Att testa mindre – och veta mer,
med tillämpade statistiska metoder

Fredrik Carlemalm

Inledning

- **Tester tar traditionellt mycket tid**
  - ***I utvecklingsprojekt - och än mer i förvaltning av IT-system - är det vanligen ont om kalendertid, pengar och/eller resurser.***
  - ***Detta har blivit än värre i dagens "slimmade" organisationer, där det ofta är svårt att få testare från en stressad verksamhet. Och tester tar i sin traditionella form mycket tid.***
- **Testarbetet går fortare med statistiska metoder**
  - ***Trots den allmänna uppfattningen att man inte kan snabba på testarbetet finns det alternativ till fullskaliga, traditionella tester.***
  - ***Ett sätt är med anpassad stickprovsteknik, s.k. statistisk testning, som innebär att vi vet tillräckligt väl, att det är tillräckligt rätt.***
- **Testresultaten blir objektiva och kvantifierbara**
  - ***Än bättre: vi får ett objektivt, mätbart kvalitetsmått i testerna! Vi kan alltså testa mindre – och veta betydligt mer.***
  - ***Vi kan lättare – och snabbare – svara på om ett system tycks fungera.***

Vad innebär stickprovsmetodik?

- **Ett slumpvis urval från en population (ett antal enheter)**
  - ***Enheterna bör vara jämförbara vad gäller storlek, typ, etc. (t.ex. krav eller testfall med en viss prioritet, kodändringar, osv.)***
- **Ett antal enheter i stickprovet som matchar ambitionsnivån**
  - ***Vill vi uttala oss med 99 % konfidens, räcker det ofta med ca 50 st i urvalet***
    Vill vi däremot uppnå 99,9 % konfidens, behöver vi betydligt fler (men inte 500)
  - ***Populationens storlek påverkar inte stickprovets storlek (i nämnvärd omfattning)!!***
- **Resultatet innebär att hela populationen accepteras eller förkastas!**
  - ***Testerna är alltså med stickprovsteknik bara kontrollerande, aldrig ”avlusande”.***

När kan stickprov användas?

För att man ska kunna använda och ha nytta av stickprov förutsätts att:

1. **Relativt stora mängder av något ska undersökas**
2. **Det som ska undersökas är av någorlunda likartad natur**
3. **Om en undersökt enhet är felaktig ska gå att avgöra**
4. **Undersökningen ska vara av kontrollerande typ (och inte rättande)**
5. **Önskad tillförlitlighet ska vara rimlig, inte total säkerhet!**

Det betyder att stickprov kan vara till stor nytta särskilt på testområdet, där det ofta finns stora mängder testfall att köra och verifiera mot facit.

- **Stickprovstekniken är dock inte begränsad till urval bland testfall**
  - ***Urval bland krav som ska verifieras eller bland utförda teststeg som ska verifieras mot testfacit är andra typiska områden där stickprovstekniken kan komma till nytta.***
- **Stickprov används i praktiken ofta redan idag, om än inte så ordnat; i system- och acceptanstest går det inte att testa alla varianter av allt**
  - ***Med hjälp av exempelvis testmatriser säkerställer man då att man testar på bra sätt. Men med ordnade stickprov får vi en kvantifierbarhet.***

Typisk testtillämpning

Huvudfrågeställningar inom test är ofta:

- Hur många testfall behövs för att säkra systemet?
- Nu när vi kört igenom x fall utan större fel, räcker det?

Två huvudsakliga varianter med hjälp av stickprov:

- Hur stort stickprov behövs för att statistiskt säkerställa en viss nivå?
- Givet ett visst stickprov, vilken kvalitetsnivå har statistiskt sett nåtts?

Faktorer som styr vad och hur mycket som ska testas är:

- **Vilken kvalitetsnivå vi vill uppnå**
  - *Kan variera mellan olika delar av ett systemkomplex*
- **Hur säkra vi vill vara på att vi uppnått önskad nivå**
  - *Kan också variera mellan olika delar av ett systemkomplex*
- **Hur mycket vi kan, orkar och hinner planera i förväg**
  - *Vi kan också använda metodiken för att i efterhand räkna ut hur säkra vi är...*

Teori: Skattningsmetoder (1)

Ur ett bestånd om N likadana enheter – t.ex. samma slags programkodsändringar – tas ett stickprov som omfattar n enheter, där x av enheterna visar sig vara felaktiga. Hur stor är totalt andelen felaktiga? Dvs hur stor lär den okända felkvoten p vara?

