Statistics

Outline

- Basic concept
- Statistical inference
- Statistic tests
- Machine learning

Basic concept

- Mean
  - the average of all numbers

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - For example, the mean of five values: 4, 36, 45, 50, 75 is

$$\frac{4+36+45+50+75}{5} = \frac{210}{5} = 42.$$

- Medium
  - the middle value, very useful when extreme value exits in a set of numbers, such as 1,3,5,7,100, the average value 23.2 does not represent the true center of the distribution
  - For example, the medium of five values: 4, 36, 45, 50, 75 is 45
  - For example, the medium of six values: 4,5,7,20,25,40 is (7+20)/2=13.5

Basic concept

- Variance
  - how far a set of (random) numbers are spread out from their mean

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2$$

  - For example, the variance of five values: 4, 36, 45, 50, 75 is

$((4-42)^2+(36-42)^2+(45-42)^2+(50-42)^2+(75-42)^2)/5=528.4$

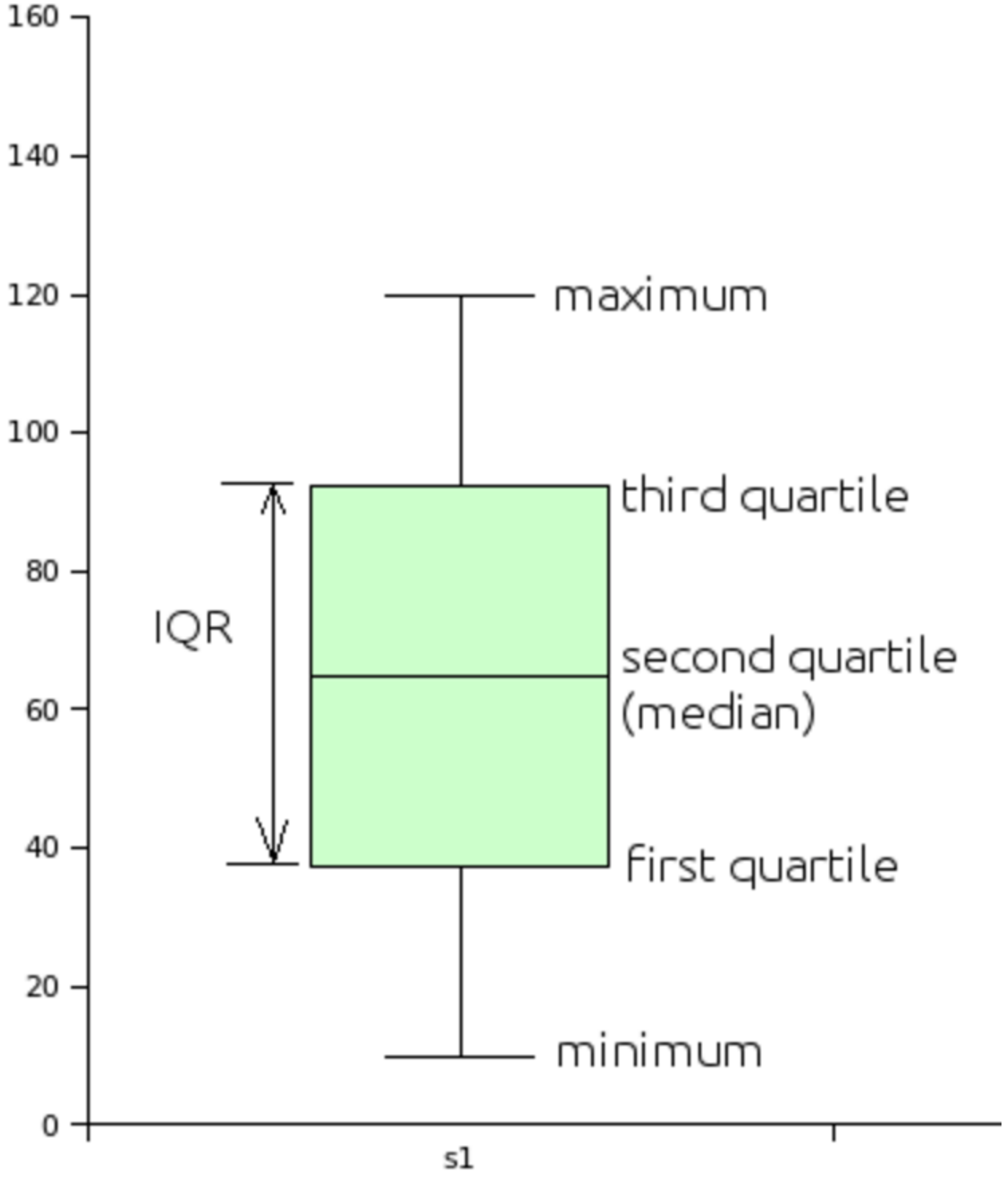
- Standard deviation
  - is the root square of variance
  - For example, the variance of five values: 4, 36, 45, 50, 75 is

$(((4-42)^2+(36-42)^2+(45-42)^2+(50-42)^2+(75-42)^2)/5)$^0.5=(528.4)^0.5=22.987

Boxplot

- Differences in the Distribution of Data
  - Range: the maximum data value minus the minimum data value.
    - The range is a useful basic statistic that provides information on the distance between the most extreme values in the data set.

