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]Corpus Linguistics Group
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Outline

Outline

Inferential statistics for continuous data

- Goal: infer (characteristics of) population distribution from small random sample, or test hypotheses about population
  - problem: overwhelmingly infinite coice of possible distributions
  - can estimate/test characteristics such as mean μ and s.d. σ
  - but H_0 doesn't determine a unique sampling distribution then
  - ☞ **parametric** model, where the population distribution of a r.v. X is completely determined by a small set of parameters

- In this session, we assume a **Gaussian population** distribution
  - estimate/test parameters μ and σ of this distribution
  - sometimes a scale transformation is necessary (e.g. lognormal)

- Nonparametric tests need fewer assumptions, but …
  - cannot test hypotheses about μ and σ
    (instead: median m, IQR = inter-quartile range, etc.)
  - more complicated and computationally expensive procedures
  - correct interpretation of results often difficult

Inferential statistics for continuous data

Rationale similar to binomial test for frequency data: measure observed **statistic** T in sample, which is compared against its **expected** value $\mathrm{E}_0[T]$ ➜ if difference is large enough, reject H_0

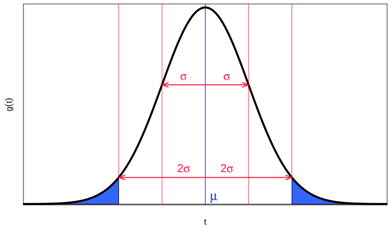
- Question 1: What is a suitable statistic?
  - depends on null hypothesis H_0
  - large difference $T - \mathrm{E}_0[T]$ should provide evidence against H_0
  - e.g. unbiased estimator for population parameter to be tested
- Question 2: what is "large enough"?
  - reject if difference is unlikely to arise by chance
  - need to compute sampling distribution of T under H_0

Inferential statistics for continuous data

- Easy if statistic T has a Gaussian distribution $T \sim N(\mu, \sigma^2)$
  - μ and σ^2 are determined by null hypothesis H_0
  - reject H_0 at two-sided significance level $\alpha = .05$ if $T < \mu - 1.96\sigma$ or $T > \mu + 1.96\sigma$
- This suggests a standardized **z-score** as a measure of extremeness:

$$Z := \frac{T - \mu}{\sigma}$$

- Central range of sampling variation: $|Z| \leq 1.96$



Notation for random samples

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

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

  ☞ specific items $\omega_1, \ldots, \omega_n$ are irrelevant, we are only interested in their properties $x_i = X(\omega_i)$, $y_i = Y(\omega_i)$, etc.
- 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
  ☞ each random experiment now yields complete sample of size n

---

Outline

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 $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are **independent**, then

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

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 mean 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

Outline

A test for the variance

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

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

- 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:

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

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

A test for the variance

- Note that the distribution of $Z_1^2 + \cdots + Z_n^2$ does not depend on the population parameters μ and σ^2 (unlike V)
- 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 $\mathrm{E}\left[\sum_i Z_i^2\right] = n$ and variance $\mathrm{Var}\left[\sum_i Z_i^2\right] = 2n$ ➜ confirms $\mathrm{E}[V] = n\sigma^2$

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/σ_0^2 can easily 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

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)

# 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(.025, df=5, lower.tail=FALSE)` # two-sided: $V / \sigma_0^2 > n$

`> qchisq(.025, df=5, lower.tail=TRUE)` # two-sided: $V / \sigma_0^2 < n$

The sample variance

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

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

- But there are two problems:
  - ☞ $X_i - \bar{X} \sim N(0, \sigma^2)$ not guaranteed because $\bar{X} \neq \mu$
  - ☞ terms are no longer i.i.d. because $\bar{X}$ depends on all X_i

The sample variance

- We can easily 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. $X_1, X_2 \sim N(\mu, \sigma^2)$
- Can also show that V' and $\bar{X}$ are independent
  - follows from independence of $X_1 - X_2$ and $X_1 + X_2$
  - this is only the case for independent Gaussian variables (Geary 1936, p. 178)

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
  - sample variance ($\frac{1}{n-1}$) vs. 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

---

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

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 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 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 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 α (in R: `-qt(α/2, n−1)`)
  - ☞ very 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 $V \sim \chi^2_k$ are independent r.v.

- $T \sim t_{n-1}$ under $H_0 : \mu = \mu_0$ because the unknown population variance σ^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 $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

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

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
- 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 $n > 30$

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

☞ this is the origin of the "± 2 standard deviations" rule of thumb

Confidence intervals

- Can you work out a similar confidence interval for σ^2?
- Test hypotheses $H_0 : \sigma^2 = a$ for different values $a > 0$
  - ☞ Which H_0 are rejected given the observed sample variance s^2?
- If H_0 is true, we have the sampling distribution

$$Z^2 := (n-1)S^2/a \sim \chi^2_{n-1}$$

- Reject H_0 if $Z^2 > C_1$ or $Z^2 < C_2$ (not symmetric)
- Solve inequalities to obtain confidence interval

$$(n-1)s^2/C_1 \leq \sigma^2 \leq (n-1)s^2/C_2$$