Brief Article

The Author

May 25, 2016

Chapter 1

Why should we use data science?

—

Solution 1.7-1
All of the choices is the correct answer.

—

Solution 1.7-2
The level of measurement of marital status is nominal.

—

Solution 1.7-3
The level of measurement of marital status is category - nominal.

—

Solution 1.7-4
The number of reported robberies in June 2014 is Kalamazoo County is a continuous variable.

—

Solution 1.7-5

Numeric - interval

—

Solution 1.7-6
Numeric - ratio

—

Solution 1.7-7
all adult residents of the U.S.

—

Solution 1.7-8
birth weight.

—

Solution 1.7-9

Variable: Homicide rate
Level of Measurement: Interval-ratio
Type: Continuous
Application: Descriptive (two variables)
—

Solution 1.7-10

Variable: party, gender, opinion
Level of Measurement: nominal, nominal, ordinal
Type: discrete, discrete, discrete
Application: inferential, NA, NA

Chapter 2

Dataset Descriptive Information

—

Solution 2.6-1

$$(2.)p = \frac{250 - 195}{250} = 0.22 \tag{2.1}$$

—

Solution 2.6-2

$$(3.)ratio = \frac{195}{250 - 195} = 3.55 \tag{2.2}$$

—

Solution 2.6-3

$$(2.)rate = \frac{13}{25000} \times 100000 = 52 \tag{2.3}$$

—

Solution 2.6-4

$$(1.)PC = \frac{(83 - 89)}{89} \times 100 = -6.7\% \tag{2.4}$$

—

Solution 2.6-5
(1.) The percentage of nurses who are female is 36.8%. **(2.)** The proportion of orderlies who are males is 0.367. **(3.)** Ratio is 18 females docs to 83 males docs or approximately 1 female doc for every 5 males docs. **(4.)** Percentage of females on the staff is 43.3%.

—

Solution 2.6-6
The measure of the lost hours due to traffic is interval-ratio.

—

Solution 2.6-7

$$pc = \frac{169.53 - 159.90}{159.90} \times 100 = 6.02\% \tag{2.5}$$

—

Solution 2.6-8

Min. 1st Qu. Median Mean 3rd Qu. Max. 15.70 17.60 18.30 18.24 19.28 20.70

—

Solution 2.6-9
$p = \frac{(19+15)}{50} \times 100 = 68$

—

Solution 2.6-10
violent crime rate is [1] 5.625563

property crime rate is [1] 46.97676

Chapter 3

Measures of Location

—

Solution 3.10-1

Compute the mean, median, and mode for the weekly grocery budget mean = 32.167, median = 32.5, mode = 35

—

Solution 3.10-2

The third value must be 22 so that the average is 20 for all three quizzes.

—

Solution 3.10-3

The mean is 472.6 The median is 555 There is no mode since there are no duplicates.

1. The greater value is the median (555). **2.** There is a negative skewness in the middle half the of the dataset. (Refer to Figure 3.1.)

3. When we compare the median to the mean, we find the median (555) is greater the mean (472.6). Therefore, we can say the the dataset is left skewed.

—

Solution 3.10-4

(1.) Mode **(2.)** Median **(3.)** Mean **(4.)** Mean **(5.)** Median **(6.)** Mode

—

Solution 3.10-5
Birth Mode="North"; **Legal** Median=2.5; **Expense** Mean=48.5; **Movies** Mean=5.8; **Food** Median=6; **Religion** Mode="Protestant"

—

Solution 3.10-6

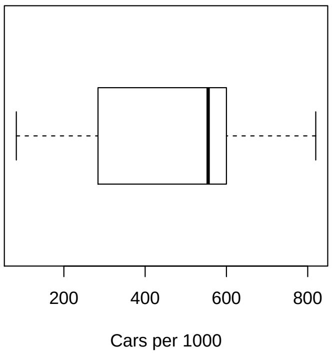

Figure 3.1: Boxplot of Cars per 1000

The mean for 2005 is 55.4. The median for 2005 is 54.5. The mean for 2015 is 57.1. The median for 2015 is 57.

