

STAT 135, Concepts of Statistics

Helmut Pitters

Introduction

Department of Statistics
University of California, Berkeley

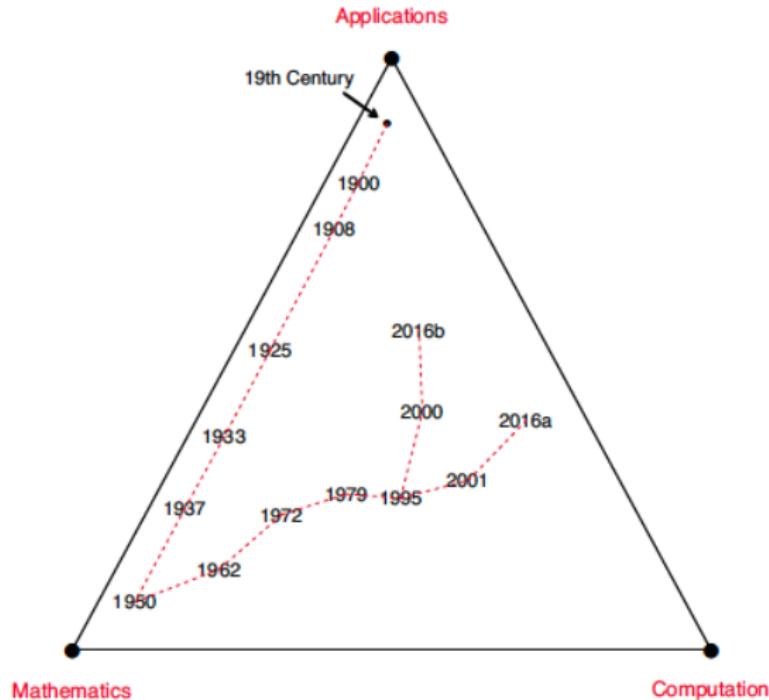
January 16, 2017

Introduction

Statistics is the science of learning from experience, particularly experience that arrives a little bit at a time: the successes and failures of a new experimental drug, the uncertain measurements of an asteroid's path toward Earth. . . .

Efron, Hastie 2016 - Computer Age Statistical Inference

Introduction



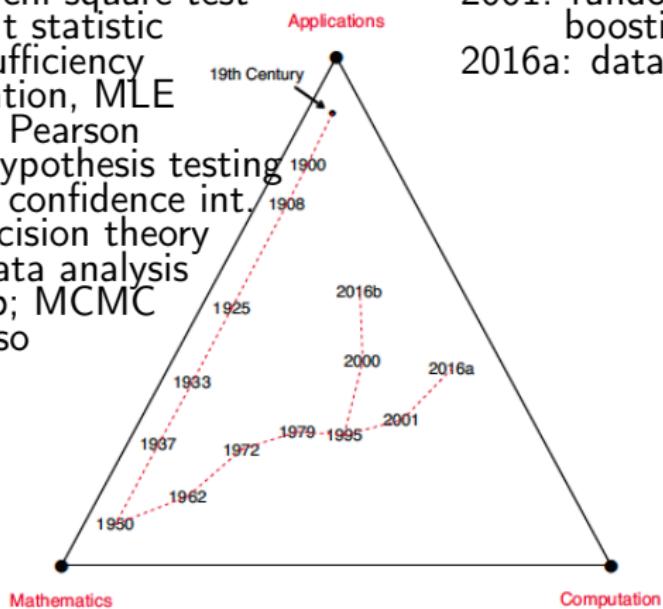
Development of the statistics discipline since the end of the nineteenth century, as discussed in the text.

Figure: From Efron, Hastie 2016

Introduction

- 1900: Pearson; chi square test
- 1908: Student; t statistic
- 1925: Fisher; sufficiency
F information, MLE
- 1933: Neyman, Pearson
optimal hypothesis testing
- 1937: Neyman; confidence int.
- 1950: Wald; decision theory
- 1962: Tukey; data analysis
- 1979: bootstrap; MCMC
- 1995: FDR; lasso

- 2001: random forests
boosting, neural nets
- 2016a: data science



Development of the statistics discipline since the end of the nineteenth century, as discussed in the text.

Figure: From Efron, Hastie 2016

Introduction and setup

We are interested in studying a (usually large) *population* of N individuals, e.g.

- ▶ California residents,
- ▶ patients worldwide diagnosed with some form of cancer,
- ▶ customers of a large mall,
- ▶ clients of life insurance company,
- ▶ facebook users, etc.

In particular, we are interested in certain characteristics

$$x_1, x_2, \dots, x_N$$

of the individuals, where

x_i = characteristic of i th individual,

e.g.

- ▶ age,
- ▶ expected lifetime,
- ▶ monthly income,
- ▶ number of friends/relationships.

Introduction and setup

Example (Speed of light)

Consider 60 of Michelson's measurements of the speed of light (observations=values listed+299000km/s).

Table: Velocity of light. Michelson, 1879.

850	960	880	890	890	740	940	880	810	850
840	900	960	880	810	780	1070	940	860	820
810	930	880	720	800	760	850	800	720	770
810	950	850	620	760	790	980	880	860	740
810	980	900	970	750	820	880	840	950	760
850	1000	830	880	910	870	980	790	910	870

[Example:histogram]

Introduction and setup

Remark

Notice: characteristics x_1, \dots, x_N are *not random*, but deterministic quantities.

However, usually one does not have access to all of the information

$$x_1, \dots, x_N$$

about the population, but only to a (randomly chosen) sample

$$x_{j_1}, x_{j_2}, \dots, x_{j_n} \subseteq \{x_1, \dots, x_N\}.$$

It is often convenient to summarize important features of the population in a few numbers, referred to as *population parameters*. We now turn to summaries that are often used in statistics.

Summarizing data.

Measures of location

Measures of location. Mean.

Arithmetic mean. The *arithmetic mean*

$$\bar{x} := \frac{1}{N} \sum_{i=1}^N x_i$$

or *average value* is probably the most common measure of location.
[Histogram: mean <> center of mass]

Measures of location. Mean.

Example (Speed of light)

Consider 60 of Michelson's measurements of the speed of light (observations=values listed+299000km/s), none of which is completely accurate due to the sensitivity of the measuring apparatus.

Table: Velocity of light. Michelson, 1879.

850	960	880	890	890	740	940	880	810	850
840	900	960	880	810	780	1070	940	860	820
810	930	880	720	800	760	850	800	720	770
810	950	850	620	760	790	980	880	860	740
810	980	900	970	750	820	880	840	950	760
850	1000	830	880	910	870	980	790	910	870

Expect $\bar{x} = 856.33$ to be a more accurate measure than each of the individual observations.

Measures of location. Mean.

Example (Speed of light)

Table: Velocity of light. Michelson, 1879.

850	960	880	890	890	740	940	880	810	850
840	900	960	880	810	780	1070	940	860	820
810	930	880	720	800	760	850	800	720	770
810	950	850	620	760	790	980	880	860	740
810	980	900	970	750	820	880	840	950	760
850	1000	830	880	910	870	980	790	910	870

Q: Consider the histogram of Michelson's observations. Roughly, what kind of shape do you expect to see? Why?

Measures of location. Mean.

A drawback of \bar{x} is its sensitivity to outliers, as the next example shows.

Example (Family incomes)

Incomes of five Berkeley families¹ are

\$90k \$70k \$77k \$85k \$300k.

Average

$$\bar{x} = 124.4k$$

of these incomes substantially influenced by single family with highest income. In fact, \bar{x} is greater than the income of any of the other four families.

A data value which is extreme w.r.t. the bulk of other values is called an *outlier*.

¹<http://www.city-data.com/income/income-Berkeley-California.html>

Measures of location. Median.

Median. Arrange the items in ascending order

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(N)}.$$

(In particular, $x_{(1)}$ is the smallest value, $x_{(N)}$ is the largest value.)

Definition

The *median* is defined to be the value of the middle item if N is odd, and the average of the values of the two middle items if N is even.

Notice that changing extreme values (e.g. smallest value $x_{(1)}$ or largest value $x_{(N)}$) does not affect the median. For this reason the median is said to be a *robust* measure of location.

Measures of location. Median.

Example (Family incomes)

The ordered family incomes are

\$70k \$77k \$85k \$90k \$300k

with median

\$85k.

This number seems to summarize the five family incomes more appropriately.

Measures of location. Trimmed mean.

For $\alpha \in [0, 1]$ the $100\alpha\%$ *trimmed mean*, denoted \bar{x}_α , is calculated by discarding the

$$\begin{cases} \text{lowest } 100\alpha\%, \text{ and the} \\ \text{highest } 100\alpha\% \end{cases} \quad (1)$$

observations, and computing the mean of the remaining observations.

Commonly the value of α is taken between 0.1 and 0.2.

Measures of location. Mode.

Definition

The *mode* of a set of data (if it exists) is the value that occurs with greatest frequency.

Notice that there may be two or more observations that occur with greatest frequency, in which case the data is said to be *bimodal*, respectively *multimodal*.

In the extreme case that all observed values are different,
e.g. family incomes

\$90k \$70k \$77k \$85k \$300k.

the mode is not defined.

Measures of location. Mode.

Example

Organization is to hold a national meeting. 80 members of organization (picked randomly) are asked to indicate their preference of city:

city	frequency
Miami	16
New Orleans	24
New York	12
San Francisco	28

Mode is San Francisco, the most preferred city.

Notice that in this example, and generally for qualitative data, the notion of a mean or median does not make any sense. The mode is therefore particularly suitable for qualitative data.

Measures of location.

There is no single “best” measure of location; instead, the choice of which measure(s) of location to use depends on the problem at hand and the purpose to which the measure is employed.

To summarize data, it is useful to compute different measures of location and to compare with graphical displays (e.g. histogram) of the data.

Summarizing data.

Measures of dispersion

Measures of dispersion. Percentile.

Measures of dispersion are used to quantify how much “spread-out” data are.

Do you recall (STAT 134) a quantity measuring the “spread” of a random variable X about its mean $\mathbb{E}[X]$?

Measures of dispersion. Percentile.

Percentile: statistical measure locating values in data set that are not necessarily central locations.

Provides information regarding how data are spread over an interval from lowest to highest value.

Definition (Percentile)

A p th percentile of a data set is a value such that *at least* $p\%$ of the items take on this value or less and *at least* $(100 - p)\%$ of the items taken on this value or more.

Example (Admission test scores)

Admission test scores of universities and colleges often reported in percentiles.

See e.g. <http://www.collegesimply.com/guides/1600-on-the-sat/>.

Measures of dispersion. Percentile.

Example (Family income)

\$70k \$77k \$85k \$90k \$300k.

Here a 20th percentile is \$73.5k, a 40th percentile is \$81k, etc.

Measures of dispersion. Percentile.

Calculating p th percentile of x_1, \dots, x_N .

1. Arrange data values in increasing order: $x_{(1)} \leq \dots \leq x_{(N)}$.
2. Compute index

$$i := \frac{p}{100}N.$$

3. The p th percentile is

$$\begin{cases} x_{(\lceil i \rceil)} & \text{if } i \text{ is not an integer} \\ \frac{x_{(i)} + x_{(i+1)}}{2} & \text{otherwise.} \end{cases}$$

Here $\lceil x \rceil$ denotes the smallest integer greater than or equal to x .

Measures of dispersion. Range.

The range is possibly the simplest measure of dispersion.

Definition (Range)

If the data x_1, \dots, x_N consist of real numbers, their range is defined to be the difference

$$x_{(N)} - x_{(1)}$$

between the maximal and the minimal value.

Notice however, that this measure is extremely sensitive to outliers.

Example (Family income)

\$90k \$70k \$77k \$85k \$300k.

The range here is $\$300k - \$70k = \$230k$.

Measures of dispersion. Variance.

The *variance*

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

is the most common measure of dispersion together with the *standard deviation s*.

Measures of dispersion. Sample variance.

Example (Family income)

\$70k \$77k \$85k \$90k \$300k.

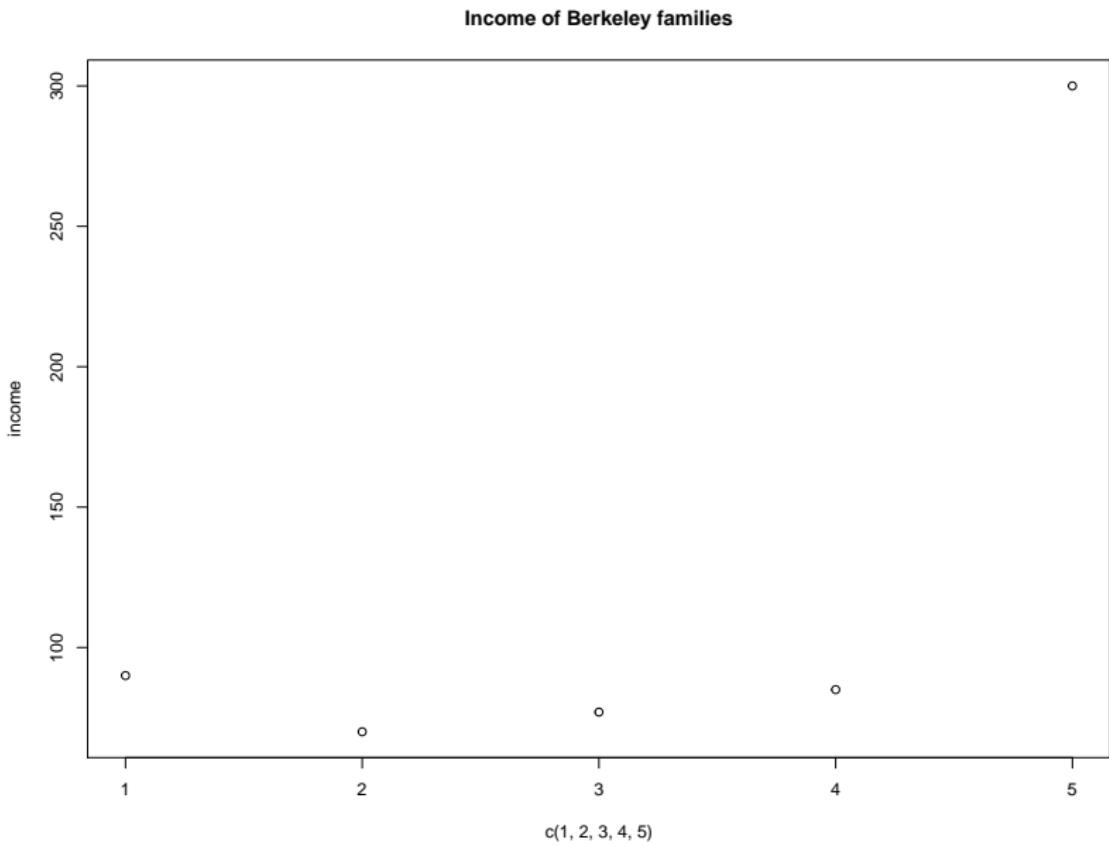
Here the standard deviation is

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} = \$98.46k.$$

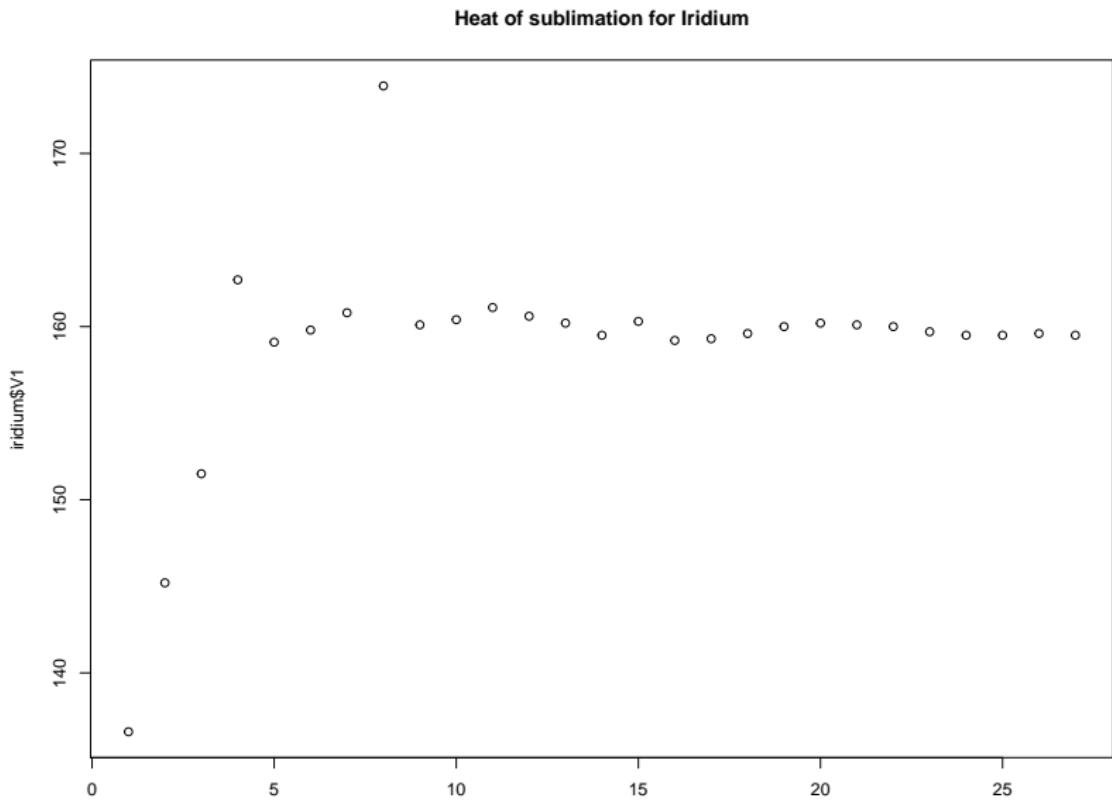
The standard deviation is rather sensitive to outliers, and we will encounter more robust measures of dispersion (e.g. interquartile range) later.

Graphical methods

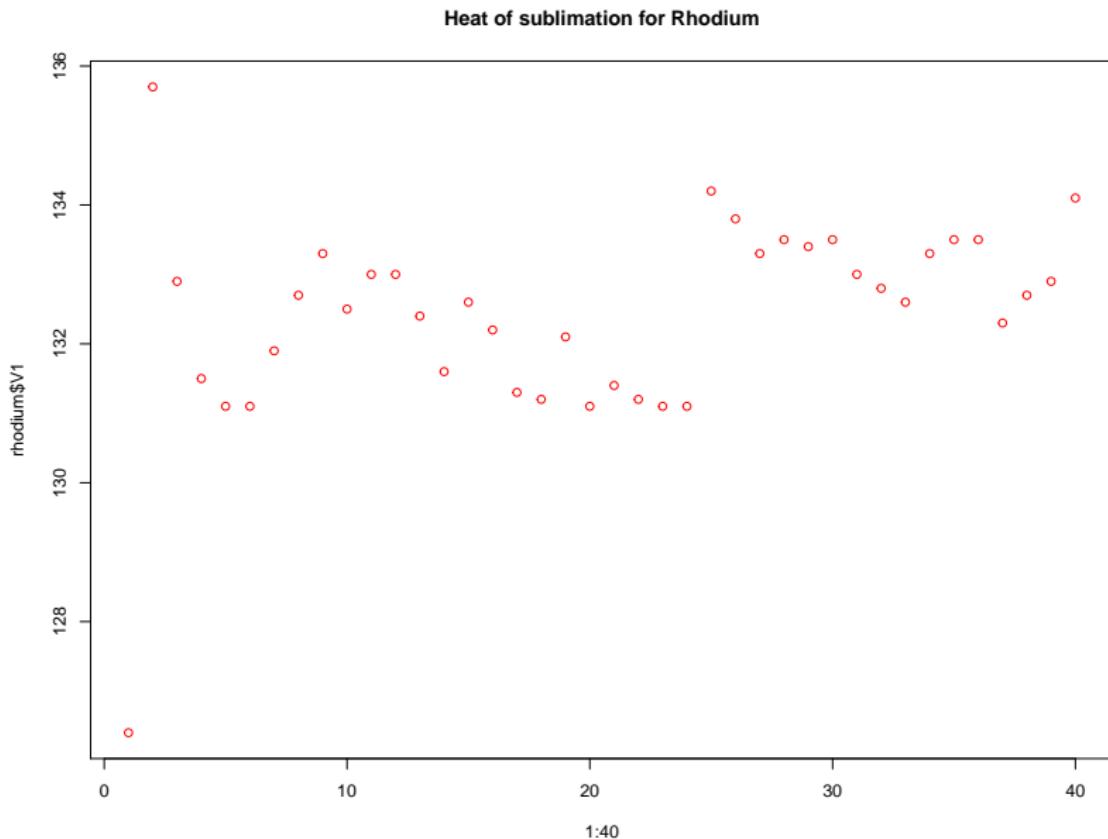
It can often be useful to plot the data x_1, \dots, x_n in sequential order.



We can immediately find the outliers in the data set on heat of sublimation of Iridium.



And we immediately see that the sublimation points for Rhodium are more dispersed.



STAT 135, Concepts of Statistics

Helmut Pitters

Simple random sampling

Department of Statistics
University of California, Berkeley

January 25, 2017

Simple random sampling

Simple random sampling

Simple random sampling: motivation

Usually, we do not know

N : population size

x_1, \dots, x_N : individual characteristics

Can sample some of the individuals and record their characteristics.
However, sampling all individuals is usually not feasible, e.g. it might

- ▶ be too costly, time-consuming (e.g. polling opinions of all US citizens)
- ▶ require to destroy products (e.g. canned food)
- ▶ be impossible, since we have no means to observe whole population (e.g. population of atlantic cod)



Simple random sampling



9-year old Antonio Martinez of San Lorenzo caught a 12 lb., 7 oz., 27" trout at Don Castro using power bait on 4/6/2008!!

Want to estimate average weight of trouts in Lake Don Castro.
How?

Catch a trout, record its weight w_1 , and release the fish again.
Maybe this fish was comparatively small/big.
Let's iterate this procedure, until we recorded the weights

$$w_1, w_2, \dots, w_n$$

of a “sufficiently large” number n of fishes.
With this knowledge, how do we estimate the weight of a trout?

Simple random sampling: setup

Mostly, one is interested in *population parameters*

$$\mu := \frac{1}{N} \sum_{i=1}^N x_i \quad \text{population mean}$$

$$\sigma^2 := \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2 \quad \text{population variance}$$

$$\tau := \sum_{i=1}^N x_i \quad \text{population total}$$

which are unknown.

Goal: Want to estimate (or “learn”) population parameters by studying a sample of n individuals drawn randomly from the population.

Simple random sampling: setup

Sample n individuals *randomly with replacement* and record their characteristics

$$X_1, X_2, \dots, X_n.$$

In other words: X_1 is chosen at random from x_1, \dots, x_N , and so is X_2, \dots, X_n , and the X_i s are independent.

Here, sampling introduces the randomness.

Later, we'll also be interested in *sampling without replacement* (referred to as simple random sampling).

Remark

Always consider carefully whether sampling is conducted with or without replacement.

Simple random sampling

Intuitively, we expect

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i \quad \text{sample mean}$$

to be a good estimator for population mean μ .

With some work we can show rigorously that this intuition is good!

Simple random sampling

Notation: suppose there are m different values in x_1, x_2, \dots, x_N ; denote them by

$$\zeta_1, \zeta_2, \dots, \zeta_m.$$

Let's say the value ζ_i appears n_i times in the population, i.e.

$$n_i := \#\{j : x_j = \zeta_i\}.$$
¹

By construction of X_1 ,

$$\mathbb{P}\{X_1 = \zeta_i\} = \frac{n_i}{n}.$$

We can now study the distribution of X_1 (and \bar{X}).

¹# A denotes the number of elements in the set A .

Simple random sampling

For the mean of X_1 we find

$$\mathbb{E}[X_1] = \sum_{i=1}^m \zeta_i \mathbb{P}\{X_1 = \zeta_i\} = \sum_{i=1}^m \zeta_i \frac{n_i}{N} = \frac{1}{N} \sum_{j=1}^N x_j = \mu.$$

It is now straightforward to compute the mean of \bar{X} :

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mu,$$

since X_1, \dots, X_n all have the same distribution.

This makes our intuition precise: if we were to study the population by drawing SRSs many times the average of the sample means would be μ .

E.g. think of a large number of scientists independently studying the same population by drawing SRSs.

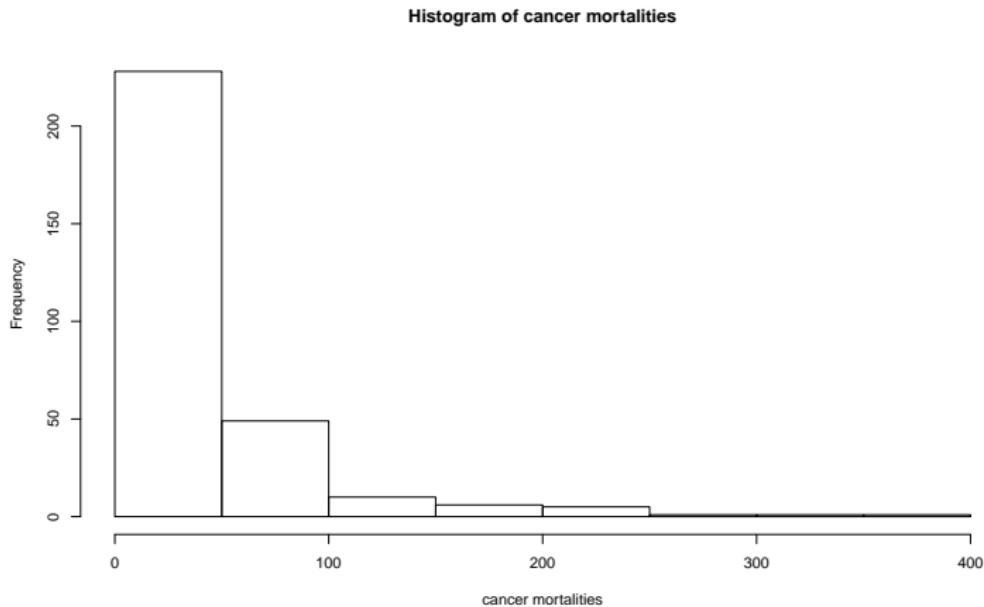
Simple random sampling

The results and calculations derived so far are still correct if our sampling is *without* replacement.

Why?

Simple random sampling

data: Values for breast cancer mortality 1950–1960 for 301 counties in North Carolina, South Carolina and Georgia.²



²You can find these data in `data/cancer.txt`, on bCourses.

Simple random sampling

data: Values for breast cancer mortality 1950–1960 for 301 counties in North Carolina, South Carolina and Georgia.³

$$\text{population mean } \mu = 39.86$$

The following are five means computed from five independent SRSs from the population, each of size 20:

$$64.7, 22.2, 29.3, 30.4, 42.2.$$

³You can find these data in `data/cancer.txt`, on bCourses.

Aside: Estimating population size

Usually, we do not even know the population size N .



9-year old Antonio Martinez of San Lorenzo caught a 12 lb., 7 oz., 27" trout at Don Castro using power bait on 4/6/2008!!

Lake Don Castro contains a population of trout, the number of which (call it N) is unknown. Come up with a simple method to estimate N .

You are allowed to catch (some) fish, and mark them with a color.

Simple random sampling

On average, sample mean agrees with population mean:

$$\mathbb{E}\bar{X} = \mu.$$

How accurate is \bar{X} as an estimator for μ ?

A reasonable measure for the accuracy of \bar{X} is its standard deviation or *standard error*

$$\sigma_{\bar{X}} := \sqrt{\text{Var}(\bar{X})}.$$

Since

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n} \text{Var}(X_1),$$

let's compute $\text{Var}(X_1)$.

Simple random sampling

We find

$$\begin{aligned}\text{Var}(X_1) &= \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \sum_{i=1}^m \zeta_i^2 \frac{n_i}{N} - \mu^2 \\ &= \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2 = \sigma^2,\end{aligned}$$

and therefore

$$\text{Var}(\bar{X}) = \frac{1}{n} \text{Var}(X_1) = \frac{\sigma^2}{n},$$

so \bar{X} has standard error

$$\boxed{\sigma_{\bar{X}} = \sqrt{\text{Var}(\bar{X})} = \frac{\sigma}{\sqrt{n}}}.$$

What does this formula tell us about how accurate we can estimate (“guess”) μ from \bar{X} ?

Review: Covariance.

Recall from STAT 134: *Covariance* of two random variables X and Y is defined by

$$\text{Cov}(X, Y) := \mathbb{E} [(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The covariance can be interpreted as a measure for joint variability or degree of linear association.

If $\text{Cov}(X, Y) = 0$, we call X and Y *uncorrelated*.

While independence of X, Y implies $\text{Cov}(X, Y) = 0$, the converse is not true.

Review: Covariance.

Recall from STAT 134 some useful formulas:

$$\text{Cov}(X, X) = \text{Var}(X)$$

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{Cov}(X, Y) = \text{Cov}(Y, X) \quad (\text{symmetry})$$

$$\text{Cov}(aX + b, Y) = a \text{Cov}(X, Y) \quad (\text{multilinearity})$$

Simple random sampling

What if instead we sample without replacement?

Now

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) + \frac{1}{n^2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \text{Cov}(X_i, X_j),\end{aligned}$$

and we need to find $\text{Cov}(X_i, X_j)$ for $i \neq j$.

Simple random sampling

Lemma

For $i \neq j$ we have

$$\text{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}. \quad (1)$$

Proof.

By definition of covariance and using $\mathbb{E}X_i = \mu$,

$$\begin{aligned}\text{Cov}(X_i, X_j) &= \mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j = \mathbb{E}X_iX_j - \mu^2 \\ \mathbb{E}X_iX_j &= \sum_{k,l=1}^m \zeta_k \zeta_l \mathbb{P}\{X_i = \zeta_k, X_j = \zeta_l\}.\end{aligned}$$

□

Proof.

$$\begin{aligned}\mathbb{P} \{X_i = \zeta_k, X_j = \zeta_l\} &= \mathbb{P} \{X_j = \zeta_l\} \mathbb{P} \{X_i = \zeta_k | X_j = \zeta_l\} \\ &= \begin{cases} \frac{n_l n_k}{N(N-1)} & k \neq l \\ \frac{n_l(n_k-1)}{N(N-1)} & k = l. \end{cases}\end{aligned}$$

Write $\frac{n_l(n_k-1)}{N(N-1)} = \frac{n_l n_k}{N(N-1)} - \frac{n_l}{N(N-1)}$ to obtain

$$\begin{aligned}\mathbb{E} X_i X_j &= \frac{1}{N(N-1)} \sum_{l=1}^m \zeta_l n_l \sum_{k=1}^m \zeta_k n_k - \frac{1}{N(N-1)} \sum_{l=1}^m \zeta_l^2 n_l \\ &= \frac{N}{N-1} \mu^2 - \frac{1}{N-1} (\sigma^2 + \mu^2) = \mu^2 - \frac{\sigma^2}{N-1}.\end{aligned}$$

The claim follows. □

Theorem

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma^2}{n} \frac{N-n}{N-1}. \quad (2)$$

Remark

- $1 - \frac{n-1}{N-1}$ is the *finite population correction*
- if *sampling fraction*

$$\frac{n}{N}$$

is small, the standard error is close to the one for sampling with replacement:

$$\sigma_{\bar{X}} \approx \frac{\sigma}{\sqrt{n}},$$

and hardly depends on N .

Proof.

Recall

$$\text{Var}(X_i) = \sigma^2, \quad \text{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}.$$

Hence

$$\begin{aligned}\text{Var}(\bar{X}) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) + \frac{1}{n^2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \text{Cov}(X_i, X_j) \\ &= \frac{\sigma^2}{n} - \frac{n(n-1)}{n^2} \frac{\sigma^2}{N-1} \\ &= \frac{\sigma^2}{n} \left(1 - \frac{n-1}{N-1}\right).\end{aligned}$$

□

Simple random sampling

We now turn to an example where we sample without replacement.

Example (Hospital discharges⁴)

For instructional purposes we study an example where we have access to the entire population—this will note be the case in real studies.

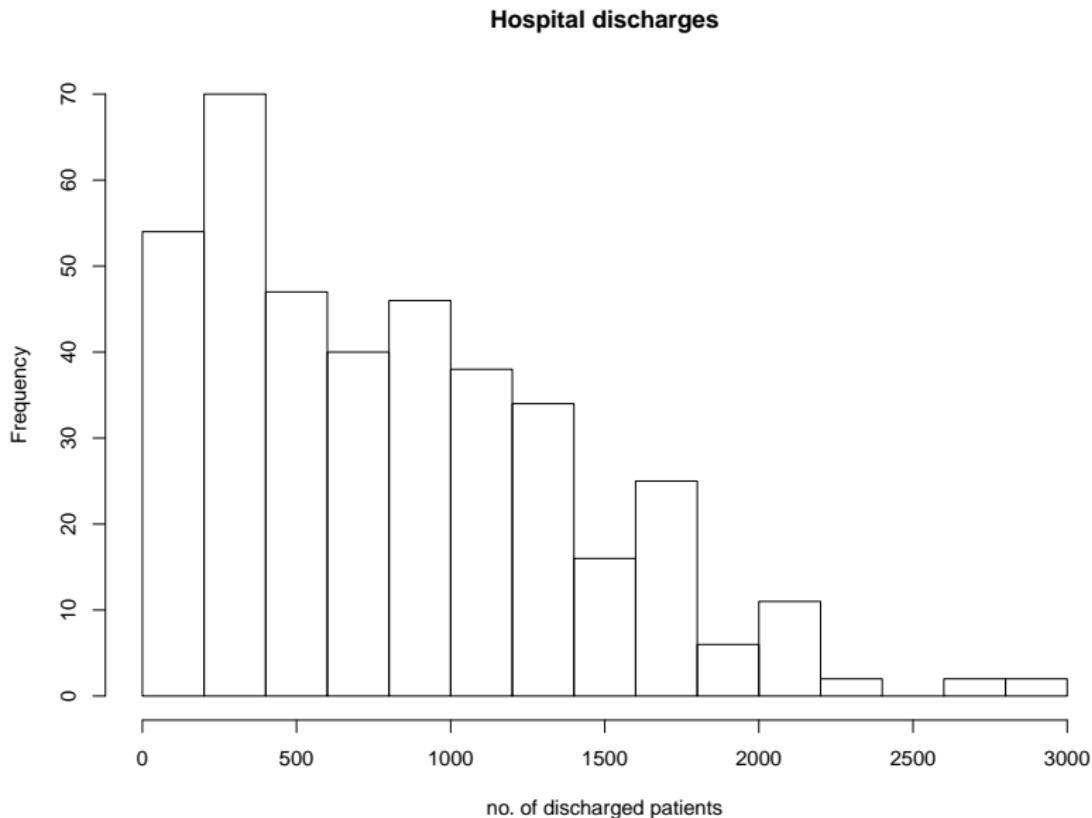
population: $N = 393$ short stay hospitals

$x_i :=$ # patients discharged from i th hospital during January 1968

$$\mu = 814.6$$

⁴Find the data in `data/hospitals.txt`.

