CS4220 Knowledge Discovery Methods for Bioinformatics

Unit 1a: Essence of Knowledge Discovery (Part A: Basic Statistics)

Wong Limsoon

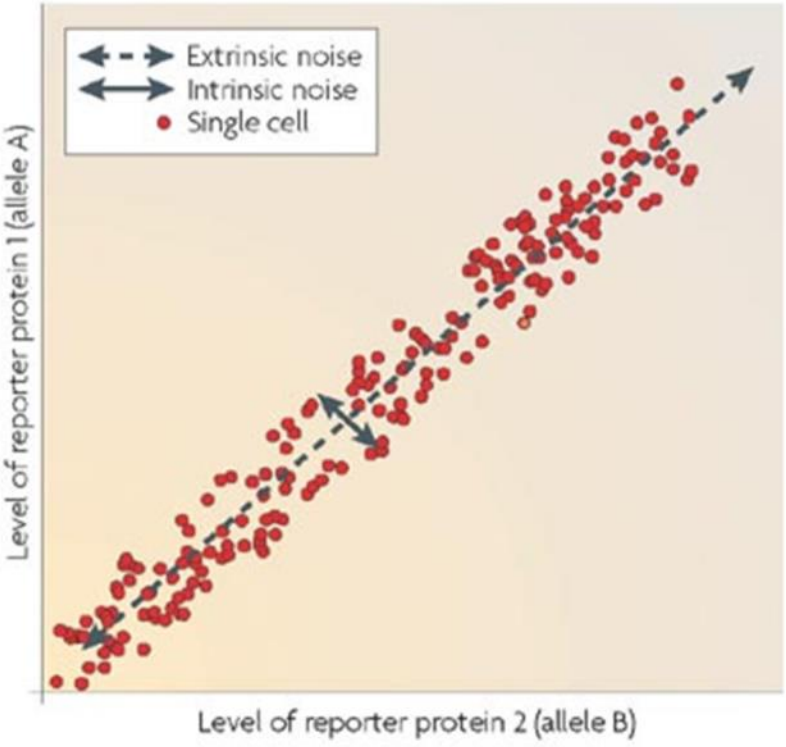
NUS National University of Singapore

Outline

- **Basics of biostatistics**
- **Statistical estimation**
- **Hypothesis testing**
  - Measurement data: z-test, t-test
  - Categorical data: χ2-test, Fisher's exact test
  - Non-parametric methods
- **Ranking and rating**
- **Principal component analysis**
- **Summary**

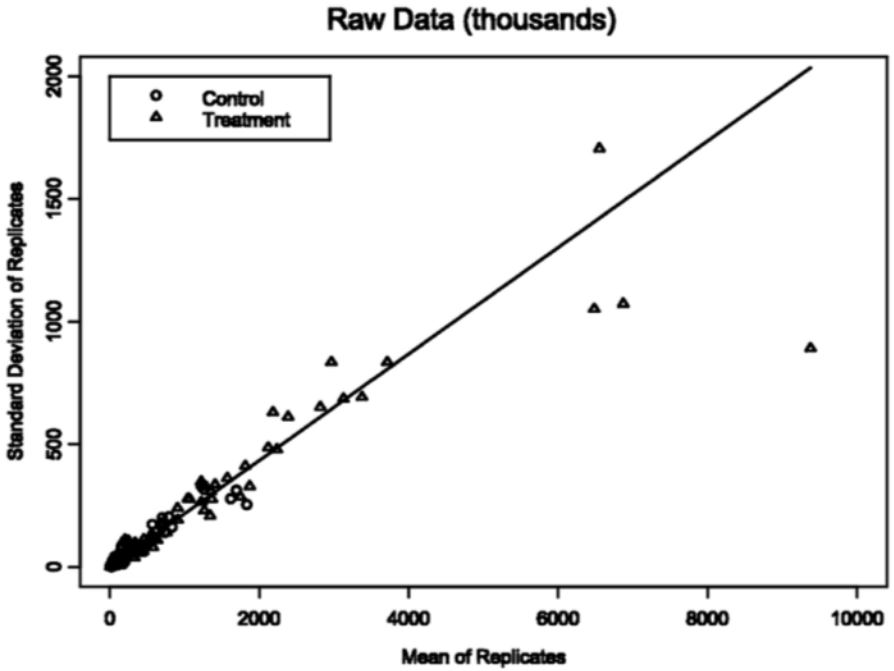
Why need biostatistics?

Intrinsic & extrinsic noise



Nat Rev Genet, 9:583-593, 2008

Measurement errors



J Comput Biol, 8(6):557-569, 2001

Why need to learn biostatistics?

- **Essential for scientific method of investigation**
  - Formulate hypothesis
  - Design study to objectively test hypothesis
  - Collect reliable and unbiased data
  - Process and evaluate data rigorously
  - Interpret and draw appropriate conclusions

- **Essential for understanding, appraisal and critique of scientific literature**

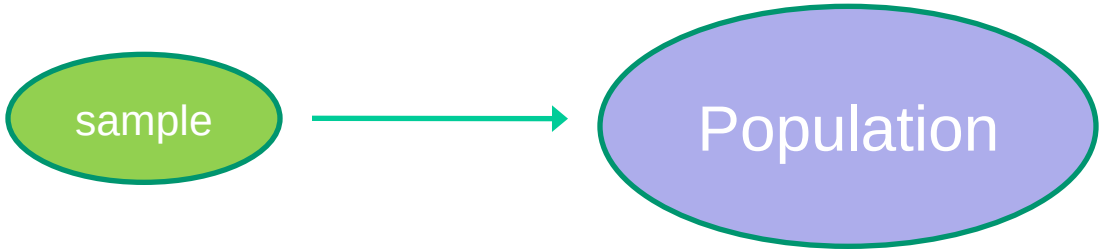
Type of statistical variables

- **Descriptive (categorical) variables**
  - Nominal variables (no order between values): gender, eye color, race group, …
  - Ordinal variables (inherent order among values): response to treatment: none, slow, moderate, fast

- **Measurement variables**
  - Continuous measurement variable: height, weight, blood pressure …
  - Discrete measurement variable (values are integers): number of siblings, the number of times a person has been admitted to a hospital …

Types of statistical methods

- **Descriptive statistical methods**
  - Provide summary indices for a given data, e.g. arithmetic mean, median, standard deviation, coefficient of variation, etc.
- **Inductive (inferential) statistical methods**
  - Produce statistical inferences about a population based on information from a sample derived from the population, need to take variation into account



Estimating population values from sample values

Summarizing data

- **Statistic is "making sense of data"**
- **Raw data have to be processed and summarized before one can make sense of data**
- **Summary can take the form of**
  - Summary index: using a single value to summarize data from a study variable
  - Tables
  - Diagrams

Summarizing categorical data

patient	gender	status
1	Male	alive
2	female	alive
3	male	dead
4	female	alive
etc	etc	etc

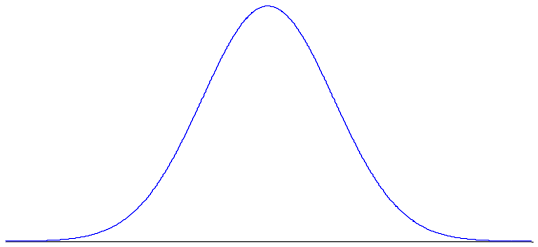
	Dead	Alive	Total
Female	12	25	37
male	23	26	49
Total	35	51	86

- **Proportion is a fraction & the numerator is a subset of the denominator**
  - proportion dead = 35/86 = 0.41
- **Odds are fractions where the numerator is not part of the denominator**
  - Odds in favor of death = 35/51 = 0.69
- **Ratio is a comparison of two numbers**
  - ratio of dead: alive = 35: 51
- **Odds ratio: commonly used in case-control study**
  - Odds in favor of death for females = 12/25 = 0.48
  - Odds in favor of death for males = 23/26 = 0.88
  - Odds ratio = 0.88/0.48 = 1.84

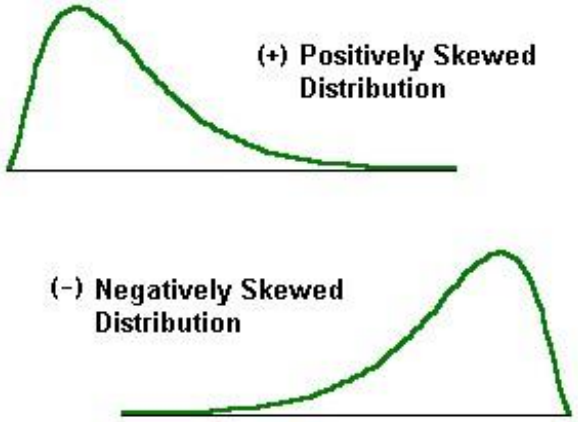
Summarizing measurement data

- **Distribution patterns**
  - Symmetrical (bell-shaped) distribution
    - **e.g. normal distribution**
  - Skewed distribution
  - Bimodal and multimodal distribution
- **Indices of central tendency**
  - Mean, median
- **Indices of dispersion**
  - Variance, standard deviation, coefficient of variance

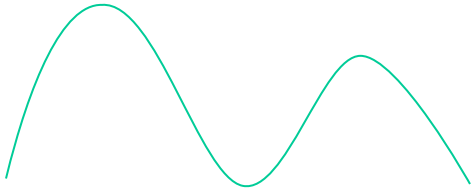
Distribution patterns



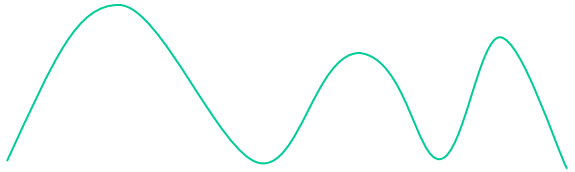
symmetrical distribution



Skewed distribution



Bimodal



Multimodal

Indices of central tendency

- **(Arithmetic) mean: Average of a set of values**

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Mean is sensitive to extreme values**
- **Example: blood pressure reading**

x1	87	87
x2	95	95
x3	98	98
x4	101	101
x5	105.0	1050
mean	97.2	286.2

Robust measure of central tendency

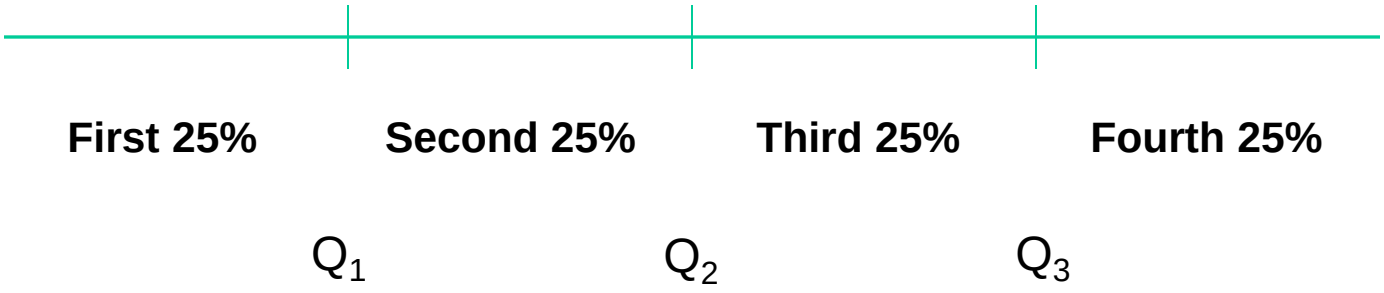
- **Median: The number separating the higher half of a sample, a population, or a population from the lower half**

- **Median is less sensitive to extreme values**

	x1	87	87
	x2	95	95
Median is unchanged →	x3	**98**	**98**
	x4	101	101
	x5	105.0	1050

Indices of central tendency: Quantiles

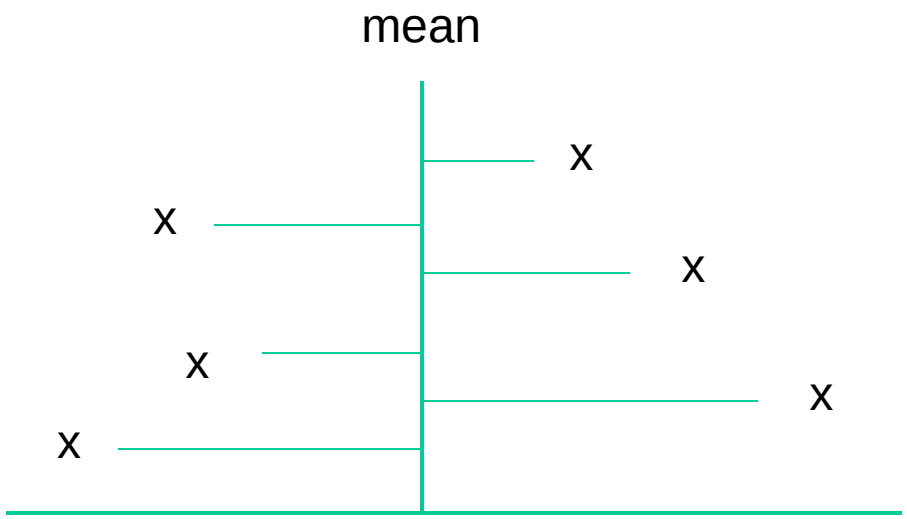
- **Quantiles: Dividing the distribution of ordered values into equal-sized parts**
  - Quartiles: 4 equal parts
  - Deciles: 10 equal parts
  - Percentiles: 100 equal parts



