Advanced Statistical Theory

Main source: lecture notes of Ryan Martin, Department of Statistics, North Carolina, 2017

Summarized by Ido Greenberg, 2019

Contents

Summary

1. **Statistics**: conversion of information in observed data X into a meaningful summary of the evidence supporting the truthfulness of various hypotheses related to a parameter of interest θ.
   a. Fundamental principle: X behaves differently for different θs (through $p_\theta(X)$).

Exponential Families, Sufficiency and Information

2. ***Sufficient statistic***: carries all info about θ – the distribution depends on θ only through it.
   a. ***Ancillary statistic*** (e.g. $sd(X)$ for location) – independent of θ, though may indicate the amount of uncertainty.
   b. ***Minimal sufficient*** T (e.g. $\bar{X}$ for location) – "smaller" than any other sufficient U: T=h(U).
   c. ***Complete sufficient*** (e.g. ΣX_i in binomial) – contains exactly all info about θ (without any ancillary info). In particular, ancillary statistic can't add any info regarding uncertainty.
3. ***Exponential families***: $p_\theta(x) = h(x)e^{\langle \eta(\theta), T(x)\rangle - A(\theta)}$
   a. Cover most of the popular distributions (though not uniform).
   b. Among families with const support (e.g. uniform is excluded) – **only exp. families have small-dimensional sufficient statistic** independently of the data dimension.
   c. The sufficient statistic wrt θ is **simply** $\boldsymbol{T(X) \coloneqq \left(\sum_{i=1}^n T_j(X_i)\right)_{j=1}^d}$.
   d. T is **complete** iff the family is ***full-rank*** ($\dim\theta = \dim\big(\eta(\theta)\big)$, e.g. unlike $N(\theta, \theta^2)$).
4. ***Fisher Information***: $I_X(\theta)_{ij} = E_\theta\left[\left(\frac{\partial \log p_\theta(X)}{\partial \theta_i}\right)\left(\frac{\partial \log p_\theta(X)}{\partial \theta_j}\right)\right] =_{(\text{if } i=j)} E_\theta\left[\left(\frac{\partial \log p_\theta(X)}{\partial \theta_i}\right)^2\right]$
   a. **"How much X varies with $\boldsymbol{\theta}$"**.
      i. KL-divergence ($E_{p_1}\left[\log\left(\frac{p_1(X)}{p_2(X)}\right)\right]$) satisfies $\boldsymbol{K(p_\theta, p_{\theta+\epsilon}) \approx \epsilon^T I(\theta)\epsilon}$.
      ii. E.g. $I_X(\mu) = 1/\sigma^2$ for $N(\mu, \sigma^2)$.
   b. $\boldsymbol{I_{T(X)} \leq I_X}$ with equality iff T is sufficient (for $dim\theta = 1$).
   c. Increases linearly with **amount of independent data**.
   d. ***Cramer-Rao Theorem***: $\boldsymbol{V_\theta\big(T(X)\big) \geq \big(g'(\theta)\big)^2 \left(I_{T(X)}(\theta)\right)^{-1}}$ $\quad (g(\theta) \coloneqq E_\theta[T])$
   e. ***Observed Fisher-Information*** (which doesn't depend on the unknown θ): $-\frac{\partial \log L(\theta)}{\partial \theta^2}|_{\hat{\theta}}$.

Likelihood

5. ***Likelihood*** = how probable X would be given θ = how plausible θ is = $\boldsymbol{L(\theta) \coloneqq p_\theta(x)}$.
6. ***MLE***: $\hat{\theta} \coloneqq \underset{\theta\in\Theta}{\operatorname{argmax}} L(\theta)$ = "best fit". ***Likelihood equation***: $\nabla \log L(\theta) = 0$.
7. A consistent sequence of solutions (there may be other local extrema for non-convex $logL$, which correspond to solutions that don't converge to the true θ^*, thus not consistent) satisfies $\boldsymbol{\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{D} N\left(0, I(\theta^*)^{-1}\right)}$ (in distribution), i.e. $\boldsymbol{err\left(\hat{\theta}_n\right) \sim \frac{1}{\sqrt{nI(\theta^*)}}}$.
8. ***Likelihood ratio***: *Wilk's theorem* for hypo. test ($H_0: \theta \in \Theta_0$): $-2\log\left(\frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta} L(\theta)}\right) \xrightarrow{D} \chi^2(||\Theta_0||)$.
9. Likelihood-based estimators may be **bad** (not unique, not exist, not consistent…), **mainly when $\boldsymbol{\dim\theta \propto \dim X}$** (e.g. estimate σ when each pair (X_{i1}, X_{i2}) has its own expectation $N(\mu_i, \sigma)$).
10. Alternative methods:
    a. ***Bootstrap***: generate many estimates $\{\hat{\theta}_b\}$ using resampling (with replacements) – and estimate the certainty of $\hat{\theta}$ using their distribution. It is difficult to tell how accurate this

distribution is (remember that from the first place the problem was not knowing the accuracy of $\hat{\theta}$…), and indeed the simple bootstrap variant is not guaranteed to work well.

b. ***Monte-Carlo*** – "exact" rather than asymptotic: calculate estimates for many generated datasets corresponding to (simple) H_0, and calculate p-value according to the quantile of the actual $\hat{\theta}$ among these estimates.

11. The ***likelihood principle***: inference about θ should only depend on (up to multiply by const) $L(\theta)$.
    a. Was thought to be resulted from more basic principles – claim that was lately refuted.
    b. Frequentist hypothesis-test asks "how probable is X as extreme as this one", which depends on probabilities of unobserved x's, which violates the principle.
    c. Bayesian inference satisfies the principle (unless the prior depends on the experiment).
    d. The principle ignores the sampling-distribution, and may lead to unreasonable results (e.g. in experiment that stops only when results are "extreme", it would ignore this fact).

Bayesian Inference

12. In addition to limit of frequency, **use probability to represent uncertainty**, e.g. of a parameter.
13. Given *prior* $\pi(\theta)$, the *posterior* $\boldsymbol{\pi_x(\theta)} = \frac{\boldsymbol{\pi(\theta)p_\theta(x)}}{\boldsymbol{p_\Pi(x)}}$ allows any kind of estimation.

	Frequentist	Bayesian
Point estimate $\hat{\theta}$	Find plausible θ (one for which the data is likely)	Find probable / probably-approximate θ
Interval estimate I_X	*Confidence interval*: generate I_X in a way that independently on θ, it would satisfy $\theta \in I_X$ with large probability	*Credible interval*: generate I_X such that $P(\theta \in I_X \mid X)$ is large
Hypothesis test H_0	Given H_0, how likely is it to observe as extreme data as X? (considering all Xs and only the tested θ)	Given X, how probable is H_0? (considering all θs and only the observed X)

14. Priors:
    a. Example – sensitivity to prior: for $X \sim N(\mu, 1)$ and prior $\mu \sim N(\mu_0, 1)$, expected error (MSE) of posterior's max is $\frac{(\mu^* - \mu_0)^2 + n}{(n+1)^2}$ (compared to $\frac{1}{n}$ for MLE).
    b. ***Prior elicitation***: quantification of domain knowledge into prior dist. (complicated).
    c. ***Conjugate priors-class*** F: $\pi \in F \Rightarrow \pi_x \in F$ (simplifying analysis).
        i. E.g. normal (wrt normal model), gamma (wrt Poisson model).
    d. ***Markov-Chain Monte-Carlo*** (***MCMC***): numerical method for estimation of $\pi_x(\theta)$.
    e. ***Improper prior***: not a probability function (e.g. const over all R).
    f. ***Jeffreys prior*** (private case of ***objective prior***): "non-informative" (minimizing KL-divergence $K(prior, posterior)$) and uniform wrt FI-based geometry; though often improper and violates the likelihood principle (since prior depends on experiment setup).
15. ***deFinetti's Theorem***: ***Exchangeability*** of $\{X_i\}$ (invariance of $p_\theta(X_1 \dots X_n)$ to their order) is in certain cases equivalent to being iid (conditionally on unknown θ).
16. ***Laplace approximation***: use optimization (finding max) to approximate certain integrals.
17. ***Bernstein-von Mises***: under some conditions, posterior mean $E_{\pi_x}[\theta]$ is **asymptotically** (1) similar to any likelihood-based estimator, and (2) normally-dist. around the true θ^* with $Var = \frac{1}{nI(\theta^*)}$.

18. ***Inferential models*** (IM) & ***belief functions*** are kind-of generalization of Bayesian inference (though not probabilistic models), which distinct between lack of info (lack of *belief*) and conflicting info (low *plausibility*).

