Statistiek I

Sampling

John Nerbonne

CLCG, Rijks*universiteit* Groningen

`http://www.let.rug.nl/nerbonne/teach/Statistiek-I/`

Overview

Humanities Statistics—Hypotheses

Lots of humanities issues are EMPIRICAL and VARIABLE

empirical — involving matters of fact, not purely conceptual

variable — issues that may be decided in different ways for different individual cases

We regard these as **hypotheses** to be tested.

Examples of empirical, variable **hypotheses**:

- sex is related to verbal fluency
- web sites with banners get more attention
- grammatical structure influences language processing

Statistical analysis needed for EMPIRICAL, VARIABLE hypotheses.

Hypothesis Testing

We begin with a research question, which we try to formulate as a **hypothesis**

- sex is related to verbal fluency
- web sites with banners get more attention
- grammatical structure influences language processing

Normally, we need to translate this to a concrete form before statistics are useful

- men and women score differently on tests of verbal fluency
- web sites with banners are revisited more often
- object relative clauses (i.e., those in which relative pronouns are grammatical objects) take longer to read than subject relative clauses

Abstraction

Given a research question, translated into a concretely testable hypothesis

- web sites with banners are revisited more often

= "**all** web sites with banners are revisited more often than web sites without banners"?
—probably not. The data is variable, there are other factors:

- amount of information (library system)
- value of information (Centraal Bureau voor Statistiek)
- changeability of data (weather, flight arrivals)

We normally need statistics to abstract away from the variability of the observations.

- web sites with banners are revisited more often **on average**

Subject Matter

- web sites with banners are revisited more often **on average**

We **must** study this on the basis of a limited number of web sites — a SAMPLE. But we're interested in the larger class of all web sites — the POPULATION.
The hypothesis concerns the population, which is studied through a representative sample.

- **men and women** differ in verbal fluency (study based on 30 men and 30 women)
- **web sites** with banners are revisited to more often (studied on the basis of 30 web sites)
- **object relative clauses** take longer to read than **subject relative clauses** (studied on the basis of 30 people's reading of 20 relative clauses of each type).

Analysis

Given a research question, translated into a concretely testable hypothesis, expressed abstractly

- web sites with banners are revisited more often **on average**

You measure rates of revisiting for a randomly selected group of sites, with and without banners.
Will any difference in averages (in the right direction) be proof?
—probably not. Very small differences might be due to **chance**.
We normally need statistics to analyse results.

- STATISTICALLY SIGNIFICANT results are those unlikely to be due to chance.

Samples and Populations

Selecting a sample from a population includes an element of chance—which individuals are studied?
Fortunately, we know a lot about the likely relation between samples and populations — the **Central Limit Theorem**
Central Limit Theorem relates sample means to likely population mean.
To understand it, imagine all the possible samples one might use, and all those sample means—the **distribution of the sample means**.

Central Limit Theorem

Background: Population standard deviation must be known (e.g., as for standardized tests—IQ, CITO, ...)

- Sample means ($\bar{x}$) are **always** be normally distributed.
- Mean of samples means is population mean.

$$m_{\bar{x}} \; = \; \mu$$

- Standard deviation (sd) among samples is systematically smaller than σ (population sd) among individuals.

$$SE = s_{\bar{x}} \; = \; \frac{\sigma}{\sqrt{n}}, \text{ where } n \text{ is sample size}$$

Central Limit Theorem: Sample mean has dist. $N(\mu, \sigma/\sqrt{n})$.

