Cumulative Distribution Function
or Distribution Function (CDF) → always area to left

$P(X \le x)$ value
random variable

$F(x) = P(X \le x)$

$F(2) = P(X \le 2)$ $\quad P(Z \le 1.96) = 0.95$

8/29 Let $X_1 \dots X_n$ follows a distribution f (unknown)

Assumption:

(1) The distribution is continuous

Always median for distribution
there may not be a mean

To test for median (θ)

$H_0: \theta = \theta_0$

$H_A: \theta > \theta_0$

distribution shape doesn't matter

Look at sign of $X_i - \theta_0$ $\quad X_i$ = each observation $\quad \forall\, i = 1, 2, \dots n$ (for all)

If $X_i > \theta_0$ then $X_i - \theta_0$ is $+$

If $X_i < \theta_0$ then $X_i - \theta_0$ is $-$

We would observe a sequence of $+$ and $-$ signs.

Consider the random variable Y_i which is such that

$$Y_i = \begin{cases} 1 & \text{if } X_i - \theta_0 > 0 \\ 0 & \text{if } X_i - \theta_0 < 0 \end{cases}$$

Hence under H_0 (if θ_0 is the true median)

$$P[Y_i = 1] = 0.5 = P[X_i - \theta_0 > 0]$$

$\therefore$ we consider the random variable $Y = \sum_{i=1}^{n} Y_i$ and $Y \sim Bin(n, 0.5)$ under H_0.

Test $H_0: p = 0.5$
$H_A: p > 0.5$

changed into a binomial dist. - original dist doesn't matter
So distribution free test

Reject H_0 at a level of significance α if $\sum_{i=1}^{n} y_i > c \ni$

$$P\left[\sum_{i=1}^{n} y_i > c \,\middle|\, p = 0.5\right] \le \alpha$$

Example 2.2 (pg-26-27) Weeks: 49, 58, 75, 110, 112, 132, 151, 276, 281, 362

$H_0: \theta = 200$

$H_A: \theta \neq 200$

x ← c( )

qqnorm (x)

shapiro.test (x)

p-value = 0.105

borderline, use nonparametric cuz fewer assumptions

Sign Test

look at signs $Y = 3$

$H_0: p = 0.5$

$H_A: p \neq 0.5$

Under H_0, $Y \sim Bin(10, 0.5)$

$P[Y=y] = \binom{10}{y}(0.5)^{10}$

y ← ifelse(x > 200, 1, 0)

Y ← sum (y)

p-value

$2 \times P[Y \leq 3]$ dbinom

$2 \times 0.172 = 0.3438$

do not reject

9/12 $F(x) = P[X \le x] = \begin{cases} \frac{1}{50}x^2, & 0 \le x \le 5 \\ \frac{2}{5}x - \frac{1}{50}x^2 - 1, & 5 < x \le 10 \end{cases}$ 2.11

R

uniform graph: curve (x/10, 0, 10)
curve (x²/50, 0, 5, add=T
col="blue")
curve(2/5x − x² − 1, 5, 10, add=T
col="blue")

$n = 12$ $\frac{12 \times 13}{2} = 78$

symmetric about 39

ex.) $H > 15$

$S_+ = 14$ $S_- = 64$

↖ in table $P[S_+ \le 14) = .026$

Not equal - two tailed test $P[S_+ \le 14) = 2(0.026)$

$N = 12$ $H_0: \theta = 15$

$S_+ = 14$ $H_A: \theta \ne 15$

Acymptotic results:

If S is either S_+ or S_-

$E(S_+) = \frac{n(n+1)}{4}$

$V(S_+) = \frac{n(n+1)(2n+1)}{24}$

$V(x) = E(x^2) - E^2(x)$

$Z = \frac{S_+ - 39}{12.75} \sim N(0,1)$ by Central Limit Theorem

$E(S_+) = \frac{12(13)}{4} = 39$

$V(S_+) = \frac{12(13)(25)}{24} = 162.5$

$Z = \frac{S_+ - E(S_+)}{\sqrt{V(S_+)}} \sim N(0,1)$ for large N

$Z = \frac{S_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$

$N = 12$ is okay

$N = 4$ not so much

Table H

observed $S_+ = 14$ $\quad Z = \frac{14-39}{12.75} = 1.96$ using continuity correction $\frac{14.5-39}{12.75} = -1.92$

9/14 Confidence Intervals

Walsh averages

$$X_1 \le X_2 \le X_3 \le X_4$$

$$\frac{X_1+X_1}{2}, \frac{X_1+X_2}{2}, \frac{X_1+X_3}{2}, \frac{X_1+X_4}{2} \qquad \frac{X_i+X_j}{2}, \; j \ge i$$

$$\frac{X_2+X_2}{2}, \frac{X_2+X_3}{2}, \frac{X_2+X_4}{2}$$

$\vdots$

point estimate for population median = median of the Walsh averages

Walsh average R code

```
x =
y = sort(x)
mat = matrix(NA, n, n)
for(j in (i+1):n)
{  mat[i,i] = y[i]
 mat[i,j] = [y[i] + y[j])/2
}
}

for(i in 1:n)
{
```

