## TRANSFORMING RVs

of continuous type only.

The main issue: Given the distribution of $X$, find the distribution of $Y \equiv \frac{1}{1+X}$ (any expression involving $X$).

This is called univariate transformation.

Later, we also discuss bivariate transformations, say $U \equiv \frac{X}{X+Y}$ (or any other expression involving $X$ and $Y$).

Another simple example: $V \equiv X + Y$ (adding two RVs is tricky).

Let us start with

### UNIVARIATE TRANSFORMATION

There are two basic techniques:

**Distribution-Function** $(F)$ **Technique** which works as follows:

Take $Y \equiv g(X)$. We can find its distribution function by

$$F_Y(y) = \Pr(Y < y) = \Pr[g(X) < y]$$

i.e. solving the $g(X) < y$ inequality for $X$ (usually an interval), and then integrating $f(x)$ over this interval.

: EXAMPLES:

1) Consider $X \in \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$ [two-directional pointer attached to a spinning wheel]. Find the distribution of

$$Y = b \tan(X) + a$$

(location of a dot the pointer would leave on a screen placed $b$ units from the wheel's center, with a scale whose origin is $a$ units off the center).

Solution:

$$F_X(x) = \frac{x + \frac{\pi}{2}}{\pi} \equiv \frac{x}{\pi} + \frac{1}{2}$$

where $-\frac{\pi}{2} < x < \frac{\pi}{2}$.

$$\begin{aligned} F_Y(y) &= \Pr[b\tan(X) + a < y] = \\ \Pr[X &< \arctan(\frac{y-a}{b})] = F_X[\arctan(\frac{y-a}{b})] = \\ & \frac{1}{\pi}\arctan\left(\frac{y-a}{b}\right) + \frac{1}{2} \end{aligned}$$

where $-\infty < y < \infty$.

This implies that

$$f_Y(y) = \frac{1}{\pi b} \cdot \frac{1}{1 + \left(\frac{y-a}{b}\right)^2} = \frac{b}{\pi} \cdot \frac{1}{b^2 + (y-a)^2}$$

It looks similar to Normal ('bell-shaped'), yet these two are world apart.

The name of this new distribution is **Cauchy** [notation: $\mathcal{C}(a, b)$].

Since the $\int_{-\infty}^{\infty} y \cdot f_Y(y)\, dy = \infty - \infty$, the Cauchy distribution does not have a mean (also, its variance is infinite). As a consequence, the central limit theorem does not apply.

Yet the distribution has a clear center (at $y = a$) and width ($\pm b$). These are the median $\tilde{\mu}_Y = a$ [verify by solving $F_Y(\tilde{\mu}) = \frac{1}{2}$] and semi-inter-quartile range (quartile deviation) $\frac{Q_U - Q_L}{2}$ where $Q_U$ and $Q_L$ are the upper and lower quartiles [defined by $F(Q_U) = \frac{3}{4}$ and $F(Q_L) = \frac{1}{4}$] - in this case, $Q_L = a - b$ and $Q_U = a + b$.

The simplest case is $\mathcal{C}(0, 1)$, whose pdf equals

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

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

Solution: From graph, $0 < Y < 1$.

Secondly,

$$\begin{aligned} F_X(x) &= 6\int (x - x^2)\, dx = \\ 6(\frac{x^2}{2} - \frac{x^3}{3}) &= 3x^2 - 2x^3 \end{aligned}$$

And finally:

$$\begin{aligned} F_Y(y) &\equiv \Pr(Y < y) = \Pr(X^3 < y) = \\ \Pr(X &< y^{\frac{1}{3}}) = F_X(y^{\frac{1}{3}}) = 3y^{\frac{2}{3}} - 2y \end{aligned}$$

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: $F_X(x) = x$ for $0 < x < 1$.

Then:

$$\begin{aligned} F_Y(y) &= \Pr(-\ln X < y) = \\ \Pr(X &> e^{-y}) = 1 - F_X(e^{-y}) = 1 - e^{-y} \end{aligned}$$

where $y > 0$.

The corresponding pdf is thus $e^{-y}$ (exponential distribution, with $\mu = 0$).

Note that $Y = -\beta \cdot \ln X$ would result in exponential distribution with $\mu = \beta$.

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

Solution:

$$\begin{aligned} F_Y(y) &= \Pr(Z^2 < y) = \\ \Pr(-\sqrt{y} &< Z < \sqrt{y}) = F_Z(\sqrt{y}) - F_Z(\sqrt{y}) \end{aligned}$$

Since we don't have an explicit expression for $F_Z(z)$, it would appear that we are stuck.

But, we can get the corresponding $f_Y(y)$ by a simple differentiation:

$$\begin{aligned} \frac{dF_Z(\sqrt{y})}{dy} - \frac{dF_Z(-\sqrt{y})}{dy} &= \\ \frac{y^{-\frac{1}{2}}}{2} f_Z(\sqrt{y}) + \frac{y^{-\frac{1}{2}}}{2} f_Z(-\sqrt{y}) &= \frac{y^{-\frac{1}{2}} e^{-\frac{y}{2}}}{\sqrt{2\pi}} \end{aligned}$$

where $y > 0$.

