# Statistics I Lecture Notes

Notes by Prof. Matthias Troffaes

Typeset in LaTeX by Tom Stoneham

## Contents

# Course Introduction

This course is delivered by Matthias Troffaes (matthias.troffaes@durham.ac.uk). Office hours are Mondays, from 08:40 - 10:40, in CM304.

Tutorials will occur once every two weeks, starting in week 12.

Problems classes occur once every two weeks in the Friday lecture spot, starting in week 14.

Homework will be set every Friday, and is to be handed in to your tutor by next Friday at 17:00.

We're working from a new course. Problem sheets and solutions are available on DUO, as are the lecture notes Matthias is working from. Not all past exam questions are going to be directly relevant to the course: it's mostly the more advanced questions from the problem sheets we should use for practice.

We will follow *Probability & Statistics*, 2012, 4ed, by DeGroot and Schervish, for most lectures, referred to in these notes as [DS12]. This is available as an e-book on DUO. For the first few lectures, we will follow *Applied Statistics & Probability for Engineers*, 2003, 3ed, by Montgomery & Runger, referred to as [MR03]. While there is not an e-book freely available, there are a number of copies available from the library. Furthermore, the library have actually scanned the first two chapters of [MR03], and these are available to read via DUO, under "Library Resources".

n.b: Only one person can "borrow" the DS12 ebook at a time, so email the library staff if you can't access it.

Throughout these notes, we will highlight certain information as follows:

**Definition: Introduction**

**Definitions** are in dark blue boxes, with the word being defined highlighted in bold.

**Related Questions**

References to related questions on the problem sheets are in light blue.

**Warning**

Any warnings/caveats related to a section's content are placed in red.

**Further Reading**

Further reading (from [DS12], [MR03]) will be placed in green.

# Formulae

This isn't in the original lecture notes, but I thought collecting up some of the useful formulae from throughout the notes might be a good idea. Please message/email me (tom@sto.neh.am) if there's something that should be included.

## Tables of Values

**Common Inverse Normal Values:**

| $1-\alpha$ | 0.80 | 0.90 | 0.95 | 0.98 | 0.99 |
|---|---|---|---|---|---|
| $z_{\alpha/2}$ | 1.28 | 1.64 | 1.96 | 2.33 | 2.58 |

## Quantiles for Chi-Square

| $n$ | 0.005 | 0.01 | 0.025 | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 | 0.95 | 0.975 | 0.99 | 0.995 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 0.00 | 0.00 | 0.02 | 0.10 | 0.45 | 1.32 | 2.71 | 3.84 | 5.02 | 6.63 | 7.88 |
| 2 | 0.01 | 0.02 | 0.05 | 0.21 | 0.58 | 1.39 | 2.77 | 4.61 | 5.99 | 7.38 | 9.21 | 10.60 |
| 3 | 0.07 | 0.11 | 0.22 | 0.58 | 1.21 | 2.37 | 4.11 | 6.25 | 7.81 | 9.35 | 11.34 | 12.84 |
| 4 | 0.21 | 0.30 | 0.48 | 1.06 | 1.92 | 3.36 | 5.39 | 7.78 | 9.49 | 11.14 | 13.28 | 14.86 |
| 5 | 0.41 | 0.55 | 0.83 | 1.61 | 2.67 | 4.35 | 6.63 | 9.24 | 11.07 | 12.83 | 15.09 | 16.75 |
| 6 | 0.68 | 0.87 | 1.24 | 2.20 | 3.45 | 5.35 | 7.84 | 10.64 | 12.59 | 14.45 | 16.81 | 18.55 |
| 7 | 0.99 | 1.24 | 1.69 | 2.83 | 4.25 | 6.35 | 9.04 | 12.02 | 14.07 | 16.01 | 18.48 | 20.28 |
| 8 | 1.34 | 1.65 | 2.18 | 3.49 | 5.07 | 7.34 | 10.22 | 13.36 | 15.51 | 17.53 | 20.09 | 21.95 |
| 9 | 1.73 | 2.09 | 2.70 | 4.17 | 5.90 | 8.34 | 11.39 | 14.68 | 16.92 | 19.02 | 21.67 | 23.59 |
| 10 | 2.16 | 2.56 | 3.25 | 4.87 | 6.74 | 9.34 | 12.55 | 15.99 | 18.31 | 20.48 | 23.21 | 25.19 |
| 11 | 2.60 | 3.05 | 3.82 | 5.58 | 7.58 | 10.34 | 13.70 | 17.28 | 19.68 | 21.92 | 24.72 | 26.76 |
| 12 | 3.07 | 3.57 | 4.40 | 6.30 | 8.44 | 11.34 | 14.85 | 18.55 | 21.03 | 23.34 | 26.22 | 28.30 |
| 13 | 3.57 | 4.11 | 5.01 | 7.04 | 9.30 | 12.34 | 15.98 | 19.81 | 22.36 | 24.74 | 27.69 | 29.82 |
| 14 | 4.07 | 4.66 | 5.63 | 7.79 | 10.17 | 13.34 | 17.12 | 21.06 | 23.68 | 26.12 | 29.14 | 31.32 |
| 15 | 4.60 | 5.23 | 6.26 | 8.55 | 11.04 | 14.34 | 18.25 | 22.31 | 25.00 | 27.49 | 30.58 | 32.80 |
| 16 | 5.14 | 5.81 | 6.91 | 9.31 | 11.91 | 15.34 | 19.37 | 23.54 | 26.30 | 28.85 | 32.00 | 34.27 |
| 17 | 5.70 | 6.41 | 7.56 | 10.09 | 12.79 | 16.34 | 20.49 | 24.77 | 27.59 | 30.19 | 33.41 | 35.72 |
| 18 | 6.26 | 7.01 | 8.23 | 10.86 | 13.68 | 17.34 | 21.60 | 25.99 | 28.87 | 31.53 | 34.81 | 37.16 |
| 19 | 6.84 | 7.63 | 8.91 | 11.65 | 14.56 | 18.34 | 22.72 | 27.20 | 30.14 | 32.85 | 36.19 | 38.58 |
| 20 | 7.43 | 8.26 | 9.59 | 12.44 | 15.45 | 19.34 | 23.83 | 28.41 | 31.41 | 34.17 | 37.57 | 40.00 |
| 21 | 8.03 | 8.90 | 10.28 | 13.24 | 16.34 | 20.34 | 24.93 | 29.62 | 32.67 | 35.48 | 38.93 | 41.40 |
| 22 | 8.64 | 9.54 | 10.98 | 14.04 | 17.24 | 21.34 | 26.04 | 30.81 | 33.92 | 36.78 | 40.29 | 42.80 |
| 23 | 9.26 | 10.20 | 11.69 | 14.85 | 18.14 | 22.34 | 27.14 | 32.01 | 35.17 | 38.08 | 41.64 | 44.18 |
| 24 | 9.89 | 10.86 | 12.40 | 15.66 | 19.04 | 23.34 | 28.24 | 33.20 | 36.42 | 39.36 | 42.98 | 45.56 |
| 25 | 10.52 | 11.52 | 13.12 | 16.47 | 19.94 | 24.34 | 29.34 | 34.38 | 37.65 | 40.65 | 44.31 | 46.93 |
| 26 | 11.16 | 12.20 | 13.84 | 17.29 | 20.84 | 25.34 | 30.43 | 35.56 | 38.89 | 41.92 | 45.64 | 48.29 |
| 27 | 11.81 | 12.88 | 14.57 | 18.11 | 21.75 | 26.34 | 31.53 | 36.74 | 40.11 | 43.19 | 46.96 | 49.64 |
| 28 | 12.46 | 13.56 | 15.31 | 18.94 | 22.66 | 27.34 | 32.62 | 37.92 | 41.34 | 44.46 | 48.28 | 50.99 |
| 29 | 13.12 | 14.26 | 16.05 | 19.77 | 23.57 | 28.34 | 33.71 | 39.09 | 42.56 | 45.72 | 49.59 | 52.34 |
| 30 | 13.79 | 14.95 | 16.79 | 20.60 | 24.48 | 29.34 | 34.80 | 40.26 | 43.77 | 46.98 | 50.89 | 53.67 |
| 35 | 17.19 | 18.51 | 20.57 | 24.80 | 29.05 | 34.34 | 40.22 | 46.06 | 49.80 | 53.20 | 57.34 | 60.27 |
| 40 | 20.71 | 22.16 | 24.43 | 29.05 | 33.66 | 39.34 | 45.62 | 51.81 | 55.76 | 59.34 | 63.69 | 66.77 |
| 45 | 24.31 | 25.90 | 28.37 | 33.35 | 38.29 | 44.34 | 50.98 | 57.51 | 61.66 | 65.41 | 69.96 | 73.17 |
| 50 | 27.99 | 29.71 | 32.36 | 37.69 | 42.94 | 49.33 | 56.33 | 63.17 | 67.50 | 71.42 | 76.15 | 79.49 |
| 55 | 31.73 | 33.57 | 36.40 | 42.06 | 47.61 | 54.33 | 61.66 | 68.80 | 73.31 | 77.38 | 82.29 | 85.75 |
| 60 | 35.53 | 37.48 | 40.48 | 46.46 | 52.29 | 59.33 | 66.98 | 74.40 | 79.08 | 83.30 | 88.38 | 91.95 |
| 70 | 43.28 | 45.44 | 48.76 | 55.33 | 61.70 | 69.33 | 77.58 | 85.53 | 90.53 | 95.02 | 100.43 | 104.21 |
| 80 | 51.17 | 53.54 | 57.15 | 64.28 | 71.14 | 79.33 | 88.13 | 96.58 | 101.88 | 106.63 | 112.33 | 116.32 |
| 90 | 59.20 | 61.75 | 65.65 | 73.29 | 80.62 | 89.33 | 98.65 | 107.57 | 113.15 | 118.14 | 124.12 | 128.30 |
| 100 | 67.33 | 70.06 | 74.22 | 82.36 | 90.13 | 99.33 | 109.14 | 118.50 | 124.34 | 129.56 | 135.81 | 140.17 |

## Distributions

**Binomial**

$$X \sim \text{Bin}(n,p) \iff p(x) = \binom{n}{x} p^x (1-p)^{n-x} \ \forall\ x \in \{0,1,...,n\}$$

$$\mathbb{E}(X) = np$$

$$\mathbb{V}\text{ar}(X) = np(1-p)$$

**Poisson**

$$X \sim \text{Po}(\lambda) \iff p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \ \forall\ x \in \{0,1,2,...\}$$

$$\mathbb{E}(X) = \lambda$$

$$\mathbb{V}\text{ar}(X) = \lambda$$

**Normal**

$$X \sim \mathcal{N}(\mu, \sigma^2) \iff f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \ \forall\ x \in \mathbb{R}$$

$$\mathbb{E}(X) = \mu$$

$$\mathbb{V}\text{ar}(X) = \sigma^2$$

**Exponential**

$$X \sim \text{Exp}(\lambda) \iff f(x) = \lambda e^{-\lambda x} \ \forall\ x \in (0,\infty)$$

$$\mathbb{E}(X) = \frac{1}{\lambda}$$

$$\mathbb{V}\text{ar}(X) = \frac{1}{\lambda^2}$$

**Beta**

$$X \sim \text{Beta}(\alpha,\beta) \iff f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \ \forall\ x \in [0,1]$$

$$\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$$

$$\mathbb{V}\text{ar}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

**Gamma**

$$X \sim \text{Gamma}(\alpha,\beta) \iff f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \ \forall\ x \in (0,\infty)$$

$$\mathbb{E}(X) = \frac{\alpha}{\beta}$$

$$\mathbb{V}\text{ar}(X) = \frac{\alpha}{\beta^2}$$

We may scale a random variable $Y \sim \text{Gamma}(\alpha,\beta)$ as follows:

$$aY \sim \text{Gamma}\left(\alpha, \frac{\beta}{\alpha}\right)$$

**Lomax**

$$X \sim \text{Lomax}(\alpha, \beta) \iff f(x) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}} \ \forall\ x \in (0, \infty)$$

$$\mathbb{E}(X) = \frac{\beta}{\alpha - 1}$$

$$\mathbb{V}\text{ar}(X) = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$$

**Chi-Square**

$$X \sim \chi^2(n) \iff X \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right) \quad \forall\, n > 0$$

$$\mathbb{E}(X) = n$$

$$\mathbb{V}\text{ar}(X) = 2n$$

If $X_i \sim \chi^2(n_i)$ are independent, then:

$$\sum_{i=1}^{k} X_i \sim \chi^2\left(\sum_{i=1}^{k} n_i\right)$$

If $X \sim \mathcal{N}(0, 1)$ then:

$$X^2 \sim \chi^2(1)$$

and if $X_i \sim \mathcal{N}(0, 1)$ are IID, then:

$$\sum_{i=1}^{k} X_i^2 \sim \chi^2(k)$$

## Frequentist Approach

**Estimators:** For any estimator $\widehat{T}$ of $t(\Theta)$, we have:

- **Bias** $:= \mathbb{E}(\widehat{T}|\theta) - t(\theta)$
- **Standard error** $:= \sqrt{\mathbb{V}\text{ar}(\widehat{T}|\theta)}$
- **Mean square error** $:= \mathbb{E}((\widehat{T} - t(\theta))^2|\theta)$
- **Relation of Mean Square Error, Standard Error, and Basis**:

$$\text{mean square error} = (\text{standard error})^2 + (\text{bias})^2$$

$$\mathbb{E}((\widehat{T} - t(\theta))^2|\theta) = \mathbb{V}\text{ar}(\widehat{T}|\theta) + (\mathbb{E}(\widehat{T}|\theta) - t(\theta))^2$$

**Central Limit Theorem:** For an infinite sequences of random variables $X_1, X_2, ...$ IID conditional on some random variable $\Theta$, with the functions $\mu(\theta) := \mathbb{E}(X_i|\theta), \sigma^2(\theta) := \mathbb{V}\text{ar}(X_i|\theta)$, we have that the sample mean $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ will follow:

$$\lim_{n\to\infty} \mathbb{P}\left(Z_h = \frac{\overline{X}_n - \mu(\Theta)}{\sigma(\Theta)/\sqrt{n}} \leq z \middle| \theta\right) = \Phi(z)$$

where $\Phi$ is the cumulative distribution function of $\mathcal{N}(0, 1)$. This leads to the approximate result:

$$\overline{X}_n|\theta \sim \mathcal{N}\left(\mu(\theta), \frac{\sigma^2(\theta)}{n}\right)$$

**Confidence Interval** For $X_1, ..., X_n$ IID conditional on $\Theta$, with $\mathbb{E}(X_i|\theta) = \theta, \mathbb{V}\text{ar}(X_i|\theta) = \sigma^2$, the $100 \cdot (1-\alpha)\%$ confidence interval on $\Theta$ is given by:

$$\left[\overline{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

**Frequentist Prediction Interval** Assuming that $X_1, ..., X_n$ are IID conditional on $\theta$ with $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$, where $\sigma$ is independent of $\theta$, then the $100 \cdot (1-\alpha)\%$ frequentist prediction interval for $X_{n+1}$ is given by:

$$\left[\overline{x}_n - z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}, \overline{x}_n + z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}\right]$$

## Bayesian Approach

**Posterior Distribution** Assuming $X_1, ..., X_n$ are IID conditional on $\Theta$ with assumed pdf $f(x|\theta)$ and prior pdf for $\Theta$ given by $f(\theta)$. Then, $\Theta$ has posterior pdf given by:

$$\begin{aligned} f(\theta|x_1, ..., x_n) &= \frac{f(\theta)\prod_{i=1}^{n} f(x_i|\theta)}{\int f(\theta')\prod_{i=1}^{n} f(x_i|\theta')d\theta'} \\ &\overset{(\theta)}{\propto} f(\theta) \cdot \prod_{i=1}^{n} f(x_i|\theta) \end{aligned}$$

**Likelihood**

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

**Credible Interval** For $X_1, ..., X_n$ IID conditional on $\theta$, with $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ and posterior distribution $\Theta|x_1, ..., x_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, we have that the $100 \cdot (1-\alpha)\%$ credible interval for the parameter $\theta$ is given by:

$$\left[\mu_n - z_{\alpha/2}\sigma_n, \mu_n + z_{\alpha/2}\sigma_n\right]$$

**General Posterior Distribution for Normal Sampling** Assume $X_i$ are IID conditional on $\Theta$, distributed according to $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ with prior distribution $\Theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, where $\sigma, \sigma_0, \mu_0$ are known constants. Then we have:

$$\begin{aligned} \mu_n &:= \frac{\mu_0/\sigma_0^2 + n\overline{x}_n/\sigma^2}{1/\sigma_0^2 + n/\sigma^2} \\ \sigma_n^2 &:= \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \\ \Theta|x_1, ..., x_n &\sim \mathcal{N}(\mu_n, \sigma_n^2) \end{aligned}$$

