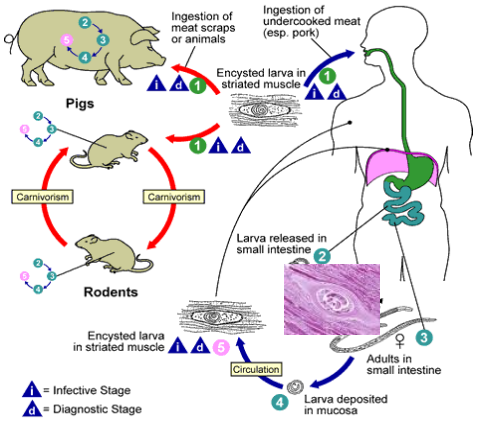
Analysis of data

Deterministic approach (SPSS)

María J. Vilar

Dep. of Food Hygiene and Environmental Health



Lecture
24. 04. 2013

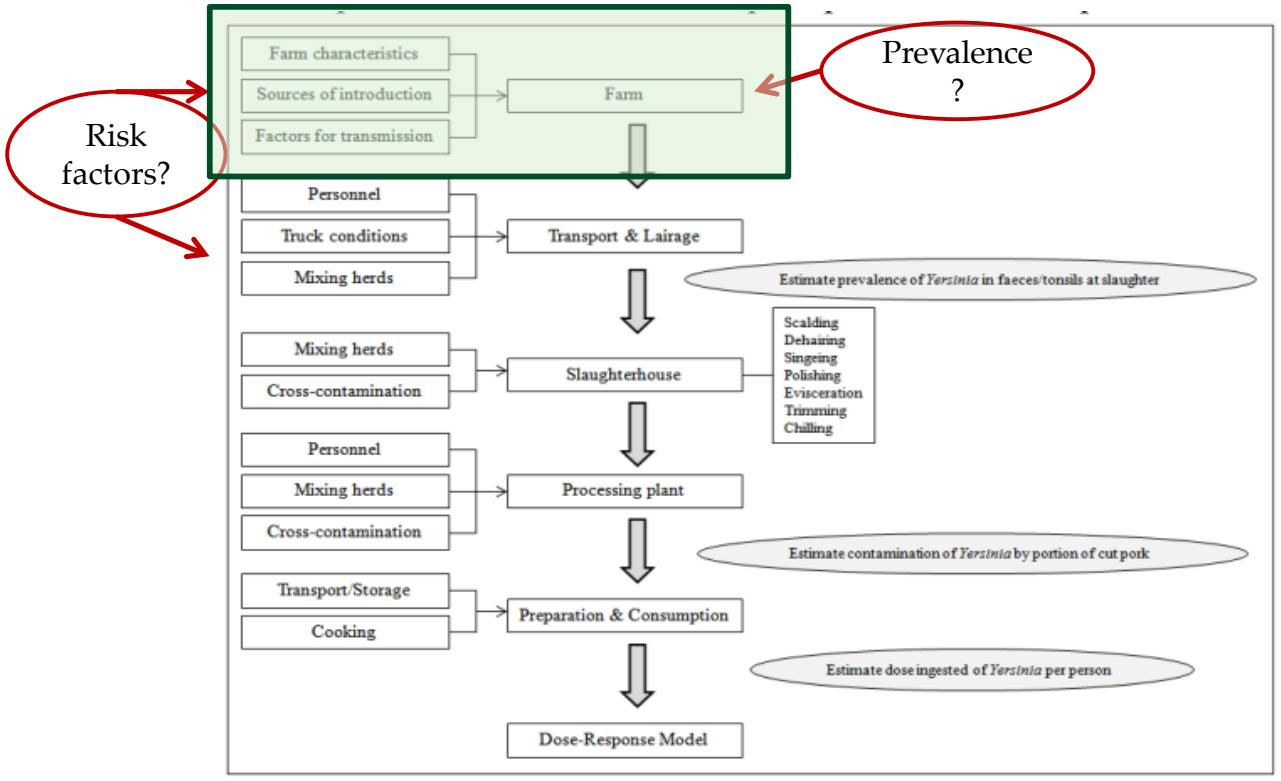
At the end of the course, the student should be able to:

- Differentiate between deterministic and stochastic analysis
- Identify how variables are associated
- Define and explain of statistical significance
- Classify the distributions (normal, binomial, Poisson)

Statistical analysis

Objective: identify those factors that cause disease/presence of a foodborne pathogen

Deterministic and Stochastic models



Associations between variables

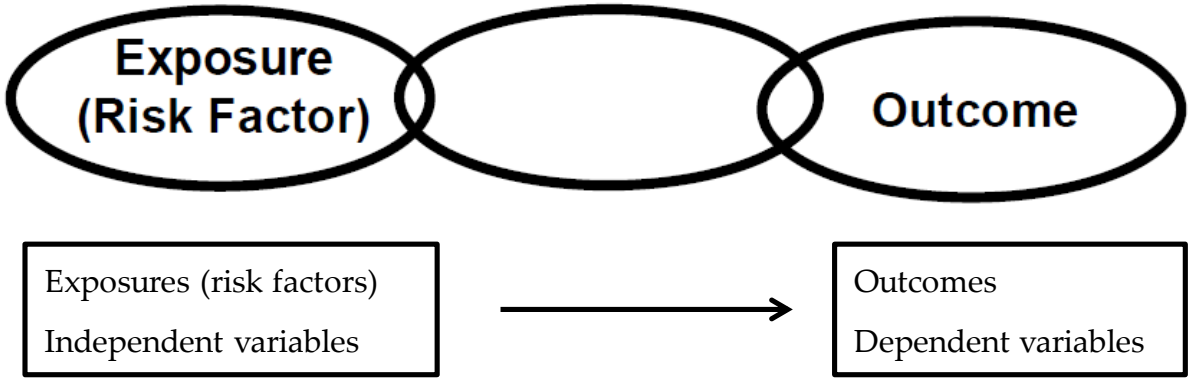
Discrete : dichotomous, categorical (nominal), ordinal

