

Hypotesetests, fejltyper og p-værdier

Søren Højsgaard
Institut for Matematiske Fag, Aalborg Universitet

October 25, 2018

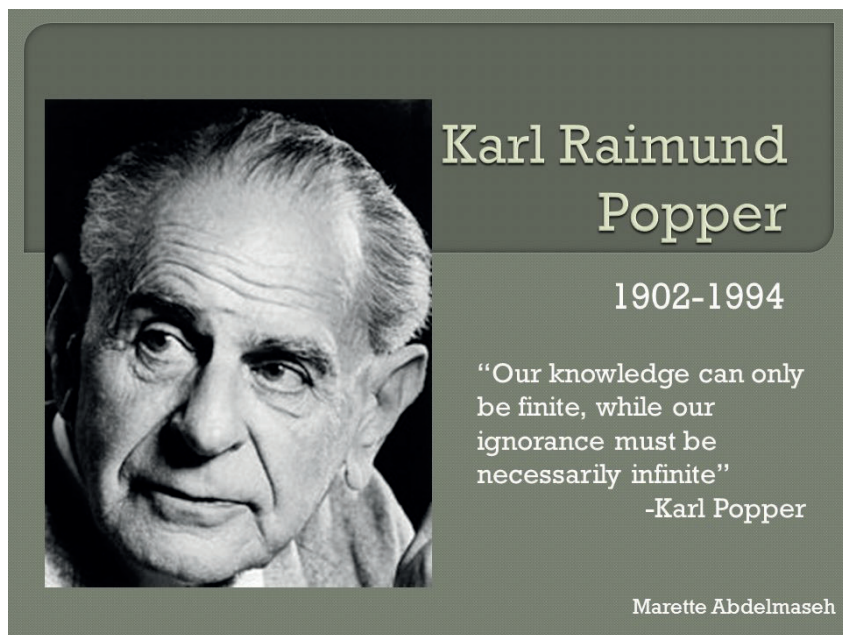
Statistisk test

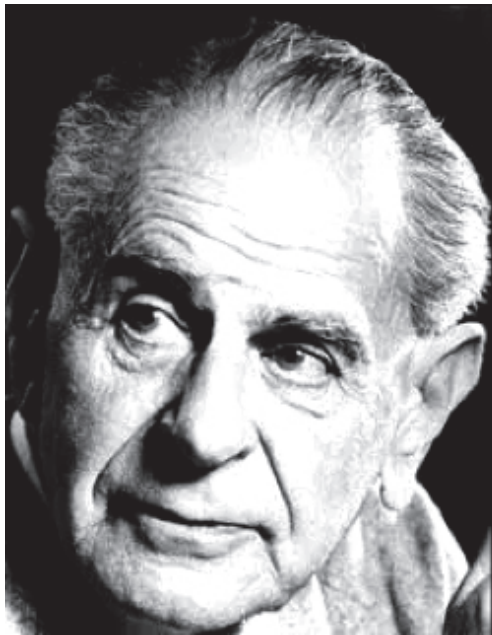
- Et statistisk test er en konfrontation af virkeligheden (data) med en teori (model).
- Laves med det formål at forsøge at falsificere modellen.
- Alternativt: Man bruger data til at “bevise” at man ikke har ret.
- NB: I statistik hedder det ET test; ikke EN test!

Karl Popper (1902-1994)

Passer ind Karl Poppers (1902-1994) videnskabsteori:

- Man kan ikke empirisk verificere videnskabelige teorier; kun falsificere dem.
- Videnskabelige fremskridt sker ved at “man har en teori indtil den bliver falsificeret”





Science may be described as the art of systematic over-simplification — the art of discerning what we may with advantage omit.

(Karl Popper)

- Se “Conjectures and Refutations” og “The Logic of Scientific Discovery”
- Se også “The Open Society and its Enemies”

Statistikens tolkningsregel

I statistikken anlægger man følgende tolkningsregel:

Det usandsynlige sker ikke

- Altså, hvis man observerer data der, hvis modellen er rigtig, er meget usandsynlige, så forkaster man modellen.
- Nødvendigt med sådan en tolkningsregel, for ellers kan man aldrig ad statistisk vej erkende noget som helst! Man vil jo altid kunne hævde, at det foreliggende datasæt blot er et *uheldigt* udfald, som ganske vist er usandsynligt men dog muligt.
- Eksempel: At slå 20 gange “krone” i 20 kast med en fair mønt sker med $\text{ssh} \approx 10^{-6}$; dvs. ca. 1 ud af 100.000 gange. 20 gange krone er et *muligt* udfald, men det er ikke videre *sandsynligt*. Så derfor har vi mere fidus til at mønten ikke er fair – altså at modellen er forkert.

Fødes der lige mange drenge og piger?

