MATH 2P82
MATHEMATICAL STATISTICS
(Lecture Notes)

Contents

Chapter 1 PROBABILITY REVIEW

Basic Combinatorics

Number of permutations of n distinct objects: $n!$

Not all distinct, such as, for example *aaabbc*:

$$\frac{6!}{3!2!1!} \overset{def.}{=} \binom{6}{3,2,1}$$

or

$$\frac{N!}{n_1!n_2!n_3!.....n_k!} \overset{def.}{=} \binom{N}{n_1, n_2, n_3, ..., n_k}$$

in general, where $N = \sum_{i=1}^{k} n_i$ which is the total word length (MULTINOMIAL COEFFICIENT).

Selecting r out of n objects (without duplication), counting all possible arrangements:

$$n \times (n-1) \times (n-2) \times .... \times (n-r+1) = \frac{n!}{(n-r)!} \overset{def.}{=} P_r^n$$

(NUMBER OF PERMUTATIONS).

Forget their final arrangement:

$$\frac{P_r^n}{r!} = \frac{n!}{(n-r)!r!} \overset{def.}{=} C_r^n$$

(NUMBER OF COMBINATIONS). This will also be called the BINOMIAL COEFFICIENT.

If we can duplicate (any number of times), and count the arrangements:

$$n^r$$

Binomial expansion

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^{n-i} y^i$$

Multinomial expansion

$$(x+y+z)^n \sum_{\substack{i,j,k \geq 0 \\ i+j+k=n}} \binom{n}{i,j,k} x^i y^j z^k$$

$$(x+y+z+w)^n = \sum_{\substack{i,j,k,\ell \geq 0 \\ i+j+k+\ell=n}} \binom{n}{i,j,k,\ell} x^i y^j z^k w^\ell$$

etc.

Random Experiments (Basic Definitions)

Sample space

is a collection of all possible outcomes of an experiment.

The individual (*complete*) outcomes are called SIMPLE EVENTS.

Events

are SUBSETS of the sample space (A, B, C,...).

Set Theory

The old notion of:	is (are) now called:
Universal set Ω	Sample space
Elements of Ω (its individual 'points')	Simple events (complete outcomes)
Subsets of Ω	Events
Empty set $\emptyset$	Null event

We continue to use the word INTERSECTION (notation: $A \cap B$, representing the collection of simple events common to both A *and* B), UNION ($A \cup B$, simple events belonging to either A *or* B or *both*), and COMPLEMENT ($\overline{A}$, simple events *not* in A). One should be able to visualize these using Venn diagrams, but when dealing with more than 3 events at a time, one can tackle problems only with the help of

Boolean Algebra

Both $\cap$ and $\cup$ (individually) are COMMUTATIVE and ASSOCIATIVE.

Intersection is DISTRIBUTIVE over union: $A \cap (B \cup C \cup ...) = (A \cap B) \cup (A \cap C) \cup ...$

Similarly, union is distributive over intersection: $A \cup (B \cap C \cap ...) = (A \cup B) \cap (A \cup C) \cap ...$

Trivial rules: $A \cap \Omega = A$, $A \cap \emptyset = \emptyset$, $A \cap A = A$, $A \cup \Omega = \Omega$, $A \cup \emptyset = A$, $A \cup A = A$, $A \cap \overline{A} = \emptyset$, $A \cup \overline{A} = \Omega$, $\overline{\overline{A}} = A$.

Also, when $A \subset B$ (A is a SUBSET of B, meaning that every element of A also belongs to B), we get: $A \cap B = A$ (the smaller event) and $A \cup B = B$ (the bigger event).

DeMorgan Laws: $\overline{A \cap B} = \overline{A} \cup \overline{B}$, and $\overline{A \cup B} = \overline{A} \cap \overline{B}$, or in general

$$\overline{A \cap B \cap C \cap ...} = \overline{A} \cup \overline{B} \cup \overline{C} \cup ...$$

and vice versa (i.e. $\cap \leftrightarrow \cup$).

A and B are called (mutually) **exclusive** or DISJOINT when $A \cap B = \emptyset$ (no overlap).

Probability of Events

Simple events can be assigned a **probability** (relative frequency of its occurrence in a *long* run). It's obvious that each of these probabilities must be a non-negative number. To find a probability of *any other* event A (not necessarily simple), we then add the probabilities of the simple events A consists of. This immediately implies that probabilities must follow a few basic rules:

$$\begin{aligned} \Pr(A) &\geq 0 \\ \Pr(\emptyset) &= 0 \\ \Pr(\Omega) &= 1 \end{aligned}$$

(the relative frequency of all Ω is obviously 1).

We should mention that $\Pr(A) = 0$ does not necessarily imply that $A = \emptyset$.

Probability rules

$\Pr(A\cup B)=\Pr(A)+\Pr(B)$ but *only* when $A\cap B=\emptyset$ (*disjoint*). This implies that $\Pr(\overline{A})=1-\Pr(A)$ as a special case.

This also implies that $\Pr(A\cap\overline{B})=\Pr(A)-\Pr(A\cap B)$.

For any A and B (possibly overlapping) we have

$$\Pr(A\cup B)=\Pr(A)+\Pr(B)-\Pr(A\cap B)$$

Can be extended to: $\Pr(A\cup B\cup C)=\Pr(A)+\Pr(B)+\Pr(C)-\Pr(A\cap B)-\Pr(A\cap C)-\Pr(B\cap C)+\Pr(A\cap B\cap C)$.

In **general**

$$\begin{aligned}\Pr(A_1\cup A_2\cup A_3\cup\ldots\cup A_k) &= \sum_{i=1}^{k}\Pr(A_i)-\sum_{i<j}^{k}\Pr(A_i\cap A_j)+\sum_{i<j<\ell}^{k}\Pr(A_i\cap A_j\cap A_\ell)-\ldots\\ &\quad\pm\Pr(A_1\cap A_2\cap A_3\cap\ldots\cap A_k)\end{aligned}$$

The formula computes the probability that *at least one* of the A_i events happens.

The probability of getting *exactly one* of the A_i events is similarly computed by:

$$\begin{aligned}&\sum_{i=1}^{k}\Pr(A_i)-2\sum_{i<j}^{k}\Pr(A_i\cap A_j)+3\sum_{i<j<\ell}^{k}\Pr(A_i\cap A_j\cap A_\ell)-\ldots\\ &\pm k\Pr(A_1\cap A_2\cap A_3\cap\ldots\cap A_k)\end{aligned}$$

Important result

Probability of any (Boolean) expression involving events A, B, C, ... can be *always* converted to a linear combination of probabilities of the individual events and their simple (non-complemented) *intersections* ($A\cap B$, $A\cap B\cap C$, etc.) only.

Probability tree

is a graphical representation of a two-stage (three-stage) random experiment.(effectively its sample space - each complete path being a simple event).

The individual branch probabilities (usually simple to figure out), are the so called CONDITIONAL PROBABILITIES.

Product rule

$$\begin{aligned}\Pr(A\cap B) &= \Pr(A)\cdot\Pr(B|A)\\ \Pr(A\cap B\cap C) &= \Pr(A)\cdot\Pr(B|A)\cdot\Pr(C|A\cap B)\\ \Pr(A\cap B\cap C\cap D) &= \Pr(A)\cdot\Pr(B|A)\cdot\Pr(C|A\cap B)\cdot\Pr(D|A\cap B\cap C)\\ &\vdots\end{aligned}$$

Conditional probability

The general definition:

$$\Pr(B|A)\equiv\frac{\Pr(A\cap B)}{\Pr(A)}$$

All basic formulas of probability remain true. *conditionally*, e.g.: $\Pr(\overline{B}|A)=1-\Pr(B|A)$, $\Pr(B\cup C|A)=\Pr(B|A)+\Pr(C|A)-\Pr(B\cap C|A)$, etc.

Total-probability formula

A PARTITION represents chopping the sample space into several smaller events, say A_1, A_2, A_3, ..., A_k, so that they

(i) don't overlap (i.e. are all mutually exclusive): $A_i \cap A_j = \emptyset$ for any $1 \leq i, j \leq k$

(ii) cover the whole Ω (i.e. 'no gaps'): $A_1 \cup A_2 \cup A_3 \cup ... \cup A_k = \Omega$.

For any partition, and an unrelated even B, we have

$$\Pr(B) = \Pr(B|A_1) \cdot \Pr(A_1) + \Pr(B|A_2) \cdot \Pr(A_2) + ... + \Pr(B|A_k) \cdot \Pr(A_k)$$

Independence

of two events is a very natural notion (we should be able to tell from the experiment): when one of these events happens, it does not effect the probability of the other. Mathematically, this is expressed by either

$$\Pr(B|A) \equiv P(B)$$

or, equivalently, by

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

Similarly, for three events, their MUTUAL independence means

$$\Pr(A \cap B \cap C) = \Pr(A) \cdot \Pr(B) \cdot \Pr(C)$$

etc.

Mutual independence of A, B, C, D, ... **implies** that any event build of A, B, ... must be independent of any event build out of C, D, ... [as long as the two sets are *distinct*].

Another **important result** is: To compute the probability of a Boolean expression (itself an event) involving only mutually independent events, it is sufficient to know the events' *individual* probabilities.

Discrete Random Variables

A RANDOM VARIABLE yields a *number*, for every possible outcome of a random experiment.

A **table** (or a **formula**, called PROBABILITY FUNCTION) summarizing the information about

1. possible outcomes of the RV (numbers, arranged from the smallest to the largest)

2. the corresponding probabilities

is called the PROBABILITY DISTRIBUTION.

Similarly, DISTRIBUTION FUNCTION: $F_x(k) = \Pr(X \leq k)$ computes *cumulative* probabilities.

Bivariate (joint) distribution

of **two** *random variables* is similarly specified via the corresponding PROBABILITY FUNCTION

$$f(i,j) = \Pr(X = i \cap Y = j)$$

with the **range** of possible i and j values. One of the two ranges is always '*marginal*' (the limits are constant), the other one is '*conditional*' (i.e. both of its limits may depend on the value of the other random variable).

Based on this, one can always find the corresponding MARGINAL DISTRIBUTION of X:

$$f_x(i) = \Pr(X = i) = \sum_{j|i} f(i,j)$$

and, similarly, the marginal distribution of Y.

Conditional distribution

of X, given an (observed) value of Y, is defined by

$$f_x(i|Y = \mathbf{j}) \equiv \Pr(X = i \,|\, Y = \mathbf{j}) = \frac{\Pr(X = i \cap Y = \mathbf{j})}{\Pr(Y = \mathbf{j})}$$

where i varies over its *conditional* range of values (given $Y = \mathbf{j}$).

Conditional distribution has all the properties of an ordinary distribution.

Independence

of X and Y means that the outcome of X cannot influence the outcome of Y (and vice versa) - something we can gather from the experiment.

This implies that $\Pr(X = i \cap Y = j) = \Pr(X = i) \times \Pr(Y = j)$ for *every* possible combination of i and j

Multivariate distribution

is a distribution of three of more RVs - conditional distributions can get rather tricky.

Expected Value of a RV

also called its **mean** or **average**, is a number which corresponds (empirically) to the average value of the random variable when the experiment is repeated, independently, infinitely many times (i.e. it is the *limit* of such averages). It is computed by

$$\mu_x \equiv \mathbb{E}(X) \equiv \sum_i i \times \Pr(X = i)$$

(weighted average), where the summation is over all possible values of i.

In general, $\mathbb{E}[g(X)] \neq g\left(\mathbb{E}[X]\right)$.

But, for a LINEAR TRANSFORMATION,

$$\mathbb{E}(aX + c) = a\mathbb{E}(X) + c$$

Expected values related to X and Y

In **general** we have

$$\mathbb{E}\left[g(X,Y)\right]=\sum_i\sum_j g(i,j)\times \Pr(X=i\cap Y=j)$$

This would normally *not* equal to $g(\mu_x,\mu_y)$, except:

$$\mathbb{E}\left[aX+bY+c\right]=a\mathbb{E}(X)+b\mathbb{E}(Y)+c$$

The previous **formula** easily extends to any number of variables:

$$\boxed{\mathbb{E}\left[a_1X_1+a_2X_2+...+a_kX_k+c\right]=a_1\mathbb{E}(X_1)+a_2\mathbb{E}(X_2)+...+a_k\mathbb{E}(X_k)+c}$$

(**no** *independence* necessary).

When X and Y are *independent*, we also have

$$\mathbb{E}(X\cdot Y)=\mathbb{E}(X)\cdot\mathbb{E}(Y)$$

and, in **general**:

$$\mathbb{E}\left[g_1(X)\cdot g_2(Y)\right]=\mathbb{E}\left[g_1(X)\right]\cdot\mathbb{E}\left[g_2(Y)\right]$$

Moments (univariate)

Simple:

$$\mathbb{E}(X^n)$$

Central:

$$\mathbb{E}\left[(X-\mu_x)^n\right]$$

Of these, the most important is the **variance** of X:

$$\mathbb{E}\left[(X-\mu_x)^2\right]=\mathbb{E}(X^2)-{\mu_x}^2$$

Its square root is the **standard deviation** of X, notation: $\sigma_x=\sqrt{\mathrm{Var}(X)}$ (this is the Greek letter 'sigma').

The interval $\mu-\sigma$ to $\mu+\sigma$ should contain the 'bulk' of the distribution – anywhere from 50 to 90%.

When $Y\equiv aX+c$ (a linear transformation of X), we get

$$\mathrm{Var}(Y)=a^2\mathrm{Var}(X)$$

which implies

$$\sigma_y=|a|\cdot\sigma_x$$

Moments (bivariate or 'joint')

Simple:

$$\mathbb{E}(X^n\cdot Y^m)$$

Central

$$\mathbb{E}\left[(X-\mu_x)^n\cdot(Y-\mu_y)^m\right]$$

The most important of these is the **covariance** of X and Y:

$$\mathrm{Cov}(X,Y)\equiv\ \mathbb{E}\left[(X-\mu_x)\cdot(Y-\mu_y)\right]\equiv\mathbb{E}(X\cdot Y)-\mu_x\cdot\mu_y$$

It becomes *zero* when X and Y are *independent*, but: zero covariance does *not* necessarily imply independence.

A related quantity is the correlation coefficient between X and Y:

$$\rho_{xy} = \frac{\text{Cov}(X,Y)}{\sigma_x \cdot \sigma y}$$

(this is the Greek letter 'rho'). The absolute value of this coefficient cannot be greater than 1.

Variance of $aX + bY + c$

is equal to

$$a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y)$$

Independence would make the last term *zero.*

Extended to a linear combination of *any number* of random variables:

$$\boxed{\begin{gathered}\text{Var}(a_1X_1 + a_2X_2 + ...a_kX_k + c) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + ... + a_k^2\text{Var}(X_k)\\ +2a_1a_2\text{Cov}(X_1,X_2) + 2a_1a_3\text{Cov}(X_1,X_3) + ... + 2a_{k-1}a_k\text{Cov}(X_{k-1},X_k)\end{gathered}}$$

Moment generating function

is defined by

$$M_x(t) \equiv \mathbb{E}\left[e^{t\,X}\right]$$

where t is an arbitrary (real) parameter.

Main results

1. $$\mathbb{E}(X^k) = M_x^{(k)}(t=0)$$

   or, in words, to get the k^{th} simple moment differentiate the corresponding MGF k times (with respect to t) and set t equal to zero.

2. For two *independent* RVs we have:

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

   This result can be extended to *any number* of mutually independent RVs:

   $$M_{X+Y+Z}(t) = M_X(t) \cdot M_Y(t) \cdot M_Z(t)$$

   etc.

3. And, finally

   $$M_{aX+c}(t) = e^{ct} \cdot M_X(at)$$

Probability generating function

is defined by

$$P_x(s) = \Pr(X=0) + \Pr(X=1)\ s + \Pr(X=2)\ s^2 + \Pr(X=3)\ s^3 + ...$$

is a somehow easier concept (applicable to integer-valued RVs only). We also have

$$P_{X+Y}(s) = P_X(s) \cdot P_Y(s)$$

Conditional expected value

$$\mathbb{E}(X|Y=\mathbf{j}) = \sum_i i \times \Pr(X=i \mid Y=\mathbf{j})$$

(summing over the corresponding conditional range of i values), etc.

Common discrete distributions

First, the UNIVARIATE type:

Binomial

Total number of successes in a series of n independent trials with two possible outcomes (success or failure, having probabilities of p and q, respectively).

$$f(i) = \binom{n}{i} p^i q^{n-i} \qquad \text{where} \qquad 0 \le i \le n$$

Expected value (mean):

$$np$$

Variance:

$$npq$$

Geometric

The number of trials to get the first success, in an independent series of trials.

$$f(i) = pq^{i-1} \qquad \text{where} \qquad i \ge 1$$

The mean

$$\frac{1}{p}$$

and variance:

$$\frac{1}{p}\left(\frac{1}{p}-1\right)$$

This time, we also have

$$F(j) = \Pr(X \le j) = 1 - q^j \qquad \text{where} \qquad j \ge 1$$

Negative Binomial

The **number of trials** until (and including) the k^{th} success is obtained. It is a sum of k *independent* random variables of the *geometric* type.

$$f(i) = \binom{i-1}{k-1} p^k q^{i-k} \equiv \binom{i-1}{i-k} p^k q^{i-k} \qquad \text{where} \qquad i \ge k$$

The mean

$$\frac{k}{p}$$

and variance:

$$\frac{k}{p}\left(\frac{1}{p}-1\right)$$

$$F(j) = 1 - \sum_{i=0}^{k-1} \binom{j}{i} p^i q^{j-i} \qquad \text{where} \qquad j \ge k$$

Hypergeometric

Suppose there are N objects, K of which have some *special* property. Of these N objects, n are *randomly* selected [SAMPLING WITHOUT REPLACEMENT]. X is the number of 'special' objects found in the sample.

$$f(i) = \frac{\binom{K}{i} \times \binom{N-K}{n-i}}{\binom{N}{n}} \qquad \text{where} \qquad \max(0, n-N+K) \le i \le \min(n, K)$$

The mean

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

and variance:

$$n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$$

Note the similarity (and difference) to the binomial npq formula.

Poisson

Assume that customers arrive at a store randomly, at a *constant* rate of φ per hour. X is the number of customers who will arrive during the next T hours. $\lambda = T \cdot \varphi$.

$$f(i) = \frac{\lambda^i}{i!}e^{-\lambda} \qquad \text{where} \qquad i \ge 0$$

Both the mean and the variance of this distribution are equal to λ.

The remaining two distributions are of the MULTIVARIATE type.

Multinomial

is an extension of the binomial distribution, in which each *trial* can result in 3 (or more) possible outcomes (not just S and F). The trials are repeated, independently, n times; this time we need three RVs X, Y and Z, which count the total number of outcomes of the first, second and third type, respectively.

$$\boxed{\Pr(X = i \cap Y = j \cap Z = k) = \binom{n}{i,j,k} p_x^i \, p_y^j \, p_z^k}$$

for any non-negative integer values of i, j, k which add up to n. This formula can be easily extended to the case of 4 or more possible outcomes.

The **marginal distribution** of X is obviously *binomial* (with n and $p \equiv p_x$ being the two parameters).

We also need

$$\text{Cov}(X, Y) = -np_x \, p_y$$

etc.

Multivariate Hypergeometric

is a simple extension of the univariate hypergeometric distribution, to the case of having thee (or more) types of objects. We now assume that the total number of objects of each type is K_1, K_2 and K_3, where $K_1 + K_2 + K_3 = N$.

$$\Pr(X = i \cap Y = j \cap Z = k) = \frac{\binom{K_1}{i}\binom{K_2}{j}\binom{K_3}{k}}{\binom{N}{n}}$$

where X, Y and Z count the number of objects of Type 1, 2 and 3, respectively, in the sample. Naturally, $i+j+k=n$. Otherwise, i, j and k can be any non-negative integers for which the above expression is meaningful (i.e. no negative factorials).

The **marginal distribution** of X (and Y, and Z) is *univariate* hypergeometric (of the old kind) with obvious parameters.

$$\operatorname{Cov}(X,Y) = -n \cdot \frac{K_1}{N} \cdot \frac{K_2}{N} \cdot \frac{N - n}{N-1}$$

Continuous Random Variables

Any real value from a certain interval can happen. $\Pr(X = x)$ is always equal to zero (we have lost the individual probabilities)! Instead, we use

Univariate probability density function (pdf)

formally defined by

$$f(x) \equiv \lim_{\varepsilon\to 0} \frac{\Pr(x \le X < x+\varepsilon)}{\varepsilon}$$

Given $f(x)$, we can compute the probability of any *interval* of values:

$$\Pr(a < X < b) = \int_a^b f(x)\,dx$$

Note that $f(x)$ is frequently defined in a piecewise manner.

Distribution Function

$$F(x) \equiv \Pr(X \le x) = \int_{-\infty}^{x} f(u)\,du$$

which is quite crucial to us now (without it, we cannot compute probabilities).

Bivariate (multivariate) pdf

$$f(x,y) = \lim_{\substack{\varepsilon\to 0\\ \delta\to 0}} \frac{\Pr(x \le X < x+\varepsilon \cap y \le Y < y+\delta)}{\varepsilon \cdot \delta}$$

which implies that the **probability** of (X,Y)-values falling inside a 2D **region** $\mathcal{A}$ is computed by

$$\iint_{\mathcal{A}} f(x,y)\,dxdy$$

Similarly for three or more variables.

Marginal Distributions

Given a bivariate pdf $f(x,y)$, we can eliminate Y and get the marginal pdf of X by

$$f(x) = \int_{\text{All } y|x} f(x,y)\,dy$$

The *integration* is over the *conditional* range of y given x, the result is valid in the *marginal* range of x.

Conditional Distribution

is the distribution of X *given* that Y has been observed to result in a specific value $\mathbf{y}$. The corresponding conditional pdf of X is computed by

$$f(x \mid Y = \mathbf{y}) = \frac{f(x, \mathbf{y})}{f(\mathbf{y})}$$

valid in the corresponding conditional range of x values.

Mutual Independence

implies that $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$, with all the other consequences (same as in the discrete case), most notably $f(x \mid Y = \mathbf{y}) = f_X(x)$.

Expected value

of a continuous RV X is computed by

$$\mathbb{E}(X) = \int\limits_{\text{All } x} x \cdot f(x)\, dx$$

Similarly:

$$\mathbb{E}[g(X)] = \int\limits_{\text{All } x} g(x) \cdot f(x)\, dx$$

where $g(..)$ is an arbitrary function.

In the bivariate case:

$$\mathbb{E}[g(X, Y)] = \iint\limits_{\mathcal{R}} g(x, y) \cdot f(x, y)\, dx\, dy$$

Simple *moments*, central moments, variance, covariance, etc. are defined in exactly *same* manner as in the discrete case. Also, all previous formulas for dealing with *linear combinations* of RVs (expected value, variance, covariance) still hold, without change.

Also, the Moment Generating Function is defined in the analogous manner as is defined via:

$$M_x(t) \equiv \mathbb{E}(e^{tX}) = \int\limits_{\text{All } x} e^{tx} \cdot f(x)\, dx$$

with all the previous results still being correct.

Common Continuous Distributions

First, the univariate case:

Name	**Notation**	**Range**	$f(x)$
Uniform	$\mathcal{U}(a,b)$	$a < x < b$	$\frac{1}{b-a}$
Normal	$\mathcal{N}(\mu,\sigma)$	$-\infty < x < \infty$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$
Exponential	$\mathcal{E}(\beta)$	$x > 0$	$\frac{1}{\beta}\exp\left[-\frac{x}{\beta}\right]$
Gamma	$\mathsf{gamma}(\alpha,\beta)$	$x > 0$	$\frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}\exp\left[-\frac{x}{\beta}\right]$
Beta	$\mathsf{beta}(k,m)$	$0 < x < 1$	$\frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)}\cdot x^{k-1}(1-x)^{m-1}$
Chi-square	χ^2_m	$x > 0$	$\frac{x^{m/2-1}}{\Gamma(m/2)2^{m/2}}\exp\left[-\frac{x}{2}\right]$
Student	t_m	$-\infty < x < \infty$	$\frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})\sqrt{m\pi}}\cdot\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}}$
Fisher	$\mathsf{F}_{k,m}$	$x > 0$	$\frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\,\Gamma(\frac{m}{2})}\left(\frac{k}{m}\right)^{\frac{k}{2}}\cdot\frac{x^{\frac{k}{2}-1}}{(1+\frac{k}{m}x)^{\frac{k+m}{2}}}$
Cauchy	$\mathcal{C}(a,b)$	$-\infty < x < \infty$	$\frac{b}{\pi}\cdot\frac{1}{b^2+(y-a)^2}$

Name	$F(x)$	**Mean**	**Variance**
Uniform	$\frac{x-a}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	Tables	μ	σ^2
Exponential	$1-\exp\left[-\frac{x}{\beta}\right]$	β	β^2
Gamma	Integer α only	$\alpha\beta$	$\alpha\beta^2$
Beta	Integer k, m only	$\frac{k}{k+m}$	$\frac{k\,m}{(k+m+1)(k+m)^2}$
Chi-square	Tables	m	$2m$
Student	Tables	0	$\frac{m}{m-2}$
Fisher	Tables	$\frac{m}{m-2}$	$\frac{2\,m^2(k+m-2)}{(m-2)^2\,(m-4)\,k}$
Cauchy	$\frac{1}{2}+\frac{1}{\pi}\arctan(\frac{y-a}{b})$	×	×

We need only one bivariate example:

Bivariate Normal distribution has, in general, 5 parameters, μ_x, μ_y, σ_x, σ_y, and ρ. Its joint pdf can be simplified by introducing $Z_1 \equiv \frac{X-\mu_x}{\sigma_x}$ and $Z_2 \equiv \frac{Y-\mu_y}{\sigma_y}$ (standardized RVs), for which

$$f(z_1,z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}}\cdot\exp\left[-\frac{z_1^2-2\rho z_1 z_2+z_2^2}{2(1-\rho^2)}\right]$$

Its marginal distributions are both Normal, so it the conditional distribution:

$$\mathrm{Distr}(X|Y=\mathbf{y}) \equiv \mathcal{N}(\mu_x+\sigma_x\rho\frac{\mathbf{u}-\mu_y}{\sigma_y}, \sigma_x\sqrt{1-\rho^2})$$

Transforming Random Variables

i.e. if $Y = g(X)$, where X has a given distribution, what is the distribution of Y?

