

## Unit 3: Inferential Statistics for Continuous Data

### Statistics for Linguists with R – A SIGIL Course

Designed by Marco Baroni<sup>1</sup> and Stefan Evert<sup>2</sup>

<sup>1</sup>Center for Mind/Brain Sciences (CIMEC)  
University of Trento, Italy

<sup>2</sup>Corpus Linguistics Group  
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

## Outline

### Inferential statistics

#### Preliminaries

### One-sample tests

Testing the mean  
Testing the variance  
Student's  $t$  test  
Confidence intervals

### Two-sample tests

Comparing the means of two samples  
Comparing the variances of two samples  
The paired  $t$  test for related samples  
Multiple comparisons

## Outline

### Inferential statistics

#### Preliminaries

### One-sample tests

Testing the mean  
Testing the variance  
Student's  $t$  test  
Confidence intervals

### Two-sample tests

Comparing the means of two samples  
Comparing the variances of two samples  
The paired  $t$  test for related samples  
Multiple comparisons

## Inferential statistics for continuous data

- ▶ Goal: infer (characteristics of) population distribution from small random sample, or test hypotheses about population
  - ▶ problem: overwhelmingly infinite choice of possible distributions
  - ▶ can estimate/test characteristics such as mean  $\mu$  and s.d.  $\sigma$
  - ▶ but  $H_0$  doesn't determine a unique sampling distribution
  - ▶ only makes sense for **parametric** model
- ▶ In this session, we assume a **Gaussian population** distribution
  - ▶ estimate/test parameters  $\mu$  and  $\sigma$  of this distribution
  - ▶ sometimes a scale transformation is necessary (e.g. lognormal)
- ▶ Nonparametric tests need fewer assumptions, but ...
  - ▶ cannot test hypotheses about  $\mu$  and  $\sigma$  (instead: median, IQR = inter-quartile range, etc.)
  - ▶ more complicated and computationally expensive procedures
  - ▶ correct interpretation of results often difficult

## A note on extremeness

- ▶ Rationale similar to binomial test for frequency data: measure observed **statistic** in sample, which is compared against **expected** value → if difference is large, reject  $H_0$
- ▶ Crucial question: what is “large enough”?
  - ☞ reject if difference is unlikely to arise by chance
- ▶ Measuring the extremeness of a single item sampled from  $\Omega$ 
  - ▶ If someone is 195 cm tall, would we consider him unusual?
  - ▶ no absolute scale → “ordinary” defined by central range, i.e. how tall the majority of people we meet are (say, 95%)
  - ▶ for Gaussian distribution: range from  $\mu - 1.96\sigma$  to  $\mu + 1.96\sigma$
- ▶ This suggests the **z-score** measure of extremeness:

$$Z(\omega) := \frac{X(\omega) - \mu}{\sigma}$$

with central range characterised by  $|Z| \leq 1.96$

## Notation for random samples

- ▶ Random sample of  $n \ll m$  items
  - ▶ e.g. participants of survey, Wikipedia sample, ...
  - ▶ recall importance of completely random selection
- ▶ Sample described by observed values of r.v.  $X, Y, Z, \dots$ :

$$x_1, \dots, x_n; \quad y_1, \dots, y_n; \quad z_1, \dots, z_n$$

(don't know which  $\omega \in \Omega$  were selected →  $x_i$  instead of  $X(\omega)$ )

- ▶ Mathematically,  $x_i, y_i, z_i$  are realisations of random variables

$$X_1, \dots, X_n; \quad Y_1, \dots, Y_n; \quad Z_1, \dots, Z_n$$

- ▶  $X_1, \dots, X_n$  are independent from each other and each one has the same distribution  $X_i \sim X$  → **i.i.d.** random variables
  - ☞ this is the formal definition of a random sample

## Outline

Inferential statistics  
Preliminaries

## One-sample tests

Testing the mean  
Testing the variance  
Student's  $t$  test  
Confidence intervals

## Two-sample tests

Comparing the means of two samples  
Comparing the variances of two samples  
The paired  $t$  test for related samples  
Multiple comparisons