$\binom{n}{2} + n$

$= \frac{n(n+1)}{2}$

total # of walsh averages

$x =$

$m_0 = 15$ - R code on canvas

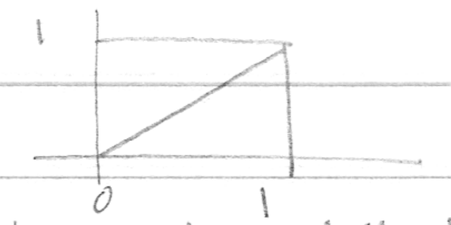
```
d = x - m0
abs(d)
rank(abs(d))
z = ifelse(d>0,1,0)
cbind(d, abs(d), rank(abs(d)), z)
Splus = sum(rank(abs(d))*z)
Sminus = sum(rank(abs(d))*(1-z))
```

CI - find 0.025 in table
lower bound is one above it

PP plot – $X_1 \dots X_n$

order $X_{(1)} \dots X_{(n)}$

$X_1 \quad F_1(x_1) \quad G_1(x_1)$



two cdfs plotted

Nonparametric in the sense that linearity is being tested

Test for Randomness

- Based on runs

Too few (or many) runs reject randomness

M M | M M | M

How many runs? (gaps / bars)

Result: Let there be n items of 1 type and r boxes. How many distinguishable ways are there to put n items in r boxes so that no box is empty? $(n \geq r)$ 7 items, 4 boxes

$\binom{n-1}{r-1}$

Result 2: Let there be n_1 items of type 1 and n_2 items of type 2. The number of ways we can have r_1 runs of type 1 & r_2 runs of type 2 $= \binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}$

$r_1 \quad r_2 = r_1$ or $r_2 = r_1 \pm 1$

consider cases where $|r_1 - r_2| = \pm 1$, $r_1 = r_2$

If there are n_1 items of type 1, then total # of ways to have r_1 runs is $\binom{n_1-1}{r_1-1}$

9/26 Goodness of Fit

K-S - Lilliefor's (estimate the parameters)

$(F_{(n)}x - F_{(0)}x)$ $\qquad$ $F_n(x) - \widehat{F_0(x)}$

D_n $\qquad$ D_n

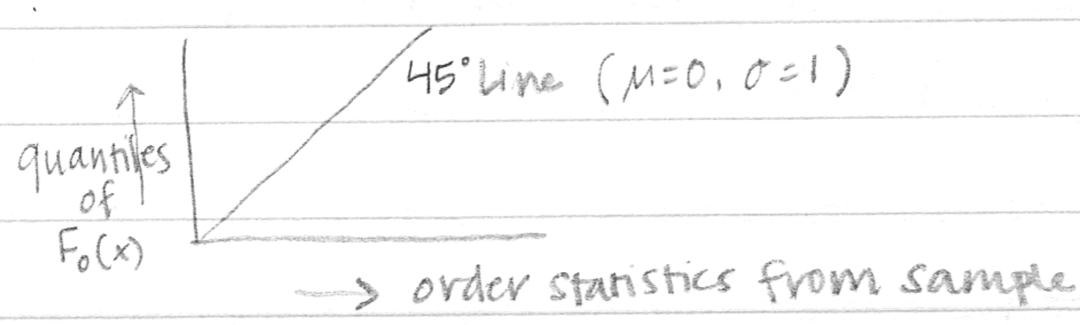
Test for normality

- Anderson-Darling (Cramer-von-miss)
- Shapiro - Wilk

$H_0: F_X(x) = F_0(x)$ $\qquad$ $X_1, X_2, \ldots X_n \sim F_X(x)$

$H_A: F_X(x) \neq F_0(x)$

QQ plot



Corresponding to the percentile p, let the quantile be $F_0^{-1}(p)$ under H_0 when $0<p<1$.

$\Phi(0) = \frac{1}{2}$

$\Phi^{-1}(\frac{1}{2}) = 0$

The plot of the order statistics against $F_0^{-1}(p)$ will be the qq plot.

Example - Testing for normality (unknown μ & σ)

$$P[X \leq x] = P[Z \leq x-\mu/\sigma] = \Phi(x-\mu/\sigma)$$

If $X \sim N(\mu, \sigma^2)$ then $F_X(x) = \Phi(x-\mu/\sigma)$

Let $F_X(x) = p$ then $\Phi(x-\mu/\sigma) = p \Rightarrow \frac{x-\mu}{\sigma} = \Phi^{-1}(p)$

$\Rightarrow x = \mu + \sigma\Phi^{-1}(p)$

relationship would be linear as in qq plot σ = slope, μ = intercept

- order samples

i	$x_{(i)}$	$\frac{i}{n+1}$ (proportion)	$\Phi^{-1}(\frac{i}{n+1})$

plot $x_{(i)}$ vs. $\Phi^{-1}(\frac{i}{n+1})$

plot $\rightarrow$ then fit regression line

10/3 More than one category $\sum_1^k n_i = n$

calculate proportion $p_i = \frac{n_i}{n}$

asymptotic

ex from book AAAABACCAAAAADDAA

$n=17$ $n_1=1$ $n_2=2$ $n_3=2$ $n_4=12$ $R=7$

$p_1=\frac{1}{17}$ $p_2=\frac{2}{17}$ $p_3=\frac{2}{17}$ $p_4=12/17$

$E(R) = n\left(1-\sum_1^k p_i^2\right)+1$

$V(R) = n\left(\sum(p_i^2-2p_i^3)+\left(\sum_{i=1}^k p_i^2\right)^2\right)$ $Z = \frac{R-E(R)}{\sqrt{V(R)}}$

reject for small values of R

Two-Sample Problems

standardized form $\frac{X-E(x)}{\sqrt{V(x)}}$

R

rchisq

χ^2_n $E(x)=n$ $V(x)=2n$

standardize χ^2 $\frac{X-n}{\sqrt{2n}}$ X= rchisq(500, 10)

