Contents

Beginning of Aug.23, 2021

0.1 Review of Probability

Definition 0.1.1: Axioms of Probability

Assume there exists some (nonempty) *universal set* Ω that contains all other sets (events). We denote $\mathbb{P}$ as a probability law on Ω. The following are the three axioms of $\mathbb{P}$:

(1) For all subsets $A \subset \Omega$, $0 \leqslant \mathbb{P}(A) \leqslant 1$.

(2) $\mathbb{P}$ is (countably) additive for *disjoint* sets, i.e., if $\{A_n\}$ are disjoint then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

(3) $\mathbb{P}(\Omega) = 1$.

From (2) we immediately see that $\mathbb{P}(A \cup B) \leqslant \mathbb{P}(A) + \mathbb{P}(B)$ and $=$ can be obtained if and only if $A \cap B = \varnothing$.

Definition 0.1.2: Conditional Probability

If $A, B \subset \Omega$, we define the **conditional probability** as

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Definition 0.1.3: Continuous Random Variable, CRV

A **random variable** is a function $X : \Omega \to \mathbb{R}$. We say X is **continuous** if

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(x)\,\mathrm{d}x$$

for some $f_X : \mathbb{R} \to [0, \infty)$ and for all $-\infty \leqslant a \leqslant b \leqslant \infty$. We call f_X the **PDF, probability density function**. We also define the **CDF, cumulative density function**, by $F_X(t) = \mathbb{P}(X \leqslant t)$.

Example 0.1.4. Suppose X is uniformly distributed in $[0, 1]$. Then for any $0 \leqslant a \leqslant b \leqslant 1$,

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b \mathrm{d}x = b - a,$$

indeed a CRV.

Definition 0.1.5: Independence of Finitely Many Sets

Let $A_1, \ldots, A_n \subset \Omega$. We say they are **independent** if for any $S \subset \{1, 2, \ldots, n\}$,

$$\mathbb{P}(\bigcap_{i \in S}(A_i)) = \prod_{i \in S} \mathbb{P}(A_i).$$

Definition 0.1.6: Independence of Countably Many Sets

We say $\{A_n\}$ are independence if, for all $n \geqslant 1$, $A_1, \ldots, A_n$ are independent.

Beginning of Aug.25, 2021

0.2 Some Random Variables

Example 0.2.1: Bernoulli. Let $0 < p < 1$. Define a random variable by $\mathbb{P}(X = 0) = 1 - p$ and $\mathbb{P}(X = 1) = p$.

Example 0.2.2: Binomial. Let $n \in \mathbb{N}$. For $0 \leqslant k \leqslant n$, let $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$. *"Number of heads flipped among n biased coin flips."*

Example 0.2.3: Geometric. For $k \in \mathbb{N}$, define $\mathbb{P}(X = k) = (1-p)^{k-1} p$. *"Number of coin clips needed to see heads for the first time."*

Definition 0.2.4: Normal Random Variable

Let $\mu \in \mathbb{R}$ and $\sigma > 0$ be two parameters. A random variable X is said to be **normal** or **Gaussian** if X has the X has pdf

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Definition 0.2.5: Gamma Function

For all $\alpha > 0$, define the **Gamma function**

$$\Gamma(\alpha) := \int_0^\infty e^{-x} x^{\alpha-1} \, \mathrm{d}x.$$

Integration by parts suggests that Γ interpolates the factorial: $\Gamma(1) = 1$ and $\Gamma(n+1) = (n+1)\Gamma(n)$.

Definition 0.2.6: Gamma Distribution & Chi-Squared Distribution

Let $\alpha, \beta > 0$. We say X is an (α, β) distributed **Gamma random variable** if X has the pdf

$$f_X(x) = \frac{x^{\alpha-1} \exp(-x/\beta)}{\beta^\alpha \Gamma(\alpha)} \cdot \chi_{[0,\infty)}(x). \tag{1}$$

For example, if $\alpha = p/2$ and $\beta = 2$, we get a **chi-squared** distribution. Its pdf with p **degrees of freedom** is

$$f_X(x) = \frac{x^{p/2-1} \exp(-x/2)}{2^{p/2}\Gamma(p/2)} \cdot \chi_{[0,\infty)}(x). \tag{2}$$

(2) is the distribution of a sum of p independent, squared, standard Gaussian distributions ($\mu = 0, \sigma = 1$).

Definition: (1.36) Indicator Functions

Let $A \subset \Omega$. We define the **indicator function** $\chi_A : \Omega \to \{0,1\}$ by

$$\chi_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

Definition: (1.37) Expected Values

Let $\mathbb{P}$ be a probability law on Ω. Let $X : \Omega \to [0,\infty)$ be a (nonnegative) random variable. The **expected value** is defined by

$$\mathbb{E}(X) := \int_0^\infty \mathbb{P}(X > t)\, dt.$$

If $X : \Omega \to \mathbb{R}$ and if $\mathbb{E}|X| < \infty$, define

$$\mathbb{E}(X) := \mathbb{E}(\max(X,0)) - \mathbb{E}(\max(-X,0))$$

If X is discrete, $\mathbb{E}(X) = \sum_{k\in\mathbb{R}} k \cdot \mathbb{P}(X = k)$.
If X is continuous with PDF f_X, $\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$.

Proposition: (1.43)

The expected value of (finite) sums is the (finite) sum of expected values:

$$\mathbb{E}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbb{E}(X_i).$$

Definition 0.2.7: Variance & Standard Deviation

The **variance** of a random variable is defined by

$$\mathrm{var}(X) := \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \mathbb{E}(X - \mathbb{E}(X))^2.$$

and the **standard deviation** is defined to be the square root of above.
Important property of variance:
$var(aX + b) = a^2$
$var(x)$

Definition 0.2.8: Joint PDF

Let X, Y be random variables and let $f_{X,Y} : \mathbb{R}^2 \to [0,\infty)$ be their **joint pdf**. Then

$$\mathbb{P}(a \leqslant X \leqslant b, c \leqslant Y \leqslant d) = \int_a^b \int_c^d f_{X,Y}(x,y)\, dy\, dx$$

and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x = 1.$$

Definition 0.2.9: Marginal Densities

Continuing on the previous example, the **marginal** of X is given by

$$f_X(x) := \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y,$$

i.e., we "fix" x and integrate $f_{X,Y}$ over all possible values of y. Likewise for $f_Y(y)$.
The density of the conditional $X \mid Y = y$ is

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Corollary 0.2.10

$\mathbb{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{x,y}(X,Y)\,\mathrm{d}x\,\mathrm{d}y$. Similarly, if $g:\mathbb{R}^2\to\mathbb{R}$, then

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Definition 0.2.11: Independence of Random Variables

Let $X_1,\dots,X_n:\Omega\to\mathbb{R}$. We say they are **independent** if

$$\mathbb{P}_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n}\mathbb{P}_{X_i}(x_i) \qquad x_i\in\mathbb{R}.$$

Beginning of Aug.27, 2021

Theorem: (1.58) Independence and Variances

If $X_1,\dots,X_n$ are independent random variables, then the variance of their sum is the sum of their variances:

$$\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n}\operatorname{var}(X_i).$$

Proposition: (1.60) Independence and Expected Values

If $X_1,\dots,X_n$ are independent random variables, then

$$\mathbb{E}(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n}\mathbb{E}(X_i).$$

In general, this is not true. There exist X,Y such that $\mathbb{E}(XY)\neq\mathbb{E}(X)\mathbb{E}(Y)$. For example let $X=Y$ and $\mathbb{P}(X=1)=\mathbb{P}(X=-1)=1/2$. Then $\mathbb{E}(XY)=1$ and $\mathbb{E}(X)\mathbb{E}(Y)=0$.

Definition 0.2.12: Covariance

The **covariance** of X and Y is given by

$$\operatorname{cov}(X,Y) := \mathbb{E}[(X-\mathbb{E}(X)(Y-\mathbb{E}(Y)].$$

By definition, $\operatorname{cov}(X,X) = \operatorname{var}(X)$.

Definition 0.2.13: Correlation

The **correlation** coefficient of X, Y is given by

$$\operatorname{corr}(X,Y) := \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}.$$

Notice that covariance is affected by scaling but correlation is not (unless there involves a change of sign). For example, $\operatorname{cov}(10X,Y) = 10\operatorname{cov}(X,Y)$ *whereas* $\operatorname{corr}(10X,Y) = \operatorname{corr}(X,Y)$. (We also assume that $\operatorname{var}(X), \operatorname{var}(Y) \neq 0$.)

Correlation is invariant under scalar multiplication. Let $t \neq 0$. Then

$$\operatorname{corr}(tX,Y) = \frac{\operatorname{cov}(tX,Y)}{\sqrt{\operatorname{var}(tX)}\sqrt{\operatorname{var}(Y)}} = \frac{\mathbb{E}(t(X-\mathbb{E}(X))(Y-\mathbb{E}(Y)))}{|t|\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}} = \frac{t\operatorname{cov}(X,Y)}{|t|\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}} = \operatorname{sgn}(t)\operatorname{corr}(X,Y).$$

□

Theorem 0.2.14: Cauchy-Schwarz

For all X, Y, $|\mathbb{E}(XY)| \leqslant \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}$. This immediately implies that for any X, Y, $-1 \leqslant \operatorname{corr}(X,Y) \leqslant 1$. Heuristically: *if* $\operatorname{corr}(X,Y)$ *is close to* 1*, then approximately* $Y = aX + b$ *for* $a > 0$*. If the scatterplot of* X, Y *are everywhere randomly, then their correlation is close to* 0.

Theorem 0.2.15: Total Expectation Theorem

Let $A_1, \ldots, A_n \subset \Omega$ be a partition of Ω (i.e., union being Ω and pairwise disjoint). Then, for all $B \subset \Omega$,

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \cap A_i) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)$$

(assuming $\mathbb{P}(A_i) \neq 0$ for all i for the sake of well-definedness).

Theorem: (1.78)

Let X, Y be continuous random variables with joint PDF. Then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y)\, \mathrm{d}y,$$

where

$$\mathbb{E}(X \mid Y = y) = \int_{-\infty}^{\infty} x f_{X|Y=y}(x \mid y)\, \mathrm{d}x \qquad f_{X|Y=y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Definition: (1.90) Convex Functions

$\varphi : \mathbb{R} \to \mathbb{R}$ is **convex** if for any $x, y \in \mathbb{R}$ and $0 \leqslant t \leqslant 1$,

$$\varphi(tx + (1-t)y) \leqslant t\varphi(x) + (1-t)\varphi(y).$$

Theorem: (1.91) Jensen's Inequality

Let X be a random variable and let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex. Then

$$\varphi(\mathbb{E}X) \leqslant \mathbb{E}\varphi(X).$$

Corollary 0.2.16

Let $\varphi(t) = t^2$. By Jensen's inequality,

$$(\mathbb{E}X)^2 \leqslant \mathbb{E}X^2.$$

Beginning of Aug.30, 2021

Proposition: (1.92) Markov's Inequality

If $\mathbb{E}|X| < \infty$ (if infinity then the claim is trivial), then

$$\mathbb{P}(|X| > t) \leqslant \frac{\mathbb{E}|X|}{t}.$$

Also, for $n \in \mathbb{Z}_+$,

$$\mathbb{P}(|X| > t) \leqslant \frac{\mathbb{E}(|X|^n)}{t^n}.$$

Corollary: (1.97) Chebyshev's Inequality

If $\mathbb{E}X < \infty$ and $\operatorname{var}(X) < \infty$, then applying Markov's inequality to $X - \mathbb{E}X$, $n = 2$, gives

$$\mathbb{P}(|X - \mu| > t) \leqslant \frac{\operatorname{var}(X)}{t^2}.$$

Proposition 0.2.17

Let $X_1, \ldots, X_n$ be i.i.d. with finite variance. Then

$$\mathbb{P}\Big(\Big|\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\text{sample mean}} - \mathbb{E}X_i\Big| > t\Big) \leqslant \frac{\operatorname{var}(\sum_{i=1}^{n} X_i/n)}{t^2} = \frac{1}{n} \cdot \frac{\operatorname{var}(X_1)}{t^2}.$$

Note that as $n \to \infty$, the probability tends to 0. This gives the Weak LLN.

Theorem 0.2.18: Weak Law of Large Numbers (WLLN)

Let $X_1, ..., X_n$ be i.i.d. with finite variance. Then

$$\lim_{n\to\infty} \mathbb{P}(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X_i\right| > t) \leqslant \lim_{n\to\infty} \frac{1}{n}\frac{\operatorname{var}(X_1)}{t^2} = 0.$$

To put formally, $\sum_{i=1}^{n} X_i/n$ **converges in probability** to $\mathbb{E}X_1$ as $n \to \infty$.

Remark. For example, if $X_1, ..., X_n$ are the poll results for n random people in California, the larger n is, the more likely it is accurate. Furthermore, note that this does *not* rely on the population size of California!

Definition: (1.106) Convolution

Let $f, g : \mathbb{R} \to \mathbb{R}$. We define the **convolution** of f and g to be

$$(f * g)(t) := \int_{-\infty}^{\infty} f(x)g(t-x)\,\mathrm{d}x.$$

Proposition: (1.107)

Let X and Y be *independent* random variables with PDFs f_X and f_Y. Then,

$$f_{x+y} = f_X * f_Y.$$

Corollary 0.2.19

The sum of two independent Gaussians is a Gaussian.

Definition 0.2.20: Modes of Convergence

(1) (Def 2.1, used in SLLN) We say random variables $Y_n : \Omega \to \mathbb{R}$ **converges almost surely** (a.s.) to $Y : \Omega \to \mathbb{R}$ if

$$\mathbb{P}(\lim_{n\to\infty} Y_n = Y) = 1$$

or in more details,

$$\mathbb{P}(\{\omega \in \Omega \mid \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1.$$

(Given $\epsilon > 0$, there exists $N \in N$ such that if $n \geqslant N$ then $Y_n \to Y$ on $S \subset \Omega$ where $\mathbb{P}(\Omega - S) = 0$.)

(2) (Def 2.2, used in WLLN) We say $\{Y_n\}$ **converges in probability** to Y if, given $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| > \epsilon) = 0.$$

(3) (Def 2.3, used in CLT) We say $\{Y_n\}$ **converges in distribution** to Y if "the limit of the CDF is the CDF of the limit" – to put formally, for all $t \in \mathbb{R}$ such that $S \mapsto \mathbb{P}(Y \leqslant S)$ is continuous at $s = t$

$$\lim_{n\to\infty} \mathbb{P}(Y_n \leqslant t) = \mathbb{P}(Y \leqslant t).$$

Theorem 0.2.21: WLLN Restated

Let $\{X_n\}$ be i.i.d. with finite variance. Then

$$\frac{X_1 + ... + X_n}{n}$$

converges in probability to $\mathbb{E}X_1$ as $n \to \infty$.

Theorem 0.2.22: Central Limit Theorem (CLT)

The previous theorem states how $X_1 + ... + X_n$ looks like, and this one states how far it would deviate from μ. Let $\{X_n\}$ be i.i.d. with finite *nonzero* variance. For convenience write $\mu := \mathbb{E}X_1$ and $\sigma := \sqrt{\mathrm{var}(X_1)}$. Then

$$\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}}$$

(note that the numerator has mean 0) converges in distribution to a standard Gaussian.

Beginning of Sept.1, 2021

Theorem: (2.11) Strong Law of Large Numbers, SLLN

Let $X_1, X_2, ...$ be i.i.d. with $\mu := \mathbb{E}X_1$ finite. Then

$$\mathbb{P}(\lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} = \mu) = 1.$$

Theorem: (2.17) CLT Restated

Let $X_1, X_2, ...$ be i.i.d. with $0 < \mathrm{var}(X_1) < \infty$. Let $\mu := \mathbb{E}X_1$ and $\sigma := \sqrt{\mathrm{var}(X_1)}$. Then, for all $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}(\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \leqslant t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(-s^2/2)\, \mathrm{d}s = \mathbb{P}(Z \leqslant t).$$

Remark. For example, if we let $n = 1$ million, $t = 1$, and X_i the result representing a coin flip (1 for heads, -1 for tails), then $\mu = \mathbb{E}X_1 = 0$ and $\mathrm{var}(X_1) = \mathbb{E}(X_1 - \mathbb{E}X_1)^2 = \mathbb{E}X_1^2 = 1$, so $\sigma = 1$. By the CLT,

$$\mathbb{P}(\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \leqslant 1) = \mathbb{P}(\frac{X_1 + ... + X_n}{\sqrt{n}} \leqslant 1) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1} \exp(-s^2/2)\, \mathrm{d}s \approx 0.84.$$

Note that $(X_1 + ... + X_n)/\sqrt{n} \leqslant 1$ means $X_1 + ... + X_n \leqslant \sqrt{n} = 1000$. Hence if we flip 1 million coins, the probability that (#heads $-$ #tails $\leqslant 1000$) is approximately 0.84 or, equivalently, the probability that we get $\leqslant 500500$ heads is approximately 0.84.

