Statistical Thinking

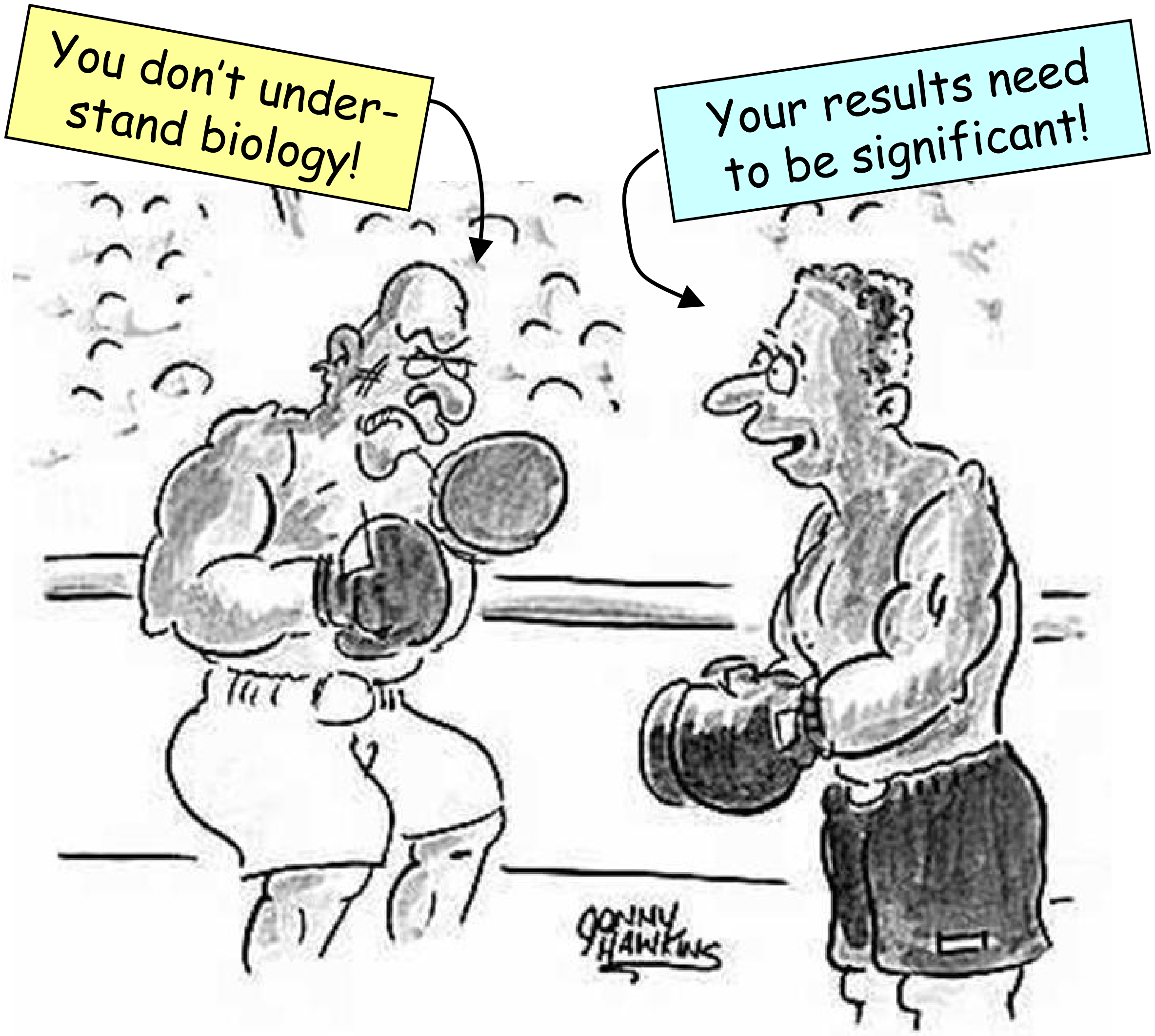
Jurg Ott

Laboratory of Statistical Genetics

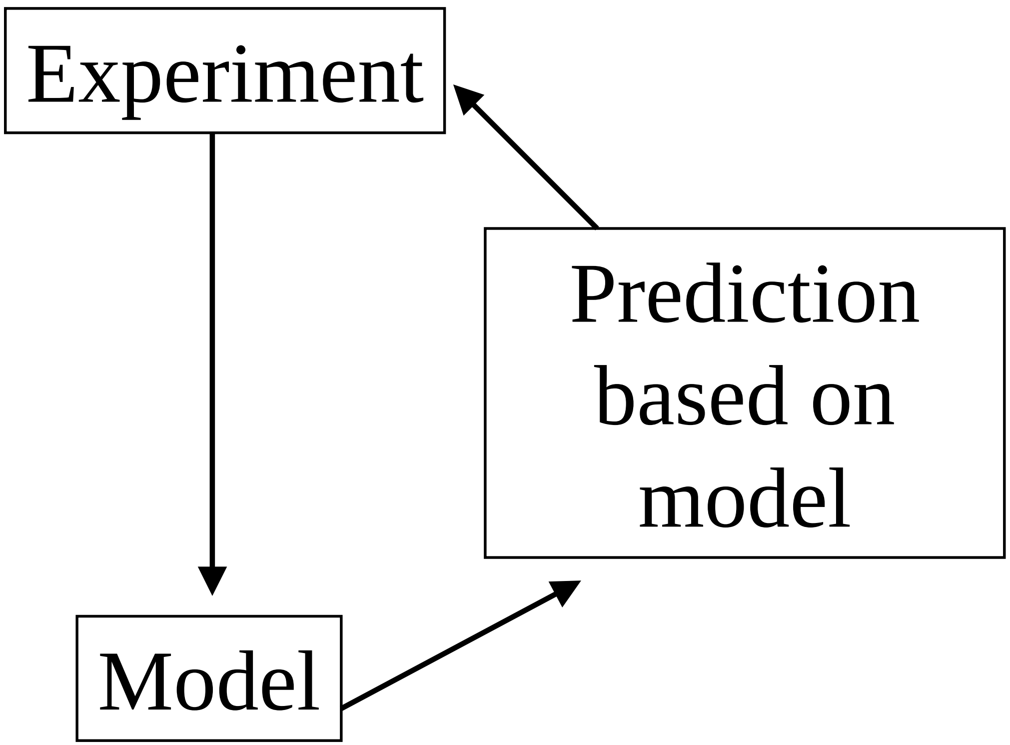
o

Ph.D. in Zoology (1967)

M.S. in Biomathematics (1972)


You don't under-
stand biology!
Your results need
to be significant!
JONNY HAWKINS

Working with Models



Statisticians also work with Models and “do” Experiments!

- **Elements of models: Random variables.** Important: Find good models describing biological phenomena accurately, making good predictions
- **Experiments**:
  - “Gedankenexperiment” – Think what happens given model assumptions
  - Computer simulation exhibits properties of models

Random Variables

- Random variable (RV) = variable that assumes different values with some defined probabilities (at random).
- Fully defined by totality of possible values and associated probabilities of occurrence = distribution of RV
- “Mind Children”, Hans Moravec, Harvard University Press, 1988

Two types of RVs

- Qualitative RV – discrete values, e.g., number of offspring
- Quantitative RV – infinite number of values possible within some range, e.g., expression level of a gene

X = number of years of age completed			
0	1	2	3…

- Underlying mechanism leading to given age distribution?
- Biologist thinks in terms of genes or hormones or … (longevity gene!)
- Epidemiologist thinks of nutrition …
- Statistician thinks is terms of a RV, a simple underlying “law”

The random variable X

Age class	New York
0-20	2,642,309
21-64	5,562,105
65+	1,109,821
Total	9,314,235
0-20	28%
21-64	60%
65+	12%
Total	100%

- X = number of individuals in given age class
- Estimation: Take sample and count
- Example: Population census, year 2000
- Underlying “law” with unknown but estimable age distribution

The random variable X

Age class	New York	Salt Lake City, Ogden
0-20	2,642,309	489,545
21-64	5,562,105	733,869
65+	1,109,821	110,500
Total	9,314,235	1,333,914
0-20	28%	37%
21-64	60%	55%
65+	12%	8%
Total	100%	100%