Two techniques to deal with this, one uses $F(x)$, the other one $f(x)$ - this only for one-to-one transformations.

This can be generalized to: Given the joint distribution of X and Y (usually independent), find the distribution of $g(X,Y)$.

Examples

g:	Distribution:
$-\beta \cdot \ln \mathcal{U}(0,1)$	$\mathcal{E}(\beta)$
$\mathcal{N}(0,1)^2$	χ_1^2
$\mathcal{N}_1(0,1)^2 + \mathcal{N}_2(0,1)^2 + ... + \mathcal{N}_m(0,1)^2$	χ_m^2
$\overline{\mathcal{C}(a,b)}$	$\mathcal{C}(a,b)$
$\frac{\mathcal{E}_1(\beta)}{\mathcal{E}_1(\beta)+\mathcal{E}_2(\beta)}$	$\mathcal{U}(0,1)$
$\frac{\mathsf{gamma}_1(k,\beta)}{\mathsf{gamma}_1(k,\beta)+\mathsf{gamma}_2(m,\beta)}$	$\mathsf{beta}(k,m)$
$\frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi_m^2}{m}}}$	t_m
$\frac{\chi_k^2}{\chi_m^2} \cdot \frac{m}{k}$	$\mathsf{F}_{k,m}$

Chapter 2 TRANSFORMING RANDOM VARIABLES

of *continuous* type *only* (the less interesting discrete case was dealt with earlier).

The main issue of this chapter is: Given the distribution of X, find the distribution of $Y \equiv \frac{1}{1+X}$ (an expression involving X). Since only one 'old' RV variable (namely X) appear in the definition of the 'new' RV, we call this a UNIVARIATE transformation. Eventually, we must also deal with the so called BIVARIATE transformations of two 'old' RVs (say X and Y), to find the distribution of a 'new' RV, say $U \equiv \frac{X}{X+Y}$ (or any other expression involving X and Y). Another simple example of this bivariate type is finding the distribution of $V \equiv X + Y$ (i.e. we will finally learn how to *add* two random variables).

Let us first deal with the

Univariate transformation

There are two basic techniques for constructing the new distribution:

Distribution-Function (F) Technique

which works as follows:

When the new random variable Y is defined as $g(X)$, we find its distribution function $F_Y(y)$ by computing $\Pr(Y < y) = \Pr[g(X) < y]$. This amounts to *solving* the $g(X) < y$ *inequality* for X [usually resulting in an interval of values], and then integrating $f(x)$ over this interval [or, equivalently, substituting into $F(x)$].

EXAMPLES:

1. Consider $X \in \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$ [this corresponds to a spinning wheel with a two-directional 'pointer', say a laser beam, where X is the pointer's angle from a fixed direction when the wheel stops spinning]. We want to know the distribution of $Y = b\tan(X) + a$ [this represents the location of a dot our laser beam would leave on a screen placed b units from the wheel's center, with a scale whose origin is a units off the center]. Note that Y can have any real value.

   Solution: We start by writing down $F_X(x) =$ [in our case] $\frac{x+\frac{\pi}{2}}{\pi} \equiv \frac{x}{\pi} + \frac{1}{2}$ when $-\frac{\pi}{2} < x < \frac{\pi}{2}$. To get $F_Y(y)$ we need: $\Pr[b\tan(X) + a < y] = \Pr[X < \arctan(\frac{y-a}{b})] = F_X[\arctan(\frac{y-a}{b})] = \frac{1}{\pi}\arctan(\frac{y-a}{b}) + \frac{1}{2}$ where $-\infty < y < \infty$. Usually, we can relate better to the corresponding $f_Y(y)$ [which tells us what is likely and what is not] $= \frac{1}{\pi b} \cdot \frac{1}{1+(\frac{y-a}{b})^2} =$

   $$\frac{b}{\pi} \cdot \frac{1}{b^2 + (y-a)^2} \qquad (f)$$

   [any real y]. Graphically, this function looks very similar to the Normal pdf (also a 'bell-shaped' curve), but in terms of its properties, the new distribution turns out to be totally different from Normal, [as we will see later].

The name of this new distribution is **Cauchy** [notation: $\mathcal{C}(a,b)$]. Since the $\int\limits_{-\infty}^{\infty} y \cdot f_Y(y)\,dy$ integral leads to $\infty - \infty$, the Cauchy distribution does *not* have a mean (consequently, its variance is infinite). Yet it possesses a clear *center* (at $y = a$) and *width* ($\pm b$). These are now identified with the *median* $\tilde{\mu}_Y = a$ [verify by solving $F_Y(\tilde{\mu}) = \frac{1}{2}$] and the so called *semi-inter-quartile range* (QUARTILE DEVIATION, for short) $\frac{Q_U - Q_L}{2}$ where Q_U and Q_L are the UPPER and LOWER QUARTILES [defined by $F(Q_U) = \frac{3}{4}$ and $F(Q_L) = \frac{1}{4}$]. One can easily verify that, in this case, $Q_L = a - b$ and $Q_U = a + b$ [note that the semi-inter-quartile range contains exactly 50% of all probability], thus the quartile deviation equals to b. The most *typical* (’standardized’) *case* of the Cauchy distribution is $\mathcal{C}(0,1)$, whose pdf equals

$$f(y) = \frac{1}{\pi} \cdot \frac{1}{1+y^2}$$

Its ’rare’ ($<\frac{1}{2}\%$) values start at ± 70, we need to go beyond ± 3000 to reach ’extremely unlikely’ ($<10^{-6}$), and only ∓ 300 billion become ’practically impossible’ (10^{-12}). Since the mean does not exist, the central limit theorem breaks down [it is no longer true that $\bar{Y} \to \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$, there is no μ and σ is infinite]. Yet, $\bar{Y}$ must have some well defined distribution. We will discover what that distribution is in the next section.

2. Let X have its pdf defined by $f(x) = 6x(1-x)$ for $0 < x < 1$.Find the pdf of $Y = X^3$.

   Solution: First we realize that $0 < Y < 1$. Secondly, we find $F_X(x) = 6\int\limits_0^x (x - x^2)\,dx = 6(\frac{x^2}{2} - \frac{x^3}{3}) = 3x^2 - 2x^3$. And finally: $F_Y(y) \equiv \Pr(Y < y) = \Pr(X^3 < y) = \Pr(X < y^{\frac{1}{3}}) = F_X(y^{\frac{1}{3}}) = 3y^{\frac{2}{3}} - 2y$.This easily converts to $f_Y(y) = 2y^{-\frac{1}{3}} - 2$ where $0 < y < 1$ [zero otherwise]. (Note that when $y \to 0$ this pdf becomes infinite, which is OK).

3. Let $X \in \mathcal{U}(0,1)$. Find and identify the distribution of $Y = -\ln X$ (its range is obviously $0 < y < \infty$).

   Solution: First we need $F_X(x) = x$ when $0 < x < 1$. Then: $F_Y(y) = \Pr(-\ln X < y) = \Pr(X > e^{-y})$ [note the sign reversal] $= 1 - F_X(e^{-y}) = 1 - e^{-y}$ where $y > 0$ ($\Rightarrow f_Y(y) = e^{-y}$). This can be easily identified as the exponential distribution with the mean of 1 [note that $Y = -\beta \cdot \ln X$ would result in the exponential distribution with the mean equal to β].

4. If $Z \in \mathcal{N}(0,1)$, what is the distribution of $Y = Z^2$.

   Solution: $F_Y(y) = \Pr(Z^2 < y) = \Pr(-\sqrt{y} < Z < \sqrt{y})$ [right?] $= F_Z(\sqrt{y}) - F_Z(\sqrt{y})$. Since we don’t have an explicit expression for $F_Z(z)$ it would appear that we are stuck at this point, but we can get the corresponding $f_Y(y)$ by a simple differentiation: $\frac{dF_Z(\sqrt{y})}{dy} - \frac{dF_Z(-\sqrt{y})}{dy} = \frac{1}{2}y^{-\frac{1}{2}} f_Z(\sqrt{y}) + \frac{1}{2}y^{-\frac{1}{2}} f_Z(-\sqrt{y}) =$

$\dfrac{y^{-\frac{1}{2}}e^{-\frac{y}{2}}}{\sqrt{2\pi}}$ where $y > 0$. This can be identified as the *gamma* distribution with $\alpha = \frac{1}{2}$ and $\beta = 2$ [the normalizing constant is equal to $\Gamma(\frac{1}{2}) \cdot 2^{\frac{1}{2}} = \sqrt{2\pi}$, check].

Due to its importance, this distribution has yet another name, it is called the **chi-square distribution** with *one degree of freedom*, or χ_1^2 for short. It has the expected value of $(\alpha \cdot \beta =)$ 1, its variance equals $(\alpha \cdot \beta^2 =)$ 2, and the MGF is $M(t) = \frac{1}{\sqrt{1-2t}}$. ■

General Chi-square distribution

(This is an extension of the previous example). We want to investigate the RV defined by $U = Z_1^2 + Z_2^2 + Z_3^2 + ... + Z_n^2$, where Z_1, Z_2, $Z_3, ...Z_n$ are *independent* RVs from the $\mathcal{N}(0,1)$ distribution. Its **MGF** must obviously equal to $M(t) = \dfrac{1}{(1-2t)^{\frac{n}{2}}}$; we can thus identify its distribution as *gamma*, with $\alpha = \frac{n}{2}$ and $\beta = 2$ ($\Rightarrow$ mean $= n$, variance $= 2n$). Due to its importance, it is also called the chi-square distribution with n (integer) degrees of freedom (χ_n^2 for short).

Probability-Density-Function (f) Technique

is a bit faster and usually somehow easier (technically) to carry out, but it works for *one-to-one* transformations *only* (e.g. it would not work in our last $Y = Z^2$ example). The procedure consists of three simple steps:

(i) Express X (the 'old' variable) in terms of y the 'new' variable [getting an expression which involves only Y].

(ii) Substitute the result [we will call it $x(y)$, switching to small letters] for the argument of $f_X(x)$, getting $f_X[x(y)]$ – a function of y!

(iii) Multiply this by $\left|\frac{dx(y)}{dy}\right|$. The result is the pdf of Y. ■

In **summary**

$$f_Y(y) = f_X[x(y)] \cdot \left|\frac{dx(y)}{dy}\right|$$

EXAMPLES (we will redo the first three examples of the previous section):

1. $X \in \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$ and $Y = b\tan(X) + a$.

   Solution: (i) $x = \arctan(\frac{y-a}{b})$, (ii) $\frac{1}{\pi}$, (iii) $\frac{1}{\pi} \cdot \frac{1}{b} \cdot \frac{1}{1+(\frac{y-a}{b})^2} = \frac{b}{\pi} \cdot \frac{1}{b^2+(y-a)^2}$ where $-\infty < y < \infty$ [check].

2. $f(x) = 6x(1-x)$ for $0 < x < 1$ and $Y = X^3$.

   Solution: (i) $x = y^{1/3}$, (ii) $6y^{1/3}(1-y^{1/3})$, (iii) $6y^{1/3}(1-y^{1/3}) \cdot \frac{1}{3}y^{-2/3} = 2(y^{-1/3}-1)$ when $0 < y < 1$ [check].

3. $X \in \mathcal{U}(0,1)$ and $Y = -\ln X$.

   Solution: (i) $x = e^{-y}$, (ii) 1, (iii) $1 \cdot e^{-y} = e^{-y}$ for $y > 0$ [check].

This does appear to be a fairly fast way of obtaining $f_Y(y)$. ■

And now we extend all this to the

Bivariate transformation

Distribution-Function Technique

follows essentially the same pattern as the univariate case:

The new random variable Y is now defined in terms of two 'old' RVs, say X_1 and X_2, by $y \equiv g(X_1, X_2)$. We find $F_Y(y) = \Pr(Y < y) = \Pr[g(X_1, X_2) < y]$ by realizing that the $g(X_1, X_2) < y$ inequality (for X_1 and X_2, y is considered fixed) will now result in some *2-D region*, and then integrating $f(x_1, x_2)$ over this region.

Thus, the technique is simple in principle, but often quite involved in terms of technical details.

EXAMPLES:

1. Suppose that X_1 and X_2 are independent RVs, both from $\mathcal{E}(1)$, and $Y = \dfrac{X_2}{X_1}$.

   Solution: $F_Y(y) = \Pr\left(\dfrac{X_2}{X_1} < y\right) = \Pr(X_2 < yX_1) = \iint\limits_{0<x_2<yx_1} e^{-x_1-x_2}\,dx_1\,dx_2 = \int\limits_0^\infty e^{-x_1} \int\limits_0^{yx_1} e^{-x_2}\,dx_2\,dx_1 = \int\limits_0^\infty e^{-x_1}(1-e^{-yx_1})\,dx_1 = \int\limits_0^\infty (e^{-x_1} - e^{-x_1(1+y)})\,dx_1 = 1 - \dfrac{1}{1+y}$, where $y > 0$. This implies that $f_Y(y) = \dfrac{1}{(1+y)^2}$ when $y > 0$. (The median $\tilde{\mu}$ of this distribution equals to 1, the lower and upper quartiles are $Q_L = \dfrac{1}{3}$ and $Q_U = 3$).

2. This time Z_1 and Z_2 are independent RVs from $\mathcal{N}(0,1)$ and $Y = Z_1^2 + Z_2^2$ [here, we know the answer: χ_2^2, let us proceed anyhow].

   Solution: $F_Y(y) = \Pr(Z_1^2 + Z_2^2 < y) = \frac{1}{2\pi} \iint\limits_{z_1^2+z_2^2<y} e^{-\frac{z_1^2+{}_2^2}{2}}\,dz_1\,dz_2 = \frac{1}{2\pi} \int\limits_0^{2\pi} \int\limits_0^{\sqrt{y}} e^{-\frac{r^2}{2}} \cdot r\,dr\,d\theta =$[substitution: $w = \frac{r^2}{2}$] $\int\limits_0^{\frac{y}{2}} e^{-w}\,dw = 1 - e^{-\frac{y}{2}}$ where (obviously) $y > 0$. This is the *exponential* distribution with $\beta = 2$ [not χ_2^2 as expected, how come?]. It does not take long to realize that the two distributions are identical.

3. (**Sum of two independent RVs**): Assume that X_1 and X_2 are independent RVs from a distribution having L and H as its lowest and highest possible value, respectively. Find the distribution of $X_1 + X_2$ [finally learning how to *add* two RVs!].

   Solution: $F_Y(y) = \Pr(X_1 + X_2 < y) = \iint\limits_{\substack{x_1+x_2<y \\ L<x_1,x_2<H}} f(x_1) \cdot f(x_2)\,dx_1\,dx_2 =$
   $$\begin{cases} \int\limits_L^{y-L} \int\limits_L^{y-x_1} f(x_1) \cdot f(x_2)\,dx_2 dx_1 & \text{when } y < L+H \\ 1 - \int\limits_{y-H}^{H} \int\limits_{y-x_1}^{H} f(x_1) \cdot f(x_2)\,dx_2 dx_1 & \text{when } y > L+H \end{cases}$$
   . Differentiating this with respect to y (for the first line, this amounts to: substituting $y - L$ for x_1

and dropping the dx_1 integration – contributing zero in this case – plus: substituting $y - x_1$ for x_2 and dropping dx_2; same for the first line, except that we have to *subtract* the second contribution) results in $f_Y(y) =$ $\begin{cases} \int\limits_{L}^{y-L} f(x_1)\cdot f(y-x_1)\,dx_1 \text{ when } y < L+H \\ \int\limits_{y-H}^{H} f(x_1)\cdot f(y-x_1)\,dx_1 \text{ when } y > L+H \end{cases}$ or, equivalently,

$$f_Y(y) = \int\limits_{\max(L,y-H)}^{\min(H,y-L)} f(x)\cdot f(y-x)\,dx$$

where the y-range is obviously $2L < y < 2H$. The right hand side of the last formula is sometimes referred to as the CONVOLUTION of two pdfs (in general, the two fs may be distinct).

Examples:

- In the specific case of the **uniform** $\mathcal{U}(0,1)$ distribution, the last formula yields, for the pdf of $Y \equiv X_1 + X_2$:
$f_Y(y) = \int\limits_{\max(0,y-1)}^{\min(1,y)} dx = \begin{cases} \int\limits_0^y dx = y & \text{when} \quad 0 < y < 1 \\ \int\limits_{y-1}^{1} dx = 2-y & \text{when} \quad 1 < y < 2 \end{cases}$ ['triangular' distribution].

- Similarly, for the 'standardized' **Cauchy** distribution $\left[f(x) = \frac{1}{\pi}\cdot\frac{1}{1+x^2}\right]$, we get: $f_{X_1+X_2}(y) = \frac{1}{\pi^2}\int\limits_{-\infty}^{\infty}\frac{1}{1+x^2}\cdot\frac{1}{1+(y-x)^2}\,dx = \frac{2}{\pi}\cdot\frac{1}{4+y^2}$ [where $-\infty < y < \infty$].

  The last result can be easily converted to the pdf of $\bar{X} = \frac{X_1+X_2}{2}$ [the *sample mean* of the two random values], yielding $f_{\bar{X}}(\bar{x}) = \frac{2}{\pi}\cdot\frac{1}{4+(2\bar{x})^2}\cdot 2 = \frac{1}{\pi}\cdot\frac{1}{1+\bar{x}^2}$. Thus, the sample mean $\bar{X}$ has the *same* Cauchy distribution as do the two individual observations (the result can be extended to *any number* of observations). We knew that the Central Limit Theorem $[\bar{X} \mathrel{\widetilde{\in}} \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})]$ would not apply to this case, but the actual distribution of $\bar{X}$ still comes as a big surprise. This implies that the sample mean of even millions of values (from a Cauchy distribution) cannot estimate the center of the distribution any better than a *single* observation [one can verify this by actual simulation]. Yet, one feels that there must be a way of substantially improving the estimate (of the location of a laser gun hidden behind a screen) when going from a single observation to a large sample. Yes, there is, if one does not use the sample *mean* but something else; later on we discover that the sample *median* will do just fine. ■

Pdf (Shortcut) Technique

works a bit faster, even though it may *appear* more complicated, as it requires the following (several) steps:

1. The procedure can work only for **one-to-one** ('invertible') transformations. This implies that the new RV $Y \equiv g(X_1, X_2)$ must be accompanied by yet another *arbitrarily* chosen function of X_1 and/or X_2 [the original Y will be called Y_1, and the auxiliary one Y_2, or vice versa]. We usually choose this second (auxiliary) function in the simplest possible manner, i.e. we make it equal to X_2 (or X_1):

2. **Invert** the transformation, i.e. solve the two equations $y_1 = g(x_1, x_2)$ and $y_2 = x_2$ for x_1 and x_2 (in terms of y_1 and y_2). Getting a unique solution guarantees that the transformation is one-to-one.

3. **Substitute** this solution $x_1(y_1, y_2)$ and $x_2(y_2)$ into the joint pdf of the 'old' X_1, X_2 pair (yielding a function of y_1 and y_2).

4. Multiply this function by the transformation's **Jacobian** $\begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$. The result is the joint pdf of Y_1 and Y_2. At the same time, establish the region of possible (Y_1, Y_2) values in the (y_1, y_2)-plane [this is often the most difficult part of the procedure].

5. Eliminate Y_2 [the 'phoney', auxiliary RV introduced to help us with the inverse] by integrating it out (finding the Y_1 **marginal**). Don't forget that you must integrate over the *conditional* range of y_2 *given* y_1.

EXAMPLES:

1. X_1, $X_2 \in \mathcal{E}(1)$, independent; $Y = \frac{X_1}{X_1+X_2}$ [the time of the first 'catch' relative to the time needed to catch two fishes].

   Solution: $Y_2 = X_2 \Rightarrow x_2 = y_2$ and $x_1 y_1 + x_2 y_1 = x_1 \Rightarrow x_1 = \frac{y_1 \cdot y_2}{1-y_1}$. Substitute into $e^{-x_1-x_2}$ getting $e^{-y_2\left(1+\frac{y_1}{1-y_1}\right)} = e^{-\frac{y_2}{1-y_1}}$, multiply by $\begin{vmatrix} y_2 \frac{1-y_1+y_1}{(1-y_1)^2} & \frac{y_1}{1-y_1} \\ 0 & 1 \end{vmatrix} = \frac{y_2}{(1-y_1)^2}$ getting $f(y_1, y_2) = \frac{y_2}{(1-y_1)^2}\, e^{-\frac{y_2}{1-y_1}}$ with $0 < y_1 < 1$ and $y_2 > 0$. Eliminate Y_2 by $\int\limits_0^\infty \frac{y_2}{(1-y_1)^2}\, e^{-\frac{y_2}{1-y_1}}\, dy_2 = \frac{1}{(1-y_1)^2} \cdot (1-y_1)^2 \equiv 1$ when $0 < y_1 < 1$ [recall the $\int\limits_0^\infty x^k\, e^{-\frac{x}{a}}\, dx = k! \cdot a^{k+1}$ formula]. The distribution of Y is thus $\mathcal{U}(0,1)$. Note that if we started with X_1, $X_2 \in \mathcal{E}(\beta)$ instead of $\mathcal{E}(1)$, the result would have been the same since this new $Y = \frac{X_1}{X_1+X_2} \equiv \frac{\frac{X_1}{\beta}}{\frac{X_1}{\beta}+\frac{X_2}{\beta}}$ where $\frac{X_1}{\beta}$ and $\frac{X_2}{\beta} \in \mathcal{E}(1)$ [this can be verified by a simple MGF argument].

2. Same X_1 and X_2 as before, $Y = \frac{X_2}{X_1}$.

   Solution: This time we reverse the labels: $Y_1 \equiv X_1$ and $Y_2 = \frac{X_2}{X_1} \Rightarrow x_1 = y_1$ and $x_2 = y_1 \cdot y_2$. Substitute into $e^{-x_1-x_2}$ to get $e^{-y_1(1+y_2)}$, times $\begin{vmatrix} 1 & 0 \\ y_2 & y_1 \end{vmatrix} = y_1$ gives the joint pdf for $y_1 > 0$ and $y_2 > 0$. Eliminate y_1 by $\int\limits_0^\infty y_1 e^{-y_1(1+y_2)}\, dy_1 =$

$\frac{1}{(1+y_2)^2}$, where $y_2 > 0$. Thus, $f_Y(y) = \frac{1}{(1+y)^2}$ with $y > 0$ [check, we have solved this problem before].

3. In this example we introduce the so called Beta **distribution**

Let X_1 and X_2 be independent RVs from the gamma distribution with parameters (k, β) and (m, β) respectively, and let $Y_1 = \frac{X_1}{X_1+X_2}$.

Solution: Using the argument of Example 1 one can show that β 'cancels out', and we can assume that $\beta = 1$ without affecting the answer. The definition of Y_1 is also the same as in Example 1 $\Rightarrow x_1 = \frac{y_1 y_2}{1-y_1}$, $x_2 = y_2$, and the Jacobian $=$ $\frac{y_2}{(1-y_1)^2}$. Substituting into $f(x_1, x_2) = \frac{x_1^{k-1} x_2^{m-1} e^{-x_1-x_2}}{\Gamma(k)\cdot\Gamma(m)}$ and multiplying by the Jacobian yields $f(y_1, y_2) = \dfrac{y_1^{k-1} y_2^{k-1} y_2^{m-1} e^{-\frac{y_2}{1-y_1}}}{\Gamma(k)\Gamma(m)(1-y_1)^{k-1}} \cdot \dfrac{y_2}{(1-y_1)^2}$ for $0 < y_1 < 1$ and $y_2 > 0$. Integrating over y_2 results in: $\dfrac{y_1^{k-1}}{\Gamma(k)\Gamma(m)(1-y_1)^{k+1}} \int\limits_0^\infty y_2^{k+m-1} e^{-\frac{y_2}{1-y_1}}\, dy_2 =$

$$\frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \cdot y_1^{k-1}(1-y_1)^{m-1} \qquad \text{(f)}$$

where $0 < y_1 < 1$.

This is the **pdf** of a new two-parameters (k and m) distribution which is called beta. Note that, as a by-product, we have effectively proved the following formula: $\int\limits_0^1 y^{k-1}(1-y)^{m-1} dy = \frac{\Gamma(k)\cdot\Gamma(m)}{\Gamma(k+m)}$ for any $k, m > 0$. This enables us to find the distribution's **mean**: $\mathbb{E}(Y) = \frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \int\limits_0^1 y^k (1-y)^{m-1}\, dy =$ $\frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \cdot \frac{\Gamma(k+1)\cdot\Gamma(m)}{\Gamma(k+m+1)} =$

$$\frac{k}{k+m} \qquad \text{(mean)}$$

and similarly $\mathbb{E}(Y^2) = \frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \int\limits_0^1 y^{k+1}(1-y)^{m-1}\, dy = \frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \cdot \frac{\Gamma(k+2)\cdot\Gamma(m)}{\Gamma(k+m+2)} =$ $\frac{(k+1)\, k}{(k+m+1)\,(k+m)} \Rightarrow Var(Y) = \frac{(k+1)\, k}{(k+m+1)\,(k+m)} - (\frac{k}{k+m})^2 =$

$$\frac{k\, m}{(k+m+1)\,(k+m)^2} \qquad \text{(variance)}$$

Note that the distribution of $1 - Y \equiv \frac{X_2}{X_1+X_2}$ is also beta (why?) with parameters m and k [reversed].

We learn how to compute related **probabilities** in the following set of Examples:

(a) $\Pr(X_1 < \frac{X_2}{2})$ where X_1 and X_2 have the gamma distribution with parameters $(4, \beta)$ and $(3, \beta)$ respectively [this corresponds to the probability that Mr.A catches 4 fishes in less than half the time Mr.B takes to catch 3].

Solution: $\Pr(2X_1 < X_2) = \Pr(3X_1 < X_1 + X_2) = \Pr(\frac{X_1}{X_1+X_2} < \frac{1}{3}) = \frac{\Gamma(4+3)}{\Gamma(4)\cdot\Gamma(3)} \int\limits_0^{\frac{1}{3}} y^3(1-y)^2 dy = 60 \times \left[\frac{y^4}{4} - 2\frac{y^5}{5} + \frac{y^6}{6}\right]_{y=0}^{\frac{1}{3}} = 10.01\%$.

(b) Evaluate $\Pr(Y < 0.4)$ where Y has the beta distribution with parameters $(\frac{3}{2}, 2)$ [half-integer values are not unusual, as we learn shortly].

