Statistical Learning

- Alberto Roverato -

A.A. 2024 – 2025

Overview – First semester (Mod. A)

1. Basic elements

2. Exploring data with graphs and numerical summaries

3. R software environment for statistical computing and graphics

4. Quick probability review
   - the Bayes theorem
   - random variables

5. Statistical inference
   - point estimation and sampling distributions
   - interval estimation
   - hypothesis testing
   - inferential tools for some specific problems
   - some general ways to build inferential tools

6. Exam: written test

Teaching materials – First semester (Mod. A)

- Course slides & handouts
  (open-book exam: a printed version of these slides can be used during the first partial, i.e. MOD-A, exam).

- Navidi, W.
  Statistics for Engineers and Scientists
  McGraw-Hill

- Software:
  - R: `https://www.r-project.org/`
  - Rstudio: `https://www.rstudio.com/`

Overview – Second semester (Mod. B)

1. Linear models and generalized linear models
2. Discriminant analysis and Classification
3. Prediction
4. Cross validation and Variable selection
5. $\vdots$
6. Exam: tba

Teaching materials – Second semester (Mod. B)

- Course slides & handouts.
- James, G., Witten, D., Hastie, T. and Tibshirani, R.
  An Introduction to Statistical Learning
  with applications in R (2nd ed.) Springer 2021.
  `https://www.statlearning.com/`
- Hastie, T., Tibshirani, R. and Friedman, J.
  The Elements of Statictical Learning (2nd ed.). 2009.
  `https://web.stanford.edu/~hastie/ElemStatLearn/`

Basic Elements

What is statistics

Statistics is the art and science of designing studies and analyzing the data that those studies produce. Its ultimate goal is translating data into knowledge and understanding of the world around us.

In short, statistics is the art and science of learning from data.

Example of data matrix: Cars93

Data from 93 cars on sale in the USA in 1993 (93 rows and 27 columns)

```
  Manufacturer        Model      Type  Passengers  Length  Weight    Origin

1        Acura      Integra     Small           5     177    2705   non-USA
2        Acura       Legend   Midsize           5     195    3560   non-USA
3         Audi           90   Compact           5     180    3375   non-USA
.......

50       Lexus        SC300   Midsize           4     191    3515   non-USA
51     Lincoln  Continental   Midsize           6     205    3695       USA
52     Lincoln     Town_Car     Large           6     219    4055       USA
53       Mazda          323     Small           4     164    2325   non-USA
.......

90  Volkswagen       Passat   Compact           5     180    2985   non-USA
91  Volkswagen      Corrado    Sporty           4     159    2810   non-USA
92       Volvo          240   Compact           5     190    2985   non-USA
93       Volvo          850   Midsize           5     184    3245   non-USA
```

Basic definitions

- Statistical unit: unit in a statistical analysis refers to one member of a set of entities being studied.

- Population: collection of all items of interest or under investigation. (May be very large or even virtually infinite.)

- Parameter: specific characteristic of a population.

- Sample: observed subset of the population.
  n represents the sample size.

- Statistic: specific characteristic of a sample.

Variables

- A variable is any entity, measured on statistical units, that can take on different values ("different" as opposed to constant).

- Values are not always numerical. For instance, the variable *origin* consists of two text values: '*USA*' and '*non-USA*'.

Types of variables

- Categorical variables are not measured on a numerical scale. Take values in a set of categories. A categorical variable is called

  (i) ordinal when there exists a natural ordering of categories. For ex:
  Patient condition: excellent, good, fair, poor.
  Government spending: too high, about right, too low.

  (ii) nominal when categories do not have any natural orderings. For ex:
  Transport to work: car, bus, bicycle, walk.
  Favorite music: rock, classical, jazz, country, folk, pop.

- Quantitative variables are measured on a numerical scale. A quantitative variable is called

  (i) discrete when it can take on only a finite or countably infinite number of values;

  (ii) continuous when it can take on any value in a certain range.

Descriptive vs inferential statistics

- Descriptive statistics and EDA=Exploratory Data Analysis
  - Graphical and numerical procedures to summarize data.
- Inferential Statistics
  - Estimation.
  - Decision.
  - Prediction.
  - ⋮

Exploring data with graphs and numerical summaries

Exploratory Data Analysis (EDA)

EDA refers to the critical process of performing initial investigations on data, with the help of summary statistics and graphical representations. EDA is a way of getting an overview of the quality and nature of the data, it is used for seeing what the data can tell us *before* we begin studying it with formal statistical models.

EDA may:

- detect obvious errors in the data;
- check that assumptions underlying formal analyses are plausible;
- uncover patterns in the data and suggest how data should be modelled;
- provide new directions for scientific inquiries.

Modern statistical software

- Provide interactive environments for exploration of data and models
- create static and dynamic graphics
- fit and critique complex models
- are extensible high level programming languages
- object oriented code

R and RStudio

- R is a free software environment for statistical computing and graphics
  https://www.r-project.org/

- RStudio is an integrated development environment (IDE) for R
  https://posit.co/downloads/

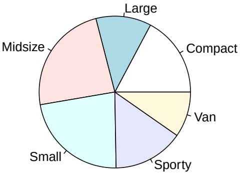
Frequency table for a categorical variable

`Type`

category	freq.	rel.freq
Compact	16	0.17
Large	11	0.12
Midsize	22	0.24
Small	21	0.22
Sporty	14	0.15
Van	9	0.10
	93	1.00

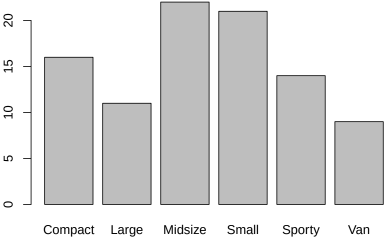
Graphs for categorical variables

Pie chart



`pie(tb.Type)`

Bar chart



`barplot(tb.Type)`

More than one variable

Tables for more two or more variables are called cross-classified tables.

	Compact	Large	Midsize	Small	Sporty	Van
USA	7	11	10	7	8	5
non-USA	9	0	12	14	6	4

Graphs for continuous variables

1. Univariate: histograms, density curves, boxplots, quantile-quantile plots.
2. Bivariate: scatter plots (with smooth lines), side-by-side boxplots.
3. Several variables: scatter plot matrices, lattice or trellis plots,...

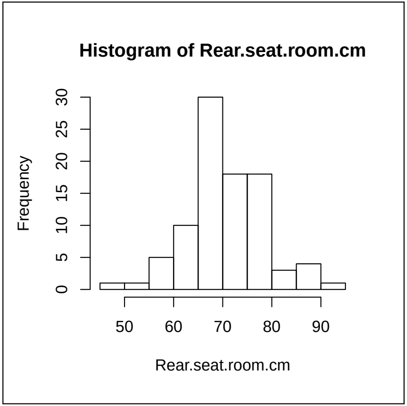
Histograms

Given a set of breakpoints covering the data, we can count how many data points fall into each interval (or class interval),

- for a frequency/relative-frequency histogram: draw a rectangle for each class with the class interval as the base and the height equal to the count/proportion of data points in the class;
- for a density histogram: draw a rectangle for each class with the class interval as the base and the area equal to the proportion of data points in the class.

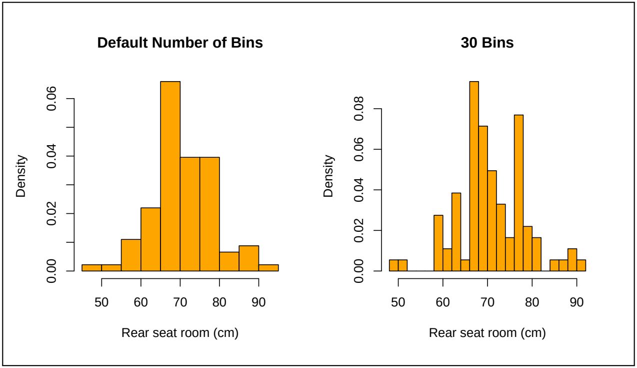
`R` default is with equi-spaced breaks, and plots the counts in the cells defined by breaks. The density histogram is default with non-equi-spaced breaks.

Histograms (cntd)



```
> Rear.seat.room.cm <- Rear.seat.room * 2.54
> hist(Rear.seat.room.cm)
```

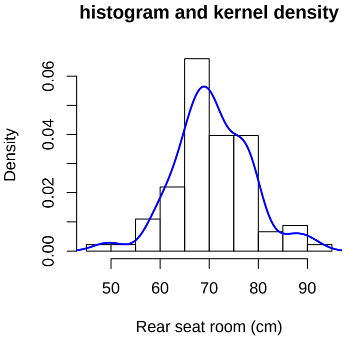
Histograms (cntd)



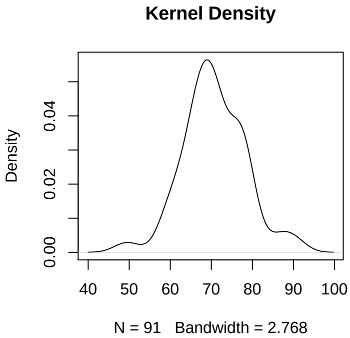
```
> par(mfrow=c(1,2))
> hist(Rear.seat.room.cm, prob=TRUE, xlab="Rear seat room (cm)",
  main="Default Number of Bins", col="orange")
> hist(Rear.seat.room.cm, prob=TRUE, xlab="Rear seat room (cm)",
  main="30 Bins", breaks=30,col="orange")
> par(mfrow=c(1,1))
```

Estimating the density curve

It is like smoothing an histogram



Estimating the density curve (cntd)



```
> dens <- density(Rear.seat.room.cm, na.rm=TRUE)
> plot(dens,main="Kernel Density")
```

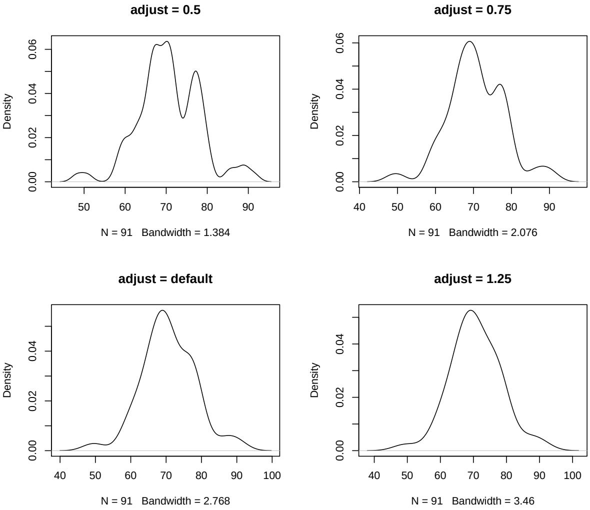
Kernel density estimation

The bandwidth of the kernel has a strong influence on the resulting estimate. Real features vs artifacts of the estimation.

```
density(y, bw, adjust, kernel, window)
```

- `bw`: bandwidth controls amount of smoothing
- `adjust`: actual bandwidth is `adjust*bw`
- `kernel, window`: smoothing kernel
- see `help(density)`: for other options

Adjusting bandwidth



```
> dens <-  density(Rear.seat.room.cm, adjust=0.5)
> plot(dens,main="adjust=0.5")
```

Numerical summaries

Numerical summaries can be used to get a feel for the location and spread of data points, and are often used when there are several data sets which we wish to compare.

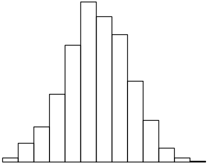
Measures of centrality

Suppose we have a set of measurements $x_1, x_2, \ldots, x_n$.

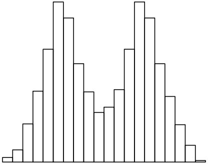
Measures of centrality include:

- the (arithmetic) mean: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$;
- the median: equal to $x_{(n+1)/2}$ if n is odd (the "middle" observation), and the average of $x_{(n/2)}$ and $x_{(n/2+1)}$ if n is even;
- the mode: the category/value that occurs most often, that is with highest frequency - not necessarily unique. For continuous variables the "modal class" is typically computed.

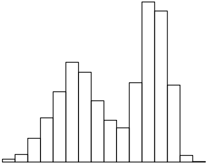
Multimodal distributions



Unimodal distribution



Bimodal distribution

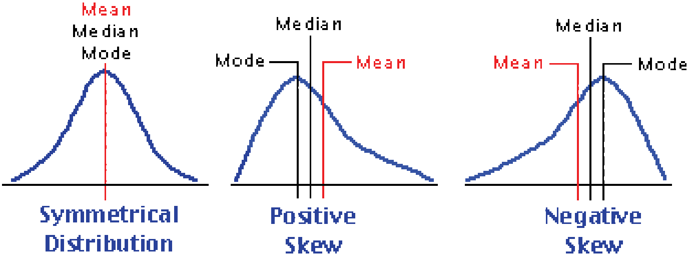


Bimodal distribution?

Strictly speaking, the distribution on the right is unimodal, but it is common practice to refer to this kind of behaviour as bimodal to emphasize that there are two peaks.

Symmetry

- Symmetric distribution: mean $\approx$ median;
- Right/positive-skewed distribution: mean $>$ median;
- Left/negative-skewed distribution: mean $<$ median;



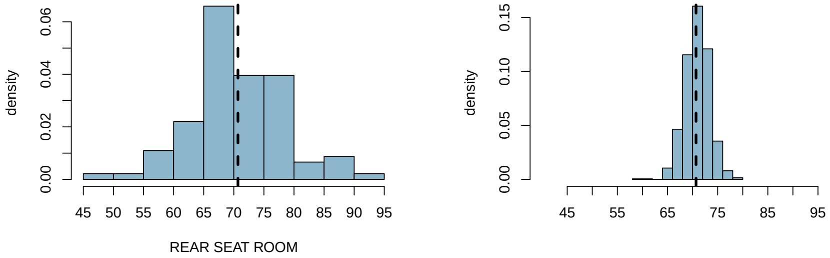
Notes

- Mean and standard deviation are sensitive to outliers.
- Symmetric data: mean and median should be approximately equal.
- Skewed data: median is more appropriate as a summary.

Variability

Variability refers to the extent to which values differ from one another. The terms variability, spread, and dispersion are synonyms.

Variability can also be thought of as how much values differ from their mean, that is how spread out a distribution is. Example: same mean different dispersion,



Measures of variability: variance and standard deviation

- Variance:

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

- the standard deviation s is the square root of the variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Measures of variability: interquartile range

- the interquartile range (IQR):

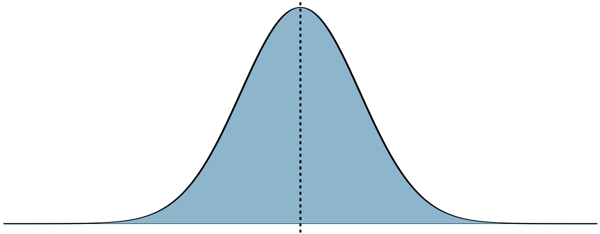
$$IQR = q_{3/4} - q_{1/4}$$

.

- Note that $q_{1/4}$ and $q_{3/4}$ are sometimes referred to as the lower and upper "hinges" respectively. In this context, the IQR is called the H–distance.
- Recall that the α quantile, q_α, is defined for $0 \leq \alpha \leq 1$ such that a proportion α of the data is less than q_α and proportion $1 - \alpha$ is greater than q_α.
- The quartiles are defined as the quantiles $q_{1/4}$ and $q_{3/4}$.