—

Solution 3.10-7

The mean is [1] 31.8
The median is [1] 35

—

Solution 3.10-8

The mean is 28.72, and the median is 30

—

Solution 3.10-9

The pretest mean is 9.3333333
The pretest median is 10
The posttest mean is 12.9333333
The pretest median is 12

—

Solution 3.10-10

The sex ed pretest mean is 9.3333333
The sex ed posttest mean is 12.9333333.

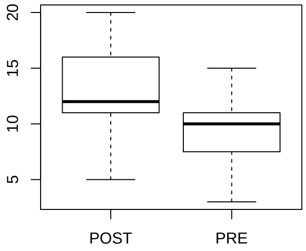

Figure 3.2: Box plot of Post - Pre Test Scores

After reviewing the box plot of the differences between posttest and pretest scores, it appears that the students learned the subject material.

Chapter 4

Measures of Spread

—

Solution 4.10-1

The range is $(X_{max} - X_{min}) = 40 - 25 = 15$

—

Solution 4.10-2

LB = Q_1 - 1.5(IQR) = 28 - 1.5 (35 - 28) = 17.5
UB = Q_3 + 1.5(IQR) = 35 + 1.5 (35 - 28) = 45.5

An outlier is any value that is less than 17.5 or greater than 45.5; therefore, no outliers in this set of data.

—

Solution 4.10-3

The standard deviation is 236.7524373

—

Solution 4.10-4

The range is 14.
The IQR is 4.775
The standard deviation is 3.2241029
The coefficient of variation is 38.8583832
The variance is 10.3948396

—

Solution 4.10-5

The LB outlier is No LB outlier and the UP outlier is No UB outlier.

—

Solution 4.10-6

First we must determine the data types for each variable:

- height is numeric, ratio continuous
- GPA is numeric, interval
- gender is categorical, nominal
- major is categorical, nominal

Note: numerical variables can use the mean, standard deviation, median, histograms, and line charts; while categorical variables can use frequency tables, mode, bar and pie charts. For example, the graphics for gender and major should be a bar or pie chart because they are categorical variables, and the graphics for GPA and height should be box plots or histograms.

—

Solution 4.10-7

The range is 41, interquartile range is 15, standard deviation is 12.6918609, and variance is 161.0833333.

—

Solution 4.10-8

The range is 41, interquartile range is 25, standard deviation is 13.6763787, and variance is 187.0433333.

—

Solution 4.10-9

The pretest range is 12 and standard deviation is 3.6187343
The posttest range is 15 and standard deviation is 4.558613

—

Solution 4.10-10

The sex education pretest mean is 9.3333333
The sex education posttest mean is 12.9333333.
After reviewing the box plot of the differences between posttest and pretest scores, it appears that the students learned the subject material.

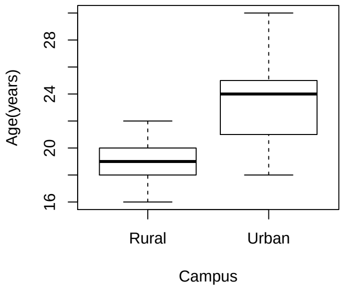

Figure 4.1: Box plot of Post - Pre Test Scores

Chapter 5

Normal Curve

—

Solution 5.8-1

$$SD = \frac{(41-32)}{3} = 3 \tag{5.1}$$

—

Solution 5.8-2

$$z = \frac{173-153}{11} = 1.82 \tag{5.2}$$

—

Solution 5.8-3

$z = \frac{143-155}{12} = -1$
$P[z < -1] = 15.87$

—

Solution 5.8-4

$x = \mu + Z\sigma$
$x = 155 + 1.28 \times (12)$
$x = 170.36$

—

Solution 5.8-5

$x = \mu + Z\sigma$
$x = 155 - 0.6745 \times (12)$
$x = 146.91$

—

Solution 5.8-6

The value of the 85^{th} percentile is 11.6414612.

There are three countries: USA(17.1), Netherland (12.9), and France (11.7).

—

Solution 5.8-7

The value of the 15^{th} percentile is 4.9585388.

There are six countries: Indonesia (3.1), Saudi Arabia (3.2), Venezuela (3.6), India (4.0), Malaysia (4.0), and Thailand (4.6).

—

Solution 5.8-8

The probability of having a mean less than 7.3 ($P[\bar{x} < 7.3]$) is 0.3782144.

—

Solution 5.8-9

The probability of having a mean greater than 9.3 ($P[\bar{x} > 9.3]$) is 0.6217856.

