

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 of 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.

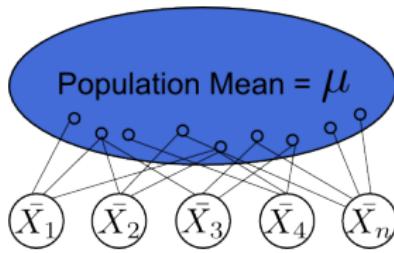
- Point Estimate: Estimate a value for a given population parameter
- Statistical Interval: Estimate a interval of possible/probable values for a given population parameter;

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:

$$x = \{x_1, x_2, \dots, x_N\};$$

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 variance, 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, \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:

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

and:

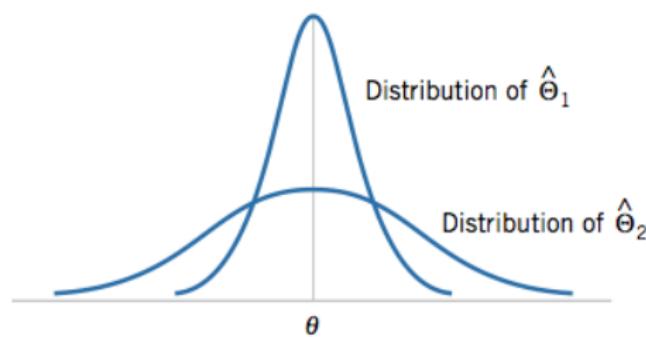
$$E[s^2] = E\left[\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2\right] = \sigma^2$$

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, \dots, x_n be a sequence of independent and identically distributed (iid) random variables, with mean μ and finite variance σ^2 . Then:

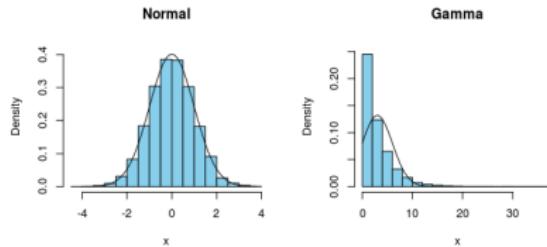
$$z_n = \frac{\sum_{i=1}^n (x_i) - n\mu}{\sqrt{n\sigma^2}}$$

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

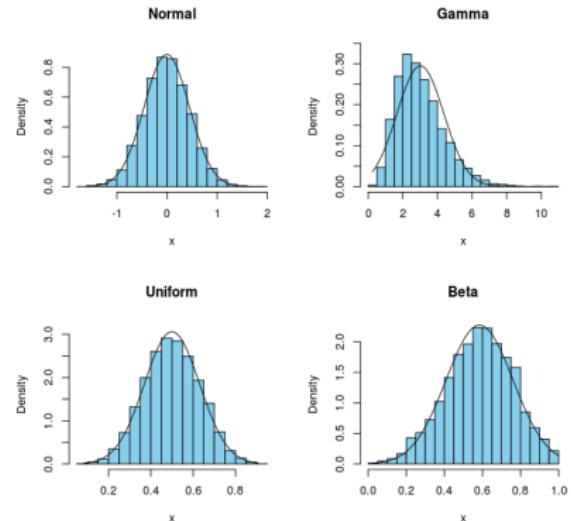
This is the **Central Limit Theorem**.

