Summer Science Camp, Tjärnö, 8 August 2012

Varför statistik?

Serik Sagitov

http://www.math.chalmers.se/∼serik/

Avdelningen

för matematisk statistik

Matematiska Vetenskaper

Chalmers Tekniska Högskola och Göteborgs Universitet

DET FINNS INGA DUMMA FRÅGOR, BARA DUMMA SVAR!

Enligt www.scb.se ordet statistik har två betydelser.

- Statistik beskriver verkligheten. Statistik är sifferuppgifter som beskriver en sak eller en verksamhet. (Ex: med en yta på 2.7 millioner kv. km. är Kazakhstan världens 9-de största land.)
- Statistik i praktiken. Statistik är också metoder för att samla in, bearbeta och analysera material.

> Det finns tre typer av lögn: lögner, förbannade lögner och statistik". Benjamin Disraeli

Matematiken bakom statistiska metoder

- Matematisk statistik
- Sannolikhetsteori

Temat för SSC 2012

Vi kommer att studera fastsittande organismer, både alger och evertebrater och forma studier kring:

- Är det skillnad i förmågan att sitta fast hos individer av samma art som lever på både skyddade och exponerade områden i havet?
- Hur ser det ut i fält?
- Hur viktig är formen för fastsittande, vid ytan och under vattnet?
- Är bottenbeskaffenheten viktig för fastsittande?

Vi kommer att forma flera frågeställningar runt detta tema som innefattar både biologi och fysik. Vi kommer också att ha Bitr professor Serik Sagitov till vår hjälp med statistik.

Exempel. Man vill jämföra vilopuls hos män och kvinnor

X = vilopuls hos kvinnor
Y = vilopuls hos män

$X > Y$ eller $X < Y$ eller $X = Y$? Men vad är egentligen X och Y?

Gunde Svan hade 32bpm i vilopuls när han var aktiv idrottare.

Variation i vilopuls mellan människor orsakad av en mängd faktorer

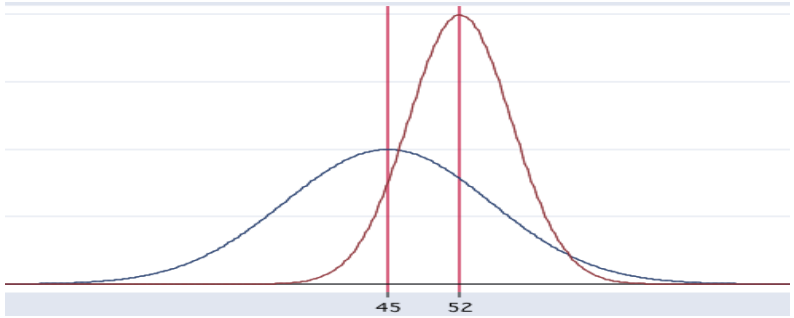
- ålder
- hur vältränad man är
- kroppsvikt
- frisk eller har infektion
- medfödda faktorer
- man eller kvinna

Populations medelvärde

X representerar en population av kvinnor och Y representerar en population av män

X och Y är *stokastiska variabler*.

Två populationsfördelningar skall jämföras.