❖ Punktskattning

- *För tillräckligt stora stickprov kan man givetvis approximera p med x / n (= p*). Men hur bra är det? Vågar vi lita på den skattningen?*

❖ Intervallskattning

- *Ett statistiskt sätt att kunna uttala sig om hur säkra vi är på vår approximation, fortfarande $p \approx x / n$, är att inrymma p i ett konfidensintervall:*

  $x / n - d_1 < p < x / n + d_2$, som med viss säkerhet täcker p.

  Beroende på hur noga det är, väljer vi d_1 och d_2 så att vi täckt in p med t.ex. 95 eller 99 % säkerhet. Ju större säkerhet som krävs, desto större intervall behöver vi ta till.

  För att vi ska kunna välja riktiga värden på d_1 och d_2 måste vi emellertid också veta något om hur sannolikheterna är fördelade i intervallet. Kommer det att vara lika stor chans att hitta x felaktiga enheter som att finna $x + nd_2$ eller noll stycken?

Teori: Skattningsmetoder (2)

- **Hypotesprövning**
  - *Ett mer avancerat sätt att tänka är med hjälp av s.k. hypotesprövning.*

    Sätt upp hypotesen att $p <= p_0$, där p_0 är ett bestämt tal, 0 – 100 %, t.ex. 1 %.
    Vi vill pröva hypotesen genom att använda stickprovet, så att vi accepterar hypotesen om $x <= x_0$, där x_0 är ännu ett bestämt tal, 0 – 100 %, t.ex. 0,5 %.

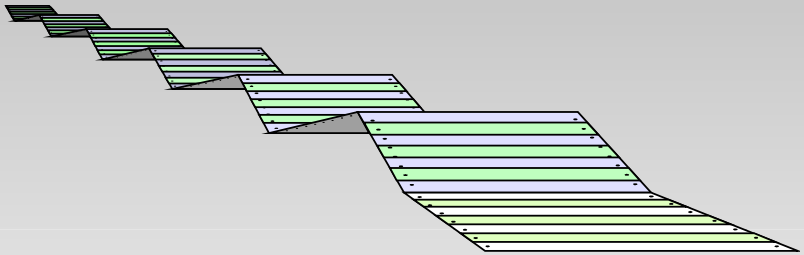
    (Det är ej nödvändigt att p_0 och x_0 är lika, även om de i det enklaste fallet är det.)

    Detta förfarande kallas signifikanstest och är i vissa fall ett utmärkt alternativ till intervallskattning. För våra behov kommer dock intervallskattning att räcka bra.
  - *Denna text utgör teoretisk bakgrund och gås därför ej igenom i detalj.*

- **Vilka matematiska formler som gäller och hur de kan approximeras beskrivs i appendix längst bak i bildmaterialet…**

Praktik: Standard-användning!

- **Standard i stället för beräkning**
  - *Med en lämplig standard för stickprovstagning kan vi undvika egna kalkyler:*
    - *ISO 2859*
    - *ANSI / ASQ Z1.4*
    - *BS6001*
    - *DIN40.080*
    - *NFX06-022*
    - *UN148-42*
    - *KS A 3109*
    - *MIL-STD-105E (”pappa” och motsvarighet till alla de andra - men gratis!!)*



  - *MIL-STD-105E har i flera decennier använts av USA:s militär, men också civilt, för att kontrollera allt ifrån gevärspatroner till datamängder.*
  - *Länk till automatkalkyl: http://www.sqconline.com/mil-std-105.html*
- **Läsanvisning / -tips:**
  - *I synnerhet avsnitt 4.9.1 i standard MIL-STD-105E är av central praktisk betydelse. Avsnitten 3.1 - 3.3, 4.3 - 4.8, 4.10 och 4.12 bör också studeras.*

Praktik: Tillvägagångssätt

1. **Bestäm storleken på beståndet och på dess eventuella delar**
   - *T.ex. antalet krav, funktioner, ändringar eller testfall*
2. **Bestäm acceptabel kvalitetsnivå, AQL.**
   - *AQL = 1 % motsvarar 99 % konfidens att beståndet är minst 99 % OK.*
3. **Bestäm inspektionsnivå, i normalfall II (enligt standarden).**
   - *S.k. specialinspektioner kan snabba upp förfarandet. De minskar mängden av stickprov, men ökar även risken.*
   - *Om specialinspektion genomförs, stäm gärna av mot AOQL, acceptabel slutlig ('utgående') kvalitetsnivå, enligt tabell.*
4. **Bestäm typ av stickprov (single, double eller multiple).**
   - *Med annat än 'single' får vi en bild redan vid första omgången.*
5. **Läs av värden i automatkalkyl eller tabell för aktuellt stickprov.**
   - *Beroende på upplägg och resultat, ändra vid behov stickprovstyp.*

