The Newsvendor and its variants

Outline

Managerial motivation

The basic model

Worked example by hand (discrete demand)

Case study: the quantile rule in the continuous case

The scenario-based linear formulation

Service levels

Risk aversion: mean-CVaR

Multi-product with a shared budget

Exercises

Learning objectives (1/2)

At the end of this chapter you will be able to:

- derive and apply the quantile rule;
- turn a stochastic problem into a scenario-based LP with the positive-part trick;

Learning objectives (2/2)

At the end of this chapter you will be able to:

- quantify the value of modelling uncertainty (VSS);
- incorporate service levels and risk aversion.

Prerequisites

- From this module, Chapter 2 «Background: linear programming and sensitivity analysis»: shadow prices and their validity ranges
- From this module, Chapter 3 «The solver: building, solving, interpreting»: building and solving an LP in `gurobipy`

Managerial motivation

Introduction: Managerial motivation

- The Newsvendor ("newspaper seller"): choosing a quantity *before* observing demand
- Class of the problem: **1D convex / scenario-based stochastic LP**
- It is the gateway to **stochastic optimization**

A very common decision

- Ordering too much generates **unsold items**; too little, **lost sales**
- It applies to fashion and seasonal goods, fresh and perishable products, drugs, hotel capacity, spare parts
- Its didactic strength is a clear-cut result:

the optimal quantity is not the average demand

We decide first and find out later

- The model is more than a century old and remains the foundation of inventory management for single-cycle products
- It is the first encounter with the idea that *we decide first and find out later*
- And with the tools needed to do it well: scenarios, value of the stochastic solution, risk measures

Summary: Managerial motivation

- A single quantity chosen before knowing demand
- Two possible errors: unsold items and lost sales
- The optimal quantity is **not** the average demand

The basic model

Introduction: The basic model

- We formulate the Newsvendor as the minimization of an expected cost
- A single variable: the order quantity q
- We derive the **quantile rule**, which gives the optimum in closed form

The problem in words

- *We decide* a single quantity q, before knowing demand
- *The objective* is to minimize the total expected cost of the error: having ordered too much (unsold items) plus having ordered too little (lost sales)
- *Constraints* are absent in the basic model ($q \geq 0$)
- They appear in the variants: shared budget, service levels, capacity limits

Data (input of the model) (1/2)

Symbol	Type	Meaning
D	random var. ≥ 0	demand, with cumulative distribution function F
c_u	$\in \mathbb{Q}_{>0}$	unit *underage* cost (unmet demand): lost margin plus any penalty (€)
c_o	$\in \mathbb{Q}_{>0}$	unit *overage* cost (unsold unit): cost minus salvage value (€)

Data (input of the model) (2/2)

- D is an exception to the convention on capital letters, to follow the literature: the random variable in upper case, its realizations in lower case (d_s)
- With selling price p, purchase cost c and salvage value $v \leq c$ (all in $\mathbb{Q}_{>0}$):

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

SAPIENZA UNIVERSITÀ DI ROMA

Decision variable and model

We introduce a non-negative variable:

$$q = \text{quantity ordered } \textit{before} \text{ observing demand } D$$

Newsvendor model in cost form

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

$$\text{subject to} \quad q \geq 0. \tag{7.1b}$$

Objective function and constraint

- The objective function is **convex** and minimizes the expected cost of the error:
  - overage $(q - D)^+$ charged at c_o
  - underage $(D - q)^+$ charged at c_u
- The symbol $(x)^+ = \max\{x, 0\}$ denotes the positive part
- The constraint $q \geq 0$ defines the variable of the model

The quantile rule

If F is continuous, the optimum satisfies the **quantile rule**:

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

Where the quantile rule comes from

- The argument works on the marginal unit: ordering the q-th extra unit
  - yields c_u if demand absorbs it (probability $1 - F(q)$)
  - costs c_o if it remains unsold (probability $F(q)$)
- It is worth doing as long as $c_u\,(1 - F(q)) \geq c_o\,F(q)$, that is, as long as

