Apunts del Taller de Nous Usos de la Informàtica

Jordi Vitrià

Universitat de Barcelona

10 de setembre de 2019

UNIVERSITAT DE BARCELONA

Lliçó: Hipòtesis, inferències i A-B Testing

Per què cal fer inferència estadística?

- Les dades no parlen, és l'analista que les fa parlar.
- Les dades no sempre són concluents sobre una qüestió i diferents analistes poden obtenir diferents conclusions en funció del mètode d'anàlisi usat, de l'enginyeria de característiques usada, etc.
- L'inferència estadística ens ajuda, a vegades, a resoldre part d'aquests problemes proveïnt-nos d'una metodologia i d'eines per manegar l'incertesa inherent a les dades.

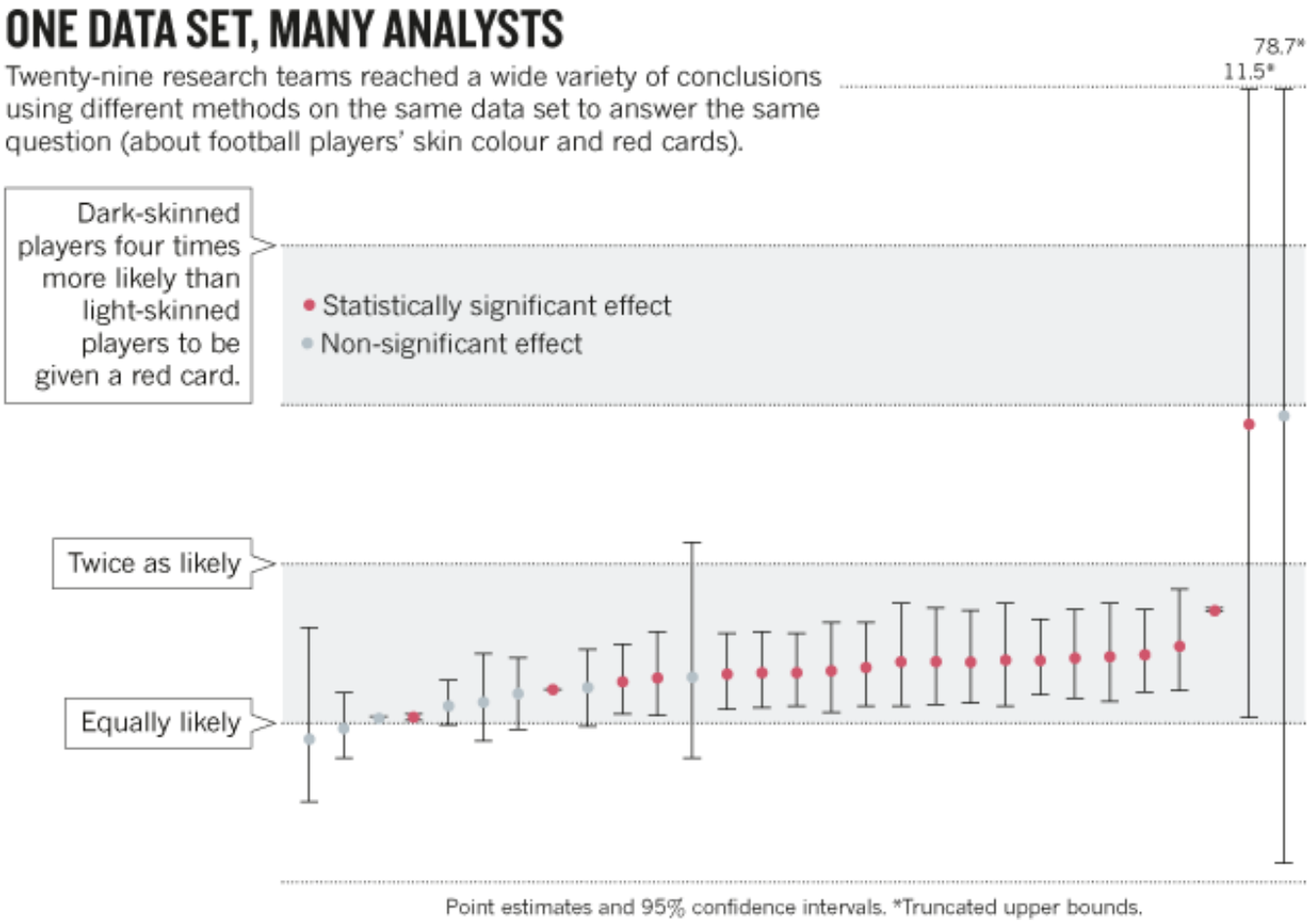
Per què cal fer inferència estadística?

- La ciència de dades ha de seguir una metodologia que minimitzi els errors i les sobre-interpretacions a partir de les dades, tot i que mai tindrem un 100% de seguretat sobre les conclusions.
- Moltes tasques habituals en ciència de dades és poden formular en tres passes:
  1. Formular una hipòtesis,
  2. Fer un experiment per recollir dades,
  3. Validar i interpretar el resultat.
- Les hipòtesis es formulen sempre ABANS de realitzar l'experiment.
- L'experiment es dissenya en funció de la hipòtesis.
- La hipòtesis es valida usant tècniques estadístiques.