This can be identified as the gamma distribution with $\alpha = \frac{1}{2}$ and $\beta = 2$.

Its moment generating function is $(1-2t)^{-1/2}$

Due to its importance, it is also called **chi-square distribution** with one degree of freedom ($\chi^2_1$).

Chi-square **distribution** with $n$ degrees of freedom is obtained by adding, independently, $Z_1^2 + Z_2^2 + Z_3^2 + .... + Z_n^2$, where each $Z_i \in \mathcal{N}(0,1)$.

A simple moment-generating-function argument shows that it is, effectively, the gamma distribution with $\alpha = \frac{n}{2}$, $\beta = 2$.

Notation: $\chi^2_n$.

**Probability-Density-Function** $(f)$ **Technique** is a bit faster, but it works for one-to-one transformations only.

It consists of four steps:

(i): Find $X$ in terms of $Y$.

(ii): Substitute the result (switch to small letters) for the argument of $f_X(x)$.

(iii): Differentiate the result of (i) with respect to $y$.

(iv): Multiply (ii) by the absolute value from (iii)

This yields directly the pdf of $Y$.

: 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) $\frac{1}{b} \cdot \frac{1}{1+(\frac{y-a}{b})^2} = \frac{b}{b^2+(y-a)^2}$

   (iv) $\frac{b}{\pi} \cdot \frac{1}{b^2+(y-a)^2}$[check].

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

   Solution:

   (i) $x = y^{1/3}$

   (ii) $6y^{1/3}(1-y^{1/3})$

   (iii) $\frac{1}{3}y^{-2/3}$

   (iv) $2(y^{-1/3}-1)$ [check].

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

   Solution:

   (i) $x = e^{-y}$

   (ii) 1

   (iii) $-e^{-y}$

   (iv) $e^{-y}$ [check].

**BIVARIATE TRANSFORMATION**

$Y$ is now a function of two 'old' RVs, say $X_1$ and $X_2$ whose joint distribution is given.

**Distribution-Function Technique** follows essentially the same pattern as the univariate case, i.e.

$$F_Y(y) = \Pr(Y < y) = \Pr[g(X_1, X_2) < y]$$

Realize that $g(X_1, X_2) < y$ results in a 2-D region, over which $f(x_1, x_2)$ needs to be integrated.

: 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:

$$F_Y(y) = \Pr\left(\frac{X_2}{X_1} < y\right) = \Pr(X_2 < yX_1) = \iint\limits_{0<x_2<yx_1} e^{-x_1-x_2}\, dx_1\, dx_2 = 1 - \frac{1}{1+y}$$

where $y > 0$. This implies that $f_Y(y) = \dfrac{1}{(1+y)^2}$

2) This time $Z_1$ and $Z_2$ are independent RVs from $\mathcal{N}(0,1)$ and

$$Y = Z_1^2 + Z_2^2$$

(we already know the answer: $\chi^2_2$).

Solution:

$$F_Y(y) = \Pr(Z_1^2 + Z_2^2 < y) = \frac{1}{2\pi} \iint\limits_{z_1^2+z_2^2<y} e^{-\frac{z_1^2+z_2^2}{2}}\, dz_1\, dz_2 =$$

$$\frac{1}{2\pi} \int_0^{2\pi} \int_0^{\sqrt{y}} e^{-\frac{r^2}{2}} \cdot r\, dr\, d\theta = \int_0^{\frac{y}{2}} e^{-w}\, dw = 1 - e^{-\frac{y}{2}}$$

where $y > 0$.

This is exponential distribution with $\beta = 2$ (not $\chi^2_2$ as expected, how come?).

3) **Sum of two independent RVs**: Assume that $X_1$ and $X_2$ are independent RVs from the same distribution, having $L$ and $H$ as its lowest and highest possible value. Find the distribution of $X_1 + X_2$.

Solution:

$$F_Y(y) = \Pr(X_1 + X_2 < y) =$$

$$\iint\limits_{\substack{x_1+x_2<y \\ L<x_1,x_2<H}} f(x_1)f(x_2)\,dx_1\;dx_2 \;=$$

$$\begin{cases} \int\limits_{L}^{y-L}\int\limits_{L}^{y-x_1} f(x_1)f(x_2)\,dx_2\,dx_1 & \text{when } y < L+H \\ 1 - \int\limits_{y-H}^{H}\int\limits_{y-x_1}^{H} f(x_1)f(x_2)\,dx_2\,dx_1 & \text{when } y > L+H \end{cases}$$

Differentiating with respect to $y$ results in $f(x) =$

$$\begin{cases} \int\limits_{L}^{y-L} f(x_1)f(y-x_1)\,dx_1 & \text{when } y < L+H \\ \int\limits_{y-H}^{H} f(x_1)f(y-x_1)\,dx_1 & \text{when } y > L+H \end{cases}$$

or, equivalently,

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

where $2L < y < 2H$.

