TAMS79: Föreläsning 5
Väntevärde och varians

Johan Thim (johan.thim@liu.se)

10 november 2018

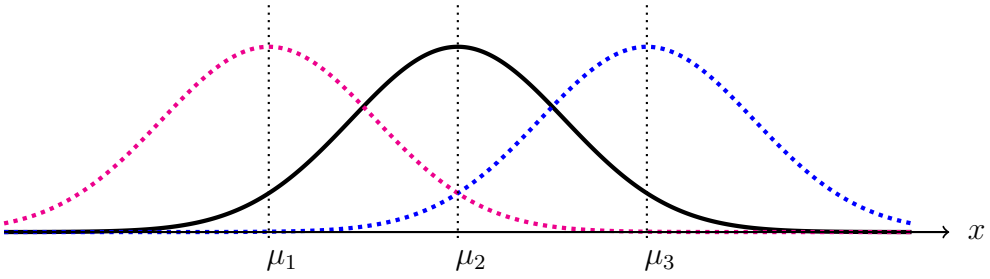
Väntevärde

Definition. Väntevärdet $E(X)$ av en stokastisk variabel X definieras som

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx \qquad \text{respektive} \qquad E(X) = \sum_k k p_X(k)$$

för kontinuerliga och diskreta variabler.

Andra vanliga beteckningar: μ eller μ_X. Väntevärdet är ett lägesmått som anger vart sannolikhetsmassan har sin tyngdpunkt (jämför med mekanikens beräkningar av tyngdpunkt). Om fördelningen är symmetrisk blir det i mitten, men är fördelningen skev blir det annorlunda. I figuren nedan ser vi tre täthetsfunktioner som har samma form men olika väntevärden. De är helt enkelt translationer av samma funktion i detta fall.



Exempel

Vid tillämpningar tolkas ofta väntevärdet som just det *förväntade värdet* för en stokastisk variabel.

(i) Om $X \sim \text{Exp}(\lambda)$ är livslängden för en lampa så är $E(X)$ den förväntade tiden en lampa vi plockar ur kartongen klarar.

(ii) Om $X \sim N(\mu, \sigma)$ är koncentrationen i en flaska salpetersyra så är $E(X)$ den koncentration vi förväntar oss när vid tar ned flaskan från hyllan.

(iii) Om X är antalet kast med en tärning innan vi får en 6:a för första gången så är $E(X)$ det förväntade antalet kast innan vi ser den första 6:an.

Observera att väntevärdet är ett reellt tal, så det kan mycket väl vara så att en variabel (då oftast en diskret sådan) inte *kan* anta sitt väntevärde.

Exempel

Kasta en 4-sidig tärning och låt X vara utfallet $1, 2, 3$ eller 4. Beräkna $E(X)$.

Lösning. Vi antar att tärningen är ärlig så $p_X(k) = 1/4$ för $k = 1, 2, 3, 4$. Då blir

$$E(X) = \sum_{k=1}^{4} k p_X(k) = \frac{1+2+3+4}{4} = \frac{5}{2}.$$

Det förväntade resultatet är alltså 2.5. Knappast ett resultat vi förväntar oss vid ett enskilt kast!

Vad är *egentligen* ett väntevärde?

Vi har gjort en definition av begreppet väntevärde ovan, så det är den som gäller. Men åtminstone följande tolkningar eller alternativa definitioner finns.

(i) Ett sannolikhetsviktat medelvärde av de värden X kan anta.

(ii) Integralen av X med avseende på sannolikhetsmåttet: $\int_\Omega X(\omega)\, dP(\omega)$.

(iii) Masscentrum för sannolikhetsfördelningen.

(iv) Det värde X hamnar på i snitt vid väldigt många upprepningar.

Vad punkt (ii) betyder kräver mer analys än vi har tillgång till. Hur hanterar vi då funktioner av stokastiska variabler på ett smidigt sätt? Följande sats ger svaret (ofta känd som *the law of the unconscious statistician*), men resultatet behöver egentligen lite diskussion.

