

University of Washington



STATISTICS

Elements of Statistical Methods Association (Ch 14)

Fritz Scholz

Spring Quarter 2010

May 28, 2010

Association Between X and Y

In the previous chapter we examined dependence or association between two partitions of a sample space.

With 2 random variables X and Y , defined on the same sample space S , we could use each of them to partition the sample space, e.g., $A_i = \{s \in S : X(s) \in (a_i, a_{i+1}]\}$ for nonoverlapping intervals $(a_i, a_{i+1}]$ spanning the whole of R , using $(-\infty, a_1)$ and (a_{r-1}, ∞) to keep the number of partition sets finite. Similarly using Y to get B_j .

While the G^2 and X^2 tests could be used to examine dependence or association between X and Y , these tests don't use any order relationships in the values of X and Y , i.e., the partition sets could be permuted around without changing the values of G^2 and X^2 .

Also, the choice of partition intervals is somewhat arbitrary.

Joint Distribution of X and Y

We will deal with continuous r.v.s X and Y defined on the same sample space S .

For any rectangular B region in R^2

$$B = [a, b] \times [c, d] = \{(x, y) : a \leq x \leq b, c \leq y \leq d\} \subset R^2$$

we need to assign a probability

$$P((X, Y) \in B) = P(X \in [a, b], Y \in [c, d])$$

In analogy with the case of a univariate continuous r.v. we use a **joint probability density function** $f(x, y)$ defined over $(x, y) \in R^2$ and define $P((X, Y) \in B)$ as the volume under the function f over the region B ,

$$P((X, Y) \in B) = \text{Volume}_B(f) = \int_a^b \int_c^d f(x, y) dx dy = \int \int_B f(x, y) dx dy$$

The right most expression holds for any Borel set B , not just for rectangles.

$f(x, y)$ is a joint density whenever $f(x, y) \geq 0$ for all $(x, y) \in R^2$

and $\text{Volume}_{R^2}(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = P((X, Y) \in R^2) = 1$

Marginal Distributions and Independence

From the joint density $f(x, y)$ we get the **marginal densities** of X alone and Y alone, respectively as

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy = \text{area over } y\text{-axis under } f(x, y) \text{ for fixed } x$$

and

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \text{area over } x\text{-axis under } f(x, y) \text{ for fixed } y$$

because (interpreting volumes under $f(x, y)$ as the accumulation of areas along x - or y -axis)

$$F_X(a) = P(X \leq a) = P((X, Y) \in (-\infty, a] \times \mathbf{R}) = \int_{-\infty}^a \int_{-\infty}^{\infty} f(x, y) dy dx = \int_{-\infty}^a f_x(x) dx$$

$$F_Y(b) = P(Y \leq b) = P((X, Y) \in \mathbf{R} \times (-\infty, b]) = \int_{-\infty}^b \int_{-\infty}^{\infty} f(x, y) dx dy = \int_{-\infty}^b f_y(y) dy$$

X and Y are independent, if and only if $f(x, y) = f_x(x) \cdot f_y(y)$ for all $(x, y) \in \mathbf{R}^2$

Joint Density Contours

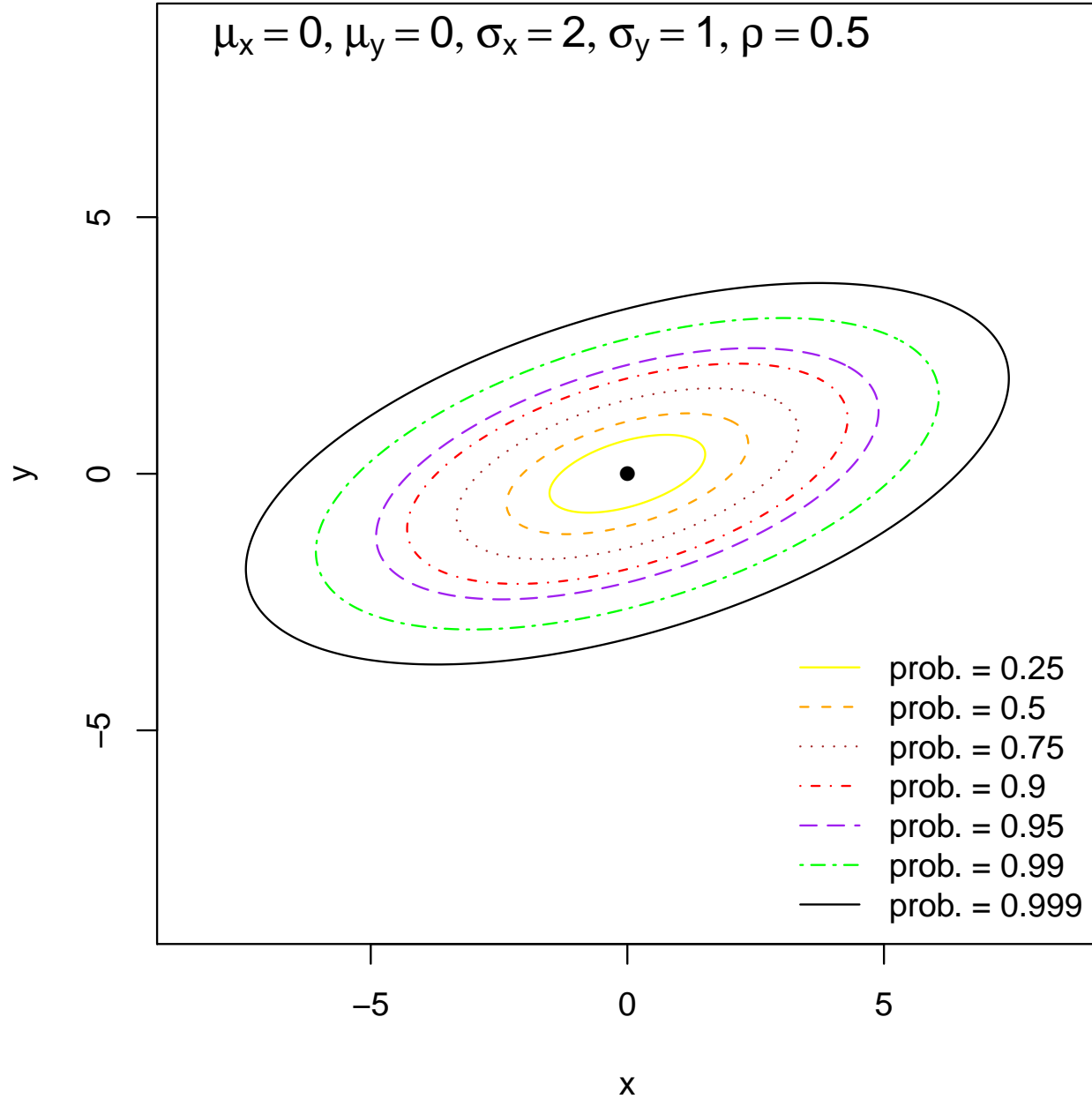
A joint density $f(x, y)$ can be viewed as providing an elevation $f(x, y) \geq 0$ at each point in the plane R^2 .

