Wilcoxian Rank Sum

Drug company with a new painkiller.

$n_1 = 15$ $n_2 = 15$

Responses = 1-5 (1=no effect, 5=fabulous)

New drug: 3,5,4,3,2,5,1,4,5,3, 3, 5,5,4

Aspirin: 4,1,3,2,4,1,3,4,2,2, 2,4, 3,4,5

H_0: location of 2 pops. is same

H_1: location of pop 1 is to the right of pop 2

Test statistic: $Z = \frac{T - E(T)}{\sigma_T}$

$E(T) = \frac{n_1(n_1 + n_2 + 1)}{2}$ $\sigma_T = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$

Rejection region: $Z > Z_\alpha = Z_{0.05} = 1.645$

If alpha is 2-tailed then p-value = 2P(Z.....)

Step 1: Rank observations

New Drug	Rank	Aspirin	Rank
3	12	4	19.5
5	27	1	2
4	19.5	3	12
3	12	2	6
2	6	4	19.5
5	27	1	2
1	2	3	12
4	19.5	4	19.5
5	27	2	6
3	12	2	6
3	12	2	6
5	27	4	19.5
5	27	3	12
5	27	4	19.5
4	19.5	5	27
Rank Sums	$T_1 = 276.5$		$T_2 = 188.5$

Step 2: Plug & Chug

$E(T) = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{15(31)}{2} = 232.5$

$\sigma_T = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(15)(15)(31)}{12}} = 24.1$

$Z = \frac{T - E(T)}{\sigma_T} = \frac{276.5 - 232.5}{24.1} = 1.83$

p-value = P(Z>1.83) = 0.5 - 0.4664 = 0.0336

z is greater than Z_α so we can conclude that the new drug is perceived to be more effective than aspirin.

1 Compare 2 pops
2 Data type: ranked or quantitative but nonnormal
3 Independent samples

Lilliefors Test

x	S(x)	F(x)	|F(x) – S(x)|
80	0.1	0.0367	0.0633
89	0.2	0.1539	0.0461
93	0.3	0.2514	0.0486
97	0.4	0.3745	0.0255
102	0.5	0.5398	0.0398
103	0.6	0.5753	0.0247
105	0.7	0.6406	0.0594
108	0.8	0.7324	0.0674
109	0.9	0.7852	0.1148
121	1.0	0.9591	0.0409

$F(x) = P(x <= 80) = \left[\frac{x-\mu}{\sigma} <= \frac{80 - 100.8}{11.62}\right] = P(Z <= -1.79) = 0.0367$

$F(x) = P(x <= 89) = \left[\frac{x-\mu}{\sigma} <= \frac{89 - 100.8}{11.62}\right] = P(Z <= -1.02) = 0.1539$

⋮

$F(x) = P(x <= 121) = \left[\frac{x-\mu}{\sigma} <= \frac{121 - 100.8}{11.62}\right] = P(Z <= 1.74) = 0.9591$

Once the table (left) is done, find the largest |F(x) – S(x)|. In this case, it is 0.1148 = D.

S(80) = 1/10 = 0.1
S(89) = 2/10 = 0.2
⋮
S(121) = 10/10 = 1.0

S(x) is the proportion of sample values that is <= x.

H_0: The data are normally distributed

H_1: The data are not normally distributed

Rejection Region: $D > D_{n,\alpha}$

0.1148 < 0.258... therefore can't reject H_0.

Simple Regression

Regression Line:

$y = \beta_0 + \beta_1 x + \varepsilon$

$b_0 = \bar{y} - b_1\bar{x}$

$b_1 = \frac{COV(X,Y)}{S^2_x}$

$y = 6{,}533 - 0.0312x$

$b_0 = 5{,}411.41 - (-0.0312)(36{,}009.45) = 6{,}533$

$b_1 = \frac{-1{,}356{,}256}{43{,}528{,}690} = -0.0312$

sample covariance:

$COV(X,Y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ $\bar{x} = \frac{\Sigma x_i}{n}$ $\bar{y} = \frac{\Sigma y_i}{n}$

$COV(X,Y) = \frac{-134{,}269{,}298}{99} = -1{,}356{,}256$

$\bar{x} = \frac{3{,}600{,}945}{100} = 36{,}009.45$ $\bar{y} = \frac{541{,}141}{100} = 5{,}411.41$

sample variance:

$s^2_x = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$ $s^2_x = \frac{4{,}309{,}340{,}277}{99} = 43{,}528{,}690$

$s^2_y = \frac{\Sigma(y_i - \bar{y})^2}{n-1}$ $s^2_y = \frac{6{,}434{,}890}{99} = 64{,}999$

residual = observed – predicted $(y_i - \hat{y}_i)$

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$SSE = (n-1)\left[s^2_y - \frac{(COV(X,Y))^2}{s^2_x}\right]$ $SSE = (99)\left[64{,}999 - \frac{(-1{,}356{,}256)^2}{43{,}528{,}690}\right] = 2{,}251{,}363$

Sign Test

H_0: location of 2 pops. is same (equivalent to testing p = 0.5)

H_1: location of Euro pop is to the right of USA pop.

