Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

Lecture 27: Introduction to Bayesian Ideas in Statistics

Relevant textbook passages:
Larsen–Marx [7]: Sections 5.3, 5.8, 5.9, 6.2

27.1 Priors and posteriors

Larsen–Marx [7]: § 5.8, pp. 333–345

The Bayesian approach essentially treats the parameters θ as subject to the laws of probability. We do not have to believe that the parameters are the result of a random experiment by Nature. Rather we assign probabilities representing degree of belief to the parameters *a priori* and use evidence to update the probabilities *a posteriori*. That is, we start with a probability density φ on the set Θ of parameters. It is represent the statistician's degree of belief about which value of the parameter is the true parameter that generates the data. The density φ is called the **prior probability density** on Θ.

The joint density or likelihood $f(\boldsymbol{x};\theta)$ of the data vector $\boldsymbol{x}$ under parameter θ is treated as a conditional probability density. After observing $\boldsymbol{x}$, we use Bayes' Law to compute the **posterior probability** on Θ, which your textbook [7] refers to as g, but I shall refer to as $\varphi(\cdot \,|\, \boldsymbol{x})$.

The posterior density given observation $\boldsymbol{x}$ is given by Bayes' Law as

$$\varphi(\theta_0 \,|\, \boldsymbol{x}) = \frac{f(\boldsymbol{x};\theta_0)\varphi(\theta_0)}{\int_\Theta f(\boldsymbol{x};\theta)\varphi(\theta)\,d\theta}$$

$$\propto f(\boldsymbol{x};\theta_0)\varphi(\theta_0).$$

[You have already done similar calculations with some of the urn problems earlier in the course.]

A **uniform prior** or **uninformative prior** has $\varphi(\theta)$ independent of θ, in which case the posterior density is proportional to the likelihood function. If Θ is unbounded, there is not really a constant probability density, since $\int_\Theta \varphi\,d\theta = \infty$, so such a "prior" is called an **improper prior**. Once evidence (data) has been obtained, the resulting posterior distribution may no longer be improper.

Actually, the description of Bayesian methodology here is perhaps simplistic. There is still an ongoing debate about Bayesian methods, what they are, and whether they should be used. The exchange between Brad Efron [3] Herman Chernoff [1], and Dennis Lindley [8], among others, though nearly twenty years old still is relevant.

27.2 ⋆ Bayesian Updating and Martingales

Starting with a prior φ_0 on Θ, we get an observation X_1, which leads to a posterior φ_1 on Θ. This new probability φ_1 on Θ is random, since it depends on the random variable X_1. We can ask, what is the expectation φ_1?

To answer this, let us treat the problem as a tree diagram problem, just as we did in Section 4.8. To keep things simple, I'll treat the case where Θ is finite, and $\mathcal{X}$ is discrete. Imagine

drawing θ from Θ according to the probability φ_0, and then drawing $\boldsymbol{x}$ from $\mathcal{X}$ according to the conditional probability $p(\boldsymbol{x} \mid \theta)$. The joint distribution of $(\boldsymbol{x}, \theta)$ is

$$p_0(\boldsymbol{x}, \theta) = p(\boldsymbol{x} \mid \theta)\varphi_0(\theta).$$

Let us write

$$f_0(\boldsymbol{x}) = \sum_{\theta \in \Theta} p(\boldsymbol{x} \mid \theta)\varphi_0(\theta),$$

the marginal probability of $\boldsymbol{x}$.

Then we may write the Bayesian updating formula as, for each $\theta \in \Theta$,

$$\varphi_1(\theta \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \theta)\varphi_0(\theta)}{f_0(\boldsymbol{x})}. \tag{1}$$

Now the expectation of the random variable $\varphi_1(\theta \mid X_1)$ is given by

$$\boldsymbol{E}\, \varphi_1(\theta \mid X_1) = \sum_{\boldsymbol{x} \in \mathcal{X}} \varphi_1(\theta \mid \boldsymbol{x}) f_0(\boldsymbol{x}),$$

so by (1)

$$\boldsymbol{E}\, \varphi_1(\theta \mid X_1) = \sum_{\boldsymbol{x} \in \mathcal{X}} \varphi_1(\theta \mid \boldsymbol{x}) f_0(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \frac{p(\boldsymbol{x}; \theta)\varphi_0(\theta)}{f_0(\boldsymbol{x})} f_0(\boldsymbol{x}) = \varphi_0(\theta) \underbrace{\sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}; \theta)}_{=1} = \varphi_0(\theta).$$

That is, the expected value of the posterior on θ is just the prior.

