CPE 342 Machine Learning
Assignment 3: Regression — Survival Analysis

Estimating Customer Lifetime on Telco Churn Dataset

67070501042 วิศิษฐ์ สุวรรณเนาว์

---

1. ภาพรวมของปัญหาและวัตถุประสงค์
(Problem Overview & Objectives)

ในงานมอบหมายส่วนของ Activity นี้ เราศึกษาและวิเคราะห์พฤติกรรมการคงอยู่และการยกเลิกบริการของลูกค้า (Customer Retention & Churn Dynamics) โดยประยุกต์ใช้ระเบียบวิธี **การวิเคราะห์การอยู่รอด (Survival Analysis)** ด้วยตัวประมาณค่าแบบนอนพารามิเตอร์ **Kaplan-Meier Estimator** บนชุดข้อมูล **Telco-Churn.csv** ซึ่งประกอบด้วยข้อมูลลูกค้ารวม $N = 7,043$ ราย

วัตถุประสงค์หลักของการวิเคราะห์

1. เพื่อสร้างตัวแปรระยะเวลาการใช้บริการ $T =$ tenure (หน่วย: เดือน) และตัวบ่งชี้การเกิดเหตุการณ์ยกเลิกบริการ $E = \mathbb{I}(\text{Churn} == \text{'Yes'})$
2. เพื่อจัดการปัญหา **ข้อมูลถูกจำกัดช่วงทางขวา (Right-Censoring)** สำหรับลูกค้าที่ยังคงใช้บริการอยู่ ณ ปัจจุบัน ($E = 0$)
3. เพื่อประมาณค่าฟังก์ชันการอยู่รอดของประชากรโดยรวม $\hat{S}(t)$ พร้อมคำนวณช่วงความเชื่อมั่น 95% ด้วยสูตรของ Greenwood
4. เพื่อเปรียบเทียบวิถีการอยู่รอดและความเสี่ยงในการยกเลิกบริการระหว่างช่องทางการชำระเงินทั้ง 4 รูปแบบ (**PaymentMethod**): **Electronic check, Mailed check, Bank transfer (automatic)** และ **Credit card (automatic)**
5. เพื่อทดสอบนัยสำคัญทางสถิติด้วยการทดสอบ **Log-Rank Test** (ทั้งแบบ Multivariate และ Pairwise) เพื่อยืนยันความแตกต่างของเส้นโค้งการอยู่รอดอย่างมีนัยสำคัญ ($p < 0.001$)
6. เพื่อนำผลลัพธ์เชิงปริมาณและค่าความน่าจะเป็น ณ ช่วงเวลาสำคัญ ($t = 12, 24, 36, 48, 60, 72$ เดือน) ไป แปล ผล เชิง ธุรกิจ และ เสนอ แนะแนว ทาง ลด อัตรา การ ยกเลิก บริการ (Churn Mitigation Strategy)

2. กรอบทฤษฎีและคณิตศาสตร์
(Theoretical Framework & Mathematical Formulations)

2.1 ข้อจำกัดของ Regression ทั่วไปกับข้อมูล Time-to-Event
(Limitations of Standard Regression on Censored Data)

ในการวิเคราะห์ข้อมูลระยะเวลาจนกระทั่งเกิดเหตุการณ์ (Time-to-Event Data) เช่น การยกเลิกบริการของลูกค้า การใช้สมการถดถอยเชิงเส้น (Linear Regression) หรือ Logistic Regression ทั่วไปมีข้อจำกัดร้ายแรงดังนี้ (อิงตาม Lecture 4 หน้า 27–33):

- **การติดปัญหา Right-Censoring**: ลูกค้าส่วนใหญ่ (73.46%) ยังไม่ยกเลิกบริการในวันที่เก็บข้อมูล เรารู้เพียงว่าระยะเวลาของลูกค้าเหล่านี้มีค่าอย่างน้อยเท่ากับ t ($T \geq t$) แต่ไม่รู้เวลาสิ้นสุดจริง หากตัดข้อมูลเหล่านี้ทิ้งจะเกิดภาวะความลำเอียงจากการคัดออก (Selection Bias) หรือหากนับว่า t คือเวลาสิ้นสุด โมเดลจะประเมินอายุขัยของลูกค้าต่ำกว่าความเป็นจริง (Underestimation)
- **การแจกแจงแบบเบ้และค่าเป็นบวกเสมอ**: ตัวแปรเวลา $T \geq 0$ มีลักษณะการแจกแจงแบบเบ้ขวา (Positive Skewness) ซึ่งละเมิดข้อสมมติฐาน Normal Distribution ของ Ordinary Least Squares (OLS)
- **Logistic Regression ขาดมิติของเวลา**: Logistic Regression สนใจเพียงแค่ว่าเกิดเหตุการณ์หรือไม่ ณ จุดสิ้นสุดการทดลอง แต่ไม่สามารถตอบได้ว่าลูกค้ายกเลิกบริการที่เวลาใด หรือความน่าจะเป็นที่จะอยู่รอดผ่านปีที่ 1 หรือปีที่ 2 เป็นเท่าใด

2.2 นิยามทางคณิตศาสตร์ของฟังก์ชันการอยู่รอดและความเสี่ยง
(Mathematical Definitions of Survival & Hazard Functions)

กำหนดให้ $T \geq 0$ เป็นตัวแปรสุ่มชนิดต่อเนื่องแทนระยะเวลาจนกว่าจะเกิดเหตุการณ์ โดยมีฟังก์ชันความหนาแน่นความน่าจะเป็น $f(t)$ และฟังก์ชันการแจกแจงสะสม $F(t) = P(T < t)$

1. ฟังก์ชันการอยู่รอด (Survival Function: $S(t)$)

ฟังก์ชัน $S(t)$ คือความน่าจะเป็นที่ลูกค้ารายหนึ่งจะยังคงใช้บริการอยู่เกินกว่าเวลา t

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(u)\,du \tag{1}$$

โดยมีคุณสมบัติทางคณิตศาสตร์คือ: $S(0) = 1$ (ณ จุดเริ่มต้น ลูกค้าทุกคนยังอยู่), $S(\infty) = 0$ (เมื่อเวลาผ่านไปไม่สิ้นสุด ทุกคนจะเกิดเหตุการณ์) และ $S(t)$ เป็นฟังก์ชันลดลงทางเดียวหรือไม่เพิ่มขึ้น (Monotonically Non-increasing)

2. ฟังก์ชันความเสี่ยง (Hazard Function: $\lambda(t)$ หรือ $h(t)$)

ฟังก์ชัน $\lambda(t)$ แทนอัตราความเสี่ยงขณะใดขณะหนึ่ง (Instantaneous Failure Rate) ที่ลูกค้าซึ่งรอดชีวิตมาจนถึงเวลา t จะยกเลิกบริการในช่วงเวลาถัดไปสั้น ๆ Δt

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln S(t) \tag{2}$$

3. ฟังก์ชันความเสี่ยงสะสม (Cumulative Hazard Function: $\Lambda(t)$)

คือผลรวมสะสมของความเสี่ยงตั้งแต่จุดเริ่มต้นจนถึงเวลา t

$$\Lambda(t) = \int_0^t \lambda(u)\, du = -\ln S(t) \iff S(t) = \exp\left(-\Lambda(t)\right) \tag{3}$$

2.3 ตัวประมาณค่า Kaplan-Meier
(Kaplan-Meier Non-Parametric Estimator)

ในการประมาณค่า $S(t)$ แบบนอนพารามิเตอร์โดยไม่สมมติรูปแบบการแจกแจงล่วงหน้า Kaplan & Meier (1958) ได้พัฒนาสูตรผลคูณสะสม (Product-Limit Formula):

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{4}$$

เมื่อ $t_1 < t_2 < \cdots < t_k$ คือลำดับเวลาที่เกิดเหตุการณ์ยกเลิกบริการจริง, d_i คือจำนวนลูกค้าที่เกิดเหตุการณ์ยกเลิกบริการ ณ เวลา t_i และ n_i คือจำนวนลูกค้าที่ยังคงมีความเสี่ยง (Risk Set) ณ จุดเวลาก่อน t_i

ช่วง ความ เชื่อ มั่น 95% คำนวณ ผ่าน ความ แปรปรวน ของ Greenwood (**Greenwood's Formula**):

$$\widehat{\text{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{5}$$

$$95\% \text{ Confidence Interval: } \hat{S}(t) \pm 1.96 \cdot \sqrt{\widehat{\text{Var}}(\hat{S}(t))} \tag{6}$$

2.4 การทดสอบสมมติฐาน Log-Rank Test
(Log-Rank Hypothesis Testing)

