Statistics MM3: Confidence Intervals

Lecturer: Israel Leyva-Mayorga

email: ilm@es.aau.dk

AALBORG UNIVERSITY
DENMARK

Connectivity

Schedule

1. Introduction to statistics
2. Parameter estimation
3. **Confidence intervals**
4. Hypothesis testing 1
5. Hypothesis testing 2
6. Regression
7. Workshop: wrap-up and exam problems

Outline

Recap on sampling and parametric estimation

What are confidence intervals?

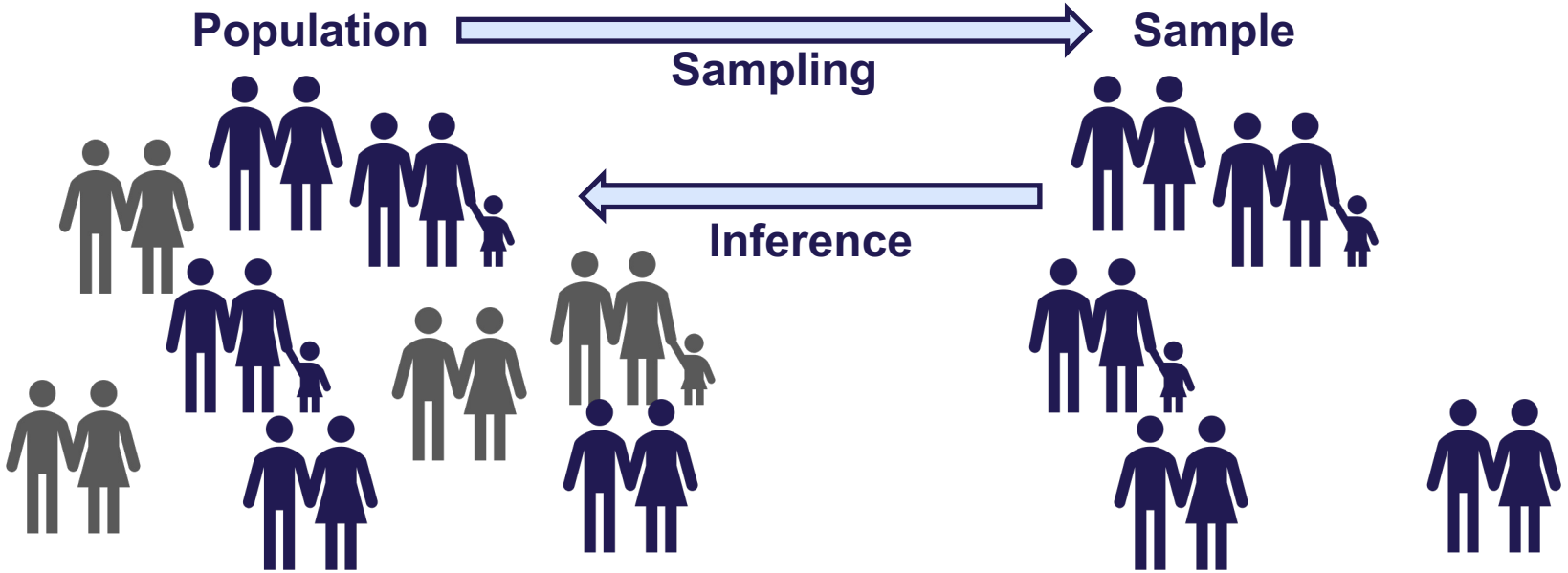
Finding confidence intervals

Estimation of difference of means

Recap on sampling and parametric estimation

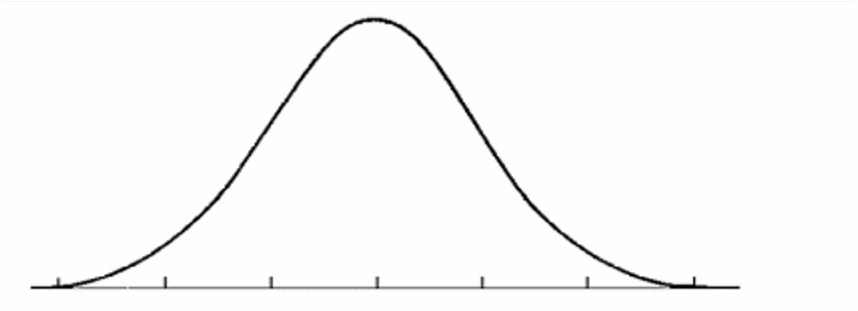
Sampling

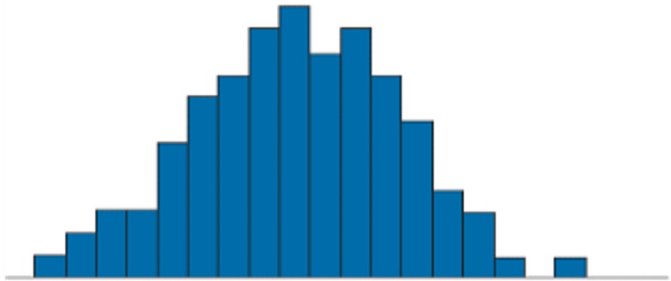
If we cannot measure the whole population, we use a smaller sample



How do we create a sample?

