Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www.yildiz.edu.tr/~naydin

1

Introduction to statistics

2

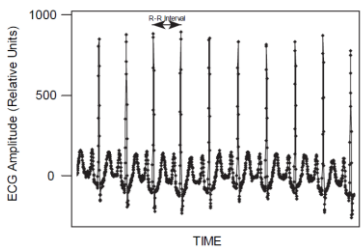
Introduction

- All sorts of data are collected from
  - patients, animals, cell counters, microarrays, imaging systems, pressure transducers, bedside monitors, manufacturing processes, material testing systems, and other measurement systems
- that support a broad spectrum of
  - research,
  - design,
  - manufacturing environments.
- Ultimately, the reason for collecting data is to make a decision.

3

Introduction

- Example of an ECG recording



- R-R interval is defined as the time interval between successive R waves of the QRS complex,

4

Introduction

- A normally functioning heart exhibits considerable variability in beat-to-beat intervals.
  - variability reflects the body's continual effort to maintain homeostasis
    - so that the body may continue to perform its most essential functions and supply the body with the oxygen and nutrients required to function normally.
- It has been demonstrated through biomedical research that there is a loss of heart rate variability associated with some diseases,
  - such as diabetes and ischemic heart disease.

5

Introduction

- Researchers seek to determine
  - if this difference in variability between normal subjects and subjects with heart disease is significant
    - meaning, it is due to some underlying change in biology and not simply a result of chance
  - whether it might be used to predict the progression of the disease.
- One will note that the probability model changes
  - as a consequence of changes in the underlying biological function or process.

6

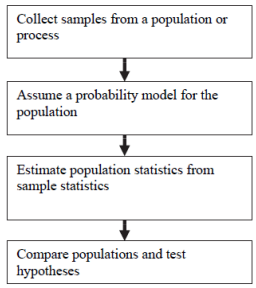
Introduction

- To make sound decisions in the context of the uncertainty with some level of confidence,
  - we need to assume some probability models for the populations from which the samples have been collected.
- Once we have assumed an underlying model,
  - we can select the appropriate statistical tests for comparing two or more populations
  - then we can use these tests to draw conclusions about our hypotheses for which we collected the data in the first place

7

Introduction

- The steps for performing statistical analysis of data.



8

Collecting Data and Experimental Design

- The value of any statistical analysis is only as good as the data collected
- Because we are using data or samples to draw conclusions about entire populations or processes,
  - it is critical that the data collected are representative of the larger, underlying population.
    - we must have enough samples to represent the variability of the underlying population

9

Collecting Data and Experimental Design

- Capturing variability is often the greatest challenge that scientists/engineers face in collecting data and using statistics to draw meaningful conclusions.
  - The experimentalist must ask questions such as the following:
    - What type of person, object, or phenomenon do I sample?
    - What variables that impact the measure or data can I control?
    - How many samples do I require to capture the population variability to apply the appropriate statistics and draw meaningful conclusions?
    - How do I avoid biasing the data with the experimental design?

10

Collecting Data and Experimental Design

- Experimental design is the most critical step to support the statistical analysis
  - that will lead to meaningful conclusions and hence sound decisions.
- Two elements of experimental design that are critical to prevent biasing the data or selecting samples that do not fairly represent the underlying population are
  - randomization
  - blocking.

11

Collecting Data and Experimental Design

- Randomization
  - the process by which we randomly select samples or experimental units from the larger underlying population such that we maximize our chance of capturing the variability in the underlying population.
- Blocking
  - the arranging of experimental units in groups (blocks) that are similar to one another
    - Blocking will help to eliminate the effect of intersubject variability.

12

Collecting Data and Experimental Design

- Some important concepts and definitions to keep in mind when designing experiments:
  - experimental unit
    - the item, object, or subject to which we apply the treatment and from which we take sample measurements
  - randomization
    - allocate the treatments randomly to the experimental units
  - blocking
    - assigning all treatments within a factor to every level of the blocking factor

13

Collecting Data and Experimental Design

- The experimentalist must always think about how representative the sample population is with respect to the greater underlying population.
- Because it is virtually impossible to test every member of a population, the scientist/engineer must often collect data from a much smaller sample drawn from the larger population.
- It is important, if the statistics are going to lead to useful conclusions,
  - that the sample population captures the variability of the underlying population.

14

Why statistics?