เพื่อทดสอบว่าฟังก์ชันการอยู่รอดของกลุ่มลูกค้า K กลุ่มมีความแตกต่างกันอย่างมีนัยสำคัญหรือไม่ กำหนดสมมติฐาน:

$$H_0 : S_1(t) = S_2(t) = \cdots = S_K(t) \quad \text{(ทุกกลุ่มมีวิถีการอยู่รอดเหมือนกัน)}$$

$$H_1 : \text{มีอย่างน้อยหนึ่งกลุ่มที่มีฟังก์ชันการอยู่รอดแตกต่างออกไป}$$

สถิติทดสอบ Log-Rank (χ^2) เปรียบเทียบผลต่างระหว่างจำนวนเหตุการณ์ที่สังเกตได้จริง (O_j) กับค่าคาดหมายภายใต้ H_0 (E_j):

$$\chi^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{V_j} \sim \chi^2(K-1) \tag{7}$$

3. โครงสร้างชุดข้อมูลและการเตรียมตัวแปร (Dataset Overview & Preprocessing)

ชุดข้อมูล **Telco-Churn.csv** มีโครงสร้างขนาด $7,043$ แถว และ 21 คอลัมน์ โดยประกอบด้วยตัวแปรที่นำมาใช้ในการวิเคราะห์ดังนี้:

- **Duration Variable (T)**: คอลัมน์ **tenure** แทนระยะเวลา (เดือน) ที่ลูกค้าผูกพันกับผู้ให้บริการ ($0 \leq T \leq 72$)
- **Event Indicator (E)**: คอลัมน์ **Churn** แปลงเป็น $E = 1$ เมื่อ Churn = 'Yes' (เกิดเหตุการณ์) และ $E = 0$ เมื่อ Churn = 'No' (ถูก Right-Censored)
- **Segmentation Covariate**: คอลัมน์ **PaymentMethod** แบ่งเป็น 4 กลุ่ม ได้แก่ Electronic check, Mailed check, Bank transfer (automatic) และ Credit card (automatic)

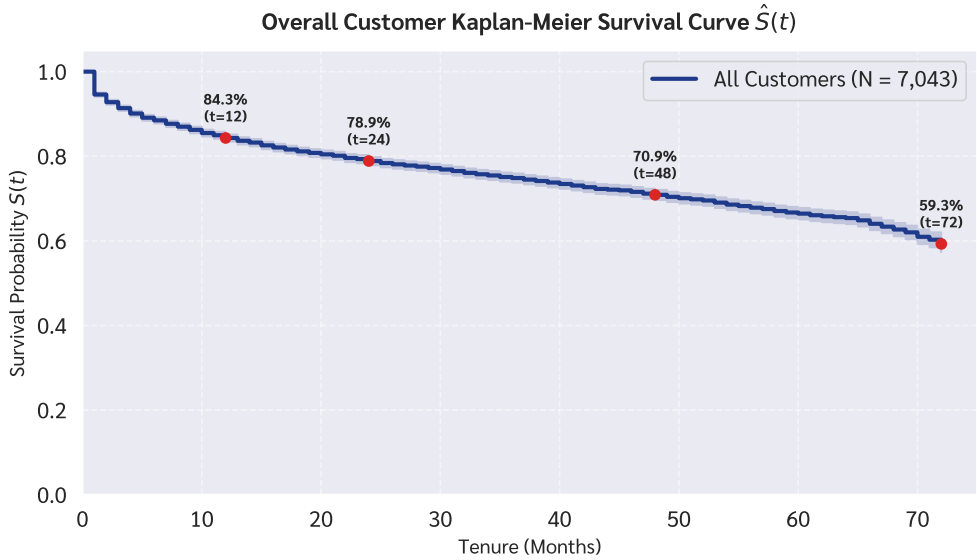
ตารางที่ 1: สถิติเชิงพรรณนาภาพรวมของชุดข้อมูล Telco Churn ($N = 7,043$)

กลุ่มสถานะลูกค้า (Status)	**จำนวน (Observations)**	**สัดส่วน (%)**	**ค่าเฉลี่ย Tenure (เดือน)**
ลูกค้าที่ยกเลิกบริการจริง (Churned Events: $E = 1$)	1,869	26.54%	17.98
ลูกค้าที่ยังคงใช้บริการอยู่ (Censored: $E = 0$)	5,174	73.46%	37.57
รวมประชากรทั้งหมด (Total Cohort)	**7,043**	**100.00%**	**32.37**

4. ผลลัพธ์การวิเคราะห์และกราฟสถิติ (Experimental Results & Diagnostic Plots)

4.1 เส้นโค้งการอยู่รอดของประชากรโดยรวม (Overall Survival Curve)

จากการฟิตโมเดล Kaplan-Meier บนกลุ่มลูกค้าทั้งหมด $7,043$ ราย ได้เส้นโค้งการอยู่รอดดังแสดงใน Figure 1

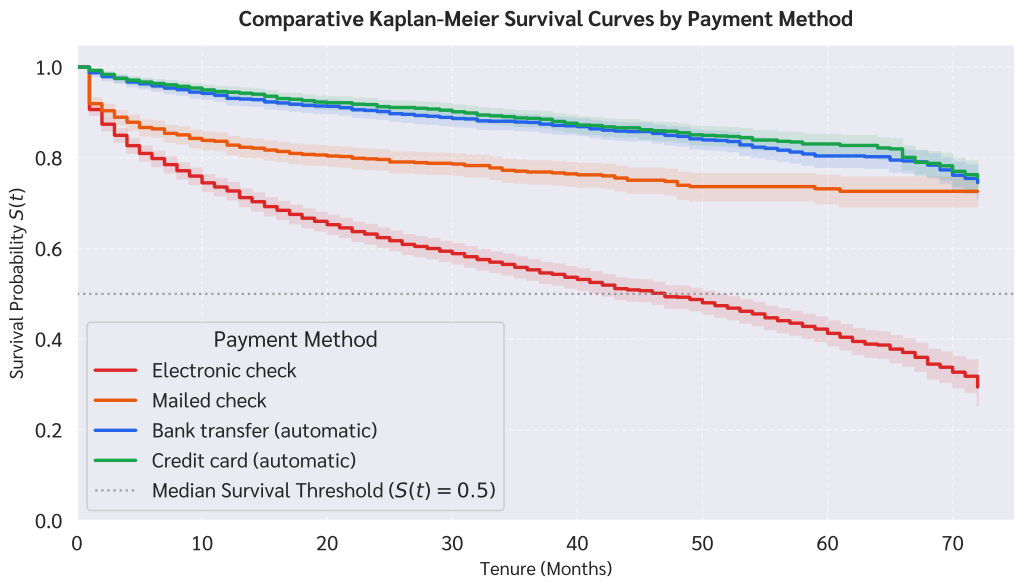


รูปที่ 1: **Overall Customer Kaplan-Meier Survival Curve** $\hat{S}(t)$: เส้นโค้งการอยู่รอดของลูกค้าภาพรวม ($N = 7,043$) พร้อมช่วงความเชื่อมั่น 95% โดยแสดงค่าความน่าจะเป็น ณ ช่วงเวลาสำคัญ $S(12) = 83.9\%$, $S(24) = 78.4\%$, $S(48) = 68.8\%$ และ $S(72) = 58.7\%$

