WEO Research Workshop

Seoul, 17 November 2018





Biostatistician's role in study design

Dr Sunny H Wong

MBChB, DPhil, FRCPEd, FRCPath, FHKCP, FHKAM

Assistant Professor Institute of Digestive Disease The Chinese University of Hong Kong



Breath of biostatistics

- Descriptive statistics (data types, central tendency, dispersion, exploratory data analysis)
- Probability distributions & confidence intervals
- Hypothesis testing (null hypothesis, type I and II errors, sample size, power)
- Inferential statistics (t-test, chi-square, trend test, Fisher's test, log rank test, comparative data analysis)
- Correlation & regressions
- Multiple comparisons & corrections
- Survival analysis
- Meta-analysis
- Bayesian statistics
- Others (diagnosis, public health, bioinformatics)

Outline

- 1. Study design and power
- 2. Descriptive statistics
- 3. Inferential statistics
- 4. Software and tips

Study design and power

Concepts of statistics

 Science of (1) collecting and analyzing data to help make decisions in uncertainty, and (2) to make inference about a phenomenon observed





Types of clinical studies



Power calculation



Descriptive statistics

Types of variables

- Nominal categories (e.g. gender, ethnicity)
- Ordinal categories with a trend (e.g. cancer stage, grade)
- Numerical / scalar quantitative
 - Continuous scale (e.g. age, height)
 - Discrete scale (e.g. number of polyp)

Box-and-whisker plot



Bar chart

- Height is the mean
- How about error bars?
 - Standard deviation (SD)
 - Standard error of mean (SE / SEM)
 - Confidence interval (95% CI)

Histogram



Q: What are the differences between bar chart and histogram? (P.T.O.)

Bar chart and histogram



Histogram



Bonus Q: what is a Pareto chart?

Normal distribution



Gaussian distribution (z-statistics)

> de Moivre, Gauss & Laplace





- Histogram inspection
 - n>30
 - Fitting shape
- Quantile-quantile plot
- Formal tests
 - Komolgorov-Smirnov test
 - Shapiro Wilk test

QQ plot



3

Q: What is the hospital length-of-stay distribution? (right skewed distribution)

Inferential statistics

Sampling

e.g. SBP of the Hong Kong population



Sampling distribution



Central limit theorem



Distribution of the means



Student's t distribution

• Described by William Sealy Gosset

Resemble normal distribution if sample size is large (n>30)





Hypothesis testing

- Confirmatory data analysis to determine the probability that a given hypothesis is true
- Null hypothesis ' H_0 ': statement of no differences or association between variables
- Alternative hypothesis 'H₁': statement of differences or association between variables
- Type I (alpha) and Type II (beta) error

P-value



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Probability of mistakenly rejecting the null hypothesis (α)

Parametric tests

- Parametric tests assume data comes from a population
 - With known probability distribution (e.g. normal)



 Based on a fixed set of parameters (e.g. mean, SD)

Non-parametric tests usually less powerful (values discarded)

Comparative tests (interval)



A walk-through of rank sum test

Group A	Group B	Rank A	Rank B		Α	В
87	71	19	9	Total rank	127	148
72	42	10	1	Median	74	75.5
94	69	22	8	n	11	12
49	97	2	23			
56	78	4	14.5	U(A)	71	
88	84	20	17	U(B)	61	
74	57	12	5	U statistic	61	
61	64	6	7	<i>P</i> -value	0.76	
80	78	16	14.5			
52	73	3	11			
75	85	13	18			
	91		21			

Comparative tests (categorical)

- Examine differences between *observed* and *expected* counts
- Two assumptions
 - Independence of observations
 - Count of all cells >5