Väntevärde och funktioner av stokastiska variabler

Sats. Låt $Y = g(X)$ och $W = h(U, V)$. I de kontinuerliga fallen blir

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \qquad \text{och} \qquad E(W) = \iint_{\mathbf{R}^2} h(x,y) f_{U,V}(x,y)\, dxdy,$$

och om X, U, V är diskreta:

$$E(Y) = \sum_k g(k) p_X(k) \qquad \text{och} \qquad E(W) = \sum_j \sum_k h(j,k) p_{U,V}(j,k).$$

Det kontinuerliga fallet (med täthetsfunktion) kommer vi inte åt på något annat sätt än att egentligen ta satsen ovan som definition. Om vi tänker på punkt (ii) i rutan föregående satsen, så förefaller det ganska rimligt. Sammansättningen $g(Y)$ till exempel är en ny stokastisk variabel och då skulle

$$E(g(Y)) = \int_\Omega g(Y)(\omega)\, dP(\omega),$$

vilket är precis hur satsen tolkas. I det diskreta fallet kan vi faktiskt producera ett bevis:

$$\begin{aligned} E(g(X)) &= \sum_k k P(g(X) = k) = \sum_k k \sum_{m\,:\,g(m)=k} P(X = m) \\ &= \sum_m \sum_{k\,:\,g(m)=k} k P(X = m) = \sum_m P(X = m) \sum_{k\,:\,g(m)=k} k = \sum_m g(m) P(X = m), \end{aligned}$$

om vi antar att serien är absolutkonvergent i det oändliga fallet så vi kan byta summationsordningen. Rent praktiskt kan beräkningarna gå till på följande sätt.

Exempel

Låt $f_{X,Y}(x,y) = 2$ om $0 < y < x < 1$ och $f_{X,Y}(x,y) = 0$ för övrigt. Bestäm väntevärdet $E(XY + Y^2X)$.

Lösning: Vi använder satsen ovan:

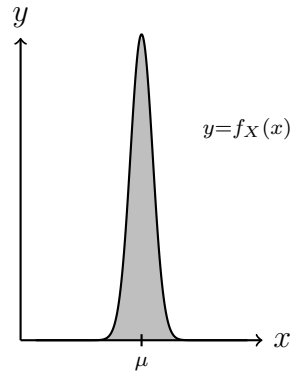
$$\begin{aligned} E(XY+Y^2X) &= \iint_{\mathbf{R}^2} (xy+y^2x) f_{X,Y}(x,y)\,dxdy = 2\int_0^1 \int_0^x (xy+y^2x)\,dydx \\ &= 2\int_0^1 \left[\frac{xy^2}{2} + \frac{xy^3}{3}\right]_0^x dx = \int_0^1 \left(x^3 + \frac{2x^4}{3}\right) dx \\ &= \frac{1}{4} + \frac{2}{15} = \frac{23}{60}. \end{aligned}$$

5.1 Varians och standardavvikelse

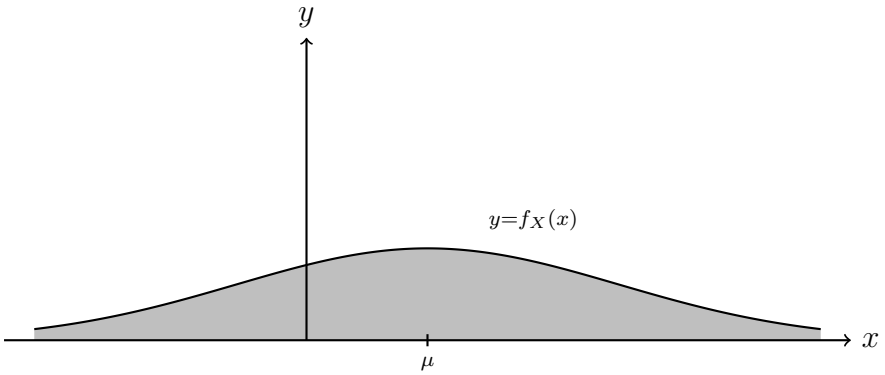
Varians och standardavvikelse

Definition. Låt X vara en stokastisk variabel med $|E(X)| < \infty$. **Variansen** $V(X)$ definieras som $V(X) = E((X - E(X))^2)$. **Standardavvikelsen** $D(X)$ definieras som $D(X) = \sqrt{V(X)}$.