—note importance of sample size

z-Tests

Given a RANDOMLY SELECTED SAMPLE, we know

distribution it is one of a normally distributed population of samples

mean $m_{\bar{x}} = \mu$ —the mean of such samples will be the population mean

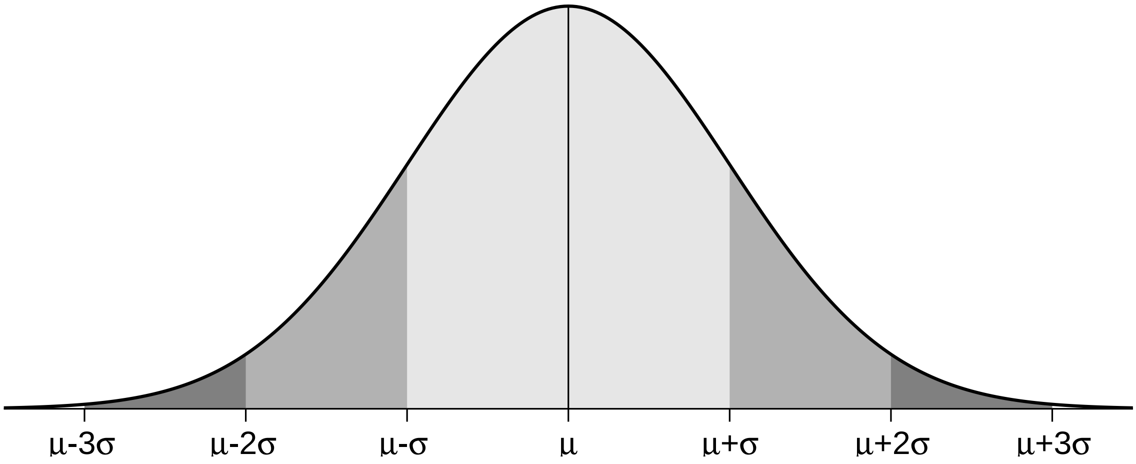
standard deviation $\mathrm{sd}_{\bar{x}} = \sigma/\sqrt{n}$ —the standard deviation of the sample means (the STANDARD ERROR) will be less population's standard deviation by a factor of $1/\sqrt{n}$

These facts allow us to reason about the population.

The reasoning will always include a *probability* that population has a mean of a given size.

An essential assumption is that the sample is randomly selected. We can't correct for biased data—even unintentionally biased.

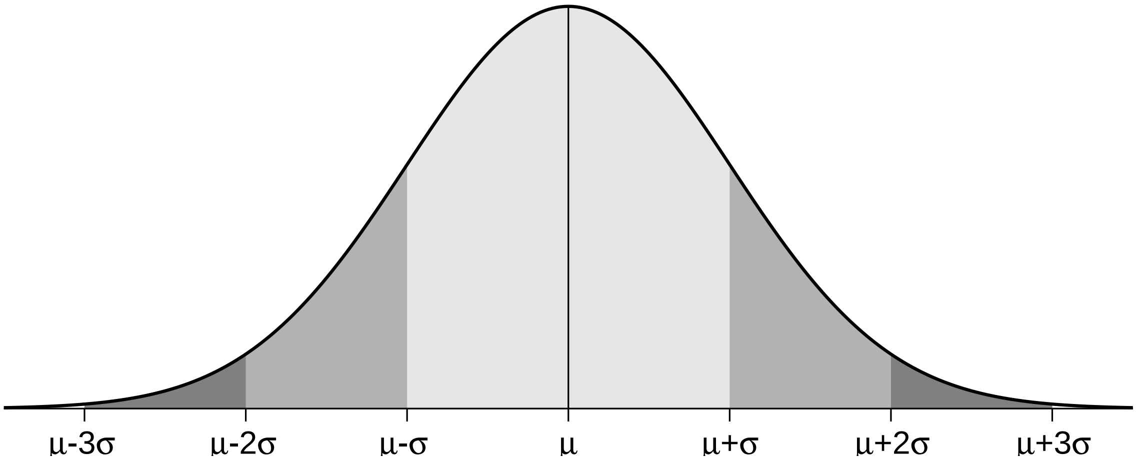
Normal Distribution (Review)



We consider an element *x* within a normal distribution, esp. the probability of *x* having a value near the mean.