- Degree of freedom
 - *df* = (*R*-1) x (*C*-1)
 - No of free variables

	Drug A	Drug B	
Cured	20	10	30
Not cured	12	22	34
	32	32	64

χ^2 distribution and df

Degrees of

df = 1

df = 2 df = 3

Freedom 0.99 0.95 0.90 0.75 0.50 0.25 0.10 0.05 1 0.000 0.004 0.016 0.102 0.455 1.32 2.71 3.84 2 0.020 0.103 0.211 0.575 1.386 2.77 4.61 5.99 3 0 352 0 584 7.81 0 115 1 2 1 2 2 366 4 11 6.25 4 0.297 0.711 9.49 1.064 1.923 3.357 5.39 7.78 5 0.554 1.145 1.610 2.675 4.351 6.63 9.24 11.07 6 0.872 1.635 12.59 2.204 3.455 5.348 7.84 10.64 7 1.239 2.167 2.833 4.255 6.346 9.04 12.02 14.07 8 1.647 2,733 3,490 5.071 7.344 10.22 13.36 15.51 9 2 088 3.325 16.92 4.168 5.899 8.343 11.39 14.68 10 2.558 3.940 4.865 6.737 9.342 12.55 15.99 18.31 11 3.053 4.575 5.578 7.584 10.341 13.70 17.28 19.68 12 3.571 5.226 6.304 8.438 11.340 14.85 18.55 21.03 13 4.107 5.892 12.340 7.042 9.299 15.98 19.81 22.36 14 4.660 6.571 7,790 10.165 13.339 17.12 21.06 23.68 15 5.229 7.261 8.547 11.037 14.339 22.31 25.00 18.25 16 5.812 7.962 9.312 11.912 15.338 19.37 23.54 26.30 17 6.408 8.672 10.085 12.792 16.338 20.49 24.77 27.59 18 7.015 9.390 28.87 10.865 13.675 17.338 21.60 25.99 19 7.633 10.117 11.651 14.562 18.338 22.72 27.20 30.14 20 8.260 10.851 12.443 15.452 19.337 23.83 28.41 31.41

Probability of a larger value of x⁴

$$df = 5$$

$$df = 10$$

$$\chi^{2}$$

$$0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad 13 \quad 14 \quad 15$$

 $\chi_e^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

Correlation



Pearson's correlation coefficient (parametric)

Spearman's rank correlation coefficient (non-parametric)

Regression

- Estimating relationships between variables (between dependent and independent variables)
- Logistic regression
- Linear regression $\hat{Y} = bX + a$



Correlation/association ≠ causality



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

Q: What are the differences between correlation and regression?

Tips and software

Study design



- Most important is your research question
- Consider
 - Time
 - Effort
 - Infrastructure
 - Clinical ethics, governance and compliance

Power calculators



GPower: <u>http://www.gpower.hhu.de/</u> CUHK CCRB: <u>http://www2.ccrb.cuhk.edu.hk/web/</u> Clin Calc: <u>http://clincalc.com/stats/samplesize.aspx</u>

