Il Newsvendor e le sue varianti

Indice

Motivazione gestionale

Il modello base

Esempio svolto a mano (domanda discreta)

Caso di studio: la regola del quantile in continuo

La formulazione lineare a scenari

Livelli di servizio

Avversione al rischio: media-CVaR

Multiprodotto con budget condiviso

Esercizi

Obiettivi (1/2)

Alla fine di questo capitolo sarai in grado di:

- ricavare e applicare la regola del quantile;
- trasformare un problema stocastico in un LP a scenari con il trucco della parte positiva;

Obiettivi (2/2)

Alla fine di questo capitolo sarai in grado di:

- quantificare il valore di modellare l'incertezza (VSS);
- incorporare livelli di servizio e avversione al rischio.

Prerequisiti

- Da questo modulo, Capitolo 2 «Richiami: programmazione lineare e analisi di sensitività»: i prezzi ombra e i loro intervalli di validità
- Da questo modulo, Capitolo 3 «Il solver: costruire, risolvere, interpretare»: la costruzione e la soluzione di un LP in `gurobipy`

Motivazione gestionale

Introduzione: Motivazione gestionale

- Il Newsvendor ("venditore di giornali"): scegliere una quantità *prima* di osservare la domanda
- Classe del problema: **convesso 1D / LP stocastico a scenari**
- È la porta d'ingresso dell'**ottimizzazione stocastica**

Una decisione comunissima

- Ordinare troppo genera **invenduto**; troppo poco, **vendite perse**
- Si applica a moda e stagionali, freschi e deperibili, farmaci, capacità alberghiera, componenti di ricambio
- Il suo punto di forza didattico è un risultato netto:

la quantità ottima non è la domanda media

Si decide prima e si scopre poi

- Il modello ha più di un secolo e resta il fondamento della gestione delle scorte per prodotti a ciclo singolo
- È il primo incontro con l'idea che *si decide prima e si scopre poi*
- E con gli strumenti per farlo bene: scenari, valore della soluzione stocastica, misure di rischio

Riepilogo: Motivazione gestionale

- Una quantità scelta prima di conoscere la domanda
- Due errori possibili: invenduto e vendite perse
- La quantità ottima **non** è la domanda media

Il modello base

Introduzione: Il modello base

- Formuliamo il Newsvendor come minimizzazione di un costo atteso
- Una sola variabile: la quantità ordinata q
- Ricaviamo la **regola del quantile**, che dà l'ottimo in forma chiusa

Il problema a parole

- *Decidiamo* una sola quantità q, prima di conoscere la domanda
- *L'obiettivo* è minimizzare il costo atteso complessivo dell'errore: aver ordinato troppo (invenduto) più aver ordinato troppo poco (vendite perse)
- *Vincoli* non ce ne sono nel modello base ($q \geq 0$)
- Compaiono nelle varianti: budget condiviso, livelli di servizio, limiti di capacità

Dati (input del modello) (1/2)

Simbolo	Tipo	Significato
D	v. a. ≥ 0	domanda, con funzione di ripartizione F
c_u	$\in \mathbb{Q}_{>0}$	costo unitario di *underage* (domanda non servita): margine perso più eventuale penalità (€)
c_o	$\in \mathbb{Q}_{>0}$	costo unitario di *overage* (unità invenduta): costo meno valore di recupero (€)

Dati (input del modello) (2/2)

- D è un'eccezione alla convenzione sulle maiuscole, per aderenza alla letteratura: variabile aleatoria maiuscola, realizzazioni minuscole (d_s)
- Con prezzo di vendita p, costo d'acquisto c e valore di recupero $v \leq c$ (tutti in $\mathbb{Q}_{>0}$):

$$c_u = p - c \qquad\qquad c_o = c - v$$

Variabile decisionale e modello

Introduciamo una variabile non negativa:

$$q = \text{quantità ordinata } \textit{prima} \text{ di osservare la domanda } D$$

