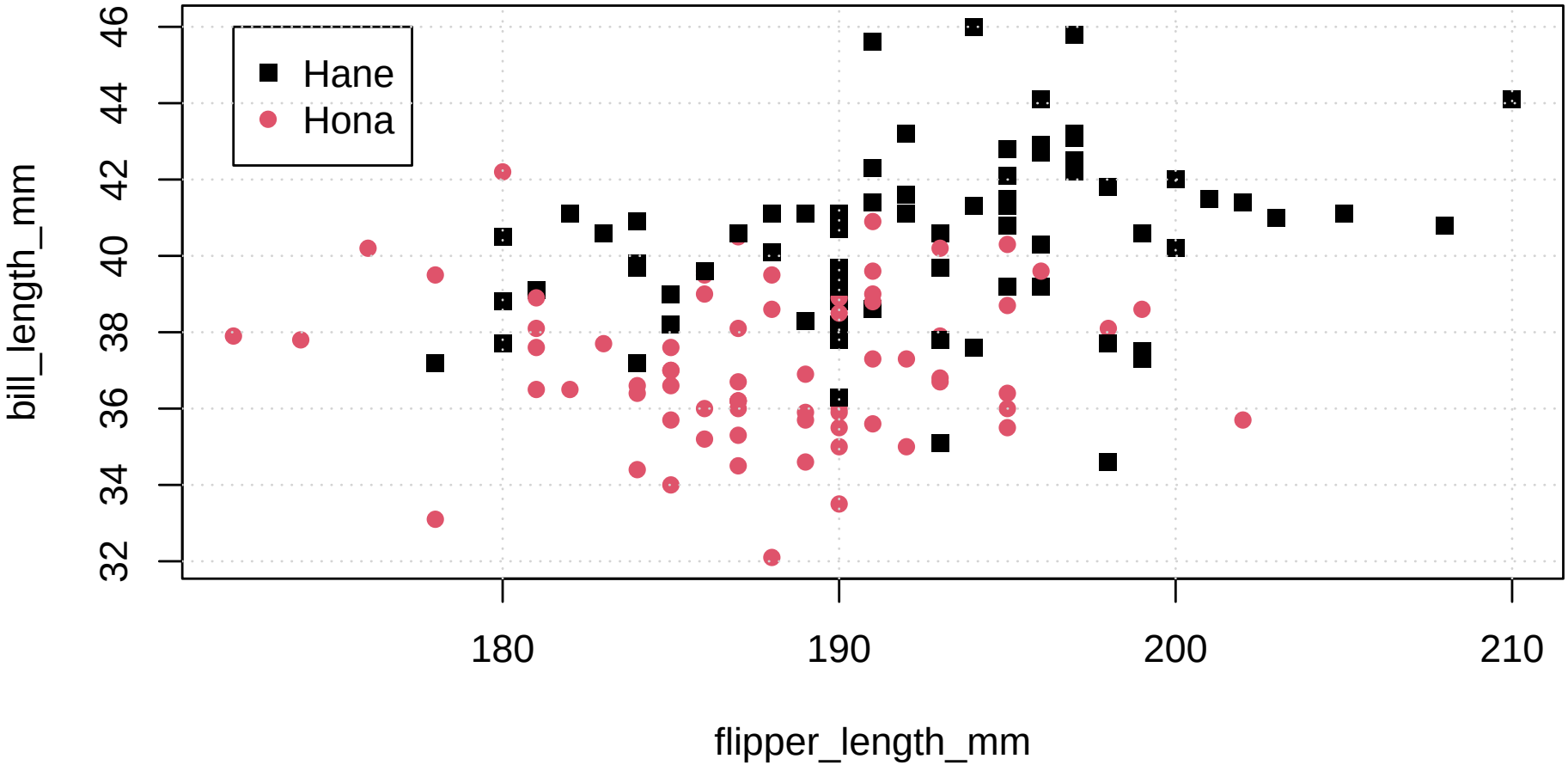
Statistik för Biologer
F3: Hypotesprövning och Konfidensintervall