So if we continue, and observe X_2, and update $\varphi_1(\theta \mid X_1)$ to $\varphi_2(\theta \mid X_2$, we get a sequence of random variables

$$\varphi_1(\theta \mid X_1),\ \varphi_2(\theta \mid X_2),\ \varphi_3(\theta \mid X_3),\ \ldots\ .$$

The argument just given shows that this sequence is a martingale. If we take as our state variable the probability $\varphi_t(\cdot \mid X_t)$, the sequence is a Markov chain as well.

27.2.1 Remark Note that I do not say that my posterior will be the same as my prior. I am saying that my posterior is a random variable that depends on X_1, but the ex ante expected value of the posterior on θ is the same as my prior.

Since probabilities are bounded, the Martingale Convergence Theorem 15.8.1 implies that for each $\theta \in \Theta$

$$P\Big(\lim_{t \to \infty} \varphi_t(\theta \mid X_t) \text{ exists} \Big) = 1.$$

That is our posteriors will converge as we get more evidence.

But to what do the posteriors converge? Imagine that there is a true parameter θ^* such that each X_i has distribution $p(\boldsymbol{x} \mid \theta^*)$, and the X_is are independent then under mild regularity conditions (akin to those for conditions of maximum likelihood estimators),

$$\varphi_t(\cdot \mid X_t) \xrightarrow{\mathcal{D}} \delta_{\theta^*},$$

where δ_{θ^*} assigns probability one to θ^*. The proof of this is beyond the scope of this course, but see, e.g., Degroot [2, Chapter 10].

The arguments I have just given work with densities on Θ as well as the finite case, but require a bit more care in the arguments.

The preceding argument may seem a bit abstract, so here is a concrete example.

27.2.2 Example Urn 1 has 2 Black balls and 1 White ball. Urn 2 has 1 Black ball and 4 White balls. I believe that Urn 1 has been chosen with probability p_0 where $0 < p_0 < 1$.

What do I expect now (before any ball has been chosen) my belief p_1 to be after one sample is drawn from the Urn? There are two possible outcomes of the draw: B and W, and

$$P(B) = P(B \mid 1)P(1) + P(B \mid 2)P(2) = \frac{2}{3}p_0 + \frac{1}{5}(1 - p_0) = \frac{1}{5} + \frac{7}{15}p_0$$

and

$$P(W) = P(W \mid 1)P(1) + P(W \mid 2)P(2) = \frac{1}{3}p_0 + \frac{4}{5}(1 - p_0) = \frac{4}{5} - \frac{7}{15}p_0.$$

If B is the outcome, my posterior belief, applying Bayes' Law, is

$$P(1 \mid B) = P(B \mid 1)\frac{P(1)}{P(B)} = \frac{2}{3}\frac{p_0}{\frac{1}{5} + \frac{7}{15}p_0} = \frac{10p_0}{3 + 7p_0}$$

and if W is the outcome, my posterior belief, applying Bayes' Law, is

$$P(1 \mid W) = P(W \mid 1)\frac{P(1)}{P(W)} = \frac{1}{3}\frac{p_0}{\frac{4}{5} - \frac{7}{15}p_0} = \frac{5p_0}{12 - 7p_0}.$$

So

$$\begin{aligned}\boldsymbol{E}\left(p_1 \mid p_0\right) &= P(1 \mid B)P(B) + P(1 \mid W)P(W)\\ &= \frac{10p_0}{3 + 7p_0}\left(\frac{1}{5} + \frac{7}{15}p_0\right) + \frac{5p_0}{12 - 7p_0}\left(\frac{4}{5} - \frac{7}{15}p_0\right)\\ &= \frac{2}{3}p_0 + \frac{1}{3}p_0 = p_0.\end{aligned}$$

The same argument applies into the future. That is, the probability p_t, the probability I attach to urn 1 after t independent samples (with replacement) is a martingale. □

27.3 ⋆ Conjugate priors

An important tool for Bayesian statistics is that of a **conjugate prior**. A parametric family of distributions is conjugate to a likelihood function if the posterior belongs to the family whenever the prior does. In other words, the prior looks like the result of having seen a prior history of the data generating process.

