Elementary maths for GMT

Probability and Statistics

Part 3: Multidimensional Statistics

Covariance

- The covariance is the extent to which two variables vary together. The variance is a special case of covariance
- Let X and Y be two real-valued random variables
- Covariance definition

$$Cov(X,Y) = E((X - E(X)) \times (Y - E(Y))$$

or, equivalent

$$Cov(X,Y) = E(X \times Y) - E(X) \times E(Y)$$

- Reminder
  - $Var(X) = E(X^2) - E(X)^2$
  - $Cov(X,X) = Var(X)$

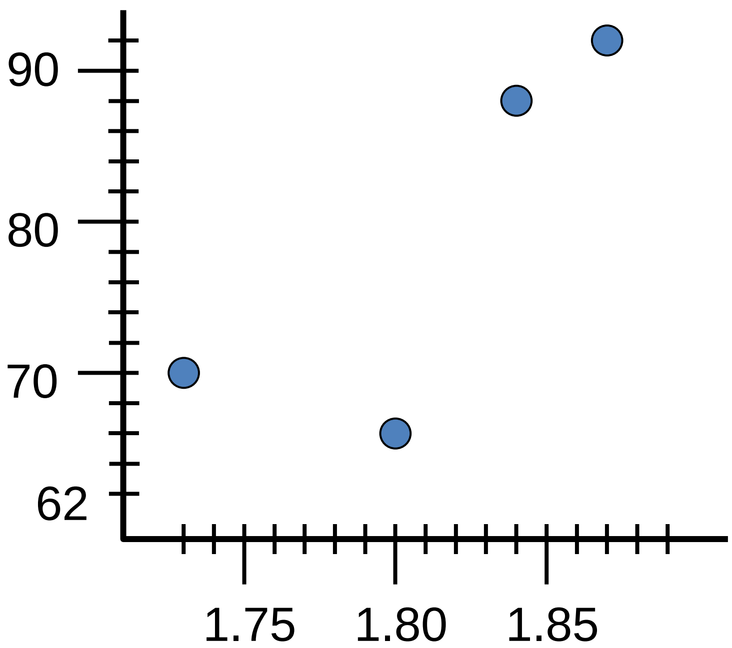
Covariance: example 1

- Suppose some measurements are
  $(X, Y) = (length, weight)$:
  $\{(1.80, 66), (1.87, 92), (1.84, 88), (1.73, 70)\}$
- $E(X) = 1.81\ m$
- $E(Y) = 79\ kg$
- $E(X \times Y) = \frac{1}{n}\sum_{i=1}^{n} x_i \times y_i = 143.465$
- $Cov(X, Y) = E(X \times Y) - E(X) \times E(Y) = 143.465 - 1.81 \times 79 = 0.475\ kg.m$

