Random Vectors

\( \mathbf{x} \) is a \( p \times 1 \) random vector with a pdf probability density function \( f(\mathbf{x}): \mathbb{R}^p \to \mathbb{R} \). Many books write \( \mathbf{X} \) for the random vector and \( \mathbf{X} = \mathbf{x} \) for the realization of its value. 

\[
E[\mathbf{X}] = \int \mathbf{x} \, f(\mathbf{x}) \, d\mathbf{x} = \mu.
\]

Theorem: \( E[A\mathbf{x} + b] = A E[\mathbf{x}] + b \)

Covariance Matrix \( E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)'] = \text{var}(\mathbf{x}) = \Sigma \) (note the location of transpose)

Theorem: \( \Sigma = E[\mathbf{xx}'] - \mu\mu' \)

If \( \mathbf{y} \) is a random variable: covariance \( C(\mathbf{x}, \mathbf{y}) = E[(\mathbf{x} - \mu)(\mathbf{y} - \nu)' \]

Theorem: For constants \( a, A \), \( \text{var}(a'\mathbf{x}) = a'\Sigma a \), \( \text{var}(A\mathbf{x} + b) = A\Sigma A' \), \( C(\mathbf{x}, \mathbf{x}) = \Sigma \), \( C(\mathbf{x}, \mathbf{y}) = C(\mathbf{y}, \mathbf{x})' \)

Theorem: If \( \mathbf{x}, \mathbf{y} \) are independent RVs, then \( C(\mathbf{x}, \mathbf{y}) = 0 \), but not conversely.

Theorem: Let \( \mathbf{x}, \mathbf{y} \) have same dimension, then \( \text{var}(\mathbf{x} + \mathbf{y}) = \text{var}(\mathbf{x}) + \text{var}(\mathbf{y}) + C(\mathbf{x}, \mathbf{y}) + C(\mathbf{y}, \mathbf{x}) \)

**Normal Random Vectors**

The Central Limit Theorem says that if a focal random variable \( \mathbf{x} \) consists of the sum of many other independent random variables, then the focal random variable will asymptotically have a distribution that is basically of the form \( e^{-x^2} \), which we call “normal” because it is so common.

Normal random variable has pdf 

\[
f(\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{\mathbf{x} - \mu}{\sigma}\right)^2 / 2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mathbf{x} - \mu)\mathbf{1}(\mathbf{x} - \mu)}{2\sigma^2}}
\]

Denote \( \mathbf{x} \) \( p \times 1 \) normal random variable with pdf 

\[
f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)}
\]

where \( \mu \) is the mean vector and \( \Sigma \) is the covariance matrix: \( \mathbf{x} \sim \mathcal{N}_p(\mu, \Sigma) \).

Bivariate Normal \( f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}} e^{-\frac{1}{2} \left[ \begin{array}{c} x_1 - \mu_1 \\ x_2 - \mu_2 \end{array} \right]' \left[ \begin{array}{cc} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{array} \right]^{-1} \left[ \begin{array}{c} x_1 - \mu_1 \\ x_2 - \mu_2 \end{array} \right]} \)

Recall variance \( \sigma_{11} \) is also sometimes written \( \sigma_1^2 \) and by symmetry \( \sigma_{12} = \sigma_{21} \). The correlation is \( \rho_{12} = \sigma_{12} / \sqrt{\sigma_{11}\sigma_{22}} = \sigma_{12} / (\sigma_1\sigma_2) \).

Theorem: Eigenvalue of \( \Sigma^{-1} \) is reciprocal of eigenvalue of \( \Sigma \) and eigenvectors are identical.

Proof: Let \( \Sigma^{-1}\mathbf{x} = \lambda\mathbf{x} \). Then \( \mathbf{x} = \lambda\Sigma\mathbf{x} \) or \( \Sigma\mathbf{x} = (1/\lambda)\mathbf{x} \).
Contour of constant probability is ellipsoid \((\mathbf{x} - \mathbf{\mu})' \Sigma^{-1} (\mathbf{x} - \mathbf{\mu}) = c^2\) for some \(c\). This is an ellipse centered at \(\mathbf{\mu}\) and with axis that point in the directions of the eigenvectors of \(\Sigma\) with length \(c \sqrt{\lambda_i}\), that is the axes are \(\pm c \sqrt{\lambda_i} \mathbf{e}_i\) where \(\lambda_i\) and \(\mathbf{e}_i\) are the eigenvalues and eigenvectors of the covariance matrix \(\Sigma\).

Suppose that \(\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\), then eigenvalues are defined by \((1-\lambda)^2 - \rho^2 = 0\) or \(\lambda = 1 \pm \rho\). The eigenvectors are values of \((x_1, x_2)\) such that \(\pm \rho x_1 + \rho x_2 = 0\) and \(x_1^2 + x_2^2 = 1\); these are \(1/\sqrt{2}, -1/\sqrt{2}\). If the correlation \(\rho\) is positive, then the eigenvector \((1/\sqrt{2}, 1/\sqrt{2})\) is stretched to a length greater than 1, \(1 + \rho\), while the eigenvector \((1/\sqrt{2}, -1/\sqrt{2})\) is shrunk to a length less than 1, \(1 - \rho\). See the figure below.