Q_1: first quartile
Q_2 : second quartile = median
Q_3: third quartile

Indices of dispersion

- **Summarize the dispersion of individual values from some central value like the mean**
- **Give a measure of variation**



Indices of dispersion: Variance

- **Variance : Average of squares of deviation from the mean**

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

- **Variance of a sample: Usually subtract 1 from n in the denominator**

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

effective sample size, also called degree of freedom

Indices of dispersion: Standard deviation

- **Problem with variance: Awkward unit of measurement as value are squared**
- **Solution: Take square root of variance => standard deviation**
- **Sample standard deviation (s or sd)**

$$\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Standard deviation

- **Caution must be exercised when using standard deviation as a comparative index of dispersion**

Weights of newborn elephants (kg)	
929	853
878	939
895	972
937	841
801	826

n=10, $\bar{X}$ = 887.1, sd = 56.50

Weights of newborn mice (kg)	
0.72	0.42
0.63	0.31
0.59	0.38
0.79	0.96
1.06	0.89

n=10, $\bar{X}$ = 0.68, sd = 0.255

It is incorrect to say that elephants show greater variation for birth-weights than mice because of higher standard deviation

Coefficient of variance

- **Coefficient of variance expresses standard deviation relative to its mean**

$$cv = \frac{s}{\bar{X}}$$

Weights of newborn elephants (kg)	
929	853
878	939
895	972
937	841
801	826

n=10, $\bar{X}$ = 887.1
s = 56.50, cv = 0.0637

Weights of newborn mice (kg)	
0.72	0.42
0.63	0.31
0.59	0.38
0.79	0.96
1.06	0.89

n=10, $\bar{X}$ = 0.68
s = 0.255, cv = 0.375

Mice show greater birth-weight variation

When to use coefficient of variance

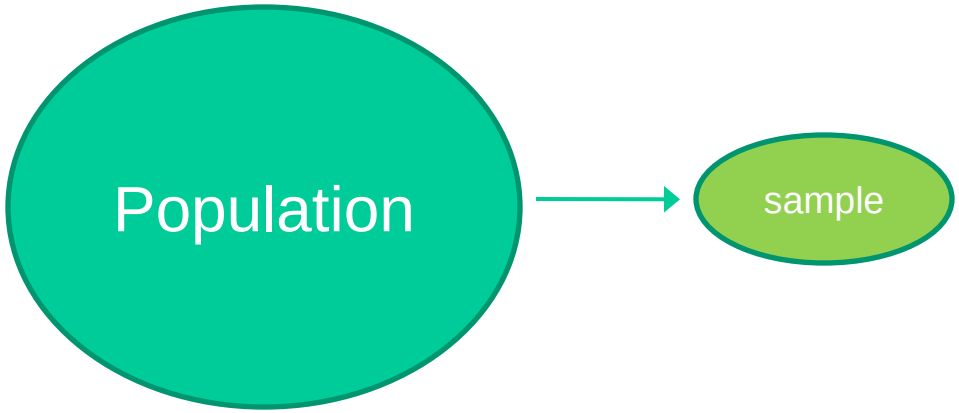
- **When comparison groups have very different means (CV is suitable as it expresses the standard deviation relative to its corresponding mean)**

- **When different units of measurements are involved, e.g. group 1 unit is mm, and group 2 unit is mg (CV is suitable for comparison as it is unit-free)**

- **In such cases, standard deviation should not be used for comparison**

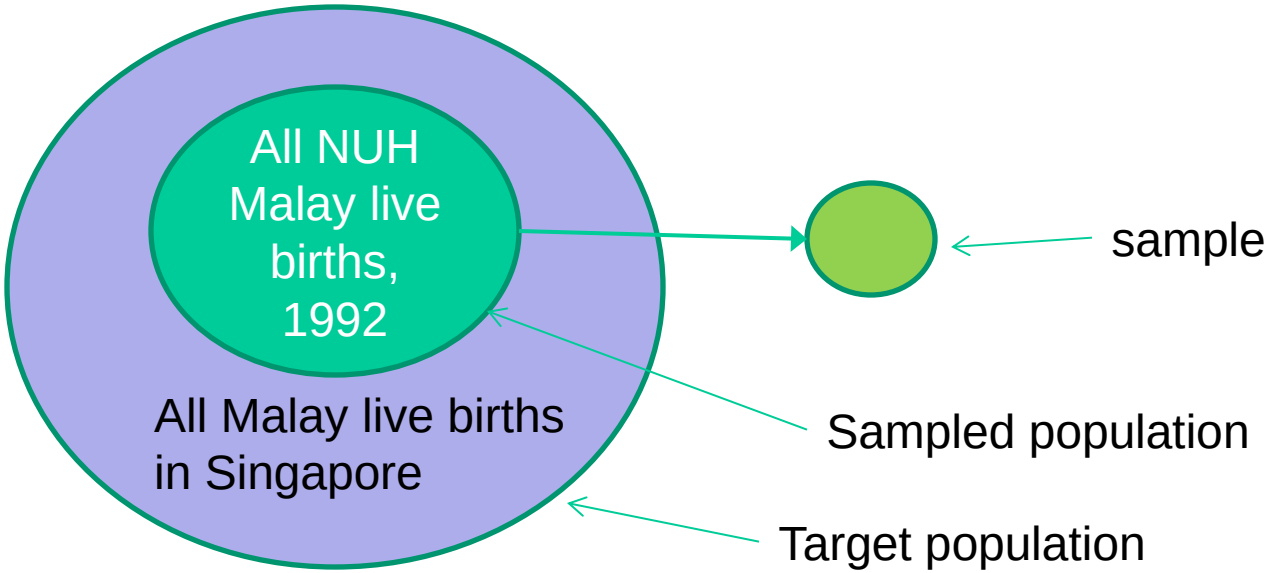
Sample and population

- **Populations are rarely studied because of logistical, financial and other considerations**
- **Researchers have to rely on study samples**
- **Many types of sampling design**
- **Most common is simple random sampling**

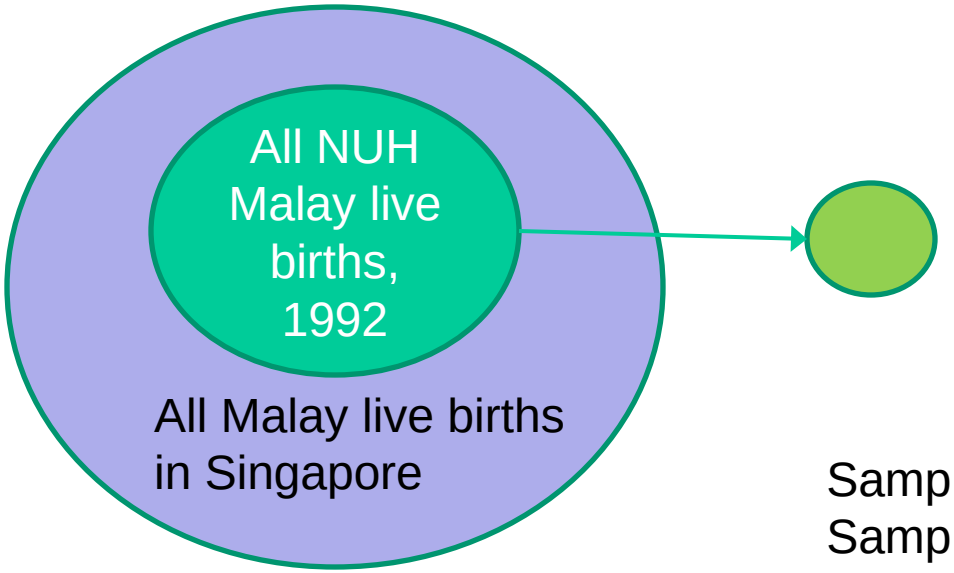


Random sampling

- **Suppose that we want to estimate the mean birth-weights of Malay male live births in Singapore**
- **Due to logistical constraints, we decide to take a random sample of 100 Malay live births at the National University Hospital in a given year**



Sample, sampled population, and target population



random sample of 100 Malay live births at NUH, 1992

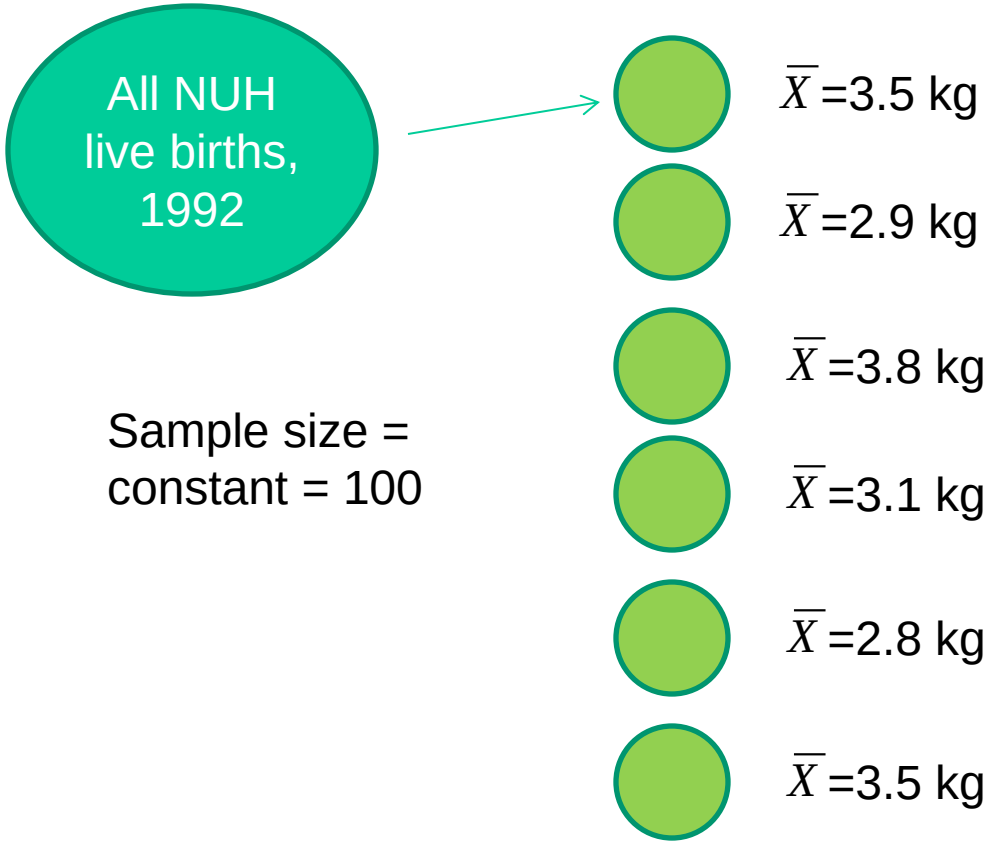
Sample mean $\overline{X}$ = 3.5 kg
Sample sd = 0.25kg

- **Suppose that we know the mean birth weight of sampled population μ to be 3.27kg with σ = 0.38kg**
- **$\overline{X} - \mu$ = 0.23kg**

Sampling error

- **Could the difference of 0.23 kg =(3.5kg –3.27kg) be real or could it be due purely to chance in sampling?**
- **'Apparent' different betw population mean and the random sample mean that is due purely to chance in sampling is called sampling error**
- **Sampling error does not mean that a mistake has been made in the process of sampling but variation experienced due to the process of sampling**
  - Sampling error reflects the difference betw the value derived from the sample and the true population value
- **The only way to eliminate sampling error is to enumerate the entire population**

Estimating sampling error

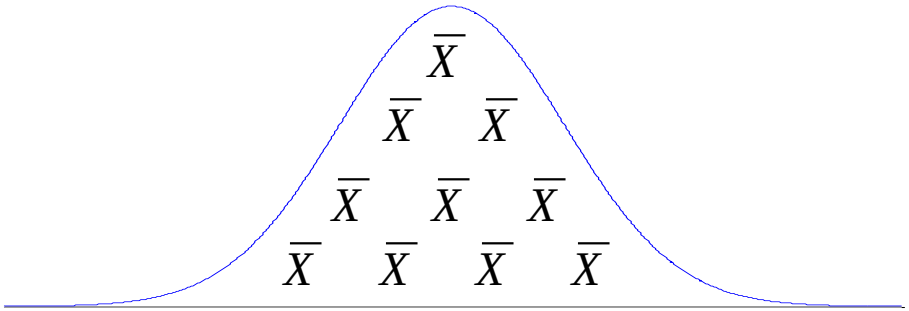


Sample size = constant = 100

Repeated sampling with replacement using the same sample size

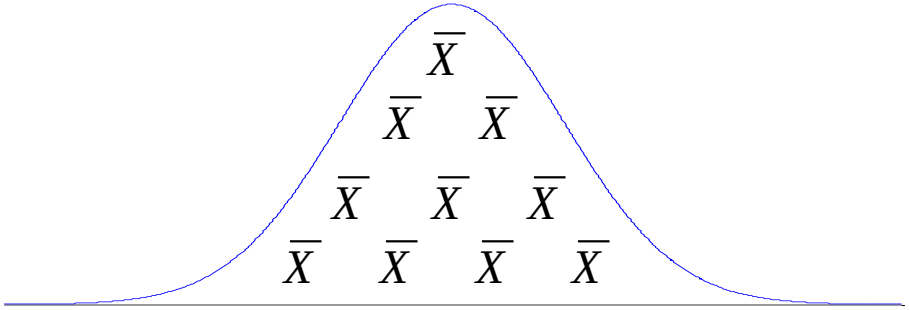
Distribution of sample means

- **Also known as sampling distribution of the mean**
- **Each unit of observation in the sampling distribution is a sample mean**
- **Spread of the sampling distribution gives a measure of the magnitude of sampling error**



