Statistical Methods in Natural Language Processing (NLP)

Class 6: Hypothesis Testing and Statistical Models

GÖTEBORGS UNIVERSITET

Charalambos (Haris) Themistocleous

Department of Philosophy, Linguistics and Theory of Science, Centre for Linguistic Theory and Studies in Probability

Introduction

- Normal Distribution
- Properties of Normal Distribution
- Hypothesis Testing
- t-tests
- Regression

Normal distribution

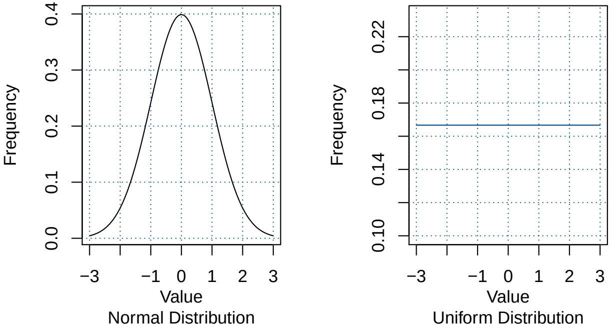


Figure: Uniform and normal distribution.

Central Limit Theorem

The basic idea is this: if there is a sufficiently large number of independent random variables, each with a finite value and finite variance, the distribution will approximate the normal distribution **whatever the underlying distribution is.**

Central Limit Theorem

Figure: Sir Francis Galton F.R.S. 1822-1911

"Order in Apparent Chaos.-I know of scarcely any-, thing so apt to impress the imagination as the wonderful form of cosmic order expressed by the " Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along."

Deviations from the normal distribution

1. Positive Distribution
2. Negative Distribution

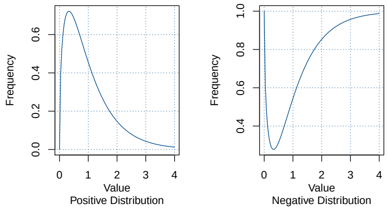


Figure: Asymmetric distributions.

Moments: Shape of the Distribution

1. *zeroth moment:* the total probability and it is equal to **1**.
2. *first moment:* **Mean**
3. *second moment:* **Variance** (standard deviation, standard error)
4. *third moment:* **Skewness**
5. *fourth moment:* **Kurtosis**

GÖTEBORGS UNIVERSITET

Measures of Central Tendency

The most common ones are

- **Mean** (trimmed mean)
- **Median**
- **Mode**

Measures of Dispersion

The most common ones are:

- Variance
- Standard deviation
- Standard Error
- Percentile
- Range, min max values
- Interquartile range

Kurtosis

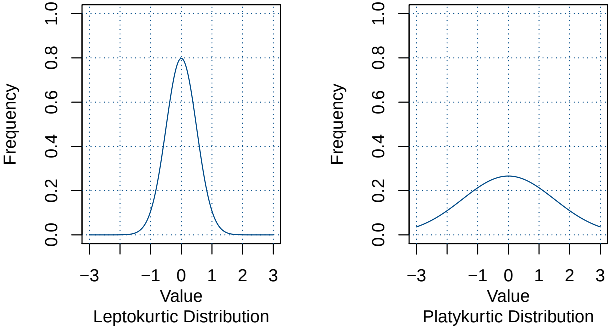


Figure: Leptokurtic distribution (left panel) and platykurtic distribution (right panel).

C. Distributions with one or more peaks

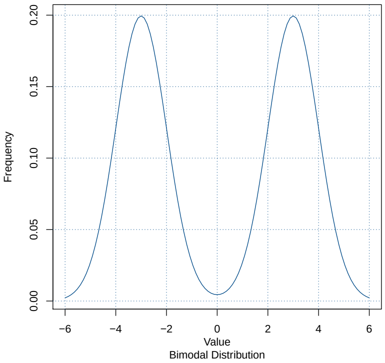


Figure: Bimodal distribution.