We randomly draw n values from the population
Each value X_i is a random variable with distribution F
We use the sample to estimate some parameter of F (inference)





Parametric estimation

We observe a sample $X_1, X_2, \dots, X_n$
Each value X_i is a random variable with **known distribution** F with parameter θ
The parameter θ is a fixed value and not a RV

From the sample with n points, we create an estimate of θ, denoted as

$$\hat{\theta}_n = h(X_1, X_2, \dots, X_n)$$

Since $\hat{\theta}_n$ depends on the sampled data, it is a RV

We hope that the estimator $\hat{\theta}_n$ is close to the real value of θ
Metrics for estimators: Bias, variance, and Mean Squared Error (MSE)
Maximum Likelihood Estimation (MLE) generates consistent estimators

What are confidence intervals?

Confidence intervals

So far, our estimators for a given parameter θ have given a single value $\hat{\theta}_n$

MLE: we get an expression for the estimator $\hat{\theta}_n$
We define the **likelihood** $\mathcal{L}_n(\theta)$ or **log-likelihood function** $\ell_n(\theta)$
These represent **how likely** is to observe a sample given that F has parameter θ
The value of θ that maximizes $\mathcal{L}_n(\theta)$ and $\ell_n(\theta)$ is the MLE
But the MLE estimator $\hat{\theta}_n$ changes with the values of the sample $X_1, X_2, \ldots, X_n$

Why giving a single value as estimator if we could give a range of values?

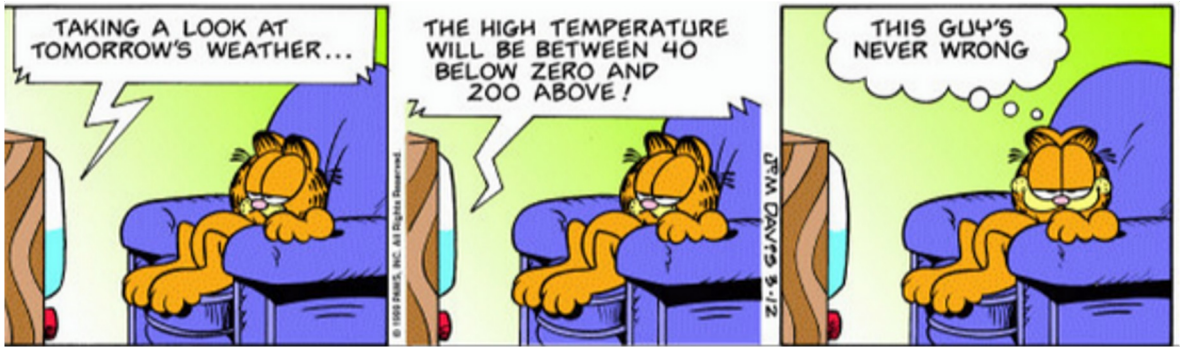
- **Confidence interval**

Confidence intervals

Provide a range of values that captures the most likely outcomes

Precision is important: You want to give the smallest range possible

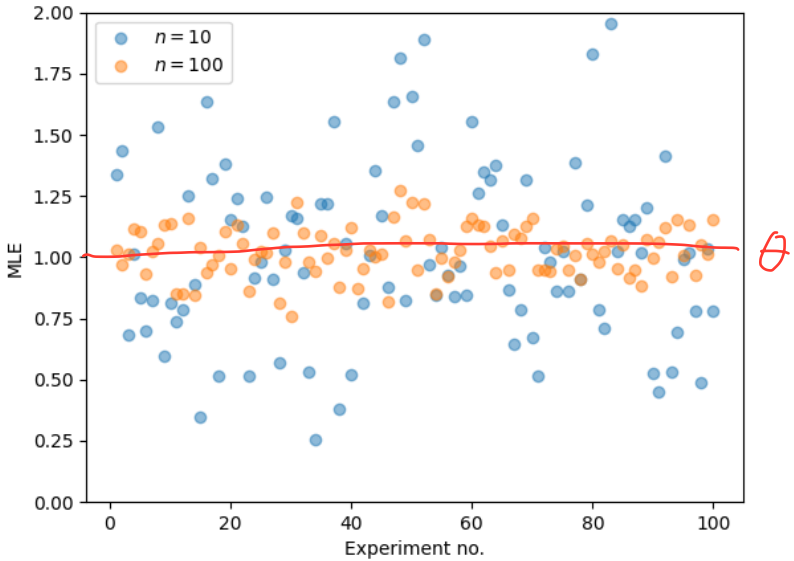
The minimum range is restriced by the variance of the distribution

