Kap. 5.2: Utvalgsfordelinger for antall og andeler

- Binære data (1/0, Ja/Nei, Suksess/Feil)
  - Utvalgsundersøkelser: Ja/Nei-spørsmål
  - Tilstedeværelse av arter: Tilstede/Ikke-tilstede (1/0)
  - Overlevelse etter behandling: Ja/Nei
  - Observator X = Antall Ja eller antall 1-ere for utvalget av størrelse *n*
  - Observator $\hat{p}=X/n$ er andel i utvalget med Ja eller 1-ere

Binomisk setting

- Fast antall observasjoner *n*
- De *n* observasjonene er uavhengige
- To mulige utfall av hver observasjon:
  - Kalles Suksess/Feil
  - Tilsvarer f.eks. Ja/Nei eller 1/0
- Sannsynlighet *p* for suksess for hver av de *n* observasjonene

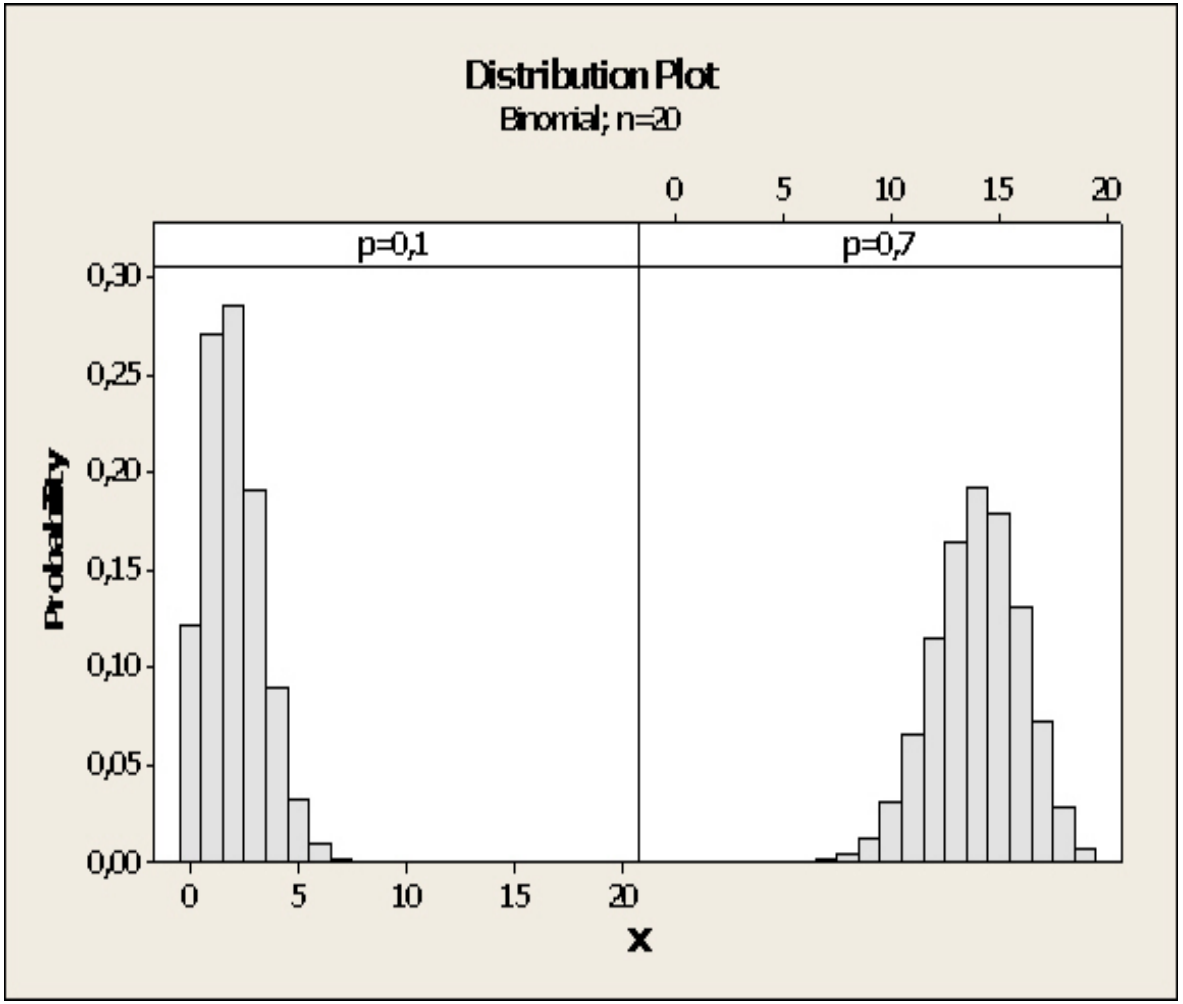
Binomisk fordeling

- Fordeling til antallet *X* av suksesser i en binomisk setting
  - Binomisk fordeling med parametre *n* (antall observasjoner) og *p* (sannsynligheten for suksess for hver observasjon)
  - Utfallsrom {*0,1,...,n*}
  - X er Bin(n,p)-fordelt
- Viktig *diskret* fordeling (sannsynlighetsfordeling for en diskret stokastisk variabel X)

Binomisk fordelte data

- Myntkast med idealisert mynt
  - Kaster en mynt n=10 ganger
  - Sannsynlighet for kron er p=0.5
  - X=Antall kron i de 10 kastene (antallet suksesser)
  - X er Bin(10,0.5)-fordelt
- Genetikk tilsier at barn av samme foreldre får gener fra foreldrene uavhengig av hverandre
  - To foreldre får n=5 barn sammen
  - Hvert barn disse foreldrene får har sannsynlighet p=0.25 for å få blodtype 0
  - X=Antall barn som får blodtype 0 (antallet suksesser)
  - X er Bin(5,0.25)-fordelt

Binomisk fordeling: Sannsynlighets-histogrammer



Utvalgsfordeling for antall suksesser

- Populasjon av størrelse N, andel suksess i populasjonen p
- Utvalg av størrelse n, observatoren X er antall suksesser i utvalget
- Utfall av 2. observasjon avhenger av utfall av 1. observasjon
- Eksempel
  - N=52 kort, trekker n=2 kort
  - P(1. Rødt)=26/52=0.5, P(2. Rødt | 1. Rødt)=25/51<0.5
- Avhengighet mellom observasjonene
- MEN: Hvis N mye større enn n $(N>20n)$, kan man neglisjere slik avhengighet, og X er tilnærmet Bin(n,p)-fordelt
- Presisjonen til denne tilnærmelsen er bedre jo større forholdet N/n er

Binomiske sannsynligheter

- Fordeling til antallet *X* av suksesser
  - Fordeling med parametre *n* (antall observasjoner) og *p* (sannsynligheten for suksess for hver observasjon)
  - Utfallsrom {*0,1,...,n*}
  - X er Bin(n,p)-fordelt
- Sannsynligheten for at X=i, for i=0,1,...,n kan finnes i tabell (Table C i boken) eller ved å bruke dataprogram
  - Avhenger kun av *n* og *p*, dvs for gitt *n* og *p* er sannsynligheten for at X=i bestemt
  - Eksempel: n=6, p=0.35, da er P(X=2)=

Binomiske sannsynligheter: Eksempel

- Genetikk tilsier at barn av samme foreldre får gener fra foreldrene uavhengig av hverandre
  - To foreldre får n=5 barn sammen
  - Hvert barn disse foreldrene får har sannsynlighet p=0.25 for å få blodtype 0
  - X=Antall barn som får blodtype 0 (antallet suksesser)
  - X er Bin(5,0.25)-fordelt
- Hva er sannsynligheten for at minst 2 av barna får blodtype 0?

TABLE C

Binomial probabilities (continued)

Entry is $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

		p								
n	k	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563
	2	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5		.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0937
	6			.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6		.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7				.0001	.0002	.0006	.0016	.0037	.0078
8	0	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0313
	2	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6		.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7			.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8					.0001	.0002	.0007	.0017	.0039

(*Continued*)

Binomiske sannsynligheter: Eksempel

- Hva er sannsynligheten for at minst 2 av barna får blodtype 0?
- $P(X \geq 2) = 1 - P(X < 2) = 1 - P(X \leq 1) = 1 - (P(X=0) + P(X=1)) =$

Table C: Bare for $p \leq 0.5$

- Dersom man ser etter sannsynlighetsfordelingen til X som er binomisk fordelt med $p>0.5$:
  - Snu om på situasjonen slik at Y teller antallet feil (i stedet for suksesser)
  - Da blir $p<0.5$ for Y som teller antall feil
  - Eksempel:
    - Antall barn som ikke har blodtype 0 er Bin(5,0.75)-fordelt
    - Antall barn som har blodtype 0 er Bin(5,0.25)-fordelt
- Tenk alltid nøye igjennom hva man teller som suksess og hva den riktige p er da!

Binomiske sannsynligheter: Eksempel

- Tenk om eksempelet var formulert slik i stedet:
  - Genetikk tilsier at barn av samme foreldre får gener fra foreldrene uavhengig av hverandre. To foreldre får n=5 barn sammen. Hvert barn disse foreldrene får har sannsynlighet p=0.75 for **ikke** å få blodtype 0
  - Y=Antall barn som ikke får blodtype 0 (antallet suksesser)
- Hva er sannsynligheten for at maksimalt 3 av barna ikke får blodtype 0, dvs at $Y \leq 3$?
  - Y er Bin(5,0.75)-fordelt – kan ikke finne sannsynligheter i Table C (fordi p>0.5)!
  - Omformuler: X= antall barn som får blodtype 0, X er Bin(5,0.25)-fordelt
  - Finn sannsynligheten for at minst 2 av barna får blodtype 0

Forventning i binomisk fordeling

- Anta X er Bin(n,p)
- La S_i være en binær stokastisk variabel som indikerer om observasjon *i* er en suksess ($S_i=1$) eller ikke ($S_i=0$)
- Da er $X=S_1+S_2+...+S_n$ (antallet suksesser)
- $P(S_i=1)$ = og $P(S_i=0)$ =
- Forventningen til hver S_i er
- Forventningen til er

Varians og standardavvik i binomisk fordeling

- $\sigma_S^2 = (1-p)^2 * p + (0-p)^2 * (1-p) = p(1-p)$
- $\sigma_S = \sqrt{[p(1-p)]}$
- $X = S_1 + S_2 + ... + S_n$
- S_i-ene er uavhengige av hverandre (binomisk setting), dvs alle parvise korrelasjoner er 0
- $\sigma_X^2 = \sigma_S^{2+} \sigma_S^2 + ... + \sigma_S^2 = n*p(1-p)$
- $\sigma_X = \sqrt{[np(1-p)]}$

Andeler

- $\hat{p}=X/n$=antall suksesser/størrelse av utvalg: andelen suksesser i utvalget
- $\hat{p}$ er estimator for andelen suksesser i populasjonen
- X tar heltallsverdier mellom 0 og *n* og er Bin(n,p)-fordelt
- $\hat{p}$ tar verdier i intervallet [0,1] og er *ikke* binomisk fordelt!
- Men kan bruke forventning og varians til X til å finne forventning og varians til $\hat{p}$:
  - $\mu_{\hat{p}}=np/n=p$ - Forventningsrett estimator for p!
  - $\sigma_{\hat{p}}^2=np(1-p)/n^2=p(1-p)/n$
  - $\sigma_{\hat{p}}=\sqrt{[p(1-p)/n]}$
- Variasjonen (usikkerheten) minker med økende n!
- $\sqrt{n}$ i nevneren for standardavviket betyr at dersom vi ønsker å halvere standardavviket til $\hat{p}$, må vi firedoble utvalgsstørrelsen *n*

Tilnærming til normalfordeling

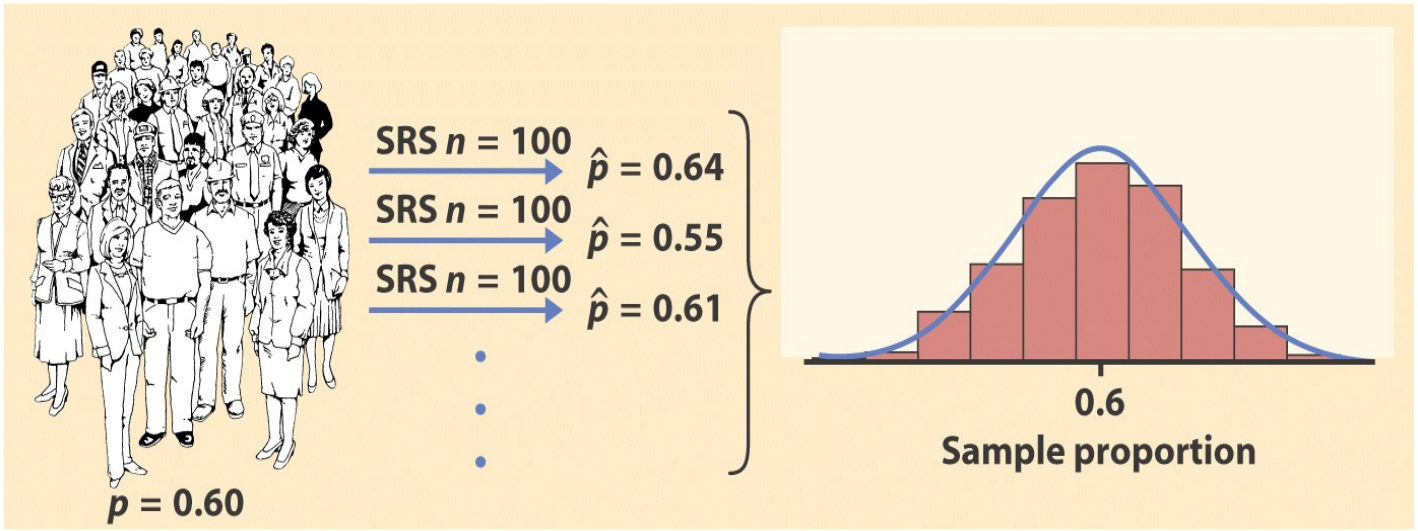
- *X* er Bin(*n*,*p*)-fordelt og *n* er stor. Da er

$$X \text{ tilnærmet } N\left(np\,,\sqrt{np\left(1-p\right)}\right)$$

$$\hat{p} \text{ tilnærmet } N\left(p\,,\sqrt{p\left(1-p\right)/n}\right)$$

- Kan brukes for å beregne sannsynligheter
- Rimelig tilnærming når np>10 og n(1-p)>10

Utvalgsfordeling: Gir svaret på hva som ville skje dersom vi så på mange utvalg med størrelse *n* fra den samme populasjonen



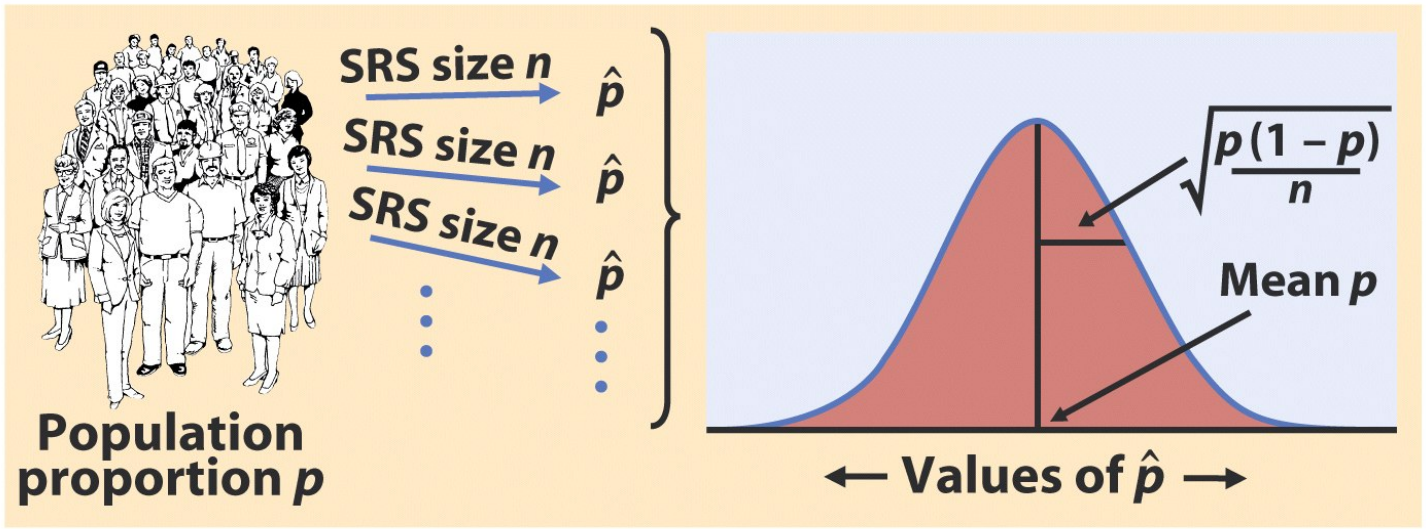


Figure 5-4
Introduction to the Practice of Statistics, Fifth Edition

Sannsynlighetshistogram og normalfordelings-tilnærmingen når X er Bin(150,0.08)-fordelt (Her: np=12, n(1-p)=138)
Sannsynlighetshistogrammet er litt høyreskjevt, noe normalfordelingen ikke kan fange opp

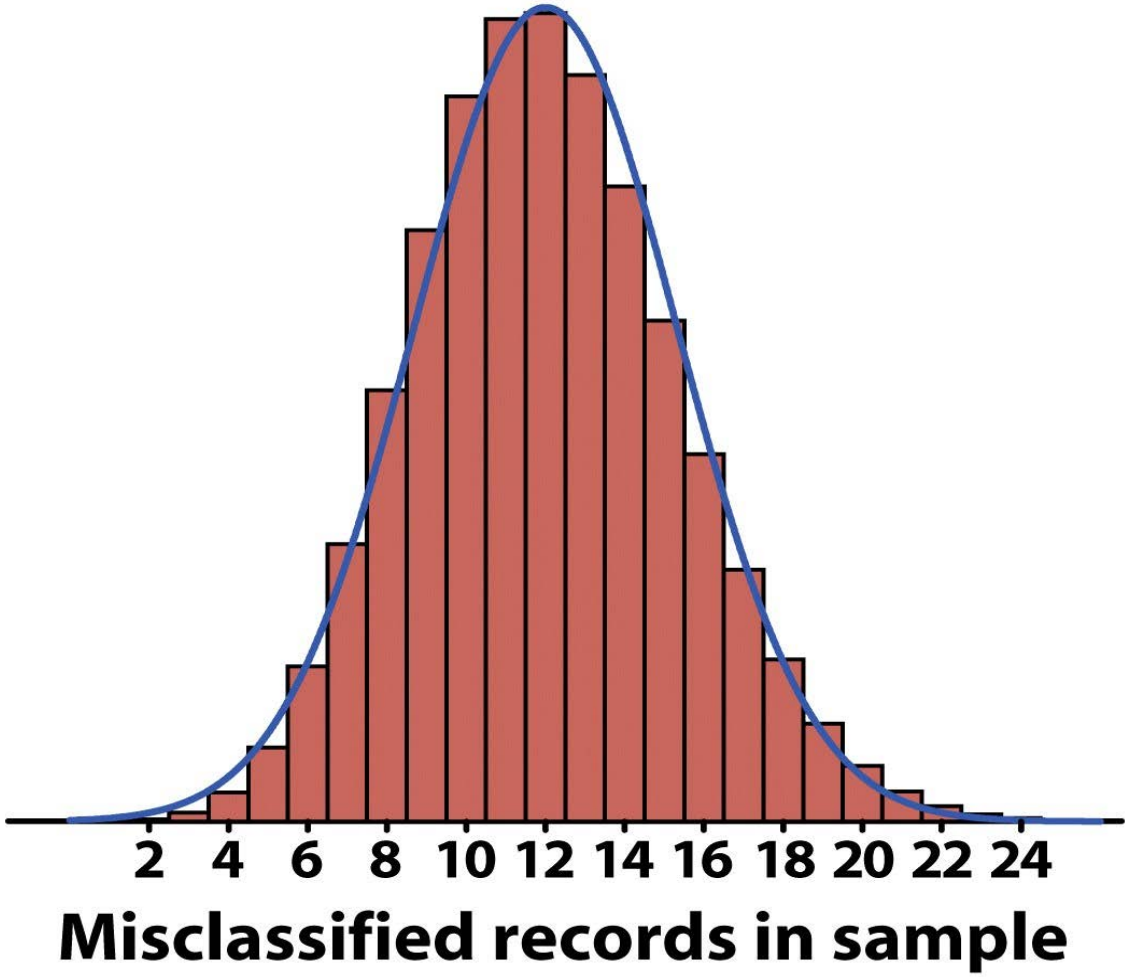


Figure 5-6
Introduction to the Practice of Statistics, Fifth Edition

Andeler og normalfordeling

- Tidligere:

$$\hat{p}\ tilnærmet\ N\left(p, \sqrt{p(1-p)/n}\right)$$

- X_i=1 hvis suksess, null ellers
- $\hat{p}$ =gjennomsnitt av x_i'ene
- Tilnærmet normalfordeling følger av sentralgrenseteoremet