UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH

RESEARCH DEDICATED TO LIFE

PhD Course

Advanced Biostatistics

Lecture 2

Probability Distributions and Interval Estimations

dr. P. Nazarov

petr.nazarov@lih.lu

23-05-2016

Outline

- **Discrete Probability distributions (L2.1)**
  - binomial, hypergeometric, Poisson
- **Continuous Probability distributions (L2.2)**
  - uniform, normal, exponential
- **Sampling distribution (L2.3)**
- **Interval estimations (L2.4)**
  - Interval estimations for means and proportions
  - Interval estimations for variance
  - Interval estimations for correlation
  - Simulation-based assignment of interval estimation for random functions

L2.1. Discrete Probability Distributions

Random Variables

Random variable
A numerical description of the outcome of an experiment.

A random variable is always a numerical measure.

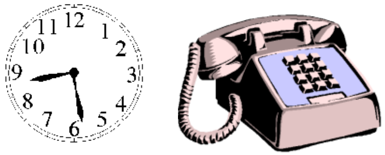
Discrete random variable
A random variable that may assume either a finite number of values or an infinite sequence of values.

Continuous random variable
A random variable that may assume any numerical value in an interval or collection of intervals.

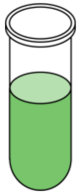
Roll a die

Number of calls to a reception per hour

Time between calls to a reception



Volume of a sample in a tube



Weight, height, blood pressure, etc

L2.1. Discrete Probability Distributions

Discrete Random Variables

Probability distribution

A description of how the probabilities are distributed over the values of the random variable.

Probability function

A function, denoted by *f(x),* that provides the probability that *x* assumes a particular value for a discrete random variable.

Number of cells under microscope

Random variable X:

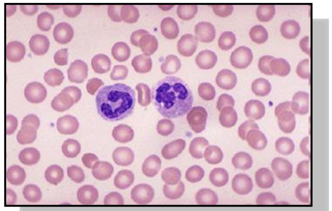
x = 0

x = 1

x = 2

x = 3

…



$$f(x) \geq 0$$

$$\sum f(x) = 1$$

Roll a die

Random variable X:

x = 1

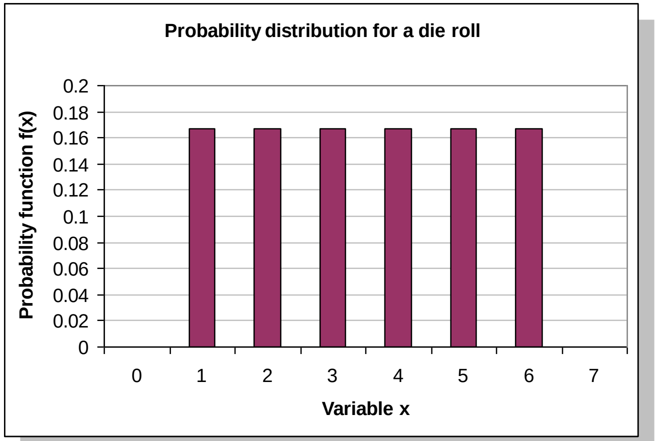
x = 2

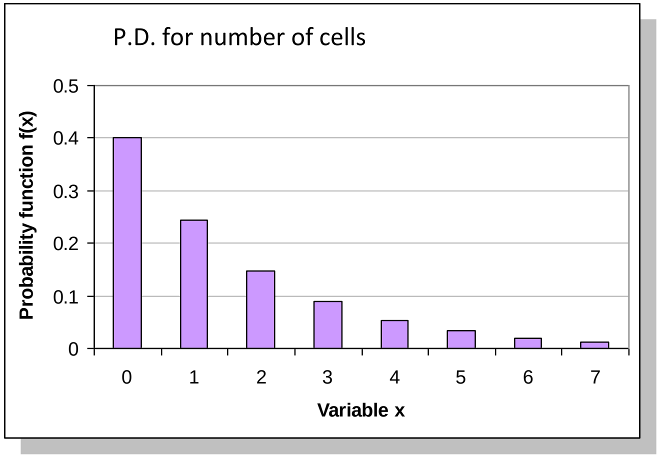
x = 3

x = 4

x = 5

x = 6





Discrete Random Variables

Expected value

A measure of the central location of a random variable, mean.

$$E(x) = \mu = \sum x f(x)$$

Variance

A measure of the variability, or dispersion, of a random variable.

$$\sigma^2 = \sum (x-\mu)^2 f(x)$$

Roll a die

Random variable X:

x = 1

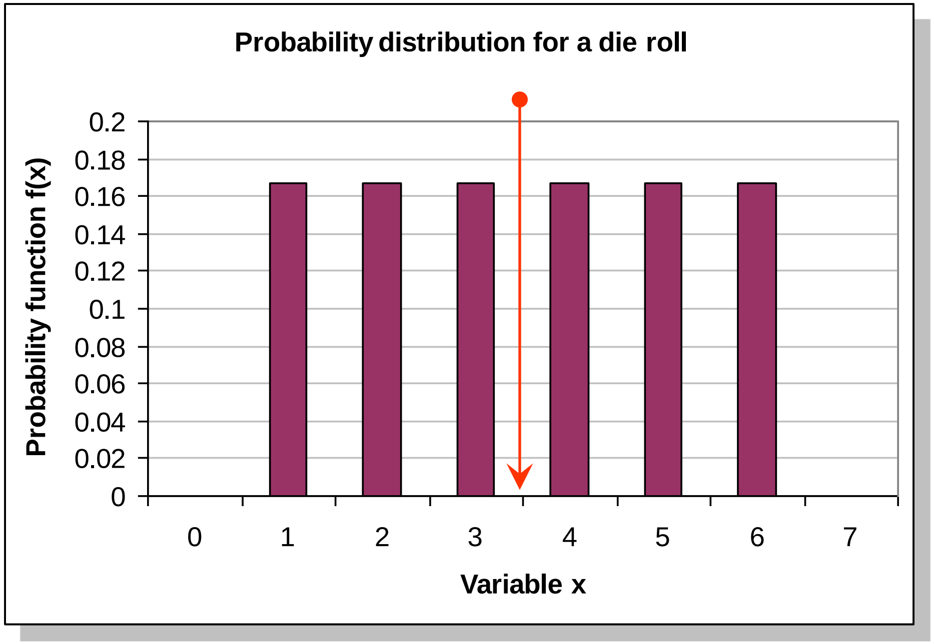
x = 2

x = 3

x = 4

x = 5

x = 6



L2.1. Discrete Probability Distributions

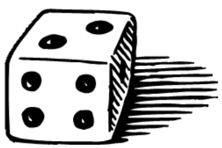
Discrete Uniform Probability Function

Discrete uniform probability distribution

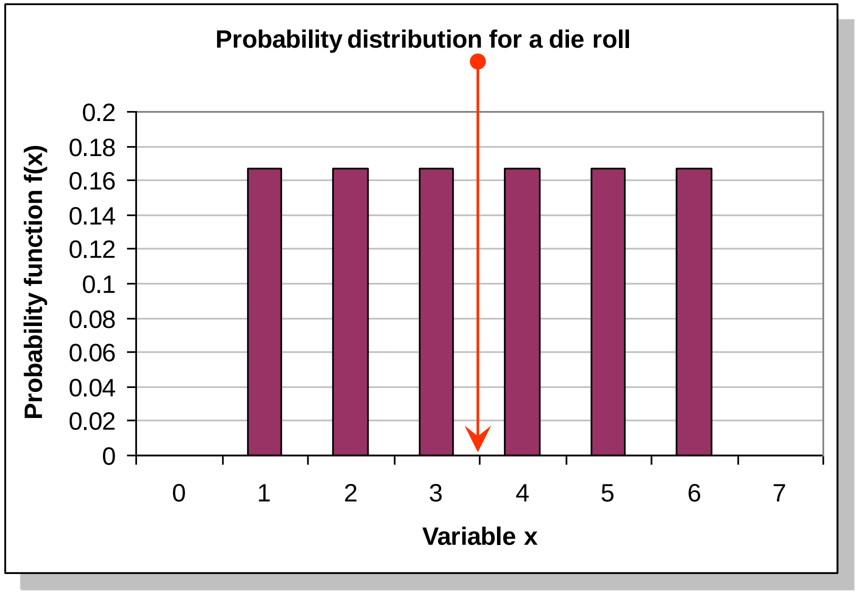
A probability distribution for which each possible value of the random variable has the same probability.

$$f(x) = \frac{1}{n}$$

n – number of values of *x*



x	*f(x)*
1	0.1667
2	0.1667
3	0.1667
4	0.1667
5	0.1667
6	0.1667



$$\mu = \sum(x_i / n) = \sum(x_i) / n$$

$\mu = 3.5$

$\sigma^2 = 2.92$

$\sigma = 1.71$

In R use:

```
ceiling(6*runif( n ))
```

Binomial Distribution

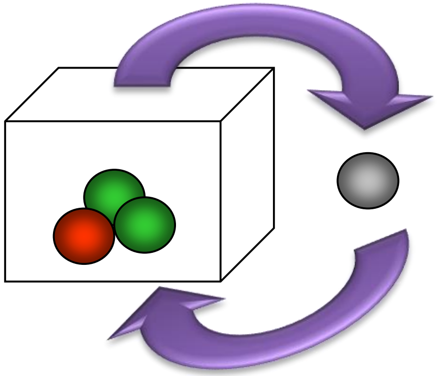
Situation

Assuming that the probability of a side effect for a patient is 0.1. What is the probability that in a group of 3 patients none, 1, 2, or all 3 will get side effects after treatment?

Binomial experiment

An experiment having the four properties:

1. The experiment consists of a sequence of n identical trials.
2. Two outcomes are possible on each trial, one called success and the other failure.
3. The probability of a success p does not change from trial to trial. Consequently, the probability of failure, $1–p$, does not change from trial to trial.
4. The trials are independent.



n trials

Binomial Distribution

Binomial probability distribution

A probability distribution showing the probability of *x* successes in *k* trials of a binomial experiment, when the probability of success *p* does not change in trials.

Probability distribution for a binomial experiment

$$f(x) = C_x^k p^x (1-p)^{(k-x)}$$

$$E(x) = \mu = kp$$

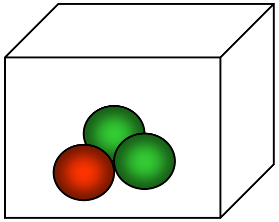
$$Var(x) = \sigma^2 = kp(1-p)$$

$$C_x^k \equiv \binom{k}{x} \equiv \frac{k!}{x!(k-x)!}$$

$$k! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot k$$

$$0! = 1$$

Probability of red *p*(red)=1/3, 3 trials are given. Random variable = number of "red" cases



$$f(2) = \frac{3!}{2!(3-2)!}\left(\frac{1}{3}\right)^2\left(1-\frac{1}{3}\right)^{(3-2)}$$

f(0) = 8/27 = 0.296
f(1) = 4/9 = 0.444
f(2) = 2/9 = 0.222
f(3) = 1/27 = 0.037

Test: $\sum$f(x) = 1

Example: Binomial Distribution

Example

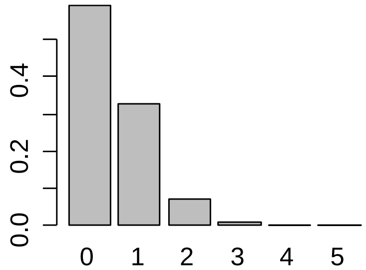
Assuming that the probability of a side effect for a patient is 0.1.

1. What is the probability to get none, 1, 2, etc. side effects in a group of 5 patients?
2. What is the probability that not more then 1 get a side effect
3. What is the expected number of side effects in the group?

$$f(x) = C_x^k p^x (1-p)^{(k-x)}$$

p = 0.1
k = 5

1. `barplot( dbinom(x = 0:5, size = 5, prob = 0.1), names.arg=0:5)`



In Excel use the function:

- `= BINOMDIST(x,n,p,false)`

In R use:

- `= dbinom(…)` – probability distribution
- `= pbinom(…)` – cumul. probability
- `= qbinom(…)` – quantiles
- `= rbinom(…)` – simulate random v.

2. `sum(dbinom(x = c(0,1), size = 5, prob = 0.1))`

3. $\mu = kp = 5*0.1 = 0.5$

Hypergeometric Distribution

Situation

There are 12 mice, of which 5 have an early brain tumor. A researcher randomly selects 3 of 12. What is the probability that none of these 3 has a tumor? What is the probability that more then 1 have a tumor?

Let's denote: ***m*** – mice with tumor, ***n*** – mice without tumor, ***k*** – number of tries

*(usually a different notation is used: **r**, **N-r** and **n**, but here let's be consistent with R)*

Hypergeometric experiment

A probability distribution showing the probability of *x* successes in *k* trials from a population *N=n+m* with *m* successes and *n* failures.