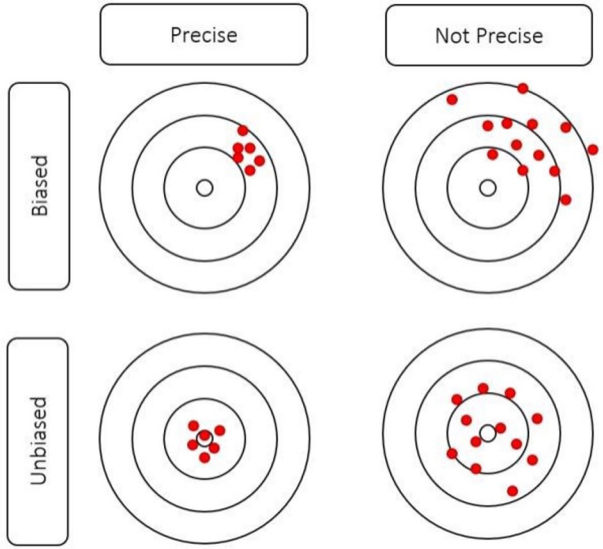


Example from the exercises

The MLE for the temperature with Gaussian noise is the sample mean $\bar{X}_n$

What is the real value of the temperature?

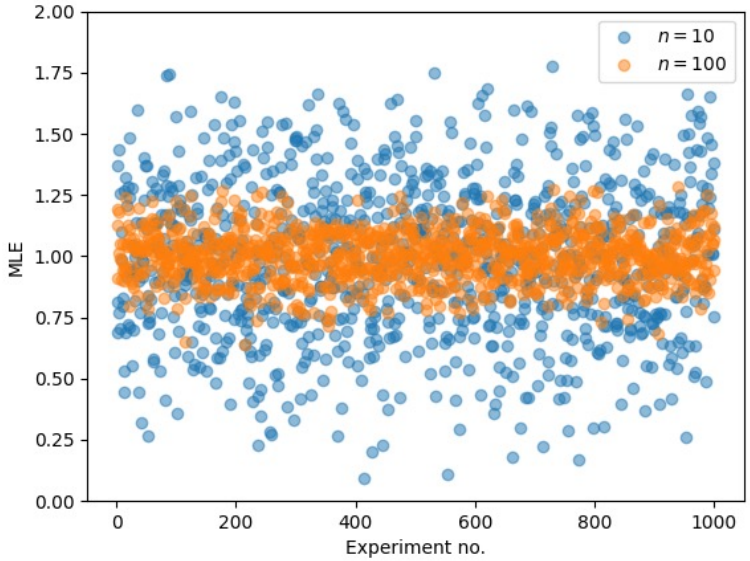


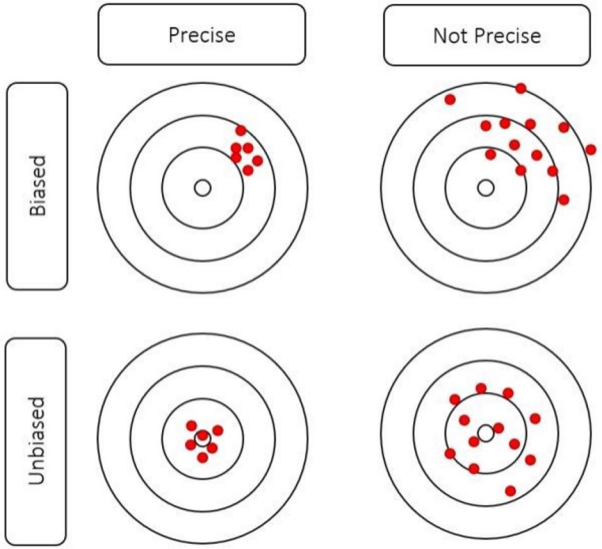


Example from the exercises

The MLE for the temperature with Gaussian noise is the sample mean $\bar{X}_n$

What is the real value of the temperature?

