Unit 3: Inferential Statistics for Continuous Data

Statistics for Linguists with R – A SIGIL Course

Designed by Marco Baroni[1] and Stefan Evert[2]

[1]Center for Mind/Brain Sciences (CIMeC)
University of Trento, Italy

[2]Institute of Cognitive Science (IKW)
University of Osnabrück, Germany

Outline

Inferential statistics
  Preliminaries

One-sample tests
  Testing the mean
  Testing the variance
  Student's t test
  Confidence intervals

Outline

Inferential statistics
  Preliminaries

One-sample tests
  Testing the mean
  Testing the variance
  Student's t test
  Confidence intervals

Inferential statistics

- Goal: infer (characteristics of) population distribution from small random sample, or test hypotheses about population
  - in particular, we will estimate/test characteristics μ and σ
    ☞ only makes sense if we have a **parametric** model
- Nonparametric tests need fewer assumptions, but ...
  - cannot test hypotheses about μ and σ
    (instead: median, IQR = inter-quartile range, etc.)
  - more complicated and computationally expensive procedures
  - correct interpretation of results often difficult
- In this session, we assume a **Gaussian population** distribution
  - sometimes a scale transformation is necessary (e.g. lognormal)

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 Ω
  - 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 $|Z| \leq 1.96$

1. NB: we can only calculate such a z-score if we know the true mean μ and s.d. σ of the Gaussian distribution!

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. $X, Y, Z, \ldots$:

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

(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

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

- $X_1, \ldots, X_n$ 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

A simple test for the mean

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

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

  - ☞ 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**

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

- 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 μ because it is **unbiased**:

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

- 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

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

1. The first question is whether it makes sense to base our test for the mean on the sample average, i.e. whether $\bar{X}$ is a sensible test statistic for μ.
2. To answer this question, we need to work out (properties of) the sampling distribution of $\bar{X}$ across different random samples of the same size n.
3. A minimal requirement is *unbiasedness*: the test statistic should not systematically indicate a wrong value for μ (possibly after suitable rescaling: $\bar{X} - \mu$ would be just as sensible as a test statistic).
4. Unbiasedness is not sufficient to ensure that a test based on $\bar{X}$ is reliable. We also need to know the variability of $\bar{X}$, or preferably its complete sampling distribution.
5. An important advantage of Gaussian random variables is that $\bar{X}$ also follows a Gaussian distribution. There are only very few other distributions with this property (e.g. Cauchy).

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

$$X_1 + \dots + X_n \sim N(n\mu, n\sigma^2)$$
$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n) \sim N(\mu, \frac{\sigma^2}{n})$$

- $\bar{X}$ has Gaussian distribution with same μ 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 μ
- If the sample size n is large enough, $\sigma_{\bar{X}} = \sigma / \sqrt{n} \to 0$ and the sample mean $\bar{x}$ becomes an accurate estimate of the true population value μ (**law of large numbers**)

The z test

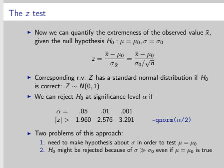
- Now we can quantify the extremeness of the observed value $\bar{x}$, given the null hypothesis $H_0 : \mu = \mu_0, \sigma = \sigma_0$

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

- 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 α if

$\alpha =$	.05	.01	.001	
$\lvert z\rvert >$	1.960	2.576	3.291	`-qnorm(`α`/2)`

- Two problems of this approach:
  1. need to make hypothesis about σ 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



1. Recall that sampling distribution of $\bar{X}$ describes how the sample average varies across different random samples of the same size n from the same population satisfying H_0 (think of many sociologists taking surveys of 100 random people each).
2. We reject H_0 if the observed value of $\bar{X}$ is sufficiently improbable according to the sampling distribution, i.e. unlikely to happen by chance if H_0 were true.
3. Note that for the binomial test in Unit 2, the sampling distribution of the passive count X was discrete. Now the test statistic $\bar{X}$ has a continuous distribution described by a density function.
4. R function `qnorm` and other functions for working with probability distributions will be explained in a moment.
5. If we overestimate $\sigma \ll \sigma_0$, the z test might fail to reject $\mu = \mu_0$ even though it clearly isn't true.

