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.

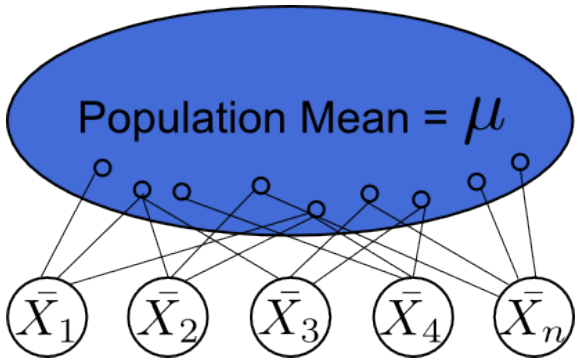
- Point Estimate: Estimate a value for a given population parameter
- Statistical Interval: Estimate a interval of possible/probable values for a given population parameter;

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 $\hat{Theta}$ 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 $x_1, \ldots, x_N$ 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:

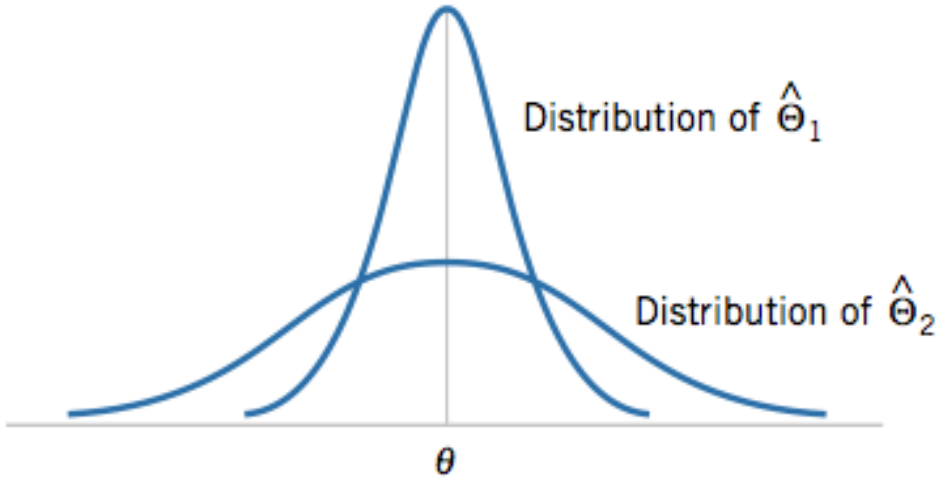
$$E[s^2] = E[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2] = \sigma^2$$

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!

