Landmålingens fejlteori
Lektion 3
Estimation af σ
Dobbeltmålinger
Geometrisk nivellement
Linearisering

Rasmus Waagepetersen - rw@math.aau.dk

Institut for Matematiske Fag
Aalborg Universitet

Repetition: Middelværdi og Varians

Sætning: Middelværdi og varians for linearkombinationer

Lad $X_1, X_2, \ldots, X_n$ være stokastiske variable. Da gælder

$$\mathbb{E}\big(a_0 + a_1 X_1 + \cdots + a_n X_n\big) = a_0 + a_1 \mathbb{E}(X_1) + \cdots + a_n \mathbb{E}(X_n)$$

Hvis $X_1, X_2, \ldots, X_n$ desuden er uafhængige gælder

$$\mathbb{V}\mathrm{ar}\big(a_0 + a_1 X_1 + \cdots + a_n X_n\big) = a_1^2 \mathbb{V}\mathrm{ar}(X_1) + \cdots + a_n^2 \mathbb{V}\mathrm{ar}(X_n)$$

Hvis X_i'erne er normalfordelte, så er summen $a_0 + a_1 X_1 + \cdots + a_n X_n$ også normalfordelt.

Estimation af middelværdi

Antag $X_1, \dots, X_n$ er uafhængige stokastiske variable med middelværdi μ og varians σ^2.

Vi estimerer μ ved gennemsnittet

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

Middelværdi og varians af $\bar{X}$:

$$\mathbb{E}\bar{X} = \mu \quad \mathbb{V}\mathrm{ar}\bar{X} = \frac{\sigma^2}{n}$$

Afhængige vs. uafhængige målinger

Lad X og Y repræsentere målinger foretaget af Bent og Børge uafhængigt af hinanden og begge med varians σ^2 (og samme middelværdi μ).

Antag Bent måler en gang til, hvor hans anden måling Z er påvirket af hans resultat for den første måling, dvs.

$$Z = X + \nu$$

hvor ν er uafhængig af X med en (lille) varians ω^2 og middelværdi nul.

Dermed er variansen på $(X + Y)/2$ lig $\sigma^2/2$ mens variansen på $(X + Z)/2$ er $\sigma^2 + \omega^2/4$. Dvs. langt den bedste præcision med uafhængige målinger.

Omvendt: $\mathbb{V}\mathrm{ar}(X - Y) = 2\sigma^2$ mens $\mathbb{V}\mathrm{ar}(X - Z) = \omega^2$, hvad der (fejlagtigt) kunne forlede landinspektørerne til at tro, at Børges målinger var mest præcise.

Hvis X_i alle normalfordelte gælder

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Hvis X_i ikke er normalfordelte

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$

hvor tilnærmelse bedre jo større n.

Konfidensinterval

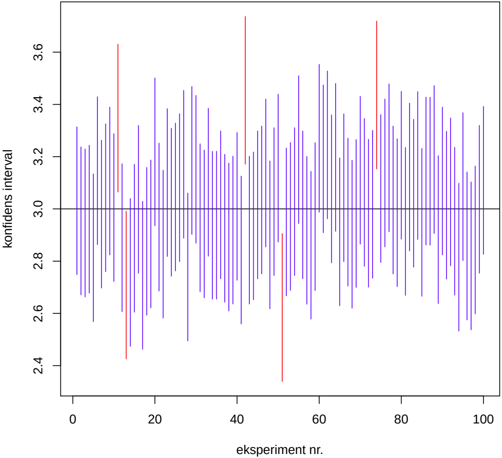
$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

indeholder μ med 95% sandsynlighed

(99% hvis 1.96 erstattes med 2.58)

Konfidensintervaller for 100 simulerede måleserier hver med 100 målinger $(X_{ij} \sim N(3{,}2),\ i{,}j = 1, \ldots, 100)$

Ca. 95% indeholder sande værdi $\mu = 3$.



Anvendelse af konfidensinterval

Enggaard A/S vil have målt en længde med en "præcision" på ±5cm.

Oversættelse til fejlteori: med meget "stor" sandsynlighed skal forskellen mellem estimat $\bar{X}$ og sand længde μ være mindre end 5cm.