Over en årrække blev disse data indsamlet (på et hospital i London) - det STORE datasæt:

| Drenge | Piger | Total |
|--------|-------|-------|
| 6389 | 6135 | 12524 |
| 0.51 | 0.49 | 1 |

Vi skal senere bruge et mindre datasæt, 10% af det store datasæt - det LILLE datasæt:

| Drenge | Piger | Total |
|--------|-------|-------|
| 639 | 614 | 1253 |
| 0.51 | 0.49 | 1 |

| Drenge | Piger | Total |
|--------|-------|-------|
| 6389 | 6135 | 12524 |
| 0.51 | 0.49 | 1 |

- Er der 50–50 chance for en dreng og en pige?
- Tydeligvis ikke – i dette datasæt.
- Men hvad med *populationen*? 51% er ikke langt fra 50% og afvigelsen kunne jo skyldes en tilfældighed.
- Spørgsmålet er: Er afvigelsen så stor, at den ikke – med rimelighed – kan tilskrives en tilfældighed?

Model for data:

For at komme videre skal vi have en model – en mekanisme, der kunne have genereret data:

Uden antagelser, ingen konklusioner. Vi antager

- Alle kvinder har samme sandsynlighed θ for at føde en dreng.
- Udkommet af alle graviditeter er uafhængige – også forskellige graviditeter for samme kvinde og også for forskellige graviditeter for samme mand.

Er disse antagelser rimelige? Tjah – måske – i hvert fald: uden antagelser, ingen konklusioner.

Fører til at antal drengefødsler X er binomialfordelt

$$X \sim \text{bin}(N, \theta), \quad N = 12524$$

Hypotesetest

- Et “fagligt” spørgsmål: Fødes der lige mange drenge og piger?
- Oversættes til statistisk spørgsmål: “Er θ (ssh for en dreng) lig med $1/2$ ”?
- Plejer at formulere det som hypotese:
 - Tester null-hypotesen: $H_0 : \theta = \theta_0$, hvor $\theta_0 = 1/2$ mod den
 - Alternative hypotese: $H_A : \theta \neq \theta_0$.
- Taler om at forkaste eller acceptere hypotesen.
- Måske burde man erstatte “acceptere” med “ikke forkaste” – men det lader sig næppe ændre.
- Poppers tankegang: At forkaste hypotesen er den **stærke konklusion**

Klassisk fremgangsmåde:

- Lad x betegne data.
- Vælg en funktion $t(x)$, der har den egenskab, at $t(x)$ er (numerisk) stor, hvis data ikke passer på modellen og lille ellers.
- Kalder $t(x)$ en teststørrelse (en: test statistic).
- Kunne f.eks. tage

$$t(x) = |x/N - \theta_0| = |x/N - 1/2|$$

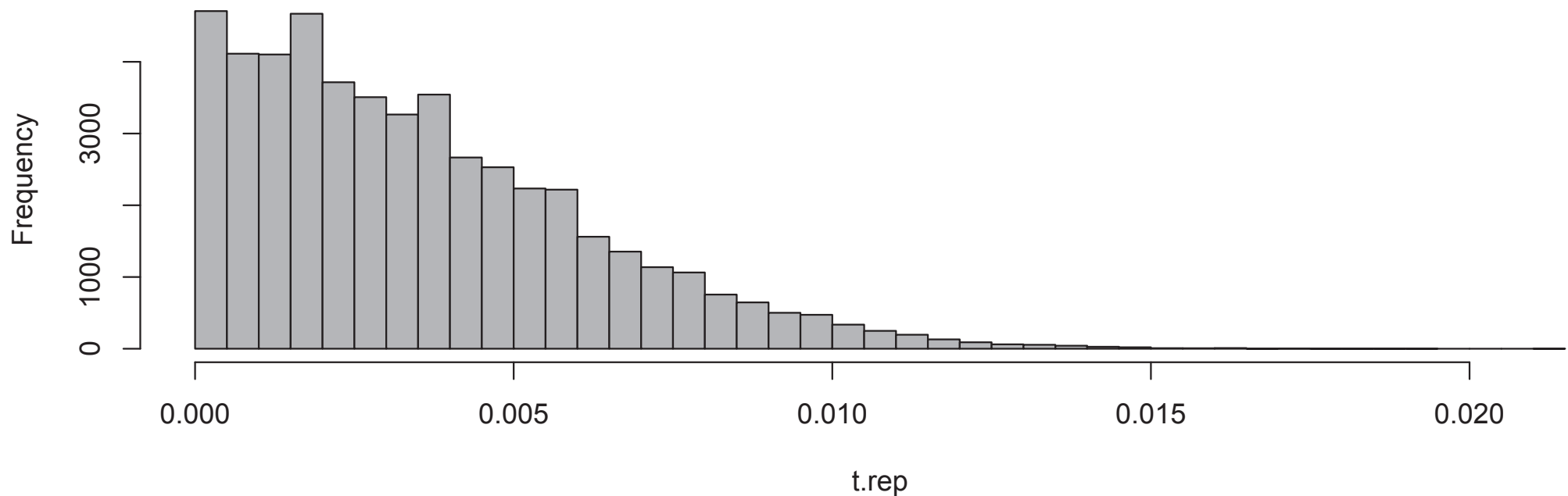
- Vi får den observerede teststørrelse $t_{obs} = t(x) = t(6389) = 0.0101$
- Er t_{obs} et stort eller lille tal?
- Svaret ligger i at spørge: Hvad er sandsynligheden for i fremtiden at se værdier af $t(x)$ der er større end t_{obs} hvis $\theta = \theta_0$?