## A simple test for the mean

- ▶ Consider simplest possible  $H_0$ : a **point hypothesis**

$$H_0: \mu = \mu_0, \sigma = \sigma_0$$

- ☞ together with normality assumption, population distribution is completely determined
- ▶ How would you test whether  $\mu = \mu_0$  is correct?
- ▶ An intuitive test statistic is the **sample mean**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{with} \quad \bar{x} \approx \mu_0 \text{ under } H_0$$

- ▶ Reject  $H_0$  if difference  $\bar{x} - \mu_0$  is sufficiently large
  - ☞ need to work out sampling distribution of  $\bar{X}$

The sampling distribution of  $\bar{X}$ 

- ▶ The sample mean is also a random variable:

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$$

- ▶  $\bar{X}$  is a sensible test statistic for  $\mu$  because it is **unbiased**:

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

- ▶ An important property of the Gaussian distribution: if  $X \sim N(\mu, \sigma_1^2)$  and  $Y \sim N(\mu, \sigma_2^2)$  are independent, then

$$\begin{aligned} X + Y &\sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\ r \cdot X &\sim N(r\mu_1, r^2\sigma_1^2) \quad \text{for } r \in \mathbb{R} \end{aligned}$$

The sampling distribution of  $\bar{X}$ 

- ▶ Since  $X_1, \dots, X_n$  are i.i.d. with  $X_i \sim N(\mu, \sigma^2)$ , we have

$$\begin{aligned} X_1 + \dots + X_n &\sim N(n\mu, n\sigma^2) \\ \bar{X} = \frac{1}{n}(X_1 + \dots + X_n) &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \end{aligned}$$

- ▶  $\bar{X}$  has Gaussian distribution with same  $\mu$  but smaller s.d. than the original r.v.  $X$ :  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

- explains why normality assumptions are so convenient
  - larger samples allow more reliable hypothesis tests about  $\mu$

- ▶ If the sample size  $n$  is large enough,  $\sigma_{\bar{X}} = \sigma/\sqrt{n} \rightarrow 0$  and the sample mean  $\bar{x}$  becomes an accurate estimate of the true population value  $\mu$  (**law of large numbers**)

## The z test

- ▶ Now we can quantify the extremeness of the observed value  $\bar{x}$ , given the null hypothesis  $H_0 : \mu = \mu_0, \sigma = \sigma_0$

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}$$

- ▶ Corresponding r.v.  $Z$  has a standard normal distribution if  $H_0$  is correct:  $Z \sim N(0, 1)$
- ▶ We can reject  $H_0$  at significance level  $\alpha$  if

$$\begin{array}{cccc} \alpha = & .05 & .01 & .001 \\ |z| > & 1.960 & 2.576 & 3.291 \end{array} \quad \text{--}\text{qnorm}(\alpha/2)$$

- ▶ Two problems of this approach:
  1. need to make hypothesis about  $\sigma$  in order to test  $\mu = \mu_0$
  2.  $H_0$  might be rejected because of  $\sigma \gg \sigma_0$  even if  $\mu = \mu_0$  is true

## Outline

Inferential statistics  
Preliminaries

## One-sample tests

Testing the mean  
Testing the variance  
Student's  $t$  test  
Confidence intervals

## Two-sample tests

Comparing the means of two samples  
Comparing the variances of two samples  
The paired  $t$  test for related samples  
Multiple comparisons

## A test for the variance

- ▶ An intuitive test statistic for  $\sigma^2$  is the sum of squares

$$V = (X_1 - \mu)^2 + \cdots + (X_n - \mu)^2$$

- ▶ Squared error  $(X - \mu)^2$  is  $\sigma^2$  on average  $\rightarrow E[V] = n\sigma^2$ 
  - ▶ reject  $\sigma = \sigma_0$  if  $V \gg n\sigma_0^2$  (variance larger than expected)
  - ▶ reject  $\sigma = \sigma_0$  if  $V \ll n\sigma_0^2$  (variance smaller than expected)
  - ▶ sampling distribution of  $V$  shows if difference is large enough

