A Quick Guide to Data Analysis Using R

Paul Testa

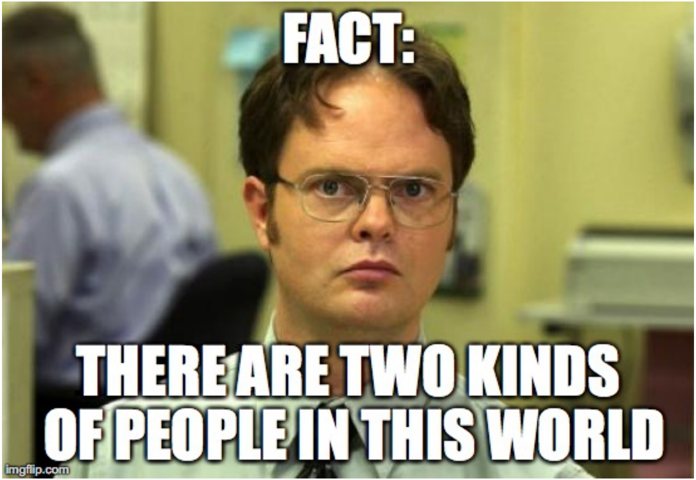
December 1, 2016

Goals

- Provide an overview of the steps to producing quantitative research
- Offer a quick tutorial on how to do such research using R

But first. . .

One piece of fundamental knowledge



Two kinds of people in this world

Two kinds of people in this world

What is it that we say we do here

What does quantitative research do?

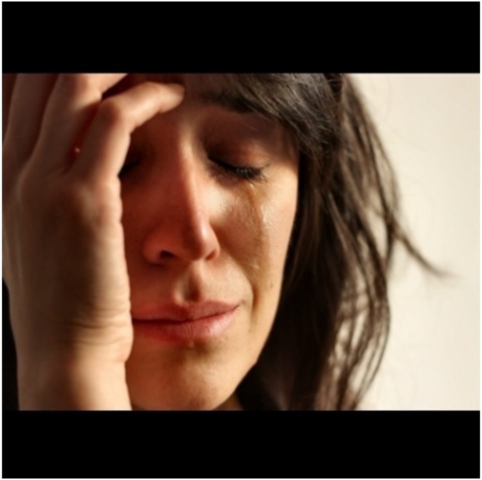
- Description
- Explanations
- Prediction

What does quantitative research do?

- Description
- Explanations
- Prediction

"All models are wrong, some models are useful" - George Box

What does quantitative research look like?



What do my papers look like?

- Introduction & Research
- Theory & Expectations
- Data & Methods
- Results & Discussion
- Conclusion

What you've done

- **Introduction & Research**
- **Theory & Expectations**
- *Data & Methods*
- Results & Discussion
- Conclusion

What we'll focus on today

- Introduction
- Theory
- **Data & Methods**
- **Results & Discussion**
- Conclusion

Introduction

Your chance to hook the reader dispel any “so what or why should we care questions” by clearly stating research question and plan for addressing it.

Data and Methods

What is Data?

What is Data?

- Information about the world

Where Does Data Come From?

Where Does Data Come From?

- Your hard work
- Other people's hard work
- A combination of both

Types of Data

- Qualitative vs Quantitave
- Levels of Measurement (NOIR)
  - Nominal: Gender, Party ID
  - Ordinal: Strongly Agree, Agree ... Disagree, Strongly Disagree
  - Interval: Temperature (Celsuis or Farenheit)
  - Ratio: Income, Height, Temperature (Kelvin)

The match between concepts and measurement is rarely perfect

Methods: How can we use data to answer questions that interest us?

- Measures of central tendency (what's typical)
  - mean, median, mode, percentiles
- Measures of dispersion (how much variation is there)
  - Variance, standard deviations
- Measures of association (how do things relate)
  - Covariance, correlation, linear regression, ...
- Methods for statistical inference
  - Confidence intervals and hypothesis tests
  - Tools for quantifying our confidence or certainty in our results

Methods: How do we know what method to use?

