

Métodos Estadísticos en Ingeniería de Software

Uso de datos en Ing de Software

- ▶ Problema
 - ▶ disponibilidad de datos de buena calidad
 - ▶ es necesario recolectarlos durante el proceso
 - ▶ se ven muy fácilmente los costos y no tanto los beneficios
- ▶ ¿ Que hacer con los datos ?

Métricas

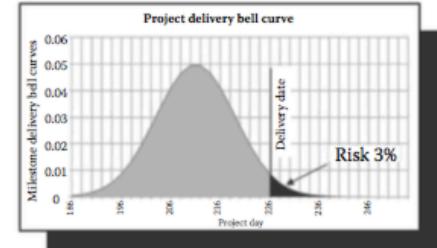
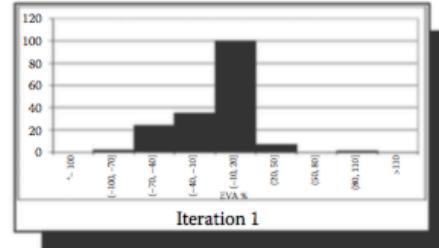
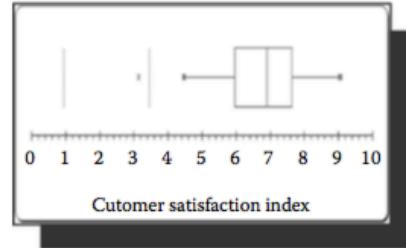
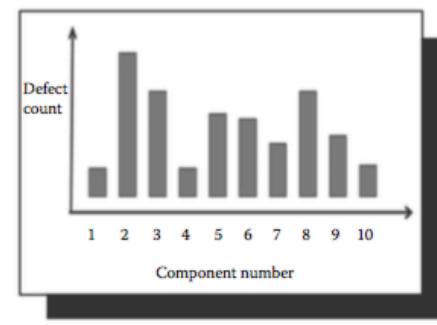
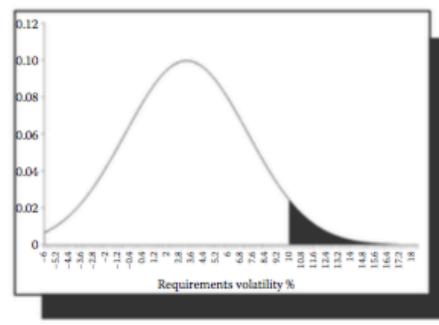
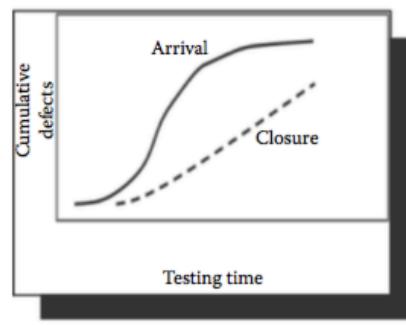
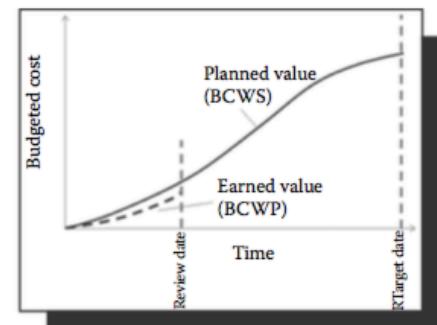
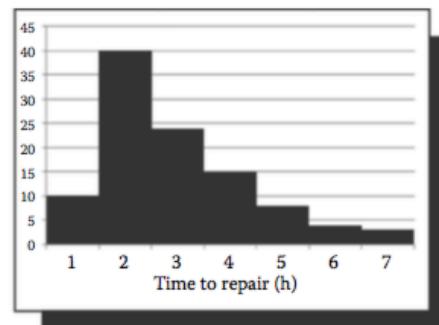
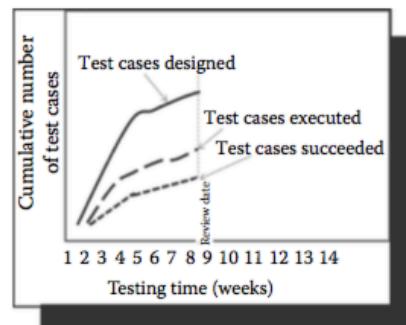
- ▶ una medida base corresponde a la data desnuda
- ▶ una métrica agrega significado
- ▶ las métricas son generadas por necesidades
- ▶ ejemplos de métricas
 - ▶ tamaño - puntos de función
 - ▶ defectos - densidad de defectos

Tipos de métrica

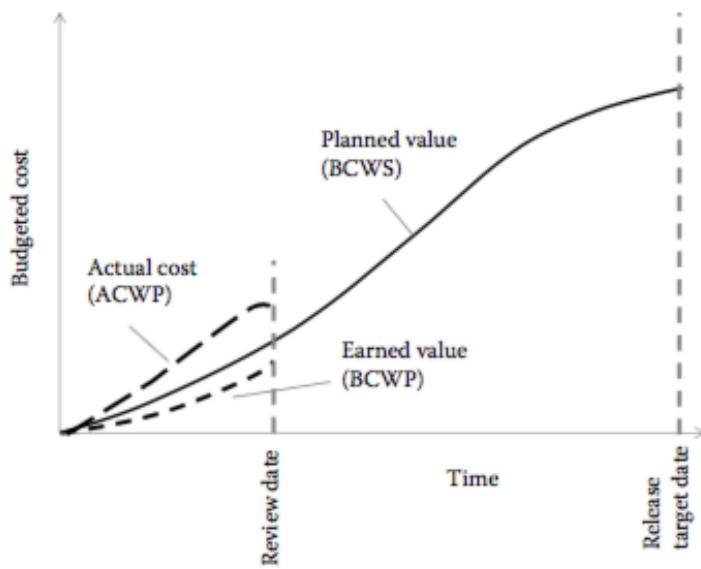
- ▶ de negocios - balance scorecard
- ▶ del proyecto - esfuerzo, calidad, productividad, satisfacción de usuarios
- ▶ del proceso - estabilidad de requisito, complejidad, efectividad de los tests
- ▶ de subprocesso - esfuerzo de diseño, esfuerzo de revisión, esfuerzo de rediseño
- ▶ del producto - tamaño del código, confiabilidad, etc

Dashboard del Proyecto

Las métricas son llevadas a representaciones gráficas



Control de Costo del Proyecto



Core Metrics

- Budgeted cost of work scheduled (BCWS) (also planned value [PV])
- Budgeted cost of work performed (BCWP) (also earned value [EV])
- Actual cost of work performed (ACWP) (also actual cost [AC])

Performance Metrics

- Cost variance = $PV - AC$
- Schedule variance = $EV - PV$
- Cost performance index (CPI) = EV/AC
- Schedule performance index (SPI) = EV/PV
- Project performance index (PPI) = $SPI \times CPI$
- To complete schedule performance index (TCSPI)

Predictive Metrics

- Budget at completion = BAC
- Estimate to complete (ETC) = $BAC - EV$
- Estimate at completion (EAC)
 - $EAC = AC + (BAC - EV)$ Optimistic
 - $EAC = AC + (BAC - EV)/CPI$ Most likely
 - $EAC = BAC/CPI$ Most likely^a simple (widely used)
 - $EAC = BAC/PPI$ Pessimistic
- Cost variance at completion (VAC) = $BAC - EAC$

Métricas del Proyecto

- ▶ Costos (días hombre)
- ▶ Calidad (número de defectos normalizado)
- ▶ Productividad (Locs/dia hombre)
- ▶ Tiempo de Reparación (desde que se detecta un error hasta que está corregido)
- ▶ Satisfacción del Cliente (encuestas)
- ▶ Volatilidad de Requerimientos

Métricas del Producto

- ▶ Tamaño del código (Locs)
- ▶ Complejidad del código (puntos de función, story points, número de clases, número de módulos)
- ▶ Densidad de defectos (por KLoc)
- ▶ Severidad de defectos

Métricas de Testing

- ▶ Estabilidad de requerimientos

$$RSI = \frac{\text{Original req} + \text{Req changed} + \text{Req added} + \text{Req deleted}}{\text{Original req}}$$

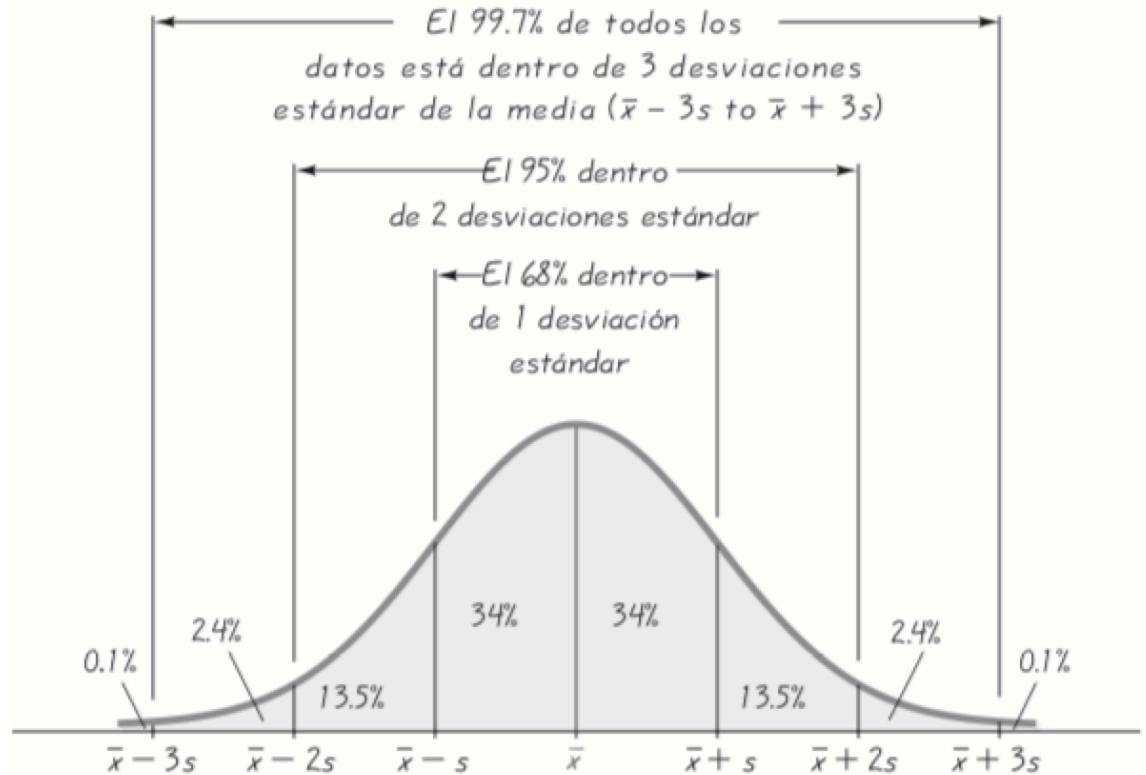
No. of Original Requirements	No. of Requirements Changed	No. of Requirements Added	No. of Requirements Deleted	Requirement Stability Index
100	15	5	5	1.25