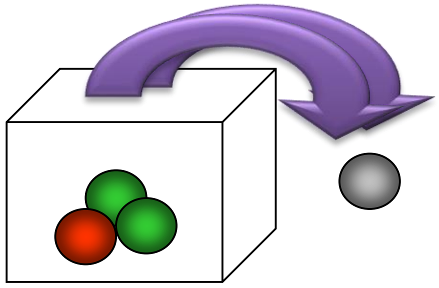
$$E(x) = \mu = k\left(\frac{m}{m+n}\right)$$

$$Var(x) = \sigma^2 = k\left(\frac{m}{m+n}\right)\left(1-\frac{m}{m+n}\right)\left(\frac{m}{m+n-1}\right)$$

$$f(x) = \frac{C_x^m C_{k-x}^n}{C_k^{n+m}}, \quad for\ 0 \le x \le m$$

In Excel use the function:

- **= HYPGEOMDIST (x,k,m,m+n)**



k trials

In R use:

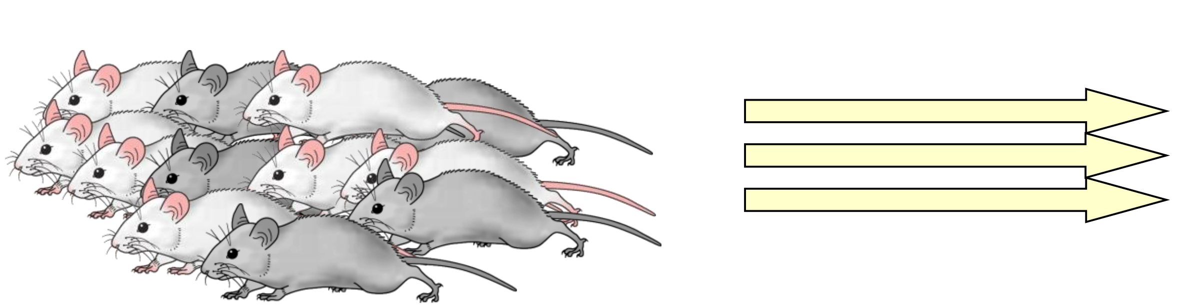
- **= dhyper(…)** – probability distribution
- **= phyper(…)** – cumul. probability
- **= qhyper(…)** – quantiles
- **= rhyper(…)** – simulate random v.

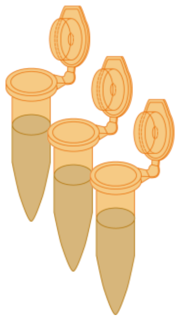
Example: Hypergeometric Distribution

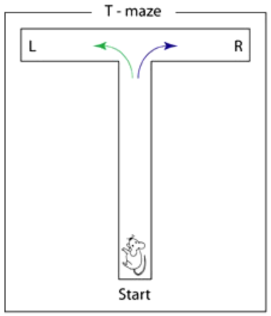
Example

There are 12 mice, of which 5 have an early brain tumor. A researcher randomly selects 3 of 12.

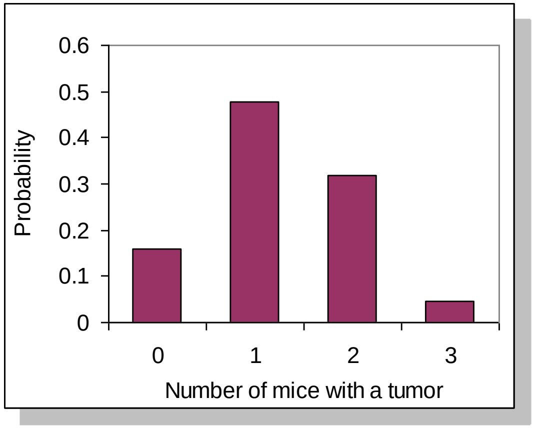
1. What is the probability that none of these 3 has a tumor?
2. What is the probability that more then 1 have a tumor?







x	f(x)
0	0.159
1	0.477
2	0.318
3	0.045



Q1.
P(0) =0.159

Q2.
P(>1) =P(2)+P(3)=0.364

Poisson Distribution

Example

Number of calls to an Emergency Service is on average 3 per hour b/w 2 a.m. and 6 a.m. of working days. What are the probabilities to have 0, 5, 10 calls in the next hour?

Poisson probability distribution

A probability distribution showing the probability of *x* occurrences of an event over a specified interval of time or space.

Poisson probability function

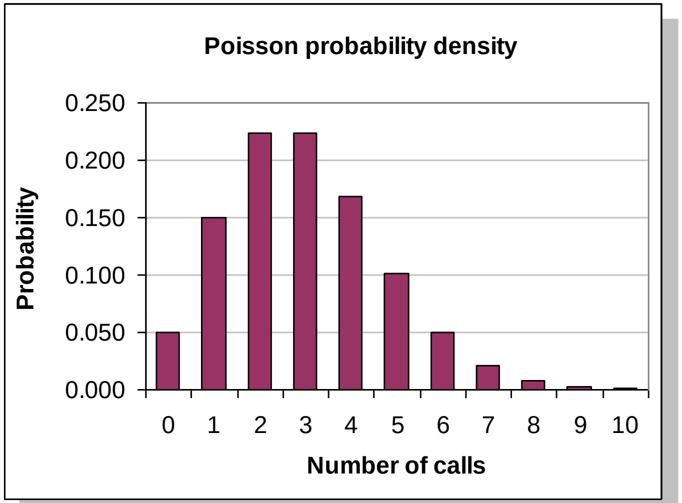
The function used to compute Poisson probabilities.

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

$$\mu = \sigma^2$$

where μ – expected value (mean)

x	f(x)
0	0.050
1	0.149
2	0.224
3	0.224
4	0.168
5	0.101
6	0.050
7	0.022
8	0.008
9	0.003
10	0.001



In Excel use the function:

- `= POISSON(x,mu,false)`

In R use :

- `= dpois(...)` – probability distr.
- `= ppois(...)` – cumul. probability
- `= qpois(...)` – quantiles
- `= rpois(...)` – simulate random v.

Note: lambda = μ in R for Poisson

Example: Poisson Distribution

Example

An ichthyologist studying the *spoonhead sculpin* catches specimens in a large bag seine that she trolls through the lake. She knows from many years experience that on averages she will catch 2 fish per trolling.

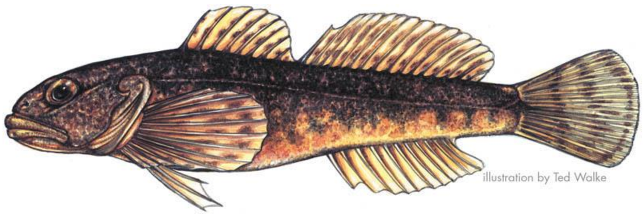
1. Draw distribution

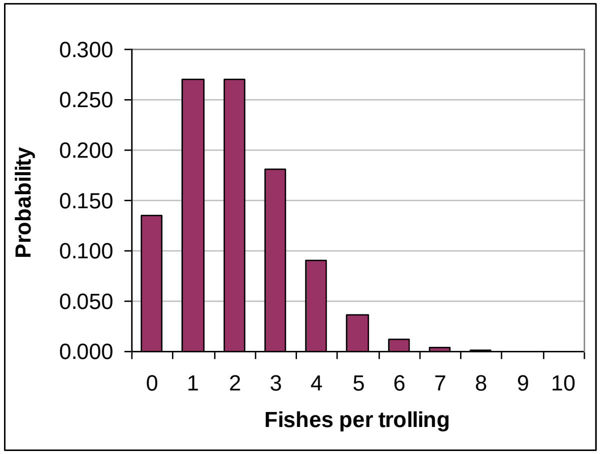
Find the probabilities of catching:

2. No fish;

3. Less than 4 fishes;

4. More then 1 fish.





Q2.
P(0) = 0.135

Q3.
P(<4) = P(0)+P(1)+P(2)+P(3)=0.857

Q4.
P(>1) =1-P(0)-P(1)=0.594

Glover , Mitchell, An Introduction to Biostatistics

Negative Binomial Distribution in R

Example

How many fruitless calls (x) to random respondents should you make in order to complete n=3 surveys? Probability that a random respondent will communicate p=0.2

Negative binomial probability distribution

A probability distribution showing the probability of ***x*** failures until ***n*** successes are achieved. Probability of a success ***p*** is constant.

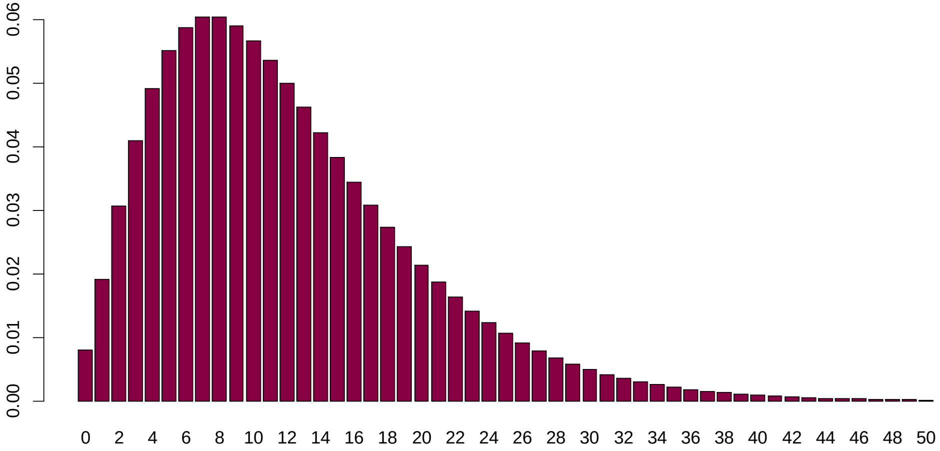
$$f(x) = \frac{\Gamma(x+n)}{\Gamma(n)x!} p^n (1-p)^x$$

$$\mu = \frac{n(1-p)}{p}$$

$$\sigma^2 = \mu + \frac{\mu^2}{n}$$

where μ – expected value (mean), σ^2 - variance

Sometimes total number of successes are measured until certain number of “failures” happens. It is just a matter of definition (see Wikipedia for other definition)



In R use :

- = **dnbinom(…)** – probability distr.
- = **pnbinom(…)** – cumul. probability
- = **qnbinom(…)** – quantiles
- = **rnbinom(…)** – simulate random v.

L2.1. Discrete Probability Distributions


```
####################################################################
# L2.1. DISCRETE PROBABILITY DISTRIBUTIONS
####################################################################

##------------------------------------------------------------------
## L2.1.1. Discrete uniform distribution
##------------------------------------------------------------------
## generate n=10 experiments with a rolling die
n=10
ceiling(6*runif(n))

##------------------------------------------------------------------
## L2.1.2. Binomial distribution
##------------------------------------------------------------------
## Assuming that the probability of a side effect for a patient
## is 0.1. What is the prob. to get 0, 1, etc. side effects in a
## group of 5 patients?
dbinom(x = 0:5, size = 5, prob = 0.1)

barplot( dbinom(x = 0:5, size = 5, prob = 0.1), names.arg=0:5)

## What is the probability that not more then 1 get a side effect
sum(dbinom(x = c(0,1), size = 5,prob = 0.1))

## What is the expected number of side effects in the group?
5*0.1 = 0.5

##------------------------------------------------------------------
## L2.1.3. Hypergeometric distribution
##------------------------------------------------------------------
## There are 12 mice, of which 5 have an early brain tumor.
## A researcher randomly selects 3 of 12.
barplot( dhyper(x=0:3, k=3, m=5, n=12-5), names.arg=0:3)

## What is the probability that none of these 3 has a tumor?
dhyper(x=0, k=3, m=5, n=12-5)

## What is the probability that more then 1 have a tumor?
sum(dhyper(x=c(2,3), k=3, m=5, n=12-5))

##------------------------------------------------------------------
## L2.1.4. Poisson distribution
##------------------------------------------------------------------
## An ichthyologist studying the spoonhead sculpin catches
## specimens in a large bag seine that she trolls through the lake.
## She knows from many years experience that on averages she will
## catch 2 fish per trolling.
m = 2
## Draw distribution
barplot( dpois(c(0:10),lambda=m), names.arg=0:10)

## Find the probabilities of catching:  No fish;
dpois(0,lambda=m)

## Find the probabilities of catching:  less then 4 fishes;
sum(dpois(c(0,1,2,3),lambda=m))

## Find the probabilities of catching:  more then 1 fish;
1-sum(dpois(c(0,1),lambda=m))

##------------------------------------------------------------------
## L2.1.5. Negative binomial distribution
##------------------------------------------------------------------
## How many fruitless calls (x) to random respondents should you
## make in order to complete  n=3 surveys? Probability that a
## random respondent will communicate  p=0.2

n = 3
p = 0.2
## Draw distribution
barplot( dnbinom(c(0:50),size=n,prob = p), names.arg=0:50,
col="#880044")

#>>>>>>>>>>>>>>>>>>>>>>>>>>>
#> please, do Task L2.1
#>>>>>>>>>>>>>>>>>>>>>>>>>>>
```

Task L2.1

L2.2. Continuous Probability Distributions

Random Variables

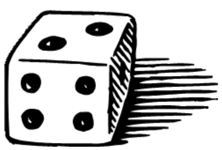
Random variable
A numerical description of the outcome of an experiment.

A random variable is always a numerical measure.

Discrete random variable
A random variable that may assume either a finite number of values or an infinite sequence of values.

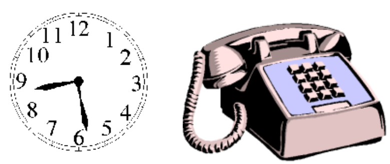
Continuous random variable
A random variable that may assume any numerical value in an interval or collection of intervals.