Statistical software



R Project for Statistical Computing

aun	maryStats.R	× HollywoodM	ovies2011 ×				
001	AIG	Source on Save	92.0			-+ Run 1	Source -
1 /	# loading	the data set					
2 1	Hollywood	Movies2011 <- red	ad.csv("HollywoodM	(ovies2011.csv")			
3							
4 1	# Calcula	iting summary stat	ts for all columns	of a data set			
5 5	summary(H	lollywoodMovies201	11)				
6							
7 1	# Calcula	iting summary stat	ts for an individu	ial column			
8 1	summary(H	lollywoodMovies201	L1\$RottenTomatoès))			
9							
10 1	# Calcula	iting specific sum	nmary statistics				
11							
12 1	# mean						
13 1	mean(Holl	ywoodMovies2011\$R	RottenTomatoes)				
14 1	mean(Holl	ywoodMovies2011\$P	RottenTomatoes, na	a.rm = TRUE)			
15							
16 #	# median						
17 1	median(Ho	llywoodMovies2011	LSRottenTomatoes,	na.rm = TRUE)			
17 18	median(Ho	llywoodMovies2011	LSRottenTomatoes,	na.rm = TRUE)			
17 1 18 10 4 8:43	# standar (Top Leve	d deviation	LSRottenTomatoes,	na.rm = TRUE)			R Scrip
17 1 18 10 4 8:43	median(Ho #_standar (Top Leve	d daviation) :	SRottenTomatoes,	na.rm = TRUE)			R Scrip
17 1 18 10 6 8:43	median(Ho ff_standor (Top Leve e ~/Docum	d daviation 0 0 nents/m107/data/ p	 SRottenTomatoes, SRottenTomatoes, 	na.rm = TRUE)	Freedoo Second	Waldara	R Scrip
17 1 18 10 8:43	median(Ho " ctondor (Top Leve e ~/Docum Genre	n lywoodMovies2011 d daviation 0 ° nents/m107/data/ ¢ TheatersOpenWeek	BOAverageOpenWee	<pre>na.rm = TRUE) k DomesticGross </pre>	ForeignGross	WorldGross	R Scrip
17 1 18 10 1 8:43	median(Ho f ctandor (Top Leve e ~/Docum Genre 1 :32	nents/m107/data/ p TheatersOpenWeek Min. : 3	BOAverageOpenWeel Min. : 1513	<pre>na.rm = TRUE) k DomesticGross Min. : 0.02 lat 0.02</pre>	ForeignGross Min. : 0.24	WorldGross Min. : 0.025	R Scrip
17 18 18 8:43 Console Action	median(Ho f standar (Top Leve e ~/Docum Genre 1 :32 7 :27 231	d deviation) = hents/m107/data/ / TheatersOpenWeek Min. : 3 1st Qu.:2550 Hodian : 2005	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779	<pre>k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median 27 36</pre>	ForeignGross Min. : 0.24 1st Qu.: 14.25	WorldGross Min. : 0.025 1st Qu.: 30.706 Medica 76 650	R Scrip
17 r 18 10 r 8:43 Console Action Comedy Drama	median(Ho f standar (Top Leve e ~/Docum Genre 1 :32 7 :27 :21 17	d deviation) = hents/m107/data/ p TheatersOpenWeek Min. : 3 1st Qu.:2550 Median :2925 Median :2828	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339	<pre>k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Median : 63.22</pre>	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mega : 46.92	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742	R Scrip
17 r 18 10 8:43 Console Action Comedy Drama Horror Thrill	median(Ho f standar (Top Leve e ~/Docum Genre 1 :32 7 :27 :21 .17 er :13	<pre>d deviation) d deviation) d ments/m107/data/ TheatersOpenWeek Min. : 3 1st Qu:2550 Median :2995 Median :2828 Srd Qu: 3400</pre>	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu : 8023	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Ou: 80.46	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.: 142.00	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu. 173.661	R Scrip
17 18 10 18 8:43 Console Action Comedy Drama Horror Thrill	median(Ho for standar (Top Leve e ~/Docum Genre 1 :32 / :27 :21 :17 .er :13 ion:12	<pre>illywoodMovies2011 id_deviation) = TheatersOpenWeek Min. : 3 1st Qu.:2550 Median :2995 Median :2985 Ard Qu.:3400 Max :4375</pre>	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu.: 8923 Max : 93230	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max : 381.01	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.:102.00 Max : 947.10	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu.: 173.691 Max : 1328.111	R Scrip