Andra vanliga beteckningar för standardavvikelsen: σ, σ_X, $\sigma(X)$.
Variansen är ett spridningsmått. Stor varians (eller standardavvikelse) betyder att sannolikhetsfördelning har stor spridning. Många värden är troliga. Liten varians betyder att fördelningen är centrerad, hög sannolikhet att hamna kring en viss punkt; se figuren nedan.



Litet $D(X) = \sigma$.



Stort $D(X) = \sigma$.

Steiners sats

Sats. $V(X) = E(X^2) - E(X)^2$.

Exempel

Låt X_1 anta värdena $\{-1, 1\}$ med $p_{X_1}(-1) = p_{X_1}(1) = 1/2$ och låt X_2 anta värdena $\{-10, 10\}$ med $p_{X_2}(-10) = p_{X_2}(10) = 1/2$. Beräkna $V(X_1)$ och $V(X_2)$.

Lösning: Det är klart att $E(X_1) = E(X_2) = 0$ (varför?), så

$$V(X_1) = E(X_1^2) - E(X_1)^2 = \frac{1}{2}(-1)^2 + \frac{1}{2}1^2 - 0 = 1$$

och

$$V(X_2) = E(X_2^2) - E(X_2)^2 = \frac{1}{2}(-10)^2 + \frac{1}{2}(10)^2 - 0 = 50 + 50 = 100.$$

Tydligt att X_2 har mycket större varians även om fördelningarna kan tyckas se snarlika ut, men sannolikheten är mycket mer utspridd för X_2.

Exempel

Låt X anta värdena $0, 1, 2, \ldots$ med sannolikheterna $p_X(k) = 2^{-k-1}$. Beräkna $E(X)$ och $V(X)$.

Lösning: Till att börja med kan vi kontrollera att p_X verkligen är en sannolikhetsfunktion. Klart att $p_X(k) \geq 0$, och summan nedan är geometrisk så

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} 2^{-k-1} = 2^{-1} \sum_{k=0}^{\infty} 2^{-k} = 2^{-1} \cdot \frac{1}{1 - 1/2} = 1.$$

Vi beräknar väntevärdet. Låt $q \in]0, 1[$ så $p_X(k) = (1-q)q^k$ med $q = 1/2$. Vi kan beräkna summan av kq^k genom följande manöver:

$$\sum_{k=0}^{\infty} kq^k = \sum_{k=1}^{\infty} kq^k = q \sum_{k=1}^{\infty} kq^{k-1} = q \sum_{k=0}^{\infty} \frac{d}{dq} q^k = q \frac{d}{dq} \sum_{k=0}^{\infty} q^k = q \frac{d}{dq} \frac{1}{1-q} = \frac{q}{(1-q)^2}.$$

Alltså blir

$$E(X) = (1-q) \sum_{k=0}^{\infty} kq^k = \frac{q}{1-q} \qquad \text{och} \qquad E(X) = 1 \text{ om } q = \frac{1}{2}.$$

För att beräkna $E(X^2)$ kikar vi på motsvarande kalkyl för andraderivatan:

$$q^2 \sum_{k=0}^{\infty} \frac{d^2}{dq^2} q^k = q^2 \sum_{k=0}^{\infty} (k^2 - k) q^{k-2} = \sum_{k=0}^{\infty} (k^2 - k) q^k = \sum_{k=0}^{\infty} k^2 q^k - \frac{q}{(1-q)^2}.$$

Således,

$$\sum_{k=0}^{\infty} k^2 q^k = \frac{q}{(1-q)^2} + q^2 \frac{d^2}{dq^2} \sum_{k=0}^{\infty} q^k = \frac{q}{(1-q)^2} + \frac{2q^2}{(1-q)^3},$$

vilket medför att

$$E(X^2) = (1-q) \sum_{k=0}^{\infty} k^2 q^k = \frac{q}{1-q} + \frac{2q^2}{(1-q)^2}$$

så

$$V(X) = E(X^2) - E(X)^2 = \frac{q}{1-q} + \frac{q^2}{(1-q)^2} = \frac{q}{(1-q)^2}.$$

och med $q = 1/2$ får vi $V(X) = 2$. Den observante läsaren kanske känner igen sannolikhetsfunktionen vi arbetar med då det är den geometriska fördelningen $X \sim \text{Geo}(1-q)$. Så vad vi visat ovan är följande:

Geometrisk fördelning

Sats. Om $X \sim \text{Geo}(p)$ så är $E(X) = \dfrac{1-p}{p}$ och $V(X) = \dfrac{1-p}{p^2}$.