Shaobo Jin

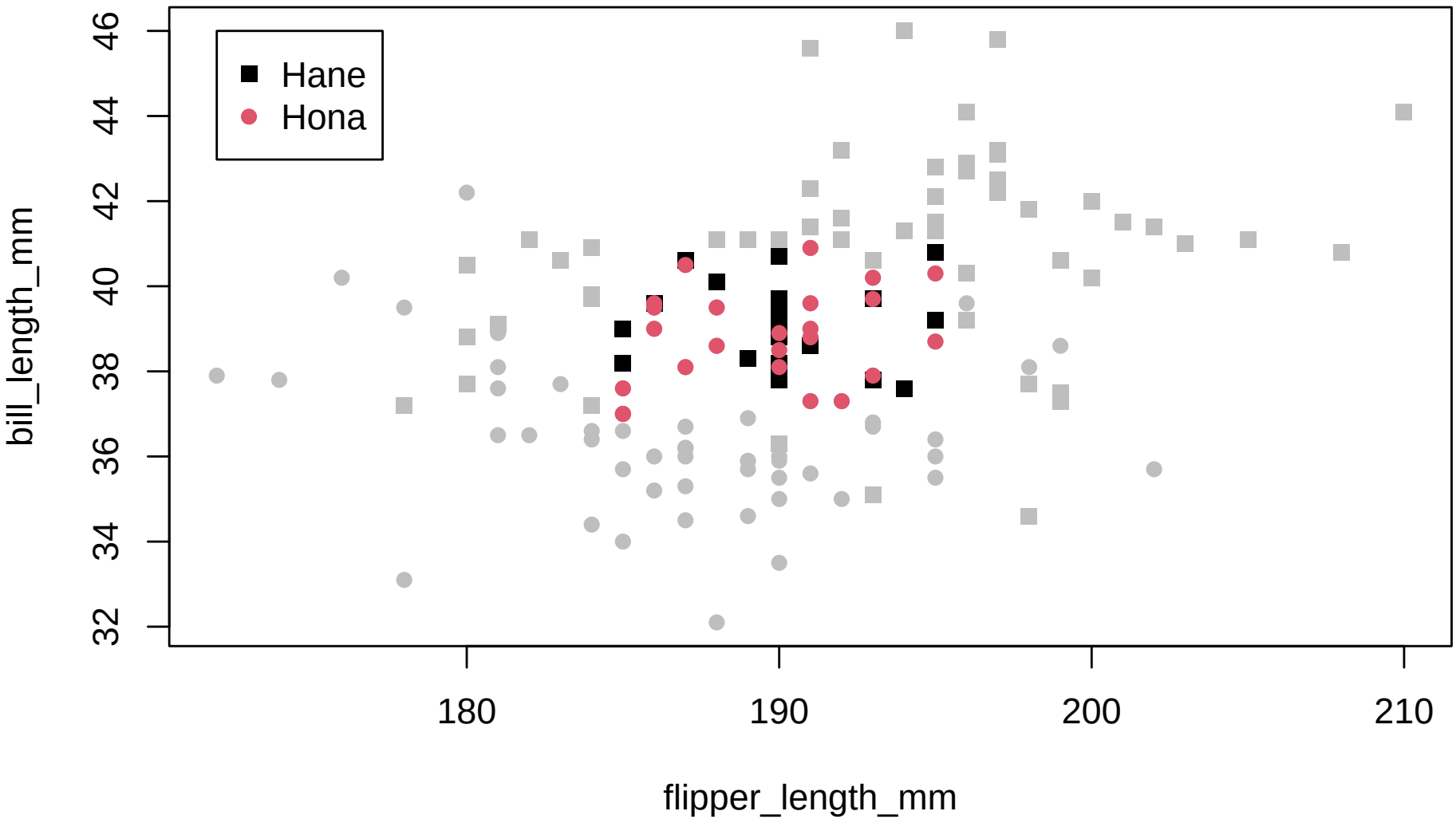
Matematiska institutionen

Kunde det blivit annorlunda?

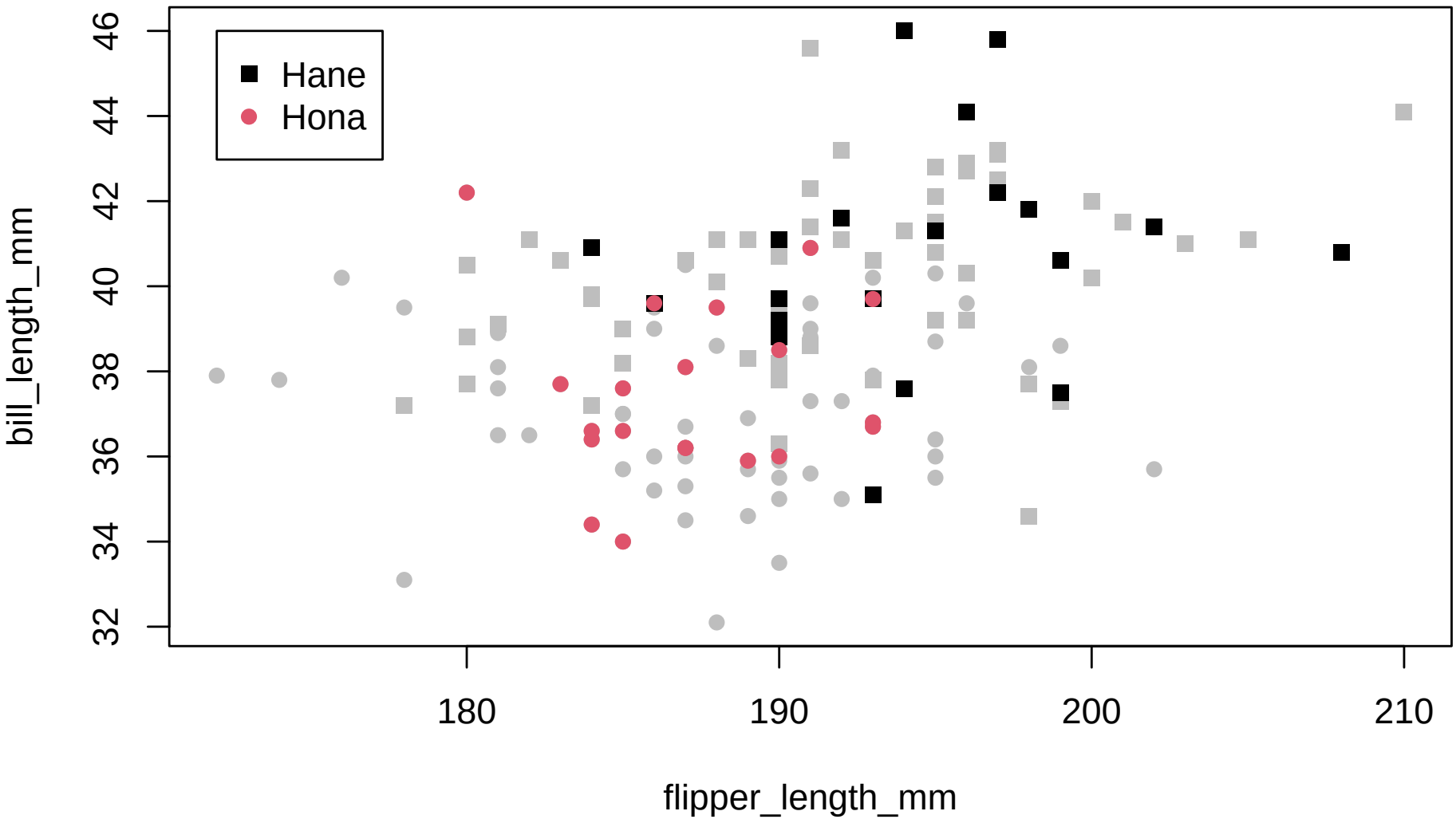
Låt oss anta att det finns totalt 152 Adeliepingviner. Men vi studerar bara 20 hanar och 20 honor. Finns storleksskillnader mellan olika kön?