Continuous

Time to event

Variable: an observable event that can vary

A variable usually is affected (Response variable) by another (Explanatory variable)



Confounding

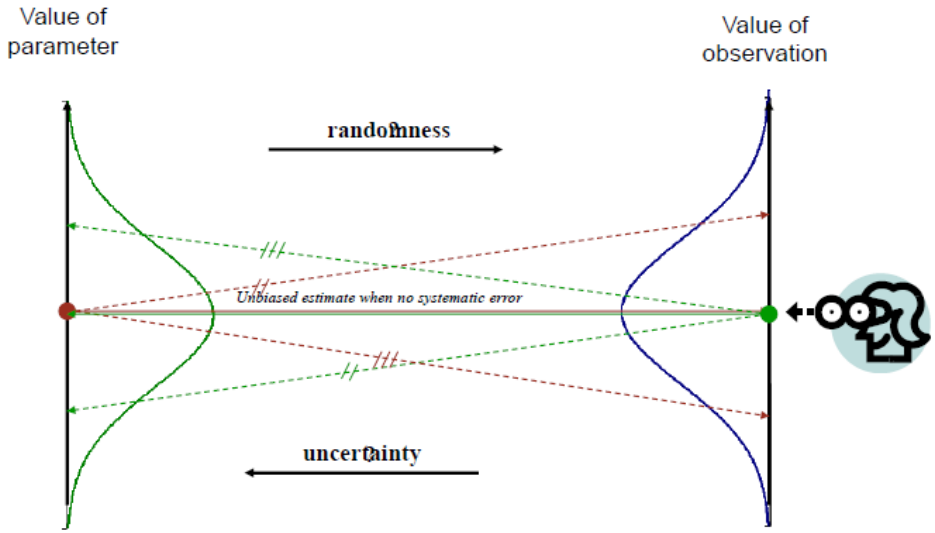
Random *vs* Uncertainty

Random or Variability: heterogeneity among individuals

Uncertainty: lack of knowledge

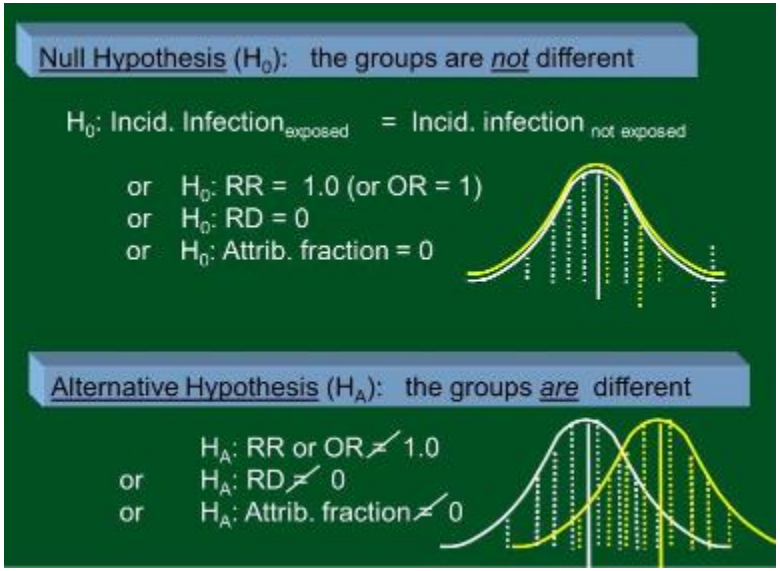
- model uncertainty
- parameter uncertainty

Impact: under or over estimation of output variance



Associations

Null hypothesis



Associations

Null hypothesis

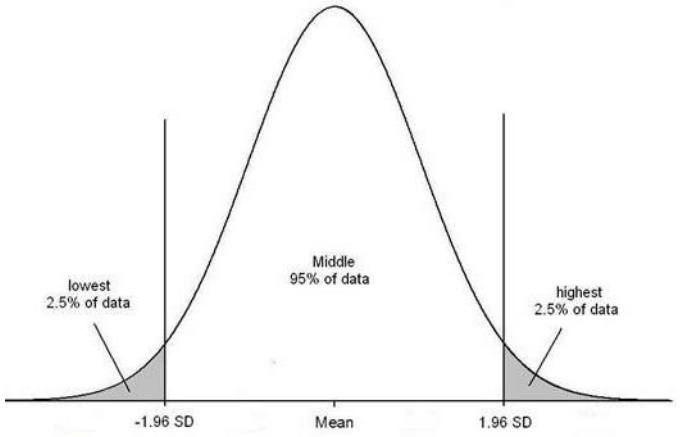
Statistical significance , p-values

usually significant test at $p<0.05$ (or $p<0.01$, $p<0.001$, ...)

