POLS/CS&SS 503:
Advanced Quantitative Political Methodology

SUM OF SQUARES, TOTAL SUM OF SQUARES, AND R^2

April 20, 2015

Jeffrey B. Arnold

UNIVERSITY of WASHINGTON
DEPARTMENT OF POLITICAL SCIENCE

CENTER for STATISTICS and the SOCIAL SCIENCES

Overview

Statistical Inference

- Population: Y
- Parameters of interest: β from $Y = \beta X + \epsilon$.
- Sample: y
- Sample statistics (estimates): b
- Since samples are random, different samples produce b?
  - How do we use the samples to the population parameters?
  - How do we quantify our uncertainty about that estimate?

Science is about Uncertainty

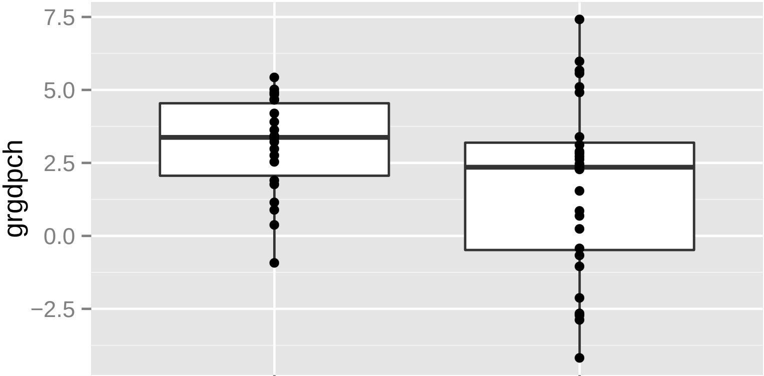
- Knowledge is never certain
- Goal: Estimating unknowns and **quantifying the uncertainty** of those estimates
- Estimates without uncertainty are incomplete at best, useless or biased at worst

The Fundamental Problem of Statistical Inference

- We have methods to calculate the probability of a sample and sample statistics **given** we know the population parameters.
- But we don't know the population parameters, so what do we do?
- Two (three) main methods
    - Frequentist: do not calculate the probability of the parameter
        - Hypothesis testing: Assume a hypothesis and check if data is consistent with it.
        - Confidence intervals: find a plausible range of parameters
    - Bayesian: calculate the probability of the parameter

Is US Economic Growth Higher Under Democratic Presidents than Republicans?

```
## Warning in loop_apply(n, do.ply): Removed 1 rows
containing non-finite values (stat_boxplot).
## Warning in loop_apply(n, do.ply): Removed 1 rows
containing missing values (geom_point).
```



Is US Economic Growth Higher Under Democratic Presidents than Republicans?

```
## Source: local data frame [2 x 4]
##
##   party     mean       sd length
## 1   Dem 3.094356 1.672123     22
## 2   Rep 1.725821 3.014028     28
```

Sampling Distribution of the Difference in Means

- Want to know $\mu_D - \mu_R$? (Difference in population means)
- What is the sample? What is the population?
- We will be making other dubious assumptions in this example: populations are independent, normal (not important).
- Estimate is $\bar{x}_D - \bar{x}_R$ (Difference in sample means)
- But sample is random, how do we characterize the uncertainty in our estimates?

Sampling Distribution of the Difference in Means

If we knew μ_D, μ_R, σ_R, σ_D, we could calculate the distribution of $\bar{x}_D - \bar{x}_R$.

$$\bar{x}_D - \bar{x}_R \sim N\left(\frac{\mu_D - \mu_R}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_D^2}{n_D}}}\right)$$

But we don't know the population ...

Logic of Significance Tests

- Assume null H_0 and alternative H_a hypotheses
- Calculate the sampling distribution of the test statistic assuming H_0 is true
- p-value is the probability of data (test statitics) equal or more extreme than the sample
- (optional) At a pre-defined significance level (α), reject H_0 if p-value less than α, fail to reject if p-value greater than α.

