Introduction to probability, statistics and data handling

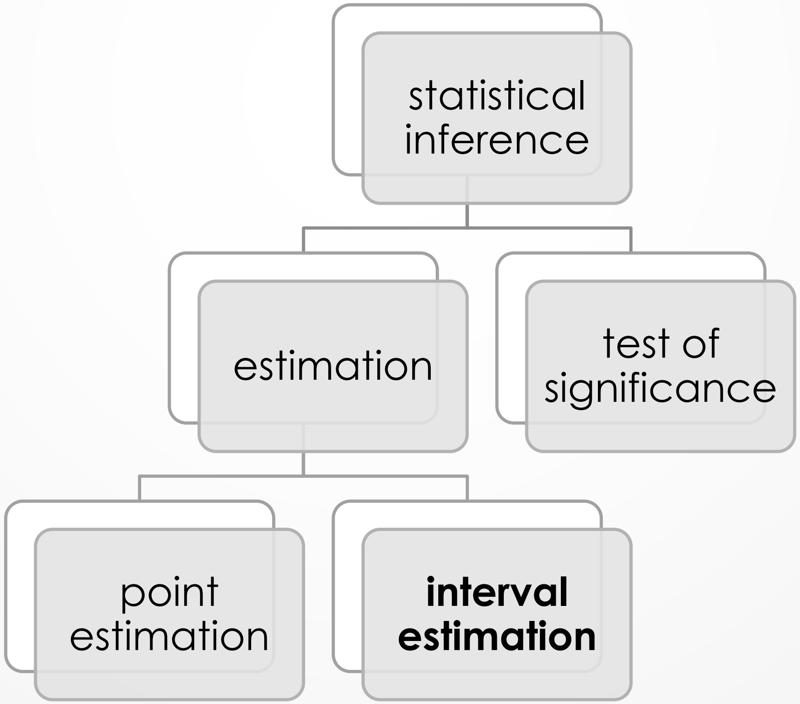
Tomasz Szumlak, Agnieszka Obłąkowska-Mucha

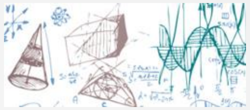
Faculty of Physics and Applied Computer Science

AGH UST Krakow

Statistical Inference

- The statistical inference consists in arriving at (quantitative) conclusions concerning a **population** where it is impossible or impractical to examine the entire set of observations that make up the population. Instead, we depend on a **subset** of observations - a **sample**.





Estimation with confidence

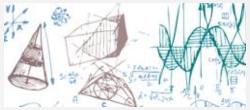
- We discussed a number of approaches to study problem of parameter estimation – we called it the **point estimation**, since we were interested only in „a value" of some parameter
- If there is a special name, it means that there must be something more… **And it is**!
- The topic of today's lecture will be the interval estimation or estimation with confidence
- We also make our first strides into hypothesis testing, for which defining the confidence if crucial

A quick one…

- **Consider the following**: we performed an experiment and got estimate on a parameter using one of the methods we learned
- Not bad… **Next we repeat the experiment and got another estimate – what should we expect**? What kind of result should be treated as „**plausible**" and what „**unlikely**"?
- It is obvious that the ability of obtaining a plausible range of values for any unknown population parameter is a powerful tool
- Remember – we are talking about a range defined in the space of model parameters (the model must be itself of course reasonable)
- The min and max limit of this plausible parameter(s) range we call **confidence limits** and the corresponding range **confidence interval**

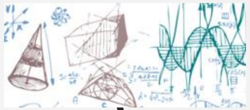
Example 1

- As usual we start discussing a bunch of experiments and discuss along the topic at hand
- **A psychology studies** were conducted to check correlation of the mental capabilities and proneness to injury among children. Total of 621 children were studied between ages of 4 and 11. The study period was divided into two intervals 4 – 7 and 8 – 11
- Say, one child experienced 3 accidents between ages 4 – 7. We could assume a Poisson model to describe the variability of the data sample. **With this single number we could still use the ML technique and obtain: $\hat{\mu} = 3$**
- We can than state that the accidents do happen but they are rather rare events
- However, it could be very useful to be able to provide some statement, with confidence 90%, that the mean lies in a range between this and that value (e.g., 1.0 – 5.0)