- Percentiles and Quartiles
  - The median is the value that is the middle value in a sorted list of values. At the median 50% of the data values are below and 50% are above. This is also called the 50th percentile for being 50% of the way "through" the data.
  - If one starts at the minimum, 25% of the way "through" the data, the point at which 25% of the values are smaller, is the 25th percentile. The value that is 25% of the way "through" the data is also called the first quartile.
  - Moving on "through" the data to the median, the median is also called the second quartile.
  - Moving past the median, 75% of the way "through" the data is the 75th percentile also known as the third quartile.



Calculation of boxplot

In general, the formula for finding the p-th percentile in an ordered data set with n values is

$$r = \frac{p}{100}(n-1)+1$$

This gives us the rank, r, of the p-th percentile.

for a set of numbers: 14,17,45,20,19,36,7,30,8

Step 1: Sort the values from the smallest to the largest:

7,8,14,17,19,20,30,36,45

Step 2: Find the minimum

equivalent to the 0th percentile.

$$r = \frac{p}{100}(n-1)+1$$

$$r = \frac{0}{100}(9-1)+1 = 1.08 \cong 1$$

This confirms that the minimum value is the first value in the list, namely 7.

Step 3: Find the maximum

equivalent to the 100th percentile

$$r = \frac{p}{100}(n-1)+1$$

$$r = \frac{100}{100}(9-1)+1 = 9$$

This confirms that the maximum value is the last (the ninth) value in the list, namely 45.

Step 4: Find the median

equivalent to the 50th percentile.

$$r = \frac{p}{100}(n-1)+1$$

$$r = \frac{50}{100}(9-1)+1 = 5$$

This shows that the median is in the middle (at the fifth position) of the ordered data set. Therefore the median value is 19.

Calculation of boxplot

Step 4: Find the 25% percentile and 75% percentile
is the value at which 25%/75% of the data values are below this value

for 7,8,14,17,19,20,30,36,45

$$r_{25}= 25/100*(9-1)+1=3$$

$$r_{75}= 75/100*(9-1)+1=7$$

For the 25th percentile the rank is 3, that is14
For 75th percentile the rank is 7, that is 30

for 3,6,7,7,9,11,12,16,31,35,40,45

$$r_{25}= 25/100*(12-1)+1=3.75$$

$$r_{75}= 75/100*(12-1)+1=9.25$$

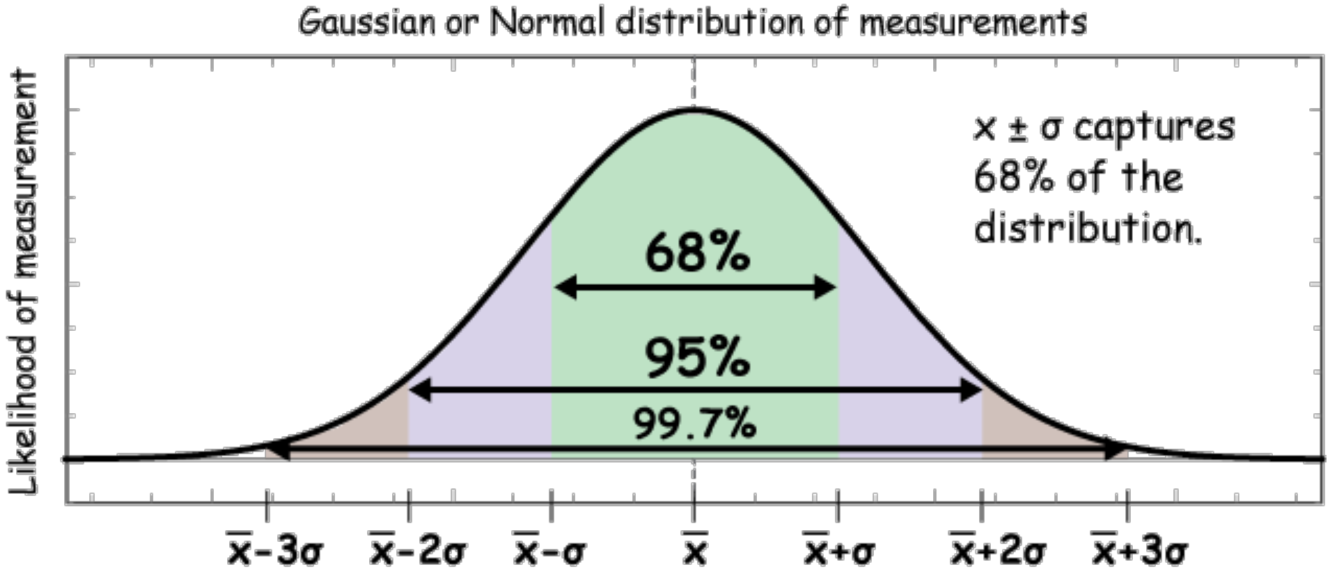
For the 25th percentile the rank is 3.75, which is between the third and fourth values. Since both these values are equal to 7, the 25th percentile is 7.
For 75th percentile the rank is 9.25, meaning between the ninth and tenth values. Therefore the 75th percentile is (31+35)/2=33

