INF5830 – 2015 FALL
NATURAL LANGUAGE PROCESSING

Jan Tore Lønning, Lecture 3, 1.9

Today: More statistics

- Binomial distribution
- Continuous random variables/distributions
- Normal distribution
- Sampling and sampling distribution
- Statistics
  - Hypothesis testing
  - Estimation
  - Known and unknown standard deviation

Last week – Probability theory

- Probability space
  - Random experiment (or trial) (no: forsøk)
  - Outcomes (utfallene)
  - Sample space (utfallsrommet)
  - An event (begivenhet)
  - Bayes theorem
- Discrete random variable
  - The probability mass function, pmf
  - The cumulative distribution function, cdf
  - The mean (or expectation) (forventningsverdi)
  - The variance of a discrete random variable X
  - The standard deviation of the random variable

Discrete random variables

Mean of a discrete random variable

- The mean (or expectation) (forventningsverdi) of a discrete random variable X:

$$\mu_X = E(X) = \sum_x p(x)x$$

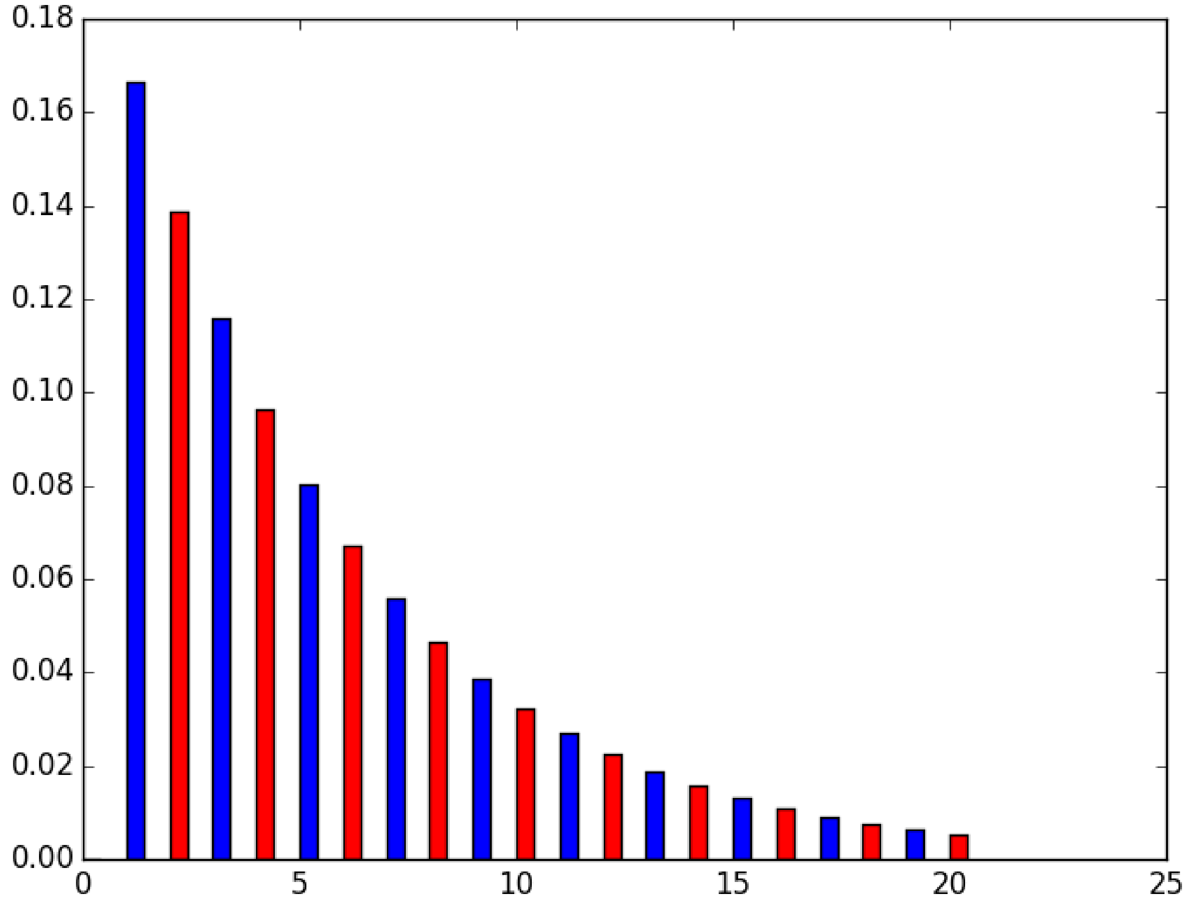
- Useful to remember

$$\mu_{(X+Y)} = \mu_X + \mu_Y$$

$$\mu_{(a+bX)} = a + b\mu_x$$

Examples:
One dice: 3.5
Two dices: 7
Ten dices: 35

Example

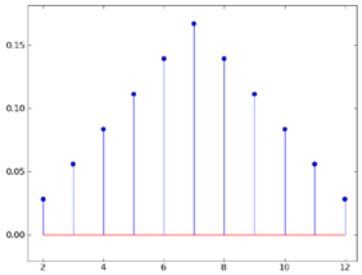


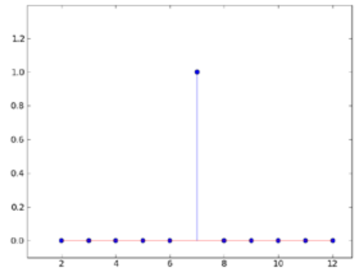
- Throwing a dice until you get 6
- P(odd) = ?
- P(even) = P(odd)*5/6
- P(even) + P(odd) = 1

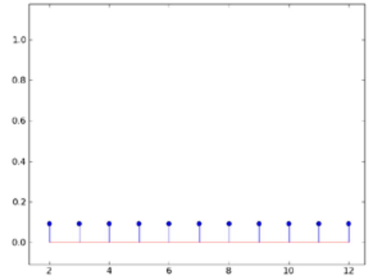
- $pmf(n) = \frac{1}{6}(\frac{5}{6})^{(n-1)}, n \geq 1$
- $\mu = 6$

More than mean

- Mean doesn't say everything
- Example
  - (1.3) The sum of the two dice, Z, i.e.
    - $p_Z(2) = 1/36, \ldots, p_Z(7) = 6/36$ etc
