Advanced Engineering Statistics

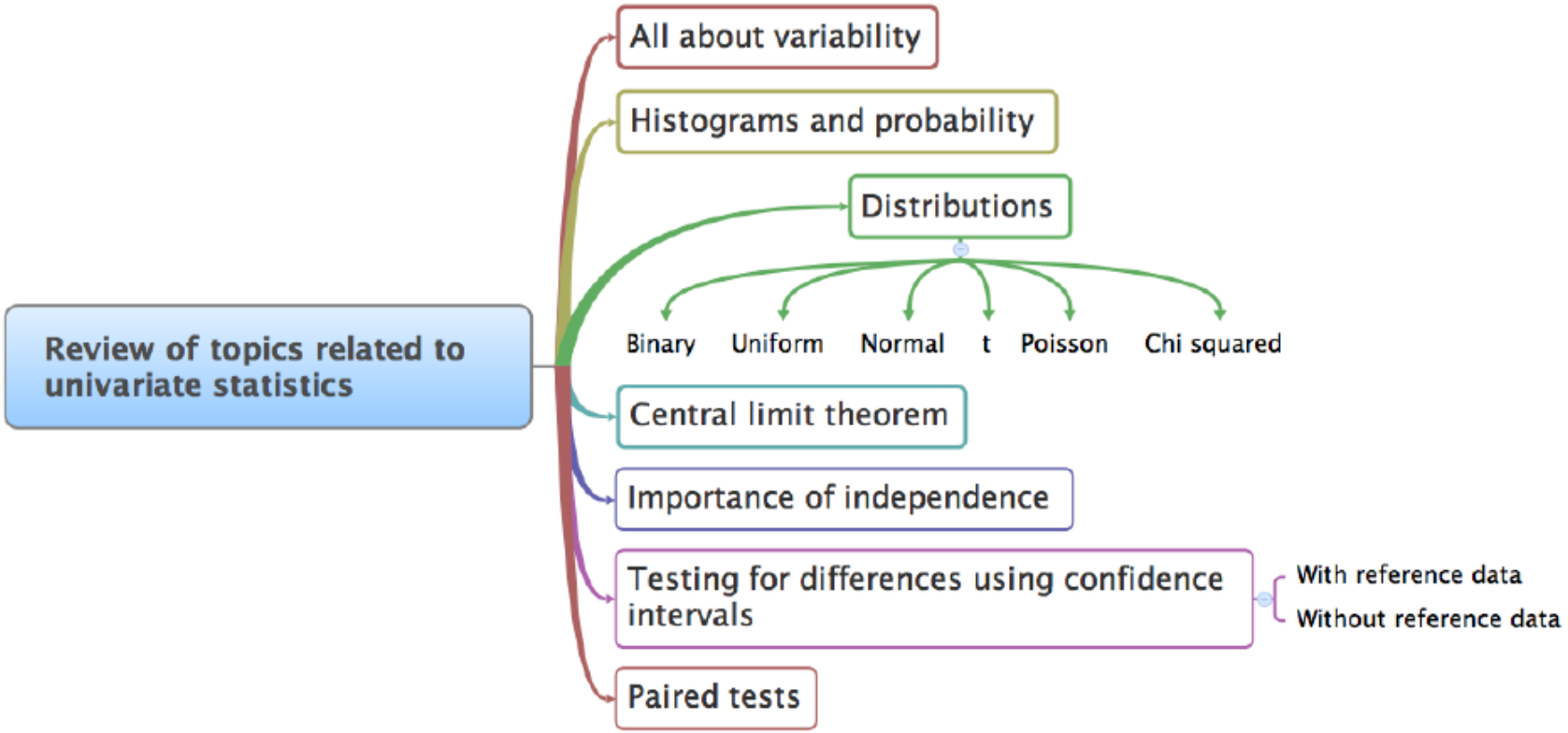
Jay Liu

Dept. Chemical Engineering

PKNU

Univariate statistics

- Overview

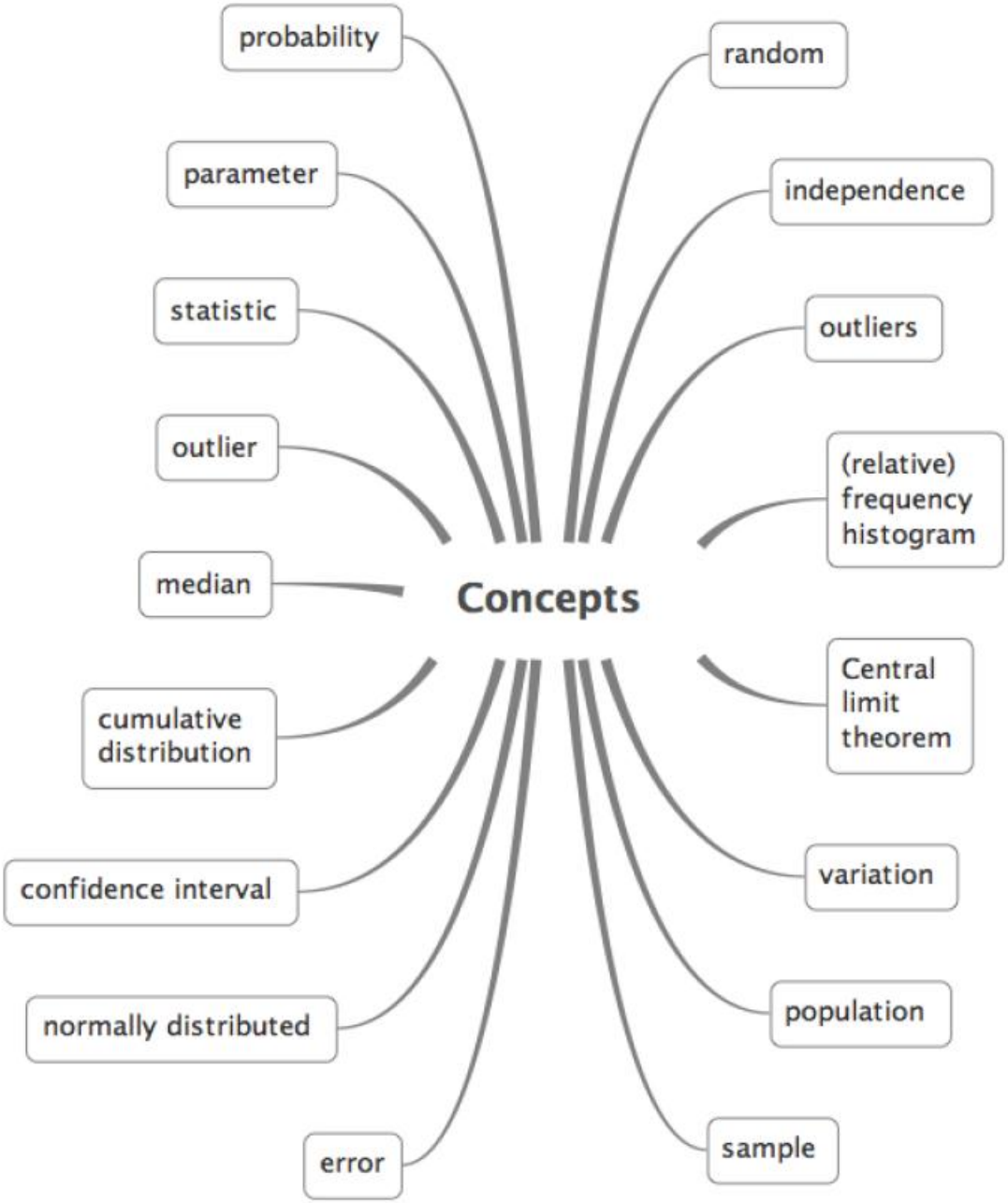


Reading: Textbook Ch. 3 ~ 5

Usage examples

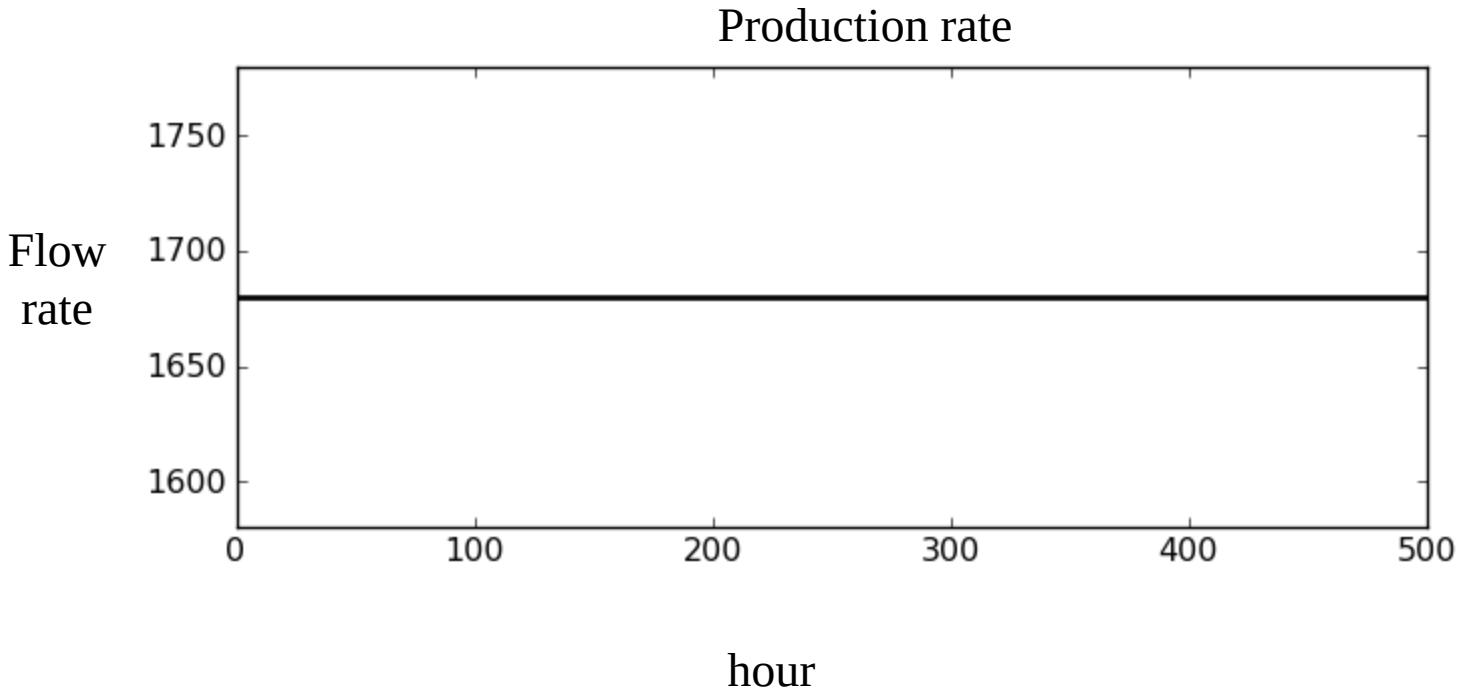
- Co-worker: Here are the yields from a batch system for the last 3 years (1256 data points)
  - yesterday's yield was less than 160g/L, something wrong?
- Yourself: I developed a new catalyst giving 95% conversion. Is this better than the previous catalyst?
- Manager: does reactor 1 have better final product purity than reactor 2?
- Potential customer: what is the 95% confidence interval for the density of your powder ingredient?