Solution: $\frac{\Gamma(\frac{7}{2})}{\Gamma(\frac{3}{2})\cdot\Gamma(2)} \int\limits_0^{0.4} y^{\frac{1}{2}}(1-y)\, dy = \frac{5}{2} \cdot \frac{3}{2} \cdot \left[\frac{y^{\frac{3}{2}}}{\frac{3}{2}} - \frac{y^{\frac{5}{2}}}{\frac{5}{2}}\right]_{y=0}^{0.4} = 48.07\%$.

(c) Evaluate $\Pr(Y < 0.7)$ where $Y \in \mathsf{beta}(4, \frac{5}{2})$.

Solution: This equals [it is more convenient to have the half-integer first] $\Pr(1-Y > 0.3) = \frac{\Gamma(\frac{13}{2})}{\Gamma(\frac{5}{2})\cdot\Gamma(4)} \int\limits_{0.3}^{1} u^{\frac{3}{2}}(1-u)^3\, du = \frac{\frac{11}{2}\cdot\frac{9}{2}\cdot\frac{7}{2}\cdot\frac{5}{2}}{3!} \left[\frac{y^{\frac{5}{2}}}{\frac{5}{2}} - 3\frac{y^{\frac{7}{2}}}{\frac{7}{2}} + 3\frac{y^{\frac{9}{2}}}{\frac{9}{2}} - \frac{y^{\frac{11}{2}}}{\frac{11}{2}}\right]_{y=0.3}^{1} = 1 - 0.3522 = 64.78\%$.

d $\Pr(Y < 0.5)$ when $Y \in \mathsf{beta}(\frac{3}{2}, \frac{1}{2})$.

Solution: $\frac{\Gamma(2)}{\Gamma(\frac{3}{2})\cdot\Gamma(\frac{1}{2})} \int\limits_0^{0.5} y^{\frac{1}{2}}(1-y)^{-\frac{1}{2}} dy = 18.17\%$ (Maple).

4. In this example we introduce the so called **Student's** or **<u>t-distribution</u>**

[notation: t_n, where n is called 'degrees of freedom' $-$ the only parameter]. We start with two independent RVs $X_1 \in \mathcal{N}(0,1)$ and $X_2 \in \chi^2_n$, and introduce a new RV by $Y_1 = \dfrac{X_1}{\sqrt{\frac{X_2}{n}}}$.

To get its **pdf** we take $Y_2 \equiv X_2$, solve for $x_2 = y_2$ and $x_1 = y_1 \cdot \sqrt{\frac{y_2}{n}}$, substitute into $f(x_1, x_2) = \dfrac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} \cdot \dfrac{x_2^{\frac{n}{2}-1} e^{-\frac{x_2}{2}}}{\Gamma(\frac{n}{2}) \cdot 2^{\frac{n}{2}}}$ and multiply by $\begin{vmatrix} \sqrt{\frac{y_2}{n}} & \frac{1}{2} \cdot \frac{y_1}{\sqrt{n y_2}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{y_2}{n}}$ to get $f(y_1, y_2) = \dfrac{e^{-\frac{y_1^2 y_2}{2n}}}{\sqrt{2\pi}} \cdot \dfrac{y_2^{\frac{n}{2}-1} e^{-\frac{y_2}{2}}}{\Gamma(\frac{n}{2}) \cdot 2^{\frac{n}{2}}} \cdot \sqrt{\frac{y_2}{n}}$ where $-\infty < y_1 < \infty$ and $y_2 > 0$. To eliminate y_2 we integrate: $\dfrac{1}{\sqrt{2\pi}\Gamma(\frac{n}{2})\, 2^{\frac{n}{2}} \sqrt{n}} \int\limits_0^{\infty} y_2^{\frac{n-1}{2}} e^{-\frac{y_2}{2}(1+\frac{y_1^2}{n})} dy_2 =$

$$\frac{\Gamma(\frac{n+1}{2})\, 2^{\frac{n+1}{2}}}{\sqrt{2\pi}\Gamma(\frac{n}{2})\, 2^{\frac{n}{2}} \sqrt{n} \left(1 + \frac{y_1^2}{n}\right)^{\frac{n+1}{2}}} =$$

$$\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \cdot \frac{1}{\left(1 + \frac{y_1^2}{n}\right)^{\frac{n+1}{2}}} \qquad (f)$$

with $-\infty < y_1 < \infty$. Note that when $n = 1$ this gives $\frac{1}{\pi} \cdot \frac{1}{1+y_1^2}$ (Cauchy), when $n \to \infty$ the second part of the formula tends to $e^{-\frac{y_1^2}{2}}$ which is, up to the normalizing constant, the pdf of $\mathcal{N}(0,1)$ [implying that $\dfrac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \underset{n\to\infty}{\longrightarrow} \dfrac{1}{\sqrt{2\pi}}$, why?].

Due to the symmetry of the distribution $[f(y) = f(-y)]$ its **mean** is zero (when is exists, i.e. when $n \geq 2$).

To compute its **variance**: $Var(Y) = \mathbb{E}(Y^2) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \int\limits_{-\infty}^{\infty} \frac{(y^2+n-n)\,dy}{\left(1+\frac{y^2}{n}\right)^{\frac{n+1}{2}}} =$ $\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left[n \cdot \frac{\Gamma(\frac{n-2}{2})\sqrt{n\pi}}{\Gamma(\frac{n-1}{2})} - n \cdot \frac{\Gamma(\frac{n}{2})\sqrt{n\pi}}{\Gamma(\frac{n+1}{2})} \right] = n \cdot \frac{\frac{n-1}{2}}{\frac{n-2}{2}} - n =$

$$\frac{n}{n-2} \qquad \text{(variance)}$$

for $n \geq 3$ (for $n = 1$ and 2 the variance is infinite).

Note that when $n \geq 30$ the t-distribution can be closely approximated by $\mathcal{N}(0,1)$.

5. And finally, we introduce the **Fisher's F-distribution**

(notation: $\mathsf{F}_{n,m}$ where n and m are its two parameters, also referred to as 'DEGREES OF FREEDOM'), defined by $Y_1 = \frac{\frac{X_1}{n}}{\frac{X_2}{m}}$ where X_1 and X_2 are *independent*, both having the chi-square distribution, with degrees of freedom n and m, respectively.

First we solve for $x_2 = y_2$ and $x_1 = \frac{n}{m}\, y_1\, y_2 \Rightarrow$ Jacobian equals to $\frac{n}{m}\, y_2$. Then we substitute into $\frac{x_1^{\frac{n}{2}-1} e^{-\frac{x_1}{2}}}{\Gamma(\frac{n}{2})\, 2^{\frac{n}{2}}} \cdot \frac{x_2^{\frac{m}{2}-1} e^{-\frac{x_2}{2}}}{\Gamma(\frac{m}{2})\, 2^{\frac{m}{2}}}$ and multiply by this Jacobian to get $\frac{(\frac{n}{m})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})\, 2^{\frac{n+m}{2}}}\, y_1^{\frac{n}{2}-1} \cdot y_2^{\frac{n+m}{2}-1} e^{-\frac{y_2\,(1+\frac{n}{m}y_1)}{2}}$ with $y_1 > 0$ and $y_2 > 0$. Integrating over y_2 (from 0 to ∞) yields the following formula for the corresponding **pdf**

$$f(y_1) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})} \left(\frac{n}{m}\right)^{\frac{n}{2}} \cdot \frac{y_1^{\frac{n}{2}-1}}{(1+\frac{n}{m}y_1)^{\frac{n+m}{2}}}$$

for $y_1 > 0$.

We can also find $\mathbb{E}(Y) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})} (\frac{n}{m})^{\frac{n}{2}} \int\limits_0^{\infty} \frac{y^{\frac{n}{2}}\, dy}{(1+\frac{n}{m}y)^{\frac{n+m}{2}}} =$

$$\frac{m}{m-2} \qquad \text{(mean)}$$

for $m \geq 3$ (the mean is infinite for $m = 1$ and 2).

Similarly $\mathbb{E}(Y^2) = \frac{(n+2)\, m^2}{(m-2)\,(m-4)\, n} \Rightarrow Var(Y) = \frac{(n+2)\, m^2}{(m-2)\,(m-4)\, n} - \frac{m^2}{(m-2)^2} = \frac{m^2}{(m-2)^2} \cdot \left[\frac{(n+2)\,(m-2)}{(m-4)\, n} - 1\right] =$

$$\frac{2\, m^2\, (n+m-2)}{(m-2)^2\, (m-4)\, n} \qquad \text{(variance)}$$

for $m \geq 5$ [infinite for $m = 1, 2, 3$ and 4].

Note that the distribution of $\dfrac{1}{Y}$ is obviously $\mathsf{F}_{m,n}$ [degrees of freedom reversed], also that $\mathsf{F}_{1,m} \equiv \dfrac{\chi_1^2}{\frac{\chi_m^2}{m}} \equiv \dfrac{Z^2}{\frac{\chi_m^2}{m}} \equiv \mathsf{t}_m^2$, and finally when both n and m are large (say > 30) then Y is **approximately normal** $\mathcal{N}\left(1, \sqrt{\frac{2(n+m)}{n\cdot m}}\right)$.

The last assertion can be proven by introducing $U = \sqrt{m}\cdot(Y-1)$, getting its pdf: (i) $y = 1+\frac{u}{\sqrt{m}}$, (ii) substituting: $\dfrac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})}\,(\frac{n}{m})^{\frac{n}{2}}\cdot\dfrac{(1+\frac{u}{\sqrt{m}})^{\frac{n}{2}-1}}{(1+\frac{n}{m}+\frac{n}{m}\frac{u}{\sqrt{m}})^{\frac{n+m}{2}}}\cdot$ $\frac{1}{\sqrt{m}}$ [the Jacobian] $= \dfrac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})\sqrt{m}}\cdot\dfrac{(\frac{n}{m})^{\frac{n}{2}}}{(1+\frac{n}{m})^{\frac{n+m}{2}}}\cdot\dfrac{(1+\frac{u}{\sqrt{m}})^{\frac{n}{2}-1}}{(1+\frac{n}{n+m}\frac{u}{\sqrt{m}})^{\frac{n+m}{2}}}$ where $-\sqrt{m} < u < \infty$. Now, taking the limit of the last factor (since that is the only part containing u, the rest being only a normalizing constant) we get [this is actually easier with the corresponding logarithm, namely $(\frac{n}{2}-1)\ln(1+\frac{u}{\sqrt{m}}) - \frac{n+m}{2}\ln(1+\frac{n}{n+m}\frac{u}{\sqrt{m}}) = -\frac{u}{\sqrt{m}} - \left[(\frac{n}{2}-1) - \frac{n^2}{2(n+m)}\right]\cdot \frac{u^2}{2m} - \ldots. = -\frac{u}{\sqrt{m}} + \frac{u^2}{2m} - \frac{n}{n+m}\frac{u^2}{4} - \ldots. \underset{n,m\to\infty}{\longrightarrow} -\dfrac{1}{1+\frac{m}{n}}\dfrac{u^2}{4}$ [assuming that the $\frac{m}{n}$ ratio remains finite]. This implies that the limiting pdf is $C\cdot e^{-\frac{u^2 n}{4(n+m)}}$ where C is a normalizing constant (try to establish its value). The limiting distribution is thus, obviously, $\mathcal{N}\left(0, \sqrt{\frac{2(n+m)}{n}}\right)$. Since this is the (approximate) distribution of U, $Y = \frac{U}{\sqrt{m}} + 1$ must be also (approximately) normal with the mean of 1 and the standard deviation of $\sqrt{\frac{2(n+m)}{n\cdot m}}$. □

We will see more examples of the F, t and χ^2 distributions in the next chapter, which discusses the importance of these distributions to Statistics, and the context in which they usually arise.

Chapter 3 RANDOM SAMPLING

A RANDOM INDEPENDENT SAMPLE (RIS) of SIZE n from a (specific) distribution is a collection of n *independent* RVs X_1, X_2, ..., X_n, each of them having the same (aforementioned) distribution. At this point, it is important to visualize these as true random variables (i.e. before the actual sample is taken, with all their would-be values), and not just as a collection of numbers (which they become eventually).

The information of a RIS is usually summarized by a handful of statistics (one is called a statistic), each of them being an expression (a transformation) involving the individual X_i's. The most important of these is the

Sample mean

defined as the usual (arithmetic) average of the X_i's:

$$\overline{X} \equiv \frac{\sum_{i=1}^{n} X_i}{n}$$

One has to realize that the sample mean, unlike the distribution's mean, is a *random variable*, with its own expected value, variance, and distribution. The obvious question is: How do these relate to the distribution from which we are sampling?

For the expected value and variance the answer is quite simple

$$\mathbb{E}\left(\overline{X}\right) = \frac{\sum_{i=1}^{n} \mathbb{E}\left(X_i\right)}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \frac{n\mu}{n} = \mu$$

and

$$\mathrm{Var}\left(\overline{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Note that this implies

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

(one of the most important formulas of Statistics).

Central Limit Theorem

The distribution of $\overline{X}$ is a lot trickier. When $n = 1$, it is clearly the same as the distribution form which we are sampling. But as soon as we take $n = 2$, we have to work out (which is a rather elaborate process) a CONVOLUTION of two such distributions (taking care of the $\frac{1}{2}$ factor is quite simple), and end up with a distribution which usually looks fairly *different* from the original. This procedure can then be repeated to get the $n = 3$, 4, etc. results. By the time we reach $n = 10$ (even though most books say 30), we notice something almost mysterious: The resulting distribution (of $\overline{X}$) will very quickly assume a *shape* which not only has nothing to do with the shape of the original distribution, it is the *same* for all (large) values of n, and (even more importantly) *for* practically *all distributions* (discrete or continuous) from which we may sample. This of course is the well known (bell-like) shape of the Normal distribution (mind you, there are other bell-look-alike distributions).

The proof of this utilizes a few things we have learned about the moment generating function:

Proof. We already know the mean and standard deviation of the distribution of $\overline{X}$ are μ and $\frac{\sigma}{\sqrt{n}}$ respectively, now we want to establish its ASYMPTOTIC (i.e. large-n) shape. This is, in a sense, trivial: since $\frac{\sigma}{\sqrt{n}} \underset{n\to\infty}{\longrightarrow} 0$, we get in the $n \to \infty$ limit a DEGENERATE (single-valued, with zero variance) distribution, with all probability concentrated at μ.

We can prevent this distribution from shrinking to a zero width by STANDARDIZING $\bar{X}$ first, i.e. defining a new RV

$$Z \equiv \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

and investigating its asymptotic distribution instead (the new random variable has the mean of 0 and the standard deviation of 1, thus its shape cannot 'disappear' on us).

We do this by constructing the MGF of Z and finding its $n \to \infty$ limit. Since $Z = \frac{\frac{\sum_{i=1}^{n}(X_i-\mu)}{n}}{\frac{\sigma}{\sqrt{n}}} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma\sqrt{n}}\right)$ (still a sum of independent, identically distributed RVs) its MGF is the MGF of $\frac{X_i-\mu}{\sigma\sqrt{n}} \equiv Y$, raised to the power of n.

We know that $M_Y(t) = 1 + \mathbb{E}(Y)\cdot t + \mathbb{E}(Y^2)\cdot\frac{t^2}{2} + \mathbb{E}(Y^3)\cdot\frac{t^3}{3!} + ... = 1 + \frac{t^2}{2n} + \frac{\alpha_3 t^3}{6n^{3/2}} + \frac{\alpha_4 t^4}{24n^2} + ...$ where α_3, α_4,... is the skewness, kurtosis, ... of the original distribution. Raising $M_Y(t)$ to the power of n and taking the $n \to \infty$ limit results in $e^{\frac{t^2}{2}}$ *regardless* of the values of α_3 and $\alpha_4, ...$ (since each is divided by higher-than-one power of n). This is easily recognized to be the MGF of the standardized (zero mean, unit variance) Normal distribution. ■

Note that, to be able to do all this, we had to assume that μ and σ are *finite*. There are (unusual) cases of distributions with an infinite variance (and sometimes also indefinite or infinite mean) for which the central limit theorem breaks down. A prime example is sampling from the Cauchy distribution, $\overline{X}$ (for any n) has the same Cauchy distribution as the individual X_i's - it does not get any narrower!

Sample variance

This is yet another expression involving the X_i's, intended as (what will later be called) an ESTIMATOR of σ^2. Its definition is

$$s^2 \equiv \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

where s, the corresponding square root, is the SAMPLE STANDARD DEVIATION (the sample variance does not have its own symbol).

To find its expected value, we first simplify its numerator:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}[(X_i-\mu)-(\bar{X}-\mu)]^2 = \sum_{i=1}^{n}(X_i-\mu)^2 - 2\sum_{i=1}^{n}(\bar{X}-\mu)(X_i-\mu) + n\cdot(\bar{X}-\mu)^2$$

This implies that

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \sum_{i=1}^{n}\mathrm{Var}(X_i)-2\sum_{i=1}^{n}\mathrm{Cov}(\bar{X},X_i)+n\cdot\mathrm{Var}(\bar{X}) = n\sigma^2+n\cdot\frac{\sigma^2}{n}-2n\cdot\frac{\sigma^2}{n} = \sigma^2(n-1)$$

since

$$\mathrm{Cov}(\bar{X},X_1) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}(X_i,X_1) = \frac{1}{n}\mathrm{Cov}(X_1,X_1)+\ 0 = \frac{1}{n}\mathrm{Var}(X_1) = \frac{\sigma^2}{n}$$

and $\mathrm{Cov}(\bar{X},X_2)$, $\mathrm{Cov}(\bar{X},X_3)$, ... must all have the same value.

Finally,

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

Thus, s^2 is a so called UNBIASED ESTIMATOR of the distribution's variance σ^2 (meaning it has the correct expected value).

Does this imply that $s \equiv \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}$ has the expected value of σ? The answer is 'no', s is (slightly) biased.

Sampling from $\mathcal{N}(\mu,\sigma)$

To be able to say anything more about s^2, we need to know the distribution form which we are sampling. We will thus assume that the distribution is Normal, with mean μ and variance σ^2. This immediately simplifies the distribution of $\overline{X}$, which must also be Normal (with mean σ and standard deviation of $\frac{\sigma}{\sqrt{n}}$, as we already know) for any sample size n (not just 'large').

Regarding s^2, one can show that it is independent of $\overline{X}$, and that the distribution of $\frac{(n-1)s^2}{\sigma^2}$ is χ^2_{n-1}. The proof of this is fairly complex.

Proof. We introduce a new set of n RVs $Y_1=\bar{X}$, $Y_2=X_2$, $Y_3=X_3$, ..., $Y_n=X_n$ and find their joint pdf by

1. solving for $\begin{cases} x_1 = ny_1 - x_2 - x_3 - \ldots - x_n \\ x_2 = y_2 \\ x_3 = y_3 \\ \ldots \\ x_n = y_n \end{cases}$

2. substituting into $\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\cdot e^{-\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$ (the pdf of the X_i's)

3. and multiplying by the Jacobian, which in this case equals to n.

Furthermore, since $\sum\limits_{i=1}^{n}(x_i-\mu)^2 = \sum\limits_{i=1}^{n}(x_i-\bar{X}+\bar{X}-\mu)^2 = \sum\limits_{i=1}^{n}(x_i-\bar{X})^2 - 2(\bar{X}-\mu)\sum\limits_{i=1}^{n}(x_i-\bar{X}) + n(\bar{X}-\mu)^2 = (n-1)s^2 + n(\bar{X}-\mu)^2$, the resulting pdf can be expressed as follows:

$$\frac{n}{(2\pi)^{\frac{n}{2}}\sigma^n}\cdot e^{-\dfrac{(n-1)s^2+n(y_1-\mu)^2}{2\sigma^2}}\qquad(dy_1dy_2....dy_n)$$

where s^2 is now to be seen as a function of the y_i's.

The *conditional pdf* of $y_2, y_3, ..., y_n|\mathbf{y}_1$ thus equals - all we have to do is divide the previous result by the marginal pdf of y_1, i.e. $\dfrac{\sqrt{n}}{(2\pi)^{\frac{1}{2}}\sigma}\cdot e^{-\dfrac{n(y_1-\mu)^2}{2\sigma^2}}$:

$$\frac{\sqrt{n}}{(2\pi)^{\frac{n-1}{2}}\sigma^{n-1}}\cdot e^{-\dfrac{(n-1)s^2}{2\sigma^2}}(dy_2....dy_n)$$

This implies that

$$\iiint\limits_{-\infty}^{\infty} e^{-\dfrac{(n-1)s^2}{2\Omega^2}}dy_2....dy_n = \frac{(2\pi)^{\frac{n-1}{2}}\Omega^{n-1}}{\sqrt{n}}$$

for *any* $\Omega > 0$ (just changing the name of σ). The last formula enables us to compute the corresponding *conditional MGF* of $\dfrac{(n-1)s^2}{\sigma^2}$ (given $\mathbf{y}_1$) by:

$$\begin{aligned}
&\frac{\sqrt{n}}{(2\pi)^{\frac{n-1}{2}}\sigma^{n-1}}\iiint\limits_{-\infty}^{\infty} e^{\dfrac{t(n-1)s^2}{\sigma^2}}\cdot e^{-\dfrac{(n-1)s^2}{2\sigma^2}}dy_2....dy_n\\
=\;&\frac{\sqrt{n}}{(2\pi)^{\frac{n-1}{2}}\sigma^{n-1}}\iiint\limits_{-\infty}^{\infty} e^{-\dfrac{(1-2t)(n-1)s^2}{2\sigma^2}}dy_2....dy_n\\
=\;&\frac{\sqrt{n}}{(2\pi)^{\frac{n-1}{2}}\sigma^{n-1}}\cdot\frac{(2\pi)^{\frac{n-1}{2}}\left(\frac{\sigma}{\sqrt{1-2t}}\right)^{n-1}}{\sqrt{n}}\\
=\;&\frac{1}{(1-2t)^{\frac{n-1}{2}}}
\end{aligned}$$

(substituting $\Omega = \frac{\sigma}{\sqrt{1-2t}}$). This is the MGF of the χ^2_{n-1} distribution, *regardless* of the value of $\mathbf{y}_1$ ($\equiv \bar{X}$). This clearly makes $\dfrac{(n-1)s^2}{\sigma^2}$ independent of $\bar{X}$. ■

The important implication of this is that $\dfrac{(\bar{X}-\mu)}{\dfrac{s}{\sqrt{n}}}$ has the $\mathtt{t}_{n-1}$ distribution.

$$\frac{(\bar{X}-\mu)}{\dfrac{s}{\sqrt{n}}}\equiv\frac{\dfrac{(\bar{X}-\mu)}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\dfrac{\frac{s^2(n-1)}{\sigma^2}}{n-1}}}\equiv\frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

Sampling without replacement

First, we have to understand the concept of a POPULATION. This is a special case of a distribution with N *equally likely* values, say x_1, x_2, ..., x_N, where N is often fairly large (millions). The x_i's don't have to be integers, they may not be all distinct (allowing only two possible values results in the hypergeometric distribution), and they may be 'dense' in one region of the real numbers and 'sparse' in another. They may thus 'mimic' just about any distribution, including Normal. That's why sometimes we use the words 'distribution' and 'population' interchangeably.

The mean and variance of this special distribution are simply

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

and

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

To generate a RIS form this distribution, we clearly have to do the so called SAMPLING WITH REPLACEMENT (meaning that each selected x_i value must be 'returned' to the population before the next draw, and potentially selected again - only this can guarantee independence). In this case, all our previous formulas concerning $\overline{X}$ and s^2 remain valid.

Sometimes though (and more efficiently), the sampling is done WITHOUT REPLACEMENT. This means that X_1, X_2, ..., X_n are no longer independent (they are still identically distributed). How does this effect the properties of $\overline{X}$ and s^2? Let's see.

The expected value of $\overline{X}$ remains equal to μ, by essentially the same argument as before (note that the proof does not require independence). Its variance is now computed by

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

since all the covariance (when $i \neq j$) have the same value, equal to

$$\begin{aligned} \text{Cov}(X_1, X_2) &= \frac{\sum_{k\neq \ell}(x_k - \mu)(x_\ell - \mu)}{N(N-1)} \\ &= \frac{\sum_{k=1}^{N}\sum_{\ell=1}^{N}(x_k - \mu)(x_\ell - \mu) - \sum_{k=1}^{N}(x_k - \mu)^2}{N(N-1)} \\ &= -\frac{\sigma^2}{N-1} \end{aligned}$$

Note that this variance is smaller (which is good) than what it was in the 'independent' case.

We don't need to pursue this topic any further.

Bivariate samples

A random independent sample of *size* n from a bivarite distribution consists of n pairs of RVs (X_1, Y_1), (X_2, Y_2), ... (X_n, Y_n), which are independent between (but not within) - each pair having *the same* (aforementioned) distribution.

We already know what are the individual properties of $\overline{X}$, $\overline{Y}$ (and of the two sample variances). Jointly, $\overline{X}$ and $\overline{Y}$ have a (complicated) bivariate distribution which, for $n \to \infty$, tends to be bivariate Normal. Accepting this statement (its proof would be similar to the univariate case), we need to know the five parameters which describe this distribution. Four of them are the marginal means and variances (already known), the last one is the correlation coefficient between $\overline{X}$ and $\overline{Y}$. One can prove that this equals to the correlation coefficient of the original distribution (from which we are sampling).

Proof. First we have

$$\mathrm{Cov}(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i) = \mathrm{Cov}(X_1, Y_1) + \mathrm{Cov}(X_2, Y_2) + \ldots.. + \mathrm{Cov}(X_n, Y_n) = n\,\mathrm{Cov}(X, Y)$$

since $\mathrm{Cov}(X_i, Y_j) = 0$ when $i \neq j$. This implies that the covariance between $\overline{X}$ and $\overline{X}$ equals $\frac{\mathrm{Cov}(X,Y)}{n}$. Finally, the corresponding correlation coefficient is: $\rho_{\overline{XY}} = \frac{\frac{\mathrm{Cov}(X,Y)}{n}}{\sqrt{\frac{\sigma_x^2}{n} \cdot \frac{\sigma_y^2}{n}}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y} = \rho_{xy}$, same as that of a single (X_i, Y_i) pair. ■

Chapter 4 ORDER STATISTICS

In this section we consider a RIS of size n from *any* distribution [not just $\mathcal{N}(\mu,\sigma)$], calling the individual observations X_1, X_2, ..., X_n (as we usually do). Based on these we define a new set of RVs $X_{(1)}$, $X_{(2)}, \ldots .X_{(n)}$ [your textbook calls them Y_1, $Y_2, \ldots Y_n$] to be the *smallest* sample value, the *second smallest* value, ..., the *largest* value, respectively. Even though the original X_i's were independent, $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$ are *strongly correlated.* They are called the first, the second, ..., and the last **order statistic**, respectively. Note that when n is odd, $X_{(\frac{n+1}{2})}$ is the sample *median* $\tilde{X}$.

