Confidence vs credibility; consistency; CLT

Christopher A. Sims
Princeton University
sims@princeton.edu

September 29, 2020

Confidence vs credibility

- Last lecture we discussed how to generate a credible set, and an HPD credible set in particular.

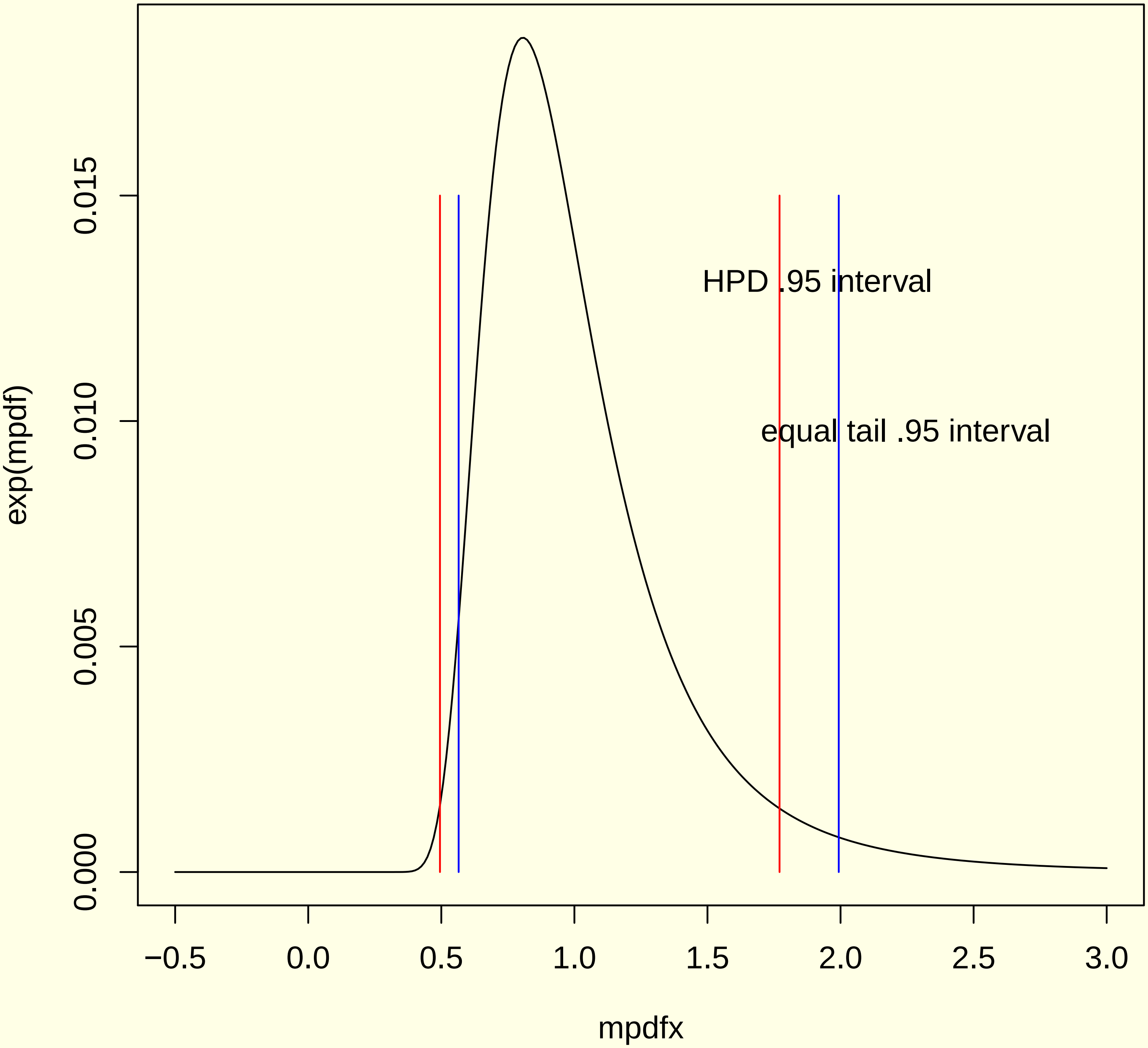
- With a one-dimensional parameter, one can also form an ”equal tail” interval. For a 95% credible set, this just puts 2.5% of the probability to the left of the interval and 2.5% to the right.

