Displays for Statistics 5401/8401

Lecture 8

September 23, 2005

Christopher Bingham, Instructor

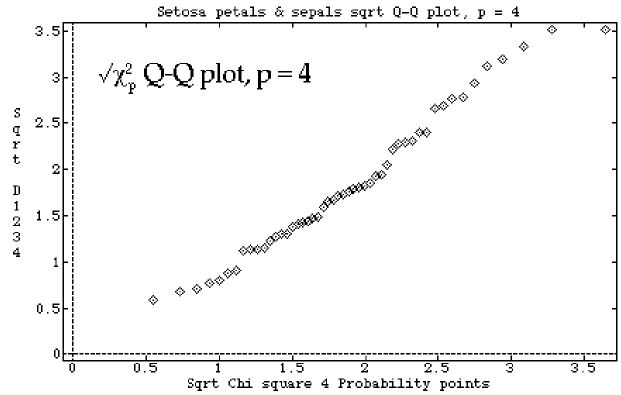
612-625-1024, kb@umn.edu
372 Ford Hall

Class Web Page

http://www.stat.umn.edu/~kb/classes/5401

Correlation normality test based on $\sqrt{\chi^2}$

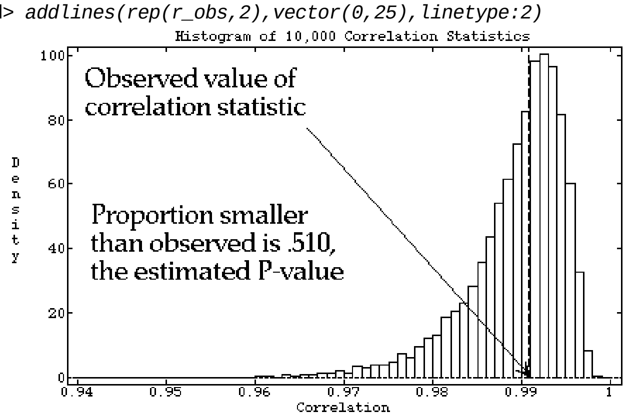
```
Cmd> d1234 <- distcomp(setosa)#4 variable dist
Cmd> p <- ncols(setosa)
Cmd> sqrtchisq4 <- sqrt(invchi((run(n)-.5)/n,p))
Cmd> plot(sqrtchisq4, sqrt(sort(d1234)),symbols:"\1",\
        xmin:0, ymin:0, ylab:"Sqrt D1234",\
        xlab:"Sqrt Chi square 4 Probability points",\
        title:"Setosa petals & sepals sqrt Q-Q plot, p = 4")
```




```
Cmd> r_obs <- cor(sqrtchisq4,sqrt(sort(d1234)))[1,2]; r_obs
(1,1)     0.99086 Observed value of correlation statistic
Cmd> M <- 10000;R <- rep(0,M) # vector to hold simulated stats
Cmd> for(i,1,M){ # compute M correlations
      @y <- matrix(rnorm(n*p),n) # rows are N_4(0,I_4)
      R[i] <- cor(sqrt(sort(distcomp(@y))), sqrtchisq4)[1,2];;
    }
Cmd> min(R)# minimum value observed in 10000 trials
(1)     0.94608      Used to set xmin on histogram
```



```
Cmd> hist(R,run(.94,1,.001),xlab:"Correlation", show:F,\
        title:"Histogram of 10,000 Correlation Statistics")
Cmd> addlines(rep(r_obs,2),vector(0,25),linetype:2)
```



Clearly r_obs = 0.9909 is not unusual.

You can estimate a P-value by counting the number of values in R less than or equal to the observed value.

```
Cmd> sum(R <= r_obs)/M # estimated P-value
(1,1)       0.5102
```

MacAnova notes

show:F in hist() suppresses immediate display.

addlines() makes the completed plot visible.

Multivariate Sampling Distributions

Inferential procedures are based on <u>sampling distributions</u> -- distributions of statistics and estimates computed from random samples.