Roll a die

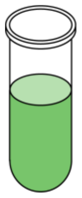


Number of calls to a reception per hour

Time between calls to a reception



Volume of a sample in a tube

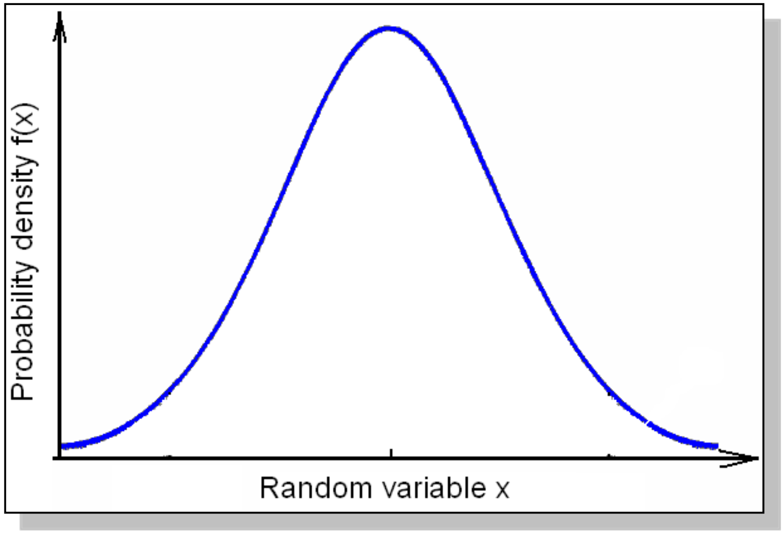


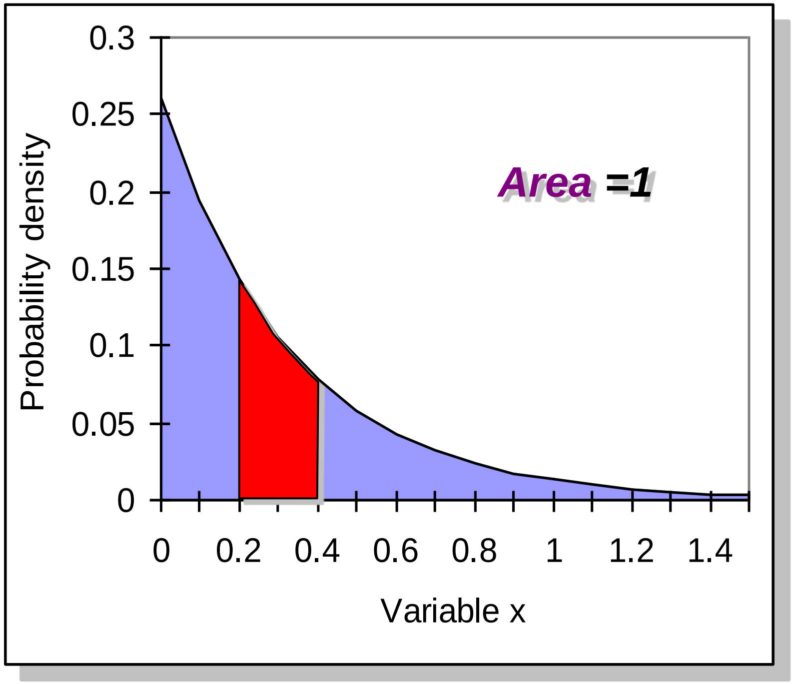
Weight, height, blood pressure, etc

Probability Density

Probability density function

A function used to compute probabilities for a continuous random variable. The area under the graph of a probability density function over an interval represents probability.





$$\int_x f(x) = 1$$

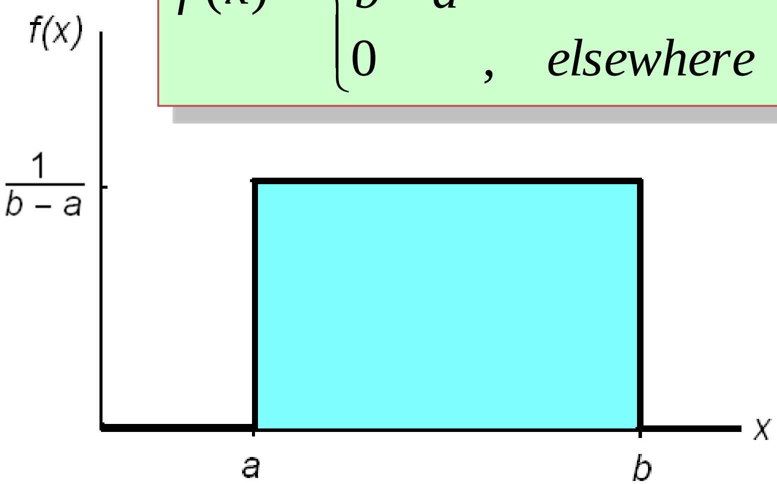
L2.2. Continuous Probability Distributions

Probability Density

Uniform probability distribution

A continuous probability distribution for which the probability that the random variable will assume a value in any interval is the same for each interval of equal length.

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for } a \le x \le b \\ 0 \quad , & \text{elsewhere} \end{cases}$$

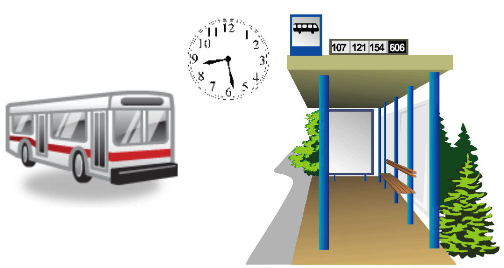


$$Var(x) = \sigma^2 = \frac{(b-a)^2}{12}$$

$$E(x) = \mu = \frac{a+b}{2}$$

In R use :

- = **dunif(…)** –probability density
- = **punif(…)** –cumul. probability
- = **qunif(…)** –quantiles
- = **runif(…)** –simulate random var.



Example

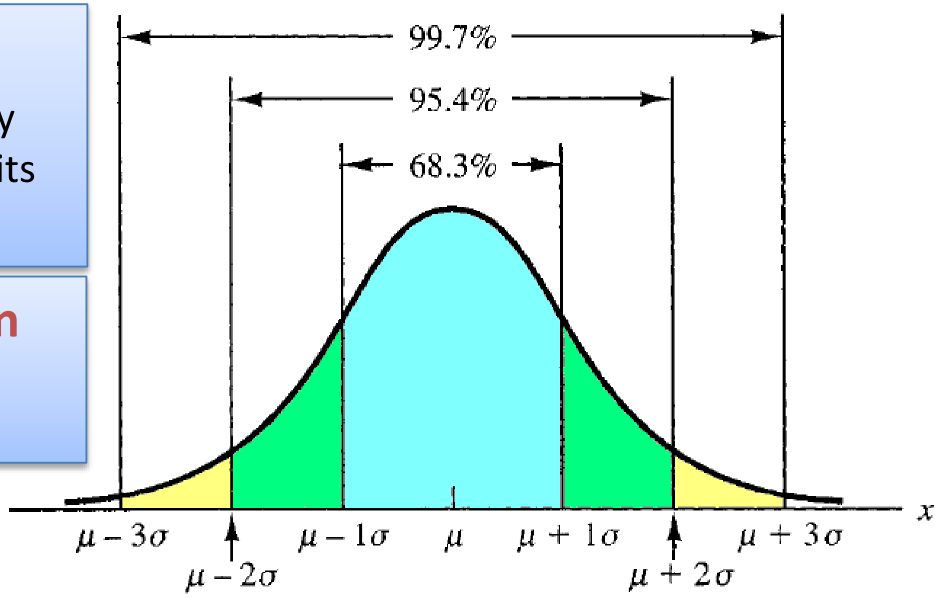
The bus 3 goes every 8 minutes. You are coming to CHL bus station, having no idea about precise timetable. What is the distribution for the time, you may wait there?

Normal Distribution

Normal probability distribution
A continuous probability distribution. Its probability density function is bell shaped and determined by its mean μ and standard deviation σ.

Standard normal probability distribution
A normal distribution with a mean of zero and a standard deviation of one.

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



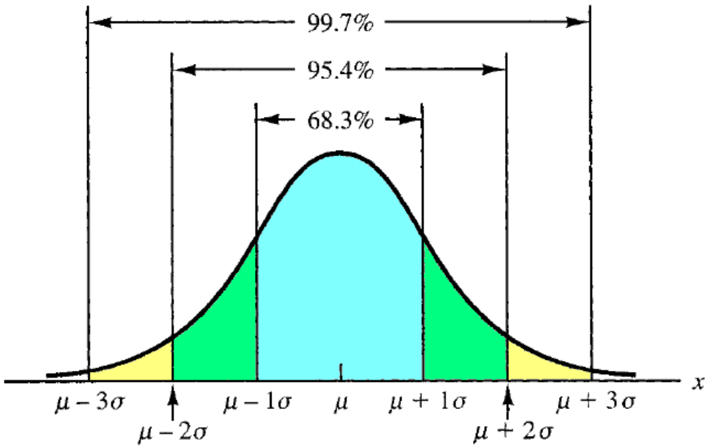
In R use :
- `= dnorm(…)` – probability density
- `= pnorm(…)` – cumul. probability
- `= qnorm(…)` – quantiles
- `= rnorm(…)` – simulate random var.

Example: Normal Distribution

Example

The volume of a liquid in bottles of one company is distributed normally with an expected value of 0.33 liter, and standard deviation of 0.01.

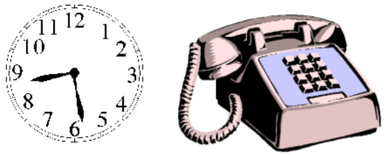
1. *Draw probability density function*
2. *Estimate the probability to buy a bottle containing less than 0.31 liters of the liquid.*
3. *Estimate the interval, in which the volumes of 95% bottles are lying (many correct solutions exist!)*



Exponential Distribution

Exponential probability distribution
A continuous probability distribution that is useful in computing probabilities for the time between independent random events.

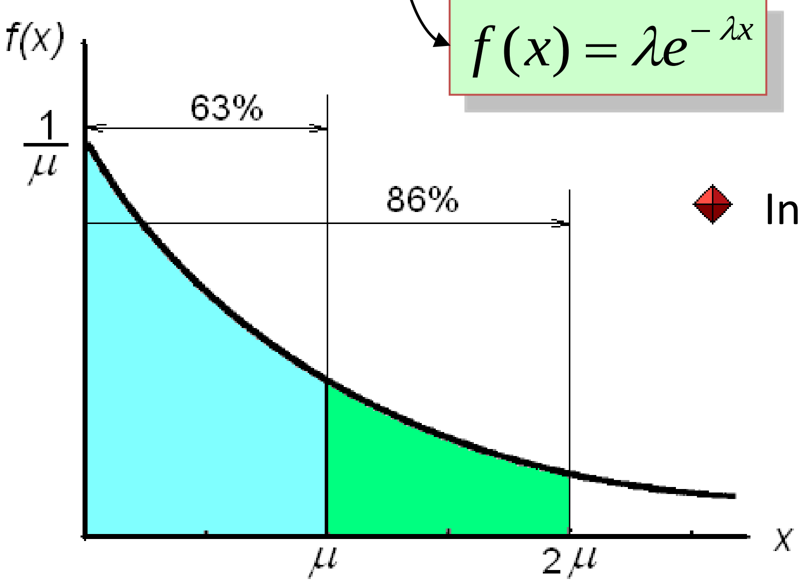
Time between calls to a reception



$$\mu = \frac{1}{\lambda} = \sigma$$

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \qquad for\ x \geq 0, \mu > 0$$

$$f(x) = \lambda e^{-\lambda x}$$



- In R exponential distribution is defined by **rate** $\lambda = 1/\mu$

In R use :

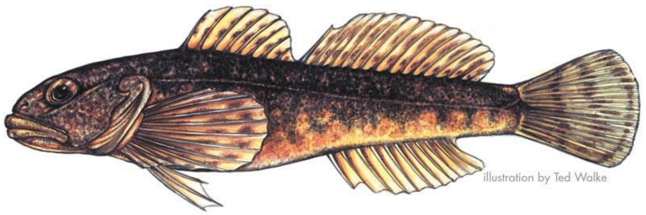
- `= dexp (...)` – probability density
- `= pexp (...)` – cumul. probability
- `= qexp (...)` – quantiles
- `= rexp (...)` – simulate random var.

Example: Exponential Distribution

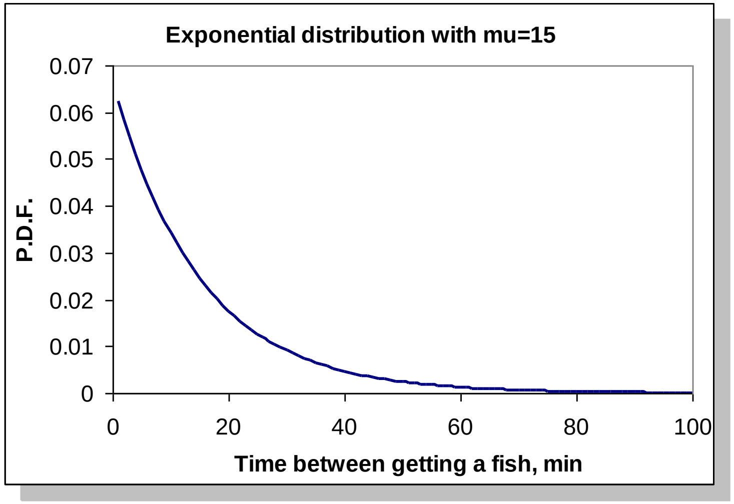
Example

An ichthyologist studying the *spoonhead sculpin* catches specimens in a large bag seine that she trolls through the lake. She knows from many years experience that on averages she will catch 2 fishes per trolling. Each trolling take ~30 minutes.