The normal (Gaussian) distribution

In many practical applications data sets exhibit a (approximately) symmetric bell-shaped distribution.



- unimodal;
- symmetric about the mean;
- bell-shaped.

Interpretation of standard deviation

As an "empirical rule", if the shape of the histogram is approximately bell-shaped, we would expect

- 68% of the data to be within 1 standard deviation of the mean;
- 95% of the data to be within 2 standard deviations of the mean;
- 99.7% of the data to be within 3 standard deviations of the mean.

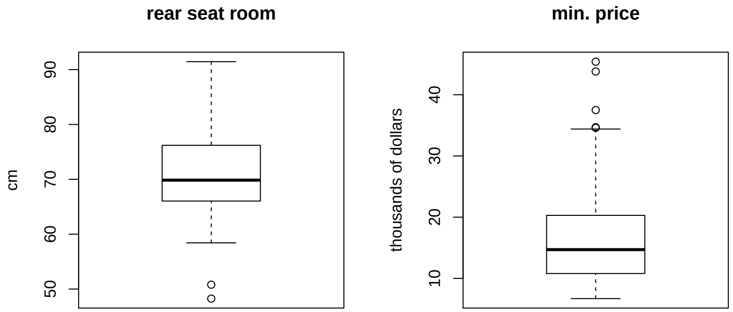
Boxplots

Simple graphical device to display the overall shape of a distribution, including the outliers.

- Box is based on upper and lower quartiles.
- Line in box indicates median.
- Whiskers based on smallest/largest observation within 1.5 × IQR of the box.
- Points for all cases beyond whiskers "outliers".

Good for showing skewness/symmetry.

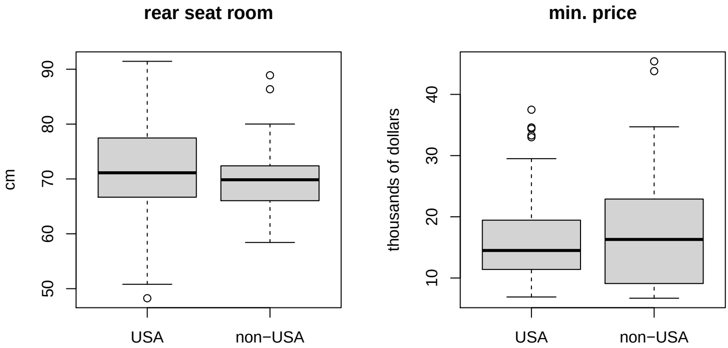
Examples of boxplots



```
> par(mfrow=c(1,2))
> boxplot(Rear.seat.room.cm, ylab="cm", main="rear seat room")
> boxplot(Min.Price, ylab="thousands of dollars", main="min. price")
> par(mfrow=c(1,1))
```

Side-by-Side boxplots

Side-by-side boxplots are useful for showing the distribution of a quantitative variable for each level of a qualitative variable.



R command

```
> par(mfrow=c(1,2))
> boxplot(Rear.seat.room.cm~Origin, col="lightgray",
  ylab="cm", main="rear seat room")
> boxplot(Min.Price~Origin, col="lightgray",
  ylab="thousands of dollars", main="min. price")
> par(mfrow=c(1,1))
```

Formula using the tilde operator "$\sim$"

$$\texttt{Rear.seat.room.cm} \sim \texttt{Origin}$$

creates a boxplot of `Rear.seat.room.cm` for each level of `Origin`

Outliers

An outlier is a piece of data that lies outside the other values in the set. A popular method used to detect outliers is:

(a) Calculate first quartile ($q_{1/4}$), third quartile ($q_{3/4}$) and the interquartile range.

(b) Compute $q_{1/4} - 1.5 \times IQR$. Compute $q_{3/4} + 1.5 \times IQR$. Anything outside this range is an outlier.

Beware: this method shouldn't be applied "blindly": If it is obvious that the outlier is due to incorrectly entered or measured data, you should drop the outlier. Otherwise, it is not legitimate to simply drop the outlier.

Possible approach: you may run the analysis both with and without the outliers so as to understand the impact of outliers on the result.

Shapes of distributions

- Overall pattern: single mode or multiple modes?
- Skewed? Left or right?
- Are there outliers?
- Changes across space/time or levels of a categorical variable.

Scatter plots

Good to explore pairwise relationships between variables

- `plot(x,y)`
- `plot(y ~ x)` Note the use of the tilde operator

To better grasp behaviour of the data, we can:

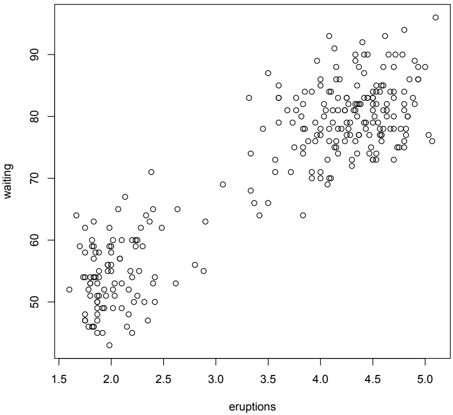
- add lines/smoothed lines;
- add legends;
- transform data.

Old Faithful Geyser data

- Old Faithful erupts every 35-120 minutes for 1.5-5 minutes to a height of 90-184 feet. The time to the next eruption is typically predicted using the duration of the current eruption. The longer the eruption lasts, the longer the interval until the next eruption.
- Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

```
> faithful
    eruptions waiting
1       3.600      79
2       1.800      54
3       3.333      74
4       2.283      62
.....
```

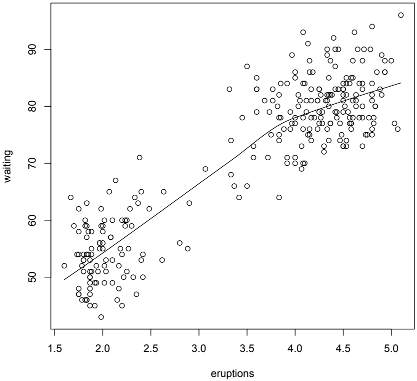
Scatter plot



```
> attach(faithful)
> plot(waiting~eruptions)
```

Adding a smooth curve

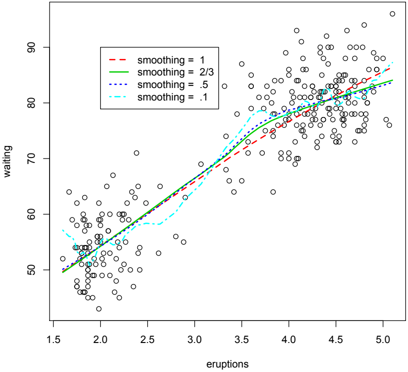
The function `scatter.smooth()` fits a smooth line to (x,y) pairs



```
> scatter.smooth(waiting~eruptions)
```

Choice of smoothing

The function has a parameter (span), which controls the degree of smoothing. The default is 2/3.



```
> lines(loess.smooth(eruptions, waiting), lty=1,col=2)
> lines(loess.smooth(eruptions, waiting, span=1), lty=2,col=3)
```

Legends

Can be added to a plot

```
legend(x,y,legend,col,lty,lwd)
```

- `x,y` location for legend
- `legend` text for legend
- `col` colors used
- `lty` line types
- `lwd` line widths

```
> legend(x=2,y=90,
    legend=c("smoothing=1","smoothing=2/3","smoothing=0.5",
  "smoothing=0.1"), col=2:5, lty=c(2,1,3,4), lwd=rep(2,4))
```

Transformations

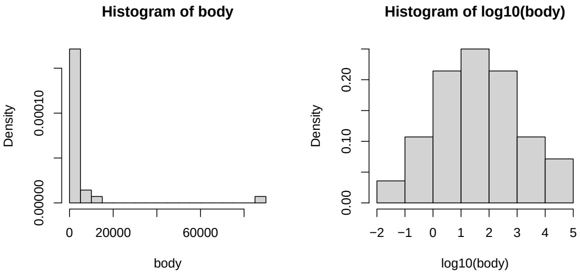
For plots to be more effective, we may need to transform variables. (Also may make normality assumptions more plausible - more on this mistery in following lectures)

Common transformations

- `log(·)`: on positive continuous measures
- `sqrt(·)`: on counts
- inverse: `(1/·)`

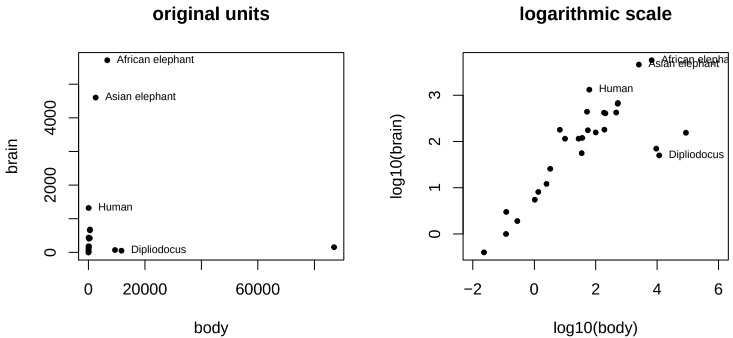
Typically, need the largest observation to be at least 10 times larger than the smallest case for transformations to be effective.

Log transformations



```
> library(MASS)
> data("Animals")
> attach(Animals)
> summary(body)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    0.02     3.10    53.83  4278.44   479.00 87000.00
> summary(log10(body))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.6383  0.4884  1.7308  1.6379  2.6798  4.9395
> 10^mean(log10(body))
[1] 43.43674
> n <- length(body)
> prod(body)^(1/n) # geometric mean
[1] 43.43674
```

Log transformations (cntd)



original units

logarithmic scale

```
> plot(body, brain, pch=20,
       main="original units")

> plot(log10(body), log10(brain), pch=20,
       main="logarithmic scale")
```

Quick Probability Review

Statistical Inference

The goal of statistical inference is to draw conclusions about a population from “representative information” about it.

A powerful way to obtain representative information about a population is through the planned introduction of chance.

Thus, probability is the foundation of statistical inference – to study the latter, we should first know the former.

Quick Probability Review

– The Bayes theorem –

The Bayes theorem

- Conditional probability:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

- The Bayes theorem:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

- The low of total probability:

$$\Pr(B) = \Pr(B \mid A)\Pr(A) + \Pr(B \mid \bar{A})\Pr(\bar{A}).$$

Assessing accuracy of a diagnostic test

A test is designed to detect the presence of a disease (e.g. breast cancer). This test also has two possible outcomes: positive $(+)$, negative $(-)$.

Because diagnostic procedures undergo extensive evaluation before they are approved for general use, the medical establishment is likely to have a fairly precise notion of the probabilities of a false positive, i.e., the probability of obtaining a positive test result given that the patient has not the diseases, and a false negative, i.e., the probability of obtaining a negative test results given that the patient has the disease.

These probabilities are conditional probabilities: the probability of a false positive is $\text{Pr}(+|\bar{D})$, and the probability of a false negative is $\text{Pr}(-|D)$.

Predictive Value of the Test

- Assume
  - $\Pr(D) = 0.001$
  - $\Pr(+|\bar{D}) = 0.015$
  - $\Pr(-|D) = 0.003$

  This implies $\Pr(\bar{D}) = .999$, $\Pr(-|\bar{D}) = .985$ and $\Pr(+|D) = .997$.

- Which is the predictive value of the test, i.e., the probability that a diagnosis of breast cancer is correct being the test positive?

The question asks to compute the quantity $\Pr(D|+)$.

Computations

By definition, we have

$$\Pr(D|+) = \frac{\Pr(D \cap +)}{\Pr(+)}$$

that can be computed by applying Bayes theorem

$$\Pr(D|+) = \frac{\Pr(+|D)\ \times\ \Pr(D)}{\Pr(+)}$$

Quick Probability Review

– Random variables –

Random variables

Informally, random variables are rules to assign numbers to experimental outcomes. In what follows, we will briefly review the most important random variables, along with `R` commands to use them.

Quick Probability Review

– Some important discrete random variables –

Bernoulli random variable

A Bernoulli trial is a random experiment with exactly two possible outcomes, “success” and “failure”. A random variable X is has a Bernoulli distribution, i.e. $X \sim \textit{Bernoulli}(\pi)$, if $X(S) = \{0, 1\}$ and

$$p_x(x;\ \pi) = \pi^x (1-\pi)^{1-x}.$$

Traditionally, we associate $X = 1$ with “success” and $X = 0$ with “failure”.

The family of probability distributions of Bernoulli trials is parameterized (indexed) by a real number $\pi \in [0, 1]$, by setting $\pi = \Pr(X = 1)$.

1. $E(X) = \pi$,
2. $Var(X) = \pi(1-\pi)$.

Expected value

The expected value of a discrete random variable X, which we will denote $E(X)$, is the probability-weighted average of the possible values of X, i.e.,

$$E(X) = \sum_{x \in X(S)} x \, p_x(x).$$

Example:.

If $X \sim \mathit{Bernoulli}(\pi)$, then $E(X) = 0 \times (1 - \pi) + 1 \times \pi = \pi$.

Variance and Standard Deviation

The variance of a discrete random variable X, which we will denote $Var(X)$, is the probability-weighted average of the squared deviations of X from $E(X)$, i.e.,

$$Var(X) = E[(X - E(X))^2] = \sum_{x \in X(S)} (x - \mu)^2 p_x(x),$$

where $\mu = E(X)$.

- The standard deviation of a random variable is the square root of its variance
- $Var(X) = E(X^2) - [E(X)]^2$

Example

If $X \sim \mathit{Bernoulli}(\pi)$, then

$$\begin{aligned} Var(X) &= E\left[(X - E(X))^2\right] \\ &= (0-\pi)^2 \times \Pr(X=0) + (1-\pi)^2 \times \Pr(X=1) \\ &= (0-\pi)^2(1-\pi) + (1-\pi)^2\pi \\ &= \pi(1-\pi)(\pi + 1 - \pi) \\ &= \pi(1-\pi). \end{aligned}$$

Expected value and variance: some properties

If a and b are two constants, then

1. $E(a\,X + b) = a\,E(X) + b$
2. $Var(a\,X + b) = a^2\,Var(X)$
3. The expected value of a sum of random variables, X_1 and X_2, equals the sum of the expected values,

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

4. furthermore, if X_1 and X_2 are independent then

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2)$$

Binomial random variable

Let $X_1, \dots, X_n$ be mutually independent Bernoulli random variables, each with success probability π. Then

$$Y = \sum_{i=1}^{n} X_i$$

is a binomial random variable, denoted $Y \sim \mathit{Binomial}(n, \pi)$.