- The equal tail interval is always at least as long as the HPD interval,

- The density function always has the same value on the boundaries of the HPD set.

- the HPD “interval” can have multiple disjoint components.

HPD and equal tail intervals for N=5, normal(mu,mu^2) model



Confidence sets

- These are mapppings from the observable data X to subsets $C(X) \subset S$, where S is the parameter space.

- $C(X)$ is a 95% **confidence set** for $\beta \in S$ iff

$$P[\beta \in C(X) \mid \beta] \geq .95$$

  for all $\beta \in S$. A $100(1-\alpha)$ percent confidence set has at least $1-\alpha$ **coverage probability**, for any $\beta \in S$.

Pre-sample probabilities

- Here there is no probability distribution on β. Only X is random. The .95 probability describes randomness in X *before* we see the data.

- Once we see X, and hence $C(X)$, there is no randomness. The probably that β is in $C(X)$, after we see X, is zero or one. Of course we don't know which it is, but this is just an unknown function of the unknown, non-random β.

- These frequentist probabilities do not describe our uncertainties about β, in other words.

How do we generate confidence sets?: a) Collecting tests

- If for every point β in a parameter space S we have a .05 level test for the point null hypothesis that β is the true parameter value, then we can get a 95% confidence interval as

$$C(X) = \{\beta \in S \mid \beta \text{ not rejected}\}\ .$$

- We can reverse this: If we have a 95% confidence set $C(X)$, we can turn it in to a test of β as a point null by rejecting when β is not in $C(X)$
- While this always works, unless we give careful attention to the alternative hypotheses in the tests, the results can be unhelpful.

How to generate confidence sets: b) pivotal statistics

- Suppose we can find a function $f(X, \beta)$ of the data and the unknown parameter whose distribution does not depend on the unknown parameter.

- This is known as a “pivot” or “pivotal statistic”. (It’s not a statistic, by the usual definition, since it is not a function of the data alone.)

- Example: In our $X \sim N(\mu, \mu^2)$ example, X/μ or, in an i.i.d. sample, $\bar{x}/\mu$, are pivots, since their distributions are $N(1, 1)$ or $N(1, 1/N)$, respectively.

Using the pivot

- Find a set A such that the probability the pivot is in A is .95.

- Then set

$$C(X) = \{\beta \in S \mid f(X,\beta) \in A\} \ .$$

- Example: In the $N(\mu,\mu^2)$ case with an i.i.d. sample, assuming $N > 1.96^2$,

$$P\left[\sqrt{N}\left(\frac{\bar{x}}{\mu} - 1\right) \in (-1.96, 1.96)\right] = .95$$

$$\therefore P\left[\mu \in \left(\frac{\bar{x}}{1 + 1.96/\sqrt{N}}, \frac{\bar{x}}{1 - 1.96/\sqrt{N}}\right)\right] = .95$$

Connection of confidence sets to credible sets

- In general there need be no connection.

- Confidence sets with high coverage probability must have a high posterior probability with high pre-sample probability.

- Credible sets with high posterior probability must have high coverage probability for a set of β's with high prior probability.

- But confidence sets can, after we see the data, be clearly unreasonable as credibility sets under any prior.

Bad confidence sets

- Suppose $X \sim N(\mu, 1)$ and we know that $\mu \in (0, 1)$. Then $X \pm 1.96$ is a 95% confidence interval for μ. It remains a 95% confidence set if we use its intersection with $(0, 1)$. (We're making the confidence set much smaller; why is it still a 95% interval?)

- But regardless of μ's value, there is some probability of observing $X < -1.96$ or $X > 2.96$. In these cases, the 95% confidence set is empty.

- This is OK, for its confidence level, since these events have low pre-sample probabiiity.

- But obviously no 95% credible set will be empty, and I've never seen an applied paper that reports an empty confidence set.

Bettable confidence sets

- Suppose a $1-\alpha$ confidence set for β has exactly $1-\alpha$ coverage probability for all $\beta \in S$.

- Now imagine that someone, after each sample draw of X has the option to bet either that the true β is in $C(X)$ or that it is not, with odds 19 to 1 if she's betting it's not in $C(X)$, 1 to 19 if she's betting it is in $C(X)$.