Praktik: Typfall

- *Antag att vi har ett systemkomplex med ca 20 enheter, i vilka ca 500 kodändringar vardera gjorts. Önskemålet är att minst 99 % av kodändringarna ska vara korrekta.*
- *Fyll i aktuella värden på SQC Online:* ***http://www.sqconline.com/mil-std-105.html***

Enter your process parameters:		
Batch size (N):	3201 to 10000	The number of items in the batch. more info...
AQL:	1.0%	The Acceptable Quality Level. more info...
Inspection Level:	II	Determines the discrimination power of the plan. more info...
Type of inspection:	Normal	Depends on the quality history. more info...
	Submit	

- ***Resultatet blir då:***

The *Single* sampling procedure is:
Sample **200*** items. If the number of non-conforming items is **5 or less** --> accept the lot. **6 or more** --> reject the lot.

Med specialinspektion (nivå S4) räcker 32 st! Då får inga fel finnas. Se vidare Appendix…

Frågeställningar (1)

- **Beror inte stickprovet på systemstorleken?**
  - *Nej, inte nämnvärt. Med största tänkbara system räcker 50 testpunkter på nivå S4.*
    Vill vi ha en högre säkerhetsnivå och kanske använda dubbla inspektioner,
    kan vi välja en nivå kring 100 testpunkter – vilket kanske blir 20 testfall.
- **Behöver vi bara ta fram 20 testfall!?**
  - *Det väsentliga är antalet testpunkter, inte hur dessa är ordnade i testfall.*
    Men det förutsätter också
    a) att vi testar systemet som en helhet.
    b) att vi inte finner några fel i testerna.

    Vi bör därför ta fram fler testfall och mer testdata än vad vi verkligen använder.
    Men vi kan också använda stickprovsteknik när det gäller avstämning mot testfacit.
- **Kan vi dela in systemet i olika delar?**
  - *Ja, det lär t.o.m. vara ett vanligt sätt att arbeta, bl.a. utifrån varierande kvalitetskrav.*
    Vi kan då tidigt stickprova ett delsystem som omfattar några procent av helheten
    och anpassa det fortsatta arbetet efter resultatet.

Frågeställningar (2)

- **Vad händer om vi hittar fel?**
  - *Om vi hittar ett enda fel med enklaste formen av inspektion förkastar vi systemet, men använder vi en lite mer avancerad form av inspektion testar vi först djupare.*
    Om systemet inte håller måttet, måste det givetvis rättas och sedan testas om.
- **Hur mycket tjänar vi?**
  - *I extremfallet, med ett stickprov på 50 st bland ca 500.000 ändringar, helt utan fel: omkring en faktor 10.000 - dvs testerna blir fullkomligt försumbara.*
  - *I normalfallet, med flera delsystem, vissa djupare tester, etc. - uppemot faktor tio.*
- **Hur säkert är detta?**
  - *Både teorierna bakom och existerande standarder, bl.a. MIL-STD-105E, är stabila.*
- **Varför testar vi inte alltid så här?**
  - *Tester används för ”avlusning” också så sent som i systemtest, inte bara för kontroll.*
  - *Tekniken är inte känd inom mjukvaruutveckling (som inte är en så mogen bransch).*
  - *Det verkar krångligt och osäkert om man inte känner den matematiska bakgrunden.*

Frågeställningar (3)

Vilka typiska nackdelar eller faror finns?

