Ref: [P. Olofsson] Probability, Statistics, and Stochastic Processes.

Chap 6. Statistical Inference.

6.1 Intro.

Statistical inference (or statistics) is a field which ties together probability models and data collection.

6.2 Pt Estimators.

Def

The collection $\{x_1, \cdots, x_n\}$ of observed values of iid random variables is called a (random) sample.

Def

A r.v. $\hat{\theta}$ is an estimator of an unknown parameter θ if it is a function of a random sample and is used to estimate θ.

The observed value of $\hat{\theta}$ is called an estimate of θ.

Def

An estimator $\hat{\theta}$ is called unbiased if $E[\hat{\theta}] = \theta$. Otherwise, it is called biased.

Def

An estimator $\hat{\theta}_n$ based on the sample $X_1, \cdots, X_n$ is called consistent if $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$.

Prop 6.1

$\hat{\theta}_n$: an unbiased estimator based on $X_1, \cdots, X_n$.

If $Var[\hat{\theta}_n] \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent.

<Pf>

By Chebyshev. #

Rmk:

① Unbiasedness and Consistency are desirable properties to check when looking at estimators.

② Q: Given unbiased consistent $\hat{\theta}_n$ and $\tilde{\theta}_n$, which should we choose?

Def

$\hat{\theta}, \tilde{\theta}$: unbiased estimators of θ.

$\hat{\theta}$ is said to be more efficient than $\tilde{\theta}$ if

$$Var[\hat{\theta}] < Var[\tilde{\theta}].$$

Prop 6.2

$X_1, \cdots, X_n$: sample w/ mean μ and variance σ^2.

$\bar{X}$: sample mean.

Then $E[\bar{X}] = \mu$ and $Var[\bar{X}] = \sigma^2/n$.

Rmk:

Given two estimators $\hat{\theta}$ and $\tilde{\theta}$, it is not always possible to find out the more efficient one since their variances may dep. on the unknown θ.

Q:

How do we know whether an estimator is optimally good?

Prop 6.3 (Cramér-Rao Lower Bound)

$\hat{\theta}_n$: an unbiased estimator of the parameter θ based on the sample $X_1, \cdots, X_n$.

Then

$$Var[\hat{\theta}_n] \geq \frac{1}{nI(\theta)}, \text{ where}$$

$$I(\theta) := -E\left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X)\right].$$

$I(\theta)$ is called the Fisher information.

<The proof of Prop 6.3 is beyond the scope of this book.>

Def

The efficiency of an unbiased estimator $\hat{\theta}_n$ is

$$e(\hat{\theta}_n) := \frac{1}{nI(\theta)Var[\hat{\theta}_n]}.$$

Rmk:

By Prop 6.3, $e(\hat{\theta}_n) \leq 1$.

Thus, if we can find an unbiased estimator $\hat{\theta}_n$ w/ $e(\hat{\theta}_n) = 1$ (or, equivalently, $Var[\hat{\theta}_n] = \frac{1}{nI(\theta)}$), then it is an optimal estimator among all unbiased estimators.

Def

$\hat{\Theta}$: an estimator.

The standard deviation of $\hat{\Theta}$, i.e.

$\boxed{\sigma_{\hat{\Theta}}} = \sqrt{Var[\hat{\Theta}]}$, is called the $\boxed{\text{standard error}}$

The $\boxed{\text{estimated standard error}}$ of $\hat{\Theta}$ is the sample standard deviation (i.e. the one computable from the observed values of the sample). (defined below).

Def

$X_1, \cdots, X_n$: random sample.

The $\boxed{\text{sample variance}}$ is defined as

$$\boxed{S^2} = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X})^2.$$

$\boxed{\text{Cor 6.1}}$ (Help compute S^2).

$$S^2 = \frac{1}{n-1} \left(\sum_{k=1}^{n} X_k^2 - n\bar{X}^2 \right).$$

The reason for dividing by $n-1$ instead of n is below:

$\boxed{\text{Prop 6.4}}$

The sample variance S^2 is an unbiased estimator of σ^2. Moreover, if $E[X_i^4] < \infty$, then it is also a consistent estimator of σ^2.