Istället för att jämföra kurvor tittar man på två populationsmedelvärden

$$\mu_1 = E[X], \quad \mu_2 = E[Y].$$

Tolkning av populationsmedelvärden

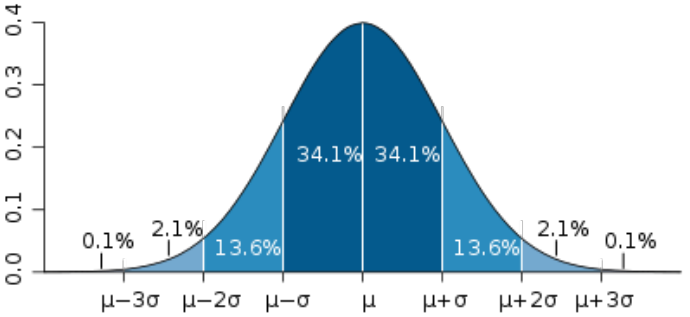
Väntevärde = Expectation

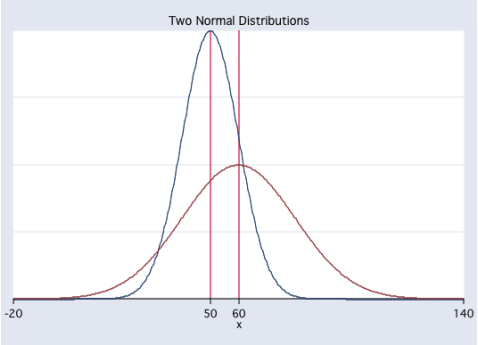
μ_1 och μ_2 är vilopuls hos kvinnor och män efter alla yttre faktorer räknades bort.

Populations standardavvikelse

Variationen inom populationerna beskrivs med standard avvikelser σ_1, σ_2 som definieras genom varianser:

$$\sigma_1^2 = E\left[(X - \mu_1)^2\right],$$
$$\sigma_2^2 = E\left[(Y - \mu_2)^2\right].$$





Om de olika värdena ligger samlade nära medelvärdet blir standardavvikelsen låg, medan värden som är spridda långt över och under medelvärdet ger en hög standardavvikelse.

Två kurvorna på bilden har $\mu_1 < \mu_2$ samt $\sigma_1 < \sigma_2$.

Om variationen saknas då $\sigma = 0$.

Normalfördelningen är en viktig fördelning inom statistik:

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

Sannolikhetsteori: Centrala gränsvärdesatsen.

Normalfördelningstabell

z	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

Två huvud problemställningar

- Uppskatta populations parametrar μ_1, μ_2.
  Beräkna ett konfidensinterval för skillnaden $(\mu_1 - \mu_2)$.
- Testa nollhypotesen $H_0 : \mu_1 = \mu_2$ mot en av alternativa hypoteser $H_1 : \mu_1 \neq \mu_2$, eller $H_1 : \mu_1 < \mu_2$, eller $H_1 : \mu_1 > \mu_2$.

Exempel. Stickprovet är 10 slumpvis valda kvinnor och lika många slumpvis valda män. I stickprovet har kvinnor i snitt 3 bpm högre vilopuls än män.

Två möjliga förklaringar:

- Slumpen har gjort att vi har hittat en skillnad på 3 bpm även om det inte finns någon skillnad mellan μ_1, μ_2.
- Nollhypotesen stämmer inte.

Stickprovs medelvärde

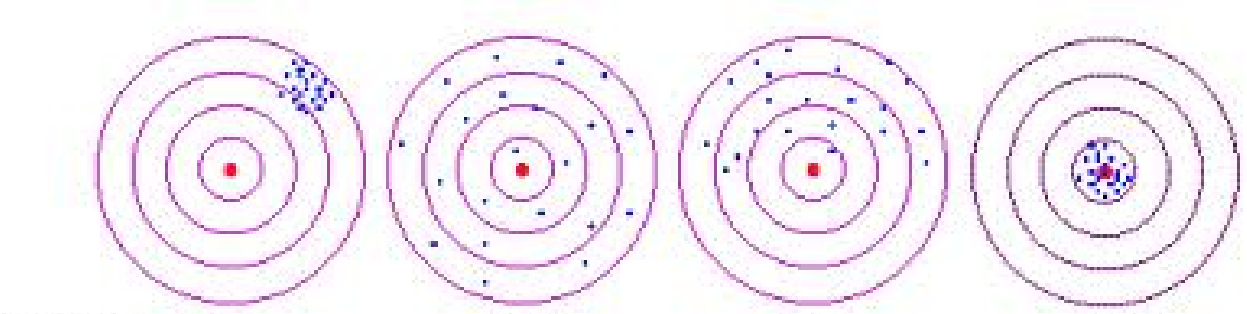
Man vill generalisera från ett stickprov till en population. Eftersom det är praktiskt omöjligt att mäta varje individ i populationen.

Stickprovsmätningar $(X_1, \ldots, X_n)$ och stickprovsmedelvärden

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

Slumpmässig urval: randomisera bort externa faktorer.