Step 1: Rank observations

Respondent	Euro	USA	Diff.
1	4	5	-1
2	2	1	1
3	5	4	1
4	3	2	1
5	2	1	1
6	5	3	2
7	1	3	-2
8	4	2	2
9	4	2	2
10	2	2	0
11	3	2	1
12	4	3	1
13	2	1	1
14	3	4	-1
15	2	1	1
16	4	3	1
17	2	1	1
18	4	3	1
19	5	4	1
20	3	1	2
21	4	2	2
22	3	3	0
23	3	4	-1
24	5	2	3
25	2	3	-1

Step 1: Calculate diff. between pairs (see left).

Step 2: Count positives.

positives: 18 X = # of positives (18)
negatives: 5 n = observations - zeros (25-2 = 23)
zeros: 2

Step 3: Plug & Chug!

$Z = \frac{X - 0.5n}{0.5\sqrt{n}} = \frac{18 - 0.5(23)}{0.5\sqrt{23}} = 2.71$

p-value = P(Z > 2.71) = 0.5 - 0.4966 = 0.0034

Test statistic: $Z = \frac{X - 0.5n}{0.5\sqrt{n}}$

Rejection region: $Z > Z_\alpha = Z_{0.05} = 1.645$

There is strong evidence that Euro is preferred over USA.

1 sample size must be >10, pops must be identical in shape and spread
2 compare 2 pops
3 data type ranked
3 matched pairs

Simple Regression

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.8063
R Square	0.6501
Adjusted R Square	0.6466
Standard Error	151.6
Observations	100

SSE: 2,251,362
SST: 6,434,890
SSR: 4,183,528
MSR = 4,183,528 = SSR/1
MSE = 22,973 = SSE/(n-2)
F = 182.1 = MSR/MSE

ANOVA

	df	SS	MS	F	Significance F
Regression	1	4183528	4183528	182.1	0.0000
Residual (error)	98	2251362	22973		
Total	99	6434890			

	Coefficients	Std Error	t Stat	P-value
Intercept	6533	84.51	77.31	0.0000
Odometer	-0.0312	0.00231	-13.49	0.0000

Car	Odometer	Price
1	37,388	$5,318
2	44,758	5,061
3	45,833	5,008
•	•	•
•	•	•
•	•	•
100	36,392	5,133

β_1 is called a stock's beta coefficient. It measures how sensitive a stock's rate of return (X) is to changes in the level of the overall market.

$\beta_1 > 1$ stock is more sensitive than market.

$\beta_1 < 1$ stock is less sensitive than market.

$\beta_1 = 2$ than Δ1% in market = Δ2% in stock.

Testing The Slope

If slope (β_1) is zero, there is no linear relationship.

Test statistic: $t = \frac{b_1 - \beta_1}{s_{b_1}}$ $t = \frac{-0.0312 - 0}{0.00231} = -13.49$

$H_0: b_1 = 0$

$H_1: b_1 != 0$

$s_{b_1} = \frac{S_\varepsilon}{\sqrt{(n-1)s^2_x}}$ $s_{b_1} = \frac{151.6}{\sqrt{(99)(43{,}528{,}690)}} = 0.00231$

$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ $s_\varepsilon = \sqrt{\frac{2{,}251{,}363}{98}} = 151.6$

Smallest value S_ε can have is zero. Smaller S_ε, better the fit of model. Judge it by comparing to $\bar{y}$.

