

# Konfidenzbereiche die auf Runden Normaldaten Basiert Sind

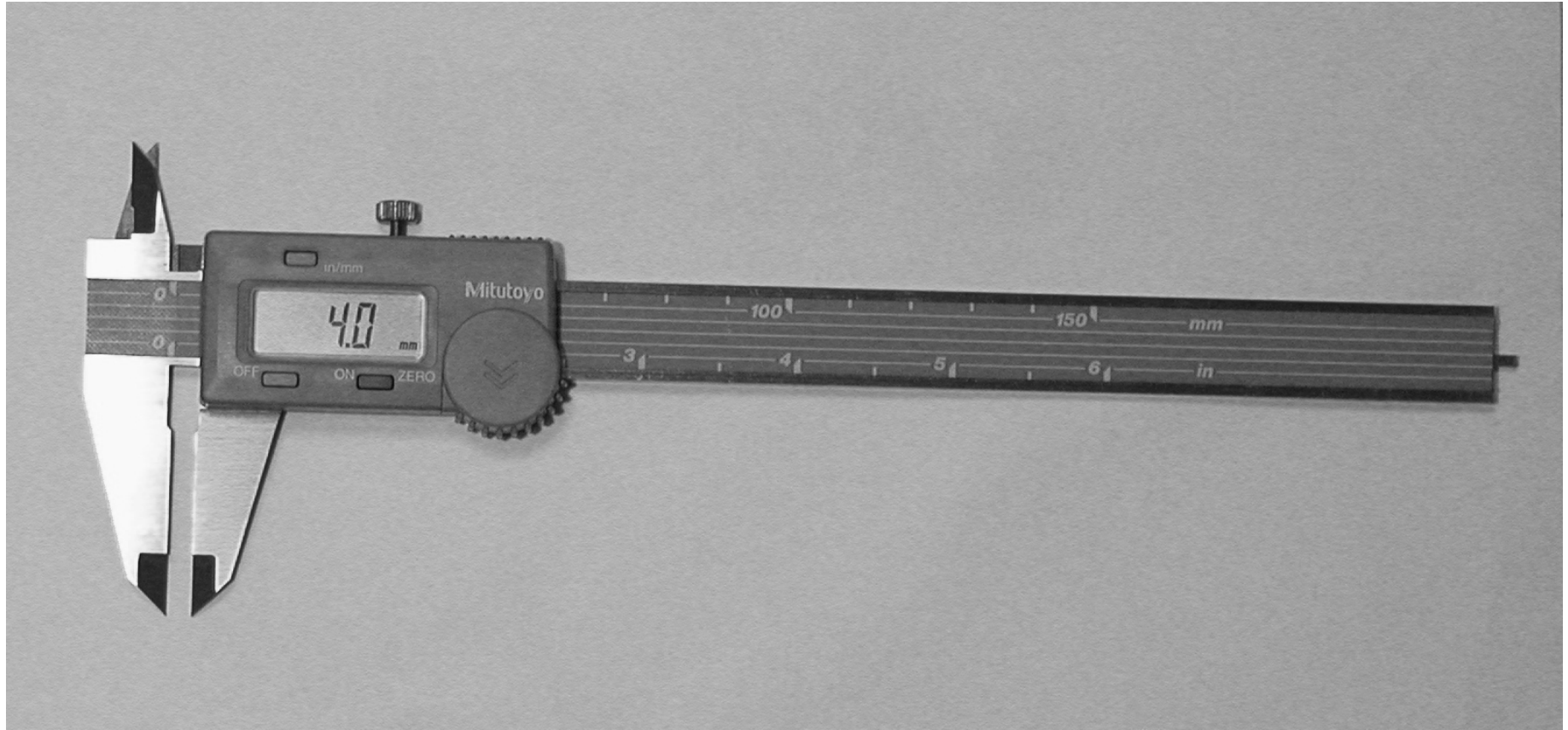
Steve Vardeman

C-S (Johnson) Lee

*(JQT 2001, Comm Stat 2002, (2003))*

Iliana Vaca

(M.S. laufend)



# Gerundete/Digitale Daten

- Kein neues Problem ... z. B. gibt es:  
Sheppard, W. (1898). "On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equidistant Divisions of a Scale." *Proceedings of the London Mathematical Society* **29**, pp. 353-380
- Messtechniker halten dieses für eine Fehlquelle, aber haben keine guten Verfahren, sie zu berücksichtigen
- Einfache statistische Methoden gehen davon aus, dass dieses Problem nicht wichtig ist

# Aber ... Stetige Verteilungen (in Einfachen Methoden)

- Beschreiben, genau genommen, nur Zufallsexperimente, die reelle Zahlen produzieren (die unendlich viele Dezimalstellen haben)
- Selbst wenn sie ein physisches Phänomen gut beschreiben, ist es nicht klar, dass sie auch gut beschreiben, was man beobachten kann

