Metody probabilistyczne i statystyka, 2022
informatyka algorytmiczna, WIiT PWr

6-Statistical Inference

Goal: parameter estimation

- population given
- distribution is known (e.g. normal distribution)
- parameters of the distribution --- to be determined

Example:

λ of the Poisson distribution?

Solution: λ=E(X), so estimate the mean

General approach: expressions for mean, variance,… may contain parameters to be estimated

Strategic question:

which function(s) apply to the sample to get a reliable information?

Methods of moments

The k-th population moment is defined as

$$\mu_k = \mathrm{E}(X^k).$$

The k-th sample moment

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Central moments

$$\mu'_k = \mathbf{E}(X - \mu_1)^k$$

$$m'_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$$

Method of moments

values computed according to theory

$$\begin{cases} \mu_1 & = & m_1 \\ \cdots & \cdots & \cdots \\ \mu_k & = & m_k \end{cases}$$

Moments computed from observation

In this system:

- concrete values on the right side
- expressions with parameters on the left side

Method of moments – example

Gamma distribution with parameters α, λ:

$$\begin{cases} \mu_1 & = & \mathrm{E}(X) & = & \alpha/\lambda & = & m_1 \\ \mu_2' & = & \mathrm{Var}(X) & = & \alpha/\lambda^2 & = & m_2'. \end{cases}$$

Example: Pareto distribution

well describes the distribution of file sizes sent on the internet

Its cdf:

$$F(x) = 1 - \left(\frac{x}{\sigma}\right)^{-\theta} \quad \text{for } x > \sigma.$$

Pareto distribution

cdf:

$$F(x) = 1 - \left(\frac{x}{\sigma}\right)^{-\theta} \quad \text{for } x > \sigma.$$

So the density is:

$$f(x) = F'(x) = \frac{\theta}{\sigma}\left(\frac{x}{\sigma}\right)^{-\theta-1} = \theta\sigma^{\theta}x^{-\theta-1}$$

Pareto distribution -- computing moments:

$$\begin{aligned}\mu_1 &= \mathrm{E}(X) = \int_\sigma^\infty x\, f(x)\, dx = \theta\sigma^\theta \int_\sigma^\infty x^{-\theta} dx \\ &= \theta\sigma^\theta \left. \frac{x^{-\theta+1}}{-\theta+1} \right|_{x=\sigma}^{x=\infty} = \frac{\theta\sigma}{\theta-1}, \quad \text{for } \theta > 1,\end{aligned}$$

$$\mu_2 = \mathrm{E}(X^2) = \int_\sigma^\infty x^2\, f(x)\, dx = \theta\sigma^\theta \int_\sigma^\infty x^{-\theta+1} dx = \frac{\theta\sigma^2}{\theta-2}, \quad \text{for } \theta > 2.$$

Pareto distribution

$$\begin{cases} \mu_1 &= \dfrac{\theta\sigma}{\theta - 1} &= m_1 \\ \mu_2 &= \dfrac{\theta\sigma^2}{\theta - 2} &= m_2 \end{cases}$$

so after some calculations:

$$\hat{\theta} = \sqrt{\frac{m_2}{m_2 - m_1^2}} + 1 \quad \text{and} \quad \hat{\sigma} = \frac{m_1(\hat{\theta} - 1)}{\hat{\theta}}.$$

