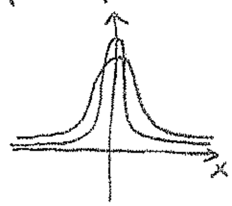
$\boxed{1}$ Review of probability distributions

Building block: $X \sim N(\mu, \sigma^2)$ $\boxed{\text{normal distribution}}$

$X \in (-\infty, \infty)$ $\quad f_X(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-(x-\mu)^2/2\sigma^2}$ density (p.d.f.)



"Bell curve" for $\mu = 0$

As $\sigma \uparrow$, hump $\downarrow$

Z - standard normal distribution $Z \sim N(0,1)$ $\quad \boxed{Z = \frac{X-\mu}{\sigma} \iff X = \mu + \sigma Z}$

$E[Z] = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0$ (since integrand is an odd function)

explanation

$$\boxed{\begin{aligned} \mathbb{P}(Z \in [a,b]) &= \mathbb{P}\left(\frac{X-\mu}{\sigma} \in [a,b]\right) = \mathbb{P}\left(X \in [a\sigma+\mu, b\sigma+\mu]\right) = \\ &= \int_{a\sigma+\mu}^{b\sigma+\mu} \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \underset{\substack{z = \frac{x-\mu}{\sigma} \\ \text{change of variable}}}{=} \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \end{aligned}}$$

$E[Z^2] = 1$. Why? Start with $1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$, integrate by parts:

$u = e^{-\frac{z^2}{2}}$ $\quad du = -z e^{-\frac{z^2}{2}} dz$

$dv = dz$ $\quad v = z$

$$\Rightarrow 1 = \frac{1}{\sqrt{2\pi}} z e^{-\frac{z^2}{2}} \Big|_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z(-z) e^{-\frac{z^2}{2}} dz =$$

$$= 0 + \int_{-\infty}^{\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = E[Z^2]$$

So, $Var(Z) = E[Z^2] - (E[Z])^2 = 1 - 0 = 1$

Hence, $E[X] = \mu + \sigma E[Z] = \mu$ (using linearity of expectation)

$E[X^2] = E[\mu^2 + 2\mu\sigma Z + \sigma^2 Z^2] = \mu^2 + \sigma^2$

$Var(X) = E[X^2] - (E[X])^2 = \sigma^2$

FACT 1 If $X_1, X_2, \dots, X_n$ are independent random variables, $X_i \sim N(\mu_i, \sigma_i^2)$, $i=1,2,\dots,n$, then $\sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2)$

FACT 2 (Law of large numbers)

If $X_1, X_2, \dots, X_n$ is an i.i.d. sample with $E[X_1] < \infty$, then the sample average $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow E[X_1]$, i.e. converges to common expectation in probability, i.e.

$$\forall \varepsilon > 0 \quad \mathbb{P}(|\overline{X_n} - E[X_1]| > \varepsilon) \longrightarrow 0 \text{ as } n \to \infty.$$

FACT 3 (Central limit theorem)

If $X_1, X_2, \dots, X_n$ is an i.i.d. sample from some continuous distribution such that $E[X_1] < \infty$ and $\sigma^2 = Var(X_1) < \infty$, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - E[X_i]) = \sqrt{n}\,(\overline{X_n} - E[X_1]) \xrightarrow{d} N(0,\sigma^2), \text{ i.e.}$$

converges in distribution to a $N(0,\sigma^2)$, where convergence in distribution means that for every interval $[a,b]$:

$$\mathbb{P}(\sqrt{n}\,(\overline{X_n} - E[X_1]) \in (a,b)) \longrightarrow \int_a^b \frac{1}{\sqrt{2\pi}\cdot\sigma}\, e^{-\frac{x^2}{2\sigma^2}}\,dx$$

$\boxed{\text{Gamma distribution}}$

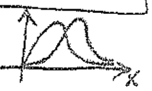
has two parameters $\underset{\text{shape}}{\alpha > 0}$, $\underset{\text{scale}}{\beta > 0}$ Before we define it, let's recall gamma function $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\,dx$

Divide both sides by $\Gamma(\alpha)$ to get: $1 = \int_0^\infty \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}\,dx$, i.e

(after substitution $x = \beta y$)

$$1 = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}\,dy$$

Define $\boxed{f(x\mid\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \text{ if } x \geq 0 \text{; and } 0 \text{ otherwise}}$

This is the p.d.f. (since it's nonnegative and it integrates to 1). 

Let $X \sim \Gamma(\alpha,\beta)$, i.e. X is a random variable with p.d.f. $f_X(x) = f(x\mid\alpha,\beta$

Properties of gamma function $\Gamma(\alpha)$

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\,dx = x^{\alpha-1}(-e^{-x})\Big|_0^{\infty} - \int_0^{\infty} (-e^{-x})(\alpha-1)x^{\alpha-2}\,dx = (\alpha-1)\int_0^{\infty} x^{(\alpha-1)-1} e^{-x}\,dx$$

$$u = x^{\alpha-1} \qquad dv = e^{-x}dx$$
$$du = (\alpha-1)x^{\alpha-2}dx \qquad v = -e^{-x}$$

i.e. $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$. Since $\Gamma(1) = \int_0^{\infty} e^{-x}dx = 1$, iterating the last identity, we get $\Gamma(n) = (n-1)!$

k^{th}-moment of Gamma distribution:

$$E[X^k] = \int_0^{\infty} x^k \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\,dx = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} x^{(\alpha+k)-1} e^{-\beta x}\,dx =$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} \int_0^{\infty} \frac{\beta^{\alpha+k}}{\Gamma(\alpha+k)} x^{(\alpha+k)-1} e^{-\beta x}\,dx = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} =$$

p.d.f. of $\Gamma(\alpha+k,\beta)$ So, it integrates to 1

$$= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \cdot \frac{1}{\beta^k} = \frac{(\alpha+k-1)(\alpha+k-2)\cdots\alpha}{\beta^k}$$

(using the property of Gamma funct

So, $\boxed{E[X] = \frac{\alpha}{\beta},\ E[X^2] = \frac{(\alpha+1)\alpha}{\beta^2},\ Var(X) = \frac{\alpha}{\beta^2} \text{ for } X \sim \Gamma(\alpha,\beta)}$

FACT 4. If $X_i \sim \Gamma(\alpha_i,\beta)$, $i = 1,2,\ldots,n$, are independent random variables, then $\sum_{i=1}^{n} X_i \sim \Gamma\left(\sum_{i=1}^{n}\alpha_i, \beta\right)$.

proof of FACT 4.: First, we find a moment generating function (m.g.f.) of $X \sim \Gamma(\alpha, \beta)$:

$$E[e^{tX}] = \int_0^\infty e^{tx} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x} dx =$$

$$= \frac{\beta^\alpha}{(\beta-t)^\alpha} \int_0^\infty \underbrace{\boxed{\frac{(\beta-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x}}}_{\substack{\text{p.d.f. of } \Gamma(\alpha, \beta-t) \\ \text{so it integrates to } 1}} dx = \left(\frac{\beta}{\beta-t}\right)^\alpha$$

The m.g.f. of $\sum_{i=1}^n X_i$, $X_i \sim \Gamma(\alpha_i, \beta)$ is:

$$E\left[e^{t\sum_{i=1}^n X_i}\right] = E\left[\prod_{i=1}^n e^{tX_i}\right] \underset{\uparrow}{=} \prod_{i=1}^n E[e^{tX_i}] = \prod_{i=1}^n \left(\frac{\beta}{\beta-t}\right)^{\alpha_i} = \left(\frac{\beta}{\beta-t}\right)^{\sum_{i=1}^n \alpha_i}$$

independence of X_i

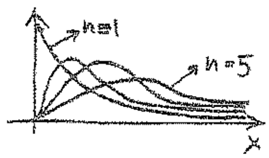
which is again the m.g.f. of a Gamma distribution $\Gamma\left(\sum_{i=1}^n \alpha_i, \beta\right)$ $\square$

$\boxed{\chi^2_n\text{-distribution}}$

(n degrees of freedom) is the distribution of $\sum_{i=1}^n X_i^2$, where $X_i \sim N(0,1)$, $i=1,\dots,n$
X_i independent

Relationship with Gamma

$\chi^2_n \equiv \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$



As $n\uparrow$, flatter w/ hump

proof: Let $X \sim N(0,1)$. Then the cumulative distribution function (c.d.f) of X^2 is

$$\mathbb{P}(X^2 \le x) = \mathbb{P}(-\sqrt{x} \le X \le \sqrt{x}) = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \text{ ; so, the p.d.f. is}$$

$$f_{X^2}(x) = \frac{d}{dx}\left(\mathbb{P}(X^2 \le x)\right) = \frac{d}{dx} \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\sqrt{x})^2}{2}} \frac{d}{dx}(\sqrt{x}) - \frac{1}{\sqrt{2\pi}} e^{-\frac{(-\sqrt{x})^2}{2}} \frac{d}{dx}(-\sqrt{x})$$

i.e. $f_{X^2}(x) = \frac{1}{\sqrt{2\pi}} x^{\frac{1}{2}-1} e^{-\frac{x}{2}}$ (after some algebraic manipulation)

Since $\Gamma(1/2) = \sqrt{\pi}$ (why?), $X^2 \sim \Gamma(1/2, 1/2)$.

Now, using FACT 4., $\sum_{i=1}^n X_i^2 \sim \Gamma(n/2, 1/2)$ where $X_i \sim N(0,1)$ i.i.d. $i=1,\dots,n$

So, $\chi^2_n \equiv \Gamma(n/2, 1/2)$

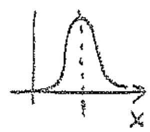
Fisher F-distribution

Let $X \sim \chi^2_k \equiv \Gamma(k/2, 1/2)$, $Y \sim \chi^2_m \equiv \Gamma(m/2, 1/2)$; X, Y independent rand. var's.

Let $Z \equiv \dfrac{X/k}{Y/m}$. Then, Z is said to have a Fisher distribution with degrees of freedom k and m, and is denoted $Z \sim F_{k,m}$.