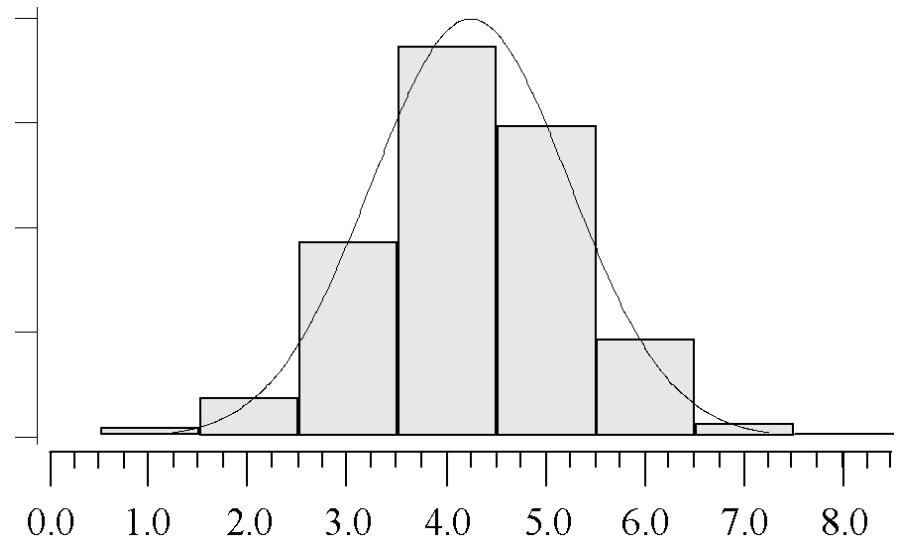
# Beobachtung “Mit Genauigkeit $\Delta$ ”

- Stichprobenwerte  $y$  werden möglicherweise in ganze Zahlen umgewandelt durch  $y' = (y - y_0) / \Delta$  so dass die Folge  
1.2, 1.2, 1.2, 1.2, 1.3, 1.3, 1.3, 1.3, 1.3, 1.3  
(mit  $\Delta = .1$  ) umgewandelt wird zu  
2, 2, 2, 2, 3, 3, 3, 3, 3, 3
- Nehmen wir an, dass ein stetig-verteiltes  $X$  in ein gerundetes/digitales  $Y$  umgewandelt wird, dann ist es möglich, dass die diskrete Verteilung von  $Y$  nicht nach der Verteilung von  $X$  aussieht

# Zwei $\Delta = 1$ “Normalfälle”

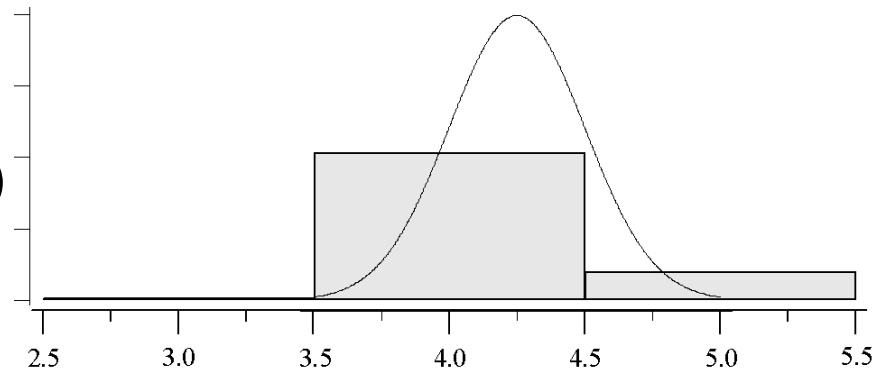
$\mu = 4.25$  und  $\sigma = 1.0$

$(\mu_Y = 4.25$  und  $\sigma_Y = 1.0809)$



$\mu = 4.25$  und  $\sigma = .25$

$(\mu_Y = 4.1573$  und  $\sigma_Y = .3678)$



# Der Schlüssel: Größe von $\frac{\sigma}{\Delta}$

- Wenn  $\sigma \geq .5\Delta$ ,  
dann ist  
$$|\mu - \mu_Y| < .005\Delta$$
- Wenn  $\sigma \approx 0$ ,  
dann kann  $|\mu - \mu_Y|$   
fast  $.5\Delta$  sein
- Wenn  $\sigma > .15\Delta$ ,  
 $\sigma_Y > \sigma$ . Für solche  $\sigma$ ,  
$$|\sigma_Y - \sigma| / \sigma$$
  - nimmt mit  $\sigma$  ab
  - ist kleiner als .141  
für  $\sigma \geq .5\Delta$
- Für kleines  $\sigma$ , kann  $\sigma_Y$   
viel größer oder viel kleiner  
als  $\sigma$  sein

# Naive Verwendung von Verfahren für Stetige Daten ...

- $\bar{y}$  schätzt  $\mu_Y$  (nicht  $\mu$ ), und wenn  $\mu_Y \neq \mu$ , dann ist das Konfidenzintervall mit den Grenzen

$$\bar{y} \pm t \frac{s_y}{\sqrt{n}}$$

zentriert in  $\mu_Y$ , und auf diese Weise ist die Vertrauenswahrscheinlichkeit fast 0

- $s_y$  schätzt  $\sigma_Y$  (nicht  $\sigma$ ), und nur wenn  $\sigma$  groß ist, ist  $(n-1)s_y^2 / \sigma^2$  ungefähr  $\chi^2$ -verteilt

