Statistical Inference

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC, USA
Spring, 2001

February 7, 2001

1. Properties of Estimators

Let X_j be a sequence of independent, identically distributed random variables taking value in some space $\mathcal{X}$, all with pdf $f(x|\theta)$ for some uncertain parameter $\theta \in \Theta$, and let $T_n(X_1, ..., X_n)$ be a sequence of estimators of θ— i.e., of functions $T_n : \mathcal{X}^n \to \Theta$ intended to satisfy $T_n(X) \approx \theta$. An example to keep in mind would be $X \sim \mathsf{No}(\theta, 1)$ with $\theta \in \Theta = \mathbb{R}$, and $T_n(X) = \overline{X}_n = (X_1 + \cdots + X_n)/n$. Let us now explore different ways of making the intention "$T_n(X) \approx \theta$" more precise.

1.1. Bias

The **Bias** of the estimator $T_n(x)$ is simply the expected difference, $\beta_n(\theta) = \mathsf{E}[T_n(X) - \theta \mid \theta]$. An estimator is called *unbiased* if $\beta_n \equiv 0$, i.e.

$$\mathsf{E}[T_n(X) \mid \theta] \equiv \theta$$

for all $n \in \mathbb{N}$ and all $\theta \in \Theta$, and is called *asymptotically unbiased* if $\beta_n \to 0$ as $n \to \infty$, i.e.

$$\lim_{n \to \infty} \mathsf{E}[T_n(X) \mid \theta] = \theta.$$

An unbiased estimator will satisfy the goal "$T_n(X) \approx \theta$" in the sense that the average value of $T_n(X)$ over many replications will be θ. This will offer little or no comfort to someone using the estimator only once, since the possibility remains that perhaps $[T_n(X)-\theta]$ will be huge and positive sometimes, huge and negative at others, with large deviations that average out to zero in the end. Asymptotic unbiasedness offers even less comfort, but its absence would be alarming. What is needed are bounds or conditions on $|T_n(X)-\theta|$.

Most of the criteria below depend, in one way or another, on the **Risk Funcion**

$$R(T_n,\theta) \equiv \mathsf{E}\left[|T_n(X)-\theta|^2 \mid \theta\right];$$

we would like this to be small.

1.2. Consistency

The estimator $T_n(x)$ is called **Consistent** if it always converges to the right answer θ as $n \to \infty$, i.e., if the random variable $|T_n(X)-\theta| \to 0$ in some sense. Each sense in which random variables may converge leads to a slightly different notion of consistency; the most commonly used is "L^2 consistency", where the requirement is that $R(T_n,\theta) \to 0$ as $n \to \infty$ for each $\theta \in \Theta$, i.e.,

$$\lim_{n\to\infty} \mathsf{E}\left[|T_n(X)-\theta|^2 \mid \theta\right] = 0.$$

By adding and subtracting the mean $\mathsf{E}[T_n(X)]$ we can see that L^2 consistency is equivalent to the two requirements $\beta_n \to 0$ and $\mathsf{V}[T_n] \to 0$ that the estimator be asymptotically unbiased and that its variance converge to zero.

A weaker requirement would be consistency *in probability*, i.e. that for each $\epsilon > 0$, $\mathsf{P}[|T_n(X)-\theta| > \epsilon \mid \theta] \to 0$ as $n \to \infty$; this is weaker than L^2 consistency by Chebychev's inequality, since $\mathsf{P}[|T_n(X)-\theta| > \epsilon] \le \mathsf{E}\left[|T_n(X)-\theta|^2\right]/\epsilon^2$.

1.3. Efficiency

Under suitable regularity conditions an L^2-consistent estimator $T_n(x)$ will have a limiting squared error of the form $R(T_n,\theta) \approx c_T/n$ for some $c_T > 0$ that may depend on θ, i.e.,

$$nR(T_n,\theta) = n\mathsf{E}\left[|T_n(X)-\theta|^2\right] \to c_T$$

for some number $c_T > 0$ (in fact they will usually obey a stronger condition, asymptotic normality, with $\mathsf{P}[\sqrt{n}(T_n(X)-\theta) \le z] \to \Phi(c_T z)$ as $n \to \infty$). Evidently it would be preferable to have c_T small. The estimator S is called "more efficient than" T if $R(S_n,\theta) \le R(T_n,\theta)$, and

$R(T_n,\theta)/R(S_n,\theta)$ is called the **relative efficiency**; it will be asymptotically $R(T_n,\theta)/R(S_n,\theta) \to c_T/c_S$ in the usual case, equal to the ratio of sample-sizes N_S/N_T needed by the two estimators to achieve the same expected-squared-error. Harold Cramèr and C.R. Rao found a lower bound $R(T_n,\theta) \geq c_I/n$ for some $c_I(\theta) > 0$, so it is possible to quantify efficiency on an absolute scale as the ratio $c_I/c_T \leq 1$; this was called the "Cramèr-Rao lower bound" in the literature until Lehmann brought it to everyone's attention that Frechèt had proven the same result earlier; now it's called the Information Inequality. I'll prove it below.