Simple random sampling



Example: hospital discharges

Population parameters

$$N = 393, \quad \mu = 814.6, \quad \sigma = 589.7$$

Drawing sample of size $n = 50$ yielded $\bar{X} = 818.0$.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{589.7}{\sqrt{50}} \sqrt{\frac{343}{392}} = 83.4 \times 0.94 \approx 78.0.$$

If we were to take samples (of size $n = 50$) repeatedly and independently, most of the sample means would be contained in

$$(\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}}) \approx (662.0, 974.0).$$

(Cf. histogram of simulated sample means)

How likely is it that this interval contains what we are actually interested in: the population mean μ ?

Review. Binomial distribution.

Example (Lake Don Castro)

Suppose a proportion $p = 0.3$ of the fishes in Lake Don Castro are trouts. If we catch $n = 20$ fishes, releasing each fish after the catch (=sampling with replacement), the number

$$T = T(n, p)$$

of trouts in Lake Don Castro is random. It has the so-called *binomial distribution* with probability mass function

$$\mathbb{P}\{T = k\} = \binom{n}{k} p^k (1 - p)^{n-k} \quad (0 \leq k \leq n).$$

Histogram: <https://www.geogebra.org/m/CmHJuJxs>

Review. Normal distribution.

Definition (Normal density)

The map

$$\phi_{\mu,\sigma}(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (-\infty < x < \infty)$$

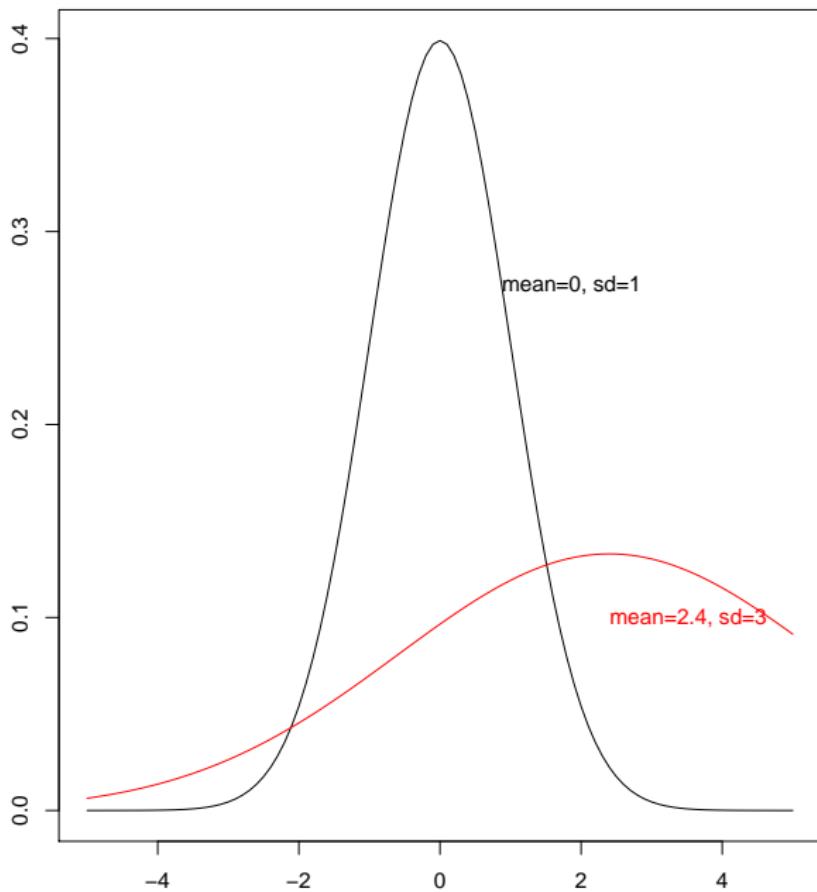
is the *density of the Normal distribution with mean μ and standard deviation $\sigma > 0$* .

For $\mu = 0$, $\sigma = 1$ we obtain the *density of the standard Normal distribution*⁵

$$\phi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (-\infty < x < \infty).$$

⁵We also refer to this density as the “normal curve.”

Normal densities



Review. Normal distribution.

Facts

1. *Total area under the curve $\phi_{\mu,\sigma}$*

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = 1.$$

2. $\phi_{\mu,\sigma}$ is symmetric about $x = \mu$, i.e.

$$\phi_{\mu,\sigma}(\mu + x) = \phi_{\mu,\sigma}(\mu - x).$$

3. *The points of inflection of $\phi_{\mu,\sigma}$ are*

$$\left(\mu \pm \sigma, \frac{1}{\sqrt{2\pi e \sigma}} \right).$$

Review. Normal distribution. Area under the curve.

Definition (Cumulative distribution function of Normal)

The area under the density $\phi(x)$ up to $x = z$ is denoted

$$\Phi(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{x^2}{2}} dx.$$

Φ is the so-called *cumulative distribution function of the standard Normal distribution*.⁶

In particular, we have $\Phi(-z) = 1 - \Phi(z)$ and $\Phi(0) = \frac{1}{2}$ from the symmetry of ϕ .

⁶There is no simple formula for the indefinite integral

$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{x^2}{2}} dx$. Values for Φ are tabulated, see Table 2 in Appendix B.

Review. Normal distribution: empirical rule.

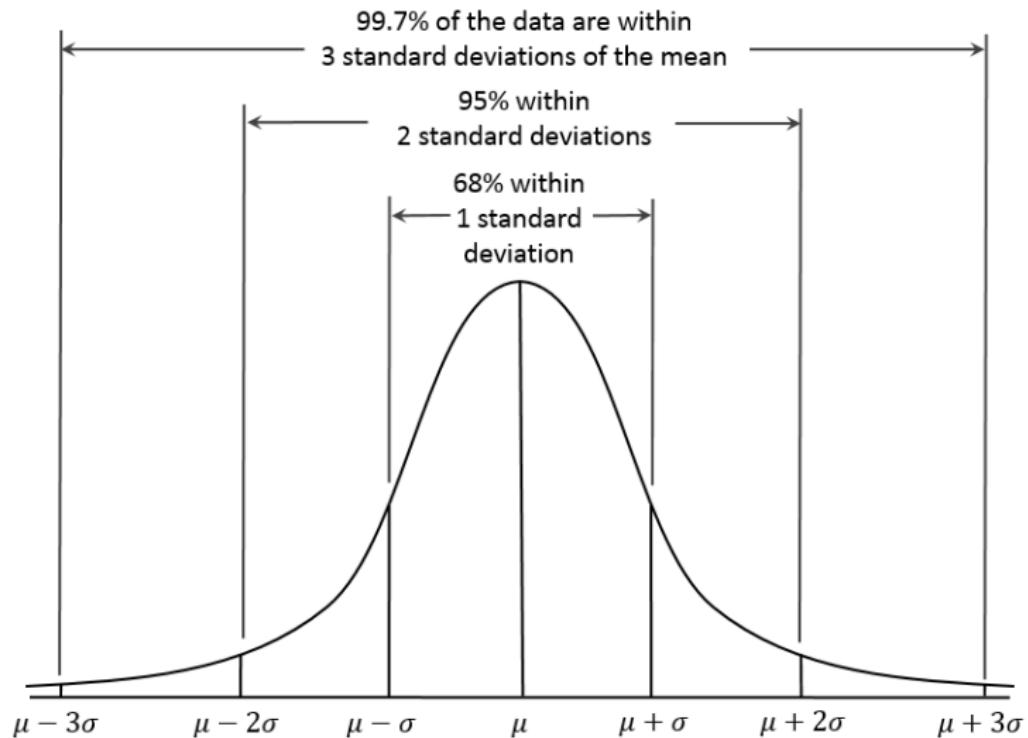


Figure: Empirical rule.

Review. Binomial distribution: Normal approximation.

For large n and p not too close to 0 or 1 the histogram of $\text{binomial}(n, p)$ can be well approximated by $\phi_{\mu, \sigma}$, the normal density with mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1 - p)}$.

[show approximation
on <https://www.geogebra.org/m/CmHJuJxs>]

Review. Binomial distribution: Normal approximation.

Fact (Normal approximation to binomial distribution)

For n independent trials with success parameter p

$$\mathbb{P}\{a \text{ to } b \text{ successes}\} \approx \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right),$$

provided n is large enough, where $\mu = np$ and $\sigma = \sqrt{np(1 - p)}$.

Review. Binomial distribution: Normal approximation.

Remark (Rule of thumb)

The normal approximation cannot always be accurate, of course. As a rule of thumb, the Normal approximation works better the larger σ is and the closer p is to $\frac{1}{2}$. A frequently used rule is that the approximation is reasonable, if

$$np > 5 \text{ and } n(1 - p) > 5.$$

Review. Binomial distribution: Fluctuation in number of successes.

Applying the Normal approximation and using the empirical rule, we have

$$\mathbb{P}\{\mu - \sigma \text{ to } \mu + \sigma \text{ successes in } n \text{ trials}\} \approx 68\%$$

$$\mathbb{P}\{\mu - 2\sigma \text{ to } \mu + 2\sigma \text{ successes in } n \text{ trials}\} \approx 95\%$$

$$\mathbb{P}\{\mu - 3\sigma \text{ to } \mu + 3\sigma \text{ successes in } n \text{ trials}\} \approx 99.7\%$$

$$\mathbb{P}\{\mu - 4\sigma \text{ to } \mu + 4\sigma \text{ successes in } n \text{ trials}\} \approx 99.99\%.$$

Typical size of fluctuation in the number of successes is

$$\sigma = \sqrt{np(1-p)}.$$

Typical size of fluctuation in the proportion of successes is

$$\frac{\sigma}{n} = \sqrt{\frac{p(1-p)}{n}}.$$

So: As n grows variability in #successes increases while variability in proportion of successes decreases.

Review. Binomial distribution: Fluctuation in number of successes.

Consider a large number n of independent trials with success probability p on each.

Fact (Square root law)

- ▶ *With high probability, the number of successes will lie in an interval centered at mean np with width a moderate multiple of \sqrt{n} .*
- ▶ *With high probability, proportion of successes will lie in an interval centered at p with width a moderate multiple of $1/\sqrt{n}$.*

In particular, as n increases, proportion of successes tends to p with high probability.

Review. Binomial distribution: Law of large numbers.

Consider n independent trials with success probability p on each.
Then for each $\varepsilon > 0$

$$\mathbb{P} \left\{ \left| \frac{\text{\#successes in } n \text{ trials}}{n} - p \right| \leq \varepsilon \right\} \rightarrow 1$$

as $n \rightarrow \infty$.

In words: as n increases, the proportion of successes in n independent trials will be very close to p with high probability.

Review. Square root law.

We are now more generally interested in distribution of

$$S_n := X_1 + X_2 + \cdots + X_n,$$

where the X_1, X_2, \dots are independent with some common distribution with $\mu = \mathbb{E}[X_1]$, $\sigma = \text{SD}(X_1)$.

$$\bar{X}_n := \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

We find

$$\mathbb{E}[S_n] = n\mu \quad \mathbb{E}[\bar{X}_n] = \mu$$

$$\text{Var}(S_n) = n\sigma^2 \quad \boxed{\text{SD}(S_n) = \sqrt{n}\sigma}$$

$$\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \quad \boxed{\text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}}$$

This is an immediate generalization of the square root law that we saw for the binomial distribution.

Review. Law of large numbers.

Let's take another careful look at

$$\bar{X} := \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

$$\mathbb{E}[\bar{X}_n] = \mu \quad \text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

As n grows large, the average

$$\bar{X}_n$$

of a sequence of independent draws

has mean μ , and its spread $\frac{\sigma}{\sqrt{n}}$ decreases.

In other words, \bar{X}_n is more and more concentrated around μ and is eventually constant.

Review. Law of large numbers.

Fact ((Weak) Law of large numbers)

Let X_1, X_2, \dots be independent r.v.s with common distribution and mean $\mu = \mathbb{E}[X_1]$. As n grows, the average

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

tends to μ with probability approaching 1.

Review. Normal approximation.

Fact (Normal approximation)

Let X_1, X_2, \dots be independent random variables with common distribution with mean $\mu := \mathbb{E}[X_1]$ and finite standard deviation $\sigma := \text{SD}(X_1) > 0$.

Then, for large n the distribution of the sum

$$S_n := X_1 + X_2 + \cdots + X_n$$

is approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$. Put differently, after standardizing S_n

$$\mathbb{P} \left\{ a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b \right\} \approx \Phi(b) - \Phi(a) \quad (a \leq b).$$

As $n \rightarrow \infty$ the error in this approximation tends to 0.

Review. Normal approximation. Simulations.

How can we statistically confirm the Normal approximation (also known as Central Limit Theorem (CLT))?

Idea: Draw large number of independent samples from quantity of interest,

$$S_n = X_1 + X_2 + \cdots + X_n,$$

and study their histogram!

1. Draw n ($=5000$) independent samples X_1, X_2, \dots, X_n from some distribution (here: geometric 0.1).
2. Compute $S_n^1 := X_1 + X_2 + \cdots + X_n$.
3. Repeat 1. and 2. r times (pick a large r , here: $r = 10000$) to obtain sums

$$S_n^1, S_n^2, \dots, S_n^r.$$

4. Draw the histogram of $S_n^1, S_n^2, \dots, S_n^r$.

[R script]

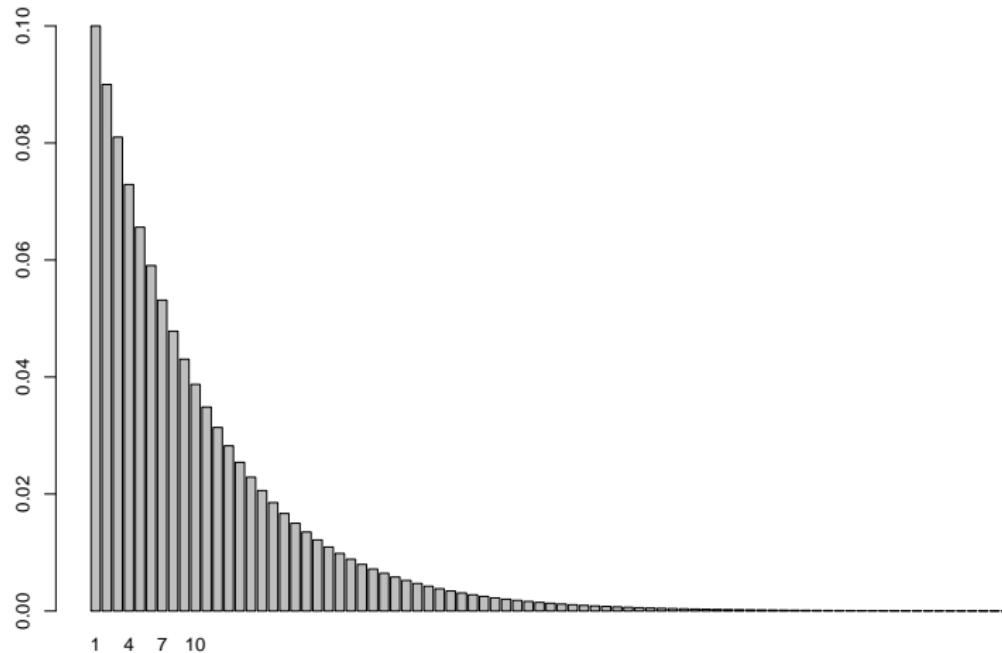


Figure: Probability mass function of geometric distribution.

Histogram of sums

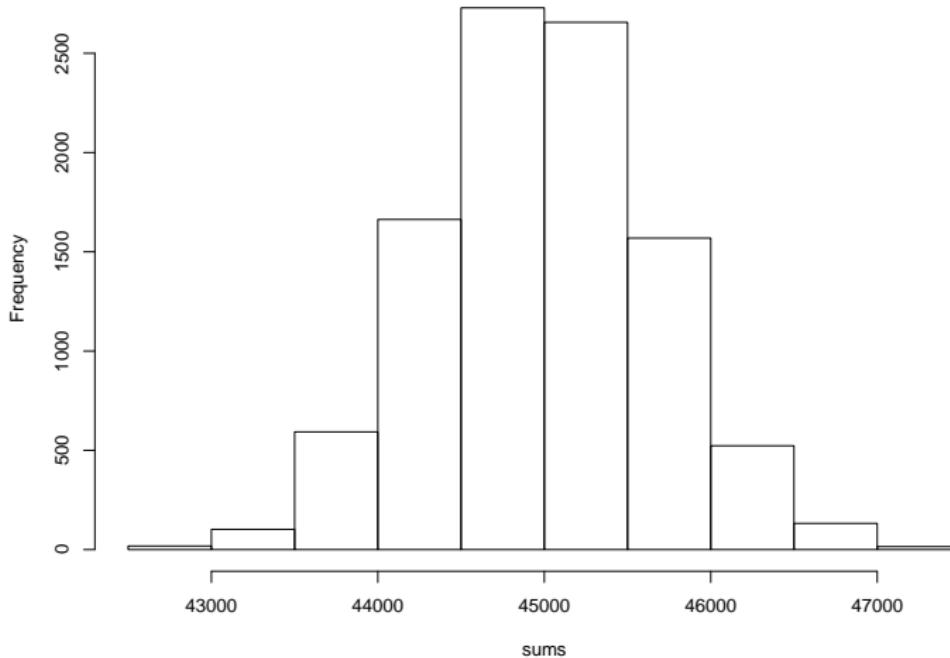


Figure: Histogram of sums $S_{5000}^1, \dots, S_{5000}^{10000}$.

[Sanity check?!]

Simple Random Sampling. Aside: drawing samples.

Example (Quality control)

Large manufacturer produces cars. Number N of cars produced in one day differs considerably from day to day and is not known beforehand. Each day a random sample of $n = 200$ cars is to be drawn (without replacement) to be tested for failures.

Cars arrive from the assembly line one by one. For each car test drivers have to decide immediately whether the car is shipped or tested. They can park n cars that are to be tested.

Study a nice algorithm that solves this problem in HW 7.7.27.

Example (Hospital discharges)

Would like to quantify how likely it is that the interval

$$(\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}}) \approx (662.0, 974.0)$$

contains μ .

To compute this probability, need distribution of \bar{X} .

In principle, we know the distribution of \bar{X} , as we can work it out in terms of the ζ_1, \dots, ζ_m . However, studying this distribution analytically is unfeasible.

[Chebyshev's inequality gives a crude upper bound on this probability.]

Recurring theme: as the setting seems to be too complicated to answer our question, let's make (reasonable) simplifying assumptions.

Example (Hospital discharges)

Simplifying assumption: suppose sample size n is large.

Idea: approximate \bar{X} by Normal distribution (provided n large enough).⁷

That is, approximately $\bar{X} \sim \mathcal{N}(\mu, \sigma_{\bar{X}}^2)$, so

$$\mathbb{P}\left\{a \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq b\right\} \approx \Phi(b) - \Phi(a).$$

Recall

$$\mathbb{E}\bar{X} = \mu = 814.6, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \approx 78.0.$$

⁷For sampling with replacement this approximation is justified by the CLT. If samples are drawn without replacement one has to work harder to show that for large n , but still small compared to N , this is still a good approximation.

Example (Hospital discharges)

Now

$$\mu \in (\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}}) \Leftrightarrow -2 \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq 2,$$

hence

$$\begin{aligned}\mathbb{P}\left\{\mu \in (\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}})\right\} &= \mathbb{P}\left\{-2 \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq 2\right\} \\ &= \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 = 0.954.\end{aligned}$$

Thus, the probability that μ differs from $\bar{X} = 818.0$ by more than $2\sigma_{\bar{X}} = 78.0$ is about 0.046 or 4.6%.

For this reason, $(\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}})$ is called a 95.4% *confidence interval* for the population mean μ .

Simple random sampling

Estimating the population variance

Simple random sampling

Saw that standard error

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}}$$

of sample mean \bar{X} (and other estimators) depends on

n and σ (and N).

However, population standard deviation σ (and N) is usually not known.

Idea: Estimate σ^2 from the data.

Natural candidate: sample variance

$$\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Theorem

$$\mathbb{E}\hat{\sigma}^2 = \sigma^2 \frac{n-1}{n} \frac{N}{N-1},$$

in particular, $\hat{\sigma}^2$ is a biased estimator for σ^2 .

Simple random sampling

Remark

For any population parameter that we might be interested in, call it θ , we could come up with an *estimator*

$$\hat{\Theta} = \hat{\Theta}(x_1, \dots, x_n)$$

for θ .

Morally speaking, $\hat{\Theta}$ is our “best guess” for θ , after seeing the data x_1, \dots, x_n .

Most of the time, we assume x_1, \dots, x_n to be samples from some r.v.s X_1, \dots, X_n . If

$$\mathbb{E}[\hat{\Theta}(X_1, \dots, X_n)] = \theta$$

then $\hat{\Theta}$ is called an *unbiased estimator* for θ .

Simple random sampling

Example

1. Since $\mathbb{E}[\bar{X}] = \mu$, \bar{X} is an unbiased estimator for μ .
2. Just saw that $\mathbb{E}[\hat{\sigma}] \neq \sigma$, so $\hat{\sigma}$ is a biased estimator for σ .

Simple random sampling

Consequently, $\hat{\sigma}^2 \frac{n}{n-1} \frac{N-1}{N}$ is an unbiased estimator for σ^2 .

Here however, we are interested in an unbiased estimator for $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \frac{N-n}{N-1}$.

Corollary

$$\begin{aligned}s_{\bar{X}}^2 &:= \frac{\hat{\sigma}^2}{n} \frac{N-n}{N-1} = \frac{\hat{\sigma}^2}{n} \frac{n}{n-1} \frac{N-1}{N} \frac{N-n}{N-1} \\ &= \frac{s^2}{n} \left(1 - \frac{n}{N}\right)\end{aligned}$$

is an unbiased estimator for $\sigma_{\bar{X}}^2$, where

$$s^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

The estimator $s_{\bar{X}}^2$ is called the estimated standard error.

Simple random sampling

Example (Hospital discharges)

Population parameters

$$N = 393, \quad \mu = 814.6, \quad \sigma = 589.7$$

Taking a sample of size $n = 50$ yields

$$\bar{X} = 815.92, \quad s \approx 565.85,$$

hence an estimated standard error of

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} \approx 74.76.$$

Recall that the true value for the standard error was

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \approx 78.0.$$

Confidence intervals.

Example: Opinion polls. Consider town of $N = 25,000$ eligible voters. Taking a simple random sample X_1, \dots, X_{1600} of size $n = 1,600 = 40^2$ we find that 917 support Democrats. Suppose we try to estimate percentage

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = p$$

of people who support Democrats, where

$$x_i = \begin{cases} 1 & \text{if } i\text{th person votes for Democrats} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \frac{2}{N} \mu \sum_{i=1}^N x_i + \mu^2 = p(1 - p).$$

Confidence intervals.

Example: Opinion polls, cont. Idea: use

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{917}{1600} \approx 0.57$$

as estimator for p . We know (simple random sampling):

$$\mathbb{E}\hat{p} = p, \quad \sigma_{\hat{p}} = \sqrt{\text{Var}(\hat{p})} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}} \approx \frac{\sigma}{40},$$

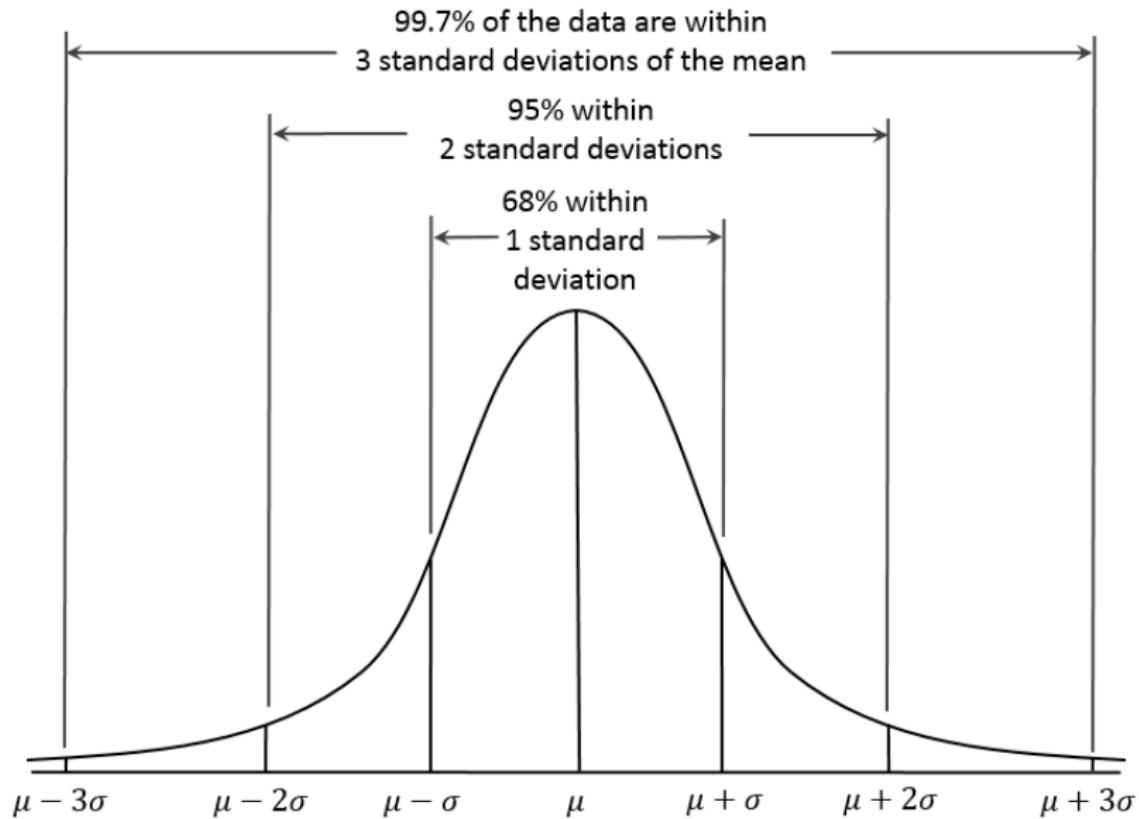
since sampling fraction $n/N = 1,600/25,000 = 0.064$ is small.
Moreover, since $\sigma = \sqrt{p(1-p)}$ not known, estimate it by

$$\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})} \approx 0.5,$$

i.e. estimate standard error $\sigma_{\hat{p}}$ by $0.0125 = 1.25\%$. Put differently,
the percentage $\hat{p} \approx 0.57$ of Democrats in the sample is likely to be
off the percentage of Democrats among all 25,000 eligible voters
by 1.25 percentage points or so.

Confidence intervals.

Recall the empirical rule for the standard Normal distribution.



Confidence intervals.

Example: Opinion polls, cont. Approximate \hat{p} ($= \bar{X}$) by $\mathcal{N}(\hat{p}, \hat{p}(1 - \hat{p})/n) = \mathcal{N}(0.57, 0.25/1600)$ (due to CLT). Hence

Table: Confidence intervals

$\hat{p} \pm 1.25\% = [0.55, 0.58]$	68.3%	confidence interval for p
$\hat{p} \pm 2 \times 1.25\% = [0.54, 0.6]$	95.5%	"
$\hat{p} \pm 3 \times 1.25\% = [0.53, 0.61]$	99.7%	"

STAT 135, Concepts of Statistics

Helmut Pitters

Parameter estimation

Department of Statistics
University of California, Berkeley

February 20, 2017

Review: Hypergeometric distribution

A population contains G good and $N - G$ bad elements.
Randomly sample $n \leq N$ elements without replacement.

$S_n :=$ number of good elements in sample

follows a hypergeometric distribution, i.e.

$$\mathbb{P}\{S_n = g\} = \frac{\binom{G}{g} \binom{N-G}{n-g}}{\binom{N}{n}}.$$

Recall:

$$\mathbb{E}S_n = np, \quad \text{Var}(S_n) = npq \frac{N-n}{N-1},$$

where $p = G/N$, $q = (N - G)/N$.

Parameter estimation

Example (Capture-recapture)

Estimating population size N .



9-year old Antonio Martinez
of San Lorenzo caught a 12
lb., 7 oz., 27" trout at Don
Castro using power bait on
4/6/2008!!

Parameter estimation.

Example (Capture-recapture)

Catch $r = 1000$ fish, mark them red, and release them.

Later, new catch of $n = 1000$ fish is made, among which $k = 100$ are found to have red marks. What can be said about the total (unknown) number N of fish in the lake?

Heuristics:

proportion of red fish in sample \approx proportion of red fish in lake,

i.e.

$$\frac{k}{n} \approx \frac{r}{N}.$$

Consequently, we expect

$$\hat{N} := \frac{n}{k}r = \frac{r}{\frac{k}{n}}$$

to be a good estimator for N .

Parameter estimation.

Example (Capture-recapture)

Clearly, $N \geq r + (n - k)$.

Before the second catch is made, the distribution

$R(n) :=$ the number of red fish in the sample

follows a hypergeometric law, i.e.

$$\mathbb{P}\{R(n) = k\} = \text{hypergeometric}(N, r, n)(k) = \frac{\binom{r}{k} \binom{N-r}{n-k}}{\binom{N}{n}}.$$

Parameter estimation.

Example (Capture-recapture)

We don't expect $\hat{N} = r + n - k = 1900$ to be a good guess for N . In fact, if \hat{N} were the actual number of fish, the outcome of our experiment would be rather unlikely, namely it would have probability

$$\begin{aligned}\text{hypergeometric}(\hat{N}, r, n)(k) &= \binom{1000}{100} \binom{900}{900} / \binom{1900}{1000} \\ &= \frac{(1000!)^2}{100!1900!},\end{aligned}$$

which has order of magnitude 10^{-430} according to Stirling's formula, $n! \sim \sqrt{2\pi n}(n/e)^n$.

Question: Which number \hat{N} should we pick as estimate for N in order to maximize likelihood of our observation?

(Notion of *maximum likelihood estimate* goes back to R. A. Fisher.)

Parameter estimation.

Example (Capture-recapture)

Let $p_N(k) := \text{hypergeometric}(N, r, n)(k)$. Consider the ratio

$$\begin{aligned}\frac{p_N(k)}{p_{N-1}(k)} &= \frac{\binom{r}{k} \binom{N-r}{n-k}}{\binom{N}{n} \binom{r}{k} \binom{N-1-r}{n-k}} \\ &= \frac{n!(N-n)!}{N!} \frac{(N-r)!}{(n-k)!(N-r-n+k)!} \\ &\quad \times \frac{(N-1)!}{n!(N-1-n)!} \frac{(n-k)!(N-1-r-n+k)!}{(N-1-r)!} \\ &= \frac{(N-r)(N-n)}{N(N-r-n+k)}.\end{aligned}$$

This yields $p_N(k)/p_{N-1}(k) < 1$ if $nr < Nk$, and
 $p_N(k)/p_{N-1}(k) > 1$ otherwise.

Thus likelihood is maximized for N the integer closest to nr/k
(= 10000).

Next: Confidence interval around \hat{N} via normal approximation

Parameter estimation

Example (Emission of alpha particles)

Radioactive material of certain mass is monitored with a Geiger counter.

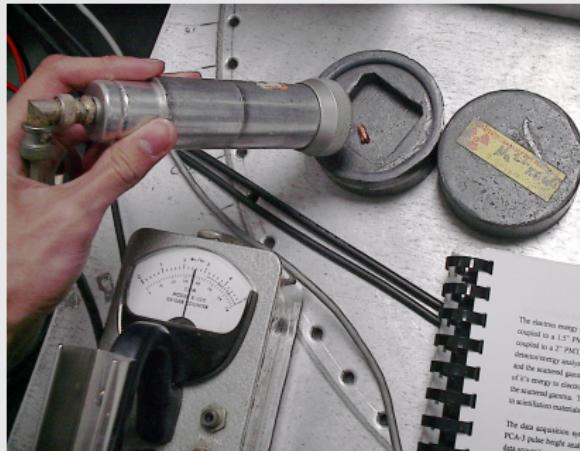


Figure: Geiger counter

Parameter estimation.

Example (Emission of alpha particles)

Assume

- ▶ rate of emission is constant over period of observation,
- ▶ particles come from large number of independent sources.

These assumptions justify model

$$N_t := \#\text{emissions during time } [0, t] \sim \text{Poisson}(\alpha t),$$

for some parameter $\alpha > 0$.

Recall that $X \sim \text{Poisson}(\alpha)$, i.e. X follows a Poisson distribution with parameter α , if

$$\mathbb{P}\{X = k\} = e^{-\alpha} \frac{\alpha^k}{k!}$$

for any non-negative integer k .

Review: Poisson distribution

$X \sim \text{Poisson}(\alpha)$, i.e. X follows a Poisson distribution with parameter α , if

$$\mathbb{P}\{X = k\} = e^{-\alpha} \frac{\alpha^k}{k!}$$

for any non-negative integer k .

$$\mathbb{E}[X] = \alpha, \quad \text{Var}(X) = \alpha.$$

Parameter estimation.

Example (Emission of alpha particles)

Data from National Bureau of Standards.

Source of alpha particles: americium 241.

10,220 times between successive emissions were recorded. Total time was subdivided into 1207 intervals of 10sec. each.

$$\text{mean emission rate} = \frac{\text{\#emissions}}{\text{total time of observation}[s]} = \frac{10220}{12070s} \approx 0.839/s$$

Hence, on average 8.39 emissions are observed in an interval.

Parameter estimation.

Example (Emission of alpha particles)

Let

$$E_i := \#\text{emissions in } i\text{th interval} \sim \text{Poisson}(10\alpha),$$

in particular $\mathbb{E}E_i = 10\alpha$. Data suggests that

$$\hat{\alpha} \approx 0.839/s$$

should be a good estimator for α .

Parameter estimation.

Example (Emission of alpha particles)

Next table shows summary of this data.

first column = number of counted emissions

second column = number of intervals with corresponding emission counts

Parameter estimation.

Example (Emission of alpha particles)

emission counts	number of intervals
0–2	18
3	28
4	56
5	105
6	126
7	146
8	164
9	161
10	123
11	101
12	74
13	53
14	23
15	15
16	9

Parameter estimation.

Example (Emission of alpha particles)

How could we compare the model with estimated parameter $\hat{\alpha} \approx 0.839/s$ to data? Notice that probability to have k emissions in interval i is

$$p(k) := \mathbb{P}\{E_i = k\} = e^{-8.39} \frac{(8.39)^k}{k!}$$

for any i . Thus number of intervals during which k emissions were counted is (E_1, E_2, \dots are i.i.d.)

$$B_k := \sum_{i=1}^{1207} \mathbf{1}\{E_i = k\} \sim \text{binomial}(1207, p(k)),$$

expected number of intervals counting k emissions is

$$\mathbb{E}B_k = 1207p(k).$$

Parameter estimation

Example (Emission of alpha particles)

emission counts	number of intervals	expected no. of intervals
0–2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	159.1
8	164	166.9
9	161	155.6
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9

From examining above table, our model seems to agree quite well with the data.

