Föreläsning 13: Repetition och diverse kommentarer

Med illustrationer från xkcd.com

David Bolin
Chalmers University of Technology
Towel day, 2015

Projektet

- Projektet ska lämnas in senast måndag 2015-06-01, kl 23:59.
- Lämna in rapporten som en PDF (inga MS Word-filer!).
- Inkludera också koden som ni använde för att utföra beräkningarna.
- Se projektbeskrivningen för bedömningskriterier samt för anvisningar till hur rapporten ska skrivas.
- I projektbeskrivningen finns en checklista för granskning av rapporter. Gå igenom och bocka av listan med frågor innan rapporten lämnas in!

Utfall och utfallsrum

slumpmässigt försök

Man brukar säga att ett slumpmässigt försök är ett försök som kan upprepas under väsentligen identiska förhållanden där utfallet av försöket inte exakt kan förutsägas i det enskilda fallet.

Utfall och utfallsrum

Resultatet av ett försök kallas för ett *utfall* ω, och mängden av alla möjliga utfall kallas för *utfallsrummet* Ω.

Händelser

Man kan också sätta samman olika utfall till *händelser*. En händelse A är en mängd av utfall, det vill säga en delmängd av utfallsrummet Ω.

Snitt, union och komplement

Låt A och B vara två händelser, vi definierar

Komplement, A^c

Mängden av alla utfall som inte finns i A. $A^c = \Omega \setminus A$.

Union, $A \cup B$

Mängden av alla utfall i A eller B.

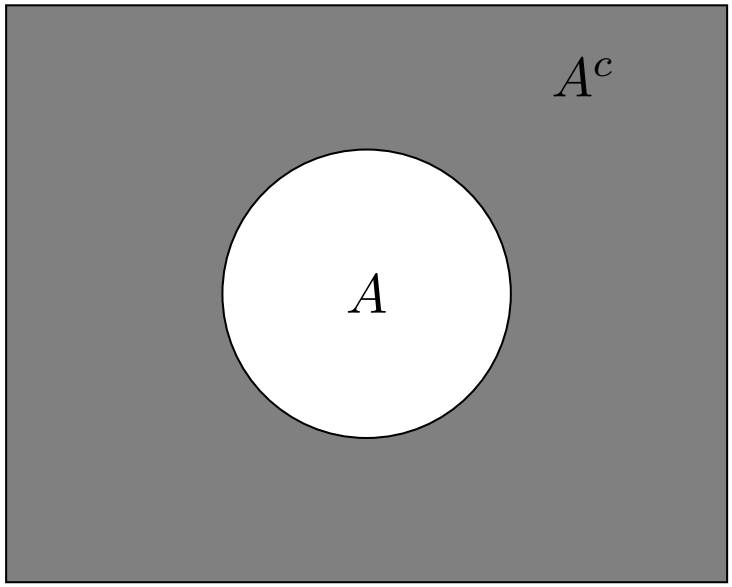
Snitt $A \cap B$

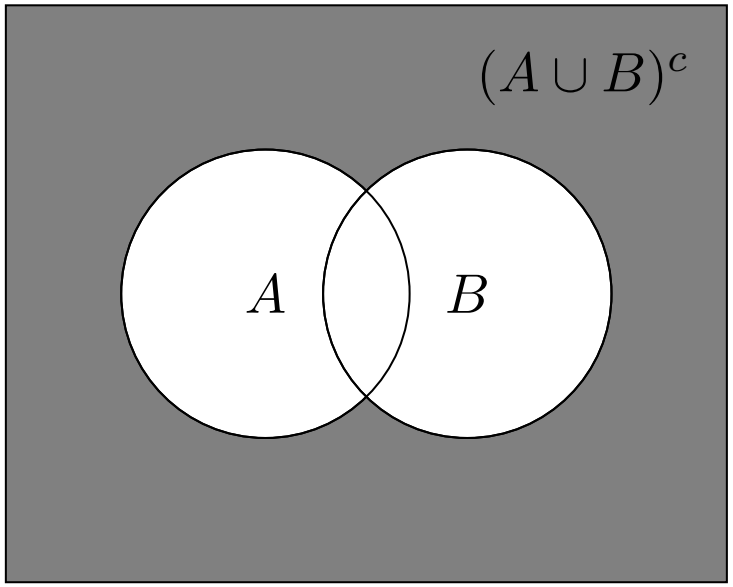
Mängden av alla utfall som finns i A och B.

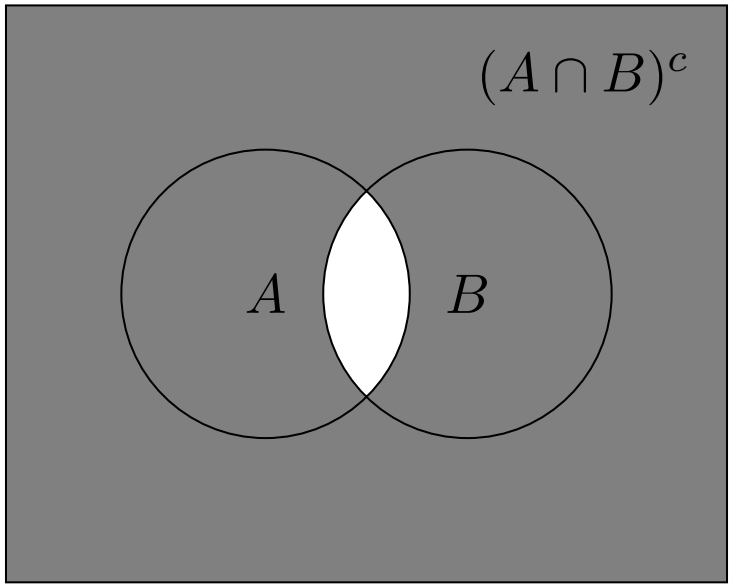
Den tomma mängden $\emptyset$

Brukar kallas för en omöjlig händelse. Om $A \cap B = \emptyset$ så säger vi att A och B är oförenliga, disjunkta, eller utesluter varandra.

Venndiagram







Kolmogorovs axiomsystem

Kolmogorovs axiomsystem

Låt Ω vara ett utfallsrum. En sannolikhet, eller sannolikhetsmått $\mathsf{P} : \Omega \to \mathbb{R}$ är en reell funktion som tar händelser i Ω som argument och returnerar ett reellt tal, och som uppfyller

1. $0 \leq \mathsf{P}(A) \leq 1$.
2. $\mathsf{P}(\Omega) = 1$.
3. Om $A_1, A_2, \ldots$ är än följd av disjunkta händelser så gäller att

$$\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i).$$

Speciellt så gäller att om händelserna A och B är disjunkta så är

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B).$$

Egenskaper hos sannolikhetsmåttet

Ur axiomen kan vi härleda följande egenskaper.

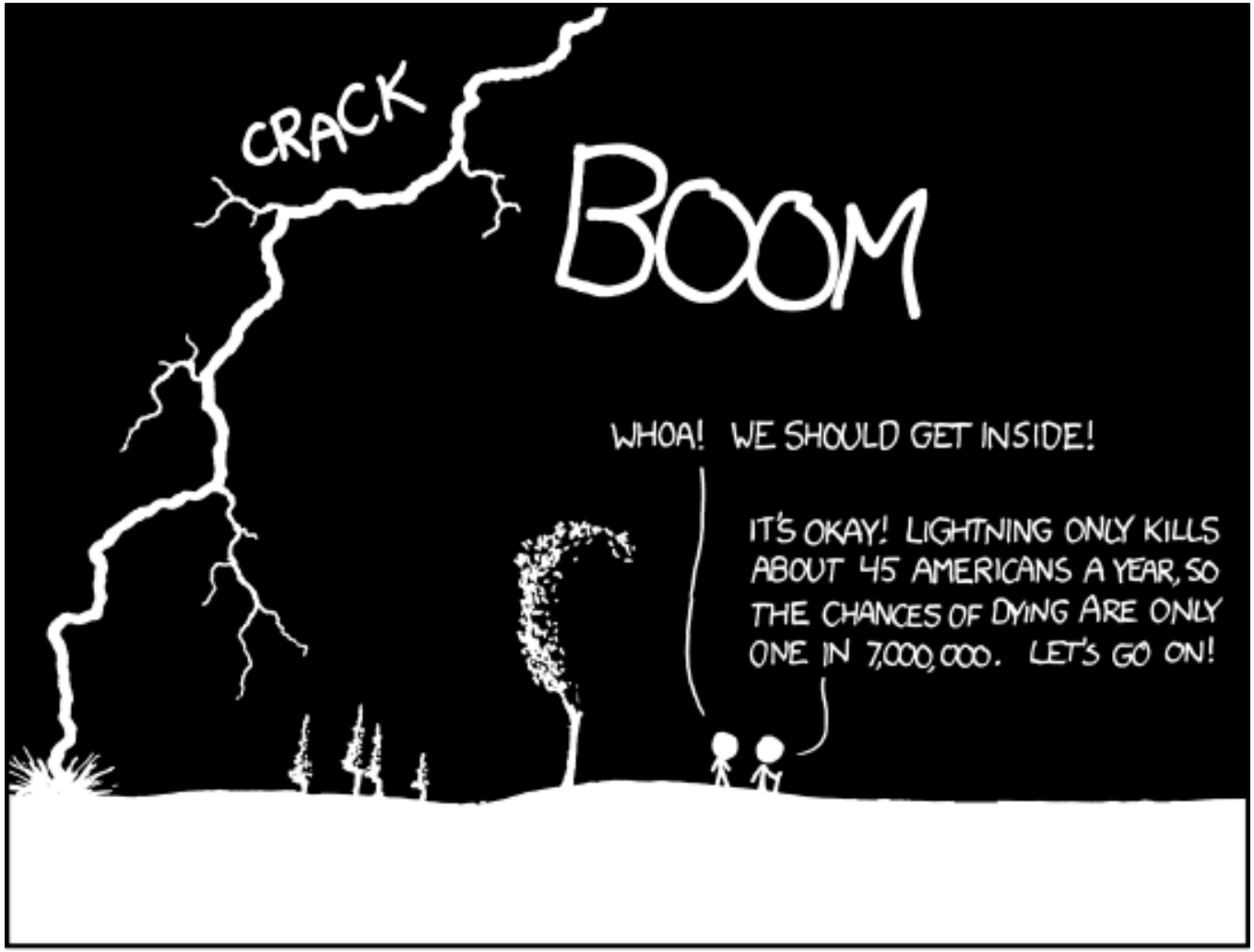
Egenskaper

För ett sannolihetsmått P gäller att

1. $\mathsf{P}(\emptyset) = 0$.
2. $\mathsf{P}(A^c) = 1 - \mathsf{P}(A)$.
3. $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$.