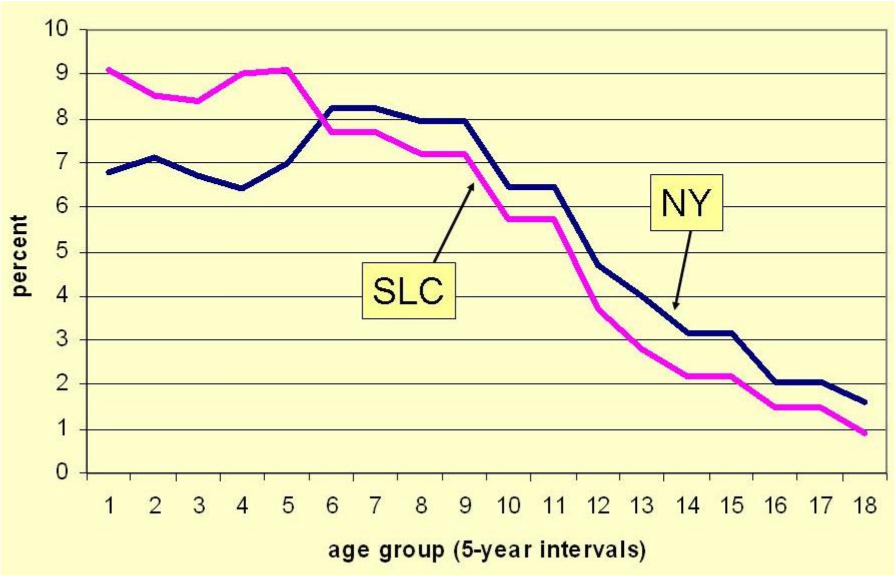
- Is underlying “force” leading to age distribution the same in NY and SLC?
- Assume “yes” and try to disprove this (null) hypothesis

Statistical Test

- Assume null hypothesis, H_0 (opposite of what you'd like to prove), and disprove it
- Why is reasoning so convoluted?
- Must make an effort to prove your case (effort = large enough sample size)

- Assume there is a difference and try to prove this is not true – simply take very small sample!

The concept of a statistic

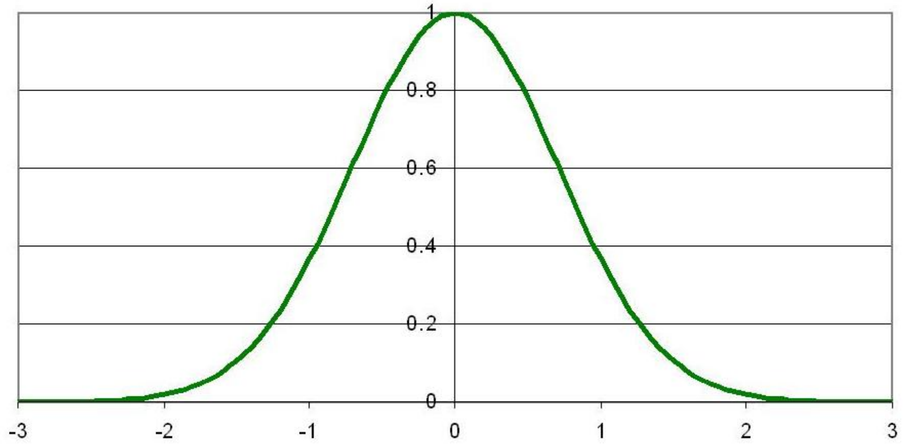


- Need quantity that summarizes the gist of what we want to show
- SLC has more young people, take mean (average) as our data summary for each city
- A statistic = function of the data that depends only on the data

- Mean depends on measurement unit (years, minutes, etc) (NY: 36 years, SLC: 31 years)
- Take scaled mean difference, t, as our quantity that characterizes the difference between the cities $\rightarrow$ t = test statistic

Working with a Test Statistic

- Test statistic = RV, known distribution if age distribution is the same (H_0 holds)
- Argument: Is the observed value of *t* compatible with what's expected under H_0?



- Find probability, *p*, that *t* is extreme if H_0 true $\rightarrow$ small *p* speaks against H_0

Predict-ions:

Poker

Public radio:
College students,
summer job

I'm feeling flu$h

BY STEPHEN HULL

POKER has come a long way from the days of playing in pub back rooms for a few pounds after hours.

Just ask Joseph Hachem – who has picked up £4.2million for winning the world's richest tournament.

Clutching a $50,000 bundle of cash, the 39-year-old kissed the money.

'A million dollars changes my life – let alone $7.5million,' he said. 'It changes everything.'

Hachem's win at the 36th World Series of Poker came after 5,619 players were whittled down to a final nine.

The Australian, a relative unknown, had already seen off some of the biggest names in the poker world.

He then had to endure a 14-hour gruelling test of nerves in the longest final table in the tournament's history.

After playing through the night, victory in the $10,000 buy-in No Limit Texas Hold 'em came as the sun rose over Los Angeles at 6.30am.

Playing against American Steven Dannenmann and with all $56 million in chips pushed into the middle of the table, he won with a seven-high straight to his opponent's pair of aces.