Decision Theory

19. For probability space $(X, \{p_\theta(x)\})$, **model decisions as actions ($\boldsymbol{a} \in \boldsymbol{A}$) with loss $\boldsymbol{L(\theta, a)}$**.
20. Statistical inference can be modeled as decision problem where loss = error.
21. ***Decision rule***: $a \coloneqq \delta(x)$, ***risk function***: $\boldsymbol{R(\theta, \delta)} \coloneqq E_\theta[L(\theta, \delta(X)]$.
22. ***Admissible rule***: not being ***dominated*** (worse risk for all θ) by any other rule.
    a. A ***(minimal) complete class*** of rules includes (only) all the admissible ones.
23. ***Rao-Blackwell***: under **convex** $L(\theta, a)$ $\forall\theta$, only $\delta = \delta$(**sufficient** statistic) can be admissible.
    a. A dominating rule is $\delta_1(t) \coloneqq E[\delta_0(X)|T = t]$.
    b. Example: for $N(\theta, 1)$, estimating $P(X < c)$ by $\delta = \text{mean}(X < c)$ is dominated by $\delta(\bar{X})$.
    c. **Randomized rules** can be formulated as function of ancillary info, hence **are inadmissible**.
    d. Stein's paradox: standard estimators (MLE, least-squares) for mean of multi-dimensional Gaussian $N(\vec{\theta}, I)$ are inadmissible wrt L2-error loss of all $\vec{\theta}$ elements simultaneously.
24. ***Bayes risk***: $\boldsymbol{r(\pi, \delta)} \coloneqq \boldsymbol{E_\pi[R(\theta, \delta)]}$, ***Bayes rule***: $\boldsymbol{\delta_\pi} \coloneqq \boldsymbol{\underset{\delta}{\text{argmin}}\, r(\pi, \delta)}$ (admissible if exists).
    a. Bayes rule can often be found through the posterior risk $E_{\pi_x}\big[L\big(\theta, \delta(x)\big)\big]$.
    b. Generalized Bayes rule: Bayes rule for improper prior.
        i. A decision rule is admissible iff it is the "limit" (in terms of risk) of generalized Bayes rules of finite-measure priors.
            1. For exponential-family model, this limit is itself a generalized Bayes rule.
25. ***Minimax decision rule***: $\boldsymbol{\underset{\delta}{\text{argmin}} \sup_\theta R(\theta, \delta)}$ (usually less practical out of adversary problems).
    a. A const-risk Bayes rule is minimax (i.e. minimax can be found by prior that makes it const).
        i. Such *least-favorable prior* π^* satisfies $\pi^* = \underset{\pi}{\text{argmax}}\, r(\pi, \delta_\pi)$.
        ii. $\bar{X}$ in location problems under the squared-error loss has const risk and is minimax.
26. Bayes minimizes average risk, and minimax minimizes max risk. Sometimes, under corresponding constraints, risk can be minimized uniformly for all θ.
27. ***Unbiased decision rule*** δ: $\forall\theta^*$: $\theta^* = \underset{\theta}{\text{argmin}}\, E_{\theta^*}\big[L\big(\theta, \delta(X)\big)\big]$.
    a. Bayes rules in estimation problems are biased.
    b. ***Lehmann-Scheffe***: admissible rule which is function of complete sufficient statistic is unique and uniformly minimizes risk among the unbiased rules.
28. ***Equivariance constraints***: if we only consider invariant decision rules (i.e. with structure of the form $\delta(gx) = \tilde{g}\delta(x)$, such as standard location & scale estimators), then under some conditions, uniform risk minimization can be achieved as Bayes rule wrt corresponding prior.
29. In hypothesis test with 0-1 loss, the risk is P(type-I err)+P(type-II err), thus ***Neyman-Pearson*** rules (which maximize power given significance, i.e. minimize type-II error given required type-I error) cover all the admissible rules.
    a. This holds for simple vs. simple and simple vs. one-sided.
    b. For simple vs. two-sided, there's generally no uniformly-most-powerful test for given significance, unless restricting to unbiased tests.

Asymptotic Theory and Optimization-Based Estimators

30. ***M- and Z-estimators*** define estimators in terms of Maximizing or Zeroizing the empirical average of some function $m_\theta(x)$ or $z_\theta(x)$ (e.g. $\hat{\theta}_n \coloneqq \underset{\theta}{\operatorname{argmax}} \overline{m_\theta(X)}$).
    a. **Many popular estimators are private cases**: MLE ($m_\theta \coloneqq \log p_\theta$), Least-Squares ($-\left(y - f_\theta(x)\right)^2$), mean ($z_{\theta(x)} \coloneqq x - \theta$), median ($sign(x-\theta)$), Huber's estimator (between mean and median).
    b. Such definition of estimator **does not require an explicit model $\boldsymbol{p_\theta(x)}$**.
    c. M- and Z-estimators are **consistent** by LLN – as long as $m_\theta(\{X_i\}_1^n)$ converges θ-uniformly.
    d. $\sqrt{\boldsymbol{n}}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}^*\right)$ is **asymptotically-normal** under quite weak conditions.
        i. Classes of z/m-functions that satisfy these conditions are called *Donsker classes*, and essentially admit "uniform CLT" (analog to uniform LLN in consistency).
31. Asymptotic theory of standard estimators (e.g. MLE) is usually based on assumptions of θ-differentiability of $p_\theta(x)$, but **for many properties, the weaker *differentiability in quadratic mean* (*DQM*) suffices**: existence of FI, Cramer-Rao inequality, local asymptotic normality, etc.
    a. In particular, asymptotic normality is claimed to be a very general property of iid models – and not a property of estimators.
32. Bayesian asymptotics can also be generalized for complicated cases (e.g. infinite-dimensional parameters, as in Neyman-Scott example), where consistency of the posterior can be defined and satisfied without the conditions of Bernstein-von Mises theorem.
33. **The consistency of the likelihood-based Bayesian posterior can be generalized to loss-function-based "pseudo-posterior", where the likelihood is replaced with $\boldsymbol{e^{-n \cdot Loss(\theta)}}$**.
    a. Equivalently, the negative log-likelihood is replaced with $n \cdot Loss(\theta)$.
    b. This **removes the dependence on model and possible nuisance parameters**, and **kind of does the final step in connecting statistical modeling to empirical-loss minimization, AKA training wrt loss function**.
    c. Syring & Martin (2015) used it for estimation of medical *Minimal Important Difference*.

Pathological Examples

34. Neyman-Scott: inconsistent MLE.
35. Binomial vs. geometric experiment: likelihood principle allowing arbitrarily significant results.
36. Stein's paradox: standard location estimators of multi-dimensional Gaussian are inadmissible.
37. Hodge's estimator: irregular estimator ("superefficient" but error does not uniformly go to 0).

Summarizer Notes

These notes are intended to keep the understanding and intuition of interesting concepts of advanced statistical theory described in the course. Since the course keeps high standards of mathematical formalism, and in order to stay loyal to the goal of the notes, many of the descriptions – including mathematical formulations – are significantly simplified in a way that may cause inaccuracies.

For example, Bayes Theorem would be presented as

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

and not as

Theorem 4.1 (Bayes's Theorem). *Under the setup described above, let Π_x denote the conditional distribution of Θ given $X = x$. Then $\Pi_x \ll \Pi$ for P_Π-almost all x, where $\mathsf{P}_\Pi = \int \mathsf{P}_\theta \, d\Pi(\theta)$ is the marginal distribution of X from model (4.1). Also, the Radon–Nikodym derivative of Π_x with respect to Π is*

$$\frac{d\Pi_x}{d\Pi}(\theta) = \frac{p_\theta(x)}{p_\Pi(x)},$$

for those x such that the marginal density $p_\Pi(x) = (d\mathsf{P}_\Pi/d\mu)(x)$ is neither 0 nor ∞. Since the set of all x such that $p_\Pi(x) \in \{0, \infty\}$ is a P_Π-null set, the Radon–Nikodym derivative can be defined arbitrarily for such x.

Background

1. ***Kullback-Leibler divergence***: $K(f,g) \coloneqq E_f\left[\log\left(\frac{f(X)}{g(X)}\right)\right] = \int \log(\frac{f}{g})df$
   a. Geometric average of the ratio between the pdfs, which measures the distance between them, though it's not a metric (non-negative but not symmetric and not satisfying the triangle inequality).
2. ***Hoeffding inequality***: if $\{Y_i\}$ are independent mean μ on a support $[a,b]$, then:

$$P(|\bar{Y}_n - \mu| > \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

   a. In particular $n \sim (b-a)^2/\epsilon^2$.
3. The "***fundamental theorem of statistics***":
   a. "***Empirical distribution function***": $\boldsymbol{\widehat{F}_n(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)}$ (for iid $\{X_i \sim F\}_{i=1}^n$)
   b. *Strong law of large numbers*: $\hat{F}_n(x) \to F(x)$ almost surely: $\forall x$: $P\left(\lim\hat{F}(x) = F(x)\right) = 1$.
   c. ***Glivenko-Cantelli***: $\left|\left|\hat{F}_n - F\right|\right|_\infty \to 0$ almost surely (i.e. uniform convergence for all x).
   d. ***Dvoretzky***: $\boldsymbol{P\left(\left|\left|\widehat{F}_n - F\right|\right|_\infty > \epsilon\right) \leq 2e^{-2n\epsilon^2}}$ (again $n \sim 1/\epsilon^2$).
4. Fisher's ***fiducial inference***:
   a. An alternative to the frequentist statistics and Bayesian statistics, which wishes to converse $P(X|\theta)$ to $P(\theta|X)$ without assuming a prior on $P(\theta)$.
   b. Demonstration on $U([0,\theta])$:
      i. $X \coloneqq \max_{1\leq i\leq n} x_i \quad \to \quad P(X \leq a\theta) = a^n \qquad (0 \leq a \leq 1)$
      ii. $\to\ P\left(\theta < \frac{X}{a}\right) = 1 - a^n$
      iii. The calculation is similar to the *pivotal method* for finding a confidence interval.
   c. Unfortunately, this approach had difficulties to prove uniqueness and additivity of the fiducial probability, and to generalize to high-dimensional parameters Θ, and thus did not manage to gain much popularity.
5. ***Statistics*** – a suggested definition: *the conversion of information in the observed data into a meaningful summary of the evidence supporting the truthfulness of various hypotheses related to the parameter of interest*.
6. **Statistical inference** through parametric family of distributions:
   a. Assumption: $x \sim f_\theta$ (where f_θ is a family of distributions).
   b. Goal: estimate θ.
   c. Strategy: analyze available data $X = \{x_i\}$ (typically assumed to be iid), while exploiting the assumption that $\boldsymbol{p_\theta(X)}$ **significantly depends on** $\boldsymbol{\theta}$, i.e. various values of X imply different underlying values of θ.