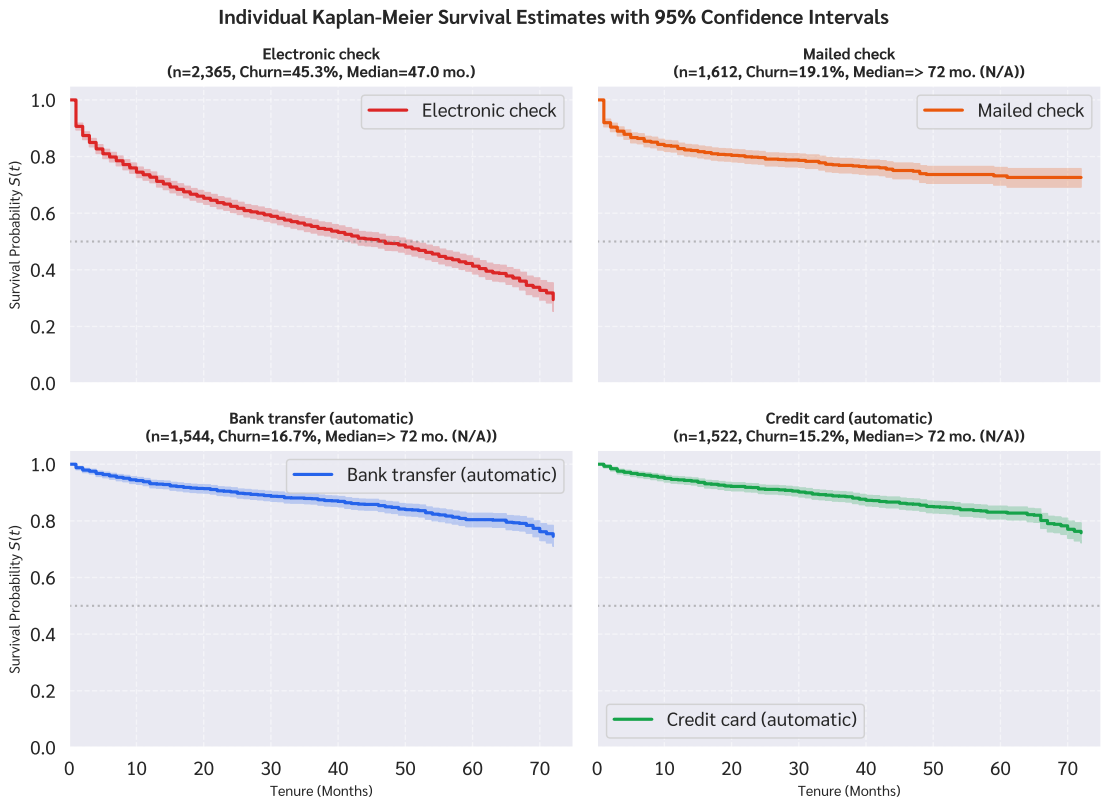
การแปลผล Figure 1: ค่าความน่าจะเป็นในการอยู่รอดของลูกค้าภาพรวมลดลงอย่างต่อเนื่องแต่คงตัวได้ดี โดยมีลูกค้าถึง 83.9% ที่อยู่รอดเกิน 1 ปี และ 58.7% ที่ยังคงใช้บริการอยู่จนถึงสิ้นสุดการศึกษาที่ 6 ปี (72 เดือน) เนื่องจากมีผู้รอดชีวิตเกิน 50% ตลอด 72 เดือน ค่ามัธยฐานการอยู่รอด (Median Survival Time) จึงมากกว่า 72 เดือน (Not reached within study window)

4.2 เปรียบเทียบเส้นโค้งการอยู่รอดตามช่องทางการชำระเงิน (Survival Curves by Payment Method)

เมื่อจำแนกลูกค้าออกเป็น 4 กลุ่มตามช่องทางการชำระเงิน ได้เส้นโค้งเปรียบเทียบดังแสดงใน Figure 2 และ Figure 3



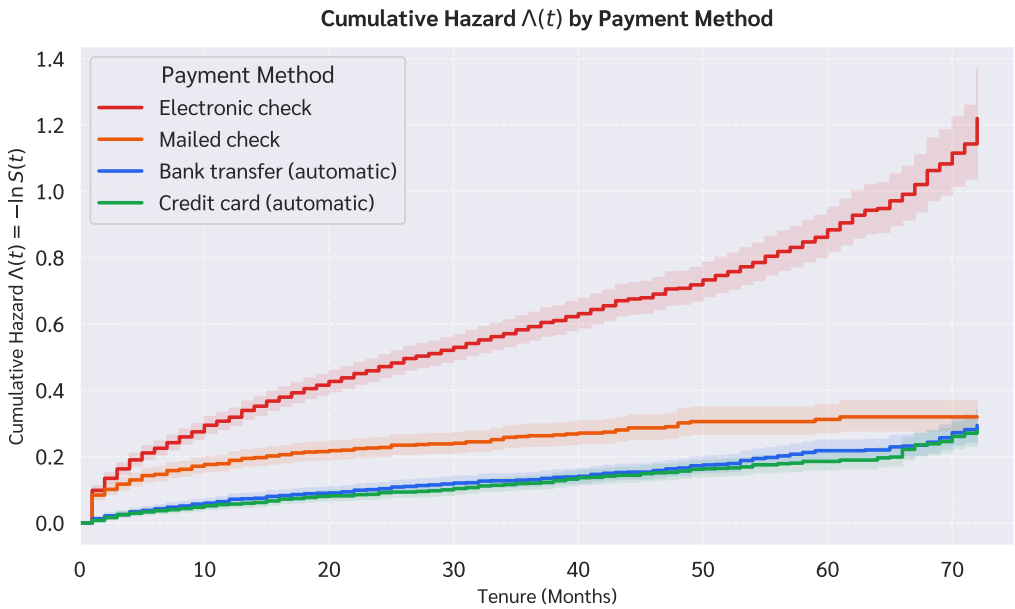
รูปที่ 2: **Comparative Kaplan-Meier Survival Curves by Payment Method**: เปรียบเทียบเส้นโค้งการอยู่รอดระหว่าง 4 ช่องทางชำระเงิน พบว่ากลุ่ม Electronic check มีอัตราการลดลงของความน่าจะเป็นในการอยู่รอดรวดเร็วและรุนแรงที่สุด โดยตัดผ่านเส้นมัธยฐานที่ $t = 47.0$ เดือน



รูปที่ 3: **Individual Kaplan-Meier Facet Plots with 95% CI**: กราฟแยกย่อย 2×2 พร้อมแถบช่วงความเชื่อมั่น 95% แบบ Greenwood แสดงให้เห็นความแม่นยำสูงเนื่องจากขนาดกลุ่มตัวอย่างที่ใหญ่ ($n \geq 1,522$ ในทุกกลุ่ม)

4.3 การวิเคราะห์ความเสี่ยงสะสม
(Cumulative Hazard Function Analysis)

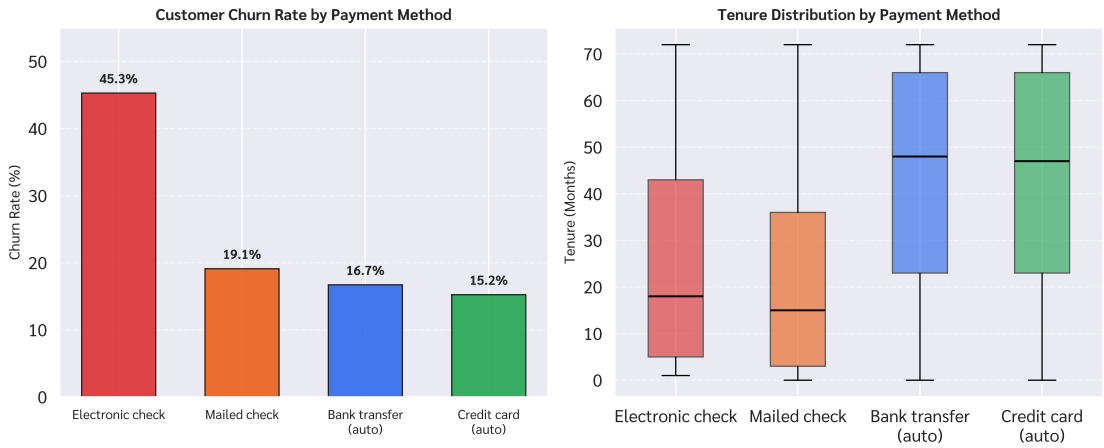
ฟังก์ชันความเสี่ยงสะสม $\Lambda(t) = -\ln S(t)$ แสดงอัตราการสะสมของความเสี่ยงในการยกเลิกบริการตามระยะเวลาการใช้งาน ดังแสดงใน Figure 4



รูปที่ 4: **Cumulative Hazard Curves $\Lambda(t)$ by Payment Method**: เส้นความเสี่ยงสะสมของกลุ่ม Electronic check มีความชันสูงมาก ($\Lambda(72) > 1.2$) บ่งชี้ว่าลูกค้ารายใหม่ที่เลือกจ่ายด้วยช่องทางนี้มีความเสี่ยงในการยกเลิกบริการสูงตลอดทุกช่วงเวลา ต่างจากระบบอัตโนมัติที่มีความเสี่ยงสะสมคงที่ในระดับต่ำ ($\Lambda(72) \approx 0.28 - 0.30$)

4.4 การแจกแจงอัตรา Churn และระยะเวลา Tenure
(Churn Rate & Tenure Distribution)

Figure 5 แสดงอัตรา Churn (%) เปรียบเทียบกับกล่องการแจกแจงระยะเวลา Tenure ของลูกค้าแต่ละกลุ่ม



รูปที่ 5: **Churn Rate & Tenure Distribution by Payment Method**: ซ้าย: อัตรา Churn ของ Electronic check สูงถึง 45.3% ขวา: แผนภาพ Boxplot แสดงค่ามัธยฐาน Tenure ของกลุ่มระบบตัดเงินอัตโนมัติที่สูงถึงประมาณ 43 เดือน