Ideally we'd like to quantify precisely "how well" the model fits.
There might be a slightly different model that fits better?!

We will see measures for the "goodness of fit" of a model later on.

Parameter estimation. Setup

Let us think about previous examples from more abstract point of view. Have observations

$$x_1, x_2, \dots, x_n$$

(e.g. number of emissions in certain time interval, whether or not caught fish is red, political opinion of voter, etc.)

which we regard as observed values of some random variables

$$X_1, X_2, \dots, X_n.$$

Parameter estimation. Setup

In general, X_1, \dots, X_n are not simple random draws from finite population. They could be

- ▶ i.i.d. $\sim \mathcal{N}(\mu, \sigma)$,
- ▶ i.i.d. $\sim \text{Poisson}(\lambda)$,
- ▶ i.i.d. $\sim \text{Gamma}(\alpha, \lambda)$,
- ▶ generated by simple random sampling from specific population of N individuals with characteristics x_1, \dots, x_N ,
- ▶ generated by sampling with replacement,
- ▶ etc.

We will focus on models where common distribution \mathbb{P}_θ of X_i depends on some parameter, generically called $\theta > 0$.
(These are so-called *parametric models*.)

Parameter estimation. Setup

Having observed data $x = (x_1, \dots, x_n)$, what can we infer about θ ?
Would like to make statements about which values of θ are plausible, based on x .

Assume that

- ▶ distribution of X is known up to some parameter θ
(e.g. distribution of X could be exponential with parameter θ).
- ▶ we have observed data x that we use to
 - ▶ construct a point estimate $\hat{\theta}$ of the value of θ ,
 - ▶ construct a confidence interval of (plausible) values for θ ,
 - ▶ test a hypothesis about θ .

Instead of “guessing” an estimator for θ as in the previous examples, we would like to have a more principled approach.

Remark

Notice that in general θ may not be a single real number, but could be an element of a more abstract set Θ (*=parameter space*).

E.g. $\mathbb{P}_{\mu, \sigma^2} \sim \mathcal{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma)$.

Parameter estimation. Method of moments

Random variable X has k -th moment (provided it exists)

$$\mu_k := \mathbb{E}[X^k].$$

Our goal will be to express θ in terms of the moments of X_1 of lowest possible order.

We will then use the k -th sample moment

$$\hat{\mu}_k := \frac{1}{n} \sum_{i=1}^n X_i^k$$

as an estimator for μ_k , and thereby find an estimator for θ .

Once we have an estimator for θ , we will be interested in its accuracy (standard error), and, more generally, in studying its distribution, the so-called *sampling distribution*.

Parameter estimation. Method of moments

Example (Capture-recapture)

$$X_i := \begin{cases} 1 & i\text{th fish in sample has red mark} \\ 0 & \text{otherwise.} \end{cases}$$

X_1, \dots, X_n is SRS from population of fishes

We are interested in parameter $\theta = N$. From

$$\mu_1 = \mathbb{E}[X_1] = \frac{r}{N},$$

we find $N = r/\mu_1$. Using $\hat{\mu}_1 = \bar{X} = k/n$ as estimator for μ , we find

$$\hat{N} = \frac{r}{\bar{X}} = r \frac{n}{k}$$

as estimator for N .

This is the so-called *method of moments estimator* for N .

Parameter estimation. Method of moments

Example

$$X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

Interested in parameter $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$.

$$\mu_1 = \mu, \quad \mu_2 = \text{Var}(X_1) + \mathbb{E}[X_1]^2 = \sigma^2 + \mu^2,$$

thus

$$\theta_1 = \mu_1, \quad \theta_2 = \mu_2 - \mu^2$$

and substituting sample moments, we obtain the estimators

$$\hat{\theta}_1 = \hat{\mu}_1 = \bar{X}, \quad \hat{\theta}_2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

We know: $\bar{X} \sim N(\mu, \sigma^2/n)$.

We will now see that $\hat{\theta}_1, \hat{\theta}_2$ are independent, and $n\hat{\theta}_2/\sigma^2 \sim \chi_{n-1}^2$.

Parameter estimation. Method of moments

Example

In order to study sampling distribution of estimator

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

need some results on distributions derived from the normal.

Review: Gamma distribution.

G has $\text{Gamma}(\alpha, \lambda)$ distribution, if its probability density is given by

$$f_G(t) = \begin{cases} \frac{(\lambda t)^{\alpha-1}}{\Gamma(\alpha)} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

and moments

$$\mathbb{E}[G^k] = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\lambda^k}.$$

For integer α can interpret G as waiting time until we see first event in Poisson Process of intensity λ .

If $G_1 \sim \text{Gamma}(\alpha_1, \lambda)$, $G_2 \sim \text{Gamma}(\alpha_2, \lambda)$ are independent, then

$$G_1 + G_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, \lambda).$$

Review: chi squared distribution

Recall:

X_1, X_2, \dots, X_n i.i.d. standard normals.

$$X_1^2 + X_2^2 + \cdots + X_n^2 \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right) = \chi_n^2$$

chi squared distribution with n degrees of freedom (df).

Aside: distributions derived from normal

$$X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$

An important result states that

\bar{X} and $(X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X})$ are independent.¹

This immediately implies that sample mean \bar{X} and sample variance

$$S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

are independent.

¹We will not prove this result in STAT 135.

Moment generating function

The *moment generating function (MGF)* $M_X(t)$ of r.v. X is defined as

$$M_X(t) := \mathbb{E}[e^{tX}],$$

if this mean exists.

Fact

If the MGF of X exists in an open interval containing 0, it uniquely determines the probability distribution of X .

Example (MGF Gamma distribution)

$G \sim \text{Gamma}(\alpha, \lambda)$.

$$\begin{aligned} M_G(t) &= \mathbb{E}[e^{tG}] = \int_0^\infty e^{ts} f_G(s) ds = \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)} \int_0^\infty e^{ts} s^{\alpha-1} \lambda e^{-\lambda s} ds \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty s^{\alpha-1} e^{-(\lambda-t)s} ds = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\lambda-t)^\alpha} = \left(\frac{\lambda}{\lambda-t}\right)^\alpha. \end{aligned}$$

Moment generating function

Fact

Suppose r.v.s X, Y are independent and their MGFs exist on open interval containing 0. Then

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

on the common interval where both M_X and M_Y exist.

The proof is straightforward.

Aside: distributions derived from normal

We now have the tools to derive an important result about the distribution of the sample variance S^2 .

Theorem

For $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$,

$$(n - 1)S^2 / \sigma^2 \sim \chi_{n-1}^2.$$

Remark

From this theorem we get the previous claim for the distribution of the rescaled sample

$$n\hat{\theta}_2^2 / \sigma^2 \sim \chi_{n-1}^2,$$

where

$$n\hat{\theta}_2^2 / \sigma^2 = n\hat{\sigma}^2 / \sigma^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Aside: distributions derived from normal

Proof.

Let us replace \bar{X} in $(n - 1)S^2/\sigma^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$ by μ .

$$\begin{aligned} L &:= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 / \sigma^2. \end{aligned}$$

Notice that RHS is sum of independent r.v.s

□

Aside: distributions derived from normal

Proof.

$$L = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 / \sigma^2.$$

Letting $R := n(\bar{X} - \mu)^2 / \sigma^2 = (\frac{X-\mu}{\sigma/\sqrt{n}})^2$, we have

$$M_L(t) = M_{(n-1)S^2/\sigma^2} M_R(t),$$

hence

$$M_{(n-1)S^2/\sigma^2}(t) = \frac{M_L(t)}{M_R(t)} = \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{n}{2}} / \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{1}{2}} = \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{n-1}{2}},$$

hence $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.

We used here $M_X(t) = \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{n}{2}}$ for $X \sim \chi_n^2$. □

Parameter estimation

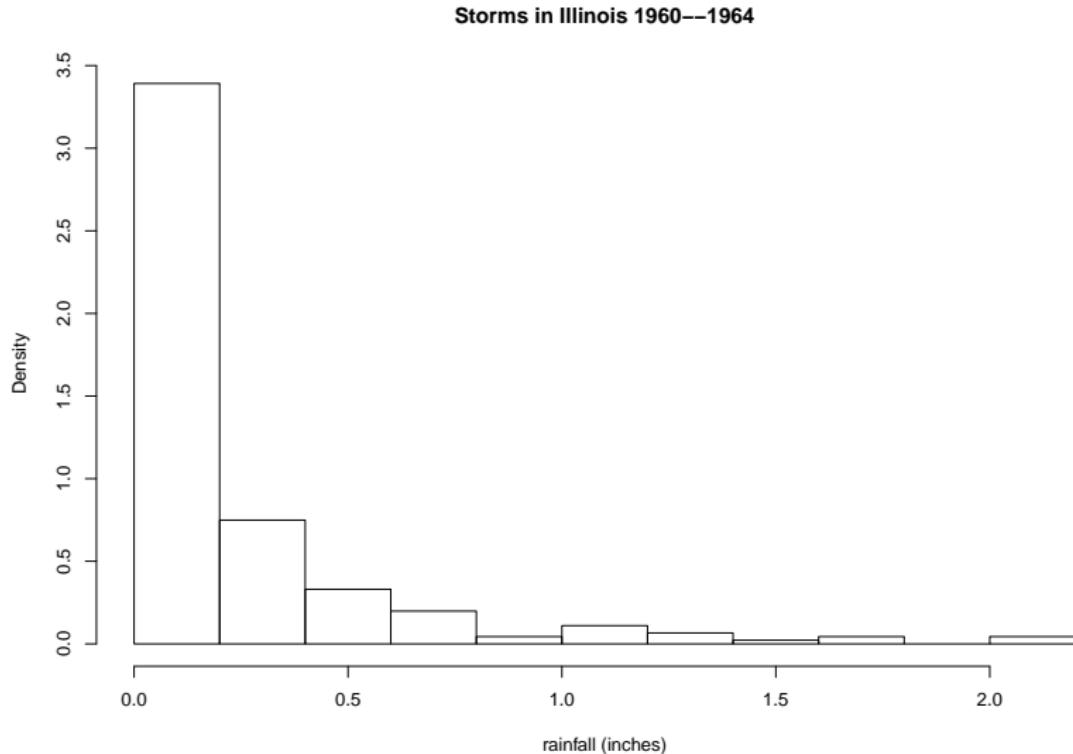


Figure: Precipitation in 227 Illinois storms.

Parameter estimation

Example (Illinois storms)

Would like to fit a probability distribution to the data. Since histogram is skewed, try to fit a gamma distribution.

Find parameters of gamma distribution via method of moments.

Parameter estimation

Example (Illinois storms)

Recall: $G \sim \Gamma(\alpha, \lambda)$ has moments

$$\mathbb{E}[G^k] = \frac{\Gamma(\alpha + k)}{\Gamma(k)\lambda^k},$$

hence (since $\Gamma(x + 1) = x\Gamma(x)$)

$$\mu_1 = \mathbb{E}[G] = \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)\lambda} = \frac{\alpha}{\lambda},$$

$$\mu_2 = \mathbb{E}[G^2] = \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)\lambda^2} = \frac{(\alpha + 1)\alpha}{\lambda^2} = \frac{\alpha^2 + \alpha}{\lambda^2},$$

and solving for α and λ , we have $\mu_2 = \mu_1^2 + \mu_1/\lambda$, hence

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2} = \frac{\mu_1}{\sigma^2} \quad \text{and} \quad \alpha = \lambda\mu_1 = \frac{\mu_1^2}{\sigma^2}.$$

Parameter estimation

Example (Illinois storms)

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2} = \frac{\mu_1}{\sigma^2} \quad \text{and} \quad \alpha = \lambda\mu_1 = \frac{\mu_1^2}{\sigma^2}$$

Substituting sample moments for population moments we obtain the method of moments estimators

$$\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2} \quad \hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}.$$

[show R script]

From data we find $\bar{X} = 0.224$, $\hat{\sigma} = 0.366$, hence

$$\hat{\lambda} = 1.672 \quad \hat{\alpha} = 0.374.$$

Parameter estimation

Example (Illinois storms)

$$\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2} \quad \hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}$$

Studying sampling distributions of $\hat{\lambda}$ and $\hat{\alpha}$ analytically (even working out standard errors) seems hopeless, as they are complicated functions of the observations.

Luckily, can still study sampling distribution with so-called (*parametric*) *bootstrap*, a versatile simulation method, thanks to computers.

Bootstrap

Example (Illinois storms)

Idea: Suppose we knew true values of α and λ , let's call them α_0 and λ_0 .

Could then simulate many drawings of samples, $i = 1, \dots, 1000$ say,

$$X_1^{(i)}, X_2^{(i)}, \dots, X_n^{(i)} \sim \text{Gamma}(\alpha_0, \lambda_0)$$

of size $n = 227$ and compute estimator $\hat{\alpha}_i^*$ for each of them.

Histogram of the $\hat{\alpha}_i^*$ should then be a good approximation of the sampling distribution of $\hat{\alpha}_{\text{MoM}}$.

Bootstrap

Example (Illinois storms)

Consequently, standard error

$$s_{\hat{\alpha}} := \sqrt{\frac{1}{B} \sum_{i=1}^B (\alpha_{MoM,i}^* - \bar{\alpha})^2}, \quad \left(\bar{\alpha} := \frac{1}{B} \sum_{i=1}^B \alpha_{MoM,i}^* \right)$$

of (simulated) estimators $\hat{\alpha}_1^*, \hat{\alpha}_2^*, \dots, \hat{\alpha}_B^*$ should be good approximation of standard error of $\hat{\alpha}$.

Problem: We don't know true values α_0, λ_0 .

Bootstrap

Example (Illinois storms)

However, since we don't know α_0 nor λ_0 , we use their method of moments estimates instead.

[show histogram and standard deviation of α_i^* in R]

Maximum likelihood estimator

Maximum likelihood estimators. Setup

Assume data to be observations of r.v.s

$$X_1, X_2, \dots, X_n.$$

Additionally, assume joint distribution \mathbb{P}_θ of (X_1, \dots, X_n) has probability density f .

Given observations $X_1 = x_1, \dots, X_n = x_n$ let

$$\text{lik}(\theta) := f(x_1, \dots, x_n | \theta)$$

denote the *likelihood* of θ .

Think of x_1, \dots, x_n as being fixed, and of θ as varying.

We ask: what is the most likely value for θ , given the observed data?

Maximum likelihood estimators.

Definition

The value $\hat{\theta}$ that maximizes the likelihood function, i.e.

$$\hat{\theta} = \arg \max_{\theta} \text{lik}(\theta)$$

is called the *maximum likelihood estimate (MLE)* for θ .

Maximum likelihood estimators.

Example (Proportion of defectives)

We know that a proportion p of products from a certain manufacturer are defective, but we don't know the value of p .

We independently sample $n = 100$ of the manufacturer's products and find $S_n = 37$ of them to be defective.

What is your "best" guess for p ?

Maximum likelihood estimators.

Example (Proportion of defectives)

Now, let's find the MLE for p . Let

$$X_i := \begin{cases} 1 & \text{if } i\text{th product is defective} \\ 0 & \text{otherwise.} \end{cases}$$

$$X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bernoulli}(p)$$

We find for the likelihood function

$$\text{lik}(p) = f(X_1, \dots, X_n | p) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} = p^{S_n} (1-p)^{n-S_n},$$

where $S_n := \sum_{i=1}^n X_i$.

Maximum likelihood estimators.

Example (Proportion of defectives)

likelihood function

$$\text{lik}(p) = p^{S_n} (1-p)^{n-S_n}$$

The log likelihood function is

$$l(p) = \log \text{lik}(p) = S_n \log p + (n - S_n) \log(1 - p),$$

and its derivative

$$\frac{d}{dp} l(p) = \frac{S_n}{p} - \frac{n - S_n}{1 - p}$$

has root

$$\hat{p} = \frac{1}{n} S_n = \bar{X},$$

the MLE for p .

[Ex: Show that S_n/n is also the method of moments estimator for p .]

Maximum likelihood estimators.

Suppose that $\hat{\theta}_n$ is an estimator of θ based on a sample of size n .
The sequence $(\hat{\theta}_n)$ of estimators is called *consistent*, if we have convergence

$$\hat{\theta}_n \rightarrow \theta$$

(we are not specific about mode of convergence here: could be in probability, or distribution).

Example (Proportion of defectives)

Law of Large numbers implies

$$\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mathbb{E}[X_1] = p$$

as $n \rightarrow \infty$, thus $\hat{p} = \hat{p}_n$ (really the sequence (\hat{p}_n)) is a consistent estimator for p .

Maximum likelihood estimators.

Example (MLE of normal distribution)

$$X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma)^2$$

$$f(x_1, \dots, x_n | \mu, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_i - \mu}{\sigma})^2}$$

To find the maximum of the likelihood function it is often convenient to maximize the *log likelihood function*

$$\begin{aligned} l(\mu, \sigma) &:= \log \text{lik}(\mu, \sigma) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \\ &= n \log \frac{1}{\sqrt{2\pi}} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \end{aligned}$$

²Here we have $\theta = (\mu, \sigma)$.

Maximum likelihood estimators.

Example (MLE of normal distribution)

$$l(\mu, \sigma) = n \log \frac{1}{\sqrt{2\pi}} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

Solving for roots of partial derivatives we obtain

$$0 = \frac{\partial}{\partial \mu} l(\mu, \sigma) = -\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) \rightsquigarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

$$0 = \frac{\partial}{\partial \sigma} l(\mu, \sigma) = -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^n (X_i - \mu)^2 \rightsquigarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

as MLEs for μ and σ . Know sampling distribution of \bar{X} if σ^2 is known.

Bayesian estimators

Conditional densities.

Example (Bayesian inference)

Often in medical problems it is assumed that a drug is effective with some (unknown) probability Π in each treatment, independently across treatments.

One challenge is to estimate (“learn”) effectiveness Π of a drug from the results of n treatments.

If we have no prior knowledge about Π , seems reasonable to assume it is a random number distributed uniformly on $[0, 1]$.
[In Bayesian jargon, this is called the *uninformative prior*.]

$$X := \# \text{ effective treatments}^3$$

Goal: “Update” distribution of Π given the observed number of effective treatments X .

³What is your guess for the distribution of X ?

Conditional densities.

Example (Bayesian inference)

More formally, we want to find conditional distribution

$$f_{\Pi}(p|X = x)$$

of effectiveness given $X = x$ effective treatments.

Maybe we can find joint density of (Π, X) first?

From the setup of the experiment

$$f_X(x|\Pi = p) = \binom{n}{x} p^x (1-p)^{n-x} \quad (0 \leq x \leq n)$$

hence, joint density of (Π, X) is

$$f(p, x) = f_X(x|\Pi = p) f_{\Pi}(p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

for $0 \leq p \leq 1, 0 \leq x \leq n$, since $\Pi \sim U(0, 1)$.

Conditional densities.

Example (Bayesian inference)

For $a, b > 0$ the distribution on $(0, 1)$ with density

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$$

is called the *beta(a, b) distribution*. Notice that this implies

$$\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

$$\begin{aligned} f_X(x) &= \int_0^1 f(p, x)dp = \binom{n}{x} \int_0^1 p^x(1-p)^{n-x}dp \\ &= \frac{n!}{x!(n-x)!} \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} = \frac{1}{n+1}, \end{aligned}$$

i.e. the number of effective treatments has uniform distribution.

Conditional densities.

Example (Bayesian inference)

For the density of the “updated effectiveness” we obtain

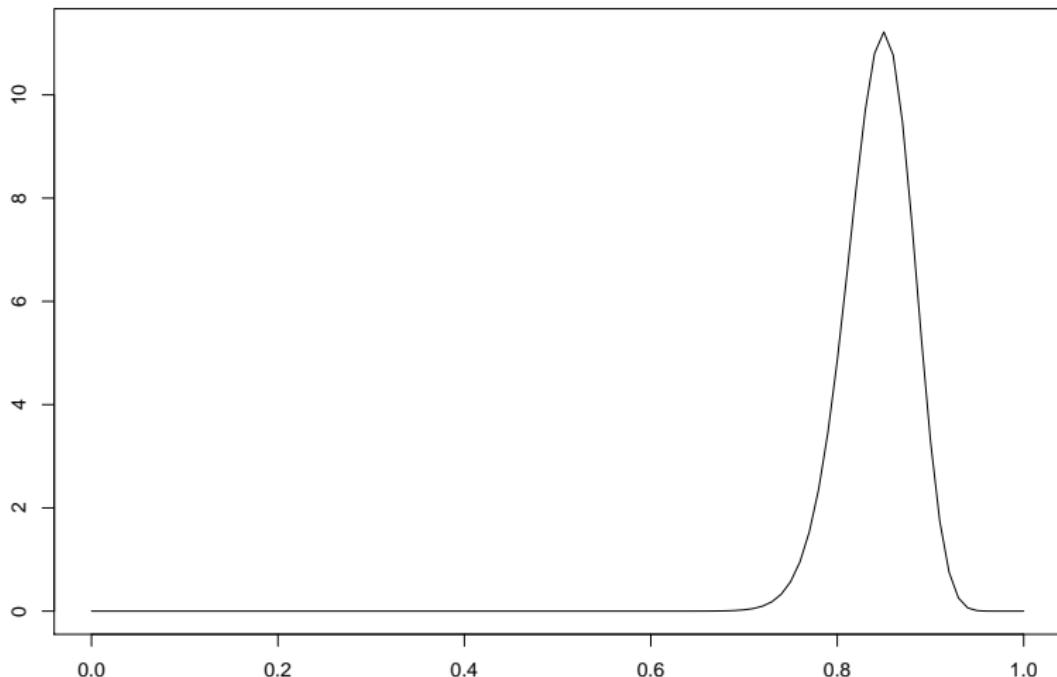
$$\begin{aligned}f_{\Pi}(p|X = x) &= \frac{f(p, x)}{f_X(x)} = (n + 1) \binom{n}{x} p^x (1 - p)^{n-x} \\&= \frac{\Gamma(n + 2)}{\Gamma(x + 1)\Gamma(n - x + 1)} p^x (1 - p)^{n-x},\end{aligned}$$

the density of a $\text{beta}(x + 1, n - x + 1)$ distribution.

$X = 85$ out of $n = 100$ treatments are found to be effective.

Given this observation, we update our prior distribution to
 $(\Pi|X = 85) \sim \text{beta}(86, 16)$, the so-called *posterior distribution*.

beta(86, 16) density



Conditional densities.

Example (Bayesian inference)

Posterior distribution:

$$(\Pi|X = x) \sim \text{beta}(x + 1, n - x + 1)$$

A natural estimator for the effectiveness Π given that out of n treatments $X = x$ were effective, is the mean of the posterior distribution⁴

$$\begin{aligned}\mathbb{E}[\Pi|X = x] &= \int_0^1 t \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} t^x (1-t)^{n-x} dt \\ &= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \frac{\Gamma(x+2)\Gamma(n-x+1)}{\Gamma(n+3)} = \frac{x+1}{n+2}\end{aligned}$$

This is called the *posterior mean* for Π .

⁴Or the mode.

More examples

Parameter estimation

Following example appears in different disguises in science, medicine, engineering, etc. We consider settings where observations can be assigned to different categories (categorial data).

Example (Emergency room)

Emergency room of large hospital assigns patients to one of three categories:

1. Stable. No immediate treatment required.
2. Serious. Immediate treatment not required, but patient needs to be monitored until physician available.
3. Critical. Patient's life endangered without immediate treatment.

Parameter estimation

Example (Emergency room)

Hospital records over past week show that

- ▶ 300 patients were classified as stable,
- ▶ 180 patients classified serious,
- ▶ 120 patients classified critical,

To ensure optimal organization, administration needs to estimate long-run frequency p_i of patients that are classified in category i .

Find estimators for p_1, p_2, p_3 .

(Make a guess before we proceed formally!)

Parameter estimation

Review: multinomial distribution. Think of n different marbles that we paint in c different colors. We paint marbles independently one after the other, with

$$\mathbb{P}\{\text{a particular marble is painted in color } i\} = p_i$$

$(\sum_i p_i = 1, p_i \geq 0)$. Let

$$X_i := \# \text{ marbles of color } i.$$

Then⁵

$$\mathbb{P}\{X_1 = x_1, \dots, X_c = x_c\} = \binom{n}{x_1, \dots, x_n} \prod_{i=1}^c p_i^{x_i}$$

if $\sum_i x_i = n$ and = 0 otherwise. The vector (X_1, \dots, X_n) has a *multinomial distribution* with parameters (n, p_1, \dots, p_c) denoted

$$(X_1, \dots, X_c) \sim \text{multinomial}(n, p_1, \dots, p_c).$$

⁵ $\binom{n}{x_1, \dots, x_n} = n!/(x_1!x_2!\cdots x_c!)$ is the multinomial coefficient.

Parameter estimation

Let us work out the MLE for p_1, p_2, \dots, p_c .

Notice that (X_1, \dots, X_c) are not i.i.d.

Want to maximize log likelihood

$$\begin{aligned} l(p_1, \dots, p_c) &= \log f(x_1, \dots, x_c | p_1, \dots, p_c) \\ &= \log \binom{n}{x_1, \dots, x_c} \prod_{i=1}^c p_i^{x_i} \\ &= \log n! - \sum_{i=1}^c \log x_i! + \sum_{i=1}^c x_i \log p_i \end{aligned}$$

subject to $p_1 + p_2 + \dots + p_c = 1$.

Maximize instead

$$L(p_1, \dots, p_c; \lambda) = \log n! - \sum_{i=1}^c \log x_i! + \sum_{i=1}^c x_i \log p_i + \lambda \left(\sum_{i=1}^c p_i - 1 \right).$$

Parameter estimation

Maximize instead

$$L(p_1, \dots, p_c; \lambda) = \log n! - \sum_{i=1}^c \log x_i! + \sum_{i=1}^c x_i \log p_i + \lambda \left(\sum_{i=1}^c p_i - 1 \right).$$

For any $j = 1, \dots, c$ the partial derivative

$$\frac{d}{dp_j} L = \frac{x_j}{p_j} + \lambda$$

has root

$$\hat{p}_j = -\frac{x_j}{\lambda},$$

and summing both sides w.r.t. j we obtain

$$1 = -\frac{1}{\lambda} \sum_{j=1}^c x_j = -\frac{n}{\lambda} \text{ hence } \lambda = -n.$$

where we used the constraint $\sum_j \hat{p}_j = 1$. MLE $\boxed{\hat{p}_j = \frac{x_j}{n}}.$

What can we say about the sampling distribution of \hat{p}_j ?

Parameter estimation

Example (Emergency room)

Recall hospital records:

- ▶ 300 patients were classified as stable,
- ▶ 180 patients classified serious,
- ▶ 120 patients classified critical.

Find MLEs

$$\hat{p}_1 = \frac{300}{600} = 50\% \quad \hat{p}_2 = \frac{180}{600} = 30\% \quad \hat{p}_3 = \frac{120}{600} = 20\%.$$

Ex: Work out the method of moments estimator for p_i .

STAT 135, Concepts of Statistics

Helmut Pitters

Confidence intervals

Department of Statistics
University of California, Berkeley

February 16, 2017

Confidence intervals.

Example: Opinion polls. Consider town of $N = 25,000$ eligible voters. Taking a simple random sample X_1, \dots, X_{1600} of size $n = 1,600 = 40^2$ we find that 917 support Democrats. Suppose we try to estimate percentage

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = p$$

of people who support Democrats, where

$$x_i = \begin{cases} 1 & \text{if } i\text{th person votes for Democrats} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \frac{2}{N} \mu \sum_{i=1}^N x_i + \mu^2 = p(1 - p).$$

Confidence intervals.

Example: Opinion polls, cont. Idea: use

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{917}{1600} \approx 0.57$$

as estimator for p . We know (simple random sampling):

$$\mathbb{E}\hat{p} = p, \quad \sigma_{\hat{p}} = \sqrt{\text{Var}(\hat{p})} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}} \approx \frac{\sigma}{40},$$

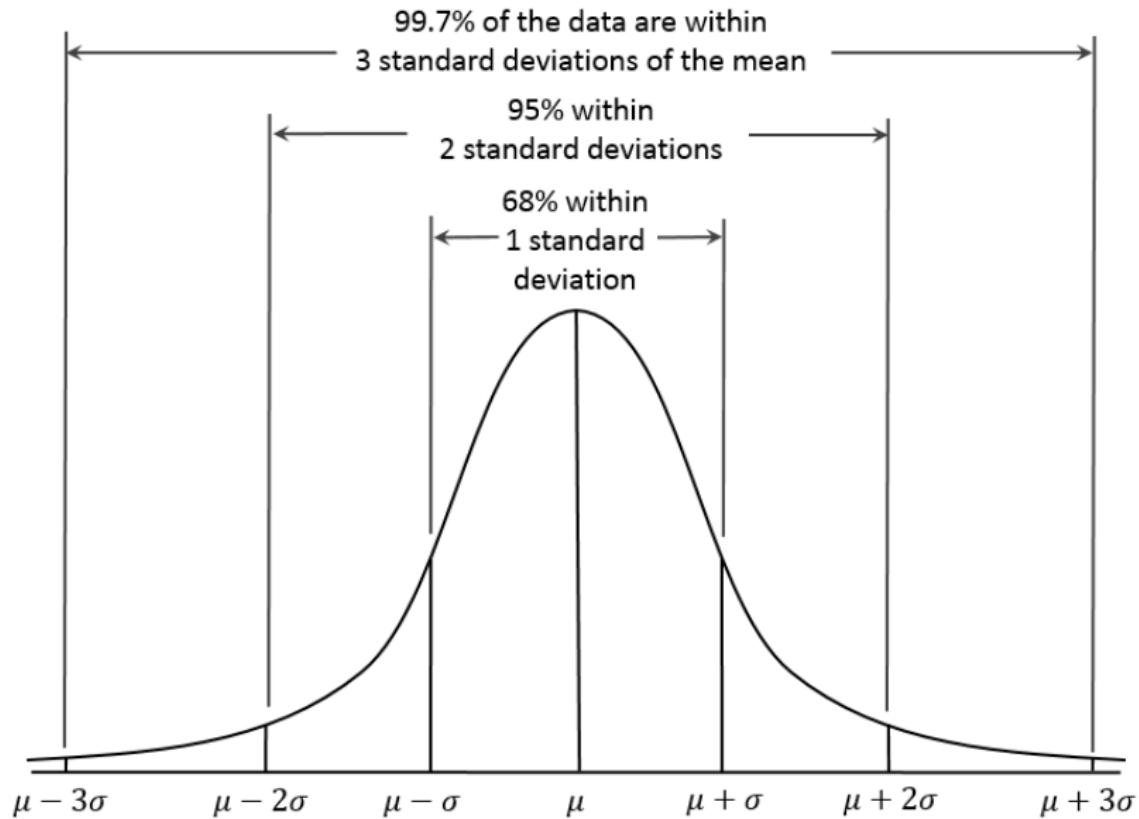
since sampling fraction $n/N = 1,600/25,000 = 0.064$ is small.
Moreover, since $\sigma = \sqrt{p(1-p)}$ not known, estimate it by

$$\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})} \approx 0.5,$$

i.e. estimate standard error $\sigma_{\hat{p}}$ by $0.0125 = 1.25\%$. Put differently,
the percentage $\hat{p} \approx 0.57$ of Democrats in the sample is likely to be
off the percentage of Democrats among all 25,000 eligible voters
by 1.25 percentage points or so.

Confidence intervals.

Recall the empirical rule for the standard Normal distribution.



Confidence intervals.

Example: Opinion polls, cont. Approximate \hat{p} ($= \bar{X}$) by $\mathcal{N}(\hat{p}, \hat{p}(1 - \hat{p})/n) = \mathcal{N}(0.57, 0.25/1600)$ (due to CLT). Hence

Table: Confidence intervals

$\hat{p} \pm 1.25\% = [0.55, 0.58]$	68.3%	confidence interval for p
$\hat{p} \pm 2 \times 1.25\% = [0.54, 0.6]$	95.5%	"
$\hat{p} \pm 3 \times 1.25\% = [0.53, 0.61]$	99.7%	"

Statistic.

Call a map $T(x_1, \dots, x_n)$ of given data x_1, \dots, x_n a *statistic*. Usually, regard these data as observed values of some random variables X_1, \dots, X_n .

Moreover, in the case at hand, one needs to specify the (joint) distribution of the X_i that depends on some parameter, generically called $\theta > 0$.

Confidence intervals.

Consider a parameter θ that we want to estimate.¹ Suppose $a(X), b(X)$ are statistics such that

$$a(x) \leq b(x)$$

for all observations x generated by some random variable X , and that on seeing data $X = x$ we infer

$$a(X) \leq \theta \leq b(X).$$

If

$$\mathbb{P}\{a(X) \leq \theta \leq b(X)\} = 1 - \alpha$$

does not depend on θ , the random interval

$$[a(X), b(X)]$$

is called a $100(1 - \alpha)\%$ confidence interval² for θ .

¹E.g. think of θ as a population parameter in simple random sampling.

²Typically α is taken to be 0.05 or 0.01 so that probability that confidence interval contains θ is high.

Confidence intervals.

Example. X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with *unknown* μ and *known* σ^2 .

Want to find 95% confidence interval for μ . Recall that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ and suppose $a \leq b$ are such that

$$\mathbb{P} \left\{ a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b \right\} = 1 - \alpha$$

which is equivalent to

$$\mathbb{P} \left\{ \bar{X} - b \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - a \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha.$$

Due to symmetry of Normal distribution, length of confidence interval minimized for $-a = b$. Since $\Phi(1.96) = 0.975$,

$$\left[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right]$$

is a 95% confidence interval for μ .

t distribution

Definition

Let N, X be independent random variables such that

$$N \sim \mathcal{N}(0, 1) \quad X \sim \chi_n^2.$$

The distribution of

$$\frac{N}{\sqrt{X/n}}$$

is called a *t distribution with n degrees of freedom*.

