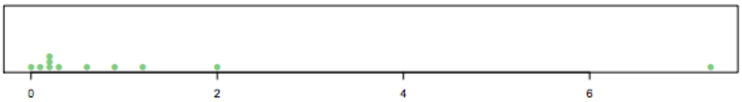
# Statistics Cheat Sheet

## Descriptive Statistics

- Dotplots: Used for reasonably small data sets
  - Represent each dot above corresponding location on measurement scale
  - Stack dots vertically for one or more occurrence

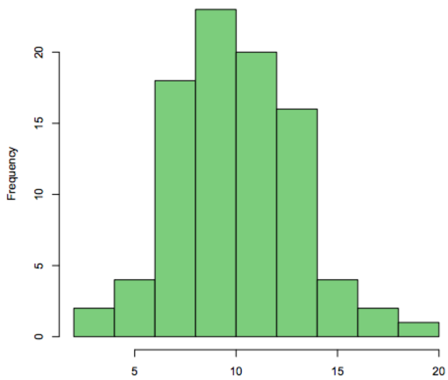



  - Extremely different points = outlier
- Stem and leaf plot: Separate observation into two parts
  - Stem: all but last digit – vertical column in increasing order (left)
  - Leaf: final digit – right of the stem, in increasing order our of the stem

```
0 | 4
1 | 1345678889
2 | 1223456666777889999
3 | 0112233344555666677777888899999
4 | 111222223344445566666677788888999
5 | 00111222223345566666777788899
6 | 01111244455666778
```

  - Identifies typical values, shows extent of spread of typical values, presence of gaps in data, extent of symmetry, number and location of peaks, outlying values
  - Can round and truncate to avoid irrelevant information
  - Can split each stem to give detail in distribution
  - Can make back to back stemplots to compare distributions
  - Matlab command: round(sort(data)*10)/10

- Histograms
  - Determine frequency observations in a class – draw rectangles around corresponding frequencies

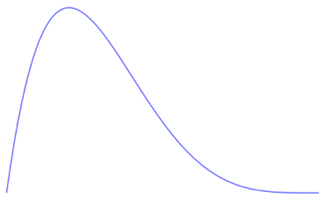



  - Matlab command: histogram(data)
    skewness(data)

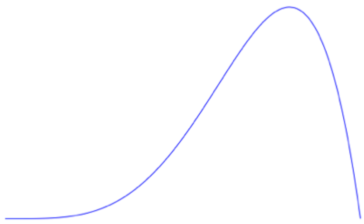
$\Rightarrow$ For density histogram: histogram(Inflow,5,'Normalization','pdf')

- Provides visual representation of shape of distribution

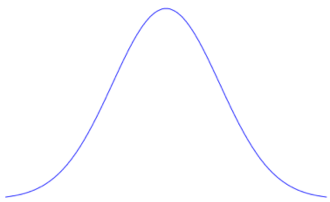
Example: Skewed to the right



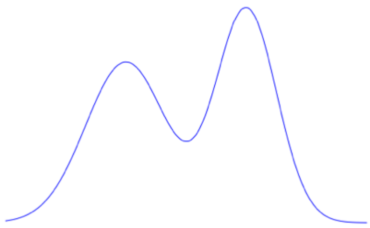
Example: Skewed to the left



Example: Symmetric shape



Example: Bimodal



- Density histogram: Rectangular heights are densities of each histograms (not frequencies) – relative frequency of class is the proportion of observations in that class. (frequency of class divided by observations)

  $\Longrightarrow$ Density = relative frequency of class/class width

  Total area of rectangles = 1 -> sum of all relative frequencies = 1

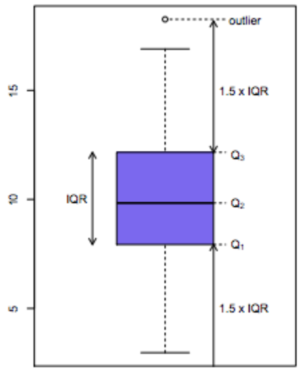
➢ Mean: Most frequently used measure of centre – arithmetic average of n observations
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $Matlab: mean(x)$

➢ Median: Divides data into two equal parts (half below and half above)
- If n is **odd**: $m = x_{\frac{n+1}{2}}$
- If n is **even**: $\frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$
- Matlab: median(x)

➢ Quantiles and percentiles: divide samples into more than two parts
- First/lower quartile: Median of **lower half** of data
- Third/upper quartile: Median of **upper half** of data
- Five number summary: $\{x_{(1)}, q_1, m, q_3, x_{(n)}\}$
- Matlab commands: [x(1), quantile(x,0.25),median(x),quantile(x,0.75),x(n)];

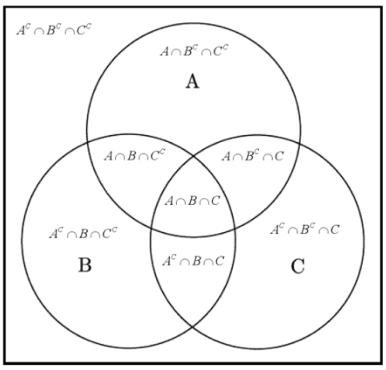
➢ Measures of variability: Observations deviations in the mean for given data
- Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Standard deviation = $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Matlab: var(x), std(x)
- Interquartile range: measure of variability related to sample mean and quartiles (difference between upper and lower quartiles)

- Iqr = $q_3$-$q_1$ – describes amount of variation in middle half of observations
  - Can detect outliers: observations that are **1.5*iqr** from closest quartile (extreme outlier is 3*iqr from closest quartile)
- Boxplot: Graphical representation of five number summary
- Central box spans quartiles
- Line in box Is median
- Lines outside box extend to data points which are not outliers (within 1.5*iqr)