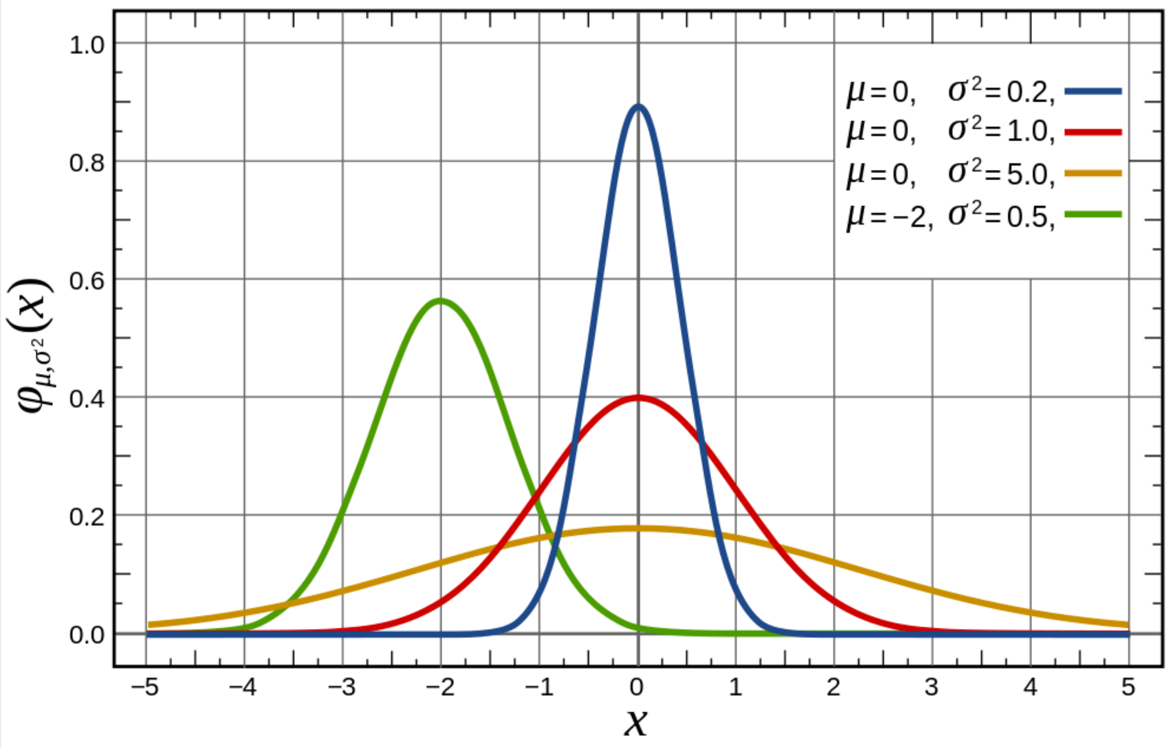
Normal distribution

- Normal (Gaussian) Distribution
  - a very common continuous probability distribution
  - The probability density of the normal distribution is:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- μ is mean or expectation of the distribution (and also its median and center)
- σ is standard deviation, measures how wide/narrow the distribution is
- σ^2 is variance

