	19987071
Afghanistan	2000	2666	200095360
Brazil	1999	30737	172006362
Brazil	2000	80488	174004898
China	1999	210258	1272015272
China	2000	210766	1280008583

observations

country	year	cases	population
Afghanistan	1999	181	19987071
Afghanistan	2000	2666	200095360
Brazil	1999	30737	172006362
Brazil	2000	80488	174004898
China	1999	210258	1272015272
China	2000	210766	1280008583

values

Example: Does a full moon affect behaviour?

- Many people believe that the moon influences the actions of some individuals.
- A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks.
- Days were classified as moon days if they were in a 3-day period centered at the day of the full moon.
- For each patient, the average number of disruptive behaviors was computed for moon days and for all otherdays.

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26
6	3.67	0.11
7	4.67	0.30

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32

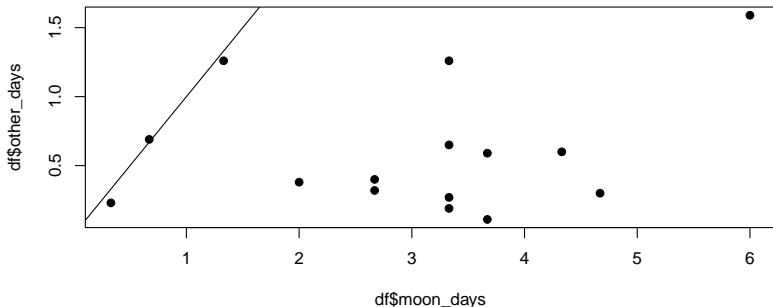
Question: Can I plot the data?

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32

Question: Can I plot the data?

```
plot(df$moon_days, df$other_days, pch = 19)  
abline(a=0,b=1)
```



Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

Question: Can I fit a meaningful regression model directly to the variables in the data?

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

Question: Can I fit a meaningful regression model directly to the variables in the data?

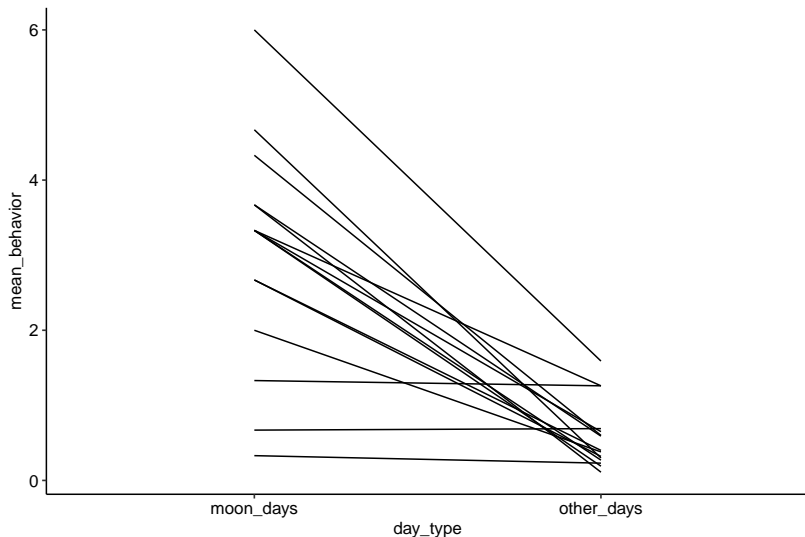
```
## Call: lm(formula = moon_days ~ other_days, data = df)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.56         0.66    3.9   0.002
## other_days       0.79         0.91    0.9   0.402
##
## Residual standard error: 1.5 on 13 degrees of freedom
## Multiple R-squared:  0.055, Adjusted R-squared: -0.018
## F-statistic: 0.75 on 1 and 13 DF,  p-value: 0.4
```

Is it tidy?

patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

Plotting with tidy data

```
ggplot(data = df_tidy, mapping = aes(x = day_type, y = mean_behavior, group = patient)) + geom_line()
```