Att dessa är uppfyllda inses enkelt genom att rita Venndiagram!

Betingad sannolikhet



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Betingade sannolikheter

Vi vill ibland beräkna sannolikheten för en händelse A givet att vi vet att en annan händelse B har inträffat. Vi vill då veta den så kallade betingade sannolikheten för A givet B.

Betingad sannolikhet

Antag att $\mathsf{P}(B) > 0$. Den betingade sannolikheten för A givet B definieras som

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}.$$

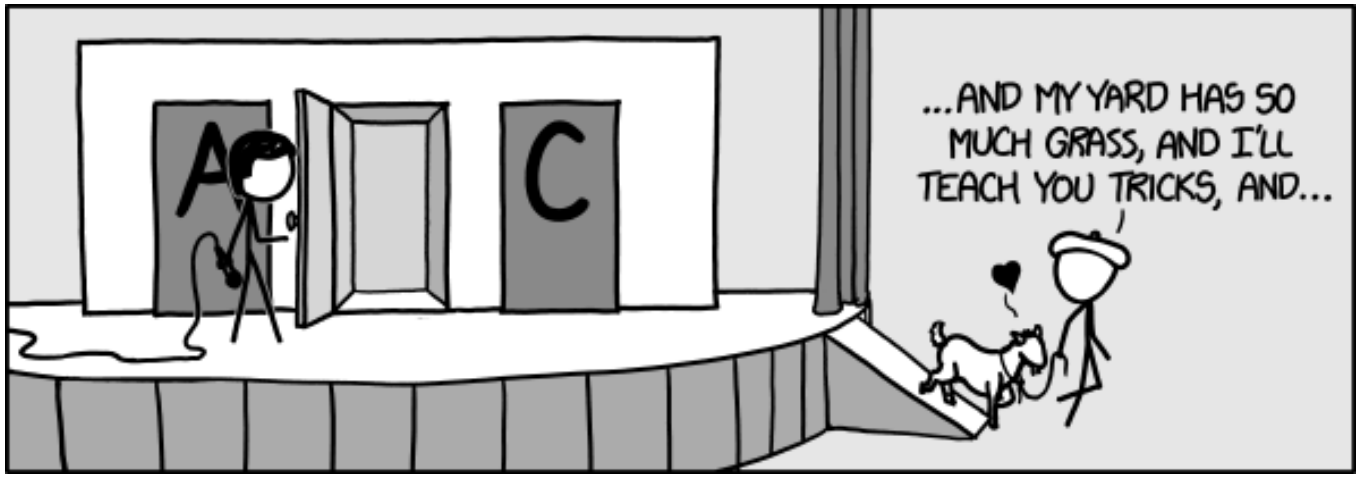
Multiplikationssatsen för sannolikheter

Om A och B är händelser så gäller att

$$\mathsf{P}(A \cap B) = \mathsf{P}(B|A)\mathsf{P}(A) = \mathsf{P}(A|B)\mathsf{P}(B).$$

Multiplikationssatsen är speciellt användbar för att beräkna sannolikheten för successiva händelser som påverkar varandra.

Bayes sats



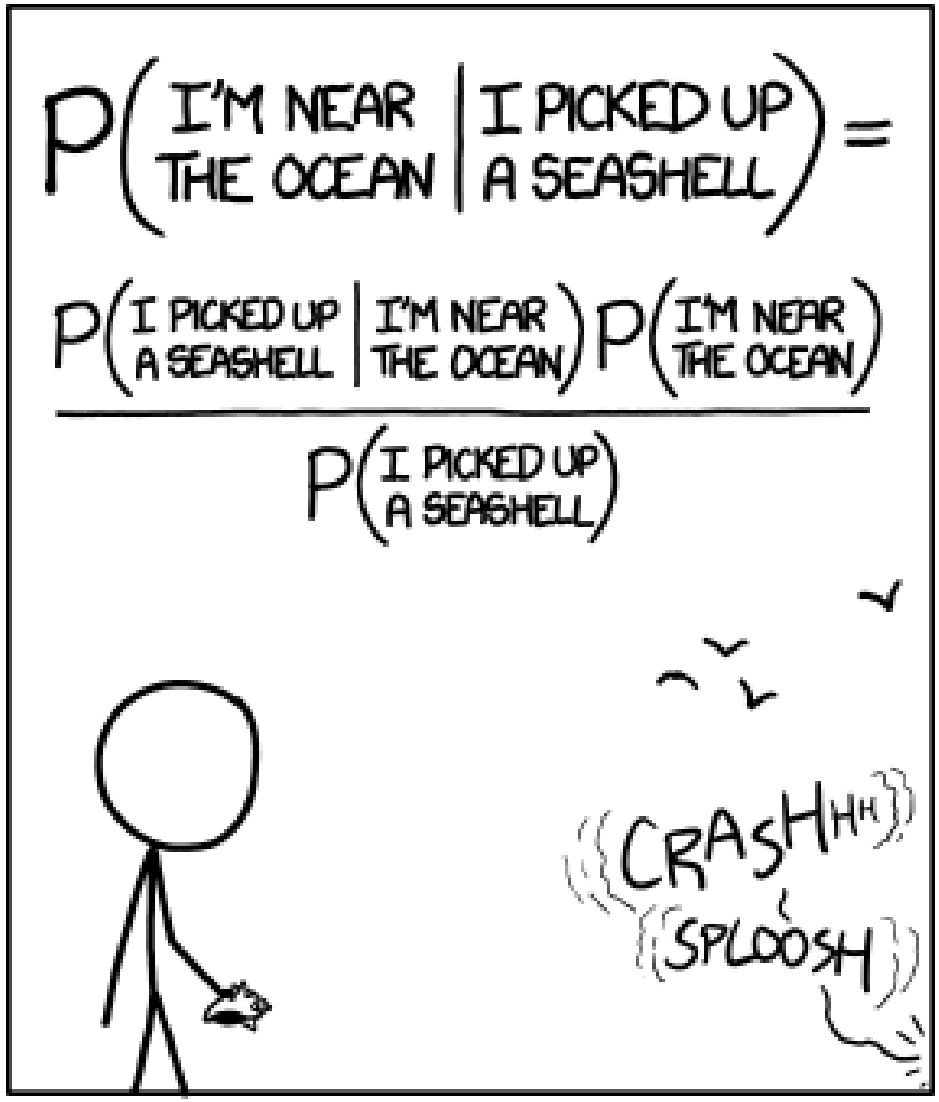
Bayes sats

För två händelser A och B gäller att

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(B|A)\mathsf{P}(A)}{\mathsf{P}(B)}$$

Man har ofta nytta av att skriva om nämnaren $\mathsf{P}(B)$ som

$$\mathsf{P}(B) = \mathsf{P}(B|A)\mathsf{P}(A) + \mathsf{P}(B|A^c)\mathsf{P}(A^c)$$



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

Slumpvariabler

En slumpvariabel beskriver utfallet av ett slumpmässigt försök med ett numeriskt värde

Slumpvariabel

En stokastisk variabel (slumpvariabel) X är en funktion som tar element från Ω som argument och som returnerar ett reellt tal.

Vi betecknar ofta slumpvariabler med stora bokstäver, X, Y och Z. Utfall betecknas ofta med respektive små bokstäver, x, y och z.

Diskreta slumpvariabler

En diskret slumpvariabel kan endast anta ett ändligt eller uppräkneligt antal värden, i allmänhet någon delmängd av heltalen.

Sannolikhetsfunktionen

Sannolikhetsfunktion

Till en diskret stokastisk variabel X definierar vi sannolikhetsfunktionen $p(k)$ genom $p(k) = \mathrm{P}(X = k)$.

Eftersom $p(k)$ beskriver sannolikheter måste vi ha att

- $p(k) \geq 0$ för alla k.
- $\sum_{k=-\infty}^{\infty} p(k) = 1$.

Dessa två villkor är både nödvändiga och tillräckliga för att en funktion $p(k)$ ska vara en sannolikhetsfunktion.

Vi kan också visa att

$$\mathrm{P}(m \leq X \leq n) = \sum_{k=m}^{n} p(k)$$

om m och n är heltal.

Kontinuerliga fördelningar

I praktiken stöter vi ofta på situationer då de möjliga värdena är alla värden i ett intervall, tex $[0, 1]$, eller är alla positiva reella tal.

Kontinuerlig slumpvariabel

En kontinuerlig slumpvariabel kan anta alla värden i något (eller några) intervall av reella tal och sannolikheten för att den antar varje specifikt värde är noll.

Kontinuerliga slumpvariabler kan beskrivas med täthetsfunktionen.

Täthetsfunktion

Låt X vara en kontinuerlig slumpvariabel, en funktion $f(x)$ så att

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)dx = 1, \quad \mathsf{P}(a \leq X \leq b) = \int_{a}^{b} f(x)dx,$$

kallas en täthetsfunktion.

Fördelningsfunktionen

Ett annat vanligt sätt att beskriva en stokastisk variabel är att använda den så kallade fördelningsfunktionen.