**Sequential Updating Theorem**

$$\begin{aligned} f(\theta|x_1, ..., x_{n+1}) &= \frac{f(\theta|x_1, ..., x_n)f(x_{n+1}|\theta)}{f(x_{n+1}|x_1, ..., x_n)} \\ &\overset{(\theta)}{\propto} f(\theta|x_1, ..., x_n)f(x_{n+1}|\theta) \end{aligned}$$

**Posterior Prediction**

$$X_{n+1}|x_1, ..., x_n \sim \mathcal{N}(\mu_n, \sigma^2 + \sigma_n^2)$$

Assuming $X_1, ..., X_n$ are IID conditional on $\theta$, and $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ with posterior distribution $\Theta|x_1, ..., x_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, then the $100 \cdot (1-\alpha)\%$ posterior prediction interval is given by:

$$\left[\mu_n - z_{\alpha/2}\sqrt{\sigma^2 + \sigma_n^2}, \mu_n + z_{\alpha/2}\sqrt{\sigma^2 + \sigma_n^2}\right]$$

**Prior Prediction** Assuming $X_1, ..., X_n$ are IID conditional on $\theta$, and $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ with prior distribution $\Theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, then the $100 \cdot (1-\alpha)\%$ prior prediction interval is given by:

$$\left[\mu_0 - z_{\alpha/2}\sqrt{\sigma^2 + \sigma_0^2}, \mu_0 + z_{\alpha/2}\sqrt{\sigma^2 + \sigma_0^2}\right]$$

**Bernoulli Distribution with Beta Prior** If $X_i|\theta \sim \text{Bernoulli}(\theta)$ IID conditional on $\Theta$ with prior $\Theta \sim \text{Beta}(\alpha_0, \beta_0)$ for $\alpha_0, \beta_0 > 0$, then:

$$\Theta|x_1, ..., x_n \sim \text{Beta}\left(\alpha_0 + \sum_{i=1}^{n} x_i, \beta_0 + n - \sum_{i=1}^{n} x_i\right)$$

**Exponential Distribution with Gamma Prior** If $X_i|\theta \sim \text{Exp}(\theta)$ IID conditional on $\Theta$ with prior $\Theta \sim \text{Gamma}(\alpha_0, \beta_0)$, we have the posterior:

$$\Theta|x_1, ..., x_n \sim \text{Gamma}\left(\alpha_0 + n, \beta_0 + \sum_{i=1}^{n} x_i\right)$$

And the posterior predictive:

$$X_{n+1}|x_1, ..., x_n \sim \text{Lomax}\left(\alpha_0 + n, \beta_0 + \sum_{i=1}^{n} x_i\right)$$

## Bayes Estimators

**Standard Loss Functions**

- **Squared Error Loss**

$$L(\theta, \widehat{\theta}) := (\theta - \widehat{\theta})^2$$

- **Absolute Error Loss**

$$L(\theta, \widehat{\theta}) := |\theta - \widehat{\theta}|$$

**Posterior Expected Loss**

$$\mathbb{E}(L(\Theta, \widehat{\theta})|x_1, ..., x_n) := \int_{-\infty}^{\infty} L(\theta, \widehat{\theta}) f(\theta|x_1, ..., x_n) d\theta$$

**Bayes Estimate in Squared Error Loss** Bayes estimate for $\Theta$ is given by posterior expectation of $\Theta$:

$$\delta(x_1, ..., x_n) = \mathbb{E}(\Theta|x_1, ..., x_n)$$

**Bayes Estimate in Absolute Error Loss** Bayes estimate for $\Theta$ is posterior median of $\Theta$:

$$\mathbb{P}(\Theta \leq \delta(x_1, ..., x_n)|x_1, ..., x_n) = \frac{1}{2}$$

## Maximum Likelihood Estimators

In general, the maximum likelihood of a parameter $\Theta$ is given by:

$$\widehat{\Theta} = \arg\max_{\theta} f(X_1, ..., X_n|\theta)$$

where $\arg\max_{\theta} g(\theta) :=$ the value of $\theta$ for which $g(\theta)$ is maximised.

**Log Likelihood**

$$L(\theta) := \log f(x_1, ..., x_n|\theta)$$

**MLE in Exponential Sampling** For $X_i|\theta \sim \text{Exp}(\theta)$ IID conditional on $\theta$, with $n$ observations, we have the likelihood:

$$f(x_1, ..., x_n) = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

and the log likelihood:

$$L(\theta) = n \log \theta - \theta \sum_{i=1}^{n} x_i$$

So we have the MLE:

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i}$$

**MLE in Bernoulli Sampling** For $X_i|\theta \sim \text{Bernoulli}(\theta)$ IID conditional on $\theta$, with $n$ observations, we have the likelihood:

$$f(x_1, ..., x_n|\theta) = \theta^y (1-\theta)^{n-y}$$

and the log likelihood:

$$L(\theta) = y \log \theta + (n-y) \log(1-\theta)$$

where:

$$y := \sum_{i=1}^{n} x_i$$

So we have the MLE:

$$\widehat{\theta} = \frac{y}{n}$$

**MLE in Normal Sampling with Unknown Mean, Known Variance** For $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ IID conditional on $\theta$ with unknown $\theta$, known $\sigma^2$, we have the log likelihood:

$$L(\theta) \overset{(\theta)}{\propto} -\frac{1}{2}\frac{n}{\sigma^2}(\theta - \overline{x}_n)^2$$

So we have the MLE:

$$\widehat{\theta} = \overline{x}_n$$

**MLE in Normal Sampling with Unknown Mean, Unknown Variance** For $X_i|\theta \sim \mathcal{N}(\mu, \sigma^2)$ IID conditional on $\theta$ with unknown $\theta, \sigma^2$, we have the likelihood:

$$f(x_1, ..., x_n|\mu, \tau) \overset{(\mu,\tau)}{\propto} \tau^{\frac{n}{2}} \exp\left(-\frac{1}{2}\tau \sum_{i=1}^{n}(x_i - \mu)^2\right)$$

and the log likelihood:

$$L(\mu, \tau) = \frac{n}{2}(\log\tau - \tau((\mu - \overline{x}_n)^2 + ...)) + ...$$

where $\tau$ is the precision, $\tau := \frac{1}{\sigma^2}$. So we have the MLEs:

$$\widehat{\mu} = \overline{x}_n$$
$$\widehat{\tau} = \frac{S_n^2}{n}$$

where $S_n^2 := \sum_{i=1}^{n}(x_i - \overline{x}_n)^2$

**Unbiased Estimation of Variance** For $X_1, ..., X_n$ IID conditional on $\theta$, with:

$$\begin{aligned}
\mu(\theta) &:= \mathbb{E}(X_i|\theta), \\
\sigma^2(\theta) &:= \mathbb{V}\text{ar}(X_i|\theta), \\
\overline{X}_n &:= \frac{1}{n}\sum_{i=1}^{n} X_i, \\
S_n^2 &:= \sum_{i=1}^{n}(X_i - \overline{X}_n)^2
\end{aligned}$$

we have the unbiased estimators:

$$\widehat{\mu(\theta)} = \overline{X}_n$$
$$\widehat{\sigma^2\theta} = \frac{S_n^2}{n-1}$$

# 1 Simple Frequentist Estimation and Prediction

## 1.1 Statistical Modelling and Inference

**Further Reading**

[MR03, section 7.1]
[DS12, section 7.1]

Consider the following practical scientific questions:

- By how much will the sea level rise in the next 50 years?
- What's the effectiveness of a new cancer treatment?
- What's the biological impact of introducing a non-native species to an environment?
- How much energy will a new wind farm produce, if built in a certain location?
- What is the distribution of dark matter in the universe?

What do these situations have in common?

**Definition: Uncertainty, Data, and Models**

Each situation involves:

- **Uncertainty**: There's no exact answer, due to a lack of knowledge, and due to randomness.
- **Data**: Empirical observations and expert knowledge.
- **Model**: Some idea of how the world behaves, and how data are correlated. May be thought of as a specific way of expressing the objective part of expert knowledge. This is based in physics, biology, engineering, etc.

We will use probability theory to tackle these types of question.

Probability theory involves a number of concepts we will make use of:

**Definition: Probability Theory Concepts**

The **possibility space**, notated $\Omega$, is the set of all possible outcomes. This can be huge, for example, the set of all possible distributions of dark matter; or smaller, like if we were to represent sea level rise by a single number.
We don't normally specify possibility spaces directly, but instead focus on **random variables**. A random variable is a *function* from $\Omega \mapsto \mathbb{R}$ (or $\mathbb{R}^k$). This can be observed. Random variables will always be denoted by capital letters: $X, Y, \Theta, ....$
An specific observed value of a random variable will be denoted with a lower-case letter: $x, y, \theta, ....$
A **statistical model** consists of an identification of:

- Relevant random variables (both observable and hypothetically observable) including the data.

  For example: the expansion coefficient of water, actual rise, global temperature.
- Parameters (both known and unknown), which we may learn about, but not observe directly.

  For example: The likelihood of recovery following the use of a certain treatment.

  It's important to note that we may treat uncertain parameters as random variables.
- A joint probability distribution, expressed through probability mass functions (pmfs) and probability density functions (pdfs), on *all* random variables, and possibly on all unknown paramters.

**Warning**

Random variables *only* correspond to observable (or *hypothetically* observable) quantities.

**Related Questions**

Exercises 1 and 2 on the problem sheet give you textual descriptions of some scenarios, and ask you to identify the statistical model.

**Definition: Statistical inference**

A **statistical inference** is a procedure which produces a probabilistic statement about any part of a statistical model.
A **probabilistic statement** is just the probability of an event, a mean, a variance, etc; i.e: Anything involving a mathematical statement of probability.

### 1.1.1 Example: Smartphone Batteries

*Consider a smartphone battery production line. Every 50th battery is destructively tested for "quality". Quality, here, is just some proxy for the battery lifetime. We're given the following data for the quality of the last 10 tested batteries:*

| Battery | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 90 | 86 | 82 | 77 | 94 | 90 | 87 | 90 | 86 | 86 |

*A battery is deemed faulty if quality is less than 80. Identify the relevant statistical model.*

We first need to identify the relevant random variables:

- The quality of the tested batteries (the *data*, known): $X_1, ..., X_n$
- The quality of the untested batteries (hypothetically observable, unknown): $X_{n+1}, X_{n+2}, ...$

$n = 10$ is the sample size.

We now need to assign some joint distribution. We might conjecture the following model, for example:

The $X_i$ are identically distributed (come from the same probability distribution), according to some probability density function $f(\cdot|\theta)$, for example, the normal distribution, with $\mathbb{E}(X_i|\theta) = \theta, \mathbb{V}\text{ar}(X_i|\theta) = 5^2$. This is just an assumption which seems reasonable when we look at the data, and is not devised from any mathematical process. $\theta$ is an unknown parameter representing the mean of $X_i$. $\Theta$ is a random variable representing $\theta$.

Here, $\theta \in \mathbb{R}$, but in general, you can have $\theta \in \mathbb{R}^k$ (i.e: multiple parameters).

**Warning**

We must assume that $X_i$ are independent *conditional on* $\Theta$.
Why can't we assume unconditional independence? Let's assume for now that all variables are discrete, for simplicity.
Say that I observe $X_1$ to learn about $X_2$. So we're trying to find:

$$\mathbb{P}(X_2 = x_2|X_1 = x_1) = \frac{\mathbb{P}(X_2 = x_2, X_1 = x_1)}{\mathbb{P}(X_1 = x_1)}$$

If we assume the variables are unconditionally independent, we then have:

$$\begin{aligned}\mathbb{P}(X_2 = x_2 | X_1 = x_1) &= \frac{\mathbb{P}(X_2 = x_2)\mathbb{P}(X_1 = x_1)}{\mathbb{P}(X_1 = x_1)} \\ &= \mathbb{P}(X_2 = x_2)\end{aligned}$$

So we've learned absolutely nothing about $X_2$!
If, on the other hand, we assume conditional independence on $\Theta$, we have:

$$\begin{aligned}\mathbb{P}(X_2 = x_2 | X_1 = x_1) &= \sum_\theta \mathbb{P}(X_2 = x_2 | X_1 = x_1, \Theta = \theta) \cdot \mathbb{P}(\Theta = \theta | X_1 = x_1) \\ &= \sum_\theta \mathbb{P}(X_2 = x_2 | \Theta = \theta) \cdot \mathbb{P}(\Theta = \theta | X_1 = x_1) \\ &= \sum_\theta \mathbb{P}(X_2 = x_2 | \Theta = \theta) \cdot \frac{\mathbb{P}(X_1 = x_1 | \Theta = \theta)\mathbb{P}(\Theta = \theta)}{\mathbb{P}(X_1 = x_1} \\ &\neq \mathbb{P}(X_2 = x_2) \quad \textit{(generally)}\end{aligned}$$

So we can learn about $X_2$ given the value of $X_1$! This is actually the *only* way to enable learning, by DeFinetti's representation theorem.

### 1.1.2 Estimation and Prediction

Typically, we perform statistical inference about either unknown parameters, such as $\Theta$, or about future observations, such as $X_{n+1}$.

**Definition: Estimation, Prediction**

**Estimation** specifically refers to statistical inference about unknown parameters, like $\Theta$.
**Prediction** refers to statistical inference about future observations.

## 1.2 Point Estimation

**Further Reading**

[MR03, section 7.2]

Let's return to the battery example. Consider the sample mean:

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

This would be a good choice to estimate $\theta$, because:

$$\begin{aligned}\mathbb{E}(\overline{X}_n | \theta) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i | \theta) \\ &= \frac{1}{n} n\theta \\ &= \theta\end{aligned}$$

We can also use the sample mean to estimate the variance:

$$\begin{aligned}\mathbb{V}\text{ar}(\overline{X}_n|\theta) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\text{ar}(X_i|\theta)\\ &= \frac{1}{n^2}n5^2\\ &= \frac{5^2}{n}\end{aligned}$$

Note that $\mathbb{V}\text{ar}(\overline{X}_n|\theta) \to 0$ as $n \to \infty$. So, we might wish to use $\overline{X}_n$ as an approximation for $\Theta$.

**Definition: Statistic, Estimator, Point Estimate**

A **statistic** is a real-valued function of the data, for example, $\overline{X}_n$.
An **estimator**, $\widehat{T}$, is a statistic which is meant to approximate some real-valued function, $t(\Theta)$, of the parameters. Remember that the parameters are random variables, so a function of a parameter is also a random variable. Usually, $t(\Theta)$ is just the identity function. We write $\widehat{\Theta}$ for an estimator of $\Theta$.
A **point estimate**, $\widehat{t}$ for $t(\Theta)$ is just a specific realisation of an estimator $\widehat{T}$ for $t(\Theta)$ after observing the data (the actual value of $\widehat{T}$). We write $\widehat{\theta}$ for a point-estimate of $\Theta$.

In our battery example, $\widehat{\Theta} = \overline{X}_{10}$ is an estimator for $\Theta$. $\widehat{\theta} = \frac{1}{10}(90 + 86 + ...) = 86.8$ is a point estimate of $\Theta$.

### 1.2.1 Properties of Estimators

**Definition: Bias, Errors**

For any estimator $\widehat{T}$ of $t(\Theta)$, we define:

- **Bias** $:= \mathbb{E}(\widehat{T}|\theta) - t(\theta)$

  An estimator with zero bias $\forall\ \theta$ is called **unbiased**.

  We saw that, for the sample mean, the conditional expectation is equal to the value we want to estimate, so we would say this estimator is unbiased.

- **Standard error** $:= \sqrt{\mathbb{V}\text{ar}(\widehat{T}|\theta)}$.

  We can think of this as how much an estimator will vary, its conditional standard deviation.

  We want for an estimator to have a low standard error.

- **Mean square error** $:= \mathbb{E}((\widehat{T} - t(\theta))^2|\theta)$

  This is what we (usually) *really* want to minimise for a good estimator. We prefer estimators with a low mean square error.