Example 2

- **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



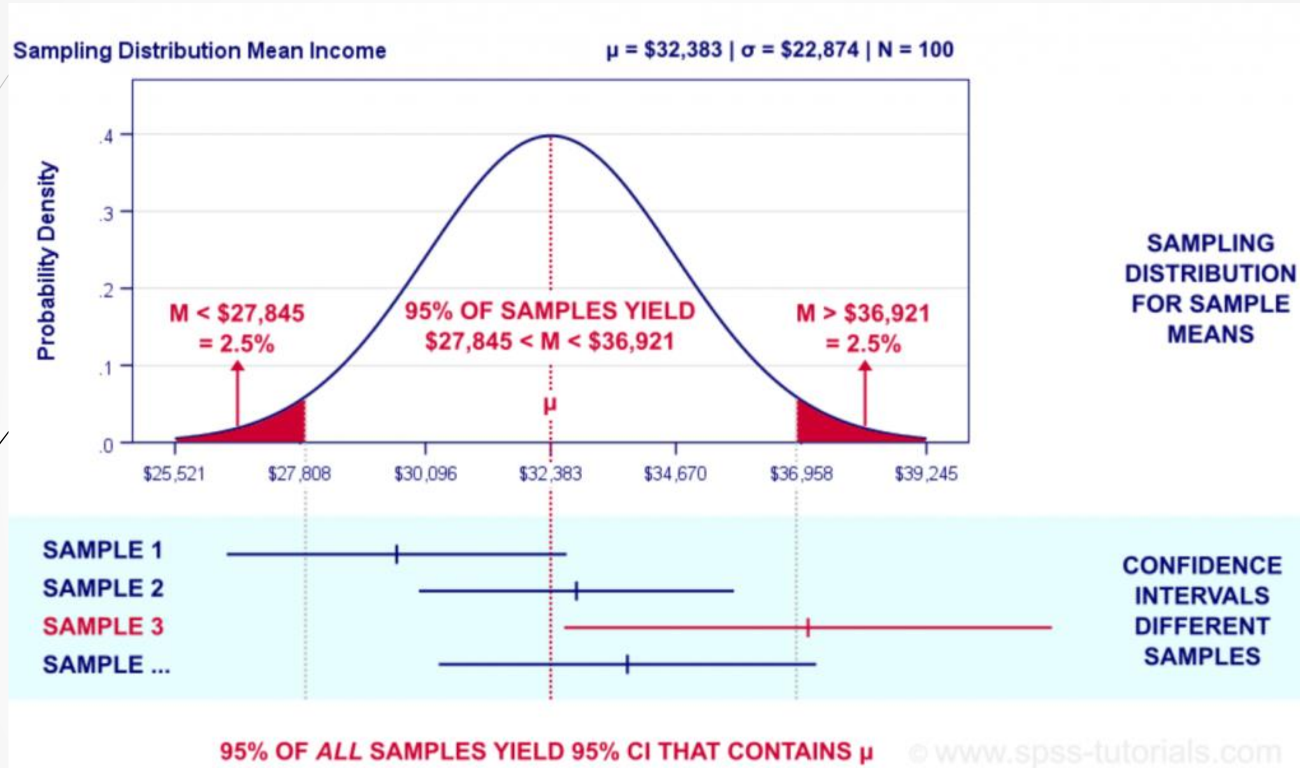
Example 3 (especially for you…)