Example of the Central Limit Theorem

sample size = 1



sample size = 5



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

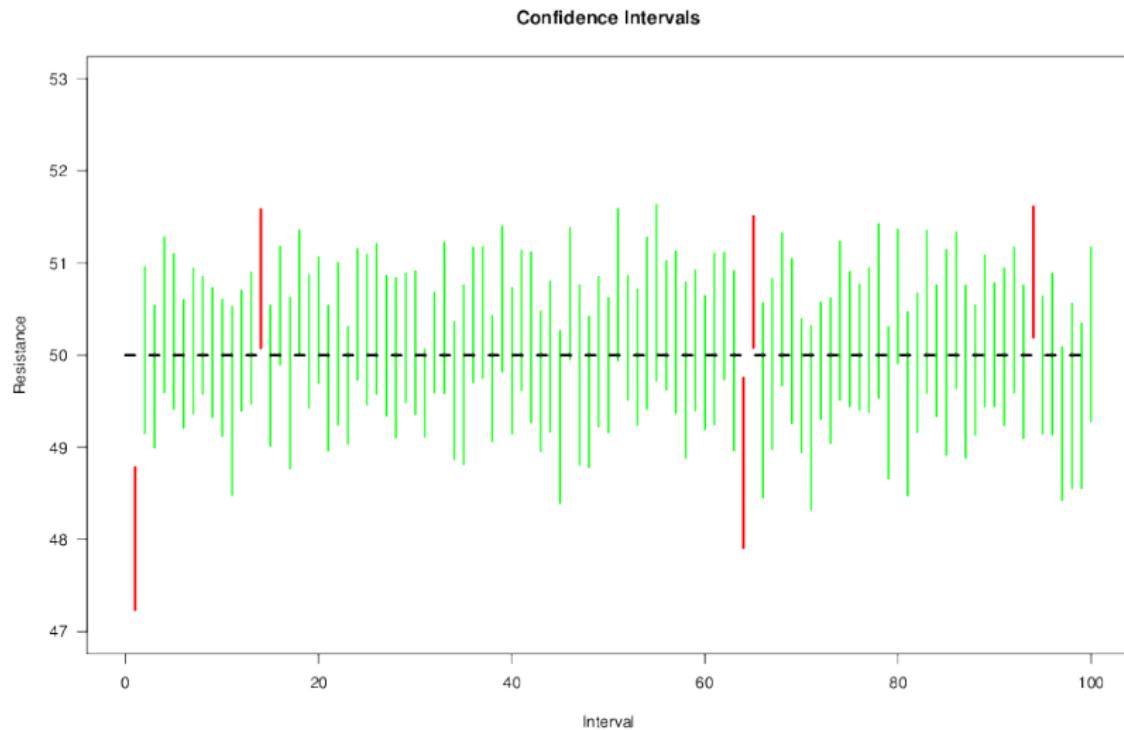
Definition

“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:

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

where $\chi_{\alpha/2; N-1}^2$ and $\chi_{1-\alpha/2; N-1}^2$ 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 hypothetical 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(0.975)  
> right <- mean.estimator + (sd.estimator/sqrt(sample.size))*qt(0.975)  
  
> 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 hypothetical 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(0.975)
> right <- mean.estimator + (sd.estimator/sqrt(sample.size))*qt(0.975)

