Statistical Inference for Everybody and a Linguist

It wasn't me!

Textbooks in Language Sciences

language science press

Textbooks in Language Sciences

In this series:

1. Müller, Stefan. Grammatical theory: From transformational grammar to constraint-based approaches.
2. Schäfer, Roland. Einführung in die grammatische Beschreibung des Deutschen.
3. Freitas, Maria João & Ana Lúcia Santos (eds.). Aquisição de língua materna e não materna: Questões gerais e dados do português.
4. Roussarie, Laurent. Sémantique formelle: Introduction à la grammaire de Montague.
5. Kroeger, Paul. Analyzing meaning: An introduction to semantics and pragmatics.
6. Ferreira, Marcelo. Curso de semântica formal.
7. Stefanowitsch, Anatol. Corpus linguistics: A guide to the methodology.
8. Müller, Stefan. Chinese fonts for TBLS 8 not loaded! Please set the option tblseight in main.tex for final production.
9. Kahane, Sylvain & Kim Gerdes. Syntaxe théorique et formelle. Vol. 1: Modélisation, unités, structures.

ISSN: 2364-6209

Statistical Inference for Everybody and a Linguist

It wasn't me!

language
science
press

It wasn't me! 2025. *Statistical Inference for Everybody and a Linguist* (Textbooks in Language Sciences). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/

ISBN: (Digital)
(Hardcover)
(Softcover)
ISSN: 2364-6209
DOI:
Source code available from www.github.com/langsci/
Errata: paperhive.org/documents/remote?type=langsci&id=

Cover and concept of design: Ulrike Harbort
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
Grünberger Str. 16
10243 Berlin, Germany
http://langsci-press.org

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

Preface iii

Acknowledgments v

1 Inferences About Means 1
- 1.1 Differing means . . . 1
  - 1.1.1 Introducing the Logic . . . 1
  - 1.1.2 Doing the Maths . . . 5
- 1.2 Interpreting the Results . . . 9
  - 1.2.1 The Distribution of P Values . . . 9

Index 11
- Name index . . . 11

Preface

Acknowledgments

1 Inferences About Means

1.1 Differing means

Problem Statement

Let's assume you know the mean reaction time for a critical region when native speakers process a certain type of relative clause. This mean reaction time and the corresponding variance in measurements are extremely well established parameters. They were predicted by a robust theory of syntactic processing, and this prediction has been corroborated by a large number of diverse experiments. For an emergent subtype of this kind of relative clause, the theory predicts considerably higher precessing effort and thus longer reaction times. You conduct an experiment and measure reaction times in the critical region. Which outcomes of the experiment would you interpret as indidcating that reaction times are indeed longer for the emergent type of relative clause?

1.1.1 Introducing the Logic

The Problem Statement exemplifies a common question: Given a known mean value, do means under a specific condition diverge from this known mean? In this section, we show through simple frequentist reasoning how measurements from experiments can provide evidence to tackle such questions. The simplest test for such tasks is the **z-Test**. Notice that for the z-Test to be applicable, the given population mean (and the corresponding variance) must be truly known, which is why the Problem Statement stresses that the mean was predcited by a robust theory and that the prediction was tested in a long series of experiments. If these conditions are not met, other tests apply, and we're going to introduce such tests as we go along.

z-Test

For the sake of illustration, let's assume that the population mean is $\mu = 120$ (for example milliseconds) and the population variance is $\sigma^2 = 16$, which corre-