Chi-square test

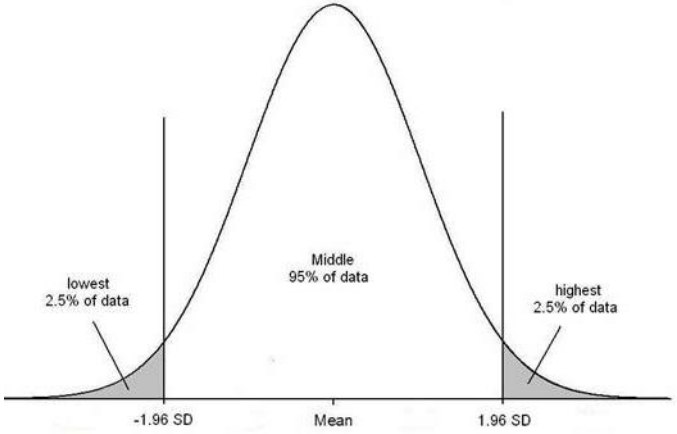
$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Fisher's exact test



Confidence intervals (CI)

A range around a point estimate within which the true value is likely to lie with a specified degree of probability, assuming there is no systematic error



Confidence intervals (CI)

CI of a mean

for n ≥ 30	for n < 30
$\bar{X} = Z\frac{s}{\sqrt{n}}$	$\bar{X} = t\frac{s}{\sqrt{n}}$

CI of a proportion

$$p \pm Z \times \sqrt{\frac{p(1-p)}{n}}$$

CI for two independent samples

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

if n1 > 30 and n2 > 30	if n1 < 30 and n2 < 30
$(\bar{X}_1 - \bar{X}_2) \pm Z\ SD\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	$(\bar{X}_1 - \bar{X}_2) \pm t\ SD\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
	df= $n_1 + n_2 - 2$

Confidence intervals (CI)

CI for two independent samples

Example

	Male			Female		
Characteristic	n_1	$\bar{X}_1$	s_1	n_2	$\bar{X}_2$	s_2
C.1	1623	128.2	17.5	1911	126.5	20.1
C.2	1622	75.6	9.8	1910	72.6	9.7
C.3	1544	192.4	35.2	1766	207.1	36.7
C.4	1612	194.0	33.8	1894	157.7	34.6
C.5	1545	68.9	2.7	1781	53.4	2.5
C.6	1545	28.8	4.6	1781	27.6	5.9

$$SD = \sqrt{\frac{(1623-1)17.5^2 + (1911-1)20.1^2}{1623+1911-2}}$$

$$(128.2 - 126.5) \pm 1.96 * 19 \sqrt{\frac{1}{1623} + \frac{1}{1911}}$$

(0.44 , 2.96)

Confidence intervals (CI)

CI for two independent samples

Example

Characteristic	n_1	$\bar{X}_1$	s_1	n_2	$\bar{X}_2$	s_2	Difference 95%CI
C.1	1623	128.2	17.5	1911	126.5	20.1	**(0.44, 2.96)**
C.2	1622	75.6	9.8	1910	72.6	9.7	**(2.38, 3.67)**
C.3	1544	192.4	35.2	1766	207.1	36.7	**(-17.16, -12.24)**
C.4	1612	194.0	33.8	1894	157.7	34.6	**(33.98, 38.53)**
C.5	1545	68.9	2.7	1781	53.4	2.5	**(5.31, 5.66)**
C.6	1545	28.8	4.6	1781	27.6	5.9	**(0.76, 1.48)**

Confidence intervals (CI)

CI of a mean

for n ≥ 30	for n < 30
$\bar{X} = Z\frac{s}{\sqrt{n}}$	$\bar{X} = t\frac{s}{\sqrt{n}}$

CI of a proportion

$$p \pm Z \times \sqrt{\frac{p(1-p)}{n}}$$

CI for two independent samples

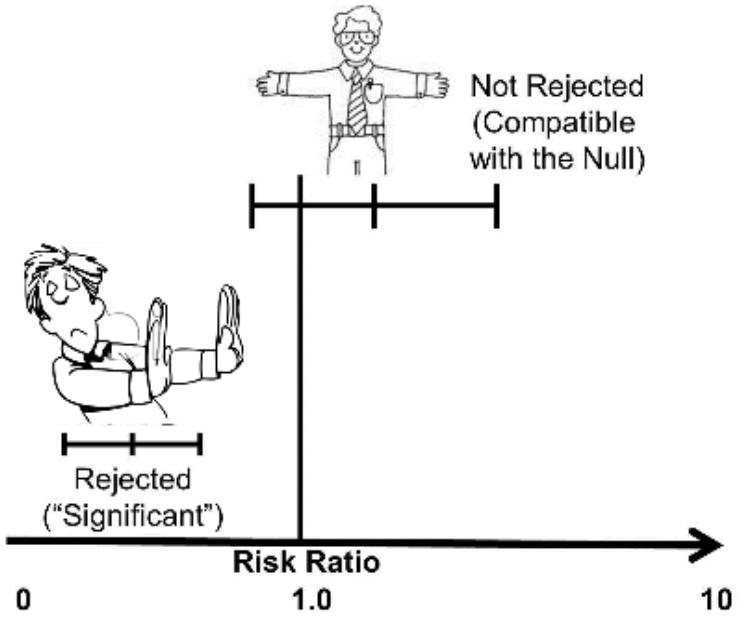
$$SD = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

if n1 > 30 and n2 > 30	if n1 < 30 and n2 < 30
$(\bar{X}_1 - \bar{X}_2) \pm Z\ SD\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	$(\bar{X}_1 - \bar{X}_2) \pm t\ SD\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
	df= $n_1 + n_2 - 2$

