Statistiek I

Sampling

John Nerbonne

CLCG, Rijks*universiteit* Groningen

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

Overview

Samples and Populations

Selecting a sample from a population includes an element of chance—which individuals are studied?

Big question: How to interpret samples scores?

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

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

university of groningen
Samples and Populations Confidence Intervals Hypotheses One-sided vs. two-sided Statistical Significance Error Types

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

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

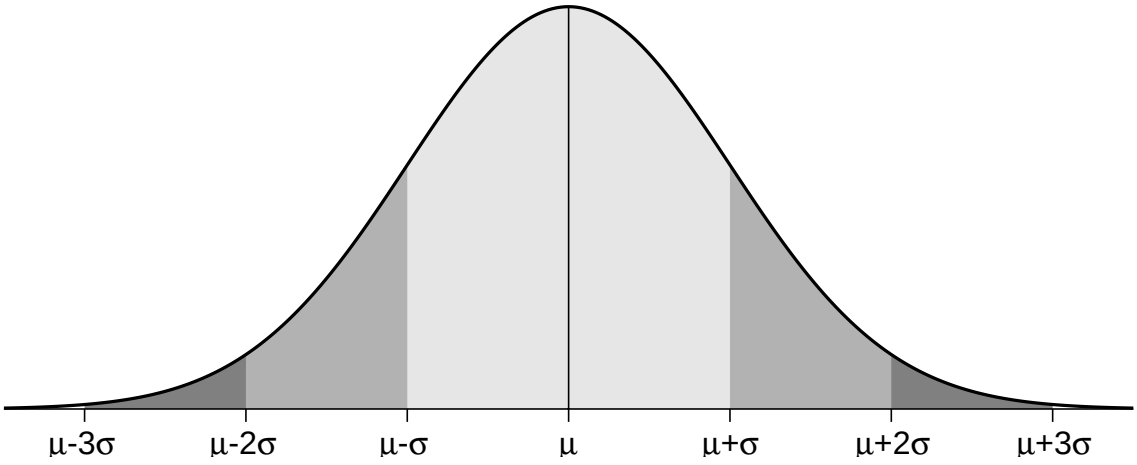
These facts allow us to reason about the population.

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

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

RANDOM SELCTION helps avoid bias.

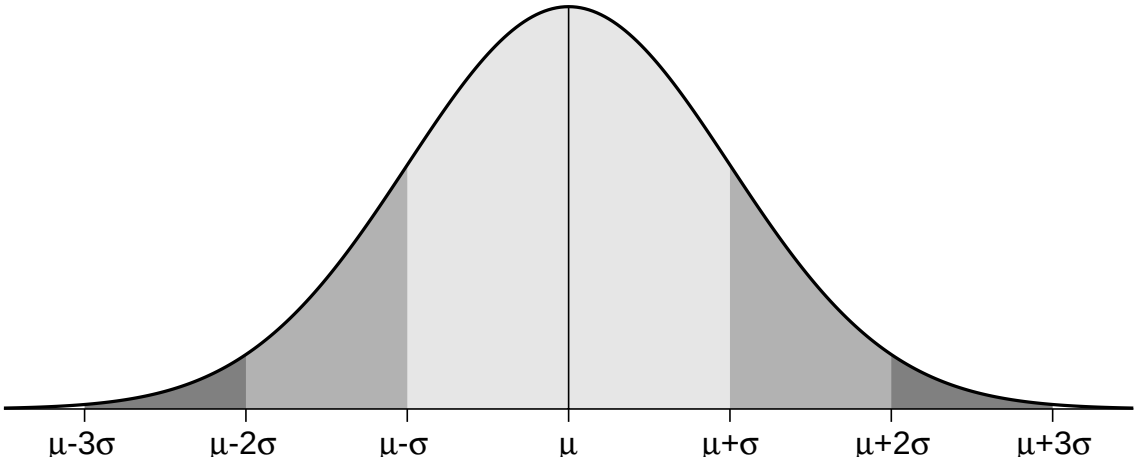
Normal Distribution (Review)



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 $\sigma = 1$:

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

Remarks on Central Limit Theorem

We examine the distribution of sample means, i.e. among all the possible samples (of a given size n) in the population:

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

- note importance of sample size
  - hard work pays off (in exactness)
  - but it doesn't pay off quickly ($\sqrt{n}$)
- what about population size?