## Elements of probability

- Events: Subset of sample space of a random experiment
  - Union: $E_1 \cup E_2$ – either $E_1$ or $E_2$ occurs: $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ for **mutually exclusive** events only
  - Intersection $E_1 \cap E_2$ – Both $E_1$ and $E_2$ occur
  - Compliment: $E^c$ – Event does not occur
  - $E_1 \subseteq E_2 \Rightarrow E_1$ implies $E_2$
  - $E_1 \cap E_2 = \varphi \Rightarrow$ mutually exclusive events (they cannot occur together)
  - De Morgan's laws: $(E_1 \cup E_2)$ c $= E^c{}_1 \cap E^c{}_2$
    $(E_1 \cap E_2)^c = E^c{}_1 \cup E^c{}_2$




  - $P(E^c) = 1 - P(E)$
  - $P(\varphi) = 0$
  - $E_1 \subseteq E_2 \Rightarrow P(E_1) \leq P(E_2)$ (increasing measure)
  - $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$ for **independent** events
- Assigning probabilities: probability = proportion of occurrences of an event
  - Multiplication rule: for operation with sequence of steps, total number of ways of completing operation: $n_1 * n_2 * \ldots * n_k$
  - Permutations: Order sequence of elements in a set:
    $P_n = n*(n-1)*(n-2)*\ldots*2*1=n!$

- Combinations: subset of elements selected from a larger set
  $\binom{n}{r} = C_r^n = \frac{n!}{r!\,(n-r)!}$

➢ Conditional probability: Probability of A, given that B has occurred

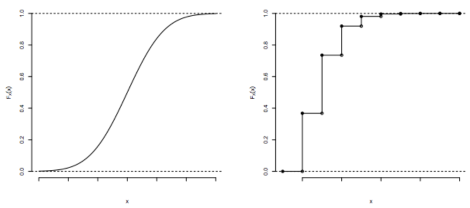
- $P(A|B) = \frac{P(A\cap B)}{P(B)}$ if $P(B)>0$
- Bayes first rule: if $P(A)>0$ and $P(B)>0$:
  $P(B|A) = P(A|B) * P(B)/P(A)$

➢ Independence
- Two events A and B are **independent** if and only if $P(A \cap B) = P(A)*P(B)$
  $P(A|B) = P(A)$ and $P(B|A) = P(B)$ (the probability of the occurrence of one event is unaffected by the other)

➢ Partition: Sequence of events $E_1, E_2, \ldots, E_n$ is called a partition of S
- Law of total probability: $P(A) = \sum_{i=1}^{n} P(A|E_i) * P(E_i)$
- Bayes second rule: for a given partition of S:
  $P(E_i|A) = \frac{P(A|E_i)P(E_i)}{\sum_{j=1}^{n} P(A|E_j)P(E_j)} \Rightarrow P(E|A) = \frac{P(A|E)P(E)}{P(A|E)P(E)+P(A|E^c)(1-P(E))}$

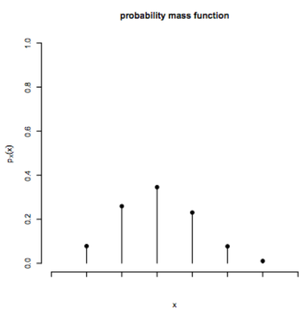
**Random Variables**

➢ Cumulative distribution function: cdf of random variable X defined for any real number. By $F(X) = P(X<=x)$
- For any $a<=b$, $P(a<X<=b) = F(b)-F(a)$
- F is a non-decreasing function



Continuous distribution → continuous r.v.        Discrete distribution → discrete r.v.

➢ Discrete random variables: Assumes a finite number of variables
- $S_X = \{x_1, x_2, \ldots\}$
- Probability mass function of discrete random variable X defined for any real number x, by $p(x) = P(X=x)$, sum of $p(x) = 1$




➢ Continuous random variables: Needs to exist a nonnegative function f(x) for all real x

- $P(X \in B) = \int_B f(x)dx$
- Consequence if P(X=x) = 0 for any x

➢ Probability density function: $F(X) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$
f(x) = dF(x)/dx = F'(x)

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

➢ Expectation: expectation or mean of a random variable E(x) or μ
Discrete r.v: $E(X) = \sum x * p(x)$
Continuous r.v: $E(X) = \int_{Sx} xf(x)dx$

- Function of a random variable:
discrete r.v: $E(g(X)) = \sum g(x) * p(x)$
Continuous r.v $E(g(X)) = \int_{Sx} g(x)f(x)dx$
- Linear transformation: E(aX+b) = aE(X)+b

➢ Variance: Var(x) or $\sigma^2 \Rightarrow$ Var(x) = $E((x-\mu)^2)$
Discrete r.v: $Var(X) = \sum (x-\mu)^2 * p(x)$
Continuous r.v: $Var(X) = \int_{Sx} (x-\mu)^2 f(x)dx$

- $Var(X) = E(X^2)-(E(X))^2 = E(X^2) - \mu^2$
- Standard deviation σ = sqrt(Var(X))
- Linear transformation: $Var(aX+b) = a^2Var(X)$

➢ Standardisation: $Z = \frac{X-\mu}{\sigma}$

- $E(Z) = (1/\sigma)E(X) - \frac{\mu}{\sigma} = 0$
- $Var(Z) = (1/\sigma^2)Var(X) = \sigma^2/\sigma^2 = 1$

➢ Covariance: Covariance of to random variables X and Y ⇒ Cov(X,Y) = E((X-E(X))(Y-E(Y))

- $\mathbb{C}ov(X, Y) = \mathbb{C}ov(Y, X)$
- $\mathbb{C}ov(X, X) = \mathbb{V}ar(X)$
- $\boxed{\mathbb{C}ov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}$
- $\mathbb{C}ov(aX + b, cY + d) = ac\,\mathbb{C}ov(X, Y)$
- $\mathbb{C}ov(X_1 + X_2, Y_1 + Y_2)$
$= \mathbb{C}ov(X_1, Y_1) + \mathbb{C}ov(X_1, Y_2) + \mathbb{C}ov(X_2, Y_1) + \mathbb{C}ov(X_2, Y_2)$

- Correlation: $\rho = \frac{Cov(X,Y)}{\sqrt{Var(x)Var(y)}}$, correlation does not prove causation
Close p is to 1 = stronger the linear relationship

### Special random variables

➢ Binomial distribution: Outcome experiment is classified as either **success or failure** ⇒ Success has probability π with n independent repetitions of the experiment

