Introduction to statistical methods for language and literature

Diana Santos

ILOS

d.s.m.santos@ilos.uio.no

Spring 2016, week 2

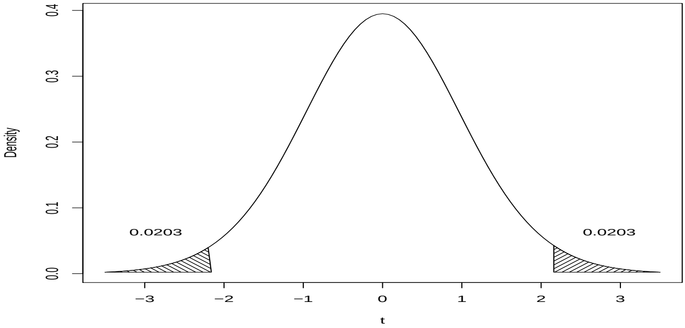
Statistics as a topic

- Empirical observation led to probability distributions being discovered.
- These distributions, which you will meet now, are mathematical functions which can describe (some) aspects of reality.
- R has these built in, and in fact has a family of commands for each distribution. And one should also talk about a family of distributions (one for each parameter).
- Statistical methods, or statistics, is the art of using these distributions on practical problems, by using STATISTICS, which are numbers obtained from your material.

Statistics uses a branch of mathematics called “probability theory”.

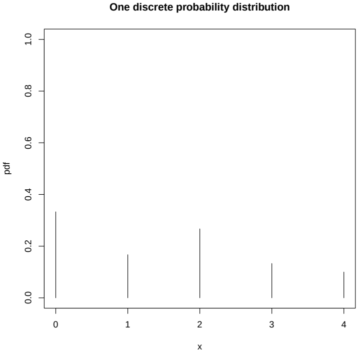
How to use a probability distribution

The area (or the sum) gives us the probability of the events/facts/objects we got.
Less or equal than x, or in the interval $]x_1, x_2[$.



How to use a discrete probability distribution

The sum gives us the probability of the events/facts/objects we got.



x	0	1	2	3	4
$p(x)$	.333	.167	.267	.133	.1

Commands in R

Generally

```
dNAVN(x,PARAMETRE)
qNAVN(quantil,PARAMETRE)
pNAVN(x,PARAMETRE)
rNAVN(antall,PARAMETRE)
```

```
dbinom(x,n,p)
qbinom(q,n,p)
pbinom(x,n,p)
rbinom(antall,n,p)

dpois(x,lambda)
qpois(q,lambda)
ppois(x,lambda)
rpois(antall,lambda)
```

Other families: NORM, CHISQ, T, F, MULTINOM ...

The Poisson distribution

is used to describe how (integer) quantities are distributed in time or space, when the events are independent of each other.

- how many students in a class per day
- how many cases of a disease in one month
- How many defects in a box with 100 products
- how many cars in a bridge per day
- how many cases of a word in texts

The cases where the Poisson distribution can be used are cases where we can count positive instances, but do not know how to count negative instances. For example: falling from horses, suicide cases.

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

The Poisson distribution

Try out the Poisson distribution in R. See the figures for different λ.

Visualization with R

One of the big pluses of R is its visualization capabilities. In addition, R can find most of the figure parameters alone.

```
plot()
boxplot()
hist()
assocplot()
barplot()
```

Other distributions

F, χ^2, t, exponential, γ, Weibull, lognormal, uniform, beta; binomial, multinomial, hypergeometric, negative binomial.
The nine first are continuous, the four last are discrete.

Parametric methods

Parametric methods use known probability distributions, where it is enough to know the parameters. (Families of distributions because one has one distribution per value of parameter.)

Distribution	Parameters
Binomial	p
Poisson	λ
Gaussian	μ, σ
t	ν - degrees of freedom
F	ν_1, ν_2
χ^2	ν - degrees of freedom

There are other ways of doing statistics, which do not require that one knows the probability density function: the non-parametric methods.

Exercises for next class

- Choose two numbers as parameters for Poisson and create the figure of the corresponding Poisson distribution.
- Draw also the normal distribution with variance and average equal to the above numbers.
- Find the probability, for each of the four cases, of $x <= number/2$

Write the commands you used in R as a text file, and put them in Fronter.

But we have just some numbers...