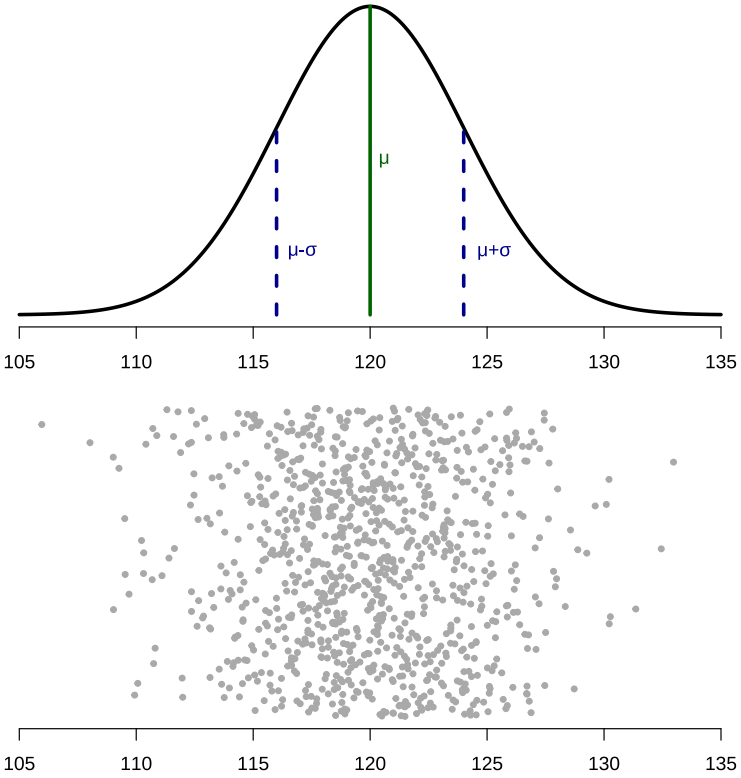


Figure 1.1: Theoretical population distribution for a normal distribution with μ=120 and σ=4 and a simulated random sample of n=1000 measurements from the population

sponds to a standard deviation of $\sigma = 4$. If the population values are generated according to a normal distribution, values are distributed according to the bell curve in Figure 1.1.

To recapitulate, this curve plots the probability density for the known population (for example, of reaction times). In the simplest terms, it shows for each measurement (x-axis) the probability with which it occurs (y-axis).[1] Informally speaking, the curve shows that if we measure random values from this population, the probability of measuring a value close to μ is highest, and measurements deviate on average by the standard deviation σ from μ. The dashed lines show

[1]Technically, the probability of each point measurement is 0, and non-zero probabilities are only defined as integrals of the density curve for intervals. This mathematical detail is mostly irrelevant for practical applications, but it should be kept in mind.

the standard deviation in each direction from the mean. As a result, a very much expected sample of $n = 1000$ measurements is shown in the form of the point cloud below the curve. The measurements are indeed centred around the mean, and they seem to follow the normal distribution.

If, however, we draw a sample from a different population where the true mean is higher (for example because we're measuring reaction times under a condition that is more difficult to process) we expect samples to turn out differently and have a higher sample mean compared to the known population mean. However, this expectation can be treacherous because individual samples are not in any way *guaranteed* to represent their population well, as we have shown in Chapter **??**. Very similar to Ronald A. Fisher in his experiment with Muriel Bristow (see Chapter **??**), we need to ask whether the actual sample warrants any inference regarding the underlying mechanism by being very much unexpected (albeit not impossible) under the assumption that the desired inference is not correct. In the case of the reaction times described in the Problem Statement, the desired inference is that reaction times are higher with the emergent subtype of relative clauses because of assumed processing penalties. However, especially if our sample is small, inferring anything from a specific result is tricky, as will be shown. Figure 1.2 shows a possible outcome with $n = 16$.

Let's call the sample plotted in Figure 1.2 x, a vector of 16 measurements x_1 through x_{16}. The mean of x is $\bar{x} = 122.1$. As we've shown, inferences in frequentist logic (see Chapter **??**) are always made by taking into account what the outcome of an experiment could have been under one or several possibly correct hypotheses. In the case at hand, we're interested in the hypothesis that the true mean under the experimental condition—call it μ_1—is larger than the known mean μ. In other words, we would like to gather evidence in support of H, where H: $\mu_1 > \mu$. For several reasons, we cannot gather evidence that supports this hypothesis directly. First, μ_1 is obviously not observable. It's a hypothesised mean in a population that exists as separate from the known population if H is correct. If that population is not substantially different from the known one, then we have $\mu_1 = \mu$. Given that μ_1 is a non-observable, all we've got are 16 data points from x and the sample mean $\bar{x}$ calculated from them. While we certainly hope that $\bar{x}$ is a good indicator of the true value μ_1, we have no guarantees whatsoever that it actually is. Second, as we're still Fisherians on this page of this book, we have no formal method of gathering positive evidence. Only Neyman-Pearson philosophy and the Severity approach will give us this power (pun intended) later in Chapter **??**, but with some important caveats. Therefore, we can only check whether the data x are in accord with a **Null** Hypothesis (*Null* for short) **Null**