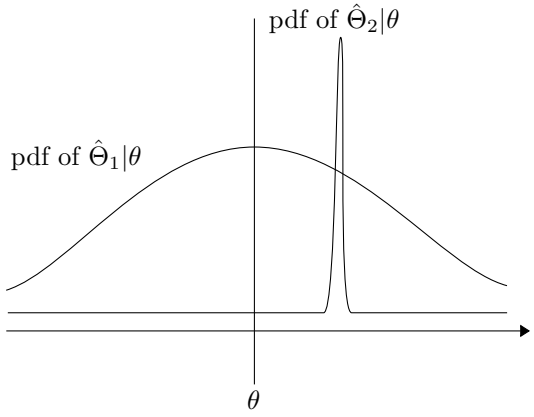
### 1.2.2 Theorem: Relation of Mean Square Error, Standard Error, and Bias

$$\text{mean square error} = (\text{standard error})^2 + (\text{bias})^2$$
$$\mathbb{E}((\widehat{T} - t(\theta))^2|\theta) = \mathbb{V}\text{ar}(\widehat{T}|\theta) + (\mathbb{E}(\hat{T}|\theta) - t(\theta))^2$$

**Related Questions**

Exercise 3 involves proving this relation. The solution is on DUO if you really want to be sure about the proof.

This theorem is important because, since we're trying to minimise mean square error, we might actually prefer a biased estimator, provided its standard error is lower.




In this case, for example, we might choose to use the clearly biased estimator $\widehat{\Theta}_2$, since its variance being so low may lead to a lower mean square error than $\widehat{\Theta}_1$.

### 1.2.3 Theorem: Sample Mean Is Minimum Variance Unbiased Estimator

Consider some sequence of random variables $X_1, X_2, ...$, with $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ where the $X_i$ are independent, identically distributed (IID) conditional on $\Theta$ and $\sigma > 0$ is some known constant. Then, $\overline{X}_n$ is the minimum variance unbiased estimator of $\Theta$ (in essence: the sample mean is the best estimator we can construct of $\Theta$).

**Related Questions**

The proof of this theorem is not given, but a related (simpler) proof is required for exercises 3 to 7.

### 1.2.4 Bias and Error in the Battery Example

In the battery example, we showed:

$$\mathbb{E}(\overline{X}_n|\theta) = \theta$$
$$\mathbb{V}\text{ar}(\overline{X}_n|\theta)\frac{5^2}{n}$$

So $\overline{X}_n$ is an unbiased estimator of $\Theta$.

$\overline{X}_n$ has standard error $\frac{5}{\sqrt{n}}$, so the mean square error is $\frac{5^2}{n}$.

## 1.3 Interval Estimation

**Further Reading**

[MR03, section 8.1, 8.2.1]

### 1.3.1 Central Limit Theorem (CLT)

This is a key result from probability theory.

Consider an infinite sequence of random variables, $X_1, X_2, ....$ We will assume they are IID conditional on some random variable $\Theta$.

We will define the following functions:

$$\mu(\theta) := \mathbb{E}(X_i|\theta)$$
$$\sigma^2(\theta) := \mathbb{V}\text{ar}(X_i|\theta)$$

Note that neither $\mu$ nor $\theta$ depend on $i$ due to the IID assumption.

We will further define the following random variables:

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$Z_n := \frac{\overline{X}_n - \mu(\Theta)}{\sigma(\Theta)/\sqrt{n}} \quad \textit{(the standardised sample mean)}$$

Then, $\forall\, z \in \mathbb{R}$, and all possible values of $\theta$, we have:

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z|\theta) = \Phi(z)$$

where $\Phi$ is the cumulative distribution function of $\mathcal{N}(0,1)$.

The practical implication of this theorem is that, for large $n$, we have the approximate result:

$$\overline{X}_n|\theta \sim \mathcal{N}\left(\mu(\theta), \frac{\sigma^2(\theta)}{n}\right)$$

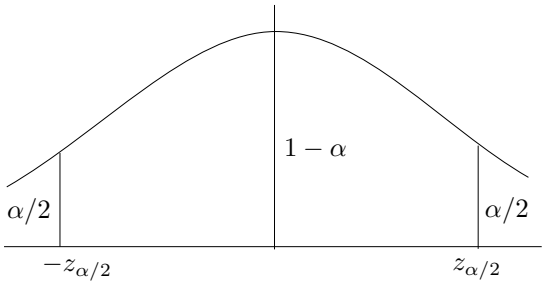
This approximation improves with a larger value of $n$, or as the distribution of $X_i|\theta$ is closer to the normal.

### 1.3.2 Application of CLT to Battery Example

In our battery example, $\mu(\theta) = \theta, \sigma^2(\theta) = 5^2$. So, $\overline{X}_{10}|\theta \sim \mathcal{N}\left(\theta, \frac{5^2}{10}\right)$ approximately.

Can we exploit the CLT to make stronger probabilistic statements about $\overline{X}_{10}$ and $\Theta$?

Let's assume the conditions of the central limit theorem, with $\mu(\theta) = \theta$ and $\sigma^2(\theta) = \sigma^2$ (i.e: $\sigma^2$ is constant). Fix any $\alpha \in [0,1]$ and let $z_{\alpha/2} := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$




By the CLT:

$$\mathbb{P}(|Z_n| \leq z_{\alpha/2}|\theta) = 1 - \alpha$$

By the definition of $Z_n$:

$$\mathbb{P}\left(\left|\frac{\overline{X}_n - \theta}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}|\theta\right) = 1 - \alpha$$

or equivalently:

$$\mathbb{P}\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\theta\right) = 1-\alpha$$

This is a **confidence interval**, which we will define after a caveat.

**Warning**

Note that, here, the $\theta$ on the left is a constant, as we're conditional on $\theta$, not a random variable, and $X_n$ is a random variable. So, if we have a value for the sample mean, we can construct the interval.
This does *not* say that the probability that $\theta$ lies in this specific interval is $1-\alpha$.
See section 1.3.3 for more information on the issues with thinking about confidence intervals like this.

**Definition: Confidence Interval**

Assume $X_1, X_2, ..., X_n$ are IID, conditional on $\Theta$, and assume that $\mathbb{E}(X_i|\theta) = \theta$, $\mathbb{V}\text{ar}(X_i|\theta) = \sigma^2 > 0$ is known, and constant independent of $\theta$. Then, $\forall\, \alpha \in [0,1]$:

$$\left[\overline{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

is a $100 \cdot (1-\alpha)\%$ **confidence interval** on $\Theta$.

Common values for $z_{\alpha/2}$:

| $1-\alpha$ | 0.80 | 0.90 | 0.95 | 0.98 | 0.99 |
|---|---|---|---|---|---|
| $z_{\alpha/2}$ | 1.28 | 1.64 | 1.96 | 2.33 | 2.58 |

Now that we've defined what a confidence interval means, we should return to the battery example. In this example, $n = 10, \overline{x}_{10} = 86.8, \sigma = 5$. For a 95% confidence interval, we need to calculate:

$$\overline{x}_n - 1.96\frac{\sigma}{\sqrt{n}} = 83.7$$
$$\overline{x}_n + 1.96\frac{\sigma}{\sqrt{n}} = 89.9$$

So the 95% confidence interval for $\Theta$ is given by $[83.7, 89.9]$.

**Related Questions**

Problem 11 is highly relevant here, and is the closest so far to an actual exam question we would expect to see.
Problem 9 also relies on confidence intervals and would be useful to attempt.

### 1.3.3 What Does a Confidence Interval Mean?

In the battery example, does $[83.7, 89.9]$ really capture the uncertainty of $\Theta$? In particular, does the event $\Theta \in [83.7, 89.9]$ have probability 0.95? *No!*

$$\mathbb{P}\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\theta\right) = 1-\alpha$$

The probability in this equation, which we used to define a confidence interval, is *conditional on knowing* $\Theta = \theta$. This isn't really what we want! We want it to be conditional on $\overline{X}_n = \overline{x}_n$.

Instead, we *only* know that $\forall\, \theta \in \mathbb{R}$, the following event:

$$\theta \in \left[\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

has probability 0.95, conditional on $\Theta = \theta$.

We used the distribution of $\overline{X}_n$ conditional on $\Theta = \theta$. We *really* want the distribution of $\Theta$ conditional on $X_1 = x_1, ..., X_n = x_n$.

How can we do that? We'll use Bayes's theorem in a later section. However, the computations are much harder, and we must be able to treat $\Theta$ as a random variable, which we've managed to avoid while thinking about confidence intervals.

## 1.4 Interval Prediction

**Further Reading**

[MR03, section 8.6]

Let's think about the battery example again. What can we say about the quality level of an untested battery, $X_{n+1}$? In other words, having observed $X_1, ..., X_n$, what can we say about $X_{n+1}$?

In prediction, we can no longer rely on the central limit theorem as we did in estimation, and must instead make another assumption about the distribution of $X_n$. So, we will assume normality, again, IID conditional on $\theta$:

$$X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$$

We then have:

$$\begin{aligned}\mathbb{E}(X_{n+1} - \overline{X}_n|\theta) &= \mathbb{E}(X_{n+1}|\theta) - \mathbb{E}(\overline{X}_n|\theta) \\ &= \theta - \theta \\ &= 0\end{aligned}$$

$$\begin{aligned}\mathbb{V}\text{ar}(X_{n+1} - \overline{X}_n|\theta) &= \mathbb{V}\text{ar}(X_{n+1}|\theta) + \mathbb{V}\text{ar}(\overline{X}_n|\theta) && \textit{(by IID of } X_1, ..., X_{n+1}\textit{)} \\ &= \sigma^2 + \frac{\sigma^2}{n} \\ &= \sigma^2\left(1 + \frac{1}{n}\right)\end{aligned}$$

Note that $X_{n+1}$ is distributed normally, and each $X_i$ is normal, so the sample mean (a sum of normals) will also be normally distributed, and the difference $X_{n+1} - \overline{X}_n$ will also have a normal distribution as follows:

$$X_{n+1} - \overline{X}_n \sim \mathcal{N}\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right)$$

and consequently, via a similar calculation to how we analysed confidence intervals, we have:

$$\mathbb{P}\left(\overline{X}_n - z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \overline{X}_n + z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}}\middle|\theta\right) = 1 - \alpha$$

**Definition: Frequentist Prediction Interval**

Assume $X_1, X_2, ..., X_n$ are IID, conditional on $\theta$, and assume that $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$, where $\sigma$ is independent of $\theta$. Then, $\forall\, \alpha \in [0,1]$:

$$\left[\overline{x}_n - z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}},\ \overline{x}_n + z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}\right]$$

is a $100 \cdot (1-\alpha)\%$ **frequentist prediction interval** for $X_{n+1}$.

**Warning**

Similar concerns to those involved in interval estimation apply: We used the distribution of $\overline{X}_n$ and $X_{n+1}$ conditional on $\Theta = \theta$, but we actually want the distribution of $X_{n+1}$ conditional on the data: $X_1 = x_1, ..., X_n = x_n$!

### 1.4.1 Interval Prediction in the Battery Example

We have $n = 10, \overline{X}_n = 86.8, \sigma = 5$, so we attain:

$$\overline{x}_n - 1.96\sigma\sqrt{1+\frac{1}{n}} = 76.5$$

$$\overline{x}_n + 1.96\sigma\sqrt{1+\frac{1}{n}} = 97.1$$

Recall that a battery was designated faulty if its quality was less than 80, so the 95% prediction interval includes values less than this threshold. As such, we'd expect to see more failures than we might be comfortable with, and we might wish to redesign the battery production process.

The interval $[76.5, 97.1]$ is a 95% frequentist prediction interval for $X_{n+1}$.

## 1.5 Parameters as Random Variables

**Further Reading**

[DS12, section 7.1]

For what we've done so far, we have avoided having to consider $\Theta$ as a random variable, because we only used $\mathbb{P}(\cdot|\theta)$, which we could consider as a distribution parametrised by $\theta$. This may be notated (and in textbooks, commonly is notated) as $\mathbb{P}_\theta(\cdot)$, or even just $\mathbb{P}(\cdot)$.

Doing this is appropriate if we only want to make probability statements indexed by $\theta$. However, this is very limiting.

If we want to make more advanced probability statements, can we treat $\Theta$ as a random variable? *Yes*, provided that $\Theta$ is (hypothetically) observable.

How can we make $\Theta$ observable? Consider, for instance, the example in which we observed the quality of batteries. By the strong rule of large numbers, we have:

$$\mathbb{P}\left(\Theta = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_i\right) = 1$$

This just means that as we take a very large value of $n$ for our sample size, we expect the sample average to converge to $\Theta$.

So in this specific example, we can treat $\Theta$ as if it were the observable quantity $\frac{1}{n}\sum_{i=1}^{n} X_i$ for very large $n$.

This limit construction is possible for a very wide range of practical statistical models (and, in particular, every model that we'll encounter in first year).

# 2 Prior and Posterior Distributions

## 2.1 Prior Distributions

**Further Reading**

[DS12, section 7.2]

**Definition: Prior PDF/PMF**

The **prior pdf or pmf** of a *parameter* $\Theta$ is the pdf/pmf of $\Theta$ in advance of observing any data. This can be used to quantify uncertainty in advance of observing actual values. Mathematically, this is the *unconditional* marginal pdf/pmf of $\Theta$.
We don't have prior pdf/pmfs for all random variables, only parameters.

### 2.1.1 Examples

Let's go back to the battery example we looked at throughout section 1, which we modelled as follows:

$$X_i|\theta \sim \mathcal{N}(\theta, 5^2)$$

Prior to seeing the data, the manager supposes that the parameter $\Theta$ is normally distributed, with mean 90 and standard deviation 10. So we have:

$$\Theta \sim \mathcal{N}(90, 10^2), \implies f(\theta) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\theta - 90}{10}\right)^2\right)$$

This is the prior pdf for $\Theta$. *It is unconditional on any* $X_i$.

**Warning**

The point of a prior pdf/pmf is that it is entirely unconditional on any $X_i$. You *cannot* use any data parameters observed in a sample to determine the prior distribution.

Let's move on to another example. Consider the outcome $X$ of a coin toss. $X|\theta \sim \text{Bin}(\theta, 1)$, where $X = 1$ represent heads, and $X = 0$ tails and either:

- The coin is fair: $\theta = \frac{1}{2}$,
- Or the coin has two heads: $\theta = 1$

Before we do any experimentation, we'll make a judgement about whether we think the coin is fair or not. Let's suppose we know the coin is fair with 80% probability. We can express this information using a probability mass function:

$$p(\theta) = \begin{cases} 0.8, & \theta = \frac{1}{2} \\ 0.2, & \theta = 1 \end{cases}$$

This is a prior pmf on $\Theta$.

We now have two examples of prior pdf/pmfs, and the distribution of data. Upon observing data, can we update our potential pdf/pmfs to quantify the uncertainty in $\Theta$?

## 2.2 Posterior Distribution and Likelihood

**Definition: Posterior Distribution**

The **posterior distribution** of a *parameter* $\Theta$ is the pdf or pmf of $\Theta$ after observing the data (i.e: the *conditional* pdf/pmf of $\Theta$ given $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$, or any other observed random variables comprising the data, although in this course the observed random vars will always be in this form).

Usually, the posterior distribution is calculated via Bayes's theorem.

### 2.2.1 Posterior Distribution for the Coin Toss Example

Consider again the coin toss problem, and say that we observe heads: $X = 1$. This provides some evidence that it might be more likely that there are two heads.

Then, by Bayes's theorem, we have:

$$\begin{aligned}\mathbb{P}\left(\Theta = \frac{1}{2}\middle| X = 1\right) &= \frac{\mathbb{P}\left(X = 1|\Theta = \frac{1}{2}\right)\mathbb{P}\left(\Theta = \frac{1}{2}\right)}{\mathbb{P}\left(X = 1|\Theta = \frac{1}{2}\right)\mathbb{P}\left(\Theta = \frac{1}{2}\right) + \mathbb{P}(X = 1|\Theta = 1)\mathbb{P}(\Theta = 1)} \\ &= \frac{\frac{1}{2} \cdot 0.8}{\frac{1}{2} \cdot 0.8 + 1 \cdot 0.2} \\ &= \frac{2}{3}\end{aligned}$$

We could also evaluate the probability that $\Theta$ is 1 by performing the same calculation to attain $\mathbb{P}(\theta = 1|X = 1) = \frac{1}{3}$. So the probability that the coin is biased has increased, which makes intuitive sense.

It's also worth checking for yourself that $\mathbb{P}\left(\Theta = \frac{1}{2}|x = 0\right) = 1$.