17 18 10 8:43 Console Action Comedy Drama Horror Thrill Animat (Other	median(Ho f standar (Top Leve e ~/Docum Genre 1 :32 / :27 :21 :17 :17 :er :13 :ion:12 · :14	llywoodMovies2011 d deviation) = TheatersOpenWeek Min. : 3 1st Qu.:2550 Median :2995 Mean :2828 3rd Qu.:3400 Max. :4375 NA's :16	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu.: 8923 Max. : 93230 NA's : 16	<pre>na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max. : 381.01 NA's '2</pre>	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.:102.00 Max. : 947.10 NA's : 15	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu.: 173.691 Max. :1328.111 NA's : 2	R Scrip
17 18 10 8:43 Console Action Comedy Drama Horror Thrill Animat (Other Bu	median(Ho f standar (Top Leve e ~/Docum Genre 1 :32 / :27 :21 . :17 er :13 .ion:12 .) :14 widget	d deviation) c hents/m107/data/ / TheatersOpenWeek Min. : 3 1st Qu.:2550 Median :2995 Mean :2828 3rd Qu.:3400 Max. :4375 NA's :16 Profitability	BOAverageOpenWeel Min. : 1513 Ist Qu.: 3779 Median : 5630 Mean : 8339 3rd Qu.: 8923 Max. : 93230 NA's :16 OpeningWeekend	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max. :381.01 NA's :2	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.:102.00 Max. : 947.10 NA's :15	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu.: 173.691 Max. : 1328.111 NA's :2	R Scrip
17 1 18 10 10 10 10 10 10 10 10 10 10	median(Hc f etandar (Top Leve e ~/Docum Genre 1 :32 / :27 :21 1 :17 er :13 :ion:12) :14 idget : 0.20	<pre>d deviation) : d deviation) : TheatersOpenWeek Min. : 3 1st Qu::2550 Median :2995 Median :2828 3rd Qu::3400 Max. :4375 NA's :16 Profitability Min. : 0.000</pre>	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu.: 8923 Max. :93230 NA's :16 OpeningWeekend Min. : 0.00	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max. :381.01 NA's :2	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.: 102.00 Max. : 947.10 NA's :15	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu.: 173.691 Max. :1328.111 NA's :2	R Scrip
17 r 18 10 10 10 10 10 10 10 10 10 10	median(Hc	<pre>illywoodMovies2011 id deviation) : TheatersOpenWeek Min. : 3 Ist Qu.:2550 Median :2995 Median :2995 Medan :2828 3rd Qu.:3400 Max. :4375 NA's :16 Profitability Min. : 0.000 Ist Ou.: 1.0655</pre>	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu.: 8923 Max. : 93230 NA's :16 OpeningWeekend Min. : 0.00 1st Ou.: 7.71	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max. : 381.01 NA's :2	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.: 102.00 Max. : 947.10 NA's :15	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu.: 173.691 Max. :1328.111 NA's :2	R Scrip
17 18 10 18 10 18 10 10 10 10 10 10 10 10 10 10	median(Hc	llywoodMovies2011 d deviation) = TheatersOpenWeek Min. : 3 1st Qu.:2550 Median :2995 Mean :2828 3rd Qu.:3400 Max. :4375 NA's :16 Profitability Min. : 0.000 1st Qu.: 1.065 Median : 2.199	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu.: 8923 Max. : 93230 NA's :16 OpeningWeekend Min. : 0.00 1st Qu.: 7.71 Median : 13.10	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max. :381.01 NA's :2	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.:102.00 Max. :947.10 NA's :15	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu.: 173.691 Max. :1328.111 NA's :2	R Scrip
