

Direct Calculation of Mean and Variance of Δ .

Let $\delta_1, \dots, \delta_N$ denote random variables and define the variance statistic by

$$\Delta = \frac{1}{N} \sum_{i=1}^N \left(\delta_i - \frac{1}{N} \sum_{j=1}^N \delta_j \right)^2.$$

We will first give an elementary derivation for the mean and variance of Δ assuming that the δ_i are independent normal variables with mean μ_i and common standard deviation σ . This will involve lots of algebraic manipulation, and afterwards a more direct derivation will be given, using matrices and moment generating functions, without assuming that the δ_i are independent.

1. Algebraic Derivation with Independent Hypothesis.

To complete the calculation efficiently and avoid error, rearrange Δ as

$$\begin{aligned} \Delta &= \frac{1}{N} \sum_{i=1}^N \left(\sum_{j=1}^N \frac{\delta_i - \delta_j}{N} \right)^2 \\ &= \frac{1}{N^3} \sum_{i=1}^N \sum_{j,k} (\delta_i - \delta_j)(\delta_i - \delta_k) \\ &= \frac{1}{N^3} \sum_{i,j,k} (\delta_i - \delta_j)(\delta_i - \delta_k). \end{aligned}$$

Observing the pattern, it is not difficult to derive

$$\Delta^2 = \frac{1}{N^6} \sum_{i,j,k,l,m,n} (\delta_i - \delta_j)(\delta_i - \delta_k)(\delta_l - \delta_m)(\delta_l - \delta_n).$$

We expand these expressions noting that each is a sum of power products of the δ_i .

$$N^3 \Delta = N^2 \sum_i \delta_i^2 - N \sum_{i,j} \delta_i \delta_j,$$

therefore

$$N^3 \Delta = (N^2 - N) \sum_i \delta_i^2 - N \sum_{i \neq j} \delta_i \delta_j.$$

Similarly Δ^2 is expanded, the result is

$$N^6 \Delta^2 = N^4 \sum_{i,j} \delta_i^2 \delta_j^2 - 2N^3 \sum_{i,j,k} \delta_i^2 \delta_j \delta_k + N^2 \sum_{i,j,k,l} \delta_i \delta_j \delta_k \delta_l,$$

and so

$$\begin{aligned} N^6 \Delta^2 &= N^2 \sum_{i,j,k,l \text{ not equal}} \delta_i \delta_j \delta_k \delta_l + (-2N^3 + 6N^2) \sum_{i,j,k \text{ not equal}} \delta_i^2 \delta_j \delta_k \\ &\quad + (N^4 - 2N^3 + 3N^2) \sum_{i \neq j} \delta_i^2 \delta_j^2 \\ &\quad + (-4N^3 + 4N^2) \sum_{i \neq j} \delta_i^3 \delta_j + (N^4 - 2N^3 + N^2) \sum_i \delta_i^4. \end{aligned}$$

To take the expectation of these expressions, it is necessary to find the expectation of products of the δ . Since δ with different subscripts are independent variables, this comes down to knowing the expectation of the powers of each δ_i from 1 to 4.

To begin with, recall that if Z is normally distributed with mean 0 and standard deviation 1, then

$$EZ = 0 \quad EZ^2 = 1 \quad EZ^3 = 0 \quad EZ^4 = 3.$$

We apply this to $Z = (\delta_i - \mu_i)/\sigma$

Obviously $E\delta_i = \mu_i$. Since $EZ^2 = 1$ we have

$$\begin{aligned} E\left(\frac{\delta_i - \mu_i}{\sigma}\right)^2 &= 1 \\ E(\delta_i^2 - 2\delta_i\mu_i + \mu_i^2) &= \sigma^2 \\ E\delta_i^2 &= \sigma^2 + \mu_i^2 \end{aligned}$$

Next, since $EZ^3 = 0$,

$$\begin{aligned} E\left(\frac{\delta_i - \mu_i}{\sigma}\right)^3 &= 0 \\ E(\delta_i^3 - 3\delta_i^2\mu_i + 3\delta_i\mu_i^2 - \mu_i^3) &= 0 \end{aligned}$$

Solving for $E\delta_i^3$ and expanding,

$$\begin{aligned} E\delta_i^3 &= 3E\delta_i^2\mu_i - 3E\delta_i\mu_i^2 + \mu_i^3 \\ &= 3(\sigma^2 + \mu_i^2)\mu_i - 3\mu_i\mu_i^2 + \mu_i^3 \\ &= 3\sigma^2\mu_i + \mu_i^3 \end{aligned}$$