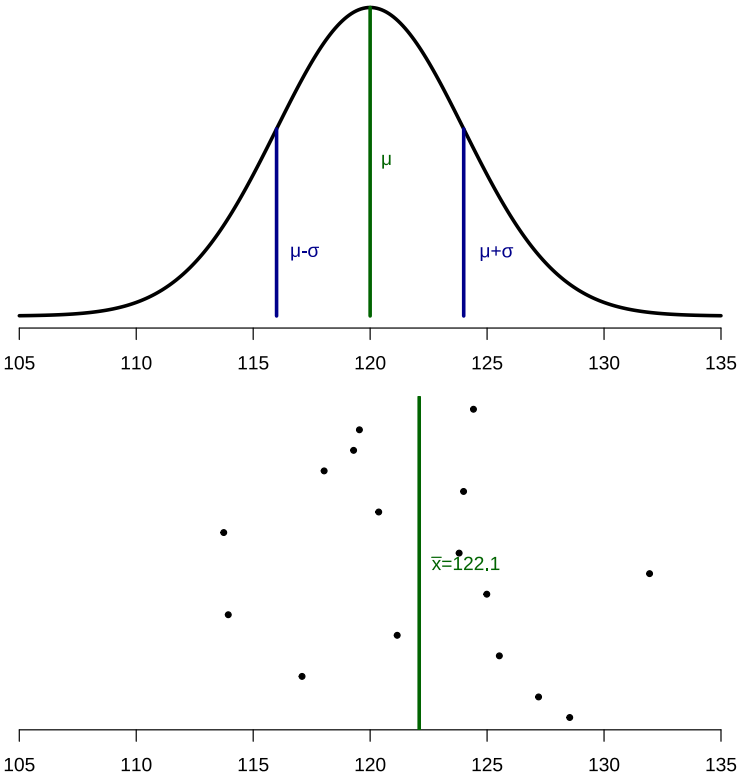


Figure 1.2: Theoretical population distribution for a normal distribution with μ=120 and σ=4 and a simulated random sample of n=16 measurements from some population

N: $\mu_1 \not> \mu$.[2]

To test this hypothesis, the Fisherian framework assigns a certain well-defined kind of probability to (obtaining) the data x given that N is correct, formally $p(x|N)$ or *the probability of x given N*. This probability can be used to assess whether the data x are in accord with the Null N. Before moving on to the calculations, it is vital to consider the question of which inferences are warranted in case x is or isn't in accord with N. First, what do we infer from the data x (in other words, from our experiment) if they are compatible with N? The answer

[2]The Null Hypothesis N is sometimes designated as H_0, in which case H is often called H_1. This nomenclature is—in our view—where the confusion between Fisher and Neyman-Pearson begins. Furtermore, as the Fisherian Null Hypothesis is not a proper hypothesis anyway but rather a non-hypothesis, we call it Null or N.

is: absolutely nothing! If the data are in accord with N, we haven't found any evidence that it is not the case that μ_1 is not greater than μ, and that's the end of it. If that sounds uselessly messed up and disappointing, that's because it is.[3] If you infer anything from such a result, you're not only wrong, but also Baeysians with Dutch names will come to haunt you (or at least unfollow you on the messaging platform of your choice)—and rightly so. Second, what do we infer from the data x if they are not compatible with N? This is the much more interesting case, but it's difficult to define the admissible inerences without creating false ideas if it's done without the maths and without a deeper look at the way H and N were set up. Let's say rather informally that in such a case we've found some evidence in support of H because N and H partition the range of possible values of μ_1: either it's greater than μ (H) or it is not greater than μ (N). Finding no accordance with N despite serious attempts to do so (see Chapter **??**) provides at least some indication that H might be correct. If you're looking for proof of anything, we recommend that you stick to pure theory, logic, theoretical maths, or pseudoscience. There is no proof to be found in (non-trivial) experiments, and statistical inferences are weak and fragile.

1.1.2 Doing the Maths

We have argued that in Fisherian inference, we have to asses whether x is compatible or in accord with N. But how do we do this? The most naive but not at all wrong thing to do is calculate the difference between the known population mean μ and the mean of the obtained sample $\bar{x}$. In our case, this is $\bar{x} - \mu = 2.1$. Clearly, a minimal requirement for any further calculations is that this difference is positive. If it were negative, the sample could hardly be interpreted as evidence against N: $\mu_1 \not\succ \mu$.[4]

However, solid empirical inference requires us to evaluate how significant this difference actually is. This evaluation follows the same logic as in our introduction to Fisher's philosophy (Chapter **??**). In the analysis of the Tea Tasting experiment, we asked how often someone who merely guesses would classify one, two, three, or four cups correctly by chance. In the case at hand, simply counting events is not informative as the events are occurrences of specific values, namely individual reaction times. Hence, the question becomes: how often would we expect to see a sample of 16 measurements with a sample mean of 122.1 or larger

[3] Whether you're a linguist or not, please consider that *finding no evidence that A is true* is not the same as *finding evidence that A is false.*

[4] This way of putting it is slightly sloppy and informal. We will return to this notion and make it more precise. However, in practice it is blatantly obvious that we would never take an experiment that showed lower reaction times as evidence for higher reaction times, etc.