Sampling distribution of the mean

- **Central limit theorem: When sample sizes are large, sampling distribution generated by repeated random sampling with replacement is invariably a normal distribution regardless of the shape of the population distribution**



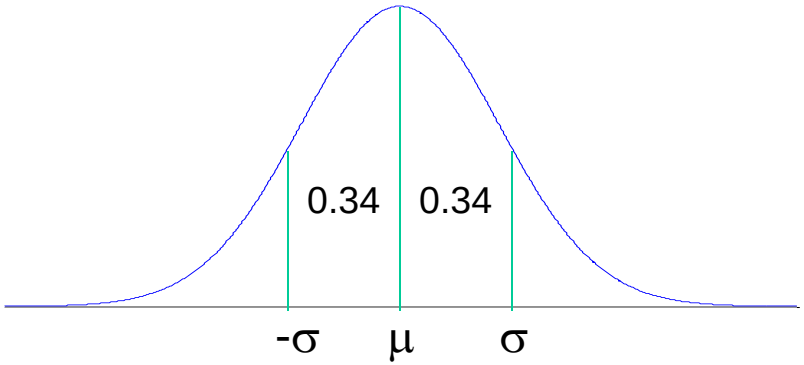
- **Mean of sampling distribution = population mean = μ**
- **Standard error of the sample mean =**

$$S.E._{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Standard deviation vs. standard error

- **Standard deviation (s.d.) tells us variability among individuals**
- **Standard error (S.E.$_{\bar{X}}$) tells us variability of sample means**
- **Standard error of the mean = S.E.$_{\bar{X}}$ =** $\frac{\sigma}{\sqrt{n}}$
  - σ : standard deviation of the population

Properties of normal distribution



- **Unimodal and symmetrical**
- **Probability distribution**
  - Area under normal curve is 1
- **For a normal distribution w/ mean μ and standard deviation σ**
  - $\mu \pm 1\sigma$ is ~68% of area under the normal curve
  - $\mu \pm 1.96\sigma$ is ~95% of area under the normal curve
  - $\mu \pm 2.58\sigma$ is ~99% of area under the normal curve

Roadmap

- **Basics of biostatistics**
- **Statistical estimation**
- **Hypothesis testing**
  - Measurement data
  - Categorical data
  - Non-parametric methods
- **Ranking and rating**
- **Principal component analysis**
- **Summary**

Central dogma of biostatistics: Estimation and hypothesis testing

- **Statistical estimation**
  - Estimating population parameters based on sample statistics
  - Application of the “confidence interval”

- **Hypothesis testing**
  - Testing certain assumptions about the population by using probabilities to estimate the likelihood of the results obtained in the sample(s) given the assumptions about the population
  - Application of “Test for statistical significance”

Statistical estimation

- **Two ways to estimate population values from sample values**
  - Point estimation
    - **Using a sample statistic to estimate a population parameter based on a single value**
      - *e.g. if a random sample of Malay births gave $\overline{X}$ =3.5kg, and we use it to estimate μ, the mean birthweight of all Malay births in the sampled population, we are making a point estimation*
    - **Point estimation ignores sampling error**
  - Interval estimation
    - **Using a sample statistic to estimate a population parameter by making allowance for sample variation (error)**

Interval estimation

- **Provide an estimation of the population parameter by defining an interval or range within which the population parameter could be found with a given probability or likelihood**
- **This interval is called confidence interval**
- **In order to understand confidence interval, we need to return to the discussion of sampling distribution**

Central limit theorem

- **Repeated sampling with replacement gives a distribution of sample means which is normally distributed and with a mean which is the true population mean, μ**

- **Assumptions:**
  - Large and constant sample size
  - Repeated sampling with replacement
  - Samples are randomly taken

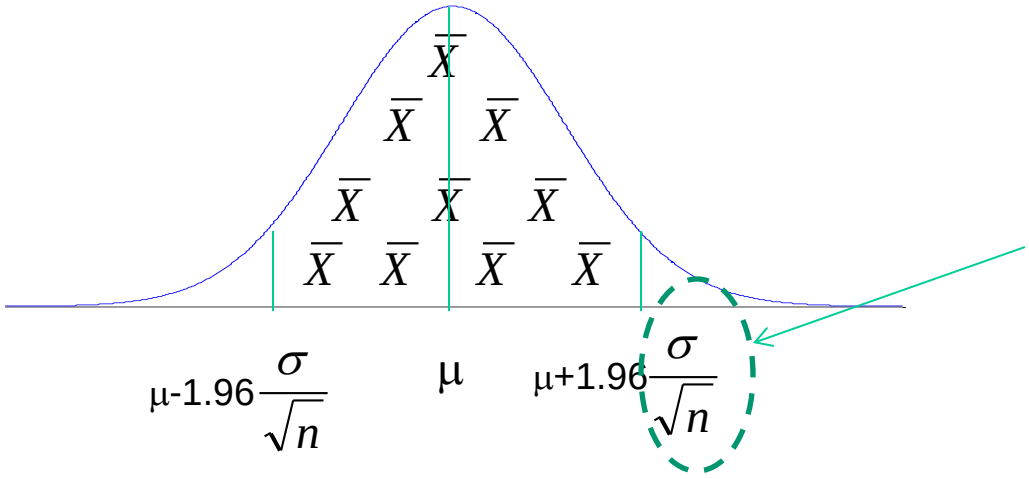
Sampling distribution of the mean

- **95% sample means of the sampling distribution can be found within the limit of $\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$**

- **Can be rewritten as**

$$\Pr\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} <= \mu <= \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) = 0.95$$

95% confidence interval

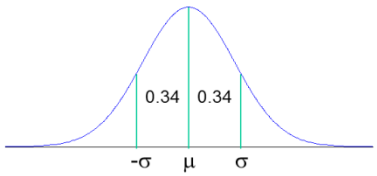


Standard error of the sample mean S.E. $\bar{X}$

Time for Exercise #1

- **Slide #28 says μ±1.96σ is 95% of the area under the normal curve. Can you explain why Slide #34 says 95% of sample means, which are normally distributed, can be found within μ±1.96σ/sqrt(n)?**

Properties of normal distribution



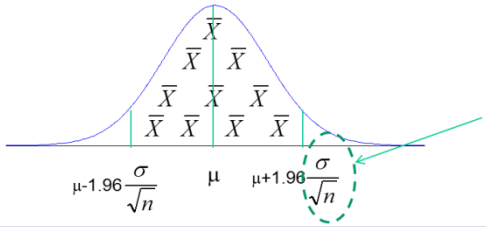
- **Unimodal and symmetrical**
- **Probability distribution**
  - Area under normal curve is 1
- **For a normal distribution w/ mean μ and standard deviation σ**
  - μ±1σ is ~68% of area under the normal curve
  - μ±1.96σ is ~95% of area under the normal curve
  - μ±2.58σ is ~99% of area under the normal curve

Sampling distribution of the mean

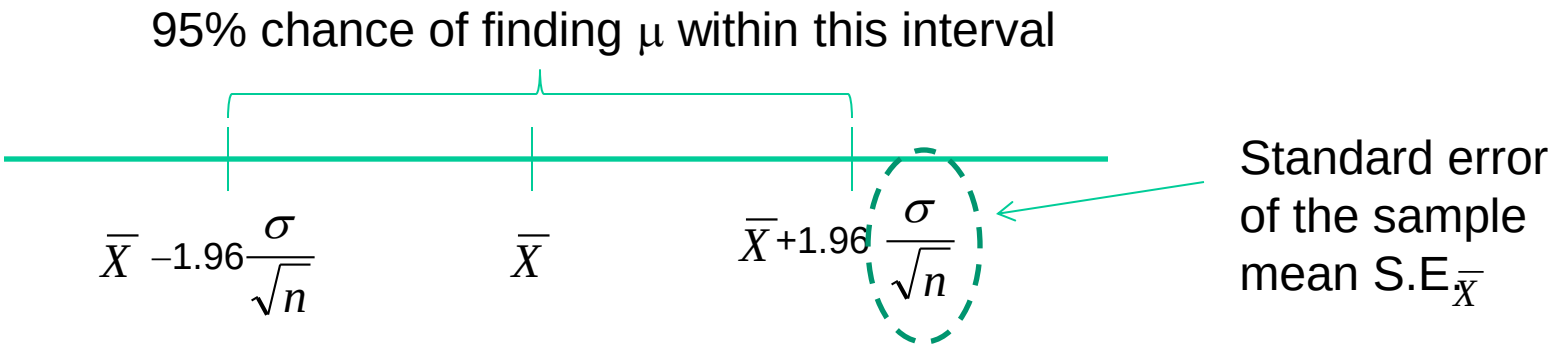
- **95% sample means of the sampling distribution can be found within the limit of** $\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$
- **Can be rewritten as**

$$\Pr\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} <= \mu <= \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

95% confidence interval



95% confidence interval



- **The 95% confidence interval gives an interval of values within which there is a 95% chance of locating the true population mean μ**

Estimating standard error

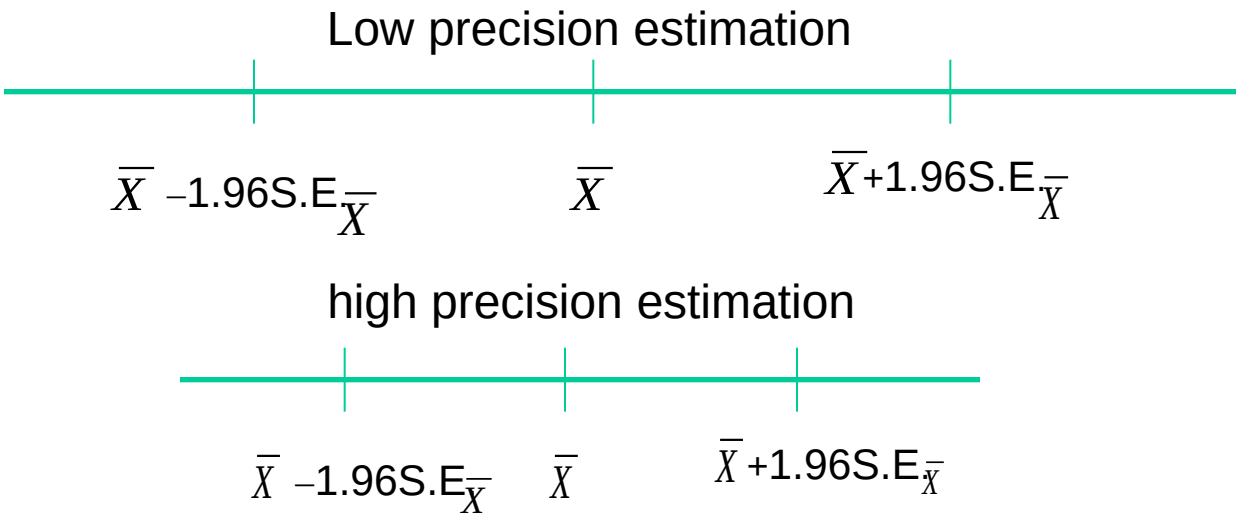
- **The sampling distribution is only a theoretical distribution as in practice we take only one sample and not repeated sample**

- **Hence S.E.$_{\bar{x}}$ is often not known but can be estimated from a single sample**

$$S.E._{\bar{X}} = \frac{s}{\sqrt{n}}$$

where s is sample standard deviation and n is the sample size

Precision of statistical estimation



- **Width of a confidence interval gives a measure of the precision of the statistical estimation**

⇒ **Estimation of population value can achieve higher precision by minimizing S.E.$_{\bar{X}}$, which depends on the population s.d. and sample size, that is S.E.$_{\bar{X}}$ can be minimized by maximizing sample size (up to a certain point)**

Roadmap

- **Basics of biostatistics**
- **Statistical estimation**
- **Hypothesis testing**
  - Measurement data
  - Categorical data
  - Non-parametric methods
- **Ranking and rating**
- **Principal component analysis**
- **Summary**

Hypothesis testing

- **Another way of statistical inference in which we want to ask a question like**
  - How likely is the mean systolic blood pressure from a sampled population (e.g. biomedical researchers) the same as those in the general population?
    - **i.e.** $\mu_{researchers} = \mu_{general}$ **?**
  - Is the difference between $\bar{X}_1$ and $\bar{X}_2$ statistically significant for us to reject the hypothesis that their corresponding u_1 and u_2 are the same?
    - **i.e.** $\mu_{male} = \mu_{female}$ **?**

Steps for test of significance

- **A test of significance can only be done on a difference, e.g. difference betw $\bar{X}$ and μ, $\bar{X}_1$ and $\bar{X}_2$**

