RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Probability and Statistics with Small Data

Ryan Womack
Rutgers University

Executive Education

Probability and Statistics for Small Data

Rutgers Executive MBA

November 1, 2017

Ryan Womack(rwomack@rutgers.edu)
Data Librarian, Rutgers University

October 9, 2017

Case Study - Price Quotes

- Two employees generate different average price quotes.
- How do we know if the difference is *important*?
- *Statistical significance* is one way to evaluate this.
- A *statistically significant* result is unlikely to have happened randomly.
- It does *not* mean that the effect is large or meaningful.

R software

- R is an open-source statistical software environment
- It is heavily used in data science and data analytics
- Favors rapid development of new methods and tools
- Can prioritize cutting-edge research over stability
- We will use R to generate results today, but not try to learn it
- See my R guide if you want to learn more.

Probability

Consider a deck of cards with the usual four suits (clubs, diamonds, hearts, spades) and no jokers.

- The chance of drawing a club from a standard deck is 1/4.
- What is the chance of drawing 2 clubs in two draws?
- If I drew 3 clubs in a row, could I conclude that the deck is stacked?
- Probability is the formal, mathematical study of the chances of events occuring.
- Mathematical statistics makes extensive use of the field of probability. Probability and statistics are intertwined, but separate fields of study.

Samples and Distributions

- One statistical method is to imagine that any random selection of data comes from an underlying larger, *true* population
- Note: If we can effortlessly measure the entire population, we do not need statistical estimates.
- We attempt to understand the true population by sampling and describing our sample
- The mean and standard deviation summarize our sample.
- $\frac{sd}{\sqrt{n}}$ is the formula for *standard error of the mean,* where sd is standard deviation and n is the size of the sample.
- This estimates the accuracy of our sample. As n increases, we get a better estimate. But we must balance with cost of acquiring and analyzing data.

Confidence Intervals

Confidence intervals are the most common way to express the accuracy or uncertainty of an estimate.

- They are directly related to the standard error of the mean.
- We typically see a 95% confidence interval. This is equivalent to the range of from ˜2 (1.96) standard deviations below the mean to ˜2 standard deviations above the mean.
- In this case, the CI formula is then mean plus or minus 2 standard errors, $\bar{x} \pm 2se$
- But we can set the confidence interval to be what we need it to be, depending on the use case.

How to interpret a Confidence Interval

- Exact interpretation is tricky, or hard to keep straight.
- We must remember it is a sampling concept.
- When we construct a 95% confidence interval, it will contain the true mean 95% of the time (in 19 of 20 samples). But we cannot actually say that the CI for the one sample we have just taken contains the true mean.
- That is a bit frustrating and unsatisfying (see Bayesian approach later), but it is accurate.
- Width of the CI indicates the uncertainty of the estimate.
- A value outside of the CI of the mean is unlikely to be the true mean of the data.

Likelihood

Much of probability and statistics is built off of similar concepts.

- Likelihood is another method of understanding how a sample represents a population.
- If I draw 4 cards from a deck, and they are all clubs, it is *more likely* that the deck either contains all clubs or a much higher proportion of clubs than a standard deck.
- We can construct a mathematical expression of likelihood to express this.

Tests of significance

When we want a formal check on whether data is statistically significant, we turn to tests.

- The t-test is a primary example.
- Named after the Student's t-distribution (who was this Student?)
- It is a test based on the sample mean. A one-sample test checks whether the mean is equivalent to some value. A two-sample test compares two population means.
- Note that the Central Limit Theorem implies that under usual conditions the sample mean is normally distributed (if the sample is sufficiently large), even if the underlying variable is not.
- So if the value of the t-statistic is high in absolute value, it is far from normal.
- The p-value expresses this numerically.

Other Tests

Most other common statistical tests operate on the same principle.

- The *chi-square(d)* test is used to check whether patterns in categorical data are random or non-random.
- The *F-test* and *Analysis of Variance (ANOVA)* also fall into this category.
- Each is appropriate in different circumstances.
- There are many specific variants of tests designed to deal with different assumptions about the underlying data.
- Your statistician can advise you on these!