Fördelningsfunktionen

Låt X vara en slumpvariabel. Dess fördelningsfunktion ges då av $F(x) = \mathsf{P}(X \leq x)$. Om X är en diskret variabel har vi

$$F(x) = \sum_{k \leq x} p(k),$$

och om X är kontinuerlig har vi

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

Väntevärde

Väntevärdet är ”tyngdpunkten” i fördelningen.

Väntevärde

Väntevärdet för en diskret fördelning definieras som

$$\mathsf{E}(X) = \sum_{k=-\infty}^{\infty} kp(k).$$

På motsvarande sätt definieras väntevärdet för en kontinuerlig slumpvariabel som

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Varians och standardavvikelse

Varians

Variansen av en stokastisk variabel definieras som $\mathsf{V}(X) = \mathsf{E}[(X-\mu)^2]$, där μ är väntevärdet av X.

Vi ser alltså att variansen definieras som väntevärdet av den kvadratavvikelsen av X från dess väntevärde. Vi beräknar variansen som

$$\mathsf{V}(X) = \begin{cases} \sum_{k=-\infty}^{\infty}(k-\mu)^2 p_X(k), & \text{om } X \text{ är diskret} \\ \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx, & \text{om } X \text{ är kontinuerlig.} \end{cases}$$

Ett enklare sätt är ofta att beräkna variansen som

$$\mathsf{V}(X) = \mathsf{E}(X^2) - \mu^2$$

Standardavvikelsen av en stokastisk variabel ges av $\sigma = \sqrt{\mathsf{V}(X)}$.

Normalfördelningen

Normalfördelningen

En kontinuerlig slumpvariabel X är normalfördelad, $\mathrm{N}(\mu, \sigma^2)$, med parametrar $\mu \in \mathbb{R}$ och $\sigma > 0$, om den har täthetsfunktionen

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

Fördelningsfunktionen ges av

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) dy$$

Parametrar

Om $X \sim \mathrm{N}(\mu, \sigma^2)$ har vi att $\mathrm{E}(X) = \mu$ och $\mathrm{V}(X) = \sigma^2$.

Beräkning av sannolikheter

Standardiserad normalfördelning

En slumpvariabel Z sägs ha en standardiserad normalfördelning om $Z \sim \mathsf{N}(0,1)$. Vi beteckningar fördelningsfunktion och täthetsfunktion för denna fördelning med $\varphi(x)$ respektive $\Phi(x)$.

Sats

Om $X \sim \mathsf{N}(\mu, \sigma^2)$ så gäller att $aX + b \sim \mathsf{N}(a\mu + b, a^2\sigma^2)$.

Detta betyder att om $X \sim \mathsf{N}(\mu, \sigma^2)$

- kan vi skriva $X = \mu + \sigma Z$ där $Z \sim \mathsf{N}(0,1)$.
- har vi att $Z = (X - \mu)/\sigma \sim \mathsf{N}(0,1)$.

Vi kan använda detta för att beräkna sannolikheter:

$$\mathsf{P}(X < x) = \mathsf{P}(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}) = \mathsf{P}(Z < \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma}).$$

Centrala gränsvärdessatsen

En av de viktigaste satserna inom statistikteorin, och en av anledningarna till varför normalfördelningen är så viktig.

CGS

Låt $X_1, \ldots, X_n$ vara oberoende och likafördelade slumpvariabler med väntevärde μ och varians $\sigma^2 < \infty$. Då gäller att

$$\mathsf{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x), \quad \text{då } n \to \infty.$$

Om n är stort har vi enligt satsen att

- $\sum_{i=1}^n X_i$ är approximativt $\mathsf{N}(n\mu, n\sigma^2)$-fördelad.
- $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ är approximativt $\mathsf{N}(\mu, \sigma^2/n)$-fördelad.

Hur stor n måste vara beror på fördelningen av X_i.

Tvådimensionella fördelningar

Definition

En två dimensionell slumpvariabel (X, Y) tillordnar två numeriska värden till varje utfall i Ω.

För diskreta variabler har vi sannolikhetsfunktionen

$$p_{X,Y}(i, j) = \mathsf{P}(X = i, Y = j).$$

För en sannolikhetsfunktion har vi att $p_{X,Y}(i, j) \geq 0$ och $\sum_{i,j} p_{X,Y}(i, j) = 1$. För kontinuerliga variabler har vi en täthetsfunktion $f_{X,Y}(x, y)$ som är sådan att

1. $f_{X,Y}(x, y) \geq 0$,
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$, och
3. $\mathsf{P}(a \leq X \leq b \text{ och } c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dx dy$.

Marginalfördelningar

Givet en diskret tvådimensionell slumpvariabel (X, Y) definieras marginalfördelningarna för X och Y som

$$p_X(i) = \sum_{j=-\infty}^{\infty} p_{X,Y}(i,j), \qquad p_Y(j) = \sum_{i=-\infty}^{\infty} p_{X,Y}(i,j)$$

I det kontinuerliga fallet har vi på motsvarande sätt

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Väntevärden av funktioner

Väntevärdet av en funktion $H(X, Y)$ definieras som

$$\mathsf{E}(H(X,Y)) = \begin{cases} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} H(i,j) p_{X,Y}(i,j), & \text{diskret} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x,y) f_{X,Y}(x,y) dxdy, & \text{kontinuerlig} \end{cases}$$

Beroende

Oberoende slumpvariabler

Två slumpvariabler X och Y sägs vara oberoende om deras täthetsfunktion (eller sannolikhetsfunktion) kan skrivas som produkten av marginalfördelningarna. Det vill säga, för det diskreta fallet

$$p_{X,Y}(i,j) = p_X(i)p_Y(j)$$

och för det kontinuerliga fallet

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

Kovarians

Kovariansen mellan X och Y definieras som $\mathsf{C}(X,Y) = \mathsf{E}[(X-\mu_X)(Y-\mu_Y)]$ där $\mu_X = \mathsf{E}(X)$ och $\mu_Y = \mathsf{E}(Y)$.

Kovarians

Sats

Kovariansen kan beräknas som

$$\mathsf{C}(X, Y) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y)$$

Sats

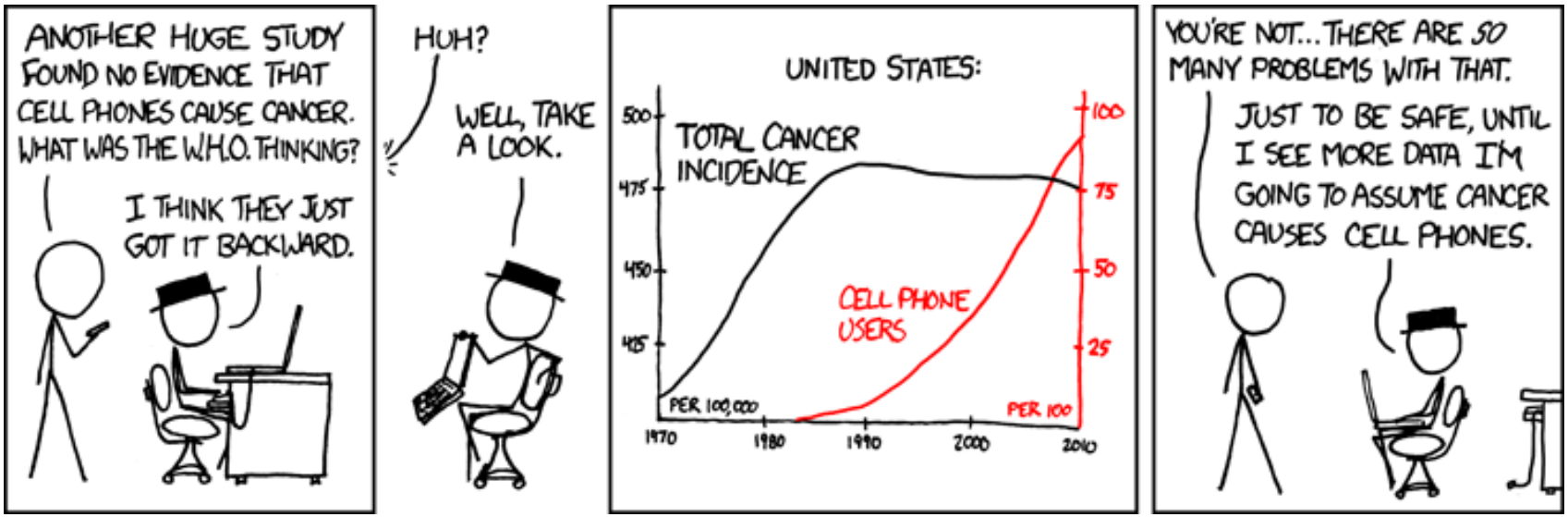
Om X och Y är oberoende så är $\mathsf{C}(X, Y) = 0$ och $\mathsf{E}(XY) = \mathsf{E}(X)\mathsf{E}(Y)$.