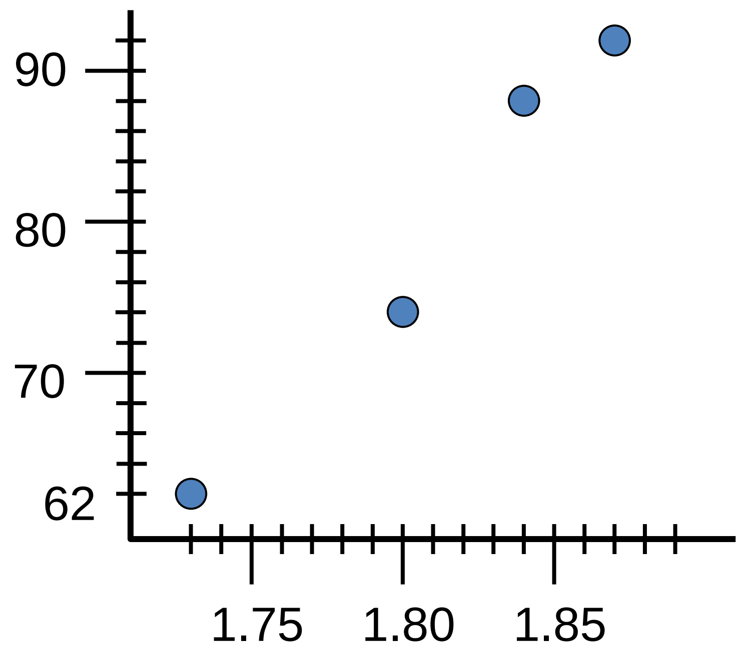
Covariance: example 2

- Suppose some measurements are $(X, Y) = (length, weight)$:
  $\{(1.80, 74), (1.87, 92), (1.84, 88), (1.73, 62)\}$
- $E(X) = 1.81\ m$
- $E(Y) = 79\ kg$
- $E(X \times Y) = 143.605$
- $Cov(X, Y) = E(X \times Y) - E(X) \times E(Y) = 143.605 - 1.81 \times 79 = 0.615\ kg.m$
- The covariance is larger so the variables X and Y vary more together than in example 1

Covariance: examples 1 and 2



$Cov = 0.475$



$Cov = 0.615$

Correlation

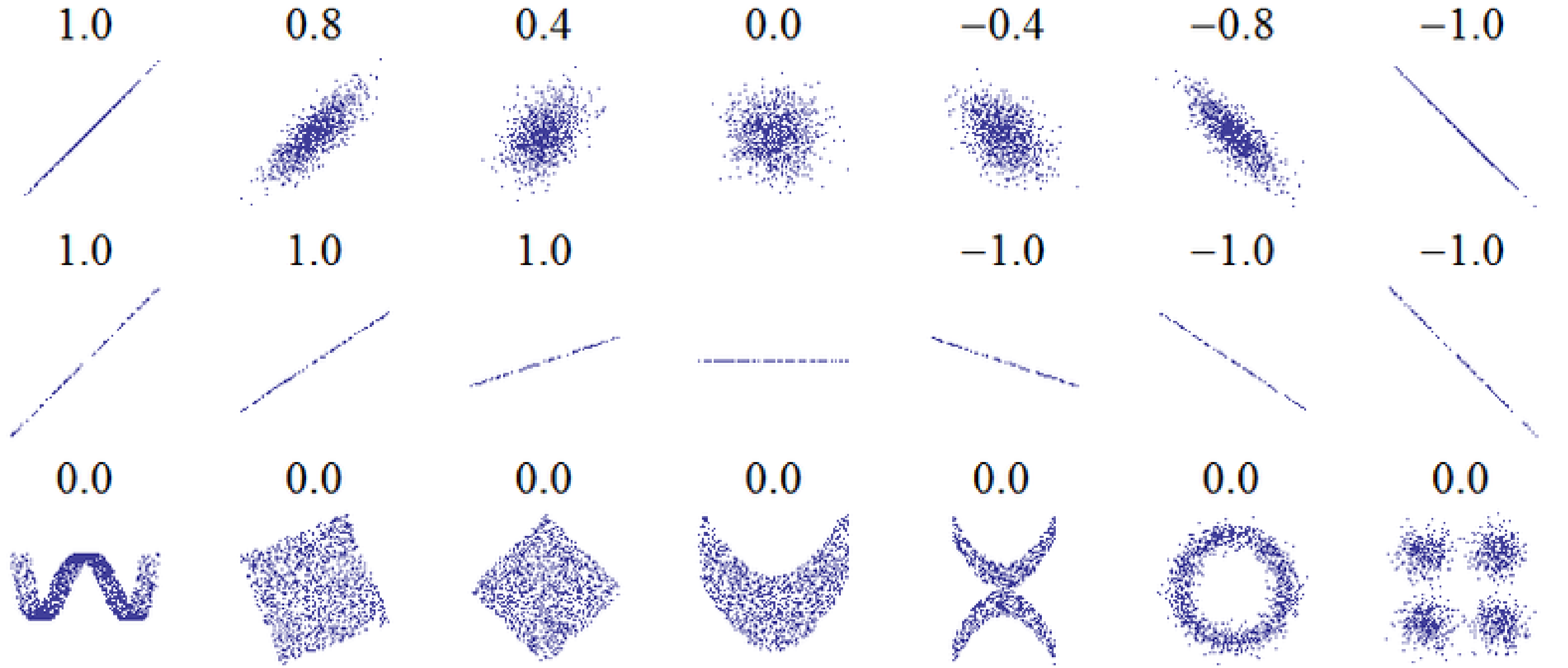
- The correlation is a measure for the degree in which two variables X and Y depend on each other
- Most common measure is the Pearson correlation coefficient

$$corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \times \sigma_Y}$$

- Is always between -1 and +1
- Is dimensionless (unlike covariance)