### 2.2.2 Theorem: Posterior Distribution in General

Assume $X_1, ..., X_n$ are IID, conditional on $\Theta$, with pdf $f(x|\theta)$ and $\Theta$ has a prior pdf $f(\theta)$. Then, $\Theta$ has posterior pdf given by:

$$f(\theta|x_1, ..., x_n) = \frac{f(\theta)\prod_{i=1}^{n} f(x_i|\theta)}{\int f(\theta')\prod_{i=1}^{n} f(x_i|\theta')d\theta'}$$

$\theta'$ is just a dummy variable over which we're integrating here.

If we instead have a probability mass function for the data, the formula is exactly the same, we just replace $f(x_i|\theta)$ with $p(x_i|\theta)$.

If the prior distribution is a pmf rather than a pdf, replace $f(\theta)$ with $p(\theta)$, replace $f(\theta|x_1, ..., x_n)$ with $p(\theta|x_1, ..., x_n)$, and note that the integral $\int ...d\theta'$ becomes a sum $\sum_{\theta'}$.

**Proof** Given a model as above with IID observations and a prior pdf, we can write down the full joint density for the model by the definition of conditional density:

$$\begin{aligned}f(x_1, ..., x_n, \theta) &= f(x_1, ..., x_n|\theta) \cdot f(\theta) \\ &= f(\theta) \cdot \prod_{i=1}^{n} f(x_i|\theta) \qquad \textit{(by IID of } X_i \textit{ cond on } \Theta\textit{)}\end{aligned}$$

Also, we have that the integral of this product over $\theta$ will give us the joint distribution of the $X_i$s, by the partition theorem:

$$\begin{aligned} f(x_1, ..., x_n) &= \int f(x_1, ..., x_n, \theta')d\theta' \\ &= \int f(\theta') \cdot \prod_{i=1}^{n} f(x_i|\theta')d\theta' \end{aligned}$$

Finally, by the definition of conditional density, we have that:

$$\begin{aligned} f(\theta|x_1, ..., x_n) &= \frac{f(x_1, ..., x_n, \theta)}{f(x_1, ..., x_n)} \\ &= \frac{f(\theta) \prod_{i=1}^{n} f(x_i|\theta)}{\int f(\theta') \prod_{i=1}^{n} f(x_i|\theta')d\theta'} \end{aligned} \qquad \square$$

The proof is very similar if we use pmfs instead.

**Tackling the Integral** The tricky part of using this theorem to calculate posterior distributions tends to be evaluating the integral in the denominator. It is useful to note, however, that it only depends on the data $x_1, ..., x_n$, not the parameter $\theta$. It only acts to normalise the numerator (to make sure the integral of the density equals 1).

We can exploit this fact to write the following:

$$f(\theta|x_1, ..., x_n) \overset{(\theta)}{\propto} f(\theta) \cdot \prod_{i=1}^{n} f(x_i|\theta)$$

**Warning**

The symbol $\overset{(\theta)}{\propto}$ means "is equal to, up to a factor independent of $\theta$". So the constant of proportionality (the *normalisation constant*) can depend on anything except $\theta$.

**Definition: Likeliness Function, Likelihood**

We define the **likeliness function** or **likelihood** to be the conditional pdf/pmf of the data given the parameter:

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

where the $X_i$ are IID are conditional on $\Theta$.
So, the posterior is proportional to the prior distribution, multiplied by the likelihood.

### 2.2.3 Assorted Lemmata

For working with normal priors and likelihoods, the following results are very useful:

**Lemma 2.2.3(a)** Let's say we have the following:

$$f(y|z) \overset{(y)}{\propto} \exp\left(-\frac{1}{2}\left(\frac{y-\mu(z)}{\sigma(z)}\right)^2\right)$$

This is equivalent to saying:

$$Y|z \sim \mathcal{N}(\mu(z), \sigma^2(z))$$

The proof of this is trivial given the definition of the normal distribution.

**Lemma 2.2.3(b)** Let's say we have a sum of the following form:

$$\sum_{i=1}^{n} a_i(\theta - b_i)^2$$

We then have the following:

$$\sum_{i=1}^{n} a_i(\theta - b_i)^2 = \left(\sum_{i=1}^{n} a_i\right) \cdot \left(\theta - \frac{\sum\limits_{i=1}^{n} a_i b_i}{\sum\limits_{i=1}^{n} a_i}\right)^2 + ... \qquad \textit{(all other terms are independent of } \theta\textit{)}$$

**Lemma 2.2.3(c)** If we just have two terms in the previous sum:

$$a_1(\theta - b_1)^2 + a_2(\theta - b_2)^2 = (a_1 + a_2) \cdot \left(\theta - \frac{a_1 b_1 + a_2 b_2}{a_1 + a_2}\right)^2 + \left(\frac{1}{a_1} + \frac{1}{a_2}\right)^{-1} \cdot (b_1 - b_2)^2$$

### 2.2.4 Posterior Distribution for the Battery Example

We'll again assume that the distribution of quality is normal, IID conditional on $\theta$:

$$X_i|\theta \sim \mathcal{N}(\theta, 5^2)$$

What is the likelihood (up to $\overset{(\theta)}{\propto}$)?

$$\begin{aligned}
f(x_1, ..., x_n|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\
&= \prod_{i=1}^{n} \frac{1}{5\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \theta}{5}\right)^2\right) \\
&\overset{(\theta)}{\propto} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{5^2}(\theta - x_i)^2\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{n}{5^2}\left(\theta - \frac{\sum_{i=1}^{n}\frac{x_i}{5^2}}{n/5^2}\right)^2 + F\right)\right) \quad \textit{(by lemma 2.2.3(b), } F \textit{ independent of } \theta\textit{)} \\
&\overset{(\theta)}{\propto} \exp\left(-\frac{n}{2 \cdot 5^2}\left(\theta - \overline{x}_n\right)^2\right) \\
&= \exp\left(-\frac{10}{2 \cdot 5^2}(\theta - 86.8)^2\right)
\end{aligned}$$

So, up to a factor independent of $\theta$, our likelihood only depends on the data through $\overline{x}_n$. Consequently, so will the posterior! Therefore, for this model, $\overline{X}_n$ is called a **sufficient statistic**: We don't need to know the full data to make inferences about $\Theta$, we only need to know the sample mean $\overline{X}_n$!

In order to find a posterior distribution, we need to suppose a prior distribution for $\Theta$, which is based on expert knowledge. Let's say an expert gives the following prior distribution:

$$\Theta \sim \mathcal{N}(90, 10^2)$$

What is the posterior density for $\Theta$?

$$\begin{aligned}
f(\theta|x_1,...,x_n) &\overset{(\theta)}{\propto} f(\theta)f(x_1,...,x_n|\theta) \\
&\overset{(\theta)}{\propto} \exp\left(-\frac{1}{2}\left(\frac{\theta-90}{10}\right)^2\right)\cdot\exp\left(-\frac{n}{2\cdot 5^2}(\theta-\bar{x}_n)^2\right) \\
&= \exp\left(-\frac{1}{2}\left[\left(\frac{\theta-90}{10}\right)^2+\frac{n}{5^2}(\theta-\bar{x}_n)^2\right]\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{1}{10^2}+\frac{n}{5^2}\right)\left(\theta-\frac{\frac{90}{10^2}+\frac{n\bar{x}_n}{5^2}}{\frac{1}{10^2}+\frac{n}{5^2}}\right)^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{1}{1.56}\right)^2\left(\theta-\frac{\frac{90}{10^2}+\frac{n\bar{x}_n}{5^2}}{\frac{1}{10^2}+\frac{n}{5^2}}\right)^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{1}{1.56}\right)^2(\theta-86.9)^2\right)
\end{aligned}$$

$$\implies \Theta|x_1,...,x_n \sim \mathcal{N}(86.9, 1.56^2)$$

**Definition: Credible Interval**

An interval $[l,u]$ is called a $100\cdot(1-\alpha)\%$ **credible interval** for $t(\Theta)$ given $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ when the posterior probability of $t(\Theta) \in [l,u]$ is $1-\alpha$.

$$\mathbb{P}(l \leq t(\Theta) \leq u | X_1 = x_1, ..., X_n = x_n) = 1-\alpha$$

$X_1, ..., X_n$ could be any other variable that comprise the data.
For $X_1, ..., X_n$ IID conditional on $\theta$, and assume $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ with posterior distribution $\Theta|x_1,...,x_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, we have that the $100\cdot(1-\alpha)$ credible interval for the parameter $\theta$ is given by:

$$[\mu_n - z_{\alpha/2}\sigma_n, \mu_n + z_{\alpha/2}\sigma_n]$$

We can now construct genuinely useful probability intervals for $\Theta$ conditional on the data!

$$\mathbb{P}(86.9 - 1.96\cdot 1.56 \leq \Theta \leq 86.9 + 1.96\cdot 1.56 | x_1,...,x_n) = 0.95$$

$[83.8, 90.0]$ is a 95% credible interval on $\theta$.

**Warning**

We should note that the 95% credible interval for $\theta$, $[83.8, 90.0]$, is very similar to the 95% confidence interval, $[83.7, 89.9]$, obtained earlier from the same data. This is not a coincidence as we will see later, but we *must* keep in mind the different interpretations of the two intervals.

To bring into sharper relief the difference between a confidence interval and a credible interval, let's consider what happens when we have a more informative prior, say:

$$\Theta \sim \mathcal{N}(90, 0.1^2)$$

We have the same calculations as before, leaving us with:

$$\Theta | x_1, ..., x_n \sim \mathcal{N}(89.99, 0.0998^2)$$

This is almost unchanged from the prior, due to the extremely low variance of the prior reflecting strong prior knowledge in the value of $\Theta \approx 90$.

### 2.2.5 Theorem: General Case for Normal Sampling

Consider a general situation as before, with $\sigma, \sigma_0$ and $\mu_0$ known constants, $X_i$ IID conditional on $\Theta$, distributed as follows:

$$X_i | \theta \sim \mathcal{N}(\theta, \sigma^2)$$

and prior distribution:

$$\Theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

Then:

$$\Theta | x_1, ..., x_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

where:

$$\mu_n := \frac{\mu_0/\sigma_0^2 + n\bar{x}_n/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}$$
$$\sigma_n^2 := \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

These formulae should probably just be memorised.

The proof is similar to the battery example given earlier.

So if $\frac{1}{\sigma_0^2} \ll \frac{n}{\sigma^2}$, then the data will drive the posterior, and the credible interval will be more similar to the confidence interval.

Whereas, if $\frac{1}{\sigma_0^2} \gg \frac{n}{\sigma^2}$, then the prior will drive the posterior.

If $\frac{1}{\sigma_0^2} \approx \frac{n}{\sigma^2}$, then both the data and the prior will influence the posterior.

**Warning**

Thinking about the prior/posterior in this way really makes it clear why $\sigma_0, \mu_0$ (the std. dev. and mean for the prior) should not be based on the data! They must reflect the uncertainty about $\theta$ as if you had not seen the data.

## 2.3 Sequential Updating and Prediction

Let's look at the posterior after a single observation, $X_1 = x_1$:

$$f(\theta | x_1) \overset{(\theta)}{\propto} f(\theta) f(x_1 | \theta)$$

Now after the second observation, $X_2 = x_2$:

$$\begin{aligned} f(\theta|x_1, x_2) &\overset{(\theta)}{\propto} f(\theta)f(x_1|\theta)f(x_2|\theta) \\ &\overset{(\theta)}{\propto} f(\theta|x_1)f(x_2|\theta) \end{aligned}$$

So, we can treat the posterior after our first observation as our prior for the posterior following the second! Similarly, after the third observation, $X_3 = x_3$:

$$\begin{aligned} f(\theta|x_1, x_2, x_3) &\overset{(\theta)}{\propto} f(\theta)f(x_1|\theta)f(x_2|\theta)f(x_3|\theta) \\ &\overset{(\theta)}{\propto} f(\theta|x_1, x_2)f(x_3|\theta) \end{aligned}$$

### 2.3.1 The Sequential Updating Theorem

$$f(\theta|x_1, ..., x_{n+1}) \overset{(\theta)}{\propto} f(\theta|x_1, ..., x_n)f(x_{n+1}|\theta)$$

We can use the posterior after $n$ observations as a prior for updating with observation $n + 1$.

The constant of proportionality has a special meaning, encapsulated in the following theorem.

### 2.3.2 Theorem: Predictive Distribution from Sequential Updates

$$f(\theta|x_1, ..., x_{n+1}) = \frac{f(\theta|x_1, ..., x_n)f(x_{n+1}|\theta)}{f(x_{n+1}|x_1, ..., x_n)}$$

So the constant of proportionality is the predictive distribution! We will use this to predict the next observation given the previous observations.

**Proof**

$$\begin{aligned} f(\theta|x_1, ..., x_{n+1})f(x_{n+1}|x_1, ..., x_n) &= f(\theta, x_{n+1}|x_1, ..., x_n) \\ &= f(x_{n+1}|\theta, x_1, ..., x_n)f(\theta|x_1, ..., x_n) \\ &= f(x_{n+1}|\theta)f(\theta|x_1, ..., x_n) \qquad \textit{(by IID on } \theta\textit{)} \quad \square \end{aligned}$$

**Definition: Prior Predictive PDF**

We define the **prior predictive pdf** to be $f(x_{n+1})$, and the **posterior predictive pdf** to be $f(x_{n+1}|x_1, ..., x_n)$.
We use the prior predictive pdf to predict $X_{n+1}$ before having seen the data, and the posterior predictive to predict $X_{n+1}$ after having seen the data.

### 2.3.3 Prediction in the Battery Example

Consider again the battery example with $X_i$ IID conditional on $\Theta$ distributed according to:

$$X_i|\theta \sim \mathcal{N}(\theta, 5^2)$$

with the prior distribution for $\Theta$ given by:

$$\Theta \sim \mathcal{N}(90, 10^2)$$

Let's find $f(x_{n+1}|x_1,...,x_n)$ for our specific data. One way to do so is to use theorem 2.3.2, and evaluate up to $\overset{(x_{n+1})}{\propto}$:

$$f(x_{n+1}|x_1,...,x_n) = \frac{f(\theta|x_1,...,x_n)f(x_{n+1}|\theta)}{f(\theta|x_1,...,x_{n+1})}$$

Note that, since the left hand side is not a function of $\theta$, you should either see that factors of $\theta$ on the right hand side cancel out, or just fix a value of $\theta$ where $f(\theta|x_1,...,x_{n+1}) > 0$.

$$f(x_{n+1}|x_1,...,x_n) \overset{(x_{n+1})}{\propto} \frac{f(x_{n+1}|\theta)}{f(\theta|x_1,...,x_{n+1})}$$

Alternatively, we could also evaluate, up to $\overset{(x_{n+1})}{\propto}$:

$$f(x_{n+1}|x_1,...,x_n) = \int f(x_{n+1}|\theta)f(\theta|x_1,...,x_n)d\theta$$

So we have:

$$\begin{aligned}
f(x_{n+1}|x_1,...,x_n) &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{x_{n+1}-\theta}{\sigma}\right)^2 \cdot \frac{1}{\sigma_n\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{\theta-\mu_n}{\sigma_n}\right)^2 d\theta \\
&\overset{(x_{n+1})}{\propto} \int_{-\infty}^{\infty} \exp -\frac{1}{2}\left(\left(\frac{x_{n+1}-\theta}{\sigma}\right)^2 + \left(\frac{\theta-\mu_n}{\sigma_n}\right)^2\right) d\theta \\
&= \int_{-\infty}^{\infty} \exp -\frac{1}{2}\left[\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2}\right)\left(\theta - \frac{x_{n+1}/\sigma^2 + \mu_n/\sigma_n^2}{1/\sigma^2+1/\sigma_n^2}\right)^2 + (\sigma^2+\sigma_n^2)^{-1}(x_{n+1}-\mu_n)^2\right] d\theta \\
&= \exp\left(-\frac{1}{2}\frac{(x_{n+1}-\mu_n)^2}{\sigma^2+\sigma_n^2}\right) \cdot \int_{-\infty}^{\infty} \exp -\frac{1}{2}\left(\frac{\theta-\mu_{n+1}}{\sigma_{n+1}}\right)^2 d\theta \\
&\overset{(x_{n+1})}{\propto} \exp -\frac{1}{2}\frac{1}{\sigma^2+\sigma_n^2}(x_{n+1}-\mu_n)^2
\end{aligned}$$

$$\implies X_{n+1}|x_1,...,x_n \sim \mathcal{N}(\mu_n, \sigma^2+\sigma_n^2)$$

**Definition: Posterior and Prior Prediction Intervals**

Assume $X_1, X_2, ..., X_n$ are IID conditional on $\theta$, and assume $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$, with posterior distribution $\Theta|x_1,...,x_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$.Then, $\forall\, \alpha \in [0,1]$:

$$\left[\mu_n - z_{\alpha/2}\sqrt{\sigma^2+\sigma_n^2}, \mu_n + z_{\alpha/2}\sqrt{\sigma^2+\sigma_n^2}\right]$$

is a $100\cdot(1-\alpha)\%$ **posterior prediction interval** for $X_{n+1}$.
The $100\cdot(1-\alpha)\%$ **prior prediction interval** given a prior distribution $\Theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$ is given by:

$$\left[\mu_0 - z_{\alpha/2}\sqrt{\sigma^2+\sigma_0^2}, \mu_0 + z_{\alpha/2}\sqrt{\sigma^2+\sigma_0^2}\right]$$

Substituting our specific data, we have $\mu_n = 86.9$, $\sqrt{\sigma^2+\sigma_n^2} = 5.24$. So:

$$\mathbb{P}(86.9 - 1.96\cdot 5.24 \le X_{n+1} \le 86.9 + 1.96\cdot 5.24 | x_1,...,x_n) = 0.95$$

and the 95% posterior prediction interval for $X_{n+1}$ is:

$$[76.6, 97.2]$$

This is very similar to the 95% frequentist prediction interval $[76.5, 97.1]$ for the same data obtained earlier.