Suppose $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ are a *random sample* from a p-dimensional multivariate distribution with

$$E[\mathbf{x}] = \boldsymbol{\mu}_x \text{ and } V[\mathbf{x}] = \boldsymbol{\Sigma}_x$$

Facts: The mean vector and variance matrix of

$\overline{\mathbf{x}} = (1/n)\sum_{1\le i\le n}\mathbf{x}_i$ are

- $\boldsymbol{\mu}_{\overline{x}} = E[\overline{\mathbf{x}}] = \boldsymbol{\mu}_x$
- $\boldsymbol{\Sigma}_{\overline{x}} = V[\overline{\mathbf{x}}] = (1/n)\boldsymbol{\Sigma}_x$

When p = 1, this is familiar:

- $\mu_{\overline{x}} = E[\overline{x}] = \mu_x$
- $\sigma_{\overline{x}}^2 = V[\overline{x}] = \sigma_x^2/n$

Don't forget of the *conceptual* difference between $\boldsymbol{\mu}_{\overline{x}}$ and $\boldsymbol{\mu}_x$.

More generally, suppose $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ are p×1 random vectors such that

- They are *independent* but may have differing mean vectors:
$$E[\mathbf{x}_i] = \boldsymbol{\mu}_i,\ i = 1,\ldots,n$$
- They have the *same* variance matrix:
$$V[\mathbf{x}_i] = \boldsymbol{\Sigma},\ i = 1,\ldots,n.$$

As usual, we collect the $\mathbf{x}_i$'s in a data data matrix $\mathbf{X} = [\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n]' = [\mathbf{X}_1,\ldots,\mathbf{X}_p]$, with rows corresponding to cases.

Similarly, collect the mean vectors in to a mean matrix

$$\mathbf{M} = E[\mathbf{X}] = \begin{bmatrix} \boldsymbol{\mu}_1' \\ \boldsymbol{\mu}_2' \\ \boldsymbol{\mu}_2' \\ \ldots \\ \boldsymbol{\mu}_n' \end{bmatrix} = [\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_n]'$$

When $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_n = \boldsymbol{\mu}$, $\mathbf{M} = \mathbf{1}_n\boldsymbol{\mu}'$.

Suppose $\mathbf{A} = [a_1, \ldots, a_n]'$ and $\mathbf{B} = [b_1, \ldots, b_n]'$ are vectors of constants for each case.

Then $\sum_{1\le i\le n} a_i\mathbf{x}_i = \mathbf{X}'\mathbf{A}$, and $\sum_{1\le i\le n} b_i\mathbf{x}_i = \mathbf{X}'\mathbf{B}$, are linear combinations of the $\mathbf{x}_i$'s with

- $E[\sum_{1\le i\le n} a_i\mathbf{x}_i] = \sum_{1\le i\le n} a_i\boldsymbol{\mu}_i = \mathbf{M}'\mathbf{A} = (\mathbf{A}'\mathbf{M})'$
- $V[\sum_{1\le i\le n} a_i\mathbf{x}_i] = (\sum_{1\le i\le n} a_i^2)\boldsymbol{\Sigma} = \|\mathbf{A}\|^2\boldsymbol{\Sigma}$
- $\mathrm{Cov}[\sum_{1\le i\le n} a_i\mathbf{x}_i, \sum_{1\le i\le n} b_i\mathbf{x}_i] = \mathrm{Cov}[\mathbf{X}'\mathbf{A}, \mathbf{X}'\mathbf{B}] =$
$$(\textstyle\sum_{1\le i\le n} a_i b_i)\boldsymbol{\Sigma} = (\mathbf{A}'\mathbf{B})\boldsymbol{\Sigma} = (\mathbf{B}'\mathbf{A})\boldsymbol{\Sigma}$$

The last is shorthand for

$$\mathrm{Cov}[\textstyle\sum_{1\le i\le n} a_i x_{ij}, \sum_{1\le i\le n} b_i x_{ik}] = \mathrm{Cov}[\mathbf{A}'\mathbf{X}_j, \mathbf{B}'\mathbf{X}_k] =$$
$$(\textstyle\sum_{1\le i\le n} a_i b_i)\sigma_{jk} = \mathbf{A}'\mathbf{B}\sigma_{jk}$$
$$j = 1,\ldots,p,\ k = 1,\ldots,p$$

Note: When $\mathbf{A}$ and $\mathbf{B}$ are *orthogonal* ($\mathbf{AB} = 0$), $\mathbf{X}'\mathbf{A}$ and $\mathbf{X}'\mathbf{B}$ are uncorrelated.