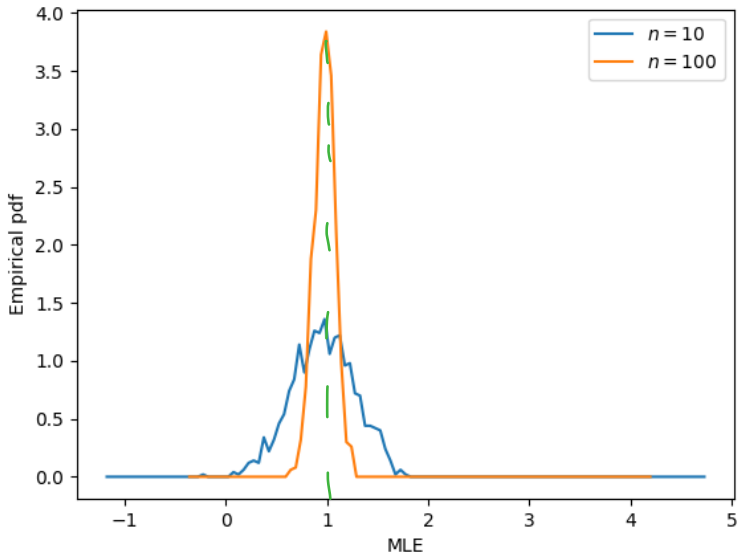


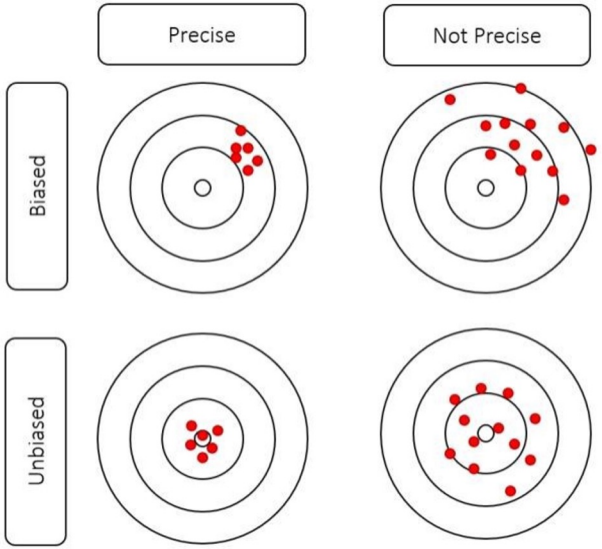


Example from the exercises

The MLE for the temperature with Gaussian noise is the sample mean $\bar{X}_n$

What is the real value of the temperature?





Confidence intervals

We know that the estimator is a RV, so it must have a distribution

The confidence interval (CI) is the best answer to the question:
What is the range of values $C_{1-\alpha} = (a, b)$ around the estimate $\hat{\theta}_n$ such that we are confident with probability $1 - \alpha$ that the true value θ is inside the range?

$$P(\theta \in C_{1-\alpha}) \geq 1 - \alpha$$

Usually $\alpha = 0.05$ so we look at the 95% confidence interval

$$C_{0.95} = (a, b)$$

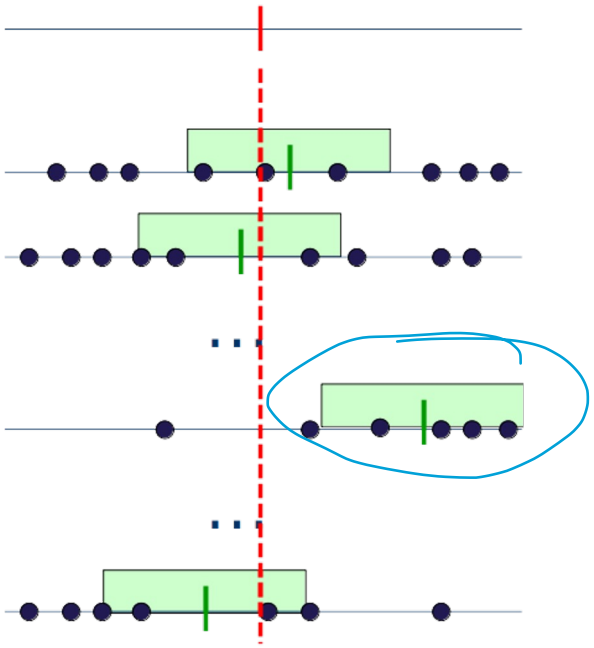
Graphical representation

The red line is the true value of the parameter

A total of m experiments are conducted:

- Take 10 measurements: RVs $X_1, X_2, \dots, X_{10}$
- Use the sample mean to estimate the parameter
- Calculate the 95% CI around the sample mean

If m is large, 95% of the CIs will contain the real value



Finding confidence intervals

Confidence intervals for normal RVs with known σ^2

If $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$, we know that the MLE of the μ is the sample mean

