

Lecture 1

Probability Theory and Statistics

Random Numbers
Distributions
Samples
Confidence Intervals

6.3.2002

Copyright Teemu Kerola 2002

1

Random Numbers

- Ref: Law-Kelton -91, Chapter 4
- Random number (satunnaisluku)
- Sample space (otosavaruus)
 - discrete $S = \{1, 2, 3, 4, 5, 6\}$
 - continuous $S = [0, 1]$ $S = [0, 1)$
- Random variable (satunnaismuuttuja)
 - one way to generate random numbers
 - determined by a function or rule: $\mathbf{R} \rightarrow \mathbf{S}$
 - random variables: $X, Y=5X, Z=g(X,Y)$

6.3.2002

Copyright Teemu Kerola 2002

2

Random Variable X

- Distribution function $F(x) = P(X \leq x)$
- Density function $f(x) = P(X=x) = P(x)$
- Both $F(x)$ and $f(x)$ determine the (same) distribution

Prob

- X discrete?

$$- P(a < X \leq b) = \sum_{a+1}^b f(x) = F(b) - F(a)$$

dice:
 $P(X=6) = 1/6$

- X continuous?

$$- P(a < X \leq b) = \int_a^b f(x) = F(b) - F(a)$$

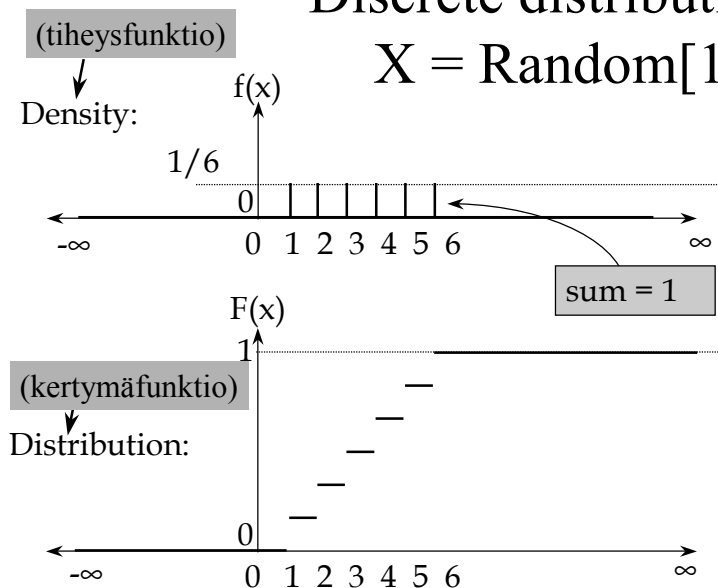
temperature:
 $P(20 \leq X \leq 24)$
 $= 80\%$

