

The binomial formula

Suppose that n is the number of trials, as for example, rolling a die ten times. Let k be equal to the number of times a given event is to occur, for example, getting two ones, and p is the probability that the event will occur on any particular trial. The *binomial formula* can be written as

$$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Assumptions in the application of the binomial formula

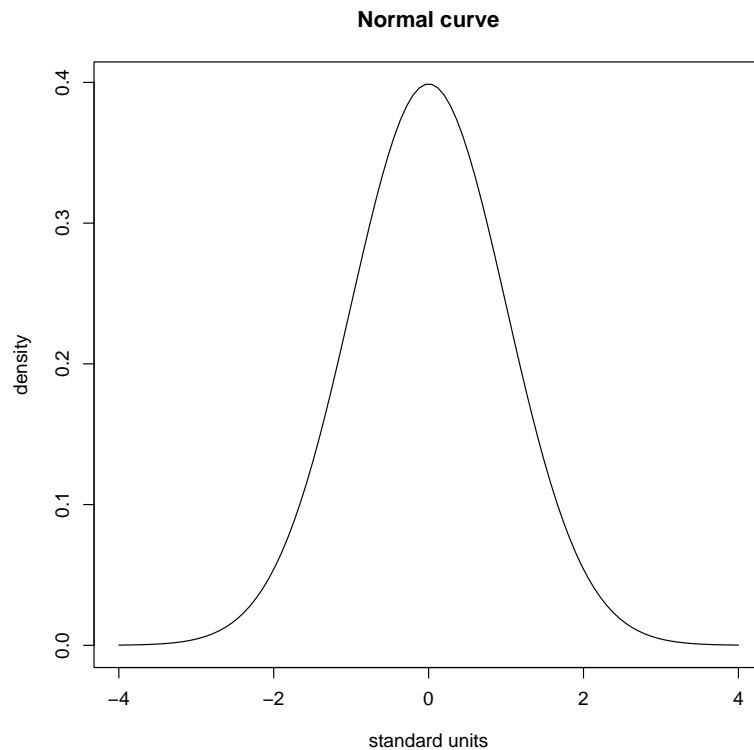
1. The value of n must be fixed in advance
2. p must be equal from trial to trial
3. The trials are independent

The normal density

The **Gaussian or normal** curve corresponds to the following formula

$$y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad e = 2.71828\dots$$

and corresponds to the graph



The area below the curve is equal to one. We observe that the curve is symmetric around zero and that most of the area is concentrated between -4 and 4 . The probability of an interval is the corresponding area under the curve.

Doing calculations with the normal curve requires the use of a table. Tables are available for the standard normal curve and they require that observations be transformed to standard units.

Given a list of numbers, we convert to standard units by subtracting the average and dividing by the SD

- $P((0, z)) = 1/2 \times P((-z, z))$
- $P((-z, x)) = P((-z, 0)) + P((0, x))$
- $P(> z) = 1/2 \times (P(< -z) + P(> z))$
- $P(< -z) + P(> z) = 1 - P((-z, z))$
- $P(< z) = P(< 0) + P((0, z))$
- $P((z, x)) = 1/2 \times (P((-x, x)) - P((-z, z))$

The Central Limit Theorem

In general it is true that the probability histogram of the sum of draws from a box of tickets will be approximated by the normal curve. This is a mathematical fact that can be expressed and proved as a theorem.

The Central Limit Theorem. When drawing at random with replacement from a box, the probability histogram for the sum will follow a normal curve, in the limit. This is even if the probability histogram of the contents of the box does not have a probability histogram that is approximately normal

The reason why the CLT is used as an approximation for distributions of lists of numbers is that it often happens that the uncertainty in the data can be thought of as the sum of several sources of randomness.

Percentages and Confidence Intervals

A box contains tickets with 0's and 1's.

The SD of the box is given by $\sqrt{(\text{fraction of 1's}) \times (\text{fraction of 0's})}$

The SE for the **sum** of 1's is $\sqrt{\text{number of draws}} \times \text{SD}$ (**square root law**)

The SE for the **percentage** of 1's is

$$\frac{\text{SE for the sum of 1's}}{\text{number of draws}} \times 100\%$$

- Sample percentage ± 1 SE is a 68% confidence interval of the population percentage.
- Sample percentage ± 2 SE is a 95% confidence interval of the population percentage.
- Sample percentage ± 3 SE is a 99.7% confidence interval of the population percentage.

Expected values and Standard errors

If we draw many times from a Box model we might add the values of draws or calculate the average of draws.

The expected value for the sum of draws =
number of draws \times average of the box

The expected value of the average of draws = average of the
box

The normal approximation can be used to calculate the chance of getting specific sum or averages.