Example z= (x-10) / sqrt(2·10)

$X \sim N(0,1)$

$Z = \frac{Y-E(Y)}{\sqrt{V(Y)}}$ ~ standardized chi square

$H_0: F_X(t) = F_Y(t)\ \forall t$ $X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8$

$H_A: F_X(t) \neq F_X(t)$ $Y_1\ Y_2\ Y_3\ Y_4\ Y_5\ Y_6\ Y_7\ Y_8$

Run test: combine to make one sample

arrange in ascending order

R

X = c(-1.91, -1.22, -0.96, -0.72, 0.14, 0.82, 1.45, 1.86)

y = c(4.9, 7.25, 8.04, 14.10, 18.30, 21.21, 23.1, 28.12)

V=18

testing between
X & Z

$Z = (y - 18) / 6$
$XZ = c(X, Z)$
sort(XZ)

From $\chi^2 = 0$, normal = 1

0 1 0 0 1 1 1 0 0 1 0 1 0 1 0 1 R = 12 n = 16

when H_0 is false, expect
smaller values of R R = total # of runs

under H_0 = large runs
this is left-tail test

Table D - runs $n_1 = 8$ $n_2 = 8$
at 0.05 $R \leq 5$ critical value = reject
When R = 12 do not reject H_0

Wald-Wolfitz Run Test

H_0: two dist. identical?
* doesn't matter which distribution

Different technique - Kolmogrov-Smirnov test for Two Samples
$D_{nm} = \max |F_{1n}(x) - F_{2n}(x)|$ * don't need same sample size

- Mix all values, then sort observations

t	# X ≤ t	$F_m(t)$	# Y ≤ t	$F_n(t)$	$F_m(t) - F_n(t)$
-2.18	0	0	1	1/8	-1/8
-1.91	1	1/8	1	1/8	0
-1.79	1	1/8	2	2/8	-1/8
-1.66	1	1/8	3	3/8	-2/8

F_n: ECDF of X
F_m: " Y

from sorted xz list

then take |d|

10/10 $U_m = W_{mn} - \frac{m(m+1)}{2}$ (x-values) $U_m + U_n = mn$

$U_n = W_{nm} - \frac{n(n+1)}{2}$ (y-values)

+

$$U_m + U_n = \frac{(m+n)(m+n+1)}{2} - \frac{m(m+1)}{2} - \frac{n(n+1)}{2} = mn$$

Ties

ex. A: 15.7, 14.8, 14.8, 16.1, 15.3, 13.9, 17.2, 14.9

B: 13.7, 14.1, 14.7, 15.4, 15.6, 16.1, 14.8, 12.9, 15.1, 14

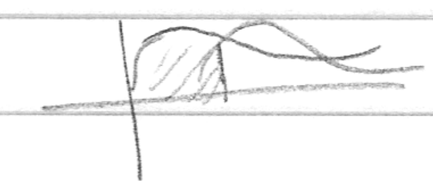
how many less than

$U_m = 9 + 5.5 + 5.5 + 9.5 + \ldots$ $D_{ij} = \begin{cases} 1 & y < x_i \\ 0.5 & y_i = x_i \\ 0 & y_j > x_j \end{cases}$

Median Test

$H_0: P[X<\delta] = P[y<\delta]$

$H_A:$ $>$



Let U be # of observations in $x_1, \ldots x_m$ that are less than the median

The distribution of U is given by $P[U=u] =$

Let t be the # of obs. < the median in the combined sample

$x_1 \ldots x_m$ & $y_1 \ldots y_n$ median $= \frac{m+n}{2}$

Rejection region: less than median $\longrightarrow \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{m+n}{t}}$ hypergeo.

If y is stochastically larger than x (ie if the difference in the median between y & x) is positive then large values of U will reject H_0.

$t-u \geq 0$ $u = 0, 1 \ldots \min(m,t)$

$u \leq t$ $u \leq m$

10/12 Ex. Sample 1: 3, 4, 9, 10 (x) ← # of nonconforming items

Sample 2: 1, 2, 5, 7, 8 (y)

$H_0: \theta_x = \theta_y$

$H_A: \theta_x > \theta_y$

$N = m+n$

m = size of x values

n = size of y values

t = # of obs. < median of combined sample

$t = \begin{cases} 0, \ldots, N/2 & \text{if } N \text{ is even} \\ 0, \ldots, N-1/2 & \text{if } N \text{ is odd} \end{cases}$

U = # of X obs less than median of combined sample

$$P(U=u) = \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{m+n}{t}}$$

If x is stochastically larger than Y (if median of x is higher than median of Y), H_0 will be rejected for smaller values of U.

ex from above $m=4$, $n=5$ combined: 1, 2, 3, 4, 5, 7, 8, 9, 10

$t=4$

$U=2$

$$P(U \le 2 \mid H_0) = \frac{\binom{4}{0}\binom{5}{4} + \binom{4}{1}\binom{5}{3} + \binom{4}{2}\binom{5}{2}}{\binom{9}{4}}$$

$$= \frac{105}{126} = \text{p-value}$$

cannot reject

Not on This test

$$P(V=j) = \frac{\binom{r+j}{j}\binom{r+n-j}{n-j}}{\binom{m+n}{n}}$$

$m = 2r$

$n = 2r+1$

$y_1 \ldots y_{2r+1}$

Two-Sample Scale Problem (Spread)