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

That is, from the **Central Limit Theorem**, a sum of RVs is normal distributed
Therefore, the variance of our estimator is

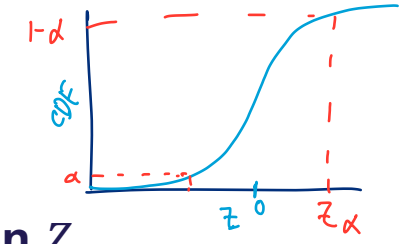
$$\text{var}(\hat{\mu}_n) = \sigma^2/n$$

$$\text{std}(\hat{\mu}_n) = \sqrt{\text{var}(\hat{\mu}_n)} = \frac{\sigma}{\sqrt{n}}$$

From 1st lecture (slide 36):
For normal RVs, 95% of outcomes are between $\hat{\mu}_n - \mathbf{1.96}\sigma$ and $\hat{\mu}_n + \mathbf{1.96}\sigma$

$$C_{0.95} = \left(\hat{\mu}_n - \mathbf{1.96}\frac{\sigma}{\sqrt{n}}, \hat{\mu}_n + \mathbf{1.96}\frac{\sigma}{\sqrt{n}}\right)$$

Finding arbitrary confidence intervals

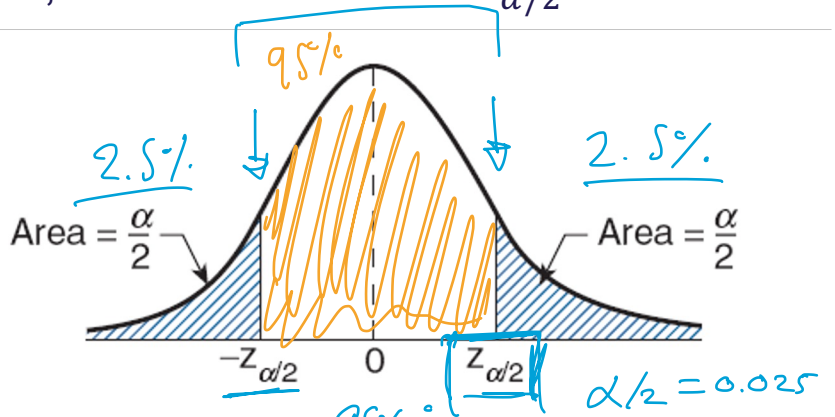


We define the standard normal complementary **quantile function** Z_α

$$Z_\alpha = \inf\{z \in R : \Phi(z) \geq 1-\alpha\}, \qquad \alpha \in [0,1]$$

The total area of the region outside the CI is α, so we need to find $Z_{\alpha/2}$

$$C_{1-\alpha} = \left(\hat{\mu}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$



How do I get $Z_{\alpha/2}$ easily?

In Matlab: norminv($\alpha/2, \mu, \sigma$)

In Python: X.ppf($\alpha/2$) with X being a **scipy.stats.norm** object

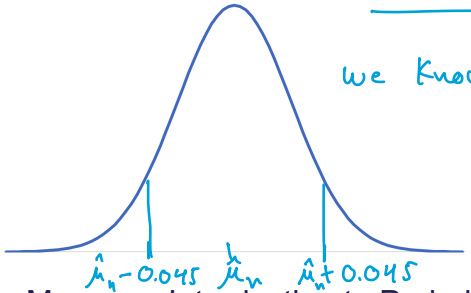
$\alpha/2 = 0.025$

95%: $\alpha = 0.05 \Rightarrow 5\%$

$1-\alpha = 0.95 \Rightarrow 95\%$

Example: Something smells fishy!

From past experience, it is known that the weights of the salmon grown at a commercial hatchery are normal distributed.
The mean of the weight changes per season but the standard deviation remains fixed at 0.135 kg.
We want to estimate the mean weight for the present season with a sample size n
How large should n be if we want to make sure that the estimate of the mean weight is correct within ± 0.045 kg with 95% confidence?



Example: Something smells fishy!

We know that $\sigma = 0.135$

We want to have $C_{0.95} = (\hat{\mu}_n - 0.045, \hat{\mu}_n + 0.045)$

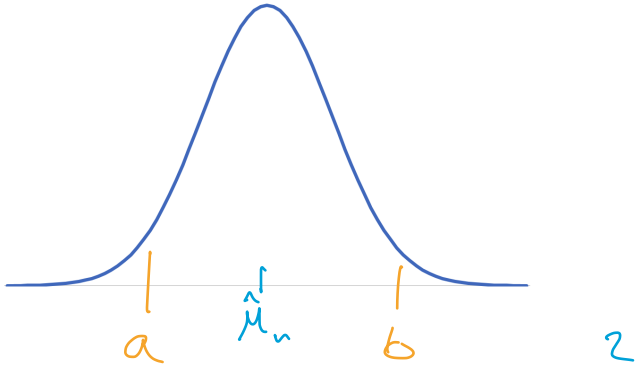
What should be the value of n?