Fortolker vi "stor" sandsynlighed som 99.9% skal der altså gælde at $3.29\sigma/\sqrt{n} < 5$.

Eks. $n = 2$ og $\sigma = 5$cm.

Da gælder $3.29\sigma/\sqrt{2} = 11.63$. Dvs. kravet er ikke opfyldt.

Mulig løsning: vælg n så

$$3.29\sigma/\sqrt{n} <= 5 \Leftrightarrow n > (3.29\sigma/5)^2$$

Da skal vi have $n \geq 11$.

Konfidensinterval: generelt set-up

Antag at θ er en ukendt størrelse som estimeres af $Y \sim N(\theta, \tau^2)$.

Da er et 95% konfidensinterval for θ givet ved

$$[Y - 1.96\tau; Y + 1.96\tau]$$

Y kunne eksempelvis være et empirisk gennemsnit af målinger $X_1, \ldots, X_n$ med middelværdi θ og varians σ^2.

Da har vi $\tau^2 = \sigma^2/n$ som før.

Estimation af varians

I nogle tilfælde er variansen σ^2 ukendt. Da må vi estimere σ^2 ud fra data.

Som estimator for σ^2 anvendes S^2:

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

Dette estimat er også centralt, dvs.

$$\mathbb{E}(S^2) = \sigma^2$$

(sætning 17 - vi skipper bevis)

Bemærk: S^2 er "empirisk" version af $\mathbb{E}(X-\mu)^2$

Estimater

Har vi observeret data kan vi estimere μ og σ^2 med $\bar{x}$ og s^2. Her udskiftes de stokastiske variable X_i i $\bar{X}$ og S^2 med de observerede x_i,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

Både $\bar{X}$ og S^2 er stokastiske variable (transformationer af X_i'erne), mens $\bar{x}$ og s^2 er realisationer af disse,

$$\begin{array}{ccccc} X_1 & \dots & X_n & \bar{X} & S^2 \\ \downarrow & & \downarrow & \downarrow & \downarrow \\ x_1 & \dots & x_n & \bar{x} & s^2 \end{array}$$

Eksempel - fortsat

Fra Eksempel 1 i noterne kan vi estimere μ med $\bar{x}$ og σ^2 med s^2.

$$\bar{x} = \frac{1}{10}(164.508 + 164.509 + \cdots + 164.514 + 164.513) = 164.511 \text{ gon}$$

$$\sum_{i=1}^{10} x_i^2 = 164.508^2 + 164.509^2 + \cdots + 164.514^2 + 164.513^2 = 270638.7 \text{ gon}^2$$

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

$$= \frac{1}{10-1}\left(270638.7 - 10 \cdot (164.511)^2\right) = (0.00298)^2 \text{ gon}^2$$

Størrelsen s^2 er et mål for nøjagtigheden af vores observationer.
Jo mindre desto mere nøjagtige er vores målinger.

Approksimativt 95% konfidensinterval (erstatter ukendt σ med s):

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} = [164.509; 164.513]$$

Matlab: Stikprøvegennemsnit og -varians

Data:

```
>> x = [164.508,164.509,164.511,164.507,
            164.510,164.511,164.517,164.510,164.514,164.513];
```

Beregn stikprøvegennemsnit $\bar{x}$:

```
>> mean(x)
ans =
  164.5110
```

Beregn stikprøvevariansen s^2:

```
>> var(x)
ans =
   8.8889e-06
```

Beregn spredningen s:

```
>> std(x)
ans =
   0.0030
```

Bonus: Beregn gennemsnit af observation 4 til 7:

```
>> mean(x(4:7)) ans = 164.5113
```

Estimatorer - kendt middelværdi μ

I situationer hvor vi kender μ (fx. på en øvelsesbane hvor sande længder og vinkler er kendt) bruger vi estimatet $\hat{s}$ til at estimere målingernes nøjagtighed:

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

I disse situationer er $\hat{s}^2$ et centralt estimat for σ^2. Dvs:

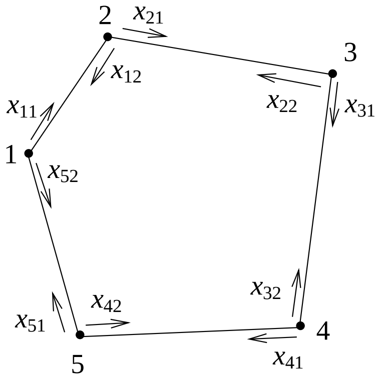
$$\mathbb{E}(\hat{S}^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left((X_i - \mu)^2\right) = \frac{1}{n}n\sigma^2 = \sigma^2.$$

Bemærk at $\mathbb{V}\mathrm{ar}(\hat{S}^2) \leq \mathbb{V}\mathrm{ar}(S^2)$, dvs. $\hat{s}^2$ er et mere nøjagtigt estimat end s^2.

(gavnligt at bruge al den viden, der er til rådighed)

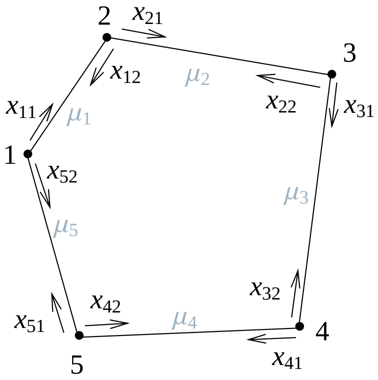
Dobbeltmålinger

Udgangspunktet er $2n$ målinger hvor målingerne to og to måler samme størrelse, altså n forskellige størrelser i alt.



Eks.: siderne i et polygon hvor hver sidelængde måles to gange.

Dobbeltmålinger



$$
\begin{array}{ccccccc}
X_{11} & X_{12} & X_{21} & X_{22} & \dots & X_{n1} & X_{n2} \\
\multicolumn{2}{c}{\downarrow} & \multicolumn{2}{c}{\downarrow} & \dots & \multicolumn{2}{c}{\downarrow} \\
x_{11} & x_{12} & x_{21} & x_{22} & \dots & x_{n1} & x_{n2}
\end{array}
$$

Uafhængige målinger af samme kvalitet, dvs. ens varians σ^2.

De n forskellige størrelser har sande værdier $\mu_1, \dots, \mu_n$, dvs:

$$\mathbb{E}(X_{11}) = \mathbb{E}(X_{12}) = \mu_1 \quad \dots \quad \mathbb{E}(X_{n1}) = \mathbb{E}(X_{n2}) = \mu_n.$$

Estimat af μ_i

Til at estimere μ_i anvendes estimatet $\bar{x}_i = \frac{1}{2}(x_{i1} + x_{i2})$.
For den tilsvarende estimator $\bar{X}_i = \frac{1}{2}(X_{i1} + X_{i2})$ gælder der:

$$\begin{aligned}\mathbb{E}(\bar{X}_i) = \mathbb{E}\left(\frac{1}{2}(X_{i1} + X_{i2})\right) &= \frac{1}{2}\left(\mathbb{E}(X_{i1}) + \mathbb{E}(X_{i2})\right) \\ &= \frac{1}{2}\left(\mu_i + \mu_i\right) = \mu_i\end{aligned}$$

Dvs. $\bar{x}_i$ er et centralt estimat af μ_i. Desuden gælder:

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\bar{X}_i) = \mathbb{V}\mathrm{ar}\left(\frac{1}{2}(X_{i1} + X_{i2})\right) &= \frac{1}{2^2}\left(\mathbb{V}\mathrm{ar}(X_{i1}) + \mathbb{V}\mathrm{ar}(X_{i2})\right) \\ &= \frac{1}{4}\left(\sigma^2 + \sigma^2\right) = \frac{\sigma^2}{2}.\end{aligned}$$

Estimat af σ^2

Idet X_{i1} og X_{i2} måler samme størrelse μ siger deres differens $Y_i = X_{i1} - X_{i2}$ noget om målekvaliteten.