Y counts the total number of successes in n Bernoulli trials, therefore it should be apparent that $Y(S) = \{0, 1, \dots, n\}$. The pdf is

$$p_{Y}(y; n, \pi) = \binom{n}{y} \pi^{y} (1 - \pi)^{n-y}$$

Mean and variance

If $Y \sim Binomial(n, \pi)$, i.e, $Y = \sum_{i=1}^{n} X_i$ with X_i independent, identically distributed $Bernoulli(\pi)$, then

1. $E(Y) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \pi = n\,\pi$,

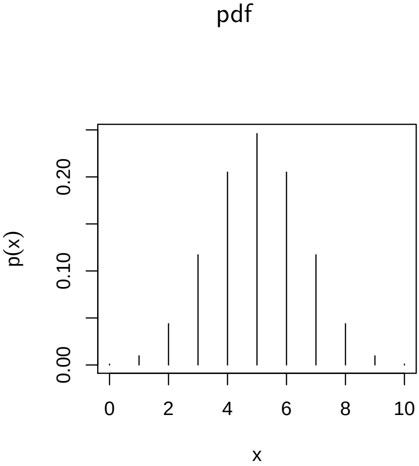
2. $Var(Y) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = n\,\pi\,(1-\pi)$.

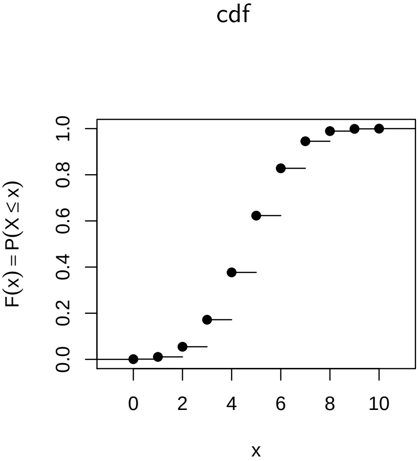
R commands to get....

```
n <- 10
x <- 0:n
p <- 0.5
pdf <- dbinom(x, n, p)
cdf <- pbinom(x, n, p)

plot(x, pdf, type = "h",
     xlab = expression(x), ylab = expression(p(x)))

plot(stepfun(x, c(0, cdf)), pch=19, vertical=FALSE,
     xlab = expression(x),  ylab = expression(F(x)==\Prob(X<=x)), main = " ")
```

For more details on the `expression()` function try `demo(plotmath)`

....these two graphs





Example

AA Airlines flies planes that seat 58 passengers. Years of experience have revealed that 20 percent of the persons who purchase tickets fail to claim their seat. (Such persons are called “no shows”.) Because of this phenomenon, AA routinely overbooks its flights, i.e., AA typically sells more than 58 tickets per flight. If more than 58 passengers show, then the “extra” passengers are “bumped” to another flight. Suppose that AA sells 64 tickets for a certain flight from Rome to London. How might AA estimate the probability that at least one passenger will have to be bumped?

Effect of the parameters: π

Try this to see how the pdf changes by changing the probability of success.

```
> binom.plot <- function(p) {
    plot(0:10, dbinom(0:10,10,p), ylim=c(0,0.5), type="h")
    Sys.sleep(0.1)
  }

> ignore <- sapply((0:100)/100, binom.plot)
```

Effect of the parameters: *n*

And this to see how the pdf changes by changing the number of trials.

```
> binom.plot <- function(n) {
    plot(0:n, dbinom(0:n,n,0.5), type="h")
    Sys.sleep(0.1)
  }

> ignore <- sapply(1:100, binom.plot)
```

The graph very quickly converges to a shape that changes very little except that it gradually becomes narrower, with respect to the range of possible values. This is the graphical manifestation of the *Central Limit Theorem*.

Poisson random variable

A random variable Y has Poisson distribution if $Y(S) = \{0, 1, \dots\}$ and

$$p_{Y}(y; \lambda) = \frac{\lambda^{y}}{y!} e^{-\lambda},$$

where $\lambda > 0$.

We write $Y \sim Poisson(\lambda)$.

Poisson random variable (cntd)

The Poisson distribution provides a model for the distribution of counts of randomly occurring events. Some examples of random variables which generally obey a Poisson distribution include:

- the number of decaying particles in a radioactive sample in a given time interval;
- the number of failures per day of a system;
- the number of people in a community who are older than 100 years;
- the number of bacterial colonies in a Petri dish;
- the number of nucleotide base substitutions in a gene over a period of time.

Mean and variance and more

1. $E(Y) = \lambda$.

2. $Var(Y) = \lambda$.

3. If $Y_1 \sim Poisson(\lambda_1)$ and $Y_2 \sim Poisson(\lambda_2)$, with Y_1 and Y_2 independent, then $Y_1 + Y_2 \sim Poisson(\lambda_1 + \lambda_2)$.

Poisson approximation to the Binomial

- The Poisson distribution is a good approximation to the binomial distribution when π is small.

- The approximation is better for large n.

- If p is small, then the binomial probability of exactly k successes is approximately the same as the Poisson probability of k with $\lambda = n\pi$.

Example

Suppose that $Y \sim \mathit{Binomial}(1000, 0.01)$. Find $\Pr(Y = 8)$. The exact calculation is:

$$\Pr(Y = 8) = \frac{1000!}{8!\ 992!} 0.01^8 0.99^{992} \doteq 0.112824.$$

Working with large factorials can be messy. The Poisson approximation

uses $\lambda = 1000 \times 0.01 = 10$ and is

$$\Pr(Y = 8) \approx \frac{10^8}{8!} e^{-10} \doteq 0.112599.$$

```
> dbinom(8, 1000, 0.01)
> dpois(8, 1000 * 0.01)
```

Quick Probability Review

– Some important continuous random variables –

Uniform random variable

A continuous random variable X has a uniform distribution on $X(S) = [a; b] \ -\infty < a < b < +\infty$, $X \sim U(a, b)$, if

$$p_x(x; a, b) = \frac{1}{b - a}$$

1. $E(X) = (a + b)/2$.
2. $Var(X) = (b - a)^2/12$.

Exponential random variable

A continuous random variable X has an exponential distribution with rate $\lambda,\ \lambda > 0$, i.e., $X \sim \mathit{Exp}(\lambda)$, if $X(S) = (0, +\infty)$ and the pdf of X is

$$p_x(x; \lambda) = \lambda\ e^{-\lambda x}.$$

1. $E(X) = \frac{1}{\lambda}$
2. $\mathit{Var}(X) = \frac{1}{\lambda^2}$

Exponential random variable (cntd)

The exponential random variable is used to model a continuous random variable whose density decreases away from 0. Some examples include:

- Customer service: time on hold at a help line
- Neurobiology: time until the next neuron fires
- Seismology: time until the next earthquake
- Medicine: remaining years of life for a cancer patient
- Ecology: dispersal distance of a seed

Be careful

The mean of the distribution, say θ, is the reciprocal of the rate λ, i.e., $\theta = 1/\lambda$.

Often, the mean is the parameter w.r.t. which the family of distributions is characterized.

The continuous random variable X has an exponential distribution with mean θ, $\theta > 0$, i.e., $X \sim \mathit{Exp}(\theta)$, if

$$p_x(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}.$$

Gamma random variable

Let $X_1, \ldots, X_n$ be mutually independent exponential rv's, each with rate $\lambda,\ \lambda > 0$. Then

$$Y = \sum_{i=1}^{n} X_i$$

is a gamma random variable, denoted $Y \sim Gamma(n, \lambda)$.

$$p_{_Y}(y; n, \lambda) = \frac{\lambda^n y^{n-1}}{\Gamma(n)} e^{-\lambda y}.$$

The $\Gamma(\cdot)$ function is a special mathematical function for which

1. $\Gamma(\alpha + 1) = \alpha \times \Gamma(\alpha),\ \alpha > 0$
2. $\Gamma(n) = (n-1)!$ if $n \in \{1, 2, 3, \ldots\}$

Gamma random variable (cntd)

The n parameter does not need to be an integer.

Indeed $Y \sim Gamma(\alpha, \lambda)$ has the following pdf

$$p_{_Y}(y; \alpha, \lambda) = \frac{\lambda^{\alpha} y^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda y}.$$

1. $E(Y) = \frac{\alpha}{\lambda}$
2. $Var(Y) = \frac{\alpha}{\lambda^2}$

Gamma random variable (cntd)

The Gamma distribution arises naturally in processes for which the waiting times between Poisson distributed events are relevant. Some examples include:

- the time it takes to recruit patients into a clinical trial;
- flow of objects through manufacturing and distribution processes;
- the load on many web servers;
- aggregate insurance claims;
- amount of rainfall accumulated in a reservoir.

But it has also a value in its own.

Normal distribution

A continuous random variable X is normally distributed with mean $\mu \in R$ and variance $\sigma^2 > 0$, denoted $X \sim N(\mu, \sigma^2)$, if $X(S) = (-\infty; +\infty)$ and the pdf of X is

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

By varying μ and σ^2 the normal distribution gives rise to the most important family of distributions in probability and statistics.

Normal distribution: properties

Many useful properties of normal distributions can be deduced directly from the pdf.

1. $p_x(x; \mu, \sigma^2) > 0$.
   It follows that, for any nonempty interval (a, b),
   $\Pr(X \in (a, b)) = \int_a^b p_x(x; \mu, \sigma^2) dx > 0$.

2. p_x is symmetric about μ.

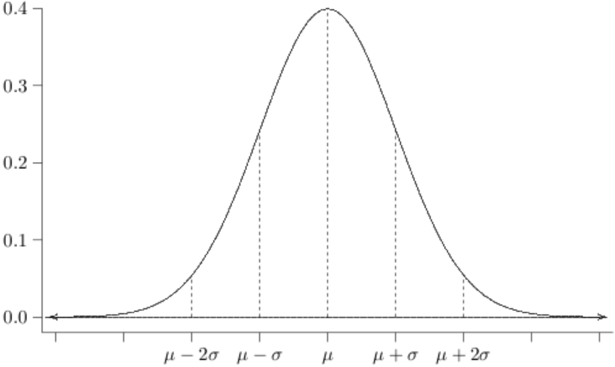
3. $p_x(x; \mu, \sigma^2)$ decreases as $|x - \mu|$ increases. In fact, the decrease is very rapid. We express this by saying that p_x has very light tails.

4. $\Pr(\mu - \sigma < X < \mu + \sigma) \doteq 0.683$.

5. $\Pr(\mu - 2\sigma < X < \mu + 2\sigma) \doteq 0.954$.

6. $\Pr(\mu - 3\sigma < X < \mu + 3\sigma) \doteq 0.997$.

Normal distribution (cntd)



Standard normal distribution

The standard normal distribution is the distribution $N(0, 1)$.

The following result is of enormous practical value.

If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

The transformation $Z = \frac{X-\mu}{\sigma}$ is called conversion to standard units. This transformation allows us to easily compute probabilities for arbitrary normal distributions.

In the following, let $\Phi(\cdot)$ denote the cdf of $Z \sim N(0, 1)$.

Examples

If $X \sim N(1,4)$, then what is the probability that X assumes a value no more than 3?

$\Pr(X \leq 3) = \Pr(\frac{X-1}{2} \leq \frac{3-1}{2}) = \Pr(Z \leq 1) = \Phi(1)$.

```
> pnorm(1)
[1] 0.8413447
> pnorm(3,mean=1,sd=2)
[1] 0.8413447
```

Examples (cntd)

If $X \sim N(2, 16)$, then what is the probability that X assumes a value between 0 and 10?

$\Pr(0 < X < 10) = \Phi(2) - \Phi(-0.5)$.

```
> pnorm(2)-pnorm(-.5)
[1] 0.6687123
```

Quick Probability Review

– Checking normality –

Empirical cdf

- Sometimes, we want to assess whether a data set is well modeled by a given distribution and, if not, how it differs.
- Most often, such distribution is the normal one.
- One obvious way to assess Normality is by looking at histograms or density estimates. But the answer is often not obvious from the figure. A better way to assess Normality is by using the empirical distribution function, that is the empirical quantiles.

Empirical quantiles

- Let us recall the definition of a quantile of level α, q_α say, for a generic continuous rv X

$$\Pr(X \leq q_\alpha) = F_X(q_\alpha) = \alpha.$$

- By applying the definition to the empirical cdf, it emerges that y is the empirical quantile of level $\frac{\#\{x \leq y\}}{n}$.

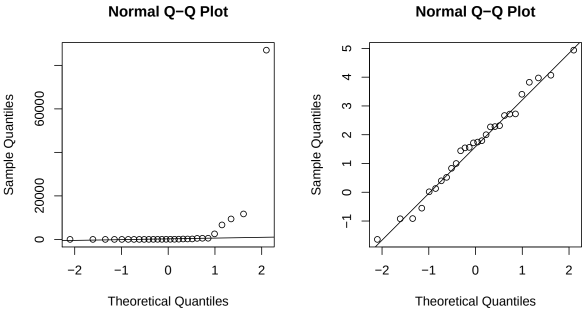
QQ plots

- QQ plots are used to compare the shapes of two distributions, most commonly by plotting the observed quantiles of an empirical distribution against the corresponding quantiles of a theoretical distribution.

- If the shape of the empirical distribution resembles that of the theoretical distribution, then the points in the QQ plot should tend to fall on a straight line.

- If a normal distribution is taken to be the reference distribution, the QQ plot is called a normal probability plot.

Example



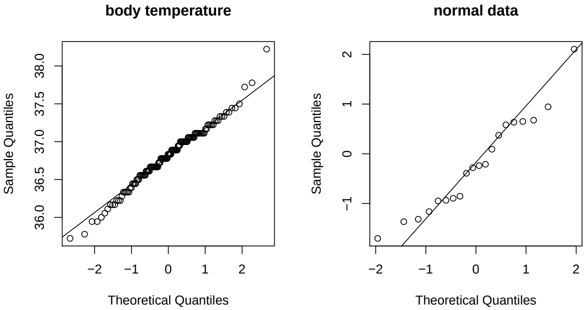
```
> library(MASS)
> data("Animals")
> attach(Animals)
> par(mfrow=c(1, 2))
> qqnorm(body)
> qqline(body)
> qqnorm(log10(body))
> qqline(log10(body))
> par(mfrow=c(1,1))
```

A word of caution

- When using normal probability plots, one must guard against overinterpreting slight departures from linearity. Remember: some departures from linearity will result from sampling variation.

- Consequently, before drawing definitive conclusions, the wise data analyst will generate several random samples from the theoretical distribution of interest in order to learn how much sampling variation is to be expected.

Body temperature and “true” normal data



```
> par(mfrow=c(1, 2))
> qqnorm(temp.C, main="body temperature")
> qqline(temp.C)
> x <- rnorm(20)
> qqnorm(x, main="normal data")
> qqline(x)
> par(mfrow=c(1,1))
```

Quick Probability Review

– Distributions related to the normal one –

Distributions related to the normal one

A number of important probability distributions can be derived by considering various functions of normal random variables. These distributions play important roles in statistical inference. They are rarely used to describe data; rather, they arise when analyzing data.

Chi-Squared distributions

Suppose that $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$ rv's and consider the continuous random variable

$$Y = Z_1^2 + \cdots + Z_n^2.$$

It is $Y(S) = [0; \infty)$. We want to know the distribution of Y.

The distribution of Y belongs to a family of probability distributions called the chi-squared family. This family is indexed by a single real-valued parameter, $\nu = n$, called the degrees of freedom parameter. We will denote a chi-squared distribution with ν degrees of freedom by χ^2_ν.

- The chi-squared distribution is a special case of the Gamma distribution, $\chi^2_\nu = Gamma(\frac{\nu}{2}, \frac{1}{2})$;
- we write $Y \sim \chi^2_\nu$.

