Posterior summaries

- **In classical statistics**, we have *estimators* for parameters. These are functions of data, e.g. mean of observations, or sample variance.
  - Parameter is thought fixed but unknown.
  - Data is random, therefore estimator is random.

Posterior summaries

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

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 $\delta_x = E(\theta|X)$, the posterior mean.

Posterior summaries

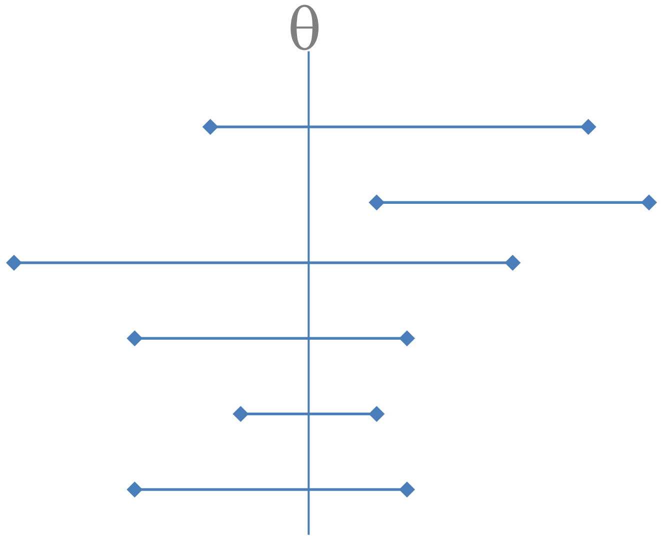
- But if our loss function is $L(\theta,\delta_x)=|\theta-\delta_x|$ then we should choose δ_x = posterior median, to minimize posterior loss (for any x).

- And if $L(\theta,\delta_x)= 1_{\{\theta=\delta x\}}(\delta_x)$ ”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 the complete information contained in a posterior distribution.

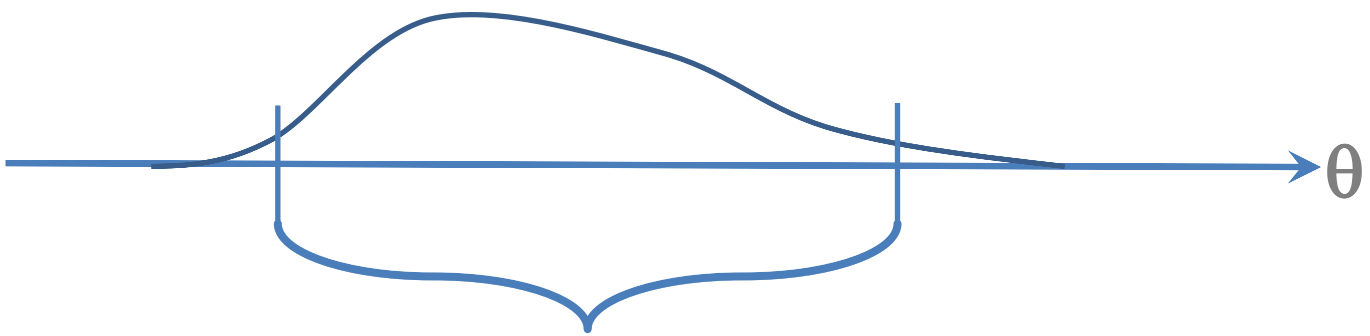
Posterior summaries

- Compare: classical 95% Conf. Interval?
  - **In classical statistics**: confidence interval 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.