- ▶ Efectividad del test

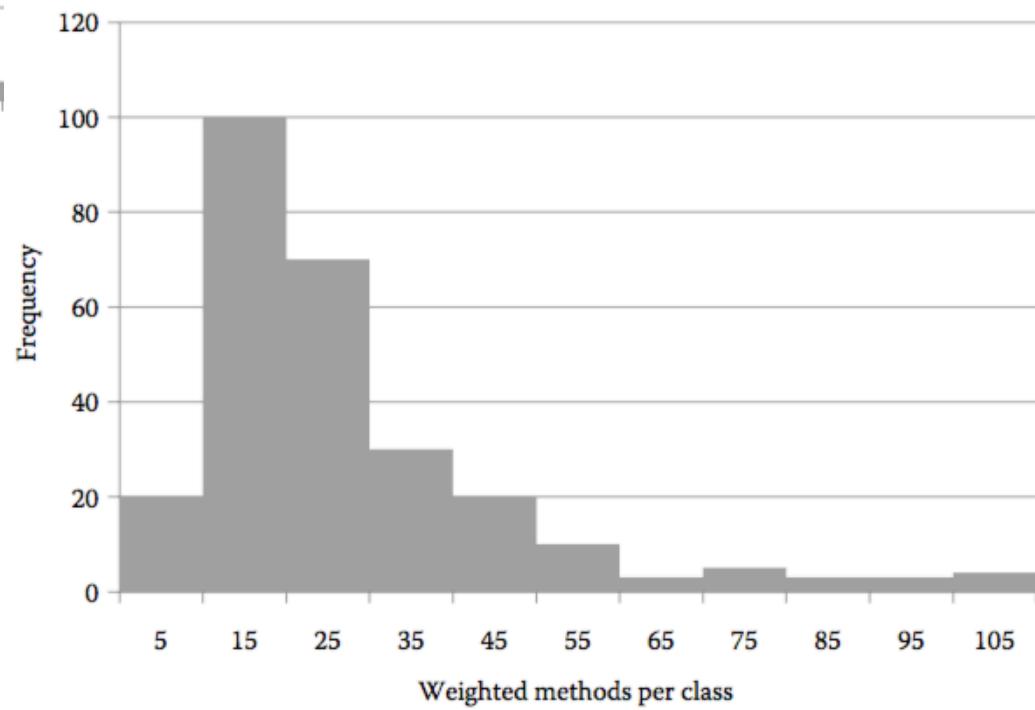
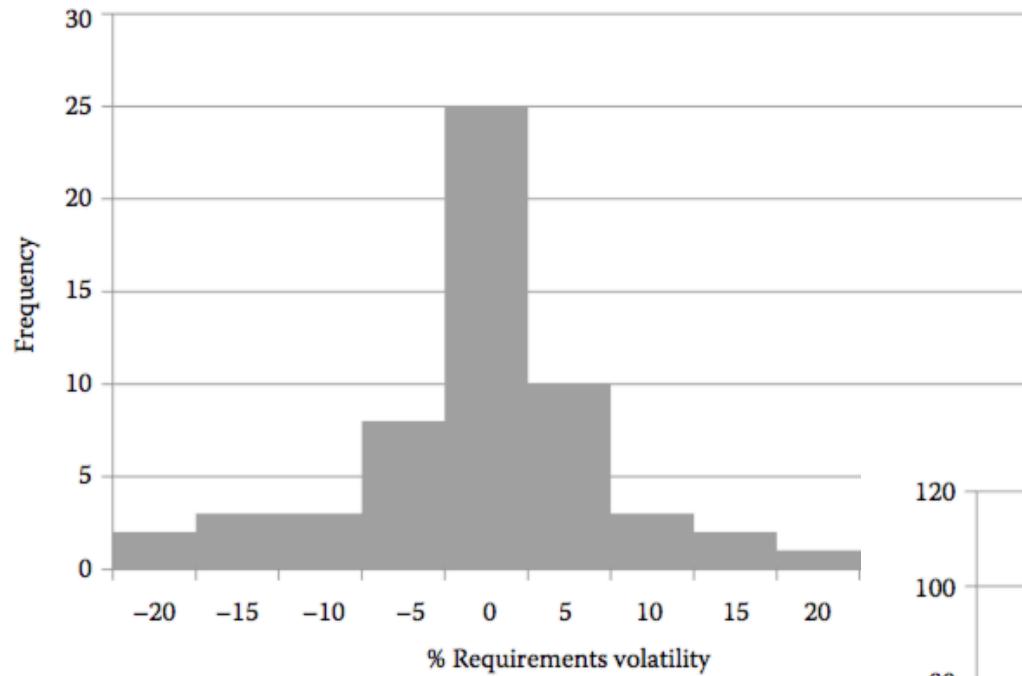
$$\text{Test effectiveness} = \frac{\text{Defects found by tests}}{\text{Defects found by tests} + \text{Defects found by business users}}$$

Estadísticos para describir y comparar datos

- ▶ Tendencia central
 - ▶ media
 - ▶ mediana
 - ▶ moda
- ▶ Variación
 - ▶ rango
 - ▶ desviación standard (95% de datos entre $\bar{x} - \sigma$ y $\bar{x} + \sigma$)
 - ▶ varianza



El útil y versátil histograma

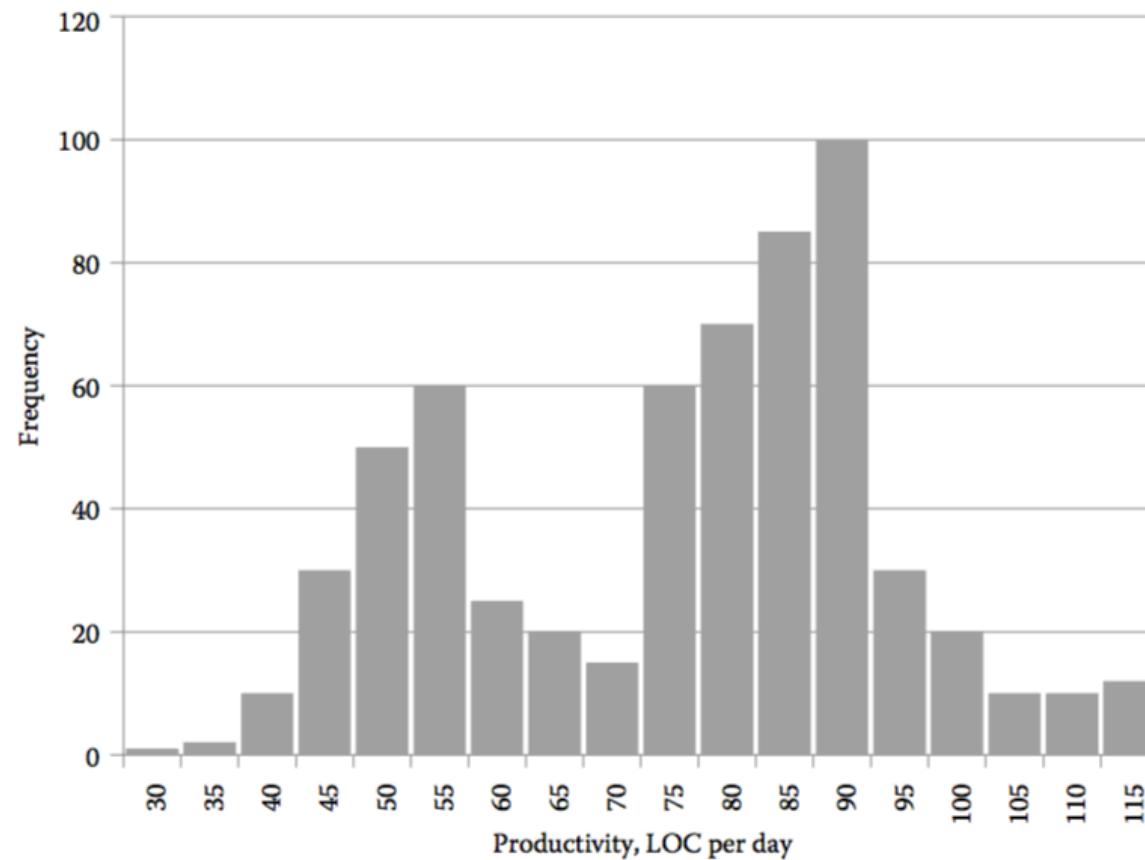


¿ Cuantos intervalos si data es de 1 a N?