Vid slumpmässig urval $\bar{X}$ är en *väntevärdesriktig punktskattning* av μ. Skattningen $\bar{X}$ innehåller inget systematiskt fel bara slumpfel.



Systematiskt fel och slumpfel

Stickprovsvarians $s^2 = \frac{(X_1-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n-1}$

Vid slumpmässig urval punktskattningen s^2 av σ^2 är väntevärdesriktig.

Medelfel av punktskattningen $\bar{X}$:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

var s är stickprovs standardavvikelse beräknad genom stickprovsvariansen.

Stickprovsstorlek n avgör hur stort utrymme får slumpfaktorer:

- ju större stickprov desto mindre påverkan,
- dock bör man inte misstro ett ganska litet men korrekt draget stickprov.

Konfidensintervall för medelvärde

	Populations...	Stickprovs...
...medelvärde	μ	$\bar{X}$
...standardavvikelse	σ	s
...varians	σ^2	s^2

Medelfel : $s_{\bar{X}} = \frac{s}{\sqrt{n}}$

Approximativt 95% konfidensintervall för μ är $\bar{X} \pm 1.96 \cdot s_{\bar{X}}$.

Givet X är normalfördelad, det exakta 95% KI för μ är $\bar{X} \pm t_{n-1} \cdot s_{\bar{X}}$

antal frihetsgrader k	3	5	9	14	18	200
koefficient t_k	3.182	2.571	2.262	2.145	2.101	1.972

T-FÖRDELNINGSTABELL

One Tail	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	
Two Tails	0.20	0.10	0.05	0.02	0.01	0.002	0.001	
D 1	3.078	6.314	12.71	31.82	63.66	318.3	637	1
E 2	1.886	2.920	4.303	6.965	9.925	22.330	31.6	2
G 3	1.638	2.353	3.182	4.541	5.841	10.210	12.92	3
R 4	1.533	2.132	2.776	3.747	4.604	7.173	8.610	4
E 5	1.476	2.015	2.571	3.365	4.032	5.893	6.869	5
E 6	1.440	1.943	2.447	3.143	3.707	5.208	5.959	6
S 7	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041	8
O 9	1.383	1.833	2.262	2.821	3.250	4.297	4.781	9
F 10	1.372	1.812	2.228	2.764	3.169	4.144	4.587	10
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437	11
F 12	1.356	1.782	2.179	2.681	3.055	3.930	4.318	12
R 13	1.350	1.771	2.160	2.650	3.012	3.852	4.221	13
E 14	1.345	1.761	2.145	2.624	2.977	3.787	4.140	14
E 15	1.341	1.753	2.131	2.602	2.947	3.733	4.073	15
D 16	1.337	1.746	2.120	2.583	2.921	3.686	4.015	16
O 17	1.333	1.740	2.110	2.567	2.898	3.646	3.965	17
M 18	1.330	1.734	2.101	2.552	2.878	3.610	3.922	18
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883	19
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850	20
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819	21
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792	22
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768	23
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745	24
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725	25
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707	26
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690	27
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674	28
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659	29
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646	30
32	1.309	1.694	2.037	2.449	2.738	3.365	3.622	32

Konfidensintervall för skillnaden

Antaganden

1. två OBEROENDE slumpmässiga stickprov
   $(X_1, \ldots, X_n), \quad (Y_1, \ldots, Y_m)$
2. både X och Y är normalfördelade möjligen med olika μ_1, μ_2
3. dock med samma standardavvikelse $\sigma_1 = \sigma_2 = \sigma$

Poolad sticksprovs varians $s^2 = \frac{n-1}{n+m-2} s_x^2 + \frac{m-1}{n+m-2} s_y^2$

Två stickprovsvarianser

$$s_x^2 = \frac{(X_1-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n-1}, \quad s_y^2 = \frac{(Y_1-\bar{Y})^2+\ldots+(Y_m-\bar{Y})^2}{m-1}.$$

Exakt KI för $(\mu_1 - \mu_2)$ är $(\bar{X} - \bar{Y}) \pm t_{m+n-2} \cdot s \cdot \frac{n+m}{nm}$

Exempel. Stickprovet är 10 slumpvis valda kvinnor och lika många slumpvis valda män.