Two special cases of this:

(a) In the specific case of the **uniform** $\mathcal{U}(0,1)$ distribution, the last formula yields:

$$f_Y(y) = \int_{\max(0,y-1)}^{\min(1,y)} dx =$$

$$\begin{cases} \int_0^y dx = y & \text{when} \quad 0 < y < 1 \\ \int_{y-1}^1 dx = 2 - y & \text{when} \quad 1 < y < 2 \end{cases}$$

(b) Similarly, for the 'standardized' **Cauchy** distribution with $f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$, we get:

$$\begin{aligned} f_Y(y) &= \frac{1}{\pi^2} \int_{-\infty}^{\infty} \frac{1}{1+x^2} \cdot \frac{1}{1+(y-x)^2}\, dx = \\ & \frac{2}{\pi} \cdot \frac{1}{4+y^2} \end{aligned}$$

where $-\infty < y < \infty$.

The last result can be easily converted to the pdf of $\bar{X} = \frac{X_1+X_2}{2} = \frac{Y}{2}$, yielding

$$f_{\bar{X}}(\bar{x}) = \frac{1}{\pi} \cdot \frac{1}{1+\bar{x}^2}$$

Thus, the sample mean $\bar{X}$ has the same Cauchy distribution as do the two individual observations (the result can be extended to any sample size)!

**Pdf (Shortcut) Technique** is more powerful, but requires several steps:

(i) Since it can work only for **one-to-one** transformations, the new random variable $Y \equiv g(X_1, X_2)$ must be extended by $Y_2 \equiv X_2$.

(ii) **Invert** the transformation, i.e. solve $y_1 = g(x_1, x_2)$ and $y_2 = x_2$ for $x_1$ and $x_2$ (in terms of $y_1$ and $y_2$).

(iii) **Substitute** this solution into the joint pdf of the 'old' $X_1, X_2$ pair.

(iv) Further multiply by the transformation's **Jacobian** (in absolute value)

$$\left| \begin{array}{cc} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{array} \right|$$

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 (this is often the most difficult part of the procedure).

(v) Eliminate $Y_2$ 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) $X_1$, $X_2 \in \mathcal{E}(1)$ and independent,

$$Y = \frac{X_1}{X_1 + X_2}$$

Solution:

(ii)

$$\begin{array}{rcl} y_1 & = & \dfrac{x_1}{x_1 + x_2} \\ y_2 & = & x_2 \end{array}$$

yields

$$\begin{aligned} x_1 &= \frac{y_1 y_2}{1-y_1} \\ x_2 &= y_2 \end{aligned}$$

(iii) Substitute this into

$$f(x_1, x_2) = \exp(-x_1 - x_2)$$

getting

$$\exp\left[-y_2\left(\frac{y_1}{1-y_1}+1\right)\right] = \exp\left(-\frac{y_2}{1-y_1}\right)$$

(iv) Further multiply by Jacobian

$$\left|\begin{array}{cc} \dfrac{1-y_1+y_1}{(1-y_1)^2}\, y_2 & \dfrac{y_1}{1-y_1} \\ 0 & 1 \end{array}\right| = \frac{y_2}{(1-y_1)^2}$$

to get

$$f(y_1, y_2) = \frac{y_2}{(1-y_1)^2}\, \exp\left(-\frac{y_2}{1-y_1}\right)$$

for $0 < y_1 < 1$ and $y_2 > 0$.

(v) Eliminate $Y_2$ by

$$\frac{1}{(1-y_1)^2}\int_0^\infty y_2\, \exp\left(-\frac{y_2}{1-y_1}\right)\, dy_2 = 1$$

for $0 < y_1 < 1$.

The distribution of $Y$ is thus $\mathcal{U}(0,1)$.

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

Solution:

(ii)

$$\begin{aligned} x_1 &= \frac{y_2}{y_1} \\ x_2 &= y_2 \end{aligned}$$

(iii) Substitute into $\exp(-x_1 - x_2)$ to get

$$\exp[-y_2(1 + \frac{1}{y_1})]$$

(iv) times

$$\left| \begin{array}{cc} -\frac{y_2}{y_1^2} & \frac{1}{y_1} \\ 0 & 1 \end{array} \right| = \frac{y_2}{y_1^2}$$

yields the joint pdf for $y_1 > 0$ and $y_2 > 0$.

(v) Eliminate $y_2$ by

$$\frac{1}{y_1^2} \int_0^\infty y_2 \exp[-y_2(1 + \frac{1}{y_1})]\, dy_2 = \frac{1}{(1+y_1)^2}$$

where $y_1 > 0$ [check].

