Statistical Data Analysis

289E Obervables nonnegative. They are ordered according to their magnitudes as $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

$$S_i = \sum_{i \le j} x_{(1)} \tag{1}$$

The curve obtained by connecting $f_i = S_i/S_N$ is called Lorenz's curve, which is a convex curve connecting the origin and $(1,1)$. The area enclosed by the diagonal and the curve has an area equal to the mean difference

$$\delta = \sum_{i,j} \frac{|x_i - x_j|}{N^2} \tag{2}$$

divided by 4 times the expectation value. $G = \delta/\langle x \rangle$ is called *Gini's coefficient.*

289G Moment generating function $M(\theta)$

$$M(\theta) = \int e^{\theta x} f(x) dx = E(e^{\theta x}). \tag{3}$$

$\varphi(t) = M(it)$ is called the *characteristic function.*
Probability generating function

$$P(t) = \sum_i f_i t^i. \tag{4}$$

Factorial moment generating function

$$\tilde{M}(t) = P(t+1) = \sum_i \frac{t^i}{i!} f_i. \tag{5}$$

Factorial cumulant can be defined through

$$\tilde{K}(t) = \log \tilde{M}(t) = \sum_i \kappa_{(i)} \frac{t^i}{i!}. \tag{6}$$

The Poisson distribution is characterized by $\kappa_{(j)} = 0$ for all $j \ge 2$.

289H Regression line $y = a + bx$ (or better $\boldsymbol{y} = bvx + \boldsymbol{a}$, where vectors denote samples as $\boldsymbol{x} = (x_1, \cdots, x_N)$): $\sum (y_i - ax_i - b)^2$ is minimized.

$$b = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle / \sigma_x^2, a = \langle y \rangle - b \langle x \rangle. \tag{7}$$

$e_i = y_i - a - bx_i$ is called the residual.

$$r_{xy}^2 = 1 - \sum e_i^2 / \sum (y_i - \langle y \rangle)^2. \tag{8}$$

We can also define the regression of y to x: $x = c + dy$. The two regression curves cross at $(\langle x \rangle, \langle y \rangle)$ and

$$|1/d| = |b/r_{xy}^2| \ge |b|. \tag{9}$$

The regression analysis was initiated by F. Galton.

289J We can generalize the above to more variable cases than two:
The matrix $R = matr.(r_{ij})$, where r_{ij} is the correlation coefficient between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ (where components are the sample observables), is called the *correlation matrix*. *Multiple correlation coefficients* $\hat{R}_i$ is defined as

$$\hat{R}_i = \sqrt{1 - 1/(R^{-1})_{ii}} \tag{10}$$

Here, R^{-1} is the inverse of R, so if R_{ii} is the cofactor of R wrt to r_{ii}, we can write

$$\hat{R}_i = \sqrt{1 - |R|/R_{ii}} \tag{11}$$

This is the correlation between $\boldsymbol{x}_i$ and the linear regression estimate. For $\boldsymbol{x}_k = (x_{k1}, \cdots, x_{kN})$ (N is the total number of samples), it is obtained by minimizing

$$Q = \sum_{s=1}^{N} \left(x_{ks} - a_0 - \sum_{j=1}^{k-1} a_j x_{js} \right)^2 . \tag{12}$$

The coefficients $\boldsymbol{a} = (a_1, \cdots, a_{k-1})$ are determined by

$$V\boldsymbol{a} = \boldsymbol{V}_k, \tag{13}$$

where V is the the correlation matrix for $i, \cdots, k-1$, and $\boldsymbol{V}_k = (V_{k1}, \cdots, V_{kk-1})$. a_0 is determined by

$$a_0 = \langle x_k \rangle - \boldsymbol{a} \cdot \langle \boldsymbol{x} \rangle. \tag{14}$$

The minimized Q reads

$$Q = (1 - R_k^2)\sigma_k^2. \tag{15}$$

• Reconsiderations on regression
When we have a collection of points in a plane, we have a contour distribution. The accuracy of the contour must be commensurate with the reliability of the data. The backbone of the contour is the regression curve.

Coarse-grained contour may be obtained by truncated Fourier expansion. Where to truncate? What is the substitute of AIC? AIC does not pay much attention to the reliability of the points. It determines the degree of the polynomial by the number of data points.
Possible strategy:
(1) Make a coarse grained pattern by Fourier truncation, or moving spatial average.
(2) Choose the core.
(3) Each point is assigned the contour value (larger).
Thus, the key problem is to find reliable distribution cloud.
One answer to this problem is the resampling a la Efron. For a given coarse-graining scale, if the resampling does not change the distribution appreciably that is the desired coarse-grained result.