<u>Observation 1</u>. Since $X \sim \chi^2_k$, then $X \equiv \sum_{i=1}^{k} X_i^2$, $X_i \sim N(0,1)$ i.i.d. By Law of large numbers (FACT 2): $\frac{1}{k}\sum_{i=1}^{k} X_i^2 \longrightarrow E[X_1^2] = 1$, as $k \to \infty$.

So, as $k, m \to \infty$, $X/k \to 1$, $Y/m \to 1$ $\Rightarrow$ Z will concentrate around 1.

p.d.f. of Z is 

As $k, m \to \infty$, hump $\uparrow$ and $\rightarrow\leftarrow$ (narrower).

<u>Observation 2</u> $F_{k,m}(c, \infty) = F_{m,k}(0, \frac{1}{c})$. Why? $F_{k,m}(c,\infty) = \mathbb{P}\left(\frac{X/k}{Y/m} \geq c\right) = \mathbb{P}\left(\frac{Y/m}{X/k} \leq \frac{1}{c}\right) = F_{m,k}(0,$

· What is the p.d.f. of $Z \sim F_{k,m}$? First, we compute the p.d.f. of $\frac{X}{Y} = \frac{k}{m} Z$.

$$\left.\begin{aligned} f_X(x) &= \frac{(1/2)^{k/2}}{\Gamma(k/2)}\, x^{k/2-1} e^{-x/2}, \quad x \geq 0 \\ f_Y(y) &= \frac{(1/2)^{m/2}}{\Gamma(m/2)}\, y^{m/2-1} e^{-y/2}, \quad y \geq 0 \end{aligned}\right\} \text{ the p.d.f.'s of } X, Y, \text{ respectively.}$$

To find the p.d.f. of X/Y, first write the c.d.f. Since $X, Y > 0$, $X/Y > 0$, so for $t \geq 0$:

$$\mathbb{P}(X/Y \leq t) = \mathbb{P}(X \leq tY) = \int_0^\infty \int_0^{ty} f_{X,Y}(x,y)\,dx\,dy, \text{ where}$$

$f_{X,Y}(x,y)$ is the joint density of X and Y.

<u>But</u>, X, Y are independent, so $f_{X,Y}(x,y) = f_X(x) f_Y(y)$. Hence,

$$\mathbb{P}(X/Y \leq t) = \int_0^\infty \int_0^{ty} f_X(x) f_Y(y)\,dx\,dy, \text{ and}$$

$$f_{X/Y}(t) = \frac{d}{dt}\mathbb{P}(X/Y \leq t) = \frac{d}{dt}\int_0^\infty \int_0^{ty} f_X(x) f_Y(y)\,dx\,dy = \int_0^\infty f_X(ty)\cdot f_Y(y)\, y\,dy =$$

$$= \int_0^\infty \frac{(1/2)^{k/2}}{\Gamma(k/2)} (ty)^{k/2-1} e^{-(ty)/2} \cdot \frac{(1/2)^{m/2}}{\Gamma(m/2)} y^{m/2-1} e^{-y/2}\, y\,dy =$$

$$= \frac{(1/2)^{(k+m)/2}}{\Gamma(k/2)\Gamma(m/2)}\, t^{k/2-1} \int_0^\infty y^{(k+m)/2-1} e^{-((t+1)y)/2}\,dy =$$

$$= \frac{(1/2)^{(k+m)/2}}{\Gamma(k/2)\Gamma(m/2)} \cdot t^{k/2-1} \cdot \frac{\Gamma\left(\frac{k+m}{2}\right)}{\left(\frac{t+1}{2}\right)^{(k+m)/2}} \int_0^\infty \underbrace{\boxed{\frac{\left(\frac{t+1}{2}\right)^{(k+m)/2}}{\Gamma\left(\frac{k+m}{2}\right)} \cdot y^{(k+m)/2-1} e^{-\left(\frac{t+1}{2}\right)y}}}_{\text{p.d.f. of } \Gamma\left(\frac{k+m}{2}, \frac{t+1}{2}\right) \text{ So integral } = 1} dy$$

Hence, $$f_{X/Y}(t) = \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} \cdot t^{\frac{k}{2}-1} \cdot (1+t)^{-\frac{k+m}{2}}$$

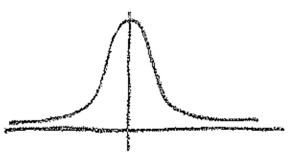
Since $\mathbb{P}(Z \le t) = \mathbb{P}\left(\frac{X}{Y} \le \frac{kt}{m}\right) \Rightarrow f_Z(t) = \frac{d}{dt}\mathbb{P}(Z \le t) = f_{X/Y}\left(\frac{kt}{m}\right) \cdot \frac{k}{m}$

Finally, the p.d.f. of $Z \sim F_{k,m}$ is:

$$f_Z(t) = \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma(k/2)\Gamma(m/2)} \cdot k^{\frac{k}{2}} m^{\frac{m}{2}} t^{\frac{k}{2}-1} (m+kt)^{-\frac{k+m}{2}} \equiv f_{k,m}(t)$$

Student t_n-distribution

distribution of a random variable $T \equiv \dfrac{X_1}{\sqrt{\left(\sum_{i=1}^n Y_i^2\right)/n}}$, where $X_1, Y_1, \ldots, Y_n$ are all i.i.d. $N(0,1)$



As $n \uparrow$, hump goes up, and it approaches the standard normal distribution.

What is the p.d.f. of T? $\mathbb{P}(-t \le T \le t) = \mathbb{P}(T^2 \le t^2) = \mathbb{P}\left(\dfrac{X_1^2}{\frac{1}{n}\sum_{i=1}^n Y_i^2} \le t^2\right)$

$$\int_{-t}^{t} f_T(x)\,dx = \int_0^{t^2} f_{1,n}(x)\,dx$$

Since $\dfrac{X_1^2}{\frac{1}{n}\sum_{i=1}^n Y_i^2} \sim F_{1,n}$

Taking $\frac{d}{dt}$ of both sides: $f_T(t) + f_T(-t) = f_{1,n}(t^2) \cdot 2t$

t_n-distribution is symmetric (because the numerator has symmetric distribution $N(0,1)$); hence,

$f_T(t) = f_T(-t)$ and thus

$$f_T(t) = f_{1,n}(t^2) \cdot t, \text{ i.e. } f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(1/2)\Gamma(n/2)} \frac{1}{\sqrt{n}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

Sample mean and sample variance

Given a sample $y_1, y_2, \ldots, y_n$ (independent y_i's) from an (unknown) distribution,
the sample mean is : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

the sample variance is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Theorem If $y_1, y_2, \ldots, y_n$ are independent sample from $N(\mu, \sigma^2)$,

then $\bar{y} \sim N(\mu, \frac{\sigma^2}{n})$

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$$ ← Comment: "Loss" of one degree of freedom is due to the linear constraint $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$ in estimating μ by $\bar{y}$.

$\bar{y}$ and s^2 are independent

proof: This will be a special case of a much more general result on multiple linear regression (next time). It is also proved in Section 3 (Theorem 1)

Multivariate distributions

If $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$ is a $k\times 1$ random vector, then its expectation is $E[X] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_k] \end{pmatrix}$

and its covariance matrix is $\mathrm{Cov}(X) = E\left[(X-E[X])(X-E[X])'\right]$ ($k\times k$ matrix)

Covariance matrix is always symmetric, i.e. $(\mathrm{Cov}(X))' = \mathrm{Cov}(X)$ and nonnegative definite, i.e. for any $k\times 1$ non-random vector a, we have:

$$a'\mathrm{Cov}(X)\,a = E\left[a'(X-E[X])(X-E[X])'a\right] \underset{\text{using } \|v\|^2 = v'v}{=} E\left[\|a'(X-E[X])\|^2\right] \geq 0.$$

How does the covariance of X change once we multiply X by a non-random $n\times k$ matrix A?
Let $Y = AX$ (Y is $n\times 1$ vector). The covariance of Y will be an $n\times n$ matrix:

$$\mathrm{Cov}(Y) = E\left[(Y-E[Y])(Y-E[Y])'\right] = E\left[(AX-E[AX])(AX-E[AX])'\right] \underset{A \text{ non-random}}{=}$$

$$= E\left[A(X-E[X])(A(X-E[X]))'\right] \underset{(CD)'=D'C'}{=} E\left[A(X-E[X])(X-E[X])'A'\right] =$$

$$= A\,E\left[(X-E[X])(X-E[X])'\right]A' = A\,\mathrm{Cov}(X)A'$$

Hence, $\boxed{\mathrm{Cov}(AX) = A\,\mathrm{Cov}(X)A'}$

k-variate normal distribution

A $k\times 1$ random vector $Z = (Z_1, \ldots, Z_k)'$ is said to have the k-variate standard normal distrib[ution] if $Z_1, Z_2, \ldots, Z_k$ are independent $N(0,1)$. The density of Z is given by:

$$f_Z(z) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} = (2\pi)^{-k/2} e^{-\|z\|^2/2}, \quad z = (z_1, \ldots, z_k) \in \mathbb{R}^k$$

The distribution of a $n\times 1$ vector AZ, where A is a non-random $n\times k$ matrix is called a k-variate normal distribution with mean $0_{n\times 1}$ and covariance
$\Sigma = \mathrm{cov}(AZ) = A\,\mathrm{cov}(Z)A' = AA'$ (since $\mathrm{Cov}(Z) = I_{k\times k}$)
and is denoted simply by $N(0, \Sigma)$.

Note: 0 here denotes an $n\times 1$ vector of zeroes
Σ is an $n\times n$ matrix

One can also shift this distribution by an $n\times 1$ vector μ. Letting $Y = \mu + AZ$, Y is said to have a k-variate normal distribution $N(\mu, \Sigma)$. (here, again $\Sigma = AA'$)

Notice that in the definition and final notation for, say, $N(0,\Sigma)$ we assumed that the distribution depends only on a covariance matrix Σ and does not depend on the construction, i.e. does not depend on the choice of Z and A. We could have started with an m-variate standard normal vector $\tilde{Z}$ and a non-random $n\times m$ matrix B so that the covariance matrix of $B\tilde{Z}$ again happens to be equal to Σ, i.e. so that $\mathrm{Cov}(B\tilde{Z}) = BB' = \Sigma \; (=AA')$.