# Inferenz-Machine: die “Korrekte” Likelihood

- (Gerundete Daten) Likelihood für eine einfache normalverteilte Stichprobe

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^n P_{\mu, \sigma} (y_i - .5\Delta < X < y_i + .5\Delta) \\ &= \prod_{i=1}^n \left( \Phi \left( \frac{y_i + .5\Delta - \mu}{\sigma} \right) - \Phi \left( \frac{y_i - .5\Delta - \mu}{\sigma} \right) \right) \end{aligned}$$

- Log-likelihood

$$l(\mu, \sigma) = \log L(\mu, \sigma)$$

- Profile-log-likelihood

$$l^*(\mu) = \sup_{\sigma > 0} l(\mu, \sigma)$$

$$l^{**}(\sigma) = \sup_{\mu} l(\mu, \sigma)$$

# Einfache Grenzverteilungen

Wenn

$$M = \sup_{(\mu, \sigma)} l(\mu, \sigma)$$

(max (sup) log-likelihood), dann gilt

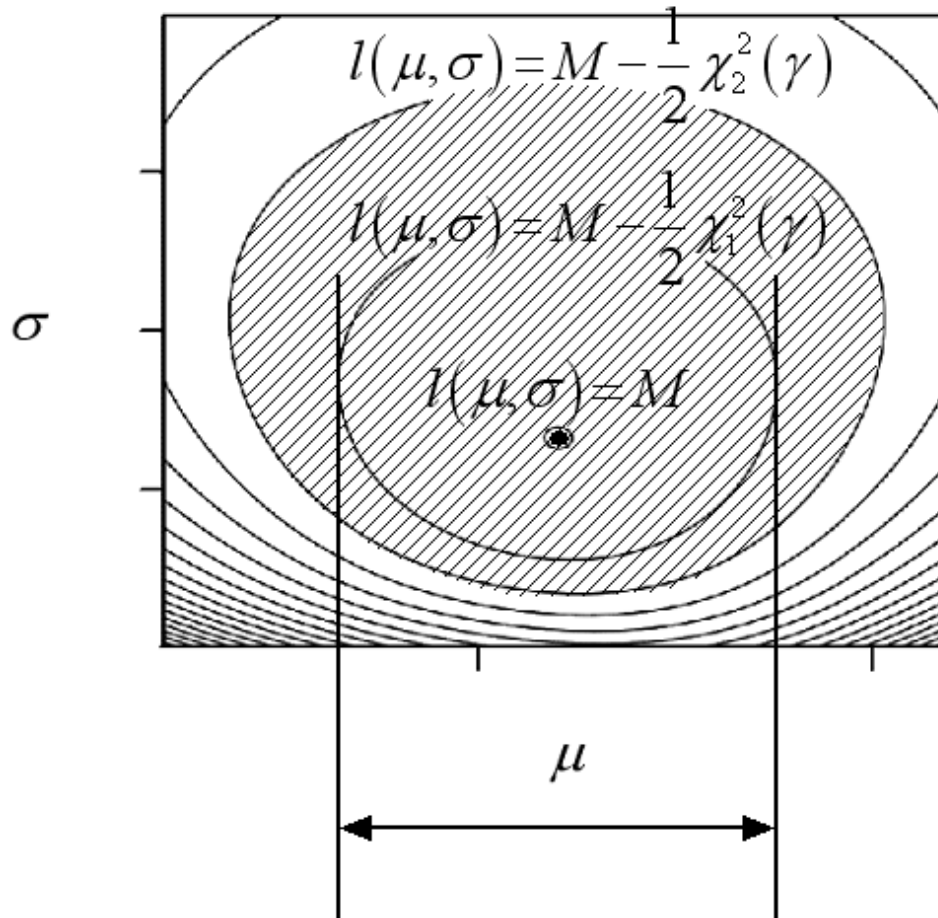
$$2(M - l(\mu, \sigma)) \xrightarrow[n \rightarrow \infty]{\mathcal{L}_{(\mu, \sigma)}} \chi_2^2$$