- ▶ Rewrite  $V$  in the following way:

$$V = \sigma^2 \left[ \left( \frac{X_1 - \mu}{\sigma} \right)^2 + \cdots + \left( \frac{X_n - \mu}{\sigma} \right)^2 \right] \\ = \sigma^2 (Z_1^2 + \cdots + Z_n^2)$$

with  $Z_i \sim N(0, 1)$  i.i.d. standard normal variables

## A test for the variance

- ▶ Statisticians have worked out the distribution of  $\sum_{i=1}^n Z_i^2$  for i.i.d.  $Z_i \sim N(0, 1)$ , known as the **chi-squared distribution**

$$\sum_{i=1}^n Z_i^2 \sim \chi_n^2$$

with  $n$  **degrees of freedom** ( $df = n$ )

- ▶ The  $\chi_n^2$  distribution has expectation  $E[\sum_i Z_i^2] = n$  and variance  $\text{Var}[\sum_i Z_i^2] = 2n \rightarrow$  confirms  $E[V] = n\sigma^2$

## A test for the variance

- ▶ Under  $H_0 : \sigma = \sigma_0$ , we have

$$\frac{V}{\sigma_0^2} = Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2$$

- ▶ Appropriate rejection thresholds for the test statistic  $V/\sigma_0^2$  can easily be obtained with R
  - ▶  $\chi_n^2$  distribution is not symmetric, so one-sided tail probabilities are used (with  $\alpha' = \alpha/2$  for two-sided test)
- ▶ Again, there are two problems:
  1. need to make hypothesis about  $\mu$  in order to test  $\sigma = \sigma_0$
  2.  $H_0$  easily rejected for  $\mu \neq \mu_0$ , even though  $\sigma = \sigma_0$  may be true

## Intermission: Distributions in R

- ▶ R can compute density functions and tail probabilities or generate random numbers for a wide range of distributions
- ▶ Systematic naming scheme for such functions:
  - d**norm() density function of Gaussian (normal) distribution
  - p**norm() tail probability
  - q**norm() quantile = inverse tail probability
  - r**norm() generate random numbers
- ▶ Available distributions include Gaussian (**norm**), chi-squared (**chisq**),  $t$  (**t**),  $F$  (**f**), binomial (**binom**), Poisson (**pois**), ...
  - ▶ you will encounter many of them later in the course
- ▶ Each function accepts distribution-specific parameters

## Intermission: Distributions in R

```
> x <- rnorm(50, mean=100, sd=15) # random sample of 50 IQ scores
> hist(x, freq=FALSE, breaks=seq(45,155,10)) # histogram

> xG <- seq(45, 155, 1) # theoretical density in steps of 1 IQ point
> yG <- dnorm(xG, mean=100, sd=15)
> lines(xG, yG, col="blue", lwd=2)

# What is the probability of an IQ score above 150?
# (we need to compute an upper tail probability to answer this question)
> pnorm(150, mean=100, sd=15, lower.tail=FALSE)

# What does it mean to be among the bottom 25% of the population?
> qnorm(.25, mean=100, sd=15) # inverse tail probability
```

## Intermission: Distributions in R

```
# Now do the same for a chi-squared distribution with 5 degrees of freedom
# (hint: the parameter you're looking for is df=5)

> xC <- seq(0, 10, .1)
> yC <- dchisq(xC, df=5)
> plot(xC, yC, type="l", col="blue", lwd=2)

# tail probability for  $\sum_i Z_i^2 \geq 10$ 
> pchisq(10, df=5, lower.tail=FALSE)

# What is the appropriate rejection criterion for a variance test with  $\alpha = 0.05$ ?
> qchisq(.05, df=5, lower.tail=FALSE) # one-sided test
```

## The sample variance

- Idea: replace true  $\mu$  by sample value  $\bar{X}$  (which is a r.v.!)
 
$$V' = (X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2$$

⚠ terms are no longer i.i.d. because  $\bar{X}$  depends on all  $X_i$

- We can work out the distribution of  $V'$  for  $n = 2$ :

$$\begin{aligned} V' &= (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 \\ &= (X_1 - \frac{X_1 + X_2}{2})^2 + (X_2 - \frac{X_1 + X_2}{2})^2 \\ &= (\frac{X_1 - X_2}{2})^2 + (\frac{X_2 - X_1}{2})^2 = \frac{1}{2}(X_1 - X_2)^2 \end{aligned}$$

where  $X_1 - X_2 \sim N(0, 2\sigma^2)$  for i.i.d.  $X_1, X_2 \sim N(\mu, \sigma^2)$