Univariate pdf

To find the (marginal) **pdf** of a single order statistic $X_{(i)}$, we proceed as follows:

$$f_{(i)}(x) \equiv \lim_{\triangle\to 0} \frac{\Pr(x \le X_{(i)} < x+\triangle)}{\triangle} = \lim_{\triangle\to 0} \tbinom{n}{i-1,1,n-i} F(x)^{i-1} \tfrac{F(x+\triangle)-F(x)}{\triangle} \left[1-F(x+\triangle)\right]^{n-i}$$

[$i-1$ of the original observations must be smaller than x, one must be between x and $x+\triangle$, the rest must be bigger than $x+\triangle$] =

$$\frac{n!}{(i-1)!(n-i)!} F(x)^{i-1} \left[1-F(x)\right]^{n-i} f(x) \qquad (f)$$

It has the *same range* as the original distribution.

Using this formula, we can compute the **mean** and **variance** of any such order statistic; to answer a related **probability** question, instead of integrating $f_{(i)}(x)$ [which would be legitimate but tedious] we use a different, simplified approach.

EXAMPLES:

1. Consider a RIS of size 7 from $\mathcal{E}(\beta = 23\,\text{min}.)$ [seven fishermen independently catching one fish each].

   (a) Find $\Pr(X_{(3)} < 15$ min.) [the third catch of the group will not take longer than 15 min.].

   Solution: First find the probability that *any* one of the original 7 *independent* observations is <15 min. [using $F(x)$ of the corresponding exponential distribution]: $\Pr(X_i < 15 \text{ min.}) = 1 - e^{-\frac{15}{23}} = 0.479088 \equiv p$. Now interpret the same sampling as a *binomial* experiment, where a value smaller than 15 min. defines a *success,* and a value bigger than 15 min. represents a 'failure'. The question is: what is the probability of getting *at least* 3 successes (right)? Using binomial probabilities (and the complement shortcut) we get $1 - \left[q^7 + 7pq^6 + \binom{7}{2}p^2q^5\right] = 73.77\%$.

   (b) Now, find the mean and standard deviation of $X_{(3)}$.

   Solution: First we have to construct the corresponding pdf. By the above formula, this equals: $\frac{7!}{2!4!}(1-e^{-\frac{x}{\beta}})^{3-1}(e^{-\frac{x}{\beta}})^{7-3} \cdot \frac{1}{\beta}e^{-\frac{x}{\beta}} = \frac{105}{\beta}(1-e^{-\frac{x}{\beta}})^2 e^{-\frac{5x}{\beta}}$

$[x > 0]$ where $\beta = 23$ min. This yields the following mean: $105 \int_0^\infty x \cdot (1 - e^{-\frac{x}{\beta}})^2 e^{-\frac{5x}{\beta}} \frac{dx}{\beta} = 105\beta \int_0^\infty u \cdot (1 - e^{-u})^2 e^{-5u} du = 105\beta \int_0^\infty u \cdot (e^{-5u} - 2e^{-6u} + e^{-7u}) du = 105\beta \times [\frac{1}{5^2} - 2\frac{1}{6^2} + \frac{1}{7^2}] = 11.72$ min. [recall the $\int_0^\infty u^k e^{-\frac{u}{a}} du = k!\, a^{k+1}$ formula]. The second sample moment $\mathbb{E}(X_{(3)}^2)$ is similarly $105\beta^2 \int_0^\infty u^2 \cdot (e^{-5u} - 2e^{-6u} + e^{-7u}) du = 105\beta^2 \times 2\,[\frac{1}{5^3} - 2\frac{1}{6^3} + \frac{1}{7^3}] = 184.0 \Rightarrow \sigma_{X_{(3)}} = \sqrt{184 - 11.72^2} = 6.830$ min.

Note that if each of the fisherman continued fishing (when getting his first, second, ... catch), the distribution of the time of the third catch would be gamma$(3, \frac{23}{7})$, with the mean of 9.86 min. and $\sigma = \sqrt{3} \times \frac{23}{7} = 5.69$ min. [similar, but shorter than the original answer].

(c) Repeat both (a) and (b) with $X_{(7)}$.

Solution: The probability question is trivial: $\Pr(X_{(7)} < 15 \text{ min.}) = p^7 = 0.579\%$. The new pdf is: $7(1 - e^{-\frac{x}{\beta}})^6 \cdot \frac{1}{\beta} e^{-\frac{x}{\beta}}$ $[x > 0]$. $\mathbb{E}(X_{(7)}) = 7\beta \int_0^\infty u \cdot (1 - e^{-u})^6 e^{-u} du = 7\beta \times [1 - 6\frac{1}{2^2} + 15\frac{1}{3^2} - 20\frac{1}{4^2} + 15\frac{1}{5^2} - 6\frac{1}{6^2} + \frac{1}{7^2}] = 59.64$ min. and $\mathbb{E}(X_{(7)}^2) = 7\beta^2 \times 2\,[1 - 6\frac{1}{2^3} + 15\frac{1}{3^3} - 20\frac{1}{4^3} + 15\frac{1}{5^3} - 6\frac{1}{6^3} + \frac{1}{7^3}] = 4356.159 \Rightarrow \sigma = \sqrt{4356.2 - 59.64^2} = 28.28$ min.

Note: By a different approach, one can derive the following general formulas (applicable only for sampling from an *exponential* distribution):

$$\begin{aligned} \mathbb{E}(X_{(i)}) &= \beta \sum_{j=0}^{i-1} \frac{1}{n-j} \\ Var(X_{(i)}) &= \beta^2 \sum_{j=0}^{i-1} \frac{1}{(n-j)^2} \end{aligned}$$

Verify that they give the same answers as our lengthy integration above.

2. Consider a RIS of size 5 form $\mathcal{U}(0,1)$. Find the mean and standard deviation of $X_{(2)}$.

   Solution: The corresponding pdf is equal to $\frac{5!}{1!3!} x(1-x)^3$ $[0 < x < 1]$ which can be readily identified as beta$(2,4)$ [for this *uniform* sampling, $X_{(i)} \in$ beta$(i, n+1-i)$ in general]. By our former formulas $\mathbb{E}(X_{(2)}) = \frac{2}{2+4} = \frac{1}{3}$ and $Var(X_{(2)}) = \frac{2\times4}{(2+4)^2(2+4+1)} = \frac{2}{63} = 0.031746 \Rightarrow \sigma_{X_{(2)}} = 0.1782$ (no integration necessary).

   Note: These results can be easily extended to sampling from any uniform distribution $\mathcal{U}(a,b)$, by utilizing the $Y \equiv (b-a)X + a$ transformation.

Sample median

is obviously the most important sample statistic; let us have a closer look at it.

For **small samples**, we treat the sample median as one of the order statistics. This enables us to get its mean and standard deviation, and to answer a related probability question (see the previous set of examples).

When n is **large** (to simplify the issue, we assume that n is odd, i.e. $n \equiv 2k+1$) we can show that the sample median is *approximately Normal*, with the mean of $\tilde{\mu}$ (the *distribution*'s median) and the standard deviation of

$$\frac{1}{2f(\tilde{\mu})\sqrt{n}}$$

This is true even for distributions whose *mean does not exist* (e.g. Cauchy).

Proof: The sample median $\tilde{X} \equiv X_{(k+1)}$ has the following pdf: $\frac{(2k+1)!}{k!\cdot k!}F(x)^k[1-F(x)]^k f(x)$. To explore what happens when $k \to \infty$ (and to avoid getting a degenerate distribution) we introduce a new RV $Y \equiv (\tilde{X}-\tilde{\mu})\sqrt{n}$ [we assume that the standard deviation of $\tilde{X}$ decreases, like that of $\bar{X}$, with $\frac{1}{\sqrt{n}}$; this guess will prove correct!]. We build the pdf of Y in the usual three steps:

1. $x = \frac{y}{\sqrt{n}} + \tilde{\mu}$
2. $\frac{(2k+1)!}{k!\cdot k!}\, F(\frac{y}{\sqrt{2k+1}}+\tilde{\mu})^k\, [1-F(\frac{y}{\sqrt{2k+1}}+\tilde{\mu})]^k\, f(\frac{y}{\sqrt{2k+1}}+\tilde{\mu})$
3. multiply the last line by $\frac{1}{\sqrt{2k+1}}$.

To take the limit of the resulting pdf we first expand $F(\frac{y}{\sqrt{2k+1}}+\tilde{\mu})$ as $F(\tilde{\mu}) + F'(\tilde{\mu})\frac{y}{\sqrt{2k+1}} + \frac{F''(\tilde{\mu})}{2}\frac{y^2}{2k+1} + ... =$

$$\frac{1}{2} + f(\tilde{\mu})\frac{y}{\sqrt{2k+1}} + \frac{f'(\tilde{\mu})}{2}\frac{y^2}{2k+1} + ... \qquad (F)$$

$\Rightarrow 1 - F(\frac{y}{\sqrt{2k+1}}+\tilde{\mu}) \approx \frac{1}{2} - f(\tilde{\mu})\frac{y}{\sqrt{2k+1}} - \frac{f'(\tilde{\mu})}{2}\frac{y^2}{2k+1} + ...$. Multiplying the two results in $F(\frac{y}{\sqrt{2k+1}}+\tilde{\mu})\,[1-F(\frac{y}{\sqrt{2k+1}}+\tilde{\mu})] \approx \frac{1}{4} - f(\tilde{\mu})^2\frac{y^2}{2k+1} + ...$ [the dots imply terms proportional to $\frac{1}{(2k+1)^{3/2}}$, $\frac{1}{(2k+1)^2}$, ...; these cannot effect the subsequent limit].

Substituting into the above pdf yields:

$$\frac{(2k+1)!}{2^{2k}\cdot k!\cdot k!\cdot\sqrt{2k+1}} \times [1-4f(\tilde{\mu})^2\frac{y^2}{2k+1}+....]^k\, f(\frac{y}{\sqrt{2k+1}}+\tilde{\mu})$$

[we extracted $\frac{1}{4}$ from inside the brackets]. Taking the $k \to \infty$ limit of the expression to the right of × [which carries the y-dependence] is trivial: $e^{-2f(\tilde{\mu})^2y^2}f(\tilde{\mu})$. This is [up to the normalizing constant] the pdf of $\mathcal{N}(0, \frac{1}{2f(\tilde{\mu})})$ [as a by-product, we derived the so called Wallis formula: $\frac{(2k+1)!}{2^{2k}\cdot k!\cdot k!\cdot\sqrt{2k+1}} \underset{k\to\infty}{\longrightarrow} \sqrt{\frac{2}{\pi}}$, to maintain proper normalization]. And, since $\tilde{X} = \tilde{\mu} + \frac{Y}{\sqrt{n}}$, the distribution of the sample median must be, approximately, $\mathcal{N}(\tilde{\mu}, \frac{1}{2f(\tilde{\mu})\sqrt{n}})$. □

EXAMPLES:

1. Consider a RIS of size 1001 from the Cauchy distribution with $f(x) = \frac{1}{\pi}\cdot\frac{1}{1+x^2}$. Find $\Pr(-0.1 < \tilde{X} < 0.1)$.

   Solution: We know that $\tilde{X} \approx \mathcal{N}(0, \frac{1}{2\cdot\frac{1}{\pi}\cdot\sqrt{1001}} = 0.049648)$. Thus $\Pr(\frac{-0.1}{0.049648} < \frac{\tilde{X}}{0.049648} < \frac{0.1}{0.049648}) = \Pr(-2.0142 < Z < 2.0142) = 95.60\%$.

   Note that $\Pr(-0.1 < \bar{X} < 0.1) = \frac{1}{\pi}\arctan(x)\Big|_{x=-0.1}^{0.1} = 6.35\%$ only (and it does not improve with n). So, in this case, the sample median enables us to estimate the center of the Cauchy distribution much more accurately then the sample mean would (but don't generalize this to other distributions).

2. Sampling from $\mathcal{N}(\mu, \sigma)$, is it better to estimate μ by the sample mean or by the sample median (trying to find the best ESTIMATOR of a parameter will be the issue of the subsequent chapter)?

   Solution: Since $\bar{X} \in \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ and $\tilde{X} \approx \mathcal{N}(\mu, \frac{1}{2\cdot\frac{1}{\sqrt{2\pi}\sigma}\cdot\sqrt{n}} = \sqrt{\frac{\pi}{2}}\cdot\frac{\sigma}{\sqrt{n}})$, it is obvious that $\tilde{X}$'s standard error is $\sqrt{\frac{\pi}{2}} = 1.253$ times *bigger* than that of $\bar{X}$. Thus, this time, we are better off using $\bar{X}$. [To estimate μ to the same accuracy as $\overline{X}$ does, $\tilde{X}$ would have to use $\frac{\pi}{2} = 1.57$ times bigger sample; the sample mean is, in this case, 57% more EFFICIENT than the sample median].

3. Consider a RIS of size 349 from a distribution with $f(x) = 2x$ $(0 < x < 1)$. Find $\Pr(\tilde{X} < 0.75)$.

   Solution: From $F(x) = x^2$ we first establish the distribution's median as the solution to $x^2 = \frac{1}{2} \Rightarrow \tilde{\mu} = \frac{1}{\sqrt{2}}$ [the corresponding $f(\tilde{\mu})$ is equal to $\sqrt{2}$]. Our probability thus equals $\Pr(\frac{\tilde{X}-\frac{1}{\sqrt{2}}}{\frac{1}{2\cdot\sqrt{2}\cdot\sqrt{349}}} < \frac{0.75-\frac{1}{\sqrt{2}}}{\frac{1}{2\cdot\sqrt{2}\cdot\sqrt{349}}}) \approx \Pr(Z < 2.26645) = 98.83\%$ [The exact probability, which can be evaluated by computer, is 99.05%].

   Subsidiary: Find $\Pr(\bar{X} < 0.75)$.

   Solution: First we need $\mathbb{E}(X) = \int\limits_0^1 2x\cdot x\,dx = \frac{2}{3}$ and $Var(X) = \int\limits_0^1 2x\cdot x^2dx - (\frac{2}{3})^2 = \frac{1}{18}$. We know that $\bar{X} \approx \mathcal{N}(\frac{2}{3}, \frac{1}{\sqrt{18}\cdot\sqrt{349}} = 0.0126168) \Rightarrow \Pr(\frac{\bar{X}-\frac{2}{3}}{0.0126168} < \frac{0.75-\frac{2}{3}}{0.0126168}) = \Pr(Z < 6.6049) = 100\%$. ■

Bivariate pdf

We now construct the **joint distribution** of *two* order statistics $X_{(i)}$ and $X_{(j)}$ $[i < j]$. By our former definition, $f(x,y) = \lim\limits_{\substack{\Delta\to 0\\ \varepsilon\to 0}} \frac{\Pr(x\le X_{(i)}<x+\Delta\,\cap\, y\le X_{(j)}<y+\varepsilon)}{\Delta\cdot\varepsilon}$. To make the event in parentheses happen, exactly $i-1$ observations must have a value less than x, 1 observation must fall in the $[x, x+\Delta)$ interval, $j-i-1$ observations must be between $x+\Delta$ and y, 1 observation must fall in $[y, y+\varepsilon)$ and $n-j$ observations must be bigger than $y+\varepsilon$. By our multinomial formula, this equals $\binom{n}{i-1,1,j-i-1,1,n-j} F(x)^{i-1}\,[F(x+\Delta)-F(x)]\,[F(y)-F(x+\Delta)]^{j-i-1}\,[F(y+\varepsilon)-F(y)]\,[1-F(y+\varepsilon)]^{n-j}$. Dividing by $\Delta\cdot\varepsilon$ and taking the two limits yields

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!}F(x)^{i-1}\,f(x)\,[F(y)-F(x)]^{j-i-1}\,f(y)\,[1-F(y)]^{n-j}$$

with $L < x < y < H$, where L and H is the lower and upper limit (respectively) of the original distribution.

Let us discuss two *important*

Special Cases
of this formula:

1. **Consecutive** order statistics, i and $i+1$:

$$f(x,y) = \frac{n!}{(i-1)!(n-i-1)!} F(x)^{i-1} [1-F(y)]^{n-i-1} f(x)\, f(y)$$

where $L < x < y < H$ [x corresponds to $X_{(i)}$, y to $X_{(i+1)}$].

- This reduces to $\frac{n!}{(i-1)!(n-i-1)!} x^{i-1}(1-y)^{n-i-1}$ with $0 < x < y < 1$ when the distribution is **uniform** $\mathcal{U}(0,1)$. Based on this, we can
- find the distribution of $U = X_{(i+1)} - X_{(i)}$:

Solution: We introduce $V \equiv X_{(i)}$. Then

(i) $y = u + v$ and $x = v$,

(ii) the joint pdf of u and v is $f(u,v) = \frac{n!}{(i-1)!(n-i-1)!} v^{i-1} (1-u-v)^{n-i-1} \cdot 1$ [Jacobian] where $0 < v < 1$ and $0 < u < 1-v \Leftrightarrow 0 < v < 1-u$ and $0 < u < 1$,

(iii) the marginal pdf of u is $\frac{n!}{(i-1)!(n-i-1)!} \int\limits_0^{1-u} v^{i-1} (1-u-v)^{n-i-1}\, dv = n(1-u)^{n-1}$ for $0 < u < n$ [with the help of $\int\limits_0^a v^{i-1}(a-v)^{j-1}\, dv = a^{i+j-1} \int\limits_0^a (\frac{v}{a})^{i-1}(1-\frac{v}{a})^{j-1} \frac{dv}{a} = a^{i+j-1} \int\limits_0^1 y^{i-1}(1-y)^{j-1}\, dy = a^{i+j-1} \frac{\Gamma(i)\Gamma(j)}{\Gamma(i+j)}$].

The corresponding distribution function is $F(u) = 1-(1-u)^n$ for $0 < u < 1$ (the same, regardless of the i value).

To see what happens to this distribution in the $n \to \infty$ limit, we must first introduce $W \equiv U \cdot n$ (why?). Then, clearly, $F_W(w) = \Pr(U < \frac{w}{n}) = 1-(1-\frac{w}{n})^n$ for $0 < w < n$. In the $n \to \infty$ limit, this $F_W(w)$ tends to $1-e^{-w}$ for $w > 0$ [the exponential distribution with $\beta = 1$]. This is what we have always used for the time interval between two consecutive arrivals (and now we understand why). We note in passing that a similar results holds even when the original distribution is not uniform (the inter-arrival times are still exponential, but the corresponding β values now depend on whether we are in the slack or busy period).

EXAMPLE:

100 students choose, *independently* and *uniformly*, to visit the library between 12 a.m. and 1 p.m. Find $\Pr(X_{(47)} - X_{(46)} > 3$ min.) [probability that the time interval between the 46th and 47th arrival is at least 3 minutes].

Solution: Based on the distribution function just derived, this equals $\Pr[X_{(47)} - X_{(46)} > \frac{3}{60}$ hr.$] = 1 - F(\frac{1}{20}) = (1-\frac{1}{20})^{100} = 0.592\%$. ■

2. **First and last** order statistics, $i = 1$ and $j = n$:

$$f(x, y) = n(n-1)\,[F(y) - F(x)]^{n-2}\,f(x)\,f(y)$$

where $L < x < y < H$.

- Based on this result, you will be asked (in the assignment) to investigate the distribution of the SAMPLE RANGE $X_{(n)} - X_{(1)}$.
- When the sampling distribution is $\mathcal{U}(0,1)$, the pdf simplifies to: $f(x, y) = n(n-1)\,[y - x]^{n-2}$, where $0 < x < y < 1$. For this special case we want to
- find the distribution of $U \equiv \frac{X_{(1)}+X_{(n)}}{2}$ [the MID-RANGE value]:

Solution: $V \equiv X_{(1)} \Rightarrow$

(i) $x = v$ and $y = 2u - v$,

(ii) $f(u, v) = 2n(n-1)\,(2u - 2v)^{n-2}$, where $0 < v < 1$ and $v < u < \frac{v+1}{2}$ [visualize the region!]

(iii) $f(u) = 2^{n-1}\,n(n-1) \int\limits_{\max(0,2u-1)}^{u} (u-v)^{n-2}\,dv = 2^{n-1}\,n \times \left\{ \begin{array}{cc} u^{n-1} & 0 < u < \frac{1}{2} \\ (1-u)^{n-1} & \frac{1}{2} < u < 1 \end{array} \right. \Rightarrow$

$$\left. \begin{array}{c} F(u) \\ 1 - F(u) \end{array} \right\} = 2^{n-1} \times \left\{ \begin{array}{cc} u^{n} & 0 < u < \frac{1}{2} \\ (1-u)^{n} & \frac{1}{2} < u < 1 \end{array} \right. .$$

Pursuing this further: $\mathbb{E}(U) = \frac{1}{2}$ [based on the $f(\frac{1}{2} + u) \equiv f(\frac{1}{2} - u)$ symmetry] and $Var(U) = \int\limits_0^1 (u - \frac{1}{2})^2\,f(u)\,du =$

$n \int\limits_0^{\frac{1}{2}} \left(u - \frac{1}{2}\right)^2 (2u)^{n-1}\,du + n \int\limits_{\frac{1}{2}}^{1} \left((1-u) - \frac{1}{2}\right)^2 (2(1-u))^{n-1}\,du = 2^n n \int\limits_0^{\frac{1}{2}} \left(\frac{1}{2} - u\right)^2 u^{n-1}\,du =$

$2^n n \frac{\Gamma(3)\Gamma(n)}{\Gamma(n+3)} \left(\frac{1}{2}\right)^{n+2} = \frac{1}{2(n+2)(n+1)} \Rightarrow \sigma_U = \frac{1}{\sqrt{2(n+2)(n+1)}}$.

These results can be now easily extended to cover the case of a general uniform distribution $\mathcal{U}(a, b)$ [note that all it takes is the $X_G \equiv (b-a)X + a$ transformation, applied to each of the $X_{(i)}$ variables, and consequently to U]. The results are now

$$\begin{aligned} \mathbb{E}(U_G) &= \frac{a+b}{2} \\ \sigma_{U_G} &= \frac{b-a}{\sqrt{2(n+2)(n+1)}} \end{aligned}$$

This means, as an estimator of $\frac{a+b}{2}$, the mid-range value is a lot better (judged by its standard error) than either $\bar{X} \approx \mathcal{N}(\frac{a+b}{2}, \frac{b-a}{\sqrt{12n}})$ or $\tilde{X} \approx \mathcal{N}(\frac{a+b}{2}, \frac{b-a}{2\sqrt{n}})$.

EXAMPLE:

Consider a RIS of size 1001 from $\mathcal{U}(0,1)$. Compare

- $\Pr(0.499 < \frac{X_{(1)}+X_{(1001\,)}}{2} < 0.501) = 1 - \frac{1}{2}(2 \times 0.499)^{1001} - \frac{1}{2}(2 \times 0.499)^{1001}$ [using $F(u)$ of the previous example] $= 86.52\%$

- $\Pr(0.499 < \bar{X} < 0.501) \simeq \Pr\left(\frac{0.499-0.5}{\frac{1}{\sqrt{12\times 1001}}} < \frac{\bar{X}-\frac{1}{2}}{\frac{1}{\sqrt{12\times 1001}}} < \frac{0.501-0.5}{\frac{1}{\sqrt{12\times 1001}}}\right) =$ $\Pr\left(-.1095993 < Z < .1095993\right) = 8.73\%$

- $\Pr(0.499 < \tilde{X} < 0.501) \simeq \Pr\left(\frac{0.499-0.5}{\frac{1}{2\sqrt{1001}}} < \frac{\tilde{X}-\frac{1}{2}}{\frac{1}{2\sqrt{1001}}} < \frac{0.501-0.5}{\frac{1}{2\sqrt{1001}}}\right) =$ $\Pr\left(-0.063277 < Z < 0.063277\right) = 5.05\%$.

This demonstrates that, for a *uniform* distribution, the mid-range value is a lot more likely to 'find' the true center than either the sample mean or the sample median. ■

Chapter 5 **ESTIMATING DISTRIBUTION PARAMETERS**

Until now we have studied PROBABILITY, proceeding as follows: we assumed *parameters* of all distributions to be *known* and, based on this, *computed probabilities* of various outcomes (in a random experiment). In this chapter we make the essential transition to STATISTICS, which is concerned with the exact opposite: the random experiment is performed (usually many times) and the individual outcomes recorded; based on these, we want to *estimate* values of the distribution *parameters* (one or more). Until the last two sections, we restrict our attention to the (easier and most common) case of estimating only ONE PARAMETER of a distribution.

EXAMPLE: How should we estimate the mean μ of a Normal distribution $\mathcal{N}(\mu, \sigma)$, based on a RIS of size n? We would probably take $\bar{X}$ (the sample mean) to be a 'reasonable' ESTIMATOR of μ [note that this name applies to the *random variable* $\bar{X}$, with all its *potential* (would-be) values; as soon as the experiment is completed and a particular value of $\bar{X}$ recorded, this *value* (i.e. a specific *number*) is called an ESTIMATE of μ]. ■

There is a few **related issues** we have to sort out:

- How do we know that $\bar{X}$ is a 'good' estimator of μ, i.e. is there some sensible set of *criteria* which would enable us to judge the *quality* of individual estimators?

- Using these criteria, can we then find the *best* estimator of a parameter, at least in some restricted sense?

- Would not it be better to use, instead of a single number [the so called POINT ESTIMATE, which can never precisely agree with the *exact* value of the unknown parameter, and is thus in this sense always wrong], an *interval* of values which may have a good chance of *containing* the correct answer?

The rest of this chapter tackles the first two issues. We start with

A few definitions

First we allow an **estimator** of a parameter θ to be *any* 'reasonable' combination (transformation) of X_1, X_2, ..., X_n [our RIS], say $\hat{\Theta}(X_1, X_2, ..., X_n)$ [the sample mean $\frac{X_1+X_2+...+X_n}{n}$ being a good example]. Note that n (being known) can be used in the expression for $\hat{\Theta}$; similarly, we can use values of other parameter if these are known [e.g.: in the case of hypergeometric distribution, N is usually known and only K needs to be estimated; in the case of negative binomial distribution k is given and p estimated, etc.]. Also note that some parameters may have only integer values, while others are real; typically, we concentrate on estimating parameters of the latter type.

