Kapitel 3, Diskreta fördelningar

Lantz: *Grundläggande Statistisk Analys*

Kapitel 3 Lärandemål (LM)

LM 3.1: Kunna skilja på diskreta och kontinuerliga slumpvariabler.

LM 3.2: Beskriva fördelningen för en diskret slumpvariabel.

LM 3.3: Beräkna och tolka beskrivande mått för en diskret slumpvariabel.

LM 3.4: Beskriva binomialfördelningen och beräkna sannolikheter för denna.

Kapitel 3 Lärandemål (LM)

LM 3.5: Beskriva Poissonfördelningen och beräkna sannolikheter för denna.

LM 3.6: Beskriva den hypergeometriska fördelningen och beräkna sannolikheter för denna

LM 3.7: Beskriva den geometriska fördelningen och beräkna sannolikheter för denna

LM 3.8: Beskriva negativ binomialfördelning och beräkna sannolikheter för denna

Hur mycket personal behövs?

- Sture är manager på ett lokalt Starbucks. Pga svag ekonomi och högre bensin- och matpriser tillkännagav Starbucks att man skulle stänga 500 caféer i USA.
- Sture café ska fortsätta vara öppet, men han funderar på hur nedläggningar av närliggande caféer kommer påverka hans affärer.
- En typisk Starbuckskund besöker kedjan 15-18 gånger per månad.
- Baserat på detta tror Sture att kunder kommer besöka hans café 18 ggr på 30 dagar i medeltal.

Hur mycket personal behövs?

- Sture behöver bestämma hur mycket personal som behövs.
  - För många anställda blir kostsamt.
  - För få anställda kan resultera i att tappa kunder som blir sura över att få vänta för länge på service.

- Med förståelse för sannolikhetsfördelningen för ankommande kunder kan Sture:
  - Beräkna förväntat antal kunder under en viss tidsperiod.
  - Beräkna sannolikheten att en typisk kund besöker kedjan ett visst antal gånger under en viss tidsperiod.

3.1 Slumpvariabler och diskreta fördelningar

LM 3.1 Skilja på diskreta och kontinuerliga slumpvariabler.

- **Slumpvariabel**
  - En funktion som ger numeriska värden åt utfall hos ett experiment.
  - Skrivs med stora bokstäver, till exempel X.

- Värden som slumpvariabeln kan ta skrivs med små bokstäver:

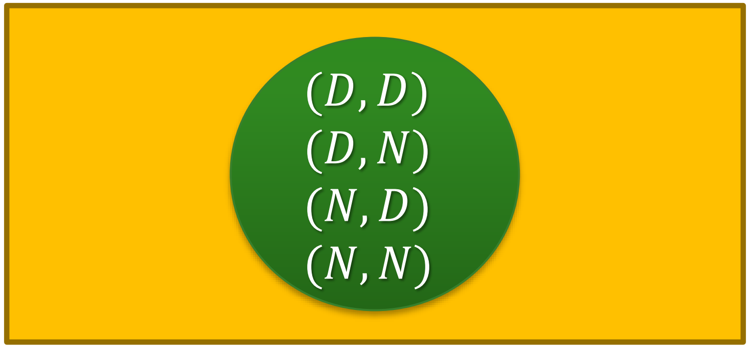
$$x_1, x_2, x_3, \ldots$$

LM 3.1 3.1 Slumpvariabler och diskreta fördelningar

- Slumpvariabler kan klassas som:
  - **Diskreta**
    - Kan anta ett uppräkneligt antal olika värden (i separerade punkter).
  - **Kontinuerliga**
    - Kan ta överuppräkneligt (det går inte att skapa en lista) många värden inom något intervall.

LM 3.1 3.1 Slumpvariabler och diskreta fördelningar

- Betrakta ett experiment där två tröjor tas ut från en produktionslinje där en tröja kan vara defekt (D) eller icke-defekt (N).
  - Här är utfallsrummet:
  - Slumpvariabeln X är antalet defekta tröjor.
  - Slumpvariabeln tar värden i mängden $\{0, 1, 2\}$.