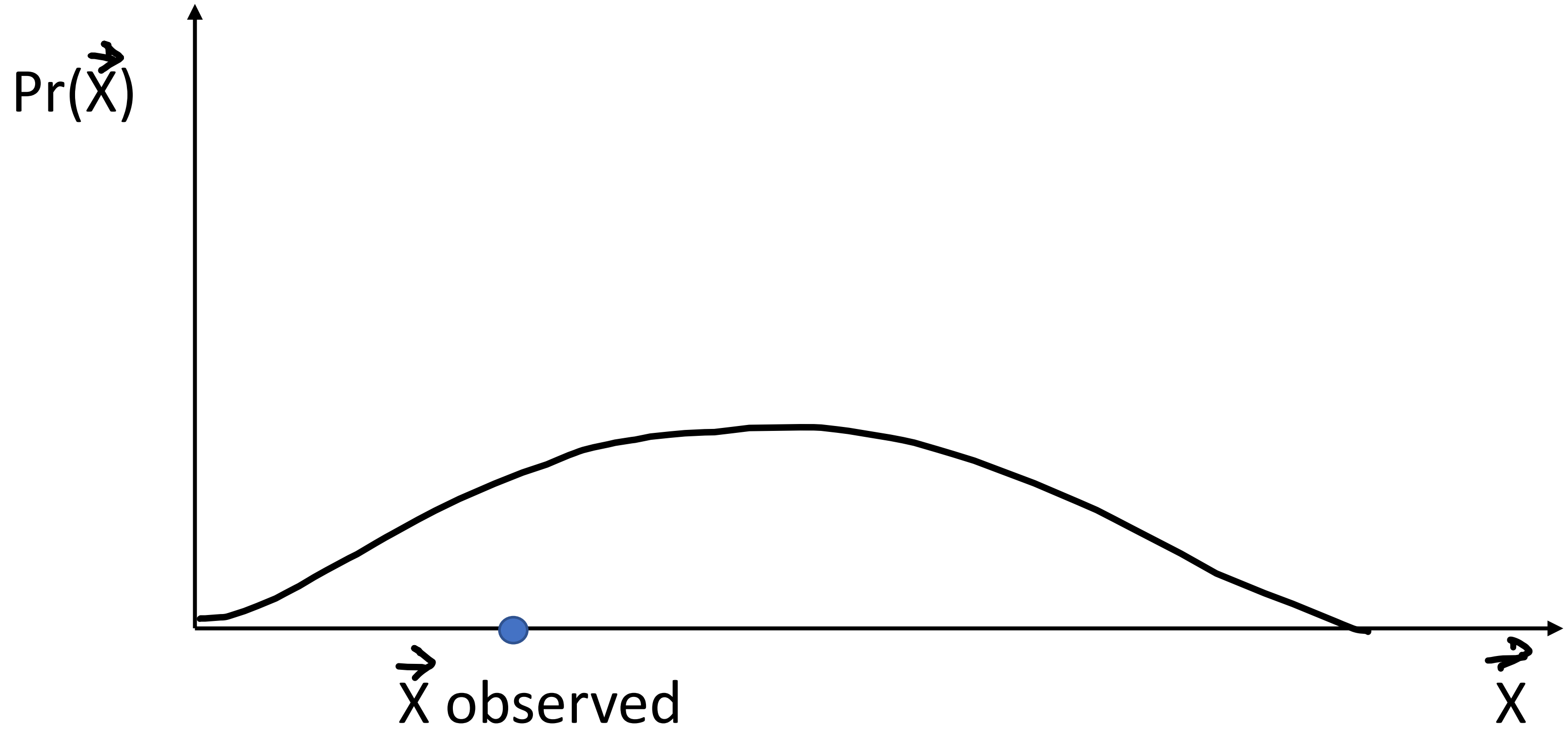
Method of Maximum Likelihood

Sample: $X_1, \dots, X_n$
Distribution: with unknown parameter λ

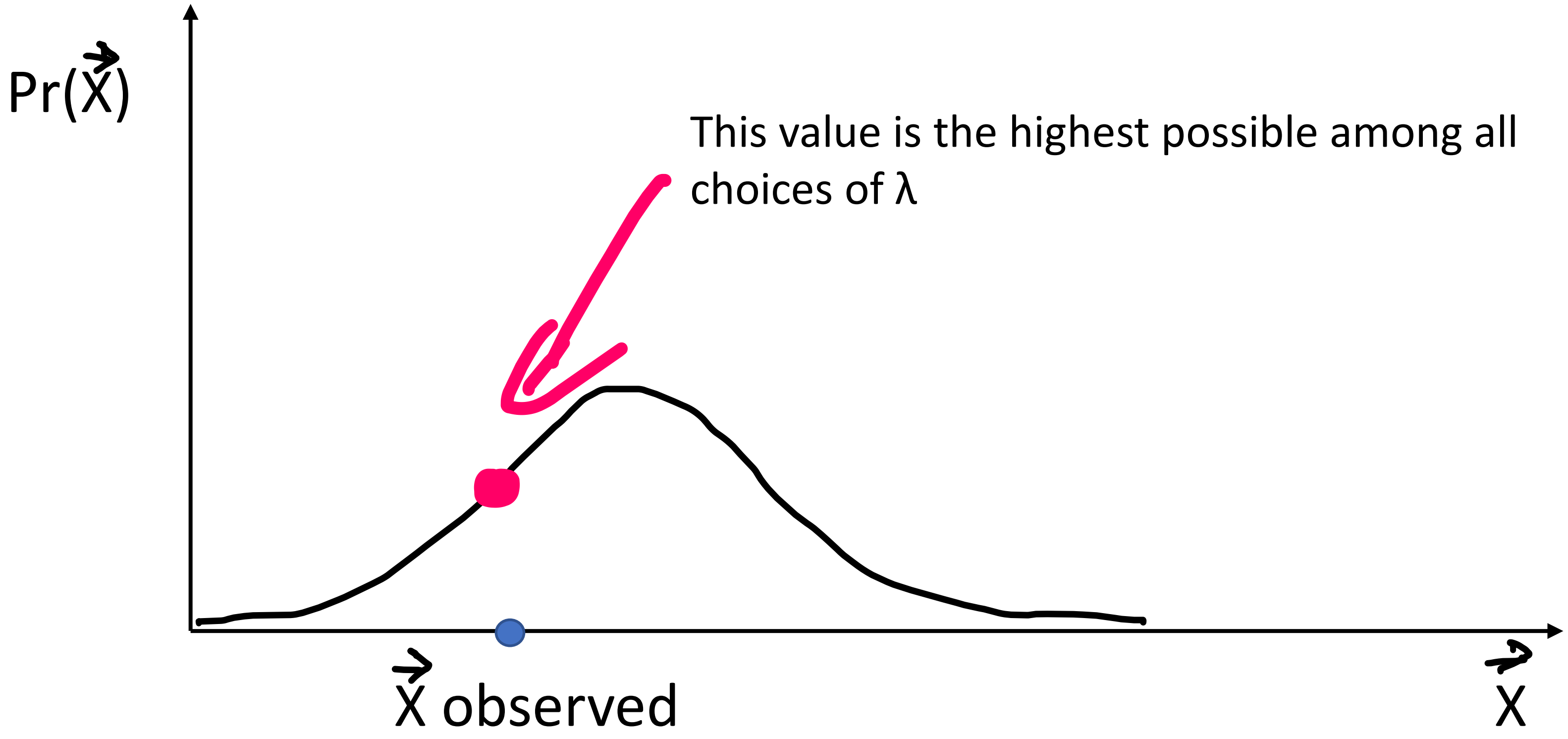
What is the value of λ?

Method:
find λ for which obtaining $X_1, \dots, X_n$ has the highest probability

For a choice of parameter λ:



parameter λ is chosen so that :



Method of Maximum Likelihood –Discrete Case

The goal is to maximize:

$$P\{X=(X_1,\ldots,X_n)\}=P(X)=P(X_1,\ldots,X_n)=\prod_{i=1}^{n}P(X_i),$$

Trick: it is easier to maximize a sum than a product, so take logarithms:

$$\ln\prod_{i=1}^{n}P(X_i)=\sum_{i=1}^{n}\ln P(X_i)$$

Method of Maximum Likelihood –example Poisson distribution

Probability: $$P(x) = e^{-\lambda} \frac{\lambda^x}{x!},$$

logarithms: $$\ln P(x) = -\lambda + x \ln \lambda - \ln(x!).$$

Maximize: $$\ln P(X) = \sum_{i=1}^{n} (-\lambda + X_i \ln \lambda) + C = -n\lambda + \ln \lambda \sum_{i=1}^{n} X_i + C,$$

Finding local maximum: $$\frac{\partial}{\partial \lambda} \ln P(X) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0.$$