  - (3.2) p_2 given by:
    - $p_2(7)=1$
    - $p_2(x)= 0$ for $x \neq 7$
  - (3.3) p_3 given by:
    - $p_3(x)= 1/11$ for $x = 2,3,\ldots,12$
  - Have the same mean but are very different







Variance

- The variance of a discrete random variable X

$$Var(X) = \sigma^2 = \sum_x p(x)(x-\mu)^2$$

- Observe that

$$Var(X) = E((X - E(X))^2)$$

- It may be shown that this equals $E(X^2) - (E(X))^2$
- The standard deviation of the random variable

$$\sigma = \sqrt{Var(X)}$$

Examples of variance

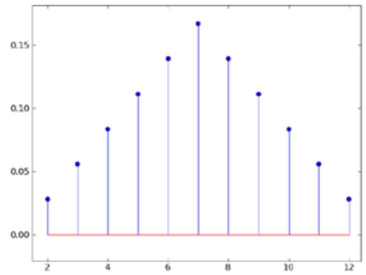
- Throwing one dice
  - $\mu = (1+2+..+6)/6=7/2$
  - $\sigma^2 = ((1- 7/2)^2 +(2-7/2)^2+\ldots(6-7/2)^2)/6 = (25+9+1)/4*3=35/12$
- (Ex 1.3) Throwing two dice: $\sigma^2 = 35/6$
- (Ex 3.2) p_2, where $p_2(7)=1$ has variance 0
- (Ex 3.3) p_3, the uniform distribution, has variance:
  - $((2-7)^2+\ldots(12-7)^2)/11 = (25+16+9+4+1+0)*2/11 = 10$

Probability distributions

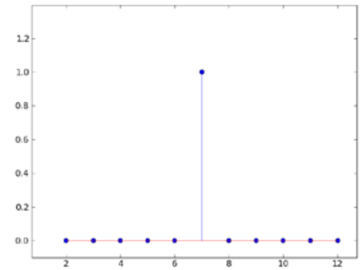
Sannsynlighetsfordelinger

Examples of distributions

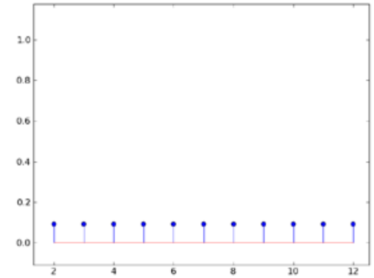
- (1.3) The sum of the two dice, Z, i.e.
  - $p_Z(2) = 1/36, \ldots, p_Z(7) = 6/36$ etc



- (3.2) p_2 given by:
  - $p_2(7)=1$
  - $p_2(x)= 0$ for $x \neq 7$



- (3.3) p_3 given by:
  - $p_3(x)= 1/11$ for $x = 2,3,\ldots,12$



Bernoulli trial

- One experiment, two outcomes
- $\Omega_X = \{0, 1\}$
- Write p for p(1)
- Then p(0) = 1-p

Examples:
- Flipping a fair coin, p=1/2
- Rolling a dice, getting a 6, p=1/6

- The mean/expectation: 0*p(0)+1*p(1)=0+p=p
- Variance $Var(X) = \sigma^2 = \sum_x p(x)(x-\mu)^2 =$