Vi ved

$$\mathbb{E}(Y_i) = \mathbb{E}(X_{i1}) - \mathbb{E}(X_{i2}) = \mu_i - \mu_i = 0$$

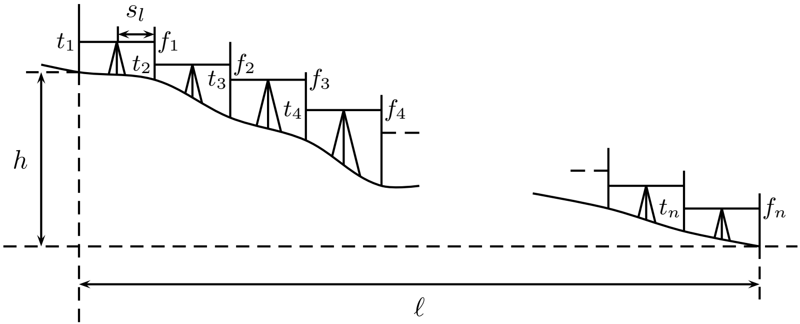
$$\mathbb{V}\mathrm{ar}(Y_i) = \mathbb{V}\mathrm{ar}(X_{i1}) + \mathbb{V}\mathrm{ar}(X_{i2}) = \sigma^2 + \sigma^2 = 2\sigma^2.$$

Dvs. vi kender den *sande* middelværdi af Y_i, $\mathbb{E}(Y_i) = \mu_Y = 0$. Derfor er

$$\hat{s}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_Y)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - x_{i2})^2$$

et centralt estimat af $2\sigma^2$, hvorved $\hat{s}_Y^2/2$ er et estimat for σ^2 (variansen på den enkelte måling).

Geometrisk nivellement



t_i : stadieaflæsning ved tilbagesigte
f_i : stadieaflæsning ved fremsigte

$$h = (t_1 - f_1) + (t_2 - f_2) + \cdots + (t_n - f_n) = \sum_{i=1}^{n} t_i - f_i$$

Totallængden ℓ er opdelt i $2n$ stykker af længde s, dvs. $\ell = 2ns_l$.

Modellen

Vi antager af t_i og f_i er realisationer af uafhængige stokatiske variable T_i og F_i med samme varians σ_a^2, $i = 1, \ldots, n$. Denne antagelse kan begrundes med at sigteafstanden er fast og den samme for alle observationer.

Ydermere bliver h således en realisation af den stokastiske variabel

$$H = \sum_{i=1}^{n} (T_i - F_i).$$

Dvs:

$$\begin{array}{cccccc} T_1 & F_1 & \ldots & T_n & F_n & H \\ \downarrow & \downarrow & & \downarrow & \downarrow & \downarrow \\ t_1 & f_1 & \ldots & t_n & f_n & h \end{array}$$

Variansen af højdemålingen

Variansen af H bestemmes ved

$$\begin{aligned}\mathbb{V}\mathrm{ar}(H) = \mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{n}(T_i - F_i)\right) &= \sum_{i=1}^{n}\left[\mathbb{V}\mathrm{ar}(T_i) + \mathbb{V}\mathrm{ar}(F_i)\right] \\ &= \sum_{i=1}^{n}(\sigma_a^2 + \sigma_a^2) \\ &= 2n\sigma_a^2\end{aligned}$$

Fra tidligere slide ved vi, at $2n = \ell/s_l$. Der gælder derfor

$$\sigma_\ell^2 = \mathbb{V}\mathrm{ar}(H) = \ell\sigma_a^2/s_l.$$

Erfaringer viser at $\sigma_a/\sqrt{s_l}$ kun i ringe grad afhænger af s_l når $s_l < 100$ m.

Således indføres **kilometerspredningen** $\sigma_k = \sigma_a/\sqrt{s_l}$, dvs. $\sigma_a = \sigma_k\sqrt{s}_l$.

Heraf følger

$$\sigma_\ell^2 = \mathbb{V}\mathrm{ar}(H) = \ell\sigma_a^2/s_l = \ell\sigma_k^2$$

Konfidensinterval for højden

Spredningen på et geometrisk nivellement over længden ℓ er således $\sigma_\ell = \sqrt{\ell}\sigma_k$.