Posterior summaries

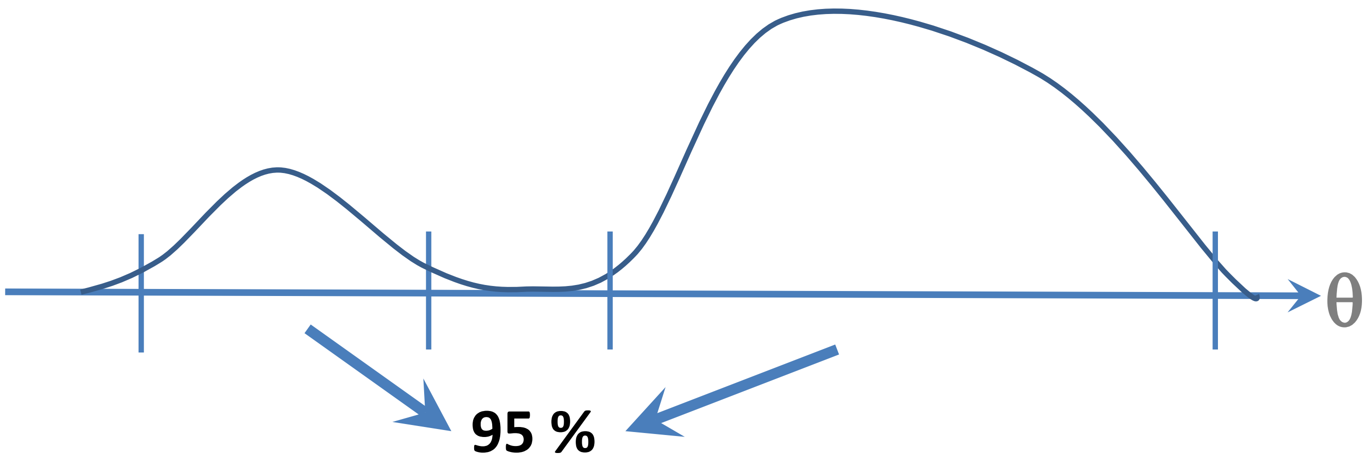
- 95% Credible interval.
  - **In Bayes**: credible interval is an interval in which the parameter is with 95% probability, **given this actual data we now had.**



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

Posterior summaries

- 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

- Generally:
  - 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).

Further use of posteriors

- **Hypotheses:**
  - About a parameter: ”$\theta<0$”
  - Compute $P(\theta < 0 \mid X)$, the cumulative density at 0.

  - $P(H_0 \mid X)$ and $P(H_1 \mid X)$ possible to compute if ”H” is a region of parameter space .
  - We do not reject or accept a H, just calculate its probability, given evidence.

Further use of posteriors

- **Hypotheses:**
  - Sometimes used: posterior odds $P(H_0 \mid X)/P(H_1 \mid X)$.
  - If ”>1”, shows support for H_0.
  - **Bayes factor**: a ratio of prior and posterior odds
  - $BF = [\ P(H_0 \mid X)/P(H_1 \mid X)\ ] \ / \ [\ P(H_0)/P(H_1)\ ]$

    $= [P(H_0 \mid X)\ P(H_1)\ ] \ / \ [P(H_1 \mid X)\ P(H_0)]$

Posterior odds = Prior odds x BF

This is a different way of expressing Bayes theorem:
BF expresses how much data change prior odds.

Further use of posteriors

- **Hypotheses:**
  - A point hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$
  - We must have positive probability $P(H_0)=1-P(H_1)$
  - The BF then becomes the same as 'likelihood ratio'

$$\frac{P(\theta = \theta_0 \mid X)}{P(\theta = \theta_1 \mid X)} = \frac{P(\theta = \theta_0)}{P(\theta = \theta_1)} \boxed{\frac{p(X \mid \theta = \theta_0)}{p(X \mid \theta = \theta_1)}}$$

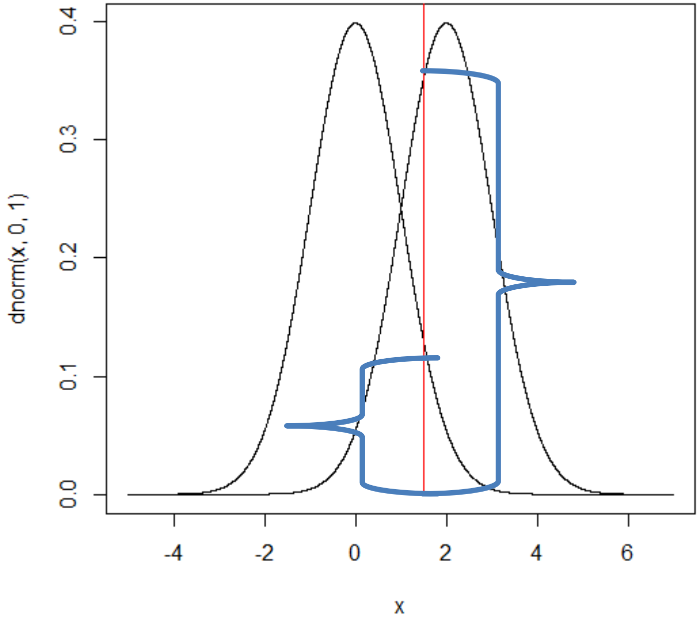
  Because constant p(X) cancels out.