1. **Decide the difference to be tested, e.g. difference in pulse rate betw those who were subjected to a stress test and the controls, on the assumption that the stress test significantly increases pulse rate (the hypothesis)**
2. **Formulate a Null Hypothesis, e.g. no difference in pulse rate betw the two groups,**
   - i.e. H0: $\mu_{test}=\mu_{control}$
   - Alternative hypothesis H1 : $\mu_{test} \neq \mu_{control}$

Steps for test of significance

3. **Carry out the appropriate test of significance. Based on the test statistic (result), estimate the likelihood that the difference is due purely to sample error**
4. **On the basis of likelihood of sample error, as measured by the P-value, decide whether to reject or not reject the Null Hypothesis**
5. **Draw the appropriate conclusion in the context of the biomedical problem, e.g. some evidence, from the dataset, that subjects who underwent the stress test have higher pulse rate than the controls on average**

Test of significance: An example

- **Suppose a random sample of 100 Malay male live births delivered at NUH gave a sample mean weight of 3.5kg with an sd of 0.9kg**

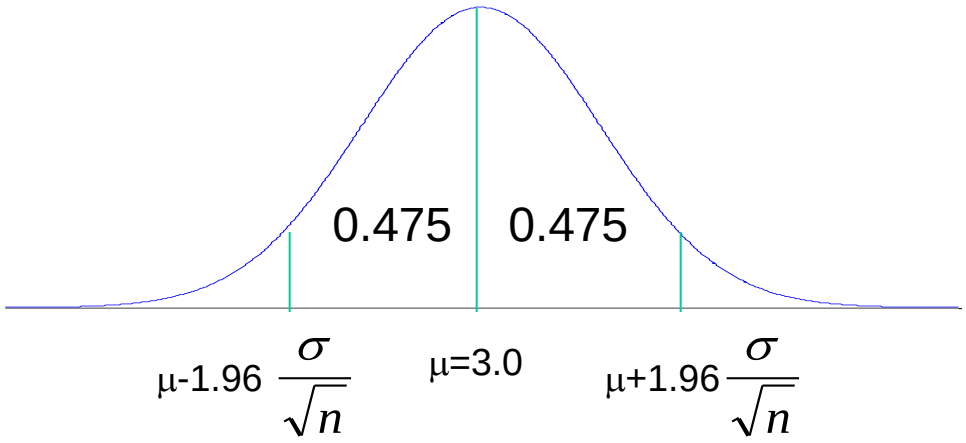
- **Question of interest: What is the likelihood that the mean birth weight from the sample population (all Malay male live birth delivered at NUH) is the same as the mean birth weight of all Malay male live births in the general population, after taking sample error into consideration?**

Test of significance: An example

- **Suppose: $\bar{X}$ = 3.5kg, sd = 0.9kg,**
  μ_{pop} = 3.0 kg σ_{pop} = 1.8kg
- **Difference betw means = 3.5–3.0 = 0.5kg**
- **Null Hypothesis, H0: $\mu_{NUH} = \mu_{pop}$**
- **Test of significance makes use of the normal distribution properties of the sampling distribution of the mean**

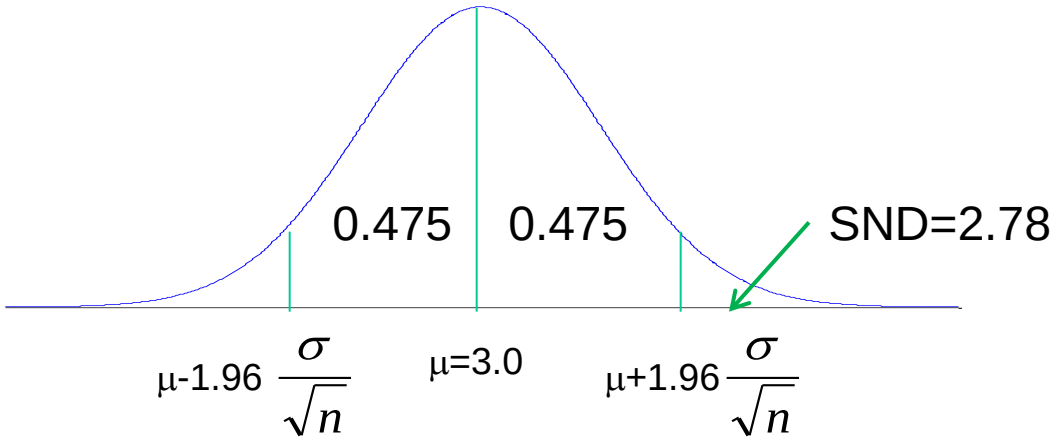
Test of significance

- **Where does our sample mean of 3.5kg lie on the sampling distribution?**
- **If it lies outside the limit of $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$, the likelihood that the sample belongs to the same population is equal or less than 0.05 (5%)**



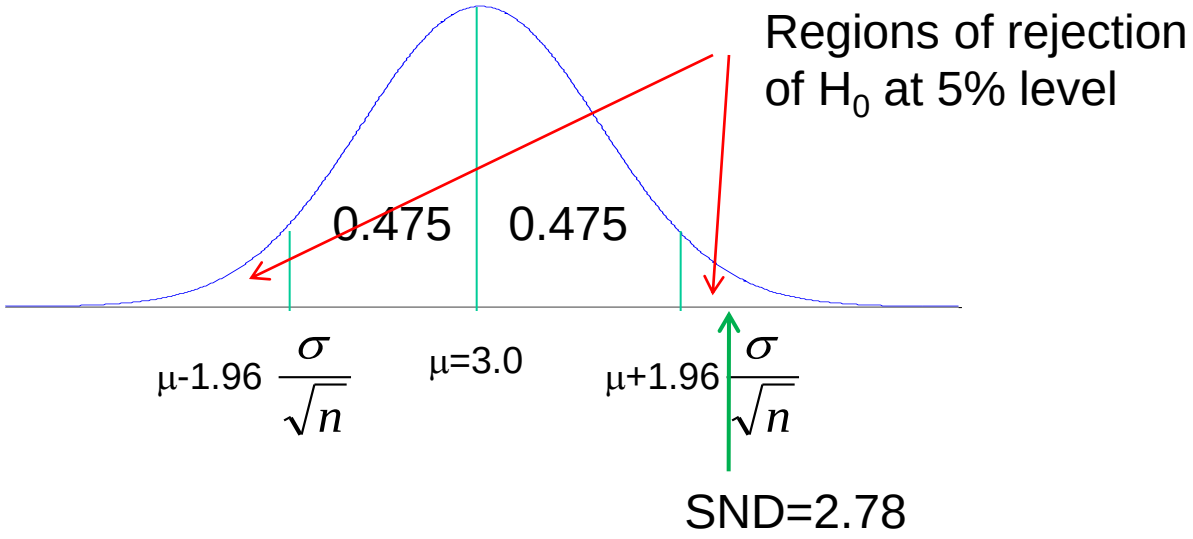
Test of significance: z-test

- **Test is given by $\frac{\bar{X} - \mu}{SE_{\bar{X}}}$ = standard normal deviate (SND or z)**
- **SND expresses the difference in standard error units on a standard normal curve with μ= 0 and σ=1**
- **For our example, SND = $\frac{3.5 - 3.0}{1.8/\sqrt{100}}$ = 2.78**



Test of significance

- **There is a less than 5% chance that a difference of this magnitude (0.5kg) could have occurred on either side of the distribution, if the random sample of the 100 Malay males had come from a population whose mean birth weight is the same as that of the general population**



NORMAL CURVE AREAS

Entries in the body of the Table give the area under the Standard Normal Curve from 0 to z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0596	.0636	.0675	.0714	.0753
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

SND(z) = 2.78

Pr (two-tailed)
= 2 x (0.5 -0.4973)
= 2 x 0.0027
= 0.0054

One-tailed vs. two-tailed test

- **So far, we have been testing for a difference that can occur on both sides of the standard normal distribution**

- **In our example, the z-value of 2.78 gave a P-value of 0.0054**

- **For a one-tailed test, a z-score of 2.78 will give a P-value of 0.0027 (=0.0054/2). It occurs when we are absolutely sure that the mean birth weight of our sample always exceeds that of the general population**

One-tailed vs. two-tailed tests

- **Two-tailed tests are conventionally used because most of the time, we are not sure of the direction of the difference**
- **One-tailed test are used only when we can anticipate a priori the direction of a difference**
- **One-tailed tests are tempting because they are more likely to give a significant result**
- **Given the same z-score, the P-value is halved for one tailed test**
- **It also mean that they run a greater risk of rejecting the Null Hypothesis when it is in fact correct --- type I error**

Type I and type II errors

- **If the difference is statistically significant, i.e. H0 is incorrect, failure to reject H0 would lead to type II error**

Conclusion from hypothesis testing	True situation: Difference exists (H_0 is incorrect)	True situation: No difference (H_0 is correct)
Difference exists (reject H_0)	Correct action (power or 1-β)	Type I or α error
No difference (Accept H_0)	Type II or β error	Correct action

Statistical significance vs. clinical significance

- **We should not be obsessed with carrying out test of significance**
  - A statistically significant result can have little or no clinical significance

- **Example: Given large sample sizes, a difference in 5 beats per minutes in pulse rate in a clinical trial involving two drugs can give a statistically significant difference when the average difference may hardly bring about a drastic metabolic change between the two groups**

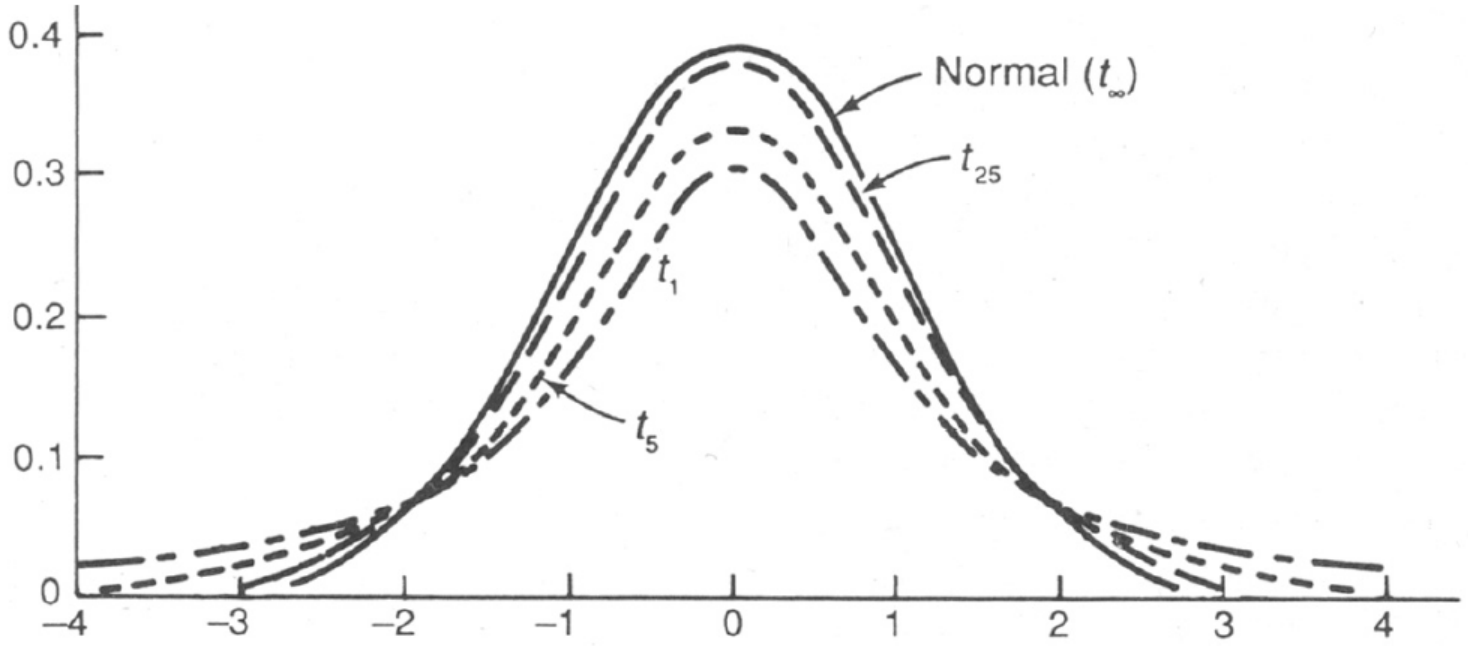
t-test

- **The assumption that the sampling distribution will be normally distributed holds for large samples but not for small samples**
- **Sample size is large, use z-test**
- **t-test is used when sample size is small**
  - Statistical concept of t-distribution
  - Comparing means for 2 independent groups
    - **unpaired t-test**
  - Comparing means for 2 matched groups
    - **paired t-test**

t-distribution

- **Sampling distribution based on small samples will be symmetrical (bell shaped) but not necessarily normal**
- **Spread of these symmetrical distributions is determined by the specific sample size. The smaller the sample size, the wider the spread, and hence the bigger the standard error**
- **These symmetrical distributions are known as Student's t-distribution or simply, t-distribution**
- **The t-distribution approaches the normal distribution when sample size tends to infinity**

Family of t-distributions



t-test for 2 independent samples or unpaired t-test

- $\bar{X}_1 - \bar{X}_2$ **= 0.08157-0.03943 = 0.04**
- **Question: What is the probability that the difference of 0.04 units between the two sample means has occurred purely by chance, i.e. due to sampling error alone?**

Blood Pb concentrations

	Battery workers (occupationally exposed)	Control (not occupationally exposed)
	0.082	0.040
	0.080	0.035
	0.079	0.036
	0.069	0.039
	0.085	0.040
	0.09	0.046
	0.086	0.040
mean	**0.08157**	**0.03943**
std dev	**0.0067047**	**0.0035523**

Unpaired t-test

- **We are testing the hypothesis that battery workers could have higher blood Pb levels than the control group of workers as they are occupationally exposed**

- **Note: conventionally, a P-value of 0.05 is generally recognized as low enough to reject the Null Hypothesis of "no difference"**

Blood Pb concentrations

Battery workers (occupationall y exposed)	Control (not occupationall y exposed)
0.082	0.040
0.080	0.035
0.079	0.036
0.069	0.039
0.085	0.040
0.09	0.046
0.086	0.040
mean **0.08157**	**0.03943**
std dev **0.0067047**	**0.0035523**

Unpaired t-test

- **Null Hypothesis: No difference in mean blood Pb level between battery workers and control group, i.e.**

  H0: $\mu_{battery} = \mu_{control}$

- **t-score is given by**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE_{(\bar{X}_1 - \bar{X}_2)}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

  with (n_1+n_2-2) degrees of freedom

Unpaired t-test

- **For the given example**

$$t = \frac{0.08157 - 0.03943}{0.002868}$$

= 14.7 with 12 d.f.