1.4. Robustness

Sometimes a probability model is in doubt, or just wrong— we may believe (or prefer to act as if) $X_j \sim f(x \mid \theta)$, for example, but make our inference about $\theta \in \Theta$ on the basis of observations $X_j \sim f^*(x \mid \theta)$ from a different family of distributions [add an example about outliers here]. An estimator T_n is called **robust** if it still satisfies $T_n(X) \approx \theta$, even for data $X_j \sim f^*(x \mid \theta)$ from a somewhat different distribution. It is hard to be more precise about the meaning without the context of a specific example; we'll return to this later.

1.5. Admissibility

An estimator T is called (squared-error) **Admissible** if there does *not* exist another S satisfying $R(S,\theta) < R(T,\theta)$ for all $\theta \in \Theta$ (more precisely, satisfying $R(S,\theta) \leq R(T,\theta)$ for all θ and $R(S,\theta') < R(T,\theta')$ for at least one $\theta \in \Theta$). It can be argued that one should never use an *in*admissible estimator T, since another S exists that is never worse and is sometimes better— but, perhaps astonishingly, Charles Stein and W. James showed that one of the most commonly-used and recommended estimators, $\overline{X}_n$ for the normal mean, is inadmissible in $p > 3$ dimensions (see below).

1.6. Bayes Risk

In the Bayesian paradigm the parameter θ is an uncertain quantity, so estimator features like unbiasedness have no appeal at all; conversely the sample-size n and the data $X = (x_1, \cdots, x_n)$ are both observed and hence not uncertain, so features about averages over other possible X's or limits as $n \to \infty$ have little appeal either. A desirable Bayesian property for an estimator $T_n()$ would be that $T_n(X) \approx \theta$ in the sense that, **given** n **and** X, the probability that θ lies far from the observed number $T_n(x)$ is small, or the expected squared distance is small. The most frequently cited quantity is the Bayes risk for specified prior distribution $\pi(d\theta)$,

$$\begin{aligned} r(T_n, \pi) &= \int_\Theta R(T_n, \theta)\, \pi(d\theta) \\ &= \iint_{\Theta \times \mathcal{X}^n} |T_n(x) - \theta|^2 \, f(x \mid \theta) dx\, \pi(d\theta) \\ &= \iint_{\Theta \times \mathcal{X}^n} |T_n(x) - \theta|^2 \, \pi(d\theta \mid x)\, f(x)\, dx \end{aligned}$$

which is evidently minimized over all possible estimators T_n by the Bayesian posterior mean estimator,

$$T_n^\pi(x) \equiv \mathsf{E}[\theta \mid X_n = x] = \int_\Theta \theta\, \pi(d\theta \mid x) = \frac{\int_\Theta \theta\, f_n(x \mid \theta) \pi(d\theta)}{\int_\Theta f_n(x \mid \theta) \pi(d\theta)}.$$

It is a remarkable fact that *every* unique Bayes estimator is admissible. Suppose, for contradiction, that for some prior distributibution $\pi(d\theta)$ on Θ the Bayes posterior mean T^π were *not* admissible; then there would exist another estimator S with $R(S, \theta) \le R(T^\pi, \theta)$ and $R(S, \theta^*) < R(T^\pi, \theta^*)$ for some $\theta^* \in \Theta$. After integrating over Θ with respect to the prior, this gives $r(S, \pi) \le r(T^\pi, \pi)$, so S too attains minimum Bayes risk— and by uniqueness must be equal to T^π, contradicting $R(S, \theta^*) < R(T^\pi, \theta^*)$. The standard method for constructing admissible estimators is to look at Bayesian posterior means, for a range of possible prior distributions $\pi(d\theta)$.

2. Normal Distribution Inference

Let $X = (X_1, \cdots, X_n)$ be a random sample from the Normal $\mathsf{No}(\mu, \sigma^2)$ distribution; the joint pdf (hence likelihood) is

$$
\begin{aligned}
f(x|\mu,\sigma^2) &= (2\pi\sigma^2)^{-n/2}e^{-\sum(x_i-\mu)^2/2\sigma^2}\\
&= (2\pi\sigma^2)^{-n/2}e^{-\sum(x_i-\overline{X}_n)^2/2\sigma^2-n(\overline{X}_n-\mu)^2/2\sigma^2}\\
&= (2\pi\sigma^2)^{-n/2}e^{-(n/2\sigma^2)[S_n^2+(\overline{X}_n-\mu)^2]}, \text{ where}\\
\overline{X}_n &= \frac{1}{n}\sum x_i \text{ and}\\
S_n^2 &= \frac{1}{n}\sum(x_i-\overline{X}_n)^2
\end{aligned}
$$

