

## Examen de la asignatura "Estadística aplicada a las ciencias sociales"

Profesor Josu Mezo. Examen del 21 de enero de 2010

### Recordatorio de fórmulas (no todas son necesarias)

$$\bar{x} = \frac{\sum x_i}{N} \quad s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} \quad ET = \sqrt{\frac{pq}{n}}$$

$$\bar{x}_c = \sum c_i f_i \quad s_c = \sqrt{\sum (c_i - \bar{x}_c)^2 f_i} \quad ET = \frac{s_x}{\sqrt{n}}$$

$$ET = \frac{\hat{s}_x}{\sqrt{n}} \cdot \sqrt{1-f} \quad ET = \sqrt{\frac{pq}{n}} \times \sqrt{1-f}$$

$$z = \frac{\text{estimador} - \theta_0}{ET} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \quad z = \frac{\text{estimador} - \theta_0}{ET} = \frac{\bar{x} - m}{\hat{s}_x / \sqrt{n}}$$

#### Pregunta nº 1 (5 puntos).

En una base de datos de películas de cine estrenadas en España en 2009, se encuentran, entre otras, las siguientes variables. Clasifícalas según sean de escala nominal, ordinal, o de intervalo, y en el último caso, según sean discretas o continuas (en el sentido "práctico", no en el sentido teórico de la expresión).

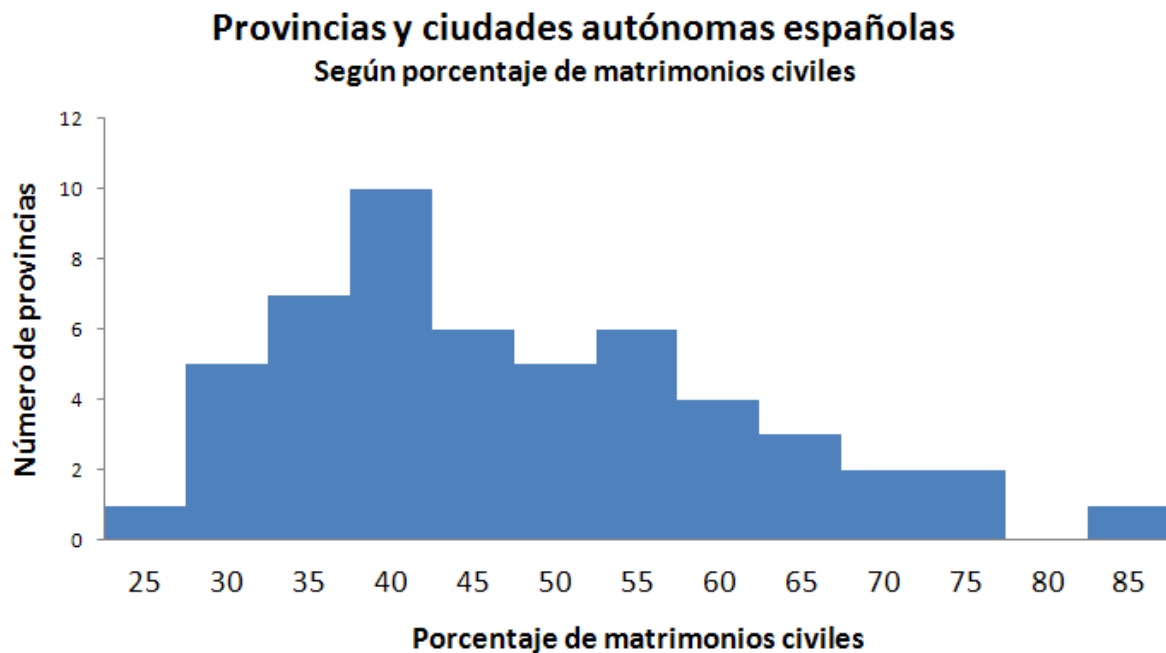
- Número de salas en las que se estrenó **Escala de intervalo, continua**
- Recaudación en euros, en el primer fin de semana **Escala de intervalo, continua**
- Puesto en el ranking de recaudaciones del primer fin de semana **Escala ordinal**
- Número de versiones en las que se estrenó (doblada a español, a catalán, con subtítulos, 3-D...; los valores van de 1 a 7) **Escala de intervalo, discreta**
- País de producción **Escala nominal**

#### Pregunta nº 2 (9 puntos)

El gráfico siguiente representa la distribución de las 52 provincias y ciudades autónomas de España por porcentaje de matrimonios civiles (por tramos de 5 puntos porcentuales, el dato que aparece en el eje de abscisas es el límite superior del tramo representado sobre él). A) ¿Cómo se llama ese tipo de gráfico? B) Explica qué es lo que se aprende al ver este gráfico.