- Då dessa är de enda möjliga värdena är X en diskret slumpvariabel.

3.1 Slumpvariabler och diskreta fördelningar

LM 3.2 Beskriva fördelningen för en diskret slumpvariabel.

- Varje slumpvariabel är “kopplad” till en (sannolikhets)**fördelning** som fullständigt beskriver slumpvariabeln.
  - En **massfunktion** används för att beskriva diskreta slumpvariabler.
  - En **täthetsfunktion** används för att beskriva kontinuerliga slumpvaribler.
  - En **(kumulativ) fördelningsfunktion** kan användas för att beskriva både diskreta och kontinuerliga slumpvariabler.

LM 3.2

3.1 Slumpvariabler och diskreta fördelningar

- **Massfunktionen** för en (diskret) slumpvariabel X ges av

$$P(X = x)$$

- **Fördelningsfunktionen** för X ges av

$$P(X \leq x)$$

LM 3.2 3.1 Slumpvariabler och diskreta fördelningar

- Två viktiga egenskaper hos diskreta fördelningar:
  - Sannolikheten för varje värde x är ett värde mellan 0 och 1, alltså

$$0 \leq P(X = x) \leq 1$$

  - Summan av sannolikheterna är 1. Alltså,

$$\sum P(X = x) = 1$$

där summan tas över alla möjliga värden x.

LM 3.2 3.1 Slumpvariabler och diskreta fördelningar

- En diskret fördelning kan ses som en tabell, algebraiskt eller grafiskt.
- Om vi till exempel slår en sexsidig tärning kan vi beskriva fördelningen enligt tabellen:

x	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

- Varje utfall har sannolikhet 1/6. Massfunktionen för X (antalet prickar tärningen visar) ges alltså av tabellen.

LM 3.2 3.1 Slumpvariabler och diskreta fördelningar

- Vi kan också göra tabeller för fördelningsfunktionen.
  - Om vi även här betraktar kast med sexsidig tärning får vi fördelningsfunktionen enligt

x	1	2	3	4	5	6
$P(X \leq x)$	1/6	2/6	3/6	4/6	5/6	6/6

- Fördelningsfunktionen ger alltså sannolikheten att X blir mindre än eller lika med x.
  Till exempel,

$$P(X \leq 4) = \frac{2}{3}$$

LM 3.2 3.1 Slumpvariabler och diskreta fördelningar

- En fördelning kan uttryckas algebraiskt.
- För en sexsidig tärning ges fördelningen för antalet prickar, X, av:

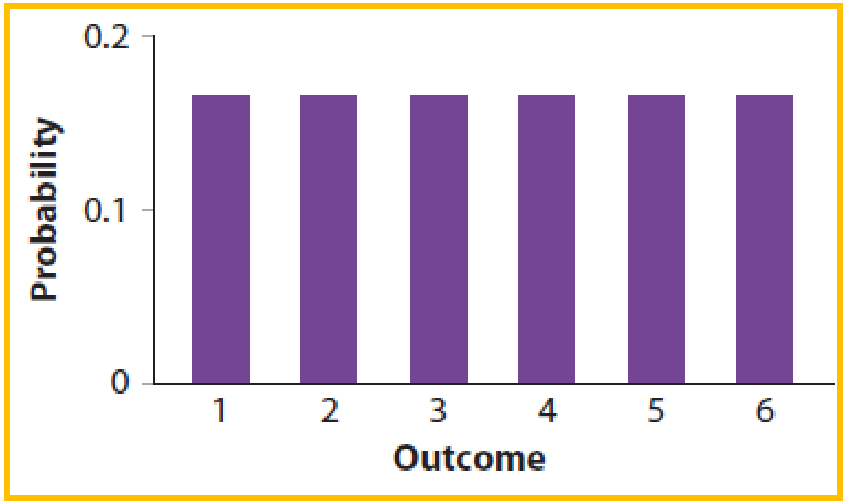
$$P(X = x) = \begin{cases} 1/6 & om\ x = 1,2,3,4,5,6 \\ 0 & annars \end{cases}$$