Modello Newsvendor in forma di costo

$$\min \quad c_o \, \mathbb{E}\big[(q - D)^+\big] + c_u \, \mathbb{E}\big[(D - q)^+\big] \tag{7.1a}$$

$$\text{soggetto a} \quad q \geq 0. \tag{7.1b}$$

Funzione obiettivo e vincolo

- La funzione obiettivo è **convessa** e minimizza il costo atteso dell'errore:
  - eccedenza $(q - D)^+$ pagata c_o
  - carenza $(D - q)^+$ pagata c_u
- Il simbolo $(x)^+ = \max\{x, 0\}$ indica la parte positiva
- Il vincolo $q \geq 0$ definisce la variabile del modello

La regola del quantile

Se F è continua, l'ottimo soddisfa la **regola del quantile**:

$$F(q^*) = \alpha^* = \frac{c_u}{c_u + c_o} \qquad \Longrightarrow \qquad \boxed{q^* = F^{-1}(\alpha^*)}$$

Da dove viene la regola del quantile

- Si ragiona sull'unità marginale: ordinare la q-esima unità in più
  - rende c_u se la domanda la assorbe (probabilità $1 - F(q)$)
  - costa c_o se resta invenduta (probabilità $F(q)$)
- Conviene finché $c_u\,(1 - F(q)) \geq c_o\,F(q)$, cioè finché

$$F(q) \leq \frac{c_u}{c_u + c_o}$$

- Carenza molto più costosa dell'eccedenza: α^* alto, si ordina un quantile alto
- Invenduto costoso (deperibili): si ordina poco

Riepilogo: Il modello base

- Costo atteso $c_o\,\mathbb{E}[(q-D)^+] + c_u\,\mathbb{E}[(D-q)^+]$, convesso in q
- $c_u = p - c$ (carenza), $c_o = c - v$ (eccedenza)
- Regola del quantile: $q^* = F^{-1}\big(c_u/(c_u + c_o)\big)$

Esempio svolto a mano (domanda discreta)

Introduzione: Esempio svolto a mano

- Tre scenari di domanda equiprobabili
- Confrontiamo il costo atteso dei candidati naturali
- Verifichiamo che la regola del quantile (versione discreta) dà lo stesso risultato

Esempio: tre scenari equiprobabili (1/3)

Dati: $D \in \{80, 100, 120\}$ con probabilità $1/3$; $c_u = 9$, $c_o = 4$, quindi

$$\alpha^* = 9/13 = 0{,}6923$$

Passo 1 — costo atteso per i tre candidati naturali.

$$C(80) = \tfrac{1}{3}\left[0 + 9 \cdot 20 + 9 \cdot 40\right] = 180$$

$$C(100) = \tfrac{1}{3}\left[4 \cdot 20 + 0 + 9 \cdot 20\right] = \tfrac{260}{3} = 86{,}7$$

$$C(120) = \tfrac{1}{3}\left[4 \cdot 40 + 4 \cdot 20 + 0\right] = 80$$

Esempio: tre scenari equiprobabili (2/3)

Passo 2 — regola del quantile discreta.

- Serve il più piccolo q con $F(q) \geq 0{,}6923$:

$$F(80) = 1/3$$
$$F(100) = 2/3 = 0{,}667 < 0{,}6923$$
$$F(120) = 1$$

- Quindi $q^* = 120$: coerente con il passo 1

Esempio: tre scenari equiprobabili (3/3)

Passo 3 — lettura.

- Ordinare il *massimo* storico nonostante la media sia 100
- Con $c_u = 9 \gg c_o = 4$, restare corti costa più del doppio che restare lunghi
- Chi ordina "la media" paga 86,7 contro 80
- L'incertezza si gestisce, non si ignora

Riepilogo: Esempio svolto a mano

- Con $\alpha^* = 0{,}6923$ la quantità ottima è $q^* = 120$, il massimo storico
- Ordinare la media costa 86,7 contro 80
- Regola del quantile discreta: il più piccolo q con $F(q) \geq \alpha^*$