- Driven by the question you want to ask.
- Helpful to think in terms of models (simplications of the world)
- Later, we'll look at the relationship between support for Trump and education In general we might a negative relationship

$$\text{Support Trump} \sim \underbrace{\text{Education}}_{(-)}$$

Methods: Regression

- We can estimate this model:

$$\text{Support Trump} \sim \underbrace{\text{Education}}_{(-)}$$

- Using Ordinary Least Squares Regression, tool for describing how the mean of one variable changes (linearly) with changes in other variable(s).
- We can also estimate models that ask how support for trump changes conditional on both education and gender and controlling for the relationships between education and gender:

$$\text{Support Trump} \sim \underbrace{\text{Education}}_{(-)} + \underbrace{\text{Male}}_{(+)}$$

An aside on “co ntrolling for X”

Statistical Inference: Some Definitions

- **Population:** all the elements of a set of data. The thing we're interested in. Typically unobserved or unknown.
  - **Parameters:** measurable characteristics of the population (e.g. expected value μ, variance, σ^2) that are typically unknown or unknowable.
- **Samples:** a subset of the population of size N from which we try to learn things about our population of interest
- **Statistical Inference:** The process of learning characteristics of the population from a sample
  - **Estimand:** The thing we're trying to estimate.
  - **Estimators:** a rule for calculating an estimate of a given quantity based on observed data.
  - **Statistic:** a measured characteristics of the sample (an estimate)

Sampling distributions

Since we sampled from a random variable (the population distribution of adult male heights), our sample is a random variable, and our estimate (a function of a random variable) is itself a random variable with it's own **sampling distribution.** That is if we'd taken a different sample, we would have gotten a slightly different statistic.

Statistical Inference from Sampling Distributions

The characteristics of this sampling distribution are governed by the underlying population, and so we can use statistics calculated from our sample to make probabilistic statements (inferences) about characteristics of the underlying/unobserved population of interest. We'll do so using two common tools:

- confidence intervals
- hypothesis tests

Confidence Intervals

Components of a Confidence Interval for a Sample Mean

- μ and σ the population mean and standard deviation (unknown)
- n sample size
- $\bar{x}$ sample mean
- $\hat{\sigma}$ sample variance
- $se = \hat{\sigma}/\sqrt{n}$ the standard error: the standard deviation of the sampling distribution
- α a significance level determining used to determine the width of the confidence interval (e.g. $(1-\alpha) \times 100$ percent c.i.)
- t a critical value determined by n and α, for finite sample using a t distribution with degrees of freedom $n-1$

Constructing a Confidence Interval

A 95% confidence interval for $\bar{x}$

$$\bar{x} \pm t * se$$

Example: 100 draws from N(0,1)

```
set.seed(123)
# 100 draws from normal, calculate means
n<-100; x<-rnorm(n); xbar<-mean(x)
# Cacluate SE and critical value
se<-sd(x)/sqrt(n); t<-qt(.975,99)*se
# Caclculate 95 CI
ci<-xbar+c(-t,t)
ci
```