The first thing one should understand, is that there are concepts to describe a population (a theoretical distribution), and concepts to describe a sample (an empirical distribution).

Central tendency (Sentralitet) or location Average (gjennomsnitt), median, mode: when one can only use one number to describe many points

Dispersion or variability Variance, standard deviation (standardavvikk), range(variasjonsbredde), interquartile interval (interkvartil-intervall), variance coeffient (variasjonskoeffisient)

Bias (forventningsretthet) Measures (positive or negative asymmetry): the curve goes more to the right or to the left?

Kurtosis Measures the peakedness (spissing) of a distribution, compared with the normal distribution

We have just some numbers...

- In inferential statistics we try to find the “real” (theoretical, population) values by measuring some observations (empirical, sample) which give us **estimates**.
- **Estimators** is what we use to try to reach/guess the underlying parameters, which are the real numbers we are interested in.
- (To find good estimators, there are two (mathematical) methods: the method of moments, and maximum likelihood estimation (MLE).)

Which numbers? I

There are concepts to describe a population, and concepts to describe a sample. Starting with our samples:

sample mean, $\bar{x}$ $\frac{\sum_{i=1}^{n} x_i}{n}$

median the $\frac{n+1}{2}$ th ordered value, or the average of the two $\frac{n}{2}$ and $\frac{n}{2}+1$ ordered values

sample variance, s^2 $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

sample standard deviation, s $s=\sqrt{s^2}$

Exercise for home: compute all these statistics “by hand” for the Portuguese summer (but you can/should use `sum()`, `sort()` for vectors, and `sqrt()`)

Populations: Which numbers? I

expected value or mean, $E(X)$ or μ_x $\sum_{x \in D} x \times p(x)$

variance, $V(X)$ or σ^2 $\sum_{x \in D}(x - \mu)^2 \times p(x) = E[(X - \mu)^2]$

standard deviation, $SD(X)$ or σ $\sigma = \sqrt{\sigma^2}$

skewness $\frac{E[(X-\mu)^3]}{\sigma^2}$

Actually, $V(X) = E[(X^2)] - [E(X)]^2$
For known probability distributions we (mathematicians) know how to compute these numbers. Or, these numbers are related to their parameters.
Examples:

- for the binomial: $E(X) = np; V(X) = np(1 - p); \sigma_X = sqrt(npq)$
- for the Poisson: $E(X) = V(X) = \lambda$
- for the normal: μ and σ are parameters

The Gaussian (or normal) distribution

The most important distribution (for reasons that will be explained later on) is $N(\mu, \sigma)$.

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/(2\sigma^2)}$$

e: natural logarithm base ca. 2.71828
What is commonly used is the "standard normal distribution", with zero mean and variance 1, which is often called $Z(0, 1)$.

$$Z(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

F-distribution

(after Fisher)

- F-distribution is used to compare variances. The estimator is a variance ratio.
- It has two parameters: degrees of freedom in the numerator, and in the denominator.
- It measures homogeneity.
- It can be used to compare two samples of the same population, or two populations. It is used in connection with the conditions for applying ANOVA.

χ^2 distribution

The probability density of the χ^2 function is, for $x > 0$:

$$\frac{1}{2^{\mu/2}\Gamma(\mu/2)} x^{\mu/2-1} e^{-x/2}$$

And this is the empirical statistic one collects from the data:

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

It is used in three different situations.

- To assess how correct an expectation is ("goodness of fit")
- To check whether two variables in a contingency table are independent or not
- To decide whether a population is homogeneous regarding a particular variable

To use χ^2 the variables should be normal, the measures should be independent, and all expected values should be greater than 5.

Appetizer: Test independence

If we have two properties that were measured for the same subjects/objects, we can try to assess whether they are independent or not.

- hair colour and eye colour
- gender and use of adjectives
- smoking and cancer

```
chisq.test(TABELL)
```

A little more about the χ-square

- It is related to the normal distribution, because the square of a normal variable has a chi-square distribution (χ^2) with one degree of freedom.
- Also the sum of independent chi-square variables (with respectively ν_1 and ν_2 degrees of freedom) is also chi-square distributed with number of degrees of freedom $\nu = \nu_1 + \nu_2$.
- Therefore, also the sum of n independent chi-square variables also chi-squared distributed, with $\nu = n$ degrees of freedom.