Caso di studio: la regola del quantile in continuo

Introduzione: la regola del quantile in continuo

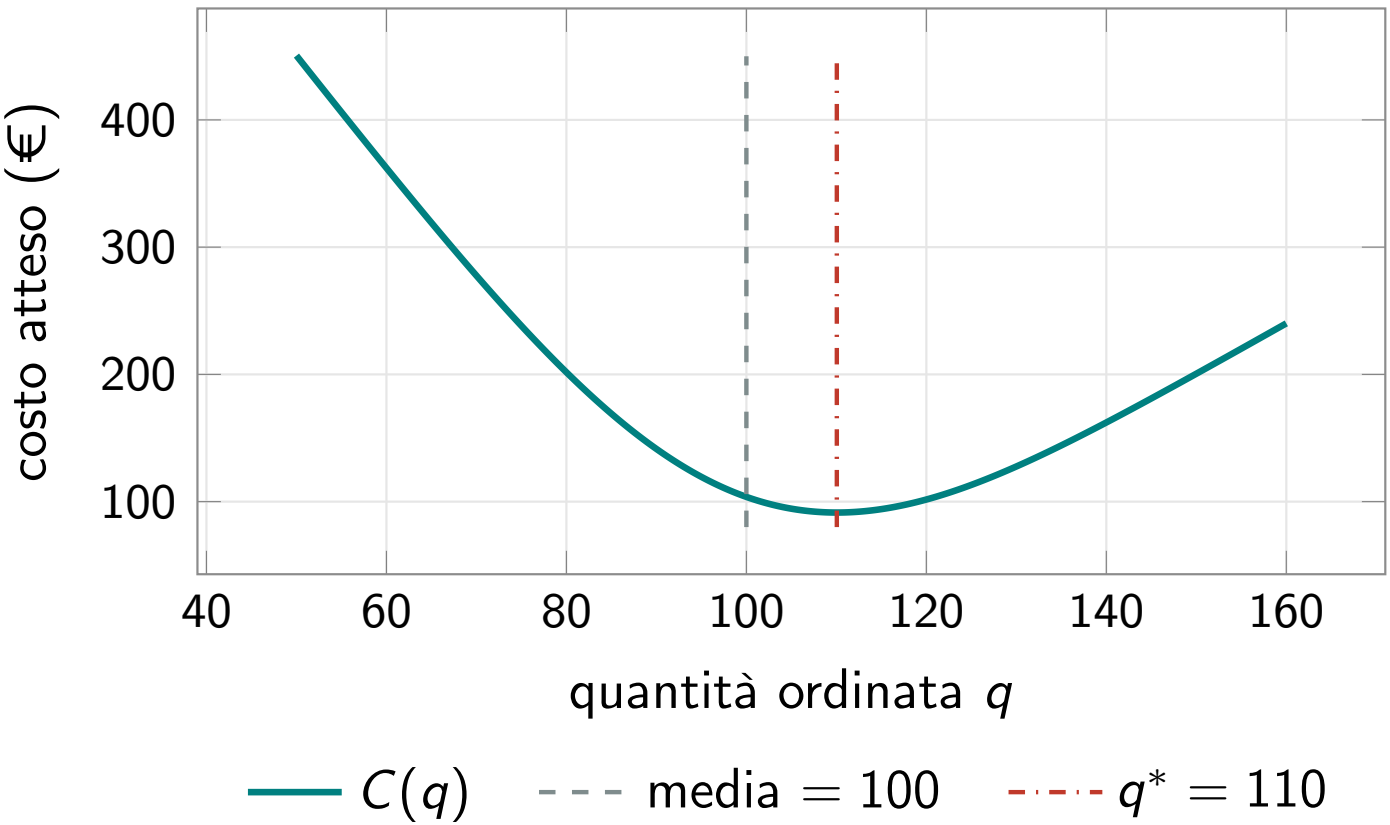
- Una panetteria: $p = 15$, $c = 6$, $v = 2$ € $\Rightarrow c_u = 9$, $c_o = 4$
- Domanda normale con $\mu = 100$, $\sigma = 20$
- Applichiamo la regola del quantile e vediamo quanto è stabile la soluzione stimata da scenari

Risultati

Panetteria: $p = 15$, $c = 6$, $v = 2$ € $\Rightarrow c_u = 9$, $c_o = 4$; domanda normale con $\mu = 100$, $\sigma = 20$.