Första Urvalet



Andra Urvalet



Vad Ska Vi Göra?

Hur ska vi kunna känna oss säkra på att den skillnad vi tycker oss se beror på biologi och inte på slumpen?

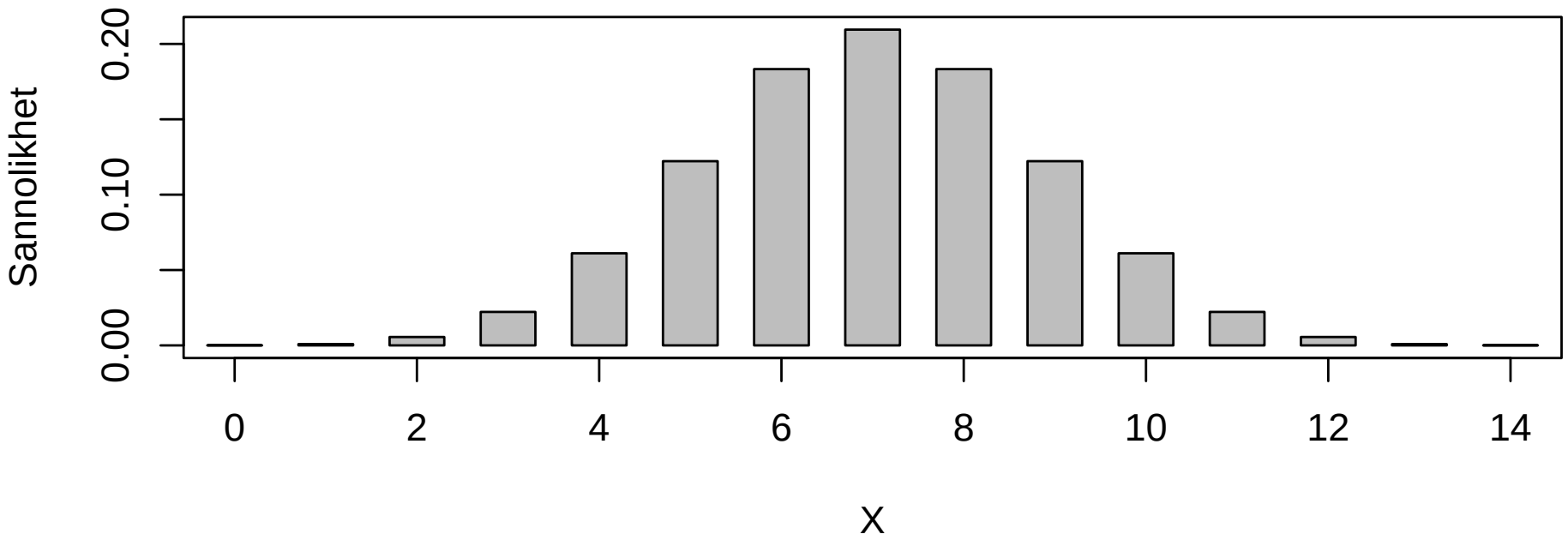
1. Idé för att få statistiskt säkerställda resultat är hypotesprövning
2. Idé för att beskriva hur stor skillnaden är konfidensintervall

