Statistics 5041
8. Means and Variances

Gary W. Oehlert
School of Statistics
313B Ford Hall
612-625-1557
gary@stat.umn.edu

Consider a set of data: $x_1, x_2, \ldots, x_n$. We all know about sample means and sample standard deviations:

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ s &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2} \end{aligned}$$

Our book has done something *evil*. They define s as

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

Later, in some places, they redefine with $n-1$.
There are some reasons to use n (it's maximum likelihood for the normal), and some to use $n-1$ (it's unbiased).

I'll use $n-1$.

It's easy to go back and forth; just make sure you know which one you are working with.
To repeat:

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ s &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2} \end{aligned}$$

The sample variance is just s^2.
Univariate Standardization. Subtract out the mean and divide by the standard deviation; the result has mean 0 and standard deviation 1. Often represented by z.

$$z_i = \frac{x_i - \overline{x}}{s}$$

Multivariate Data. Let $\mathbf{X}$ be an $n \times p$ matrix of data: n cases with p variables.

$$\mathbf{X} = \begin{bmatrix} \vec{\mathbf{X}}_1' \\ \vec{\mathbf{X}}_2' \\ \vdots \\ \vec{\mathbf{X}}_n' \end{bmatrix} = \left[\check{\mathbf{X}}_1, \check{\mathbf{X}}_2, \ldots, \check{\mathbf{X}}_p\right]$$

$\vec{\mathbf{X}}_i$ is the data (column) vector for the ith case; its transpose is the ith row of $\mathbf{X}$. x_{ij} is the element in row i, column j.
Note: an individual data case is a column vector, but it appears in the data matrix as a row.
Let $\overline{x}_j$ be the mean of $\check{\mathbf{X}}_j$, the jth column of $\mathbf{X}$.

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Let $\overline{\mathbf{x}}$ be the column vector of column means

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix}$$

Note again: $\overline{\mathbf{x}}$ is a column vector.

$$\overline{\mathbf{x}} = \mathbf{X}'\mathbf{1}_n \ \frac{1}{n}$$

```
Cmd> readdata("",dradius,radius,\
dhumerus,humerus,dulna,ulna)
Read from file "/cdrom/T1-8.DAT"
Column 1 saved as REAL vector dradius
Column 2 saved as REAL vector radius
Column 3 saved as REAL vector dhumerus
Column 4 saved as REAL vector humerus
Column 5 saved as REAL vector dulna
Column 6 saved as REAL vector ulna

Cmd> X <- hconcat(dradius,radius,dulna,ulna)

Cmd> describe(X,mean:T)
(1)    0.8438   0.81832    0.7044   0.69384

Cmd> X'%*%rep(1,25)/25
(1,1)       0.8438
(2,1)      0.81832
(3,1)       0.7044
(4,1)      0.69384

Cmd> tabs(X,mean:T)
(1)    0.8438   0.81832    0.7044   0.69384
```

$\mathbf{S}$ is the *sample variance matrix* of $\mathbf{X}$.

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \dots & s_{1p} \\ s_{21} & s_{22} & s_{23} & \dots & s_{2p} \\ s_{31} & s_{32} & s_{33} & \dots & s_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \dots & s_{pp} \end{bmatrix}$$

s_{jj} is the sample variance of $\check{\mathbf{X}}_j$, the jth variable, and s_{ij} is the sample *covariance* of $\check{\mathbf{X}}_j$ and $\check{\mathbf{X}}_k$, the jth and kth variables.

$$s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$$

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$$

Note that $s_{jk} = s_{kj}$, so $\mathbf{S}$ is symmetric.
Let $\mathbf{D}$ be the matrix of deviations from variable means: $d_{ij} = x_{ij} - \overline{x}_j$ or

$$\mathbf{D} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}' = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}$$

Then

$$\begin{aligned}\mathbf{S} &= \frac{1}{n-1}\mathbf{D}'\mathbf{D} = \frac{1}{n-1}\sum_{i=1}^{n}\vec{\mathbf{D}}_i'\vec{\mathbf{D}}_i \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\vec{\mathbf{X}}_i - \overline{\mathbf{x}})(\vec{\mathbf{X}}_i - \overline{\mathbf{x}})'\end{aligned}$$

$\mathbf{D}'\mathbf{D}$ is sometimes called the sums of squares and crossproducts matrix (SSCP) of the deviations.

```
Cmd> setoptions(format:"f8.4")

Cmd> covar(X)
component: n
(1)  25.0000
component: mean
(1,1)   0.8438   0.8183   0.7044   0.6938
component: covariance
(1,1)   0.0130   0.0104   0.0091   0.0080
(2,1)   0.0104   0.0114   0.0085   0.0089
(3,1)   0.0091   0.0085   0.0116   0.0081
(4,1)   0.0080   0.0089   0.0081   0.0106

Cmd> tabs(X,covar:T)
(1,1)   0.0130   0.0104   0.0091   0.0080
(2,1)   0.0104   0.0114   0.0085   0.0089
(3,1)   0.0091   0.0085   0.0116   0.0081
(4,1)   0.0080   0.0089   0.0081   0.0106

Cmd> I <- dmat(25,1);one <- rep(1,25);\
D <- (I-one*one'/25)%*%X

Cmd> sum(D)
```

```
(1,1)   0.0000   0.0000   0.0000   0.0000

Cmd> D'%*%D/24
(1,1)   0.0130   0.0104   0.0091   0.0080
(2,1)   0.0104   0.0114   0.0085   0.0089
(3,1)   0.0091   0.0085   0.0116   0.0081
(4,1)   0.0080   0.0089   0.0081   0.0106
```