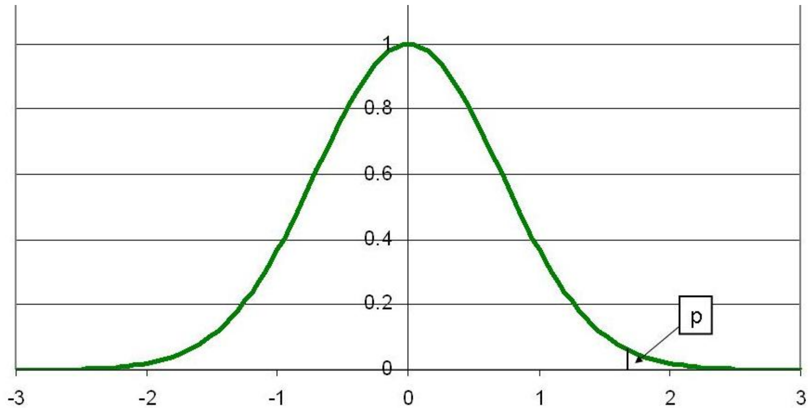
'Thank you, America,' Hachem shouted in delight.

The Lebanese-born professional moved to Australia with his family in 1972. He gave up a 13-year career as a chiropractor three years ago to play the game for a living.

It's all mine: Joseph Hachem kisses a stack of $100 dollar bills after his win

Picture: Reuters

Prediction: Gold Medal Winners

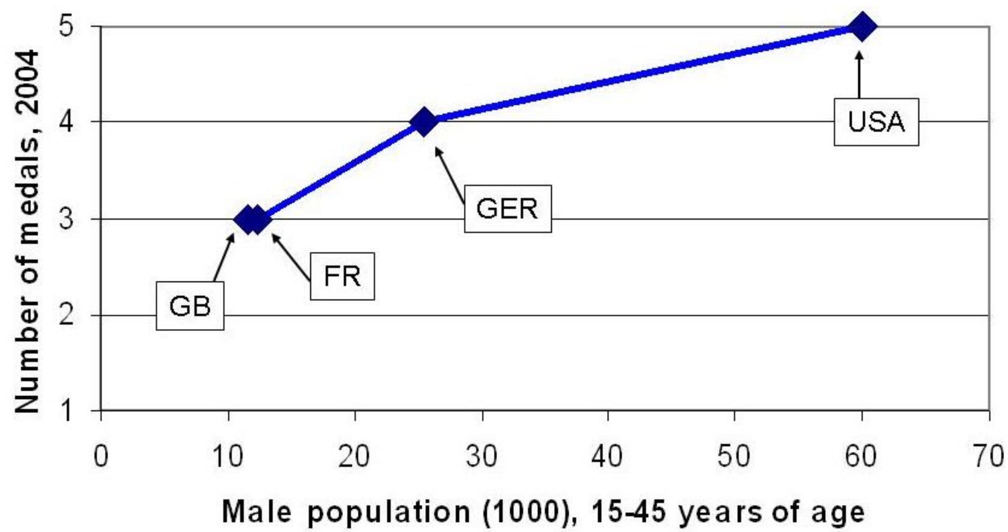


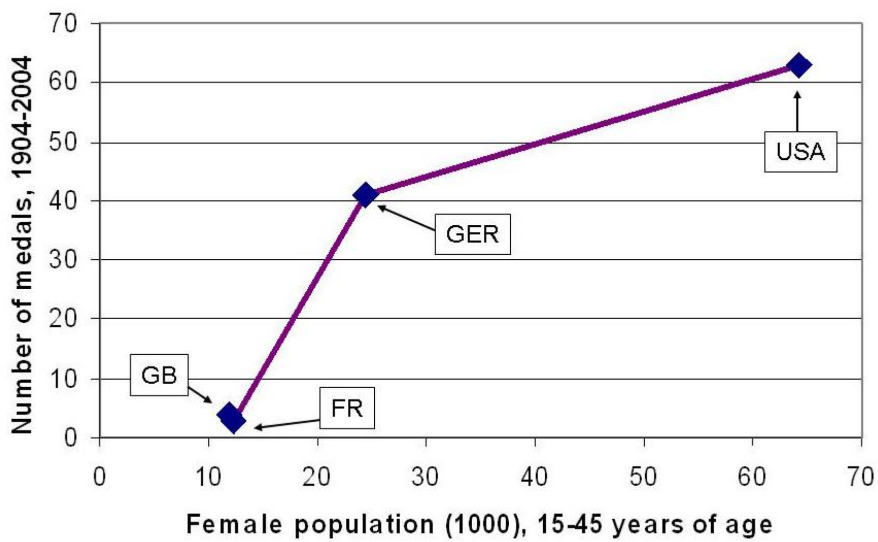
- **Olympic games**: US often picks up many gold medals. Determining factors?