3) Introducing '**Student**' or **t-distribution**:

Start with two independent RVs $X_1 \in \mathcal{N}(0,1)$ and $X_2 \in \chi^2_n$, define

$$Y_1 = \frac{X_1}{\sqrt{\frac{X_2}{n}}}$$

(ii)

$$\begin{aligned} x_1 &= y_1\sqrt{\frac{y_2}{n}} \\ x_2 &= y_2 \end{aligned}$$

(iii) Substitute into

$$\begin{aligned} f(x_1,x_2) &= \frac{\exp(-\frac{x_1^2}{2})}{\sqrt{2\pi}} \cdot \frac{x_2^{\frac{n}{2}-1}\exp(-\frac{x_2}{2})}{\Gamma(\frac{n}{2})\cdot 2^{\frac{n}{2}}} = \\ & \frac{\exp(-\frac{y_1^2 y_2}{2n})}{\sqrt{2\pi}} \cdot \frac{y_2^{\frac{n}{2}-1}\exp(-\frac{y_2}{2})}{\Gamma(\frac{n}{2})\cdot 2^{\frac{n}{2}}} \end{aligned}$$

(iv) and multiply by

$$\left| \begin{array}{cc} \sqrt{\frac{y_2}{n}} & \frac{y_1}{2\sqrt{n\,y_2}} \\ 0 & 1 \end{array} \right| = \sqrt{\frac{y_2}{n}}$$

(v) Eliminate $y_2$:

$$\frac{1}{\sqrt{2\pi}\Gamma(\frac{n}{2})\,2^{\frac{n}{2}}\sqrt{n}} \int_0^\infty {y_2}^{\frac{n-1}{2}} \exp[-\tfrac{y_2}{2}(1+\tfrac{y_1^2}{n})]\,dy_2 =$$

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

where $-\infty < y_1 < \infty$.

Note that when $n = 1$ this equals $\frac{1}{\pi} \cdot \frac{1}{1+y_1^2}$ (Cauchy), when $n \to \infty$, the second part of the formula tends to $e^{-\frac{y_1^2}{2}}$ (standardized Normal).

When $n \geq 2$, the mean of this distribution is 0. when $n \geq 3$, its variance equals

$$\frac{n}{n-2}$$

4) Introducing **Fisher's** or F-**distribution**

defined by

$$Y_1 = \frac{\frac{X_1}{n}}{\frac{X_2}{m}} \equiv \frac{m}{n} \cdot \frac{X_1}{X_2}$$

where $X_1$ and $X_2$ are independent, both having the chi-square distribution, with degrees of freedom $n$ and $m$, respectively.

(ii)

$$\begin{aligned} x_1 &= \frac{n}{m}\, y_1\, y_2 \\ x_2 &= y_2 \end{aligned}$$

(iii) Substitute into

$$\frac{x_1^{\frac{n}{2}-1} \exp(-\frac{x_1}{2})}{\Gamma(\frac{n}{2})\, 2^{\frac{n}{2}}} \cdot \frac{x_2^{\frac{m}{2}-1} \exp(-\frac{x_2}{2})}{\Gamma(\frac{m}{2})\, 2^{\frac{m}{2}}} =$$

$$\frac{(\frac{n}{m})^{\frac{n}{2}-1}}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})\, 2^{\frac{n+m}{2}}}\; y_1^{\frac{n}{2}-1} \cdot y_2^{\frac{n+m}{2}-2} \exp\left[-\frac{y_2(1+\frac{n}{m}y_1)}{2}\right]$$

(iv) multiply by

$$\begin{vmatrix} \frac{n}{m} \cdot y_2 & .... \\ 0 & 1 \end{vmatrix} = \frac{n}{m} \cdot y_2$$

where $y_1 > 0$ and $y_2 > 0$.

(v) Integrating over $y_2$ (from 0 to $\infty$) yields:

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

where $y_1 > 0$.

## SAMPLING FROM A DISTRIBUTION

A random independent sample (RIS) of size $n$ is a collection of $n$ **independent** RVs $X_1$, $X_2$, ..., $X_n$, each having the same distribution.

A function of these is called a **statistic**. The most important of these is

**SAMPLE MEAN** is defined as the usual average of the $X_i$'s:

$$\overline{X} \equiv \frac{X_1 + X_2 + ... + X_n}{n}$$

Unlike the distribution's mean, this is a random variable, having a distribution of its own.

How does the distribution of $\overline{X}$ relate to the distribution of the individual $X_i$ s?

In terms of the expected value and variance, the answer is simple:

$$\mathbb{E}\left(\overline{X}\right) = \frac{\mathbb{E}\left(X_1 + X_2 + ... + X_n\right)}{n} = \frac{\mu + \mu + ... + \mu}{n} = \mu$$

and

$$\begin{aligned} &\mathrm{Var}\left(\overline{X}\right) = \frac{1}{n^2}\mathrm{Var}\left(X_1\right) + \frac{1}{n^2}\mathrm{Var}\left(X_2\right) + ... + \frac{1}{n^2}\mathrm{Var}\left(X_n\right) \\ &= \frac{\sigma^2 + \sigma^2 + ... + \sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

This implies

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

## CENTRAL LIMIT THEOREM

The **shape** of the $\overline{X}$ distribution is more tricky.

When $n = 1$, we have the original shape.

For $n = 2$, the distribution looks already quite different.

When we reach $n = 10$, a bell-shaped curve is quite apparent.

And, at $n = 30$, the match with the Normal distribution is almost perfect.