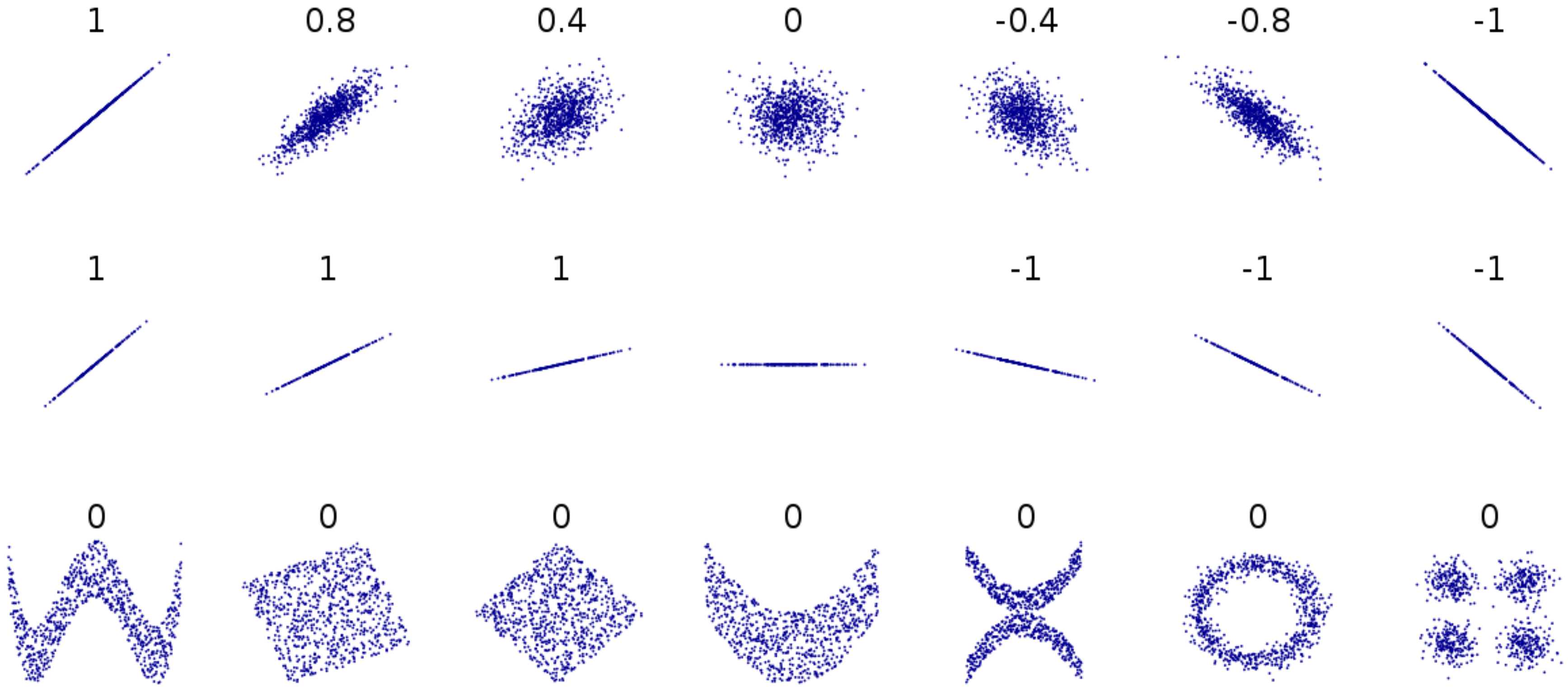




Correlation

- Correlation
  - helps quantify the linear dependence between two quantities
  - -1<=ρ<=1
  - -1: indirect relationship; 1: direct relationship

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n\sigma_x\sigma_y}$$



P value

- P value
  - the probability for a given statistical model that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two compared groups) would be the same as or more extreme than the actual observed results.
  - It is the degree of confidence with which we can reject the null hypothesis (H0)
  - It is the measure of evidence that we have against the null hypothesis (H0)
  - p-value can range anywhere between 0 & 1 and is interpreted in the following way (Here 5 % is called the significance level)-
    - A small p-value (typically ≤ 0.05 or 5 %) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
    - A large p-value (> 0.05 or 5 %) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
    - p-values very close to the cutoff (0.05 0r 5 %) are considered to be marginal (could go either way). Always report the p-value so your readers can draw their own conclusions

How to calculate P value

- Calculate P value
  - p- value is calculated for various distributions such as normal, t-distribution, chi square distribution using standard tables by finding the area under the graph.

P-Value Approach

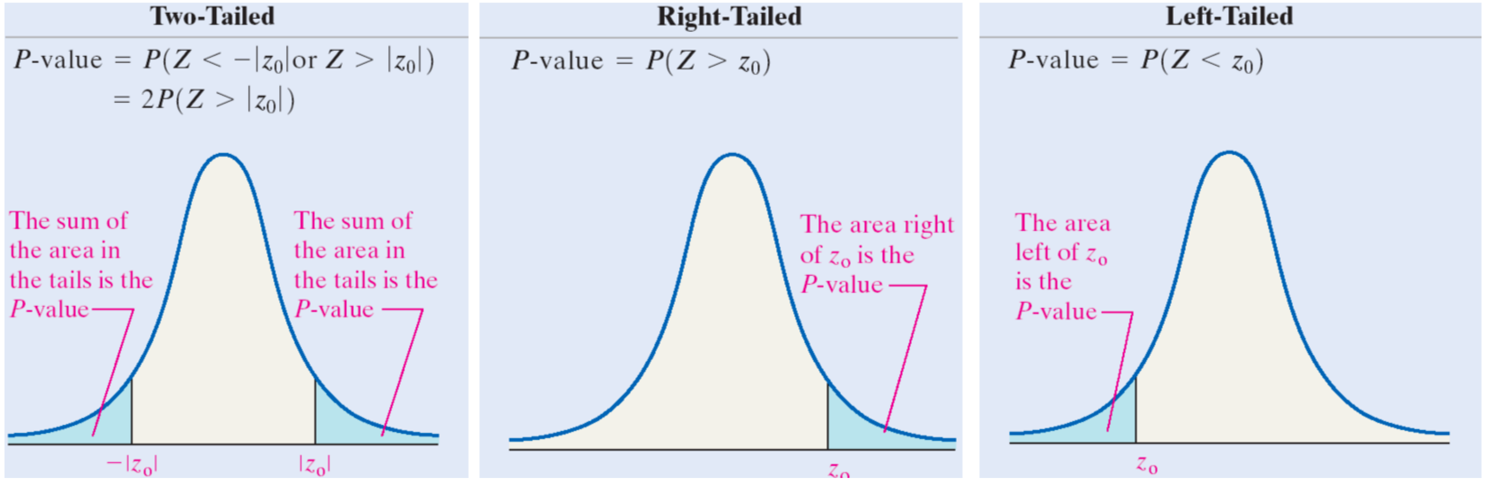
Assume that the null hypothesis is true.

The P-Value is the probability of observing a sample mean that is as or more extreme than the observed.

How to compute the P-Value for each type of test:

Step 1: Compute the test statistic $z_0 = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$

Two-tail Right Tail Left Tail



In the graphs the shaded area is called the ‘rejection region’ and it corresponds to the Z0 value (Area = 0.05 0r 5 % for single tailed test and Area = 0.025 or 2.5 % for two tailed test) if Z (test-statistic) lies in the rejection region it means p < 0.05 for single tailed and p <0.025 for two tailed and hence we can reject the null hypothesis.