It is desirable to characterize this as a fixed point point of coarse-graining and resampling. One idea may be to use resampling as a scaling. The density of the points determines the natural cutoff scale. Coarse-grain at this scale. Then. reduce the number of points. This should correspond to decimation. Thus, a fixed point may be obtained.
Q: AIC etc: what is the relation between this approach and the standard information based approach?
Q: What is the Stückelberg-Petermann formalism? □

289K Let $\boldsymbol{i} = (i_1, \cdots, i_k)$ be the qualitative categorization of samples. $N_{\boldsymbol{i}}$ is the number of samples in the category $\boldsymbol{i}$. The table of $N_{\boldsymbol{i}}$ is called the *k-way contingency table.*

289M If the class is naturally ordered, we say that the observation is performed according to the *ordinal scale.* To study the relationships among different ordinal scales, sometimes scores are attached to each class according to each ordinal scale, These scores can be chosen to maximize the correlation among different ordinal scales. For the case with two scales, let C be the contingency table. Then, the Perron-Frobenius eigenvector for CC^T is the answer. In this case no original ordering is respected.

• What should we do with ordinal data?
Two scale cases: two categories are embedded into 1-space to define coordinates. Then, each sample has a coordinate (x, y). This way, we have a distribution of the samples in 2-space that relate xd and y. □

Statistical Inference

286A Mathematically, this is to find a probabilistic model describing the observed data. Let X
be the observed values. We wish to make $f(x; \theta, \eta)$, where θ is the unknown variables on
which we wish to estimate and η are parameters describing experiments, etc. Assuming the
286B observed data as iid sample, estimate f.

The Bayesian approach assumes that there is a certain prior density $\pi(\theta, \eta)$ for parameters and use the Bayes theorem

$$P(\theta, \eta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta, \eta)\pi(\theta, \eta)}{\int d\theta d\eta\, f(\boldsymbol{x}|\theta, \eta)\pi(\theta, \eta)} \tag{16}$$

P as the subjective probability can be axiomatized based on the consistency of individual behaviors.[1]

286C The problem of statistical inference may be formulated as a problem of making a rule to choose an element from the set of possible conclusions. The rule is evaluated by the probability derived from the observed sample set. From the objective probability point of view, this probability is nothing but the relative frequency, so for each set of observed samples, the probability does not have any direct sense. Thus, Neyman regarded statistical inference not as inductive inference but as inductive action; in contrast, Fisher maintained that statistical inference is an inductive inference and its purpose is to derive the most appropriate conclusions.

Fisher asserted that all the information must be used. From this he derived: ∗ Principle of sufficiency: if there is a sufficient statistics (→ 293E), all the inferences must depend on it.

[1] L. J. Savage, *The foundation of statistics* (Wiley, 1954; Dover, 1972). cf. http://plato.stanford.edu/entries/epistemology-bayesian/ Difficulties in the Theory of Personal Probability Leonard J. Savage Philosophy of Science, Vol. 34, No. 4 (Dec., 1967) , pp. 305-310; Implications of Personal Probability for Induction Leonard J. Savage Journal of Philosophy, Vol. 64, No. 19, Sixty-Fourth Annual Meeting of the American Philosophical Association, Eastern Division (Oct. 5, 1967) , pp. 593-607 ; Reading Suggestions for the Foundations of Statistics Leonard J. Savage American Statistician, Vol. 24, No. 4 (Oct., 1970) , pp. 23-27

$*$ Principle of conditionality: if there is an ancillary statistic that do not depend on the parameters of the distribution, inference must be performed with the conditional distribution conditioned by the ancillary statistics.

286E The idea of population and the test of goodness of fit were due to K. Pearson, who generalized Galton's regression and correlation ideas. Now, the population is understood as a probability space with a measure containing unknown parameters.

293A Statistic

Statistic is a function of observables (sample values). It is introduced to characterize the sample distribution and to distill the information wrt the population parameters (true distributions).

293B Let $(\Omega, \mathcal{B}, P)$ be a probability space, which is understood as a population. A random sample of size n from the population is an iid stochastic variables $X_1, \cdots, X_n$. The probability space induced by this sample is called the n-sample space. Sample values are described in terms of lower case letters. Statistic Y is a measurable function of X. P (or Φ_n, the distribution of $\{X_k\}$) may be know to belong to a parameter family P_θ ($\theta \in \Theta$; Θ is called the parameter space. The most general formulation is in terms of a parameter set of probability measures on some measurable space.