<Pf>

1° Rmk: For a r.v. X, w/ mean μ and variance σ^2, we have $E[X^2] = \mu^2 + \sigma^2$.

$$E[S^2] = \frac{1}{n-1} \left[\sum_{k=1}^{n} E[X_k^2] - n \cdot E[\bar{X}^2] \right]$$

$$= \frac{1}{n-1} \left[n \cdot (\mu^2 + \sigma^2) - n \cdot (\mu^2 + \frac{\sigma^2}{n}) \right] = \sigma^2.$$

Thus S^2 is unbiased. #

2° To prove S^2 is consistent, use Chebyshev. #

Def

The square root of S^2, denoted $\boxed{S}$, is called the $\boxed{\text{sample standard deviation}}$.

$\boxed{\text{Prop}}$

S is NOT an unbiased estimator of σ.

<Pf>

$\because$ S is random $\therefore$ $Var[S] > 0$.

Note $Var[S] = E[S^2] - E[S]^2 = \sigma^2 - E[S]^2$.

$\Rightarrow$ $E[S] = \sqrt{\sigma^2 - Var[S]} < \sigma$. Thus S is biased. #

6.3 Confidence Intervals.

Idea: It's desirable to supplement an estimator w/ an error bound.

Def

$X_1, \cdots, X_n$: random sample.

θ : unknown parameter.

T_1, T_2 : two functions of the sample s.t.

$$P(T_1 \le \theta \le T_2) = q.$$

Then we say that the interval $[T_1, T_2]$ is a $\boxed{\text{confidence interval}}$ for θ w/ confidence level q and write $\boxed{T_1 \le \theta \le T_2 \ (q)}$.

i.e. A confidence interval is a random interval containing θ w/ prob. q.

Def

Once the values of $X_1, \cdots, X_n$ are observed, T_1 and T_2 can be computed, and $[T_1, T_2]$ is then called an $\boxed{\text{observed confidence interval}}$.

Rmk:

Confidence interval can be viewed as an estimator of an interval while the estimator introduced before is one for an unknown value. Thus, $\boxed{\text{point estimation}}$ and $\boxed{\text{interval estimation}}$ are used for distinguishing these two.

Procedures:

1. Determine the dist. of an estimator $\hat{\Theta}$.

Its dist. involves the unknown θ.

2. Choose a confidence level q

3. Use 1. to obtain T_1, T_2 s.t.

$$P(T_1 \le \theta \le T_2) = q.$$

Often, $[T_1, T_2]$ is of the form $[\hat{\Theta} - R, \hat{\Theta} + R]$.

In this case, we write $\boxed{\theta = \hat{\Theta} \pm R \ (q)}$

Rmk: Standard values of q: 0.90, 0.95, 0.99, etc.

6.3.1. Confidence Interval for the Mean in the Normal Dist. w/ Known Variance.

$\boxed{\text{Prop 6.5}}$

$X_1, \cdots, X_n$: a sample from $N(\mu, \sigma^2)$ w/ σ^2 : known.

Then a $(100q)\%$ confidence interval for μ is

$$\mu = \bar{X} \pm z\frac{\sigma}{\sqrt{n}} \quad (q)$$

where z is such that $\Phi(z) = \frac{1+q}{2}$.

6.3.2 Confidence Interval for an Unknown Probability.

p : probability of an event A.

Repeat experiment n times and observe A in X times of these n times.

Then $X \sim \text{bin}(n, p)$.

$\hat{p} := X/n$.

By Central Limit Theorem, $\hat{p} \overset{d}{\approx} N\left(p, \frac{p(1-p)}{n}\right)$.