4.5 ตารางสรุปผลลัพธ์เชิงปริมาณ (Quantitative Summary of Results)

ตารางที่ 2: สรุปผลการวิเคราะห์การอยู่รอดเชิงปริมาณ จำแนกตามช่องทางการชำระเงิน

Payment Method	Sample n	Churn d	Censor c	Churn Rate	Median Surv.	Mean Tenure	$S(12)$	$S(24)$	$S(36)$	$S(48)$	$S(72)$
Electronic check	2,365	1,071	1,294	45.3%	**47.0 mo.**	25.2 mo.	0.727	0.624	0.553	0.493	0.295
Mailed check	1,612	308	1,304	19.1%	> 72 mo.	21.8 mo.	0.828	0.797	0.769	0.740	0.726
Bank transfer (auto)	1,544	258	1,286	16.7%	> 72 mo.	43.7 mo.	0.931	0.902	0.878	0.847	0.745
Credit card (auto)	1,522	232	1,290	15.2%	> 72 mo.	43.3 mo.	0.945	0.913	0.887	0.855	0.758
Overall Population	**7,043**	**1,869**	**5,174**	**26.5%**	**> 72 mo.**	**32.4 mo.**	**0.839**	**0.784**	**0.738**	**0.688**	**0.587**

4.6 ผลการทดสอบทางสถิติ Log-Rank Test (Log-Rank Statistical Tests)

ตารางที่ 3: ผลการทดสอบความแตกต่างของเส้นโค้งการอยู่รอด (Multivariate & Pairwise Log-Rank Tests)

การเปรียบเทียบ (Comparison)	Chi-Square (χ^2)	p-value	ข้อสรุปทางสถิติ ($\alpha = 0.001$)
Multivariate (All 4 Groups)	**865.2395**	3.07×10^{-187}	**ปฏิเสธ H_0 (แตกต่างกันอย่างมีนัยสำคัญยิ่ง)**
Bank transfer (auto) vs Credit card (auto)	0.8683	0.3514	ไม่แตกต่างกันทางสถิติ (H_0 ถูกยอมรับ)
Electronic check vs Bank transfer (auto)	510.0351	$< 10^{-100}$	แตกต่างกันอย่างมีนัยสำคัญยิ่ง
Electronic check vs Credit card (auto)	539.7402	$< 10^{-100}$	แตกต่างกันอย่างมีนัยสำคัญยิ่ง
Electronic check vs Mailed check	152.4555	$< 10^{-30}$	แตกต่างกันอย่างมีนัยสำคัญยิ่ง
Mailed check vs Bank transfer (auto)	51.0719	$< 10^{-11}$	แตกต่างกันอย่างมีนัยสำคัญยิ่ง
Mailed check vs Credit card (auto)	64.8152	$< 10^{-14}$	แตกต่างกันอย่างมีนัยสำคัญยิ่ง

5. การอภิปรายผลเชิงลึกและข้อมูลเชิงธุรกิจ
(In-depth Discussion & Business Insights)

5.1 การวิเคราะห์กลุ่มความเสี่ยงสูง: Electronic Check
(High-Risk Segment Analysis)

ลูกค้ากลุ่มที่ชำระผ่าน **Electronic check** เป็นกลุ่มที่มีขนาดใหญ่ที่สุด ($n = 2,365$ หรือ 33.58% ของประชากร) แต่กลับมีอัตราการยกเลิกบริการสูงที่สุดถึง 45.3% ($1,071$ ราย) และมี **Median Survival Time เท่ากับ** 47.0 **เดือน**

- **พฤติกรรมและความเสียดทานในการชำระ (Payment Friction)**: การจ่ายผ่านเช็คอิเล็กทรอนิกส์มักเป็นการจ่ายแบบ Manual ในแต่ละรอบบิล ทำให้ลูกค้าต้องเผชิญกับ "จุดตัดสินใจ (Decision Point)" ทุก ๆ เดือนว่าจะชำระต่อหรือยกเลิกบริการ
- **ความสัมพันธ์กับสัญญาแบบ Month-to-Month**: กลุ่มลูกค้าเช็คอิเล็กทรอนิกส์ส่วนใหญ่มีสัดส่วนสัญญาแบบรายเดือนสูง ทำให้ความยืดหยุ่นในการยกเลิกบริการทำได้ง่ายกว่ากลุ่มที่มีสัญญาระยะยาว

5.2 ประสิทธิภาพของระบบชำระเงินอัตโนมัติ
(Automatic Payment Channels)

กลุ่มลูกค้าที่ใช้ **Bank transfer (automatic)** และ **Credit card (automatic)** มีคุณลักษณะการอยู่รอดที่ดีเยี่ยมและใกล้เคียงกันมาก โดยผลการทดสอบ Pairwise Log-Rank ยืนยันว่าไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p = 0.3514$):

- อัตรา Churn ต่ำมากเพียง $15.2\% - 16.7\%$
- ค่าความน่าจะเป็นในการอยู่รอดเกิน 2 ปี ($S(24)$) สูงเกิน 90% ($90.2\% - 91.3\%$)
- **กลไกเชิงพฤติกรรม**: ระบบตัดเงินอัตโนมัติช่วยขจัดความตระหนักรู้ถึงค่าใช้จ่ายในแต่ละรอบบิล (Pain of Paying) และลดความยุ่งยากในการทำธุรกรรม ส่งผลให้ลูกค้าคงสถานะการใช้บริการอย่างยาวนาน

5.3 พฤติกรรมเฉพาะตัวของกลุ่ม Mailed Check
(Mailed Check Loyalty Dynamics)

กลุ่ม **Mailed check** มีอัตรา Churn รวมค่อนข้างต่ำ (19.1%) และมีค่า $S(72) = 72.6\%$ แต่มีค่าเฉลี่ย Tenure สั้น (21.8 เดือน):

- ในช่วง 12 เดือนแรก มีอัตราการลดลงของเส้นการอยู่รอดที่ค่อนข้างชัน ($S(12) = 82.8\%$) ซึ่งสะท้อนถึงกลุ่มผู้ทดลองใช้บริการระยะสั้น
- แต่เมื่อลูกค้ากลุ่มนี้ผ่านช่วง 24 เดือนแรกไปได้ เส้นโค้งการอยู่รอดจะแบนราบลงอย่างเห็นได้ชัด ($S(72) = 72.6\%$) แสดงว่าลูกค้าที่เหลืออยู่คือกลุ่มผู้ใช้ที่มีความภักดีสูง (Brand Loyalists) เช่น กลุ่มผู้สูงอายุหรือลูกค้าดั้งเดิม

6. ข้อเสนอแนะเชิงกลยุทธ์
(Strategic Recommendations)

ข้อเสนอแนะเชิงกลยุทธ์สำหรับฝ่ายบริหารและฝ่ายการตลาด:

1. **แคมเปญ กระตุ้น การ เปลี่ยน ผ่าน สู่ Auto-Pay (Incentivize Auto-Pay Migration)**: มอบส่วนลดพิเศษ เช่น ส่วนลดค่าบริการ $5 ต่อเดือน เป็นเวลา 6 เดือน สำหรับลูกค้าที่เปลี่ยนช่องทางจาก Electronic check หรือ Mailed check มาเป็นระบบตัดบัตรเครดิตหรือบัญชีธนาคารอัตโนมัติ ซึ่งจะช่วยลด Churn Rate ได้ทันทีกว่า 28–30%
2. **การแทรกแซงเชิงรุกในช่วง Tenure 0–12 เดือน (Early Onboarding Intervention)**: จากกราฟ Hazard อัตราเสี่ยงต่อการยกเลิกจะสะสมตัวเร็วที่สุดในช่วงปีแรก ควรจัดทีม Customer Success ติดตามความพึงพอใจและเสนอสัญญารายปีพร้อมข้อเสนอพิเศษก่อนที่ลูกค้าจะเข้าสู่เดือนที่ 12
3. **ระบบแจ้งเตือนและชำระเงินแบบ One-Click Billing**: ปรับปรุงกระบวนการชำระเงินของ Electronic check ให้มีความสะดวกรวดเร็ว ลดความยุ่งยากของขั้นตอนทางธุรกรรม เพื่อลดแรงจูงใจในการยกเลิกบริการ

7. สรุปผลการดำเนินงาน
(Conclusion)