Et 95% konfidens interval for den ukendte højde er

$$H \pm 1.96\sqrt{\ell}\sigma_k$$

(jf. tidligere slide om konfidensinterval i generelt set-up)

Eksempel: $h = 12000\text{mm}$, $\sigma_k = 3\text{mm}/\sqrt{\text{km}}$ og $\ell = 4\text{km}$. (Realiseret) konfidensinterval

$$12000 \pm 1.96\sqrt{4\text{km}}3\text{mm}/\sqrt{\text{km}} = [11988; 12012]$$

Lineær transformation (repetition)

Sætning: Lineær transformation af SV

Hvis X er en stokastisk variabel, og $a,b \in \mathbb{R}$, så gælder

$$\begin{aligned} \mathbb{E}(aX+b) &= a\mathbb{E}(X)+b, \\ \mathbb{V}\mathrm{ar}(aX+b) &= a^2\mathbb{V}\mathrm{ar}(X). \end{aligned}$$

Spørgsmål:

- Hvordan håndterer vi ikke-lineære transformationer ?

Vilkårlig transformation af X

Lad X være en stokastisk variabel med

$$\mathbb{E}(X) = \mu \text{ og } \mathbb{V}\mathrm{ar}(X) = \sigma^2.$$

samt tæthedsfunktion $f(x)$.

$Y = g(X)$: vilkårlig **differentiabel transformation** af X.

Middelværdien af $Y = g(X)$:

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Problem: Middelværdien $\mathbb{E}(Y)$ er ofte vanskelig at beregne.
Løsning: Vi lineariserer transformationen $g(X)$.

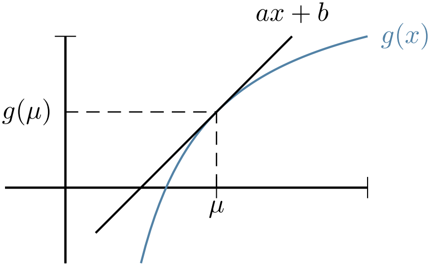
Eksempel på transformationer i landmåling: trigonometriske funktioner, afstandsformel.

Linearisering

Lineær approximation af g omkring μ:

$$
\begin{aligned}
Y = g(X) &\approx g(\mu) + g'(\mu)(X - \mu) \\
&= g'(\mu)X - g'(\mu)\mu + g(\mu) \\
&= aX + b,
\end{aligned}
$$

hvor $a = g'(\mu)$ og $b = -g'(\mu)\mu + g(\mu)$.



Linearisering

Vi har en approksimation af $g(x)$:

$$Y \approx aX + b,$$

hvor $a = g'(\mu)$ og $b = -g'(\mu)\mu + g(\mu)$.

Heraf følger approksimativ middelværdi og varians for Y:

$$\begin{aligned}\mathbb{E}(Y) &\approx a\mathbb{E}(X) + b \\ &= g'(\mu)\mu - g'(\mu)\mu + g(\mu) \\ &= g(\mu)\end{aligned}$$

$$\begin{aligned}\mathbb{V}\mathrm{ar}(Y) &\approx a^2\mathbb{V}\mathrm{ar}(X) \\ &= g'(\mu)^2\sigma^2,\end{aligned}$$

hvor approximationerne er gode, hvis σ er lille.

Linearisering: Eksempel

Antag $X \sim \mathcal{N}(\mu,\sigma^2)$ og $Y = g(X) = \exp(X)$.

En linearisering af $\exp(x)$ omkring $x = \mu$ giver:

- $g'(x) = \frac{d}{dx}\exp(x) = \exp(x)$
- $g(x) \approx g(\mu) + g'(\mu)(x-\mu) = \exp(\mu) + \exp(\mu)(x-\mu)$.

