Experiment Planning and Design

Lecture 4: Statistical Concepts

Claus Aranha
caranha@cs.tsukuba.ac.jp

Department of Computer Science

2015-05-12

Notes

- Sorry about the sickness last class; Let's try Class 3 again!
- No class on May 19th and May 26th;

Class Outline

- Random Variables
- Point Estimators
- Interval Estimators
- Hypothesis Testing

The goal of this class is to allow you to do a simple analysis of the data going into, and coming out of an experiment.

Introduction: Probability vs Statistics

Probability

Given the pool, what are the odds of drawing a combination of certain colors?

Statistics

Given the colors of a few balls drawn, what can I know about the pool?

Statistical Inference: Using *samples* to draw conclusions about *populations*

Population, Sample and Observation

"A population is a large set of objects of a similar nature which is of interest as a whole". It can be an actual set (all balls in the pool), or an hypothetical one (all possible outcomes for an experiment).

A sample is a subset of a population. "A sample is chosen to make inferences about the population by examining or measuring the elements in the sample"

An observation is a single element of a given sample, an individual data point. An observation can also be considered as a sample of size one.

Glossary of statistical terms: `http://www.statistics.com/glossary`

Population, Sample and Observation

Let's remember Alice and Bob's experiments

Alice and Bob build spam filter programs. They test their programs by counting how many spam the system catches in a day.

Observation

Sample

Population

Population, Sample and Observation

Let's remember Alice and Bob's experiments

Alice and Bob build spam filter programs. They test their programs by counting how many spam the system catches in a day.

Observation

If we count the number of spam caught by a system in one day, that is one observation.

If we count the number of spam caught by a system another day, that is a second observation

Sample

Population

Population, Sample and Observation

Let's remember Alice and Bob's experiments

Alice and Bob build spam filter programs. They test their programs by counting how many spam the system catches in a day.

Observation

Sample

If we count the number os spam caught every day for a week, we will have seven observations. That is a Sample

Population

Population, Sample and Observation

Let's remember Alice and Bob's experiments

Alice and Bob build spam filter programs. They test their programs by counting how many spam the system catches in a day.

Observation

Sample

Population

If we know ALL possible results for ALL possible days, that is the Population

In practice, it is usually impossible to KNOW the population, but we want to learn as much as possible from it, by observing samples.

Point and Interval Estimates

Two central concepts of Statistical Inference are point estimators and statistical intervals

Both terms refer to the idea of using information obtained from a sample to infer values about parameters of the population.

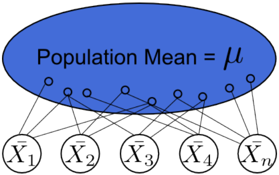
- Point Estimate: Estimate a value for a given population parameter
- Statistical Interval: Estimate a interval of possible/probable values for a given population parameter;

Point Estimates, Statistics, and Sampling distributions

Suppose one wants to obtain a point estimate for the mean of a given population. We take a sample of the population, and calculate the mean of that sample.

However, a random sample from a population results in a random variable! Any function of the sample - any *statistic* - is also a random variable.

This means that statistics calculated from samples will also have their own probability distributions, called sampling distributions.



See D.W. Stockburger: http://www.psychstat.missouristate.edu/introbook/sbk19.htm

I heard you like statistics!

So in order to specify parameters of the population (such as means, deviation, etc), we draw a random sample and calculate the parameters from it.

But because the sample is random, the parameter calculated from the sample will also have its own statistics!

Everything is easier with an R example

```
> population <- rnorm(100) # Pretend you don't know this!
> x1 <- sample(population,5)
> x2 <- sample(population,5)
> x3 <- sample(population,5)
> x1
[1]  0.6028260  0.1333065  1.1145946 -0.8675467 -0.4329469
> c(pop=mean(population),x1=mean(x1),x2=mean(x2),x3=mean(x3))
        pop          x1          x2          x3
 0.05722922  0.11004669 -0.10459150  0.12630965
> c(mean(c(mean(x1),mean(x2),mean(x3))),sd(c(mean(x1),mean(x2),mea
[1] 0.04392161 0.12887292
```

Point Estimators

A Point Estimator is a statistic which provides the value of maximum plausibility for a given (unknown) population parameter θ.

Consider a random variable X distributed according to a given $f(X|\theta)$ (a population which distribution is controlled by this parameter)

Now consider also a random sample from this variable: $x = \{x_1, x_2, \ldots, x_N\}$;

A given function $\hat{\Theta} = h(x)$ is called a *point estimator* of the parameter θ, and a value returned by this function for a given sample is referred to as a *point estimate* $\hat{\theta}$ of the parameter.

What does this mean?

A Point Estimator is a function that, given a sample, generates an estimated parameter for the distribution from which the sample was obtained.