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

- Underage far more expensive than overage: α^* is high, one orders a high quantile
- Unsold items expensive (perishables): one orders little

Summary: The basic model

- Expected cost $c_o\,\mathbb{E}[(q-D)^+] + c_u\,\mathbb{E}[(D-q)^+]$, convex in q
- $c_u = p - c$ (underage), $c_o = c - v$ (overage)
- Quantile rule: $q^* = F^{-1}\big(c_u/(c_u+c_o)\big)$

Worked example by hand (discrete demand)

Introduction: Worked example by hand

- Three equally likely demand scenarios
- We compare the expected cost of the natural candidates
- We check that the quantile rule (discrete version) gives the same answer

Example: three equally likely scenarios (1/3)

Data: $D \in \{80, 100, 120\}$ with probability $1/3$; $c_u = 9$, $c_o = 4$, hence

$$\alpha^* = 9/13 = 0.6923$$

Step 1 — expected cost for the three natural candidates.

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

$$C(100) = \tfrac{1}{3}\left[4 \cdot 20 + 0 + 9 \cdot 20\right] = \tfrac{260}{3} = 86.7$$

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

Example: three equally likely scenarios (2/3)

Step 2 — discrete quantile rule.

- We need the smallest q with $F(q) \geq 0.6923$:

$$F(80) = 1/3$$
$$F(100) = 2/3 = 0.667 < 0.6923$$
$$F(120) = 1$$

- Hence $q^* = 120$: consistent with step 1

Example: three equally likely scenarios (3/3)

Step 3 — interpretation.

- Ordering the historical *maximum* even though the mean is 100
- With $c_u = 9 \gg c_o = 4$, running short costs more than twice as much as running long
- Whoever orders "the mean" pays 86.7 against 80
- Uncertainty is managed, not ignored

Summary: Worked example by hand

- With $\alpha^* = 0.6923$ the optimal quantity is $q^* = 120$, the historical maximum
- Ordering the mean costs 86.7 against 80
- Discrete quantile rule: the smallest q with $F(q) \geq \alpha^*$

