공정모형 및 해석

Jay Liu

Dept. Chemical Engineering

PKNU

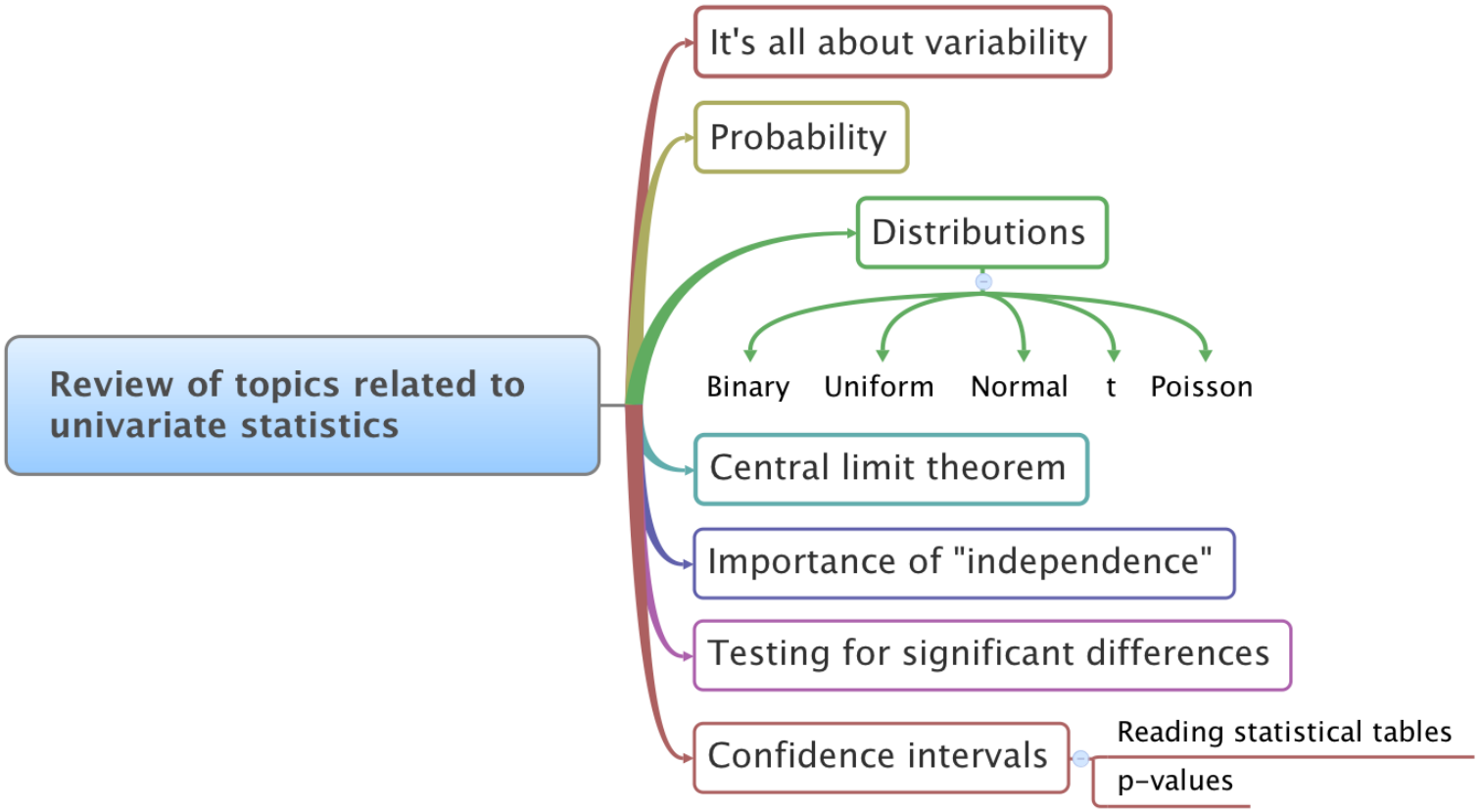
Exercise – use of Pareto chart

- The company you work for manufactures metal bookcases. During final inspection, a certain number of bookcases are rejected due to scratches, peel, smudge, or other. You count the number of times each defect occurred, then you enter the name of the defect each time it occurs into a worksheet column called Damage.

No.	Defect	No.	Defect	No.	Defect	No.	Defect
1	Scratch	11	Peel	21	Peel	31	Other
2	Scratch	12	Peel	22	Peel	32	Scratch
3	Peel	13	Scratch	23	Other	33	Scratch
4	Peel	14	Scratch	24	Other	34	Peel
5	Smudge	15	Peel	25	Scratch	35	Peel
6	Scratch	16	Scratch	26	Scratch	36	Peel
7	Other	17	Smudge	27	Peel	37	Smudge
8	Other	18	Scratch	28	Scratch	38	Smudge
9	Peel	19	Peel	29	Smudge	39	Smudge
10	Peel	20	Peel	30	Scratch	40	Other

Univariate statistics

- Overview



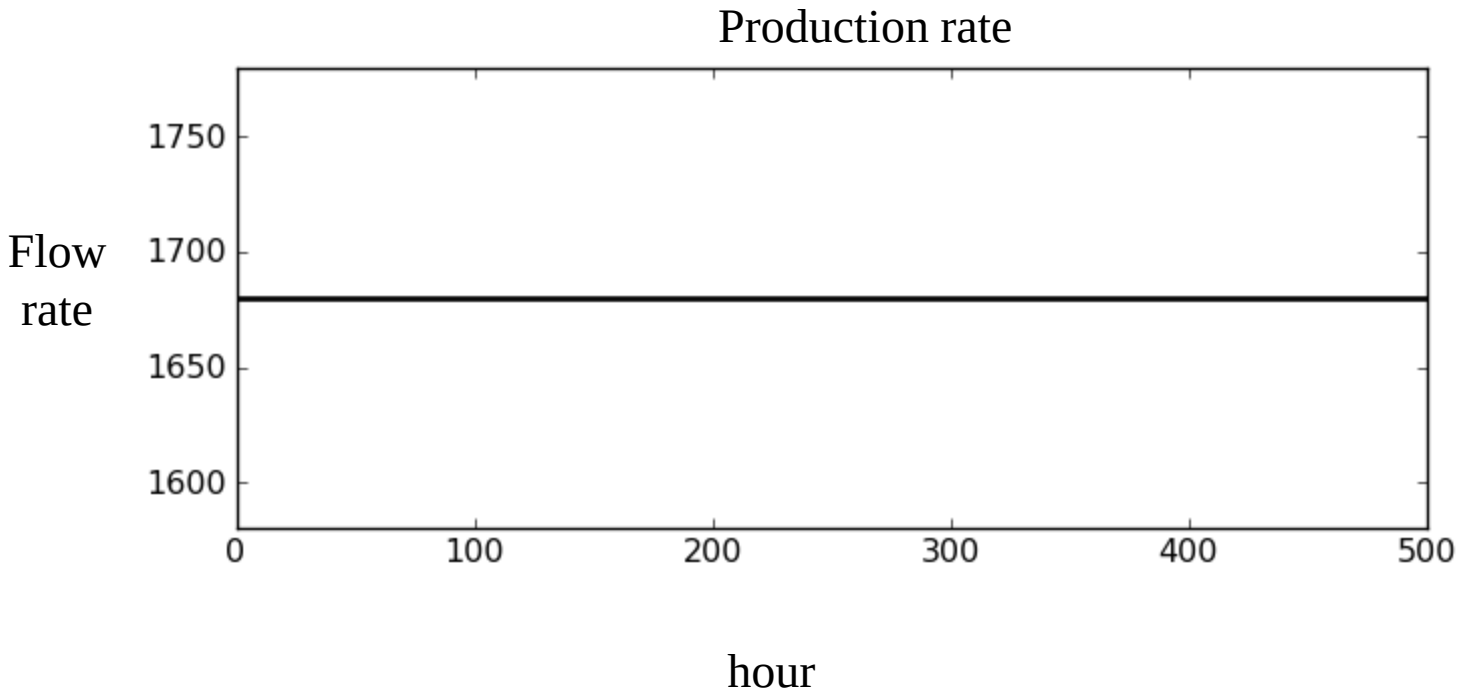
Reading: Minitab manual p77-97 (or use Minitab "help")

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 50%, is this rare occasion?
- 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?

Variability

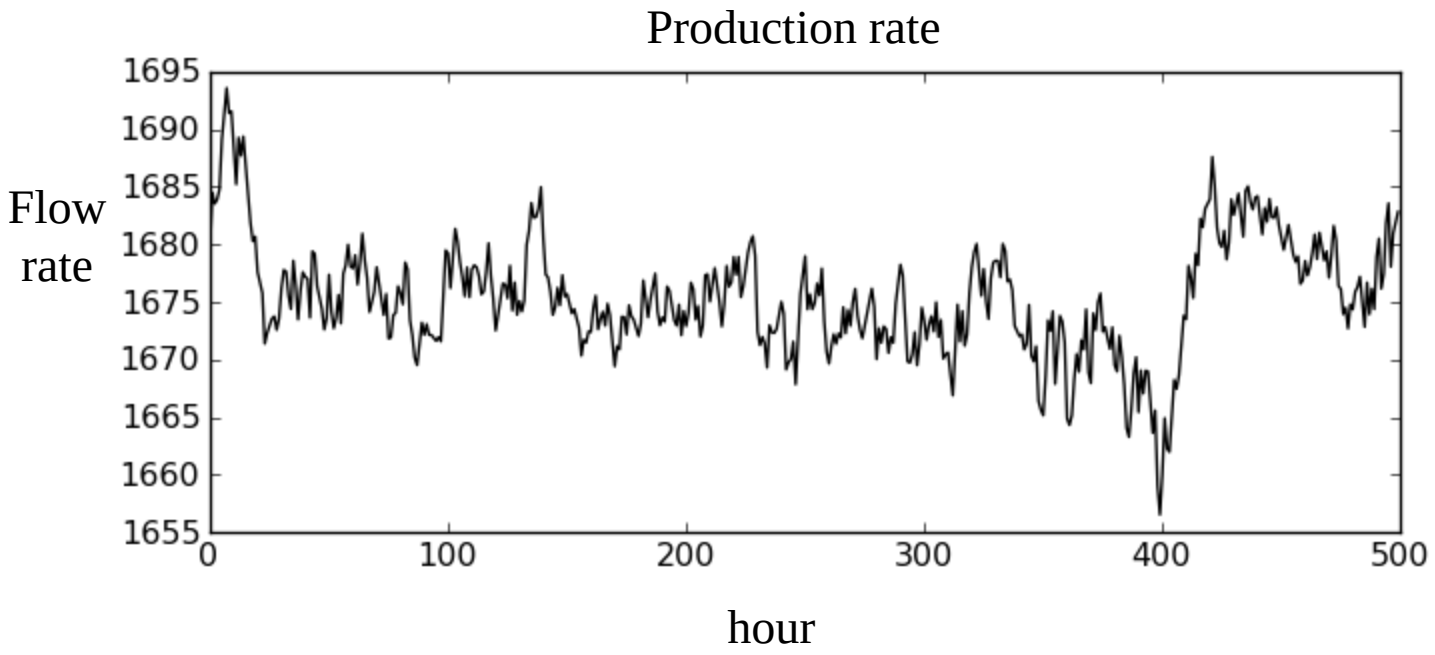
- Life will be pretty boring if



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

Variability

- We have plenty of variability in our recorded data:



This is because … Variation in raw material properties

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

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

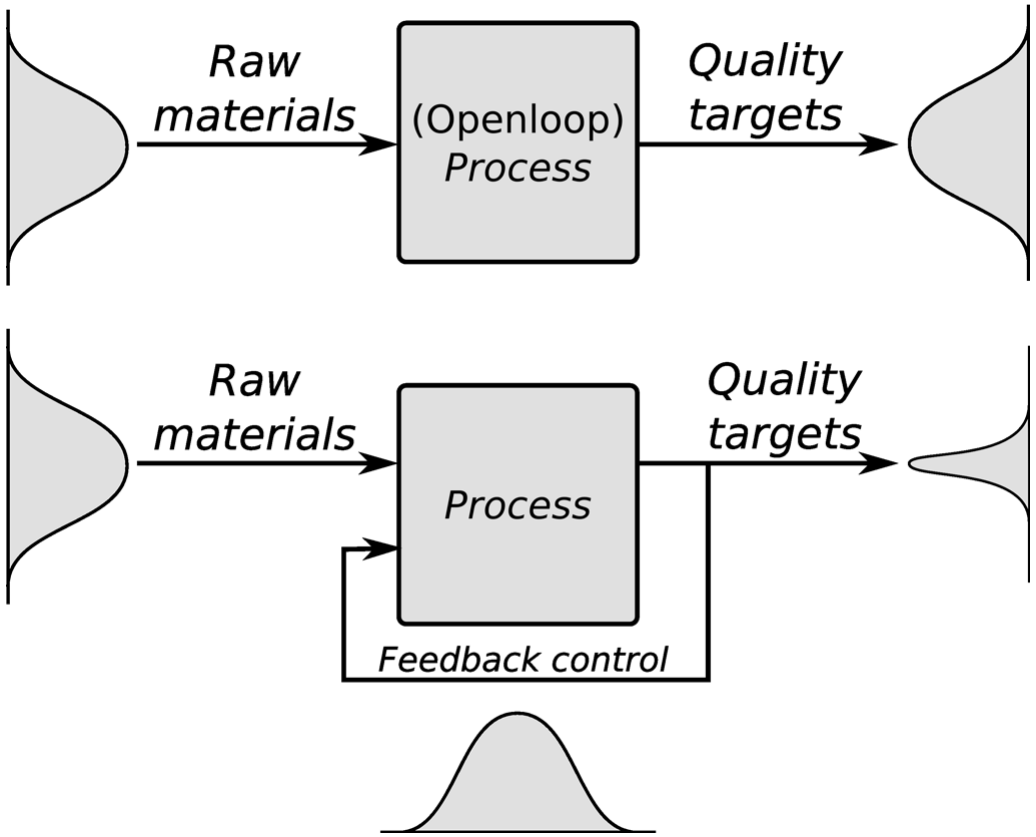
The high cost of variability in your final product

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

The high cost of variability in your raw materials

- Example



This course is about variability

This section discusses

1. Visualizing the variability
2. Quantifying variability, then comparing variability

Following sections

- Least Squares: variation in one variable affects another
- DOE : intentionally introduce variation to learn about process

※ SPC : construct monitoring charts to track variability

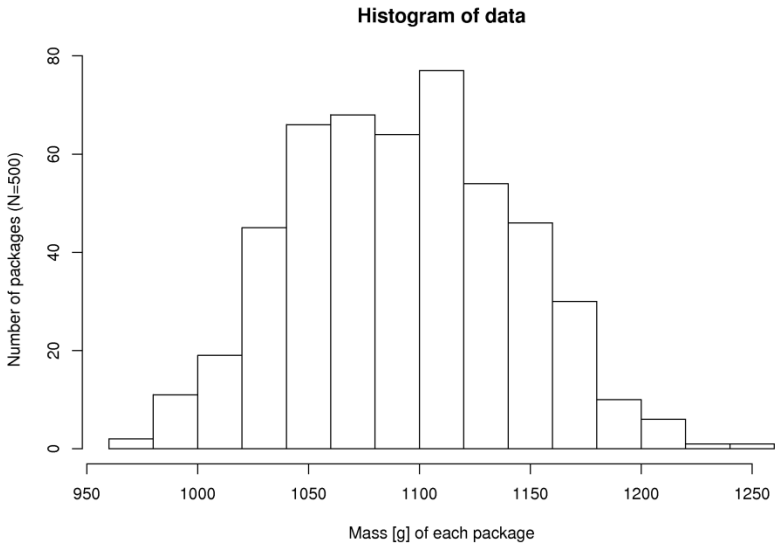
※ Multivariate: dealing with multiple variables, simultaneously extracting information

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)



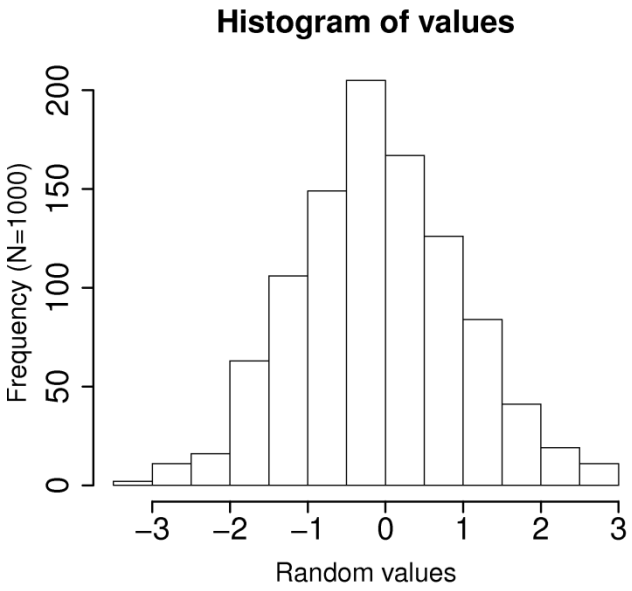
A rule of thumb: $\#\ bins \approx \sqrt{n}$

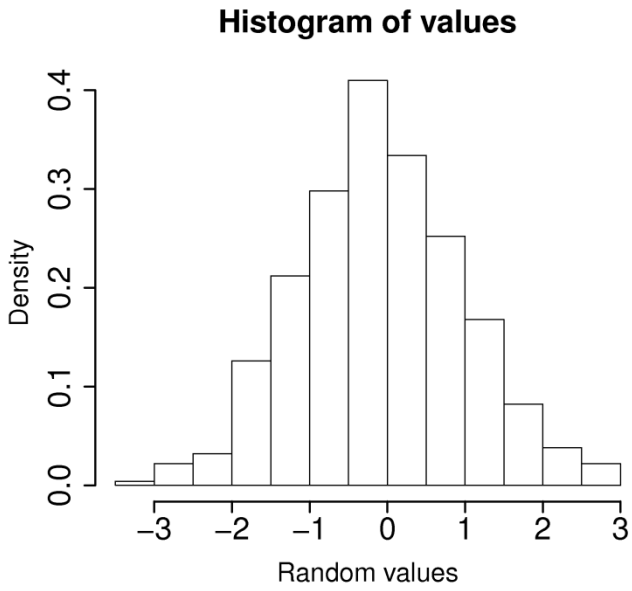
Histograms

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, and 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

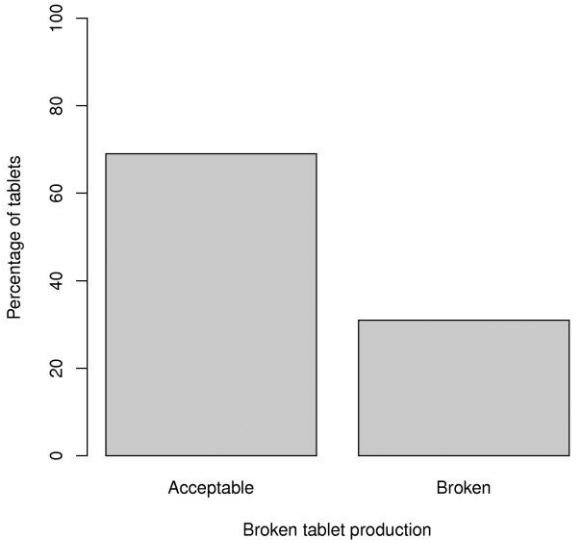
Distributions

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

Binary (Bernoulli distribution)

- 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: Each sample of water has a 10% chance of containing a particular organic pollutant. Find the probability that in the next 18 samples, exactly 2 contain the pollutant.

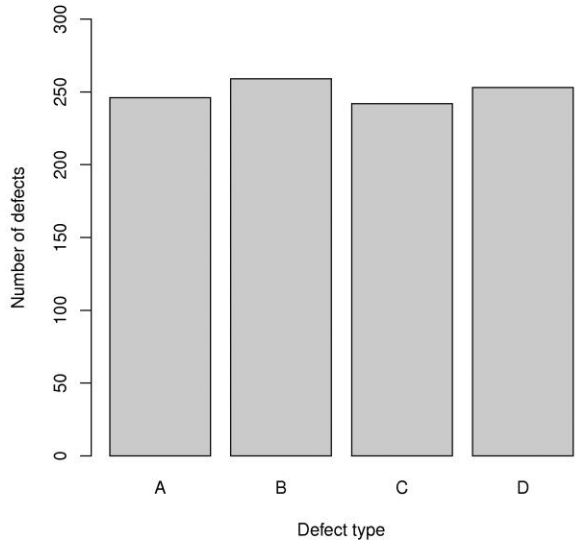
Uniform distribution

- 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:



Normal distribution

normal PDF
(probability density function)

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

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

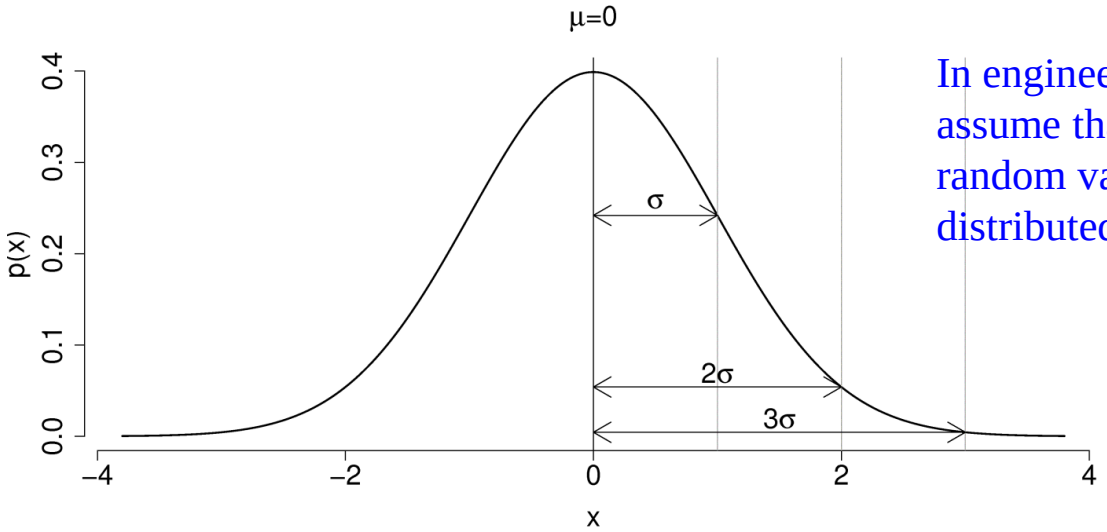
x: variable of interest

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.

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

The Standard Normal Distribution

- Any probability density function $f(x)$ and probability P calculated from it satisfy followings

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

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

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

$$P(X \ge a) = 1 - P(X \le a)$$

$$P(X \le -a) = P(X \ge a) = 1 - P(X \le a)$$

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

The Standard Normal Distribution

- Any probability density function $f(x)$ and probability P calculated from it satisfy followings

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

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

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

$$P(X \ge a) = 1 - P(X \le a)$$

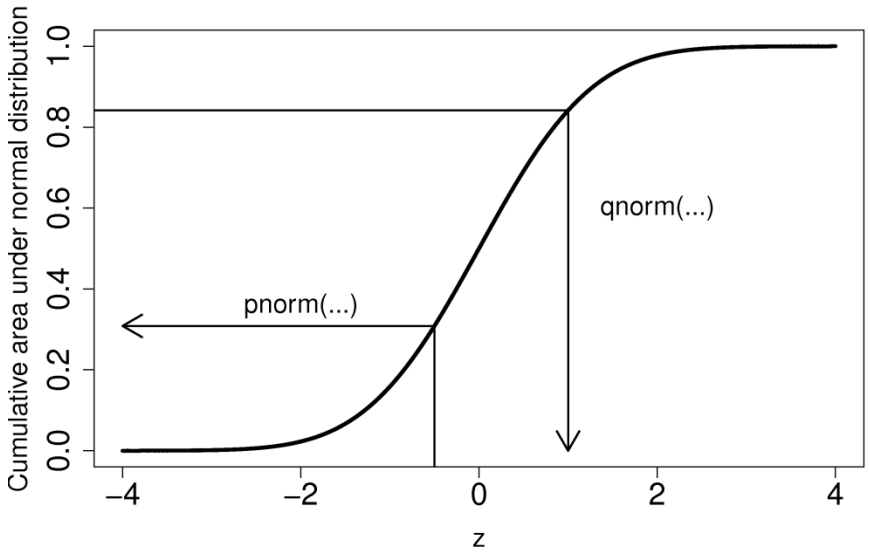
$$P(X \le -a) = P(X \ge a) = 1 - P(X \le a)$$

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

Cumulative Density Function

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

- 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.



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, $T \sim N(40,5^2) \rightarrow P(37 \leq x \leq 43.5)$?

In minitab, “calc”→“probability distributions”

[FYI] Nomenclature

- Mean
  - Measure of location (position)
  - Population mean $\mu = E(X) = \frac{1}{N}\sum x \quad or \quad = \int_{-\infty}^{\infty} x f(x)\,dx \quad or \quad = \sum_{x} x f(x)$
  - 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:

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

[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:

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

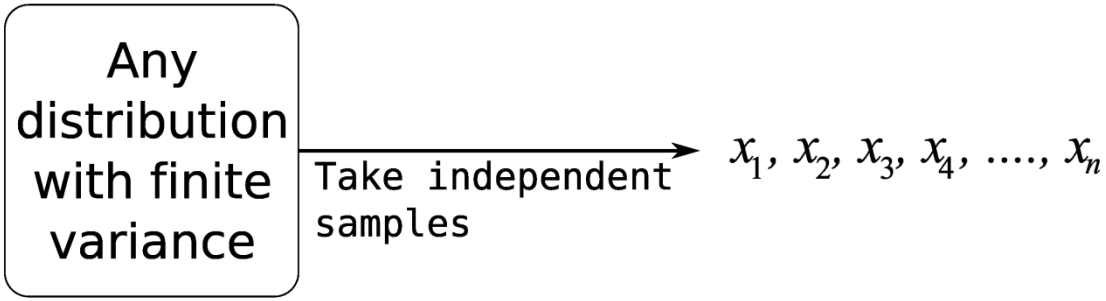
[FYI] Nomenclature

- 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