Regression with tidy data

```
fit <- lme4::lmer(mean_behavior ~ day_type + (1|patient), data = df_tidy)
summary(fit)

## Linear mixed model fit by REML ['lmerMod']
## Formula: mean_behavior ~ day_type + (1 | patient)
## Data: df_tidy
##
## REML criterion at convergence: 90.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.27236 -0.30142 -0.04023  0.48540  2.44753
##
## Random effects:
## Groups Name Variance Std.Dev.
## patient (Intercept) 0.1563  0.3954
## Residual 1.0659  1.0324
## Number of obs: 30, groups: patient, 15
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)      3.0220     0.2854  10.587
## day_typeother_days -2.4327     0.3770  -6.453
##
## Correlation of Fixed Effects:
##              (Intr)
## dy_typthr_d -0.660
```

Not tidy vs. tidy data

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

Not tidy vs. tidy data

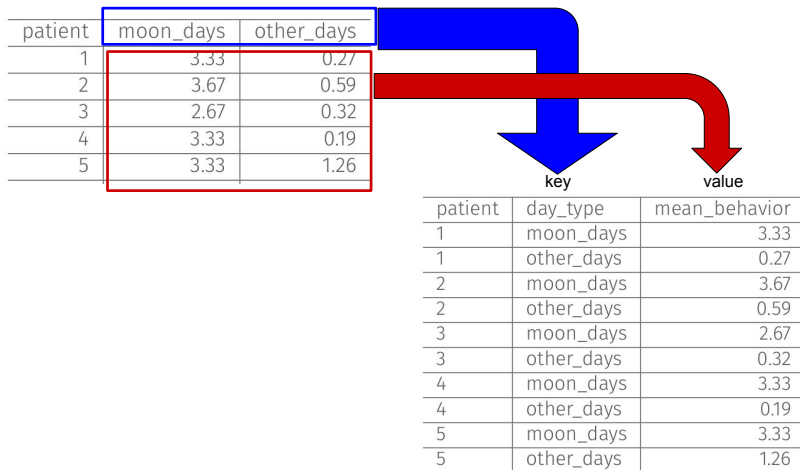
patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

Not tidy

tidy

patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

tidyr::pivot_longer()



```
tidyr::pivot_longer(data = df, cols = -patient, names_to = "day_type", values_to = "mean_behavior")
```


Traditional stats textbook

CHAPTER 7

Hypothesis Testing: One-Sample Inference / 211

- 7.1 Introduction / 211
- 7.2 General Concepts / 211
- 7.3 One-Sample Test for the Mean of a Normal Distribution: One-Sided Alternatives / 214
- 7.4 One-Sample Test for the Mean of a Normal Distribution: Two-Sided Alternatives / 222
- 7.5 The Relationship Between Hypothesis Testing and Confidence Intervals / 229
- 7.6 The Power of a Test / 232
- 7.7 Sample-Size Determination / 239
- 7.8 One-Sample χ^2 Test for the Variance of a Normal Distribution / 245
- 7.9 One-Sample Inference for the Binomial Distribution / 249
- 7.10 One-Sample Inference for the Poisson Distribution / 259
- 7.11 Case Study: Effects of Tobacco Use on Bone-Mineral Density in Middle-Aged Women / 265
- 7.12 Derivation of Selected Formulas / 265
- 7.13 Summary / 267

Problems / 269

CHAPTER 8

Hypothesis Testing: Two-Sample Inference / 279

- 8.1 Introduction / 279
- 8.2 The Paired t Test / 281
- 8.3 Interval Estimation for the Comparison of Means from Two Paired Samples / 285
- 8.4 Two-Sample t Test for Independent Samples with Equal Variances / 286
- 8.5 Interval Estimation for the Comparison of Means from Two Independent Samples (Equal Variance Case) / 290
- 8.6 Testing for the Equality of Two Variances / 292
- 8.7 Two-Sample t Test for Independent Samples with Unequal Variances / 298