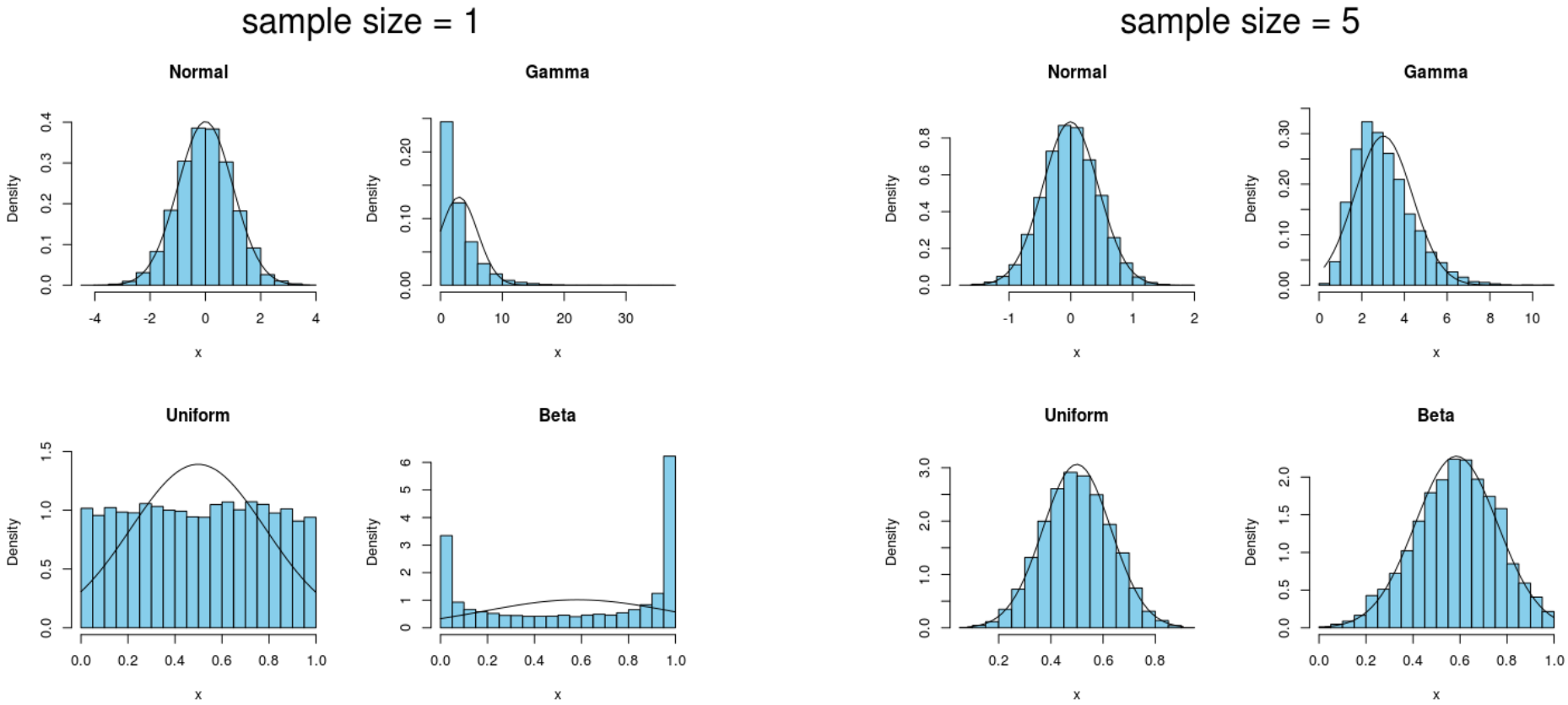
More generally, let $x_1, \ldots, x_n$ be a sequence of independent and identically distributed (iid) random variables, with mean μ and finite variance σ^2. Then:

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

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)
```

Implications of the Central Limit Theorem

The CLT is one of the most useful properties for statistical inference. The CLT allows the use of techniques based on the Gaussian distribution, even when the population under study is not normal.

For “well-behaved” distributions (continuous, symmetrical, unimodal) even small sample sizes are enough to justify invoking the CLT and using parametric techniques.

For an interactive demonstration of the CLT, check:
`http://drwho.cpdee.ufmg.br:3838/CLT/`

mini-break, questions?

Statistical Intervals

Statistical Intervals are important in quantifying the uncertainty associated to a given estimate;

Example: Coaxial cable factory

A coaxial cable manufacturing operation produces cables with a target resistance of 50Ω and a standard deviation of 2Ω. Assume that the resistance values of the cables produced can be well modeled by a normal distribution.

Suppose that we take a sample of $N = 25$ cables produced, and the sample mean is $\bar{x} = 48$. Given the variability of the sample, it is likely that this value is not exactly the true value μ.

How can we quantify the uncertainty of this estimate?

Definition of Statistical Intervals

Statistical Intervals define regions that are likely to contain the true value of an estimated parameter.

More formally, it is generally possible to quantify the level of uncertainty associated with the estimation, which allows the derivation of sound conclusions at predefined levels of certainty.

Example

We estimate that the value of the mean of this population is between 5.3 and 7.8, and we have a 95% confidence on the method used to generate this interval.

The most common types of interval are:

1. Confidence Intervals;
2. Tolerance Intervals;
3. Prediction Intervals;

Confidence Intervals

Confidence Intervals quantify the degree of uncertainty associated with the estimation of the population parameter, such as the mean or the variance.

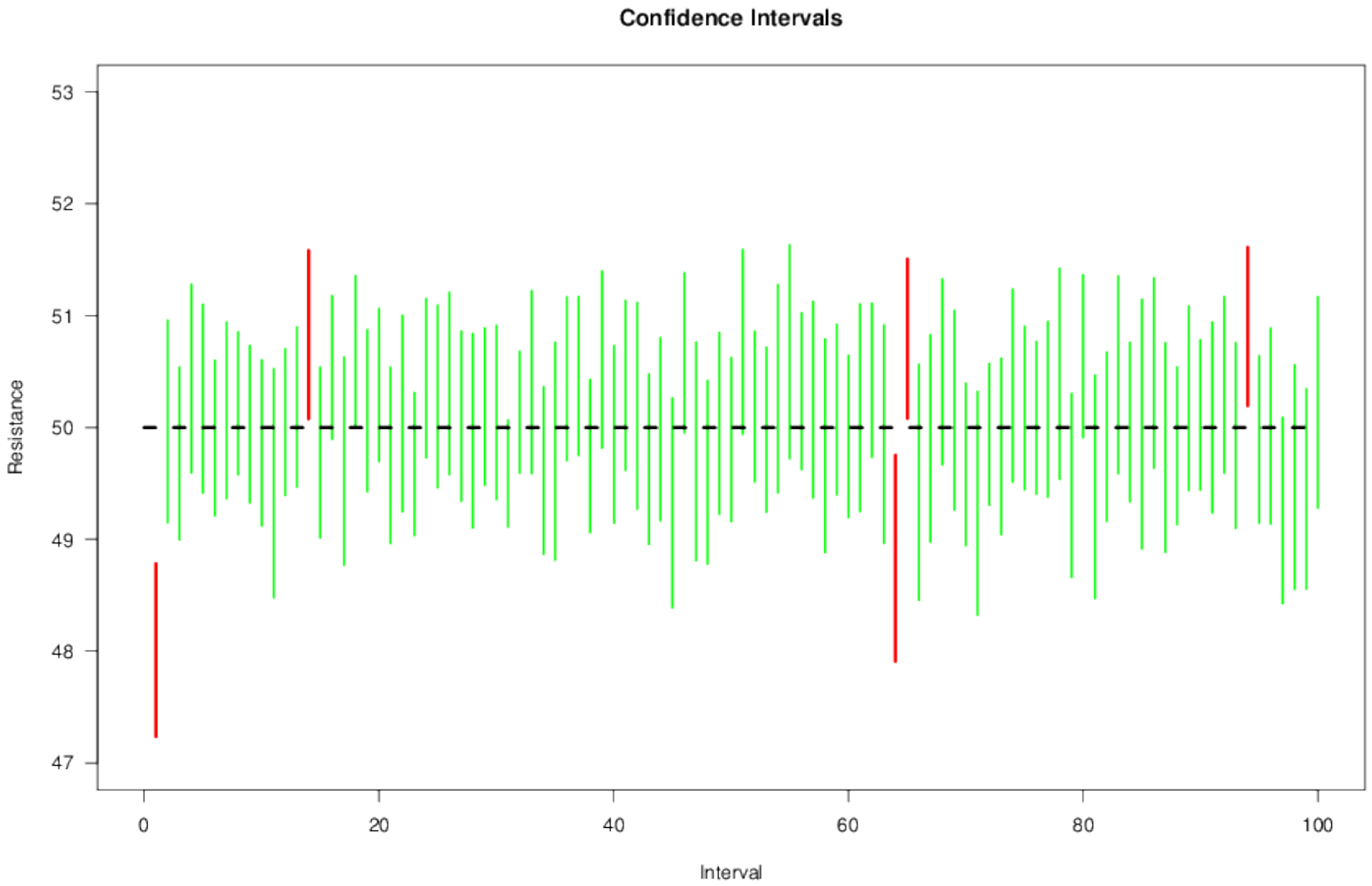
Definition

"The interval that contains the true value of a given population with a confidence level of $100(1-\alpha)$"

- Wrong: "there is a 95% chance that the interval contains the true population mean":
- Right: "The method used to derive the interval has a hit rate of 95%" - i.e., the interval generated has a 95% chance of capturing the true population parameter;

It is easier to understand if you think about confidence in the **method**, not in the interval.

Example, 100 $CI_{.95}$ for a sample of 25 observations



CI on the Mean of a Normal Variable

The two-sided $CI_{(1-\alpha)}$ for the mean of a normal population with known variance σ^2 is given by:

$$\bar{x} - \frac{\sigma}{\sqrt{N}} Z_{(\alpha/2)} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{N}} Z_{(\alpha/2)}$$

where $(1-\alpha)$ is the confidence level and $z_{(\alpha/2)}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.
For the more usual case with an unknown variance,

$$\bar{x} - \frac{s}{\sqrt{N}} t_{(\alpha/2;N-1)} \leq \mu \leq \bar{x} + \frac{s}{\sqrt{N}} t_{(\alpha/2;N-1)}$$

where $t_{(\alpha/2;N-1)}$ is the corresponding quantile of the t distribution with $N-1$ degrees of freedom.

CI on the Variance of a Normal Variable

In the same way, a two-sided confidence interval on the variance of a normal variable can be easily calculated:

$$\frac{(N-1)s^2}{\chi^2_{\alpha/2;N-1}} \leq \sigma^2 \leq \frac{(N-1)s^2}{\chi^2_{1-\alpha/2;N-1}}$$

where $\chi^2_{\alpha/2;N-1}$ and $\chi^2_{1-\alpha/2;N-1}$ are the upper and lower $(\alpha/2)$-quantiles of the χ^2 distribution with $N-1$ degrees of freedom.

Calculating the CI with R

Remember that we don't want to do all these calculations by hand! It is important to understand what they mean, but in practice you will do something like this:

```
> population <- rnorm(5000) # our hypotethical population
> sample.size = 20
> x1 <- sample(population,sample.size) # replace with experiment

> mean.estimator <- mean(x1)
> sd.estimator <- sd(x1)

> left <- mean.estimator - (sd.estimator/sqrt(sample.size))*qt(0.9
> right <- mean.estimator + (sd.estimator/sqrt(sample.size))*qt(0.

> c(left,right)
[1] -0.5218866  0.3356534
```

What if we want a smaller Interval?

One way to decrease the size of the confidence interval, without losing confidence, is increasing the size of a sample. This has its own problems which we will see in the future (e.g. cost of sampling).

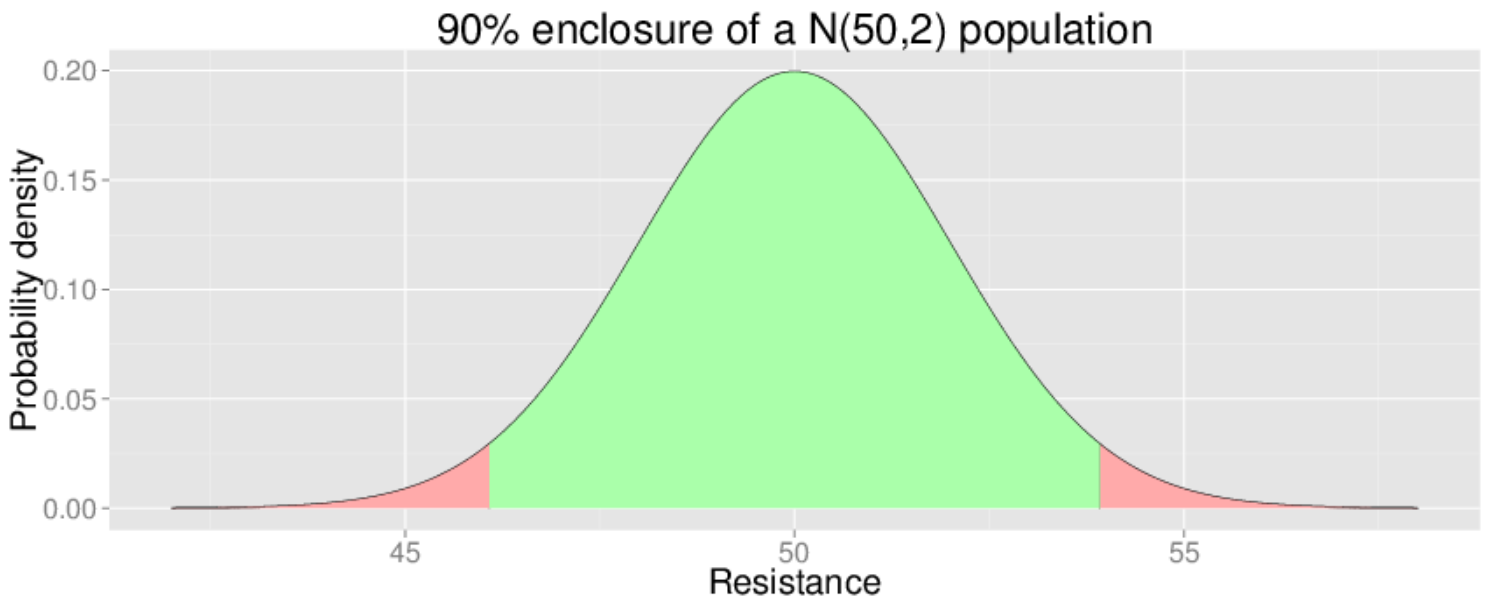
```
> population <- rnorm(5000) # our hypotethical population
> sample.size = 100 # INCREASED SAMPLE SIZE
> x1 <- sample(population,sample.size) # replace with experiment

> mean.estimator <- mean(x1)
> sd.estimator <- sd(x1)

> left <- mean.estimator - (sd.estimator/sqrt(sample.size))*qt(0.9
> right <- mean.estimator + (sd.estimator/sqrt(sample.size))*qt(0.