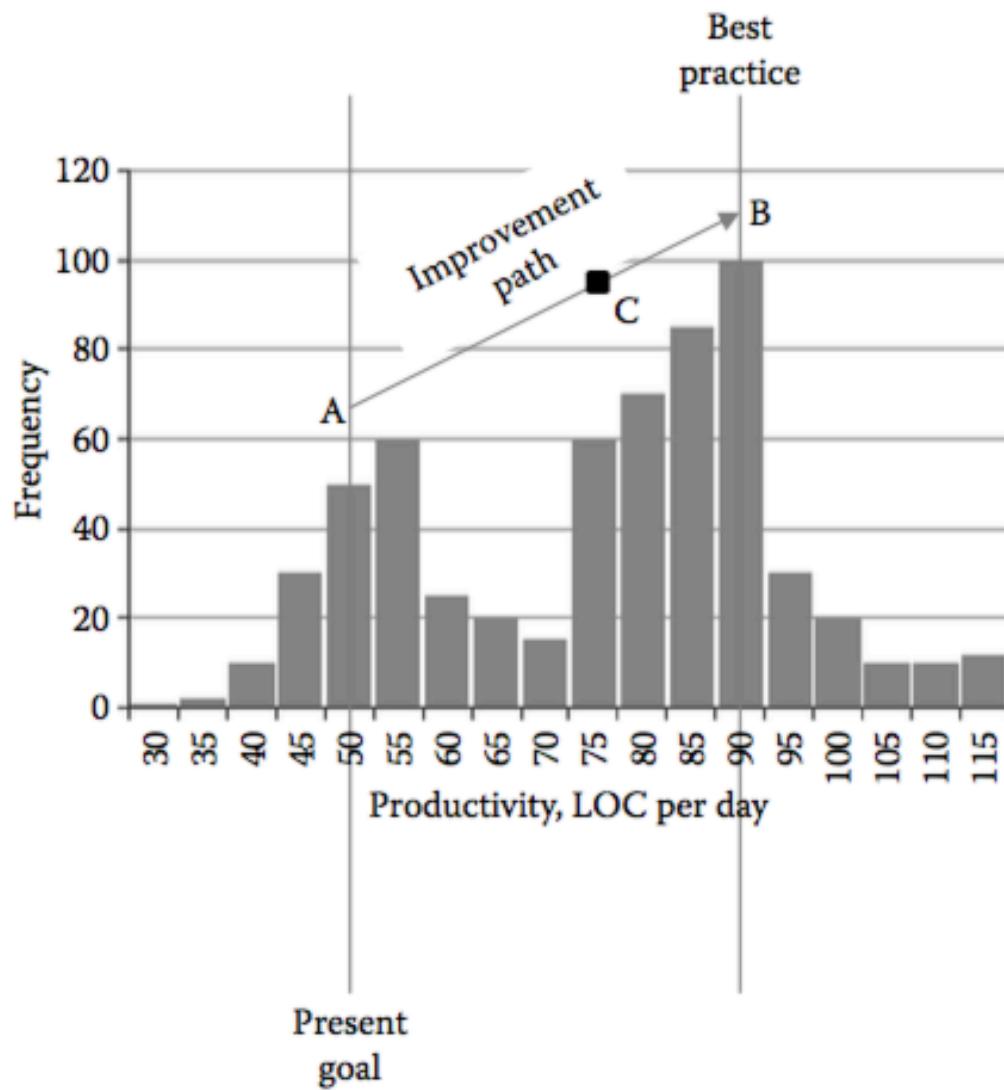
- ▶ Forma varía bastante según cuantos intervalos consideremos
- ▶ Hay varias reglas posibles dado un número n de puntos
 - ▶ $N = n^{0.5}$
 - ▶ $N = \log_2 n + 1$
- ▶ Ejemplo: $N = 100 \Rightarrow 8 \text{ a } 10$

Forma del histograma

- ▶ Forma del histograma puede ya decir bastante
- ▶ Por ejemplo un histograma bimodal para productividad puede deberse a dos equipos de muy distinto desempeño



Mejorar Productividad



Distribuciones de Probabilidad

- ▶ Función que entrega las probabilidades que puede tomar cada valor de la variable aleatoria X
 - ▶ número de defectos en un módulo
 - ▶ número de veces que se cumple con los plazos
- ▶ Hay de dos tipos
 - ▶ discretas (variable discreta)
 - ▶ contínuas (variable continua)

Más sobre Distribuciones de Probabilidad

- ▶ Variable Aleatoria - valor del resultado de un experimento o procedimiento (x)
- ▶ Distribución de Probabilidad - probabilidad de cada valor de la variable aleatoria
 - ▶ la suma de prob para cada valor de x es 1
 - ▶ la prob de un valor individual de x es un valor entre 0 y 1

Distribución Geométrica

- › Sabemos que la probabilidad de obtener un 6 con un dado es 1/6

¿ Cual es la probabilidad de obtener un 6 justo al tercer intentos ?

X = número de intentos

$$P(X = 6) = 5/6 * 5/6 * 1/6 = 0.116$$

Probabilidad de obtener un éxito dado k intentos

$$P(X = k) = (1 - p)^{k-1} p$$

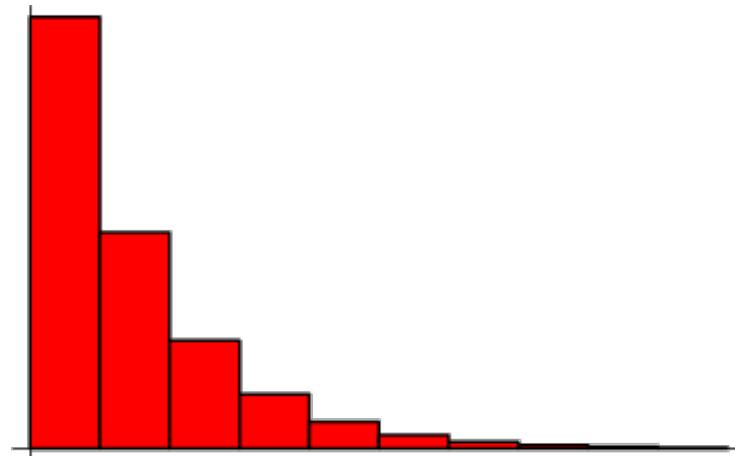
Por ejemplo si la probabilidad de que exista una componente defectuosa es 0.2, entonces la probabilidad de que se descubra a medida que probamos componentes es

$$P(X = 1) = 0.2 * 0.8^0 = 0.20$$

$$P(X = 2) = 0.2 * 0.8^1 = 0.16$$

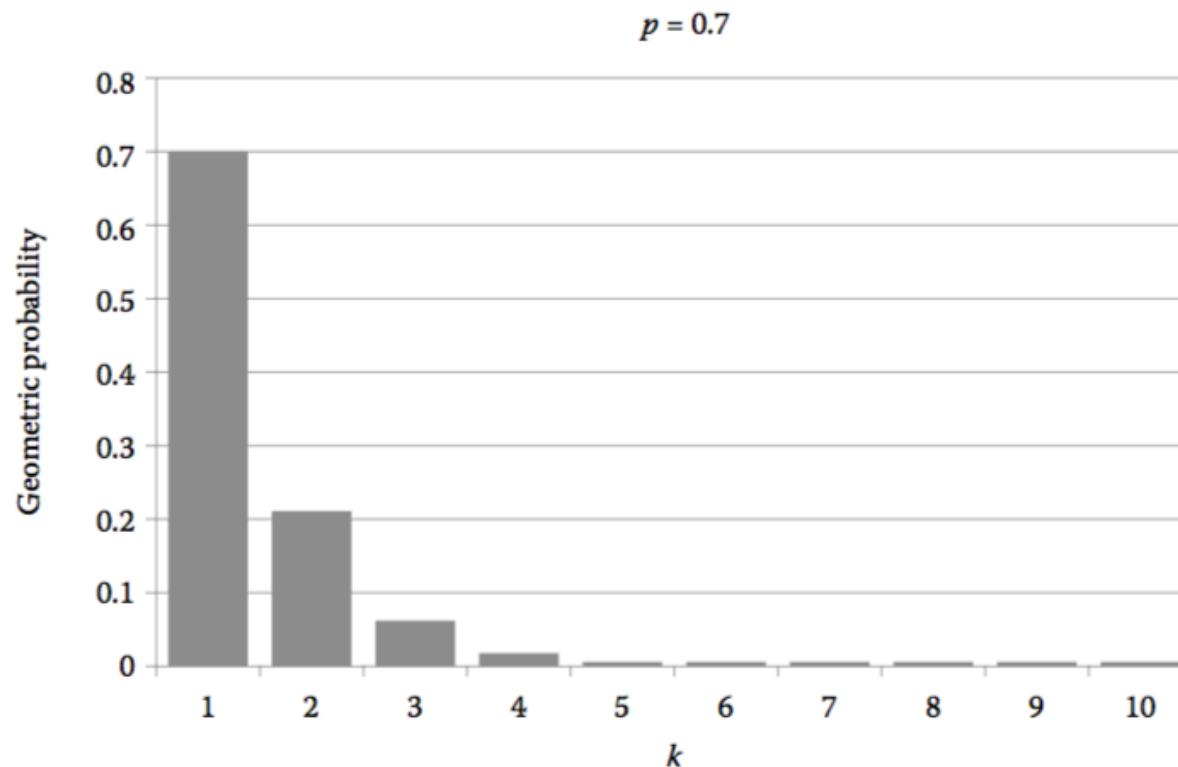
$$P(X = 3) = 0.2 * 0.8^2 = 0.13$$

$$P(X = 4) = 0.2 * 0.8^3 = 0.10$$



Aplicación de Distribución Geométrica

Probabilidad de un diseño libre de errores se considera en 0.7.
¿ Cuál es la probabilidad de necesitar 4 intentos para encontrar uno ?