**Proof** needs moment generating function:

First, we standardize $\overline{X}$

$$\begin{aligned} Z \equiv \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} &= \frac{\frac{\sum_{i=1}^{n}(X_i - \mu)}{n}}{\frac{\sigma}{\sqrt{n}}} = \\ \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma\sqrt{n}}\right) &\equiv \sum_{i=1}^{n} Y_i \end{aligned}$$

We know that

$$\begin{aligned} M_Y(t) &= 1 + \mathbb{E}[Y]\, t + \mathbb{E}[Y^2]\, \frac{t^2}{2} + \mathbb{E}[Y^3]\, \frac{t^3}{3!} + ... \\ &= 1 + \frac{t^2}{2n} + \frac{\alpha_3 t^3}{6n^{3/2}} + .... \end{aligned}$$

where $\alpha_3$ is the skewness of the original distribution.

The MGF of $Z$ is this $M(t)$, raised to the power of $n$. When $n \to \infty$, we get

$$\lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + \frac{\alpha_3 t^3}{6n^{3/2}} + ....\right)^n = \exp\left(\frac{t^2}{2}\right)$$

which is the MFG of $\mathcal{N}(0, 1)$.

**SAMPLE VARIANCE** is defined by

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

($s$ is then the **sample standard deviation**).

We first expand its numerator

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

then take its expected value:

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

since

$$\begin{aligned}\mathrm{Cov}(\bar{X}, X_1) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}(X_i, X_1) = \\ \frac{1}{n}\mathrm{Cov}(X_1, X_1) &= \frac{1}{n}\mathrm{Var}(X_1) = \frac{\sigma^2}{n}\end{aligned}$$

and the same for $\mathrm{Cov}(\bar{X}, X_2)$, $\mathrm{Cov}(\bar{X}, X_3)$.

We have thus shown that

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

$s^2$ can thus be called an **unbiased estimator** of the distribution's variance $\sigma^2$.

Does this imply that $s$ has the expected value of $\sigma$?

The answer is 'no'.

**SAMPLING FROM $\mathcal{N}(\mu, \sigma)$**

To be able to say anything more about $s^2$, we need to be more specific about the distribution form which the sample is taken.

In this section, we assume that this distribution is Normal.

This immediately implies that the distribution of $\overline{X}$ is Normal for **any** $n$.

Regarding $s^2$, one can show that it is **independent** of $\overline{X}$, and that the distribution of $\dfrac{(n-1)\,s^2}{\sigma^2}$ is $\chi^2_{n-1}$.

The proof of this is fairly complex.

The important implication of all this is that

$$\frac{(\bar{X}-\mu)}{\frac{s}{\sqrt{n}}}$$

has the $\mathsf{t}_{n-1}$ distribution.

**Proof:**

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

**SAMPLING WITHOUT REPLACEMENT** Suppose we select a random sample $X_1$, $X_2$, .... $X_n$, from a **population** of $N$ numbers, say $x_1$, $x_2$, ..., $x_N$ (these don't need to be integers, they may also not be all distinct, and they may be 'dense' in one region and 'sparse' in another - they may thus closely resemble any distribution, including Normal).

Assuming that each number of the population has the same chance of being selected, the mean and variance of the distribution of the $X_i$ s

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

and

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

If the sample is to be independent, the sampling has to be done **with replacement** (meaning that each selected $x_i$ value must be 'returned' to the population before the next draw, to be considered again). In this case, our previous formulas concerning $\overline{X}$ and $s^2$ remain valid.

If the sampling is done **without replacement**, $X_1$, $X_2$, ..., $X_n$ are no longer independent (they are still identically distributed). How does this effect the properties of $\overline{X}$?

The expected value of $\overline{X}$ remains equal to $\mu$, by essentially the same argument as before (the proof does not require independence).

Its variance is now computed by

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

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

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

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

**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).

We already know the individual properties of $\overline{X}$, $\overline{Y}$ (and of $s_x^2$ and $s_y^2$).

Jointly, $\overline{X}$ and $\overline{Y}$ have a distribution which, for $n \to \infty$, tends to be bivariate Normal (proof similar to the univariate case).

As we know, this distribution has five parameters - four of them are the marginal means and standard deviations ($\mu_x$, $\mu_y$, $\frac{\sigma_x}{\sqrt{n}}$ and $\frac{\sigma_y}{\sqrt{n}}$), the last one is the **correlation coefficient** between $\overline{X}$ and $\overline{Y}$. Let's try to derive it.

Clearly

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

This implies that the covariance between $\overline{X}$ and $\overline{X}$ equals $\dfrac{\mathrm{Cov}(X,Y)}{n}$.

Therefore,

$$\rho_{\overline{XY}} = \frac{\frac{\mathrm{Cov}(X,Y)}{n}}{\sqrt{\frac{\sigma_x^2}{n} \cdot \frac{\sigma_y^2}{n}}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y} = \rho_{xy}$$

same as that of a single $(X_i, Y_i)$ pair!

## ORDER STATISTICS

Consider RIS of size $n$ from some distribution.