- Reasons for using statistical data summary and analysis:
  - The real world is full of random events that cannot be described by exact mathematical expressions
  - Variability is a natural and normal characteristic of the natural world
  - We like to make decisions with some confidence.
    - This means that we need to find trends within the variability

15

Questions to address

- There are several basic questions we hope to address when using numerical and graphical summary of data:
  - Can we differentiate between groups or populations?
    - probably the most frequent aim of biomedical research
  - Are there correlations between variables or populations?
  - Are processes under control?
    - Such a question may arise if there are tight controls on the manufacturing specifications for a medical device

16

Graphical summarization of data

- Before blindly applying the statistical analysis, it is always good to look at the raw data,
  - usually in a graphical form,

  and then use graphical methods to summarize the data in an easy to interpret format.
  - A Picture is worth a thousand word
- The types of graphical displays that are most frequently used by scientists/engineers
  - scatterplots, time series, box-and-whisker plots, and histograms.

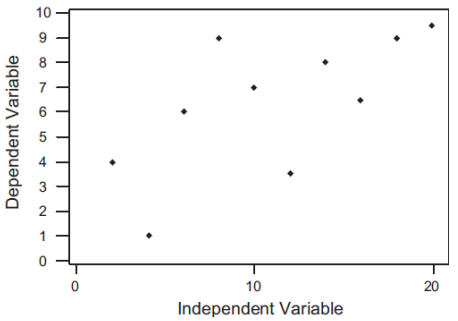
17

Scatterplots

- graphs the occurrence of one variable with respect to another.
- In most cases, one of the variables may be considered the independent variable
  - such as time or subject number,
- the second variable is considered the dependent variable.
- Next slide illustrates an example of a scatterplot for two sets of data.

18

Scatterplots



19

Scatterplots

- In general, we are interested in whether there is a predictable relationship that
  - maps our independent variable
    - such as respiratory rate
  - into our dependent variable
    - such as heart rate
- If there is a linear relationship between the two variables,
  - the data points should fall close to a straight line

20

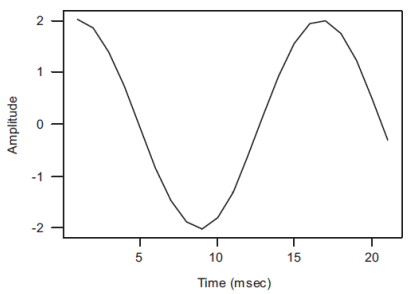
Time Series

- used to plot the changes in a variable as a function of time.
- The variable is usually
  - a physiological measure,
    - such as electrical activation in the brain or hormone concentration in the blood stream, that changes with time.
- Next slide illustrates an example of a time series plot.

21

Time Series

- a simple sinusoid function changing with time.



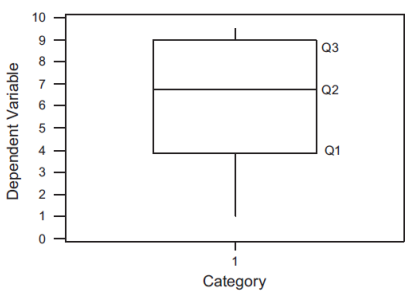
22

Box-and-Whisker Plots

- illustrate the 1st, 2nd, and 3rd quartiles as well as the minimum and maximum values of the data collected.
  - The 2nd quartile (Q2) is also known as the median of the data.
  - The 1st quartile (Q1) can be thought of as the median value of the samples that fall below the 2nd quartile.
  - The 3rd quartile (Q3) can be thought of as the median value of the samples that fall above the 2nd quartile.
- Box-and-whisker plots are useful in that they highlight whether there is skew to the data or any unusual outliers in the samples.

23

Box-and-Whisker Plots



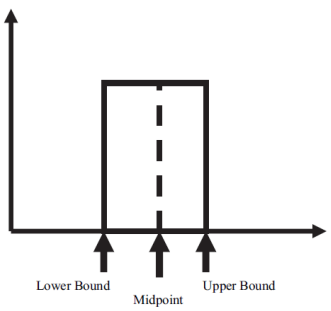
24

Histogram

- defined as a frequency distribution.
- Given N samples or measurements, x_i ranging from X_{min} to X_{max}, the samples are grouped into nonoverlapping intervals (bins), usually of equal width.
  - Typically, the number of bins is on the order of 7–14, depending on the nature of the data.
  - In addition, we typically expect to have at least 3 samples per bin.
    - Sturgess'rule may also be used to estimate the number of bins and is given by $k = 1 + 3.3 \log(n)$.
      - where k is the number of bins and n is the number of samples.

