Advanced Statistics

Dr. Syed Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics,** 13th Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

- Elementary Statistics, 14th Edition, Mario F. Triola

These notes contain material from the above resources.

Example: Use Python to find the value of the linear correlation coefficient *r* for the Powerball jackpot amounts and numbers of tickets listed in Table 1.

Table 1 Powerball Tickets Sold and Jackpot Amounts

Jackpot	334	127	300	227	202	180	164	145	255
Tickets	54	16	41	27	23	18	18	16	26

```
import numpy as np
# Given data set
jackpot_amounts = [334, 127, 300, 227, 202, 180, 164, 145, 255]
tickets_sold = [54, 16, 41, 27, 23, 18, 18, 16, 26]

# Calculate correlation coefficient
correlation_coefficient = np.corrcoef(jackpot_amounts, tickets_sold)[0, 1]

# Print the result
print(f"Correlation Coefficient: {correlation_coefficient:.2f}")

# Correlation Coefficient: 0.95
```

x(Jackpot)	y(Tickets)	x^2	y^2	xy
334	54	111,556	2916	18,036
127	16	16,129	256	2032
300	41	90,000	1681	12,300
227	27	51,529	729	6129
202	23	40,804	529	4646
180	18	32,400	324	3240
164	18	26,896	324	2952
145	16	21,025	256	2320
255	26	65,025	676	6630
$\sum x$ = 1934	$\sum y$ = 239	$\sum x^2$= 455,364	$\sum y^2$= 7691	$\sum xy$ = 58,285

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{9(158{,}2852) - (1934)(239)}{\sqrt{9(455{,}364) - (1943)^2}\sqrt{9(7651) - (239)^2}}$$

$$r = \frac{62{,}339}{\sqrt{357{,}920}\sqrt{12{,}098}} = 0.947$$

Method 2

$$r = \frac{\sum Z_x Z_y}{n-1}$$

$$Z_x = \frac{x - \bar{x}}{s_x}$$

$$Z_y = \frac{y - \bar{y}}{s_y}$$

Method 2

$$s_x = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

or

$$s_x = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i)^2\}}$$

$$s_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n-1}}$$

or

$$s_y = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} y^2{}_i - (\sum_{i=1}^{n} y_i)^2\}}$$

$$s_x = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i)^2\}}$$

$$s_x = \sqrt{\frac{1}{9(9-1)}\{9(455,364) - (1934)^2\}}$$

$$s_x = 70.5061$$

$$s_y = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} y^2{}_i - (\sum_{i=1}^{n} y_i)^2\}}$$

$$s_y = \sqrt{\frac{1}{9(9-1)}\{9(7691) - (239)^2\}}$$

$$s_y = 12.9626$$

$\bar{x} = 214.8889$

$\bar{y} = 26.5556$

$$Z_x = \frac{x - \bar{x}}{s_x}$$

$$Z_y = \frac{y - \bar{y}}{s_y}$$

$$Z_x = \frac{334 - 214.8889}{70.5061}$$

$$Z_y = \frac{54 - 26.5556}{12.9626}$$

$Z_x = 1.6894$

$Z_y = 2.1172$

x	y	$Z_x = \frac{x - \bar{x}}{s_x}$	$Z_y = \frac{y - \bar{y}}{s_y}$	$Z_x Z_y$
334	54	1.6894	2.1172	3.57680
127	16	−1.2465	− 0.8143	1.01502
300	41	1.2071	1.1143	1.34507
227	27	0.1718	0.0343	0.00589
202	23	−0.1828	−0.2743	0.05014
180	18	−0.4948	−0.6600	0.32657
164	18	−0.7218	−0.6600	0.47639
145	16	−0.9912	−0.8143	0.80713
255	26	0.5689	−0.0429	− 0.02441
				$\sum(Z_x Z_y) =$ 7.57862

$$r = \frac{\sum Z_x Z_y}{n-1}$$

$$r = \frac{7.57862}{9-1}$$

$$r = 0.947$$

Interpreting *r*: Explained Variation

- If we conclude that there is a **linear correlation between *x* and *y***, we can find a linear equation that expresses *y* in terms of *x*, and that equation can be used to predict values of *y* for given values of *x*.

- But a **predicted value of *y* will not necessarily be the exact result** that occurs because in addition to *x*, there are other factors affecting *y*, such as random variation and other characteristics not included in the study.

- The value of r^2 is **the proportion of the variation in *y*** that is explained by the linear relationship between *x* and *y*.