Remark. Let $X_1, X_2, ...$ be defined as above. Then $(X_1 + ... + X_n - n\mu)/(\sigma\sqrt{n})$ has mean 0 (obvious) and

variance 1:

$$\begin{aligned}\operatorname{var}(\ldots) &= \frac{1}{\sigma^2 n}\operatorname{var}(X_1 + \ldots + X_n - n\mu)\\ &= \frac{1}{\sigma^2 n}\operatorname{var}(X_1 + \ldots + X_n) = \frac{n\operatorname{var}(X_1)}{\sigma^2 n} = 1.\end{aligned}$$

The random variables $Z_n := (X_1 + \ldots + X_n - n\mu)/(\sigma\sqrt{n})$ appear in the CLT. We have shown that Z_n's have the same mean and variance as the standard Gaussian. That $\operatorname{var}(Z_n) = 1$ explains why we chose to put $\sqrt{n}$ in the denominator, not n, which might have seemed to be more natural on first glance.

Theorem: (2.30) Berry-Esseén Theorem

Let $\sigma > 0$ and $X_1, X_2, \ldots$ be i.i.d. with mean zero so that $\mathbb{E}X_1^2 = \sigma^2$. Furthermore, assume that $\mathbb{E}|X_1|^3 < \infty$. Let Z be a standard Gaussian random variable. Then, for all $n > 1$,

$$\begin{aligned}\left|\mathbb{P}\left(\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leqslant t\right) - \mathbb{P}(Z \leqslant t)\right| &= \left|\mathbb{P}\left(\frac{X_1 + \ldots + X_n}{\sigma\sqrt{n}} \leqslant t\right) - \mathbb{P}(Z \leqslant t)\right|\\ &\leqslant \frac{\mathbb{E}|X_1|^3}{\sigma^3\sqrt{n}}.\end{aligned}$$

This provides an improvement of the CLT.

End of Review for 407

Chapter 1

Random Samples

Definition: (3.1) Random Sample

A **random sample** of size n is a sequence $X_1, ..., X_n$ of i.i.d. random variables.

Definition: (3.2) Statistic

A **statistic** is a function of a random variable:
Let $X_1, ..., X_n$ be a random sample of size n and let $t : \mathbb{R}^n \to \mathbb{R}^k$. A statistic is a random variable of the form

$$Y := t(X_1, ..., X_n)$$

(where the output of Y contains k numbers). The distribution of Y is called the **sampling distribution**.

Example: (3.3, 3.4). The **sample mean** of $X_1, ..., X_n$, denoted $\overline{X}$, is the following statistic:

$$\overline{X} := \frac{X_1 + ... + X_n}{n}.$$

For $n \geqslant 2$, **sample standard deviation**, denoted S, is the following statistic:

$$S := \left(\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \right)^{1/2}.$$

The **sample variance** is simply S^2.

Example: (3.5) Why $n-1$ in Sample Sample Deviation?. Let $X_1, ..., X_n$ be i.i.d. with $\mu := \mathbb{E}X_1 \in \mathbb{R}$ and $\sigma := \sqrt{\mathrm{var}(X_1)} < \infty$. Then

(1) $\mathbb{E}S^2 = \sigma^2$. (If we divide by anything other than $n-1$, there will be extra constants involved.)

(2) $\mathrm{var}(\overline{X}) = \sigma^2/n$.

Beginning of Sept.3, 2021

Proposition: (3.7)

Let $n \geqslant 2$ and let $X_1 ..., X_n$ be a random sample from a Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance σ^2. Then

(1) $\overline{X}$ and S are independent;

(2) $\overline{X}$ is also Gaussian with mean μ and variance σ^2/n;

(3) $(n-1)S^2/\sigma^2$ has the same distribution as χ^2_{n-1} (chi-squared with degrees of freedom $n-1$).

Proof for n = 2. By definition $\overline{X} = (X_1 + X_2)/2$ and $S = \sqrt{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2}$, or, after rewriting $\overline{X}$,

$$S = \sqrt{(X_2 - X_1)^2/4 + (X_2 - X_1)^2/4} = \sqrt{(X_1 - X_2)^2/2}.$$

Note that it suffices to show that $X_1 + X_2$, $X_1 - X_2$ are independent (given both are i.i.d. Gaussians). On one hand,

$$\mathbb{E}(X_1 + X_2)(X_1 - X_2) = \mathbb{E}X_1^2 - \mathbb{E}X_2^2 = 0.$$

On the other hand, X_1, X_2 have the same mean, which implies $\mathbb{E}(X_1 - X_2) = 0$ so $\mathbb{E}(X_1 + X_2)\mathbb{E}(X_1 - X_2) = 0$. In general, $\mathbb{E}X\mathbb{E}Y = \mathbb{E}(XY)$ does not imply independence of X and Y (the converse does), this implication is in fact true given that X and Y are Gaussians!

Gaussian Independence Proof. WLOG assume that X_1, X_2 are standard Gaussians. By assumption X_1, X_2 are independent and we want to show that $X_1 + X_2, X_1 - X_2$ are independent. By definition of independence,

$$\mathbb{P}((X_1, X_2) \in A) = \frac{1}{2\pi} \iint_A \exp(-(x_1^2 + x_2^2)/2)\, \mathrm{d}x_1 \mathrm{d}x_2. \tag{1}$$

In order to show $X_1 + X_2, X_1 - X_2$ are independent, we want to show that their JPDF is the product of their PDFs, i.e., we need to show that, for $B, C \in \mathbb{R}$,

$$\mathbb{P}(X_1 + X_2 \in B, X_1 - X_2 \in C) = \mathbb{P}(X_1 + X_2 \in B)\mathbb{P}(X_1 - X_2 \in C).$$

Manipulating the LHS,

$$\begin{aligned}
\mathbb{P}(X_1 + X_2 \in B, X_1 - X_2 \in C) &= \mathbb{P}(\langle (X_1, X_2), (1,1)\rangle \in B, \langle (X_1, X_2), (1,-1)\rangle \in C) \\
&= \iint_{\substack{\text{Rotation} \\ \text{of } B\times C}} \frac{1}{2\pi} \exp(-(x_1^2 + x_2^2)^2/2)\, \mathrm{d}x_1 \mathrm{d}x_2 \\
&= \iint_{B\times C} \frac{1}{2\pi} \exp(-(x_1^2 + x_2^2)/2)\, \mathrm{d}x_1 \mathrm{d}x_2 \\
[(1)] &= \int_B \frac{1}{\sqrt{2\pi}} \exp(-x_1^2/2)\, \mathrm{d}x_1 \int_C \frac{1}{\sqrt{2\pi}} \exp(-x_2^2/2)\, \mathrm{d}x_2 \\
&= \mathbb{P}(\langle (X_1 X_2), (1,1)\rangle \in B)\mathbb{P}(\langle (X_1, X_2), (1,-1)\rangle \in C).
\end{aligned}$$

We can generalize this to $n > 2$: if X_i are independent (standard) Gaussians, then $X - X_i$ are independent. □

□

Beginning of Sept.8, 2021

Remark. S is *not* Gaussian. To find its distribution, one way is to notice that both S and S^2 are positive, so

$$\mathbb{P}(S \leqslant t) = \mathbb{P}(S^2 \leqslant t^2) = \mathbb{P}((n-1)S^2/\sigma^2 \leqslant (n-1)t^2/\sigma^2) = \mathbb{P}(\chi^2_{n-1} \leqslant (n-1)t^2/\sigma^2) = \int_0^{(n-1)t^2/\sigma^2} f_{\chi^2_{n-1}}(x)\,\mathrm{d}x.$$

Differentiating the above expression (along with chain rule) would give us the pdf of S.

1.1 Student's t-distribution

Recall that

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

has mean 0 and variance 1, given $X_1, \ldots, X_n$ are i.i.d. with finite mean and variance in $(0, \infty)$. Dividing both the numerator and the denominator by n gives

$$\frac{(X_1 + \ldots + X_n)/n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Now suppose that μ, σ are unknown, and we want to find them using $X_1, \ldots, X_n$.
It may be annoying to have two unknowns in one such equation, so sometimes we replace σ by the sample standard deviation, S, so that μ is the only free parameter, despite the fact that we don't know σ either:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

If $X_1, \ldots, X_n$ are i.i.d. *Gaussians*, then the above quotient has the **Student's t-distribution**.

Proposition: (3.7) Student's t-distribution

Let X be a standard Gaussian random variable. Let Y be a chi-squared random variable with p degrees of freedom. Assume X and Y are independent. Then

$$\frac{X}{\sqrt{Y/p}}$$

has a **student's t-distribution** with

$$f_{X/\sqrt{Y/p}}(t) = \frac{\Gamma((p+1)/2)}{\sqrt{p\pi}\,\Gamma(p/2)}\left(1 + \frac{t^2}{p}\right)^{-(p+1)/2}, \qquad t \in \mathbb{R}.$$

Proof. First we define $Z := \sqrt{Y/p}$. It follows that for any $y > 0$,

$$\begin{aligned} f_Z(y) &= \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Z \leqslant y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Y \leqslant y^2 p) = \frac{\mathrm{d}}{\mathrm{d}y} \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\int_0^{y^2 p} \frac{x^{p/2-1}e^{-x/2}}{2^{p/2}\Gamma(p/2)}\,\mathrm{d}x \\ &= 2yp \cdot p^{p/2-1}y^{p-2}e^{-y^2 p/2} \cdot \frac{1}{2^{p/2}\Gamma(p/2)} \\ &= 2y^{p-1}p^{p/2}e^{-y^2 p/2} \cdot \frac{1}{2^{p/2}\Gamma(p/2)}. \end{aligned}$$

Now we look at the CDF of X/Z:

$$\mathbb{P}(X/Z \leqslant t) = \mathbb{P}(X \leqslant tZ) = \iint_{\substack{x\leqslant ty\\ y>0}} f_{X,Z}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$
$$[\text{independence}] = \iint_{\dots} f_X(x)f_Z(y)\,\mathrm{d}x\mathrm{d}y. \tag{Δ}$$

Now we apply change of variable $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$ by $\varphi(a,b) = (ab, a)$ and $\varphi^{-1}(x,y) = (y, x/y)$. Then

$$|\mathcal{J}(a,b)| = \text{abs value of } \begin{vmatrix} b & 1 \\ a & 0 \end{vmatrix} = |a|.$$

Then (Δ) becomes

$$\begin{aligned}\iint_{\dots} f(x,y)\,\mathrm{d}x\mathrm{d}y &= \iint_{\varphi^{-1}(\dots)} f(\varphi(a,b))|\mathcal{J}(,)|\,\mathrm{d}a\mathrm{d}b \\ &= \iint_{\substack{b\leqslant t\\ a>0}} |a| f_X(ab) f_Z(b)\,\mathrm{d}a\mathrm{b} \\ &= \int_{b=-\infty}^{b=t}\int_{a=0}^{a=\infty} |a| f_X(ab) f_Z(b)\,\mathrm{d}a\,\mathrm{d}b. \end{aligned} \tag{$\square$}$$

Recall that we will eventually $\mathrm{d}/\mathrm{d}t$ everything — this is exactly why we want $b = t$ as the upper limit of the outer integral: the derivative of this integral becomes

$$\begin{aligned} f_{X/Z}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(X/Z \leqslant t) \\ &= \int_{a=0}^{a=\infty} |a| f_X(at) f_Z(a)\,\mathrm{d}a \\ &= \int_0^\infty a \frac{1}{\sqrt{2\pi}} e^{-a^2t^2/2} p^{p/2} a^{p-1} e^{-a^2 p/2}\,\mathrm{d}a \cdot \frac{1}{2^{p/2-1}\Gamma(p/2)} \\ &= \frac{p^{p/2}}{\sqrt{2\pi}2^{p/2-1}\Gamma(p/2)} \int_0^\infty a^p \exp\left(-\frac{a^2}{2}(t^2+p)\right)\mathrm{d}a \\ &= \frac{p^{p/2}}{\sqrt{2\pi}2^{p/2-1}\Gamma(p/2)} \int_0^\infty x^{p/2} \cdot \frac{1}{2\sqrt{x}} \cdot \exp\left(-\frac{x}{2}(t^2+p)\right)\mathrm{d}x \\ &= \frac{p^{p/2}}{\sqrt{2\pi}2^{p/2}\Gamma(p/2)} \int_0^\infty x^{(p-1)/2} e^{-(t^2+p)x/2}\,\mathrm{d}x \end{aligned}$$

where the integrand is related to a Gamma distribution with parameters $\alpha - 1 = (p-1)/2$ and $\beta = 2/(t^2+p)$. Therefore it evaluates to $\beta^\alpha\Gamma(\alpha)$. Hence

$$f_{X/Z}(t) = \frac{1}{\sqrt{2\pi}} \frac{p^{p/2}\beta^\alpha\Gamma(\alpha)}{2^{p/2}\Gamma(p/2)} = \dots = \frac{\Gamma((p+1)/2)}{\sqrt{p}\sqrt{\pi}\Gamma(p/2)} \left(1 + \frac{t^2}{p}\right)^{-(p+1)/2}.$$

□

Beginning of Sept.10, 2021

The Delta Method

If $X_1, \dots, X_n$ are i.i.d., we have a "good" way to estimate the mean in the sense that

$$\mathbb{E}\overline{X} = \mathbb{E}X_1 \qquad \mathrm{var}(\overline{X}) = \frac{\sigma^2}{n}.$$

What about $1/\mu$ or μ^2 or other functions?

Theorem: (3.13) Delta Method

Let $\theta \in \mathbb{R}$, and let $Y_1, Y_2, \ldots$ be random variables (not necessarily i.i.d.!) such that

$$\sqrt{n}(Y_n - \theta)$$

converges in distribution to a Gaussian with mean 0 and positive variance. Let $f : \mathbb{R} \to \mathbb{R}$. Assume $f'(\theta)$ exists. Then

$$\sqrt{n}(f(Y_n) - f(\theta))$$

converges in distribution to a Gaussian with mean 0 and variance $\sigma^2(f'(\theta))^2$.

Example: (3.14). If we let $f(x) = 1/x$, let $\overline{X} = Y_n$, and assume that $\sqrt{n}(\overline{X} - \mu)$ converges in distribution to a Gaussian with mean 0 and positive variance, then the Delta method says that

$$\sqrt{n}(f(Y_n) - f(\mu))$$

converges in distribution to a Gaussian with mean 0 and variance $\sigma^2\mu^{-4}$. Therefore $1/\overline{X}$ has expected value $\approx 1/\mu$ and variance $\approx n^{-1}\sigma^2\mu^{-4}$. As $n \to \infty$ the variance becomes small, so $1/\overline{X}$ is a "good" estimate of $1/\mu$. Upshot: $1/\overline{X}$ might still be a *biased* estimate of $1/\mu$, but as $n \to \infty$ the limit becomes $1/\mu$. In other words, $1/\overline{X}$ are an *asymptotically unbiased* estimate of $1/\mu$.

Beginning of Sept.13, 2021

Proof Sketch. By Taylor expansion around θ,

$$f(y) = f(\theta) + f'(\theta)(y - \theta) + \text{Error}.$$

Substituting Y_n into the above equation and making some arrangements,

$$\sqrt{n}(f(Y_n) - f(\theta)) = \sqrt{n}f'(\theta)(Y_n - \theta) + \text{Error}.$$

The claim "then follows" as the error $\to 0$. □

Remark. Currently, if $f'(\theta) = 0$, we get a mean zero variance zero Gaussian, which is simply a constant random variable. The theorem below fixes this issue:

Theorem: (3.16) Second-Order Delta Method

Following the previous theorem, if we further assume that $f'(\theta) = 0$ but $f''(\theta)$ exists and is nonzero, then

$$n(f(Y_n) - f(\theta))$$

converges in distrbution to a chi-squared random variable with *one* degree of freedom (χ_1^2), multiplied by $\sigma^2 f''(\theta)/2$.

Proof Sketch. Like above we have

$$f(Y_n) = f(\theta) + \underbrace{f'(\theta)}_{=0}(y-\theta) + \frac{1}{2}f''(\theta)(y-\theta)^2 + \text{Error}$$

so

$$n(f(Y_n) - f(\theta)) = \frac{n}{2}f''(\theta)(Y_n-\theta)^2 + \text{Error}.$$

As $n \to \infty$, $(\sqrt{n}(Y_n-\theta))^2 f''(\theta)$ converges in distribution to the square of a mean zero Gaussian, multiplied by $\sigma^2 f''(\theta)/2$. □

Example 1.1.1. Let $X_1, X_2, \ldots$ be i.i.d., let $f(x) := x^2$, and let $Y_n := (X_1 + \ldots + X_n)/n$. Then the second-order Delta method says that $n(Y_n^2 - 0)$ converges in distribution to χ_1^2 multiplied by $\sigma^2 f''(0)/2$, i.e.,

$$E(nY_n^2) \approx \frac{1}{2}\sigma^2 \cdot 2 = \sigma^2$$

In other words, $\mathbb{E}Y_n^2 \approx \sigma^2/n$. Also,

$$\operatorname{var}(nY_n)^2 \approx \operatorname{var}(\chi_1^2 \cdot \sigma^2 f''(0)/2) = \sigma^4 \operatorname{var}(\chi_1^2) = 2\sigma^4,$$

so $\operatorname{var}(Y_n)^2 \approx 2\sigma^4/n^2$.