Define $X_{(1)}$, $X_{(2)}$, ....$X_{(n)}$ to be the smallest, the second smallest, ..., the largest observation, respectively (they will be 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 CASE

It is fairly easy to find $\Pr[X_{(i)} < x]$. This simply means that, out of the original $n$ independent observations, $i$ or more are smaller than $x$. The probability that any one of these $n$ is smaller than $x$ is $p = F(x)$. The answer is thus:

$$\Pr[X_{(i)} < x] = \sum_{j=i}^{n} \binom{n}{j} F(x)^j [1 - F(x)]^{n-j}$$

To get the corresponding **pdf**, we have to differentiate this with respect to $x$. This is a bit tricky, but the answer is:

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

It has the same range as the original distribution.

Using these formulas, we can easily answer any **probability** question, and compute the **mean** and **variance** of this distribution.

## EXAMPLES:

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

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

Solution: Find the probability that any one of the original 7 independent observations is $< 15$ min.:

$$p = \Pr(X_i < 15 \text{ min.}) = 1 - e^{-\frac{15}{23}} = 0.479088$$

Getting 3 or more successes equals

$$1 - \left[ q^7 + 7pq^6 + \binom{7}{2} p^2 q^5 \right] = 73.77\%$$

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

Solution: We know that

$$\begin{aligned} f_{(3)}(x) &= \frac{7!}{2!4!} (1 - e^{-\frac{x}{\beta}})^{3-1} (e^{-\frac{x}{\beta}})^{7-3} \cdot \frac{1}{\beta} e^{-\frac{x}{\beta}} \\ &= \frac{105}{\beta} (1 - e^{-\frac{x}{\beta}})^2 e^{-\frac{5x}{\beta}} \end{aligned}$$

where $\beta = 23$ min.

The corresponding mean is

$$\mu_{(3)} = \frac{105}{23} \int_0^\infty x(1 - e^{-\frac{x}{23}})^2 e^{-\frac{5x}{23}} dx = 11.719 \text{ min.}$$

the standard deviation:

$$\sigma_{(3)} = \sqrt{\frac{105}{23} \int_0^\infty (x - 11.719)^2 (1 - e^{-\frac{x}{23}})^2 e^{-\frac{5x}{23}} dx} = 6.830 \text{ min.}$$

2) Consider a RIS of size 5 form $\mathcal{U}(0, 1)$.

Find $\Pr[X_{(2)} > 0.3]$. This implies that 1 or fewer observations are less than 0.3 .

$p = \Pr(X < 0.3) = 0.3$ .

$$\Pr[X_{(2)} > 0.3] = 0.7^5 + 5 \times 0.7^4 \times 0.3 = 58.82\%$$

Find the mean and standard deviation of $X_{(2)}$.

The corresponding pdf is

$$f_{(2)}(x) = \frac{5!}{1!3!} x(1 - x)^3$$

for $0 < x < 1$.

$$\begin{aligned} \mu_{(2)} &= 20 \int_0^1 x^2 (1 - x)^3 dx = \frac{1}{3} \\ \sigma_{(2)} &= \sqrt{20 \int_0^1 (x - \frac{1}{3})^2 x(1 - x)^3 dx} = 0.1782 \end{aligned}$$

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

For **small samples**, we treat it as one of the order statistics.

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

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

This is true for all distributions.

Proof: The sample median $\tilde{X} \equiv X_{(k+1)}$ has the following pdf:

$$\frac{n!}{k! \cdot k!} F(x)^k [1 - F(x)]^k f(x)$$

We introduce a new RV $Y \equiv (\tilde{X} - \tilde{\mu})\sqrt{n}$ and find its pdf in the usual four steps:

(i) Solve for $x = \tilde{\mu} + \frac{y}{\sqrt{n}}$

(ii) Substitute:

$$\frac{n!}{k!k!} F(\tilde{\mu} + \frac{y}{\sqrt{n}})^k [1 - F(\tilde{\mu} + \frac{y}{\sqrt{n}})]^k f(\tilde{\mu} + \frac{y}{\sqrt{n}})$$

(iii) $\frac{dx}{dy} = \frac{1}{\sqrt{n}}$

(iv)

$$\frac{n!}{k!k!\sqrt{n}} F(\tilde{\mu} + \frac{y}{\sqrt{n}})^k [1 - F(\tilde{\mu} + \frac{y}{\sqrt{n}})]^k f(\tilde{\mu} + \frac{y}{\sqrt{n}})$$