Tankegangen er nu:

- Lad os nu antage, at der findes en afkrog et sted på jorden, hvor vi (af en eller anden grund) ved, at i denne afkrog er hypotesen sand, dvs. $\theta = \theta_0 = 1/2$.
- I denne afkrog gentager vi studiet M gange 1) venter på, at $N = 12524$ børn er født og 2) noterer os antal drenge x^j for $j = 1, \dots, M$.
- Beregn $t(x^j)$ for hvert x^j og tegn et histogram af $t(x^j)$ 'erne.

Vi behøver ikke lede efter denne afkrog på jorden; computeren er opfundet og vi kan lave studiet ved simulation (et “in silico trial”):

Histogram of t.rep



- Tænk på, at vi skal tage en beslutning: **Acceptere H_0** eller **forkaste H_0** .
- Laver en beslutningsregel: Forkast H_0 hvis $t(x)$ er “stor”; mere konkret: forkast H_0 hvis $t(x) \geq c$ hvor c er et tal, kaldet den **kritiske værdi**.
- Der er to typer fejl vi kan begå:
- Forkaste H_0 selvom H_0 er sand; kaldes **type-I** fejl
- Acceptere H_0 selvom H_0 er falsk; kaldes **type-II** fejl

- Man fastlægger ofte at ssh for at begå en type-I fejl skal være mindre end et tal α ; f.eks. $\alpha = 0.05$.

$$Pr_{\theta_0}(\text{Forkaste } H_0) \leq \alpha$$

hvor $Pr_{\theta_0}()$ indikerer, at ssh er beregnet for $\theta = \theta_0$.

- Hvis beslutningsreglen er "Forkast H_0 hvis $t(x) \geq c$ så kan vi finde c fra:

$$Pr_{\theta_0}(t(X) \geq c) \leq \alpha$$

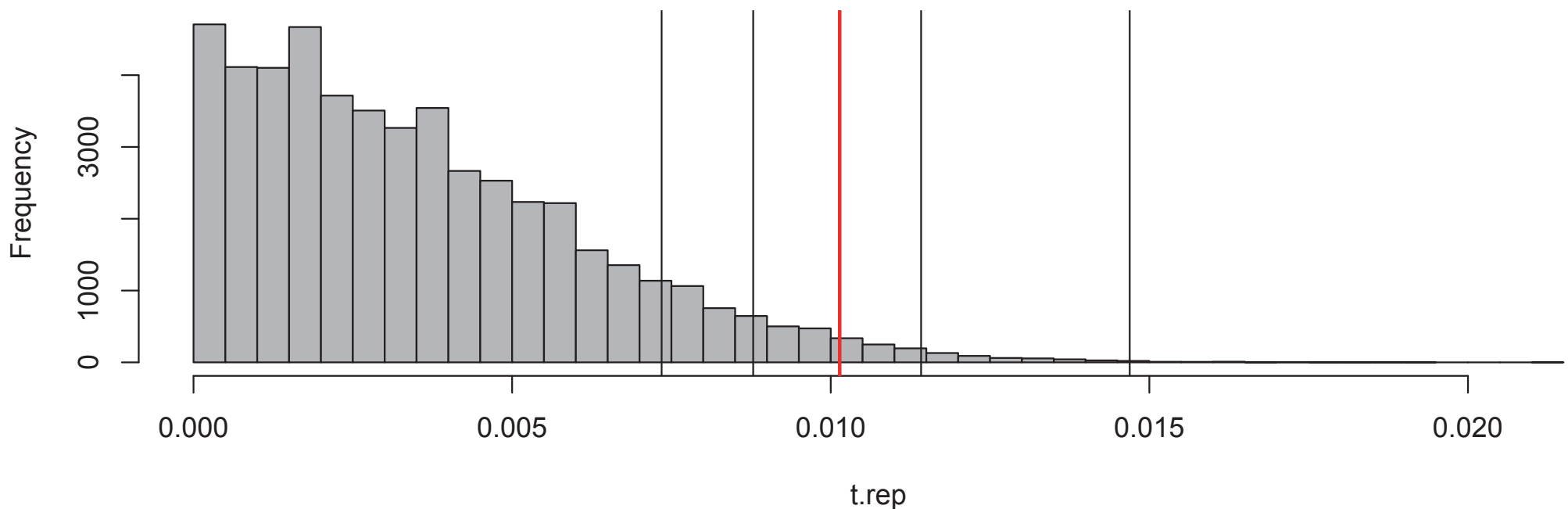
- Beslutningsreglen bliver så: Forkast H_0 hvis $t(x) \geq c$.
- Hvis $t(x) \geq c$ siger man, at **testet er signifikant på niveau α** .

- For hvert α finde den kritiske værdi c_α :

| | | | | |
|----|----------|----------|----------|----------|
| ## | 0.1 | 0.05 | 0.01 | 0.001 |
| ## | 0.007346 | 0.008783 | 0.011418 | 0.014692 |

Sammenholder med $t_{obs} = 0.0101$

Histogram of t.rep



Vi siger at testet er **signifikant på niveau 5%** (men ikke signifikant på niveau 1%).

Ofte bruger man **signifikansniveauerne** 0.10, 0.05, 0.01 og 0.001 – men der er altså intet “guddommeligt” over disse tal; de har alene historiske grunde.