Exemples

- Treuen els àrbitres de futbol més tarjetes als jugadors de pell fosca que als jugadors de pell clara?
- Hi ha relació entre la quantitat de xocolata que es menja en un pais i el nivell d'intel·ligència dels seus habitants?

Exemples


ONE DATA SET, MANY ANALYSTS
Twenty-nine research teams reached a wide variety of conclusions using different methods on the same data set to answer the same question (about football players' skin colour and red cards).
78.7*
11.5*
Dark-skinned players four times more likely than light-skinned players to be given a red card.
Statistically significant effect
Non-significant effect
Twice as likely
Equally likely
Point estimates and 95% confidence intervals. *Truncated upper bounds.

Exemples

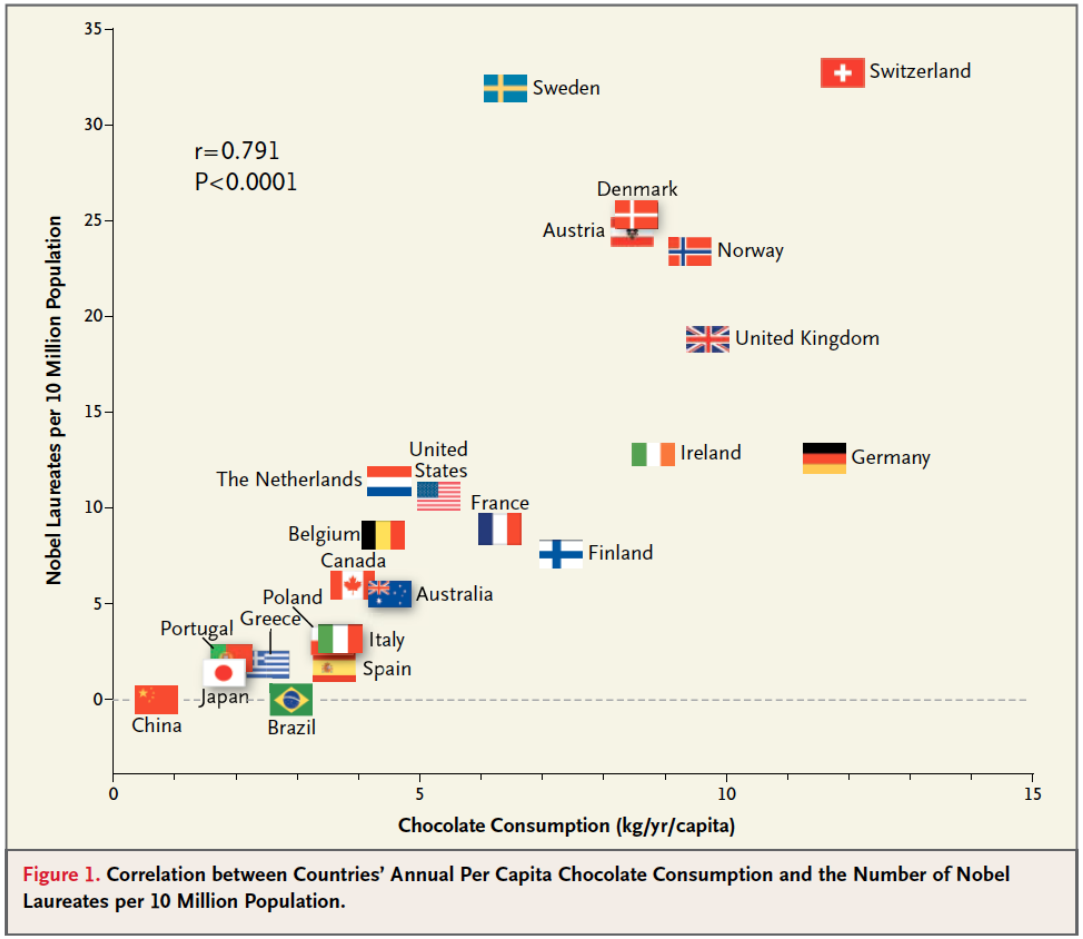


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Exemples


IMPROVE YOUR CLINICAL STRATEGIES
BROWSE CLINICAL COLLECTIONS
Welcome Guest | Renew, Subscribe or Create Account | Sign In
The NEW ENGLAND JOURNAL of MEDICINE
SUBSCRIBE OR RENEW
Includes NEJM iPad Edition, 20 FREE
Online CME Exams and more >>
HOME
ARTICLES & MULTIMEDIA
ISSUES
SPECIALTIES & TOPICS
FOR AUTHORS
CME
Keyword, Title, Author, or Citation
Advanced Search
This article is available to subscribers.
Sign in now if you're a subscriber.
Access this article:
SUBSCRIBE TO NEJM
NEW: Includes NEJM iPad Edition
Or purchase this article - $20
Print Subscriber? Activate your online access now.
Free Preview
PRINT | E-MAIL | DOWNLOAD CITATION | PERMISSIONS
OCCASIONAL NOTES
Chocolate Consumption, Cognitive Function, and Nobel Laureates
Franz H. Messerli, M.D.
N Engl J Med 2012; 367:1562-1564 | October 18, 2012 | DOI: 10.1056/NEJMon1211064
Share:
Chocolate consumption could hypothetically improve cognitive function not only in individuals but in whole populations. Could there be a correlation between a country's level of chocolate consumption and its total number of Nobel laureates per capita?
MEDIA IN THIS ARTICLE
FIGURE 1
Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.
ARTICLE ACTIVITY
38 articles have cited this article
Why Subscribe?
Your NEJM subscription includes:
• NEJM iPad Edition
• NEJM.org (1990-present)
• 20 FREE Online CME Exams
• 50 FREE Archive Views (1812-1989)
SUBSCRIBE >>
Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.
This article was published on October 10, 2012, at NEJM.org.
Dr. Messerli reports regular daily chocolate consumption, mostly but not exclusively in the form of Lindt's dark varieties.
SOURCE INFORMATION
From St. Luke's–Roosevelt Hospital and Columbia University, New York.
Access this article: Subscribe to NEJM | Purchase this article