Confidence Interval

Two ways to interpret sample means

- Confidence Intervals
- Hypothesis Testing

Confidence interval: true population mean is within an error margin of measured sample mean.

Confidence Interval

Example: you want to know how many hours per week a student in CIW 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 randomly chosen 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

- $\sigma = 1\text{hr.}/\text{wk}$, $n = 100$
- 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 samples will fall in interval $m \pm 1\text{sd (SE)} = m \pm 0.1$.
95% of samples will fall within $m \pm 2\text{sd (SE)} = m \pm 0.2$.

We are 95% confident that μ is in:
$5\text{hr.}/\text{wk.} \pm 0.2\text{hr.}/\text{wk.} = (4.8\text{hr.}/\text{wk.}, 5.2\text{hr.}/\text{wk.})$

where 5hr./wk. is the estimate, & 0.2hr./wk. the error margin

How much do students work (per week)?

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

We can specify many confidence intervals.

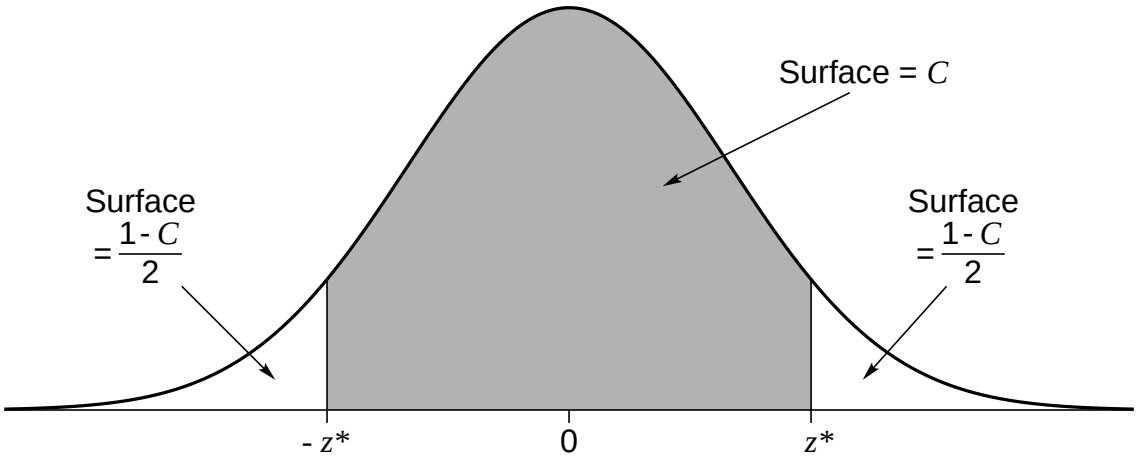
68%conf. interval	$m \pm 1\sigma(\text{SE})$	$4.9 \leq m \leq 5.1$	$(4.9, 5.1)$
95%conf. interval	$m \pm 2\sigma(\text{SE})$	$4.8 \leq m \leq 5.2$	$(4.8, 5.2)$
99.7%conf. interval	$m \pm 3\sigma(\text{SE})$	$4.7 \leq m \leq 5.3$	$(4.7, 5.3)$

Don't forget: SE is a kind of standard deviation (σ), nl. the standard deviation in the distribution of sample means.

Note that SE estimated from samples is $\text{sd}/\sqrt{n-1}$, but we assume that σ is known here.