In terms of the population probability measure, population descriptive statisitcs may be introduced as usual (called *population characteristics*; mean , variance, etc.)

293E Let $(\mathcal{X}, \mathcal{U})$ be a measurable space and $\mathcal{P}$ be a family of probability measures on it. Then, $(\mathcal{X}, \mathcal{U}, \mathcal{P})$ is called a *statistical structure*. A subfamily $\mathcal{B}$ is sufficient for $\mathcal{P}$, if for any $A \in \mathcal{U}$ we can define conditional probability $P_\theta(\cdot|B)$ for any θ.
Let t be a statistic. If $\mathcal{B}(t)$ (a family defined by the subset of the population characterized
by t) is sufficient, t is called a *sufficient statistic*. In short, if the set of samples compatible
293H with $t \cup A$ for any $A \in \mathcal{U}$ is measurable, then we say t is sufficient. t is ancillary, if $P_\theta(t)$ is
independent of θ.

338A Sample Distribution

If $\{X_i\}$ is iid n samples from $N(0,1)$,

$$Y = \sum_{i=1}^{n} X_i^2 \tag{17}$$

obeys the *chi-square distribution* of degree of freedom n $\chi^2(n)$:

$$f_n(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2} e^{-y/2}. \tag{18}$$

If $X \in N(0,1)$ and $Y \in \chi^2 n$,

$$T = X/\sqrt{Y/n} \tag{19}$$

obeys $t(n)$ (the *t-distribution* of degree of freedom n).

Let $X_i \in N(\mu, \sigma^2)$. The sample variance

$$S^2 = \sum_i (X_i - \langle X \rangle)^2/(n-1) \tag{20}$$

Then, $(n-1)S^2 \in \chi^2(n-1)$ and the sample t-statistics

$$T = \sqrt{n}(\langle X \rangle - \mu)/\sqrt{S^2} \tag{21}$$

obeys $t(n-1)$ (according to Student). $\langle X \rangle$ and S^2 are statistically independent (according to Fisher).[2]

If $X \in \chi^2(m)$ and $Y \in \chi^2(n)$, the distribution of $Z = (X/m)/(Y/n)$ obeys the F-distribution of degree of freedom $(m.n)$.

338C Generalization to multidimensional cases is straightforward.

338D If the number of samples is large, we may use CLT.

338E

$$F_n(x) = frac1n \sum_{i=1}^{n} \Theta(x - x_i) \tag{22}$$

is called the *empirical distribution functions.* If $X \in F$, then almost surely (Glivenko-Cantelli's theorem)

$$\|F(x) - F_n(x)\|_\infty \to 0 \tag{23}$$

$\sqrt{n}(F_n(t) - F(t))$ converges in law to the Brownian bridge. This fact can be used to test a hypothesis on F.

Let $\{X_i\} \in F$ and $\{Y_i\} \in G$ are random samples. Then, the distribution of

$$D_{mn} = sup_x |F_m(x) - G_n(x)| \tag{24}$$

does not depend on G nor F. This is the basis of the *Kolmogorov-Smirnov test.*

338F Let $X_i \in F$ which is with mean μ and variance σ^2. If F is absolutely continuous and has ν-th moment, then $\sum(X_i - \mu)/\sqrt{n}$ obeys F_n such that

$$F_n(x) = \Phi(x) + \sum_{k=1}^{\nu-2} R_k(x) \left(\frac{1}{\sqrt{n}}\right)^k \varphi(x) + B(1/\sqrt{n})^{\nu-1}. \tag{25}$$

This is called the *Edgeworth expansion.* Here, φ is the density distribution of Φ, which is $N(0,1)$, and B is a term bounded by a certain constant determined by F and ν. R_k is a polynomial dependent on the cumulants γ_n of F. In particular,

$$\begin{aligned} R_1 &= -\gamma_3(x^2-1)/6, & (26)\\ R_2 &= -\gamma_4(x^3-3x)/24 - \gamma_3^2(x^5 - 10x^3 + 15x)/72. & (27)\end{aligned}$$

There is a version for F on $\boldsymbol{Z}$. Using the Edgeworth expansion, it is possible to determine y such that $F_n(y) = \Phi(x)$ where

$$x = y + \sum_{k=1}^{\nu-2} A_k(y) n^{-k/2} \tag{28}$$

[2] If $\mu \neq \mu_0$ (the sample mean), $T \in t(n-1, \sqrt{n}(\mu - \mu_0)/\sigma)$: non-central t.