25

Histogram



- One bin of a histogram plot
- The bin is defined by
  - a lower bound,
  - a midpoint,
  - an upper bound

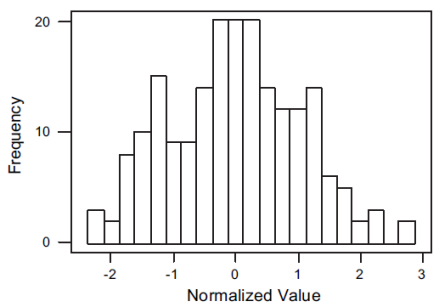
26

Histogram

- constructed by plotting the number of samples in each bin.
  - horizontal axis,
    - the sample value,
  - the vertical axis,
    - the number of occurrences of samples falling within a bin
- Next slide illustrates a histogram for 1000 samples drawn from a normal distribution with mean (μ) = 0 and standard deviation (σ) = 1.0.

27

Histogram



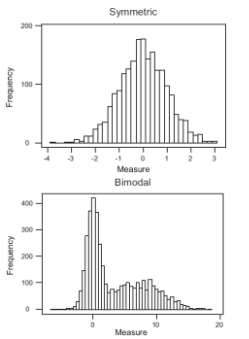
28

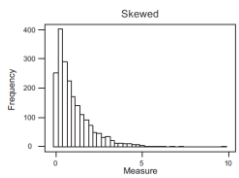
Histogram

- Two useful measures in describing a histogram:
  - the absolute frequency in one or more bins
    - f_i = absolute frequency in ith bin
  - the relative frequency in one or more bins
    - f_i / n = relative frequency in ith bin,
      - where n is the total number of samples being summarized in the histogram
- The histogram can exhibit several shapes
  - symmetric, skewed, or bimodal.

29

Histogram





- In each case, 2000 samples were drawn from the underlying populations.

30

Histogram

- For example, a skewed histogram may be attributed to the following:
  - mechanisms of interest that generate the data
    - e.g., the physiological mechanisms that determine the beat-to-beat intervals in the heart
  - an artifact of the measurement process or a shift in the underlying mechanism over time
    - e.g., there may be time-varying changes in a manufacturing process that lead to a change in the statistics of the manufacturing process over time
  - a mixing of populations from which samples are drawn
    - this is typically the source of a bimodal histogram

31

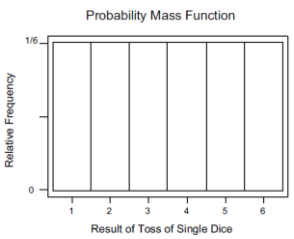
Histogram

- The histogram is important because it serves as
  - a rough estimate of the true probability density function or
  - probability distribution of the underlying random process from which the samples are being collected.
- The probability density function or probability distribution is a function that quantifies the probability of a random event, *x*, occurring.
  - When the underlying random event is discrete in nature,
    - we refer to the probability density function as the probability mass function

32

Histogram

- The probability density function for a discrete random variable (probability mass function).

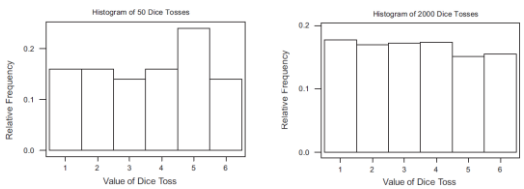


- In this case, the random variable is the value of a toss of a single dice.
  - Note that each of the six possible outcomes has a probability of occurrence of 1 of 6.
- This probability density function is also known as a uniform probability distribution.

33

Histogram

- Histograms representing the outcomes of experiments in which a single dice is tossed 50 and 2000 times, respectively



  - Note that as the sample size increases, the histogram approaches the true probability distribution (uniform probability distribution)

34

Histogram

- Scientists/engineers are trying to make decisions about populations or processes to which they have limited access.
- Thus, they design experiments and collect samples that they think will fairly represent the underlying population or process.
- Regardless of what type of statistical analysis will result from the investigation or study,
  - all statistical analysis should follow the same general approach:
    - Measure a limited number of representative samples from a larger population.
    - Estimate the true statistics of larger population from the sample statistics.