$$P(\mu - \sigma \leq x \leq \mu + \sigma) = 68\%$$
$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 95\%$$
$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = 99.7\%$$

Normal Distribution (Review)



If we convert *x* to a “standard *z* score” ($z = x - \mu/\sigma$), where $\mu = 1$ and $\sigma = 1$:

$$P(-1 \leq z \leq 1) = 68\%$$
$$P(-2 \leq z \leq 2) = 95\%$$
$$P(-3 \leq z \leq 3) = 99.7\%$$

Example Application of z-test

You suspect that CALL programs may be effective for young children (since they can be initiated before reading, and look like computer games, need little supervision, ...).
You have a standard test for English proficiency, where $\mu = 70, \sigma = 14$. You apply the same test to 49 randomly chosen schoolchildren who've had a CALL program at home for three years. Result: $\bar{x} = 74$
Since this is a sample, we compute a standard error SE $= \sigma/\sqrt{n} = 14/\sqrt{49} = 2$. We see that this sample is two standard deviations above the established population mean!

Since this is a sample mean, it is normally distributed, so that we can conclude that this sample is at the 97.5%-ile of all such samples.
There is only a 2.5% probability that the sample mean would be this high by chance.

CALL Conclusions

You apply a test to kids who've used a CALL program, the result is a z-score of 2, and the chance of this is 2.5%. It's very unlikely that this arose by chance (it would happen once every forty times).
Conclusion: the CALL programs are probably helping.
Notate bene: it is possible that the programs are not helping at all, and that the sample happened to include lots of proficient kids. ...There might be many confounding factors.
(Try to think of some.)

Importance of Sample Size

Suppose you had applied the test to only 9 kids who've used a CALL program with the same result of 74, where the test (as above) has scores with $\mu = 70,\ \sigma = 14$.
Then standard error would be greater: SE $= \sigma/\sqrt{n} = 14/\sqrt{9} = 4.7$. In this case the sample ($\bar{x} = 74$) would be less than 1 SE above the population mean ($\mu = 70$), i.e., at less than the 68$^{\text{th}}$ percentile — not very surprising. Samples means this high are found 32% of the time by chance.
Then we'd have no reason to suspect any special effect of CALL programs.
This could be a CHANCE EFFECT

Analysing the Reasoning

Statisticians have analyzed this reasoning in the following way.
We always imagine two hypotheses about the data, a NULL HYPOTHESIS, H_0, and an alternative, H_a. In the CALL example:

$H_0 : \mu_{\text{CALL}} = 70$
$H_a : \mu_{\text{CALL}} > 70$

H_a looks right, since $74 > 70$. But this is insufficient evidence, since some differences could be due to chance.
We formulate a null hypothesis in order to measure the likelihood of the data we collect.

Logically, we'd prefer to formulate $H_0 : \mu_{\text{CALL}} \leq 70$, exactly the negation of H_a. But we usually see '$=$' in formulations.

The Reasoning

$H_0 : \mu_{\text{CALL}} = 70$
$H_a : \mu_{\text{CALL}} > 70$

We reason as follows: if H_0 is right, what is the chance p of a random sample with $\bar{x} = 74$? We obtain a p-value by converting the score to a z-score, and checking its probability in a table.

$$z_x = (x - \mu)/\sigma$$
$$z_{74} = (74 - 70)/2 = 2$$

A check in the tables for the standard normal distribution show $P(z \geq 2) = 0.025$, and so the chance of the sample is just $P(\bar{x} = 74) = 0.025$. This is the p-VALUE, aka MEASURED SIGNIFICANCE LEVEL, or *overschrijdingskans*.

If H_0 is correct, and kids with CALL experience have the same language proficiency as others, then the observed sample would be expected only 2.5% of the time. As always, small values of p are strong evidence against the null hypothesis.

Statistically Significant?