$\boxed{\text{Prop 6.6}}$

An approximate $(100q)\%$ confidence interval for p is

$$p = \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (\approx q),$$

where $\Phi(z) = \frac{1+q}{2}$.

Rmk: $\pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is called the $\boxed{\text{margin of error}}$ or the $\boxed{\text{sampling error}}$

6.3.3 One-Sided Confidence Intervals.

Def

① A confidence interval of the form $T_1 \le \theta \le T_2$ or $\theta = \hat{\theta} \pm R$ is called $\boxed{\text{two-sided}}$.

② That of the form $\theta \le \hat{\theta} + R$ or $\theta \ge \hat{\theta} - R$ is called $\boxed{\text{one-sided}}$.

6.4 Estimation Methods.

Q: How to find an estimator?

6.4.1 The Method of Moments.

Def

X : r.v.

The $\boxed{\text{rth moment}}$ of X is

$$\boxed{\mu_r} := E[X^r].$$

Def

$X_1, \cdots, X_n$: a random sample.

The $\boxed{\text{rth sample moment}}$ is

$$\boxed{\hat{\mu}_r} := \frac{1}{n}\sum_{k=1}^{n} X_k^r.$$

Prop

$E[\hat{\mu}_r] = \mu_r$ and $\text{Var}[\hat{\mu}_r] = \frac{1}{n}\text{Var}[X^r]$.

Def

Suppose the unknown parameter θ can be expressed as a function of the first j moments, say $\theta = g(\mu_1, \cdots, \mu_j)$. The estimator $\hat{\theta} = g(\hat{\mu}_1, \cdots, \hat{\mu}_j)$ is then called the $\boxed{\text{moment estimator}}$ of θ.

$\boxed{\text{Procedure: (Method of Moments)}}$

Start by computing μ_1, then $\mu_2, \cdots$, until some μ_j so that $\theta = g(\mu_1, \cdots, \mu_j)$ for some function g.

Rmk:

① Moment estimators may be biased.

② No guarantee-type theorem stated in the book. But, clearly, if g is a cont. fun, then $\hat{\theta}$ is guaranteed to be consistent.

6.4.2 Maximum Likelihood.

Motivation: Look at Example 6.13.

Roughly speaking, we want to choose the parameter θ making the observed sample most likely.

Def

$X_1, \cdots, X_n$: a random sample from a distributⁿ having pmf or pdf f_θ.

The fun.

$$\boxed{L(\theta)} = \prod_{k=1}^{n} f_\theta(X_k)$$

is called the likelihood function.

$\hat{\theta} := \underset{\theta}{\operatorname{argmax}}\, L(\theta)$ is called the

maximum-likelihood estimator (MLE) of θ.

Def (Motivation: Product is harder to deal with).

$\boxed{\ell(\theta)}$

$:= \log(L(\theta))$, is called the log-likelihood function.

Prop 6.7 (Quite Obvious!!)

If $\hat{\theta}$: MLE of θ and g is 1-1, then $g(\hat{\theta})$ is MLE of $g(\theta)$.

Rmk:

MLE is of central importance in statistics. Moreover, it is optimal asymptotically in the following sense:

Prop 6.8

$\hat{\theta}_n$: MLE of θ based on the sample $X_1, \cdots, X_n$.

Assume the Fisher Information $I(\theta)$ exists.

Then, as $n \to \infty$,

(i) $E[\hat{\theta}_n] \to \theta$. (asymptotically unbiased)

(ii) $\hat{\theta}_n$: consistent.

(iii) $e(\hat{\theta}_n) \to 1$. (asymptotically efficient)

(iv) $\sqrt{nI(\theta)}\,(\hat{\theta}_n - \theta) \xrightarrow{d} N(0,1)$.

$\left(\hat{\theta}_n \overset{d}{\approx} N\left(\theta, \frac{1}{nI(\theta)}\right)\right)$.