Case study: the quantile rule in the continuous case

Introduction: the quantile rule in the continuous case

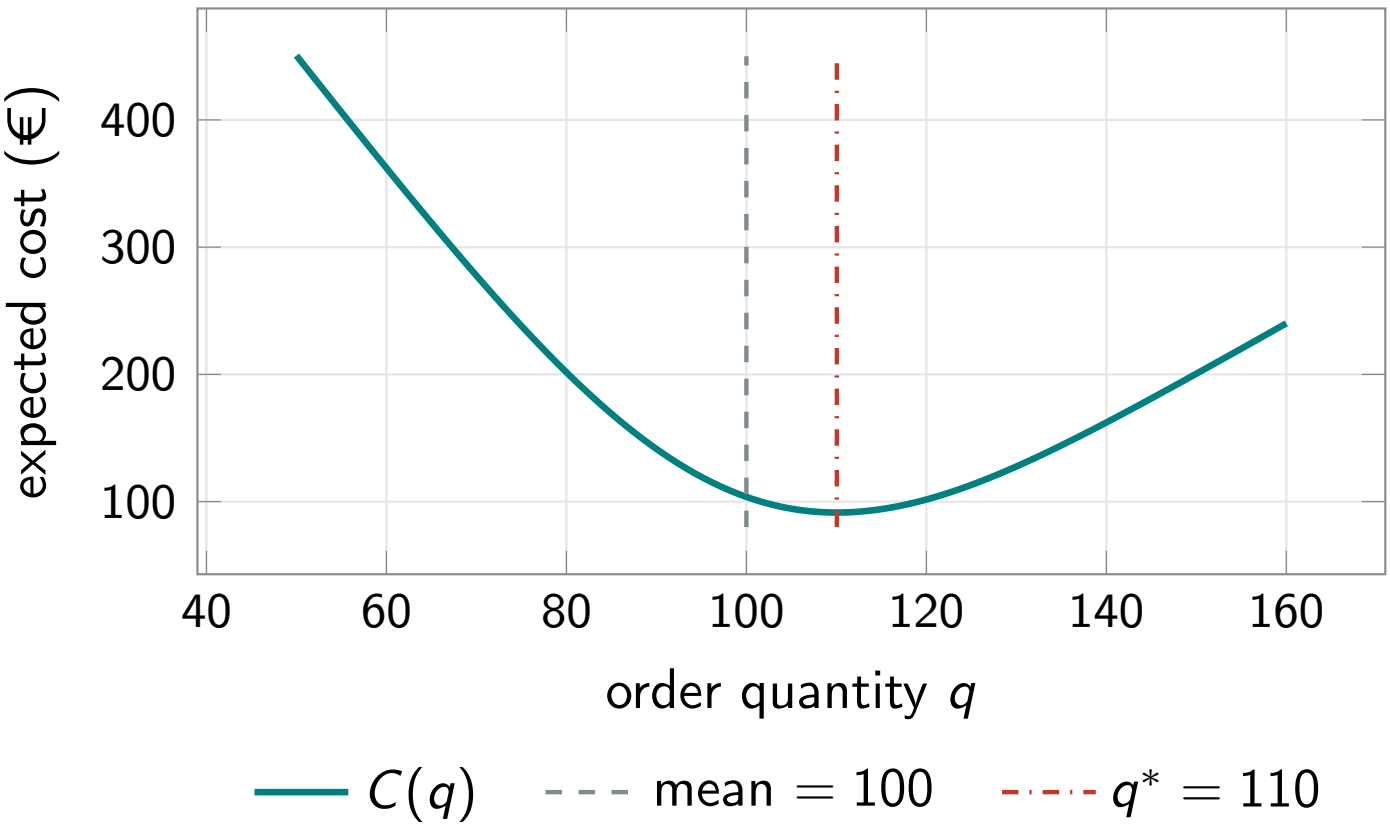
- A bakery: $p = 15$, $c = 6$, $v = 2$ € $\Rightarrow c_u = 9$, $c_o = 4$
- Normally distributed demand with $\mu = 100$, $\sigma = 20$
- We apply the quantile rule and see how stable the solution estimated from scenarios is

Results

Bakery: $p = 15$, $c = 6$, $v = 2$ € $\Rightarrow c_u = 9$, $c_o = 4$; normally distributed demand with $\mu = 100$, $\sigma = 20$.