To narrow down our choices (we are after 'sensible', 'good' estimators) we first insist that our estimators be **unbiased**

$$\mathbb{E}(\hat{\Theta}) = \theta$$

(having the *exact* long-run *average*), or at least ASYMPTOTICALLY UNBIASED , i.e.

$$\mathbb{E}(\hat{\Theta}) \underset{n\to\infty}{\longrightarrow} \theta$$

(being unbiased in the large-sample limit).

The $\mathbb{E}(\hat{\Theta}) - \theta$ difference is called a BIAS of an estimator [$\equiv 0$ for unbiased estimators, usually proportional to $\frac{1}{n}$ for asymptotically unbiased estimators], and can be normally removed with a bit of effort (i.e. constructing unbiased estimators is not a major challenge).

EXAMPLE:

Propose an estimator for the variance σ^2 ($\equiv$ our θ) of a Normal $\mathcal{N}(\mu, \sigma)$ distribution, assuming that the value of μ is also unknown.

We can start with $\hat{\Theta} = \dfrac{\sum\limits_{i=1}^{n}(X - \bar{X})^2}{n}$ and show [as we did in a previous chapter] that $\mathbb{E}(\hat{\Theta}) = \frac{n-1}{n}\sigma^2$. Our estimator is thus asymptotically unbiased only. This bias can be easily removed by defining a new estimator $s^2 \equiv \frac{n}{n-1}\hat{\Theta}$ [the sample variance] which is fully unbiased. Since $\frac{n-1}{\sigma^2}s^2 \in \chi^2_{n-1}$, we can establish not only that $\mathbb{E}(s^2) = \frac{\sigma^2}{n-1} \cdot n - 1 = \sigma^2$ (unbiased), but also that $Var(s^2) = (\frac{\sigma^2}{n-1})^2 \cdot 2(n-1) = \frac{2\sigma^4}{n-1}$, which we need later.

- Supplementary: Does this imply that s is an unbiased estimator of σ? The answer is 'No', as we can see from $\mathbb{E}\left(\sqrt{\chi^2_{n-1}}\right) = \dfrac{1}{\Gamma(\frac{n-1}{2})\, 2^{\frac{n-1}{2}}} \int\limits_0^\infty \sqrt{x}{\cdot}x^{\frac{n-3}{2}} e^{-\frac{x}{2}}\, dx = \dfrac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \Rightarrow \mathbb{E}(s) = \dfrac{\sigma}{\sqrt{n-1}} \cdot \dfrac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \approx \sigma(1 - \frac{1}{4n} - \frac{7}{32n^2} + ...)$. We know how to fix this: use $\hat{\Theta} \equiv \sqrt{\frac{n-1}{2}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} s$ instead, it is a fully unbiased estimator of σ. ■

Yet, making an estimator unbiased (or at least asymptotically so) is not enough to make it even acceptable (let alone 'good'). Consider estimating μ of a distribution by taking $\hat{\Theta} = X_1$ (the first observation only), throwing away X_2, $X_3, ...X_n$ [most of our sample!]. We get a fully unbiased estimator which is evidently unacceptable, since we are wasting nearly all the information contained in our sample. It is thus obvious that being unbiased is only *one* essential ingredient of a good estimator, the other one is its *variance* (a square of its standard error). A good estimator should not only be unbiased, but it should also have a variance which is as small as possible. This leads to two new definitions:

Consistent estimator is such that

1. $\mathbb{E}(\hat{\Theta}) \underset{n\to\infty}{\longrightarrow} \theta$ [asymptotically unbiased], and
2. $Var(\hat{\Theta}) \underset{n\to\infty}{\longrightarrow} 0$.

This implies that we can reach the exact value of θ by indefinitely increasing the sample size. That sounds fairly good, yet it represents what I would call '*minimal standards*' (or less), i.e. every 'decent' estimator is consistent; that by itself does not make it particularly good.

Example: $\hat{\Theta} = \dfrac{X_2 + X_4 + X_6 + ...X_n}{\frac{n}{2}}$ [n even] is a consistent estimator of μ, since its asymptotic (large n) distribution is $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{\frac{n}{2}}})$. Yet, we are wasting one half of our sample, which is unacceptable.

Minimum variance unbiased estimator (MVUE or 'best' estimator from now on) is an unbiased estimator whose *variance* is better or equal to the variance of any other *unbiased* estimator [uniformly, i.e. for *all* values of θ]. (The restriction to unbiased estimators is essential: an arbitrary *constant* may be totally nonsensical as an estimator (in all but 'lucky-guess' situations), yet no other estimator can compete with its variance which is identically equal to zero).

Having such an estimator would of course be ideal, but we run into two difficulties:

1. The variance of an estimator is, in general, a *function* of the unknown parameter [to see that, go back to the s^2 example], so we are comparing functions, not values. It may easily happen that two unbiased estimators have variances such that one estimator is better in some range of θ values and worse in another. Neither estimator is then (uniformly) better than its counterpart, and the 'best' estimator *may* therefore *not exit* at all.
2. Even when the 'best' estimator exists, how do we know that it does and, more importantly, how do we find it (out of the multitude of all unbiased estimators)?

To partially answer the last issues: luckily, there is a *theoretical* lower *bound* on the variance of all unbiased estimators; when an estimator achieves this bound, it is automatically MVUE. The relevant details are summarized in the following Theorem:

Cramér-Rao inequality

When estimating a parameter θ which does *not* appear in the *limits* of the distribution (the so called **regular case**), by an *unbiased* estimator $\hat{\Theta}$, then

$$Var(\hat{\Theta}) \geq \frac{1}{n\mathbb{E}\left[\left(\frac{\partial \ln f(x|\theta)}{\partial\theta}\right)^2\right]} \equiv \frac{1}{-n\mathbb{E}\left[\frac{\partial^2 \ln f(x|\theta)}{\partial\theta^2}\right]} \qquad \text{(C-R)}$$

where $f(x|\theta)$ stands for the old $f(x)$ – we are now emphasizing its *functional* dependence on the parameter θ. As θ is fixed (albeit unknown) and not 'random' in any sense, this is *not* to be confused with our conditional-pdf notation.

Proof: The joint pdf of X_1, X_2, ..., X_n is $\prod_{i=1}^{n} f(x_i|\theta)$ where $L < x_1, x_2, ...x_n < H$. Define a new RV $U \equiv \sum_{i=1}^{n} U_i \equiv \sum_{i=1}^{n} \frac{\partial \ln f(X_i|\theta)}{\partial\theta} = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \ln f(X_i|\theta) = \frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i|\theta) = \frac{\frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(X_i|\theta)}{\prod_{i=1}^{n} f(X_i|\theta)} \Rightarrow \mathbb{E}(U) = \sum_{i=1}^{n} \mathbb{E}(U_i) = n \int_L^H \frac{\partial \ln f(x|\theta)}{\partial\theta} \cdot f(x|\theta)\, dx = n \int_L^H \frac{\partial f(x|\theta)}{\partial\theta}\, dx = n\frac{\partial}{\partial\theta} \int_L^H f(x|\theta)\, dx = n\frac{\partial}{\partial\theta}(1) = 0$ and $Var(U) = \sum_{i=1}^{n} Var(U_i) = n\mathbb{E}\left[\left(\frac{\partial \ln f(X|\theta)}{\partial\theta}\right)^2\right]$. We also know that $\mathbb{E}(\hat{\Theta}) = \int_L^H ... \int_L^H \hat{\Theta} \cdot \prod_{i=1}^{n} f(x_i|\theta)\, dx_1 dx_2 ....dx_n = \theta$ [unbiased]. Differentiating this equation with respect to θ yields: $\mathbb{E}(\hat{\Theta}{\cdot}U) = 1 \Rightarrow Cov(\hat{\Theta}, U) = 1$. But we know that $Cov(\hat{\Theta}, U)^2 \le Var(\hat{\Theta}) \cdot Var(U) \Rightarrow Var(\hat{\Theta}) \ge \frac{1}{n\mathbb{E}\left[\left(\frac{\partial \ln f(X|\theta)}{\partial\theta}\right)^2\right]}$, which is the C-R bound. Differentiating $\int_L^H f(x|\theta)\, dx = 1$ yields $\int_L^H \frac{\partial f(x|\theta)}{\partial\theta}\, dx \equiv \int_L^H \frac{\partial}{\partial\theta} \ln f(x|\theta) \cdot f(x|\theta)\, dx = 0$, and once more:
$\int_L^H \left[\frac{\partial^2}{\partial\theta^2} \ln f(x|\theta) \cdot f(x|\theta) + \frac{\partial}{\partial\theta} \ln f(x|\theta) \cdot \frac{\partial}{\partial\theta} f(x|\theta)\right]\, dx = 0 \Rightarrow \mathbb{E}\left[\left(\frac{\partial \ln f(X|\theta)}{\partial\theta}\right)^2\right] \equiv -\mathbb{E}\left[\frac{\partial^2 \ln f(X|\theta)}{\partial\theta^2}\right]$. Equivalently, we can then say that $Var(\hat{\Theta}) \ge \frac{1}{-n\mathbb{E}\left[\frac{\partial^2 \ln f(X|\theta)}{\partial\theta^2}\right]}$ [we will use CRV as a shorthand for the last expression]. Note that this proof holds in the case of a discrete distribution as well (each integration needs to be replaced by the corresponding summation). □

Based on this C-R bound we define the so called **efficiency** of an *unbiased* estimator $\hat{\Theta}$ as the ratio of the theoretical variance bound CRV to the actual variance of $\hat{\Theta}$, thus:

$$\frac{CRV}{Var(\hat{\Theta})}$$

usually expressed in percent [we know that its value cannot be bigger that 1, i.e. 100%]. An estimator whose variance is as small as RCV is called EFFICIENT [note that, from what we know already, this makes it automatically the MVUE or 'best' estimator of θ]. An estimator which reaches 100% efficiency only in the $n \to \infty$ limit is called ASYMPTOTICALLY EFFICIENT.

One can also define RELATIVE EFFICIENCY of two estimators with respect to one another as $\frac{Var(\hat{\Theta}_2)}{Var(\hat{\Theta}_1)}$ [this is the relative efficiency of $\hat{\Theta}_1$ compared to $\hat{\Theta}_2$ – note that the variance ratio is reversed!].

EXAMPLES:

1. How good is $\bar{X}$ as an estimator of μ of the Normal distribution $\mathcal{N}(\mu,\sigma)$.

   Solution: We know that its variance is $\frac{\sigma^2}{n}$. To compute the C-R bound we do $\frac{\partial^2}{\partial\mu^2}\left[-\ln(\sqrt{2\pi}\sigma)-\frac{(x-\mu)^2}{2\sigma^2}\right] = -\frac{1}{\sigma^2}$. Thus CRV equals $\dfrac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n}$ implying that $\bar{X}$ is the best (unbiased) estimator of μ.

2. Consider a RIS of size 3 form $\mathcal{N}(\mu,\sigma)$. What is the relative efficiency of $\frac{X_1+2X_2+X_3}{4}$ [obviously unbiased] with respect to $\bar{X}$ (when estimating μ)?

   Solution: $Var(\frac{X_1+2X_2+X_3}{4}) = (\frac{\sigma^2}{16}+\frac{4\sigma^2}{16}+\frac{\sigma^2}{16}) = \frac{3}{8}\sigma^2$.

   Answer: $\dfrac{\frac{\sigma^2}{3}}{\frac{3}{8}\sigma^2} = \frac{8}{9} = 88.89\%$.

3. Suppose we want to estimate p of a Bernoulli distribution by the experimental proportion of successes, i.e. $\hat{\Theta} = \dfrac{\sum\limits_{i=1}^{n} X_i}{n}$. The mean of our estimator is $\frac{np}{n} = p$ [unbiased], its variance equals $\frac{npq}{n^2} = \frac{pq}{n}$ [since $\sum\limits_{i=1}^{n} X_i$ has the binomial distribution]. Is this the best we can do?

   Solution: Let us compute the corresponding CRV by starting from $f(x) = p^x(1-p)^{1-x}$ $[x = 0,1]$ and computing $\frac{\partial^2}{\partial p^2}[x\ln p + (1-x)\ln p] = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2} \Rightarrow \mathbb{E}\left[\frac{X}{p^2}+\frac{1-X}{(1-p)^2}\right] = \frac{1}{p}+\frac{1}{1-p} = \frac{1}{pq} \Rightarrow CRV = \frac{pq}{n}$. So again, our estimator is the best one can find.

4. Let us find the efficiency of $\bar{X}$ to estimate the mean β of the exponential distribution, with $f(x) = \frac{1}{\beta}e^{-\frac{x}{\beta}}$ $[x > 0]$.

   Solution: $\frac{\partial^2}{\partial\beta^2}\left[-\ln\beta - \frac{x}{\beta}\right] = \frac{1}{\beta^2} - \frac{2x}{\beta^3} \Rightarrow \mathbb{E}\left[\frac{2X}{\beta^3}-\frac{1}{\beta^2}\right] = \frac{1}{\beta^2} \Rightarrow CRV = \frac{\beta^2}{n}$. We know that $\mathbb{E}(\bar{X}) = \frac{n\beta}{n} = \beta$ and $Var(\bar{X}) = \frac{n\beta^2}{n^2} = \frac{\beta^2}{n}$.

   Conclusion: $\bar{X}$ is the best estimator of β.

5. Similarly, how good is $\bar{X}$ in estimating λ of the Poisson distribution?

   Solution: $\frac{\partial^2}{\partial\lambda^2}[x\ln\lambda - \ln(x!) - \lambda] = -\frac{x}{\lambda^2} \Rightarrow \mathbb{E}\left[\frac{X}{\lambda^2}\right] = \frac{1}{\lambda} \Rightarrow CRV = \frac{\lambda}{n}$. Since $\mathbb{E}(\bar{X}) = \frac{n\lambda}{n} = \lambda$ and $Var(\bar{X}) = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$, we again have the best estimator.

6. Let us try estimating θ of the uniform distribution $\mathcal{U}(0,\theta)$. This is *not* a *regular* case, so we don't have CRV and the concept of (absolute) efficiency. We propose, and compare the quality of, two estimators, namely $2\bar{X}$ and $X_{(n)}$ [the largest sample value].

   To investigate the former one we need $\mathbb{E}(X_i) = \frac{\theta}{2}$ and $Var(X_i) = \frac{\theta^2}{12} \Rightarrow \mathbb{E}(2\bar{X}) = \frac{2n\theta}{2n} = \theta$ [unbiased] and $Var(2\bar{X}) = \frac{4n\theta^2}{12n^2} = \frac{\theta^2}{3n}$ [consistent].

   As to $X_{(n)}$, we realize that $\frac{X_{(n)}}{\theta} \in \mathsf{beta}(n,1) \Rightarrow \mathbb{E}(\frac{X_{(n)}}{\theta}) = \frac{n}{n+1}$ and $Var(\frac{X_{(n)}}{\theta}) = \frac{n}{(n+1)^2\,(n+2)}$ [$X_{(n)}$ is consistent, but unbiased only asymptotically] $\Rightarrow \frac{n+1}{n}X_{(n)}$ is an unbiased estimator of θ, having the variance of $\frac{\theta^2}{(n+2)\,n}$. Its relative efficiency with respect to $2\bar{X}$ is therefore $\frac{n+2}{3}$ i.e., in the large-sample limit, $\frac{n+1}{n}X_{(n)}$ is 'infinitely' more efficient than $2\bar{X}$. But how can we establish whether $\frac{n+1}{n}X_{(n)}$ is the 'best' unbiased estimator, lacking the C-R bound? Obviously, something else is needed for cases (like this) which are not regular. This is the concept of

Sufficiency

which, in addition to providing a new criterion for being the 'best' estimator (of a regular case or *not*), will also help us *find* it (the C-R bound does not do that!).

Definition: $\hat{\Phi}(X_1, X_2, ...X_n)$ is called a **sufficient statistic** (*not* an *estimator* yet) for estimating θ iff the joint pdf of the sample $\prod\limits_{i=1}^{n} f(x_i|\theta)$ can be factorized into a product of a function of θ and $\hat{\Phi}$ only, times a function of all the x_is (but *no* θ), thus:

$$\prod_{i=1}^{n} f(x_i|\theta) \equiv g(\theta, \hat{\Phi}) \cdot h(x_1, x_2, ...x_n)$$

where $g(\theta, \hat{\Phi})$ must fully take care of the joint pdf's θ dependence, *including* the range's limits (L and H). Such $\hat{\Phi}$ (when it exists) 'extracts', from the RIS, all the information relevant for estimating θ. All we have to do to convert $\hat{\Phi}$ into the best possible *estimator* of θ is to make it *unbiased* (by some transformation, which is usually easy to design).

One can show that, if this transformation is *unique*, the resulting estimator is MVUE (best), even if it does not reach the C-R limit (but: it must be efficient at least asymptotically). To prove uniqueness, one has to show that $\mathbb{E}\left\{u(\hat{\Phi})\right\} \equiv 0$ (for each value of θ) implies $u(\hat{\Phi}) \equiv 0$, where $u(\hat{\Phi})$ is a function of $\hat{\Phi}$.

EXAMPLES:

1. Bernoulli distribution: $\prod\limits_{i=1}^{n} f(x_i|p) = p^{x_1+x_2+....+x_n}(1-p)^{n-x_1-x_2-...-x_n}$ is a function of p and of a *single combination* of the x_is, namely $\sum\limits_{i=1}^{n} x_i$. A sufficient statistic for estimating p is thus $\sum\limits_{i=1}^{n} X_i$ [we know how to make it into an *unbiased* estimator].

2. Normal distribution: $\prod\limits_{i=1}^{n} f(x_i|\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum\limits_{i=1}^{n} x_i^2}{2\sigma^2}\right) \times \exp\left(-\frac{n\mu^2 - 2\mu\sum\limits_{i=1}^{n} x_i}{2\sigma^2}\right)$ where the first factor (to the left of ×) contains no μ and the second factor is a function of only a single combination of the x_is, namely their sum. This leads to the same conclusion as in the previous example.

3. Exponential: $\prod\limits_{i=1}^{n} f(x_i|\beta) = \frac{1}{\beta^n}\exp\left(-\frac{1}{\beta}\sum\limits_{i=1}^{n} x_i\right) \Rightarrow$ ditto.

4. Referring to the same exponential distribution: what if the parameter to estimate is $\theta \equiv \frac{1}{\beta}$ (the rate of arrivals) rather than β itself. Then $\prod\limits_{i=1}^{n} f(x_i|\theta) = \theta^n \exp\left(-\theta \sum\limits_{i=1}^{n} x_i\right) \Rightarrow \sum\limits_{i=1}^{n} X_i \in \mathsf{gamma}(n, \frac{1}{\theta})$ is also a sufficient statistic for estimating θ, but now it is a lot more difficult to make it unbiased. By direct integration, we get: $\mathbb{E}\left(\frac{1}{\sum\limits_{i=1}^{n} X_i}\right) = \frac{\theta^n}{(n-1)!}\int\limits_0^\infty \frac{1}{u}\cdot u^{n-1}e^{-\theta\, u}\, du = \frac{(n-2)!\theta^n}{(n-1)!\theta^{n-1}} = \frac{\theta}{n-1} \Rightarrow \frac{n-1}{\sum\limits_{i=1}^{n} X_i}$ is an unbiased estimator of θ. Its variance can be shown (by a similar integration) to be equal to $\frac{\theta^2}{n-2}$, whereas the C-R bound yields $\frac{\theta^2}{n}$ [verify!]. Thus the 'efficiency' of $\frac{n-1}{\sum\limits_{i=1}^{n} X_i}$ is $\frac{n-2}{n}$, making it only *asymptotically* efficient [it is still the MVUE and therefore the *best* unbiased estimator in existence, i.e. 100% efficiency is, in this case, an impossible goal].

5. Gamma(k,β): $\prod\limits_{i=1}^{n} f(x_i|\beta) = \frac{\left(\prod\limits_{i=1}^{n} x_i\right)^{k-1}}{(k-1)!^n} \times \frac{\exp\left(-\frac{1}{\beta}\sum\limits_{i=1}^{n} x_i\right)}{\beta^{kn}}$, which makes $\sum\limits_{i=1}^{n} X_i$ a sufficient statistics for estimating β [similarly, $\prod\limits_{i=1}^{n} X_i$ would be a sufficient statistics for estimating k]. Since $\mathbb{E}(\sum\limits_{i=1}^{n} X_i) = nk\beta$, $\frac{\sum_{i=1}^{n} X_i}{nk}$ is the corresponding unbiased estimator. Its variance equals to $\frac{nk\beta^2}{(nk)^2} = \frac{\beta^2}{nk}$, which agrees with the C-R bound (verify!).

6. We can show that $X_{(n)}$ is a sufficient statistic for estimating θ of the uniform $\mathcal{U}(0,\theta)$ distribution.

   Proof: Introduce $G_{a,b}(x) \equiv \begin{cases} 0 & x < a \\ 1 & a \le x \le b \\ 0 & x > b \end{cases}$. The joint pdf of $X_1, X_2, \ldots, X_n$ can be written as $\dfrac{1}{\theta^n}\prod\limits_{i=1}^{n} G_{0,\theta}(x_i) = \dfrac{1}{\theta^n} G_{-\infty,\theta}(x_{(n)}) \times G_{0,\infty}(x_{(1)})$ where the first factor is a function of θ and $x_{(n)}$ *only*. □

   Knowing that $\mathbb{E}(X_{(n)}) = \frac{n}{n+1}\theta$ [as done earlier], we can easily see that $\frac{n+1}{n}X_{(n)}$ is an *unbiased* estimator of θ. Now we also know that it is the *best* estimator we can find for this purpose. ■

The only difficulty with the approach of this section arises when a sufficient statistic does not exist (try finding it for the Cauchy distribution). In that case, one can resort to using one of the following two techniques for finding an estimator of a parameter (or joint estimators of two or more parameters):

Method of moments

is the simpler of the two; it provides adequate (often 'best') estimators in most cases, but it can also, on occasion, result in estimators which are pathetically inefficient. It works like this: set each of the following expressions: $\mathbb{E}(X)$, $Var(X)$,

$\mathbb{E}\left[(X-\mu)^3\right]$, etc. [use as many of these as the number of parameters to be estimated – usually one or two] equal to its empirical equivalent, i.e.

$$\begin{aligned}
\mathbb{E}(X) &= \frac{\sum\limits_{i=1}^{n} X_i}{n} \; (\equiv \bar{X}) \\
Var(X) &= \frac{\sum\limits_{i=1}^{n} (X_i - \bar{X})^2}{n} \; (\equiv S^2) \\
\mathbb{E}\left[(X-\mu)^3\right] &= \frac{\sum\limits_{i=1}^{n} (X_i - \bar{X})^3}{n}
\end{aligned}$$

etc., then *solve* for the unknown parameters. The result yields the corresponding estimators (each a function of $\bar{X}$, S^2, etc. depending on the number of parameters). These will be asymptotically (but not necessarily fully) unbiased, consistent (but not necessarily efficient nor MVUE). The method fails when $\mathbb{E}(X)$ does not exist (Cauchy).

EXAMPLES:

One Parameter

1. Exponential $\mathcal{E}(\beta)$ distribution; estimating β.

   Solution: $\mathbb{E}(X) = \beta = \bar{X} \Rightarrow \hat{\beta} = \bar{X}$.

2. Uniform $\mathcal{U}(0, \theta)$ distribution; estimating θ.

   Solution: $\mathbb{E}(X) = \frac{\theta}{2} = \bar{X} \Rightarrow \hat{\theta} = 2\bar{X}$ [a very inefficient estimator].

3. Geometric distribution; estimate p.

   Solution: $\mathbb{E}(X) = \frac{1}{p} = \bar{X} \Rightarrow \hat{p} = \frac{1}{\bar{X}}$. One can show that $\mathbb{E}(\frac{1}{\bar{X}}) = p + \frac{pq}{n} - \frac{pq(p-q)}{n^2} + ...$ [biased].

   The following adjustment would make it into an unbiased estimator: $\hat{p} = \dfrac{1-\frac{1}{n}}{\bar{X}-\frac{1}{n}} \equiv \dfrac{n-1}{\sum\limits_{i=1}^{n} X_i - 1}$. Its variance is $\frac{p^2q}{n} + \frac{2p^2q^2}{n^2} + ...$ whereas the C-R bound equals to $\frac{p^2q}{n}$, so $\hat{p}$ is only asymptotically efficient.

4. Distribution given by $f(x) = \frac{2x}{a} e^{-\frac{x^2}{a}}$ for $x > 0$; estimate a.

   Solution: $\mathbb{E}(X) = \int\limits_0^\infty \frac{2x^2}{a} e^{-\frac{x^2}{a}} dx = \int\limits_0^\infty \sqrt{au} e^{-u}\, du$ [using the $u = \frac{x^2}{a}$ substitution] $= \sqrt{a}\Gamma(\frac{3}{2}) = \frac{1}{2}\sqrt{a\pi}$. Making this equal to $\bar{X}$ and solving for a yields: $\hat{a} = \frac{4\bar{X}^2}{\pi}$. Since $\mathbb{E}[\bar{X}^2] = \frac{n}{n^2}\mathbb{E}[X_1^2] + \frac{n(n-1)}{n^2}\mathbb{E}[X_1 \cdot X_2] = \frac{a}{n} + \frac{n-1}{n} \cdot \frac{a\pi}{4} \Rightarrow \mathbb{E}[\hat{a}] = a + \frac{a}{n}(\frac{4}{\pi} - 1)$, the estimator is unbiased only asymptotically [dividing it by $1 + \frac{1}{n}(\frac{4}{\pi} - 1)$ would fully remove the bias]. Establishing the asymptotic efficiency of the last (unbiased) estimator would get a bit messy (the more adventurous students may like to try it).

5. gamma(α, β): estimate β assuming α known.

   Solution: $\mathbb{E}(X) = \alpha\beta = \bar{X} \Rightarrow \widehat{\beta} = \frac{\bar{X}}{\alpha}$.

6. gamma(α, β): estimate α assuming β known.

   Solution: $\mathbb{E}(X) = \alpha\beta = \bar{X} \Rightarrow \widehat{\alpha} = \frac{\bar{X}}{\beta}$.