- If she can make positive expected returns, the confidence interval is bettable. It is not bettable if and only if it is also a 95% credible set for some prior.

Norets and Mueller

- A paper by these authors, referenced on the course syllabus, studies the situation where the confidence set does not have uniform coverage, but instead has probability $1-\alpha$ or above for every β.

- Then, instead of allowing the bettor to bet either way, for or against the truth being in $C(X)$, after seeing X, the better is allowed only to bet against the confidence set, but with the option of not betting at all.

- In this case, if the bettor cannot make money, the $1-\alpha$ confidence set $C(X)$ must correspond to a $1-\alpha$ credible set $B(X)$ such that $B(X) \subset C(X)$ for every X.

Confidence sets for functions of the parameter vector

- Suppose $f(\beta)$ is a function of β that has the same value at more than one value of β.

- The leading example of this is a single element of the parameter vector: $f(\vec{\beta}) = \beta_1$, for example, with $\vec{\beta} \in \mathbb{R}^k$, $k > 1$, for example.

- Then the technique of collecting all the non-rejected β's under point-null tests does not work, because the set of β's at which $f(\beta)$ has a particular value includes many $\vec{\beta}$'s, each of which is likely to imply a different coverage probability.

Credible sets for functions of the parameter vector

- Forming credibile sets for functions $f(\beta)$ is in principle straightforward. Analytically, it requires working out a change of variables with Jacobian term, then integrating out all the components of the tranformed parameter except $f(\beta)$.

- Usually, though, it is easier to generate simulated draws $\{\beta_i\}$ from the posterior distribution, calculating $f(\beta_i)$ for these draws, and then constructing an HPD set from those draws.

Strong law of large numbers

$$\frac{\sum_{i=1}^{N} X_i}{N} \xrightarrow[N\to\infty]{a.s.} E[X_i]\,.$$

We are assuming that all the X_i variables have the same, finite expectation, and that they are i.i.d. with the same distribution. No other assumptions are required, and the result can be proved with much weaker assumptions.

The definition of "a.s." (for "almost sure") convergence of a sequence $\{X_i\}_{i=1}^{\infty}$ of random variables to another random variable (or constant) Y is that $P[X_i \xrightarrow[i\to\infty]{} Y] = 1$, where the "$\rightarrow$" inside the brackets is ordinary convergence of real numbers.

Consistency

- If β is an unknown parameter and $\left\{\hat{\beta}_N\right\}$ is a sequence of estimators, based on samples of size N, of β such that $\hat{\beta}_N \xrightarrow{a.s.} \beta$, we say that $\hat{\beta}_N$ is a **strongly consistent** estimator of β.

  =item Another type of convergence of random variables is $X_i \xrightarrow{P} Y$, which means that for every $\varepsilon > 0$, $P[|X_i - Y| > \varepsilon] \to 0$. This requires that X_i is likely to be close to Y for large i, but makes no claim about whether the full sequence $\{X_i\}$ converges to the realized value of Y. "$\xrightarrow{P}$" is called **convergence in probability**.

- $\hat{\beta}_N \xrightarrow{P} \beta$ means $\hat{\beta}_N$ is **weakly consistent**.

Example of convergence in probability that's not convergence a.s.

i.i.d. X_i with $P[X_i = 1] = 1 - 2^{-j}$ for i between 2^j and 2^{j+1}, with $X_i = 0$ when it is not equal to 1. X_i clearly converges in probability to 1, but within every $(2^j, 2^{j+1})$ interval of i values, it has 2^j chances to produce a zero. Since the probability of a zero at each draw of X_i in this interval of i's is 2^{-j}, there is a good chance that there will be a zero in the sequence, for every j. (In fact, for large j the probability of a zero in the j'th subsequence is very close to $1/e$.) Since for every j there is a probability bounded away from zero that X_i is zero, there is no chance that X_i gets close to 1 and stays there, as would be required for a.s. convergence. The probability that we will eventually see another $X_i = 0$ is always 1, at every i no matter how large.

