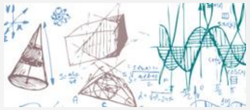


Introduction to probability, statistics and data handling

Tomasz Szumlak, Agnieszka Obłąkowska-Mucha
Faculty of Physics and Applied Computer Science

AGH UST Krakow

Type I and Type II error

- A Type I error means rejecting the null hypothesis when it's actually true. It means concluding that results are **statistically significant** when, in reality, they came about purely by chance or because of unrelated factors.This kind of error is called **Type I**.
- If we accept a hypothesis which is not true, we say we made **Type II** error.

You decide to get tested for pregnancy. There are two errors that could potentially occur:

Type I error (false positive): the test result says you are pregnant, but you actually don't.

Type II error (false negative): the test result says you are not pregnant, but you actually are.

Test Results	Reality	
	Pregnant (=1)	Not Pregnant (=0)
Positive (=1)	Number of True Positives (TP)	Number of False Positives (FP)
Negative (=0)	Number of False Negatives (FN)	Number of True Negatives (TN)

https://neeraj-kumar-vaid.medium.com/statistical-performance-measures-12bad66694b7

Type I and Type II error

- A Type I error means rejecting the null hypothesis when it's actually true. It means concluding that results are **statistically significant** when, in reality, they came about purely by chance or because of unrelated factors.This kind of error is called **Type I**.
- If we accepte a hypothesis which is not true, we say we made **Type II** error.

You decide to get tested for pregnancy. There are two errors that could potentially occur:

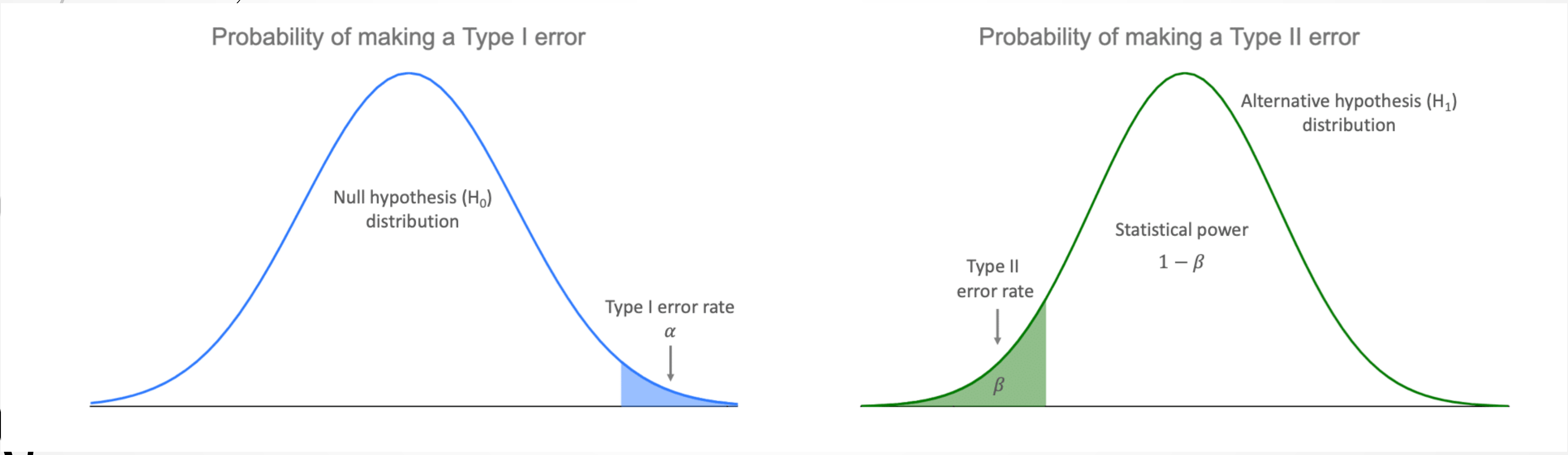
Type I error (false positive): the test result says you are pregnant, but you actually don't.

Type II error (false negative): the test result says you are not pregnant, but you actually are.

https://neeraj-kumar-vaid.medium.com/statistical-performance-measures-12bad66694b7

Type I and Type II error

- Imagine that we rejected a hypothesis and it happens to be true. This kind of error is called **Type I**.
- If we accepted a hypothesis which is not true, we say we made **Type II** error.



What can happen?

- No matter what we do, in real life we are never going to know for certain that we made the right choice or we in fact made one of the errors
- Remember Type I is to **REJECT** the null when it is true and Type II is **NOT TO REJECT** the null when it is false in reality
- Let's start from a generic (and easy) example, then we try to make a number of general statements and we get back to some more examples…

- Say, we are testing (for normal distribution) the following:

  H_0: $\mu = 14$ H_1: $\mu > 14$
- After applying our „procedure" we decided to reject the hypo at significance level $\alpha = 0.05$
- Either of the two happened: **the null is wrong** and we did good or the **null is true** and we made a Type I error

What can happen?

- After applying our „procedure" we decided to accept the hypo at significance level $\alpha = 0.05$
- Either of the two happened: **the null is ok** and we make a good decision or the **null is false** and we made a Type II error
- Is it a bit confusing…? Yes, and we can summarise this using the CONFUSION MATRIX – now you know why we use this name…

Ground truth

	H_0 **is false**	H_0 **is true**
Reject	Super	Type I error
Do not reject	Type II error	Super

