Statistical Data Analysis – Revision Lec. 2

Goodness of fit / significance test

(→ p-value)

p_H = Prob (data in region of equal or lesser compat. w/ H | H)

↖ more compat. w/ some H'

$\neq P(H)$

data values where p-value

$p_H \leq \alpha$ = a critical region for a test of H of size α

If $p_H \leq \alpha$ $\Rightarrow$ reject H.

Parameter estimation

given $x \sim f(x|\theta)$

construct $\hat{\theta}(\vec{x})$ = estimator for θ

↖ is a r.v.

$\hat{\theta} \sim g(\hat{\theta}|\theta)$

bias $b = E[\hat{\theta}] - \theta$

variance $V[\hat{\theta}]$

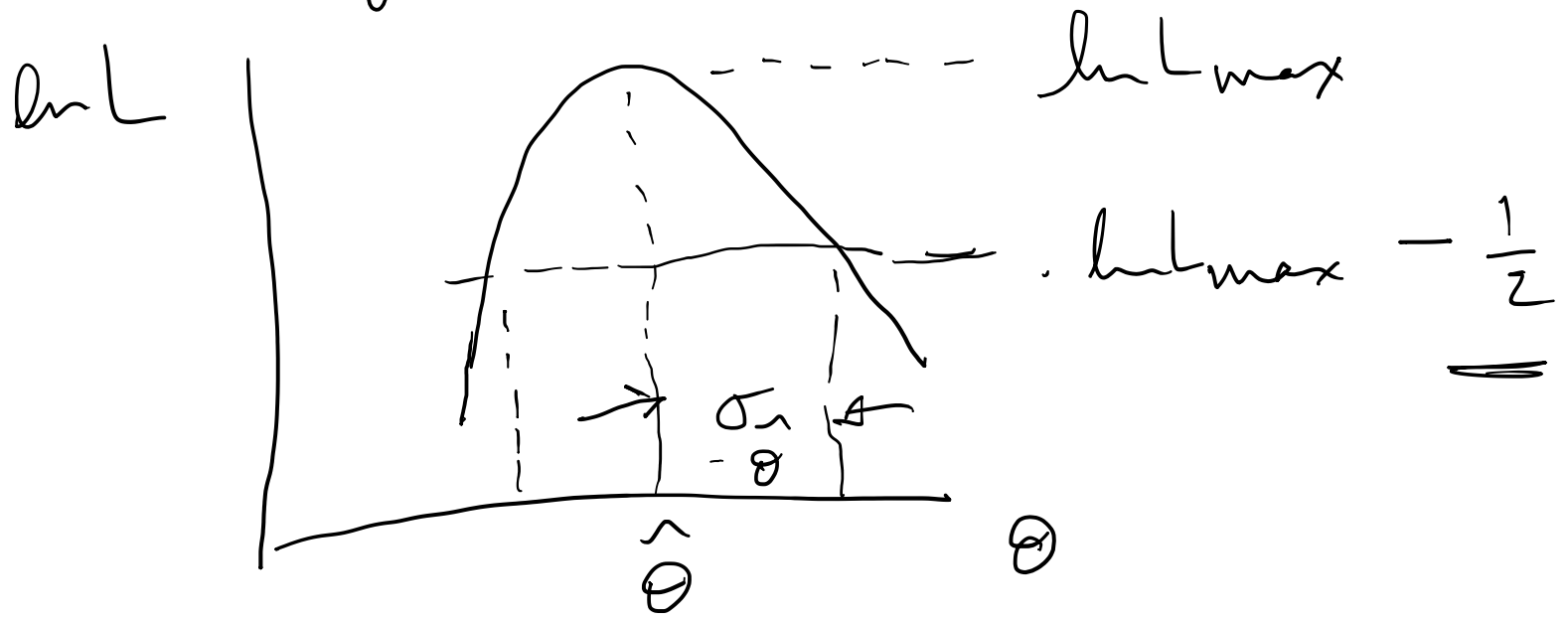
} competing measures of quality

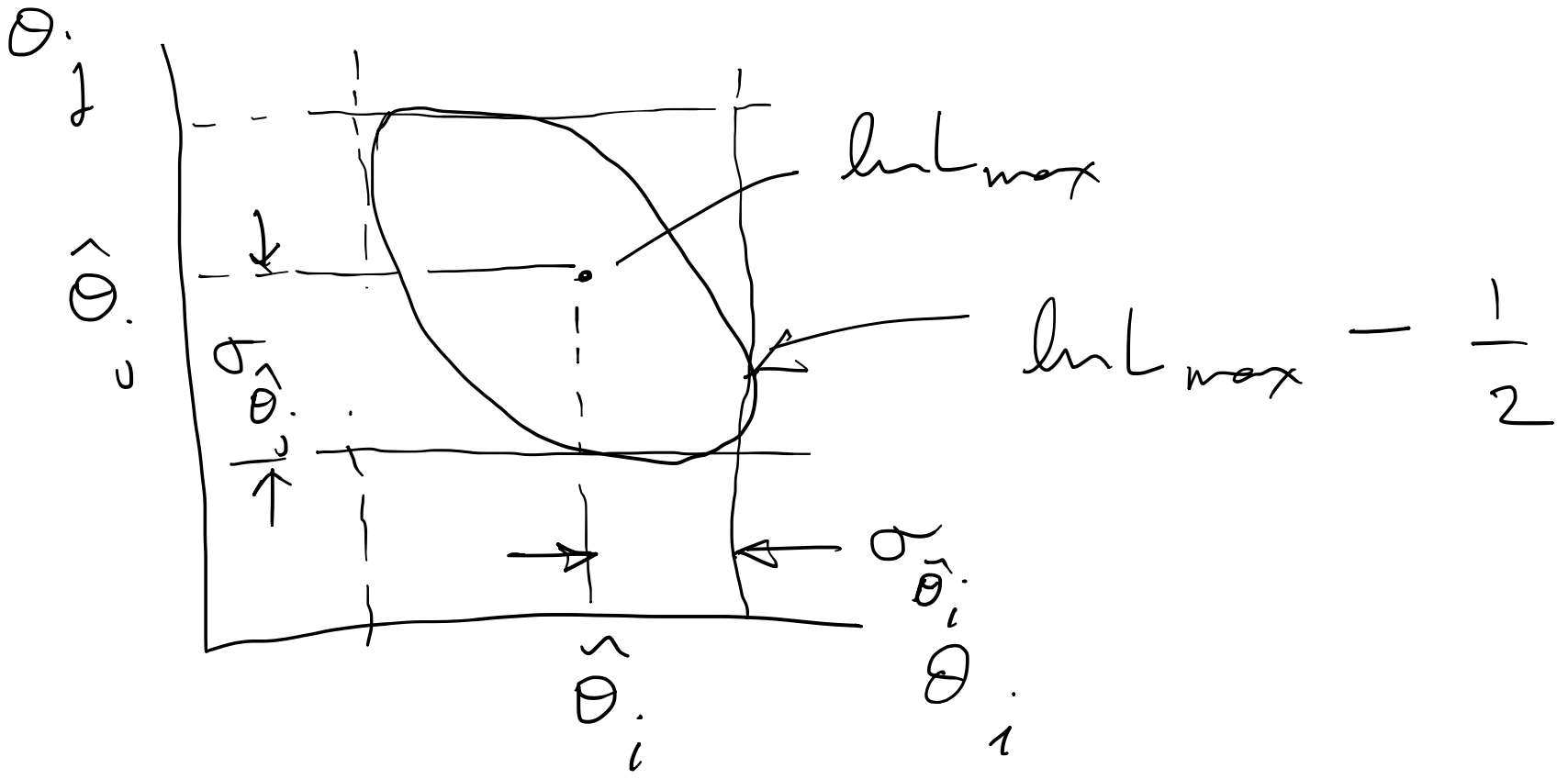
Likelihood $L(\theta) = P(x \mid \theta)$

↑ fixed.

Often use $\ln L(\theta)$

Method of max. likelihood: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$




θ_j
$\hat{\theta}_j$
$\sigma_{\hat{\theta}_j}$
$\ln L_{max}$
$\ln L_{max} - \frac{1}{2}$
$\sigma_{\hat{\theta}_i}$
$\hat{\theta}_i$
θ_i

Information inequality — often b small

$$V[\hat{\theta}] \geq \text{MVB} = -\frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