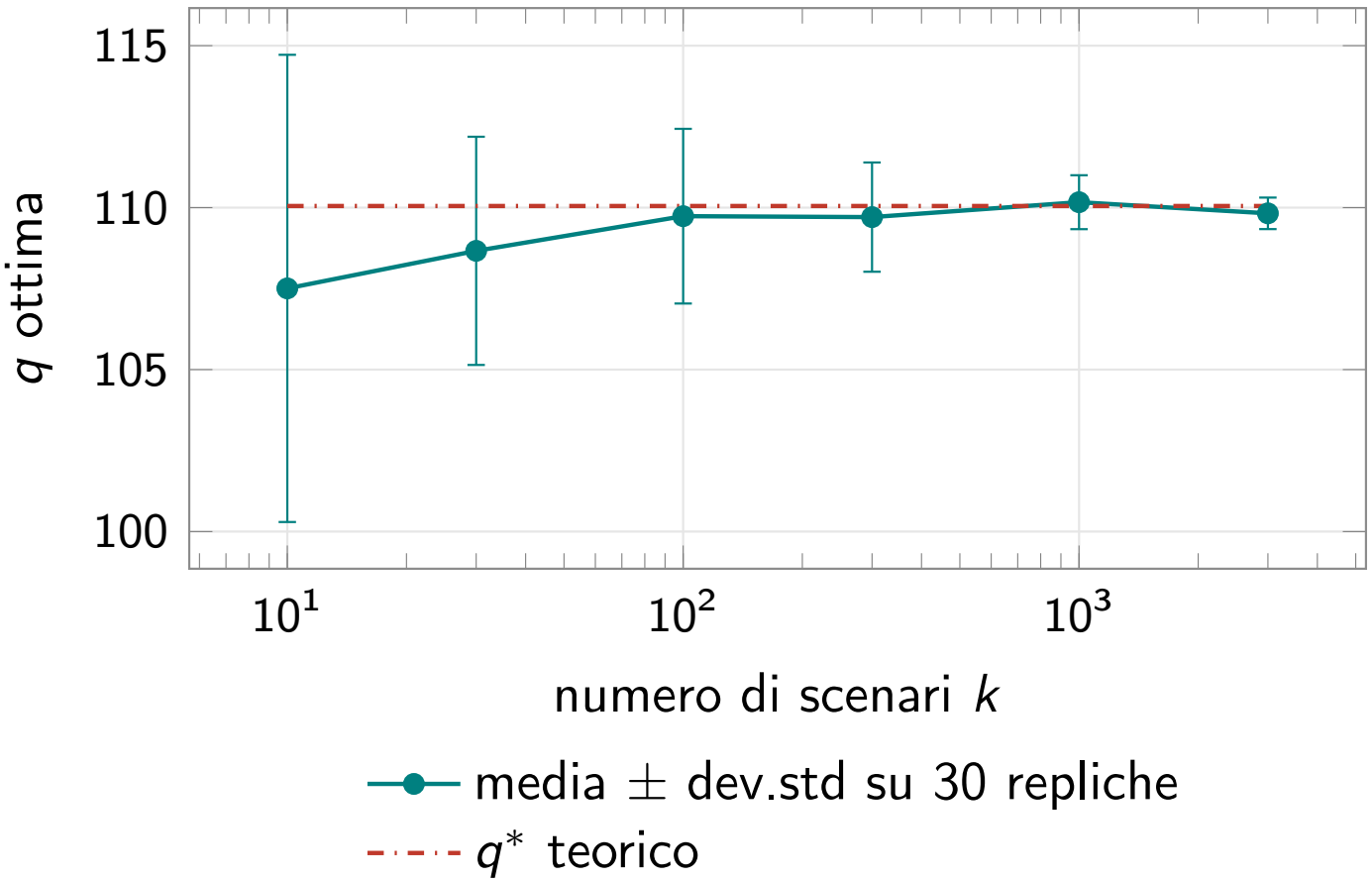
```
alpha* = 9/13 = 0,6923
q* = F^-1(0,6923) = 110,05 unita'   (media = 100)
costo atteso in q*: 91,43 EUR    in q = media: 103,72 EUR
```

Il minimo è al 69° percentile



Il costo atteso è convesso; il minimo cade in $q^* = 110$, non sulla media 100.