Ett annat lägesmått än väntevärdet är medianen.

Median

Definition. En **median** för en stokastisk variabel X är ett tal $m \in \mathbf{R}$ så att

$$P(X \leq m) = P(X \geq m) = \frac{1}{2}.$$

Observera att medianen inte behöver vara entydig!

Medianen för en Exponentialfördelning

Låt $X \sim \text{Exp}(\lambda)$. Beräkna medianen och väntevärdet för X.

Lösning: Vi räknar ut fördelningsfunktionen för X. Om $x > 0$,

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \int_0^x \lambda e^{-\lambda t}\,dt = 1 - e^{-\lambda x}.$$

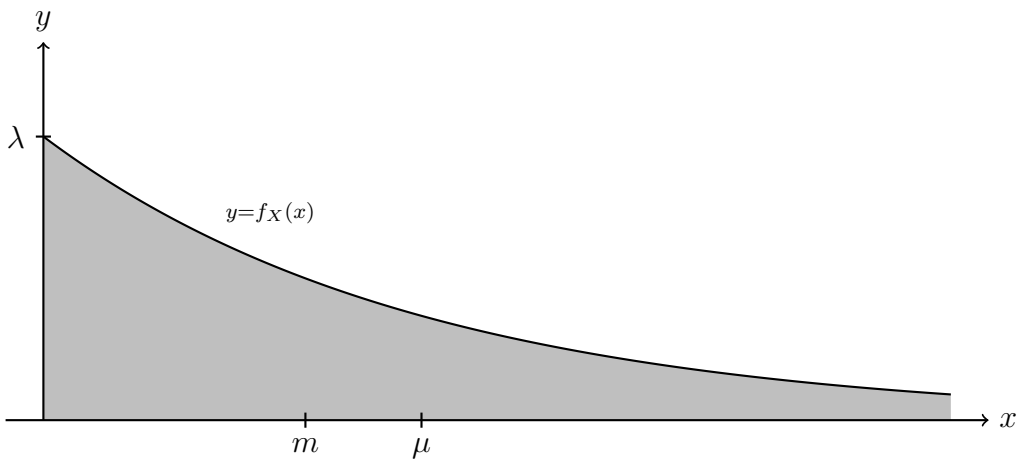
Medianen finner vi ur ekvationen $F_X(m) = 1/2$, dvs

$$1 - e^{-\lambda m} = \frac{1}{2} \quad \Leftrightarrow \quad e^{-\lambda m} = \frac{1}{2} \quad \Leftrightarrow \quad m = \frac{\ln 2}{\lambda}.$$

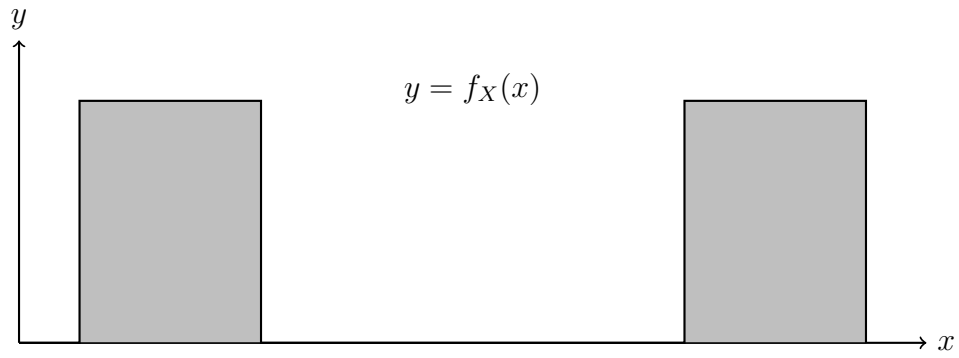
Jämför detta med väntevärdet för X:

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x}\,dx = \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx = 0 - 0 + \left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^\infty = \frac{1}{\lambda}.$$

Medianen och väntevärdet behöver alltså *inte* vara samma sak!