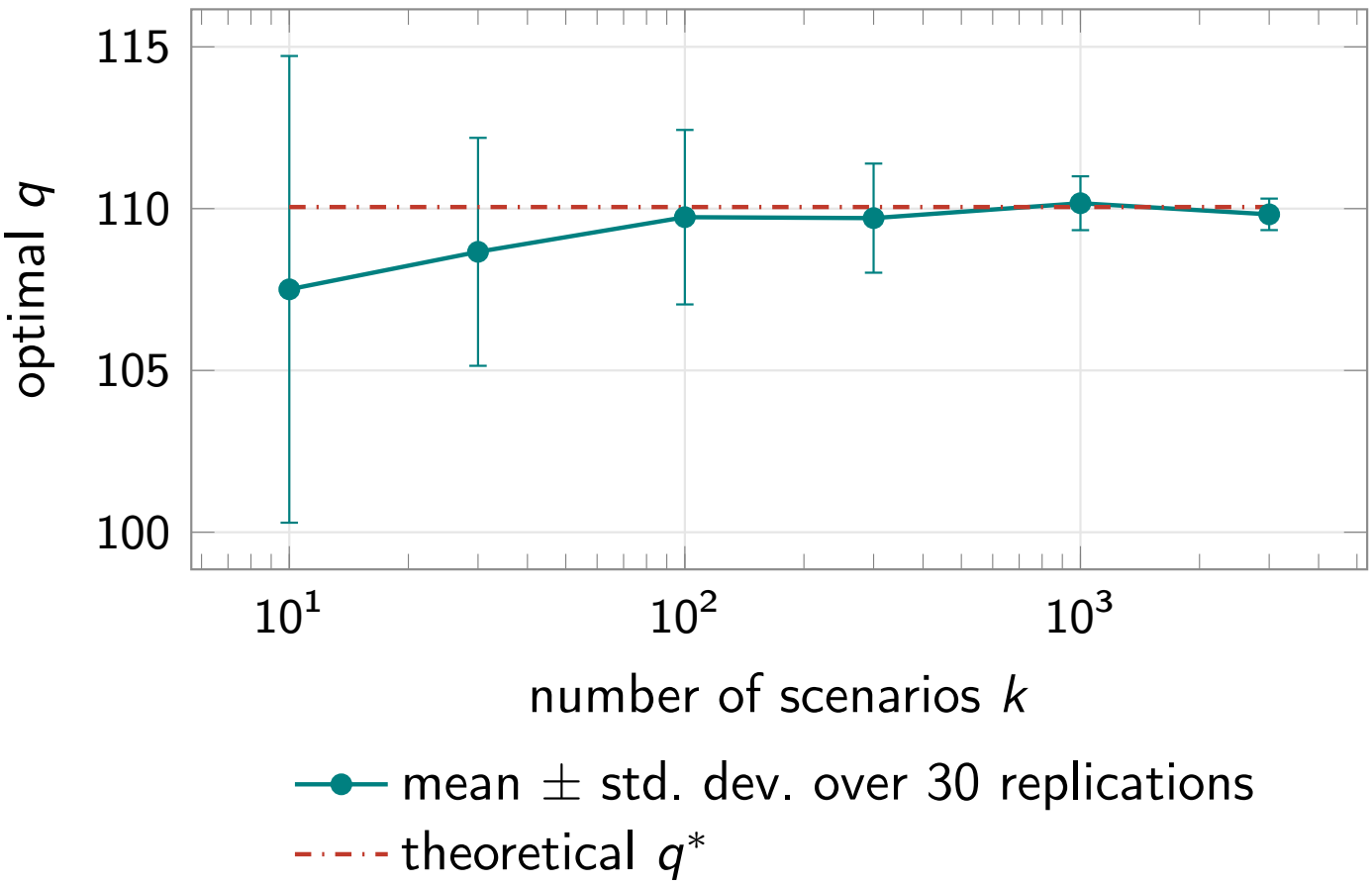
```
alpha* = 9/13 = 0.6923
q* = F^-1(0.6923) = 110.05 units   (mean = 100)
expected cost at q*: 91.43 EUR    at q = mean: 103.72 EUR
```

Minimum at the 69th percentile



The expected cost is convex; the minimum falls at $q^* = 110$, not at the mean 100.

Stability of the scenario-based solution



Reading the figures

- The expected cost $C(q)$ is convex and the minimum falls at $q^* = 110$ (69th percentile of demand), not at the mean 100
- To the right of the mean one pays less: at q^* the expected cost is 91.43 € against 103.72 € at the mean
- The optimal q estimated from k sampled scenarios (mean and standard deviation over 30 replications) shows that with few scenarios the decision is **unstable**

Summary: the quantile rule in the continuous case

- $\alpha^* = 0.6923$, $q^* = 110.05$ against a mean of 100
- Ordering the mean costs 12.29 € more per cycle
- Solutions estimated from few scenarios swing a lot

The scenario-based linear formulation

Introduction: The scenario-based linear formulation

- When F is not known in closed form — the normal situation in a company — one uses **scenarios**
- Historical data, samples from an estimated distribution, or weighted scenarios
- We obtain an **LP** thanks to the positive-part trick
- We measure the value of the stochastic solution (VSS)