Stabilità della soluzione a scenari



Lettura dei grafici

- Il costo atteso $C(q)$ è convesso e il minimo cade in $q^* = 110$ (69° percentile della domanda), non sulla media 100
- A destra della media si paga meno: in q^* il costo atteso è 91,43 € contro 103,72 € nella media
- La q ottima stimata da k scenari campionati (media e deviazione standard su 30 repliche) mostra che con pochi scenari la decisione è **instabile**

Riepilogo: la regola del quantile in continuo

- $\alpha^* = 0{,}6923$, $q^* = 110{,}05$ contro una media di 100
- Ordinare la media costa 12,29 € in più a ciclo
- Le soluzioni stimate da pochi scenari oscillano molto

La formulazione lineare a scenari

Introduzione: La formulazione lineare a scenari

- Quando F non è nota in forma chiusa — il caso normale in azienda — si usano gli **scenari**
- Dati storici, campioni da una distribuzione stimata, o scenari pesati
- Otteniamo un **LP** grazie al trucco delle parti positive
- Misuriamo il valore della soluzione stocastica (VSS)

Nuovi dati e variabili

- Numero di scenari $k \in \mathbb{Z}_{\geq 1}$, indicizzati da $s \in \{1, 2, \ldots, k\}$
- Per ogni scenario: domanda $d_s \in \mathbb{Q}_{\geq 0}$ e probabilità $\pi_s \in \mathbb{Q}_{\geq 0}$, con $\sum_{s=1}^{k} \pi_s = 1$
- Oltre a q, introduciamo $2\,k$ variabili non negative:

$$\begin{cases} o_s = \text{eccedenza (unità invendute) nello scenario } s \\ u_s = \text{carenza (domanda non servita) nello scenario } s \end{cases}$$

per ogni $s \in \{1, 2, \ldots, k\}$

Il modello LP

Modello LP del Newsvendor a scenari

$$\min \quad \sum_{s=1}^{k} \pi_s \left(c_o \, o_s + c_u \, u_s \right) \tag{7.2a}$$

$$\text{soggetto a} \quad o_s \geq q - d_s, \quad \forall s \in \{1, 2, \ldots, k\}, \tag{7.2b}$$

$$u_s \geq d_s - q, \quad \forall s \in \{1, 2, \ldots, k\}, \tag{7.2c}$$

$$q \geq 0, \tag{7.2d}$$

$$o_s, \; u_s \geq 0, \qquad \forall s \in \{1, 2, \ldots, k\}. \tag{7.2e}$$

Funzione obiettivo e vincoli

- La funzione obiettivo lineare è la media pesata, sugli scenari, dei costi di eccedenza e di carenza: la versione "a scenari" del costo atteso
- I vincoli $o_s \geq q - d_s$ e $u_s \geq d_s - q$ definiscono eccedenza e carenza di ogni scenario a partire dall'unica vera decisione q
- Insieme alle non negatività realizzano le parti positive ($2\,k$ vincoli lineari)
- $q \geq 0$ e $o_s, u_s \geq 0$ definiscono le variabili

Perché è lineare: il trucco delle parti positive

- $(q - d_s)^+$ non è lineare
- All'*ottimo* la variabile o_s, che compare nell'obiettivo con coefficiente positivo, si schiaccia sul massimo tra $q - d_s$ e 0
- Il vincolo $o_s \geq q - d_s$ insieme a $o_s \geq 0$ fa esattamente il lavoro di $\max\{q - d_s, 0\}$ senza variabili binarie
- Stesso meccanismo per u_s
- Il trucco torna identico nel **CVaR** e nella **SVM**