- Formeln ger till exempel att

$$P(X = 4) = 1/6 \text{ och } P(X = 7) = 0$$

LM 3.2 3.1 Slumpvariabler och diskreta fördelningar

- En fördelning kan uttryckas grafiskt.
  - Värdena x som X kan ta sätts på den horisontella axeln och motsvarande sannolikheter sätts på den vertikala axeln.
  - En linje/stapel dras så att dess höjd representerar sannolikheten för x.
  - Som exempel kan vi ta tärningskasten:
  - Detta är en **likformig** fördelning då alla staplar har samma höjd



LM 3.2 3.1 Slumpvariabler och diskreta fördelningar

- Exempel: Betrakta fördelningen för hur många kredikort besökarna av sidan bankrate.com bär med sig:
  - Är detta en valid fördelning?
  - Vad är sannolikheten för inget kort?
  - Vad är sannolikheten för färre än två kort?
  - Vad är sannolikheten för minst två kort?

Number of Credit Cards	Percentage
0	2.5%
1	9.8
2	16.6
3	16.5
4*	54.6

*denotes 4 or more credit cards.

SOURCE: www.bankrate.com, Financial Literacy Series, 2007.

LM 3.2 3.1 Slumpvariabler och diskreta fördelningar

- Exempel: Betrakta fördelningen för hur många kredikort besökarna av sidan bankrate.com bär med sig:
  - Ja, eftersom $0 \leq P(X = x) \leq 1$ och $\sum P(X = x) = 1$.
  - $P(X = 0) = 0.025$
  - $P(X < 2) = P(X = 0) + P(X = 1)$ $= 0.025 + 0.098 = 0.123$
  - $P(X \geq 2) = P(X = 2) + P(X = 3) + P(X \geq 4)$ $= 0.166 + 0.165 + 0.546 = 0.877$

  Alternativt, $P(X \geq 2) = 1 - P(X < 2) = 1 - 0.123 = 0.877$

Number of Credit Cards	Percentage
0	2.5%
1	9.8
2	16.6
3	16.5
4*	54.6

*denotes 4 or more credit cards.

SOURCE: www.bankrate.com, Financial Literacy Series, 2007.

3.2 Väntevärde, Varians och Standardavvikelse

LM 3.3 Beräkna och tolka deskriptiva mått för en diskret slumpvariabel.

- Beskrivande mått för en slumpvariabel inkluderar
  - Lägesmåttet **väntevärde (expected value)**
  - Spridningsmåttet **varians**
  - Spridningsmåttet **standardavvikelse (standard deviation)**

3.2 Väntevärde, Varians och Standardavvikelse

- **Väntevärde** $\Longleftrightarrow$ **Populationsmedelvärde**

  $$E(X) \Longleftrightarrow \mu$$

  - $E(X)$ är "långtidsmedelvärdet" av oändligt många upprepade experiment, alltså det tal medelvärdet av observationer av slumpvariabeln kommer närma sig då antalet observationer växer. Ibland betcknar vi väntevärdet med μ (my) och kallar det populationsmedelvärde
  - För en diskret slumpvariabel X som tar värden $x_1, x_2, x_3, \ldots$ med sannolikheterna $P(X = x_i)$, är väntevärdet

$$E(X) = \mu = \sum_i x_i P(X = x_i)$$

3.2 Väntevärde, Varians och Standardavvikelse

- **Varians** and **Standardavvikelse**
  - För en diskret slumpvariabel X som tar värden $x_1, x_2, x_3, \ldots$ med sannolikheterna $P(X = x_i)$, är variansen

$$V(X) = \sigma^2 = \sum_i (x_i - \mu)^2 P(X = x_i)$$

$$= \sum_i x_i^2 P(X = x_i) - \mu^2$$

  - Standardavvikelsen är kvadratroten av variansen

$$S(X) = \sigma = \sqrt{\sigma^2}$$

LM 3.3

3.2 Väntevärde, Varians och Standardavvikelse

- Exempel: Brad Williams, som är bilhandlare i Chicago, bestämmer sig för att införa följande “incentive scheme” för sina säljare.