(parametric = F-test)

Assume median is the same

distribution of ranks

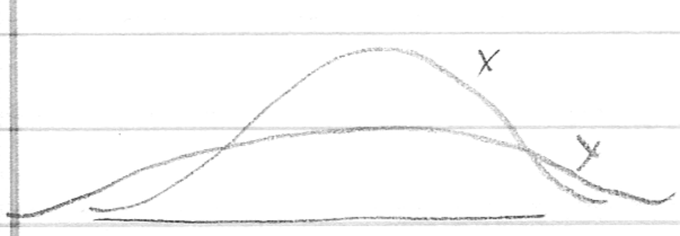
Ansari-Bradley Test - David-Barton

$d_i = i - \frac{N+1}{2}$

i = rank of the ith order statistic

$N = n + m$

$\frac{N+1}{2}$ = mid-rank



Test statistic: S = sum of $|d_i|$'s corresponding to x sample

ex A: 35, 66, 58, 83, 71

B: 46, 56, 60, 49 equal variability?

$n = 4$

$m = 5$

$N = 9$ mid-rank = 5

order: 35, 46, 49, 56, 58, 60, 66, 71, 83 ranks

$S = |2-5| + |3-5| + |4-5| + |6-5| = 3+2+1+1 = 7$

(S values large if variability is different)

In R: exactRankTests

ansari.test(x, y, exact=T)

<u>Siegel-Tukey Test</u> test of spread assume same median

order: 35, 46, 49, 56, 58, 60, 66, 71, 83 $m=4$

(rank) as such: 1 4 5 8 9 7 6 3 2 $n=5$

$S_n = 24$ S_n has same distribution as Wilcoxon Rank Sum test statistic - use Table J

sum of ranks

p-value = 0.412 = 2 × 0.206

10/24 Kruskall-Wallis test review

1	2	3	. . .	k
X_{11}	X_{21}	X_{31}		X_{k1}
X_{12}	X_{22}			
$\vdots$	$\vdots$			
X_{1n}				

$H_0: \theta_1 = \theta_2 \ldots = \theta_k$

H_A: At least one inequality

$N = \sum_1^k n_i$ $\qquad$ $t = \#$ obs < median of combined

$u_i = \#$ obs < t in i^{th} group i

$$P(u_1 = u_1, u_2 = u_2 \ldots u_k = u_k) = \frac{\binom{n_1}{u_1}\binom{n_2}{u_2}\binom{n_3}{u_3}\cdots\binom{n_k}{t - \sum_1^{k-1} u_i}}{\binom{N}{t}}$$

reject if one group greater than rest

Approximate Test

χ^2 test

Goodness of Fit

f_{ij} = obs frequency of jth category in i^{th} sample

e_{ij} = expected " "

n_i = size of ith sample (category)

$$\sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}} \sim \chi^2_{k-1}$$

$k-1$ ↑ group

test statistic: $\displaystyle\sum_1^k \sum_1^2 \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$ $\qquad$ $N = \sum_1^k n_i$

either above or below median

$$= \sum_{i=1}^k \left[\frac{(f_{i1} - e_{i1})^2}{e_{i1}} + \frac{(f_{i2} - e_{i2})^2}{e_{i2}} \right] = \sum_{i=1}^k \left[\frac{(u_i - n_i \frac{t}{N})^2}{n_i \frac{t}{N}} + \frac{\left[(n_i - u_i) - \frac{n_i(N-t)}{N}\right]^2}{\frac{n_i(N-t)}{N}} \right]$$

$$= N \sum_{i=1}^k \left[\frac{(u_i - n_i t/N)^2}{n_i t} + \frac{(\frac{n_i t}{N} - u_i)^2}{n_i (N-t)} \right]$$

$$= N \sum_{i=1}^k \left[\frac{(u_i - \frac{n_i t}{N})^2}{n_i t} + \frac{(u_i - \frac{n_i t}{N})^2}{n_i (N-t)} \right] = N \sum \frac{(u_i - \frac{n_i t}{N})}{n_i} \left(\frac{1}{t} + \frac{1}{N-t} \right)$$

$$= \frac{N^2}{t(N-t)} \sum_{i=1}^k \frac{(u_i - \frac{n_i t}{N})^2}{n_i}$$

$$= Q$$

under H_0, $Q \sim \chi^2_{k-1}$

see book pg. 203-206

$$S = \sum_{i=1}^{t} R_i^2 - \frac{b^2 t (t+1)^2}{4}$$

$$E(S) = \frac{bt(t^2-1)}{12} \qquad V(S) = \frac{t^2 b(b-1)(t+1)^2}{72}(t-1)$$

$$E\left(\frac{12S}{bt(t+1)}\right) = t-1 \qquad V\left(\frac{12S}{bt(t+1)}\right) = \frac{2(b-1)}{b}$$

$$T = \frac{12S}{bt(t+1)} = \frac{12\sum R_i^2}{bt(t+1)} - 3b(t+1) \qquad V(T) = 2(t-1)\left(1-\frac{1}{b}\right) = 2(t-1)$$

as $b \to \infty$

Ex. Questioning children — rank within block

	1	2	3	4
1	1	2	3	4
2	1	2.5	2.5	4
3	1	2	4	3
4	1	2	4	3
5	1	2	3	4
6	1	2	3	4
7	1	2.5	4	2.5
8	1	2	3	4
	8	17	26.5	28.5

block on day — ties okay using χ^2

10/31 $H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4$ — Block design = Friedman's