Concepts


Concepts
probability
parameter
statistic
outlier
median
cumulative distribution
confidence interval
normally distributed
error
random
independence
outliers
(relative) frequency histogram
Central limit theorem
variation
population
sample

Variability

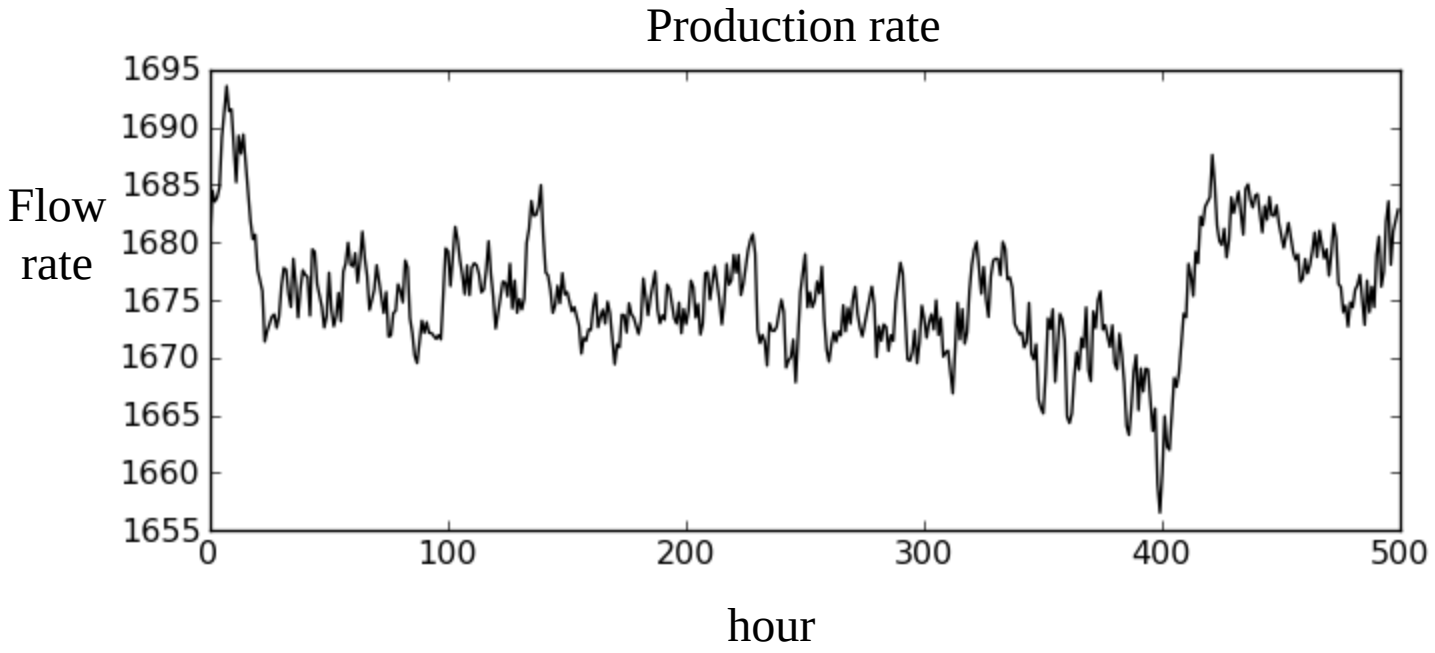
- Life will be pretty boring if



No plant engineer is needed (and this course would be unnecessary)

Variability

- But in reality, we have plenty of **variability** in our recorded data:



※ In most engineering work, the data is subject to variability/error. Therefore, many of the variables we will work with will be **random variables**. The statistics we evaluate will also be random variables.

Variability

This is because …

Variation in raw material properties

Production disturbances, Feedback control, Operating staff, Measurement and sampling variability, …

All this variability keep us process engineers **employed**, but it comes at a **price**.

The high cost of variability in your final product

Customers expect both uniformity and low cost when they buy your product. Variability defeats both objectives.

1. Customer totally unable to use your product:

   Ex1. A polymer with viscosity too high

   Ex2. Oil that causes pump failure

2. Your product leads to poor performance.

   Ex1. Customer must put in more energy (melting point too high)

   Ex2. Longer reaction times for off-spec catalyst

3. Your brand can be diminished. (ex. Toyota quality issues in 2010)

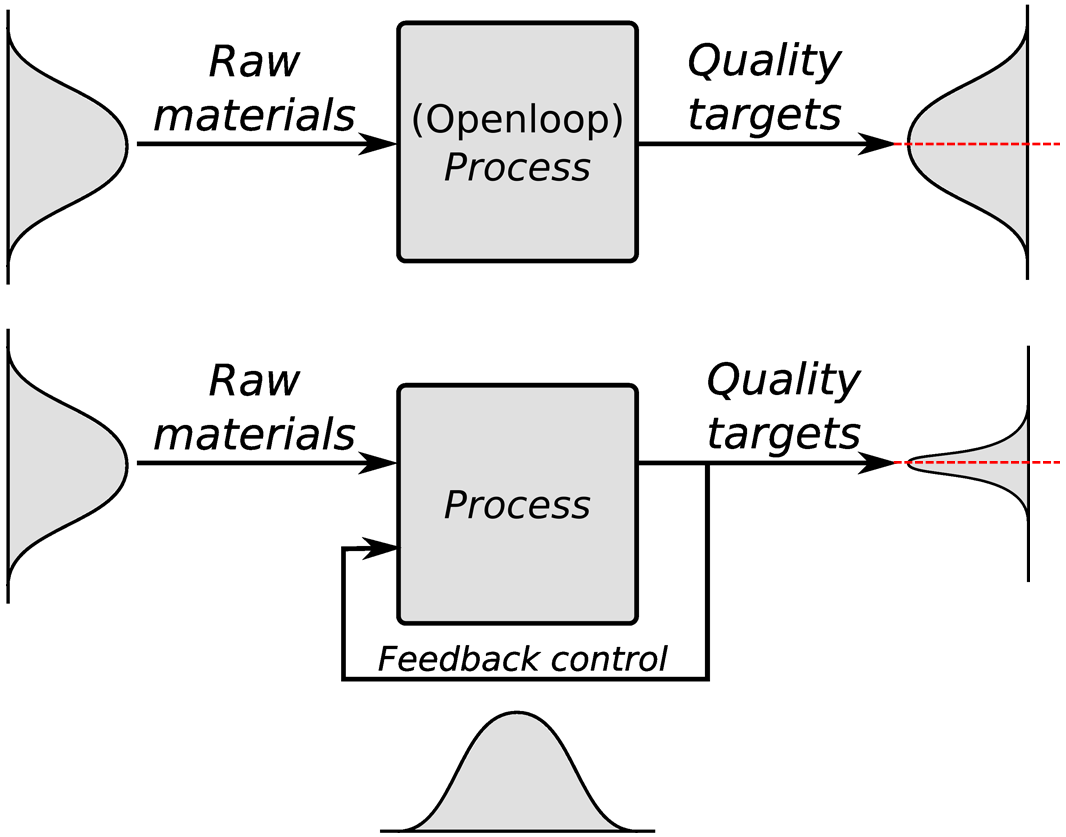
The high cost of variability in your final product