Two Parameters

1. For $\mathcal{N}(\mu, \sigma)$, estimate both μ and σ.

   Solution: $\mathbb{E}(X) = \mu = \bar{X}$ and $Var(X) = \sigma^2 = S^2 \Rightarrow \hat{\mu} = \bar{X}$ and $\hat{\sigma} = \sqrt{S^2}$ [the latter being unbiased only asymptotically].

2. For $\mathcal{U}(a, b)$ estimate both a and b.

   Solution: $\mathbb{E}(X) = \frac{a+b}{2} = \bar{X}$ and $Var(X) = \frac{(b-a)^2}{12} = S^2 \Rightarrow \hat{a} = \bar{X} + \sqrt{3S^2}$ and $\hat{b} = \bar{X} - \sqrt{3S^2}$ [this would prove to be a very *inefficient* way of estimating a and b].

3. Binomial, where *both* n and p need to be estimated.

   Solution: $\mathbb{E}(X) = np = \bar{X}$ and $Var(X) = npq = S^2 \Rightarrow \hat{p} = 1 - \frac{S^2}{\bar{X}}$ and $\hat{n} = \dfrac{\bar{X}}{1 - \frac{S^2}{\bar{X}}}$ (rounded to the nearest integer). Both estimators appear biased when explored by computer SIMULATION [generating many RISs using the binomial distribution with specific values of n and p, then computing $\hat{n}$ and $\hat{p}$ to see how they perform; in this case $\hat{p}$ is consistently overestimated and $\hat{n}$ underestimated].

4. beta(n, m), estimate both n and m.

   Solution: $\mathbb{E}(X) = \frac{n}{n+m} = \bar{X}$ [$\Rightarrow \frac{m}{n+m} = 1 - \bar{X}$] and $Var(X) = \frac{nm}{(n+m)^2(n+m+1)} = S^2 \Rightarrow \hat{n} = \bar{X} \cdot \left[\frac{\bar{X}(1-\bar{X})}{S^2} - 1\right]$ and $\hat{m} = (1 - \bar{X}) \cdot \left[\frac{\bar{X}(1-\bar{X})}{S^2} - 1\right]$.

5. gamma(α, β): estimate both parameters.

   Solution: $\mathbb{E}(X) = \alpha\beta = \overline{X}$ and $Var(X) = \alpha\beta^2 = S^2 \Rightarrow \widehat{\beta} = \frac{S^2}{\overline{X}}$ and $\widehat{\alpha} = \frac{\overline{X}^2}{S^2}$.

Maximum-likelihood technique

always performs very well; it *guarantees* to find the '*best*' estimators under the circumstances (even though they may be only asymptotically unbiased) – the major difficulty is that the estimators may turn out to be rather *complicated* functions of the X_is (to the extent that we may be able to find them only numerically, via computer optimization).

The **technique** for finding them is rather simple (in principle, not in technical detail): In the joint pdf of X_1, X_2, ..., X_n, i.e. in $\prod\limits_{i=1}^{n} f(x_i|\theta_1, \theta_2, ...)$, replace x_i by the actually observed value of X_i and maximize the resulting expression (called the LIKELIHOOD FUNCTION) with respect to θ_1, θ_2, ... The corresponding (optimal) θ-values are the actual parameter estimates. Note that it is frequently easier (yet equivalent) to maximize the *natural logarithm* of the likelihood function instead.

EXAMPLES:

One Parameter

1. Exponential distribution, estimating β.

   Solution: We have to maximize $-n\ln\beta - \frac{\sum_{i=1}^n X_i}{\beta}$ with respect to β. Making the corresponding first derivative equal to zero yields: $-\frac{n}{\beta} + \frac{\sum_{i=1}^n X_i}{\beta^2} = 0 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^n X_i}{n}$ [same as the method of moments].

2. Uniform distribution $\mathcal{U}(0,\theta)$, estimate θ.

   Solution: We have to maximize $\dfrac{1}{\theta^n} G_{-\infty,\theta}(X_{(n)}) \cdot G_{0,\infty}(X_{(1)})$ with respect to θ; this can be done by choosing the smallest possible value for θ while keeping $G_{-\infty,\theta}(X_{(n)}) = 1$. This is achieved by $\hat{\theta} = X_{(n)}$ [any smaller value of θ and $G_{0,\theta}(X_{(n)})$ drops down to 0]. We already know that this estimator has a small $\propto \frac{1}{n}$ bias and also how to fix it.

3. Geometric distribution, estimating p.

   Solution: Maximize $n\ln p + (\sum\limits_{i=1}^n X_i - n)\ln(1-p) \Rightarrow \frac{n}{p} - \frac{\sum_{i=1}^n X_i - n}{1-p} = 0 \Rightarrow \hat{p} = \frac{n}{\sum_{i=1}^n X_i}$ [same as the method of moments].

4. The distribution is given by $f(x) = \frac{2x}{a}e^{-\frac{x^2}{a}}$ for $x > 0$, estimate a.

   Solution: Maximize $n\ln 2 - n\ln a + \ln\prod\limits_{i=1}^n X_i - \frac{\sum_{i=1}^n X_i^2}{a} \Rightarrow -\frac{n}{a} + \frac{\sum_{i=1}^n X_i^2}{a^2} = 0 \Rightarrow \hat{a} = \frac{\sum_{i=1}^n X_i^2}{a} \equiv \overline{X^2}$. Since $\mathbb{E}(X_i^2) = a$ [done earlier], $\hat{a}$ is an unbiased estimator. Based on $\frac{\partial^2}{\partial a^2}[\ln(2X) - \ln a - \frac{X^2}{a}] = \frac{1}{a^2} - 2\frac{X^2}{a^3}$ (whose expected value equals to $-\frac{1}{a^2}$) the C-R bound is $\frac{a^2}{n}$. Since $Var(\overline{X^2}) = \frac{Var(X^2)}{n} = \frac{\mathbb{E}(X^4) - a^2}{n} = \frac{2a^2 - a^2}{n} = \frac{a^2}{n}$, our estimator is 100% efficient.

5. Normal distribution $\mathcal{N}(\mu,\sigma)$, assuming that μ is known, and σ^2 is to be estimated [a rather unusual situation].

   Solution: Maximize $\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2}$ with respect to $\sigma \Rightarrow -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^3} = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i-\mu)^2}{n}$ [clearly an unbiased estimator]. To assess its efficiency: C-R bound can be computed based on $\frac{\partial^2}{(\partial\sigma^2)^2}\left[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{2\sigma^2} - \frac{(x-\mu)^2}{\sigma^6}$, whose expected value is $-\frac{1}{2\sigma^4} \Rightarrow \frac{2\sigma^4}{n}$. Since $Var(\hat{\sigma}^2) = \mathbb{E}\left[\left(\frac{\sum_{i=1}^n [(X_i-\mu)^2-\sigma^2]^2}{n}\right)\right] = \frac{\mathbb{E}[(X-\mu)^4] - 2\mathbb{E}[(X-\mu)^2]\sigma^2 + \sigma^4}{n} = \frac{3\sigma^4 - 2\sigma^4 + \sigma^4}{n} = \frac{2\sigma^4}{n}$, our estimator is 100% efficient.

6. gamma(α,β): estimate β.

   Solution: Maximize $(\alpha - 1)\ln\prod\limits_{i=1}^n X_i - \frac{\sum_{i=1}^n X_i}{\beta} - n\ln\Gamma(\alpha) - n\alpha\ln\beta$. The β derivative yields: $\frac{\sum_{i=1}^n X_i}{\beta^2} - \frac{n\alpha}{\beta} = 0 \Rightarrow \widehat{\beta} = \frac{\overline{X}}{\alpha}$ (same as the method of moments).

7. gamma(α,β): estimate α.

Solution: Maximize $(\alpha-1)\ln\prod\limits_{i=1}^{n} X_i - \frac{\sum_{i=1}^{n} X_i}{\beta} - n\ln\Gamma(\alpha) - n\alpha\ln\beta$. The α derivative yields: $\ln\prod\limits_{i=1}^{n} X_i - n\,\psi(\alpha) - n\,\ln\beta = 0$ [where $\psi(\alpha)$ is the so called Euler's psi function] $\Rightarrow \psi(\alpha) = \sqrt[n]{\prod\limits_{i=1}^{n}\frac{X_i}{\beta}}$ (this is the geometric mean of the $\frac{X_i}{\beta}$ values) $\Rightarrow \widehat{\alpha} = \psi^{-1}\left(\sqrt[n]{\prod\limits_{i=1}^{n}\frac{X_i}{\beta}}\right)$.

8. Cauchy distribution with $f(x) = \frac{1}{\pi}\cdot\frac{1}{1+(x-a)^2}$, estimate a [the *location* of the 'laser gun', *knowing* its (unit) *distance* behind a screen]. Note that the method of moments would not work in this case.

   Solution: Maximize $-n\ln\pi - \sum\limits_{i=1}^{n}\ln[1+(X_i-a)^2] \Rightarrow \sum\limits_{i=1}^{n}\dfrac{X_i-a}{1+(X_i-a)^2} = 0$. This equation would have to be solved, for a, *numerically* [i.e. one would need a computer].

   Would this give us something substantially better than our (sensible but ad hoc) *sample median* $\tilde{X}$? Well, we know that the new estimator is asymptotically efficient, i.e. its variance approaches the C-R bound of $\dfrac{1}{n\mathbb{E}\left[\left(\frac{\partial\ln f}{\partial a}\right)^2\right]} = \dfrac{1}{\frac{n}{\pi}\int_{-\infty}^{\infty}\frac{4(x-a)^2\,dx}{[1+(x-a)^2]^3}} = \frac{2}{n}$. The variance of $\tilde{X}$ was $\frac{1}{4nf(a)^2} = \frac{\pi^2}{4n}$, so its relative efficiency is $\frac{8}{\pi^2} = 81.06\%$. The loss of 19% efficiency seems an acceptable trade off, since $\tilde{X}$ is so much easier to evaluate and (which is another substantial advantage over the 'best' estimator), it does *not* require the knowledge of the distance of the laser gun from the screen.

Two-parameters

1. The distribution is $\mathcal{N}(\mu,\sigma)$, estimate both μ and σ.

   Solution: Maximize $\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}$ by setting both derivatives equal to zero, i.e. $\frac{\sum_{i=1}^{n}(X_i-\mu)}{2\sigma^2} = 0$ and $-\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^3} = 0$, and solving for $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = \sqrt{S^2}$ (same as when using the method of moments).

2. Uniform $\mathcal{U}(a,b)$, estimate both limits a and b.

   Solution: Maximize $\frac{1}{(b-a)^n}\prod\limits_{i=1}^{n} G_{a,b}(X_i)$ by choosing a and b as close to each other as the G-functions allow (before dropping to zero). Obviously, a cannot be any bigger that $X_{(1)}$ and b cannot be any smaller than $X_{(n)}$, so these are the corresponding estimators [both slightly biased, but we know how to fix that]. These estimators are much better then what we got from the method of moments.

3. gamma(α,β), estimate both parameters.

Solution: Maximize $(\alpha-1)\ln\prod_{i=1}^{n}X_i-\frac{\sum_{i=1}^{n}X_i}{\beta}-n\ln\Gamma(\alpha)-n\alpha\ln\beta$. The two derivatives are: $\ln\prod_{i=1}^{n}X_i-n\,\psi(\alpha)-n\,\ln\beta=0$ and $\frac{\sum_{i=1}^{n}X_i}{\beta^2}-\frac{n\alpha}{\beta}=0$. Solving them jointly can be done only numerically.

4. Binomial distribution, with both n and p to be estimated.

   Solution: Maximize $N\ln(n!)-\ln\prod_{i=1}^{N}X_i!-\ln\prod_{i=1}^{N}(n-X_i)!+\ln p\sum_{i=1}^{N}X_i-\ln(1-p)\,(Nn-\sum_{i=1}^{N}X_i)$, where N is the sample size. Differentiating, we get $[\frac{\partial}{\partial n}:]$ $N\psi(n+1)-\sum_{i=1}^{N}\psi(n-X_i+1)-N\ln(1-p)=0$ and $[\frac{\partial}{\partial p}:]$ $\frac{\sum_{i=1}^{N}X_i}{p}-\frac{Nn-\sum_{i=1}^{N}X_i}{1-p}=0$. One can solve the second equation for $p=\frac{\sum_{i=1}^{N}X_i}{Nn}$, then substitute into the first equation and solve, *numerically*, for n. This would require a help of a computer, which is frequently the price to pay for high-quality estimators. ■

Chapter 6 CONFIDENCE INTERVALS

The last chapter considered the issue of so called POINT ESTIMATES (good, better and best), but one can easily see that, even for the best of these, a statement which claims a parameter, say μ, to be close to 8.3, is not very informative, unless we can specify what 'close' means. This is the purpose of a CONFIDENCE INTERVAL, which requires quoting the estimate together with specific limits, e.g. 8.3 ± 0.1 (or $8.2 \leftrightarrow 8.4$, using an interval form). The limits are established to meet a certain (usually 95%) LEVEL OF CONFIDENCE (not a probability, since the statement does not involve any randomness - we are either 100% right, or 100% wrong!).

The level of confidence ($1 - \alpha$ in general) corresponds to the original, A-PRIORI probability (i.e. before the sample is even taken) of the procedure to get it right (the probability is, as always, in the random sampling). To be able to calculate this probability *exactly*, we must know what distribution we are sampling from. So, until further notice, we will assume that the distribution is Normal.

CI for mean μ

We first assume that, even though μ is to be estimated (being unknown), we still **know** the exact (population) value of σ (based on past experience).

We know that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{6.1}$$

is standardized normal (usually denoted Z). This mean that

$$\Pr\left(\left|\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right) = \Pr\left(\left|\overline{X} - \mu\right| < z_{\alpha/2} \cdot \sigma/\sqrt{n}\right) = 1 - \alpha$$

where $z_{\alpha/2}$ (the so called CRITICAL VALUE) is easily found from tables (such as the last row of Table IV). Note that in general

$$\Pr(Z > z_a) = a$$

Usually, we need $\alpha/2 = 0.025$, which corresponds to 95% probability (eventually called confidence).

The random variable of the last statement is clearly $\overline{X}$ (before a sample is taken, and the value is computed). Assume now that the (random independent) sample has been taken, and $\overline{X}$ has been computed to have a specific value (8.3 say). The inequality in parentheses is then either true or false - the only trouble is that it contains μ whose value we don't know! We can thus solve it for μ, i.e.

$$\boxed{\overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \overline{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}}$$

and interpret this as a $100 \cdot (1 - \alpha)\%$ *confidence interval* for the exact (still unknown) value of μ.

σ unknown

In this case, we have to replace σ by the next best thing, which is of course sample standard deviation s. We know than, that the distribution of

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \tag{6.2}$$

changes from $\mathcal{N}(0,1)$ to t_{n-1}. This means that we also have to change $z_{\alpha/2}$ to $\mathsf{t}_{\alpha/2,n-1}$, the rest remains the same. A $100\cdot(1-\alpha)\%$ confidence interval for μ is then constructed by

$$\boxed{\overline{X}-\mathsf{t}_{\alpha/2,n-1}\cdot s/\sqrt{n}<\mu<\overline{X}+\mathsf{t}_{\alpha/2,n-1}\cdot s/\sqrt{n}}$$

Large-sample case

When n is 'large' ($n \geq 30$), there is little difference between $z_{\alpha/2}$ and $\mathsf{t}_{\alpha/2,n-1}$, so we would use $z_{\alpha/2}$ in either case.

Furthermore, both (6.1) and (6.2) are *approximately* Normal, even when the population is *not* (and, regardless what the distribution *is*). This means we can still construct an approximate confidence interval for μ (using σ if it's known, s when it isn't - $z_{\alpha/2}$ in either case).

Difference of two means

When two populations are Normal, with the same σ but potentially different μ, we already know (assuming the two samples be independent) that

$$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \tag{6.3}$$

is standardized normal (Z).

Proof. $\overline{X}_1 \in \mathcal{N}(\mu_1, \frac{\sigma}{\sqrt{n_1}})$ and $\overline{X}_2 \in \mathcal{N}(\mu_2, \frac{\sigma}{\sqrt{n_2}})$ implies $\overline{X}_1-\overline{X}_2 \in \mathcal{N}\left(\mu_1-\mu_2, \sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}\right)$ ■

The confidence interval for $\mu_1-\mu_2$ is thus

$$\boxed{\overline{X}_1-\overline{X}_2-z_{\alpha/2}\cdot\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}<\mu_1-\mu_2<\overline{X}_1-\overline{X}_2+z_{\alpha/2}\cdot\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

When σ is **not known**, (6.3) changes to

$$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \tag{6.4}$$

where

$$s_p \equiv \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

is called the POOLED sample standard deviation.

(6.4) now has the $\mathsf{t}_{n_1+n_2-2}$ distribution.

Proof. We need to proof that $\frac{(n_1+n_2-2)\cdot s_p^2}{\sigma^2} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{\sigma^2} \in \chi^2_{n_1+n_2-2}$ (automatically independent of $\overline{X}_1 - \overline{X}_2$). This follows from the fact that $\frac{(n_1-1)s_1^2}{\sigma^2} \in \chi^2_{n_1-1}$ and $\frac{(n_2-1)s_2^2}{\sigma^2} \in \chi^2_{n_2-1}$, and they are independent of each other. ■

The corresponding confidence interval is now

$$\boxed{\overline{X}_1 - \overline{X}_2 - \mathsf{t}_{\alpha/2,n_1+n_2-2} \cdot s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \mu_1 - \mu_2 < \overline{X}_1 - \overline{X}_2 + \mathsf{t}_{\alpha/2,n_1+n_2-2} \cdot s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

When the two σ's are **not identical** (but both **known**), we have

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} \in \mathcal{N}(0,1)$$

and the corresponding confidence interval:

$$\boxed{\overline{X}_1 - \overline{X}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{X}_1 - \overline{X}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

When the σ's are **unknown** (and have to be replaced by s_1 and s_2), we end up with a situation which has no simple distribution, *unless* both n_1 and n_2 are 'large'. In that case, we (also) don't have to worry about the normality of the population, and construct an *approximate* CI by:

$$\boxed{\overline{X}_1 - \overline{X}_2 - z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{X}_1 - \overline{X}_2 + z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

Proportion(s)

Here, we construct a CI for the p parameter of a binomial distribution. This usually corresponds to sampling, from an 'infinite' population with a certain percentage (or PROPORTION) of 'special' cases. We will deal only with the **large** n situation.

The X_1, X_2, ... X_n or our RIS now have values of either 1 (special case) or 0. This means that $\overline{X}$ equals to the SAMPLE PROPORTION of special cases, also denoted by $\widehat{p}$. We know that $\widehat{p}$ is, for large n, approximately Normal, with mean p and standard deviation of $\frac{p(1-p)}{n}$. One can actually take it one small step further, and show that

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \,\tilde{\in}\, \mathcal{N}(0,1)$$

One can thus construct an *approximate* CI for p by

$$\boxed{\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} < p < \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}$$

Similarly, for a **difference** between two p values (having two *independent* samples), we get the following *approximate* CI

$$\boxed{\widehat{p}_1 - \widehat{p}_2 - z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}+\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} < p_1 - p_2 < \widehat{p}_1 - \widehat{p}_2 + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}+\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}}$$

Variance(s)

We have to go back to assuming sampling from $\mathcal{N}(\mu, \sigma)$. To construct a $100\cdot(1-\alpha)\%$ confidence interval for the population variance σ^2, we just have to recall that

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

This implies that

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

where $\chi^2_{1-\alpha/2,n-1}$ and $\chi^2_{\alpha/2,n-1}$ are two critical values of the χ^2_{n-1} distribution (Table V). This time, they are both positive, with no symmetry to help.

The corresponding confidence interval for σ^2 is then

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

(the bigger critical value first).

To construct a CI for σ, we would just take the square root of these.

σ ratio

A CI for a ratio of two σ's (not a very common thing to do) is based on

$$\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_1^2}{\sigma_1^2}} \in \mathsf{F}_{n_1-1,n_2-1}$$

(assuming *independent* samples). This readily implies

$$\Pr\left(\mathsf{F}_{1-\frac{\alpha}{2},n_1-1,n_2-1} < \frac{s_1^2\cdot\sigma_2^2}{s_2^2\cdot\sigma_1^2} < \mathsf{F}_{\frac{\alpha}{2},n_1-1,n_2-1}\right)$$

which yields the final result:

$$\boxed{\frac{1}{\mathsf{F}_{\frac{\alpha}{2},n_1-1,n_2-1}}\cdot\frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\cdot\frac{1}{\mathsf{F}_{1-\frac{\alpha}{2},n_1-1,n_2-1}} = \frac{s_1^2}{s_2^2}\cdot\mathsf{F}_{\frac{\alpha}{2},n_2-1,n_2-1}}$$

Note that

$$\begin{aligned}\frac{\alpha}{2} &= \Pr\left(\mathsf{F}_{n_2-1,n_1-1} > \mathsf{F}_{\frac{\alpha}{2},n_2-1,n_2-1}\right) = \Pr\left(\frac{1}{\mathsf{F}_{n_1-1,n_2-1}} > \mathsf{F}_{\frac{\alpha}{2},n_2-1,n_2-1}\right)\\ &= \Pr\left(\mathsf{F}_{n_1-1,n_2-1} < \frac{1}{\mathsf{F}_{\frac{\alpha}{2},n_2-1,n_2-1}}\right)\end{aligned}$$

implies that

$$\Pr\left(\mathsf{F}_{n_1-1,n_2-1} > \frac{1}{\mathsf{F}_{\frac{\alpha}{2},n_2-1,n_2-1}} \equiv \mathsf{F}_{1-\frac{\alpha}{2},n_1-1,n_2-1}\right) = 1-\frac{\alpha}{2}$$

The critical values are in Table VI (but only for 90% and 98% confidence levels)!

Chapter 7 TESTING HYPOTHESES

Suppose now that, instead of trying to estimate μ, we would like it to be equal to (or at least reasonably close to) some desired, specific value called μ_0. To test whether it is (the so called NULL HYPOTHESIS, say H_0: $\mu = 500$) or is not (the ALTERNATE HYPOTHESIS H_A: $\mu \neq 500$) can be done, in this case, in one of two ways:

1. Construct the corresponding CI for μ, and see whether it contains 500 (if it does, ACCEPT H_0, otherwise, REJECT it). The corresponding α (usually 5%) is called the LEVEL OF SIGNIFICANCE.

2. Compute the value of the so called TEST STATISTIC

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

(it has the Z distribution *only* when H_0 is true) and see whether it falls in the corresponding ACCEPTANCE REGION $[-z_{\alpha/2}, z_{\alpha/2}]$ or REJECTION REGION (outside this interval).

Clearly, the latter way of performing the test is *equivalent* to the former. Even though it appears more elaborate (actually, it is a touch easier computationally), it is the 'standard' way to go.

Test statistics are usually constructed with the help of the corresponding LIKELIHOOD FUNCTION, something we learned about in the previous chapter.

There are two types of error we can make:

- Rejecting H_0 when it's true - this is called TYPE I ERROR.

- Accepting it when it's false - TYPE II ERROR.

The probability of making Type I error is obviously equal to α (under our control).

The probability of Type II error (β) depends on the actual value of μ (and σ) - we can compute it and plot it as a function of μ (OC curve) - when μ approaches (but is not equal to) μ_0, this error clearly reaches $1 - \alpha$. Equivalently, they sometimes plot $1 - \beta$ (the POWER FUNCTION) instead - we like it big (close to 1).

Two notes concerning alternate hypotheses:

When H_A consists of (infinitely) many possibilities (such as our $\mu \neq 500$ example), it is called COMPOSITE. This is the usual case.

When H_A considers only one specific possibility (e.g. $\mu = 400$), it is called SIMPLE. In practice, this would be very unusual - we will not be too concerned with it here.

To us, a more important distinction is this:

Sometimes (as in our example), the alternate hypothesis has the $\neq$ sign, indicating that we don't like a deviation from μ_0 either way (e.g. 500 mg is the amount of aspirin we want in one pill - it should not be smaller, it should not be bigger) - this is called a TWO-SIDED hypothesis.

Frequently (this is actually even more common), we need to make sure that μ meets the specifications one way (amount of coke in one bottle is posted as 350 mL, we want to avoid the possibility that $\mu < 350$ - the ONE-SIDED alternate hypothesis). In this case, the null hypothesis is sometimes still stated in the old manner of H_0: $\mu = 350$, sometimes they put is as H_0: $\mu \geq 350$. In any case, the null hypothesis *must* always have the = sign!

When the alternate hypothesis is one-sided, so is the corresponding rejection region (also called ONE-TAILED), which would now consist of the $(-\infty, z_\alpha)$ interval - note that now a single tail get the full α. Note that now the correspondence between this test and a CI for μ becomes more complicated (we would normally *not* use CI in this case).

In either case (one or two sided), these is yet another alternate (but equivalent) way of performing the test (bypassing the critical region). It works like this:

- For a two-sided test, we compute the value of the test statistic (let us call it $\mathbf{t}$), which it then converted into the so called P-VALUE, thus:

$$P = 2\Pr(Z > |\mathbf{t}|)$$

  When this P value is less than α, we reject H_0 (accept otherwise).

- For a one-sided test, whenever $\mathbf{t}$ is on the H_0 side, we accept H_0 without having to compute anything. When $\mathbf{t}$ is on the H_A side, we compute

$$P = \Pr(Z > |\mathbf{t}|)$$

  and reject H_0 when P is smaller than α, accept otherwise.

Tests concerning mean(s)

We need to specify the assumptions, null hypothesis, test statistic, and its distribution (under H_0) - the rest is routine.