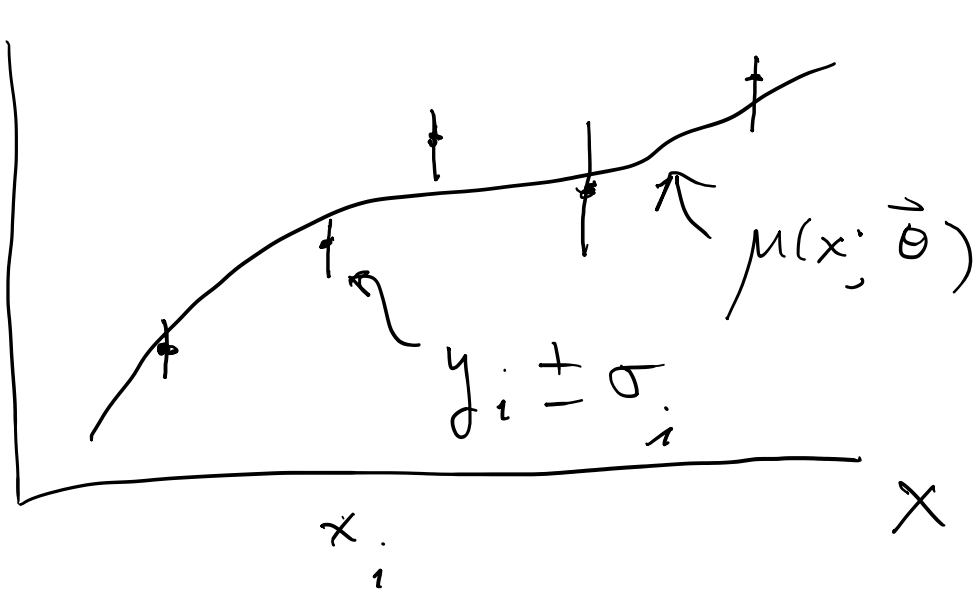
$\uparrow$ approx. equality

$$V[\hat{\theta}] \approx -\frac{1}{E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]}, \qquad \hat{V}[\hat{\theta}] \approx -\frac{1}{\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)_{\hat{\theta}}}$$

Least Squares

$y_i \sim \text{Gauss}(\mu(x_i;\vec{\theta}), \sigma_i)$

↖ indep.



$$-2\ln L(\vec{\theta}) = \underbrace{\sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\vec{\theta}))^2}{\sigma_i^2}}_{=\chi^2(\vec{\theta})} + C$$

Extended ML

$$n \sim \text{Poisson}(\nu)$$

$$x_1, \ldots, x_n \qquad x \sim f(x; \vec{\theta}) \Big\} \text{ indep.}$$

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i; \vec{\theta})$$

↑ extended

Interval estimation

Freq. int est.: Test each value of θ (size α)

Set of θ values not reject $\equiv$

confidence int for θ @ CL $= 1-\alpha$

$\Uparrow$ $p_\theta > \alpha$

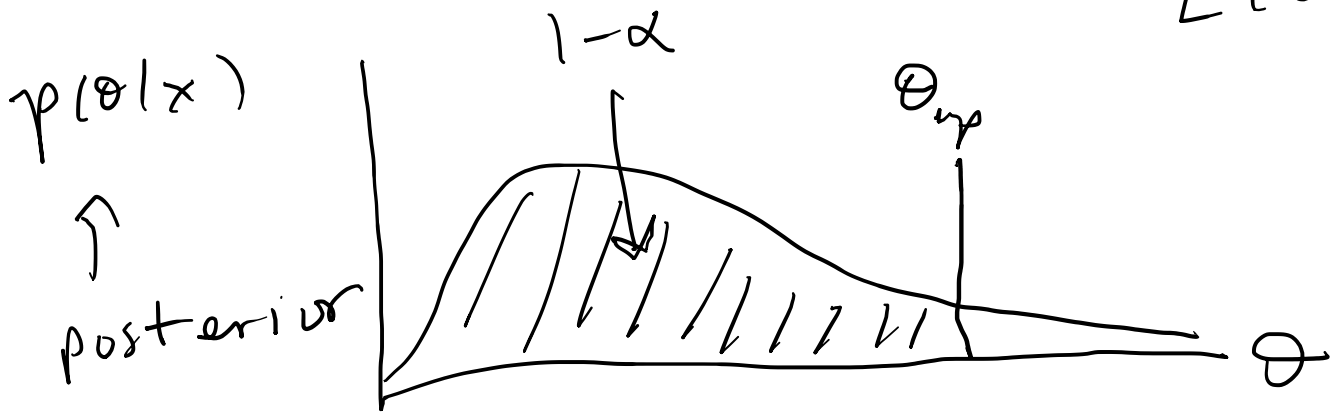
To find boundary of conf. int

set $p_\theta = \alpha$ & solve for θ.