6.3.2002

Copyright Teemu Kerola 2002

3

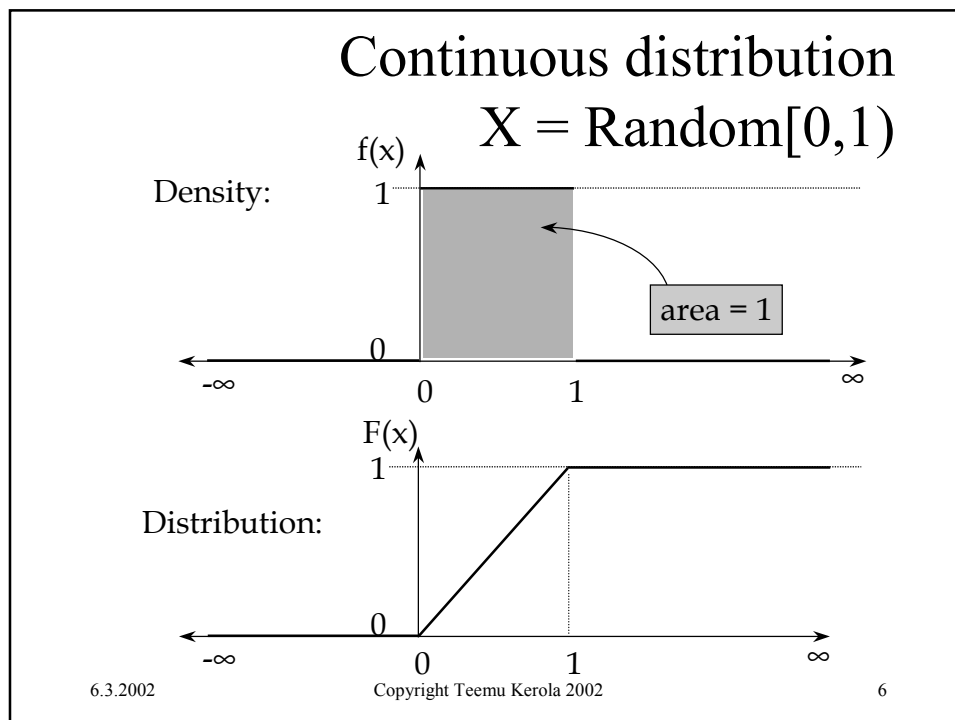
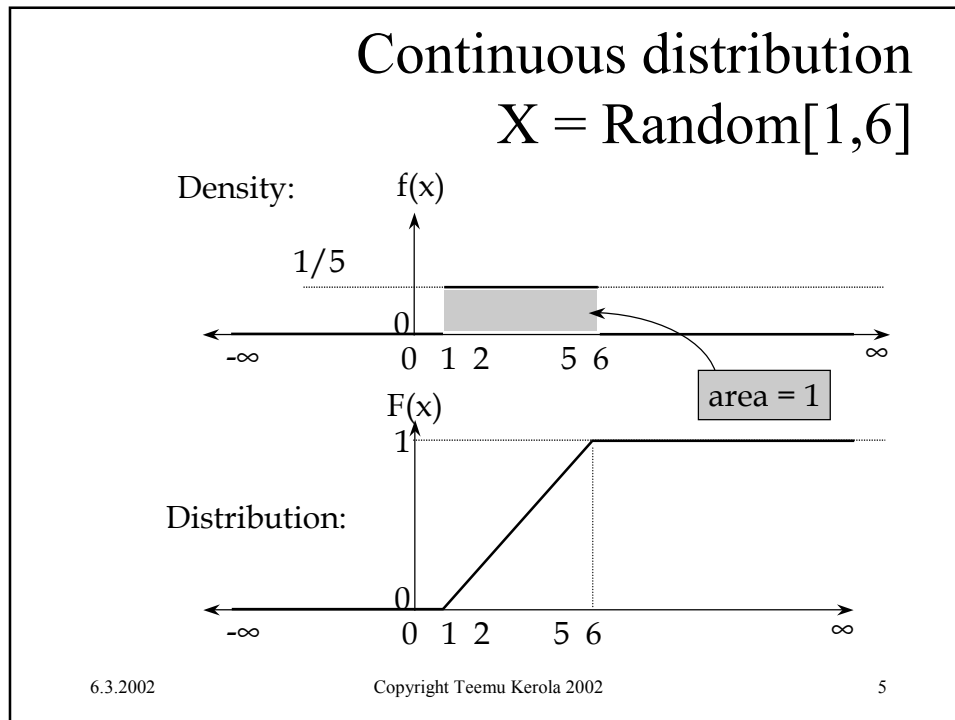
Discrete distribution $X = \text{Random}[1,6]$