CI for two dependent samples

if n > 30	if n > 30
$\bar{X}_d \pm Z\frac{S_d}{\sqrt{n}}$	$\bar{X}_d \pm t\frac{S_d}{\sqrt{n}}$
	df= n-1

Statistical significance (p-values) and Confidence intervals



Associations

Null hypothesis

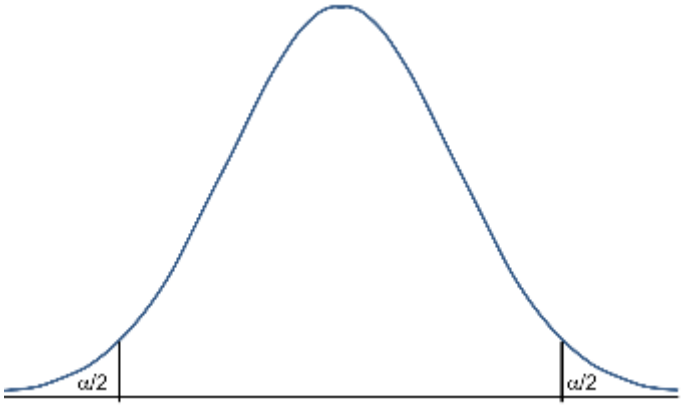
Statistical significance , p-values

Errors of inference

Type I error $\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$

Type II error $\beta = P(\text{type II error}) = P(\text{do not reject } H_0 \mid H_0 \text{ is false})$

Power



Associations

Null hypothesis

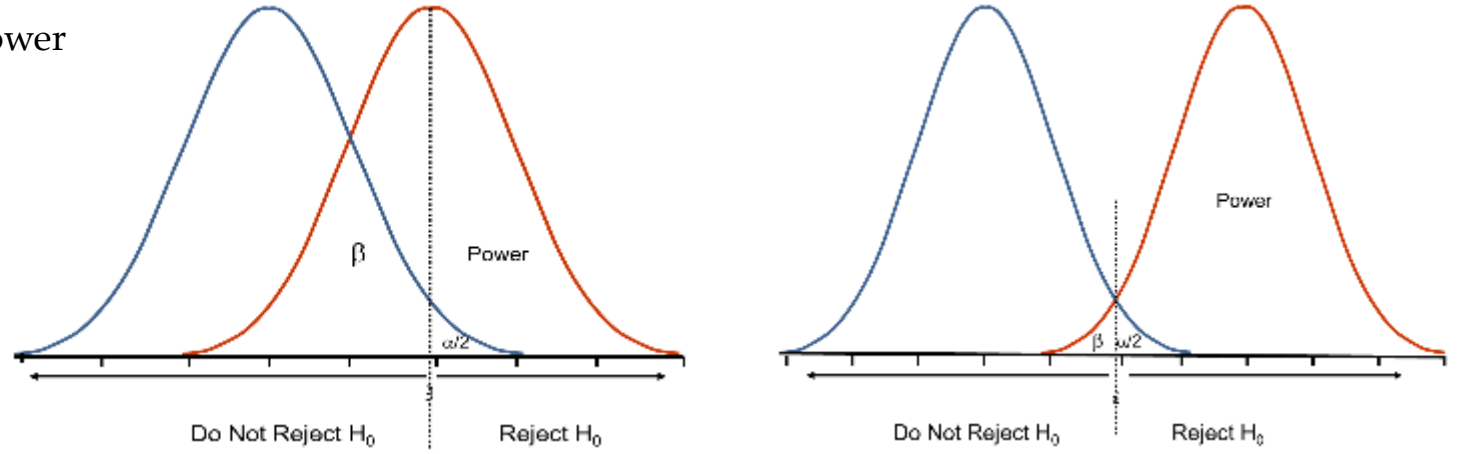
Statistical significance , p-values

Errors of inference

Type I error $\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$

Type II error $\beta = P(\text{type II error}) = P(\text{do not reject } H_0 \mid H_0 \text{ is false})$

Power



Associations

Null hypothesis

Statistical significance , p-values

Errors of inference

One- or Two tailed test

Independent and related samples

Parametric and non-parametric techniques

Measures of position (location)

Mean ($\bar{x}$)

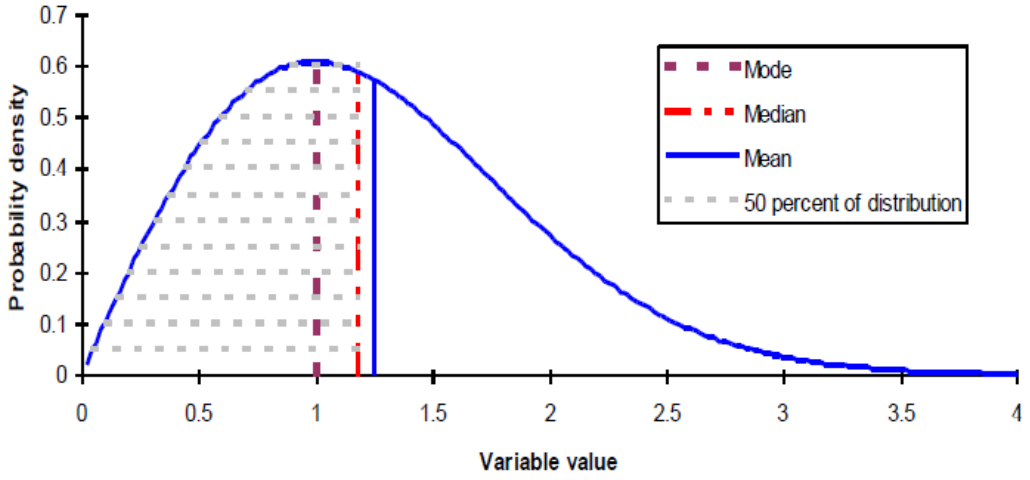
$$\mu = \sum_{i=1}^{n} x_i . p_i \qquad \text{for discrete distributions}$$