Find the probability of catching no fish in the next hour



1. Let's calculate rate λ for this situation: $\lambda = 2 / 30 = 1/15$ minutes $^{-1}$



2.Calculate:

$$P(x \geq 60) = 1 - P(x \leq 60) = 1 - F(60) = e^{-\frac{60}{15}} \approx 0.02$$

L2.2. Continuous Probability Distributions

UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

```
################################################################################
# L2.2. CONTINUOUS PROBABILITY DISTRIBUTIONS
################################################################################

##------------------------------------------------------------------------------
## L2.2.1. Uniform distribution
##------------------------------------------------------------------------------
## Values from uniform distribution b/w 1 and 2:
runif(n=10,min=1, max=2)

##------------------------------------------------------------------------------
## L2.2.2. Normal distriburion
##------------------------------------------------------------------------------
## The volume of a liquid in bottles of one company is distributed normally with
## an expected value of 0.33 liter, and standard deviation of 0.01.

## Draw probability density function
x11()
x=seq(0.29,0.37,by=0.0001)
plot(x,dnorm(x,mean=0.33,sd=0.01),type="l",lwd=2)

## Estimate the probability to buy a bottle with > 0.31 liters of the liquid
pnorm(q=0.31,mean=0.33, sd=0.01)

## Estimate the interval, in which the volumes of 95% bottles are lying
## from 5% percentile to +Inf
c(qnorm(p=0.05,mean=0.33, sd=0.01),Inf)
## from 0 to 95% percentile
c(0, qnorm(p=0.95,mean=0.33, sd=0.01))
## (optimal) b/w 2.5% and 97.5% percentiles
c(qnorm(p=0.025,mean=0.33, sd=0.01), qnorm(p=0.975,mean=0.33, sd=0.01))

##------------------------------------------------------------------------------
## L2.2.3. Exponential distribution
##------------------------------------------------------------------------------
## An ichthyologist studying the spoonhead sculpin catches specimens in a large
## bag seine that she trolls through the lake. She knows from many years
## experience that on averages she will catch 2 fishes per trolling. Each
## trolling take ~30 minutes.

## Draw probability density function
x11()
x=0:100
plot(x,dexp(x,rate=1/15),type="l",lwd=2)

## Find the probability of catching no fish in the next hour
##   if pexp(60,rate=1/15) - prob to cach a fishe before 60 minutes, then
1-pexp(60,rate=1/15)

#>>>>>>>>>>>>>>>>>>>>>>>>>>>
#> please, do Task L2.2
#>>>>>>>>>>>>>>>>>>>>>>>>>>>
```

Task L2.2

L2.3. Sampling and Sampling Distribution

Population and Sample

Population parameter

A numerical value used as a summary measure for a population (e.g., the population mean μ, variance σ^2, standard deviation σ)

POPULATION

μ – mean

σ^2 – variance

N – number of elements (usually $N=\infty$)

SAMPLE

m, $\bar{x}$ – mean

s^2 – variance

n – number of elements

Sample statistic

A numerical value used as a summary measure for a sample (e.g., the sample mean m, sample variance s^2, and sample standard deviation s)

All existing laboratory *Mus musculus*

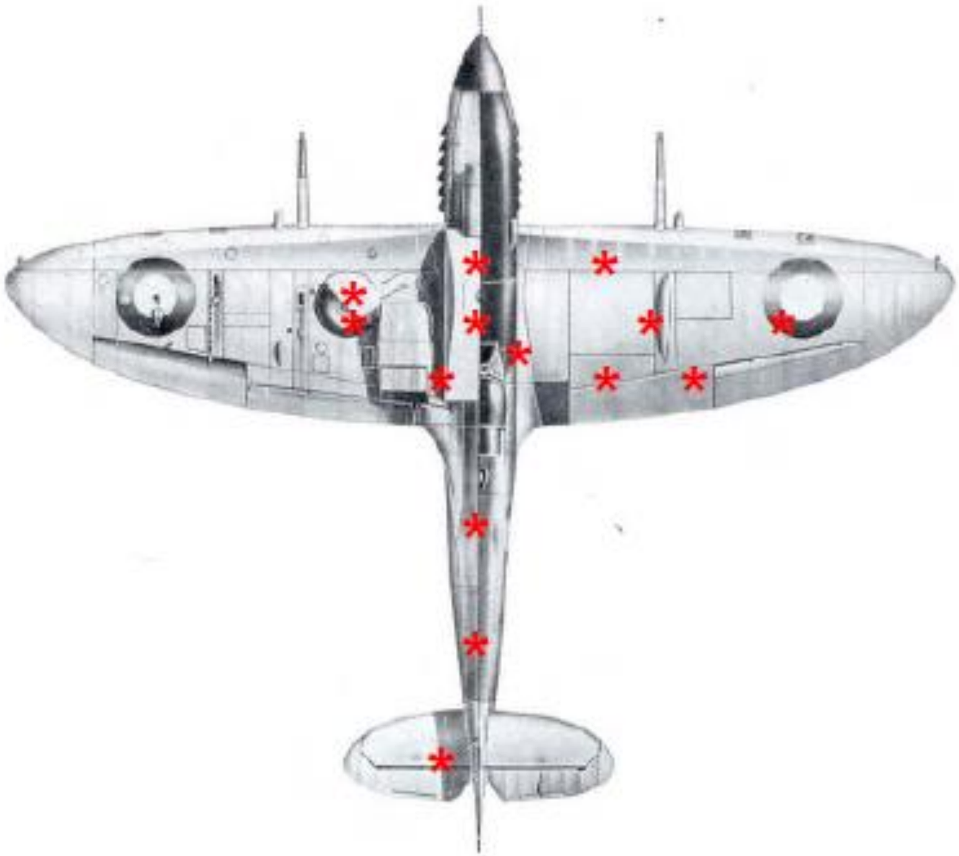
mice.txt

790 mice from different strains

http://phenome.jax.org

ID	Strain	Sex	Starting age	Ending age	Starting weight	Ending weight	Weight change	Bleeding time	Ionized Ca in blood	Blood pH	Bone mineral density	Lean tissues weight	Fat weight
1	129S1/SvImJ	f	66	116	19.3	20.5	1.062	64	1.2	7.24	0.0605	14.5	4.4
2	129S1/SvImJ	f	66	116	19.1	20.8	1.089	78	1.15	7.27	0.0553	13.9	4.4
3	129S1/SvImJ	f	66	108	17.9	19.8	1.106	90	1.16	7.26	0.0546	13.8	2.9
368	129S1/SvImJ	f	72	114	18.3	21	1.148	65	1.26	7.22	0.0599	15.4	4.2
369	129S1/SvImJ	f	72	115	20.2	21.9	1.084	55	1.23	7.3	0.0623	15.6	4.3
370	129S1/SvImJ	f	72	116	18.8	22.1	1.176		1.21	7.28	0.0626	16.4	4.3
371	129S1/SvImJ	f	72	119	19.4	21.3	1.098	49	1.24	7.24	0.0632	16.6	5.4
372	129S1/SvImJ	f	72	122	18.3	20.1	1.098	73	1.17	7.19	0.0592	16	4.1
4	129S1/SvImJ	f	66	109	17.2	18.9	1.099	41	1.25	7.29	0.0513	14	3.2
5	129S1/SvImJ	f	66	112	19.7	21.3	1.081	129	1.14	7.22	0.0501	16.3	5.2
10	129S1/SvImJ	m	66	112	24.3	24.7	1.016	119	1.13	7.24	0.0533	17.6	6.8
364	129S1/SvImJ	m	72	114	25.3	27.2	1.075	64	1.25	7.27	0.0596	19.3	5.8
365	129S1/SvImJ	m	72	115	21.4	23.9	1.117	48	1.25	7.28	0.0563	17.4	5.7
366	129S1/SvImJ	m	72	118	24.5	26.3	1.073	59	1.25	7.26	0.0609	17.8	7.1
367	129S1/SvImJ	m	72	122	24	26	1.083	69	1.29	7.26	0.0584	19.2	4.6
6	129S1/SvImJ	m	66	116	21.6	23.3	1.079	78	1.15	7.27	0.0497	17.2	5.7
7	129S1/SvImJ	m	66	107	22.7	26.5	1.167	90	1.18	7.28	0.0493	18.7	7
8	129S1/SvImJ	m	66	108	25.4	27.4	1.079	35	1.24	7.26	0.0538	18.9	7.1
9	129S1/SvImJ	m	66	109	24.4	27.5	1.127	43	1.29	7.29	0.0539	19.5	7.1

Be Careful with Sampling!



Were to put additional protection?

L2.3. Sampling and Sampling Distribution

Making a Random Sampling

`mice.txt`

790 mice from different strains

http://phenome.jax.org

ID	Strain	Sex	Starting age	Ending age	Starting weight	Ending weight	Weight change	Bleeding time	Ionized Ca in blood	Blood pH	Bone mineral density	Lean tissues weight	Fat weight
1	129S1/SvImJ	f	66	116	19.3	20.5	1.062	64	1.2	7.24	0.0605	14.5	4.4
2	129S1/SvImJ	f	66	116	19.1	20.8	1.089	78	1.15	7.27	0.0553	13.9	4.4
3	129S1/SvImJ	f	66	108	17.9	19.8	1.106	90	1.16	7.26	0.0546	13.8	2.9
368	129S1/SvImJ	f	72	114	18.3	21	1.148	65	1.26	7.22	0.0599	15.4	4.2
369	129S1/SvImJ	f	72	115	20.2	21.9	1.084	55	1.23	7.3	0.0623	15.6	4.3
370	129S1/SvImJ	f	72	116	18.8	22.1	1.176		1.21	7.28	0.0626	16.4	4.3
371	129S1/SvImJ	f	72	119	19.4	21.3	1.098	49	1.24	7.24	0.0632	16.6	5.4
372	129S1/SvImJ	f	72	122	18.3	20.1	1.098	73	1.17	7.19	0.0592	16	4.1
4	129S1/SvImJ	f	66	109	17.2	18.9	1.099	41	1.25	7.29	0.0513	14	3.2
5	129S1/SvImJ	f	66	112	19.7	21.3	1.081	129	1.14	7.22	0.0501	16.3	5.2
10	129S1/SvImJ	m	66	112	24.3	24.7	1.016	119	1.13	7.24	0.0533	17.6	6.8
364	129S1/SvImJ	m	72	114	25.3	27.2	1.075	64	1.25	7.27	0.0596	19.3	5.8
365	129S1/SvImJ	m	72	115	21.4	23.9	1.117	48	1.25	7.28	0.0563	17.4	5.7
366	129S1/SvImJ	m	72	118	24.5	26.3	1.073	59	1.25	7.26	0.0609	17.8	7.1
367	129S1/SvImJ	m	72	122	24	26	1.083	69	1.29	7.26	0.0584	19.2	4.6
6	129S1/SvImJ	m	66	116	21.6	23.3	1.079	78	1.15	7.27	0.0497	17.2	5.7
7	129S1/SvImJ	m	66	107	22.7	26.5	1.167	90	1.18	7.28	0.0493	18.7	7
8	129S1/SvImJ	m	66	108	25.4	27.4	1.079	35	1.24	7.26	0.0538	18.9	7.1
9	129S1/SvImJ	m	66	109	24.4	27.5	1.127	43	1.29	7.29	0.0539	19.5	7.1

1. Load the data

2. Use `sample()` to get either a sample of a column or index for the matrix

```
sample(Mice$Ending.weight, size=20)
# or
idx = sample(1:nrow(Mice), size=20)
```

3. Assume that these mice is a population with size N=790. Build 5 samples with n=20

4. Calculate m, s for ending weight and p – proportion of males for each sample

Point estimator

The sample statistic, such as m, s, or p, that provides the point estimation the population parameters μ, σ, π.

In R use :

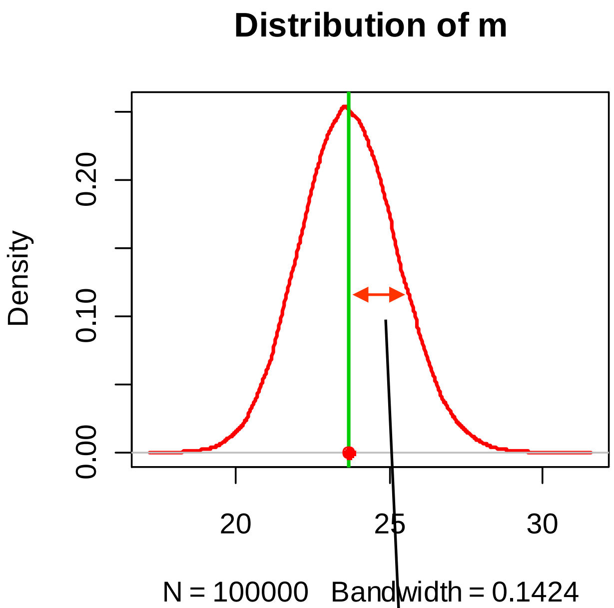
◆ = **`sample(…)`**

standard random sampling – without replacement (do not pick the same object twice)

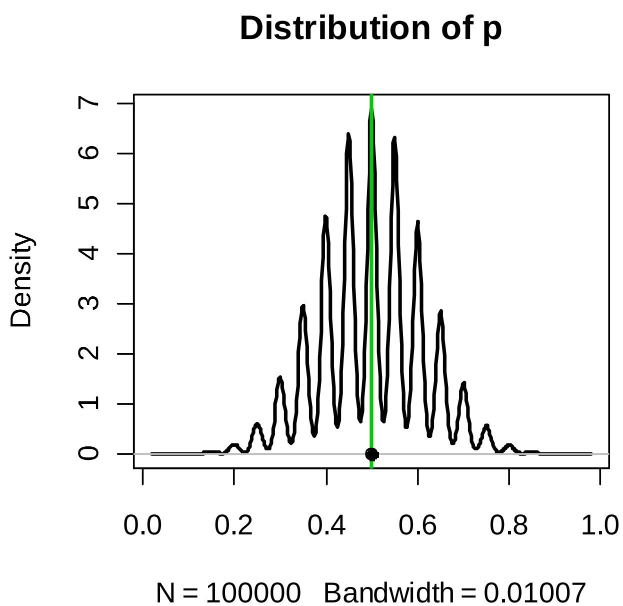
Making a Random Sampling

Sampling distribution

A probability distribution consisting of all possible values of a sample statistic.