If, instead, we had a prior pdf for $\Theta$ such that:

$$\Theta \sim \mathcal{N}(90, 0.1^2)$$

we would have $\mu_n = 89.99, \sigma_n = 0.0998$, as seen earlier, so the 95% posterior prediction interval is:

$$[89.99 - 1.96\sqrt{5^2 + 0.0998^2}, 89.99 + 1.96\sqrt{5^2 + 0.0998^2}] = [80.188, 99.792]$$

Due to the very low variance in the prior pdf, we would expect this prediction interval to be almost identical to the 95% prior prediction interval. This would be:

$$[90 - 1.96\sqrt{5^2 + 0.1}, 90 + 1.96\sqrt{5^2 + 0.1}] = [80.198, 99.802]$$

**Related Questions**

These questions are probably appropriate for the whole of section 2.
[DS12, section 7.2, exercises 1, 2, 3, 4, 5, 6, 7, 10, 11]

# Aside: Summary of Sections 1 and 2

## Frequentist Method

- Needs a likelihood, given by:

$$\prod_{i=1}^{n} f(x_i|\theta)$$

- We constructed confidence intervals, conditioned on the parameter $\theta$:

$$\mathbb{P}(L \leq \theta \leq U|\theta) = 1 - \alpha$$

  where $L, U$ are functions of the data, e.g: $\overline{X}_n \pm z_{\alpha/2}, ...$

- We don't have a sense of having a distribution for a predictive quantity based on the data in the frequentist approach.

- We constructed frequentist prediction intervals:

$$\mathbb{P}(L \leq X_{n+1} \leq U|\theta) = 1 - \alpha$$

## Bayesian Method

- Needs a posterior $\propto$ prior $\times$ likelihood:

$$f_{\alpha_0}(\theta|x_1, ..., x_n) \overset{(\theta)}{\propto} f_{\alpha_0}(\theta) \prod_{i=1}^{n} f(x_i|\theta)$$

  Here, $\alpha_0$ is a hyper-parameter, like $\mu_0, \sigma_0$: It's a parameter that determines the distribution of other parameters.

- We constructed credible intervals, conditioned on the data:

$$\mathbb{P}(l \leq \Theta \leq u|x_1, ..., x_n) = 1 - \alpha$$

  where $l, u$ are functions of the data and the hyper-parameters (so $\alpha_0, x_1, ..., x_n$).

- We looked at the posterior predictive distribution:

$$f_{\alpha_0}(x_{n+1}|x_1, ..., x_n) \overset{(x_{n+1})}{\propto} \frac{f(x_{n+1}|\theta)}{f_{\alpha_0}(\theta|x_1, ..., x_{n+1})}$$

- We constructed posterior prediction intervals, also conditioned on the data:

$$\mathbb{P}(l \leq X_{n+1} \leq u|x_1, ..., x_n) = 1 - \alpha$$

- We looked at the prior predictive distribution:

$$f_{\alpha_0}(x_{n+1}) \overset{(x_{n+1})}{\propto} \frac{f(x_{n+1}|\theta)}{f_{\alpha_0}(\theta)}$$

- We constructed prior predictive intervals:

$$\mathbb{P}(l' \leq X_{n+1} \leq u') = 1 - \alpha$$

  where $l', u'$ are functions of $\alpha_0$ only.

# 3 Conjugate Distributions

## 3.1 Sampling from a Normal with Known Variance

We saw that, in the context of the battery example, if $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ IID conditional on $\Theta$, and $\Theta \sim \mathcal{N}(\mu_0, \sigma_0)$, then the posterior distribution for $\Theta$ is also normal:

$$\Theta|x_1, ..., x_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

with simple formulae for $\mu_n, \sigma_n$ as functions of $\mu_0, \sigma_0, x_1, ..., x_n$ to update from the prior to the posterior.

The property that the posterior is from the same class of distribution as the prior is not unique to this scenario!

## 3.2 Sampling from a Bernoulli Distribution

### 3.2.1 Example: Clinical Trial

150 patients are randomly selected, and receive different treatments for depression: Imipramine (treatment 1), lithium carbonate (treatment 2), a combination of the two (treatment 3), or a placebo (treatment 4). The following table shows the number of patients who suffered a relapse within three years:

| Treatment | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| Relapse | 18 | 13 | 22 | 24 | 77 |
| No relapse | 22 | 25 | 16 | 10 | 73 |
| Total | 40 | 38 | 38 | 34 | 150 |

For now, we'll just focus on treatment 1, and try to apply Bayesian methods in order to make statistical inferences. Let $\Theta$ be the proportion of patients from the general population suffering from depression who do not relapse under this treatment.

The first step is to identify the statistical model: We will define the random variable as follows:

$$X_i = \begin{cases} 0, & \text{if patient relapses} \\ 1 & \text{if patient does not relapse} \end{cases}$$

$$X_i|\theta \sim \text{Bernoulli}(\theta)$$

(so $\mathbb{P}(X_i = 1|\theta) = \theta, \mathbb{P}(X_i = 0|\theta) = 1 - \theta$).

We will assume the following prior distribution for $\Theta$:

$$\Theta \sim \text{Beta}(\alpha, \beta), \quad \alpha > 0, \beta > 0$$

**Definition: Beta Distribution**

Let $\alpha > 0, \beta > 0$. We say that $\theta$ follows a **Beta distribution**, $\text{Beta}(\alpha, \beta)$, when $\Theta$ has the following pdf:

$$f(\theta) := \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}, & 0 \leq \theta \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

N.b: $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is the normalisation constant.
For the beta distribution, we have the following expressions:

$$\mathbb{E}(\Theta) = \frac{\alpha}{\alpha+\beta}$$
$$\mathbb{V}\text{ar}(\Theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

You don't need to remember these formulae, they'll be given to you if you have to use them in an exam setting.
$\Gamma$ is the **gamma function**. We don't need to know much about it, other than that for $z > 0$, we have $\Gamma(z) > 0$. Again, if we need to use its properties in an exam, we will be given them. There is not a closed form expression for its entire domain.

Returning to the clinical trial example, let's look at the likelihood. As this is a discrete case, we'll have a pmf rather than a pdf, which by the IID cond on $\Theta$ assumption, is given by:

$$\begin{aligned} p(x_1, ..., x_n|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^y(1-\theta)^{n-y}, \quad \left(\textit{where } y := \sum_{i=1}^{n} x_i\right) \end{aligned}$$

So $Y_i = \sum_{n=1}^{n} X_i$ is a sufficient statistic, and since we have the data, we have its value.
Now let's find the posterior distribution:

$$\begin{aligned} f(\theta|x_i, ..., x_n) &\overset{(\theta)}{\propto} f(\theta)p(x_1, ..., x_n|\theta) \\ &\overset{(\theta)}{\propto} \begin{cases} \theta^{\alpha+y-1} \cdot (1-\theta)^{\beta+n-y-1}, & \theta \in [0,1] \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

So the posterior distribution must be a $\text{Beta}(\alpha + y, \beta + n - y)$ distribution!

### 3.2.2 Theorem

If $X_i|\theta \sim \text{Bernoulli}(\theta)$ IID conditional on $\Theta$, and $\Theta \sim \text{Beta}(\alpha_0, \beta_0)$ for some constants $\alpha_0, \beta_0 > 0$ then we've shown that $\Theta|x_1, ..., x_n \sim \text{Beta}(\alpha_n, \beta_n)$ where, with $y := \sum_{i=1}^{n} x_i$, we have:

$$\alpha_n := \alpha_0 + y$$
$$\beta_n := \beta_0 + n - y$$

### 3.2.3 Application of Bayesian Stats to Clinical Trial

In the clinical trial above, if $\Theta \sim \text{Beta}(1, 1)\ (= \mathcal{U}(0, 1))$, then for the first treatment, since $n = 40, y = 22$, we have that:

$$\Theta|x_1, ..., x_n \sim \text{Beta}(1 + 22, 1 + 40 - 22) = \text{Beta}(23, 19)$$

For the posterior expectation/variance, we have:

$$\mathbb{E}(\Theta|x_1, ..., x_n) = \frac{1+22}{1+22+1+40-22} = \frac{23}{42} = 0.45$$

$$\mathbb{V}\text{ar}(\Theta|x_1,...,x_n) = \frac{23 \cdot 19}{42^2 \cdot 43} = 0.076^2$$

For comparison with the prior, we had $\mathbb{E}(\Theta) = \frac{1}{2}, \mathbb{V}\text{ar}(\Theta) = 0.29^2$. So the uncertainty has really decreased.

For the credible interval, unfortunately, there's no closed form when using the beta distribution.

## 3.3 Conjugate Families and Hyper-Parameters

**Definition: Conjugate Families**

Assume we have data $X_i$, IID conditional on the parameter $\Theta$ with pdf $f(x|\theta)$, or pmf $p(x|\theta)$. Let $f_\alpha(\theta)$ represent a family of densities for $\Theta$, indexed by the *hyper-parameter* $\alpha \in \mathcal{A} \subseteq \mathbb{R}^k$. Then this family is said to be a **conjugate family of priors** for sampling from $f(x|\theta)$ if $\forall\, \alpha_0 \in \mathcal{A}$, and all $n \in \mathbb{N}$, and all possible samples $x_1, ..., x_n$, then $\exists\, \alpha_n \in \mathcal{A}$ such that:

$$f(\theta) = f_{\alpha_0}(\theta) \implies f(\theta|x_1,...,x_n) = f_{\alpha_n}(\theta)$$

i.e: Whenever the prior belongs to the family, then so does the posterior.

The benefit of using a conjugate family is that *updating the distribution is reduced to updating just the hyper-parameters!*

Let's consider, for example, the case using the normal distribution. $\mathcal{N}(\mu_0, \sigma_0^2)$, with hyperparameters $\mu_0 \in \mathbb{R}, \sigma_0 > 0$, is a conjugate family for $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$, assuming that $\sigma$ is known.

$$\mu_n := \frac{\mu_0/\sigma_0^2 + n\overline{x}_n/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}$$

$$\sigma_n^2 := \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

Now considering a beta distribution, $\text{Beta}(\alpha_0, \beta_0)$ with hyper-parameters $\alpha_0, \beta_0 > 0$, we see that this forms a conjugate family for $X_i|\theta \sim \text{Bernoulli}(\theta)$, with:

$$\alpha_n := \alpha_0 + \sum_{i=1}^{n} x_i$$

$$\beta_n := \beta_0 + n - \sum_{i=1}^{n} x_i$$

## 3.4 Sampling from an Exponential Distribution

**Further Reading**

[DS12]

### 3.4.1 Example: Lifetimes of Electrical Components

Consider a company selling electrical components. Each component is assumed to fail with probability $\theta dt$ in any time interval $[t, t+dt]$ for small values of $dt$, where $\theta$ is an unknown parameter. So the component lifetimes $X_1, X_2, ...$ are exponentially distributed, as follows:

$$X_i|\theta \sim \text{Exp}(\theta)$$

in which $\theta$ is called the rate parameter. The pdf for this distribution is given by:

$$f(x_i|\theta) = \begin{cases} \theta \exp(-\theta x_i), & x_i \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

We make the usual assumption that the $X_i$ are IID conditional on $\Theta$.

A priori, the company judges that $\mathbb{E}(\Theta) = 0.5, \mathbb{V}\text{ar}(\Theta) = 0.5^2$. We then observe: $X_1 = 3, X_2 = 1.5, X_3 = 2.1$. What can we say about future failures?

We need to identify:

1. A family of conjugate priors for this case (i.e: the case of sampling from an exponential distribution)
2. The posterior pdf of $\Theta$
3. The posterior *predictive* pdf of $\Theta$

**Definition: Gamma Distribution**

We say that $\Theta \sim \text{Gamma}(\alpha, \beta)$ for $\alpha, \beta > 0$ if the pdf is given by:

$$f(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, & \theta \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

This is the family of **gamma distributions**.
For a gamma distribution, we have the following results:

- $\mathbb{E}(\Theta) = \frac{\alpha}{\beta}$
- $\mathbb{V}\text{ar}(\Theta) = \frac{\alpha}{\beta^2}$
- If $\alpha = 1$, then we have $\text{Gamma}(1, \beta) = \text{Exp}(\beta)$

In this example, if $\Theta \sim \text{Gamma}(\alpha_0, \beta_0)$ for some known constants $\alpha_0, \beta_0 > 0$, then for $\theta \geq 0$, the posterior distribution follows:

$$\begin{aligned} f(\theta|x_1, ..., x_n) &\overset{(\theta)}{\propto} f(\theta) \prod_{i=1}^{n} f(x_i|\theta) \\ &\overset{(\theta)}{\propto} \theta^{\alpha_0 - 1} e^{-\beta_0 \theta} \prod_{i=1}^{n} \theta e^{-\theta x_i} \\ &= \theta^{\alpha_0 + n - 1} \exp\left( - \left( \beta_0 + \sum_{i=1}^{n} x_i \right) \theta \right) \end{aligned}$$

This is a $\text{Gamma}(\alpha_n, \beta_n)$ pdf, with:

$$\begin{aligned} \alpha_n &:= \alpha_0 + n \\ \beta_n &:= \beta_0 + \sum_{i=1}^{n} x_i \end{aligned}$$

So, the family of Gamma distributions is *conjugate for exponential sampling.*

In this example, we need that $\mathbb{E}(\Theta) = \frac{\alpha_0}{\beta_0} = \frac{1}{2}$, and $\mathbb{V}\text{ar}(\Theta) = \frac{\alpha_0}{\beta_0^2} = \frac{1}{4}$, so we can very easily see that $\alpha_0 = 1, \beta_0 = 2$.

The posterior distribution will be a $\text{Gamma}(\alpha_n, \beta_n)$ distribution, with:

$$\alpha_n = \alpha_0 + n = 1 + 3 = 4$$
$$\beta_n = \beta_0 + \sum_{i=1}^{n} x_i = 8.6$$

Note that, from the posterior distribution, we therefore have $\mathbb{E}(\Theta|x_1, ..., x_n) = \frac{4}{8.6} \approx 0.47 \approx \frac{1}{2}$, so the expectation is still pretty close to the prior distribution. However, when considering the variance, $\mathbb{V}\text{ar}(\Theta|x_1, ..., x_n) = \frac{4}{8.6^2} = 0.054 \ll \frac{1}{4}$, so the variance has greatly decreased.