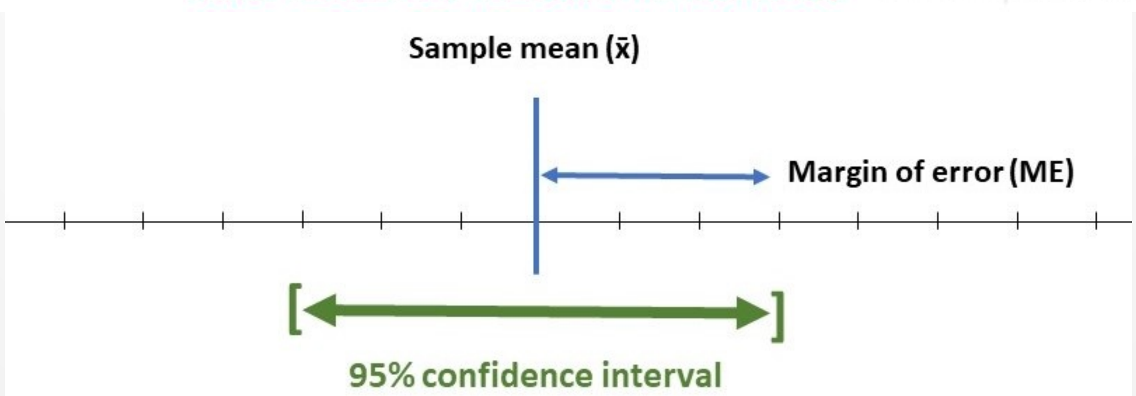
- **Lecture absences** of 113 student from a course A were noted over a period of two semesters (total of 24 lectures). In this case we could try to use a binomial model for the number of total missed lectures
- Say we divide the tested period into two parts – semester one and semester two. The corresponding number of lectures were 11 and 13
- The data showed that one student missed four (hmm… a lot) lectures, so the proportion of missed lectures would be $p = \frac{4}{11} = 0.364$
- One could question whether the binomial model is the best one to describe this data. It may so happen that some of the students are committed to the course…

Confidence

- Statistical statements regarding R.Vs. and probability should always be interpreted in terms of model parameters and confidence
- **We express the confidence using fractional numbers** (%). So, we could say, for instance, a $\kappa\%$ confidence interval for parameter θ (based on an actual observation) is the interval from θ_{-} to θ_{+}, where $\kappa\% \rightarrow 99\%, 95\%, 90\%, \ldots$
- Its meaning is as follow: if we observe an event with the prob. of 95% we say it is reasonable, on the other hand if this is just 5% it should be considered **unlikely**
- So, what left now is to evaluate the confidence interval, we reserve for example **5%** of probability for „strange" events and consider both cases too-low-strange and too-high-strange
- This is, so called, **two tailed** or two sided confidence interval and we have reserved **2.5%** probability for very high and very low results

Confidence intervals





Confidence intervals

- We got the impression that reporting the value of an estimator (e.g. $\hat{X}$) tells us nothing about the magnitude of the discrepancy that may exist between the estimator and the estimated parameter ($E[X] = \mu$).
- What would be the "confidence interval" of the estimated PARAMETER?
- WE MAY DEFINE the confidence interval in the following manner:

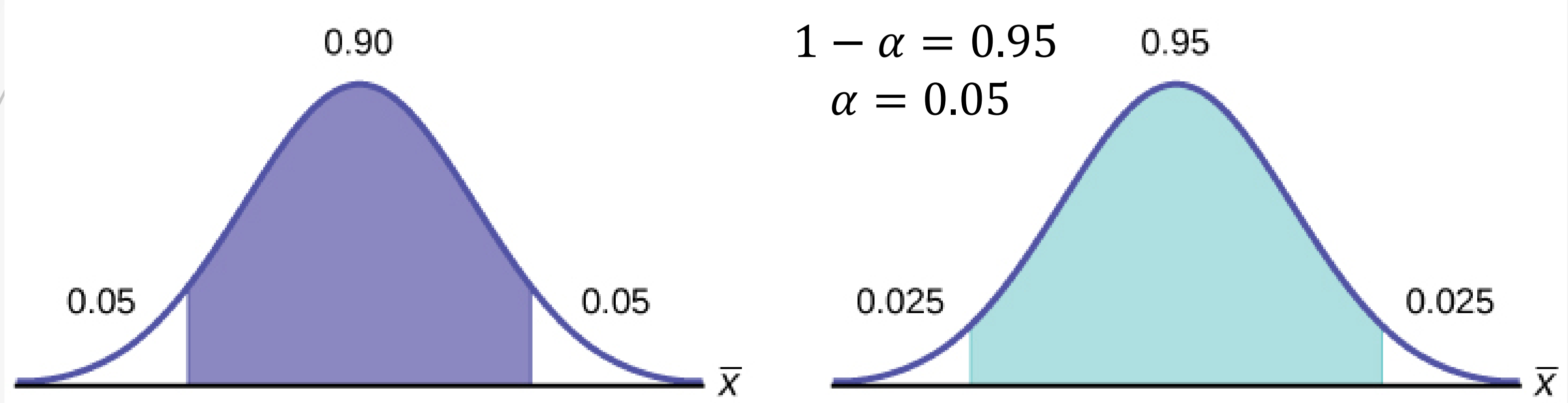
  1. we start with choosing a value $\mathbf{1-\alpha}$ of the **confidence level** as:

  $$1-\alpha, \qquad 0<\alpha<1$$

  1. usually $\alpha = 0.01;\ \mathbf{0.05};\ 0.1$
  2. the confidence interval Δ is chosen in such a way that the probability for Δ to cover the unknown parameter (like μ or σ^2) is $1-\alpha$

$C.I.$ for the normal distribution

- We already know a lot about evaluating probabilities using the normal distribution



Confidence Level	Alpha	Alpha/2	z alpha/2
90%	10%	5.0%	1.645
95%	5%	2.5%	1.96
98%	2%	1.0%	2.326
99%	1%	0.5%	2.576

$C.I.$ for the normal distribution

- Using the plot or the table from the previous slide we write for the critical values $z_c = \pm 1.96$, which corresponds to the confidence level of 95%:

$$P\left(-1.96 \leq \frac{\bar{X} - \boldsymbol{\mu}}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

- As usual, there are some tricks… For instance if we knew the distribution variance $\boldsymbol{\sigma}$ (remember the normal model has two parameters!) we could immediately solve these inequalities

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \boldsymbol{\mu} \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

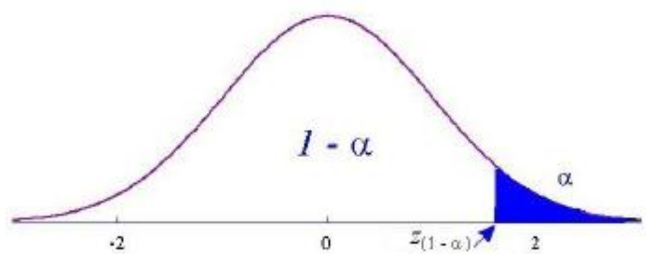
- This is a random interval, defined around the sample mean, which contains the unknown population mean with the probability of 95%. So, the 95% C.I. for μ is given by

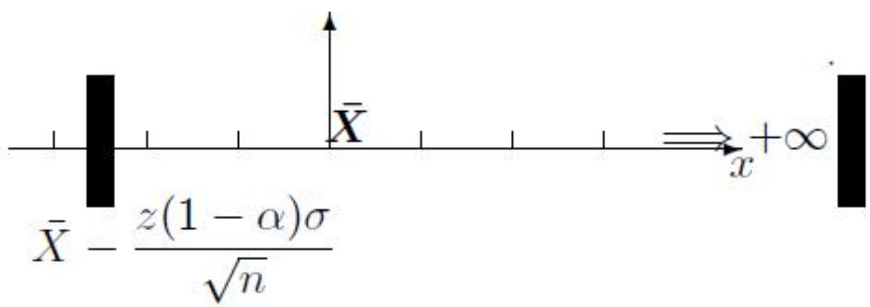
$$C.I._{95\%}^{\mathcal{N}} = \left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