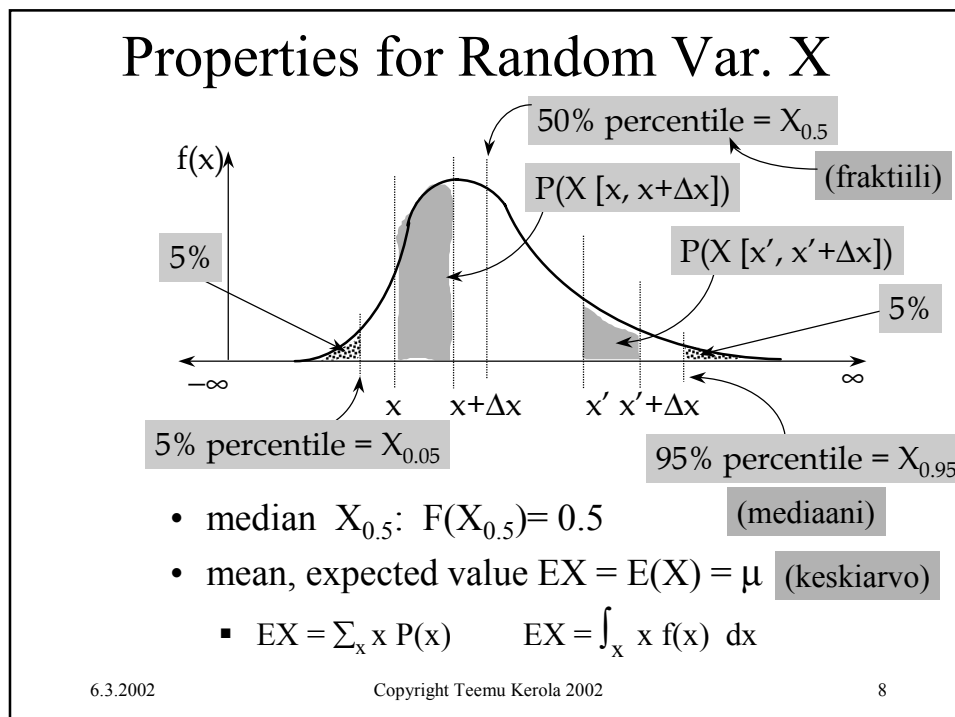
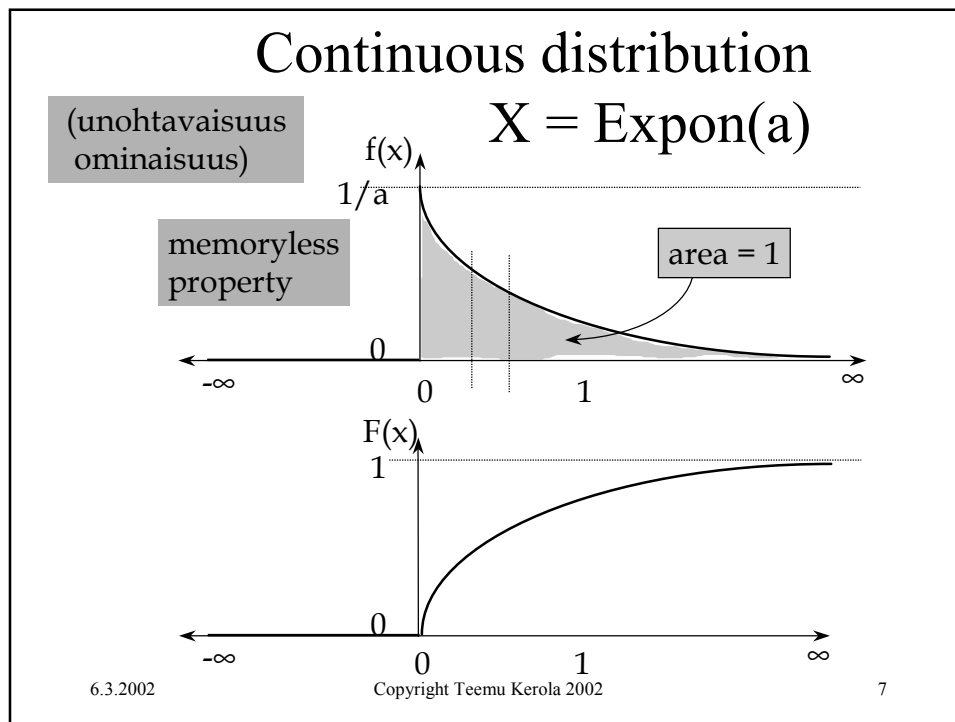