CHAPTER 11

Regression and Correlation Methods / 457

- 11.1 Introduction / 457
- 11.2 General Concepts / 458
- 11.3 Fitting Regression Lines—The Method of Least Squares / 461
- 11.4 Inferences About Parameters from Regression Lines / 465
- 11.5 Interval Estimation for Linear Regression / 475
- 11.6 Assessing the Goodness of Fit of Regression Lines / 481
- 11.7 The Correlation Coefficient / 485
- 11.8 Statistical Inference for Correlation Coefficients / 490
- 11.9 Multiple Regression / 502
- 11.10 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 519
- 11.11 Partial and Multiple Correlation / 526
- 11.12 Rank Correlation / 529

Copyright 2011 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has determined that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

x Contents

- 8.8 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 305
- 8.9 Estimation of Sample Size and Power for Comparing Two Means / 307
- 8.10 The Treatment of Outliers / 312
- 8.11 Derivation of Equation 8.13 / 319
- 8.12 Summary / 320

Problems / 320

This course

CHAPTER 7

Hypothesis Testing: One-Sample Inference / 211

- 7.1 Introduction / 211
- 7.2 General Concepts / 212
 - 7.2.1 One-Sample z Test for the Variance
 - 7.2.2 General Concepts / 212
- 7.3 One-Sample t Test for the Mean of a Normal Distribution / 214
 - 7.3.1 One-Sample t Test for the Mean of a Normal Distribution / 214
 - 7.3.2 One-Sample t Test for the Mean of a Normal Distribution / 214
- 7.4 One-Sample F Test for the Variance of a Normal Distribution / 222
 - 7.4.1 One-Sample F Test for the Variance of a Normal Distribution / 222
 - 7.4.2 One-Sample F Test for the Variance of a Normal Distribution / 222
- 7.5 The Power of a Test / 229
- 7.6 Sample Size Determination / 229
- 7.7 Problems / 269

CHAPTER 8

Hypothesis Testing: Two-Sample Inference / 279

- 8.1 Introduction / 279
- 8.2 General Concepts / 281
 - 8.2.1 General Concepts / 281
 - 8.2.2 General Concepts / 281
- 8.3 Interval Estimation for the Comparison of Means from Two Paired Samples / 285
 - 8.3.1 Interval Estimation for the Comparison of Means from Two Paired Samples / 285
 - 8.3.2 Interval Estimation for the Comparison of Means from Two Paired Samples / 285
- 8.4 Two-Sample t Test for Independent Samples with Equal Variances / 286
 - 8.4.1 Two-Sample t Test for Independent Samples with Equal Variances / 286
 - 8.4.2 Two-Sample t Test for Independent Samples with Equal Variances / 286
- 8.5 Two-Sample t Test for Independent Samples with Unequal Variances / 292
 - 8.5.1 Two-Sample t Test for Independent Samples with Unequal Variances / 292
 - 8.5.2 Two-Sample t Test for Independent Samples with Unequal Variances / 292
- 8.6 Two-Sample F Test for Independent Samples with Unequal Variances / 298
 - 8.6.1 Two-Sample F Test for Independent Samples with Unequal Variances / 298
 - 8.6.2 Two-Sample F Test for Independent Samples with Unequal Variances / 298
- 8.7 Problems / 312