$z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ } std of $\hat{\mu}_n$

1.96

$1.96 \frac{\sigma}{\sqrt{n}} \leq 0.045$

$n = 35$



$n \geq \left(\frac{1.96 \times 0.135}{0.045} \right)^2$

$n \geq 34.57$

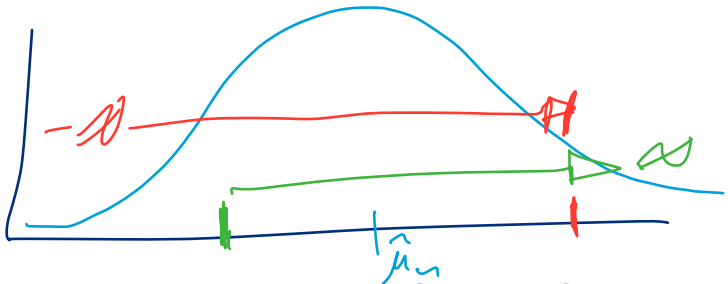
One-sided CI

We can find a point a_l so that the real mean lies above a_l with $1 - \alpha$ confidence

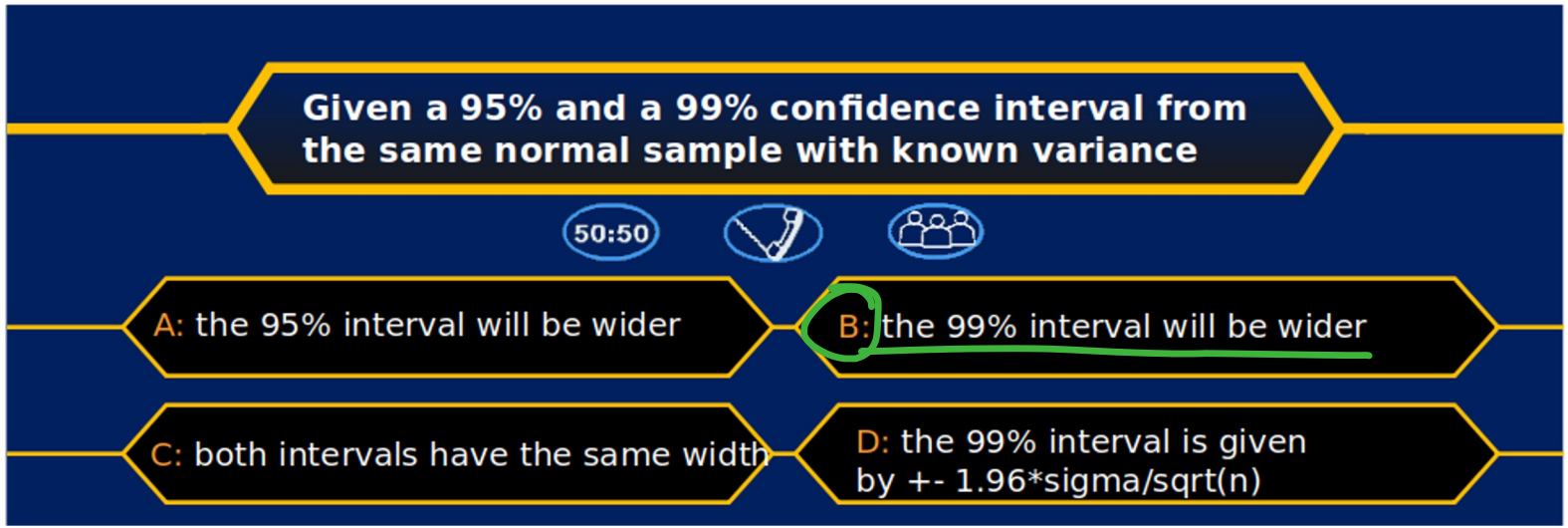
$$a_l = \hat{\mu}_n - Z_\alpha \frac{\sigma}{\sqrt{n}}$$

Or a point b_u so that the real mean lies below b_u with $1 - \alpha$ confidence

$$b_u = \hat{\mu}_n + Z_\alpha \frac{\sigma}{\sqrt{n}}$$



Quiz


Given a 95% and a 99% confidence interval from the same normal sample with known variance
50:50
A: the 95% interval will be wider
B: the 99% interval will be wider
C: both intervals have the same width
D: the 99% interval is given by +- 1.96*sigma/sqrt(n)

Confidence intervals for normal RVs with unknown σ^2

If we knew the real value of σ, we could calculate the $1-\alpha$ CI as

$$C_{1-\alpha} = \left(\hat{\mu}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

But now we don't know σ beforehand, what can we do?