- For X = number of successes, binomial random variable: **X ~ Bin(n, π)**
- Binomial PMF is given by $p(x) = \frac{n!}{x!(n-x)!} \times \pi^x (1-\pi)^{n-x}$ $for\ x = 0,1, \ldots, n$ n and x are binomial coefficients

- If n = 1, use Bernoulli distribution (X ~ Bern(π))

$$p(x) = \begin{cases} 1-\pi & \text{if } x = 0 \\ \pi & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

- $X_1 + X_2 \sim Bin(n_1 + n_2, \pi)$
- u = E(X) = nπ
- $\sigma^2$ = Var(x) = nπ(1-π)
- For X ~ Bin(20,0.1)

a) $\mathbb{P}(X = 2) = \binom{20}{2} 0.1^2 0.9^{18} = 0.2852$ Matlab: binopdf(2,20,0.1)

b) $\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) - \mathbb{P}(X = 3) - \mathbb{P}(X = 4)$ Matlab: 1-binocdf(4,10,0.1)

c) $\mathbb{P}(X > 10) = \mathbb{P}(X = 11) + \mathbb{P}(X = 12) + \ldots + \mathbb{P}(X = 20) = \ldots$

- Matlab for P(5<=Y<15) = binocdf(14,20,0.1)-binocdf(4,10,0.1)
- Matlab for P(5<Y<=15) = binocdf(15,10,0.1)-binocdf(5,10,0.1)

➢ Poisson distribution: Interest in **number of occurrences** of some random phenomenon in a fixed **period of time**

- For X = number of occurrences, Poisson random variable: **X ~ P(λ)**
- p(x) = $\frac{e^{-\lambda}\lambda^x}{x!}$ $for\ x = 0,1,2,3, \ldots$ ⇒ λ must satisfy λ>0
- E(X) = λ
- Var(X) = λ
- X~P(20)
  - P(X<=10) = poisscdf(10,20)
  - P(X<10) = poisscdf(9,20)
  - P(X>=10) = 1-poisscdf(9,20)

➢ Uniform distribution: **continuous distribution**

- A random variable is uniformly distribution over an interval [α,β]
  $X \sim U_{[\alpha,\beta]}$

if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } x \in [\alpha, \beta] \\ 0 & \text{otherwise} \end{cases} \qquad (\rightarrow S_X = [\alpha, \beta])$$

Constant density → $X$ is just as likely to be "close" to any value in $S_X$.

By integration, it is easy to show that

$$F(x) = \begin{cases} 0 & \text{if } x < \alpha \\ \frac{x-\alpha}{\beta-\alpha} & \text{if } \alpha \leq x \leq \beta \\ 1 & \text{if } x > \beta \end{cases}$$

- E(X) = (α+β)/2
- Var(X) = $(\beta-\alpha)^2/12$
- P(a < X < b) = (b − a)/(β − α)
- For $X \sim U_{[-1.1]}$
  - P(X<0) = P(X<=0) = unifcdf(0,-1,1)
  - P(−0.9 ≤ X ≤ 0.8) = P(X ≤ 0.8) − P(X < −0.9) = unifcdf(0.8,-1,1)-unifcdf(-0.9,-1,1)
  - the value of x such that P(−x ≤ X ≤ x) = 0.9 = unifinv(0.95,-1,1)

i) $\mathbb{P}(X > 0.2125) = 1 - \mathbb{P}(X \leq 0.2125)$

```
>> 1-unifcdf(0.2125,0.205,0.215)
ans =
0.2500
```

ii) What is $x$ to have $\mathbb{P}(X > x) = 0.1$? That means : $\mathbb{P}(X \leq x) = 0.9$

```
>> unifinv(0.9,0.205,0.215)
ans =
0.2140
```

⇝ $x = 0.214$ is the thickness exceeded by 10% of the wafers

- Exponential distribution: Interest in random amount of time **before the first** occurrence of a given phenomenon over a unit **period of time**
    - Random variable is an exponential random variable with parameter μ
      X ~ Exp(μ)

      if its probability density function is given by

      $$f(x) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad (\to S_X = \mathbb{R}^+)$$

      By integration, we find

      $$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\frac{x}{\mu}} & \text{if } x \geq 0 \end{cases}$$

    - E(X) = μ
    - Var(X) = $\mu^2$
    - For X ~ Exp(2)

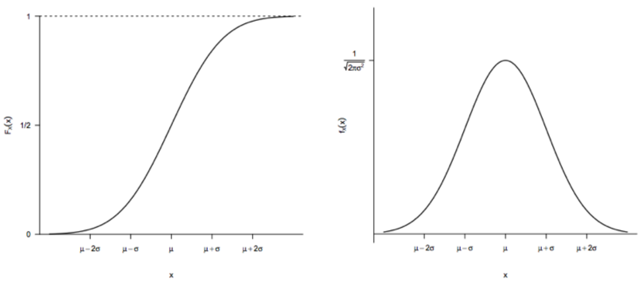
      i) $\mathbb{P}(W \leq 2)$
```
>> expcdf(2,2)
ans =
0.6321
```
      ii) $\mathbb{P}(W < 2) = \mathbb{P}(W \leq 2)$, as Exponential is a continuous distribution. For that matter, the MATLAB command is the same to compute $\mathbb{P}(W \leq w)$ and $\mathbb{P}(W < w)$ : `expcdf`

      iii) $\mathbb{P}(10 < W < 13) = \mathbb{P}(W < 13) - \mathbb{P}(W \leq 10)$
```
>> expcdf(13,2)-expcdf(10,2)
ans =
0.0052
```
      iv) $\mathbb{P}(W > -5) = 1 - \mathbb{P}(W \leq -5)$ must be 1, as an Exponential random variable can only assume non negative values. Indeed :
```
>> 1-expcdf(-5,2)
ans =
1
```
      For μ = 1/0.0003

      Find $x$ such that $\mathbb{P}(X > x) = 0.95$. That means that $\mathbb{P}(X \leq x) = 0.05$
```
>> expinv(0.05,1/0.0003)
ans =
170.9776
```
      $\rightsquigarrow$ 95% of the fans will last longer than 170.98 hours