CHAPTER 11

Regression and Correlation Methods / 457

- 11.1 Introduction / 457
- 11.2 General Concepts / 458
 - 11.2.1 General Concepts / 458
 - 11.2.2 General Concepts / 458
- 11.3 Fitting Regression Lines—The Method of Least Squares / 461
 - 11.3.1 Fitting Regression Lines—The Method of Least Squares / 461
 - 11.3.2 Fitting Regression Lines—The Method of Least Squares / 461
- 11.4 Inferences About Parameters from Regression Lines / 465
 - 11.4.1 Inferences About Parameters from Regression Lines / 465
 - 11.4.2 Inferences About Parameters from Regression Lines / 465
- 11.5 Interval Estimation for Linear Regression / 475
 - 11.5.1 Interval Estimation for Linear Regression / 475
 - 11.5.2 Interval Estimation for Linear Regression / 475
- 11.6 Assessing the Goodness of Fit of Regression Lines / 481
 - 11.6.1 Assessing the Goodness of Fit of Regression Lines / 481
 - 11.6.2 Assessing the Goodness of Fit of Regression Lines / 481
- 11.7 The Correlation Coefficient / 485
 - 11.7.1 The Correlation Coefficient / 485
 - 11.7.2 The Correlation Coefficient / 485
- 11.8 Statistical Inference for Correlation Coefficients / 490
 - 11.8.1 Statistical Inference for Correlation Coefficients / 490
 - 11.8.2 Statistical Inference for Correlation Coefficients / 490
- 11.9 Multiple Regression / 502
 - 11.9.1 Multiple Regression / 502
 - 11.9.2 Multiple Regression / 502
- 11.10 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 519
 - 11.10.1 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 519
 - 11.10.2 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 519
- 11.11 Partial and Multiple Correlation / 526
 - 11.11.1 Partial and Multiple Correlation / 526
 - 11.11.2 Partial and Multiple Correlation / 526
- 11.12 Rank Correlation / 529
 - 11.12.1 Rank Correlation / 529
 - 11.12.2 Rank Correlation / 529

x Contents

- 8.8 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 305
- 8.9 Estimation of Sample Size and Power for Comparing Two Means / 307
- 8.10 The Treatment of Outliers / 312
- 8.11 Derivation of Equation 8.13 / 319
- 8.12 Summary / 320
- Problems / 320

Statistical concepts

RESULTS The total populations were 462 445 in the Iowa border counties and 272 385 in the Illinois border counties. Population density was higher in the Iowa counties (114.2 people per square mile) than in the Illinois counties (78.2 people per square mile). Trends of cumulative COVID-19 cases per 10 000 residents for the Iowa and Illinois border counties were comparable before the Illinois stay-at-home order, which went into effect at 5:00 PM on March 21 (March 15 to March 21: 0.024 per 10 000 residents vs 0.026 per 10 000 residents). After that, cases increased more quickly in Iowa and more slowly in Illinois. Within 10, 20, and 30 days after the enactment of the stay-at-home order in Illinois, the difference in cases was -0.51 per 10 000 residents (SE, 0.09; 95% CI, -0.69 to -0.32; $P < .001$), -1.15 per 10 000 residents (SE, 0.49; 95% CI, -2.12 to -0.18; $P = .02$), and -4.71 per 10 000 residents (SE, 1.99; 95% CI, -8.64 to -0.78; $P = .02$), respectively. The estimates indicate excess cases in the border Iowa counties by as many as 217 cases after 1 month without a stay-at-home order. This estimate of excess cases represents 30.4% of the 716 total cases in those Iowa counties by that date. Sensitivity analyses addressing differences in timing of closing schools and nonessential businesses and differences in county population density and poverty rates between the 2 states supported these findings.

4

⁴ <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2766229>

Statistical concepts

Table 1. Difference-in-Differences Estimates of COVID-19 Cases Comparing Border Counties in Iowa With Those in Illinois Before and After the Stay-at-Home Order Was Issued in Illinois^a

Period	Difference in COVID-19 cases per 10 000 residents ^b	Heteroskedasticity robust SE (95% CI) ^c	P value	Excess cases in Iowa border counties	Excess cases as proportion of total cases, %
3/22-3/26	-0.14	0.04 (-0.23 to -0.06)	.001	6	32.4
3/27-3/31	-0.51	0.09 (-0.69 to -0.32)	<.001	24	38.0
4/01-4/05	-0.41	0.17 (-0.74 to -0.07)	.02	19	15.2
4/06-4/10	-1.15	0.49 (-2.12 to -0.18)	.02	53	17.8
4/11-4/15	-3.35	1.19 (-5.70 to -0.99)	.006	154	30.0
4/16-4/20	-4.71	1.99 (-8.64 to -0.78)	.02	217	30.4

Abbreviation: COVID-19, coronavirus disease 2019.