$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$



$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

$$E(m) = \mu$$

$$E(p) = \pi$$

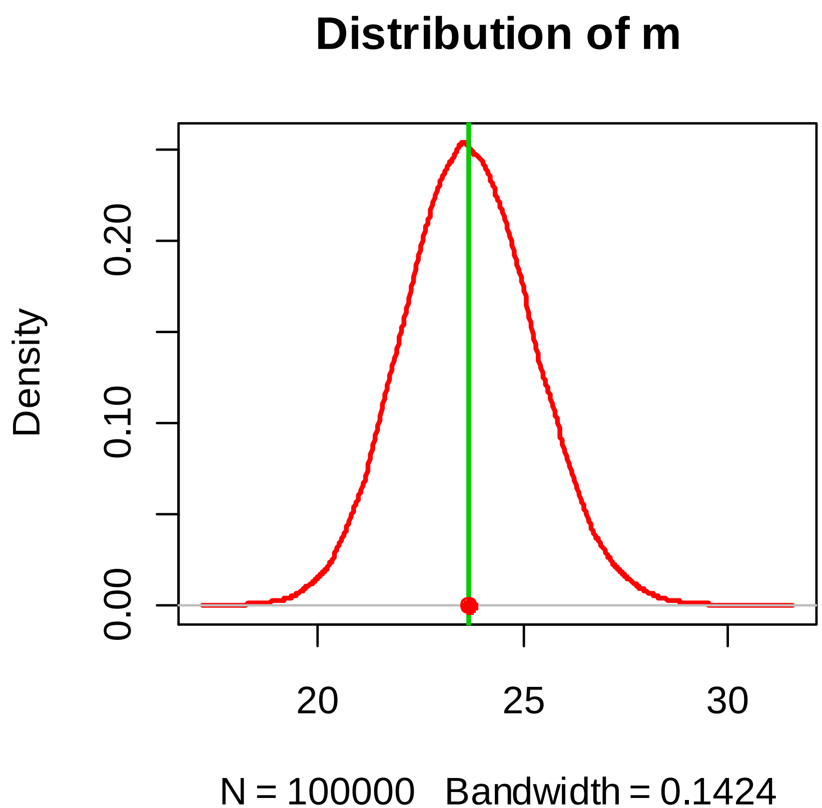
Standard error

The standard deviation of a point estimator.

Unbiased Point Estimator: mean

Unbiased

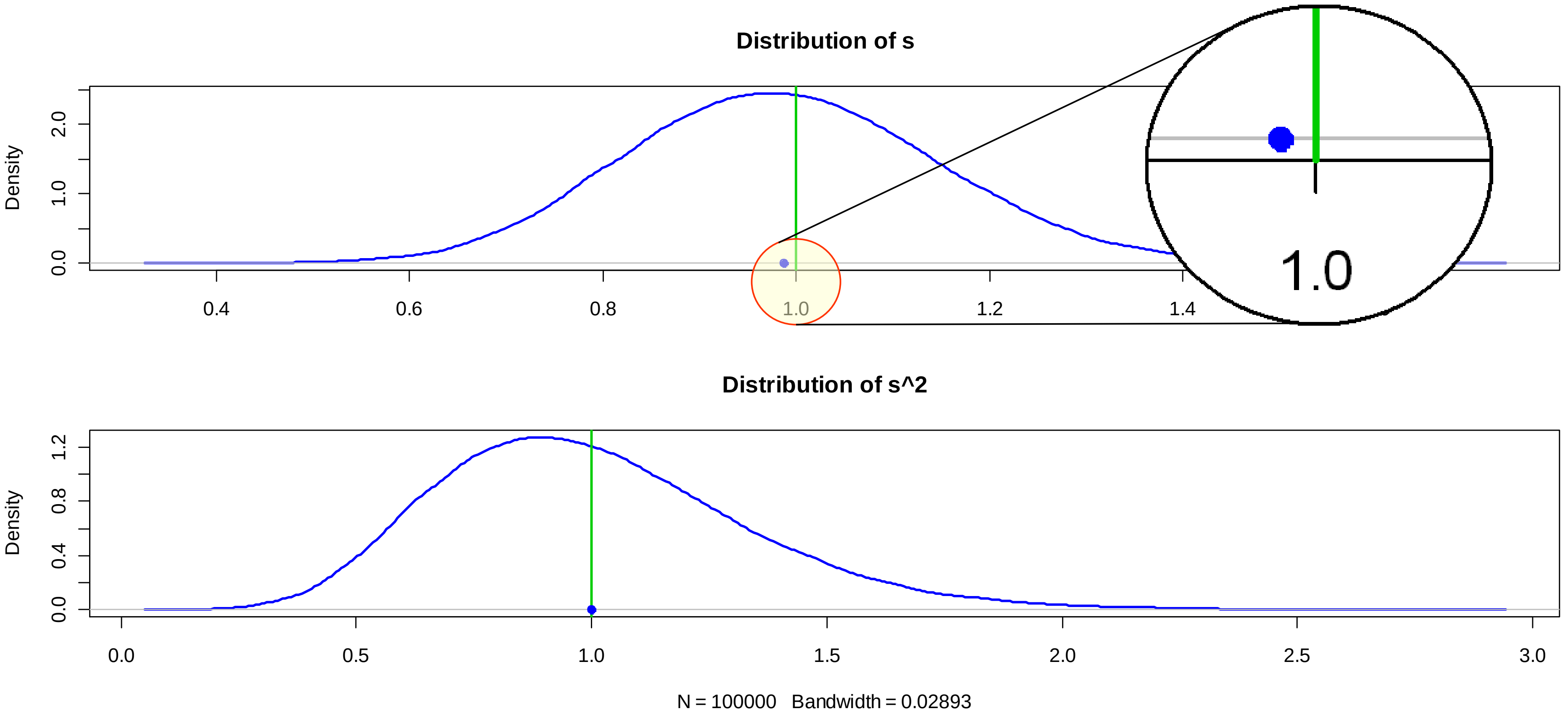
A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.



Unbiased Point Estimator: variance

Unbiased

A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.

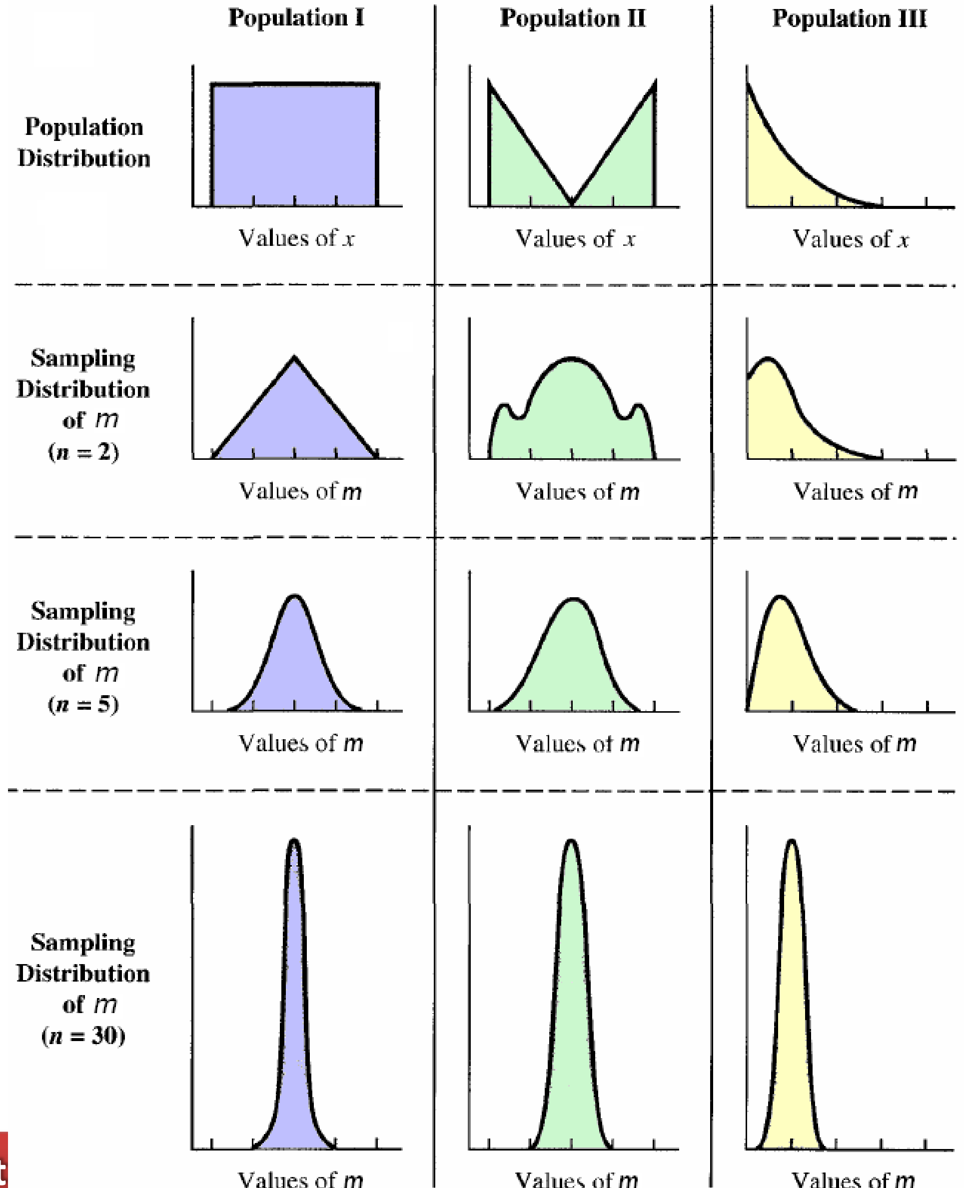


Central Limit Theorem

Central limit theorem

In selecting simple random sample of size *n* from a population, the *sampling distribution of the sample mean m* ***can be approximated by a normal distribution*** as the sample size becomes large

In practice, if the sample size is *n*>30, the normal distribution is a good approximation for the sample mean for any initial distribution.

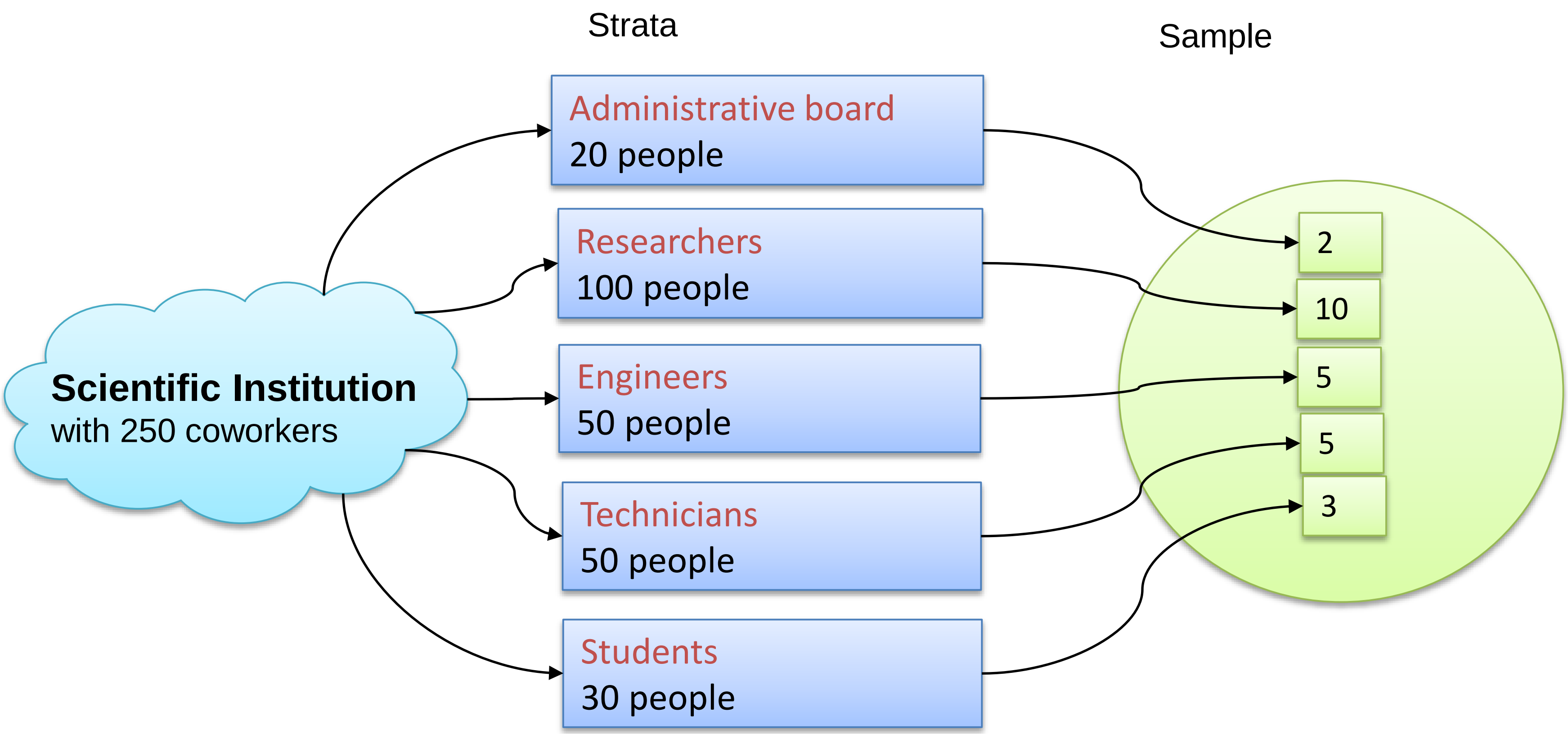


L2.3. Sampling and Sampling Distribution

Other Sampling Methods: Stratified

Stratified random sampling

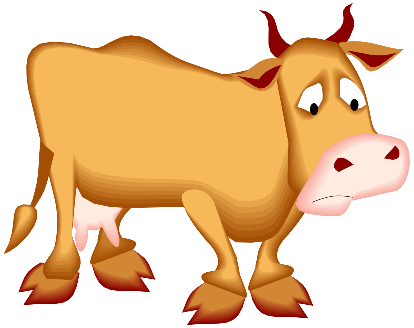
A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.



The Wisdom of the Crowd

The wisdom of the crowd
is the process of taking into account the collective opinion of a group of individuals rather than a single expert to answer a question. A large group's aggregated answers to questions involving quantity estimation has generally been found to be as good as, and often better than, the answer given by any of the individuals within the group.