Bläckfisken Paul

Vi antar att

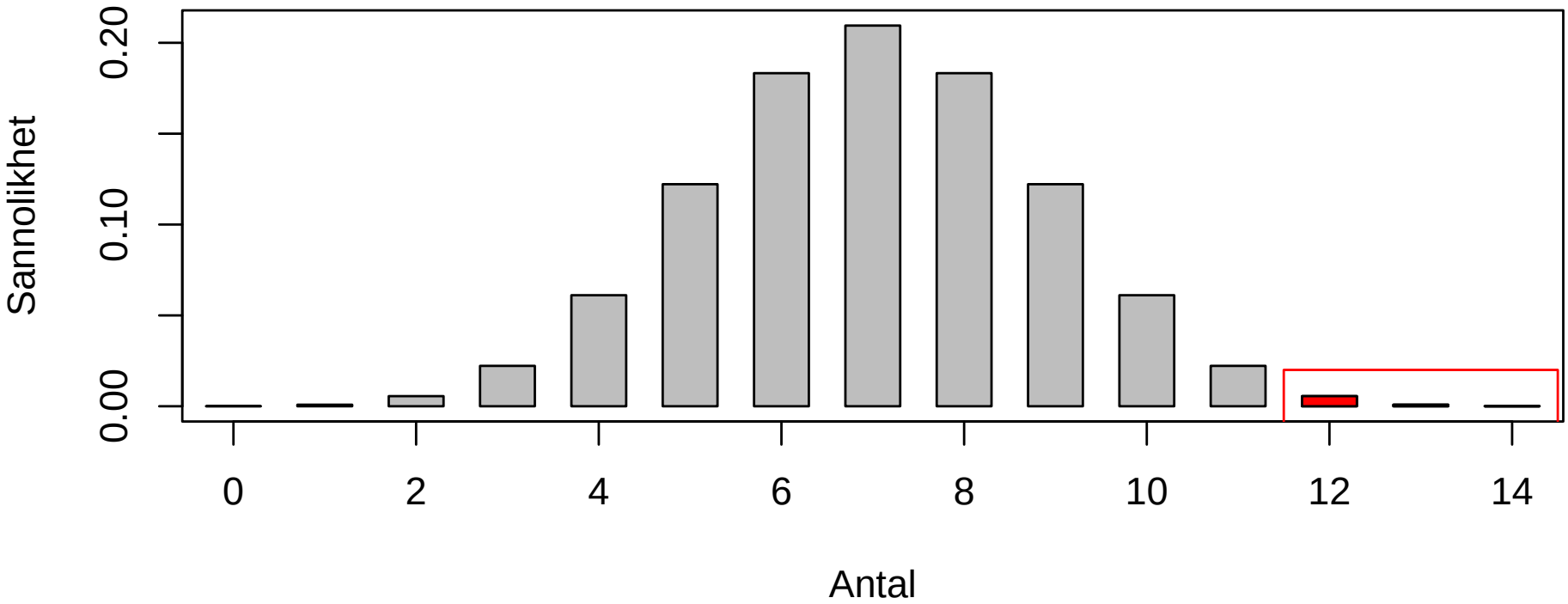
1. Bläckfisken Paul tippar $n = 14$ matcher
2. Varje gånger är sannolikheten att tippa rätt vinnare $p = 0.5$
3. Varje ny tippning är oberoende av tidigare tippningar

Antalet rätt X som Paul får är $\text{Bin}(14, 0.5)$.



Ett Minst Lika Extremt Resultat

- Någon påstår att sannolikheten att tippar rätt vinnare är större än 0.5.
- Ju mer Pauls resultat avviker från vad man skulle förvänta sig, desto starkare belägg för att $p > 0.5$.
- Våra data visar att Paul tippade 12 av 14 rätt. Så 12 eller fler rätt är "minst lika extremt" som Pauls resultat:



Hur "Extremt" Var Pauls Resultat?

Antag att $X \sim \text{Bin}(14, 0.5)$. Sannolikheten att vara minst lika extremt som Pauls resultat är

$$P(X \geq 12) = P(X = 12) + P(X = 13) + P(X = 14).$$

```
dbinom(12, 14, 0.5) + dbinom(13, 14, 0.5) +
    dbinom(14, 14, 0.5)
```

```
## [1] 0.006469727
```

Eller

```
1 - pbinom(11, 14, 0.5)
```

```
## [1] 0.006469727
```

Terminologi

I **statistisk hypotesprövning** utvärderar vi **nollhypotes** H_0: en hypotes som ska motbevisas. Till exempel,

- Varje gång är sannolikheten att tippa rätt vinnare $p = 0.5$
- Det finns ingen skillnad i vikt mellan honor och hanar
- Pingviner är längre än 195mm

Om nollhypotesen inte stämmer så tror vi istället på **alternativhypotesen** H_1 (eller H_A)

- Varje gång är sannolikheten att tippa rätt vinnare inte $p = 0.5$
- Det finns en skillnad i vikt mellan honor och hanar
- Pingviner är inte längre än 195mm

p-Värdet

För att utvärdera nollhypotesen H_0 brukar vi beräkna **p-värdet** - sannolikheten att få ett resultat som är minst lika extremt som det observerade, om H_0 är sann.