Bernoulli trial

- One experiment, two outcomes
- $\Omega_X=\{0, 1\}$
- Write p for p(1)
- Then p(0) = 1-p

Examples:
- Flipping a fair coin, p=1/2
- Rolling a dice, getting a 6, p=1/6

- The mean/expectation: 0*p(0)+1*p(1)=0+p=p
- Variance $Var(X) = \sigma^2 = \sum_x p(x)(x-\mu)^2 =$

$$(1-p)(0-p)^2 + p(1-p)^2 = p(1-p)$$

Binomial distribution

- Binomial distribution (binomisk fordeling)
- Conducting *n* Bernoulli trials with the same probability and counting the number of successes

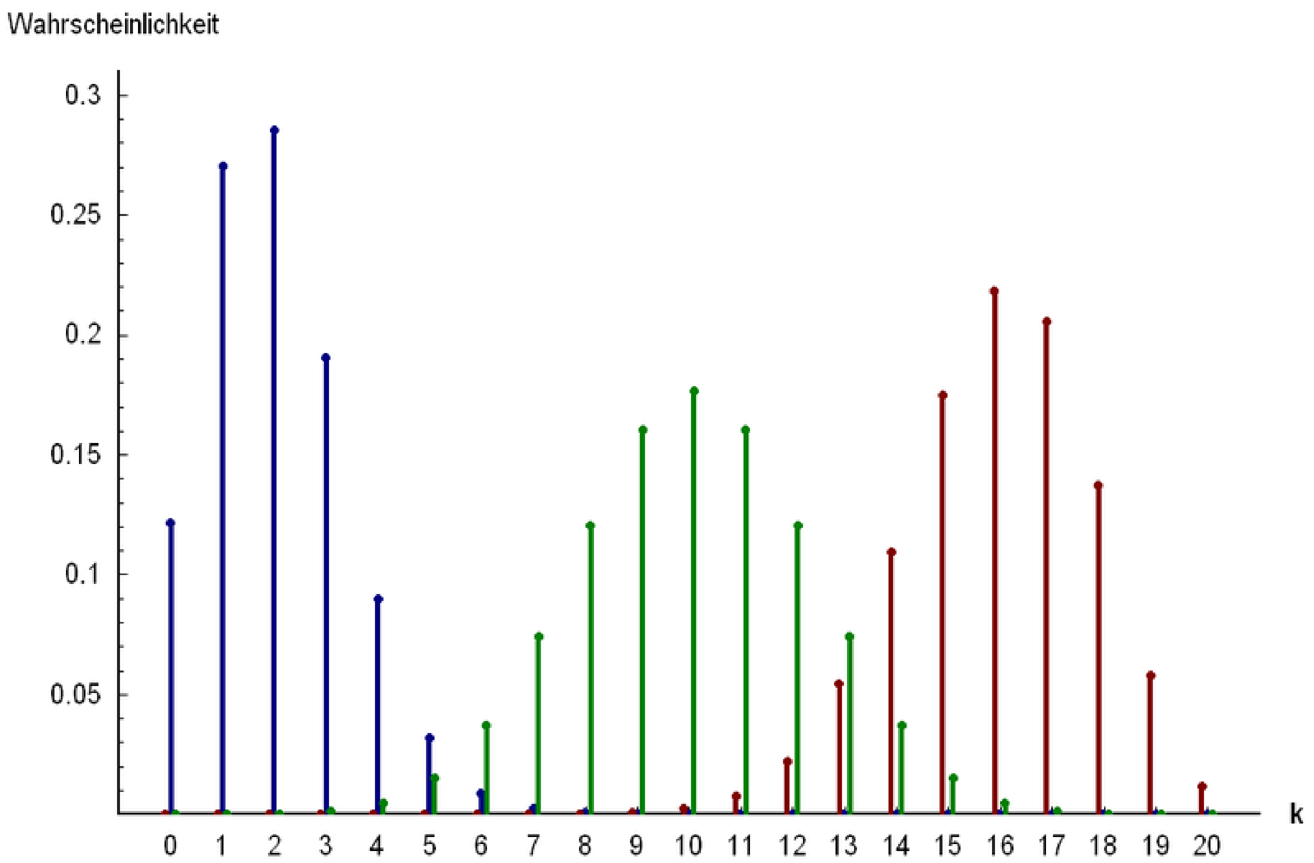
- Example flipping a fair coin *n* times, p(k):
  - n=2: p(0)=1/4, p(1)=1/2, p(2) =1/4
  - n=3: p(0)=1/8, p(1)=3/8, p(2)=3/8, p(3)=1/8
  - n=4: (1,4,6,4,1)/16
  - n=5: (1,5,10,5,1)/32
- n: $p(k) = \binom{n}{k}\left(\frac{1}{2}\right)^n$ where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Binomial distribution