Com es valida una hipòtesis?

Seguirem el raonament de l'**estadística freqüentista**, que és la més extesa. Hi ha altres formes de fer-ho, com l'estadística Bayesiana, que no tractarem.

- L'estadística freqüentista parteix del fet de suposar que hi ha una *població* (infinita) de la que prenem una *mostra* (finita). La població esta regida per una funció de distribució (caracteritzada per uns paràmetres) desconeguda i l'única forma (aproximada) de saber quins són aquests paràmetres és calcular-los per la mostra.
- El que és segur és que si estimem els paràmetres a partir de la mostra, ens podem aproximar al seu valor, però hi ha una incertesa inevitable.

Com es valida una hipòtesis?

Per entendre aquest procés, el concepte bàsic és el de **funció de distribució mostral**.

- Per exemple, suposem que volem calcular quina és la durada mitjana d'un embaràs als EEUU. Com que és impossible obtenir totes les dades per calcular la mitjana, enviem una sèrie d'entrevistadors per tot el territori que han d'aconseguir la durada de 1000 embaraços.
- Des d'un punt de vista estadístic, tots els embaraços dels EEUU constitueixen la població, i el conjunt de dades dels entrevistadors constitueix la mostra.
- La pregunta que ens podem fer és: quina relació hi ha entre la mitjana de la mostra i la mitjana de la població?
- La resposta ens la dóna la **funció de distribució mostral** de la mitjana.

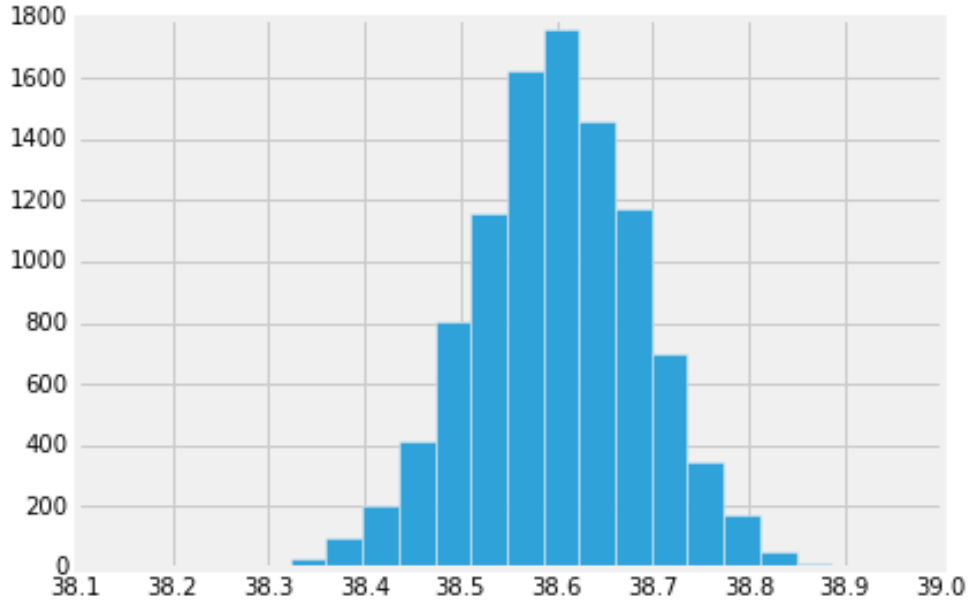
Com es valida una hipòtesi?

La **funció de distribució mostral** de la mitjana es podria (imaginàriament) calcular així:.

- Enviem 10.000 d'entrevistadors per tot el territori que han d'aconseguir la durada de 1000 embaraços cada un. Amb això tenim 10.000 mostres de mida 1000.
- Per cada mostra calculem la mitjana.
- Construïm la funció de distribució de les mitjanes calculades.

La funció que hem construït conté la informació necessària per mesurar l'incertesa associada al càlcul de la mitjana de la població a partir d'una mostra de 1000 elements.

Com es valida una hipòtesi?



Aquesta funció ens permet calcular la variança de l'estimació, la probabilitat de que l'estimació sigui més gran o igual que un determinat valor, etc.

Com es valida una hipòtesi?