I stickprovet har kvinnor i snitt $\bar{X} - \bar{Y} = 3$ bpm högre vilopuls än män.

Två stickprovs standard avvikelser $s_x = 10$, $s_y = 9$,

Poolad sticksprovs varians

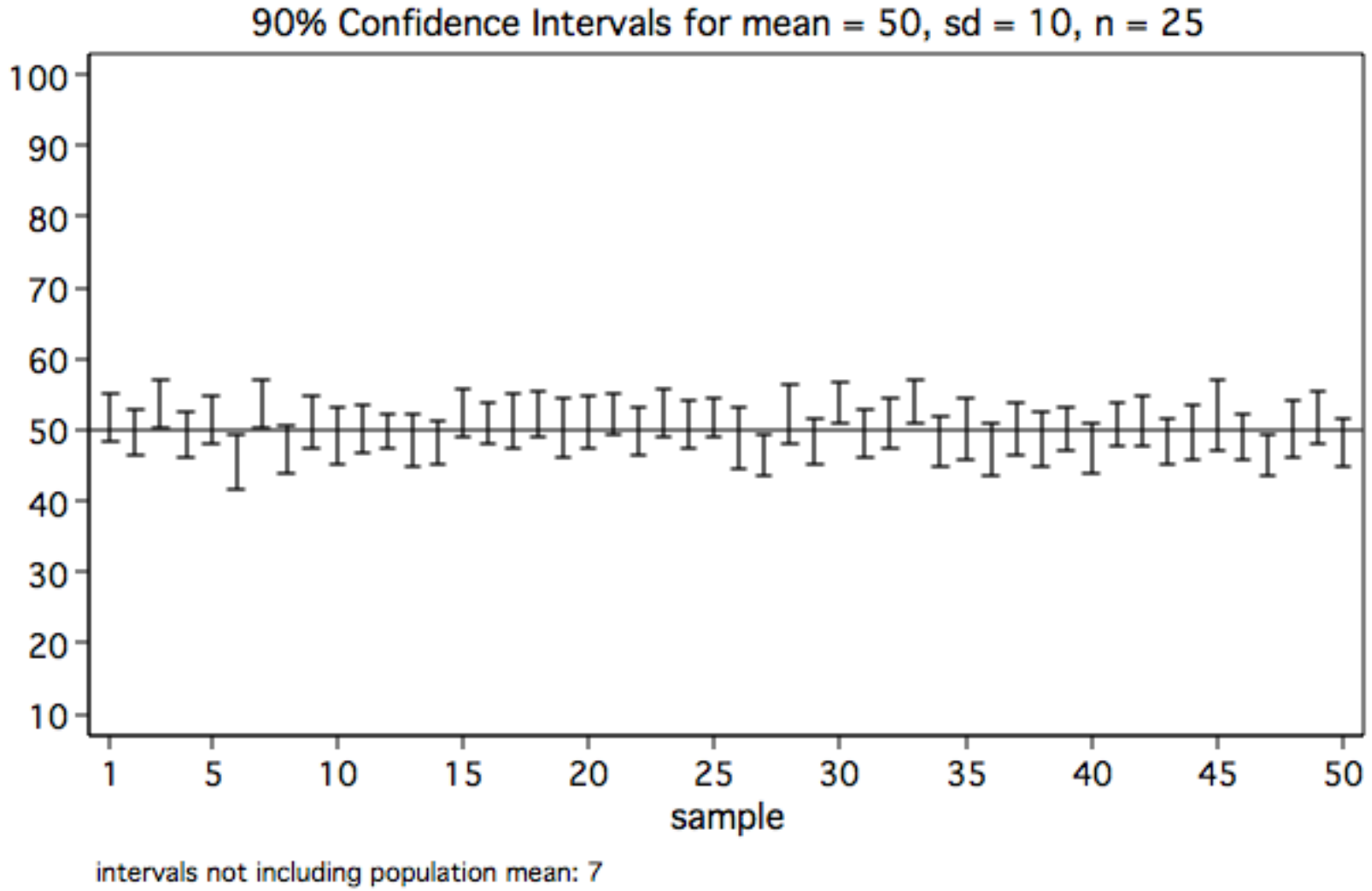
$$s^2 = \frac{n-1}{n+m-2} s_x^2 + \frac{m-1}{n+m-2} s_y^2 = \frac{10^2 + 9^2}{2} = 90.5 = (9.51)^2.$$