Variability also has these costs:

1. Inspection costs:
   - Too expensive and inefficient to test every product
   - Low variability means you don't need to inspect every product
2. Off-specification products cost you and customer money:
   - Reworked
   - Disposed
   - Sold at a loss

High variability in raw materials

- Example



This course is about variability

This section discusses

1. Visualizing, quantifying, and then comparing variability

Following sections

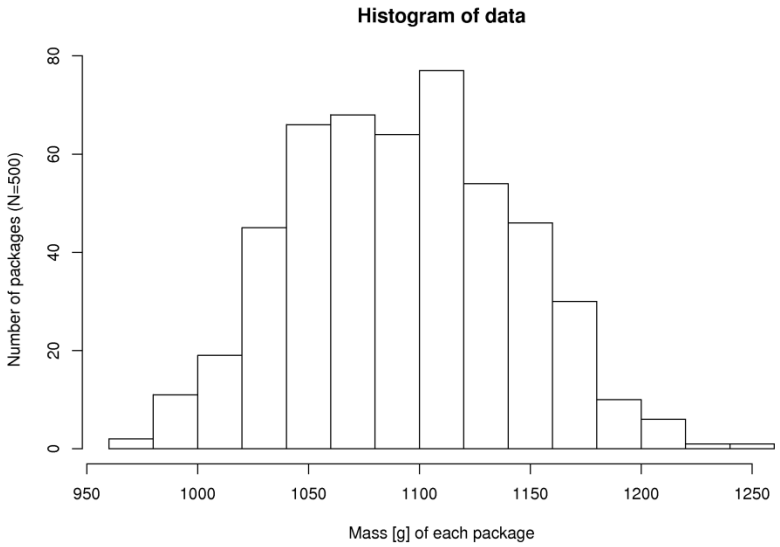
- SPC : construct monitoring charts to track variability
- Least Squares: variation in one variable affects another
- DOE : intentionally introduce variation to learn about process
- Multivariate: dealing with multiple variables, simultaneously extracting information

Histograms

Histogram: graphical summary of the variation in a measured variable

Shows number of samples that occur in a *category*: called a *frequency distribution*

Continuous variables: create category bins (usually equal-size)



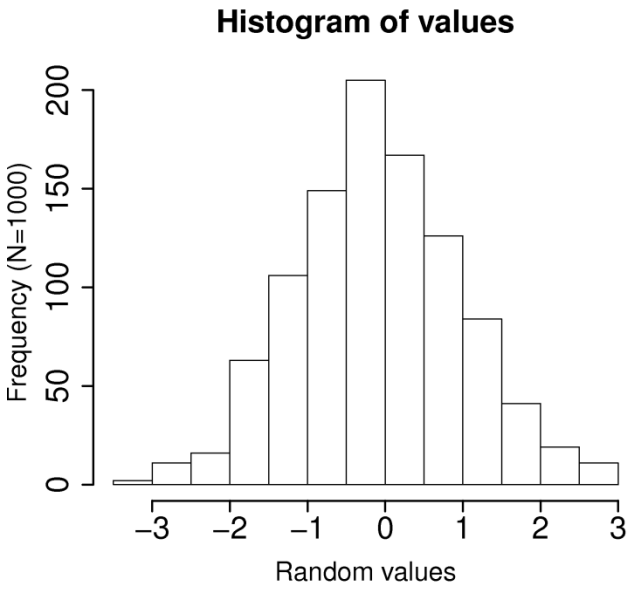
A rule of thumb: $\#\ bins \approx \sqrt{n}$

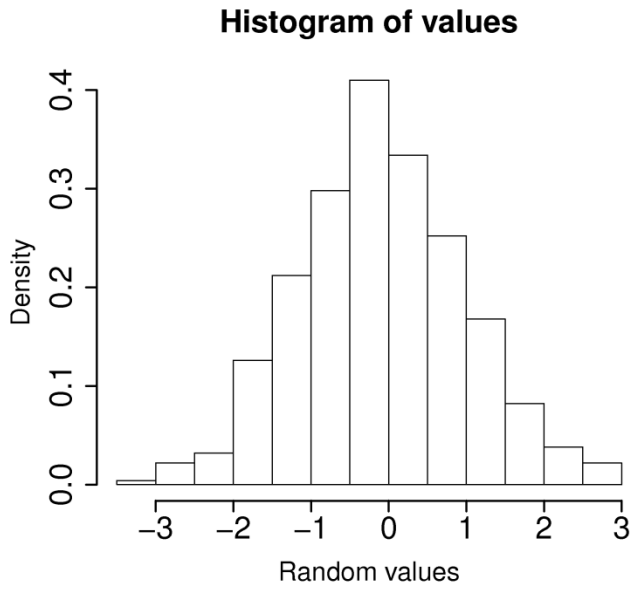
Histograms

A relative frequency is sometimes preferred:

- we do not need to report the total number of observations, N
- it can be compared to other distributions
- if N is large enough, then the relative frequency histogram starts to resemble the population's distribution
- the area under the histogram is equal to 1 (related to probability)

$$relative\ \ frequency = \frac{frequency}{n}$$





[FYI] Summary statistics

- Given a large table of values, it is often difficult to visually arrive at any meaningful information.
- Often we try to summarize the information in a set of data by condensing all of the information into a couple of statistics that will give us a feel for the behavior of the system from which the data were sampled.
- As a minimum, to characterize a dataset, we usually look for a measure of location and a measure of variability.
  - **Measures of location:** mean, median, mode
  - **Measures of variability:** variance, range, standard deviation

[FYI] Nomenclature

- Population
  - Large collection of potential measurements (not necessary to be infinite, a large N works well)
- Sample
  - Collection of observations that have actually occurred (n samples)
- Parameter
  - Value that describes the population's distribution
- Statistic
  - An *estimate* of one of the population's parameters

[FYI] Nomenclature

- Outliers
  - A point that is unusual, given the context of the surrounding data
    - 4024, 5152, 2314, 6360, 4915, 9552, 2415, 6402, 6261
    - 4, 61, 12, 64, 4024, 52, -8, 67, 104, 24
- Median (location)
  - Robust statistic: insensitive (robust) to outliers in the data
- Mode (location)
  - The most frequently occurring data (in a distribution)
- Range (variability)
  - the difference between the largest and the smallest values

Probability distributions

- Just a review; please read textbook for more details
- Focus on when to use the distribution
- And how the distribution looks

Probability distributions

- Probability distribution
  - a description of the set of possible values of X, along with the probability associated with each of the possible values
- Probability density function
  - a function that enables the calculation of probabilities involving a continuous random variable X. It is analogous to the probability mass function of a discrete random variable. It is usually denoted by f(x). The area under the curve f(x) between x1 and x2 defines the probability of obtaining a value of X in the interval [x1, x2].

Probability distributions

- A probability density function must satisfy the following properties:

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(u)du$$

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$$

$$P(X \geq a) = 1 - P(X \leq a)$$

$$P(X \leq -a) = P(X \geq a) = 1 - P(X \leq a)$$

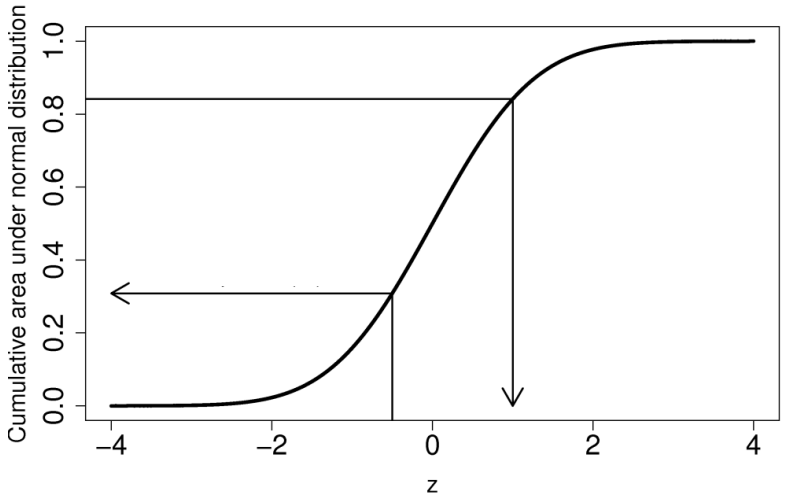
$$P(X \leq a) = P\left(z \leq \frac{a - \mu}{\sigma}\right)$$

Probability distributions

- Cumulative Density Function
  - Usually denoted by F(x) and defined by:

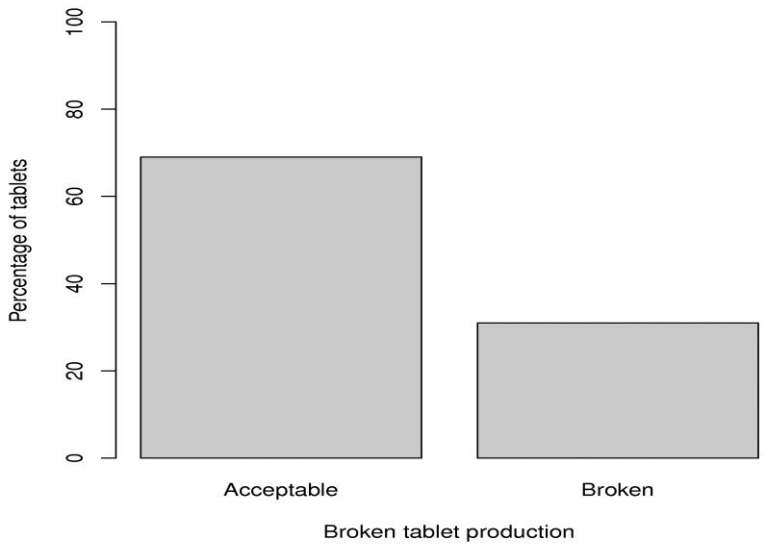
$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du, \quad f(x) = \frac{dF(x)}{dx}$$

- Cumulative distribution: area underneath the distribution function
- Inverse cumulative distribution: we know the area, but want to get back to the value along the x-axis.