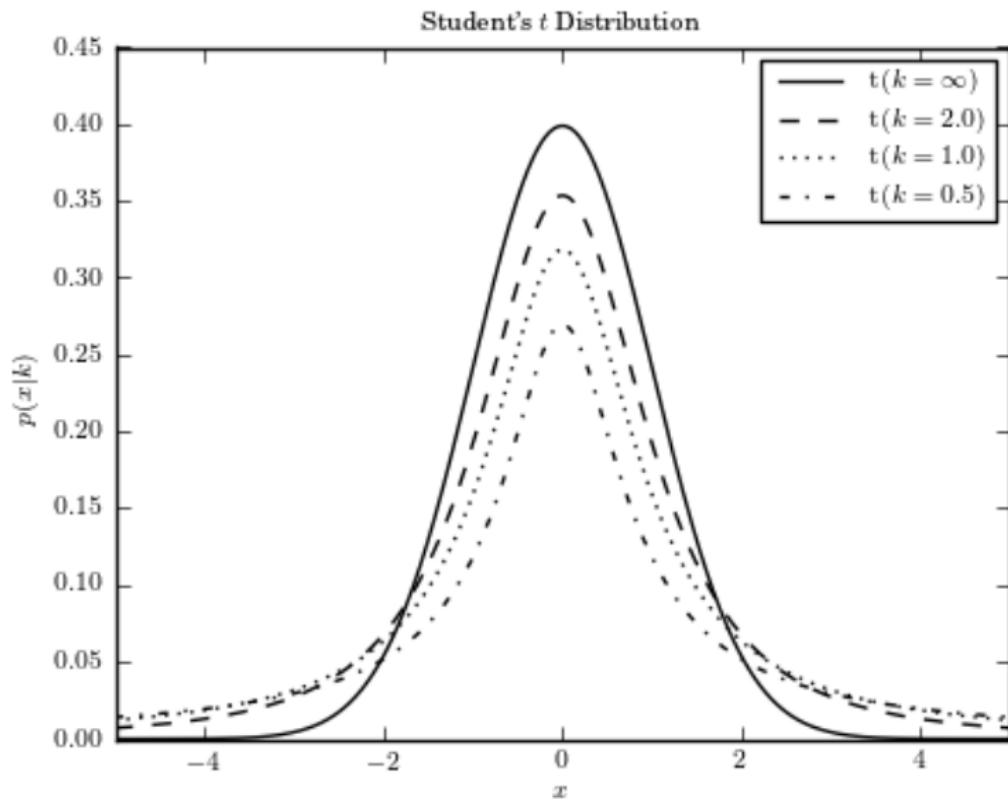
Fact

One can show that the t distribution has density

$$f(t) = \begin{cases} \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} & t \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Notice that the density of the t distribution is symmetric about 0.

t distribution



Confidence intervals.

Example. X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with *both* μ and σ^2 unknown. Want to find (shortest) 95% confidence interval for μ . From our results on distributions derived from Normal, we obtain:³

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n), \quad (n-1)S^2/\sigma^2 \sim \chi_{n-1}^2,$$

and

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

follows Student's t-distribution with $n - 1$ degrees of freedom.⁴

³Sample variance was defined to be $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

⁴However, if sample size n is large, a Normal approximation might be considered where σ^2 is estimated by the sample variance.

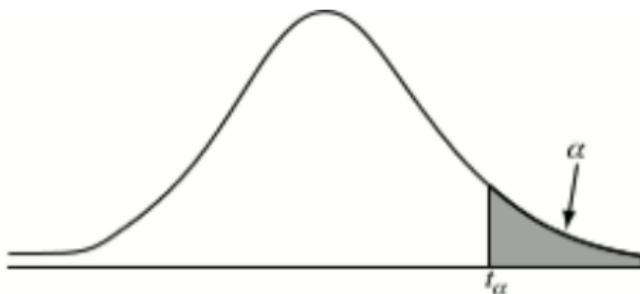
Confidence intervals.

Want 95% confidence interval which is symmetric about 0,
i.e. want b such that

$$\mathbb{P}\{-b \leq T_{n-1} \leq b\} = 95\%,$$

where T_{n-1} follows Student's t distribution with $n - 1$ degrees of freedom. Find from tabulated values of percentiles of t distribution...

Confidence intervals.



Values of α for one-tailed test and $\alpha/2$ for two-tailed test

df	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	$t_{0.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930

Confidence intervals.

Find from tabulated values of percentiles of t distribution
for $n = 11$

$$\mathbb{P} \left\{ -2.201 \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq 2.201 \right\} = 95\%,$$

i.e.

$$\left[\bar{X} - 2.201 \frac{S}{\sqrt{n}}, \bar{X} + 2.201 \frac{S}{\sqrt{n}} \right]$$

is a 95% confidence interval for μ .

Confidence intervals.

How can we find a confidence interval for σ^2 ?

Recall

$$(n - 1)S^2 / \sigma^2 \sim \chi_{n-1}^2.$$

Now define x_α by

$$\mathbb{P}\{X \leq x_\alpha\} = \alpha,$$

where X is some r.v. with distribution χ_{n-1}^2 . Then

$$\begin{aligned}\alpha &= \mathbb{P}\left\{x_{\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq x_{1-\alpha/2}\right\} \\ &= \mathbb{P}\left\{\frac{(n-1)S^2}{x_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{x_{\alpha/2}}\right\},\end{aligned}$$

thus

$$\left[\frac{(n-1)S^2}{x_{1-\alpha/2}}, \frac{(n-1)S^2}{x_{\alpha/2}}\right]$$

is a α -confidence interval for σ^2 .

STAT 135, Concepts of Statistics

Helmut Pitters

Sufficiency

Department of Statistics
University of California, Berkeley

February 23, 2017

Sufficiency.

Example (Coin tossing)

Flip coin 10 times and record pattern HHTHHHTHTT.

1. Natural guess for probability p for heads?

$$\frac{\text{\# of heads in HHTHHHTHTT}}{10}?$$

2. Imagine we throw the coin 10^6 times

HTTHHHHHHHTTT ...

Pointless to analyze details of corresponding pattern of heads and tails.

To estimate p , seems sufficient to know number (statistic)

$$h(\text{HTTHHHHHHHTTT} \dots)$$

of heads observed. $h(\dots)$ is said to be a *sufficient statistic* for n

Sufficiency.

From sample

$$(X_1, X_2, \dots, X_n) \sim \mathbb{P}_\theta$$

want to learn θ .

If sample size n is large, may be hard to interpret list of numbers x_1, x_2, \dots, x_n . Instead, might be enough to consider some key features, e.g.

mean, standard deviation, $x_{(1)} = \min_i x_i$, $x_{(n)} = \max_i x_i$,

etc. that are functions of the data (“statistics” in statistical jargon).

Statistics reduce/compress the data.

Natural questions:

- ▶ How can we compress data without compromising quality of inference?
- ▶ Is there an “optimal” method to compress? If so, how can we find it?

Sufficiency.

More generally: A statistic $T(X_1, \dots, X_n)$ is called *sufficient for θ* if any inference about θ depends on X_1, \dots, X_n only via $T(X_1, \dots, X_n)$.

Definition (Sufficient statistic)

Statistic $T(X_1, \dots, X_n)$ is called *sufficient statistic* for θ if conditional distribution of X_1, \dots, X_n given $T = t$, i.e.

$$\mathbb{P}_\theta\{X_1 \in \cdot, \dots, X_n \in \cdot | T = t\}$$

does not depend on θ for any value of t .

In other words: Inference of θ is not improved by gaining more information about X_1, \dots, X_n than is contained in $T(X_1, \dots, X_n)$.

Sufficiency.

Example (Coin tossing)

Consider again n independent tosses of a coin that shows up heads w.p. p . Let

$$X_i := \begin{cases} 1 & \text{coin shows heads in } i\text{th toss} \\ 0 & \text{otherwise.} \end{cases}$$

Argued earlier that, intuitively,

$$H := H(X_1, \dots, X_n) := \sum_{i=1}^n X_i = \# \text{ of heads}$$

should be sufficient statistic for p .

Sufficiency.

Example (Coin tossing)

Does H satisfy definition of sufficiency?

$$\mathbb{P}\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | H = h\}$$

$$= \frac{\mathbb{P}\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}}{\mathbb{P}\{H = h\}} = \frac{p^h(1-p)^{n-h}}{{n \choose h} p^h(1-p)^{n-h}} = {n \choose h}^{-1}$$

does not depend on p , therefore H is sufficient stat. for p .

Remark

Notice that sufficient statistic need not be unique, e.g. statistic $2H$ would do just as well as H .

Sufficiency.

Theorem (Factorization theorem)

Statistic T is sufficient for θ , if and only if $f(x|\theta)$ can be written as

$$f(x|\theta) = g(T(x), \theta)h(x). \quad (1)$$

Remark

Recall that MLE for θ is the value $\hat{\theta}$ that maximizes $f(x|\theta)$.

Suppose T is sufficient for θ . Because of the factorization theorem, $\hat{\theta}$ maximizes $f(x|\theta)$ if and only if it maximizes $g(T(x), \theta)$, in other words, MLE is a function of the sufficient statistic $T(X)$.

Sufficiency.

Proof of factorization theorem.

We prove this theorem only for the discrete case. Suppose $f(x|\theta) = \mathbb{P}_\theta\{X = x\}$ satisfies above factorization and $T(X) = t$. Then

$$\begin{aligned}\mathbb{P}_\theta\{X = x|T(X) = t\} &= \frac{\mathbb{P}_\theta\{X = x\}}{\mathbb{P}_\theta\{T(X) = t\}} \\ &= \frac{g(T(x), \theta)h(x)}{\sum_{x: T(x)=t} g(T(x), \theta)h(x)} = \frac{g(t, \theta)h(x)}{\sum_{x: T(x)=t} g(t, \theta)h(x)} \\ &= \frac{h(x)}{\sum_{x: T(x)=t} h(x)}, \text{ and this quantity does not depend on } \theta.\end{aligned}$$

□

Sufficiency.

Proof.

Suppose now that T is sufficient and $T(X) = t$. Then

$$\mathbb{P}_\theta\{X = x\} = \mathbb{P}_\theta\{X = x | T(X) = t\}\mathbb{P}_\theta\{T(X) = t\},$$

where the first factor does not depend on θ (by sufficiency), hence factorization is given by

$$h(x) := \mathbb{P}_\theta\{X = x | T(X) = t\}$$

and

$$g(T(X), \theta) := \mathbb{P}_\theta\{T(X) = t\}.$$



Sufficiency.

Example (uniform, one parameter). Let X_1, \dots, X_n be i.i.d. with uniform distribution on $[0, \theta]$. Want to estimate unknown θ .

Write $x = (x_1, \dots, x_n)$.

$$f(x|\theta) = \prod_{i=1}^n \frac{1}{\theta} \mathbf{1}_{[0,\theta]}(x_i) = \theta^{-n} \mathbf{1}_{[0,\theta]}(\max_i x_i) \quad \text{for } x_1, \dots, x_n \geq 0,$$

where

$$\mathbf{1}_A(z) := \begin{cases} 1 & \text{if } z \in A \\ 0 & \text{otherwise} \end{cases}$$
 denotes the indicator of A .

$T(x) := \max_i x_i$ is sufficient statistic for θ , since

$$\begin{aligned} g(T(x), \theta) &:= \theta^{-n} \mathbf{1}_{[0,\theta]}(T(x)) \\ h(x) &:= 1. \end{aligned}$$

Moreover, $\max_i x_i$ is the MLE for θ , since it maximizes $f(x|\theta)$.

Sufficiency.

Example (Poisson). Let X_1, \dots, X_n be i.i.d. with $\text{Poisson}(\lambda)$ distribution. Want to estimate unknown λ .

$$f(x|\lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_i x_i}}{\prod_i x_i!} = g(T(x), \lambda)h(x),$$

where

$$T(x) := \sum_i x_i$$

$$g(T(x), \lambda) := e^{-\lambda n} \lambda^{T(x)}$$

$$h(x) := \frac{1}{\prod_i x_i!}.$$

By the factorization theorem, $\sum_i X_i$ is a sufficient statistic for λ .

Sufficiency.

Definition

If X is an estimator for θ , its *mean squared error* is defined by

$$\boxed{\text{MSE}(X) := \mathbb{E}(X - \theta)^2}$$

and often used to measure the accuracy of an estimate.

If $\mathbb{E}(X - \theta)^2 < \infty$ we have

$$\mathbb{E}(X - \theta)^2 = \text{Var}(X) + b^2(\theta, X),$$

where

$$b(\theta, X) := \mathbb{E}[X] - \theta$$

is the *bias* of X .

Sufficiency.

The next theorem shows that if we look for an estimator with small MSE, it is enough to consider estimators that are functions of sufficient statistics.

Theorem (Rao-Blackwell)

Let $\hat{\theta}$ be an estimator of θ such that $\mathbb{E}\hat{\theta}^2 = \mathbb{E}_\theta\hat{\theta}^2 < \infty$ for all θ .

Suppose that T is sufficient for θ , and define $\tilde{\theta} := \mathbb{E}[\hat{\theta}|T]$ to be the conditional expectation of $\hat{\theta}$ given T . Then, for all θ

$$\text{MSE}(\tilde{\theta}) = \mathbb{E}(\tilde{\theta} - \theta)^2 \leq \mathbb{E}(\hat{\theta} - \theta)^2 = \text{MSE}(\hat{\theta}).$$

The inequality is strict unless $\tilde{\theta} = \hat{\theta}$.

Sufficiency.

Proof.

(of Rao-Blackwell thm.) From the tower property of conditional expectation (i.e. $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$),

$$\mathbb{E}\hat{\theta} = \mathbb{E}[\mathbb{E}[\hat{\theta}|T]] = \mathbb{E}\tilde{\theta}.$$

Consequently, $\tilde{\theta}$ and $\hat{\theta}$, have the same bias, and

$$\text{MSE}(\tilde{\theta}) - \text{MSE}(\hat{\theta}) = \text{Var}(\tilde{\theta}) - \text{Var}(\hat{\theta}).$$

Recall the conditional variance formula

$$\begin{aligned}\text{Var}(\hat{\theta}) &= \text{Var}(\mathbb{E}[\hat{\theta}|T]) + \mathbb{E}[\text{Var}(\hat{\theta}|T)] \\ &= \text{Var}(\tilde{\theta}) + \mathbb{E}[\text{Var}(\tilde{\theta}|T)] \geq \text{Var}(\tilde{\theta}),\end{aligned}$$

with equality if and only if $\text{Var}(\tilde{\theta}) = 0$, in other words, $\hat{\theta}$ is a function of T . □

STAT 135, Concepts of Statistics

Helmut Pitters

Hypothesis testing 1

Department of Statistics
University of California, Berkeley

February 28, 2017

Hypothesis testing.

So far concerned ourselves with estimating population parameters.

Another important pillar of classical statistical inference:
testing whether or not an hypothesis about a population parameter
has to be rejected when checked against data.

Contexts in which hypothesis testing might appear in applications:

- ▶ Does specific drug/medical treatment have a positive effect on patients' health?
- ▶ Does specific ad increase sales of a product?
- ▶ Establish the authorship of documents.
- ▶ Is specific die fair?

Hypothesis testing.

Example

Political candidate T claims to gain 50% of votes in city election.

Conservative hypothesis: assume T is right, i.e.

“null hypothesis” $H_0: p = \frac{1}{2}$,

where

p = proportion of supporters of T in the electorate.

Might be skeptical of T's claim and seek to support

“alternative hypothesis” $H_A: p < \frac{1}{2}$.

Among $n = 15$ randomly selected eligible voters, 8 favor T. Does this data support T's claim? Could high percentage $8/15 > 0.5$ of supporters be explained just by the chance of sampling?

Hypothesis testing.

Example

To decide whether or not to reject H_0 , define *test statistic*

$$S_n := \# \text{ supporters of T in sample of size } n.$$

Provided H_0 is true (i.e. $p = 0.5$),

$$S_n \sim \text{binomial}(n, 0.5),$$

Distribution of S_n under H_0 is called *null distribution*.

Hypothesis testing.

Example

Small values of S_n contradict H_0 , large values of S_n support H_0 .

If S_n is "small enough," want to reject H_0 .

How should we decide what a "small enough" value for S_n is?

Hypothesis testing.

Example

k	0	1	2	3	4	5		
$\mathbb{P}\{S_n \leq k\}$	0.0	0.0	0.004	0.018	0.059	0.151		
k	6	7	8	9	10	11	12	13
$\mathbb{P}\{S_n \leq k\}$	0.304	0.5	0.696	0.849	0.941	0.982	0.996	1.0

Table: Cumulative distribution function of Binomial(15, 0.5) (rounded to 3 decimals).

Decision rule. Want to find critical value k such that

$$\begin{cases} H_0 \text{ is rejected if } S_n \leq k \\ H_0 \text{ is not rejected if } S_n > k. \end{cases} \quad (1)$$

Hypothesis testing.

Remark (Types of errors)

Notice: irrespective of decision rule, there are two types of errors we can make:

	H_0 true	H_0 false
reject H_0	type I error	correct decision
do not reject H_0	correct decision	type II error

Table: Two types of errors in hypothesis testing.

Usually, H_0 taken to be a conservative hypothesis,
→ type I error considered "more serious" than type II error.

Example (Murder trial)

Null hypothesis should be: "accused is innocent" ('in dubio pro reo'), since type I error: "innocent person is convicted" considered more serious than type II error: "murderer is acquitted."

Hypothesis testing.

Remark

	H_0 true	H_0 false
reject H_0	type I error	correct decision
do not reject H_0	correct decision	type II error

Table: Two types of errors in hypothesis testing.

Probability

$$\alpha := \mathbb{P}\{\text{reject } H_0 | H_0\} = \mathbb{P}\{\text{type I error}\}$$

is called the *significance level* of a test.

Hypothesis testing.

In Neyman-Pearson paradigm of hypothesis testing controlling these two types of errors is central.

One starts by specifying significance level α . Commonly used values are $\alpha = 0.05$ or $\alpha = 0.01$.

Hypothesis testing.

Example

Let us fix

significance level $\alpha = 0.018$.

Let

$$k = 3$$

i.e. reject H_0 if $S_n \leq 3$. Then¹

$$\mathbb{P}\{\text{reject } H_0 | H_0\} = \mathbb{P}\{S_n \leq 3 | H_0\} = \sum_{i=0}^3 \text{binomial}(15, 0.5)(i) = 0.018,$$

Our risk to conclude that T will lose, if, in fact, he wins, is 0.018,
i.e. we'd make this error in 18 out of 1.000 cases.

¹For large values of n and k find these values in table of binomial probabilities or via R.

Hypothesis testing.

Example (Speed limit)

Consider certain stretch of a highway.

Original speed limit: 65mph

original avg. speed: 63mph.

On this stretch a new speed limit is set:

55mph.

In sample of $n = 100$ cars:

new avg. speed: $\bar{X}_n = 61.4\text{mph}$ with standard deviation $SD=4.6\text{mph}$

Suppose speeds X_1, \dots, X_{100} are i.i.d. $\mathcal{N}(\mu, \sigma^2)$.

Was avg. speed genuinely reduced, or is reduced speed in sample due to chance in sampling?

(Find a test with significance level $\alpha = 0.01$.)

Hypothesis testing.

Example (Speed limit)

Find a test with significance level $\alpha = 0.01$.

- ▶ $H_0 : \mu = 63.0\text{mph}$ (avg. speed unchanged)
- ▶ $H_A : \mu < 63.0\text{mph}$
- ▶ Under H_0 , distribution of test statistic

$$Z := \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim \mathcal{N}(0, 1) \quad \text{is approximately standard normal.}^2$$

- ▶ Small values of Z contradict H_0 , that is we're looking for r s.t.

$$\alpha = \mathbb{P}\{\text{type I error}\} = \mathbb{P}\{Z \leq (-\infty, r] | H_0\},$$

Look up α -percentile normal distribution to find $r \approx -2.32$.

- ▶ Decision rule > next slide

²Would use t-distribution for small sample size n .

Hypothesis testing.

Example (Speed limit)

- Decision rule for test with significance level $\alpha = 0.01$ is

$$\begin{cases} \text{reject } H_0 & \text{if } S \leq -2.32 \\ \text{do not reject } H_0 & \text{if } S > -2.32. \end{cases}$$

Hypothesis testing.

Example (Speed limit)

- Decision rule for test with significance level $\alpha = 0.01$ is

$$\begin{cases} \text{reject } H_0 & \text{if } S \leq -2.33 \\ \text{do not reject } H_0 & \text{if } S > -2.33. \end{cases}$$

- From data we find:

$$Z = \frac{61.4 - 63.0}{4.6/10} = -3.48 \leq -2.33$$

and hence reject H_0 based on the data.

Notice that rejecting H_0 is not a statement about whether or not the new speed limit *caused* the reduction of avg. speed.
(There could be numerous other reasons causing the reduction that are unknown to us.)

Hypothesis testing.

Terminology of Neyman-Pearson paradigm. Consider data x_1, \dots, x_n modeled as random samples from r.v.s X_1, \dots, X_n with common distribution \mathbb{P}_θ that depends on some parameter $\theta \in \Theta$.³
E.g.

- ▶ $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$, $\Theta = (-\infty, \infty)$.
- ▶ $\mathbb{P}_\theta = \text{Exponential}(\theta)$, $\Theta = (0, \infty)$.
- ▶ $\mathbb{P}_\theta = \text{binomial}(N, \theta)$, $\Theta = [0, 1]$.

³Tests dealing with such parameterized models are referred to as parametric tests.

Hypothesis testing.

Terminology of Neyman-Pearson paradigm. Two hypotheses to be examined on the basis of data:

null hypothesis	$H_0: \theta \in \Theta_0$
-----------------	----------------------------

This is usually a conservative hypothesis that is not to be rejected unless there is clear evidence to do so.

alternative hypothesis	$H_A: \theta \in \Theta_A$
------------------------	----------------------------

H_A specifies the kind of departure from H_0 one is interested in.
One has $\Theta_0 \cup \Theta_A = \Theta$ and $\Theta_0 \cap \Theta_A = \emptyset$.

If hypothesis specifies precisely one distribution, e.g.

$$H_A: \text{Poisson}(2.7)$$

it is called *simple*, otherwise it is called *composite*.

Hypothesis testing.

Terminology of Neyman-Pearson paradigm. Based on a *test statistic* $T = T(X_1, \dots, X_n)$ a test is defined by the *rejection region*, C say, and the decision rule

$$\begin{cases} \text{reject } H_0 & \text{if } T(X) \in C \\ \text{do not reject } H_0 & \text{if } T(X) \notin C \end{cases} \quad (2)$$

The complement \bar{C} of C is referred to as the *acceptance region*. As before

$$\alpha := \mathbb{P}\{\text{type I error}\},$$

and

$$\beta := \mathbb{P}\{\text{type II error}\} = \mathbb{P}\{\text{accept } H_0 | H_0 \text{ is false}\}.$$

The probability

$$1 - \beta = \mathbb{P}\{\text{reject } H_0 | H_0 \text{ is false}\}$$

to reject H_0 when it is false is called the *power* of the test.

STAT 135, Concepts of Statistics

Helmut Pitters

Hypothesis testing 2

Department of Statistics
University of California, Berkeley

March 7, 2017

Hypothesis testing.

Example (Adjusting lab equipment)

Hospital lab uses instrument to determine hemoglobin levels in blood samples. Instrument needs to be calibrated on regular basis—many states require daily checks of instruments. To this end, lab makes measurements using sample from standard blood supply.

On particular day technician carries out n independent measurements of standard blood supply with known mean hemoglobin level 15.1.

$H_0: \mu = 15.1$ (no adjustment needed)

$H_A: \mu \neq 15.1$ (instrument needs to be readjusted)

Hypothesis testing.

Example (Adjusting lab equipment)

Lab assumes $\bar{X}_n \sim \mathcal{N}(\mu, 0.16)$ and does not adjust instrument if error is within two standard deviations, i.e. has acceptance region

$$\bar{C} := 15.1 \pm 2 \times 0.4 = (14.3, 15.9)$$

yielding significance level

$$\begin{aligned}\alpha &= \mathbb{P}\{\text{type I error}\} = 1 - \mathbb{P}\{14.3 \leq \bar{X}_n \leq 15.9 | H_0\} \\ &= 1 - \left(\Phi\left(\frac{15.9 - 15.1}{0.4}\right) - \Phi\left(\frac{14.3 - 15.1}{0.4}\right) \right) \approx 0.046\end{aligned}$$

Hypothesis testing.

Example (Adjusting lab equipment)

However, probability β of type II error is not uniquely determined, since alternative

$$H_A: \mu \neq 15.1$$

contains more than one distribution. For any particular value $\mu^* \neq 15.1$ can find the power

$$\begin{aligned}1 - \beta &= 1 - \mathbb{P}\{\text{type II error}\} = \mathbb{P}\{\text{accept } H_0 | \mu = \mu^*\} \\&= 1 - \mathbb{P}\{14.3 \leq \bar{X}_n \leq 15.9 | \mu = \mu^*\} \\&= 1 - \Phi\left(\frac{15.9 - \mu^*}{0.4}\right) + \Phi\left(\frac{14.3 - \mu^*}{0.4}\right),\end{aligned}$$

therefore power ($= 1 - \beta$) is a function of μ .

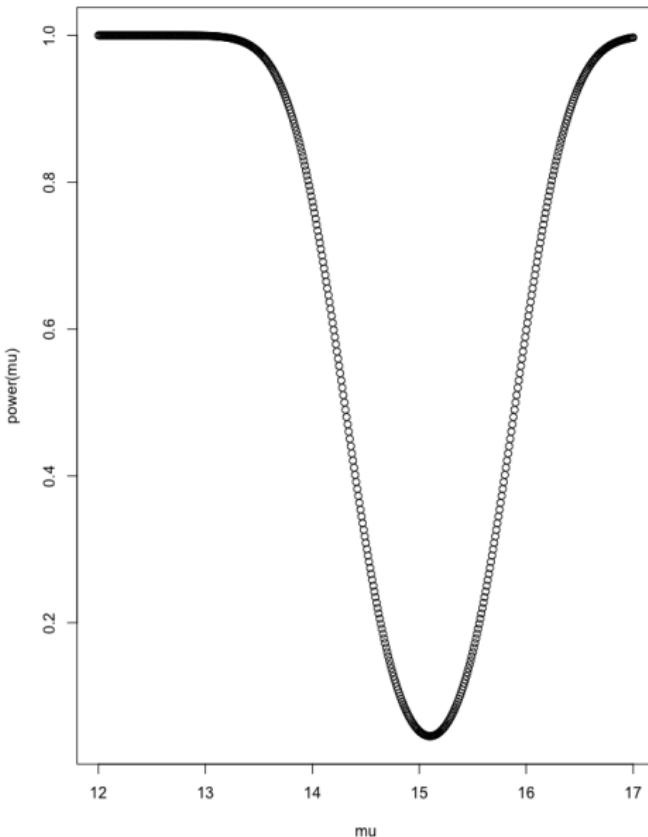


Figure: The power as a function of μ .

Hypothesis testing.

Example

Consider two coins, one black, one white, s.t.

$$\mathbb{P}_b\{\text{heads}\} = \mathbb{P}\{\text{heads}|\text{black coin thrown}\} = 0.5 \quad (1)$$

$$\mathbb{P}_w\{\text{heads}\} = \mathbb{P}\{\text{heads}|\text{white coin thrown}\} = 0.7. \quad (2)$$

One coin tossed $n = 10$ times, shows H heads.

Based on H , how would you decide which coin was used?

h	0	1	2	3	4	5
$\mathbb{P}_b\{H = h\}$	0.001	0.0098	0.0439	0.1172	0.2051	0.2461
$\mathbb{P}_w\{H = h\}$	0.0	0.0001	0.0014	0.009	0.0368	0.1029
h	6	7	8	9	10	
$\mathbb{P}_b\{H = h\}$	0.2051	0.1172	0.0439	0.0098	0.001	
$\mathbb{P}_w\{H = h\}$	0.2001	0.2668	0.2335	0.1211	0.0282	

Table: Probabilities of Binomial(10, p) for $p = 0.5$ and $p = 0.7$.

Hypothesis testing.

Example 4. Based on H , how would you decide which coin was tossed?

h	0	1	2	3	4	5
$\mathbb{P}_b\{H = h\}$	0.001	0.0098	0.0439	0.1172	0.2051	0.2461
$\mathbb{P}_w\{H = h\}$	0.0	0.0001	0.0014	0.009	0.0368	0.1029
h	6	7	8	9	10	
$\mathbb{P}_b\{H = h\}$	0.2051	0.1172	0.0439	0.0098	0.001	
$\mathbb{P}_w\{H = h\}$	0.2001	0.2668	0.2335	0.1211	0.0282	

Table: Probabilities of Binomial(10, p) for $p = 0.5$ and $p = 0.7$ (rounded to 3 decimals).

Suppose $H = 3$, then

$$\frac{\mathbb{P}_b\{H = 3\}}{\mathbb{P}_w\{H = 3\}} = \frac{0.1172}{0.009} \approx 13.0,$$

i.e. data suggest it is about 13 times more likely that black coin was thrown.

Hypothesis testing.

Example 4.

h	0	1	2	3	4	5
$\mathbb{P}_b\{H = h\}$	0.001	0.0098	0.0439	0.1172	0.2051	0.2461
$\mathbb{P}_w\{H = h\}$	0.0	0.0001	0.0014	0.009	0.0368	0.1029
h	6	7	8	9	10	
$\mathbb{P}_b\{H = h\}$	0.2051	0.1172	0.0439	0.0098	0.001	
$\mathbb{P}_w\{H = h\}$	0.2001	0.2668	0.2335	0.1211	0.0282	

Table: Probabilities of Binomial(10, p) for $p = 0.5$ and $p = 0.7$ (rounded to 3 decimals).

On the other hand, if $H = 9$, then

$$\frac{\mathbb{P}_b\{H = 9\}}{\mathbb{P}_w\{H = 9\}} = \frac{0.0098}{0.1211} \approx 0.081,$$

i.e. data suggest it is about $1/0.081 \approx 12$ times more likely that white coin was thrown.

Hypothesis testing.

Example 4.

h	0	1	2	3	4	5
$\frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}}$	165.38	70.88	30.38	13.02	5.58	2.39
h	6	7	8	9	10	
$\frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}}$	1.02	0.44	0.19	0.08	0.03	

Table: Likelihood of Binomial(10, 0.5) vs. Binomial(10, 0.7) (rounded to 3 decimals).

Large values of so-called *likelihood ratio*

$$\frac{\mathbb{P}_b\{H = h\}}{\mathbb{P}_w\{H = h\}}$$

support

null hypothesis H_0 : black coin was tossed (i.e. $p = 0.5$),

whereas small values support

alternative H_A : white coin was tossed (i.e. $p = 0.7$).

Hypothesis testing.

Example 4.

h	0	1	2	3	4	5
$\frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}}$	165.38	70.88	30.38	13.02	5.58	2.39
h	6	7	8	9	10	
$\frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}}$	1.02	0.44	0.19	0.08	0.03	

Table: Likelihood of Binomial(10, 0.5) vs. Binomial(10, 0.7) (rounded to 3 decimals).

The value $k = 6$ is critical in that

$$\begin{cases} \frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}} > 1 & \text{for } h \leq k \\ \frac{\mathbb{P}_b\{H=h\}}{\mathbb{P}_w\{H=h\}} < 1 & \text{for } h > k, \end{cases}$$

in other words, observing $h \leq 6$ (just) suggests that it is more likely that the black coin was tossed.

Hypothesis testing.

Example 4. Taking rejection region $C = \{7, 8, 9, 10\}$ yields a significance level

$$\alpha = \mathbb{P}\{\text{type I error}\} = \mathbb{P}_b\{H > 6\} = 0.18$$

with type II error probability

$$\beta = \mathbb{P}\{\text{type II error}\} = \mathbb{P}_w\{H \leq 6\} = 0.35.$$

If we are not willing to risk rejecting H_0 when, in fact, the black coin was tossed, with probability 0.18, but we are willing to risk this error with probability 0.01, i.e. $\alpha = 0.01$, we need to shrink the rejection region.

Setting rejection region to $C = \{9, 10\}$ yields

$$\alpha = \mathbb{P}_b\{H \geq 9\} = 0.01,$$

and type II error probability

$$\beta = \mathbb{P}_w\{H \leq 8\} = 0.85.$$

Hypothesis testing.

Example 4. More generally, for rejection region $C := \{k, \dots, 10\}$ we obtain

$$\alpha = \mathbb{P}\{\text{type I error}\} = \mathbb{P}_b\{k, \dots, 10\} = \sum_{i=k}^{10} \binom{10}{i} \left(\frac{1}{2}\right)^{10},$$

and

$$\beta = \mathbb{P}\{\text{type II error}\} = \mathbb{P}_w\{1, \dots, k-1\} = \sum_{i=1}^{k-1} \binom{10}{i} (0.7)^i (0.3)^{10-i}.$$

This calculation shows how reducing the type I error probability (by choosing a larger value for k) is at the cost of increasing the type II error probability, and vice versa.

Hypothesis testing.

Likelihood ratio test for simple hypotheses. The previous example is an instance of the so-called *likelihood ratio test*. In general, for any two simple hypotheses

$$H_0: \theta = \theta_0 \quad H_A: \theta = \theta_A$$

the *likelihood ratio (LR)* is defined to be

$$\Lambda := \frac{\text{lik}(\theta_0)}{\text{lik}(\theta_A)} = \frac{f(x_1, \dots, x_n | \theta_0)}{f(x_1, \dots, x_n | \theta_A)}.$$

The corresponding test with decision rule

$$\begin{cases} \text{reject } H_0 & \text{if } \Lambda < K \\ \text{accept } H_0 & \text{otherwise} \end{cases}$$

for some constant K is called a the *likelihood ratio test (LRT)*. As in the previous example, large values of Λ suggest that data support H_0 , whereas small values suggest that data support H_A over H_0 .

Hypothesis testing.

In general, even if we fix a significance level α , there can be a number of hypothesis tests with this level. Ideally, among all tests with level α we'd like to find one (there might be several) that minimizes the type II error probability, respectively maximizes power.

In the setting where both the null hypothesis and the alternative are simple, the next theorem shows that the optimal test (in the sense above) is the likelihood ratio test.

Hypothesis testing.

Lemma (Neyman-Pearson lemma)

Consider simple hypotheses H_0 and H_A and the corresponding likelihood ratio test with significance level α and power $1 - \beta$. Then, any other test with significance level at most α has power less than or equal to $1 - \beta$.

(without proof)

Hypothesis testing.

Example 5. Consider random sample X_1, \dots, X_n
s.t. $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ with known variance σ^2 . The hypotheses are

$$H_0: \mu = \mu_0 \quad H_A: \mu = \mu_A$$

for some given $\mu_0 > \mu_A$. Likelihood ratio

$$\Lambda = \frac{\text{lik}(\theta_0)}{\text{lik}(\theta_A)} = \frac{e^{-\frac{1}{2} \sum_{i=1}^n (X_i - \mu_0)^2 / \sigma^2}}{e^{-\frac{1}{2} \sum_{i=1}^n (X_i - \mu_A)^2 / \sigma^2}} = e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n [(X_i - \mu_A)^2 - (X_i - \mu_0)^2]}.$$

Here and in general, the likelihood ratio is a complicated statistic.
However, one can often find a simpler statistic which determines
the LR.

Hypothesis testing.