- H_0: Paul tippar rätt med $p \leq 0.5$ mot H_1: $p > 0.5$.
  Ju fler rätt Paul tippar, desto extremare är resultatet.
- H_0: Paul tippar rätt med $p = 0.5$ mot H_1: $p \neq 0.5$.
  Ju fler/mindre rätt Paul tippar, desto extremare är resultatet.
- H_0: det finns ingen skillnad i vikt mellan honor och hanar
  H_1: det finns en skillnad i vikt.
  Ju större den observerade skillnaden mellan grupperna är, desto extremare är resultatet.

Hur Använder Vi p-Värdet?

Om p-värdet är mindre än en bestämd gräns (ofta 0.05 eller 0.01) säger vi att resultatet är **signfikant** eller (mer sällan) **statistiskt säkerställt**. Vi anser då att vi har statistiska belägg för att H_0 inte stämmer och säger att vi förkastar H_0.

Exempel: Paul

Sannolikheten att någon lyckas tippa minst lika bra som Paul gjorde, är mindre än 0.01. Slutsatsen är att sannolikheten att tippa rätt vinnare varje gånger är högre än 0.5.

Interleukin 6

Interleukin 6 (IL-6) är ett protein som är inblandat i bland annat många autoimmuna sjukdomar, cancer och depression. I en klinisk studie mättes uttrycksnivåerna av IL-6 i plasma hos en grupp patienter före och efter behandling. Nollhypotesen H_0 var att behandlingen inte påverkar nivåerna av IL-6.

Vi kan tänka oss olika sorters alternativhypoteser här:

- Behandlingen förändrar nivåerna av IL-6 (**dubbelsidig hypotes**)
- Behandlingen sänker nivåerna av IL-6 (**enkelsidig hypotes**)
- Behandlingen ökar nivåerna (**enkelsidig hypotes**)

Vilken alternativhypotes som används bestäms innan vi tittar på data!

Våra Data

x (före)	y (efter)
9.103086	**9.417651**
10.184849	9.981753
11.587845	8.607305
8.869624	7.960331
9.919748	**10.782229**
10.132420	6.688931
10.707955	9.878605
9.760302	9.035807
11.984474	10.012829
9.861213	9.432265

2 patienter hade högre IL-6 efter behandlingen, medan övriga 8 hade lägre nivå efter behandlingen.

Statistisk Modell: Teckentestet

H_0 : behandling påverkar inte IL-6-nivåerna

- Om H_0 är sann så är det lika sannolikt att $y_i < x_i$ som att $y_i > x_i$.
- Vi upprepar försöket "se om IL-6 har en lägre uttrycksnivå efter behandling" 10 oberoende gånger.
- Varje gång kontrollerar vi om uttrycksnivån är lägre efter behandling
  - om $y_i - x_i$ är positivt eller negativt
- Låt U =antal gånger som uttrycksnivån är lägre efter behandling.
- Under H_0 gäller att U är binomialfördelat med parametrar $n = 10$ och $p = 0.5$.

Att Beräkna p-Värdet: Enkelsidig Hypotes

Hur extremt var resultatet? Vad som menas med "extremt" bestäms av alternativhypotesen! Låt

$$U \quad = \quad \text{antal gånger som uttrycksnivån är lägre efter behandling.}$$

- H_1: Behandlingen sänker nivåerna av IL-6 (**enkelsidig hypotes**)
- Ovanligt höga värden på U tyder på att H_0 är fel.
- p-värdet blir:

$$P(U \geq 8) \quad = \quad P(U = 8) + P(U = 9) + P(U = 10).$$

```
dbinom(8, 10, 0.5) + dbinom(9, 10, 0.5) +
    dbinom(10, 10, 0.5)
## [1] 0.0546875
```

- p-värdet är större än 0.05, så vi kan inte förkasta H_0.

Att Beräkna p-Värdet: Dubbelsidig Hypotes

Hur extremt var resultatet? Vad som menas med "extremt" bestäms av alternativhypotesen! Låt

$$U \quad = \quad \text{antal gånger som uttrycksnivån är lägre efter behandling.}$$

- H_1: Behandlingen förändrar nivåerna av IL-6 (**dubbelsidig hypotes**)
- Ovanligt låga eller höga värden på U tyder på att H_0 är fel.
- I det här fallet brukar p-värdet beräknas som 2 gånger p-värdet för den enkelsidiga alternativhypotes som stämmer bäst med data, dvs

$$2 \cdot P\left(U \geq 8\right).$$

```
2.0 * (dbinom(8, 10, 0.5) + dbinom(9, 10, 0.5) +
       dbinom(10, 10, 0.5))
## [1] 0.109375
```

Nackdelar Med Teckentestet

Teckentestet tar inte hänsyn till hur stora skillnaderna mellan mätningarna är.