- Binomial distribution (binomisk fordeling)
- General form:
  - $0<p<1$
  - *n* a natural number
- B(n,p) is given by $$b(k;n,p)=\binom{n}{k}p^k(1-p)^{(n-k)}$$

for k = 0, 1, ...n, where $$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

Binomial distribution



- n = 20
- p = 0.1 (blue), p = 0.5 (green) and p = 0.8 (red)

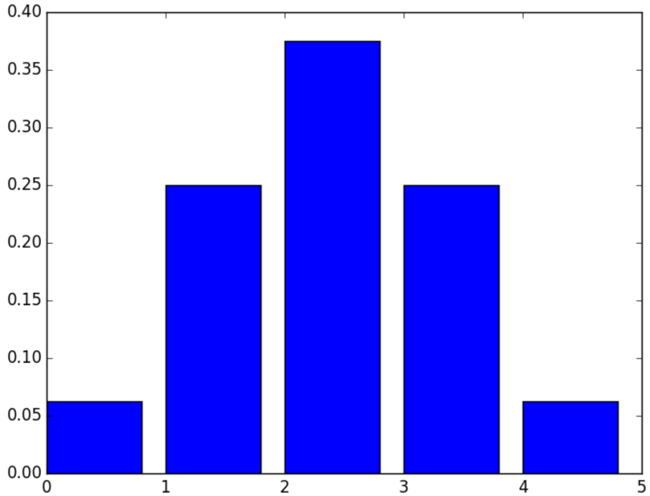
Binomial distribution

- Mean/expectation, μ, of B(n,p) is *np*
  - *n* Bernoulli trials
  - Each Bernoulli trial has mean *p*
- The variance is *np(1-p)*
  - Because the Bernoulli trials are independent
  - Each Bernoulli trial has variance *p(1-p)*

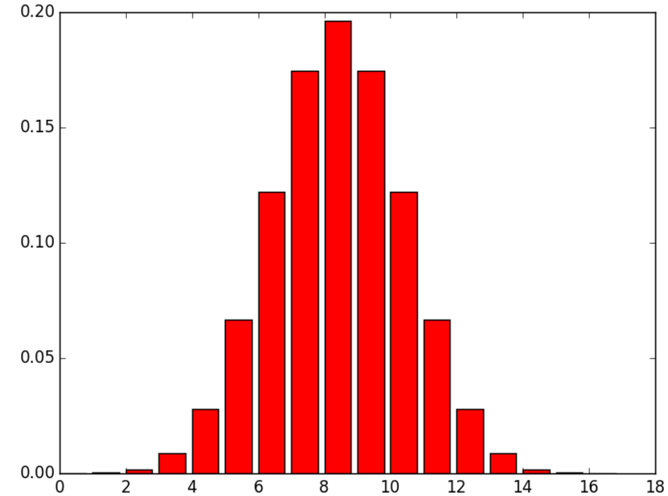
The variance of the sum of two <u>independent</u> random variables is the sum of their variances

p=0.5

N=4:

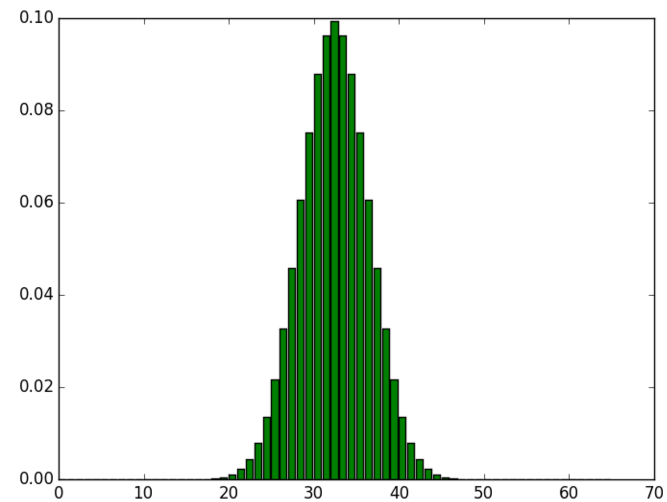


N=16:



N	1	4	16	64	256
σ^2	0.25	1	4	16	64
σ	0.5	1	2	4	8

N=64:



- The relative variation gets smaller with growing N
- The pmf graph approaches a bell shape