6.3.2002

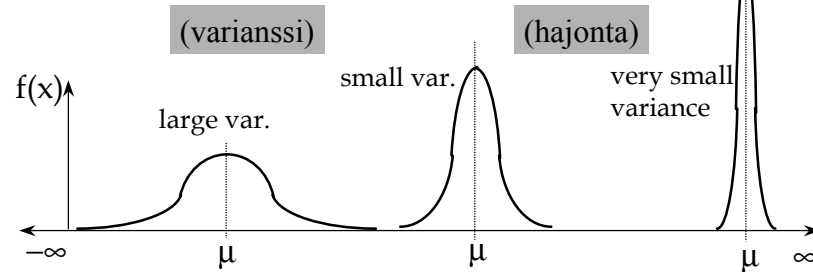
Copyright Teemu Kerola 2002

4





Variance and Deviation



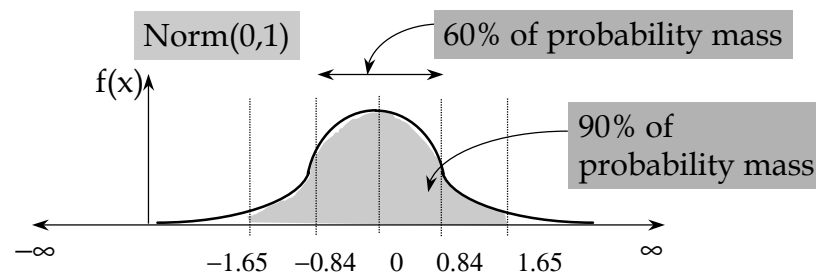
- average $E(X) = EX = \mu = \int_{-\infty}^{\infty} x f(x) dx$
– describes where the values are located
- variance $\text{Var}(X) = \sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2$
– describes how close the values are to average value
- Standard deviation $\sigma = \sqrt{\sigma^2}$

6.3.2002

Copyright Teemu Kerola 2002

9

Standard Normal Distribution (2)



- Normal distribution $\text{Norm}(\mu, \sigma^2)$
- Standard Normal Distribution = Z-distribution = $\text{Norm}(0, 1) = N(0, 1)$
- Z-percentiles $Z_{0.95} = 1.645 = Z_{1-\alpha/2}$ for $\alpha = 10\%$
 $Z_{0.80} = 0.84$

6.3.2002

Copyright Teemu Kerola 2002

10

Covariance

(kovarianssi)

- X and Y are random variables
 - Does X depend on Y or vice versa or what?
- Covariance $\text{Cov}(X, Y) = C_{XY}$

$$= E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$$
- $C_{XY} = C_{YX}$
- $C_{XX} = \text{Cov}(X, X) = \text{Var}(X) = \sigma^2$
 - $C_{XY} = 0$ X, Y uncorrelated X large => Y ???
 - $C_{XY} > 0$ X, Y positively correlated X large => Y large
 - $C_{XY} < 0$ X, Y negatively correlated X large => Y small

6.3.2002

Copyright Teemu Kerola 2002

11

Correlation

(korrelaatio)

- Problem: C_{XY} varies much in magnitude
 - can not deduct much from its magnitude
- Solution: scale with (geom. aver) variance
- Correlation
$$\text{Cor}(X, Y) = \rho_{XY} = \frac{C_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} \in [-1, 1]$$
 - $\rho_{XY} = 0$ X, Y uncorrelated
 - $\rho_{XY} > 0$ X, Y positively correlated
 - $\rho_{XY} < 0$ X, Y negatively correlated
 - $\rho_{XY} = 0.98$ X, Y highly positively correlated
- Correlation does not mean causal dependence
 - E.g., eating ice cream vs. drowning accidents

6.3.2002

Copyright Teemu Kerola 2002

12

Independent X and Y ?