Both constructions give the same multivariate normal distribution $N(0,\Sigma)$ according to our definition. Why are the distributions of AZ and $B\tilde{Z}$ even the same? We show the proof here in the case when A and B are both $n\times n$ invertible matrices (and $Z, \tilde{Z}$ are n-variate standard normal vectors); the proof in general is a bit more complicated.

Let's calculate the p.d.f. of AZ. For every set $\Omega \subseteq \mathbb{R}^n$, we can write:

$$\mathbb{P}(AZ\in\Omega) = \mathbb{P}(Z\in A^{-1}\Omega) = \int_{A^{-1}\Omega} (2\pi)^{-n/2} e^{-\frac{1}{2}\|z\|^2}\,dz \underset{\substack{y=Az\\ z=A^{-1}y\\ \text{change of variables}}}{=} \int_{\Omega} (2\pi)^{-n/2}\frac{1}{|\det(A)|}\cdot e^{-\frac{1}{2}\|A^{-1}y\|^2}\,dy$$

Now, $\det(\Sigma) = \det(AA') = \det(A)\det(A') = (\det(A))^2$

$$\|A^{-1}y\|^2 = (A^{-1}y)'(A^{-1}y) = y'(A^{-1})'(A^{-1}y) \underset{\substack{(A^{-1})'=(A')^{-1}\\ \text{when } A \text{ invertible}}}{=} y'(A')^{-1}A^{-1}y \underset{(CD)^{-1}=D^{-1}C^{-1}}{=} y'(AA')^{-1}y = y'\Sigma^{-1}y$$

So, $$\mathbb{P}(AZ\in\Omega) = \int_{\Omega} (2\pi)^{-n/2}\frac{1}{\sqrt{\det(\Sigma)}}\, e^{-\frac{1}{2}y'\Sigma^{-1}y}\,dy$$

So, random vector AZ has the density $(2\pi)^{-n/2}\cdot\dfrac{1}{\sqrt{\det(\Sigma)}}\, e^{-\frac{1}{2}y'\Sigma^{-1}y}$, which depends only on Σ, and not on A! Hence, AZ and $B\tilde{Z}$ must have the same density/distribution, and our definition of multivariate normal distribution is valid, since it depends only on Σ not on particular choice of A (and Z).

One nice consequence of this discussion is the density function of a k-variate normal distribut

Let $Y\sim N(\mu,\Sigma)$. Then,

$$\boxed{f_Y(y) = (2\pi)^{-k/2}\cdot\frac{1}{\sqrt{\det(\Sigma)}}\, e^{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)}, \quad y\in\mathbb{R}^k.}$$

For $k=2$, this can be written as:

$$f_Y(y_1,y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot e^{-\frac{\frac{(y_1-\mu_1)^2}{\sigma_1^2} - 2\rho\cdot\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}}{2(1-\rho^2)}}$$

known as the bivariate normal density function.

$\sigma_i^2 = \mathrm{Var}(Y_i)$

(ρ correlation coeff. between Y_1 and Y_2)

Question : Given a symmetric non-negative definite $n\times n$ matrix Σ, how does one find a matrix A such that $\Sigma = AA'$? One can use, for example, the eigenvalue decomposition $\Sigma = QDQ'$, where Q is orthogonal, D is diagonal (with eigenvalues $\lambda_1, \dots, \lambda_n$ of Σ on its diagonal). If $D^{1/2}$ denotes the diagonal matrix with $\sqrt{\lambda_i}$ on the diagonal, one can take $A = QD^{1/2}$ or $A = QD^{1/2}Q'$ (so that $AA' = \Sigma$).

FACT 5. Let $Y \sim N(0_{k\times 1}, \Sigma_{k\times k})$. Let M be an $m\times k$ non-random matrix. Then $MY \sim N(0_{m\times 1}, M\Sigma M')$. } "linear transfor of normal is again normal"

proof: $Y = AZ$ for some $k\times k$ matrix A such that $\Sigma = AA'$ and a k-variate standard normal Z. Then $MY = M(AZ) = (MA)Z$ is, by definition, m-variate normal with mean $0_{m\times 1}$ and $\mathrm{cov}(MY) = (MA)(MA)' = MAA'M' = M\Sigma M'$.

FACT 6. Let $Z \sim N(0_{k\times 1}, I_{k\times k})$ and let Q be an orthogonal $k\times k$ matrix. Then $QZ \sim N(0_{k\times 1}, I_{k\times k})$. } "orthogonal transform of a standard normal is again standard normal"

proof: Recall that a $k\times k$ matrix Q is orthogonal when one of the following properties hold

A) $Q^{-1} = Q'$ (and hence $|\det(Q)| = 1$)

B) rows/columns of Q form an orthonormal basis in $\mathbb{R}^k$

C) for any $x \in \mathbb{R}^k$ we have $\|Qx\| = \|x\|$, i.e. Q preserves the length of vectors

Geometrically, orthogonal transformations represent linear transformations that preserve distance between points, such as rotations and reflections.

$$\forall \Omega \subseteq \mathbb{R}^k : \mathbb{P}(QZ \in \Omega) = \mathbb{P}(Z \in Q^{-1}\Omega) = \int_{Q^{-1}\Omega} f_Z(z)\,dz \underset{\uparrow}{=} \int_{\Omega} \frac{f_Z(Q^{-1}x)}{|\det(Q)|}\,dx$$

Change of var: $x = Qz$, $z = Q^{-1}x$

Since $|\det(Q)| = 1$ and $\|Q^{-1}x\| = \|x\|$, we get:

$$f_Z(Q^{-1}x) = (2\pi)^{-k/2} e^{-\|Q^{-1}x\|^2/2} = (2\pi)^{-k/2} e^{-\|x\|^2/2} = f_Z(x)$$

Hence, $\mathbb{P}(QZ \in \Omega) = \int_{\Omega} f_Z(x)\,dx = \mathbb{P}(Z \in \Omega) \quad \forall \Omega \subseteq \mathbb{R}^k$

So, $QZ \sim N(0_{k\times 1}, I_{k\times k})$.

FACT 7 Uncorrelated components of a multivariate normal vector are independent.

FACT 8 Multivariate CLT (central limit theorem)

Suppose $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$ is a random $k\times 1$ vector with covariance Σ (and $E[X_i^2] < \infty$)

Let $Y_1, Y_2, \dots, Y_n$ be a sequence of i.i.d. copies of X. Then

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - E[Y_i]) \xrightarrow{d} N(0_{k\times 1}, \Sigma) \quad \text{as } n \to \infty$$

where the convergence $\xrightarrow{d}$ in distribution means that for any set $\Omega \subseteq \mathbb{R}^k$

$$\lim_{n\to\infty} \mathbb{P}(S_n \in \Omega) = \mathbb{P}(Z \in \Omega) \text{ for a random vector } Z \sim N(0, \Sigma)$$

Sample mean and covariance matrix from a multivariate normal distribution

Let $Y_1, \dots, Y_n$ be independent $m\times 1$ random $N(\mu, \Sigma)$ vectors with $n > m$ and positive definite Σ. Define

the sample mean vector $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

the sample covariance matrix $\frac{W}{n-1}$, where $W = \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})'$

Generalizing the corresponding results on page 7 in the case $m=1$, the following facts are known

FACT 9 $\bar{Y} \sim N(\mu, \frac{\Sigma}{n})$

$\bar{Y}$ and $\frac{W}{n-1}$ are independent

Question: How do we generalize $\sum_{i=1}^{n} (y_i - \bar{y})^2/\sigma^2 = (n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$ to the multivariate case

We need to generalize the χ^2-distribution to the multivariate case.

Wishart distribution

Let $Y_1, \dots, Y_n$ be independent $N(0_{m\times1}, \Sigma)$. The random matrix $W = \sum_{i=1}^{n} Y_i Y_i^T$ is said to have a Wishart distribution, denoted by $W_m(\Sigma, n)$.

Recall that $\chi_n^2 \equiv \Gamma(n/2, 1/2)$, so the density of $W_1(\sigma^2, n) \equiv \chi_n^2$ is

$$f_{W_1}(w) = w^{(n-2)/2} e^{-\frac{w}{2\sigma^2}} \cdot \frac{1}{(2\sigma^2)^{n/2}\Gamma(n/2)}$$

The density of the Wishart distribution $W_m(\Sigma, n)$ generalizes this to:

$$f(W) = \frac{(\det(W))^{(n-m-1)/2} e^{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}W)}}{(2^m \det(\Sigma))^{n/2}\,\Gamma_m(n/2)}$$

for all positive definite matrices W

where $\Gamma_m(\cdot)$ is the multivariate gamma function.

$$\Gamma_m(t) = \pi^{m(m-1)/4} \cdot \prod_{i=1}^{m} \Gamma\left(t - \frac{i-1}{2}\right)$$

Wishart distribution has many applications in CAPM testing, and in Bayesian statistics.

<u>FACT 10</u> - If $W \sim W_m(\Sigma, n)$, then $E[W] = n\Sigma$.

- If $W_1, W_2, \dots, W_k$ are independent with $W_i \sim W_m(\Sigma, n_i)$, then
$$\sum_{i=1}^{k} W_i \sim W_m\left(\Sigma, \sum_{i=1}^{k} n_i\right)$$
- If $W \sim W_m(\Sigma, n)$ and A is a nonrandom $m\times m$ nonsingular matrix, then $AWA' \sim W_m(A\Sigma A', n)$.
- $W := \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})' \sim W_m(\Sigma, n-1)$.

$\boxed{\text{Multivariate } t\text{-distribution}}$

Let $Z \sim N(0_{m\times 1}, \Sigma)$ and $W \sim W_m(\Sigma, k)$ be independent. Then $(W/k)^{-1/2} \cdot Z$ is said to have the m-variate t-distribution with k degrees of freedom.

The density function is

$$f(t) = \frac{\Gamma\left(\frac{k+m}{2}\right)}{(\pi k)^{\frac{m}{2}}\Gamma(k/2)}\left(1+\frac{\|t\|^2}{k}\right)^{-\frac{k+m}{2}}, \quad t \in \mathbb{R}^m$$

(which reduces to the expression on bottom of page 6. in the case $m=1$).

m-variate t-distribution is used in risk management (statistical models for VaR) and t-copulas.

The square of a t_k-distributed random variable (i.e. univariate with k degrees of freedom) is actually $F_{1,k}$-distributed.