$$\mu = \int_{\min}^{\max} x . f(x) . dx \qquad \text{for continuous distributions}$$

$$\bar{x} = \frac{\sum x}{n}$$

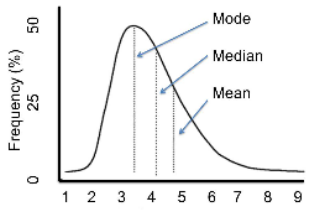
Median (Q_2)

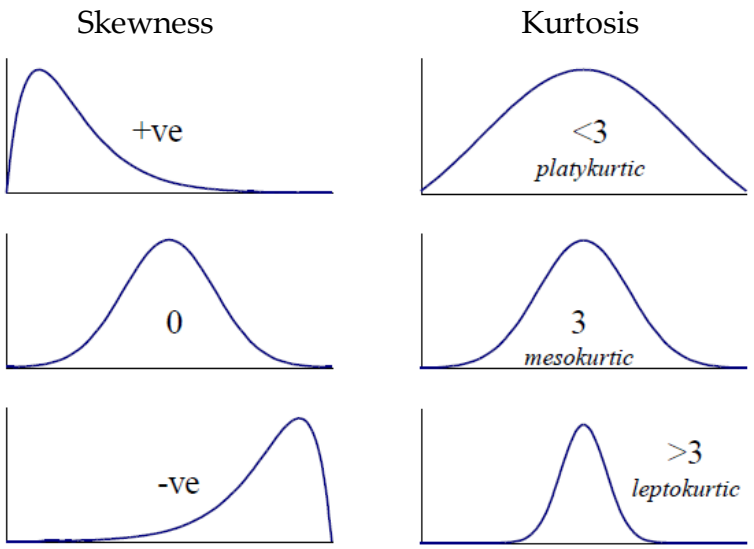
Mode



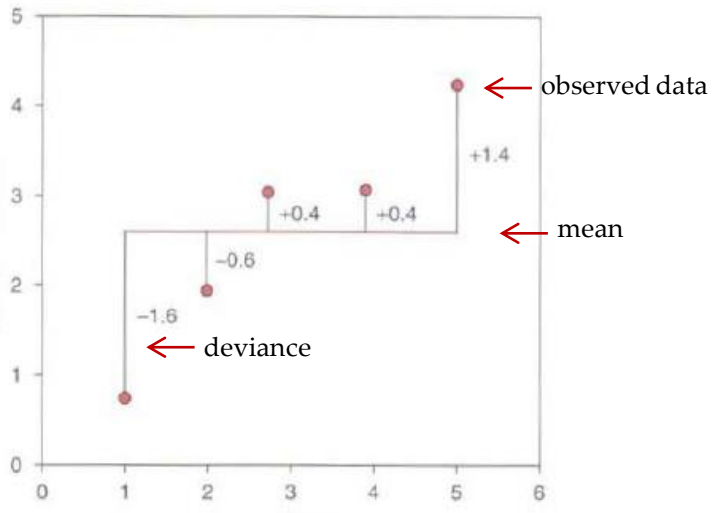
Measures of shape (of distributions)

	Discrete	Continuous
Skewness	$S = \frac{\sum_{i=1}^{n} (x_i - \mu)^3 . p_i}{\sigma^3}$	$S = \frac{\int_{\min}^{\max} (x - \mu)^3 . f(x) . dx}{\sigma^3}$
Kurtosis	$K = \frac{\sum_{i=1}^{n} (x_i - \mu)^4 . p_i}{\sigma^4}$	$K = \frac{\int_{\min}^{\max} (x - \mu)^4 . f(x) . dx}{\sigma^4}$





Measures of spread (deviation)



Variance $s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$

Standard deviation $\sigma^2 = \sqrt{s^2}$

Measures of spread (deviation)

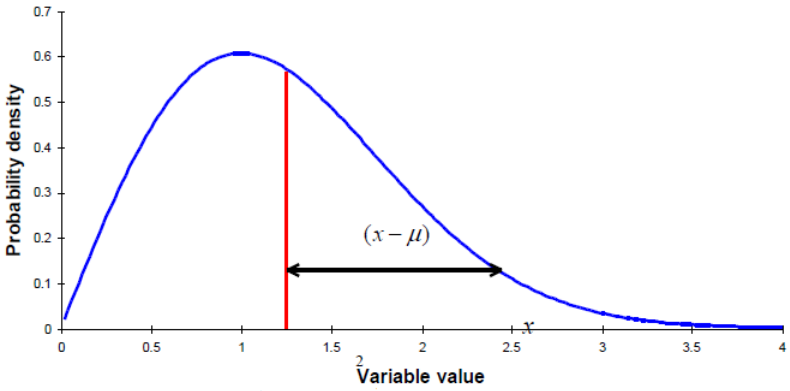
Variance (s^2)