H_A: At least one $\neq$

$R_i^2 = 64 \quad 289 \quad 702.25 \quad 812.25$

$$T = \frac{12\sum_1^t R_i^2}{bt(t+1)} - 3b(t+1) \sim \chi^2_{t-1}$$

$$T = \frac{12(1867.5)}{8(4)(5)} - (3 \times 8 \times 5) = 20.06$$

Table B

$\chi^2 < 0.001$

Reject

With ties $T = \dfrac{b(t-1)(S_t - C)}{S_r - C}$

C = correction factor $= \dfrac{bt(t+1)}{4}$

S_r = sum of squares of ranks $= \frac{1}{b}\sum_1^t R_i^2$

Page Test

$H_0: Q_1 = Q_2 = \dots = Q_t$

$H_A: Q_1 \le Q_2 \dots \le Q_t$

with at least one strict $\neq$

$P = R_1 + 2R_2 + 3R_3 \dots tR_t$

	1	2	. . .	t	
1					rank
2					within
⋮					blocks
b					
	R_1	R_2	...	R_t	

Ex.

	A	B	C	D
1	3	4	2	1
2	4	2	1	3
3	4	2	3	1
4	4	1	3	2
5	2	4	3	1
6	4	3	1	2
R_i	21	16	13	10
W_i	4	3	2	1

H_0:

$H_A: Q_D < Q_C < Q_B < Q_A$

Weight 1, 2, 3, 4

$P = 21(4) + 16(3) + 13(2) + 10(1) = 168$

Large sample test

$E(P) = \frac{tb(t+1)^2}{4}$

$V(P) = \frac{b(t^3-t)^2}{144(t-1)} = \frac{bt^2(t+1)^2(t-1)}{144}$

$Z = \frac{P - E(P)}{\sqrt{V(P)}}$

Table Q (Page's)

$E(P) = 150$

$V(P) = 50$

cont. corr.

$Z = \frac{(168 - 0.5) - 150}{\sqrt{50}}$

$= 2.48$

p-value $= 0.0066$

In R

Friedman.test( )[data in matrix form]

DescTools - PageTest(y, groups, blocks)

	1	2 trt	3	
Data	A		1	Block #
	B		1	
	C		1	
	D		1	

End of Ch. 7

11/2 Chapter 10

Correlation coefficient $r = \frac{cov(x,y)}{S_x S_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$

Some good properties that measures of association should have:

- it should lie between -1 & $+1$
- $+1$ should represent a perfect concordance
  if $x_i < x_j$ then $y_i < y_j$
- -1 should represent a perfect discordance
  if $x_i < x_j$ then $y_i > y_j$
- a value of 0 should indicate "independence"
- the relation between x and y will be the same as $-x$ and $-y$
- the relation between x and y will be the same as y and x (communative)
- the measure should be invariant to order preserving transformations

Ranks also associated

$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$

$R_i = rank(x_i)$ ie intelligence, degree of satisfaction

$S_i = rank(y_i)$

Spearman Rank Correlation

$$r_s = \frac{\sum (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum (R_i - \bar{R})^2}\sqrt{\sum (S_i - \bar{S})^2}}$$

X	R	y	S
x_1	R_1	y_1	S_1
x_2	R_2	y_2	S_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	R_n	y_n	S_n

- Assume no ties -

$\sum_1^n R_i = \frac{n(n+1)}{2}$ $\quad \sum_1^n S_i = \frac{n(n+1)}{2}$

$\bar{R} = \frac{n+1}{2}$ $\quad \bar{S} = \frac{n+1}{2}$

$$\sum_1^n (R_i - \bar{R})^2 = \sum_1^n R_i^2 - 2\bar{R}\sum R_i + \sum_1^n \bar{R}^2 = \frac{n(n+1)(2n+1)}{6} - \frac{2(n+1)}{2} \cdot \frac{n(n+1)}{2} + \frac{n(n+1)^2}{4}$$

$$= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)^2}{4} = \ldots \text{ math } \ldots$$

$$= \frac{n(n^2-1)}{12} = \sum_1^n (R_i - \bar{R})^2 \qquad \sum_1^n (S_i - \bar{S})^2 = \frac{n(n^2-1)}{12}$$

$$\sum_1^n (R_i - \bar{R})(S_i - \bar{S}) = \sum^n R_i S_i - \bar{S}\sum R_i - \bar{R}\sum S_i + n\bar{S}\bar{R} = \sum_1^n R_i S_i - \frac{n(n+1)^2}{4}$$

$$r_s = \frac{\sum R_i S_i - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}} = \frac{12\sum R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1}$$

Example ranking tea 1 = best

Tea	A	B	C	D	E	F	G	H	I
Judge 1	1	5	9	7	4	6	8	2	3
Judge 2	4	3	6	8	2	7	9	1	5

product of ranks: $4 + 15 + 54 + 56 + 8 + 42 + 72 + 2 + 15 = 268$

$n = 9$

$$r_s = \frac{12 \times 268}{9 \times 80} - \frac{30}{8} = 0.716$$

Table M

p-value = .018

11/7 More on concordance

$(x_1,y_1), (x_2,y_2) \ldots (x_n,y_n)$ continuous bivariate

Between a pair (x_i,y_i) & (x_j,y_j) there is concordance if

$x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$

Let p_c be the probability of concordance

$$P_c = P[\{(x_i < x_j) \cap (y_i < y_j)\} \cup \{(x_i > x_j) \cap (y_i > y_j)\}]$$
$$= P[(x_i < x_j) \cap (y_i < y_j)] + P[(x_i > x_j) \cap (y_i > y_j)] \quad (i)$$