Morris DeGroot [2] devotes Chapter 9 to conjugate priors. Here are a few examples.

27.3.1 Example (Binomial(n, p) Likelihood) The likelihood function for p given a sample of n independent Bernoulli(p) trials $X_1, \ldots, X_n$ can be based on the sufficient statistic $k = \sum_{i=1}^n X_i$:

$$L(p; k, n) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

If the prior density φ on p is a Beta(s, f) density

$$\varphi(p) \propto p^{s-1}(1 - p)^{f-1}$$

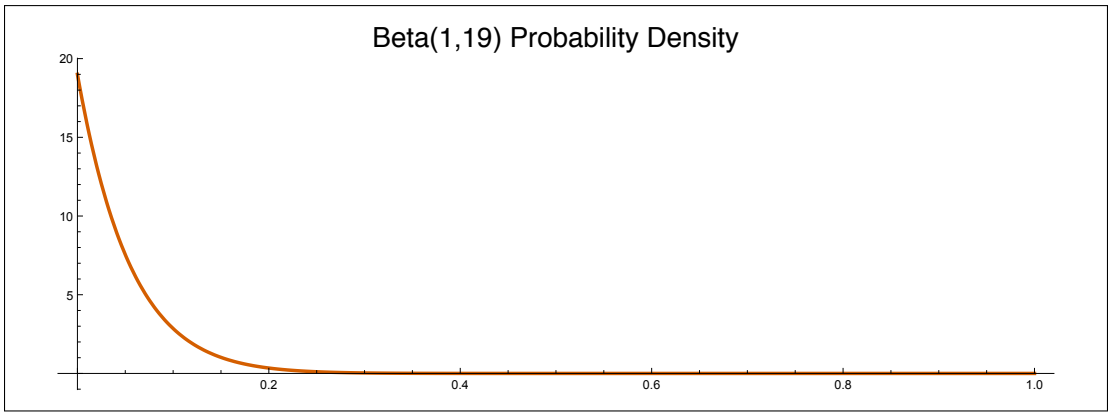
on $[0, 1]$, then the posterior density satisfies

$$\varphi(p \mid k) \propto p^k (1 - p)^{n-k}\, p^{s-1}(1 - p)^{f-1} = p^{s+k-1}(1 - p)^{f+(n-k)-1},$$

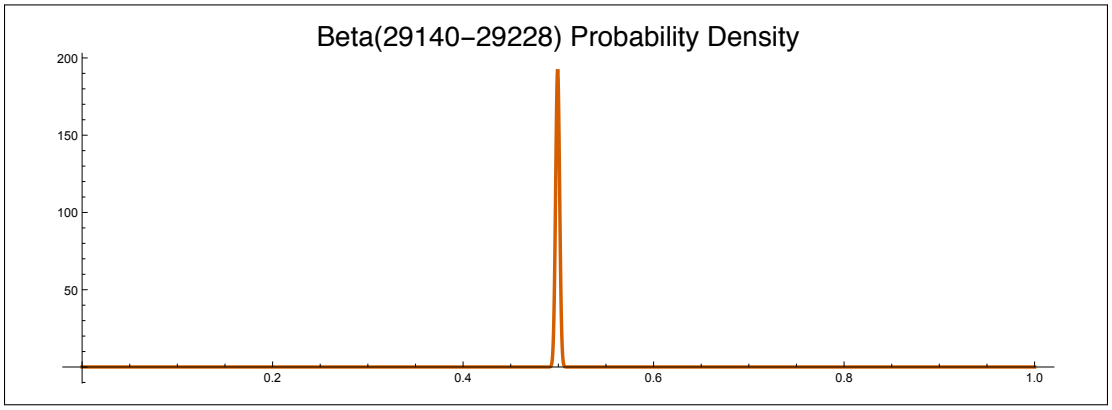
which is a Beta$(s + k, f + n - k)$ density. The interpretation of the parameters in the Beta distribution is this. If the expected number of successes is $s + 1$ and the expected number of

failures is $f+1$, then the density of p is Beta(s,f). The mean is $s/(s+f)$. So starting with a Beta(s,f) is like starting with a prior history of $s-1$ successes and $f-1$ failures. The Beta$(1,1)$ distribution is the uniform $U[0,1]$ distribution.