Think about

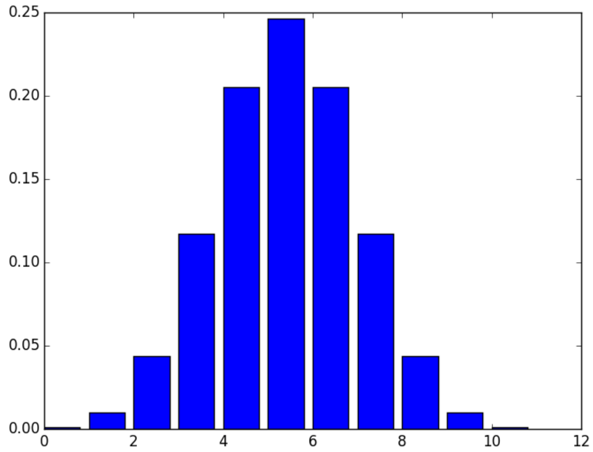
- Flip a coin 10 times, count the number of heads
- You expect 5 heads, but not exactly 5
  - 6 is OK
- When do you start to worry whether the coin is unfair?
  - 8 heads?
  - 9 heads?

- This is the task for inferential statistics

Tossing a fair(?) coin

- The cumulative distribution function: ``How likely is it to get N or fewer tails?´´

10:



N	pmf(N)	cdf(N)
0	0.001	0.001
1	0.010	0.011
2	0.044	0.055
3	0.117	0.172
4	0.205	0.377
5	0.246	0.623
6	0.205	0.828
7	0.117	0.945
8	0.044	0.989
9	0.010	0.999
10	0.001	1.000

SciPy

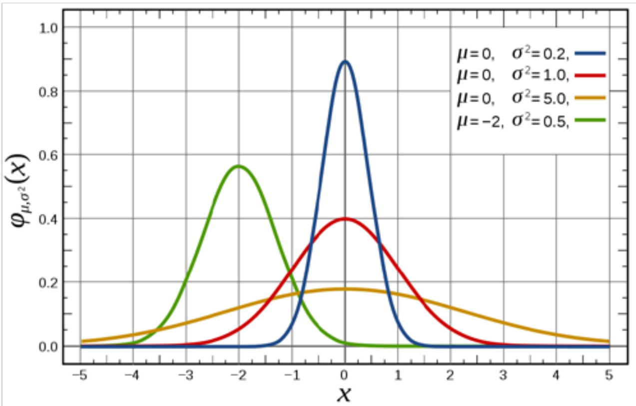
- import scipy
- from scipy import stats
- bin10 = stats.binom(10, 0.5) # N=10, p=0.5
- bin10.pmf(3) # probability mass of 3
- bin10.cdf(3)   # cumulative distribution function at 3
- bin10.var()   # variance
- bin10.std()   # standard deviation

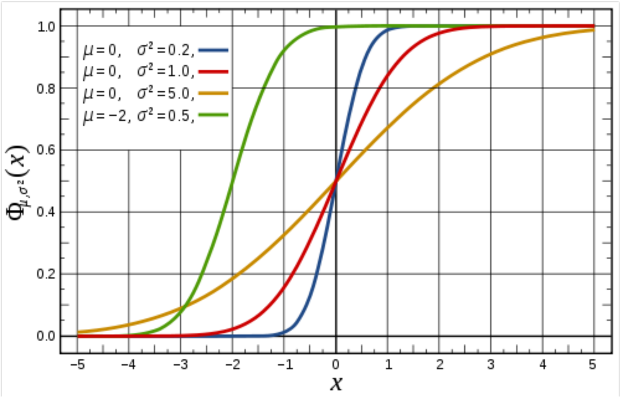
Continuous random variables

Continuous random variables

- $P(X=a) = 0$ for all values a
- The probability mass function does not make sense
- The cumulative distribution function, cdf, given by $F(a) = P(X \leq a)$ makes sense
- $P(a \leq x \leq b) = F(b) - F(a)$
- To calculate expectation and variance we must use integration instead of (infinite) sums.
  - We skip the details!

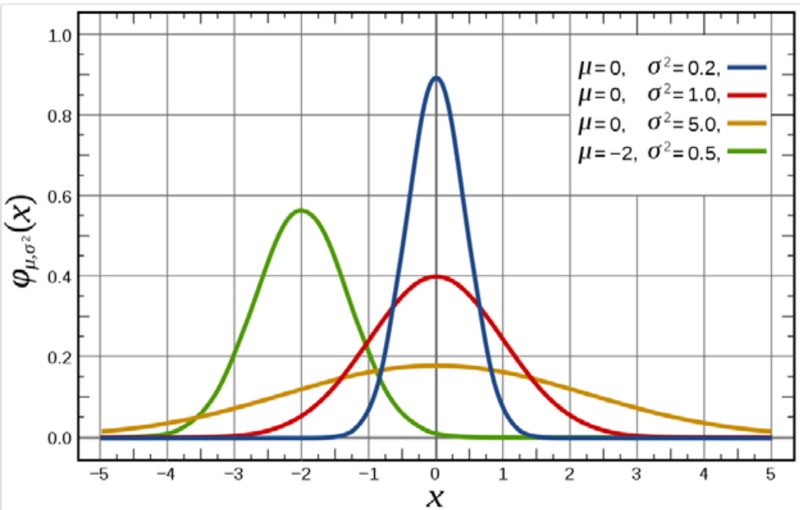
Probability density function





- The derivative of the cdf, F', is called the probability density function, pdf (sannsynlighetstetthet)
- We draw curves for pdf-s
- The pdf has a similar relationship to the cdf in the continuous case as the pma has in the discrete case