The Standard Error is the size of the **Chance Error** after many draws.

$$\text{SE for the sum} = \sqrt{\text{number of draws}} \times \text{SD of the box}$$

$$\text{SE for average} = \frac{\text{SE for sum}}{\text{number of draws}} = \frac{\text{SD of the box}}{\sqrt{\text{number of draws}}}$$

$$\text{SE for count} = \text{SE for sum from a 0-1 box}$$

$$\text{SE for percent} = \frac{\text{SE for count}}{\text{number of draws}} \times 100\% = \frac{\text{SD of the box}}{\sqrt{\text{number of draws}}} \times 100\%$$

Test of Significance

- set up the null hypothesis
- pick a test statistics to measure the difference between the data and what is expected under the null hypothesis
- compute the test statistics and the corresponding observed significance level.

In general we are calculating a **test statistics** given by

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

which is referred to as the **z-test**.

The observed significance level is the chance of getting a test statistics as extreme or more than the observed one. This is usually denoted as P and referred to as the *P-value*

The smaller the P-value, the stronger the evidence against the null, but

The P-value is NOT the chance of the null hypothesis being right

The *t*-test

Step 1: Consider a different estimate of the SD

$$SD^+ = \sqrt{\frac{\text{number of measurements}}{\text{number of measurements} - 1}} \times SD$$

Notice that $SD^+ > SD$.

Step 2:

$$t = \frac{\text{observed} - \text{expected}}{SE^+}$$

where SE^+ corresponds to SD^+ .

Step 3: To find the observed significance level we can not use the normal curve any more. We need to use a **Student's *t* curve**. This curve depends on the **degrees of freedom** (DF). These are calculated as

$$\text{degrees of freedom} = \text{number of measurements} - 1$$

The Chi-square test

$$\chi^2 = \text{sum of } \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

When testing for independence in a table with m rows and n columns, there are $(m - 1) \times (n - 1)$ DF

H_0 : the two variables in the table are independent

H_1 : the two variables in the table are not independent

$$\text{expected value of one cell in the table} = \frac{\text{row total} \times \text{column total}}{\text{total of the table}}$$

(This calculation assumes that the two variables are independent)

Correlation

The correlation coefficient gives a measure of the linear association of two variables

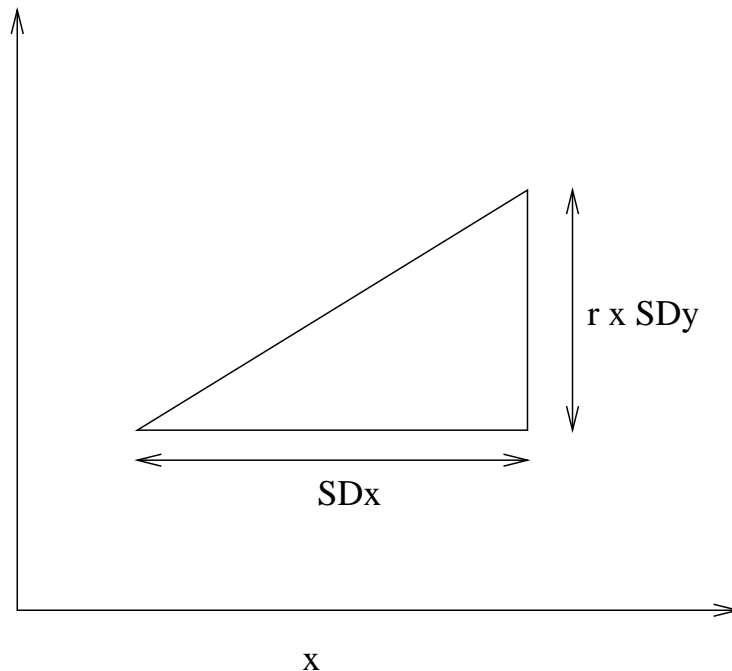
The correlation coefficient is usually denoted by r and takes values between -1 and 1

- The correlation is not affected when the two variables are interchanged.
- The correlation is not changed if the same number is added to all the values of one of the variables.
- The correlation is not changed if all the values of one of the variables is multiplied by the same positive number. It will change sign if the number is negative.
- The correlation coefficient is 1 if the variables have perfect positive linear association and -1 if they have perfect negative linear association.