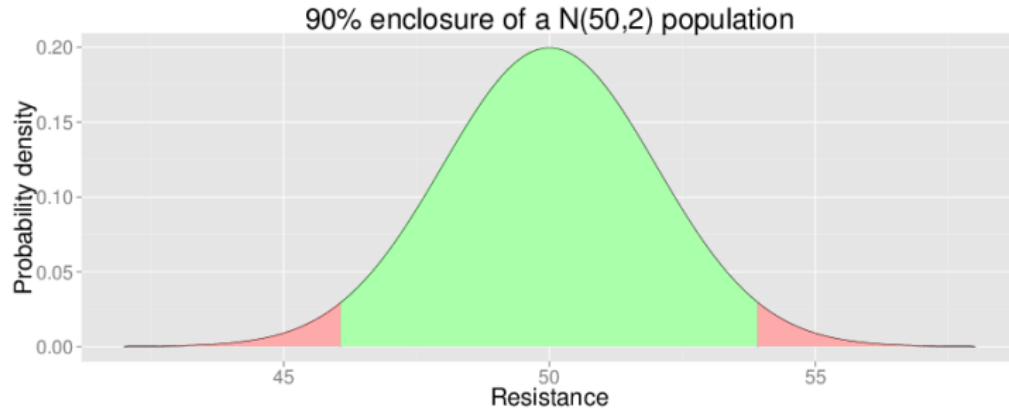
> c(left,right)
[1] -0.1163320 0.2011607 # DECREASED CONFIDENCE INTERVAL
```

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:

$$\bar{x} \pm \sqrt{(N-1)(N+z_{(\alpha/2)}^2)} N_{\chi_{(\gamma;N-1)}^2}$$

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 building a program that should finish its operation in between 10 and $30\mu s$. 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 interested 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:

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

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.

Did you understand everything?
Another mini-break!

Introduction to Statistical Inference

Point Estimators and Statistical Intervals belong to a branch of statistics known as *descriptive statistics*. That is, these methods are focused on accurately describing characteristics such as the location and uncertainty about a given population parameter;

While these concepts are certainly important, in many cases description is not enough – someone may need the decision-making tools to deal with information from random samples, tools that allow a research to perform **inference** with a quantifiable degree of certainty.

Statistical Hypotheses

A **hypothesis** is a proposed explanation for an observable phenomenon.
Scientific hypotheses must satisfy (at least) two conditions:

- Testability;
- Falsifiability;

The **hypothetico-deductive** model of construction of scientific knowledge includes:

- Formulation of a falsifiable hypotheses;
- Refutation or corroboration of the hypotheses by the data;
- Comparison between alternative hypotheses – principle of parsimony (Ockham's razor);
- Predictive power;

Definitions of Statistical Hypotheses

Statistical hypotheses are defined as objective statements about parameters of one or more populations;

Attention: the statements in statistical hypotheses are about parameters of the population or model, **not the sample.**

On frequentist approaches, the formal test of hypotheses involves the contrast between null and alternate hypotheses.

Null hypothesis (H_0)

- Absence of effects;
- Conservative model;
- Point value for the parameter.

Example: $H_0 : \mu = 25$

Null hypothesis (H_A)

- Presence of some effect;
- Existence of something “new”;
- Interval value for the parameter.

Example: $H_A : \mu \neq 25$

Statistical Hypotheses: Working them out

Determination of the reference value for the null hypothesis H_0 :

- Previous knowledge about the process (investigation of changes);
- Value obtained from theory or models (model validation);
- Project requirements (investigation of system compliance);

Hypothesis testing involves:

- Obtaining the sample;
- Calculation of test statistics;
- Decision based on the computed value;

Statistical Hypothesis: an example



Suppose you are a large-scale consumer of green peas, and you want to determine if the 500g packages from a supplier really contain their nominal weight (at least on average).

In this case, the null hypothesis could be defined as: *the average net weight of a package is 500g*, and the alternative of interest could be expressed as the complementary inequality.