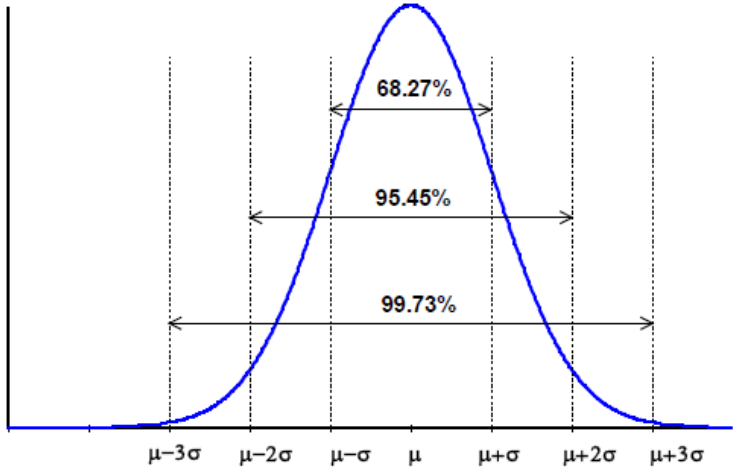
$$V = \sum_{i=1}^{n} (x_i - \mu)^2 . p_i \qquad \text{discrete}$$

$$V = \int_{\min}^{\max} (x - \mu)^2 . f(x).dx \qquad \text{continuous}$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Standard deviation (σ^2) $\quad \sigma^2 = \sqrt{s^2}$

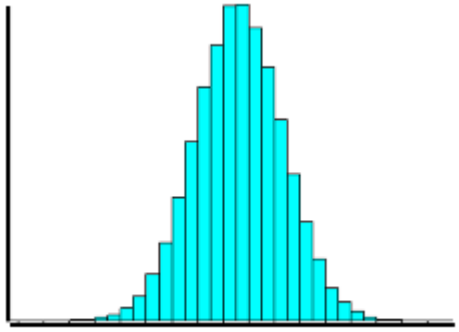




Why are important the measures of position, shape and spread?

- $Outcome_i = (model_i) + error_i$
- Indicates which distribution we have

Central Limit Theorem



Population mean $\mu = \frac{\sum X}{N}$ $\mu_{\bar{x}} = \mu$

Population standard deviation $\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem

For large samples, the distribution of the sample means is approx. normally distributed with a mean $\mu_{\bar{x}} = \mu$ and a standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

The Central Limit Theorem for large samples $Z = \dfrac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

For standard normal distribution, P(-1.96 <Z<1.96) = 0.95; there is a 95% probability that a standard normal variable Z will fall between -1,96 and 1.96

$\bar{x} \pm 1.96\,\sigma/\sqrt{n}$

The general confidence interval is

Point estimate ± Z x SD (point estimate)

$$\bar{x} = Normal\left(\mu, \sigma/\sqrt{n}\right)$$

$$\Sigma = n\bar{x} = Normal(n\mu, \sqrt{n}\sigma)$$

Central Limit Theorem

$$\bar{x} \pm 1.96\,\sigma/\sqrt{n}$$

The general confidence interval is

Point estimate ± Z x SD (point estimate)

Desired confidence interval	Z score
90%	1.645
95%	1.96
99%	2.576

t distribution

Point estimate ± t x SD (point estimate)

Central Limit Theorem

Example

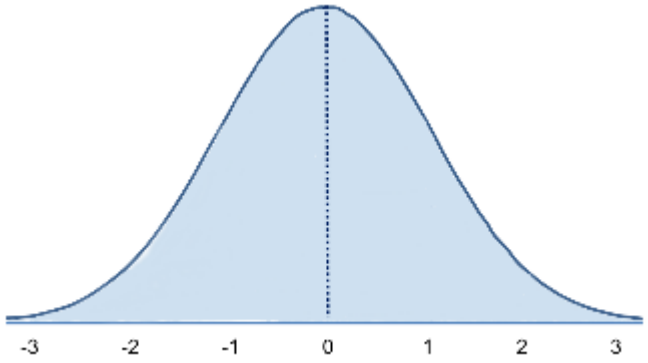
Data from the a study found that subjects over a specific age had a mean level of X = 54 and a standard deviation of 17. Suppose a researcher has 40 individuals over this specific age and wants to determine the probability that the mean levels of X for this sample of 40 individuals is 60 mg/dl or more (i.e., low risk). Probability questions about a sample mean can be addressed with the Central Limit Theorem, as long as the sample size is sufficiently large. In this case n=40, so the sample mean is likely to be approximately normally distributed, so we can compute the probability of levels of X>60 by using the standard normal distribution table.