Exponential Families, Sufficiency and Information

Sufficiency

1. *Statistic*: a function of the data T(X).
2. ***Sufficient statistic*** wrt parameter θ: $f_\theta\big(X\big|T(X)\big)$ is independent on θ, i.e. the statistic carries all info about θ.
   a. E.g. for iid Bernoulli's and $T(X) \coloneqq \sum_1^n x_i = t$: $\mathrm{P}_\theta(X = x \mid T(X) = t) = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}$
   b. ***Neyman-Fisher Factorization*** *Theorem*: T(X) is sufficient iff $f_\theta(X) = g_\theta\big(T(X)\big)h(X)$.
      i. I.e. f(X) depends on θ only "through" (i.e. as function of) T(X).
         1. Note: given T, f indeed does not longer depend on θ.
      ii. Allows detection of sufficient statistics directly from the pdf using dirty algebra.
         1. E.g. by writing uniform distribution as:
            $$p_\theta(x) = \prod_{i=1}^{n} \theta^{-1} I_{(0,\theta)}(x_i) = \theta^{-n} I_{(0,\theta)}(\max x_i)$$
         2. Similarly, $\big(\bar{X}, s^2(X)\big)$ is sufficient for $\theta = (\mu, \sigma^2)$ in Normal distribution.
3. ***Minimal sufficient statistic*** T: any other sufficient U contains all its information, i.e. T=h(U).
   a. Theorem: [T is sufficient] AND [T(x)=T(y) for any x,y that don't distinguish between different θs (i.e. $p_\theta(x)/p_\theta(y)$ is independent on θ)] ➔ [T is minimal].
4. ***Ancillary statistic***: statistic whose distribution is independent of θ.
   a. Ancillary information: information which is "not about" θ.
   b. **Conditioning on ancillary statistics**: although ancillary info can't improve any point-estimator $\hat{\theta}$, it **may still carry information regarding $\boldsymbol{Var(\theta)}$** (i.e. the confidence of the estimation), hence ancillary info should not be disregarded without consideration.
      i. Example: iid X1,X2 with $P(\theta - 1) = P(\theta + 1) = 0.5$. $\bar{X}$ is a sufficient statistic and an unbiased estimator of θ, but the ancillary statistic $X_2 - X_1$ can still tell whether $X_1 \neq X_2$ (hence the estimate is exact) or not.
      ii. In practice, one may infer regarding a parameter using a less-noisy subset of the data (even though it throws data away…), where the subset is possibly determined by conditioning on an ancillary statistic.
5. ***Complete sufficient statistic***: contains exactly all the info in X about θ, without ancillary info.
   a. Formal definition: $E_\theta[f(T)] = 0 \rightarrow f \equiv 0$ (i.e. each "feature" of T has unique info)
   b. **Complete ➔ minimal**.
      i. The opposite isn't true: minimal sufficient statistic may contain ancillary info.
   c. **Any complete sufficient statistic T is independent of any ancillary statistic U**.
      i. In particular, in case of complete statistic T there's **no point in conditioning** on ancillary statistics, since it would not affect T in any way, including confidence.

Exponential families

1. ***Exponential families***: families that can be written as $\boldsymbol{p_\theta(x) = h(x)e^{\langle \eta(\theta), T(x)\rangle - A(\theta)}}$.
   a. The dimension d of the inner product $\langle \eta(\theta), T(x)\rangle = \sum_{j=1}^{d} \eta_j(\theta) T_j(x)$ often corresponds to the vector space of θ (i.e. just a product if θ is scalar).
   b. Representation:
      i. $e^{-A(\theta)}$ may be replaced by $a(\theta)$.

   ii. When considering expectation wrt θ, $h(x)$ may be absorbed into $d\mu(x)$ within the integral of the expectation.
2. Most known families (Normal, Exponential, Poisson, Binomial, Geometric, Beta, Gamma) can be written as exponential (dirty work). **Uniform distribution cannot**.
3. ***Full rank*** exponential family:
   a. Definition: $\eta(\theta)$ has non-empty interior and (T1…Td) are not linearly dependent.
   b. Equivalently: $S \coloneqq \{\eta(\theta_2) - \eta(\theta_1) \mid \theta_j \in \Theta\}$ spans the whole R^d.
   c. Most known exponential families are full rank. $\boldsymbol{N(\theta, \theta^2)}$ **is not**, and belongs to ***curved exponential families***, whose natural parameter space is a curve (or some other set) whose effective dimension is smaller than the actual dimension d.
4. For exponential family of the form $p_\theta(x) = a(\theta)e^{\langle\theta,x\rangle}$ (just d-D exponential dist.?):
   a. $E_\theta(X_i) = -\frac{\partial}{\partial\theta_i}\log a(\theta)$
   b. $C_\theta(X_i, X_j) = -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log a(\theta)$
5. For any exponential family and iid data $X \coloneqq (X_1 \dots X_n)$, the mean of each dimension over all the samples $\boldsymbol{T(X)} \coloneqq \left(\sum_{i=1}^{n} \boldsymbol{T_j(X_i)}\right)_{j=1}^{d}$ **is sufficient**.
   a. If the family is **full rank – T is complete sufficient**.
   b. In the curved family $\boldsymbol{N(\theta, \theta^2)}$, T is minimal but not complete, and **there's no complete statistic**.
6. ***Characterization of sufficiency*** **– mostly only exponential families allow sufficient statistic with small dimension independently of the data dimension**:
   a. If: [$f_\theta(x)$ is continuously differentiable wrt x] AND [$X_1 \dots X_n$ are iid] AND [$supp\big(f_\theta(x)\big)$ is independent of θ];
      i. Note: **uniform distribution**'s support depends on θ, hence it's **excluded**.
   b. Then: **[there exists sufficient statistic of dimension k]** $\Leftrightarrow$ **[**$\boldsymbol{f_\theta(x)}$ **is exponential family with** $\boldsymbol{d \le k}$**]**.

Fisher information

1. ***FI regularity conditions***: $supp(p_\theta)$ is independent of θ + some weak differentiability conditions.
   a. All the discussion in this section assumes the FI conditions on p_θ.
2. ***Score vector***: $s_X(\theta) \coloneqq \nabla_\theta \log p_\theta(X)$
   a. $E_\theta[s_X(\theta)] = 0$ (derived from conservation of the sum $\int p_\theta(X)d\mu(X) = 1$)
3. ***Fisher information*** – the covariance matrix of the score:
   a. $$I_X(\theta)_{ij} = \mathrm{C}_\theta\left(\frac{\partial \log p_\theta(X)}{\partial\theta_i}, \frac{\partial \log p_\theta(X)}{\partial\theta_j}\right)$$
      i. In particular: $\boldsymbol{I_X(\theta)_{ii} = E_\theta\left[\left(\frac{\partial \log p_\theta(X)}{\partial\theta_i}\right)^2\right]}$
      ii. Interpretation: **how much** $\boldsymbol{X}$ **varies with** $\boldsymbol{\theta}$.
   b. Equivalently: $$I_X(\theta)_{ij} = -\mathrm{E}_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\right\}$$
      i. In particular, for **iid** variables: $\boldsymbol{I_{X_1\dots X_n}(\theta) = nI_{X_1}(\theta)}$.
      ii. In general, **information is accumulated** with increase of data.

iii. Example – $N(\mu, \sigma^2)$ (known σ): $I_X(\mu) = -E_\mu\left[\frac{\partial^2}{\partial\mu^2}\left(C(\sigma) - \frac{(x-\mu)^2}{2\sigma^2}\right)\right] = 1/\sigma^2$.

4. For any statistic T, $I_X(\theta) - I_T(\theta)$ is positive semidefinite, and $I_X(\theta) - I_T(\theta) \equiv 0 \Leftrightarrow$ T is sufficient.
   a. In particular for $dim\theta = 1$: $\boldsymbol{I_{T(X)} \leq I_X}$**, with equality iff T is sufficient**.
5. ***Cramer-Rao Theorem***: for a statistic T of **iid** $X_1 \dots X_n$ with $E_\theta[T] = g(\theta)$:

$$\boldsymbol{V_\theta(T) \geq \left(g'(\theta)\right)^2 \left(nI(\theta)\right)^{-1}}$$

   a. Note: if T is an unbiased estimator of θ, then $E[T] = \theta$ and thus $g' \equiv 1$.
   b. Note: $I(\theta)$ here actually means $I_{observations}(\theta)$. In particular, in experimental design we wish our observations to maximize FI in order to allow accurate estimation of θ.
      i. "Maximize" FI in the non-scalar case is usually associated with some functional of FI, e.g. its determinant(?!).
6. KL-divergence of p_θ from a slight variant of it satisfies $\boldsymbol{K(p_\theta, p_{\theta+\epsilon}) \approx \epsilon^T I(\theta) \epsilon}$ as $\epsilon \to 0$, which is another way to see **FI** as **the sensitivity of the distribution of X to $\boldsymbol{\theta}$**.
7. Cramer-Rao Theorem makes FI a consensus under FI regularity conditions, which covers most known families. In other cases (e.g. uniform distribution) there are attempts to generalize, e.g. based on the last property. One such generalization is "Hellinger information" (with the underlying *Hellinger distance* $h^2(\theta, \theta') \coloneqq \int \left(p_\theta^{1/2} - p_{\theta'}^{1/2}\right)^2$ rather than KL-divergence), which is currently being formed by the course author.