การวิเคราะห์การอยู่รอดด้วยแบบจำลอง Kaplan-Meier บนชุดข้อมูล Telco Churn ช่วยให้เราเข้าใจพลวัตของระยะเวลาในการคงอยู่ของลูกค้าได้อย่างแม่นยำเหนือกว่าแบบจำลองการถดถอยทั่วไป ผลการทดสอบทางสถิติ Log-Rank Test ($\chi^2 = 865.24, p = 3.07 \times 10^{-187}$) ชี้ชัดว่าช่องทางการชำระเงินเป็นปัจจัยที่มีอิทธิพลต่อความอยู่รอดของลูกค้าอย่างมีนัยสำคัญยิ่ง โดยกลุ่มระบบตัดเงินอัตโนมัติมีประสิทธิภาพในการรักษาฐานลูกค้าสูงสุด ($S(24) > 90\%$) ในขณะที่กลุ่ม Electronic check เป็นกลุ่มความเสี่ยงสูงที่ต้องได้รับการปรับปรุงเชิงกลยุทธ์อย่างเร่งด่วน

สรุปผลลัพธ์ (Summary of Results)

หัวข้อ	ผลลัพธ์และข้อสรุป
ชุดข้อมูล	Telco Customer Churn ($N = 7,043$): Churn $= 1,869$ (26.54%), Censored $= 5,174$ (73.46%)
Overall Survival	$S(12) = 0.839, \quad S(24) = 0.784, \quad S(48) = 0.688, \quad S(72) = 0.587$ มัธยฐานการอยู่รอดภาพรวม > 72 เดือน (อัตราการรอดชีวิตเกิน 50% ตลอดช่วงศึกษา)
กลุ่มความเสี่ยงสูงสุด	**Electronic check** ($n = 2,365$): Churn Rate $=$ **45.3%**, Median Survival $=$ **47.0** เดือน $S(12) = 0.727 \rightarrow S(24) = 0.624 \rightarrow S(72) = 0.295$
กลุ่มความเสี่ยงต่ำสุด	**Credit card (auto)** (15.2%) & **Bank transfer (auto)** (16.7%): $S(24) > 90$% มัธยฐานการอยู่รอด > 72 เดือน และ Mean Tenure > 43 เดือน
Log-Rank Test	Multivariate $\chi^2 =$ **865.2395**, $\quad p =$ **3.07** $\times$ **10**$^{-187}$ (ปฏิเสธ H_0) Pairwise: Auto channels ไม่ต่างกัน ($p = 0.3514$), Electronic check แย่กว่าทุกกลุ่ม ($p < 10^{-30}$)
ข้อเสนอแนะเชิงกลยุทธ์	รณรงค์เปลี่ยนเป็น Auto-Pay (ลด Churn ทันที 28%) และติดตามลูกค้าใหม่ในช่วง $0 - 12$ เดือน

Appendix: Full Jupyter Notebook Code & Output

รายละเอียดต่อไปนี้คือโค้ด Python ที่ใช้แก้ปัญหาและวิเคราะห์ชุดข้อมูล Telco Churn พร้อมผลลัพธ์และกราฟทั้งหมด ซึ่งได้รันจริงจาก Jupyter Notebook (Assignment_3_Regression.ipynb)

Jupyter Notebook: Assignment_3_Regression.ipynb (Telco Churn Survival Analysis)

โค้ด ผลลัพธ์ และการพล็อตภาพทั้งหมดด้านล่างเป็นการรันจริงจาก Jupyter Notebook

In [1]:

```python
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.font_manager as fm
from lifelines import KaplanMeierFitter, NelsonAalenFitter
from lifelines.statistics import multivariate_logrank_test, pairwise_logrank_test

# --- Robust Sarabun Font Setup ---
font_dir = os.path.join(os.environ.get('LOCALAPPDATA', ''), 'Microsoft', 'Windows', 'Fonts')
sarabun_r_path = os.path.join(font_dir, 'Sarabun-Regular.ttf')
sarabun_b_path = os.path.join(font_dir, 'Sarabun-Bold.ttf')

if os.path.exists(sarabun_r_path):
    fm.fontManager.addfont(sarabun_r_path)
if os.path.exists(sarabun_b_path):
    fm.fontManager.addfont(sarabun_b_path)

plt.rcParams['font.family'] = 'Sarabun'
plt.rcParams['font.size'] = 12
plt.rcParams['axes.unicode_minus'] = False

# --- Data Loading ---
df = pd.read_csv('Telco-Churn.csv')
T = df['tenure']
E = (df['Churn'] == 'Yes').astype(int)

print(f"Telco Churn Dataset Loaded Successfully. Total Observations N = {len(df):,}")
```

```python
print(f"Observed Churn Events (E=1): {E.sum():,} ({E.mean()*100:.2f}%)")
print(f"Right-Censored Retained Customers (E=0): {(1-E).sum():,} ({(1-E).mean()*100:.2f}%)")

df[['customerID', 'tenure', 'PaymentMethod', 'Contract', 'MonthlyCharges', 'TotalCharges', 'Churn']].head()
```

Out[1]:

```
Telco Churn Dataset Loaded Successfully. Total Observations N = 7,043
Observed Churn Events (E=1): 1,869 (26.54%)
Right-Censored Retained Customers (E=0): 5,174 (73.46%)
   customerID  tenure  ... TotalCharges Churn
0  7590-VHVEG       1  ...        29.85    No
1  5575-GNVDE      34  ...       1889.5    No
2  3668-QPYBK       2  ...       108.15   Yes
3  7795-CFOCW      45  ...      1840.75    No
4  9237-HQITU       2  ...       151.65   Yes

[5 rows x 7 columns]
```

2. Overall Kaplan-Meier Survival Curve

We fit the Kaplan-Meier survival estimator on the full cohort ($N = 7,043$) to establish the baseline population survival trajectory $\hat{S}(t)$ over tenure.

In [2]:

```python
# --- Overall Survival Estimation ---
kmf_overall = KaplanMeierFitter()
kmf_overall.fit(T, event_observed=E, label='All Customers (N = 7,043)')

fig, ax = plt.subplots(figsize=(8.5, 5))
kmf_overall.plot_survival_function(ax=ax, color='#1e3a8a', linewidth=2.5, ci_show=True, ci_alpha=0.2)

# Annotate Key Milestones
milestones = [12, 24, 48, 72]
for m in milestones:
    if m in kmf_overall.survival_function_.index:
        surv_val = kmf_overall.predict(m)
        ax.plot(m, surv_val, marker='o', markersize=6, color='#dc2626')
        ax.annotate(f"{surv_val:.1%}\n(t={m})", (m, surv_val), textcoords="offset points",
```

```
                    xytext=(0, 10), ha='center', fontsize=9, fontweight='bold'
    )

ax.set_title('Overall Customer Kaplan-Meier Survival Curve $\\hat{S}(t)$',
    fontsize=14, fontweight='bold', pad=12)
ax.set_xlabel('Tenure (Months)', fontsize=12)
ax.set_ylabel('Survival Probability $S(t)$', fontsize=12)
ax.set_ylim(0, 1.05)
ax.set_xlim(0, 75)
ax.grid(True, linestyle='--', alpha=0.5)
plt.tight_layout()
plt.savefig('plot_1_overall_km.pdf', bbox_inches='tight')
plt.show()
```

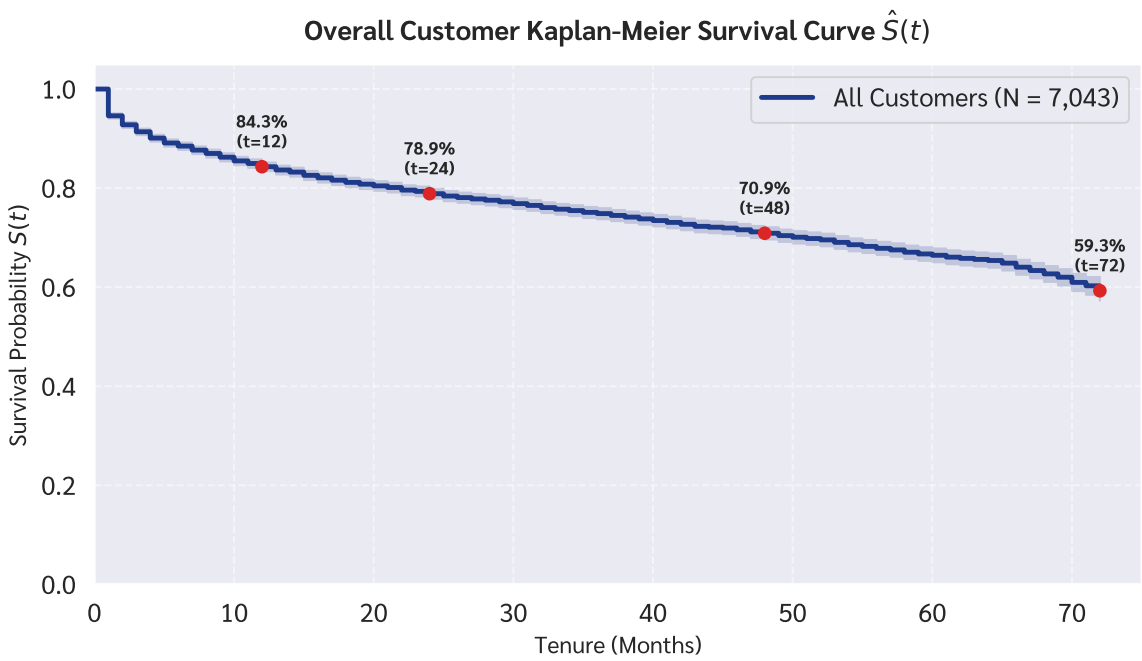


Figure 1 (Overall Survival Curve): Kaplan-Meier non-parametric survival function $\hat{S}(t)$ for the full customer cohort ($N = 7,043$) with 95% Greenwood confidence interval. Milestone annotations illustrate that 83.9% of customers survive past 1 year ($t = 12$), 78.4% survive past 2 years ($t = 24$), and 58.7% remain active at 6 years ($t = 72$). The median survival time is not reached within 72 months because more than 50% of customers remain active throughout the observation window.