35

- Once the researcher has estimated the sample statistics from the sample population,
  - he or she will try to draw conclusions about the larger (true) population.
- The most important question to ask when reviewing the statistics and conclusions drawn from the sample population is
  - how well the sample population represents the larger, underlying population.

36

DESCRIPTIVE STATISTICS

- Once the data have been collected, we use some basic descriptive statistics to summarize the data.
- Basic descriptive statistics include the following general measures:
  - central tendency,
  - variability,
  - correlation.

37

DESCRIPTIVE STATISTICS

- There are a number of descriptive statistics
  - that help us to picture the distribution of the underlying population.
- Ultimate goal is to
  - assume an underlying probability model for the population and then
  - select the statistical analyses that are appropriate for that probability model.

38

DESCRIPTIVE STATISTICS

- The underlying model for any sample, or measure (the outcome of the experiment) is as follows:

  $X = \mu \pm$ individual differences $\pm$ situational factors $\pm$ unknown variables,
  - where X is our measure or sample value and is influenced by μ, which is the true population mean;
  - individual differences such as genetics, training, motivation, and physical condition;
  - situational factors, such as environmental factors; and
  - unknown variables such as unidentified/nonquantified factors that behave in an unpredictable fashion from moment to moment.

39

Measures of Central Tendency

- A central tendency is a central or typical value for a probability distribution.
  - also called a center or location of the distribution.
- Measures of central tendency are often called averages.
- There are several measures that reflect the central tendency
  - sample mean,
  - sample median,
  - sample mode.

40

Mean

- In mathematics, mean has several different definitions depending on the context.
- In probability and statistics,
  - mean and expected value are synonymous
- In the case of a discrete probability distribution of a random variable X,
  - the mean is equal to the sum over every possible value weighted by the probability of that value;

$$\mu = \sum_{i} xP(x)$$

41

Mean

- For a data set, the terms arithmetic mean, mathematical expectation, and sometimes average are used synonymously to refer to a central value of a discrete set of numbers
  - specifically, the sum of the values divided by the number of values.
- If the data set were based on a series of observations obtained by sampling from a statistical population,
  - the arithmetic mean is termed the sample mean to distinguish it from the population mean

42

Mean

- Outside of probability and statistics, a wide range of other notions of mean are often used in geometry and analysis:
  - Pythagorean means
    - Arithmetic mean, Geometric mean, Harmonic mean
  - Generalized means
    - Power mean, f-mean
  - Weighted arithmetic mean
  - Truncated mean
  - Interquartile mean
  - Fréchet mean
  - …

43

Mean

- Arithmetic mean (or simply mean) of a sample $x_1, x_2, \ldots, x_n$, usually denoted by $\bar{x}$

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product, e.g. rates of growth

$$\bar{x} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

44

Mean

- Harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, for example speed

$$\bar{x} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

- Arithmetic mean, Geometric mean, and Harmonic mean satisfy these inequalities:

  Arithmetic mean ≤ Geometric mean ≤ Harmonic mean

- Equality holds if and only if all the elements of the given sample are equal

45

Mean

- Weighted arithmetic mean is used if one wants to combine average values from samples of the same population with different sample sizes

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

- The weights w_i represent the sizes of the different samples.
- In other applications they represent a measure for the reliability of the influence upon the mean by the respective values

46

Mean (Arithmetic)

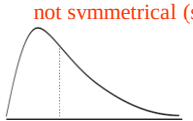
- It is used when the spread of the data is fairly similar on each side of the mid point,
  - when the data are "normally distributed".

- If a value (or a number of values) is a lot smaller or larger than the others, "skewing" the data, the mean will then not give a good picture of the typical value.

47

Median

- Sometimes known as the mid-point.
  - It is used to represent the average when the data are not svmmetrical (skewed distribution)



- The median value of a group of observations or samples, x_i, is the middle observation when samples, x_i, are listed in descending order.

- Note that if the number of samples, n, is odd, the median will be the middle observation.
- If the sample size, n, is even, then the median equals the average of two middle observations.
- Compared with the sample mean, the sample median is less susceptible to outliers.

48

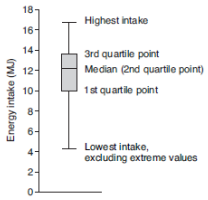
Median

- The median may be given with its inter-quartile range (IQR).
- The 1st quartile point has the 1⁄4 of the data below it
- The 3rd quartile point has the 3⁄4 of the sample below it
- The IQR contains the middle 1⁄2 of the sample
- This can be shown in a "box and whisker" plot.