Però a la realitat, l'avaluació de la incertesa seguint el mètode de construcció que hem vist de la **funció de distribució mostral** del paràmetre d'interès d'un problema no és factible!
Depenent del problema, les opcions factibles són dues: calcular de forma teòrica (mètode clàssic) o de forma empírica (mètode alternatiu) una *aproximació de la funció de distribució mostral* de la mesura que estem analitzant.
Llavors podrem calcular, a partir de l'aproximació, la probabilitat de que el resultat sigui producte de la *casualitat* i emetre una **proposició** sobre el resultat de l'anàlisi.

Com es valida una hipòtesi?

Anem a veure tres casos que exemplifiquen aquest procés:

- Com generar una proposició sobre un cas en el que disposem d'un model teòric que ens permet deduïr la funció de distribució mostral de l'esdeveniment d'interès (monedes). Aquest cas es dona poc a la realitat. En aquest cas no cal ni tant sols calcular una aproximació de la funció de distribució mostral, atès que podem calcular explícitament la probabilitat de que el resultat sigui una casualitat!
- Com generar una proposició sobre la diferència entre dues mitjanes. És un cas molt important en ciència de dades i fins i tot en disseny web.
- Com generar una proposició sobre un paràmetre sobre le qual no disposem d'un model teòric que ens permeti usar la funció de distribució mostral.

Problema 1: Inferència Estadística

When spun on edge 250 times, a Belgian one-euro coin came up heads 140 times and tails 110. *It looks very suspicious to me*, said Barry Blight, a statistics lecturer at the London School of Economics. *If the coin were unbiased the chance of getting a result as extreme as that would be less than 7%.*

"The Guardian", 4 de gener de 2002.

El mètode clàssic: plantejament

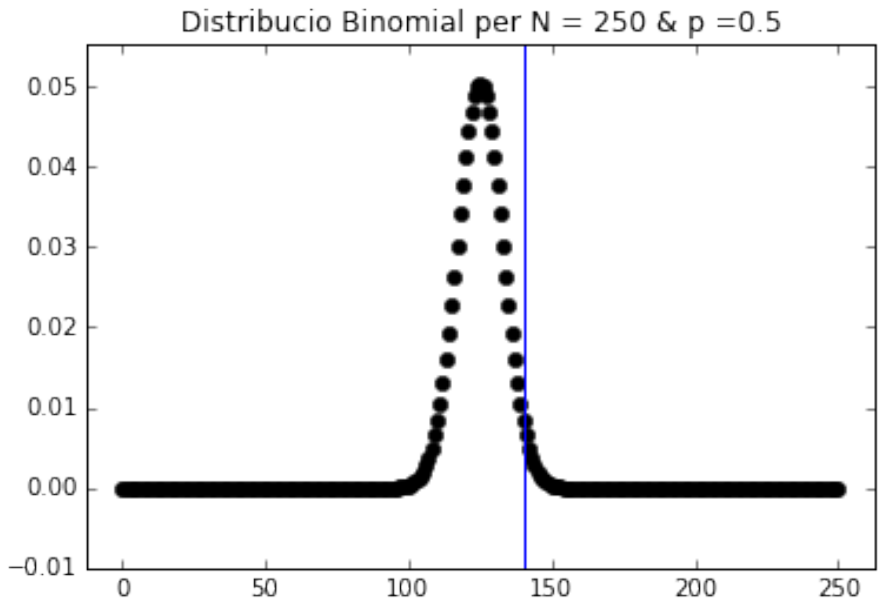
1. Assumim una posició *escèptica* respecte al resultat. En aquest cas, com que és una moneda de curs legal, la posició escèptica és assumir que la probabilitat de cara o creu és la mateixa. Aquesta posició s'anomena la **hipòtesis nula**. La hipòtesis contraria s'anomena **hipòtesis alternativa**.
2. L'experiment és llençar-la 250 vegades i recollir els resultats.
3. Segons el plantejament clàssic de l'estadística, la validació consisteix en **evaluar la probabilitat del resultat obtingut (o més intens) sota la hipòtesis nula** (o el que és el mateix, quina és la probabilitat que el resultat sigui fruit de la casualitat). Si aquesta probabilitat és alta, descartem la hipòtesi alternativa. En cas contrari, no es pot descartar.

El mètode clàssic: validació de la hipòtesi

En el cas de la moneda, la probabilitat de la hipòtesis nula es pot calcular explícitament

- $P(+) = \frac{1}{2}$
- $P(++) = (\frac{1}{2})^2$
- $P(2_+, 1_\times) = P(++\times) + P(+\times+) + P(\times++) = \frac{3}{8}$
- ...
- La funció de distribució de probabilitats que modela el cas de la moneda és la Binomial: $P(N_+, N_\times) = \binom{N}{N_+}(\frac{1}{2})^{N_+}(1-\frac{1}{2})^{N_\times}$, on $\binom{N}{N_+}$ és el nombre de combinacions de N en N_+, (N elements presos de N_+ en N_+) i $(\frac{1}{2})^{N_+}$ és la probabiliat de N_+ cares i $(1-\frac{1}{2})^{N_\times}$ és la probabilitat de $N_\times$ creus.

El mètode clàssic: validació de la hipòtesi



La línia blava correspon a $N_+ = 140$. Si sumem tot el que queda a la dreta tenim $P(N_+ \geq 140) = 0.033$ sota la hipòtesis nula.