- Es un histograma
- Lo que se aprende es que el porcentaje de matrimonios civiles en las provincias y ciudades autónomas varía bastante, desde una provincia donde está entre el 25 y el 30% hasta otra donde está entre el 80 y el 85%, y una considerable dispersión entre los valores intermedios. El porcentaje más común es entre 35 y 40 (10 casos) y los valores cercanos a ese son también bastante frecuentes con 7 provincias o ciudades entre 30 y 35%, 6 entre 40 y 45. Más de la

mitad de las provincias tienen entre 25 y 50% de matrimonios civiles, y números decrecientes de provincias tienen porcentajes mayores del 55%, hasta llegar al caso único que tiene entre 80 y 85%.



**Pregunta n° 3** (4 puntos) Responde y explica tus respuestas:

En el gráfico anterior, dadas sus características:

- ¿De qué número crees tú que estará más cerca la media, de 40 ó de 50?
- ¿Y cuál será mayor, la media o la mediana?
- ¿Y cuál será más representativo del conjunto de los valores, la media o la mediana?

A) En principio, y sin hacer cálculos, podemos suponer que dada la distribución asimétrica hacia la derecha, los valores altos “tirarán de la media” hacia valores separados de los más comunes, que rondan el valor 40, como hemos visto en la respuesta anterior. Por eso, es más probable que la media esté más cerca de 50 que de 40.

B) Por la misma razón (distribución asimétrica hacia la derecha) la media será mayor que la mediana. De hecho la mediana la podemos calcular fácilmente: si son 52 casos, la mediana será el valor intermedio entre los casos 26 y 27, ordenados de menor a mayor. Contando las frecuencias absolutas encontramos que los casos 26 y 27 están ambos en la clase 40-45%, y por tanto esa será la mediana. La media tendrá seguramente un valor algo superior.

c) Por las mismas razones, como pasa en general en las distribuciones asimétricas, la mediana será seguramente más representativa de los valores comunes que la media, que está “descentrada”.

**Pregunta n° 4 (21 puntos)**

La siguiente tabla presenta la distribución por edades de las madres que tuvieron un niño en España en 2007. Calcula

Edad de la madre	Porcentaje
15 a 19	2,9%
20 a 24	9,8%
25 a 29	23,4%
30 a 34	38,4%
35 a 39	21,5%
40 a 44	3,7%
45 a 49	0,2%
Total	100,0%

- a) La edad media de las madres que tuvieron un niño
- b) La desviación típica de esa edad
- c) El coeficiente de variación
- d) La edad mediana
- e) El primer y tercer cuartil
- f) El rango intercuartílico
- g) La moda

Edad de la madre	Marca de clase ( $c_i$ )	Frecuencia relativa ( $f_i$ )	$c_i * f_i$	$(c_i - \text{media})^2 * f_i$	Frecuencia relativa acumulada ( $F_i$ )
15 a 19	17	0,029	0,493	5,569	0,029
20 a 24	22	0,098	2,156	7,689	0,127
25 a 29	27	0,234	6,318	3,483	0,361
30 a 34	32	0,384	12,288	0,501	0,745
35 a 39	37	0,215	7,955	8,111	0,96
40 a 44	42	0,037	1,554	4,593	0,997
45 a 49	47	0,002	0,094	0,521	1
Total		1	Suma=media=30,858	Suma=varianza=30,467	

- a) La media es 30,858 (sumo los productos de cada clase por su frecuencia relativa, que es el porcentaje convertido en proporción)
- b) La varianza es 30,467 (es la suma de los cuadrados de las diferencias entre cada marca de clase y la media, multiplicados por sus frecuencias relativas) y la desviación típica es su raíz cuadrada, es decir 5,52
- c) El coeficiente de variación es la desviación típica partida por la media es decir,  $5,52/30,858=0,179$
- d) La mediana es el valor que tiene una frecuencia relativa acumulada 0,5, que en este caso es la clase de 30 a 34 años, o la marca de clase 32
- e) El primer cuartil es el valor con la frecuencia relativa acumulada 0,25, que en este caso es la clase de 25 a 29 años, o la marca de clase 27  
El tercer cuartil es el valor con la frecuencia relativa acumulada 0,75, que en este caso es el la clase 35 a 39, o la marca de clase 37.
- f) El rango intercuartílico es la diferencia entre el primer y el tercer cuartil, por lo tanto en este caso  $37-27=10$  años
- g) La moda es el valor más frecuente, y en este caso coincide con la mediana, ya que es el valor 30 a 34 (o la marca de clase 32) que tiene la frecuencia relativa mayor (38,4% ó 0,384).