These results *are not valid*

- When $\mathbf{x}_i$ and $\mathbf{x}_j$ are correlated for $i \ne j$
- When $V[\mathbf{x}_i]$ is not constant

When $a_i = 1/n$, so $\mathbf{A} = n^{-1}\mathbf{1}_n$, you get the familiar result:

- $\sum_{1\le i\le n} a_i\mathbf{x}_i = \overline{\mathbf{x}}$
- $V[\overline{\mathbf{x}}] = V[\sum_{1\le i\le n} a_i\mathbf{x}_i] = \sum_{1\le i\le n} a_i^2\boldsymbol{\Sigma} = (1/n)\boldsymbol{\Sigma}$

Note that

$$\mathbf{X}'\mathbf{A} = (\mathbf{A}'\mathbf{X})' = [\mathbf{A}'\mathbf{X}_1, \ldots, \mathbf{A}'\mathbf{X}_p]' \text{ is a}$$

p-vector with elements

$$\mathbf{A}'\mathbf{X}_j = \textstyle\sum_{1\le i\le n} a_i x_{ij},\ j = 1, \ldots, p$$

$\mathbf{A}'\mathbf{X} = (\mathbf{X}'\mathbf{A})'$ is a row vector with the same elements

Multivariate Central Limit Theorem

As before, let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from a random vector with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$.

1. As $n \to \infty$, ("for large n") $\sqrt{n}(\overline{\mathbf{x}} - \boldsymbol{\mu})$ is approximately $N_p(\mathbf{0},\boldsymbol{\Sigma})$.

Informally, you can interpret this as:

When n is "*large*", $\overline{\mathbf{x}}$ is approximately $N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$

This is the *multivariate central limit theorem* (CLT).

As in the univariate case, there is no universal rule of thumb as to what constitutes "large." Generally you need somewhat larger n than for the univariate CLT.

2. A more general CLT shows that, as $n \to \infty$, many <u>vector statistics</u>

$$\mathbf{y} = \mathbf{g}(\mathbf{X}) = [g_1(\mathbf{X}), g_2(\mathbf{X}), \ldots, g_q(\mathbf{X})]'$$

computed from a data matrix with *independent* rows are approximately multivariate normal.

That is, if $\mathbf{y}$ has dimension q, as $n \to \infty$, $\mathbf{y}$ is approximately $N_q(E[\mathbf{y}], V[\mathbf{y}])$.

In many cases, $V[\mathbf{y}] \cong (1/n)\mathbf{\Sigma}^*$ for some variance matrix $\mathbf{\Sigma}^*$. Sometimes $\mathbf{\Sigma}^* = \mathbf{\Sigma}$ or $\mathbf{\Sigma}^*$ is depends on $\mathbf{\Sigma}$.

Example:

Suppose p = 2 and s_{11} and s_{22} are sample variances and r_{12} = sample correlation between x_1 and x_2.

Then for large n

$$\mathbf{y} = [\sqrt{s_{11}}, \sqrt{s_{22}}, \tanh^{-1} r_{12}]'$$

is approximately $N_3(E[\mathbf{y}], V[\mathbf{y}])$, where

$$E[\mathbf{y}] \cong [\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \tanh^{-1}\rho_{12}]'.$$

and $V[\mathbf{y}] \cong \mathbf{\Sigma}^*/n$ where $\mathbf{\Sigma}^*$ can be expressed in terms of moments of $\mathbf{y}$ (in terms of $\mathbf{\Sigma}$ when $\mathbf{x}$ is normal).