Rejection region: $t > t_{\alpha/2,n-2} = t_{0.025,98} \sim 1.984$ or $t < -t_{\alpha/2,n-2} \sim -1.984$

There is overwhelming evidence of a linear relationship. ($t < t_{\alpha/2,n-2}$, $-13.49 < -1.984$)

Wilcoxian Signed Rank Sum Test For Matched Pairs

Step 1: Rank observations

Worker	8am arrive	flextime	Diff.	|Diff.|	Rank
1	34	31	3	3	21.0
2	35	31	4	4	27.0
3	43	43	-1	1	(4.5)
4	46	44	2	2	13.0
5	16	15	1	1	4.5
6	26	28	-2	2	(13.0)
7	68	63	5	5	31.0
8	38	39	-1	1	(4.5)
9	61	63	-2	2	(13.0)
10	52	54	-2	2	(13.0)
11	68	65	3	3	21.0
12	13	12	1	1	4.5
13	69	71	-2	2	(13.0)
14	18	13	5	5	31.0
15	53	55	-2	2	(13.0)
16	18	19	-1	1	(4.5)
17	41	41	3	3	21.0
18	25	23	2	2	13.0
19	17	14	3	3	21.0
20	26	21	5	5	31.0
21	44	40	4	4	27.0
22	30	33	-3	3	(21.0)
23	19	18	1	1	4.5
24	48	51	-3	3	(21.0)
25	29	33	-4	4	(27.0)
26	24	21	3	3	21.0
27	51	50	1	1	4.5
28	40	38	2	2	13.0
29	26	22	4	4	27.0
30	20	19	1	1	4.5
31	19	21	-2	2	(13.0)
32	42	38	4	4	27.0

Step 2: Plug & Chug!

$T^+ = 367.5$ (sum of positives)
$T^- = 160.5$ (sum of negatives)
(zeros would be eliminated)

Test statistic: $Z = \frac{T - E(T)}{\sigma_T}$

Where:

$T = T^+ = 367.5$

$E(T) = \frac{n(n+1)}{4} = \frac{32(33)}{4} = 264$

$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{32(33)(65)}{24}} = 53.48$

$Z = \frac{T - E(T)}{\sigma_T} = \frac{367.5 - 264}{53.48} = 1.94$

rejection region: $Z > Z_{\alpha/2} = Z_{0.025} = 1.96$
or, $Z < -Z_{\alpha/2} = -1.96$

p-value = 2P(Z> 1.94) = 2(0.5 - 0.4738) = 0.0524

Because Z>1.96, we can say there is not enough evidence to say that flextime commutes are different from the 8am schedule.

1 compare 2 pops
2 data type: quantitative
3 matched pairs

Coefficient of Determination

Measures strength of linear relationship.

Test statistic: $R^2 = \frac{(COV(X,Y))^2}{S^2_x S^2_y} = 1 - \frac{SSE}{\Sigma(y_i - \bar{y})^2} = \frac{SSR}{\Sigma(y_i - \bar{y})^2}$

$R^2 = \frac{(COV(X,Y))^2}{S^2_x S^2_y} = \frac{(-1{,}356{,}256)^2}{(43{,}528{,}690) \times (64{,}999)} = 0.6501 = 65.01\%$

65% of the variation in price is due to odometer reading.

Testing Coefficient of Correlation

Population correlation coefficient = ρ.

When there is no linear relationship between 2 variables, ρ=0.

$H_0: \rho = 0$ $H_1: \rho != 0$

Test statistic: $t = r\sqrt{\frac{n-2}{1-r^2}} = (-0.806)\sqrt{\frac{100-2}{1-(-0.806)^2}} = -13.49$

$r = \frac{COV(X,Y)}{S_x S_y} = \frac{(-1{,}356{,}256)}{(6597.6) \times (254.9)} = -0.806$

Rejection region: $t > t_{\alpha/2,n-2} = t_{0.025,98} \sim 1.984$ or $t < -t_{0.025,98} = -1.984$

There is overwhelming evidence to infer the two variables are linearly related. (–13.49 < –1.984)

Residuals

Standardized residuals for point *i*: $i = \frac{r_i}{S_\varepsilon}$

Std. dev. of the *i*th residual:

$s_{r_i} = S_\varepsilon\sqrt{1 - h_i}$

$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s^2_x}$

Prediction Interval

Predicting the particular value of y for a given x

$\hat{y} \pm t_{\alpha/2,n-2} S_\varepsilon\sqrt{1 + \frac{1}{n} + \frac{(X_g - \bar{X})^2}{(n-1)S^2_X}}$

X_g = given value of X

$\hat{y} = b_0 + b_1 X_g$

Prediction Interval

Dealer wants to predict price of *single* car.

$\hat{y} = 6{,}533 - 0.0312x$
$= 6{,}533 - 0.0312(40{,}000)$
$= 5{,}285$

5,288 +/- 303

Xg = 40,000 (miles)
$S_\varepsilon = 151.6$
$S^2_X = 43{,}528{,}690$
$\bar{X} = 36{,}009$
$t_{\alpha/2,n-2} = \sim 1.984$

Interval Estimator

Estimating the expected value of y for a given x.