Conclusions using data

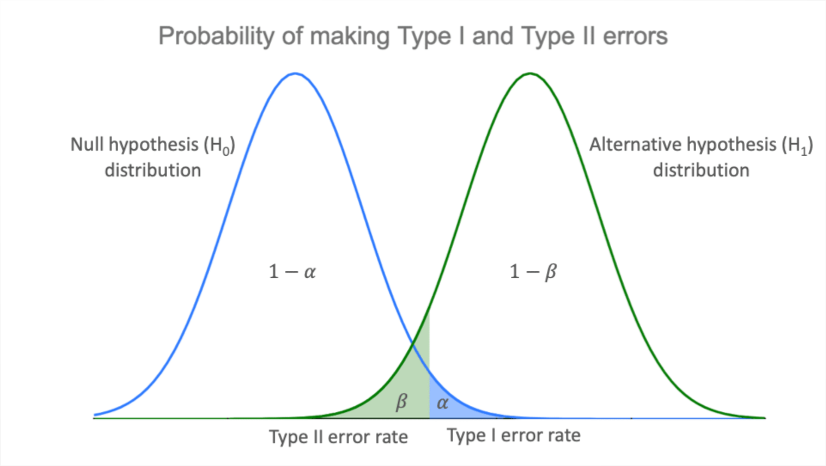
Note, using data we in fact make calculations, the ground truth remains, however, unknown…

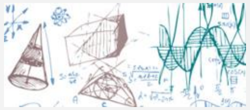
A trial example

- This is a classic example: criminal trial

$$H_0: \boldsymbol{not\ guilty}\ (\boldsymbol{no\ crime}),$$
$$H_1: \boldsymbol{guilty}(\boldsymbol{criminal})$$

- In a typical jurisdiction system we have a rule „not guilty until proven otherwise", so we need to have very strong evidences to convict
- Type I: rule guilty when in reality a person is innocent
- Type II: not guilty when a person committed a crime
- Definitely, we want to make sure, that we never convict an innocent person (well, in real life we do…)
- What we can do – we should decrease the possibility of making Type I error as much as possible





Significance and power

- The probability of **Type I error is called the significance level** of the statistical test (we know that already), we write this as

$$\boldsymbol{P\left(Type\ I\ error \middle| H_0\right) = \alpha}$$

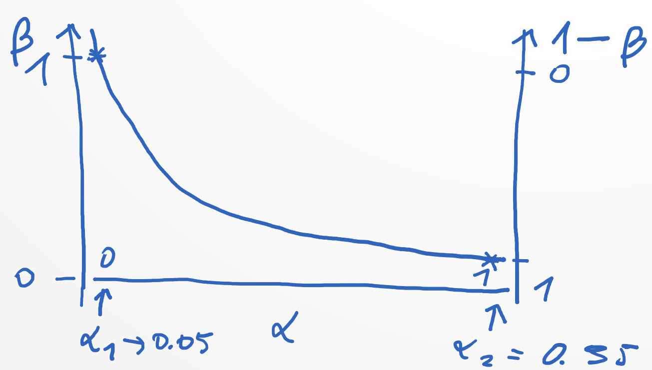
- Usually we start the test procedure by setting the value of α
- Now, we call the probability of making a Type II error as β
- There is a complication – it is much harder to define this error, it depend on the true value of the testing parameter, the size of the data sample and on the significance level itself!
- The **POWER** of a statistical test is **the probability of rejecting the null hypothesis when it is false**, we write

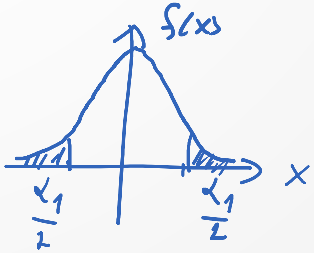
$$\boldsymbol{\mathcal{R} = 1 - P(Type\ II\ error) = 1 - \beta}$$

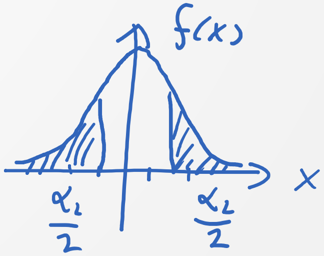
- The power depends on β – and we still do not know how to calculate it…

Intricate interplay of α and β

- There is some intuitive understanding about β.
- So, why not pick-up some very, very tiny value for the significance? The critical region would be very small, thus we would make almost impossible to reject the null hypo. We also say, that the efficiency of the selection is very high, but the probability of making the Type II error will also be very high (or we say, the purity of the selection is low)
- Conversely, for the large values of α the probability of Type II errors will go down
- Also, from this discussion you should start to understand, that the dependency is not trivial (i.e., it is not just some linear function or such)







Examples!

- Let's start calculating something!
- **Ex. 1** Say, we consider a sampling from a normal distribution with known variance $\sigma = 18$, the size of the data sample $n = 36$. Our hypos

$$H_0\colon \mu = 40, H_1\colon \mu < 40$$

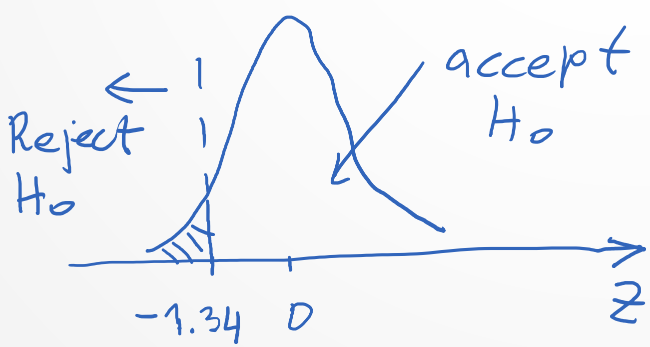
- We set the significance of the test: $\alpha = 0.09$ and our test statistics is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