High (low) density in the vicinity of (x, y) means high (low) probability in the same vicinity of (x, y)

We can, as in topographic maps, consider equal density contour lines to get an appreciation of the probability distribution described by $f(x, y)$.

$$\left\{ (x, y) \in R^2 : f(x, y) = c \right\} = \text{contour set for the density } f(x, y) \text{ at level } c$$

Bivariate Normal Probability Contours



Positive and Negative Association

Assume $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 1)$, but not necessarily independent.

To get a sense of the dependence between X and Y consider $E(XY)$.

There is **positive association** between X and Y when positive (negative) values of X tend to be associated with positive (negative) values of Y .

In either case XY will tend to be mostly positive and thus $E(XY) > 0$.

$$(+) \cdot (+) > 0 \quad \text{and} \quad (-) \cdot (-) > 0.$$

There is **negative association** between X and Y when positive (negative) values of X tend to be associated with negative (positive) values of Y .

In either case XY will tend to be mostly negative and thus $E(XY) < 0$.

$$(+) \cdot (-) < 0 \quad \text{and} \quad (-) \cdot (+) < 0.$$

Product-Moment Correlation Coefficient

When $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ we measure dependence in terms of

standard units $X' = \frac{X - \mu_x}{\sigma_x}$ and $Y' = \frac{Y - \mu_y}{\sigma_y}$

via $\rho = \rho(X, Y) = E(X'Y') = E \left[\left(\frac{X - \mu_x}{\sigma_x} \right) \left(\frac{Y - \mu_y}{\sigma_y} \right) \right]$

ρ is called the [product-moment correlation coefficient](#).

This definition holds generally, whether the marginal distributions are normal or not.

All that is needed is that means and variances exist and are finite.

For constants $a > 0, b, c > 0, d$ we have $\rho(aX + b, cY + d) = \rho(X, Y)$,

i.e., scaling and shifting the random variables does not affect their correlation.

Properties Concerning $\rho = \rho(X, Y)$

Assume that X and Y have finite variance σ_x^2 and σ_y^2

1. $-1 \leq \rho \leq 1$

2. $\rho = \pm 1 \iff (Y - \mu_y)/\sigma_y = \pm(X - \mu_x)/\sigma_x$

Y is a linear function of X with positive/negative slope.

3. If X and Y are independent, then $\rho = 0$.

The converse is not necessarily true, but \downarrow

4. If X and Y are bivariate normal (elliptical density contours) and $\rho = 0$, then X and Y are independent.

The parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ determine a unique bivariate normal pdf.

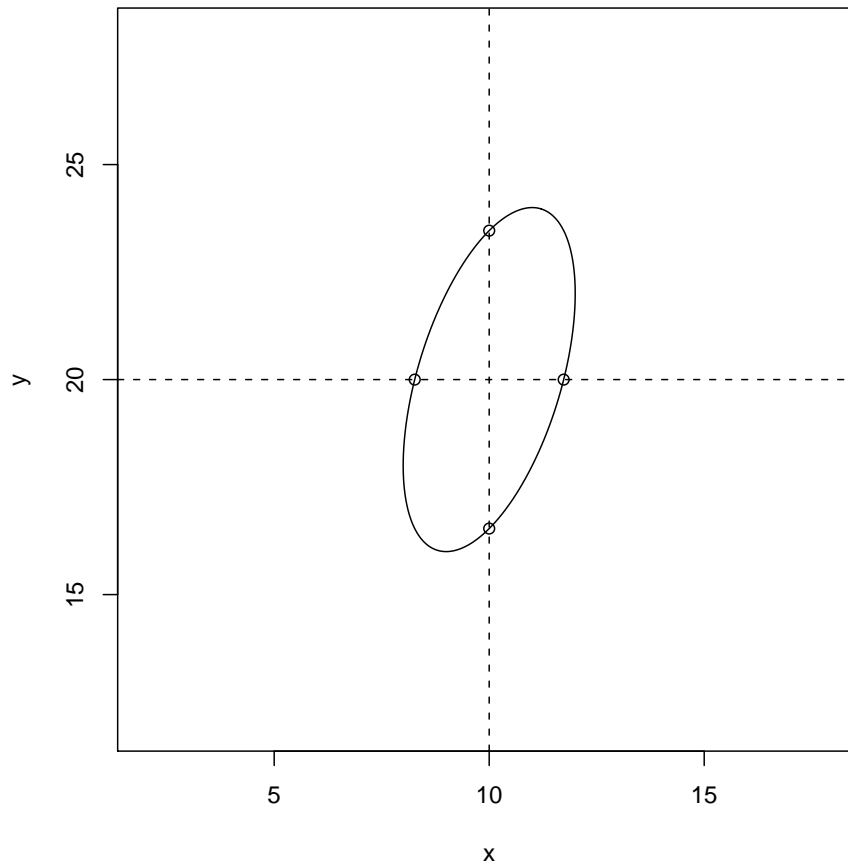
Closeness of $|\rho|$ to 1 drives the narrowness of the contour ellipses.