Interval estimate is narrower than prediction interval.

☺ $\hat{y} \pm t_{\alpha/2,n-2} S_\varepsilon\sqrt{\frac{1}{n} + \frac{(X_g - \bar{X})^2}{(n-1)S^2_X}}$

Interval Estimator

Dealer wants to estimate *average price of a large lot* of cars.

5,288 +/- 35

Spearman Rank Correlation Coefficient

Employee	Score	Rank (a)	Rating	Rank (b)
1	59	9	3	10.5
2	47	3	2	3.5
3	58	8	4	17.0
4	66	14	3	10.5
5	77	20	2	3.5
6	57	7	4	17.0
7	62	12	3	10.5
8	68	16	3	10.5
9	69	17	5	19.5
10	36	1	1	1.0
11	48	4	3	10.5
12	65	13	3	10.5
13	51	5	2	3.5
14	61	11	3	10.5
15	40	2	3	10.5
16	67	15	4	17.0
17	60	10	2	3.5
18	56	6	3	10.5
19	76	19	3	10.5
20	71	18	5	19.5

Population Spearman rank correlation coefficient = r_s.

$H_0: \rho_s = 0$ α = 0.05 $s_a = 5.92$

$H_1: \rho_s != 0$ n = 20 $s_b = 5.50$

Cov(a,b) = 12.34

$r_s = \frac{COV(a,b)}{S_a S_b} = \frac{12.34}{(5.92)(5.50)} = 0.379$

Rejection region: $r_s > 0.450$ or $r_s < -0.450$

There is not enough evidence to believe scores and ratings are related. (0.379 != 0.450 so cannot reject H_0).

Kruskall-Wallis Test

Step 1: Rank observations

4pm-mid.		Mid.-8am		8am-4pm	
4	27.0	3	16.5	3	16.5
4	27.0	4	27.0	1	2.0
3	16.5	2	6.5	3	16.5
4	27.0	2	6.5	2	6.5
3	16.5	3	16.5	1	2.0
3	16.5	4	27.0	3	16.5
3	16.5	3	16.5	4	27.0
3	16.5	3	16.5	2	6.5
2	6.5	2	6.5	4	27.0
3	16.5	3	16.5	1	2.0
$T_1 = 186.5$		$T_2 = 156.0$		$T_3 = 122.5$	

1 compare 2 or more pops
2 data type: ranked or quantitative but not normal
3 independent samples

Step 2: Plug & Chug

H_1: At least 2 pop locations differ

H_0: location of all 3 pops is the same

Test statistic: $H = \left[\frac{12}{n(n+1)}\sum_{j=i}^{k}\frac{T^2_j}{n_j}\right] - 3(n+1)$

$= \left[\frac{12}{n(n+1)}\left(\frac{T^2_1}{n_1} + \frac{T^2_2}{n_2} + \frac{T^2_3}{n_3}\right)\right] - 3(n+1)$

$= \left[\frac{12}{30(30+1)}\left(\frac{186.5^2}{10} + \frac{156^2}{10} + \frac{122.5^2}{10}\right)\right] - 3(30+1)$

$= 2.64$

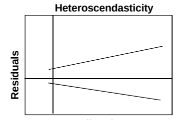
Rejection Region: $H > X^2_{\alpha,k-1} = X^2_{0.05,2} = 5.99147$

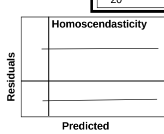
There is not enough evidence to infer a difference exists between the 3 shifts. (i.e. 2.64 is not > 5.99147)

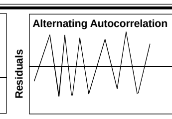
Testing For Normality:

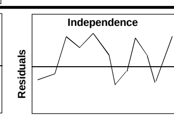
- Lillefors
- Chi-Squared test
- Hetero/homoscendasticity
- Autocorrelation

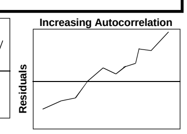
Heteroscendasticity











Friedman Test

Manager

Applicant	1	(rank)	2	(rank)	3	(rank)	4	(rank)
1	2	(3)	1	(1)	2	(3)	2	(3)
2	4	(4)	2	(1.5)	3	(3)	2	(1.5)
3	2	(2)	2	(2)	2	(2)	3	(4)
4	3	(3.5)	1	(1)	3	(3.5)	2	(2)
5	3	(2.5)	2	(1)	3	(2.5)	5	(4)
6	2	(1.5)	2	(1.5)	3	(3)	4	(4)
7	4	(2)	1	(1)	5	(3.5)	5	(3.5)
8	3	(2.5)	2	(1)	5	(4)	3	(2.5)
	$T_1 = 21$		$T_2 = 10$		$T_3 = 24.5$		$T_4 = 24.5$	