Student's t distributions

Now let $Z \sim \mathit{Normal}(0,1)$ and $Y \sim \chi^2_\nu$ be independent random variables and consider the continuous random variable

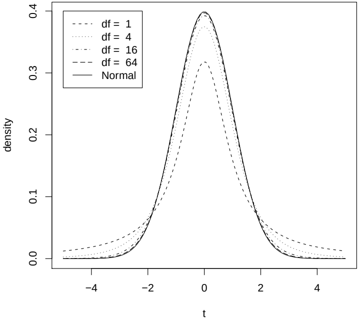
$$T = \frac{Z}{\sqrt{Y/\nu}}.$$

The set of possible values of T is $T(S) = (-\infty, +\infty)$. We are interested in the distribution of T.

The distribution of T is called a t distribution with ν degrees of freedom. We will denote this distribution by t_ν.

t distributions (cntd)

1. The t densities are unimodal and symmetric about 0, but have less mass in the middle and more mass in the tails than the $N(0,1)$ density.
2. In the limit, as $\nu \to \infty$, the t_ν density appears to approach the $N(0,1)$ density.



Fisher's F distributions

Finally, let $Y_1 \sim \chi^2_{\nu_1}$ and $Y_2 \sim \chi^2_{\nu_2}$ be independent random variables and consider the continuous random variable

$$F = \frac{Y_1/\nu_1}{Y_2/\nu_2}.$$

The set of possible values of F is $F(S) = [0, \infty)$. We are interested in the distribution of F.

The distribution of F is called an F distribution with ν_1 and ν_2 degrees of freedom. We will denote this distribution by F_{ν_1,ν_2}.

Quick Probability Review

– Linear combination of random variables –

Linear combination of random variables

A linear combination of the random variables $X_1, \dots, X_n$ is a random variable defined as

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b$$

where $a_1, \dots, a_n, b$ are real constants.

Relevant instances of linear combination of rv's:

- $S_n = \sum_{i=1}^n X_i$
- $\bar{X}_n = \frac{S_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$

Expected value of a linear combination of random variables

The expected value of $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b$ is

$$E(Y) = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n) + b.$$

For example, if $X_1, \ldots, X_n$ are identically distributed with $E(X_i) = \mu$ then

- $E(S_n) = n \times \mu$
- $E(\bar{X}_n) = \mu$

Variance of a linear combination of random variables

If the random variables $X_1, \ldots, X_n$ are independent, the variance of $Y = a_1 X_1 + a_2 X_2 \cdots + a_n X_n + b$ is equal to

$$Var(Y) = a_1^2 \; Var(X_1) + a_2^2 \; Var(X_2) \cdots + a_n^2 \; Var(X_n).$$

For example, if $X_1, \ldots, X_n$ are i.i.d. with $Var(X_i) = \sigma^2$ then

- $Var(S_n) = n \times \sigma^2$
- $Var(\bar{X}_n) = \frac{\sigma^2}{n}$

Linear combination of normally distributed random variables

If $X_1, \dots, X_n$ are independent and normally distributed then the linear combination

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b$$

is normally distributed.

For example, if $X_1, \dots, X_n$ are i.i.d., normally distributed, with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ then

- $S_n \sim N(n\mu, n\sigma^2)$
- $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Behaviour of the sample mean

Assume the true population is $N(5,1)$.

```
> curve(dnorm(x,5,1), xlim=c(2,8), ylim=c(0,10), col=2)
> samplingdist <- function(n) {
curve(dnorm(x,5,1), xlim=c(2,8), ylim=c(0,10), col=2)
curve(dnorm(x,5,1/sqrt(n)), xlim=c(2,8), ylim=c(0,10), add=TRUE)
Sys.sleep(0.2)}
> ignore <- sapply(1:20, samplingdist)
```

The Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be any sequence of i.i.d. random variables having finite mean μ and finite variance σ^2 and let

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Then

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} N(0,1) \quad \text{if } n \to \infty.$$

This gives a useful approximation for the distribution of $\bar{X}$, as

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \dot{\sim} N(0,1) \ ,$$

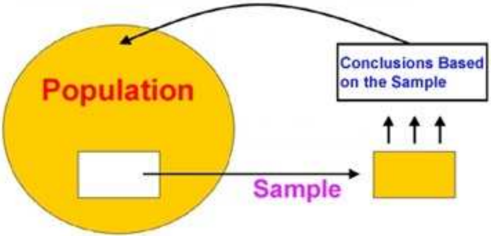
or, equivalently, $\bar{X} \dot{\sim} N(\mu, \sigma^2/n)$.

First steps in statistical inference

– Introduction –

Inferential Statistics

Statistical inference is the branch of statistics concerned with drawing conclusions and/or making decisions concerning a population based only on sample data.



The sample

- Census survey: attempt to gather information from each and every unit of the population of interest;
- sample survey: gathers information from only a subset of the units of the population of interest.

Why to use a sample?

1. Less time consuming than a census;
2. less costly to administer than a census;
3. measuring the variable of interest may involve the destruction of the population unit;
4. a population may be “infinite”.

Selecting a sample

A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.

Simple random sample: Each unit is chosen randomly and entirely by chance, such that each unit has the same probability of being chosen at any stage during the sampling process, and each subset of n units has the same probability of being chosen for the sample as any other subset of n individuals. For instance:

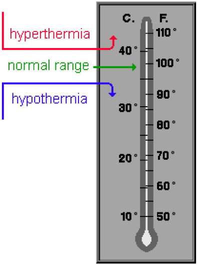
- extraction with replacement;
- extraction without replacement.

For large populations (compared to the sample size) the difference between these two sampling techniques is negligible. In the following we will always assume that samples are extracted with replacement from the population of interest.

Example: body temperature

Body temperature (degrees Celsius) in the general population of adult individuals:

1. What is the "typical" (that is the mean) body temperature of the population?
2. At what temperature should we consider someone's temperature to be "abnormal"?
3. Is the true population mean 36.75 °C?



Example: body temperature (cntd)

Formalization of the problem:

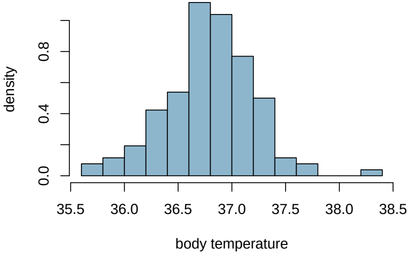
- Population: healthy adult individuals;
- Variable of interest, X: body temperature in degrees Celcius;
- Parameters of interest:
  - the population mean μ, that is $E(X)$, can be used to represent the "typical" body temperature;
  - the population variance σ^2, that is $Var(X)$, can be used to identify "normal" deviations about the mean.

Example: body temperature (cntd)

- Sample of $n = 130$ adult individuals

$$x_1 = 36.56 \quad x_2 = 35.94 \quad x_3 = 37.22 \quad \ldots \quad x_{130} = 36.06$$

- sample statistics: $\bar{x} = 36.81$ and $s = 0.41$;
- histogram of the data,



Example: proportion of “defective” items

- A clothing store chain regularly buys from a supplier large quantities of a certain piece of clothing. Each item can be classified either as good quality or top quality. The agreements require that the delivered goods comply with standards predetermined quality. In particular, the proportion of good quality items must not exceed 25% of the total.

- A large supply was delivered today:

  1. What is the proportion of good quality items of today supply?

  2. Should today supply be rejected because the proportion of good quality items exceeds 25%?

Example: proportion of defective items (cntd)

Formalization of the problem:

- Population: all the pieces of clothes of the consignment;
- Variable of interest, X: good/top quality of the item (binary variable);
- Parameter of interest:
  - proportion π of good quality items in the population.
  - If we set

    $$\text{good-quality} = 1 \quad \text{and} \quad \text{top-quality} = 0$$

    then $\pi = E(X)$.

Statistical inference: proportion of defective items (cntd)

- Sample of $n = 40$ items extracted from the consignment;

$$x_1 = 1 \quad x_2 = 0 \quad x_3 = 0 \quad \ldots \quad x_{40} = 1$$

- sample statistic: $p = \dfrac{11}{40} = 0.275$

- notice that:

$$\begin{aligned} p &= \text{proportion of good quality items in the sample} \\ &= \frac{\text{number of good quality items in the sample}}{n} \\ &= \frac{\sum_{i=1}^{40} x_1}{n} \\ &= \bar{x} \end{aligned}$$

Probabilistic representation of a sample

- The probability distribution of a random variable X is a mathematical abstraction of an experimental procedure for sampling from a population.

- When we extract a sample unit, we observe one of the possible values of X.

- To distinguish an observed value of a random variable from the random variable itself, we designate random variables by uppercase letters (X, Y, $\ldots$) and observed values by corresponding lowercase letters (x, y, $\ldots$).

Probabilistic representation of a sample (cntd)

- Before the sample is observed the sampling values are unknown and the sample can be regarded as a sequence of n random variables,

$$X_1,\ X_2,\ \ldots,\ X_n$$

- After the sample is observed the sampling values are known and therefore the sample is a sequence of values (each being the realization of a corresponding random variable),

$$x_1,\ x_2,\ \ldots,\ x_n.$$

- Easier to deal with this framework when $X_1, X_2, \ldots, X_n$ are i.i.d.

Data

- From now on, we will be concerned with experiments that are replicated a fixed number of times, n say. By replication, we mean that each repetition of the experiment is performed under identical conditions and that the repetitions are mutually independent. We will denote our sample as $x_1, \ldots, x_n$.

- This sample will be used to

  (i) to assess assumptions about the population (for example, to decide whether or not the population can plausibly be modeled by a normal distribution);

  (ii) to draw inferences about the population from which was sampled (for example, to guess the value of the population mean).

First steps in statistical inference

– Point estimation and sampling distributions –

Point estimation problem

Problem:

- unknown population parameter of interest such as, for example, μ, σ^2 or π;
- random sample from the population

$$x_1, x_2, \ldots, x_n$$

The objective of point estimation is to use the sample data to calculate a single value which is to serve as a estimate of an unknown population parameter.

Point estimation: example

- Body temperature:
  - to estimate μ, the population mean body temperature, we can use the corresponding sample mean $\bar{x} = 36.81$;
  - to estimate σ, the population standard deviation, we can use the corresponding sample standard deviation $s = 0.41$.

- Proportion of defective items:
  - to estimate π, the population proportion of good quality items, we can use the corresponding sample proportion $p = 0.275$.

Point estimation

To estimate the value of a population parameter, we compute a corresponding characteristic of the sample, referred to as a sample statistic.

- In the two examples considered, it is easy to find reasonable estimates of the unknown parameters. These are obtained by equating the population mean and variance with the corresponding sample statistics (recall that π is a mean), that is

$$\widehat{\mu} = \bar{x} \qquad \widehat{\sigma}^2 = s^2 \qquad \widehat{\pi} = p$$

- this is an application of the so called method of moments;
- we need general methods to build point estimates which, ideally, can be used in all estimation problems and that have general properties that we can rely on.

Assessing the uncertainty of point estimators

- The point estimates differ somewhat from the corresponding population parameters;
- this difference is to be expected because a sample, and not a census of the entire population, is being used to develop the point estimates;
- this difference is due to sampling variability and thus called sampling error;
- we would like to know how close to the true population parameter the point estimate is likely to be;
- assessing the uncertainty of point estimators is a fundamental aspect of point estimation procedures.

Sampling distribution of a statistic

- The sampling distribution of a statistic is the distribution of that statistic considered as a random variable (when derived from a random sample of size n);

- it may be considered as the distribution of the statistic for all possible samples from the same population of a given sample size.

- The sampling distribution depends on:

    1. the underlying distribution of the population;
    2. the statistic being considered;
    3. the sampling procedure employed;
    4. the sample size used.

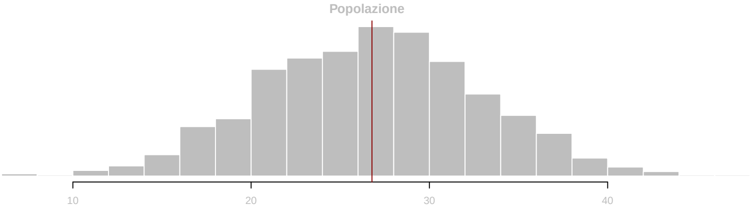
Behaviour of the sample mean

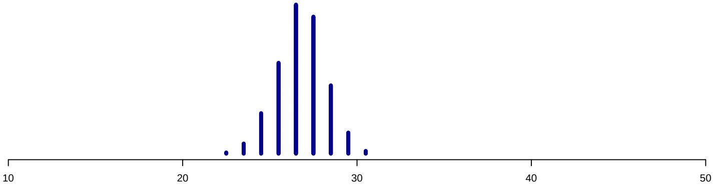
Let us first concentrate on the mean μ, that we have estimated with $\hat{\mu}$.

To answer the previous questions, we should analyze the behaviour of the sample means over all the possible samples of size n taken from a $N(\mu, \sigma^2)$.

Let us assume to sample 20 observations from a population with mean 26.778. We repeat the sampling 1000 times.

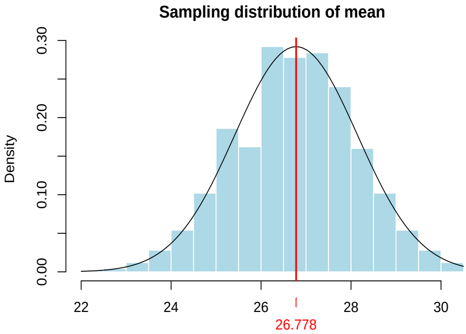
Example





Sampling distribution

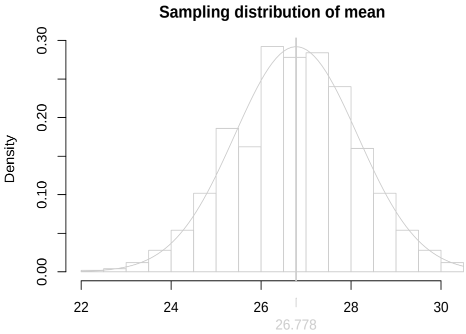
The 1000 means computed on our 1000 samples have the following distribution



- Values around the true mean, 26.78, are more frequent than values far apart.
- Assuming to have an infinity of samples, we can switch to a density representation.

Sampling distribution

The 1000 means computed on our 1000 samples have the following distribution



In reality

- the mean of the population is unknown,
- only one sample is observed.

Sampling distribution

- Mathematically, the simulation exercise that we performed is equivalent to, first, defining a new random variable, the estimator

$$\hat{\mu} = \bar{X},$$

of which the observed value $\bar{x}$ is a realization, and, then, to studying its distribution.

- This can be a difficult task. Still, in some situations, it is analytically tractable.

- If the distribution of X is normal, then also $\bar{X}$ follows a normal distribution, that is,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Standard error of the mean

- The standard deviation of $\bar{X}$ is referred to as the standard error of the mean

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- the standard error is a measure of the uncertainty associated with the estimation process, due to sampling variability.

- the value of the standard error of the mean is helpful in determining how far the sample mean may be from the population mean.