Bivariate Normal Concentration Ellipse

The contour level ellipse passing through the points

$$(\mu_x \pm \sqrt{1 - \rho^2} \cdot \sigma_x, \mu_y) \quad \text{and} \quad (\mu_x, \mu_y \pm \sqrt{1 - \rho^2} \cdot \sigma_y)$$

is called the **concentration ellipse**.



Bivariate Normal Samples

Such bivariate samples of paired observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, drawn independently from a bivariate population, are usually displayed in a two column format.

This facilitates easier comparison within each pair and across pairs.

The table below shows the two midterm scores for a class of 20 students

x	y	x	y	x	y	x	y
87	87	82	66	84	75	94	69
25	57	94	86	99	92	99	98
76	91	89	74	92	55	63	81
84	67	92	92	74	74	82	80
91	67	76	85	84	74	91	85

While it is easy to contemplate the two scores for any given student, it becomes harder to grasp the meaning for the totality of scores.

⇒ scatter plot or scatter diagram.

Scatter Plot Code

Getting the data: Download `midterms351.dat`

from <http://mypage.iu.edu/~mtrosset/StatInfeR.html>

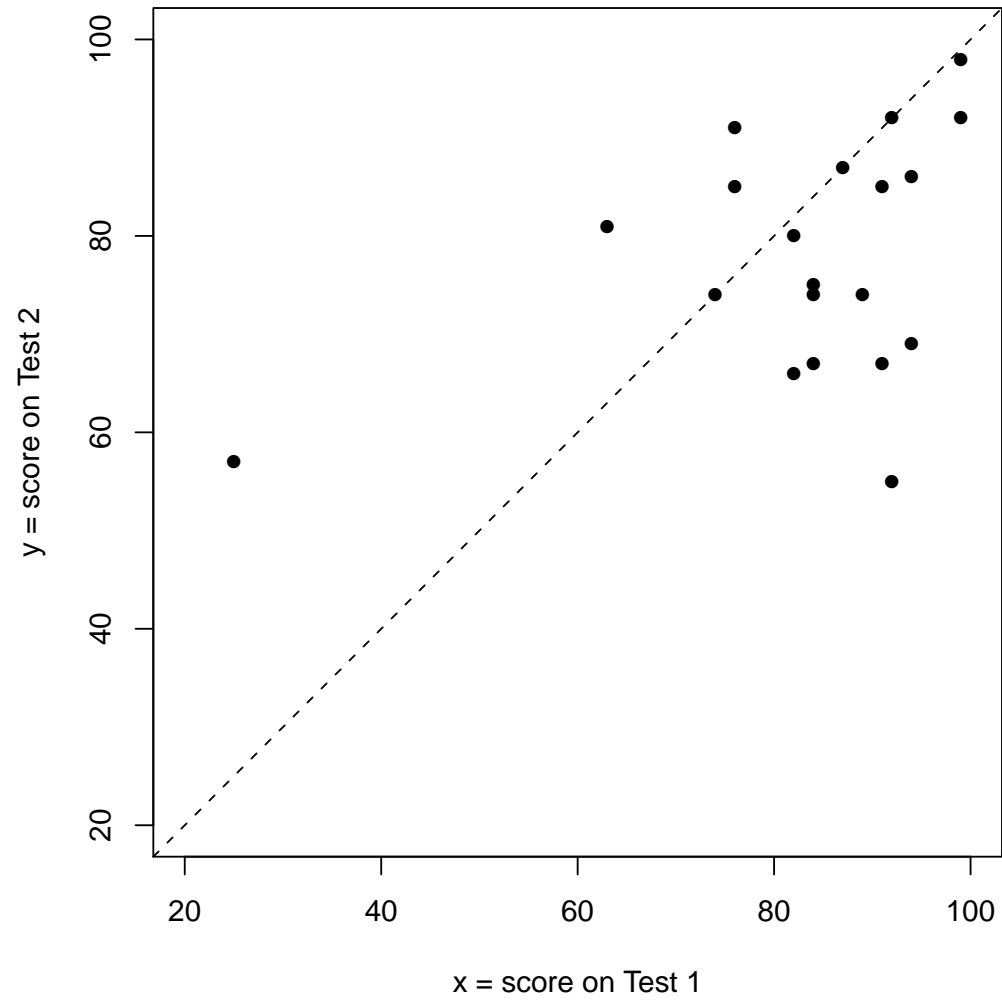
Getting the data into **R**, assuming the file is in the directory from which **R** started

```
midterms <- read.table("midterms351.dat")
```

The function for the scatter plot:

```
midterms.plot <- function (dat = midterms, PDF = F) {  
  if (PDF == T) pdf(file = "scatterplot.pdf", width = 6)  
  par(pty = "s")  
  plot(dat[, 1], dat[, 2], xlim = c(20, 100), ylim = c(20, 100),  
        xlab = "x = score on Test 1", ylab = "y = score on Test 2",  
        pch = 16)  
  if (PDF == T) dev.off() }  
}
```

Scatter Plot



Comments on Scatter

We seem to have a clear outlier.

About 5 students had consistent scores on both midterms.

4 students improved on the second midterm,
while 11 dropped in their performance.

The point pattern does not appear to be elliptical
(even without the outlier).