Theorems: The moment generating function (mgf) for multivariate normal is

\[
\phi_X(t) = \mathbb{E}[e^{t' \mathbf{x}}] = e^{t' \mathbf{\mu} + \frac{1}{2} t' \Sigma t}.
\]

For \(\mathbf{x} \sim \mathcal{N}(\mathbf{\mu}, \Sigma)\) \(\Rightarrow x_i \sim \mathcal{N}(\mu_i, \sigma_{ii})\)

\(y = a' \mathbf{x} \Rightarrow y \sim \mathcal{N}(a' \mathbf{\mu}, a'a)\) and mgf \(\phi_y(t) = e^{t a' \mathbf{\mu} + \frac{1}{2} t^2 a'a} \Sigma a\)

\(x_1 | x_2 \sim \mathcal{N}(\mu_1 + \sigma_{12}/\sigma_{22}, \sigma_{11} - \sigma_{12}^2/\sigma_{22})\)

Theorem \((\mathbf{x} - \mathbf{\mu})' \Sigma^{-1} (\mathbf{x} - \mathbf{\mu}) \sim \chi^2_p\)

Proof: (note:chi-sq is the sum of the squares of independent normals). Using spectral decomposition, \((\mathbf{x} - \mathbf{\mu})' \Sigma^{-1} (\mathbf{x} - \mathbf{\mu}) = (\mathbf{x} - \mathbf{\mu})' (\mathbf{P} \Lambda \mathbf{P}')^{-1} (\mathbf{x} - \mathbf{\mu}) = (\mathbf{x} - \mathbf{\mu})' \Lambda^{-1/2} \mathbf{P}' \Lambda^{-1/2} \mathbf{P} (\mathbf{x} - \mathbf{\mu})\). From above \(\Lambda^{-1/2} \mathbf{P}' (\mathbf{x} - \mathbf{\mu}) \sim \mathcal{N}(0, \Lambda^{-1/2} \mathbf{P}' \Sigma \mathbf{P} \Lambda^{-1/2} = \mathcal{N}(0, \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = \mathcal{N}(0, I))\), so quadratic form is the sum of independent squared normal random variables. QED

Normal data matrix \(X = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{M} \\ \mathbf{x}_n' \end{bmatrix}\) where \(\mathbf{x}_i\) is iid \(\mathcal{N}_p(\mathbf{\mu}, \Sigma)\). This is a \(n \times p\) matrix of random variables.

Each row is independent of other rows and identically distributed.

\[
P(X) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp(-\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{\mu})' \Sigma^{-1} (\mathbf{x}_i - \mathbf{\mu}) / 2)
\]

\[
= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left(- \text{tr}\left(\Sigma^{-1} (\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \mathbf{\mu})(\bar{\mathbf{x}} - \mathbf{\mu})') / 2\right)\right)
\]

Note: \(x'Ax = \text{tr}(x'Ax) = \text{tr}(Ax'x)\)

**Aside: On Union Intersection Tests**

\(\mathbf{x} \sim \mathcal{N}(\mathbf{\mu}, I) \Rightarrow y = a' \mathbf{x} \sim \mathcal{N}(a' \mathbf{\mu}, a'a)\)

\(H_0: \mathbf{\mu} = 0 \iff y = a' \mathbf{x} \sim \mathcal{N}(0, a'a)\) for all \(a\).

\(H_{oa}: a' \mathbf{\mu} = 0\)
$H_0 = \cap H_{0a}$ note: if you find one $a$ that violates $H_{0a}$ then $H_0$ cannot be true.

Let’s test $H_{0a}$ using $z_a = y_a / \sqrt{a'a}$. The rejection region is $R_a = \{ z_a | z_a^2 > c^2 \}$. What about $H_a$?
$R = \cup R_a$. So $H_0$ is accepted if and only if $z_a^2 < c^2$ for all $a$. The worst case scenario is $\max_a z_a^2$. So, if $\max_a z_a^2 < c^2$ then this will be true for all $a$. Suppose that we have independent draws of a random vector. Let $x_j$ be the jth draw. Define $y = a'x_j$. Then we know that $\bar{y} = a'\bar{x}$ and $s_y^2 = a'Sa$. Compute $t^2 = \frac{n(a'(\bar{x} - \mu))^2}{a'Sa}$. Following the Union Intersection test procedure we would like to find the value of $a$ that is the worst case scenario. The Cauchy-Schwartz inequality helps here: $(x'y)^2 \leq (x'x)(y'y)$ (this is a consequence of $x'y = ||x|| ||y|| \cos \theta$): $(a' (\bar{x} - \mu))^2 \leq (a'Sa)((\bar{x} - \mu)'S^{-1}(\bar{x} - \mu))$ and can only “=” if $a = S^{-1}(\bar{x} - \mu)$.