Moregenerally, if t has the m-variate t-distribution with k degrees of freedom ($k \geq m$), then $\frac{k-m+1}{km}\|t\|^2$ has the $F_{m,k-m+1}$ distribution. (*)

Now, lets go back to the sample setup:

Let $Y_1, \dots, Y_n$ independent $m\times 1$ random $N(\mu, \Sigma)$ vectors with $n > m$ and positive definite Σ. We know that $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\frac{W}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})(Y_i-\bar{Y})'$ are independent; $\bar{Y} \sim N\left(\mu, \frac{\Sigma}{n}\right)$; $W \sim W_m(\Sigma, n-1)$.

Define the <u>Hotelling's T^2-statistic</u> (famous in multivariate hypothesis testing)

$$T^2 = n(\bar{Y}-\mu)'\left(\frac{W}{n-1}\right)^{-1}(\bar{Y}-\mu)$$

Note that $\left(\frac{W}{n-1}\right)^{-1/2}\left(\sqrt{n}(\bar{Y}-\mu)\right) \sim \left(\frac{W_m(\Sigma, n-1)}{n-1}\right)^{-1/2} N(0, \Sigma)$ has the m-variate t-distribution with $n-1$ degrees of freedom

Then, according to (*), $\frac{n-1-m+1}{(n-1)m}T^2 = \frac{n-m}{(n-1)m}T^2 \sim F_{m,n-1-m+1}$, i.e. $\boxed{\frac{n-m}{(n-1)m}T^2 \sim F_{m,n-m}}$

2 Method of maximum likelihood

Given data of any kind, we're often faced with the following questions:

A) How to estimate the unknown parameters of a distribution given the data from it?

B) How good are these estimates; are they close to the actual "true" parameters?

C) Does the data come from a particular type of distribution; for example, normal or gamma

First, we'll keep it simple and study Questions A) and B), while assuming that we know what type of distribution the sample comes from (so we only do not know the parameters of this distribution).

Consider a family of distributions $\mathbb{P}_\theta$ indexed by a parameter (which, in general, could be a vector of parameters) θ that belongs to a set Θ. For example, we could be considering a family of normal distributions $N(\mu, \sigma^2)$ in which case $\theta = (\mu, \sigma^2)'$.

Let $f_\theta(x_1, x_2, \dots, x_n)$ be the joint density function of $X_1, \dots, X_n$. The <u>likelihood function</u> based on the observations $X_1, \dots, X_n$ is $L(\theta) = f_\theta(X_1, \dots, X_n)$ and the MLE (<u>maximum likelihood estimate</u>) $\hat{\theta}$ of θ is the value of θ that maximizes $L(\theta)$, over all $\theta \in \Theta$.

More often than not, the sample $X_1, \dots, X_n$ is assumed to be independent(*); so

$L(\theta) = f_\theta(X_1) \cdot f_\theta(X_2) \cdot \dots \cdot f_\theta(X_n)$, where $f_\theta(x)$ is the p.d.f. of the distribution $\mathbb{P}_\theta$

(Make sure you understand that $X_1, \dots, X_n$ are given; so L is a function of θ only!)

Intuitively, the likelihood function is the probability to observe the sample $X_1, \dots, X_n$ when the unknown parameters of the distribution are equal to θ.

When finding the MLE, it is sometimes easier to maximize the <u>log-likelihood function</u>

$$\ell(\theta) = \log f_\theta(X_1, \dots, X_n) \quad \text{instead} \quad (\underline{\text{Note}}: \log x \text{ is an increasing function})$$

When $X_1, \dots, X_n$ are independent, then $\ell(\theta) = \sum_{i=1}^{n} \ell_\theta(X_i)$, where $\ell_\theta(x) = \log f_\theta(x)$

Let's do several examples of calculating the MLE!

(*) Without assuming that X_i are independent, law of large numbers and CLT could not be applied. However, one could still use martingale strong laws and central limit theorems; and most of the results here would still hold, under some regularity conditions.

Example 1. Bernoulli distribution $B(p)$

$X \sim B(p) \qquad 0 \le p \le 1 \qquad \mathbb{P}(X=1)=p$

$\mathbb{P}(X=0)=1-p$

p.d.f $\quad f_p(x) = \begin{cases} p, & \text{if } x=1 \\ 1-p, & \text{if } x=0 \end{cases} \quad$ OR $\quad f_p(x) = p^x(1-p)^{1-x}$

likelihood function $L(p) = \prod_{i=1}^{n} f_p(X_i) = p^{\#\text{of 1's in } X_1,\ldots,X_n} \cdot (1-p)^{\#\text{of 0's in } X_1,\ldots,X_n}$

$$L(p) = p^{X_1+\ldots+X_n}(1-p)^{n-(X_1+\ldots+X_n)}$$

log-likelihood function $\ell(p) = \left(\sum_{i=1}^{n} X_i\right)\log p + \left(n - \sum_{i=1}^{n} X_i\right)\log(1-p)$

$\frac{d}{dp}(\ell(p)) = 0 \Rightarrow \frac{1}{p}\sum_{i=1}^{n} X_i - \left(n - \sum_{i=1}^{n} X_i\right)\cdot \frac{1}{1-p} = 0$ $\qquad$ Solve for p to get:

$$p = \frac{X_1+\ldots+X_n}{n} = \overline{X}$$

Therefore, the proportion of successes $\hat{p} = \overline{X}$ in the sample is the MLE for p, which is perfectly intuitive. Note that by the law of large numbers (FACT 2), we have $\hat{p} = \overline{X} \to \mathbb{E}[X_1] = p$ (in probability), which means that our MLE will approximate the unknown parameter p well when we get more and more data. More about this in a second, once we start talking about consistency of the MLE.

Example 2. Normal distribution $N(\mu, \sigma^2)$ $\qquad$ Let $\theta = (\mu, \sigma^2)'$

p.d.f. $\quad f_\theta(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

$$\ell(\theta) = \sum_{i=1}^{n}\left(\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\sigma - \frac{(X_i-\mu)^2}{2\sigma^2}\right)$$

$$\ell(\theta) = n\log\left(\frac{1}{\sqrt{2\pi}}\right) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

We need to maximize $\ell(\theta)$ over $\Theta = \{(\mu,\sigma^2)' \mid \mu \in (-\infty,\infty), \sigma^2 > 0\} \subseteq \mathbb{R}^2$

The usual approach would be to find the MLE by solving the equation $\nabla \ell(\theta) = 0_{2\times 1}$, where $\nabla \ell$ is the gradient vector of partial derivatives

$$\nabla \ell = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \end{pmatrix} \equiv \begin{pmatrix} \frac{\partial \ell}{\partial \mu} \\ \frac{\partial \ell}{\partial (\sigma^2)} \end{pmatrix}$$; but in this case, we the "conditional log likelihood method".

First, for any σ^2, we minimize $\sum_{i=1}^{n} (X_i - \mu)^2$ over μ:

$$\frac{d}{d\mu} \sum_{i=1}^{n} (X_i - \mu)^2 = 0 \quad \text{gives} \quad -2 \sum_{i=1}^{n} (X_i - \mu) = 0, \text{ i.e. } \boxed{\hat{\mu} = \overline{X}.}$$

We plug in this estimate in the log-likelihood function to obtain the conditional log-likelihood function:

$$n \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2$$, which needs to be maximized over σ^2. Letting $\gamma \equiv \sigma^2$ in the last expression and taking $\frac{d}{d\gamma}$, we get:

$$-\frac{n}{2\gamma} + \frac{1}{2\gamma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = 0, \text{ i.e. } \hat{\gamma} \equiv \boxed{\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2},$$

which is only slightly different from the sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

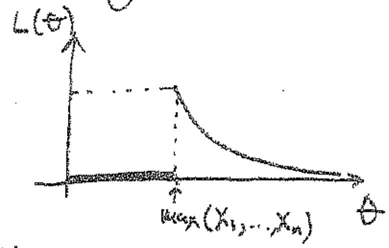
Example 3. Uniform distribution $U[0, \theta]$

p.d.f. $f_\theta(x) = \begin{cases} 1/\theta, & \text{if } 0 \le x \le \theta \\ 0, & \text{otherwise} \end{cases}$

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i) = \frac{1}{\theta^n} \underbrace{I}_{\text{indicator random variable for the event in parentheses}} (X_1 \in [0, \theta] \text{ and } X_2 \in [0, \theta] \text{ and } \dots \text{ and } X_n \in [0, \theta])$$

Simpler way to write it:

$$L(\theta) = \frac{1}{\theta^n} I(\max(X_1, \dots, X_n) \le \theta) = \begin{cases} 0, & \text{if } \theta < \max(X_1, \dots, X_n) \\ \frac{1}{\theta^n}, & \text{if } \theta \ge \max(X_1, \dots, X_n) \end{cases}$$



In this example, no need to go to $\ell(\theta)$. Also, we cannot differentiate $L(\theta)$ w.r.t. θ.

Nonetheless, it's easy to see how to maximize the likelihood function!

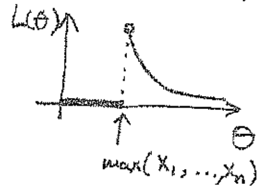
Set $\hat{\theta} = \max(X_1, \dots, X_n)$. This is the MLE for θ!

Note: It is often not easy to find the MLE as in the above examples, so numerical procedur (such as Newton-Raphson) need to be used. Also, MLE does not always exist!

Here is an artificial example based on Example 3.

Consider $\mathbb{P}_\theta$ to be $U[0,\theta)$ uniform on $[0,\theta)$ (where θ is unknown).

Then, similarly as before, $L(\theta) = \frac{1}{\theta^n} I\,(\max(X_1,\dots,X_n) < \theta)$ and the maximum at the point $\hat{\theta} = \max(X_1,\dots,X_n)$ is not achieved.



Question: Why are MLE's good?

Because of consistency and asymptotic normality. Next, we explain these concepts in the univariate case, i.e. when θ is just 1-dimensional.

The multivariate case (when θ is a vector of unknown parameters) is very similar and will be mentioned at the end of this section.