Statistical inference

- Statistical inference
  - the process of deducing properties of an underlying distribution by analysis of data

- Statistical hypothesis
  - a hypothesis that is testable on the basis of observing a process that is modeled by a set of random variables

- Procedure of a hypothesis test
  - State null and alternative hypothesis
    - Null hypothesis: H_0, is usually the hypothesis that sample observations result purely from chance.
    - Alternative hypothesis:H_a, is the hypothesis that sample observations are influenced by some non-random cause.
  - Assume null hypothesis is true, calculate the value of the test statistic (such as p value)
  - Based on p value to determine which hypothesis is true
    - if $p>0.05$, retain H_0
    - if $p<0.05$, reject H_0 in favor of H_a

One sample T-test

- One sample T-test
  - Determine whether the mean of a group differs from a specified value
  - The test statistic for a One Sample t Test is denoted t, which is calculated using the following formula:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

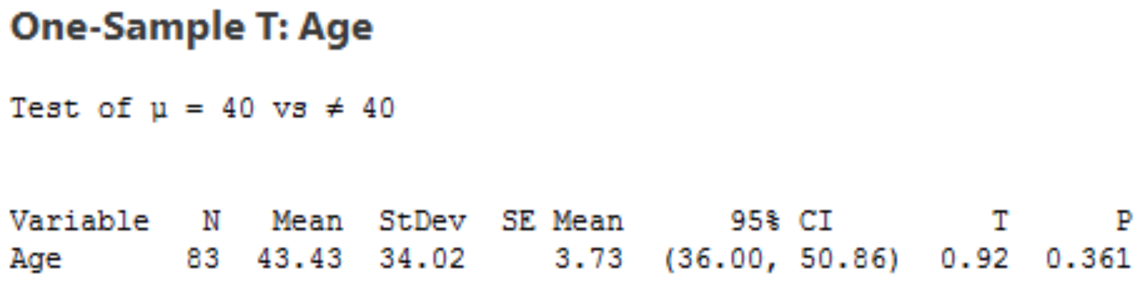
  - where

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

  - where
    - μ = Proposed constant for the population mean
    - $\bar{x}$ = Sample mean
    - n = Sample size (i.e., number of observations)
    - s = Sample standard deviation
    - $s_{\bar{x}}$ = Estimated standard error of the mean (s/sqrt(n))
  - The calculated t value is then compared to the critical t value from the t distribution table with degrees of freedom df = n - 1 and chosen confidence level. If the calculated p value<0.05, then we reject the null hypothesis.

Example of one sample T-test

- One sample T-test
  - Null hypothesis: the mean of the age is 40
  - Alternative hypothesis: the mean of the age is not 40
  - conduct one sample t-test

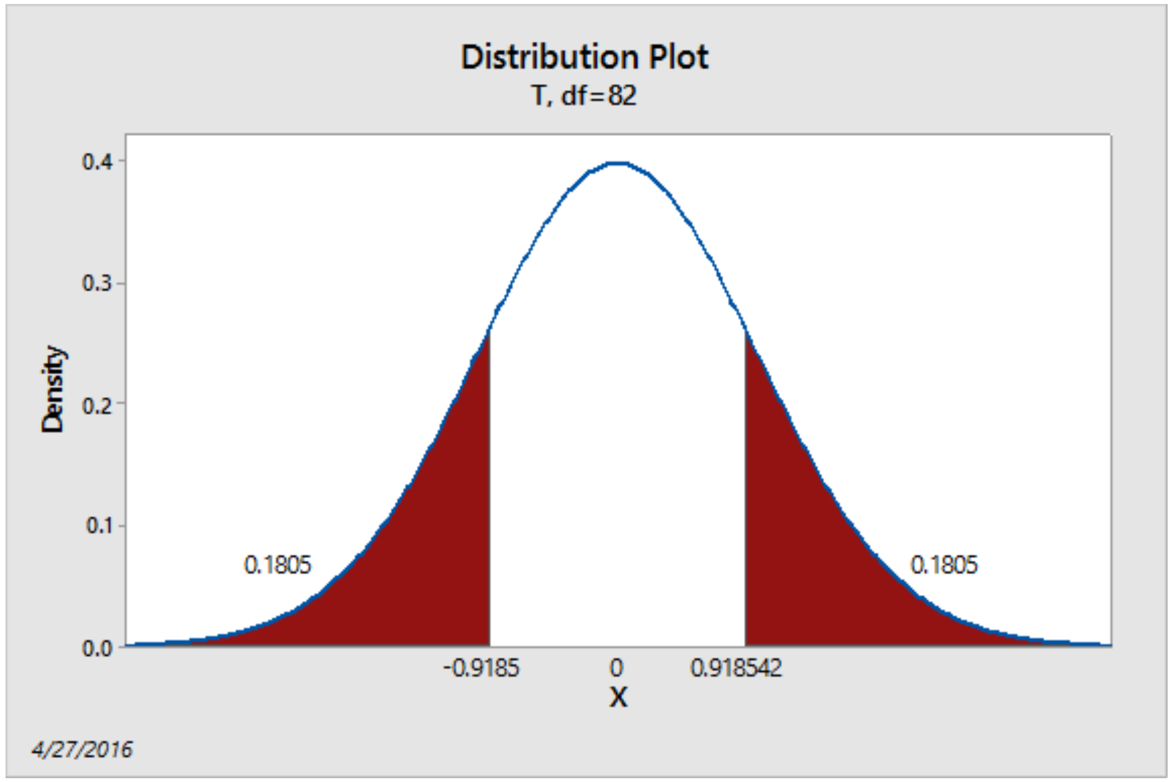


One-Sample T: Age

```
Test of μ = 40 vs ≠ 40


Variable   N   Mean  StDev  SE Mean       95% CI          T      P
Age       83  43.43  34.02     3.73  (36.00, 50.86)  0.92  0.361
```

- The output from the 1-sample t test above gives us all the information we need to plug the values into our formula:
- Sample mean: 43.43
- Sample standard deviation: 34.02
- Sample size: 83
- We also know that our target or hypothesized value for the mean is 40.
- Using the numbers above to calculate the t-statistic we see: t = (43.43-40)/34.02/√83) = 0.918542 (which rounds to 0.92)



- We add together the probabilities from both tails, 0.1805 + 0.1805 and that equals 0.361, that is the p-value, which is greater than 0.05
- Based on P value, retain H_0
- conclude that the average age of bears is 40.