—

Solution 5.8-10

The probability of having a mean between 7.3 and 9.3 ($P[7.3 < \bar{x} < 9.3]$) is 0.2435711.

Chapter 6

Inferential Statistics

—

Solution 6.9-1
The population for this sample survey is all adult resident of the United States.

—

Solution 6.9-2
This is an example of stratified random sampling.

—

Solution 6.9-3
This method is called systemic random sampling.

—

Solution 6.9-4
This is an example of cluster sampling.

—

Solution 6.9-5
This is an example of simple random sampling.

—

Solution 6.9-6
What proportion will exceed 72.0 inches?

$SE = \frac{SD}{\sqrt{n}} = \frac{2.9}{\sqrt{10}} = 0.9171$

$Z = \frac{(X-\mu)}{SE}$

$Z = \frac{(72.0-69.1)}{0.9171} = 3.1623$

$P[Z > 3.1623] = 0.0008$

—

Solution 6.9-7
What is the chance that this group will average over 220?

$SE = \frac{SD}{\sqrt{n}} = \frac{28}{\sqrt{40}} = 4.4272$

$Z = \frac{(X-\mu)}{SE}$

$Z = \frac{(220-210)}{4.4272} = 2.2588$
$P[Z > 2.2588] = 0.0119$

—

Solution 6.9-8
What is the chance that the contributions will exceed 130 (4,680/36) dollars?
$SE = \frac{SD}{\sqrt{n}} = \frac{40}{\sqrt{36}} = 6.6667$
$Z = \frac{(X-\mu)}{SE}$
$Z = \frac{(130-120)}{6.6667} = 1.5$
$P[Z > 1.5] = 0.06681$

—

Solution 6.9-9
What is the chance that the contributions will be less than 115 (4,140/36) dollars?
$SE = \frac{SD}{\sqrt{n}} = \frac{40}{\sqrt{36}} = 6.6667$
$Z = \frac{(X-\mu)}{SE}$
$Z = \frac{(115-120)}{6.6667} = -0.75$
$P[Z < -0.75] = 0.2266$

—

Solution 6.9-10
What is the chance that the contributions will be between 115 (4140/36) and 125 (4500/36) dollars?
$SE = \frac{SD}{\sqrt{n}} = \frac{40}{\sqrt{36}} = 6.6667$
$Z = \frac{(X-\mu)}{SE}$
$Z_1 = \frac{(115-120)}{6.6667} = -0.75$
$Z_2 = \frac{(125-120)}{6.6667} = 0.75$
$P[-0.75 < Z < 0.75] = 0.5467$

Chapter 7

Estimation

—

Solution 7.5-1
The point estimate is $\bar{X} = 3.38$.

—

Solution 7.5-2
$SE = \frac{s}{\sqrt{n}} = \frac{0.30}{\sqrt{100}} =$ [1] 0.03

—

Solution 7.5-3

The critical value is [1] 1.984217

—

Solution 7.5-4
SE = [1] 0.03

ME = [1] 0.05952651

—

Solution 7.5-5

$CI = \bar{X} \pm ME = 3.38 \pm 0.0595$
$CI = (3.38 - 0.0595, 3.38 + 0.0595)$
$CI = (3.32, 3.44)$

—

Solution 7.5-6
The point estimate for Brazil is $p = \frac{1367}{1486} =$ [1] 0.9199192

—

Solution 7.5-7
The standard error of the estimate for China is

[1] 0.007550553

—

Solution 7.5-8

The critical value for all nations is [1] 1.959964

—

Solution 7.5-9

$ME = (CV)(SE) = 1.96 \times 0.0069 =$ [1] 0.01355432

—

Solution 7.5-10

The 95 percent CI is [1] 0.943 [1] "+/-" [1] 0.01016095

Chapter 8

Testing One Sample Hypotheses

—

Solution 8.7-1

$H_0 : \mu = 6.2$ vs. $H_A : \mu \neq 6.2$

—

Solution 8.7-2

$SE = \frac{0.7}{\sqrt{25}} = 0.14$
$t = \frac{(5.9-6.2)}{0.14} = -2.14$

—

Solution 8.7-3

From the distribution of t, using row df = 24 and column 0.025, CV(t) = 2.064.