are the maximum likelihood estimates (MLE's) for μ and σ^2, respectively. Evidently the likelihood depends on the data only through these statistics; since S_n^2 depends only on $[X-\overline{X}_n]$, a normal vector independent of $\overline{X}_n$, it follows that the random variables $\overline{X}_n$ and S_n^2 are independent. Their distributions are

$$
\overline{X}_n \sim \mathsf{No}(\mu,\sigma^2/n) \qquad S_n^2 \sim \mathsf{Ga}\left(\frac{n-1}{2},\frac{2\sigma^2}{n}\right),
$$

respectively, so the re-scaled quantity

$$
Y \equiv \frac{n}{\sigma^2}S_n^2 \sim \mathsf{Ga}\left(\frac{n-1}{2},2\right) = \chi^2_{n-1}
$$

has a χ^2_ν distribution with $\nu=(n-1)$ degrees of freedom,

$$
Z \equiv \frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}
$$

has a standard $\mathsf{No}(0,1)$ distribution, and

$$
t \equiv \frac{\overline{X}_n-\mu}{S/\sqrt{n-1}} = \frac{Z}{\sqrt{Y/(n-1)}}
$$

has a distribution, "Student's t_{n-1}", that doesn't depend on μ or σ^2 at all. William Gossett, writing under the *nôm de plume* "Student" because his employer, the Guiness brewery, didn't allow its employees to publish, saw that this could provide a basis for inference about a normal mean μ when the variance σ^2 is unknown, and computed the probability density function

$$
f_\nu(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left[1+\frac{t^2}{\nu}\right]^{-\frac{\nu+1}{2}}.
$$

It follows that for any $n \in \mathbb{N}$ and $\alpha \in (0,1)$ we can find the number $t_{\alpha/2} =$ `qt(1-alpha/2,nu)` such that $\mathsf{P}[|t| > t_{\alpha/2}] = \alpha$ and note that

$$\begin{aligned}
1-\alpha &= \mathsf{P}[-t_{\alpha/2} \le t \le t_{\alpha/2}] \\
&= \mathsf{P}[\frac{-t_{\alpha/2}S_n}{\sqrt{n-1}} \le \overline{X}_n - \mu \le \frac{t_{\alpha/2}S_n}{\sqrt{n-1}}] \\
&= \mathsf{P}\left[\mu \in \left(\overline{X}_n - \frac{t_{\alpha/2}S_n}{\sqrt{n-1}},\ \overline{X}_n + \frac{t_{\alpha/2}S_n}{\sqrt{n-1}}\right)\right],
\end{aligned}$$

giving a random interval $\overline{X}_n \pm t_{\alpha/2}S_n/\sqrt{n-1}$ (called a *confidence interval*) which will contain the uncertain quantity μ with the specified probability $1-\alpha$. It is sometimes useful to notice that if $t \sim t_\nu$ then t^2 has an F^1_ν distribution and that $t^2/(t^2+\nu) \sim \mathsf{Be}(1/2, \nu/2)$; the latter allows one to use the incomplete Beta function to compute t probability integrals in C or Fortran.

In the limit as $\nu \to \infty$ the density converges $f_\nu(t) = c_\nu(1+t^2/\nu)^{-(\nu+1)/2} \to c_\infty e^{-t^2/2}$ to the standard normal distribution, so in the limit the number $t_{\alpha/2} =$ `qt(1-alpha/2,nu)` is approximately $z_{\alpha/2} =$ `qnorm(1-alpha/2)`, but for any $\nu < \infty$ we have $t_{\alpha/2} > z_{\alpha/2}$ so approximate intervals of the form $\overline{X}_n \pm z_{\alpha/2}S_n/\sqrt{n-1}$ would be too short and would fail to contain μ with probability greater than α.

Note: As an estimator of σ^2 the maximum likelihood estimator $S_n^2 = \sum(x_i - \overline{X}_n)^2/n$ is biased, since $\beta_n = \mathsf{E}[S^2 \mid \sigma] - \sigma^2 = (n-1)\sigma^2/n - \sigma^2 = -\sigma^2/n \ne 0$ (it is *asymptotically* unbiased, of course). Some authors prefer to define "sample variance" by $S_n^2 \equiv \sum(x_i - \overline{X}_n)^2/(n-1)$, which does satisfy $\mathsf{E}[S_n^2 \mid \sigma^2] \equiv \sigma^2$; our intervals above will be recovered if we replace each $\sqrt{n-1}$ with $\sqrt{n}$.