Calculating

Two functions related to the normal distribution is dnorm (), which produces normal probability density distribution and rnorm (), which produces random values to a normal distribution. For -1 standard deviation, the probability of the standard distribution or the cumulative probability is 0.16. In other words, to 16% of the values will be less than -1.

```
pnorm (−2)
[1] 0.02275013
Caution! The opposite of pnorm function () is qnorm ():
qnorm (0.02275013)
[1] −2
```

Standard normal distribution

The standard normal distribution is a normal distribution with **arithmetic mean**$=0$ and **standard deviation** $=1$.

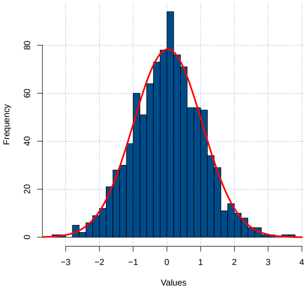


Figure: Standard normal distribution.

Standard Scores or z-scores

The z-score is calculated using the following:

$$z = \frac{x - \mu}{\sigma}$$

where: μ is the mean of the population. σ is the standard deviation of the population.

Using the z-score

- we can calculate the probability that a value is included in the normal distribution.
- we can compare two values that come from different normal distributions.

Models: Bad and non so Bad

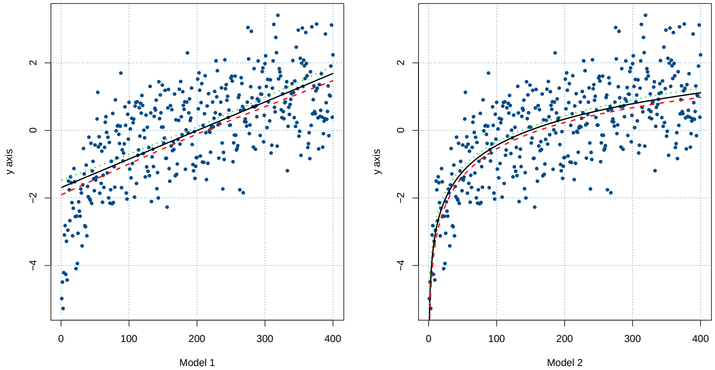


Figure: Model 1 and Model 2

GÖTEBORGS UNIVERSITET

Error

$$result_x = (model_x) + error_x$$

Error

$$model_x = result_x - error_x$$

Error

$$error_x = result_x - model_x$$

Null and alternative hypothesis

A. The **null hypothesis** states that there is no difference between the two results:

$$GroupA - GroupB = 0$$

B. The **alternative hypothesis** states that there is a significant difference between the two results.

$$GroupA - GroupB \neq 0$$

The research hypotheses guide the design of any experiment.

Confidence Intervals

Ronald Aylmer Fisher (1890–1962):
Trust the **95%**!

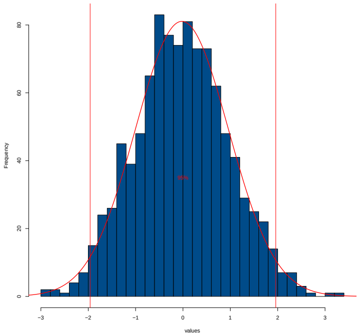


Figure: Confidence Intervals 95%

α level

α-level $=$ .05 or 95% of the distribution. To reject the null hypothesis, we need to estimate that a certain value appears with a probability, a.k.a. *p-value*, which is smaller than the α-level. When it is greater we need to accept the null hypothesis, that the two measurements come from the same distribution.

Statistical power

The b-level is .2 or 20% and shows result 1 b ' ie. 1-0.2 = 0.8 or 80%. If we find that the statistical power (statistical power) is 80% or more then we can feel secure that any results can be found really is. Since we know the level of a 'and level b ', then we can use the previous survey to calculate the size of the effect that we hope to find in the experiment. The most important of these is that with their help we can calculate the number of participants in the experiment.

GÖTEBORGS UNIVERSITET

Calculation error: Error type *a* and *b*

Measurement error is the difference between the value measured and the answers we receive.