- **P-value <0.001, reject Null hypothesis**

⇒Some evidence, from the data, that battery workers in our study have higher blood Pb level than the control group on average

Blood Pb concentrations

	Battery workers (occupationally exposed)	Control (not occupationally exposed)
	0.082	0.040
	0.080	0.035
	0.079	0.036
	0.069	0.039
	0.085	0.040
	0.09	0.046
	0.086	0.040
mean	**0.08157**	**0.03943**
std dev	**0.0067047**	**0.0035523**

t-table

From our example:
t=14.7 with 12 d.f.

Value far exceeds 4.318, the critical value for statistical significance at the Pr=0.001 (0.1%) level when df=12
i.e. Pr < 0.001

	Probability			
df	.05	.02	.01	.001
1	12.706	31.821	63.657	636.619
2	4.303	6.965	9.925	31.598
3	3.182	4.541	5.841	12.924
4	2.776	3.747	4.604	8.610
5	2.571	3.365	4.032	6.869
6	2.447	3.143	3.707	5.959
7	2.365	2.998	3.499	5.408
8	2.306	2.896	3.355	5.041
9	2.262	2.821	3.250	4.781
10	2.228	2.764	3.169	4.587
11	2.201	2.718	3.106	4.437
12	2.179	2.681	3.055	4.318
13	2.160	2.650	3.012	4.221
14	2.145	2.624	2.977	4.140
15	2.131	2.602	2.947	4.073
16	2.120	2.583	2.921	4.015
17	2.110	2.567	2.898	3.965
18	2.101	2.552	2.878	3.922
19	2.093	2.539	2.861	3.883
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .
25	2.060	2.485	2.787	3.725
26	2.056	2.479	2.779	3.707
27	2.052	2.473	2.771	3.690
28	2.048	2.467	2.763	3.674
29	2.045	2.462	2.756	3.659
30	2.042	2.457	2.750	3.646
40	2.021	2.423	2.704	3.551
60	2.000	2.390	2.660	3.460
120	1.980	2.358	2.617	3.373
α	1.960	2.326	2.576	3.291

Unpaired t-test assumptions

- **Data are normally distributed in the population from which the two independent samples have been drawn**
- **The two samples are random and independent, i.e. observations in one group are not related to observations in the other group**
- **The 2 independent samples have been drawn from populations with the same (homogeneous) variance, i.e. $\sigma_1=\sigma_2$**

Paired t-test

- **Previous problem uses un-paired t-test as the two samples were not matched**
  - i.e. the two samples were independently derived
- **Sometimes, we may need to deal with matched study designs**

Patient	Fasting cholesterol	Postprandial cholesterol
1	198	202
2	192	188
3	241	238
4	229	226
5	185	174
6	303	315

Study involves 6 subjects acting as their own control (best match)

Paired t-test

- **Null hypothesis: No difference in mean cholesterol levels between fasting and postprandial states**

  H0: $\mu_{fasting} = \mu_{postprandial}$

Patient	Fasting cholesterol	Postprandial cholesterol	Difference (d)
1	198	202	-4
2	192	188	+4
3	241	238	+3
4	229	226	+3
5	185	174	+11
6	303	315	-12

$\bar{d}$ = 0.833
s_d = 7.885
n= 6

Paired t-test

- **t-score given by**

$$t = \frac{\bar{d}}{SE_{\bar{d}}} = \frac{\bar{d}}{s_d / \sqrt{n}}$$

$$= \frac{0.833}{3.219} = 0.259$$

with (n-1) degrees of freedom, where n is the # of pairs

Patient	Difference (d)
1	-4
2	+4
3	+3
4	+3
5	+11
6	-12

$\bar{d}$ = 0.833
s_d = 7.885
n= 6

t-table

From our example:
t=0.259 with 5 d.f.

Value is very much lower than 2.571, the critical value for statistical significance at the Pr=0.05 (5%) level when df=5
i.e. Pr > 0.05

	Probability			
df	.05	.02	.01	.001
1	12.706	31.821	63.657	636.619
2	4.303	6.965	9.925	31.598
3	3.182	4.541	5.841	12.924
4	2.776	3.747	4.604	8.610
5	2.571	3.365	4.032	6.869
6	2.447	3.143	3.707	5.959
7	2.365	2.998	3.499	5.408
8	2.306	2.896	3.355	5.041
9	2.262	2.821	3.250	4.781
10	2.228	2.764	3.169	4.587
11	2.201	2.718	3.106	4.437
12	2.179	2.681	3.055	4.318
13	2.160	2.650	3.012	4.221
14	2.145	2.624	2.977	4.140
15	2.131	2.602	2.947	4.073
16	2.120	2.583	2.921	4.015
17	2.110	2.567	2.898	3.965
18	2.101	2.552	2.878	3.922
19	2.093	2.539	2.861	3.883
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .
25	2.060	2.485	2.787	3.725
26	2.056	2.479	2.779	3.707
27	2.052	2.473	2.771	3.690
28	2.048	2.467	2.763	3.674
29	2.045	2.462	2.756	3.659
30	2.042	2.457	2.750	3.646
40	2.021	2.423	2.704	3.551
60	2.000	2.390	2.660	3.460
120	1.980	2.358	2.617	3.373
α	1.960	2.326	2.576	3.291

Paired t-test

Patient	Fasting cholesterol	Postprandial cholesterol
1	198	202
2	192	188
3	241	238
4	229	226
5	185	174
6	303	315

- **Action: Should not reject the Null Hypothesis**

- **Conclusion: Insufficient evidence, from the data, to suggest that postprandial cholesterol levels are, on average, higher than fasting cholesterol levels**

Common errors relating to t-test

- **Failure to recognize assumptions**
  - If assumption does not hold, explore data transformation or use of non-parametric methods
- **Failure to distinguish between paired and unpaired designs**

Roadmap

- **Basics of biostatistics**
- **Statistical estimation**
- **Hypothesis testing**
  - Measurement data
  - Categorical data
  - Non-parametric methods
- **Ranking and rating**
- **Principal component analysis**
- **Summary**

Hypothesis testing involving categorical data

- **Chi-square test for statistical association involving 2x2 tables and RxC tables**
  - Testing for associations involving small, unmatched samples
  - Testing for associations involving small, matched samples

Association

- **Examining relationship betw 2 categorical variables**
- **Some examples of association:**
  - Smoking and lung cancer
  - Ethic group and coronary heart disease
- **Questions of interest when testing for association betw two categorical variables**
  - Does the presence/absence of one factor (variable) influence the presence/absence of the other factor (variable)?
- **Caution**
  - presence of an association does not necessarily imply causation

Relating to comparison betw proportions

Treatment	Improvement	No improvement	Total
Arthritic drug	18	6	24
placebo	9	11	20
Total	27	17	44

- **Proportion improved in drug group = 18/24 = 75%**
- **Proportion improved in placebo group = 9/20 = 45.0%**

- **Question: What is the probability that the observed difference of 30% is purely due to sampling error, i.e. chance in sampling?**
- **Use $\chi 2$ –test**

Chi-square test for statistical association

treatment	Improvement	No improvement	Total
Arthritic drug	18 (a)	6 (b)	24
placebo	9 (c)	11 (d)	20
Total	27	17	44

- **Prob of selecting a person in drug group = 24/44**
- **Prob of selecting a person with improvement = 27/44**
- **Prob of selecting a person from drug group who had shown improvement= (24/44)*(27/44) = 0.3347 (assuming two independent events)**
- **Expected value for cell (a) =0.3347*44 = 14.73**

Chi-square test for statistical association

treatment	Improvement	No improvement	Total
Arthritic drug	18 (14.73)	6 (9.27)	24
placebo	9 (12.27)	11 (7.73)	20
Total	27	17	44

- **General formula for $\chi 2$**

$$\chi^2 = \sum \frac{(obs - \exp)^2}{\exp}$$

- **$\chi 2$ –test is always performed on categorical variables using absolute frequencies, never percentage or proportion**

Chi-square test for statistical association

- **For the given problem:**

$$\sum \frac{(obs - \exp)^2}{\exp} = \frac{(18-14.73)^2}{14.73} + \frac{(6-9.27)^2}{9.27} + \frac{(9-12.27)^2}{12.27} + \frac{(11-7.73)^2}{7.73}$$

$$= 4.14 \text{ with } 1 \text{ degree of freedom}$$

- **χ2 degree of freedom is given by:**

  (no. of rows – 1)*(no. of cols – 1)

  = (2 – 1)*(2 – 1) = 1

18	6	24
9	11	20
27	17	44

How many of these 4 cells are free to vary if we keep the row and column totals constant?

χ^2 table

Critical values in the distributions of chi-squared for different degrees of freedom

df	Probability .05	.02	.01	.001
1	3.841	5.412	6.635	10.827
2	5.991	7.824	9.210	13.815
3	7.815	9.837	11.345	16.266
4	9.488	11.668	13.277	18.467
5	11.070	13.388	15.086	20.515
6	12.592	15.033	16.812	22.457
7	14.067	16.622	18.475	24.322
8	15.507	18.168	20.090	26.125
9	16.919	19.679	21.666	27.877
10	18.307	21.161	23.209	29.588
11	19.675	22.618	24.725	31.264
12	21.026	24.054	26.217	32.909
13	22.362	25.372	27.688	34.528
14	23.585	26.873	29.141	36.123
15	24.996	28.259	30.578	37.697
16	26.296	29.633	32.000	39.252
17	27.587	30.995	33.409	40.790
18	28.869	32.346	34.805	42.312
19	30.144	33.687	36.191	43.820
20	31.410	35.020	37.566	35.315
21	32.671	36.343	38.932	46.797
22	33.924	37.659	40.289	48.268
23	35.172	38.968	41.638	49.728
24	36.415	40.270	42.980	51.179
25	37.652	41.566	44.314	52.620
26	38.885	42.856	45.642	54.052
27	40.113	44.140	46.963	55.476
28	41.337	45.419	48.278	56.893
29	42.557	46.693	49.588	58.302
30	43.773	47.962	50.892	59.703

observed χ^2 value of 4.14 exceeds critical value of 3.841 for P=0.05 but not critical value of 5.412 for P=0.02 at 1 d.f.

i.e. $0.05 > P > 0.02$

Chi-square test for statistical association

- **Probability of getting an observed difference of 30% in improvement rates if the Null Hypothesis of no association is correct is betw 2% and 5%**

- **Hence, there is some statistical evidence from this study to suggest that treatment of arthritic patient with the drug can significantly improve grip strength**

Yate's correction for continuity

- In the $\chi 2$ test, we are using a discrete statistic which is approx by a continuous $\chi 2$ distribution. To correct for the use of the discrete statistic, a correction is applied to the original $\chi 2$ value to improve the fit

$$\chi_c^2 = \sum \frac{(|obs - \exp| - 0.5)^2}{\exp}$$

- Yate's correction for continuity is particularly useful when dealing with small sample size studies
- Yate's correction does not apply to contingency tables larger than 2x2. For non-2x2 tables, low cell frequencies are resolved by pooling (collapsing) adjacent cells

Extending to RxC tables

Type of vaccines	Had flu	Avoided flu	total
I	43	237	280
II	52	198	250
III	25	245	270
IV	48	212	260
V	57	233	290
Total	225	1125	1350

- **Null Hypothesis assumes all vaccines tested had equal efficacy**

Computation of the $\chi 2$

Type of vaccines	Had flu	$(O-E)^2/E$	Avoided flu	$(O-E)^2/E$
I	43 (46.7)	0.293	237 (233.3)	0.059
II	52 (41.7)	2.544	198 (208.3)	0.509
III	25 (45.0)	8.889	245 (225.0)	1.778
IV	48 (43.3)	0.510	212 (216.7)	0.102
V	57 (48.3)	1.567	233 (241.7)	0.313
Total	225	13.803	1125	2.761