El mètode clàssic: validació de la hipòtesi

Arribats a aquest punt l'estadística clàssica fa aquest raonament:

1. La probabilitat de tenir 140 cares o més sota la hipòtesis nula (la moneda està ben feta) és del 3%. Per tant, la probabilitat de tenir un resultat tant estrany com aquest era del 7%.
2. Aquesta probabilitat és petita... però...
3. Que fem, rebutgem l'hipòtesis nula i acceptem que la hipòtesi alternativa o no?.

L'estadística clàssica assumeix que la probabilitat d'una hipòtesis és petita si és menor que 0.05. Aquest valor és arbitrari però és el que s'usa a la pràctica.
IMPORTANT: Això vol dir que acceptem un marge d'error del 5% quan acceptem la hipòtesis alternativa!

El mètode alternatiu: Simulació

Si sabem **simular** els esdeveniments, podem construïr directament la funció de distribució mostral del paràmetre d'interès! Només cal programar-ho, calcular $P(N_+ \geq 140)$ i obtindrem el mateix resultat:

In [9]:
```
import numpy as np

M=0
for i in range(1000000):
    trials = np.random.randint(2,size=250)
    if (trials.sum() >= 140):
        M += 1
p = M/1000000.0
print p
```

```
0.033097
```

Què penseu que passa si fem més simulacions?

Problema 2: A/B Testing o com triem la millor opció?

Traffic is randomly assigned to each page variant based upon a predetermined weighting. For example, if you are running a test with 2 page variants, you might split the traffic 50 — 50 or 60 — 40. Visitors are typically cookied so that they will always see the same version of the page (to maintain the integrity of the test). Then, you can log the time each user spent at each page (assuming that more time is better). At last, you analyze the log to make a decision.

Recollida de mostres.

Suposem que ho fem per dues pàgines, la A i la B, i recollim el temps que alguns usuaris passen a cada una d'elles:

A	84	72	57	46	63	76	99	91				
B	81	69	74	61	56	87	69	65	66	44	62	69

Segons aquestes dades, el temps mitjà que un usuari passa a A és 73.5, i a B és 66.9.
Fins a quin punt podem estar segurs que A és millor que B? Dit d'una altra manera: fins a quin punt la diferència observada (que és 6.6) indica que A és millor que B?

El mètode clàssic: diferència entre mitjanes.

1. Ara tenim un problema de **diferència entre mitjanes**.
2. Assumim una posició *escèptica* respecte al resultat. En aquest cas, la posició escèptica és que el canvi de disseny no té efecte (positiu o negatiu) sobre els usuaris.
3. L'experiment és la recollida de mostres que hem vist (8 valors per A i 12 per B).
4. Segons el plantejament clàssic de l'estadística, la validació consisteix en **evaluar la probabilitat del resultat obtingut sota la hipòtesis nula**. Si aquesta probabilitat és alta (> 0.05), descartem la hipòtesi alternativa. En cas contrari, no es pot descartar.
5. Si volem seguir la metodologia, el que hauriem de fer és veure quina és la funció de distribució de probabilitats de la diferència entre dues mitjanes i calcular la probabilitat de que una diferència d'aquest estil sigui més gran (o més petita) que 6.6.

El mètode clàssic: diferència entre mitjanes.

La distribució de la diferència entre les mitjanes es pot considerar com la distribució que es produiria si repetim els següents tres passos una i altra vegada :

1. Mostreja n_1 valors de la població 1 (8 usuaris d'A) n_2 valors de la població 2 (12 usuaris de B) .
2. Calcula les mitjanes de les dues mostres ($\hat{\mu}_1$ i $\hat{\mu}_2$).
3. Calcula la diferència entre les mitjanes ($\hat{\mu}_1 - \hat{\mu}_2$).

La distribució de les diferències entre mitjanes es pot construir a partir de repeticions d'aquest experiment, però per obtenir una bona aproximació caldria fer MOLTS experiments, cosa que no és factible.

El mètode clàssic: diferència entre mitjanes.

El mètode clàssic de l'estadística es basa en es pot demostrar que la distribució mostral de la diferència entre les mitjanes de dues distribucions segueix una distribució normal $N(\mu, \sigma^2)$, on:

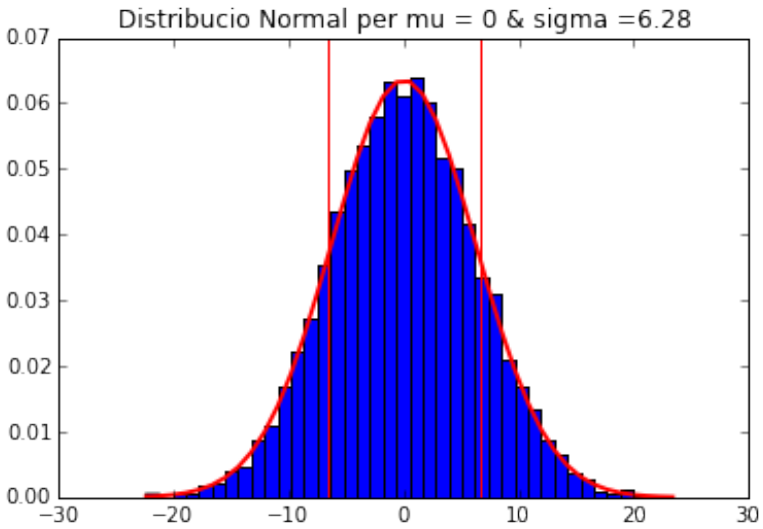
- $\mu = \mu_1 - \mu_2$, on μ_1 és la mitjana de la primera distribució i μ_2 la de la segona,
- $\sigma^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, on σ_1^2 és la mitjana de la primera distribució, σ_2^2 la de la segona i
- n_1 i n_2 són el nombre d'elements de la mostra de cada distribució respectivament.