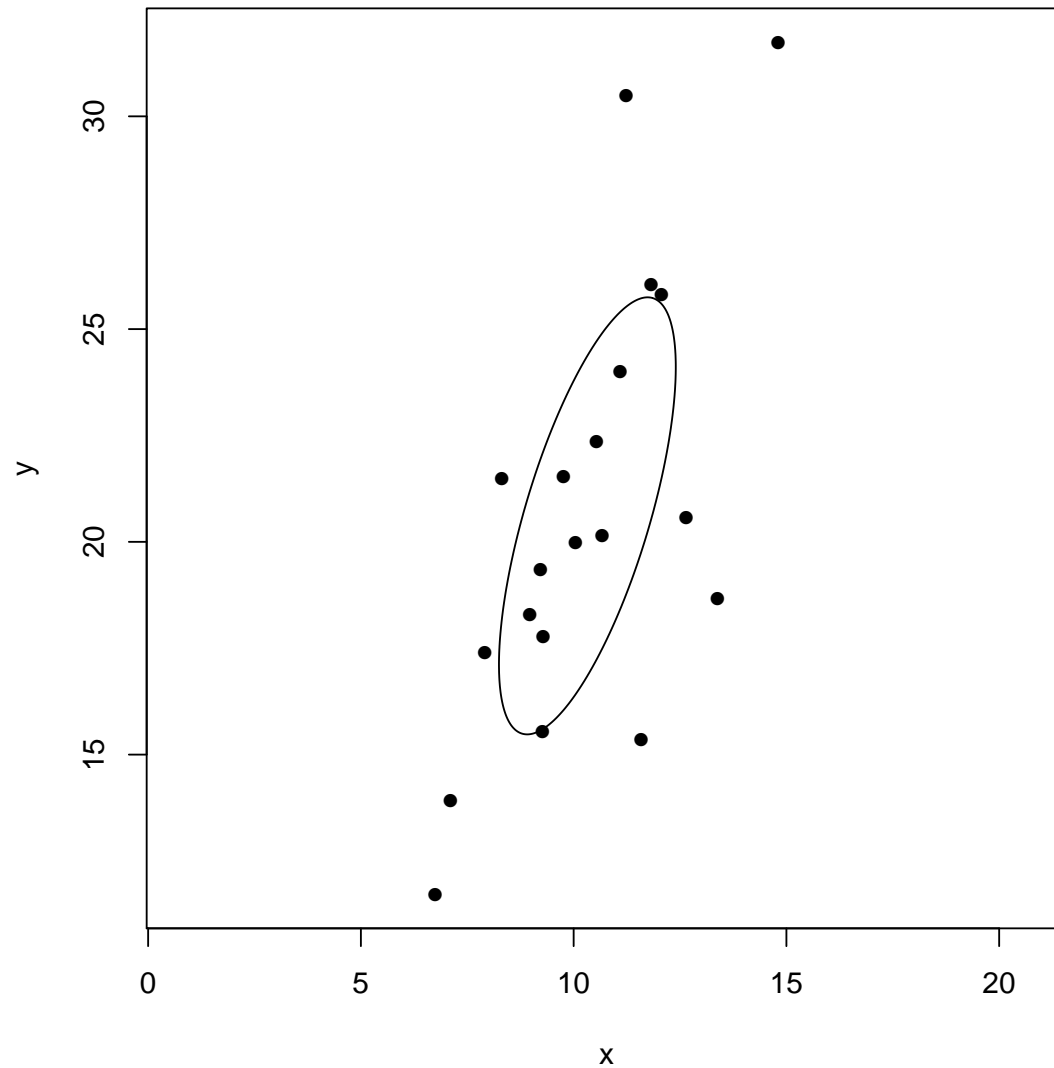
However, the sample size is small to make any hard pronouncements
against bivariate normality.

The next slide shows a random sample of size $n = 20$ from a
bivariate normal distribution with superimposed concentration ellipse.

```
binorm.scatter(binorm.sample(pop=c(10,20,4,16,.5),20))
```

Scatter Plot for Bivariate Normal Data

Scatter Diagram



Bivariate Normal Samples via `binorm.sample`

```
> pop <- c(10,20,4,16,.5) # setting bivariate normal parameters

> binorm.sample(pop,5) # getting a sample of size n = 5
      [,1]      [,2]
[1,] 12.94626 23.29288
[2,] 11.22774 18.59080
[3,] 13.54964 26.17255
[4,] 11.03156 17.25417
[5,] 10.85842 25.18908
```

The output is given in the form of a 2 column data matrix.

`binorm.sample` can be obtained from

<http://mypage.iu.edu/~mtrosset/StatInfeR.html>

via the link [binorm.R](#).

Pearson's Product-Moment Correlation Coefficient

We discussed how to estimate the population means μ_x, μ_y using the sample means \bar{x} and \bar{y} and the population variances σ_x^2 and σ_y^2 using the sample variances s_x^2 and s_y^2 . The plug-in estimate of the population correlation coefficient is

$$\begin{aligned}\hat{\rho} &= \frac{1}{n} \sum_{i=1}^n \left[\left(\frac{x_i - \hat{\mu}_x}{\hat{\sigma}_x} \right) \left(\frac{y_i - \hat{\mu}_y}{\hat{\sigma}_y} \right) \right] \\ &= \frac{1}{n} \sum_{i=1}^n \left[\left(\frac{x_i - \bar{x}}{\sqrt{(n-1)s_x^2/n}} \right) \left(\frac{y_i - \bar{y}}{\sqrt{(n-1)s_y^2/n}} \right) \right] \\ &= \frac{1}{n-1} \sum_{i=1}^n \left[\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right]\end{aligned}$$

called **Pearson's product-moment correlation coefficient**, usually denoted by r .

$\hat{\rho} \xrightarrow{P} \rho$ as $n \rightarrow \infty$, i.e., $\hat{\rho}$ is a consistent estimator of ρ .