Suppose we start with a prior heavily biased toward 0, say Beta$(1,19)$ which has mean $1/20$ and density:

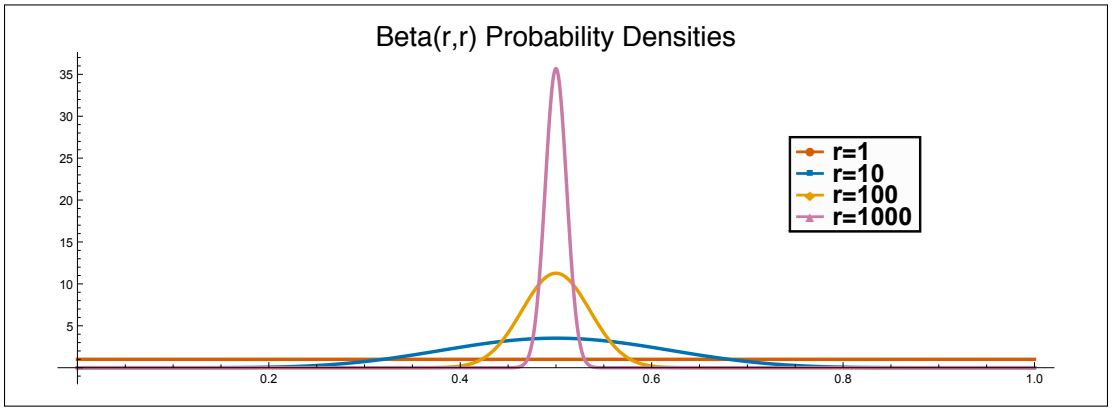


After the coin tossing data (29,140 successes and 29,228 failures) the posterior density is Beta$(29140, 29228)$:



That is, the data essentially drown out the prior.

The next figure shows the effect on the posterior of more data.



□

Here are a couple of other examples of conjugate priors, taken from DeGroot [2]. He describes many more examples, including the conjugate for a Normal with known standard deviation.

- **Exponential**(λ) (cf. DeGroot [2, p. 166])

Let $X_1, \dots, X_n$ be independent and identically distributed random variables with an Exponential density. The likelihood function in terms of the sufficient statistic $T = \sum_{i=1}^n X_i$ is

$$L(\lambda \,|\, T, n) \propto \lambda^n e^{-\lambda T}.$$

The conjugate prior φ for λ on $(0, \infty)$ is a Gamma(n_0, T_0) density with $n_0 > 0$, $T_0 > 0$,

$$\varphi(\lambda) \propto \lambda^{n_0 - 1} e^{-\lambda T_0},$$

and the posterior density for λ is a Gamma$(n_0 + n, T_0 + T)$ density,

$$\varphi(\lambda \,|\, k, n) \propto \lambda^{n_0 + n - 1} e^{-\lambda(T_0 + T)}.$$

- **Poisson**(μ) (cf. DeGroot [2, p. 164])

Let $X_1, \dots, X_n$ be independent and identically distributed random variables with a Poisson probability mass function, with unknown parameter μ, and let $k = \sum_{i=1}^n X_i$. The likelihood function satisfies

$$L(\mu; k, n) \propto \mu^k e^{-n\mu}.$$

The conjugate prior φ for μ on $(0, \infty)$ is a Gamma(k_0, n_0) density with $k_0 > 0$, $n_0 > 0$

$$\varphi(\mu) \propto \mu^{k_0 - 1} e^{-n_0 \mu},$$

and the posterior density for μ is a Gamma$(k_0 + k, n_0 + n)$ density,

$$\varphi(\mu \,|\, k, n) \propto \mu^{k_0 + k - 1} e^{-(n_0 + n)\mu}$$

27.4 Loss functions

Bayesian posteriors are often used to create point estimates by minimizing the posterior expected values of a **loss function**. A loss function L is a function of both the parameter and the estimate, satisfying $L(\hat{\theta}, \theta) \geqslant 0$ and $L(\theta, \theta) = 0$. The associated **risk function** is defined by

$$\int_\Theta L(\hat{\theta}, \theta) \varphi(\theta \,|\, \boldsymbol{x})\, d\theta.$$

When $\hat{\theta}$ is chosen to minimize the risk, it is called a **Bayesian estimate**.