- *Kravställningen måste vara tydlig - vilka delar kräver vilken kvalitetsnivå?*
- *Ambitionsnivån kan sättas för högt eller lågt, av okunskap eller girighet. Särskilt alltför glesa stickprov gör lätt att vi får en felaktig uppfattning...*
- *Systemet består av många olika slags, icke jämförbara enheter.*
- *Vi kan ha otur – eller förstått fel – och ta ett missvisande stickprov. T.ex. stickprov bland testfall, som totalt ej täcker systemfunktionaliteten...*
- *Testerna kan inte användas för att avlusa ett dåligt konstruktionsarbete. I många organisationer tror man fortfarande att testerna är till för detta.*
- *Vi kan ha förberett för få testfall för fördjupade tester, om kvaliteten är låg.*
- *Metodiken accepteras inte av testarna (eller andra viktiga intressenter i test).*
- ***Otillräckliga kunskaper i statistik gör att stickprovstekniken tillämpas fel.***

Appendix

Teori och exempel

Appendix: Statistisk metodik (1)

- **Dragning utan återläggning**
  - ***Att ta stickprov bland ett antal (likartade) kodändringar för att se om något av stickproven är felaktigt motsvarar ett av statistikernas favoritexempel; att blint plocka ut olikfärgade kulor ur en urna, utan återläggning, för att efteråt se vilka kulor som har en viss av färgerna.***

    Talar vi om två färger, t.ex. svart och vit, kan felaktiga kodändringar sägas motsvara svarta kulor.

- **Hypergeometrisk fördelning**
  - ***Det vi då har är en hypergeometrisk fördelning, med en sannolikhet $p_X(k)$ enligt följande formel:***

$$p_X(k) = \binom{Np}{k}\binom{Nq}{n-k} / \binom{N}{n} \; ; \; 0 <= k <= Np, \; 0 <= n-k <= Nq$$

$$\binom{Np}{k} = \frac{(Np)!}{k!(Np-k)!}, \; etc.$$

Appendix: Statistisk metodik (2)

Binomialapproximation

- *För stora N - och förutsatt att kvoten n / N < ca 10% - kan ovanstående formel approximeras med en binomialfördelning; $p_X(k)$ = Bin (n, p):*

$$p_X(k) \approx \binom{n}{k} p^k q^{n-k} \; ; \; 0 <= k <= n$$

Normalapproximation

- *För stora n och N kan på motsvarande sätt en normalapproximation (dvs den så nyttiga normalfördelningskurvan) användas:*

$$P(a < X <= b) \approx \Phi\left(\frac{b-np}{\sqrt{npq}}\right) - \Phi\left(\frac{a-np}{\sqrt{npq}}\right) \; ; \; X \in Norm(np, \sqrt{npq})$$

- *Detta förutsätter att standardavvikelsen är tillräckligt stor, vilket i praktiken ställer krav på att npq > ca 10*
- *Vanligen förfinar man ytterligare approximationen med s.k. halvkorrektion*

Appendix: Statistisk metodik (3)

Poissonfördelning

- *Om p är litet och N mycket större än n, så att p + n / N < ca 10 %, kan vi approximera $p_X(k)$ med en Poissonfördelning, Po (np):*

$$p_X(k) \approx \frac{e^{-np}(np)^k}{k!} = Po\,(np);\ k = 0, 1, 2, \ldots$$

Poisson-/normalfördelning

- *Är dessutom np tillräckligt stort, minst ca 15, med värdet på p fortfarande litet och därmed q = ca 1, kan vi i formlerna för Normalapproximation ersätta npq med np och sålunda få:*

$$P\,(a < X <= b) \approx \Phi\left(\frac{b + ½ - np}{\sqrt{np}}\right) - \Phi\left(\frac{a + ½ - np}{\sqrt{np}}\right), \text{ med halvkorrektion; } X \in Norm\,(np, \sqrt{np}\,)$$

- *Detta är nog att betrakta som överkurs…*
  Det viktigaste är vilka möjligheter vi har att använda oss av normalfördelningen.

Appendix: Normalfördelning

❖ Normalapproximation

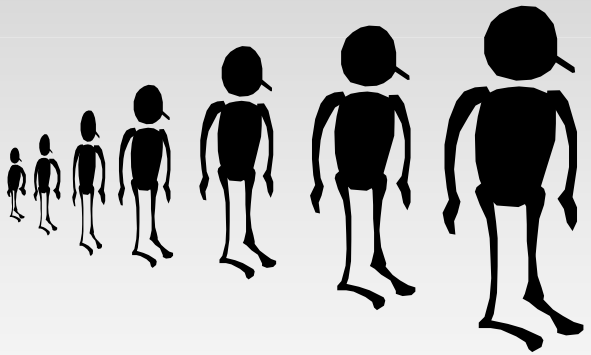
$$P(a < X <= b) \approx \Phi\left(\frac{b-np}{\sqrt{npq}}\right) - \Phi\left(\frac{a-np}{\sqrt{npq}}\right); X \in Norm(np, \sqrt{npq})$$

