STA 104 Applied Nonparametric Statistics

Chapter 1: Introduction

Xiner Zhou

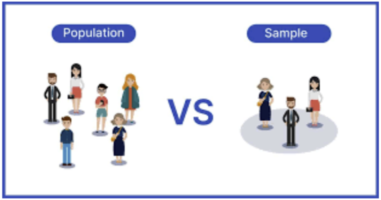
Department of Statistics, University of California, Davis

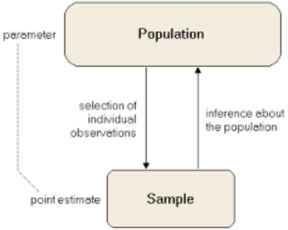
Table of contents

Parametric and Nonparametric Methods

Statistics or Statistical Inference:

to make inferences about a larger potentially observable collection of data called a population, using a sample.





We associate distributions with populations.

- families of distributions (parameters)
  - normal $\mathrm{N}\left(\mu, \sigma^2\right)$
  - $\mathrm{B}(n, p)$ distribution
  - Uniform
  - multinomial
  - Poisson
  - exponential
  - gamma
  - beta
  - Cauchy

Parametric methods:

Given a set of random sample from some population with a distribution that is assumed to be a member of a family such as the normal or binomial, to estimate or test hypotheses about the unknown parameters.

For a sample from a normal distribution

- sample mean is a point (i.e., a single value) estimate of the parameter μ
- CI: $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$
- z or t-test provides a measure of the strength of the evidence provided by a sample in support of an a priori hypothesized value μ_0 for population mean

Normal distribution? theoretical grounds, past experience

Central limit theorem justifies such a use of the normal distribution: asymptotic approximations

Parametric inference may be inappropriate or even impossible.

- no obvious family of distributions that provides our data
- no clearly defined parameters about which we can make inferences

$\Rightarrow$ Nonparametric methods= distribution-free methods

Nonparametric methods= distribution-free methods

Make inferences about parameters in wider sense:

> we do not assume our samples are associated with any prespecified family of distributions

Does not mean assumption free:

> always make some assumptions about the underlying population distribution
>
> e.g. Nonparametric test: methods can be applied to samples from populations having distributions only specified in broad terms
>
> - e.g., as being continuous, symmetric
> - distribution-free property: distribution of test statistic is the same no matter what the population distribution may be

Robust:

> do not depend critically on the correctness of an assumption that samples come from a distribution in a particular family

Historical Notes

- Francis Galton (1892): developed a method for classifying and assess agreement between patterns (categorical data) on fingertips

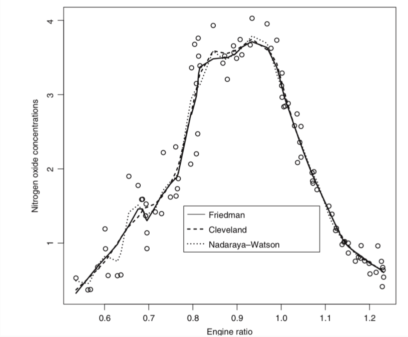
- Karl Pearson (1900): chisquared goodness-of-fit test applicable to any discrete distribution
- Spearman (1904): rank correlation coefficient
- Ronald Fisher and E.J.G. Pitman in the 1930s: permutation tests

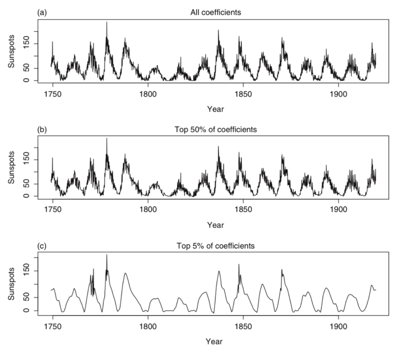
Historical Notes

- Friedman, Smirnov, Wilcoxonin 1930s
  - observations consisting simply of preferences or ranks could be used in permutation tests to make some inferences
  - even if we have precise measurements, we sometimes lose little useful information by ranking them in increasing order of magnitude and basing analyses on these ranks.
  - when assumptions of normality are not justified, analyses based on ranks may be the most efficient available and robust
- Hodges and Lehmann (1963) : interval estimation
- modern, computer intensive procedures of bootstrapping introduced by Efron (1979)

Historical Notes

- Nonparametric Regression
  - Local averaging
  - Local regression
  - Kernel smoothing
  - Wavelets





Nonparametric advantage?

ill-founded hopes that data would fit a restricted mathematical model with few parameters, and emphasis on simplifying concepts such as linearity, have often been replaced by the use of robust methods.

Strength: when insufficient theory or data to justify, or to test compatibility with, specific distributional models.

Binomial Problem

Setting

We observe the outcomes of n independent Bernoulli trials [1].