Ett annat exempel, där medianen *inte* är entydigt definierad:



Vart är medianen??

5.2 Räknelagar

För väntevärdet gäller bland annat följande regler.

Linjäritet och oberoende produkt

Låt X och Y vara stokastiska variabler. Då gäller

(i) $E(aX + b) = aE(X) + b$ för alla $a, b \in \mathbf{R}$;

(ii) $E(aX + bY) = aE(X) + bE(Y)$ för alla $a, b \in \mathbf{R}$;

(iii) Om X och Y är oberoende gäller $E(XY) = E(X)E(Y)$.

För variansen kan vi visa följande:

Låt X och Y vara stokastiska variabler. Då gäller

(i) $V(aX + b) = a^2V(X)$ för alla $a, b \in \mathbf{R}$;

(ii) $V(aX \pm bY) = a^2V(X) + b^2V(Y) + 2ab(E(XY) - E(X)E(Y))$ för alla $a, b \in \mathbf{R}$;

(iii) Om X och Y är oberoende gäller $V(aX \pm bY) = a^2V(X) + b^2V(Y)$.

Varianser adderas alltid!

Observera att det *alltid* blir ett plustecken mellan varianserna för linjär-kombinationer:

$$V(aX \pm bY) = a^2V(X) + b^2V(Y).$$

Vi kommer *aldrig* att bilda skillnader mellan varianser!

Det följer att standardavvikelsen för en linjärkombination $aX+bY$ av två oberoende stokastiska variabler ges av $\sigma_{aX+bY} = \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2}$.

Exempel

Låt X och Y vara oberoende med $E(X) = 2$, $E(X^2) = 8$, $E(Y) = -1$, $V(Y) = 2$. Beräkna $E(XY)$, $E(X^2Y^2)$ och $D(2X - 3Y)$.

Då variablerna är oberoende blir $E(XY) = E(X)E(Y) = 2 \cdot (-1) = -2$, och

$$E(X^2Y^2) = E(X^2)E(Y^2) = 8E(Y^2) = 8(V(Y) + E(Y)^2) = 8 \cdot 3 = 24.$$

Vidare erhåller vi

$$V(2X - 3Y) = 2^2V(X) + (-3)^2V(Y) = 4(8 - 2^2) + 9 \cdot 2 = 34,$$

så $D(2X - 3Y) = \sqrt{34}$.

Exempel

Låt X och Y vara oberoende likafördelade variabler med $p_X(k) = p_Y(k) = 1/4$ om $k = 1, 2$ och $p_X(k) = p_Y(k) = 1/2$ om $k = 0$. Bestäm sannolikhetsfuntionen för $Z = X + Y$, väntevärdet $E(X + Y)$ samt variansen $V(X + Y)$.

Vi börjar med att bestämma $p_Z(k)$. Man kan göra detta på samma sätt som i exemplet från föregående föreläsning (med summor istället för integraler), men då vi har så få möjliga värden på variablerna är det enklare att ställa upp en tabell.

$Z = X + Y = n$	Par (j, k) så $j + k = n$	Total sannolikhet
-1	–	0
0	$(0, 0)$	$\frac{1}{2}\frac{1}{2} = \frac{1}{4}$
1	$(0, 1)$, $(1, 0)$	$2 \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{4}$
2	$(0, 2)$, $(1, 1)$, $(2, 0)$	$2 \cdot \frac{1}{2}\frac{1}{4} + \frac{1}{4}\frac{1}{4} = \frac{5}{16}$
3	$(1, 2)$, $(2, 1)$	$2 \cdot \frac{1}{4}\frac{1}{4} = \frac{1}{8}$
4	$(2, 2)$	$\frac{1}{4}\frac{1}{4} = \frac{1}{16}$
5	–	0

Två sätt att beräkna väntevärdet:

$$E(Z) = E(X + Y) = E(X) + E(Y) = 2E(X) = 2\sum_{k=0}^{2} kp_X(k) = 2\left(0 + \frac{1}{4} + \frac{2}{4}\right) = \frac{3}{2},$$

$$E(Z) = \sum_{k=0}^{4} kp_Z(k) = \frac{1}{4} + \frac{10}{16} + \frac{3}{8} + \frac{4}{16} = \frac{24}{16} = \frac{3}{2}.$$