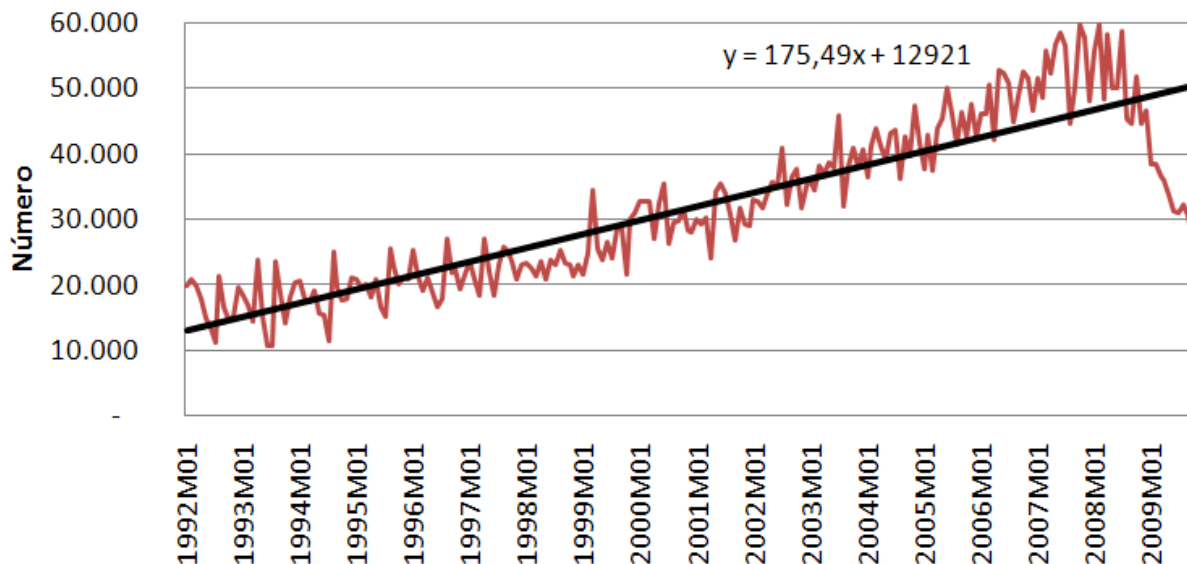
**Pregunta n° 5 (21 puntos)**

El gráfico siguiente representa el número de viviendas terminadas de construir cada mes en España, desde enero de 1992 a octubre de 2009, y la línea de tendencia (en línea recta),

acompañada de la ecuación que expresa la tendencia ( $y = 175,49x + 12921$ ).

- Explica la información que obtienes al ver el gráfico sobre la evolución de la construcción de vivienda en España en el periodo representado
- Teniendo en cuenta que el mes 0 sería diciembre de 1991; ¿cuál sería la predicción de la línea de tendencia para diciembre de 2010? ¿Y para diciembre de 2012?
- ¿Es esperable que esas predicciones se parezcan mucho a la realidad? Explica tu respuesta.

### Viviendas terminadas en España por meses 1992-2009 (oct)



A) El gráfico muestra cómo en los últimos años el número de viviendas construidas cada mes en España ha crecido mucho, desde unos 15.000 al mes, a comienzos de los noventa, hasta unas 55.000 al mes, en la última parte de los años 2000, sufriendo sin embargo, en los dos últimos años, una caída muy brusca, de forma que en poco más de un año habría descendido desde unos 50.000 a menos de 30.000 viviendas terminadas cada mes (entre primeros de 2008 y finales de 2009).