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

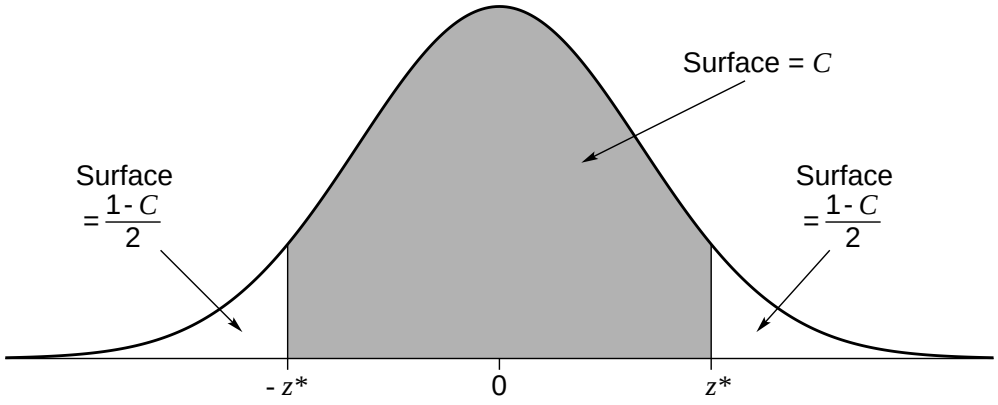
Confidence Interval



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 check whether a (background) hypothesis is compatible with a sample result (to refute a hypothesis at level α, a sample statistic outside the central $1-\alpha$ is needed—i.e., outside the $1-\alpha$ CONFIDENCE INTERVAL).

Instead of confidence intervals, we often interpret samples as tests of hypotheses about populations.

Humanities Statistics—Hypotheses

Lots of humanities issues are EMPIRICAL and VARIABLE

empirical — involving matters of fact, not purely conceptual

variable — different individual cases vary

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.

Field emphasizes the need to derive hypotheses from theory (Chap.1), but in applied studies theory is often too inexplicit and/or incomplete

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 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 w. banners are revisited more often than web sites w.o. 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 need statistics to abstract away from the variability of the observations (what Field calls "unssytematic variation", Chap.1).

- 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** (ünsystematic variation).

We normally need statistics to analyse results.

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

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** (ünsystematic variation).

We normally need statistics to analyse results.

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

Example Application of *z*-test

You suspect that CALL may be effective for young kids (as CALL can be used early on, and looks like games, needs little supervision, ...).

You have a standard test for proficiency, where $\mu = 70, \sigma = 14$. You apply the test to 49 randomly chosen children who've had a CALL program for three years. Result: $\bar{x} = 74$

Compute standard error SE $= \sigma/\sqrt{n} = 14/\sqrt{49} = 2$ for the sample. It is thus two σ's (standard deviations) above $\mu = 70$, 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.

Only 2.5% probability that the sample mean would be so high by chance.

CALL Conclusions

Kids who've used a CALL program score two σ's above the mean, (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, e.g. the sample happened to include lots of proficient kids. ... There might be many CONFOUNDING factors.

(Try to think of some.)

Note: HIDDEN VARIABLE are those **not** used in a study, CONFOUNDS are influential hidden variables. (Field, Chap.1)

university of groningen
Samples and Populations Confidence Intervals Hypotheses One-sided vs. two-sided Statistical Significance Error Types

CALL Conclusions

Kids who've used a CALL program score two σ's above the mean, (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, e.g. the sample happened to include lots of proficient kids. ... There might be many CONFOUNDING factors.

(Try to think of some.)

Note: HIDDEN VARIABLE are those **not** used in a study, CONFOUNDS are influential hidden variables. (Field, Chap.1)

Importance of Sample Size

Suppose you had tested only 9 kids who've used CALL, still with $\bar{x} = 74$, where the test still has $\mu = 70,\ \sigma = 14$.

Then standard error increases: SE $= \sigma/\sqrt{n} = 14/\sqrt{9} = 4.7$. In this case the sample ($\bar{x} = 74$) is less than 1 SE above population mean ($\mu = 70$), i.e., at less than the 84^{th} percentile — not very surprising. Samples means this high are found 16% 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 (due to unsystematic variation).

university of groningen
Samples and Populations Confidence Intervals Hypotheses One-sided vs. two-sided Statistical Significance Error Types

Importance of Sample Size

Suppose you had tested only 9 kids who've used CALL, still with $\bar{x} = 74$, where the test still has $\mu = 70,\ \sigma = 14$.

Then standard error increases: SE $= \sigma/\sqrt{n} = 14/\sqrt{9} = 4.7$. In this case the sample ($\bar{x} = 74$) is less than 1 SE above population mean ($\mu = 70$), i.e., at less than the 84^{th} percentile — not very surprising. Samples means this high are found 16% 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 (due to unsystematic variation).

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*

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

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

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

The tables show $P(z \geq 2) = 0.025$, and so the chance of sample this extreme is $P(\bar{x} = 74) = 0.025$. This is the p-VALUE, aka MEASURED SIGNIFICANCE LEVEL, or *overschrijdingskans*.

If H_0 were correct, and kids with CALL experience had the same language proficiency as others, then the observed sample would be expected only 2.5% of the time. Small p-values are strong evidence *against* the null hypothesis.