  - But: how big (small) BF is big (small) enough ?
  - Composite hypothesis, one-sided, two-sided…

Further use of posteriors

- **$X \sim N(\theta,1)$, data: $X=1.5$**
  - A point hypothesis $H_0 : \theta = 0$ against $H_1 : \theta = 2$.
  - Assume prior $p(\theta=0)=p(\theta=2)=0.5$
  - Then, posterior odds = likelihood ratio.
  - Conversion to probability :

    p = 1/(1+odds)



$$\frac{P(\theta=\theta_0 \mid X)}{P(\theta=\theta_1 \mid X)} = \frac{P(\theta=\theta_0)}{P(\theta=\theta_1)} \boxed{\frac{p(X \mid \theta=\theta_0)}{p(X \mid \theta=\theta_1)}}$$

Further use of posteriors

- **Predictions:**
  - It is rather easy to compute predictive distribution of X based on **given parameters θ** and the model $P(X|\theta)$. And likewise for any function $g(X)$.
    - Assuming you can generate samples from $P(X|\theta)$.
    - This would not take into account the uncertainty about parameters θ.
  - Aim: to compute posterior predictive distribution $\mathbf{P(X_{new} \mid X_{obs})}$
  - This gives prediction based on the past data, not based on assumed parameter estimates.

Predictive distributions

- Consider series of observations: $X_1,\ldots,X_n$ and a model $p(X_i \mid \theta)$ so that X_i are conditionally independent, given θ.

  Posterior predictive distribution of X_{n+1} :

$$p(X_{n+1} \mid X_1,\cdots,X_n) = \int p(X_{n+1},\theta \mid X_1,\cdots,X_n)d\theta$$

$$= \int p(X_{n+1} \mid \theta, X_1,\cdots,X_n)\,p(\theta \mid X_1,\cdots,X_n)d\theta$$

$$= \int \underbrace{p(X_{n+1} \mid \theta)}_{\text{Our model}}\,\underbrace{p(\theta \mid X_1,\cdots,X_n)}_{\text{Posterior of } \theta}d\theta$$

Predictive distributions

- **Likewise:**
  Prior predictive distribution of X_{n+1} :

$$p(X_{n+1}) = \int p(X_{n+1}, \theta) d\theta = \int \underbrace{p(X_{n+1} \mid \theta)}_{\text{Our model}} \underbrace{p(\theta)}_{\text{Prior of } \theta} d\theta$$

- *"With the predictive approach parameters diminish in importance, especially those that have no physical meaning. From the Bayesian viewpoint, such parameters can be regarded as just place holders for a particular kind of uncertainty on your way to making good predictions".* (Draper 1997, Lindley 1972).

Predictive distributions

- Note also, directly from Bayes:

$$p(X) = \frac{p(X \mid \theta)\, p(\theta)}{p(\theta \mid X)}$$

- by inserting *prior*, *posterior*, *model of X*, we find prior predictive density of X.
- Similarly,

$$p(X_{n+1} \mid X_1, \cdots, X_n) = \frac{p(X \mid \theta, X_1, \cdots, X_n)\, p(\theta \mid X_1, \cdots, X_n)}{p(\theta \mid X_1, \cdots, X_{n+1})}$$

Predictive distributions

- **Let's try with binomial model.**
  - **Assume we have a posterior which is beta(α,β).**
  - **'Old data' is then included in α,β.**