The next slide shows how to estimate all 5 parameters using `binorm.estimate`.

binorm.estimate and binorm.scatter

The featured functions `binorm.estimate` and `binorm.scatter` can be obtained from <http://mypage.iu.edu/~mtrosset/StatInfer.html> via the link [binorm.R](#).

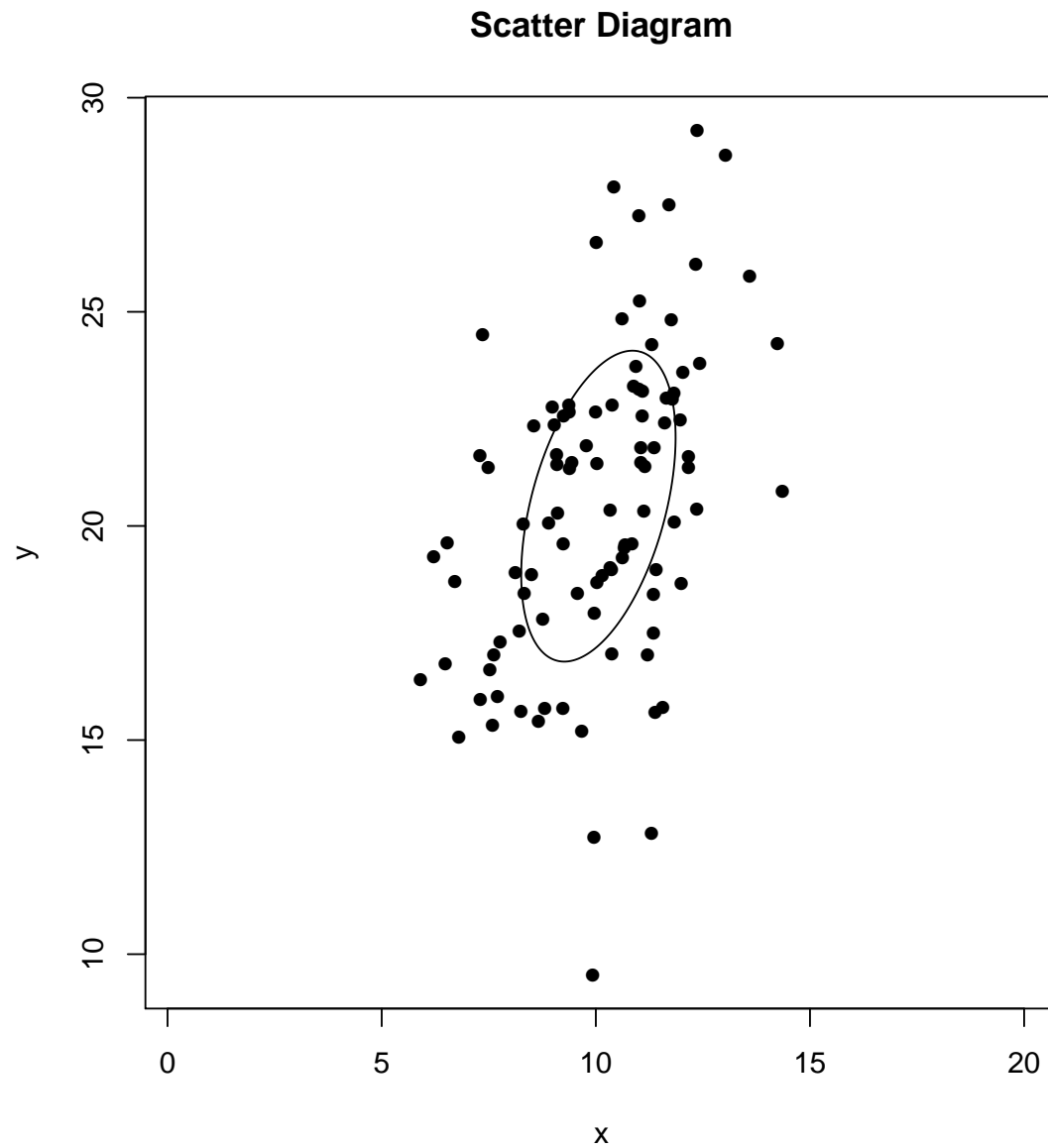
```
> Data <- binorm.sample(pop,100) # generating sample, n = 100

> binorm.estimate(Data) # estimating parameters
[1] 10.0596963 20.4624716  3.2477105 13.1775460  0.4419401

> binorm.scatter(Data) # making the scatter plot, see next slide
```

To get the fatter plotting dots I changed `pch="."`, `cex=2` to `pch=16` inside `binorm.scatter`

Scatter Plot for Bivariate Normal Data



Testing $H_0 : \rho = 0$

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution, i.e., the pairs are independent with common bivariate normal distribution.

When X_i and Y_i are uncorrelated, i.e., independent, then $\hat{\rho}$ as a function of $(X_1, Y_1), \dots, (X_n, Y_n)$ is a random variable and we have

$$T = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t(n-2)$$

Clearly, large values of $|\hat{\rho}|$, or equivalently large values of $|T|$, are significant.

We can assess the significance probability for any observed value r of $\hat{\rho}$ via

$$\begin{aligned} \mathbf{p}(r) &= P_{\rho=0}(|\hat{\rho}| \geq |r|) = P_{\rho=0}\left(|T| \geq \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}\right) = P_{\rho=0}(|T| \geq |t|) \\ &= 2 * \text{pt}(-\text{abs}(t), n-2) = 2 * \text{pt}(-\text{abs}(r) * \text{sqrt}((n-2)/(1-r^2)), n-2) \end{aligned}$$

Reject H_0 at level α whenever $\mathbf{p}(r) \leq \alpha$.