	Correct zero hypothesis	False zero hypothesis
Accept null hypothesis	correct	**Type I error**
Rejection of the null hypothesis	**Type II error**	correct

Type I and Type II errors

1. In case of error type a ' (Type I error) believe that there is influence of the population and does not exist . According to the criterion of Fisher (see . To Section 14.5.5), the probability that the error is the level a ' . 0.5 (5%).
2. If we consider that there is no impact on the population and the fact is , that the error is called error type b (Type II error). The probability of a error ytou is .2% (20%). In any case we want to reduce the probability for this the wrong .

Systematic and non-systematic variation

Equally important is the error or deviation from the model to understand how good a statistical model. There are two types of variation:

1. the **systematic variation**: it occurs from the experimental modification, e.g., Drug vs. Placebo.
2. the **non-systematic variation**: physiological differences between patients or subjects. A doctor conducting an experiment might want the non-systematic variation to be as small as possible.

$$\frac{systematic\ variation}{non-systematic\ variation}$$

Effect Size

The effect size r measures the strength of a test. The Cohen (1992) suggests the following scale of effect sizes:

Effect size	R
No result	0
little effect	0.10
medium effect	0.30
big effect	0.50
perfect result	1

Table: Effect size

Students t-Test

Figure: William Sealy Gosset a.k.a., Student (1876-1937).

Student is the pen name of William Sealy Gosset (1876-1937). Gosset worked as a chemist at the Guiness Brewing Company (Arthur Guinness & Son) in Dublin, Ireland. The Company Guiness had banned employees from publishing their work because earlier another employee had published corporate secrets. Gosset requested permission to publish his research, which he argued was not a threat to the company. Eventually, he was given permission to publish his work. Nevertheless, because Guiness did not want to encourage other employees from doing the same, asked Gosset to use a pseudonym. Despite this, other chemists in the company followed the Gosset's example and published the work using nicknames such as 'Mathetes' (which means Student in Greek) and 'Sophister' (see., Hotelling, 1930).

$t-test$

1. Independent $t - test$
2. Dependent $t - test$

One sample $t - test$

Answer to the question: Is a specified value equal to the sample mean?

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where

- μ_0: specified value.
- $\overline{x}$: is the sample mean or the observed value.
- s: is the sample standard deviation.
- n: is the sample size..

Independent t-test

$$t = \frac{Mean\ of\ observed\ values - Mean\ of\ expected\ values}{Standard\ Error\ (SE)\ of\ the\ difference\ between\ the\ two\ means}$$

Independent t-test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{2/n}}$$

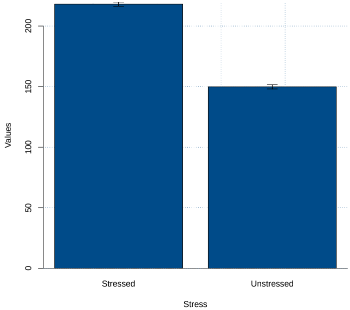
here s_p is called pooled standard deviation and is calculated as follows:

$$s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$$

GÖTEBORGS UNIVERSITET

A Simple Problem

- **Null Hypothesis**: The stressed and unstressed syllables will have the same duration.
- **Alternative Hypothesis**: The stressed and unstressed syllables will have significantly different duration.



Independent T-test

```
> t.test(Duration ~ Stress, paired=TRUE)

        Paired t-test

data:  Duration by Stress
t = -9.0734, df = 149, p-value = 6.33e-16
alternative hypothesis: true difference in means is not
95 percent confidence interval:
 -60.77536 -39.03797
sample estimates:
mean of the differences
              -49.90667
```