49

Median (example)

- A dietician measured the energy intake over 24 hours of 50 patients on a variety of wards. One ward had two patients that were "nil by mouth". The median was 12.2 megajoules, IQR 9.9 to 13.6. The lowest intake was 0, the highest was 16.7.
- This distribution is represented by the box and whisker plot below.

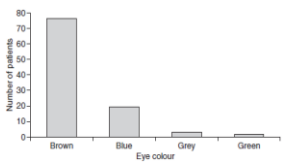


- Box and whisker plot of energy intake of 50 patients over 24 hours.
- The ends of the whiskers represent the maximum and minimum values, excluding extreme results like those of the two "nil by mouth" patients.

50

Mode

- the most common of a set of events
  - used when we need a label for the most frequently occurring event
    - Example: An eye clinic sister noted the eye colour of 100 consecutive patients. The results are shown below

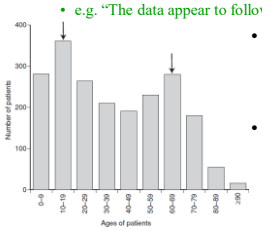


- Graph of eye colour of patients attending an eye clinic.
- In this case the mode is brown, the commonest eye colour.

51

Mode

- You may see reference to a bi-modal distribution.
  - Generally when this is mentioned in papers it is as a concept rather than from calculating the actual values,
    - e.g. "The data appear to follow a bi-modal distribution".

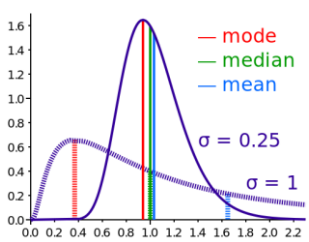


- Graph of ages of patients with asthma in a practice
  - The arrows point to the modes at ages 10–19 and 60–69.
- Bi-modal data may suggest that two populations are present that are mixed together,
  - so an average is not a suitable measure for the distribution.

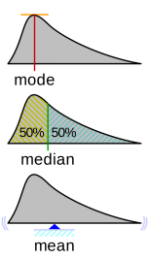
52

Mean, Median, Mode

- Comparison of the arithmetic mean, median and mode of two skewed (log-normal) distributions.



- Geometric visualisation of the mode, median and mean of an arbitrary probability density function.



53

An applicaton of mean: moving AVERAGE filter

- Highlights trends in a signal (smoothing, blurring)

$$x_n : n = 1,\ldots,N$$

$$y_n = \sum_{j=-k}^{k} w_j x_{n+j} \quad : \quad n = k+1,\ldots,N-k,$$

k: pozitif integer, w_j: weights, $\Sigma\, w_j = 1$

- Algorithm

```
for n=1:N
        y(n)=0.5*(x(n)+x(n+1));
end
```

- Example (2 point moving AVERAGE filter)

$x=(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, \ldots)$

$y=([x_1+x_2]/2, [x_2+x_3]/2, [x_3+x_4]/2, \ldots)$

54

Complementary procedure: moving DIFFERENCE filter

- Removes trends from a signal (sharpening, deblurring)
- 1st order differencing

  $Dy_t=y_t-y_{t-1}$
- Higher order differences (2nd order)

  $D^2y_t= D(Dy_t)=Dy_t-Dy_{t-1}=y_t-2y_{t-1}+y_{t-2}$
- Example (1st order moving DIFFERENCE filter)

  $x=(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, \ldots)$

  $y=([x_2-x_1], [x_3-x_2], [x_4-x_3], \ldots)$

55

Moving MEDIAN filtering

- Useful in impulsive noise removal (image processing, moving/sliding MEDIAN filtering)
- Example (3 point moving MEDIAN filtering)

  $x=(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, \ldots)$

  $y=(\text{med}[x_1,x_2,x_3], \text{med}[x_2,x_3,x_4], \text{med}[x_3,x_4,x_5], \ldots)$
- If a window with even number of samples are selected median is average of two mid-point samples

56

Convolution

- Is a mathematical way of combining two signals to form a third signal
- Is the relationship between a system's input signal, output signal, and impulse response

$$y(k)=\sum_{n=0}^{N-1} h(n)x(k-n)$$

57

Correlation

- Is a measure of similarities of two signals (cross-correlation)