6.4.3 Evaluation of Estimators w/ Simulation.

Motivation: Estimators derived using Method of moments or maximum-likelihood can have complicated expressions as functions of the sample.

Solution:

Use simulation to evaluate such estimators.

6.4.4. Bootstrap Simulation.

Motivation:

Simulation method in 6.4.3 can be used only when we have full knowledge of the underlying distribution of a sample. (i.e. even though θ is unknown, we still know the form of the dist. for any given θ).

Sol'n:

Instead of the true distribution, we use:

Def

$x_1, \cdots, x_n$: observed sample.

The distribution (CDF)

$$\boxed{\hat{F}_n(x)} = \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \le x\}}$$

is called the empirical distribution function.

Rmk:

① If n is large, then $\hat{F}_n(x)$ approximates the true distribution $F(x)$ reasonably well.

② Using $\hat{F}_n(x)$ to do simulation is called bootstrap simulation.

③ We can also use bootstrap simulation to obtain confidence intervals. (approximately)

④ The simulation in 6.4.3 can be used to achieve any precision in the estimate while the bootstrap simulation is limited by the size of the sample in hand.

6.5 Hypothesis Testing.

Idea: We want to test whether a hypothesis should be "rejected". (Philosophy: Reject a hypothesis if we have "sufficiently high" probability indicating its invalidity.)

Procedure (Hypothesis Testing) (Motivation Version)

① Make an assumption. (hypothesis)

② If a "very unlikely" outcome is observed, decide that the assumption is false.

③ Realize that there is a small risk we are wrong.

Def

θ: unknown parameter.

Want to decide whether θ equals some θ_0.

Thus, formulate the **null hypothesis** that $\theta = \theta_0$, denoted $\boxed{H_0 : \theta = \theta_0}$.

In conjunction, we also have an **alternative hypothesis**, denoted $\boxed{H_A}$.

If we formulate $H_A: \theta > \theta_0$ (or $H_A: \theta < \theta_0$), then such H_A is called **one-sided**.

A **two-sided** alternative hypothesis is $H_A: \theta \neq \theta_0$.

Def

A **test statistic** $\boxed{T}$ is a function of the sample. (to be used to test H_0).

The **significance level** $\boxed{\alpha}$ and **critical region** $\boxed{C}$ are determined such that $P(T \in C) = \alpha$, under the assumption H_0 is true.

If $T \in C$, **reject** H_0 in favor of H_A.

Otherwise, we say that we **accept** H_0.

Rmk:

① The significance level α is the risk we are willing to take to reject a hypothesis that is in fact true.

② "Accepting" a hypothesis does NOT mean we prove it. It is only that the data do not support a rejection.

③ It is often the alternative hypothesis that we want to prove.

④ (Philosophy) It is easier to falsify a hypothesis than to prove it.

Procedure: (Hypothesis Testing). (data, α not involved)

① State H_0 and H_A.

② Find T and decide for what values it rejects H_0 in favor of H_A. (i.e. the distribution of T under H_0) (under H_0.)

③ Choose α and find C s.t. $P(T \in C) = \alpha$ under H_0. (use the result of ②) (α involved)

④ Compute T. (data involved) If $T \in C$, reject H_0 in favor of H_A; otherwise accept H_0.

Rmk:

Hypothesis testing is the same as computing confidence interval.

6.5.1 Large Sample Tests.

Prop 6.10.

$\hat{\theta}$: MLE of an unknown θ based on a large sample $X_1, \cdots, X_n$.

Wish to test $H_0: \theta = \theta_0$ versus $H_A: \theta \neq \theta_0$.

The test statistic is

$$Z = \sqrt{n I(\theta_0)}\,(\hat{\theta} - \theta_0)$$

where $I(\theta_0)$: Fisher information at θ_0.

Then we reject H_0 on level $\approx \alpha$ if

$$|Z| \geq c$$

where $\Phi(c) = 1 - \alpha/2$.