Bonus (in $1,000s)	Performance Type	Probability
$10	Superior	0.15
$6	Good	0.25
$3	Fair	0.40
$0	Poor	0.20

  - Beräkna väntevärdet för (årlig) bonus.
  - Beräkna variansen och standardavvikelsen för (årlig) bonus.

LM 3.3 3.2 Väntevärde, Varians och Standardavvikelse

- Lösning: Låt slumpvariabeln X beteckna den årliga bonusen (i tusentals dollar) för en säljare.

Value, x_i	Probability, $P(X = x_i)$	Weighted Value, $x_i P(X = x_i)$	Weighted Squared Deviation, $(x_i - \mu)^2 P(X = x_i)$
10	0.15	$10 \times 0.15 = 1.5$	$(10 - 4.2)^2 \times 0.15 = 5.05$
6	0.25	$6 \times 0.25 = 1.5$	$(6 - 4.2)^2 \times 0.25 = 0.81$
3	0.40	$3 \times 0.40 = 1.2$	$(3 - 4.2)^2 \times 0.40 = 0.58$
0	0.20	$0 \times 0.20 = 0$	$(0 - 4.2)^2 \times 0.20 = 3.53$
		Total = 4.2	Total = 9.97

  - $E(X) = \mu = \sum x_i P(X = x_i) = 4.2$ \$4,200
  - $V(X) = \sigma^2 = \sum (x_i - \mu)^2 P(X = x_i) = 9.97$ $(\$1,000)^2$.
  - $S(X) = \sigma = \sqrt{\sigma^2} = \sqrt{9.97} = 3.16$ \$3,160

3.2 Väntevärde, Varians och Standardavvikelse

- Räkneregler.
  - Givet en slumpvariabler X och två konstanter a och b

    - Är väntevärdet av $aX + b$

$$E(aX + b) = aE(X) + b$$

    - Är variansen för $aX + b$

$$V(aX + b) = a^2V(X)$$

3.4 Binomialfördelningen

LM 3.4 Beskriva och beräkna sannolikheter för binomialfördelningen.

- En **binomialfördelad slumpvariabel** ges av antalet lyckade försök bland n försök i en s k Bernoulliprocess.
  - En **Bernoulliprocess** består av en serie om n oberoende och identiska försök av ett experiment sådana att för varje försök:
    - Finns bara två möjliga utfall, lyckat eller misslyckat:
      p betecknar sannolikheten för lyckat försök och
      $q = 1 - p$ betecknar sannolikheten för misslyckat försök
    - Varje gång försöket upprepas är sannolikheterna för lyckat och misslyckat samma som i tidigare försök.

LM 3.4 3.4 Binomialfördelningen

- En **binomialfördelning** beskrivs enklast med massfunktionen.
  - För en binomialfördelad slumpvariabel X, ges massfunktionen av

$$P(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

för $x = 0,1, \dots, n$

LM 3.4 3.4 Binomialfördelningen

- För binomialfördelningen gäller:
  - Väntevärdet: $E(X) = \mu = np$
  - Variansen: $V(X) = \sigma^2 = npq$
  - Standardavvikelsen: $S(X) = \sigma = \sqrt{npq}$

LM 3.4 3.4 Binomialfördelningen

- Exempel: Ungefär 20% av arbetarna i USA är rädda att de aldrig ska kunna gå i pension. Antag att ett stickprov om 10 dras.
  - Vad är sannolikheten att ingen av arbetarna är rädd att inte kunna gå i pension?
    - Lösning: Med n=10 och p=0.2 får vi

$$
\begin{aligned}
P(X=0) &= \frac{10!}{0!(10-0)!} \times (0.20)^0 \times (0.80)^{10} \\
&= \frac{10 \times 9 \times \cdots \times 1}{(1) \times (10 \times 9 \times \cdots \times 1)} \times 1 \times (0.80)^{10} = 1 \times 1 \times 0.1074 \\
&= 0.1074.
\end{aligned}
$$

3.5 Poissonfördelningen

LM 3.5 Beskriva Poissonfördelningen och beräkna sannolikheter för den.