Point estimation of a mean with σ^2 known: example

In the "body temperature" example, assume that the body temperature X is such that

- $X \sim N(\mu, \sigma^2)$
- the value of the variance is known and equal to $\sigma^2 = 0.20$ so that $\sigma = 0.45$

Then a point estimate of the typical body temperature, μ, is

- $\widehat{\mu} = \bar{x} = 36.81$
- and the standard error of this estimate is

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{130}} = \frac{0.45}{\sqrt{130}} = 0.039$$

Point estimation of a mean with σ^2 unknown

- Typically the value of σ^2 is not known;
- in this case we estimate it as $\hat{\sigma}^2 = s^2$;
- this can be used, for instance, to estimate the standard error of $\hat{\mu}$

$$\widehat{SE}(\bar{X}) = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}.$$

- In the "body temperature" example, the standard error can be estimated as

$$\widehat{SE}(\bar{X}) = \frac{0.39}{\sqrt{130}} = 0.034$$

Normal rv's

If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then

1. $\bar{X} \sim N(\mu, \sigma^2/n)$, or equivalently $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
2. $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ where $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$
3. $\bar{X}$ and S^2 are independent rv's
4. $\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

First steps in statistical inference

– Confidence intervals for the mean –

Interval estimation

- A point estimate consists of a single value, so that
  - if $\bar{X}$ is a point estimator of μ then it holds that

$$\Pr(\bar{X} = \mu) = 0$$

  - more generally, $\Pr(\hat{\theta} = \theta) = 0$.

- Interval estimation is the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter.

Confidence interval for the mean of a normal population (σ known)

- $X_1, \ldots, X_n$ simple random sample with $X_i \sim N(\mu, \sigma^2)$;
- assume σ known;
- a point estimator of μ is

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- the standard error of the estimator is $SE(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$

Formal derivation of the 95% CI for μ: σ known

It holds that

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{(recall that} \quad \frac{\sigma}{\sqrt{n}} \text{ is the Standard Error)}$$

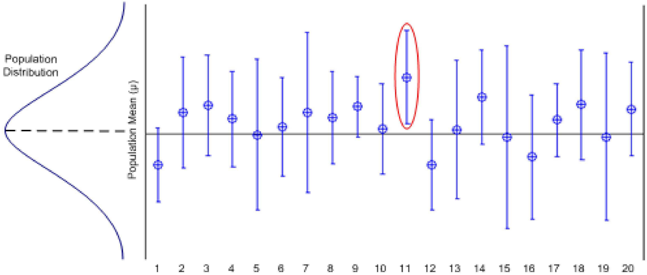
so that

$$\begin{aligned}
0.95 &= \Pr\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) \\
&= \Pr\left(-1.96\,\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq 1.96\,\frac{\sigma}{\sqrt{n}}\right) \\
&= \Pr\left(-\bar{X}-1.96\,\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}+1.96\,\frac{\sigma}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{X}-1.96\,\frac{\sigma}{\sqrt{n}} \leq \ \mu \ \leq \bar{X}+1.96\,\frac{\sigma}{\sqrt{n}}\right)
\end{aligned}$$

Interpretation of confidence intervals

Probability is associated with the procedure that leads to the derivation of a confidence interval, not with the interval itself. A specific interval either will contain or will not contain the true parameter, and no probability is involved in a specific interval.

If the procedure were repeated on a large number of different samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward 0.95 (more generally, $1 - \alpha$).



Confidence interval for μ: level $1-\alpha$ and σ known

A confidence interval at the (confidence) level $1-\alpha$, or $(1-\alpha)\%$, for μ is given by

$$\left(\bar{X} - z_{1-\alpha/2}\, SE; \quad \bar{X} + z_{1-\alpha/2}\, SE\right)$$

Since $SE = \frac{\sigma}{\sqrt{n}}$ then

$$\left(\bar{X} - z_{1-\alpha/2}\, \frac{\sigma}{\sqrt{n}}; \quad \bar{X} + z_{1-\alpha/2}\, \frac{\sigma}{\sqrt{n}}\right)$$

Margin of error

- The confidence interval

$$\bar{x} \pm z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

- can also be written as $\bar{x} \pm ME$ where

$$ME = z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

  is called the margin of error.

- the interval width is equal to twice the margin of error.

- Reducing the margin of error: the margin of error can be reduced, without changing the accuracy of the estimate, by increasing the sample size ($n \uparrow$).

Confidence interval for μ with σ unknown

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$;
- $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$;
- in this case the standard error is unknown and needs to be estimated.

$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ is estimated by $\widehat{SE}(\bar{X}) = \frac{\hat{\sigma}}{\sqrt{n}}$ where $\hat{\sigma} = S$

- and it holds that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Confidence interval for μ with σ unknown

$$1-\alpha \quad = \quad \Pr\left(-t_{n-1;1-\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{n-1;1-\alpha/2}\right)$$

where $t_{n-1;1-\alpha/2}$ is the value such that the area under the t pdf, with $n-1$ d.f. between $t_{n-1;1-\alpha/2}$ and $+\infty$ is equal to $1-\alpha/2$.

Hence, a confidence interval at the level $(1-\alpha)$ for μ is

$$\left(\bar{X} - t_{n-1;1-\alpha/2}\,\frac{S}{\sqrt{n}}; \quad \bar{X} + t_{n-1;1-\alpha/2}\,\frac{S}{\sqrt{n}}\right)$$

First steps in statistical inference

– Testing an hypothesis on the mean–

Example of decision problem

From Wikipedia:

> *Normal human body temperature, also known as normothermia or euthermia, is the typical temperature range found in humans. The normal human body temperature range is typically stated as* $36.5 - 37.0^\circ C$

Problem concerning the “typical” (that is “mean”) body temperature of healthy adult individuals.

Scientific hypothesis: the typical (mean) temperature is 36.75°C. Is that true for the population of interest?

- We want to make a decision about the scientific hypothesis;
- the decision needs to be made on the basis of the information provided by a simple random sample, which in this case has size $n = 130$.

Statistical hypotheses

A decisional problem in specified by means of two statistical hypotheses:

- the null hypothesis H_0: this will often be the hypothesis of "no change" or "no effect".
- The null hypothesis is contrasted to the "opposite" hypothesis, the alternative hypothesis H_1 (also denoted by H_a).

The two hypotheses concern the value of an unknown population parameter, for instance μ,

$$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu \neq \mu_0 \end{cases}$$

where μ_0 is the hypothesized value, for example in the body temperature example $\mu_0 = 36.75$.

Observed value of the sample mean

- the observed value of the sample mean is $\bar{x}$ (in the example $\bar{x} = 36.81$);
- even in the case where H_0 is true, $\bar{x}$ cannot be expected to be identically equal to μ_0 (that is to 36.75 in the example);
- under H_0 the expected value of $\bar{X}$ is equal to μ_0 and the difference between μ_0 and $\bar{x}$ is uniquely due to the sampling error.

Hence the sampling error is

$$\bar{x} - \mu_0$$

that is

observed value *"minus"* expected value

(in the example $36.81 - 36.75 = 0.06$).

Test of hypothesis

- A test of a statistical hypothesis is a rule for rejecting H_0, based on the observations $x_1, x_2, \ldots, x_n$.

- For example, body temperatures are assumed to be normally distributed. We know from the theory that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

- so that, if the null hypothesis was true, we know that

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

  our statistic has a t-distribution with $n - 1$ degrees of freedom.

- T is called a test statistic.

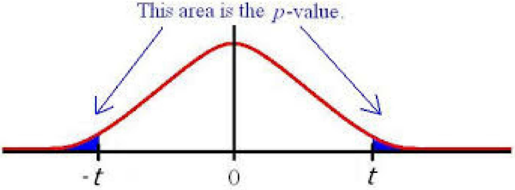
The p-value

The p-value, also called observed level of significance is the probability of obtaining a value of the test statistic more extreme than the observed sample value, under H_0, formally

$$p = \Pr(|T| > |t^{obs}| \mid H_0).$$

where, in this case H_0 means $\mu = \mu_0$

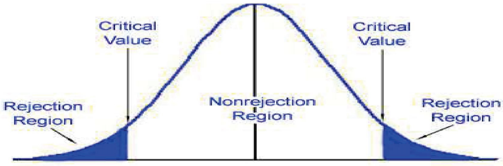
- A small p-value indicates strong evidence against the null hypothesis, so one should reject the null hypothesis.
- A large p-value indicates weak evidence against the null hypothesis, so one should fail to reject the null hypothesis.



Decision rule

1. Set a significance level α

2. determine the critical value c such that $\Pr(|T| > c \mid \mu = \mu_0) = \alpha$



3. Reject H_0 if the observed value of $|T|$ is greater than c.

- Note that: we reject the null hypothesis at level α if and only if the p-value is less than or equal to α;

Outcomes and probabilities

There are two possible "states of the world" and two possible decisions. This leads to four possible outcomes.

	H_0 TRUE	H_0 FALSE
H_0 IS REJECTED	Type I error $\boldsymbol{\alpha}$	OK
H_0 IS NOT REJECTED	OK	Type II error $\boldsymbol{\beta}$

The probability of the type I error (significance level) of the test and can be arbitrarily fixed (typically 5%).

Significance level

Note that there is no “magic” significance level at which null hypotheses should be automatically rejected. Often people will use a significance level of $\alpha = 0.05$ at which they reject the null hypothesis, however this is completely arbitrary. The significance level used should always take into account the consequences of rejecting the null-hypothesis.

At the least you should qualify any statements, for example:

$0.01 < p$-value < 0.05 evidence for rejection;

$0.001 < p$-value < 0.01 strong evidence for rejection;

p-value < 0.001 very strong evidence for rejection.

Beware of confusing statistical and practical significance.

One-sample Test of Location

Suppose $X_1, X_2, \dots, X_n$ are iid $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. To test $H_0 : \mu = \mu_0$ against the alternative, we use the test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}, \text{ under } H_0.$$

Example. For the body temperature, we wish to test $H_0 : \mu = 36.75$ against $H_1 : \mu \neq 36.75$. The value of the test statistic is 1.54 and the p-value is

$$p = \Pr(|T| > 1.54 \mid \mu = 36.75) = 0.13$$

Hence, there is no evidence for rejection of the null hypothesis.

Different forms of the alternative hypotheses

Two sided hypothesis:

- $\boxed{H_1 : \mu \neq \mu_0}$
- $p - \text{value} = \Pr(|T| > |t^{obs}|)$

One-sided hypothesis (right)

- $\boxed{H_1 : \mu > \mu_0}$
- $p - \text{value} = \Pr(T > t^{obs})$

One-sided hypothesis (left)

- $\boxed{H_1 : \mu < \mu_0}$
- $p - \text{value} = \Pr(T < t^{obs})$

First steps in statistical inference

– The case of non normal populations –

More on the sampling distribution of the mean

We saw that, if $X_1, \dots, X$ are i.i.d. $N(\mu, \sigma^2)$, then

1. $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$,
2. $\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

But what if $X_1, \dots, X_n$ are i.i.d. from a different distribution?

Some results help to find an answer, at least approximately, in the cases where the distribution of an estimator is difficult to work out directly.

Remember the Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be any sequence of i.i.d. random variables having finite mean μ and finite variance σ^2 and let $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$. Then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1) \quad \text{if } n \to \infty.$$

This gives a useful approximation for the distribution of $\bar{X}$, as

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1) \ ,$$

or, equivalently, $\bar{X} \dot{\sim} N(\mu, \sigma^2/n)$.

Example

To see that more clearly, let's make 1000 simulations of the sample mean in $n = 5; 10; 25; 100$ Bernoulli trials X_i with $\pi = 0.2$.

By the CLT, we know that

$$\hat{\pi} = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{d} N(\pi, \pi(1-\pi)/n),$$

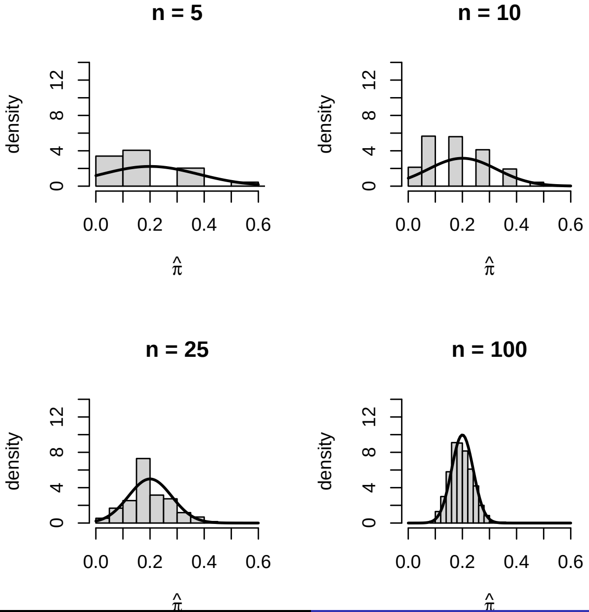
We will build the sampling distribution of $\hat{\pi}$ by simulation, make

histograms of those simulations and overlay them with Normal densities.

R commands for sample size 5

```
> sampsize <- 5
> n.sim <- 1000
> p.true <- .2
> x <- matrix (rbinom(n.sim*sampsize, 1, p.true),
+                nrow=n.sim, ncol=sampsize)
> hat.p <- apply (x, 1, mean)
> hist (hat.p, prob=T,
+          xlim=c(0,.6), xlab=expression(hat(pi)),
+          ylim=c(0,14), ylab="density", main="n = 5")
> m <- p.true
> sd <- sqrt(p.true*(1-p.true)/sampsize)
> curve(dnorm(x, m, sd), lty=1,  add=T, lwd=2)
```

Distributions of the sample mean



Interpretation

Notice that the Normal approximation is not very good for small n. That's because the underlying distribution is highly skewed, nothing at all like a Normal distribution.

But for large n, the Normal approximation is quite good. That's the Central Limit Theorem kicking in.

CLT in practice

In pratice, statisticians use a result (due to Slutsky) according to which, if $\hat{\sigma}$ is a "good guess" for σ (or, more formally , $\hat{\sigma}$ converges in probability to σ) then

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \dot{\sim} N(0, 1) \ ,$$

for n sufficiently big.

Equivalently,

$$\bar{X} \dot{\sim} N(\mu, \hat{\sigma}^2/n).$$

The result tells us that we can use the estimate of the variance to compute a good approximation to the sampling distribution of $\hat{\mu}$. For small n, that estimate doesn't help us much. But for n big, it tell us a lot about the accuracy of the rv $\hat{\mu}$, and the Normal approximation computed from sample is a good match to the sampling distribution of $\hat{\mu}$.

CLT (cntd)

This result is of considerable consequence. It states that, as we average more and more X_i, the average values that we observe tend to be distributed closer and closer around the theoretical average of the X_i. This property of the sample mean strengthens our contention that the behavior of $\bar{X}$ provides more and more information about the value of μ as n increases.

Of course, we require some sense of how many random variables must be averaged in order for the normal approximation to be reasonably accurate. This does depend on the distribution of the random variables, but a popular rule of thumb is that the normal approximation can be used if $n > 30$.

Vitamin C and the Common Cold

In a Canadian experiment, 407 volunteers received at the beginning of the winter a supply of vitamin C pills, adequate to last through the entire cold season at 1000mg per day.

At the end of the season, each subject was interviewed to decide whether the subject had or had not suffered a cold during the period. 302 subjects reported to have suffered a cold.