1.2 Simulation of Random Variables

When we simulate random quantities on a computer, the numbers generated are not actually random, as computers cannot store arbitrary real numbers. Instead, what's generated are **pseudorandom**. We check whether a PRNG (pseudorandom random number generator) behaves like a random variable by checking if it agrees with the LLN and the CLT.

Example: (3.18) Simulating Discrete RVs. If we are able to use computer to generate a uniformly distributed random variable U in $(0,1)$, we can simulate a discrete random variable by partitioning $(0,1)$ into subintervals, each corresponding to an outcome of the discrete random variable, based on the probability of each. For example,

$$X(U) := \begin{cases} 1 & U < 1/3 \\ 2 & 1/3 \leqslant U < 2/3 \\ 3 & 2/3 \leqslant U < 1 \end{cases}$$

simulates a discrete random variable that takes values in $\{1,2,3\}$, each with probability $1/3$.

Beginning of Sept.15, 2021

Example 1.2.1: Simulating Continuous RVs. If the CDF of a continuous random variable is given by F, then we can simulate it using the inverse F^{-1}. To put formally:

Let X be a continuous random variable. Let $F(t) := \mathbb{P}(X \leqslant t)$. If F^{-1} exists and if U is a uniform random variable on $(0,1)$, then $F^{-1}(U)$ is a random variable with

$$F^{-1}(U \leqslant t) = F(t).$$

Proof. $\mathbb{P}(F^{-1}(U) \leqslant t) = P(F(F^{-1}(U)) \leqslant F(t)) = P(U \leqslant F(t)) = F(t)$. □

1.3 Parameter Estimation

A basic problem in statistics is to fit data to an unknown probability distribution. For example, if we have data distribution of *some* unknown Gaussian distribution, what are some ways to figure out the mean and variance?

Stated formally, let $X_1, ..., X_n$ be a random variable of size n from a family of distributions $\{f_\theta : \theta \in \Theta\}$. (For example we can think of f_θ as a PDF or a PMF.)
If we are "guessing" the parameters of a Gaussian, Θ would be $\mathbb{R} \times [0, \infty)$ and θ would be of form (μ, σ^2):

$$\{f_{\mu,\sigma^2}(x) : (\mu, \sigma^2) \in \mathbb{R}^2, \mu \in \mathbb{R}, \sigma^2 > 0\}.$$

Definition 1.3.1: Estimator

If Y is a statistic that is used to estimate a parameter θ, then we call Y a **point estimator** or **estimator**. (Some examples include sample mean and sample variance we've previously talked about.)

Example 1.3.2. Let $X_1, ..., X_{20}$ be a random sample of 20 from a Gaussian with unknown mean and variance. Then the family $\{f_\theta : \theta \in \Theta\}$ is of form

$$\left\{ \frac{1}{2\pi\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) : \mu \in \mathbb{R}, \sigma^2 > 0 \right\}.$$

An estimator for the mean μ is $\overline{X} := (X_1 + ... + X_{20})/20$ (this is a good one!), and a not-so-good one is for example $X_1 + X_2$. Similarly, an estimator for σ^2 is

$$\frac{1}{19} \sum_{i=1}^{20} (X_i - \overline{X})^2.$$

Of course we can also have "worse" estimators too.

Definition 1.3.3: Unbiased Estimator

Let $X_1, ..., X_n$ be a random sample of size n from a family of distributions $\{f_\theta : \theta \in \Theta\}$. Let $Y = t(X_1, ..., X_n)$ be an estimator for $g(\theta)$ (e.g., in the Gaussian example we can have an estimator for not only μ but also μ^2, or also any function of μ). Here $t : \mathbb{R}^n \to \mathbb{R}^k$ and $g : \theta \to \mathbb{R}^k$. We say Y is **unbiased** if

$$\mathbb{E}_\theta(Y) = g(\theta) \qquad \text{for all } \theta \in \Theta.$$

In other words, the expected value of the estimator is exactly what it estimates.

Remark. Sample mean and sample variance are unbiased.

Besides asking "*how good an estimator is*", another natural question arises — "*how to get a good estimator?*"

Example 1.3.4. Let $X_1, ..., X_n$ be a random sample of size n. By the Weak LLN, if $\mathbb{E}_\theta|X_2| < \infty$ for all $\theta \in \Theta$, then the sample mean $(X_1 + ... + X_n)/n$ is not only unbiased but also converges *in probability* to the constant variable $\mathbb{E}_\theta X_1$. We say this estimator is **consistent**.
More generally, for $j \in \mathbb{N}$, if $\mathbb{E}_\theta|X_1|^j < \infty$, then

$$M_j(\theta) := \frac{1}{n}\sum_{i=1}^{n} X_i^j,$$

the **sample j^{th} moment**, is also consistent: $M_j(\theta)$ converges *in probability* to $\mu_j(\theta) := \mathbb{E}_\theta X_1^j$ as $n \to \infty$.

Definition: (4.5) Methods of Moments

Suppose we want to estimate $g(\theta)$ and suppose there exists $h : \mathbb{R}^j \to \mathbb{R}^k$ such that

$$g(\theta) = h(\mu_1, ..., \mu_j).$$

Then the estimator $h(M_1, ..., M_j)$ is called the **method of moments** estimator for $g(\theta)$.

Example 1.3.5. Let $g(\theta)$ be the variance. We know $\text{var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$. Then the MoM for $g(\theta)$ is $M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$.

Beginning of Sept.17, 2021

Definition: (4.3) Consistency

Let $Y_1, Y_2, ...$ be a sequence of estimators for $g(\theta)$. We say $Y_1, Y_2, ...$ is **consistent for $g(\theta)$** if $Y_1, Y_2, ...$ converges in probability to the constant random variable $g(\theta)$ with respect to f_θ.

Example: (4.6). Following the previous example, define

$$Y_n := \sqrt{\sum_{i=1}^{n} X_i^2/n - (\sum_{i=1}^{n} X_i/n)^2}.$$

Since $(a, b) \mapsto \sqrt{a - b^2}$ is continuous, and since $\sum_{i=1}^n X_i^2/n$ and $\sum_{i=1}^n X_i/n$ converge to $\mathbb{E}X^2$ and $\mathbb{E}X$ respectively, we claim that $Y_n \to \sqrt{\mathbb{E}X^2 - (\mathbb{E}X)^2}$ as $n \to \infty$. This implies that Y_n is *consistent*.
However, Y_n *is* biased! Take $n = 1$ and X the uniform distribution on $[0, 1]$. Then

$$\mathbb{E}X = \frac{1}{2}, \mathbb{E}X^2 = \frac{1}{3}, \text{var}(X) = \frac{1}{12}, \text{ and } \sigma = \frac{1}{2\sqrt{3}}.$$

On the other hand,

$$\mathbb{E}\sqrt{X^2 - X^2} = 0.$$

Therefore Y_n is *consistent but biased*.

Example 1.3.6: Unbiased but inconsistent estimator. Let $X_1, ..., X_n$ be i.i.d. uniform on $(0,1)$. Let $Y_n := X_1$ for every n. Then $\mathbb{E}Y_n = \mathbb{E}X = 1/2$ for all n, meaning that it is unbiased, yet they do not converge to the constant variable $1/2$ as X_1 itself isn't a constant variable, so Y_n is inconsistent.

Example: (4.7) . Suppose $X_1, ..., X_n$ is a random variable sample of size n, uniform on $[0,\theta]$ where $\theta > 0$ is unknown. What is a MoM estimator for θ?
Since $\mathbb{E}X_1 = \theta/2$, the MoM estimator for θ is given by $Y_n := 2\mu_1 = 2\sum_{i=1}^n X_i/n$.
Clearly Y_n is unbiased: $\mathbb{E}Y_n = \theta$ for all n. It is also consistent: since $\sum_{i=1}^n X_i/n$ converges in probability to $\mathbb{E}X = \theta/2$, multiplying both sides by 2 gives our desired claim.

Beginning of Sept.20, 2021

Example: (4.7). Let $X_1, ..., X_n$ be a random sample of size n from a uniform distribution on $[0,\theta]$, where θ is unknown. We showed last time that

$$Y_n := \frac{2}{n}\sum_{i=1}^n X_i$$

is an unbiased and consistent estimator of θ. Also,

$$\operatorname{var}(Y_n) = \frac{4}{n^2}\cdot n \cdot \operatorname{var}(X_1) = \frac{\theta^2}{3n}.$$

It turns out that there is an even "better" unbiased, consistent estimator

$$\left(1+\frac{1}{n}\right)X_{(n)} = \left(1+\frac{1}{n}\right)\max_{1\leqslant i\leqslant n} X_i \tag{Δ}$$

with *smaller* variance:

$$\operatorname{var}((1+1/n)X_{(n)}) = \frac{\theta^2}{n(n+2)}.$$

This means, in some sense, that this estimator is much more accurate than the previous one.
Upshot: *the method of moments may not give the best estimator.*

Proof sketch. For convenience call the estimator in (Δ) as $\hat{\theta}$. Then (recall that $\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geqslant t)\,\mathrm{d}t$)

$$\begin{aligned}
\mathbb{E}(\hat{\theta}) &= \left(1+\frac{1}{n}\right)\int_0^\theta \mathbb{P}(X_{(n)} > t)\,\mathrm{d}t \\
&= \left(1+\frac{1}{n}\right)\int_0^\theta \mathbb{P}(X_1 > t, ..., X_n > t\,\mathrm{d}t \\
&= \left(1+\frac{1}{n}\right)\int_0^\theta 1-(t/\theta)^n\,\mathrm{d}t \\
&= \left(1+\frac{1}{n}\right)(\theta - \int_0^\theta (t/\theta)^n\,\mathrm{d}t) \\
&= \left(1+\frac{1}{n}\right)(\theta - \theta^{-n}\theta^{n+1}/(n+1)) \\
&= \theta\cdot\frac{n+1}{n}\cdot\frac{n}{n+1} = \theta.
\end{aligned}$$

For variance, recall that [!] $\mathbb{E}X^n = \int_0^\infty nx^{n-1}\mathbb{P}(X \geqslant t)\,\mathrm{d}t$, so

$$\begin{aligned}\operatorname{var}(\hat{\theta}) &= \frac{(n+1)^2}{n^2}\mathbb{E}X_{(n)}^2 - \theta^2 \\ &= \frac{(n+1)^2}{n^2}\int_0^\theta 2t\mathbb{P}(X_{(n)} > t)\,\mathrm{d}t - \theta^2 \\ &= \frac{(n+1)^2}{n^2}\int_0^\theta 2t(1-(t/\theta)^n)\,\mathrm{d}t - \theta^2 = \ldots = \frac{\theta^2}{n(n+2)}.\end{aligned}$$

It turns out that among all unbiased estimators, $\hat{\theta}$ has the smallest variance. We say $(1+1/n)X_{(n)}$ is the **uniform minimum variance unbiased estimator** (UMVU estimator) for θ. □

1.4 Sufficient Statistics

Definition: (4.9) Sufficient Statistic

Suppose $X = (X_1, \ldots, X_n)$ be a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of PDFs or PMFs. Let $t : \mathbb{R}^n \to \mathbb{R}^k$ so that $Y := t(X_1, \ldots, X_n)$ is a statistic. We say Y is a **sufficient statistic** *for the parameter* θ if, for all $y \in \mathbb{R}^k$ and for all $\theta \in \Theta$, the conditional distribution of $X = (X_1, \ldots, X_n)$ given $Y = y$ does not depend on θ.
That is, Y provides sufficient information to estimate (not determine!) what θ is from our sample $X_1, \ldots, X_n$.

Example: (4.10). Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli distribution with unknown parameter $\theta \in (0,1)$. Claim: $Y_n := X_1 + \ldots + X_n$ is sufficient for θ.

Proof. By definition we compute the probability of $X = (x_1, \ldots, x_n)$ conditioned on $Y = y$ (assuming by definition that $x_i \in \{0,1\}$). Note that Y is a binomial with parameters n and p.

$$\mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \mid Y = y) = \frac{\mathbb{P}(X_1, \ldots, X_n, Y) = (x_1, \ldots, x_n, y)}{\mathbb{P}(Y = y)}.$$

If $y \neq x_1 + \ldots + x_n$ then the numerator is just 0, which does not depend on θ, and we are done with the proof. If $y = x_1 + \ldots + x_n$, then the numerator is just $\mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n))$, and so

$$\begin{aligned}\ldots &= \frac{\mathbb{P}(X_1, \ldots, X_n) = (x_1, \ldots, x_n)}{\mathbb{P}(Y = y)} \\ &= \frac{\prod_{i=1}^n \mathbb{P}(X_i = x_i)}{P(Y = y)} = \frac{\prod_{i=1}^n \theta^{x_i}(1=\theta)^{1-x_i}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} = \binom{n}{y}^{-1}.\end{aligned}$$

□

Beginning of Sept.22, 2021

Example: (4.11). Let $X_1, \ldots, X_n$ be i.i.d. Gaussians with known variance $\sigma^2 > 0$ and unknown mean. We will show that

$$Y := \frac{1}{n}(X_1 + \ldots + X_n)$$

is sufficient for μ.

Proof. Let $x_1, ..., x_n \in \mathbb{R}$ and let $y \in \mathbb{R}$. Then Y is Gaussian with mean ν and variance σ^2/n. We have

$$\begin{aligned}
f_{X_1,...,X_n|Y}(x_1,...,x_n \mid y) &= \frac{f_{X_1,...,X_n,Y}(x_1,...,x_n,y)}{f_Y(y)} \\
&= \frac{f_{X_1,...,X_n,Y}(x_1,...,x_n,(x_1+...+x_n)/n)}{f_Y(y)} \\
&\overset{*}{=} \frac{f_{X_1,...,X_n}(x_1,...,x_n)}{f_Y(y)} \\
&= \frac{\prod_{i=1}^n f_{X_i}(x_i)}{f_Y(y)} = \frac{\prod_{i=1}^n \exp(-(x_i-\mu)^2/(2\sigma^2))/(\sqrt{2\pi}\sigma)}{\exp\left(-\frac{((x_1+...+x_n)/n-\mu)^2}{2\sigma^2/n}\right)/(\sqrt{2\pi}\sigma/\sqrt{n})} \\
&= \frac{\sigma^{-n}(2\pi)^{-n/2}}{n^{1/2}\sigma^{-1}(2\pi)^{-1/2}} \frac{\exp(-(x_1^2+...+x_n^2)/(2\sigma^2) - n\mu^2/(2\sigma^2) + \sum_{i=1}^n x_i\cdot\mu/\sigma^2)}{\exp(-y^2 n/(2\sigma^2) - n\mu^2/(2\sigma^2) + n\mu y/\sigma^2)} \\
&= \frac{\sigma^{-n}(2\pi)^{-n/2}}{n^{1/2}\sigma^{-1}(2\pi)^{1/2}} \exp(-\sum_{i=1}^n x_i^2)/(2\sigma^2)/\exp(-y^2 n/(2\sigma^2)).
\end{aligned}$$

where $(*)$ is because Y is a function of $X_1, ..., X_n$ (so once $x_1, ..., x_n$ have been determined, y is automatically chosen). Since the last expression does not depend on μ, we have shown that Y is sufficient for μ. □

Theorem: (4.12) Factorization Theorem

This theorem provides an "easy" way to find or identify sufficient statistics. Suppose $X_1, ..., X_n$ is a random sample from $\{f_\theta : \theta \in \Theta\}$, where f_θ is a joint PDF of $X_1, ..., X_n$. Suppose $Y = t(X_1, ..., X_n)$ is a statistic and $t : \mathbb{R}^n \to \mathbb{R}^k$. Then Y is sufficient for θ if and only if there exist $h : \mathbb{R}^n \to [0, \infty)$ and $g_\theta : \mathbb{R}^k \to [0, \infty)$ such that

$$f_\theta(x) = g_\theta(t(x)) \cdot h(x) \qquad \text{for all } \theta \in \Theta.$$

Problem: (HW3 p4)

Let $\theta \in \mathbb{R}$ be an unknown parameter. Consider the density

$$f_\theta(x) := \begin{cases} \exp(-(x-\theta)) & x \geqslant \theta \\ 0 & x < \theta. \end{cases}$$

Suppose $X_1, ..., X_n$ is a random sample of size n such that each X_i has density f_θ. Show that $X_{(1)} = \min_{1 \leqslant i \leqslant n} X_i$ is a sufficient statistic for θ.