$$2(M - l^*(\mu)) \xrightarrow[n \rightarrow \infty]{\mathcal{L}_\mu} \chi_1^2$$

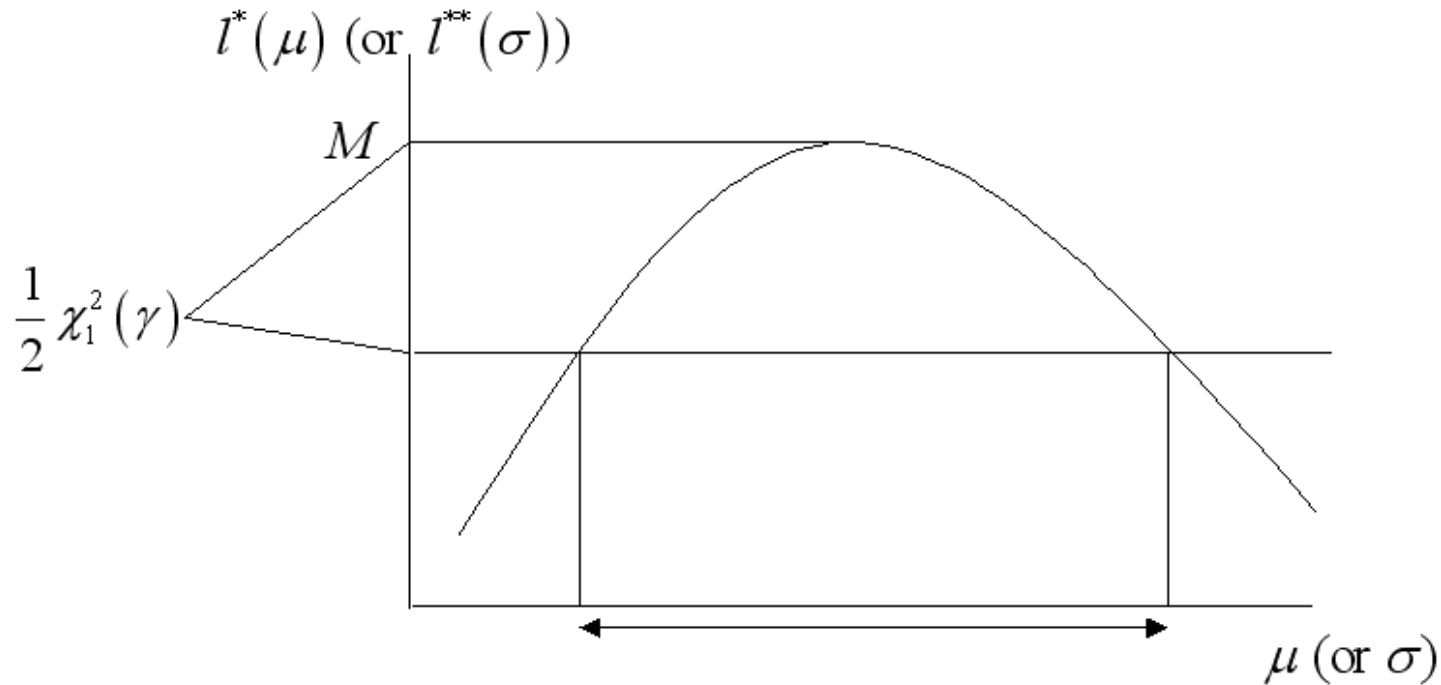
$$2(M - l^{**}(\sigma)) \xrightarrow[n \rightarrow \infty]{\mathcal{L}_\sigma} \chi_1^2$$

# Asymptotische Konfidenzbereiche

- Bereich für  $(\mu, \sigma)$  schattiert; Intervall für  $\mu$  (ähnliches Intervall für  $\sigma$ )



- Dazugehörige Skizze für Profile-log-likelihoods und Vertrauensintervalle für  $\mu$  oder  $\sigma$



# Wenn man die asymptotischen Konfidenzbereiche in die Praxis umsetzt ...

- Die eigentlichen Sicherheitsschwellen sind viel zu klein
  - für  $\mu$  wenn  $\sigma$  *groß* ist
  - für  $\sigma$  wenn  $\sigma$  *entweder* groß oder (etwas) klein ist
  - für  $(\mu, \sigma)$  wenn  $\sigma$  *groß* ist
- *Abschätzung* der Bereiche ist nicht immer so klar (die Log-likelihood sieht nicht immer so schön aus)

# Unser Plan (im Rückblick)

- Das Verhalten der Log-likelihood und der profile-log-likelihoods für kleines  $n$  verstehen
- Irgendwie die unzulänglichen Sicherheitsschwellen in Ordnung zu bringen ...  
 $\chi_1^2(\gamma)$  und  $\chi_2^2(\gamma)$  durch passende (größere) Werte ersetzen zusammen mit ????????? ...
  - **Grundlegende Idee:** für *großes*  $\sigma$ , hat vielleicht  $2(M - l(\mu, \sigma))$  im Wesentlichen dieselbe Verteilung wie die dazugehörige Zufallsvariable, die auf den genauen  $x$ -Werten basiert

# Die Natur der Log-likelihood

- Diese hängt von der Spannweite  $R$  ab, und nur wenn  $R \geq 2\Delta$ , ist sie “zahm” (hügelförmig)