Korrelation

Den så kallade korrelationskoefficienten definieras som

$$\rho(X, Y) = \frac{\mathsf{C}(X, Y)}{\sqrt{\mathsf{V}(X)\mathsf{V}(Y)}}.$$

Korrelation



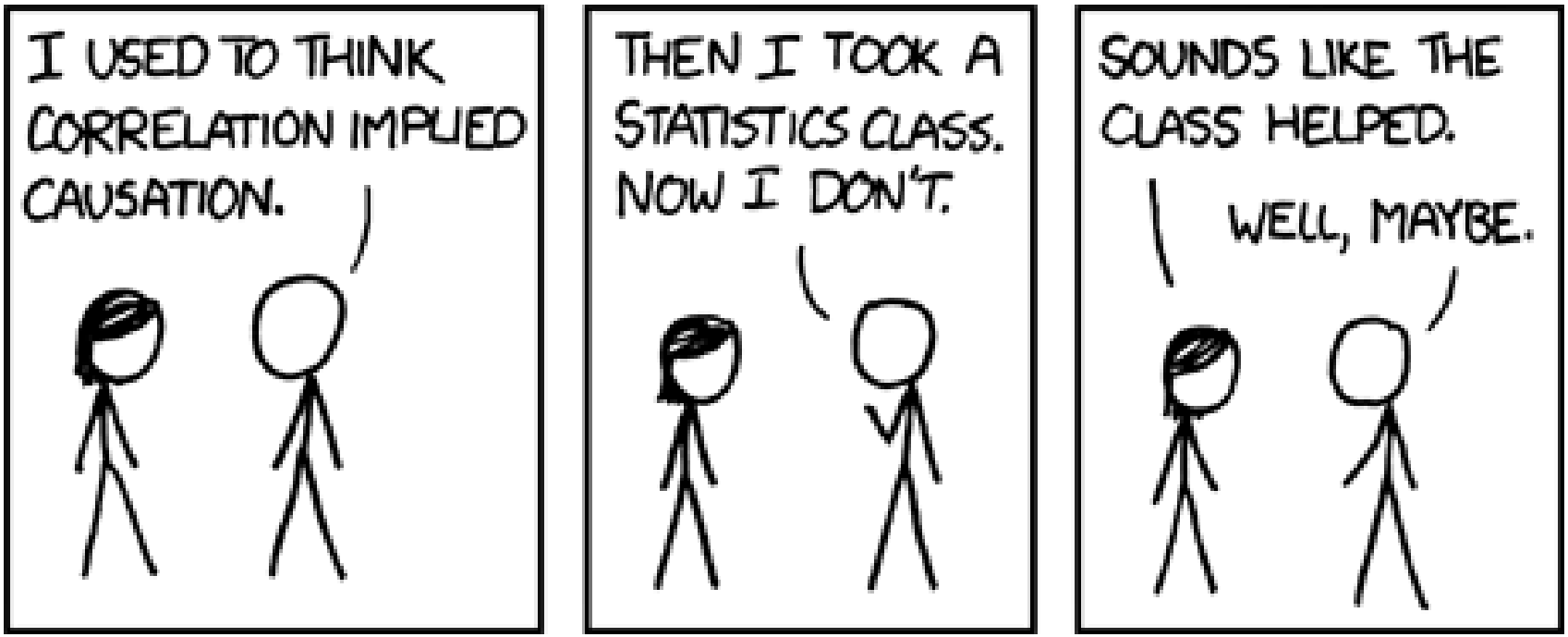
Korrelation beskriver linjärt beroende mellan två variabler.

Korrelation säger inget om orsakssamband (kausalitet)!

(men telefoner $\Rightarrow$ cancer kanske är troligare än cancer $\Rightarrow$ telefoner)

Vi kan också ha att all samvariation kan förklaras av en tredje icke uppmätt variabel.

Korrelation



Statistikteori

Punktskattningar

Stickprov

Ett stickprov av storlek n är n oberoende observationer av en slumpvariabel X. Vi kan alltså skriva stickprovet som observationer av n slumpvariabler $X_1, \ldots, X_n$ där alla X_i är oberoende och likafördelade.

Skattning

En skattning av en parameter θ är en funktion $\hat{\theta}(X_1, \ldots, X_n)$ av observationerna.

Två egenskaper vi vill att vår skattare ska ha är att

- Den ska vara *väntevärdesriktig* (unbiased på engelska), vilket betyder att vi vill att $\mathrm{E}(\hat{\theta}(X_1, \ldots, X_n)) = \theta$.
- Den ska ha låg varians om n är stort, helst vill vi att $\mathrm{V}(\hat{\theta}(X_1, \ldots, X_n)) \to 0$ då $n \to \infty$.

Skattning av väntevärde och varians

Låt $x_1, \ldots, x_n$ vara observationer av oberoende och likafördelade s.v. $X_1, \ldots, X_n$ med väntevärde μ och standardavvikelse σ. Väntevärdesriktiga skattningar av μ och σ^2 är då

- $\mu^* = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$
- $(\sigma^2)^* = S^2 = \frac{Q}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Om $X_i \in \mathsf{N}(\mu, \sigma^2)$ så har vi $\mu^* \sim \mathsf{N}(\mu, \frac{\sigma^2}{n})$ och $\frac{Q}{\sigma^2} \in \chi^2(n-1)$

Låt $x_{i1}, \ldots, x_{in_i}$ vara ober. obs. från $\mathsf{N}(\mu_i, \sigma)$ då $i = 1, \ldots, k$. Då är en poolad variansskattning

- $(\sigma^2)^* = S_p^2 = \frac{Q}{f} = \frac{(n_1-1)\,S_1^2+\cdots+(n_k-1)\,S_k^2}{(n_1-1)+\cdots+(n_k-1)}$

Eftersom $X_{ij} \sim \mathsf{N}(\mu_i, \sigma^2)$ så har vi $\frac{Q}{\sigma^2} \sim \chi^2(f)$

Maximum likelihood-metoden

Tanken bakom maximum likelihood metoden är att vi vill hitta det parametervärde som mest troligast producerade det stickprov vi har. Metoden är baserad på den så kallade likelihood-funktionen, som för ett stickprov $x_1 \ldots, x_n$ är

$$L(\theta) = \prod_{i=1}^{n} f(x_i)$$

där $f(x)$ är täthetsfunktionen för fördelningen och θ är parametrarna vi vill skatta. I det diskreta fallet byter vi täthetsfunktionen mot sannolikhetsfunktionen.

ML skattare

Maximum likelihood-skattaren av en parameter θ ges av

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta)$$

Konfidensintervall

Konfidensintervall

Låt $X_1, \ldots, X_n$ vara slumpvariabler med en fördelning som har θ som en parameter med θ_0 som sant okänt värde. Ett $100(1-\alpha)\%$ konfidensintervall för θ med konfidensgraden $1-\alpha$ är ett intervall $[a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]$ sådant att

$$\mathsf{P}(a \leq \theta_0 \leq b) = 1 - \alpha.$$

- (a, b) är ett slumpmässigt intervall, eftersom a och b är slumpvariabler som beror av $X_1, \ldots, X_n$.
- Konfidensintervallet ska alltså tolkas som att om vi gör upprepade mätningar av $X_1, \ldots, X_n$ och bildar konfidensintervallet för alla dessa mätningar så kommer $100(1-\alpha)\%$ av dessa intervall täcka det sanna värdet θ_0.

Konfidensintervall

Om $\theta^*(X_1, \ldots, X_n) \in \mathsf{N}(\theta, \mathsf{V}(\theta^*))$ så har vi

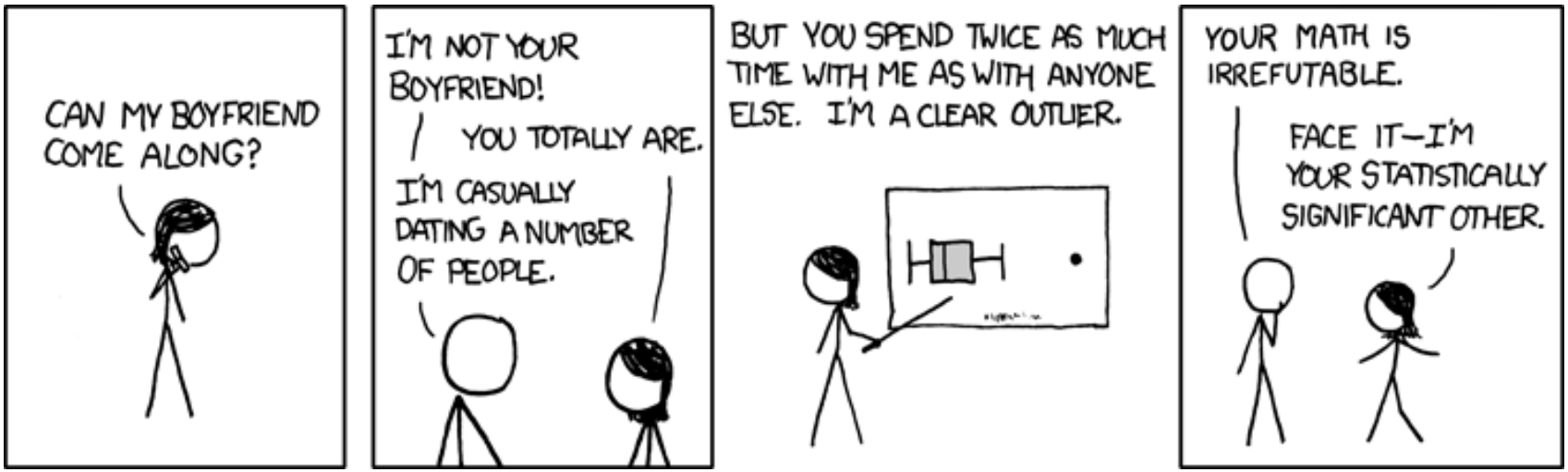
- $I_\theta = (\theta^* \pm \lambda_{\alpha/2} \cdot \sqrt{\mathsf{V}(\theta^*)})$ om $\mathsf{V}(\theta^*)$ är känd
- $I_\theta = (\theta^* \pm t_{\alpha/2}(f) \cdot s \cdot c)$ om $\mathsf{V}(\theta^*) = c^2 \cdot \sigma^2$ där $\sigma^2 = \mathsf{V}(X_i)$, c är en konstant och $(\sigma^2)^* = S^2 = \frac{Q}{f}$ med $\frac{Q}{\sigma^2} \sim \chi^2(f)$

Om $\theta^*(X_1, \ldots, X_n) \approx \mathsf{N}(\theta, \mathsf{V}(\theta^*))$ enligt CGS (el. dyl.) så kan vi använda normalbaserade konfidensintervall approximativt.