To find the posterior predictive pdf, we may use integration. For $x_{n+1} \geq 0$, we have:

$$\begin{aligned} f(x_{n+1}|x_1, ..., x_n) &= \int_0^\infty f(x_{n+1}|\theta) \cdot f(\theta|x_1, ..., x_n) d\theta \\ &\overset{(x_{n+1})}{\propto} \int_0^\infty \theta \exp(-\theta x_{n+1}) \theta^{\alpha_n - 1} \exp(-\beta_n \theta) d\theta \\ &= \int_0^\infty \theta^{\alpha_n} \exp(-(\beta_n + x_{n+1})\theta) d\theta \end{aligned}$$

The integrand here is the pdf of a Gamma$(\alpha_n + 1, \beta_n + x_{n+1})$ distribution, up to some normalisation constant, so the integral is the Gamma function over such a constant:

$$\begin{aligned} &= \frac{\Gamma(\alpha_n + 1)}{(\beta_n + x_{n+1})^{\alpha_n + 1}} \\ &\overset{(x_{n+1})}{\propto} (\beta_n + x_{n+1})^{-(\alpha_n + 1)} \end{aligned}$$

An alternative method for finding this result is to consider:

$$\begin{aligned} f(x_{n+1}|x_1, ..., x_n) &= \frac{f(x_{n+1}|\theta) f(\theta|x_1, ..., x_n)}{f(\theta|x_1, ..., x_{n+1})} \\ &= \frac{\theta e^{-x_{n+1}\theta} \beta_n^{\alpha_n} / \Gamma(\alpha_n) \cdot \theta^{\alpha_n - 1} e^{-\beta_n \theta}}{\beta_{n+1}^{\alpha_{n+1}} / \Gamma(\alpha_{n+1}) \cdot \theta^{\alpha_{n+1} - 1} e^{-\beta_{n+1}\theta}} \\ &= \frac{\Gamma(\alpha_n + 1)}{\Gamma(\alpha_n)} \cdot \frac{\beta_n^{\alpha_n}}{(\beta_n + x_{n+1})^{\alpha_n + 1}} \end{aligned}$$

This step is allowed because we have $\alpha_{n+1} = \alpha_n + 1, \beta_{n+1} = \beta_n + x_{n+1}$.

We then use a "magical" result that is left unproven for this course: the first term equals $\alpha_n$:

$$f(x_{n+1}|x_1, ..., x_n) = \frac{\alpha_n \beta_n^{\alpha_n}}{(\beta_n + x_{n+1})^{\alpha_n + 1}}$$

This is the same result.

Note that the posterior predictive pdf found here is *not* exponential! It follows the **Lomax Distribution**: $X_{n+1}|x_1, ..., x_n \sim \text{Lomax}(\alpha_n, \beta_n)$. For the Lomax distribution, we have:

- $\mathbb{E}(X_{n+1}|x_1, ..., x_n) = \frac{\beta_n}{\alpha_n - 1}$ if $\alpha_n > 1$
- $\mathbb{V}\text{ar}(X_{n+1}|x_1, ..., x_n) = \frac{\beta_n^2 \alpha_n}{(\alpha_n - 1)^2 (\alpha_n - 2)}$ if $\alpha > 2$

In our case, the posterior predictive pdf of $X_4$ is:

$$f(x_4|x_1, x_2, x_3) \overset{(x_4)}{\propto} (8.6 + x_4)^{-5}$$

with:

$$\mathbb{E}(x_4|x_1, ..., x_3) = \frac{8.6}{3} = 2.87$$
$$\mathbb{V}\text{ar}(x_4|x_1, ..., x_3) = \frac{8.6^2 \cdot 4}{3^2 \cdot 2} = 4.05^2$$

**Further Reading**

[DS12, section 7.3] contains many more examples

**Related Questions**

[DS12, section 7.3, exercises 1 - 13, 17, 19, 20]
Extra exercises: [DS12, section 7.3, exercises 14, 15, 16, 23, 24]

# 4 Bayes Estimators

**Further Reading**

[DS12, section 7.4]

Suppose that, instead of reporting the entire posterior distribution $f(\theta|x_1, ..., x_n)$, we we only want to report a summary (i.e: a single number, which is a function of the data such that this value is somehow "close" to the parameter $\Theta$). In other words, we want to find an *estimator* for $\Theta$:

$$\widehat{\Theta} := \delta(X_1, ..., X_n)$$

We will find this based on the posterior distribution. We want to choose $\delta$ such that $\delta(X_1, ..., X_n) - \Theta$ is close to zero.

We formally define this concept of "distance" with a loss function.

## 4.1 Loss Functions

**Definition: Loss Function**

A **loss function** $L(\theta, \widehat{\theta})$ is just a real-valued function of two variables.
The idea is that if $\Theta = \theta$ and we choose $\widehat{\theta}$ as our estimate, then we lose the amount $L(\theta, \widehat{\theta})$.
Typically, the larger the distance between $\theta$ and $\widehat{\theta}$, the larger the loss $L(\theta, \widehat{\theta})$ will be.

### 4.1.1 Examples of Loss Functions

- **Squared Error Loss**

$$L(\theta, \widehat{\theta}) := (\theta - \widehat{\theta})^2$$

- **Absolute Error Loss**

$$L(\theta, \widehat{\theta}) := |\theta - \widehat{\theta}|$$

- Arbitrary example

$$L(\theta, \widehat{\theta}) := \begin{cases} 3(\widehat{\theta} - \theta)^2, & \theta \geq \widehat{\theta} \\ (\widehat{\theta} - \theta)^2, & \text{otherwise} \end{cases}$$

  Note that this is asymmetric, and could make sense if underestimating is more costly than overestimating (e.g: sea level rise).

We will mostly focus on squared error loss.

## 4.2 Bayes Estimator

If we have the posterior pdf $f(\theta|x_1, ..., x_n)$, then for any $\widehat{\theta}$, we can find the posterior expected loss.

**Definition: Posterior Expected Loss, Bayes Esimators and Estimates**

The **posterior expected loss** is given by:

$$\mathbb{E}(L(\Theta, \widehat{\theta})|x_1, ..., x_n) := \int_{-\infty}^{\infty} L(\theta, \widehat{\theta}) f(\theta|x_1, ..., x_n) d\theta$$

If working with a pmf, replace the integral with a sum.
Let $L$ be any loss function. For any $x_1, ..., x_n$, let $\delta(x_1, ..., x_n)$ denote the value $\widehat{\theta}$ for which the posterior expected loss is minimised. Then $\delta(X_1, ..., X_n)$ is called the **Bayes Estimator** of $\Theta$, an $\delta(x_1, ..., x_n)$ is called the **Bayes Estimate** of $\widehat{\Theta}$.

In other words, $\delta$ is chosen such that:

$$\mathbb{E}(L(\Theta, \delta(x_1, ..., x_n))|x_1, ..., x_n) = \min \mathbb{E}(L(\Theta, \widehat{\Theta})|x_1, ..., x_n)$$

Note that $\delta$ depends on:

- The data, $x_1, ..., x_n$
- The posterior distribution of $\Theta$
- The loss function $L$

#### 4.2.1 Theorem: Bayes Estimate in Squared Error Loss

Under squared error loss, the Bayes estimate for $\Theta$ is the posterior expectation for $\Theta$, given by:

$$\delta(x_1, ..., x_n) = \mathbb{E}(\Theta|x_1, ..., x_n)$$

**Proof** For simplicity of exposition, we will omit the conditioning on $x_1, ..., x_n$.
We need to show $\mathbb{E}((\Theta - \widehat{\theta})^2)$ is minimised for $\widehat{\theta} = \mathbb{E}(\Theta)$. Indeed:

$$\begin{aligned}\mathbb{E}((\Theta - \widehat{\theta})^2) &= \mathbb{E}((\Theta - \mathbb{E}(\Theta) + \mathbb{E}(\Theta) - \widehat{\theta})^2) \\ &= \mathbb{E}((\Theta - \mathbb{E}(\Theta))^2) + 2\mathbb{E}(\Theta - \mathbb{E}(\Theta))(\mathbb{E}(\Theta) - \widehat{\theta}) + (\mathbb{E}(\Theta - \widehat{\theta})^2)\end{aligned}$$

#### 4.2.2 Example

Consider $X_i|\theta \sim$ Bernoulli IID conditional on $\Theta$, with prior $\Theta \sim \text{Beta}(\alpha, \beta)$.

By conjugacy, we have the posterior $\Theta|x_1, ..., x_n \sim \text{Beta}(\alpha_n, \beta_n)$ with $\alpha_n := \alpha_0 + \sum_{i=1}^{n} x_i, \beta_n := \beta_0 + n - \sum_{i=1}^{n} x_i$.
So the Bayes estimate under squared error loss is:

$$\mathbb{E}(\Theta|x_1, ..., x_n) = \frac{\alpha_n}{\alpha_n + \beta_n} = \frac{\alpha_0 + \sum_{i=1}^{n} x_i}{\alpha_0 + \beta_0 + n}$$

by the expression of the expectation of the beta distribution. The Bayes estimator for $\Theta$ is:

$$\widehat{\Theta} := \frac{\alpha_0 + \sum_{i=1}^{n} X_i}{\alpha_0 + \beta_0 + n}$$

For instance, in our depression relapse problem, we had $\alpha_0 = \beta_0 = 1, n = 40, \sum_{i=1}^{n} x_i = 22$, so $\widehat{\theta} = 23/42 = 0.55$ is our Bayes estimate.

### 4.2.3 Theorem: Bayes Estimate in Absolute Error Loss

Under absolute error loss, the Bayes estimate for $\Theta$ is the posterior median of $\Theta$, i.e: $\delta(x_1, ..., x_n)$ is chosen such that:

$$\mathbb{P}(\Theta \leq \delta(x_1, ..., x_n)|x_1, ..., x_n) = \frac{1}{2}$$

**Proof**

Further Reading

[DS12, theorem 4.5.3]

# 5 Maximum Likelihood Estimators

## 5.1 Definitions

As discussed previously, the **likelihood** is the pmf or pdf of the data, given the parameter, written as follows via conditional independence:

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) (= l(\theta))$$

Bayes' theorem gives us that the posterior is proportional to the likelihood multiplied by the prior. But what do we do if we don't have a prior? Is it possible to construct an estimator for $\Theta$ based on just the likelihood? This problem is what we'll consider in this section.

**Definition: Arg max, maximum likelihood estimator**

For any function $g(\theta)$, let $\arg\max_{\theta} g(\theta)$ denote the value of $\theta$ for which $g(\theta)$ is maximised.

$\theta* = \arg\max_{\theta} g(\theta)$ when $g(\theta*) = \max_{\theta} g(\theta)$.

We define the **maximum likelihood estimator** (MLE) of $\Theta$ to be:

$$\widehat{\Theta} = \delta(X_1, ..., X_n) = \arg\max_{\theta} f(X_1, ..., X_n|\theta)$$

After observing $X_1 = x_1, ..., X_n = x_n$, then:

$$\widehat{\theta} = \delta(x_1, ..., x_n) = \arg\max_{\theta} f(x_1, ..., x_n|\theta)$$

is the **maximum likelihood estimate**.

The interpretation of this is that $\widehat{\theta}$ is the value which maximises the probability of the data conditional on $\theta$.

## 5.2 Log Likelihood

In practice, we often work with the **log likelihood**, given by:

$$L(\theta) := \log f(x_1, ..., x_n|\theta)$$

and then maximise $L(\theta)$ instead. Since log is a strictly increasing function, this is equivalent to maximising $l(\theta)$.

How do we maximise such a function? We'll just use differentiation: We need to find $\widehat{\theta}$ such that $L'(\widehat{\theta}) = 0$ and $L''(\widehat{\theta}) < 0$.

### 5.2.1 Log Likelihood in Component Lifetimes Example

We have $X_i|\theta \sim \text{Exp}(\theta), \theta > 0$, with the observations $X_1 = 3, X_2 = 1.5, X_3 = 2.1$, and $X_i$ IID conditional on $\theta$.

We therefore have the likelihood:

$$\begin{aligned} f(x_1, x_2, x_3|\theta) &= \prod_{i=1}^{3} \theta e^{-\theta x_i} \\ &= \theta^3 e^{-\theta \sum_{i=1}^{3} x_i} \\ &= \theta^3 e^{-6.6\theta} \end{aligned}$$

And the log likelihood:

$$L(\theta) = 3 \log \theta - 6.6\theta$$

We wish to maximise this.

$$L'(\theta) = \frac{3}{\theta} - 6.6, \quad L''(\theta) = -\frac{3}{\theta^2}$$

At $\theta = \widehat{\theta}$, we have $L'(\widehat{\theta}) = 0 \iff \widehat{\theta} = 0.455$. It's clear that $L''(\theta) < 0 \, \forall \, \theta$, so the MLE is $\widehat{\theta} = 0.455$.

## 5.3 Bernoulli Sampling

Let's say we have $X_i \in \{0, 1\}, X_i|\theta \sim \text{Bernoulli}(\theta)$, so $\mathbb{P}(X_i = 1|\theta) = \theta, \mathbb{P}(X_i = 0|\theta) = 1 - \theta$.
Therefore, the likelihood is:

$$\begin{aligned} f(x_1, ..., x_n|\theta) &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i} \cdot (1-\theta)^{n - \sum x_i} \end{aligned}$$

So $f(x_1, ..., x_n|\theta) = \theta^y (1-\theta)^{n-y}$ where $y = \sum_{i=1}^{n} x_i$, and we have the log likelihood:

$$L(\theta) = y \log \theta + (n - y) \log(1 - \theta)$$

To find the maximum, first consider the boundary cases:

- For $y = 0$, it's clear that $L(\theta)$ is monotone decreasing in $\theta$, so the maximum is at $\widehat{\theta} = 0$
- For $y = n$, we have that $L(\theta)$ is monotone increasing in $\theta$, so $\widehat{\theta} = 1$

Otherwise, differentiate:

$$L'(\theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta}$$

At $\theta = \widehat{\theta}$:

$$\begin{aligned} &\frac{y}{\widehat{\theta}} - \frac{n-y}{1-\widehat{\theta}} = 0 \\ \iff &y(1-\widehat{\theta}) - (n-y)\widehat{\theta}) = 0 \\ \implies &n\widehat{\theta} = y \\ \implies &\widehat{\theta} = \frac{y}{n} \end{aligned}$$

This makes sense, since it's the sample proportion. We should also check that $L''(\widehat{\theta}) < 0$, but we won't bother here.

So, the MLE $\widehat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Let's now compare this with the Bayes estimator under squared error loss, with prior $\Theta \sim \text{Beta}(\alpha_0, \beta_0)$, which will be the posterior expectation:

$$\mathbb{E}(\theta|x_1, ..., x_n) = \frac{\alpha_0 + \sum_{i=1}^{n} x_i}{\alpha_0 + \beta_0 + n}$$

This Bayesian estimate will tend to the MLE $\widehat{\theta}$ as $\alpha_0 \to 0, \beta_0 \to 0$. (We can't set $\alpha_0 = 0$ or $\beta_0 = 0$ as $\text{Beta}(0, 0)$ is not a proper distribution).

## 5.4 Normal Sampling with Unknown Mean and Known Variance

When $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ with unknown $\theta$ and known $\sigma^2$, we saw in section 2.2.4 that:

$$f(x_1, ..., x_n|\theta) \overset{(\theta)}{\propto} \exp\left(-\frac{1}{2}\frac{n}{\sigma^2}(\theta - \overline{x}_n)^2\right)$$

So:

$$L(\theta) = -\frac{1}{2}\frac{n}{\sigma^2}(\theta - \overline{x}_n)^2 + D$$

where $D$ are terms that do not depend on $\theta$.

Clearly, this function is maximised for $\widehat{\theta} = \overline{x}_n$. So the maxmimum likelihood estimator for $\Theta$ is $\overline{X}_n$ (i.e: the sample mean, regardless of the variance).

Compare this with the Bayes estimate:

$$\mathbb{E}(\Theta|x_1, ..., x_n) = \frac{\mu_0/\sigma_0^2 + n\overline{x}_n/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}$$

Again, we recover the maximum likelihood estimator as $\sigma_0 \to \infty$ (we can't set $\sigma_0 = \infty$ as $\mathcal{N}(\mu_0, \infty)$ is not a valid normal distribution).

## 5.5 Normal Sampling with Unknown Mean and Unknown Variance/Precision

It is slightly more convenient to work with the precision rather than the standard deviation if variance is unknown.