Outline

Inferential statistics
Preliminaries

One-sample tests
Testing the mean
Testing the variance
Student's t test
Confidence intervals

A test for the variance

- An intuitive test statistic for σ^2 is the sum of squares

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

- Squared error $(X - \mu)^2$ is σ^2 on average ➜ $\mathrm{E}[V] = n\sigma^2$
  - 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:

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

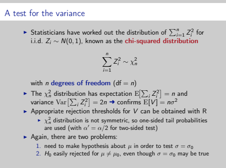
with $Z_i \sim N(0,1)$ i.i.d. standard normal variables

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 χ_n^2 distribution has expectation $\mathrm{E}[\sum_i Z_i^2] = n$ and variance $\mathrm{Var}[\sum_i Z_i^2] = 2n$ ➜ confirms $\mathrm{E}[V] = n\sigma^2$
- Appropriate rejection thresholds for V can be obtained with R
  - χ_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 μ 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



1. This variance test is even more sensitive to a violation of the assumption $\mu = \mu_0$ than the z test is against $\sigma = \sigma_0$.
2. We need to estimate true μ from the sample data!
3. But first, let us take a short intermission: This is a good moment for a hands-on session on working with distributions in R.

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:

`dnorm()`	density function of Gaussian (normal) distribution
`pnorm()`	tail probability
`qnorm()`	quantile = inverse tail probability
`rnorm()`	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)

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

pchisq(10, df=5, lower.tail=FALSE)  # tail prob. for ∑_i Z_i^2 ≥ 10

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

The sample variance

- Idea: replace true μ by sample value $\bar{X}$ (which is a r.v.!)

$$V' = (X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2$$

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

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

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

  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: $\mathrm{E}[S^2] = \sigma^2$
- We can use S^2 to test $H_0 : \sigma = \sigma_0$ without making any assumptions about the true mean μ ➜ **chi-squared test**

- Remarks
  - compare sample variance ($\frac{1}{n-1}$) with population variance ($\frac{1}{m}$)
  - χ^2 distribution doesn't have parameters σ^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

Chi-squared test of variance in R

```
> x <- Survey$height  # sample data: n items
> n <- length(x)

# Chi-squared test for a hypothesis about the s.d. (with unknown mean)
# H0 : σ = 13 (one-sided test against σ > σ0)
> sigma0 <- 13
> S2 <- sum((x - mean(x))^2) / (n-1)  # unbiased estimator of σ^2
> X2 <- (n-1) * S2 / sigma0^2     # has χ^2 distribution under H0
> pchisq(X2, df=n-1, lower.tail=FALSE)

# Here's a trick to carry out a 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

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 σ^2
- In the z-score for $\bar{X}$

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

replace the unknown true s.d. σ by the 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 distriution 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

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

Student's **t distribution** with $\mathrm{df} = n - 1$ 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 $\mathrm{df} = n - 1$ and desired significance level α (in R: `-qt(`$\alpha/2$`, `$n-1$`)`)
  - ☞ close to z-score thresholds for $n > 30$

1. TODO – mathematical explanation of the "miracle" that cancels out σ^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.)
# H0 : μ = 165 cm
> mu0 <- 165
> x.bar <- mean(x) # sample mean x̄
> s2 <- sum((x - x.bar)^2) / (n-1) # sample variance s^2
> s2 <- sd(x)^2 # easier with built-in function (check equality!)
> t.score <- (x.bar - mu0) / sqrt(s^2 / n) # t statistic
> print(t.score) # positive indicates μ > μ0, negative μ < μ0
> -qt(0.05/2, n-1) # two-sided rejection threshold for |t| at α = .05
# 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 μ!
```


Outline

Inferential statistics
Preliminaries

One-sample tests
Testing the mean
Testing the variance
Student's t test
Confidence intervals

Confidence intervals

- If we do not have a specific H_0 to start from, estimate **confidence interval** for μ or σ^2 by inverting hypothesis tests
  - in principle same procedure as for binomial confidence intervals
  - implemented in R for t test and chi-squared test
- For t test, confidence interval has a particularly simple form, so we can carry out the procedure by hand
  - we'll write $\hat{\mu} = \bar{x}$ here to emphasise its use as estimator
- Given $H_0 : \mu = a$ for some $a \in \mathbb{R}$, we reject H_0 if

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

  with $C \approx 2$ for $\alpha = .05$ and $n > 30$
- This leads to $\hat{\mu} - C \cdot s/\sqrt{n} \leq \mu \leq \hat{\mu} + C \cdot s/\sqrt{n}$
  - ☞ origin of the "± 2 standard deviations" rule

Switch to blackboard mode . . .