- En binomialfördelad slumpvariabel räknar antalet lyckade försök i en serie (oberoende försök) av fix längd.
- En **Poissonfördelad** slumpvariabel räknar antalet lyckade försök/händelser i ett visst tids- (eller rums-) intervall.
- Exempel:
  - *Med avseende på tid*—antal bilar som åker över Brooklyn Bridge mellan 9:00 och 10:00 en måndagmorgon.
  - *Med avseende på rum*—antal defekter i en 50 meter lång tygrulle.

LM 3.5 3.5 Poissonfördelningen

- Ett slumpförsök är en **Poissonprocess** om:
  - Antal lyckade försök inom ett visst tidsintervall är ett heltal ej mindre än noll.
  - Antal lyckade försök i icke-överlappande tidsintervall är oberoende.
  - Sannolikheten att få lyckade försök i något intervall är den samma för alla intervall av samma storlek och proportionell mot intervallets storlek.

LM 3.5 3.5 Poissonfördelningen

- För en Poissonfördelad slumpvariabel X, har vi att sannolikheten för x lyckade försök i ett visst tidsintervall är

$$P(X = x) = e^{-\mu} \frac{\mu^x}{x!}$$

där μ är medelantalet lyckade försök och $e \approx 2.718$ är basen för den naturliga logaritmen.

LM 3.5 3.5 Poissonfördelningen

- För en Poissonfördelad slumpvariabel, X, har vi:
  - Väntevärdet: $E(X) = \mu$
  - Variansen: $V(X) = \mu$
  - Standardavvikelsen: $S(X) = \sqrt{\mu}$

LM 3.5 3.5 Poissonfördelningen

- Exempel: Starbucks igen; Sture tror att en genomsnittlig Starbuckskund “snittar på” 18 besök under en 30-dagars period.
  - Hur många besök kan Sture förvänta sig över en femdagarsperiod av en genomsnittlig kund?

Given the rate of 18 visits over a 30-day month, we can write the mean for the 30-day period as $\mu_{30} = 18$. For this problem, we compute the proportional mean for a 5-day period as $\mu_5 = 3$ because $\frac{18 \text{ visits}}{30 \text{ days}} = \frac{3 \text{ visits}}{5 \text{ days}}$.

  - Vad är sannolikheten en kund besöker kedjan fem gånger under en femdagarsperiod?

$$P(X = 5) = \frac{e^{-3}3^5}{5!} = \frac{(0.0498)(243)}{120} = 0.1008$$

3.5 Binomialfördelningen och Poissonfördelningen

- Om $X \sim B(n, p)$ där n är stort, p är litet och np är lagom stort gäller att

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \approx e^{-np} \frac{(np)^x}{x!}$$

- Tumregler:
  - $n > 10$
  - $p < 0.1$
  - $0.01 < np < 50$

3.6 Hypergeometrisk fördelning

LM 3.6 Beskriva den hypergeometriska fördelningen och beräkna sannolikheter för den.

- En binomialfördelad slumpvariabel X räknar antal lyckade försök i en serie om n försök som är
  - Oberoende
  - Sannolikheten för lyckat försök är konstant mellan försöken.
- En **hypergeometriskt fördelad** slumpvariabel räknar antalet lyckade försök i en serie om n försök som inte kan antas oberoende.

LM 3.6 3.6 Hypergeometrisk fördelning

- Hypergeometrisk fördelning kan man använda när man vill beskriva dragning utan återläggning där populationsstorleken påverkar förutsättningarna för dragningarna.
  - För en **hypergeometrisk slumpvariabel** X, är sannolikheten för x lyckade försök bland n slumpmässigt valda;

$$P(X = x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}$$

för $x = 0,1, \ldots, n$ om $n \leq R$ och $x = 0,1, \ldots, R$ om $n > R$
där N är populationsstorleken och R är antal lyckade i populationen.

LM 3.6 3.6 Hypergeometrisk fördelning

- För en hypergeometriskt fördelad slumpvariabel är:
  - **Väntevärdet**: $E(X) = n\frac{R}{N}$
  - **Variansen**: $V(X) = n\frac{R}{N}\left(1 - \frac{R}{N}\right)\frac{N-n}{N-1}$
  - **Standardavvikelsen**: $S(X) = \sqrt{n\frac{R}{N}\left(1 - \frac{R}{N}\right)\frac{N-n}{N-1}}$

LM 3.6 3.6 Hypergeometrisk fördelning

- Exempel: I en jourlivs i Majorna, inspekterar föreståndaren fem slumpmässigt valda mango i en låda om 20 mango för att se om det finns transportskador. Antag att lådan innehåller precis två skadade mango.
  - Vad är sannolikheten att en av fem inspekterade mango är transportskadad?
  - The probability that one out of five mangoes is damaged is $P(X = 1)$. We calculate

$$P(X = 1) = \frac{\binom{2}{1}\binom{20-2}{5-1}}{\binom{20}{5}} = \frac{\left(\frac{2!}{1!1!}\right)\left(\frac{18!}{4!14!}\right)}{\left(\frac{20!}{5!15!}\right)} = \frac{(2)(3060)}{15{,}504} = 0.3947.$$

Therefore, the likelihood that exactly one out of five mangoes is damaged is 39.47%.

3.7 Geometrisk fördelning

LM 3.7 Beskriva den geometriska fördelningen och beräkna sannolikheter för den.

- För en Bernoulliprocess kan man fråga sig: Vad är sannolikheten att vi behöver göra x försök tills vi får ett lyckat försök

- En **geometriskt fördelad** slumpvariabel räknar antalet försök som krävs till och med det då vi får det första lyckade i en Bernoulliprocess.

LM 3.7

3.7 Geometrisk fördelning

- Massfunktionen
  - För en **geometriskt fördelad slumpvariabel** X, är sannolikheten att det krävs x försök för att få ett lyckat;

$$P(X = x) = pq^{x-1}$$

LM 3.7

3.7 Geometrisk fördelning

- För en geometriskt fördelad slumpvariabel är:
  - **Väntevärdet**: $E(X) = \frac{1}{p}$
  - **Variansen**: $V(X) = \frac{1-p}{p^2}$
  - **Standardavvikelsen**: $S(X) = \frac{\sqrt{1-p}}{p}$

LM 3.7

3.7 Geometrisk fördelning

- Vad är sannolikheten att man får ett jobb i femte försöket om vi antar att antalet jobbansökningar tills man får ett jobb är $G(0.2)$?
- Lösning: $P(X = 5) = 0.2 \cdot 0.8^4 \approx 0.08$

3.8 Negativ binomialfördelning

LM 3.8 Beskriva negativ binomialfördelning och beräkna sannolikheter för den.

- För en Bernoulliprocess kan man fråga sig: Vad är sannolikheten att vi behöver göra x försök som måste göras tills vi har s lyckade försök

- En slumpvariabel med negativ binomialfördelning räknar antalet försök som krävs till och med det då vi har s lyckade i en Bernoulliprocess.

LM 3.8 3.8 Negativ binomialfördelning

- Massfunktionen
  - För en **slumpvariabel** X med **negativ binomial-fördelning**, är sannolikheten att det krävs x försök för att få s lyckade;

$$P(X = x) = \binom{x-1}{s-1} p^s q^{x-s}$$

OBS! Annorlunda parametrisering I SciPy!

$$P(X = x) = \binom{x+s-1}{s-1} p^s q^x$$

LM 3.8 3.8 Negativ binomialfördelning

- För en slumpvariabel med negativ binomialfördelning är:
  - **Väntevärdet**: $E(X) = \frac{s}{p}$
  - **Variansen**: $V(X) = \frac{s(1-p)}{p^2}$
  - **Standardavvikelsen**: $S(X) = \frac{\sqrt{s(1-p)}}{p}$

LM 3.8

3.8 Negativ binomialfördelning

- Vad är sannolikheten att man i sjunde tärningskastet får den andra sexan?
- Lösning:

$$P(X = x) = \binom{x-1}{s-1} p^s q^{x-s} = \binom{7-1}{2-1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{7-2}$$

$$= 6 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^5 \approx 0.067$$