Lecture 6: Statistical Inference

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

10/27/2016

Overview

1. Inference for Numerical Data
2. Inference for Categorical Data
3. Bias
4. Exercise - Inference for numerical and categorical data

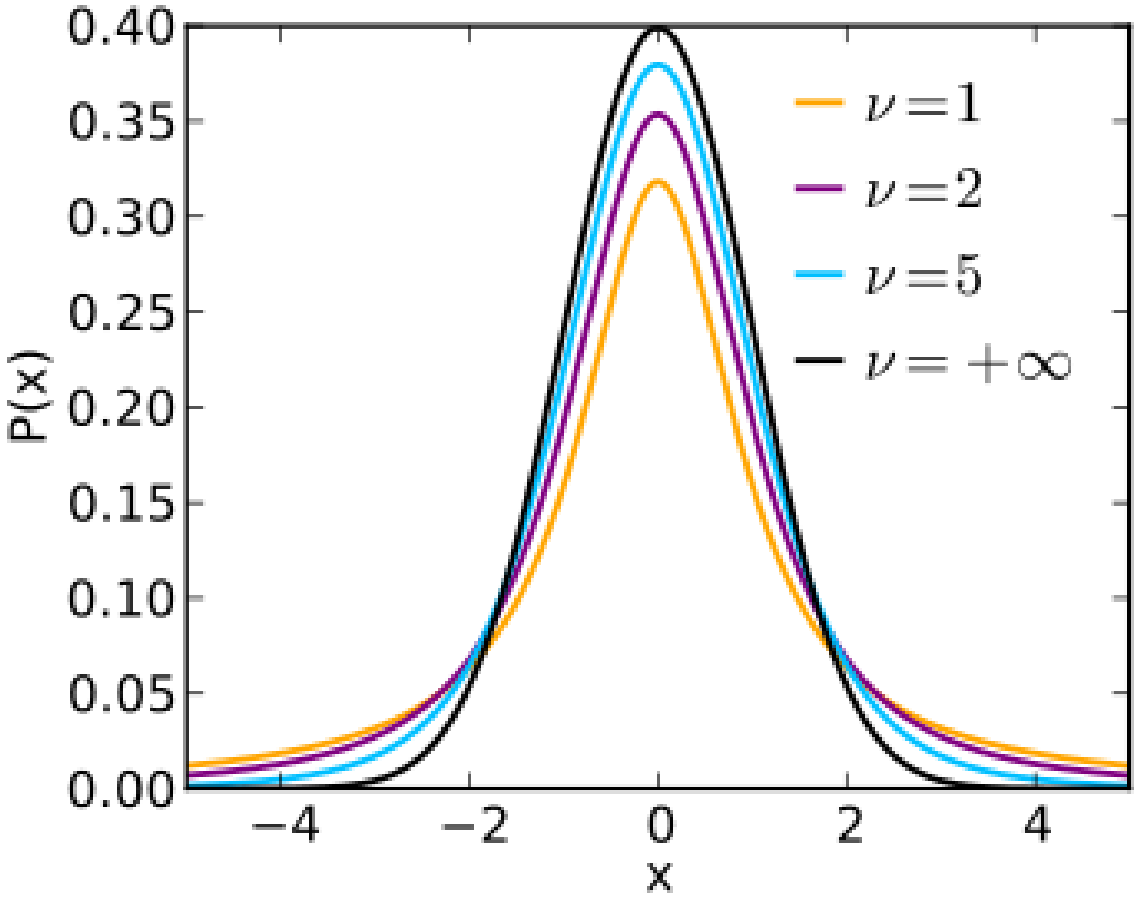
Statistical Inference

1. Determine which point estimate or test statistic is useful.
2. Identify an appropriate distribution for the point estimate or test statistic.
3. Create a confidence interval or hypothesis test using the chosen distribution.