Correlation

- Pearson's correlation coefficients



Estimator for $Cov(X, Y)$

- As seen previously $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_S)^2$ is an unbiased estimator for variance from a sample

- An unbiased estimator for covariance based on a sample is

$$\widehat{\sigma_{xy}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

because $E\left(\widehat{\sigma_{xy}}\right) = \sigma_{xy}$

Covariance matrix

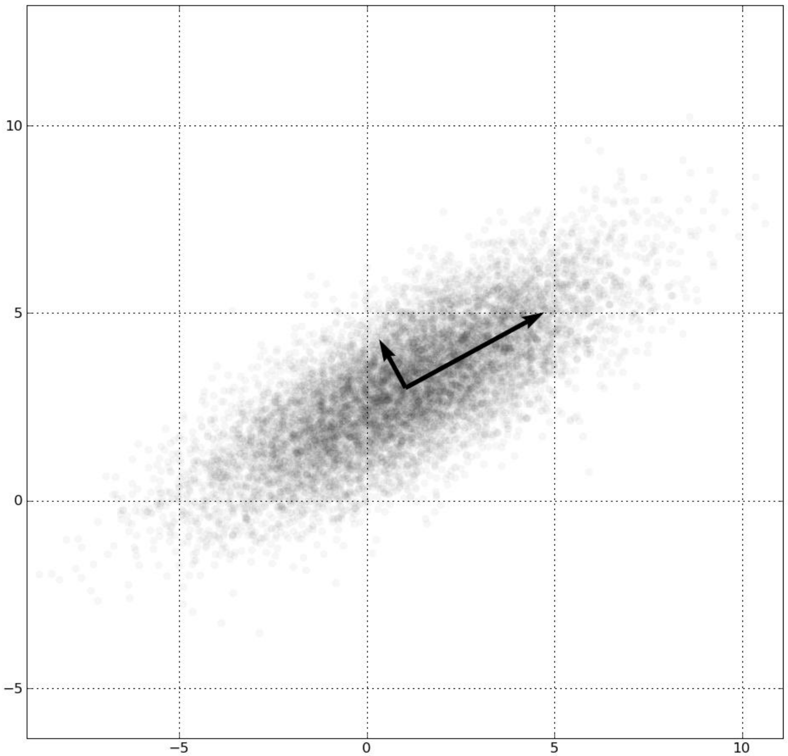
- The covariance of each pair of variables can be stored in a matrix
  - Diagonal terms: $E(x_i x_i) - E(x_i)E(x_i) = Var(x_i)$
  - Other terms: $E\left(x_i x_j\right) - E(x_i)E\left(x_j\right) = Covar\left(x_i, x_j\right)$

$$\begin{bmatrix} Var(x_1) & Cov(x_1, x_2) & \dots & Cov(x_1, x_d) \\ Cov(x_1, x_2) & Var(x_2) & \cdots & Cov(x_2, x_d) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_1, x_d) & Cov(x_2, x_d) & \cdots & Var(x_d) \end{bmatrix}$$

- The covariance matrix is symmetric

Covariance matrix

- Useful for analyzing relations between variables
- Example: Principal Component Analysis (PCA)
  - Uses covariance in combination with eigenvectors
  - Span an orthonormal basis of the covariance matrix where the covariance between new axes is minimal



Tests in statistics

- Null-hypothesis, denoted H_0 is the statement that assumes there is no relationship or effect
- With a test, the null-hypothesis may be rejected or not
- We need a pre-specified significance level for this
- A result is significant if is unlikely that it occurred by chance
- An alternative hypothesis is denoted H_1 and can only be accepted when H_0 can be rejected