Distribución Binomial

- ▶ Un caso particular de distribución (muy importante) en cuando el resultado puede tomar solo dos valores
 - ▶ Ilueve y no llueve
 - ▶ pasó el test o no pasó el test
 - ▶ éxito o fracaso
- ▶ La probabilidad de éxito (p) en cada ensayo es siempre la misma
- ▶ $P(x)$ es la probabilidad de x éxitos en n ensayos

Probabilidad de obtener k aciertos en n intentos

- ▶ Probabilidad de obtener k aciertos en n intentos

$$P(X = k) = C_k^n p^k (1 - p)^{n-k} \quad \text{Mean} = np$$

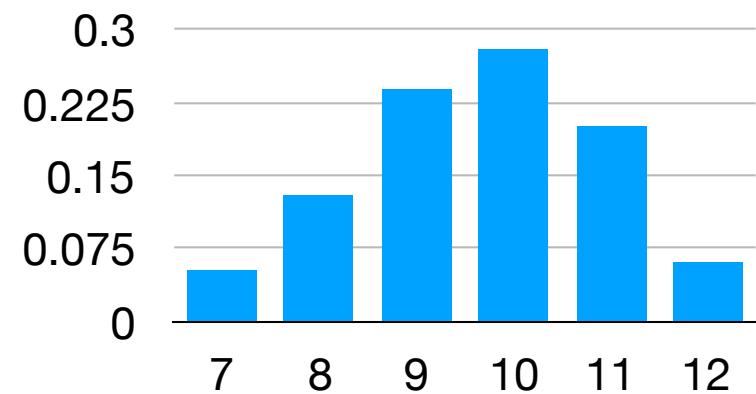
$$\text{Variance} = np(1 - p)$$

- ▶ Ejemplo: Probabilidad de seleccionar a 7 hombres y 5 mujeres de población que son 80% hombres

$$P(X = 7) = C_7^{12} 0.8^7 (1 - 0.8)^{12-7} = 0.053$$

$$\text{Media} = 12 * 0.8 = 9.6$$

$$\text{Varianza} = 9.6 * 0.2 = 1.9$$



Ejemplo de Ingeniería de Software

De datos históricos, probabilidad de cumplir SLA en un proyecto de mejora es del 70%

¿ Cual es la probabilidad de no satisfacer SLA en 5 de las 20 entregas planeadas para el año ?

$$P(X = 10) = C_5^{20} 0.3^5 (0.7)^{15} = 0.18$$
$$p = 0.3 \quad np = 20 * 0.3 = 6$$

Probabilidad que fallen 5 de las 20

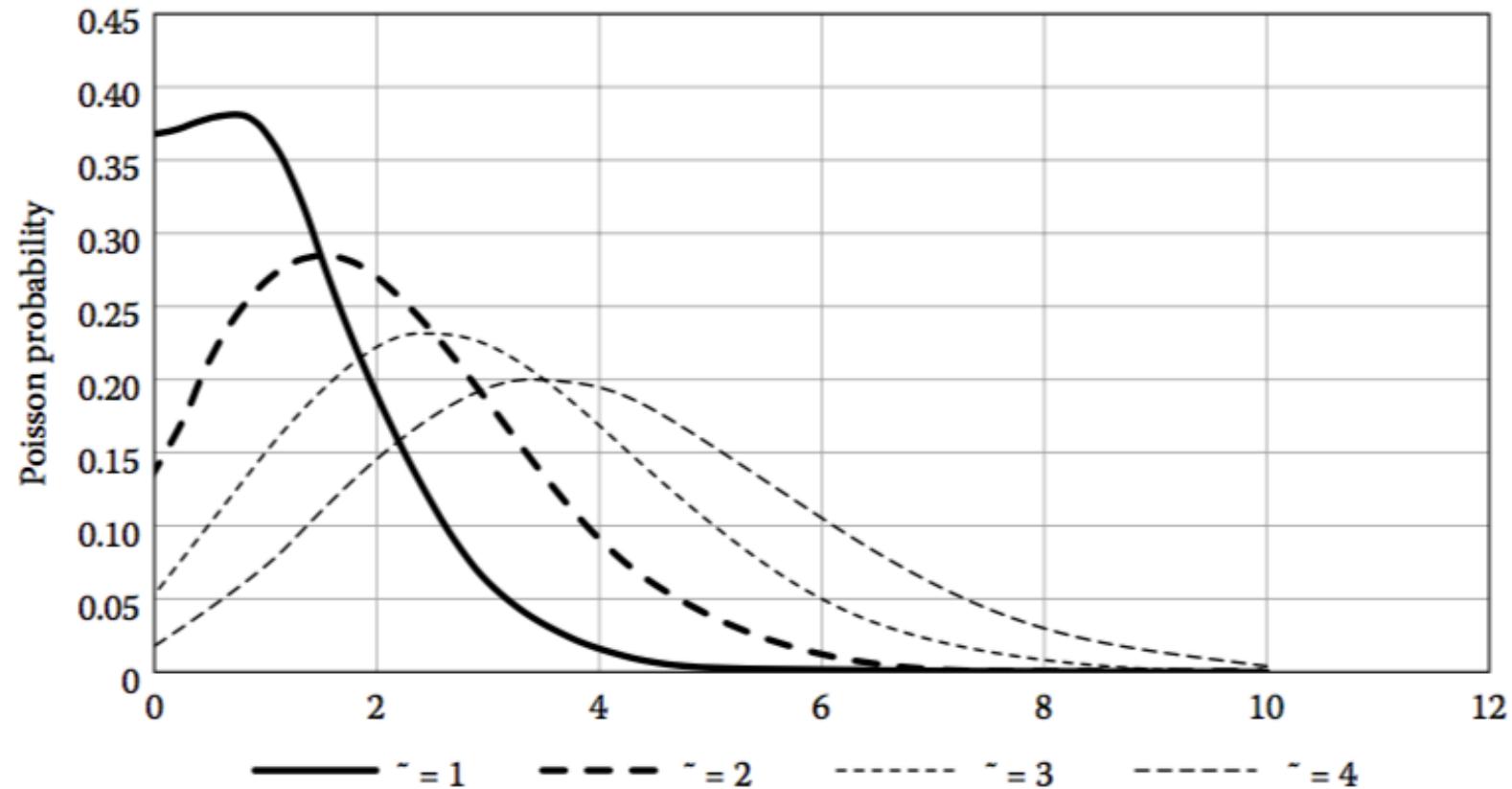
Distribución de Poisson

- ▶ Distribución de probabilidad discreta que se aplica a las ocurrencias de algún suceso durante un intervalo específico
- ▶ x es el número de veces que el suceso ocurre en el intervalo

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

- ▶ media es μ , desviación standard es $\sqrt{\mu}$
- ▶ intervalo puede ser no solo tiempo sino distancia, área, etc
- ▶ a diferencia de binomial, solo depende de la media y no de la probabilidad p o tamaño n
- ▶ no hay límite superior para x

Distribución de Poisson



Aplicación

En testing se está encontrando en promedio 0.3 defectos por módulo

El release incluye 100 módulos

"Intervalo" es el módulo, el suceso es encontrar un defecto

¿ Cual es la probabilidad de encontrar un defecto en un módulo dado ?

Usando Poisson $\lambda = 0.3$, $\sigma = 0.548$ $P(x) = \frac{0.3^1 e^{-0.3}}{1!} = 0.086$

El máximo número de defectos que deberíamos esperar por módulo es

$$0.3 + 3 * 0.548 = 1.94 \quad P(1) = \frac{0.2^1 e^{-0.2}}{1!} = 0.24$$

Otra

El número de emails que llegan al server en un período de 15 minutos puede describirse por una distribución de Poisson con una media de 2 (192 mensajes diarios)

- a) ¿ Cual es la probabilidad de no se reciba ningún mensaje en un intervalo de 15 minutos ?
- b) ¿ Como es la distribución de mensajes que llegan en una hora ?

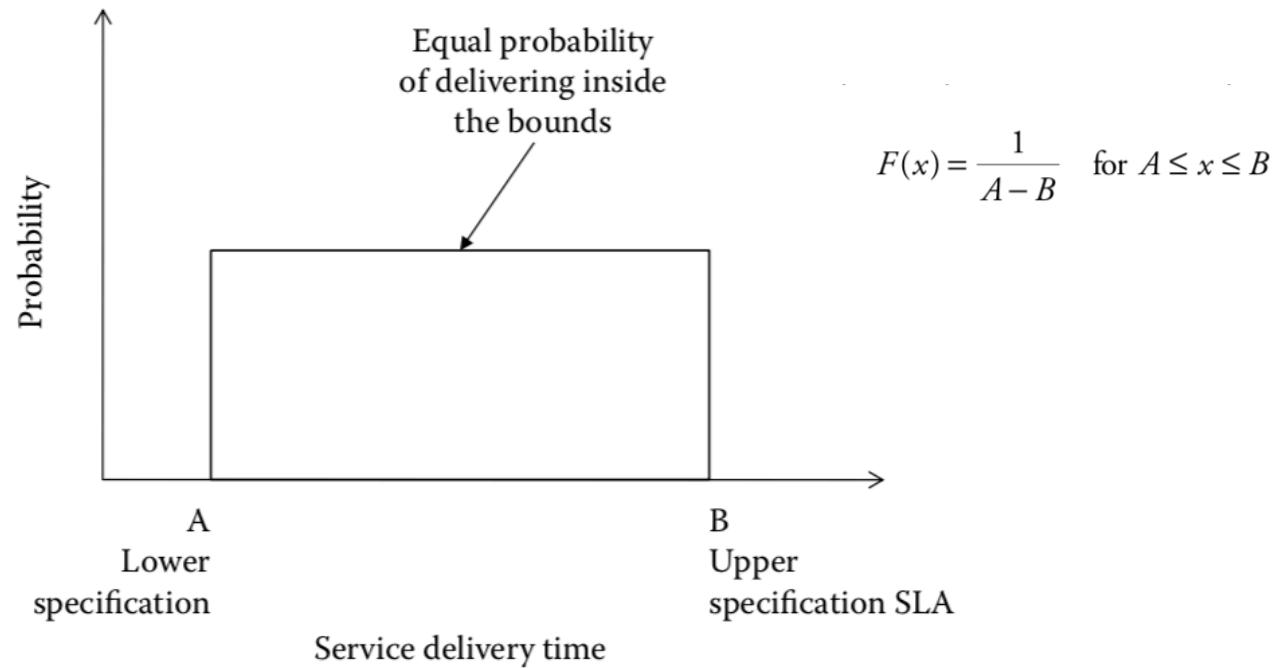
$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

$$\mu = 2 \quad X = 0$$

a) $p(x=0) = e^{-2} = 0.135$

- b) Corresponde a una distribución de Poisson con media 8 (2 en 15 minutos)