Distributions

- Normal distribution: large sample, independent observations
- Student's t-distribution: small sample, independent observations, observations come from a nearly normal distribution
- F-distribution: Compare means of more than two groups using ANOVA
- χ^2 distribution: categorical data

Student's t-distribution



Degrees of freedom: $\nu = n - 1$

Test statistics: $T_{df} = \frac{\text{point estimate - null value}}{SE}$

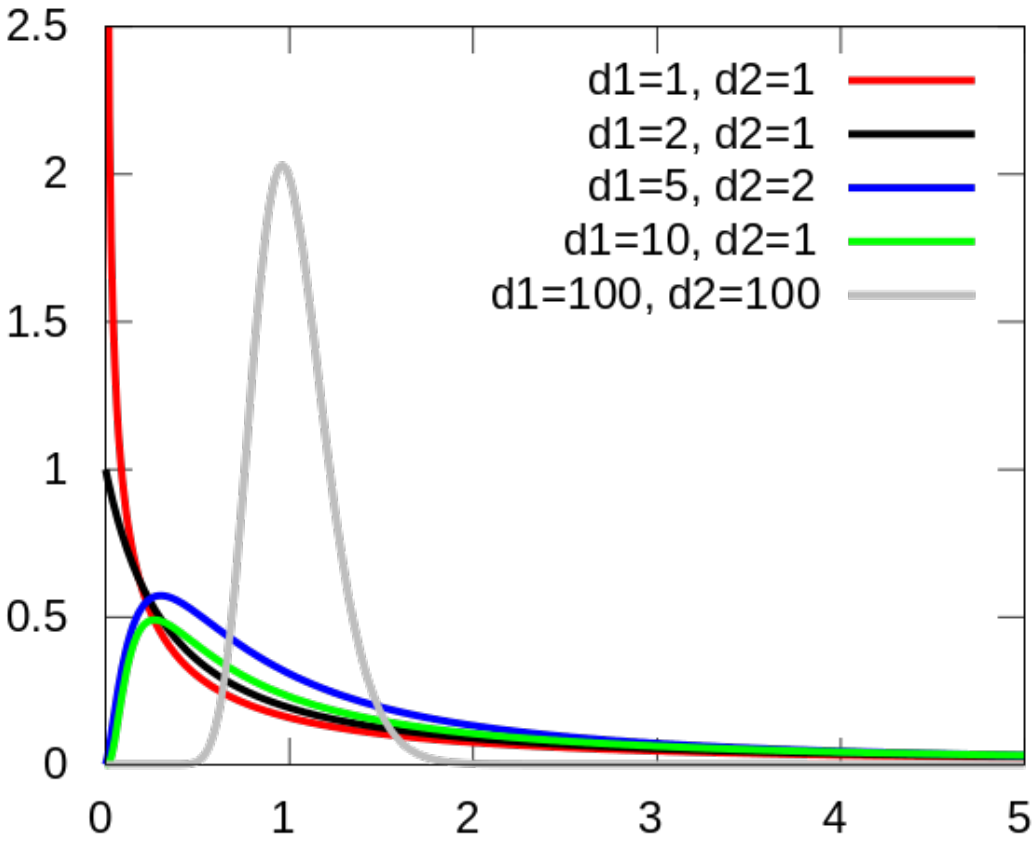
Analysis of Variability (ANOVA)

- Anova tests if means across many groups are equal.
- Null hypothesis: All means are equal. ($H_0 : \mu_1 = \mu_2 = ...$)
- Alternative hypothesis: All means are not equal.
- F statistic:

$$F = \frac{\text{Variation among sample means}}{\text{Variation within groups}}$$

- To reject H_0, $p - value < \alpha$, requires $F \gg 0$.
- ANOVA can only provide evidence that sample means are different among subgroups, but not which means are different.
- With an $\alpha = 0.05$, there is a 5% chance of Type 1 error for *each* ANOVA test performed. Performing multiple pair-wise tests to determine which sample means differ would lead to balooning error rate, so to find which means differ, use $\alpha^* = \alpha/K$ where $K = \frac{k(k-1)}{2}$, the number of possible pairs.