**Definition: Precision**

For a given standard deviation $\sigma$, we define the **precision**, $\tau$, as follows:

$$\tau := \frac{1}{\sigma^2}$$

Our parameter is now a vector $\theta = (\mu, \tau)$ with 2 components. We assume that $X_i|\mu, \tau \sim \mathcal{N}(\mu, \frac{1}{\tau})$ are IID conditional on $\Theta$. The likelihood is:

$$f(x_1,...,x_n|\mu,\tau) = \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x_i-\mu)^2\right)$$
$$\overset{(\mu,\tau)}{\propto} \tau^{\frac{n}{2}} \exp\left(-\frac{1}{2}\tau\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

So, by lemma 2.2.3(b):

$$\begin{aligned} L(\mu,\tau) &= \frac{n}{2}\log\tau - \frac{1}{2}\tau\sum_{i=1}^{n}(x_i-\mu)^2 + D_1 \\ &= \frac{n}{2}\left(\log\tau - \tau((\mu-\overline{x}_n)^2 + D_2)\right) + D_3 \end{aligned}$$

where $D_1, D_2, D_3$ are the terms that do not depend on $\mu$ or $\tau$.
Clearly, for any $\tau > 0$, this is maximised for $\widehat{\mu} = \overline{x}_n$. To find $\widehat{\tau}$, we must maximise:

$$L(\widehat{\mu},\tau) = \frac{n}{2}\log\tau - \frac{1}{2}\tau\sum_{i=1}^{n}(x_i-\widehat{\mu})^2 + ...$$

over $\tau$.

$$\frac{\partial L(\widehat{\mu},\tau)}{\partial\tau} = \frac{n}{2\tau} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x}_i)^2 = 0$$

This is satisfied for:

$$\frac{1}{\widehat{\tau}} = \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2 = \frac{S_n^2}{n}$$

where $S_n^2 = \sum_{i=1}^{n}(x_i-\overline{x}_n)^2$.
Note that:

$$\frac{\partial^2 L(\widehat{\mu},\tau)}{\partial\tau^2} = -\frac{n}{2\tau^2} < 0 \quad \forall\,\tau > 0$$

So yes, we have found a maximum.

## 5.6 Unbiased Estimation of Variance

**Further Reading**

[DS12, section 8.7]

### 5.6.1 Theorem

Assume we have $X_1, ..., X_n$ IID conditional on some parameter $\Theta$. Let the following definitions hold:

$$\begin{aligned}\mu(\theta) &:= \mathbb{E}(X_i|\theta),\\ \sigma^2(\theta) &:= \mathbb{V}\text{ar}(X_i|\theta),\\ \overline{X}_n &:= \frac{1}{n}\sum_{i=1}^{n} X_i,\\ S_n^2 &:= \sum_{i=1}^{n}(X_i - \overline{X}_n)^2\end{aligned}$$

Then, $\overline{X}_n$ is an unbiased estimator for $\mu(\Theta)$ and $\frac{S_n^2}{n-1}$ is an unbiased estimator for $\sigma^2(\Theta)$.

**Warning**

This means that, for normal sampling with unknown $\mu, \theta^2$, the maximum likelihood estimator for $\sigma^2$, namely $\frac{S_n^2}{n}$, is biased! We should instead use $\frac{S_n^2}{n-1}$.

**Proof** The first part of the proof is easier:

$$\mathbb{E}(\overline{X}_n|\theta) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i|\theta) = \frac{1}{n}\cdot n\mu(\theta) = \mu(\theta)$$

For the second part, note that:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \mu(\theta))^2 &= \sum_{i=1}^{n}(X_i - \overline{X}_n + \overline{X}_n - \mu(\theta))^2\\ &= \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + n(\overline{X}_n - \mu(\theta))^2 + 2\sum_{i=1}^{n}(X_i - \overline{X}_n)(\overline{X}_n - \mu(\theta))\\ &= S_n^2 + n(\overline{X}_n - \mu(\theta))^2\end{aligned}$$

So:

$$\begin{aligned}\mathbb{E}(S_n^2|\theta) &= \sum_{i=1}^{n}\mathbb{E}((X_i - \mu(\theta))^2|\theta) - n\mathbb{E}((\overline{X}_n - \mu(\theta))^2|\theta)\\ &= n\sigma^2(\theta) - n\frac{\sigma^2(\theta)}{n}\\ &= (n-1)\sigma^2(\theta)\end{aligned}$$

# 6 Sampling Distributions

## 6.1 Introduction

**Further Reading**

[DS12, section 8.1]

We may ask: prior to seeing the data in a sample, what can we say about the statistical properties of an estimator $\widehat{\Theta}$ (which could be Bayesian, MLE, etc) for a parameter, either conditionally on $\Theta = \theta$, or unconditionally?

### 6.1.1 Examples: Depression Drug Trial

Let's consider once again the example of depression treatment, where we're considering how likely relapse is under different forms of treatment. We modelled each patient as $X_i|\theta \sim \text{Bernoulli}(\theta)$, where $\mathbb{P}(\text{patient relapses}|\theta) = \theta$. We might take $\widehat{\Theta}$ to be the maximum likelihood estimator $\sum_{i=1}^{n} \frac{x_i}{n}$ as an estimator for $\Theta$. How good is this estimator?

To know this, before we see any data, we could be interested in (for example, in the *frequentist approach*) $\mathbb{P}(|\widehat{\Theta} - \theta| \leq 0.1|\theta)$ for all $\theta \in [0, 1]$. If this probability is small, this isn't good: it means the estimator is further from the true value. If the probability is large, then we are confident that the true value of $\theta$ will be "close" to the actually observed value $\widehat{\theta}$.

In order to actually calculate this probability, we need the distribution of $\widehat{\Theta}$ given $\Theta = \theta$.

If we take a *Bayesian approach*, we'll have a prior on $\Theta$, and might want to know:

$$\mathbb{P}(|\widehat{\Theta} - \Theta| \leq 0.1) = \int_0^1 \mathbb{P}(|\widehat{\Theta} - \theta| \leq 0.1|\theta) \cdot f(\theta) d\theta$$

Once again, we need the distribution of $\widehat{\Theta}$ conditional on $\Theta = \theta$. In the case of the depression treatment trial, we have $n\widehat{\Theta}|\theta \sim \text{Bin}(n, \theta)$, so:

$$p(\widehat{\theta}|\theta) = \binom{n}{n\widehat{\theta}} \cdot \theta^{n\widehat{\theta}} \cdot (1-\theta)^{n(1-\widehat{\theta})}, \quad \widehat{\theta} = 0, \frac{1}{n}, \frac{2}{n}, ..., 1$$

determines the distribution of $\widehat{\Theta}|\theta$.

### 6.1.2 Definition

**Definition: Sampling Distribution**

Let $T := t(X_1, X_2, ..., X_n, \Theta)$ be any function of the data and (optionally) the parameter. Then, the distribution of $T$, conditional on $\Theta = \theta$, is called the **sampling distribution** of $T$.

### 6.1.3 Example: Sample Mean and the CLT

As another example, let's look at $X_1, ..., X_n$ IID conditional on $\Theta$. By the central limit theorem, with:

$$\mu(\theta) := \mathbb{E}(X_i|\theta), \quad \sigma^2(\theta) := \mathbb{V}\text{ar}(X_i|\theta) > 0$$

then (approximately, for large values of $n$), we have the sampling distribution for $\theta$, given by:

$$\overline{X}_n|\theta \sim \mathcal{N}\left(\mu(\theta), \frac{\sigma^2(\theta)}{n}\right)$$

**Warning**

Remember that this is *approximation*: it won't hold for too small a value of $n$. For $n \geq 30$, for most distributions, this approximation works fairly well.

### 6.1.4 Example: Electrical Component Lifetimes

Recall the example from section 3.4.1 of electrical component lifetimes. We modelled lifetimes as $X_i|\theta \sim \text{Exp}(\theta)$. Under a $\text{Gamma}(\alpha_0, \beta_0)$ prior for $\Theta$, we saw that $\Theta|x_1, ..., x_n \sim \text{Gamma}(\alpha_n, \beta_n)$ with $\alpha_n := \alpha_0 + n, \beta_n := \beta_0 + \sum_{i=1}^{n} x_i$. Therefore, $\mathbb{E}(\Theta|x_1, ..., x_n) = \frac{\alpha_0+n}{\beta_0+\sum_{i=1}^{n} x_i}$. A Bayes estimator is therefore given by:

$$\widehat{\Theta} := \frac{\alpha_0 + n}{\beta_0 + \sum_{i=1}^{n} x_i} = \frac{4}{2 + X_1 + X_2 + X_3}$$

What can we say about the sampling distribution of $\widehat{\Theta}$? Note the following useful fact: If $X_i|\theta \sim \text{Exp}(\theta)$, then $\sum_{i=1}^{n} X_i|\theta \sim \text{Gamma}(n, \theta)$. So, for any $a > 0$:

$$\begin{aligned}
\mathbb{P}(\widehat{\Theta} \leq a|\theta) &= \mathbb{P}(4 \leq a(2 + X_1 + X_2 + X_3)|\theta) \\
&= \mathbb{P}(\frac{4}{a} - 2 \leq X_1 + X_2 + X_3|\theta) \\
&= 1 - G\left(\frac{4}{a} - 2|\theta\right)
\end{aligned}$$

where $G$ is the CDF of the Gamma distribution, given here by:

$$G\left(\frac{4}{a} - 2 \middle| \theta\right) = \begin{cases} \exp\left(-\theta\left(\frac{4}{a} - 2\right)\right) \cdot \sum_{k=0}^{2} \theta\left(\frac{4}{a} - 2\right)^k / k!, & 0 < a < 2 \\ 1, & a \geq 2 \end{cases}$$

We can then use this to find:

$$\begin{aligned}
\mathbb{P}(|\widehat{\Theta} - \theta| \leq 0.1|\theta) &= \mathbb{P}(\widehat{\Theta} \leq \theta + 0.1|\theta) - \mathbb{P}(\widehat{\Theta} \leq \theta - 0.1|\theta) \\
&= g(\theta) \\
\implies \mathbb{P}(|\widehat{\Theta} - \theta| \leq 0.1) &= \int_0^\infty g(\theta) \cdot f(\theta) d\theta \\
&\approx 0.478
\end{aligned}$$

Let's also look at the maximum likelihood estimator:

$$\widehat{\Theta} := \frac{n}{\sum_{i=1}^{n} X_i} = \frac{3}{X_1 + X_2 + X_3}$$

Let's say we look at $\mathbb{P}\left(\left|\frac{\widehat{\Theta}}{\theta} - 1\right| \leq 0.1|\theta\right)$ (i.e: the relative error instead of the absolute).

We begin with a useful fact: If we have $Y \sim \text{Gamma}(\alpha, \beta)$, and we wish to rescale $Y$, then we have:

$$aY \sim \text{Gamma}\left(\alpha, \frac{\beta}{a}\right), \quad \forall\, a > 0$$

We have seen that $X_1 + X_2 + X_3|\theta \sim \text{Gamma}(3, \theta)$, and from the definition of $\widehat{\Theta}$, $\frac{\theta}{\widehat{\Theta}} = \frac{\theta}{3}(X_1 + X_2 + X_3)$. We can use these observations to write:

$$\frac{\theta}{\widehat{\Theta}}|\theta \sim \text{Gamma}(3, 3)$$

So the sampling distribution of $\frac{\widehat{\Theta}}{\theta}$ *does not depend on* $\theta$*!* Any quantity or random variable whose sampling distribution doesn't depend on the parameter(s) is called a **pivotal quantity**. So:

$$\mathbb{P}\left(\left|\frac{\widehat{\Theta}}{\theta} - 1\right| \leq 0.1 \middle| \theta\right) = \mathbb{P}\left(\frac{\widehat{\Theta}}{\theta} \leq 1.1 \middle| \theta\right) - \mathbb{P}\left(\frac{\widehat{\Theta}}{\theta} \leq 0.9 \middle| \theta\right) \approx 0.134$$

So the probability we wanted to find is 0.134, regardless of the value of $\theta$. We may note that this is not a great estimator.

Consequently, we also have that (regardless of the prior distribution of $\Theta$):

$$\mathbb{P}\left(\left|\frac{\widehat{\Theta}}{\Theta} - 1\right| \leq 0.1\right) \approx 0.134$$

## 6.2 The Chi-Square Distribution

**Further Reading**

[DS14, section 8.2]

From the exercises, you should have shown that the maximum likelihood estimator for $\sigma^2$ when $X_i|\sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ (assuming that $\mu$ is known) is equal to:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{S_n^2(\mu)}{n}$$

(since the sum itself here is the definition of $S_n^2$).

The sampling distribution of $\frac{S_n^2(\mu)}{\sigma^2}$ is a well known object: it is a $\text{Gamma}(n/2, 1/2)$ distribution. Because it comes up so often, we give it a special name.

**Definition: Chi-Square Distribution**

For any $n > 0$ (where $n$ is typically an integer), we say that $X$ has **Chi-Square Distribution**, with $n$ degrees of freedom, and write:

$$X \sim \chi^2(n)$$

if $X \sim \text{Gamma}(n/2, 1/2)$.
For this distribution, we have the following expressions, by the corresponding definitions for the Gamma distribution:

$$\mathbb{E}(X) = \frac{n/2}{1/2} = n$$

$$\mathbb{V}\text{ar}(X) = \frac{n/2}{(1/2)^2} = 2n$$

The moment generating function is given by:

$$M_X(t) = \mathbb{E}(e^{tX}) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}, \quad -\frac{1}{2} < t < \frac{1}{2}$$

### 6.2.1 Theorem: Sum of Chi-Square Distributions

If $X_i \sim \chi^2(n_i)$ are independent, then:

$$\sum_{i=1}^{k} X_i \sim \chi^2\left(\sum_{i=1}^{k} n_i\right)$$

**Proof** By the properties of the moment generating functions, we know that:

$$M_{\sum_{i=1}^k X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(\frac{1}{1-2t}\right)^{\sum_{i=1}^k n_i/2}$$

which is the moment generating function of a $\chi^2\left(\sum_{i=1}^{k} n_i\right)$ distribution.

### 6.2.2 Theorem: Relation Between Normal and Chi-Square

If $X \sim \mathcal{N}(0,1)$, then $X^2 \sim \chi^2(1)$.

**Proof** Let $Y := X^2$. Then, for any $y \geq 0$:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) \end{aligned}$$

Note that $\frac{d}{dy}\sqrt{y} = \frac{1}{2\sqrt{y}}$, and $\frac{d}{dz}\Phi(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$. Therefore:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\left(F_Y(y)\right) \\ &= \frac{1}{2\sqrt{y}}\phi(\sqrt{y}) + \frac{1}{2\sqrt{y}}(\phi(-\sqrt{y})) \\ &= \frac{\phi(\sqrt{y})}{\sqrt{y}} \\ &\overset{(y)}{\propto} y^{-\frac{1}{2}} e^{-\frac{1}{2}y} \end{aligned}$$

Recall that if $Z \sim \text{Gamma}(\alpha, \beta)$, then $f(z) \overset{(z)}{\propto} z^{\alpha-1}e^{-\beta z}$ so this final answer is the density of $\text{Gamma}(1/2, 1/2) = \chi^2(1)$.

### 6.2.3 Theorem: Further Relations to Normal

If $X_i \sim \mathcal{N}(0,1)$ are IID, then $\sum_{i=1}^{k} X_i^2 \sim \chi^2(k)$.

**Proof** Simply combine the previous two theorems.

#### 6.2.4 Theorem:

If $X_i|\sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ are IID conditional on $\sigma^2$, with a known value of $\mu$, then:

$$\frac{S_n^2(\mu)}{\sigma^2}|\sigma^2 \sim \chi^2(n)$$

A consequence of this theorem is as follows:

$$\mathbb{E}\left(\frac{S_n^2(\mu)}{n}\right) = \sigma^2$$

Or, in other words, $S_n^2(\mu)$ is an unbiased estimator for $\sigma^2$.

**Proof** Note that $\frac{S_n^2(\mu)}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$. Now note that $\frac{X_i-\mu}{\sigma} \sim \mathcal{N}(0,1)$ and use the previous theorem for the proof.

### 6.3 Joint Sampling Distribution of Sample Mean and Sample Variance

**Further Reading**

[DS12, section 8.3]

At the start of term, we constructed confidence intervals of the form $\overline{x}_n \pm a\sigma$, which has the property that:

$$\mathbb{P}(\overline{X}_n - a\sigma \leq \theta \leq \overline{X}_n + a\sigma|\theta) = 1 - \alpha$$

for a specified value of $\alpha \in [0,1], a = \frac{z_{\alpha/2}}{\sqrt{n}}$.

