SPEARMAN'S RANK CORRELATION

A measure of association for two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ is the **Pearson Product Moment Correlation Coefficient**, r, where

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}\, SS_{yy}}}$$

where

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad SS_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

This quantity measures the **linear** association between the X and Y variables. A more general measure of association between the samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ is the **Spearman Rank Correlation Coefficient**, r_S, which computes the correlation between the **ranks** of the data.

The Spearman Rank Correlation Coefficient is computed as follows:

1. Assign ranks $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ to the data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ separately by sorting each sample into ascending order and assigning the ranks in order.

2. Compute r_S as

$$r_S = \frac{SS_{uv}}{\sqrt{SS_{uu}\, SS_{vv}}}$$

   where

$$SS_{uu} = \sum_{i=1}^{n}(u_i - \overline{u})^2 \qquad SS_{vv} = \sum_{i=1}^{n}(v_i - \overline{v})^2 \qquad SS_{uv} = \sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v})$$

If there are no ties in the data, then

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \text{where} \qquad d_i = u_i - v_i \qquad i = 1, \ldots, n$$

Tests for r_S : If the population correlation is ρ, then we may test the hypothesis

$$H_0 \ : \ \rho = 0$$

against different hypotheses using the table of the null distribution in McClave and Sincich. If Spearman_α is the α tail quantile of the null distribution

$H_a \ : \ \rho > 0$	Reject H_0 if $r_S > \text{Spearman}_\alpha$
$H_a \ : \ \rho < 0$	Reject H_0 if $r_S < -\text{Spearman}_\alpha$
$H_a \ : \ \rho \neq 0$	Reject H_0 if $\lvert r_S \rvert > \text{Spearman}_{\alpha/2}$

Country	Latitude x	Rank u	DZT Rate y	Rank v
Portugal	40	1.5	6.5	2.0
Greece	40	1.5	8.8	13.0
Spain	41	3.0	5.9	1.0
Bulgaria	42	4.0	7.0	3.0
Italy	44	5.0	8.6	11.5
France	47	6.5	7.1	4.0
Switzerland	47	6.5	8.1	7.5
Austria	48	8.0	7.5	6.0
Belgium	51	9.5	7.3	5.0
FR Germany	51	9.5	8.2	9.0
Holland	52	11.5	8.1	7.5
GDR	52	11.5	9.1	16.0
England & Wales	53	13.5	8.9	14.5
Ireland	53	13.5	11.0	18.0
Scotland	56	15.5	8.9	14.5
Denmark	56	15.5	9.6	17.0
Sweden	60	17.0	8.6	11.5
Norway	61	18.0	8.3	10.0
Finland	62	19.0	12.1	19.0

EXAMPLE Latitude and dizygotic twinning rates
The relationship between the geographical latitude of a country and its dizygotic twinning (DZT) rate is to be investigated. The data are presented and plotted below.

Reference: James, W.H. (1985) Dizygotic twinning, birth weight and latitude, *Annals of Human Biology*, **12**, 5, pp. 441-447.

For these data

$$r_S = \frac{SS_{uv}}{\sqrt{SS_{uu}\,SS_{vv}}} = \frac{384.5}{\sqrt{567 \times 568.5}} = 0.677 \qquad r = \frac{SS_{xy}}{\sqrt{SS_{xx}\,SS_{yy}}} = \frac{118.4}{\sqrt{866.105 \times 38.88}} = 0.645$$

indicating a strong positive association.

```
> cor(DZT$Lat,DZT$DZTRate,method="spearman")

+ [1] 0.677235

> cor(DZT$Lat,DZT$DZTRate,method="pearson")

+ [1] 0.645213
```

```
> par(mar=c(4,4,2,2),pty='s')
> plot(DZT$Lat,DZT$DZT,pch=20,xlab="Latitude (degrees)",
+      ylab="DZT Rate (per 1000, standardized)",
+      xlim=range(37,65),ylim=range(5,12.5))
> DZT.ylab<-DZT$DZT+0.25
> DZT.ylab[c(11,13,18)]<-DZT.ylab[c(11,13,18)]-0.45
> text(DZT$Lat,DZT.ylab,DZT$Country,cex=0.6)
```