We have H_0, H_a and a way to calculate the chances of samples assuming H_0. In the CALL example, we know that 49-element samples have a dist. $N(70, 14/\sqrt{49})$

$H_0 : \mu_{\text{CALL}} = 70$
$H_a : \mu_{\text{CALL}} > 70$

The classical test specifies a level of likelihood that must be attained for a test to count as significant, the threshold SIGNIFICANCE LEVEL, or α-LEVEL. This is a level which the p-value is compared against. Most common are $\alpha = 0.05$ and $\alpha = 0.01$, but stricter levels may be required if important decisions depend on results.

The p-value is the chance of encountering the sample, assuming that the H_0 is right. The α-level is the threshold beyond which we regard the result as significant.

Is the *p*-value below α?

$H_0 : \mu_{\text{CALL}} = 70$
$H_a : \mu_{\text{CALL}} > 70$

Given the sample of 49 with mean $m = 74$ in the dist. $N(70, 14/\sqrt{49})$, we calculate $p = 0.025$. This is below 0.05, but not below 0.01.

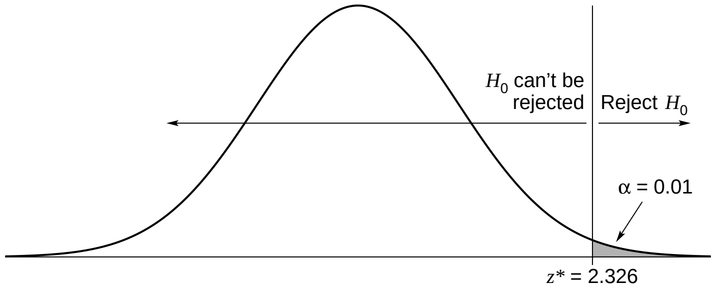
So the result was SIGNIFICANT AT THE $\alpha = 0.05$ LEVEL, but not at the 0.01-level.

Reminder: work out the pencil-&-paper exercise on sampling statistics!

Summary of Significance Tests

Step 1 Formulate H_0, H_a—your research question.
Test statistic (e.g., sample mean) is specified as is underlying dist. (assuming H_0).

Step 2 Specify the α-level—the level at which H_0 will be rejected.



the α-level of 0.01 for a test based on the normal distribution.

Step 3 Calculate the statistic which the test uses (e.g., mean).

Step 4 Calculate the *p*-value, and compare it to the α-level.

Summary of Significance Tests

- Step 1 Formulate H_0, H_a—your research question.
- Step 2 Specify the α-level—the level at which H_0 will be rejected.

Some books recommend that Step 2 include a computation of the “critical values” of the test statistic—the values which will lead to rejection of H_0. At $\alpha = 0.05$, the critical region is $z|P(z) \leq 0.05$, i.e. $z \geq 1.65$. We can translate this back to raw scores by using the z formula.

$$\begin{aligned}
&z_x = (x - \mu)/\sigma \\
&1.65 = (x - 70)/2 \\
&3.3 = x - 70 \\
&x = 73.3
\end{aligned}$$

Implicitly done by statistical software, so we omit it hence.

One-sided z-tests

Formally, our CALL example is a z-test because it is based on a normal distribution whose mean μ and sd σ are known.
In every case we calculate the mean of a random sample m, and a z-value based on it, where z is, as usual, $z = (m - \mu)/(\sigma/\sqrt{n})$
It can take many forms, depending on which values of z are predicted in the H_a.