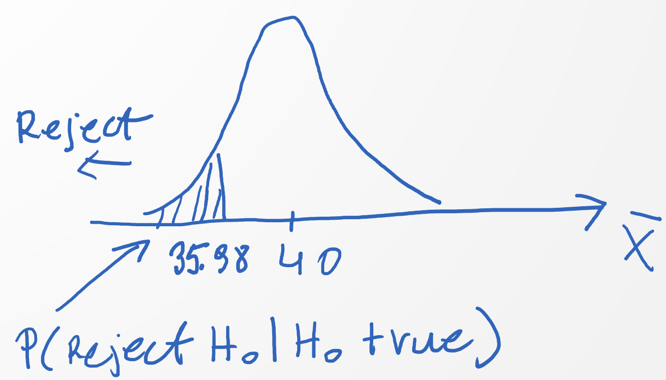
$$\bar{X} = \mu_0 + Z\frac{\sigma}{\sqrt{n}} = 40 + \frac{18}{\sqrt{36}}(-1.34)$$

$$\boldsymbol{\bar{X} = 35.98}$$

- The critical region



$\alpha = 0.09$



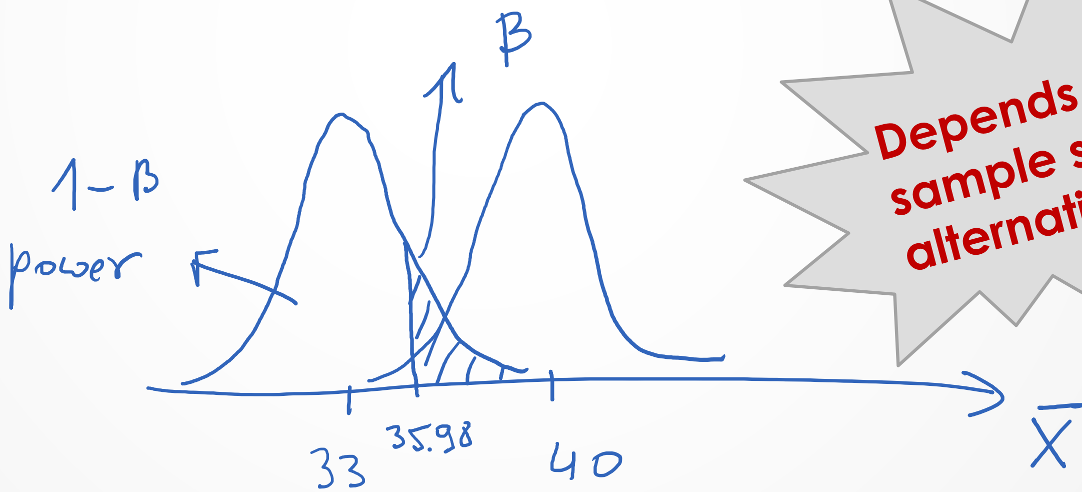
Examples!

- Now, what if the true mean is $\mu = 33$? What is the Type II error in this case?

  $P(Accept\ H_0 | \ \mu = 33)$

- We reject the H_0 hypo if the mean is less than 35.98, so the Type II error will occur with the probability

$$P(\bar{X} > 35.98 \mid H_1) = P(\bar{X} > 35.98 \mid \mu = 33)$$



Depends on the sample size and alternative!

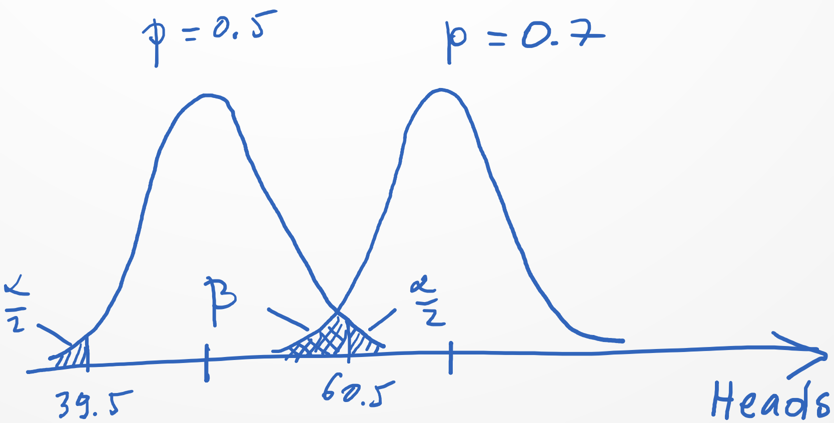
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{35.98 - 33}{18/6} = \frac{2.98}{3} = 0.99 \rightarrow P(Z > 0.99\ |H_1) = 0.16$$

β

Examples

- Note, that when we calculate power and probability of the Type II error we use the alternative hypothesis – we say the calculations are done under the alternative hypothesis

- **Ex. 2** Let's go back to the example with the coin (last lecture) we mentioned the possibility of accepting it if the true probability of heads is $p = 0.7$. This time we calculate the exact value of the probability of accepting the null hypo (i.e., the coin is fair)
- Our hypos are: $H_0 : p = 0.5, H_1 : p = 0.7$
- The numbers we worked out: for the sample of **100 tosses**, we accept the null if we get between **39.5 and 60.5 heads**. This gave us, in turn, the significance $\boldsymbol{\alpha = 0.0358}$.



Examples

- The right-hand side distribution represent 100 tosses with the probability of getting head set to 0.7
- Let's work out the probability of the Type II error

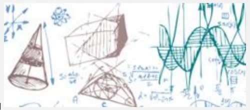
$$\mu_1 = np = 100 \cdot 0.7 = 70, \sigma_1 = \sqrt{npq} = 4.58$$

- Now, we need to translate the critical points of the null hypo into Z-score for the alternative hypo

$$z_{+/-}^{(1)} = \frac{h_{+/-} - \mu_1}{\sigma_1} \to z_{+}^{(1)} = \frac{60.5 - 70}{4.58} = -2.07, z_{-}^{(1)} = \cdots = -6.66$$

$$\beta = P(-6.66 \le z_1 \le -2.07 | H_1) = 0.019$$

- There are slim chances, we accept the null (the coin is fair) when it is loaded with the probability of getting heads 0.7
- There is more to this example… When discussing the $C.I.$ we just picked two values for heads and checked the probability – from this we were able to get the significance level – **we say the decision rule was given**