$\boxed{\text{Consistency}}$ ("no bias") We say that the MLE $\hat{\theta}$ is consistent if $\hat{\theta} \to \theta_0$ in probability, as $n \to \infty$, where θ_0 is the true value of the unknown parameter of the distribution of the sample.

("$\hat{\theta} \to \theta_0$ in prob. as $n \to \infty$" means "$\forall \varepsilon > 0 \quad \mathbb{P}(|\hat{\theta} - \theta_0| > \varepsilon) \to 0$ as $n \to \infty$")

$\boxed{\text{Asymptotic normality}}$ We say that $\hat{\theta}$ is asymptotically normal if as $n \to \infty$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^2_{\theta_0})$$

for some $\sigma^2_{\theta_0}$ which is called the asymptotic variance of the estimator $\hat{\theta}$.

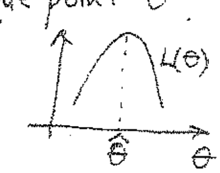
σ_{θ_0} is also known as the standard error of $\hat{\theta}$.

Actually, we will show that $\sigma^2_{\theta_0} = \frac{1}{I(\theta_0)}$, where $I(\theta_0)$ will be defined later as the Fisher information.

Let's attempt to prove consistency, at least intuitively. Assume that the likelihood function $L(\theta) = \prod_{i=1}^{n} f_\theta(X_i)$ is smooth and its maximum is achieved at a unique point $\hat{\theta}$.

$\hat{\theta}$ also maximizes $\ell(\theta)$ (the log-likelihood), as well as

$$\ell_n(\theta) := \frac{1}{n}\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i) \quad \text{(the log-likelihood "normalized" by } 1/n\text{)}$$



Aside from $\ell_n(\theta)$, we also introduce $\mathcal{L}(\theta) := E_{\theta_0}[\ell_\theta(X)]$, i.e.

$$\mathcal{L}(\theta) = \int \ell_\theta(x) f_{\theta_0}(x)\,dx$$

Note that while ℓ_n depends on the sample $X_1, X_2, \dots, X_n$; $\mathcal{L}(\theta)$ only depends on θ.

By <u>FACT 2</u> (Law of large numbers), for any θ, we have

$$\ell_n(\theta) \longrightarrow E_{\theta_0}[\ell_\theta(X)] = \mathcal{L}(\theta). \quad (*)$$

<u>Claim</u> $\mathcal{L}(\theta) \le \mathcal{L}(\theta_0)$ for every θ.

proof: Recall <u>Jensen's inequality</u>: Suppose Z has a finite mean and $\varphi: \mathbb{R} \to \mathbb{R}$ is convex, i.e. $\lambda\varphi(x) + (1-\lambda)\varphi(y) \ge \varphi(\lambda x + (1-\lambda)y)$ $\forall 0<\lambda<1, \forall x, y \in \mathbb{R}$. Then $E[\varphi(Z)] \ge \varphi(E[Z])$.

Since log is a concave function (i.e. $-\log$ is convex), then by Jensen's inequality for $Z \equiv \frac{f_\theta(X)}{f_{\theta_0}(X)}$, we get: $E_{\theta_0}[-\log Z] \ge -\log(E_{\theta_0}[Z])$, i.e.

$E_{\theta_0}[\log Z] \le \log(E_{\theta_0}[Z])$, i.e.

$$E_{\theta_0}\left[\log\left(\frac{f_\theta(X)}{f_{\theta_0}(X)}\right)\right] \le \log\left(\int \frac{f_\theta(x)}{\cancel{f_{\theta_0}(x)}} \cancel{f_{\theta_0}(x)}\,dx\right) = \log\Big(\overset{1}{\overset{\|}{\int f_\theta(x)\,dx}}\Big) = \log 1 = 0$$

(integrating the p.d.f.)

|| linearity of expectation

$E_{\theta_0}[\log(f_\theta(X))] - E_{\theta_0}[\log(f_{\theta_0}(X))]$

|| definition of ℓ

$E_{\theta_0}[\ell_\theta(X)] - E_{\theta_0}[\ell_{\theta_0}(X)]$

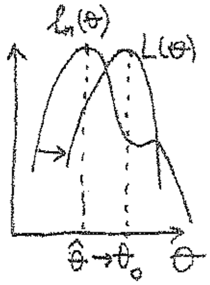
|| definition of $\mathcal{L}$

$\mathcal{L}(\theta) - \mathcal{L}(\theta_0)$

So, $\mathcal{L}(\theta) - \mathcal{L}(\theta_0) \le 0$, i.e. $\mathcal{L}(\theta) \le \mathcal{L}(\theta_0)$ $\square$.

<u>Theorem</u> Under some regularity conditions on the family of distributions (which we don't sta[te] here), the MLE $\hat{\theta}$ is consistent, i.e. $\hat{\theta} \to \theta_0$ as $n \to \infty$.

(intuitive) proof: Let's summarize what we discussed so far:



1. $\hat\theta$ is the maximizer of $l_n(\theta)$ (by definition of MLE)
2. θ_0 is the maximizer of $\mathcal{L}(\theta)$ by previous claim.
3. $\forall \theta$, we have $l_n(\theta) \to \mathcal{L}(\theta)$ by $(*)$.

Since two functions l_n and $\mathcal{L}$ are getting closer, the points of maximum should also get closer, which means $\hat\theta \to \theta_0$ $\square$

Before we discuss asymptotic normality, let's define Fisher Information:

Definition (univariate case) Fisher information of a random variable X with distribution $\mathbb{P}_{\theta_0}$ from the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is defined by:

$$I(\theta_0) = E_{\theta_0}\left[\left(\frac{d}{d\theta} l_\theta(X)\Big|_{\theta=\theta_0}\right)^2\right]$$

Question: How to think of Fisher information intuitively?

It can be interpreted as a measure of how quickly the p.d.f. will change when one changes the parameter θ near θ_0 slightly.

$$\frac{d}{d\theta} l_\theta(X)\Big|_{\theta=\theta_0} = \frac{d}{d\theta}\left(\log f_\theta(x)\right)\Big|_{\theta=\theta_0} = \frac{\left(\frac{d}{d\theta} f_\theta(x)\right)\Big|_{\theta=\theta_0}}{f_{\theta_0}(x)}$$

When we square this and take expectation, i.e. average over all x, we get an "average" version of this measure. So, if Fisher information is large, this means that the distribution will change quickly when we move the parameter away from its true value θ_0, so the distribution with parameter θ_0 is "quite different" and can be "well distinguished" from the distributions with parameters away from θ_0. If Fisher information is small, this means that the distribution is "very similar" to the distributions with parameters away from θ_0 and thus more difficult to distinguish; so our estimation will be worse.

There is an alternative formula for Fisher information, which is actually used more often than the definition above.

<u>Formula for $I(\theta_0)$</u>: $\quad I(\theta_0) = -E_{\theta_0}\left[\left(\frac{\partial^2}{\partial\theta^2} l_\theta(x)\right)\Big|_{\theta=\theta_0}\right]$

proof: First, let's find an "appropriate" expression for $\left(\frac{\partial^2}{\partial\theta^2} l_\theta(x)\right)\Big|_{\theta=\theta_0}$ to be used in this proof

$$\left(\frac{\partial}{\partial\theta} l_\theta(x)\right)\Big|_{\theta=\theta_0} = \frac{\left(\frac{\partial}{\partial\theta} f_\theta(x)\right)\Big|_{\theta=\theta_0}}{f_{\theta_0}(x)}$$

using the definition of $l_\theta(x)$ as $\log f_\theta(x)$.

Quotient rule then gives:

$$\left(\frac{\partial^2}{\partial\theta^2} l_\theta(x)\right)\Big|_{\theta=\theta_0} = \frac{\left(\frac{\partial^2}{\partial\theta^2} f_\theta(x)\right)\Big|_{\theta=\theta_0}}{f_{\theta_0}(x)} - \left(\frac{\left(\frac{\partial}{\partial\theta} f_\theta(x)\right)\Big|_{\theta=\theta_0}}{f_{\theta_0}(x)}\right)^2, \text{ i.e.}$$

$$\left(\frac{\partial^2}{\partial\theta^2} l_\theta(x)\right)\Big|_{\theta=\theta_0} = \frac{\left(\frac{\partial^2}{\partial\theta^2} f_\theta(x)\right)\Big|_{\theta=\theta_0}}{f_{\theta_0}(x)} - \left(\left(\frac{\partial}{\partial\theta} l_\theta(x)\right)\Big|_{\theta=\theta_0}\right)^2 \quad \oplus$$

Now,

$$E_{\theta_0}\left[\left(\frac{\partial^2}{\partial\theta^2} l_\theta(x)\right)\Big|_{\theta=\theta_0}\right] = \int \left(\frac{\partial^2}{\partial\theta^2} l_\theta(x)\right)\Big|_{\theta=\theta_0} \cdot f_{\theta_0}(x)\,dx =$$

$$= \int \frac{\left(\frac{\partial^2}{\partial\theta^2} f_\theta(x)\right)\Big|_{\theta=\theta_0}}{\cancel{f_{\theta_0}(x)}} \cdot \cancel{f_{\theta_0}(x)}\,dx - \int\left(\left(\frac{\partial}{\partial\theta} l_\theta(x)\right)\Big|_{\theta=\theta_0}\right)^2 f_{\theta_0}(x)\,dx =$$

$$= \int \left(\frac{\partial^2}{\partial\theta^2} f_\theta(x)\right)\Big|_{\theta=\theta_0} dx - E_{\theta_0}\left[\left(\frac{\partial}{\partial\theta} l_\theta(x)\Big|_{\theta=\theta_0}\right)^2\right] = 0 - I(\theta_0) = -I(\theta_0)$$

Why is $\int \left(\frac{\partial^2}{\partial\theta^2} f_\theta(x)\right)\Big|_{\theta=\theta_0} dx = 0$?

Since $f_\theta(x)$ is a p.d.f., one has: $\int f_\theta(x)\,dx = 1$, so $\frac{\partial}{\partial\theta}\int f_\theta(x)\,dx = 0$, i.e.