Aside: It is unfortunate that L is used to denote the loss function in this context, and it is used to denote the likelihood function in other contexts. What can I say? Decision theorists and economists frequently use **utility functions**, which are measures of gains rather than losses, and maximize expected utility rather than minimize expected loss.

27.4.1 Proposition (Larsen and Marx [7, Theorem 5.8.1, pp. 342–344])

a. *When $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, the risk minimizing $\hat{\theta}$ is the median of $\varphi(\theta \,|\, \boldsymbol{x})$.*

b. *When the loss is the square error, $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, the risk minimizing $\hat{\theta}$ is the mean of g.*

Proof: a. The risk function is given by

$$\int |\hat{\theta} - \theta| \varphi(\theta \,|\, \boldsymbol{x})\, d\theta = \int \Big((\hat{\theta} - \theta) \mathbf{1}_{(-\infty, \hat{\theta}]}(\theta) + (\theta - \hat{\theta}) \mathbf{1}_{[\hat{\theta}, \infty)}(\theta) \Big) \varphi(\theta \,|\, \boldsymbol{x})\, d\theta$$

Differentiating with respect to $\hat{\theta}$ gives

$$\int \Big(\mathbf{1}_{(-\infty, \hat{\theta}]}(\theta) - \mathbf{1}_{[\hat{\theta}, \infty)}(\theta) \Big) \varphi(\theta \,|\, \boldsymbol{x})\, d\theta$$

which is negative for $\hat{\theta} <$ median, positive for $\hat{\theta} >$ median, equal to zero at the median of the posterior.

b. The risk function is given by

$$\int (\hat{\theta} - \theta)^2 \varphi(\theta \,|\, \boldsymbol{x})\, d\theta.$$

The first order condition for a minimum is obtained by differentiating under the integral sign to get: $\int 2(\hat{\theta} - \theta) \varphi(\theta \,|\, \boldsymbol{x})\, d\theta = 0$. The solution is $\hat{\theta} = \int \theta \varphi(\theta \,|\, \boldsymbol{x})\, d\theta$. That is, $\hat{\theta}$ is the posterior mean.

∎

27.5 ⋆ Appendix: The Bayesian Bookie

27.5.1 Statistical inference: the game

Freedman and Purves [4] caricature statistical inference in terms of the following game.

1. The Master of Ceremonies chooses an urn θ_0 from a set Θ of urns, draws a sample x from the urn according to the probability measure $p_\theta(x)$, and exhibits the sample to the Bettor and the Bookie.

2. A Bookie posts prices q for lottery tickets that pay off 1 unit in case θ is the urn, for each $\theta \in \Theta$. The Bookie must buy and/or sell tickets at these prices.

3. The Bettor buys a portfolio of lottery tickets. The Bettor may also sell tickets to the Bookie at the same price the Bookie sells them.

4. The MC reveals the urn θ_0, and the tickets are payed off.

The reason this is a caricature is that in the real world of statistical inference, there is never an MC to reveal θ_0.

27.5.2 Strategies

The Bettor and the Bookie choose their strategies for the game in advance of playing, so they must decide what to do for each possible sample that could be observed.

The Bookie chooses $q \geqq 0 \in \boldsymbol{R}^{\Theta \times \mathcal{X}}$. For each $x \in \mathcal{X}$ and $\theta \in \Theta$, $q(\theta, x)$ is the price he sets, after having seen the sample x, for a lottery ticket that pays \$1 if the chosen urn is θ.

Then Bettor then chooses wagers $w \in \boldsymbol{R}^{\Theta \times \mathcal{X}}$, and buys $w(\theta, x)$ θ-tickets, after having the seen the sample x and the prices q.

Under these strategies, the expected payoff to the Bettor when θ is the selected urn is just

$$\sum_{x \in \mathcal{X}} \left(\sum_{t \in \Theta} \big(\mathbf{1}_t(\theta) - q(t, x) \big) w(t, x) \right) p_\theta(x).$$

27.5.1 Bayesian updating theorem *Either*
(i) *The Bookie chooses some prior P on Θ and sets prices according to the posterior $P(\theta|x)$,*

$$P(\theta|x) = \frac{p_\theta(x)P(\theta)}{\sum_{t\in\Theta} p_t(x)P(t)}.$$

Or else
(ii) *There is a betting strategy that gives the Bettor a positive expected payoff regardless of which urn θ is selected by the MC.*

Note that this result does not say that the MC actually selected the urn at random according to P—it is merely a device to calculate the prices to avoid (ii).