The normal distribution

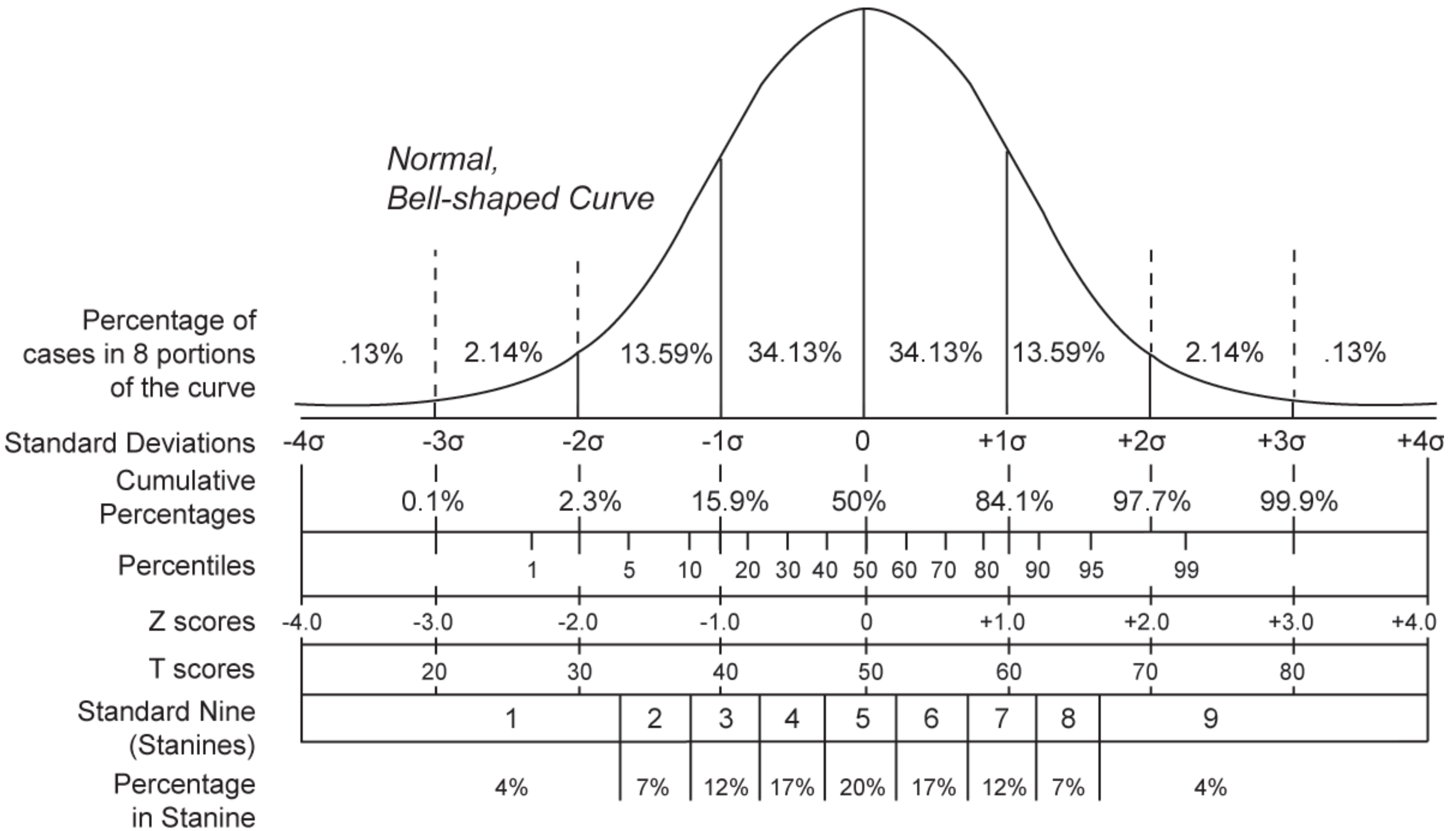


z-score relates the general case to the standard case

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

		Standard norm.dist. (red curve)	General norm.dist N(μ,σ)
Scary formula	(Don't have to remember)	$f(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
Important			
Mean		0	μ
Standard deviation		1	σ