F-distribution



Degrees of freedom:

$d_1 = k - 1$, k is the number of groups

$d_2 = n - 1$, n is the total sample size

F-statistic

$$F = \frac{\text{Variation among sample means}}{\text{Variation within groups}} = \frac{MSG}{MSE}$$

$$SSG = \sum_{i=1}^{k} n_k(\bar{x}_i - \bar{x})^2$$

$$MSG = \frac{SSG}{k-1}$$

$$SST = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$SSE = SST - SSG$$

$$MSE = \frac{SSE}{d_2 - d_1}$$

Categorical Data

	nationality	response	year
0	Afghanistan	non-atheist	2012
1	Afghanistan	non-atheist	2012
2	Afghanistan	non-atheist	2012
3	Afghanistan	non-atheist	2012
4	Afghanistan	non-atheist	2012

What is the parameter of interest?

- **Parameter of interest**: Proportion of global population that is atheist, p.
- **Point estimate**: Proportion of sample who are atheist, $\hat{p}$.
- Confidence interval: $\hat{p} \pm ME = \hat{p} \pm \text{critical value} \times SE_{\hat{p}}$

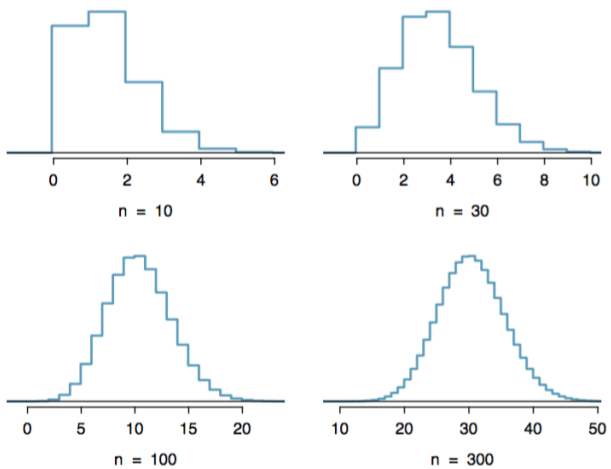
$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Central Limit Theorem for Proportions

Sample proportions are nearly normally distributed with mean equal to the population mean, p, and standard deviation equal to the standard error, $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Conditions

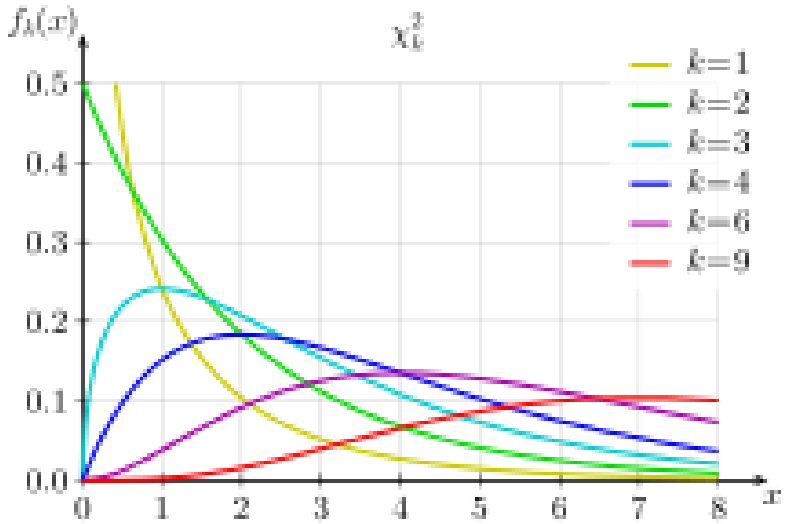
- Independent observations
- At least 10 each "successes" and "failures"



Binomial distribution with $p = 0.10$, n shown below histogram. [Diez, 2016]

χ^2 distribution

χ^2-test used for more than two categories.



hrefhttps://commons.wikimedia.org/w/index.php?curid=9884213By Geek3 - Own work, CC BY 3.0

Bias

- Selection
  - Data
  - Non-response (or voluntary response)
  - Convenience sample
- Confirmation
- Reporting
- Recall

References

 David Diez, Christopher Barr, & Mine Çetinkaya-Rundel (2015)
OpenIntro Statistics, OpenIntro

Recommended Reading

OpenIntro Statistics, Chapters 5-6

Exercise

Lesson6_StatisticalInference.ipynb

- Inference for numerical data
- Inference for categorical data