Möjligen kan man tycka att det andra alternativet ser enklare ut, men då har man glömt bort allt arbete som gick åt till att skapa tabellen ovan. Det första alternativet är nästan alltid det enklaste. Variansen beräknar man enklast enligt följande

$$V(Z) = V(X + Y) = \big[\text{oberoende variabler}\big] = V(X) + V(Y) = 2V(X) = \frac{1}{8}$$

eftersom vi vet att

$$V(X) = E(X^2) - E(X)^2 = \sum_{k=0}^{2} k^2 p_X(k) - \frac{9}{16} = \left(0 + \frac{1}{4} + \frac{4}{4}\right) - \frac{9}{16} = \frac{1}{16}.$$

5.3 Medelvärdet

För att hitta en bra gissning (*skattning*) på väntevärdet brukar man använda medelvärdet av de observationer man har tillgång till.

Medelvärde

Låt $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ vara medelvärdet av n stycken oberoende likafördelade stokastiska variabler sådana att $E(X_i) = \mu$ och $V(X_i) = \sigma^2$ för alla i. Vad är $E(\overline{X})$? Vad är $\lim_{n\to\infty} V(\overline{X})$?

Lösning: Eftersom väntevärdesoperatorn är linjär så gäller att

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu.$$

Detta kallas att $\overline{X}$ är en *väntevärdesriktig skattning* av μ; ni kommer att stöta på detta fler gånger i senare kurser.
Då variablerna är oberoende kan vi göra en liknande kalkyl för variansen:

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Här ser vi att $V(\overline{X}) \to 0$ då $n \to \infty$. Även detta är ett exempel på ett fenomen som kommer visa sig viktigt i statistisk inferens (nämligen *konsistens*). Vi kommer även stöta på det här igen i föreläsning 8, när vi visar att $\overline{X}$ *konvergerar* mot väntevärdet i sannolikhet (vad nu det innebär).

5.4 Bökigt exempel

Två sätt att räkna ut $E(g(X))$

Låt $\Theta \sim \mathrm{Re}(0, \pi/4)$ vara likformigt fördelad och definiera $Y = \cos\Theta$. Vad blir $E(Y)$?

Lösning: Det enklaste sättet är att använda satsen från början av föreläsningen:

$$E(Y) = E(\cos\Theta) = \int_{-\infty}^{\infty} \cos\theta f_\Theta(\theta)\, d\theta = \frac{4}{\pi}\int_0^{\pi/4} \cos\theta\, d\theta = \frac{4}{\pi}\left(\frac{\sqrt{2}}{2} - 0\right) = \frac{2\sqrt{2}}{\pi}.$$

Den andra varianten börjar med beräkning av täthetsfunktionen för $Y = \cos\Theta$. Vi ställer upp fördelningsfunktionen först:

$$F_Y(y) = P(Y \le y) = P(\cos\Theta \le y) = P(\Theta \ge \arccos y) = 1 - P(\Theta < \arccos y) = 1 - F_\Theta(\arccos y).$$

Här har vi utnyttjat att $\cos\theta$ är avtagande för $\theta \in [0, \pi/4]$. Vi kan nu derivera fram $f_Y(y)$:

$$f_Y(y) = F_Y'(y) = -F_\Theta'(\arccos y)\cdot\frac{-1}{\sqrt{1-y^2}} = \frac{f_\Theta(\arccos y)}{\sqrt{1-y^2}}$$
$$= \begin{cases} \frac{4}{\pi}\frac{1}{\sqrt{1-y^2}}, & 0 \le \arccos y \le \frac{\pi}{4} \quad \Leftrightarrow \quad \frac{\sqrt{2}}{2} \le y \le 1 \\ 0, & \text{för övrigt.} \end{cases}$$

Vi kan nu beräkna $E(Y)$ enligt definitionen:

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)\, dy = \frac{4}{\pi} \int_0^{\sqrt{2}/2} \frac{y}{\sqrt{1-y^2}}\, dy = \frac{4}{\pi} \left[-\sqrt{1-y^2} \right]_0^{\frac{\sqrt{2}}{2}} = \frac{2\sqrt{2}}{\pi}.$$

Vilket metod tycker du är enklast?