—

Solution 8.7-4

Since the test statistic, $|t| = |-2.14| > t_{.025} = 2.064$
Therefore, reject H_0, there is difference.

—

Solution 8.7-5

$H_0 : P_u = 0.55$ vs. $H_A : P_u \neq 0.55$

—

Solution 8.7-6

$SE = \sqrt{\frac{(.55(1-.55))}{150}} = 0.0406$
$z = \frac{(.6-.55)}{0.0406}$
$z = 1.23$

—

Solution 8.7-7

Standard Normal distribution, $CV(z) = 1.96$

—

Solution 8.7-8

Conclude there is no difference, since the population, 55, percent is within the confidence interval. The 95 percent CI is (0.5516, 0.6784).

—

Solution 8.7-9

No Error was committed.

—

Solution 8.7-10

$H_0 : \mu = 2.5$ vs. $H_1 : \mu > 2.5$
$SE = \frac{0.75}{\sqrt{60}} = 0.0968$
$T = \frac{(2.6-2.5)}{0.0968} = 1.0328$
$T_{0.05,59} = 1.6711$
Since the $T < T_{0.05,59}$; fail to reject H_0.
Seniors are not significant greater than the student body.

Chapter 9

Testing Two Sample Hypotheses

—

Solution 9.6-1
$H_0 : \mu_1 = \mu_2$ vs. $H_0 : \mu_1 \neq \mu_2$

—

Solution 9.6-2
From the t-distribution table, choose df row 10 and column significance level 0.05.
Student's t distribution, $CV = 1.8125$

—

Solution 9.6-3
The test statistics is $|-1.66|$ and significance level is 1.8125. Since the test statistic is less than the significance level, fail to reject H_0.

—

Solution 9.6-4
$H_0 : P_1 = P_2$ vs. $H_A : P_1 \neq P_2$

—

Solution 9.6-5
Standard Normal (z) distribution, $Z = \pm 1.96$.

—

Solution 9.6-6
Conclude that there is a difference, since the test statistic (z = 2.04) is greater than the critical value (z = 1.96). Therefore, reject H_0

—

Solution 9.6-7
$H_0 : P_1 = P_2$ vs. $H_A : P_1 > P_2$

—

Solution 9.6-8
Standard Normal (z) distribution, $Z = \pm 1.645$, because it is a upper-tailed test.

—

Solution 9.6-9

Conclude that there is no difference since the p-value is greater than 0.05.

—

Solution 9.6-10

Conclude that there is no difference.

Chapter 10

Testing Equality of two or more Proportions

—

Solution 10.5-1

1. Significance test:

   Pearson's Chi-squared test

   data: mtxA X-squared = 1.6192, df = 2, p-value = 0.445

   Reviewing the results of the chi-square test, the test statistic ($\chi^2 = 1.6192$), degrees of freedom = 2, and the p-value = 0.445 which is greater than 0.05, we conclude that there is no difference in GPA from students who live on-campus or off-campus.

2. column percents

   The on-campus group is most likely to have a high GPA.

—

Solution 10.5-2

1. Significance test:

   Pearson's Chi-squared test

   data: mtxA X-squared = 13.983, df = 4, p-value = 0.007349

   Reviewing the results of the chi-square test, the test statistic ($\chi^2 = 13,983$), degrees of freedom = 4, and the p-value = 0.007349 which is less than 0.05, we conclude that there is a difference in quality of life and level of satisfaction with the neighborhood.

2. column percents

   The quality of life group is most likely to have a high level of neighborhood satisfaction.

—

Solution 10.5-3

Reviewing the results of the table, the independent variable is *gender*.

—

Solution 10.5-4

The hypotheses for this scenario is H_0 : Sex and having a gun in home are independent, vs. H_1 : Sex and having a gun in home are dependent.

—

Solution 10.5-5

Yes the figures look like the above figures.

—

Solution 10.5-6

Dependent Variable	Chi-square	Degrees of Freedom	Significance
have gun in home	28.351	2	0.000

Table 10.1: Record the results.

—

Solution 10.5-7

For a sample of 1711 subjects, there was a significant relationship between *gender* and *have gun in home* (chi-square = 28.351, degree of freedom = 2, and p-value = 0.000). According to the GSS2014 survey 73.7 percent of women respondents said either no or refused to have a gun in the home vs. 63.5 percent of men respondents.