Om $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$ med $(\sigma^2)^* = S^2 = \frac{Q}{f}$ och $\frac{Q}{\sigma^2} \in \chi^2(f)$ så

$$I_{\sigma^2} = \left(\frac{f \cdot s^2}{\chi^2_{\alpha/2}(f)}, \frac{f \cdot s^2}{\chi^2_{1-\alpha/2}(f)} \right)$$

Hypotesprövning



Vi har ett stickprov med någon fördelning med okänd parameter θ. Vi vill testa *nollhypotesen*

$$H_0 : \theta = \theta_0$$

mot *mothypotesen*

$$H_1 : \theta \neq \theta_0.$$

Om testet bekräftar att $\theta \neq \theta_0$ säger vi att H_0 förkastas till förmån för H_1. Man brukar då också säga att θ är signifikant skild från θ_0.


I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT YEARS AGO.

Konfidensintervallmetoden

Om vi bildar ett 95% konfidensintervall I_θ för θ kan vi direkt använda detta för att göra hypotestestet.

- Vi förkastar H_0 om intervallet inte innehåller θ_0.
- Om intervallet däremot innehåller θ_0 kan vi inte förkasta H_0.

Konfidensgraden vi använder när vi beräknar konfidensintervallet är också konfidensgraden för hypotestestet vi gör.

Definition

Felrisken eller signifikansnivån definieras som

$$\alpha = \mathsf{P}(H_0 \text{ förkastas } | H_0 \text{ sann}).$$

Översatt till konfidensintervall är signifikansnivån

$$\alpha = \mathsf{P}(\theta \notin I_\theta \text{ om } \theta = \theta_0) = \mathsf{P}(\theta_0 \notin I_\theta).$$

Typ 2 fel

Felet vi gör om vi förkastar H_0 trots att den är sann brukar kallas för ett ”Typ 1 fel”.

Den andra sortens fel vi kan göra är ”Typ 2 fel”:

Definition

Under hypotesprövning säger vi att vi gör ett typ 2 fel om vi inte förkastar H_0 trots att H_1 är sann. Sannolikheten för att göra ett typ 2 fel brukar betecknas med β:

$$\beta = \mathrm{P}(H_0 \text{ förkastas inte } | H_1 \text{ sann}).$$

Sannolikheten för att göra ett sådant fel betecknas med β. Sannolikheten att *inte* göra ett typ 2 fel kallas för testets *styrka*, och ges alltså av $1 - \beta$.

Teststorheter

Definition

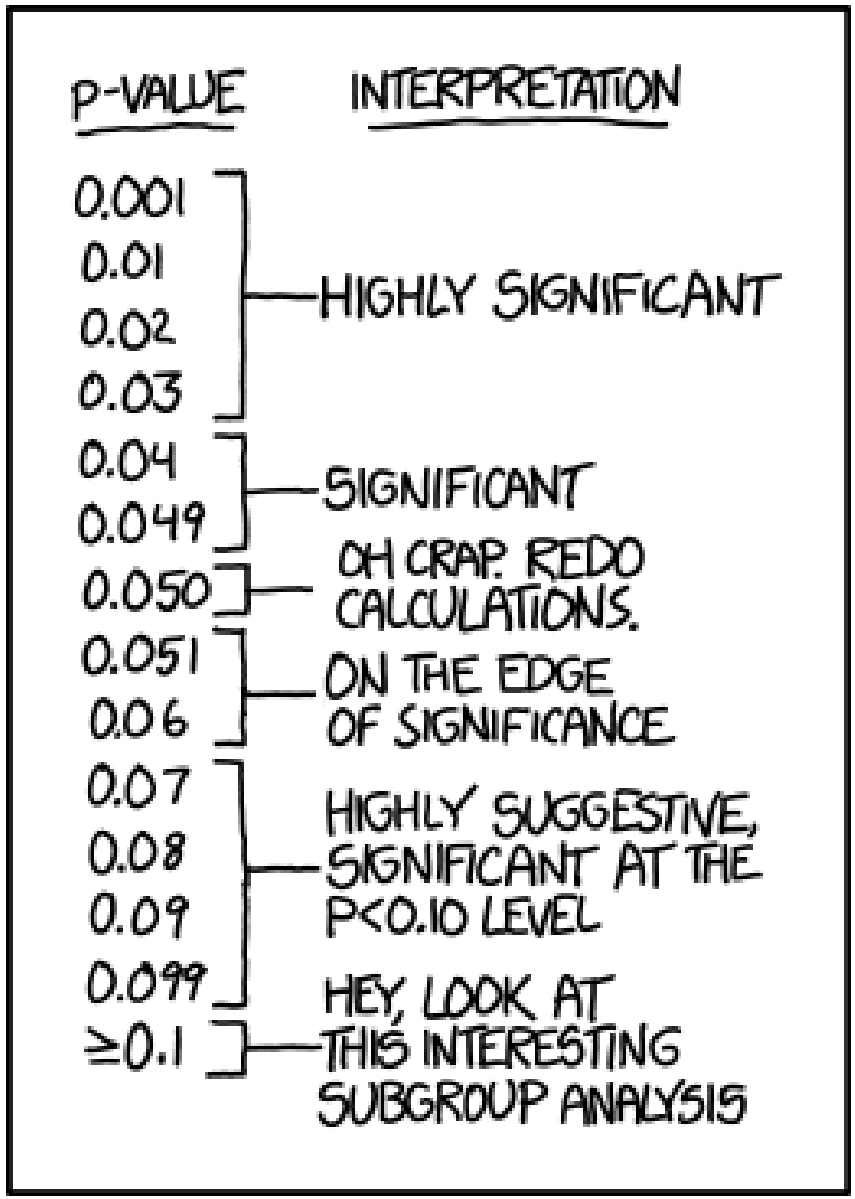
En teststorhet $T = T(X_1, \ldots, X_n)$ är en funktion av observationerna, och alltså en slumpvariabel. $T_{obs} = T(x_1, \ldots, x_n)$ är ett observerat värde av teststorheten för givna observationer.

I allmänhet innehåller teststorheten en skattning θ^* av parametern.

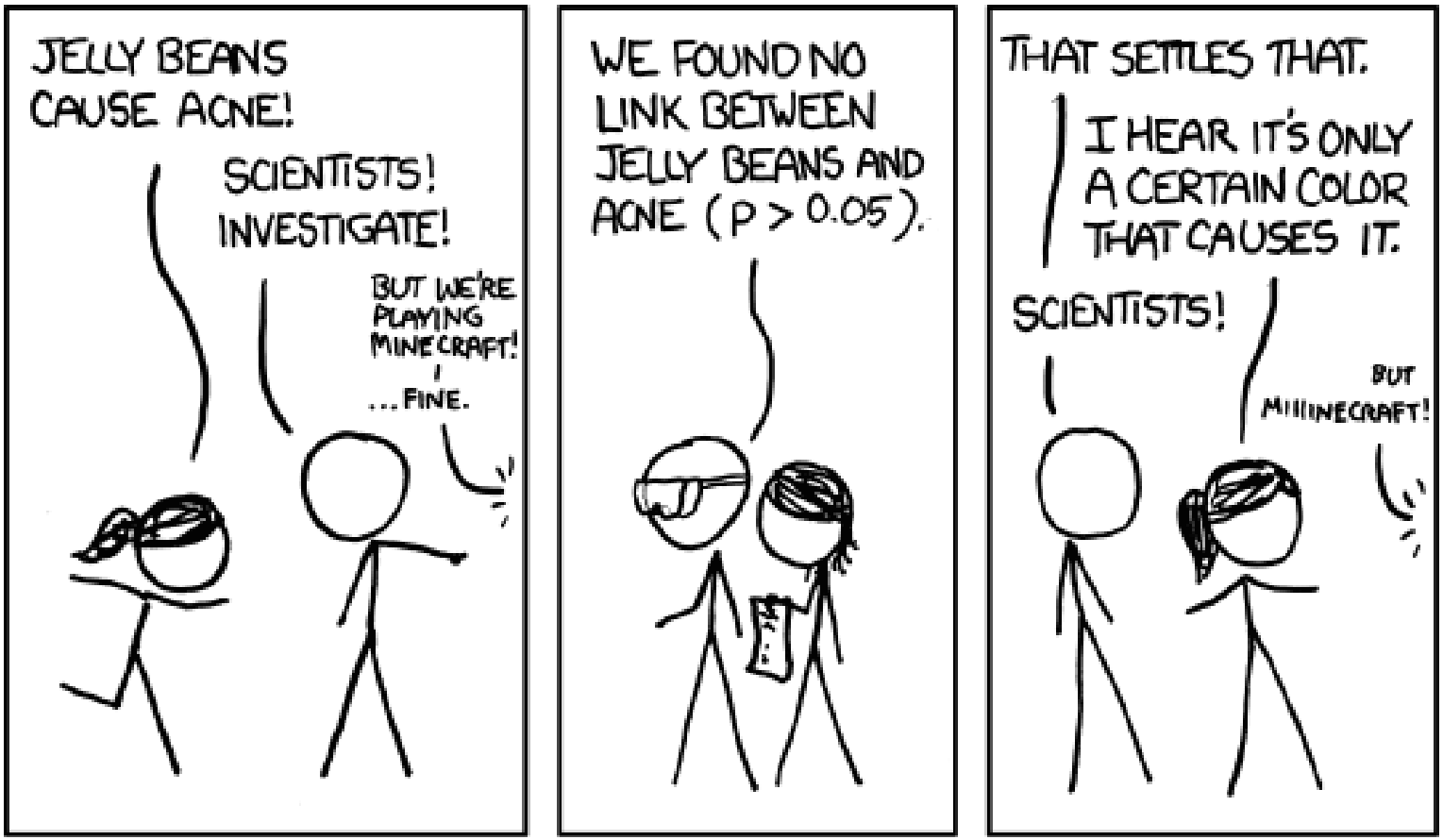
Givet att vi har formulerat vår teststorhet gör vi testet genom att beräkna ett p-värde.