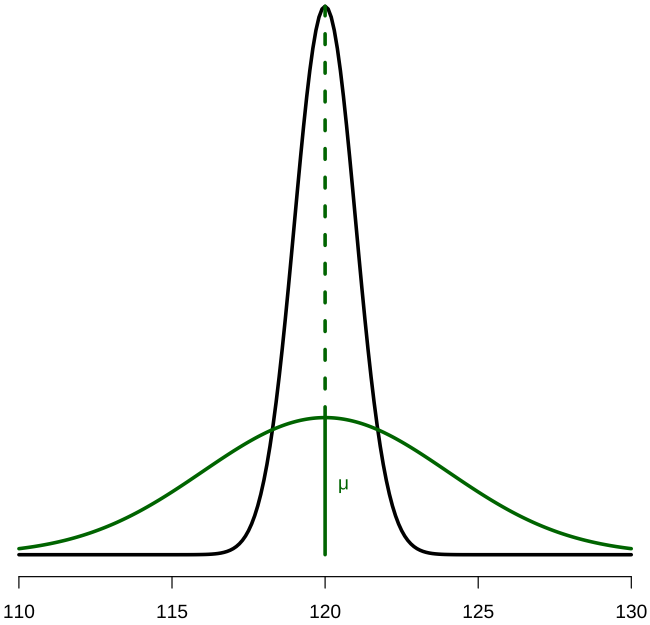


Figure 1.3: Theoretical population distribution for a normal distribution with μ=120 and σ=4 (green) the distribution of sample means for samples from this distribution with n=16 (black)

if the true mean is that specified by the Null, which is $\mu = 120$? Luckily, we have already introduced the tool that we need: the **standard error** of the mean. The standard error for $n = 16$ and the known variance $s = 4$ (see p. 1) tells us how much samples of this size differ from the true mean on average. The standard error of the mean is:

standard error

$$S(n, \sigma) = \frac{4}{\sqrt{16}} = 1$$

Remember what the standard error is all about (Chapter **??**). If the mean in a population is μ and the standard deviation is σ, then the sample means $\bar{x}_i$ of repeated samples of size n are themselves normally distributed with the standard error S being the standard deviation of that normal distribution. Furthermore,

keep in mind that we're talking about the distribution of the known population. Under the Null, it is also the distribution from which our small sample was drawn. Figure 1.3 contrasts the density of the distribution of individual data points (in our example: individual reaction times) with the much narrower distribution of sample means. Mathematically, it is narrower because the standard error is always smaller or equal to the standard deviation. Intuitively, it should be narrower because on average sample means from samples with $n > 1$ approximate the true mean better than single measurements.[5]

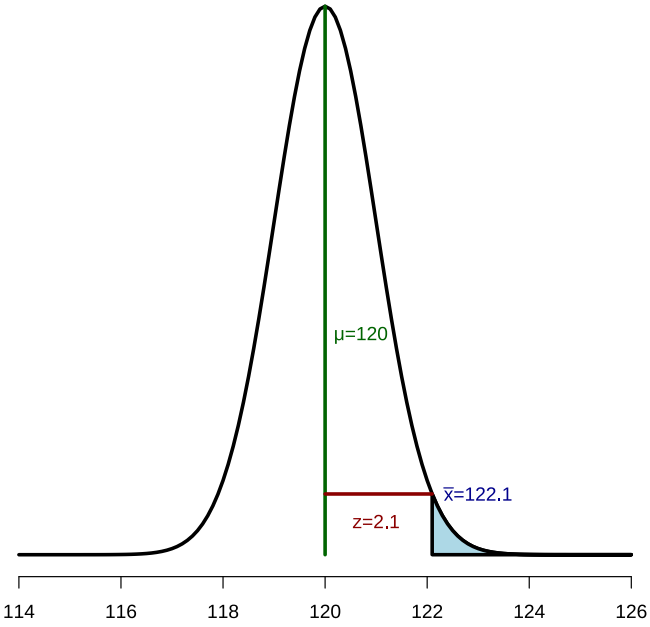
Since (i) the distribution of sample means is normal, (ii) we know its mean, (iii) we know its variance/standard deviation, we can calculate how many samples of infinitely many samples (or at least a lot of samples) drawn from the known population have a mean of 122.1 or larger. In other words, we can calculate how many sample means would deviate by 2.1 from the population mean anyway due to expected sampling error if we took a lot of samples of size $n = 16$. This is exactly parallel to the argument regarding the Tea Tasting experiment where we calculated how often we could expect certain outcomes anyway, even if the person performing the Tea Tasting task had no ability to detect which liquid was poured into the cup first. It gives us a very precise and well-defined measure of how surprising the obtained result would be were the Null true.