- ^a The regression model was estimated separately for each of 5-day period. The regression was estimated using least squares weighted by the 2019 county population. The regression adjusted for county and day fixed effects. The number of county-day observations was 180 for each regression.
- ^b This indicates the estimated difference-in-differences association of a stay-at-home order with COVID-19 cases in a given period relative to March 15 to March 21 (ie, the period before the stay-at-home order in Illinois was enacted).
- ^c Heteroskedasticity robust SEs were estimated because homoscedasticity is rejected for all post-period regressions.

5

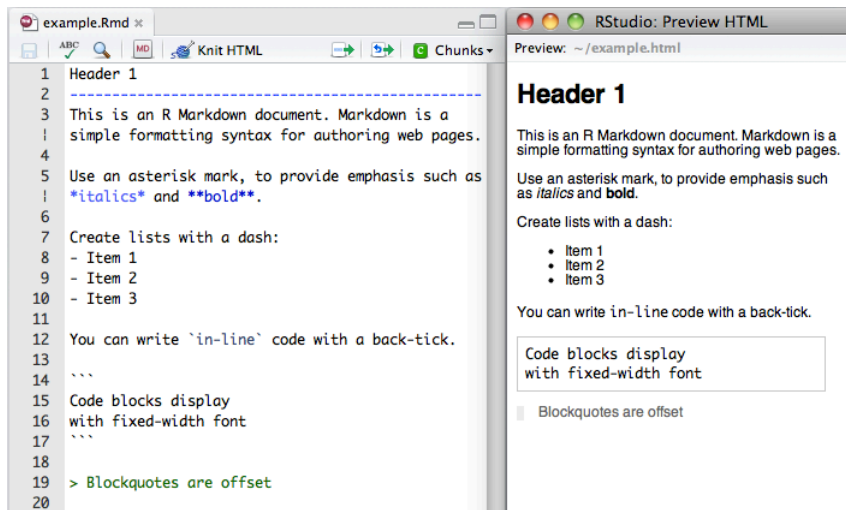
⁵ <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2766229>

Copy paste ad nauseam

The screenshot illustrates a workflow involving two Microsoft Office applications. In the background, a Microsoft Excel window titled 'Import Sample.xlsx' is open, displaying a table with project data. A red arrow points from the 'Paste' button in the Excel ribbon to the 'Paste' button in the foreground Word window. The foreground Word window, titled 'Select Table from Graph.doc [Compatibility Mode]', has its 'Paste' button highlighted, and a 'Paste Options' menu is open, showing icons for different paste formats. The Word window also displays a table with the same data as the Excel window.

Designer	Project	Due Date	Budget
Anderson	Christmas Village	Nov 28	\$ 26,000.00
Cochran	Arts Festival	Sep 15	\$ 31,000.00
Fotou	Botanical Gardens	Aug 10	\$ 18,000.00
Heath	LeMieux Galleries	Oct 15	\$ 22,000.00
Hill	Beau Rivage	Nov 01	\$ 24,500.00

Markdown: HTML without knowing HTML



The image shows a screenshot of the RStudio interface. On the left, the 'example.Rmd' file is open in the editor. The code is as follows:

```
1 Header 1
2 -----
3 This is an R Markdown document. Markdown is a
4 | simple formatting syntax for authoring web pages.
5 Use an asterisk mark, to provide emphasis such as
6 | italics and bold.
7 Create lists with a dash:
8 - Item 1
9 - Item 2
10 - Item 3
11
12 You can write in-line code with a back-tick.
13
14 ```
15 Code blocks display
16 with fixed-width font
17 ```
18
19 > Blockquotes are offset
20
```