Taking this worst case scenario, the $\max_a t^2 = n (\bar{x} - \mu)'S^{-1}(\bar{x} - \mu)$.

Theorem: The interval $a'\bar{x} \pm \sqrt{\frac{p(n-1)}{n-p} \frac{a'Sa}{F_{p,n-p} (\alpha)}}$ will contain $a'\mu$ a fraction $1-\alpha \%$ of the time, simultaneously for all possible $a$.

Comparison of Traditional and Simultaneous Confidence Intervals

Suppose that you had $H_{0i}: \mu_i = 0$ for $i = 1, 2, \ldots, p$. If you ignore the fact that there are several simultaneous test, you would do this one variable at a time, computing confidence intervals:

$\bar{x}_i \pm \sqrt{s_{ii}/n} \cdot t_{n-1}(\alpha/2)$.

As we have seen before, the confidence for these as a whole is not $1-\alpha \%$, but rather $(1-\alpha)^p$: for 6 variables $0.95^6 = 0.75$. Hence the rectangular region sketched out by these intervals is really only a $75\%$ confidence region. If you had 13 variables, then this region will capture the truth in all dimensions only $50\%$ of the time. We have a false sense of high accuracy.

The above Union Intersection test would sequentially set $a' = (0, 0, 0, 1, 0, 0, \ldots)$ where the 1 is in the $i$th entry and then calculate the intervals

$\bar{x}_i \pm \sqrt{s_{ii}/n} \cdot \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)}$.

These simultaneous intervals will be much wider than the above, but we can then say that there is a $95\%$ confidence that all variables will be covered by the combined rectangle. These intervals are the “shadows of the $95\%$ confidence ellipse in a $p$-dimensional space.

How much wider are these simultaneous intervals? It depends on $n$ and $p$. As you can see the intervals are much wider, making it very difficult to say with high confidence, “All elements of my theory are true.”

<table>
<thead>
<tr>
<th>$n$</th>
<th>$t_{n-1}(0.025)$</th>
<th>$p=4$</th>
<th>$p=10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>2.145</td>
<td>4.14</td>
<td>11.52</td>
</tr>
<tr>
<td>25</td>
<td>2.064</td>
<td>3.60</td>
<td>6.39</td>
</tr>
<tr>
<td>50</td>
<td>2.010</td>
<td>3.31</td>
<td>5.05</td>
</tr>
<tr>
<td>100</td>
<td>1.970</td>
<td>3.19</td>
<td>4.61</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.960</td>
<td>3.08</td>
<td>4.28</td>
</tr>
</tbody>
</table>
**Generalization of t-test to T^2-test**

Neither the traditional nor the simultaneous interval tests take into account that the variables may be correlated with one another. The 95% confidence region should not be a rectangle, but rather an ellipse. How should we handle this? This is not that complicated.

In the single normal variable case, we test using \( t \equiv \frac{\bar{x} - \mu}{s / \sqrt{n}} \). When we have several variables that we want to combined without having \(-\)'s canceling \(+\)'s, we use the Hotelling T^2-distribution of the variable \( t^2 \equiv \frac{\bar{x} - \mu}{s^2 / n} \). For p-variate normal vector case, the equivalent statistic is

\[
T^2_{p,n-1} = \frac{(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu)}{1/n} = n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu).
\]

This statistic has normals squared on the top (the x'x terms) and normals squared in the bottom (since S is composed from squared normals). That is, it is the ratio of \( \chi^2 \)'s and hence has the F-distribution:

\[
T^2_{p,n-1} \sim \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha).
\]

Thus if we had a null hypothesis that the p-variate variable \( x \) had mean \( \mu \), then we would construct the above T^2 statistic and see if it exceeded the critical value found in an F-distribution table. This will tells us whether our theoretical value \( \mu \) is covered by the confidence ellipsoid 1-\( \alpha \)% of the time in repeated samples. The 1-\( \alpha \)% confidence ellipsoid has axes determined by

\[
n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu) \leq \chi^2 = \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha).
\]

That is they are determined by starting at \( \bar{x} \) and going \( \pm \sqrt{\lambda_i} \ c / \sqrt{n} = \pm \sqrt{\lambda_i} \ \sqrt{\frac{(n-1)p}{n(n-p)}} F_{p,n-p}(\alpha) \) units along the eigenvectors \( e_i \). This is better than doing p separate t-tests of each variable, since it uses all the information in S, including the fact that some variables are highly correlated.

In summary, do not claim when you study p variables and all of them fit your theory that you are 95% confident in your theory. Apparent confidence is not real confidence. On the other hand, even if one-at-a-time you cannot reject the null, you still may be able to with 95% confidence state, “There are some elements of this theory that must be true, I just cannot tell you which ones.” In the above graph, the true \( \mu \) is inside the apparent 95% confidence interval, so you apparently cannot reject any element \( \mu_i \), but \( \mu \) is outside the 95% confidence ellipsoid, so you reject \( \mu \) as a whole.