Test example

- Null-hypothesis H_0: a coin is fair. Significance level required set at 0.05
- Possible outcome of 6 tosses to be all the same (either heads or tails) has a probability of $2/64 = 1/32 \approx 0.03$, assuming H_0
- Possible outcome of 6 tosses to be five vs. one *or more extreme* has a probability of $12/64 + 2/64 \approx 0.22$, assuming H_0
- In the first experiment, H_0 is rejected since $P(outcome|H_0) \approx 0.03 < 0.05$, so the coin is biased
- In the second experiment we do not reject H_0

The *t*-test

- Founder is William Sealy Gosset
- Worked at the Guinness brewery to control quality of beer
- Wrote under the pseudonym “Student”
- Mostly worked during tea (t) time
- Hence known as the Student’s *t*-test
- Goal of the t-test: test the validity of a null hypothesis

The *t*-test

Commonly performed *t*-tests:

- Compare the mean of a data set to a constant value and check whether the difference is significant
  - one-sample location test
- Compare the means of two data sets and check whether the difference is significant
  - two-sample location test

The *t*-test

- Examples
  - Calculate whether the average weight of a package of pasta really is 500 gr. or smaller (one-sample location test)
  - Calculate whether a weight reduction treatment is successful by comparing means before and after treatment (two-sample location test)
  - Calculate whether a novel algorithm produces significant better results than its prior version or its competitors (two-sample location test with golden standard data)

1-sided vs. 2-sided tests

- A 1-sided test is used when you know beforehand that, if there is an effect of your treatment, one sample mean should definitely be greater / smaller than the other

- A 2-sided test is used when you don't know beforehand which way the effect should go, if your treatment has an effect at all

- We look at 1-sided tests only in EMGMT

The *t*-test

Conditions:

- Population(s) should follow a normal distribution
- In case of a single population, the population variance can be unknown; in that case, the (unbiased) estimator for the variance is used:

$$\widehat{\sigma_x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_S)^2$$

One-sample *t*-test

- Test whether the mean equals a constant value μ_0, variance unknown, $H_0\colon \mu = \mu_0$
- Against hypothesis: $H_1\colon \mu < \mu_0$ or $H_1\colon \mu > \mu_0$
- The statistics then is $T = \frac{\bar{X} - \mu_0}{\hat{\sigma}}\sqrt{n}$
- We compare this T value against a value from the table using the degrees of freedom ($df = sample\ size - 1$) and the significance level
- We reject H_0 if the probability to get T is smaller than the significance level

One-sample *t*-test

- “Sunshine” DVD players should last on average 5 years. A test on 20 DVD players reveals they lasted on average 4.9 years with $\hat{\sigma} = 1.5$ years.

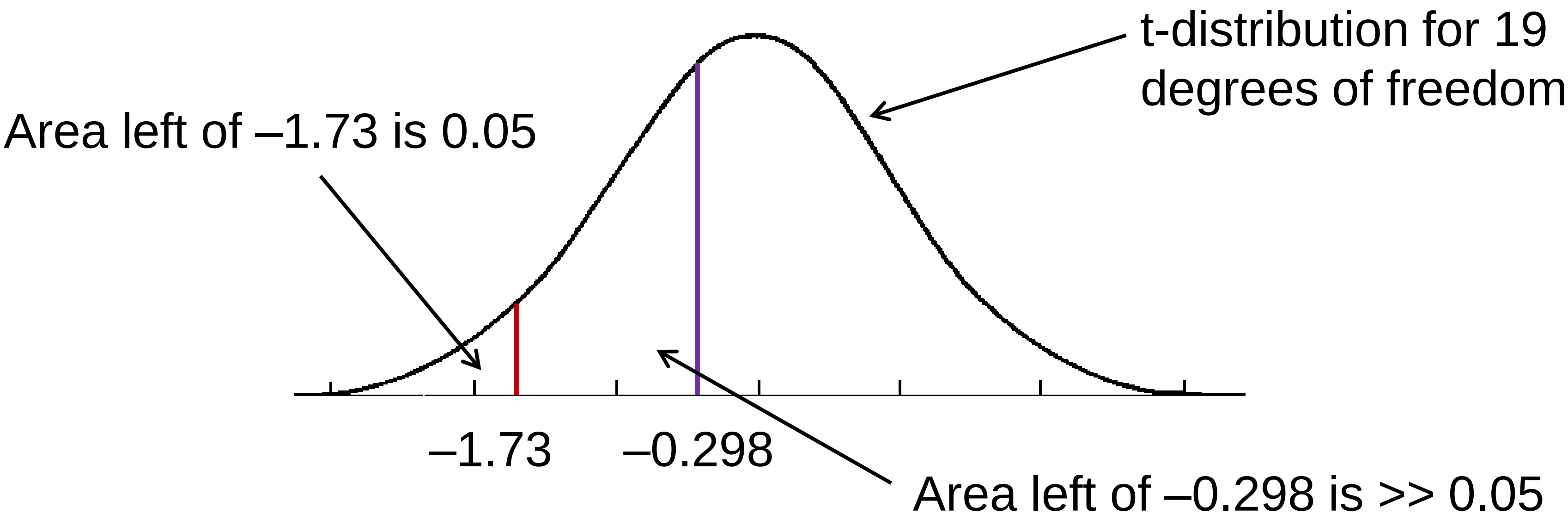
  Test if the actual average life is significantly smaller (with significance level set at 0.05)