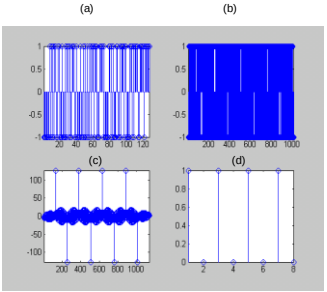
$$r_{xy}(k)=\sum_{n=0}^{N-1} x(n)y(k+n)$$

- Is a way to detect a known waveform in a noisy background (matched filter)
- Algorithm

```
for k=1:K+N-1
    for n=1:N
        y(k)=y(k)+a(n)*b(k+n-1);
    end
end
```

58

A correlation example



(a) A PN code
(b) A noisy binary signal (10101010) coded by the PN code in (a)
(c) Result of the correlation between (a) and (b)
(d) Recovered signal (10101010) after thresholding

59

Measures of Variability

- When summarizing the variability of a population or process, we typically ask,
  - "How far from the center (sample mean) do the samples (data) lie?"
- To answer this question, we typically use the following estimates that represent the spread of the sample data:
  - interquartile ranges,
  - sample variance,
  - sample standard deviation.

60

Measures of Variability

- The interquartile range is the difference between the 1st and 3rd quartiles of the sample data.
- For sampled data, the median is also known as the 2nd quartile, Q2.
- Given Q2, we can find the 1st quartile, Q1, by simply taking the median value of those samples that lie below the 2nd quartile.
- We can find the 3d quartile, Q3, by taking the median value of those samples that lie above the 2nd quartile.

61

Measures of Variability

- As an illustration, we have the following samples:
  1, 3, 3, 2, 5, 1, 1, 4, 3, 2.
- If we list these samples in descending order,
  5, 4, 3, 3, 3, 2, 2, 1, 1, 1,
- the median value (2nd quartile) for these samples is 2.5.
- The 1st quartile, Q1, can be found by taking the median of the following samples,
  2.5, 2, 2, 1, 1, 1,
  which is 1.5.
- The 3d quartile, Q3, may be found by taking the median value of the following samples:
  5, 4, 3, 3, 3, 2.5,
  which is 3.
- Thus, the interquartile range,
  Q3 − Q1 = 3 − 1.5 = 2.

62

Measures of Variability

- Sample variance, s^2, is defined as the "average distance of data from the mean"
- The formula for estimating s^2 from a collection of samples, x_i, is

$$s^2 = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})^2$$

- Sample standard deviation, s, which is more commonly referred to in describing the variability of the data is

$$s = \sqrt{s^2}$$

63

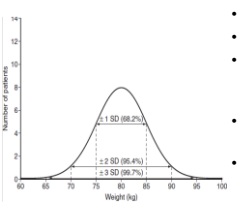
Measures of Variability

- Standard deviation (SD) is used for data which are "normally distributed", to provide information on how much the data vary around their mean.
  - SD indicates how much a set of values is spread around the average.
    - A range of one SD above and below the mean (abbreviated to ± 1 SD) includes 68.2% of the values.
    - ± 2 SD includes 95.4% of the data.
    - ± 3 SD includes 99.7%.

64

Measures of Variability

- Example:
- Let us say that a group of patients enrolling for a trial had a normal distribution for weight. The mean weight of the patients was 80 kg. For this group, the SD was calculated to be 5 kg.
- normal distribution of weights of patients enrolling in a trial with mean 80 kg, SD 5 kg.



- 1 SD below the average is 80 – 5 = 75 kg.
- 1 SD above the average is 80 + 5 = 85 kg.
- ± 1 SD will include 68.2% of the subjects, so 68.2% of patients will weigh between 75 and 85 kg.
- 95.4% will weigh between 70 and 90 kg (± 2 SD).
- 99.7% of patients will weigh between 65 and 95 kg (± 3 SD)

65

Measures of Variability

- It is important to note that for normal distributions (symmetrical histograms), sample mean and sample deviation are the only parameters needed to describe the statistics of the underlying phenomenon.
- Thus, if one were to compare two or more normally distributed populations, one only need to test the equivalence of the means and variances of those populations.