One-sided CI

1. LOWER one-sided confidence interval: $\alpha_1 = 0 \quad z(\alpha_1) = -\infty$
$z(\alpha_2) = z(1-\alpha)$; the interval is:

$$\left(\bar{X} - z(1-\alpha)\frac{\sigma}{\sqrt{n}},\ +\infty\right)$$



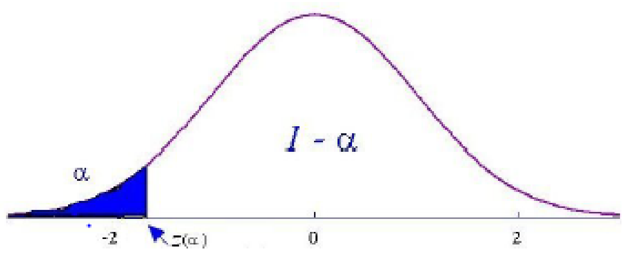


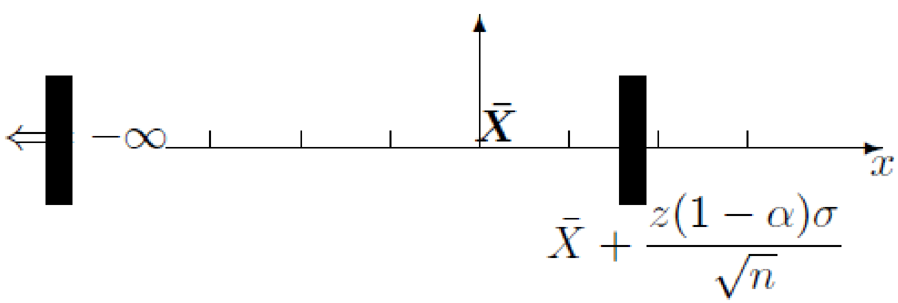
we may be $1-\alpha$ certain that μ is **no less** than $\bar{X} - \dfrac{z(1-\alpha)\sigma}{\sqrt{n}}$

One-sided CI

2. UPPER one-sided confidence interval $\alpha_2 = 0 \quad z(1-\alpha_2) = \infty$
the interval is:

$$\left(-\infty,\ \bar{X} - z(\alpha)\frac{\sigma}{\sqrt{n}}\right) \equiv \left(-\infty,\ \bar{X} + z(1-\alpha)\frac{\sigma}{\sqrt{n}}\right)$$





we may be $1-\alpha$ certain that μ is **not greater** than $\bar{X} + \dfrac{z(1-\alpha)\sigma}{\sqrt{n}}$

Interpretation of $C.I.$

- Observe, we are able to define a $C.I.$ using the formula describing the model and „probability points" that follow from definition of the confidence level
- If we obtain a single measurement then we will get an $C.I.$ spanning $0.27X$ to $39.5X$
- The proper interpretation is, that this interval **contains** the unknown parameter with probability **0.95**
- In other words: if we repeat an experiment 100 times, and calculate each time the $C.I.$ (random interval) then we should expect that about 95 times the unknown parameter will be inside these intervals
- The parameter is a number and the confidence statement is made based on properties of the random interval – it may or may not contain the parameter!

$C.I.$ for the normal distribution

- Imagine that we want to test the accuracy of some timer device using a more accurate one (like an digital stop-watch)
- It could go, for instance, like that – we set the tested timer to 5 minutes and we measure the actual time interval
- Assume that the observed data variation is a consequence of the scale precision (you may not be able to set the actual time) and the precision of the time mechanism - $\mathcal{N}(\mu, \sigma^2)$
- Say we made n observations and obtained sample mean and sample variance $\bar{x} = 294.8$, $s^2 = 3.12$ respectively
- We know, that if one draws a sample from a normal distribution the sampling distribution of $\bar{X}$ is also normal

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow Z = \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) \sim \mathcal{N}(0,1)$$

- And in general we can write:

$$P\left(-z_c \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_c\right) = 1 - \alpha$$

$C.I.$ for the normal distribution

- A general formula that can be applied for the normal distribution for its mean is then

$$C.I._{100\cdot(1-\alpha)\%}^{\mathcal{N}} = \left(\bar{X} - z_c \frac{\sigma}{\sqrt{n}}, \bar{X} + z_c \frac{\sigma}{\sqrt{n}}\right)$$

- Nice, but… what **if we do not know the distribution** variance (and we usually do not)? The most sensible approach would be to use the sample variance to estimate σ^2

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2 \rightarrow E[S^2] = \sigma^2$$