The central limit theorem

There are many versions of this, relaxing its assumptions in various directions. An easily stated version that we can understand with our tools is:

Theorem. *If $\{X_i\}_{i=1}^N$ is an i.i.d. sequence of k-dimensional random vectors with mean zero and $k \times k$ covariance matrix Σ, then*

$$\frac{\sum_{i=1}^N X_i}{\sqrt{N}} \xrightarrow[N \to \infty]{D} N(0, \Sigma) \; .$$

Of course we still have to define what "$\xrightarrow{D}$" means.

Convergence in distribution

The sequence of random variables X_i converges in distribution to the distribution of the random variable Y if for every bounded, continuous function f, $E[f(X_i)] \to E[f(Y)]$.

For one-dimensional X_i, $X_i \xrightarrow{D} Y \Leftrightarrow F_{X_i}(a) \to F_Y(a)$ at all points of continuity of F_Y, where $F_Y[a] = P[Y \leq a]$ is the cdf (cumulative distribution function) of Y.

Relations among convergences

$$a.s. \Rightarrow P \Rightarrow D\ .$$

If f is continuous,

$$X_i \xrightarrow{Z} Y \Rightarrow f(X_i) \xrightarrow{Z} f(Y)\ ,$$

where Z is any of $a.s.$, P, or D.

Examples of convergence

Check for which (if any) of the types of convergence these X_i's converge to Y.

- $X_i = Y + \frac{1}{i}$

- $P[X_i = 1] = 1 - 1/i$; $P[X_i = 0] = 1/i$, $Y \equiv 1$. (What happens to $E[X_i]$ in this case?)

- X_i i.i.d. $N(0,1)$, independent of Y, which is also $N(0,1)$.

One more type of convergence

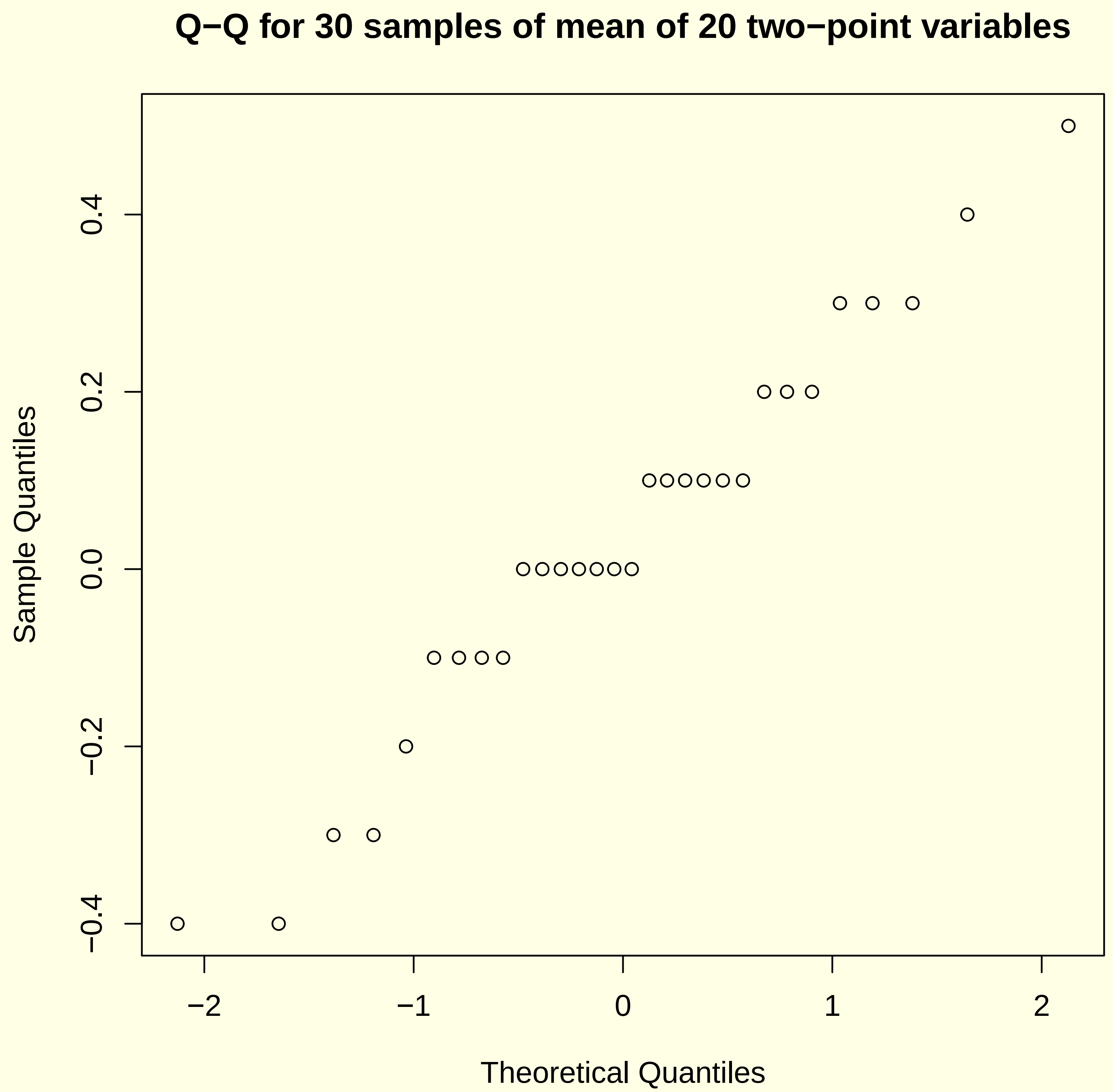
- $X_i \xrightarrow{q.m.} Y$ iff $E[(X_i - Y)^2] \to 0$. This is **quadratic mean** convergence. It implies, but is not implied by, convergence in probability, and neither implies, nor is implied by, $a.s.$ convergence.