A/B Testing

A/B Testing, or Online Controlled Experiments,is an important variant of significance testing that has emerged in internet applications.

- Also called split tests, randomized experiments, or other terminology.
- Because it is a *randomized, controlled* experiment, *causality* can be established.
- The only expected difference between two groups is whether they receive the *control* or *treatment* (A/B)
- So we can test whether the observed value of some measure of interest shows a *statistically significant difference* between the two groups (for example, with 2-sample t-test).
- If there is a difference, the treatment should be responsible for it.
- Statistics helps us establish appropriate sample sizes with enough power to generate a statistically significant result.

Linear Regression

Regression analysis is the general term for modeling the relationship among variables

- Least squares, or ordinary least squares (OLS), is just the most common method to fit a line to points. Least squares or OLS regression is sometimes used loosely as a substitute for linear regression.
- Linear regression - just fitting a line
- Simple linear regression is the relationship of one *explanatory* (independent) variable to one *response* (dependent) variable.
- Multiple linear regression allows for many explanatory variables to predict a response.
- The result is a linear equation.

Case Study - Cell Phone Service

We explore a dataset describing the percentage of poor quality cell phone calls, wind speed, and barometric pressure.

- Does wind speed and atmospheric pressure affect call quality?
- How can we estimate the effect?
- Applied linear regression modeling

Linear Regression components

Each of these plays a role in understanding the overall meaning of a regression analysis

- Statistical significance (p-value)
- Explanatory power (R and R^2)
- Model fit (regression diagnostics)
- Effect size (variable coefficients)
- Uncertainty (confidence intervals)

Linear Regression - Modeling Process

- Modeling process is iterative.
- Requires consideration of many variable combinations and evaluation of fit (e.g. Stepwise regression, AIC).
- Always some subjectivity in the design, which depends on the goal of the study.
- Tradeoff between simplicity, robustness, and accuracy in modeling.
- Cost and time of collecting input data is also a major consideration.

Regression - other models

Of course a straight line may not always represent the best fit, or be suitable to the data

- Generalized linear models allow any family of shapes to fit the data
- Logistic regression modeling is useful for discrete outcomes (Yes/No, Pass/Fail)
- Many other methods as appropriate