Two sample T-test

- Two sample T-test
  - Determine whether the means of two independent groups differ
  - Assumptions:
    - The populations from which the samples are drawn are normally dist.
    - The standard deviations of the two populations are equal.
    - Sample observations are randomly drawn and independent.
  - Hypothesis:
    - H0: Variable A and Variable B have same mean values
    - Ha: Variable A and Variable B have different mean values
  - Calculate t test statistic:
    - The calculated t value is then compared to the critical t value from the t distribution table with degrees of freedom df = n - 1 and chosen confidence level. If the calculated p value<0.05, then we reject the null hypothesis.

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{n_1+n_2-2}$$

F test

- F test
  - Test if two population variances are equal
  - Assumptions:
    - The populations from which the samples are drawn are normally dist.
    - Sample observations are randomly drawn and independent.
  - Hypothesis:
    - H0: Variable A and Variable B have same variances
    - Ha: Variable A and Variable B have different variances
  - Calculate F test statistic:
    - Compare the test statistic value with a standard table of F-values to determine whether the test statistic surpasses the threshold of statistical significance (yielding a significant p-value).

      If the calculated p value<0.05, then we reject the null hypothesis.

$$F^* = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1}$$

One-Way ANOVA test

- One-Way ANOVA
  - Asses the equality of means of two or more groups, when two groups, equal to two-sample T test.
  - Assumptions:
    - The populations from which the samples are drawn are normally dist.
    - The standard deviations of the populations are equal.
    - Sample observations are randomly drawn and independent.
  - Calculate F test statistic:
    - $MS_{BetweenGroups}$: between-group mean square value
    - $MS_{WithinGroups}$:within-group mean square value

$$F^* = \frac{MS_{BetweenGroups}}{MS_{WithinGroups}} \sim F_{k-1,N-k}$$

Example of One-Way ANOVA

- One-Way ANOVA
  - Null hypothesis: for the overall F-test for this experiment would be that all three levels of the factor produce the same response, on average.
  - To calculate the F-ratio:
    - Step 1: Calculate the mean within each group:

$$\overline{Y}_1 = \frac{1}{6}\sum Y_{1i} = \frac{6+8+4+5+3+4}{6} = 5$$
$$\overline{Y}_2 = \frac{1}{6}\sum Y_{2i} = \frac{8+12+9+11+6+8}{6} = 9$$
$$\overline{Y}_3 = \frac{1}{6}\sum Y_{3i} = \frac{13+9+11+8+7+12}{6} = 10$$

    - Step 2: Calculate the overall mean:

$$\overline{Y} = \frac{\sum_i \overline{Y}_i}{a} = \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3}{a} = \frac{5+9+10}{3} = 8$$

where *a* is the number of groups.

    - Step 3: Calculate the "between-group" sum of squared differences:

$$S_B = n(\overline{Y}_1 - \overline{Y})^2 + n(\overline{Y}_2 - \overline{Y})^2 + n(\overline{Y}_3 - \overline{Y})^2$$
$$= 6(5-8)^2 + 6(9-8)^2 + 6(10-8)^2 = 84$$

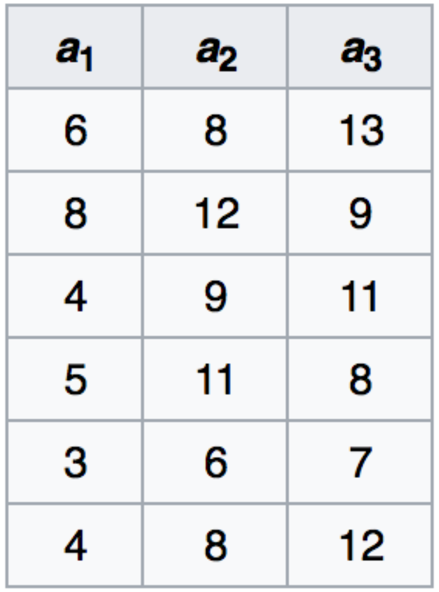
where *n* is the number of data values per group.