Here q = 3 and $g_1(\mathbf{X}) = \sqrt{s_{11}}$ $g_2(\mathbf{X}) = \sqrt{s_{22}}$, $g_3(\mathbf{X}) = \tanh^{-1} r_{12}$.

Note: $\tanh^{-1} r = (1/2)(\log(1+r) - \log(1-r))$

MacAnova

Function `z <- atanh(r)` computes $z = \tanh^{-1}(r)$ and `r <- tanh(z)` computes $r = (e^z - e^{-z})/(e^z + e^{-z})$ from z.

The transformation of r

$$z = z(r) \equiv \tanh^{-1} r \equiv 0.5*\log((1+r)/(1-r))$$

is the **Fisher z-transformation** for correlation coefficients.

When X_1 and X_2 are bivariate normal, the distribution of z(r) is very closely approximated by to $N_1(\tanh^{-1}\rho, 1/(n-3))$.

Because $V(z) \cong 1/(n-3)$ doesn't depend on ρ, you can use z(r) for inference about ρ from one or more bivariate random samples.

Examples: Confidence limits for ρ

$\tanh(z_L) \le \rho \le \tanh(z_U)$, where

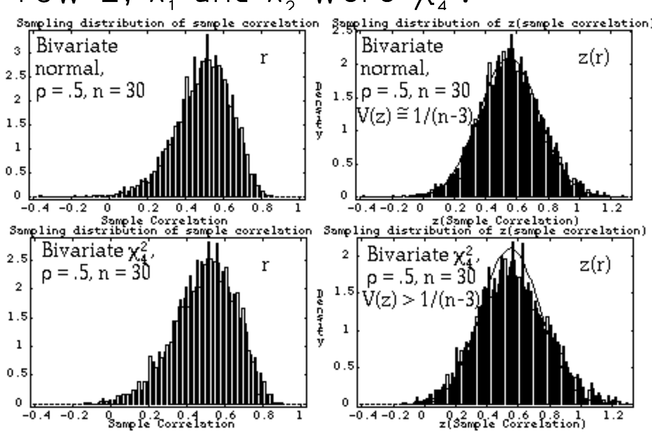
$$(z_L, z_U) = z(r) \pm z_{\alpha/2}/\sqrt{(n-3)}$$

Test statistic for H_0: $\rho_1 = \rho_2$

$$Z = (z(r_1) - z(r_2))/\sqrt{\{1/(n_1-3)+1/(n_2-3)\}}$$

With non-normal data, z(r) is often close to normality but with $V(z) \ne 1/(n - 3)$.

These graphs from simulation display the distribution of r and z(r) for n = 30 with $\rho = .5$. In row 1, (x_1,x_2 were N(0,1); in the row 2, x_1 and x_2 were χ_4^2.



Although the distribution of r is skewed, the distribution of z(r) is nearly normal.

Why did I choose $\sqrt{s_{11}}$, $\sqrt{s_{22}}$ and $\tanh^{-1}r$ for this example? It might seem more natural to use the variances and covariances s_{11}, s_{22}, and s_{12}.

In fact, as $n \to \infty$, $[s_{11}, s_{22}, s_{12}]'$ is approproximately $N_3([\sigma_{11}, \sigma_{22}, \sigma_{12}]', \mathbf{\Sigma}^{**}/n)$, where, when $\mathbf{x}$ is bivariate normal, $\mathbf{\Sigma}^{**}$ depends on $\mathbf{\Sigma}$.

However, you need a larger n for $[s_{11}, s_{22}, s_{12}]'$ to be approximately N_3 than for $[\sqrt{s_{11}}, \sqrt{s_{22}}, \tanh^{-1}r_{12}]'$.

The CLT and the generalized CLT are important because of the following related **facts**.