Likelihood

Likelihood-Based Methods

1. ***Likelihood***: $\boldsymbol{L(\theta) \coloneqq p_\theta(x)}$.
   a. Terminology coined by Fisher (1973): intuitively similar to probability (describing **how *plausible* $\boldsymbol{\theta}$ is**), yet not a probability function (e.g. $\sum L(\theta) \neq 1$).
2. ***Maximum Likelihood Estimation*** (***MLE***): $\hat{\theta} \coloneqq \underset{\theta \in \Theta}{\operatorname{argmax}} L(\theta)$.
   a. Model which is most likely to have produced the observed x ("**best fit**").
   b. Not necessarily minimizing expected estimation error.
   c. The ***likelihood equation***: $\boldsymbol{\nabla l(\theta) = 0}$ $\quad$ $\boldsymbol{(l \coloneqq \log L)}$
   d. Convergence of $\hat{\theta} \in \Theta \subset R$: under certain conditions (mainly smoothness and constant $supp(p_\theta)$), if $\hat{\theta}_n \coloneqq \hat{\theta}(x_1 \dots x_n)$ is a consistent sequence of solutions to the likelihood eq.:
   $\boldsymbol{\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \to N\left(0, I(\theta^*)^{-1}\right)}$ in distribution **under $\boldsymbol{p_{\theta^*}}$** (where $\theta^* \in int(\Theta)$)
   In other words, $\boldsymbol{err\left(\hat{\theta}\right) \sim \frac{1}{\sqrt{nI(\theta^*)}}}$.
      i. Since θ^* and $I(\theta^*)$ are unknown, the ***observed Fisher Information*** $\boldsymbol{-l_n''\left(\hat{\theta}_n\right)}$ often replaces the original FI.
3. ***Likelihood ratio tests***:
   a. Hypothesis test with the statistic $\boldsymbol{T_n(X_n, \Theta_0) \coloneqq \sup_{\theta \in \Theta_0} L(\theta) \,/\, \sup_{\theta \in \Theta} L(\theta)}$.
   b. Wilk's statistic: $W_n \coloneqq -2logT_n$.
      i. ***Wilk's theorem***: $\boldsymbol{W_n \to \chi^2(||\Theta_0||)}$ under $\theta \in \Theta_0$.
4. **Many desired properties are not universally satisfied for MLE: uniqueness, existence, asymptotic normality, consistency**.
   a. Example (***Neyman-Scott***): for $X_{i1}, X_{i2} \sim N(\mu_i, \sigma^2)$, MLE yields $\hat{\mu}_i = \frac{X_{i1}+X_{i2}}{2}$, $\hat{\sigma} = \frac{1}{n}\sum(X_{i2} - X_{i1})^2$, and $\hat{\sigma}^2 \to \frac{1}{2}\sigma^2 \neq \sigma^2$ is not consistent.
   b. **This anomaly is due to the dimension of the nuisance parameter ($\{\boldsymbol{\mu_i}\}$) increasing with n**. In particular, the MLE estimates $\hat{\mu}_i$ and $\hat{\sigma}$ together, while $Var(\hat{\mu}_i) = const \nrightarrow 0$. This can be easily solved by the transformation $Y_i \coloneqq X_{i2} - X_{i1}$.
   c. Yet, the author claims that ML is reliable only in cases of very regular distributions, in which it is not needed anyway.
   d. Bottom line – **avoid assuming likelihood-based methods to work as expected**.

Alternative and Advanced Methods

1. ***Bootstrap***: use resampling (with replacement) to generate confidence tests & intervals based on a single dataset. Simplest form:
   a. The data $X = X_1 \dots X_n$ leads to $\hat{\theta}$.
   b. For b=1:B: resample (with replacement) $\{X_{bi}^*\}_{i=1}^n$, leading to $\hat{\theta}_b^*$.
   c. Approximate the distribution of $\hat{\theta}$ using $\left\{\hat{\theta}_b^*\right\}$, and estimate confidence accordingly.
   d. This is justified by the fundamental theory of statistics – empirical distribution convergence to the true distribution – though it's **unclear how it can be assumed** in advance that there's sufficient data for approximate bootstrap confidence estimation, without assuming that $\hat{\theta}$ is a good approximation as well anyway.

   e. **Basic bootstrap doesn't work universally** (some specific failures are known), hence **can't be used blindly**. More advanced and non-intuitive bootstrap methods solve some of the problems, and also guarantee higher-order accuracy(?).
2. ***Monte-Carlo*** sampling:
   a. "*The focus on asymptotic theory is arguably driven by tradition – when … were no computers available, … only asymptotic analytical approximations were possible.*"
   b. Given some $H_0: \theta = \theta_0$, sample many datasets $X_1 \dots X_n \sim p_{\theta_0}$, calculate the corresponding test statistics $\{U_i\}_1^n$, and determine p-value according to the quantile of the actual statistic U among the simulated distribution of $\{U_i\}$.
      i. This allows ***exact inference*** – without asymptotic analysis.
      ii. ***Plausibility functions*** are mentioned but not clearly explained.
   c. If H_0 is more general and does not fully specifies θ, then such simulation is problematic. Sometimes it can be shown that the statistic is independent of the unspecified components of θ. Alternatively, it is possible to run simulations with various parametric assumptions.
3. **Marginal & conditional likelihood**:
   a. Goal – handle nuisance parameters: **for $\boldsymbol{\theta = (\psi, \lambda)}$, infer on $\boldsymbol{\psi}$** with minimal effect of the uncertainty in λ (which normally affects the whole MLE $\hat{\theta}$).
   b. If the data X can be represented (one-to-one mapping) as (S, T) such that $p_\theta(X) = p_\theta(S,T) = \boldsymbol{p_\theta(S) p_\psi(T|S)}$ (i.e. only ψ affects T) then inference about ψ can be done using conditioning on S through the marginal distribution of T.
   c. Note: such decomposition may throw away information about ψ in $p_\theta(S)$, but the elimination of the nuisance parameter λ may be worth it.
   d. Example – Neyman-Scott: the decomposition $S_i \coloneqq X_{i1} + X_{i2}, T \coloneqq X_{i2} - X_{i1}$ derives exactly the solution suggested above.
   e. The challenge is to find such decomposition X=(S,T). A semi-general method is available for the subset of exponential families that satisfy $p_\theta(x) \sim e^{\langle \psi, T(x) \rangle + \langle \lambda, S(x) \rangle - A(\psi, \lambda)}$. An application in the context of logistic regression was given by Boos & Stefanski (2013).
4. **Asymptotic expansions**:
   a. According to CLT, $S_n = \sqrt{n}(\bar{X} - \mu)/\sigma \to N(0,1)$. It is claimed that this is a 2nd-degree Taylor approximation of the actual distribution (derived from its moment-generating function), and that more accurate approximations (which are beneficial for finite n…) can be achieved using higher orders (higher moments).
   b. Note: both MLE & Wilk's statistics are private cases: inference about them typically uses Normal approximation of the statistic's distribution.

The likelihood principle

1. ***Likelihood principle***: two datasets which derive equivalent likelihoods (i.e. $L_1(\theta) = c \cdot L_2(\theta)$), should result in the same inference regarding θ.
2. ***Sampling-distribution***: the distribution of a statistic of data assuming some probabilistic model.
   a. Sampling-distribution-based method: inference by comparing the value of a test statistic to its sampling-distribution (**e.g. hypothesis tests**, and in particular likelihood-ratio test).

   b. The sampling-distribution of a statistic depends on the experiment's setup (in particular on the stop condition), thus the inference from certain data may depend on the setup, hence **sampling-distribution-based inferential methods violate the likelihood principle**.
3. History of necessity of the likelihood principle:
   a. ***Sufficiency principle***: two datasets which derive the same sufficient statistic T, should result in the same inference regarding θ.
   b. ***Conditionality principle***: if two experiments are considered, and only one is randomly (independently of θ) chosen, the inference regarding θ should be based only on the actually-conducted experiment.
   c. *Birnbaum* (1962): sufficiency & conditionality ➔ likelihood.
      i. Since both are typically considered necessary (unclear to me why sufficiency is more consensus than likelihood…), it means that the likelihood principle is necessary as well.
      ii. **The Bayesian approach is the only known one which satisfies the likelihood principle** (unless the prior depends on the experiment as in objective priors). In particular, frequentist methods are "illogical" in the sense of these principles.
      iii. *Evans* (2013) & *Mayo* (2014): Birnbaum's claim is actually **false**.
4. Example: for Bernoulli with $H_0: \theta = \frac{1}{2}, H_A: \theta > \frac{1}{2}$, the data X=(1,0) can lead to different p-values:
   a. T = **num of 1s in 2 samples**: $P_{\theta_0}(T \geq 1) = P(1) + P(>1) = 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = \frac{3}{4}$
   b. T = **num of 1s until first 0**: $P_{\theta_0}(T \geq 1) = P(1) + P(>1) = \frac{1}{4} + \sum_2^\infty \left(\frac{1}{2}\right)^{T+1} = \frac{2}{4}$
   c. Note: the likelihoods $L_1(\theta) = P_\theta(T=1) = 2 \cdot \theta(1-\theta)$ and $L_2(\theta) = P_\theta(T=1) = \theta(1-\theta)$ are indeed equivalent, so the likelihood principle is indeed violated.
   d. The key is that the question "how likely was it to observe data D(X) as extreme or more than X" is not well defined – **D(X) = "as extreme as X=(1,0)" refers to unobserved (and arguably arbitrary) dataset, which is different in each case**:

	X = #1 in 2 samples		X = #1 until 0	
T = 0	{(0,0)}	P = ¼	{(0)}	P = ½
T = 1	{(1,0), (0,1)}	P = ½	{{1,0)}	P = ¼
T > 1	{(1,1)}	P = ¼	{(1,1,0), (1,1,1,0)…}	P = ¼

   e. Note that each experiment drew the X=(1,0) from a different probability space, hence the result receives different interpretation. In particular, **the second experiment's probability space includes more (probable) elements corresponding to T=0**, and less elements corresponding to T=1, thus the experiment design affects the expected results.
   f. The **key difference between the frequentist and the Bayesian approaches** in this context is that the former asks **"how likely is X for this θ compared to other Xs?"** ($P(T > t)$) while the latter asks **"how likely is X for this θ compared to other θs?"** ($P(\theta)$).
5. Is it really a problem that the p-value depends on the context of the experiment?
   a. Consider **T = rate of 1s until [rate of 1s > $1/2 + 2\sigma_0/\sqrt{n} = 1/2 + 1/\sqrt{n}$]**.
      i. Note: this stops within finite time with probability 1.
   b. When stops after n repetitions, the **result of the experiment is guaranteed to reject H_0** with $\alpha = 5\%$ if interpreted as a standard n-repetitions experiment. Thus, **it is clear that**

the inference must consider the context in which the data was generated, namely the sampling distribution.