Proof. We first write $f_\theta(x) = \exp(-(x-\theta))\chi_{[\theta,\infty)}(x_i)$. Since X_i's are i.i.d., for $\vec{x} := (x_1, ..., x_n) \in \mathbb{R}^n$,

$$f_\theta(\vec{x}) = \prod_{i=1}^n f_\theta(x_i) = \exp(n\theta - \sum_{i=1}^n x_i) \prod_{i=1}^n \chi_{[\theta,\infty)}(x_i).$$

Note that $f_\theta(\vec{x}) \neq 0$ if and only if $x_i \geqslant \theta$ for all i, i.e., $X_{(1)} \geqslant \theta$. That is,

$$f_\theta(\tilde{x}) = \exp(n\theta - \sum_{i=1}^n x_i)\chi_{[\theta,\infty)}(X_{(1)}).$$

Therefore, $f_\theta(\tilde{x})$ admits a factorization

$$f_\theta(\tilde{x}) = \underbrace{\exp(n\theta)\chi_{[\theta,\infty)}(X_{(1)})}_{g_\theta(X_{(1)})} \cdot \underbrace{\exp(-\sum_{i=1}^{n} x_i)}_{h(x)},$$

which by the factorization theorem shows $X_{(1)}$ is sufficient. □

Proof of Factorization Theorem (discrete): sufficient ⇒ factorization. Suppose Y is sufficient for θ. Let $x \in \mathbb{R}^n$. Then (the starred equation is again because both sides are equivalent)

$$\begin{aligned} f_\theta(x) &= \mathbb{P}_\theta(X = x) \overset{*}{=} \mathbb{P}_\theta(X = x \text{ and } t(X) = t(x)) \\ &= \mathbb{P}_\theta(Y = t(x)) \cdot \mathbb{P}_\theta(X = x \mid Y = t(x)). \end{aligned}$$

Since Y is sufficient by assumption, the second term $\mathbb{P}_\theta(X = x \mid Y = t(x))$ does *not* depend on θ, i.e., it is a function of x only. We have therefore obtained our factorization. □

Remark: (4.13). If we let $t(x) := x$ for all $x \in \mathbb{R}^n$, then the statistic $t(X_1, ..., X_n) = (X_1, ..., X_n)$ is always trivially sufficient. Therefore we always have *a* sufficient static. Our goal is find a *minimal* sufficient statistic, using as little information as possible.

1.5 Evaluating Estimators

Beginning of Sept.24, 2021

Definition: (4.15) UMVU

Let $X_1, ..., X_n$ be i.i.d. from a distribution in $\{f_\theta : \theta \in \Theta\}$. Let $g : \Theta \to \mathbb{R}$, let $t \in \mathbb{R}^n \to \mathbb{R}$, and let $Y := t(X_1, ..., X_n)$ is unbiased. We say Y is **uniformly minimum variance unbiased** (UMVU) if for any other unbiased estimator Z for $g(\theta)$, we have

$$\text{var}_\theta(Y) \leqslant \text{var}_\theta(Z) \qquad \text{for all } \theta \in \Theta.$$

Remark. If we have an UMVU, we obtain the "best" unbiased estimator possible.

Example: (4.16) UMVU might not exist. Unfortunately UMVU might not exist. Suppose X is a binomial with known parameter n but unknown $\theta \in (0, 1)$ and we want an estimator for $g(\theta) := \theta/(1-\theta)$.
There is not even any unbiased estimator for $g(\theta)$! For any estimator $Y = t(X)$,

$$\mathbb{E}_\theta Y = \mathbb{E}_\theta t(X) = \sum_{i=1}^{n} \binom{n}{i} t(i)\theta^i(1-\theta)^i,$$

a polynomial of θ, whereas $\theta/(1-\theta)$ itself isn't. Therefore it's impossible to have $E_\theta Y = \theta/(1-\theta)$ for all $\theta \in (0, 1)$.

Example 1.5.1. Even if an unbiased estimator exists, a UMVU might not exist. Let Y, Z be unbiased for $g(\theta)$ where $\theta \in [0,1]$. It could happen that

$$\mathrm{var}_0(Y) < \mathrm{var}_0(Z) \qquad \text{and} \qquad \mathrm{var}_1(Z) < \mathrm{var}_0(Y).$$

Question. If an UMVU exists, how do we find it in practice? The following provides a possible method.

Theorem: (4.17) Rao-Blackwell

Let Z be sufficient for for $\{f_\theta : \theta \in \Theta\}$, and let Y be any unbiased estimator for $g(\theta)$. Define $W := E_\theta(Y \mid Z)$. (*Since Z is sufficient for θ, W is in fact not a function of θ.*) Let $\theta \in \Theta$ with $\mathrm{var}_\theta(Y) < \infty$. Then

$$\mathrm{var}_\theta(W) \leqslant \mathrm{var}_\theta(Y)$$

with equality only when $W = Y$.

Proof. By conditional Jensen's inequality with $\varphi(x) := x^2$,

$$(W-\theta)^2 = (\mathbb{E}_\theta(Y \mid Z) - \theta)^2 \leqslant \mathbb{E}((Y-\theta)^2 \mid Z).$$

Taking $\mathbb{E}$ of both sides gives (the first $\leqslant$ and the second $=$ are mentioned in HW3 p2)

$$\mathrm{var}_\theta(W) \leqslant \mathbb{E}(W-\theta)^2 = \mathbb{E}(\mathbb{E}_\theta[(Y-\theta)^2 \mid Z]) = \mathbb{E}_\theta(Y-\theta)^2 = \mathrm{var}_\theta(Y).$$

□

Remark: (4.21). If Y is unbiaed, then $E_\theta W = E_\theta(E_\theta[Y \mid Z]) = \mathbb{E}_\theta Y = \theta$, so W is always unbiased.

Beginning of Sept.29, 2021

Example: (4.23). Let $X_1, \ldots, X_n$ be i.i.d. with unknown mean μ. We compute $\mathbb{E}(X_1 \mid \sum_{i=1}^n X_i)$.

Solution. For $1 \leqslant k < \ell \leqslant n$, the joint distribution $(X_k, \sum_{i=1}^n X_i)$ is the same as that of $(X_\ell, \sum_{i=1}^n X_i)$, so

$$W := \mathbb{E}(X_1 \mid \sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i \mid \sum_{i=1}^n X_i) = \frac{1}{n}\mathbb{E}(\sum_{i=1}^n X_i \mid \sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n X_i.$$

In this case, $\mathrm{var}(X_1) = \sigma^2$ but $\mathrm{var}(W) = \sigma^2/n$. Rao-Blackwell gives an estimator with much smaller variance! (In fact we are not explicitly using Rao-Blackwell; $\sum_{i=1}^n X_i$ may not be sufficient (see quiz3 prep p1), but nevertheless W defined this way gives us a better estimator.)

1.6 Efficiency of Estimators

Previously we've talked about what is and how to find a good estimator. Now we turn our focus to "what makes an estimator good?"

Definition: (4.24) Fisher Information

Let $\{f_\theta : \theta \in \Theta\}$ be a family of multivariable PDFs or PMFs. Let $\theta \subset \mathbb{R}$. Let X be a random vector with distribution f_θ. The **Fisher information** of the family to be

$$I(\theta) := I_X(\theta) := \mathbb{E}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) \right)^2, \qquad \text{for all } \theta \in \Theta,$$

if it exists and is finite.

Remark. In order to define $I(\theta)$, the set $\{x \in \mathbb{R}^n : f_\theta(x) > 0\}$ should *not* depend on θ.

Example: (4.25). Let $\sigma > 0$. Let $f_\theta(x) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2} \right)$ for all $x \in \mathbb{R}$, $\theta \in \mathbb{R}$. (In other words we have Gaussians.) We have

$$\log f_\theta(x) = \log\left(\frac{1}{\sigma\sqrt{2\pi}} \right) \cdot -\frac{(x-\theta)^2}{2\sigma^2}$$

so

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) = \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{-(X-\theta)^2}{2\sigma^2},$$

and so

$$I(\theta) = \mathbb{E}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \frac{-(X-\theta)^2}{2\sigma^2} \right)^2 = \mathbb{E}_\theta \left(\frac{X-\theta}{\sigma^2} \right)^2 = \frac{1}{\sigma^4} \operatorname{var}(X-\theta) = \frac{1}{\sigma^2}.$$

Remark. When σ is small, f_θ looks more like a sharp bump than a flat curve. A small σ corresponds to a larger $I(\theta)$ which gives us "more information" about how the random variable is distributed.

Remark. Without the square,

$$\mathbb{E}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) \right) = \int_{\mathbb{R}^n} \frac{\mathrm{d}/\mathrm{d}\theta f_\theta(x)}{f_\theta(x)} f_\theta(x)\, \mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\theta} \int_{\mathbb{R}^n} f_\theta(x)\, \mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\theta}(1) = 0.$$

Therefore, treating $\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X)$ as a random variable,

$$I(\theta) = \mathbb{E}_\theta(\ldots)^2 = \operatorname{var}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) \right).$$

Remark. Alternatively,

$$\begin{aligned}\mathbb{E}_\theta\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f_\theta(X)\right) &= \int_{\mathbb{R}^n}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\mathrm{d}/\mathrm{d}\theta f_\theta(x)}{f_\theta(x)}f_\theta(x)\,\mathrm{d}x\\ &= \int_{\mathbb{R}^n}\frac{f_\theta(x)\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f_\theta(x)-(\frac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x))^2}{(f_\theta(x))^2}f_\theta(x)\,\mathrm{d}x\\ &= \int_{\mathbb{R}^n}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f_\theta(x)-\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(x)\right)^2 f_\theta(x)\,\mathrm{d}x\\ &= \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}(1)-\int_{\mathbb{R}^n}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(x)\right)^2 f_\theta(x)\,\mathrm{d}x = 0-I(\theta) = -I(\theta).\end{aligned}$$

Proposition: (4.26)

Let X, Y be independent where their distributions are from $\{f_\theta : \theta\in\Theta\}$ and $\{g_\theta : \theta\in\Theta\}$ respectively. Then

$$I_{(X,Y)}(\theta) = I_X(\theta)+I_Y(\theta).$$

Proof. Using the variance expression,

$$\begin{aligned}I_{(X,Y)}(\theta) &\overset{*}{=} \operatorname{var}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log(f_\theta(X)g_\theta(Y))\right) = \operatorname{var}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(X)+\log g_\theta(X)\right)\\ &\overset{*}{=} \operatorname{var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right)+\operatorname{var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log g_\theta(X)\right) = I_X(\theta)+I_Y(\theta).\end{aligned}$$

(The starred equations are because of independence.) □

Beginning of Oct.1, 2021

Theorem: (4.28) Cramér-Rao / Information Inequality

Let $X:\Omega\to\mathbb{R}^n$ be a random variable with distribution from $\{f_\theta:\theta\in\Theta\}$, $\Theta\subset\mathbb{R}$. Let $Y := t(X)$ be a statistic. For $\theta\in\Theta$, define $g(\theta) := \mathbb{E}_\theta Y$. Then

$$\operatorname{var}_\theta(Y)\geqslant\frac{|g'(\theta)|^2}{I_X(\theta)}\qquad\text{for all }\theta\in\Theta.$$

In particular if Y is *unbiased* then $g(\theta)=\theta$ and $g'(\theta)=1$, so

$$\operatorname{var}_\theta(Y)\geqslant\frac{1}{I_X(\theta)}\qquad\text{for all }\theta\in\Theta.$$

In both cases, "=" happens only when $\dfrac{\mathrm{d}/\mathrm{d}\theta(\log f_\theta(X))}{Y-\mathbb{E}_\theta Y}\in\mathbb{R}$ for some $\theta\in\Theta$.

This theorem provides a lower bound on the variance of unbiased estimators of θ — in general, we cannot get estimators with arbitrarily small variance.

Remark: (4.29). If $X_1,\ldots,X_n$ are i.i.d. and $X=(X_1,\ldots,X_n)$, then (by last proposition) $I_X(\theta)=nI_{X_1}(\theta)$. If $\mathbb{E}_\theta Y=\theta$, then $\operatorname{var}_\theta(Y)\geqslant 1/(nI_{X_1}(\theta))$ for all $\theta\in\Theta$.

Proof. Define $g(\theta)$, Y, and t accordingly. If X is continuous (similar for discrete),

$$\begin{aligned}
|g'(\theta)| &= \left|\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathbb{R}^n} f_\theta(x)t(x)\,\mathrm{d}x\right| = \left|\int_{\mathbb{R}^n}\frac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x)t(x)\,\mathrm{d}x\right| \\
&\underset{*}{=} \left|\int_{\mathbb{R}^n}\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(x))\,t(x)f_\theta(x)\,\mathrm{d}x\right| \\
&\overset{*}{=} \left|\mathrm{cov}(\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(X)), t(X))\right| \\
&\leqslant \left(\mathrm{var}_\theta(\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(X)))\right)^{1/2}\mathrm{var}_\theta(t(X))^{1/2} \\
&= \sqrt{I_X(\theta)}\sqrt{\mathrm{var}_\theta Y}.
\end{aligned}$$

For $\underset{*}{=}$: $\dfrac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(x)) = \dfrac{1}{f_\theta(x)}\dfrac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x)$ [note that $t(x)$ is treated as a constant when doing $\mathrm{d}/\mathrm{d}\theta$], and for $\overset{*}{=}$: if $\mathbb{E}W = 0$, then $\mathrm{cov}(W,Z) = \mathbb{E}[(W-\mathbb{E}W)(Z-\mathbb{E}Z)] = \mathbb{E}[W(Z-\mathbb{E}Z)] = \mathbb{E}(WZ)$.

Note that equality in Cramér-Rao happens if and only if the Cauchy-Schwarz step is attained, i.e., when

$$\frac{\mathrm{d}/\mathrm{d}\theta(\log f_\theta(X)) - \mathbb{E}(\ldots)}{t(X)-\mathbb{E}(t_\theta(X))} = \frac{\mathrm{d}/\mathrm{d}\theta(\log f_\theta(X))}{Y-\mathbb{E}_\theta Y} \text{ is a constant.}$$

□

Example: (4.30). Let $f_\theta(x) := \theta x^{\theta-1}\chi_{(0,1)}(x)$ for $x\in\mathbb{R}$ and $\theta>0$. Then for $x\in(0,1)$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(x) = \frac{\mathrm{d}}{\mathrm{d}\theta}\log(\theta x^{\theta-1}) = \frac{\mathrm{d}}{\mathrm{d}\theta}\left[\log\theta + (\theta-1)\log x\right] = \frac{1}{\theta}+\log x.$$

Then if $X_1,\ldots,X_n$ are i.i.d., for $(x_1,\ldots,x_n)\in(0,1)^n$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^n f_\theta(x_i) = \sum_{i=1}^n(\theta^{-1}+\log x_i) = n\left(\frac{1}{\theta}+\frac{1}{n}\log\sum_{i=1}^n x_i\right).$$

By Cramér-Rao, any multiple of $\dfrac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^n f_\theta(X_i)$ (plus a constant) is UMVU for $\mathbb{E}_\theta Y$.

For example, since $\mathbb{E}(\dfrac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^n f_\theta(X_i)) = 0$, we know $\mathbb{E}\sum_{i=1}^n\log X_i = -n/\theta$. Hence if we define $Y := -\dfrac{1}{n}\log\prod_{i=1}^n X_i$, its expected valve is $1/\theta$, and we claim that this is UMVU of its expectation.

Beginning of Oct.4, 2021

1.7 Maximum Likelihood Estimator (MLE)

Definition 1.7.1: Likelihood Function

Let $X_1, ..., X_n$ be i.i.d. from $f_\theta \in \{f_\theta : \theta \in \Theta\}$. The joint distribution of $X_1, ..., X_n$, by independence, is $\prod_{i=1}^n f_\theta(x_i)$. Fix $(x_1, ..., x_n) \in \mathbb{R}^n$. We define the **likelihood function** $\ell : \Theta \to [0, \infty)$ by $\ell(\theta) := \prod_{i=1}^n f_\theta(x_i)$.

Definition: (4.32) MLE

The **maximum likelihood estimator** (**MLE**) Y is an estimator that maximizes the likelihood function.
In other words, $Y = t(X), t : \mathbb{R}^n \to \Theta$, $X = (X_1, ..., X_n)$, and $t(x_1, ..., x_n)$ is defined to be any value of $\theta \in \Theta$ that maximizes $\ell(\theta) = \prod_{i=1}^n f_\theta(x_i)$.
(The θ maximizing $\ell(\theta)$ might not exist; even if it exists, it might not be unique.)

Remark: (4.33). Since $\log$ is monotone, whatever maximizes $\ell(\theta) = \prod_{i=1}^n f_\theta(x_i)$ also maximizes $\log \ell(\theta) = \sum_{i=1}^n \log f_\theta(x_i)$ and vice versa. Sometimes it might be more convenient to maximize the latter.

Example: (4.34). Let $X_1, .., X_n$ be i.i.d. from f_θ with $f_\theta(x_i) = \chi_{[\theta,\theta+1)}(x_i)$, i.e., X is uniform on $[\theta, \theta+1]$. Then the joint PDF is $\prod_{i=1}^n \chi_{x_i \in [\theta,\theta+1]}$.
Suppose for example that $x_1 = ... = x_n = 0$. Then $\ell(\theta) = \chi_{0 \in [\theta,\theta+1]} = \chi_{\theta \in [-1,0]}$, so any $\theta \in [-1, 0]$ is a MLE in this case. Uncountably many!

Example: (4.41). Consider a Gaussian with unknown $\mu \in \mathbb{R}$ and unknown $\sigma^2 > 0$ so $\theta = (\mu, \sigma)$. Find its MLE.