- Normal distribution: random variable is normally distributed with parameters μ and σ: **X ~ N(μ,σ)**
    - Probability density function:

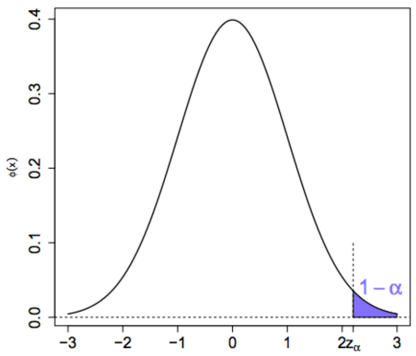
      $$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



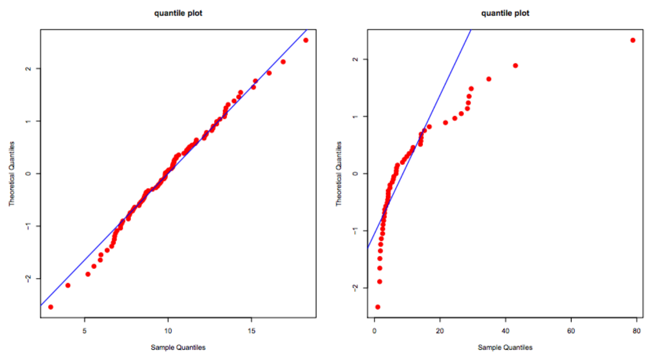

cdf $F(x)$ pdf $f(x) = F'(x)$

    - E(X) = μ
    - Var(X) = $\sigma^2$

- Standardisation: if X ~ N(μ,σ) then $Z = \frac{x-\mu}{\sigma} \sim N(0,1)$
  - This transforms X into a standard normal random variable Z
- P(x<=Z<y) in matlab: normcdf(y)-normcdf(x)
  - For P(X<=x) = P((x-μ)/σ)
  - For y = P(X<=x) = P(Z<=(x-μ)/σ)
    ⇒ norminv(y) = z
    (x-μ)/σ = z ⇒ can solve for μ
- Quantiles: P(Z>$z_a$) = 1 - α or P(Z<$z_a$) = α




- Checking normal distribution
  - Can check density histogram follows a bell-shaped curve
  - Quantile plots: More reliable for smaller sample sizes (matlab: qqplot(data))
    - Compares the data ordered from smallest to largest, if the sample comes from a normal distribution, the points should follow approximately a straight line.




→ the normality assumption appears acceptable for the first data set, not at all for the second

**Inferences concerning a mean**

- Point estimation: An estimator y is a random variable, which has its mean, variance and probability distribution, known as sampling distribution.
  - To estimate population mean μ, use sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
    This derives to E(X) = μ and Var(X) = σ²/n
  - $X_i \sim N(\mu,\sigma/n^{1/2})$
- Properties of estimators: An estimator Y of y is said to be unbiased if an only if it means is equal to y such that E(Y) = y
  - To say its unbiased means on the average, its values will equal the parameter its supposed to estimate.

- If not biased, E(Y)-y is called the biased estimator (systematic error)
- Estimators with smaller variances are more likely to produce estimates close to the true value y.
  - Good way to check is to show variance decreases to 0 as n increases.

➢ Standard error: the standard error of an estimator Y is its standard deviation sd(Y)
- Standard error of sample mean: from E(X) and var(X), sd(X) = $\sigma/n^{1/2}$
- Sample standard deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - Where sd(X) = $s/n^{1/2}$

➢ Confidence intervals: Interval for which we can assert a reasonable degree of certainty that will contain the true value of the proportion under consideration.
- Short interval implies precise estimation, wide interval shows there is a great deal of uncertainty concerning the parameter we are estimating.
- At a confidence level of 100*(1-α)% ⇒ we are that percent confident that our true value of the parameter is included into the interval [a,b]

➢ Confidence interval on the mean of a normal distribution with variance known:
- If we desire a confidence interval for μ of level 100*(1-α)% from a random sample: P(L<=μ<=U)= 1-α
  - Since Z ~ N(0,1): P($-z_{1-\alpha/2} \le \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le z_{1-\alpha/2}$) = 1-α
- Two sided Confidence interval of level 100*(1-α)% for μ is given by:

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- If one sided, then $z_{1-\alpha}$
- Matlab: [mean(x)-norminv(1-α/2)*(std(x)/sqrt(n)), mean(x)+norminv(1-α/2)*(std(x)/sqrt(n))]
- Since the error e = $|\bar{x}-\mu|$ is less than $z_{1-\alpha/2}*\sigma/n^{1/2}$ then sampling size is

$$n = \left(\frac{z_{1-\frac{\alpha}{2}}\sigma}{e}\right)^2$$

➢ Central limit theorem: Asserts that the sum of a large number of independent random variables has a distribution that is approximately normal
- If $X_1, X_2, \ldots, X_n$ is a random sample taken from a population with mean μ and finite difference $\sigma^2$ and given $\bar{x}$ is the sample mean, then limiting distribution is: $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ as n goes to infinity, is the standard normal distribution
- $X_i \sim \mathrm{Exp}(\lambda) \Rightarrow (\mu = 1/\lambda, \sigma = 1/\lambda) \Rightarrow \frac{\sqrt{n}\left(\bar{X}-\frac{1}{\lambda}\right)}{\frac{1}{\lambda}} \sim N(0,1)$
- $X_i \sim U_{[a,b]} \Rightarrow (\mu = (a+b)/2, \sigma = (b-a)/12^{1/2}) \Rightarrow \frac{\sqrt{n}\left(\bar{X}-\frac{a+b}{2}\right)}{\frac{b-a}{\sqrt{12}}} \sim N(0,1)$
- $X_i \sim \mathrm{Bern}(\pi) \Rightarrow (\mu = \pi, \sigma = \sqrt{\pi(1-\pi)}) \Rightarrow \frac{\sqrt{n}(\bar{X}-\pi)}{\sqrt{\pi(1-\pi)}} \sim N(0,1)$
- The larger the n, the better the normal approximation
- The closer the population distribution is to being normal, the more rapidly the distribution of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ approaches normality as n gets large.