Il modello in Gurobi

```
q = m.addVar(name="q")
o = m.addVars(S, name="o")      # eccedenza per scenario
u = m.addVars(S, name="u")      # carenza per scenario
m.addConstrs((o[s] >= q - dom[s] for s in range(S)))
m.addConstrs((u[s] >= dom[s] - q for s in range(S)))
m.setObjective(gp.quicksum((Co*o[s] + Cu*u[s]) / S
                           for s in range(S)), GRB.MINIMIZE)
```

Risultati

```
LP con k = 600 scenari: q = 108,61
quantile empirico al 69,23%: 108,60   (coincidono: e' un teorema)
VSS: decisione q = E[D] = 100 costa 99,50
     q stocastica costa 88,33
     valore della soluzione stocastica = 11,17 EUR per ciclo
```

Insight manageriale (1/2)

- Il *valore della soluzione stocastica* (VSS) risponde alla domanda del capo: “perché complicarsi la vita con gli scenari?”
- Qui vale 11,17 € a ciclo, l’11% del costo
- È il guadagno di *modellare* l’incertezza invece di sostituirla con la media

Insight manageriale (2/2)

- La figura sulla stabilità dà l'altra metà del messaggio
- Sotto i 100 scenari la q ottima balla di ± 3 unità tra una stima e l'altra
- Gli scenari sono anch'essi un campione, e la decisione eredita la loro varianza

Riepilogo: La formulazione lineare a scenari

- Con k scenari il Newsvendor diventa un LP con $2k+1$ variabili
- Le parti positive si linearizzano senza variabili binarie
- L'LP riproduce il quantile empirico (è un teorema)
- VSS $= 11{,}17$ € a ciclo: modellare l'incertezza conviene

Livelli di servizio

Introduzione: Livelli di servizio

- Spesso i contratti impongono un **livello di servizio**
- *Cycle service level*: probabilità di non andare in stock-out
- *Fill rate*: quota di domanda servita in media
- Vediamo quanto costano e perché non vanno confusi

Il costo dei livelli di servizio

```
cycle service level 90%: q = 125,63
                         (costo +23,41 EUR sull'ottimo economico)
cycle service level 95%: q = 132,90  (costo +45,59)
cycle service level 99%: q = 146,53  (costo +95,56)
fill rate in q*: 96,06%
probabilita' di non-stockout in q*: 69,23%
```

Non confondere i livelli di servizio

Osservazione 1: Non confondere i livelli di servizio

In $q^* = 110$ la probabilità di evitare lo stock-out è solo il 69%, ma il fill rate (quota di domanda servita in media) è il 96%: misure diverse che i contratti spesso confondono.

- Promettere "95% di probabilità di copertura totale" costa $+45{,}59$ € a ciclo rispetto all'ottimo economico
- Promettere "95% di fill rate" è quasi gratis
- Leggere bene lo SLA prima di firmarlo

Riepilogo: Livelli di servizio

- Alzare il cycle service level costa sempre di più: $+23$, $+46$, $+96$ € per 90%, 95%, 99%
- In q^*: non-stockout 69%, ma fill rate 96%
- Due misure diverse: leggere bene lo SLA

Avversione al rischio: media-CVaR

Introduzione: Avversione al rischio

- Il costo atteso non distingue tra tante piccole perdite e rari disastri
- Con la formulazione lineare del CVaR (qui solo anticipata) minimizziamo