To take the limit of the resulting pdf, we first expand

$$\begin{aligned} F(\tilde{\mu} + \frac{y}{\sqrt{n}}) \;\; &\simeq \;\; F(\tilde{\mu}) + F'(\tilde{\mu})\frac{y}{\sqrt{n}} + \frac{F''(\tilde{\mu})}{2}\frac{y^2}{n} + .... = \\ &\frac{1}{2} + f(\tilde{\mu})\frac{y}{\sqrt{n}} + \frac{f'(\tilde{\mu})}{2}\frac{y^2}{n} + .... \end{aligned}$$

which implies that

$$1 - F(\tilde{\mu} + \frac{y}{\sqrt{n}}) \simeq \frac{1}{2} - \ f(\tilde{\mu})\frac{y}{\sqrt{n}} - \ \frac{f'(\tilde{\mu})}{2}\frac{y^2}{n} + \ ...$$

and

$$F(\tilde{\mu} + \frac{y}{\sqrt{n}}) \left[1 - F(\tilde{\mu} + \frac{y}{\sqrt{n}})\right] \simeq \frac{1}{4} - \ f(\tilde{\mu})^2\frac{y^2}{n} + .... =$$
$$\frac{1}{4}\left(1 - \ \frac{4f(\tilde{\mu})^2 y^2}{n} + ...\right)$$

And finally the limit of:

$$\frac{(2k+1)!}{4^k k! k! \sqrt{2k+1}}\left(1 - \frac{4f(\tilde{\mu})^2 y^2}{2k+1} + ...\right)^k \ f(\tilde{\mu} + \frac{y}{\sqrt{2k+1}})$$
$$= \text{const} \times \exp(-2f(\tilde{\mu})^2 y^2) \times f(\tilde{\mu})$$

How does this compare with the general Normal pdf of

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

Clearly, $\mu = 0$ and $\sigma^2 = \frac{1}{4f(\tilde{\mu})^2}$ or $\sigma = \frac{1}{2f(\tilde{\mu})}$.

So $Y \in \mathcal{N}(0, \frac{1}{2f(\tilde{\mu})})$, which implies that $\widetilde{X} = \tilde{\mu} + \frac{Y}{\sqrt{n}}$ is Normal with the mean of $\tilde{\mu}$ and standard deviation of $\frac{1}{2\sqrt{n}f(\tilde{\mu})}$.

## EXAMPLES:

1) Consider a RIS of size 1001 from 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 $\Pr(-0.1 < \tilde{X} < 0.1) =$

$$\frac{1}{0.049648 \times \sqrt{2\pi}} \int_{-0.1}^{0.1} \exp\left(-\frac{x^2}{2 \times 0.049648^2}\right) dx$$
$$= \ 95.60\%$$

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 $n$).

2) Sampling from $\mathcal{N}(\mu, \sigma)$, is it better to estimate $\mu$ by the sample mean or by the sample median?

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 $\bar{X}$. To estimate $\mu$ to the same accuracy as $\bar{X}$ does, $\tilde{X}$ would have to use $\frac{\pi}{2} = 1.57$ times bigger sample.

3) Consider a RIS of size 349 from a distribution with $f(x) = 2x$ $(0 < x < 1)$. Find $\Pr(\tilde{X} < 0.7)$.

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 $\sqrt{2}$, which means $\tilde{\sigma} = \frac{1}{2\sqrt{2} \times \sqrt{349}} = 0.018925$, and $\Pr(\tilde{X} < 0.75) =$

$$\begin{aligned} & \frac{1}{0.018925 \times \sqrt{2\pi}} \int_{-\infty}^{0.7} \exp\left(-\frac{(x - \frac{1}{\sqrt{2}})^2}{2 \times 0.018925^2}\right) dx \\ = \; & 35.36\% \end{aligned}$$

Subsidiary: Find $\Pr(\bar{X} < 0.7)$.

Solution: The distribution being sampled has $\mu = \int_0^1 2x \cdot x \, dx = \frac{2}{3}$ and $\sigma^2 = \int_0^1 2x \cdot x^{\,2} dx - (\frac{2}{3})^2 = \frac{1}{18}$.

We know that $\bar{X} \approx \mathcal{N}(\frac{2}{3}, \frac{1}{\sqrt{18} \cdot \sqrt{349}} = 0.0126168)$, therefore $\Pr(\bar{X} < 0.75) =$

$$\begin{aligned} & \frac{1}{0.0126168 \times \sqrt{2\pi}} \int_{-\infty}^{0.7} \exp\left(-\frac{(x - \frac{2}{3})^2}{2 \times 0.0126168^2}\right) dx \\ = \; & 99.59\% \end{aligned}$$

**BIVARIATE CASE**

**Joint pdf** of two order statistics $X_{(i)}$ and $X_{(j)}$ $(i < j)$ is

$$f(x_i, x_j) = \lim_{\substack{\Delta \to 0 \\ \varepsilon \to 0}} \frac{\Pr[\,(x_i \le X_{(i)} < x_i + \Delta) \cap (x_j \le X_{(j)} < x_j + \varepsilon)\,]}{\Delta \cdot \varepsilon}$$

i.e.

| Interval | # of observation |
|---|---|
| $L \leftrightarrow x_i$ | $i-1$ |
| $x_i \leftrightarrow x_i + \Delta$ | $1$ |
| $x_i + \Delta \leftrightarrow x_j$ | $j-i-1$ |
| $x_j \leftrightarrow x_j + \varepsilon$ | $1$ |
| $x_j + \varepsilon \leftrightarrow H$ | $n-i-j$ |