Example-CI for Poisson DF

- **Mining disasters** (more than 10 victims) registered in a country in Europe were studied between 15 Mar 1851 and 22 Mar 1962
- Total of 191 accidents occurred – that is a lot…
- The first accident after the 15 Mar 1851 occurred 157 days after. Say, we stopped the study after the second accident. We would then obtain a single observation on a R.V. X. For this discussion we assume that it follows the exponential distribution
- **We have now an estimate on the mean time interval expected between accidents**
- Again – it would be really useful to be able to define an interval of confidence for this average time between disasters

Exponential distribution

- Let's evaluate the C.I. for the mining accidents example
- We assumed that the R.V. T follows the exponential model, we have a single observation of $t = 157$ days
- We ask for $C.L. = 100(1-\alpha)\% = 90\%$, $\alpha = \frac{1}{2}\alpha = 0.05$

$$P(T \le t) = 1 - e^{-\frac{t}{\mu}} = 0.05 \rightarrow \frac{t}{\mu} = -\ln(0.95) \rightarrow \mu_+ = 3060\ (days)$$

$$P(T \ge t) = e^{-\frac{t}{\mu}} = 0.05 \rightarrow \frac{t}{\mu} = -\ln(0.05) \rightarrow \mu_- = 52.4\ (days)$$

- As another summary we should stress, that evaluation of the **C.I. requires: data sample, a model** (to evaluate probabilities) and the parameter we want to evaluate

Remainder

- If σ is not known and is determined based on data:
  - for big samples $n > 30$ we may estimate σ with afair accuracy by its unbiased estimator:

$$\sigma \approx S^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

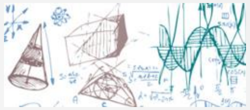
so the two-sided (symetric) interval is:

$$\left(\bar{x} - z(1-\frac{\alpha}{2})\frac{S^*}{\sqrt{n}},\ \bar{x} + z(1-\frac{\alpha}{2})\frac{S^*}{\sqrt{n}}\right)$$

  - if the sample is not too numerous we introduce t RV:

$$t = \frac{\bar{X} - \mu}{S}\sqrt{n-1} = \frac{\bar{X} - \mu}{S^*}\sqrt{n} \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\bar{X} - t(1-\frac{1}{2}\alpha, n-1)\frac{S}{\sqrt{n-1}} < \mu < \bar{X} + t(1-\frac{1}{2}\alpha, n-1)\frac{S}{\sqrt{n-1}}$$



Examples – C.I.

- ***Ex. 4*** A sample of $\mathrm{N} = 10$ measurements of the diameter of a ball bearing gave a mean $\bar{d} = 10.731\ cm$ and a standard deviation $s = 0.147\ cm$. Find (a) 95% and (b) 99% c.i. for the actual diameter.

a) In this case the confidence interval is given by: $C.I. = \bar{X} \pm t_{0.975} \frac{S}{\sqrt{N-1}}$.
And since the sample size $N = 10$, the number of degrees of freedom is: $\upsilon = 10 - 1 = 9$:

$$C.I._{0.975}^{(d)} = 10.731 \pm 2.26 \frac{0.147}{\sqrt{9}} = 10.731 \pm 0.11\ cm$$

b) Now, the 99% confidence interval: $t_{c(0.995)} = 3.25$ for $\upsilon = 10 - 1 = 9$:

$$C.I._{0.995}^{(d)} = 10.731 \pm 3.25 \frac{0.147}{\sqrt{9}} = 10.731 \pm 0.159\ cm$$

Note the precision of the measurement – we use a very accurate device (but that makes perfect sense – ball bearing may be a critical component of an aircraft engine).

Sample variance

- If $\{X_1, X_2, \cdots X_n\}$ denote R.Vs for a random sample of size n, the R.V. giving the variance of the sample (the sample variance) is defined as:

$$S^2 = \frac{1}{n}[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2]$$

- We already know, that $E[\bar{X}] = \mu$, is this the same for $E[S^2] = \sigma^2$?
  - A little digression – whenever the expected value of a statistics **is equal** to the corresponding **population parameter**, we call this statistics **an unbiased estimator**. Its value is then an unbiased estimate of the respective parameter
- Unfortunately, it can be proved that for the sample variance, we have some bias:

$$E[S^2] = \mu_{S^2} = \frac{n-1}{n}\sigma^2$$

- However, an unbiased variance estimator $\hat{\sigma}^2$, based on data, is easy to find:

$$\hat{S}^2 \equiv \hat{\sigma}^2 = \frac{n}{n-1}S^2 = \frac{\mathbf{1}}{\boldsymbol{n-1}}[(\boldsymbol{X_1} - \bar{\boldsymbol{X}})^{\mathbf{2}} + (\boldsymbol{X_2} - \bar{\boldsymbol{X}})^{\mathbf{2}} + \cdots + (\boldsymbol{X_n} - \bar{\boldsymbol{X}})^{\mathbf{2}}]$$

for large samples the diffrenece between S^2 and $\hat{S}^2$ is small.