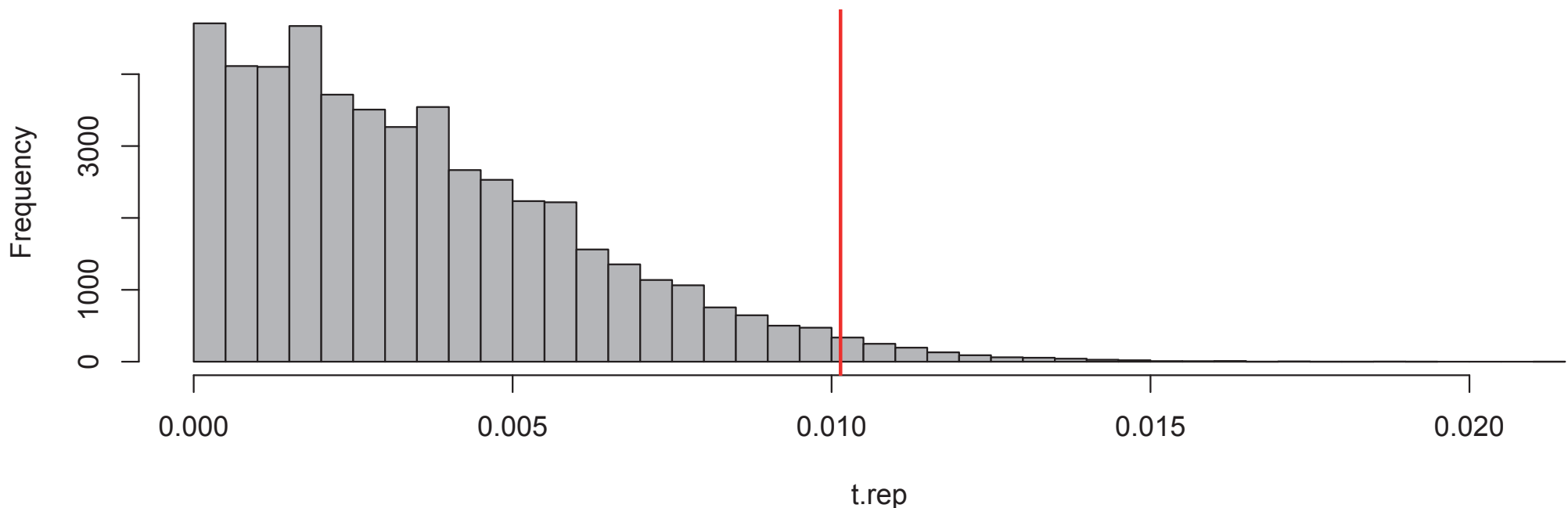
p -værdier

En lidt anden tilgang er: Beregn p -værdien (også kaldet testsandsynligheden) der er defineret som

$$p = Pr_{\theta_0}(t(X) \geq t_{obs})$$

– altså ssh for at observere en værdi at teststørrelsen $t()$ der er større end den vi aktuelt står med. Vi får p -værdien er 0.0235

Histogram of t.rep



- På den måde kan man sige, at p -værdien er et mål for “graden af evidens mod en hypotese”.
- Giver i nogle sammenhænge langt mere mening end at gøre problemet til et beslutningsproblem.

Fejltyper: type-1 fejl og type-2 fejl

Verdens sande tilstand - og de beslutninger vi træffer.

| | H_0 hypotese accepteres | H_0 hypotese forkastes |
|-------------------------|---------------------------|--------------------------|
| H_0 hypotese er sand | | type-I fejl |
| H_0 hypotese er falsk | type-II fejl | |

At forkaste H_0 selvom H_0 er sand kaldes en type-1 fejl

At acceptere H_0 selvom H_0 er falsk kaldes en type-2 fejl

Som ved retssag:

| | | |
|--------------------------|-------------------------|------------------------|
| | H_0 uskyld accepteres | H_0 uskyld forkastes |
| H_0 anklagede uskyldig | | “justitsmord” |
| H_0 anklagede skyldig | en “fejl” vi lever med | |

Man lader tvivlen komme den anklagede til gode: “Man er uskyldig til noget andet er bevist”.

Med mindre der er stærk evidens (data) mod H_0 så accepterer man H_0 .

Fortolkning af p -værdier

- Tilbage til det oprindelige spørgsmål: Er der 50-50 chance for en dreng og en pige?
- En p -værdi kan opfattes som et mål for evidensen MOD en hypotese: En lille p -værdi indikerer stor evidens mod hypotesen.
- Her er p -værdien lille så det får os til at tvivle på hypotesen.
- Kan vi deraf konkludere, at null-hypotesen $H_0 : \theta = \theta_0 = 1/2$ er falsk? Har vi “bevist”, at $\theta \neq 1/2$.
- Nej. Hvis $\theta = 1/2$, så er sandsynligheden for at observere 6389 drenge i 12524 graviditeter er 0.00054 eller knapt 1 ud af 2000 gange. Det er en lille sandsynlighed, bevarer, men det er afgjort muligt selv hvis hypotesen er sand.
- Der er dog mange studier, der peger på, at der fødes flere drenge end piger.

