## INF5830-2015 FALL NATURAL LANGUAGE PROCESSING

Jan Tore Lønning, Lecture 5, 14.9

## Today - more statistics

$\square$ Sampling distributions
$\square$ Normal distributions
$\square$ The effect of sample size
$\square$ Hypothesis testing
$\square$ Estimation
$\square$ With known standard deviation
$\square$ With unknown standard deviation

## Sampling distribution

Utvalgsfordeling

## Sampling - empirically

## Goal:

$\square$ make assertions about a whole population
$\square$ from observations of a sample (utvalg)
$\square$ A simple random sample (SRS) (tilfeldig utvalg):

1. Each individual has equal chance of being chosen (unbiased/forventningsrett)
2. Selection of the various individuals are independent

## Sampling distributions - Example

- Height: X
- assume $N(180,6)$
- (Var=36)
- Randomly choose 100.
$\square$ Add their heights:

$$
S=X_{1}+X_{2}+\ldots+X_{n}
$$

$\square$ A new random variable (all such samples)
$\square \operatorname{Exp}(S)=n^{*} \mu=18000(\mathrm{~cm})$
$\square \operatorname{Var}(S)=100^{*} \operatorname{Var}(X)=3600$

- $\sigma_{S}=10 \times \sigma_{X}=60(\mathrm{~cm})$



## Sampling distributions - Example

- Height: X
- assume $N(180,6)$
- (Var=36)
- Randomly choose 100.
$\square$ Add their heights: $\mathrm{S}=\mathrm{X}_{1}+\mathrm{X}_{2}+\ldots+\mathrm{X}_{\mathrm{n}}$
$\square$ A new random variable (all such samples)
- $\operatorname{Exp}(S)=n * \mu=18000(c m)$
- $\operatorname{Var}(\mathrm{S})=100 * \operatorname{Var}(\mathrm{X})=3600$
- $\sigma_{S}=10 \times \sigma_{X}=60(\mathrm{~cm})$
$\square$ The mean of the samples:
$\square \overline{\mathrm{X}}=\mathrm{S} / \mathrm{n}$
- A new random variable (all such means of samples of 100)
$\square \operatorname{Exp}(\bar{X})=\mu=180(\mathrm{~cm})$
$\square \sigma_{\bar{X}}=\frac{1}{100} \times \sigma_{S}=0.6(\mathrm{~cm})$


## Sampling distributions

## Let

$\square X$ be a random variable for a population with exp: $\mu$, std: $\sigma$
$\square$ Let $S=X_{1}+X_{2}+\ldots+X_{n}$ i.e. each $X_{i}$ equals $X$
$\square$ Let: $\bar{X}=S / n$

## Then:

$\square \operatorname{Exp}(S)=n^{*} \mu$
$\square \operatorname{Exp}(\bar{X})=\mu$
$\operatorname{Var}(S)=\sigma_{S}^{2}=n \times \operatorname{Var}(X)=n \times \sigma_{X}^{2}$
$\operatorname{Var}(\bar{X})=\sigma_{\bar{X}}^{2}=\frac{1}{n^{2}} \times \operatorname{Var}(S)=\frac{1}{n} \times \sigma_{X}^{2}$
ㅁ

$$
\sigma_{\bar{X}}=\frac{1}{\sqrt{n}} \times \sigma_{X}
$$

## The form of the distribution

$\square$ If the Xi -s are independent and normally distributed, then $\bar{X}$ is normally distributed (as expected)
$\square$ (More surprisingly) Even though the Xi -s themselves are not normally distributed: for large $n-s, \bar{X}$ is approximately normally distributed
$\square=$ Central Limit Theorem

## Example: throwing the dice until a 6

## Number of samples: 1000





$E(\bar{X})=E(X)=\mu=6$
$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}=\frac{\sqrt{6 \times 5}}{\sqrt{n}}$

4


Trivia question: How many dice throws,roughly?

## So what?

$\square$ This means that
$\square$ Given a population P with
$\square$ a known mean and
$\square$ known standard deviation

- and we meet a set of objects $S$ of some size,
$\square$ We can say something about how likely/unlikely it is that this set $S$ is a simple random sample from $P$


## Normal distribution

$\square$ We will take a closer look on the normal distribution for one individual before considering samples.

## Example height (contd.)

$\square$ Tallness of Norwegian young men (rough numbers):
$\square$ Normal distribution
$\square \mu=180 \mathrm{~cm}$
$\square \sigma=6 \mathrm{~cm}$
$\square$ How many are taller than 190cm?
$\square$ First calculate the z-score
(how many standard deviations is this?)
$\square Z=\frac{x-\mu}{\sigma}=\frac{190-180}{6}=1.67$

$\square$ Use software, calculator or table to find the corresponding probability

- Here 0.0475


## 68\%-95\%-99.7\%



## Look up

$\square$ Statistical table

- course.shufe.edu.cn/ipkc/irilx/ref/StaTable.pdf
$\square$ SciPy
ㅁ >>>import scipy