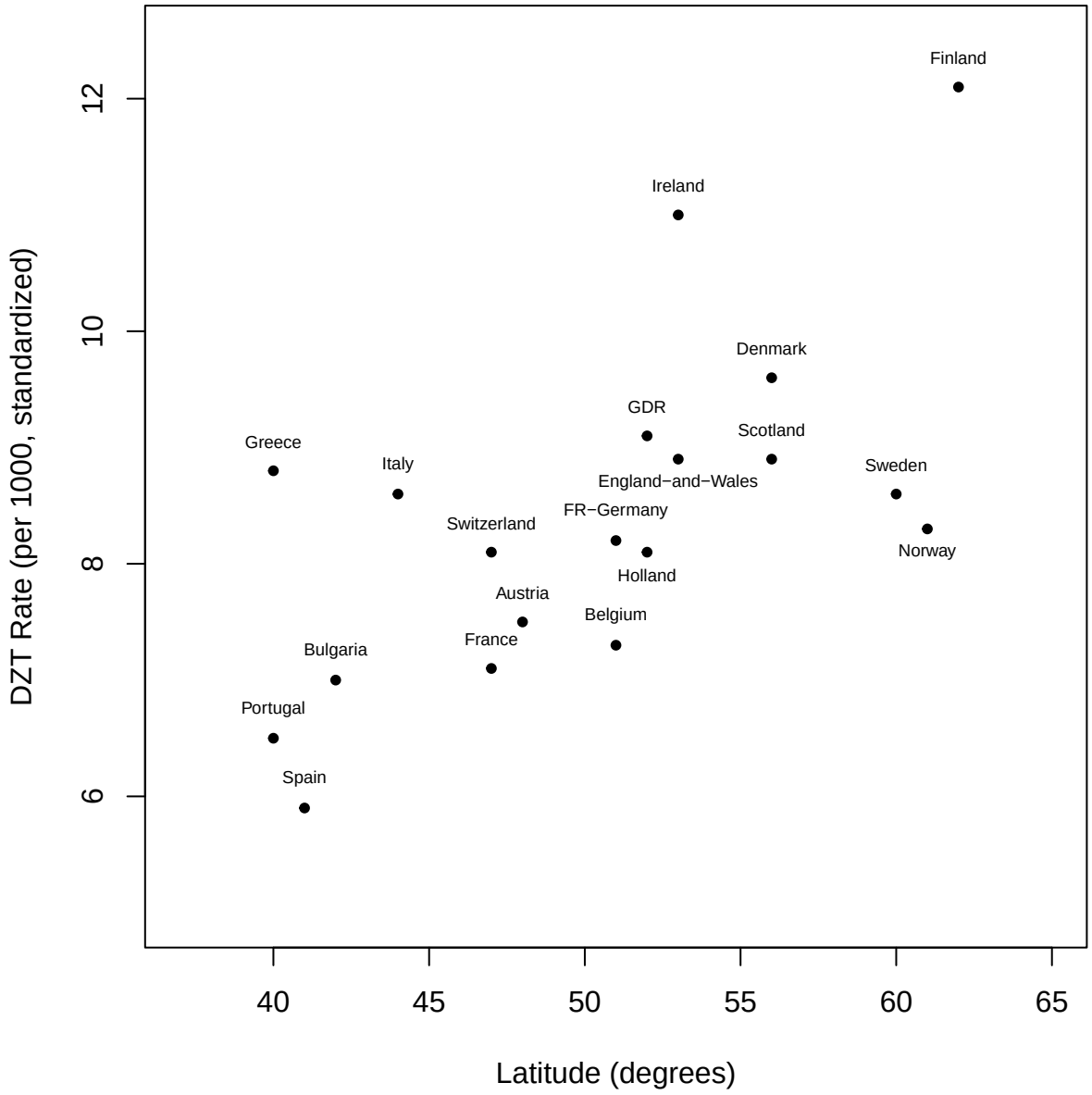


Figure 27: Latitude and dizygotic twinning rates: raw data

RANDOMIZATION AND PERMUTATION PROCEDURES

Randomization or **Permutation** procedures are useful for computing **exact** null distributions for non-parametric test statistics when sample sizes are small.

Suppose that two data samples $x_1 \ldots, x_{n_1}$ and $y_1 \ldots, y_{n_2}$ (where $n_1 \geq n_2$) have been obtained, and we wish to carry out a comparison of the two populations from which the samples are drawn. The Wilcoxon test statistic, W, is the sum of the ranks for the second sample. The permutation test proceeds as follows:

1. Let $n = n_1 + n_2$. If there are no ties, the pooled and ranked samples will have ranks
$$1 \quad 2 \quad 3 \quad \ldots \quad n$$
2. The test statistic is $W = R_2$, the rank sum for sample two items. For the observed data, W will be the sum of n_2 of the ranks given in the list above.
3. If the null hypothesis
$$H_0 \ : \text{ No difference between population 1 and population2}$$
were **true**, then there should be **no pattern** in the group labels when sorted into ascending order; the sorted data would give rise a **random** assortment of group 1 and group 2 labels.
4. To obtain the exact distribution of W under H_0 (for the assessment of statistical significance), we could compute W for all possible permutations of the group labels, and then form the probability distribution of the values of W. We call this the **permutation null distribution**.
5. But W is a rank sum, so we can compute the permutation null distribution simply by tabulating **all possible subsets** of size n_2 of the set of ranks $\{1, 2, 3, \ldots, n\}$.
6. There are
$$\binom{n}{n_2} = \frac{n!}{n_1!\, n_2!} = N$$
say possible subsets of size n_2; for $n = 6$ and $n_2 = 2$, the number of subsets of size n_2 is
$$\binom{8}{2} = \frac{8!}{6!\, 2!} = 28$$
However, the number of subsets increases dramatically as n increases; for $n_1 = n_2 = 10$, so that $n = 20$, the number of subsets of size n_2 is
$$\binom{20}{10} = \frac{20!}{10!\, 10!} = 184756$$
7. The exact rejection region and p-value are computed from the permutation null distribution. Let $W_i, i = 1, \ldots, N$ denote the value of the Wilcoxon statistic for the N possible subsets of the ranks of size n_2. The probability that the test statistic, W, is less than or equal to w is
$$\Pr[W \leq w] = \frac{\text{Number of } W_i \leq w}{N}$$
We seek the values of w that give the appropriate rejection region, $\mathcal{R}$, so that
$$\Pr[W \in \mathcal{R}] = \frac{\text{Number of } W_i \in \mathcal{R}}{N} = \alpha$$
It may not be possible to find critical values, and define $\mathcal{R}$, so that this probability is **exactly** α as the distribution of W is **discrete**.

EXAMPLE Simple Example
Suppose $n_1 = 7$ and $n_2 = 3$. There are

$$\binom{10}{3} = \frac{10!}{7!\,3!} = 120$$

subsets of the ranks $\{1, 2, 3, \ldots, 10\}$ of size 3. The subsets and rank sums are listed below:

Ranks			W	Ranks			W	Ranks			W	Ranks			W
1	2	3	6	1	7	8	16	2	7	10	19	4	6	7	17
1	2	4	7	1	7	9	17	2	8	9	19	4	6	8	18
1	2	5	8	1	7	10	18	2	8	10	20	4	6	9	19
1	2	6	9	1	8	9	18	2	9	10	21	4	6	10	20
1	2	7	10	1	8	10	19	3	4	5	12	4	7	8	19
1	2	8	11	1	9	10	20	3	4	6	13	4	7	9	20
1	2	9	12	2	3	4	9	3	4	7	14	4	7	10	21
1	2	10	13	2	3	5	10	3	4	8	15	4	8	9	21
1	3	4	8	2	3	6	11	3	4	9	16	4	8	10	22
1	3	5	9	2	3	7	12	3	4	10	17	4	9	10	23
1	3	6	10	2	3	8	13	3	5	6	14	5	6	7	18
1	3	7	11	2	3	9	14	3	5	7	15	5	6	8	19
1	3	8	12	2	3	10	15	3	5	8	16	5	6	9	20
1	3	9	13	2	4	5	11	3	5	9	17	5	6	10	21
1	3	10	14	2	4	6	12	3	5	10	18	5	7	8	20
1	4	5	10	2	4	7	13	3	6	7	16	5	7	9	21
1	4	6	11	2	4	8	14	3	6	8	17	5	7	10	22
1	4	7	12	2	4	9	15	3	6	9	18	5	8	9	22
1	4	8	13	2	4	10	16	3	6	10	19	5	8	10	23
1	4	9	14	2	5	6	13	3	7	8	18	5	9	10	24
1	4	10	15	2	5	7	14	3	7	9	19	6	7	8	21
1	5	6	12	2	5	8	15	3	7	10	20	6	7	9	22
1	5	7	13	2	5	9	16	3	8	9	20	6	7	10	23
1	5	8	14	2	5	10	17	3	8	10	21	6	8	9	23
1	5	9	15	2	6	7	15	3	9	10	22	6	8	10	24
1	5	10	16	2	6	8	16	4	5	6	15	6	9	10	25
1	6	7	14	2	6	9	17	4	5	7	16	7	8	9	24
1	6	8	15	2	6	10	18	4	5	8	17	7	8	10	25
1	6	9	16	2	7	8	17	4	5	9	18	7	9	10	26
1	6	10	17	2	7	9	18	4	5	10	19	8	9	10	27

There are 22 possible rank sums, $\{6, 7, 8, \ldots, 25, 26, 27\}$; the number of times each is observed is displayed in the table below, with the corresponding probabilities and cumulative probabilities.