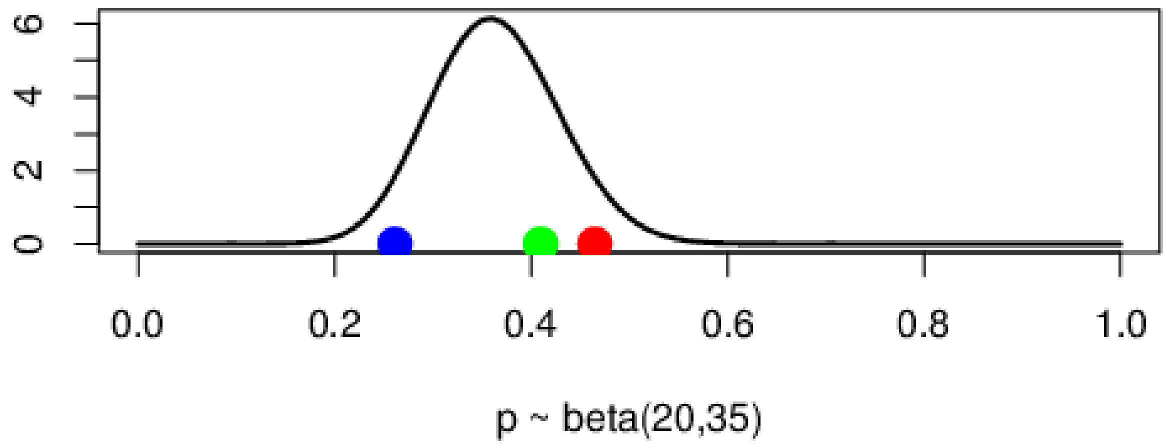
$$p(X \mid \alpha,\beta) = \int p(X,\theta \mid \alpha,\beta)\,d\theta = \int \underbrace{p(X \mid \theta)}_{\text{Binomial}(N,\theta)}\,\underbrace{p(\theta \mid \alpha,\beta)}_{\text{Beta}(\alpha,\beta)}\,d\theta$$

  - **This can be solved as:**

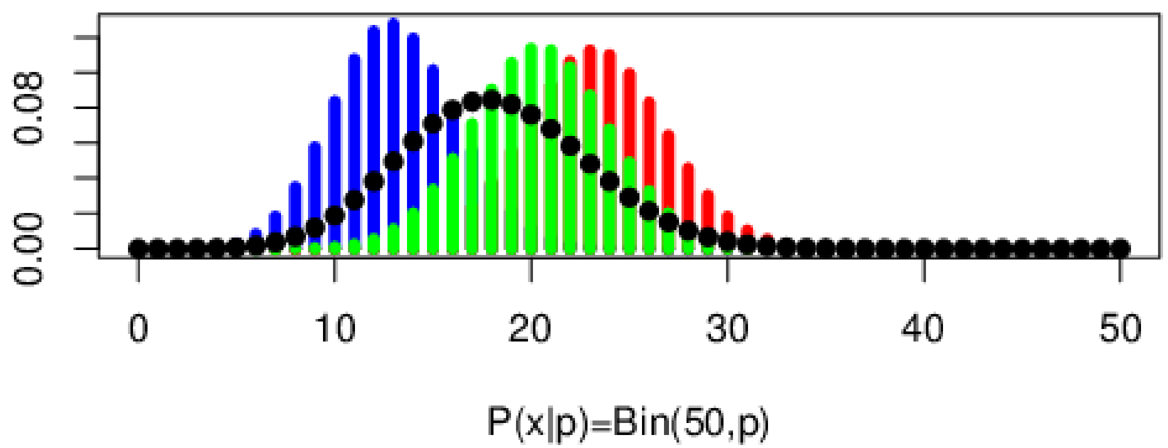
$$p(X \mid \alpha,\beta) = \binom{N}{X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

A = X+α, B = N-X+β

A BETA-BINOMIAL distribution.



6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0
p ~ beta(20,35)



0.08
0.00
0
10
20
30
40
50
P(x|p)=Bin(50,p)

Predictive distributions

- **With Poisson model:**
  - **Assume we have a prior which is gamma(α,β).**
  - **If posterior, 'old data' is included in α,β.**

$$p(X \mid \alpha,\beta) = \int p(X,\lambda \mid \alpha,\beta)\, d\theta = \int \underbrace{p(X \mid \lambda)}_{\text{Poisson}(\lambda)}\, \underbrace{p(\lambda \mid \alpha,\beta)}_{\text{gamma}(\alpha,\beta)}\, d\lambda$$

- The solution is **NEGATIVE BINOMIAL distribution:**

$$p(X \mid \alpha,\beta) = \binom{\alpha + X - 1}{X} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{X}$$

Predictive distributions

- **To solve predictive means, variances:**
  - **Use $E(X) = E(E(X|\theta))$**
  - **Use $V(X) = E(V(X|\theta)) + V(E(X|\theta))$**

  - **For example, with Poisson + Gamma:**
  - **$E(X) = \alpha/\beta$**
  - **$V(X) = \alpha/\beta + \alpha/\beta^2$**

  - **By including parameter uncertainty $p(\theta)$ to a model $p(X|\theta)$ we get models $p(X) = \int p(X|\theta)p(\theta)d\theta$, suitable for e.g. overdispersed data.**