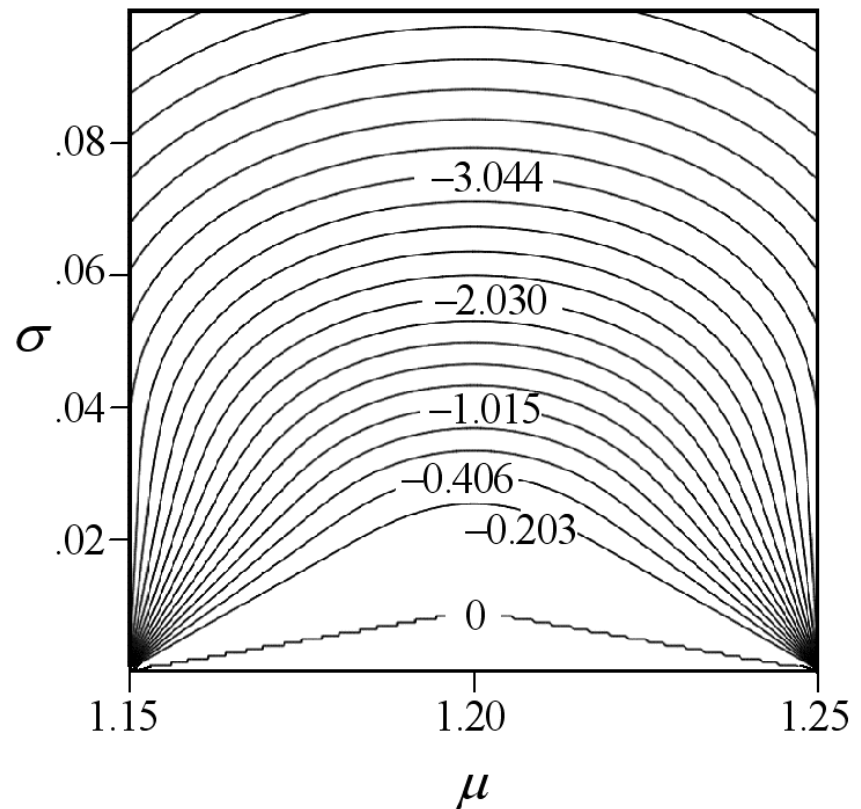
– Ein  $R=0$  Fall:

$n = 10$  Beobachtungen,

alle = 1.2,  $\Delta = .1$ ;

(Basis 10) Log-

likelihood



– Ein  $R=\Delta$  Fall:

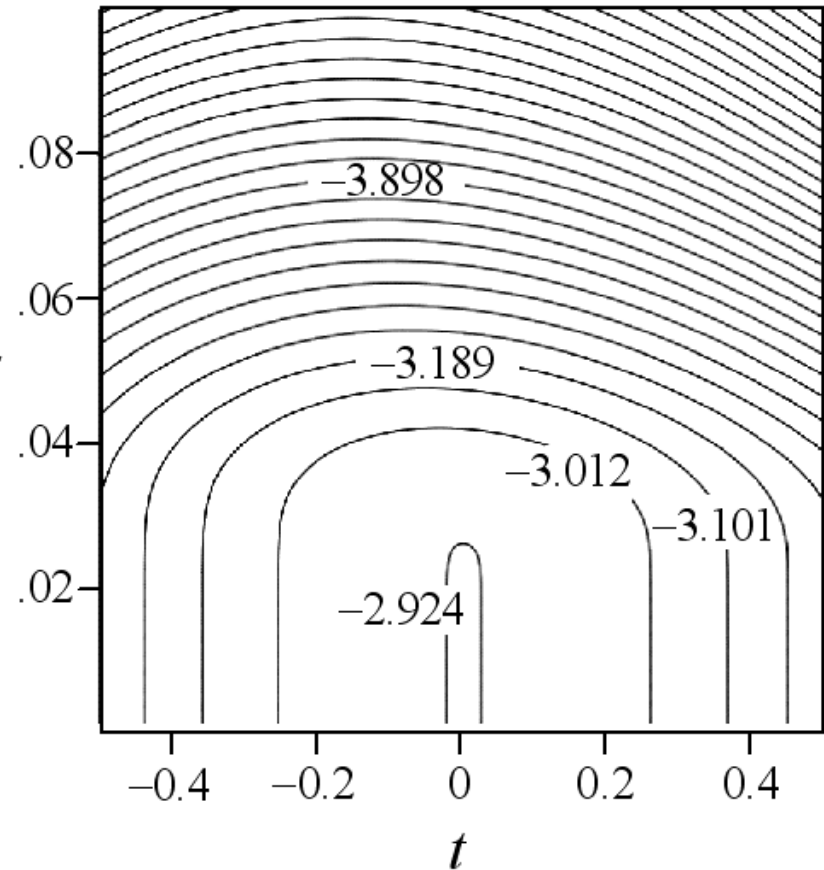
**das anfängliche**

**Beispiel** mit  $\Delta = .1$ ,  $\sigma$

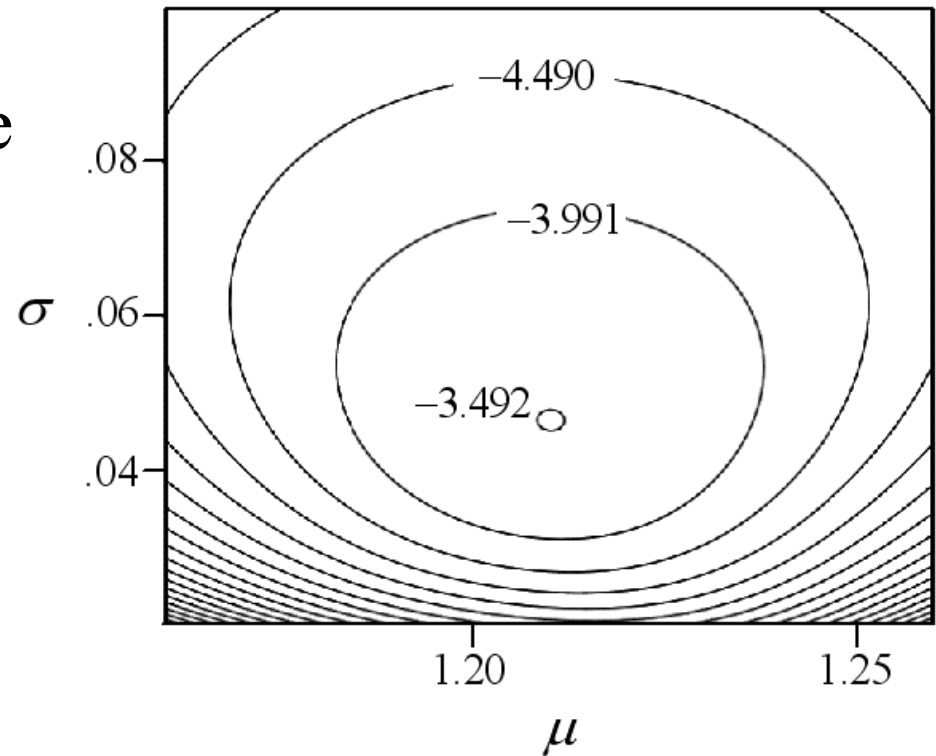
(Basis 10)