2.1. Example: Estimating the Normal Mean

The most obvious estimator of the normal mean μ is the sample mean $T_n^1(x) = \overline{X}_n = \sum_{i=1}^n x_i/n$, but I would also like to consider two others: the sample *median* $T_n^2(x) = X_{(m+1)}$, the $m+1$'st-smallest observation if $n = 2m+1$ is odd, or $T_n^2(x) = (X_{(m)} + X_{(m+1)})/2$, the average of the two middle values, if $n = 2m$ is even; and, for any $\xi \in \mathbb{R}$ and $\tau > 0$, the weighted average $T_n^3(x) = [n\sigma^{-2}\overline{X}_n + \tau^{-2}\xi]/[n\sigma^{-2} + \tau^{-2}]$ (we'll see later that this is the **Bayes estimator** when σ^2 is regarded as fixed and μ is accorded a Normal $\mu \sim \mathsf{No}(\xi, \tau^2)$ prior distribution). To simplify life we'll take $n = 2m+1$

odd, so that

$$T_n^1(x) = \frac{\sum_{i=1}^n x_i}{n} \qquad T_n^2(x) = X_{(m+1)} \qquad T_n^3(x) = \frac{\sigma^{-2}\sum_{i=1}^n x_i + \tau^{-2}\xi}{n\sigma^{-2} + \tau^{-2}}.$$

2.1.1. Bias

The means are $\mathsf{E}[T_n^1] = \mu$ and (by symmetry) $\mathsf{E}[T_n^2] = \mu$, while evidently $\mathsf{E}[T_n^3(x)] = (n\tau^2\mu + \sigma^2\xi)/(n\tau^2 + \sigma^2) = \mu + (\xi - \mu)/(n(\tau/\sigma)^2 + 1)$, so T_1 and T_2 are unbiased while T^3 is only asymptotically unbiased.

2.1.2. Consistency

The sample mean $\overline{X}_n \sim \mathsf{No}(\mu, \sigma^2/n)$ has a normal distribution with mean μ and variance σ^2/n, so T^1 is L^2 consistent with $\mathsf{E}[|T_n^1 - \mu|^2] = \sigma^2/n \to 0$. A little more arithmetic shows that $\mathsf{E}[|T_n^3 - \mu|^2] = (n\sigma^2 + r^2\delta^2)/(n + r)^2$, where $\delta \equiv (\xi - \mu)$ and $r \equiv (\sigma^2/\tau^2)$, so $\mathsf{E}[|T_n^3 - \mu|^2] \to 0$ as well.

The median is more fun. Recall that the Beta distribution $\mathsf{Be}(\alpha, \beta)$ has mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, or $\frac{1}{2}$ and $\frac{1}{4(2m+3)}$ in the symmetric case $\alpha = \beta = m + 1$, and that $\mathsf{Be}(\alpha, \beta)$ is asymptotically normal as $\alpha + \beta \to \infty$. From the earlier class notes we have

```
pbinom(x,n,p)=1-pbeta(p,x+1,n-x)
```

for any number $p \in (0, 1)$ and integers $0 \le x \le n < \infty$, so for any number $a \in \mathbb{R}$ the probability that the median $X_{(m+1)} \le \mu + a\sigma/\sqrt{n}$ is the same as the probability that at least $m + 1$ of the $n = 2m + 1$ X_i's are less than $\mu + a\sigma/\sqrt{n}$, an event with probability

$$\begin{aligned}
\mathsf{P}[X_{(m+1)} \le \mu + a\sigma/\sqrt{n}] &= \sum_{k=m+1}^{n} \binom{n}{k} \Phi(a/\sqrt{n})^k \Phi(-a/\sqrt{n})^{n-k} \\
&= \texttt{1-pbinom(m, n, pnorm(a/sqrt(n)))} \\
&= \texttt{pbeta(pnorm(a/sqrt(n)), m+1, m+1)} \\
&\approx \texttt{pbeta(0.5+dnorm(0)*a/sqrt(n), m+1, m+1)} \\
&\approx \Phi\left(\frac{(\frac{1}{2} + \phi(0)a/\sqrt{n}) - \frac{1}{2}}{\sqrt{1/4(2m+3)}}\right) \\
&= \Phi\left(2\sqrt{2m+3}\;\phi(0)a/\sqrt{2m+1}\right) \\
&\approx \Phi\left(2\phi(0)\,a\right),
\end{aligned}$$

so the median $\tilde{X}_n \equiv X_{(m+1)}$ has a limiting distribution given by

$$\sqrt{n}(\tilde{X}_n - \mu) \sim \mathsf{No}(0, \sigma^2/4\phi(0)^2)$$

approximately for large n. Of course we could simplify this distribution to $\mathsf{No}(0, \sigma^2\pi/2)$ using $\phi(0) = 1/\sqrt{2\pi}$, but in fact the result is true more generally: for any $X_j \sim F(x)$ with $F(\theta) = 1/2$ and $f(\theta) = F'(\theta) > 0$, the median $\tilde{X}_n \equiv X_{(m+1)}$ has the limiting distribution $\sqrt{n}(\tilde{X}_n - \theta) \sim \mathsf{No}(0, 1/4f(\theta)^2)$. In particular, asymptotically we have $\mathsf{E}[|T_n^2 - \mu|^2] \approx \sigma^2\pi/2n \to 0$, so T^2 is consistent too.