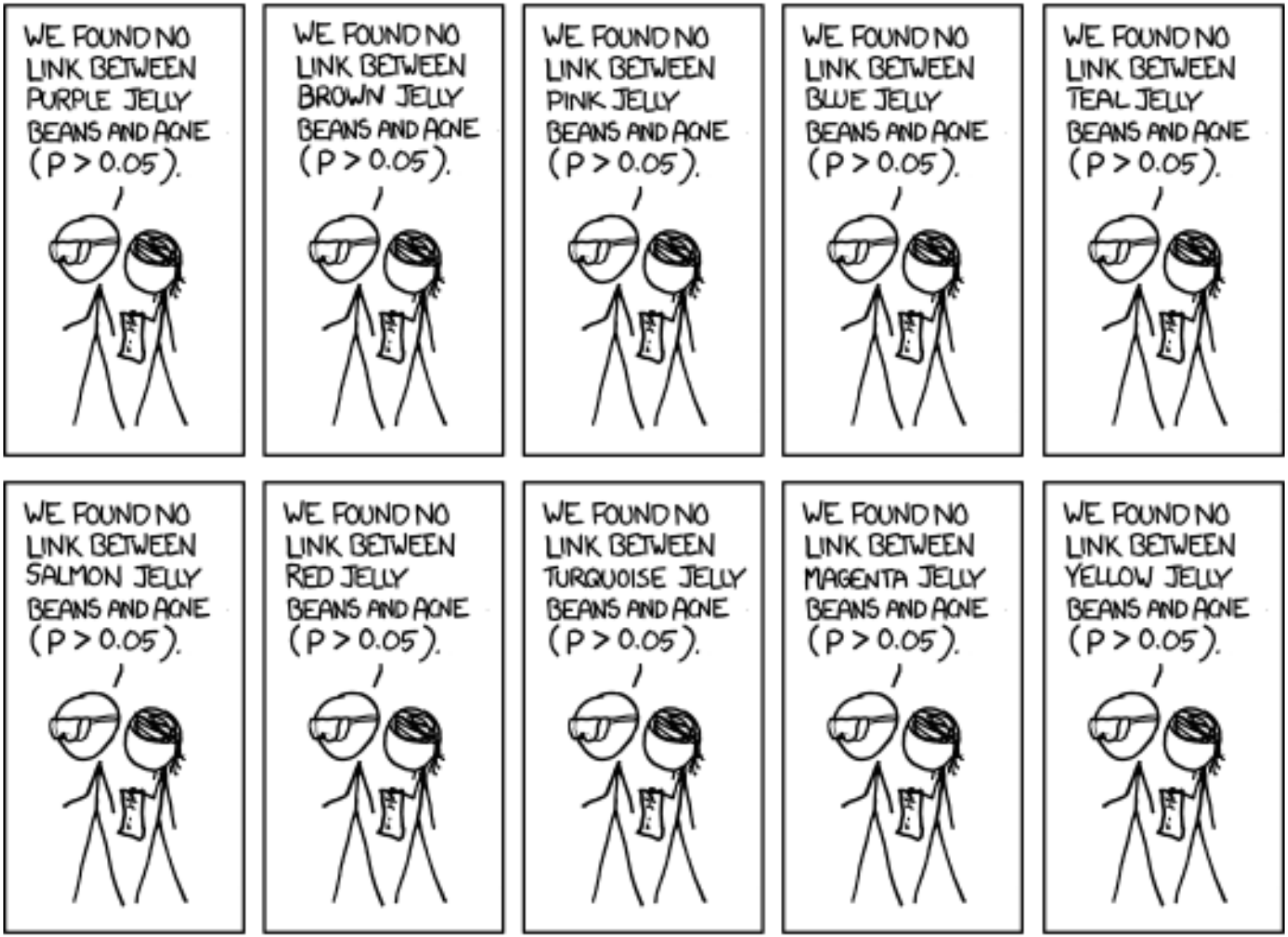
Definition

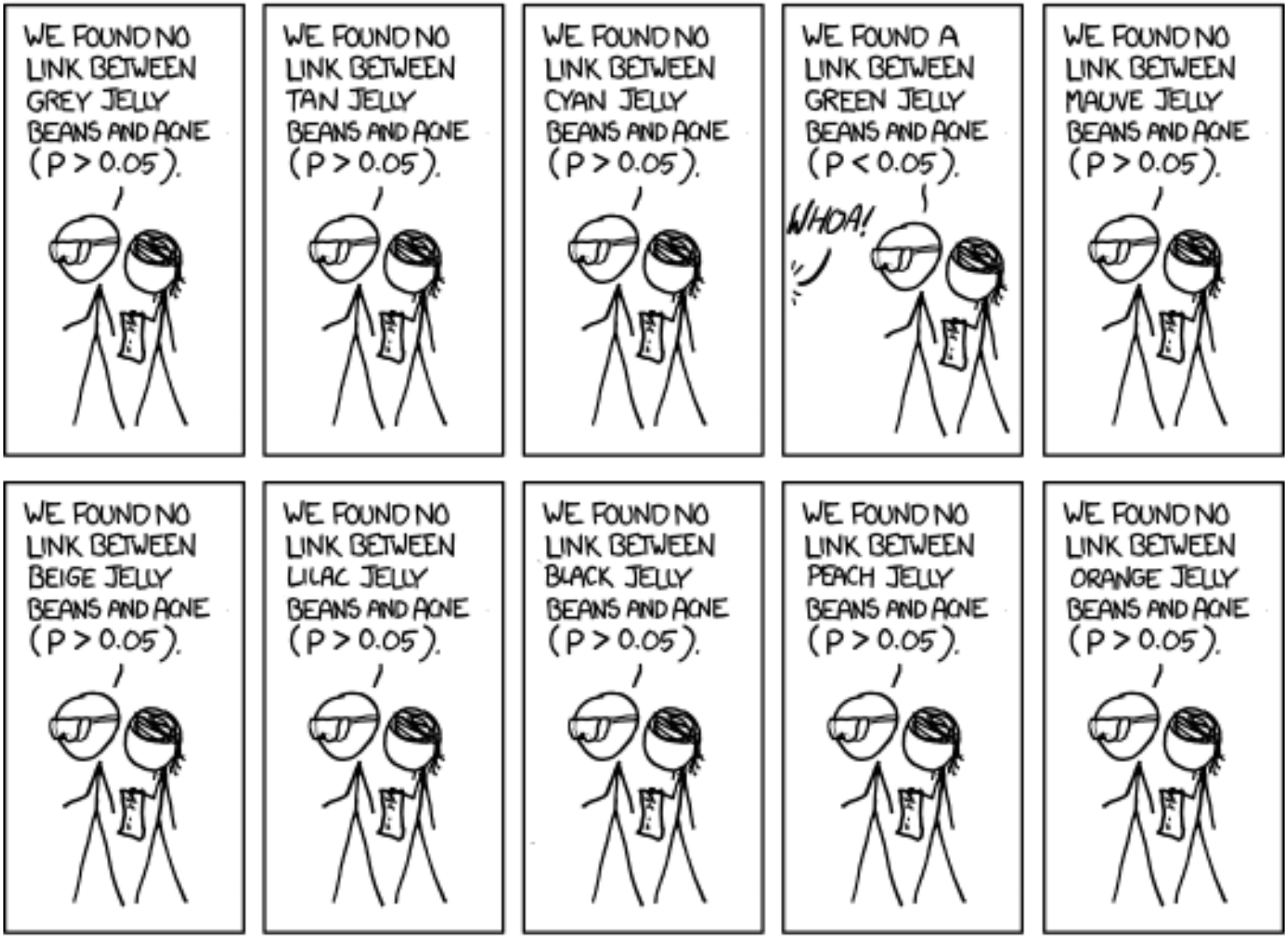
p-värdet eller signifikanssannolikheten definieras som sannolikheten under nollhypotesen att vi får ett värde $|T|$ som är lika stort eller större än det observerade värdet $|T_{obs}|$.