- Some tider fortolkes en p -værdi fejlagtigt som noget i retningen af *“ p -værdien er sandsynligheden for at hypotesen er sand.”*
- Dette er forkert: Sandsynligheder er noget vi knytter til fænomener, hvor der er usikkerhed om udkommet (kast med en mønt eller en terning).
- Der er ingen usikker om hypotesen: Hypotesen er enten sand eller falsk (vi ved bare ikke hvad den er, for vi har ingen guddommelig indsigt).

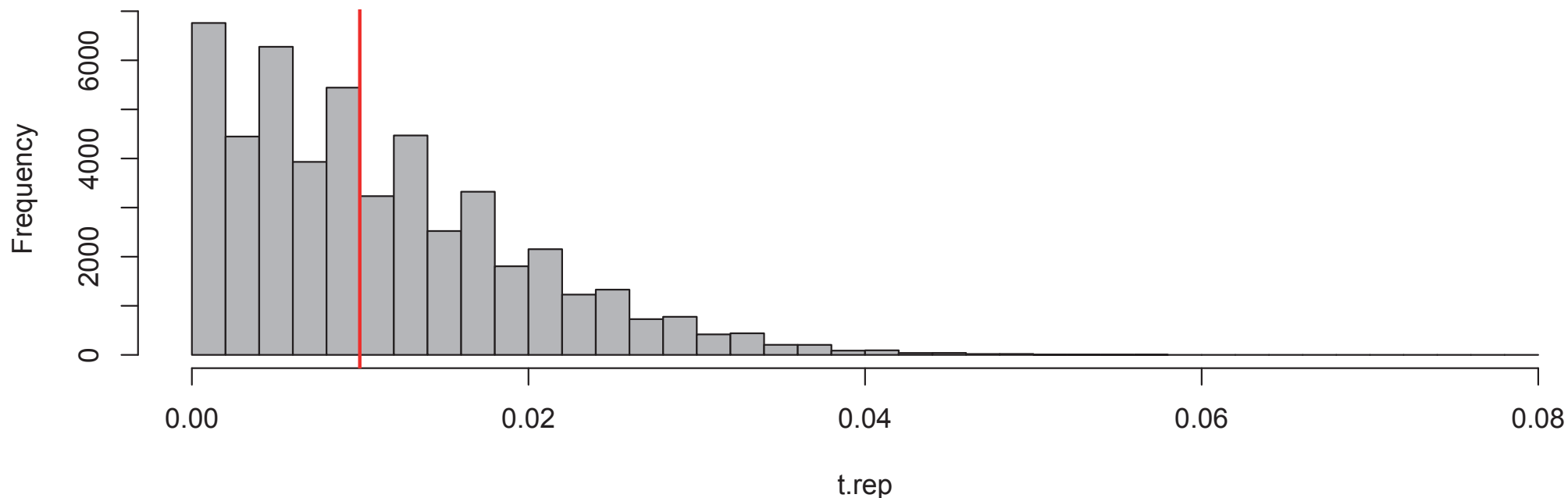
Effekten af stikprøvestørrelsen

Antag at data var følgende:

| Drenge | Piger | Total |
|--------|-------|-------|
| 639 | 614 | 1253 |
| 0.51 | 0.49 | 1 |

Dvs vi har kun 10% of data, men andelen af drenge er stadig 0.51.

Histogram of t.rep



Hvad kan vi så konkludere?

t_{obs} er som før (på nær nogle decimaler): 0.01 men

- Med 12524 børn og 6389 drenge er der stærk evidens mod hypotesen $\theta = \frac{1}{2}$. p -værdien er 2%.
- Med 1253 børn og 639 drenge er p -værdien er meget stor: 0.4975 så der er meget lidt evidens mod hypotesen.
- I begge tilfælde er andelen af drenge 0.51. Hvad skal vi mene om dette?

- Vi formulerer en hypotese om “verdens sande tilstand”, og dernæst “spørger vi data” om der er evidens **mod** denne hypotese.
 - Hvis der ikke er evidens mod hypotesen, kan det være fordi hypotesen er sand, eller
 - fordi der ikke er tilstrækkeligt data (information) til at komme med denne evidens (altså til at “påvise” at hypotesen er forkert)

Mere poetisk:

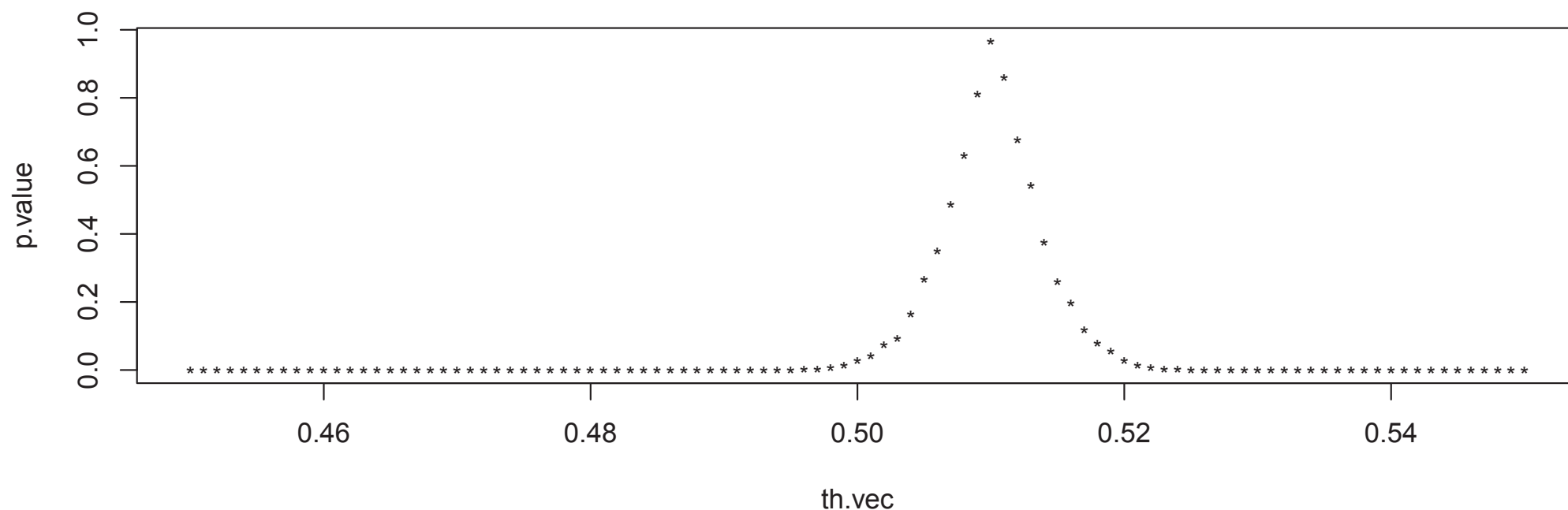
“Absence of evidence (of an effect) is NOT the same as evidence of absence (of an effect).”

Test og konfidensinterval – to sider af samme sag

Vi testede ovenfor hypotesen $\theta = \theta_0$ hvor $\theta_0 = 1/2$.

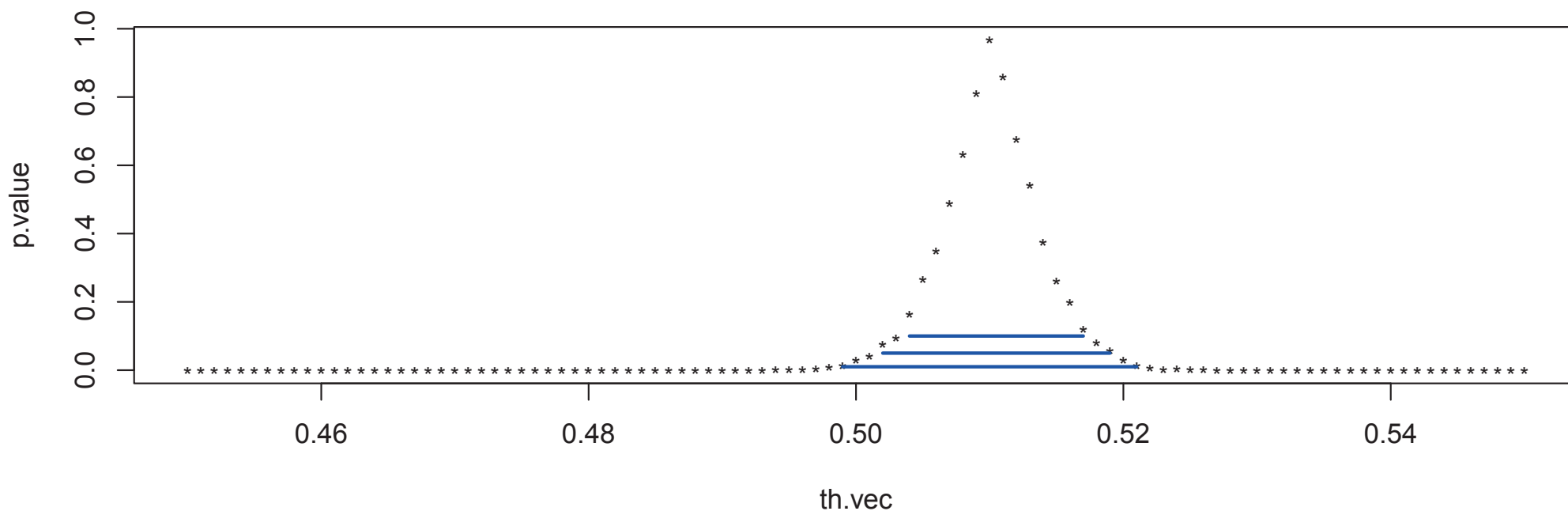
Vi kunne teste samme hypotese for mange andre værdier af θ_0 .

For hver værdi af θ_0 beregner vi p -værdien og plotter mod θ_0



Husk: Små p -værdier er evidens mod hypotesen.

Indlæg intervaller hvor p -værdien er større end 0.01, 0.05 og 0.10.