c. Mayo (2014): *For the sampling theorist, to report a 1.96 standard deviation difference known to have come from optional stopping, just the same as if the sample size had been fixed, is to discard relevant information for inferring inconsistency with the null, while "according to any approach that is in accord with the strong likelihood principle, the fact that this particular stopping rule has been used is irrelevant." [Cox and Hinkley (1974), page 51.]*.

d. Actually there's a **lot of (not always consistent) material out there about Bayesians vs. frequentists dealing with stopping rules**, and it **seems to depend a lot on how exactly each of the two tries to handle the stopping rule**. See for example 1,2,3.

   i. In particular, **Bayesian approach can handle this by using an objective prior** (see Bayesian Inference \ Choice of priors) such as Jeffreys prior $\propto \sqrt{I_X(p)}$, though it **costs in loss of the likelihood principle**.

e. **Fisher information** of the **binomial** test is $\frac{\mathbf{2}}{\boldsymbol{p(1-p)}}$, while for the **geometric** one it is $\frac{\mathbf{1}}{\boldsymbol{p(1-p)^2}}$.

   i. Note: both tests are similarly informative for $p \sim \frac{1}{2}$, while the latter is better as $p \to 1$, i.e. it has better resolution to distinguish 0.9 from 0.8 than 0.2 from 0.1 – which isn't surprising as for larger p we'd expect to essentially have more data.

6. **In summary** of this section, I'd say that:
   a. The likelihood principle doesn't make much sense, as it essentially ignores the mechanism which generates the data, and leads to paradoxes of fake significance.
   b. With careful formulation, probably both frequentists and Bayesians can handle this issue, though it feels less natural in the Bayesian framework.
   c. In any case, we should never forget:

"We have to remember that frequentism and Bayesianism are different things, that answer different questions, whose basic object of study just happens to have the same name – probability – but is not the same thing at all. To a frequentist, it's a limiting frequency; to a Bayesian, it's a measure of uncertainty. They agree a lot, but sometimes they don't." (rationalistramble)

Bayesian Inference

Background

1. ***Frequentist approach***:
    a. Given some procedure, study a random variable (point estimator, confidence interval, etc.) in terms of behavior in repeated sampling.
    b. **Can't say anything about the probability that some hypothesis is true** – only likelihood of data given that hypothesis.
2. ***Bayesian approach***:
    a. **Axiom: uncertainties can only be described with probability**.
    b. *Prior distribution* of a parameter expresses one's uncertainty and belief rather than values of the parameter in some repeated sampling procedure.
        i. Similarly to "there's 50-50 chance that I'll come to the party".
        ii. Allows to express researcher's beliefs regarding the phenomenon.
    c. Parameters are still considered constant unknowns rather than random variables, but the axiom derives similar mathematical treatment for both types.

Bayesian analysis

1. Notation: θ=value, Θ=variable (also parameter space, distinguish will be clear in context).
2. **Hierarchical model**: $\boldsymbol{\Theta \sim \Pi}$ (***prior distribution***) and $\boldsymbol{X|(\Theta=\theta) \sim p_\theta(x)}$.
3. Inference is based on ***posterior distribution*** $\pi_x(\theta)$ through ***Bayes theorem***: $\pi_x(\theta) = \frac{\pi(\theta)p_\theta(x)}{p_\Pi(x)} \propto \pi(\theta)p_\theta(x)$.
4. **Marginalization** ($\theta=(\psi,\lambda)$ where only ψ is of interest) – straightforward from the rules of probability: $\pi_x(\psi) = \int \pi_x(\psi,\lambda)d\lambda$.
    a. **Prediction** as a private case of marginalization: $\pi_{\{x_i\}_1^n}(x_{n+1}) = \int \pi_{\{x_i\}_1^n}(\theta, x_{n+1})d\theta$.
    b. Claim: **any "rolling" prediction rule** (i.e. $X_1 \sim p_0$ & after n observations $X_{n+1} \sim p_{\{x_i\}_1^n} = f(x_1 \dots x_n)$) **which permits change of observations order** (i.e. f is invariant to input permutations) **must be a private case of such a Bayesian rule**.
5. **Point estimation**:
    a. $\boldsymbol{\hat{\theta}_{mean}} \coloneqq E[\Theta|X=x] = \int \theta d\Pi_x(\theta)$ (minimizing L2 error)
    b. $\boldsymbol{\hat{\theta}_{mode}} \coloneqq \underset{\theta}{\operatorname{argmax}}\, \pi_x(\theta)$ (essentially ML that assumes prior on θ)
6. **Set (interval) estimation**:
    a. $\boldsymbol{1-\alpha}$***-credible set***: $C \subset \Theta$ such that $\Pi_x(C) = 1-\alpha$.
    b. **Centered confidence interval**: $C = [quantile_{\Pi_x}(\alpha/2)\,, quantile_{\Pi_x}(1-\alpha/2)]$
    c. **Highest-density confidence set**: $C = \{\theta : \pi_x(\theta) \geq c_\alpha\}$ (c_α is chosen such that $\Pi_x(C) = 1-\alpha$)
    d. Note: unlike frequentist confidence interval, the ***coverage probability*** (given the true θ_0 – the probability that C will contain θ_0) is **not necessarily** $\boldsymbol{1-\alpha}$ (due to **sensitivity to the prior** – e.g. if $\Pi(\theta_0) = \epsilon \ll 1$, then C is less likely to contain θ_0).
7. **Hypothesis test**: $\theta \in H$ for some $H \subset \Theta$ is **rejected if** $\boldsymbol{\alpha > \Pi_x(H)} = \frac{\int_H p_\theta(x)d\Pi(\theta)}{p_\Pi(x)}$.

	Frequentist	Bayesian
Point estimate $\widehat{\boldsymbol{\theta}}$	Find θ for which the data is likely	Find probable / probably-approximate θ
Interval estimate $\boldsymbol{I_X}$	***Confidence interval***: generate I_X in a way that independently on θ, it would satisfy $\theta \in I_X$ with large probability	***Credible interval***: generate I_X such that $P(\theta \in I_X \mid X)$ is large
Hypothesis test $\boldsymbol{H_0 \subset \Theta}$	Given H0, how likely is it to observe as extreme data as X? (considering all Xs and only the tested θ)	Given X, how probable is H0? (considering all θs and only the observed X)

Choice of priors

1. Accuracy can be very **sensitive to the choice of prior**. For example:
   a. Let $\mu \sim N(0,1)$, $X \sim N(\mu, 1)$, $\mu_0 :=$true μ, $\bar{X} =$ ML estimator, $\hat{\mu} := \operatorname{argmax}(\pi_x(\mu))$.
   b. $\boldsymbol{MSE}(\bar{\boldsymbol{X}}) := E_{\mu_0}[(\bar{X} - \mu_0)^2] = \frac{\mathbf{1}}{\boldsymbol{n}}$, $\boldsymbol{MSE}(\hat{\boldsymbol{\mu}}) = \frac{\boldsymbol{\mu_0^2 + n}}{\boldsymbol{(n+1)^2}}$.
2. ***Prior elicitation***: detailed quantification of domain knowledge in terms of prior distribution – usually impractical.
3. ***Conjugate priors-class***: family of distributions F such that $\Pi \in F \Rightarrow \Pi_x \in F$. Makes analysis easy. For example:
   a. $X \sim N(\Theta, \sigma^2)$ (known σ), $\Theta \sim N(\omega, \tau^2)$ ➔ $\Theta|\bar{X} \sim N\left(\frac{n\tau^2}{\sigma^2+n\tau^2}\bar{x} + \frac{\sigma^2}{\sigma^2+n\tau^2}\omega, \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\right)$.
   b. $X \sim Pois(\Theta)$, $\Theta \sim Gamma(a, b)$ ➔ $\Theta|\bar{X} \sim Gamma\left(n\bar{x} + a, \frac{1}{n+1/b}\right)$.
   c. Conjugate classes can be extended by mixtures of conjugate priors.
4. Nowadays it is unnecessary to choose priors just for convenience, since numerical methods can be used to estimate Π_x, e.g. ***Markov-Chain Monte-Carlo*** (***MCMC***).
5. It is also possible to try several different priors and hope for consistent inference.
6. ***Improper prior***: prior which is not a probability function (e.g. uniform $\pi(\mu) = 1$).
   a. Posterior can still be calculated (in the example simply $\pi_x \propto ML$).
   b. In general, probability theory can be extended to improper priors (keeping much of the theory and in particular Bayes theorem) by either allow infinite probabilities or remove countable additivity (only assume finite additivity).
7. ***Objective prior***:
   a. ***Jeffreys prior***: $\boldsymbol{\pi(\theta) \propto \sqrt{\det(I_X(\theta))}}$.
      i. Θ is **distributed uniformly** wrt the geometry induced by Riemannian metric (determined by FI).
         1. Euclidean-uniform under FI=const, as in location parameters, and in particular improper in these cases.
      ii. **Non-informativity**: Minimizing asymptotic KL-divergence between prior & posterior.
      iii. Credible sets approximately satisfy coverage properties of frequentist methods.
   b. All objective priors (e.g. Jeffreys, invariant priors) are improper under common models.

c. Bayesian inference satisfies the likelihood principle (inference depending only on likelihood of data) only given a certain prior. **Objective priors** depend on the experimental setup (e.g. through FI as in Jeffreys prior) and thus **do not respect the likelihood principle**.