Interpreting *r*: Explained Variation

Example: Using the 9 pairs of data from Table 1, we get a linear correlation coefficient of $r = 0.947$. What proportion of the variation in the numbers of tickets sold can be **explained by the variation** in jackpot amounts?

Interpreting *r*: Explained Variation

With $r = 0.947$, we get $r^2 = 0.897$.

INTERPRETATION We conclude that **0.897 (or about 90%)** of the variation in **the numbers of tickets sold can be explained by the linear relationship between jackpot amounts and numbers of tickets sold**.

This implies that about **10% of the variation in the numbers of tickets sold cannot be explained** by the linear correlation between the two variables.

Interpreting *r* with Causation: Don't Go There!

In the previous example, we concluded that there is sufficient evidence to support the claim of a linear correlation between **lottery jackpot amounts** and **numbers of tickets sold**.

We should *not* make any conclusion that includes a statement about a **cause-effect relationship** between the two variables. We should **not conclude that an increase in the jackpot amount will cause ticket sales to increase**. See the first of the following common errors, and know this:

Correlation does not imply causality (the relationship between cause and effect).!

Common Errors Involving Correlation

Here are three of the most common errors made in interpreting results involving correlation:

1. ***Assuming that correlation implies causality.*** One classic example involves paired data consisting of the **stork population** in Oldenburg, Germany and the **number of human births**. For the years of 1930 to 1936, the data suggested a linear correlation.

Bulletin: Storks do not actually cause births, and births do not cause storks. Both variables were affected by another **variable lurking** in the background. (A **lurking variable** is one that **affects the variables being studied but is not included in the study**.) Here, an **increasing human population** resulted in more births and increased construction of **thatched roofs** that attracted **storks!**

Stork

Thatched roof

Common Errors Involving Correlation

2. *Using data based on averages.* Averages suppress individual variation and may **inflate the correlation coefficient**. One study produced a **0.4 linear correlation coefficient** for paired data relating income and education among individuals, but the **linear correlation coefficient became 0.7** when regional averages were used.

3. *Ignoring the possibility of a nonlinear relationship.* If there is no linear correlation, there might be some other correlation that is not linear.

Example: Use the paired data from Table 1 on the previous slide to conduct a **formal hypothesis test of the claim that there is a linear correlation between lottery jackpot amounts and numbers of tickets sold**. Use a 0.05 significance level with the *P*-value method of testing hypotheses.

Hypothesis Test for Correlation

Given **r =** 0.947 , **n = 9**

1. **We state our hypothesis as**:

H_0: $\rho = 0$ (There is no linear correlation.)

H_1: $\rho \neq 0$ (There is a linear correlation.)

2. **The level of significance is set** $\alpha = 0.05$.

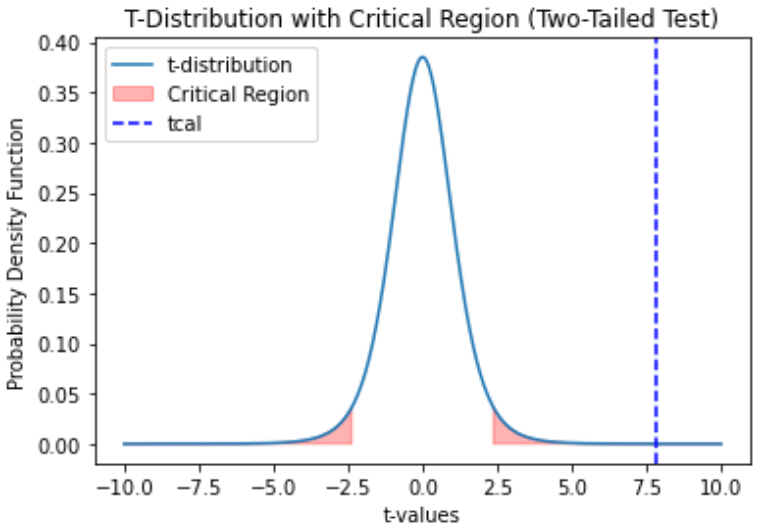
3. **Test statistic to be used is** $t_{cal} = \dfrac{r}{\sqrt{\dfrac{1 - r^2}{n-2}}}$

4. **Calculations:**

$$t_{cal} = \frac{0.947}{\sqrt{\dfrac{1 - (0.947)^2}{9-2}}} = 7.800$$

5. Critical region:

$|t_{cal}| > t_{tab}$, where $t_{tab} = t_{(\alpha/2,\, n-2)}$



$t_{tab} = t_{(0.0250,\, 7)} = 2.365$

$7.8000 > 2.365$ (True)