66

Gaussian Distribution…

- The spread (distribution) of data may be rectangular, skewed, Gaussian, or other.
- The Gaussian distribution is given by

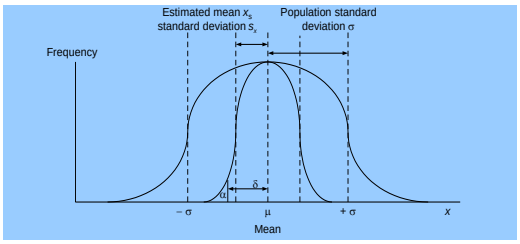
$$f(X) = \frac{e^{-(X-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma}$$

where μ is the true mean and σ is the true standard deviation of a very large number of measurements.

67

…Gaussian Distribution

- For the normal distribution, 68% of the data lies within ±1 standard deviation. By measuring samples and averaging, we obtain the estimated mean , which has a smaller standard deviation s_x. α is the tail probability that x_s does not differ from μ by more than δ.



68

Poisson Probability…

- The Poisson probability density function is another type of distribution.
  - It can describe, among other things, the probability of radioactive decay events, cells flowing through a counter, or the incidence of light photons.
- The probability that a particular number of events K will occur in a measurement (or during a time) having an average number of events m is

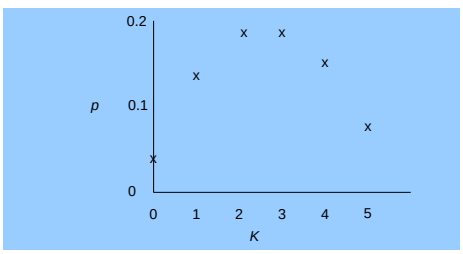
$$p(K,m) = \frac{e^{-m}m^K}{K!}$$

- The standard deviation of the Poisson distribution is $\sqrt{m}$

69

…Poisson Probability

- A typical Poisson distribution for m = 3.



70

Hypothesis testing…

- In hypothesis testing, there are two hypotheses.
  - H_0, the null hypothesis,
    - a hypothesis that assumes that the variable in the experiment will have no effect on the result
  - H_a is the alternative hypothesis that states that the variable will affect the results.
- For any population, one of the two hypotheses must be true.
- The goal of hypothesis testing is to find out which hypothesis is true by sampling the population.
- In reality, H_0 is either true or false and we draw a conclusion from our tests of either true or false.
- This leads to four possibilities (next slide)

71

…Hypothesis testing…

- The four outcomes of hypothesis testing.

Conclusion	Real situation	
	H_0 true	H_a true
Accept H_0	Correct decision	Type II error, $p = b$
Reject H_0	Type I error, $p = a$	Correct decision

72

…Hypothesis testing…

- Equivalent table of the table given in previous slide for results relating to a condition or disease.

Test result	Has condition?	
	No	Yes
Negative	True negative (TN)	False negative (FN)
Postitive	False positive (FP)	True positive (TP)

73

…Hypothesis testing…

- The terms in the Table in previous slide are useful for defining measures that
  - describe the proportion of, for example, a disease in a population and the success of a test in identifying it.
- Incidence
  - is the number of cases of a disease during a stated period, such as *x* cases per 1000 per year.

74

…Hypothesis testing…

- Prevalence
  - the number of cases of a disease at a given time such as *y* cases per 1000.
- It is all diseased persons divided by all persons.

$$\text{Prevalence} = \frac{TP + FN}{TN + TP + FN + FP}$$

75

…Hypothesis testing…

- Sensitivity
  - the probability of a positive test result when the disease is present.
  - Among all diseased persons, it is the percent who test positive.

$$\text{Sensitivity} = \frac{TP}{TP + FN} 100\%$$

- Specificity
  - the probability of a negative diagnostic result in the absence of the disease.
  - Among all normal persons, it is the percent who test negative.

$$\text{Specificity} = \frac{TN}{TN + FP} 100\%$$

76

…Hypothesis testing…

- Considering only those who test positive,
  - positive predictive value (PPV) is the ratio of patients who have the disease to all who test positive.

$$PPV = \frac{TP}{TP + FP} 100\%$$

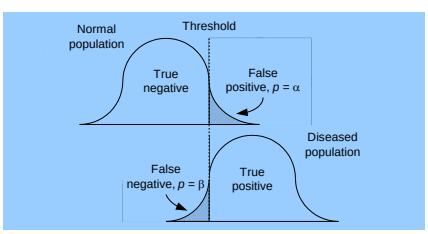
- Considering only those who test negative,
  - negative predictive value (NPV) is the ratio of nondiseased patients to all who test negative.

$$NPV = \frac{TN}{TN + FN} 100\%$$

77

…Hypothesis testing

- The test result threshold is set to minimize false positives and false negatives.



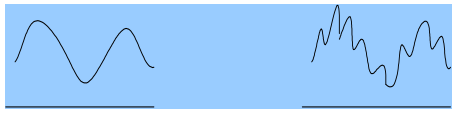
78

Errors in measurements…

- When we measure a variable, we seek to determine the true value.
- This true value may be corrupted by a variety of errors.
- For example
  - Movement of electrodes on the skin may cause an undesired added voltage called an artifact.
  - Electric and magnetic fields from the power lines may couple into the wires and cause an undesired added voltage called interference
  - Thermal voltages in the amplifier semiconductor junctions may cause an undesired added random voltage called noise.
  - Temperature changes in the amplifier electronic components may cause undesired slow changes in voltage called drift.
- We must evaluate each of these error sources to determine their size and what we can do to minimize them.

79

…Errors in measurements…



(a) (b)

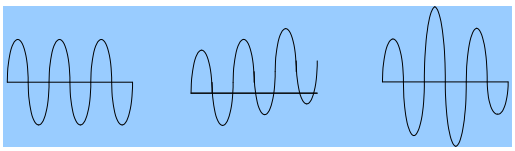
(a) Signals without noise are uncorrupted.

(b) Interference superimposed on signals causes error.

Frequency filters can be used to reduce noise and interference.

80

…Errors in measurements…



(a) (b) (c)

(a) Original waveform.

(b) An interfering input may shift the baseline.

(c) A modifying input may change the gain.

81

Accuracy and precision…

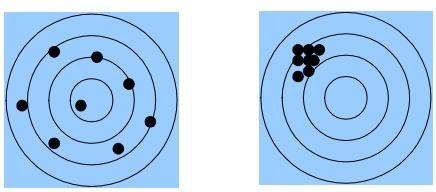
- Resolution
  - the smallest incremental quantity that can be reliably measured.
    - a voltmeter with a larger number of digits has a higher resolution than one with fewer digits.
  - However, high resolution does not imply high accuracy.
- Precision
  - the quality of obtaining the same output from repeated measurements from the same input under the same conditions.
  - High resolution implies high precision.
- Repeatability
  - the quality of obtaining the same output from repeated measurements from the same input over a period of time.

82

…Accuracy and precision…

- Data points with

(a) low precision (b) high precision



83

…Accuracy and precision…

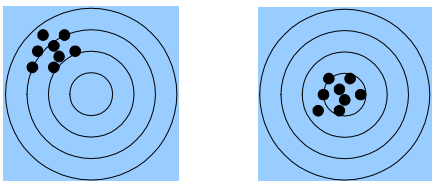
- Accuracy
  - the difference between the true value and the measured value divided by the true value.
- Obtaining the highest possible precision, repeatability, and accuracy is a major goal in bioinstrumentation design.

84

…Accuracy and precision…

- Data points with

(a) low accuracy (b) high accuracy



85

Calibration…

- Measuring instruments should be calibrated against a standard that
  - has an accuracy 3 to 10 times better than the desired calibration accuracy.
- The accuracy of the standard should be traceable to the institutions regulating the standards
  - National Institute of Standards and Technology, TSI, etc. .

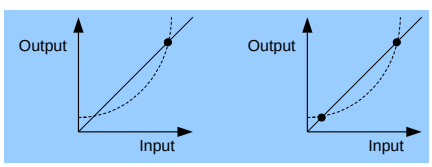
86

Calibration…

- If the instrument is linear,
  - its output can be set to zero for zero input.
  - Then a one-point calibration defines the calibration curve that plots output versus input (next slide).
- If the linearity is unknown,
  - a two-point calibration should be performed and these two points plus the zero point plotted to ensure linearity (next slide).
- If the resulting curve is nonlinear,
  - many points should be measured and plotted to obtain the calibration curve.
- If the output cannot be set to zero for zero input,
  - measurements should be performed at zero and full scale for linear instruments and at more points for nonlinear instruments.
- Calibration curves should be obtained at several expected temperatures to determine temperature drift of the zero point and the gain.

87

…Calibration



(a) (b)

(a) The one-point calibration may miss nonlinearity.
(b) The two-point calibration may also miss nonlinearity.

88