⚠ one can also show that  $X_1 - X_2$  and  $\bar{X}$  are independent

## The sample variance

- We now have

$$V' = \sigma^2 \left( \frac{X_1 - X_2}{\sigma\sqrt{2}} \right)^2 = \sigma^2 Z^2$$

with  $Z^2 \sim \chi_1^2$  because of  $X_1 - X_2 \sim N(0, 2\sigma^2)$

- For  $n > 2$  it can be shown that

$$V' = \sum_{i=1}^n (X_i - \bar{X})^2 = \sigma^2 \sum_{j=1}^{n-1} Z_j^2$$

with  $\sum_j Z_j^2 \sim \chi_{n-1}^2$  independent from  $\bar{X}$

- proof based on multivariate Gaussian and vector algebra
- notice that we “lose” one degree of freedom because one parameter ( $\mu \approx \bar{x}$ ) has been estimated from the sample

## Sample variance and the chi-squared test

- ▶ This motivates the following definition of **sample variance**  $S^2$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

with sampling distribution  $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$

- ▶  $S^2$  is an unbiased estimator of variance:  $E[S^2] = \sigma^2$
- ▶ We can use  $S^2$  to test  $H_0: \sigma = \sigma_0$  without making any assumptions about the true mean  $\mu$  → **chi-squared test**
- ▶ Remarks
  - ▶ sample variance  $(\frac{1}{n-1})$  vs. population variance  $(\frac{1}{n})$
  - ▶  $\chi^2$  distribution doesn't have parameters  $\sigma^2$  etc., so we need to specify the distribution of  $S^2$  in a roundabout way
  - ▶ independence of  $S^2$  and  $\bar{X}$  will play an important role later

## Sample data for this session

```
# Let us take a reproducible sample from the population of Ingary
> library(SIGIL)
> Census <- simulated.census()
> Survey <- Census[1:100, ]

# We will be testing hypotheses about the distribution of body heights
> x <- Survey$height # sample data: n items
> n <- length(x)
```

## Chi-squared test of variance in R

```
# Chi-squared test for a hypothesis about the s.d. (with unknown mean)
# H0: σ = 12 (one-sided test against σ > σ0)
> sigma0 <- 12 # you can also use the name σ0 in a Unicode locale
> S2 <- sum((x - mean(x))^2) / (n-1) # unbiased estimator of σ^2
> S2 <- var(x) # this should give exactly the same value
> X2 <- (n-1) * S2 / sigma0^2 # has χ^2 distribution under H0
> pchisq(X2, df=n-1, lower.tail=FALSE)

# How do you carry out a one-sided test against σ < σ0?

# Here's a trick for an approximate two-sided test (try e.g. with σ0 = 20)
> alt.higher <- S2 > sigma0^2
> 2 * pchisq(X2, df=n-1, lower.tail=!alt.higher)
```

## Outline

Inferential statistics  
Preliminaries

## One-sample tests

Testing the mean  
Testing the variance  
Student's t test  
Confidence intervals

## Two-sample tests

Comparing the means of two samples  
Comparing the variances of two samples  
The paired t test for related samples  
Multiple comparisons

Student's *t* test for the mean

- ▶ Now we have the ingredients for a test of  $H_0 : \mu = \mu_0$  that does not require knowledge of the true variance  $\sigma^2$

- ▶ In the z-score for  $\bar{X}$

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

replace the unknown true s.d.  $\sigma$  by the unbiased sample estimate  $\hat{\sigma} = \sqrt{S^2}$ , resulting in a so-called **t-score**:

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}}$$

- ▶ William S. Gosset worked out the precise sampling distribution of  $T$  and published it under the pseudonym "Student"

Student's *t* test for the mean

- ▶ Because  $\bar{X}$  and  $S^2$  are independent, we find that

$$T \sim t_{n-1} \quad \text{under } H_0 : \mu = \mu_0$$

Student's **t distribution** with  $df = n - 1$  degrees of freedom

- ▶ In order to carry out a one-sample *t* test, calculate the statistic

$$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}}$$

and reject  $H_0 : \mu = \mu_0$  if  $|t| > C$

- ▶ Rejection threshold  $C$  depends on  $df = n - 1$  and desired significance level  $\alpha$  (in R: `-qt( $\alpha/2$ ,  $n - 1$ )`)

close to z-score thresholds for  $n > 30$

The mathematical magic behind Student's *t* test

- ▶ Student's *t* distribution characterizes the quantity

$$\frac{Z}{\sqrt{V/k}} \sim t_k$$

where  $Z \sim N(0, 1)$  and  $V \sim \chi_k^2$  are **independent** r.v.