The classic wisdom-of-the-crowds finding involves point estimation of a continuous quantity. At a 1906 country fair in Plymouth, eight hundred people participated in a contest to estimate the weight of a slaughtered and dressed ox. Statistician **Francis Galton** observed that the median guess, 1207 pounds, was accurate within 1% of the true weight of 1198 pounds.



http://www.youtube.com/watch?v=r-FonWBEb0o

L2.3. Sampling and Sampling Distribution

```
################################################################################
# L2.3. SAMPLING AND SAMPLING DISTRIBUTION
################################################################################

## load the data
Mice=read.table("http://edu.sablab.net/data/txt/mice.txt",header=T,sep="\t")
str(Mice)

sample(Mice$Ending.weight, size=3)

## run 5 times. see the variability in m and s
idx = sample(1:nrow(Mice), size=20)
mean(Mice$Ending.weight[idx])
sd(Mice$Ending.weight[idx])
```

Task L2.3.

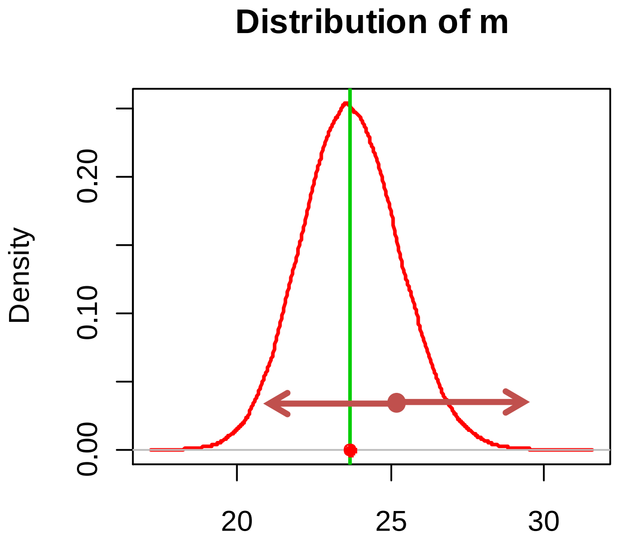
Interval Estimation

Interval estimate

An estimate of a population parameter that provides an interval believed to contain the value of the parameter. For the interval estimates of mean and proportion it has the form: **point estimate ± margin of error.**

Margin of error

The ± value added to and subtracted from a point estimate in order to develop an interval estimate of a population parameter.



$$\mu = m \pm merror$$

σ known

The condition existing when historical data or other information provides a good value for the population standard deviation prior to taking a sample. The interval estimation procedure uses this known value of σ in computing the margin of error.

σ unknown

The condition existing when no good basis exists for estimating the population standard deviation prior to taking the sample. The interval estimation procedure uses the sample standard deviation ***s*** in computing the margin of error.

Confidence and Confidence Interval

Confidence level

The confidence associated with an interval estimate. For example, if an interval estimation procedure provides intervals such that 95% of the intervals formed using the procedure will include the population parameter, the interval estimate is said to be constructed at the 95% confidence level.

Confidence interval

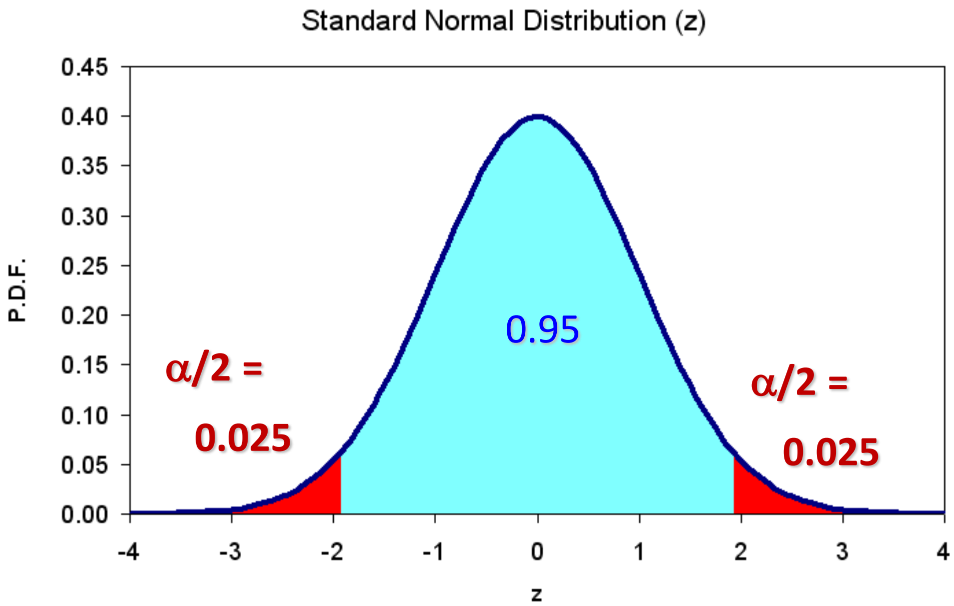
Another name for an interval estimate.

$$\mu = m \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

For 95 % confidence α = 0.05, which means that in each tail we have 0.025. Corresponding $z_{\alpha/2}$ = 1.96

In R use (for $z_{\alpha/2}$):

```
= -qnorm(α/2)
```

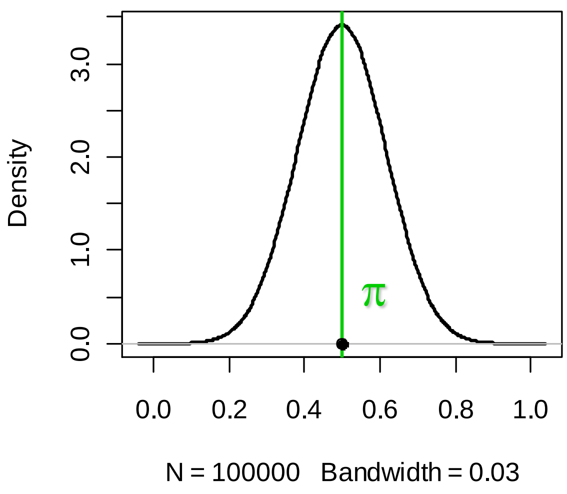


L2.4. Interval Estimation

Confidence and Confidence Interval

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \rightarrow \sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Sampling distribution for proportion *p*



$$\pi = p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

if $np \geq 5$ and $n(1-p) \geq 5$

Practical Work

`pancreatitis.txt`

Define a 95% confidence interval for never-smoking proportion of people coming to a hospital

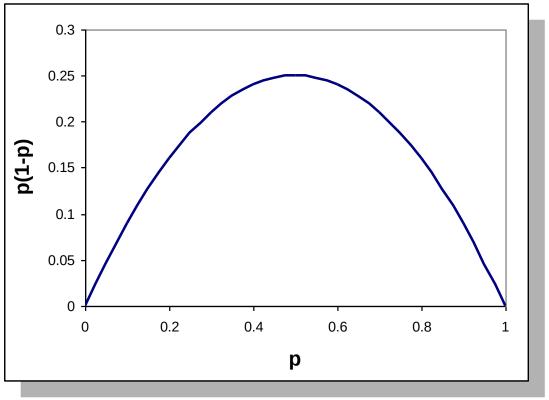
n=	270
p(never)=	**0.214815**
sp=	0.024994
E=	**0.048988**

for 95% confidence $z_{0.025} = 1.96$

$\pi = 21.5 \pm 4.9$ %

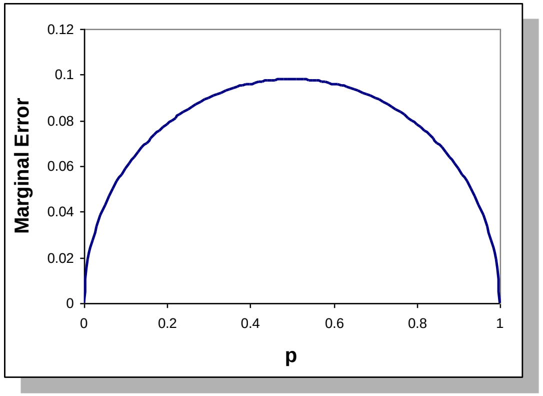
Population Proportion: Some Practical Aspects

$$\pi = p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$



1. The normal distribution is applicable only when enough data points are observed. The rule of thumb is: $np \geq 5$ and $n(1-p) \geq 5$

2. The maximal marginal error is observed when p=0.5



3. The estimation of the sample size can be obtained:

$$n = \frac{z_{\alpha/2}^2\, p(1-p)}{E^2}$$

$np \geq 5$ and $n(1-p) \geq 5$

where p is a best guess for π or the result of a preliminary study

L2.4. Interval Estimation

Population Mean: σ Unknown

Assume that we have a sample of 20 mice and would like to estimate an average size of a mice in population.

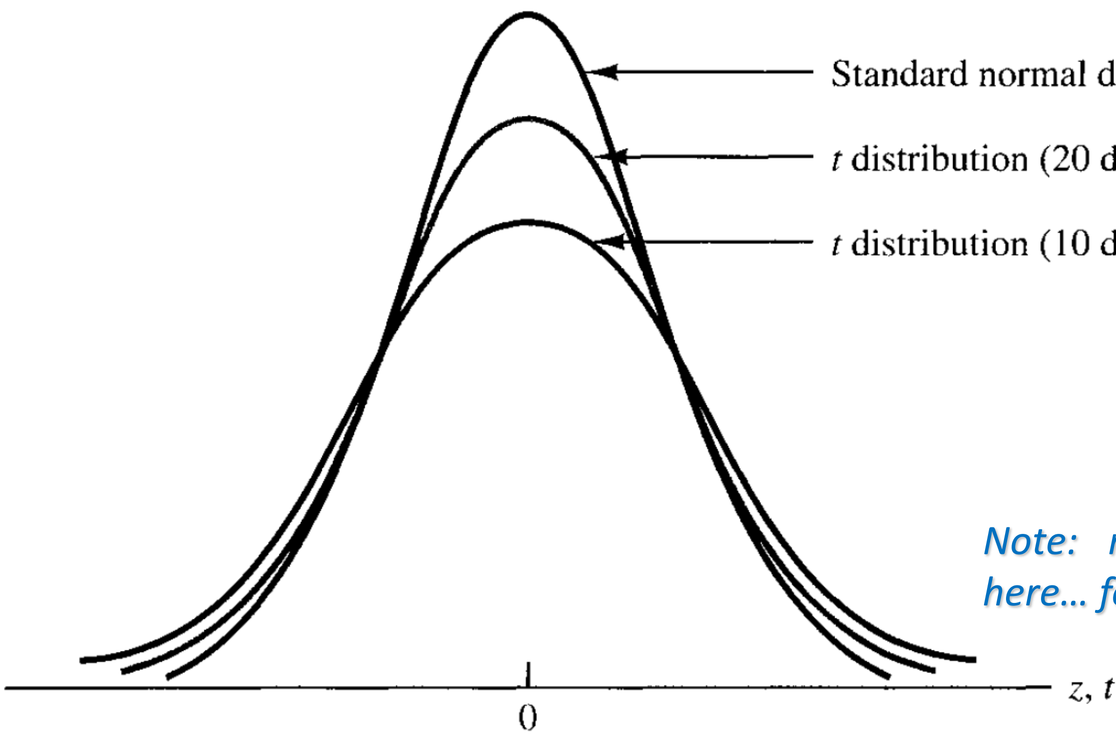
Weight
39.9
19.8
32.4
21
27.5
20.8
21.3
40
10.7
22.6
27
10.8
20.9
14.7
31.4
17.2
11.4
19.1
31.3
14.8

$m = 22.73$

$s = 8.84$

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

As we replace $\sigma \to s$, we introduce an additional error and this change the distribution from z to t **(Student)**



Note: not a realistic scale here… for illustration only

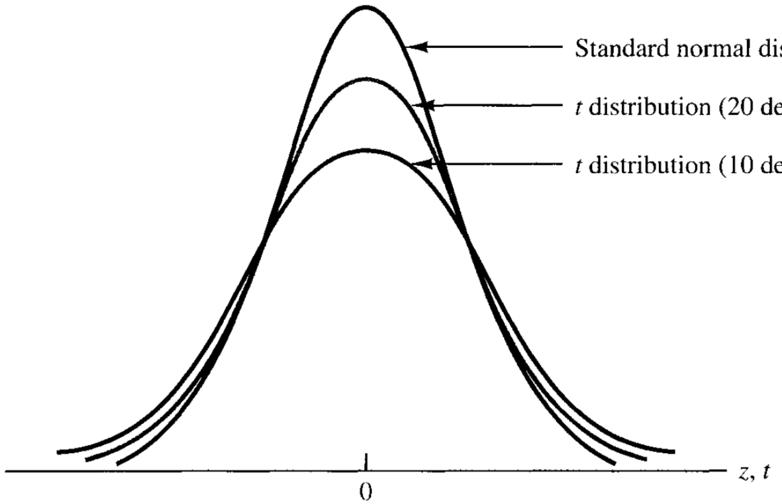
Population Mean: σ Unknown

t-distribution

A family of probability distributions that can be used to develop an interval estimate of a population mean whenever the population standard deviation σ is unknown and is estimated by the sample standard deviation s.

Degrees of freedom

A parameter of the *t*-distribution. When the *t* distribution is used in the computation of an interval estimate of a population mean, the appropriate *t* distribution has $n - 1$ degrees of freedom, where n is the size of the simple random sample.