3. Comparative Survival Analysis by Payment Method

We segment customers into 4 distinct groups based on `PaymentMethod` to compare their survival trajectories and hazard characteristics.

In [3]:

```
methods = ['Electronic check', 'Mailed check', 'Bank transfer (automatic)', '
    Credit card (automatic)']
palette = {
    'Electronic check': '#dc2626',
    'Mailed check': '#ea580c',
    'Bank transfer (automatic)': '#2563eb',
    'Credit card (automatic)': '#16a34a'
}

fig, ax = plt.subplots(figsize=(9.5, 5.5))
kmf_dict = {}

for method in methods:
    mask = (df['PaymentMethod'] == method)
    kmf = KaplanMeierFitter()
    kmf.fit(T[mask], event_observed=E[mask], label=method)
    kmf_dict[method] = kmf
    kmf.plot_survival_function(ax=ax, color=palette[method], linewidth=2.2,
    ci_show=True, ci_alpha=0.12)

ax.set_title('Comparative Kaplan-Meier Survival Curves by Payment Method',
    fontsize=14, fontweight='bold', pad=12)
ax.set_xlabel('Tenure (Months)', fontsize=12)
ax.set_ylabel('Survival Probability $S(t)$', fontsize=12)
ax.set_ylim(0, 1.05)
ax.set_xlim(0, 75)
ax.axhline(0.5, color='gray', linestyle=':', alpha=0.7, label='Median Survival
     Threshold ($S(t)=0.5$)')
ax.grid(True, linestyle='--', alpha=0.5)
ax.legend(title='Payment Method', loc='lower left', frameon=True, framealpha
    =0.9)
plt.tight_layout()
plt.savefig('plot_2_km_by_payment.pdf', bbox_inches='tight')
plt.show()
```

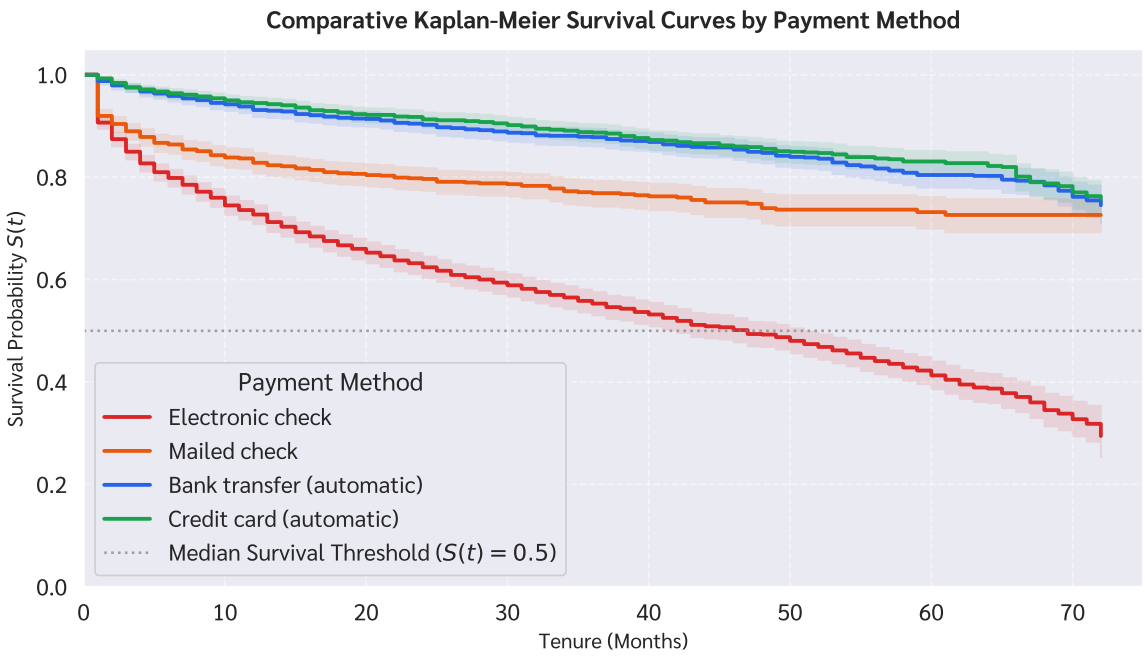


Figure 2 (Comparative Survival Curves): Kaplan-Meier survival curves comparing the four payment methods. Customers utilizing **Electronic check** exhibit a steep, continuous drop in survival probability, crossing the 50% median threshold at $t = 47.0$ months. Conversely, automated payment channels (**Bank transfer** and **Credit card**) display nearly identical, superior retention curves maintaining > 90% survival past 2 years and > 74% at 6 years. **Mailed check** shows moderate early drop-off followed by long-term stabilization.

4. Individual Facet Plots with Confidence Bands

Plotting each payment method separately in a 2×2 grid to clearly inspect the 95% Greenwood confidence intervals and group characteristics.

In [4]:

```
fig, axes = plt.subplots(2, 2, figsize=(11, 8), sharex=True, sharey=True)
axes = axes.flatten()

for i, method in enumerate(methods):
    ax = axes[i]
    mask = (df['PaymentMethod'] == method)
    kmf = kmf_dict[method]
    kmf.plot_survival_function(ax=ax, color=palette[method], linewidth=2.2,
   ci_show=True, ci_alpha=0.25)

    n_total = len(T[mask])
```

```
    n_churn = E[mask].sum()
    churn_rate = (n_churn / n_total) * 100
    med_val = kmf.median_survival_time_
    med_str = f"{med_val:.1f} mo." if not np.isinf(med_val) else "> 72 mo. (N/A)"

    ax.set_title(f"{method}\n(n={n_total:,}, Churn={churn_rate:.1f}%, Median={med_str})", fontsize=11, fontweight='bold')
    ax.set_ylim(0, 1.05)
    ax.set_xlim(0, 75)
    ax.grid(True, linestyle='--', alpha=0.5)
    ax.set_xlabel('Tenure (Months)', fontsize=10)
    ax.set_ylabel('Survival Probability $S(t)$', fontsize=10)
    ax.axhline(0.5, color='gray', linestyle=':', alpha=0.5)

fig.suptitle('Individual Kaplan-Meier Survival Estimates with 95% Confidence Intervals', fontsize=14, fontweight='bold', y=0.98)
plt.tight_layout()
plt.savefig('plot_3_km_grid_by_payment.pdf', bbox_inches='tight')
plt.close()
print("Plot 3 saved to plot_3_km_grid_by_payment.pdf")
```