- Vi utnyttjar inte all information vi har i våra data!
- Vi använder inte själva värden utan tecken!
  - $x = 12$ och $y = 6$ ledar till ett negativt $y - x$
  - $x = 6.1$ och $y = 6$ ledar också till ett negativt $y - x$
  - Är det rimligt att de behandlas lika?

Egenskaper hos ett test

	Sanning	
Vår slutsats	H_0 är rätt	H_1 är rätt
H_0 är rätt		**fel**
H_1 är rätt	**fel**	

Vi vill att vårt test:

1. Inte ska få oss att felaktigt förkasta H_0.
   1. **signifikansnivån** α - sannolikheten att förkasta H_0 om H_0 är sann.
   2. α är den gränsen som p-värdet ska ligga under för att resultatet ska vara signifikant
2. Ska utnyttja informationen i våra data så bra som möjligt, så att vi kan förkasta H_0 om den inte stämmer.
   1. Mäts genom testets **styrka** - sannolikheten att förkasta H_0 om H_1 är sann.

Paul Igen

Utfall	p-värde
0	1.000
1	1.000
2	0.999
3	0.994
4	0.971
5	0.910
6	0.788
7	0.605
8	0.395
9	0.212
10	0.090
11	**0.029**
12	0.006
13	0.001
14	0.000

- α är den gränsen som p-värdet ska ligga under för att resultatet ska vara signifikant.
- Vid $\alpha = 0.05$ förkastas H_0 om Paul får minst 11 rätt.

Styrka för Testet

Styrkan hos ett statistiskt hypotestest är sannolikheten att förkasta H_0 om H_1 är sann. I Pauls fall förkastas H_0: $p = 0.5$ om antalet rätt X är minst 11.

Om sanningen är $p = 0.8$ blir testets styrka

$$P(X \geq 11) \quad = \quad P(X = 11) + P(X = 12) + P(X = 13) + P(X = 14),$$

då $X \sim \text{Bin}(14, 0.8)$.

```
dbinom(11, 14, 0.8) + dbinom(12, 14, 0.8) +
    dbinom(13, 14, 0.8) + dbinom(14, 14, 0.8)
```

```
## [1] 0.6981899
```

Typ I-Fel och Typ II-Fel

	Sanning	
Vår slutsats	H_0 är rätt	H_1 är rätt
H_0 är rätt		**Typ II-fel**
H_1 är rätt	**Typ I-fel**	

1. Att felaktigt förkasta H_0 trots att H_0 är sann kallas för ett **typ I-fel** (ett falskt positivt resultat)
   - Låg signifikansnivå α ger låg risk för typ I-fel
2. Att felaktigt fortsätta tro på H_0 trots att H_1 är sann kallas för ett **typ II-fel** (ett falskt negativt resultat)
   - Hög styrka ger låg risk för typ II-fel

När vi minskar α så minskar vi också styrkan

- Risken för typ I-fel blir lägre, men risken för typ II-fel blir högre
- Vi väljer en acceptabel risk för typ I-fel, typiskt $\alpha = 0.05$ eller $\alpha = 0.01$.

Att Få Hög Styrka

Flera saker påverkar testets styrka:

- Ju starkare effekten är, desto större blir testets styrka
  - Om Paul har 60% chans att tippa rätt blir styrkan 12%
  - Om Paul har 90% chans att tippa rätt blir styrkan 96%
- Större stickprov ger normalt högre styrka
  - Om stickprovsstorleken är låg kan styrkan bli så låg att det inte är värt att genomföra studien!
- Vid jämförelse av två (eller flera) grupper får vi ofta högre styrka om grupperna är balanserade, dvs. har samma antal observationer
- Olika test kan ha olika hög styrka i olika situationer
  - **t-testet** (Föreläsning 4) har ofta högre styrka än teckentestet eftersom det använder mer information från våra data
- Bra försöksplanering är viktigt för att få ett test med hög styrka!

Signifikant och Betydelsefull Effekten

Att resultatet av ett hypotestest är **signifikant** betyder inte att den upptäckta effekten är stor att vara betydelsefull. Att enbart titta på p-värden räcker inte - vi måste också titta på hur stor effekten är!

Vara Betydelsefull?

I en studie fann man att träd med en mutation i genen LfMYB113 fällde sina löv tidigare på hösten ($p = 0.002$). Skillnaden mot träd utan mutationen var dock bara 0.4 dagar.

Konfidensintervall

Definition

Ett **konfidensintervall** för en okänd parameter θ med **konfidensgrad** $1-\alpha$ är ett intervall

1. vars gränser beräknas utifrån data: $A(X_1, X_2, ..., X_n)$ och $B(X_1, X_2, ..., X_n)$ är funktioner av data.
2. som med sannolikhet $1-\alpha$ kommer att täcka det sanna värden på θ:

$$P[A(X_1, X_2, ..., X_n) < \theta < B(X_1, X_2, ..., X_n)] = 1-\alpha.$$

Intervallet

$$A(X_1, X_2, ..., X_n) < \theta < B(X_1, X_2, ..., X_n)$$

säges vara ett $1-\alpha$ **konfidensintervall** för parametern θ.