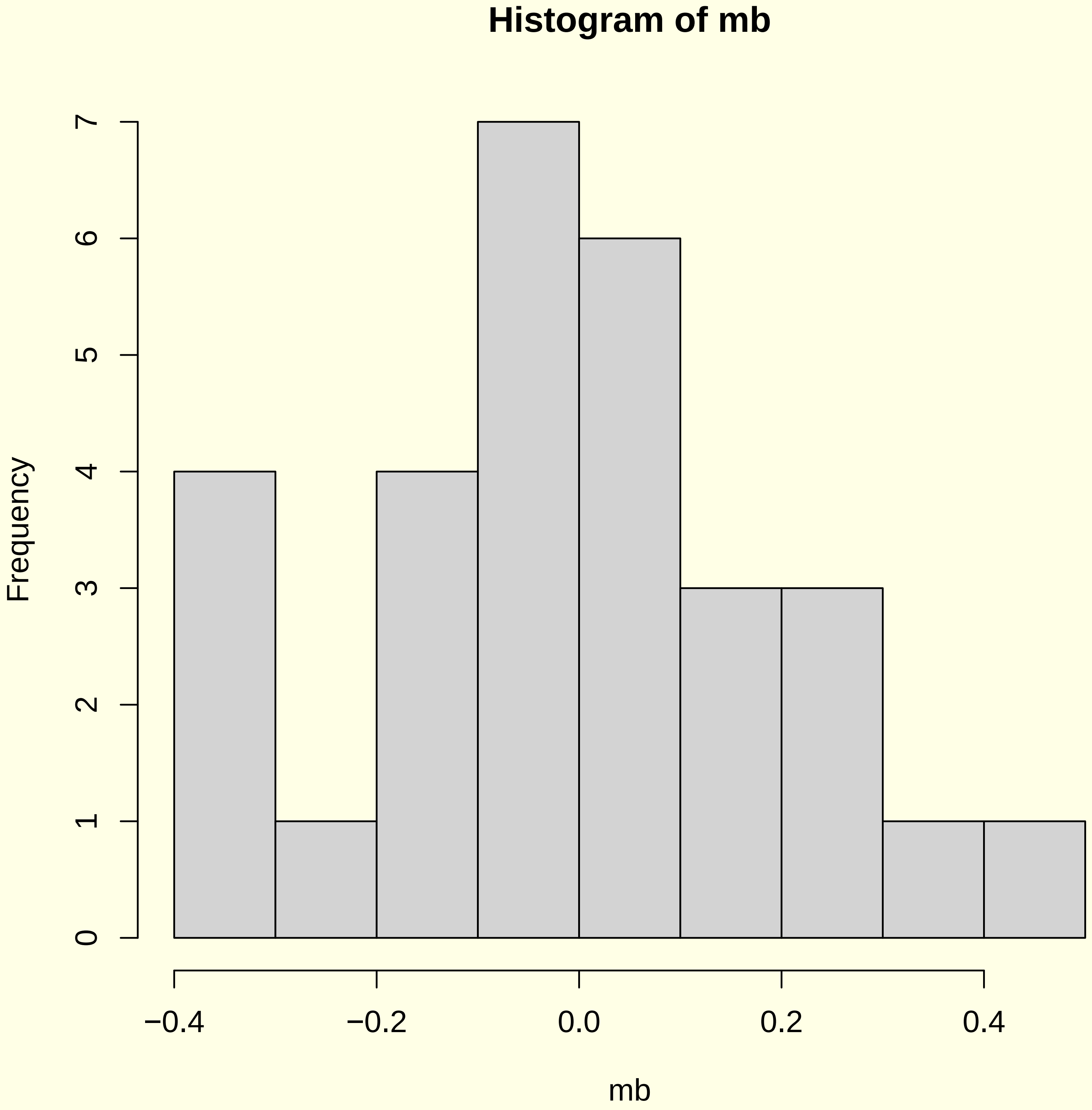
- Often we're working with random variables that we assume have finite variance. Then proving $q.m.$ convergence can be a handy way of proving convergence in probability.