Significance Test

- Null hypothesis: $H_0 : \mu_D - \mu_R = 0$
- Alternative hypothesis: $H_a : \mu_D - \mu_R \neq 0$
- The test statistic is

$$t = \frac{\bar{x}_D - \bar{x}_R}{se} \tag{1}$$

  where

$$se = \frac{\sigma_D}{}$$

- Which is distributed

$$t \sim N\left(0, \frac{\sigma_R^2}{n_R} + \frac{\sigma_D^2}{n_D}\right)$$

- But don't know σ_R^2 and σ_D^2.
- Use s_R^2, s_D^2, t distributed Student's t to acccount for uncertainty from estimating standard deviations.

t-distribution

See |https://jrnold.shinyapps.io/tdist|

t-tests for difference of means in R

```
##
##  Welch Two Sample t-test
##
## data:  grgdpch by party
## t = 2.0366, df = 43.679, p-value = 0.04778
## alternative hypothesis: true difference in means is no
## 95 percent confidence interval:
##  0.01400377 2.72306582
## sample estimates:
## mean in group Dem mean in group Rep
##          3.094356          1.725821
```

Significance Tests

- Two approaches:
  - Fisher: p-value represents the level of evidence against H_0
  - Neyman-Pearson: choose a significance level α and reject null hypothesis if p-value is less than α.
- When making a decision of reject / not reject:
  - Type I error: H_0 true, reject H_0
  - Type II error: H_0 false, fail to reject H_0
- Power $1 - \mathrm{Pr}(\text{Type II error})$
- Tests generally focus on Type I error
  - "conservative" – is it really?
  - much harder to calculate Type II errors

How to report a confidence interval

- Either of
  - Democratic presidents enjoyed growth rates 1.37 points higher (95% CI: 0.01 to 2.72) than their Republican counterparts.
  - Democrats enjoyed 1.37 points higher growth than Republicans, with a 95 percent confidence interval of 0.01 to 2.72.
- We could calculate any CI we wish: 90 percent, 80 percent, 50 percent, etc.
- The most commonly used are: 90, 95, and 99.

Logic of Confidence Intervals

- Find a plausible range of values of the parameter: $[\bar{x}_{lower}, \bar{x}_{upper}]$
- Only know probability of data given parameter value, so cannot calculate a probability distribution for a parameter value (Bayesian approach)
- Frequentist approach: method to generate intervals which contain the true parameter μ in C% of the samples.

What a $100(1-\alpha)\%$ confidence interval means

Coverage A $100(1-\alpha)\%$ confidence interval for a parameter θ, is an interval generated by a method that generates intervals that include the true parameter θ in $100(1-\alpha)\%$ of samples.

Rejection Region A $100(1-\alpha)\%$ confidence interval such that $H_0 : \theta = \theta'$ cannot be rejected at the α significance level for all values of θ' in the interval, and $H_0 : \theta = \theta'$ is rejected for all values of θ' outside the interval. (not all confidence intervals have this property).

Confidence levels for difference in means

To get a $100(1-\alpha)\%$ confidence interval for a difference of means

$$\bar{x}_D - \bar{x}_R \pm t_{\alpha/2,\nu}\sqrt{\frac{s_D^2}{n_D} + \frac{s_R^2}{n_R}}$$

where $t_{\alpha/2,\nu}$ is a critical value of the t distribution such that the tails area of the distribution is α. The value of ν is complicated.

Conditional Probability

$$p(A|B) = \frac{p(A\&B)}{P(B)}$$

- What if A and B are independent? $P(A|B) = P(A)$
- What is the sampling distribution?

Bayes Rule

$$
\begin{aligned}
p(A|B) &= \frac{p(B|A)p(A)}{\sum_{A'} p(B|A')p(A')} \\
&= \frac{p(B|A)p(A)}{\sum_{A'} p(B|A')p(A')} \\
&\propto p(B|A)p(B)
\end{aligned}
$$

Confidence Intervals vs. Significance Tests

- Problems with **both**
  - Simply commitment to a certain error rate, given **assumptions**. Does not account for **model uncertainty**.
  - “File drawer problem”, “fishing”: even if it makes sense on an individual test, multiple testing within a research project + selecting on significant results can result in
- Problems with signficance tests that CI overcome
  - tests are “weak” -
  - confidence intervals focus more on substantive significance (parameter values); p-values ignore all substantive signifance.

Confidence Intervals vs. Significance Tests

- Confidence intervals often misinterpreted
- Definition of confidence interval is awkward and not exactly what we want, so often interpreted as probability interval
- But which is clearer?
  - Compared to Republicans, the effect of Democratic presidents on the economy is significantly positive at the 0.05 level.
  - Democratic presidents enjoyed 1.37 points higher growth than Republicans, with a 95 percent confidence interval of 0.01 to 2.72.

Inference and Bayes Rule

$$
\begin{aligned}
p(H|D) &= \frac{p(H|D)p(D)}{\sum_{H'} p(D|H')p(H')} \\
&= \frac{p(D|H)p(H)}{p(D)} \\
&\propto p(D|H)p(H)
\end{aligned}
$$

- $p(D|H)$ is the likelihood (related to the sampling distribution).
- Where does $p(H)$ come from?