(*Cornish-Fisher expansion*). For $\nu = 3$ there is a practical recipe (*Wilson-Hilferty's approximation*)[3]:

$$\sqrt{9n/2}\left[(X/n)^{1/3} - 1 + 2/9n\right] \tag{29}$$

obeys $N(0,1)$. This is even reliable for small n.

338G Ordinal statistics $X_{(1)} \le \cdots \le X_{(n)}$ can be constructed from n samples. The simultaneous density distribution function for p of them $Y_1 = X_{(\alpha)}, \cdots, Y_p = X_{(\eta)}$ can be obtained explicitly in terms of F.

Gnedenko proved the following:[4]
Let $\{a_n\}$ be a real sequence and $\{b_n\}$ a positive sequence. If $X_{(n)} - a_n/b_n$ converges to a nondegenerate distribution G, then there are only three cases modulo positions and scales:

$$\begin{aligned} G_1(x) &= \exp(-x^{-\gamma})\Theta(x), & (30)\\ G_2(x) &= \exp(-(-x)^{-g})(1-\Theta(x)), & (31)\\ G_3(x) &= \exp(-e^{-x}). & (32) \end{aligned}$$

The domains of attraction for G_i are completely characterized. The distributions of $X_{(n)} - X_{(1)}$ (the sample range) and $(X_{(n)} + X_{(1)})/2$ are also characterized.

338H If F has a bounded density and characteristic function is definable, then $X_1, \cdots, X_9$ can uniquely determine F.[5]

Statistical Model

290A According to Fisher the process of statistical inference consists of the following three steps:
(1) Specifying the model,
(2) Estimating the unknown parameters,
(3) Test of fitness.
Thus, the process is realized as the introduction and selection of the models.

290B We must be able to compare distributions. The KL entropy

$$B(f;g) = \int dx f(x) \log \frac{g(x)}{f(x)} \tag{33}$$

This is essentially the average of the logarithmic likelihood. $\log g$ gives the unbiased estimator of the logarithmic likelihood of the true distribution.
Comment: This is justified better by Sanov's theorem.

290F Let $f(x|\theta)$ be a parameter family of distributions. $\log f(x|\theta)$. θ that maximizes this is called the *maximum likelihood estimate* of θ and is denoted by $\theta(x)$.
1)*Log likelihood ratio analysis* The ratio is defined as $f(x|\theta(x))/f(x|\theta'(x))$. χ^2-test is used,

[3] PNAS **17**, (1931).

[4] Ann. Math. **44** (1943).

[5] Yu. V. Prokhorov, TPA **10** 438 (1965) (Russian original).

but the choice of the significant level is not unique.
2) Minimizing AIC: AIC $= -2\log f(x|\theta(x))+d(\theta)$, where d is the dimension of the parameter space. If we regard $\exp(-\mathrm{AIC}/2)$ as the likelihood of $f(\cdot|\theta)$, we wish to maximize it.

290G We can use Bayesian models assuming *a priori* distribution $\pi(\theta)$ for the parameters.

$$\int d\theta\, f(x|\theta)\pi(\theta) \tag{34}$$

is the Bayesian likelihood. The point of view common to AIC and Bayes is that the goodness of the statistical model is determined by the balance between the information given by data and the complexity of model. This point of view was obtained first by introducing the concept of information into the statistical modeling.[6]

• Information requires basic uniformity. That is, the reason for $-\log p$ is a good measure of information presupposes that the uniform state is the least informative. This depends on how you define the *a priori* distribution. Therefore, we need a different more objective approach driven by data. We should notice that statistical mechanics has the same defect. It must be formulated in a much more data driven fashion; that is, thermodynamically. □

Statistical Estimation

287A This is the problem of inferring the parameters or the value of a function $g(\theta)$ at the true parameter value from the samples.

287B The the *point estimation* is to determine $g(\theta) = \phi(x)$ when the sampled value of X is x. $\phi(X)$ is called the *estimator* of $g(\theta)$.

287C The estimator should exhibit some sort of unbiased nature wrt g. If $E_\theta(\phi(X)) = g(\theta)$ for all θ, φ is called the *(mean) unbiased estimator* for g. If g has such φ, g is said to be *estimable.*

$$b(\theta) = E_\theta(\phi(X)) - g(\theta) \tag{35}$$

is called the *bias*. Estimators with smaller variances are desirable.
Theorem [Rao-Blackwell] Let $T = t(X)$ be a sufficient statistic (→), and $\varphi(X)$ be an arbitrary unbiased estimator for g. Let $\psi(t) = E(\varphi(X)|T = t)$. Then $\varphi^*(X) = \psi(t(X))$ is also an unbiased estimator for g, and $V_\theta(\varphi^*) \leq V_\theta(\varphi)$. The equality holds if $\varphi = varphi*$ (a.e. wrt $\mathcal{P}$).□

287D The lower bound of the variation of unbiased estimator: the variation of the estimator at θ_0 with the smallest variation (locally best unbiased estimator) exists:
Theorem [Barankin] If $E([p_\theta(x)/p_{\theta_0}(x)]^2) < +\infty$, then in $\mathcal{M}$ (= the set of unbiased estimators of $g(\theta)$ for which the variation is finite at θ_0) exist the local best unbiased estimator.