The between-group degrees of freedom is one less than the number of groups

$$f_b = 3 - 1 = 2$$

so the between-group mean square value is

$$MS_B = 84/2 = 42$$



a_1	a_2	a_3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

Example of One-Way ANOVA

- One-Way ANOVA
  - Step 4: Calculate the "within-group" sum of squares. Begin by centering the data in each group

a_1	a_2	a_3
6−5=1	8−9=−1	13−10=3
8−5=3	12−9=3	9−10=−1
4−5=−1	9−9=0	11−10=1
5−5=0	11−9=2	8−10=−2
3−5=−2	6−9=−3	7−10=−3
4−5=−1	8−9=−1	12−10=2

  - The within-group sum of squares is the sum of squares of all 18 values in this table

$$\begin{aligned} S_W =&(1)^2+(3)^2+(-1)^2+(0)^2+(-2)^2+(-1)^2+\\ &(-1)^2+(3)^2+(0)^2+(2)^2+(-3)^2+(-1)^2+\\ &(3)^2+(-1)^2+(1)^2+(-2)^2+(-3)^2+(2)^2\\ =&\,1+9+1+0+4+1+1+9+0+4+9+1+9+1+1+4+9+4\\ =&\,68 \end{aligned}$$

  - The within-group degrees of freedom is

$$f_W = a(n-1) = 3(6-1) = 15$$

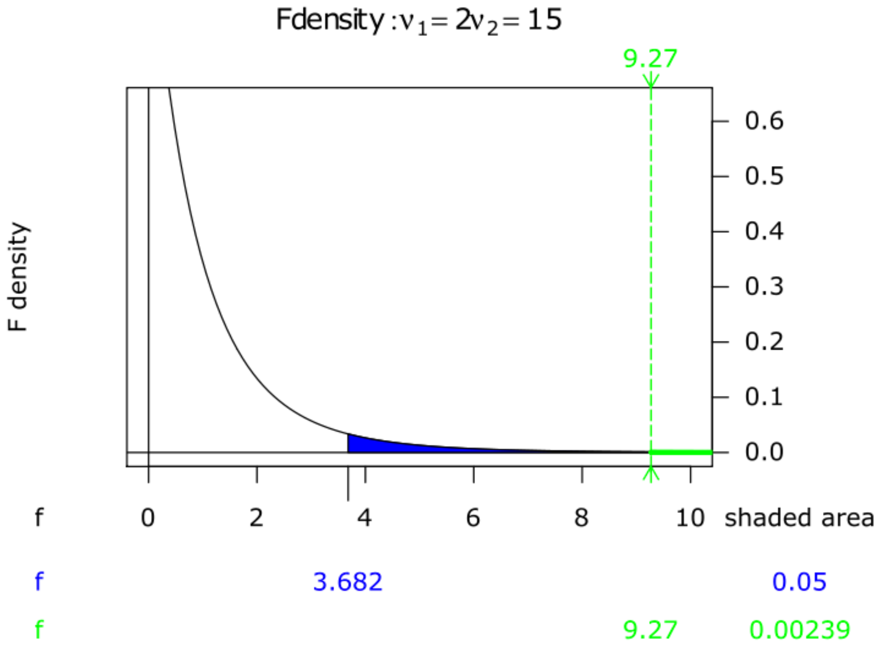
  - The within-group mean square value is is

$$MS_W = S_W / f_W = 68/15 \approx 4.5$$

  - Step 5: The F-ratio is

.

$$F = \frac{MS_B}{MS_W} \approx 42/4.5 \approx 9.3$$



- The critical value is the number that the test statistic must exceed to reject the test. In this case, Fcrit(2,15) = 3.68 at α = 0.05. Since F=9.3 > 3.68, the results are significant at the 5% significance level. One would reject the null hypothesis, concluding that there is strong evidence that the expected values in the three groups differ. The p-value for this test is 0.002.

Chi-square Test of Independence

- Chi-square Test of Independence
  - Test the independence of two categorical variables
  - Assumptions: Sample observations are randomly drawn and independent
  - Hypothesis:
    - H0: Variable A and Variable B are independent.
    - Ha: Variable A and Variable B are not independent.
  - Calculate chi-square test statistic:
    - Step 1: Calculate degrees of freedom. The degrees of freedom (DF) is equal to:

$$DF = (r - 1) * (c - 1)$$

- where r is the number of levels for one categorical variable, and c is the number of levels for the other categorical variable.

color	size
red	big
green	small
green	small
blue	small
red	medium
green	big
blue	medium

r=3
c=3
DF=(3-1)*(3-1)=4

Chi-square Test of Independence

- Step 2: Calculate expected frequencies. The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute r * c expected frequencies, according to the following formula.

$$E_{r,c} = (n_r * n_c) / n$$

  - where Er,c is the expected frequency count for level r of Variable A and level c of Variable B, nr is the total number of sample observations at level r of Variable A, nc is the total number of sample observations at level c of Variable B, and n is the total sample size.

color	size
red	big
green	small
green	small
blue	small
red	medium
green	big
blue	medium

arrangement and combination of A and B: 9 types

red+small: Er,c=(2*3)/7

red+medium: Er,c=(2*2)/7

red+big: Er,c=(2*2)/7

green+small: Er,c=(3*3)/7

green+medium: Er,c=(3*2)/7

green+big: Er,c=(3*2)/7

blue+small: Er,c=(2*3)/7

blue+medium: Er,c=(2*2)/7

blue+big: Er,c=(2*2)/7

Chi-square Test of Independence

- Step 3: Calculate test statistic. The test statistic is a chi-square random variable (X^2) defined by the following equation.

$$X^2 = \Sigma \, [\, (O_{r,c} - E_{r,c})^2 / E_{r,c} \,]$$

  - where Or,c is the observed frequency count at level r of Variable A and level c of Variable B, and Er,c is the expected frequency count at level r of Variable A and level c of Variable B.