Exchangeability

1. $X_1 \dots X_n$ are ***exchangeable*** if their joint distribution is invariant to permutations, i.e. their order is "irrelevant".
   a. An infinite series of variables is exchangeable if any finite subset is exchangeable.
2. Despite looking much **weaker assumption than iid**, infinite exchangeable series turn out to be conditionally iid under certain conditions:
   a. ***deFinetti's theorem***: binary variables $X_1 \dots X_n$ are exchangeable iff there's random variable Θ in $[0,1]$ such that $X_i|(\Theta = \theta) \sim Ber(\theta)$.
   b. Note: exchangeability derives existence of prior but does not help to find it.
3. Generalizations to non-binary variables (e.g. Hewitt-Savage) are available (but complicated).

Asymptotic Bayesian theory

1. ***Laplace approximation*** ($\theta \in R^p, \hat{\theta} \coloneqq \underset{\theta}{\operatorname{argmax}}\, h(\theta)$):

$$\int q(\theta) e^{nh(\theta)} d\theta = q\left(\hat{\theta}\right) e^{nh(\hat{\theta})} \cdot \sqrt{\left(\frac{2\pi}{n}\right)^p \frac{1}{\det\left(-h''\left(\hat{\theta}\right)\right)}} \cdot \left(1 + O\left(\frac{1}{n}\right)\right)$$

   a. Allows **calculation of integral without integration – only optimization** ($\operatorname{argmax} h$).
   b. Useful for a variant of applications (e.g. *Stirling's approximation* of n!), and in particular common integration problems in Bayesian statistics.
2. ***Bernstein-von Mises theorem***: **the posterior of Θ** is asymptotically normally-**distributed around any consistent estimate $\widehat{\boldsymbol{\theta}}$**, with **variance** $\frac{\mathbf{1}}{\boldsymbol{nI(\theta^*)}}$ (θ^* is apparently the true value).
   a. More specifically: under certain conditions (mainly smoothness and constant $supp(p_\theta)$, and the prior being continuous & positive at θ^*), any consistent sequence $\hat{\theta}_n$ of solutions to the likelihood equation satisfies $\sqrt{n}\left(\Theta - \hat{\theta}_n\right) \to N(0, 1/I(\theta^*))$ in P_{θ^*}-probability.
3. Conclusion: **the posterior mean is asymptotically similar to any consistent likelihood-based estimate – almost independently of the prior**.
   a. Specifically: under the same conditions as above, and assuming $E_\Pi(\theta) < \infty$, the **posterior mean $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}} \coloneqq \boldsymbol{E}[\boldsymbol{\Theta}|\boldsymbol{X}]$** satisfies $\sqrt{\boldsymbol{n}}\left(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right) \to \mathbf{0}$ and $\sqrt{\boldsymbol{n}}\left(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}^*\right) \to \boldsymbol{N}\left(\mathbf{0}, \mathbf{1}/\boldsymbol{I}(\boldsymbol{\theta}^*)\right)$.

Belief functions, additional topics and references

1. Bayesian methodology and philosophy: Berger (1985), Ghosh (2006).
2. Bayesian modeling and methodology: Gelman (2004).
3. Hypothesis tests with $\pi(H_0) = 0$ (e.g. $H_0 = \{\theta_0\}$): Bayes factors, Ghosh (2006).
   a. Bayes factors are not functions of the posterior, thus not really Bayesian.
4. Monte-Carlo numerical integration: Robert & Casella (2004).
5. ***Belief functions*** and ***inferential models*** (***IM***): Martin and Liu (2013, 2015).
   a. If nothing supports either A or A^C, belief functions allow expressing small evidence of any of the two, instead of the strict tradeoff $P(A) + P(A^C) = 1$ existing in probability.

Dempster-Shafer theory of belief functions

6. ***Dempster-Shafer theory*** suggests a way to summarize evidence in favor of various possibilities within a discrete space X (e.g. red / yellow / green), and express degrees of belief (/confidence/trust) in subsets of these possibilities.
7. ***Mass*** $m: 2^X \to [0,1]$ represents the evidence in favor of each specific subset (e.g. non-green evidence increases red-or-yellow mass), assigns 0 mass to empty set, and is normalized to sum 1. It is yet interpreted as normalized amount of evidence and NOT as probability, since e.g. mass(yellow-or-red) may be smaller than mass(red).
    a. ***Dempster's rule*** can define masses as combination of info from multiple sources. It can be seen as a generalization of Bayes rule. Jøsang claimed it should be only interpreted as fusion of belief-constraints from different sources.
8. ***Belief* (/*support*) = how strongly we believe it to be true** = total amount of supporting evidence = sum of masses of contained subsets.
    a. E.g. belief of yellow-or-red = masses sum of yellow, red & yellow-or-red.
9. ***Plausibility* = lack of conflict with observed evidence** = 1 – amount of evidence against = 1 – sum of masses of non-intersecting subsets.
    a. E.g. plausibility of yellow-or-red = 1 – mass of green = masses sum of yellow, red, yellow-or-red, red-or-green & yellow-or-green.
    b. In particular $pl(A) = 1 - bel(A^C)$.
10. While belief and plausibility can be generalized to infinite X, mass can't.
11. **In summary, Dempster-Shafer theory allows distinction between evidence in favor (belief) and lack of evidence against (ignorance)**.
    a. **Lack of evidence** can be expressed as most mass going to the large subsets (red-or-yellow-or-green) and result in high plausibility but low belief for every possibility. In other words, **the gap between plausibility and belief expresses ignorance**.
    b. On the other hand, much **conflicting / contradicting info** can be expressed as mass going to non-intersecting subsets, increasing each's belief but also reducing each other's plausibility.
12. A newer attempt – apparently to summarize belief and plausibility into posterior info which does not depend on a subjective prior – was published by Martin & Liu in 2013 (*Inferential Models: A Framework for Prior-Free Posterior Probabilistic Inference*).

Decision Theory

Basic definitions

1. **<u>Goal</u>: make decisions that minimize some cost under uncertainty**.
2. Can be seen as game theory problem where players are the statistician and the "nature" (who knows the true value of the parameter).
3. Setup:
   a. **Probability space** $(X, \mathrm{A}, \{p_\theta : \theta \in \Theta\})$
   b. ***Action space*** A
   c. ***Loss function*** $L: \Theta \times A \to [0, \infty)$ (cost of an action given the parameter value)
4. Examples:
   a. Hypothesis test: $\Theta = \{0,1\}$ (H_0/H_1), $A = \{0,1\}$ (accept/reject), $L(0,0) = L(1,1) = 0$, $L(0,1) = L(1,0) = 1$.
   b. Point estimation (of $\psi(\theta)$): $A = \psi(\Theta)$ (estimate), $L(\theta, a) = \big(a - \psi(\theta)\big)^2$ (squared-error of estimate).
5. ***Decision rule*** – choose action (or randomized action) according to observed data:
   a. Deterministic: $\delta: X \to A$ $\qquad L = L\big(\theta, \delta(x)\big)$
   b. Randomized: $\delta: X \to$ measure functions on A $\qquad L = E_\delta[L(\theta, a)] = \int_A L(\theta, a)\delta(x)(da)$
      i. Note: randomized decision rules are sometimes used in discrete hypothesis tests to allow choosing a specific confidence level – where none of the possible discrete rejection-thresholds corresponds to this confidence level.
6. ***Risk function*** – expected loss of a decision rule:
   a. $R(\theta, \delta) \coloneqq E_\theta[L(\theta, \delta(X)] = \int_X L\big(\theta, \delta(x)\big) P_\theta(dx)$.
   b. In general, there's no uniform risk minimization $\underset{\delta}{\operatorname{argmin}}\, R(\theta, \delta)$ for all θ.

Admissibility

1. δ is ***inadmissible*** if it is ***dominated*** by some δ' (i.e. $\forall\theta: R(\theta, \delta') \le R(\theta, \delta)$, with strict inequality for some θ).
   a. Inadmissible decisions rules usually don't need to be considered.
   b. Note: in point-estimation of θ, $\delta(x) \equiv 42$ is admissible (better than any other rule if $\theta = 42$), yet it isn't a reasonable rule.
2. ***Rao-Blackwell***: **under convex loss function, only functions of sufficient statistics can be admissible deterministic decision rules**.
   a. Specifically – for action space $A \subset R^d$ – a dominating rule is $\delta_1(t) \coloneqq E[\delta_0(X)|T = t]$.
      i. That's expectation over the action space, which is independent of θ as long as T is sufficient (thus p_θ depends on X only through T).
      ii. Actually all rules are functions of the data X which is a sufficient statistic, so it's not very clear… maybe minimal sufficient?
   b. Example – estimating $P(X < c) = \Phi(c - \theta)$ for $X_1 \dots X_n \sim N(\theta, 1)$. A natural estimator is $mean(X < c) = 1/n \cdot \sum I_{(-\infty, c)}(x_i)$, but it's not a function of $T = \bar{X}$, thus for the convex loss $\big(a - \Phi(c - \theta)\big)^2$, a dominating rule is $E[\delta_0(X)|t] = \frac{1}{n}\sum E\big[I_{(-\infty,c)}(X_i)|t\big] = P(X_1 < c|t) = \cdots$.