Binary (Bernoulli distribution)

For binary events: event A and event B



- Pass/fail, or yes/no system
- p(pass) + p(fail) = 1

$$P(r) = \frac{N!}{r!(N-r)!}\pi^{r}(1-\pi)^{N-r}$$

- Example: What is the probability of obtaining exactly 3 heads if a fair coin is flipped 6 times?

$$P(3) = \frac{6!}{3!\,(6-3)!}\,(0.5)^{3}\,(1-0.5)^{6-3}$$

$$= \frac{6 \times 5 \times 4 \times 3 \times 2}{(3 \times 2)(3 \times 2)}\,(0.125)(0.125) = 0.3125.$$

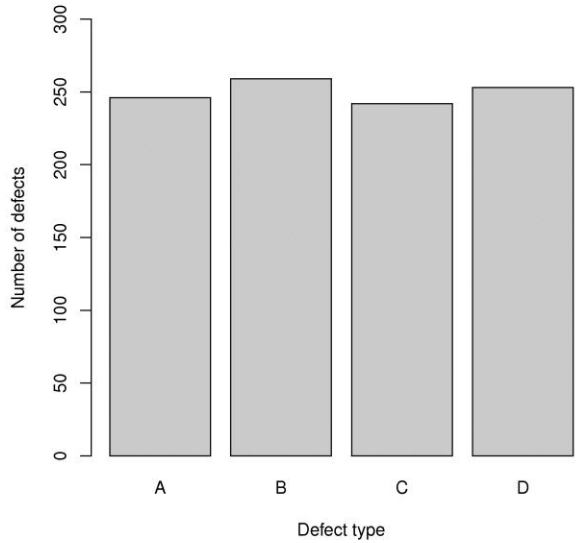
Uniform distribution

- Each outcome is equally as likely to occur as all the others.

  The classic example is dice: each face is equally as likely.

  (This sort of phenomena is not often found in practice)

- Probability distribution for an event with 4 possible outcomes:



Poisson distribution

- Distribution of counts in the cases such as
  - Particles contamination in semiconductor manufacturing
  - Flaws in rolls of textiles
  - Atomic particles emitted from a specimen

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 1, 2, 3, \ldots$$ λ: average flaws/particles

- Example
  - In a copper wire manufacturing, suppose that the number of flaws follows a Poisson distribution with a mean of 2.3 flaws per millimeter. Determine the probability of exactly 2 flaws in 1 millimeter of wire.

$$P(X = 2) = \frac{e^{-2.3}2.3^{2}}{2!} = 0.265$$

Normal distribution

normal PDF
(probability density function)

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

$x \sim N(\mu, \sigma^2)$

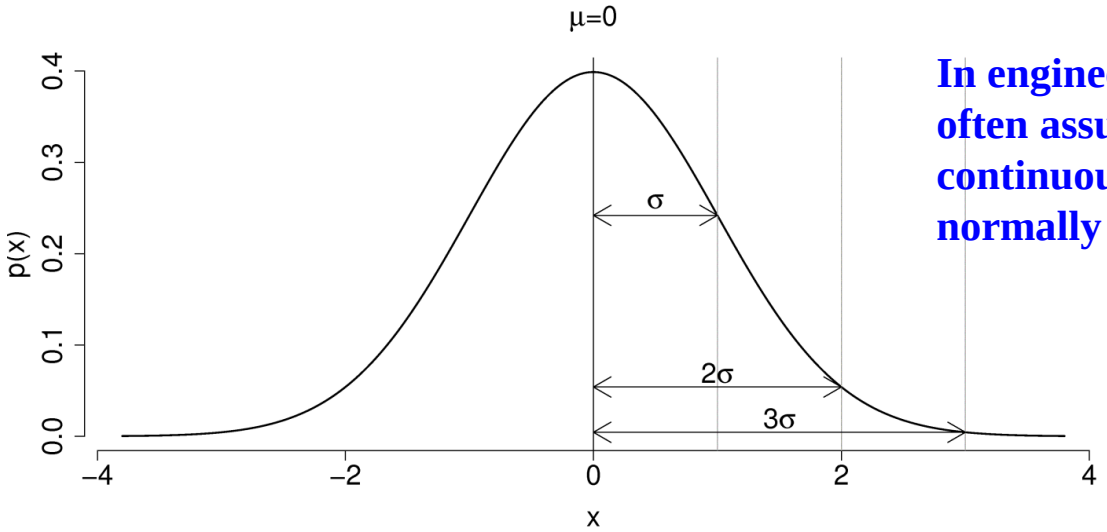
x: variable of interest

f(x) probability of obtaining that x

μ population mean for variable x

σ population standard deviation (positive)

Distribution is symmetric about μ



In engineering applications we often assume that measured continuous random variables are normally distributed. Why?

The Standard Normal Distribution

- The standard normal distribution refers to the normal distribution with mean zero and variance one.
- The standard normal distribution is important in that we can use tabulated values of the cumulative standard normal distribution for any normally distributed random variable by first standardizing it. We **standardize** a random variable X that is $N(\mu, \sigma^2)$ using:

$$Z = \frac{X - \mu}{\sigma}$$

- Units of *z* if *x* were measured in kg, for example?
- Standardization allows us to straightforwardly compare 2 variables that have different means and spreads

Self exercise

- Assume x = biological activity of a drug, $x \sim N(26.9, 9.3)$. Probability of $x \leq 30.0$?
- Assume x = yield from batch process, $x \sim N(85\ \text{g/L}, 16\ \text{g/L})$. Proportion of yield values between 77 and 93 g/L ?

Recommendation: make sure you can read a statistical table

Tables of the Normal Distribution

Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Exercise

- The temperature of a heated flotation cell under standard operatingconditions is believed to fluctuate as a normal p.d.f. with a mean value of 40 degrees Celsius and a standard deviation of 5 degrees Celsius. What is the probability that the next measured temperature will lie between 37 degrees Celsius and 43.5 degrees Celsius?