$\int \frac{\partial}{\partial\theta} f_\theta(x)\,dx = 0$ and $\int\left(\frac{\partial^2}{\partial\theta^2} f_\theta(x)\right)\Big|_{\theta=\theta_0} dx = 0$. $\square$

Theorem 2 $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$ as $n \to \infty$

Note: Larger $I(\theta_0)$ means better MLE, since its asymptotic variance/dispersion around the true value of the parameter is smaller.

proof: Recall from the proof of Theorem 1 that $\hat{\theta}$ maximizes $l_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(x_i)$

so: $$\left(\frac{\partial}{\partial\theta} l_n(\theta)\right)\Big|_{\theta=\hat{\theta}} = 0.$$

By Taylor's theorem (or by mean value theorem: $f(a) = f(b) + f'(c)(a-b)$ for some $c \in [a,b]$)

we get that for some $\theta^* \in [\hat{\theta}, \theta_0]$: $\longleftarrow$ $f(\theta) = \frac{\partial}{\partial\theta} l_n(\theta)$, $a = \hat{\theta}$, $b = \theta_0$

$$0 = \left(\frac{\partial}{\partial\theta} l_n(\theta)\right)\Big|_{\theta=\hat{\theta}} = \left(\frac{\partial}{\partial\theta} l_n(\theta)\right)\Big|_{\theta=\theta_0} + \left(\frac{\partial^2}{\partial\theta^2} l_n(\theta)\right)\Big|_{\theta=\theta^*} \cdot (\hat{\theta}-\theta_0)$$

so $$\hat{\theta}-\theta_0 = -\frac{\left(\frac{\partial}{\partial\theta} l_n(\theta)\right)\Big|_{\theta=\theta_0}}{\left(\frac{\partial^2}{\partial\theta^2} l_n(\theta)\right)\Big|_{\theta=\theta^*}}, \text{ i.e. } \boxed{\sqrt{n}(\hat{\theta}-\theta_0) = -\frac{\sqrt{n}\left(\frac{\partial}{\partial\theta} l_n(\theta)\right)\Big|_{\theta=\theta_0}}{\left(\frac{\partial^2}{\partial\theta^2} l_n(\theta)\right)\Big|_{\theta=\theta^*}}} \quad \circledast$$

Recall from the proof of Theorem 1 that θ_0 maximizes $\mathcal{L}(\theta) = E_{\theta_0}[l_\theta(x)]$, so

$$\left(\frac{\partial}{\partial\theta} \mathcal{L}(\theta)\right)\Big|_{\theta=\theta_0} = E_{\theta_0}\left[\left(\frac{\partial}{\partial\theta} l_\theta(x)\right)\Big|_{\theta=\theta_0}\right] = 0 \qquad \oplus$$

Consider the numerator in $\circledast$. We have:

$$\sqrt{n}\left(\frac{\partial}{\partial\theta} l_n(\theta)\right)\Big|_{\theta=\theta_0} = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta} l_\theta(x_i)\right)\Big|_{\theta=\theta_0} - 0\right) =$$

using $\oplus$ here $\longrightarrow$ $$= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta} l_\theta(x_i)\right)\Big|_{\theta=\theta_0} - E_{\theta_0}\left[\left(\frac{\partial}{\partial\theta} l_\theta(x_1)\right)\Big|_{\theta=\theta_0}\right]\right)$$

By FACT3 (central limit theorem)

$$\sqrt{n}\left(\frac{\partial}{\partial\theta} l_n(\theta)\right)\Big|_{\theta=\theta_0} \xrightarrow{d} N\left(0, Var_{\theta_0}\left(\left(\frac{\partial}{\partial\theta} l_\theta(x_1)\right)\Big|_{\theta=\theta_0}\right)\right)$$

Consider the denominator in ⊛. By FACT 2 (Law of large numbers), for every θ:

$$\frac{\partial^2}{\partial\theta^2} l_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} l_\theta(X_i) \longrightarrow E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2} l_\theta(X_1)\right]$$

Since $\theta^* \in [\hat\theta, \theta_0]$ and since (by Theorem 1) $\hat\theta \mapsto \theta_0$, then also $\theta^* \to \theta_0$, and:

$$\left(\frac{\partial^2}{\partial\theta^2} l_n(\theta)\right)\Bigg|_{\theta=\theta^*} \longrightarrow E_{\theta_0}\left[\left(\frac{\partial^2}{\partial\theta^2} l_\theta(X_1)\right)\Bigg|_{\theta=\theta^*}\right] \longrightarrow E_{\theta_0}\left[\left(\frac{\partial^2}{\partial\theta^2} l_\theta(X_1)\right)\Bigg|_{\theta=\theta_0}\right], \text{ i.e.}$$

by Formula for $I(\theta_0)$:

$$\left(\frac{\partial^2}{\partial\theta^2} l_n(\theta)\right)\Bigg|_{\theta=\theta^*} \longrightarrow -I(\theta_0).$$

<u>Slutsky's theorem</u>: If $Z_n \xrightarrow{d} Z$; $W_n \to \omega$ in probability (where ω is nonrandom),

then $W_n Z_n \xrightarrow{d} \omega Z$

Using this theorem and our convergence results for the numerator & denominator of ⊛, we get:

$$\sqrt{n}\,(\hat\theta - \theta_0) \xrightarrow{d} N\left(0, \frac{Var_{\theta_0}\left(\left(\frac{\partial}{\partial\theta} l_\theta(X_1)\right)\Big|_{\theta=\theta_0}\right)}{(I(\theta_0))^2}\right)$$

Finally, $Var_{\theta_0}\left(\left(\frac{\partial}{\partial\theta} l_\theta(X_1)\right)\Big|_{\theta=\theta_0}\right) = E_{\theta_0}\left[\left(\left(\frac{\partial}{\partial\theta} l_\theta(X_1)\right)\Big|_{\theta=\theta_0}\right)^2\right] - \left(E_{\theta_0}\left[\left(\frac{\partial}{\partial\theta} l_\theta(X_1)\right)\Big|_{\theta=\theta_0}\right]\right)^2 =$

$$= I(\theta_0) - 0 = I(\theta_0).$$

by definition of Fisher information and by ⊛

So, $\sqrt{n}\,(\hat\theta - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$ □

Let's do some examples.

Example 4. The family of Bernoulli distributions $B(p)$ has p.d.f. $f_p(x) = p^x(1-p)^{1-x}$.

So, $\ell_p(x) = \log f_p(x) = x\log p + (1-x)\log(1-p)$

$$\frac{\partial}{\partial p}\ell_p(x) = \frac{x}{p} - \frac{1-x}{1-p} \quad ; \quad \frac{\partial^2}{\partial p^2}\ell_p(x) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

Fisher information $I(p)$ by formula on page 20. is:

$$I(p_0) = -E_{p_0}\left[\left(\frac{\partial^2}{\partial p^2}\ell_p(x)\right)\Big|_{p=p_0}\right] = \frac{E_{p_0}[X]}{p_0^2} + \frac{1-E_{p_0}[X]}{(1-p_0)^2} = \frac{p_0}{p_0^2} + \frac{1-p_0}{(1-p_0)^2}$$

$$\Rightarrow I(p_0) = \frac{1}{p_0(1-p_0)}$$

We know from example 1 that the MLE $\hat{p} = \overline{X}$. Asymptotic normality of $\hat{p}$ now states that:

$$\sqrt{n}(\hat{p} - p_0) \xrightarrow{d} N(0, p_0(1-p_0)) \quad \text{as } n\to\infty$$

(Which, of course, also follows directly from the CLT! Check!)

Example 5. The family of exponential distributions $\Xi(\alpha)$ has p.d.f.:

$$f_\alpha(x) = \begin{cases} \alpha e^{-\alpha x}, & \text{if } x \geq 0 \\ 0, & \text{if } x<0 \end{cases}$$

Check that the MLE $\hat{\alpha} = \frac{1}{\overline{X}}$!

$$\ell_\alpha(x) = \log f_\alpha(x) = \log\alpha - \alpha x$$

$$\frac{\partial}{\partial\alpha}\ell_\alpha(x) = \frac{1}{\alpha} - x \quad ; \quad \frac{\partial^2}{\partial\alpha^2}\ell_\alpha(x) = -\frac{1}{\alpha^2} \quad \text{(does not depend on } X\text{)}$$

$$I(\alpha_0) = -E_{\alpha_0}\left[\left(\frac{\partial^2}{\partial\alpha^2}\ell_\alpha(x)\right)\Big|_{\alpha=\alpha_0}\right] = -E_{\alpha_0}\left[-\frac{1}{\alpha_0^2}\right] = \frac{1}{\alpha_0^2}$$

So, the asymptotic normality gives:

$$\sqrt{n}\left(\frac{1}{\overline{X}} - \alpha_0\right) \xrightarrow{d} N(0, \alpha_0^2) \quad \text{as } n\to\infty$$

Example 6. The family of normal distributions $N(\mu, \sigma^2)$ with known σ^2 (but unknown μ).

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

$$l_\mu(x) = \log f_\mu(x) = -\frac{1}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \frac{1}{2\sigma^2}(X-\mu)^2$$

$$\frac{\partial}{\partial\mu} l_\mu(x) = \frac{1}{\sigma^2}(X-\mu) \quad ; \quad \frac{\partial^2}{\partial\mu^2} l_\mu(x) = -\frac{1}{\sigma^2} \quad \text{(does not depend on } x\text{)}$$

$$I(\mu_0) = -E_{\mu_0}\left[\left(\frac{\partial^2}{\partial\mu^2} l_\mu(x)\right)\Big|_{\mu=\mu_0}\right] = -E_{\mu_0}\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2}$$

So, the asymptotic normality gives:

$$\sqrt{n}(\hat{\mu}-\mu_0) \xrightarrow{d} N(0,\sigma^2). \qquad \text{Now, the MLE } \hat{\mu} = \bar{X} \text{ (check! why?)}$$

$$\sqrt{n}(\bar{X}-\mu_0) \xrightarrow{d} N(0,\sigma^2), \text{ i.e. } \bar{X} \xrightarrow{d} N\left(\mu_0, \frac{\sigma^2}{n}\right), \text{ as } n\to\infty$$

This matches the usual statistician's statement that the standard error of the MLE is $\frac{\sigma}{\sqrt{n}}$, when σ is known.

Question: What if θ, the vector of unknown parameters, is, say, p-dimensional?
(multivariate case) Not much changes! The proofs and results are very similar.