- The normal approximation is valid whenever sample size n >= 30

- Student t-distribution: Random variable T, from a normal population that follows students t-distribution with ν degrees of freedom such that **$T \sim t_\nu$**
  - $E(T) = 0$ and $Var(T) = \nu/(\nu-2)$ for $\nu>2$
  - Student's t-distribution has a heavier tail that a normal distribution
  - Quantiles: critical t value $\Rightarrow t_{\nu;1-\alpha} = -t_{\nu;\alpha}$
  - For n >= 2, $T = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$ and thus:
  - $P(-t_{n-1;1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq t_{n-1;1-\alpha/2} = 1-\alpha$
  - Confidence interval:

$$\boxed{\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]}$$

  - Matlab: [mean(x)-tinv(1-α/2.n-1)*(std(x)/sqrt(n)), mean(x)+yinv(1-α/2,n-1)*(std(x)/sqrt(n))]

- Predicting future observation: Predicting an $X_{n+1}$ value for a single future observation where $X^*$ is a statistic of the predictor $X_{n+1}$
  - $E(X_{n+1}-X^*) = 0$ and $Var(X_{n+1}-X^*) = \sigma^2(1+1/n)$
  - $Z = \frac{X_{n+1}-\bar{X}}{\sigma\sqrt{1+\frac{1}{n}}} \sim N(0,1)$ and $T = \frac{X_{n+1}-\bar{X}}{S\sqrt{1+\frac{1}{n}}} \sim t_{n-1}$
  - Confidence intervals:

$$\left[\bar{x} - z_{1-\alpha/2}\,\sigma\sqrt{1+\tfrac{1}{n}}, \bar{x} + z_{1-\alpha/2}\,\sigma\sqrt{1+\tfrac{1}{n}}\right]$$

$$\left[\bar{x} - t_{n-1;1-\alpha/2}\,s\sqrt{1+\tfrac{1}{n}}, \bar{x} + t_{n-1;1-\alpha/2}\,s\sqrt{1+\tfrac{1}{n}}\right]$$

- Inferences concerning proportions: Focuses around π, the proportion of the population that has characteristic of interest. **$X \sim Bern(\pi)$**
  - Sample proportion: $\hat{p} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - $E(\hat{p}) = \pi$ and $Var(\hat{p}) = \pi(1-n)/n$
  - $\frac{\sqrt{n}(\hat{p}-\pi)}{\sqrt{\pi(1-\pi)}} \sim N(0,1) \Rightarrow P(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\hat{p}-\pi)}{\sqrt{\pi(1-\pi)}} \leq z_{1-\alpha/2} = 1-\alpha$
  - Confidence interval:

$$\boxed{\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]}$$

  - Choice of sample size:

$$n = \left(\frac{Z_{1-\frac{\alpha}{2}}}{2e}\right)^2$$

- Hypothesis testing: Requires the decision of which 2 parameters are true
  - Null hypothesis $H_0$: μ = x ⇒ we assume this is true unless we have enough evidence otherwise and thus:
  - Alternative hypothesis: could be $H_a$: μ≠x (two sided alternatives) or $H_a$: μ>x or μ<x (one sided alternatives)
  - Errors: P(type 1 error): P(reject $H_0$ when it is true) = α (reduce by increasing acceptance region and increase significance level α)

P(type II eror): P(fail to reject $H_0$ when it is false) = $\beta$ (reduce by opposite of above)

- If population follow a normal distribution with known standard deviation $\sigma$, at significance level $\alpha$, the decision rule is:

$$\text{reject } H_0 \text{ if } \bar{x} \notin \left[\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

  - We can tolerate being wrong at $\alpha$% of most cases

➢ P-value: Smallest level of significance that would lead to rejection of $H_0$ with the observed sample – the probability that the test statistic will take on a value that is at least extreme as the observed value when $H_0$ is true

- When testing $H_0$: $\mu = \mu_0$ against $H_a$: $\mu \neq \mu_0$ then need $z_0$ as z-score
  $z_0 = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$
  p = $2*(1-\phi(|z_0|)$: $2*(1\text{-normcdf}(z_0))$ (matlab)
- If $p < \alpha$ then reject $H_0$, if $p >= \alpha$ then do not reject $H_0$

1. State the null and alternative hypotheses: $H_0$ and $H_a$
2. Determine the rejection criterion
3. Compute the appropriate test statistic and determine its distribution
4. Calculate the $p$-value using the test statistics computed
5. Conclusion: reject/do not reject $H_0$, relate back to the research question

➢ One sided alternatives: When testing $H_a$: $\mu > \mu_0$, we reject $H_0$ if $\bar{X} > \mu_0 + z_{1-\alpha/2}*(\sigma/n^{1/2})$

- p = $1-\phi(z_0)$ (1-normcdf($z_0$))

➢ Unknown standard deviation: Now work with $t_0 = \frac{\sqrt{n}(\bar{X}-\mu_0)}{s}$

- For a two sided hypothesis test, rejection criterion:

$$\text{reject } H_0 \text{ if } \bar{x} \notin \left[\mu_0 - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \mu_0 + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

- p = $2*P(T>|t_0|)$ ($2*(1\text{-tcdf}(t_0))$)
- For one sided alternatives, rejection criterion:

$$\text{reject } H_0 \text{ if } \bar{x} > \mu_0 + t_{n-1,1-\alpha}\frac{s}{\sqrt{n}} \quad \text{or} \quad \text{reject } H_0 \text{ if } \bar{x} < \mu_0 - t_{n-1,1-\alpha}\frac{s}{\sqrt{n}}$$

and the associated $p$-values are

$$p = \mathbb{P}(T > t_0) \quad \text{or} \quad p = \mathbb{P}(T < t_0)$$

➢ Hypothesis tests for a proportion: Consider two sided hypothesis test where $H_0$: $\pi = \pi_0$ against $H_a$: $\pi \neq \pi_0$, rejection criterion:

$$\text{reject } H_0 \text{ if } \hat{p} \notin \left[\pi_0 - z_{1-\alpha/2}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}, \pi_0 + z_{1-\alpha/2}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}\right]$$

- $z_0 = \frac{\sqrt{n}(\hat{p}-\pi_0)}{\sqrt{\pi_0(1-\pi_0)}} \Rightarrow p = 2*(1-\phi(|z_0|)$

- For a one sided test for $H_0$: $\pi=\pi_0$ against $H_a$: $\pi>\pi_0$ or $H_a$: $\pi<\pi_0$, then

$$\text{reject } H_0 \text{ if } \hat{p} > \pi_0 + z_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

or

$$\text{reject } H_0 \text{ if } \hat{p} < \pi_0 - z_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

will have **approximate significance level** $\alpha$.