Solution: $$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

Method of Maximum Likelihood – continuous case

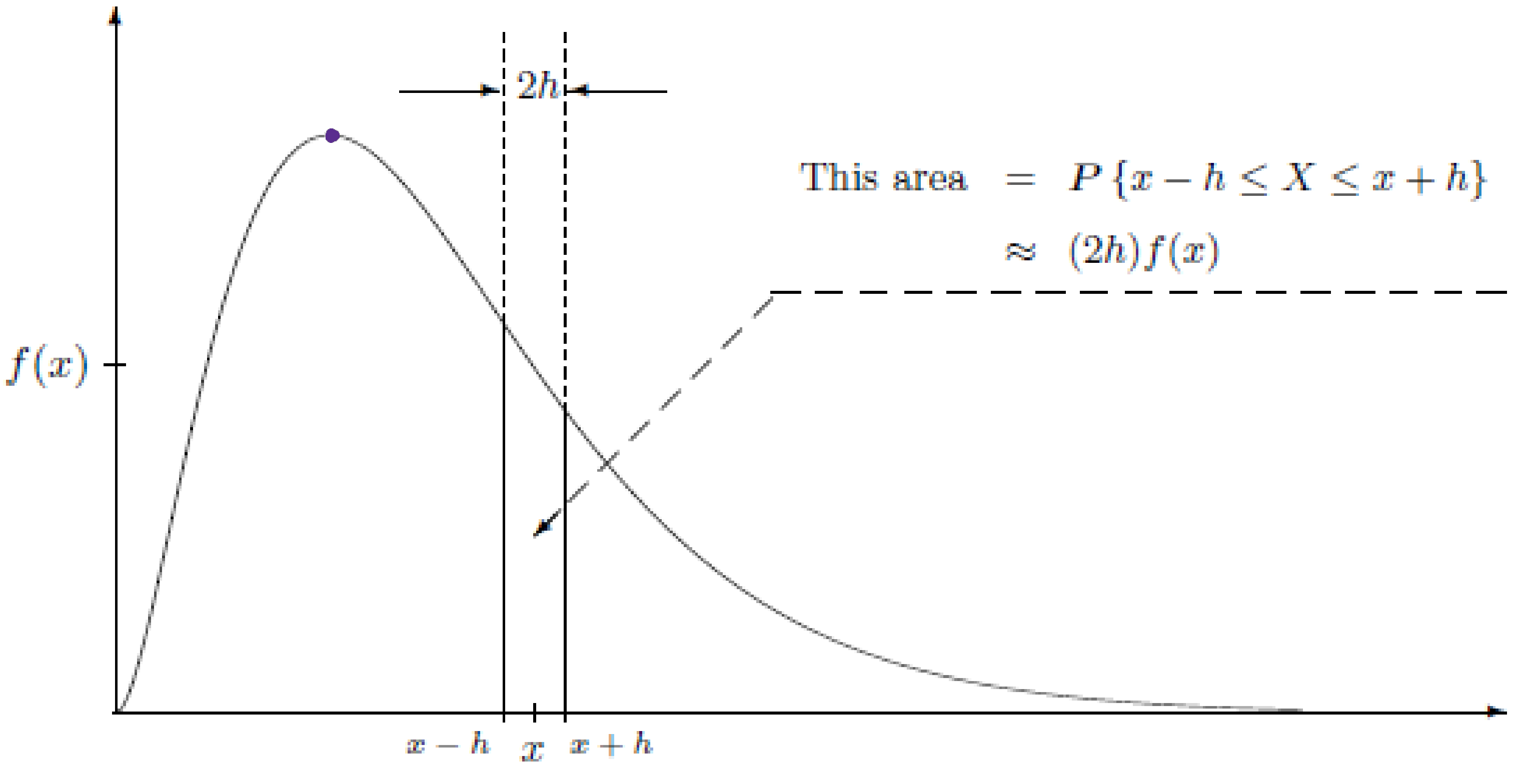


FIGURE 9.1: *Probability of observing "almost" $X = x$.*

Conclusion: take parameters such that f(X) is maximal

Method of Maximum Likelihood – example: exponential density

density: $f(x) = \lambda e^{-\lambda x},$

ln(sample density): $\ln f(X) = \sum_{i=1}^{n} \ln\left(\lambda e^{-\lambda X_i}\right) = \sum_{i=1}^{n} (\ln \lambda - \lambda X_i) = n \ln \lambda - \lambda \sum_{i=1}^{n} X_i.$

Find maximum of ln(*f(X)*):

derivative: $\frac{\partial}{\partial \lambda} \ln f(X) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0,$

solution: $\hat{\lambda} = \frac{n}{\sum X_i} = \frac{1}{\bar{X}}.$

To be checked:
what happens for λ =0 and infinity,
(the maximum is not always where f'(x)=0)

Estimating estimator's error

estimator is a random variable

Question: how concentrated is the estimator value around the true value

Example: Poisson distribution

already we have obtained an estimator for λ:

$$\sigma = \sqrt{\lambda} \text{ for the Poisson}(\lambda) \qquad \hat{\lambda} = \bar{X}$$

Approach 1:

$$\sigma(\hat{\lambda}) = \sigma(\bar{X}) = \sigma/\sqrt{n} = \sqrt{\lambda/n},$$

Then we replace λ by its estimator:

$$s_1(\hat{\lambda}) = \sqrt{\frac{\bar{X}}{n}} = \frac{\sqrt{\sum X_i}}{n}.$$

Example Poisson distribution

Which approach is better?

It depends.

Each method is only an estimation … and not the true value.

Approach 2:

We know that $\sigma(\bar{X}) = \sigma/\sqrt{n},$

So put: $s(\bar{X}) = s/\sqrt{n}$

… and use unbiased estimator for s:

$$s_2(\hat{\lambda}) = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n(n-1)}}.$$

Confidence interval

An interval $[a, b]$ is a $(1 - \alpha)100\%$ **confidence interval** for the parameter θ if it contains the parameter with probability $(1 - \alpha)$,

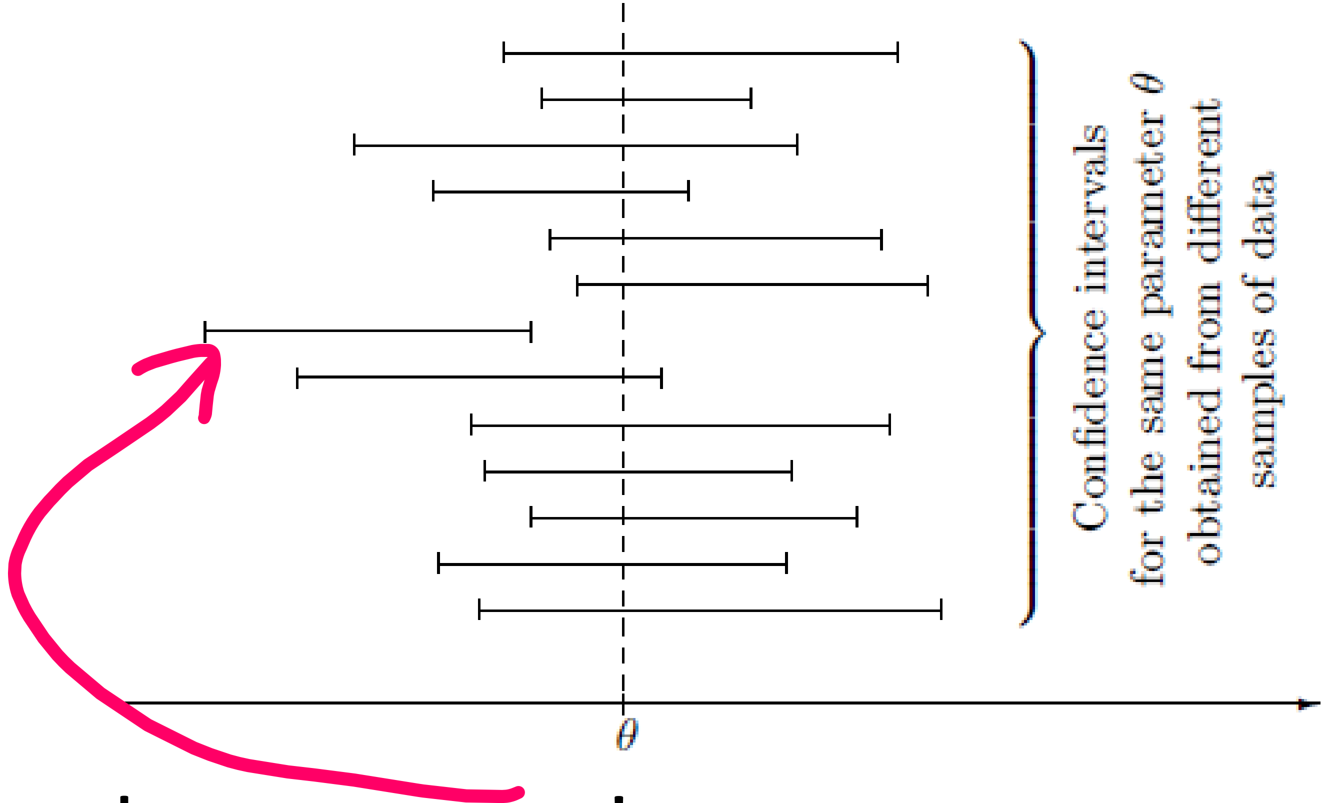
$$\boldsymbol{P}\{a \leq \theta \leq b\} = 1 - \alpha.$$

The **coverage probability** $(1 - \alpha)$ is also called a **confidence level**.