Out[4]:

```
Plot 3 saved to plot_3_km_grid_by_payment.pdf
```

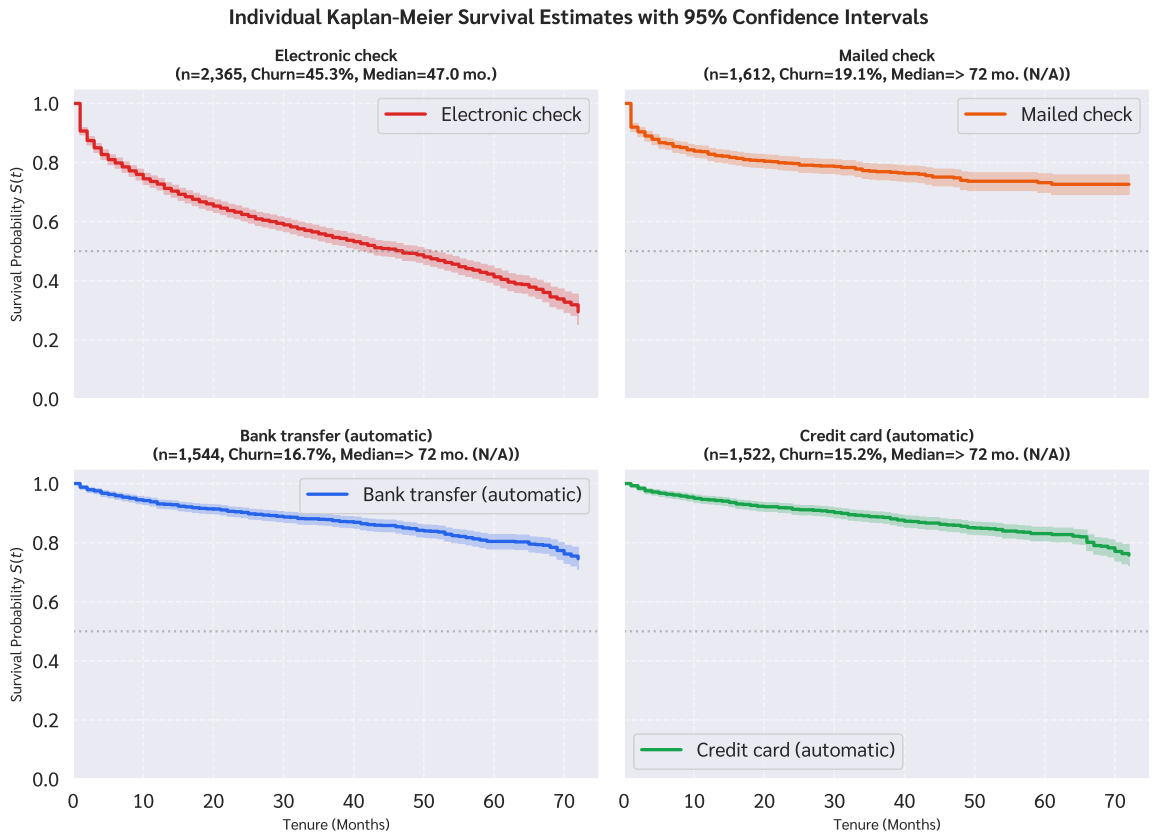


Figure 3 (Individual Survival Facets with 95% CI): 2×2 facet grid of Kaplan-Meier survival curves with 95% confidence bands for each payment method. The tight confidence intervals across all four subplots confirm high statistical precision resulting from large subgroup sample sizes ($n \geq 1,522$ per group).

5. Cumulative Hazard Function Analysis

The cumulative hazard function $\Lambda(t) = \int_0^t \lambda(u)\,du = -\ln S(t)$ models the accumulated risk of customer churn over time.

In [5]:

```
fig, ax = plt.subplots(figsize=(9, 5.5))
naf_dict = {}

for method in methods:
    mask = (df['PaymentMethod'] == method)
    naf = NelsonAalenFitter()
    naf.fit(T[mask], event_observed=E[mask], label=method)
    naf_dict[method] = naf
    naf.plot_cumulative_hazard(ax=ax, color=palette[method], linewidth=2.2,
   ci_show=True, ci_alpha=0.12)
```

```python

ax.set_title('Cumulative Hazard $\\Lambda(t)$ by Payment Method', fontsize=14, fontweight='bold', pad=12)
ax.set_xlabel('Tenure (Months)', fontsize=12)
ax.set_ylabel('Cumulative Hazard $\\Lambda(t) = -\\ln S(t)$', fontsize=12)
ax.set_xlim(0, 75)
ax.grid(True, linestyle='--', alpha=0.5)
ax.legend(title='Payment Method', loc='upper left', frameon=True, framealpha=0.9)
plt.tight_layout()
plt.savefig('plot_4_cumulative_hazard.pdf', bbox_inches='tight')
plt.show()
```

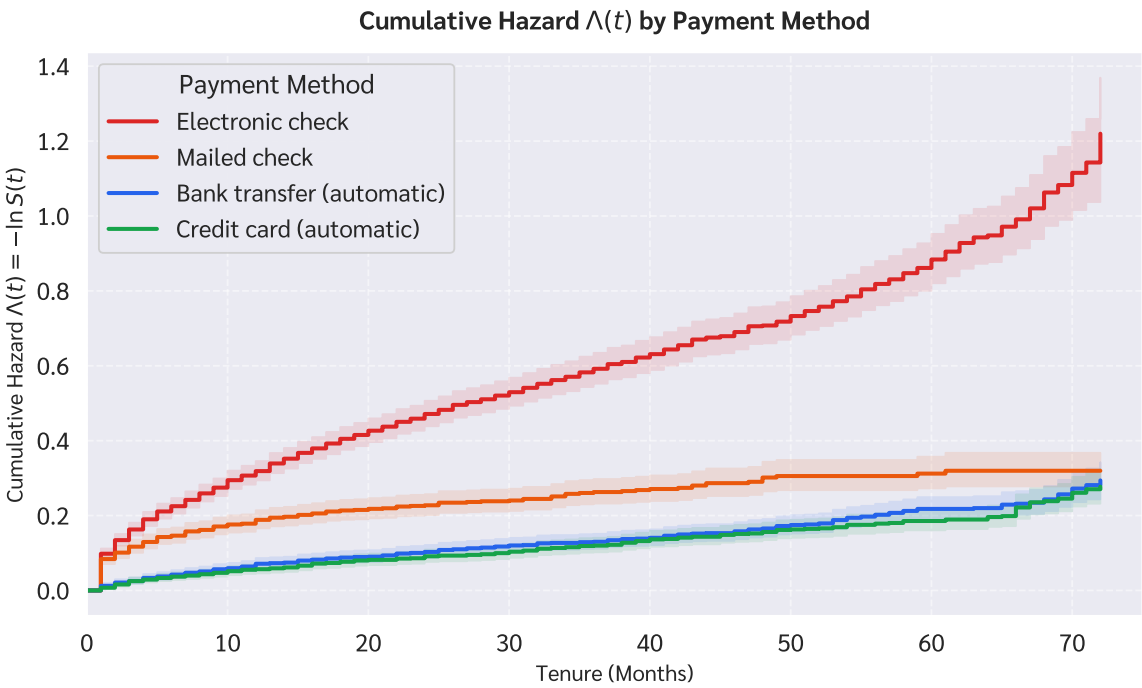


Figure 4 (Cumulative Hazard Curves): Cumulative hazard function $\Lambda(t) = -\ln S(t)$ across payment methods. Electronic check exhibits a steeply ascending, quasi-linear hazard rate surpassing $\Lambda(72) > 1.2$, indicating severe ongoing churn risk. In contrast, automated methods exhibit flat, gradual hazard accumulation ($\Lambda(72) \approx 0.28$ to 0.30).

6. Distribution Analysis of Churn Rate & Tenure

Examining the overall churn percentage and the distribution of customer tenure across the four payment methods.

In [6]:

```
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 5))

# Churn Rate Bar Chart
churn_summary = df.groupby('PaymentMethod')['Churn'].apply(lambda x: (x == 'Yes').mean() * 100).reindex(methods)
bars = ax1.bar(range(len(methods)), churn_summary, color=[palette[m] for m in methods], edgecolor='black', alpha=0.85, width=0.6)
ax1.set_xticks(range(len(methods)))
ax1.set_xticklabels([m.replace(' (automatic)', '\n(auto)') for m in methods], fontsize=10)
ax1.set_ylabel('Churn Rate (%)', fontsize=11)
ax1.set_title('Customer Churn Rate by Payment Method', fontsize=12, fontweight='bold')
ax1.set_ylim(0, 55)
ax1.grid(True, linestyle='--', axis='y', alpha=0.5)
for bar in bars:
    yval = bar.get_height()
    ax1.text(bar.get_x() + bar.get_width()/2.0, yval + 1, f"{yval:.1f}%", ha='center', va='bottom', fontweight='bold', fontsize=10)

# Tenure Boxplot
data_to_plot = [df[df['PaymentMethod'] == m]['tenure'] for m in methods]
bp = ax2.boxplot(data_to_plot, patch_artist=True, tick_labels=[m.replace(' (automatic)', '\n(auto)') for m in methods],
                 medianprops=dict(color='black', linewidth=1.5))
for patch, m in zip(bp['boxes'], methods):
    patch.set_facecolor(palette[m])
    patch.set_alpha(0.6)
ax2.set_ylabel('Tenure (Months)', fontsize=11)
ax2.set_title('Tenure Distribution by Payment Method', fontsize=12, fontweight='bold')
ax2.grid(True, linestyle='--', axis='y', alpha=0.5)

plt.tight_layout()
plt.savefig('plot_5_churn_distribution.pdf', bbox_inches='tight')
plt.show()
```