- **$\chi 2$ =13.803+2.761 = 16.564 with 4 d.f.**

χ^2 table

Critical values in the distributions of chi-squared for different degrees of freedom

df	Probability .05	.02	.01	.001
1	3.841	5.412	6.635	10.827
2	5.991	7.824	9.210	13.815
3	7.815	9.837	11.345	16.266
4	9.488	11.668	13.277	18.467
5	11.070	13.388	15.086	20.515
6	12.592	15.033	16.812	22.457
7	14.067	16.622	18.475	24.322
8	15.507	18.168	20.090	26.125
9	16.919	19.679	21.666	27.877
10	18.307	21.161	23.209	29.588
11	19.675	22.618	24.725	31.264
12	21.026	24.054	26.217	32.909
13	22.362	25.372	27.688	34.528
14	23.585	26.873	29.141	36.123
15	24.996	28.259	30.578	37.697
16	26.296	29.633	32.000	39.252
17	27.587	30.995	33.409	40.790
18	28.869	32.346	34.805	42.312
19	30.144	33.687	36.191	43.820
20	31.410	35.020	37.566	35.315
21	32.671	36.343	38.932	46.797
22	33.924	37.659	40.289	48.268
23	35.172	38.968	41.638	49.728
24	36.415	40.270	42.980	51.179
25	37.652	41.566	44.314	52.620
26	38.885	42.856	45.642	54.052
27	40.113	44.140	46.963	55.476
28	41.337	45.419	48.278	56.893
29	42.557	46.693	49.588	58.302
30	43.773	47.962	50.892	59.703

observed χ^2 value of 16.564 with 4 d.f. exceeds critical value of 13.277 for P=0.01 but not critical value of 18.467 for P=0.001.

i.e. $0.01 > P > 0.001$

Computation of the $\chi 2$

Type of vaccines	Had flu	$(O-E)^2/E$	Avoided flu	$(O-E)^2/E$
I	43 (46.7)	0.293	237 (233.3)	0.059
II	52 (41.7)	2.544	198 (208.3)	0.509
III	**25 (45.0)**	**8.889**	**245 (225.0)**	**1.778**
IV	48 (43.3)	0.510	212 (216.7)	0.102
V	57 (48.3)	1.567	233 (241.7)	0.313
Total	225	13.803	1125	2.761

- **Vaccine III contributes to the overall $\chi 2$= (8.889+1.778)/16.564 = 64.4%**

$\chi 2$ with Vaccine III removed

Type of vaccines	Had flu	Avoided flu	total
I	43	237	280
II	52	198	250
IV	48	212	260
V	57	233	290

- **$\chi 2$ =2.983 with 3 d.f.**
- **0.1<p<0.5, not statistically significant**

Vaccine III vs. rest

Type of vaccines	Had flu	Avoided flu	total
III	25	245	270
I, II, IV, V	200	880	1080
Total	225	1125	1350

- $\chi 2 = 12.7$ with 1 d.f.
- $P < 0.001$
- There appear to be strong statistical evidence that the protective effect of vaccine III is significantly better than the other vaccines

Handling extremely small samples

- **For extremely small samples, $\chi 2$ -test even with Yate's correction is NOT recommended**
- **Fisher's exact test should be used when there are small expected frequencies**
  - Involves calculating the exact probability of a table as extreme or more extreme than the one observed, given that the null hypothesis is correct
- **When to use fisher's exact test (rule of thumb)**
  - When the overall sample size < 20
  - Overall sample size is between 20 and 40 and the smallest of the four expected value < 5

Calculating Fisher's exact probability

- **Exact probability of observing a particular set of frequencies in a 2x2 table when the row and column totals are fixed is given by the hypergeometric distribution**

$$P(x=k)=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

	Yes	No	Total
Yes	k	m-k	**m**
No	n-k	N+k-m-n	**N-m**
total	**n**	**N-n**	**N**

Comparing proportion for matched data

- **100 women in a fertility drug trial were matched in pairs for age, race group and duration of marriage. By random allocation, one woman in each pair was given a fertility drug while the other was given a placebo.**
- **Success is recorded if, within 12 months, a study subject became pregnant and failure otherwise**
- **Point to note about the study**
  - Matched design
  - Compare proportion of successes between fertility drug and placebo

McNemar's test

		placebo: success	placebo: failure
drug	success	20 (a)	12(b)
	failure	2 (c)	16 (d)

- **McNemar's test (based on discordant pairs)**

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{81}{14} = 5.79$$

- **0.01<p<0.02**
- **Strong statistical evidence that the fertility drug produces a higher success rate than the placebo**

Roadmap

- **Basics of biostatistics**
- **Statistical estimation**
- **Hypothesis testing**
  - Measurement data
  - Categorical data
  - Non-parametric methods
- **Ranking and rating**
- **Principal component analysis**
- **Summary**

Why non-parametric methods

- **Certain statistical tests like the t-test require assumptions of the distribution of the study variables in the population**
  - t-test requires the underlying assumption of a normal distribution
  - Such tests are known as parametric tests

- **There are situations when it is obvious that the study variable cannot be normally distributed, e.g.,**
  - # of hospital admissions per person per year
  - # of surgical operations per person

Why non-parametric methods

- **The study variable generates data which are scores and so should be treated as a categorical variable with data measured on ordinal scale**
  - E.g., scoring system for degree of skin reaction to a chemical agent:
    - **1: intense skin reaction**
    - **2: less intense reaction**
    - **3: No reaction**
- **For such type of data, the assumption required for parametric tests seem invalid => non-parametric methods should be used**
- **Aka distribution-free tests, because they make no assumption about the underlying distribution of the study variables**

Wilcoxon rank sum test (aka Mann-Whitney U test)

- **Non-parametric equivalent of parametric t-test for 2 independent samples (unpaired t-test)**

- **Suppose the waiting time (in days) for cataract surgery at two eye clinics are as follows:**

Patients at clinic A (n_A=18)	1, 5, 15, 7, 42, 13, 8, 35, 21, 12, 12, 22, 3, 14, 4, 2, 7, 2
Patients at clinic B (n_B=15)	4, 9, 6, 2, 10, 11, 16, 18, 6, 0, 9, 11, 7, 11, 10

Wilcoxon rank sum test

1. **Rank all observations (n_A+n_B) in ascending order (least time to longest) along with the group identity each observation belongs**
2. **Resolve tied ranks by dividing sum of the ranks by the number of entries for a particular set of ties, i.e. average the ranks**

time	rank	clinic	time	rank	clinic
0	1	B	8	15	A
1	2	A	9	16.5	B
2	4	A	9	16.5	B
2	4	B	10	18.5	B
2	4	A	10	18.5	B
3	6	A	11	21	B
4	7.5	A	11	21	B
4	7.5	B	11	21	B
5	9	A	12	23.5	A
6	10.5	B	12	23.5	A
6	10.5	B	13	25	A
7	13	A	etc	etc	etc
7	13	A			
7	13	B			

Wilcoxon rank sum test

3. **Sum up ranks separately for the two groups. If the two populations from which the samples have been drawn have similar distributions, we would expect the sum of ranks to be close. If not, we would expect the group with the smaller median to have the smaller sum of ranks**
4. **If the group sizes in both groups are the same, take the group with the smaller sum of ranks**

   If both groups have unique sample sizes, then use the sum of ranks of the smaller group
5. **Test for statistical significance**

Wilcoxon rank sum test

- **In this example**
  - sum of group A ranks = 324.5
  - sum of group B ranks = 236.5
- **T= 236.5 (sum of ranks of the smaller group)**
- **If $n=n_A+n_B$ <=25, then looking up table giving critical values of T for various size of n_A and n_B**
- **If n>25, we assume that T is practically normally distributed with**

$$\mu_T = \frac{n_A(n_A + n_B + 1)}{2}, \text{ where } n_A < n_B$$

$$SE_T = \sqrt{\frac{n_B \mu}{6}}$$

Wilcoxon rank sum test

- **For our problem, T=236.5, n_A=18, n_B=15**

$$z = \frac{T - \mu_T}{SE_T} = \frac{236.5 - 255}{27.66} = 0.67$$

- **Result is not statistically significant at 5% (P=0.05) level**

$\Rightarrow$ **No strong evidence to show that the difference in waiting time for the two clinics are statistically significant**

Wilcoxon matched pairs signed ranks test

- **Non-parametric equiv of parametric paired t-test**

- **Suppose the anxiety scores recorded for 10 patients receiving a new drug and a placebo in random order in a cross-over clinical trial are:**

Patients	1	2	3	4	5	6	7	8	9	10
Drug score	19	11	14	17	23	11	15	19	11	8
Placebo score	22	18	17	19	22	12	14	11	19	7

- **Question: Is there any statistical evidence to show that the new drug can significantly lower anxiety scores when compared with the placebo?**

Wilcoxon matched pairs signed ranks test

1. **Take the difference for each pair of readings**

Patients	1	2	3	4	5	6	7	8	9	10
Drug score	19	11	14	17	23	11	15	19	11	8
Placebo score	22	18	17	19	22	12	14	11	19	7
difference	-3	-7	-3	-2	1	-1	1	8	-8	1

Wilcoxon matched pairs signed ranks test

2. Rank the differences from the smallest to the largest, ignoring signs and omitting 0 differences

Patients	1	2	3	4	5	6	7	8	9	10
Drug score	19	11	14	17	23	11	15	19	11	8
Placebo score	22	18	17	19	22	12	14	11	19	7
difference	-3	-7	-3	-2	1	-1	1	8	-8	1
rank	6.5	8	6.5	5	2.5	2.5	2.5	9.5	9.5	2.5

Wilcoxon matched pairs signed ranks test

3. Put back the signs to the ranks

Patients	1	2	3	4	5	6	7	8	9	10
Drug score	19	11	14	17	23	11	15	19	11	8
Placebo score	22	18	17	19	22	12	14	11	19	7
difference	-3	-7	-3	-2	1	-1	1	8	-8	1
Rank -	6.5	8	6.5	5		2.5			9.5	
Rank +					2.5		2.5	9.5		2.5

Wilcoxon matched pairs signed ranks test

4. **Add up ranks of positive differences and ranks of negative differences. Call the sum of the smaller group T**

Patients	1	2	3	4	5	6	7	8	9	10
Drug score	19	11	14	17	23	11	15	19	11	8
Placebo score	22	18	17	19	22	12	14	11	19	7
difference	-3	-7	-3	-2	1	-1	1	8	-8	1
Rank -	6.5	8	6.5	5		2.5			9.5	
Rank +					2.5		2.5	9.5		2.5

- **Sum of + ranks: 17 (n+ =4)**
- **Sum of – ranks: 38 (n– = 6)**
- **T (sum of ranks of smaller group) = 17**

Wilcoxon matched pairs signed ranks test

5. **Test for statistical significance**
   - If n≤25, then look up table giving critical values of T for various size of n
   - If n>25, we can assume that T is practically normally distributed with

$$\mu_T = \frac{n(n+1)}{4}$$

$$SE_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{\mu_T(2n+1)}{6}}$$

- **For our problem, T=17 and n=10, hence we look up table**

critical value for P=0.05 at N=10 is 8 (for 2-tailed test)

Note that critical values go progressively smaller as P gets smaller

Table B **Table of Critical Values of T in the Wilcoxon's Matched-Pairs Signed-Ranks Test**

	Level of significance for one-tailed test		
	0.025	0.01	0.005
N	Level of significance for two-tailed test		
	0.05	0.02	0.01
6	0	-	-
7	2	0	-
8	4	2	0
9	6	3	2
→ 10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Wilcoxon matched pairs signed ranks test

- **For our problem, we found that T value of 17 is higher than the critical value for statistical significance at the 5% level**

$\Rightarrow$ **There is insufficient evidence to show that the new drug can significantly lower anxiety scores than the placebo. Therefore, we cannot rule out the possibility that the observed differences among scores are due to sampling error.**

Non-parametric vs. parametric methods

- **Advantages:**
  - Do not require the assumption needed for parametric tests. Therefore useful for data which are markedly skewed
  - Good for data generated from small samples. For such small samples, parametric tests are not recommended unless the nature of population distribution is known
  - Good for observations which are scores, i.e. measured on ordinal scale
  - Quick and easy to apply and yet compare quite well with parametric methods

Non-parametric vs. parametric methods

- **Disadvantages**
  - Not suitable for estimation purposes as confidence intervals are difficult to construct
  - No equivalent methods for more complicated parametric methods like testing for interactions in ANOVA models
  - Not quite as statistically efficient as parametric methods if the assumptions needed for the parametric methods have been met

Roadmap

- **Basics of biostatistics**
- **Statistical estimation**
- **Hypothesis testing**
  - Measurement data
  - Categorical data
  - Non-parametric methods
- **Ranking and rating**
- **Principal component analysis**
- **Summary**