- **H_0**: Size of country, i.e. population size, *n*
- **Model**: X = fitness, normally distributed; *p* = probability an individual has X > threshold to be able to win gold medal
- P(≥1 medal) = 1 – P(no medals) = $1 - (1 - p)^n \approx np$: Number of gold medals is proportional to population size if H_0 is true

Number of Gold Medal Winners

- Males, 2004
- Prediction of linearity quite accurate





- Females, 1904-2004
- Factors other than country size must also be important (high GDP)

Bernard & Busse (2004) *Revue of Economics and Statist* 86, 413-417

Two Camps of Statisticians

- **Likelihood** types (RA Fisher). Model for X = number of gold medal winners: $P(x) = e^{-\lambda}\lambda^{x}/x!$, $x = 0, 1, 2, \ldots$, with λ being a parameter, an unknown constant (here: mean).
- **Bayesian** concept. Parameters are RVs, with a prior distribution. Their estimation leads to their posterior distribution whose mean or mode is then taken as their best value.

Multiple Testing

- Each test: significance level, $\alpha = P(\text{sig}|H_0)$
- n tests, Prob(≥1 of them are significant), $p = 1 - (1 - \alpha)^n \approx n\alpha$.
- Bonferroni correction: Choose α such that, e.g., $p = 0.05 \rightarrow \alpha = p/n$
- Very conservative, particularly for dependent tests.
- Solutions: Randomization (permutation based) tests. FDR.

False Discovery Rate, FDR

Devlin et al. (2003); Storey & Tibshirani (2003) *PNAS* **100**, 9440

	Test not signif.	Test sig-nificant	# tests
H_0 true	U	V	m_0
H_0 false	T	S	m_1
	m - R	R	m

- Avg. significance level = V/m_0 (false pos.)
- Avg. FDR = V/R (need estimate)

Evaluating FDR (Benjamini & Hochberg)

SNP1	SNP2	p_i	rank,i	0.10*i/m
TP53-1_17p13	CBS-1_21q22	**0.000002**	1	**0.000026**
F2_11p11	HSPA1B_06p21	**0.000049**	2	**0.000052**
CETP-1_16q21	CBS-1_21q22	**0.000055**	3	**0.000078**
TP53-1_17p13	CBS-2_21q22	**0.000094**	4	**0.000104**
EDNRA_04q31	CBS-2_21q22	0.000180	5	0.000131
IL4RA_03p26	PON1-3_07q21	0.000367	6	0.000157
EDNRA_04q31	CBS-1_21q22	0.000433	7	0.000183
CD14_05q31	F2_11p11	0.000488	8	0.000209
SCNN1A-2_12p1	HSPA1B_06p21	0.000600	9	0.000235
TNFR1_12p13	NPPA-2_01p36	0.000713	10	0.000261
…	…	…	…	…
CBS-1_21q22	AGTR1_03q21	0.999957	3824	0.099896
CBS-1_21q22	NPPA-2_01p36	0.999990	3825	0.099922
EDN1_06p24	APOE-2_19q13	0.999997	3826	0.099948
NPPA-1_01p36	ITGA2B_17q21	0.999998	3827	0.099974
FCER1B_11q13	LDLR_19p13	0.999998	3828	0.10

For smallest *p*, FDR is The same as Bonferroni Corrected *p*-value.

Here, work with 0.10 sig. level

Benjamini et al. (2001) *Behavioral Brain Research* **125**, 279

The results of comparing 17 exploratory behavior measures between eight C57 and eighht BALB mice

Measure	Observed P-values	Rank (i)	Bonferroni threshold	FDR (BH
Lingering time (prop.)	0.000001	1	0.0029	0.0029
Lingering speed (cm/s)	0.000013	2	0.0029	0.0058
Early activity in move segments (m)	0.000065	3	0.0029	0.0088
Early activity (m)	0.00063	4	0.0029	0.0117
Spread of lingering (cm)	0.0008	5	0.0029	0.0147
Dynamics of activity	0.0017	6	**0.0029**	0.0176
Dynamics of diversity	0.0032	7	0.0029	0.0205
Number of excursions	0.0065	8	0.0029	0.0235
Movement speed (cm/s)	0.0148	9	0.0029	**0.0264**
Spread of move segments	0.049	10	0.0029	0.0294
Stops per excursions (upper quartile)	0.094	11	0.0029	0.0323
Center activity (prop.)	0.11	12	0.0029	0.0352
Center rest (prop.)	0.15	13	0.0029	0.0382
Activity (m)	0.24	14	0.0029	0.0411
Lingering activity (prop.)	0.45	15	0.0029	0.0441
Diversity	0.56	16	0.0029	0.047
Lingering at home base (prop.)	0.87	17	0.0029	0.05