- >>> from scipy import stats
- >>> stats.norm.cdf(10/6)
- 0.9522096477271853
- >>> 1-stats.norm.cdf(10/6)
- 0.047790352272814696
$\square \ggg$ stats.norm.cdf(190,180,6)
- 0.9522096477271853


## Table

$\square$ Given probability p , for which h is $\mathrm{P}(\mathrm{x}>\mathrm{h})<\mathrm{p}$ ?
$\square$ Standardize, calculate the Z-score: $z=\frac{x-\mu}{\sigma}$
$\square P(x>h)=P\left(\frac{x-\mu}{\sigma}>\frac{h-\mu}{\sigma}\right)=P\left(z>\frac{h-\mu}{\sigma}\right)$
$\square$ Use table or software to look up $z$
$\square \mathrm{h}=\sigma \mathrm{z}+\mu$

| Probability <br> p-value | Z-score | centimeters | height |
| :--- | :--- | :--- | :--- |
| 0.1 | 1.28 | 7.68 | 187.68 |
| 0.05 | 1.645 | 9.87 | 189.87 |
| 0.01 | 2.326 | 14 | 194 |
| 0.001 | 3.091 | 18.5 | 198.5 |

## Look up

$\square$ Statistical table
口 course.shufe.edu.cn/ipkc/irilx/ref/StaTable.pdf
$\square$ SciPy

- >>> stats.norm.ppf(.999)

- array(3.0902323061678132)


# The effect of the sample size 

## Relationship $\mathrm{h}, \mathrm{n}$ for $\mathrm{P}(\overline{\mathrm{x}}<\mathrm{h})<0.01$

| sample size | $\sigma /$ sqri(n) | z-value | centimeters | height |
| :--- | :--- | :--- | :--- | :--- |
| 1 | 6 | -2.326 | 14 | 194 |
| 4 | 3 |  | 7 | 187 |
| 36 | 1 |  | 2.3 | 182.3 |
| 100 | $6 / 10$ |  | 1.4 | 181.4 |
| 1000 | 0.19 |  | 0.42 | 180.42 |

## Relationship h, p when $\mathrm{n}=100$

| probability | z-value | centimeters | height |
| :--- | :--- | :--- | :--- |
| 0.1 | 1.28 | 0.768 | 180.768 |
| 0.05 | 1.645 | 0.987 | 180.987 |
| 0.01 | 2.326 | 1.4 | 181.4 |
| 0.001 | 3.091 | 1.85 | 181.85 |

## Hypothesis testing

$\square$ A population P2
$\square$ Could be:
■ Norw. males 50ys olds in 2007

- Norw. females $18 y s$ olds in 2007

■ Norw. males $18 y s$ olds in 1957

- Swe. males 18 ys olds in 2007
- Etc.
$\square$ Are the individuals in P2 shorter than they in P?
$\square$ Pick a random sample $\left\{x_{1}, x_{2}, \ldots, x_{n}\right\}$ from P2
$\square$ Null hypothesis, $H_{0}: \mu_{\mathrm{P} 2}=\mu$
$\square$ Hypothesis, $\mathrm{H}_{\mathrm{a}}: \mu_{\mathrm{P} 2}<\mu$
$\square$ We formulate the question: What is the chance $\left\{x_{1}, x_{2}\right.$, $\left.\ldots, x_{n}\right\}$ could have been a s.r.s. from $P$ ?
$\square$ For example,
- If we take a SRS from P2 of $\mathrm{n}=100$ individuals,
$\square$ and we find $\bar{x}=178.5$,
$\square$ we can conclude there is less than 0.01 chance that $\left\{x_{1}, x_{2}\right.$, $\left.\ldots, x_{n}\right\}$ is a s.r.s. from $P$
$\square$ In other words, if P and P2 had been equal (w.r.t. height), there is less than $1 \%$ chance that we would have chosen such a SRS


## Conclusion, jargon

$\square$ In other words, if P and P2 had been equal (w.r.t. height), there is less than $1 \%$ chance that we would have chosen such a SRS
$\square$ The p -value is less than 0.01
$\square$ The hypothesis $\mathrm{H}_{\mathrm{a}}$ is significant at level 0.01
$\square$ and 0.05
$\square$ But not 0.001

## Recipe

$\square$ Formulate $\mathrm{H}_{\mathrm{a}}$ and $\mathrm{H}_{0}$
$\square$ Sample an appropriate SRS and find its mean value, $\bar{X}$
$\square$ Calculate the Z-score: $Z=\frac{x-\mu}{\sigma}$
$\square H_{a}: \mu_{\mathrm{P} 2}<\mu$ is $\mathrm{P}(\mathrm{X}<\mathrm{z})$
$\square>$ similarly:
$\square \mathrm{H}_{\mathrm{a}}: \mu_{\mathrm{P} 2}=/=\mu$ is $2 \times \mathrm{P}(\mathrm{X}>|\mathrm{z}|)$

## Remarks

|  |  | Truth |  |
| :--- | :--- | :--- | :--- |
|  |  | H0 | Ha |
| Decision | Not rejecting <br> HO |  | Type II error |
|  | Reject HO | Type I error <br> Prob. p-value |  |

$\square$ There is a chance of probability $p$ that we erroneously reject HO (Type I error)
$\square$ The test does not estimate type II error
$\square$ Says nothing about how much the difference is between P2 and $P$
$\square$ Many possible banana skins: E.g. is the sample really random?