In order to do this, we assumed $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ are IID conditional on $\Theta$, and $\sigma^2$ was known. However, knowing the standard deviation is *highly* unrealistic in practise. Instead, we may wish to make statements about intervals of the form $\overline{X}_n \pm a\sqrt{\frac{S_n^2}{n-1}}$ (recall that we just demonstrated $\frac{S_n^2}{n}$ is an unbiased estimator for $\sigma^2$). To do this, we need the *joint* distribution of $\overline{X}_n$ and $S_n^2$.

#### 6.3.1 Theorem: Independence of Sample Mean and Variance for Normal Sampling

If $X_i|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ are IID conditional on the random variables $\mu, \sigma^2$ (n.b: in our capital notation for random variables, we should write the RVs as $M, \Sigma^2$, but no-one does this), then $\overline{X}_n|\mu, \sigma^2 \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ with $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. Further:

$$\frac{S_n^2}{\sigma^2}|\mu, \sigma \sim \chi^2(n-1), \quad S_n^2 := \sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

**Warning**

Take note especially of the fact that the parameter here is $n-1$, not $n$: Using the sample mean rather than $\mu$ means we lose a degree of freedom.

Then, $\overline{X}_n$ and $S_n^2$ are i*ndependent conditional on $\mu$ and $\sigma^2$*! This is kind of intuitive: learning about the mean tells us nothing about the variance, and vice-versa.

**Proof** The main trick used in this proof is non-obvious: We transform the initial variables $X_1, ..., X_n$ as follows.

$$B := \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 1 & 1 & -2 & 0 & \dots & 0 \\ 1 & 1 & 1 & -3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & \dots & -n+1 \\ 1 & 1 & 1 & 1 & \dots & 1 \end{bmatrix}$$

So the entries of each row of $B$, other than the final one, will sum to 0.

Let $B_i$ denote the $i$th row of $B$. Note that the inner product of two rows is:

$$B_i B_j^T = \begin{cases} 0, & i \neq j \\ i + i^2, & i = j < n \\ n, & i = j = n \end{cases}$$

Therefore, two distinct rows are orthogonal. We will now further demonstrate they are *orthonormal*. Define $A_i := \frac{B_i}{\sqrt{B_i B_i^T}}$, and form the matrix:

$$A := \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{bmatrix}$$

Then:

$$A_i A_j^T = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

So $AA^T = I$, and $A$ is therefore called **orthonormal**. Orthonormal matrices are interesting because they correspond to a change in basis representing a rotation.

Define $Z_i := \frac{X_i - \mu}{\sigma}$, and note that $Z_i | \mu, \sigma^2 \sim \mathcal{N}(0, 1)$ are IID conditional on $\mu, \sigma^2$. Define:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} := A \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots Z_n \end{bmatrix}$$

Now note the following facts:

1.

$$\begin{aligned}\sum_{i=1}^{n} Y_i^2 &= Y^T Y \\ &= (Z^T A^T)(AZ) \\ &= Z^T (A^T A) Z \\ &= Z^T Z \\ &= \sum_{i=1}^{n} Z_i^2\end{aligned}$$

2. The joint density of the $Z_i$ is given by:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \overset{(z)}{\propto} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} z_i^2\right)$$

This is invariant under rotation, as it is only a function of the distance to the origin. Therefore, $z \mapsto y = Az$ is a rotation, and the joint density of the $Y_i$ is:

$$\overset{(y)}{\propto} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right)$$

So the $Y_i|\mu, \sigma^2 \sim \mathcal{N}(0,1)$ are IID conditional on $\mu, \sigma^2$

3.

$$\begin{aligned}Y_n &= \sum_{i=1}^{n} \frac{Z_i}{n} \\ &= \sqrt{n}\overline{Z}_n \\ &= \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\end{aligned}$$

So:

$$\overline{X}_n|\mu, \sigma^2 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

# Problems Classes

## 07-Feb-2020

**11** *$n = 100$ random samples of water from a fresh water lake were taken and the calcium concentration (given in milligrams per litre) measured. A 95% confidence interval on the mean calcium concentration $\Theta$ is $[0.49, 0.82]$, where the standard deviation is assumed to be known*

*(a) Consider the following statement: There is a 95% chance that $\Theta$ is between 0.49 and 0.82. Is this statement correct? Explain your answer*

No, the statement is not correct, as the confidence interval is conditioned on $\theta$, so we are making a probability statement about the *data*, not the parameter!

$$\mathbb{P}(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

From the information we know, the probability $\mathbb{P}(0.49 \leq \Theta \leq 0.82)$ could be anywhere between 0 and 1, and actually depends on the prior.

Also, $\mathbb{P}(0.49 \leq \theta \leq 0.82|\theta)$ will be either 0 or 1 depending on $\theta$.

*(b) The process of taking $n = 100$ random samples of water from the lake and computing a 95% confidence interval on $\Theta$ is repeated 1000 times, each time obtaining a slightly different interval due to randomness across the different samples. One of these intervals is $[0.49, 0.82]$. The true value $\theta$ of $\Theta$ is then revealed. Someone claims that precisely 950 of the 1000 95% confidence intervals computed earlier should contain $\theta$. Is this correct? Explain your answer.*

This is *not* correct, although we could say that it's "almost" correct. We need to remember that, ultimately, the number of intervals that will contain $\theta, Y$, is a random variable, and is not known in advance. $Y$ may be thought of as a sum of Bernoulli variables:

$$Y = \sum_{i=1}^{1000} I_{A_i}, \quad A_i = \{\overline{X}_{n,i} - 1.96\frac{\sigma}{\sqrt{n}} \leq \Theta \leq ... + 1.96...\}$$

If conditional on $\Theta = \theta, Y$ will still not have a fixed value! Recall:

$$I_{A_i}(\omega) := \begin{cases} 1, & \omega \in A_i \\ 0, & \omega \notin A_i \end{cases}$$

Note that $\mathbb{P}(A_i|\theta) = 0.95$.

*(c) Let $Y$ denote the number of 1000 95% confidence intervals that contain $\Theta$. Identify the unconditional distribution of $Y$, and compute its unconditional mean and standard deviation. Finally, identify two numbers $a, b$ such that (approximately) $\mathbb{P}(a \leq Y \leq b) = 0.95$.*

By (b), we have:

$$Y|\theta \sim \text{Bin}(1000, 0.95)$$

Because this conditional distribution doesn't depend on $\Theta$, by the partition theorem, we have:

$$p(y) = \int_0^1 p(y|\theta) f(\theta) d\theta = p(y|\theta)$$

Note that $p(y|\theta)$ doesn't depend on $\theta$, so we have

$$p(y) = p(y|\theta) \int_0^1 f(\theta) d\theta$$

So, $Y \sim \text{Bin}(1000, 0.95)$.

Note that $\mathbb{E}(Y) = 1000 \cdot 0.95 = 950, \mathbb{V}\text{ar}(Y) = 1000 \cdot 0.95 \cdot 0.05 = 6.89^2$.

Since we have a large $n$, we may approximate the binomial distribution by the normal distribution, so approximately:

$$Y \sim \mathcal{N}(950, 6.89^2)$$

Therefore, a 95% interval on $Y$ can be calculated:

$$\mathbb{P}(950 - 1.96 \cdot 6.89 \leq Y \leq 950 + 1.96 \cdot 6.89) = 0.95$$
$$[936.5, 963.5]$$

**22** *Consider again the conditions of exercise 21, and assume the same prior distribution of $\Theta$. Suppose now, however, that six observations are selected at random from the uniform distribution $[\theta - 1/2, \theta + 1/2]$, and their values are $11.0, 11.5, 11.7, 11.1, 11.4, 10.3$. Determine the posterior distribution.*

We have $X_i|\theta \sim \mathcal{U}(\theta - 1/2, \theta + 1/2)$, and the prior $\Theta \sim \mathcal{U}(10, 20)$.

$$f(\theta|x_1, ..., x_n) \overset{(\theta)}{\propto} f(\theta) \prod_{i=1}^{n} f(x_i|\theta)$$

where $f(\theta) \overset{(\theta)}{\propto} \begin{cases} 1, & \theta \in [10, 20] \\ 0, & \text{otherwise} \end{cases}$.

$$\implies f(x_i|\theta) = \begin{cases} 1, & x_i \in [\theta - 1/2, \theta + 1/2] \\ 0, & \text{otherwise} \end{cases}$$

$x_i \in [\theta - 1/2, \theta + 1/2] \iff \theta \in [x_i - 1/2, x_i + 1/2]$, so we have:

$$f(\theta|x_1, ..., x_n) \overset{(\theta)}{\propto} \begin{cases} 1, & \theta \in [10, 20] \wedge \theta \in [x_i - 1/2, x_i + 1/2] \; \forall \, i = 1, ..., n \\ 0, & \text{otherwise} \end{cases}$$

The first condition of this piecewise function holds $\iff \theta \in [10, 20] \cap \bigcap_{i=1}^{n} [x_i - 1/2, x_i + 1/2] \iff \theta \in [10] \cap [\max_{i=1}^{n} x_i - 1/2, \min_{i=1}^{n} x_i + 1/2]$. Let $x^* = \max_{i=1}^{n} x_i, x_* = \min_{i=1}^{n} x_i$ So we have:

$$\Theta|x_1, ..., x_n \sim \mathcal{U}(x^* - 1/2, x_* + 1/2) = \mathcal{U}(11.2, 11.4)$$

**33** *Let $\theta$ denote the average number of defects per 100 feet of a certain type of magnetic tape. Suppose that the value of $\theta$ is unknown and that the prior distribution of $\Theta$ is the gamma distribution with parameters $\alpha_0 = 2, \beta_0 = 10$. When a 1200-foot roll of this tape is inspected, exactly four defects are found. Determine the posterior distribution of $\Theta$.*

**Related Questions**

Consider exercise 31 first!

We could pretend that instead of one 1200 foot roll, we instead have 12 one foot rolls, with a total of $\sum_{i=1}^{12} x_i = 4$ defects. Therefore, by exercise 31, the posterior distribution is:

$$\text{Gamma}\left(\alpha_0 + \sum_{i=1}^{n} x_i, \beta_0 + n\right) = \text{Gamma}(3 + 4, 1 + 12) = \text{Gamma}(7, 13)$$

## 21-Feb-2020

**38** *In a clincal trial, let the probability of successful outcome $\Theta$ have a prior distribution that is the uniform distribution on the interval $[0, 1]$, which is also the beta distribution with parameters 1 and 1. Suppose that the first patient has a successful outcome. Find the Bayes estimate of $\Theta$ under squared error loss.*

We have the prior $\Theta \sim \mathcal{U}(0,1) = \text{Beta}(1,1)$, and the sampling $X_i|\theta \sim \text{Bernoulli}(\theta)$. We have $n = 1, x_1 = 1$. From lectures, we have $\widehat{\theta} = \mathbb{E}(\Theta|x_1, ..., x_n)$, so the question is actually just asking us to find the posterior expectation under squared error loss.

By conjugacy, we have $\Theta|x_1 \sim \text{Beta}(1+1, 1+1-1) = \text{Beta}(2,1)$, since $\alpha_1 = \alpha_0 + x_1, \beta_1 = \beta_0 + n - x_1$.

By the properties of the beta distribution, we have:

$$\mathbb{E}(\Theta|x_1) = \frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{2}{3}$$

**41** *A random sample of size $n$ is taken from the Bernoulli distribution with parameter $\theta$, which is unknown. The prior distribution for $\theta$ is a beta distribution for which the mean is $\mu_0$. Show that the mean of the posterior expectation of $\Theta$ will be a weighted average of the form $\gamma_n \overline{x}_n + (1-\gamma_n)\mu_0$, and show that $\gamma_n \to 1$ as $n \to \infty$.*

We have sampling $X_i|\theta \sim \text{Bernoulli}(\theta)$, with prior $\Theta \sim \text{Beta}(\alpha_0, \beta_0)$. Therefore, $\mathbb{E}(\Theta) = \frac{\alpha_0}{\alpha_0+\beta_0} = \mu_0$.

By conjugacy, we have:

$$\begin{aligned}
\mathbb{E}(\Theta|x_1, ..., x_n) &= \frac{\alpha_n}{\alpha_n + \beta_n} \\
&= \frac{\alpha_0 + \sum_{i=1}^{n} x_i}{\alpha_0 + \beta_0 + n} \\
&= \frac{\alpha_0}{\alpha_0 + \beta_0 + n} + \frac{n}{\alpha_0 + \beta_0 + n}\overline{x}_n
\end{aligned}$$

Is the denominator $\alpha_0 + \beta_0 + n$ equal to $\gamma_n$? Yes, if the first fraction $\frac{\alpha_0}{\alpha_0+\beta_0+n}$ is equal to $(1-\gamma_n)\mu_0$.

We can check this:

$$\begin{aligned}
(1-\gamma_n)\mu_0 &= \frac{\alpha_0 + \beta_0}{\alpha_0 + \beta_0 + n} \cdot \frac{\alpha_0}{\alpha_0 + \beta_0} \\
&= \frac{\alpha_0}{\alpha_0 + \beta_0 + n}
\end{aligned}$$

So yes, we've demonstrated that this is a weighted average.

**49** *It is not known what proportion $\Theta$ of the purchases of a certain brand of breakfast cereal are made by women, and what proportion are made by men. In a random sample of 70 purchases of this cereal, it was found that 58 were made by women and 12 were made by men.*

*(a) Find the maximum likelihood estimator of $\Theta$.*

We are doing Bernoulli sampling. Let $X_i$ equal 1 if female, 0 if male. Then $X_i|\theta \sim \text{Bernoulli}(\theta)$.

For this sampling, the MLE is therefore $\widehat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} \approx 0.83$.

*(b) Now suppose that it is additionally known that $\frac{1}{2} \leq \Theta \leq \frac{2}{3}$. What is now the maximum likelihood likelihood estimator of $\Theta$? (Hint: When is $L'(\theta) > 0$?)*

Let's answer the hint first, since if we want to maximise this function, we just need to maximise the log likelihood.

$$L(\theta) = y \log \theta + (n-y)\log(1-\theta), \quad y = \sum_{i=1}^{n} x_i$$

$$\begin{aligned}
& L'(\theta) > 0 \\
\iff & \frac{y}{\theta} - \frac{n-y}{1-\theta} > 0 \\
\iff & (1-\theta)y - \theta(n-y) > 0 \\
\iff & y - n\theta > 0 \\
\iff & \theta < \frac{y}{n} = \frac{58}{70} \approx 0.83
\end{aligned}$$

So the maximum of the log likelihood $L(\theta)$ over $\left[\frac{1}{2}, \frac{2}{3}\right]$ is achieved at $\widehat{\theta} = \frac{2}{3}$ (as the gradient is always positive, so the function is strictly increasing).

**50** *Suppose that $X_1, ..., X_n$ form a random sample from a Poisson distribution for which the mean $\theta$ is unknown, with $\theta > 0$.*

*(a) Determine the maximum likelihood estimate of $\Theta$, assuming that at least one of the oserved values is different from 0.*

We have sampling given by $X_i|\theta \sim \text{Po}(\theta)$, so $p(x_i|\theta) = \theta^{x_i} \frac{e^{-\theta}}{x_i!}$, and $\log p(x_i|\theta) = x_i \log \theta - \theta - \log(x_i!)$. So:

$$\begin{aligned}
L(\theta) &= \log p(x_1, ..., x_n|\theta) \\
&= \log \prod_{i=1}^{n} p(x_i|\theta) \\
&= \sum_{i=1}^{n} \log p(x_i|\theta) \\
&= \sum_{i=1}^{n} (x_i \log \theta - \theta - \log(x_i!)) \\
&= \log \theta \sum_{i=1}^{n} x_i - n\theta - D
\end{aligned}$$

where $D$ is a constant that doesn't depend on $\theta$.

Therefore:

$$L'(\theta) = \frac{\sum_{i=1}^{n} x_i}{\theta} - n = \frac{n\overline{x}_n}{\theta} - n$$

$$L''(\theta) = -\frac{n\overline{x}_n}{\theta^2}$$

So $L'(\widehat{\theta}) = 0$ when $\widehat{\theta} = \overline{x}_n$, provided $x_n > 0$ (so when at least one $x_i > 0$). Further $L''(\overline{\theta}) < 0$, so it is a maximum.

*(b) Show that the maximum likelihood estimate of $\Theta$ does not exist if every observed value is 0.*

If all $x_i = 0$, then $\overline{x}_n = 0$, and $L'(\theta) = -n < 0$. This achieves a maximum near $\theta = 0$, but for $\theta = 0, \text{Po}(\theta)$ is not a valid distribution, as $\widehat{\theta}$ does not exist in this case!