17 18 10 18 10 18 10 18 10 10 10 10 10 10 10 10 10 10	median(Hc condense e ~/Docum Genre 1 :32 7 :27 :17 er :13 :ion:12) :14 idget : 0.20 .: 20.25 1 : 36.50 : 53.48	llywoodMovies2011 d deviation) = TheatersOpenWeek Min. : 3 1st Qu.:2550 Median :2995 Median :2995 Mean :2828 3rd Qu.:3400 Max. :4375 NA's :16 Profitability Min. : 0.000 1st Qu.: 1.065 Median : 2.199 Mean : 3.315	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu.: 8923 Max. : 93230 NA's :16 OpeningWeekend Min. : 0.00 1st Qu.: 7.71 Median : 13.10 Mean : 20.34	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max. :381.01 NA's :2	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.:102.00 Max. :947.10 NA's :15	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu.: 173.691 Max. :1328.111 NA's :2	R Scrip
17 18 10 18 10 18 10 10 10 10 10 10 10 10 10 10	median(Hc condition condition from Leve e ~/Docum Genre 1 :32 / :27 :21 :17 :17 :17 :17 :17 :17 :17 :21 :17 :21 :17 :21 :17 :21 :21 :17 :21 :21 :21 :21 :21 :21 :21 :21	llywoodMovies2011 d deviation): TheatersOpenWeek Min. : 3 1st Qu.:2550 Median :2995 Median :2995 Mean : 2828 3rd Qu.:3400 Max. :4375 NA's :16 Profitability Min. : 0.000 1st Qu.: 1.065 Median : 2.199 Mean : 3.315 3rd Qu.: 3.667	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu.: 8923 Max. : 93230 NA's :16 OpeningWeekend Min. : 0.00 1st Qu.: 7.71 Median : 13.10 Mean : 20.34 3rd Qu.: 25.00	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max. :381.01 NA's :2	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.:102.00 Max. : 947.10 NA's :15	WorldGross Min. : 0.025 1st Qu.: 30.706 Mean : 150.742 3rd Qu.: 173.691 Max. :1328.111 NA's :2	R Scrip
17 18 10 18 10 18 10 18 13 13 13 13 13 13 13 13 13 13	median(Hc f ctondor (Top Leve e -/Docun Genre 1 :22 7 :27 :21 :17 :17 :21 :17 :17 :21 :17 :17 :21 :17 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :17 :21 :14 :20 :20 :20 :20 :21 :21 :21 :21 :21 :21 :21 :21	llywoodMovies2011 d deviation) c TheatersOpenWeek Min. : 3 1st Qu.:2550 Median :2995 Median :2995 Max. :4375 NA's :16 Profitability Min. : 0.000 1st Qu.: 1.0655 Median : 2.199 Mean : 3.315 3rd Qu.: 3.667 Max. : 64.673	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu.: 8923 NA's :16 OpeningWeekend Min. : 0.00 1st Qu.: 7.71 Median : 13.10 Mean : 20.34 3rd Qu.: 25.00 Max. :169.19	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max. : 381.01 NA's :2	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.:102.00 Max. : 947.10 NA's :15	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu.: 173.691 Max. :1328.111 NA's :2	R Scrip
17 18 10 18 10 10 10 10 10 10 10 10 10 10	median(Hc construction) median(Hc constru	llywoodMovies2011 d deviation) TheatersOpenWeek Min. : 3 1st Qu.:2550 Median :2995 Mean :2828 3rd Qu.:3400 Max. :4375 NA's :16 Profitability Min. : 0.000 1st Qu.: 1.065 Median : 2.199 Mean : 3.315 3rd Qu.: 3.667 Max. :64.673 NA's :2	BOAverageOpenWeel Min. : 1513 1st Qu.: 3779 Median : 5686 Mean : 8339 3rd Qu.: 8923 Max. :93230 NA's :16 OpeningWeekend Min. : 0.00 1st Qu.: 7.71 Median : 13.10 Mean : 20.34 3rd Qu.: 25.00 Max. :169.19 NA's :3	na.rm = TRUE) k DomesticGross Min. : 0.02 1st Qu.: 19.03 Median : 37.35 Mean : 63.22 3rd Qu.: 80.46 Max. :381.01 NA's :2	ForeignGross Min. : 0.24 1st Qu.: 14.25 Median : 47.00 Mean : 96.92 3rd Qu.: 102.00 Max. : 947.10 NA's :15	WorldGross Min. : 0.025 1st Qu.: 30.706 Median : 76.659 Mean : 150.742 3rd Qu.: 173.691 Max. :1328.111 NA's :2	R Scrip

Statistical software





Questions?