2.1.3. Efficiency

The Information Inequality

Let $f(x \mid \theta)$ be a density function with the property that $\log f(x \mid \theta)$ is differentiable in θ throughout the open p-dimensional parameter set $\Theta \subset \mathbb{R}^p$; then the **score statistic** (or score function) is defined by

$$Z(X) \quad \equiv \quad \nabla_\theta \log f(x \mid \theta) = \frac{\nabla_\theta f(x \mid \theta)}{f(x \mid \theta)}$$

and the **Fisher** (or **Expected**) **Information** matrix is defined by

$$I(\theta) \quad \equiv \quad \mathsf{E}\left[Z(X)Z(X)' \mid \theta\right];$$

if we may exchange integration with differentiation then we can calculate

$$\begin{aligned}
\mathsf{E}[Z_i(X) \mid \theta] &= \int_{\mathcal{X}} [\frac{d}{d\theta_i} \log f(x \mid \theta)]\, f(x \mid \theta)\, dx \\
&= \int_{\mathcal{X}} \frac{\frac{d}{d\theta_i} f(x \mid \theta)}{f(x \mid \theta)}\, f(x \mid \theta)\, dx \\
&= \int_{\mathcal{X}} \frac{d}{d\theta_i} f(x \mid \theta)\, dx \\
&= \frac{d}{d\theta_i} \int_{\mathcal{X}} f(x \mid \theta)\, dx \\
&= 0
\end{aligned}$$

and hence $\mathsf{E}[Z(X) \mid \theta] = 0$ and $\mathsf{Cov}[Z(X) \mid \theta] = \mathsf{E}\left[Z(X)Z(X)' \mid \theta\right] = I(\theta)$; taking another derivative with respect to θ_j of the equation $\mathsf{E}[Z_i(X) \mid \theta] = 0$

gives, by the product rule,

$$
\begin{aligned}
0 &= \frac{d}{d\theta_j}\mathsf{E}[Z_i(X) \mid \theta] \\
&= \frac{d}{d\theta_j}\int_{\mathcal{X}}[\frac{d}{d\theta_i}\log f(x \mid \theta)]\, f(x \mid \theta)\, dx \\
&= \int_{\mathcal{X}}[\frac{d^2}{d\theta_i\, d\theta_j}\log f(x \mid \theta)]\, f(x \mid \theta)\, dx + \int_{\mathcal{X}}[\frac{d}{d\theta_i}\log f(x \mid \theta)]\,[\frac{d}{d\theta_j}\log f(x \mid \theta)]\, f(x \mid \theta)\, dx \\
&= \mathsf{E}\left[\frac{d^2}{d\theta_i\, d\theta_j}\log f(x \mid \theta)\right] + I(\theta),
\end{aligned}
$$

so we may also compute the Fisher Information as

$$
I(\theta) \quad = \quad \mathsf{E}\left[-\nabla^2_\theta \log f(X \mid \theta)\right],
$$

the matrix of expected negative second derivatives of the log likelihood with respect to θ.

Now let $\Theta \subset \mathbb{R}$ be one-dimensional and let T be any statistic with finite expectation $\psi(\theta) \equiv \mathsf{E}[T(X) \mid \theta]$, and assume additionally that ψ is differentiable throughout Θ to justify exchanging integration and differentiation as follows:

$$
\begin{aligned}
\psi'(\theta) &= \frac{d}{d\theta}\int_{\mathcal{X}} T(x)\, f(x \mid \theta)\, dx \\
&= \int_{\mathcal{X}} T(x)\, \frac{d}{d\theta} f(x \mid \theta)\, dx \\
&= \int_{\mathcal{X}} T(x)\, Z(x)\, f(x \mid \theta)\, dx \\
&= \mathsf{E}\left[T(X)\, Z(X) \mid \theta\right] = \mathsf{Cov}\left[T(X)\, Z(X)\right],
\end{aligned}
$$

so the score statistic $Z(X) \equiv \frac{d}{d\theta}\log f(x \mid \theta)$ has mean zero, variance $I(\theta)$, and covariance $\psi'(\theta) = \mathsf{Cov}[T(X), Z(X)]$ with $T(X)$; by the Covariance Inequality $|\mathsf{Cov}(T,Z)|^2 \leq \mathsf{V}(T)\,\mathsf{V}(Z)$ (Minkowski's inequality), we can conclude that $|\psi'(\theta)|^2 \leq I(\theta)\mathsf{V}(T(X))$, or that

$$
\mathsf{V}(T(X)) \geq \frac{|\psi'(\theta)|^2}{I(\theta)};
$$

in particular, any unbiased estimator T of θ must have risk

$$
R(T, \theta) \geq \frac{1}{I(\theta)}
$$

bounded below by the celebrated Information Inequality.