Finally, from $EZ^4 = 3$,

$$\begin{aligned} E\left(\frac{\delta_i - \mu_i}{\sigma}\right)^4 &= 3 \\ E(\delta_i^4 - 4\delta_i^3\mu_i + 6\delta_i^2\mu_i^2 - 4\delta_i\mu_i^3 + \mu_i^4) &= 3\sigma^4. \end{aligned}$$

Solving for $E\delta_i^4$ and expanding,

$$\begin{aligned} E\delta_i^4 &= 3\sigma^4 + 4E\delta_i^3\mu_i - 6E\delta_i^2\mu_i^2 + 4E\delta_i\mu_i^3 - \mu_i^4 \\ &= 3\sigma^4 + 4(3\sigma^2\mu_i + \mu_i^3)\mu_i - 6(\sigma^2 + \mu_i^2)\mu_i^2 + 4\mu_i\mu_i^3 - \mu_i^4 \\ &= 3\sigma^4 + 6\sigma^2\mu_i^2 + \mu_i^4. \end{aligned}$$

We use these expressions in the expanded form for Δ and Δ^2 , noting that if $i \neq j$ then δ_i and δ_j are independent. Thus

$$\begin{aligned} N^3 E\Delta &= (N^2 - N) \sum_i E\delta_i^2 - N \sum_{i \neq j} E\delta_i E\delta_j \\ &= (N^2 - N) \sum_i (\sigma^2 + \mu_i^2) - N \sum_{i \neq j} \mu_i \mu_j \\ &= (N^2 - N)N\sigma^2 + ((N^2 - N) \sum_i \mu_i^2 - N \sum_{i \neq j} \mu_j \mu_i) \end{aligned}$$

Now note that the sums in parentheses correspond to the value of $N^3\Delta$ with $\delta_i = \mu_i$, we will denote this by Δ_0 . Finally we have

$$E\Delta = \frac{N-1}{N}\sigma^2 + \Delta_0.$$

While there is a more straightforward way of deriving $E\Delta$, finding $E\Delta^2$ is more cumbersome. The method being used, however, is not much more difficult when applied to $E\Delta^2$. Taking expectations of the expression,

$$\begin{aligned} N^6 E\Delta^2 &= N^2 \sum_{i,j,k,l \text{ not equal}} E\delta_i E\delta_j E\delta_k E\delta_l + (-2N^3 + 6N^2) \sum_{i,j,k \text{ not equal}} E\delta_i^2 E\delta_j E\delta_k \\ &\quad + (N^4 - 2N^3 + 3N^2) \sum_{i \neq j} E\delta_i^2 E\delta_j^2 \\ &\quad + (-4N^3 + 4N^2) \sum_{i \neq j} E\delta_i^3 E\delta_j + (N^4 - 2N^3 + N^2) \sum_i E\delta_i^4 \end{aligned}$$

Substituting the expressions for the expectations,

$$\begin{aligned}
N^6 E\Delta^2 &= N^2 \sum_{i,j,k,l \text{ not equal}} \mu_i \mu_j \mu_k \mu_l + (-2N^3 + 6N^2) \sum_{i,j,k \text{ not equal}} (\sigma^2 + \mu_i^2) \mu_j \mu_k \\
&\quad + (N^4 - 2N^3 + 3N^2) \sum_{i \neq j} (\sigma^2 + \mu_i^2)(\sigma^2 + \mu_j^2) + (-4N^3 + 4N^2) \sum_{i \neq j} (3\sigma^2 \mu_i + \mu_i^3) \mu_j \\
&\quad + (N^4 - 2N^3 + N^2) \sum_i (3\sigma^4 + 6\sigma^2 \mu_i^2 + \mu_i^4)
\end{aligned}$$

Note that all of the terms that do not involve σ group together to form $N^6 \Delta_0^2$ similar to the previous case. One is left with

$$\begin{aligned}
N^6 (E\Delta^2 - \Delta_0^2) &= (-2N^3 + 6N^2)(N-2) \sum_{i \neq j} \sigma^2 \mu_i \mu_j \\
&\quad + (N^4 - 2N^3 + 3N^2) \sum_{i \neq j} (\sigma^4 + \sigma^2 \mu_i^2 + \sigma^2 \mu_j^2) + (-4N^3 + 4N^2) \sum_{i \neq j} 3\sigma^2 \mu_i \mu_j \\
&\quad + (N^4 - 2N^3 + N^2) \sum_i (3\sigma^4 + 6\sigma^2 \mu_i^2)
\end{aligned}$$