The associated **approximate *p*-values** will be

$$p = 1 - \Phi(z_0) \quad \text{or} \quad p = \Phi(z_0)$$

**Inferences concerning a difference of means**

- Based around interest in comparing two different populations assuming the two samples studied are independent
- Hypothesis test: Null hypothesis will be $H_0$: $\mu_1 = \mu_2$
  $H_a$: $\mu_1 \neq \mu_2$ (two sided alternative) or
  $H_a$: $\mu_1 > \mu_2$ or $H_a$: $\mu_1 < \mu_2$ (one sided alternative)
  - Sampling distribution:

    $$\boxed{\bar{X}_1 - \bar{X}_2 \overset{(a)}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)}$$

  - Rejection criterion:

    $$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 \notin \left[-z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

  - $Z_0 = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \Rightarrow p = 2*(1-\phi(|z_0|)$
  - For one sided alternatives:

    Similarly, for the one-sided test with alternative $H_1 : \mu_1 > \mu_2$, the decision rule is

    $$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 > z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

    and the associated *p*-value is

    $$p = 1 - \Phi(z_0),$$

    while for the one-sided test with alternative $H_1 : \mu_1 < \mu_2$, the decision rule is

    $$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 < -z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

    and the associated *p*-value is

    $$p = \Phi(z_0)$$

  - Confidence interval for $\mu_1$-$\mu_2$:

    $$\left[(\bar{x}_1 - \bar{x}_2) - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

- Hypothesis test for $\mu_1=\mu_2$ ($\sigma_1 = \sigma_2$) we have:

$$\boxed{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{(a)}{\sim} t_{n_1+n_2-2}}$$

Where $S_p$ is the pooled standard deviation: $S_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

reject $H_0 : \mu_1 = \mu_2$ if

$$\bar{x}_1 - \bar{x}_2 \notin \left[-t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\, t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$

- $t_0 = \frac{(\bar{X}_1-\bar{X}_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \Rightarrow p = 2*(T>|t_0|)$ where $T \sim t_{n1+n2-2}$

One-sided versions of this test are also available. For the alternative $H_a : \mu_1 > \mu_2$, the decision rule is

$$\text{reject } H_0 : \mu_1 = \mu_2 \text{ if } \bar{x}_1 - \bar{x}_2 > t_{n_1+n_2-2;1-\alpha}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

and the associated $p$-value is

$$p = 1 - \mathbb{P}(T < t_0),$$

whereas for the alternative $H_a : \mu_1 < \mu_2$, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 < -t_{n_1+n_2-2;1-\alpha}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

and the associated $p$-value is

$$p = \mathbb{P}(T < t_0)$$

- Two sided confidence interval for $\mu_1$-$\mu_2$:

$$[(\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, [(\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2;1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}]$$

while two $100 \times (1-\alpha)\%$ one-sided confidence intervals for $\mu_1 - \mu_2$ are

$$\left(-\infty, (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\alpha} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$

and

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\alpha} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, +\infty\right)$$

- Hypothesis test for $\mu_1$=$\mu_2$ ($\sigma_1 \neq \sigma_2$) we have:

$$\boxed{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \overset{a}{\sim} t_\nu}$$

  - Degrees of freedom is now:

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

  - Confidence interval:

$$\left[\bar{x}_1 - \bar{x}_2 \pm t_{\nu;1-\alpha/2} \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

- Paired observations: Used to deal with "before and after" data
  - Consider differences $D_i = X_{i1}-X_{i2} = Y_{i1} - Y_{i2}$
    $\mu_D = \mu_1 - \mu_2$
  - Hypothesis test: $H_0$: $\mu_1 = \mu_2$ against $H_a$: $\mu_1>\mu_2$
  - Need to take sample difference and then find the mean $\bar{d}$ and the sample standard deviation s, then the rejection criterion is:

    reject $H_0$ if $\bar{d} > t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}$

  - The p-value is: $P(T>|t_0|)$ where $t_0 = \frac{\sqrt{n}\bar{d}}{s}$

## Regression Analysis

- Simple linear regression model: If points lie randomly around a straight line, it is reasonable to assume X and Y are related
  - Regression Model: $Y = \beta_0 + \beta_1 X + \varepsilon$, where the slope $\beta_1$ and intercept $\beta_0$ are regression coefficients.
  - $E(\varepsilon) = 0 \Rightarrow \mu_Y = \beta_0+\beta_1 X$ and $Var(\varepsilon) = \sigma^2$
  - Random error is normally distributed: $\varepsilon \sim N(0,\sigma)$, $Y|(X=x) \sim N(\beta_0+\beta_1 X,\sigma)$
- Least squares estimators
  - $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$ $(= \sum_{i=1}^{n} X_i^2 - \frac{(\sum_i X_i)^2}{n})$
  - $S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})$ $(= \sum_{i=1}^{n} X_i^2 - \frac{(\sum_i X_i)^1(\sum_i Y_i)^1}{n})$
  - $\beta_1 = S_{XX}/S_{XY}$ and $\beta_0 = \bar{Y} - \frac{S_{XX}}{S_{XY}}\bar{X}$
  - $\bar{X} = \frac{\sum_i X_i}{n}$
- Estimating $\sigma^2$: $\varepsilon = Y-(\beta_0+\beta_1 X)$
  - Residuals of fitted model:

    $$\bar{e} = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = \bar{y} - (\hat{b}_0 + \hat{b}_1\bar{x}) = 0$$