S^2 is unbiased for normal variance,

Sampling dist. of variances (lab)

- With such unbiased estimator, we have:

$$E[\hat{S}^2] = \sigma^2$$

- In order to create the sampling distribution of variances we do exactly the same as for the SDoM, we take all the possible samples of size n, that can be drawn from a population and calculate their variances
- One change is, that instead of looking directly at the distribution of the sample variance, we look at the R.V.:

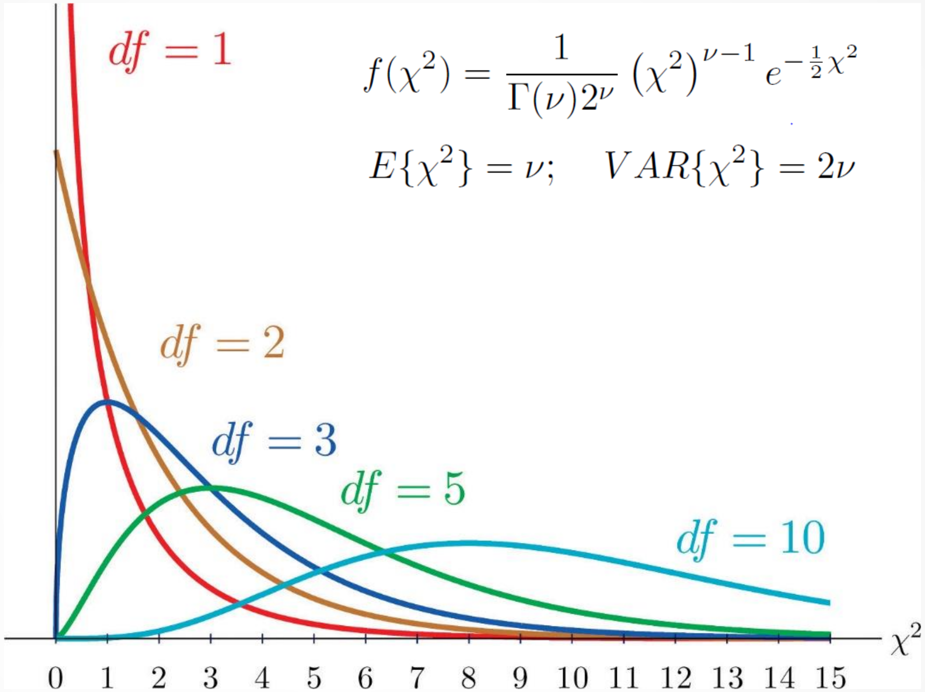
$$\chi^2 \equiv \frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{\sigma^2}$$

new RV: χ^2 = (coefficient) x sum of random variables)

what is the μ and σ of this new RV?

- **Theorem 6**. If a random samples of size n are taken from a population having a normal distribution, than the sampling variable $\frac{nS^2}{\sigma^2}$ has a $\boldsymbol{\chi^2}$ **distribution** with $n-1$ degrees of freedom

χ^2 distribution (lab)



- This is another very popular distribution in Statistics!
- The mathematical formula describing it is quite complex, again we are going to use tabulated values when solving problems!

$C.I.$ for the variance

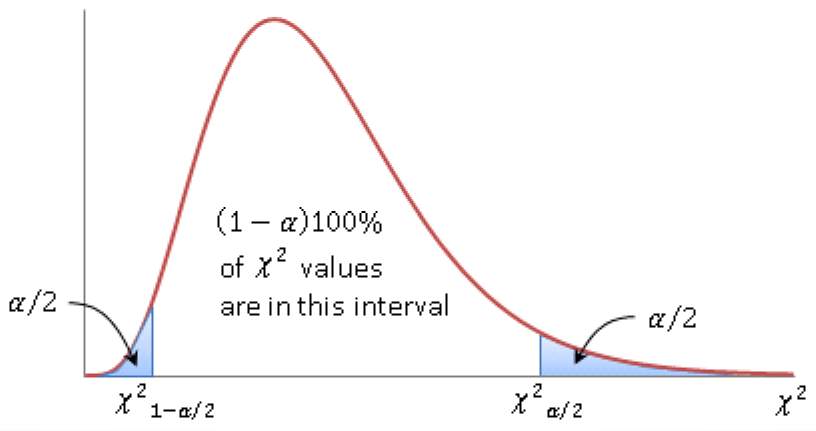
- Imagine that a company is delivering composite fibres for aircraft wings. In that case a great care should be taken to produce fibres that do not vary too much in tensile strength (expressed in kg)
- A sample of 10 fibres were taken and tested, the results were as follow $\bar{x} = 150.72\ kg$ and $s^2 = 37.75\ kg^2$. **Our mission is to find a confidence interval for the variance**
- We assume that the parent distribution of the fibre strength is normal, thus the sampling distribution of variance should follow the $\chi^2(\nu = n-1)$ distribution

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(\nu = n-1)$$

- The χ^2 is a family of curves and for increasing number of degrees of freedom it is getting more and more symmetric

$C.I.$ for the variance

- Again, the game is to find critical points using a given model (in this case the chi-squared)



$$P\left(\chi^2_{c-} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{c+}\right) = 1-\alpha \to \chi^2_{c-} = q_{\frac{1}{2}\alpha}, \chi^2_{c+} = q_{1-\frac{1}{2}\alpha}$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{c+}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{c-}}\right) = 1-\alpha$$

$$C.I.^{\chi^2(v)}_{100\cdot(1-\alpha)\%} = \left(\frac{(n-1)S^2}{\chi^2_{c+}}, \frac{(n-1)S^2}{\chi^2_{c-}}\right)$$

$C.I.$ for the variance

- Getting back to the fibre strength example, we are searching for $C.I._{90\%}^{\chi^2(9)}$, the critical points (from tables) $\chi^2_{5\%} = 3.325$ and $\chi^2_{95\%} = 16.919$ for $\chi^2(\nu = 9)$ distribution
- Our probability statement then is

$$P\left(3.325 \leq \frac{9s^2}{\sigma^2} \leq 16.919\right) = 0.9$$

$$C.I._{90\%}^{\chi^2(9)} = \left(\frac{(n-1)S^2}{\chi^2_{c+}}, \frac{(n-1)S^2}{\chi^2_{c-}}\right) = \left(\frac{9s^2}{16.919}, \frac{9s^2}{3.325}\right)$$

$$= (0.53s^2, 2.71s^2) = \cdots$$

- For the timer example, this would give us

$$C.I._{90\%}^{\chi^2(10)} = \left(\frac{10 \cdot 3.12}{18.307}, \frac{10 \cdot 3.12}{3.247}\right) = (1.70,9.38)$$

Examples – C.I.