El mètode clàssic: diferència entre mitjanes.

En el nostre cas i com que estem sota la hipòtesis nula:

1. $\mu_1 = \mu_2$ i per tant $\mu = 0$.
2. $\sigma_1^2 = \sigma_2^2$ i per tant σ^2 es pot estimar com $\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}$, on $\hat{\sigma}^2$ és la variança de les dades de les dues mostres agregades. En el nostre problema aquest valor és 6.28.

El mètode clàssic: diferència entre mitjanes.



Com que per nosaltres les dues poblacions no tenen un significat especial, hem de veure quina probabilitat hi ha de que el resultat de la diferència sigui $+6.6$ o -6.6 (doncs això depèn de quina mostra considerem primer). Al gràfic podem observar que aquest valor està al voltant d'un 30%, que és un valor molt alt. Per això diem que no podem rebutjar la hipòtesi nula.

El mètode clàssic: diferència entre mitjanes.

Si no podem rebutjar la hipòtesis nula vol dir que no hi ha evidència que una pàgina sigui millor que l'altra.

Preguntes importants: Podria canviar aquesta conclusió si trobéssim una diferència de 6.6 amb una mostra amb n_1 i n_2 molt més grans? Si realment les pàgines són equivalents, que observariem en el valor de la diferència si tenim una mostra amb n_1 i n_2 molt més grans?

El mètode alternatiu: Shuffling

Però hi ha un model alternatiu més directe, basat en la següent consideració:

- Si les etiquetes realment no importen (hipòtesis nula), llavors redistribuir-les entre les dades no ha de tenir cap efecte en la distribució mostral de la diferència entre mitjanes.

Llavors, podem aplicar el següent procediment una sèrie de vegades:

1. Barrejar (*shuffling*) les etiquetes.
2. Recalculem les mitjanes i calculem la seva diferència.

Si això ho fem moltes vegades podem construir la distribució mostral de la diferència entre mitjanes, contar quantes vegades surt una diferència més gran que la observada i assignar aquesta probabilitat al valor observat.

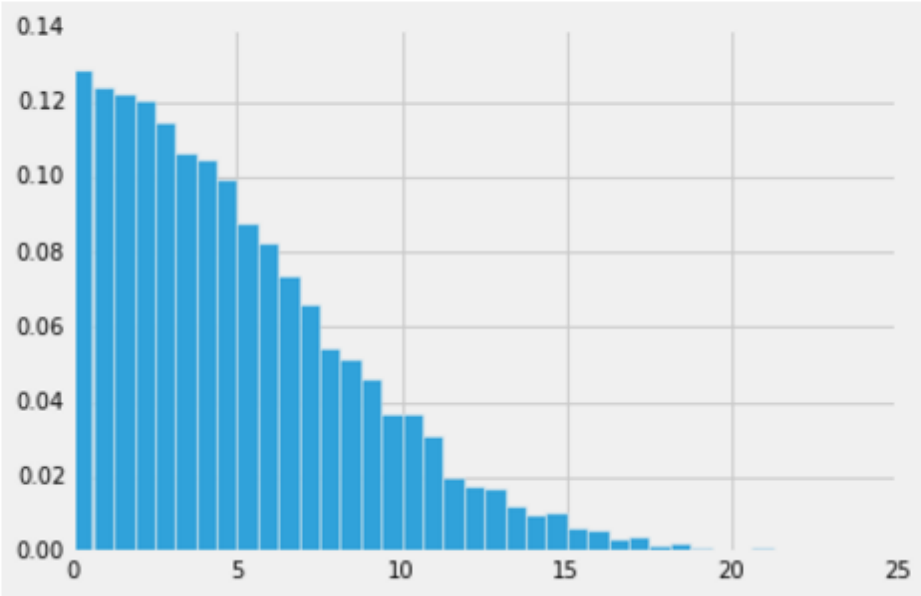
El mètode alternatiu: Shuffling

In [10]:
```
import numpy as np
import random
A = np.array([84 , 72 , 57 , 46 , 63 , 76 , 99 , 91 ])
B = np.array([81 , 69 , 74 , 61 , 56 , 87 , 69 , 65 , 66 , 44 , 62 , 69 ])
m= len(A)
n= len(B)
p = abs(A.mean() - B.mean())
pool = np.concatenate([A,B])
np.random.shuffle(pool)
```

El mètode alternatiu: Shuffling

In [11]:
```
N = 10000
diff = range(N)
for i in range(N):
    p1 = [random.choice(pool) for _ in xrange(m)]
    p2 = [random.choice(pool) for _ in xrange(n)]
    diff[i] = abs(np.mean(p1)-np.mean(p2))

diff2 = np.array(diff)
w1 = np.where(diff2 > p)[0]
len(w1)
with plt.style.context('fivethirtyeight'):
    plt.hist(diff, bins=40, normed=True)
```



El mètode alternatiu: Shuffling

In [12]:
```
print 'p-value (Simulation)=', len(w1)/float(N), '(', len(w1)/float(N)*100 ,'%)',
if len(w1)/float(N)<0.05:
    print 'The effect is likely'
else:
    print 'The effect is not likely'
```

```
p-value (Simulation)= 0.2902 ( 29.02 %) Difference = 6.58333333333
The effect is not likely
```

Problema 3: Estimació de paràmetres.