- H_a predicts high m CALL programs *improve* foreign language ability of children. $p = P(Z \geq z)$
- H_a predicts low m Brocoli eaters have *low* chlosterol levels. $p = P(Z \leq z)$

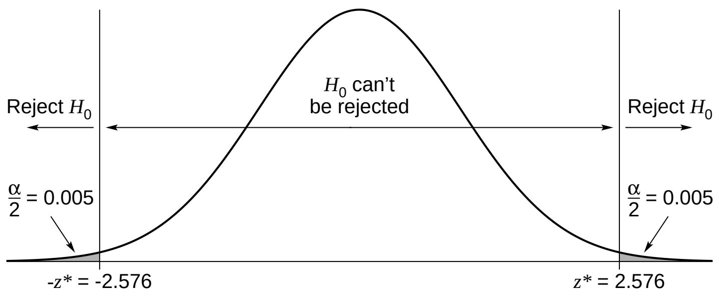
These are called ONE-SIDED tests because H_0 will be rejected on the basis of p values on one side of the distribution.
But sometimes H_a doesn’t predict high or low, just *different*.

Two-sided z-tests

Sometimes H_a doesn't predict high or low, just *different*.
Example You wish to use a children's test for aphasia developed in the UK (after translation). The test developers claim that scores are distributed $N(100, 10)$ on nonaphasic children. To validate its use after translation, you could test it on 25 normal Dutch children.
In this case H_0 predicts that $\mu_O = \mu_T$ (translation has same mean as original), and $H_a : \mu_O \neq \mu_T$, but without specifying whether μ_T is higher or lower than μ_O.
Suppose we again require a significance level of $\alpha = 0.01$.
In this case, both *extremely high* and *extremely low* sample means give reason to reject H_0.

Two-sided z-tests

H_a predicts extreme m at $\alpha = 0.01$, we need a $\bar{x}$ in the most extreme 1% of the distribution in order to reject H_0, i.e. in the highest 0.5% or the lowest 0.5%.



the most extreme 1% of the normal distribution is divided into the lowest 0.5% and the highest 0.5%. p-values must reflect the probability of $Z \geq |z|$, either $Z \geq z$ or $Z \leq -z$. z-values in both tails of the distribution give grounds for rejection of H_0.

Understanding Significance

Recall the language teaching example. Hypotheses were:

$$\begin{array}{l} H_0 : \mu_{\text{CALL}} = 70 \\ H_a : \mu_{\text{CALL}} > 70 \end{array}$$

Given a 49-element samples, we have dist. $N(70, 14/\sqrt{49})$ The sample mean of $m = 74$ has a measured significance level of $p = 0.025$. This is significant at the $\alpha = 0.05$, but not at the level of $\alpha = 0.01$.
If you're sure of $m = 74$, and if you wanted significance at $\alpha = 0.01$, you *could* ask how large the sample would need to be.

Chasing Significance

If you're sure of $m = 74$, and if you wanted significance at $\alpha = 0.01$, you *could* ask how large the sample would need to be.
$\alpha = 0.01$ corresponds to $z = 2.33$ (tables), so we can derive:

$$\begin{array}{rcl} z & = & (\bar{x} - \mu)/(\sigma/\sqrt{n}) \\ 2.33 & = & (74 - 70)/(14/\sqrt{n}) \\ & = & 4\sqrt{n}/14 \\ \sqrt{n} & = & (2.33 \times 14)/4 \\ n & \approx & 67 \end{array}$$

A sample size of 67 would show significance at the $\alpha = 0.01$ level assuming the sample mean stayed at $\bar{x} = 74$.
Would it be sensible to collect the extra data?

Understanding Significance

Is it sensible to collect the extra data to “push” a result to significance?
No. At least, usually not.
The real result is the extent of the difference (4 pt.) This does not change in the hypothetical example. You have to know whether this difference in ability has consequences (e.g., in the policies of the school you work in, or in buying software for your kids, or whatever).
“Statistically significant” implies that an effect probably is not due to chance, but the effect can be very small.

This is a two-edged sword: just because an effect was not demonstrated to be statistically significant doesn’t mean that nothing important is going on. It means you’re not sure.

Misuse of Significance

Garbage in, garbage out If the experiment poorly designed, or the data is poorly collected, no amount of statistical sophistication can repair the situation.