<Pf>

Apply Prop 6.8 (iv). #

6.5.2 Test for an Unknown Probability.

Prop 6.11

p: unknown probability, estimated by $\hat{p}$, the relative frequency based on n indep. trials. (large)

Wish to test $H_0: p = p_0$ versus $H_A: p \neq p_0$.

Test statistic: $T = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$.

Rmk: If "normal" doesn't work well, use "binomial".

Then we reject H_0 on level $\approx \alpha$ if $|T| \geq c$ where $\Phi(c) = 1 - \alpha/2$.

<Pf> CLT. #

6.6 Further Topics in Hypothesis Testing.

6.6.1 P-Values.

Def

Given a data set. dep. on the observed data

The P-value of a test is the lowest significance level on which we can reject H_0.

i.e. $\min\{\alpha \mid \alpha$ is a significance level on which H_0 is rejected$\}$.

Rmk:

① If p is the P-value, then we reject on level α if $\alpha \geq p$. The smaller P-value is, the more confident we reject H_0.

② The P-value can be viewed as the "minimum risk" we are taking when rejecting H_0.

6.6.2. Data Snooping.

Data snooping: to first look at the data and then formulate and test the hypothesis.

Rmk:

For a hypothesis test to be meaningful, we should avoid data snooping.

6.6.3 The Power of a Test.

Def Given a hypothesis test.

Type I error = Pr[a true null hypothesis is rejected]

Type II error = Pr[a false null hypothesis is NOT rejected]

Rmk:

Type I error is exactly the significance level.

Def

Suppose we are testing the null hypo. $H_0: \theta = \theta_0$.

Fix a significance level α.

The function does not dep. on the observed data, but we need to fix the data size.

$$g(\theta) := \Pr[\text{reject } H_0 \text{ if the true para. is } \theta]$$

is called the power function of the test.

Rmk: (procedure of computing $g(\theta)$).

① Under H_0, find C s.t. $P_{H_0}(T \in C) = \alpha$.

② For θ, compute P_θ.

③ $g(\theta) = P_\theta(C)$.

Rmk:

$g(\theta_0) = \alpha$.

6.6.4. Multiple Hypothesis Testing.

① **Bonferroni correction:**

Motivating issue:

Suppose we have n (maybe dependent) hypothesis to test. If we set each significance level to α, then the total significance level (risk) could be (if all of them are indep.) $1-(1-\alpha)^n$, which may be large.

Sol'n:

Let $A_i = \{\text{true } H_0^{(i)} \text{ but rejected}\}^c$.

Fix a const. significance $\underbrace{\text{level}}_{\alpha}$ for test i. (i.e. $P(A_i) = 1-\alpha$)

Let α_m (m stands for "multiple".) be the significance level. Suppose we want to achieve $\alpha_m \ngtr \alpha'$.

Then

$$\alpha_m = P(\{\exists\, i \text{ s.t. true } H_0^{(i)} \text{ but rejected}\})$$

$$= P\Big(\bigcup_i A_i^c\Big) \leq \sum_i P(A_i^c) = n\alpha.$$

Thus, making $\alpha \leq \alpha'/n$, we achieve the desired α'. #

This is Bonferroni correction.

Disadvantage:

α'/n may be too small to reject any $H_0^{(i)}$.

② **Bonferroni-Holm correction:**

Procedure:

1° Calculate p-value of each single test and order them as $p_{(1)} \leq \cdots \leq p_{(n)}$.

2° For $i = 1, 2, \cdots, n$.

If $p_{(i)} \leq \alpha'/(n-i+1)$, reject $H_0^{(i)}$

Else, break and accept $H_0^{(i)}, H_0^{(i+1)}, \cdots, H_0^{(n)}$.