1. Obtain an upper bound for σ, such that $\sigma \leq \sigma_u$
   - In this case, the CI is valid but **conservative**
2. Estimate σ
   - In this case, we obtain an approximation for the CI
   - **Problem:** If we're not careful, the level of confidence will be less than the claimed one

Example: The presidential election

We are hired to estimate the results of an election with two candidates A and B

We ask n voters: Who will you vote for?

Our sample is $X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p)$

$X_i = 1$ if voter will vote for candidate A and is 0 otherwise

How do we calculate the confidence interval?

We know that $\mathbb{E}[X_i] = p$ and that $\text{var}[X_i] = \sigma^2 = p(1-p)$

So, we need p to estimate σ but p is what we want to estimate!

Approach 1: finding an upper bound for σ

We know that $\sigma^2 = p(1-p)$, when do we have the largest $\sigma_u \geq \sigma$?

Define $f(p) = p(1-p)$

Find the maximum for $f(p)$ by doing

$$\frac{df(p)}{dp} = \frac{d\big(p(1-p)\big)}{dp} = p(-1) + (1)(1-p) = 1 - 2p = 0$$

$1 = 2p$

$p^* = \frac{1}{2}$

So, the maximum variance is obtained when $p = \frac{1}{2}$, for which $\sigma_u = \sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)} = \sqrt{\frac{1}{2} - \frac{1}{4}} = \frac{1}{2}$

And we obtain a conservative CI as

$$C_{1-\alpha} = \left(\hat{p}_n - Z_{\alpha/2}\frac{\sigma_u}{\sqrt{n}}, \hat{p}_n + Z_{\alpha/2}\frac{\sigma_u}{\sqrt{n}}\right) = \left(\hat{p}_n - \frac{Z_{\alpha/2}}{2\sqrt{n}}, \hat{p}_n + \frac{Z_{\alpha/2}}{2\sqrt{n}}\right)$$

Approach 2: Estimate σ from $\hat{p}_n$

We know that the MLE for Bernoulli RVs is the sample mean

$$\hat{p}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

So, we use this to estimate σ^2 from $\hat{p}_n$ as

$$\hat{\sigma}_n^2 = \hat{p}_n(1-\hat{p}_n) = \bar{X}_n(1-\bar{X}_n)$$

With this, we obtain an approximate CI in the form

$$C_{1-\alpha} = \left(\hat{p}_n - Z_{\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{p}_n + Z_{\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}\right) = \left(\hat{p}_n - Z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \hat{p}_n + Z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right)$$

General approximation of the CI by estimating σ

The latter formulation is only valid for Bernoulli RVs

But **we know** that the sample variance is an unbiased approximator for σ

$$\hat{\sigma}_n^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2$$

So we can can obtain a **general approximation of the CI** as

$$C_{1-\alpha} = \left(\hat{p}_n - Z_{\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{p}_n + Z_{\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}\right) = \left(\hat{p}_n - Z_{\alpha/2}\frac{S_n}{\sqrt{n}}, \hat{p}_n + Z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right)$$

An example with numbers

A manufacturer produces computer chips.
Eash chip is of sufficient quality with probability p.
An initial sample of 30 chips was taken to get $\hat{p}_n = 26/30$
Assuming that $\hat{p}_n$ is accurate, how large should the sample be to obtain a 99% confidence interval whose length is approximately 0.05?

1. Using the conservative approach and the fact that $Z_{0.005} = 2.576$

$$\frac{Z_{\alpha/2}}{2\sqrt{n}} \leq 0.025 \Longrightarrow n \geq \left(\frac{2.576}{0.05}\right)^2 = 2654.31, \text{ so } n = 2655$$

2. Using the approximation for Bernoulli RVs

$$Z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leq 0.025 \Longrightarrow n \geq \left(\frac{2.576}{0.025}\right)^2 \hat{p}_n(1-\hat{p}_n) = 1226.88, \text{ so } n = 1227$$

The problem of small sample size pt. 1

With a small sample size we cannot get a good estimate for the variance

Recall that, if the sample size is large

$$\hat{\mu}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

And if $X \sim N(\mu, \sigma^2)$, then $aX + b$ has mean $a\mu + b$ and standard deviation $|a|\sigma$

Therefore, the RV

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1)$$

The problem of small sample size pt. 2

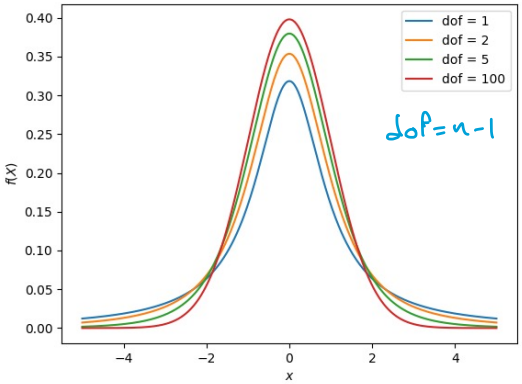
Problem: We only have an estimate for σ which is not accurate for low n

How can we estimate the confidence interval?