➢ ***I praktiken använder vi vanligen formeln med b = ∞ eller a = b, vilket kan slås upp i statistiktabeller och skrivs på följande sätt:***

$P(X > \lambda_\alpha) = \alpha$, *för* $b = \infty$

respektive

$P(-\lambda_{\alpha/2} < X < \lambda_{\alpha/2}) = 1 - \alpha$, *för* $a = b$



- ***Vanliga och för oss användbara värden på a är 5 %, 1 % och 0,1 % - typiska acceptabla felnivåer.***

Appendix: Intervallskattning

Konfidensintervall, normalfördelning

- ***Om vi vill sätta upp ett konfidensintervall för den nyss nämnda fördelningen, får vi (med normalapproximation,*** $X \in Norm(np, \sqrt{npq})$, $X/n \in Norm(p, \sqrt{pq/n})$***):***

$$I_p = (p^* - \lambda_{\alpha/2}\sqrt{p^*(1-p^*)/n}\ ,\ p^* + \lambda_{\alpha/2}\sqrt{p^*(1-p^*)/n}\ )$$

- ***Motsvarande för en fördelning nära nollpunkten, med ett enkelsidigt intervall:***

$$I_p = (0,\ p^* + \lambda_{\alpha}\sqrt{p^*(1-p^*)/n}\ )$$

Exempel

- ***Antag att vi skulle finna fem fel i ett stickprov på 250. Uttrycket under roten blir då 0,02 x 0,98 / 250, vilket ger ett värde på rotuttrycket på ca 0,00885. Skulle vi nu vilja uttrycka oss med 95 % säkerhet, letar vi i en normalfördelningstabell upp*** $\alpha = 0{,}05$ ***(eller snarare*** $\alpha/2 = 0{,}025$***) och får*** $\lambda_{\alpha/2} = 1{,}9600$***.***
- ***Med 95% säkerhet ligger då felkvoten i intervallet 0,020 ± 0,017, dvs 0,3 - 3,7 %.***

Appendix: Stickprovsstorlek

Bestämning av stickprovsstorlek

- *Genom att istället föreskriva intervallets längd, liksom konfidensnivån, kan stickprovsstorleken styras.*
- *För att detta ska bli effektivt bör vi dock ha en uppfattning om storleken på p*. I annat fall bör vi anta värsta möjliga utfall, att p* = 1 - p* = 0.5, så att det under rotuttrycket för intervallet står 0,25 / n.*

Exempel

- *Antag att vi med 99 % säkerhet ska kunna säga att p* är korrekt, med 1 % marginal åt varje håll. Rimligen är inte mer än 1 % av beståndet felaktigt; snarare rör det sig om en tiondels procent…*
- ***Intervallängden = 2 x 0,01 =*** $2\ \lambda_{\alpha/2}\sqrt{0{,}01(1-0{,}01)/n} = (\alpha/2 = 0{,}005)\ 2 \text{ x } 2{,}5758 \text{ x } \sqrt{0{,}0099/n}$

  $n = 0{,}0099 \text{ x } (2{,}5758 / 0{,}01)^2 = 656$

- ***Egentligen bör vi räkna med ett enkelsidigt intervall så nära nollpunkten, vilket ger:***
- ***Intervallängden = 0,01*** $= \lambda_{\alpha}\sqrt{0{,}01(1-0{,}01)/n} = (\alpha = 0{,}01)\ 2{,}3263 \text{ x } \sqrt{0{,}0099/n}$

  $n = 0{,}0099 \text{ x } (2{,}3263 / 0{,}01)^2 = 536$

Appendix: Exempel (1)

Normalinspektion, enhetsvis, singelprov

- *Antag att vi har ett systemkomplex med ca 20 enheter, i vilka ca 500 kodändringar vardera gjorts. Önskemålet är att 99 % av kodändringarna ska vara korrekta.*
- *Enligt Table I, General Inspection Level II, ska Sampling Plan L väljas.*
  För den enklaste sortens plan, Single Sampling Plan, tas ett stickprov om 200, varefter enheten accepteras, om färre än sex fel hittas.

Specialinspektion, enhetsvis, singelprov

- *Emellertid kan vi även välja att göra en specialinspektion enligt någon av nivåerna S-1 - S-4 i tabellen, för s.k. specialinspektion.*

  (Detta förutsätter att vi kan acceptera en större risk, men det kan vi parera genom att sänka nivån för vilket slutresultat som är acceptabelt, den s.k. AOQL, i Table V-A).