From this several theorems are derived. For example, *Cramér-Rao's inequality*; if X is the size n random sampling from the population with density f:

$$V_{\theta_0}(\varphi(X)) \geq [g'(\theta_0)]^2/nI(\theta_0), \tag{36}$$

where I is the Fisher information:

$$I(\theta) = E_\theta\left[\left(\frac{\partial \log f(x,\theta)}{\partial\theta}\right)^2\right] \tag{37}$$

[6] Akaike et al., Surikagaku **218** (1981) 7-66.

287K If the number of samples is large, then the distribution may be approximated by the asymptotic distribution. Let φ_n be the estimator on $X = \{X_1, \cdots, X_n\}$. If $\varphi_n \to g$ (in probabilty) φ_n is called the *consistent estimator*. If $n^{1/2}(\varphi_n - g)$ converges in law to a normal distribution, φ_n is called *asymptotically normally distributed*. If it is also consistent, it is called a *CAN estimator*.

If Can estimator has the asymptotic variation $I(\theta)^{-1}$ is it called a *BAN estimator* (best...) or *asymptoptically effective estimator*.

287M The *maximum likelihood estimator* is $\theta(X)$ that maximizes $P_theta(X)$.
Theorem [Wald] (Under some technical conditions) if $P_{\theta_n}(\{X_n\})/P_{\theta_0}(\{X_n\}) \geq c > 0$ then $\theta_n \to \theta_0$ (a.e.).

A condition for the ML estimator to be BAN is also given by Cramér.

Statistical Decision Function

285A This theory was initiated by Wald to unify statistics as mathematics. A *sample space* $(\mathcal{X}, \mathcal{B})$ is a measurable space, on which a family of probability measures $\mathcal{P}$ is defined. $x \in \mathcal{X}$ is called the *sample point*, and the true measure $P \in \mathcal{P}$ is assumed. If $\mathcal{P}$ is parameterized as P_θ ($\theta \in \Omega$, which is called the *parameter space*), the θ satisfying $P = P_\theta$ is called the *true value of a parameter*. When we observe x, we do something to determine θ. Let $\mathcal{A}$ be the set of actions we take to this end. We make it a measurable space $(\mathcal{A}, \mathcal{U})$, and statistical decision function δ ($\in D$; D is called the *space of decitsion function*) is a map from $\mathcal{X}$ to a probability measure on $(\mathcal{A}, \mathcal{U})$. $\delta(C|x)$ is understood as the probability of taking an action in C when x is observed.

When $P = P_\theta$, the loss due to the action $a \in \mathcal{A}$ is expressed by a *loss function* $w(\theta, a)$, which is positive and measurable wrt to a. The *risk function* is defined as

$$r(\theta\delta) = \int_{\mathcal{X}} P_\theta(dx) \int_{\mathcal{A}} \delta(da|x) w(\theta, a). \tag{38}$$

The statisitical decision problem is $(\mathcal{X}, \mathcal{B}, \mathcal{P}, \Omega, \mathcal{A}, \mathcal{U}, w, D)$.

Inference problems can be formulated as decision problems, but nothing seems fundamentally new.

285H Statistical decision problems may be understood as games between Nature and the statistician. The strategy of Nature is the true P or true θ, and tha of the statistician is d. The price the statistician pays is $r(\theta, \delta)$. *A priori* distribution is the mixed strategy of Nature. The minimax solution is the minimax decition of δ.

Statistical Hypothesis Testing

A *test function* $\phi(x)$ is a probability assigned to a sample point x such that the null hypothesis is rejected with this probability. Let P_θ be the true probability measure. Even if the hypothesis θ is correct, the probability of its rejection is

$$E_\theta(\phi) = \int P_\theta(dx)\phi(x). \tag{39}$$

If the rejection is decided with a predetermined level α it is called the level α test.
Error of the first kind: rejecting correct hypothesis

Error of the second kind: accepting the wrong hypothesis
$E_\theta(\phi)$ is the probability

$\boldsymbol{M}$ultivariate Aanalysis

258B