Which is a reasonable estimate of the probability π of catching a cold while on a vitamin C regimen?

Vitamin C and the Common Cold

Assume that $X_1, \dots, X_n$, with $n = 407$ are i.i.d *Bernoulli*(π). Then, $Y = \sum_{i=1}^n X_i \sim \mathit{Binomial}(n, \pi)$.

It is

$$\hat{\pi} = \frac{\sum_{i=1}^n X_i}{n} = \frac{Y}{n}.$$

A confidence interval for π of approximate level 0.95 is therefore

$$\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n} = (0.70, 0.78).$$

Test for a proportion

- Null hypotheses: $H_0 : \pi = \pi_0$;

- Under H_0 the sampling distribution of $\hat{\pi}$ is approximately normal with expected value $E(\hat{\pi}) = \pi_0$ and standard error

$$SE(\hat{\pi}) = \sqrt{\frac{\pi_0 \, (1 - \pi_0)}{n}}$$

Note that under H_0 there are no unknown parameters.

Another distribution: exponential distribution

Assume to have a simple random sample from an exponential distribution.

The exponential rv with mean θ has pdf

$$p_x(x, \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \ \theta > 0, \quad x > 0.$$

The mean of X is θ and the variance is θ^2.

We can estimate the mean θ with the sample mean $\bar{x}$, i.e., $\hat{\theta} = \bar{x}$, and the variance with $\hat{\theta}^2 = \bar{x}^2$,

From CTL, we have

$$\bar{X} \dot{\sim} N(\theta, \hat{\theta}^2/n),$$

and

$$\Pr\left(\bar{X} - 1.96\frac{\hat{\theta}}{\sqrt{n}} \leq \theta \leq \bar{X} + 1.96\frac{\hat{\theta}}{\sqrt{n}}\right) \approx .95$$

Example: life times

The lifetime in months of one electronic component is measured on 8 products. The following sample is obtained.

$$6,\ 0,\ 1,\ 7,\ 3,\ 5,\ 2,\ 1$$

It is believed that the mean lifetime is around 2 months. Is there any evidence against this hypothesis?

Example: solution

It is $\bar{x} = 3.125$ so that a confidence interval of approximate level 0.95 is (0.95, 5.29).

A test statistics can be constructed

$$T = \frac{\bar{X} - \theta}{\hat{\theta}/\sqrt{n}} \dot{\sim} N(0,1)$$

It is $t^{obs} = 1.018234$ with a p-value $= 0.31$.

This value is considerably high, so the null hypothesis is not rejected.

First steps in statistical inference

– More general problems –

So far...

We have seen inference (estimation, CI, tests) on the mean of a population. The steps we have taken are

1. definition of an estimator, i.e., a function of the observations, the sample mean in our case;
2. construction of its sampling distribution;
3. inference.

Take home message

The sampling distribution of the estimator is our mean of building inference tools. Accuracy of inference depends on the behavior of the distribution of estimator, where it is located and how spread it is: if the distribution is tightly concentrated around the true value, inference is highly accurate.

By following the same lines...

...we can tackle a wealth of problems, such as, for example,

1. inference on the variance of Normal populations;
2. compare two variances;
3. compare two means in:
    - one sample
    - two sample

Inference on the variance of a Normal population

If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then

1. $\bar{X}$ and $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ are independent rv's,
2. $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$,

So we have an estimator, i.e., s^2, and a sampling distribution for a monotone transformation of it, i.e., $\dfrac{(n-1)S^2}{\sigma^2}$.

According to what we saw, this should be enough.

95% CI for the variance of a Normal population

Need to find

$$\Pr\left(k_1 < \frac{(n-1)S^2}{\sigma^2} < k_2\right) = 0.95$$

Hence,

$$\begin{aligned} 0.95 &= \Pr\left(\chi^2_{n-1;0.025} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1;0.975}\right) \\ &= \Pr\left(\frac{1}{\chi^2_{n-1;0.975}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_{n-1;0.025}}\right) \\ &= \Pr\left(\frac{(n-1)S^2}{\chi^2_{n-1;0.975}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1;0.025}}\right) \end{aligned}$$

$(1-\alpha)\%$ CI for the variance of a Normal population

Need to find

$$\Pr\left(k_1 < \frac{(n-1)S^2}{\sigma^2} < k_2\right) = 1-\alpha$$

Hence,

$$\begin{aligned}
1-\alpha &= \Pr\left(\chi^2_{n-1;\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1;1-\alpha/2}\right) \\
&= \Pr\left(\frac{1}{\chi^2_{n-1;1-\alpha/2}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_{n-1;\alpha/2}}\right) \\
&= \Pr\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right)
\end{aligned}$$

CI for the variance of a Normal population

An confidence interval of exact level $(1-\alpha)$ is therefore given by

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}, \;\; \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} \right)$$

Of course, tests on σ^2 can also be derived.

Exercise: waiting time

A post office is not only interested in reducing the waiting times of its customers but also that the waiting times of different customers are as similar as possible to each other.

With individual lines at its various windows, the post office finds that the normally distributed waiting times for customers on Friday afternoon has mean $\mu = 15.3$ and standard deviation $\sigma = 7.2$ minutes.

The post office experiments with a single, main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have mean $\bar{x} = 14.8$ and standard deviation of $s = 4.9$ minutes.

With a significance level of 5%, test the claim that a single line implies lower variation among waiting times for customers.

Waiting time: solution

The system of hypothesis is

$$\begin{cases} H_0 : & \sigma^2 = 7.2^2 \\ H_1 : & \sigma^2 < 7.2^2 \end{cases}$$

Statistical test

$$t^{obs} = \frac{(n-1)s^2}{7.2^2} = \frac{24 \times 4.9^2}{7.2^2} = 11.11$$

and from the χ^2_{24} we can see that $p - \text{value} = 0.012$ and therefore at the level 5% we reject H_0.

Exercise: glucose monitor

Diabetic patients monitor their blood sugar levels with a home glucose monitor which analyses a drop of blood from a finger stick. Although the monitor gives precise results in a laboratory, the results are too variable when it is used by patients.

A new monitor is developed to improve the precision of the assay results under home use. Home testing on the new monitor is done by 25 persons using drops from a sample having a glucose concentration of 118 mg/dl. If $\sigma < 10$ mg/dl, then the precision of the new device under home use is better than the current monitor.

The readings from 25 tests have mean $\bar{x} = 118.5$ and standard deviation $s = 6.2$. Test $H_0 : \sigma = 10$ vs $H_1 : \sigma < 10$ at the 0.10 level.

Glucose monitor: solution

Assuming that data come from a Normal distribution, the test statistic is

$$t^{obs} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{24 \times 6.2^2}{10^2} = 9.2$$

Small values of T are evidence against H_0. The lower critical point is

$$\chi^2_{n-1;\alpha} = \chi^2_{24;0.1} = 15.65868$$

The p-value is given by

$$P(T < 9.2 \mid H_0) = 0.00286.$$

This value is considerably less than 0.1, so the null hypothesis is rejected at the level 0.1.

The two-sample t-test

Suppose we want to compare the locations of

- two samples
  - $x_1, x_2, \ldots, x_{n_1}$
  - $y_1, y_2, \ldots, y_{n_2}$
- assume that the observations are realizations of independent normal random variables
  - $X_i \sim N(\mu_1, \sigma_1^2), \quad i = 1, \ldots, n_1$
  - $Y_j \sim N(\mu_2, \sigma_2^2), \quad j = 1, \ldots, n_2$

The two-sample t-test (cont.)

We are interested in testing the hypothesis

$$\begin{cases} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 \neq \mu_2 \end{cases}$$

We have

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

Firstly, we will consider the case that we know (or have tested) that the variances are the same, that is

$$\sigma_1^2 = \sigma_2^2 = \sigma^2.$$

The two-sample t-test (homoschedasticity)

Hence, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ we have

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2,\;\; \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

Assuming that H_0 holds,

$$T = \frac{(\bar{X} - \bar{Y})}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where S^2 is the pooled estimate of the variance, i.e.

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

.

The two-sample t-test: example

Consider the results of two experiments (which we know should have the same variance) to measure the concentration of a chemical:

Exp. 1	22	19	35	11	21	10
Exp. 2	33	11	20	38		

We wish to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$.

The observed value of the test statistic T is $t^{obs} = -0.87$ which, on comparing with a t_8 distribution, gives a p-value of 0.41. Hence there is not enough evidence to reject H_0.

Test of equal variances

Suppose, as before, that we have two independent samples from two normal populations. We wish to test

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

In this case, we can use the test statistic

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1} \text{ under } H_0.$$

The further the value of the test statistic is from 1, the stronger the evidence for unequal variances.

Returning to the chemical example, we might wish to test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$. In this case, the observed value of the test statistic $s_1^2/s_2^2 = 0.545$, and comparing this with a $F_{5,3}$ distribution gives a p-value of 0.52.

The Welch's t-test for two-samples (heteroschedasticity)

If the variances are not the same, then we can use the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

- its distribution under H_0 can be approximated by a t_ν distribution
- ν is given by an expression involving the n_i and s_i^2 (no further details here)
- more robust than the standard t-test
- this is the default in R: `t.test(x, y, var.equal = FALSE)` is the same as `t.test(x, y)`

The paired t-test

Example. Two types of rubber were randomly assigned to the left and right shoes of 10 boys and the relative wear on each measured.

Boy	1	2	3	4	5	6	7	8	9	10
Left	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
Right	14.0	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

We wish to test the hypothesis of no difference in mean of wear, $H_0 : \mu_1 = \mu_2$, against $H_1 : \mu_1 \neq \mu_2$.

The paired t-test

Here, we have paired data $(x_i, y_i), i = 1, \dots, 10$, realizations of paired rv $(X_i, Y_i), i = 1, \dots, 10$. As X and Y are measured on the same subject, the two samples are not independent and the two sample t-test cannot be applied.

A possible solution relies on considering the differences $D_i = X_i - Y_i$.

If $D_i = X_i - Y_i$, $i = 1, \dots, 10$, are iid from a $N(\mu_D, \sigma_D^2)$, with $\mu_D = \mu_1 - \mu_2$ then the hypothesis $H_0 : \mu_1 = \mu_2$, can be verified by testing $H_0 : \mu_D = 0$, through a one sample location test on D_i, $i = 1, \dots, 10$.

In practice

From the data we have $\bar{d} = -0.41$ and $s_D^2 = 0.15$, so $t^{obs} = -3.34$. The p-value

$$P(|T| > 3.34 | \mu_D = 0) = 0.0085,$$

so there is strong evidence to reject the hypothesis that wear is the same for both types of rubber.

```
> x <- c(13.2, 8.2, 10.9, 14.3, 10.7, 6.6, 9.5, 10.8,
    8.8, 13.3)
> y <- c(14.0, 8.8, 11.2, 14.2, 11.8, 6.4, 9.8, 11.3,
    9.3, 13.6)
> t.test(x,y,paired=TRUE)
 Paired t-test
data:  x and y
t = -3.3489, df = 9, p-value = 0.008539
alternative hypothesis: true difference in means is
   not equal to 0
```

Vitamin C and the Common Cold

The Canadian experiment on the use of vitamin C for preventing colds was much more complex. It involved 818 volunteers, randomly divided into two groups, a vitamin C group, and a placebo group, which received inert pills (appropriately flavored). Neither the subject nor the doctor knew the treatment that the subject received (double-blind experiment).

Results are shown in the following table.

	Outcome		
	Cold	No Cold	Totals
Placebo	335	76	411
Vitamin C	302	105	407
Totals	637	181	818

Is there evidence of a difference between the two population proportions? If yes, can this difference be attributed to differences in treatments?

Vitamin C and the Common Cold (cntd)

Assume that $X_1, \ldots, X_{n_c}$, with $n_c = 407$ are i.i.d *Bernoulli*(π_c) $(c =$ vitamin C$)$ and $Y_1, \ldots, Y_{n_p}$, with $n_p = 411$ are i.i.d *Bernoulli*(π_p) $(p =$ placebo$)$, independent from the X_i's

A test statistic for the hypothesis $H_1 : \pi_p > \pi_c$ can be built on exploiting the result

$$\hat{\pi}_p - \hat{\pi}_c \,\dot{\sim}\, N\left(\pi_p - \pi_c, \frac{\pi_p(1-\pi_p)}{n_p} + \frac{\pi_c(1-\pi_c)}{n_c}\right).$$

Under $H_0 : \pi_p = \pi_c$, the best estimate of the standard error of $\hat{\pi}_p - \hat{\pi}_c$ is obtained by using $\hat{\pi}_{pooled}$, i.e., the sample proportion from the combined sample.

The test statistic is

$$T = \frac{\hat{\pi}_p - \hat{\pi}_c}{\sqrt{\frac{\hat{\pi}_{pooled}(1-\hat{\pi}_{pooled})}{n_p} + \frac{\hat{\pi}_{pooled}(1-\hat{\pi}_{pooled})}{n_c}}}$$

which has an approximate standard Normal distribution under H_0.

Vitamin C and the Common Cold (cntd)

$$\hat{\pi}_c = \frac{302}{407} = 0.742$$

$$\hat{\pi}_p = \frac{335}{411} = 0.815$$

$$\hat{\pi}_{pooled} = \frac{302 + 335}{407 + 411} = 0.7789$$

$$\hat{\pi}_p - \hat{\pi}_c = 0.073$$

$$se = \sqrt{\frac{\hat{\pi}_{pooled}(1 - \hat{\pi}_{pooled})}{n_p} + \frac{\hat{\pi}_{pooled}(1 - \hat{\pi}_{pooled})}{n_c}} = 0.029$$

$$t^{obs} = \frac{0.073}{0.029} = 2.517241$$

Some general ways to build inference tools

Open questions

- What if the parameter of interest is not the mean/the population is not normal?
  How to construct estimators for general parameters in general models?

- What if we have more than one estimator?
  How to compare estimators?

- What if we cannot analytically derive a (approximate) sampling distribution of the estimator?
  How to construct sampling distributions in general situations?

First steps in statistical inference

– Bias and mean squared error –

Comparing variances of estimators

- Let θ be the parameter to estimate on the basis of a sample $X_1, \ldots, X_n$.

- Let $\hat{\theta}^{(a)} = T^{(a)}(X_1, \ldots, X_n)$ and $\hat{\theta}^{(b)} = T^{(b)}(X_1, \ldots, X_n)$ be two different estimators of θ.

- How can we choose between $\hat{\theta}^{(a)}$ and $\hat{\theta}^{(b)}$?

- Does it make sense to compare estimators through their variances $Var(\hat{\theta}^{(a)})$ and $Var(\hat{\theta}^{(b)})$?

- We can note that if we set $\hat{\theta}^{(c)} =$ constant regardless of the sample, for instance $\hat{\theta}^{(c)} = 10$, then it holds that $Var(\hat{\theta}^{(c)}) = 0$. When comparing estimators we should also consider their bias.

Let us start to be more formal: the bias

Let θ indicate the parameter to estimate and $\hat{\theta} = T(X_1, \dots, X_n)$ be the estimator.

Definition

The bias of $\hat{\theta}$ is the difference between its expectation and the "true" values, i.e.,

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

An estimator $\hat{\theta}$ is unbiased for θ if $E(\hat{\theta}) = \theta \quad \forall\, \theta$.