3. Given a randomized decision rule, a new probability space $\tilde{X} = (X, U)$ can be built such that U expresses the randomization of the action, and wrt this space the decision rule is deterministic, and not a function of a sufficient statistic. Thus, **all randomized decision rules are (at least weakly) inadmissible under convex loss function**.
    a. In particular loss of estimation is usually convex (e.g. L2-error) – hence doesn't require randomized estimator.
    b. Stein's paradox: standard estimators (MLE, least-squares) for mean of multi-dimensional Gaussian $N(\vec{\theta}, I)$ are inadmissible wrt L2-error loss of all $\vec{\theta}$ elements simultaneously.
4. ***Complete class*** of decision rules: one which covers all the admissible rules.
    a. E.g. all function-of-sufficient-statistic decision rules under convex loss, or all deterministic rules under such loss.
    b. ***Minimal complete class***: no subset of it is still complete.
        i. Such class is actually the set of all admissible decision rules.

Minimizing average risk

1. ***Bayes risk***: $\boldsymbol{r(\Pi, \delta)} \coloneqq \boldsymbol{E_{\Pi}[R(\theta, \delta)]}$ wrt some prior $\pi(\theta)$.
2. ***Bayes rule***: $\boldsymbol{\delta_{\Pi}} \coloneqq \boldsymbol{\operatorname{argmin}_{\delta} r(\Pi, \delta)}$ (if exists).
3. ***Posterior risk***: $\int_{\Theta} L(\theta, \delta(x)) \Pi_x(d\theta)$.
    a. By change of integration order, $r(\Pi, \delta) = \int_X (posterior\ risk) P_{\Pi}(dx)$, thus minimizing posterior risk is useful to find the Bayes rule. For example:
        i. Squared error: $\delta_{\Pi}(x) = E(\theta|x)$ =mean of posterior.
        ii. Absolute error: $\delta_{\Pi}(x)$ =median of posterior.
4. Any (continuous with finite risk) **Bayes rule is admissible** (since any dominating rule would in particular be better wrt the Bayes risk).
    a. Does any admissible rule minimize average (Bayes) risk according to some weights (prior)?
5. ***Generalized Bayes rule***: Bayes rule wrt some measure $\boldsymbol{\Pi}$ **which isn't necessarily a probability** (in particular X may have infinite measure).
    a. It is argued that such improper prior is clearly legit here since it's a technical tool used to build decision rules with "good" properties, without any probabilistic interpretation.
    b. Posterior may be improper as well. Admissibility is still guaranteed if R is Π-integrable, but this often doesn't hold – e.g. for $X \sim N(\mu, 1)$ and $\pi(\mu) = 1$, risk of $\delta(x) = \bar{x}$ (which is also MLE) is constant, thus not integrable.
6. **A decision rule is admissible if it is the "limit" (in terms of risk) of generalized Bayes rules of finite-measure priors** (as long as these measures don't "neglect" open sets).
    a. Specifically: if R is continuous, $\{\Pi_s\}_1^{\infty}$ are finite (not necessarily 1-sum) measures with existing generalized Bayes rules δ_{π_s}, $\liminf \Pi_s > 0$ for any open set, and $r(\Pi_s, \delta) - r(\Pi_s, \delta_{\Pi_s}) \to 0$, then δ is admissible.
    b. If case of **exponential-family model**, the series of measures converges (at least in subsequence) $\Pi_{s_n} \to \Pi$, thus **this limit is itself a generalized Bayes rule** δ_{Π}.
    c. All these **"limits" of generalized Bayes rules form a complete class**.
    d. Example: for $X \sim N(\mu, 1)$ and $\Pi_s \coloneqq \sqrt{s} N(0, s) \to const$, MLE is $\delta(x) = x$ and generalized Bayes rule is $\delta_s(x) = \frac{s}{s+1} x$. Bayes risks difference wrt squared-error loss ($(\delta(x) - \mu)^2$) is $\sqrt{s} - s^{3/2}/(s+1) \to 0$, so MLE is admissible.

e. **Generalized admissibility theorems are available for standard estimators in exponential families**.

Minimizing maximum risk

1. ***Minimax decision rule*** δ_0: $\forall\delta\colon \sup_\theta R(\theta,\delta_0) \le \sup_\theta R(\theta,\delta)$.
   a. Decision mechanism which is focused on **minimizing worst-case scenario's cost**.
   b. Originated in zero-sum games & adversary situations – often less relevant in statistical decision problems.
   c. In certain problems, **where conventional estimators like MLE are disadvantageous (e.g. if the parameter's dimension increases with the data)**, then ***asymptotically minimax*** procedures can provide **useful benchmark**.
2. Theorem: a **constant-risk Bayes rule** ($\sup_\theta R(\theta,\delta_\Pi) = r(\Pi,\delta_\Pi) \coloneqq E_\Pi[R(\theta,\delta_\Pi)]$) **is minimax**.
   a. Accordingly, **minimax rules can be found by setting parameters of the prior such that the risk will be independent of $\boldsymbol{\theta}$**.
   b. ***Least-favorable prior***: such "worst-case scenario" prior, for which average risk = max risk.
      i. Least-favorable prior Π induces the Bayes rule δ_Π with the largest risk $r(\Pi,\delta_\Pi)$.
3. The standard estimator $\bar{X}$ in location problems under the squared-error loss is a minimax rule (as admissible rule with constant risk).
4. For 1 sample of d-dimensional $X\sim N_d(\theta,\Sigma)$ (with known Σ), X is a minimax estimator under any loss $L(\theta,a) = W(a-\theta)$ for bowl-shaped (i.e. symmetrically-increasing around the origin) W.

Minimizing risk under constraints

1. Minimizing a function of the risk (mean/max over θ as above) is required since no decision rule minimizes the risk uniformly for all θ. However, **uniform minimization of risk is sometimes possible for a constrained class of decision rules**.
2. Unbiasedness constraints:
   a. Reminder – *unbiased estimator*: $\forall\theta\colon E_\theta[\hat{\theta}] = \theta$. In other words, if θ is correct, then it is "expected" to be estimated. A generalization for decision rules says that if θ is correct, then its expected Loss is the smallest in Θ.
   b. ***Unbiased decision rule*** $\delta\colon \boldsymbol{E_\theta}[L(\boldsymbol{\theta},\delta(X))] \le \boldsymbol{E_\theta}[L(\boldsymbol{\theta'},\delta(X))] \quad \forall\theta'$
   c. Theorem: **Bayes estimators are biased** (up to some degenerated priors).
   d. ***Lehmann-Scheffe***: by Rao-Blackwell (above), only a function of sufficient statistic T can be admissible rule δ. **If T is also complete and only unbiased rules are considered, then $\boldsymbol{\delta}$ is unique and uniformly minimizes the risk**.
      i. Specifically: in estimation problem (of some $g(\theta)$) with complete sufficient statistic T and convex loss, if an unbiased estimator exists, then it's essentially unique, a function of T, and uniformly minimizes the risk.
      ii. Note: unbiased estimator does not always exist (e.g. estimating $1/\theta$ in $Bin(n,\theta)$).
3. ***Equivariance constraints***: if we only consider invariant decision rules (i.e. with structure of the form $\delta(gx) = \tilde{g}\delta(x)$, such as standard location & scale estimators), then under some conditions, uniform risk minimization can be achieved as Bayes rule wrt corresponding prior.
   a. *Invariant function* satisfies $f(gx) = f(x)$ wrt some group of transformations $\{g\colon X\to X\}$.
   b. ***Equivariant function*** satisfies $\boldsymbol{f(gx) = \tilde{g}f(x)}$ wrt some two groups of transformations.

   i. E.g. standard **location estimator** $\bar{x}$ wrt shifts $\{g_d(x) = x - d\}_d$ and $\tilde{g} = g$: if x is shifted by d then so is the estimate. Also **scale estimator** wrt multiplication.
c. *Invariant decision problem*: only equivariant decision rules; and loss function doesn't change under transformed θ along with correspondingly-transformed a.
d. **In invariant decision problem**, under certain assumptions, if Bayes rule wrt a certain prior (right invariant Haar prior) exists, then it **minimizes risk uniformly over equivariant rules**.
   i. ***Haar measure***: measure of volume of sets which is invariant to some chosen group of topological operations (e.g. Lebesgue measure wrt constant additions).

4. **Type I error constraints**:
   a. In hypothesis tests, the tradeoff between type-I & type-II errors is usually handled by fixing type-I error and minimizing type-II error.
   b. For **simple-vs.-simple** hypotheses, ***Neyman-Pearson*** tells that for certain significance α, the most powerful test is given by determining a threshold k_α for the likelihood ratio $p_1(x)/p_0(x)$ – smaller decides to accept H_0, larger decides to accept H_1, and equal (in discrete-data problems, where it is not zero-measure case) decides to randomize.
   c. **In decision problem of simple-vs.-simple hypothesis testing with 0-1 loss, the risk is the sum of type-I & type-II error probabilities, hence the rules δ_α corresponding to Neyman-Pearson cover all the admissible rules** (any other rule with significance α has smaller power than δ_α, so its risk is larger).
   d. **In one-sided problem (e.g. $H_1: \theta > \theta_0$) Neyman-Pearson can often be generalized** to depend on θ_0 only (from the simple H_0).
   e. **In two-sided problems there is generally no uniformly-most-powerful test** for a given α (since after determining α there's still a DoF for tradeoff between lower threshold (determining power for $\theta < \theta_0$) and upper threshold (determining power for $\theta > \theta_0$)). However, **there's often a uniformly-most-powerful <u>unbiased</u> test** (the unbiasedness determines the tradeoff between lower & upper thresholds).

