Classical estimators

- **In classical statistics**, we have *estimators* for parameters. These are functions of data (=X).
  - e.g. observed sample mean

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

    is an estimator for population (true) mean θ=E(X), or sample variance s^2 is an estimator for σ^2 = Var(X).
  - Parameter is thought fixed but unknown.
  - Data X are random, therefore estimator is random.
  - Uncertainty about θ is *'indirectly'* described by distribution of X, even after observing X.

Posterior summaries

- **In Bayes**: posterior density describes *directly* our uncertainty about the unknown parameter θ, after observing data X.
  - Observed data X are fixed (=evidence).
  - Parameter is random, because it is *still* uncertain.
  - Probability is a measure of uncertainty, posterior distribution is complete description for that, given the X we had.
  - Mode = the 'most probable' value.
  - Mean = expected value, if you'd make a bet.
  - Median = with 50% probability, it's below this.

Posterior summaries

- **The best summary of a posterior distribution is to show the distribution itself graphically.**
- BUT: in practice, you often need to report just numbers in a table.
- Also: multidimensional distribution cannot be shown in a figure anyway.
- How to summarize a distribution?
  → *this can be seen as a decision problem…*

Posterior summaries

- **Comparison of mean, median, mode**:
  - Define a loss function $L(\theta,\delta_x)$ to describe the loss due to estimating θ by point estimate δ_x based on data x.
  - For any x, choose δ_x to minimize the posterior loss

$$E(L(\theta,\delta_x)\,|\,x) = \int L(\theta,\delta_x)\,p(\theta\,|\,x)\,d\theta$$

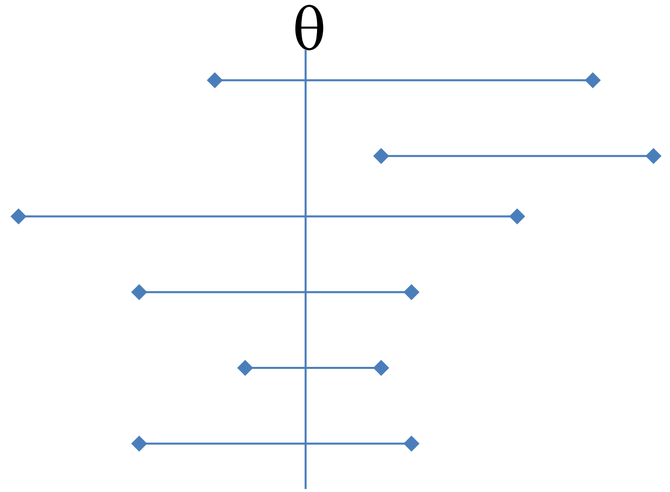
  - If the loss function is quadratic $L(\theta,\delta_x)=(\theta-\delta_x)^2$ then the posterior loss becomes $V(\theta|X)+(E(\theta|X)-\delta_x)^2$ which is minimized by choosing $\delta_x = E(\theta|X)$, that is, the posterior mean.

Posterior summaries

- But if our loss function is $L(\theta,\delta_x)=|\theta-\delta_x|$ then we should choose δ_x = posterior median, to minimize posterior loss (for any x).
- And if $L(\theta,\delta_x)= 1_{\{\theta\neq\delta x\}}(\delta_x)$ "all-or-nothing error", then the choice would be posterior mode.
- E.g. if you prefer choosing posterior mean, this means that you **behave as if** you had a quadratic loss function.
- No point value can fully convey all information contained in a posterior distribution.