When there is discordance: $x_i < x_j$ and $y_i > y_j$ or $x_i > x_j$ and $y_i < y_j$

$$P_D = P[(x_i < x_j) \cap (y_i > y_j)] + P[(x_i > x_j) \cap (y_i < y_j)]$$

$P(A \cap \bar{B}) = P(A) - P(A \cap B)$

$$P[\overset{A}{(x_i < x_j)} \cap \overset{\bar{B}}{(y_i > y_j)}] = P[(x_i < x_j)] - P[(x_i < x_j) \cap (y_i - y_j)]$$
$$P[(x_i > x_j) \cap (y_i < y_j)] = P[(x_i > x_j)] - P[(x_i > x_j) \cap (y_i > y_j)]$$

substitutin in

$$= \underbrace{P[(x_i < x_j)] - P[(x_i < x_j) \cap (y_i - y_j)]}_{=1} + P[(x_i > x_j)] - P[(x_i > x_j) \cap (y_i > y_j)]$$
$$= 1 - [P[(x_i < x_j) \cap (y_i < y_j)] + P[(x_i > x_j) \cap (y_i > y_j)]] \quad (ii)$$

(as a parameter)

Kendall's τ for the population is defined as: $\tau = P_c - P_D$

From (i) & (ii), $P_D = 1 - P_c$ $\qquad \tau = P_c - (1 - P_c) = 2P_c - 1$

$= (1 - P_D) - P_D = 1 - 2P_D$

t_k is a point estimate for τ

is an (unbiased) estimator $\qquad$ * don't need to worry about ties

If $(x_1,y_1), (x_2,y_2) \ldots (x_n,y_n)$ form a random sample

from a bivariate normal distribution with ρ as corr. coef,

$\tau = 0$ does not imply independence

$$\tau = \frac{2}{\pi} \sin^{-1} \rho$$

If X and Y are independent, $P_c = P(x_i < x_j)P(y_i < y_j) + P(x_i > x_j)P(y_i > y_j)$

$\tau = 0$

$$P_D = 1 - P(x_i < x_j)P(y_i < y_j) - P(x_i > x_j)P(y_i > y_j)$$
$$P_c - P_D = 2[P(x_i < x_j)P(y_i < y_j) + P(x_i > x_j)P(y_i > y_j)] - 1$$

if x and y are iid $N(0,1)$ then $P(x_1 > x_2) = P(x_2 > x_1)$

$$P_c = P(x_i < x_j)[P(y_i < y_j) + P(y_i > y_j)]$$

Spearman's Rank Correlation

11/9 (x_1, y_1) (x_2, y_2) $(x_3\ y_3)$

r_s	$R(x)$	$R(y)$	d_i
	$R(x_1)$	$R(y_1)$	$R(x_1) - R(y_1)$
	$R(x_2)$	$R(y_2)$	$R(x_2) - R(y_2)$
	$R(x_3)$	$R(y_3)$	$R(x_3) - R(y_3)$

$$R_S = 1 - \frac{6\sum D_i^2}{n(n^2-1)} = 1 - \frac{6\sum (i - S_i)^2}{n(n^2-1)} \qquad D_i = i - S_i$$

Tea Example

Reorder Judge 1 $(i:)$	1	2	3	4	5	6	7	8	9
corresponding Judge 2 $(S_i:)$	4	1	5	2	3	7	8	9	6

can have 9! different responses

$$P[R=r] = \frac{\#\text{ of ways we can get } R=r}{n!}$$

Example:

ranks for x:	1	2	3	$\sum d_i^2$	r_s
y:	1	2	3	0	1
	1	3	2	2	$1 - \frac{6\times 2}{3\times 8} = \frac{1}{2}$
	2	1	3	2	1/2
	2	3	1	6	-1/2
	3	1	2	6	-1/2
	3	2	1	8	-1

R	$P(R=r_s)$
-1	1/6
-0.5	2/6
0.5	2/6
1	1/6

conjugate pairs y_i are reversed

$(1, S_1), (2, S_2) \ldots (n, S_n)$

$(1, S_n), (2, S_{n-1}) \ldots (n, S_1)$

sums of d_i^2 for each pair will be same

$$\sum_1^n d_i^2 = \sum_1^n (i - S_i)^2$$

$$\sum d_i'^2 = \sum (i - S_{n-i+1})^2 = \sum (n - i + 1 - S_i)^2$$

$$d_i + d_i' = n + 1 - 2S_i$$

$$d_i - d_i' = 2i - (n+1)$$

$$\sum (d_i + d_i')^2 = \sum [2S_i - (n+1)]^2 = 4\sum \left(S_i - \frac{n+1}{2}\right)^2$$

$$\sum (d_i - d_i')^2 = \sum [2i-(n+1)]^2 = 4\sum \left(i - \frac{n+1}{2}\right)^2 = \frac{n(n^2-1)}{3}$$

$$\sum \left(i-\frac{n+1}{2}\right)^2 = \sum \left[i^2 - (n+1)i + \frac{(n+1)^2}{4}\right] = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)^2}{4}$$