Example 5. Since

$$\begin{aligned}\sum_{i=1}^n [(X_i - \mu_A)^2 - (X_i - \mu_0)^2] &= -2\mu_0 n \bar{X}_n + n\mu_0^2 + 2\mu_A n \bar{X}_n - n\mu_A^2 \\ &= 2n\bar{X}_n(\mu_0 - \mu_A) + n\mu_0^2 - n\mu_A^2,\end{aligned}$$

the likelihood ratio is small for large values of \bar{X}_n , i.e. LRT rejects H_0 if

$$\Lambda < K \quad \text{or, equivalently, if} \quad \bar{X}_n > K'$$

for some values of K, K' .

In other words, instead of the LR we might as well use the sample mean \bar{X}_n to construct the decision rule (replacing the constant K by K'). This allows us to work out K , since we know the sampling distribution of \bar{X}_n .

Hypothesis testing.

To proceed, let us fix a significance level, say $\alpha = 0.01$. In order for the LRT to have significance level α , we have to solve

$$\begin{aligned} 0.01 &= \alpha = \mathbb{P}\{\text{reject } H_0 | H_0\} = \mathbb{P}\{\bar{X}_n > K' | H_0\} \\ &= \mathbb{P}\left\{\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} > \frac{K' - \mu_0}{\sigma/\sqrt{n}} | H_0\right\} = 1 - \Phi\left(\frac{K' - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

for K' , i.e. $K' = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.99) + \mu_0$, where $\Phi^{-1}(x)$ denotes the quantile function of the standard normal distribution.¹ For the power, we find

$$\begin{aligned} 1 - \beta &= \mathbb{P}\{\text{reject } H_0 | \mu = \mu_A\} = \mathbb{P}\{\bar{X}_n > K' | \mu = \mu_A\} \\ &= 1 - \Phi\left(\frac{K' - \mu_A}{\sigma/\sqrt{n}}\right) \end{aligned}$$

According to the Neyman-Pearson lemma, there is no other test with significance level $\leq \alpha = 0.01$ that has a power greater than $1 - \Phi\left(\frac{K' - \mu_A}{\sigma/\sqrt{n}}\right)$.

¹The quantile function of the standard normal distribution is sometimes called the *probit* function. The corresponding command in R is qnorm.

STAT 135, Concepts of Statistics

Helmut Pitters

Hypothesis testing 3

Department of Statistics
University of California, Berkeley

March 14, 2017

Hypothesis testing.

Uniformly most powerful test. Neyman-Pearson lemma gives most powerful test provided both hypotheses are simple. In general not possible to find most powerful test for composite hypotheses. However, if null is simple, can extend theory to so-called uniformly most powerful tests.

Definition (Uniformly most powerful test)

A test

$$H_0: \theta = \theta_0 \quad H_A: \theta \in \Theta_A$$

with simple null is called *uniformly most powerful (UMP)* if it is most powerful for every simple alternative in H_A , i.e. if for any $\theta_a \in \Theta_A$ it is most powerful for

$$H_0: \theta = \theta_0 \quad H_A: \theta = \theta_a.$$

Hypothesis testing.

Generalized likelihood ratio test. If hypotheses are composite

$$H_0: \theta \in \Theta_0 \quad H_A: \theta \in \Theta_A$$

cannot expect to find most powerful test. However, we can still compare likelihoods of null and alternative, both evaluated at their maximal value. This leads to what is called the generalized likelihood ratio tests which are similarly important to testing as MLEs are in estimation.

Definition (Generalized likelihood ratio test)

The *generalized likelihood ratio test (GLRT)* rejects H_0 for small values of the *generalized likelihood ratio (GLR)*

$$\Lambda := \frac{\sup_{\theta \in \Theta_0} \text{lik}(\theta)}{\sup_{\theta \in \Theta} \text{lik}(\theta)},$$

where $\Theta = \Theta_0 \cup \Theta_A$.

As before, reject H_0 for small values of Λ .

Hypothesis testing.

Generalized likelihood ratio test. To work out the critical value of the GLRT for fixed level α we would need to know the sampling distribution $(\Lambda|H_0)$ of Λ under H_0 .

This distribution has no simple form in general, but $-2 \log \Lambda$ can be approximated by a chi squared distribution for large sample size.

Fact

Under smoothness conditions on $f(x|\theta)$ the null distribution of

$$-2 \log \Lambda$$

is asymptotically (as $n \rightarrow \infty$) distributed according to a chi squared distribution with

$$\#\text{free parameters in } \Theta - \#\text{free parameters in } \Theta_0$$

degrees of freedom.

Hypothesis testing.

Goodness of fit

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit).
Emergency room of large hospital assigns patients to one of three categories:

1. Stable. No immediate treatment required.
2. Serious. Immediate treatment not required, but patient needs to be monitored until physician available.
3. Critical. Patient's life endangered without immediate treatment.

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit).

Hospital records over past year show that

- ▶ 50% of patients classified stable,
- ▶ 30% of patients classified serious,
- ▶ 20% of patients classified critical,

hence if we took a random sample of n patients and observed

N_i patients in category i

(where $\sum_i N_i = n$) we expect that

$$(N_1, N_2, N_3)$$

follows a multinomial distribution with parameters
(3, 50%, 30%, 20%).

Hypothesis testing.

Review: multinomial distribution. Think of n different marbles that we paint in c different colors. We paint marbles one after the other, with

$$\mathbb{P}\{\text{a particular marble is painted in color } i\} = p_i$$

(with $\sum_i p_i = 1$, $p_i \geq 0$) independent of the colors in which the other marbles are painted. Let

$$X_i := \# \text{ marbles of color } i.$$

Then

$$\mathbb{P}\{X_1 = x_1, \dots, X_c = x_c\} = \binom{n}{x_1, \dots, x_n} \prod_{i=1}^c p_i^{x_i}$$

if $\sum_i x_i = n$ and $= 0$ otherwise. The vector (X_1, \dots, X_n) has a *multinomial distribution* with parameters (n, p_1, \dots, p_c) denoted

$$(X_1, \dots, X_n) \sim \text{multinomial}(n, p_1, \dots, p_c).$$

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit).
Emergency room's good reputation resulted in increased number of patients.

Important question for organization: has increased number of patients also brought about a change in distribution of patients among categories?

I.e. want to test

$$H_0: (p_1, p_2, p_3) = (0.5, 0.3, 0.2) \quad \text{vs.} \quad H_A: (p_1, p_2, p_3) \neq (0.5, 0.3, 0.2).$$

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit).

Record categories of n incoming patients, modeled as independent random draws from 1, 2, 3.

$$O_i := \# \text{ patients in sample of category } i$$

Under null expect

$$E_i := \mathbb{E}[N_i] = np_i$$

patients in category i .¹ Pearson's chi-squared test is based on test statistic

$$\sum_{i=1}^c \frac{(O_i - E_i)^2}{E_i},$$

which can be approximated by a χ_{c-1}^2 distribution if sample size n is large enough.²

¹Recall that the marginal distribution X_i in the multinomial distribution $(X_1, \dots, X_c) \sim \text{multinomial}(c, p_1, \dots, p_c)$ has a $\text{binomial}(n, p_i)$ distribution.

²As a rule of thumb: approximation is good if $O_i \geq 5$.

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit). A random sample of $n = 200$ patients is taken and their observed frequencies are as in the table below

	stable	serious	critical
observed frequencies (O_i)	98	48	54
expected frequencies (E_i)	100	60	40

Table: Frequencies of patients in different categories.

Notice that if $(p_1, p_2, p_3) = (0.5, 0.3, 0.2)$, we expect

$$E_1 := \mathbb{E}[N_1] = 200 \times 0.5 = 100 \text{ patients in category 1}$$

$$E_2 := \mathbb{E}[N_2] = 200 \times 0.3 = 60 \text{ patients in category 2}$$

$$E_3 := \mathbb{E}[N_3] = 200 \times 0.2 = 40 \text{ patients in category 3}$$

since $N_i \sim \text{binomial}(200, p_i)$.

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit).

Heuristically, if there are large differences

$$|O_i - E_i|$$

between observed and expected numbers of patients for some categories, we reject the null. However, if these differences are small, the data do not provide sufficient evidence to reject H_0 . Pearson's chi squared test is based on the test statistic

$$X^2 := \sum_{i=1}^c \frac{(O_i - E_i)^2}{E_i},$$

which is well approximated by a χ^2_{c-1} distribution provided the sample size n is large. Null is rejected for large values of X^2 .

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit).

Define the α -percentile of χ_n^2 by

$$\mathbb{P}\{Y \leq \chi_n^2(\alpha)\} = \alpha$$

for $Y \sim \chi_n^2$.

Fix significance level $\alpha = 0.05$. We reject H_0 if

$$X^2 > \chi_{c-1}^2(1 - \alpha) = \chi_2^2(0.95) = 5.99.^3$$

In our example

$$X^2 = \frac{(98 - 100)^2}{100} + \frac{(48 - 60)^2}{60} + \frac{(54 - 40)^2}{40} = 7.34,$$

thus we reject H_0 on the basis of the data, i.e. there is strong evidence that the distribution of patients among categories has changed.

³Find percentiles of chi squared distribution tabulated or evaluate them via statistical software package.

Hypothesis testing.

Pearson's chi squared test that we saw in the last example can be applied to a broad range of settings.

Many observations in social and physical sciences are not numerical, but can be assigned to different categories. This is called categorial or enumerative data. E.g.

- ▶ brand of motor vehicle on certain highway section
- ▶ classification of documents according to topics
- ▶ classification of animals according to species
- ▶ blood type of a person: A, B, AB, O

Hypothesis testing.

Pearson's chi squared test can be readily generalized to some arbitrary number c of categories.

Consider population with each individual belonging to one of c categories

(e.g. residents of California with blood types A, B, AB, O), and let

$$p_i := \# \text{ relative frequency of items in category } i.$$

Draw random sample of size n .

$$N_i := \# \text{ items of category } i \text{ in sample.}$$

$$(N_1, \dots, N_c) \sim \text{multinomial}(n, p_1, \dots, p_c).$$

The hypotheses

$$H_0: (p_1, \dots, p_c) = (\pi_1, \dots, \pi_c) \quad H_A: (p_1, \dots, p_c) \neq (\pi_1, \dots, \pi_c)$$

(for some $\pi_i \geq 0$ s.t. $\sum_i \pi_i = 1$) can then be tested via Pearson's chi squared test.

Hypothesis testing.

A note on Rémy's algorithm for generating random binary trees

Erkki Mäkinen

em@cs.uta.fi

Dept. of Computer and Information Sciences,
P.O. Box 607, FIN-33014 University of Tampere, Finland

Jarmo Siltaneva

Jarmo.Siltaneva@tt.tampere.fi

Information Technology Center, City of Tampere,
Lenkkielijäankatu 8, Finn-Medi 2, FIN-33520 Tampere, Finland

Abstract. This note discusses the implementation of Rémy's algorithm for generating unbiased random binary trees. We point out an error in a published implementation of the algorithm. The error is found by using the χ^2 -test. Moreover, a correct implementation of the algorithm is presented.

Hypothesis testing.

Neyman-Pearson paradigm somewhat rigid in that it requires to either reject or not reject the null hypothesis. Instead, might be interested in measuring strength of evidence against H_0 .

Hypothesis testing.

Example: $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$ i.i.d. Test

$$H_0: \mu = 0 \quad \text{vs.} \quad H_A: \mu = 1$$

with decision rule:

reject H_0 if $\bar{X}_n > c(\alpha)$, yielding level α test.

On observing $\bar{X}_n = x$ we define

$$p^* := \mathbb{P}\{\bar{X}_n > x | H_0\}.$$

Decision rule can now be equivalently written as

$$\begin{cases} \text{reject } H_0 & \text{if } p^* \leq \alpha \\ \text{do not reject } H_0 & \text{otherwise} \end{cases}$$

that is p^* contains all the information we need about the sample to make the decision.

Hypothesis testing.

Definition (p-value)

The p-value is the smallest value of α for which the null hypothesis is rejected.

Remark

Suppose $T = T(X_1, \dots, X_n)$ is continuous test statistic and rejection region is of the form

$$\{T > t\} \quad (\text{for some } \textit{critical value } t).$$

Can interpret p-value

$$\mathbb{P}\{T(X_1, \dots, X_n) > T(x_1, \dots, x_n) | H_0\}$$

as probability under the null hypothesis of observing a value of the test statistic as or more extreme than the observation $T(x_1, \dots, x_n)$.⁴

⁴Likewise for rejection region $\{T < t\}$.

Hypothesis testing.

Example 6 (Pearson's chi squared test for goodness of fit).
For the p-value we find

$$\mathbb{P}\left\{X^2 \geq 7.34\right\} = 0.025,$$

that is under the null hypothesis we would observe a value greater or equal to 7.34 in about 3 of 100 cases. In other words, the data suggest that H_0 is not very likely.

STAT 135, Concepts of Statistics

Helmut Pitters

Hypothesis testing 4

Department of Statistics
University of California, Berkeley

March 16, 2017

Hypothesis testing.

Likelihood ratio test for multinomial distribution. Last example: Pearson's chi squared test for goodness of fit of a multinomial distribution.

Now: LRT for multinomial distribution.

Let

$$M := \{(p_1, \dots, p_m) : p_i \geq 0, \sum_i p_i = 1\}$$

denote the set of all possible probabilities for a multinomial distribution with m cells. Consider hypotheses¹

$$H_0: p \in M_0 := \{(p_1(\theta), \dots, p_m(\theta)) \in M : \theta \in \Theta_0\},$$

vs.

$$H_A: (p_1, \dots, p_m) \in M$$

such that $(p_1, \dots, p_m) \neq (p_1(\theta), \dots, p_m(\theta))$ for all $\theta \in \Theta_0$.

¹Notice that the cell probabilities p_i depend on some parameter θ .

Hypothesis testing.

Likelihood ratio test for multinomial distribution. For random sample of size n LR is found to be

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} \text{lik}(p(\theta))}{\sup_{p \in M} \text{lik}(p)} = \prod_{i=1}^m \left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right)^{x_i}, \quad \text{lik}(p) = \binom{n}{x_1, \dots, x_m} \prod_i p_i^{x_i},$$

where $\hat{p}_i = x_i/n$ and $\hat{\theta}$ is MLE for θ under restriction $\theta \in \Theta_0$.²

Therefore obtain

$$-2 \log \Lambda = -2n \sum_{i=1}^m \hat{p}_i \log \left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right) = 2 \sum_{i=1}^m O_i \log \frac{O_i}{E_i}$$

where

$$O_i = n\hat{p}_i = \# \text{ observed counts in category } i,$$

$$E_i = np_i(\hat{\theta}) = \# \text{ expected counts in category } i.$$

One can show that as sample size n grows without bounds

$-2 \log \Lambda$ and Pearson's chi squared statistics

are asymptotically equivalent under H_0 .

²IAPT suggests a connection between testing a hypothesis and estimating a

Hypothesis testing.

Example: Mendel's pea experiment. Gregor Mendel supported his theory of inheritance with his famous pea experiment, in which he crossed round yellow seeds with wrinkled green seeds. According to his theory, the relative frequencies of peas of different types are

round yellow	round green	wrinkled yellow	wrinkled green
$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$

Table: Relative frequencies of types of beans.

In $n = 556$ trials Mendel observed

$$O = (315, 101, 108, 32) \quad \text{peas of each type.}$$

Under

$$H_0: p = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right),$$

the hypothesis claimed by Mendel's theory, we'd expect

$$E = 556 \times \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right) = (312.75, 104.25, 104.25, 34.75)$$

peas of each type.

Hypothesis testing.

Example: Mendel's pea experiment. Evaluating Pearson's chi squared statistic yields

$$X^2 = \sum_{i \in \{\text{type of peas}\}} \frac{(O_i - E_i)^2}{E_i} = \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} \\ + \frac{(108 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} \approx 0.604.$$

Since large values of X^2 contradict H_0 , the p-value is

$$\mathbb{P}\{Y \geq 0.604\} = 0.896$$

where $Y \sim \chi_3^2$ has chi squared distribution with 3 df (= 4 categories of peas - 1).

That is in about 90% of cases we see for X^2 a value of 0.604 or larger. This result is in very good agreement with H_0 (and hence Mendel's theory) — too good to be true?

Hypothesis testing.

Example: Mendel's pea experiment. Notice that with Λ denoting the likelihood ratio, we obtain

$$-2 \log \Lambda = 2 \sum_{i \in \{\text{type of peas}\}} O_i \log \frac{O_i}{E_i} \approx 0.618$$

which is very close to the value of Pearson's chi squared statistic X^2 .

Hypothesis testing.

“Recipe” for testing hypotheses.

1. Specify null hypothesis H_0 and alternative hypothesis H_A .
2. Specify test statistic T that discriminates between H_0 and H_A .
3. Specify “extreme” values of T under H_0 in direction of H_A , suggesting that H_A better explains data than H_0 .
E.g. reject H_0 for
 - ▶ large values of T , or
 - ▶ small values of T , or
 - ▶ large values of $|T - t|$ for some t ,
 - ▶ etc.
4. If significance level α given, can work out rejection region if null distribution (of T) is known or can be approximated.
5. Evaluate T from data and
 - ▶ decide whether to reject H_0 , or
 - ▶ work out p-value.

Hypothesis testing.

Duality of confidence intervals and hypothesis tests. Next theorem shows that procedures of testing a hypothesis and estimating a parameter are in fact intimately related. To see this, have to slightly generalize the notion of confidence interval to a confidence set.

Recall that $X = (X_1, \dots, X_n) \sim \mathbb{P}_\theta$ denotes vector from which the sample is drawn whose joint distribution depends on some (unknown) parameter $\theta \in \Theta$.³ Let's assume Θ is a subset of \mathbb{R} .

Definition (Confidence set)

A (random) subset $C(X)$ of \mathbb{R} is called an α *confidence set for θ* if

$$\mathbb{P}\{\theta \in C(X)\} = \alpha.$$

³Think of specific examples as the ones we saw in the introduction of the terminology in the Neyman-Pearson paradigm.

Hypothesis testing.

Duality of confidence intervals and hypothesis tests.

Theorem

(I) Suppose that for each $\theta_0 \in \Theta$ there exists a level α test of

$$H_0: \theta = \theta_0$$

with acceptance region $A(\theta_0)$, i.e.

test rejects H_0 if $X \notin A(\theta_0)$.

Then

$$C(X) := \{\theta : X \in A(\theta)\}$$

is a $1 - \alpha$ confidence set for θ .

Proof. By definition of $C(X)$

$$\mathbb{P}\{\theta_0 \in C(X) | \theta = \theta_0\} = \mathbb{P}\{X \in A(\theta_0) | \theta = \theta_0\} = 1 - \alpha.$$

Hypothesis testing.

Duality of confidence intervals and hypothesis tests.

Theorem

(II) In the other direction, start with $1 - \alpha$ confidence set $C(X)$ for θ . A level α test for

$$H_0: \theta = \theta_0$$

is given by decision rule

$$\text{reject } H_0 \text{ if } X \notin A(\theta_0) := \{x: \theta_0 \in C(x)\}$$

Proof. By definition of $A(\theta_0)$

$$\mathbb{P}\{X \notin A(\theta_0) | \theta = \theta_0\} = \mathbb{P}\{\theta_0 \notin C(X) | \theta = \theta_0\} = \alpha.$$

STAT 135, Concepts of Statistics

Helmut Pitters

Goodness of fit: further techniques.

Department of Statistics
University of California, Berkeley

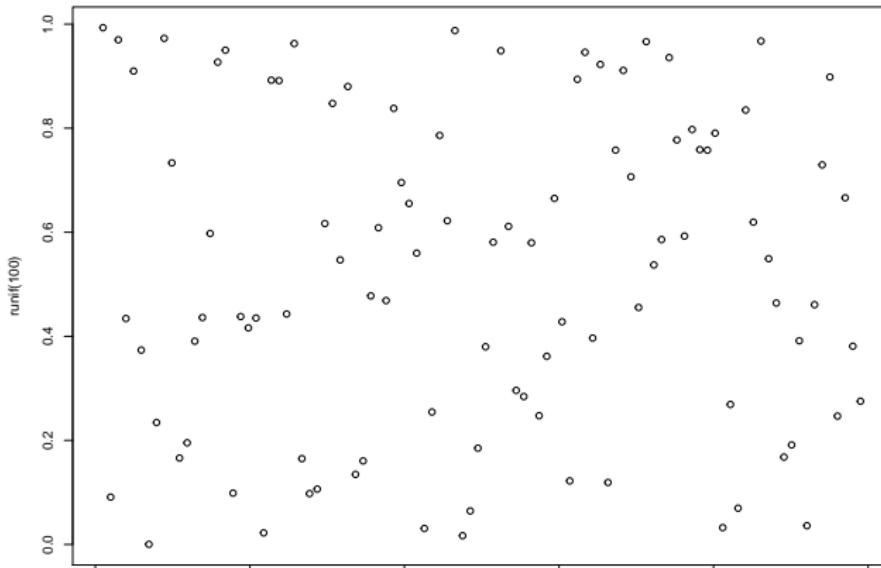
March 29, 2016

Goodness of fit: Probability plots.

Consider independent random samples X_1, \dots, X_n that we conjecture to have uniform[0, 1] distribution.

We are interested in a graphical method that allows to check qualitatively whether our conjecture is at all reasonable.

Figure shows a plot of the pairs $(k/100, X_k)$ for $1 \leq k \leq 100$.



Goodness of fit: Probability plots.

Order the X_1, \dots, X_n in increasing order to obtain order statistics

$$X_{(1)} < \dots < X_{(n)}.^1$$

Recall from Stat 134: $X_{(k)} \sim \text{beta}(k, n - k + 1)$, in particular

$$\mathbb{E}X_{(k)} = \frac{k}{n+1}.$$

The points

$$\left(\frac{k}{n+1}, X_{(k)}\right) \quad (1 \leq k \leq 100)$$

should be spread close to their averages

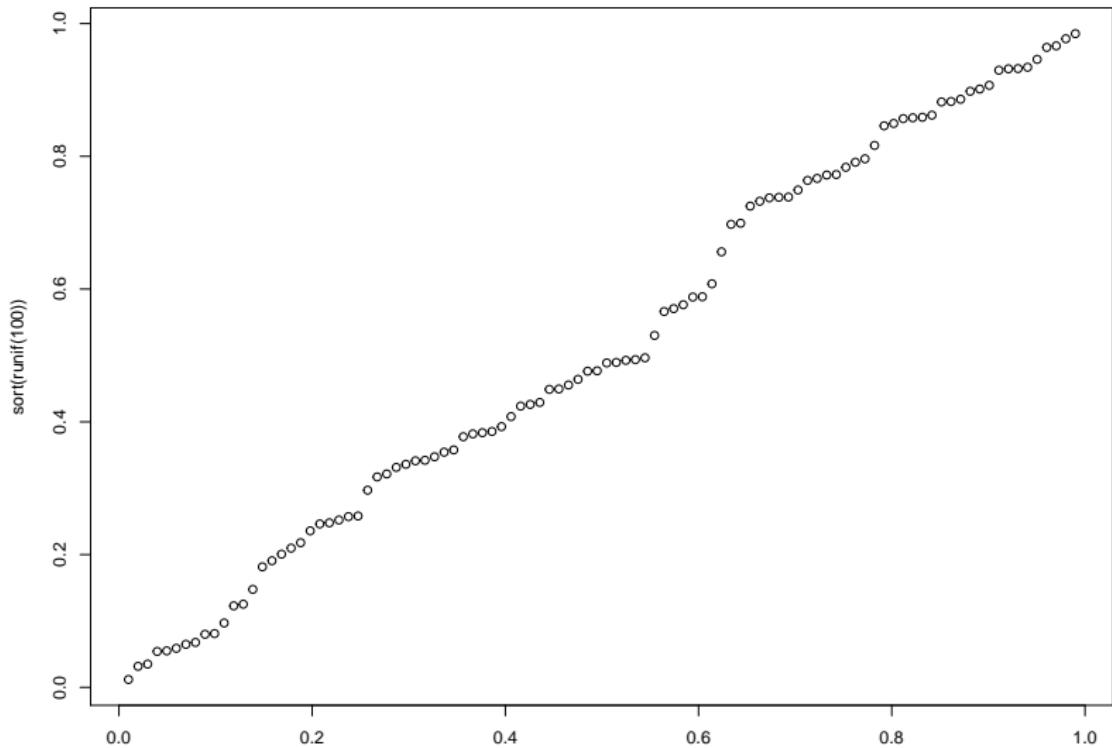
$$\left(\frac{k}{n+1}, \mathbb{E}X_{(k)}\right) = \left(\frac{k}{n+1}, \frac{k}{n+1}\right)$$

which lie on the straight line through $(0, 0)$ with slope 1.

¹In particular $X_{(1)} = \min(X_1, \dots, X_n)$, $X_{(n)} = \max(X_1, \dots, X_n)$.

Goodness of fit: Probability plots.

Figure shows a plot of the pairs $(k/101, X_{(k)})$ for $1 \leq k \leq 100$.



Goodness of fit: Probability plots.

What if the common distribution of X_1, \dots, X_n is not uniform $[0, 1]$?

A simple observation allows us to extend this graphical method to general distributions on \mathbb{R} .

Let X denote a real r.v. with *cumulative distribution function (cdf)*

$$F(t) := \mathbb{P}\{X \leq t\} \quad (t \in \mathbb{R}).$$

Goodness of fit: Probability plots.

Lemma

If X is continuous and F is strictly increasing, then $F(X)$ is a real random variable with distribution uniform[0, 1].

Proof.

Notice first that $0 \leq F(t) \leq 1$ for all $t \in \mathbb{R}$. Moreover, for $0 \leq u \leq 1$

$$\mathbb{P}\{F(X) \leq u\} = \mathbb{P}\{X \leq F^{-1}(u)\} = F(F^{-1}(u)) = u,$$

where F^{-1} denotes the right inverse of F . □

Goodness of fit: Probability plots.

Back to our problem: random sample X_1, \dots, X_n , and we conjecture that their common cdf is F .

Provided our conjecture is true, the $F(X_1), \dots, F(X_n)$ are i.i.d. uniform[0, 1]. Consequently, the points

$$\left(\frac{k}{n+1}, F(X_{(k)}) \right) \quad (1 \leq k \leq n)$$

should be spread close to a linear function.

Alternatively, can plot

$$\left(F^{-1}\left(\frac{k}{n+1}\right), X_{(k)} \right) \quad (1 \leq k \leq n)$$

where F^{-1} denotes the right inverse of F .

Goodness of fit: Probability plots.

Probability plots of theoretical distributions against each other

- ▶ uniform-uniform
- ▶ uniform-normal
- ▶ normal-normal
- ▶ normal-uniform

Probability plots of data against theoretical distributions

- ▶ percentage of manganese in iron (data: manganese.txt)
- ▶ strength of Kevlar/epoxy, a material used in space shuttle (kevlar.txt)
- ▶ Michelson's measurements of light speed (michelson.txt)
- ▶ precipitation in Illinois storms (illinois60.txt, ..., illinois64.txt)

STAT 135, Concepts of Statistics

Helmut Pitters

Summarizing data 2 & nonparametric bootstrap

Department of Statistics
University of California, Berkeley

March 21, 2017

Summarizing data.

In what follows we consider data x_1, \dots, x_n . The data are not necessarily considered as observations of an underlying probabilistic model.

So far encountered

- ▶ measures of location
(trimmed) mean, median, mode
- ▶ measures of spread
percentiles, range, variance
- ▶ and some graphical methods.

We turn to some more detailed summaries.

Tukey's 5-number summary.

A *5-number summary* of a data set (of real numbers) consists of the

1. minimal value, $x_{(1)}$,
2. first quartile (25th percentile), sometimes called the *lower hinge*,
3. median,
4. third quartile (75th percentile), sometimes called the *upper hinge*, and
5. maximal value, $x_{(n)}$.

Example (Heat of sublimation for Iridium and Rhodium)

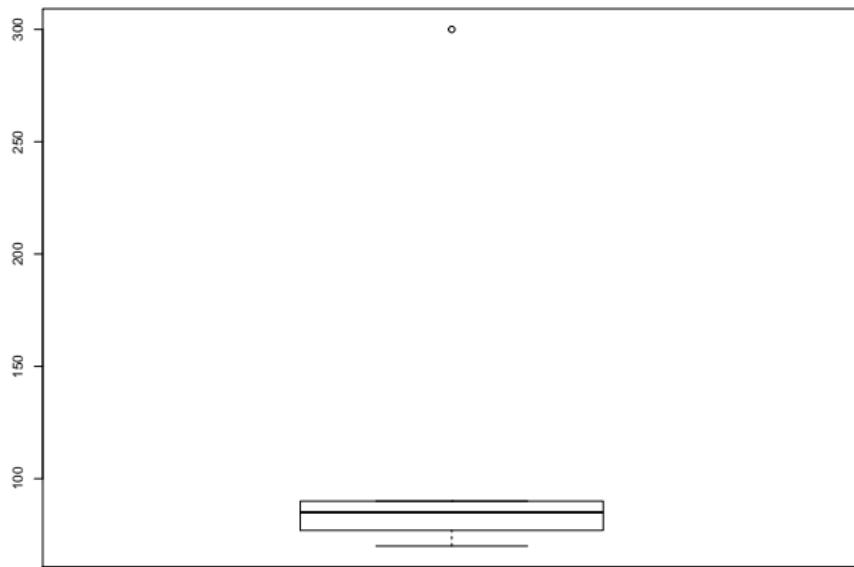
The 5-number summaries for the data sets on heat of sublimation for

1. Iridium: 136.60, 159.50, 159.80, 160.25, 173.90
2. Rhodium: 126.40, 131.45, 132.65, 133.30, 135.70

Box-and-Whisker plot.

A Box-and-Whisker plot is essentially an easy-to-read graphical display of a 5-number summary.

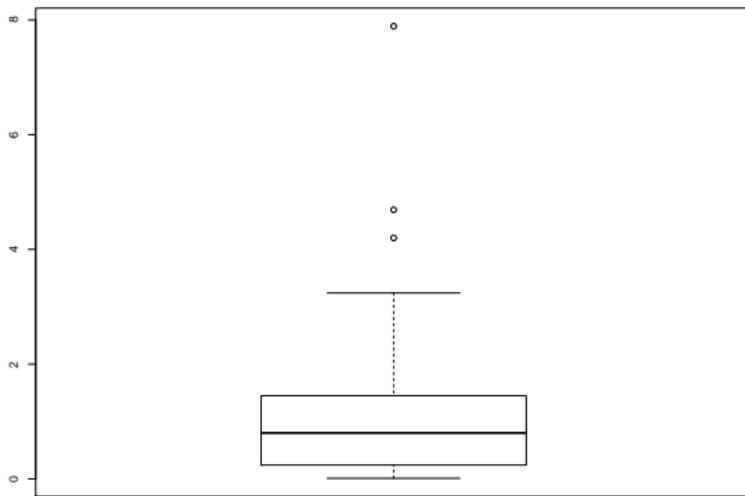
Below is a boxplot of the data on family income.



Box-and-Whisker plot.

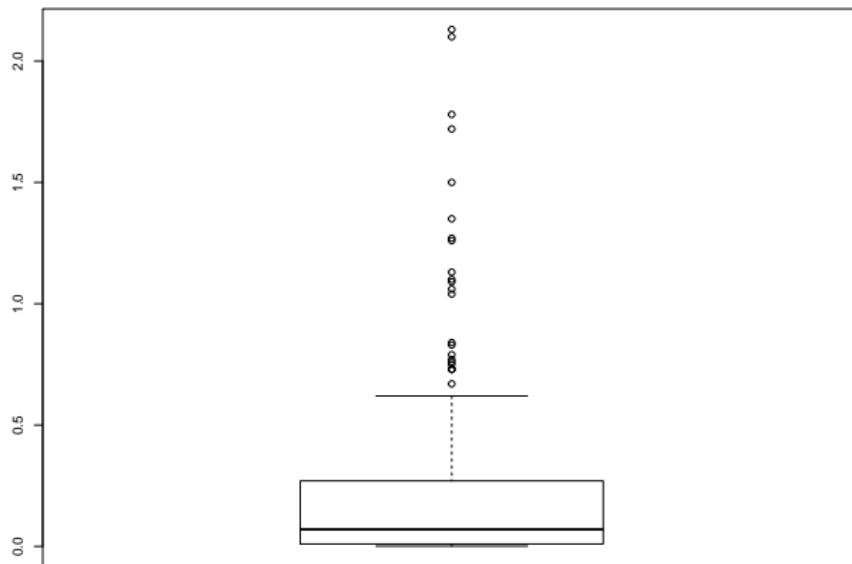
A boxplot of the times to failure [hours] data of 76 strands of Kevlar material used in space shuttles.

Horizontal bars in the box indicate the lower hinge, median and upper hinge. The two remaining bars indicate the most extreme data points within distance of 1.5 of the upper/lower quartile. Dots indicate outliers.



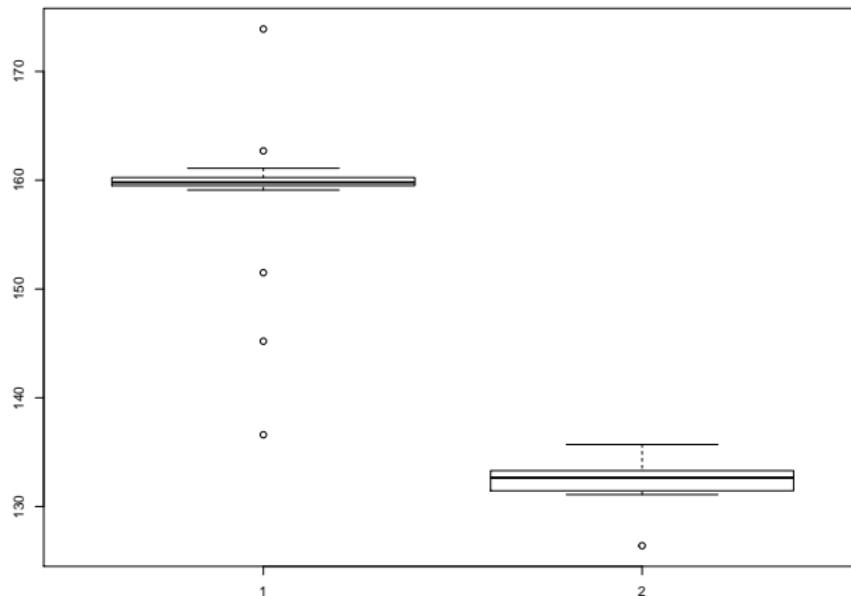
Box-and-Whisker plot.

A boxplot of the precipitation data for storms in Illinois.