Proof: Condition (ii) is equivalent to the matrix inequality

$$\begin{matrix} \\ t \end{matrix}\!\!\begin{bmatrix} & \overset{(\theta,x)}{\vdots} & \\ \cdots & \big(\mathbf{1}_\theta(t) - q(\theta,x)\big)p_t(x) & \cdots \\ & \vdots & \end{bmatrix}\begin{bmatrix} \vdots \\ w(\theta,x) \\ \vdots \end{bmatrix} \gg 0,$$

where rows are indexed by $t \in \Theta$ and columns are indexed by $(\theta, x) \in \Theta \times \mathcal{X}$.

Gordan's Alternative 27.5.2 below, asserts that the alternative to (ii) is the existence of a probability vector $P \in \boldsymbol{R}^\Theta$ such that for each column $(\theta, x) \in \Theta \times \mathcal{X}$,

$$\sum_{t\in\Theta}\big(\mathbf{1}_\theta(t) - q(\theta,x)\big)p_t(x)P(t) = 0.$$

In other words,

$$p_\theta(x)P(\theta) = \sum_{t\in\Theta} q(\theta,x)p_t(x)P(t),$$

or

$$q(\theta,x) = \frac{p_\theta(x)P(\theta)}{\sum_{t\in\Theta} p_t(x)P(t)} = P(\theta|x),$$

which is (i). ∎

The proof relied on this result due to P. Gordan [6], which is a form of a **Theorem of the Alternative**. See David Gale [5, Chapter 2] or my on-line notes for a proof.

27.5.2 Gordan's Alternative *Let A be an $m \times n$ matrix. Exactly one of the following alternatives holds. Either there exists $x \in \boldsymbol{R}^{\mathrm{n}}$ satisfying*

$$Ax \gg 0. \tag{2}$$

or else there exists $p \in \boldsymbol{R}^{\mathrm{m}}$ satisfying

$$\begin{aligned} pA &= 0 \\ p &> 0. \end{aligned} \tag{3}$$

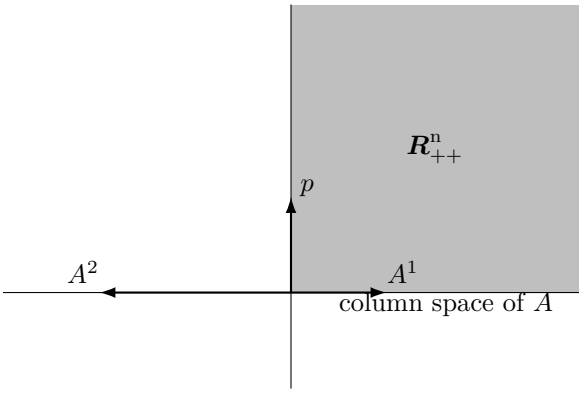


Figure 27.1. Geometry of the Gordan Alternative

Bibliography

[1] H. Chernoff. 1986. [Why isn't everyone a Bayesian?]: Comment. *American Statistician* 40(1):5–6. http://www.jstor.org/stable/2683106

[2] M. H. DeGroot. 1970. *Optimal statistical decisions*. New York: McGraw-Hill.

[3] B. Efron. 1986. Why isn't everyone a Bayesian? *American Statistician* 40(1):1–5. http://www.jstor.org/stable/2683105

[4] D. A. Freedman and R. A. Purves. 1969. Bayes' method for bookies. *Annals of Mathematical Statistics* 40(4):1177–1186. http://www.jstor.org/stable/2239586

[5] D. Gale. 1989. *Theory of linear economic models*. Chicago: University of Chicago Press. Reprint of the 1960 edition published by McGraw-Hill.

[6] P. Gordan. 1873. Über die auflösung linearer Gleichungen mit reelen Coefficienten [On the solution of linear inequalities with real coefficients]. *Mathematische Annalen* 6(1):23–28. DOI: 10.1007/BF01442864

[7] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[8] D. V. Lindley. 1986. [Why isn't everyone a Bayesian?]: Comment. *American Statistician* 40(1):6–7. http://www.jstor.org/stable/2683107

[9] ——— . 1997. [Bayes for beginners? some reasons to hesitate]: Discussion. *American Statistician* 51(3):265–266. http://www.jstor.org/stable/2684900