Remember: **we do not know for sure that the true value belongs to the confidence interval!**

Situation

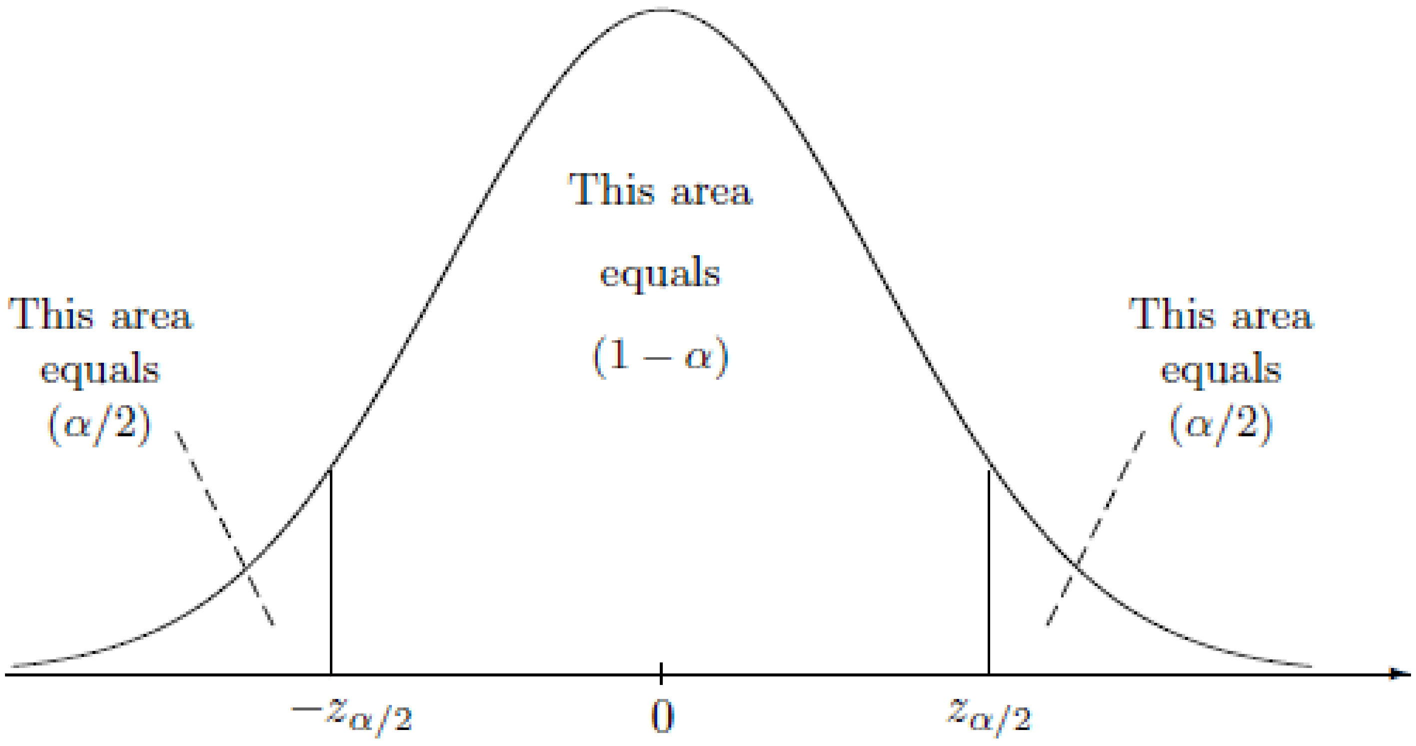
Illustration of computed confidence intervals



some intervals are ok, some are wrong!

wrong ones should occur with small pbb

Confidence interval for normal distribution



Confidence interval for unbiased estimator with normal distribution

after normalizing to Standard Normal distribution:

$$P\left\{-z_{\alpha/2} \le \frac{\hat{\theta}-\theta}{\sigma(\hat{\theta})} \le z_{\alpha/2}\right\} = 1-\alpha.$$

$$P\left\{\hat{\theta} - z_{\alpha/2}\cdot\sigma(\hat{\theta}) \le \theta \le \hat{\theta} - z_{\alpha/2}\cdot\sigma(\hat{\theta})\right\} = 1-\alpha.$$

Confidence interval [a,b] where:

$$\begin{aligned} a &= \hat{\theta} - z_{\alpha/2}\cdot\sigma(\hat{\theta}) \\ b &= \hat{\theta} + z_{\alpha/2}\cdot\sigma(\hat{\theta}) \end{aligned}$$

Application: confidence level for a sample mean

it applies for:

- sum of (a few) random variables with normal distribution
- a large number of samples for any random variable (due to CLT the sum ≈ normal distribution)

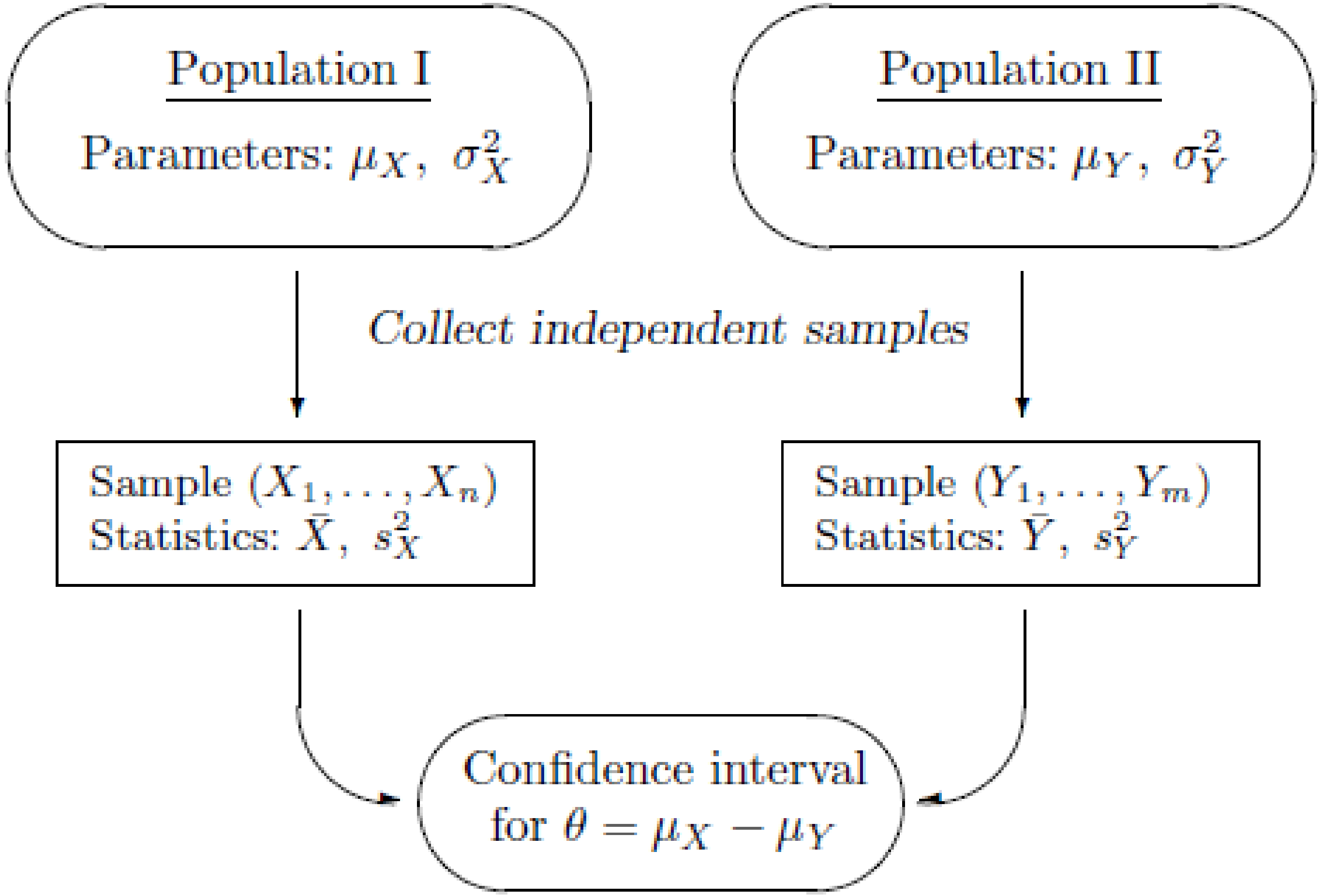
Recall that:

$$\begin{aligned} \mathbf{E}(\bar{X}) &= \mu \\ \sigma(\bar{X}) &= \sigma/\sqrt{n}. \end{aligned}$$

So the confidence interval with endpoints: $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

omitted slide

Confidence interval for difference between two means:



Steps

omitted slide

1. estimator of mean value: $\hat{\theta} = \bar{X} - \bar{Y}$ (it is unbiased)

2. if the sample is large, then approximately normal distribution
3. estimate variance:

$$\sigma(\hat{\theta}) = \sqrt{\mathrm{Var}\left(\bar{X} - \bar{Y}\right)} = \sqrt{\mathrm{Var}\left(\bar{X}\right) + \mathrm{Var}\left(\bar{Y}\right)} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

4. Confidence interval with endpoints:

$$\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

How big should be the sample size?

good question,

- **if we have to pay for each X_i**
- **or getting a new sample is problematic or impossible**

(like finding the next skeleton of Tyranosaurus to estimate their height)

How big should be the sample size?

Confidence interval depends on sample size n and normal distribution:

$$\text{margin} = z_{\alpha/2} \cdot \sigma/\sqrt{n}.$$

So we have a simple rule:

In order to attain a margin of error Δ for estimating a population mean with a confidence level $(1-\alpha)$,

a sample of size $\quad n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{\Delta}\right)^2 \quad$ is required.

so reducing Δ by factor 0.1 increases n by factor 100 (costs!)

Confidence interval for unknown variance

Example: population with fraction *p* of objects with property *A*

Sample proportion: $\hat{p} = \dfrac{\text{number of sampled items from } A}{n}$

So: $X_i = \begin{cases} 1 & \text{if} \quad i \in A \\ 0 & \text{if} \quad i \notin A \end{cases}$

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

Finally: $\hat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ ← This is never higher than 0.25

Problem for small sample size

Then the estimation of variance is quite poor! 😠

What to do??

Recall normalization (for normal distribution):

$$Z = \frac{\hat{\theta} - \mathbf{E}(\hat{\theta})}{\sigma(\hat{\theta})} = \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})},$$

For small sample we consider so called T-ratio:

$$t = \frac{\hat{\theta} - \theta}{s(\hat{\theta})}$$

Student's distribution

Introduced by W. Gosset (pseudonym Student):

for T-ratio: $$t = \frac{\hat{\theta} - \theta}{s(\hat{\theta})}$$

computed for a sample of size *n* for random variable with normal distribution

Subtle issue: T-ratio is not normal (the denominator is also an estimator!)

True distribution: Student's distribution with "*n-1* degrees of freedom"

Using Students distribution:

For each *n*:

A table with precomputed values for any confidence interval

– then follow the same steps as for normal distribution to get the confidence interval:

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Only this has changed

omitted slide

Example: *X-Y* for random variables *X,Y* with variance *σ*:

assumption: $$\sigma_X^2 = \sigma_Y^2 = \sigma^2.$$

sample variance: $$s_p^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2} = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

Also: $$\sigma(\hat{\theta}) = \sqrt{\mathrm{Var}\left(\bar{X} - \bar{Y}\right)} = \sqrt{\mathrm{Var}\left(\bar{X}\right) + \mathrm{Var}\left(\bar{Y}\right)} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

finally: confidence interval from Student's distribution:

$$\bar{X} - \bar{Y} \pm t_{\alpha/2}\, s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

easy..

Omitted slide

Example: difference between two variables with the different variance:

problem: not the Student distribution anymore!
no compact and clean solution

Approximation (only to see):

1. computing "degree of freedom"

$$\nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}}.$$

2. Proceed with formulas for Student's distribution with this degree

$$\bar{X} - \bar{Y} \pm t_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

Hypothesis testing

Population -- claimed property H_0

-- alternative property H_1

so that both cannot hold at the same time

Case 1: unrealistic

Data from the whole population available:

one can say which of them is false

Case 2: real life

Only a sample is available

H_0 or H_1 is true?

Example

H_0= the proportion of defect chips is 3%
H_1 = the proportion of defect chips is >3%

Test outcomes

	Result of the test	
	Reject H_0	Accept H_0
H_0 is true	*Type I error*	correct
H_0 is false	correct	*Type II error*

Examples: biometric recognition, AI is full of such situations

(e.g., H_0= „face seen by the smartphone is the face of the smartphone owner")

Significance level of a test (poziom istotności)

For type 1 error:

$$\alpha = P\{\text{reject } H_0 \mid H_0 \text{ is true}\}$$

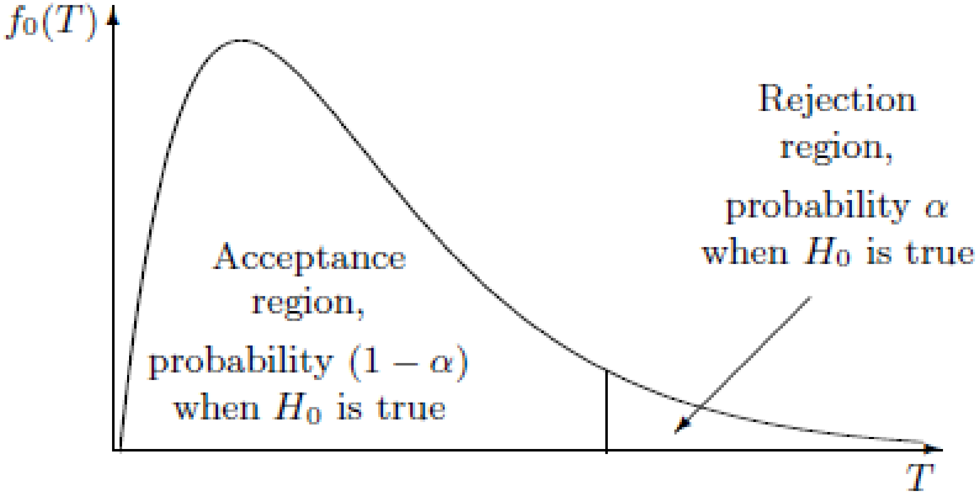
Power of the test

Alternative hypothesis H_A with parameters θ

$$p(\theta) = P\{\text{reject } H_0 \mid \theta;\ H_A \text{ is true}\}.$$

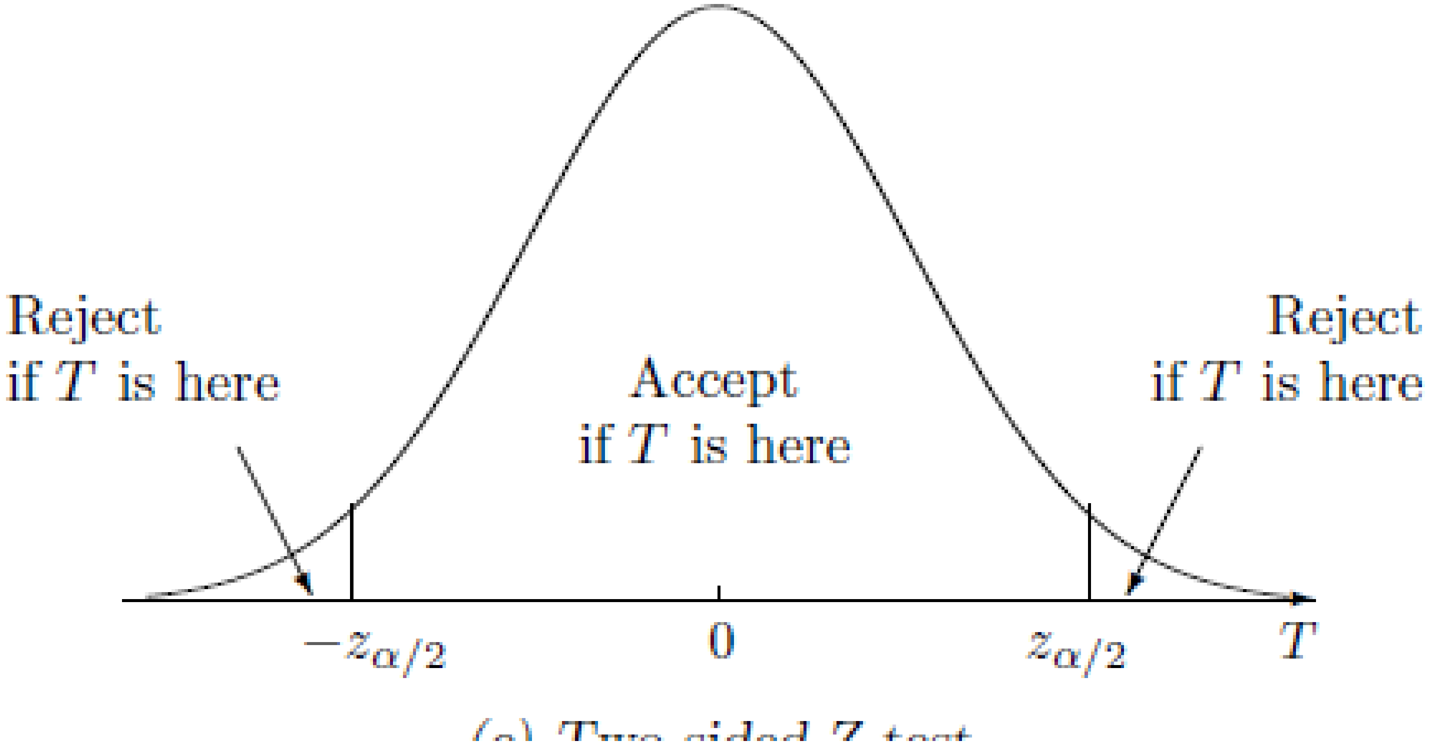
General approach

- H_0 corresponds to some distribution F_0
- define statistic T
- define acceptance and rejection regions so that probability of values from rejection regions is at most α



$$\begin{aligned} \text{Significance level} &= P\{\text{ Type I error }\} \\ &= P\{\text{ Reject } \mid H_0\} \\ &= P\{T \in \mathcal{R} \mid H_0\} \\ &= \alpha. \end{aligned}$$

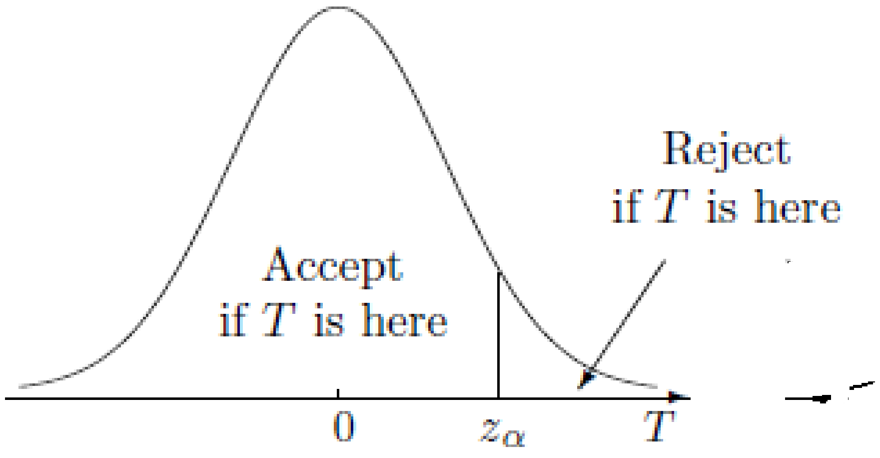
For normal distribution mean 0 – two sided Z test



Right tail alternative

(a) A level α test with a **right-tail alternative** should

$$\begin{cases} \text{reject } H_0 & \text{if} \quad Z \geq z_\alpha \\ \text{accept } H_0 & \text{if} \quad Z < z_\alpha \end{cases}$$