Suposem que vull estimar el nombre mitjà de clients que entren a una botiga durant els dissabtes i recullo aquestes dades durant 20 dissabtes:

48	24	32	61	51	12	32	18	19	24
21	41	29	21	25	23	42	18	23	13

Quina és la mitjana? Quina és l'incertesa sobre la seva estimació?

El mètode clàssic: Estimació de la mitjana.

L'estadística freqüentista respon a les dues preguntes amb dues fòrmules, una sobre quina és la millor estimació possible (segons una sèrie d'assumpcions no trivials) de la mitjana $\hat{\mu}$ a partir d'una mostra de N elements $\{x_i\}$ i una altra sobre l'error estàndard $\sigma_{\hat{\mu}}$ d'aquesta estimació:

$$\hat{\mu} = \frac{1}{N}\sum_{1}^{N} x_i = 28.9$$

$$\sigma_{\hat{\mu}} = \frac{1}{\sqrt{N}}\sqrt{\frac{1}{N-1}\sum_{1}^{N}(x_i - \hat{\mu})^2} = 3.0$$

El mètode alternatiu: Bootstrap.

Podem intentar la via de simular la mostra, però no tenim un model generador de l'entrada de clients a la meva botiga!

El mètode de *bootstrap* ens permet crear una aproximació robusta de la distribució mostral de la mitjana a partir d'aplicar un **mostreig aleatori amb reemplaçament**[1].

[1] Donat un conjunt d'elements i un nombre N, el mostreig aleatori amb reemplaçament consisteix en (1) Assignar un nombre enter a cada element, (2) Seleccionar N elements del conjunt (alguns d'ells possiblement repetits) mitjançant la generació de N nombres aleatoris de l'interval d'enters $(1, \ldots, N)$.

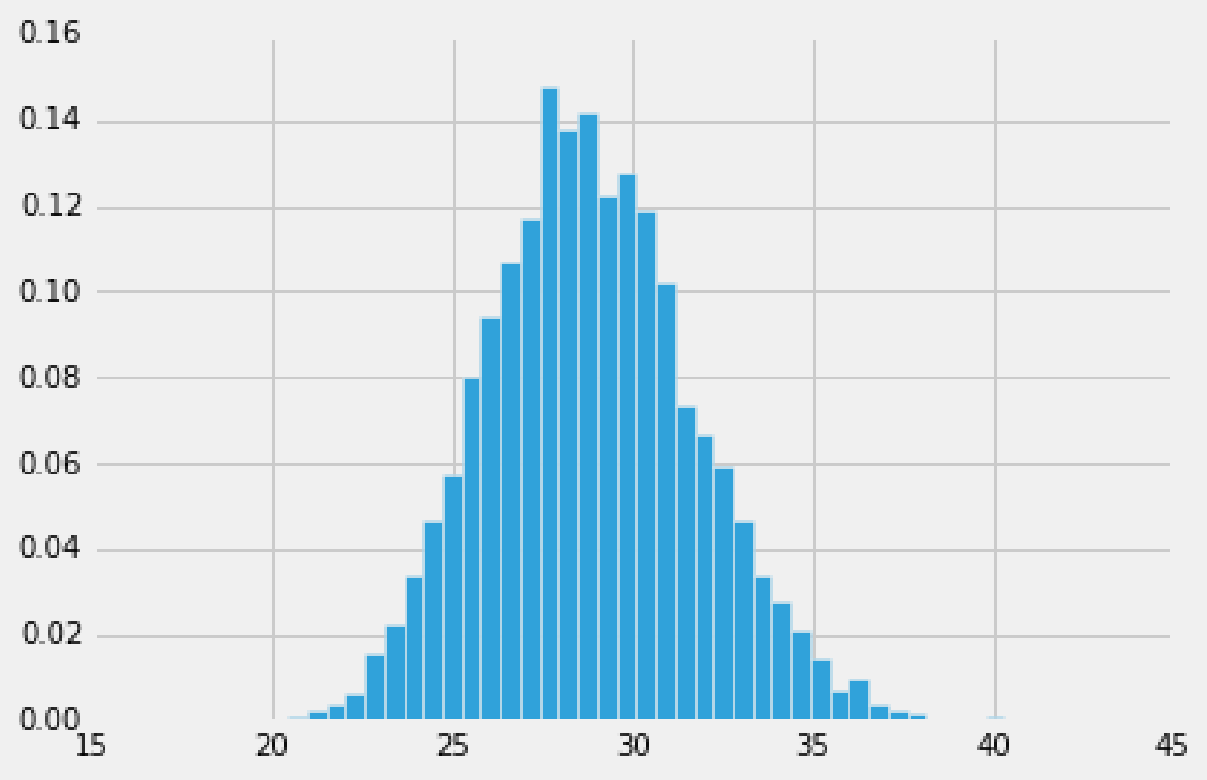
El mètode alternatiu: Bootstrap.