> c(left,right)
[1] -0.1163320  0.2011607 # DECREASED CONFIDENCE INTERVAL
```

Tolerance Intervals

"A tolerance interval is a **enclosure** interval for a specified proportion of the samplede population, not its mean or standard deviation. For a specified confidence level, you may want to determine lower and upper bounds such that a given percent of the population is contained within them."

J.G. Ramirez: `https://www.sas.com/resources/whitepaper/wp_4430.pdf`

90% enclosure of a N(50,2) population



Tolerance Intervals

The common practice in engineering of defining specification limits by adding $\pm 3\sigma$ to a given estimate of the mean arises from this definition - for a normally-distributed population, approximately 99.75% of the observations will fall within these limits.
However, in most cases the true variance is unknown. So we have to use its estimate, s^2, and compensate for the uncertainty in this estimation. The two-sided tolerance interval is given as:

$$\bar{x} \pm \sqrt{(N-1)(N+z^2_{(\alpha/2)})N\chi^2_{(\gamma;N-1)}}$$

In which γ is the proportion of the population to be enclosed, and $1-\alpha$ is the desired confidence level for the interval.

Another Example

We are bulding a program that should finish its operation in between 10 and 30μs. An initial analysis would be to run the programs a few times to calculate the tolerance interval for its running time.

```
> runtime <- c(14.92869, 13.65345, 14.63093, 14.38412,
               14.98059, 13.92460, 14.81254, 14.26117
               13.31676, 19.80000)
> df <- length(runtime)
> prop <- 0.9
> conf <- 0.95
> spread <- sqrt(((df-1)*(df+qnorm(conf/2)^2))/(df*qchisq(prop,df=df-1)))
> left <- mean(runtime) - spread
> right <- mean(runtime) + spread
> c(left,right)
[1] 14.08623 15.64972
```

Prediction Intervals

Prediction intervals quantify the uncertainty associated with forecasting the value of a future observation;

Essentially, one is intersted in obtaining an interval within which he or she can declare that the next observation will fall with a given probability;

For a normal distribution, we have:

$$\bar{x} - t_{(\alpha/2;N-1)} s/sqrt1 + \frac{1}{N} \leq X_{N+1} \leq \bar{x} + t_{(\alpha/2;N-1)} s/sqrt1 + \frac{1}{N}$$

which is similar to the confidence interval for the mean, but adding 1 to the term within the square root to account for the prediction noise.

Wrapping up

Statistical intervals quantify the uncertainty associated with different aspects of estimation;

Reporting intervals is always better than point estimates, as it provides to you (and your readers) the necessary information to quantify the location and spread of your estimated values;

The correct interpretation is a little tricky, but it is essential in order to derive the correct conclusions based on the statistical interval of interest.

Related reading:

- J.G. Ramirez, Statistical Intervals: Confidence, Prediction, Enclosure: `https://www.sas.com/resources/whitepaper/wp_4430.pdf`
- D.C. Montgomery and G.C. Runger, “Applied Statistics and Probability for Engineers”, chapter 8, 3rd Ed., Wiley 2005.

Homework 2

Basic Data analysis

- Load two data files related to your research into R data frames (experiment results, data sets, etc);
- Find out, for each data set: variables, their means, variances, maximum and minimum values;
- Plot relevant plots to characterize these data sets;
- Compare these data sets using statistical estimators (point estimators, interval estimators, etc)
- Describe your findings;

Submission materials

- Files 1..n: text files containing the data used;
- File n+1: R file (text file) containing the tasks above; R file must contain comments explaining what each command block does.

If you don't have useable data related to your research, you can use the following data sets:

Credits

- 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

This lecture notes is a derived work of

Felipe Campelo (2015), “Lecture Notes on Design and Analysis of Experiments”
Online: `https://github.com/fcampelo/Design-and-Analysis-of-Experiments`