One-sample *t*-test

- $H_0: \mu = \mu_0$ and $H_1: \mu < \mu_0$
  We test H_1 against H_0, we reject H_0 for too small T
- Our value of $T = (4.9 - 5)/1.5 \times \sqrt{20} = -0.298$
- Table look-up $df = 19$ and significance is 0.05 gives us $T(0.05,19) = 1.73$
- Meaning that the area of the tail of the *t*-distribution with $19\ df$ is 0.05 in the interval $[1.73, \infty)$ (recall that the area is a probability)
- Since it is symmetric, this also holds for $(-\infty, -1.73]$

One-sample *t*-test

- Since $P\big(x \in (-\infty, -1.73]\big) = 0.05$, we can observe that $P\big(x \in (-\infty, -0.298]\big) \gg 0.05$, so the outcome of the test ($\mu = 4.9$ and $\sigma = 1.5$ or more extreme) is not so unlikely that one thinks it happens less than 5% of the times



One-sample *t*-test

- $H_0: \mu = \mu_0$ and $H_1: \mu < \mu_0$
  We test H_1 against H_0, we reject H_0 for too small T
- $T = -0.298$ and $T(0.05,19) = 1.73$
- Note $P(T(19) \geq 1.73) = P(T(19) \leq -1.73) = 0.05$
- Since T is not smaller than -1.73, we cannot reject the null hypothesis, therefore we cannot prove H_1

One-sample *t*-test

- “Sunshine” DVD players should last on average 5 years. A test on 20 DVD players reveals they lasted on average 4.6 years with $\hat{\sigma} = 1$ years.

Test if the actual average life is significantly smaller (with significance level set at 0.05)