Version von

$$l(1.25 + (t + .25)\sigma, \sigma)$$



- Ein  $R=2\Delta$  Fall:
    - $\Delta=.1$ , ein Wert
    - 1.1, sieben Werte
    - 1.2, und zwei Werte
    - 1.3; (Basis 10)
- Log-likelihood



# Schätzung des Mittelwerts $\mu$

- Hier ist die “genaue-Werte- $x$ ” Version von  $2(M - l^*(\mu))$  die Zufallsvariable

$$n \ln \left( 1 + \frac{\left( \frac{\bar{x} - \mu}{s_x / \sqrt{n}} \right)^2}{n-1} \right)$$

und vielleicht kann man

$$c_n(\gamma) = n \ln \left( 1 + \frac{t_{n-1}^2 \left( \frac{1+\gamma}{2} \right)}{n-1} \right)$$

anstatt  $\chi_1^2(\gamma)$  benutzen

- Simulationsstudien zeigen, dass diese Ersetzung sehr wirksam ist
  - Die Intervalle sind von konservativ (für kleines  $\sigma$ ) bis genau (für großes  $\sigma$ )
  - Weil

$$2\left(M - l^*(\mu)\right) \xrightarrow[n \rightarrow \infty]{\mathcal{L}_\mu} \chi_1^2 \quad \text{und} \quad c_n(\gamma) \xrightarrow[n \rightarrow \infty]{} \chi_1^2(\gamma)$$

sind Sicherheitsschwellen korrekt für großes  $n$

- Für großes  $R$  sind die Grenzen ungefähr

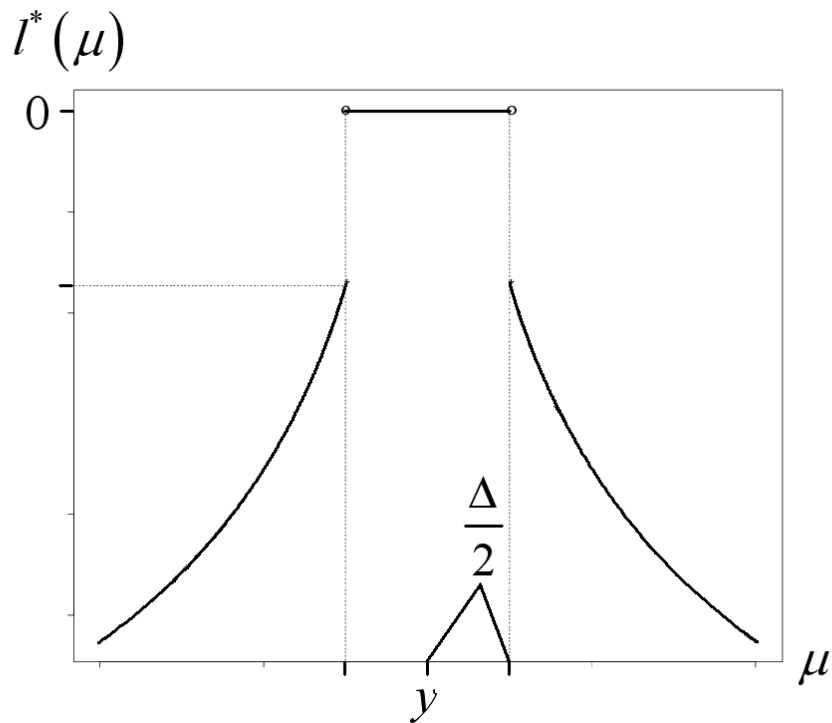
$$\bar{y} \pm t \frac{s_y}{\sqrt{n}}$$

- Für übliche Niveaus  $\gamma$  und mäßige Stichprobenumfänge  $n$ , sind  $R = 0$  Intervalle