95% konfidensintervall för $(\mu_1 - \mu_2)$ är

$$(\bar{X} - \bar{Y}) \pm t_{m+n-2} \cdot s \cdot \frac{n+m}{nm} = 3 \pm 2.101 \cdot 9.51 \cdot \frac{20}{100} = 3 \pm 4.$$

Konfidensintervallet innehåller 0, vi kan inte förkasta nollhypotesen.

Konfidensintervalls mening



Länkar

http://www.topendsports.com/testing/heart-rate-resting-chart.htm

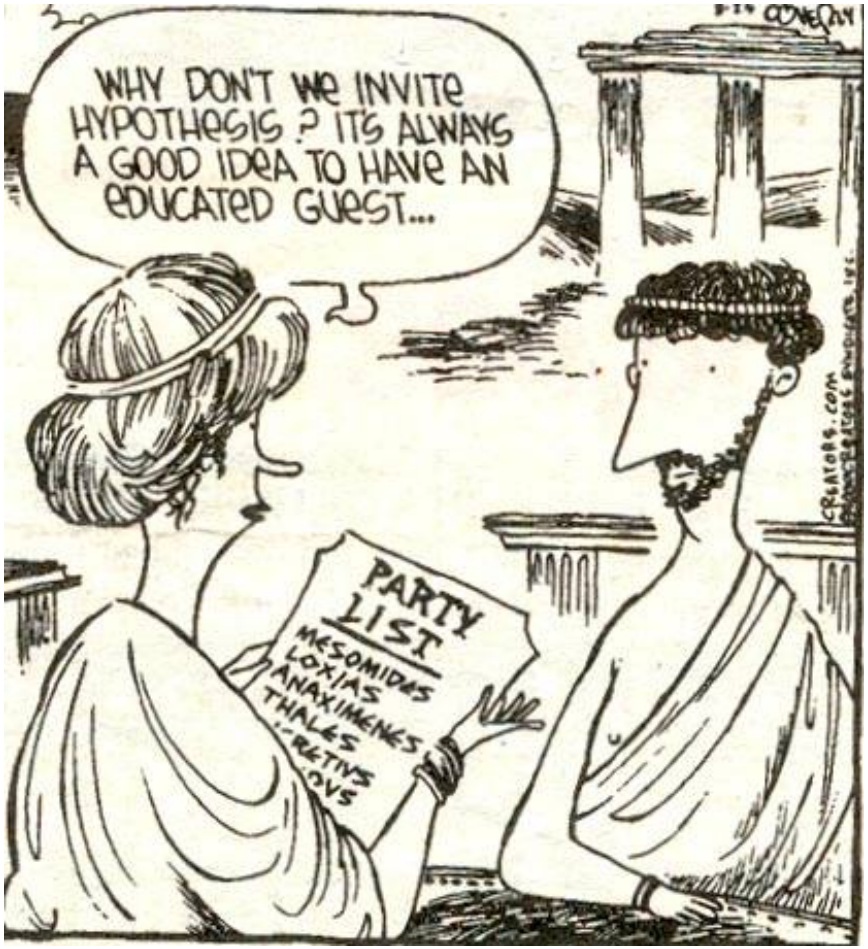
http://bookboon.com/se/studentlitteratur/statistik/

http://www.scb.se/

http://www.hh.se/download/18.23c46aaf13349ab6d418000865/

"If you're not prepared to be wrong, you'll never come up with anything original.
"http://www.youtube.com/watch?v=iG9CE55wbtY&feature=plcp

Tack för mig!



Educated guess = Kvalificerad gissning