68% - 95% - 99.7%


Normal, Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ -3σ -2σ -1σ 0 +1σ +2σ +3σ +4σ
Cumulative Percentages
0.1% 2.3% 15.9% 50% 84.1% 97.7% 99.9%
Percentiles
1 5 10 20 30 40 50 60 70 80 90 95 99
Z scores
-4.0 -3.0 -2.0 -1.0 0 +1.0 +2.0 +3.0 +4.0
T scores
20 30 40 50 60 70 80
Standard Nine (Stanines)
1 2 3 4 5 6 7 8 9
Percentage in Stanine
4% 7% 12% 17% 20% 17% 12% 7% 4%

Example

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

- Tallness of Norwegian young men (rough numbers):
  - μ = 180 cm
  - σ = 6cm
  - z = (186-180)/6=1 (standard deviation)
  - (100-68)/2%= 16% are taller than 186cm

  - How many are taller than 190cm?
  - z = (190-180)/6 = 1.67
  - Prob. = 0.0475 (from table or software)

Sampling distribution

Utvalgsfordeling

Sampling - empirically

Goal:
- make assertions about a whole population
- from observations of a sample (utvalg)

- 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
- Not as simple as it sounds (c.f. the current election polls):
  - Various methods to rescue
  - E.g. choose from known groups, weigh by group size (gender, age, home town, etc.)

Sampling in Language Technology

- You want to take a simple random sample of words from a corpus?
  - Can you use the *n* first sentences?
  - Can you use a random sample of *n* sentences?
- How can you build a corpus (sample) which gives a random sample of Norwegian texts?

Sampling distributions – Example

- Height: X
  - assume N(180, 6)
  - (Var=36)
- Randomly choose 100.
- Add their heights:
  $S = X_1 + X_2 + \ldots + X_n$
- A new random variable (all such samples)
  - Exp(S) = n*μ= 18000 (cm)
  - Var(S) = 100*Var(X) = 3600
  - $\sigma_S = 10 \times \sigma_X = 60\ (cm)$

Source: Wikipedia

Sampling distributions – Example

- Height: X
  - assume N(180, 6)
  - (Var=36)
- Randomly choose 100.
- Add their heights:
  $S = X_1 + X_2 + \ldots + X_n$
- A new random variable
  (all such samples)
  - Exp(S) = n*μ= 18000 (cm)
  - Var(S) = 100*Var(X) = 3600
  - $\sigma_S = 10 \times \sigma_X = 60\ (cm)$

- The mean of the samples:
- $\overline{X} = S/n$
- A new random variable
  (all such means of samples of 100)
  - Exp(S) = μ= 180 (cm)
  - $\sigma_{\bar{X}} = \frac{1}{100} \times \sigma_S = 0.6\ (cm)$

Sampling distributions

- Let
  - X be a random variable for a population with exp: μ, std: σ
  - Let $S = X_1 + X_2 + \ldots + X_n$, i.e. each X_i equals X
  - Let : $\overline{X} = S/n$
- Then:
  - $Exp(S) = n * \mu$
  - $Exp(\overline{X}) = \mu$
  - $$Var(S) = \sigma_S^2 = n \times Var(X) = n \times \sigma_X^2$$
  - $$Var(\overline{X}) = \sigma_{\overline{X}}^2 = \frac{1}{n^2} \times Var(S) = \frac{1}{n} \times \sigma_X^2$$
  - $$\sigma_{\overline{X}} = \frac{1}{\sqrt{n}} \times \sigma_X$$