Distribución Uniforme (Contínua)



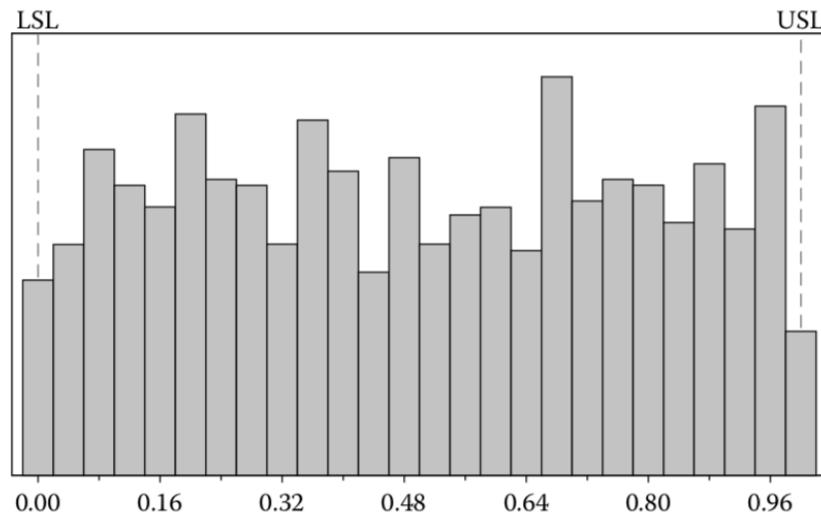
$$\text{Mean} = (A + B)/2$$

$$\text{Median} = (A + B)/2$$

$$\text{Range} = B - A$$

$$\text{Variance} = \frac{(B - A)^2}{12}$$

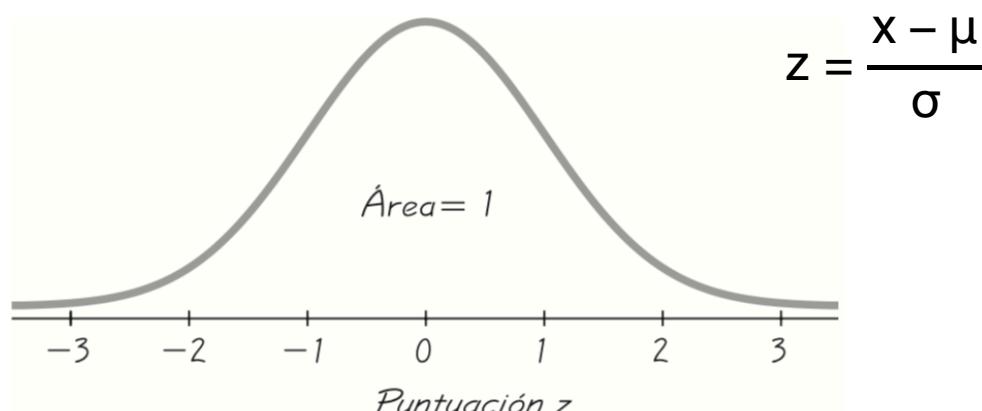
Randon Numbers



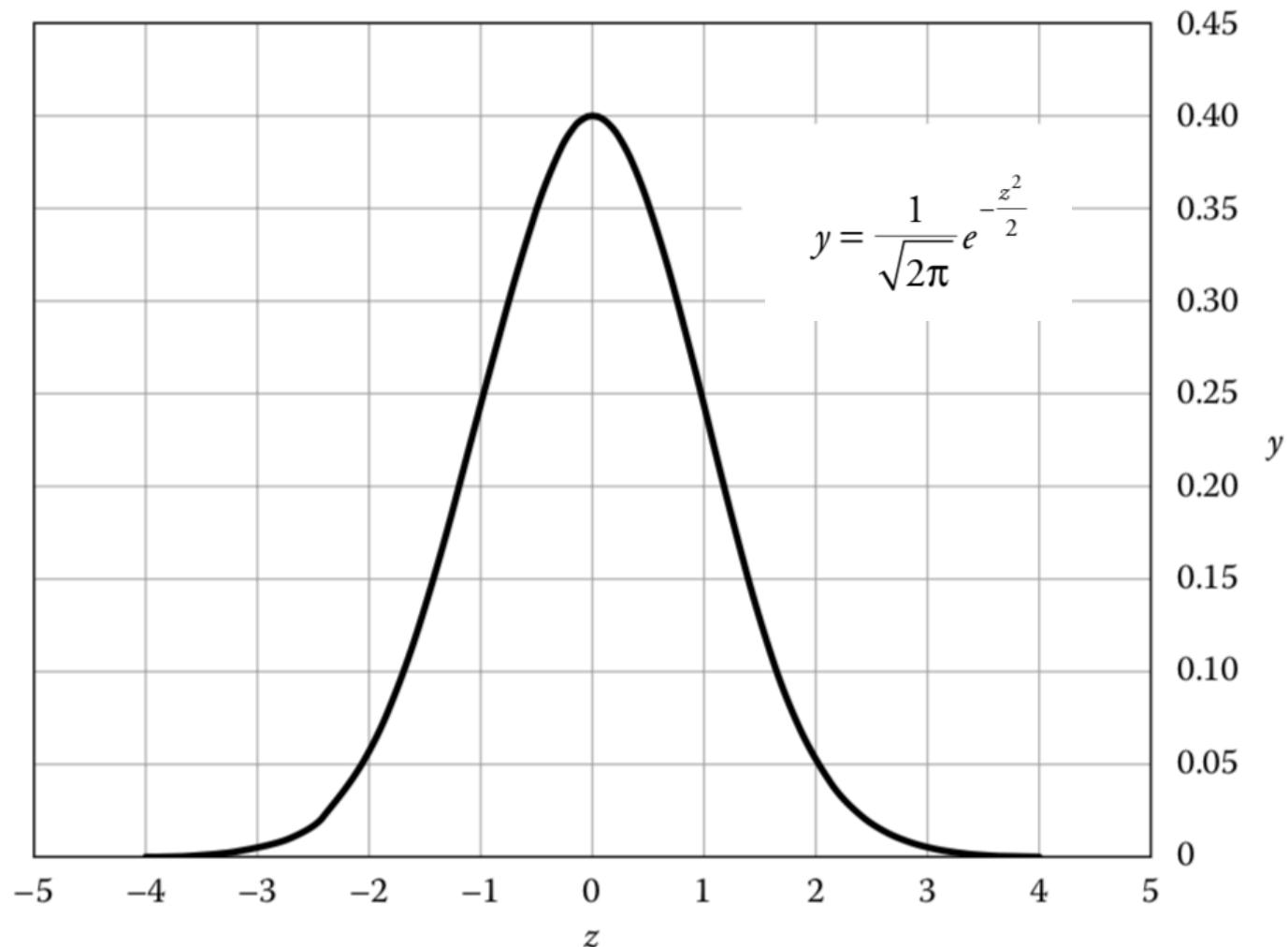
Distribución Normal (Contínua)