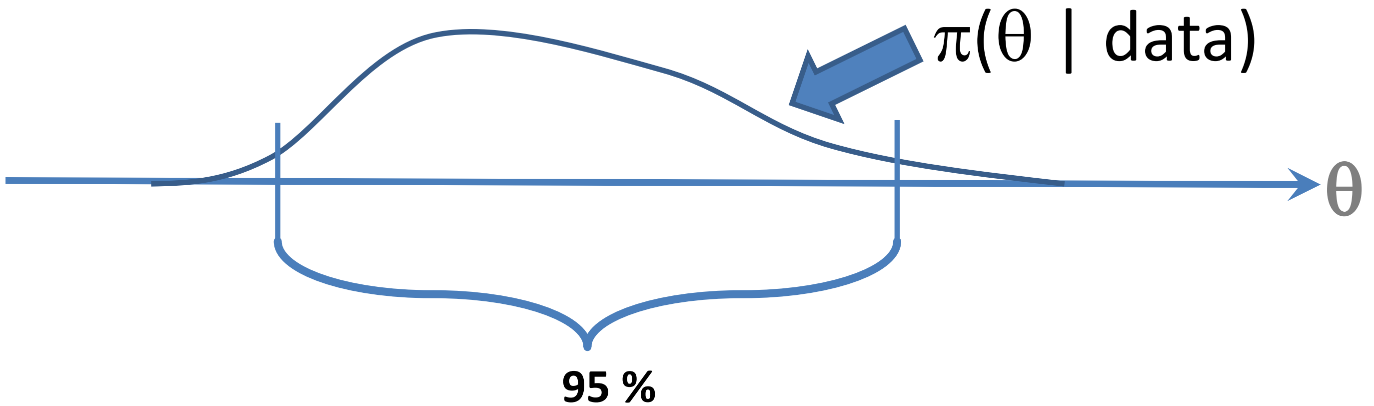
Intervals instead of point estimates

- Classical 95% Confidence Interval?
  - **In classical statistics**: confidence interval (luottamusväli) is a function of data, therefore random.
  - With 95% frequency, the interval will cover the true parameter value, in the long run. (If the experiment is repeated). i.e. we are 95% **confident** of this.



Intervals from posterior distribution

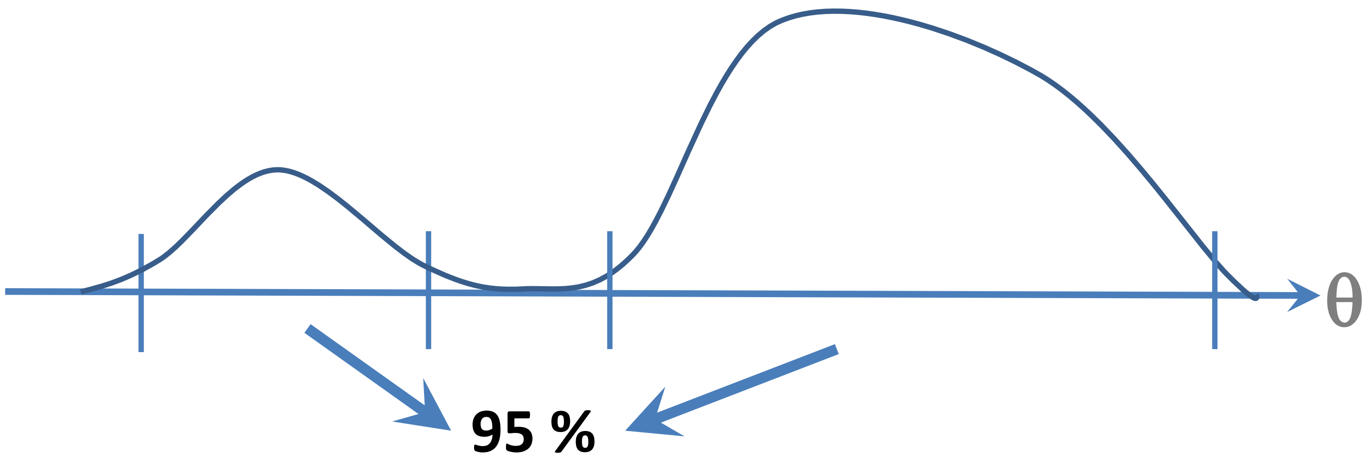
- 95% Credible Interval. (AKA credibility interval, probability interval, todennäköisyysväli, luottoväli)
  - **In Bayes**: credible interval is an interval in which the parameter is with 95% probability, **given the actual data we now had.**



  - **Can choose 95% interval in many ways, though.**

Intervals from posterior distribution

- 95% Credible Interval.
  - Posterior density can be bimodal or multimodal.
  - CI does not need to be a connected set.



  - A shortest possible interval with a given probability is **Highest Posterior Density Interval**

Posterior summaries

- Mode, mean, median are often readily available if we have analytic solution as a standard density (when using conjugate priors)
  - If we need to approximate the density by Monte Carlo sample (or MCMC), then mean ≈ sample mean, median ≈ sample median, but mode is less straightforward. (Highest point of histogram, or estimated density).
    - Also HPD-interval needs more effort then.

Posterior summaries

- Summaries improve with more data:
  - With little data → posterior is dictated by prior
  - With enough data → posterior is dictated by data
  - Savage: ”When they have little data, scientists disagree and are subjectivists; when they have piles of data, they agree and become objectivists”.

  - **$V(\theta) = E(V(\theta \mid X)) + V(E(\theta \mid X))$** which means that posterior variance **$V(\theta \mid X)$** is *expected* to be smaller than the prior variance $V(\theta)$. (But sometimes it can increase).