L2.4. Interval Estimation

Population Mean: σ Unknown

Weight
39.9
19.8
32.4
21
27.5
20.8
21.3
40
10.7
22.6
27
10.8
20.9
14.7
31.4
17.2
11.4
19.1
31.3
14.8

m =	22.73
s =	8.84

s(m) =	1.98
t =	2.09
m.e. =	4.14

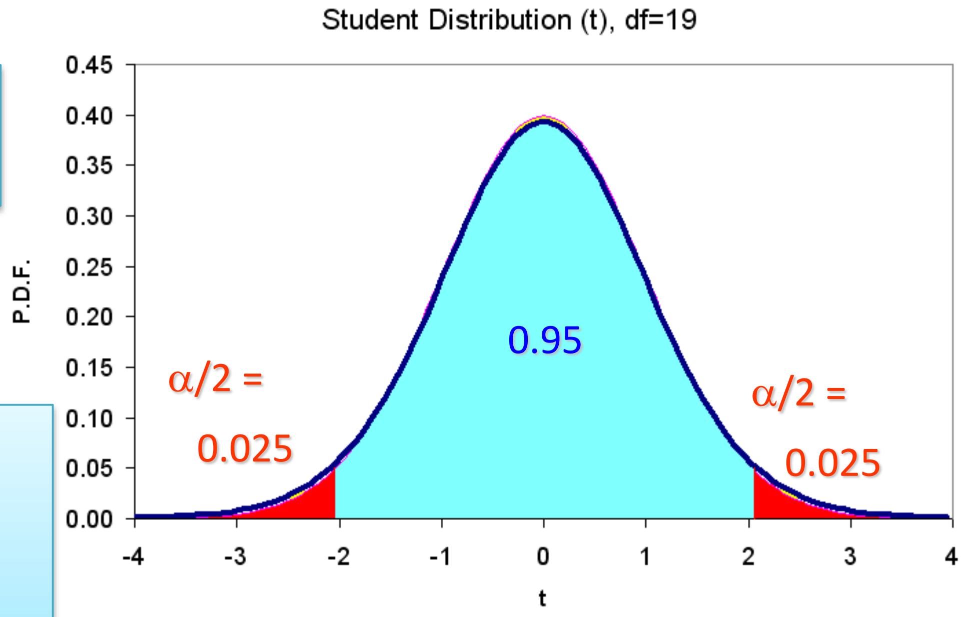
$$\mu = m \pm t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}$$

In R use (for $t_{\alpha/2}$):

- `-qt(α/2, n-1)`

In R use :

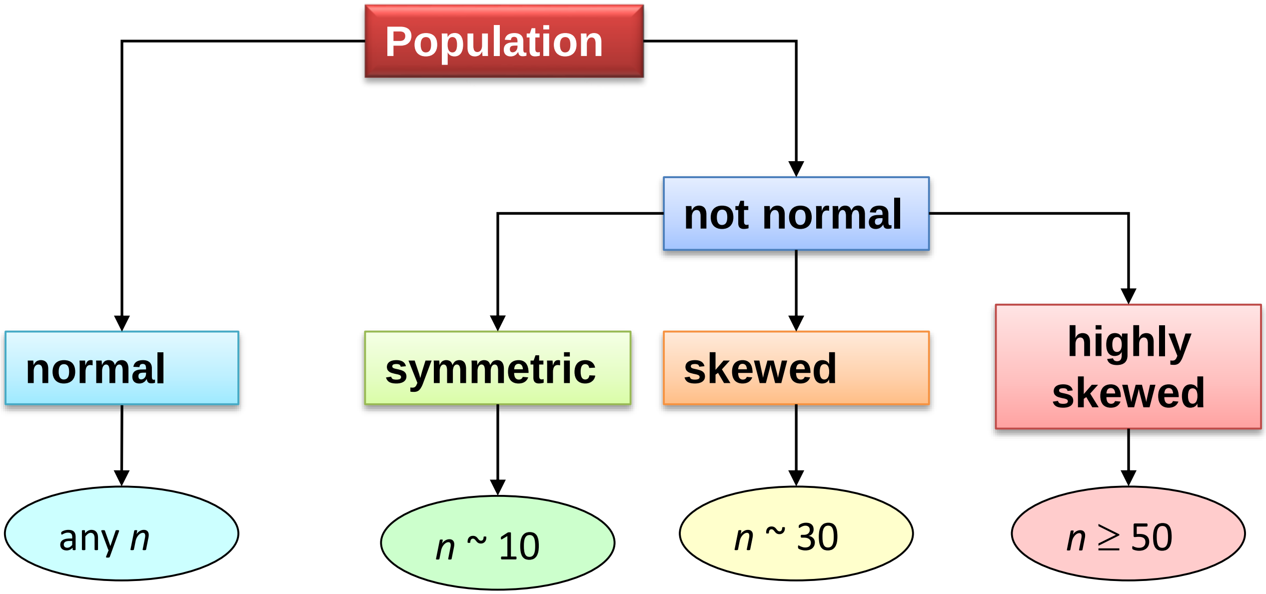
- `dt(…)` – probability density
- `pt(…)` – cumul. probability
- `qt(…)` – quantiles
- `rt(…)` – simulate random var.



Practical Advices

$$\mu = m \pm t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}$$

Advice 1



Advice 2

if n >100 you can use z-statistics instead of t-statistics (error will be <1.5%)

Determining Sample Size

Let's focus on another aspect: how to select a proper number of experiments.

$$\mu = m \pm E(n,\sigma)$$
$$E(n,\sigma) = E$$
$$n - ?$$

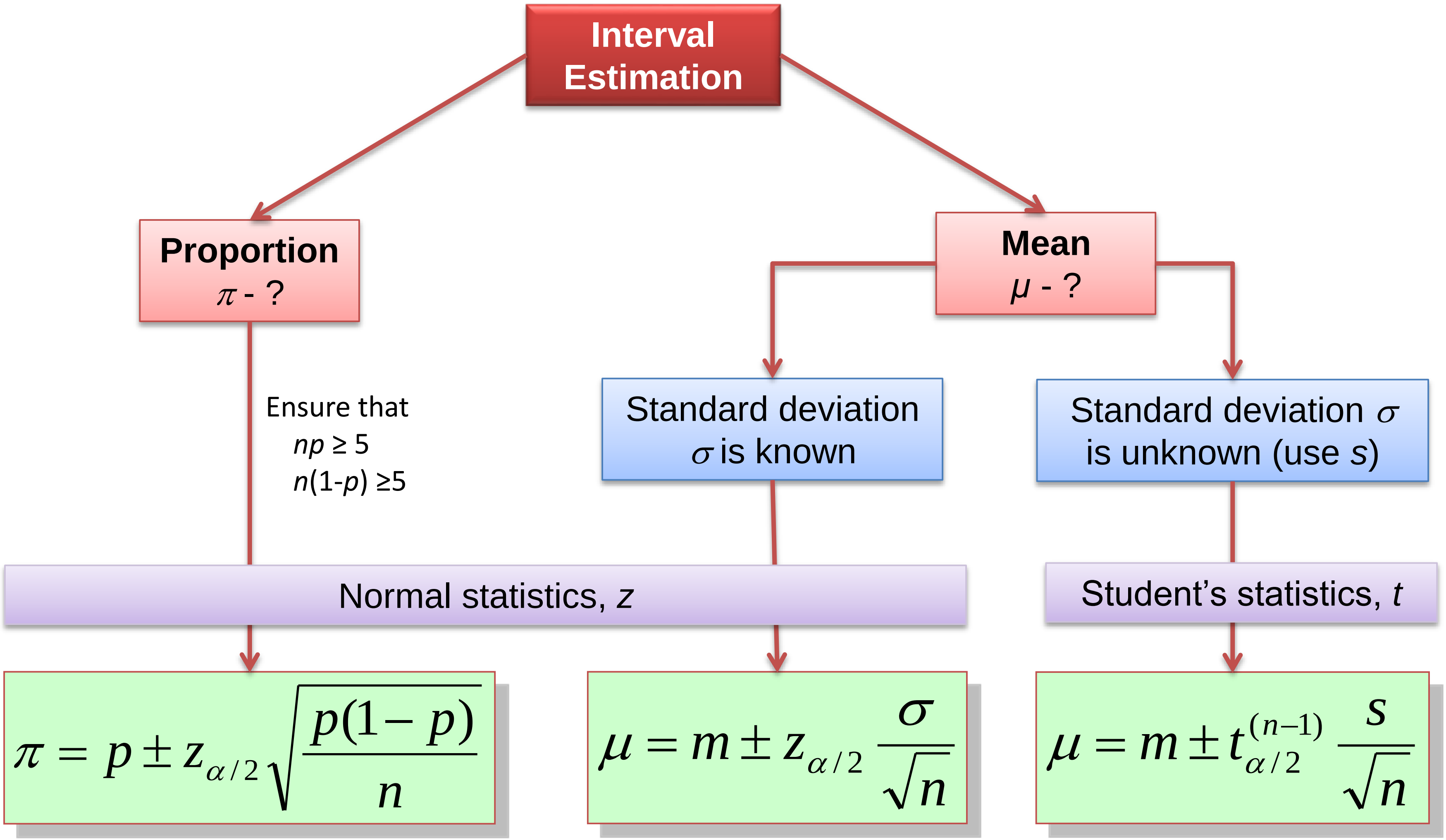
$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

$$\boxed{n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}}$$

$$\boxed{n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2}}$$

Pipeline: Interval Estimations for Mean and Proportions



$z_{\alpha/2}$ = -qnorm(a/2)

$t_{\alpha/2}^{(n-1)}$=-qt(a/2,n-1)

Interval Estimation for Variance

Variance
A measure of variability based on the squared deviations of the data values about the mean.

population

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

sample

$$s^2 = \frac{\sum (x_i - m)^2}{n-1}$$

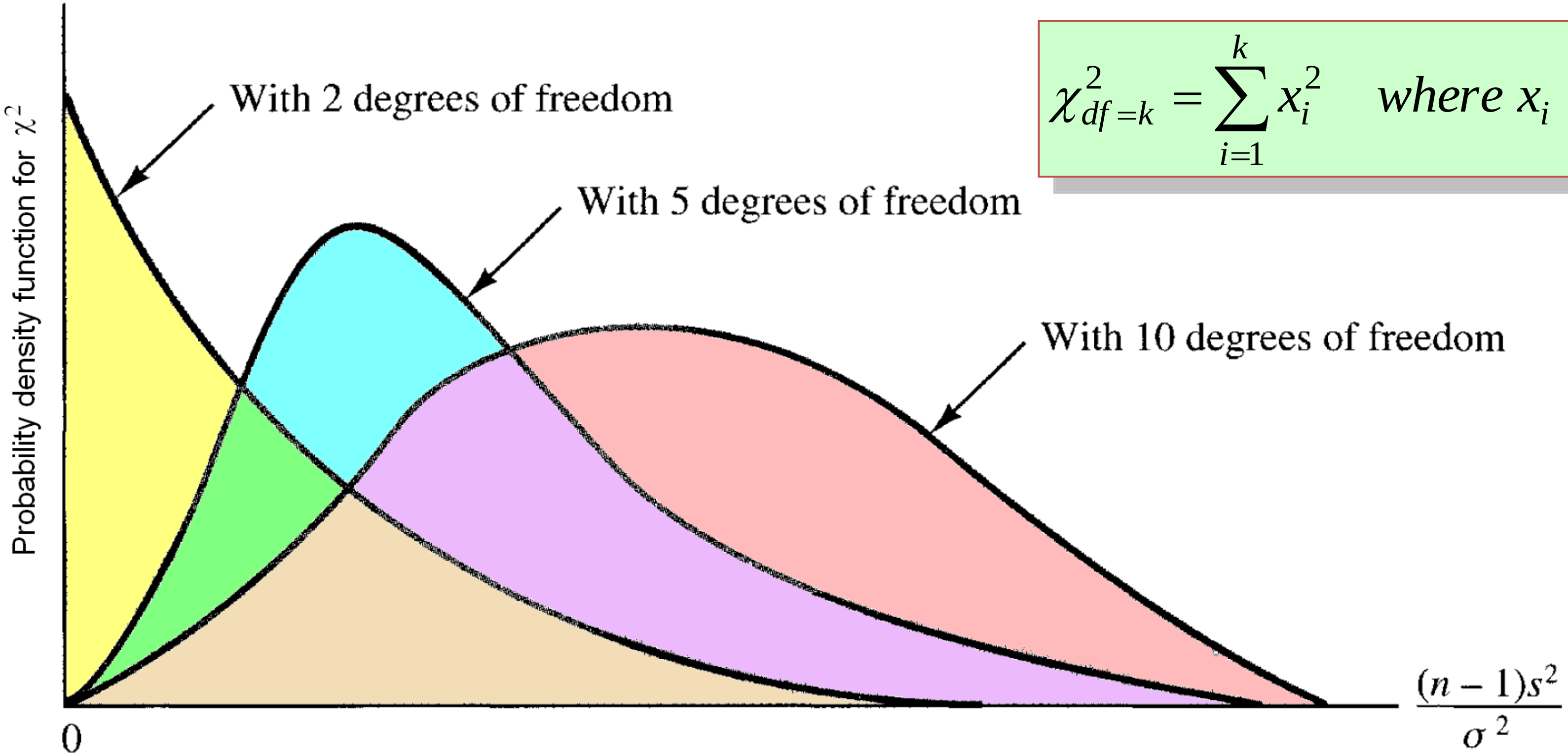
The interval estimation for variance is build using the following measure:

$$(n-1)\frac{s^2}{\sigma^2}$$

Sampling distribution of $(n-1)s^2/\sigma^2$
Whenever a simple random sample of size n is selected from a normal population, the sampling distribution of $(n-1)s^2/\sigma^2$ has a chi-square distribution (χ^2) with n-1 degrees of freedom.

$$(n-1)\frac{s^2}{\sigma^2} = \chi^2_{df=n-1}$$

Interval Estimation for Variance



$$\chi^2_{df=k} = \sum_{i=1}^{k} x_i^2 \quad where\ x_i - normal$$

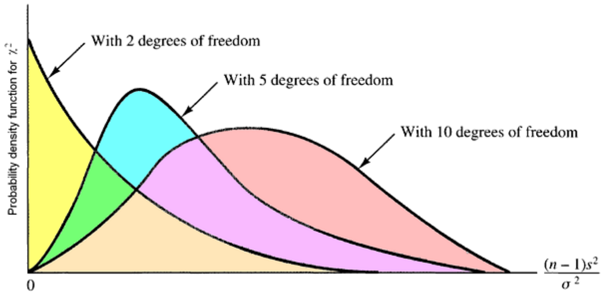
χ^2 distribution works only for sampling from normal population