$$= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n(n+1)}{2}\left[\frac{2n+1}{3} - \frac{n+1}{2}\right] = \frac{n(n^2-1)}{12}$$

$$\sum [(d_i + d_i') + (d_i + d_i')]^2 = 4\sum d_i^2$$

$$= \sum (d_i + d_i')^2 + \sum (d_i - d_i')^2 + 2\sum (d_i^2 - d_i'^2)$$

$$= \frac{2n(n^2-1)}{3} + 2\left(\sum d_i^2 - \sum d_i'^2\right)$$

$$2\sum d_i^2 = \frac{n(n^2-1)}{3} + \sum d_i^2 - \sum d_i'^2$$

$$\sum d_i^2 + \sum d_i'^2 = \frac{n(n^2-1)}{3}$$

$$= \frac{n(n^2-1)}{6} + \frac{n(n^2-1)}{6}$$

$$\sum d_i^2 - \frac{n(n^2-1)}{6} = \frac{n(n^2-1)}{6} - \sum d_i'^2$$

$$r_s = 1 - \frac{\sum d_i^2}{\frac{n(n^2-1)}{6}} = \frac{\frac{n(n^2-1)}{6} - \sum d_i^2}{\frac{n(n^2-1)}{6}}$$

$$r_s' = \frac{\frac{n(n^2-1)}{6} - \sum d_i'^2}{\frac{n(n^2-1)}{6}}$$

$$r_s + r_s' = 0$$

symmetric about 0

$$P[X=c] = P[X=-c]$$

$$P[R=r_s] = P\left[R = \frac{\frac{n(n^2-1)}{6} - \sum d_i^2}{\frac{n(n^2-1)}{6}}\right]$$

$$= P\left[R = \frac{\sum d_i'^2 - \frac{n(n^2-1)}{6}}{\frac{n(n^2-1)}{6}}\right]$$

$$= P\left[R = -\left(1 - \frac{\sum d_i'^2}{\frac{n(n^2-1)}{6}}\right)\right]$$

$$= P[R = -r_s']$$

Probability Models

Sheldon Ross 10

11/14 Fisher's Exact Test

- test for binomial proportions (last class)

$H_0: p_1 = p_2$

$H_A: p_1 > p_2$ or $p_1 < p_2$

$$P(X_{11} = n_{11}) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n}{n_{\cdot 1}}}$$

	Pop 1	Pop 2	
S	n_{11}	n_{12}	$n_{1\cdot}$
F	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	n

- row totals, column totals fixed - if one observation known, all others fixed (that observation is hyp)

ex. placebo

$$P[X_n = n_{11}] = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n}{n_{\cdot 1}}}$$

	New Drug		
Placebo	Some Relief		
Some relief	20	15	35
No relief	35	30	65
	55	45	100

David Salzburg

Lady Testing Tea

$$P[X_n \geq 20] = \frac{\binom{35}{20}\binom{65}{35}}{\binom{100}{55}} + \dots$$

$P[X \geq 20] = 1 - [X \leq 19] = 1 - \text{phyper}(19, 35, 65, 55)$

Could also do Pearson's χ^2 test for this

12.8 - independent - regular χ^2 test if $H_A: p \neq p_2$

$$P[X = x] = \frac{\binom{45}{x}\binom{75}{45-x}}{\binom{120}{45}}$$

Find $P[X \leq 14]$

use Fisher's exact for one sided *

ex. Business sizes

	Size S	M	L
Response	125	81	40

χ^2 test $\quad \chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad e_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} \quad n = \sum_i \sum_j n_{ij}$

suppose $r = 2$

	1	2	
y_1	n_{11}	n_{12}	$n_{1\cdot} - y_1$
y_2	n_{21}	n_{22}	$n_{2\cdot} - y_2$
	$\vdots$	$\vdots$	

ex.

NR	R
75	125
119	81
160	40

$$\chi^2 = \frac{1}{\hat{p}(1-\hat{p})} \sum \frac{y_i^2}{n_i} - \frac{n\hat{p}}{1-\hat{p}}$$

$$\chi^2 = \sum \frac{(y_i - n_i \hat{p})^2}{n_i \cdot \hat{p}(1-\hat{p})} \quad \hat{p} = \sum \frac{y_i}{n}$$

$$n = 600 \quad \hat{p} = \frac{75 + 119 + 160}{600} = 0.59$$

$$\frac{\sum y_i^2}{n_i} = \frac{75^2 + 119^2 + 160^2}{200} = 226.93$$

11/28 **Regression**

Review - best fit line minimizes residuals $\hat{\beta}_1 = r\frac{S_y}{S_x}$ (Pearson's corr.) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$y = \alpha + \beta_1 x + \varepsilon$ $\varepsilon \sim \text{iid } N(0, \sigma^2)$ assume: equal variances

Why is this assumption necessary?

to approximate t dist

$H_0: \beta = \beta_0$

$H_A: \beta \gtreqless \beta_0$

$t = \frac{\hat{\beta}_1 - \beta_0}{se} \sim t_{n-2}$

Non-parametric = cannot assume $\varepsilon \sim N$

Assume $\varepsilon \sim$ iid some symmetric

(Rank based approach)