Si nos fijamos algo más en detalle vemos que entre 1992 y 1996 la construcción estuvo más o menos estable alrededor de las 15.000 viviendas mensuales, y es a partir de la segunda fecha cuando se observa un primer “salto”, estabilizándose las viviendas terminadas ligeramente por encima de las 20.000 mensuales, cuando claramente se empiezan a construir más de 20.000 viviendas mensuales. Entre 1999 y 2002 se da un nuevo salto, con valores rondando las 30.000 viviendas mensuales, y es sobre todo entre 2002 y 2008 que se dio un crecimiento firme y continuado que duplicó el número de viviendas terminadas cada mes (de unas 30.000 a unas 60.000 al mes). Después, como se ha indicado arriba, ha habido un cambio muy brusco, con una vuelta, en apenas año y medio, a los valores del año 2002 (unas 30.000 viviendas mensuales).

B) Al ser diciembre de 1991 el mes cero, para diciembre de 2010 habrían transcurrido  $19 \times 12$  meses, es decir, 228 meses.

Por lo tanto, el valor previsto por la ecuación sería

$$y = (175,49 \times 228) + 12921 = 52.933 \text{ viviendas}$$

En diciembre de 2012 habrán transcurrido 24 meses más, es decir, que sería el mes 252. Por lo tanto, el valor predicho por la ecuación será:

$$y = (175,49 \times 252) + 12921 = 57.145 \text{ viviendas}$$

c) Obviamente, es muy poco probable (en el caso de 2010 casi diríamos que seguro) que los valores predichos por la ecuación de la renta de tendencia no se darán en los meses de diciembre de 2010 y 2012. Esto se debe a que la tendencia lineal, por definición, no puede recoger el cambio de tendencia que se ha producido en los dos últimos años. Al calcular una tendencia promedio de un periodo largo, el resultado predicho es el que se derivaría de que las cosas, siguieran siendo, en los próximos meses y años como han sido, en promedio, en los últimos 17. Pero ya sabemos que en los dos últimos años las cosas ya no han sido como en los 15 años anteriores, sino muy, muy diferentes. Por lo tanto, la predicción es errónea, y no nos sirve para imaginar ni siquiera aproximadamente lo que estará sucediendo a finales de 2010 o de 2012.

**Pregunta nº 6 (6 puntos)**

Explica qué es un muestreo sistemático. ¿Es un muestreo probabilístico? ¿En qué circunstancias es seguro usarlo, y en cuáles no se debe hacer?

El muestreo sistemático consiste en dividir el tamaño de la población ( $N$ ) entre el tamaño de la muestra ( $n$ ) y con ello obtener el factor de elevación ( $f=N/n$ ), que es el número de veces que la población contiene a la muestra.

A continuación se obtiene un número aleatorio entre 1 y  $f(x)$ , y a partir de ahí se constituye una muestra que estará formada por los casos que en una lista ordenada y numerada tengan el número de orden  $x, x+f, x+2f, x+3f...$  hasta obtener  $n$  casos.

Es un muestreo probabilístico, porque todos los miembros de la población tienen la misma probabilidad de aparecer en la muestra ( $1/f$ ).

Es seguro utilizarlo si la lista de miembros de la población está ordenada por algún criterio que no tenga nada que ver con los fenómenos estudiados en el estudio (por ejemplo, personas por orden alfabético).

Pero no es seguro si se dan dos condiciones:

A) el orden de los casos puede tener algo que ver con el fenómeno estudiado, de forma cíclica: por ejemplo, si los casos son unidades temporales (horas, días o meses) y están ordenados cronológicamente;

b) el factor de elevación es un múltiplo de un número que contribuye a la formación de esos ciclos con los que están ordenados los casos (por ejemplo el factor de elevación es múltiplo de siete, y las unidades de análisis son días: todos los casos estudiados corresponderán al mismo día de la semana; lo mismo con meses y múltiplos de doce).

**Pregunta nº 7 (10 puntos)**

En una base de datos de los municipios de toda España la media del número de vehículos por 100 habitantes es de 45, con una desviación típica de 8 (datos ficticios). Suponiendo que la distribución de esa variable fuera normal, y utilizando, cuando sea necesario, la tabla de probabilidades de los valores de  $z$  en una distribución normal estándar, que tienes reproducida al final del examen, calcula

- ¿Qué proporción de los municipios tienen más de 50 vehículos por 100 habitantes?
- ¿Qué proporción tiene menos de 40?
- ¿Qué proporción de los municipios tiene entre 45 y 55 vehículos por 100 habitantes?
- ¿Cuál es el primer cuartil de la tasa de vehículos por 100 habitantes de los municipios?