Expanding the sums of constants,

$$\begin{aligned}
N^6 (E\Delta^2 - \Delta_0^2) &= (-2N^3 + 6N^2)(N-2) \sigma^2 \sum_{i \neq j} \mu_i \mu_j \\
&\quad + (N^4 - 2N^3 + 3N^2)(N(N-1)\sigma^4 + 2(N-1)\sigma^2 \sum_i \mu_i^2) + (-4N^3 + 4N^2) 3\sigma^2 \sum_{i \neq j} \mu_i \mu_j \\
&\quad + (N^4 - 2N^3 + N^2)(N 3\sigma^4 + 6\sigma^2 \sum_i \mu_i^2)
\end{aligned}$$

Now collecting similar terms together

$$\begin{aligned}
N^6 (E\Delta^2 - \Delta_0^2) &= ((N^4 - 2N^3 + 3N^2)N(N-1) + (N^4 - 2N^3 + N^2)N 3) \sigma^4 \\
&\quad + ((-2N^3 + 6N^2)(N-2) + (-4N^3 + 4N^2) 3) \sigma^2 \sum_{i \neq j} \mu_i \mu_j \\
&\quad + ((N^4 - 2N^3 + 3N^2) 2(N-1) + (N^4 - 2N^3 + N^2) 6) \sigma^2 \sum_i \mu_i^2
\end{aligned}$$

And simplifying the coefficients,

$$\begin{aligned}
N^6 (E\Delta^2 - \Delta_0^2) &= N^4(N-1)(N+1)\sigma^4 \\
&\quad - 2N^3(N+1)\sigma^2 \sum_{i \neq j} \mu_i \mu_j \\
&\quad + 2N^3(N-1)(N+1)\sigma^2 \sum_i \mu_i^2.
\end{aligned}$$

Rewriting this,

$$N^6 (E\Delta^2 - \Delta_0^2) = N^4(N-1)(N+1)\sigma^4 + 2N^2(N+1)\sigma^2((N^2 - N) \sum_i \mu_i^2 - N \sum_{i \neq j} \mu_i \mu_j)$$

Again the sum in parentheses, is just $N^3 \Delta_0$ so

$$E\Delta^2 = \frac{(N-1)(N+1)}{N^2} \sigma^4 + 2 \frac{N+1}{N} \sigma^2 \Delta_0.$$

The reader following this far should have no difficulty seeing that

$$Var(\Delta) = E\Delta^2 - (E\Delta)^2 = 2\frac{N-1}{N^2}\sigma^4 + \frac{4}{N}\sigma^2\Delta_0.$$

2. Moment Generating Function Derivation without Independent Hypothesis.

Now assume that the vector $\delta = (\delta_1, \dots, \delta_N)$ has a multivariate normal distribution with mean $\mu = (\mu_1, \dots, \mu_N)$ and covariance matrix \mathbf{V} whose i, j entry is $E(\delta_i - \mu_i)(\delta_j - \mu_j)$. The density function is

$$\frac{1}{(2\pi)^{n/2}|\mathbf{V}|^{1/2}} \exp\left(-\frac{(\delta - \mu)' \mathbf{V}^{-1}(\delta - \mu)}{2}\right)$$

(the prime denotes transpose) and the moment generating function is

$$M(\mathbf{t}) = \exp(\mathbf{t}'\mu + \mathbf{t}'\mathbf{V}\mathbf{t}/2).$$

Now consider the quadratic form $\delta' \mathbf{A} \delta$ where \mathbf{A} is a real symmetric matrix. The moment generating function is $E \exp(t\delta' \mathbf{A} \delta)$ which can be written

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}|\mathbf{V}|^{1/2}} \exp\left(t\delta' \mathbf{A} \delta - \frac{(\delta - \mu)' \mathbf{V}^{-1}(\delta - \mu)}{2}\right) d\delta_1 \cdots d\delta_N \\ &= |\mathbf{V}| |\mathbf{V}^{-1} - 2t\mathbf{A}|^{-1/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(2t\mu' \mathbf{A} \delta - t\mu' \mathbf{A} \mu) \\ &\quad \frac{1}{(2\pi)^{n/2}|\mathbf{V}^{-1} - 2t\mathbf{A}|^{-1/2}} \exp\left(-\frac{(\delta - \mu)' (\mathbf{V}^{-1} - 2t\mathbf{A})(\delta - \mu)}{2}\right) d\delta_1 \cdots d\delta_N \\ &= |\mathbf{V}| |\mathbf{V}^{-1} - 2t\mathbf{A}|^{-1/2} \exp(-t\mu' \mathbf{A} \mu) E \exp(2t\mu' \mathbf{A} \delta). \end{aligned}$$