Effect of sample size

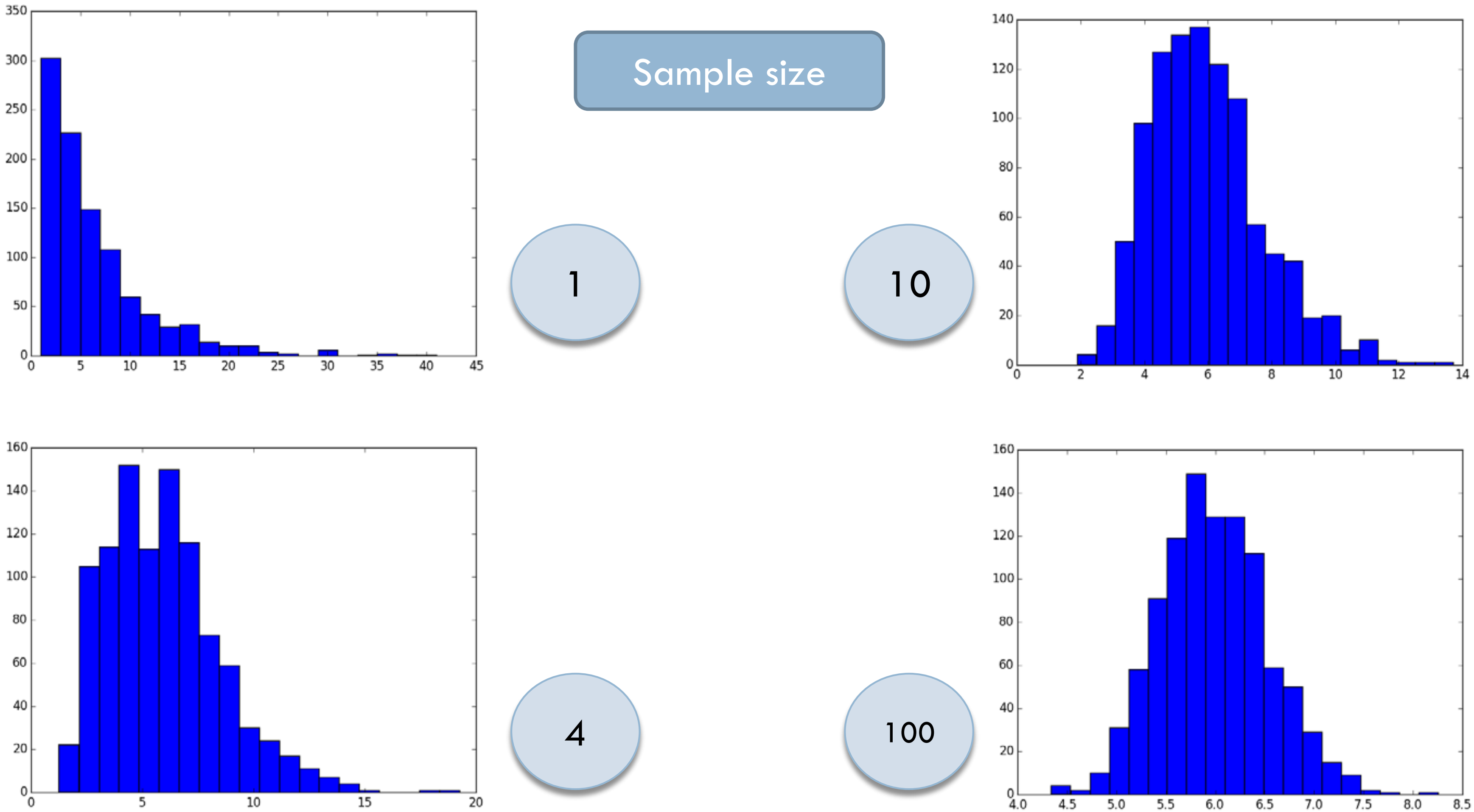
Sample size	1	4	16	100	400	1600
Standard dev.	6	3	1.5	0.6	0.3	0.15

The form of the distribution

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

Example: throwing the dice until a 6

Number of samples: 1000



Binomial distribution

$$b(k;n,p)=\binom{n}{k}p^k(1-p)^{(n-k)}$$

Population: all Bernoulli trials with probability *p*.
Sample: *n* such trials
Example: Throwing a dice *n* times, counting the number of 6-s (success)

- Number of successes: X
- Random variable over all series of *n* trials
- Binomial distribution (binomisk fordeling): B(n,p)
- E(X)= *np*
- Var(X)= *np(1-p)*
- $\sigma_X = \sqrt{np(1-p)}$
- Approximated by N(*np*, $\sqrt{np(1-p)}$) for large n

- Proportion of success: $\hat{p}$=X/n
- E($\hat{p}$) = E(X/n) = *np*/*n* = *p*
- $Var(\hat{p}) = \sigma_X^2/n^2 =$ $np(1-p)/n^2 = p(1-p)/n$
- $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \frac{\sigma_Y}{\sqrt{n}}$
- Approximated by N(*p*, $\sqrt{p(1-p)/n}$) for large n

Rule of thumb:
np>10 and
n(1-p)>10

Example

- Example:
  - p = 0.8

You have a classifier which you think is 80 % correct.
What can you expect of this classifier from samples of various sizes?

N	E(X)	Var(X)	SD(X)	$\mu \pm 2\sigma$	E($\hat{p}$) =E(X/n)	Var($\hat{p}$)	SD($\hat{p}$)	$\mu \pm 2\sigma$
1	0.8	0.16	0.4		0.8	0.16	0.4	
25	20	4	2		0.8	0.0064	0.08	
100	80	16	4	[72, 88]	0.8	0.0016	0.04	[.72,.88]
2500	2000	400	20	[1960, 2040]	0.8	0.000064	0.008	
10000	8000	1600	40	[7920,8080]	0.8	0.000016	0.004	[.792,.808]