The *correlation* between variables j and k is

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} = \frac{< \check{\mathbf{D}}_j, \check{\mathbf{D}}_k >}{||\check{\mathbf{D}}_j||\ ||\check{\mathbf{D}}_k||} = \cos(\theta)$$

where θ is the angle between $\check{\mathbf{D}}_j$ and $\check{\mathbf{D}}_k$.
$\mathbf{R}$ is the *sample correlation matrix* of $\mathbf{X}$.

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1p} \\ r_{21} & r_{22} & r_{23} & \dots & r_{2p} \\ r_{31} & r_{32} & r_{33} & \dots & r_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \dots & r_{pp} \end{bmatrix}$$

Let $\mathbf{B}$ be the $p \times p$ diagonal matrix with $1/\sqrt{s_{jj}}$ on the diagonal. Then

$$\mathbf{R} = \mathbf{BSB}$$

Also, $\mathbf{Y}$ with

$$\mathbf{Y} = (\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}')\mathbf{B} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{XB}$$

is the matrix of data after univariate standardization.

```
Cmd>Cmd> cor(X)
(1,1)   1.0000   0.8518   0.7437   0.6779
(2,1)   0.8518   1.0000   0.7422   0.8098
(3,1)   0.7437   0.7422   1.0000   0.7289
(4,1)   0.6779   0.8098   0.7289   1.0000

Cmd> S <- tabs(X,covar:T)

Cmd> B <- dmat(diag(S)^-.5)

Cmd> B %*% S %*% B
(1,1)   1.0000   0.8518   0.7437   0.6779
(2,1)   0.8518   1.0000   0.7422   0.8098
(3,1)   0.7437   0.7422   1.0000   0.7289
(4,1)   0.6779   0.8098   0.7289   1.0000
```

```
Cmd> Y <- (X - sum(X)/25)%*%B

Cmd> tabs(Y,covar:T)
(1,1)   1.0000   0.8518   0.7437   0.6779
(2,1)   0.8518   1.0000   0.7422   0.8098
(3,1)   0.7437   0.7422   1.0000   0.7289
(4,1)   0.6779   0.8098   0.7289   1.0000
```

Let $\mathbf{C}$ be a $q \times p$ matrix. Consider using $\mathbf{C}$ to make a linear transformation of our variables

$$\mathbf{Y} = \mathbf{X}\mathbf{C}'$$

or

$$\vec{\mathbf{Y}}_i = \mathbf{C}\vec{\mathbf{X}}_i$$

What are the mean and variance of the new variables?
It's pretty simple:

$$\overline{\mathbf{y}} = \mathbf{C}\overline{\mathbf{x}}$$

and

$$\mathbf{S}_y = \mathbf{C}\mathbf{S}_x\mathbf{C}'$$

$$\begin{aligned}
\overline{\mathbf{y}} &= \frac{1}{n}\mathbf{Y}'\mathbf{1}_n \\
&= \frac{1}{n}(\mathbf{X}\mathbf{C}')'\mathbf{1}_n \\
&= \frac{1}{n}\mathbf{C}\mathbf{X}'\mathbf{1}_n \\
\overline{\mathbf{y}} &= \mathbf{C}\overline{\mathbf{x}}
\end{aligned}$$

$$\mathbf{D}_y = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{Y} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}\mathbf{C}' = \mathbf{D}_x\mathbf{C}'$$

So

$$\mathbf{S}_y = \frac{1}{n-1}\mathbf{D}_y'\mathbf{D}_y = \frac{1}{n-1}\mathbf{C}\mathbf{D}_x'\mathbf{D}_x\mathbf{C}' = \mathbf{C}\mathbf{S}_x\mathbf{C}'$$