Fisher information matrix is a $p\times p$ matrix

$$I(\theta_0) = -\left(E_{\theta_0}\left[\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} l_\theta(x)\right)\Big|_{\theta=\theta_0}\right]\right)_{1\le i,j\le p}$$

Asymptotic variance now states that

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N\left(0, \left(I(\theta_0)\right)^{-1}\right) \quad \text{as } n\to\infty$$

Here, θ_0 denotes the $p\times 1$ vector of true values of unknown parameters.

FACT 11 (no proof here)

$$(\hat{\theta}-\theta_0)'\left(-\left(\nabla^2 l(\theta)\right)\Big|_{\theta=\hat{\theta}}\right)(\hat{\theta}-\theta_0) \xrightarrow{d} \chi^2_p \quad \text{as } n\to\infty$$

where $\nabla^2 l(\theta) = \left(\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\right)_{1\le i,j\le p}$ is the Hessian $p\times p$ matrix of second partial derivativ

$-\nabla^2(l(\theta))\Big|_{\theta=\hat{\theta}}$ is called the observed Fisher information matrix

3 CONFIDENCE INTERVALS

Set-up: MLE's for normal distribution and their distribution

Consider a random sample $X_1, X_2, \dots X_n \sim N(\mu, \sigma^2)$, μ, σ^2 unknown.

From Example 2 Section 2: MLE for μ: $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

MLE for σ^2: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \overset{\text{some algebra}}{=} \overline{X^2} - (\overline{X})^2$, where $\overline{X^2} = \frac{1}{n}\sum X_i^2$

Question: How close are these estimates to the true values?

Note: By Law of Large Numbers $\overline{X} \to \mu$, $\overline{X^2} - (\overline{X})^2 \to \sigma^2$ as $n \to \infty$ but how quick is the convergence? Can we say more?

Answer: Want to construct intervals of values that will with certain pre-specified probability contain the true value of the $\underbrace{\text{parameter}}_{\text{unknown}}$.

Note: Unlike in Bayesian theory where the parameter is considered random and interval fixed, and where we talk about a probability that the unknown parameter belongs to a fixed ("credible") interval; HERE, the unknown parameter is fixed, while the endpoints of the interval are random and have a probability distribution.

We start with

Theorem 1 If $X_1, \dots, X_n$ are i.i.d. $\sim N(0,1)$, then the sample mean $\overline{X}$ and the MLE variance $\overline{X^2} - (\overline{X})^2$ are independent and

$$\sqrt{n}\,\overline{X} \sim N(0,1) \quad \text{and} \quad n(\overline{X^2} - (\overline{X})^2) \sim \chi^2_{n-1}$$

Chi-squared distribution with $n-1$ degrees of freedom

In other words, $\hat{\mu}$ and $\hat{\sigma}^2$ are independent,

$$\sqrt{n}\,\hat{\mu} \sim N(0,1) \quad \text{and} \quad n\hat{\sigma}^2 \sim \chi^2_{n-1}$$

Note: This is actually Theorem on page 7 (Section 1)! Notice that there is a slight difference between the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and the MLE variance $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

proof: Consider $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = VX = \begin{pmatrix} 1/\sqrt{n} & \dots & 1/\sqrt{n} \\ & ? & \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$

some orthogonal transformation of X, where the first row of V is taken to be $(1/\sqrt{n} \dots 1/\sqrt{n})'$ and the remaining rows are to be any orthogonal basis in the hyperplane orthogonal to this unit vector.

Now, $Y_1, \dots, Y_n$ are also i.i.d. standard normal (FACT 6.) and moreover,

$$Y_1 = \frac{1}{\sqrt{n}} X_1 + \dots + \frac{1}{\sqrt{n}} X_n = \sqrt{n}\,\overline{X}, \text{ i.e. } \boxed{\overline{X} = \frac{1}{\sqrt{n}} Y_1} \quad (1)$$

Also, $n\left(\overline{X^2} - (\overline{X})^2\right) = X_1^2 + \dots + X_n^2 - \left(\frac{1}{\sqrt{n}}(X_1 + \dots + X_n)\right)^2 = X_1^2 + \dots + X_n^2 - Y_1^2$

V is orthogonal, so $\|Y\| = \|VX\| = \|X\|$, i.e. $Y_1^2 + \dots + Y_n^2 = X_1^2 + \dots + X_n^2$,

so $$\boxed{n\left(\overline{X^2} - (\overline{X})^2\right) = Y_1^2 + \dots + Y_n^2 - Y_1^2 = Y_2^2 + \dots + Y_n^2} \quad (2)$$

So, (1) & (2) $\Rightarrow$ $\overline{X}$ and $\overline{X^2} - (\overline{X})^2$ are independent,

$\sqrt{n}\,\overline{X} = Y_1 \sim N(0,1)$ and $n\left(\overline{X^2} - (\overline{X})^2\right) \sim \chi^2_{n-1}$ $\square$

<u>Corollary 1.</u> If $X_1, \dots, X_n$ are i.i.d. $\sim N(\mu, \sigma^2)$, then the MLE's $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = \overline{X^2} - (\overline{X})^2$ are independent and

$$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \sim N(0,1), \quad \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

<u>Note:</u> We know the complete joint distribution of $\hat{\mu}$ and $\hat{\sigma}^2$.

OK! So, let's observe a sample $X_1, \dots, X_n$ with distribution $\mathbb{P}_{\theta_0}$ from a parametric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ and θ_0 is unknown

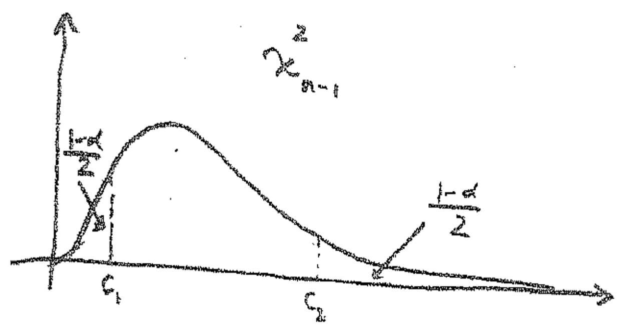
Given a confidence parameter $\alpha \in [0,1]$ (usually $\alpha = 0.95$), if exist two statistics $S_1 = S_1(X_1,\dots,X_n)$ & $S_2 = S_2(X_1,\dots,X_n)$ such that probability $\mathbb{P}_{\theta_0}(S_1 \le \theta_0 \le S_2) \ge \alpha$, then we call $[S_1, S_2]$ an α-confidence interval for the unknown parameter θ_0.

Let $X_1,\dots,X_n$ i.i.d $N(\mu,\sigma^2)$; μ, σ^2 unknown.

Corollary 1 $\Rightarrow$ $A = \dfrac{\sqrt{n}(\hat{\mu}-\mu)}{\sigma} \sim N(0,1)$ and $B = \dfrac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$.
A&B independent

So, we can represent A&B as $A = Y_1$, $B = Y_2^2 + \dots + Y_n^2$ for some $Y_1,\dots,Y_n$ i.i.d. $N(0,1)$

Choose points c_1, c_2 so that $P(c_1 \le B \le c_2) = \alpha$, i.e. the area btw. c_1 and c_2 is α, i.e. the area in each tail is $(1-\alpha)/2$. (see the Note on the bottom)



For those values of c_1 and c_2, we can GUARANTEE w/ confidence α that

$$c_1 \le B = \frac{n\hat{\sigma}^2}{\sigma^2} \le c_2$$

Solve this for the unknown parameter σ^2 $\Rightarrow$ the α-confidence interval for σ^2 is

$$\left[\frac{n\hat{\sigma}^2}{c_2}, \frac{n\hat{\sigma}^2}{c_1}\right]$$ where c_1, c_2 are such that

$$P(c_1 \le Z \le c_2) = \alpha \quad \text{where } Z \sim \chi^2_{n-1}$$

Note. Def. the q^{th} quantile u of a probability distribution of a continuous random variable U is defined by $\mathbb{P}(U \le u) = q$.

So, c_1 is nothing else but the $(1-\alpha)/2$-quantile of χ^2_{n-1}; while c_2 is the $\frac{1+\alpha}{2}$-quantile.

Sometimes one writes $c_1 \equiv \chi^2_{n-1;(1-\alpha)/2}$ and $c_2 = \chi^2_{n-1;(1+\alpha)/2}$

Next let's find an α-confidence interval for μ.

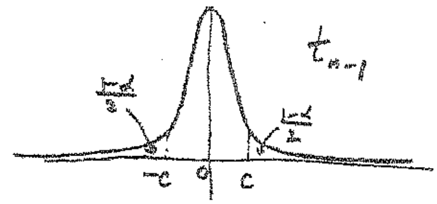
Consider $$\frac{A}{\sqrt{\frac{B}{n-1}}} = \frac{Y_1}{\sqrt{\frac{1}{n-1}(Y_2^2+\dots+Y_n^2)}} \sim t_{n-1}$$ student t-distr. w/ $n-1$ degrees of freedom

But, $$\frac{A}{\sqrt{\frac{B}{n-1}}} = \frac{\sqrt{n}\cdot\frac{\hat\mu-\mu}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{n\hat\sigma^2}{\sigma^2}}} = \frac{\sqrt{n-1}}{\hat\sigma}(\hat\mu-\mu)$$

So, choose $\overbrace{c}^{\text{a constant}}$ such that the area in each tail of the t_{n-1}-distr. is $\frac{1-\alpha}{2}$.