$$y = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

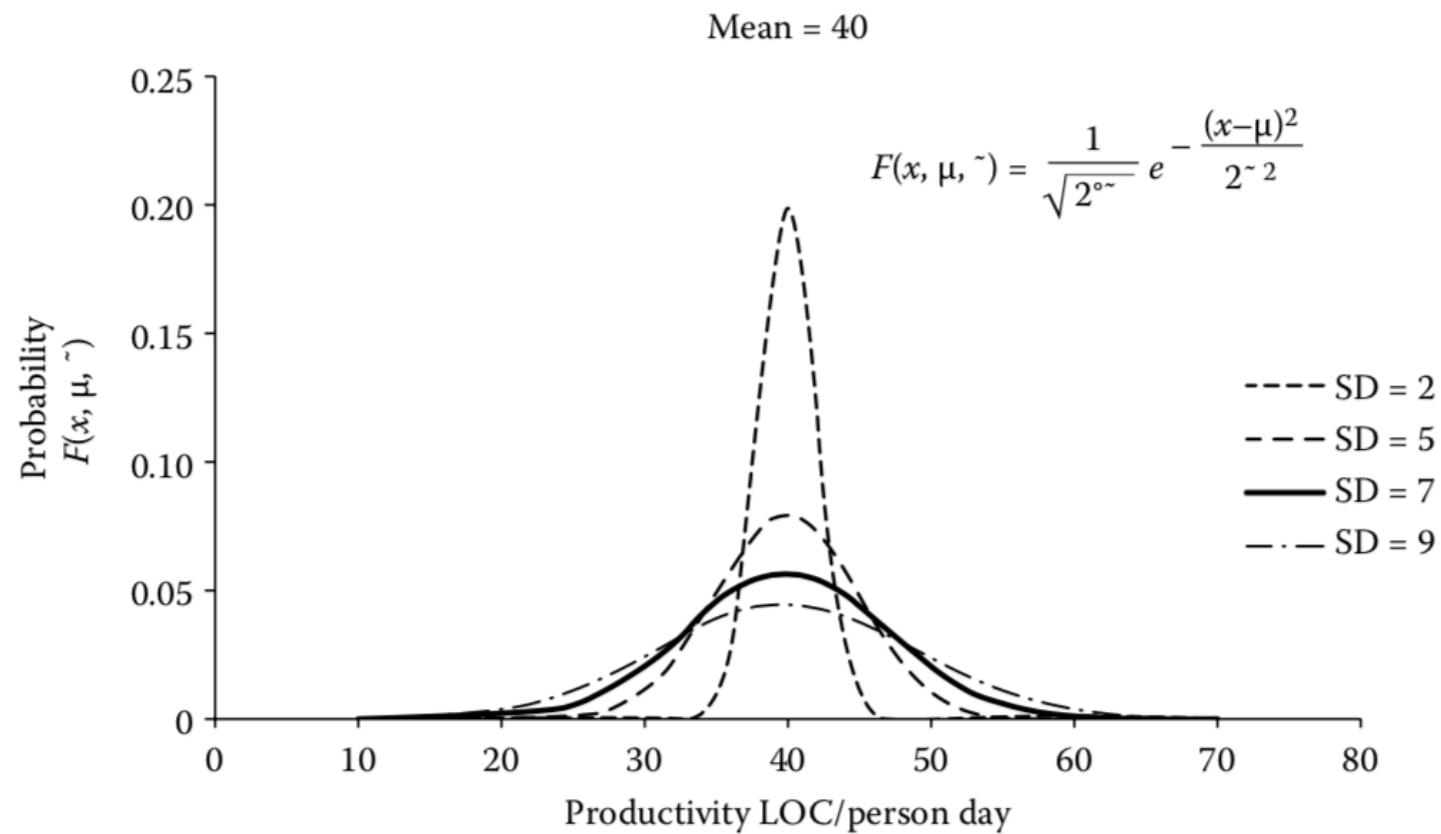
- ▶ Distribución de probabilidad continua
- ▶ Parámetros μ (media) y σ (desviación standard)
- ▶ Distribución normalizada con $\mu = 0$ y $\sigma = 1$



Distribución Normalizada



Productividad en LOC/persona



Ejemplo de Aplicación

La distribución del monto de almacenamiento usado por los usuarios del correo electrónico de una empresa se aproxima a una normal con media en 55MB y desviación estándar de 30 MB

- a) ¿ Por qué esta distribución es imposible que sea correcta ?
- b) ¿ Que proporción de usuarios requiere 75MB o más ?
- c) Si la empresa quiere establecer cuotas (límites de almacenamiento) cual debería ser el valor si se quiere que solo un 1% de la gente exceda la cuota

X es variable aleatoria con distribución normal de media μ y desviación σ

$z = (x - \mu)/\sigma$ (variable estandarizada, media 0)

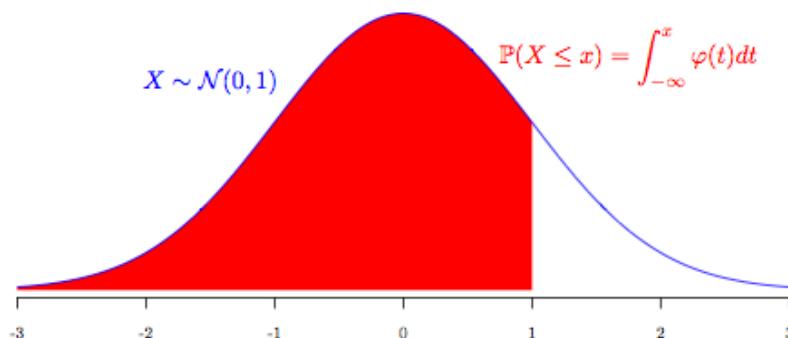
$z = (x - 55)/30$

a) Sabemos que no es posible que haya usuarios con uso de espacio negativo ($x < 0$)

Pero $x < 0 \Rightarrow z < -1.83$ y para ese valor la distribución no da cero sino un valor pequeño ($1 - 0.9664 = 0.033$)

b) $p(x > 75) = p(z > 2/3) = 1 - 0.75 = 0.25$ 25%

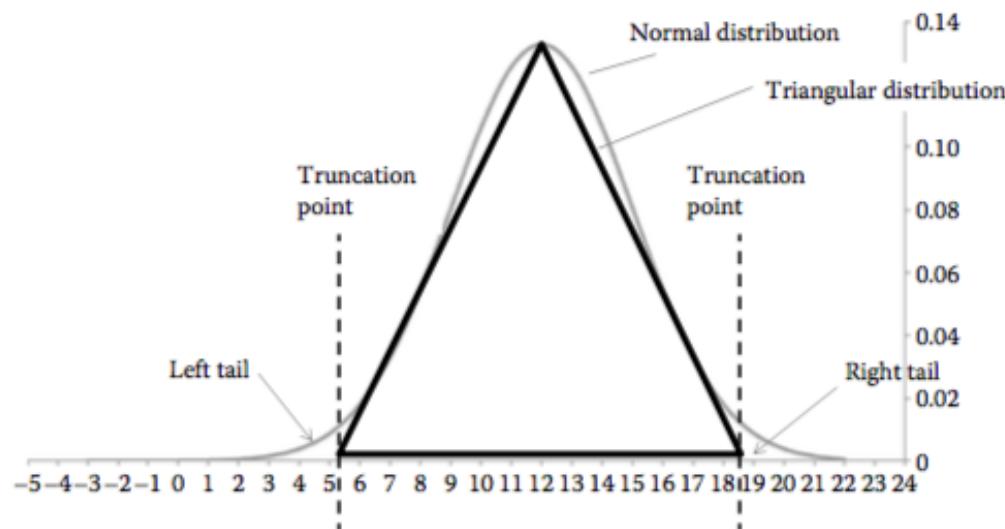
c) $p(z < z_0) = .99 \Rightarrow z_0 = 2.33 \Rightarrow$ cuota = 2.33 *30 + 55 = 125 MB



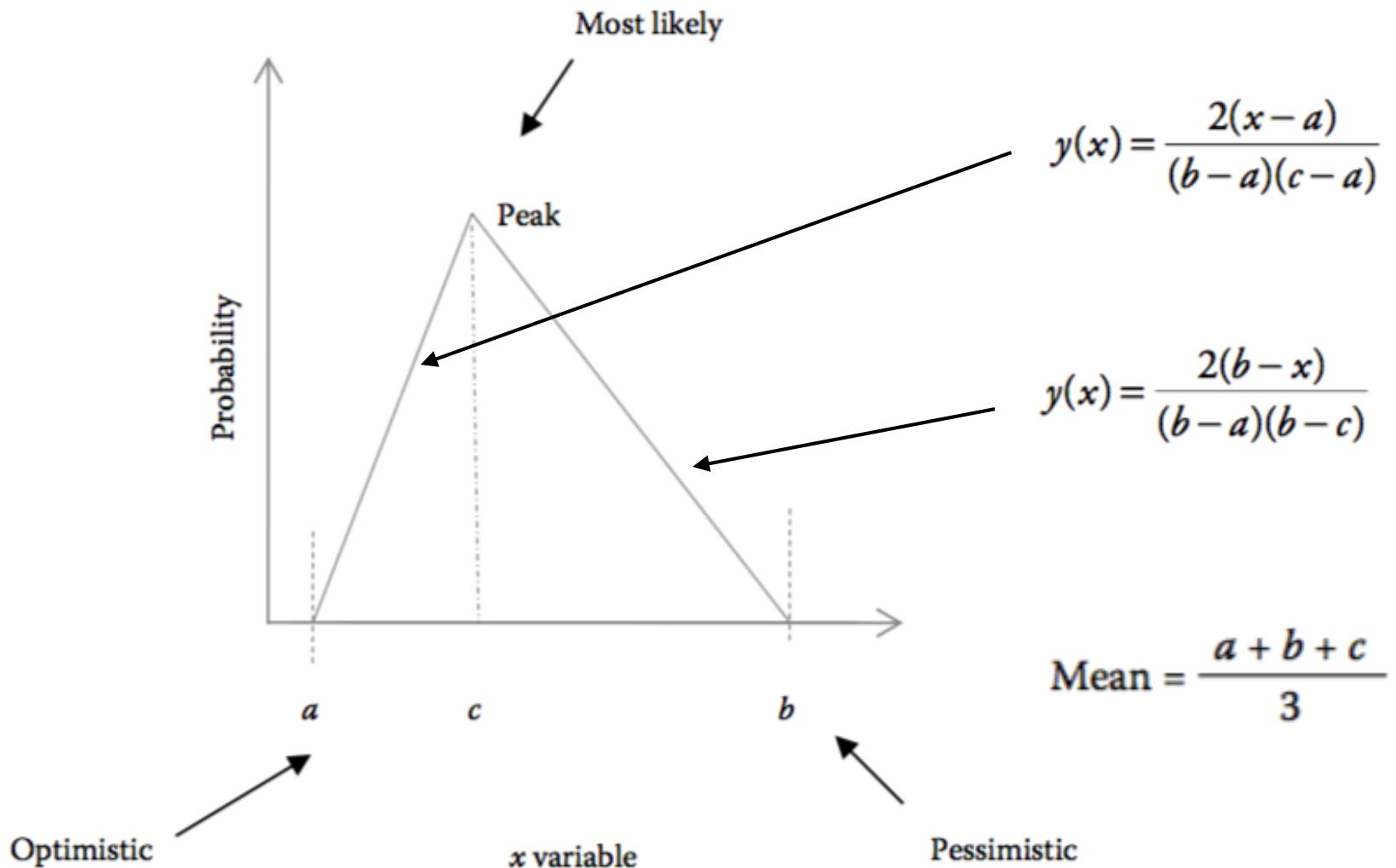
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Distribución Triangular