Using Critical Values

With $q_t = qt(1 - \alpha/2, n - 2)$ we reject H_0 at level $\alpha = \alpha$, whenever

$$\begin{aligned} \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \geq q_t &\iff \frac{r^2(n-2)}{1-r^2} \geq q_t^2 \iff r^2(n-2) \geq (1-r^2)q_t^2 = q_t^2 - r^2q_t^2 \\ &\iff r^2(n-2+q_t^2) \geq q_t^2 \iff r^2 \geq \frac{q_t^2}{n-2+q_t^2} \end{aligned}$$

Strong Association Evidence $\stackrel{??}{=} \text{Strong Association}$

Strong evidence that association is present does not necessarily mean that there is evidence of a strong association.

The attribute “strong” modifies two different words: evidence and association.

Strong association means that X and Y are highly correlated, i.e., knowing X will give us a good idea about the value of Y . This occurs when $|\rho|$ is close to 1.

Strong evidence means that we have sufficient data to reject the hypothesis $H_0 : \rho = 0$. However, $|\rho|$ may still be quite small (\Rightarrow weak association).

The multiplier $\sqrt{n-2}$ in the denominator of T will lead to rejection of H_0 , even for relatively small $\hat{\rho}$.

Demonstration Code

```
StrongEvidence <- function(n=300, rho=.1, iseed=13) {  
  set.seed(iseed)  
  pop <- c(0,0,1,1, rho)  
  data <- binorm.sample(pop, n)  
  est <- binorm.estimate(data)  
  binorm.scatter(data)  
  r <- est[5]  
  pval.r <- 2*pt(-abs(r)*sqrt((n-2)/(1-r^2)), n-2)  
  list(pval.r=pval.r, estimates = est)  
}
```

Illustrations

```
> StrongEvidence(n=300, rho=.1, iseed=13)
```

```
$pval.r
```

```
[1] 0.006001744 # strong evidence, but weak association.
```

```
$estimates
```

```
[1] -0.09926213 -0.00642963 1.12713397 0.98752296 0.15829714
```

```
> StrongEvidence(n=10, rho=.8, iseed=42)
```

```
$pval.r
```

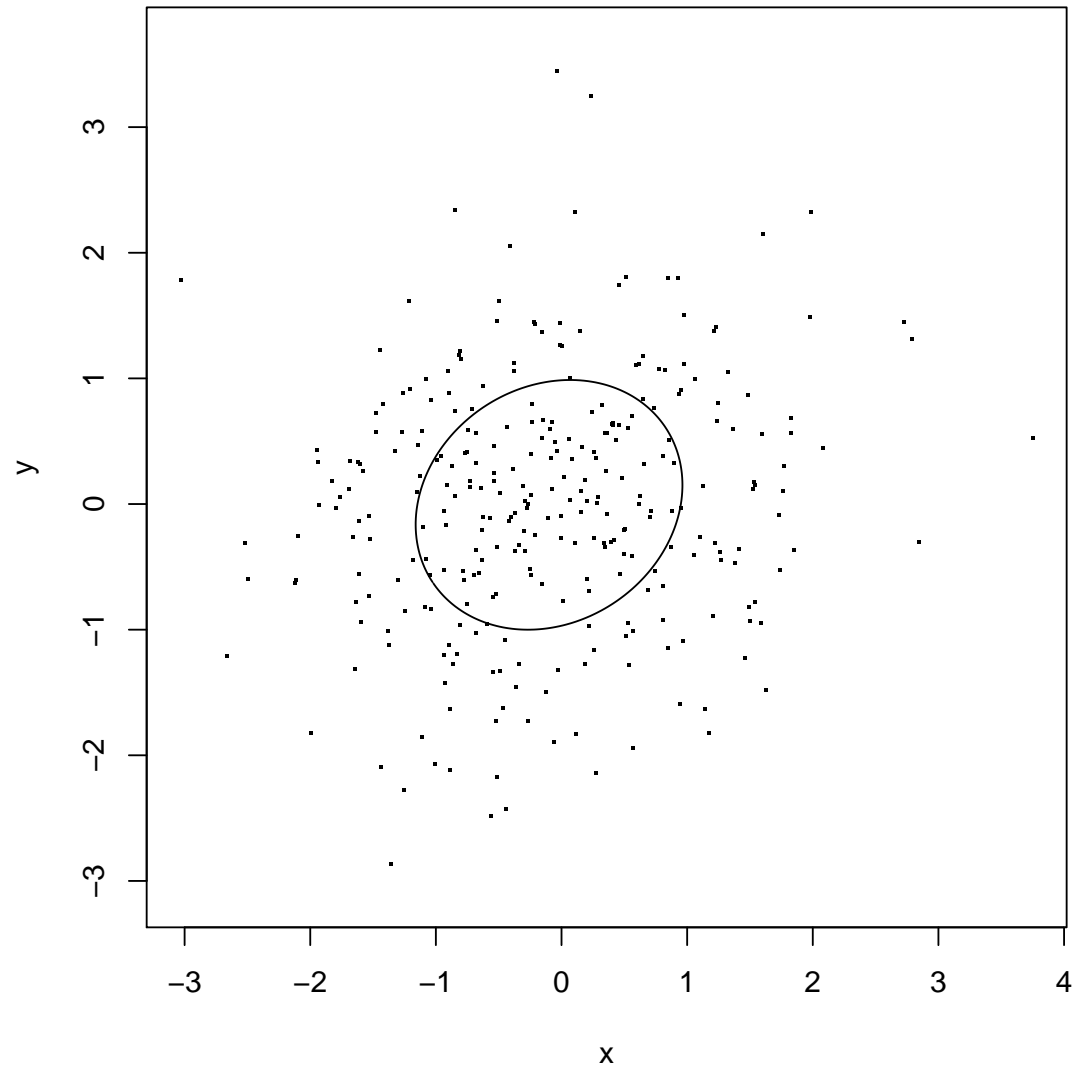
```
[1] 0.2132450 # strong association, but weak evidence.
```

```
$estimates
```

```
[1] 0.5709009 0.4675217 1.2005104 0.5876052 0.4313588
```