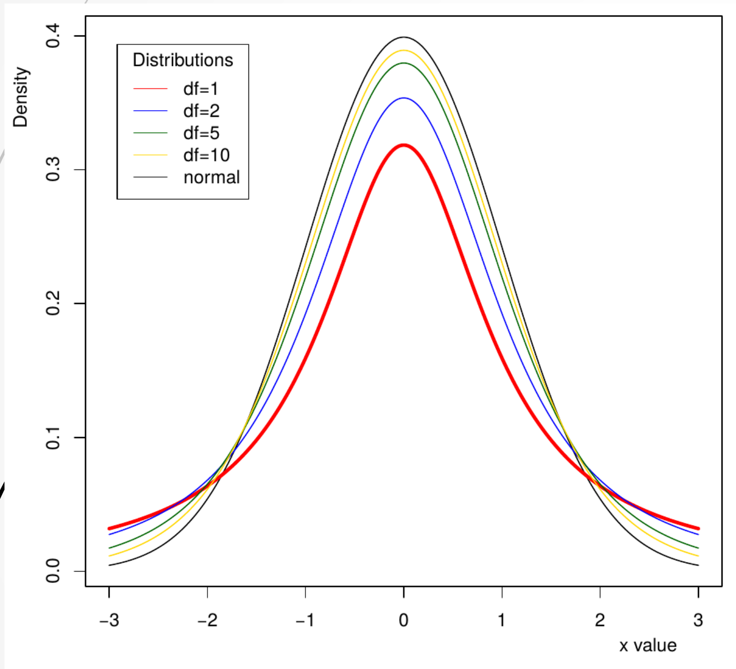
- We define a new $\mathrm{R.V.}$ T

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \rightarrow P\left(-t \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t\right) = 1 - \alpha$$

- The R.V. T follows the **Student's t-distribution** (actually there is a whole family of distribution) $T \sim t(\nu)$

t-distribution

- t-distribution is similar to the normal on (obviously!)



	50%	90%	95%	99%	99.9%
DF=5	0.73	2.02	2.57	4.03	6.87
DF=10	0.70	1.81	2.23	3.17	4.59
DF=20	0.69	1.72	2.09	2.85	3.85
DF=30	0.68	1.70	2.04	2.75	3.65
DF=50	0.68	1.68	2.01	2.68	3.50
(Normal)	0.67	1.64	1.96	2.58	3.29

- The larger the ν the more resemblance to the normal curve
- We use tables to evaluate the critical values t_c for a given confidence levels, let's continue on the next slide…

$C.I.$ for t-distribution

- Start with some formalities… If we draw a sample of size n from a normal distribution with the mean μ, the $R.V.\ T$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t \quad (\nu = n - 1)$$

- Where $\bar{X}$ is the sample mean and S its standard deviation

$$P\left(-t_c \le \frac{\bar{X} - \mu}{S/\sqrt{n}} \le t_c\right) = 1 - \alpha$$

$$P\left(\bar{X} - t_c \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_c \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

- And the $C.I.$ is centred about the sample mean, which contains the true unknown population parameter μ with probability $1 - \alpha$

$$C.I._{100\cdot(1-\alpha)}^{t} = \left(\bar{x} - t_c \frac{S}{\sqrt{n}}, \bar{x} + t_c \frac{S}{\sqrt{n}}\right)$$

$C.I.$ for t-distribution

- For our timer example, let's pick up the confidence level to be 90% and assume that collected data sample in $n = 11$

$$\frac{1}{2}\alpha = 0.05 \rightarrow t_c = \pm 1.81$$

These guys will set the unit!

$$C.I._{90\%}^{t(10)} = \left(\bar{x} - t_c \frac{S}{\sqrt{n}}, \bar{x} + t_c \frac{S}{\sqrt{n}}\right) =$$

$$= \left(294.8 - 1.81\frac{1.77}{\sqrt{11}}, 294.8 + 1.81\frac{1.77}{\sqrt{11}}\right) =$$

$$= (293.8, 295.8)$$

- With larger data sample, our C.I. is now nicely narrow (so, we add some predictability actually)
- Also, note that 300 second (this was the setting on the timer) is not included inside the interval
- Is it an indication that the device goes consistently early?