[1] -0.0907 0.2715

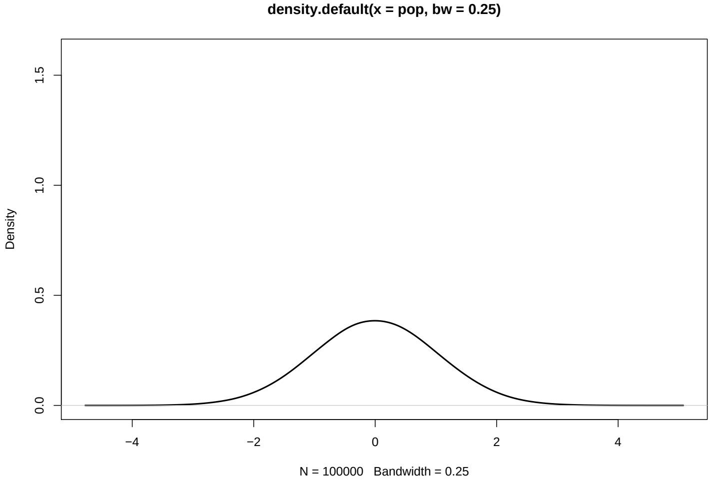
```
# Check
t.test(x)$conf.int
```

[1] -0.0907 0.2715 attr(,"conf.level") [1] 0.95

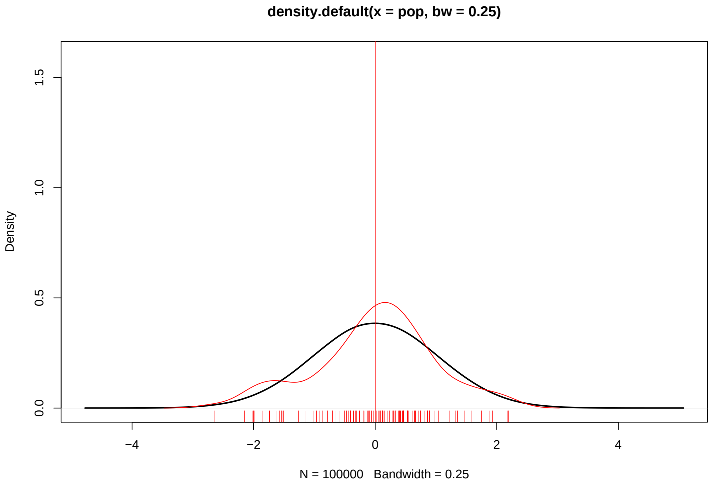
The confidence is about the interval, not the point estimate

95 percent of the intervals constructed in this manner will contain the population value.

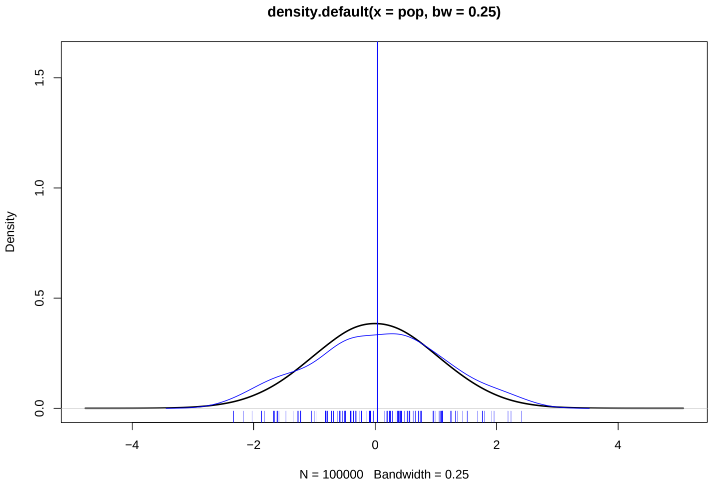
Example: Population 100,000 units N(0,1)



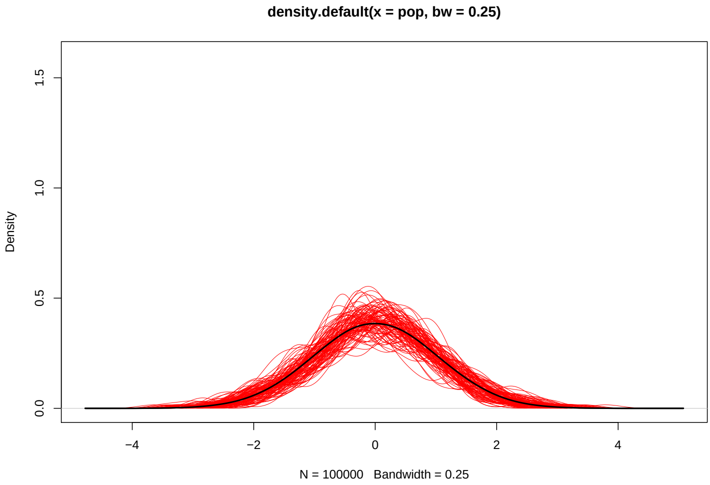
One Sample (n=100) from Population (N=100,000)