3. When a q-vector $\mathbf{y}$ of estimates or statistics computed from a random sample, is *approximately* N_q, then

   $T^2 \equiv d(\mathbf{y},E[\mathbf{y}])^2 = (\mathbf{y}-E[\mathbf{y}])'\{V[\mathbf{y}]\}^{-1}(\mathbf{y}-E[\mathbf{y}])$

   is *approximately* distributed as χ_q^2

4. In ***large samples***, when $\mathbf{y}$ is approximately N_q and when $\hat{V}[\mathbf{y}]$ is a <u>consistent estimator</u> of $V[\mathbf{y}]$,

   $T^2 \equiv (\mathbf{y} - E[\mathbf{y}])'\{\hat{V}[\mathbf{y}]\}^{-1}(\mathbf{y} - E[\mathbf{y}])$

   is <u>approximately</u> χ_q^2 ($\mathbf{y}$ is a q-vector)

This generalizes the fact that in many cases $t^2 = \{(\hat{\theta} - \theta)/\hat{\sigma}_{\hat{\theta}}\}^2 \mathrel{\tilde{=}} \chi_1^2$ for large n, where $\hat{\theta}$ is an estimate with estimated variance $\hat{\sigma}_{\hat{\theta}}^2$.

Example: ***Large sample test*** of multivariate mean:

- $\mathbf{y} = \overline{\mathbf{x}}$ with $E[\mathbf{y}] = \boldsymbol{\mu}$, $\hat{V}[\mathbf{y}] = \hat{V}[\overline{\mathbf{x}}] = n^{-1}\mathbf{S}$.

Then, q = p and

$$\begin{aligned} T^2 &= T^2(\boldsymbol{\mu}) = (\overline{\mathbf{x}} - \boldsymbol{\mu})'\hat{V}[\overline{\mathbf{x}}]^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \\ &= (\overline{\mathbf{x}} - \boldsymbol{\mu})'\{\mathbf{S}/n\}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \\ &= n(\overline{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \\ &\mathrel{\tilde{=}} \chi_p^2 \end{aligned}$$

A *large sample* test of H_0: $\boldsymbol{\mu}_x = \boldsymbol{\mu}_0$ with significance level α is

"Reject H_0: $\boldsymbol{\mu}=\boldsymbol{\mu}_0$ when $T^2(\boldsymbol{\mu}_0) > \chi_p^2(\alpha)$".

Vocabulary

$T^2(\boldsymbol{\mu}_0)$ is the <u>one-sample</u> **Hotelling's T^2** statistic for testing H_0: $\boldsymbol{\mu}_x = \boldsymbol{\mu}_0$.

When p = 1, $T^2 = t^2$, where

$$t = (\overline{x} - \mu_0)/(s_x/\sqrt{n})$$

is the usual one sample t-statistic.

MacAnova

You can compute $\chi_p^2(\alpha)$ by

`invchi(1-alpha,p)`

or

`invchi(alpha,p,upper:T)`

You can compute T^2 using `hotellval()`

```
Cmd> irisdata <- read("","t11_05",quiet:T)
Read from file "TP1:Stat5401:Stat5401F05:Data:JWData5.txt"
Cmd> setosa <- irisdata[irisdata[,1] == 1,-1]
Cmd> stats <- tabs(setosa, mean:T,covar:T)
Cmd> ybar <- stats$mean; s <- stats$covar
Cmd> ybar # sample mean vector
(1)       5.006       3.428       1.462       0.246
Cmd> mu_0 <- vector(4.5,3,2,1) # hypothesized μ
Cmd> n <- nrows(setosa); vhat <- s/n
Cmd> tsq <- (ybar - mu_0)' %*% solve(vhat) %*% (ybar - mu_0)
Cmd> tsq # T^2 computed by white box method
(1,1)      28.102
Cmd> hotellval(setosa - mu_0')#  T^2 by black box method
(1,1)      28.102
Cmd> cumchi(tsq,ncols(setosa),upper:T) # P-value
(1,1)   1.1891e-05       Strong evidence against H0: μ = μ0
Cmd> tval(setosa - mu_0') # univariate t-statistics
(1)      -3.8917      -1.3431      -1.5472      -3.6232
```

MacAnova

- `solve(A)` computes $\mathbf{A}^{-1}$
- `solve(A,b)` or `A %\% b` computes $\mathbf{A}^{-1}\mathbf{b}$