Scatterplots

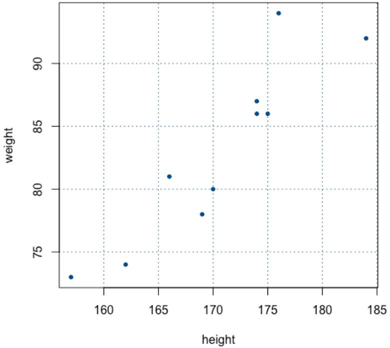


Figure: Weights and Heights

Scatterplots

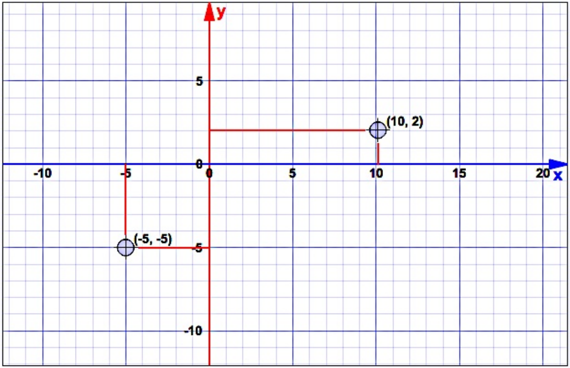


Figure:

Covariance

$$cov(x, y) = \frac{\sum (x - \bar{x}) \cdot (y - \bar{y})}{n - 1}$$

To expand the parentheses part

$$(x - \bar{x}) \cdot (y - \bar{y}) = xy - x\bar{y} - \bar{x}y + \bar{x}\bar{y}$$

An R function

An example of a simple function

```
covariance <- function(x,y)
{
if(length(x)!=length(y))
stop ("The two variables must have the same lenght")
sum((x-mean(x))*(y-mean(y)))/(length(x)-1)
}
```

or simply use the built in function cov(x,y)!

Problem with Covariance

```
x <- c(1,2,3,4)
y <- c(3,3,4,3)
cov(x,y)
[1] 0.1666667
```

but say we have the covariance:

```
x <- c(1,2,3,4)*100
y <- c(3,3,4,3)*100
cov(x,y)
[1] 1666.667
```

To standardize these measurements we use correlation.

Correlation

$$r = \frac{cov(x, y)}{\sqrt{{s^2}_x \cdot {s^2}_y}}$$

High Correlation vs. Low Correlation

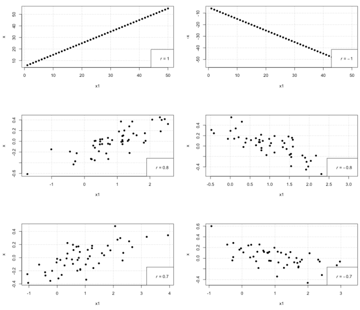


Figure: Correlation $r = 1, .8, .7$

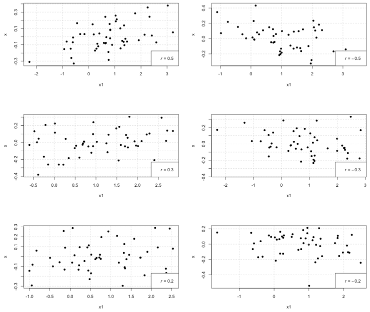


Figure: Correlation $r = .5, .3, .2$

Regression

- Response Variables:
  - This is what you measure.
  - It goes to the ordinate (the y axis of the graph)
- Explanatory Variables
  - These are all the conditions that explain the variation that occurs.
  - This goes on the abscissa (the x axis of the graph)

Conducting Linear Regression

$$y = (a + bx) + \epsilon$$

1. y is the response variable.
2. a is the intercept (when $y = 0$)
3. b is the slope/gradient $b = \frac{y}{x}$
4. ϵ is the error.

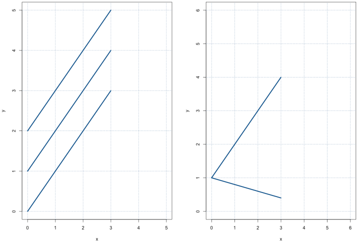


Figure: Regression Line Coefficients

GÖTEBORGS UNIVERSITET

Linear Regression

The mean describes the data using a single point. Another way to describe the data is using a line. Since many possible lines can describe the data, the regression analysis aims to find the best one that can model the data.

