Statistics 2

Overview Statistics I

Casper Albers & Jorge Tendeiro

Lecture 0, 2019 – 2020

university of groningen

Statistics II

- Statistics II will introduce a range of inferential methods for finding relations between variables in a wide range of practical settings.
- These methods all continue from the basics of statistical inference you have learned in Statistics I:
  Hypothesis testing, confidence intervals, p-values, the t-test, normal distribution, z-scores, checking assumptions,...
- In Statistics II we assume you have active knowledge of these topics.
- Recap this material in the textbook (Chapters 4, 5, 6).

Goal of today: A refresher of these topics.

Overview

Inference: To derive as a conclusion from facts or premises.

Confidence intervals (CIs)

- An x% CI contains an (unknown) population parameter with x% certainty.
- When repeating the study many times, about x% of the CIs will contain the parameter.

Hypothesis testing

- The probability of the current sample result (or more extreme) is so small, under the null hypothesis, that it is unlikely that the population parameter has a certain value (defined under $\mathcal{H}_0$).

Inferential statistics – Example

Inferring number of sex partners

How many male sex partners did females have since their 18th birthday (age 23-29)?

N	Mean	SD
129	6.6	13.3

95% CI $= (4.4, 8.8)$

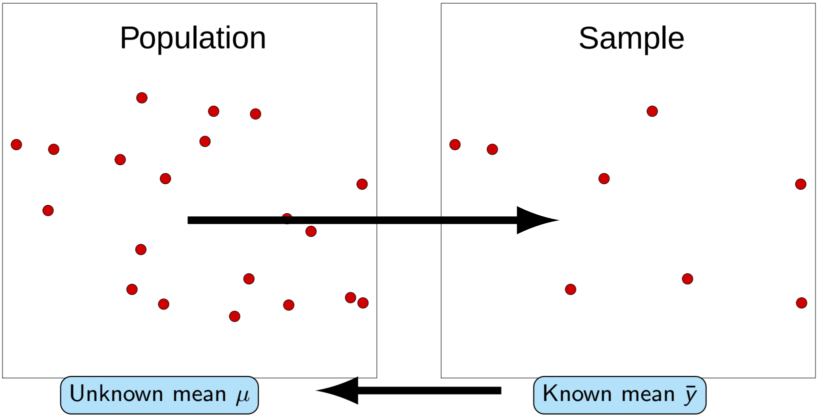
We are 95% confident that μ, the population mean number of sex partners, lies in this interval.

$\mathcal{H}_0 : \mu = 1$ **vs** $\mathcal{H}_a : \mu > 1$

- $t(128) = 4.78$, $p < .001$.
- If $\mathcal{H}_0$ were true, the sample result would be extremely unlikely. Thus, we reject $\mathcal{H}_0$.

Population and sample

Use a fact about a sample to estimate the truth about the whole population.



Population and sample

Example: The sample mean $\overline{y}$ and the population mean μ.

The sample mean $\overline{y}$ can be used to:

- Estimate μ.
- Make probabilistic statements about μ:
  - "The 95% CI for μ is $(4.4, 8.8)$."
  - We reject the hypothesis that $\mu = 1$ at $\alpha = 5\%$.

To make such probabilistic statements we need knowledge about the sampling distribution of the statistic.

Sampling distribution

Understanding the sampling distribution of the sample mean (for a fixed sample size n):

1. Collect a sample. Compute the sample mean: $\overline{y}_1$.
2. Collect a sample. Compute the sample mean: $\overline{y}_2$.
3. Collect a sample. Compute the sample mean: $\overline{y}_3$.
4. ... (say, some hundreds of times)

This provides a set of sample means: $\{\overline{y}_1, \overline{y}_2, \overline{y}_3, \ldots\}$.
This set of scores has a certain distribution:

The sampling distribution of the mean.

Of course, the same principle generalizes to any statistic other than the sample mean.

Sampling distribution

Hence,

> *The sampling distribution is the probability distribution of a statistic in the sample.*

What do we know about the sampling distribution of $\overline{y}$ that allows us using it to estimate μ?