Behind the scenes

For mathematicians, Devore and Berk (2007) say that

> *the chi-squared, t, and F distributions are "distributions based on a normal random sample"*

- for the distribution of the sample variance, one needs the distribution of sums of squares of normal variables -> the χ^2 distribution
- to use the sample standard deviation in a measure of precision for the mean $\overline{X}$, we need a distribution that combines the square root of a chi-squared variable with a normal variable -> the t distribution
- to compare two independent sample variances, we need the distribution of the ratio of two independent chi-squared variables -> the F distribution

Back to normal?

This (also called Gaussian) is the best-known distribution of all. It is especially important because of the **Law of the big numbers**:

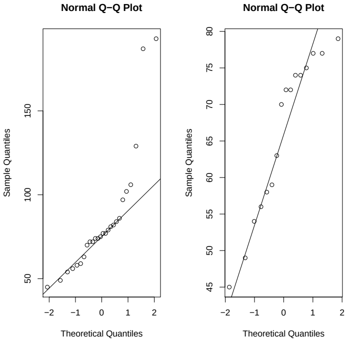
> *Even if one population is not normal (ly distributed), the samples averages are normally distributed.*

But note that the normal distribution cannot be used for words. There we have a LNRE case "large numbers of rare events": each word is so rare that most of the cases it does not occur.
In text, the averages increase with larger samples, so we cannot do expect an average distribution.

Testing normality...

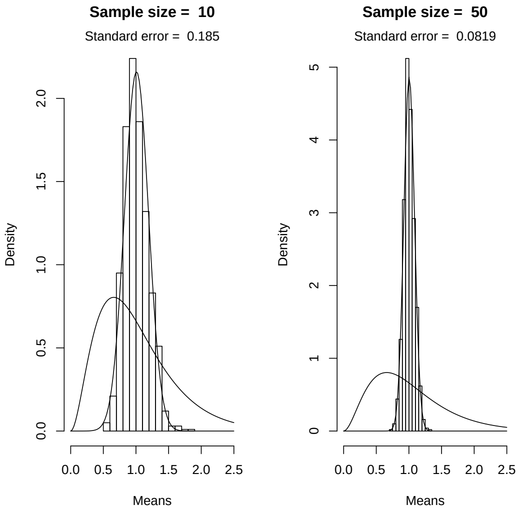
How can one test whether the phenomenon one is interested in is normal? Quantile test: compare the (theoretical) normal distribution with our data:

```
par(mfrow=c(1,2))
qqnorm(vot01)
qqline(vot01)
qqnorm(vot01[vot01<80])
qqline(vot01[vot01<80])
```



Central limit teorem

```
par(mfrow=c(1,2))
source("central.limit")
central.limit(10)
central.limit(50)
```



What does the Normal Curve “Mean”?

Julian Simon:

> *The most important characteristic of the Normal Distribution is that its occurrence is entirely caused by the researcher, and its appearance “means” that the researcher may consider that his research work is complete.*

(Julian Simon, 1968:436) (the same argument applies to the Poisson and the logarithmic-normal distributions.)

Student's t

It is used to test an hypothesis about means of a normal population, based on a small sample (so that the sample variance of the sample is not near the population variance).

The t statistic:

$$t = \frac{\overline{x}-\mu}{\frac{s_{\overline{x}}}{\sqrt{n}}}$$

It is also the test to compare the means of two independent samples with the same variance, the test statistic being the difference of the means in terms of the number of standard errors by which the two sample means are separated.

$$t = \frac{\overline{y_A}-\overline{y_B}}{\sqrt{\frac{{s_A}^2}{n_A}+\frac{{s_B}^2}{n_B}}}$$

Examples of use of a t-test

Differences between means: is the mean in our sample just like the known mean (12)?

```
t.test(HEDGES, mu=12)
```

Are F1 frequencies of men and women different?

```
t.test(F1S~gender, paired=FALSE)
t.test(F1S[gender=="M"], F1S[gender=="F"], paired=FALSE)
```

Are the length differences in translation consistently higher?

```
t.test(Length~OrigOrTrans, paired=TRUE)
```

Summing up

- You met a lot of probability and statistical actors for the first time.
- You got the distinction between sample and population
- You learned to compute some (descriptive) statistics from your sample.
- You got the idea that statistical science/engineering will be based on tests