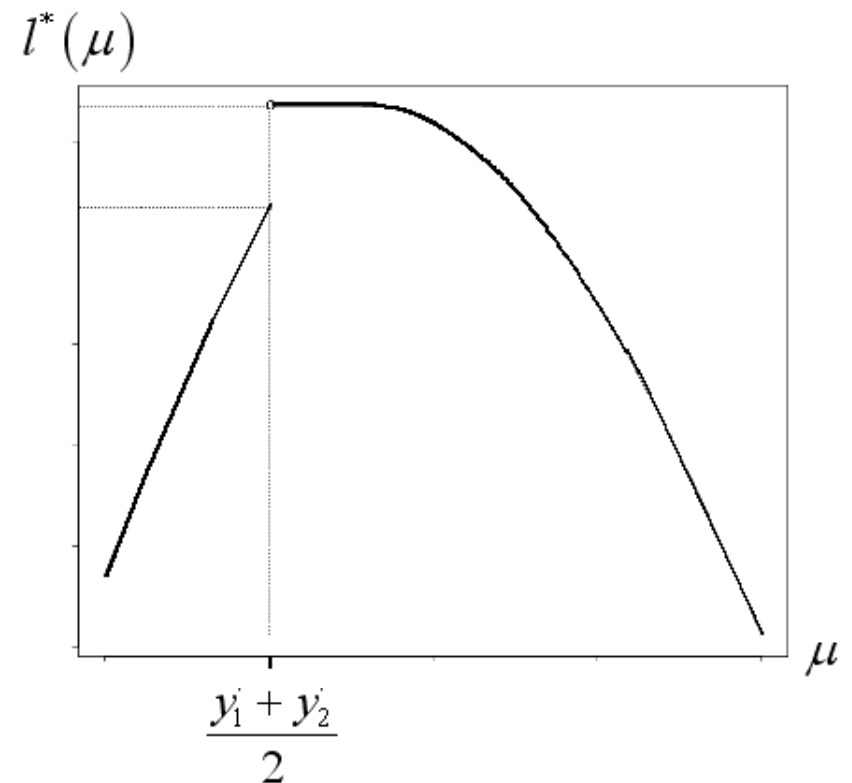
$$\left( y - \frac{\Delta}{2}, y + \frac{\Delta}{2} \right)$$

- Profile-log-likelihoods für  $R=0$  und  $R=\Delta$   
(Skizzen)

$R=0$



$R=\Delta$



# Schätzung der Standardabweichung $\sigma$

- Zunächst betrachten wir für großes  $\sigma$  die unzulänglichen Sicherheitsschwellen ... wir bemerken, dass die “genaue-Werte- $\chi$ ” Version von  $2(M - l^{**}(\sigma))$  die Zufallsvariable

$$n \ln \left( \frac{n\sigma^2}{(n-1)s_x^2} \right) + \frac{(n-1)s_x^2}{\sigma^2} - n$$

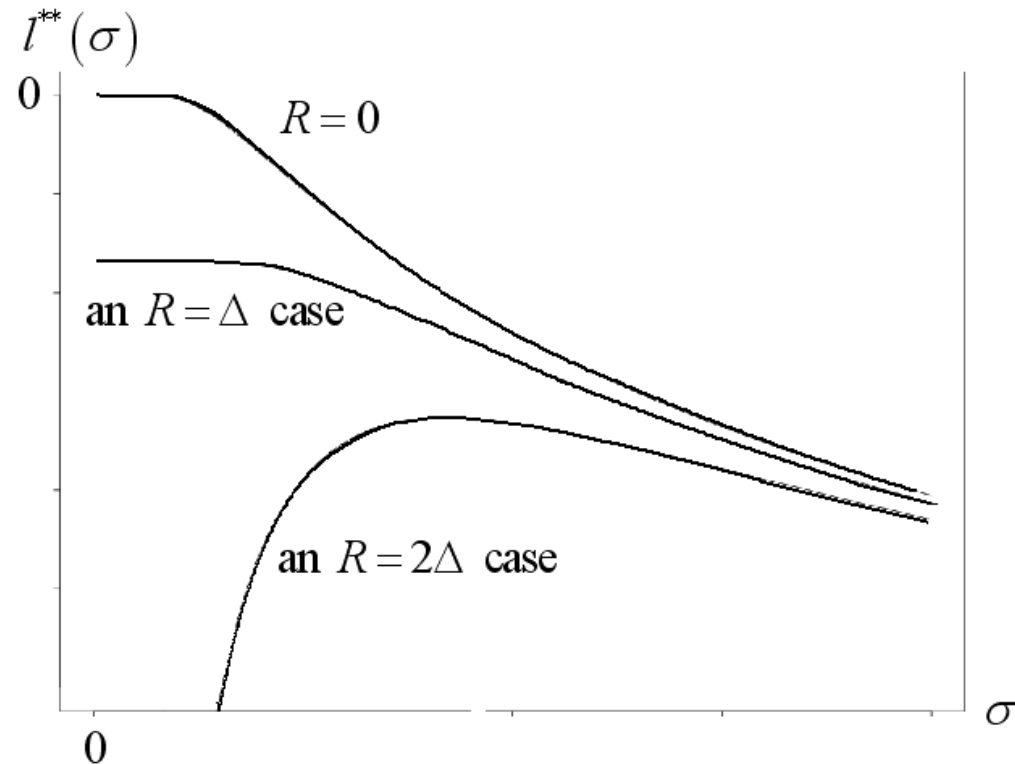
ist und diese hat die Verteilung von

$$U_n = n \ln \left( \frac{n}{W} \right) + W - n \quad \text{wo } W \sim \chi_1^2$$