Ranking and rating

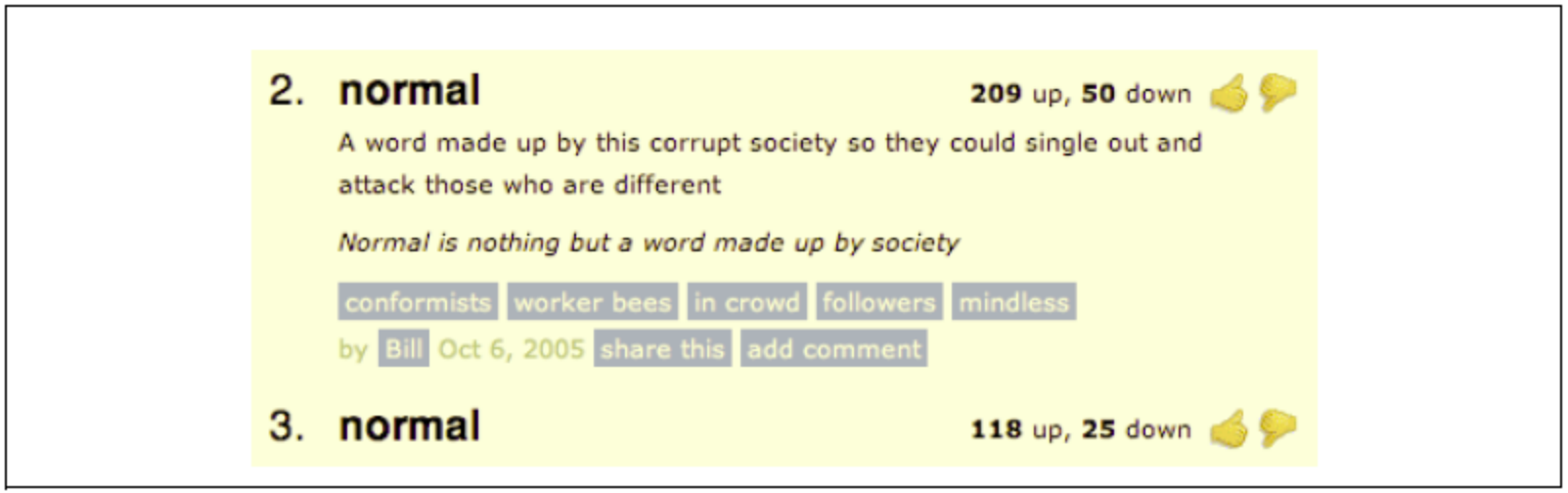
- **PROBLEM: You are a web programmer. You have users. Your users rate stuff on your site. You want to put the highest-rated stuff at the top and lowest-rated at the bottom. You need some sort of "score" to sort by**

PROBLEM: You are a web programmer. You have users. Your users rate stuff on your site. You want to put the highest-rated stuff at the top and lowest-rated at the bottom. You need some sort of "score" to sort by.

WRONG SOLUTION #1: Score = (Positive ratings) - (Negative ratings)

Why it is wrong: Suppose one item has 600 positive ratings and 400 negative ratings: 60% positive. Suppose item two has 5,500 positive ratings and 4,500 negative ratings: 55% positive. This algorithm puts item two (score = 1000, but only 55% positive) above item one (score = 200, and 60% positive). WRONG.

Sites that make this mistake: Urban Dictionary



PROBLEM: You are a web programmer. You have users. Your users rate stuff on your site. You want to put the highest-rated stuff at the top and lowest-rated at the bottom. You need some sort of "score" to sort by.

WRONG SOLUTION #2: Score = Average rating = (Positive ratings) / (Total ratings)

Why it is wrong: Average rating works fine if you always have a ton of ratings, but suppose item 1 has 2 positive ratings and 0 negative ratings. Suppose item 2 has 100 positive ratings and 1 negative rating. This algorithm puts item two (tons of positive ratings) below item one (very few positive ratings). WRONG.

Sites that make this mistake: Amazon.com

PROBLEM: You are a web programmer. You have users. Your users rate stuff on your site. You want to put the highest-rated stuff at the top and lowest-rated at the bottom. You need some sort of "score" to sort by.

- **A possible solution is the lower bound of the normal approximation interval**

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

where $\hat{p}$ is the proportion of successes in a Bernoulli trial process estimated from the statistical sample, $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of a standard normal distribution, α is the error percentile and n is the sample size. For example, for a 95% confidence level the error (α) is 5%, so $1-\alpha/2 = 0.975$ and $z_{1-\alpha/2} = 1.96$.

- **An improvement is the lower bound of the Wilson interval**

$$\frac{\hat{p} + \frac{1}{2n}z^2_{1-\alpha/2} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1 + \frac{1}{n}z^2_{1-\alpha/2}}$$

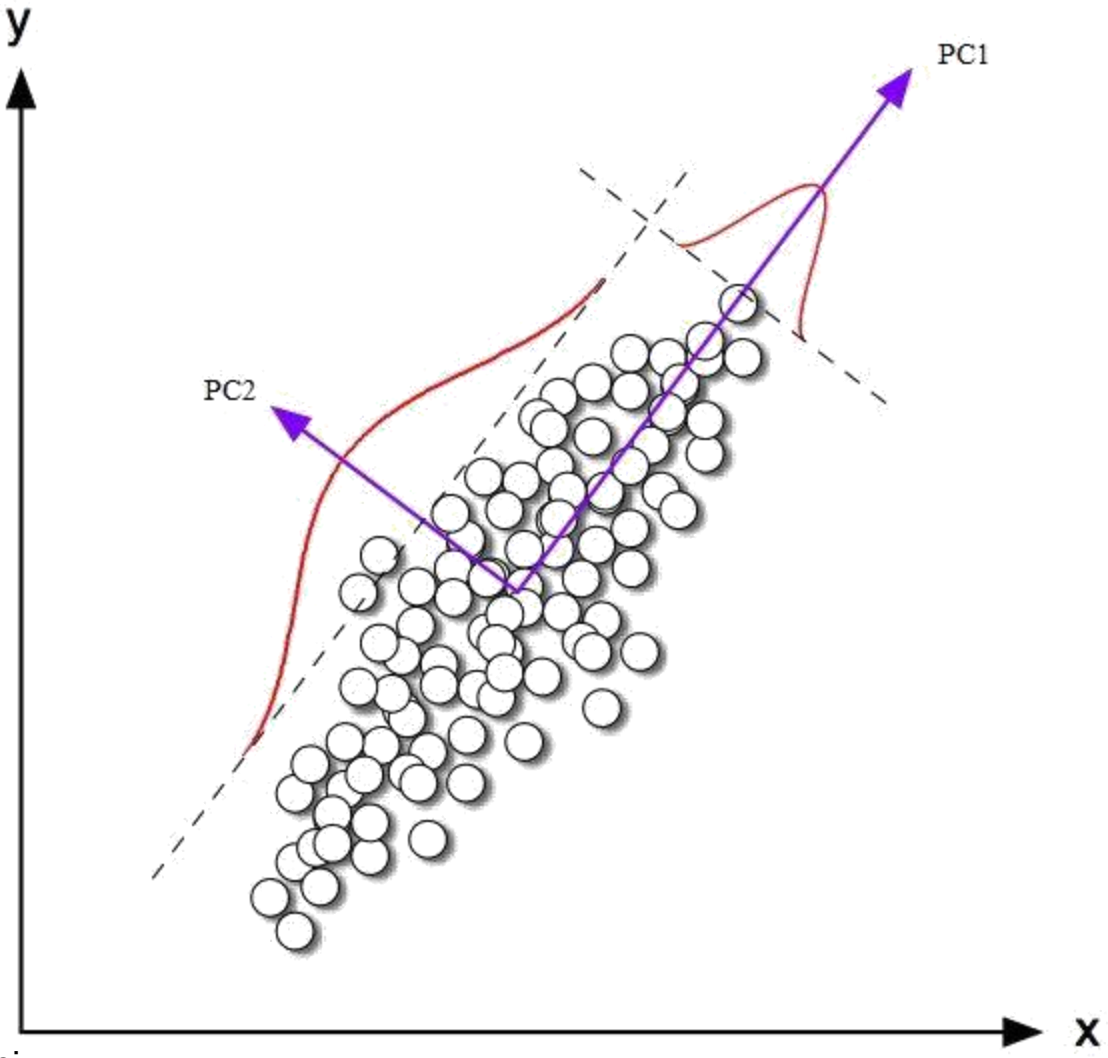
Time for Exercise #2

- **You are a web programmer. You have users. Your users rate stuff on your site. You want to put the highest-rated stuff at the top and lowest-rated at the bottom. What "score" would you suggest to sort your stuff by?**

Roadmap

- **Basics of biostatistics**
- **Statistical estimation**
- **Hypothesis testing**
  - Measurement data
  - Categorical data
  - Non-parametric methods
- **Ranking and rating**
- **Principal component analysis**
- **Summary**

Principal Component Analysis



Credit: Alessandro Giuliani

PCA, a la Pearson (1901)

)(98)(

SULLE FUNZIONI BILINEARI

DI

E. BELTRAMI

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y,$$
$$\text{or} \quad z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

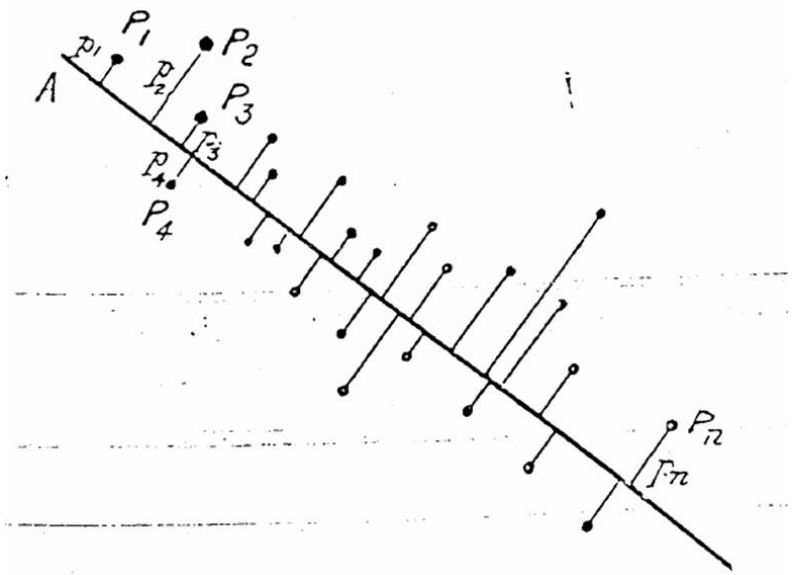
where $y, x, z, x_1, x_2, \ldots x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \ldots a_n$

For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

$$U = S(p^2) = \text{a minimum.}$$

If y were the dependent variable, we should have made

$$S(y' - y)^2 = \text{a minimum}$$



Credit: Alessandro Giuliani

Credit: Marloes Maathuis

NUS National University of Singapore

PCA, in modern English ☺

Introduction

- Technique quite old: Pearson (1901) and Hotelling (1933), but still one of the most used multivariate techniques today
- Main idea:
  - Start with variables $X_1, \ldots, X_p$
  - Find a *rotation* of these variables, say $Y_1, \ldots, Y_p$ (called principal components), so that:
    - $Y_1, \ldots, Y_p$ are uncorrelated. Idea: they measure different dimensions of the data.
    - $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \text{Var}(Y_p)$. Idea: Y_1 is most important, then Y_2, etc.

9 / 33

Definition of PCA

- Given $X = (X_1, \ldots, X_p)'$
- We call $a'X$ a standard linear combination (SLC) if $\sum a_i^2 = 1$
- Find the SLC $a'_{(1)} = (a_{11}, \ldots, a_{p1})$ so that $Y_1 = a'_{(1)}X$ has maximal variance
- Find the SLC $a'_{(2)} = (a_{12}, \ldots, a_{p2})$ so that $Y_2 = a'_{(2)}X$ has maximal variance, subject to the constraint that Y_2 is uncorrelated to Y_1.
- Find the SLC $a'_{(3)} = (a_{13}, \ldots, a_{p3})$ so that $Y_3 = a'_{(3)}X$ has maximal variance, subject to the constraint that Y_3 is uncorrelated to Y_1 and Y_2
- Etc...

10 / 33

1st Principal Component

- How to combine the scores on 5 different exams to a total score? One could simply take the average. But it may be better to use the first principal component

- How to combine different cost factors into a cost of living index? Use first principal component

- The first principal component maximizes the variance, it spreads out the scores as much as possible

Credit: Marloes Maathuis

2nd and Other Principal Components

- **When all measurements are positively correlated,** the 1st principal component is often some kind of average of the measurements
  - Size of birds
  - Severity index of psychiatric symptoms, …

- **The 2nd and other principal components give important info about the remaining pattern**
  - Shape of birds
  - Pattern of psychiatric symptoms, …

Credit: Marloes Maathuis

Uses of Principal Components

- Dimension reduction
  - Summarize the data with a smaller number of variables, losing as little info as possible
  - Graphical representations of data

- Input for regression analysis
  - Highly correlated explanatory variables are problematic in regression analysis
  - One can replace them by their principal components, which are uncorrelated by definition

Credit: Marloes Maathuis

Growth, 1960, **24**, 339-354.