$$\chi^2_{0.025} \le \frac{(n-1)\hat{S}^2}{\sigma^2} \le \chi^2_{0.975}$$

- **Ex. 7** The standard deviation of heights of 16 male students chosen at random in a school of 1000 male students yielded: $s = 5.9\ cm$. Find (a) 95% and (b) 99% c.i. of the standard deviation for all male students at this school. Assume that height is normally distributed.
  - The critical values in this case are: $S\frac{\sqrt{n}}{\chi_{0.975}}$ and $S\frac{\sqrt{n}}{\chi_{0.025}}$.
  - From a table for $\upsilon = 15$:

    $\chi^2_{0.975} = 27.5 \rightarrow \chi_{0.975} = 5.24$ and $\chi^2_{0.025} = 6.26 \rightarrow \chi_{0.975} = 2.50$

$$C.I._{0.975}^{(s)} = S\frac{\sqrt{n}}{\chi_{0.025}} < s < S\frac{\sqrt{n}}{\chi_{0.975}} = 4.9 < s < 9.4\ cm$$

  - The 99% c.i.:

$$C.I._{0.975}^{(s)} = S\frac{\sqrt{n}}{\chi_{0.005}} < s < S\frac{\sqrt{n}}{\chi_{0.995}} = 4.1 < s < 11.0\ cm$$

χ^2 for fitting

- χ^2 distribution should be always associated with a RV which describes the dispersion of the square of the deviations of a an RV around a fixed point.
- Now: what if the point is the „TRUE" expected value of X, i.e. μ_X?

It means that the variable:

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - E\{X\})^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \mu_X)^2}{\sigma^2}$$

has a χ^2 distribution with $\nu = n$ (!) degrees of freedom

- One can use this to determine wheather the data fit a particular distribution (goodness-of-fit test).

the test distribution is χ^2 distribution with a test statistic for goodness-of-fit test:

$$\chi^2 = \sum_k \frac{(O_k - E_k)^2}{E_k}$$

E_k - TRUE value (or expected), for instance from theory or expected based on other very precise measurements

Differences of means

- **Ex. 4** Two classes attended the same course and took an exam. The mean mark for the first class (40 students) was 74 points with a standard deviation of 8, the second (50 students) scored the mean of 78 points with the standard deviation of 7. Can we claim that one of the class is significantly better than other? We set the $\alpha = 0.05$

- H_0: $\mu_1 = \mu_2$ - no significant difference, just fluctuation
- H_1: $\mu_1 \neq \mu_2$ - the second class is better

Differences of means

- **Ex. 4** Two classes attended the same course and took an exam. The mean mark for the first class (40 students) was 74 points with a standard deviation of 8, the second (50 students) scored the mean of 78 points with the standard deviation of 7. Can we claim that one of the class is significantly better than other? We set the $\alpha = 0.05$

- H_0: $\mu_1 = \mu_2$ - no significant difference, just fluctuation
- H_1: $\mu_1 \neq \mu_2$ - the second class is better
- The test statitstics is:

Differences of means

- **Ex. 4** Two classes attended the same course and took an exam. The mean mark for the first class (40 students) was 74 points with a standard deviation of 8, the second (50 students) scored the mean of 78 points with the standard deviation of 7. Can we claim that one of the class is significantly better than other? We set the $\alpha = 0.05$

- H_0: $\mu_1 = \mu_2$ - no significant difference, just fluctuation
- H_1: $\mu_1 \neq \mu_2$ - the second class is better
- The test statititstics is: the difference of means

$$\bar{X}_1 \to N\left(\mu, \frac{\sigma_1}{\sqrt{n_1}}\right) \qquad \bar{X}_2 \to N\left(\mu, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

$$\bar{X}_1 - \bar{X}_2 \to N\big(0, \sigma(\bar{X}_1 - \bar{X}_2)\big) \qquad \sigma(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_{\bar{X}_1} - \mu_{\bar{X}_2})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}}$$

$$Z_{st} = \frac{RV - \mu}{\sigma}$$

Differences of means

- **Ex. 4** Two classes attended the same course and took an exam. The mean mark for the first class (40 students) was 74 points with a standard deviation of 8, the second (50 students) scored the mean of 78 points with the standard deviation of 7. Can we claim that one of the class is significantly better than other? We set the $\alpha = 0.05$

- H_0: $\mu_1 = \mu_2$ - no significant difference, just fluctuation
- H_1: $\mu_1 \neq \mu_2$ - the second class is better
- The test statitstics is the difference of means

$$\mu_{\bar{X}_1-\bar{X}_2} = 0, \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}} = \sqrt{\frac{64}{40} + \frac{49}{50}} = 1.606$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \left(\mu_{\bar{X}_1} - \mu_{\bar{X}_2}\right)}{\sigma_{\bar{X}_1-\bar{X}_2}} = \frac{74 - 78}{1.606} = -2.49$$

$$z_{1,2} = \pm 1.96 \rightarrow \cancel{H_0}$$ rejected @ 5%

we can use z-scores for normal distribution (big sample)

- p-value: $P(Z \leq -2.49) + P(Z \geq 2.49) = 0.0128$

Examples – C.I.