$$\begin{cases} H_0 : \mu = 500g \\ H_1 : \mu \neq 500g \end{cases}$$

Suppose still that you get 10 randomly selected packs from the supplier (in other words, $N = 10$), and you then measure their contents using a calibrated scale;

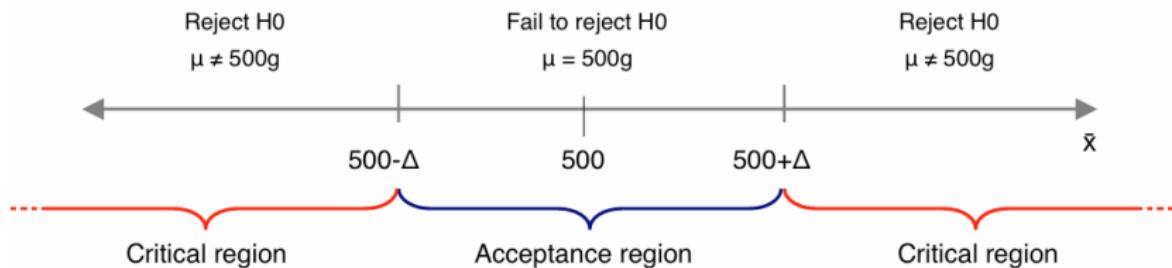
Statistical Hypothesis: an example

Since the sample mean \bar{x} is a good estimator of the real mean μ , we can assume:

- If $\bar{x} \cong 500g$ - then H_0 is corroborated;
- If $\bar{x} \ll 500g$ or $\bar{x} \gg 500g$ - H_0 is refuted;

This suggests that we use the value of \bar{x} as the basis for a statistical test.

First, we need to define a **critical region** for the rejection of H_0 .



The big question: how do we define Δ ?

Type I error

Type I error (AKA false positive): rejecting the null hypothesis when it is true.

The probability of occurrence of a false positive in any hypothesis testing procedure is generally known as the **significance level** of the test. It is generally represented by the Greek letter α :

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$$

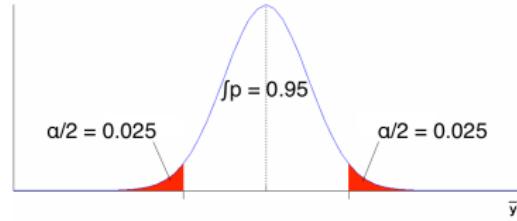
Another frequently used term is the **confidence level** of the test, given by $(1 - \alpha)$

Type I error

For a given sample, the selected value of α defines the critical threshold for the rejection of H_0 . Bigger α – lower threshold – higher probability of a type I error.

If H_0 is true (i.e., if $\mu = 500g$), the distribution of values of \bar{x} is approximately normal (remember the CLT), with average 500 and variance given by s^2/n .

For a Type-I error probability $\alpha = 0.05$, the critical values of the distribution of \bar{x} are those for which the probability density within the acceptance region is $1 - \alpha = 0.95$.



Type II error

Type II error (AKA. false negative): failure to reject the null hypothesis when it is false.

The probability of occurrence of a false negative in any hypothesis testing procedure is generally represented by β :

$$\beta = P(\text{type II error}) = P(\text{not reject } H_0 | H_0 \text{ is false})$$

The quantity $(1 - \beta)$ is known as the **power of the test**, and quantifies its sensitivity to effects that violate the null hypothesis.

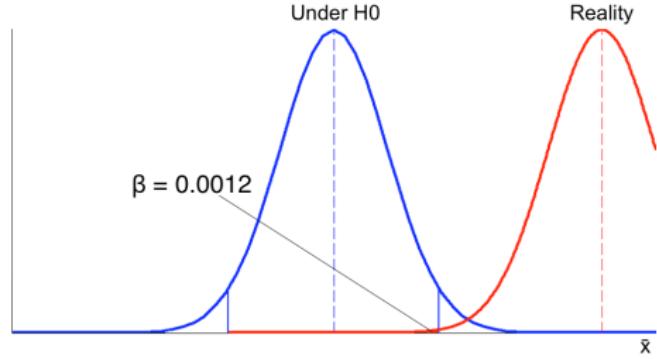
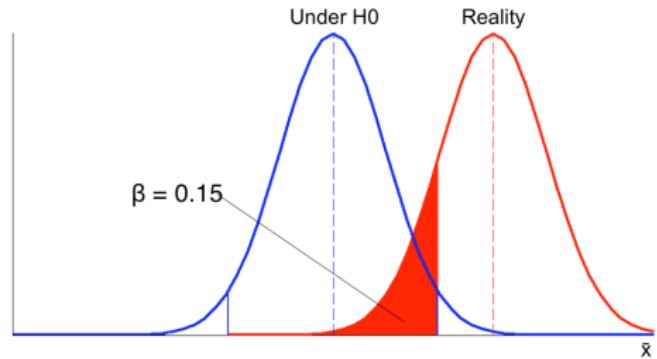
In other words

A type II error usually happens when the difference between the effect and the null hypothesis is too small to be detected. In this case, the testing procedure is not sensitive enough.

Type II error

Unlike the Type-I error, the definition of the Type-II error rate requires further specification of the value of the parameter being investigated;

The probability of failing to reject a false H_0 is strongly dependent on the magnitude of the difference between the value under H_0 and the real value of the parameter.



Type II error

The power of a test is governed by several factors:

- **Controllable factors**: significance level α , sample size N ;
- **Uncontrollable factors**: real value of the parameter in the population;

If H_0 is false, the smaller the magnitude of the difference between the real value of the parameter and the one under the null hypothesis, the greater the probability of a type II error.