Assume:	H_0	T	Distribution of T
Normal population, σ known	$\mu = \mu_0$	$\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	Z
Normal population, σ unknown	$\mu = \mu_0$	$\dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$	t_{n-1}
Any population, large n, σ unknown	$\mu = \mu_0$	$\dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$	Z
Two Normal populations, same unknown σ	$\mu_1 = \mu_2$	$\dfrac{(\overline{X}_1 - \overline{X}_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$\mathsf{t}_{n_1+n_2-2}$

Concerning variance(s)

Assume:	H_0	T	Distribution of T
Normal population	$\sigma = \sigma_0$	$\frac{(n-1)s^2}{\sigma_0^2}$	χ^2_{n-1}
Two Normal populations	$\sigma_1 = \sigma_2$	$\frac{s_1^2}{s_2^2}$	F_{n_1-1,n_2-1}

Concerning proportion(s)

Assume:	H_0	T	Distribution of T
One population, large n	$p = p_0$	$\frac{\widehat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Z (approximate)
k populations, large samples	$p_1 = p_2 = \ldots = p_k$	$\frac{\sum_{i=1}^k n_i(\widehat{p}_i - \widehat{\widehat{p}})^2}{\widehat{\widehat{p}}(1-\widehat{\widehat{p}})}$	χ^2_{k-1} (approximate)

Contingency tables

Here, we have two (nominal scale) attributes (e.g. cities and detergent brands - see your textbook), and we want to know whether they are *independent* (i.e. customers in different cities having the same detergent preferences - H_0, or not - H_A).

Example 1

	Brand A	*Brand B*	*Brand C*
Montreal	*87*	*62*	*12*
Toronto	*120*	*96*	*23*
Vancouver	*57*	*49*	*9*

The numbers are called OBSERVED FREQUENCIES, denoted o_{ij} (i is the row, j the column label).

First, we have to compute the corresponding EXPECTED FREQUENCIES (assuming independence) by

$$e_{ij} = \frac{(\sum_{j=1}^{c} o_{ij}) \cdot (\sum_{i=1}^{r} o_{ij})}{\sum_{j=1}^{c} \sum_{j=1}^{r} o_{ij}}$$

To be able to proceed, these must be all bigger than 5.

The test statistic equals

$$\sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

and has (under H_0), approximately, the $\chi^2_{(r-1)(c-1)}$, where r (c) is the number of rows (columns) respectively.

Goodness of fit

This time, we are testing whether a random variable has a specific distribution, say Poisson (we will stick to discrete cases).

First, based on the data, we estimate the value of each unknown parameter (this distribution has only one), based on what we learned in the previous chapter.

Then, we compute the expected frequency of each possible outcome (all integers, in this case), by

$$e_i = n \times f(i \mid \widehat{\lambda}) = n \times \frac{\widehat{\lambda}^i}{i!} e^{-\widehat{\lambda}} \qquad\qquad i = 0, 1, 2, ...$$

where n is the *total* frequency and $\widehat{\lambda} = \frac{\sum_i i \cdot o_i}{\sum_i o_i}$ is the usual λ estimator. We have to make sure that none of the expected frequencies is less than 5 (otherwise, we *pool* outcomes to achieve this).

The test statistic is

$$T = \sum_i \frac{(o_i - e_i)^2}{o_i}$$

Under H_0 (which now states: the distribution is Poisson, with unspecified λ), T has the χ^2 distribution with $m-1-p$ degrees of freedom, where m is the number of possible outcomes (after pooling), and p is the number of (unspecified) parameters, to be estimated based on the original data (in this case, $p = 1$).

Chapter 8 LINEAR REGRESSION AND CORRELATION

We will first consider the case of having one 'independent' (REGRESSOR) variable, called x, and a 'dependent' (RESPONSE) variable y. This is called

Simple regression

The **model** is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{8.1}$$

where $i = 1, 2, ..., n$, making the following assumptions:

1. The values of x are measured 'exactly', with no random error. This is usually so when we can choose them at will.

2. The ε_i are normally distributed, independent of each other (uncorrelated), having the expected value of 0 and variance equal to σ^2 (the same for each of them, regardless of the value of x_i). Note that the actual value of σ is usually not known.

The two regression coefficients are called the SLOPE AND INTERCEPT. Their actual values are also unknown, and need to be estimated using the empirical data at hand.

To find such ESTIMATORS, we use the

Maximum likelihood method

which is almost always the best tool for this kind of task. It guarantees to yield estimators which are ASYMPTOTICALLY UNBIASED, having the smallest possible variance. It works as follows:

1. We write down the joint probability density function of the y_i's (note that these *are* random variables).

2. Considering it a function of the parameters (β_0, β_1 and σ in this case) *only* (i.e. 'freezing' the y_i's at their *observed* values), we *maximize* it, using the usual techniques. The values of β_0, β_1 and σ to yield the maximum value of this so called LIKELIHOOD FUNCTION (usually denoted by $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}$) are the actual estimators (note that they will be functions of x_i *and* y_i).

Note that instead of maximizing the likelihood function itself, we may choose to maximize its logarithm (which must yield the same $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}$).

Least-squares technique

In our case, the Likelihood function is:

$$L = \frac{1}{(\sqrt{2\pi}\sigma)^n} \prod_{i=1}^{n} \exp\left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right]$$

and its logarithm:

$$\ln L = -\frac{n}{2}\log(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

To maximize this expression, we first differentiate it with respect to σ, and make the result equal to zero. This yields:

$$\widehat{\sigma}_m = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n}}$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the values of β_0 and β_1 which minimize

$$SS_e \equiv \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

namely the sum of squares of the *vertical* deviations (or RESIDUALS) of the y_i values from the fitted straight line (this gives the technique its name).

To find $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we have to differentiate SS_e, separately, with respect to β_0 and β_1, and set each of the two answers to zero. This yields:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1}^{n} y_i - n\,\beta_0 - \beta_1\sum_{i=1}^{n} x_i = 0$$

and

$$\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1}^{n} x_i\,y_i - \beta_0\sum_{i=1}^{n} x_i - \beta_1\sum_{i=1}^{n} x_i^2 = 0$$

or equivalently, the following so called

Normal equations

$$\begin{aligned} n\,\beta_0 + \beta_1\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \beta_0\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i\,y_i \end{aligned}$$

They can be solved easily for β_0 and β_1 (at this point we can start calling them $\widehat{\beta}_0$ and $\widehat{\beta}_1$):

$$\widehat{\beta}_1 = \frac{n\sum\limits_{i=1}^{n} x_i\,y_i - \sum\limits_{i=1}^{n} x_i \cdot \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} \equiv \frac{S_{xy}}{S_{xx}}$$

and

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} \tag{8.2}$$

meaning that the regression line passes through the $(\overline{x}, \overline{y})$ point, where

$$\overline{x} \equiv \frac{\sum\limits_{i=1}^{n} x_i}{n}$$

and

$$\overline{y} \equiv \frac{\sum\limits_{i=1}^{n} y_i}{n}$$

Each $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is clearly a linear combination of normally distributed random variables, their joint distribution is thus of the bivariate normal type.

Statistical properties of the estimators

First, we should realize that it is the y_i (not x_i) which are random, due to the ε_i term in (8.1) - both β_0 and β_1 are also fixed, albeit unknown parameters. Clearly then

$$\mathbb{E}(y_i - \overline{y}) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 \overline{x}) = \beta_1 (x_i - \overline{x})$$

which implies

$$\mathbb{E}\left(\widehat{\beta}_1\right) = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x}) \cdot \mathbb{E}(y_i - \overline{y})}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} = \beta_1$$

Similarly, since $\mathbb{E}(\overline{y}) = \beta_0 + \beta_1 \overline{x}$, we get

$$\mathbb{E}\left(\widehat{\beta}_0\right) = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0$$

Both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are thus UNBIASED estimators of β_0 and β_1, respectively.

To find their respective variance, we first note that

$$\widehat{\beta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \equiv \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})\, y_i}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}$$

(right?), based on which

$$\operatorname{Var}\left(\widehat{\beta}_1\right) = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2 \cdot \operatorname{Var}(y_i)}{\left(\sum\limits_{i=1}^{n} (x_i - \overline{x})^2\right)^2} = \frac{\sigma^2 S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

>From (8.2) we get

$$\operatorname{Var}\left(\widehat{\beta}_0\right) = \operatorname{Var}(\overline{y}) - 2\overline{x}\operatorname{Cov}(\overline{y}, \widehat{\beta}_1) + \overline{x}^2 \operatorname{Var}\left(\widehat{\beta}_1\right)$$

We already have a formula for $\operatorname{Var}\left(\widehat{\beta}_1\right)$, so now we need

$$\operatorname{Var}(\overline{y}) = \operatorname{Var}(\overline{\varepsilon}) = \frac{\sigma^2}{n}$$

and

$$\operatorname{Cov}(\overline{y},\widehat{\beta}_1)=\operatorname{Cov}\left(\frac{\sum\limits_{i=1}^{n}\varepsilon_i}{n},\frac{\sum\limits_{i=1}^{n}(x_i-\overline{x})\,\varepsilon_i}{S_{xx}}\right)=\frac{\sigma^2\sum\limits_{i=1}^{n}(x_i-\overline{x})}{S_{xx}}=0$$

(uncorrelated). Putting these together yields:

$$\operatorname{Var}\left(\widehat{\beta}_0\right)=\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)$$

The covariance between $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is thus equals to $-\overline{x}\operatorname{Var}(\widehat{\beta}_1)$, and their correlation coefficient is

$$\frac{-1}{\sqrt{1+\frac{1}{n}\cdot\frac{S_{xx}}{\overline{x}^2}}}$$

Both variance formulas contain σ^2, which, in most situations, must be replaced by its ML estimator

$$\widehat{\sigma}_m^2=\frac{\sum\limits_{i=1}^{n}(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}{n}\equiv\frac{SS_e}{n}$$

where the numerator defines the so called RESIDUAL (ERROR) SUM OF SQUARES. It can be rewritten in the following form (replacing $\widehat{\beta}_0$ by $\overline{y}-\widehat{\beta}_1\overline{x}$):

$$\begin{aligned}SS_e &= \sum_{i=1}^{n}(y_i-\overline{y}+\widehat{\beta}_1\overline{x}-\widehat{\beta}_1x_i)^2=\sum_{i=1}^{n}\left[y_i-\overline{y}+\widehat{\beta}_1(\overline{x}-x_i)\right]^2\\ &= S_{yy}-2\widehat{\beta}_1S_{xy}+\widehat{\beta}_1^2S_{xx}=S_{yy}-2\frac{S_{xy}}{S_{xx}}S_{xy}+\left(\frac{S_{xy}}{S_{xx}}\right)^2S_{xx}\\ &= S_{yy}-\frac{S_{xy}}{S_{xx}}S_{xy}=S_{yy}-\widehat{\beta}_1S_{xy}\equiv S_{yy}-\widehat{\beta}_1^2S_{xx}\end{aligned}$$

Based on (8.1) and $\overline{y}=\beta_0+\beta_1\overline{x}+\overline{\varepsilon}$ (from now on, we have to be very careful to differentiate between β_0 and $\widehat{\beta}_0$, etc.), we get

$$\mathbb{E}(S_{yy})=\mathbb{E}\left\{\sum_{i=1}^{n}\left[\beta_1(x_i-\overline{x})+(\varepsilon_i-\overline{\varepsilon})\right]^2\right\}=\beta_1^2\,S_{xx}+\sigma^2(n-1)$$

(the last term was derived in MATH 2F81). Furthermore,

$$\mathbb{E}\left(\widehat{\beta}_1^2\right)=\operatorname{Var}(\widehat{\beta}_1)-\mathbb{E}(\widehat{\beta}_1)^2=\frac{\sigma^2}{S_{xx}}-\beta_1^2$$

Combining the two, we get

$$\mathbb{E}(SS_e)=\sigma^2(n-2)$$

Later on, we will be able to prove that $\frac{SS_e}{\sigma^2}$ has the χ^2 distribution with $n-2$ degrees of freedom. It is also *independent* of each $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

This means that there is a slight bias in the $\widehat{\sigma}_m^2$ estimator of σ^2 (even though the bias disappears in the $n \to \infty$ limit - such estimators are called ASYMPTOTICALLY UNBIASED). We can easily fix this by defining a new, fully unbiased

$$\widehat{\sigma}^2 = \frac{SS_e}{n-2} \equiv MS_e$$

(the so called MEAN SQUARE) to be used instead of $\widehat{\sigma}_m^2$ from now on.

All of this implies that both

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{MS_e\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}}$$

and

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_e}{S_{xx}}}} \tag{8.3}$$

have the Student $\mathtt{t}$ distribution with $n-2$ degrees of freedom. This can be used either to construct the so called CONFIDENCE INTERVAL for either β_0 or β_1, or to test any HYPOTHESIS concerning β_0 or β_1.

Confidence intervals

Knowing that (8.3) has the $\mathtt{t}_{n-2}$ distribution, we must then find two values (called CRITICAL) such that the probability of (8.3) falling inside the corresponding interval (between the two values) is $1-\alpha$. At the same time, we would like to have the interval as short as possible. This means that we will be choosing the critical values symmetrically around 0; the positive one will equal to $\mathtt{t}_{\frac{\alpha}{2},n-2}$, the negative one to $-\mathtt{t}_{\frac{\alpha}{2},n-2}$ (the first index now refers to the area of the remaining TAIL of the distribution) - these critical values are widely tabulated.

The statement that (8.3) falls in the interval between the two critical values of $\mathtt{t}_{n-2}$ is equivalent (solve the corresponding equation for β_1) to saying that the value of β_1 is in the following range

$$\widehat{\beta}_1 \pm \mathtt{t}_{\frac{\alpha}{2},n-2}\sqrt{\frac{MS_e}{S_{xx}}}$$

which is our $(1-\alpha)\cdot 100\,\%$ confidence interval.

Similarly, we can construct a $1-\alpha$ level-of-confidence interval for $\widehat{\beta}_0$, thus:

$$\widehat{\beta}_0 \pm \mathtt{t}_{\frac{\alpha}{2},n-2}\sqrt{MS_e\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$$

Note that, since $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are not independent, making a joint statement about the two (with a specific level of confidence) is more complicated (one has to construct a confidence ellipse, to make it correct).

Constructing a $1-\alpha$ confidence interval for σ^2 is a touch more complicated. Since $\frac{SS_e}{\sigma^2}$ has the χ^2_{n-2} distribution, we must first find the corresponding two critical values. Unfortunately, the χ^2 distribution is not symmetric, so for these two we have to take $\chi^2_{\frac{\alpha}{2},n-2}$ and $\chi^2_{1-\frac{\alpha}{2},n-2}$. Clearly, the probability of a χ^2_{n-2} random variable falling between the two values equals $1-\alpha$. The resulting interval may not be the shortest of all these, but we are obviously quite close to the right solution; furthermore, the choice of how to divide α between the two tails remains simple and logical.

Solving for σ^2 yields

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

as the corresponding $(1-\alpha)\cdot 100\,\%$ confidence interval.

Correlation

Suppose now that *both* x and y are random, normally distributed with (bivariate) parameters μ_x, μ_y, σ_x, σ_y and ρ. We know that the *conditional* distribution of y *given* x is also (univariate) normal, with the following conditional mean and variance:

$$\mu_y + \rho\,\sigma_y \frac{x-\mu_x}{\sigma_x} \equiv \beta_0 + \beta_1\, x \tag{8.4}$$

$$\sigma_y^2\,(1-\rho^2)$$

The usual $\widehat{\beta}_0$ and $\widehat{\beta}_1$ estimators are still the 'best' (maximizing the likelihood function), but their statistical properties are now substantially more complicated.

Historical comment: Note that by reversing the rôle of x and y (which is now quite legitimate - the two variables are treated as 'equals' by this model), we get the following regression line:

$$\mu_{x\,|\,y} = \mu_x + \rho\,\sigma_x \frac{y-\mu_y}{\sigma_y}$$

One can easily see that this line is inconsistent with (8.4) - it is a lot *steeper* when plotted on the same graph. Ordinary regression thus tends, in this case, to distort the true relationship between x and y, making it either more flat or more steep, depending on which variable is taken to be the 'independent' one.

Thus, for example, if x is the height of fathers and y that of sons, the regression line will have a slope less than 45 degrees, implying a false averaging trend (regression towards the mean, as it was originally called - and the name, even though ultimately incorrect, stuck). The fallacy of this argument was discovered as soon as someone got the bright idea to fit y against x, which would then, still falsely, imply a tendency towards increasing diversity.

One can show that the ML technique would use the usual $\overline{x}$ and $\overline{y}$ to estimate μ_x and μ_y, $\sqrt{\frac{S_{xx}}{n-1}}$ and $\sqrt{\frac{S_{yy}}{n-1}}$ (after unbiasing) to estimate σ_x and σ_y, and

$$r \equiv \frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}} \tag{8.5}$$

as an estimator of ρ (for some strange reason, they like calling the estimator r rather than the usual $\widehat{\rho}$). This relates to the fact that

$$\frac{S_{xy}}{n-1}$$

is an unbiased estimator of $\mathrm{Cov}(X,Y)$.
Proof.

$$\mathbb{E}\left\{\sum_{i=1}^{n}\left[x_i-\mu_x-(\overline{x}-\mu_x)\right]\left[y_i-\mu_y-(\overline{y}-\mu_y)\right]\right\}=$$
$$\sum_{i=1}^{n}\left[\mathrm{Cov}(X,Y)-\frac{\mathrm{Cov}(X,Y)}{n}-\frac{\mathrm{Cov}(X,Y)}{n}+\frac{\mathrm{Cov}(X,Y)}{n}\right]=$$
$$n\,\mathrm{Cov}(X,Y)\,(1-\frac{1}{n})=\mathrm{Cov}(X,Y)\,(n-1)$$

■

Investigating statistical properties of r, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ *exactly* is now short of impossible (mainly because of dividing by $\sqrt{S_{xx}}$, which is random) - now, we have to resort to LARGE-SAMPLE approach, to derive ASYMPTOTIC formulas only (i.e. expanded in powers of $\frac{1}{n}$), something we will take up shortly.

This is also how one can show that

$$\text{arctanh } r$$

approaches, for 'large' n, the Normal distribution (with the mean of arctanh $\rho+\frac{\rho}{2n}+...$ and variance of $\frac{1}{n-3}+...$) a lot faster than r itself. Utilizing this, we construct an approximate CI for arctanh ρ:

$$\text{arctanh } r-\frac{r}{2n}\pm\frac{z_{\alpha/2}}{\sqrt{n-3}}$$

and consequently for ρ (take tanh of each limit).

Squaring the r estimator yields the so called COEFFICIENT OF DETERMINATION

$$r^2=\frac{S_{yy}-S_{yy}+\dfrac{S_{xy}^2}{S_{xx}}}{S_{yy}}=1-\frac{SS_E}{S_{yy}}$$

which tells us how much of the original y variance has been removed by fitting the best straight line.

Multiple regression

This time, we have k independent (regressor) variables $x_{1,}$ x_2,..., x_k; still only one dependent (response) variable y. The model is

$$y_i=\beta_0+\beta_1x_{1,i}+\beta_2x_{2,i}+...+\beta_kx_{k,i}+\varepsilon_i$$

with $i = 1, 2, ..., n$, where the first index labels the variable, and the second the observation. It is more convenient now to switch to using the following matrix notation

$$\mathbf{y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are (column) vectors of length n, $\boldsymbol{\beta}$ is a (column) vector of length $k+1$, and $\mathbb{X}$ is a n by $k+1$ matrix of observations (with its first column having all elements equal to 1, the second column being filled by the observed values of x_1, etc.). Note that the exact values of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are, and will always remain, unknown to us (thus, they must not appear in any of our computational formulas).

To minimize the sum of squares of the residuals (a SCALAR quantity), namely

$$\begin{aligned}&(\mathbf{y} - \mathbb{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbb{X}\boldsymbol{\beta}) = \\ &\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbb{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbb{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbb{X}^T\mathbb{X}\boldsymbol{\beta}\end{aligned}$$

(note that the second and third terms are identical - why?), we differentiate it with respect to each element of $\boldsymbol{\beta}$. This yields the following vector:

$$-2\mathbb{X}^T\mathbf{y} + 2\mathbb{X}^T\mathbb{X}\boldsymbol{\beta}$$

Making these equal to zero provides the following maximum likelihood (least square) estimators of the regression parameters:

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y} \equiv \boldsymbol{\beta} + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\varepsilon}$$

The last form makes it clear that $\widehat{\boldsymbol{\beta}}$ are *unbiased* estimators of $\boldsymbol{\beta}$, normally distributed with the variance-covariance matrix of

$$\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$$

The 'fitted' values of $\mathbf{y}$ (let us call them $\widehat{\mathbf{y}}$), are computed by

$$\widehat{\mathbf{y}} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}\ \boldsymbol{\beta} + \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\varepsilon} \equiv \mathbb{X}\ \boldsymbol{\beta} + \mathbb{H}\ \boldsymbol{\varepsilon}$$

where $\mathbb{H}$ is clearly *symmetric* and *idempotent* (i.e. $\mathbb{H}^2 = \mathbb{H}$). Note that $\mathbb{H}\mathbb{X} = \mathbb{X}$.

This means that the residuals e_i are computed by

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbb{I} - \mathbb{H})\boldsymbol{\varepsilon}$$

($\mathbb{I} - \mathbb{H}$ is also idempotent). Furthermore, the covariance (matrix) between the elements of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and those of $\mathbf{e}$ is:

$$\begin{aligned}&\mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\mathbf{e}^T\right] = \mathbb{E}\left[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbb{H})\right] = \\ &(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right](\mathbb{I} - \mathbb{H}) = \mathbb{O}\end{aligned}$$

which means that the variables are uncorrelated and therefore *independent* (i.e. each of the regression-coefficient estimators is independent of each of the residuals – slightly counter-intuitive but correct nevertheless).

The sum of squares of the residuals, namely $\mathbf{e}^T\mathbf{e}$, is equal to

$$\boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbb{H})^T(\mathbb{I} - \mathbb{H})\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbb{H})\boldsymbol{\varepsilon}$$

Divided by σ^2:

$$\frac{\boldsymbol{\varepsilon}^T(\mathbb{I}-\mathbb{H})\boldsymbol{\varepsilon}}{\sigma^2} \equiv \mathbf{Z}^T(\mathbb{I}-\mathbb{H})\mathbf{Z}$$

where $\mathbf{Z}$ are standardized, independent and normal.

We know (from matrix theory) that any symmetric matrix (including our $\mathbb{I}-\mathbb{H}$) can be written as $\mathbb{R}^T\mathbb{D}\mathbb{R}$, where $\mathbb{D}$ is *diagonal* and $\mathbb{R}$ is *orthogonal* (implying $\mathbb{R}^T \equiv \mathbb{R}^{-1}$). We can then rewrite the previous expression as

$$\mathbf{Z}^T\mathbb{R}^T\mathbb{D}\mathbb{R}\mathbf{Z} = \widetilde{\mathbf{Z}}^T\mathbb{D}\widetilde{\mathbf{Z}}$$

where $\widetilde{\mathbf{Z}} \equiv \mathbb{R}\mathbf{Z}$ is still a set of standardized, independent Normal random variables (since its variance-covariance matrix equals $\mathbb{I}$). Its distribution is thus χ^2 if and only if the diagonal elements of $\mathbb{D}$ are all equal either to 0 or 1 (the number of degrees being equal to the *trace* of $\mathbb{D}$).

How can we tell whether this is true for our $\mathbb{I}-\mathbb{H}$ matrix (when expressed in the $\mathbb{R}^T\mathbb{D}\mathbb{R}$ form) *without* actually performing the diagonalization (a fairly tricky process). Well, such a test is not difficult to design, once we notice that $(\mathbb{I}-\mathbb{H})^2 = \mathbb{R}^T\mathbb{D}\mathbb{R}\mathbb{R}^T\mathbb{D}\mathbb{R} = \mathbb{R}^T\mathbb{D}^2\mathbb{R}$. Clearly, $\mathbb{D}$ has the proper form (only 0 or 1 on the main diagonal) if and only if $\mathbb{D}^2 = \mathbb{D}$, which is the same as saying that $(\mathbb{I}-\mathbb{H})^2 = \mathbb{I}-\mathbb{H}$ (which we already know is true). This then implies that the sum of squares of the residuals has χ^2 distribution. Now, how about its degrees of freedom? Well, since the trace of $\mathbb{D}$ is the same as the trace of $\mathbb{R}^T\mathbb{D}\mathbb{R}$ (a well known property of trace), we just have to find the trace of $\mathbb{I}-\mathbb{H}$, by

$$\begin{aligned}&\mathrm{Tr}\,[\mathbb{I}-\mathbb{H}] = \mathrm{Tr}\,(\mathbb{I}_{n\times n}) - \mathrm{Tr}\,(\mathbb{H}) = n - \mathrm{Tr}\left(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right) = \\ &n - \mathrm{Tr}\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\right) = n - \mathrm{Tr}\left(\mathbb{I}_{(k+1)\times(k+1)}\right) = n-(k+1)\end{aligned}$$

i.e. the number of observations minus the number of regression coefficients.

The sum of squares of the residuals is usually denoted SS_e (for 'error' sum of squares, even though it is usually called RESIDUAL SUM OF SQUARES) and *computed* by

$$\begin{aligned}&(\mathbf{y}-\mathbb{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbb{X}\widehat{\boldsymbol{\beta}}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbb{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbf{y} + \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbb{X}\widehat{\boldsymbol{\beta}} = \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbb{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbf{y} + \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbb{X}\widehat{\boldsymbol{\beta}} \equiv \\ &\mathbf{y}^T\mathbf{y} - \widehat{\boldsymbol{\beta}}^T\mathbb{X}^T\mathbf{y}\end{aligned}$$

We have just proved that $\frac{SS_e}{\sigma^2}$ has the χ^2 distribution with $n-(k+1)$ degrees of freedom, and is independent of $\widehat{\boldsymbol{\beta}}$. A related definition is that of a RESIDUAL (error) MEAN SQUARE

$$MS_e \equiv \frac{SS_e}{n-(k+1)}$$

This would clearly be our unbiased estimator of σ^2.

Various standard errors

We would thus construct a confidence interval for any one of the β coefficients, say β_j, by

$$\widehat{\beta}_j \pm \mathsf{t}_{\frac{\alpha}{2},n-k-1} \cdot \sqrt{C_{jj}\cdot MS_E}$$

where $\mathbb{C} \equiv (\mathbb{X}^T\mathbb{X})^{-1}$.

Similarly, to test a hypothesis concerning a single β_j, we would use

$$\frac{\widehat{\beta}_j - \beta_{i_0}}{\sqrt{C_{jj} \cdot MS_E}}$$

as the test statistic.

Since the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is $\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$, we know that

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbb{X}^T \mathbb{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2}$$

has the χ^2_{k+1} distribution. Furthermore, since the β's are independent of the residuals,

$$\frac{\dfrac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbb{X}^T \mathbb{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{k+1}}{\dfrac{SS_E}{n-k-1}}$$

must have the $\mathsf{F}_{k+1,n-k-1}$ distribution. This enables us to construct a CONFIDENCE ELLIPSE (ellipsoid) simultaneously for all parameters or, correspondingly, perform a single test of H_0: $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$.