where the expectation is over a random variable with mean μ and covariance matrix $\mathbf{V}^{-1} - 2t\mathbf{A}$. Therefore we have

$$M(t) = |I - 2t\mathbf{V}\mathbf{A}|^{-1/2} \exp(t\mu' \mathbf{A} \mu + 2t^2 \mu' \mathbf{A} \mathbf{V} \mathbf{A} \mu)$$

where $\mathbf{W}^{-1} = \mathbf{V}^{-1} - 2t\mathbf{A}$. Note that $M(0) = 1$ as it must. The mean and variance are given by $E\Delta = M'(0)$ and $Var\Delta = M''(0) - M'(0)^2$. Suppose $\mathbf{V}\mathbf{A}$ is diagonalized by \mathbf{L} with eigenvalues $\lambda_1, \dots, \lambda_N$, then $|I - 2t\mathbf{V}\mathbf{A}| = |I - 2t\mathbf{L}'\mathbf{V}\mathbf{A}\mathbf{L}| = \prod (1 - 2t\lambda_i)$. Take the logarithmic derivative of $M(t)$,

$$\frac{M'(t)}{M(t)} = \sum \frac{\lambda_i}{1 - 2t\lambda_i} + \mu' \mathbf{A} \mu + 4t\mu' \mathbf{A} \mathbf{V} \mathbf{A} \mu + 2t^2 \mu' \mathbf{A} \frac{d\mathbf{W}}{dt} \mathbf{A} \mu.$$

Call this $S(t)$ and take its derivative with respect to t . Substituting in $t = 0$ (note $\mathbf{W}(0) = \mathbf{V}(0)$) and ignoring the terms with t^2 ,

$$S'(0) = \sum 2\lambda_i^2 + 4\mu' \mathbf{A} \mathbf{V} \mathbf{A} \mu.$$

The final simplification to note is that $\sum \lambda_i = tr(\mathbf{V}\mathbf{A})$ and $\sum \lambda_i^2 = tr(\mathbf{V}\mathbf{A})^2$. Therefore,

$$\begin{aligned} E\Delta &= tr(\mathbf{V}\mathbf{A}) + \mu' \mathbf{A} \mu \\ Var\Delta &= 2tr(\mathbf{V}\mathbf{A})^2 + 4\mu' \mathbf{A} \mathbf{V} \mathbf{A} \mu. \end{aligned}$$

Let us confirm that this reduces to the equations for a noncentral χ^2 distribution when $\mathbf{V} = \sigma^2 I$ and $\mathbf{A} = \frac{1}{N}(I - \frac{1}{N}\mathbf{1})$, which is the case considered in the first part. In any case $\mu' \mathbf{A} \mu = \Delta_0$. And $tr(\mathbf{V}\mathbf{A}) = \sigma^2 tr \frac{1}{N}(I - \frac{1}{N}\mathbf{1}) = \sigma^2 \frac{N-1}{N}$, and this agrees with $E\Delta$. Continuing, $(\mathbf{V}\mathbf{A})^2 = \sigma^4 \mathbf{A}^2 = \frac{\sigma^4}{N} \mathbf{A}$ and so the trace is $\sigma^2 \frac{N-1}{N}$. In the second term note that $\mathbf{A}\mathbf{V}\mathbf{A} = \sigma^2 \mathbf{A}^2 = \frac{\sigma^2}{N} \mathbf{A}$ so multiplying by μ on both sides simply gives $4\frac{\sigma^2}{N}\Delta_0$ establishing agreement with $Var\Delta$.

It may be useful to have a derivation of $M(t)$ that finds all of the coefficients. To show how this can be done, expand the factors in a Taylor series. To begin with, recall

$$(1 - x)^{-1/2} = 1 + \frac{1}{2}x + \frac{3}{8}x^2 + o(x^3).$$

Therefore

$$(1 - 2t\lambda_i)^{-1/2} = 1 + t\lambda_i + \frac{3}{2}t^2\lambda_i^2 + o(t^3)$$

and

$$\prod_i (1 - 2t\lambda_i)^{-1/2} = 1 + t \sum_i \lambda_i + t^2 \left(\frac{3}{2} \sum_i \lambda_i^2 + \sum_{i < j} \lambda_i \lambda_j \right) + o(t^3).$$