Bayesian interval est.

$$p(\theta|x) \propto P(x|\theta)\,\pi(\theta)$$

$$\pi(\theta) \propto L(\theta)$$



$$\int_{-\infty}^{\theta_{up}} p(\theta|x)\,d\theta = 1-\alpha \quad \leftarrow \text{solve for } \theta_{up}$$

More on intervals (multi param.)

Wilks' thm

$\vec{\theta} = (\theta_1, \ldots, \theta_n) \rightarrow t_{\vec{\theta}} = -2 \ln \dfrac{L(\vec{\theta})}{L(\hat{\vec{\theta}})}$

→ in "asymptotic limit", $f(t_{\vec{\theta}} | \vec{\theta})$ is χ^2_n ← d.o.f.

↑ large sample, $\hat{\vec{\theta}} \sim$ Gauss

⇒ multiparam conf. region

$\ln L(\vec{\theta}) = \ln L_{\max} - \frac{1}{2} F^{-1}_{\chi^2_n}(1-\alpha)$

← = # of param.

Special case $n = 1$, $1 - \alpha = 68.3\%$

$$F^{-1}_{\chi^2_1}(0.683) = 1$$

$$\ln L(\theta) = \ln L_{\max} - \frac{1}{2}$$

$$\left[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}\right]$$ is a 68.3% CL conf. int.

Bayes factors

$$B_{ij} = \frac{L_i}{L_j} = \frac{\int p(x|H_i,\theta)\,\pi(\theta)\,d\theta}{\int p(x|H_j,\eta)\,\pi(\eta)\,d\eta}$$

marginal likelihoods

$$B_{ji} = \frac{1}{B_{ij}}$$

hard to compute

use e.g. harmonic mean importance sampling

2015 Exam #4

$n \sim \text{Poisson}(\mu + \nu) \qquad \mu + \nu \geq 0$

a) suppose ν known

$$L(\mu) = P(n|\mu) = \frac{(\mu+\nu)^n}{n!} e^{-(\mu+\nu)}$$

$$\ln L(\mu) = n \ln(\mu+\nu) - (\mu+\nu) + C$$

↑ doesn't depend on params.

b) $\frac{\partial \ln L}{\partial \mu} = \frac{n}{\mu+\nu} - 1 \overset{\text{set}}{=} 0$

$\Rightarrow \quad \hat{\mu} = n - \nu$

$E[\hat{\mu}] = E[n-\nu] = E[n] - \nu = \mu + \nu - \nu = \mu$

$\Rightarrow \quad b = E[\hat{\mu}] - \mu = 0$

$V[\hat{\mu}] = V[n-\nu] = V[n] = E[n] = \mu + \nu$ (for Poisson)

$\sigma_{\hat{\mu}} = \sqrt{\mu+\nu}$

c) Test $\mu = 0$ w/ $t_0 = -2 \ln \lambda(0)$

$$\lambda(\mu) = \frac{L(\mu)}{L(\hat{\mu})}$$

$\hat{\mu} = n - \nu$

For $n \neq 0$: $t_0 = -2\left[\ln L(0) - \ln L(\hat{\mu})\right]$ ($\hat{\mu} = n-\nu$)

$$= -2\left[n \ln \nu - \nu - n \ln(n - \nu + \nu) + n - \nu + \nu\right]$$

$$= -2\left[n \ln \frac{\nu}{n} + n - \nu\right]$$

$n = 0$, $L(n=0|\mu) = \frac{(\mu+\nu)^0}{0!} e^{-(\mu+\nu)}$

$$\ln L(n=0|\mu) = -(\mu+\nu)$$

$$t_0(n=0) = -2\left[\ln L(n=0|\mu=0) - \ln L(n=0|\mu=\hat{\mu}=-\nu)\right]$$

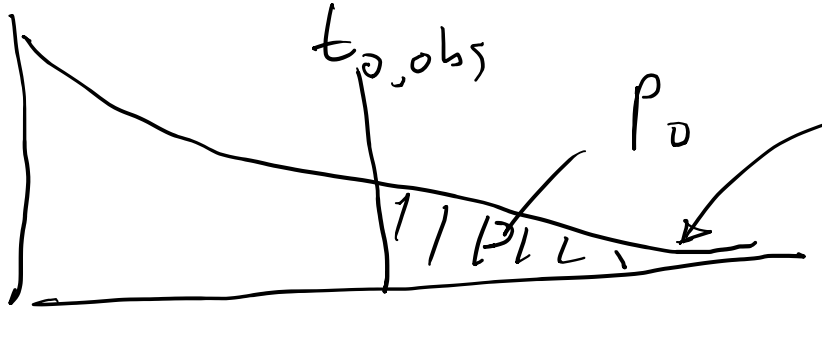
$$= -2\left[-(0+\nu) + (-\nu+\nu)\right]$$