(riippumat-
tomuus)

- Random variables X, Y
- Joint prob. density function $f(X,Y)$:

$$P(X \in A, Y \in B) = \int_A \int_B f(x,y) dx dy$$
- X and Y are independent if

$$f(x,y) = f_X(x) \cdot f_Y(y)$$

(yhteisjakauman
tiheysfuntio)

where

$$f_X(x) = \int_{-\infty, \infty} f(x,y) dy$$

$$f_Y(y) = \int_{-\infty, \infty} f(x,y) dx$$

marginal prob.
density functions

6.3.2002

Copyright Teemu Kerola 2002

13

Set of Random Variables

- Many random variables X_i or X_i
 - $i \in [1, N]$ $i \in [1, \infty]$
- $E(X_i) = EX_i = \mu_{X_i} = \mu_i$
- $\text{Var}(X_i) = \text{Var}(X_i) = \sigma_{X_i}^2 = \sigma_i^2$
- $\text{Cov}(X_i, X_j) = \text{Cov}(X_i, X_j) = C_{X_i, X_j} = C_{i,j} = C_{ij}$
- $\text{Cor}(X_i, X_j) = \text{Cor}(X_i, X_j) = \rho_{X_i, X_j} = \rho_{i,j} = \rho_{ij}$
- All comparisons pair-wise!

6.3.2002

Copyright Teemu Kerola 2002

14

IID Set of Random Variables

- Random variables $X_1, X_2, \dots$
- X_i 's identically distributed?

$$\begin{aligned} EX_i &= \mu_i = \mu & \forall i = 1, 2, \dots \\ \text{Var}(X_i) &= \sigma_i^2 = \sigma^2 & \forall i = 1, 2, \dots \\ f_{X_i}(x) &= f(x) & \forall i = 1, 2, \dots \end{aligned}$$
- X_i 's independent?

$$X_i \text{ and } X_j \text{ are independent } \forall i, j \in \{1, 2, \dots\}$$
- Now, X_i 's are IID
 - independent and identically distributed (= IID)
 - almost never the case in real life
 - almost always needed for the math to work out!

6.3.2002

Copyright Teemu Kerola 2002

15

Sample

(otos)

- Problem: What is the distribution of X ?
- Solution: try it out and see
- Sample: one set of values for X
 - throw a dice for 100 times? 10000 times?
 - measure response time from 100 trials?
Each trial 10 values? 1000 values?
 - write down sales sum for next 10 years?
- New question: what can we tell about X from the sample?
 - What are $EX=\mu$ and $\text{Var}(X)=\sigma^2$?

6.3.2002

Copyright Teemu Kerola 2002

16

Sample Properties

- Sample Set $S = \{ X_i \mid i=1, \dots, n \}$ (otosjoukko)
- Sample points X_i (otospisteet)
 - should be set of IID random variables
 - $X_i \sim X$, i.e.,
 X_i has the (unknown) same distribution than X
 $EX_i = EX = \mu$ and $\text{Var}(X_i) = \text{Var}(X) = \sigma^2$
- Sample mean $\bar{X}(n) = \frac{\sum_i X_i}{n}$ (otoskeskiarvo) (harhaton estimaatti)

is unbiased estimator for $EX = \mu$: $E\bar{X}(n) = \mu$

 - with large n , sample mean is close to μ
 - how close? how large n ?

6.3.2002

Copyright Teemu Kerola 2002

17

Sample Properties

- Variance for sample mean: $\text{Var}(\bar{X}(n)) = \frac{\sigma^2}{n}$
 - problem: $\text{var}(X) = \sigma^2$ is unknown
 - solution: use sample variance $s^2(n)$ instead:
$$s^2(n) = \frac{\sum_i [X_i - \bar{X}(n)]^2}{n-1} = \frac{\sum_i X_i^2 - n\bar{X}^2(n)}{n-1}$$

(otosvarianssi)

 - now, $\widehat{\text{Var}}(\bar{X}(n)) = \frac{s^2(n)}{n} = \frac{\sum_i [X_i - \bar{X}(n)]^2}{n(n-1)}$

is unbiased estimator for $\text{Var}(\bar{X}(n))$

 - know now how good estimator sample mean is

6.3.2002

Copyright Teemu Kerola 2002

18

Central Limit Theorem (CLT)

- “When n becomes large, normalized sample mean becomes normally distributed”

$$\frac{\bar{X}(n) - \mu}{\sqrt{\sigma^2 / n}} \xrightarrow{n \rightarrow \infty} N(0,1)$$

(keskeinen
raja-arvolause)

- Problem: what if n is not very large?
- Solution: use Student distribution instead:

$$\frac{\bar{X}(n) - \mu}{\sqrt{s^2 / n}} \approx T(n-1)$$

(vapausaste)

where $n-1$ is the degrees of freedom for T

- Percentiles for $N(0,1)$ and $T(k)$ are tabulated

6.3.2002

Copyright Teemu Kerola 2002

19

How to Use Central Limit Theorem?