Suppose we want to keep all four variables, and add the average radius and average ulna.

```
Cmd> C
(1,1)   1.0000   0.0000   0.0000   0.0000
(2,1)   0.0000   1.0000   0.0000   0.0000
(3,1)   0.0000   0.0000   1.0000   0.0000
(4,1)   0.0000   0.0000   0.0000   1.0000
(5,1)   0.5000   0.5000   0.0000   0.0000
(6,1)   0.0000   0.0000   0.5000   0.5000
```

```
Cmd> case1 <- X[1,]';case1
(1,1)   1.1030
(2,1)   1.0520
(3,1)   0.8730
(4,1)   0.8720

Cmd> C%*%case1
(1,1)   1.1030
(2,1)   1.0520
(3,1)   0.8730
(4,1)   0.8720
(5,1)   1.0775
(6,1)   0.8725

Cmd> Y <- X%*%C'

Cmd> setoptions(format:"f6.3")

Cmd> describe(Y,mean:T)
(1)  0.844  0.818  0.704  0.694  0.831  0.699

Cmd> C %*% describe(X,mean:T)
(1,1)   0.844
(2,1)   0.818
(3,1)   0.704
(4,1)   0.694
(5,1)   0.831
(6,1)   0.699

Cmd> tabs(Y,covar:T)
(1,1)  0.013  0.010  0.009  0.008  0.012  0.009
(2,1)  0.010  0.011  0.009  0.009  0.011  0.009
(3,1)  0.009  0.009  0.012  0.008  0.009  0.010
(4,1)  0.008  0.009  0.008  0.011  0.008  0.009
(5,1)  0.012  0.011  0.009  0.008  0.011  0.009
(6,1)  0.009  0.009  0.010  0.009  0.009  0.010

Cmd> C%*%S%*%C'
(1,1)  0.013  0.010  0.009  0.008  0.012  0.009
(2,1)  0.010  0.011  0.009  0.009  0.011  0.009
(3,1)  0.009  0.009  0.012  0.008  0.009  0.010
(4,1)  0.008  0.009  0.008  0.011  0.008  0.009
(5,1)  0.012  0.011  0.009  0.008  0.011  0.009
(6,1)  0.009  0.009  0.010  0.009  0.009  0.010

Cmd> Sy <- tabs(Y,covar:T)
```

```
Cmd> eig<-eigen(Sy)

Cmd> eig$values
(1)  0.057 0.005  0.003  0.001  0.000  0.000

Cmd> eig$vectors[,vector(5,6)]
(1,1)        0.221      0.34326
(2,1)        0.221      0.34326
(3,1)      0.34326       -0.221
(4,1)      0.34326       -0.221
(5,1)       -0.442     -0.68652
(6,1)     -0.68652        0.442

Cmd> Y %*% eig$vectors[,vector(5,6)]
 (1,1)   3.8852e-16  -2.7037e-16
 ... lots more zeros

Cmd> Y %*% vector(.5,.5,0,0,-1,0)
 (1,1)  -1.1102e-16
 ... lots more zeros

Cmd> Y %*% vector(0,0,.5,.5,0,-1)
 (1,1)  -5.5511e-17
 ... lots more zeros
```

Eigenvectors are not unique for repeated eigenvalues.
The *Generalized Variance* of $\mathbf{S}$ is the determinant $|\mathbf{S}|$.
Consider the ellipsoid formed by $x'\mathbf{S}^{-1}x \leq c^2$. The volume of this ellipsoid is proportional to

$$|\mathbf{S}|^{.5}c^p$$

So (square root) generalized variance more or less us how much volume of space is within Mahalanobis distance c of the center. Singular $\mathbf{S}$ would have zero thickness in some direction, and thus zero volume.

```
Cmd> det(S)
(1)   6.0542e-10

Cmd> prod(eigenvals(S))
(1)   6.0542e-10

Cmd> det(Sy)
(1)  -4.6369e-45

Cmd> prod(eigenvals(Sy))
(1)    7.288e-45
```

Random Vectors and Matrices

If the elements of a matrix or vector are random variables, then you have a random matrix or random variable. Random variables have expectations, variances, covariances, and so on, so we can compute the same kinds of summaries for random matrices.

Let $\mathbf{X}$ and $\mathbf{Y}$ be $n \times p$ random matrices. Let $\mathbf{B}$ ($m \times n$) and $\mathbf{C}$ ($p \times q$) be fixed matrices.

$$E(\mathbf{X}) = \begin{bmatrix} E(x_{11}) & E(x_{12}) & \dots & E(x_{1p}) \\ E(x_{21}) & E(x_{22}) & \dots & E(x_{2p}) \\ \dots & \dots & \dots & \dots \\ E(x_{n1}) & E(x_{n2}) & \dots & E(x_{np}) \end{bmatrix}$$

$$E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$$

$$E(\mathbf{BXC}) = \mathbf{B}E(\mathbf{X})\mathbf{C}$$

Let x be a random p-vector.

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_p) \end{bmatrix} = E(x)$$

is the (population or theoretical) mean vector.

$$\Sigma = E(x - \mu)(x - \mu)' = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}$$

is the (population or theoretical) variance matrix.

(Population) correlations are

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

Note: if x_i and x_j are independent, then $\sigma_{ij} = 0$.

The converse is not true in general, but is true for multivariate normally distributed random variables.

$\mathbf{C}$ ($q \times p$) not random.

$$E(\mathbf{C}x) = \mathbf{C}E(x) = \mathbf{C}\mu$$

$$\text{Var}(\mathbf{C}x) = \mathbf{C}\text{Var}(x)\mathbf{C}' = \mathbf{C}\Sigma\mathbf{C}'$$