(solution)

Interpretation: Temperature, $T \sim N(40,5^2)$ → $P(37 \le x \le 43.5)$?

In minitab, “calc”→“probability distributions”

[FYI] Nomenclature

- Mean
  - Measure of location (position)
  - Population mean $\mu = E(X) = \frac{1}{N}\sum x \quad or \quad = \int_{-\infty}^{\infty} xf(x)\,dx \quad or \quad = \sum_{x} xf(x)$
  - Sample mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Variance
  - Measure of spread, or variability
  - Population variance $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad V(X) = \sigma_x^2 = E(X-\mu)^2 = \sum_{x}(x-\mu)^2 f(x)$

    $or \quad = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx$
  - Sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

[FYI] Nomenclature

- Expected value

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx$$

  - The mean and variance are special cases of this general definition
  - Properties of Expectations and Variances:

$$E[cX] = cE[X]$$
$$E[X + Y] = E[X] + E[Y]$$
$$V(cX) = c^2 V(x)$$
$$V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$$

[FYI] Nomenclature

- Covariance
  - Covariance is a measure of the linear association between random variables.
  - Population covariance:

$$\sigma_{XY}^2 = E\{(X-\mu_X)(Y-\mu_Y)\} = E(XY) - \mu_X \mu_Y$$

  - Sample covariance:

$$\hat{\sigma}_{XY}^2 = s_{XY}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{n-1}$$

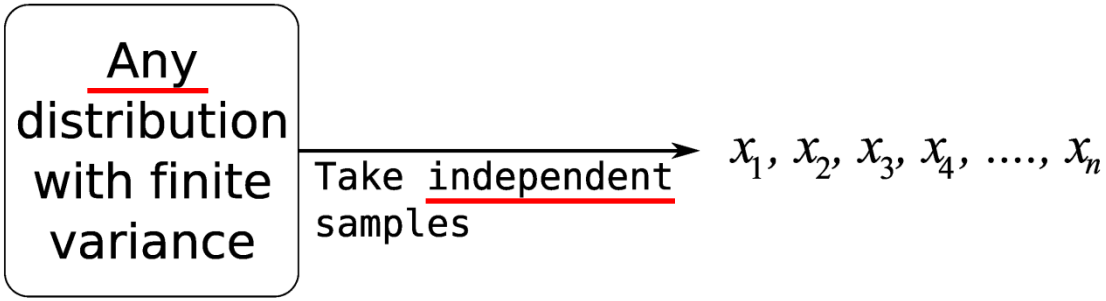
[FYI] Nomenclature

- Correlation
  - a scaled version of covariance. The scaling is done so that the range of ρ is [-1, 1] .
  - Population correlation: $\rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y}$
  - Sample correlation:

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})\right)}}$$

Central limit theorem

- Central limit theorem
  - The average of a sequence of values from *any distribution* will approach the normal distribution, provided the original distribution has finite variance.



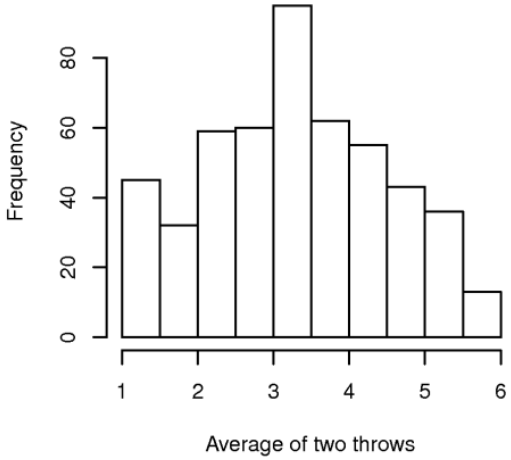
  - If $x_1, x_2, x_3, \ldots, x_n$ are taken from a population with mean μ and finite variance σ^2. Then as $n \rightarrow \infty$,

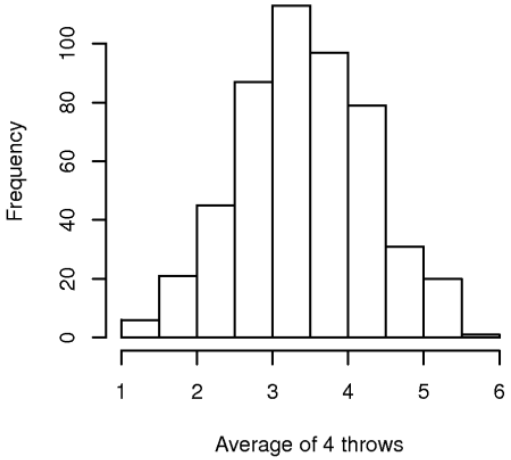
    sample mean $\bar{x}$ approaches to normal distribution.

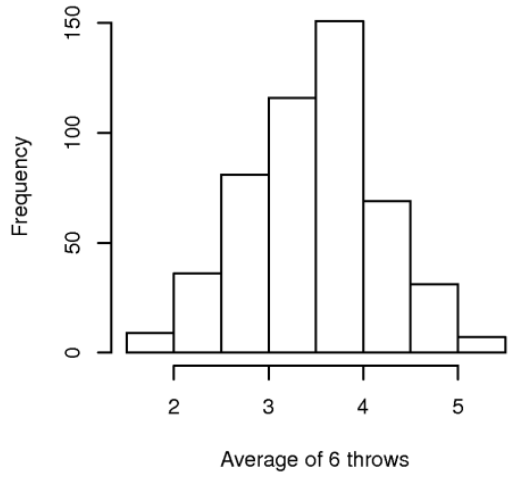
    $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ approaches to standard normal distribution.