Chapter 9 **ANALYSIS OF VARIANCE**

One-way ANOVA

Suppose we have k Normal populations, having the same σ, but potentially different means μ_1, $\mu_2, ..\ \mu_k$. We want to test whether all these means are identical (the null hypothesis) or not (alternate).

We do this by first selecting a random independent sample of size n, independently from each of the k populations (note that, to simplify matters, we use the same sample size for all), and estimating each of the μ_i's by

$$\bar{X}_{(i)} \equiv \frac{\sum_{j=1}^{n} X_{ij}}{n}$$

(these of course will always be different from each other).

Secondly (to decide whether 'what we see is what we get'), we need a test statistic which meets the following two conditions:

1. It is *sensitive to any deviation* from the null hypothesis (it should return a *small* value when H_0 holds, a *large* value otherwise)

2. It has a *known distribution* under H_0 (to decide what is 'small' and what is 'large').

The RV which meets these objectives is

$$T = \frac{\dfrac{n\sum\limits_{i=1}^{k}(\bar{X}_{(i)} - \overline{X})^2}{k-1}}{\dfrac{\sum\limits_{i=1}^{k} s_{(i)}^2}{k}} \tag{9.1}$$

where $\overline{X}$ is the GRAND mean (mean of means) of all the nk observations put together, and $\bar{X}_{(i)}$ and $s_{(i)}^2$ are the individual sample means and variances, where $i = 1, 2, ...\ k$. Let us recall that $\bar{X}_{(i)} \in \mathcal{N}(\mu_i, \frac{\sigma}{\sqrt{n}})$ and $\frac{n-1}{\sigma^2} s_{(i)}^2 \in \chi_{n-1}^2$, for each i.

Note that the numerator of the formula will be small when the population means are identical (the sample means will be close to each other, and to their grand mean), becoming large when they are not. On the other hand the denominator of the formula (the average of the individual sample variances) merely estimates the common σ^2, and is totally insensitive to potential differences between population means.

To figure out the distribution of this test statistic *when* H_0 is *true*, we notice that $\bar{X}_{(1)}$, $\bar{X}_{(2)}$, ..., $\bar{X}_{(k)}$ effectively constitute a RIS of size k from $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$, and $\dfrac{\sum\limits_{i=1}^{k}(\bar{X}_{(i)} - \overline{X})^2}{k-1}$ is thus the corresponding sample variance. This means that $\dfrac{\sum\limits_{i=1}^{k}(\bar{X}_{(i)} - \overline{X})^2}{\sigma^2/n}$ has the χ_{k-1}^2 distribution.

Similarly $\dfrac{(n-1)\sum\limits_{i=1}^{k} s_{(i)}^2}{\sigma^2}$ is just a sum of k independent χ_{n-1}^2 RVs, whose distribution is $\chi_{k(n-1)}^2$ (the degrees of freedom simply add up). The ratio of $\dfrac{n}{\sigma^2}\dfrac{\sum\limits_{i=1}^{k}(\bar{X}_{(i)}-\overline{X})^2}{k-1}$ by $\dfrac{1}{k\sigma^2}\sum\limits_{i=1}^{k} s_{(i)}^2$ (note that σ^2 cancels out, leading to T) has therefore the distribution of $\dfrac{\dfrac{\chi_{k-1}^2}{k-1}}{\dfrac{\chi_{k(n-1)}^2}{k(n-1)}} \equiv \mathsf{F}_{k-1,k(n-1)}$.

The only thing left to do is to figure out some efficient way to compute T (this used to be important in the pre-computer days, but even current textbooks cannot leave it alone - like those silly tables of Poisson distribution in the Appendix). It is not difficult to figure out that

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\overline{X})^2 = n\sum_{i=1}^{k}(\bar{X}_{(i)}-\overline{X})^2 + (n-1)\sum_{i=1}^{k} s_{(i)}^2$$

or $SS_T = SS_B + SS_W$, where the subscripts stand for TOTAL, BETWEEN (or treatment) and WITHIN (or error) sum of squares, respectively.

Furthermore, one can show that

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}\right)^2}{kn} \tag{9.2}$$

and

$$SS_B = \frac{\sum_{i=1}^{k}\left(\sum_{j=1}^{n} X_{ij}\right)^2}{n} - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}\right)^2}{kn}$$

which is how these two quantities are efficiently computed (with $SS_W = SS_T - SS_B$).

The whole computation (of T) is the summarized in the following table:

Source	df	SS	MS	T
Between	$k-1$	SS_B	$MS_B = \frac{SS_B}{k-1}$	$\frac{MS_B}{MS_W}$
Within	$k(n-1)$	SS_W	$MS_W = \frac{SS_W}{k(n-1)}$	
Total	$kn-1$	SS_T		

Two-way ANOVA

In the previous section, the population index ($i = 1, 2, ... k$) can be seen as a (nominal scale) variable, which is, in this context called a FACTOR (e.g. labelling the city from which the observation is taken). In some situations, we may need more than one factor (e.g. white, black, Hispanic) - we will only discuss how to deal with **two**. (For a sake of example, we will take the response variable X to represent a person's salary).

No interaction

Our DESIGN will first:

1. assume that there is no INTERACTION between the factors (meaning that racial biases - if they exist - do not vary from city to city).
2. randomly select only one representative for each CELL (one employee of each race form every city).

The former implies that $X_{ij} \in \mathcal{N}(\mu + \alpha_i + \beta_j, \sigma)$, where $\sum_{i=1}^{k} \alpha_i = 0$ and $\sum_{j=1}^{m} \beta_j = 0$ (k and m is the number of LEVELS of the first and second factor).

To estimate the individual parameters, we would clearly use

$$\begin{aligned}
\widehat{\mu} &= \overline{X} \equiv \frac{\sum_{j=1}^{m} \sum_{i=1}^{k} X_{ij}}{mk} \\
\widehat{\alpha}_i &= \bar{X}_{(i\bullet)} - \overline{X} \equiv \frac{\sum_{j=1}^{m} X_{ij}}{m} - \overline{X} \\
\widehat{\beta}_j &= \bar{X}_{(\bullet j)} - \overline{X} \equiv \frac{\sum_{i=1}^{k} X_{ij}}{k} - \overline{X}
\end{aligned}$$

This time, we can test several null hypotheses at once: One stating that all the α's equal to zero (no difference between cities), the other claiming that same for the β's (no difference between racial groups), and the last one setting them all (α's and β's) to zero.

The total sum of squares is computed as before (see 9.2), except that n changes to m. Similarly, one can show that now

$$SS_T = SS_A + SS_B + SS_E \tag{9.3}$$

where SS_A (SS_B) is the sum of squares due to the first (second) factor and computed by

$$\begin{aligned}
SS_A &= m \sum_{i=1}^{k} \widehat{\alpha}_i^2 = \frac{\sum_{i=1}^{k} \left(\sum_{j=1}^{m} X_{ij} \right)^2}{m} - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{m} X_{ij} \right)^2}{km} \\
SS_B &= k \sum_{j=1}^{m} \widehat{\beta}_j^2 = \frac{\sum_{j=1}^{m} \left(\sum_{i=1}^{k} X_{ij} \right)^2}{k} - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{m} X_{ij} \right)^2}{km}
\end{aligned}$$

and SS_E is the ERROR (residual) sum of squares (computed form 9.3).

The summary will now look as follows:

Source	df	SS	MS	T
Factor A	$k-1$	SS_A	$MS_A = \frac{SS_A}{k-1}$	$\frac{MS_A}{MS_E}$
Factor B	$m-1$	SS_B	$MS_B = \frac{SS_B}{m-1}$	$\frac{MS_B}{MS_E}$
Error	$(k-1)(m-1)$	SS_E	$MS_E = \frac{SS_E}{(k-1)(m-1)}$	
Total	$km-1$	SS_T		

With interaction

Now, we assume a possible interaction between the two factors (the pattern of racial bias may differ between cities), which necessitates selecting more than one (say n) random employees from each cell. The (theoretical) mean of the $X_{ij\ell}$ distribution will now equal to $\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, where $\sum_{i=1}^{k} (\alpha\beta)_{ij} = 0$ for each j and $\sum_{j=1}^{m} (\alpha\beta)_{ij} = 0$ for each i.

The corresponding estimators are now

$$\begin{aligned}
\widehat{\mu} &= \overline{X} \equiv \frac{\sum_{\ell=1}^{n} \sum_{j=1}^{m} \sum_{i=1}^{k} X_{ij\ell}}{nmk} \\
\widehat{\alpha}_i &= \bar{X}_{(i\bullet\bullet)} - \overline{X} \equiv \frac{\sum_{\ell=1}^{n} \sum_{j=1}^{m} X_{ij\ell}}{nm} - \overline{X} \\
\widehat{\beta}_j &= \bar{X}_{(\bullet j\bullet)} - \overline{X} \equiv \frac{\sum_{\ell=1}^{n} \sum_{i=1}^{k} X_{ij\ell}}{nk} - \overline{X} \\
(\widehat{\alpha\beta})_{ij} &= \bar{X}_{(ij\bullet)} - \overline{X} - \widehat{\alpha}_i - \widehat{\beta}_j \equiv \frac{\sum_{\ell=1}^{n} X_{ij\ell}}{n} - \overline{X} - \widehat{\alpha}_i - \widehat{\beta}_j
\end{aligned}$$

For the total sum of squares, we now get

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

where

$$\begin{aligned}
SS_T &= \sum_{\ell=1}^{n} \sum_{j=1}^{m} \sum_{i=1}^{k} (X_{ij\ell} - \overline{X})^2 \\
SS_A &= nm \sum_{i=1}^{k} \widehat{\alpha}_i^2 \\
SS_B &= nk \sum_{j=1}^{m} \widehat{\beta}_i^2 \\
SS_{AB} &= n \sum_{j=1}^{m} \sum_{i=1}^{k} (\widehat{\alpha\beta})_{ij}^2
\end{aligned}$$

In summary:

Source	df	SS	MS	T
Factor A	$k-1$	SS_A	$MS_A = \frac{SS_A}{k-1}$	$\frac{MS_A}{MS_E}$
Factor B	$m-1$	SS_B	$MS_B = \frac{SS_B}{m-1}$	$\frac{MS_B}{MS_E}$
Interaction	$(k-1)(m-1)$	SS_{AB}	$MS_{AB} = \frac{SS_{AB}}{(k-1)(m-1)}$	$\frac{MS_{AB}}{MS_E}$
Error	$km(n-1)$	SS_E	$MS_E = \frac{SS_E}{km(n-1)}$	
Total	$kmn-1$	SS_T		

Chapter 10 NONPARAMETRIC TESTS

These don't make any assumption about the shape of the distribution from which we sample (they are equally valid for distributions of any shape). As a result, they may not be as powerful ('sharp') as test designed for a specific (usually Normal) distribution.

Sign test

The null hypothesis states that the population median equals a specific number, H_0: $\tilde{\mu} = \tilde{\mu}_0$ If we throw in an assumption that the distribution is symmetric, the median is the same as the mean, so we can restate it in those terms.

We also assume that the distribution is continuous (or essentially so), so that the probability of any observation being exactly equal to $\tilde{\mu}_0$ is practically zero (if we do get such a value, we would have to discard it).

The test statistic (say B) is simply the number of observations (out of n) which are bigger than $\tilde{\mu}_0$ (sometimes, these are represented by + signs, thus the name of the test). Its distribution is, under H_0, obviously Binomial, where n is the number of trials, and $p = \frac{1}{2}$. The trouble is that, due to B's discreteness, we cannot arbitrarily set the value of α, and have to settle for anything reasonable close to, say 5%. That's why, in this case, we are better off simply stating the corresponding P value.

When n is 'large', it is permissible to approximate the Binomial distribution by Normal, which leads to a modified test statistic

$$T = \frac{B - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{2B - n}{\sqrt{n}}$$

with critical values of $\pm z_{\alpha/2}$ (two-sided test), or either z_α or $-z_{\alpha/2}$ (one sided test).

The previous test is often used in the context of so called PAIRED SAMPLES (such as taking a blood pressure of individuals *before* and *after* taking some medication). In this case, we are concerned with the distribution of the *difference* in blood pressure, testing whether its population median stayed the same (the null hypothesis), or decreased (alternate). This time, we assign + to increase, and − to decrease, the rest is the same.

Signed-rank test

A better (more powerful) test is, under the same circumstances, the so called WILCOXON SIGNED-RANK TEST.

First, we compute the differences between individual observations and $\tilde{\mu}_0$ (in the case of a ONE-SAMPLE test), or between the paired observations (paired-sample test). Then, we RANK (i.e. assign 1, 2,... n) the (absolute value) differences (discarding zero differences, and assigning the corresponding rank average to any ties). The test statistic equals the sum of these ranks of all *positive* differences (denoted T^+).

The distribution of T^+ under H_0 (which states that the median difference equals zero) is not one of our 'common' cases, that's why its critical values are tabulated in Table X. We of course are in a good position to compute the corresponding P value ourselves (with the help of Maple). All we need to do is to assign a random sign (with equal probability for + and −) to the first n integers.

It's quite easy to show that the mean and variance of the T^+ distribution are $\frac{n(n+1)}{4}$ and $\frac{n(n+1)(2n+1)}{24}$ respectively.

Proof. We can write $T^+ = 1 \cdot X_1 + 2 \cdot X_2 + 3 \cdot X_3 + ... + n \cdot X_n$, where the X_i's are independent, having the Bernoulli distribution with $p = \frac{1}{2}$. This implies the mean of $\frac{1+2+3+...n}{2} = \frac{n(n+1)}{4}$. Similarly, $\text{Var}(T^+) = \frac{1^2+2^2+3^2+...+n^2}{4} = \frac{n(n+1)(2n+1)}{24}$.

To derive formulas for $s_1 \equiv \sum_{i=1}^n i$ and $s_2 \equiv \sum_{i=1}^n i^2$, we proceed as follows: $\sum_{i=0}^n (1+i)^2 = s_2 + (n+1)^2$, but is also equals (by expanding) to $n+1+2s_1+s_2$. Make these two equal, and solve for s_1.

Similarly $\sum_{i=0}^n (1+i)^3 = s_3 + (n+1)^3 = n+1+3s_1+3s_2+s_3$. Since s_3 cancels out, and we already know what s_1 is, we can solve for s_2. ■

For $n \geq 15$, it is quite legitimate to treat the distribution of T^+ as approximately Normal.

Rank-sum tests

Mann-Whitney

Suppose we have two distributions (of the same - up to a 'shift' - shape) and the corresponding *independent* (no longer paired) samples. We want to test whether the two sample means are identical (the null hypothesis) against one the three possible (>, < or ≠) alternate hypotheses.

We do this by ranking the $n_1 + n_2$ observations pooled together (as if a single sample), then we compute the sum of the ranks 'belonging' to the first sample, usually denoted W_1. The corresponding test statistic is

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2}$$

Under $H_{0,}$ the distribution of U_1 is symmetric (even though 'uncommon'), with the smallest possible value of 0 and the largest equal to $\frac{(n_1+n_2)(n_1+n_2+1)}{2} - \frac{n_2(n_2+1)}{2} - \frac{n_1(n_1+1)}{2} = n_1 n_2$. Its critical values are listed in Table XI. To compute them, one has to realize that the distribution of W_1 is that of the sum of randomly selected n_1 integers out of the first $n_1 + n_2$ (again, we may try doing this with our own Maple program).

It is reasonable to use the Normal approximation when both n_1 and n_2 are bigger than 8. The expected value of U_1 is $\frac{n_1 n_2}{2}$ (the center of symmetry), its variance is equal to $\frac{n_1 n_2 (n_1+n_2+1)}{12}$.

Proof. Suppose the numbers are selected randomly, one by one (*without* replacement). W_1 is then equal to $X_1 + X_2 + ... + X_{n_1}$, where X_i is the number selected in the i^{th} draw. Clearly, $\mathbb{E}(X_i) = \frac{n_1+n_2+1}{2}$ for each i. This implies that the expected value of W_1 is $n_1 \frac{n_1+n_2+1}{2}$, and that of U_1 equals $n_1 \frac{n_1+n_2+1}{2} - \frac{n_1(n_1+1)}{2} = \frac{n_1 n_2}{2}$.

Similarly, $\mathrm{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} = \frac{N^2-1}{12}$, where $N \equiv n_1 + n_2$, and $\mathrm{Cov}(X_i, X_j) = \frac{N(N+1)^2}{4(N-1)} - \frac{(N+1)(2N+1)}{6(N-1)} - \frac{(N+1)^2}{4} = -\frac{N+1}{12}$ for any $i \neq j$. This means that the variance of W_1 (and also of U_1) is $n_1 \mathrm{Var}(X_i) + n_1(n_1 - 1)\mathrm{Cov}(X_i, X_j) = \frac{n_1 n_2 (n_1+n_2+1)}{12}$. ■

Kruskal-Wallis

This is a generalization of the previous test to the case of more than two (say k) same-shape populations, testing whether all the means are identical (H_0) or not (H_A). It is a non-parametric analog to ANOVA.

Again, we rank all the $N \equiv n_1 + n_2 + ... + n_k$ observations pooled together, then compute the sum of the resulting ranks (say R_i) individually for each sample. The test statistic is

$$T = \frac{12}{N(N+1)} \cdot \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

and has, approximately (for large n_i), the χ^2_{k-1} distribution.

Proof. First we show that T can be written as

$$\frac{12}{N(N+1)} \cdot \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{N+1}{2} \right)^2$$

This follows from:
$\sum_{i=1}^k n_i \left(\frac{R_i}{n_i} - \frac{N+1}{2} \right)^2 = \sum_{i=1}^k \left[\frac{R_i^2}{n_i} - (N+1)R_i + n_i \frac{(N+1)^2}{4} \right] = \sum_{i=1}^k \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{2} + \frac{N(N+1)^2}{4} = \sum_{i=1}^k \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}$

It was already shown in the previous section that, for large n_i and N, $S_i \equiv \frac{R_i}{n_i} - \frac{N+1}{2}$ is approximately Normal with zero mean and variance equal to $\frac{(N-n_i)(N+1)}{12 n_i}$. Similarly, $\mathrm{Cov}(S_i, S_j) = -\frac{N+1}{12}$. This means that the variance-covariance matrix of the$\sqrt{\frac{12 n_i}{N(N+1)}} S_i$'s is

$$\mathbb{I} - \begin{bmatrix} \sqrt{\frac{n_1}{N}} \\ \sqrt{\frac{n_2}{N}} \\ \vdots \\ \sqrt{\frac{n_k}{N}} \end{bmatrix} \otimes \begin{bmatrix} \sqrt{\frac{n_1}{N}} & \sqrt{\frac{n_2}{N}} & \cdots & \sqrt{\frac{n_k}{N}} \end{bmatrix}$$

This matrix is clearly idempotent, which makes the sum of squares of the$\sqrt{\frac{12 n_i}{N(N+1)}} S_i$'s into a χ^2 type RV. The degrees of freedom are given by the Trace of the previous matrix, which is $k - 1$. ■

Run test

This is to test whether a *sequence* of observations constitutes a random *independent* sample or not. We assume that the observations are of the success (S) and failure (F) type - any other sequence can be converted into that form, one way or another.

A series of *consecutive* successes (or failures) is called a RUN. Clearly, in a truly random sequence, runs should never be 'too long' (but not consistently 'too short', either). This also means that, in a random sequence with n successes and

m failures, we should not have too many or too few runs in total (the total number of runs will be our test statistic T).

This time it is possible to derive a formula for the corresponding distribution:

- The sample space consists of $\binom{n+m}{n}$ equally likely possibilities (there are that many 'words' with n letters S and m letters F).

- To partition n letters S into k groups of at least one letter, we must first 'reserve' one letter for each group. This leaves us with $n-k$ to further distribute among the k groups (the 'circle and bar' game), which can be done in $\binom{n-1}{k-1}$ ways. Similarly, to partition m letters F into k groups can be done in $\binom{m-1}{k-1}$ ways. Finally, we have to decide whether to start with an S of F run (2 choices). The probability that T equals $2k$ is thus computed by
$$f(2k) = \frac{2\binom{n-1}{k-1}\binom{m-1}{k-1}}{\binom{n+m}{n}}$$
for $k = 1,\ 2, ...,\ \min(n,m)$.

- The other possibility (in addition to having the *same* number of S and F runs) is that these differ by one (i.e. k S runs and $k+1$ F runs, or the other way around). The probability of T equal to $2k+1$ is thus
$$f(2k+1) = \frac{\binom{n-1}{k}\binom{m-1}{k-1} + \binom{n-1}{k-1}\binom{m-1}{k}}{\binom{n+m}{n}}$$
where $k = 1,\ 2, ...,\ \max(\min(n, m-1), \min(n-1, m))$.

Based on these formulas, we can easily compute (with the help of Maple) tables of the corresponding distribution, and figure out the critical values for any particular n, m and α.

We can also find the mean and variance of the corresponding distribution, with the help of
$$\mu = \sum_{\text{All } k} 4k\frac{\binom{n-1}{k-1}\binom{m-1}{k-1}}{\binom{n+m}{n}} + \sum_{\text{All } k}(2k+1)\frac{\binom{n-1}{k}\binom{m-1}{k-1} + \binom{n-1}{k-1}\binom{m-1}{k}}{\binom{n+m}{n}} = \frac{2nm}{n+m} + 1$$
and
$$\begin{aligned}&\sum_{\text{All } k} 8k^2\frac{\binom{n-1}{k-1}\binom{m-1}{k-1}}{\binom{n+m}{n}} + \sum_{\text{All } k}(2k+1)^2\frac{\binom{n-1}{k}\binom{m-1}{k-1} + \binom{n-1}{k-1}\binom{m-1}{k}}{\binom{n+m}{n}} \\ = &\ \frac{4nm(n+1)(m+1) + (n+m)^2 - 10nm - n - m}{(n+m)(n+m-1)}\end{aligned}$$
which results in
$$\begin{aligned}\sigma^2 &= \frac{4nm(n+1)(m+1) + (n+m)^2 - 10nm - n - m}{(n+m)(n+m-1)} - \left(\frac{2nm}{n+m} + 1\right)^2 \\ &= \frac{2nm(2nm-n-m)}{(n+m)^2(m+m-1)}\end{aligned}$$

For both n and m bigger than 9, the distribution of T is approximately Normal. This is when the formulas for μ and σ^2 would come handy.

(Sperman's) rank correlation coefficient

All we have to do is to rank, individually, the X and Y observations (from 1 to n), and compute the *regular* correlation coefficient between the ranks. This simplifies to

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where d_i is the difference between the ranks of the X and Y observation of the i^{th} pair.

Proof. Let $\hat{X}_i$ and $\hat{Y}_i$ denote the ranks. We know that, individually, their sum is $\frac{n(n+1)}{2}$, and their sum of squares equals to $\frac{n(n+1)(2n+1)}{6}$. Furthermore

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (\hat{X}_i - \hat{Y}_i)^2 = \frac{n(n+1)(2n+1)}{3} - 2\sum_{i=1}^{n} \hat{X}_i \hat{Y}_i$$

This implies

$$\begin{aligned} r_S &= \frac{\sum_{i=1}^{n} \hat{X}_i \hat{Y}_i - \frac{(\sum_{i=1}^{n} \hat{X}_i)(\sum_{i=1}^{n} \hat{Y}_i)}{n}}{\sqrt{\left[\sum_{i=1}^{n} \hat{X}_i^2 - \frac{(\sum_{i=1}^{n} \hat{X}_i)^2}{n}\right] \cdot \left[\sum_{i=1}^{n} \hat{Y}_i^2 - \frac{(\sum_{i=1}^{n} \hat{Y}_i)^2}{n}\right]}} \\ &= \frac{\frac{n(n+1)(2n+1)}{6} - \frac{\sum_{i=1}^{n} d_i^2}{2} - \frac{n(n+1)^2}{4}}{\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}} = 1 - \frac{\frac{\sum_{i=1}^{n} d_i^2}{2}}{\frac{n(n^2-1)}{12}} \end{aligned}$$

■

For relatively small n, we can easily construct the distribution of r_S, assuming that X and Y are *independent* (and design the corresponding test for testing that as the null hypothesis).

When n is 'large' (bigger than 10), the distribution of r_S is approximately Normal. To be able to utilize this, we need to know the corresponding mean and variance under H_0. These turn out to be 0 and $\frac{1}{n-1}$, respectively.

Proof. What we need is

$$\begin{aligned} \mathbb{E}(d_i^2) &= \frac{\sum_{k=1}^{n} \sum_{\ell=1}^{n} (k-\ell)^2}{n^2} = \frac{n^2-1}{6} \\ \mathbb{E}(d_i^4) &= \frac{\sum_{k=1}^{n} \sum_{\ell=1}^{n} (k-\ell)^4}{n^2} = \frac{(n^2-1)(2n^2-3)}{30} \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}(d_i^2 d_j^2) = \frac{\sum_{k=1}^{n} \sum_{\ell=1}^{n} (k-j)^2 \sum_{K \neq k} \sum_{L \neq \ell} (K-L)^2}{n^2(n-1)^2} \\ = \frac{\sum_{k=1}^{n} \sum_{\ell=1}^{n} (k-j)^2 \sum_{K=1}^{n} \sum_{L=1}^{n} (K-L)^2}{n^2(n-1)^2} \\ - \frac{2\sum_{k=1}^{n} \sum_{\ell=1}^{n} (k-j)^2 \sum_{L=1}^{n} (k-L)^2}{n^2(n-1)^2} + \frac{\sum_{k=1}^{n} \sum_{\ell=1}^{n} (k-j)^4}{n^2(n-1)^2} \\ = \frac{(5n^3 - 7n^2 + 18)(n+1)}{180} \end{aligned}$$

implying: $\mathrm{Var}(d_i^2) = \frac{7n^2-13}{180}(n^2-1)$ and $\mathrm{Cov}(d_i^2, d_j^2) = -\frac{2n^2-5n-13}{180}(n+1)$.

Based on this,

$$\mathbb{E}\left(\sum_{i=1}^{n} d_i^2\right) = \frac{n(n^2-1)}{6}$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{n} d_i^2\right) = n\frac{7n^2-13}{180}(n^2-1) - n(n-1)\frac{2n^2-5n-13}{180}(n+1) = \frac{n^2(n^2-1)^2}{36(n-1)}$$

■