- ***Ex. 6*** A sample of 150 brand A light bulbs showed a mean life-time of 1400 hours and a standard dev. of 120 h. A sample of 200 brand B light bulbs showed the corresponding values of 1200 h and 80 h. Find (a) 95% and (b) 99% $C.I.$ for the difference of the mean life-times of the populations of brands A and B.

  a) The confidence interval for the difference in means can be written as:

$$\bar{L}_A - \bar{L}_B \pm z_c \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = 1400 - 1200 \pm 1.96 \sqrt{\frac{120^2}{150} + \frac{80^2}{100}} = 200 \pm 24.8\, h$$

We can be 95% confident that the difference in population means lies between 175 and 225 h

  b) And for the 99% $C.I.$:

Differences of means

- What would be a change if ***σ* is unknown?**

- The test statistics is the same: the difference of means

$$\bar{X}_1 - \bar{X}_2 \to N\big(0, \sigma(\bar{X}_1 - \bar{X}_2)\big)$$

  - but now $\sigma(\bar{X}_1 - \bar{X}_2)$ is estimated:

$$\sigma(\bar{X}_1 - \bar{X}_2) \to \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2 \frac{1}{n_1}$$

  - and test startistic is t-score:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \left(\mu_{\bar{X}_1} - \mu_{\bar{X}_2}\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\frac{s_1^2}{n_1} + \frac{1}{n_2 - 1}\frac{s_2^2}{n_2}}$$

- and we need to use t-Student distribution with ν degrees of freedom

Differences of means

- **How we can compare two independent Population Proportion?**
- There are a few assumptions that must be fulfilled:
  - samples should be random and independent
  - number of successes >5

This enables using the normal dist, approximation.
The pooled proportion RV:

$$P = \frac{x_A + x_B}{n_A + n_B}$$

the difference:

$$\Delta_P = P'_A - P'_B \qquad \Delta_P \to N(0, \sigma_\Delta)$$

$$\sigma_\Delta = \sqrt{P(1-P)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

and test statistics (z-score):

$$Z = \frac{(P'_A - P'_B) - (P_A - P_B)}{\sqrt{P(1-P)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

H_0: $P_A = P_B$, so $\Delta_P = 0$
H_a: $P_A \neq P_B$, so $\Delta_P \neq 0$

Differences of means: Population Proportion

- **Example:** Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve *A* cracked under 4,500 psi. Six out of a random sample of 100 of Valve *B* cracked under 4,500 psi. Test at a 5% level of significance.

$$P = \frac{x_A + x_B}{n_A + n_B} = \frac{15+6}{100+100} = 0.105 \qquad P'_A = \frac{15}{100} \qquad P'_B = \frac{6}{100}$$

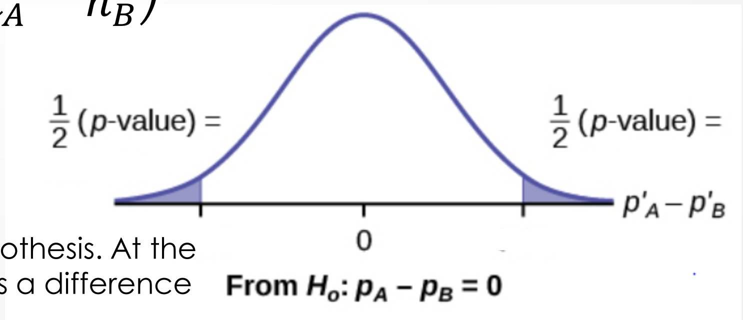
the difference: $\Delta_P = P'_A - P'_B = 0.21 \quad \Delta_P \to N(0, \sigma_\Delta)$

and test statistics (z-score):

$$Z = \frac{(P'_A - P'_B) - (P_A - P_B)}{\sqrt{P(1-P)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \qquad \sigma_\Delta = \sqrt{P(1-P)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

H_0: $P_A = P_B$, so $\Delta_P = 0$

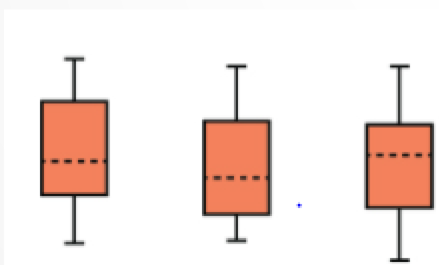
H_a: $P_A \neq P_B$, so $\Delta_P \neq 0$



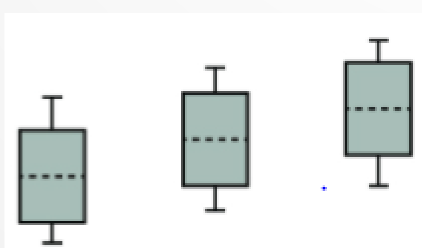
The *p*-value is 0.0379, so we can reject the null hypothesis. At the 5% significance level, the data support that there is a difference in the pressure tolerances between the two valves.

C.I. for variance

- In case of comparing the mean values of two populations we introduced how the sampling distribution of differences (and sums…) of means can be obtained. And what about the distribution of **differences of variances** that can be denote as: $S_1^2 - S_2^2$?
- For hypothesis tests comparing averages between more than two groups: "Analysis of Variance" (ANOVA).
- The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different.



H_0: $\mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. Differences are due to random variation.



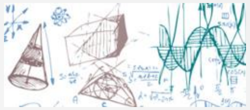
If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means.
Differences are too large to be due to random variation.

C.I. for variance ratios

- It turns out, that such distribution is **surprisingly** troublesome! Instead we can use **another statistic**: $\boldsymbol{F \propto S_1^2/S_2^2}$. We conclude that if the ratio is small/large the difference between variance is large, conversely if the ratio is close to unity the difference should be small, and we have a theorem…
- **Theorem 1**. Consider, we have **two random and independent** samples, of size n and m respectively, drawn from two normal populations with variances σ_1^2 and σ_2^2. It can be shown that the statistic

$$F = \frac{\dfrac{mS_1^2}{(m-1)\sigma_1^2}}{\dfrac{nS_2^2}{(n-1)\sigma_2^2}} = \frac{\dfrac{\hat{S}_1^2}{\sigma_1^2}}{\dfrac{\hat{S}_2^2}{\sigma_2^2}}$$

has the F distribution with $m-1, n-1$ degrees of freedom



C.I. for variance ratios

- Let's repeat our scheme again. We denote by $F_{0.01}$ and $F_{0.99}$ the values of F R.V. for which 1% of the area lies in each tail of the F distribution, then with 98% confidence we have the following

$$F_{0.01} \leq \frac{\frac{\hat{S}_1^2}{\sigma_1^2}}{\frac{\hat{S}_2^2}{\sigma_2^2}} \leq F_{0.99}$$

- So, a 98% C.I. for the variance ratio σ_1^2/σ_2^2 of two populations is given by:

$$\frac{1}{F_{0.99}}\frac{\hat{S}_1^2}{\hat{S}_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{F_{0.01}}\frac{\hat{S}_1^2}{\hat{S}_2^2}$$

- The respective critical values are tabulated. The way we read the tables are a bit difficult but we do that during the tutorial, so don't worry!