Solution. Here we maximize $\log \ell(\theta)$:

$$\log \ell(\theta) = \log \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \sum_{i=1}^n \left[-\log \sigma - \frac{\log 2\pi}{2} - \frac{(x_i-\mu)^2}{2\sigma^2}\right].$$

Computing its partials,

$$\frac{\partial}{\partial \mu} \log \ell(\theta) = \frac{x_i - \mu}{\sigma^2} \qquad \frac{\partial}{\partial \sigma} \log \ell(\theta) = \sum_{i=1}^n -\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}.$$

Setting them to 0, we obtain

$$\mu = \frac{1}{n}\sum_{i=1}^n x_i \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2.$$

(Note that we did not get $1/(n-1)$ for σ^2, but nevertheless this is still pretty good.)
Now that we found a critical point, we need to verify that it is a maximum. Write $\alpha := 1/\sigma^2$. Then

$$\log \ell(\theta) = \frac{1}{2}\left(\sum_{i=1}^n \log \alpha - \log 2\pi - \alpha(x_i - \mu)^2\right)$$

For fixed α, $\log \ell(\theta)$ is strictly concave function of μ; likewise, fixing μ, $\log \ell(\theta)$ is a strictly concave function of α, so the critical point must have been a global maximum. We have therefore found *the* (only) MLE:

$$\theta = (\mu, \sigma^2) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i, \left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \frac{1}{n} \sum_{i=1}^{n} X_i)^2 \right)^{1/2} \right).$$

□

Beginning of Oct.6, 2021

Definition 1.7.2: Convex & Strictly Convex Functions

A function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is **convex** if for all $x \neq y$ and $\lambda \in [0,1]$,

$$\varphi(\lambda x + (1-\lambda)y) \leqslant \lambda \varphi(x) + (1-\lambda)\varphi(y).$$

φ is said to be **strictly convex** if the above holds with $\leqslant$ replaced by $<$.
Replacing $\leqslant$ and $<$ by $\geqslant$ and $>$, we obtain the definitions of concave and strictly concave functions.

Definition: (4.35) Log-Concavity

We say $\varphi : \mathbb{R}^n \to (0, \infty)$ is **log-concave** if $\log \varphi$ is concave. We sayφ is **strictly log-concave** if $\log \varphi$ is.

Proposition: (4.36) MLE and Log-Concavity

Let $f_\theta : \mathbb{R} \to [0, \theta)$ be a family of PDFs where $\theta \in \Theta \subset \mathbb{R}^k$. If $\theta \mapsto f_\theta(x_i)$ is strictly log-concave for every $i \in \{1, ..., n\}$, then the likelihood function

$$\ell(\theta) : \theta \mapsto \prod_{i=1}^{n} f_\theta(x_i)$$

has at most one maximum value.

Proof. The log of the likelihood function, $\log \ell(\theta)$, is $\sum_{i=1}^{n} \log f_\theta(x_i)$. By assumption this is the sum of strictly concave functions so it itself is also strictly concave. But a strictly concave function has at most one maximum (if $x \neq y$ are both maxima then $1/2(x+y)$ takes a higher function value by strict concavity, contradiction). □

Beginning of Oct.8, 2021

Example: (4.42). Let $X_1, ..., X_n$ be i.i.d. uniform on $[0, \theta]$. Let $x_1, x_2, ..., x_n \in \mathbb{R}$ be given. For a MLE, we need to find θ maximizing

$$\ell(\theta) = \theta^{-n} \chi_{0 \leqslant X_{(1)}, X_{(n)} \leqslant \theta}.$$

To maximize this, of course we need the indicator function to be 1. While keeping this true, we need θ^{-n} to be as large as possible, so θ needs to be as small as possible. Thus the MLE for θ is simply $X_{(n)}$.
Recall that the UMVU in this case is $(1 + 1/n)X_{(n)}$. Hence our MLE is asymptotically equivalent though biased.

Example: (4.43). Let $X_1, ..., X_n$ be i.i.d. from the exponential density $\chi_{x>0}\theta e^{-\theta x}$ with θ unknown. The log of the likelihood function is

$$\log \ell(\theta) = \log \chi_{x_i>0 \forall i}(n \log \theta - \theta \sum_{i=1}^{n} x_i)$$

so, assuming $x_i > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log \ell(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i.$$

Setting this to 0, we see $\theta := n/\sum_{i=1}^{n} x_i$ is a critical point (the only one), and it is clear that $(\log \ell(\theta))' < 0$ when $\theta < n/\sum_{i=1}^{n} x_i$ and > 0 when $>$. Thus we have found the *unique* maximum of $\log \ell(\theta)$ and

$$Y := \frac{n}{\sum_{i=1}^{n} X_i} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1}$$

is *the* MLE for θ.

How good are the estimators Y_n?
Recall that $\mathbb{E}X_1 = 1/\theta$ and $\mathrm{var}(X_1) = 1/\theta^2$. The CLT states that

$$\sqrt{n}(\overline{X}_n - \theta^{-1})$$

converges in distribution to a Gaussian with mean 0 and variance $1/\theta^2$. Using Delta method with $g(x) = 1/x, g'(x) = -1/x^2$, we see

$$\sqrt{n}(1/\overline{X_n} - g(1/\theta)) = \sqrt{n}(1/\overline{X}_n - \theta) = \sqrt{n}(Y_n - \theta)$$

converges in distribution to a Gaussian with mean 0 and variance $(g'(1/\theta))^2/\theta^2 = \theta^2$ as $n \to \infty$. This shows the Y_n's are asymptotically unbiased and consistent. Hence

$$\mathrm{var}(Y_n) = \mathrm{var}(1/\overline{X}_n) \approx \frac{\theta^2}{n}.$$

(More rigorously, $\mathrm{var}(Y_N) = \theta^2/n(1 + \mathcal{O}(1))$.) On the other hand, Cramér-Rao says

$$\mathrm{var}(Y) \geqslant \frac{1}{I_Y(\theta)} = \frac{1}{\mathrm{var}(\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{n}{\theta} - \sum_{i=1}^{n} x_i\right])} = \frac{1}{\mathrm{var}(\frac{\mathrm{d}}{\mathrm{d}\theta}[-\sum_{i=1}^{n} x_i])} = \frac{\theta^2}{n},$$

so the MLE is pretty close to the UMVU (if there is any) too.

Example: (4.44). Continuation of the previous example: the MLE for $e^{-\theta}$ would simple be e^{-Y} (because $\exp$ is a bijection; see proposition below).

Proposition: (4.45) Functional Equivariance of MLE

Let $g : \Theta \to \Theta'$ be a bijection. Then Y is the MLE for $\theta \Rightarrow g(Y)$ is the MLE of $g(\theta)$.

Proof. Since g is bijective, we write $\ell(\theta)$ as $\ell(g^{-1}g(\theta))$. Then $\ell(\theta)$ attains maximum at $\theta = x \Leftrightarrow \ell(g^{-1}(g(\theta)))$ attains maximum when $g^{-1}(g(\theta)) = x$, i.e., when $g(\theta) = g(x)$. □

Remark. Under some technical assumptions, MLE is *always* consistent, asumptotically unbiased, and achieves Cramér-Rao *equality*.

Chapter 2

Hypothesis Testing

Recall MLE asks "what is a good estimator of an unknown parameter?"
Hypothesis testing asks "does an unknown parameter lie in some range $[a, b]$ with at least $90\&$ certainty?"

Definition: (5.11) Null Hypothesis, Alternative Hypothesis

Let $\{f_\theta : \theta \in \Theta\}$ be a family of distributions. Let $\Theta_0 \subset \Theta$. A **null hypothesis** H_0 is an event of form

$$\{\theta \in \Theta_0\}.$$

Define $\Theta_1 := \Theta_0^c$ so $\Theta = \Theta_0 \sqcup \Theta_1$. The **alternative hypothesis** H_1 is the event $\{\theta \in \Theta_1\}$.

Goal: test whether or not H_0 is true or false.
Let $X : \Omega \to \mathbb{R}^n$ be a random variable with distribution $f_\theta \in \{f_\theta : \theta \in \Theta\}$.

Definition: (5.4) Critical/Rejection Region

Let H_0 be a null hypothesis. A **hypothesis test** of H_0 vs. H_1 is specified by a subset $C \subset \mathbb{R}^n$. The set C is called the **critical region** or the **rejection region**.

(1) If $X \notin C$, we accept H_0.

(2) If $X \in C$, we reject H_0 and assert that H_1 is true.

The complement $C^c \subset \mathbb{R}^n$ is called the **acceptance region**. The performance of the test is quantified by the **power function** $\beta : \Theta \to [0, 1]$ by

$$\beta(\theta) := \mathbb{P}_\theta(X \in C) = 1 - \mathbb{P}_\theta(X \notin C).$$

Remark. Ideally, we want to find a "perfect" test in the sense that $\beta(\theta) = 0$ for all $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for all $\theta \in \Theta_1$, i.e., if the null hypothesis is true then we accept it with probability 1 and if it is false, we accept it with probability 0. However, this might not always happen.

Definition: (5.5) Type II Error: *false negative*

A **type II error** for a hypothesis test occurs when $X \notin C$ with positive probability but H_0 is actually false. That is, $\beta(\theta) < 1$ for some $\theta \in \Theta_1$. In other words, H_0 is accepted to be true whereas it is actually false. The quantity $1 - \beta(\theta)$ is the probability of occurrence of a type II error for $\theta \in \Theta_1$.

Definition: (5.6) Type I Error: *false positive*

A **type I error** for a hypothesis test occurs when $X \in C$ with positive probability but H_1 is actually false. That is, $\beta(\theta) > 0$ for some $\theta \in \Theta_0$. In other words, H_1 is accepted to be true whereas it is actually false. The value of $\beta(\theta)$ is the probability of occurrence of a type I error for $\theta \in \Theta_0$.

Definition 2.0.1: Significance Level

The **significance level** α is defined as

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta).$$

This shows the "worst" probability of a type I error (false positive) occurring.

Beginning of Oct.13, 2021

Example: (5.7). Let X be a binomial r.v. with parameters $n = 5$ and $\theta \in [0,1] =: \Theta$. We let $H_0 := \{0 \leqslant \theta \leqslant 1/2\}$ and $H_1 := \{1/2 < \theta \leqslant 1\}$.

If θ is small, we expect X to take smaller values more likely, so a "good" hypothesis test should use a rejection region corresponding to large values of X.

We first let the rejection region to be $C := \{5\}$. That is,

$$\text{If } X \notin C, \text{ i.e., if } 0 \leqslant X \leqslant 4, \text{ accept } H_0$$
$$\text{If } X \in C, \text{ i.e., if } X = 5, \text{ reject } H_0.$$

In this case

$$\beta(\theta) = \mathbb{P}_\theta(X \in C) = \mathbb{P}_\theta(X = 5) = \theta^5.$$

Then

$$\alpha = \sup_{[0,1/2]} \theta^5 = 2^{-5}.$$

The worst probability of a type I error happening is pretty small. However, type II errors are much more likely to happen: for small $\theta > 0.5$, $1 - \beta(\theta)$ is not close to 1. For example $1 - \beta(0.6) \approx 0.92$.

Now instead consider another test and let $C = \{3, 4, 5\}$. In this case

$$\beta(\theta) = \mathbb{P}_\theta(X \geqslant 3) = \theta^5 + 5\theta^4(1-\theta) + 10\theta^3(1-\theta)^2.$$

For this one, $\alpha = 1/2$ is worse, but type II errors are better: for example $1 - \beta(0.6) \approx 0.32$.

Question. Is there a "best" hypothesis test?

Definition: (5.8) Uniformly Most Powerful Test (UMP)

Let $\Theta_0 \subset \Theta$ and denote $\Theta_1 := \Theta_0^c$. Let H_0 be the hypothesis $\theta \in \Theta_0$ and H_1 be $\{\theta \in \Theta_1\}$. Let $\mathcal{T}$ be a family of hypothesis tests. A hypothesis test in $\mathcal{T}$ with power function $\beta(\theta)$ is called the **uniformly most powerful class $\mathcal{T}$ test** if $\beta(\theta) \geqslant \beta'(\theta)$ for all $\theta \in \Theta_1$ for every $\beta'(\theta)$ corresponding to a hypothesis test in $\mathcal{T}$.

Remark. Since UMP only focuses on Θ_1, it is sometimes helpful to fix $\alpha > 0$ and let $\mathcal{T}$ be the class of all hypothesis tests with significance level $\leqslant \alpha$.

Remark. The existence of a UMP for general Θ is a difficult question, but if Θ consists of exactly two points, UMP always exists, and we can explicitly construct them.

2.1 Neyman-Pearson Testing

Lemma: (5.9) Neyman-Pearson

Let $\Theta = \{\theta_0, \theta_1\}$, $\Theta_0 := \{\theta_0\}$, and $\Theta_1 := \{\theta_1\}$. Let H_0 be the hypothesis $\{\theta = \theta_0\}$ and H_1 be $\{\theta = \theta_1\}$. Let $\{f_{\theta_0}, f_{\theta_1}\}$ be two multivariable PDFs or PMFs. Fix $k \geqslant 0$. Define the **likelihood ratio test** with rejection region C in the following way:

$$C := \{x \in \mathbb{R}^n : f_{\theta_1}(x) > k f_{\theta_0}(x)\}. \tag{1}$$

As usual, define

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = \mathbb{P}_{\theta_0}(X \in C). \tag{2}$$

Let $\mathcal{T}$ be the class of hypothesis tests with significance levels $\leqslant \alpha$. Then:

- (Sufficiency) Any hypothesis test satisfying (1) is a UMP class $\mathcal{T}$ test.
- (Necessity) If there exists a hypothesis test satisfying (1) and (2) with $k > 0$, then any UMP class $\mathcal{T}$ test has significance level equal to α, and any UMP class $\mathcal{T}$ test satisfies (1), except possibly on a null set D with $\mathbb{P}_{\theta_1}(X \in D) = 0$.

Beginning of Oct.18, 2021

Proof. Assume f_θ's are PDFs. Also recall that $\alpha = \sup_{\theta\in\Theta_0} \beta(\theta) = \beta(\theta_0)$. Let $\beta(\theta)$ be the power function of the test corresponding to C, and let C' be the rejection region of *any* UMP class $\mathcal{T}$ test with $\beta'(\theta)$ being its power function. By definition of C,

$$[\chi_C(x) - \chi_{C'}(x)][f_{\theta_1}(x) - kf_{\theta_0}(x)] \geqslant 0.$$

(If $x \in C$ then the second term $\geqslant 0$ and the first term is $1 - \chi_{C'}$, also nonnegative. Likewise if $x \notin C$, the second term < 0 and the first term $\leqslant 0$.) Therefore,

$$\begin{aligned} 0 &\leqslant \int_{\mathbb{R}^n} [\chi_C(x) - \chi_{C'}(x)][f_{\theta_1}(x) - kf_{\theta_0}(x)]\,\mathrm{d}x \\ &= \mathbb{P}_{\theta_1}(X \in C) - \mathbb{P}_{\theta_1}(X \in C') - k[\mathbb{P}_{\theta_0}(X \in C) - \mathbb{P}_{\theta_0}(X \in C')] \\ &= \beta(\theta_1) - \beta'(\theta_1) - k[\beta(\theta_0) - \beta'(\theta_0)]. \end{aligned} \tag{3}$$

By definition of $\mathcal{T}$, the significance level of the test corresponding to C' is $\leqslant \alpha$, so $\beta(\theta_0) - \beta'(\theta_0) \geqslant 0$. (3) therefore implies $\beta(\theta_1) - \beta'(\theta_1) \geqslant 0$, i.e., the C test is UMP class $\mathcal{T}$.

For necessity, we now show that if C' is UMP class $\mathcal{T}$ then C' corresponds to a likelihood ratio test. Since the previous part implies C must be UMP too, $\beta(\theta_1) = \beta'(\theta_1)$. Therefore (3) implies

$$0 - k[\beta(\theta_0) - \beta'(\theta_0)] \geqslant 0 \implies \beta(\theta_0) - \beta'(\theta_0) = \alpha - \beta'(\theta_0) \leqslant 0 \implies \alpha \leqslant \beta'(\theta_0).$$

Since C' is UMP class $\mathcal{T}$, by assumption its significance level is (again) $\leqslant \alpha$, so $\beta'(\theta_0) \leqslant \alpha$. Thus we must have $\beta'(\theta_0) = \alpha$.

Now we have $\beta(\theta_1) = \beta(\theta_1)$ and $\beta(\theta_0) = \beta'(\theta_0)$. Hence (3) is zero:

$$\int_{\mathbb{R}^n} [\chi_C(x) - \chi_{C'}(x)][f_{\theta_1}(x) - kf_{\theta_0}(x)]\,\mathrm{d}x = 0.$$

Since the integrand is nonnegative, this implies that it is nonnegative almost everywhere. □

Example: (5.12). Suppose X is binomial with parameters 2 and $\theta \in \{1/2, 3/4\}$. Let H_0 be $\theta = 1/2$ and let H_1 be $\theta = 3/4$. The lemma says we simply need a likelihood ratio test to determine the UMP among tests with an upper bound on significance level. Note that X only takes three values:

$$\frac{f_{3/4}(0)}{f_{1/2}(0)} = \frac{1}{4} \qquad \frac{f_{3/4}(1)}{f_{1/2}(1)} = \frac{3}{4} \qquad \frac{f_{3/4}(2)}{f_{1/2}(2)} = \frac{9}{4}.$$

Thus,

(1) If $3/4 < k \leqslant 9/4$, then H_0 is rejected if and only if $X = 2$, and this test is the unique UMP for tests with significance level $\leqslant \mathbb{P}_{1/2}(X = 2) = 1/4$.