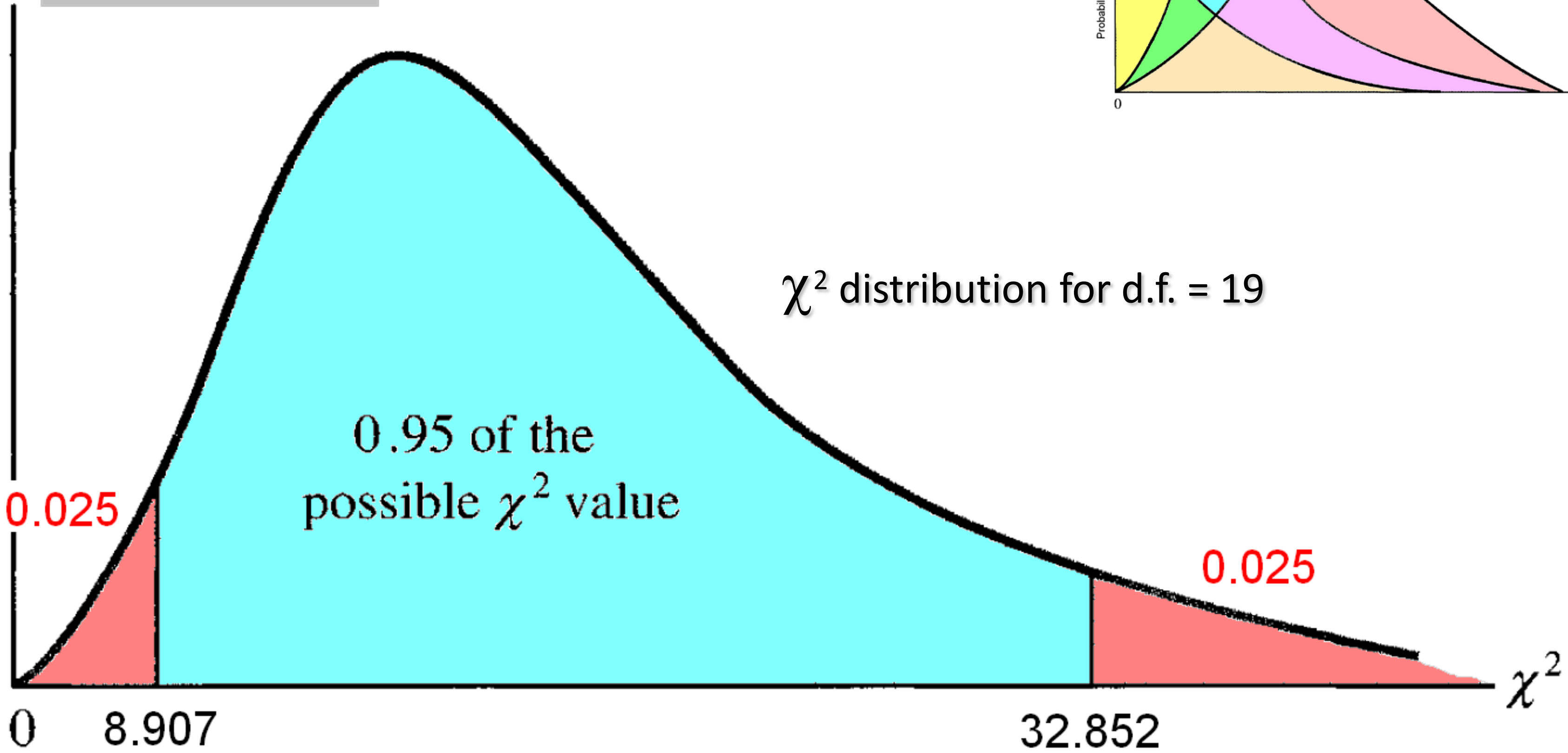
In R use :

- `dchisq(…)` – probability density
- `pchisq(…)` – cumul. probability
- `qchisq(…)` – quantiles
- `rchisq(…)` – simulate random var.

Interval Estimation for Variance

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2}$$

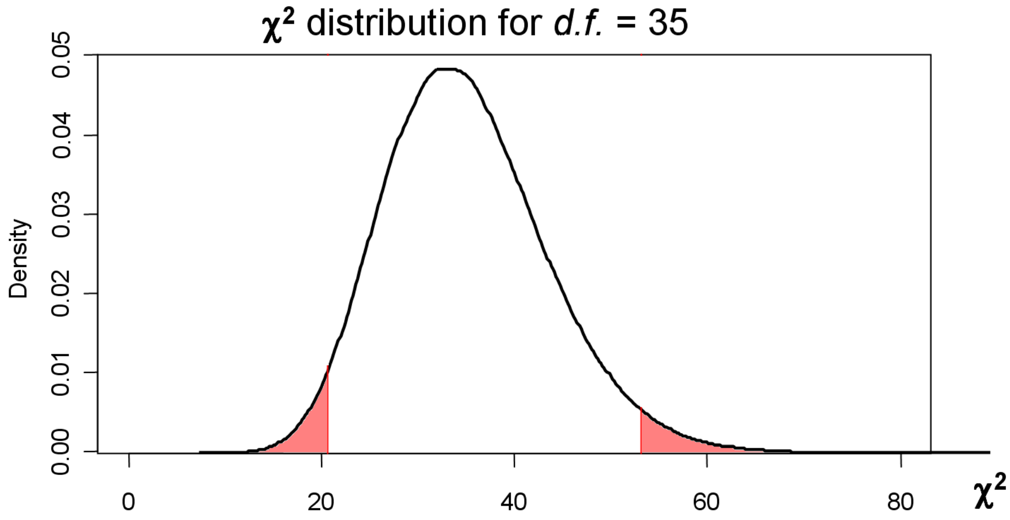




Interval Estimation for Variance

$$\chi^2_{1-\alpha/2} \le (n-1)\frac{s^2}{\sigma^2} \le \chi^2_{\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$



Suppose sample of $n = 36$ coffee cans is selected and $m = 2.92$ and $s = 0.18$ lbm is observed. Provide 95% confidence interval for the standard deviation

$$\frac{(36-1)0.18^2}{53.203} \le \sigma^2 \le \frac{(36-1)0.18^2}{20.569}$$

$$0.0213 \le \sigma^2 \le 0.0551$$

$$\boxed{0.146 \le \sigma \le 0.235}$$

In R use (for $\alpha/2$):

- `qchisq(α/2, n-1)`
- `qchisq(1-α/2, n-1)`

There can be an inversion of quantiles b/w R and Excel! Check!

L2.4. Interval Estimation

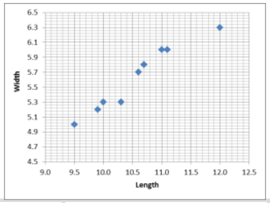
Interval Estimation for Correlation

Fisher's transformation connects normal z-values and correlation coefficients. We assume Z is normally distributed and its standard error can be calculated as shown below:

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \quad \Leftrightarrow \quad r = \frac{e^{2Z}-1}{e^{2Z}+1}$$

$$\sigma_Z = \sqrt{\frac{1}{n-3}}$$

Then confidence intervals with Fisher's transformation can be obtained:



1. Transform correlation r → Z
2. Calculate standard deviation for Z using equation σ_Z
3. Calculate upper and lower limits of Z:

   $Z_{min\,/\,max} = Z \pm z_{\alpha/2}\,\sigma_Z = Z \pm 1.96\,\sigma_Z$
4. Transform $Z_{min/max}$ back into $r_{min/max}$

n = 10

r=	0.969226	
Fisher's Z=	2.079362	
sZ=	0.377964	
	Lower	Upper
Limits Z	1.338552	2.820172
Limits r	0.871324	0.992922

Interval Estimation for Random Function

Distribution of sum or difference of 2 normal random variables

The sum/difference of 2 (or more) normal random variables is a normal random variable with **mean equal to sum/difference** of the means and **variance equal to SUM** of the variances of the compounds.

$$x \pm y \rightarrow Normal\ distribution$$

$$E[x \pm y] = E[x] \pm E[y]$$

$$\sigma^2_{x \pm y} = \sigma^2_x + \sigma^2_y$$

Distribution of sum of squares on *k* standard normal random variables

The sum of squares of *k* standard normal random variables is a χ^2 with *k* degree of freedom.

$$if \quad x_1, \ldots, x_k \rightarrow Normal\ distribution$$

$$\sum_{i=1}^{k} x_i^2 \rightarrow \chi^2 \quad with\ d.f. = k$$

What to do in more complex situations?

$$\frac{x}{y} \rightarrow ?$$

$$\sqrt{x} \rightarrow ?$$

$$\log(|x|) \rightarrow ?$$

Interval Estimation for Random Function

Try to solve analytically?

Simplest case. E[x] = E[y] = 0

Ratio distribution

From Wikipedia, the free encyclopedia

A **ratio distribution** (or *quotient distribution*) is a probability distribution constructed as the distribution of the ratio of random variables having two other known distributions. Given two random variables X and Y, the distribution of the random variable Z that is formed as the ratio

$$Z = X/Y$$

is a *ratio distribution*.

$$p_Z(z) = \frac{b(z) \cdot c(z)}{a^3(z)} \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y} \left[2\Phi\left(\frac{b(z)}{a(z)}\right) - 1\right] + \frac{1}{a^2(z) \cdot \pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2}+\frac{\mu_y^2}{\sigma_y^2}\right)}$$

where

$$a(z) = \sqrt{\frac{1}{\sigma_x^2}z^2 + \frac{1}{\sigma_y^2}}$$

$$b(z) = \frac{\mu_x}{\sigma_x^2}z + \frac{\mu_y}{\sigma_y^2}$$

$$c(z) = e^{\frac{1}{2}\frac{b^2(z)}{a^2(z)} - \frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2}+\frac{\mu_y^2}{\sigma_y^2}\right)}$$

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du$$

L2.4. Interval Estimation

Interval Estimation for Random Function

Two rates where measured for a PCR experiment: experimental value (X) and control (Y). 5 replicates where performed for each.

From previous experience we know that the error between replicates is normally distributed.

Q: provide an interval estimation for the fold change X/Y (α=0.05)

#	Experiment	Control
1	215	83
2	253	75
3	198	62
4	225	91
5	240	70

Mean	226.2	76.2
StDev	21.39	11.26

Mean	226.2	76.2
StDev	**9.57**	**5.03**

Standard error (st.dev. of mean)

1. Calculate ***m***, ***s*** and ***sm*** (standard error)
2. Simulate outcomes of large number of 5-rep. experimets

```
X = rnorm(n, mean=226.2, sd=9.57)
Y = rnorm(n, mean=76.2, sd=5.03)
```

3. Calculate the function X/Y and estimate quantiles, which cut 95% of the results

```
quantile(X/Y,prob=c(0.025,0.975))
```

Alternative: in case of a ratio – log your data…

UNIVERSITÉ DU LUXEMBOURG

L2.4. Interval Estimation

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

```
################################################################################
# L2.4. INTERVAL ESTIMATION
################################################################################

##------------------------------------------------------------------------------
## L2.4.2. Interval estimation for proportion
##------------------------------------------------------------------------------

Pan=read.table("http://edu.sablab.net/data/txt/pancreatitis.txt",header=T,
               sep="\t")
## this is not completely correct as we pool control and experiomental group.
## try to avoid on practice
x=Pan$Smoking == "Never"
n=length(x)
p=sum(x)/n
sp = sqrt(p*(1-p)/n)
E=-qnorm(0.025)*sp

##------------------------------------------------------------------------------
## L2.4.3. Interval estimation for mean
##------------------------------------------------------------------------------
m=22.73
s=8.84
n=20
sm=s/sqrt(20)
E=-qt(0.025,n-1)*sm

##------------------------------------------------------------------------------
## L2.4.4. Interval estimation for variance
##------------------------------------------------------------------------------
n=36
s=0.18
a=0.05
## limits (in Excel values are inverted)
sqrt((n-1)*s^2 / qchisq(1-a/2, n-1))
sqrt((n-1)*s^2 / qchisq(a/2, n-1))

#>>>>>>>>>>>>>>>>>>>>>>>>>>
#> please, do Task L2.4
#>>>>>>>>>>>>>>>>>>>>>>>>>>
```

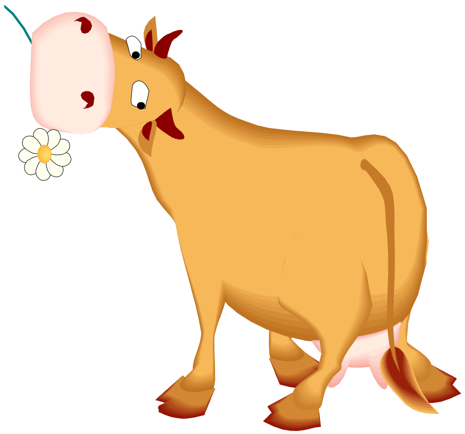
```
##################################################################
# L2.6. INTERVAL ESTIMATION FOR A RANDOM FUNCTION
##################################################################
n=10^5
## simultate mean X for n times
X = rnorm(n,mean=226.2, sd=9.57)
## simultate mean Y for n times
Y = rnorm(n,mean=76.2, sd=5.03)
## estimte confidence intervals by quantiles
plot(density(X/Y))
q95 = quantile(X/Y,prob=c(0.025,0.975))
abline(v=q95,lty=2)
lq95 = exp(quantile(log(X/Y),prob=c(0.025,0.975)))
print(q95)

## can we use analytical solution with log?
m = mean(log(X)) - mean(log(Y))
s = sqrt(var(log(X)) + var(log(Y)))
lim = c(m - s*1.96, m + s*1.96)
print(exp(lim))
```

Task L2.4.

Questions ?

Thank you for your attention



to be continued…