- ▶ Aproximación
- ▶ Simples
- ▶ Frecuentemente usada como aproximación a una normal



Se adapta bien a estimaciones



Ejemplo

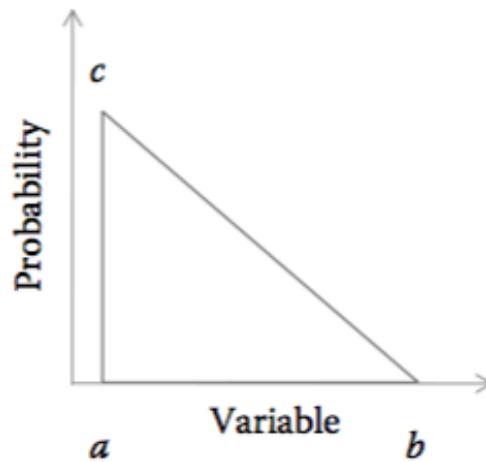
- ▶ Estimaciones de expertos
 - ▶ optimista 25 días
 - ▶ pesimista 50 días
 - ▶ mas probable 30 días
- ▶ Media = $(25 + 50 + 30)/3 = 35$ días
- ▶ PERT da $(25 + 4 * 30 + 50)/6 = 32.5$ días

Procesos Sesgados

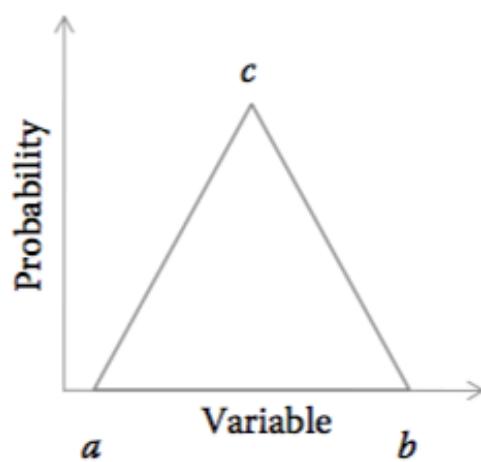
c Process mode

a Lower boundary of process

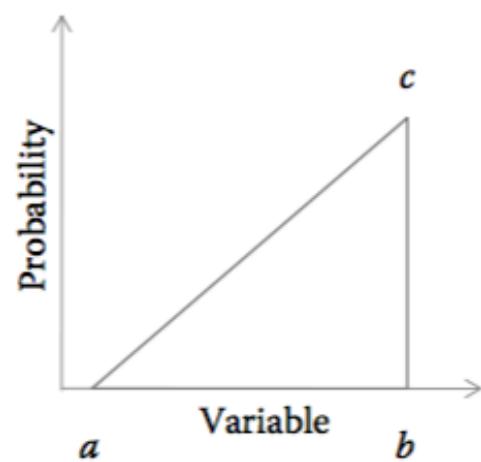
b Upper boundary of process



This is a model of process aligned to some convenient lower bound.



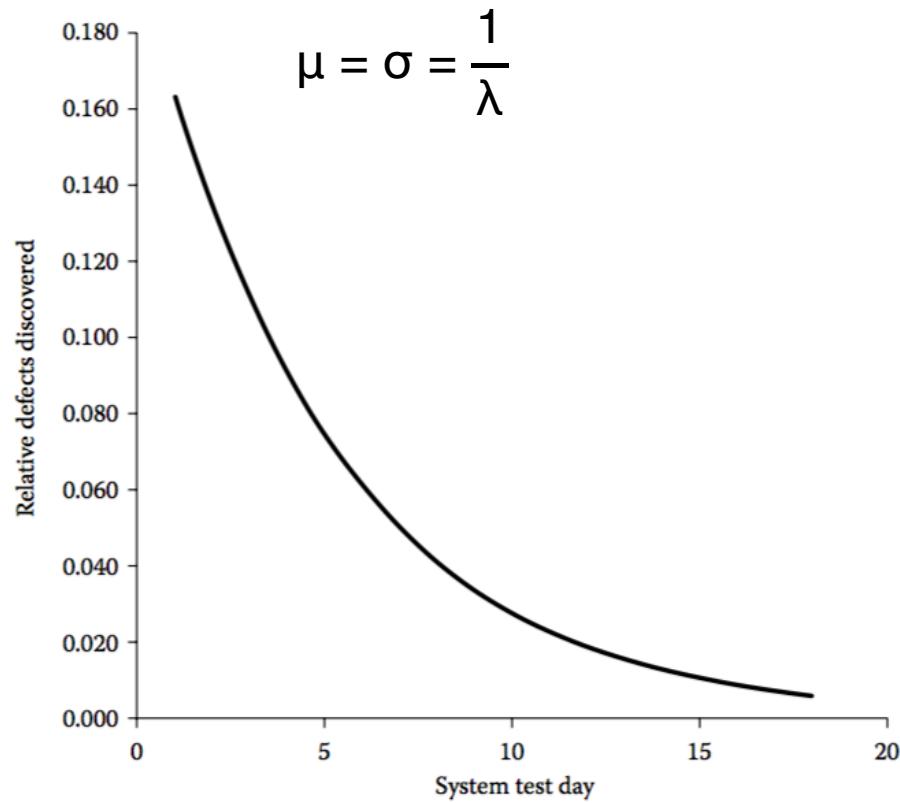
This is a model of a centered process.



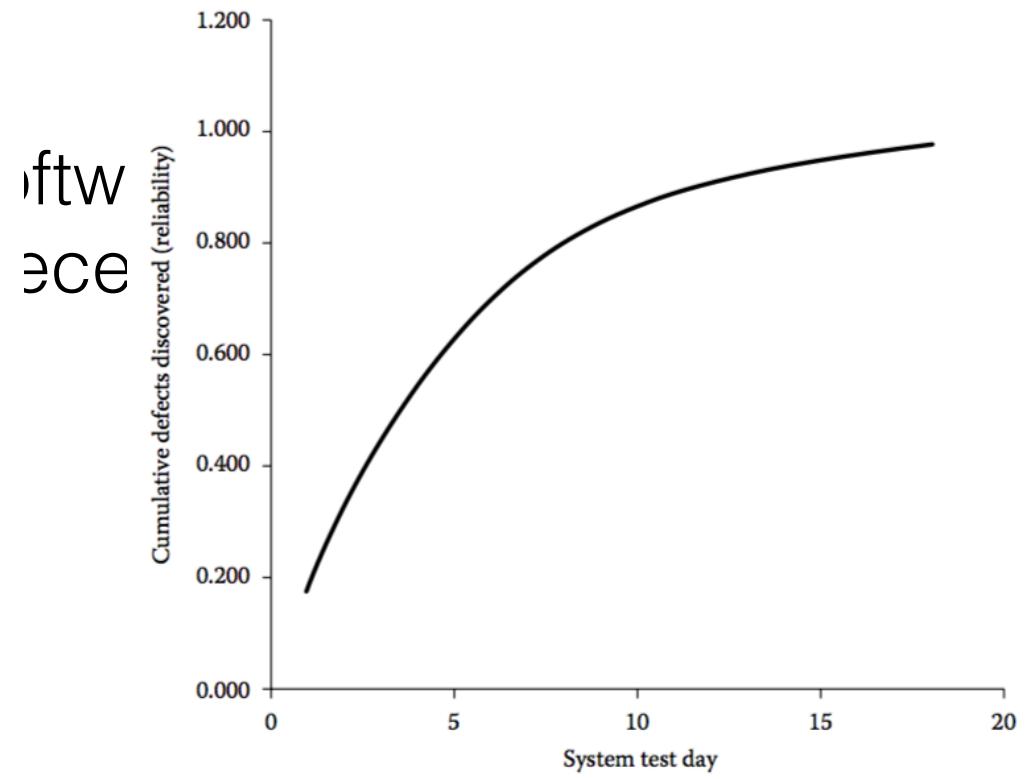
This is a model of a process stretching to the maximum tolerance.