However!

If the difference between H_0 and the parameter is too small, the practical importance of the effect (and the importance of the type II error) also gets insignificant.

Considerations about type I, II errors

Type I error (α) depends only on the distribution of the null hypothesis – easier to control;

Type II error (β) depends on the real value of the parameter – more difficult to specify and control;

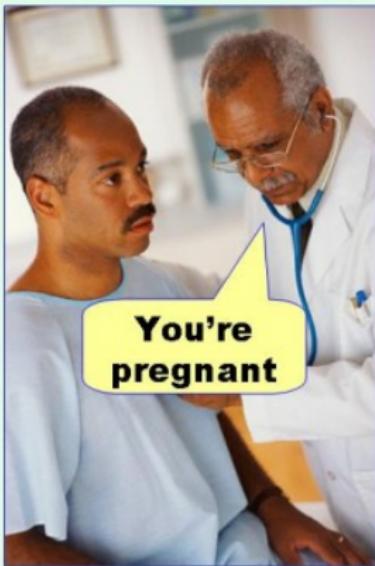
These characteristics lead to the following classification of the conclusions obtained from hypothesis testing:

- Rejection of H_0 – strong conclusion;
- Failure to reject H_0 – weak conclusion (but can be strengthened);

It is important to remember that failing to reject H_0 does not mean that there is evidence in favor of H_0 – it only suggests that it is a better model than the alternative.

Easy way to remember type I, II errors

Type I error
(false positive)



Type II error
(false negative)



Also – mini break!

General Procedure for Hypothesis testing

- Identify the parameter of interest;
- Define H_0 and H_A (one or two sided)
- Determine desired α, β ;
- Define minimally interesting effect δ^* ;
- Calculate sample size;
- Determine the test statistic and critical region;
- Compute the statistic;
- Decide whether or not to reject H_0 ;



Image (c) Roots Run Deep Winery: <http://www.rootsrundep.com/hypothesis.html>

Hypothesis testing: Mean of a normal dist, variance known



Back to the green peas example, we want to determine if there is any significant deviation on the mean weight of the packages. Assume for now that the variance of the process is known. The test hypothesis are defined as:

$$\begin{cases} H_0 : \mu = 500g \\ H_1 : \mu \neq 500g \end{cases}$$

Let the desired significance level be $\alpha = 0.05$;

Given these characteristics, we expect that the sampling distribution of \bar{X} is also normal, with variance $\text{Var}(\bar{X}) = \sigma^2/n$. If H_0 is true, the mean of the population is $\mu_{\bar{X}} = \mu_0 = 500$;

Hypothesis testing: Mean of a normal dist, variance known

Based on these characteristics, the variable

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

will be distributed according to the standard normal, $N(0,1)$, but **only if H_0 is true.**

This result implies a probability of $(1 - \alpha)$ that Z_0 will fall within the range $(\pm z_{\alpha/2})^b$ if H_0 is true. This provides a selection criterion between H_0 and H_1 :

- If $|z_0| > z_{\alpha/2}$, we reject H_0 at the confidence level $1 - \alpha$;
- Otherwise, there is not enough evidence to reject H_0

Hypothesis testing: Mean of a normal dist, variance known

Assume that we got $\bar{x} = 496.48$ from our $n = 10$ observations, and that $\sigma = 10g$. In this case.

$$z_0 = \frac{496.48 - 500}{10/\sqrt{10}} = -1.113$$

The critical values of the standard normal distribution are taken from its quantiles, so $\pm z_{\sigma/2} = \pm z_{0.025} = \pm 1.96$;
Since $|z_0| < Z_{0.025}$, we can conclude that there is not enough evidence to reject H_0 at the 95% confidence level.

Hypothesis testing: Mean of a normal dist, variance known