Asymptotic Theory and Optimization-Based Estimators

M- and Z-estimators

1. $Pf \coloneqq \int f dP$ (*deFinetti notation* for probability measure P).
2. ***Empirical distribution***: $P_n \coloneqq \frac{1}{n}\sum \delta_{X_i}$, ***empirical average***: $P_n f = \frac{1}{n}\sum f(X_i)$.
3. ***Z-estimator***: $\{\boldsymbol{\theta}: \boldsymbol{Z_n(\theta)} = \mathbf{0}\}$ for $Z_n(\theta) \coloneqq P_n z_\theta$ for some $z_\theta: X \to R$.
4. ***M-estimator***: $\boldsymbol{\operatorname{argmax}_\theta M_n(\theta)}$ for $M_n(\theta) \coloneqq P_n m_\theta$ for some $m_\theta: X \to R$.
    a. May be equivalent to Z-estimator of derivative of M_n (if smooth with single extremum).
5. Examples:
    a. MLE: $m_\theta \coloneqq \log p_\theta$.
    b. Least-squares: $m_\theta(x,y) \coloneqq -\big(y - f_\theta(x)\big)^2$ (under the model $E[Y|x] = f_\theta(x)$).
    c. Median: $z_\theta \coloneqq \chi_{x>\theta} - 1_{x<\theta}$ (generalization for quantiles is possible).
    d. Location estimation: mean ($z_{\theta(x)} \coloneqq x - \theta$) and median ($sign(x-\theta)$) can be generalized to $z_\theta(x) = g(x-\theta)$, e.g. ***Huber's estimator*** $g_k(u) \coloneqq$if ($|u| \le k$) u else const, which allows a controllable tradeoff between the smooth mean and the outlier-insensitive median.
6. M- and Z-estimators **do not require a model $\boldsymbol{p_\theta(x)}$**, which both saves efforts and prevents model-biases.
7. **<u>Consistency</u>**:
    a. By the **Law of Large Numbers**: $\boldsymbol{M_n(\theta)} \to \boldsymbol{M(\theta)} = Pm_\theta$ in probability, pointwise in θ.
    b. If the convergence is uniform in θ (e.g. if $\forall x$: $m_\theta(x)$ is continuous in compact domain of θ), and if M has essentially unique maximum θ^*, then a sequence with $M_n(\hat{\theta}_n) \ge M_n(\theta^*) - o_P(1)$ satisfies $\boldsymbol{\hat{\theta}_n} \to \boldsymbol{\theta^*}$ in probability.
    c. MLE & non-linear least squares can be proved to be consistent (under corresponding conditions) as private cases.
8. **<u>Asymptotic normality</u>**:
    a. Z-estimator: $\sqrt{\boldsymbol{n}}\left(\boldsymbol{\hat{\theta}_n} - \boldsymbol{\theta^*}\right) \to \boldsymbol{N}\left(0, \frac{Pz_{\theta^*}^2}{\dot{Z}(\theta^*)^2}\right)$ in distribution, if: $\hat{\theta}_n \to \theta^*$; finite 2nd moment ($Pz_{\theta^*}^2 < \infty$); and $z_\theta(x)$ satisfies Lipschitz condition (bounded differences) wrt θ and is differentiable at θ^*.
        i. Note: **these conditions are quite weak** – MLE asymptotic normality traditionally assumes stronger conditions (two continuous derivatives rather than just Lipschitz condition) (how can it be if MLE is a private case?).
    b. M-estimators have similar asymptotic normality with slightly more complicated technicalities.
    c. ***Donsker classes***: classes of functions $\{z_\theta: \theta \in \Theta\}$ or $\{m_\theta: \theta \in \Theta\}$ for which the asymptotic normality holds.
        i. Note: asymptotic normality essentially means "θ-uniform Central Limit Theorem". Indeed, **asymptotic normality can be derived from "uniform CLT" just as consistency was derived from "uniform LLN"**, though the definition of uniform CLT is trickier.

Asymptotic normality and optimality

1. Conventional regularity conditions include some smoothness of the model $p_\theta(x)$ (two continuous derivatives or something?). Such conditions are sometimes not satisfied or even hard to define (if Θ isn't a standard space).
2. ***Differentiability in Quadratic Mean*** (***DQM***): **weak & robust definition for θ-smoothness of $\sqrt{p_\theta}$**, based on integration (not derivation) on 2nd-order Taylor expansion, with $\dot{l}_\theta$ as non-differentiable analog of the score $\frac{\dot{p}_\theta}{p_\theta} = \frac{\partial}{\partial\theta}\log p_\theta$: $\quad \exists \dot{l}_\theta: \ \int_X \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}\dot{l}_\theta\sqrt{p_\theta}h\right]^2 d\mu = o(h^2)$.
   a. Continuously-differentiable $\sqrt{p_\theta}$ ➔ DQM, exponential family ➔ DQM.
3. **DQM is sufficient for many properties**, e.g. zero mean & finite variance of the score function (the latter is actually FI), and adjusted version of Cramer-Rao inequality.
4. ***Local asymptotic normality***: incomplete section, but probably means to say that CLT (mean ~ Normal around the expectation) is guaranteed under minimal conditions based on DQM.
   a. In particular, asymptotic normality is claimed to be a very general property of iid models – and not a property of estimators.
5. Irregular example – Hodge's estimator: $\hat{\theta}_n^H := \begin{cases} \hat{\theta}_n & |\hat{\theta}_n| \geq \frac{1}{n^{1/4}} \\ 0 & |\hat{\theta}_n| < \frac{1}{n^{1/4}} \end{cases}$.
   a. If $\theta^* \neq 0$ it's asymptotically identical to $\hat{\theta}$, and if $\theta^* = 0$ then asymptotically $\hat{\theta}^H \equiv \theta^*$.
   b. Such "superefficiency" is generally possible only in zero-measure subsets of Θ.
   c. For any $\theta^* \neq 0$, $\hat{\theta}_n^H$ is indeed n-asymptotically identical to $\hat{\theta}_n$, but for any n, $\hat{\theta}_n^H$ has certain θ^* with larger errors, and in particular the max error $\max_{\theta^*\in\Theta} n \cdot E[MSE]$ is unbounded (the MSE does not θ^*-uniformly goes to 0), making Hodge's estimator irregular.

More Bayesian asymptotics

1. Bernstein-von-Mises theorem guarantees consistency & asymptotic normality of posterior mean under certain conditions. These conditions may be violated in complex models (e.g. infinite-dimensional parameters, as in Neyman-Scott example). Fortunately, more robust results are available in such cases.
2. ***Consistent posterior*** Π_n (X was omitted from the notation for convenience): $\Pi_n(U^C) \to 0$ for any neighborhood U of θ^* (all posterior mass is asymptotically concentrated around θ^*).
3. ***KL-condition***: $\forall \epsilon > 0: \Pi(\{\theta: K(p_{\theta^*}, p_\theta)\}) > 0$ (K = KL-divergence).
   a. I.e. positive prior probability to any "KL-neighborhood" of θ^*).
   b. Since θ^* is unknown, that should hold for any $\theta^* \in \Theta$.
4. ***Well-separated*** θ^*: $\quad \forall U_{\theta^*}: \inf_{\theta\notin U} K(p_{\theta^*}, p_\theta) > 0$ (all θs out of neighborhood are KL-far).
5. KL-condition + good-separation + some unclear uniform LLN:
   a. ➔ **consistent posterior**.
   b. ➔ if prior mean exists, then **posterior mean satisfies $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}} \to \boldsymbol{\theta}^*$** with probability 1.
6. **Likelihood-free Bayes posterior**: consistency holds for any $\widetilde{\Pi}_n(A) \propto \int_A e^{-nL_n(\theta)}\Pi(d\theta)$ ("pseudo-posterior"), where $L_n(\theta) = P_n k_\theta$ is the empirical version of some **loss function** $L(\theta) = Pk_\theta$ which is minimized at $\theta = \theta^*$.
   a. **Standard Bayes posterior is a private case with log-likelihood loss** $L_n = -\frac{1}{n}\log L(\theta)$.

   b. **Preventing necessity of model, likelihood, and marginalization** (i.e. if $\theta = (\psi, \lambda)$ with only ψ of interest, there's no need to estimate all θ only to estimate ψ).
7. Example – Syring & Martin (2015):
   a. Frequentist hypothesis test for effectiveness of a medical treatment can only verify significance of effect (i.e. whether it exists) – not its magnitude or meaning.
   b. ***Minimal clinically important difference*** (***MCID***) is the minimal treatment outcome ($X \in X \subset R$) required to make an "important" difference $Y \in \{-1,1\}$, often measured through either survey among patients (*anchor based*) or expert panel (*Delphi method*).
   c. Desired MCID can be defined as t for which $X > t \Leftrightarrow P(Y = 1) > 1/2$.
      i. This corresponds to minimizing the loss $L_n = P(Y \neq sign(X - t))$.
   d. Standard Bayesian approach may use logistic model $p_{\alpha,\beta}(Y = 1|X = x) = \frac{1}{1+e^{-(\alpha+\beta x)}}$ and calculate posterior for $T \coloneqq -\alpha/\beta$. This requires:
      i. Assuming the logistic model.
      ii. Choosing priors for α and β.
      iii. Estimating both α and β just to derive $T = -\alpha/\beta$.
   e. Specifically, Syring & Martin simulated data from underlying model which is very different from the logistic model, yielding poor Logistic-model-based Bayesian posterior, hence demonstrating the sensitivity to the choice of model (section 3.1 in the paper cited above).
   f. They indeed used the loss function described above as an alternative to the negative log-likelihood in Bayes posterior, and achieved better posterior of T.
      i. This requires only a prior for T.
      ii. It is worthy to note that from the first place T was essentially defined by the loss function, which was justified, but gave clear advantage to the loss-based method.
8. **Essentially, the likelihood-free Bayes posterior just assigns probabilistic meaning to a variant of the loss function ($e^{-nL_n(\theta)}$), and naturally guarantees consistency**.
   a. Indeed, if you need to choose a parameter for decision-making, and you measure your success by a known loss function, and you don't know very well how to model your data-generating-mechanism, and you want to enjoy something-looking-like-distribution of the "correct" parameter – then likelihood-free Bayes posterior may be very convenient.