No “significance hunting” Hunting among dozens of variables is *likely* to turn up some extreme results. Multiple tests need to be analyzed especially if statistical significance is to be claimed.
Looking at many variables *can* be useful in early stages of investigation—before hypothesis testing.

Power of Statistical Tests Some tests are more sensitive than others, and this makes them more useful. Relatively insensitive tests may show no significance even when an effect is genuine.
More formally, the discriminatory power of a test is likelihood that H_0 will be rejected when H_a is true.

Confidence Interval

An alternative view of statistical significance.

Example: you want to know how many hours per week a student in information science works (outside of study, to earn money). You know the standard deviation for the university is approx. 1 hr./week

- $\sigma = 1\text{hr.}/\text{wk}$
- collect info from 100 people
- calculcate $m = 5\text{hr.}/\text{wk}$
- therefore $\mu = 5\text{hr.}$, SE is $1\text{hr.}/\sqrt{100} = 0.1\text{hr.}$

Sample is **randomly chosen**, thus subject to random error. It is one of many samples (whose theoretical distribution you know).

How certain are you of this estimate?

Confidence Interval

- $\sigma = 1\text{hr.}/\text{wk}$
- collect info from 100 people
- calculcate $m = 5\text{hr.}/\text{wk}$
- therefore estimate $\mu \approx 5\text{hr.}$, SE is $1\text{hr.}/\sqrt{100} = 0.1\text{hr.}$

Since it is part of a normal distribution, we can apply the usual reasoning to obtain an ERROR MARGIN. For example:

68% of all elements of this distribution will fall in the interval
$m \pm 1\text{sd} = m \pm 0.1$.

95% of all elements of this distribution will fall in the interval
$m \pm 2\text{sd} = m \pm 0.2$.

We are 95% confident that μ is in the interval
$5\text{hr.}/\text{wk.} \pm 0.2\text{hr.}/\text{wk.} \;=\; (4.8\text{hr.}/\text{wk.}, 5.2\text{hr.}/\text{wk.})$ where 5hr./wk. is the estimate, & 0.2hr./wk. the error margin

Confidence Interval

Example: how much do students work (per week)?

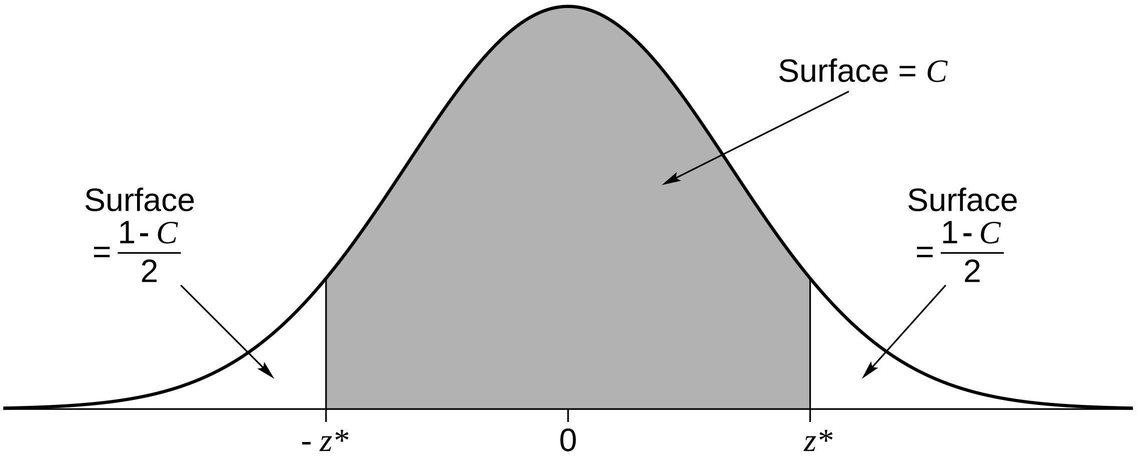
- $n = 100, \sigma = 1/\text{wk}$
- calculcate $m = 5hr, /\text{wk}$
- therefore $\mu \approx 5\text{hr.}/\text{wk.}$, SE is $1\text{hr.}/\text{wk.}/\sqrt{100} = 0.1\text{hr.}/\text{wk.}$

We can specify many confidence intervals.