Regression, caveats

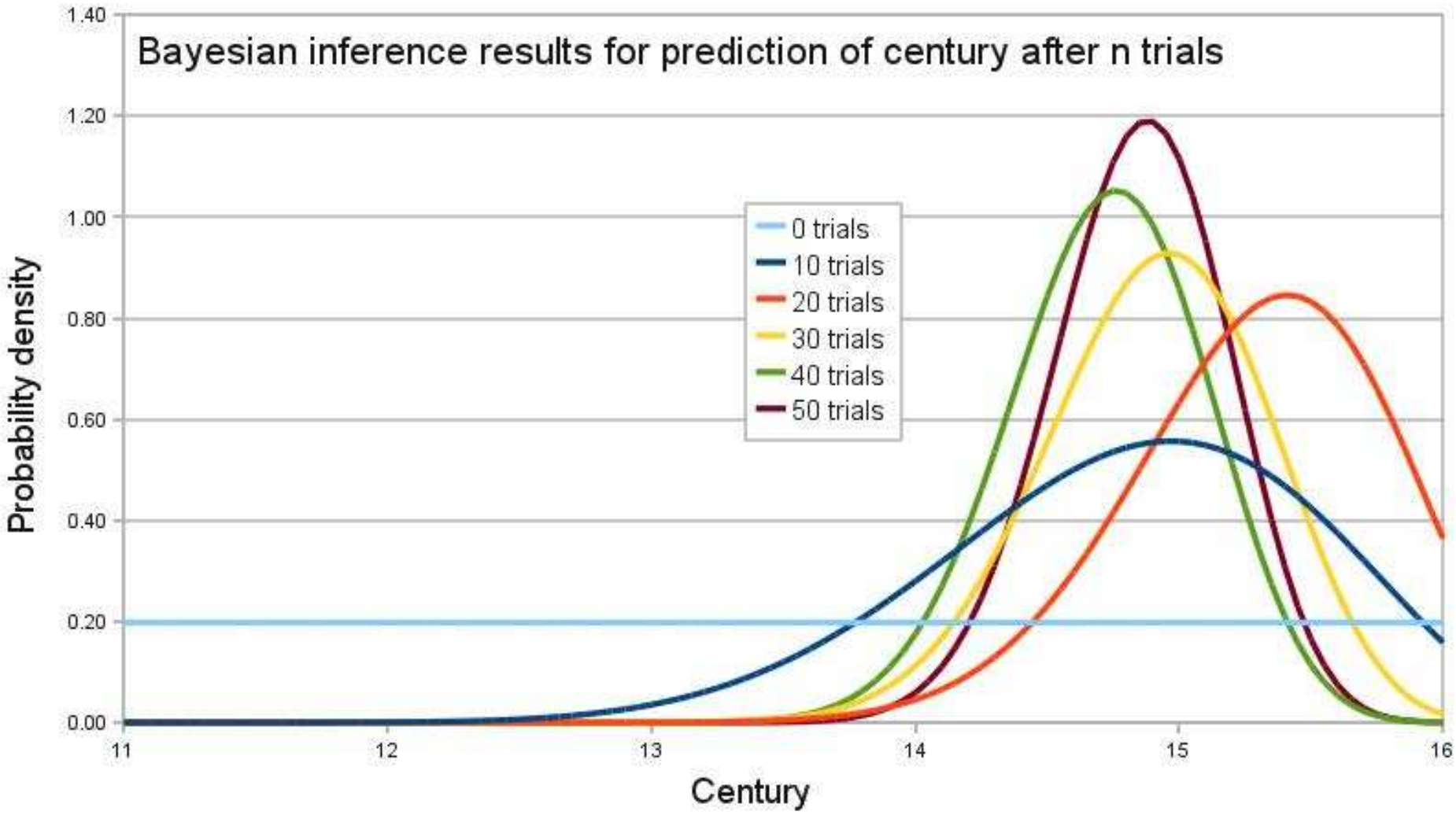
- Correlation is not causation.
- Interpretation of a model depends on some external theory about how the situation works.
- Fit is only for one set of data at one point in time.
- Other kinds of analysis (*time series*) are needed to capture changing relationships.
- Advances in computation have made it much easier to apply Bayesian analysis.

Bayesian Approach

The classical approach to probability that we have discussed up to now is also called the *frequentist* approach. *Bayesian analysis* is a different method.

- In Bayesian analysis we specify a *prior* distribution, which expresses our knowledge before any new data.
- The new data is used to modify the prior, producing a *posterior* distribution that summarizes our knowledge.
- Although this approach offers less theoretical precision than frequentism, it has many real-world benefits. Prior knowledge can be important, and help us to arrive at a faster solution to a problem.

Bayesian, cont.



Bayesian, cont.

- There is some subjectivity in defining the prior and model selection, but this critique of Bayesianism can be exaggerated.
- The posterior distribution is often more effective for modeling outcomes than an estimated mean and confidence interval.
- Advances in computation have made it much easier to apply Bayesian analysis.

Randomized Controlled Trials

How can we firmly establish causation?

- We must isolate bias and spurious associations in the data.
- We must assure that the only difference between two groups is the intervention.
- Full randomization of subject participants is necessary to do this.
- *Clinical trials* are randomized controlled trials in a biomedical setting.

Conclusion

Many tools are available to us from the domain of statistics to understand the variability, reliability, and impact of sampled data. These techniques will also work on big data providing certain conditions are met.

- Descriptive statistics
- Statistical tests of significance, analysis of variance
- Regression and other modeling techniques
- Bayesian methods