Small sample distribution for normal $\mathbf{x}$

5. When $\mathbf{x}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\overline{\mathbf{x}}$ is $N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$, for any n
- $T^2 \equiv (\overline{\mathbf{x}} - \boldsymbol{\mu})'\{\mathbf{S}/n\}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})$ is distributed, for any $n > p$, as

  $$\{(pf_e)/(f_e - p + 1)\}F_{p,f_e-p+1} \qquad f_e = n - 1$$

  Put another way,

  $$((f_e-p+1)/(f_e p))\, T^2 = F_{p,f_e-p+1}$$

- This is a ***small sample*** result which *requires normality* to be exactly correct
- It is quite robust against non-normality. That is, it at least approximately "works as advertised" even when the data are not normal, except when n is very small.

The denominator degrees of freedom are $f_e - (p - 1)$: In a certain sense you lose a d.f. for each dimension after the first.

Small sample test of H_0: $\boldsymbol{\mu}_x = \boldsymbol{\mu}_0$,

"Reject H_0 when

$$((f_e-p+1)/(f_e p))T^2(\boldsymbol{\mu}_0) > F_{p,f_e-p+1}(\alpha)"$$

You can compute $F_{p,f_e-p+1}(\alpha)$ by

```
invF(1-alpha,p,fe-p+1)
```

or

```
invF(alpha,p,fe-p+1,upper:T)
```

For large n (large f_e), the *small* sample $\{(pf_e)/(f_e - p + 1)\}F_{p,f_e-p+1}$ distribution is consistent with the *large* sample χ_p^2 distribution:

- For large f_e,
  $(f_e p)/(f_e - p + 1) = p(1 + (p-1)/f_e) \cong p$
- $F_{p,f_e-p+1} \cong F_{p,\infty} = \chi_p^2/p$.

So

$$T^2 = ((f_e p)/(f_e-p+1))F_{p,f_e-p+1}$$
$$\cong pF_{p,f_e-p+1} \cong pF_{p,\infty} \cong \chi_p^2$$

Here's a slightly less artificial example with the iris data.

The variables are sepal length, sepal width, petal length and petal width.

A hypothesis conceivably of interest might be that the mean sepal lengths = mean sepal widths *and* mean petal lengths = mean petal widths.

Symbolically this is

$$H_0: \mu_1 = \mu_2,\ \mu_3 = \mu_4$$

or

$$H_0: \mu_1 - \mu_2 = 0 \text{ and } \mu_3 - \mu_4 = 0$$

or

$$H_0: \boldsymbol{\mu}_y = \mathbf{0}, \text{ where } \mathbf{y} = \begin{bmatrix} x_1 - x_2 \\ \\ x_3 - x_4 \end{bmatrix}.$$

H_0 is a hypothesis about the *shape* of the sepals and petals (probably a very implausible one).


```
Cmd> Y <- hconcat(setosa[,1] - setosa [,2],\
        setosa [,3] - setosa [,4])

Cmd> t_sq <- hotellval(Y - 0); t_sq
(1,1)      4012.1

Cmd> p <- ncols(Y); fe <- n - 1; vector(p, fe)
(1)          2         49

Cmd> invchi(.01,p,upper:T) # ChiSq_2(.01)
(1)      9.2103             large sample 1% critical value

Cmd> (p*fe/(fe-p+1))*invF(.01,p, fe-p+1, upper:T)
(1)      10.365             small sample 1% critical value

Cmd> f <- ((fe-p+1)/(p*fe))*t_sq; f
(1,1)      1965.1           F form of T^2

Cmd> invF(.01,p, fe-p+1, upper:T) # F_2_48(.01)
(1)      5.0767             small sample 1% crit. val. for F

Cmd> cumF((fe - p + 1)*t_sq/(fe*p),p,fe-p+1,upper:T)
(1,1)  9.0628e-47
```

T^2 much much larger than $\chi_2^2(.01)$ = 9.2103 and $((f_e-p+1)/(f_e p))T^2$ is far beyond $F_{2,48}(.01) = 5.0767$.