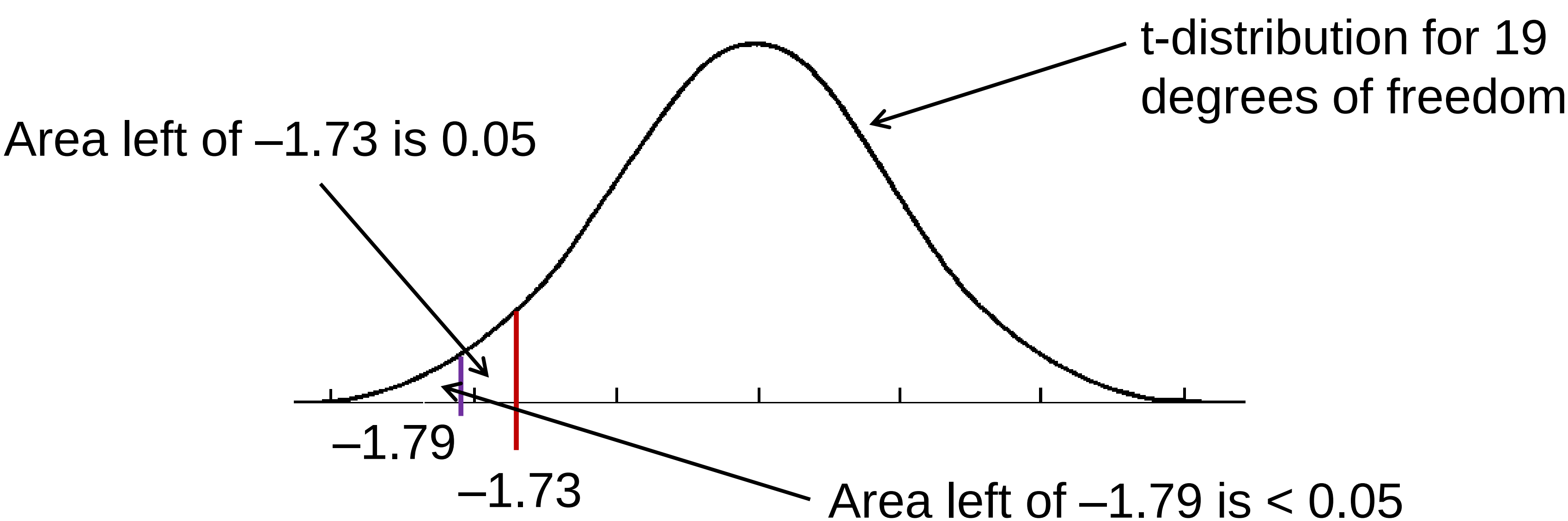
One-sample *t*-test

- Our value of $T = (4.6 - 5)/1 \times \sqrt{20} = -1.79$
- Table look-up $df = 19$ and significance is 0.05 gives us $T(0.05{,}19) = 1.73$ and we still have $P(T(19) \geq 1.73) = P(T(19) \leq -1.73) = 0.05$

- But now since T is smaller than -1.73, we can reject the null hypothesis, and accept H_1

One-sample *t*-test

- Since $P\big(x \in (-\infty, -1.73]\big) = 0.05$, we can observe that $P\big(x \in (-\infty, -1.79]\big) < 0.05$, so the outcome of the test ($\mu = 4.6$ and $\sigma = 1$ or more extreme, given the null hypothesis) is more unlikely than 5% of the times



Two-sample (unpaired) *t*-test

- Check whether the mean between two sample sets (X and Y) of size n and m is equal
- The statistics is

$$T = \frac{\bar{X} - \bar{Y}}{S_{XY}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

- Where S_{XY} is the unbiased weighted standard deviation

$$S_{XY} = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}$$

- Degrees of freedom: $n + m - 2$

Two-sample (unpaired) *t*-test

- Example: To check whether a new engine really uses less gas (with significance 5%), we determine how many liters are needed to perform a distance of 100 km (two groups of 10 cars; $n = m = 10$)
  - New engine (X): $mean = 5.2$, $S_X = \sigma_X = 0.8$
  - Old engine (Y): $mean = 5.5$, $S_Y = \sigma_Y = 0.5$

Two-sample (unpaired) *t*-test

- $Z = mean(X) - mean(Y)$
- $H_0 : Z = 0,\ H_1 : Z < 0$
- $S_{XY} = 0.667$
- $T = (5.2 - 5.5)/0.667 \times \sqrt{5} = -1.0056$
- $P(T(18) \geq 1.73) = P(T(18) \leq -1.73) = 0.05$
- We reject the null hypothesis for too small values of T
- Since $-1.0056 > -1.73$, the null hypothesis is not rejected

Two-sample (unpaired) *t*-test

- Example: To check whether a new engine really makes it use less gas (with significance 5%), we determine how many liters are needed for a distance of 100 km (two groups of 15 cars; $n = m = 15$)
  - New engine (X): $mean = 5.2$, $S_X = \sigma_X = 0.8$
  - Old engine (Y): $mean = 5.5$, $S_Y = \sigma_Y = 0.5$