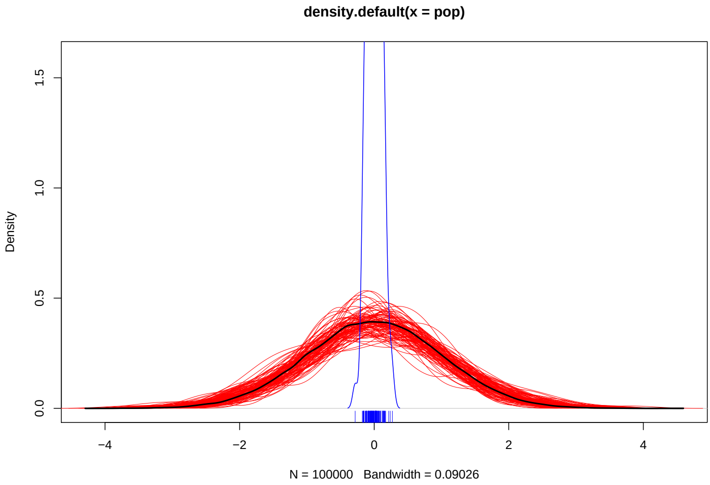
Another Sample (n=100) from Population (N=100,000)



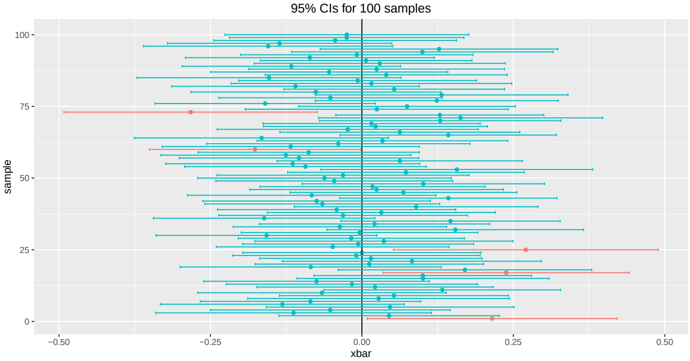
Distribution of 100 Samples, N=100

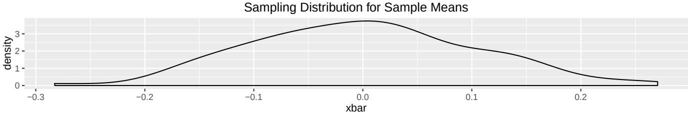


Distribution of Sample Means



Confidence intervals





Hypothesis Testing

Components of a Hypothesis Test

- Construct a hypothesis (H_0) and it's alternative (H_1)
- Choose a test statistic T
- Determine the distribution of T under H_0
- Is the observed value of T likely to occur under H_0?
  - Yes? Fail to reject H_0
  - No? Reject H_0
- p-value: conditional probability of observing a T at least as extreme under H_0

Hypothesis Test for Proportions

- $H_0 : p = 0.5$ and $H_1 : p \neq 0.5$
- Test statistic: $\bar{X}$
- Standardize to compare to reference (normal) distribution by CLT:

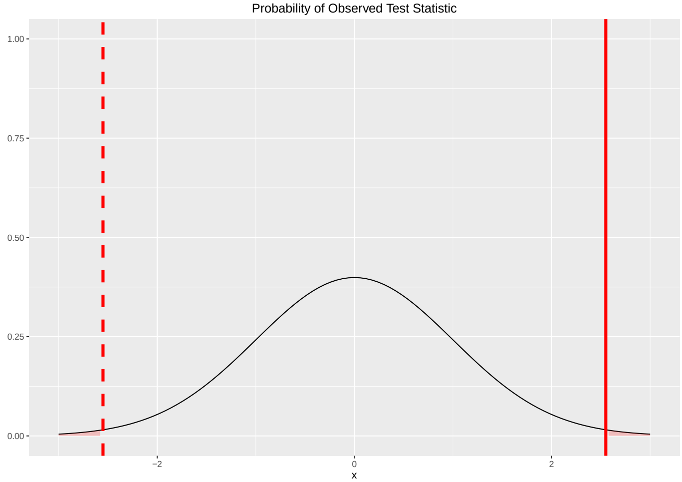
$$Z = \frac{\bar{X} - p_0}{s.e.} = \frac{\bar{X} - p_0}{\sqrt{p_0(1 - p0)/n}} \approx N(0, 1)$$

- Is Z unusual?
  - Reject H_0 if $|Z_obs| > Z_{\alpha/2}$
  - Where $p(reject|H_0) = \alpha$