[Could add examples, No+Po+(perhaps) Exponential Family; at least, mention that $I_n(\theta) = nI(\theta)$ for iid samples]

Efficiency of the Mean and Median

The Fisher Information for n observations from the $\mathsf{No}(\theta, \sigma^2)$ distribution is $I_n(\theta) = n\sigma^{-2}$, so no unbiased estimator can have risk less than $1/I(\theta) = \sigma^2/n$; this bound is attained by the sample mean T_n^1, so no estimator of θ is more efficient than the sample mean $T_n^1 = \overline{X}_n$, with $\mathsf{E}[|T_n^1 - \mu|^2] = \sigma^2/n$. The relative efficiency of the sample median $T_n^2 = \tilde{T}_n$ then is

$$\frac{R(T_n^1, \mu)}{R(T_n^2, \mu)} = \frac{\mathsf{E}[|T_n^1 - \mu|^2]}{\mathsf{E}[|T_n^2 - \mu|^2]} \approx \frac{\sigma^2/n}{\sigma^2\pi/2n} = \frac{2}{\pi},$$

so the sample median will require a sample about $\pi/2 \approx 1.57$ times (57% more) observations than $f_n(t)$ forwould the sample mean to achieve equally small squared errors.

The Bayes estimator has efficiency

$$\frac{R(T_n^1, \mu)}{R(T_n^3, \mu)} = \frac{\mathsf{E}[|T_n^1 - \mu|^2]}{\mathsf{E}[|T_n^3 - \mu|^2]} = \frac{\sigma^2/n}{(n\sigma^2 + r^2\delta^2)/(n+r)^2} = \frac{(n+r)^2\sigma^2}{n^2\sigma^2 + nr^2\delta^2} \to 1,$$

so the Bayes estimate T^3 is also has the highest asymptotic efficiency possible.

Pitman noticed that, if $\sqrt{n}(T_n - \theta) \Rightarrow \mathsf{No}(0, c)$ in a strong enough sense that even the *density* function converges pointwise, then the constant c can be recovered from the pdf $f_n(t)$ for T_n by

$$c_T = \lim_{n\to\infty} \frac{n}{2\pi f_n(\theta)^2}.$$

Applying this to the sample median of $n = 2m + 1$ samples from a distribution with CDF $F(t)$ and pdf $f(t) = F'(t) > 0$, with $F(\theta) = 1/2$, and recalling Stirling's approximation $n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n}$ (with relative error no larger than $e^{1/12n}$),

$$
\begin{aligned}
f_n(\theta+\frac{t}{\sqrt{n}}) &= n\binom{2m}{m}\frac{1}{\sqrt{n}}f(\theta+\frac{t}{\sqrt{n}})\,F(\theta+\frac{t}{\sqrt{n}})^m\,\big(1-F(\theta+\frac{t}{\sqrt{n}})\big)^m \\
&\approx \sqrt{n}\frac{(2m)!}{m!^2}f(\theta)\big(1/2+\frac{t}{\sqrt{n}}\,f(\theta)\big)^m\,\big(1/2-\frac{t}{\sqrt{n}}\,f(\theta)\big)^m \\
&\approx \sqrt{n}\frac{\sqrt{2\pi}(2m)^{2m+1/2}\,e^{-2m}}{(\sqrt{2\pi}(m)^{m+1/2}\,e^{-m})^2\;2^{2m}}\Big(1-4\,\frac{t^2}{n}f(\theta)^2\Big)^m \\
&= \sqrt{\frac{2n}{\pi(n-1)}}f(\theta)\Big(1-4\,\frac{t^2}{n}f(\theta)^2\Big)^{(n-1)/2} \\
&\approx \sqrt{\frac{4\,f(\theta)^2}{2\pi}}\;\exp\big(-2\,t^2 f(\theta)^2\big),
\end{aligned}
$$

a normal density function in t with mean 0 and variance $1/4f(\theta)^2$, so the asymptotic relative efficiency of the Median with respect to the sample Mean (which by the Central Limit Theorem satisfies $\sqrt{n}(\overline{X}_n-\theta)\Rightarrow \mathsf{No}(0,\sigma^2)$) is $\mathsf{ARE}=4f(\theta)^2\sigma^2$. For normal $f(\cdot)=\mathsf{No}(\theta,\sigma^2)$, $f(\theta)^2=1/2\pi\sigma^2$ giving an ARE of $2/\pi$ for the Median, as before.