Point Estimators

Point estimation problems arise frequently in all areas of science and engineering, whenever there is a need for estimating a parameter of a population:

- The population mean, μ;
- The population variange, σ^2;
- a population proportion, p;
- the difference in the means of two populations, $\mu_1 - \mu_2$;
- etc...

For each cases (and many others) there are multiple ways of performing the estimation task. We choose the estimators based on its statistics.

Multiple estimators?

We always consider only one definition for estimators (e.g., the mean). But we can be creative and invent others!

$$\mu = \sum_{i=0}^{N} \frac{x_i}{N}$$

$$\mu' = \frac{max(x) - min(x)}{2}$$

Evaluating Estimators

A good estimator should consistently generate estimates that are close to the real value of the parameter θ.

We say that an estimator $Th\hat{e}ta$ is unbiased for a parameter θ if:

$$E[\hat{\Theta}] = \theta$$

or, equivalently:

$$E[\hat{\Theta}] - \theta = 0.$$

The difference $E[\hat{\Theta}] - \theta$ is referred as the bias of an estimator.

Evaluating Estimators

The usual estimators for mean and variance are unbiased estimators;
Let $x_1, \dots, x_N$ be a random sample from a given population X, which is characterized by its mean μ and variance σ^2. In this situation, it is possible to show that:

$$E[\bar{x}] = E[\frac{1}{N}\sum_{i=1}^{N} x_i] = \mu$$

and:

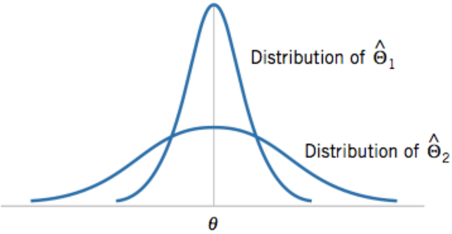
$$E[s^2] = E[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2] = \sigma^2$$

See this link for an example proof:

`http://isites.harvard.edu/fs/docs/icb.topic515975.files/Proof%20that%20Sample%20Variance%20is%20Unbiased.pdf`

Evaluating Estimators (2)

There usually exists more than one unbiased estimator for a parameter θ. One way to choose which to use is to select the one with the smallest variance. This is generally called the *minimal-variance unbiased estimator* (MVUE).



MVUE have the ability of generating estimates $\hat{\theta}$ that are relatively close to the real value.

Distribution of samples

Even for an arbitrary population, the sampling distribution of means tends to be approximately normal (with $E[\bar{x}] = \mu$ and $s_{\bar{x}} = \sigma^2/N$

Warning! Maths!

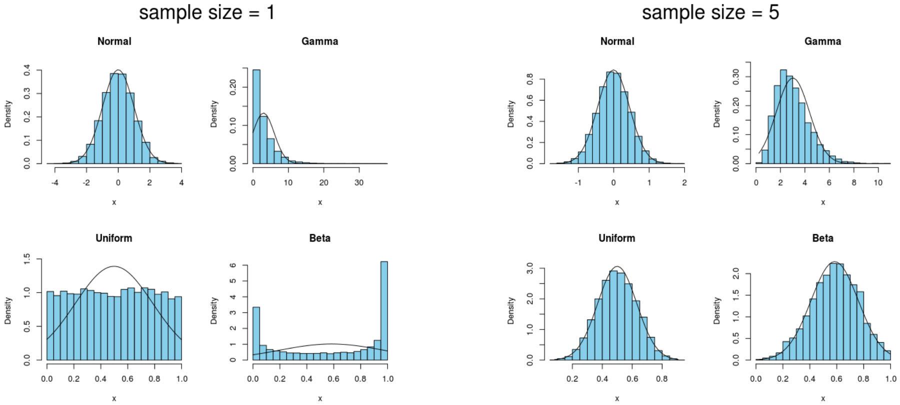
More generally, let $x_1, \ldots, x_n$ be a sequence of independent and identically distributed (iid) random variables, with mean μ and finite variance σ^2. Then:

$$z_n = \frac{\sum_{i=1}^{n}(x_i) - n\mu}{\sqrt{n\sigma^2}}$$

is distributed approximately as a standard normal variable. That is, $z_n \ N(0,1)$

This is the Central Limit Theorem.

Example of the Central Limit Theorem




```
# Load the Teaching Demos library if you don't have it.
> install.packages("TeachingDemos")
> library(TeachingDemos)

> clt.examp()
> clt.examp(5)
```

Homework 2

- Pool of balls image: `http://goo.gl/y8doaN`
- Green ball: `http://goo.gl/Fb8z68`
- MVUE image: D.C.Montgomery, G.C. Runger, "Applied Statistics and Probability for Engineers", Wiley 2003