In [21]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

xbar = [0]*10000
X = [48, 24, 32, 61, 51, 12, 32, 18, 19, 24, 21, 41, 29, 21, 25, 23, 42, 18, 23, 13]
for i in range(10000):
    sample = [X[_] for _ in np.random.randint(20, size=20)]
    xbar[i] = np.mean(sample)
print np.mean(xbar), np.std(xbar)
with plt.style.context('fivethirtyeight'):
    plt.hist(xbar, bins=40, normed=True)
```

```
28.815425 2.88070144572
```

El mètode alternatiu: Bootstrap.



Entren 29 ± 3 persones cada dissabte a la botiga!

Altres aplicacions del bootstraping.

El mètode de boostraping es pot aplicar per mesurar l'incertesa d'estadístics més complexes, com per exemple a la regressió linial.

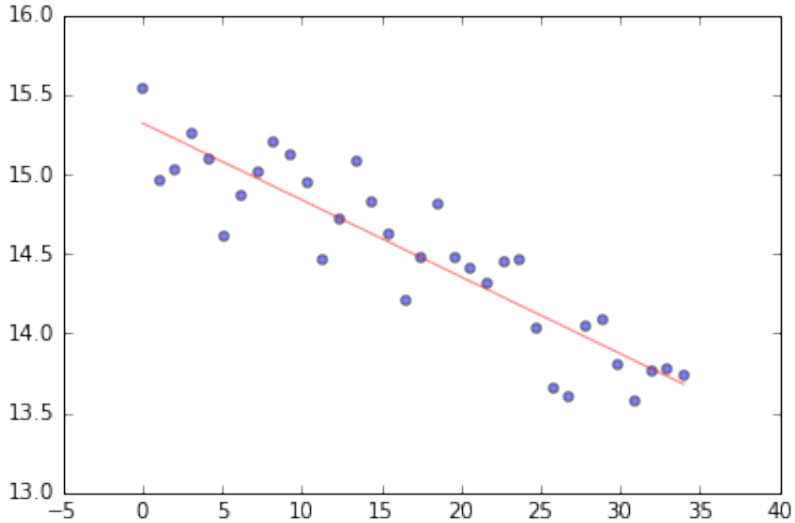
Altres aplicacions del bootstraping.

In [44]:
```
from sklearn.linear_model import LinearRegression

y = [15.54,14.96,15.03,15.26,15.10,14.61,14.86,15.02,15.20,15.12,14.95,14.46,1
x = np.linspace(0, 34, 34)
x = [[x[i]] for i in range(len(x))]

model = LinearRegression()
model.fit(x,y)
y_hat = model.predict(x)
plt.scatter(x, y, alpha=0.5)
plt.plot(x, y_hat, 'r', alpha=0.5)
```

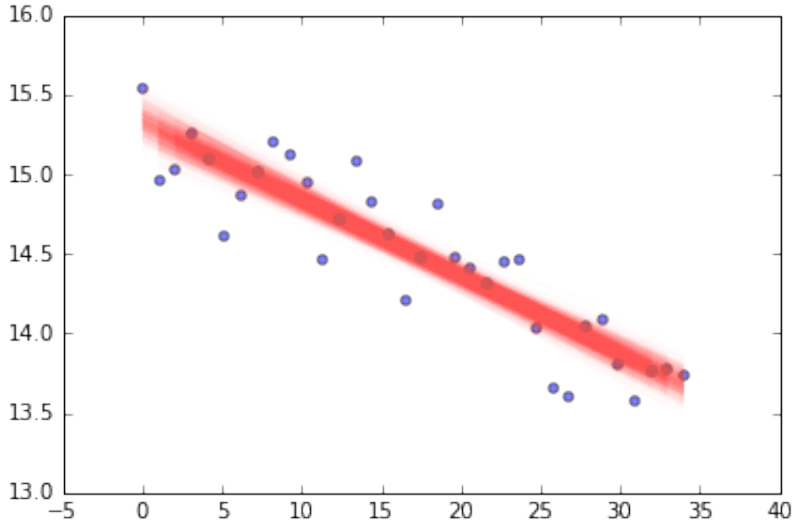
Altres aplicacions del bootstraping.



Altres aplicacions del bootstraping.

In [56]:
```
for i in range(1000):
    boot = np.random.randint(34, size=34)
    samplex = [x[_] for _ in boot]
    sampley = [y[_] for _ in boot]
    model.fit(samplex,sampley)
    y_hat = model.predict(samplex)
    plt.plot(samplex, y_hat, 'r', alpha=0.01)
```

Altres aplicacions del bootstraping.



Reflexions finals.

- L'estadística ens ajuda a quantificar l'incertesa d'un resultat, però el significat d'un resultat no depèn de les dades ni de la validació dels resultats, sinó de l'analista.
- La casualitat existeix: fins i tot quan l'estadística calcula una probabilitat petita per l'efecte observat sota la hipòtesis nula, el resultat pot ser no real!
- Sempre cal ser escèptic i no fer proposicions massa definitives sobre els resultats. La única forma d'augmentar la certesa sobre algun efecte és repetir l'experiment moltes vegades.