Regression

The regression line for y on x estimates the average value of y corresponding to each value of x

Associated with an increase of y one SD in x there is an increase of $r \times$ SDs in y on average.



error = actual value of y - predicted value of y

$$\text{RMS error} = \sqrt{1 - r^2} \times SD \text{ of } y$$

The average of the residuals is 0 and the regression plot for the residuals is horizontal

The formula for the slope of a regression line is

$$\frac{r \times \text{SD of } y}{\text{SD of } x}$$

The intercept of the regression line is the predicted value of y for $x = 0$.

Among all possible lines through a cloud, the regression line is the one that has the smallest RMS error in predicting y from x .

Problem 1: A simple random sample of size 400 was taken from the population of all manufacturing establishments in a certain state. The results are that 16 establishments had 250 employees or more.

1. Estimate the percentage of manufacturing establishments with 250 employee or more.

4%

2. Attach a standard error to the estimate.

$$\frac{\sqrt{.04 \times .96}}{\sqrt{400}} \approx .01$$

Problem 2: A simple random sample of 1,000 persons is taken to estimate the percentage of Democrats in a large population. It turns out that 543 of the people in the sample are Democrats.

1. Calculate the sample percentage and SE for the sample percentage.

The sample percentage is given by:

$$(543/1,000) \times 100\% = 54.3\%$$

The SE for the sample percentage is:

$$\sqrt{(0.543) \times (0.457)} / \sqrt{1,000} \times 100\% = 1.6\%$$

2. Find an approximate 95% confidence interval for the percentage of Democrats in the population.

Two SE's corresponds to 3.2%. Thus a 95% confidence interval is given by (51.1%,57.5%)

Problem 3: The speed of light is measured 25 times by a new procedure. The 25 measurements are recorded and show no trend or pattern. The average of the measurements is 299,789.2 kilometers per second and the SD is 12 kilometers per second. Find an approximate 95% confidence interval for the speed of light.

1. Calculate the SE of the average.

The SE is given by $12/\sqrt{25} = 12/5 = 2.4$.

2. Find an approximate 95% confidence interval for the speed of light.

Two SE's correspond to 4.8 km per second. Thus a 95% confidence interval is given by (299,784.4 , 299,794).

Problem 4: Find the area under a Student's t curve with 3 degrees of freedom in the following cases:

1. To the right of 2.35.

5%

2. To the left of -2.35.

5%

3. Between -2.35 and 2.35.

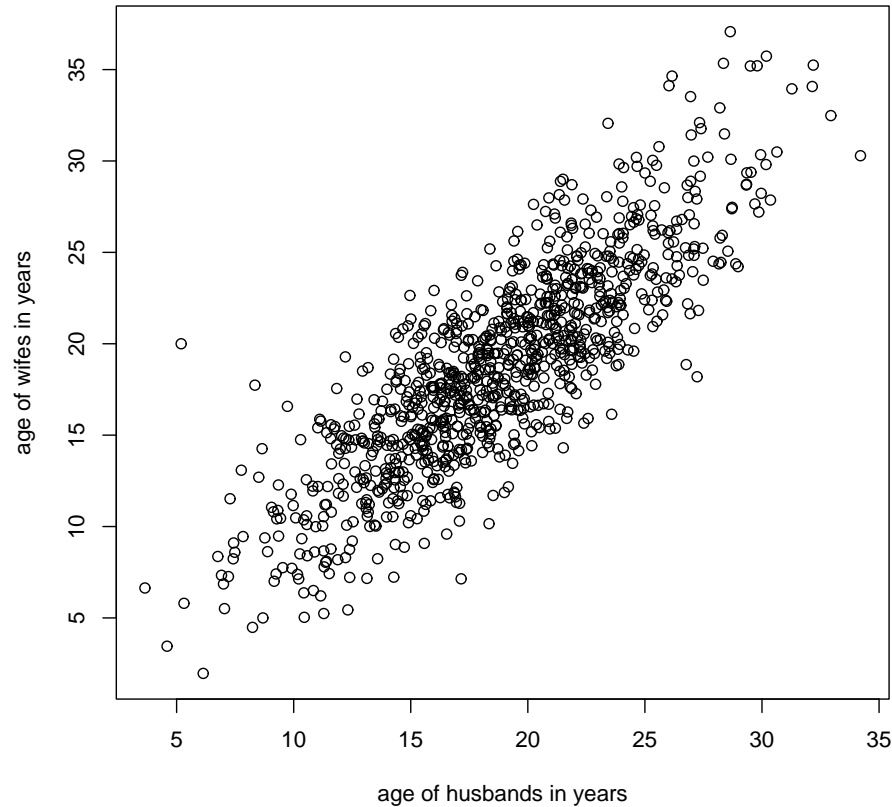
90%

4. Are these values higher or lower than the ones that correspond to the standard normal curve?

a) and b) are smaller for the normal, as a consequence, c) is larger.