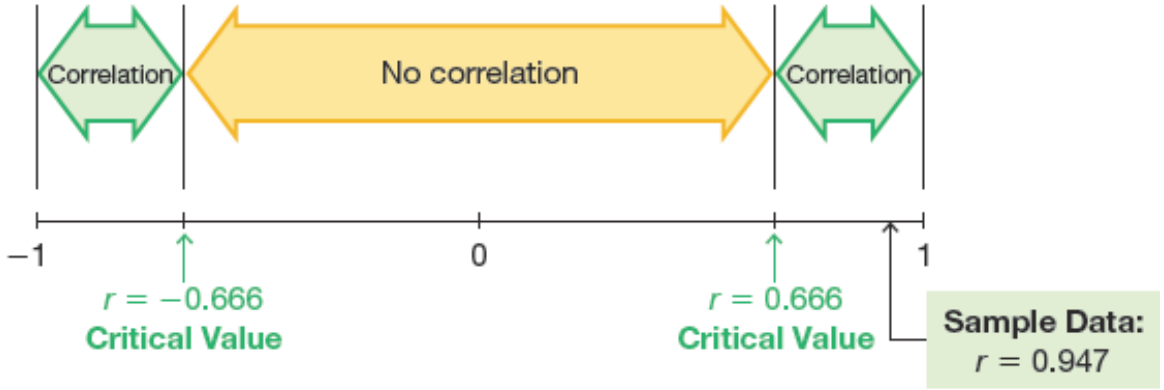
6. Conclusion: Since t_{cal} is greater than the t_{tab}, so we reject H_o.

- There is **sufficient evidence** to support the claim of a linear correlation.

Method 2

5. Critical region:

Table A-6 yields a critical value of $r = 0.666$. We can now compare the computed value of $r = 0.947$ to the **critical values of** ± 0.666, as shown in the figure.



6. Conclusion: Because the above figure shows that the computed value of $r = 0.947$ lies beyond the upper critical value, we conclude that *there is sufficient evidence to support the claim of a linear correlation between Powerball jackpot amounts and numbers of lottery tickets sold.*

Interpreting the Linear Correlation Coefficient *r*

P-value $\leq \alpha$**:** Supports the claim of a linear correlation.

P-value $> \alpha$**:** Does not support the claim of a linear correlation.

Conclusion based on Table A-6:

Correlation If the computed linear correlation coefficient *r* lies in the left tail at or below the leftmost critical value or if it lies in the right tail at or above the rightmost critical value (**that is,** $|r| \geq$ **critical value**), conclude that there is sufficient evidence to support the claim of a linear correlation.

No Correlation If the computed linear correlation coefficient lies *between* the two critical values (that is, $|r| >$ **critical value**), conclude that there is not sufficient evidence to support the claim of a linear correlation.

OR

5. p-value:

With n - 2 = 7 degrees of freedom, using Table-A3 shows that the test statistic of t = 7.800 yields a P-value that is less than 0.01

P-value $\leq \alpha$**:** Supports the claim of a linear correlation.

P-value $> \alpha$**:** Does not support the claim of a linear correlation.

0.01 (or P-value is less than 0.01) $\leq$ 0.05 (True)

6. Conclusion: There is **sufficient evidence** to support the claim of a linear correlation.

Conclusion based on Table A-6:

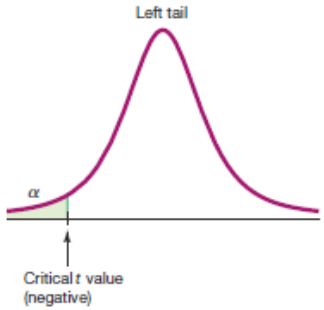
Correlation If the computed linear correlation coefficient *r* lies in the left or right tail region at or beyond the critical value for that tail, conclude that there is sufficient evidence to support the claim of a linear correlation.

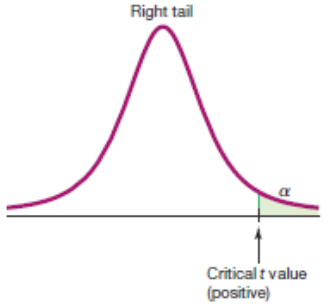
No Correlation If the computed linear correlation coefficient lies between the two critical values, conclude that there is not sufficient evidence to support the claim of a linear correlation.

TABLE A-6 Critical Values of the Pearson Correlation Coefficient r

n	$\alpha = .05$	$\alpha = .01$
4	.950	.990
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

NOTE: To test H_0: $\rho = 0$ (no correlation) against H_1: $\rho \neq 0$ (correlation), reject H_0 if the absolute value of r is greater than or equal to the critical value in the table.