Konfidensintervall för Pauls Förmåga

- Vi vill veta vad sannolikheten p att Paul lyckades tippa rätt vinnare i en fotbollsmatch var.
- Vår bästa **skattningen** är $\hat{p} = 12/14 \approx 0.86$.
- Ett 95% konfidensintervall ($\alpha = 0.05$) för p är $(0.60, 0, 96)$.
- Tolkning: utifrån våra data bedömer vi att Pauls sannolikhet att tippa rätt låg mellan 60% och 90%.
- Tolkning av konfidensgrad: metoden som vi använt för att beräkna intervallet prickar rätt i 95% av alla studier. I 5% av alla studier kommer den att missa det sanna värdet på parametern.
- Vi kommer att återkomma till hur vi kan beräkna konfidensintervall.

Differens

Vi tar ut åsnepingvinernas vikt från våra pingvindata:

```
library(palmerpenguins)
gentoo <- subset(penguins, species == "Gentoo")
```

Medelvärdet i våra data beror på mätskalan (na.rm tar bort saknade värden)

```
mean(gentoo$body_mass_g, na.rm = TRUE)    # gram

## [1] 5076.016

mean(gentoo$body_mass_g/1000, na.rm = TRUE) # kg

## [1] 5.076016
```

Ändra Enhet: Variansen

Variansen mäter spridning och skalas om när vi ändrar enhet:

```
# Varians och standardavvikelse i gram:
var(gentoo$body_mass_g, na.rm = TRUE)

## [1] 254133.2
```

```
# Varians och standardavvikelse i kg:
var(gentoo$body_mass_g / 1000, na.rm = TRUE)

## [1] 0.2541332
```

Räkneregel:

$$s^2\left(a \cdot x_1, ..., a \cdot x_n\right) \quad = \quad a^2 s^2\left(x_1, ..., x_n\right),$$

Ändra Enhet: Standardavvikelsen

Standardavvikelsen skalas om också när vi ändrar enhet:

```
# Varians och standardavvikelse i gram:
sd(gentoo$body_mass_g, na.rm = TRUE)

## [1] 504.1162
```

```
# Varians och standardavvikelse i kg:
sd(gentoo$body_mass_g / 1000, na.rm = TRUE)

## [1] 0.5041162
```

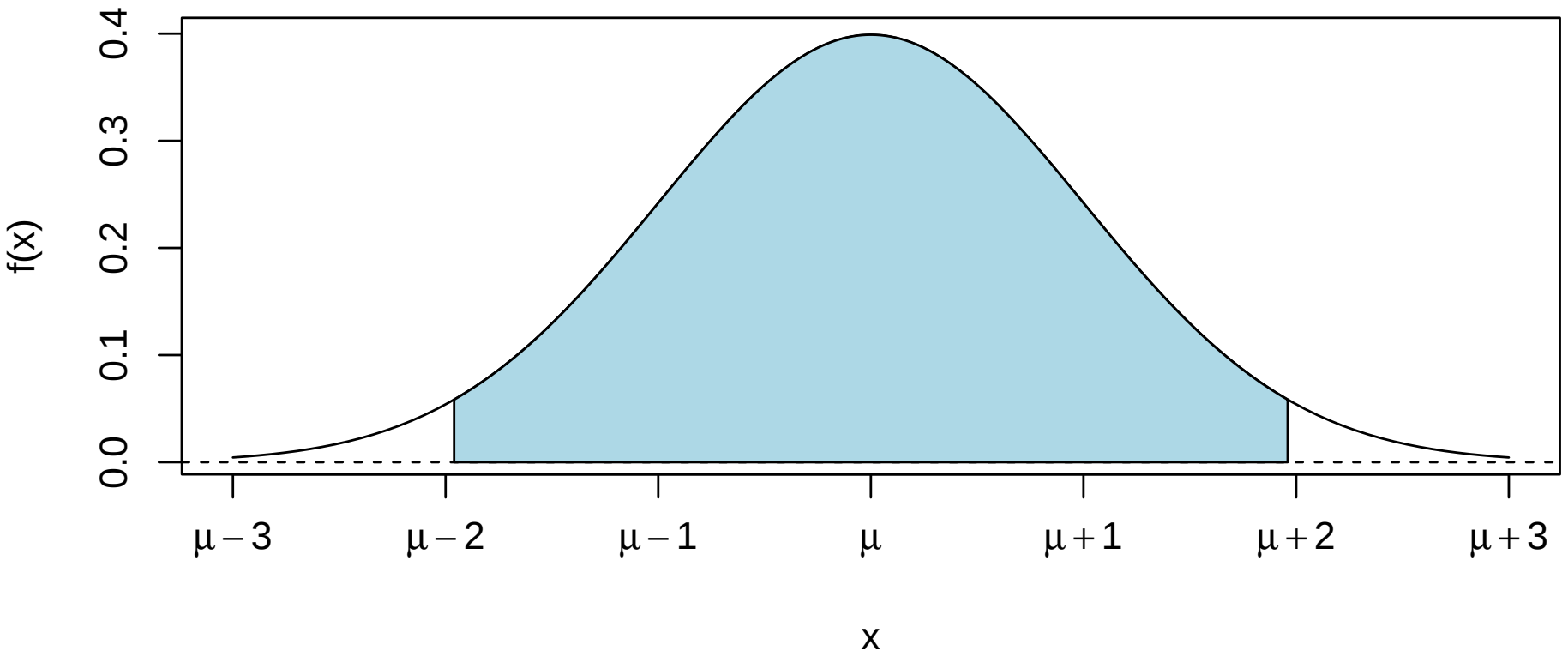
Räkneregel:

$$s\left(a \cdot x_1, ..., a \cdot x_n\right) \quad = \quad a \times s\left(x_1, ..., x_n\right).$$