Chapter 11

Testing Equality of three or more Averages

—

Solution 11.4-1

$H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_a : at least one mean is different.

—

Solution 11.4-2

Since the $F_{obtained} = 1.84 < F_{critical} = 3.68$, fail to reject H_0

—

Solution 11.4-3

The dependent variable is *age* and the independent variable is *confidence in the banking and financial systems.*

Chapter 12

Testing the relationship between numerical variables

—

Solution 12.5-1
$BAC = -0.012 + 0.017(9) = 0.14$

—

Solution 12.5-2
A moderately strong positive straight-line relationship between number of beers and BAC.

—

Solution 12.5-3
The correllation coefficient is $r = 0.875 = \sqrt{0.765}$ where r^2 is the coefficient of determination.

Chapter 13

Workshops

—

Solution 13.1-1
all adults with normal vision.

—

Solution 13.1-2
the 1,347 teachers who mail back the questionnaire.

—

Solution 13.1-3
three

—

Solution 13.1-4

Rate of California is [1] 8.768816
Rate of Florida is [1] 23.75374
Rate of Illinois is [1] 10.39501
Rate of Nevada is [1] 42.45283
Therefore, Nevada has the highest number of death row prisoners.

—

Solution 13.1-5

Rate of Michelle's income is [1] -76.19048

—

Solution 13.1-6

[1] 78
is not equal to 100%.

—

Solution 13.1-7

A good choice of a graph would be a bar chart.

—

Solution 13.1-8

right skewed, mean, median.

—

Solution 13.1-9

counts or percents, mean, median.

—

Solution 13.1-10
Min. 1st Qu. Median Mean 3rd Qu. Max. -2.0000 0.0000 1.0000 0.8182 2.0000 3.0000
$Q_1 = 0$

—

Solution 13.2-1

Min. 1st Qu. Median Mean 3rd Qu. Max. 1.00 2.00 4.00 5.00 4.75 22.00
Therefore, $Q_3 = 5$

—

Solution 13.2-2

Therefore, IQR = [1] 2.75

—

Solution 13.2-3

Standard deviation would change.

—

Solution 13.2-4

The median will be larger than the mean if the distribution is left skewed.

—

Solution 13.2-5

You made an error in your calculations.

—

Solution 13.2-6

all the observations have the same value.

—

Solution 13.2-7

The box in each box plot marks the range covered by the middle half of the data.

—

Solution 13.2-8

mean of curve A is less than mean of curve B and standard deviation of curve A is less than standard deviation of curve B.

—

Solution 13.3-1

The mean of the normal distribution is 50.

—

Solution 13.3-2

The standard deviation of the normal distribution is 10.

—

Solution 13.3-3

A number with 60 percent of the data above it is the 40^{th} percentile.

—

Solution 13.3-4

the standard deviation of the test scores is [1] 15

—

Solution 13.3-5

$P[-1 < z < 2] = P[z < 2] - P[z < -1] =$ [1] 0.8185946

—

Solution 13.3-6

$P[\bar{x} > 12] \times 1000 = (1 - P[\bar{x} < 12]) \times 1000 = (1 - P[z < .5]) \times 1000 =$ [1] 308.5375

—

Solution 13.3-7

So the median score on the exam is equal to 500.

—

Solution 13.3-8

The percent of scores are higher is [1] 0.0249979

—

Solution 13.3-9

the proportion of exceptional students among male SAT takers is about [1] 0.02275013

—

Solution 13.4-1

The point estimate is [1] 0.4199605

—

Solution 13.4-2

The standard error of your estimate is [1] 0.01551468

—

Solution 13.4-3

The critical value is [1] 1.644854

—

Solution 13.4-4

The margin error of your estimate is [1] 0.02551937

—

Solution 13.4-5

The confidence interval is [1] 0.3944411 [1] 0.4454798

—

Solution 13.4-6

The point estimate is [1] 114.9

—

Solution 13.4-7

The standard error of your estimate is [1] 1.789786

—

Solution 13.4-8

The critical value is [1] 2.055529

—

Solution 13.4-9

The margin of error is [1] 3.678957

—

Solution 13.4-10

The 95% CI is [1] 111.221 [1] 118.579

—

Solution 13.5-1

The population parameter of interest is $ [1] 8

—

Solution 13.5-2

The appropriate hypotheses are $H_0 : \mu = 8$ vs. $H_a : \mu \neq 8$

—

Solution 13.5-3

The test statistic is [1] 1.781538

—

Solution 13.5-4

The critical value is [1] -2.200985

—

Solution 13.5-5

Your conclusion is fail to reject H_0.