Then w/ probability α, we have

$$-c \le \frac{\sqrt{n-1}}{\hat\sigma}(\hat\mu-\mu) \le c.$$



So, the α-confidence interval for μ is $\left[\hat\mu - c\frac{\hat\sigma}{\sqrt{n-1}},\ \hat\mu + c\frac{\hat\sigma}{\sqrt{n-1}}\right]$

<u>Example</u> sample of size $n=10$ from $N(\mu,\sigma^2)$; μ,σ^2 unknown

X:

0.86
1.53
1.57
1.81
0.99
1.09
1.29
1.78
1.29
1.58

$\hat\mu = \overline{X} = 1.379$

$\hat\sigma^2 = \overline{X^2} - (\overline{X})^2 = 0.0266$ choose $\alpha = 0.95$ (95%)

Find c such that $t_9(-\infty, c) = 0.975 \to c = 2.262$

Find c_1, c_2 s.t. $\chi_9^2([0,c_1]) = 0.025 \longrightarrow c_1 = 2.7$

$\chi_9^2([0,c_2]) = 0.975 \longrightarrow c_2 = 19.02$

So, w/ prob. 95% we can guarantee that

$$\overline{X} - c\sqrt{\tfrac{1}{9}(\overline{X^2}-(\overline{X})^2)} \le \mu \le \overline{X} + c\sqrt{\tfrac{1}{9}(\overline{X^2}-(\overline{X})^2)}$$

i.e. $1.1446 \le \mu \le 1.6134$

and similarly for σ^2: $0.0508 \le \sigma^2 \le 0.3579$

4 Testing hypotheses (about normal distr. for now)

Setup: Given an i.i.d. sample $(X_1, \dots, X_n) \in S$ from $N(\mu, \sigma^2)$; μ, σ^2 unknown

sample space of all possible outcomes

We need to decide between two hypotheses about the unknown parameters, say μ. Hypotheses will be one of the following:

① $\begin{cases} H_0: \mu = \mu_0 & \text{(null)} \\ H_1: \mu \neq \mu_0 & \text{(alternative)} \end{cases}$ OR ② $\begin{cases} H_0: \mu \geq \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$ OR ③ $\begin{cases} H_0: \mu \leq \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$

Want to construct a TEST $\delta: S \to \{H_0, H_1\}$ that given an i.i.d. $(X_1, \dots, X_n) \in S$ either accepts H_0 or rejects H_0 (i.e. accepts H_1)

Choose $\alpha \in (0,1)$: level of significance for δ and we want to devise δ so that δ rejects H_0 when it's true w/ prob $\leq \alpha$, i.e.

usually 0.05

$$\mathbb{P}(\delta = H_1 \mid H_0) \leq \alpha.$$

$$\| $$

$$\sup_{(\mu, \sigma^2) \in \Omega_0} \mathbb{P}(\delta = H_1 \mid \mu, \sigma^2)$$

Ω_0 : subset of parameter space Ω s.t. $H_0: \theta \in \Omega_0$

unknown parameter

4.1. Hypothesis about mean of one normal sample

(ttest in Matlab)

We know that $\sqrt{n-1} \cdot \frac{\hat{\mu} - \mu}{\hat{\sigma}} \sim t_{n-1}$

Consider a t-statistic $\boxed{T = \sqrt{n-1} \cdot \frac{\hat{\mu} - \mu_0}{\hat{\sigma}}}$ another random var. It behaves differently depending on whether the true unknown mean $\mu = \mu_0$, $\mu < \mu_0$ or $\mu > \mu_0$. Why?

If $\mu = \mu_0$, then $T \sim t_{n-1}$. If $\mu < \mu_0$ then

$$T = \sqrt{n-1} \cdot \frac{\hat{\mu} - \mu}{\hat{\sigma}} + \sqrt{n-1} \cdot \frac{\mu - \mu_0}{\hat{\sigma}} \to -\infty$$

since the first term has t_{n-1}-distribution and the second one goes to $-\infty$. Similarly, if $\mu > \mu_0$, then $T \to +\infty$.

∴ Idea: base the tests on this statistics → so called t-tests.

① ($H_0: \mu = \mu_0$) The indication that H_0 is not true would be if $|T|$ becomes too large, i.e. $T \to \pm\infty$.

So, let's devise the following test: $\delta = \begin{cases} H_0, & \text{if } |T| \le c \\ H_1, & \text{if } |T| > c \end{cases}$

What is c? It depends on the level of significance α. We want.

$$\alpha \ge \mathbb{P}(\delta = H_1 | H_0) = \mathbb{P}(|T| > c | H_0) = t_{n-1}(|T| > c) = 2\, t_{n-1}(c, \infty) = \alpha$$

given that H_0 holds, we have $T \sim t_{n-1}$

So, from $2\, t_{n-1}(c, \infty) = \alpha$, you find c.

② ($H_0: \mu \ge \mu_0$) The indication that H_0 is not true would be if $T \to -\infty$.

So, $\delta = \begin{cases} H_0, & \text{if } T \ge c \\ H_1, & \text{if } T < c \end{cases}$ What is c? Depends on α again.

$$\alpha \ge \mathbb{P}(\delta = H_1 | H_0) = \mathbb{P}(T < c | H_0) =$$

$$= \mathbb{P}\left(T - \sqrt{n-1}\cdot\frac{\mu - \mu_0}{\hat{\sigma}} < c - \sqrt{n-1}\cdot\frac{\mu-\mu_0}{\hat{\sigma}} \,\Big|\, H_0\right) =$$

$$= \sup_{\mu \ge \mu_0} \mathbb{P}\left(T - \sqrt{n-1}\cdot\frac{\mu-\mu_0}{\hat{\sigma}} < c - \sqrt{n-1}\cdot\frac{\mu-\mu_0}{\hat{\sigma}}\right) =$$

this is t_{n-1}-distributed

this is maximized when $\mu = \mu_0$.

$$= t_{n-1}((-\infty, c]) = \alpha.$$

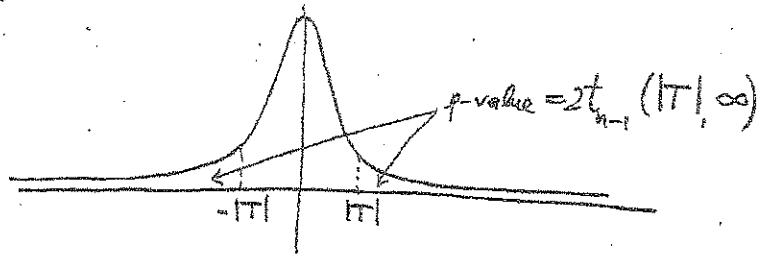
So, all you need to do is find c so that $t_{n-1}(-\infty, c) = \alpha$, and you have the t-test to test ②

For ③ similar!

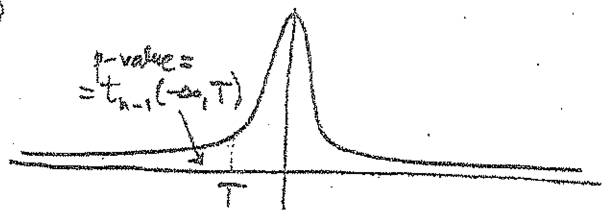
· **p-value**: Rather than specifying α and deciding whether to accept or reject H_0 at level α, we can ask "for what values of α do we reject H_0?"

p-value: the smallest value of α for which H_0 is rejected.

p-value can be understood as a probability, given that H_0 is true, to observe a value of t-statistic equally or less likely than the one that was observed. So, the small p-value means that the observed t-statistic is very unlikely under the null hypothesis, which in turn provides strong evidence against H_0.

① 

Stated differently, to perform a test using a given sample, we first find the p-value of the sample and then H_0 is rejected if we decide to use α larger than the p-value; and accept otherwise.

② 

(in this case, under H_0, $T \to +\infty$)

p-value also tells us whether the decision to accept or reject H_0 is a close call.

4.2. Hypothesis about variance of one normal sample

We know that $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$. So, similarly to 4.1, we base our tests on the following statistic:

$$Q = \frac{n\hat{\sigma}^2}{\sigma_0^2}$$

Since, $Q = \frac{n\hat{\sigma}^2}{\sigma^2} \cdot \frac{\sigma^2}{\sigma_0^2} \sim \frac{\sigma^2}{\sigma_0^2}\chi^2_{n-1}$, then Q behaves differently depending on whether $\sigma = \sigma_0$, $\sigma > \sigma_0$ or $\sigma < \sigma_0$ (exactly what we need!)

① ($H_0: \sigma = \sigma_0$) The decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } c_1 \le Q \le c_2 \\ H_1, & \text{if } Q < c_1 \text{ or } c_2 < Q \end{cases}$$

$\therefore$ Thresholds c_1, c_2 should satisfy the condition

$$\alpha = \mathbb{P}(\delta = H_1 | H_0) = \mathbb{P}(Q < c_1 | \sigma = \sigma_0) + \mathbb{P}(Q > c_2 | \sigma = \sigma_0) =$$
$$= \chi^2_{n-1}(0, c_1) + \chi^2_{n-1}(c_2, \infty)$$

So, for example, you can set $\chi^2_{n-1}(0, c_1) = \chi^2_{n-1}(c_2, \infty) = \alpha/2$.

② ($H_0: \sigma \le \sigma_0$) In this case, the decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } Q \le c \\ H_1, & \text{if } Q > c. \end{cases}$$

Threshold c should satisfy the condition

$$\alpha = \mathbb{P}(\delta = H_1 | H_0) = \sup_{\sigma \le \sigma_0} \mathbb{P}(Q > c) = \sup_{\sigma \le \sigma_0} \mathbb{P}\left(\frac{n\hat{\sigma}^2}{\sigma_0^2} > c\right) =$$

$$= \sup_{\sigma \le \sigma_0} \mathbb{P}\left(\frac{n\hat{\sigma}^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2} c\right) = \mathbb{P}\left(\frac{n\hat{\sigma}^2}{\sigma^2} > c\right) = \chi^2_{n-1}(c, \infty)$$

make this as small as possible under $\sigma \le \sigma_0$

χ^2_{n-1}-distributed

So, find c from $\alpha = \chi^2_{n-1}(c, \infty)$

③ ($H_0: \sigma \ge \sigma_0$) is similar.

4.3 Two-sample t-tests

4.3.1 Paired samples

Suppose we wish to compare returns on small-cap vs. large-cap stocks.
For each of n yrs we have the returns on a portfolio of small-cap stocks $(X_1, \ldots, X_n)$ and on a portfolio of large-cap stocks $(Y_1, \ldots, Y_n)$. Form $Z_i = X_i - Y_i$, $i = \overline{1, n}$.
We want to test $H_0: \mu_X = \mu_Y$ for the means of the two samples.
Assuming that X_i, Y_i are normal and independent, so then Z_i will be normal.
So, H_0 is equivalent to $H_0: \mu_Z = 0$. Now, do the usual t-test from 4.1 ① for $Z_1, \ldots, Z_n$