  - Number of degrees of freedom is now n-2 since there are 2 parameters
  - Unbiased estimated of $\sigma^2$:

    $$\boxed{s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2}$$

    $$\Rightarrow \quad S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

- Inferences concerning $\beta_1$: Need to consider that $\beta_1 = 0$ as a hypothesis (it does not depend on the predictor X)
  - $H_0$: $\beta_1 = 0$ against $H_a$: $\beta_1 \neq 0$ where $\sqrt{S_{xx}}\frac{\widehat{\beta_1}-\beta_1}{S} \sim t_{n-2}$
  - Rejection criterion:

    reject $H_0$ if $\hat{b}_1 \notin \left[-t_{n-2,1-\alpha/2}\frac{s}{\sqrt{S_{xx}}}, t_{n-2,1-\alpha/2}\frac{s}{\sqrt{S_{xx}}}\right]$

  - $p = 2*P(T>|t_0|) \Rightarrow t_0 = \sqrt{S_{xx}}\frac{\widehat{b_1}}{s}$

- Two sided Confidence interval for the true slope $\beta_1$:

$$\left[\hat{b}_1 - t_{n-2;1-\alpha/2}\frac{s}{\sqrt{S_{xx}}}, \hat{b}_1 + t_{n-2;1-\alpha/2}\frac{s}{\sqrt{S_{xx}}}\right]$$

➢ Inferences concerning $\beta_0$

- Two sided confidence interval:

$$\left[\hat{b}_0 - t_{n-2;1-\alpha/2} s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}, \hat{b}_0 + t_{n-2;1-\alpha/2} s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\right]$$

- Rejection criterion for the same hypothesis test:

$$\text{reject } H_0 \text{ if } \hat{b}_0 \notin \left[-t_{n-2,1-\alpha/2} s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}, t_{n-2,1-\alpha/2} s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\right]$$

$$t_0 = \frac{\hat{b}_0}{s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}} \quad \rightarrow p = 2 \times \mathbb{P}(T > |t_0|), \quad T \sim t_{n-2}$$

➢ Computer Output

```
Regression Analysis:  Y versus X

The regression equation is Y = 74.283 + 14.947 X

 Predictor      Coef   SE Coef      T      P
 Constant     74.283     1.593  46.62  0.000
 X            14.947     1.317  11.35  0.000

S = 1.087 R-Sq = 87.74% R-Sq(adj) = 87.06%
```

- First Column 'Coef' is coefficients $b_0$ and $b_1$
- Second Column 'SE Coef' Is standard errors $s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$ *and* $\frac{s}{\sqrt{S_{xx}}}$
- Third Column – T is observed $t_0$ values of test statistic
- Fourth column – p gives associated p values
- S = estimate s of $\sigma$

➢ Confidence interval on mean response: Specified at a specific value X

- Standardising unknown $\sigma$ as S: $\frac{\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}}{s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}$
- From y(x) $=b_0 + b_1x \Rightarrow$ Two sided confidence interval on the mean response Y

$$\left[\hat{y}(x) - t_{n-2;1-\alpha/2} s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}, \hat{y}(x) + t_{n-2;1-\alpha/2} s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}\right]$$

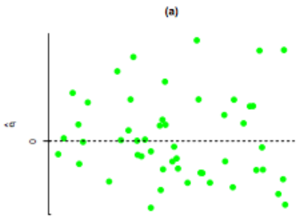
➢ Predicting new observations: Given by $Y^*(X) = \beta_0+\beta_1 x$

- Standardisation gives: $\frac{\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}}{s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}$
- Prediction interval:

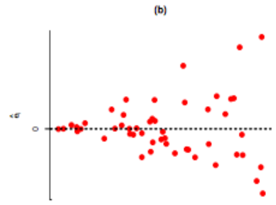
$$\left[\hat{y}(x) - t_{n-2;1-\alpha/2} s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}, \hat{y}(x) + t_{n-2;1-\alpha/2} s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}\right]$$

- CI does not tend to 0 as n increases $\Rightarrow$ Inherent variability of the new observation never vanishes.

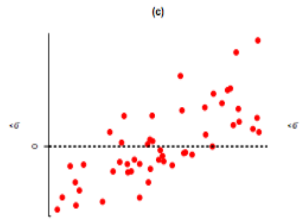
- Residual analysis: The observed residuals $e_i = y_i - y(x_i) = y_i - (b_0 + b_1 x_i)$
  - Plot the residuals in time sequence/against fitted values/against predictor values



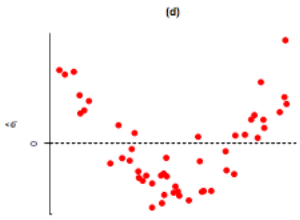

**Represents ideal solution**




**Variability changes with x and y**




**Dependence in error terms**




**Nonlinear probabilistic model**

  - Normality assumption is checked with qqplot
- Coefficient of determination: Compare $ss_r$ to $ss_t$ to judge model adequacy
  - $r^2$ = coefficient of determination: $r^2 = ss_r/ss_t$, where $ss_t = ss_r + ss_e$

1. because the $x_i$ values are different, all $Y_i$ have different means. This variability is quantified by the '**regression sum of squares**':

$$ss_r = \sum_{i=1}^{n} (\hat{y}(x_i) - \bar{y})^2$$

2. each value $Y_i$ has variance $\sigma^2$ around its mean. This variability is quantified by the '**error sum of squares**':

$$ss_e = \sum_{i=1}^{n} (y_i - \hat{y}(x_i))^2 = \sum_{i=1}^{n} \hat{e}_i^2$$

  - $r^2$ represents proportion of variability in the responses that is explained by the predictor $\Rightarrow$ taken into account in the model
  - $r^2$ near 1 = good fit to data , $r^2$ near 0 = poor fit to data
- Correlation: : $\rho = \frac{Cov(X,Y)}{\sqrt{Var(x)Var(y)}} = \frac{E((X-E(X))(Y-E(Y))}{\sqrt{E((X-E(X))^2)E((Y-E(Y))^2)}}$ used to quantify strength of linear relationship between X and Y
  - ρ close to 1 or -1 = strong linear relationship

- From $r^2 = ss_r/ss_t = s^2_{xy}/s_{xx}s_{yy}$

$$\boxed{r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}}$$

- This is the sample correlation coefficient which can be regarded as sample estimate of the proportion correlation coefficient $\rho$
- Correlation does not prove causation.

**ANOVA (analysis of variance)**

- Comparing more than 2 populations
- Use boxplots to show variability of observations **within** and **between** groups
- Each group is called a treatment that is denoted by k independent samples.
- Grand mean:

$$\bar{\bar{X}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n}$$

- ANOVA model: $X_{ij} = \mu_i + \varepsilon_{ij}$
  - $\mu_I$ = mean response for ith treatment
  - $\varepsilon_{ij}$ = individual random error component
- ANOVA hypothesis: $H_0$: $\mu_1=\mu_2=\ldots=\mu_k$ against $H_a$: not all the means are equal
- Variability decomposition
  - Total sum of squares: $SS_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{\bar{X}}\right)^2 \quad (df = n-1)$ $\Rightarrow$ Quantifies total amount of variation contained in the global sample.
  - Treatment sum of squares (variability between groups): $SS_{Tr} = \sum_{i=1}^{k} n_i\left(\bar{X}_i - \bar{\bar{X}}\right)^2 \quad (df = k-1)$
  - Error sum of squares (variability within groups): $SS_{Er} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2 \quad (df = n-k)$
  - Sum of squares identity: $SS_{Tot} = SS_{Tr} + SS_{Er}$
- Mean squared error: $MS_{Er} = \frac{SS_{Er}}{n-k}$ $\Rightarrow$ Always estimates $\sigma^2$
- Treatment mean square: $MS_{Tr} = \frac{SS_{Tr}}{k-1}$ $\Rightarrow$ Estimates $\sigma^2$ only when $H_0$ is true
- Fisher F-distribution: $F \sim F_{k-1,n-k}$
  - It can be shown that $H_0$ Is true from the ratio $F = \frac{MS_{Tr}}{MS_{Er}} = \frac{\frac{SS_{Tr}}{k-1}}{\frac{SS_{Er}}{n-k}}$
  - Rejection criterion:

$$\text{reject } H_0 \text{ if } \frac{ms_{Tr}}{ms_{Er}} > f_{k-1,n-k;1-\alpha}$$ $\Rightarrow$ Matlab: finv(1-α,k-1,n-1)

  - p-value: $p = P(X>f_0)$ where $f_0 = ms_{Tr}/ms_{Er}$ $\Rightarrow$ Matlab: fcdf($f_0$,k-1,n-1)

- ANVOA table:

| Source | degrees of freedom | sum of squares | mean square | $F$-statistic |
|---|---|---|---|---|
| Treatment | $df_{Tr} = k - 1$ | $ss_{Tr}$ | $ms_{Tr} = \frac{ss_{Tr}}{k-1}$ | $f_0 = \frac{ms_{Tr}}{ms_{Er}}$ |
| Error | $df_{Er} = n - k$ | $ss_{Er}$ | $ms_{Er} = \frac{ss_{Er}}{n-k}$ | |
| Total | $df_{Tot} = n - 1$ | $ss_{Tot}$ | | |

Note 1: $df_{Tot} = df_{Tr} + df_{Er}$ and $ss_{Tot} = ss_{Tr} + ss_{Er}$

Note 2: this table is the usual computer output when an ANOVA procedure is run

- Confidence intervals on the treatment of means: Interest about which $\mu_I$'s are different from each other
  - As $MS_{Er}$ is an unbiased estimator for $\sigma^2$ with n-k

$$\sqrt{n_i}\frac{\bar{X}_i - \mu_i}{\sqrt{MS_{Er}}} \sim t_{n-k}$$

  - Two sided confidence interval for $\mu_i$, from observed values $x_i$ and $MS_{Er}$

$$\left[\bar{x}_i - t_{n-k,1-\alpha/2}\sqrt{\frac{MS_{Er}}{n_i}}, \bar{x}_i + t_{n-k,1-\alpha/2}\sqrt{\frac{MS_{Er}}{n_i}}\right]$$

  - Confidence intervals for each group will tell which values $\mu_i$'s are much different from one another and which are close
- Pairwise comparisons: Confidence intervals on the difference of two means $\mu_i$ and $\mu_j$

$$\left[(\bar{x}_i - \bar{x}_j) - t_{n-k;1-\alpha/2}\sqrt{MS_{Er}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},\right.$$
$$\left.(\bar{x}_i - \bar{x}_j) + t_{n-k;1-\alpha/2}\sqrt{MS_{Er}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}\right]$$

  - If 0 is contained in the found interval, then conclude that $\mu_i$ and $\mu_j$ are not significantly different. Vice versa in there is no 0 in the interval.
  - Bonferonni adjustments: Gives an overall significance level $\alpha$, where pairwise comparison tests are carried out at significance level $\alpha$/K%, where K = $\binom{k}{2} = \frac{k!}{2!(k-2)!}$
  - Bonferonni-adjusted t-test: Testing $H_0$: $\mu_i = \mu_j$ againt $H_a$: $\mu_i \neq \mu_j$
    test statistic: $t_0 = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MS_{Er}(\frac{1}{n_i} + \frac{1}{n_j})}}$
  - p-value: p = 2*P(T>|$t_0$|), T ~ $t_{n-k}$ $\Rightarrow$ reject $H_0$ if p value is less than $\alpha$/K, where K is the number of pairwise comparisons
- Adequacy of ANOVA model: Based on several assumptions that should be carefully checked
  1. Normality: check by constructing a normal quantile plot for residuals.
  2. Equal variances in each group: checked by plotting residuals against treatment level $\Rightarrow$ Spread of residuals should not depend on $\bar{x}_i$
     - Rule of thumb: if ratio of largest standard deviation to smallest standard deviation is less than 2 $\Rightarrow$ validate equal variance
  3. Independence assumption: Checked by plotting residuals against time (if available) $\Rightarrow$ No pattern (negative or positive sequences) should be observed. Generally no valid way of checking this.