```
> install.packages("TeachingDemos")
> library(TeachingDemos)

> peas <- scan("Rfiles/greenpeas.txt")
> z.test (as.numeric(peas), mu=500, stdev=10)
```

```
One Sample z-test
data: as.numeric(peas)
z = -1.1131, n = 10.0000, Std. Dev. = 10.0000,
Std. Dev. of the sample mean = 3.1623,
p-value = 0.2657
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 490.282 502.678
sample estimates:
mean of as.numeric(peas)
                           496.48
```

Hypothesis testing 2: normal dist, variance unknown

Suppose now a more realistic situation: we don't know the real variance of the population. Besides, assume that we are interested in detecting only negative deviations (We only care if the real number of peas is less than the stated amount).

The test hypotheses can be defined as:

$$\begin{cases} H_0 : \mu = 500g \\ H_1 : \mu < 500g \end{cases}$$

In this second scenario, we want to be more conservative, so we pick a significance level of $\alpha = 0.01$;

It can be shown that, **if H_0 is true** then

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

Hypothesis testing 2: normal dist, variance unknown

From the same data as in the previous example, we can calculate that $\bar{x} = 496.48g$, $n = 10$, $s = 6.97g$. From this we find out that:

$$t_0 = \frac{496.48 - 500}{6.79/\sqrt{10}} = -1.5969$$

The critical value of this test statistic for the desired significance is $t_{\alpha,n-1} = t_{0.01,9} = -2.82$;

Given that $t_0 > -2.54$, we conclude that the evidence is insufficient to reject H_0 at the 99% confidence level;

In the end

From analyzing this sample of 10 peas packages, we do not have enough evidence to affirm, with a 99% confidence level, that the mean weight of pea packages is less than 500g.

Hypothesis Testing 2: calculating it with R

```
> library(TeachingDemos)
> peas <- scan("Rfiles/greenpeas.txt")
> t.test(peas, alternative="less", mu=500, conf.level=0.99)
```

One Sample t-test

```
data:  peas
t = -1.5969, df = 9, p-value = 0.07237
alternative hypothesis: true mean is less than 500
99 percent confidence interval:
      -Inf 502.6991
sample estimates:
mean of x
496.48
```

Hypothesis testing: description of results

(In)Sufficient evidence for rejecting H_0 at the significance level α

Even though this is correct (and better than just comparing single values), this description is still relatively poor:

- ① It does not provide information on the intensity of the evidence for rejection/non rejection;
- ② It imposes a predetermined significance level to the consumer of the information;
- ③ It does not provide information about the magnitude of the effect that was found, or the sensitivity of the test;

Hypothesis testing: p-values

The p-value:

The lowest significance level that would lead to the rejection of H_0 for the available data.

Can be interpreted as the probability that, if H_0 is true, that the test statistic would obtain a value at least as extreme as the one obtained.

For our previous example, the p-value could have been calculated as:

$$p = P(t_0 \leq -1.597 | H_0 = \text{TRUE}) = \int_{-\infty}^{-1.597} f(t_0) dt = 0.07237$$

It is still important to define the significance level a priori!

Beware the Caveats of hypothesis testing!

The p-value is a very powerful tool for analysing the result of a hypothesis test procedure. However, even the p-value can lead to erroneous conclusions if not analyzed carefully:

Some simple caveats

- Multiplicative effects of multiple experiments;
- Huge n sizes leading to high confidence for small effects;
- Type II errors
- Throwing away low confidence results

But you must be tired by now, so we will talk about these guys next class!

Today you learned

- Statistical population as a proxy for the unknown truth we want to understand through experiments;
- That there are different methods to infer information about the population from the experiments (samples);
- That different statistical methods have different accuracies and confidence levels, and these can be calculated;
- About hypothesis testing as a way to make decisions about a population based on a sample;
- About how to calculate and analyse a hypothesis about a parameter of the population;

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 "air quality" data set in the "datasets" R package

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
- Peas image: http://www.storko.eu/ed_files/image/green-peas.jpg

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>

Creative Commons BY-NC-SA 4.0.