(2) If $1/4 < k < 3/4$, then H_0 is rejected if and only if $X \in \{1, 2\}$, and this test is the unique UMP for tests with significance level $\leqslant \mathbb{P}_{1/2}(X \in \{1, 2\}) = 3/4$.

(3) If $0 < k \leqslant 1/4$, then the likelihood ratio test always lands in C so H_0 is always rejected. This test is the unique UMP for tests with significance level $\mathbb{P}_{1/2}(X \in \{0, 1, 2\}) = 1$.

(4) If $k > 9/4$, thent he likelihood ratio test never lands in C, so H_0 is never rejected. This test is the unique UMP for tests with significance level at most $\mathbb{P}_{1/2}(X \in \varnothing) = 0$.

Beginning of Oct.20, 2021

2.2 Hypothesis Tests & Confidence Intervals

Definition: (5.14) Confidence Interval, Confidence Region

Let $X : \Omega \to \mathbb{R}^n$ be a random variable with distribution $f_\theta \in \{f_\theta : \theta \in \Theta\}$. Let $g : \Theta \to \mathbb{R}$. Let $u, v : \mathbb{R}^n \to \mathbb{R}$ be such that $u(x) \leqslant g(x)$ for all $x \in \mathbb{R}^n$. A **$100(1-\alpha)\%$ confidence interval** for a parameter $g(\theta)$ is a random variable of form $[u(X), v(X)]$ satisfying

$$\mathbb{P}_\theta(g(\theta)) \in [u(X), v(X)] \geqslant 1 - \alpha \qquad \text{for all } \theta \in \Theta.$$

More generally, if $c : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ (power set), then a **$100(1-\alpha)\%$ confidence region** for $g(\theta)$ is a random set $c(X)$ satisfying

$$\mathbb{P}_\theta(g(\theta) \in c(X)) \geqslant 1 - \alpha \qquad \text{for all } \theta \in \Theta.$$

Example: (5.15) CLT and confidence intervals. Let $X_1, ..., X_n$ be i.i.d. with values in $[0,1]$, known $\sigma^2 \in (0,1)$, but unknown $\mu \in [0,1]$. Let $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then $\mathbb{E}X = \mu$ and $\operatorname{var}(X) = \sigma^2/n$. By Berry-Esséen (CLT with error bound),

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \right) - \mathbb{P}(Z < t) \right| \leqslant \frac{1}{\sigma^3\sqrt{n}}.$$

If we take $t = 2$ and $t = -2$ separately and subtract the results,

$$\left| \mathbb{P}\left(-2 < \frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} < 2 \right) - \mathbb{P}(-2 < Z < 2) \right| \leqslant \frac{2}{\sigma^3\sqrt{n}}.$$

That is,

$$\mathbb{P}\left(\overline{X} - \frac{2\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{2\sigma}{\sqrt{n}} \right) \geqslant \underbrace{\mathbb{P}(-2 < Z < 2)}_{\approx 0.95} - \frac{2}{\sigma^3\sqrt{n}}.$$

In our notation,

$$\mathbb{P}_\mu(\mu \in [u(X), v(X)]) \geqslant 0.95 - \frac{2}{\sigma^3\sqrt{n}}$$

where

$$u(X) = \overline{X} - \frac{2\sigma}{\sqrt{n}} \qquad v(X) = \overline{X} - \frac{2\sigma}{\sqrt{n}}.$$

Theorem: (5.16) Confidence Region / Hypothesis Test Duality

Heuristically, we have a hypothesis test if and only if we have a confidence region. Let $X : \Omega \to \mathbb{R}^n$.

(1) Fix $\alpha \in (0,1)$. Assume that for every $\theta_0 \in \Theta$, there is a hypothesis test with significance level α of hypothesis H_0 $\{\theta = \theta_0\}$. Let $C(\theta_0)$ denote the rejection region of the test. Then

$$c(X) := \{\theta \in \Theta : X \notin C(\theta)\}$$

is a $100(1-\alpha)\%$ confidence region for θ.

(2) Let $c : \mathbb{R}^n \to 2^\Theta$. Assume that $c(X)$ is a $100(1-\alpha)\%$ confidence region for θ. Define a hypothesis test with rejection region

$$C(\theta) := \{x \in \mathbb{R}^n : \theta \notin c(x)\}.$$

Then this test has significance level at most α.

Beginning of Oct.22, 2021

Proof.

(1) Since H_0 corresponds to $\{\theta = \theta_0\}$, the significance level is easy to compute:

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = \mathbb{P}_\theta(X \in C(\theta_0)).$$

Therefore by the definition $c(X) = \{\theta \in \Theta : X \notin C(\theta)\}$, we have

$$\mathbb{P}_\theta(\theta \in c(X)) = \mathbb{P}_\theta(X \notin C(\theta)) = 1 - \mathbb{P}_\theta(X \in C(\theta)) = 1 - \alpha \tag{Δ}$$

when $\theta = \theta_0$. Since by assumption (Δ) holds for every $\theta_0 \in \Theta$, this is indeed a $100(1-\alpha)\%$ confidence region, as claimed.

(2) The assumption that $c(X)$ is a $100(1-\alpha)\%$ confidence region for θ says

$$1 - \alpha \leqslant \mathbb{P}_\theta(\theta \in c(X)).$$

By the definition of $C(\theta)$, for any $\theta \in \Theta$,

$$1 - \alpha \leqslant \mathbb{P}_\theta(\theta \in c(X)) = \mathbb{P}_\theta(X \notin C(\theta)) = 1 - \mathbb{P}_\theta(X \in C(\theta)).$$

Therefore $\mathbb{P}_\theta(X \in C(\theta)) \leqslant \alpha$. Therefore, in particular

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = \mathbb{P}_{\theta_0}(X \in C(\theta_0)) \leqslant \alpha.$$

In other words this test has significance level $\leqslant \alpha$.

□

2.3 p-value

Stated informally, p-value is *a measure of belief of rejecting null hypothesis*. A small p-value corresponds to a high probability that the null hypothesis is false.

Definition: (5.17) p-value

Let $X_1, ..., X_n$ be real-valued random sample with $f_\theta \in \{f_\theta : \theta \in \Theta\}$. Define $X := (X_1, ..., X_n)$ for convenience. Let $Y := t(X)$ where $t : \mathbb{R}^n \to \mathbb{R}$. For all $c \in \mathbb{R}$, consider the hypothesis test with rejection region $\{x \in \mathbb{R}^n : t(x) \geqslant c\}$. Let $p : \mathbb{R}^n \to [0,1]$ be defined by

$$p(x) := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(t(X) \geqslant t(x)) \qquad \text{for all } x \in \mathbb{R}^n.$$

The ***p*-value** for the hypothesis test is defined to be the statistic $p(X)$.

Remark: (5.18). Fix $c \in \mathbb{R}$. By our definition of C, $\beta(\theta) = \mathbb{P}_\theta(X \in C) = \mathbb{P}_\theta(t(X) \geqslant c)$. Thus the significance level is

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \in \Theta} \mathbb{P}_\theta(t(X) \geqslant c).$$

This is very similar to the definition of $p(x)$ — $p(x)$ is equal to α where $c = t(x)$.

Also notice that as c increases, for each θ, $\mathbb{P}_\theta(t(X) \geqslant c)$ gets smaller, so the supremum gets smaller. Thus as c increases, α decreases. We say $p(x)$ is the smallest significance level such that the hypothesis test rejects the null hypothesis.

Remark: (5.19). Let $Y = t(X)$ be continuous. Fix $\theta \in \Theta$. For $c \in \mathbb{R}$, define $F_{-Y}(c) := \mathbb{P}(-Y \leqslant c)$, and for all $x \in \mathbb{R}^n$, denote

$$g_\theta(x) := \mathbb{P}_\theta(t(X) \geqslant t(x)) = \mathbb{P}_\theta(-t(X) \leqslant -t(x)) = F_{-Y}(-t(x)).$$

Then

$$\begin{aligned} g_\theta(X) &= F_{-Y}(-Y) = \mathbb{P}_\theta(F_{-Y}(-Y) \leqslant c) = \mathbb{P}_\theta(-Y \leqslant F_{-Y}^{-1}(c)) \\ &= F_{-Y}(F_{-Y}^{-1}(c)) = c. \end{aligned}$$

Therefore by definition of $p(x)$, since $p(x) = \sup \mathbb{P}_\theta(t(X) \geqslant t(x)) = \sup g_\theta(x)$, we have $p(X) \geqslant g_\theta(X)$ and so

$$\mathbb{P}_\theta(p(X) \leqslant c) \leqslant \mathbb{P}_\theta(g_\theta(X) \leqslant c) = c.$$

Therefore, when $\theta \in \Theta_0$, probability of $p(X)$ being small is small. For example $p(X) \leqslant 0.05$ has probability $\leqslant 0.05$. In other words, with probability $\geqslant 0.95$, θ is not supposed to be in Θ_0, i.e., the null hypothesis is false. *This explains why small p-value suggests a rejection of the null hypothesis.*

Example: (5.20). Let X be binomial, $n = 5$, and $\theta \in [0,1]$ unknown. Let H_0 be $\{\theta = 1/2\}$ and H_1 be $\{\theta \in [0,1] : \theta \neq 1/2\}$. For $c \in \mathbb{R}$, let the rejection regions be of form $C := \{x \in \mathbb{N} : x \geqslant c\}$.
First suppose $X = 2$. Then

$$p(2) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(X \geqslant 2) = \mathbb{P}_{1/2}(X \geqslant 2) = 1 - \mathbb{P}_{1/2}(X \leqslant 1) = 0.8125,$$

which suggests we are not at all confident in rejecting H_0.

Alternatively, suppose $X = 4$. Then the corresponding p-value is

$$p(4) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(X \geqslant 4) = \mathbb{P}_{1/2}(X \geqslant 4) = 0.1875,$$

in which case we are more confident in rejecting H_0.

Beginning of Oct.25, 2021

2.4 Generalized Likelihood Ratio Tests

Let $X_1, ..., X_n$ be i.i.d. from $f_\theta \in \{f_\theta : \theta \in \Theta\}$. In particular X_1 has distribution f_θ. The joint PDF is given by

$$\prod_{i=1}^{n} f_\theta(x_i).$$

If we have $x = (x_1, ..., x_n) \in \mathbb{R}^n$, then recall that the likelihood function $\ell : \Theta \to [0, \infty)$ is defined to be

$$\ell(\theta) := f_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i).$$

Also recall that Neyman-Pearson states that, when Θ consists of two points, then the likelihood ratio test is UMP among all tests with significance level $\leqslant \alpha$. The rejection region was defined as

$$C := \{x \in \mathbb{R}^n : f_{\theta_1} I(x) > k f_{\theta_0}(x)\}.$$

When Θ contains more than two points, we want an analogue of the region defined above. In this case, we define

$$C := \{x \in \mathbb{R}^n : \sup_{\theta \in \Theta_1} f_\theta(x) \geqslant k \sup_{\theta \in \Theta_0} f_\theta(x)\}.$$

Definition: (5.22) Generalized Likelihood Ratio Test

Let $k > 1$. The **generalized likelihood ratio test** of a hypothesis H_0 that $\{\theta \in \Theta_0\}$ is defined by the following region:

$$C := \{x \in \mathbb{R}^n : \sup_{\theta \in \Theta} f_\theta(x) \geqslant k \sup_{\theta \in \Theta_0} f_\theta(x)\}.$$

Remark. If $0 < k \leqslant 1$ then obviously $\sup_{\theta \in \Theta} \geqslant \sup_{\theta \in \Theta_0}$, and so $C = \mathbb{R}^n$ entirely. The test becomes trivial, so we restrict it to $k > 1$.

Example: (5.24). Let $X_1, ..., X_n$ be i.i.d. Gaussians with known $\sigma^2 > 0$ but unknown $\mu \in \mathbb{R}$. Fix $\mu_0 \in \mathbb{R}$. Suppose H_0 is $\mu = \mu_0$ and H_1 is $\mu \neq \mu_0$. Hence $\Theta = \mathbb{R}, \Theta_0 = \{\mu_0\}$, and $\Theta_1 = \mathbb{R} - \{\mu_0\}$. Also, for $x = (x_1, ..., x_n) \in \mathbb{R}^n$,

$$f_\mu(x) = f_\mu(x_1, ..., x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Our null region contains μ_0 only so $\sup_{\mu \in \Theta_0} f_\mu(x) = f_{\mu_0}(x)$. Also recall that the MLE in this case is the sample

mean. That is,

$$\sup_{\mu\in\Theta} f_\mu(x) = f_{\overline{\mu}}(x)$$

where $\overline{\mu} = (x_1 + \ldots + x_n)/n$. Therefore,

$$C = \{x \in \mathbb{R}^n : f_{\overline{\mu}}(x) \geqslant k f_{\mu_0}(x)\}.$$

That is,

$$\begin{aligned}
C &= \left\{x \in \mathbb{R}^n : \prod_{i=1}^n \exp\left(\frac{-(x_i-\overline{x})^2 + (x_i-\mu_0)^2}{2\sigma^2}\right) \geqslant k\right\} \\
&= \left\{x \in \mathbb{R}^n : \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n((x_i-\overline{x})^2 - (x_i-\mu_0)^2)\right) \geqslant k\right\} \\
&= \left\{x \in \mathbb{R}^n : \sum_{i=1}^n\left[(x_i - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2\right] \leqslant -2\sigma^2\log k\right\} \\
&= \left\{x \in \mathbb{R}^n : -n(\frac{1}{n}\sum_{j=1}^n x_j - \mu_0)^2 \leqslant -2\sigma^2\log k\right\} \\
&= \left\{x \in \mathbb{R}^n : \left|\frac{1}{n}\sum_{j=1}^n x_j - \mu_0\right| \geqslant \sqrt{2\sigma^2\log k/n}\right\}.
\end{aligned}$$

Intuitively, the rejection region consists of points where the sample mean is far from μ_0.

Beginning of Oct.27, 2021

Remark. Recall that the sample mean is a sufficient statistic for μ, so C is a function of a sufficient statistic. This is reasonable since a sufficient statistic has "sufficient information" to estimate μ.

Remark. Denote $X = (X_1, \ldots, X_n)$. If H_0 is true, then

$$2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} = \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^n(X_i-\mu_0)\right)^2 = \left(\frac{1}{\sigma\sqrt{n}}\sum_{j=1}^n(X_i-\mu_0)\right)^2,$$

whose distribution is the square of a standard normal, i.e., a χ_1^2. In general, even if X_i's are not Gaussian, as $n \to \infty$, the quotient above still converges to a χ_1^2.

Remark. In the Gaussian case, the p-value is

$$p(x) := \mathbb{P}_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu_0\right| \geqslant \left|\frac{1}{n}\sum_{i=1}^n x_i - \mu_0\right|\right).$$

2.5 Case Study: Alpha Particle Emissions

The following table counts particle emissions of americium 241, Am-241. During 1207 disjoint intervals of 10 seconds each, a number m of alpha particle emissions were observed.

m	0, 1 or 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	≥ 17
# of Intervals	18	28	56	105	126	146	164	161	123	101	74	53	23	15	9	5

Question. What is a good model of understanding what's going on?

We claim that the number of particle emissions in each of the intervals can be modelled as 1207 i.i.d. Poisson distributions with unknown mean $\lambda > 0$. This is sensible since we are observing "low probability events" but repeated many times.

Recall

$$\mathbb{P}_\lambda(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

Since the average number of alpha particle emitted is 8.392, our naive guess is $\lambda \approx 8.392$. (Note this is also the MLE.)

Beginning of Oct.29, 2021

For $j \geqslant 0$, let $q_j := \mathbb{P}(\#\text{emission} = j)$ so that $\sum_{j=0}^{\infty} q_i = 1$.

Let H_0 be that the emission follows a Poisson distribution, i.e., $q_j = e^{-\lambda}\lambda^j/j!$ for some $\lambda > 0$, and let the alternative hypothesis H_1 be that the emission does not follow a Poisson distribution.

We now consider a multinomial distribution with 1207 trials of rolling a 16-sided die (there are 16 columns in the table above) and use this to model the probabilities of the counts of appearances of each side. Hence, we consider the random variables $X_1, ..., X_{16}$ defined by

$$f_\theta(x) := f_\theta(x_1, ..., x_{16}) := \mathbb{P}(X_i = x_i \forall i) = 1207! \prod_{j=1}^{16} \frac{q_j^{x_j}}{x_j!}$$

subject to

$$x_j \in \mathbb{Z}^+, 1 \leqslant j \leqslant 16, \sum_{j=1}^{16} x_j = 1207$$

and $q_i = q_i(\theta)$, depending on θ.