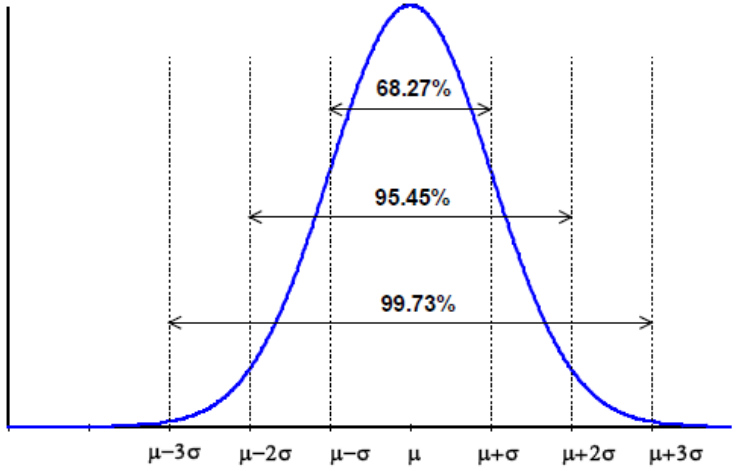
Solution

$z = \frac{60-54}{11/\sqrt{40}} = 2.22$ P (Z>2.22) = 1 – 0.9868 = 0.0132

The probability that the mean level in these 40 patients will exceed 60 is 1.32%.

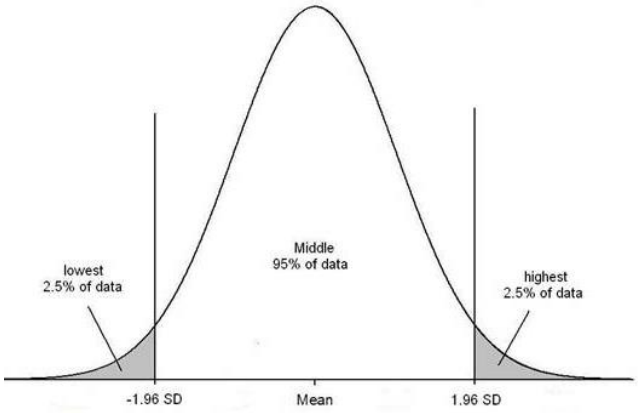
Normal distribution

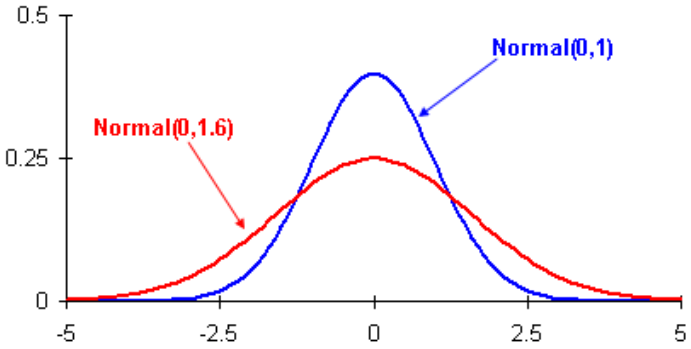




Normal distribution

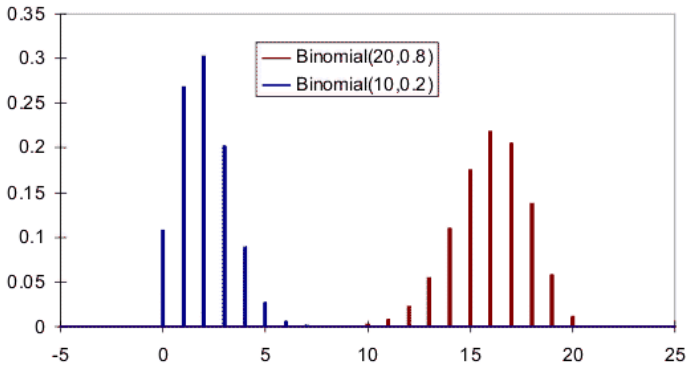
$$P(r) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$





Binomial distribution

$$Pr(r) = \frac{n!}{r!\,(n-r)!} p^r (1-p)^{n-r}$$



Mean $\mu = np$

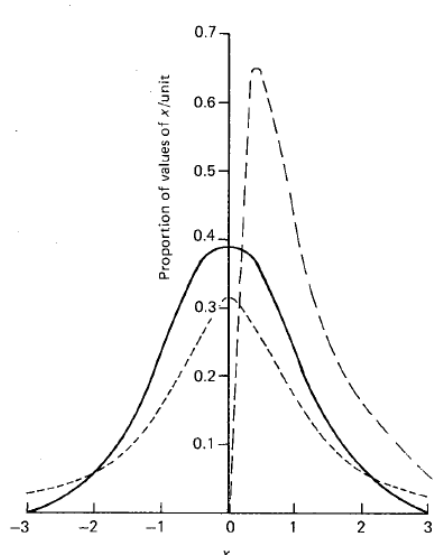
Standard deviation $\sigma = \sqrt{np(1-p)}$

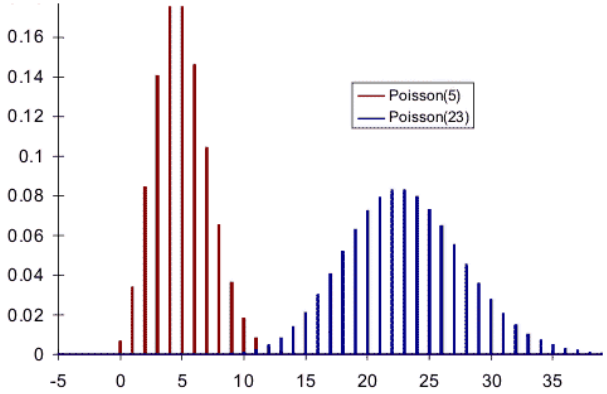
Assumptions

- Two possible outcomes (success or failure)
- Probability of success is the same for each trial
- Trials are independent

Poisson distribution

$$\Pr(r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$





Classical statistics (Deterministic, Frequentist)

A frequentist approach

Uses the information in the data

Software: SPSS, STATA, SAS, R, etc.