A) Calculamos el valor z que corresponde a 50:

$$(50-45)/8=0,625$$

El valor z es 0,625 (quiere decir que 50 está 0,625 veces la desviación típica por encima de la media).

Redondeando a dos decimales (0,63) en la tabla averiguamos que la proporción de casos con valor z igual o menor que 0,63 es 0,7357 [corregido: por error se había puesto 0,7347]

Por tanto, la proporción de casos con valor z mayor que 0,63 en una distribución normal es  $(1-0,7357)=0,2643$

Es decir, que la proporción de municipios con más de 50 vehículos por 100 habitantes es 0,2643, o el 26,43%

b) Valor z para 40:  $(40-45)/8=-5/8=-0,625$

Es el mismo valor z del ejercicio anterior, pero en negativo

Antes queríamos averiguar cuántos casos tenían un valor z mayor que 0,63 y ahora queremos averiguar cuántos tienen valor z menor que -0,63

Son dos colas de la distribución normal por encima y por debajo de un mismo valor z, en positivo y en negativo, y por tanto, lógicamente, las dos colas son iguales.

Por lo tanto, la proporción es la misma del ejercicio anterior: la proporción de municipios con menos de 40 vehículos por 100 habitantes es 0,2653 o el 26,53%

c) Como 45 es la media, y es una distribución normal, es también la mediana, y por tanto, la mitad de los casos tendrán valores menores (en proporciones: 0,5).

Ahora sólo necesitamos averiguar cuántos casos tienen valores menores o iguales a 55

Procedemos como en a)

$$(55-45)/8=10/8=1,25$$

En la tabla vemos que la proporción de casos por debajo del valor z 1,25 es 0,8944.

Pues entonces, la proporción de casos con valores entre 45 y 55 será  $0,8944-0,5=0,3944$  o bien 39,44%

d) El primer cuartil es el valor z que tenga un 0,25 de los casos con valor inferior o igual. Tiene que ser un valor z negativo, pero en la tabla sólo tenemos valores positivos. Tenemos que buscar entonces el valor z positivo tal que el 0,25 de los casos tengan un valor z superior (es decir, el valor z con la frecuencia acumulada 0,75).

Ese valor, según la tabla, es 0,67 (es el valor con la frecuencia acumulada más cercana a 0,75).

Por lo tanto, el valor z negativo con la frecuencia relativa acumulada 0,25 (el primer cuartil) será -0,67

Ya sólo tenemos que convertir el valor z en valores de la variable original (vehículos por 100 habitantes) y nos da

$$+45+(-0,67*8)=45-5,36=39,64$$

### Pregunta nº 8 (12 puntos)

En una encuesta realizada en diciembre por el CIS a una muestra aleatoria de 2.489 españoles, un 77,1% de los encuestados dijo que iba a jugar a la lotería de Navidad. Calcula:

a) El error típico de la encuesta

b) El intervalo de confianza, con un nivel de confianza del 95,5%, para la proporción de españoles que tenían intención de jugar a la lotería de Navidad.

A) El error típico de una encuesta, al tratarse de porcentajes, se calcula con la fórmula

$ET = \text{Raiz}(p \cdot q/n)$

En este caso

$\text{raiz}(0,771 \cdot 0,228/2489) = \text{raiz}(0,176/2489) = \text{raiz}(0,0000707) = 0,0084$

El error típico es 0,0084 (en proporciones) ó 0,84%

b) Con un 95,5% de confianza, el intervalo de confianza se forma por el estimado más/menos dos veces el error típico.

Por tanto, en este caso:

$77,1\% \pm 1,68\%$

o bien

$0,771 \pm 0,0168$

En definitiva el intervalo de confianza va de 75,42% a 78,78% (o, en proporciones, de 0,7542 a 0,7878)

### Pregunta nº 9 (12 puntos)

Habitualmente se suele decir que el 10% de los varones son homosexuales. Sin embargo, en la reciente Encuesta Nacional de Salud Sexual 2009, sólo un 2,8% de los 4.900 hombres entrevistados declararon que se sienten exclusivamente o habitualmente atraídos por personas de su mismo sexo.

a) Realiza un contraste de hipótesis para ver si es posible decir que, con la encuesta en la mano, podríamos rechazar la hipótesis de que el 10% de los varones son homosexuales. (8 puntos)

b) Explica por qué razones relacionadas con las características del método de encuesta es posible dudar de las conclusiones del estudio en este aspecto (4 puntos).