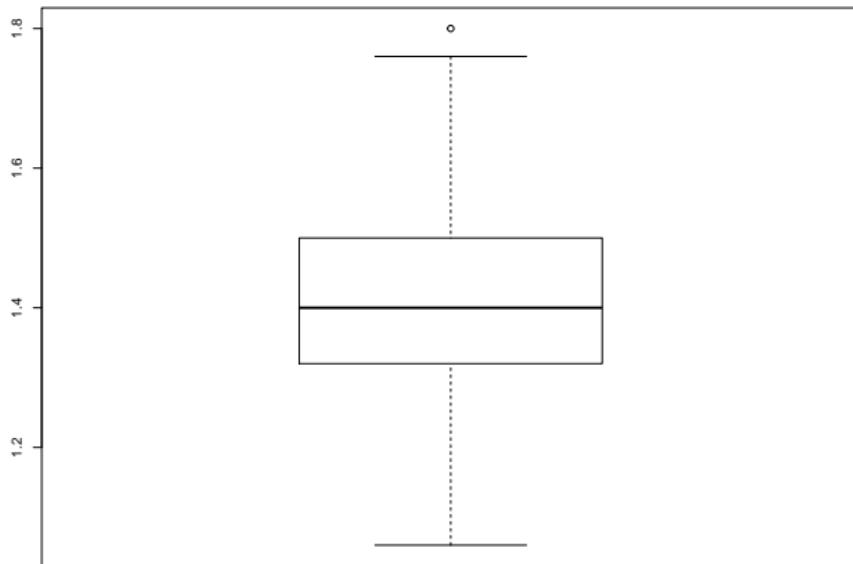
Box-and-Whisker plot.

A boxplot of the heat of sublimation for iridium (left) and rhodium (right).



Box-and-Whisker plot.

A boxplot of the heat of the percentage of manganese in iron.



Empirical cdf.

Suppose data $x = (x_1, \dots, x_n)$ are given by real numbers.

Definition (Empirical cdf)

The *empirical cumulative distribution function* (*ecdf*)¹ is defined by

$$F_n(x) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(-\infty, x]}(x_i) = \frac{1}{n} \#\{i : x_i \leq x\}$$

= relative frequency of data items with values $\leq x$.

¹In R use command `ecdf(x)` to compute the ecdf.

Empirical cdf: Family incomes

Example (Family incomes)

Incomes of five families sampled randomly from Berkeley's population² are

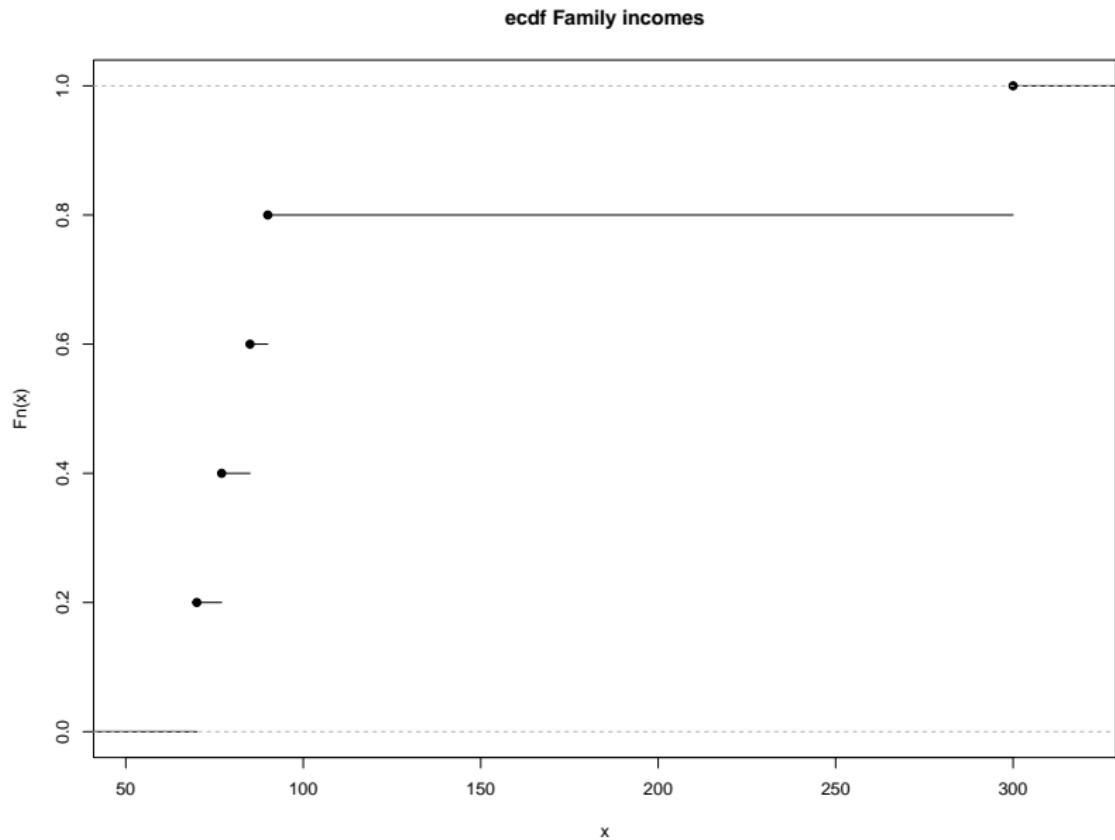
\$90k \$70k \$77k \$85k \$300k.

Average

$$\bar{x} = 124.4$$

²<http://www.city-data.com/income/income-Berkeley-California.html>

Empirical cdf: Family incomes



Empirical cdf.

Notice that the ecdf $F_n(x) = \frac{1}{n} \#\{i: x_i \leq x\}$

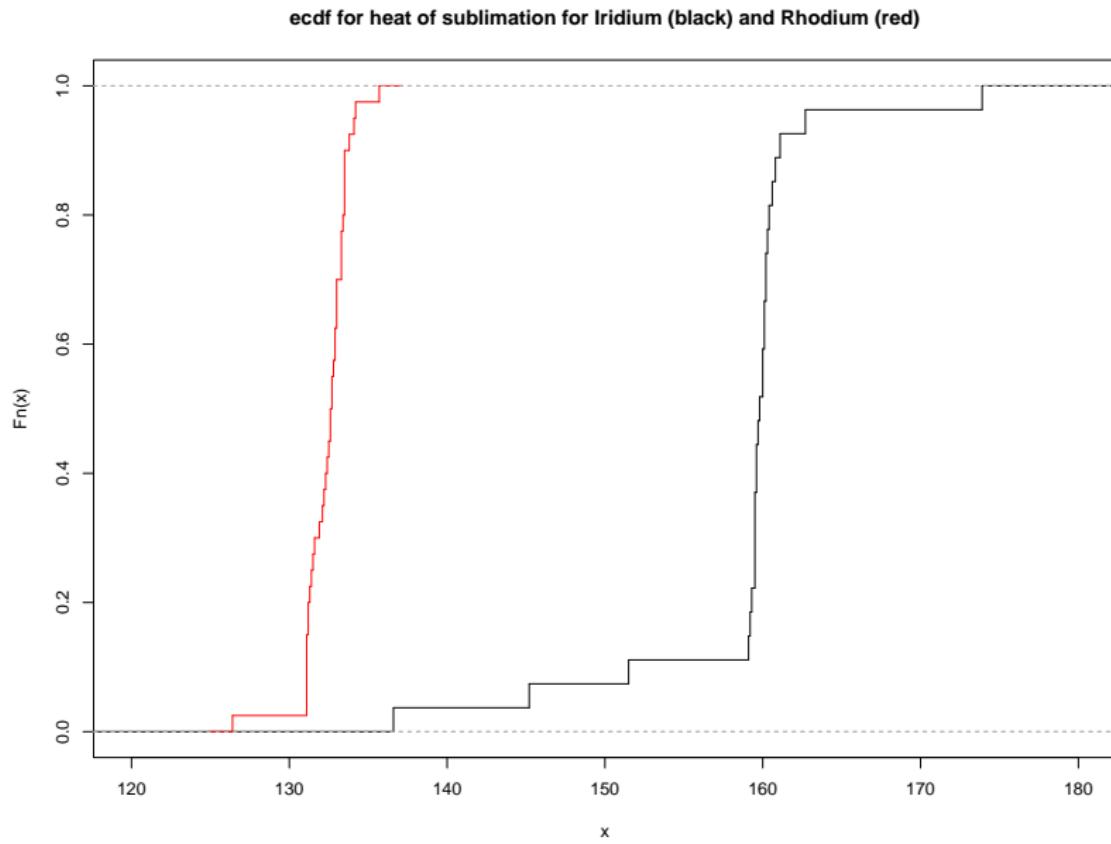
$\begin{cases} \text{has a jump at } x \text{ if there is an observation with value } x, \text{ and} \\ \text{the jump is of size } \frac{i}{n} \text{ if there are } i \text{ observations with value } x \end{cases}$

Can conveniently read off percentiles from the ecdf.

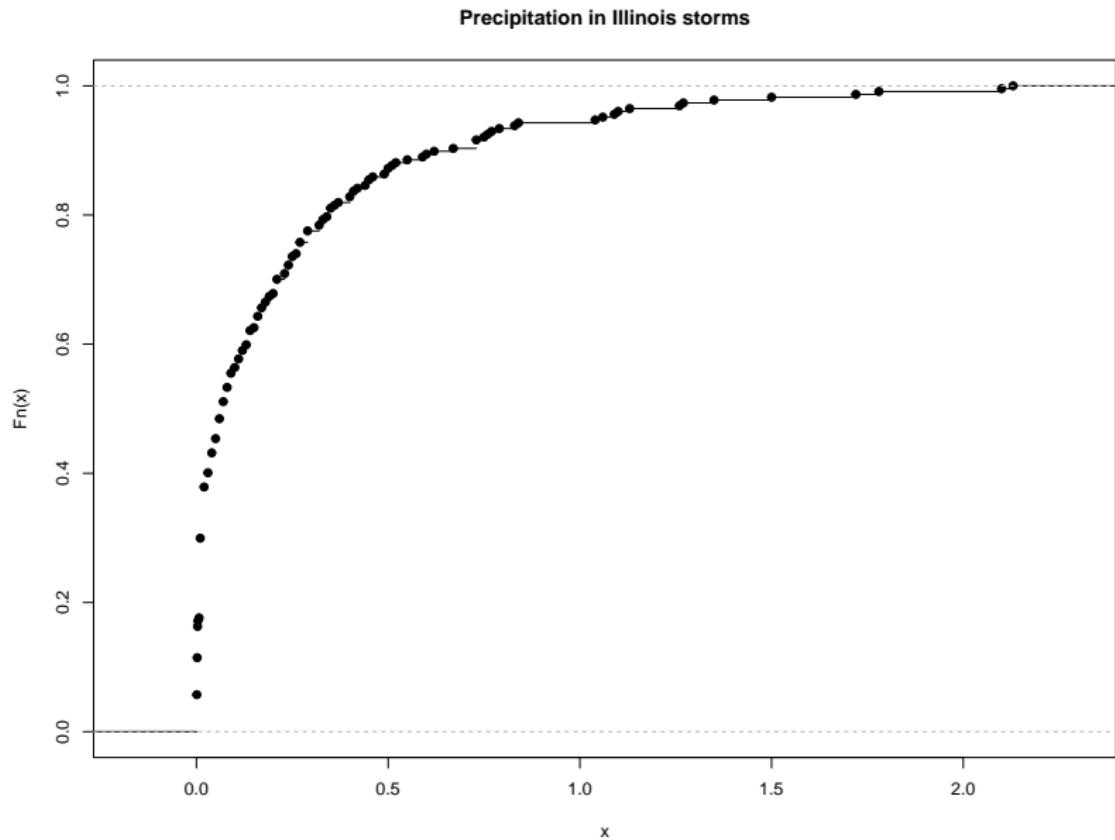
The median is close to the value x such that $F_n(x) = \frac{1}{2}$.

The p th percentile is close to the value x such that $F_n(x) = \frac{p}{100}$.

Empirical cdf. Heat of sublimation for Iridium and Rhodium.



Empirical cdf. Precipitation in Illinois storms.



Nonparametric bootstrap.

We first recall the **parametric bootstrap** method.

Suppose we are interested in a measure of location (e.g. mean, median, trimmed mean) of the data x_1, \dots, x_n . Data is assumed to be independent random sample from some (unknown) distribution with cdf F_θ .

(E.g. we assumed F_θ to be a $\text{gamma}(\alpha, \lambda)$ cdf in the precipitation data in storms in Illinois.)

Nonparametric bootstrap.

An estimator, $\hat{\theta}$ say, for the measure of location θ is usually straightforward to compute from x_1, \dots, x_n . In particular, $\hat{\theta}$ is a function of X_1, \dots, X_n and therefore random. In the parametric bootstrap approximated the distribution of $\hat{\theta}$ by bootstrap simulations as follows.

1. Use $F_{\hat{\theta}}$ as an approximation for F_{θ} .

(Since we don't know the 'true' cdf F_{θ} from which the data is sampled.)

2. Draw n independent samples x_1^*, \dots, x_n^* from $F_{\hat{\theta}}$

Compute $\theta = \theta(x_1^*, \dots, x_n^*)$ and denote this value by θ_1^* .

Repeat B times to obtain $\theta_1^*, \theta_2^*, \dots, \theta_B^*$

take empirical distribution of $\theta_1^*, \theta_2^*, \dots, \theta_B^*$ as a good approximation to distribution of $\hat{\theta}$.

The empirical distribution of $\theta_1^*, \theta_2^*, \dots, \theta_B^*$ approximates distribution of $\hat{\theta}$ better and better as we increase B .

Nonparametric bootstrap.

Nonparametric bootstrap. How does the nonparametric bootstrap differ from the parametric bootstrap?

In the nonparametric bootstrap we do *not* assume that the true cdf F from which the data x_1, \dots, x_n are sampled is contained in a parametric model F_θ for some θ .

Instead, in the nonparametric bootstrap one uses the empirical cdf

$$F_n(x) = \frac{1}{n} \#\{i : x_i \leq x\}$$

as an approximation for F .

Nonparametric bootstrap.

Consequently, the distribution of some estimator $\hat{\theta}$ is approximated in the nonparametric bootstrap as follows.

1. Use F_n as an approximation for F .

(Since we don't know the 'true' cdf F from which the data is sampled.)

2. Draw n independent samples x_1^*, \dots, x_n^* from F_n .

Compute $\theta = \theta(x_1^*, \dots, x_n^*)$ and denote this value by θ_1^* .

Repeat B times to obtain $\theta_1^*, \theta_2^*, \dots, \theta_B^*$.

Take empirical distribution of $\theta_1^*, \theta_2^*, \dots, \theta_B^*$ as a good approximation to the distribution of $\hat{\theta}$.

Again, the approximation of the distribution of $\hat{\theta}$ by empirical distribution of $\theta_1^*, \theta_2^*, \dots, \theta_B^*$ becomes better as we increase B .

Nonparametric bootstrap.

Example. Heat of sublimation for Iridium/Rhodium.

Bootstrapping mean and median.

- ▶ Which of the two estimators, for mean or median, do you expect to have a larger spread/standard error?
- ▶ Is the sampling distribution for the mean more spread-out for the Iridium or the Rhodium data?

R: bootstrap, histograms of sampling distributions.

STAT 135, Concepts of Statistics

Helmut Pitters

Comparing two populations - independent samples

Department of Statistics
University of California, Berkeley

April 12, 2017

Comparing two populations.

New treatments¹ are proposed continually:

- ▶ treatments for breast cancer, multiple sclerosis
- ▶ drugs,
- ▶ surgical techniques,
- ▶ medicine for pain relief,
- ▶ fertilizers,
- ▶ teaching methods,
- ▶ etc.

as well as new formulas for

- ▶ making bread,
- ▶ detergents,
- ▶ alloys,
- ▶ etc.

to improve some aspect of live (increase productivity/profit, reduce pain, decrease waiting time, enhance taste, etc.)

¹Treatment being understood in a very broad sense of the word.

Comparing two populations.

Before spending money on a new treatment we surely want to convince ourselves of its efficacy.



Want to assess *effects of treatment* by comparing population of *responses* of treated individuals (*treatment group*) with population of untreated individuals (*control group*). Usually, can only compare samples of different populations (examining entire population is unfeasible).

Comparing two populations.

Remark

Won't consider categorial data (e.g. little, moderate, complete relief from pain).

Numerical observations (e.g. blood pressure, yield, waiting time) will differ from individual to individual. Treatment may increase some responses and decrease others. However, we are interested in the *average response*.

Comparing two populations.

Controlled experiments vs. observational studies.

Experiments are studies where a researcher assigns treatments to cases. When this assignment is done in a random fashion, it is called a *randomized experiment*.

A study in which the researcher collects data without interfering with how the data arise is called an *observational study*. From observational studies one can possibly infer association, but not causation.

Not always possible to have randomized controls: e.g. in order to assess effect of long-term smoking on health one cannot possibly ask people to smoke for study purposes.

Summarizing data.

I Comparing two populations via independent samples

Comparing two populations.

Example (Age of customers)

Company with department stores in Atlanta, Georgia has stores in inner city and in suburban shopping centers. Customers in different shops are sampled randomly and their age is recorded.

Table: Age of customers in inner city and suburban stores.

store type	sample size	sample mean (age)	sample SD (age)
inner city	$n_1 = 60$	$\bar{x}_1 = 40$ yrs	$s_1 = 9$ yrs
suburban	$n_2 = 80$	$\bar{x}_2 = 35$ yrs	$s_2 = 10$ yrs

Table: Two types of errors in hypothesis testing.

Question. Are customers shopping in inner city stores older than customers in suburban stores?

How could one answer such a question?

Comparing two populations.

Consider n_1 independent samples

$$X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

from the treatment population, and n_2 independent samples

$$Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

from the population of controls.

As a measure for treatment effect we would like to know average difference

$$\mu_X - \mu_Y$$

of treatments and controls.

As an estimator for this difference we use

$$\bar{X}_{n_1} - \bar{Y}_{n_2}.$$

Comparing two populations.

Since $\bar{X}_{n_1} - \bar{Y}_{n_2}$ is a linear combination of independent normals, we find

$$\bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathcal{N}(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}).$$

Suppose σ_X^2, σ_Y^2 are known. Then the statistic

$$Z := \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

follows a standard normal distribution.

In particular, this allows us to work out that

$$\bar{X}_{n_1} - \bar{Y}_{n_2} \pm z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}$$

is a $(1 - \alpha)$ confidence intervals for $\mu_X - \mu_Y$.²

²We denote by $z(\alpha)$ the (100α) th percentile of the standard normal distribution, i.e. $\Phi(z(\alpha)) = \alpha$.

Comparing two populations.

Example (Age of customers)

We may not be willing to model ages of customers by normal distribution. However, since sample sizes

$$n_1 = 60 \quad n_2 = 80$$

are large, can approximate \bar{X}_{n_1} and \bar{Y}_{n_2} by normal distributions (due to CLT).

Using s_1, s_2 as approximations for σ_X, σ_Y we find the 90% confidence interval for $\mu_x - \mu_Y$ to be

$$40 - 35 \pm z(0.05) \sqrt{\frac{9^2}{60} + \frac{10^2}{80}} = 5 \pm 1.645 \times 1.61 = 5 \pm 2.65.$$

Comparing two populations.

If sample sizes are small³ a normal approximation is not accurate in general, and one has to work harder in order to find the distribution of $X_{n_1} - Y_{n_2}$.

We're going to use the *pooled sample variance*

$$s_p^2 := \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2},$$

as an estimator for the variance of $X_{n_1} - Y_{n_2}$. Here

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

denotes the sample variance of the X 's and s_Y^2 is defined in complete analogy.

³I.e. smaller than 30.

Comparing two populations.

Notice that

$$s_p^2 = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2},$$

is a weighted sum of the variances, where each variance is weighed according to the corresponding sample size.

Remark

Imagine sampling from two populations, e.g.

$$n_1 = 1000, s_1^2 = 50 \quad n_2 = 2, s_2^2 = 1.$$

Surely, the overall variance

$$\bar{X}_{1000} - \bar{Y}_2$$

will be much closer to $s_1^2 = 50$ than to $s_2^2 = 1$, as the contribution in variance from \bar{Y}_2 is comparatively negligible.

Comparing two populations.

With the methods we developed in the first chapter (distributions derived from the normal distribution), it is not hard to show the following

Fact

Provided the variances $\sigma_X^2 = \sigma_Y^2$ in the treatment and control populations are equal, the statistic

$$t := \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\mu_X - \mu_Y)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

follows Student's t distribution with $n_1 + n_2 - 2$ df.

As before, this implies that we can work out confidence intervals for $\mu_X - \mu_Y$

Comparing two populations.

Example (Household incomes of neighborhoods)

Urban planning department interested in difference between average household income for two neighborhoods. The table summarizes data collected from randomly sampled households.

	neighborhood X	neighborhood Y
sample size	$n_1 = 8$	$n_2 = 12$
s. mean (income)	$\bar{x}_1 = \$75,000$	$\bar{x}_2 = \$82,000$
s. SD	$s_X = \$2,000$	$s_Y = \$1,800$

Table: Household incomes in two neighborhoods.

Question: Construct a 95% confidence interval for $\mu_X - \mu_Y$.

Comparing two populations.

Example (Household incomes of neighborhoods)

We assume that incomes are normally distributed⁴ with equal variances $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ for both neighborhoods. Since

$$t_{18}(0.025) = 2.101 \quad s_p^2 = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2} = 1,880.31$$

we find a 95% confidence interval for $\mu_X - \mu_Y$ as

$$\begin{aligned}\bar{x}_1 - \bar{x}_2 &\pm t_{18}(0.025)s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\&= -7,000 \pm 2.101 \times 43.36 \times 0.456 \\&= -7.000 \pm 41.54\end{aligned}$$

⁴We cannot apply CLT here due to small sample sizes.

Comparing two populations.

Hypothesis tests. Since we know that the statistic t follows Student's t distribution, we can carry out hypothesis tests. Only want to introduce new treatment if there is strong evidence that it has an effect. In other words, unless there is strong evidence in the data that treatment has an effect, we assume null hypothesis

$$H_0: \mu_X = \mu_Y$$

that on average the treatment has no effect.

When comparing two populations, alternatives usually are of the form

$$\begin{cases} \mu_X \neq \mu_Y & \text{(two-sided alternative)} \\ \mu_X > \mu_Y & \text{(one-sided alternative)} \\ \mu_X < \mu_Y & " " ". \end{cases}$$

Comparing two populations.

Hypothesis tests. Suppose we study the two-sided alternative

$$H_A: \mu_X \neq \mu_Y.$$

Put differently, we reject H_0 for large values of $|t|$.

Under H_0 the t statistic

$$t = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

follows Student's t distribution with $n_1 + n_2 - 2$ df.

Consequently, the decision rule

$$\text{reject } H_0 \text{ if } |t| > t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right)$$

yields a test at level α .⁵

Question: Test whether the mean time required until pain relief is the same for both medicines at level $\alpha = 0.5$.

⁵We denote by $t_n(\alpha)$ the $(100\alpha)th$ percentile of Student's t distribution with n df.

Comparing two populations.

Example (Medicine for pain-relief)

Effectiveness of two medicines for pain-relief are compared in a medical research study. $n = 473$ patients were randomly assigned to one of two groups. Medicine 1 was prescribed to the patients of group 1, medicine 2 was prescribed to patients in the other group. Experimenters recorded the time required to receive pain relief. These data are summarized in the table.

	group 1	group 2
sample size	$n_1 = 248$	$n_2 = 225$
sample mean (time)	$\bar{x}_1 = 24.8m$	$\bar{x}_2 = 26.1m$
sample SD	$s_X = 3.3m$	$s_Y = 4.2m$

Table: Summary of times (in minutes) required to receive pain relief.

Test whether the average time until pain relief is the same for both medicines.

Comparing two populations.

Example (Medicine for pain-relief)

Want to test null hypothesis

$$H_0: \mu_1 = \mu_2$$

versus alternative

$$H_A: \mu_1 \neq \mu_2,$$

where

$\mu_i :=$ mean time until pain relief for medicine i .

Comparing two populations.

Example (Medicine for pain-relief)

Decision rule for the two-sided test at level $\alpha = 0.5$ is

$$\text{accept } H_0 \text{ if } t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) \leq t \leq t_{n_1+n_2-2}\left(1 - \frac{\alpha}{2}\right).$$

Since

$$t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) = t_{471}(0.025) \approx -1.97$$

and the t distribution is symmetric, we have

$$t_{471}(0.975) \approx 1.97$$

Comparing two populations.

Example (Medicine for pain-relief)

We reject H_0 at level $\alpha = 0.5$, since

$$\begin{aligned}s_p^2 &= \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2} = \frac{247(3.3m)^2 + 224(4.2m)^2}{248 + 225 - 2} \\&= \frac{6641.2m^2}{471} \approx 14.1m^2,\end{aligned}$$

$$\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{1}{248} + \frac{1}{225}} \approx 0.092$$

and we obtain for the t statistic (under the null)

$$t = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{24.8 - 26.1}{3.75 \times 0.092} \approx -3.77 < -1.97.$$

Comparing two populations.

Remark

Often the requirement $\sigma_X^2 = \sigma_Y^2$ that the variances in treatment and control population be equal is not met. In this case the variance of $\bar{X}_{n_1} - \bar{Y}_{n_2}$ could be estimated by the corresponding sample variance

$$\frac{s_X^2}{n} + \frac{s_Y^2}{m}.$$

However, as a rule of thumb, the statistic

$$\frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

no longer exactly follows Student's t distribution, but it can be approximated by Student's t distribution with df the integer nearest to

$$\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m} \right)^2 \Bigg/ \left(\frac{s_X^2}{n^2(n-1)} + \frac{s_Y^2}{m^2(m-1)} \right).$$

Comparing two populations.

Nonparametric tests

Comparing two populations. Wilcoxon rank-sum test.

A nonparametric statistical method does not assume the data x_1, \dots, x_n to be observations of random variables X_1, \dots, X_n whose joint distribution stems from a specific parametric family of distributions

$$P_\theta, \theta \in \Theta.$$

Do you know an example of a nonparametric method?

Comparing two populations. Wilcoxon rank-sum test.

Evidently, a nonparametric method is *per se* more widely applicable (there is no restriction on the distribution that gives rise to the data) than a parametric method.

However, one often pays a prize for this generality, e.g.

- ▶ if information about the data generating mechanism is known, one may lose power, and
- ▶ derivations are analytically more involved.

Comparing two populations. Wilcoxon rank-sum test.

So far we compared two populations (treatments and controls) by comparing independent random (unpaired) samples from these populations. The hypothesis that a treatment has no effect was formalized as

$$H_0: \mu_X = \mu_Y.$$

Moreover, we assumed

- (1) both populations to be normally distributed
(in particular $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$)

and, in the case of small sample size, we additionally assumed

- (2) $\sigma_X^2 = \sigma_Y^2$.

Comparing two populations. Wilcoxon rank-sum test.

Wilcoxon rank-sum test (proposed by Wilcoxon in 1945) does not make any of these assumptions.

Wilcoxon's insight: Instead of the observations study their *ranks*.
(Also, makes test less sensitive to outliers.)

Null hypothesis (treatment has no effect)

H_0 : the two populations are identical

implies that X_1, \dots, X_n and Y_1, \dots, Y_n have the same joint distribution.

Alternative hypothesis

H_A : the two populations are not identical.

More precisely, if F/G denote the cdfs of treatment/control population

H_A : $\begin{cases} F(x) \leq G(x) \text{ for all } x & \text{(one-sided)} \\ F(x) \leq G(x) \text{ for all } x, \text{ or } F(x) \geq G(x) \text{ for all } x. & \text{(two-sided)} \end{cases}$

Comparing two populations. Wilcoxon rank-sum test.

This time we do not construct the test statistic from

$$\bar{X}_{n_1} \quad \text{and} \quad \bar{Y}_{n_2}.$$

Instead we study the pooled observations

$$X_1, X_2, \dots, X_{n_1}, Y_1, Y_2, \dots, Y_{n_2}$$

which we denote by

$$Z_1, Z_2, \dots, Z_{n_1+n_2}.$$

Consider the order statistics

$$Z_{(1)} \leq Z_{(2)} \leq \dots \leq Z_{(n_1+n_2)}.$$

For simplicity, assume there are no ties,
i.e. $Z_{(1)} < Z_{(2)} < \dots < Z_{(n_1+n_2)}$.

Comparing two populations. Wilcoxon rank-sum test.

The *rank*⁶ of observation X_i among the pooled observations is

$$\text{rank}(X_i) := j \quad \text{if } X_i = Z_{(j)}.$$

The rank sum of the sample from the first, respectively second, population is

$$R_1 := \sum_{i=1}^{n_1} \text{rank}(X_i) \quad \text{respectively} \quad R_2 := \sum_{i=1}^{n_2} \text{rank}(Y_i).$$

In particular,

$$R_1 + R_2 = \sum_{i=1}^{n_1+n_2} i = \frac{1}{2}(n_1 + n_2)(n_1 + n_2 + 1).$$

⁶If observations $z_{(i)} = z_{(i+1)} = \dots = z_{(i+j)}$ are ties, set their ranks equal to their average rank.

Comparing two populations. Wilcoxon rank-sum test.

Example

Consider independent samples

$$5, 10 \quad 2, 7, 9$$

of sizes $n_1 = 2$ and $n_2 = 3$ from two different populations.

Ordered pooled observations and their ranks are given in the table.

2	5	7	9	10
1	2	3	4	5

Rank sums

$$R_1 = 7, \quad R_2 = 8.$$

Comparing two populations. Wilcoxon rank-sum test.

Under the null hypothesis, observation X_1 will be smaller or larger than Y_1 with equal probability, i.e.⁷

$$\mathbb{P}\{X_1 < Y_1\} = \mathbb{P}\{X_1 > Y_1\} = \frac{1}{2},$$

and this is true for any two observations X_i, Y_j , i.e.

$$\mathbb{P}\{X_i < Y_j\} = \mathbb{P}\{X_i > Y_j\} = \frac{1}{2}.$$

What is more, we will see any particular assignment of rankings to the X s (respectively Y s) with equal probability. The total number of ways to assign rankings (numbers between 1 and $n_1 + n_2$) to the X s is $\binom{n_1+n_2}{n_1}$.

⁷We assume there are no ties, e.g. the observations could be drawn from continuous populations.

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

$n = 22$ bank accounts are sampled randomly from two different branches of a bank. The records of balances are given in the table.

branch 1		branch 2		
acc.	balance	rank	acc. balance	rank
	1095	20	885	7
	955	14	850	4
	1200	22	915	8
	1195	21	950	12.5
	925	9	800	2
	950	12.5	750	1
	805	3	865	5
	945	11	1000	16
	875	6	1050	18
	1055	19	935	10
	1025	17		
	975	15		

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

The sum of ranks for each sample are

$$R_1 = 169.5, \quad R_2 = 83.5.$$

There were $n_1 = 12$ bank accounts sampled from branch 1. The minimal value for the sum of ranks for a sample of this size is

$$R_1 = 1 + 2 + \cdots + n_1 = \frac{1}{2}n_1(n_1 + 1) = 78,$$

corresponding to the fact that all observations from the first population are smaller than any observation in the second population.

R_1 close to 78 implies that branch 1 has smaller account balances, and contradicts H_0 .

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

The maximal value for the sum of ranks for a sample of size $n_1 = 12$ where the second sample is of size $n_2 = 10$ is

$$R_1 = (n_2+1) + (n_2+2) + \cdots + (n_2+n_1) = n_1 n_2 + \frac{1}{2} n_1 (n_1+1) = 198,$$

corresponding to the fact that all observations from the first population are greater than any observation in the second population.

R_1 close to 198 implies that branch 1 has greater account balances, and also contradicts H_0 .

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

Under the null the distribution R_1 is symmetric, and we expect R_1 to be close to the average

$$\frac{n_1 n_2 + n_1(n_1 + 1)}{2} = \frac{n_1(n_1 + n_2 + 1)}{2} = 138.$$

of its minimal and maximal value.

Recall that

$$R_1 = 169.5 > 138,$$

and we might suspect that branch 1 has greater account balances.

Or could this deviation be just due to chance?

—In order to answer this question, need to know null distribution of R_1 .

Comparing two populations. Wilcoxon rank-sum test.

Fact

Consider n random samples, n_1 from treatment population, n_2 from population of controls. If R_1 denotes the rank sum of the treatment group, then under the null hypothesis

$$\mathbb{E}R_1 = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\text{Var}(R_1) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

Notice that because of symmetry this fact directly yields mean and variance of R_2 (interchange n_1 and n_2 in the formulas).

Comparing two populations. Wilcoxon rank-sum test.

In practice, instead of R_1 often a function thereof is used as test statistic

(in order to exploit symmetries of distribution of R_1).

Namely,

1. Consider the pooled observations
 $Z := (X_1, \dots, X_n, Y_1, \dots, Y_n)$.
2. Let n^* denote the size of the smaller sample.
3. Calculate $R :=$ sum of ranks in Z of smaller sample.
4. Set $R' := n^*(n_1 + n_2 + 1) - R$.
5. Set $R^* := \min(R, R')$.
6. Critical values for R^* are tabulated. If R^* is too small, reject H_0 that populations are identical.

Comparing two populations. Wilcoxon rank-sum test.

Example (Balances in bank accounts)

We find

2. $n^* = 10$,
3. $R = R_2 = 83.5$,
4. $R' := n^*(n_1 + n_2 + 1) - R = 10(12 + 10 + 1) - 83.5 = 146.5$.
5. $R^* := \min(R, R') = 83.5$.
6. From Table 8 in Appendix B of Rice's textbook find 84 as critical value for R^* ($n_1 = 12$, $n_2 = 10$) for a two-sided test at level $\alpha = 0.05$. Since $R^* = 83.5 < 84$, reject H_0 .⁸

This means that, at significance level $\alpha = 0.05$, the differences in account balances we see in the data cannot be explained by chance alone. They are due to the fact that balances in the two branches⁹ are indeed different.

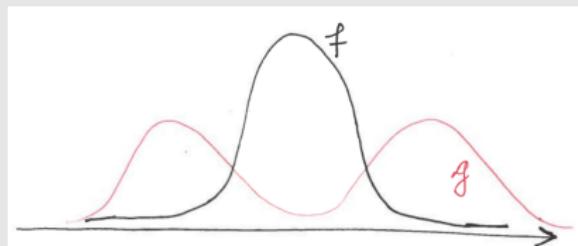
⁸Critical value for significance level $\alpha = 0.01$ is 76, so Wilcoxon does not reject at this level.

⁹More precisely: the distributions of balances in the bank accounts differ.

Comparing two populations. Wilcoxon rank-sum test.