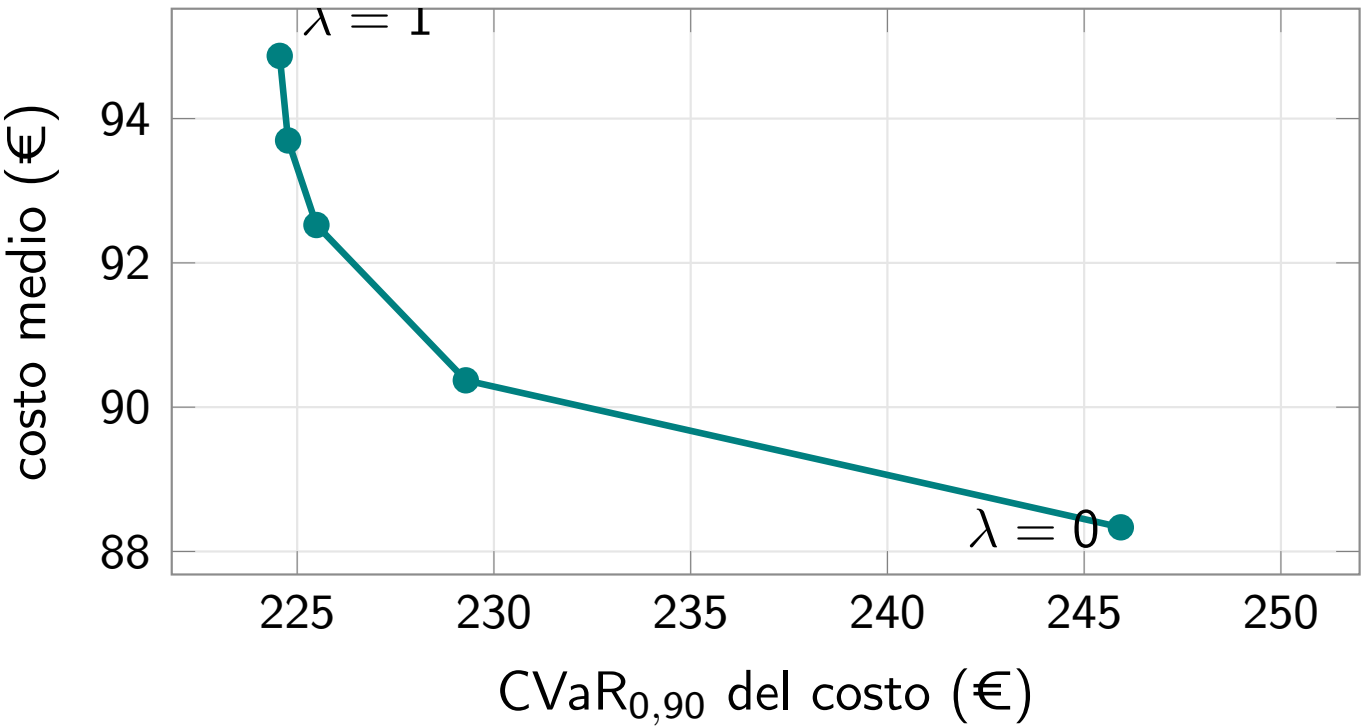
$$(1-\lambda)\,\mathbb{E}[\text{costo}] + \lambda\,\mathrm{CVaR}_{0,90}(\text{costo})$$

- Il peso $\lambda \in [0,1]$ misura l'avversione al rischio

Risultati al variare di λ

```
lambda = 0,00: q = 108,61  costo medio 88,33  CVaR90 245,93
lambda = 0,25: q = 112,45  costo medio 90,37  CVaR90 229,29
lambda = 0,50: q = 114,48  costo medio 92,52  CVaR90 225,49
lambda = 1,00: q = 116,17  costo medio 94,87  CVaR90 224,56
```

Frontiera costo medio – rischio di coda



Lettura della frontiera

- Da $\lambda = 0$ (ottimo in media, coda pesante) a $\lambda = 1$ (protezione massima)
- Il primo tratto è quasi verticale: molta protezione per pochissimo costo medio
- Il decisore avverso al rischio ordina di più (da 108,6 a 116,2 unità)
- Paga 6,5 € in più in media per tagliare di 21,4 € il costo medio dei peggiori 10% degli scenari
- Gran parte della protezione si ottiene già con $\lambda = 0{,}25$

Riepilogo: Avversione al rischio

- Obiettivo media-CVaR: compromesso tra costo atteso e rischio di coda
- Più avversione al rischio $\Rightarrow$ si ordina di più
- La frontiera è ripida all'inizio e piatta poi: basta poca avversione per molta protezione