Denote these coefficients as $\beta_0 = 1$, $\beta_1 = \sum_i \lambda_i = \text{tr} \mathbf{VA}$ and $\beta_2 = \left(\frac{3}{2} \sum_i \lambda_i^2 + \sum_{i < j} \lambda_i \lambda_j \right)$. The expression for β_2 can be rewritten as

$$\begin{aligned} \beta_2 &= \frac{1}{2} \left(\sum \lambda_i^2 + \left(\sum \lambda_i \right)^2 \right) \\ &= \frac{1}{2} \left(\text{tr}(\mathbf{VA})^2 + (\text{tr} \mathbf{VA})^2 \right). \end{aligned}$$

Next note that with $\mathbf{W} = (\mathbf{V}^{-1} - 2t\mathbf{A})^{-1}$ then for t sufficiently small,

$$\mathbf{W} = \mathbf{V}(\mathbf{I} - 2t\mathbf{VA})^{-1} = \mathbf{V} \sum_{n=0}^{\infty} (2t\mathbf{VA})^n.$$

So that

$$t\mu' \mathbf{A} \mu + 2t^2 \mu' \mathbf{A} \mathbf{W} \mathbf{A} \mu = \mu' \left(\sum_{n=0}^{\infty} (2\mathbf{VA})^n \mathbf{A} t^{n+1} \right) \mu.$$

Call $\alpha_0 = 0$, $\alpha_1 = \mu' \mathbf{A} \mu$, and $\alpha_2 = 2\mu' \mathbf{VA}^2 \mu$ so that the expression is $\alpha_0 + t\alpha_1 + t^2\alpha_2 + o(t^3)$. Finally from

$$e^x = 1 + x + \frac{1}{2}x^2 + o(x^3)$$

we find

$$\exp\left(\sum_{i>0} \alpha_i t^i\right) = 1 + \alpha_1 t + \left(\frac{1}{2}\alpha_1^2 + \alpha_2\right)t^2 + o(t^3).$$

Therefore,

$$M(t) = 1 + (\beta_1 + \alpha_1)t + \left(\alpha_1\beta_1 + \beta_2 + \frac{1}{2}\alpha_1^2 + \alpha_2\right)t^2 + o(t^3).$$

And so

$$\begin{aligned} M'(0) &= \alpha_1 + \beta_1 = \text{tr} \mathbf{VA} + \mu' \mathbf{A} \mu \\ M''(0) &= 2(\alpha_1\beta_1 + \beta_2 + \frac{1}{2}\alpha_1^2 + \alpha_2) \\ &= 2(\text{tr} \mathbf{VA})\mu' \mathbf{A} \mu \text{tr}(\mathbf{VA})^2 + (\text{tr} \mathbf{VA})^2 + (\mu' \mathbf{A} \mu)^2 + 4\mu' \mathbf{VA}^2 \mu, \end{aligned}$$

which lead to the same expressions derived for $E\Delta$ and $\text{Var}\Delta$.

3. Handling Missing Data.

To model missing data, let us appeal to a general lemma. Let X_1, \dots, X_d be d random variables, not necessarily identically distributed or independent. Let $p_1 + \dots + p_d = 1$ define a random variable S on $1, \dots, d$ and define $X = X_S$. If $M_s(t)$ is the moment generating function of X_s then $\sum p_s M_s(t)$ is the moment generating function of X . The proof is simple, namely

$$E \exp(tX) = \sum p_s E \exp(tX_s).$$

If we let EX_s and $VarX_s$ denote the expectation and variance of the random variable X_s , then we can see that

$$EX = E_S(EX_s)$$

and

$$VarX = E_S(VarX_s) + Var_S(EX_s)$$

where E_s and Var_s denote expectation and variance of the expression as a function of S . To prove this, note first $EX = M'(0) = \sum p_s M'_s(0) = \sum p_s EX_s = E_S(EX_s)$. Next, $VarX = EX^2 - (EX)^2 = M''(0) - M'(0)^2$. But $M''(0) = \sum p_s M''_s(0) = \sum p_s EX_s^2$ so $EX^2 = E_S(EX_s)^2$. So

$$\begin{aligned} VarX &= E_S(EX_s)^2 - (E_S(EX_s))^2 \\ &= E_S((EX_s)^2 - (EX_s)^2) + (E_S(EX_s)^2 - (E_S(EX_s))^2) \\ &= E_S(VarX_s) + Var_S(EX_s) \end{aligned}$$

We can apply this to the problem of missing data by considering the statistic Δ^s to be computed for a subset $s \subset \{1, \dots, d\}$ of components. In a particular data set, the observed subsets of s will vary according to when components of data are missing, which can be modelled as a distribution of sets. We therefore need to consider the statistic $\Delta = \Delta^S$ where S follows observed distribution of sets of non-missing data.