BME 2200: BME Biostatistics and Research Methods

Lecture 5: Representing data statistics

Lucas C. Parra
Biomedical Engineering Department
City College of New York

CCNY

parra@ccny.cuny.edu

Content, Schedule

1. Scientific literature:
- Literature search
- Structure biomedical papers, engineering papers, technical reports
- Experimental design, correlation, causality.

2. Presentation skills:
- Report – Written report on literature search (individual)
- Talk – Oral presentation on biomedical implant (individual and group)

3. Graphical representation of data:
- Introduction to MATLAB
- Plot formats: line, scatter, polar, surface, contour, bar-graph, error bars. etc.
- Labeling: title, label, grid, legend, etc.
- Statistics: histogram, percentile, mean, variance, standard error, box plot

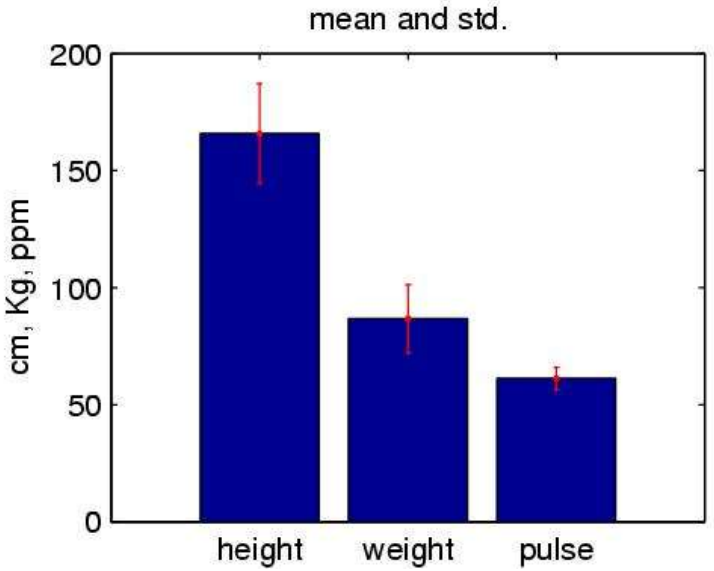
4. Biostatistics:
- Basics of probability
- t-Test, ANOVA
- Linear regression, cross-validation
- Error analysis
- Test power, sensitivity, specificity, ROC analysis

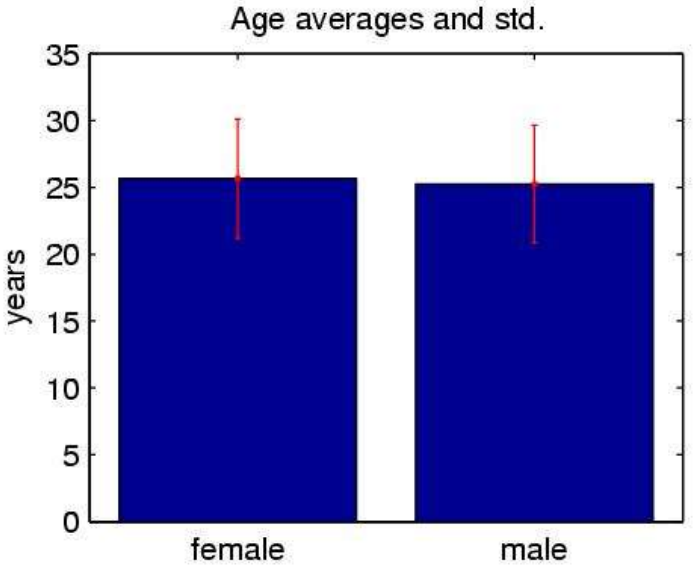
How to represent data statistics

Showing all the raw data is sometimes not useful. Showing the **mean** and **standard deviation** may be better:

$$\langle x \rangle = \frac{1}{N}\sum_{i=1}^{N} x_n \qquad s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_n - \langle x \rangle\right)^2}$$

Standard deviation measures how much data deviates from the mean.



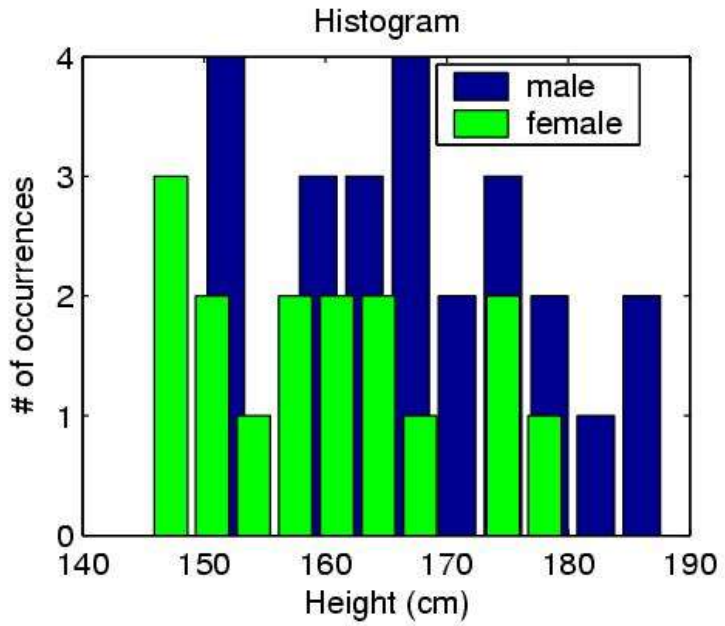


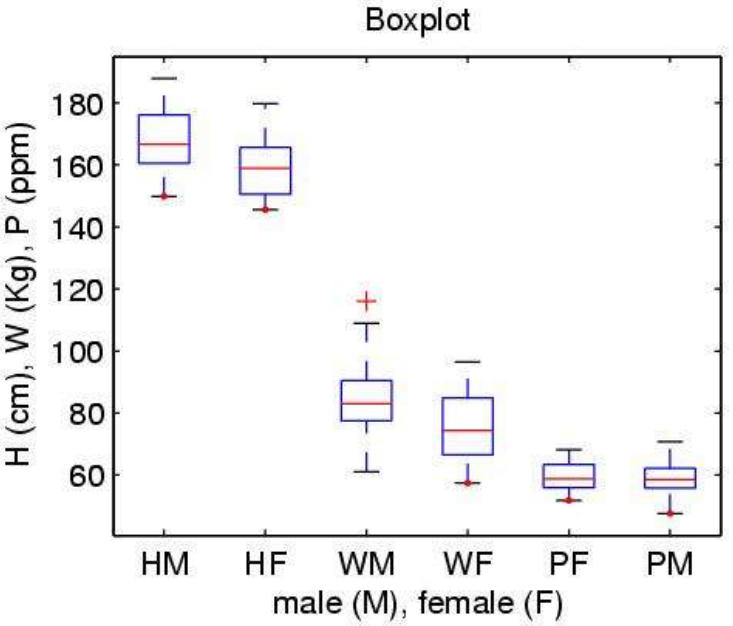

```
bar(mean(x)); errorbar(1:2,mean(x),-std(x),std(x),'.r');
```

How to represent data statistics

To get an idea on how two samples differ use **histogram**.

Or to compare more data look at some simple statistics: **standard deviation, percentile, and outliers**:



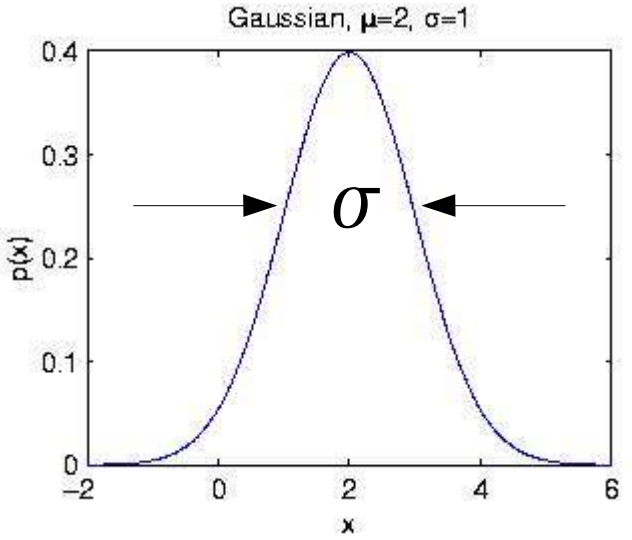



```
hist(height)
boxplot([height;weight;pulse])
```

Normal distribution – mean and std

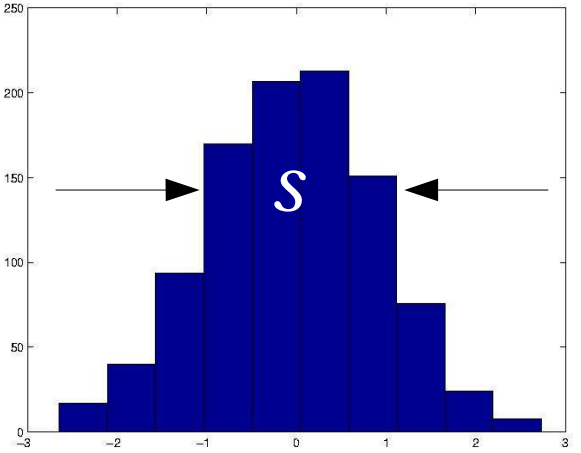
Perhaps the most important distribution is the normal distribution with Gaussian probability density function:

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\equiv N(\mu,\sigma)$$



where the mean and variance can be easily estimated from a sample:

$$\mu\approx\langle x\rangle \qquad \sigma\approx s$$

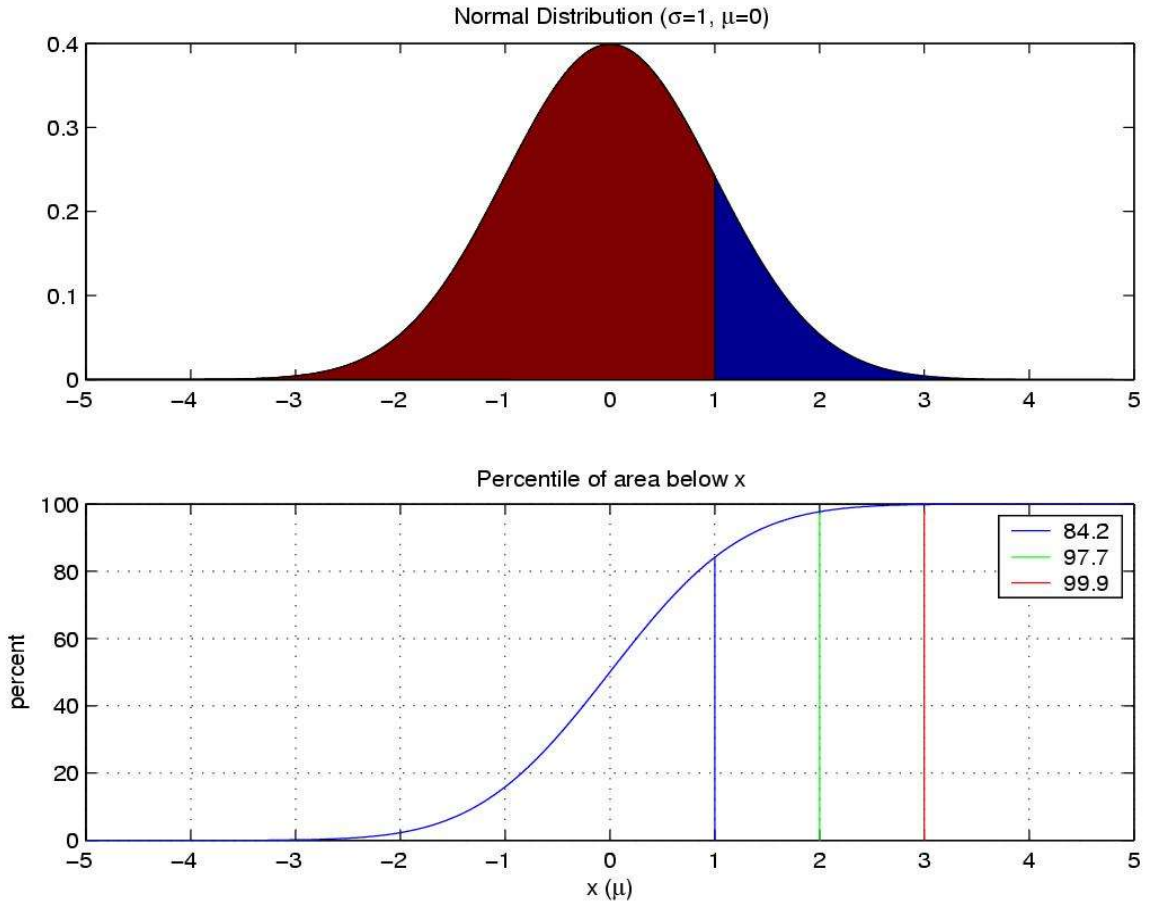


To generate a normal distributed variable with mean `m` and std `s` use:

```
x = m + s*randn(N,1);
hist(x)
```

Normal distribution - Percentile

For the normal distributed variable 68.3% of the data lies within one standard deviation and 95.5% within two std.




```
fill(x,p,1); hold on
fill([x(1:n) x(n)],[p(1:n) 0],2); hold off
```

Interpreting mean and std

When quoting a mean and standard deviation implicit one is assuming normal distributed data. This can be misleading as these examples show:

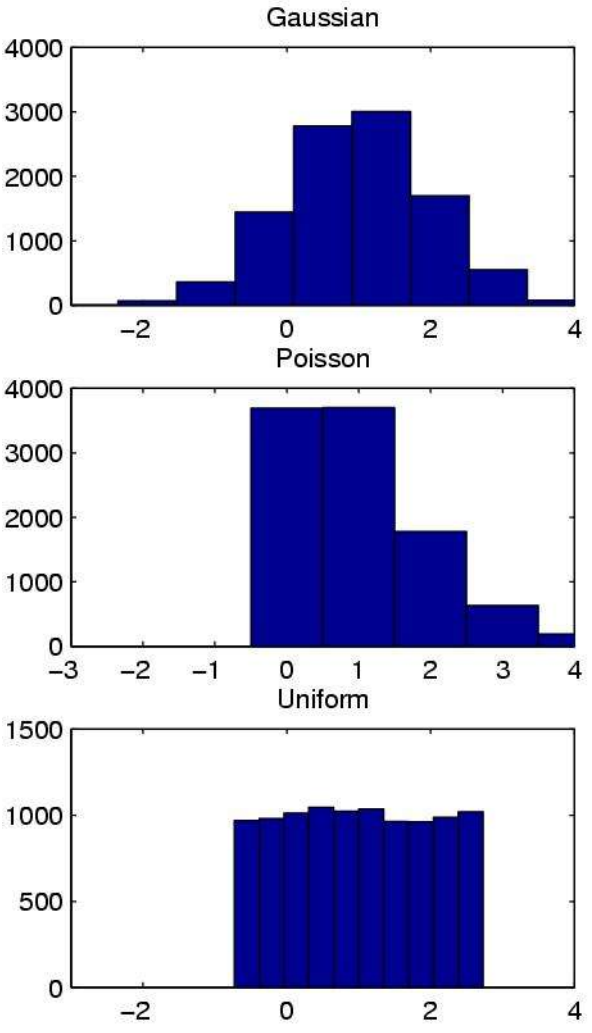
$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(k)=\frac{\lambda^k}{k!}e^{-\lambda}$$

$$p(x)=\begin{cases}1 & -0.5\le x\le 0.5\\ 0 & else\end{cases}$$

All data sampled from distributions with:

$$\mu=1 \quad \sigma=1$$



Standard deviation and standard error

Note the difference between the **standard deviation** and the **standard error** of the mean (SEM).

$$s_{\langle x \rangle} = \frac{1}{\sqrt{N}} s_x$$

Sample: A set of N data points

Standard deviation: measures the variability from one data point to the next within a given sample. Use this to compare individuals.

Standard error: estimates the variability of the mean with sample size N, i.e. How much does the mean change if I repeatedly draw a new set of N values. *Assumes Gaussian distributed data*. Use this to compare groups of individuals.

Central limit theorem

We will often be comparing groups of subjects by looking at their mean values from samples.

So how good is this Gaussian assumption in practice? In particular for the standard errors?

The answer is: **pretty good!** Provided that

- N is large and
- the samples are independent!

Thanks to the **Central Limit Theorem** which states that:

The sum of many independent values drawn from the same distribution ***is Gaussian*** distributed.

Central limit theorem

Let $X_1, X_2, \ldots$ be independently drawn samples from an arbitrary distribution with mean μ and variance σ^2.

Consider the **sample average**:

$$\langle X_n \rangle = \frac{1}{n} \sum_{k=1}^{n} X_k$$

The **Law of Large Numbers** states that sample average converges to the ensemble average

$$\lim_{n \to \infty} \langle X_n \rangle = \mu$$

The C**entral Limit Theorem** states that sample average is normal

$$\lim_{n \to \infty} p(\langle X_n \rangle) = N(\mu, \sigma / \sqrt{n})$$

Programming and Reading Assignment

Assignment 6:

- Write a MATLAB program that generates figures 3, 4, 6 using the data from the class.
- Read chapter 1 and 2 in Glantz: Biostatistics

- Voluntary: For a more in-depth study read chapter 3 and 5 in Schaum's outlines: Probability and Statistics.