We rely on the fact that the RV

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n} \sim T_{n-1}$$

has a t-distribution with $n-1$ degrees of freedom



The pdf of the t-distribution is complicated…

… but is implemented in standard mathematical libraries

The solution to the problem of small sample size

Use the t-distribution to calculate the CI

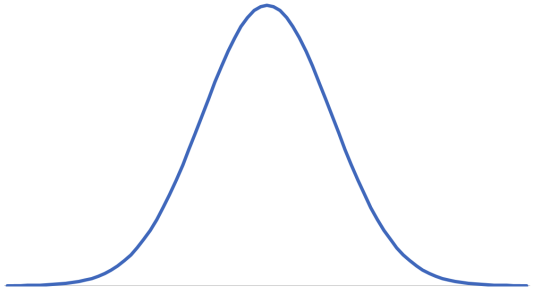
$$C_{1-\alpha} = \left(\hat{\mu}_n - T_{\frac{\alpha}{2},n-1}\frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{\mu}_n + T_{\frac{\alpha}{2},n-1}\frac{\hat{\sigma}_n}{\sqrt{n}}\right)$$

How?

Matlab: $T_{\frac{\alpha}{2},n-1}$ = tinv(1- α/2,n-1)

Python:

1. Create object X = scp.stats.t(n-1, 0,1)
2. Get quantile X.ppf(1- α/2)



Back to the chips example

A manufacturer produces computer chips.
Eash chip is of sufficient quality with probability p.
An initial sample of 30 chips was taken to get $\hat{p}_n = 26/30$
Assuming that $\hat{p}_n$ is accurate, how large should the sample be to obtain a 99% confidence interval whose length is approximately 0.05?

1. Using the conservative approach we got $n = 2655$
2. Using the approximation for Bernoulli RVs we got $n = 1227$
3. **Using the t-distribution we get**

$$T_{\frac{\alpha}{2},30-1}\frac{\hat{\sigma}_n}{\sqrt{n}} \leq 0.025 \Longrightarrow n \geq \left(\frac{2.756}{0.025}\right)^2 \hat{p}_n(1-\hat{p}_n) = 1404.3$$

So, our most accurate result is $n = 1405$

Estimation of difference of means

Introductory example pt. 1

Two different types of cable insulation have been tested.
We need to determine the voltage level at which failures occur
The tests revealed that the individual cables failed at the following voltage levels

Type A		Type B	
36	54	52	60
44	52	64	44
41	37	38	48
53	51	68	46
38	44	66	70
36	35	52	62
34	44		

Introductory example pt. 2

We know that the voltage level that cables of **type A** can whitstand is normally distributed with **unknown mean** μ_A and **known variance** $\sigma_A^2 = 40.$
We also know that the voltage level that cables of **type B** can whitstand is normally distributed with **unknown mean** μ_B and **known variance** $\sigma_B^2 = 100.$
Determine a 95% CI for $\mu_A - \mu_B$

Solution to introductory example

We know that the sum and difference of normal RV are normal, so

$$\bar{X}_n - \bar{Y}_m \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

variance

$$\mu_A - \mu_B = -13.0476$$

$$C_{0.95} = \left(\hat{\mu}_{X_n} - \hat{\mu}_{Y_m} - 1.96\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, \hat{\mu}_{X_n} - \hat{\mu}_{Y_m} + 1.96\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right)$$

$$= (-19.6042, -6.4909)$$

Summary

Summary

The Maximum Likelihood Estimator (MLE) is consistent but gives a single value

How do we know, with a high confidence, which values we might observe?

Confidence Intervals (CIs)

If we know the variance we can easily calculate exact Cis

If we don't know the variance we have options:

1. **Pessimistic:** Use the worst-case for the specific distribution
2. **Optimistic:** Estimate the variance and use the quantile for normal RVs $Z_{\alpha/2}$
3. **Best:** Estimate the variance and use the quantile for t-distribution RVs $T_{\frac{\alpha}{2},n-1}$

The methodology for CIs with difference of means $\mu_A - \mu_B$ is similar