  Nivån S-4 motsvarar plan G, där man med ett stickprov på 32 st når en nominell nivå på kvaliteten på 0,40 % och en slutlig dito på ca 1,2 %.

  Detta förutsätter dock att inga som helst fel tillåts för att enheten ska accepteras!

Appendix: Exempel (2)

- **Specialinspektion, helhetslösning, dubbelprov**
  - *Alternativt skulle vi även kunna välja att se hela systemkomplexet som ett enda paket med ändringar, där vi som nyss kan välja att följa nivån S-4. Antag vidare att vi vill använda inspektion i två steg, s.k. Double Inspection, vilket då innebär plan J.*

    Enligt denna uppnås en AOQL på 1,1 % för Double Inspection, som visar sig kräva ett första stickprov på 50 st för hela komplexet, med ett eventuellt andra prov på lika många till om (exakt) ett fel upptäcks. Upptäcks totalt två eller fler fel, förkastas allt!
- **Specialinspektion, helhetslösning, obegränsad storlek**
  - *Antag till sist att vi har ett systemkomplex med en enorm mängd enheter, där miljontals ändringar gjorts och 99 % av dessa ska vara korrekta.*
  - *Väljer vi ännu en gång nivån S-4, finner vi att det nu borde vara plan K som gäller.*

    Denna skulle kräva ett stickprov på 125 st för singelinspektion och 80 (+80) st för dubbelinspektion. AOQL visar sig också vara tillräcklig, med omkring 0,29 % för singelinspektionen och 0,67 % för dubbelinspektionen.

    (Egentligen räcker nivå S-3, som med plan H ger AOQL = 0,74 % för singelfallet, vilket kräver ett stickprov på endast 50 st !)

Appendix: Tillämpningar (1)

- **Dokumentgranskning**
  - *Ni har en färdigställd kravspecifikation, 200 sidor, i snitt 20 krav per sida. Projektledaren vill ha en statusöversikt under dagen, med 95 % konfidens.*
  - *Förslag till lösning: Specialinspektion (S-4), dubbel inspektion, 20 i varje. Godkännandenivå: Max 1 fel i inspektion 1, max 4 fel t.o.m. inspektion 2. För att ge snabbt besked krävs alltså bara att 1 % av kraven kontrolleras.*
- **Kodändring – byte av valuta**
  - *I en hel banks samtliga system ska SEK i tillämpliga fall ersättas med EUR. Ändringarna görs i huvudsak i ett annat land, med halvmaskinellt kodstöd. Dessa levereras i omgångar om ca 20.000 ändringar vardera, totalt 30 ggr. Hur mycket ska kontrolleras för att uppnå 99,9 % konfidens för allt detta?*
  - *Förslag till lösning: Normalinspektion (II), enkel inspektion per leverans. Godkännandenivå: Max 1 fel i inspektionen, som görs på 500 ändringar. Observera att det inte är lämpligt att verifiera alla leveranser samtidigt, framför allt av praktiska skäl – en leverans utgör här en typisk batch. Totalt blir det därmed 15.000 ändringar som kontrolleras – i bästa fall.*

Appendix: Tillämpningar (2)

Urval bland (täckande) testfall

- *Det finns 1.750 menyfunktioner byggda för ett system, täckta i 10.500 testfall. Hur många fall behöver man färdigställa till testerna, köra och stämma av, för att uppnå en konfidens på 99,8 % - och i en inledande test minst 99 %?*
- *Förslag till lösning: Specialinspektion (S-4), dubbel inspektion, 32 i varje, för den inledande testen, med godkännandenivå: max 0 fel i inspektion 1, max 1 fel t.o.m. inspektion 2. Normalinspektion (II), dubbel inspektion, 200 i varje, för den fullständiga testen, med samma godkännandenivå!*

  Observera att det är klokt att ta fram och förbereda fler testfall än så. Om testfallen består av unikt täckande steg, kan färre fall behövas.

Redan genomförda (täckande) tester

- *Man har kört ca 300 väl täckande testfall, om vardera fem unika teststeg. Inget enda allvarligt fel har påträffats under denna slutgiltiga testomgång. Med vilken konfidens kan systemet ifråga nu antas vara fritt från sådana fel?*
- *Förslag till lösning: Jämför med tabeller eller SQC online för 1.500 i batch. Godkännandenivå: 0 fel i inspektionen, vilket visar sig motsvara 99,99 %.*