Statistically Significant?

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

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

The classical hypothesis test specifies how *un*likely a sample must be for a test to count as significant, the threshold SIGNIFICANCE LEVEL, or α-LEVEL.

We compare the *p*-value against the required threshhold α. 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.

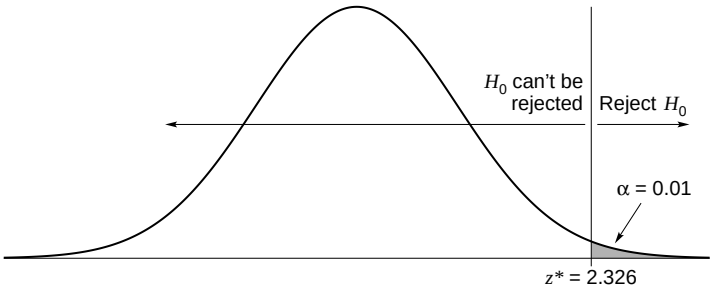
So the result was SIGNIFICANT AT THE $\alpha = 0.05$ LEVEL, but not at the 0.01-level.

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.

$$z_x = (x - \mu)/\sigma$$
$$1.65 = (x - 70)/2$$
$$3.3 = x - 70$$
$$x = 73.3$$

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

One-sided z-tests

CALL example is a z-test because it is based on a normal distribution whose mean μ and sd σ are known.
We calculate the mean of a random sample m, and a z-value based on it, where z is, as usual, $z = (m - \mu)/(\sigma/\sqrt{n})$

z-tests take many forms, depending on values predicted by H_a.

H_a predicts high m CALL *improves* language ability. $p = P(Z \geq z)$

H_a predicts low m Brocoli eaters have *low* chlosterol levels. $p = P(Z \leq z)$

—called ONE-SIDED tests because H_0 is 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 validate a test for aphasia developed in UK (after translation). Developers claim scores are distributed $N(100, 10)$ on nonaphasics. To validate its use in NL, you test it on 25 normal Dutch children.

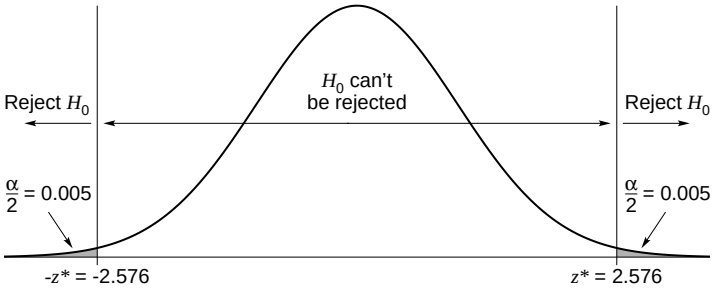
Here H_0 predicts $\mu_O = \mu_T$ (translation has same mean as original), and $H_a : \mu_O \neq \mu_T$, but without specifying whether μ_T is higher or lower than μ_O.

Suppose we 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 distribution is divided into lowest 0.5% and highest 0.5%. α-values reflect the probability of either $Z \geq z$ or $Z \leq -z$. z-values in either tail of the distribution justify rejection of H_0.

Understanding Significance

Recall the language learning example. Hypotheses were:

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

Given 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:

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

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 (EFFECT SIZE) is the difference (4 pt.), (nearly $0.3 \cdot \sigma$) constant in the hypothetical example.

"Statistically significant" implies that an effect probably is not due to chance, but the effect can be very small. You may want to know e.g. whether you should buy software for your kids, but "statistically significant" does not tell you this.

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. It could be a chance effect.

Misuse of Significance

Garbage in, garbage out If the experiment poorly designed, or the data poorly collected, statistical sophistication cannot repair the situation.

No "significance hunting" (Field, Chap.2 "cheating") Hypotheses should be formulated before data collection and analysis.

- Modern danger: Hunting among dozens of variables is *likely* to turn up *some* extreme results. Multiple tests need to be analyzed in special ways 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.

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 $\neg H_a$

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

Results with $p = 0.06$ aren't very different from $p = 0.05$, but we need a boundary. 0.05 is low because the "burden of proof" is on the alternative.

In such cases we certainly haven't **proven** H_0, only 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
$\square$ 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

Next Week

Pencil 'n paper exercise on sampling statistics.
SPSS Lab 1

t-tests