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
- Using these two examples, try to come up with 90% C.I. for the absence case

Confidence

- In order to find the confidence interval (C.I.) we solve

$$P(X \leq 3;\mu) = e^{-\mu}\left(1+\mu+\frac{\mu^2}{2!}+\frac{\mu^3}{3!}\right) = 0.025 \rightarrow \mu_- = 0.62$$

$$P(X \geq 3;\mu) = 1 - e^{-\mu}\left(1+\mu+\frac{\mu^2}{2!}+\frac{\mu^3}{3!}\right) = 0.025 \rightarrow \mu_+ = 8.8$$

- And our statistical statement would be: **a 95% confidence interval for the parameter μ of the Poisson model, evaluated using a single observation is** $(\mu_-, \mu_+) = (0.62, 8.8)$
- The obtained confidence interval is very wide – we can make it better by collecting more data!
- The grand summary of what we did: with an observation(s) on the R.V. X, assuming X follows some specified model with an unknown parameter θ, we may evaluate a 95% confidence interval for θ by solving $P(X \leq x) = \frac{1}{2}\alpha$ and $P(X \geq x) = \frac{1}{2}\alpha$. We define the confidence level as $C.L. = (1-\alpha)\%$

Interpretation of $C.I.$

- Since, we need data to construct a $C.I.$ it follows that **for different sample we obtain a different interval** – we call it random interval ($\theta_-, \theta_+$are $R.Vs.$ themselves)
- Consider again a single observation of $R.V.X$ that follows an exponential distribution (the math is very easy). The parameter is μ. Now concentrate! We can express the respective limits of the $C.I.$ using the value of that parameter

$$x_{0.025} \rightarrow 1 - e^{-\frac{x_{0.025}}{\mu}} = 0.025 \rightarrow x_{0.025} = -\ln(0.975) = 0.025\mu$$

$$x_{0.975} \rightarrow 1 - e^{-\frac{x_{0.975}}{\mu}} = 0.975 \rightarrow x_{0.975} = -\ln(0.05) = 3.69\mu$$

- Thus, we have $P(0.025\mu \leq X \leq 3.69\mu) = 0.95.$ We can also rewrite this in terms of the unknown parameter

$$P\left(\frac{X}{3.69} \leq \mu \leq \frac{X}{0.025}\right) = P(0.27X \leq \mu \leq 39.5X) = 0.95$$

$C.I.$ for the variance

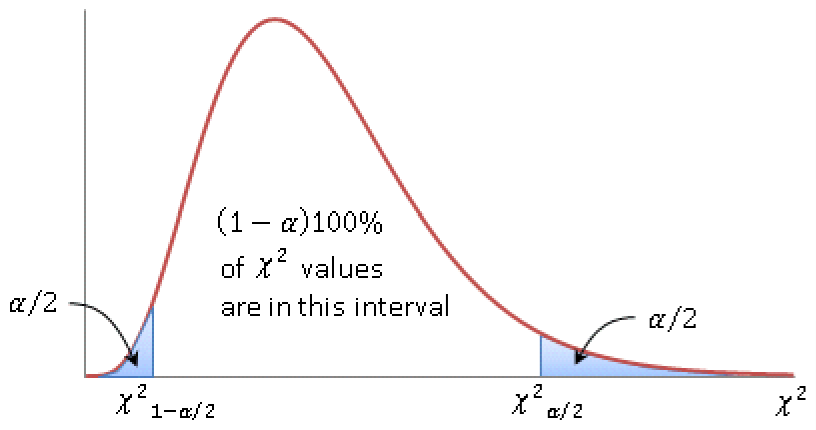
- Imagine that a company is delivering composite fibres for aircraft wings. In that case a great care should be taken to produce fibres that do not vary too much in tensile strength (expressed in kg)
- A sample of 10 fibres were taken and tested, the results were as follow $\bar{x} = 150.72\ kg$ and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

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

- 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)



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

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

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

$C.I.$ 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

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

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

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

- For the timer example, this would give us

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

- Try to work out the C.I. for the normal standard deviation