- Use CLT for percentiles

$$P(-z_{1-\alpha/2} \leq \frac{\bar{X}(n) - \mu}{\sqrt{\sigma^2 / n}} \leq z_{1-\alpha/2}) = 1 - \alpha$$

- Use unbiased estimator for σ^2

$$P(-z_{1-\alpha/2} \leq \frac{\bar{X}(n) - \mu}{\sqrt{s^2(n) / n}} \leq z_{1-\alpha/2}) = 1 - \alpha$$

- Solve for μ , get confidence interval for μ

$$P(\mu \in \left\{ \bar{X}(n) \pm z_{1-\alpha/2} \sqrt{s^2(n) / n} \right\}) = 1 - \alpha$$

(luottamusväli)

6.3.2002

Copyright Teemu Kerola 2002

20

Confidence Interval

- $(1-\alpha)$ confidence interval

$$P(\mu \in \{\bar{X}(n) \pm z_{1-\alpha/2} \sqrt{s^2(n)/n}\}) = 1 - \alpha$$

- 90% confidence interval, $\alpha = 10\%$

$$P(\mu \in \{\bar{X}(n) \pm z_{0.95} \sqrt{s^2(n)/n}\}) = 90\% \quad n > 30 ?$$

or

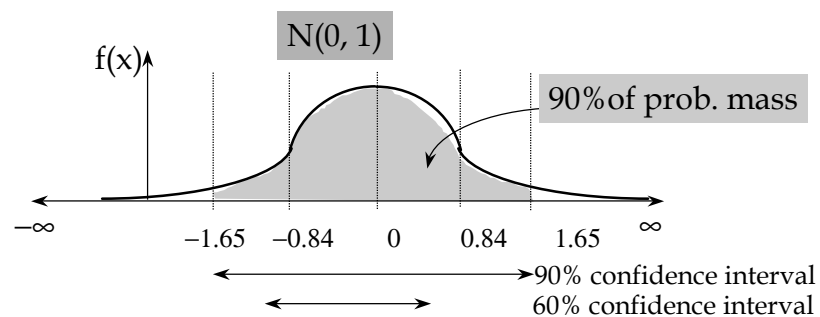
$$P(\mu \in \{\bar{X}(n) \pm t_{n-1,0.95} \sqrt{s^2(n)/n}\}) = 90\% \quad n \leq 30 ?$$

6.3.2002

Copyright Teemu Kerola 2002

21

Confidence Interval (CI)



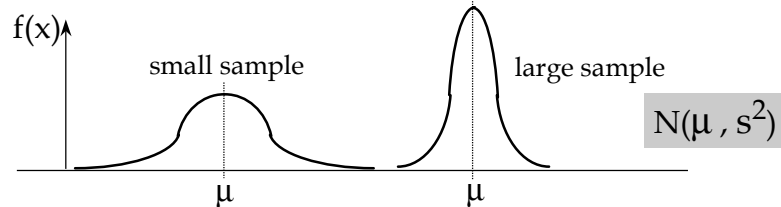
- Distribution for normalized μ based on sample
- 60% CI is narrower than 90% CI
 - 60% CI looks "better"
 - 60% CI is more likely to be wrong: 40% change that it does not to contain the true mean

6.3.2002

Copyright Teemu Kerola 2002

22

Effect of Sample Size, Strong Law of Large Numbers



- sample distribution s^2 for the μ (true, unknown mean) is based on sample
- large sample, s^2 small
- small sample, s^2 large

- Strong Law of Large Numbers:

(suurten lukujen laki)

$$\bar{X}(n) \xrightarrow{n \rightarrow \infty} \mu \text{ with probability 1}$$

6.3.2002

Copyright Teemu Kerola 2002

23

Example A

- Question: X is normally distributed.