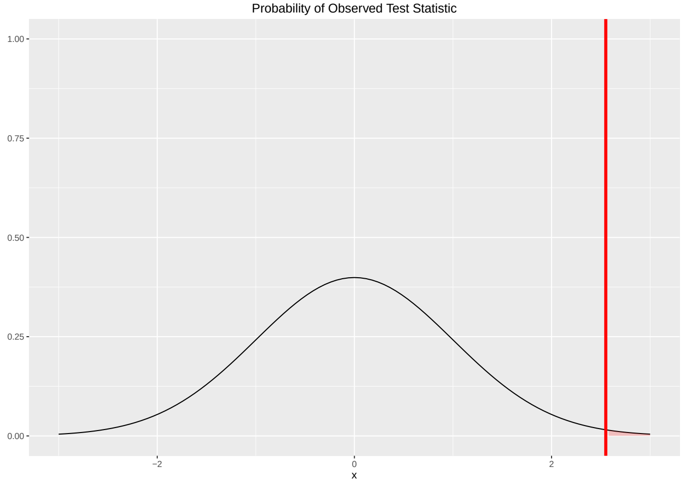
Example: Obama's Approval Rating

- $H_0 : p = 0.5$ $H_1 : p \neq 0.5$
- $\bar{X} = 0.54$ $n = 1018$
- $Z_obs = (0.54 - 0.5)/\sqrt{.5 \times .5/1018} = 2.55 > Z_{0.025} = 1.96$
- $p - value = 0.005 \times 2 = 0.01$
- Reject the null

Example: Obama's Approval Rating



One sided test P(T>t)



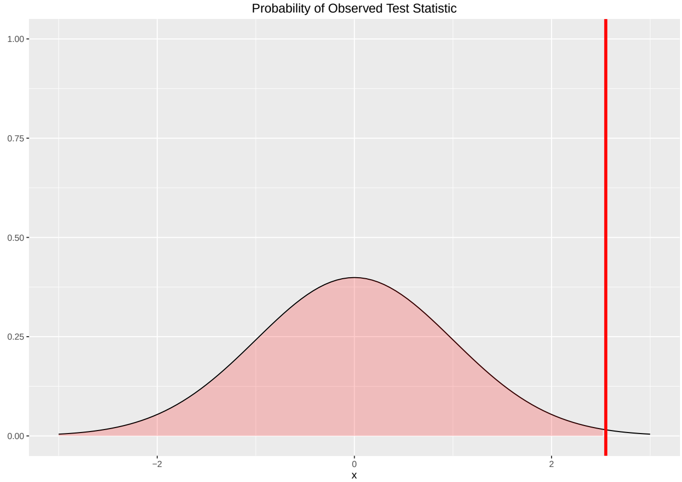
One sided test P(T<t) the other way

A Quick Guide to Data Analysis Using R

Paul Testa

Confidence Intervals

Hypothesis Testing



Type I and Type II Errors

	Reject H_0	Fail to Reject H_0
H_0 is true	Type I Error	Correct!
H_0 is false	Correct!	Type II Error

We generally choose tests to minimize Type I error. Why?

Research Question

- Why is it that most people prefer chocolate-based candy while some degenerates prefer fruit-based candy?

Theory

Two theories of the origins of candy preferences:

- *Candy preferences are innate*
  - Testa et al. (n.d.) argue people's candy preferences are largely determined by our genes, with fruit candy preferences arising from a genetic mutation
- *Candy preferences are socially constructed*
  - Atset et al. (n.d) claim that candy preferences are function of our social environment, and specifically, our desire to appear cool

Expectations

1. **Genes only** If candy perferences are primarily genetic, then once we've controlled for certain mutations, the relationship between liking digusting fruit-based candy and other social factors vanish
2. **Environment only** If candy preferences are socially determined, then genes shouldn't matter.
3. **Genes** $\times$ **Environment** Alterntively, the effects of certain genetic mutations may only be evident in certain enivornments

Some example data

- Wonka Values Survey (WVS)
  - N = 1,000, random sample of U.S. adults
- Outcome: Preference for Fruit Candy (0-1 indicator, μ=0.305)
- Key Predictors:
  - Mutant (0-1 indicator, N_mutant=493)
  - Percent of iTunes that's dubstep (μ=0.455,σ=0.162)