$Y \sim$ (some symmetrical dist.) $(\alpha + \beta x_i, \sigma)$

$E(Y|x_i) = \alpha + \beta x_i$

* median *

example

Years (x_i)	Ratio (y_i)
1	1.26
2	1.27
3	1.12
4	1.16
5	1.03

$H_0: \beta = 0$

$H_A: \beta < 0$

1- order x_i $x_1 < x_2 < \ldots < x_n$

2- find $D_i = y_i - \beta_0 x_i$

$D_j = y_j - \beta_0 x_j$

3- compute $d_{ij} = D_j - D_i$ (under H_0)

$= (y_j - y_i) - \beta_0 (x_j - x_i)$ $j > i$

Test statistic: $T_t(\beta) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}(d_{ij})$

x_i	y_i	D_i	sgn (d_{1j})	sgn (d_{2j})	sgn (d_{3j})
1	1.26	1.26			
2	1.27	1.27	1		
3	1.12	1.12	-1	-1	
4	1.16	1.16	-1	-1	1
5	1.03	1.03	-1	-1	-1	-1

$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$

x_1 compared to x_2

$T_t(\beta) = -6$ ← concordance − discordance

Kendall's $\tau = \frac{-6}{\binom{5}{2}} = -0.6$

Table L p-value = 0.117

do not reject

12/2 Cloud seeding example - confidence intervals for β

Use R code

Look on Table for p-values

n	T	P
5	1.0	0.008
	0.8	.042

← first value and last value of β

$p = 0.008$ $1 - 2 \times 0.008 = 0.984$

exact CI

$\tau = -0.6$ $p = 0.117$ $1 - 2 \times 0.117 = 0.866$

86.6% CI

Example 11.1 - using R code $y = 2.33 + 0.567x$

when plug in x, get predicted median (y).

Influential points - nonparametric procedure robust against outliers/influential points

Least Squares

Heavily affected by
- design points
- outliers

Rank-based

Heavily affected by
- design points → based on x values

$$T_n(\beta) = \sum_1^n (x_i - \bar{x}) \underbrace{R(y_i - \beta x_i)}_{\text{rank of deviations}}$$

if β true parameter, constant

value of $T_n(\beta) = 0$ because $\sum (x_i - \bar{x})$ will $= 0$

End of Ch 11

χ^2 test - two-sided

Fisher's exact - one-sided

12/5 Example - commercials

	don't remember	vaguely	remember	
A	12	15	23	50
B	15	15	20	50
total	27	30	43	100

Ordinal Variables

	Low	Med	High	Total
Under 45	2	3	5	10
Over 45	4	5	1	10
	6	8	6	20

Degree of risk aversion

χ^2 test - association

younger more risk than older? Wilcoxon Rank Sum - sum of ranks corresponding to group <45

	1	2	3
<45	2	3	5
≥45	4	5	1
	6	8	6

one sided - with ties

average of 1 ranks $\frac{1+2+3+4+5+6}{6} = 3.5$ $\qquad \frac{n(n+1)}{2n}$ $\qquad \frac{\text{first}+\text{last}}{2}$

2 ranks $\frac{7+8+9+10+11+12+13+14}{8} = (8\times 7) = 10.5$ $\qquad \frac{7+14}{2}$

3 ranks $\frac{15+16+17+18+19+20}{6} = \frac{15+20}{2} = 17.5$

Test Statistic

$W_N = (2\times 3.5) + (3\times 10.5) + (5\times 17.5) = 120$

in group <45 $\quad$ rank

No table, so use large sample test

$E(W_N) = \frac{m(N+1)}{2}$

use for ties

$V(W_N) = \frac{mn(N+1)}{12} - \frac{mn\sum t(t^2-1)}{12N(N-1)}$

# of cases of ties

$\sum t(t^2-1) =$

$6(6^2-1) + 8(8^2-1) + 6(6^2-1)$

$=$

$$V(W_N) = \frac{10 \times 10 \times 21}{12} - \frac{10 \times 10 \times 921}{12 \times 20 \times 19} = 154.74$$

$$E(W_N) = \frac{10 \times 21}{2} = 105$$

cont. correction $-.5$ right-tailed

$$Z = \frac{125.5 - 105}{\sqrt{154.74}} = 1.647$$

McNemar's - Exact

$H_0: p_{\cdot 1} = p_{1\cdot}$ or $p_{12} = p_{21}$

$H_A: p_{\cdot 1} > p_{1\cdot}$ or $p_{12} < p_{21}$

$p_{\cdot 1} < p_{1\cdot}$ or $p_{12} > p_{21}$

$p_{\cdot 1} \neq p_{1\cdot}$ or $p_{12} \neq p_{21}$ ← χ^2 test for large samples only

$p_{\cdot 1} = p_{11} + p_{21}$

$p_{1\cdot} = p_{11} + p_{12}$

$$\frac{(X_{12} - X_{21})^2}{X_{12} + X_{21}}$$

$$(X_{12} \mid X_{12} + X_{21} = S) \sim bin\left(S, \frac{p_{12}}{p_{12} + p_{21}}\right)$$

Under H_0, $p_{12} = p_{21} \Rightarrow p = \frac{1}{2}$

$H_0: p_{12} = p_{21}$

$H_A: p_{12} < p_{21}$ small values of x_{12} will reject H_0

under H_0: $X_{12} \sim bin(S, \frac{1}{2})$ Test 3 Prob 4 example

$4 + 13 = 17$

p-value $= P[X_{12} \geq 4]$

$$= \sum^{4} \binom{17}{x} \left(\frac{1}{2}\right)^{17} = 0.0245$$