Pitman also suggested an incredibly easy way to compute ARE's: under the strong conditions that the PDF of $\sqrt{n}(T_n-\theta)$ converges pointwise and in L^1 to a normal distribution with mean 0 and variance c_T, the density function $f_n(x)$ of T_n will satisfy

$$c_T=\lim_{n\to\infty}\frac{n}{2\pi f_n(\theta)^2}$$

so we can pick off the asymptotic relative efficiency simply from the value at θ of the pdf; for the median of normal deviates, this again gives us $c_T=\pi\sigma^2/2$, once again giving an ARE of $2/\pi$.

2.1.4. Robustness

If the true distribution of the X_i is not $\mathsf{No}(\mu,\sigma^2)$ but rather a mixture of 99% $\mathsf{No}(\mu,\sigma^2)$ and 1% of something with much heavier tails, like a Cauchy or even a normal distribution with much larger variance, then there is a chance that the data will include one or more "outliers" from the contaminating distribution; many of the pleasant features of $T_n^1=\overline{X}_n$ and the Bayes estimator T_n^3 are lost now, because $\overline{X}_n$ is quite sensitive to the presence of outliers— for example, the bias and expected squared error are now much larger, perhaps infinite. The median is much less effected by the contamination, and remains consistent and comparatively efficient; for this reason it is called a *robust* estimator, while the sample mean and its relatives are not. Notice now that $R(T_n^1, \mu) = \infty > R(T_n^2, \mu)$.

Tukey proposed a simple "contamination model"

$$X_i \sim \epsilon \mathsf{No}(\theta, \sigma^2) + (1-\epsilon)\mathsf{No}(\theta, \tau^2\sigma^2)$$

for some $\epsilon > 0$, $\tau \gg 1$, the ϵ-mixture of a $\mathsf{No}(\theta, \sigma^2)$ random variable with a normal distribution with the same mean but inflated variance. It is easy to see that this has again mean θ but variance $\sigma^2[1 - \epsilon + \epsilon\tau^2]$, and has a density function whose value at θ is

$$f(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\left(1 - \epsilon + \frac{\epsilon}{\tau}\right),$$

so by Pitman's argument the relative efficiency of the median to the mean are

$$4f(\theta)^2\mathsf{V}(X) = 2(1 - \epsilon + \epsilon/\tau)^2(1 - \epsilon + \epsilon\tau^2)/\pi,$$

giving $2/\pi$ for $\epsilon = 0$ but becoming arbitrarily large as $\tau \to \infty$ for any $\epsilon > 0$ indicating that the median is *more* efficient under a contamination model.

[Could add note on how Bayes estimators with respect to *flat-tailed* priors are robust, while those for conjugate priors are not; for example, Cauchy prior for normal problems, as recommended by Jeffreys.].

2.1.5. Admissibility

An estimator T is called (squared-error) **Admissible** if there does *not* exist another S satisfying $R(S, \theta) < R(T, \theta)$ for all $\theta \in \Theta$ (more precisely, satisfying $R(S, \theta) \le R(T, \theta)$ for all θ and $R(S, \theta') < R(T, \theta')$ for at least one $\theta \in \Theta$). It can be argued that one should never use an *in*admissible estimator T, since another S exists that is never worse and is sometimes better— but, perhaps astonishingly, Charles Stein and W. James showed that one of the most commonly-used and recommended estimators, $\overline{X}_n$ for the normal mean, is inadmissible in $p > 3$ dimensions, with higher risk than the "James-Stein estimator"

$$T^{\mathsf{JS}}(X) = \overline{\overline{X}}_n + \left(1 - \frac{p-3}{\mathsf{SS}_n}\right)(\overline{X}_n - \overline{\overline{X}}_n)$$

where $\overline{\overline{X}}_n$ is the grand average over not only the n observations but also *over the p components* of $\overline{X}_n$ and where $\mathsf{SS}_n = \sum_{i=1}^p (\overline{X}_n^i - \overline{\overline{X}}_n)^2$, the sum of squared differences.

The James-Stein estimator may be viewed as an emprical analogue of the Bayes shrinkage estimator T^3 introduced above, which may be written (for $\sigma^2 = 1$)

$$T^3(X) = \xi + \left(1 - \frac{1}{1 + n\tau^2}\right)(\overline{X}_n - \xi),$$

with "prior mean" $\xi = \overline{\overline{X}}_n$ taken from the componentwise average of the data and "prior variance" τ^2 taken from how variable are the p components. We will see below that **every** Bayes estimator is admissible; thus a common approach to constructing admissible estimators is to choose specific prior distributions and derive their Bayes estimators.