3° The result obtained in this procedure is called the Bonferroni-Holm correction and is guaranteed to satisfy

$$\alpha_m = P(\{\exists\, i \text{ s.t. true } H_0^{(i)} \text{ being rejected}\}) \leq \alpha'.$$

<Pf> Google Bonferroni-Holm. #

6.7 Goodness of Fit.

Q: How do we test whether our data do come from a specific distribution?

Prop 6.12

Suppose, in an experiment, an observation can fall into any of the r different categories, w/ prob. $p_1, \cdots, p_r$.

Let the experiment be repeated n times (iid) and $\boxed{X_k} := \#(\text{observations falling into category } k)$.

Then

$$\underbrace{\sum_{k=1}^{r} \frac{(X_k - np_k)^2}{np_k}}_{(*)} \overset{d}{\approx} \chi^2_{r-1}$$

where $\boxed{\chi^2_{r-1}}$ is the chi-square dist. w/ $r-1$ degrees of freedom.

Def

(*) above is often denoted by $\boxed{\chi^2}$ and written as

$$\chi^2 = \sum_{k=1}^{r} \frac{(O_k - E_k)^2}{E_k}, \text{ where } \text{"O"} = \text{"observed"}, \ \text{"E"} = \text{"expected"}.$$

Rmk:

① $(X_1, \cdots, X_r)$: multinoulli w/ para. $(n, p_1, \cdots, p_r)$.

② Rule of thumb: For Prop 6.12 to hold, we need $E_k \geq 5, \forall k$.

If not satisfied, clump categories together.

Cor 6.2

To test

H_0: the dist. is $(p_1, \cdots, p_r)$

against H_0 is not true, use the test statistic χ^2 and reject H_0 on level α if

$$\chi^2 \geq x$$

where $F_{\chi^2_{r-1}}(x) = 1 - \alpha$.

Rmk:

The above is an example of the so-called **goodness-of-fit test**.

Q: How do we test whether our data come from a certain type of distribution?

Prop 6.13

Under the same assumption of Prop 6.12 but now the prob. dep. on an unknown para. θ.

i.e. $p_1(\theta), \cdots, p_r(\theta)$.

$\hat{\theta} :=$ MLE of θ

Assume $p_1(\theta), \cdots, p_r(\theta)$ diff.

Then

$$\boxed{\sum_{k=1}^{r} \frac{(X_k - np_k(\hat{\theta}))^2}{np_k(\hat{\theta})} \overset{d}{\approx} \chi^2_{r-2}},$$

a chi-square dist. w/ $r-2$ degrees of freedom.

Rmk:

① (generalization of Prop 6.13) If there are j unknown parameters, then the result is χ^2_{r-j-1}, instead.

② Actually, we can replace MLE by any estimators satisfying certain asymptotic properties.

③ The two tests above can also apply to cont. dist. The approach is to divide the values into r regions, compute the prob. on each region, and go on as before.

6.7.1 Goodness-of-Fit Test for Independence.

A, B : events.

H_0: A and B are indep.

$p := P(A)$, $q := P(B)$.

Then

Category	$A \cap B$	$A \cap B^c$	$A^c \cap B$	$A^c \cap B^c$
Prob.	pq	$p(1-q)$	$(1-p)q$	$(1-p)(1-q)$

Apply Prop 6.13 w/ $j = 2$ (see Rmk ①); we can do hypothesis testing for indep. of A and B.

For $A_1, \cdots, A_{n_1}$, $B_1, \cdots, B_{n_2}$ w/

H_0: A_i and B_j indep. $\forall i, j$.

The same approach applies. The degree of freedom for chi-square is $n_1 \cdot n_2 - [(n_1 - 1) + (n_2 - 1)] - 1$
$= (n_1 - 1)(n_2 - 1)$.

6.7.2 Fisher's Exact Test.

Specifically designed for the case when lacking sufficient observations to achieve $n\hat{p}_i\hat{q}_j \geq 5, \forall i,j$.

Works for only testing H_0: A and B are indep.