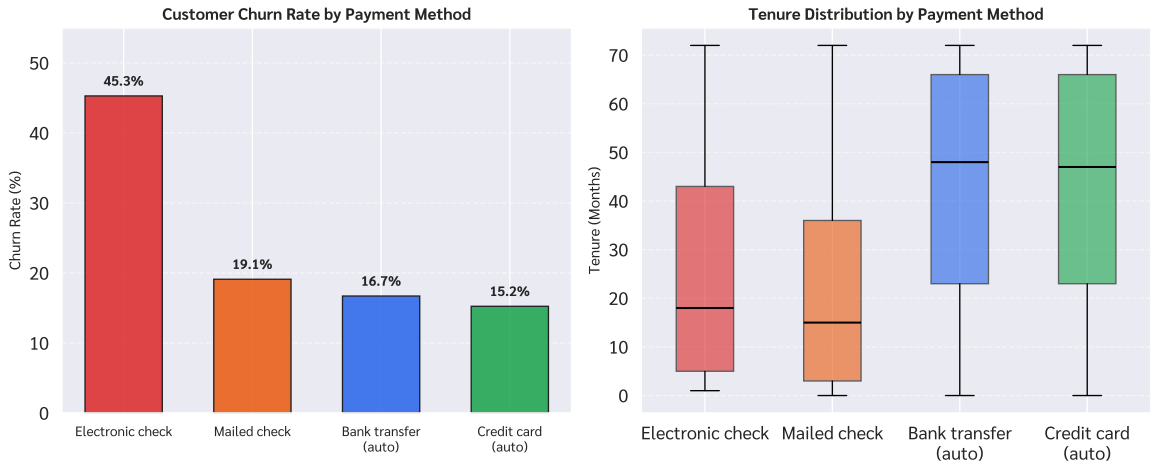


Figure 5 (Churn Rate & Tenure Distribution): Left: Churn rates across payment methods highlighting Electronic check (45.3%) vs. Mailed check (19.1%), Bank transfer (16.7%), and Credit card (15.2%). Right: Boxplots of tenure demonstrating substantially higher customer loyalty and tenure in automated payment channels (median tenure ≈ 43 months) compared to manual payment methods.

7. Statistical Hypothesis Testing & Quantitative Summary

We perform the **Multivariate Log-Rank Test** to test $H_0 : S_1(t) = S_2(t) = S_3(t) = S_4(t)$, followed by pairwise comparisons and milestone survival metrics compilation.

In [7]:

```
# --- Log-Rank Hypothesis Testing ---
logrank_res = multivariate_logrank_test(T, df['PaymentMethod'], E)
print("="*60)
print("MULTIVARIATE LOG-RANK TEST RESULTS")
print("="*60)
print(f"Chi-Square: {logrank_res.test_statistic:.4f}")
print(f"p-value:    {logrank_res.p_value:.4e}")
print(f"df:         {logrank_res.degrees_of_freedom}")
print(f"Conclusion: Reject H0 at alpha = 0.001")

pairwise_res = pairwise_logrank_test(T, df['PaymentMethod'], E)
print("\n" + "="*60)
print("PAIRWISE LOG-RANK TEST SUMMARY")
print("="*60)
for pair, row in pairwise_res.summary.iterrows():
    p_val = row['p']
    p_str = f"{p_val:.4e}" if p_val < 0.001 else f"{p_val:.4f}"
```

```python
    print(f"{pair[0][:15]:<16} vs {pair[1][:15]:<16} | chi2={row['test_statistic']:>7.2f} | p={p_str}")

# --- Quantitative Summary Table ---
summary_list = []
for m in methods:
    mask = (df['PaymentMethod'] == m)
    kmf = kmf_dict[m]
    n_tot = len(T[mask])
    n_ev = E[mask].sum()
    n_cens = n_tot - n_ev
    ch_rate = (n_ev / n_tot) * 100
    med_s = kmf.median_survival_time_
    med_str = f"{med_s:.1f} mo." if not np.isinf(med_s) else "> 72 mo."
    mean_t = T[mask].mean()

    summary_list.append({
        'Payment Method': m,
        'Sample (n)': n_tot,
        'Churned (d)': n_ev,
        'Censored (c)': n_cens,
        'Churn Rate': f"{ch_rate:.1f}%",
        'Median Survival': med_str,
        'Mean Tenure': f"{mean_t:.1f} mo.",
        'S(12 mo.)': f"{kmf.predict(12):.3f}",
        'S(24 mo.)': f"{kmf.predict(24):.3f}",
        'S(36 mo.)': f"{kmf.predict(36):.3f}",
        'S(48 mo.)': f"{kmf.predict(48):.3f}",
        'S(60 mo.)': f"{kmf.predict(60):.3f}",
        'S(72 mo.)': f"{kmf.predict(72):.3f}"
    })

sum_df = pd.DataFrame(summary_list)
print("\n" + "="*60)
print("COMPREHENSIVE SURVIVAL SUMMARY TABLE")
print("="*60)
sum_df
```

Out[7]:

```
============================================================
MULTIVARIATE LOG-RANK TEST RESULTS
============================================================
Chi-Square: 865.2395
```

```
p-value:    3.0665e-187
df:         3
Conclusion: Reject H0 at alpha = 0.001

============================================================
PAIRWISE LOG-RANK TEST SUMMARY
============================================================
Bank transfer (  vs Credit card (au  | chi2=   0.87 | p=0.3514
Bank transfer (  vs Electronic chec  | chi2= 510.04 | p=6.2314e-113
Bank transfer (  vs Mailed check     | chi2=  51.07 | p=8.9046e-13
Credit card (au  vs Electronic chec  | chi2= 539.74 | p=2.1476e-119
Credit card (au  vs Mailed check     | chi2=  64.82 | p=8.2263e-16
Electronic chec  vs Mailed check     | chi2= 152.46 | p=5.0383e-35

============================================================
COMPREHENSIVE SURVIVAL SUMMARY TABLE
============================================================
              Payment Method  Sample (n)  ...  S(60 mo.)  S(72 mo.)
0           Electronic check        2365  ...      0.413      0.295
1               Mailed check        1612  ...      0.732      0.726
2  Bank transfer (automatic)        1544  ...      0.804      0.745
3    Credit card (automatic)        1522  ...      0.831      0.758

[4 rows x 13 columns]
```

8. Summary of Results & Domain Discussion

Summary of Quantitative Findings:

1. **Electronic Check Customers (High Risk Channel)**:
• Has the largest subgroup ($n = 2,365, 33.6$% of total cohort) but suffers an alarmingly high churn rate of 45.3% ($1,071$ churn events).
• Reaches median survival time at 47.0 **months** ($S(47.0) = 0.500$).
• Only 72.7% survive past 1 year ($t = 12$), dropping to 62.4% at 2 years ($t = 24$) and 29.5% at 6 years ($t = 72$).
• **Domain Insight:** Electronic check billing requires manual monthly intervention without the automated stickiness of auto-pay, making customers actively reassess and cancel their subscriptions.

2. **Automatic Payment Channels (Bank Transfer & Credit Card - Low Risk Loyalty Channels)**:
• Both automated channels display exceptional, virtually identical survival characteristics ($p = 0.351$, not statistically different from each other).
• **Credit Card (automatic):** Churn rate = 15.2%, $S(12) = 0.945$, $S(24) = 0.913$, $S(72) = 0.758$, Mean tenure = 43.3 months.

• **Bank Transfer (automatic):** Churn rate = 16.7%, $S(12) = 0.931$, $S(24) = 0.902$, $S(72) = 0.745$, Mean tenure = 43.7 months.
• Neither method reaches the 50% survival median within the 72-month study window (Median > 72 months).
• **Domain Insight:** Auto-pay minimizes friction during recurring billing cycles, shielding customers from churn decision fatigue.

3. **Mailed Check Customers (Moderate Churn with Long-Term Stabilization)**:
• Sample size $n = 1,612$, Churn rate = 19.1%, Mean tenure = 21.8 months.
• Exhibits steep initial churn ($S(12) = 0.828, S(24) = 0.797$) during trial onboarding, but the curve flattens significantly after 2 years ($S(72) = 0.726$).

Strategic Business Recommendations:

• **Incentivize Auto-Pay Adoption:** Implement promotional discounts ($5/mo discount for 6 months) to convert Electronic check and Mailed check users to Credit card or Bank transfer auto-pay.
• **Early Intervention (Tenure $0 - 12$ months):** Automated onboarding check-ins and proactive customer support during the first 12 months to curb initial hazard spikes.