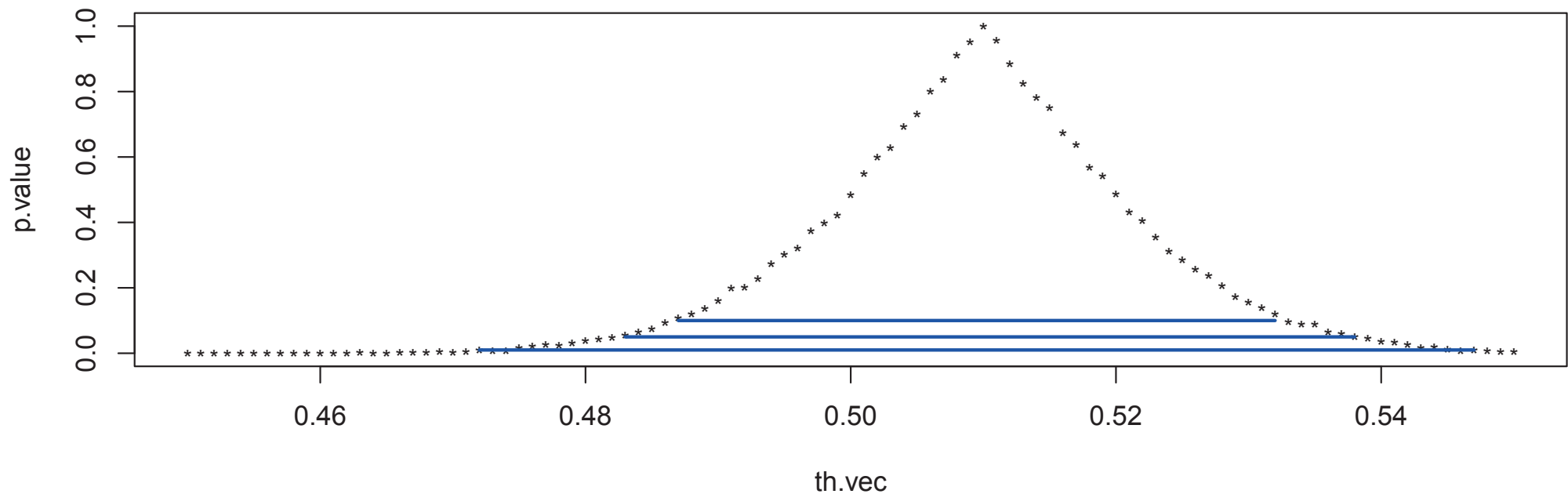


Disse intervaller er præcist 99%, 95% og 90% **konfidensintervaller**:

99% konfidensinterval: [0.499; 0.521]

95% konfidensinterval: [0.502; 0.519]

90% konfidensinterval: [0.504; 0.517]



Disse intervaller er præcist 99%, 95% og 90% **konfidensintervaller**:

99% konfidensinterval: [0.475; 0.547]

95% konfidensinterval: [0.483; 0.538]

90% konfidensinterval: [0.487; 0.532]

Statistisk signifikans, praktisk signifikans, klinisk signifikans. . .

Oprindelsen er det latinske **significantia**, der betyder **betydning**

Når man finder en statistisk signifikant “effekt” så betyder det, at den effekt man ser er for stor til – med rimelighed – at kunne tilskrives tilfældigheder.

Mange studier, der viser, at der fødes ca. 50.5% drenge og 49.5% piger.

Men når man venter et barn så tænker man, at der er 50–50 chance for hvert køn.

Den statistiske signifikans betyder altså ikke nødvendigvis så meget i praksis. . .

Man finder det samme fænomen i sundhedsverdenen: En statistisk signifikant effekt kan sagtes være så svag, at den ikke er **klinisk** relevant for patienten.

Estimation (by request)

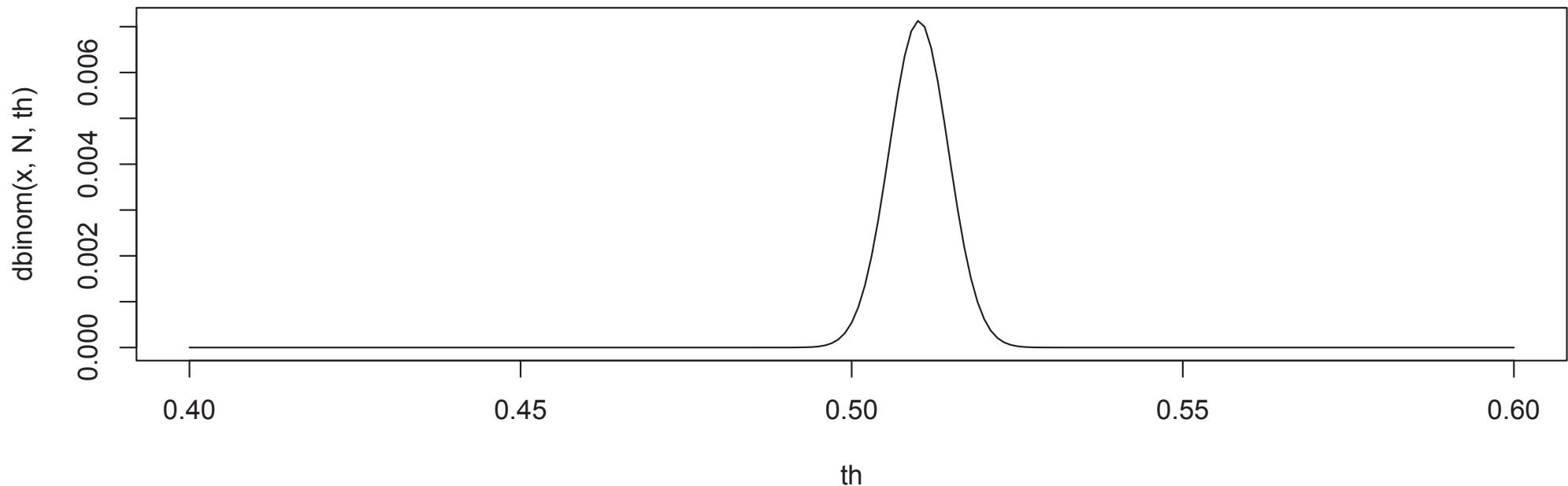
- Estimation af θ i binomialfordelingen
- Det er ikke helt oplagt at bruge mindste kvadraters metode.
- Alternativ: Maximum Likelihood Metoden:
- Model: $X \sim \text{bin}(N, \theta)$
- Binomial tæthed