($q_i(\lambda)$ denotes q_i depending on θ.) We will compute the GLR test, i.e., we compute

$$\frac{\sup_{\theta\in\Theta} f_\theta(x)}{\sup_{\theta\in\Theta_0} f_\theta(x)}.$$

For the numerator $\sup_{\theta\in\Theta} f_\theta(x)$, we want to maximize $1207! \prod_{j=1}^{16} \frac{q_j^{x_j}}{x_j!}$ above all $\{q_1, ..., q_{16}\}$ subject to $q_j \geqslant 0$ and $\sum_{j=1}^{16} q_j = 1$.

Note that

$$\frac{\partial}{\partial q_i} f_\theta(x) = 1207! \prod_{j\neq i} \frac{q_j^{x_j}}{x_j!} \cdot x_i q_i^{x_i - 1} = \frac{x_i}{q_i} f_\theta(x).$$

Using Lagrange multipliers, the constraint function $\sum_{j=1}^{16} q_i$ has partial 1 for all components; that is, to maximize f_θ, we just need the partials of f_θ to have the same number for each components. That is, the extrema is obtained when $x_1 : x_2 : ... : x_{16} = q_1 : q_2 : ... : q_{16}$, or

$$\frac{q_1}{x_1} = \frac{q_2}{x_2} = ... = \frac{q_{16}}{x_{16}} \qquad \text{and} \qquad \sum_{j=1}^{16} q_j = 1.$$

Since $\sum_{j=1}^{16} x_j = 1207$, this gives

$$q_i = \frac{x_i}{1207} \qquad 1 \leqslant i \leqslant 16.$$

Notice on the boundary of $\sum_{j=1}^{16} q_j = 1$, some $q_i = 0$, so the product $f_\theta(x)$ is zero. Also, for each θ, f_θ is continuous. We found only one critical point in the interior, so it must be a maximum!

Therefore,

$$\sup_{\theta \in \Theta} f_\theta(x) = 1207! \prod_{j=1}^{16} \frac{(x_j/1207)^{x_j}}{x_j!}. \tag{1}$$

Now we need to compute the supremum over $\theta \in \Theta_0$. In this case q_j's needs to satisfy $q_j = e^{-\lambda}\lambda^{-j}/j!$ for $1 < j < 16$. Note that q_1 and q_{16} are slightly different. The sumpremum over these is numerically computed to be $\lambda \approx 8.366$, pretty close to our original naive guess that $\lambda = 8.392$.

Beginning of Nov.1, 2021

The likelihood of the emission having a Poisson with parameter λ is the following.

$$\prod_{j=2}^{15} \frac{[e^{-\lambda}\lambda^{j+1}/(j+1)!]^{x_j}}{x_j!} \cdot \frac{(e^{-\lambda}[1+\lambda+\lambda^2/2])^{x_1}}{x_1!} \cdot \frac{[1-e^{-\lambda}\sum_{i=0}^{16}\lambda^i/i!]^{x_{16}}}{x_{16}!}$$

where the second term corresponds to the probability of having $0, 1$ or 2 emissions for a total of x_1 times and the third term represents the probability of having > 16 emissions for x_{16} times.

Therefore, the likelihood ratio $\sup_{\theta\in\Theta} f_\theta(x) / \sup_{\theta\in\Theta_0} f_\theta(x)$ is approximately

$$\left[\frac{x_1/1207}{e^{-8.37}(1+8.37+8.37^2/2)}\right]^{x_1} \left[\frac{x_{16}/1207}{e^{-8.37}\sum_{i=17}^{\infty} 8.37^i/i!}\right]^{x_{16}} \prod_{j=2}^{15}\left[\frac{x_i/1207}{e^{-8.37}8.37^{j+1}/(j+1)!}\right]^{x_j}. \tag{1}$$

We can find the asymptotic distribution of the GLR as $m = 1207$ tends to infinity.

How? If $X_1, ..., X_{16}$ are i.i.d. and $X := (X_1, ..., X_{16})$, then

$$2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} \text{ or equivalently } -2\log\frac{\sup_{\theta\in\Theta_0} f_\theta(X)}{\sup_{\theta\in\Theta} f_\theta(X)} \tag{*}$$

converges in distribution to χ^2_1 as $n = 16$ tends to ∞.

We also claim that as $m = 1207$ tends to ∞, (*) converges in distribution to a χ^2 with $16 - 1 - 1 = 14$ degrees of freedom. *If the distribution says that the observed (*) is unlikely, then we can reject the null hypothesis. We finally return to our original question — whether or not the original emission follows a Poisson distribution.*

Proof. For simplicity, let $p_j(\lambda)$ be the probability of the j^{th} entry of column occurring for a Poisson with parameter λ (e.g., $p_1(\lambda)$ corresponds to a Poisson with parameter λ evaluating to $0, 1$, or 2). Then

$$\begin{aligned}
2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} &= 2\log\prod_{i=1}^{16}\frac{(X_j/1207)^{X_j}}{(p_j(\lambda))^{X_j}} \\
&= 2\log\prod_{j=1}^{16}\left[\frac{X_j/1207}{p_j(\lambda)}\right]^{X_j} = 2\sum_{j=1}^{16} X_j \log\frac{X_j/1207}{p_j(\lambda)} \\
&= 2\cdot 1207\cdot\sum_{j=1}^{16}\frac{X_j}{1207}\log\left[\frac{X_j/1207}{p_j(\lambda)}\right].
\end{aligned}$$

If the emission resembles a Poisson, the quotient $\frac{X_j/1207}{p_j(\lambda)}$ should be close to 1. We let $a := X_j/1207$ and $b := p_j(\lambda)$. Doing Taylor expansion of $h(a) := a\log(a/b)$ around $a = b$ gives

$$h(a) = h(b) + h'(b)(b-a) + \frac{1}{2}h''(b)(b-a)^2 + \mathcal{O}((b-a)^3).$$

The first derivative at a is

$$\frac{\mathrm{d}}{\mathrm{d}a}(a\log(a/b)) = \left[\frac{a}{b} + \log(a/b)\right]_{a=b} = 1$$

At a, the second derivative is $0 + 1/a$ evaluated at b, i.e., $1/b$.

Therefore

$$h(a) \approx (a-b) + \frac{(a-b)^2}{2b}.$$

Putting this back,

$$\begin{aligned} 2\log(\ldots) &\approx 2\cdot 1207 \sum_{j=1}^{16}\left[\frac{X_j}{1207} - p_j(\lambda)\right] + 1207\sum_{j=1}^{16}\frac{(X_j/1207 - p_j(\lambda))^2}{p_j(\lambda)} \\ &= 2\cdot 1207\cdot\left[\sum_{i=1}^{16}\frac{X_j}{1207} - \sum_{j=1}^{16}p_j(\lambda)\right] + \ldots \\ &= \sum_{j=1}^{16}\frac{(X_j - 1207p_j(\lambda))^2}{1207p_j(\lambda)} = \sum_{j=1}^{16}\frac{(X_j - \mathbb{E}_\lambda X_j)^2}{\mathbb{E}_\lambda X_j}. \end{aligned}$$

This gives the **Pearson's chi-squared test statistic**! □

Beginning of Nov.3, 2021

For convenience call the last statistic S. With the corresponding $\lambda \approx 8.366$ and the observed x_i's, we have $S \approx 8.94$. We claim that S is approximately a chi-squared distribution with 14 degrees of freedom. First recall that $X_1, \ldots, X_{16}$ follow a multinomial distribution with $X_1 + \ldots + X_{16} = 1207$ and that each X_i is itself a binomial distribution. If $X_1, \ldots, X_{16}$ are independent, then

$$S = \sum_{j=1}^{16}\frac{(X_j - \mathbb{E}X_j)^2}{\mathbb{E}X_j} = \sum_{j=1}^{16}\left(\frac{X_j - \mathbb{E}X_j}{\sqrt{\mathbb{E}X_j}}\right)^2 \approx \text{sum of 16 independent Gaussians by CLT.}$$

We don't get a χ^2_{16} because $X_1, \ldots, X_{16}$ are not independent; heuristically, $X_{16} = 1207 - X_1 - \ldots - X_{15}$ and $\mathbb{E}X_{16} = 1207 - \mathbb{E}X_1 - \ldots - \mathbb{E}X_{15}$. This implies

$$(X_{16} - \mathbb{E}X_{16}) = \left(\sum_{i=1}^{15}(X_i - \mathbb{E}X_i)\right)^2$$

Therefore,

$$S = \sum_{j=1}^{15}\frac{(X_j - \mathbb{E}X_j)^2}{\mathbb{E}X_j} + \frac{(\sum_{i=1}^{15}(X_i - \mathbb{E}X_i))^2}{\mathbb{E}X_{16}}$$

and it (somehow) has a distribution of 15 independent standard squared Gaussians *if λ is fixed.* (For simplicity we denoted $\mathbb{E}_\lambda X_i$ by $\mathbb{E}X_i$.) However, λ is *not* fixed! We used the fixed λ from the MLE but λ itself should also be a function of $X_1, \ldots, X_{16}$. Therefore we lose another degree of freemdom, ending up having a chi-squared with 14 degrees of freedom.

Then the p-value is $\mathbb{P}(S \geqslant 8.94) \approx 0.835$ so the data *does probably* follow a Poisson distribution.

Chapter 3

Comparing Two Samples

3.1 Comparing Independent Gaussians

Suppose $X_1, ..., X_n$ are i.i.d. Gaussians with *unknown* $\mu_X \in \mathbb{R}$ and *known* $\sigma_X^2 > 0$. Suppose $Y_1, ..., Y_m$ are i.i.d. Gaussians with *unknown* $\mu_Y \in \mathbb{R}$ and *known* variance $\sigma_Y^2 > 0$. Also assume that X, Y are independent.

Goal. Give a confidence interval for $\mu_X - \mu_Y$.

Notice that

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j\right) - (\mu_X - \mu_Y)$$

has mean 0 and variance $\sigma_X^2/n + \sigma_Y^2/m$, so the above divided by $\sqrt{\sigma_X^2/n + \sigma_Y^2/m}$ gives a standard normal. (Recall that adding / subtracting independent Gaussians still result in a Gaussian, so this entire thing is obtained from a shifted Gaussian divided by some constant. This of course is also a Gaussian.) That is, has mean 0 and variance $\sigma_X^2/n + \sigma_Y^2/m$, so the above divided by $\sqrt{\sigma_X^2/n + \sigma_Y^2/m}$ gives a standard normal. (Recall that adding / subtracting independent Gaussians still result in a Gaussian, so this entire thing is obtained from a shifted Gaussian divided by some constant. This of course is also a Gaussian.) That is,

$$W := \frac{(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j) - \mu_X + \mu_Y}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim \mathcal{N}(0,1).$$

Therefore, $\mathbb{P}(-t \leqslant W \leqslant t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-s^2/2}\, \mathrm{d}s$, i.e., the expression below equals $\frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-s^2/2}\, \mathrm{d}s$:

$$\mathbb{P}\left(\overline{X} - \overline{Y} - t\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} < \mu_X - \mu_Y < \overline{X} - \overline{Y} + t\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right).$$

This gives a confidence interval for $\mu_X - \mu_Y$.

Beginning of Nov.8, 2021

3.2 Mann-Whitney Test

Let $m, n > 0$ be integers. Suppose we run an experiment on $m + n$ people, e.g., to cure a disease. Among all $m + n$ people, n of them are chosen uniformly at random to be in the control group, and the remaining m people are in the treatment. The null hypothesis is that the treatment has *no* effect on people.

Suppose we order the people by integer $1,...,m+n$ so that $1,...,n$ correspond to the control group and $n+1,...,n+m$ correspond to the greatment group.

Suppose the quality of outcome of the i^{th} person is $x_i \in [0,1]$ (high score good, low score bad).

Let us reorder $x_1,...,x_{m+n}$ by $x_{I_1} \leqslant x_{I_2} \leqslant ... \leqslant x_{I_{m+n}}$ where $\{I_i\} = \{1,...,m+n\}$.

We define the test statistic to be

$$Z := \sum_{j=1}^{m+n} j\chi_{\{1\leqslant I_j\leqslant n\}},$$

the sum of "ranks" of people in the control group. Ideally, if the null hypothesis is true (i.e., the treatment has no effect), then the ranks should be completely random, i.e., for each I_i, the assignment of values $\{1,...,m+n\}$ should all be equally likely. There are $\binom{m+n}{n}$ such assignments, each with probability $\binom{m+n}{n}^{-1}$. For $k>0$, write $c_{n,m,k}$ as the number of ways k can be written as a sum of n *distinct* integers among elements of $\{1,...,m+n\}$, disregarding order (because $\binom{m+n}{n}$ disregards it). Then

$$\mathbb{P}(Z=k) = \frac{c_{n,m,k}}{\binom{m+n}{n}}.$$

For small values of m,n, we can easily compute $\mathbb{P}(Z=k)$ explicitly, whereas for large m,n, CLT may help.

Observe that, given m,n, $3 \leqslant Z \leqslant 2(m+n)-1$. For exmaple:

Beginning of Nov.10, 2021

Example: (6.4). Suppose $n=2, m=3$. Then $3 \leqslant Z \leqslant 9$. The null hypothesis test is that the treatment has no effect on people. By inspection, it is easy to see that

$$\mathbb{P}(Z=k) = \begin{cases} 1/10 & k=3,4,8,9 \\ 1/5 & k=5,6,7. \end{cases}$$

Then $\mathbb{E}Z=6$. We reject the null when Z is close to 3 or 9 (i.e., when the treatment either has terrible bad effect or amazing effect). We consider the hypothesis that we reject when $|Z-6| \geqslant c$. If we observed that $Z=7$, then $|7-6|=1$, and

$$\mathbb{P}(|Z-6| \geqslant 1) = 1-\mathbb{P}(Z-6) = \frac{4}{5},$$

so we are not confident in rejecting the null. However, if we observed that $Z=9$, then

$$\mathbb{P}(|Z-6| \geqslant |9-6|) = \mathbb{P}(Z=3 \text{ or } 9) = \frac{1}{10}+\frac{1}{10} = \frac{1}{5},$$

in which case we are relatively more confident in rejecting the null.

Example 3.2.1. If $X_1,...,X_m, Y_1,...,Y_n$ are i.i.d. When m and n are large, we want to find an ultimate way to approximate Z. (Like before, we have m people in the treatment group and n in control group, and Z denotes the sum of ranks in the control group.) Then,

$$\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{X_i<Y_j} = \sum_{i=1}^{m}\sum_{j=1}^{n} 1_{X_{(i)}<Y_{(j)}},$$

as the sum simply rearranges things. Then, notice that, after fixing j, $\sum_{i=1}^{m} 1_{X_{(i)}<Y_{(j)}}$ denotes the rank of $Y_{(j)}-j$, i.e., number of $X_{(i)}$'s less than $Y_{(j)}$: among the first (rank of $Y_{(j)}$) ranks, j are from Y and the remaining

(rank of $Y_{(j)}$ minus j) must come from X's. Hence,

$$\sum_{j=1}^{n}\sum_{i=1}^{m} 1_{X_{(i)}<Y_{(j)}} = \sum_{j=1}^{n}(\text{rank of } Y_{(j)} - j).$$

Since $\sum_{j=1}^{n}(\text{rank of } Y_{(j)}) = Z$, we have

$$\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{X_i<Y_j} = Z - \frac{n(n+1)}{2}.$$

Under null hypothesis, $X_1,...,X_m,Y_1,...,Y_n$ are i.i.d., and if the distribution of X_1 is continuous then $\mathbb{E}1_{X_1<Y_1} = 1/2$. Therefore,

$$\mathbb{E}Z = \frac{m(m+1)}{2} + mn\mathbb{E}1_{X_1<Y_1} = \frac{mn}{2} + \frac{n(n+1)}{2} = \frac{n(m+n+1)}{2}.$$

To compute the variance, we first disregard the constant $n(n+1)/2$. Since $\operatorname{var} 1_{X_1<Y_1} = 1/4$, $\mathbb{E}1_{X_i<Y_j}1_{X_i<Y_k} = 1/3$ (needs X_i to be the smallest among all three), and $\mathbb{E}1_{X_i<Y_j}1_{X_j<Y_k} = (1/2)^2 = 1/4$, we have

$$\begin{aligned}
\operatorname{var}(Z) &= \operatorname{var}(\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{X_i<Y_j}) \\
&= \sum_{i.k=1}^{m}\sum_{j,\ell=1}^{n} \operatorname{cov}(1_{X_i<Y_j}, 1_{X_k<Y_\ell}) \\
&= \sum_{i=k}^{m}\sum_{j=\ell}^{n} + \sum_{i=k}^{m}\sum_{j\neq\ell}^{n} + \sum_{i\neq k}^{m}\sum_{j=\ell}^{n} + \sum_{i\neq k}^{m}\sum_{j\neq\ell}^{n} \\
&= \frac{mn}{4} + \frac{m(n^2-n)}{12} + \frac{n(m^2-m)}{12} + (m^2-m)(n^2-n)(0) \\
&= mn\frac{1+m+n}{12}.
\end{aligned}$$

Beginning of Nov.12, 2021

In summary, the following has mean zero and variance 1, assume H_0 is true:

$$\frac{Z - n(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}.$$

As $m,n \to \infty$, the above converges to a standard Gaussian random variable (so e.g. we can compute the p-value approximately).

3.3 Signed Rank Test

In this section we compare *dependent* samples.