Multiprodotto con budget condiviso

Introduzione: Multiprodotto con budget condiviso

- Tre dolci con domande *correlate* ($\rho = 0{,}7$): le feste vanno bene o male per tutti insieme
- Un budget di produzione di 1200 € condiviso
- Vediamo come il budget comprime le quantità e cosa dice il duale

Risultati

```
  prodotto   | q senza budget | q con budget
 panettone   |          111,1 |         99,3
 pandoro     |           91,5 |         75,4
 torrone     |           65,9 |         56,8
Spesa: 1200/1200   prezzo ombra del budget: -0,553
```

Lettura dei risultati (1/2)

- Il budget vincolante comprime le tre quantità sotto i rispettivi quantili ottimi
- Ma non proporzionalmente: il taglio dipende dal rapporto c_u/c_o e dal costo unitario di ciascun prodotto
- Il duale dice che un euro di budget in più ridurrebbe il costo atteso di 0,55 €

Lettura dei risultati (2/2)

- Un rendimento del 55% che giustificherebbe quasi qualunque finanziamento
- Informazione che senza il modello nessuno avrebbe quantificato
- La correlazione alta toglie il beneficio di diversificazione: i tre prodotti falliscono insieme
- Trattarli come indipendenti sottostimerebbe il rischio

Riepilogo: Multiprodotto con budget condiviso

- Un budget condiviso taglia le quantità in modo non proporzionale
- Il prezzo ombra ($-0{,}553$) misura il valore di un euro di budget in più
- Con domande correlate non c'è diversificazione

Esercizi

Introduzione: Esercizi

- Sei esercizi per consolidare i concetti del capitolo
- Regola del quantile, dati storici, deperibili, EVPI, fill rate, multiprodotto
- Si risolvono a mano e con lo script del laboratorio

Esercizio: verifica

- Il Comune impone una penalità $b = 3$ € per cliente non servito:

$$c_u = p - c + b$$

- Ricalcolare a mano α^* e q^*
- Verificare con lo script

Esercizio: dati storici

- Usare come scenari le 24 osservazioni di domanda degli ultimi due anni (generarle con lo script) invece della distribuzione
- Quanto differisce q?
- Quale approccio consiglierebbe a un'azienda che ha solo 24 osservazioni?

Esercizio: deperibili

- Il recupero diventa negativo: $v = -1$ € (smaltire costa)
- Come cambiano α^* e q^*?
- Tracciare q^* in funzione di $v \in [-2, 4]$

Esercizio: EVPI

- Calcolare l'EVPI (*Expected Value of Perfect Information*, valore atteso dell'informazione perfetta)
- Costo medio con q scelta *dopo* aver visto la domanda (cioè $q = d_s$, costo 0) contro costo della soluzione stocastica
- Quanto vale al massimo una previsione perfetta della domanda?

Esercizio: fill rate vincolato

- Aggiungere all'LP il vincolo di fill rate

$$\sum_{s=1}^{k} \pi_s u_s \leq (1 - 0{,}98)\, \mathbb{E}[D]$$

- Confrontare quantità e costo con il cycle service level al 98%

Esercizio: multiprodotto

- Rieseguire il multiprodotto con domande indipendenti ($\rho = 0$) a parità di tutto il resto
- Il costo atteso ottimo migliora o peggiora?
- Perché?

Riepilogo (1/2)

Modello base:

- Si decide q prima di osservare la domanda
- Regola del quantile: $q^* = F^{-1}\big(c_u/(c_u + c_o)\big)$, non la media

Formulazione a scenari:

- LP con variabili di eccedenza e carenza (trucco delle parti positive)
- VSS: il guadagno di modellare l'incertezza
- Pochi scenari $\Rightarrow$ decisione instabile

Riepilogo (2/2)

Varianti:

- Livelli di servizio: cycle service level e fill rate sono misure diverse
- Media-CVaR: chi è avverso al rischio ordina di più; poca avversione dà molta protezione
- Budget condiviso: il prezzo ombra misura il valore di un euro in più