$$\Pr(X = x; \theta) = \binom{N}{x} \theta^x (1 - \theta)^{N-x}$$

- Når data er observeret $x = 6389$ med $N = 12524$ så bliver ovenstående en funktion af θ alene. Man kalder så funktionen for **likelihood funktionen**

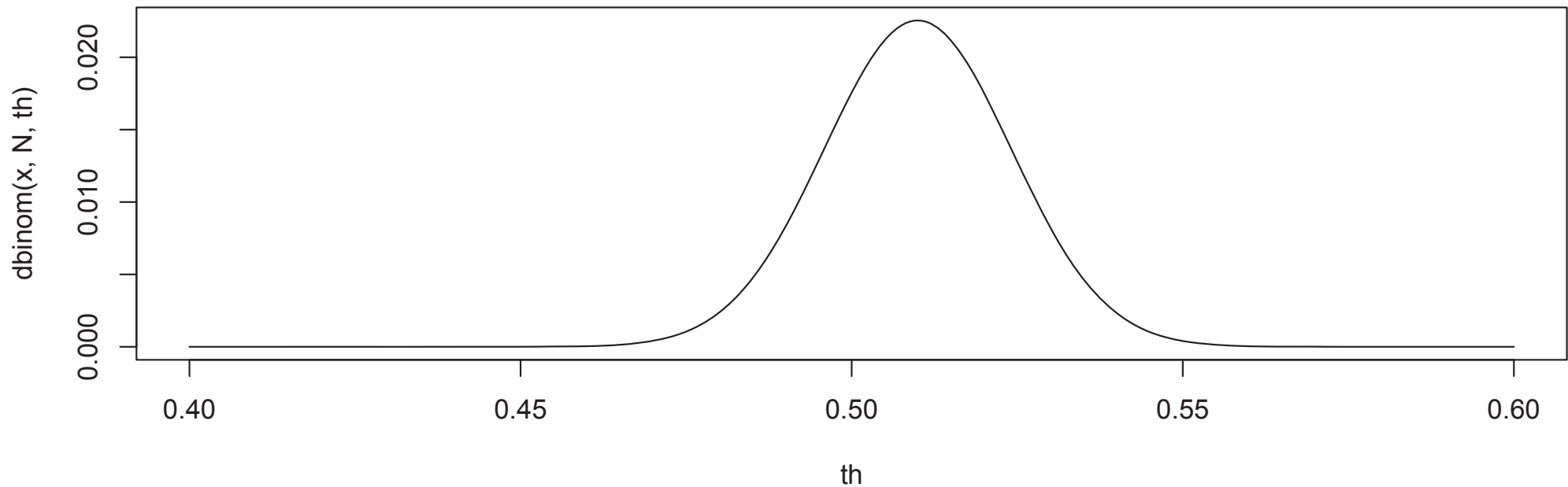
$$L(\theta) = \Pr(X = x; \theta)$$

- Plot $L(\theta)$ mod forskellige værdier af θ



- Den værdi af θ der maximerer L kaldes **maximum likelihood estimatet** (MLE).
- Det er oftest lettere at maximere $l(\theta) = \log L(\theta)$ og MLE bliver $\hat{\theta} = x/N$
- MLE er på mange måder det bedst tænkelige estimat man kan opnå.

- For det lille datasæt får vi



- Samme maximum, men likelihood funktionen er mindre peaked – svarende til at der er større usikkerhed på estimatet fordi datasættet er mindre.

Mindste kvadrater og MLE

- Hvad så med regressionsmodellen og mindste kvadrater?
- Model: $y_i = \beta_0 + \beta_1 x_i + e_i$
- Samme som at sige $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- Normalfordelingstæthed bliver

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2\right)$$

- Hvis y_1, \dots, y_N er uafhængige så er

$$f(y_1, \dots, y_N) = \prod_{i=1}^N f(y_i)$$

- Bliver helt konkret

$$f(y_1, \dots, y_N) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - (\beta_0 + \beta_1 x_i))^2\right)$$

- Antag nu at σ er kendt. Så er likelihood funktionen

$$L(\beta_1, \beta_2) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - (\beta_0 + \beta_1 x_i))^2\right)$$

- Pga af minus i eksponenten så maximeres L ved at minimere

$$\sum_{i=1}^N (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Dvs i regressionsmodellen er mindste kvadrater og MLE det samme.