Example: $X_1, X_2, \dots, X_n$ i.i.d. $N(\mu, \sigma^2)$. It is $E(\bar{X}) = \mu$, so $\bar{X}$ is unbiased.

Unbiasedness of S^2

Example: $X_1, X_2, \ldots, X_n$ i.i.d. with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$.

Let $\tilde{S}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$. Is $\tilde{S}^2$ an unbiased estimator of σ^2?

$$\begin{aligned}\tilde{S}^2 &= \frac{\sum_{i=1}^n X_i^2 + n\bar{X}^2 - 2\bar{X}\sum_{i=1}^n X_i}{n} \\ &= \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n}\end{aligned}$$

Recall that

1. $Var(X) = E(X^2) - E(X)^2$ so that $E(X^2) = Var(X) + E(X)^2$
2. $Var(\bar{X}) = E(\bar{X}^2) - E(\bar{X})^2$ so that $E(\bar{X}^2) = Var(\bar{X}) + E(\bar{X})^2$
3. $Var(\bar{X}) = \sigma^2/n$ and $E(\bar{X}) = \mu$

Unbiasedness of S^2 (cntd.)

Hence,

$$
\begin{aligned}
E(\tilde{S}^2) &= \frac{E(\sum_{i=1}^n X_i^2 - n\bar{X}^2)}{n} \\
&= \frac{\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2)}{n} \\
&= \frac{\sum_{i=1}^n (Var(X) + E(X)^2) - n(Var(\bar{X}) + E(\bar{X})^2)}{n} \\
&= \frac{n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2}{n} \\
&= \frac{(n-1)\sigma^2}{n}
\end{aligned}
$$

Unbiasedness of S^2 (cntd.)

This yields

$$E(\tilde{S}^2) = \frac{n-1}{n}\sigma^2.$$

So, $\tilde{S}^2$ is a biased estimator of σ^2 for finite n, but it is asymptotically unbiased, i.e., for $n \to +\infty$

$$E(\tilde{S}^2) \longrightarrow \sigma^2.$$

Nevertheless, it is easy to see that $S^2 = \tilde{S}^2 \frac{n}{n-1}$ is such that $E(S^2) = \sigma^2 \ \forall n$, so that S^2 is unbiased.

It could sound reasonable to use unbiasedness as a criterion to construct estimators. But this is not necessarily so great, as shown in the following example.

Example

Suppose $X \sim Poisson(\lambda)$ and for some reason (which escapes me for the moment), you want to estimate $\theta = P(X = 0)^2 = e^{-2\lambda}$.

Any unbiased estimator $\hat{\theta}$ must satisfy $E(\hat{\theta}) = \theta$, or

$$E(\hat{\theta}) = \sum_{x=0}^{\infty} \hat{\theta} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \hat{\theta} \frac{\lambda^x}{x!} = e^{-2\lambda}.$$

This implies that the following must hold $\sum_{x=0}^{\infty} \hat{\theta} \frac{\lambda^x}{x!} = e^{-\lambda}$.

The only function that can satisfy this equation is $\hat{\theta} = (-1)^x$. Thus, the only unbiased estimator is equal to 1 if X is even and -1 if it is odd, which is not sensible.

Mean squared error

We prefer estimators whose sampling distributions "cluster more closely" around the true value of θ, whatever that value might be.

MSE

The Mean Squared Error of an estimator of $\hat{\theta}$ is $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$.

It is

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2] + [E(\hat{\theta}) - \theta]^2 \\ &\qquad + 2[E(\hat{\theta}) - \theta]E[\hat{\theta} - E(\hat{\theta})] \\ &= Var(\hat{\theta}) + \text{bias}^2(\hat{\theta}) \end{aligned}$$

because $E[\hat{\theta} - E(\hat{\theta})] = 0$.

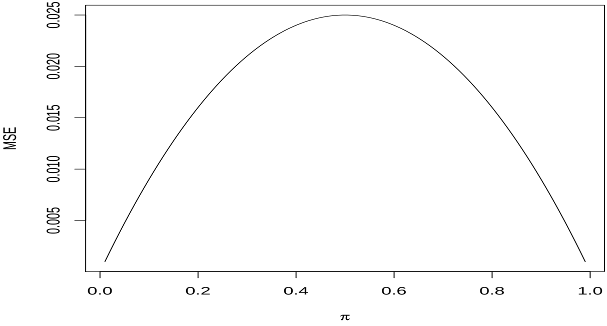
Example

Suppose $X_1, \dots, X_n$ are i.i.d. from $X \sim Bernoulli(\pi)$, and we want to estimate the mean π.

A natural estimator is $\hat{\pi} = \bar{X}$, which is unbiased since

$$E(\hat{\pi}) = E(\bar{X}) = \pi.$$

Furthermore $Var(\hat{\pi}) = Var(\bar{X}) = \pi(1-\pi)/n$ so that, for $n = 10$, $MSE(\hat{\pi})$ is



Example (cntd)

Consider an alternative estimator $\tilde{\pi} = \frac{\sum_{i=1}^{n} X_i + 1}{n+2}$.

It holds that $E(\tilde{\pi}) = E\left(\frac{\sum_{i=1}^{n} X_i + 1}{n+2}\right) = \frac{n\pi+1}{n+2}$. Therefore $\tilde{\pi}$ is biased, with

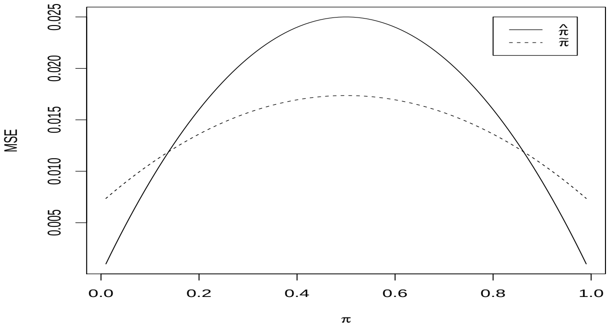
$$\text{bias}(\tilde{\pi}) = \frac{n\pi + 1}{n+2} - \pi.$$

Furthermore, $Var(\tilde{\pi}) = Var(\sum_{i=1}^{n} X_i)/(n+2)^2 = n\pi(1-\pi)/(n+2)^2$.

Hence,

$$\text{MSE}(\tilde{\pi}) = Var(\tilde{\pi}) + \text{bias}^2(\tilde{\pi}) = \frac{n\pi(1-\pi)}{(n+2)^2} + \left(\frac{n\pi+1}{n+2} - \pi\right)^2.$$

Example (cntd)



So the biased estimator has smaller MSE in much of the range of π and may be preferable if we do not think π is near 0 or 1. So our prior judgement about π might affect our choice of estimator. On the other hand, the plot is for $n = 10$ and the two curves become more and more similar as n increases.

First steps in statistical inference

– The likelihood –

A first example of likelihood function

- Assume $X \sim \textit{Poisson}(\lambda)$ with λ unknown
- we observe $n = 1$ samples from X and we obtain $x = 3$
- what does that say about λ?
- Different values of λ yield different values of $\Pr(X = 3;\ \lambda)$
- Consider the function $L(\lambda) = \Pr(X = 3; \lambda)$

$$L(1) = P(X = 3; \lambda = 1) = \frac{1^3 e^{-1}}{3!} \approx 0.06$$

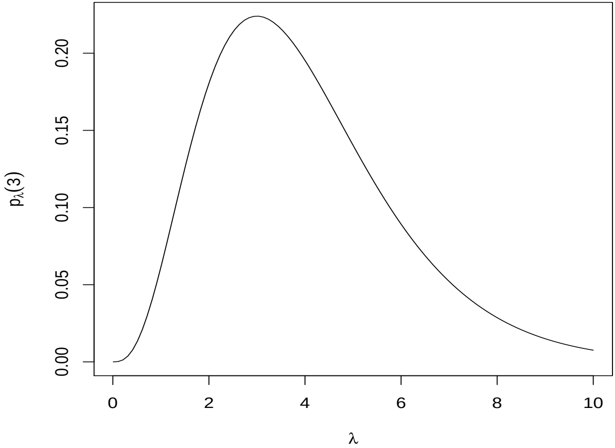
$$L(2) = P(X = 3; \lambda = 2) = \frac{2^3 e^{-2}}{3!} \approx 0.18$$

$$L(3) = P(X = 3; \lambda = 3) = \frac{3^3 e^{-3}}{3!} \approx 0.22$$

$$L(4) = P(X = 3; \lambda = 4) = \frac{4^3 e^{-4}}{3!} \approx 0.14$$

A first example of likelihood function (cntd)

Here, the value $\lambda = 3$ explains the data almost four times as well as the value $\lambda = 1$ and a bit better that the values $\lambda = 2$ and $\lambda = 4$.



Likelihood

The likelihood function is one of the most important and widely used tools for making inference. Let $X_1, \dots, X_n$ be random variables with (joint) pdf $p(x_1, \dots, x_n; \theta)$. We observe $x_1, \dots x_n$.

Definition

The likelihood of θ is $L(\theta) = p(x_1, \dots, x_n; \theta)$ as a function of θ.

If $X_1, \dots, X_n$ are i.i.d. with pdf $p(x_i; \theta)$, then

$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

.

Likelihood: interpretation

- The different values of the parameter represent different hypothesis about nature;
- a way to discriminate among them is according to how well they explain the data. This information is stored in the likelihood.
- that is, the likelihood function expresses the plausibility of different parameter values after we have observed the data;
- the parameter value θ_1 receives more support from the data than the parameter value θ_2 if $L(\theta_1) > L(\theta_2)$, or, alternatively

$$\frac{L(\theta_1)}{L(\theta_2)} > 1.$$

Maximum likelihood

Definition

The maximum likelihood estimate (MLE) of θ, $\hat{\theta}$, is the value that maximizes $L(\theta)$.

A strictly increasing transformation of $L(\theta)$ does not change the maximum (maxima) of the likelihood. So, it is often easier to maximize the *log-likelihood*.

$$\ell(\theta) = log\ L(\theta) = \sum_{i=1}^{n} log\ p_{X_i}(x_i; \theta).$$

MLE: computations in regular cases

Let $X_1, \dots, X_n$ be iid *Bernoulli*(π). Then

$$\begin{aligned} L(\pi) &= \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i} \\ &= \pi^{n\bar{x}}(1-\pi)^{n-n\bar{x}} \end{aligned}$$

$$\ell(\pi) = n\bar{x}\log\pi + n(1-\bar{x})\log(1-\pi)$$

$\ell(\pi)$ is a smooth function in π, so at the max-point, the derivative should be equal to zero (if not, the max-point is at one of the end-points).

$$\frac{d}{d\pi}\ell(\pi) = \frac{n\bar{x}}{\pi} - \frac{n(1-\bar{x})}{1-\pi}$$

Score function

The first derivative of $\ell(\theta)$ is given a special name.

Definition

The score function of θ is $\ell_*(\theta) = \frac{d}{d\theta}\ell(\theta)$.

Bernoulli example. $\ell_*(\pi)$ is zero for $\hat{\pi} = \bar{x}$. As $\frac{d^2}{d\pi^2}\ell(\pi)|_{\pi=\hat{\pi}} < 0$, $\hat{\pi} = \bar{x}$ is the maximum likelihood estimate.

The rv $\hat{\pi} = \bar{X}$ is therefore the maximum likelihood estimator.

We have $E(\hat{\pi}) = \pi$, so the estimator is unbiased.

More complex cases: muon decay

The decay of a muon into a positron, an electron neutrino, and a muon antineutrino has a distribution angle θ with density given by

$$p_X(x,\alpha) = \frac{1+\alpha x}{2} \qquad -1 \le x \le 1, \qquad -1 \le \alpha \le 1$$

with $x = \cos\theta$, where θ is the angle between the positron trajectory and the positive muon-spin and α the anisometry parameter.

It is

$$\begin{aligned} L(\alpha) &= \prod_{i=1}^{n} \frac{1+\alpha x_i}{2} \propto \prod_{i=1}^{n} (1+\alpha x_i) \\ \ell(\alpha) &= \sum_{i=1}^{n} \log(1+\alpha x_i) \\ \ell_*(\alpha) &= \sum_{i=1}^{n} \frac{x_i}{1+\alpha x_i} \end{aligned}$$

No closed expression is available for the solution of the score equation.

R code: muon decay likelihood

Data:

```
muon.data <-
c( 0.41040018,  0.91061564, -0.61106896,  0.39736684,  0.37997637, 0.34565436,
   0.01906680, -0.28765977, -0.33169289,  0.99989810, -0.35203164, 0.10360470,
   0.30573300,  0.75283842, -0.33736278, -0.91455101, -0.76222116, 0.27150040,
  -0.01257456,  0.68492778, -0.72343908,  0.45530570,  0.86249107, 0.52578673,
   0.14145264,  0.76645754, -0.65536275,  0.12497668,  0.74971197, 0.53839119)
```

Log-likelihood and score functions with their “vectorized” versions

```
# log-likelihood
lmuon <- function(alpha, data){
  l <- sum(log(1+alpha*data))-length(data)*log(2)
  return(l)
}

lmuon.v <- Vectorize(FUN=lmuon, vectorize.args = "alpha")

# score function
smuon <- function(alpha, data){
  l <- sum(data/(1+alpha*data))
  return(l)
}

smuon.v <- Vectorize(FUN=smuon, vectorize.args = "alpha")
```

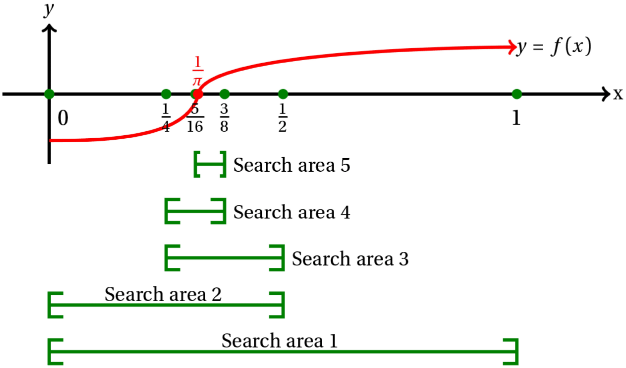
MLE: computations in more complex cases

- Optimization problems often occur in statistics. Analytic expressions for maximum likelihood estimators in complex models are usually not easily available, and numerical methods are needed.

- Different optimization methods could be considered to this aim. All methods commonly used are iterative. They start with some initial guess on the parameters. Next, a new candidate to the optimum, presumable better than the initial guess is found. This new guess is used to find an even better one, and so on until convergence is reached.

The bisection method

The bisection method is a root-finding method that applies to any continuous functions for which one knows two values with opposite signs. The method consists of repeatedly bisecting the interval defined by these values and then selecting the subinterval in which the function changes sign, and therefore must contain a root.