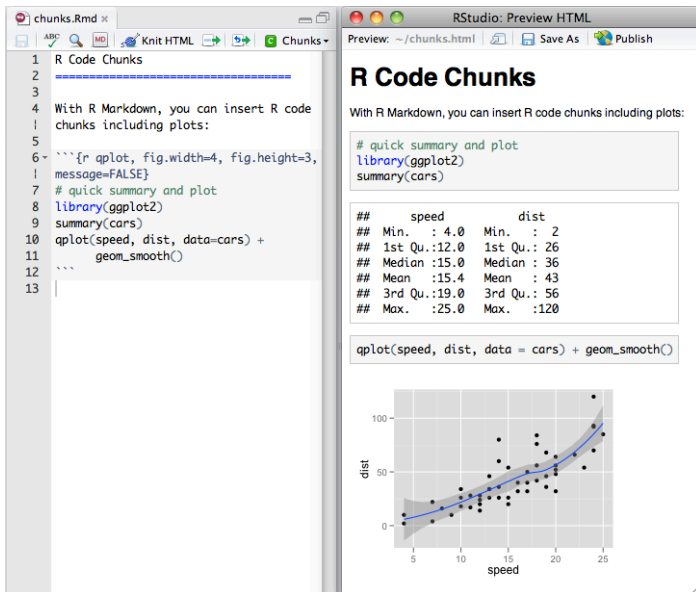
On the right, the 'RStudio: Preview HTML' window shows the rendered output of the document. The preview is located at `~/example.html`. The rendered HTML includes:

- A header

Header 1

.
- A paragraph: 'This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages.'
- Another paragraph: 'Use an asterisk mark, to provide emphasis such as *italics* and **bold**.'
- A section titled 'Create lists with a dash:' followed by a bulleted list:
 - Item 1
 - Item 2
 - Item 3
- A paragraph: 'You can write in-line code with a back-tick.'
- A code block containing the text: 'Code blocks display with fixed-width font'.
- A blockquote: 'Blockquotes are offset'.

R + Markdown = RMarkdown



The screenshot displays the RStudio interface with two main panes. The left pane, titled 'chunks.Rmd', shows the source R Markdown code. The right pane, titled 'RStudio: Preview HTML', shows the rendered HTML output of the code chunks.

Left Pane (Source Code):

```
1 R Code Chunks
2 =====
3
4 With R Markdown, you can insert R code
5 chunks including plots:
6
7 ```{r qplot, fig.width=4, fig.height=3,
8    message=FALSE}
9 # quick summary and plot
10 library(ggplot2)
11 summary(cars)
12 qplot(speed, dist, data=cars) +
13   geom_smooth()
```

Right Pane (Rendered HTML):

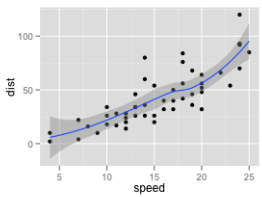
R Code Chunks

With R Markdown, you can insert R code chunks including plots:

```
# quick summary and plot
library(ggplot2)
summary(cars)
```

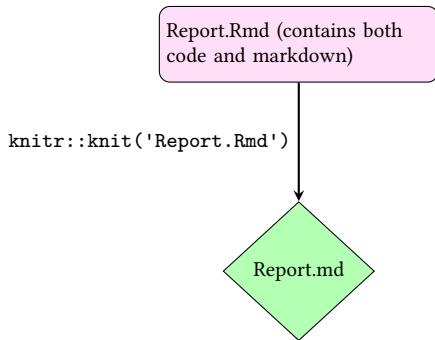
##	speed	dist
## Min.	: 4.0	Min. : 2
## 1st Qu.:	12.0	1st Qu.: 26
## Median	:15.0	Median : 36
## Mean	:15.4	Mean : 43
## 3rd Qu.:	19.0	3rd Qu.: 56
## Max.	:25.0	Max. :120

```
qplot(speed, dist, data = cars) + geom_smooth()
```



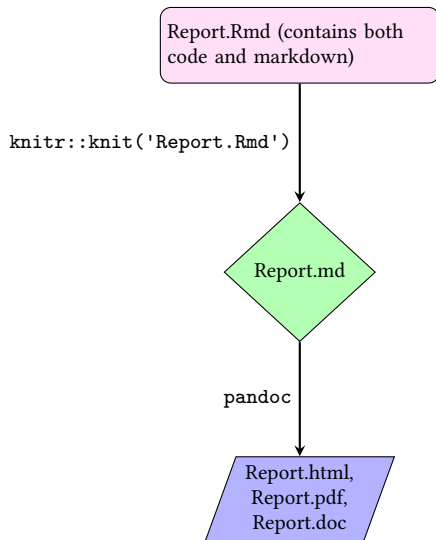
What rmarkdown does

RMarkdown example:



What rmarkdown does

RMarkdown example:

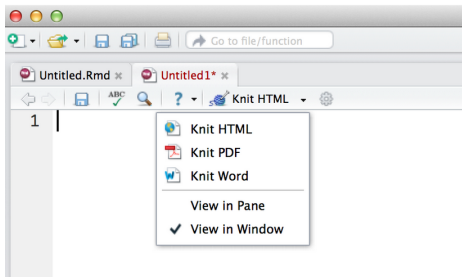


Compiling a .Rmd document

The two steps on previous slide can be executed in one command:

```
rmarkdown::render()
```

or in RStudio:



Topics by level of exposure

Level of Exposure



- Mainstream media
- Scientific Article
- Statistical Analysis
- Cleaning Data
- Collecting Data
- Research Ethics Board (REB) Approval
- Study Design
- Research Question
- Statistical Methods Development

First year courses

Level of Exposure



- Mainstream media
- Scientific Article
- Statistical Analysis
- Cleaning Data
- Collecting Data
- Research Ethics Board (REB) Approval
- Study Design
- Research Question
- Statistical Methods Development

EPIB 607/621
EPIB 613

EPIB 601/602

My area of research

Level of Exposure



- Mainstream media
- Scientific Article
- Statistical Analysis
- Cleaning Data
- Collecting Data
- Research Ethics Board (REB) Approval
- Study Design
- Research Question
- Statistical Methods Development

EPIB 607/621
EPIB 613

EPIB 601/602

What I do

Session Info

```
R version 3.6.2 (2019-12-12)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 19.10

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnu/openblas/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/libopenblas-p-r0.3.7.so

attached base packages:
[1] tools      stats      graphics  grDevices  utils      datasets  methods
[8] base

other attached packages:
[1] NCStats_0.4.7      FSA_0.8.30          forcats_0.5.0       stringr_1.4.0
[5] dplyr_1.0.2         purrr_0.3.4         readr_1.3.1         tidyr_1.1.2
[9] tibble_3.0.3        ggplot2_3.3.2.9000  tidyverse_1.3.0     knitr_1.29

loaded via a namespace (and not attached):
[1] Rcpp_1.0.4.6        lubridate_1.7.4      lattice_0.20-38      assertthat_0.2.1
[5] digest_0.6.25       R6_2.4.1             cellranger_1.1.0     plyr_1.8.6
[9] backports_1.1.9     reprex_0.3.0         evaluate_0.14        httr_1.4.1
[13] highr_0.8           pillar_1.4.6         TeachingDemos_2.12   rlang_0.4.7
[17] readxl_1.3.1        rstudioapi_0.11      minqa_1.2.4          nloptr_1.2.2.1
[21] Matrix_1.2-18       labeling_0.3          splines_3.6.2        lme4_1.1-21
[25] munsell_0.5.0       broom_0.7.0          compiler_3.6.2       modelr_0.1.5
[29] xfun_0.16           pkgconfig_2.0.3      tidyselect_1.1.0     fansi_0.4.1
[33] crayon_1.3.4        dbplyr_1.4.2         withr_2.2.0          ggpubr_0.2.5
[37] MASS_7.3-51.5       grid_3.6.2           nlme_3.1-143         jsonlite_1.7.0
[41] gtable_0.3.0        lifecycle_0.2.0      DBI_1.1.0            magrittr_1.5
[45] scales_1.1.1        cli_2.0.2            stringi_1.4.6        farver_2.0.3
[49] ggsignif_0.6.0      fs_1.3.2             xml2_1.3.0           ellipsis_0.3.1
[53] generics_0.0.2      vctrs_0.3.4          boot_1.3-24          glue_1.4.2
[57] hms_0.5.3           colorspace_1.4-1     rvest_0.3.5          haven_2.3.1
```