Two-sample (unpaired) *t*-test

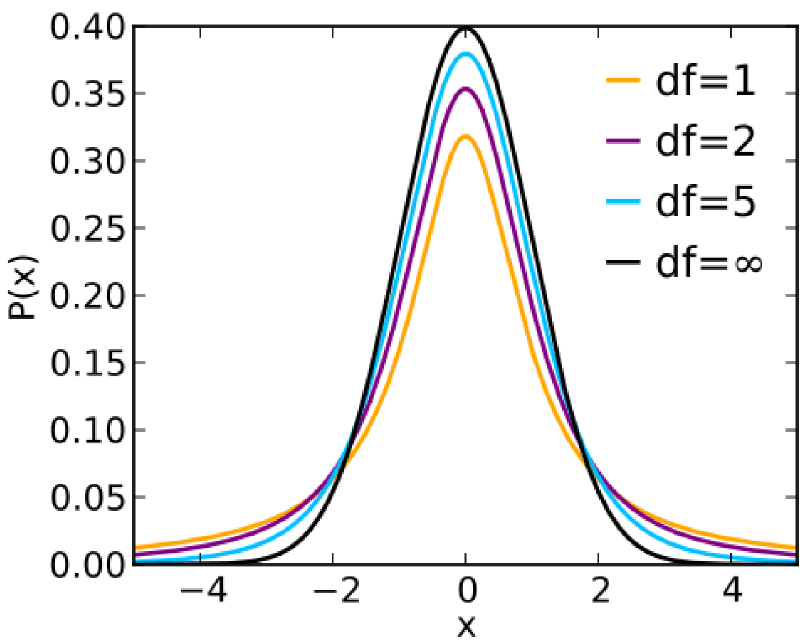
- $Z = mean(X) - mean(Y)$
- $H_0: Z = 0,\ H_1: Z < 0$
- $S_{XY} = 0.667$
- $T = (5.2 - 5.5)/0.667 \times \sqrt{7.5} = -1.232$
- $P(T(28) \geq 1.701)) = P(T(28) \leq -1.701) = 0.05$
- We reject the null hypothesis for too small values of T
- Since $-1.232 > -1.701$, the null hypothesis is not rejected

Two-sample (unpaired) *t*-test

- Note: the test statistic is such that
  - A larger difference in mean can cause H_0 to be rejected
  - A larger sample size for X and/or Y can cause H_0 to be rejected
  - A smaller standard deviation (estimate) for X and/or Y can make H_0 to be rejected

$$T = \frac{\bar{X} - \bar{Y}}{S_{XY}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$S_{XY} = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}$$



Two-sample (paired) *t*-test

- Suppose we have a set of paired samples (X_i, Y_i)
- The sample set is of size n
- We define $Z = X - Y$
- Our null hypothesis is H_0: $\mu_z = 0$
- Our test statistic is $T = \frac{\bar{Z}}{S_Z}\sqrt{n}$

  where S_Z is the unbiased estimator for the standard deviation of Z

Two-sample (paired) *t*-test

- We want to test if a diet is effective (5% significance), so we measure test subject's weights before and after the diet

Test subject	1	2	3	4	5	6	7	8	9	10
Weight before (X)	110	85	73	91	163	88	92	75	103	115
Weight after (Y)	99	83	75	86	141	79	96	70	91	102
Z = X - Y	11	2	-2	5	22	9	-4	5	12	13

- $H_0: Z = 0, H_1: Z > 0$
- We can calculate $\bar{Z} = 7.3$ and $S_Z = 7.75$

Two-sample (paired) *t*-test

- Then our value $T = 7.3/7.75 \times \sqrt{10} = 2.98$
- In the table ($df = 9, p = 0.05$), the critical value of T is 1.833, *i.e.* $P(T \geq 1.833|H_0) = 0.05$
- Since our value $T = 2.98$ is (much) larger, then $P(T \geq 2.98|H_0) < 0.05$, meaning that the probability of this outcome (or more extreme) given the null-hypothesis is less than 5% (the pre-defined significance level)
- Hence we reject the null-hypothesis, so yes, the diet is effective

Two-sample (paired) *t*-test

- Illustration