What is the mean μ of X ?

1.20, 1.50, 1.68, 1.89, 0.95,
1.49, 1.58, 1.55, 0.50, 1.09

- Answer:

– Take sample of 10 observations for X

– compute $\bar{X}(10)$ and $s^2(10)$

1.34 and 0.17

– get 90% percentile for T-distribution with 9 degrees of freedom, $t_{9,0.95} = 1.83$

– compute 90% confidence interval for μ :

$$\begin{aligned} \bar{X}(10) \pm t_{9,0.95} \sqrt{s^2(10)/10} &= 1.34 \pm 1.83 * 0.13 \\ &= (1.34 \pm 0.24) = (1.10, 1.58) \end{aligned}$$

6.3.2002

Copyright Teemu Kerola 2002

24

Example A (contd)

- 90% confidence interval for μ :

$$\bar{X}(10) \pm t_{9,0.95} \sqrt{s^2(10)/10} = 1.34 \pm 1.83 * 0.13 \\ = (1.34 \pm 0.24) = (1.10, 1.58)$$

- There is 90% chance that $1.10 \leq \mu \leq 1.58$

90% chance that estimate is correct

- There is 10% chance that $\mu < 1.10$ or $1.58 < \mu$

10% chance that estimate is wrong

6.3.2002

Copyright Teemu Kerola 2002

25

Example A (contd)

- 80% confidence interval for μ

$$\bar{X}(10) \pm t_{9,0.90} \sqrt{s^2(10)/10} = (1.34 \pm 0.18) = (1.16, 1.52)$$

- 90% confidence interval for μ

$$\bar{X}(10) \pm t_{9,0.95} \sqrt{s^2(10)/10} = (1.34 \pm 0.24) = (1.10, 1.58)$$

- 95% confidence interval for μ

$$\bar{X}(10) \pm t_{9,0.975} \sqrt{s^2(10)/10} = (1.34 \pm 0.29) = (1.05, 1.63)$$

- 99% confidence interval for μ

$$\bar{X}(10) \pm t_{9,0.995} \sqrt{s^2(10)/10} = (1.34 \pm 0.42) = (0.92, 1.76)$$

6.3.2002

Copyright Teemu Kerola 2002

26

Example B

- Question: I know the distribution.
Is the mean $\mu = 5$? (“Two-tailed hypothesis” H_0)
- Answer:
 - Take sample of 10, compute test factor t_9 :

$$t_9 = \frac{\bar{X}(10) - 5}{\sqrt{s^2(10)/10}}$$

(testisuure)

(scaled to std distr)
 - get 90% confidence interval for T-distribution with 9 degrees of freedom $(-t_{9,0.95}, t_{9,0.95}) = (-1.83, 1.83)$
 - if test factor t_9 is not in there, reject H_0 , i.e., say “no”
 - 10% chance of being wrong! (why?)
 - if t_9 is in there, accept H_0 , i.e., say “yes”

6.3.2002

Copyright Teemu Kerola 2002

27

Example C

- Question: do X and Y have same mean?
- Answer:
 - Take samples for X and Y, sample averages: $\bar{X}(n)$ and $\bar{Y}(m)$
 - get test factor t_{n+m-1} :

$$t_{n+m-1} = \frac{\bar{X}(n) - \bar{Y}(m)}{s \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where $s^2 = \frac{1}{n+m-2} (\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2)$
 - get 90% confidence interval for T-distribution
 - if t_{n+m-1} is not in there, reject H_0 , i.e., say “no”
 - 10% chance of being wrong

6.3.2002

Copyright Teemu Kerola 2002

28

Usual Assumptions That Usually Do Not Apply

- Normal distribution
 - central limit theorem
 - strong law of large numbers
- Independent samples
 - $X_1, X_2, \dots$
- Assumptions required by probability theory and statistics to work do not usually apply
- Results may still be usable
- Watch out and always check how far off you are from required assumptions

6.3.2002

Copyright Teemu Kerola 2002

29

6.3.2002

Copyright Teemu Kerola 2002

30