—

Solution 13.5-6

Population parameter of interest is 128.

—

Solution 13.5-7

The appropriate hypotheses are $H_0 : \mu = 128$ vs. $H_a : \mu > 128$

—

Solution 13.5-8

The test statistic is [1] 2.321429

—

Solution 13.5-9

The critical value is [1] 1.317836

—

Solution 13.5-10

Your conclustion is reject H_0 in favor H_a

—

Solution 13.6-1

The appropriate hypotheses are $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$

—

Solution 13.6-2

The test statistic is [1] 2.043016

—

Solution 13.6-3

The critical value is [1] 1.724718

—

Solution 13.6-4

The correct conclusion is reject H_0 in favor H_a.

—

Solution 13.6-5

It will be concluded that the two methods of learning are not equal when they are.

—

Solution 13.6-6

The hypotheses are $H_0 : p_A = p_B$ vs. $H_0 : p_A > p_B$

—

Solution 13.6-7

The test statistic is [1] 1.871063

—

Solution 13.6-8

The critical value is

—

Solution 13.6-9

The the correct conclusion in reject H_0 in favor of H_a, since the test statistic (1.87) is greater than the critical value 1.645.

—

Solution 13.6-10

It will be concluded that brand A outsells brand B when it does not.

—

Solution 13.7-1

H_0 : the type of pharmacies and waiting time are independent vs. H_a : the type of pharmacies and waiting time are dependent.

—

Solution 13.7-2

The chi-square test statistic for this data is 20.937.

—

Solution 13.7-3

The critical value for chi-square is [1] 0.3518463

—

Solution 13.7-4

The type of pharmacies and waiting time are dependent.

—

Solution 13.7-5

The hypotheses are $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ vs. H_a : at least one of the population means is different.

—

Solution 13.7-6
the test statistic 4.302

—

Solution 13.7-7

The critical value at 5 percent level of significance is [1] 3.343889

—

Solution 13.7-8

Your conclusion at the 5 percent level of significance is to reject H_0 in favor of H_a since the test statistic (4.302) is greater than the critical value (3.3439).

—

Solution 13.7-9

The median assembly time for the group who attended training program C is the highest, followed by training program D, then A and B is the lowest.

—

Solution 13.7-10

The differences for the mean assembly time for employees who attended training programs A and B do not differ as well as the mean assembly time for those who attended training programs A, C, and D. But the difference for the mean assembly time for employees who attended B and C appear to be different.

—

Solution 13.8-1

7.3

—

Solution 13.8-2

The graph shows a clear negative association

—

Solution 13.8-3

The correlation coefficient is moderately negative.

—

Solution 13.8-4

We should put hours of TV on the horizontal axis of the scatterplot of the data because it is the explanatory (independent) variable.

—

Solution 13.8-5

You conclude that people who smoke more tend to be less overweight.

—

Solution 13.8-6

This tells us that taller than average fathers tend to have taller than average sons.

—

Solution 13.8-7

Correlation coefficient between heights of fathers and heights of sons would be unchanged: equal to 0.52.

—

Solution 13.8-8

This means that the educator is confused because correlation makes no sense in this situation.

—

Solution 13.8-9

The test score goes down 1.3 points.

—

Solution 13.8-10

The correlation coefficient between hours studied and exam scores is [1] 0.9

—

Solution 13.8-11

You predict that a person with lean body mass 50 kilograms will have metabolic rate equal to [1] 1458.2

—

Solution 13.8-12

The slope of the regression line is [1] 26.9

—

Solution 13.8-13

The percent prediction to be obese in 1998 is [1] 20.128

—

Solution 13.8-14

The percent of changes in municipal bonds performance that can be explained by the straight line relationship between municipal bonds and large cap stocks is [1] 0.2025

—

Solution 13.8-15

The correlation coefficient r between a player's salary and his position makes no sense.