New data and variables

- Number of scenarios $k \in \mathbb{Z}_{\geq 1}$, indexed by $s \in \{1, 2, \ldots, k\}$
- For each scenario: demand $d_s \in \mathbb{Q}_{\geq 0}$ and probability $\pi_s \in \mathbb{Q}_{\geq 0}$, with $\sum_{s=1}^{k} \pi_s = 1$
- Besides q, we introduce $2\,k$ non-negative variables:

$$\begin{cases} o_s = \text{overage (unsold units) in scenario } s \\ u_s = \text{underage (unmet demand) in scenario } s \end{cases}$$

for every $s \in \{1, 2, \ldots, k\}$

The LP model

Scenario-based Newsvendor LP model

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

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

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

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

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

Objective function and constraints

- The linear objective function is the weighted average, over the scenarios, of the overage and underage costs: the "scenario-based" version of the expected cost
- The constraints $o_s \geq q - d_s$ and $u_s \geq d_s - q$ define the overage and the underage of each scenario starting from the only genuine decision q
- Together with the non-negativity conditions they realize the positive parts ($2\,k$ linear constraints)
- $q \geq 0$ and $o_s, u_s \geq 0$ define the variables

Why it is linear: the positive-part trick

- $(q - d_s)^+$ is not linear
- At the *optimum* the variable o_s, which appears in the objective with a positive coefficient, is pushed down onto the maximum between $q - d_s$ and 0
- The constraint $o_s \geq q - d_s$ together with $o_s \geq 0$ does exactly the job of $\max\{q - d_s, 0\}$ without binary variables
- The same mechanism applies to u_s
- The trick returns unchanged in the **CVaR** and in the **SVM**

The model in Gurobi

```
q = m.addVar(name="q")
o = m.addVars(S, name="o")      # overage per scenario
u = m.addVars(S, name="u")      # underage 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)
```

Results

```
LP with k = 600 scenarios: q = 108.61
empirical quantile at 69.23%: 108.60   (they coincide: it is a
    theorem)
VSS: decision q = E[D] = 100 costs 99.50
     stochastic q costs 88.33
     value of the stochastic solution = 11.17 EUR per cycle
```

Managerial insight (1/2)

- The *value of the stochastic solution* (VSS) answers the boss's question: "why complicate our lives with scenarios?"
- Here it is worth 11.17 € per cycle, 11% of the cost
- It is the gain from *modelling* uncertainty instead of replacing it with the mean

Managerial insight (2/2)

- The figure on stability gives the other half of the message
- Below 100 scenarios the optimal q swings by ± 3 units from one estimate to the next
- Scenarios are themselves a sample, and the decision inherits their variance

Summary: The scenario-based linear formulation

- With k scenarios the Newsvendor becomes an LP with $2k + 1$ variables
- The positive parts are linearized without binary variables
- The LP reproduces the empirical quantile (it is a theorem)
- VSS $= 11.17$ € per cycle: modelling uncertainty pays off

Service levels

Introduction: Service levels

- Contracts often impose a **service level**
- *Cycle service level*: probability of not running into a stock-out
- *Fill rate*: share of demand served on average
- We see how much they cost and why they must not be confused

The cost of service levels

```
cycle service level 90%: q = 125.63
                         (cost +23.41 EUR over the economic
                             optimum)
cycle service level 95%: q = 132.90  (cost +45.59)
cycle service level 99%: q = 146.53  (cost +95.56)
fill rate at q*: 96.06%
probability of no stockout at q*: 69.23%
```

Do not confuse the service levels

Observation 1: Do not confuse the service levels

At $q^* = 110$ the probability of avoiding a stock-out is only 69%, but the fill rate (share of demand served on average) is 96%: different measures that contracts often confuse.