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

Sampling distribution

Sampling distribution of $\overline{y}$:

- Mean := $\mu_{\overline{y}} = \mu$.
- SD := $\sigma_{\overline{y}} = \frac{\sigma}{\sqrt{n}}$: Standard Error (SE) of the mean.
- It is normally distributed if the population of y values is also normally distributed (regardless of the sample size n):

$$y_i \sim N(\mu, \sigma) \Longrightarrow \overline{y} \sim N(\mu_{\overline{y}}, \sigma_{\overline{y}}) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

- If the population of y values is not normally distributed, use the Central Limit Theorem (CLT):

For a random sample of size n from an arbitrary distribution with mean μ and standard deviation σ, the sampling distribution of the sample mean is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$, if n is large.

Q: What's the use of sampling distributions for inferential statistics?

A: Sampling distributions help quantifying which values of the statistic are most/least probable. This allows associating probabilities to sample values:

- Significance tests: p-values.
- Confidence intervals: The lower and upper boundaries.

Significance tests

Underlying principles:

- A formal procedure for comparing observed data with an hypothesis whose truth we want to assess.
- It is intended to assess the evidence provided by data against $\mathcal{H}_0$ and in favor of $\mathcal{H}_a$.

There are two types of hypotheses in significance testing:

- Null hypothesis ($\mathcal{H}_0$): Statement quantifying a value for the population parameter of interest.
- Alternative hypothesis ($\mathcal{H}_a$): Statement contradicting the null hypothesis (smaller, larger, different).

The alternative hypothesis always contradicts the null hypothesis.
Example:

$$\mathcal{H}_0 : \mu = 0 \quad \text{versus} \quad \mathcal{H}_a : \mu \neq 0$$

Significance tests

Each significance test is based on a test statistic..

General form of a test statistic for z-tests and t-tests:

$$\text{test statistic} = \frac{\text{(estimate statistic)} - (\text{hypothesized value}|\mathcal{H}_0 \text{ is true})}{\text{SE}_{\text{statistic}}}$$

From the sample

From the population

Example:

One-sample z test

- In the population: $y \sim N(\mu, \sigma)$, σ known.
- Sampling distribution: $\overline{y} \sim N(\mu, \sigma/\sqrt{n})$.

$$Z = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Significance tests: p-values

The p-value is the probability of getting an outcome as extreme or more extreme than the actually observed outcome from the sample, given that $\mathcal{H}_0$ were true.

- The smaller the p-value, the stronger the evidence against $\mathcal{H}_0$, that is, the more unlikely $\mathcal{H}_0$ is.
- What is 'small'?
  Compare p with the significance level α (e.g., $\alpha = 5\%$).

Significance tests: One-sample t test

- $y \sim N(\mu, \sigma)$. Both μ and σ unknown.
- $\mathcal{H}_0 : \mu = \mu_0$.
- Estimate σ with $s = \sqrt{\frac{1}{n-1} \sum_i (y_i - \overline{y})^2}$.
- Test statistic:
$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} \sim t(n-1).$$

The t distribution is used to compute the p-value.

Significance tests: Two-sample t test

- $y_1 \sim N(\mu_1, \sigma_1)$, $y_2 \sim N(\mu_2, \sigma_2)$. All μs and σs unknown.
- $\mathcal{H}_0 : \mu_1 = \mu_2$ or, equivalently, $\mathcal{H}_0 : \mu_1 - \mu_2 = 0$.
- Test statistic
  - If we can assume that $\sigma_1 = \sigma_2$ (ideal):

    $$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

    where s_p = pooled SD $= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}$.
  - Otherwise:

    $$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(k)$$

    (k approximated by software).

The t distribution is used to compute the p-value.

Example: ‘Sesame Street’ Data.

- Two populations:
  - Boys ($n_1 = 115$, mean $= 26.39$).
  - Girls ($n_2 = 125$, mean $= 26.98$).
- Pooled SD $= 13.30$.
- y = POSTLET, knowledge of the alphabet.

Independent Samples Test

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Diff.	SE. Diff.	95% Confidence Interval of the Difference: Lower	95% Confidence Interval of the Difference: Upper
POSTLET	Equal variances assumed	.340	238	.734	-.5847	1.72	-3.9694	2.8000
	Equal variances							

$$t = \frac{26.39 - 26.98}{13.30 \times \sqrt{1/115 + 1/125}}$$

df = 115+125-2

Don’t reject $\mathcal{H}_0$

Significance tests: Which and where

- Regression
  - t-tests: Parameters.
  - F-tests: Model fit.
- Correlation
  - t-test: Special case ($\mathcal{H}_0 : \rho = 0$).
  - z-test: Fisher's Z transformation.
- Analysis of variance (ANOVA)
  - t-tests: Contrasts, multiple comparisons.
  - F-tests: Model fit.