Foreign Language Learning Period and New Words

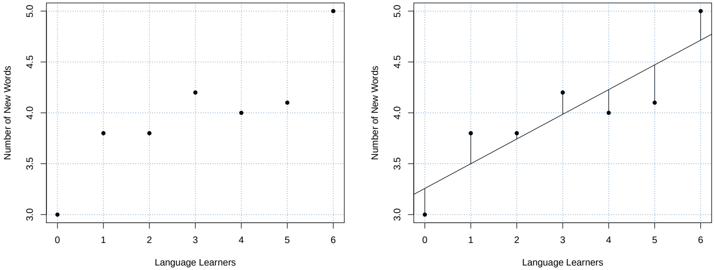


Figure: Regression Line Coefficients

Understanding the Regression Line

- The regression line provides a model of the data.
- The vertical lines indicate the differences of the regression line from the actual data. These are the residuals of the model.
- Residuals $\sum(y - \hat{y}) = 0$
- Some of the data are on the regression line. So, the line models these data perfectly.
- Most data are either over the line so the line underestimates them whereas other data are under the line so the line overestimates them.

GÖTEBORGS UNIVERSITET

Calculating the Intercept

$y = a + bx$
$y - bx = a + bx - bx$
$y - bx = a$
$\sum(y - \hat{y}) = 0$ and $\sum(y - a - bx) = 0$ both sum up to zero. To avoid this we get the square of these differences:

$$d2 = \sum(y - a - bx)^2$$

Solving the equation*

- The sum of x (Σx)
- The sum of y (Σy)
- The sum of the squares of x (Σx^2)
- The sum of squares of y (Σy^2)
- The sum of the product of x and y (Σxy)

Table: Estimating Regression Parameters

Variable	Variable in R	Formula	Result
Σx	Sx	sum(x)	21
Σy	Sy	sum(y)	27.9
Σx^2	Sxx	sum(x^2)	91
Σy^2	Syy	sum(y^2)	113.33
Σxy	Sxy	sum(x*y)	90.5

Calculating the corrected sums of the squares of $x, y, x * y$

$$SSX = \sum x^2 - \frac{\sum x^2}{n}$$

$$SSY = \sum y^2 - \frac{\sum y^2}{n}$$

$$SSXY = \sum xy - \frac{\sum x \sum y}{n}$$

Table: Estimating Regression Parameters

Variable R	Type Result	
SSX	sum(x^2)-sum(x)^2/length(x)	28
SSY	sum(y^2)-sum(y)^2/length(y)	2.128571
SSXY	sum(x*y)-sum(x)*sum(y)/length(x)	6.8
b	SSXY/SSX	0.2428571
a	sum(y)/length(y)-b*sum(x)/length(x)	3.257143

Calculating the error

Table: Estimating Regression Parameters

	Sum of Squares	Degrees of Freedom	Mean Squares	F ratio
Regression	SSR	df	SSR/df	$\text{F} = \text{df}/s^2$
Error	SSE	$n-2$	$s^2 = SSE/(n-2)$	
Total	SSY	$n-1$	–	

Calculating b. Maximum likelihood

$$b = \frac{SSXY}{SSX}$$

What is the variation we explain?

The variation we can explain is the regression sum of squares, which is:

$$SSR = \frac{SSXY^2}{SSX}$$

$SSR = b * SSXY = 0.2428571 \times 6.8 = 1.651428$ whereas the error sum of squares is the variation we cannot explain:

$$SSE = SSY - SSR \ or \ \sum (y - a - bx)^2$$

$SSE = SSY - SSR = 2.128571 - 1.651428 = 0.477143$

Calculating the error

	Sum of Squares	Degrees of Freedom	Mean Squares	F ratio
Regression	$SSR = 1.65$	df	SSR/df	$\text{F} = \text{df}/s^2$
Error	$SSE = 0.48$	$n - 2$	$s^2 = SSE/(n-2)$	
Total	SSY=SSR+SSE= 2.13	$n - 1$	–	

Degrees of freedom

1. SSR. In simple regression that we have only 1 parameter to estimate, i.e., the *b*
2. SSE. To calculate the SSE $(\sum(y-a-bx)^2)$ we need to estimate x and y as a and b are already in the data, therefore the df = n-2.
3. SSY. Since we need to estimate only y in $SSY = \sum(y-\bar{y})^2$, i.e., only one parameter $(\bar{y})$ the d.f. is n-1.