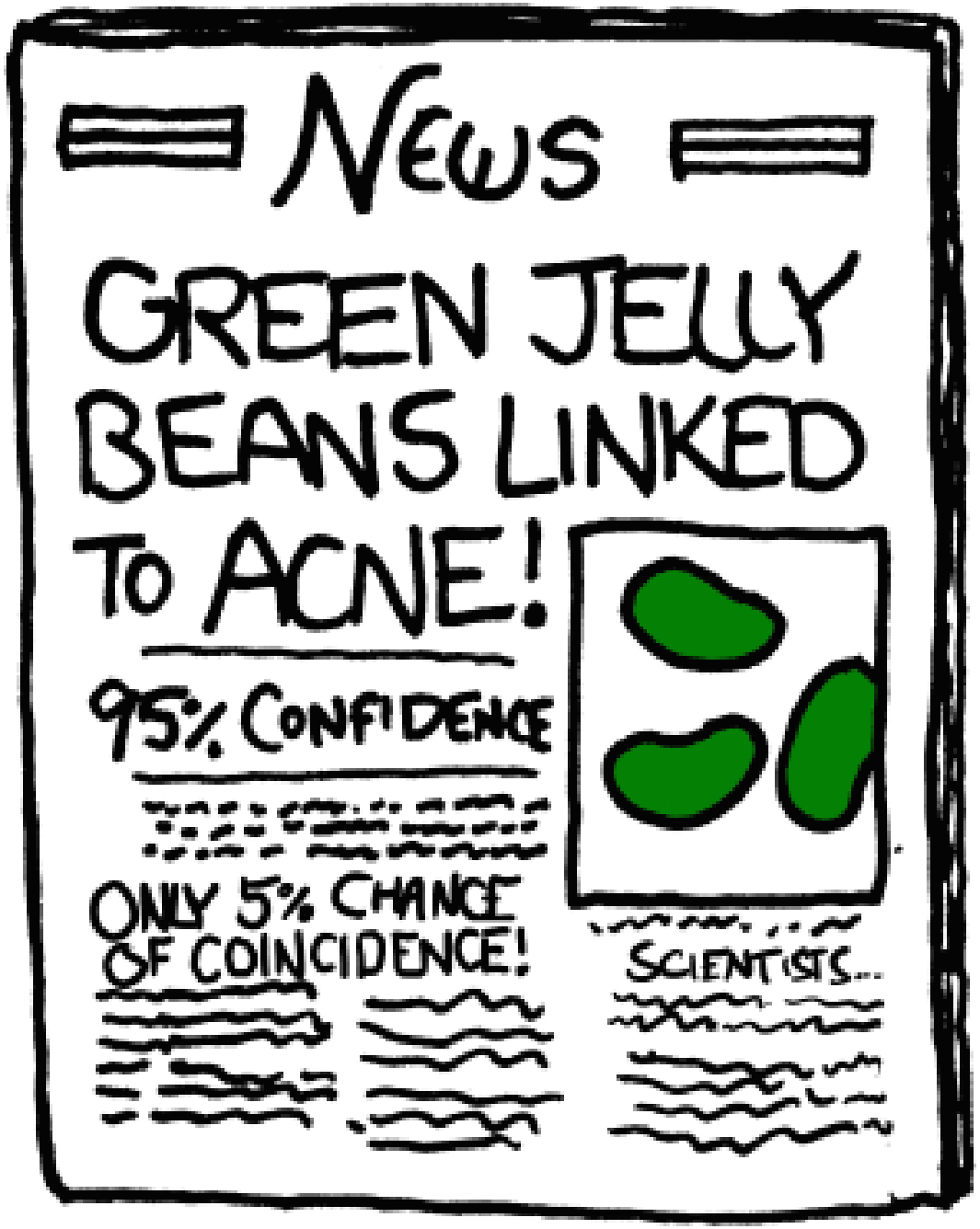

P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.099
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Multipla test och ”signifikansfiske”




WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Oberoende stickprov

Vi antar att vi har två oberoende stickprov

- n_1 observationer $X_{11}, X_{12}, \ldots, X_{1n_1}$ från $\mathrm{N}(\mu_1, \sigma_1^2)$.
- n_2 observationer $X_{21}, X_{22}, \ldots, X_{2n_2}$ från $\mathrm{N}(\mu_2, \sigma_2^2)$.

Den intressanta parametern är nu $\theta = \mu_1 - \mu_2$ som vi skattar med $\theta^* = \bar{X}_1 - \bar{X}_2$ och vi vill nu bilda ett konfidensintervall för θ samt testa hypotesen

$$H_0 : \theta = 0,$$

vilket är samma sak som att testa $\mu_1 = \mu_2$, mot

$$H_1 : \theta \neq 0$$

eller någon ensidig hypotes, $\theta > 0$ (som motsvarar $\mu_1 > \mu_2$) eller $\theta < 0$ (som motsvarar $\mu_1 < \mu_2$). Vi skiljer på tre fall:

1. σ_1 och σ_2 är kända.
2. $\sigma_1 = \sigma_2 = \sigma$ där σ är okänd.
3. σ_2 och σ_2 är okända och ej säkert lika.

Testet i de olika fallen

Vi bildar teststorheten

$$T = \frac{\theta^*}{\sqrt{\mathsf{V}(\theta^*)}}$$

- Om σ_1 och σ_2 är kända gäller att $\mathsf{V}(\theta^*) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ och T är $\mathsf{N}(0,1)$-fördelad.
- Om $\sigma_1 = \sigma_2 = \sigma$ där σ är okänd använder vi en poolad variansskattning och skattar $\mathsf{V}(\theta^*)$ med $s_p^2(1/n_1 + 1/n_2)$. T är då $t(n_1 + n_2 - 2)$-fördelad.
- Om σ_1 och σ_2 är okända och $\sigma_1 \neq \sigma_2$ skattar vi $\mathsf{V}(\theta^*)$ med $s_1^2/n_1 + s_2^2/n_2$. T är då approximativt $t(f)$-fördelad med

$$f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Jämförelse av variansen för oberoende stickprov

För att jämföra de två varianserna inför vi teststorheten

$$T = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

T är $F(n_1 - 1, n_2 - 1)$-fördelad. Vi låter $F_\alpha(f_1, f_2)$ beteckna α-kvantilen i F-fördelningen och får att ett konfidensintervall för σ_1^2/σ_2^2 ges av

$$I_{\sigma_1^2/\sigma_2^2} = \left[\frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)} \right]$$

För hypotestest av $\sigma_1^2/\sigma_2^2 = 1$ bildar vi teststorheten

$$T = s_1^2/s_2^2$$

som under H_0 är $F(n_1 - 1, n_2 - 1)$-fördelad.

Stickprov i par

En annan vanlig situation är att mätningarna uppkommer i par, till exempel om

- Man vill studera hur mycket rökare går upp i vikt när de slutar röka. Man mäter då vikten före och efter för varje person före och efter den slutar röka och jämförelsen sker för varje person.
- Man vill studera systematiska skillnader mellan två mätmetoder och använder varje metod på var och ett av ett antal prover och jämför de två metoderna för varje prov.

Modellen vi nu ansätter är att vi har n observationer i två stickprov

$$X_1, X_2, \ldots, X_n \qquad\qquad Y_1, Y_2, \ldots, Y_n$$

För varje mätning bildar vi differensen som antas vara normalfördelad:

$$D_i = X_i - Y_i \sim \mathsf{N}(\Delta, \sigma^2)$$

Vi vill nu testa om $\Delta = 0$, vilket görs som vanligt för normalfördelade mätningar.

Binomialfördelningen

Låt $X \sim \text{Bin}(n, p)$. $p^* = x/n$ är en väntevärdesriktig skattning av p med varians $\text{V}(p^*) = p(1-p)/n$.

Enligt CGS är X approximativt $\text{N}(np, np(1-p))$-fördelad om n är stor ($np(1-p) > 10$). Vi har då att

$$T = \frac{p^* - p}{\sqrt{p^*(1-p^*)/n}}$$

är approximativt $\text{N}(0, 1)$-fördelad.

Om vi vill testa

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

eller en ensidig hypotes $H_1 : p > p_0$ eller $H_1 : p < p_0$ så är $X \sim \text{Bin}(n, p_0)$ under H_0, och vi kan direkt beräkna p-värdet

Jämförelser i binomialfördelningen

Antag att $X_1 \sim \text{Bin}(n_1, p_1)$ och $X_2 \sim \text{Bin}(n_2, p_2)$ och vi vill testa om det är rimligt att anta att $p_1 = p_2$.

$p_1^* - p_2^*$ är en väntevärdesriktig skattning av $p_1 - p_2$.

Om $n_1 p_1(1-p_1)$ och $n_2 p_2(1-p_2)$ båda är stora (säg större än 10) så är $p_1^* - p_2^*$ approximativt normalfördelad. Vi har

$$\text{V}(p_1^* - p_2^*) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

ett konfidensintervall för $p_1 - p_2$ fås som

$$I_{p_1-p_2} = \left(p_1^* - p_2^* \pm z_{\alpha/2} \sqrt{\frac{p_1^*(1-p_1^*)}{n_1} + \frac{p_2^*(1-p_2^*)}{n_2}} \right)$$

Hypotestest

Om vi vill testa $H_0 : p_1 = p_2$ kan vi använda en poolad variansskattning eftersom vi har att under H_0 så gäller

$$\mathsf{V}_{H_0}(p_1^* - p_2^*) = p(1-p)(\frac{1}{n_1} - \frac{1}{n_2})$$

Här är p det gemensamma värdet under H_0, vilket skattas som

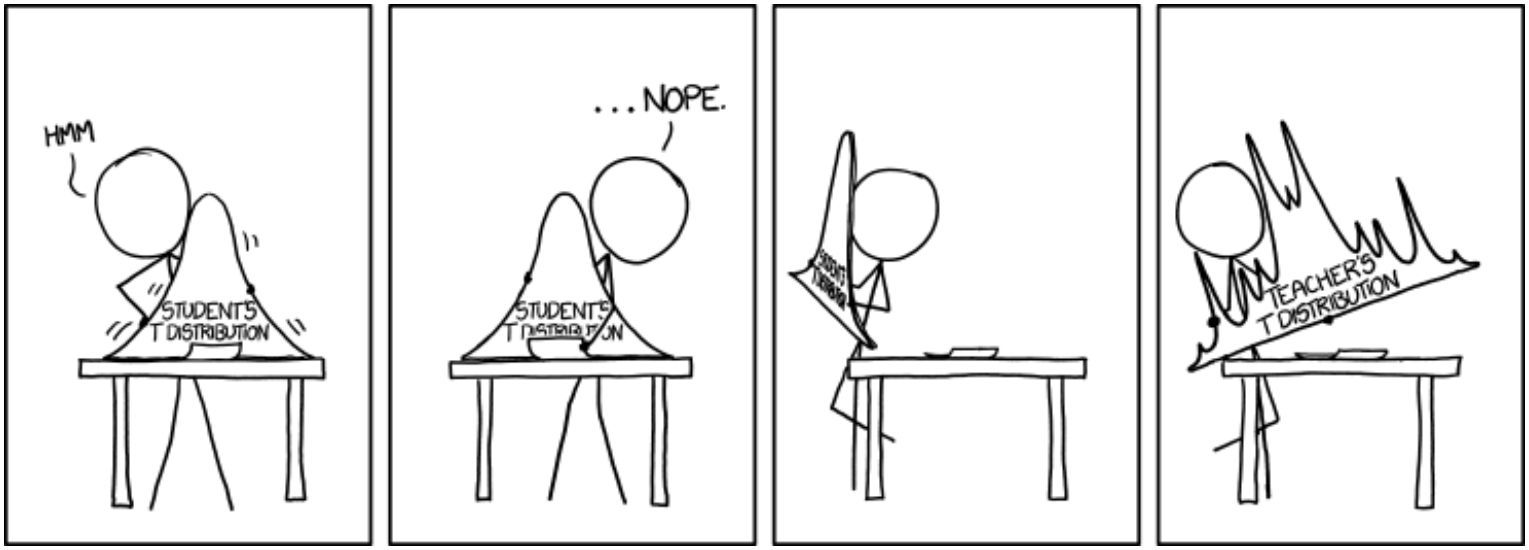
$$p^* = \frac{x_1 + x_2}{n_1 + n_2}$$

Vi bilar då teststorheten

$$T = \frac{p_1^* - p_2^*}{\sqrt{p^*(1-p^*)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

som under H_0 är approximativt $\mathsf{N}(0,1)$ fördelad.