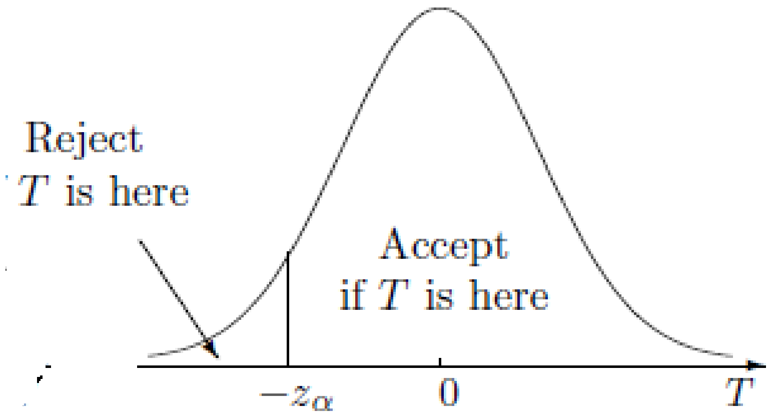


(a) *Right-tail Z-test*

Left tail alternative

With a **left-tail alternative**, we should

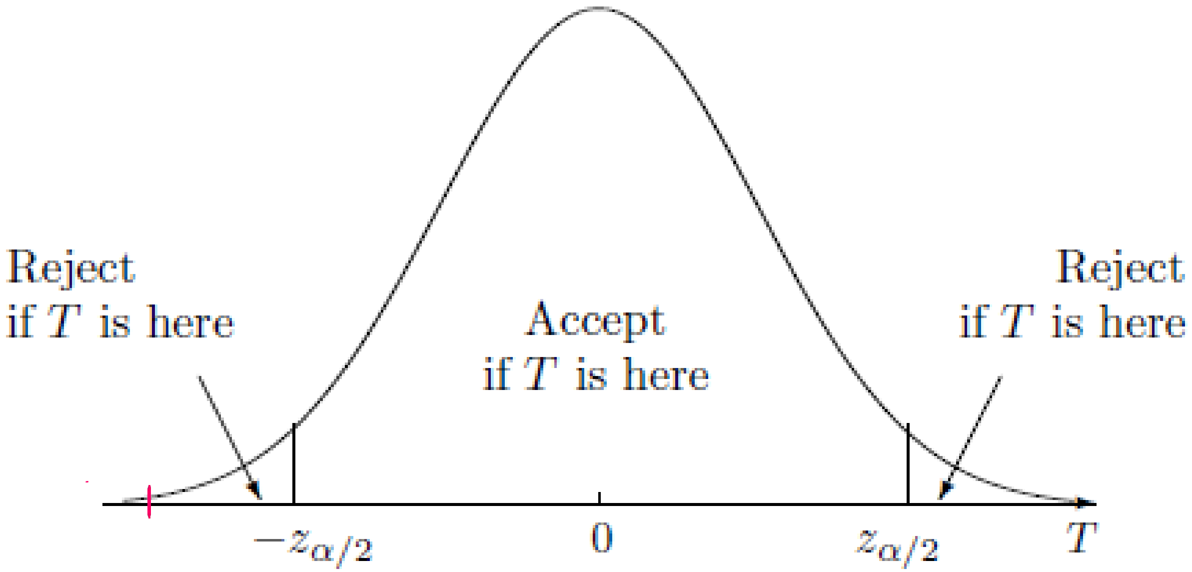
$$\begin{cases} \text{reject } H_0 & \text{if} \quad Z \le -z_\alpha \\ \text{accept } H_0 & \text{if} \quad Z > -z_\alpha \end{cases}$$



(b) *Left-tail Z-test*

Choosing α

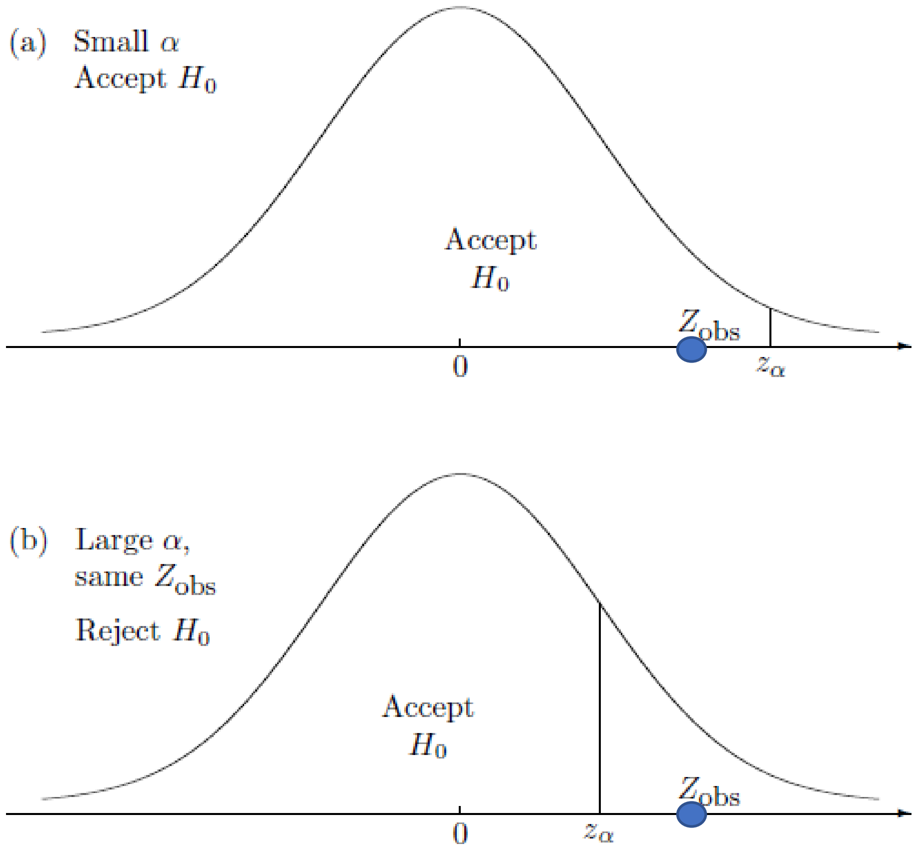
Delicate issue, a tradeoff between errors of type 1 and 2



P-value

For a given observation which values of α force rejection of H_0 and which force acceptance of H_0?

P-value is the boundary between these regions of α



P-value

Testing H_0 with a P-value

For	$\alpha < P,$	accept H_0
For	$\alpha > P,$	reject H_0

Practically,

If	$P < 0.01,$	reject H_0
If	$P > 0.1,$	accept H_0

Confidence intervals and testing for the variance

Important for making decisions based on a sample:

-- system reliability

-- quality testing

Variance unbiased estimator

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

the values $(X_i - \bar{X})^2$ are not independent:

- ❑ each X_i occurs in the sample mean
- ❑ CLT can be applied only for large *n*
- ❑ distribution of s^2 is not even symmetric

Distribution of variance?

Assumption: $X_1, \ldots, X_n$ -- independent, normally distributed with variance σ

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

is *Chi-square with* $(n-1)$ *degrees of freedom*

Density:

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \qquad x > 0,$$

Chi-square distribution

A case of Gamma distribution:

$$\text{Chi-square}(\nu) = \text{Gamma}(\nu/2, 1/2),$$

Deriving from general formulas for Gamma distribution:

$$\mathrm{E}(X) = \nu \quad \text{and} \quad \mathrm{Var}(X) = 2\nu.$$

Chi-square distribution

As you see, for a low degrees of freedom approximating with normal distribution would be a bad idea