Bayesian and Frequentist Statistics

In many research questions we are interested in the probability of the hypothesis H, given the data D: $p(H|D)$.

Frequentist: Assume a hypothesis, e.g. null hypothesis H_0, and calculate the probability of the data: $p(D|H_0)$

Bayesian: Assume a prior distribution $p(H)$ and calculate the probability of the hypothesis given the data:

$$p(H|D) \propto p(D|H)p(H)$$

My Claim: Even if using frequentist tests $p(D|H)$, a paper assigns prior probilities to hypotheses, e.g. lit review, logical arguments, etc. to make a Bayesian argument, $p(D|H)p(H)$.

Statistical Inference for OLS

- Individual Coefficients:
  - Significance tests: β_k
  - Confidence intervals
- Multiple coefficients:
  - Significance test
    - F-test on all slopes: $H_0 : \beta_1 = \ldots = \beta_k = 0$
    - F-test on subset of slopes: $H_0 : \beta_1 = \ldots = \beta_k = 0$
    - F-test on linear combinations of slopes: e.g $H_0 : \beta_1 - \beta_2 = 0$.
  - Confidence regions

Statistical Inference for Individual Coefficients for Simple Regression

If all assumptions of Gauss-Markov hold,

$$V(B) \sim N\left(\beta, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}\right)$$

If ϵ not normal, approximate as $n \to \infty$

- Test statistic for $H_0 : \beta = \beta_0$ distributed Student's t with $n - k - 1$ df.

$$t = \frac{\beta - \beta_0}{se}$$

- Confidence interval is

$$B \pm t_{\alpha/2, n-k-1} se$$

- standard error is $V(B)$ with $\hat{\sigma}_\epsilon^2$ as an estimate of σ_ϵ.

Estimate of σ_ϵ^2

Estimate σ_ϵ^2 from the regression errors

$$\hat{\epsilon} = \frac{\sum E_i^2}{n - k - 1}$$

- $n - k - 1 =$ (observations) − (variables) − (intercept) is the degrees of freedom.
- Similar to mean squared error, but to estimate population divide by degrees of freedom.

Statistical Inference for Individual Coefficients for Multiple Regression

If multiple variables, and Gauss-Markov assumptions hold, then

$$b \sim MVN\left(\beta, \sigma_\epsilon^2 \left(XX'\right)^{-1}\right)$$

- analagous to the bivariate version
    - calculates all standard errors simultaneously
    - covariances: b_i and b_j can be correlated

Statistical Inference for Individual Coefficients for Multiple Regression

Standard error for a single coefficient

$$SE(B_j) = \sqrt{se^2(X'X)^{-1}_{jj}} = \frac{1}{\sqrt{1 - R_j^2}} \frac{se}{\sqrt{\sum (x_{ij} - \bar{x}_j)^2}}$$

- Test statistic for $H_0 : \beta = \beta_0$ distributed Student's t with $n - k - 1$ df.

$$t = \frac{\beta - \beta_0}{se}$$

- Confidence interval is

$$B \pm t_{\alpha/2, n-k-1} se$$

- standard error is $V(B)$ with $\hat{\sigma}_\epsilon^2$ as an estimate of σ_ϵ.

Confidence Interval

General Definition

In repeated samples, C% of samples have a C% confidence interval that contains the population (true) parameter θ.

- Not a statement about a sample interval, statement about the method
- Each confidence interval either contains θ or not, there is no probability. Parameters are fixed, only samples are random.

Overlapping confidence intervals does not mean difference is not statistically significant

See `https://www.cscu.cornell.edu/news/statnews/Stnews73insert.pdf`

Significant and Not significant are not statistically significant

- Common example:
  - Regression with several dummy variables
  - Coefficient of dummy variable of category A (β_A) is significant at 5% level, dummy variable of category B (β_B) is not significant at the 5% level.
  - Common (wrong) interpretation: A and B are different
  - Correct procedures:
    - Sigificance test with $H_0 : \beta_A = \beta_B$
    - calculate confidence interval of $\beta_A - \beta_B$.

Statistical and Substantive Significance

- p-values are a function of estimated effect size (B) but also the sample size
- p-values only show statistical signifiance, not substantive significance.
- Confidence intervals can be more useful for displaying substantive significance

References

- Some slides derived from Christopher Adolph *Inference and Interpretation of Linear Regression*. Used with permission. <http://faculty.washington.edu/cadolph/503/topic4.pw.pdf>
- Material included from
    - Fox Ch 6, 9.3