SIZE AND SHAPE VARIATION IN THE PAINTED TURTLE.[1] A PRINCIPAL COMPONENT ANALYSIS

PIERRE JOLICOEUR AND JAMES E. MOSIMANN[2]

Walker Museum, University of Chicago
and
Institut de Biologie, Université de Montréal

(Received for publication July 11, 1960)

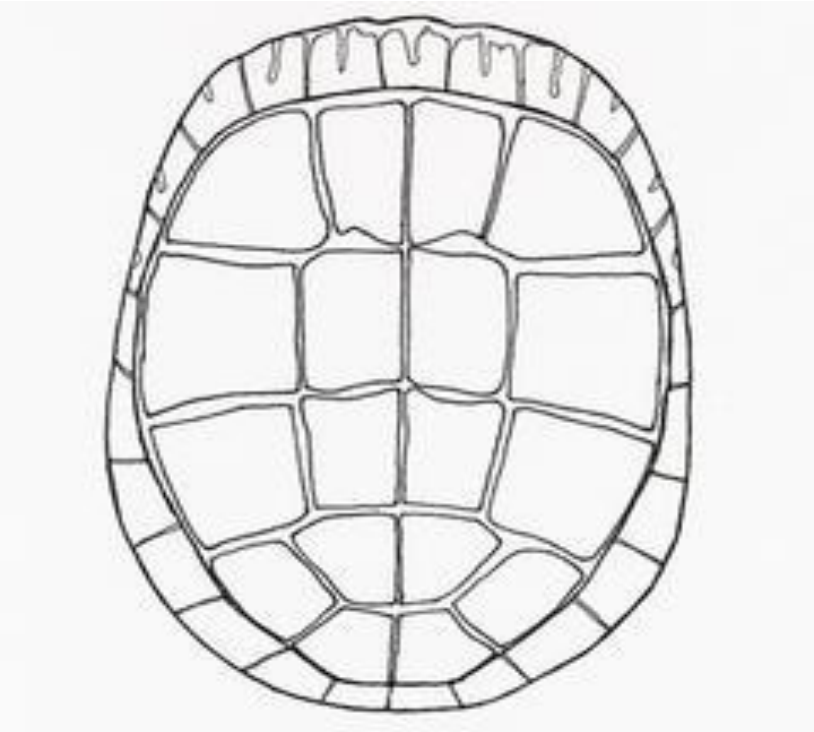
Credit: Alessandro Giuliani

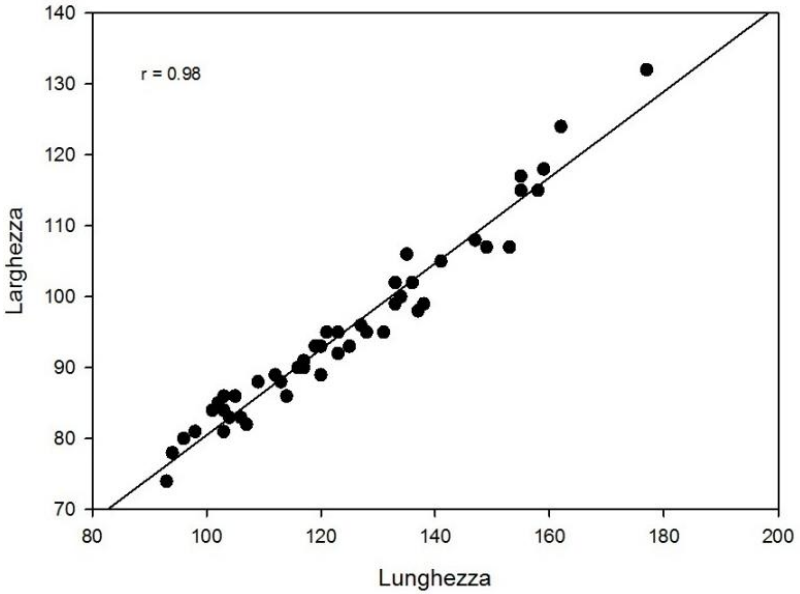
TABLE 1

CARAPACE DIMENSIONS OF PAINTED TURTLES (*Chrysemys picta marginata*) IN MM.

24 Males			24 Females		
length	width	height	length	width	height
93	74	37	98	81	38
94	78	35	103	84	38
96	80	35	103	86	42
101	84	39	105	86	40
102	85	38	109	88	44
103	81	37	123	92	50
104	83	39	123	95	46
106	83	39	133	99	51
107	82	38	133	102	51
112	89	40	133	102	51
113	88	40	134	100	48
114	86	40	136	102	49
116	90	43	137	98	51
117	90	41	138	99	51
117	91	41	141	105	53
119	93	41	147	108	57
120	89	40	149	107	55
120	93	44	153	107	56
121	95	42	155	115	63
125	93	45	155	117	60
127	96	45	158	115	62
128	95	45	159	118	63
131	95	46	162	124	61
135	106	47	177	132	67

Credit: Alessandro Giuliani





Pearson Correlation Coefficients,

	length	width	height
length	1.00000	0.97831	0.96469
width	0.97831	1.00000	0.96057
height	0.96469	0.96057	1.00000

Width = 19,94 + 0,605*Length

Credit: Alessandro Giuliani

Credit: Alessandro Giuliani

Interesting info are often in the 2nd principal component

	PC1 (98%)	PC2 (1.4%)
Length	0,992	-0,067
Width	0,990	-0,100
Height	0,986	0,168

$$PC1 = 33.78*Length + 33.73*Width + 33.57*Height$$

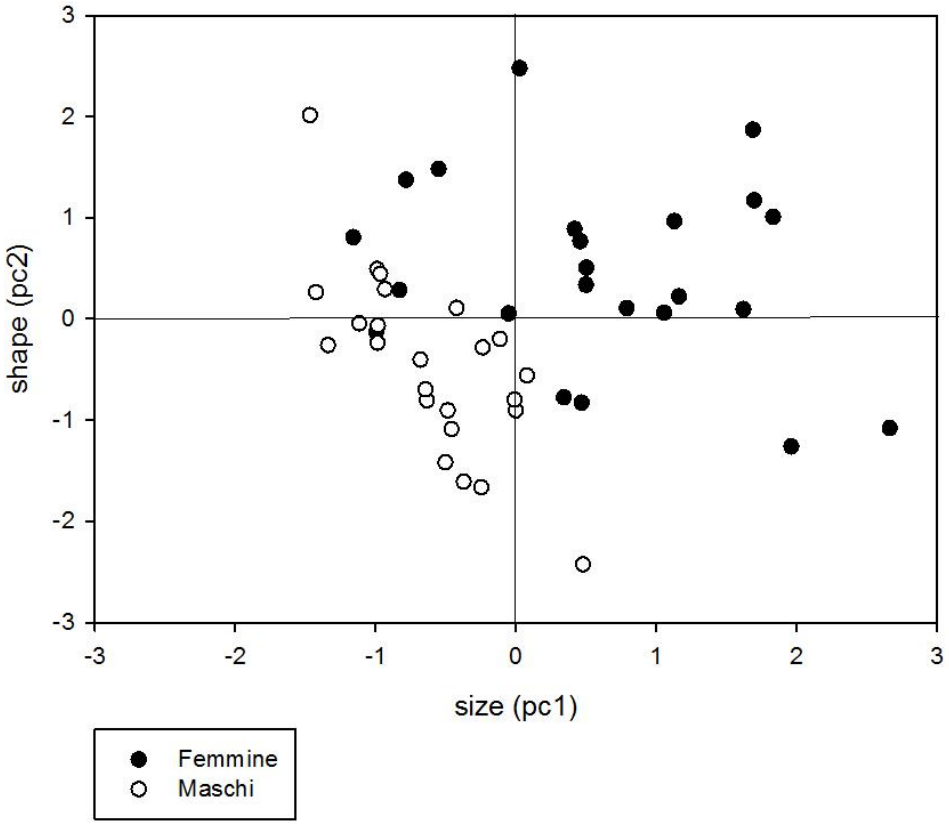
$$PC2 = -1.57*Length - 2.33*Width + 3.93*Height$$

- Presence of an overwhelming size component explaining system variance comes from the presence of a 'typical' common shape
- Displacement along pc = size variation (all positive terms)
- Displacement along pc2 = shape deformation (both positive and negative terms)

unit	sex	Length	Width	Height	PC1(size)	PC2(shape)
T25	F	98	81	38	-1,15774	0,80754832
T26	F	103	84	38	-0,99544	-0,1285916
T27	F	103	86	42	-0,7822	1,37433475
T28	F	105	86	40	-0,82922	0,28526912
T29	F	109	88	44	-0,55001	1,4815252
T30	F	123	92	50	0,027368	2,47830153
T31	F	123	95	46	-0,05281	0,05403839
T32	F	133	99	51	0,418589	0,88961967
T33	F	133	102	51	0,498425	0,33681756
T34	F	133	102	51	0,498425	0,33681756
T35	F	134	100	48	0,341684	-0,774911
T36	F	136	102	49	0,467898	-0,8289156
T37	F	137	98	51	0,457949	0,76721682
T38	F	138	99	51	0,501055	0,50628189
T39	F	141	105	53	0,790215	0,10640554
T40	F	147	108	57	1,129025	0,96505915
T41	F	149	107	55	1,055392	0,06026089
T42	F	153	107	56	1,161368	0,22145593
T43	F	155	115	63	1,687277	1,86903869
T44	F	158	115	62	1,696753	1,17117077
T45	F	159	118	63	1,833086	1,00956637
T46	F	162	124	61	1,962232	-1,261771
T47	F	177	132	67	2,662548	-1,0787317
T48	F	155	117	60	1,620491	0,09690818
T1	M	93	74	37	-1,46649	2,01289241
T2	M	94	78	35	-1,42356	0,26342486
T3	M	96	80	35	-1,33735	-0,258445
T4	M	101	84	39	-0,98842	0,49260881
T5	M	102	85	38	-0,98532	-0,2361914
T6	M	103	81	37	-1,11528	-0,0436547
T7	M	104	83	39	-0,96555	0,44687352
T8	M	106	83	39	-0,93257	0,29353841
T9	M	107	82	38	-0,98269	-0,066727
T10	M	112	89	40	-0,63393	-0,8042059
T11	M	113	88	40	-0,64405	-0,6966061
T12	M	114	86	40	-0,68078	-0,4047389
T13	M	116	90	43	-0,42133	0,10845233
T14	M	117	90	41	-0,48485	-0,9039457
T15	M	117	91	41	-0,45824	-1,0882131
T16	M	119	93	41	-0,37202	-1,610083
T17	M	120	89	40	-0,50198	-1,4175463
T18	M	120	93	44	-0,23552	-0,2831547
T19	M	121	95	42	-0,24581	-1,6640875
T20	M	125	93	45	-0,11305	-0,1986272
T21	M	127	96	45	-0,00023	-0,9047645
T22	M	128	95	45	-0,01035	-0,7971646
T23	M	131	95	46	0,079136	-0,559302
T24	M	135	106	47	0,477846	-2,4250481

Credit: Alessandro Giuliani

Female turtles are larger and have more exaggerated height ☺



Caution: PCA is not scale invariant

- Suppose we have measurements in kg and meters, and we want to have principal components expressed in grams and hectometers
- Option 1: multiply measurements in kg by 1000, multiply measurements in meters by 1/100, and then apply PCA
- Option 2: apply PCA on original measurements, and then re-scale to the appropriate units
- These two options generally give different results!

Credit: Marloes Maathuis

Caution: PCA is sensitive to outliers

Formal definition of PCA - population case

- We first consider the population case: Let $X \in \mathbb{R}^p$ be a random vector with mean μ and covariance matrix Σ (note that we don't make any assumptions about the distribution of X).
- Then the principal component transformation is the transformation

$$X \to Y = \Gamma'(X - \mu)$$

where Γ is the orthogonal matrix consisting of the standardized eigenvectors corresponding to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of Σ. Thus, $\Sigma = \Gamma\Lambda\Gamma'$, or equivalently, $\Gamma'\Sigma\Gamma = \Lambda$.

18 / 33

- PCA is sensitive to outliers, since it is based on the sample covariance matrix Σ which is sensitive to outliers

Credit: Marloes Maathuis

Roadmap

- **Basics of biostatistics**
- **Statistical estimation**
- **Hypothesis testing**
  - Measurement data
  - Categorical data
  - Non-parametric methods
- **Ranking and rating**
- **Principal component analysis**
- **Summary**

Summary

- **Statistical estimation**
  - Confidence interval
- **Ranking & rating**
  - Binomial proportion confidence intervals
    - **Normal approx interval**
    - **Wilson interval**
- **Principal component analysis**
  - Interestingness of 2nd principal component
- **Hypothesis testing**
  - Large sample size: z-test
  - Measurement data
    - **Small sample size & normal distribution: unpaired t-test & paired t-test**
  - Categorical data
    - **Small sample size: χ2-test**
    - **Extremely small sample size: Fisher's exact test**
  - Non-parametric methods
    - **Wilcoxon rank sum test**
    - **Wilcoxon matched pairs signed ranks test**

Summary

- **Many software available to do hypothesis testing**
  - MATLAB, R, SPSS …
- **More important for us to know**
  - When to use which test
  - Interpret the results and draw proper conclusions

Topics not covered

- **Summarizing data with graphs**
  - Bar charts, pie charts, histograms, boxplots, scatter plots, …
- **Hypothesis testing involving >2 samples**
  - ANOVA
- **Association on 3-way contingency tables**
  - Cochran-Mantel-Haenszel (CMH) test
- **Liner correlation and regression**
- **Survival analysis**
- **Sample size estimation**
- **……**

Acknowledgements

- **I adapted these slides from Dr Liu Guimei, who got them from Prof K. C. Lun**

- **I adapted the slides on PCA from a talk given by Dr Alessandro Giuliani**