68%conf. interval	$m \pm 1\sigma$	$4.9 \leq m \leq 5.1$	$(4.9, 5.1)$
95%conf. interval	$m \pm 2\sigma$	$4.8 \leq m \leq 5.2$	$(4.8, 5.2)$
99.7%conf. interval	$m \pm 3\sigma$	$4.7 \leq m \leq 5.3$	$(4.7, 5.3)$

Note that larger (less exact) intervals can *always* be specified at higher confidence levels. We trade confidence for precision.

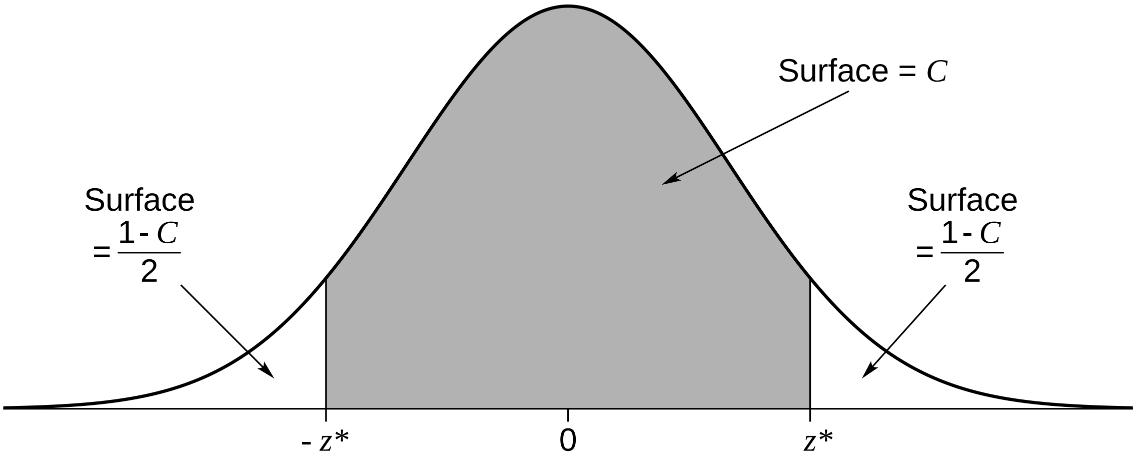
Confidence Interval



Summary

- With confidence C we identify an interval within which a mean μ is expected to fall
- Exercise 2 (single sample t-tests) involves a confidence interval where the standard deviation is effectively estimated.

Significance Tests vs. Confidence Interval



Hypothesis tests typically identify a 95% CI within which sample results should fall if H_0 is correct.
To confirm a two-sided hypothesis at level α, a sample statistic outside the central $1 - \alpha$ is needed—i.e., outside the $1 - \alpha$ CONFIDENCE INTERVAL

Hypothesis Testing

A **statistical hypothesis** concerns a population about which a hypothesis is made involving some statistic

- population (all web sites)
- parameter (statistic) (rate of revisiting)
- hypothesis (ave. rate of revisiting higher when banners used)

- **always** about populations, not just about samples
- sampling statistic identified
  - mean
  - frequencies
  - ...

Identifying Hypotheses

ALTERNATIVE HYPOTHESIS ($=$ original hypothesis) is contrasted with NULL HYPOTHESIS — hypothesis that nothing out of the ordinary is involved.

- H_a: (ave. rate of revisiting is higher when banners used)

contrasts with NULL HYPOTHESIS:

- H_0 (null hypothesis): (banners make no difference in ave. rate of revisiting)

Logically, H_0 should imply $\neg H_a$

Quantifying Significance

- STATISTICALLY SIGNIFICANT results are those unlikely to be due to chance.