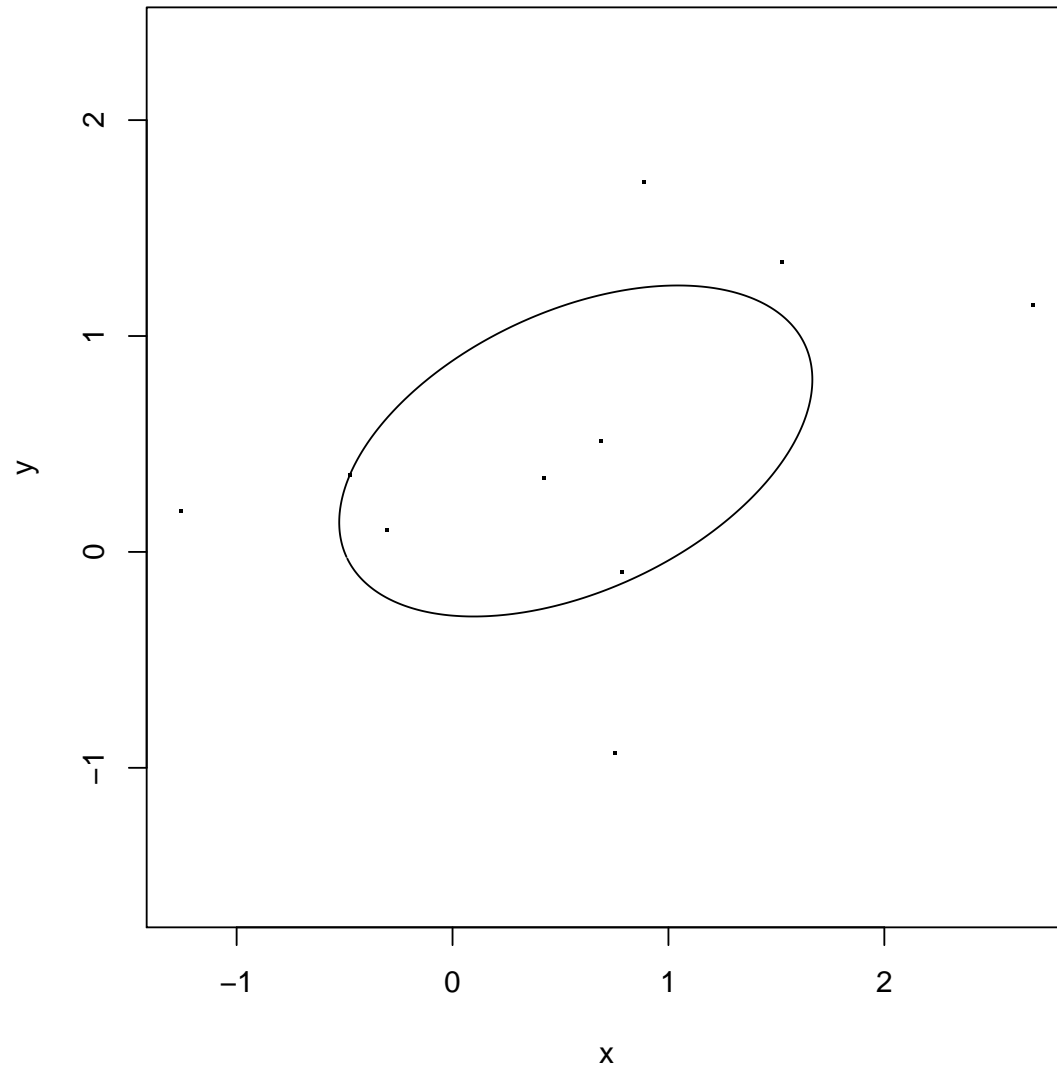

Strong Evidence of Weak Association

Scatter Diagram



Weak Evidence of Strong Association

Scatter Diagram



Confidence Intervals for ρ

Confidence intervals for ρ consist of all those values ρ_0 for which the hypothesis $H_0 : \rho = \rho_0$ is not rejected at level α when testing H_0 against $H_1 : \rho \neq \rho_0$.

The exact distribution of $\hat{\rho}$ for $\rho \neq 0$ is rather complicated and exact confidence intervals based on it appear not yet to have been implemented in [R](#).

Instead we resort to a large sample approximation based on [Fisher's \$z\$ -transform](#).

Sir Ronald Fisher was one of the founding fathers of statistics and is also famous (some say: more famous) for his work in genetics.

Fisher's Z-Transform of ρ and $\hat{\rho}$

Fisher showed that the following z -transform f , when applied to ρ and $\hat{\rho}$,

$$\zeta = f(\rho) = \frac{1}{2} \log \left(\frac{1 + \rho}{1 - \rho} \right) \quad \text{and} \quad \hat{\zeta} = f(\hat{\rho}) = \frac{1}{2} \log \left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right)$$

provides the following approximate distributional results for sufficiently large n

$$\hat{\zeta} \approx \mathcal{N} \left(\zeta, \frac{1}{n-3} \right) \quad \text{or} \quad \frac{\hat{\zeta} - \zeta}{1/\sqrt{n-3}} = \sqrt{n-3} (\hat{\zeta} - \zeta) \approx Z \sim \mathcal{N}(0, 1)$$

Note that the range of the z -transform $f(\rho)$ is $(-\infty, \infty)$ as ρ varies over $(-1, 1)$, giving the normal approximation the unlimited range it needs.