## Population estimation

$\square$ With a known mean $\mu$ :
we consider $\mathrm{P}(\overline{\mathrm{x}}<\mu-\mathrm{e})$ for means of samples.
$\square$ If we do not know the true mean $\mu$, we see that
$\square \mathrm{P}(\mu>\overline{\mathrm{x}}+e)=\mathrm{P}(\bar{x}<\mu-e)$, and
$\square \mathrm{P}(\overline{\mathrm{x}}-\mathrm{e}<\mu<\overline{\mathrm{x}}+\mathrm{e})=\mathrm{P}(\mu-\mathrm{e}<\overline{\mathrm{x}}<\mu+\mathrm{e})$

## Population estimation

$\square$ We may estimate $\mu$ :
$\square$ from a random sample $\left\{x_{1}, x_{2}, \ldots, x_{n}\right\}$
$\square$ with a certain confidence level $C$, where $0 \leq C \leq 1$ )
$\square$ if we know the true standard deviation $\sigma$
$\square$ Let
$\square z^{*}$ be such that the area under the standard normal curve between -z* and $z^{*}$ is C
$\square$ Then the level $C$ confidence interval for $\mu$ is
$\square[\bar{x}-e, \bar{x}+e]$
$\square$ where $e=z^{*} \frac{\sigma}{\sqrt{n}}$

- Normal distribution: Exact
-Always: approx. for large n

$\square$ The blue intervals indicate $[\bar{x}-e, \bar{x}+e]$ for various samples
$\square$ Some of them miss $\mu$


## Example

$\square 18$ ys old men from Finnmark
$\square$ Pick a random sample of 100 men:

- $\bar{x}=177$
$\square$ Estimate the average height for this population
- Choose confidence level 0.95

$$
\bar{x} \pm z^{*} \frac{\sigma}{\sqrt{n}}=177 \pm 1.96 \frac{6}{\sqrt{100}}=177 \pm 1.176
$$

$\square$ The $95 \%$ confidence interval for $\mu$ : [175.8, 178.2]
$\square$ This presupposes that we know $\sigma$ !
$\square$ Not normally the case

## Estimation

$\square$ How to estimate the true mean $\mu$ of a sample if the standard deviation $\sigma$ of the population is unknown?
$\square$ All we have is a sample $X=\left\{x_{1}, x_{2}, \ldots, x_{n}\right\}$
$\square$ The sample mean $\bar{x}$ is still the best estimate of the pop. mean $\mu$
$\square$ How good an estimate is this?

## Estimation

$\square$ To determine this, we try to estimate the true standard deviation of the population.
$\square$ We use the standard deviation of the sample $X$,
$\square s^{2}=\left((x 1-\bar{x})^{2}+(x 2-\bar{x})^{2}+\ldots+(x n-\bar{x})^{2}\right) /(n-1)$
$\square$ Observe ( $n-1$ ) and not $n$
$\square$ That is to compensate for using $\bar{x}$ instead of $\mu$ in the formula
$s$ is a random variable (like $\bar{x}$ ) over all s.r.samples of size $n$ $s$ is an unbiased estimator for $\sigma: \operatorname{Exp}(\mathrm{s})=\sigma$

## Estimation

$\square$ In addition we do not use the standard Zdistribution but the t -distribution for $\mathrm{n}-1$.
$\square$ Then the level $C$ confidence interval for $\mu$ is
$\square[\bar{x}-e, \bar{x}+e]$

- Where

$$
e=t * \frac{s}{\sqrt{n}}
$$

$\square$ and $t^{*}$ is the value from the $t(n-1)$ density curve for $C$

The t -distribution is similar to the z -distribution for large n . But is more picky when $t$ is small

## Example

$\square$ Assume we do not know the st.dev. 18 ys old men from Finmark
$\square$ Pick a random sample of 9 men:

- $\bar{x}=177, s=5$
$\square$ Estimate the average height for this population
- Choose confidence level 0.95

Table, or
In [78]: stats.t.ppf(.025,8)
Out[79]: -2.3060041350333709

$$
\bar{x} \pm t * \frac{s}{\sqrt{n}}=177 \pm 2.306 \frac{5}{\sqrt{9}}=177 \pm 3.843
$$

$\square$ The 95\% confidence interval for $\mu$ : [173.1, 180.9]
$\square$ Exact for normal distribution
$\square$ Approximation for large n otherwise

## T-test

$\square$ The z-significance test assumed that:
$\square$ we know the true mean and st.dev for the population, $\mu$ and $\sigma$
$\square$ If we know $\mu$ but not $\sigma$
$\square$ we use the T -test for one sample
$\square$ The difference is that we use
$\square$ the sample standard deviation $s$
$\square$ the $t(n-1)$-distribution instead of the normal distribution

## Summary

|  | $\sigma$ known | $\sigma$ unknown |
| :--- | :--- | :--- |