Distribución Exponencial (Contínua)

$$f(x) = \lambda e^{-\lambda x}$$



$$F(x) = 1 - e^{-\lambda x}$$

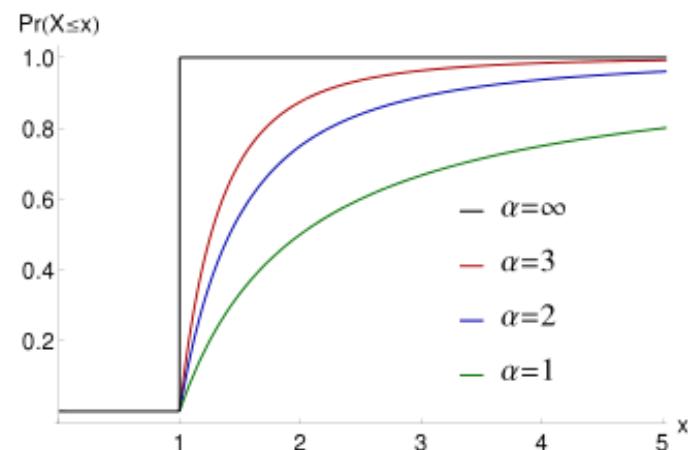
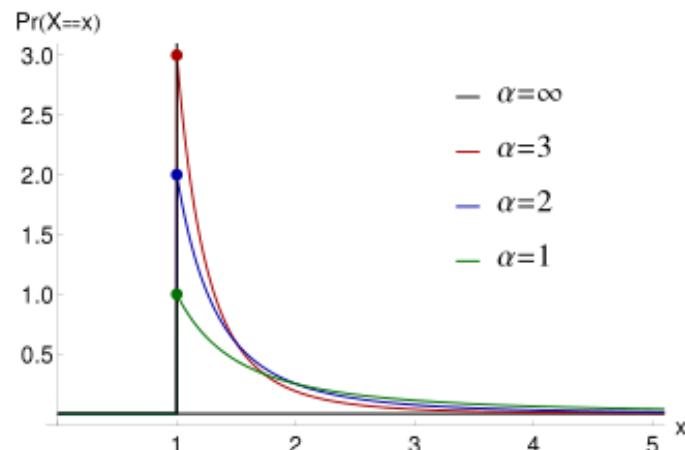


Distribución de Pareto (80-20)

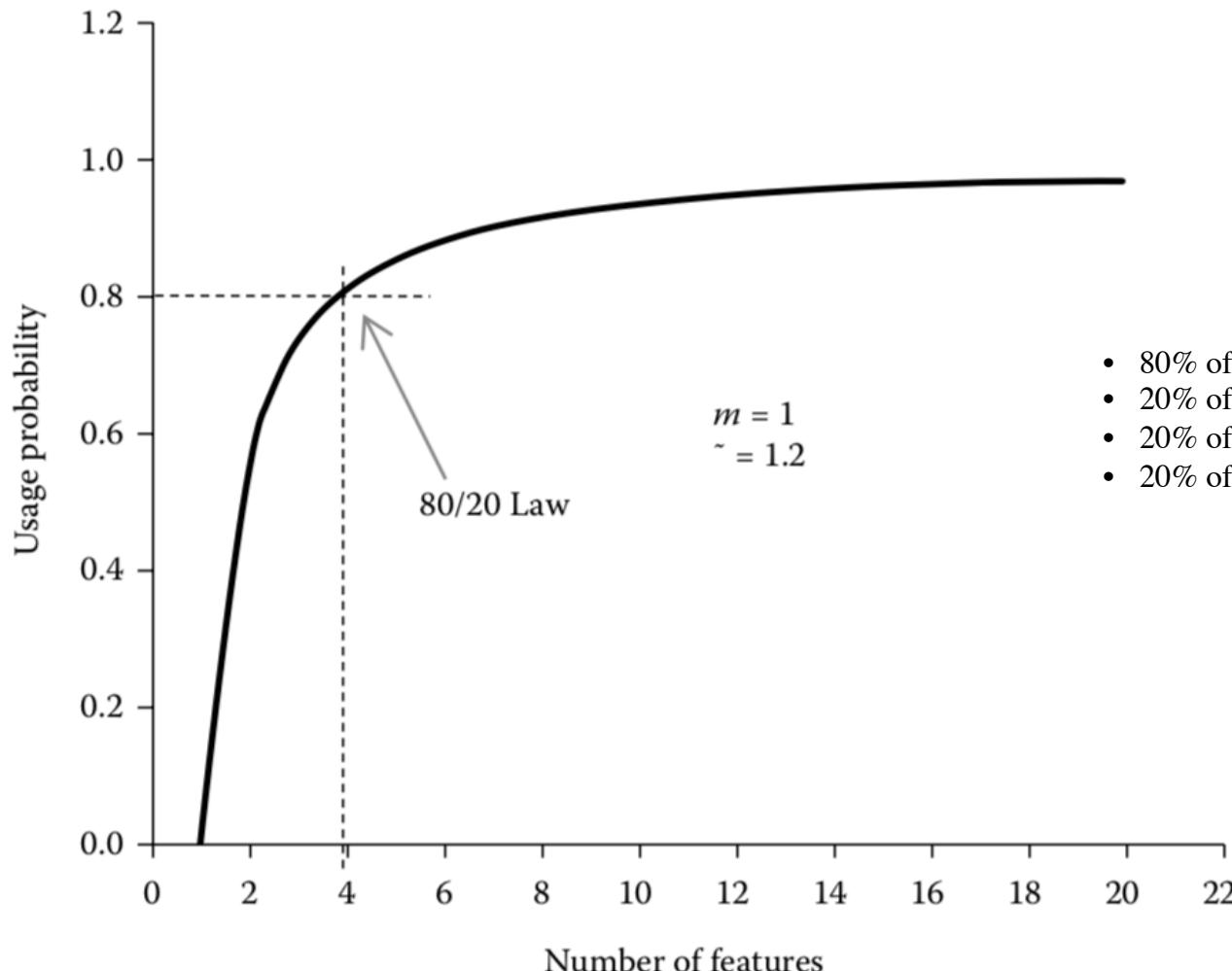
$$f(x) = \left(\frac{x_m}{x}\right)^{\alpha} \quad x \geq x_m$$

$$f(x) = 0 \quad x < x_m$$

$$F(x) = 1 - \left(\frac{x_m}{x}\right)^{\alpha} \quad x \geq x_m$$



El famoso 80/20 de Pareto



- 80% of errors and crashes come from 20% of bugs
- 20% of software components contain 80% of defects
- 20% of defects cause 80% of down time
- 20% of test cases capture 80% of defects

Pareto en Ing de Software

- ▶ recolectar info sobre errores y defectos
- ▶ trazar para cada error su causa
- ▶ seleccionar el 20% de causas que genera el 80% de los errores (Pareto)
- ▶ actuar sobre las causas
- ▶ en el fondo: enfocarse en lo más importante

Teorema de Bayes

- ▶ La probabilidad de un evento futuro está influenciada por la historia
- ▶ $P(A|B) = P(A \text{ and } B) / P(B)$
- ▶ Probabilidad condicional de A dado B es la probabilidad de que ocurran ambos dividida por la probabilidad que ocurra B
- ▶ $P(B)$ - prior probability
- ▶ $P(A)$ - posterior probability

Forma más útil

$$p(A|B) = \frac{p(A \text{ and } B)}{P(B)}$$

$$p(B|A) = \frac{p(A \text{ and } B)}{P(A)}$$

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

Se suele dar una interpretación como la probabilidad de una hipótesis a la luz de nueva data

Significado en contexto de validez de Hipótesis a la luz de nueva Data

$$p(H|D) = \frac{p(H) \cdot p(D|H)}{p(D)}$$

- $p(H)$ probabilidad de hipótesis H antes de ver la data (prior)
- $p(H|D)$ probabilidad de la hipótesis después de ver la data (posterior)
- $p(D|H)$ probabilidad de la data bajo hipótesis H (likelihood)
- $p(D)$ probabilidad de la Data ante cualquier hipótesis (constante de normalización)

Bayes con Galletitas

Jarra A - 30 de vainilla y 10 de chocolate

Jarra B - 20 de vainilla y 20 de chocolate

Si elijo una jarra al azar y saco una galleta que resulta ser de vainilla, ¿Cual es la probabilidad de que la jarra era la A?

Hipótesis previa : Jarra A es del 50% pero ahora hay nueva data

$$p(\text{vainilla}/\text{jarraA}) = 3/4$$

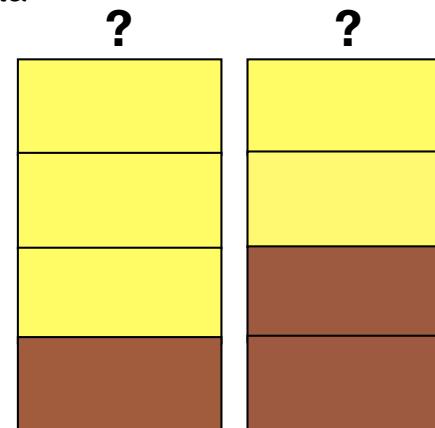
$$p(\text{vainilla}) = 4/8$$

$$p(\text{jarraA}) = 1/2$$

$$p(\text{jarraA}|\text{vainilla}) = \frac{p(\text{vainilla}) \cdot p(\text{vainilla}|\text{jarraA})}{p(\text{jarraA})}$$

$$p(\text{jarraA}/\text{vainilla}) = (5/8 * 3/4)/1/2 = 0.93$$

Es decir, si saqué una de vainilla es mas probable que haya venido de la jarra A



A person tests positive in a lab. The lab has a reputation of 99% correct diagnosis but also has false alarm probability of 5%. There is a background information that the disease occurs in 1 in 1000 people (0.1% probability). Intuitively one would expect the probability that the person has the disease is 99%, based on the lab's reputation. Two other probabilities are working in this problem: a background probability of 0.1% and a false alarm probability of 5%. Bayes theorem allows us to combine all the three probabilities and predict the chance of the person having the disease as 1.94%. This is dramatically less than an intuitive guess.

P_1 : probability of correct diagnosis

P_2 : probability of false alarm

P_3 : prevalent disease probability (background history)

Given Lab Characteristics		
P_1	Reputation of correct diagnosis	99%
P_2	False alarm probability	5%

Question: What is the probability P_0 of a person who tests positive having the disease?

P_3 Disease Probability %	P_0 Bayes Estimation Chance of Having Disease %
0.1	1.9
1.0	16.7
10.0	68.8
20.0	83.2
30.0	89.5
40.0	93.0
50.0	95.2
60.0	96.7
70.0	97.9
80.0	98.8
90.0	99.4

Note: It may be seen that posterior probability depends on prior probability.

Bayes e Ing de Software

- ▶ Estimaciones de costo - juicio a priori de experto se combina con información (data) para producir un modelo a posteriori
- ▶ Estimaciones de productividad