Suppose $X_1,...,X_n$ are i.i.d., and $Z_1,...,Z_n$ are i.i.d., but X and Z are *not* necessarily independent. For example consider a medical study where X_i denotes the blood pressure of the i^{th} patient before treatment and Z_i the one after treatment. To check the efficacy of the treatment, we examine $Z_1 - X_1,...,Z_n - X_n$. and rank them

$$|Z_{I_1} - X_{I_1}| \leqslant |Z_{I_2} - X_{I_2}| \leqslant ... \leqslant |Z_{I_n} - X_{I-n}|.$$

We consider the statistic (assuming $Z_i - X_i \neq 0$)

$$W := \sum \text{ ranks of positive } X_i - Z_i = \sum_{i=1}^{n} \max((\text{rank of } Z_i - X_i) \cdot \operatorname{sgn}(Z_i - X_i), 0).$$

Let the null hypothesis be that the treatment has *no* effect. Under this, $Z_1 - X_1$ and $X_1 - Z_1$ should have the same distribution, so $\operatorname{sgn}(Z - X)$ has probability $1/2$ of being 1 and $1/2$ for -1.
Let $Y_1, \ldots, Y_n$ be i.i.d. uniformly distributed in $\{-1, 1\}$. Then

$$W = \sum_{i=1}^{n} \max(iY_i, 0).$$

We can use convolutions to explicitly compute W, since

$$\max(iY_i, 0) = \begin{cases} i & \text{with probability } 1/2 \\ 0 & \text{with probability } 1/2. \end{cases}$$

Since $\mathbb{E}W = n(n+1)/4$ and $\operatorname{var}(W) = n(n+1)(2n+1)/24 \approx n^3/12$,

$$\frac{W_n - n(n+1)/4}{\sqrt{n^3/12}}$$

converges to a standard Gaussian as $n \to \infty$ by *Lindeberg's CLT*.

Chapter 4

Analysis of Variance, ANOVA

Beginning of Nov.15, 2021

4.1 General Linear Model

Let A be $n \times m$ with known (deterministic) constants and let $\beta \in \mathbb{R}^m$ be an unknown vector of (deterministic) constants. Let $\epsilon \in \mathbb{R}^m$ be a random vector. Suppose our observation of data is the vector $Y \in \mathbb{R}^n$ given by $Y = A\beta + \epsilon$.
Goal. Try to estimate β when we only have Y and A.

Example: (7.1) One-Way ANOVA. Let $n_1, n_2, n_3 > 0$ be integers and let $n = n_1 + n_2 + n_3$. Let $\beta_1, \beta_2, \beta_3$ be unknown real numbers.
Define

$$A := \begin{bmatrix} 1_{n_1\times 1} & 0_{n_1\times 1} & 0_{n_1\times 1} \\ 0_{n_2\times 1} & 1_{n_2\times 1} & 0_{n_2\times 1} \\ 0_{n_3\times 1} & 0_{n_3\times 1} & 1_{n_3\times 1} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

Let $\sigma^2 > 0$ be fixed. Let $Y_1, ..., Y_n$ be independent random variable such that

(1) For $1 \leqslant i \leqslant n_1$, Y_i is a Gaussian with mean β_1 and variance σ^2;

(2) For $n + 1 \leqslant i \leqslant n_1 + n_2$, Y_i is a Gaussian with mean β_2 and variance σ^2; and

(3) For each $n_1 + n_2 + 1 \leqslant i \leqslant n$, Y_i is a Gaussian with mean β_3 and variance σ^2.

Finally, let $\epsilon \in \mathbb{R}^n$ be of i.i.d. Gaussians with mean 0 and variance σ^2. Let $Y = (Y_1, ..., Y_n)^T$ so that

$$Y = A\beta + \epsilon.$$

Question. How to estimate β_j's? Is it true that $\beta_1 = \beta_2 = \beta_3$?
More generally, we could have $n_1, ..., n_m$ and consider

$$A = \begin{bmatrix} 1_{n_1\times 1} & 0_{n_1\times 1} & \cdots & 0_{n_1\times 1} \\ 0_{n_2\times 1} & 1_{n_2\times 1} & \cdots & 0_{n_2\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_m\times 1} & 0_{n_m\times 1} & \cdots & 1_{n_m\times 1} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}.$$

For example, we could set $\beta_1 = \beta_2 = \beta_3$ to be our null hypothesis. Recall that we know to test the difference of Gaussians using the difference of sample means.

Example: (7.2) Linear Regression. Another example of the general linear model: suppose we have $\beta_1, \beta_2 \in \mathbb{R}$ unknown and we have $x_1, ..., x_n \in \mathbb{R}$ constant. Fix $\sigma^2 > 0$. Define

$$A := \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Let $\epsilon \in \mathbb{R}^n$ be the column vector consisting of i.i.d. Gussians with mean 0 and variance σ^2. Then $Y = A\beta + \epsilon$ says that for all i with $1 \leqslant i \leqslant n$,

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i.$$

Example: Another View of Linear Regression – Least Squares. Let $x_1, .., , x_n, y_1, ..., y_n \in \mathbb{R}$ be given. We want to find $\beta_1, \beta_2 \in \mathbb{R}$ minimizing

$$\sum_{i=1}^{n} (y_i - (\beta_1 + \beta_2 x_i))^2.$$

Beginning of Nov.17, 2021

Back to one-way ANOVA:

Recall that $Y_i = \beta_p + \epsilon_i$ for all $m_{p-1} + 1 \leqslant i \leqslant m_p$. Correspondingly, for each j, we define

$$\overline{Y}_j := \frac{1}{n_j} \sum_{i=m_{j-1}+1}^{m_j} Y_i,$$

the sample mean of the random variables that have mean β_j. Hence $\mathbb{E}\overline{Y}_j = \beta_j$. Previously, we said that the difference of two Gaussians can be transformed into a standard Gaussian:

$$\frac{(\overline{Y}_j - \overline{Y}_k) - (\beta_j - \beta_k)}{\sigma\sqrt{1/n_j + 1/n_k}} \sim \mathcal{N}(0, 1).$$

More generally, for any linear combination $\sum_{j=1}^{p} c_j \overline{Y}_j$, we have

$$\frac{\sum_{j=1}^{p} c_j \overline{Y}_j - \sum_{j=1}^{p} c_j \beta_j}{\sigma\sqrt{\sum_{j=1}^{p} c_j^2/n_j}} \sim \mathcal{N}(0, 1).$$

On the other hand, suppose further that σ^2 is also *unknown*. For each j, we define the j^{th} sample variance to be

$$S_j^2 := \frac{1}{n_j - 1} \sum_{i=m_{j-1}+1}^{m_j} (Y_i - \overline{Y}_j)^2.$$

Then the following has a student t's distribution with $n_j + n_k - 2$ degrees of freemdom (see Quiz 5 P4):

$$\frac{(\overline{Y}_j - \overline{Y}_k) - (\beta_j - \beta_k)}{S\sqrt{1/n_j + 1/n_k}}$$

where

$$S^2 = \frac{(n_j-1)S_j^2 + (n_k-1)S_k^2}{n_j+n_k-2}.$$

More generally, the following has student t's distribution with $-p+\sum_{j=1}^{p} n_j$ degrees of freedom:

$$\frac{\sum_{j=1}^p c_j\overline{Y}_j - \sum_{j=1}^p c_j\beta_j}{S\sqrt{\sum_{j=1}^p c_j^2/n_j}} \qquad \text{where} \qquad S^2 = \frac{\sum_{j=1}^p (n_j-1)S_j^2}{-p+\sum_{j=1}^p n_j}. \tag{*}$$

Upshot. We can get confidence intervals for $\sum_{j=1}^{p} c_j\beta_j$, regardless of whether we know σ^2.

Now we test our hypothesis that $\beta_1 = \beta_2 = \ldots = \beta_p$. Note that if we consider (*) with $\sum_{j=1}^{p} c_j = 0$, then

$$\sum_{j=1}^{p} c_j\beta_j = \beta_1 \sum_{j=1}^{p} c_j = 0,$$

assuming the hypothesis is true. Conversely, if for all combination $\sum_{j=1}^{p} c_j = 0$ we have $\sum_{j=1}^{p} c_j\beta_j = 0$, then $\beta_i = \beta_j$: indeed, letting $c_1 = 1, c_2 = -1, c_j = 0$ for all $j \geqslant 3$ implies $\beta_1 = \beta_2$, and likewise all β's are the same. Hence

$$\beta_1 = \beta_2 = \ldots = \beta_p \iff \sum_{j=1}^{p} c_j\beta_j = 0 \text{ for all } \{c_j\}_{j=1}^{p} \text{ with } \sum_{j=1}^{p} c_j = 0. \tag{1}$$

Proposition: (7.4) *F*-Test

We define

$$F := \sup_{\sum_{j=1}^p c_j=0} \frac{(\sum_{j=1}^p c_j\overline{Y}_j - \sum_{j=1}^p c_j\beta_j)^2}{S^2\sum_{j=1}^p c_j^2/n_j}.$$

Idea: if all β's are equal then $F = 0$. Also this statistic looks for the "worst" violation of $\beta_1 = \ldots = \beta_p$.

The supremum can be attained and Lagrange multipliers give an explicit formula:

$$F = S^{-2}\sum_{j=1}^{p} n_j[(\overline{Y}_j - \overline{Y}) - (\beta_i - \overline{\beta})]^2$$

where $\overline{Y} := (m_p)^{-1}\sum_{j=1}^{m_p} Y_j$, $m_p = \sum_{j=1}^{p} n_j$, and $\overline{\beta} = \mathbb{E}\overline{Y} = (m_p)^{-1}\sum_{j=1}^{m_p}\mathbb{E}Y_j = (m_p)^{-1}\sum_{j=1}^{p} n_p\beta_p$.

Moreover, $F/(p-1)$ has Snedecor's F-distribution with $p-1$ and $m_p - p$ degrees of freedom.

Beginning of Nov.19, 2021

Proof. Under the null hypotehsis $\beta_1 = \ldots = \beta_p$, $\beta_i = \overline{\beta}$, so

$$F = S^{-2}\sum_{j=1}^{p} n_j(\overline{Y}_j - \overline{Y})^2.$$

F can be found by minimizing the denominator, or just $\sum_{j=1}^{p} c_j^2/n_j$, subject to (1) the top being fixed, i.e., $\sum_{j=1}^{p} c_i(\overline{Y}_j -$

β_j) and (2) $\sum_{j=1}^{p} c_j = 0$.

Recall that if we were to minimize $h(c_1, ..., c_p)$ subject to $r(c_1, ..., c_p) = s(c_1, ..., c_p) = 0$, we need to solve

$$\nabla h = \lambda_1 + \nabla r + \lambda_2 \nabla s \qquad \text{for some } \lambda_1, \lambda_2 \in \mathbb{R}.$$

In other words,

$$2c_j/n_j = \lambda_1(\overline{Y}_j - \beta_j) + \lambda_2 \qquad \text{for all } 1 \leqslant j \leqslant p.$$

Using lemma 7.6 in notes, we can find a minimum at which

$$\frac{(\sum_{j=1}^{p} c_j \overline{Y}_j - \sum_{j=1}^{p} c_j \beta - j)^2}{\sum_{j=1}^{p} c_j^2/m_j} = \sum_{j=1}^{p} n_j((\overline{Y}_j - \overline{Y}) - (\beta_i - \overline{\beta})).$$

Since $\sum_{j=1}^{p} c_j^2/n_j$ is strictly convex and we found a minimum (indeed x^2 only has a global min), it must be the unique global minimum. □

4.2 Linear Regression

Example: (7.7). Suppose we are presented with data $(x_1, y_1), ..., (x_n, y_n)$. We want to find a line $mx + b$ that fits the data "best". Among various ways to define the "wellness", a standard one is to minimize the **least squares**, i.e., to find m, b minimizing the following:

$$f(m, b) = \sum_{i=1}^{n} (y_i - (mx_i + b))^2.$$

Since this function is strictly convex, any critical point must be the global minimum. In this case,

$$m = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{x})}{\sum_{j=1}^{n}(x_i - \overline{x})^2} \qquad \text{and} \qquad b = \overline{y} - m\overline{x}.$$

Beginning of Nov.22, 2021

Alternate Presentation of Linear Regression

(7.2 revisited) Let $x_1, ..., x_n \in \mathbb{R}$. Let $\sigma^2 > 0$. Let $\beta_1, \beta_2 \in \mathbb{R}$ unknown. Let $\epsilon_1, ..., \epsilon_n$ be i.i.d. Gaussians with mean zero and variance $\sigma^2 > 0$.

We want to find β_1, β_2 so that $Y_i = \beta_1 + \beta_2 x_i + \epsilon_i$ for all i; that is, $Y_i - [\beta_1 + \beta_2 x_i] = \epsilon_i$. We consider estimators that are linear combinations of Y_i's, i.e., estimators of form $\sum_{i=1}^{n} c_i Y_i$. Goal: find unbiased estimators for β_1, β_2.

Claim. The two versions are equivalent.

Theorem: (7.8)

Let $c_1, ..., c_n \in \mathbb{R}$ be such that $\sum_{i=1}^n c_i Y_i$ is unbiased of β_2. Then

$$\text{var}(\sum_{i=1}^n c_i Y_i) \leqslant \text{var}(\sum_{i=1}^n c_i' Y_i)$$

for any other combination coefficients $c_1', ..., c_n' \in \mathbb{R}$. Furthermore,

$$\sum_{i=1}^n c_i Y_i = \frac{\sum_{i=1}^n (Y_i - \sum_{j=1}^n Y_j/n)(x_i - \sum_{j=1}^n x_j/n)}{\sum_{k=1}^n (x_k - \sum_{j=1}^n x_j/n)^2}. \tag{1}$$

Similarly, if $\sum_{i=1}^n c_i Y_i$ is unbiased for β_1, then

$$\text{var}(\sum_{i=1}^n c_i Y_i) \leqslant \text{var}(\sum_{i=1}^n c_i' Y_i)$$

and

$$\sum_{i=1}^n c_i Y_i = \overline{Y} - (1) \cdot \overline{x}.$$

Proof. (First statement.) First note that $\mathbb{E}(\sum_{i=1}^n c_i Y_i) = \sum_{i=1}^n (c_i(\beta_1 + \beta_2 x_i + \epsilon_i)) = \sum_{i=1}^n c_i(\beta_1 + \beta_2 x_i)$. By assumption this equals β_2, so

$$\sum_{i=1}^n c_i = 0 \qquad \sum_{i=1}^n c_i x_i = 1. \tag{1}$$

On the other hand $\text{var}(\sum_{i=1}^n c_i Y_i) = \sum_{i=1}^n c_i^2 \,\text{var}(Y_i) = \sigma^2 \sum_{i=1}^n c_i^2$ since $\text{var}(Y_i) = \text{var}(\beta_1 + \beta_2 x_i + \epsilon_i) = \text{var}(\epsilon_i) = \sigma^2$.

Thus, we'd like to minimize $\sigma^2 \sum_{i=1}^n c_i^2$ subject to (1). Using the lemma on Lagrange multiplier, this is mi minimized when

$$c_i = \frac{x_i - \overline{x}}{\sum_{j=1}^n (x_j - \overline{x})^2},$$

so

$$\sum_{i=1}^n c_i Y_i = \frac{\sum_{i=1}^n Y_i (x_i - \overline{x})}{\sum_{k=1}^n (x_k - \overline{x})^2}.$$

Since $\sum_{i=1}^n \overline{Y}(x_i - \overline{x}) = \overline{Y} n(\overline{x} - \overline{x}) = 0$, this can be re-written as

$$\sum_{i=1}^n c_i Y_i = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{k=1}^n (x_k - \overline{x})^2}.$$

The second statement has an analogous proof, except that now the constraints are swapped. □

4.3 Logistic Regression

We denote the **logistic function** as

$$h(x) := \frac{1}{1 + e^{-x}} \qquad \text{for all } x \in \mathbb{R}.$$

Note that $\lim_{x \to \infty} h(x) = 1$ and $\lim_{x \to -\infty} h(x) = 0$.

We use logistic regression to classify data into two bins, e.g., classifying emails into spam or "not spam" or determining if a turkey is cooked or uncooked, based on a threshold of $h(x)$.

Let $X_1, ..., X_n$ be i.i.d. and let $g : \mathbb{R} \to \{0, 1\}$ be an unknown function. Let $Y_i := g(X_i)$. (Example: assume we've never seen a turkey before; X_i = temperature of i^{th} turkey; $g(X_i) = 1$ if cooked; and $g(X_i) = 0$ if uncooked.)

Beginning of Nov.29, 2021

Note that $Y_1, ..., Y_n$ are i.i.d. Bernoulli, so there is some $p \in [0, 1]$ known such that $p = \mathbb{P}(Y_1 = 1)$. Assume that there exist $a, b \in \mathbb{R}$ such that $p \approx h(ax + b) \approx g(x)$. Then the likelihood function is

$$\ell(a, b) = \prod_{i=1}^{n} p^{y_i}(1 - p)^{1-y_i} \approx \prod_{i=1}^{n} [h(ax_i + b)]^{y_i} [1 - h(ax_i + b)]^{1-y_i}.$$

The best candidates for a, b might be given by the MLE.