R code: bisection method for muon decay

```
bsec.root.muon <- function(lower, upper, tol=10e-7){
  while(upper-lower>tol){
    mid <- (upper+lower)/2
    if (sign(smuon(mid, data=muon.data))==sign(smuon(lower, data=muon.data)))
    {
      lower <- mid
    }else{
      upper <- mid
    }
  }
  return(mid)
}
```

```
bsec.root.muon(-1, 1)
[1] 0.4943933
```

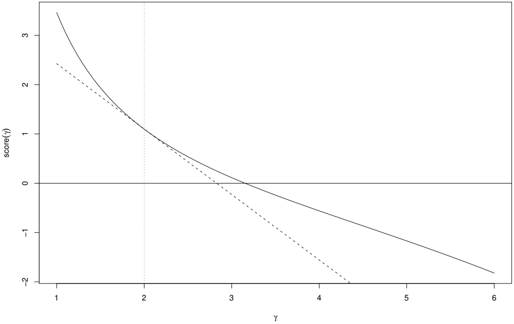
R code: functions as arguments and the three-dots construct

```
# generic function implementing the bisection method

bsec.root <- function(FUN, lower, upper, tol=10e-7, ...){
  while(upper-lower>tol){
    mid <- (upper+lower)/2
    if (sign(FUN(mid, ...))==sign(FUN(lower, ...))){
      lower <- mid
    }else{
      upper <- mid
    }
  }
  return(mid)
}

bsec.root(FUN=smuon, -1, 1, data=muon.data)
[1] 0.4943933
```

Newton Raphson

The Newton-Raphson method is one of the most used methods for optimization in statistics. It is it is based on the simple idea of linear approximation.

First step: $\gamma_1 = 2$



Muon decay: built in optimization and root finding

```
# The function uniroot searches the interval from
# lower to upper for a root (i.e., zero)

uniroot(smuon, data=muon.data, interval=c(-1,1))


minus.lmuon <- function(alpha, data) -lmuon(alpha, data)

# optimization functions that by default  perform minimization

nlminb(0.6, minus.lmuon, data=muon.data, lower=-1, upper=1)

optim (0.6, minus.lmuon, data=muon.data, lower=-1, upper=1)
```

Observed Fisher Information

The opposite of the second derivative of $\ell(\theta)$ is given a special name.

Definition

The observed Fisher information of θ is

$$j(\theta) = -\ell_{**}(\theta) = -\frac{d^2}{d\theta^2}\ell(\theta).$$

Observed Information: interpretation

In a neighborhood of $\hat{\theta}$, the log-likelihood can be approximated by a parabola where $j(\theta)$ is proportional to its curvature.

$$\ell(\theta) = \ell(\hat{\theta}) + (\theta - \hat{\theta})\ell_*(\hat{\theta}) + \frac{(\theta - \hat{\theta})^2}{2}\ell_{**}(\hat{\theta}) + \cdots$$

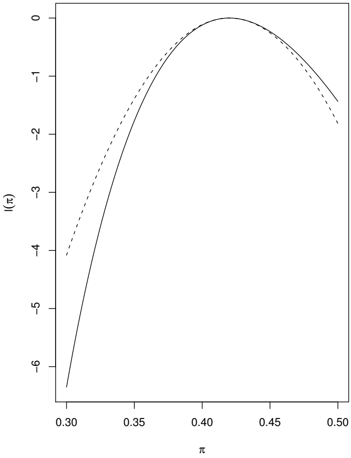
$$\ell(\hat{\theta}) - \ell(\theta) \doteq \frac{(\theta - \hat{\theta})^2}{2} j(\hat{\theta})$$

The larger $j(\hat{\theta})$, the stronger the curvature. In other words, values of θ close to $\hat{\theta}$ loose quickly the support from the sample.

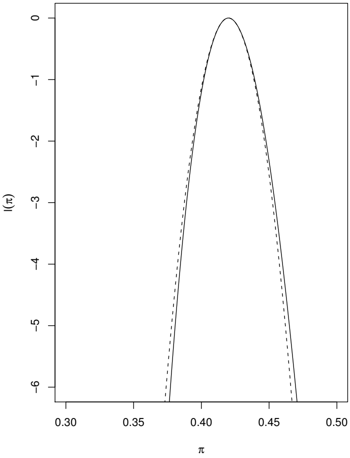
Observed Information: Example

$X_1, \dots, X_n$ from $X \sim Bernoulli(\pi)$

$n = 100, \quad \bar{x} = 0.42$



$n = 1000, \quad \bar{x} = 0.42$



Asymptotic distribution of MLEs

1. The asymptotic distribution of the maximum-likelihood estimator is established under the assumption that the log-likelihood function obeys certain regularity conditions.

2. It can be shown that, under regularity conditions, $\hat{\theta}$ is asymptotically normal with mean θ and "smallest attainable variance".

3. Such variance can be consistently estimated by the inverse of the observed information computed at $\hat{\theta}$, that is $j(\hat{\theta})^{-1}$,

$$\hat{\theta} \dot{\sim} N(\theta, j(\hat{\theta})^{-1})$$

.

Large sample properties of MLEs

Under regularity conditions, the MLE $\hat{\theta}$ satisfies the following properties:

1. is asymptotically normally distributed;
2. is asymptotically unbiased;
3. is asymptotically efficient;
4. if $\phi = h(\theta)$ where h is injective, then the mle of ϕ is $\hat{\phi} = h(\hat{\theta})$. This is called the equivariance property of MLE's.

Consequences

The result on the approximate sampling distribution of $\hat{\theta}$ is of considerable importance.

It gives the *se* of $\hat{\theta}$ and, therefore, the possibility of easily building tools for approximate inference in a variety of situations.

- θ one-dimensional

  CI of approximate level $(1-\alpha)$ $\quad \hat{\theta} \pm z_{1-\alpha/2}\ j(\hat{\theta})^{-1/2}$

  Test for $H_0 : \theta = \theta_0$ $\quad T = (\hat{\theta} - \theta_0)\ j(\hat{\theta})^{1/2} \dot{\sim} N(0,1)$

A non-regular case: How often does my bus come?

- The bus I take every day comes at fixed intervals of θ minutes.
- Assume that amount of time X that I must wait for a bus is uniformly distributed between zero and θ minutes (this assumption is not reasonable if I go to the bus stop always at the same time of the day). Formally, $X \sim U[0, \theta]$ with

$$f(x;\theta) = \frac{1}{\theta} I_{[0,\theta]}(x)$$

where $I_{[0,\theta]}(x)$ is the indicator function

$$I_{[0,\theta]}(x) = \begin{cases} 1 & \text{for } x \in [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

- What is the MLE of θ?

Non regular cases (cntd)

Let $X_1, \dots X_n$ i.i.d. $U[0, \theta]$. Here, the support of the distribution depends on the parameter θ, a condition which is not regular. Then

$$L(\theta) \quad = \quad \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[x_i,+\infty)}(\theta)$$

The product is either 0 or 1. It is 1 iff $0 \leq x_i \leq \theta$ for all i's, i.e. $0 \leq x_{(n)} \leq \theta$. Therefore

$$L(\theta) \quad = \quad \frac{1}{\theta^n} I_{[x_{(n)},+\infty)}(\theta)$$

For $\theta \geq x_{(n)}$, the likelihood is decreasing as θ increases, it is zero otherwise.

Therefore, $\hat{\theta} = x_{(n)}$. Is $\hat{\theta} = X_{(n)}$ an unbiased estimator?

Non regular cases (cntd)

We can try to answer the question by retrieving the distribution of $\hat{\theta} = X_{(n)}$. The cumulative distribution function is

$$F_{\hat{\theta}}(t) = P(X_{(n)} \leq t) = P\left(\cap_i (X_i \leq t)\right) = P(X_1 \leq t)^n = \left(\frac{t}{\theta}\right)^n$$

Differentiating, we have the pdf

$$p_{\hat{\theta}}(t;\theta) = \frac{nt^{n-1}}{\theta^n} \qquad 0 < t < \theta,$$

and

$$E(\hat{\theta}) = \int_0^\theta \frac{nt^n}{\theta^n} dt = \frac{n\theta}{n+1}$$

so $\hat{\theta}$ is biased but asymptotically unbiased.

Multiparameter case

When $\theta = (\theta_1, \ldots, \theta_p)^\top$ is multidimensional the observed Fisher information matrix is $J(\theta)$ has entries

$$[J(\theta)]_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta)$$

and under regularity conditions

$$\hat{\theta} \mathrel{\dot\sim} N_p(\theta;\ J(\hat{\theta})^{-1})$$

Consequences

- θ multi-dimensional: inference on j-th component

$$\hat{\theta}_j \dot{\sim} N(\theta_j, [J(\hat{\theta})^{-1}]_{jj})$$

CI of approximate level $(1-\alpha)$ $\quad \hat{\theta}_j \pm z_{1-\alpha/2}\sqrt{[J(\hat{\theta})^{-1}]_{jj}}$

Test for $H_0 : \theta_j = \theta_{j0}$ $\quad T = \frac{\hat{\theta}_j - \theta_{j0}}{\sqrt{[J(\hat{\theta})^{-1}]_{jj}}} \dot{\sim} N(0,1)$

Example: normal distribution

For the normal distribution the MLEs of μ and σ^2 are $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \tilde{S}^2$, respectively. The (exact) distribution of MLEs is as follows (recall that $E(\chi^2_\nu) = \nu$ and $Var(\chi^2_\nu) = 2\nu$):

- $\hat{\mu} \sim N(\mu, \sigma^2/n)$
- $\dfrac{n\hat{\sigma}^2}{\sigma^2} = \dfrac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$
- so that $E(\hat{\sigma}^2) = \dfrac{n-1}{n}\sigma^2$ and $Var(\hat{\sigma}^2) = \dfrac{2(n-1)}{n^2}\sigma^4$
- Furthermore $\hat{\mu}$ and $\hat{\sigma}^2$ are independent.

Example: normal distribution (cntd)

The asymptotic distribution of $(\hat{\mu}, \hat{\sigma}^2)$ is (bivariate) normal with expected value (μ, σ^2). Furthermore, an estimate of the variance matrix of such distribution is $j(\hat{\mu}, \hat{\sigma}^2)^{-1}$ that can be computed as follows.

Given that

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{n\tilde{s}^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}$$

then the observed information is the is the negative of the second derivative (the Hessian matrix) of $\ell(\mu, \sigma^2)$

$$J(\mu, \sigma^2) = -H(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{n(\bar{x}-\mu)}{2\sigma^4} \\ * & -\frac{n}{2\sigma^4} + \frac{n[\tilde{s}^2+(\bar{x}-\mu)^2]}{\sigma^6} \end{pmatrix}$$

Example: normal distribution (cntd)

The observed information matrix computed at $(\hat{\mu}, \hat{\sigma}^2)$ is

$$J(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix}$$

so that

$$J(\hat{\mu}, \hat{\sigma}^2)^{-1} = \begin{pmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{pmatrix}$$

Survival times for leukemia

A distribution commonly applied to survival data is the Weibull distribution which has pdf

$$p_X(x;\alpha,\beta) = \beta\alpha^{-1}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(\frac{x}{\alpha})^{\beta}},$$

where $\alpha > 0$ and $\beta > 0$. Note that if $\beta = 1$ then $X \sim \mathit{Exp}(1/\alpha)$.

The survival times observed on 16 patients (in months) are:

56 65 17 7 16 22 3 4 2 3 8 4 3 30 4 43

Want to estimate $\theta = (\alpha, \beta)$.

Survival times for leukemia (cntd)

$$\begin{aligned}
\ell(\alpha,\beta) &= n\log(\beta) - n\beta\log(\alpha) + (\beta-1)\sum_{i=1}^{n}\log(x_i) - \sum_{i=1}^{n}(x_i/\alpha)^{\beta} \\
\frac{d}{d\alpha}\ell(\alpha,\beta) &= -\frac{n\beta}{\alpha} + \frac{\beta}{\alpha}\sum_{i=1}^{n}(x_i/\alpha)^{\beta} \\
\frac{d}{d\beta}\ell(\alpha,\beta) &= \frac{n}{\beta} - n\log(\alpha) + \sum_{i=1}^{n}\log(x_i) - \sum_{i=1}^{n}(x_i/\alpha)^{\beta}\log(x_i/\alpha)
\end{aligned}$$

Maximizing the likelihood analytically is not possible. Need to work numerically.

Survival times for leukemia (cntd)

From the first equation we have

$$\hat{\alpha}(\beta) = \left(\frac{\sum_{i=1}^{n} x_i^{\beta}}{n} \right)^{1/\beta} .$$

Replacing $\hat{\alpha}(\beta)$ in $\frac{d}{d\beta}\ell(\alpha, \beta)$,

$$\frac{d}{d\beta}\ell(\alpha, \beta)|_{\hat{\alpha}(\beta)} \quad = \quad \frac{n}{\beta} + \sum_{i=1}^{n} \log(x_i) - n \frac{\sum_{i=1}^{n} x_i^{\beta} \log(x_i)}{\sum_{i=1}^{n} x_i^{\beta}}$$

we can solve the score equation with numerical methods.

Survival times for leukemia (cntd)

Once solved,

$$\hat{\alpha} = \left(\frac{\sum_{i=1}^{n} x_i^{\hat{\beta}}}{n} \right)^{1/\hat{\beta}} .$$

```
x <- c(56,65,17,7,16,22,3,4,2,3,8,4,3,30,4,43)
score <- function(beta, data)
{
    n <- length(data)
    n/beta +sum(log(data))-n*(sum(data^beta*log(data))/(sum(data^beta)))
}
hat.beta <- uniroot(score, data=x, interval=c(0.1,100))$root
hat.beta
hat.alpha <- (mean(x^hat.beta))^(1/hat.beta)
hat.alpha
```

Survival times for leukemia: R code

```
# data
x <- c(56,65,17,7,16,22,3,4,2,3,8,4,3,30,4,43)

# density of Weibull distribution
dW <- function(x, alpha, beta) beta*alpha^-1*(x/alpha)^(beta-1)*exp(-(x/alpha)^beta)

# we could also use the built-in function for the density of Weibull distribution
# shape = beta
# scale = alpha
# dweibull(x, shape, scale)

# minus-loglikelihood function
minus.loglik.W <- function(par, data){#par=c(alpha, beta)
  alpha <- par[1]
  beta  <- par[2]
  l <- sum(log(dW(data, alpha, beta)))
  return(-l)
}

# Maximum likelihood estimates of alpha and beta
opt.result <- nlminb(c(1,1), minus.loglik.W, data=x)
alpha.hat <- opt.result$par[1]
beta.hat  <- opt.result$par[2]
alpha.hat
[1] 17.20194
beta.hat
[1] 0.9218849
```

Survival times for leukemia: R code (cntd)

```
# observed information matrix evaluated at the maximum-likelihood estimate
library(numDeriv)
j <- hessian(minus.loglik.W, c(hat.alpha, hat.beta), data=x)
j

            [,1]      [,2]
[1,]  0.04595353 -0.432358
[2,] -0.43235801 36.303833

# asymptotic variance and covariance matrix
asy.var <- solve(j)
asy.var

           [,1]       [,2]
[1,] 24.5071665 0.29186642
[2,]  0.2918664 0.03102126

# standard error of the parameters
se.alpha <- sqrt(asy.var[1,1])
se.beta  <- sqrt(asy.var[2,2])

# 95% confidence intervals
hat.alpha+c(-1,1)*qnorm(0.975)*se.alpha

[1]  7.499203 26.904694

hat.beta +c(-1,1)*qnorm(0.975)*se.beta

[1] 0.5766801 1.2670913
```