Example (A cautionary example)

(X_i) i.i.d. $\sim F$ with density f (Y_i) i.i.d. $\sim G$ with density g



Densities f, g symmetric about their common mean

$$\mu = \int_{-\infty}^{\infty} tf(t)dt = \int_{-\infty}^{\infty} tg(t)dt.$$

Under H_0 expect R_1 to be close to its mean $n_1(n_1 + n_2 + 1)/2$ supporting

$$H_0: F = G.$$

But, we know that $F \neq G$ by construction! What went wrong?

Comparing two populations. Mann-Whitney test.

In 1947 Mann and Whitney proposed a different test based on ranks, which (at first sight) is not based on the rank-sum statistic. This test occurs often in the literature.

It turns out, however, that Mann and Whitney's U statistic is a simple function of the rank sum R_2 , and the two tests are therefore equivalent.

Comparing two populations. Mann-Whitney test.

Suppose the data are modeled as draws from two populations with

$$\begin{aligned}F &\text{ cdf of treatment population} \\G &\text{ cdf of control population.}\end{aligned}$$

As before, we are interested in testing

$$H_0: F = G.$$

Instead of ranking the observations, consider the probability

$$\pi := \mathbb{P}\{X < Y\}$$

that a sample X from the treatment population is smaller than a sample Y from the control population. Under H_0 we have $\pi = \frac{1}{2}$.

Comparing two populations. Mann-Whitney test.

A good estimator for

$$\pi := \mathbb{P}\{X < Y\}$$

should be the relative frequency of the pairs (X_i, Y_j) of observations such that $X_i < Y_j$. Since there are $n_1 n_2$ possible pairs in total, the estimator is

$$\begin{aligned}\hat{\pi} &:= \frac{\#\{(X_i, Y_j) : X_i < Y_j\}}{n_1 n_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{1}_{\{X_i < Y_j\}} \\ &= \frac{1}{n_1 n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \mathbf{1}_{\{X_{(i)} < Y_{(j)}\}}.\end{aligned}$$

Notice that

$$\sum_{i=1}^{n_1} \mathbf{1}_{\{X_{(i)} < Y_{(j)}\}} = \text{number of } X\text{s less than } Y_{(j)} = \text{rank}(Y_{(j)}) - j,$$

where the $-j$ accounts for $Y_{(1)}, \dots, Y_{(j)}$.

Comparing two populations. Mann-Whitney test.

Plugging

$$\sum_{i=1}^{n_1} \mathbf{1}_{\{X_{(i)} < Y_{(j)}\}} = \text{rank}(Y_{(j)}) - j$$

into the formula for $\hat{\pi}$ yields

$$\begin{aligned}\hat{\pi} &= \frac{1}{n_1 n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \mathbf{1}_{\{X_{(i)} < Y_{(j)}\}} = \frac{1}{n_1 n_2} \sum_{j=1}^{n_2} (\text{rank}(Y_{(j)}) - j) \\ &= \frac{1}{n_1 n_2} \left(\sum_{j=1}^{n_2} \text{rank}(Y_{(j)}) - \sum_{j=1}^{n_2} j \right) = \frac{1}{n_1 n_2} \left(R_2 - \frac{n_2(n_2 + 1)}{2} \right),\end{aligned}$$

where R_2 is the sum of ranks of the sample from the control population.

Comparing two populations. Mann-Whitney test.

The Mann-Whitney U statistic is defined as

$$U_Y := n_1 n_2 \hat{\pi} = \#\{(X_i, Y_j) : X_i < Y_j\} = R_2 - \frac{n_2(n_2 + 1)}{2}.$$

Corollary

From the previous fact on mean and variance of R_1 (respectively R_2) we find

$$\mathbb{E}U_Y = \frac{n_1 n_2}{2}, \quad \text{Var}(U_Y) = \text{Var}(R_2) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{2}.$$

Ex: prove these statements.

Comparing two populations. Mann-Whitney test.

It can be shown that as $n_1, n_2 \rightarrow \infty$, the null distribution of the Mann-Whitney statistic U_Y converges to a normal distribution, i.e.

$$\frac{U_Y - \mathbb{E}U_Y}{\sqrt{\text{Var}(U_Y)}} \rightarrow N,$$

under the null, where $N \sim \mathcal{N}(0, 1)$.

[Heuristics: Normal approximation to binomial distribution]

In practice, the normal approximation is already used for sample sizes n_1, n_2 greater than 10.

Comparing two populations. Mann-Whitney test.

Example (Balances in bank accounts)

For our data we find

$$U_Y = 91.5$$

and a p-value of 0.041 (e.g. using **wilcox.test** in R).

As with the Wilcoxon rank-sum test the Mann-Whitney test rejects H_0 at significance level $\alpha = 0.05$, but not at significance level $\alpha = 0.01$

Comparing two populations. Mann-Whitney test.

Bootstrapping $\hat{\pi}$. We saw the importance of the estimator

$$\hat{\pi} := \frac{\#\{(X_i, Y_j) : X_i < Y_j\}}{n_1 n_2}$$

of $\pi := \mathbb{P}\{X < Y\}$ for the Mann-Whitney test.

To bootstrap $\hat{\pi}$,

1. Approximate the

- ▶ cdf F of the treatment population by the ecdf F_{n_1} ,
- ▶ cdf G of the control population by the ecdf G_{n_2} .

2. Take independent random samples

- ▶ $x_1^*, x_2^*, \dots, x_{n_1}^*$ from F_{n_1}
- ▶ $y_1^*, y_2^*, \dots, y_{n_2}^*$ from G_{n_2} ,

and use them to compute the corresponding bootstrap sample

$$\hat{\pi}_1^* = \frac{\#\{(x_i^*, y_j^*) : x_i < y_j\}}{n_1 n_2}.$$

Repeat B times to obtain simulated samples $\hat{\pi}_1^*, \hat{\pi}_2^*, \dots, \hat{\pi}_B^*$.

3. Can now either study distribution of $\hat{\pi}_1^*, \hat{\pi}_2^*, \dots, \hat{\pi}_B^*$, or compute any of its properties.

STAT 135, Concepts of Statistics

Helmut Pitters

Comparing two populations - matched samples

Department of Statistics
University of California, Berkeley

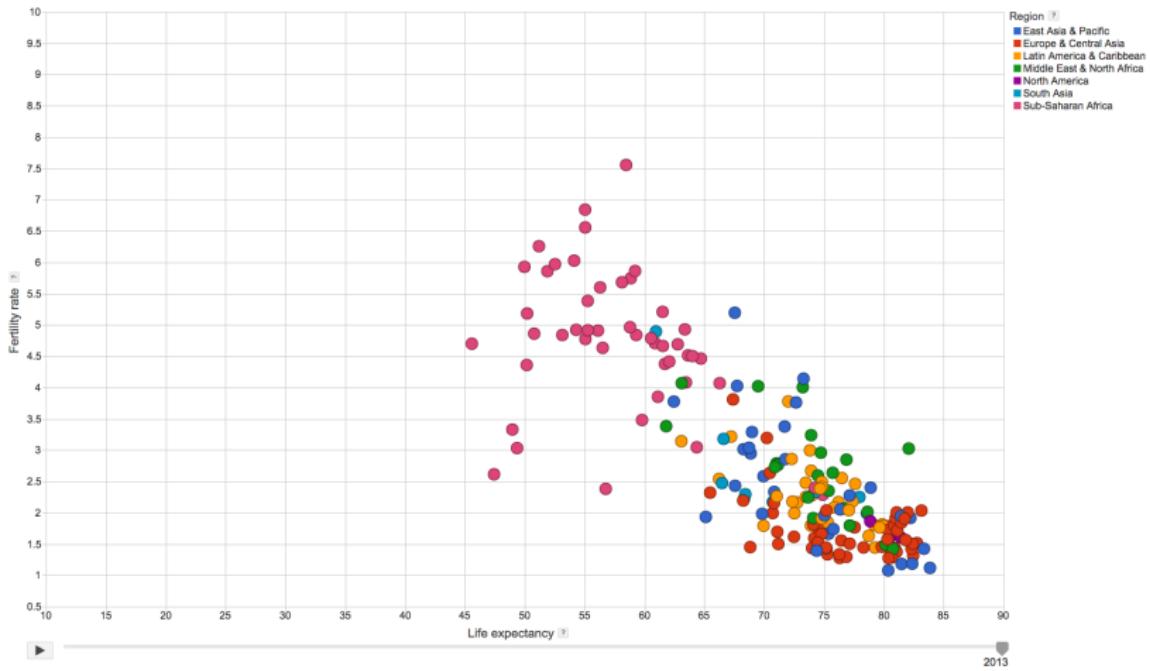
April 17, 2017

Review: Covariance and correlation.

Example: comparing production methods

Often in statistics not only interested in one random variable/population, but in relationships between different populations. – What if populations are not independent?

Relationship between two quantitative variables are usually displayed via scatterplots.



Scatterplot from Google Public Data Explorer.

Review: Covariance and correlation.

Recall from STAT 134: *Covariance* of two random variables X and Y is defined by

$$\text{Cov}(X, Y) := \mathbb{E} [(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The covariance can be interpreted as a measure for joint variability or degree of linear association.

If $\text{Cov}(X, Y) = 0$, we called X and Y *uncorrelated*. While independence of X, Y implies $\text{Cov}(X, Y) = 0$, the converse is not true.

Review: Covariance and correlation.

Recall from STAT 134 some useful formulas:

$$\text{Cov}(X, X) = \text{Var}(X)$$

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{Cov}(X, Y) = \text{Cov}(Y, X) \quad (\text{symmetry})$$

$$\text{Cov}(aX + b, Y) = a \text{Cov}(X, Y) \quad (\text{multilinearity})$$

Review: Covariance and correlation.

A drawback of the covariance is that it depends on the units in which X and Y are measured. We therefore agreed to first transform a random variable X to standard units, i.e. we center and rescale X

$$X^* := \frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}}$$

to standard units.¹ Accordingly, we defined the *correlation* of X and Y by

$$\begin{aligned}\text{Corr}(X, Y) &:= \text{Cov}(X^*, Y^*) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} \\ &= \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \mathbb{E}[X^*Y^*].\end{aligned}$$

¹In particular, $\mathbb{E}X^* = 0$, $\text{Var}(X^*) = 1$.

Review: Covariance and correlation.

We found that for any two real random variables X and Y

$$-1 \leq \text{Corr}(X, Y) \leq 1.$$

Moreover, $\text{Corr}(X, Y) = 1$ or $= -1$ implies the existence of reals a, b such that

$$Y = aX + b.$$

Review: Covariance and correlation.

For a sample $(x_1, y_1), \dots, (x_n, y_n)$ of n paired observations the *sample covariance* is defined as

$$s_{xy} := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

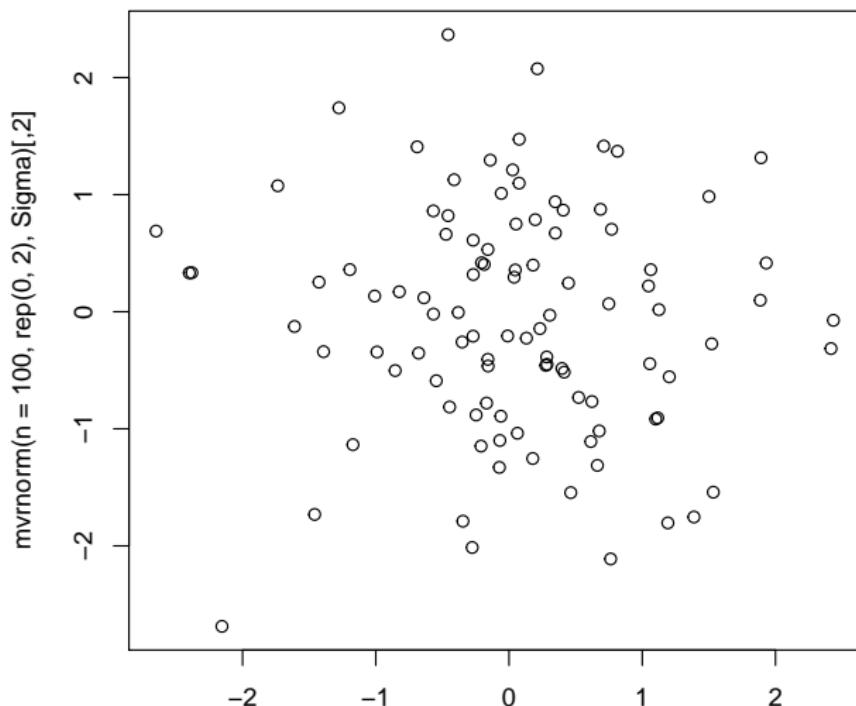
In complete analogy to the case of random variables, the *sample correlation coefficient* is defined as

$$r := \frac{s_{xy}}{s_x s_y},$$

where $s_x^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, respectively s_y denotes the *sample variance* of the x -, respectively y -sample.

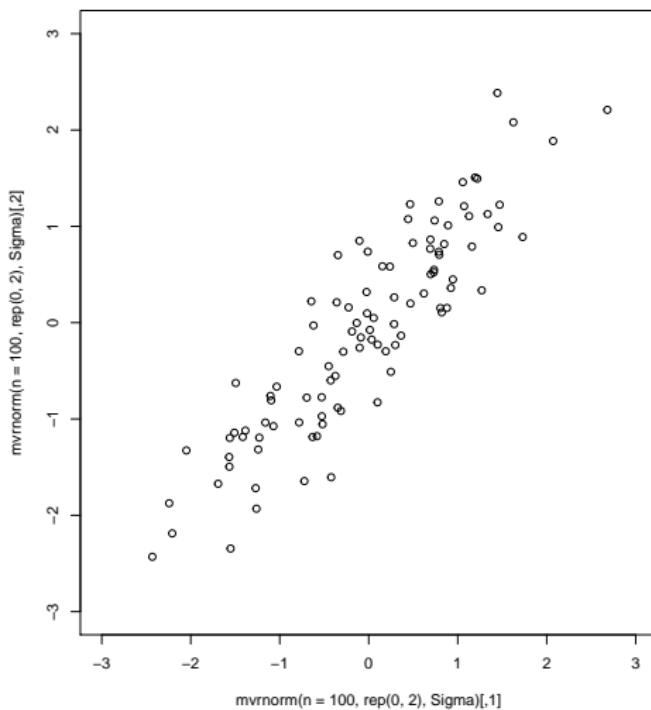
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0$.



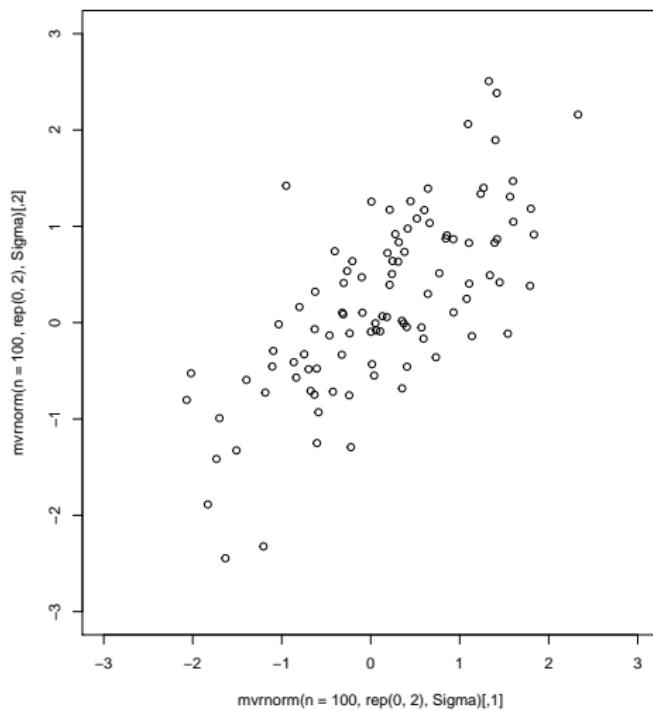
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0.9$.



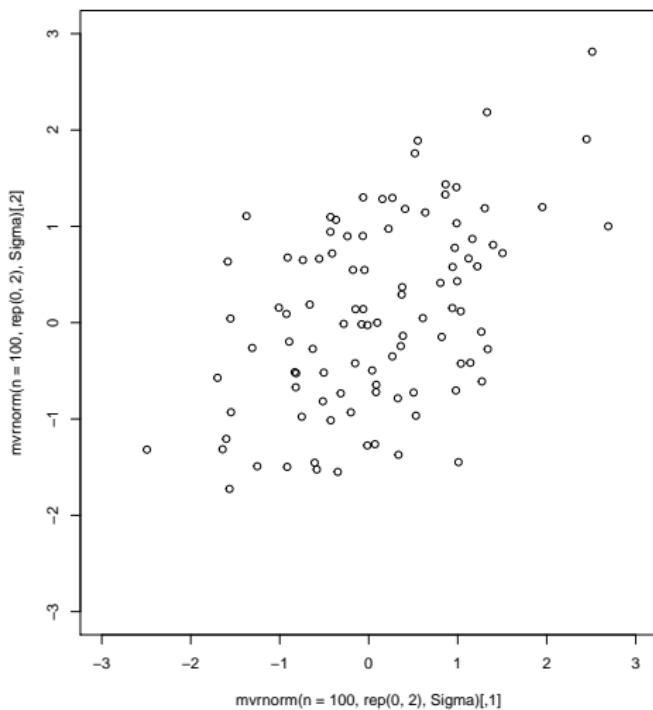
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0.7$.



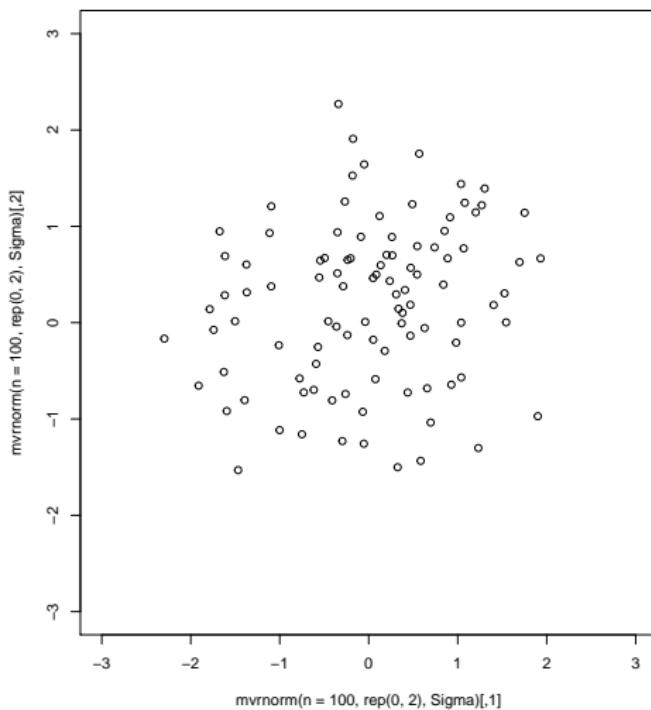
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0.5$.



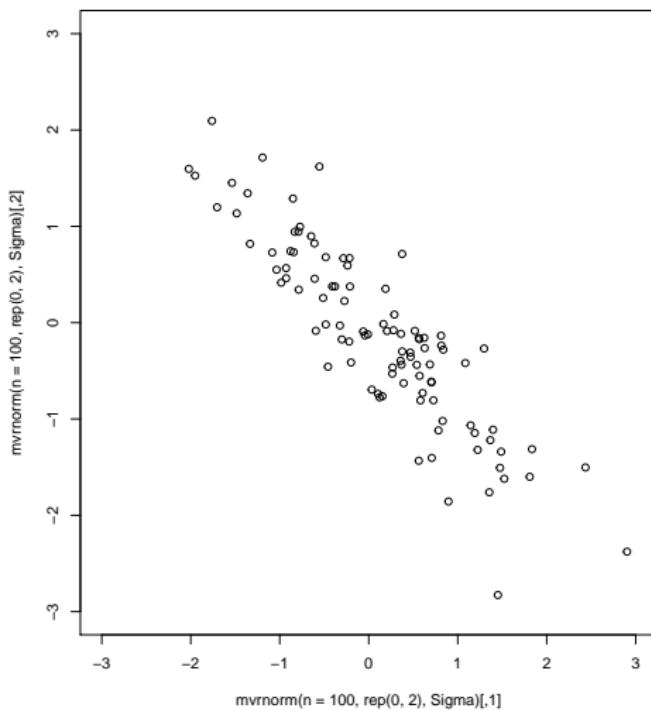
Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = 0.2$.



Review: Covariance and correlation.

100 samples from bivariate Normal distribution, $r = -0.9$.



Comparing two populations. Matched samples

Back to comparing two populations. – Why matched pairs?

Example (Comparing production methods)

Want to compare two production methods. Each of $n = 6$ workers completes task once by method 1, and once by method 2.

Completion times (t_i^1, t_i^2) for each worker $i \in \{1, \dots, n\}$

are recorded (in minutes).

method 1 (t_i^1)	method 2 (t_i^2)	difference ($t_i^1 - t_i^2$)
6.0	5.4	.6
5.0	5.2	-.2
7.0	6.5	.5
6.2	5.9	.3
6.0	6.0	.0
6.4	5.8	.6

Table: Completion times for method 1 and 2.

Comparing two populations. Matched samples

Example (Comparing production methods)

A meaningful procedure that compares method 1 and method 2 will be based on

differences in completion times $t_i^1 - t_i^2$.

Completion times not only depend on production method, but also on worker. Want to eliminate the effect of worker speed on differences $t_i^1 - t_i^2$ (and thus reduce their variance).

Not matching completion times

$$t_1^1, t_2^1, t_3^1, \dots, t_n^1, t_1^2, t_2^2, \dots, t_n^2$$

corresponds to having $2n$ workers completing tasks, which introduces additional randomness (or noise) due to individual worker speed.

Comparing two populations. Matched samples

Setup and notation. Samples $(x_1, y_1), \dots, (x_n, y_n)$ assumed to be observations from n pairs

$(X_1, Y_1), \dots, (X_n, Y_n)$ of i.i.d. random variables.

Conceptually, comparing matched samples is easier than comparing non-paired samples from two populations. Why?

Can study

$$\text{differences} \quad D_i := X_i - Y_i.$$

(If samples were not matched, which of the $n_1 n_2$ differences $X_i - Y_j$ should we consider?)

Comparing two populations. Matched samples

Population parameters

$$\mu_X := \mathbb{E}X_1, \quad \mu_Y := \mathbb{E}Y_1$$

$$\sigma_X^2 := \text{Var}(X_1), \quad \sigma_Y^2 := \text{Var}(Y_1)$$

$$\sigma_{XY} := \text{Cov}(X_i, Y_i) = \rho\sigma_X\sigma_Y,$$

where $\rho := \text{Corr}(X_i, Y_i)$. Consequently²

$$\mathbb{E}D_i = \mathbb{E}[X_i - Y_i] = \mu_X - \mu_Y$$

$$\begin{aligned}\text{Var}(D_i) &= \text{Var}(X_i - Y_i) = \text{Var}(X_i) + \text{Var}(Y_i) - 2\text{Cov}(X_i, Y_i) \\ &= \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y.\end{aligned}$$

²Generally, what can you say about noise in $X \pm Y$ based on σ_X, σ_Y , and ρ ?

Comparing two populations. Matched samples

As before, interested in null hypothesis

$$H_0: \mu_X = \mu_Y \quad \text{or} \quad \mu_X - \mu_Y = 0$$

that populations (treatment and control) do not differ,
i.e. treatment has no effect, vs. $H_A: \mu_X \neq \mu_Y$.

As estimator for $\mu_X - \mu_Y$ (and test statistic) take

$$\bar{D}_n := \frac{1}{n} \sum_{i=1}^n D_i = \bar{X}_n - \bar{Y}_n$$

with

$$\mathbb{E}\bar{D}_n = \mu_X - \mu_Y \quad \text{Var}(\bar{D}_n) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y).$$

In principle we are done: as long as we can work out distribution of \bar{D}_n , we can

- ▶ work out confidence intervals
- ▶ do hypothesis tests, etc.

Comparing two populations. Matched samples

$$\mathbb{E}\bar{D}_n = \mu_X - \mu_Y \quad \text{Var}(\bar{D}_n) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y).$$

If samples were taken independently, without matching

$$\text{Var}(\bar{X}_n - \bar{Y}_n) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2).$$

Thus matching is more effective if $\rho > 0$, i.e. if populations are positively correlated.

Comparing two populations. Matched samples

Normal distribution

Comparing two populations. Matched samples

Now assume $D_1, \dots, D_n \sim \mathcal{N}(\mu_D, \sigma_D^2)$. Let

$$\bar{d}_n := \frac{1}{n} \sum_{i=1}^n d_i \quad s_{\bar{D}_n} := \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d}_n)^2$$

If σ_D^2 is known, use statistic

$$\frac{\bar{D}_n - \mu_D}{\sigma_D} \sim \mathcal{N}(0, 1)$$

for inference.

Otherwise, statistic

$$t := \frac{\bar{D}_n - \mu_D}{s_{\bar{D}_n}/\sqrt{n}} \sim t_{n-1}$$

follows Student's t distribution with $n - 1$ df and can be used for inference.

Comparing two populations. Matched samples

Example (Comparing production methods)

If worker i is particularly quick at completing the task by method 1, we expect that this is partly due to him being a fast worker.

Therefore expect him to also be quick (in relation to other workers) at completing the task by method 2.

More formally: expect completion times (T_i^1, T_i^2) to be positively correlated. This is why its meaningful to match samples.

Comparing two populations. Matched samples

Example

Since $n = 6$ we have under null hypothesis $\mu_D = 0$

$$t = \frac{\bar{D}_n}{s_{\bar{D}_n}/\sqrt{n}} \sim t_5.$$

Since $t_5(0.025) = -2.57$, a level $\alpha = 0.05$ test of

$$H_0: \mu_D = 0 \quad \text{vs.} \quad H_A: \mu_D \neq 0$$

has decision rule

reject H_0 if $t \leq -2.57$ or $t \geq 2.57$.

Comparing two populations. Matched samples

Example

From data compute

$$\bar{d}_n = \frac{1}{n} \sum_{i=1}^n (t_i^1 - t_i^2) = 0.3$$

$$s_{\bar{D}_n} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d}_n)^2} = 0.335$$

hence

$$t = \frac{\bar{d}_n}{s_{\bar{D}_n}/\sqrt{n}} = 2.19$$

that is the test does not reject H_0 . The difference in completion times we see in the data can be explained by chance due to random sampling (at level $\alpha = 0.05$).

Comparing two populations. Matched samples

Nonparametric tests. We discuss ideas of nonparametric tests applied to two different settings of matched samples without going into details.

1. Data do not follow normal distribution

idea: rank absolute values $|D_1|, \dots, |D_n|$

null hypothesis: populations are identical

under null, $D_i = X_i - Y_i$ is symmetric about 0, and

$$\sum_{i=1}^n \text{sgn}(D_i) \text{rank}(|D_i|) \quad \text{should be close to 0,}$$

where $\text{sgn}(x)$ denotes the sign of x .

[Wilcoxon signed rank test]

2. Data are not quantitative. E.g. customers indicate their preference for one of two products.

Comparing two populations. Matched samples

Let D be a (continuous) real r.v., symmetric about 0,
i.e. $D =_d -D$.

$$\mathbb{P}\{\text{sgn}(D) = 1\} = \mathbb{P}\{D \geq 0\} = \frac{1}{2} = \mathbb{P}\{D < 0\} = \mathbb{P}\{\text{sgn}(D) = -1\},$$

that is, $\text{sgn}(D)$ has Bernoulli 1/2 distribution on $\{-1, 1\}$.
Moreover, for any $x \in \mathbb{R}$

$$\begin{aligned}\mathbb{P}\{\text{sgn}(D) = 1, |D| > x\} &= \mathbb{P}\{D > x\} \\&= \frac{1}{2}(\mathbb{P}\{D > x\} + \mathbb{P}\{-D > x\}) \\&= \frac{1}{2}\mathbb{P}\{|D| > x\} \\&= \mathbb{P}\{\text{sgn}(D) = 1\} \mathbb{P}\{|D| > x\},\end{aligned}$$

showing (with similar calculation for $\text{sgn}(D) = -1$) that $\text{sgn}(D)$ and $|D|$ are independent.

Comparing two populations. Matched samples

Back to our setting:

D_1, \dots, D_n are i.i.d., continuous, symmetric about 0.

Hence

$$(\text{sgn}(D_1), \text{sgn}(D_2), \dots, \text{sgn}(D_n)) \quad \text{and} \quad (|D_1|, |D_2|, \dots, |D_n|)$$

are both i.i.d. and independent of each other, and since $\text{rank}(|D_i|)$ only depends on $(|D_1|, \dots, |D_n|)$,

$$(\text{sgn}(D_1), \dots, \text{sgn}(D_n)) \text{ and } (\text{rank}(|D_1|), \dots, \text{rank}(|D_n|))$$

are independent of each other.

STAT 135, Concepts of Statistics

Helmut Pitters

Linear regression

Department of Statistics
University of California, Berkeley

April 20, 2017

Linear regression.

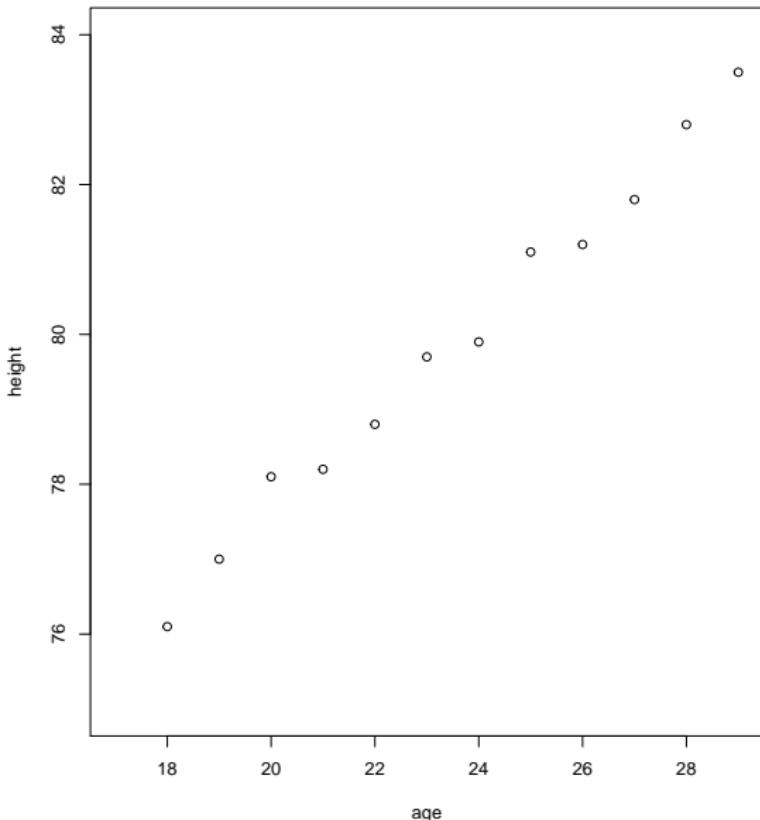
Example (Growth of Kalama children)

How (fast) do children grow?

In context of nutritional study in Egyptian village Kalama heights of $n = 161$ children¹ were recorded from their 18th to 29th month. Next figure shows scatterplot of average heights.

¹Sampled randomly among all children in Kalama in their 18th month.

Figure: Age (months) and height (cm) of 161 Kalama children.



Linear regression.

Example (Growth of Kalama children)

age (months)	height (cm)
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

Table: Age and height of Kalama 161 children.

Linear regression.

Example (Growth of Kalama children)

Consider another child, of 25 months say, (sampled randomly) from Kalama.

Question: How would you guess the child's height?

Linear regression.

Example (Growth of Kalama children)

Guessing height of 25 month old child.

Heuristic:

Scatterplot displays linear pattern—correlation coefficient
 $r = 0.994$.

- ▶ Draw “the” line suggested by the scatterplot that comes “as close as possible” to the data points (*fitting a line*).
- ▶ Read off y -value at $x = 25$.

Challenges:

- ▶ What do we mean by “the” line?
- ▶ How do we find “the” line?

Linear regression.

Example (Growth of Kalama children)

Clearly, no straight line can fit the data points

$$(x_i, y_i) = (x_i\text{th month, average age in } x_i\text{th month})$$

exactly. Some of the data points will be missed, which incurs an error.

However, a good candidate for “the line” among all straight lines should be the one that minimizes this error.

Linear regression.

There are numerous ways to define the distance between a straight line and a point (x_i, y_i) .

Suppose line we are looking for is given by

$$f(x) = \beta_1 x + \beta_0,$$

i.e. has

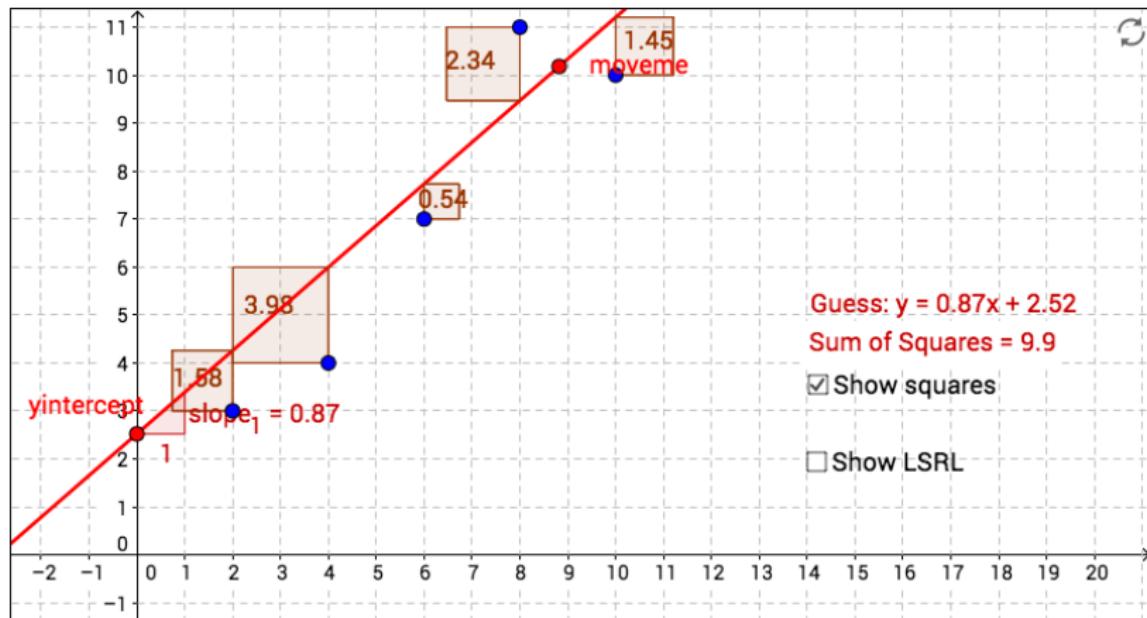
slope	β_1
intercept	β_0 .