Confidence intervals (CIs)

Underlying principles:

- Numerical interval that, with a given degree of certainty, contains the value of a population parameter.
- Confidence level: The degree of certainty, most commonly 95%.
- After repeating the experiment many times, 95% of the times the confidence interval will contain the population parameter.

$$\text{CI} = \text{estimate} \pm \underbrace{\text{margin of error}}_{\text{Critical value from sampling distribution} \times \underbrace{\text{SE}}}$$

From the sample

Confidence intervals: Mean of one population

Known σ: z confidence interval

- In the population: $y \sim N(\mu, \sigma)$, σ known.

$$\text{CI} = \overline{y} \pm z^* \times \boxed{\frac{\sigma}{\sqrt{n}}}$$

SE of the mean

$z^* =$ critical value from $N(0,1)$

Unknown σ: t confidence interval

- In the population: $y \sim N(\mu, \sigma)$. Both μ and σ unknown.
- Estimate σ with $s = \sqrt{\frac{1}{n-1}\sum_i (y_i - \overline{y})^2}$.

$$\text{CI} = \overline{y} \pm t^* \times \frac{s}{\sqrt{n}}$$

$t^* =$ critical value from $t(n-1)$

- Assume equal variances:
  $y_1 \sim N(\mu_1, \sigma)$, $y_2 \sim N(\mu_2, \sigma)$.
  μ's and σ unknown.
- Sample sizes: n_1, n_2.
- Recall the test statistic:

$$t = \frac{(\overline{y}_1 - \overline{y}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2), s_p = \text{pooled SD}$$

- Confidence interval:

$$\text{CI} = (\overline{y}_1 - \overline{y}_2) \pm t^* \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$t^* =$ critical value from $t(n_1 + n_2 - 2)$

Example: 'Sesame Street' Data

- Two populations:
  - Boys ($n_1 = 115$, mean $= 26.39$).
  - Girls ($n_2 = 125$, mean $= 26.98$).
- Pooled SD $= 13.30$.
- y = POSTLET, knowledge of the alphabet.

Independent Samples Test

		t-test for Equality of Means					95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Diff.	SE. Diff.	Lower	Upper
POSTLET	Equal variances assumed	-.340	238	.734	-.5847	1.72	-3.9694	2.8000
	Equal variances							

Significance test

Diff $\pm t^*(238) \times$ SE$_{\text{Diff}}$

Confidence intervals

The margin of error of a CI decreases (i.e., the CI becomes smaller) if:

- The confidence level decreases.
- The sample size increases.
- The SD decreases.

Possible correct interpretations of a 95% CI $= (a, b)$:

- We say that we are 95% confident that the unknown parameter lies between a and b.
- We arrived at these numbers by a method that gives correct results 95% of the time.
- In the long run, 95% of all samples lead to an interval that covers the unknown parameter.

Confidence intervals

What you cannot say about a 95% CI:

> *There is 95% probability that the unknown parameter from the population is inside the CI.*

Q: Why not?

A: The population parameter that one wishes to estimate is supposed to be fixed (and unknown, obviously).

A specific CI either contains or does not contain the parameter.

This is not a matter of probability.

Property:

A two-sided test with significance level α rejects the hypothesis

$$\mathcal{H}_0 : \mu = \mu_0$$

if and only if the value μ_0 lies outside the $(1-\alpha)\%$ CI for μ.

Example

Suppose the 95% CI for the difference between the means of two groups is

$$95\% \text{ CI } = (-3.97, 2.80).$$

It can be concluded that the null hypothesis $\mathcal{H}_0 : \mu_1 = \mu_2$ cannot be rejected for $\alpha = 5\%$ because 0 $(= \mu_1 - \mu_2)$ is contained in the CI.

Next week

Start of the course. Contents of Lecture 1:

- Rules and regulations of the course: Lectures, practicals, homework, exam.
- Overview of methods to be introduced in Statistics II.

Read:

- No new reading material. Make sure your Statistics I knowledge is up-to-date.