W	6	7	8	9	10	11	12	13	14	15	16
Frequency	1	1	2	3	4	5	7	8	9	10	10
Prob.	0.008	0.008	0.017	0.025	0.033	0.042	0.058	0.067	0.075	0.083	0.083
Cumulative Prob.	0.008	0.017	0.033	0.058	0.092	0.133	0.192	0.258	0.333	0.417	0.500

W	17	18	**19**	20	21	22	23	24	25	26	27
Frequency	10	10	**9**	8	7	5	4	3	2	1	1
Prob.	0.083	0.083	**0.075**	0.067	0.058	0.042	0.033	0.025	0.017	0.008	0.008
Cumulative Prob.	0.583	0.667	0.742	0.808	0.867	0.908	0.942	0.967	0.983	0.992	1.000

Thus, for example, the probability that $W = 19$ is 0.075, with a frequency of 9 out of 120. From this table:

$$\Pr[8 \leq W \leq 25] = 0.983 - 0.017 = 0.966$$

implying that the two-sided rejection region for $\alpha = 0.05$ is the set $\mathcal{R} = \{6, 7, 26, 27\}$.

For these sample sizes, the exact calculation can be performed using the `combinat` library

```
> n1<-7;n2<-3;n<-n1+n2
> Rvec<-c(1:n)
>
> #use the combinat library
> Cmat<-combinat::combn(n,n2)
> Scores<-apply(Cmat,2,sum)
> Prob<-as.numeric(table(Scores))/length(Scores)
> CProb<-cumsum(Prob)
> par(mar=c(4,4,1,2))
> plot(unique(Scores),Prob, ylim=range(0,0.15),
+     pch=18,xlab="W",ylab="Prob.")
```