Agree to define distance, or error, between (x_i, y_i) and $f(x)$ by squared distance

$$(y_i - f(x_i))^2 = (y_i - \beta_1 x_i - \beta_0)^2.$$

Linear regression.

Figure: Finding the regression line, <http://www.geogebra.org/m/105271>.



Linear regression.

Overall error between line and data points is sum of errors at each data point, i.e.

$$S(\beta_0, \beta_1) := \sum_{i=1}^n (y_i - f(x_i))^2 = \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0)^2.$$

So, formally our task is to find $\hat{\beta}_0, \hat{\beta}_1$ such that

$S(\beta_0, \beta_1)$ is minimised.

Then

$$\hat{r}(x) = \hat{\beta}_1 x + \hat{\beta}_0$$

is the line we are looking for, the so-called *fitted line*.

This method of minimizing $S(\beta_0, \beta_1)$ is called the *method of least squares*.

Linear regression.

Theorem

Slope and intercept of the regression line are given by²

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r \sqrt{\frac{s_{yy}}{s_{xx}}}, \quad \hat{\beta}_0 = -\hat{\beta}_1 \bar{x} + \bar{y}.$$

Proof.

Exercise. Hint: Find roots of partial derivatives

$$\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = -2 \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0)$$

$$\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = -2 \sum_{i=1}^n x_i (y_i - \beta_1 x_i - \beta_0).$$

□

²Notice that $r = 0$ if and only if $\hat{\beta}_1 = 0$.

Linear regression.

Proof.

$$S(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

For root of partial w.r.t. β_0 find

$$0 = \frac{\partial}{\partial \beta_0} S = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) \implies \beta_0 n = n\bar{y} - \beta_1 n\bar{x}.$$

Simple observation:

$$\sum_i x_i y_i - \sum_i \bar{x}\bar{y} = \sum_i (x_i - \bar{x})y_i = \sum_i (x_i - \bar{x})(y_i - \bar{y}),$$

thus, choosing $y = x$, $\sum_i x_i^2 - \sum_i \bar{x}^2 = \sum_i (x_i - \bar{x})^2$. □

Linear regression.

Proof.

$$S(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

Now find root of partial w.r.t. β_1 .

$$0 = \frac{\partial}{\partial \beta_1} S = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) x_i \text{ and, replacing } \beta_0 \text{ by } \hat{\beta}_0$$

$$0 = \sum_i y_i x_i - \beta_0 n \bar{x} - \beta_1 \sum_i x_i^2$$

$$= \sum_i y_i x_i - (\bar{y} - \beta_1 \bar{x}) n \bar{x} - \beta_1 \sum_i x_i^2$$

$$= \sum_i x_i y_i - \sum_i \bar{x} \bar{y} - \beta_1 \left(\sum_i x_i^2 - \sum_i \bar{x}^2 \right)$$

$$= \sum_i (x_i - \bar{x})(y_i - \bar{y}) - \beta_1 \sum_i (x_i - \bar{x})^2.$$

Linear regression.

Example (Growth of Kalama children)

Output of command **lm(formula = height~age)** in R:

```
--  
age <- 18:29  
height <- c(76.1, 77.0, 78.1, 78.2, 78.8, 79.7, 79.9, 81.1, 81.2,  
81.8, 82.8, 83.5)
```

Call:

```
lm(formula = height ~ age)
```

Coefficients:

(Intercept)	age
-------------	-----

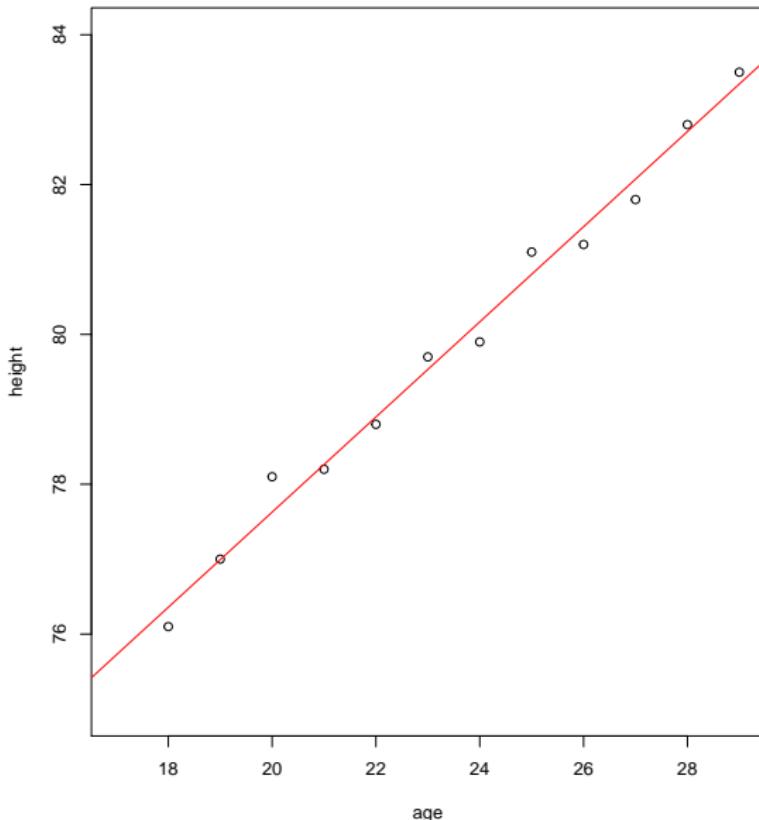
64.928	0.635
--------	-------

—

Consequently, the fitted line is given by

$$\hat{r}(x) = 0.635x + 64.93.$$

Figure: Age (months) and height (cm) of 161 Kalama children.



Simple linear regression

Simple linear regression.

So far, we have not made any assumption on the distribution underlying the data $(x_1, y_1), \dots, (x_n, y_n)$.

Standard and most popular statistical model is *simple linear regression model*:

$$(\text{linearity}) \quad y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i \in \{1, \dots, n\}$$

where $\epsilon_1, \dots, \epsilon_n$ is an i.i.d. sequence of r.v.s such that

$$\mathbb{E}\epsilon_i = 0 \quad \text{Var}(\epsilon_i) = \sigma^2.$$

x_i are assumed fixed, called *predictor* variables, and

y_i are called *response* variables

β_1 can be interpreted as average change
of y_i if x_i increases by 1 unit

β_0, β_1 are called the regression coefficients

Simple linear regression.

Remark

- ▶ Want to stress that here we model the x_i as being deterministic (think of a controlled experiment where the investigator can freely choose the x_i). More general models of linear regression allow to model the x_i as random.
- ▶ Notice that $\mathbb{E}y_i = \beta_0 + \beta_1 x_i$, $\text{Var}(y_i) = \sigma^2$. I.e. if we would repeatedly draw samples $(x_1, y_1), \dots, (x_n, y_n)$ and average out the y_i values for each x_i , we'd obtain the line $x \mapsto \beta_0 + \beta_1 x$ through the points $(x_i, \mathbb{E}y_i)$.
- ▶ Here “simple” refers to the x_i being one-dimensional.
- ▶ Unknown parameters are: $\beta_0, \beta_1, \sigma^2$. Will use $\hat{\beta}_0, \hat{\beta}_1$ as estimators for β_0, β_1 .

Simple linear regression.

If we had taken another sample $(x_1, y_1), \dots, (x_n, y_n)$ from the same model, i.e. such that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i \in \{1, \dots, n\},$$

would have obtained different values $\hat{\beta}_0, \hat{\beta}_1$ for optimal intercept and slope of regression line by method of least squares.

Suppose we wanted to use fitted line

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

to make predictions.

$\hat{y}_i := \hat{r}(x_i)$ are the *predicted/fitted* values

$\hat{\epsilon}_i := y_i - \hat{r}(x_i)$ are the *residuals*.

How reliable are $\hat{\beta}_0, \hat{\beta}_1$ as estimators for β_0, β_1 ?

Simple linear regression.

Recall estimators of intercept and slope given by least squares method:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = -\hat{\beta}_1 \bar{x} + \bar{y}.$$

Using $\sum_i (x_i - \bar{x}) = 0$ and $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, which implies $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon}$, where $\bar{\epsilon} := \frac{1}{n} \sum_{i=1}^n \epsilon_i$, we obtain

$$\mathbb{E}\hat{\beta}_1 = \mathbb{E} \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i - \beta_0 - \beta_1 \bar{x} - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1,$$

since $\mathbb{E}\epsilon_i = 0$, and therefore

$$\mathbb{E}\hat{\beta}_0 = \mathbb{E}[\bar{y} - \hat{\beta}_1 \bar{x}] = \beta_0.$$

Simple linear regression.

We just proved

Theorem

For simple linear regression, the estimators

$$\hat{\beta}_0 \text{ and } \hat{\beta}_1$$

for intercept and slope of the regression line are unbiased.

Simple linear regression.

Again, recall

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Thus, for the variance we find

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2} = \frac{n\sigma^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}.$$

Simple linear regression.

Theorem

For $\hat{\beta}_0, \hat{\beta}_1$ given by the method of least squares under simple linear regression we have

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{n\sigma^2}{n \sum_i x_i^2 - (\sum_i x_i)^2},$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \sum_i x_i}{n \sum_i x_i^2 - (\sum_i x_i)^2} = -\bar{x} \text{Var}(\hat{\beta}_1),$$

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_i x_i^2}{n \sum_i x_i^2 - (\sum_i x_i)^2} = \frac{\sum x_i^2}{n} \text{Var}(\hat{\beta}_1).$$

Simple linear regression.

Proof.

Let's first compute (co)variances of y_i, \bar{y} using the multilinearity of covariance:

$$\text{Var}(y_i) = \sigma^2 \quad \text{Cov}(y_i, y_j) = \delta_{ij}\sigma^2,$$

where δ_{ij} equals 1 if $i = j$ and 0 otherwise,

$$\text{Var}(\bar{y}) = \frac{\sigma^2}{n} \quad \text{Cov}(y_i, \bar{y}) = \frac{1}{n} \sum_{j=1}^n \text{Cov}(y_i, y_j) = \frac{\sigma^2}{n}$$

and

$$\text{Cov}(\bar{y}, \hat{\beta}_1) = \frac{\sum(x_i - \bar{x}) \text{Cov}(\bar{y}, y_i)}{\sum(x_i - \bar{x})^2} = 0, \text{ and therefore}$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) = -\bar{x} \text{Var}(\hat{\beta}_1).$$

□

Simple linear regression.

Proof.

Recall $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\begin{aligned}\text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{y}) + \text{Var}(\hat{\beta}_1 \bar{x}) - 2 \text{Cov}(\bar{y}, \hat{\beta}_1 \bar{x}) \\&= \frac{\sigma^2}{n} + \bar{x}^2 \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\&= \frac{\sigma^2(\sum(x_i - \bar{x})^2 + n\bar{x}^2)}{n \sum(x_i - \bar{x})^2} = \frac{\sum x_i^2}{n} \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\&= \frac{\sum x_i^2}{n} \text{Var}(\hat{\beta}_1).\end{aligned}$$

□

Simple linear regression.

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_i x_i^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}, \text{Var}(\hat{\beta}_1) = \frac{n\sigma^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}.$$

In practice, parameter σ^2 is usually unknown. However, since

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \hat{r}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

residuals

$$\hat{\epsilon}_i = y_i - \hat{r}(x_i)$$

should capture the information about σ^2 contained in the sample.
We'll see that

$$s^2 := \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2$$

is unbiased estimator for σ^2 . The expression $RSS := \sum_{i=1}^n \hat{\epsilon}_i^2$ is called the *residual sum of squares (RSS)*.

Simple linear regression.

Consequently, replacing σ^2 by s^2 , we estimate the variances of $\hat{\beta}_0, \hat{\beta}_1$ by

$$s_{\hat{\beta}_0}^2 = \frac{s^2 \sum_i x_i^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$
$$s_{\hat{\beta}_1}^2 = \frac{ns^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}.$$

Simple linear regression.

Recall simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where (ϵ_i) are i.i.d. r.v.s with $\mathbb{E}\epsilon_i = 0$, $\text{Var}(\epsilon_i) = \sigma^2$.

If we assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (which is the case in numerous settings), then responses y_i are normally distributed. In particular,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = -\hat{\beta}_1 \bar{x} + \bar{y}.$$

are linear combinations of normally distributed r.v.s, hence normal.
One can show that

$$\frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} \text{ and } \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{n-2}.$$

Consequently, one can construct confidence intervals and hypothesis tests for β_0, β_1 .

Simple linear regression.

Example (Growth of Kalama children)

R command: **summary(lm(formula = height~age))**.

Simple linear regression.

For calculations, the formulas

$$\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_j - \sum x_i \sum x_j y_j}{n \sum x_i^2 - (\sum x_i)^2}$$
$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_j}{n \sum x_i^2 - (\sum x_i)^2}$$

can be convenient.

Proof.

Recall $\sum_i (x_i - \bar{x}) = \sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - n^{-1}(\sum x_i)^2$.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum (x_i - \bar{x})^2},$$

now multiply both numerator & denominator by n . □

Simple linear regression.

$$\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_j - \sum x_i \sum x_j y_j}{n \sum x_i^2 - (\sum x_i)^2}$$
$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_j}{n \sum x_i^2 - (\sum x_i)^2}$$

Proof.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum y_i \sum (x_j - \bar{x})^2}{n \sum x_i^2 - (\sum x_i)^2} - \frac{n \bar{x} \sum x_i y_i - \bar{x} \sum x_i \sum y_j}{n \sum x_i^2 - (\sum x_i)^2},$$

and

$$\begin{aligned} \sum y_i \left(\sum x_j^2 - n \bar{x}^2 \right) + \sum y_i \bar{x} \sum x_j &= \sum y_i \left(\sum x_j^2 - n \bar{x}^2 + n \bar{x}^2 \right) \\ &= \sum y_i \sum x_j^2. \end{aligned}$$



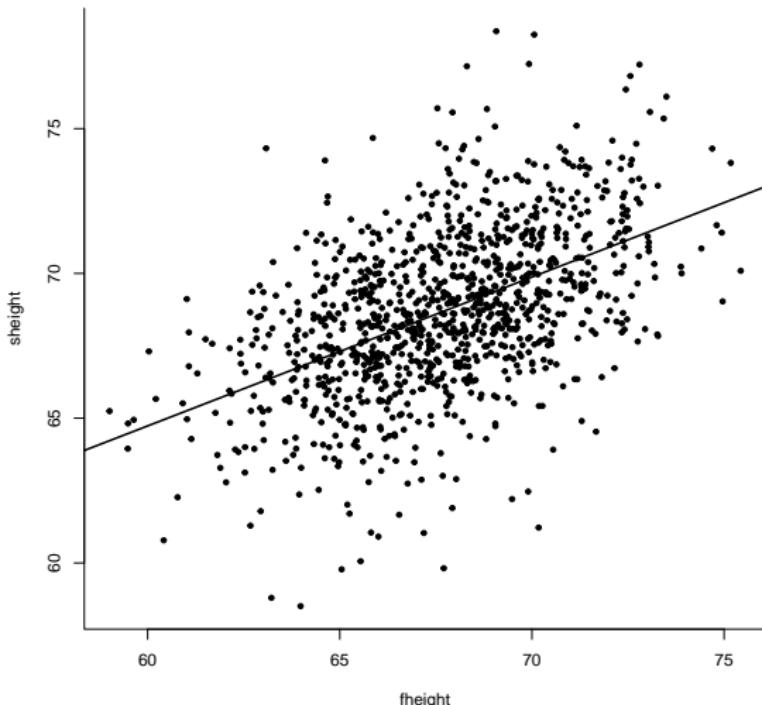
Regression to the mean.

I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning. When I had finished my enthusiastic speech, one of the most seasoned instructors in the audience raised his hand and made his own short speech, which began by conceding that positive reinforcement might be good for the birds, but went on to deny that it was optimal for flight cadets. He said, "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don't tell us that reinforcement works and punishment does not, because the opposite is the case."

Daniel Kahnemann, Nobel laureate in Economic Sciences

Galton's study of heights of fathers and sons.

Figure: Height (inch) of fathers and their sons.



Galton's study of heights of fathers and sons.

Children of larger than average parents tend to be smaller than their parents. Likewise, children of smaller than average parents tend to be larger than their parents.

Why?

Regression to the mean.

This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them. I immediately arranged a demonstration in which each participant tossed two coins at a target behind his back, without any feedback. We measured the distances from the target and could see that those who had done best the first time had mostly deteriorated on their second try, and vice versa. But I knew that this demonstration would not undo the effects of lifelong exposure to a perverse contingency.

Daniel Kahnemann, Nobel laureate in Economic Sciences
(from Kahnemann's biography at nobelprize.org)

Linear regression.

A great number of models can be cast as linear via appropriate data transformations.

E.g. consider a model of exponential growth/decay for a population

$$P_t = P_0 e^{m(t-t_0)},$$

where

P_t = size of population at time t

m = growth rate.

Question: How to estimate growth rate from given data (x_i, P_{x_i}) ?

Idea: Apply log

$$\begin{aligned}\log P_t &= \log P_0 + m(t - t_0) = \log P_0 - mt_0 + mt \\ &= \beta_0 + \beta_1 t.\end{aligned}$$

Introducing random errors (ϵ_t) the model becomes

$$\log P_t = \beta_0 + \beta_1 t + \epsilon_t$$

and can be treated as a simple linear regression model.

Simple linear regression.

Example (Repair times, 1)

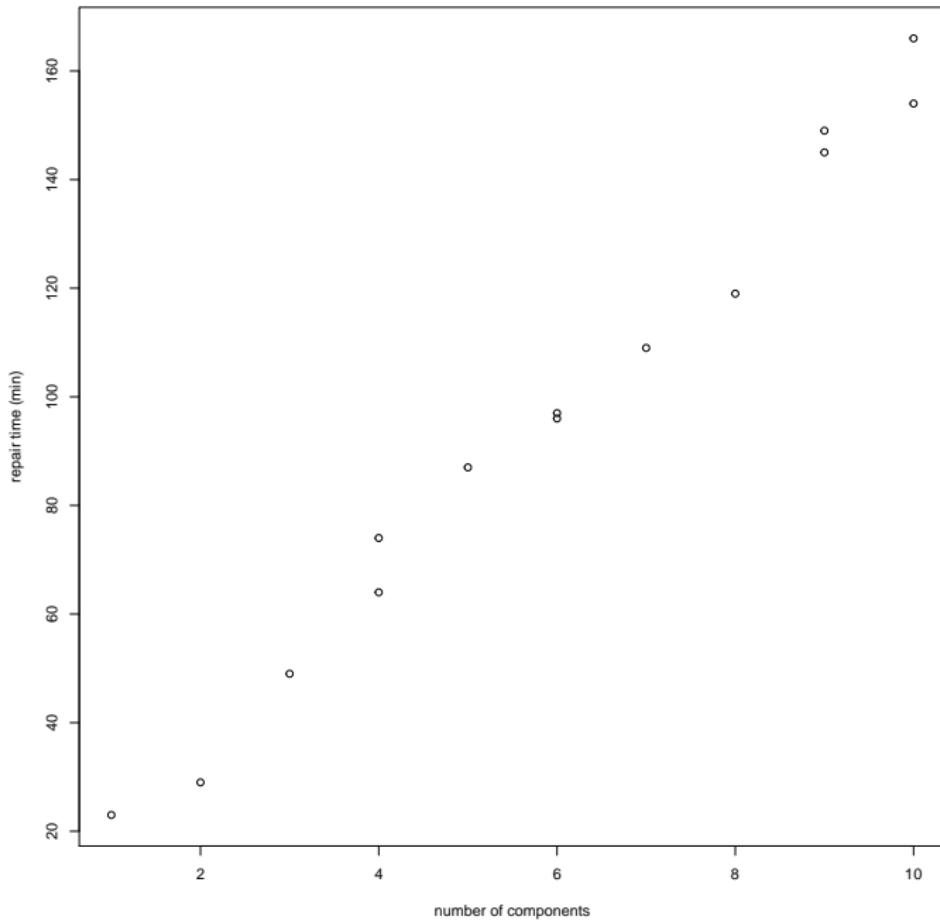
Company repairs certain devices with small number of electronic components. Important for company's allocation of resources is to understand how lengths of service calls, (y_i), depend on number of electronic components, (x_i).

x_i = number of electrical components in device for i th service

y_i = length of i th service call.

Next figure shows plot of sample of size $n = 14$.

Repair times



Simple linear regression.

Example (Repair times, 1)

From data we find

$$\bar{x} = 6, \quad \bar{y} = 97.21, \quad r = 0.996.$$

High correlation coefficient confirms linear association we see in scatter plot.

Seems appropriate to model relationship between number of components and repair time by simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where (ϵ_i) are i.i.d. $\mathbb{E}\epsilon_i = 0$, $\text{Var}(\epsilon_i) = 1$.

Simple linear regression.

Example (Repair times, 1)

For least squares estimates obtain

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \hat{y})}{\sum_i (x_i - \bar{x})^2} = 15.51,$$

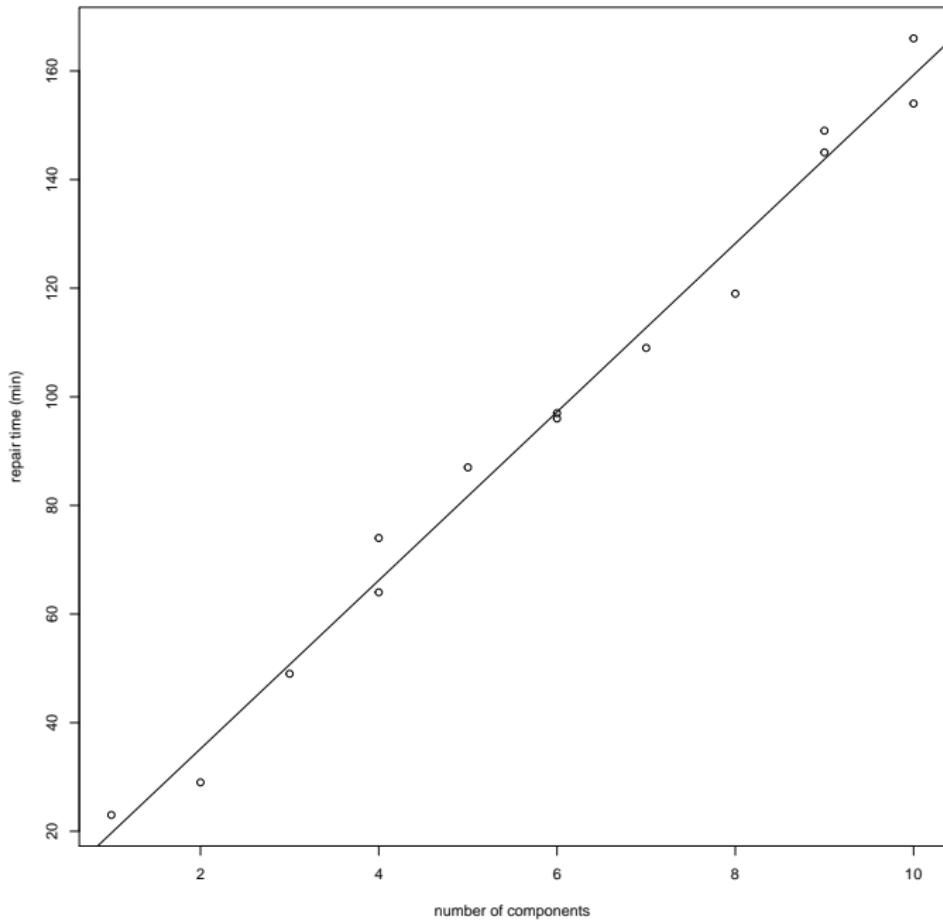
$$\hat{\beta}_0 = -\beta_1 \bar{x} + \bar{y} = 4.16,$$

and thus the regression line

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = 4.16 + 15.51mx.$$

Let's check this result informally by looking at the plot of the regression line.

Repair times



Simple linear regression.

Example (Repair times, 1)

regression line

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = 4.16m + 15.51mx.$$

Interpretation:

$\hat{\beta}_0 = 4.16m$: time needed to set up repair service
independent of number of components

$\hat{\beta}_1 = 15.5m$: avg. increase in service time for
each additional component.

However, one should be careful with interpretations. In particular
the value $\hat{r}(0) = 4.16m$ is obtained by interpolating in a region
where there are no observations (extrapolation)!

Simple linear regression.

Model deficiencies. In addition to examining correlation coefficient r and t statistics one should always consult a graphical plot of the data to see whether linear fit is actually meaningful.

To this end analysis of residuals ($\hat{\epsilon}_i$) is often essential.

We now present four rather different data sets given by Anscombe (1973), all with same r and t statistic. However, only one of them shows linear association.

R: plots of Anscombe's data sets.

Simple linear regression.

Detecting model deficiencies. To detect whether assumptions of linear model might be violated, often residuals

$$\hat{\epsilon}_i = y_i - \hat{r}(x_i)$$

are examined graphically.

E.g. residuals ($\hat{\epsilon}_i$) can be plotted against

- ▶ predictors: $(x_i, \hat{\epsilon}_i)$
- ▶ fitted values: $(\hat{y}_i, \hat{\epsilon}_i)$
- ▶ order in which observations occurred: $(i, \hat{\epsilon}_i)$.

Plots that do not appear symmetric about³ $y = 0$ or show a distinct pattern of variation suggest that (some) assumptions of linear model are violated.

³A consequence of $\mathbb{E}\epsilon_i = 0$.

Simple linear regression.

Assumption that residuals (ϵ_i) are normally distributed can be checked via probability plots.

Example (Repair times, 1)

R: Analysis of residuals.

Analysis does not contradict our assumptions of a linear model.

Whether random disturbances (ϵ_i) can be assumed to be normally distributed is not so clear.

Simple linear regression.

Since residuals have different variances, often *standardized residuals*

$$\hat{\epsilon}_i^s := \frac{\hat{\epsilon}_i}{\text{estimated SE}(\hat{\epsilon}_i)}$$

are examined instead.

Here estimated standard error of $\hat{\epsilon}_i$ is given by⁴

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

⁴We don't prove this.

Simple linear regression.

Example (Repair times, 1)

One way to assess how good predictors (x_i) explain responses (y_i) in linear model is to check hypotheses

$$H_0: \hat{\beta}_1 = 0 \quad \text{vs} \quad H_A: \hat{\beta}_1 \neq 0.$$

Null hypothesis can be interpreted as there being almost no linear association between (x_i) and (y_i)⁵.

Obtain for t-statistic ($df = 12$)

$$\frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} = 30.71, \quad (\text{since } \text{SE}(\hat{\beta}_1) = 0.505)$$

and p-value 8.92×10^{-13} . Thus H_0 can be safely rejected.

⁵Recall $\hat{\beta}_1 = rs_x/s_y$.

Simple linear regression.

Example (Repair times, 1)

A $(1 - \alpha) = 95\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm \text{SE}(\hat{\beta}_1) t_{12}\left(\frac{\alpha}{2}\right) = 15.51m \pm 0.505m \times 2.18 = [14.41m, 16.61m].$$

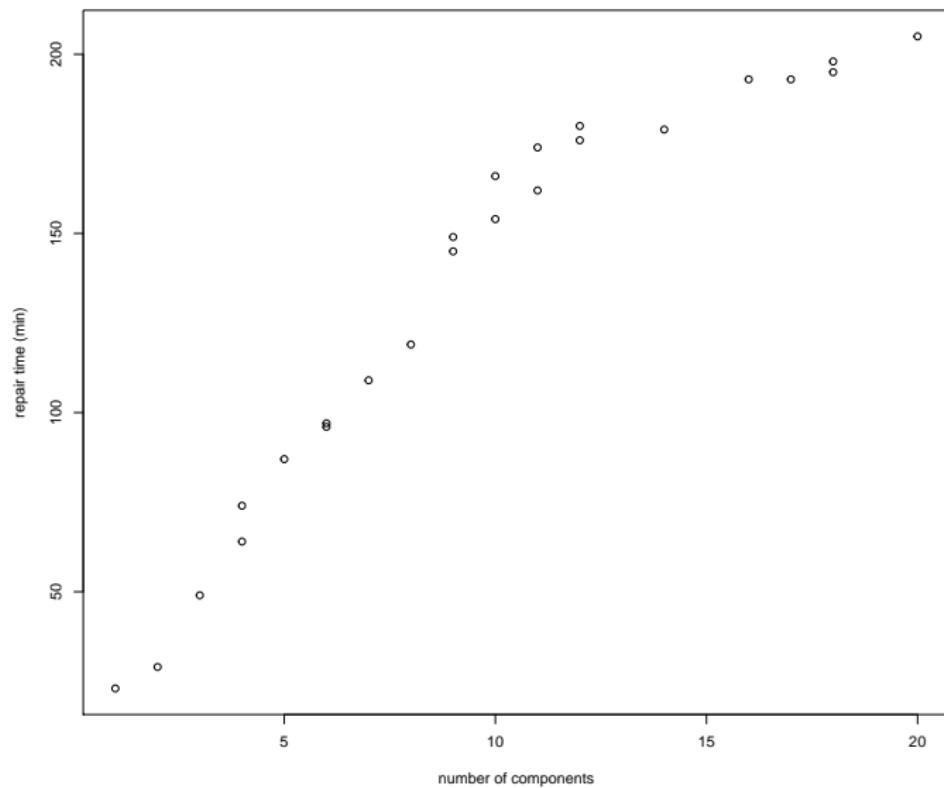
Interpretation: If we observe a large number of repair times, in 90% of cases the avg. increase in repair time per component is between 14.41m and 16.61m

Simple linear regression.

Example (Repair times, 2)

In another sampling period (same sampling method, same company) 10 further observations were taken. Next figure shows their scatter plot.

Repair times 2



Simple linear regression.

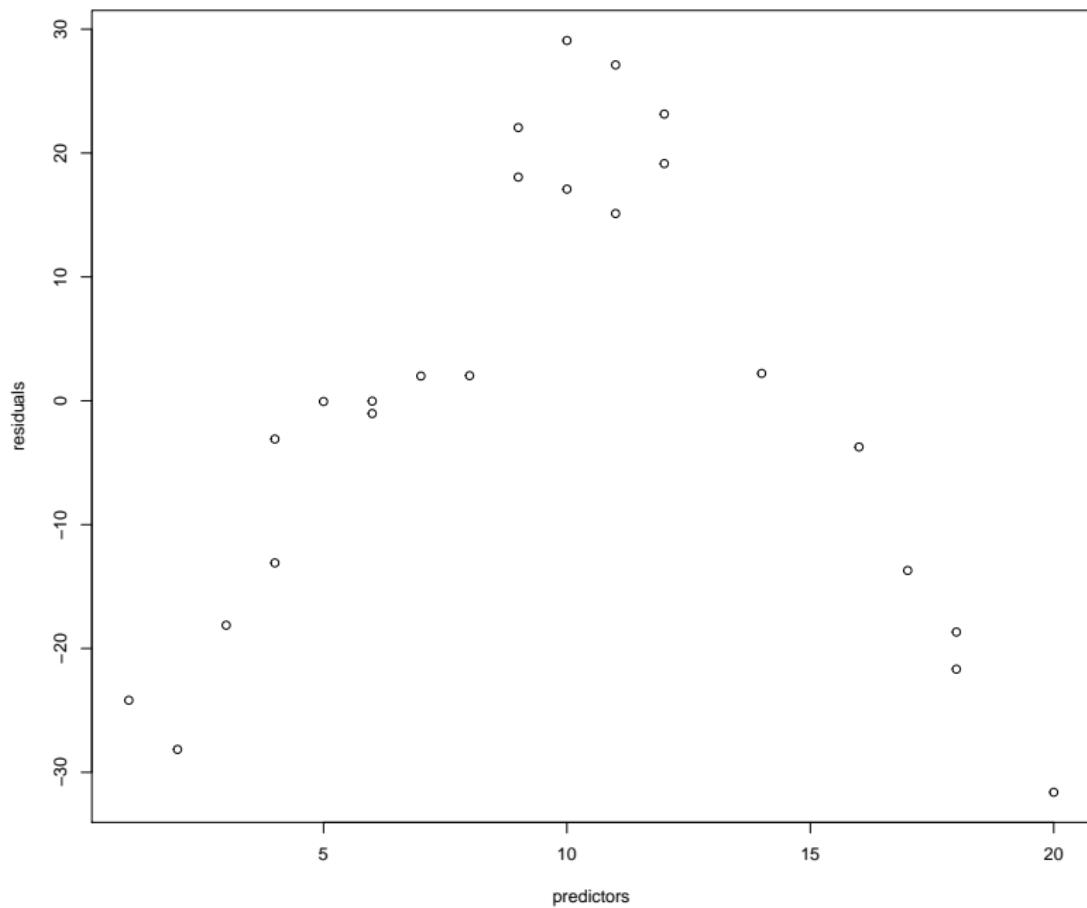
Example (Repair times, 2)

There seems to be a break point at $x \approx 12, 13$ components, where slope declines.

Suggests an effect of time efficiency when device for repair has 12 or more components.

R: analyzing residuals.

Analysis of residuals 2: residuals vs predictors



Simple linear regression.

Example (Repair times, 2)

Linear regression does not seem to be appropriate to proceed with statistical analysis.

Investigator may want to study reasons behind observed efficiency.
This can lead to observation of another important factor influencing repair time (besides # of components) -> multiple regression?!

Multiple linear regression

Multiple linear regression.

Saw before that in simple linear regression response y_i is modeled to depend on one predictor x_i in a linear fashion:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \quad (1)$$

Often don't expect that response can be well predicted by one predictor only. E.g. don't expect income of person to only depend on years of education, but also on socioeconomic status of parents, work experience, etc.

Therefore, allow y_i to depend on several predictors $x_{i,1}, \dots, x_{i,p-1}$:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1} + \epsilon_i,$$

where (ϵ_i) are i.i.d., $\mathbb{E}\epsilon_i = 0$, $\text{Var}(\epsilon_i) = \sigma^2$, and β_0, \dots, β_p are unknown parameters.

This is the so-called *multiple linear regression model*.

Multiple linear regression.

There are two particularly important statistical regimes depending on the ratio of the number of p and the number of observations:

- ▶ $\frac{p}{n} \rightarrow 0$ as $n \rightarrow \infty$. Classical regime (that we consider).
- ▶ $\frac{p}{n} \rightarrow C > 0$ as $n \rightarrow \infty$. A regime that emerged in more recent years (high dimensional data, e.g. data on health status of patient). Harder to study, less tools available, but more interesting.

Simple linear regression.

To study multiple linear regression, turns out that notions from linear algebra are very well suited. We collect some of these notions and introduce some notation.

Notions needed from linear algebra. Consider

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n.$$

Often consider x as $n \times 1$ matrix, i.e. an element of $\mathbb{R}^{n \times 1}$.

Euclidean norm: $\|x\| := \sqrt{\sum_{i=1}^n x_i^2}$.