Methods: Linear and Logistic Regressions

$$Y = \beta_0 + \beta_1 Mutant$$

$$Y = \beta_0 + \beta_1 Dubstep$$

$$Y = \beta_0 + \beta_1 Mutant + \beta_2 Dubstep$$

$$Y = \beta_0 + \beta_1 Mutant + \beta_2 Dubstep + \beta_3 Mutant \times Dubstep$$

Results: A Pretty Table

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.18^{***}	-0.03	-0.15^{***}	0.03
	(0.02)	(0.04)	(0.04)	(0.06)
Mutant	0.25^{***}		0.25^{***}	-0.12
	(0.03)		(0.03)	(0.08)
Dubstep_prop		0.73^{***}	0.73^{***}	0.33^{**}
		(0.09)	(0.08)	(0.11)
Mutant:Dubstep_prop				0.81^{***}
				(0.17)
R^2	0.08	0.07	0.14	0.16
Adj. R^2	0.08	0.06	0.14	0.16
Num. obs.	1000	1000	1000	1000
RMSE	0.44	0.45	0.43	0.42

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: OLS Estimates

Results: A Pretty Table

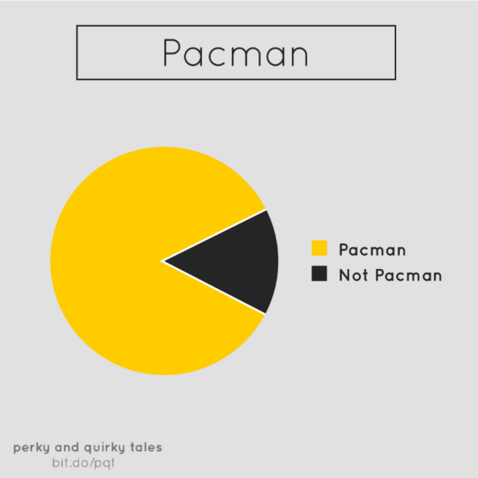
	Model 1	Model 2	Model 3	Model 4
(Intercept)	-1.52^{***}	-2.56^{***}	-3.47^{***}	-2.61^{***}
	(0.12)	(0.24)	(0.28)	(0.38)
Mutant	1.25^{***}		1.36^{***}	-0.14
	(0.15)		(0.15)	(0.51)
Dubstep_prop		3.67^{***}	4.00^{***}	2.30^{**}
		(0.47)	(0.49)	(0.73)
Mutant:Dubstep_prop				3.07^{**}
				(1.01)
AIC	1156.05	1166.93	1085.08	1077.70
BIC	1165.87	1176.75	1099.80	1097.33
Log Likelihood	-576.03	-581.47	-539.54	-534.85
Deviance	1152.05	1162.93	1079.08	1069.70
Num. obs.	1000	1000	1000	1000

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

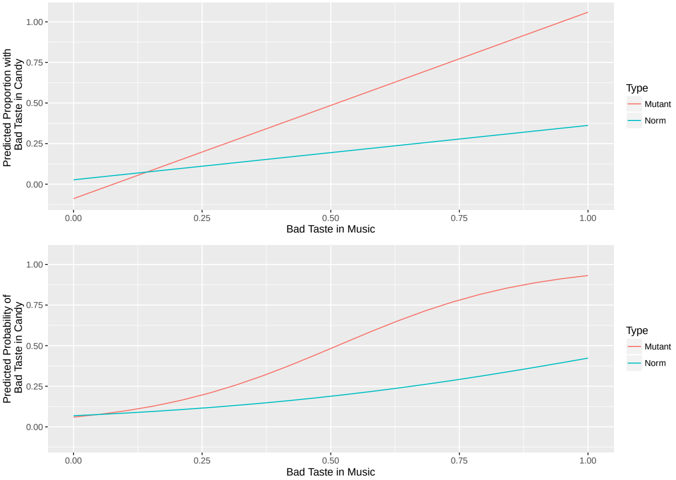
Table 3: Logistic Regression Estimates

Results: A Useless Figure

The only useful pie chart



Results: A Useful Figure



Conclusion

- Dubstep leads to moral and tooth decay

Conclusion