Examples – C.I.

- ***Ex. 8*** Two samples of sizes 16 and 10 are drawn from two normal populations. If their variances are found to be 24 and 18, respectively, find (a) 98% and (b) 90% confidence limits for the ratio of the variances
  - First we can find the modified sample variances:

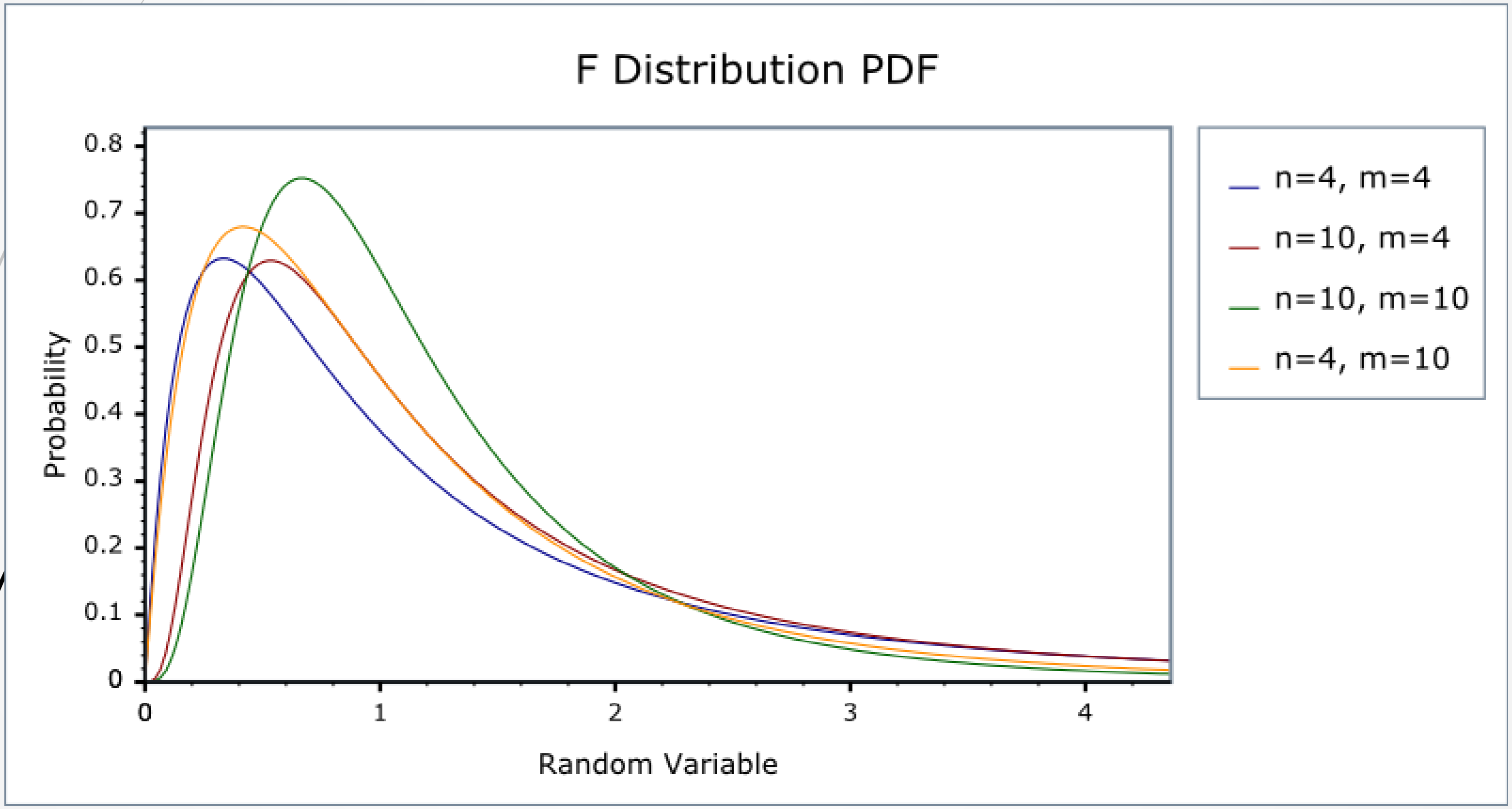
$$\hat{s}_1^2 = \frac{m}{m-1} s_1^2 = \frac{16}{15} \cdot 24 = 25.2$$

$$\hat{s}_2^2 = \frac{n}{n-1} s_2^2 = \frac{10}{9} \cdot 18 = 20$$

  - We need to use F distribution. For the 98% confidence interval we have: $F_{0.99} = 4.96$ for $\upsilon_1 = 16 - 1 = 15$ and $\upsilon_2 = 10 - 1 = 9$. And $F_{0.01} = 1/3.89$ for the same values of degrees of freedom. So, the c.i.:

$$C.I._{0.98}^{s^2} = \frac{1}{4.96} \frac{25.2}{20} < \frac{\sigma_1^2}{\sigma_2^2} < 3.89 \frac{25.2}{20} \qquad 0.28 < \frac{\sigma_1^2}{\sigma_2^2} < 4.90$$

Examples – C.I.


F Distribution PDF
Probability
Random Variable
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
n=4, m=4
n=10, m=4
n=10, m=10
n=4, m=10