	B	B^c
A	X_{11}	X_{12}
A^c	X_{21}	X_{22}

← This is called a contingency table.

Idea:

$$P(X_{11}=x_{11}) = \frac{\binom{x_{11}+x_{12}}{x_{11}}\binom{n-x_{11}-x_{12}}{x_{21}}}{\binom{n}{x_{11}+x_{21}}}. \leftarrow \boxed{\text{hypergeometric distribution.}}$$

Use this property to perform hypothesis testing. This method is called Fisher's Exact Test.

6.8 Bayesian Statistics.

Idea: Instead of viewing an unknown parameter as an unknown const., in Bayesian statistics, an unknown parameter is viewed as a random variable.

Procedure:

① Assign the unknown parameter a distribution that describe how likely we think different parameter values are.

② After gathering data, update the distribution of the parameter to the conditional distribution given the data.

Rmk:

The methods developed when assuming unknown parameters as unknown constants are often called frequentists statistics or classical statistics.

Def

The distribution we assign on the unknown parameter before gathering data is called the prior distribution (or prior, for short).

The conditional distribution obtained via the gathered data is called the posterior distribution.

Prop

θ : prob. dist.

D : data gathered.

Then $$P(\theta|D) = \frac{P(D|\theta)\cdot P(\theta)}{P(D)}$$

Rmk:

① $P(\theta)$: prior distribution.

② $P(\theta|D)$: posterior.

③ $P(D|\theta)$: prob. of the data if the parameter value is θ.

④ $P(D)$: unconditional prob. of the data.

Def

θ : parameter. D : data.

The posterior mean $E[\theta|D]$ is called a Bayes estimator of θ.

Rmk:

① When the data D is in the form of r.v., we called it estimator while if their values have been observed, we call it an estimate.

② For f : pdf or pmf, update f via
$$f(\theta|D) = \frac{P(D|\theta)\,f(\theta)}{P(D)}$$

Def

$f(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$, $0 \leq x \leq 1$, is called the beta distribution w/ nonnegative parameters α and β, where

$B(\alpha,\beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$ is called the beta function.

Rmk:

① For $\alpha,\beta \in \mathbb{N}_0$, $B(\alpha,\beta) = \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!}$.

② If $\alpha=\beta$, then it is symmetric around $1/2$. For $\alpha=\beta=1$, it is unif. on $[0,1]$.

③

Given $X \sim B(\alpha,\beta)$. Then

$$E[X] = \frac{\alpha}{\alpha+\beta}, \quad Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

④ Since Unif [0,1] stands for a certain type of noninformative prior, $B(\alpha,\beta)$ is also a popular choice of prior.

⑤ Note that $P(D)$ is a const. and hence

$$f(\theta|D) \propto P(D|\theta) f(\theta)$$

i.e. Posterior $\propto$ likelihood $\cdot$ prior.

and $\frac{1}{P(D)}$ is simply the const. making

$P(D|\theta)f(\theta)$ a prob. dist.

However, $\frac{1}{P(D)}$ is mostly challenging to compute and Markov chain Monte Carlo (MCMC) is often used to compute it numerically.

6.8.1 Noninformative Priors.

Here, two kinds are introduced.

① Uniform prior:

For cont. or discrete, for bdd or unbdd, use $f(\theta) \propto 1$.

② Invariant under transformation:

Use a prior as described.

Prop 6.15 (Jefferey's Prior).

$f(\theta) \propto \sqrt{I(\theta)}$, where

$$I(\theta) = -E\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right]$$

is the Fisher information.

Then $f(\theta)$ is invariant under parameter transformation.

6.8.2 Credibility Intervals.

This is the Bayesian analog of the confidence interval.

Def

A credibility interval of level q is an interval $[x_1, x_2]$ s.t. $P(x_1 \le \theta \le x_2 | D) = q$.

Rmk:

A more informative prior yields a narrower credibility interval.