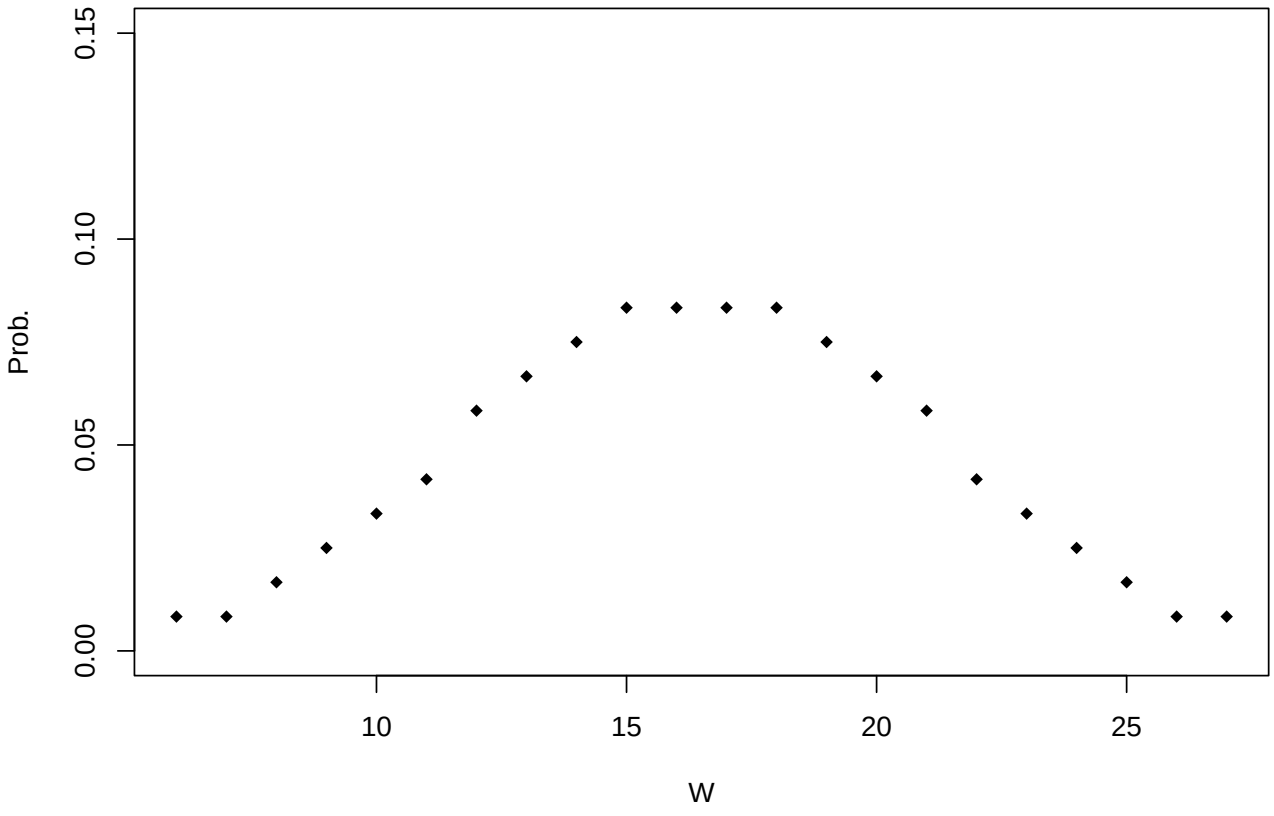


Figure 28: Permutation pmf

```
> par(mar=c(4,4,1,2))
> plot(unique(Scores),CProb,, ylim=range(0,1),
+     pch=18,xlab="W",ylab="Cumulative Prob.")
```

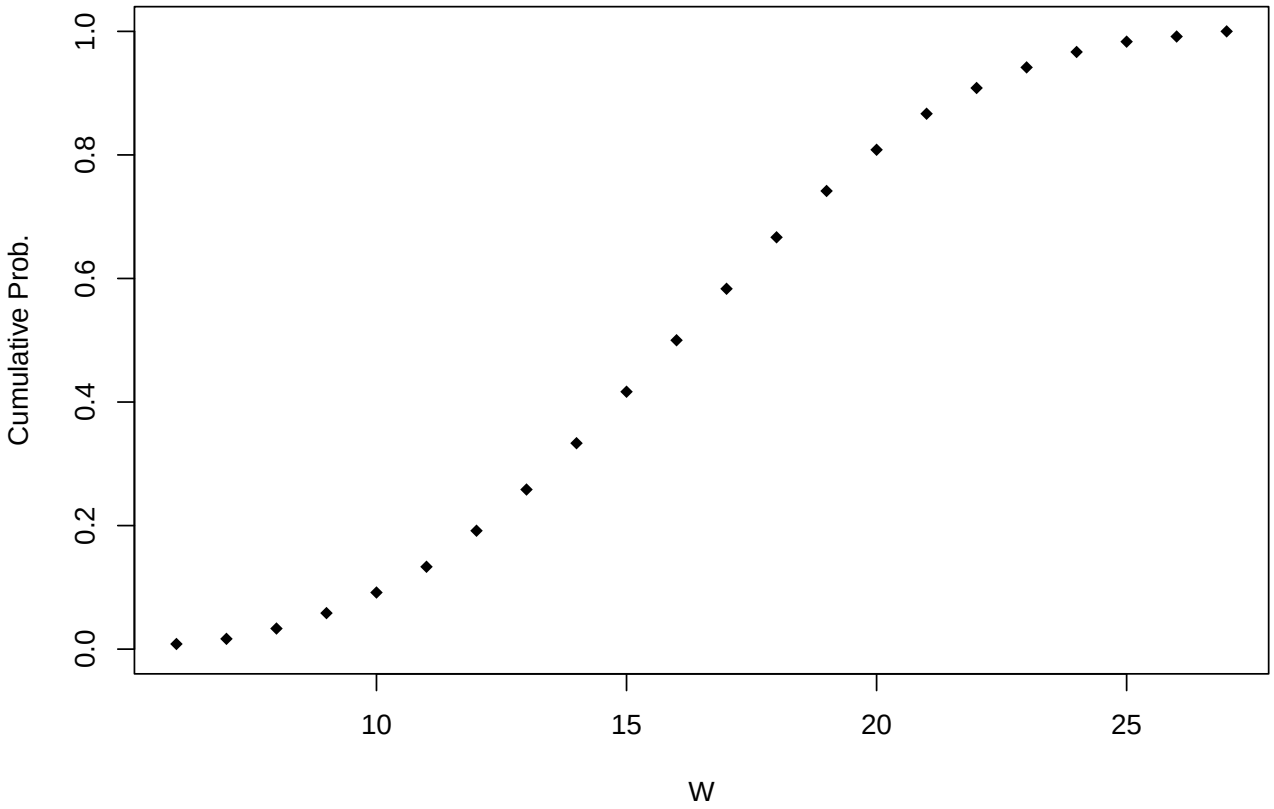


Figure 29: Permutation cdf