2.1.6. Bayes Risk

The Bayes risk $r(T_n, \pi)$ may (in principle) be calculated for any prior distribution π; the calculation is particularly easy for a

$$\begin{aligned} f_n(\theta \mid X = x) &\propto f_n(x \mid \theta)\, \pi(\theta) \\ &\propto e^{-(n/2\sigma^2)(\overline{X}_n - \theta)^2}\, e^{-(\theta-\xi)^2/2\tau^2} \\ &\propto e^{-(\theta - M)^2/2V}, \end{aligned}$$

again normal but now with mean $M \equiv \frac{n\sigma^{-2}\overline{X}_n + \tau^{-2}\xi}{n\sigma^{-2} + \tau^{-2}} = \frac{\overline{X}_n + (n\sigma^2/\tau^2)\xi}{1 + (n\sigma^2/\tau^2)}$ and variance $V = [n\sigma^{-2} + \tau^{-2}]^{-1}$. The smallest possible Bayes Risk is that of $T_n^3(X) = \mathsf{E}[\theta \mid X] = M$, namely $r(T_n^3, \pi) = V = [n\sigma^{-2} + \tau^{-2}]^{-1}$, while that of the sample mean is $r(T_n^1, \pi) = V + \left(\frac{r^2\delta^2}{r^2+n}\right) V$. Notice that $r(T_n^1, \pi)/r(\pi, T^3) \to 1$ as $n \to \infty$, so for large enough sample sizes both estimators have approximately the same Bayes risk. I'm not sure how to calculate the Bayes risk of the sample median, $r(T_n^2, \pi)$, but clearly it's larger than $r(T_n^3, \pi)$.

3. Where do Estimators Come From?

3.1. Method of Moments

Match first d population moments $\mu_j(\theta) \equiv \mathsf{E}[X^j \mid \theta]$ with sample moments $\hat{\mu}_j \equiv \frac{1}{n}\sum_{i=1}^n X_i^j$, where d is lowest value to give unique solution (usually the dimension of Θ)

3.2. Least Squares

Minimize squared Euclidean distance between observed data vector $Y = \{Y_1, ..., Y_n\}$ and expectation $\mu = \{\mu_1(\theta), ..., \mu_n(\theta)\}$ (common in regression settings)

3.3. Maximum Likelihood

Proposed by Carl Friedrich Gauss (1821), attributed to Sir Ronald Aylmer Fisher (1922). Maximize likelihood $L(\theta)$ (or its logarithm).

3.4. Location and Scale

Let X have any probability distribution with CDF $F(x)$, pdf $f(z)$, and MGF $M(t) = \mathsf{E}[e^{tZ}]$; let $a \in \mathbb{R}$ and $b > 0$. Then $Y = aX + b$ will have a continuous distribution too, with CDF $F_Y\big(a^{-1}(y-b)\big)$, pdf $a^{-1}f_Y\big(a^{-1}(y-b)\big)$, and MGF $M_Y(t) = M(bt)\,e^{ta}$. The family of distributions $y \sim f(y \mid a, b)$ is called the *location-scale family* generated by $f(x)$; familiar examples inlclude the $\mathsf{No}(\theta, \sigma)$, based on the standard $\mathsf{No}(0,1)$ distribution, and the uniform $\mathsf{Un}(b, a{+}b)$, built on the standard $\mathsf{Un}(0,1)$ distribution, but the location-scale families built on the t, Cauchy, exponential, Weibull, and other distributions arise frequently in examples and applications. If the base distribution has a well-defined mean θ and finite variance σ^2 then $Y = aX + b$ has mean $a\theta + b$ and variance $a^2\sigma^2$, so it is possible to estimate location and scale parameters a and b on the basis of sample mean $\overline{X}_n$ (or median or trimmed mean or other measure of centrality) and sample variance S_n^2 (or mean or median absolute deviation or interquartile range or other measure of scale). Any remaining ("shape") parameters are handled using other methods.

3.5. Bayesian Posterior Mean (or Mode or Median)

Choose some prior distribution $\pi(d\theta)$ and evaluate the posterior distribution $\pi(d\theta|X)$ or some measure of its center

3.6. Empirical Bayes

Begin with some family of possible prior distributions $\pi_\gamma(d\theta)$, using a data-dependent choice of γ (Note: This will violate likelihood principle)

3.7. Objective Bayes

Choose a conventional prior distribution $\pi(d\theta)$, depending on the sampling distribution $f(x|\theta)$ for the problem, then proceed as in Bayesian analysis. Most common choices are uniform $\pi(d\theta) = d\theta$ and Jeffreys $\pi(d\theta) = \sqrt{|I(\theta)|}\, d\theta$, where $I(\theta) = \nabla^2 \mathsf{E}[-\log f(X|\theta) \mid \theta]$ is the Fisher Information matrix. (Note: This will also violate likelihood principle. Also, usually $\pi(d\theta)$ is improper, i.e., $\pi(\Theta) = \infty$).

3.8. Ad Hoc