We quantify significance by estimating how likely it is that results could be due to chance.
Concretely: if the null hypothesis were true, how likely would the sample statistic be?
Example: If in fact banners make **no** difference in how often web sites are revisited, how likely is it that a sample of 20 web sites with and without banners would show that 18% of the visitors return to the former and only 13% to the latter?
p-VALUE is the chance of sample given H_0
A low p-value is evidence against H_0, and for H_a

“Significant at the 0.05 level”

We normally determine in advance which significance level is required for (probabilistic) proof.
For example, we may agree that any result with a p-value less that 0.05 is sufficient proof against the H_0 (and therefore for the H_a) that we will be convinced.
The p-value that is determined to be sufficient for the rejection of H_0 is referred to as the α-LEVEL
We may then report the results of the experiment as “significant at the $p \leq 0.05$-level” or “significant at the 0.05-level”.

Other Significant Levels

Sometimes, α is determined to be 0.01, sometimes 0.001
α is threshold of REGION OF REJECTION — score needed to reject H_0 (and accept H_a)
—low values unlikely if H_0 is true, likely if H_a true
Size of region (value of α) inversely proportional to acceptable risk (of wrongly accepting H_a)

Other Significant Levels — Example

Example: You have aphasia test, with known μ (mean), σ (standard deviation) from US, & may wish to use it in the Netherlands
H_0: $\mu_{\text{US}} = \mu_{\text{NL}}$ (same population, therefore same μ)
H_a: $\mu_{\text{US}} \neq \mu_{\text{NL}}$ (different populations, maybe due to language dependencies)
region of rejection: 0.05
—you reject H_0 even though results would be consistent 5% of time
Region of rejection variable

- perhaps new test very expensive
- perhaps this aspect of diagnosis not essential

Interpreting Results

1. Take a sample (of n aphasic patients), administer test, determine μ_{NL}.
2. Determine z score for sample of n.

$$z = \frac{\mu_{\text{NL}} - \mu_{\text{US}}}{\sigma/\sqrt{n}}$$

3. Use tables to determine chance of z score, $P(z)$. This is the p-value, the chance of the sample if $\mu_{\text{NL}} = \mu_{\text{US}} (= H_0)$
4. If sample statistic is in rejection region, e.g., $p < 0.01$, reject H_0 in favor of H_a (statistically significant)
5. If sample statistic *not* in rejection region, then either accept H_0 or suspend judgement

Possible Errors

You could, of course, be wrong.
The selection of the sample could be unlucky (unrepresentative). Possibilities:

H_0	true	false
accepted	correct	type II error
rejected	type I error	correct

Type I Errors — focus of hypothesis testing
p-value – chance of a type I error
α-level: boundary of acceptable level of type I error

Formulating Results

H_0	true	false
accepted	correct	type II error
rejected	type I error	correct

Note that results with $p = 0.06$ aren't very different from $p = 0.05$, but we need to specify a boundary. 0.05 is low because the "burden of proof" is on the alternative.
In these cases we certainly don't feel that we've **proven** H_0, only that we've failed to show convincingly that it's wrong.
We speak of "retaining H_0" ("H_0 *handhaven*").
Type II Errors (null hypothesis accepted by false)
β —probability of type II error
$1 - \beta$ —"power of statistical test" (no further mention in this course)

Degrees of Freedom

Most hypothesis-tests require that one specify Degrees of Freedom (dF) — the number of ways in which data could vary (and still yield same result).
Example: 5 data points, mean
If mean & 4 data points known, fifth is determined
Mean 6, data is $4, 5, 7, 8$ and one unknown
□ fifth $= 6$
There are **four** degrees of freedom in this set.
In general, with n numbers, $n - 1$ degrees of freedom (for the mean).
Reminder: Pencil & Paper Exercise on Sampling Statistics

Z-tests