Icke-parametriska metoder



Om vi inte hittar någon fördelning som passar när vi vill göra ett test kan vi använda någon icke-parametrisk metod.
Givet ett stickprov av storlek n vill vi testa $H_0 : M = M_0$ mot en ensidig eller tvåsidig mothypotes. Vi kan då använda ett

- Teckentest eller Wilcoxons ranktest.

Givet två oberoende stickprov $X_1, \ldots, X_m$ och $Y_1, \ldots, Y_n$ från några fördelningar X respektive Y vill vi testa $H_0 : M_X = M_Y$.

- Vi kan då använda Wilcoxons ranksummetest.

Linjär regression

Vi har talpar (x_i, Y_i) där x_i är ett fixt värde och Y_i är en slumpvariabel. Modellen vi ansätter för Y_i är

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

där ε_i är oberoende $\mathsf{N}(0, \sigma^2)$ slumpvariabeler som beskriver mätfelet och β_0 och β_1 är okända parametrar som beskriver det linjära sambandet.

MK-skattningar av β_0 och β_1

Inför

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Vi kan nu skriva upp MK skattningen av parametrarna som

$$\beta_1^* = S_{xy}/S_{xx}$$
$$\beta_0^* = \bar{y} - \beta_1^*\bar{x}$$

MK-skattning av σ^2

Variansparametern σ^2 beskriver spridningen kring linjen och skattas som

$$s^2 = \frac{Q_0}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \beta_0^* - \beta_1^* x_i)^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right).$$

Vi har att

$$\beta_1^* \sim \mathsf{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\beta_0^* \sim \mathsf{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$$

$$\mu_Y^*(x_0) = \beta_0^* + \beta_1^* x_0 \sim \mathsf{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

Prediktionsintervall

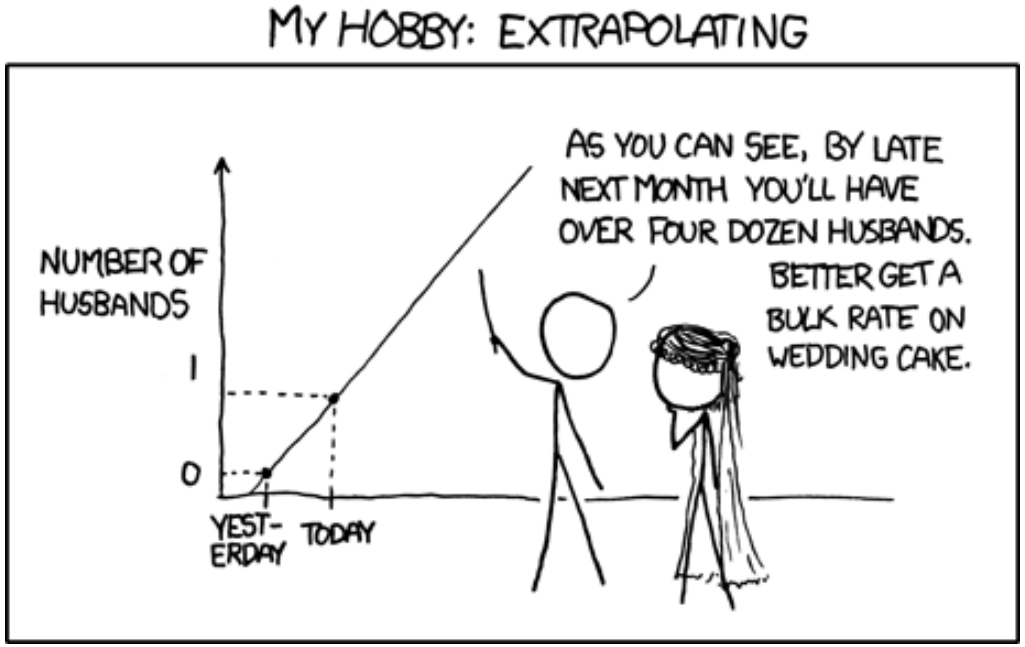
Om man är intresserad av var en framtida observation Y kommer ligga för ett visst x_0, och man använder då ett *prediktionsintervall*. Skillnaden mellan ett konfidensintervall för μ_Y och ett prediktionsintervall för Y är att konfidensintervallet talar om var väntevärdet troligen ligger, medan prediktionsintervallet talar om var en framtida observation troligen ligger.
Eftersom vi har en viss spriding kring linjen för de faktiska observationerna så måste prediktionsintervallet vara bredare än konfidensintervallet, och man kan visa att

$$Y^*(x_0) \sim \mathsf{N}\left(\beta_0 + \beta_1 x_0, \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})\right).$$

Ett prediktionsintervall bildas nu som

$$I_{Y(x_0)} = \left[\beta_0^* + \beta_1^* x_0 \pm t_{p/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right]$$

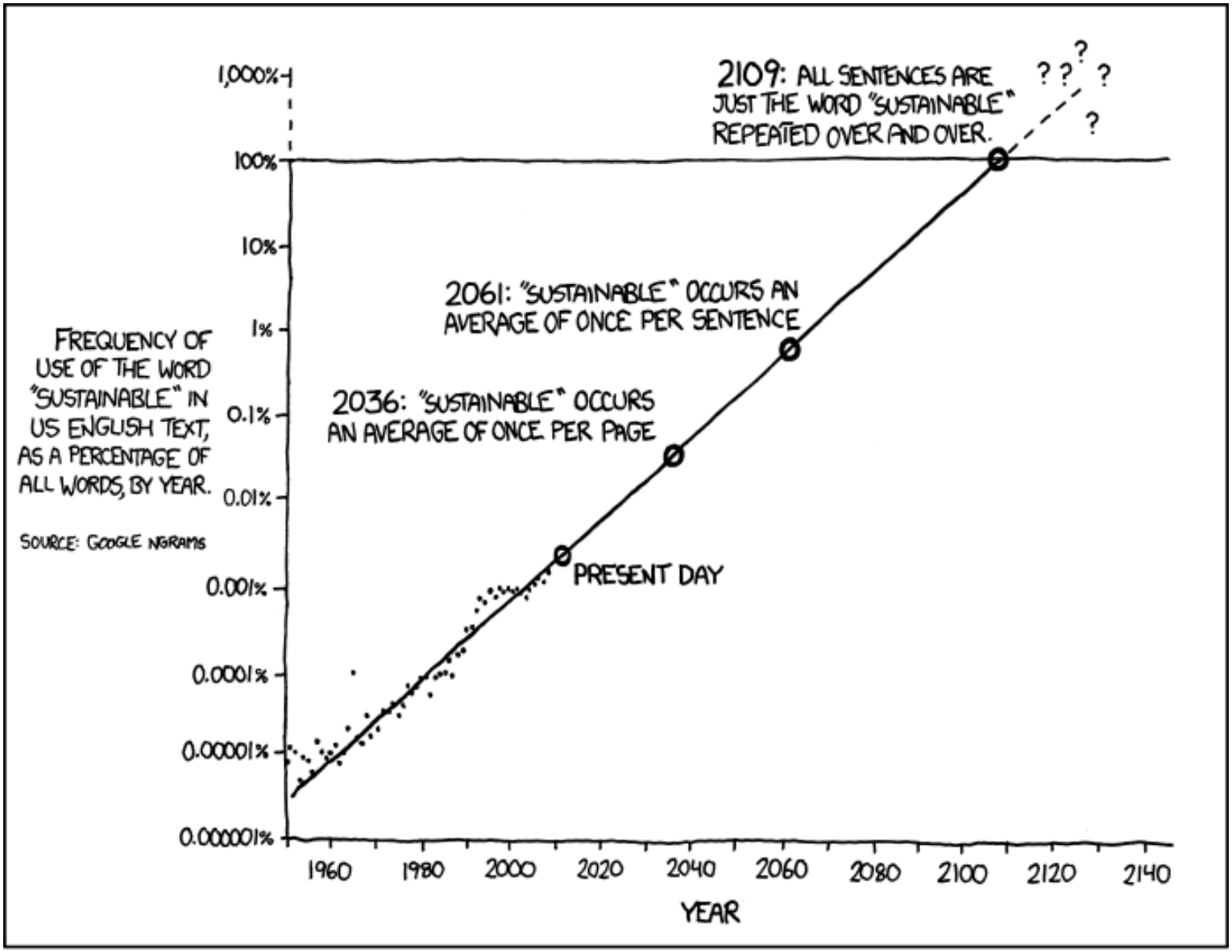
Extrapolering



Som vi tidigare nämnt är det viktigt att validera att modellantagandena är uppfyllda, genom att studera residualerna

$$e_i = y_i - \beta_0^* - \beta_1^* x_i$$

Det är också viktigt att inse att vi inte kan vara säkra på att den linjära modellen stämmer för x utanför mätområdet.



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.