Calculating the error

Table: Estimating Regression Parameters

	Sum of Squares	Degrees of Freedom	Mean Squares	F ratio
Regression	$SSR = 1.65$	1	SSR/df	$\mathrm{F} = \mathrm{df}/s^2$
Error	$SSE = 0.48$	5	$s^2 = SSE/(n-2)$	
Total	SSY=SSR+SSE= 2.13	6	–	

Calculating the error

Table: Estimating Regression Parameters

	Sum of Squares	Degrees of Freedom	Mean Squares	F ratio
Regression	$SSR = 1.65$	1	SSR/df=1.65	F $=$ df/s^2=1.73
Error	$SSE = 0.48$	5	$s^2 = SSE/(n-2)$=0.96	
Total	SSY=SSR+SSE= 2.13	6	–	

How well does the regression perform?

- The null hypothesis is that the slope $= 0$ in both cases.
- The alternative hypothesis is that the slope is greater or lower.
- So, that the there is no relationship between the two datasets.
- To estimate that we evaluate the F-ratio with the critical value of F.
- The critical value is evaluated from quantiles of F distribution.
- Traditionally, people used tables.
- R provides these estimates automatically, using the qf() function, with 1 degree of freedom in the nominator and n-1 degrees of freedom in the denominator.

How well does the regression perform?

- In our data, the critical value of the F ratio is the value of F, which will only arise by chance if the null hypothesis were true, if we had 1 degree of freedom in nominator and 5 degrees of freedom in the denominator.
- We need also define the probability that we will accept the null hypothesis: 95% or 0.95 and the alpha value is 5% or 0.5 that we will reject it.
- qf(0.95,1,5)
- [1] 6.607891

GÖTEBORGS UNIVERSITET

How well does the regression perform?

- Now we see that the F value which we found to be 17.31 is greater that the critical value.
- Usually, we estimate the probability to get a F equal to 17.31 or greater if the null hypothesis is true. To this purpose we use the 1-pf() instead of qf, which provides the so called p value.
- 1-pf(17.31,1,5)
- [1] 0.008819561
- If the p-value is lower than the .05 we need to reject the null hypothesis and accept the alternative hypothesis.

Estimating the standard errors of the slope and the intercept

- We assumed that SSY = SSR + SSE.
- To estimate the standard errors of the slope and the intercept, we start from the $s^2 = 0.0954$.

Estimating the standard errors of the slope and the intercept

So, the Standard Error of the slope b is estimated as follows:

$$se_b = \sqrt{\frac{s^2}{SSX}} = \frac{0.0954}{28} = 0.583 \tag{1}$$

The Standard Error of the intercept a is:

$$se_a = \sqrt{\frac{s^2 \cdot \sum x^2}{n \cdot SSX}} = \sqrt{\frac{0.0954 \cdot 91}{7 \cdot 28}} = 0.2105 \tag{2}$$

Conducting and Interpreting Regression Analysis in R

```
model.1 <- lm(y ~ x)
summary(model.1)

Call:
lm(formula = y ~ x)
Residuals:
       1        2        3        4        5        6        7
-0.25714  0.30000  0.05714  0.21429 -0.22857 -0.37143  0.28571

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.25714    0.21049   15.47 2.05e-05 ***
x            0.24286    0.05838    4.16  0.00882 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3089 on 5 degrees of freedom
Multiple R-squared: 0.7758,^^IAdjusted R-squared: 0.731
F-statistic: 17.31 on 1 and 5 DF,  p-value: 0.008824
```

```
  summary.aov(model.1)
            Df Sum Sq Mean Sq F value   Pr(>F)
x            1 1.6514  1.6514    17.3 0.00882 **
Residuals    5 0.4771  0.0954
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The table shows the Variation Error (s2=0.0954), the SSR (SSR=1.6514), the SSE (SSE=0.4771) and the p value which we have calculated using 1-pf.

Next Class

- Information Theory
- Entropy