- vielleicht kann man dann

$d_n(\gamma) =$  das  $\gamma$ -Quantil der Verteilung von  $U_n$   
anstatt  $\chi_1^2(\gamma)$  benutzen

- Dieses ist ein Heilmittel für großes  $\sigma$  (die Sicherheitsschwellen sind korrekt), aber es reicht nicht aus, wenn  $\sigma$  klein ist.
- Beachte: kleines  $\sigma$  bewirkt oft  $R = 0$  oder  $R = \Delta$



- Wir finden, dass die  $R = 0$  und  $R = \Delta$  Log-likelihoods (und  $\chi_1^2(\gamma)$  oder  $d_n(\gamma)$ ) produzieren (nur) obere Intervallgrenzen

$$\sigma_0 < \sigma_{\Delta,1} < \sigma_{\Delta,2} < \dots < \sigma_{\Delta, \lfloor \frac{n}{2} \rfloor}$$

wo

$\sigma_0$  = die " $R = 0$ " Grenze

$\sigma_{\Delta,j}$  = die " $R = \Delta$  und *kleinere Häufigkeit* =  $j$ " Grenze

- ????? Diese Werte durch andere (minimal) größere Werte ersetzen ???

- Eine einfache **notwendige** Bedingung für konservative oder genaue Sicherheitsschwellen, ist

$$P_{\mu,\sigma}^0 + P_{\mu,\sigma}^\Delta \leq 1 - \gamma \quad \forall \mu, \sigma \quad (*)$$

für

$$P_{\mu,\sigma}^\eta = P_{\mu,\sigma} \left( \begin{array}{l} R = \eta \text{ und das Intervall} \\ \text{umfasst } \sigma \text{ nicht} \end{array} \right)$$

- “Rohe Gewalt” Berechnungen machen (für den  $\Delta = 1$  Fall), um ersetzende Werte für  $\sigma_0$  and  $\sigma_{1,j}$  zu finden, die (\*) sicherstellen

- Finden (für  $\Delta = 1$  )

$$\sigma_0^* = \text{Minimum } \sigma \text{ so dass } \max_{\mu} P_{\mu, \sigma} (R = 0) \leq 1 - \gamma$$

- Finden (für  $\Delta = 1$  )

$$\sigma_{1,j}^* = \text{Minimum } \sigma \text{ so dass}$$

$$\max_{\mu} \left[ \begin{array}{l} P_{\mu, \sigma} (R = 0) \\ + \sum_{l=1}^j P_{\mu, \sigma} (R = 1 \text{ und } \textit{kleinere H\u00e4ufigkeit} = l) \end{array} \right] \leq 1 - \gamma$$

- $\sigma_0$  durch  $\Delta \sigma_0^*$  und  $\sigma_{\Delta, j}$  durch  $\Delta \sigma_{1, j}^*$  ersetzen

- Simulationsstudien zeigen, dass diese ( $d_n(\gamma)$  und  $R = 0$  and  $R = \Delta$ ) Ersetzungen ergeben Intervalle,
  - die selten zu kleine Sicherheitsschwellen haben und nie unannehmbar kleine
  - Die, wenn  $\sigma$  groß ist, im Mittel etwas kürzer sind als die “üblichen” auf beiden Seiten “gleich schweren” (auf gerundeten Werten basierten) Intervalle ... keine Überraschung, weil die üblichen Intervalle nicht in Bezug auf Länge optimiert sind

# Beispiel

- Für das (anfängliche) Beispiel (4 Werte 1.2 und 6 Werte 1.3) sind die 95% Intervalle
  - (1.226,1.294) für  $\mu$  (ähnlich wie bei naiver Anwendung des  $t$ -Intervalls in diesem Fall)
  - (0,.0851) für  $\sigma$  (ähnlich wie das naive einseitige  $\chi^2$ -Intervall in diesem Fall) (hier ist  $s_y = .0518$ )

# Gleichzeitige Schätzung für $(\mu, \sigma)$ (in Arbeit)

- Dieses könnte benutzt werden, z.B., um simultane Vertrauensbereiche der Verteilungsfunktion zu bekommen
- Die “genaue-Werte- $x$ ” Version von  $2(M - l(\mu, \sigma))$  ist

$$n \ln \frac{n}{\left( \frac{(n-1)s_x^2}{\sigma^2} \right)} + \left( \frac{(n-1)s_x^2}{\sigma^2} \right) - n + \left( \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right)^2$$

und diese hat die Verteilung von

$$Q_n = n \ln \frac{n}{W} + W - n + V$$

für unabhängige  $W \sim \chi_{n-1}^2$  und  $V \sim \chi_1^2$

- Numerische Berechnung der Verteilungsfunktion (und dann der Quantile) von solch einem  $Q_n$  ist einfach

- Wir erwarten, dass wenn

$q_n(\gamma) =$  das  $\gamma$ -Quantil der Verteilung von  $Q_n$

dann ist

$$\left\{ (\mu, \sigma) \mid M - l(\mu, \sigma) < \frac{1}{2} q_n(\gamma) \right\}$$

ein zuverlässiger Vertrauensbereich