- Promising "95% probability of full coverage" costs +45.59 € per cycle compared with the economic optimum
- Promising "95% fill rate" is almost free
- Read the SLA carefully before signing it

Summary: Service levels

- Raising the cycle service level costs more and more: +23, +46, +96 € for 90%, 95%, 99%
- At q^*: no stockout 69%, but fill rate 96%
- Two different measures: read the SLA carefully

Risk aversion: mean-CVaR

Introduction: Risk aversion

- Expected cost does not distinguish between many small losses and rare disasters
- With the linear formulation of the CVaR (here only anticipated) we minimize

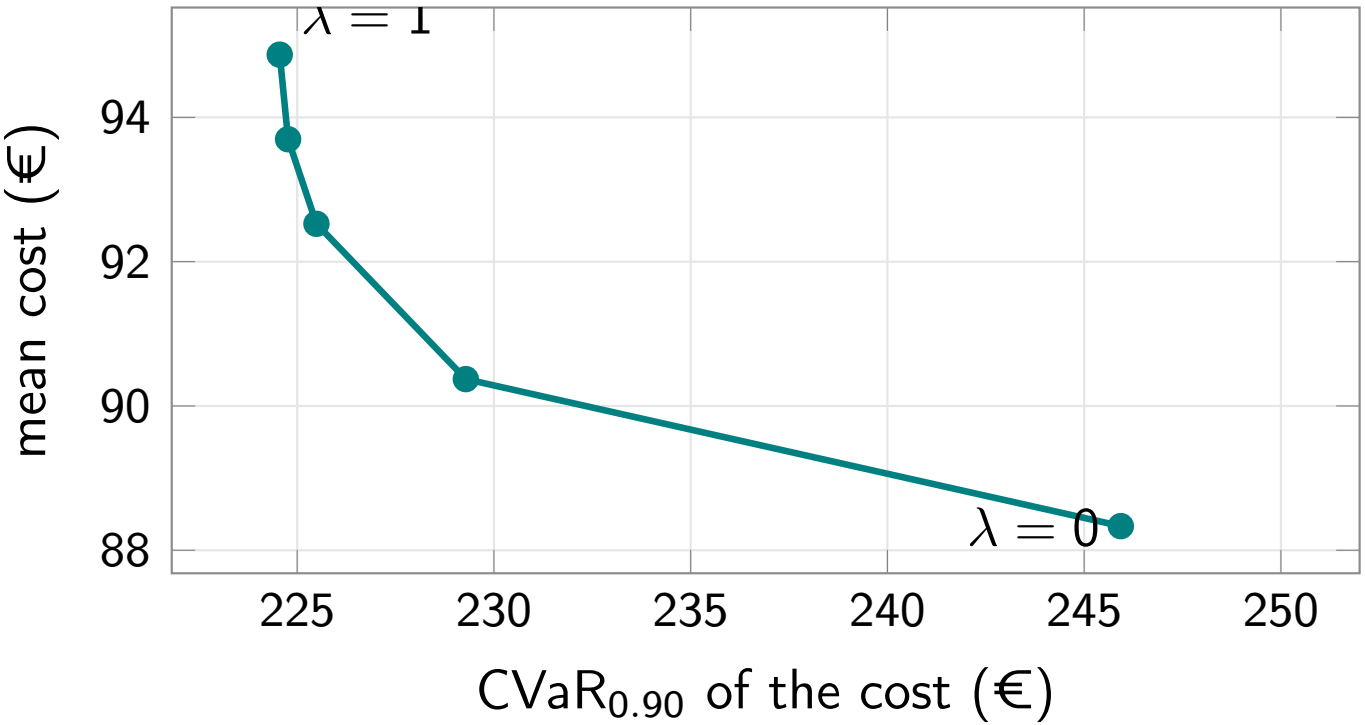
$$(1-\lambda)\,\mathbb{E}[\text{cost}] + \lambda\,\mathrm{CVaR}_{0.90}(\text{cost})$$

- The weight $\lambda \in [0,1]$ measures risk aversion

Results as λ varies

```
lambda = 0.00: q = 108.61  mean cost 88.33  CVaR90 245.93
lambda = 0.25: q = 112.45  mean cost 90.37  CVaR90 229.29
lambda = 0.50: q = 114.48  mean cost 92.52  CVaR90 225.49
lambda = 1.00: q = 116.17  mean cost 94.87  CVaR90 224.56
```