- The outcome of each trial can be classified as a success or a failure.
- The probability of a success, denoted by p, remains constant from trial to trial.
- The n trials are independent.

[1] Research problems possessing these assumptions underpinnings are common, and thus the binomial test procedures find frequent use. A particularly important special case occurs when we wish to test hypotheses about the unknown median of a population. The application of binomial test leads to the sign test.

Hypothesis

Two-Sided Test:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

One-Sided Upper-Tail Test:

$$H_0 : p = p_0 \text{ versus } H_a : p > p_0$$

One-Sided Lower-Tail Test:

$$H_0 : p = p_0 \text{ versus } H_a : p < p_0$$

Motivation

Motivation

Motivation

Motivation

Procedure

a. One-Sided Upper-Tail Test.

To test

$$H_0 : p = p_0$$

versus

$$H_1 : p > p_0$$

at the α level of significance,

Reject H_0 if $B \geq b_\alpha$; otherwise do not reject, where the constant b_α is chosen to make the type I error probability equal to α. The number b_α is the upper α percentile point of the binomial distribution with sample size n and success probability p_0.

The normal approximation:

Reject H_0 if $B^* \geq z_\alpha$; otherwise do not reject.

Procedure

b. One-Sided Lower-Tail Test.

To test

$$H_0 : p = p_0$$

versus

$$H_2 : p < p_0$$

at the α level of significance, Reject H_0 if $B \leq b_{1-\alpha}$; otherwise do not reject.

The normal approximation:

Reject H_0 if $B^* \leq -z_\alpha$; otherwise do not reject.

Procedure

c. Two-Sided Test.

To test

$$H_0 : p = p_0$$

versus

$$H_3 : p \neq p_0$$

at the α level of significance, Reject H_0 if $B \geq b_{\alpha/2}$ or $B \leq b_{1-\alpha/2}$; otherwise do not reject

The normal approximation:

Reject H_0 if $|B^*| \geq z_{\alpha/2}$; otherwise do not reject.

Example

Sensory Difference Tests

The triangle test is principally used to determine whether a sensory difference exists between two products. For example to test whether a change to a product's ingredient, process or packaging has had an impact on it's overall sensory properties.

To conduct the test; the assessors compare the test product against a control product. Each assessor is presented with three samples (two are the same and one is different) and asked to evaluate the samples from left to right, select the "different" sample and describe the difference perceived.

Out of 50 trials, there were 25 correct selections and 25 incorrect selections.

Example

We consider the binomial test of $H_0 : p = \frac{1}{3}$ versus the one-sided alternative $p > \frac{1}{3}$. We set $\alpha = .05$ for purposes illustration.

Asymptotic approximation:

(1) critical value approach:

To find $z_{.05}$, the 95^{th} upper quantile of the standard normal

```
> qnorm (.95, 0, 1)
[1] 1.644854
```

we find $z_{.05} = 1.645$. Thus Reject H_0 if $B^* \geq 1.645$; otherwise do not reject.

Example

From the data we have $n = 50$ and B (the number of correct identifications) $= 25$.

$$B^* = \frac{25 - 50\left(\frac{1}{3}\right)}{\left\{50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right\}^{1/2}} = 2.5.$$

The large sample approximation value $B^* = 2.5 > 1.645$ and thus we reject $H_0 : p = \frac{1}{3}$ in favor of $p > \frac{1}{3}$ at the approximate $\alpha = .05$ level.

Example

(2) p-value approach: To find the P-value corresponding to $B^* = 2.5$

```
> 1 -pnorm(2.5)
[1] 0.006209665
```

Example

Exact test: (1) critical value approach:

To find $b_{.05}$, the 95^{th} upper quantile of the binomial distribution

```
> qbinom(0.05, size=50, prob=1/3, lower.tail =FALSE)
[1] 22
```

we find $b_{.05} = 22$. Thus Reject H_0 if $B \geq 22$; otherwise do not reject.

Example

From the data we have $n = 50$ and B (the number of correct identifications) $= 25$.

Thus we reject $H_0 : p = \frac{1}{3}$ in favor of $p > \frac{1}{3}$ at the approximate $\alpha = .05$ level.

Example

(2) p-value approach:

To find the P-value corresponding to $B^* = 2.5$

```
> pbinom(24,50,1 / 3, lower.tail =FALSE)
[1] 0.01082668
```

Conclusion: From both the exact test and large-sample test, there is statistically significant evidence that the food being tested tastes differently.

An estimator for the probability of success with uncertain quantification

Point estimate:

$$\hat{p} = \frac{B}{n}$$

Observed Relative Frequency of Success.

Standard errorof $\widehat{p}$

$$\widehat{sd}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

measures the variability of the estimate.

Confidence interval

A $(1-\alpha)100\%$ confidence interval:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$