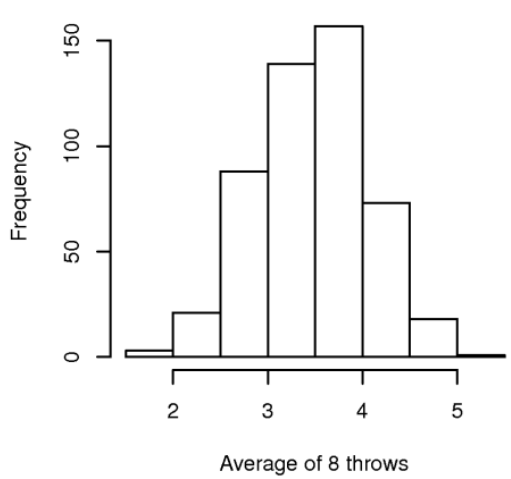
Central limit theorem

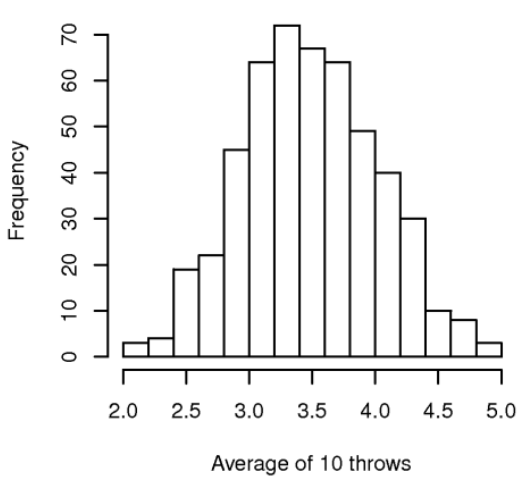
- Example: throwing dice











Statistical independence

The assumption of independence is widely used. It is a condition for the central limit theorem.

- Independence
  - The samples are *randomly* taken from a population. If two samples are independent, there is *no possible relationship between them*.
    - A questionnaire is given to students. Are the marks independent if students discuss the questionnaire prior to handing it in?
  - Often people say that random variables x and y are independent if correlation is zero. Is this enough?

$$\rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y} = 0$$

[FYI]Continuous vs. discrete variables

- Probability density function
  - For continuous random variables

(1) $f(x) \geq 0$

(2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

(3) $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx =$ area under $f(x)$ from a to b

for any a and b

- Probability mass function
  - For discrete random variables

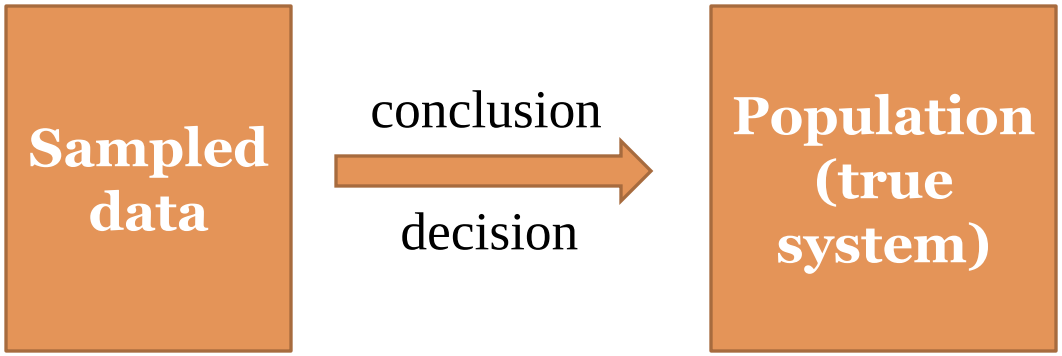
(1) $f(x_i) \geq 0$

(2) $\sum_{i=1}^{n} f(x_i) = 1$

(3) $f(x_i) = P(X = x_i)$

Statistical inference

- In engineering applications, we are more often in the position that we have a sample of data, and based on this data we want to make some statements about our belief in the population parameters (i.e. the properties of the "true" system). This is the realm of statistical inference.



ex. Comparing conversions of two different catalysts.

Sampling distribution

- In most engineering work, the data is subject to error (such as sensor noise). Therefore, many of the variables we will work with will be *random variables*. The statistics we evaluate will have a probability distribution associated with them.
- The probability distribution of a statistic (a random variable whose value is based on a sample of data) is called a sampling distribution.
  - Examples of statistics: sample mean/variance/correlation/…

Sampling distribution

- This implies that there is a certain amount of uncertainty in the value we obtain for a statistic.
- The idea is that if we collected another sample in the same way and evaluated the statistic based on the new data we would likely obtain a different value. If we continued to collect new data and evaluate the statistic we would be able to construct a histogram of all of the value of the statistic. This histogram would be an estimator of the sampling distribution of the statistic.
- The sampling distribution provides information about the amount of variation in the statistic and the nature of the variation.

Confidence interval

- Confidence intervals are an important way to quantify and state how uncertain is an estimate calculated from samples.
- Confidence intervals convey two types of information:
  - A summary of the behavior of the data in that sample
  - Indications of the characteristics of the population from which the sample was obtained.
- A confidence interval is a range calculated based on the data in a sample and assumptions about the underlying p.d.f. This interval has a specified probability of containing the true value of the parameter being studied.
- A confidence interval for a parameter is a range of plausible values for the parameter in the light of the available data.

Confidence interval - examples

- We may measure 20 temperatures in a heated vessel and calculate the mean to be 620 degrees Celsius. The mean of 620 is a (point) estimate of the "true" temperature of the vessel. We then calculate a 95% confidence interval for the mean to be C.I.=[600, 640] ^{o}C
- This says that we are 95 % confident that the true temperature of the vessel is between 600 and 640 degrees Celsius (assuming our assumptions are valid). This range also gives an indication of the amount of uncertainty in our estimate of the temperature in the vessel.
- Yet another interpretation is that if we continued to sample 20 temperatures, compute the means and confidence intervals, 95% of the confidence intervals would contain the true value of the temperature.