H_1: The location of all k pops are the same

H_0: At least 2 pop locations differ

Test statistic: $F_r = \left[\frac{12}{b(k)(k+1)}\sum_{j=i}^{k}T^2_j\right] - 3b(k+1)$

b = 8
k = 4

$= \left[\frac{12}{8(4)(5)}(21^2 + 10^2 + 24.5^2 + 24.5^2)\right] - 3(8)(5)$

$= 10.61$

Rejection Region: $F_r > X^2_{\alpha,k-1} = X^2_{0.05,3} = 7.81473$

There appears to be sufficient evidence to indicate that the manager's evaluations differ. (10.61 > 7.81473)

1 compare 2 or more pops
2 data type: ranked or quantitative but not normal
3 blocked samples

Multiple Regression

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.7246
R Square	0.5251
Adjusted R Square	0.4944
Standard Error	5.51
Observations	100

SSE: 2,825.6
SST: $5{,}949.5 = \Sigma(y_i - \bar{y})^2$
SSR: 3,123.8
MSR = 520.6 = SSR/k
MSE = 30.4 = SSE/(n-k-1)
F = 182.1 = MSR/MSE
K = 6 N = 100

ANOVA

	df	SS	MS	F	Significance F
Regression	6	3123.8	520.6	17.14	0.0000
Residual (error)	93	2825.6	30.4		
Total	99	5949.5			

	Coefficients	Std Error	t Stat	P-value
Intercept	72.45	7.89	9.18	0.0000
Rooms	-0.0076	0.0013	-6.07	0.0000
Nearest	-1.6462	0.6328	-2.60	0.0108
Office	0.0198	0.0034	5.8	0.0000
College	0.2118	0.1334	1.59	0.1159
Income	-0.4131	0.1396	-2.96	0.0039
Disttwn	0.2253	0.1787	1.26	0.2107

Regression Model is estimated by:

Margin = 72.45 – 0.0076(Rooms) – 1.6462(Nearest) + 0.0198(Office) + 0.2118(College) – 0.4131(Income) + 0.2253(Disttown)

If SSR (explained error) is high relative to SSE (unexplained error), the model is good.

Standard Error of Estimate

$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$

= 5.51

F-Test

$F = \frac{(\Sigma(y_i - \bar{y})^2 - SSE)/k}{SSE/(n-k-1)}$

= 17.14

Coefficient of Determination

$R^2 = 1 - \frac{SSE}{\Sigma(y_i - \bar{y})^2}$

= 0.5251

Coefficient of Determination (Adjusted)

Adjusted $R^2 = 1 - \frac{SSE/(n-k-1)}{\Sigma(y_i - \bar{y})^2/(n-1)}$

= 0.4944

SSE	S_ε	R^2	F	Model
0	0	1	∞	Perfect
Small	Small	Close to 1	Large	Good
Large	Large	Close to 0	Small	Poor
$\Sigma(y_i - y)^2$	$\sqrt{\frac{\Sigma(y_i - y)^2}{n-k-1}}$	0	0	Useless

* When n is large and k is small, this quantity is approx. equal to the st. dev. of y.

Confidence Interval

$\hat{y} \pm t_{\alpha/2,n-k+1}$ (s.e.(y-hat$_{mean}$))

Prediction Interval

$\hat{y} \pm t_{\alpha/2,n-k+1}\sqrt{S^2_\varepsilon + (\text{s.e.(y-hat}_{mean}))^2}$

s.e.(y-hat$_{mean}$) $= S_\varepsilon\sqrt{\frac{1}{n} + \frac{(X_g - \bar{X})^2}{SSx}}$

$SSx = \Sigma(x_i - \bar{x})^2$

Testing The Coefficients

$H_0: \beta_i = 0$ (linearly related)

$H_1: \beta_i != 0$ (not linearly related)

$t = \frac{b_i - \beta_i}{s_{b_i}}$

t is Student t distributed with d.f. = n – k – 1

$t_{\alpha,n-k-1} = \sim 1.987$ (α = 0.05)

Rejection region: t > 1.987 or t < –1.987

Example: (using t for rooms)
t = –6.07, p-value = 0
Overwhelming evidence of a linear relationship.

Coefficient Interpretation

Easy. For example: For each additional room (1), the margin changes on average by 0.0076%.