With $q_z = \text{qnorm}(1 - \alpha/2) = \text{qnorm}(1 - \alpha/2)$ we thus get the following approximate $(1 - \alpha)$ -level confidence interval for ζ

$$\hat{\zeta} \pm \frac{q_z}{\sqrt{n-3}}$$

Approximate Confidence Intervals for ρ

Since $\zeta = f(\rho)$ is a strictly increasing function of ρ , one can convert confidence intervals for ζ into corresponding confidence intervals for ρ .

We just need to invert the z -transformation $f(\rho)$.

$$\rho = f^{-1}(\zeta) = \frac{e^{2\zeta} - 1}{e^{2\zeta} + 1}$$

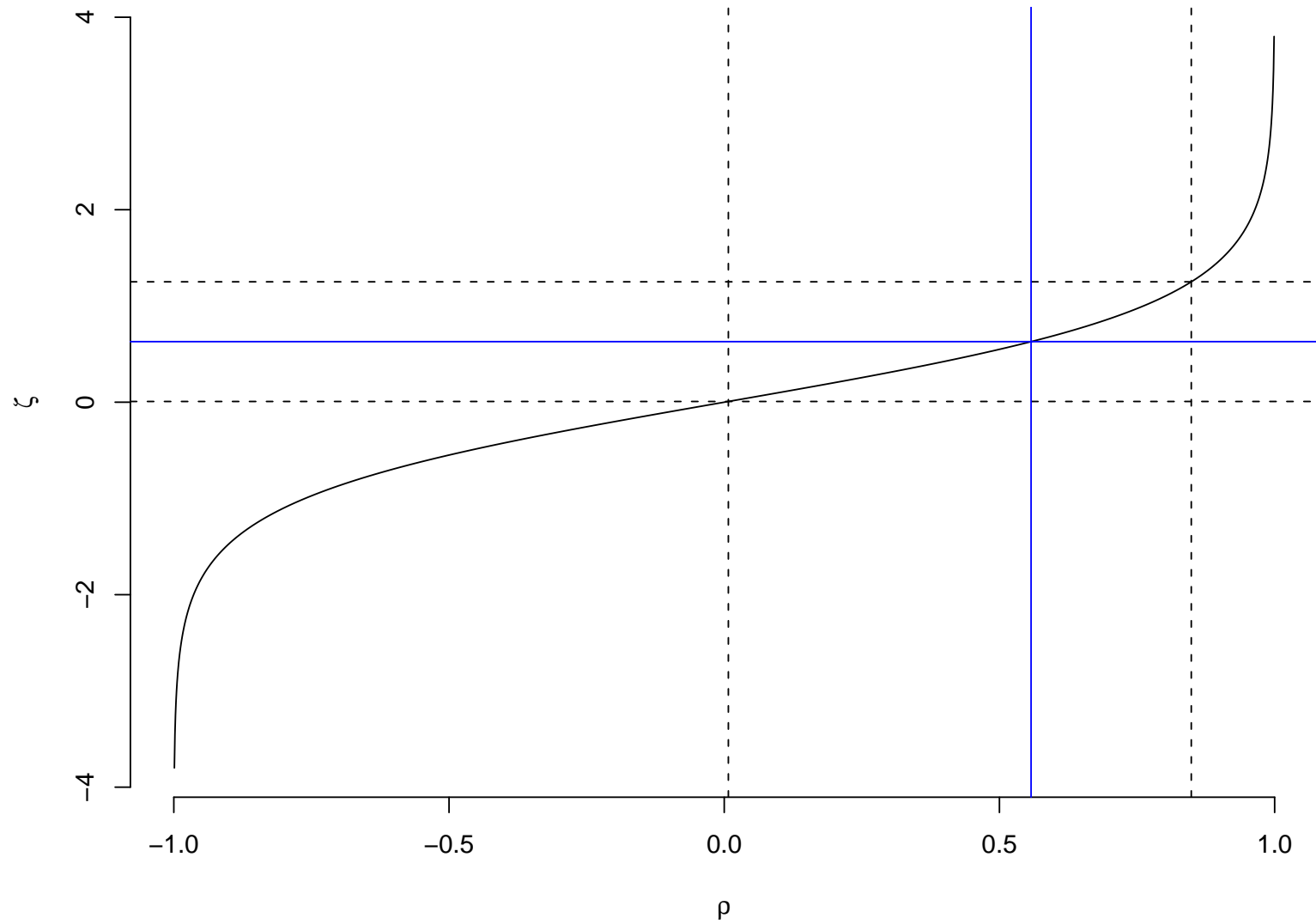
since $e^{2\zeta} = \frac{1+\rho}{1-\rho} \implies \frac{e^{2\zeta} - 1}{e^{2\zeta} + 1} = \frac{\frac{1+\rho}{1-\rho} - 1}{\frac{1+\rho}{1-\rho} + 1} = \frac{1+\rho - (1-\rho)}{1+\rho + (1-\rho)} = \rho$

If $[\zeta_L, \zeta_U]$ is an approximate $(1 - \alpha)$ -level confidence interval for ζ we get in

$$\left[f^{-1}(\zeta_L), f^{-1}(\zeta_U) \right] = \left[\frac{e^{2\zeta_L} - 1}{e^{2\zeta_L} + 1}, \frac{e^{2\zeta_U} - 1}{e^{2\zeta_U} + 1} \right]$$

an approximate $(1 - \alpha)$ -level confidence interval for ρ .

Illustration of a Sample of Size $n = 10$



Confidence Intervals for ρ

The R function `cor.test` carries out this procedure.

```
> pop <- c(10,20,4,16,.85) # setting the population parameters
> set.seed(37)             # setting the RNG seed
> out <- binorm.sample(pop,10) # getting the bivariate sample
> cor.test(out[,1],out[,2],conf.level=.9)
```

Pearson's product-moment correlation

```
data: out[, 1] and out[, 2]
t = 1.8994, df = 8, p-value = 0.09406
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.00748991 0.84853093
sample estimates:
      cor
0.5574917
```