Mean cost – tail risk frontier



Reading the frontier

- From $\lambda = 0$ (optimal on average, heavy tail) to $\lambda = 1$ (maximum protection)
- The first stretch is almost vertical: a lot of protection for very little mean cost
- The risk-averse decision maker orders more (from 108.6 to 116.2 units)
- Pays 6.5 € more on average in order to cut by 21.4 € the mean cost of the worst 10% of the scenarios
- Most of the protection is already obtained with $\lambda = 0.25$

Summary: Risk aversion

- Mean-CVaR objective: a trade-off between expected cost and tail risk
- More risk aversion $\Rightarrow$ one orders more
- The frontier is steep at the beginning and flat afterwards: a little aversion gives a lot of protection

Multi-product with a shared budget

Introduction: Multi-product with a shared budget

- Three cakes with *correlated* demands ($\rho = 0.7$): the holidays go well or badly for everyone together
- A shared production budget of 1200 €
- We see how the budget compresses the quantities and what the dual says

Results

```
  product    | q without budget | q with budget
 panettone   |            111.1 |          99.3
 pandoro     |             91.5 |          75.4
 torrone     |             65.9 |          56.8
Spend: 1200/1200   shadow price of the budget: -0.553
```

Reading the results (1/2)

- The binding budget compresses the three quantities below their respective optimal quantiles
- But not proportionally: the cut depends on the ratio c_u/c_o and on the unit cost of each product
- The dual says that one more euro of budget would reduce the expected cost by 0.55 €

Reading the results (2/2)

- A 55% return that would justify almost any financing
- Information that nobody would have quantified without the model
- The high correlation removes the diversification benefit: the three products fail together
- Treating them as independent would underestimate the risk

Summary: Multi-product with a shared budget

- A shared budget cuts the quantities in a non-proportional way
- The shadow price (-0.553) measures the value of one more euro of budget
- With correlated demands there is no diversification

Exercises

Introduction: Exercises

- Six exercises to consolidate the concepts of the chapter
- Quantile rule, historical data, perishables, EVPI, fill rate, multi-product
- They are solved by hand and with the laboratory script

Exercise: check

- The municipality imposes a penalty $b = 3$ € per unserved customer:

$$c_u = p - c + b$$

- Recompute by hand α^* and q^*
- Check with the script

Exercise: historical data

- Use as scenarios the 24 demand observations of the last two years (generate them with the script) instead of the distribution
- How much does q differ?
- Which approach would you recommend to a company that has only 24 observations?

Exercise: perishables

- The salvage value becomes negative: $v = -1$ € (disposal costs money)
- How do α^* and q^* change?
- Plot q^* as a function of $v \in [-2, 4]$

Exercise: EVPI

- Compute the EVPI (*Expected Value of Perfect Information*)
- Mean cost with q chosen *after* having seen demand (that is, $q = d_s$, cost 0) against the cost of the stochastic solution
- How much is a perfect demand forecast worth at most?

Exercise: constrained fill rate

- Add to the LP the fill rate constraint

$$\sum_{s=1}^{k} \pi_s u_s \leq (1 - 0.98)\, \mathbb{E}[D]$$

- Compare quantity and cost with the cycle service level at 98%

Exercise: multi-product

- Rerun the multi-product case with independent demands ($\rho = 0$), everything else being equal
- Does the optimal expected cost improve or worsen?
- Why?

Summary (1/2)

Basic model:

- We decide q before observing demand
- Quantile rule: $q^* = F^{-1}\big(c_u/(c_u + c_o)\big)$, not the mean

Scenario-based formulation:

- LP with overage and underage variables (positive-part trick)
- VSS: the gain from modelling uncertainty
- Few scenarios $\Rightarrow$ unstable decision

Summary (2/2)

Variants:

- Service levels: cycle service level and fill rate are different measures
- Mean-CVaR: the risk-averse decision maker orders more; a little aversion gives a lot of protection
- Shared budget: the shadow price measures the value of one more euro