Problem 5: Looking at data and making sense of them is the first step of a statistical analysis.

The scatterdiagram below shows the ages of 1,000 husbands and wives in a town in California. Explore the plot. Is there anything wrong with the data?



The range of x does not correspond to the usual range of married men. In particular, there is a 5 years old man married to a 20 years old woman.

Problem 6: True or false:

1. To make a t test with 4 measurements use a Student's t curve with 4 degrees of freedom.

F

2. For a given experiment the null hypothesis is that the average is equal to 231 units. The alternative hypothesis is that the average is above 231 units. You compute a z -test and the corresponding value P -value is 2.5%. The conclusion is that the probability that the average is equal to 231 units is 2.5%.

F

3. The R.M.S. error for a regression line of y on x is less than or equal to the SD of y .

T

4. The correlation between the daily minimum temperatures of L.A. and San Francisco is higher when measured in Fahrenheit

than when it is measured in Celsius.

F

5. The correlation between two variables is -0.92 , this implies that there is a strong negative linear association between the variables.

T

Problem 7: A newspaper article says that on average, college freshmen spend 7.5 hours a week going to parties. One administrator does not believe that these figures apply at her college, which has nearly 3,000 freshmen. She takes a simple random sample of 100 freshmen, and interviews them. On average, they report 6.6 hours a week going to parties, and the SD is 9 hours.

1. Formulate the null and the alternative hypothesis.

H_0 : The average number of hours a week that college freshmen go to parties is 7.5

H_1 : The average number of hours a week that college freshmen go to parties is less than 7.5

2. Is the difference between 6.6 and 7.5 real?

$$\frac{6.6 - 7.5}{9/\sqrt{100}} = -1$$

3. What is your conclusion?

There is not enough evidence in the data to reject the null hypothesis since the P -value is close to 0.16.

Problem 8: A statistical analysis is made of the midterm and final scores in a large class. The results are

average midterm score ≈ 60 , SD ≈ 15

average final score ≈ 65 , SD ≈ 20 , $r \approx 0.50$

1. Using the normal approximation, about what percentage of the students scored over 80 on the midterm?

80 points on the midterm corresponds to

$$\frac{80 - 60}{15} = 1.33$$

standard units. Using the normal we obtain that approximately 9% of the students scored over 80 on the midterm.

2. What is the R.M.S. error?

$$\sqrt{1 - .5^2} \times 20 = 17.32$$

3. What is the slope of the regression line?

$$\frac{0.5 \times 20}{15} = 0.67$$

4. What is the predicted final score for a student who scored 80 in the midterm?

80 points on the midterm is 1.33 SD units above average. This corresponds to $1.33 \times 0.5 = 0.67$ SD above average on the final. That corresponds to $0.67 \times 20 = 13.4$ points over average on the final, so the students that scored 80 on the midterm, scored, on average, $65 + 13.4 = 78.4$ on the final.

5. Of the students who scored 80 on the midterm, about what percentage scored over 80 on the final?

In standard units we have

$$\frac{80 - 78.4}{17.32} = 0.09$$

and there is an area of about 46% to the right of this value under the normal curve.

Problem 9: Each respondent in the Current Population Survey of March 1993 was classified as employed, unemployed or outside the labor force. The results for men in California age 35-44 can be cross tabulated by marital status as follows:

	Married	Widowed, divorced or separated	never married	Total
Employed	679 (654)	103 (109)	114 (133)	896
Unemployed	63 (68)	10 (11)	20 (14)	93
Not in the labor force	42 (62)	18 (10)	25 (13)	85
Total	784	131	159	1074

Men of different marital status seem to have different distributions of labor force status.

1. What is the null hypothesis relevant to the former table?
 H_0 : marital status and employment status are independent.
2. The expected values under the null hypothesis are given in parenthesis, calculate the χ^2 -test.

$$\begin{aligned} & \frac{(679 - 654)^2}{654} + \frac{(103 - 109)^2}{109} + \frac{(114 - 133)^2}{133} \\ & + \frac{(63 - 68)^2}{68} + \frac{(10 - 11)^2}{11} + \frac{(20 - 14)^2}{14} + \\ & \frac{(42 - 62)^2}{62} + \frac{(18 - 10)^2}{10} + \frac{(25 - 13)^2}{13} \approx 31 \end{aligned}$$

3. How many degrees of freedom has the test?

$$(3 - 1) \times (3 - 1) = 4$$

4. What are your conclusions?

The P -value is smaller than 1%, so the null hypothesis is

rejected. There is strong evidence in the data that the two variables are NOT independent.