A)

Se trata de hacer un contraste de hipótesis en el que la hipótesis nula ( $H_0$ ) es que  $p=0,1$ , siendo  $p$  la proporción de varones homosexuales.

En principio no tenemos ninguna hipótesis previa sobre si la proporción, en caso de ser distinta, es mayor o menor, así que la hipótesis alternativa ( $H_1$ ) es que  $p \neq 0,1$ , y hacemos un contraste de hipótesis bilateral.

Calculamos el valor  $z$  de la proporción estimada, usando la fórmula:

$$z = \frac{\text{estimador} - \theta_0}{ET} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0,028 - 0,1}{\sqrt{\frac{0,1 \cdot 0,9}{4900}}} = \frac{-0,072}{\sqrt{\frac{0,09}{4900}}} = \frac{-0,072}{\sqrt{0,0000184}} =$$
$$= \frac{-0,072}{0,00429} = -15,78$$

Normalmente, el paso siguiente, sería comprobar en la tabla cuál es la probabilidad de un valor  $z = -15,78$  o dicho de otra manera, ver si el valor  $z$  está dentro de la región de rechazo para un contraste bilateral.

Pero como nos ha salido un valor  $z$  tan alto, no está en las tablas (con un ordenador podría calcularse la probabilidad de que el valor  $z$  mayor o igual a 15,78 o menor igual a -15,78; al ser

un contraste bilateral sumaríamos esas dos probabilidades).

Dicho de otra forma, si la región de rechazo, en un contraste bilateral, la componen los valores de  $z$  mayores que 1,96, o menores -1,96 y hemos obtenido un valor de -15,78, claramente nuestro valor está en la región de rechazo.

Es un valor altísimo y por tanto no tenemos ninguna duda de que con seguridad, con esos datos de encuesta, podríamos rechazar con total seguridad la hipótesis nula de que el porcentaje de homosexuales entre los varones adultos es del 10%.

B) Podríamos dudar de las conclusiones del punto anterior, atendiendo a que son datos de encuesta, y que como sabemos el método de encuesta tiene algunas dificultades relacionadas con que requieren la participación voluntaria de las personas y la veracidad de sus respuestas. Dado el tema tan delicado del que estamos tratando, tan rodeado de estereotipos, tabús, etc... es posible que podamos pensar que no todas las personas se han mostrado por igual dispuestas a participar en la encuesta, y que luego, entre las que lo han hecho, las respuestas a preguntas sobre la orientación sexual pueden no ser del todo sinceras. Por esa razón, tal vez quepa pensar que el porcentaje del 2,8% que ha salido en la encuesta podría ser, tal vez, una infraestimación del verdadero porcentaje de personas que tienen una orientación sexual homosexual. Por lo tanto, la conclusión del punto anterior ha de ser tomada con cautela, por las peculiaridades del tema de que se trata que hace que dudemos más en este caso del método de encuesta que en otras circunstancias.

**Áreas bajo la curva normal estándar. Los valores de la tabla que no se muestran en negrita representan la probabilidad de observar un valor menor o igual al valor correspondiente de z**

<b>Segunda cifra decimal del valor de z</b>										
<b>z</b>	<b>0.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>
<b>0.0</b>	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
<b>0.1</b>	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
<b>0.2</b>	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
<b>0.3</b>	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
<b>0.4</b>	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
<b>0.5</b>	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
<b>0.6</b>	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
<b>0.7</b>	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
<b>0.8</b>	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
<b>0.9</b>	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
<b>1.0</b>	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
<b>1.1</b>	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
<b>1.2</b>	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
<b>1.3</b>	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
<b>1.4</b>	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
<b>1.5</b>	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
<b>1.6</b>	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
<b>1.7</b>	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
<b>1.8</b>	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
<b>1.9</b>	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
<b>2.0</b>	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
<b>2.1</b>	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
<b>2.2</b>	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
<b>2.3</b>	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
<b>2.4</b>	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
<b>2.5</b>	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
<b>2.6</b>	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
<b>2.7</b>	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
<b>2.8</b>	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
<b>2.9</b>	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
<b>3.0</b>	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
<b>3.1</b>	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
<b>3.2</b>	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
<b>3.3</b>	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
<b>3.4</b>	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998