CLT example

- On the homework you dealt with an example where the mean of a sample of size 20 from a finite mean, finite variance population was very far from normally distributed. I.e. the CLT is not providing a good approximation.

- Here's an example that's better behaved: $X_i \in \{-1, 1\}$, with probability .5 on each of these two points. I generated 30 samples of size 20 from this. Histogram and qq normal plots on next two slides.


Q−Q for 30 samples of mean of 20 two−point variables
Sample Quantiles
0.4
0.2
0.0
−0.2
−0.4
−2
−1
0
1
2
Theoretical Quantiles


Histogram of mb
Frequency
0
1
2
3
4
5
6
7
−0.4
−0.2
0.0
0.2
0.4
mb

Bernstein-von Mises theorem setup

We consider a setup where $\{X_i\}$ is a random sample of size N from a distribution with pdf $p(X, \beta)$ and

$$LH_N(\beta) = \prod_{i=1}^{N} p(X_i, \beta)$$

is the sample likelihood from a sample of size N. $\hat{\beta}_N$ is the maximum likelihood estimator of β. β_0 is the value of β for the distribution that generated the data — the “true β”.

The Bernstein-von Mises theorem

Theorem. *Suppose p has continuous second derivatives and that there exists an estimator β^*_N such that $\sqrt{N}(\beta^*_N - \beta)$ has a limiting distribution. I.e. a "root N consistent" estimator exists. Let $\nu = \sqrt{N}(\beta - \hat{\beta})$. Then*

$$LH_N(\beta) = LH_N(v/\sqrt{N} + \hat{\beta}_N)\,,$$

when normalized to integrate to one in ν, defines, as a function of ν, the pdf for a distribution that converges in distribution to $N(0, \Omega)$, where

$$\Omega = -\left(E\left[\frac{\partial^2 \log(p(X, \beta_0))}{\partial\beta\partial\beta'}\right]\right)^{-1}$$

Discussion of the theorem

- For proof of a more general version, including discussion of what happens if the data are not generated from the $p(y, \beta)$ model, see Kleijn and van der Vaart (2012).

- This theorem is interpreted by doctrinaire frequentists as “justifying” credible sets because in large samples, in standard root-N consistent setups, they become similar to confidence sets.

- From a Bayesian perspective, the result approximately justifies the usual mistaken interpretation of confidence sets as if they were credible sets, characterizing uncertainty about an unknown parameter.

- The result depends on the data actually being generated from $p(y, \beta)$ for some β.

- Unlike frequentest asymptotics for an estimator $\hat{\beta}$, the result that the likelihood will asymptotically be Gaussian in shape provides an approximation we can check for accuracy in finite samples.

*

References

KLEIJN, B. AND A. VAN DER VAART (2012): "The Bernstein-Von-Mises theorem under misspecification," *Electronic Journal of Statistics*, 6, 354–381.