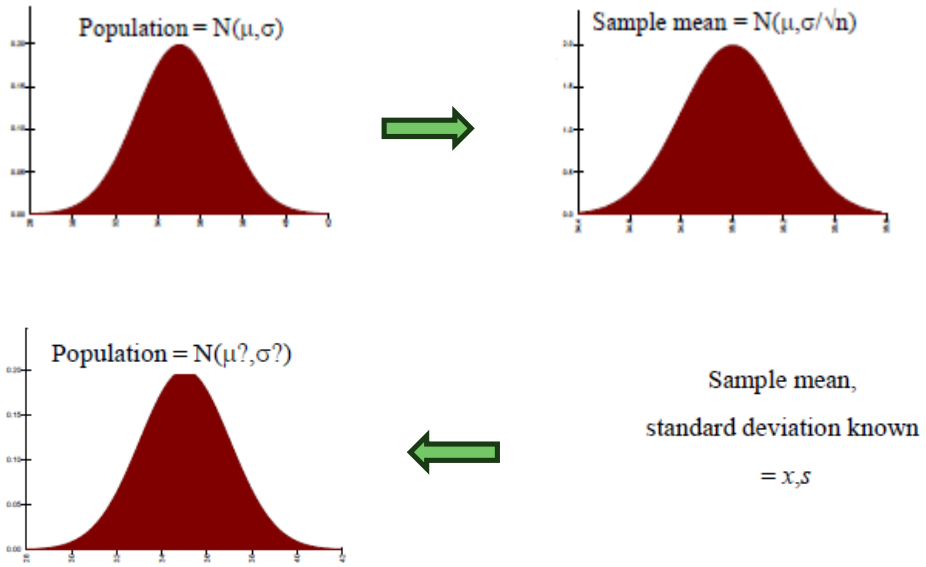


Table 14.1 Summary of some parametric statistical techniques for testing hypotheses relating to means of Normally distributed data.

Level of measurement	*Variance*	*Parametric statistical test*					*Parametric measure of association*
		One-sample case	*Two-sample case*		*Case with three or more samples*		
			Related samples	*Independent samples*	*Related samples*	*Independent samples*	
Interval and ratio	Known	Normal test	Normal test	Normal test			Correlation coefficient, ρ*
Interval and ratio	Unknown	*t*-test	*t*-test	*t*-test* (equal variance) Welch *t*-test* (unequal variances)	*F*-test (equal variance) Welch *F*-test (unequal variances)	*F*-test (equal variance) Welch *F*-test (unequal variances)	Correlation coefficient, ρ*

Classical statistics (Deterministic, Frequentist)

		Parametric test	Non-parametric test
Comparison of two means	Two experimental conditions and different participants	independent t-test	Wilcoxon rank-sum test or Mann-Whitney test
	Two experimental conditions and same participants	dependent t-test or repeated measure	Wilcoxon signed-rank test
Comparison of several means		ANOVA	Kruskal-Wallis test
		GLM: linear, multipe, logistic regression	

Categorical variables

Pearson's chi square test

Likelihood ratio (two variables)

Loglinear analyses (several variables)

Correlation

Pearson and Spearman

Level of measurement	*One-sample case*	*Non-parametric statistical test*				*Non-parametric measure of correlation*
		Two-sample case		*Case with three or more samples*		
		Related/matched samples	*Independent samples*	*Related samples*	*Independent samples*	
Nominal	Binomial test	McNemar change test	Fisher exact test for 2×2 contingency tables*	Cochran Q test	χ^2 test for $r \times k$ tables	Cramer coefficient, C Phi coefficient, r_ϕ
						The kappa coefficient of agreement, K*
	χ^2 goodness-of-fit test		χ^2 test for: 2×2 contingency tables* $r \times 2$ contingency tables			Asymmetrical association, the lambda statistic, L_B
Ordinal	Kolmogorov–Smirnov one-sample test, $D_{m,n}$	Sign test	Median test		Extension of the median test	Spearman rank-order correlation coefficient, r_s
	One-sample runs test	Wilcoxon signed ranks test, T^{+}**	Wilcoxon–Mann–Whitney test, W_x^*	Friedman two-way analysis of variance by ranks, F_r	Kruskal–Wallis one-way analysis of variance, KW	Kendall rank-order correlation coefficient, T
	Change-point test		Robust rank-order test, U	Page test for ordered alternatives, L	Jonckheere test for ordered alternatives, J	Kendall partial rank-order correlation coefficient, $T_{xy,z}$
			Kolmogorov–Smirnov two-sample test, $D_{m,n}$			Kendall coefficient of concordance, W
			Siegel–Tukey test for scale differences			Kendall coefficient of agreement, U
						Correlation between k judges and a criterion, T_c
Interval and ratio	Test for distributional symmetry	Permutation test for paired replicates	Permutation test for two independent samples			Gamma statistic, G
			Moses rank-like test for scale differences			Somers' index of asymmetric association, d_{BA}

References

Field, A. Discovering statistics using SPSS. 2nd edn. SAGE pulbications., 2005.

Rosner B. Fundamentals of Biostatistics. Belmont, CA: Duxbury-Brooks/Cole; 2006.

IBM SPSS Statistics Information Center

http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Fidh_glmm.htm

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/.