By the multinomial formula, the probability equals

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(x_i)^{i-1} \times [F(x_i + \Delta) - F(x_i)][F(x_j) - F(x_i + \Delta)]^{j-i-1} \times [F(x_j + \varepsilon) - F(x_j)][1 - F(x_j + \varepsilon)]^{n-j}$$

Dividing by $\Delta \cdot \varepsilon$ and taking the two limits yields

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(x_i)^{i-1} f(x_i) \times [F(x_j) - F(x_i)]^{j-i-1} f(x_j)[1 - F(x_j)]^{n-j}$$

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

Two important special cases of this formula:

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

$$f(x_i, x_{i+1}) = \frac{n!}{(i-1)!(n-i-1)!} \times F(x_i)^{i-1} [1 - F(x_{i+1})]^{n-i-1} f(x_i) f(x_{i+1})$$

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

$$f(x_1, x_n) = n(n-1) [F(x_n) - F(x_1)]^{n-2} f(x_1) f(x_n)$$

EXAMPLES:

1) Assuming we sample from $\mathcal{E}(1)$ and $n = 9$, find $\mathrm{Cov}(X_{(3)}, X_{(5)})$ and $\Pr(X_{(3)} > \frac{1}{2} \cap X_{(5)} < 2)$.

The corresponding bivariate pdf. is $f(x_3, x_5) =$

$$\frac{9!}{2 \times 4!}(1 - e^{-x_3})^2(e^{-x_3} - e^{-x_5})e^{-4x_5}e^{-x_3}e^{-x_5} \quad x_3 < x_5$$

$$\mu_{(3)} = 7560 \times$$

$$\int_0^\infty \int_0^{x_5} x_3 \cdot (1 - e^{-x_3})^2(e^{-x_3} - e^{-x_5})e^{-5x_5}e^{-x_3} dx_3 dx_5$$

$$= 0.37897$$

$$\mu_{(5)} = 7560 \times$$

$$\int_0^\infty \int_0^{x_5} x_5 \cdot (1 - e^{-x_3})^2(e^{-x_3} - e^{-x_5})e^{-5x_5}e^{-x_3} dx_3 dx_5$$

$$= 0.74563$$

$$\mathbb{E}[X_{(3)}X_{(5)}] = 7560 \times$$

$$\int_0^\infty \int_0^{x_5} x_3 x_5 \cdot (1 - e^{-x_3})^2(e^{-x_3} - e^{-x_5})e^{-5x_5}e^{-x_3} dx_3 dx_5$$

$$= 0.33095$$

which imply that

$$\mathrm{Cov}(X_{(3)}, X_{(5)}) = 0.33095 - 0.37897 \times 0.74563 = 0.04838$$

and, finally

$$\Pr(X_{(3)} > \frac{1}{2} \cap X_{(5)} < 2) = 7560\times$$

$$\int_{0.5}^{2}\int_{0.5}^{x_5}(1-e^{-x_3})^2(e^{-x_3}-e^{-x_5})e^{-5x_5}e^{-x_3}dx_3dx_5$$

$$= 24.15\%$$

2) Assuming that we sample from $\mathcal{U}(0,1)$, find the distribution of $Y = \frac{X_{(1)}+X_{(n)}}{2}$, the mid-range value.

This means that

$$f(x_1, x_n) = n(n-1)(x_n - x_1)^{n-2}$$

Solution: $Y_2 = X_{(n)}$. This means that $0 < y_2 < 1$ and $\frac{y_2}{2} < y_1 < y_2$.

(ii) $x_1 = 2y_1 - y_2$ and $x_n = y_2$.

(iii) $n(n-1)(2y_2 - 2y_1)^{n-2}$

(iv) $\begin{vmatrix} 2 & -1 \\ 0 & 1 \end{vmatrix} = 2$

(v)

$$f(y_1) = 2^{n-1}n(n-1)\int_{y_1}^{\min(2y_2,1)}(y_2-y_1)^{n-2}\,dy_2 =$$

$$2^{n-1}\,n\,(y_2-y_1)^{n-1}\Big|_{y_2=y_1}^{\min(2y_1,1)} =$$

$$2^{n-1}n \times \begin{cases} y_1^{n-1} & 0 < y_1 < \frac{1}{2} \\ (1-y_1)^{n-1} & \frac{1}{2} < y_1 < 1 \end{cases}$$

This implies that

$$\begin{aligned} \mathbb{E}(Y_1) &= \int_0^1 y_1 f(y_1)dy_1 = \frac{1}{2} \\ \mathrm{Var}(Y_1) &= \int_0^1 (y_1 - \tfrac{1}{2})^2 f(y_1)dy_1 = \frac{1}{2(n+2)(n+1)} \end{aligned}$$

These can be easily extended to the case of a general uniform distribution $\mathcal{U}(a, b)$:

$$\mathbb{E}\left[\frac{X_{(1)}+X_{(n)}}{2}\right] = \frac{a+b}{2}, \text{Var}\left[\frac{X_{(1)}+X_{(n)}}{2}\right] = \frac{(b-a)^2}{2(n+2)(n+1)}$$