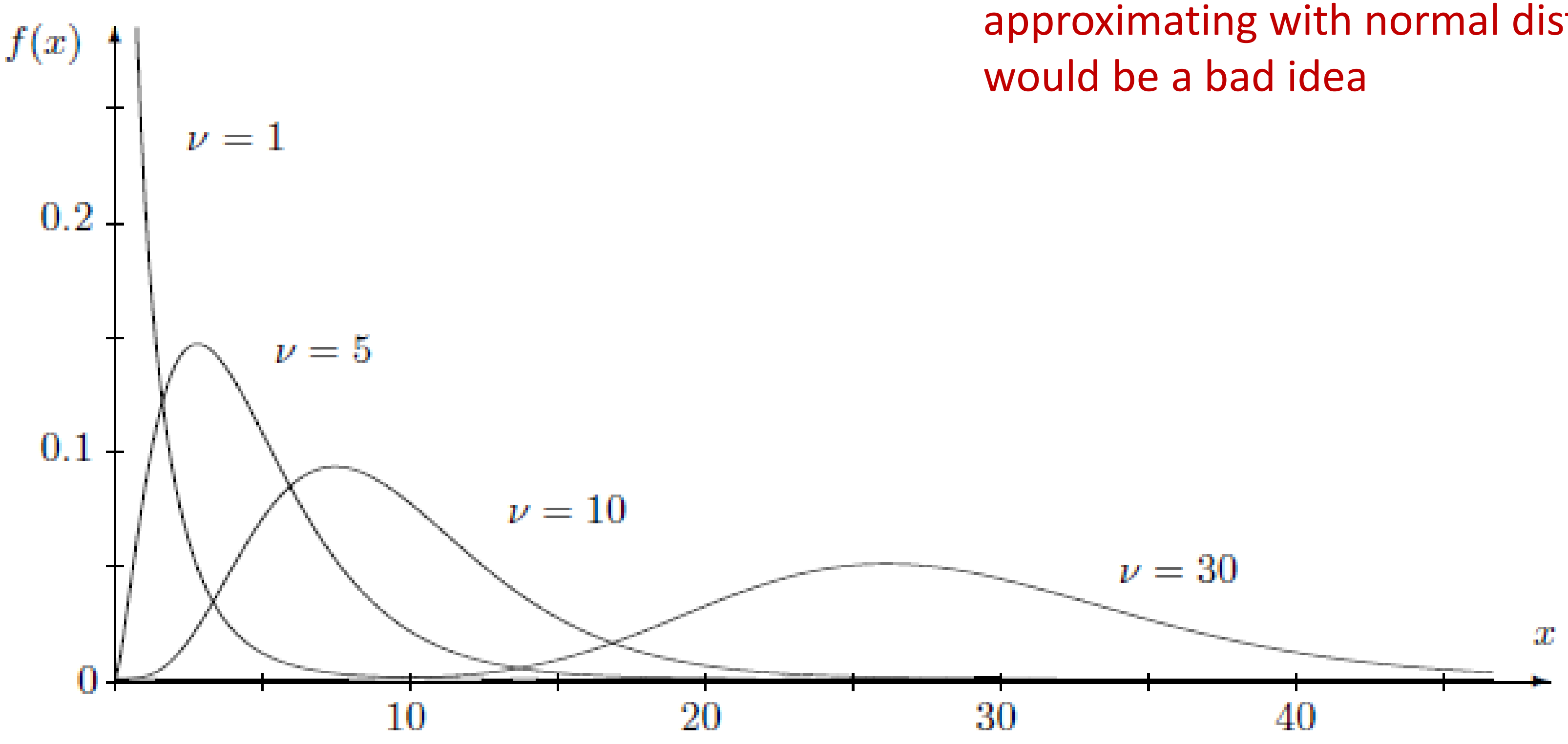
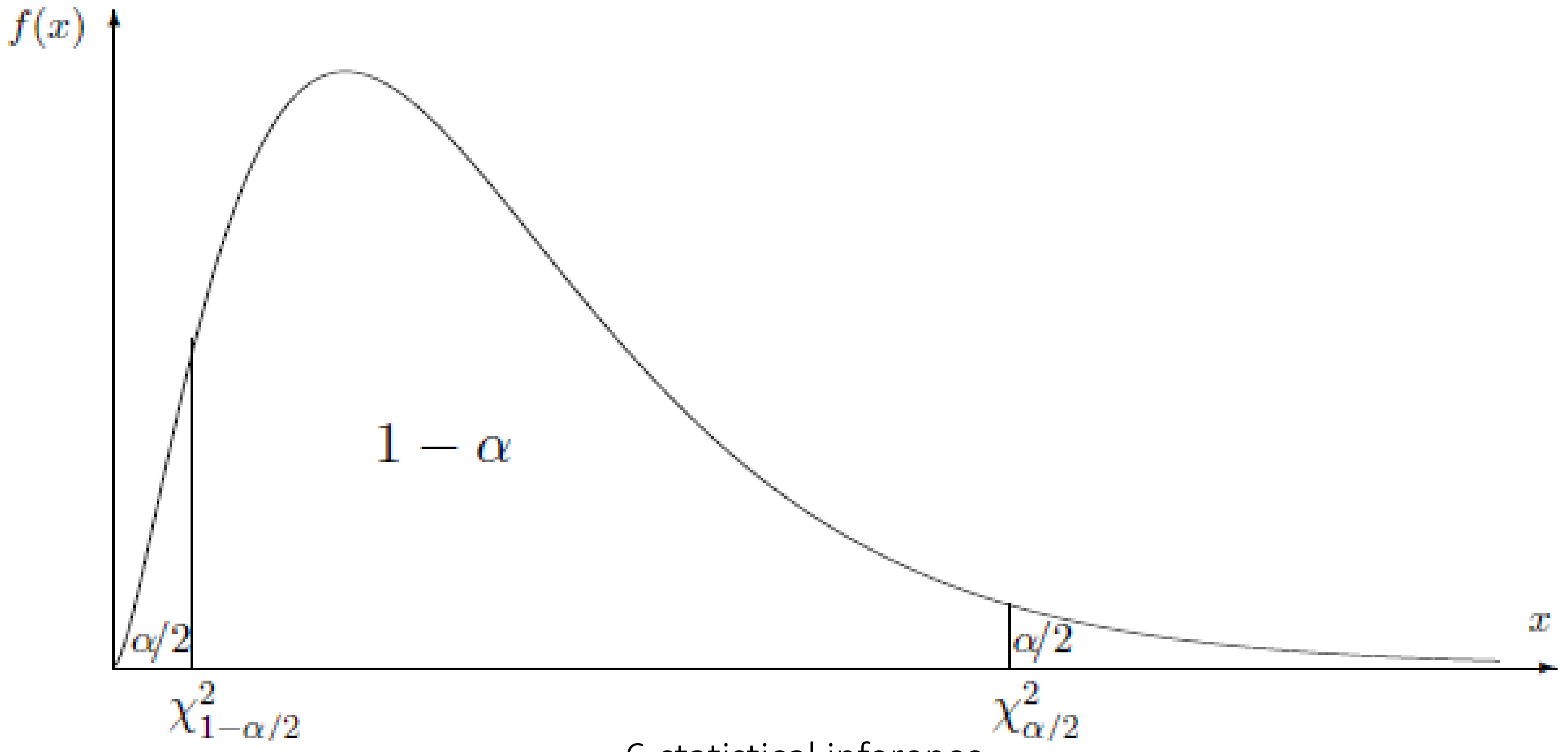


FIGURE 9.12: *Chi-square densities with ν = 1, 5, 10, and 30 degrees of freedom. Each distribution is right-skewed. For large ν, it is approximately Normal.*

Confidence interval

distribution not symmetrical, so the confidence interval is not of the form $s \mp \Delta$

- ➢ two values must be read from precomputed lookup tables



Confidence interval

Confidence interval for the variance

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$$

these values are precomputed and available from functions in many libraries

6-statistical inference