color	size
red	big
green	small
green	small
blue	small
red	medium
green	big
blue	medium

arrangement and combination of A and B: 9 types

red+small: Er,c=(2*3)/7, Or,c=0

red+medium: Er,c=(2*2)/7, Or,c=1

red+big: Er,c=(2*2)/7, Or,c=1

green+small: Er,c=(3*3)/7, Or,c=2

green+medium: Er,c=(3*2)/7, Or,c=0

green+big: Er,c=(3*2)/7, Or,c=1

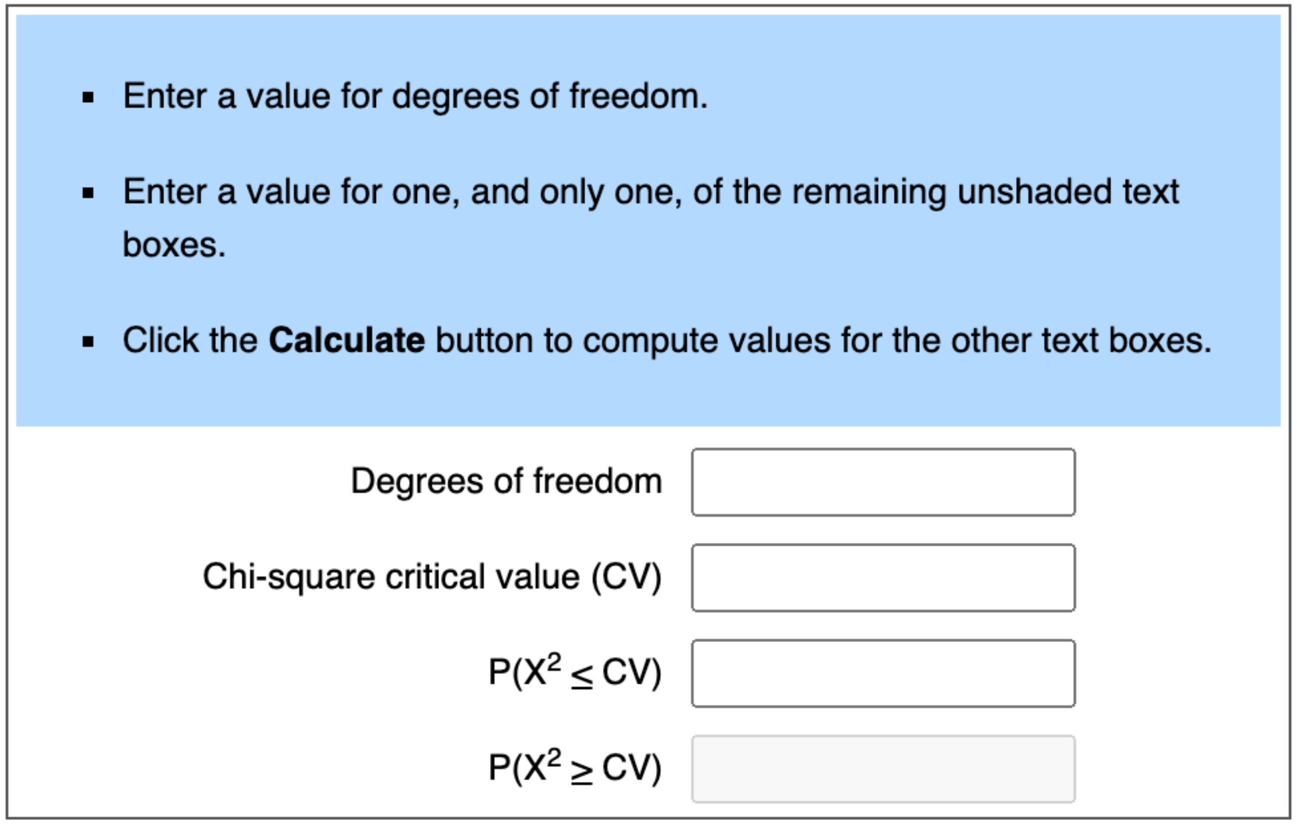
blue+small: Er,c=(2*3)/7, Or,c=1

blue+medium: Er,c=(2*2)/7, Or,c=1

blue+big: Er,c=(2*2)/7, Or,c=0

Chi-square Test of Independence

- Step 4: Calculate P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.
- Based on p value to determine which hypothesis is true
  - if $p>0.05$, retain H_0
  - if $p<0.05$, reject H_0 in favor of H_a



- Enter a value for degrees of freedom.
- Enter a value for one, and only one, of the remaining unshaded text boxes.
- Click the **Calculate** button to compute values for the other text boxes.

Degrees of freedom	
Chi-square critical value (CV)	
$P(X^2 \leq CV)$	
$P(X^2 \geq CV)$	

Machine learning

- Machine learning
  - a core sub-area of artificial intelligence as it enables computers to get into a mode of self-learning without being explicitly programmed
  - When exposed to new data, computer programs, are enabled to learn, grow, change, and develop by themselves

- Supervised learning algorithm
  - inferring a function from ‘labeled’ training data
  - Each example is a pair consisting of an input object and a desired output value
    - regression: predict a continuous output given a set of input variables
    - classification: predict a categorical output given a set of input variables

- Unsupervised learning algorithm
  - inferring a function to describe hidden structure from ‘unlabeled’ data
  - No output
    - clustering: organize objects into groups whose members are similar in some way
    - dimension reduction:convert a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely