Väntevärde, standardavvikelse och varians

Ett statistiskt material kan sammanfattas med medelvärde och standardavvikelse (varians), $\bar{x}$ och s (s^2).

På liknande sätt kan en sannolikhetsfördelning med kända förutsättningar sammanfattas med *väntevärde,* μ, och *standardavvikelse,* σ.

μ anger vilket medelvärde och

σ anger vilken standardavvikelse man kan förvänta sig att få om mäter många gånger.

Väntevärde, standardavvikelse och varians

Om ξ är en ***diskret*** stokastisk variabel med utfallsrummet $\{x_i, i = 1,...\}$.
Väntevärdet för ξ, $E[\xi]$, ofta betecknat μ, definieras då som

$$E[\xi] = \sum_i x_i P(\xi = x_i)$$

Variansen för ξ, ofta betecknad σ^2, definieras som

$$V[\xi] = E[(\xi - \mu)^2] = \sum_i (x_i - \mu)^2 P(\xi = x_i) = E(\xi^2) - \mu^2$$

Standardavvikelsen, ofta betecknad med σ, definieras som

$$\sqrt{V[\xi]} = D(\xi) = \sigma$$

Väntevärde, standardavvikelse och varians

Om ξ är en ***kontinuerlig*** stokastisk variabel med frekvensfunktionen *f(x)*. Väntevärdet för ξ, $E[\xi]$, ofta betecknad μ, definieras då som

$$\mu = E[\xi] = \int_{-\infty}^{\infty} x f(x) dx$$

Variansen för ξ, ofta betecknad σ^2 definieras som

$$\sigma^2 = V[\xi] = E[(\xi - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = E[\xi^2] - \mu^2$$

Standardavvikelsen, ofta betecknad med σ, definieras som

$$\sigma = \sqrt{V[\xi]} = D[\xi]$$

Median, kvartil och percentil

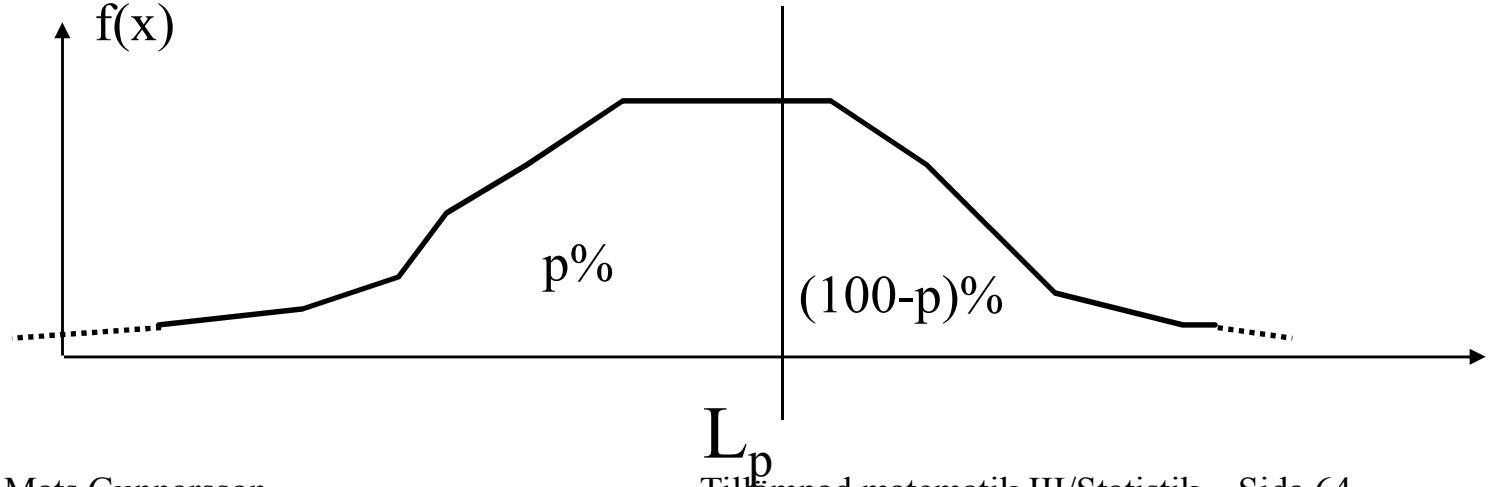
Den stokastiska variabeln ξ har fördelningsfunktionen $F(x)$.
Medianen definieras som det tal, m, som uppfyller

$$\boldsymbol{F(m) = 0,5}$$

Den stokastiska variabeln ξ har fördelningsfunktionen $F(x)$. Den p:te percentilen definieras som det tal L_p som uppfyller

$$\boldsymbol{F(L_p) = p\% = (p/100)}$$

Med kvartiler avses $Q_1 = L_{25}$, $Q_2 = L_{50}$ (medianen) och $Q_3 = L_{75}$.



Väntevärde, standardavvikelse m.m med *Mathematica*

ξ diskret stokastisk variabel med utfall $x_1, x_2, \ldots, x_n$ och given sannolikhetsfunktion $p(x_k)$.

Med *Mathematica* beräknas väntevärde och varians enligt.

x={x$_1$,x$_2$,.., x$_n$}

px={p(x$_1$),p(x$_2$),.., p(x$_n$)}

my=x.px (skalärprodukt)

varians=x^2.px-my^2

Väntevärde, standardavvikelse m.m med *Mathematica*

ξ kontinuerlig stokastisk variabel med utfall $a < x < b$ och given frekvensfunktion *f(x)*.

Med *Mathematica* beräknas väntevärde och varians direkt med definitionen

$$\mathbf{my} = \int_a^b \boldsymbol{x f(x)}\, \boldsymbol{dx}$$

$$\mathbf{varians} = \int_a^b \boldsymbol{x^2 f(x)}\, \boldsymbol{dx} - \mathbf{my}^2$$

Väntevärde, standardavvikelse m.m med *Mathematica*

För de ”kända” fördelningarna använder man

my=Mean[*fördelning*] resp.

varians=Variance[*fördelning*]

median=Median[*fördelning*]

kvartiler=Quartiles[*fördelning*]

ex.

```
Mean[BinomialDistribution[n,p]]
Variance[ExponentialDistribution[λ]]
Median[PoissonDistribution[λ]]
Quartiles[NormalDistribution[μ,σ]]
```

Några vanliga fördelningar

Fördelning	Slh funkt resp. täthet	Ω	Väntevärde	Varians
Binomial $Bin(n,p)$	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$	$k = 0, \ldots, n$	np	$np(1-p)$
Hypergeometrisk $Hyp(N,n,p)$	$p(k) = \frac{\binom{Np}{k}\binom{N-Np}{n-k}}{\binom{N}{n}}$	$k \leq Np, n-k \leq N(1-p)$	np	$\frac{N-n}{N-1} np(1-p)$
Poisson $Po(\lambda)$	$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$	$k = 0, 1 \ldots$	λ	λ
Geometrisk	$p(k) = p(1-p)^k$	$k = 0, 1 \ldots$	$(1-p)/p$	$(1-p)/p^2$
ffg	$p(k) = p(1-p)^{k-1}$	$k = 1, 2, \ldots$	$1/p$	$(1-p)/p^2$
Normal $N(m,\sigma)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$	$-\infty < x < \infty$	m	σ^2
Gamma $\Gamma(p,a)$	$f(x) = \frac{1}{a^p \Gamma(p)} x^{p-1} e^{-\frac{x}{a}}$	$x \geq 0$	ap	$a^2 p$
Exponential $Exp(\lambda)$	$f(x) = \lambda e^{-\lambda x}$	$x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Rektangel $R(a,b)$	$f(x) = \frac{1}{b-a}$	$a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$

Oberoende stokastiska variabler

Vi har *2* stokastiska variabler ξ_1,och ξ_2

Om $P(\xi_1<x_1 \text{ och } \xi_2<x_2) = P(\xi_1<x_1)P(\xi_2<x_2)$
för alla tal x_1 och x_2
så sägs ξ_1 och ξ_2 vara oberoende stokastiska variabler.

Jämför: Om $A = (\xi_1<x_1)$ och $B = (\xi_2<x_2)$,

A och B oberoende händelser gäller

$$P(\xi_1<x_1 \text{ och } \xi_2<x_2) = \mathbf{P(A\cap B) = P(A)P(B)} =$$

$$= P(\xi_1<x_1)P(\xi_2<x_2)$$

Oberoende stokastiska variabler

Vi har n stokastiska variabler $\xi_1, \xi_2, \ldots, \xi_n$

Om

$$P(\xi_1<x_1 \text{ och } \xi_2<x_2 \text{ och } \ldots \text{ och } \xi_n<x_n) =$$
$$= P(\xi_1<x_1)P(\xi_2<x_2) \ \ldots P(\xi_n<x_n)$$

för alla tal $x_1, x_2, \ldots x_n$

så är $\xi_1, \xi_2, \ldots, \xi_n$ oberoende stokastiska variabler

Sannolikheten för att $\xi_i<x_i$ påverkar inte sannolikheten för de övriga.

Räkneregler för väntevärde och varians för funktioner av stokastiska variabler

Sats 5A-C

- $E[a\xi + b] = aE[\xi] + b$
- $V[a\xi + b] = a^2 V[\xi]$

- $E[\xi_1 + \xi_2] = E[\xi_1] + E[\xi_2]$
- $V[\xi_1 + \xi_2] = V[\xi_1] + V[\xi_2]$, *om* ξ_1 *och* ξ_2 *är oberoende*

- $E[a_1\xi_1 + \ldots + a_n\xi n] = a_1 E[\xi_1] + \ldots + a_n E[\xi_n]$
- $V[a^2\xi_1 + \ldots + a_n\xi_1] = a_1^2 V[\xi_1] + \ldots + a_n^2 V[\xi_n]$,

om $\xi_1, \ldots, \xi_n$ *är oberoende*

Medelvärde av oberoende försök

Vi har n oberoende stokastiska variabler $\xi_1, \xi_2, \ldots, \xi_n$

Alla har samma väntevärde: $E[\xi_i] = \mu$

Alla har samma varians: $V[\xi_i] = \sigma^2$

Sätt $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$

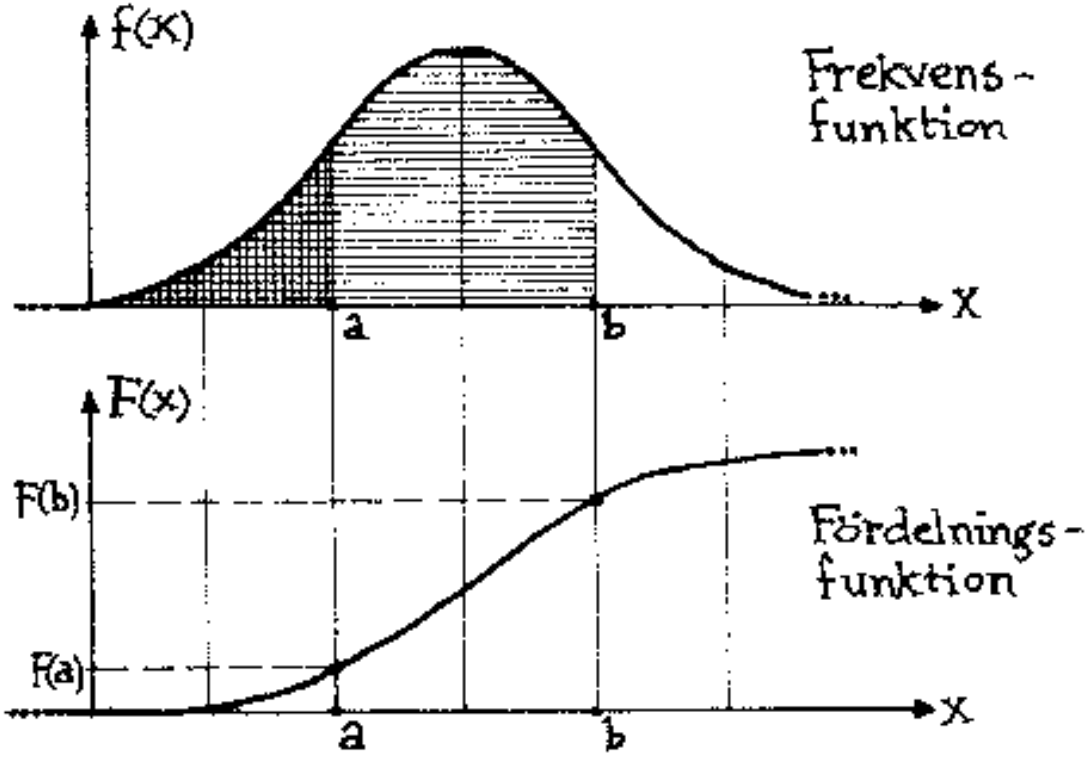
Då gäller $E[\bar{\xi}] = \mu \quad och \quad V[\bar{\xi}] = \sigma^2 / n$

Detta är tillämpligt vid till exempel upprepade mätningar på samma variabel

Normalfördelningen

Normalfördelningen är vanligt förekommande

– Den bestäms av två parametrar, väntevärde, μ, samt standardavvikelse, σ



$$\xi \in N(\mu,\sigma)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/(2\sigma^2)} dt$$

Normalfördelningen

För normalfördelningen är *F(x)* omöjlig att beräkna utan numeriska metoder (den går inte att lösa algebraiskt)

Därför finns tabeller för *N(0,1)*, vilken har fördelningsfunktionen

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

För denna finns tabeller

$$Om\ \xi \in N(\mu,\sigma)\ så\ gäller\ att\ P(\xi \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$\frac{\xi-\mu}{\sigma} \in N(0,1)$$

$$\Phi(-x) = 1 - \Phi(x)$$

Allmänna egenskaper

Sats

Om $\xi \in N(\mu,\sigma)$ och $Y = \dfrac{\xi - \mu}{\sigma}$

Då blir $Y \in N(0,1)$.

Sats

Om $\xi \in N(\mu,\sigma)$ då är $E(\xi)=\mu$ *och* $D(\xi) = \sigma$.

Dessutom gäller

$Y = a\xi + b \in N(a\mu + b;\ |a|\sigma)$

Allmänna egenskaper forts.

För alla normalfördelningar gäller:

$$P(\mu-\sigma<\xi<\mu+\sigma)=0.682$$
$$P(\mu-2\sigma<\xi<\mu+2\sigma)=0.954$$
$$P(\mu-3\sigma<\xi<\mu+3\sigma)=0.997$$

$$P(\mu-1.96\sigma<\xi<\mu+1.96\sigma)=0.95$$
$$P(\mu-2.58\sigma<\xi<\mu+2.58\sigma)=0.99$$
$$P(\mu-3.29\sigma<\xi<\mu+3.29\sigma)=0.999$$

Fler egenskaper

Sats **Om** $\xi_1 \in N(\mu_1;\sigma_1)$**,** $\xi_2 \in N(\mu_2;\sigma_2)$ **och oberoende gäller**

$$\xi_1 + \xi_2 \in N(\mu_1 + \mu_2; \sqrt{\sigma_1^2 + \sigma_2^2})$$

$$\xi_1 - \xi_2 \in N(\mu_1 - \mu_2; \sqrt{\sigma_1^2 + \sigma_2^2})$$

Sats Om $\xi_i \in N(\mu_i;\sigma_i)$ och oberoende samt $c_i \in \Re$ och är givna, $i = 1,\ldots,n$ gäller

$$\sum_{i=1}^{n} \xi_i \in N\left(\sum_{i=1}^{n} c_i\mu_i; \sqrt{\sum_{i=1}^{n} c_i^2\sigma_i^2}\right) \qquad \text{med } \mu_i = \mu \text{ fås}$$

$$\sum_{i=1}^{n} \xi_i \in N\left(n\mu; \sqrt{n}\sigma\right) \text{ och } \bar{\xi} \in N\left(\mu; \sigma/\sqrt{n}\right)$$

Centrala gränsvärdessatsen

Vi har n oberoende likafördelade stokastiska variabler

$\xi_1, \xi_2, \ldots, \xi_n$, med väntevärdet μ och standardavvikelsen σ

Om n går mot oändligheten gäller att

$$P\left(\frac{\sum_{i=1}^{n} \xi_i - n\mu}{\sigma\sqrt{n}} \leq x \right) \rightarrow \Phi(x)$$

Praktiskt: summan av antal slumpvariabler är approximativt normalfördelade om n är stort. (Tumregel $n \geq 30$)

Normalapproximationer är mycket användbara

Följder av centrala gränsvärdessatsen

Det gäller att $\overline{\xi}$ är approximativt $\mathrm{N}(\mu, \sigma/\sqrt{\mathrm{n}})$

och

$\sum_{i=1}^{n} \xi_i$ är approximativt normalfördelad $N(\mu n, \sigma\sqrt{n})$

Oavsett bakomliggande fördelning, bara *n* är tillräckligt stort, tum regel: $n > 30$

Följder av centrala gränsvärdessatsen

$$Om\, \xi \in Bin(n, p) \text{ så gäller } \xi \approx N\left(np, \sqrt{np(1-p)}\right)$$

om n är stort, tumregel : $V(\xi) = np(1-p) > 10.$

$$Om\, \xi \in Hyp(N, n, p) \text{ så gäller } \xi \approx N\left[np, \sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)}\right]$$

om n är stort, tumregel: $V(\xi) = np(1-p)\left(\frac{N-n}{N-1}\right) > 10.$

$$Om\, \xi \in Po(\lambda) \text{ så gäller } \xi \approx N\left(\lambda, \sqrt{\lambda}\right)$$

om n är stort tumregel : $V(\xi) = \lambda > 15.$

Approximationsregler - centrala gränsvärdessatsen