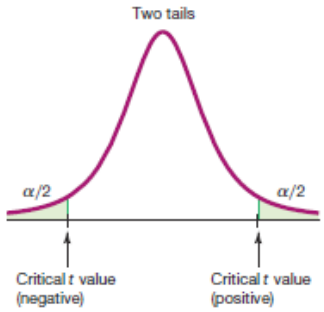


TABLE A-3 *t* Distribution: Critical *t* Values

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319
24	2.797	2.492	2.064	1.711	1.318
25	2.787	2.485	2.060	1.708	1.316
26	2.779	2.479	2.056	1.706	1.315
27	2.771	2.473	2.052	1.703	1.314
28	2.763	2.467	2.048	1.701	1.313
29	2.756	2.462	2.045	1.699	1.311
30	2.750	2.457	2.042	1.697	1.310
31	2.744	2.453	2.040	1.696	1.309
32	2.738	2.449	2.037	1.694	1.309
33	2.733	2.445	2.035	1.692	1.308
34	2.728	2.441	2.032	1.691	1.307
35	2.724	2.438	2.030	1.690	1.306
36	2.719	2.434	2.028	1.688	1.306
37	2.715	2.431	2.026	1.687	1.305
38	2.712	2.429	2.024	1.686	1.304
39	2.708	2.426	2.023	1.685	1.304
40	2.704	2.423	2.021	1.684	1.303
45	2.690	2.412	2.014	1.679	1.301
50	2.678	2.403	2.009	1.676	1.299
60	2.660	2.390	2.000	1.671	1.296
70	2.648	2.381	1.994	1.667	1.294
80	2.639	2.374	1.990	1.664	1.292
90	2.632	2.368	1.987	1.662	1.291
100	2.626	2.364	1.984	1.660	1.290
200	2.601	2.345	1.972	1.653	1.286
300	2.592	2.339	1.968	1.650	1.284
400	2.588	2.336	1.966	1.649	1.284
500	2.586	2.334	1.965	1.648	1.283
1000	2.581	2.330	1.962	1.646	1.282
2000	2.578	2.328	1.961	1.646	1.282
Large	2.576	2.326	1.960	1.645	1.282

One-Tailed Tests

The examples and exercises in this section generally involve two tailed tests, but one-tailed tests can occur with a claim of a positive linear correlation or a claim of a negative linear correlation. In such cases, the hypotheses will be as shown here.

Claim of Negative Correlation (Left-Tailed Test)	Claim of Positive Correlation (Right-Tailed Test)
$H_0: \rho = 0$	$H_0: \rho = 0$
$H_1: \rho < 0$	$H_1: \rho > 0$

Randomization Test

The **randomization method** is based on the principle that when assuming a **null hypothesis of no correlation**, we can **resample by holding the *x* data values fixed** while **randomly shuffling the order of the *y* data values**.

We do the **shuffling without replacement**, so we are working with a randomization test that can be used to test the assumption (null hypothesis) of no correlation.

Randomization Test

Shown below is Table, followed by two resamplings in which the ticket data are **shuffled in a random order.** Such **shuffles are based on the null hypothesis of no correlation**, and if we calculate the linear correlation coefficient *r* for each new shuffle, we get a list of *r* values that can be used to determine whether the actual *r* = 0.947 from the original data in **Table 1 is *significant* in the sense that it is not likely to occur by chance when there really is no correlation.**

Table 1 Powerball Tickets Sold and Jackpot Amounts

Jackpot	334	127	300	227	202	180	164	145	255
Tickets	54	16	41	27	23	18	18	16	26

Randomization Test

Random shuffling of ticket data:

Tickets	16	54	16	26	18	27	41	23	18

Another random shuffling of ticket data:

Tickets	18	16	54	27	41	23	18	16	26

Randomization Test

- Using the paired data in Table 1, we can use Python or Matlab to create **1000 samples using the preceding method of shuffling**.

- Here is one result from technology: **Among the 1000 values of *r* created by shuffling the ticket values as described above, none of them are at least as extreme as the value** of $r = 0.947$ found from the original data in Table 1.

Randomization Test

Because a result of ***r* = 0.947 never occurred among 1000 samples**, it appears that the likelihood of such an extreme value is **around 0.000**. This shows that a value such as *r* = 0.947 is significant in the sense that it is **not likely to occur by chance.**

This suggests that there ***is* a correlation** between **the jackpot amounts and the numbers of tickets sold**.