There are many ways of calculating the number of interest. Since the area under the normal curve sums up to 1 (as should be the case with probability density functions), we could integrate it over the interval $[122.1, \infty]$. Alternatively, we could use the so-called cumulative density function, to which we will return later. To make things much simpler in practice, the most widely used way in applied statistics is based on counting the distance between the distribution mean and the sample mean as multiples of the standard error, which gives us the **z-score**: **z-score**

$$z = \frac{\bar{x}-\mu}{S(n,\sigma)} = \frac{122.1-120}{1} = \frac{2.1}{1} = 2.1$$

Figure **??** shows the distribution of sample means for the known population, the true mean μ, the obtained sample value $\bar{x}$, and the area under the normal curve which defines how many samples of size $n = 16$ (in the limit) have means of $\bar{x}$ or greater. In addition, the red line measures the distance from μ to $\bar{x}$, which corresponds to the z-score. By dividing the the distance between the sample mean and the population mean with the standard error, the z-score normalises said

[5] Understanding this argument is crucial. If you're not following it, you should go back to Chapter **??** for an introduction to the distribution of sample means, especially the argument concerning samples of size $n = 1, 2, \ldots$.



distance, which itself varies with the standard deviation in the population and subsequently with the standard error. Hence, regardless of the slope of the concrete distribution, $z = 2.1$ tells us how (im)probable the sample mean (or a more extreme sample mean) is under the Null. It gives us another infamous **p-value**. In the olden days (and in this book), tables were used to look up p-values corresponding to z-scores, and modern statistics software has functions to achieve the same. We will discuss those tables later in this chapter, but for the time being, let's take it for granted that

p-value

$$Pr(\bar{x}|N) = p_{\mathbf{Norm}}(2.1) = 0.02$$

Let's go through this step by step. First, $Pr(\bar{x}|N)$ reads *the probability of the sample mean $\bar{x}$ given the Null N*. Remember that the Null was specified as N: $\mu_1 \not\succ \mu$, which reas *the mean under the condition of interest* (the mean reaction time in the emergent type of relative clause) *is not greater than the population mean*

(the reaction time in other relative clauses). Second, this probability is equated to $p_{\text{Norm}}(2.1)$, which is the p-value corresponding to the z-value of 2.1 as calculated above, which is 0.02.

1.2 Interpreting the Results

At this point, a little exercise is in order. From the following statements, choose the one which is a correct interpretation of the calculations above. There is no pressure. Nobody can read your mind, nobody even cares whether you pick a wrong statement, and you can only gain by actually thinking about each statement thoroughly. Do not decide on one statement because it is intuitively correct, but because you know why it is correct.

1. The probability that hypothesis H (reaction times are longer in the emergent type of relative clause) is true is 0.02.
2. The probability that hypothesis H (reaction times are longer in the emergent type of relative clause) is true is 0.98.
3. The results prove that the emergent type of relative clause incurs longer reaction times than other relative clauses.
4. The p-value is very small, which indicates that mean reaction times in the emergent type of relative clause are substantially longer than in other relative clauses.
5. The probability of obtaining another sample with a mean of 122.1 or greater is 0.02.
6. Based on this outcome, we can reject the possibility that reaction times under the condition of interest are actually *smaller* than in the population with a certainty of $1 - 0.02 = 0.98$ (or 98%).
7. The experiment has shown that reaction times are normally distributed.
8. The experiment provides evidence in favour of the underlying theory of linguistic processing.

1.2.1 The Distribution of P Values

Statistical Inference for Everybody and a Linguist

This book is good.