Heraf følger:

- $\mathbb{E}(Y) \approx g(\mu) = \exp(\mu)$
- $\mathbb{V}\mathrm{ar}(Y) \approx g'(\mu)^2\sigma^2 = (\exp(\mu))^2\,\sigma^2 = \exp(2\mu)\sigma^2$

Vi har derfor, at Y er tilnærmet normalfordelt, med middelværdi $\exp(\mu)$ og varians $\exp(\mu)\sigma^2$:

$$Y \approx \mathcal{N}(\exp(\mu), \exp(2\mu)\sigma^2).$$

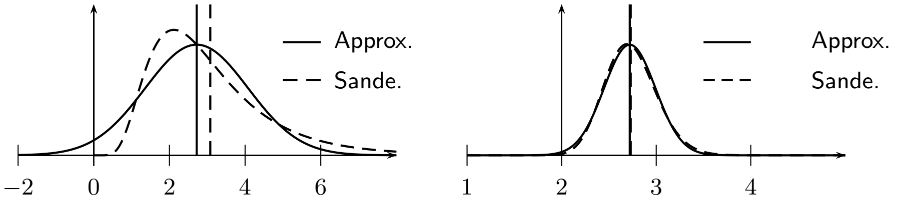
Linearisering: Eksempel (forts.)

Antag (igen) $X \sim \mathcal{N}(\mu,\sigma^2)$ og $Y = g(X) = \exp(X)$.

To eksempler, hvor $\mu = 1$, og $\sigma = 0.5$ (venstre) og $\sigma = 0.1$ (højre).

$X \sim \mathcal{N}(1; 0{,}5^2)$
$Y \approx \mathcal{N}(\exp(1); \exp(2)0{,}5^2)$

$X \sim \mathcal{N}(1; 0{,}1^2)$
$Y \approx \mathcal{N}(\exp(1); \exp(2)0{,}1^2)$



Til venstre er den relative varians for X $\sigma^2/\mu^2 = 0{,}25$ og til højre er den relative varians for X $0{,}01$. Jo mindre relativ varians jo bedre er approksimationen.

Linearisering — estimation af transformeret størrelse

Antag vi vil estimere $\theta = h(\mu)$ hvor vi kan estimere μ vha. $\bar{X}$ baseret på en stikprøve $X_1, \dots, X_n$ af uafhængige stokastiske variable med middelværdi μ og varians σ^2.

Da er vores estimat

$$\hat{\theta} = h(\bar{X})$$

Pr. linearisering og central grænseværdisætning har vi

$$\hat{\theta} \approx N(\theta, (h'(\mu))^2 \sigma^2 / n)$$

Dermed er et approksimativ 95% konfidensinterval givet ved

$$\hat{\theta} \pm 1.96 |h'(\bar{X})| \sigma / \sqrt{n}$$

I praksis er σ^2 ofte ukendt og erstattes af estimatet s^2 baseret på $X_1, \dots, X_n$.

Trigonometriske funktioner: Gon og radianer

Lad $\sin_r(x)$ og $\sin(x)$ betegne sinus når vinklen x er målt i hhv. radianer og gon. Tilsvarende for cosinus og tangens.

Vi har

$$\begin{aligned}\sin(C) &= \sin_r\left(\frac{2\pi}{400\ \text{gon}}C\right)\\ &= \sin_r\left(\frac{1}{\omega}C\right),\end{aligned}$$

hvor

$$\omega = \frac{200\ \text{gon}}{\pi},$$

er en konverterings-faktor.

Trigonometriske funktioner: Differentitation

Vi har regneregler for differentiation af trigonmetriske funktioner, når vinklen er målt i radianer. Fx.

$$\frac{\mathrm{d}\sin_r(x)}{\mathrm{d}x} = \cos_r(x).$$

Når vinklen er målt i gon får vi:

$$\frac{\mathrm{d}\sin(x)}{\mathrm{d}x} = \frac{\mathrm{d}\sin_r\left(\frac{1}{\omega}x\right)}{\mathrm{d}x} = \cos_r\left(\frac{1}{\omega}x\right)\frac{1}{\omega} = \cos(x)\frac{1}{\omega}.$$

Konverterings-faktoren $\frac{1}{\omega} = \pi/200$ gon optræder på samme måde ved differentiation af cosinus og tangens:

$$\frac{\mathrm{d}\cos(x)}{\mathrm{d}x} = -\sin(x)\frac{1}{\omega} \quad \text{og} \quad \frac{\mathrm{d}\tan(x)}{\mathrm{d}x} = \left(1+\tan(x)^2\right)\frac{1}{\omega}$$