Engenskap Hos Normalfördelning

För normalfördelningen gäller att 95% av alla observationer hamnar inom två standardavvikelser från väntevärdet:

$$\begin{aligned} P\left(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma\right) &\approx 0.95 \\ P\left(-1.96\sigma \leq X - \mu \leq 1.96\sigma\right) &\approx 0.95. \end{aligned}$$



Centrala Gränsvärdessatsen

Om X_1, X_2, ..., X_n är:

1. oberoende slumpvariabler
2. som alla har samma fördelning, med $E(X_i) = \mu$ och $V(X_i) = \sigma^2$,

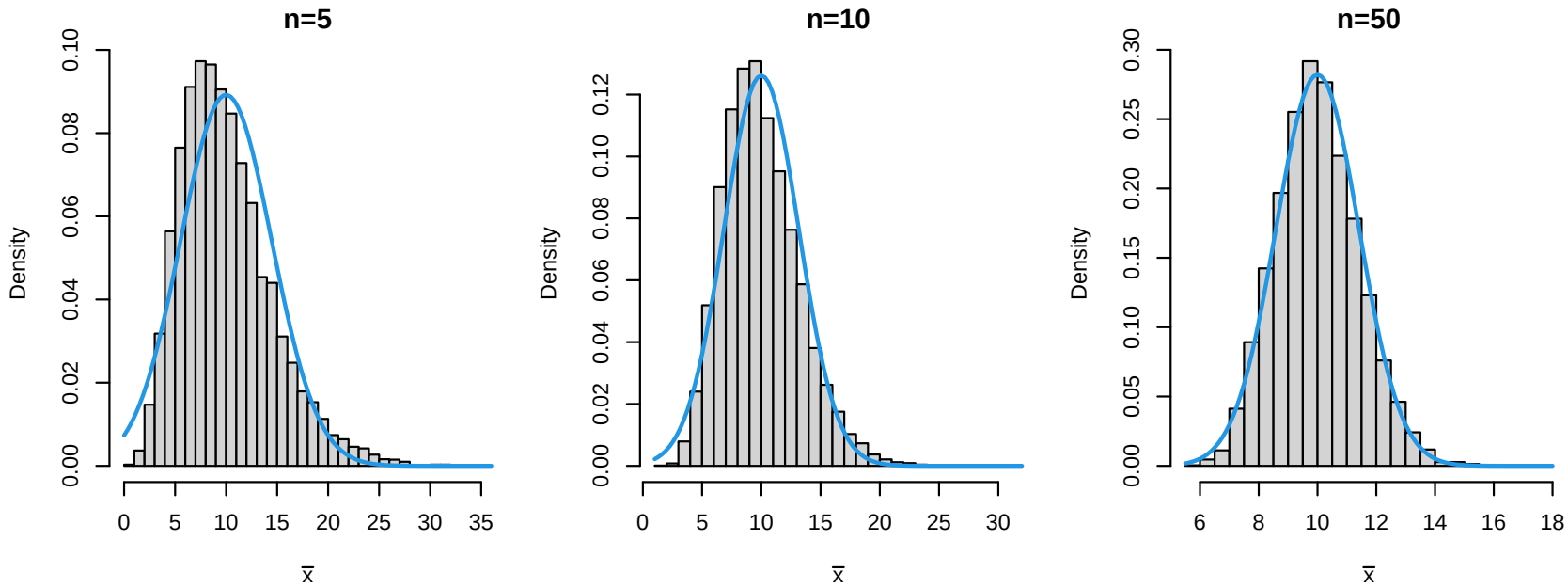
så gäller att medelvärdet $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ är approximativt normalfördelat med parametrarna μ och σ^2/n när n är tillräckligt stort.

Om n är tillräckligt stort så kommer alltså medelvärdet att ligga i spannet $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$ i ca 95% av alla studier!

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

Centrala Gränsvärdessatsen

X_1, X_2, ..., X_n är oberoende slumpvariabler som följer Exp (0.1).
Histogram av $\bar{X}$ är



Sammanfattning

1. Hypotesprövning:
   1. nollhypotes H_0 och alternativhypotes H_1.
   2. p-värdet kan används för att testa H_0 mot H_1. Ett lågt p-värde betyder att vi har statistiska belägg mot H_0.
   3. Vi väljer signifikansnivån α
   4. Testets styrka mäter hur sannolikt det är att vi lyckas förkasta H_0 om H_1 är sann
2. Konfidensintervall kvantifierar osäkerheten i vår skattning av en parameter