- ▶  $T \sim t_{n-1}$  under  $H_0 : \mu = \mu_0$  because the unknown population variance  $\sigma^2$  cancels out between the independent r.v.  $\bar{X}$  and  $S^2$

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} = \frac{\frac{\bar{X} - \mu_0}{\sigma}}{\sqrt{\frac{S^2}{n\sigma^2}}} = \frac{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}$$

with  $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$  and  $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$

One-sample *t* test in R

# we will use the same sample *x* of size *n* as in the previous example

# Student's t-test for a hypothesis about the mean (with unknown s.d.)

#  $H_0 : \mu = 165$  cm

> mu0 <- 165

> x.bar <- mean(x) # sample mean  $\bar{x}$

> s2 <- var(x) # sample variance  $s^2$

> t.score <- (x.bar - mu0) / sqrt(s2 / n) # *t* statistic

> print(t.score) # positive indicates  $\mu > \mu_0$ , negative  $\mu < \mu_0$

> -qt(0.05/2, n-1) # two-sided rejection threshold for  $|t|$  at  $\alpha = .05$

> 2 \* pt(abs(t.score), n-1, lower=FALSE) # two-sided p-value

# Mini-task: plot density function of *t* distribution for different d.f.

> t.test(x, mu=165) # agrees with our "manual" t-test

# Note that `t.test()` also provides a confidence interval for the true  $\mu$ !

## Outline

### Inferential statistics

#### Preliminaries

### One-sample tests

#### Testing the mean

#### Testing the variance

#### Student's $t$ test

#### Confidence intervals

### Two-sample tests

#### Comparing the means of two samples

#### Comparing the variances of two samples

#### The paired $t$ test for related samples

#### Multiple comparisons


## Confidence intervals

- ▶ If we do not have a specific  $H_0$  to start from, estimate **confidence interval** for  $\mu$  or  $\sigma^2$  by inverting hypothesis tests
  - ▶ in principle same procedure as for binomial confidence intervals
  - ▶ implemented in R for  $t$  test and chi-squared test
- ▶ Confidence interval has a particularly simple form for the  $t$  test
- ▶ Given  $H_0 : \mu = a$  for some  $a \in \mathbb{R}$ , we reject  $H_0$  if

$$|t| = \left| \frac{\bar{x} - a}{\sqrt{s^2/n}} \right| > C$$

with  $C \approx 2$  for  $\alpha = .05$  and  $n > 30$

$$\Rightarrow \bar{x} - C \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + C \frac{s}{\sqrt{n}}$$

 this is the origin of the “ $\pm 2$  standard deviations” rule of thumb

## Outline

### Inferential statistics

#### Preliminaries

### One-sample tests

#### Testing the mean

#### Testing the variance

#### Student's $t$ test

#### Confidence intervals

### Two-sample tests

#### Comparing the means of two samples

#### Comparing the variances of two samples

#### The paired $t$ test for related samples

#### Multiple comparisons



## Outline

### Inferential statistics

Preliminaries

### One-sample tests

Testing the mean

Testing the variance

Student's  $t$  test

Confidence intervals

### Two-sample tests

Comparing the means of two samples

Comparing the variances of two samples

The paired  $t$  test for related samples

Multiple comparisons

## Outline

### Inferential statistics

#### Preliminaries

### One-sample tests

#### Testing the mean

#### Testing the variance

#### Student's $t$ test

#### Confidence intervals

### Two-sample tests

#### Comparing the means of two samples

#### Comparing the variances of two samples

#### The paired $t$ test for related samples

#### Multiple comparisons

## Outline

### Inferential statistics

#### Preliminaries

### One-sample tests

#### Testing the mean

#### Testing the variance

#### Student's $t$ test

#### Confidence intervals

### Two-sample tests

#### Comparing the means of two samples

#### Comparing the variances of two samples

#### The paired $t$ test for related samples

#### Multiple comparisons