$$= 2\nu$$ ✓

c – ii)

$$p_0 = P(t_0 \geq t_{0,obs} \mid \mu = 0)$$

$$= 1 - \int_0^{t_{0,obs}} f(t_0 \mid \mu = 0)\, dt_0$$



$f(t_0|0) = \chi^2_1$ assuming Wilks.

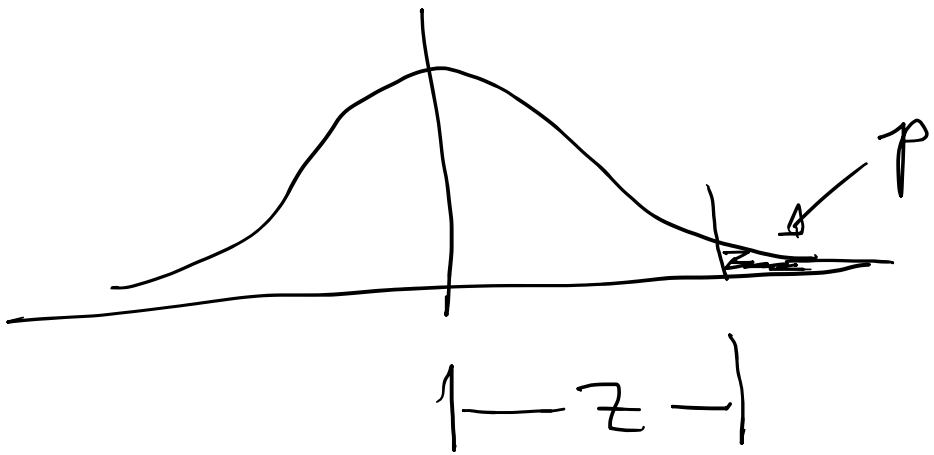
$$p_0 = 1 - F_{\chi^2_1}(t_0)\Big|_{obs}$$

Significance

$$p = 1 - \Phi(z)$$

$$z = \Phi^{-1}(1-p)$$

$$= \Phi^{-1}\left(F_{\chi^2_1}(t_0)\right)$$



d) ν = nuisance param.

$y \sim \text{Gauss}(\nu, \sigma)$ ← indep of μ

τ known

$$L(\mu, \nu) = P(n, y | \mu, \nu)$$

$$= \frac{(\mu+\nu)^n}{n!} e^{-(\mu+\nu)} \times \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(y-\nu)^2/2\sigma^2}$$

$$\ln L(\mu, \nu) = n \ln(\mu+\nu) - (\mu+\nu) - \frac{1}{2}\frac{(y-\nu)^2}{\sigma^2} + C$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{n}{\mu+\nu} - 1 \overset{\text{set}}{=} 0$$

$$\frac{\partial \ln L}{\partial \nu} = \left(\frac{n}{\mu+\nu} - 1\right) + \frac{y-\nu}{\sigma^2} \overset{\text{set}}{=} 0$$

} solve for $\mu \neq \nu$

$$\hat{\mu} = n - y$$

$$\hat{\nu} = y$$

$$V[\hat{\mu}] = V[x - y] = V[x] + V[y]$$

$$= \mu + \nu + \sigma^2$$

$$V[\hat{\nu}] = V[y] = \sigma^2$$

$$\mathrm{cov}[\hat{\mu}, \hat{\nu}] = \mathrm{cov}[x - y, y] = \mathrm{cov}[x, y] - \mathrm{cov}[y, y]$$

($\mathrm{cov}[x, y] = 0$; $\mathrm{cov}[y, y] = \sigma^2$)

$$= -\sigma^2$$

$$\rho_{\hat{\mu}\hat{\nu}} = \frac{\mathrm{cov}[\hat{\mu}, \hat{\nu}]}{\sigma_{\hat{\mu}} \sigma_{\hat{\nu}}}$$

$$= -\frac{\sigma}{\sqrt{\mu + \nu + \sigma^2}}$$